UNIVERSITÉ DE STRASBOURG

Institut de Génétique et de Biologie Moléculaire et Cellulaire

Département de Biologie et de Génomique Structurale

Laboratoire de Bioinformatique et de Génomique Intégratives

Laboratoire de l'Expression de l'Information Génétique

Thèse présentée pour obtenir le titre de

Docteur de l'Université de Strasbourg

Discipline : Sciences du Vivant et de la Santé

Spécialité : Bioinformatique Structurale

par

Laurent-Philippe Albou

Analyse intégrative des données structurales et reconnaissance de forme : application à la régulation de la transcription eucaryote

Soutenue publiquement le 29 Septembre 2010 devant le jury :

Directeur de Thèse :	**Dino MORAS**, IGBMC, Illkirch
Co-Directeur de Thèse :	**Olivier POCH**, IGBMC, Illkirch
Rapporteur interne :	**Marcel HIBERT**, Faculté de Pharmacie, Illkirch
Rapporteur externe :	**François GROS**, Académie des Sciences, Paris
Rapporteur externe :	**Joël JANIN**, IBBMC, Orsay
Examinateur :	**Daniel CAPUT**, Sanofi-Aventis R&D, Labege

à ma famille,

à ceux qui m'ont formé,

Remerciements

Avant tout, je souhaite exprimer ma profonde reconnaissance à Marcel Hibert, François Gros, Joël Janin et Daniel Caput pour avoir accepté de juger mon travail de thèse et d'avoir fait preuve de beaucoup de compréhension pour le retard avec lequel je leur remets ce manuscrit.

Je tiens également à remercier chaleureusement Dino Moras et Olivier Poch pour m'avoir accueilli au sein des équipes de l'Expression de l'Information Génétique, et de Bioinformatique et de Génomique Intégratives. Ces quatre années passées à l'IGBMC pour cette thèse m'auront beaucoup appris, aussi bien dans le domaine scientifique qu'humain. Je vous remercie tout particulièrement pour l'autonomie et la confiance que vous m'avez témoignées, et ce, malgré mes défauts et le projet d'entreprise dans lequel je me suis lancé dès 2008 et qui a pris du temps sur mes travaux. Je suis honoré d'avoir pu travailler à vos côtés, et j'espère que ce manuscrit et les résultats qui y sont présentés témoigneront de mon estime sincère. Bien que cela ait retardé la remise du manuscrit de quelques jours, je suis heureux d'avoir pu intégrer les « arbres fonctionnels » des récepteurs nucléaires, relatifs à leur site de dimérisation et à leur site de liaison au cofacteur peptidique. Cette thèse aurait été incomplète sans eux, et je pense que vous partagerez ce sentiment. J'espère que nous aurons l'opportunité dans l'année à venir d'affiner ces arbres, et de tester certaines des hypothèses qu'ils impliquent.

Merci à toi, Olivier, d'avoir été là, de m'avoir soutenu et de m'avoir fait profiter de tes connaissances sur des systèmes biologiques d'intérêts, permettant d'illustrer ou d'aider à la validation de certaines de mes approches. Merci aussi pour ta tolérance, car je sais qu'à de nombreuses reprises, le projet d'entreprise et mon absence t'ont été difficiles. J'espère que dans les mois à venir, nous aurons l'occasion d'avancer sur l'étude structurale des protéines et SNP impliquées dans des maladies monogéniques, et aussi, si le temps nous le permet, d'avancer sur Alvinella. Cependant, je ne suis pas certain de l'utilité de créer un être humain thermotolérant capable de vivre sur Mars… (comprendra qui peut !). Et finalement, bien que plus générale, mais tout aussi important, merci d'avoir vécu ton rêve de Bioinformaticien, d'avoir créé cette équipe et de m'avoir permis d'y travailler et de réaliser à mon tour mes rêves de recherche.

Merci aussi, à tous ceux qui m'ont formé, au cours de ces longues années d'études… qui, paraît-il, continuent tout au long de la vie du chercheur ? J'ai une pensée notamment pour les Professeurs de Pierre et Marie Curie, à Paris VI, dont le cursus exceptionnel de Biochimie m'avait permis de prendre conscience des aspects chimiques fondamentaux du Vivant. Bien

entendu, ces remerciements seraient incomplets sans une pensée toute particulière pour l'équipe de Bioinformatique et de Biostatistique d'Orsay, et de ses Directeurs, Michel Termier et Alain Denise. Merci Professeur, pour cette « main protectrice » que vous avez posée au-dessus de moi pour me protéger de mes erreurs. Vous voyez, l'étude des structures moléculaires peut effectivement servir à quelque chose, malgré les suspicions de certains… ! Nous n'en sommes peut-être pas encore à expliquer l'apparition de mutations dans le génome par des instabilités chimiques et structurales, mais cela viendra peut-être en son temps ?

Je dois également beaucoup à Joël Janin que je rencontrai lors d'un cours de Master II et qui me permit de faire mon stage de fin d'étude à ses côtés. Je me souviens qu'à l'époque, je n'étais principalement intéressé que par la stabilité des structures moléculaires, et les mécanismes qui permettent leurs repliements. J'avais alors encore du mal à imaginer que l'on puisse étudier les interactions moléculaires sans comprendre avant, les forces qui gouvernent leurs repliements. Vous m'avez montré, avec R. Bahadur, qu'il était non seulement possible, mais nécessaire, de les étudier, et que, malgré des données limitées (mais qui n'ont cessé de croître rapidement !), il était possible de mieux appréhender les mécanismes responsables de la reconnaissance moléculaire. Vous m'avez permis de chercher, et j'espère pouvoir le dire, de « trouver » un sens aux structures moléculaires, en élargissant mes connaissances et le champ des problématiques liées à ces structures. A toi Joël, merci pour tout.

Aussi, que serais-je sans ma famille et sans François Gros ? Merci à toi, mon père, pour m'avoir guidé tout au long de ces années et de m'avoir toujours fait confiance, tout en me laissant la liberté de choisir entre la Science et d'autres métiers ; mon cœur est à la Science, pour autant que je puisse lui apporter quelque chose. Merci à toi, ma mère, pour ta sensibilité et pour ton imagination qui ont sans doute coloré ma vision des Sciences et m'ont permis de ne pas vivre qu'en « raisonnant ». Merci à mes grands parents, disparus au cours de ma thèse et peu de temps avant, pour m'avoir montré les valeurs et morales d'un autre temps. Merci à mes frères, qui, par leurs plaisanteries constantes, me permettent de ne pas toujours me prendre trop au sérieux. Merci à toi, François, que je rencontrai un jour à l'Académie en accompagnant mon père, et qui joua tout au long de mes études et jusqu'à maintenant, un rôle de parrain et d'ami. Merci de m'avoir conforté dès les débuts, dans mon souhait de faire de la Bioinformatique, et d'avoir cru en son importance croissante pour l'étude du Vivant, ce alors même que des cursus universitaires permettant d'étudier cette discipline nouvelle, n'existaient pas encore en France.

Enfin, merci à tous mes amis, qui m'ont permis de me construire au fil des ans, et qui ont su être présents lorsque j'en avais besoin. Je sais que, pour certains d'entre vous, ce travail

nous aura éloignés un temps ; j'espère cependant que, lorsque je reviendrais, tout sourire, sonnant à votre porte, vous saurez me reconnaître malgré les années qui ont passé. J'ai une pensée notamment pour François K., cet ami qui me connaît depuis mon enfance et m'a vu grandir comme un frère. Je t'ai toujours regardé avec fierté, et si tu es à présent professeur des lycées en Biologie, j'espère qu'un jour tu poursuivras ton rêve premier de faire à ton tour de la Recherche.

Et bien entendu… Car certains d'entre vous à la lecture de ces remerciements se demanderont où ils se trouvent… ! Merci à tous les membres de l'équipe de Bioinformatique, actuelle comme passée, pour avoir été là. Merci en particulier à Nicolas G., David K., Guillaume B., Yann B., Anne F., Hoan H., Julie T., Benjamin S., Benjamin L. et Yannick N. d'avoir été là et d'avoir contribué à la bonne ambiance du laboratoire. Je pense que nous nous souviendrons tous de cette époque où nous arrivions à concilier à la fois la Recherche et une humeur joviale. Je pense notamment à ces réunions de couloir (et quel couloir !), et à ces moments où avec Nicolas et David, nous poussions la chansonnette sur des paroles de Poch (mais non Olivier, ce n'est pas toi !). Merci aussi à celle qui renverse les cafés… d'avoir été là et d'avoir contribué aux mystères planant dans le laboratoire.

Merci à Serge A., Jean-Christophe A., Jean-Paul B., Jean-Luc D., Magalie L. et Richard B., pour m'avoir permis de créer l'entreprise BIONEXT et d'avoir aidé à son développement. Le nouveau-né est à présent lauréat du Concours National de Création d'Entreprises Innovantes, les projets et développements sont là. Je souhaite que nous parvenions à réaliser nos espérances et à donner à cette entreprise une image unique, qui pourrait concilier une bonne Recherche à de bonnes applications, et ce, sans que l'une nuise à l'autre. Le temps nous le dira.

Enfin, je finirai ces remerciements par celle qui a choisi de me rejoindre il y a deux ans, rencontrée un soir sur une plage de Croatie. Merci Rania de m'avoir supporté dans ces moments d'acharnement sur ma thèse et lors de la création et du développement de l'entreprise… Merci pour ton affection, ta compréhension, pour le courage et l'honnêteté que tu mets dans ton propre travail. Merci à Elisabeth Z. qui t'aide dans ce sens et te permet de te réaliser et de prouver avec elle, ta valeur. Je sais que je ne semble pas être en toute chose, une personne modérée ; pourquoi faire quelque chose en une semaine quand on pourrait le faire en travaillant 24h d'affilée ? Mais je tente d'apprendre avec toi cette vertu : la modération. Comme le répète souvent mon père : « *qui va piano va sano* ».

Dans cette thèse, je tends à montrer qu'une molécule seule n'a pas de fonction, et que ce sont ses interactions avec les autres molécules de l'environnement qui lui confèrent ses fonctions et d'une certaine manière… la définissent.

Merci à tous de m'avoir défini.

Liste des abréviations

- PDB — Protein Data Bank
- PQS — Protein Quaternary Structure
- X-ray — Cristallographie aux rayons X
- RMN — Résonance Magnétique Nucléaire
- ADN — Acide Désoxyribonucléique
- ARN — Acide Ribonucléique
- ARNm — ARN messager
- ARNt — ARN de transfert
- miARN — Micro ARN
- siARN — ARN silencer
- RN — Récepteur Nucléaire
- RXR — Récepteur X des rétinoïdes
- RAR — Récepteur à l'acide rétinoïque
- CoA — Coactivateur
- CoR — Corépresseur
- DBD — DNA Binding Domain
- LBD — Ligand Binding Domain
- HSP — Heat Shock Protein
- HoD — Homodimère
- HeD — Hétérodimère
- CX — Crystal packing
- P.Surface — Protein Surface
- P.Core — Protein Core
- IR — Interacting Residues
- BS — Binding Sites
- Petit peptide — Polypeptide de moins de 20 acides aminés
- Peptide — Polypeptide de 20 à 60/80 acides aminés
- Protéine — Polypeptide de plus de 60/80 acides aminés
- RVdW — Rayon de Van der Waals
- RCoul — Rayon de Coulomb
- RCov — Rayon de Covalence
- MS — Surface moléculaire de Connolly
- AS — Surface accessible de Lee & Richards
- ASh — Forme alpha d'Edelsbrunner
- TAP-tag — Tandem Affinity Purification-tag
- Y2H — Yeast Two Hybride (Double hybride)

Table des matières

REMERCIEMENTS **6**

LISTE DES ABREVIATIONS **11**

TABLE DES MATIERES **13**

LISTE DES FIGURES **17**

LISTE DES TABLEAUX **20**

AVANT-PROPOS **21**

RESUME DE LA THESE **25**

I. INTRODUCTION 25
II. CARTOGRAPHIE MOLECULAIRE 25
III. ANALYSE DE LA SURFACE MOLECULAIRE 28
IV. CRIBLAGE DE STRUCTURES SUR LA BASE DE REGIONS CONTIGUËS 29
V. TRAVAUX CONNEXES 30
VI. CONCLUSIONS ET PERSPECTIVES 30

INTRODUCTION GENERALE **33**

I. UNE BREVE HISTOIRE DE LA BIOLOGIE MODERNE 33
1. De l'observation de la cellule à la Théorie Cellulaire 34
2. Du Créationnisme à une nécessité d'Evolution et d'Extinction 35
3. Hérédité, Sélection, Evolution, Environnements 36
4. Des prémices de la Biologie Moderne à une nouvelle ère de la Médecine 40
5. De la caractérisation de l'ADN à la Biologie Structurale 41
6. Des Révolutions de la Biologie Moléculaire au Génie Génétique 42
II. BIOLOGIE MOLECULAIRE 44
1. La cellule et son contenu moléculaire 44
a. L'homme et la pluralité de ses paysages cellulaires 45
b. Nombre de molécules et composition chimique des cellules 46
2. Acides nucléiques 47
a. Composition chimique 47
b. La diversité des ARN et de leurs fonctions 48
c. Structures secondaires et méthodes prédictives 50
3. Protéines 51
a. Acides aminés 53
b. Liaison peptidique et squelette polypeptidique 57
c. Structure primaire 58
d. Structure secondaire 59
e. Structure tertiaire 61
f. Structure quaternaire 64
III. BIOLOGIE STRUCTURALE 67
1. Les méthodes expérimentales 67
2. Les projets de Génomique Structurale 69
3. Les données de la Biologie Structurale en 2010 71
a. Protein Data Bank 72
b. ModBase 74

c. SNP et MutDB 75
d. SM2PH-db 76
IV. BIOINFORMATIQUE 77
1. Bioinformatique des Séquences 77
a. De la comparaison par paire des séquences à l'inférence de fonction 77
b. De la comparaison multiple de séquences à l'annotation des protéines 78
2. Bioinformatique Structurale 80
a. Prédiction des structures protéiques 80
b. Prédiction des assemblages protéiques 81
c. Perspectives et apports pour la biologie et biomédecine 81
V. LA REGULATION DE LA TRANSCRIPTION EUCARYOTE 82
1. Famille des récepteurs nucléaires 83
2. Structures et oligomérisation des récepteurs nucléaires 86
3. Structures moléculaires et Génétique 87

STRUCTURES, SURFACES ET INTERACTIONS MOLECULAIRES 90

I. LA MODELISATION DES STRUCTURES MOLECULAIRES 90
1. L'atome, ses rayons, ses volumes 91
2. Surface moléculaire et surface accessible 93
3. La tesselation de Voronoï 97
4. Le complexe de Delaunay 99
5. Le complexe alpha et la forme alpha 100
6. Distinction entre la surface et le cœur des protéines 102
II. LES METHODES D'ANALYSES DES INTERACTIONS MOLECULAIRES 103
1. Détection des résidus interagissants par distance de contact 103
2. Détection des résidus interagissants par analyse des surfaces accessibles 104
3. Détection des résidus interagissants par les tesselations de Voronoï et dérivées 105
4. Amarrage moléculaire 107
III. LES CARACTERISTIQUES GEOMETRIQUES ET PHYSICO-CHIMIQUES DES INTERFACES 110
1. Les différents types d'interfaces moléculaires 110
a. Présentation générale 110
b. Problématiques et objectifs 113
2. Caractéristiques des interfaces moléculaires 114
a. Oligomères et homodimères 114
b. Hétérodimères 117
c. Contacts et assemblages cristallins 121
d. Protéine-petit peptide 125
e. Protéine-ADN 127
f. Protéine-ARN 130
g. Résumé des caractéristiques des différents types d'interactions moléculaires 133
3. Prédiction des sites de liaisons 136
a. Trois approches pour prédire les sites de liaison 137
b. Description et analyse des surfaces moléculaires par les régions 138
c. Conservation et coévolution : exemple de RXRα 140
d. Problèmes liés à la génération de régions par les méthodes usuelles 143
e. Problèmes liés à la prédiction des sites de liaison 144
IV. LE CRIBLAGE VIRTUEL : VERS UNE NOUVELLE ERE DU DEVELOPPEMENT DE MEDICAMENTS 145
1. Le développement de médicaments, une science interdisciplinaire en révolution 146

2. Les données structurales dans le développement de médicaments ... 148
3. La bioinformatique et la chemoinformatique dans le développement de médicaments ... 149
a. Modéliser les structures de cibles ... 149
b. Identifier les sites de liaisons et protéines druggables ... 150
4. Quels autres développements ? ... 151
a. Des bases de données sur les effets secondaires et la toxicité de composés ... 151
b. L'importance du contexte cellulaire ... 151
c. Vers une nouvelle utilisation des données structurales ... 152

MODELISATION DES STRUCTURES MOLECULAIRES ... 157

I. PROBLEMATIQUE : VERS UNE MODELISATION RAPIDE ET GENERIQUE DES SURFACES MOLECULAIRES ... 157
II. ARTICLE : DEFINING AND CHARACTERIZING PROTEIN SURFACES USING ALPHA SHAPES ... 159
III. CONCLUSIONS ET PERSPECTIVES ... 171
1. Empreintes structurales ... 171
2. Régions intermédiaires et intérieures ... 172
3. Régions sous contraintes de formes ... 173
4. La complémentarité des sites de liaison de l'interface ... 174

ANALYSE HAUT DEBIT DES INTERACTIONS MOLECULAIRES ... 177

I. PROBLEMATIQUE : IDENTIFICATION DES DIFFERENTS TYPES DE MOLECULES ET D'INTERACTIONS, ET INFERENCE DE FONCTIONS PAR HOMOLOGIE ... 177
II. ARTICLE : M-ORBIS : MAPPING OF MOLECULAR BINDING SITES AND SURFACES ... 179
III. CONCLUSIONS ET PERSPECTIVES ... 193
1. Etude de la surface non-interagissante ... 193
2. La Cartographie Moléculaire par le criblage de régions fonctionnelles ... 196

CRIBLAGE HAUT DEBIT DES REGIONS MOLECULAIRES ... 198

I. PROBLEMATIQUE ... 198
II. L'IMPORTANCE DE LA NOTION DE CONTIGUÏTE POUR LE CRIBLAGE DE REGIONS ... 200
III. LE CRIBLAGE DES REGIONS MOLECULAIRES ... 201
1. Définition de la liste exhaustive des régions contiguës d'une molécule ... 201
2. Comparaison et criblage de régions moléculaires ... 202
3. Filtres pour améliorer les performances du criblage et GRID computing ... 203
IV. ARBRE FONCTIONNEL D'UNE REGION ... 205
1. Récupération des structures de récepteurs nucléaires ... 206
2. Criblage du site de dimérisation de RXRα sur la famille des récepteurs nucléaires ... 207
3. Arbre fonctionnel du site de dimérisation ... 219
4. Arbre fonctionnel du site de liaison au cofacteur ... 221
5. La famille des GR, MR, PR, AR et NURR1 ... 223
V. CRIBLAGE DE REGIONS FONCTIONNELLES ET DETECTION D'INTERACTIONS ... 224
1. Modélisation de l'environnement par les structures moléculaires ... 224
2. Perspectives sur la modélisation des environnements cellulaires à partir des structures moléculaires 225
3. Spécifité d'une région moléculaire ... 227

CONCLUSIONS GENERALES ... 228

PUBLICATION CONNEXE : SM2PH-DB ... 233

ANNEXES 243

I. ANNEXE 1 : LISTE DES 832 STRUCTURES DE RECEPTEURS NUCLEAIRES REGROUPEES PAR FAMILLE 244
II. ANNEXE 2: QUELQUES RESULTATS DU CRIBLAGE DE L'ATP SUR TOUTES LES STRUCTURES MOLECULAIRES 249
III. ANNEXE 3: QUELQUES RESULTATS DU CRIBLAGE DU DORZOLAMIDE SUR TOUTES LES STRUCTURES MOLECULAIRES 249
IV. ANNEXE 4: QUELQUES RESULTATS DU CRIBLAGE DU COACTIVATEUR 2 DES RECEPTEURS NUCLEAIRES SUR TOUTES LES STRUCTURES MOLECULAIRES (KHKILHRLLQD) 250

RÉFÉRENCES 251

Liste des Figures

FIGURE 1. DE L'OBSERVATION FINE DE L'ORGANISATION DU VIVANT A LA THEORIE CELLULAIRE. 34
FIGURE 2. LA NECESSITE D'UNE EVOLUTION COMME MOTEUR DE LA DIVERSITE. 36
FIGURE 3. FOSSILE DE DARWINIUS MASILLAE, UN ANCETRE ELOIGNE DES PRIMATES DATANT DE L'EOCENE. 37
FIGURE 4. DE L'HEREDITE ET DE LA SELECTION NATURELLE A LA THEORIE DE L'EVOLUTION SYNTHETIQUE 38
FIGURE 5. EXEMPLE DE MICROENVIRONNEMENT A L'ECHELLE DE LA CELLULE. 39
FIGURE 6. EXEMPLE DE MICROENVIRONNEMENT A L'ECHELLE MACROMOLECULAIRE. 39
FIGURE 7. EXEMPLE DE MICROENVIRONNEMENT A L'ECHELLE MOLECULAIRE. 40
FIGURE 8. LES PERES DE LA MEDECINE MODERNE. 41
FIGURE 9. DE LA DECOUVERTE DE L'ADN, FACTEUR DE L'HEREDITE, A SA STRUCTURE CHIMIQUE. 42
FIGURE 10. L'ERE DE LA BIOLOGIE MOLECULAIRE : DE L'ARN MESSAGER AUX ENZYMES DE RESTRICTION ET LIGASES, JUSQU'AU SEQUENÇAGE DES PROTEINES ET ACIDES NUCLEIQUES. 43
FIGURE 11. TROIS EXEMPLES DE TYPES CELLULAIRES. 45
FIGURE 12. STRUCTURES DU RIBOSE ET DU DESOXYRIBOSE. 47
FIGURE 13. ADN, ARN ET LEURS BASES AZOTEES. 48
FIGURE 14. STRUCTURE SECONDAIRE ET TERTIAIRE D'UN ARN DE TRANSFERT. 51
FIGURE 15. LES DIFFERENTES ECHELLES DE STRUCTURES DES PROTEINES 53
FIGURE 16. LES DEUX ENANTIOMERES L (LEVOGYRE) ET D (DEXTROGYRE) DES ACIDES AMINES. 54
FIGURE 17. DIAGRAMME DE VENN DES 20 ACIDES AMINES REGULIERS.. 56
FIGURE 18. REPRESENTATION SIMPLIFIEE D'UNE CHAINE POLYPEPTIDIQUE EN FORMATION, D'UNE LIAISON PEPTIDIQUE ET DES ANGLES DE TORSION PHI (Φ) ET PSI (Ψ) 58
FIGURE 19. OBSERVATIONS DE MOTIFS REPETES AU NIVEAU DE STRUCTURES SECONDAIRES.. 59
FIGURE 20. REPRESENTATIONS DES STRUCTURES SECONDAIRES D'HELICE ALPHA ET DE FEUILLET BETA. 60
FIGURE 21. DIAGRAMME DE RAMACHANDRAN ET EXEMPLES DE CONFORMATIONS DE DEUX RESIDUS EN FONCTION DE LEURS ANGLES DE TORSION Φ ET Ψ. 60
FIGURE 22. EXEMPLE DU COMPLEXE RAP-RAF FORMANT UN FEUILLET BETA A L'INTERFACE. 61
FIGURE 23. CLASSIFICATION HIERARCHIQUE DES STRUCTURES PROTEIQUES DANS SCOP. 63
FIGURE 24. LE NUCLEOSOME, UNE STRUCTURE QUATERNAIRE CONSTITUEE DE 8 CHAINES POLYPEPTIDIQUES D'HISTONE. 65
FIGURE 25. COMPLEXE MULTI-PROTEIQUE TFIIH OBTENU PAR CRYO-MICROSCOPIE ELECTRONIQUE 68
FIGURE 26. STRUCTURE ISOLEE ET STRUCTURE « FONCTIONNELLE ». 70
FIGURE 27. CROISSANCE DES STRUCTURES ET REPARTITION PAR RESOLUTION. 72
FIGURE 28. REDONDANCE DES CHAINES STRUCTURALES PRESENTES DANS LA PDB EN FONCTION DE LEURS IDENTITES DE SEQUENCES. 73
FIGURE 29. EXEMPLE D'ALIGNEMENT PAR PAIRE DES SEQUENCES DE RXRA ET DE RARA. 77
FIGURE 30. EXEMPLE D'ALIGNEMENT MULTIPLE DE SEQUENCES DE RECEPTEURS NUCLEAIRES. 79
FIGURE 31. ORGANISATION DES 6 DOMAINES DES RECEPTEURS NUCLEAIRES. 83
FIGURE 32. PHYLOGENIE DES 48 RECEPTEURS NUCLEAIRES DE L'HOMME. 86
FIGURE 33. LES CLASSES I ET II DES RECEPTEURS NUCLEAIRES DEFINIS A PARTIR DES PONTS SALINS ENTRE LE SITE DE DIMERISATION, LE SITE DE LIAISON AU COFACTEUR ET LA BOUCLE H8-H9. 87
FIGURE 34. DECOUVERTE DE STRUCTURES IMPLIQUEES DANS LA REGULATION DE L'EXPRESSION DES GENES ET DES PRODUITS DES GENES. 88
FIGURE 35. LES DIFFERENTS RAYONS DECRIVANT LES ATOMES. 91
FIGURE 36. SURFACE MOLECULAIRE ET SURFACE ACCESSIBLE. 94
FIGURE 37. SURFACE MOLECULAIRE DE CONNOLLY EN FONCTION DE LA POLARITE DU SOLVANT. 94
FIGURE 38. SURFACE MOLECULAIRE A DIFFERENTES RESOLUTIONS. 95

FIGURE 39. CONSTRUCTION DE LA TESSELATION DE VORONOÏ D'UN ESPACE A DEUX DIMENSIONS. 98
FIGURE 40. TESSELATION DE VORONOÏ ET COMPLEXE DE DELAUNAY D'UN ESPACE A DEUX DIMENSIONS. 100
FIGURE 41. DU COMPLEXE DE DELAUNAY AU COMPLEXE ALPHA ET A LA FORME ALPHA. 101
FIGURE 42. CORRESPONDANCES ENTRE LA SURFACE MOLECULAIRE DE CONNOLLY ET LA FORME ALPHA. 101
FIGURE 43. DETECTION DES RESIDUS INTERAGISSANTS ET DE L'AIRE D'INTERFACE A PARTIR DU CALCUL D'AIRE ACCESSIBLE. 104
FIGURE 44. TESSELATION DE VORONOÏ D'UNE INTERFACE ET DETECTION DES ATOMES ET RESIDUS INTERAGISSANTS. 106
FIGURE 45. COMPLEXE DE DELAUNAY ET COMPLEXE ALPHA D'UNE INTERFACE ET DETECTION DES ATOMES ET RESIDUS INTERAGISSANTS 107
FIGURE 46. LES SIX DEGRES DE LIBERTE DANS LA RECHERCHE DE LA CONFORMATION D'UN ASSEMBLAGE MOLECULAIRE. 108
FIGURE 47. INTERFACES SECHES ET ROLE MEDIATEUR DE L'EAU DANS LES INTERFACES HUMIDES. 111
FIGURE 48. LES DIFFERENTS TYPES D'INTERACTIONS MOLECULAIRES. 112
FIGURE 49. REGIONS DE RECONNAISSANCE DANS DES COMPLEXES PROTEASE-INHIBITEUR. 119
FIGURE 50. DISTRIBUTION DES AIRES D'INTERFACES CRISTALLINES A PARTIR DE 1320 COUPLES D'INTERFACES. 122
FIGURE 51. DISTRIBUTION DES INTERFACES BIOLOGIQUES ET CRISTALLINES EN FONCTION DE L'AIRE NON POLAIRE ET DE LA FRACTION DES ATOMES ENFOUIS 125
FIGURE 52. QUELQUES EXEMPLES D'INTERACTIONS PROTEINE-PETIT PEPTIDE. 127
FIGURE 53. TROIS GRANDS MODES DE LIAISON DE L'ADN PAR LES PROTEINES. 129
FIGURE 54. DETECTION DES RESIDUS PROTEIQUES IMPLIQUES DANS LA RECONNAISSANCE SPECIFIQUE D'UN SEGMENT D'ADN. 129
FIGURE 55. TROIS TYPES D'ASSEMBLAGES PROTEINE-ARN. 131
FIGURE 56. REPRESENTATION SCHEMATIQUE DU CALCUL ET DE LA SIGNIFICATION DE LA FREQUENCE DE CONTACT $F_{CONTACT}$ D'UN RESIDU. 136
FIGURE 57. TENDANCE DES ACIDES AMINES A FAIRE PARTIE DES INTERFACES PROTEIQUES PLUTOT QUE DE LA SURFACE NON-INTERAGISSANTE. 137
FIGURE 58. CONSERVATION DES RESIDUS DE SURFACE DU RECEPTEUR RXRA 140
FIGURE 59. ARBRE DE DECISIONS PERMETTANT DE PREDIRE LES SITES DE LIAISON. 142
FIGURE 60. DE POSSIBLES COEVOLUTIONS ENTRE LES RESIDUS IMPLIQUES DANS LA LIAISON DU LIGAND ET LES RESIDUS IMPLIQUES DANS LA DIMERISATION DE RXRA. 143
FIGURE 61. PROBLEMES LIES AUX APPROCHES USUELLES DE GENERATION DE REGIONS. 144
FIGURE 62. LES DIFFERENTES ETAPES DE LA DECOUVERTE ET DU DEVELOPPEMENT DE MEDICAMENTS. 147
FIGURE 63. PROBABILITE DE DEUX MEDICAMENTS DE PARTAGER UNE CIBLE EN FONCTION DE LA SIMILARITE DE LEURS EFFETS SECONDAIRES ET DE LEUR SIMILARITE CHIMIQUE. 153
FIGURE 64. ETUDE ET CARACTERISATION DES SURFACES MOLECULAIRES PAR LES FORMES ALPHA 158
FIGURE 65. CALCUL D'UNE DISTANCE ENTRE DEUX POINTS, DEUX ATOMES OU DEUX RESIDUS EN FONCTION DE N PROPRIETES. . 171
FIGURE 66. CALCUL D'UNE DISTANCE GEODESIQUE SEPARANT LES POINTS A ET B PAR LE PLUS COURT CHEMIN D'ARETES 172
FIGURE 67. PRINCIPE DE LA GENERATION DE REGIONS SOUS CONTRAINTES DE FORMES. 174
FIGURE 68. COMPLEMENTARITE DE L'INTERFACE DE RXRA ET RARB VUE PAR SURFACE MOLECULAIRE ET FORME ALPHA. 175
FIGURE 69. QUATRE OBSERVATIONS RAPPROCHEES D'INTERFACES MOLECULAIRES REPRESENTEES PAR FORME ALPHA. 175
FIGURE 70. PRINCIPE DE LA CARTOGRAPHIE MOLECULAIRE DES SITES DE LIAISONS ILLUSTRE AVEC LA PROTEINE P53. 178
FIGURE 71. RECHERCHE CONFORMATIONNELLE DE L'ASSEMBLAGE SPECIFIQUE DE DEUX MOLECULES. 194
FIGURE 72. LA RENCONTRE DES MOLECULES OU LE « JEU DE LA CELLULE ». 195
FIGURE 73. DEUX MOLECULES AVEC DES SEQUENCES ET STRUCTURES SECONDAIRES ET TERTIAIRES DIFFERENTES PEUVENT LIER UN MEME PARTENAIRE. 199
FIGURE 74. IMPORTANCE DE LA SELECTION DE REGIONS CONTIGUËS POUR LE CRIBLAGE VIRTUEL. 200
FIGURE 75. FILTRE POUR LA COMPARAISON DE REGIONS, BASE SUR LEURS DEGRES DE PLISSEMENTS. 204
FIGURE 76. CORRELATION ENTRE LE SCORE D'ENERGIE ET L'IDENTITE DE SEQUENCE POUR LA FAMILLE DES RECEPTEURS NUCLEAIRES. 217

FIGURE 77. CORRELATION DES COMPOSANTES DU SCORE D'ENERGIE ET L'IDENTITE DE SEQUENCE POUR LA FAMILLE DES RECEPTEURS NUCLEAIRES..... 218
FIGURE 78. EN DEHORS DES FAMILLES PROTEIQUES, AUCUNE CORRELATION ENTRE L'IDENTITE DE SEQUENCE DES CIBLES ET LA RESSEMBLANCE DE LEURS SITES DE LIAISON N'EST OBSERVABLE. 219
FIGURE 79. ARBRE FONCTIONNEL DU SITE DE DIMERISATION DES RECEPTEURS NUCLEAIRES HUMAINS..... 220
FIGURE 80. ARBRE FONCTIONNEL DU SITE DE LIAISON AU COFACTEUR DES RECEPTEURS NUCLEAIRES HUMAINS..... 222
FIGURE 81. EXTREMITE C-TERMINALE PROLONGEE D'UNE SOUS-FAMILLE DE RN MODULANT LA FORME ET LA DISTRIBUTION DES CHARGES DU SITE USUEL DE DIMERISATION. 223

Liste des tableaux

Tableau 1. Composition chimique approximative d'une cellule d'E.coli en divisions rapides. 46
Tableau 2. Les 20 acides amines reguliers et la selenocysteine. 55
Tableau 3. Bases de donnees structurales differenciant les contacts cristallins des contacts biologiques. 66
Tableau 4. Bases de donnees structurales specialisees. 71
Tableau 5. Ligands les plus representes dans la Protein Data Bank. 74
Tableau 6. Contenu de ModBase en 2009. 75
Tableau 7. Les 48 genes des recepteurs nucleaires humains. 85
Tableau 8. Differentes evaluations des rayons de Van der Waals pour differents groupes atomiques. 92
Tableau 9. Differences entre les rayons de Coulomb et les rayons de Van der Waals pour les atomes polaires. ... 93
Tableau 10. Composition en acides amines de la surface et du cœur proteique. 96
Tableau 11. Differentes approches d'amarrage proteine-proteine existantes sur internet 109
Tableau 12. Une composition differente pour la surface et les interfaces d'homodimeres 115
Tableau 13. Une composition differente pour le cœur et le bord des interfaces homodimeriques. 116
Tableau 14. Compositions des interfaces d'homodimeres et d'heterodimeres 118
Tableau 15. Une composition differente pour le cœur et le bord des interfaces heterodimeriques. 120
Tableau 16. Ressemblances et differences entre les homodimeres et heterodimeres au niveau de leurs cœurs et bords d'interfaces. 121
Tableau 17. Comparaison de proprietes geometriques et physico-chimiques des interfaces specifiques et non-specifiques. 124
Tableau 18. Differences entre les interfaces proteine-ADN et proteine-proteine. 128
Tableau 19. Liste des residus de RXRα et de PPARγ impliques dans la reconnaissance des bases azotees de l'ADN. 130
Tableau 20. Comparaison des ARN simple et double brins avec l'ADN et les heterodimeres proteiques. 132
Tableau 21. Recapitulatif des caracteristiques de differents types d'interfaces moleculaires. 134
Tableau 22. Differences entre le cœur et le bord de l'interface pour les homo et heterodimeres. 135
Tableau 23. Accessibilites, aires enfouies et conservations des residus de RXRα a l'interface RXRα - RAR. 141
Tableau 24. Sept medicaments determines par l'analyse des structures de cibles therapeutiques : le structure-based drug design. 149
Tableau 25. Les differents types d'interactions et leurs representations sur la surface proteique 178
Tableau 26. Donnees structurales des recepteurs nucleaires 206
Tableau 27. Alignement par paire du site de dimerisation de RXR 216

Avant-propos

Les travaux présentés dans ce manuscrit de thèse ont pour objectifs d'apporter de nouvelles méthodologies bioinformatiques rapides et robustes pour l'analyse et la comparaison à haut débit des structures moléculaires. En effet, les projets internationaux de génomiques structurales débutés en 2000 ont entraîné la forte croissance des structures disponibles dans la Protein Data Bank. Cette disponibilité nouvelle des données structurales représente à la fois un défi en termes d'analyse et de complexité informatiques, mais aussi une opportunité sans précédent pour mieux comprendre les mécanismes impliqués dans la stabilité des macromolécules et dans la spécificité des reconnaissances moléculaires.

L'application de ces nouvelles méthodologies, basées à la fois sur l'utilisation des complexes et formes alphas (une représentation simplifiée du volume et des surfaces moléculaires), et sur la caractérisation systématique des différents types d'interactions représentées dans la PDB a permis d'une part de mieux caractériser la surface interagissante et non-interagissante des protéines, et d'autre part, d'identifier des ressemblances fonctionnelles pour des protéines ne partageant aucune identité de séquence ou de structures secondaires et tertiaires.

En effet, les molécules sont traditionnellement comparées sur la base de leurs séquences ou de leurs structures secondaires ou tertiaires. Plus récemment, des approches permettant de comparer ces structures également sur la base de motifs tridimensionnels ont vu le jour ; néanmoins ces motifs sont généralement de petites tailles, n'assurent pas la notion de contigüité pourtant fondamentale pour la description et la caractérisation de blocs fonctionnels indivisibles, et ne conservent pas les positions atomiques. L'approche que nous proposons permet de comparer les structures moléculaires sur la base de régions contigües de tailles et de formes variées, et préserve la position des atomes tout en tolérant une certaine variabilité observée dans les structures biologiques pour accroître la sensibilité de la recherche.

La comparaison rapide de ces structures moléculaires constitue une nouvelle forme de criblage virtuel qui a de nombreuses applications pour les sciences du vivant et pour la recherche de médicaments, notamment : l'identification de nouvelles cibles de composés et de candidats médicaments ; la détermination de la spécificité (ou fréquence) d'une région moléculaire dans un environnement cellulaire ; et plus généralement, l'évaluation de l'efficacité et des effets secondaires et toxiques de molécules dans un environnement.

Pour que le criblage des molécules sur la base de leurs régions soit pertinent, il est nécessaire de modéliser des environnements en fonction de leurs contenus moléculaires :

deux organismes ou bien deux cellules différenciées n'expriment pas les mêmes ARN et protéines ; par conséquent elles constituent des milieux différents dans lesquels une molécule donnée n'exerce pas nécessairement une même fonction. Plus précisément, l'un des objectifs de cette thèse est d'illustrer le fait que les « fonctions » d'une molécule sont définies par l'ensemble des interactions auxquelles elle participe dans un environnement donné. En perspectives de ce travail de thèse, il sera donc nécessaire d'incorporer des données de génomique fonctionnelle pour rendre compte de l'existence et de la fréquence d'une molécule dans différents organismes et pour différents types de cellules différenciées.

Ce manuscrit est composé d'un résumé regroupant les principaux résultats de ce travail de thèse, suivis de cinq parties :

La première partie consiste en une introduction générale qui, après une brève revue historique de découvertes majeures ayant contribué aux ères de la biologie moléculaire et de la biologie structurale, tend à illustrer et à décrire la cellule et son contenu moléculaire, avant de détailler les données de séquences et de structures disponibles, ainsi que les outils pour les analyser. La diversité chimique des différentes molécules est également abordée afin de fournir les bases théoriques nécessaires à l'analyse structurale de leurs interactions dans la seconde partie du manuscrit. Cette première partie se conclut par des rappels sur la régulation de l'expression des gènes eucaryotes avec un parallèle sur les structures et le rôle qu'elles ont joués dans la compréhension de ces mécanismes de régulation.

La seconde partie intitulée « Structures, surfaces et interactions moléculaires » constitue un état de l'art, à la fois sur les méthodologies employées pour modéliser et analyser les structures et interactions moléculaires, mais également sur les connaissances accumulées depuis 30 ans sur les différents types d'interactions moléculaires et sur les mécanismes de reconnaissance impliqués. La dernière partie de ce chapitre trace un parallèle entre l'étude des structures et interactions moléculaires et le développement de candidats médicaments au travers de quelques exemples concrets, avant de conclure sur de nouvelles problématiques et applications touchant à l'évaluation de l'efficacité et des effets secondaires et toxiques d'une molécule dans un environnement donné. Ces problématiques et applications seront reprises dans la cinquième et dernière partie de cette thèse, avec le criblage virtuel des régions moléculaires.

La troisième partie intitulée « Modélisation des structures moléculaires » décrit l'utilisation des formes alphas pour modéliser et décrire efficacement les surfaces moléculaires et leurs régions contiguës. Construits autour d'un article publié en 2008, nous présentons quelques développements supplémentaires de ces travaux, en particulier la génération d'empreintes structurales, des régions caractérisant des propriétés remarquables des surfaces moléculaires et pouvant servir à des classifications structurales. De plus,

l'utilisation des complexes alphas plutôt que des formes alphas permet également de modéliser le volume d'une structure moléculaire, et par conséquent de calculer de façon rapide et robuste la profondeur d'un atome dans une structure. Le complexe alpha permet également de définir des régions internes, c'est-à-dire des régions ne contenant que des atomes et résidus inaccessibles au solvant. Ces régions internes peuvent être utilisées pour identifier les régions de faibles et de fortes densités, ainsi que pour étudier de possibles coévolutions entre régions distantes. Les coévolutions entre régions de protéines, plutôt que les coévolutions entre résidus de protéines pourraient permettre d'identifier plus facilement des chemins de régulation allostérique, notamment entre les sites de liaison et le cœur protéique.

La quatrième partie intitulée « Analyse haut débit des interactions moléculaires » décrit l'approche M-ORBIS que nous avons nommée « Cartographie Moléculaire » en raison de la caractérisation fonctionnelle et dynamique qu'elle procure sur les protéines. Articulés autour d'un article publié en 2010, nous présentons dans un premier temps les aspects méthodologiques qui permettent d'établir la cartographie moléculaire d'une protéine à partir de l'analyse intégrative des structures homologues de cette protéine. Dans un second temps, nous décrivons les résultats de M-ORBIS, avec en particulier la détermination que 44% de la surface protéique est capable d'interagir avec des partenaires autres que du solvant et des ions. De plus, nous montrons l'existence de sites de liaisons spécifiques (qui ne peuvent lier qu'un type de molécule), et de sites de liaisons polyvalents (qui peuvent lier plusieurs types de molécules), qui diffèrent grandement en termes de composition. En perspectives de ce chapitre, nous formulons l'hypothèse qu'une partie de la surface moléculaire a effectivement évolué sous une pression de sélection visant à défavoriser la formation d'assemblages moléculaires stables. Une telle surface non-interagissante pourrait être corrélée à la cinétique de rencontre de deux molécules dans un environnement donné, permettant aux assemblages non-spécifiques de se dissocier plus rapidement après une collision en limitant le jeu des possibles conformations à tester. Le chapitre se termine sur un développement futur de la cartographie moléculaire : pour inférer des fonctions et explorer la dynamique conformationnelle d'une protéine, M-ORBIS doit récupérer des structures homologues de cette protéine. Cette étape limite son champ d'application aux seules protéines pour lesquelles on dispose de plusieurs structures. Nous proposons une évolution de M-ORBIS permettant d'inférer des sites de liaison et partenaires associés à partir du criblage des régions contiguës composant la protéine, s'affranchissant ainsi de la nécessité d'avoir des structures homologues.

La cinquième partie intitulée « Criblage haut débit des régions moléculaires » décrit dans un premier temps les aspects méthodologiques de ce criblage, et dans un second temps, illustre la puissance de cette nouvelle approche de comparaison par plusieurs exemples de

prédiction d'interactions portant sur des assemblages protéine-protéine, protéine-petit peptide et protéine-ligand. La possibilité de récupérer l'ensemble des molécules portant une région fonctionnelle ouvre la voie à une meilleure compréhension des mécanismes de reconnaissance moléculaire, en permettant la création « d'arbres fonctionnels ». Ceux-ci regroupent les molécules en fonction de leurs capacités à partager une fonction (comme lier un partenaire, ou phosphoryler un résidu), indépendamment d'une évolution commune. Deux arbres fonctionnels des récepteurs nucléaires humains ont par exemple été définis : l'un à partir de leurs sites de dimérisation, l'autre à partir de leurs sites de liaison au cofacteur. Les arbres fonctionnels comportent à la fois des similitudes et des différences avec les arbres évolutifs. La pertinence des hypothèses que prédisent les arbres fonctionnels restent à valider expérimentalement. Par ailleurs, le criblage d'un grand nombre de régions moléculaires a également permis de mettre en évidence l'existence de régions très largement sous-représentées dans l'environnement cellulaire, et d'autres, très largement sur-représentées. Des corrélations entre la fréquence et l'aspect fonctionnel d'une région sont en cours d'étude. Le chapitre se termine sur des exemples concernant les nouvelles applications de ces criblages de régions moléculaires, notamment l'évaluation de phénomènes d'efficacité et d'effets secondaires/toxiques de composés, et précise la création de l'entreprise de biotechnologies BIONEXT, ainsi que le dépôt de deux brevets en 2009.

Résumé de la Thèse

I. Introduction

La biologie expérimentale est à présent capable de séquencer massivement des polymères biologiques tels que les protéines, ADN et ARN. Ce séquençage n'est que le premier pas nécessaire d'un domaine émergent consistant à identifier les acteurs cellulaires et à caractériser leurs fonctions à partir de données biologiques variées et au sein d'un contexte cellulaire : la biologie des systèmes.

Dans ce contexte, les structures biologiques apportent des données quantitatives et qualitatives permettant de mieux comprendre les mécanismes d'interactions, leurs diversités et leurs dynamiques. Les projets de génomiques structurales ont permis une forte croissance de ces données, passant de 10 000 structures recensées dans la Protein Data Bank (PDB) en 2000, à plus de 66 000 structures en 2010. Les structures de protéines membranaires, d'ADN, d'ARN et de macro-complexes sont également déterminées plus facilement et contribuent à accroître les connaissances, en particulier sur les voies de régulation de l'expression des gènes. L'étude des structures est de plus à l'origine de nombreux succès thérapeutiques et constitue une discipline à part entière dans la conception de médicaments : le « *structure-based drug design* ».

La Bioinformatique Structurale qui a pour rôle d'analyser ces structures, s'est en particulier intéressée aux différents types d'interactions moléculaires (protéine-protéine, protéine-ADN, protéine-ARN, protéine-peptide), en les caractérisant par des paramètres géométriques, physico-chimiques et évolutifs. Il a ainsi été possible de montrer que les interfaces d'homodimères sont plus hydrophobes, moins chargées et plus planes que les autres interfaces moléculaires. D'autres travaux ont mis en évidence la présence répétée de contacts cristallins dans les structures cristallographiques, et ont conduit à développer des approches prédictives pour discriminer les assemblages biologiques des assemblages cristallins. En 2006, date de début de cette thèse, l'étude des interactions moléculaires se concentrait sur la prédiction des assemblages moléculaires (*docking*), la prédiction des sites de liaisons sur les protéines, la présence de résidus énergiquement importants aux interfaces (*points chauds* ou *hot spots*), et le rôle de l'eau dans la médiation des interactions.

II. Cartographie Moléculaire

Ce travail de thèse a consisté dans un premier temps à assimiler l'état des connaissances sur les différents types d'interactions moléculaires, puis de concevoir des algorithmes, bases

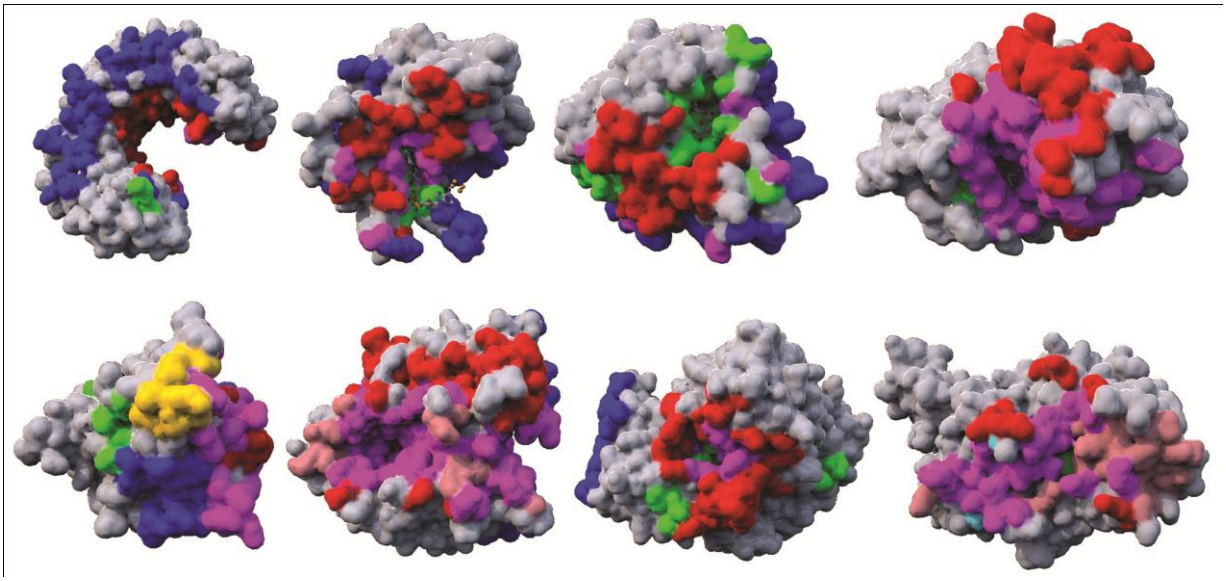
de données et interfaces de visualisation permettant l'analyse des surfaces et interfaces moléculaires, pour faire émerger des caractéristiques biologiques importantes et proposer des approches prédictives.

Alors que notre approche pour différencier les assemblages biologiques et cristallins a une précision de l'ordre de 90%, la prédiction des sites de liaisons était systématiquement limitée à 75%. L'analyse de nos résultats a montré que certaines régions des protéines étudiées étaient constamment prédites comme étant des sites de liaisons, bien que les structures étudiées alors n'en faisaient pas état. La consultation des structures proches montra que ces régions qui n'avaient pas de fonction connue dans une structure, étaient souvent des sites de liaisons dans des structures homologues. Le pourcentage approximé de la surface interagissante des protéines, alors évalué à 20% dans la littérature, devait donc être revu. Prédire qu'une large fraction de la surface est constituée de sites de liaisons est cependant moins utile pour guider les programmes d'amarrages moléculaires (*docking*) ainsi que les expériences de mutagénèses dirigées. Aussi, avons-nous réorienté notre attention sur les données structurales existantes et développé le concept de Cartographie Moléculaire [1].

La Cartographie Moléculaire est une approche haut-débit de comparaison intégrative des données structurales, qui permet de décrire les fonctions et comportements dynamiques d'une molécule. En automatisant pour la première fois la reconnaissance des types moléculaires et la caractérisation de 9 différents types d'interactions entre molécules (homodimère, hétérodimère, protéine-peptide, protéine-petit peptide, protéine-ADN, protéine-ARN, protéine-ligand, protéine-ion, protéine-solvant), notre approche a permis la description fine des 9 différents types de sites de liaisons, ainsi que la description fine des fréquences de contact des résidus interagissants et des fréquences de solvatation des résidus. La Cartographie Moléculaire fournit également pour la première fois une classification des structures en fonction de leur contexte moléculaire (molécule impliquée dans des interactions homodimériques, hétérodimériques, protéine-ADN, protéine-ARN, …). Elle permet d'étudier et de quantifier simplement des aspects dynamiques d'une molécule, tels que sa variabilité intrinsèque (ex : étudier la flexibilité du squelette pour une protéine en forme libre), ainsi que d'étudier et quantifier les changements s'opérant lors du passage d'un contexte à un autre (ex : étudier les changements de conformations lors du passage d'une forme libre à une forme liant un partenaire).

Validé sur 6 jeux non redondants d'assemblages structuraux représentatifs de différents types d'interactions, le concept de Cartographie Moléculaire que nous avons développé a permis de démontrer qu'en moyenne 44% de la surface protéique est impliquée dans des interactions avec des molécules autres que solvant et ions. Alors que près de 86% de la

surface des protéines est transitoirement hydratée, seules en moyenne 15% de ces surfaces lient des molécules d'eau de façon spécifique (non transitoire). Plus important encore, en différenciant l'ensemble des 9 types de sites de liaisons, nous avons pu mettre en évidence l'existence de recouvrements entre ces régions. Cette observation nous a conduit à la description de deux grandes familles de sites de liaisons jusque-là méconnues: des sites de liaisons spécifiques, dans le sens qu'ils ne peuvent lier qu'un seul type de molécule (protéine, ADN, ARN…), et des sites de liaisons polyvalents, dans le sens qu'ils peuvent lier au moins deux types différents de molécules. Nous avons montré que les sites de liaisons spécifiques diffèrent grandement des sites de liaisons polyvalents, notamment dans le cas des homodimères où la composition en hydrophobes initialement décrite à 39% est en moyenne de 66% pour les sites de liaisons spécifiques à la formation d'homodimères. De telles régions hydrophobes, proches de la composition du cœur protéique (corrélation de Pearson de 0.93), pourraient suggérer des interactions fortes, voire permanentes.



Cartographie Moléculaire des sites de liaisons

Les sites de liaisons pour hétérodimères sont représentés en rouge, ceux pour les homodimères et ligands sont respectivement en bleu et vert. Les sites de liaisons à l'ADN et aux peptides sont quant à eux en jaune et saumon et les sites de liaisons polyvalents sont indiqués en mauve. De haut en bas et de gauche à droite : inhibiteur de ribonucléase (1dfj) ; ferrodoxin-NADP réductase (1ewy) ; neuraminidase (3b7e) ; kinase cAMP dépendante (1ydr) ; suppresseur de tumeur p53 (1tsr) ; neurotoxine bont/A (1xtg) ; acétylcholinestérase (1fss) ; protéine G (1bof).

Finalement, la Cartographie Moléculaire a permis de montrer que les résidus composant un site de liaison pouvaient être divisés en deux familles : des résidus qui sont toujours vus interagissant dans des structures homologues présentant un même type d'interaction (un

même contexte moléculaire), et qui pourraient servir à remplir les contraintes propres au type d'interaction considérée ; et des résidus qui sont vus n'interagissant que dans une fraction des structures homologues présentant un même type d'interaction. Ces derniers pourraient notamment servir à moduler la spécificité de la reconnaissance des partenaires liant un même site de liaison. Les résultats de cette approche font l'objet d'un article sous presse dans le journal Nucleic Acid Research [1].

III. Analyse de la surface moléculaire

Une autre composante importante de ce travail a consisté à poser des bases méthodologiques rigoureuses pour l'analyse massive des données structurales, notamment des surfaces moléculaires. La surface moléculaire a longtemps été étudiée à partir des représentations de Connolly [2] et de Lee et Richards *[3]*. Ces surfaces, qui dérivent d'une approche de type « *marching cube* », permettent de modéliser les surfaces continues avec un haut niveau de détail. En 1995, le mathématicien américain Edelsbrunner introduisit en biologie une méthode de modélisation des surfaces qui simplifie leurs représentations (10 à 20 fois moins de facettes), tout en conservant la position des atomes: la forme duale, encore appelée forme alpha. Nous avons démontré la pertinence de ces formes alphas pour analyser les surfaces moléculaires, différencier les résidus de surface des résidus du cœur protéique et segmenter les surfaces en régions contiguës. Ainsi, la surface protéique est en moyenne 40% moins hydrophobe, 65% plus flexible et 75% moins conservée que le cœur protéique. La définition des régions nous a permis de développer notamment une nouvelle mesure de la courbure, non biaisée par la présence de cavités sous la surface. Cette définition de courbure a permis d'observer que le paysage des surfaces moléculaires est composé en moyenne de 38% de bosses et de 62% de crevasses. Bien qu'étant un problème de longue date, la détection des crevasses et cavités est toujours d'actualité en raison de la complexité du problème géométrique et de son importance en biologie et en pharmaceutique : les crevasses sont connues pour favoriser la liaison de ligands. La PDB référence ainsi plus de 10 000 interactions protéine-ligand, dont la plupart ont lieu dans ces régions.

Bien que moins fréquentes, les bosses sont composées de résidus qui participent statistiquement 30% plus aux interactions protéine-protéine que les résidus des crevasses. Indépendamment de leur localisation, les résidus interagissants sont en moyenne 15% plus exposés au solvant que le restant des résidus de surface. Les bosses sont plus chargées, moins hydrophobes et moins aromatiques que les crevasses, qui ressemblent davantage à la composition du cœur protéique. Ces observations ont conduit à montrer que la présence de

résidus aromatiques très exposés dans les bosses (où ils sont généralement peu fréquents), indique presque systématiquement la présence d'un site de liaison. Cette méthodologie consistant à générer des surfaces moléculaires simplifiées par le biais des formes alphas, de construire des régions et de calculer des courbures non biaisées, est rapide, robuste et applicable à tout type moléculaire, y compris ADN et ARN. Ce travail a fait l'objet d'une publication et de la couverture du journal Proteins en 2009 [4].

IV. Criblage de structures sur la base de régions contiguës

La Cartographie Moléculaire présentée précédemment repose sur la comparaison globale des structures moléculaires et de leurs séquences afin de fournir une description dynamique et fonctionnelle de la molécule étudiée. Par conséquent, une fonction d'une molécule ne peut être inférée qu'à partir d'une ressemblance suffisante avec une autre molécule, sur la quasi-totalité de sa séquence ou de sa structure. L'inférence est donc limitée par le nombre de structures proches. Faisant suite au travail réalisé sur l'analyse des surfaces, nous avons développé une méthode originale permettant de comparer les structures moléculaires sur la base de leurs régions contiguës, s'affranchissant ainsi de toute nécessité de ressemblance de séquences ou de structures secondaires et tertiaires. Ainsi, l'inférence d'une fonction n'est plus limitée par le nombre de structures proches, mais par le nombre (beaucoup plus important) de structures présentant des régions similaires.

A partir des 66 000 structures biologiques contenues dans la PDB, il a été possible de générer plus de 90 millions de régions de tailles et de formes variées ; plus d'un million de ces régions comportent une fonction connue. Comparable à un BLAST, notre approche fonctionne sur des structures et permet de récupérer l'ensemble des structures partageant une région similaire. Le procédé consiste à diviser une structure en une pluralité de régions contiguës et à cribler chacune de ces régions sur la totalité des régions connues afin, (1) d'identifier des similarités entre des molécules parfois très différentes, (2) repérer les différences fines entre sites de liaisons proches et, (3) inférer des fonctions par similarité de régions. En effet, nous avons pu montrer qu'une similarité entre régions implique très souvent une similarité de fonction (aptitude à lier un ligand ou tout autre type moléculaire). Validée à partir de la comparaison de sites de liaisons de ligands fréquemment rencontrés dans la PDB (liaison à l'ATP, FAD, MAN, etc), l'approche a une précision évaluée à 80-90% en fonction de la taille du site de liaison et des annotations disponibles ; elle présente de plus l'avantage de fournir tous les éléments permettant de retracer l'obtention d'un résultat à des fins de vérification. Le criblage de régions a permis de mettre en évidence des interactions qui n'étaient pas connues, en détaillant notamment les sites de liaisons et partenaires impliqués, et en proposant une structure de l'assemblage prédit ; des validations

expérimentales sont en cours. De plus, le criblage de multiples régions a également permis de constater que pour un contexte cellulaire donné, certaines régions sont très largement sous-représentées, alors que d'autres régions sont très largement sur-représentées.

Une base de données ne contenant que les régions moléculaires définies à partir des structures de récepteurs nucléaires a ensuite été constituée. Le criblage croisé des sites de dimérisation, ainsi que le criblage croisé des sites de liaisons aux cofacteurs a permis de construire des « arbres fonctionnels » de ces deux régions. Ces arbres réunissent les récepteurs nucléaires de façon comparable à un arbre phylogénétique, mais en fonction de la ressemblance respective de leurs sites de dimérisation, ou de leurs sites de liaisons au cofacteur. L'intérêt de ces arbres tient dans leur capacité à rassembler les molécules partageant des fonctions similaires, et non de les rassembler en fonction d'une évolution commune. Ces arbres fonctionnels ont donc une valeur prédictive que nous continuons d'évaluer, notamment pour les récepteurs orphelins. Un article présentant notre approche de criblage virtuel des régions moléculaires et son application aux récepteurs nucléaires et à quelques composés sera bientôt rédigé.

V. Travaux connexes

Dans le projet *Alvinella (étude d'un métazoaire thermotolérant)*, les approches développées au cours de cette thèse ont montré une accumulation de résidus chargés et de liaisons hydrogènes dans le cœur protéique qui pourraient contribuer à la thermotolérance de l'organisme. D'autres analyses sont en cours pour étudier les différences au niveau des sites de liaisons et régions remarquables.

Dans le projet *SM2PH (relation phénotype et mutation structurale)*, les approches développées ont permis de distinguer les mutations faisant partie du cœur et de la surface protéiques, et d'étudier les phénomènes d'accumulations de mutations dans des sites de liaisons et régions remarquables [5].

Dans le projet *CAMI (interactome du muscle)*, les approches développées ont permis de valider la présence et la localisation de sites de liaisons et de partenaires détectés par double hybride (Y2H).

VI. Conclusions et perspectives

Les protéines sont des molécules dynamiques et pluripotentes, capables de lier une grande variété de partenaires par le biais d'une grande diversité de mécanismes. Notre

analyse par Cartographie Moléculaire a permis de mieux caractériser ces molécules et leurs différents sites de liaisons. L'existence de sites de liaisons spécifiques à un type de molécule, et de sites de liaisons polyvalents capables de lier plusieurs molécules de types différents ouvre la voie à une nouvelle compréhension des mécanismes d'interactions. En particulier, il sera intéressant d'étudier la composition de ces sites de liaisons avec la vision dynamique qu'offre la Cartographie Moléculaire, pour observer notamment s'il existe des différences géométriques, physico-chimiques ou évolutives entre les résidus d'un site de liaison toujours impliqués dans un type d'interaction (pour répondre aux contraintes d'une liaison d'un type moléculaire ?), et les résidus qui ne sont impliqués que dans une fraction des structures d'un même contexte (pour moduler la spécificité de la reconnaissance des partenaires ?).

La comparaison des données structurales à l'échelle des régions marque également une nouvelle ère dans l'identification rapide et précise des fonctions moléculaires. Ces analyses permettent de mieux comprendre pourquoi des molécules proches peuvent remplir des fonctions complètement différentes alors que des molécules différentes peuvent effectuer des fonctions similaires. D'autres analyses sont en cours, notamment pour classifier et caractériser les régions sur- et sous- représentées dans un contexte cellulaire. Les régions sur-représentées pourraient notamment correspondre à une partie de la surface non-interagissante et les régions sous-représentées pourraient correspondre à des sites de liaisons essentielles à la vie cellulaire pour lesquels il est nécessaire d'éviter des interférences avec d'autres systèmes, et donc de conserver des particularités uniques.

Parallèlement à ma thèse, j'ai créé l'entreprise BIONEXT avec l'aide de plusieurs collaborateurs dont l'incubateur SEMIA. Intégrant des travaux antérieurs à la thèse, celle-ci a permis d'accélérer et de mener à bien le développement des approches de criblage et de développer plusieurs applications industrielles et thérapeutiques touchant aux problèmes d'effets secondaires, d'efficacité et de toxicité des composés. En effet, plus de 20% des protéines humaines disposent d'au moins une structure dans la banque structurale PDB, et près de 40% des protéines humaines pourraient être modélisées par homologie. BIONEXT est à présent lauréat du Concours National de Création et Développement d'Entreprises Innovantes 2010 et participe également à une ANR blanche et à une ANR BiotecS 2010, afin de favoriser les collaborations futures avec l'IGBMC, améliorer les approches de criblages et continuer d'explorer le champ des nouvelles applications.

Introduction Générale

« Rien ne naît ni ne périt, mais des choses déjà existantes se combinent, puis se séparent de nouveau »
Anaxagore, 500-428 avant J.C.

« Rien ne se perd, rien ne se crée, tout se transforme »
Lavoisier, 1743 – 1794 après J.C.

I. Une brève histoire de la Biologie Moderne

Cette première partie décrit l'historique de quelques grandes découvertes scientifiques qui ont permis l'essor de trois disciplines fondamentales pour les travaux présentés dans cette thèse : la biologie moléculaire avec l'étude des séquences et des fonctions des protéines et acides nucléiques ; la biologie structurale, avec l'élucidation de structures permettant de mieux comprendre les mécanismes impliqués dans la stabilité et la reconnaissance des molécules ; la recherche de médicaments, jusqu'à l'identification des causes moléculaires des pathologies. L'ADN et quelques concepts d'Evolution seront également présentés afin de donner un aperçu de l'origine de la variabilité génétique observée chez les êtres vivants.

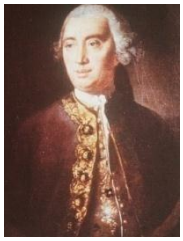
L'objectif de cette partie est de rappeler sinon de faire prendre conscience que le vivant et l'environnement sont indissociables, et que l'on ne peut définir une fonction à un gène, un ARN, une protéine ou à tout autre molécule, que pour un environnement donné. Plus précisément, nous considérerons que la fonction d'une molécule dans l'environnement est définie par la somme des interactions à laquelle elle participe.

La Biologie du grec ancien « bios » et « logos » est « l'étude de la vie » et de ce qui la compose. Dans ses origines grecques, Platon et Aristote pensaient que le monde et le vivant étaient constitués de quatre éléments : la terre, l'eau, le feu et l'air. Pour Aristote, les aliments étaient eux-mêmes formés de ces éléments et ceux-ci se dégradaient (prélude au concept de catabolisme) pour se recombiner différemment (prélude au concept de d'anabolisme et/ou de métabolisme) dans le corps humain. La transition de l'alchimie à la chimie moderne suite aux travaux de Lavoisier, permettra aux chercheurs du XX^{ème} siècle à nos jours de décrire par des formules et des équations chimiques une partie du vivant, de sa variabilité et de sa dynamique.

1. De l'observation de la cellule à la Théorie Cellulaire

La Biologie Moderne telle que nous la concevons aujourd'hui prend peut-être ses racines dans les travaux de Robert Hooke au XVII^{ème} siècle, qui inventa le premier microscope et rapporta dans son livre de 1665 des descriptions fines de plusieurs organismes (insectes, foraminifères, plumes d'oiseaux, …) [6]. Il fut également le premier à observer les cellules de plantes et à les nommer ainsi en raison de la compartimentation de l'espace qu'elles opèrent. Ses observations lui permirent de conclure que plusieurs organismes différents partageaient ces structures similaires de cellules. C'est à cette même période qu'un naturaliste hollandais du nom d'Antonie van Leeuwenhoek fit la découverte stupéfiante de petits organismes se mouvant dans l'eau et la salive [7] ; ceux-ci seraient plus tard appelés « bactéries » et « protistes ».

Ces recherches donnèrent lieu en 1838 à l'énoncé par J.M. Schleiden [8] puis en 1839 par T. Schwann [9] de la Théorie Cellulaire ; celle-ci propose que la cellule est l'unité fonctionnelle fondamentale du vivant et que celle-ci est elle-même composée de sous-compartiments appelés organites. La distinction entre cellules procaryotes (du grec pro « avant », et caryon « noyau ») et eucaryotes (eu « vrai », caryon « noyau ») ne se fera toutefois que bien plus tard, dans le courant des années 1960 avec Edouard Chatton, Roger Stanier et C.B. van Niel [10]. La subséquente division des procaryotes en bactéries et archaebactéries ne sera faite qu'en 1977 par Carl Woese à partir de la comparaison des séquences de l'ARN ribosomal 16S [11, 12].



Robert Hooke (1635-1703)

Antonie van Leeuwenhoek (1632-1723)

Matthias Jakob Schleiden (1804-1881)

Théodore Schwann (1810-1882)

Edouard Chatton (1883-1947)

Carl Woese (1928 - .)

Figure 1. De l'observation fine de l'organisation du vivant à la Théorie Cellulaire.

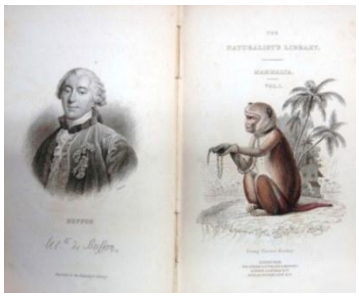
2. Du Créationnisme à une nécessité d'Evolution et d'Extinction

«… le temps et les conditions favorables sont les deux principaux moyens que la nature emploie pour donner naissance à chacune de ses créations »

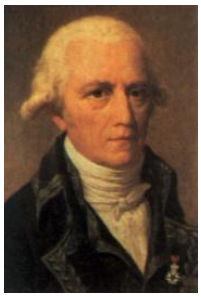
Lamarck, 1801

Parallèlement à ces observations du monde microscopique, des chercheurs du XVII^{ème} siècle furent également saisis par la diversité apparente du vivant. Désavouant les théories existantes sur l'apparition de la Vie, de l'Homme et même de la Terre (alors estimée à 6000 ans par l'Eglise Catholique), ils cherchèrent des causes naturelles plutôt que divines à cette diversité. Au nombre de ceux-ci, Georges-Louis Leclerc, comte de Buffon, publia des travaux tels que l'encyclopédie de l'Histoire Naturelle (1749-1788), précisant les similarités entre l'homme et le singe et allant jusqu'à théoriser un ancêtre commun entre ces deux espèces. Bien que Buffon eût l'intime conviction de cette évolution d'espèces en espèces, il ne put jamais expliquer quels en étaient les mécanismes et les moteurs.

L'un de ses contemporains, Jean-Baptiste Lamarck exprima en 1801 l'idée que, avec le temps, des conditions favorables pourraient conduire des espèces à évoluer et à s'adapter. La pensée qu'un organisme pourrait réagir et évoluer aux conditions de son environnement était nouvelle et en opposition aux théories de Geoffroy Saint-Hilaire. Il exprima ainsi sa théorie de l'Evolution, le Lamarckisme, dans deux principes ou lois : la première énonce que l'utilisation ou la non-utilisation d'une structure ou d'un organe l'amène à grandir ou rétrécir au cours du temps ; la seconde précise que de tels changements sont héréditaires. La conséquence de ces deux lois est le changement graduel et continu des organismes face à leurs environnements. Toutefois, Lamarck voyait dans l'Evolution du vivant, une perfection trop parfaite et trop rapide pour envisager qu'une espèce puisse s'éteindre ; il imaginait plutôt qu'une espèce qui disparaissait était en fait graduellement transformée dans une autre espèce. En cela, les pensées de Lamarck différaient profondément de celles de Georges Cuvier, un estimé collègue du Muséum National d'Histoire Naturelle de Paris qui établit, par l'étude des fossiles, que l'extinction des espèces est une réalité. Cuvier reprit les travaux de ses prédécesseurs indiquant que la Terre était beaucoup plus ancienne que ce qui était communément accepté, et introduisit le concept de « catastrophe naturelle », un événement ponctuel capable de provoquer l'extinction de plusieurs espèces. Ces théories ont plus tard été reprises pour expliquer les crises d'extinctions massives d'espèces.



Georges-Louis Leclerc, Comte de Buffon
(1707-1788)



Jean-Baptiste Pierre Antoine de Monet, Chevalier de la Marck
(1744-1829)

Jean Léopold Nicolas Frédéric Cuvier
(1769-1832)

Figure 2. La nécessité d'une évolution comme moteur de la diversité.

3. Hérédité, Sélection, Evolution, Environnements

« In the struggle for survival, the fittest win out at the expense of their rivals because they succeed in adapting themselves best to their environment»
Charles Darwin, XIX^{ème} siècle

«A species is a reproductive community of populations (reproductively isolated from others) that occupies a specific niche in nature »
Ernst Mayr, 1982

Le XIXème siècle est le siècle d'un renouvellement profond de la Biologie sur des bases modernes. Les observations de Buffon, Lamarck et Cuvier eurent probablement un rôle important dans les découvertes futures de Gregor Mendel sur l'Hérédité en 1866 [13] et de Charles Darwin sur l'Evolution en 1859 [14]. Bien que ces deux hommes de Science ne se soient pas connus, ils révélèrent tous deux des lois complémentaires qui, plus tard, formeraient les fondements de la Théorie Synthétique de l'Evolution.

Gregor Mendel s'intéressa aux lois de l'hérédité, démontrant la présence de « facteurs » transmis de génération en génération et permettant l'expression d'un phénotype. Il eut également la remarquable intuition de proposer que ces facteurs pouvaient être soit dominants (s'exprimant toujours dans le phénotype), soit récessifs (ne s'exprimant que s'ils sont présents en deux exemplaires). Cette théorie lui permit d'expliquer comment les caractères de parents pouvaient ne pas se retrouver chez leurs enfants mais parfois chez leurs petits enfants. Ainsi, Gregor Mendel mit en évidence ces « facteurs » de l'hérédité et la façon dont ils étaient transmis de génération en génération. Il restait encore à caractériser chimiquement et fonctionnellement ces facteurs de l'hérédité ainsi que de proposer des

mécanismes d'évolution et de spéciation qui expliqueraient l'apparition, et dans une certaine mesure, la disparition des espèces.

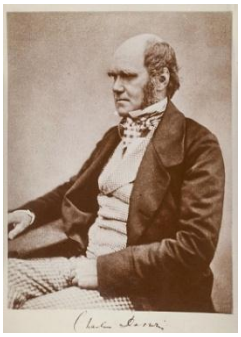
La compilation de près de 30 ans d'observations permit à Charles Darwin de présenter sa Théorie sur l'Origine des Espèces. Il y décrit l'influence de l'environnement sur l'évolution des espèces : parmi la diversité existante au sein d'une espèce, les organismes les plus adaptés arrivent quantitativement à se nourrir et à se reproduire plus facilement. Dès lors, par « Sélection Naturelle », ceux-ci colonisent davantage leur territoire et l'accumulation de ces facteurs, qui les rendent plus adaptés au milieu, les conduit parfois à créer une espèce différente. Charles Darwin, par sa théorie de la sélection, introduit l'idée que tous les organismes vivants partagent un ancêtre commun, hypothèse qui fut depuis plusieurs fois vérifiée, notamment par la découverte de l'archæoptéryx, un fossile d'hybride reptile et oiseau, ou plus récemment par la découverte du fossile de *Darwinius masillae,* un genre inconnu de primate datant de l'époque de l'Eocène, il y a environ 47 millions d'années [15].

Figure 3. Fossile de Darwinius masillae, un ancêtre éloigné des primates datant de l'Eocène. *Proche des lémuriens, Darwinius fait partie de la sous-famille des Cercamoniinae de la famille des Notharctidae. Il appartient aux Adapiformes, un groupe éteint des premiers primates.*

Ainsi, l'Evolution est la résultante de la Sélection Naturelle sur les organismes et les facteurs découverts par Mendel constituent le support de l'hérédité. Il faudra attendre les expériences de Frederick Griffith en 1928 pour découvrir et localiser l'ADN dans la cellule, alors appelé « facteur transformant ». La caractérisation biochimique de cette molécule se poursuivra en 1944 avec les travaux d'Oswald Avery et collaborateurs à l'Institut Rockefeller de New-York. C'est à cette même époque qu'un petit groupe de biologistes, dont Ernst Mayr et Ronald Aylmer Fisher, formule la « Théorie Synthétique de l'Evolution » ou « Néodarwinisme », incorporant la variabilité génétique des populations et donnant la première définition tangible et non arbitraire de l'espèce : une espèce est une population d'organismes capables de mêler leurs matériels génétiques (ne pouvant se reproduire qu'entre eux). Cette nouvelle théorie de l'Evolution sur des bases darwiniennes permet notamment d'expliquer le mécanisme de spéciation par lequel une sous-population de l'espèce se retrouve isolée et accumule des changements graduels jusqu'à devenir une

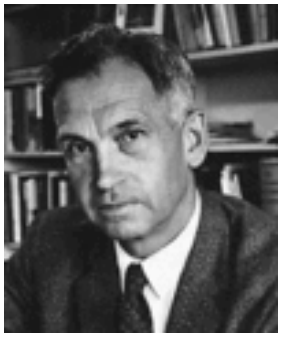
nouvelle espèce incapable de se reproduire avec la première. Ernst Mayr milita pour l'idée qu'un gène seul ne peut être le sujet de la sélection naturelle : il est toujours dépendant d'un contexte d'autres gènes, le rendant ainsi plus ou moins favorable dans le milieu.

Gregor Johann Mendel (1822-1884)



Charles Robert Darwin (1809-1882)

Ronald Aylmer Fisher (1890-1962)



Ernst Mayr (1904-2005)

Figure 4. De l'Hérédité et de la Sélection Naturelle à la Théorie Synthétique de l'Evolution.

Nous verrons par la suite que le concept de « contexte » ou « d'environnement » est en fait capital pour la détermination des fonctions de molécules : dans un milieu isolé, une molécule n'a pas de fonction ; elle est inerte. Ce sont ses interactions avec d'autres molécules de l'environnement qui lui confèrent des fonctionnalités, et qui permettent donc de la décrire et de la définir. En effet, nous redécouvrons depuis peu ce phénomène en observant notamment qu'un composé chimique n'a pas les mêmes effets sur des cellules de tissus différents, et qu'un composé a généralement de très nombreuses cibles cellulaires [16]. Ce point sera davantage étudié dans le chapitre 2, section 4 : « Le criblage virtuel : vers une nouvelle ère du développement de médicaments », ainsi que dans le chapitre « Criblage virtuel des régions moléculaires ».

Plus encore, l'évolution a façonné de nombreuses innovations qui permettent de créer des microenvironnements « contrôlés » à l'intérieur même des cellules ; c'est le cas notamment des différents organites avec, en particulier, la présence d'un noyau protégeant l'ADN et limitant ses interactions avec le restant du contenu cellulaire (Figure 5). Les complexes macromoléculaires peuvent eux aussi former des microenvironnements plus ou moins grands, comme la chambre catalytique de la protéine octamérique Pab87 [17] (Figure 6). A une échelle plus petite, il est intéressant de faire l'analogie avec les sites actifs des enzymes qui localement créent un microenvironnement favorisant une réaction chimique bien précise, tel par exemple le site catalytique de la vanillyl oxydase qui catalyse la

transformation de l'eugénol en vanille (Figure 7). L'Evolution a donc créé plusieurs échelles d'environnements permettant de contrôler l'entrée, la sortie et la concentration de molécules au sein de micro-environnements propices à des interactions particulières et à la catalyse de réactions chimiques. En raison de l'environnement hydrophobe des membranes, les protéines membranaires constituent une famille de molécules très différentes de celles présentes dans le cytoplasme des cellules. Alors que la surface des protéines cytoplasmiques est majoritairement polaire et chargée, la surface des protéines membranaires est majoritairement hydrophobe. Un autre exemple, celui des lysosomes, montre un micro-environnement dans lequel se concentrent des enzymes de dégradation, telles que les lipases, les protéases et les osidases. Avec un pH variant de 3,5 à 5, le micro-environnement (ou milieu) formé par le lysosome est propice à la dégradation des macromolécules (Figure 5). Ainsi, l'Evolution a créé et conservé des organites non seulement pour concentrer des molécules et créer des conditions propices à accélérer des interactions et réactions chimiques telles que la dégradation, mais aussi pour séparer des molécules qui pourraient soit s'autodégrader, soit perturber leurs jeux d'interactions dans la cellule. Quelques perspectives sur la compartimentation des cellules et l'existence d'une surface non-interagissante seront proposées à la fin du chapitre 4 avec le « jeu de la cellule » (Figure 64) en référence au « jeu de la vie ».

Figure 5.Exemple de microenvironnement à l'échelle de la cellule.

(source : http://kailyndix.wordpress.com/)

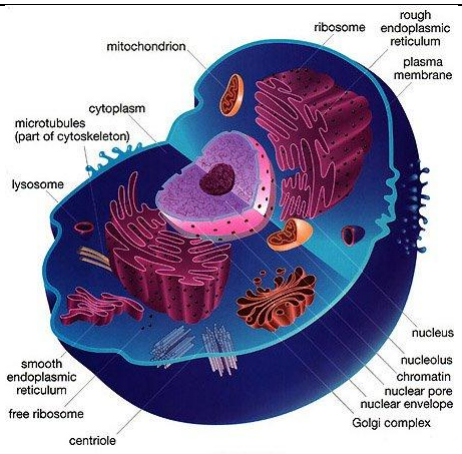


Figure 6. Exemple de microenvironnement à l'échelle macromoléculaire.

Protéine Pab87 de *Pyrococcus abyssi* avec sa chambre catalytique donnant accès à 8 sites actifs *(PDB : 2qmi)*.

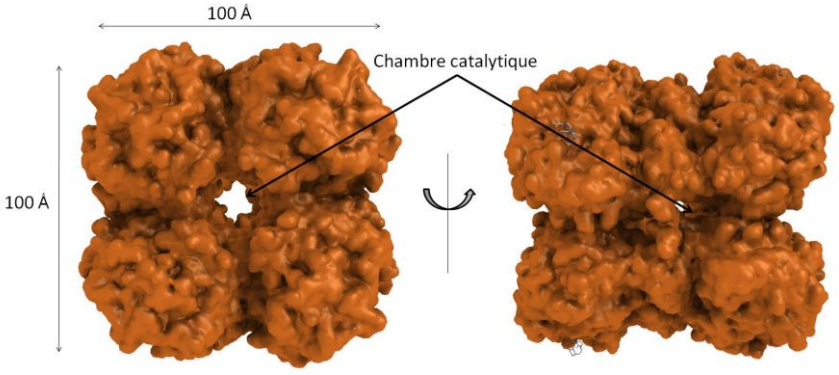
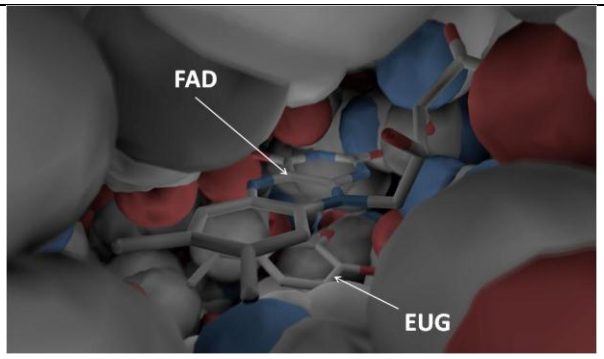


Figure 7. Exemple de microenvironnement à l'échelle moléculaire.

Cavité catalytique de la protéine vanillyl oxydase contenant la Flavin Adénine Dinucléotide (FAD) et l'eugénol (EUG).



4. Des prémices de la Biologie Moderne à une nouvelle ère de la Médecine

« La chance ne sourit qu'aux esprits bien préparés »
Louis Pasteur, XIX^{ème} siècle

Alors que les bases de l'Evolution étaient dévoilées par des chercheurs tels que Lamarck, Mendel et Darwin, d'autres comme Rudolf Virchow énoncèrent que toute cellule provient d'une cellule « *Omnis cellula e cellula* » et que la vie est la somme d'actions physiques et chimiques couplées à l'expression de l'activité cellulaire. Virchow révolutionna ainsi la Médecine en démontrant par l'observation au microscope les fondements de la pathologie cellulaire : une maladie se produit parce qu'une cellule bien portante est altérée ou perturbée [18]. Il milita activement pour les nouvelles pratiques médicales telles que l'observation clinique et la pathologie anatomique et fonda la pathologie comparative par laquelle on observe le développement de maladies chez des animaux, ce qui marque le début de l'expérimentation animale.

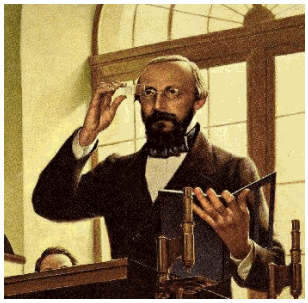
Parallèlement à ces travaux, un chercheur français du nom de Louis Pasteur invalida par une série d'expériences la théorie de la génération spontanée longtemps soutenue par les traités d'Aristote, portant ainsi un coup décisif à cette ancienne théorie de l'apparition de la Vie. Alors que Virchow pensait que les maladies ne pouvaient provenir que d'une pathologie cellulaire, Pasteur démontra que les maladies pouvaient également provenir de l'infection par des virus et par des bactéries. Pasteur et coll. révolutionnèrent ainsi la Médecine en proposant la première méthode efficace contre les infections virales : le vaccin qui consiste en l'injection de petites quantités de virus moins virulents pour préparer le système immunitaire à le reconnaître.

Si le vaccin est un traitement qui valut à Pasteur sa renommée internationale, il ne peut toutefois traiter que les infections virales. C'est en 1928 qu'un écossais, Alexandre Fleming, fit la découverte extraordinaire des antibiotiques et de la pénicilline. Cette substance,

produite par la bactérie *Penicillium notatum* dans certaines conditions de milieu, permit en effet de soigner la plupart des infections bactériennes telles que les staphylocoques [19].

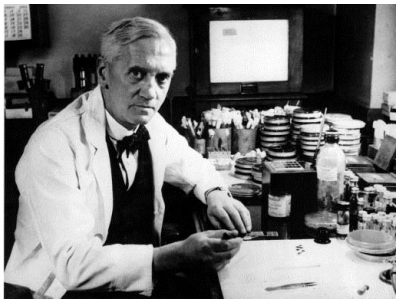
En 1949, après que James Neel ait démontré la transmission mendélienne de l'anémie falciforme ou drépanocytose, Linus Pauling démontre par une mesure de la mobilité électrophorétique (liée au nombre de groupes ionisables), que cette pathologie se développe en raison d'une altération chimique (mutation) de l'hémoglobine [20]. Par ce travail, Linus Pauling démontre et illustre pour la première fois l'origine moléculaire d'une maladie et des mutations, ce qui contribuera au rapprochement progressif de la Biologie et de la Chimie pour la compréhension du vivant et des pathologies.

La découverte de la Théorie Cellulaire, des pathologies cellulaires, des micro-organismes, des vaccins et des antibiotiques, ainsi que la description moléculaire des pathologies, sont à l'origine même de notre médecine moderne et des recherches pharmaceutiques.



Rudolf Virchow
(1821-1902)

Louis Pasteur
(1822-1895)



Alexander Fleming
(1881-1955)

Linus Pauling
(1901-1994)

Figure 8. Les pères de la médecine moderne.

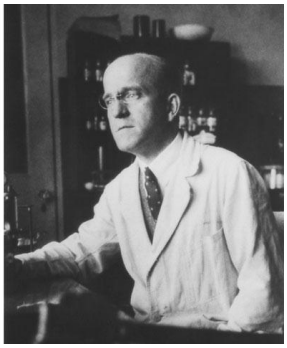
5. De la caractérisation de l'ADN à la Biologie Structurale

« La caractéristique unique, universelle et essentielle des êtres vivants est la possibilité de conserver la structure chimique (ADN) au sein de laquelle est écrit le code génétique »
Jacques Monod, 1988

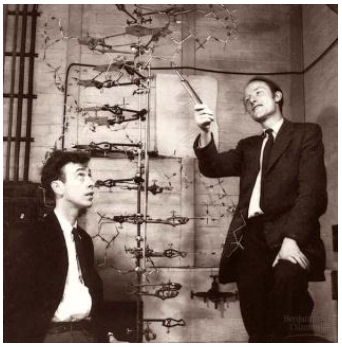
Suite à la caractérisation biochimique de l'ADN par Oswald Avery et coll. en 1944 à l'institut Rockefeller de New York, un jeune duo de Cambridge, composé de James Watson et Francis Crick, détermine à l'aide des règles d'Erwin Chargaff (rapport A/T = G/C = 1) et des clichés cristallographiques de Rosalind Franklin, la structure tridimensionnelle de l'ADN en 1953 [21]. Le vivant et l'hérédité ont dès lors une structure et une base chimique tangibles, observables et analysables. L'élucidation de cette structure permit à son tour de mieux comprendre comment le matériel génétique pouvait non seulement se reproduire au quasi-identique de génération en génération cellulaire (mitose), mais aussi, comment le

patrimoine génétique d'un individu pouvait se diviser pour former des gamètes (méiose). Ces résultats sur la structure de l'ADN, accompagnés de la découverte des structures de l'hémoglobine et de la myoglobine par Max Perutz et John Kendrew marquent le début de la Biologie Structurale [22, 23]. La structure devient une donnée biologique fiable qui permet de mieux comprendre jusqu'aux mécanismes chimiques du vivant : « la structure dicte la fonction ».

Frederick Griffith
(1879-1941)



Oswald Avery
(1877-1955)



James Dewey Watson
(1928 - .)
Francis Crick
(1916-2004)

Rosalind Elsie Franklin
(1920-1958)

Figure 9. De la découverte de l'ADN, facteur de l'hérédité, à sa structure chimique.

6. Des Révolutions de la Biologie Moléculaire au Génie Génétique

« Il ne suffit pas de «voir» un objet jusque-là invisible pour le transformer en objet d'analyse ... Il faut encore qu'une théorie soit prête à l'accueillir »

François Jacob, 1976

En 1961, deux groupes de chercheurs, l'un travaillant à Pasadena (S. Brenner, F. Jacob, et M. Meselson), l'autre à Harvard (F. Gros, H. Hiatt, W. Gilbert, C.Kurland, R.Risebrough et J.D. Watson) parviennent à isoler l'acteur du flux transitoire de l'information génétique, l'ARN messager [24, 25]. La même année, Francis Crick propose que le code génétique est constitué de triplets, c'est-à-dire qu'un acide aminé est codé par une succession de trois nucléotides appelée codon [26]. En 1965, le code génétique est entièrement décrypté notamment grâce aux efforts de M. Nirenberg, G. Khorana, et R. Holley [27-29] : il devient possible de décrire comment l'information codée sur l'acide nucléique peut être transposée en protéines. L'universalité du code génétique (bien qu'il existe plusieurs exceptions) est remarquable : en effet, pour que ces mécanismes de traductions et de synthèses soient partagés par la très large majorité des êtres vivants, ce code est nécessairement très ancien.

Ainsi, des chercheurs tels que Carl Woese, ont utilisé les séquences de l'ARN 16S du ribosome afin d'établir une phylogénie du vivant.

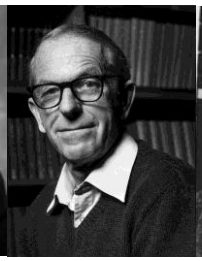
L'essor de la Biologie Moléculaire n'en est pourtant qu'à ses débuts et la discipline sera révolutionnée de nombreuses fois notamment par la découverte d'enzymes permettant de manipuler l'ADN, ou des vecteurs permettant le transfert de l'information génétique dans une cellule donnée. Ainsi, Werner Arber découvre les enzymes de restriction, des protéines capables de découper l'ADN en des positions précises, alors que Bernard Weiss et Charles Richardson découvrent les ligases, des enzymes capables de rabouter l'ADN. En reprenant les découvertes de ses contemporains, Paul Berg parvient en 1972 à faire la première recombinaison génétique, réunissant l'ADN d'une bactérie et d'un virus. En 1973, suite aux travaux de Stanley Cohen et d'Herbert Boyer, *Escherichia coli* devient le premier organisme recombiné par l'homme. En 1955, Frederick Sanger et coll. développent la première technique de séquençage des protéines en déterminant la séquence de l'insuline [30-32]. En 1977, Federick Sanger, Walter Gilbert et coll. développent une seconde technique, capable de séquencer l'ADN [33] : les molécules du vivant ont une séquence unique qui les définit, et cette base chimique tangible va ouvrir la voie à l'analyse haut débit et systématique du patrimoine génétique des êtres vivants. En 40 ans, les progrès dans la compréhension du vivant sont tels qu'en juillet 2010, l'équipe de J.C. Venter annonce avoir créé la première bactérie synthétique, c'est-à-dire, une bactérie ne contenant que l'ADN synthétique généré par ordinateur à partir d'une collection de gènes d'intérêts. La bactérie nommée « *Mycoplasma mycoides JCVI-syn1.0* » est capable d'autoréplication et exprime les phénotypes souhaités, encodés dans l'ADN synthétique [34].

François Jacob
(1920 - .)
Jacques Monod
(1910 - 1976)
André Lwoff
(1902 - 1994)

François Gros
(1925 - .)

Werner Arber
(1929 - .)

Bernard Weiss
(1925 - .)


Frederick Sanger
(1918 - .)

Walter Gilbert
(1932 - .)

Figure 10. L'ère de la biologie moléculaire : de l'ARN messager aux enzymes de restriction et ligases, jusqu'au séquençage des protéines et acides nucléiques.

II. Biologie Moléculaire

« Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself »

Francis Crick, 1988

Après une brève description de la cellule et du phénomène de mutation, cette partie est consacrée à la caractérisation chimique et fonctionnelle des deux grands types de macromolécules biologiques : les protéines et les acides nucléiques. Bien que l'importance de l'ADN ne soit plus à démontrer, nous détaillons davantage les rôles et structures des ARN pour lesquels nos connaissances n'ont cessé d'évoluer au cours des 15 dernières années. Notre objectif est de donner un aperçu du contenu cellulaire qui définit un environnement unique dans lequel les molécules « naissent, vivent, et meurent ». En effet, l'environnement est si important et indissociable du fonctionnement du vivant, que nous proposons de définir la fonction d'une molécule comme étant la somme de ses interactions dans un milieu défini. Si bien que la fonction d'une molécule (1) doit toujours être relative à un environnement cellulaire, et (2) que sa fonction peut varier d'une cellule à une autre. La compréhension et la description de ces environnements et de leurs contenus moléculaires sont une étape essentielle à toute modélisation bioinformatique, et nous sert à illustrer les problématiques de fonctions et de criblage virtuel abordées dans les chapitres 4 (Analyse haut-débit des interactions moléculaires) et 5 (Criblage virtuel haut-débit des régions moléculaires).

---

1. La cellule et son contenu moléculaire

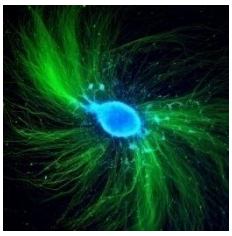
La cellule est donc l'unité fonctionnelle de base du vivant, capable de se reproduire en transmettant à chacune des cellules filles une copie quasi-exacte de son matériel génétique. A chaque génération de cellule, on estime qu'il y a une mutation toutes les 10^4 (virus) à 10^9 (bactéries) paires de bases, en fonction de l'organisme et du locus [35-38]. Chez l'homme, le taux de mutations spontanées est estimé à environ un nucléotide tous les 10^9 nucléotides répliqués [39]. Par exemple, le taux de mutation des microsatellites, très utilisés en cartographie génétique, est beaucoup plus élevé que dans le restant du génome. De façon similaire, le système immunitaire humain est capable de produire des anticorps d'une grande variété et constitue un exemple surprenant où le taux de mutations des cellules somatiques est de six à sept fois plus élevé que dans le restant du génome, sans toutefois engendrer de mutations nuisibles ou létales [40]. Si une grande partie des erreurs commises lors de la réplication de l'ADN par l'ADN polymérase est corrigée [41], des mutations

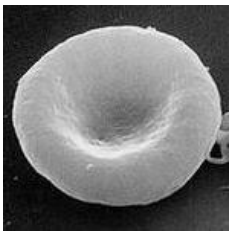
persistent qui, si elles touchent les cellules germinales, pourront être transmises aux descendants, ou alternativement, se transmettre uniquement aux cellules filles ; ces mutations induisent une variabilité génétique propice ou non aux divers milieux rencontrés, variabilité qui sera ou non conservée par la pression de Sélection Naturelle.

a. L'homme et la pluralité de ses paysages cellulaires

Ainsi, le corps humain est constitué de plus de 10 000 milliards de cellules et l'on estime que près de 20 millions de cellules se renouvellent chaque seconde. Le génome de l'Homme comprend environ 3,3 milliards de paires de bases, ce qui permet d'évaluer qu'en moyenne, trois à quatre mutations surviennent lors de chaque réplication du génome. A l'échelle du corps humain, c'est donc 60 à 80 millions de mutations qui interviennent à chaque seconde et aboutissent à une variabilité génétique à l'intérieur même du corps humain : la variabilité génétique ne s'observe pas seulement entre des organismes différents, mais également entre les cellules d'un même organisme multicellulaire.

Chez les organismes multicellulaires, les cellules vont se différencier au cours de l'embryogénèse pour créer différents types de cellules et de tissus tels que les tissus cardiaques, musculaires et neuronaux, chacun spécialisé dans l'accomplissement d'une tâche précise. Plus de 200 types cellulaires ont pu être identifiés chez l'homme. On sait que seule une fraction du patrimoine génétique est exprimée dans un type cellulaire donné, ce qui aboutit à des environnements cellulaires (une collection d'ARN exprimés, de protéines et de modifications de concentrations de sels, de pH, …) très différents. L'expression différentielle du patrimoine génétique permet parfois de différencier les cellules sur le seul critère de leur morphologie (ex : neurone, érythrocyte, fibroblaste, …).


Cellule neuronale


Cellule sanguine

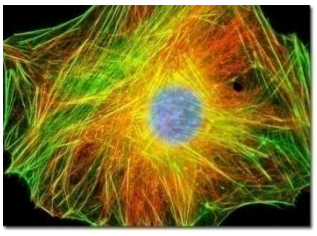

Cellule fibroblastique

Figure 11. Trois exemples de types cellulaires.

De plus, l'expression du matériel génétique d'une même cellule ou lignée cellulaire peut évoluer en fonction du temps et de son stade dans la division cellulaire. Cette expression différentielle du matériel génétique dans différents types cellulaires et à différents stades du

développement vient s'ajouter aux facteurs de compartimentation, ce qui contribue à créer un ensemble complexe d'environnements où la composition en protéines, ARN, et autres molécules varie de façon significative.

b. Nombre de molécules et composition chimique des cellules

L'hydrogène (H), l'oxygène (O), l'azote (N), le carbone (C), le souffre (S) et le phosphore (P) constituent 99% de la masse cellulaire. Une cellule est majoritairement composée d'eau, d'acides nucléiques, de protéines, de polysaccharides et de lipides. L'ADN est la plus grande des biomolécules contenues dans la cellule et est réparti dans des unités fonctionnelles capables de réplications : les chromosomes.

Molécule	Pourcentage du poids hydraté de la cellule	Différents types de molécules / cellule
Eau	70	1
Acides Nucléiques		
ADN	1	1
ARN	6	
Ribosomes		3
Transferts		40
Messagers		1000
Protéines	15	2000-3000
Polysaccharides	3	200
Lipides	2	50
Nucléotides et métabolites	0.8	200
Acides aminés et métabolites	0.8	100
Ions inorganiques	1	20

Tableau 1. Composition chimique approximative d'une cellule d'E.coli en divisions rapides. *(Source : JD Watson, Molecular Biology of the Gene, 2nd ed., Philadelphia, PA : Saunders, 1972)*

Ni la taille du génome, ni le nombre de chromosomes ne semblent être corrélés avec le degré de complexité de l'organisme : ainsi, l'homme possède 23 paires de chromosomes et 3,3 milliards de paires de bases, alors que le blé possède 6 jeux de 7 chromosomes correspondant à près de 16 milliards de paires de bases. Le génome du blé est toutefois

constitué de 80% de séquences répétées contre 50% pour le génome humain. L'état des connaissances semble à présent indiquer que la complexité de l'organisme n'est pas codée dans la taille des génomes, mais plutôt dans la complexité des interactions de ses biomolécules. Ainsi, en 2008, on évalue à 650 000 le nombre d'interactions protéine-protéine chez l'homme, et à 30 000 le nombre d'interactions protéine-protéine chez *S.cerevisae* [42]. Si la complexité d'un organisme est effectivement codée par ses réseaux d'interactions, alors le phénomène de compartimentation est sans doute un marqueur important de la complexité de l'organisme.

2. Acides nucléiques

a. Composition chimique

Les acides nucléiques sont des polymères de nucléotides constitués majoritairement d'Adénine (A), Cytosine (C), Guanine (G), Thymine (T) et Uracile (U). Il existe toutefois quelques nucléotides tels que l'hypoxanthine dont on retrouve des traces, notamment dans les anticodons d'ARNt [43]. Les règles d'Erwin Chargaff et les travaux de Watson et Crick précisent que les bases A et T s'apparient en formant 2 liaisons hydrogènes, alors que les bases C et G s'apparient en formant 3 liaisons (Figure 13).

Chacun de ces nucléotides est formé d'un groupement phosphate et d'un nucléoside lui-même formé d'une base azotée et d'un sucre (Figure 12). L'ADN ou Acide Désoxyribonucléique, et l'ARN ou Acide Ribonucléique, diffèrent chimiquement à la fois par leurs sucres (désoxyribose pour l'ADN et ribose pour l'ARN), mais aussi par le type de nucléotides qui les constituent. Ainsi l'ADN est constitué de A, C, G ou T alors que l'ARN est composé des nucléotides A, C, G et U.

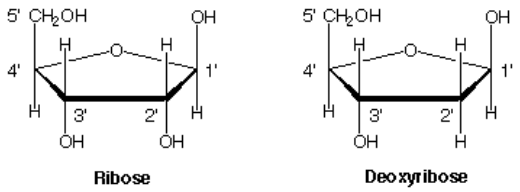


Figure 12. Structures du ribose et du désoxyribose. *Le groupement alcool –OH en 2' permet de différencier les deux sucres. L'ADN qui ne dispose pas de ce groupement alcool est plus stable que l'ARN.*

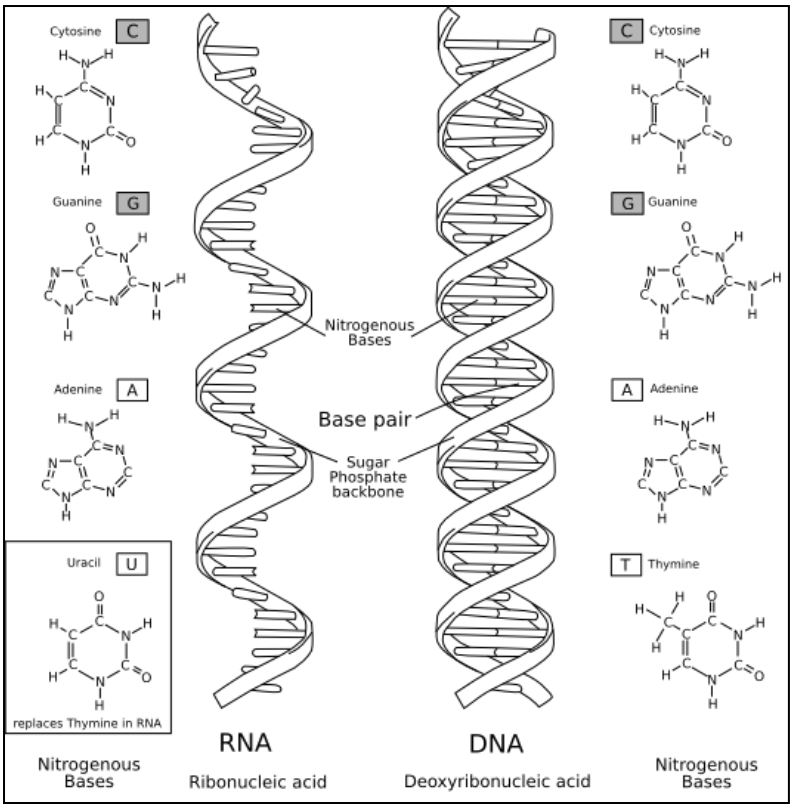


Figure 13. ADN, ARN et leurs bases azotées.

Nous verrons par la suite que les nucléotides peuvent de plus être méthylés par des ADN méthyltransférases, fournissant ainsi un mécanisme supplémentaire de régulation de l'expression des gènes. Certaines méthylations de l'ADN peuvent être transmises à la descendance, portant ainsi un phénotype particulier ou touchant à la régulation d'un gène. De tels mécanismes permettant la transmission d'un phénotype ou touchant à la régulation d'un gène sans toutefois modifier la séquence de l'ADN, sont appelés épigénétiques. Avec les mutations spontanées, les modifications épigénétiques contribuent à la variabilité de l'expression du patrimoine génétique, et donc des molécules contenues dans le milieu cellulaire.

b. La diversité des ARN et de leurs fonctions

ARNm (ARN messager): découvertes en 1961 [24, 25], les molécules d'ARNm sont des copies transitoires d'une portion d'un chromosome, plus précisément de gènes. Chez les eucaryotes, leur rôle est particulièrement important puisqu'ils sont les vecteurs nécessaires du flux de l'information génétique depuis l'ADN protégé dans le noyau jusqu'aux machineries traductionnelles (ribosomes) plus exposées de la cellule.

ARNt (ARN de transfert): la traduction, par les ribosomes, d'un ARNm en protéines requiert la mobilisation d'acides aminés libres. Ceux-ci sont portés de façon spécifique par les ARNt et sont délivrés dans le ribosome lors de la synthèse peptidique.

ARNr (ARN ribosomal): le ribosome est une gigantesque macromolécule constituée d'environ 35% de protéines et 65% d'ARN ribosomal répartis en deux sous-unités : 50S et 30S pour les procaryotes, 60S et 40S pour les eucaryotes. Le site actif est formé structuralement par de l'ARN, et contribua à prouver le concept de « ribozymes » : tout comme les protéines, les ARN sont capables de réactions enzymatiques.

miARN (micro ARN): les miARN sont de petits polymères de 22 nucléotides en moyenne qui peuvent se lier par complémentarité aux régions 3' d'ARNm. Le génome humain pourrait produire près de 1000 miARN, chacun capable de réguler l'expression de plusieurs gènes.

siARN (silencing ARN): les siARN ressemblent grandement aux miARN, toutefois ils possèdent une séquence complémentaire quasi-parfaite d'un seul ARNm. Alors qu'un miARN peut cibler les produits de plusieurs gènes, un siARN ne peut cibler que les produits d'un seul gène avec précision. Ce type d'ARN vient s'ajouter aux nombreux outils de la biologie moléculaire pour étudier le fonctionnement des gènes.

La découverte croissante de la diversité des fonctions d'ARN soutient l'hypothèse de l'origine d'un monde à ARN où celui-ci pourrait à la fois contenir le matériel génétique, l'exprimer, et procéder à des catalyses réactionnelles [44]. Toutefois, la versatilité de ces molécules et leurs fragilités auraient pu conduire à l'apparition des molécules d'ADN, plus stables et moins réactives, c'est-à-dire capables de mieux préserver les innovations de l'Evolution.

La découverte et la reconnaissance de multiples fonctions importantes de l'ARN pour la vie cellulaire sont relativement récentes. D'un point de vue structural, comme nous le verrons, il n'existe encore que peu de données (environ 1200 structures d'acides nucléiques dans la PDB). Cependant, celles-ci permettent d'aider dès maintenant à la compréhension de leurs mécanismes de repliements et d'interactions avec d'autres molécules [45]. Il sera donc de plus en plus nécessaire de tenir compte des ARN dans la modélisation bioinformatique d'environnements et de microenvironnements. Toutefois, la détermination de leurs structures constitue toujours un véritable défi en raison de leurs instabilités intrinsèques. En particulier, nous verrons dans la partie « Criblage haut-débit des régions moléculaires » que la fonction d'une molécule ainsi que la spécificité (fréquence) des régions qui la composent, sont évaluées sur une base de données formée de l'ensemble des régions de molécules présent dans l'environnement. L'apport des structures d'ARN permettra d'identifier plus

facilement des interactions protéine-ARN, mais aussi de mieux évaluer la fréquence des régions moléculaires, en tenant compte de ces molécules d'un autre genre.

c. *Structures secondaires et méthodes prédictives*

L'appariement des bases C-G, A-T et A-U, ainsi que les appariements Wobble permettent aux acides nucléiques de former des structures secondaires : les deux brins s'apparient et forment une ou plusieurs hélices compactes et flexibles à partir de liaisons hydrogènes de faibles énergies. Cet appariement relativement « fragile », permet par exemple à l'ADN de s'ouvrir localement pour laisser agir la machinerie réplicationnelle et transcriptionnelle.

Les ARN sont quant à eux souvent simple brin, et le potentiel d'appariement de leurs bases les conduit à se replier sur eux-mêmes pour former notamment des épingles à cheveux (*hairpin*) (Figure 14). Nous verrons lors de l'étude des interactions protéine-ARN que selon que l'ARN est simple ou double brin, et selon qu'il mime ou non une structure de double hélice, il n'interagira pas de la même façon avec les protéines.

La prédiction des structures secondaires et tertiaires des ARN repose sur la maximisation de ces appariements entre bases. Toutefois, la réalité biologique est plus complexe et il faut tenir compte des pseudonoeuds, des sur-enroulements de l'ARN, qui compliquent considérablement la prédiction de leurs structures. Une prochaine étape de la modélisation d'un environnement cellulaire sera atteinte lorsqu'il sera possible de modéliser et de prendre en compte la présence des ARN. Si des programmes tels que Mfold [46] et Pfold [47] permettent déjà de prédire ces structures, de nombreux développements seront toutefois nécessaires pour parvenir à des modélisations exploitables, prenant en compte l'importante flexibilité de ces molécules.

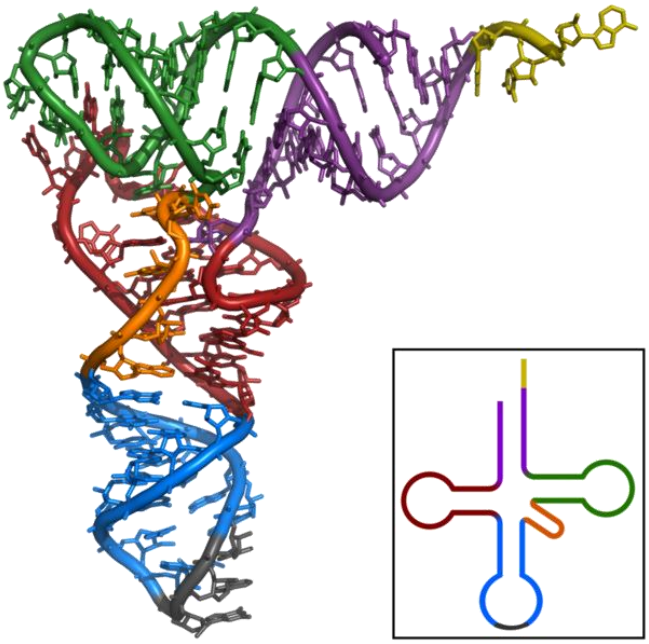


Figure 14. Structure secondaire et tertiaire d'un ARN de transfert. *Les appariements entre bases sont indiqués en rouge et forment des structures d'épingles à cheveux. Le site accepteur d'acide aminé en 3' est indiqué en jaune et le site de l'anticodon en gris. Les autres régions correspondent à différentes boucles de l'ARN, dont la boucle D en bordeaux. (Source : http://upload.wikimedia.org/wikipedia/commons/b/ba/TRNA-Phe_yeast_1ehz.png)*

3. Protéines

Les protéines sont des polymères d'acides aminés, encore appelées polypeptides. Elles sont issues de la traduction des ARNm par les ribosomes qui effectuent la synthèse peptidique. Leur diversité a longtemps marqué les esprits au point que l'on croyait jusqu'aux expériences de F. Griffith, puis de O. Avery [48], qu'elles pouvaient être le support de l'hérédité. Bien qu'elles soient le produit de l'information génétique et non le support, les protéines sont présentes dans la plupart des mécanismes cellulaires et jouent une multitude de rôles :

- **Rôles structuraux** : les protéines s'associant de façon spécifique pour former des macro-assemblages fonctionnels. Les fibres d'actine permettent par exemple la contraction musculaire, alors que les protéines ribosomales forment un gigantesque édifice comprenant plusieurs sites actifs, et capable de lier différents ARN, dont les ARN de transferts. La taille importante d'un grand nombre de protéines permet de plus de lier une pluralité de partenaires différents et de parvenir ainsi à une régulation fine de leurs fonctions.

- **Rôle dans la reconnaissance moléculaire** : de par leur diversité de taille, de forme et de répartition des charges, les protéines sont capables de former des édifices structuraux variés, et plus particulièrement de proposer des sites de liaisons potentiellement très différents les uns des autres. Ainsi, à la fois la forme globale de la protéine, mais aussi les propriétés locales de ses sites de liaison, permettent de moduler la spécificité de la reconnaissance d'un partenaire. A titre d'exemple, on peut citer les récepteurs nucléaires qui ont un repliement commun clairement défini, mais dont la spécificité de reconnaissance d'un partenaire se joue au niveau du site de liaison en modulant la distribution de propriétés géométriques et physico-chimiques.

- **Rôles enzymatiques** : la catalyse d'une réaction par enzyme peut augmenter la vitesse d'une réaction chimique de façon considérable. La réplication de l'ADN est par exemple catalysée par une polymérase spécifique qui permet à la fois de copier et de corriger ce matériel génétique.

- **Rôles de régulateurs** : les polypeptides de moins de 20 acides aminés, que nous appellerons aussi « petits peptides » dans cette thèse, sont capables de mimer les comportements régulateurs de ligands endogènes et exogènes. On notera par exemple les fragments peptidiques du coactivateur SRC-1 des récepteurs nucléaires observés dans plusieurs structures (PDB : 3kmg). Les polypeptides de plus grandes tailles, aussi appelés protéines, sont quant à eux capables de réguler des processus cellulaires, notamment par des changements conformationnels induits par des variations du milieu tel que le pH. Ainsi, la GTPase Ras passe d'une forme « on » qui signale la croissance cellulaire, à une forme « off » lorsque le GTP est hydrolysé en GDP [49].

Fin des années 1990, il fut découvert qu'une même protéine « clair de lune » (*moonlight protein*) pouvait remplir plusieurs fonctions dans des processus cellulaires distincts [50]. Les multiples fonctions de ces protéines sont parfois directement régulées par des conditions de l'environnement cellulaire. La résolution des structures des protéines moonlight I-Anil et PutA suggèrent que leurs différentes fonctionnalités s'expriment par le biais de sites de liaisons distincts [51].

Les protéines et autres polypeptides sont par ailleurs décrits par 4 niveaux de structures selon la nomenclature de Linderstrom-Lang [52] :

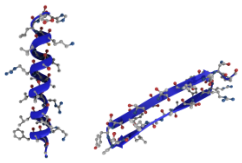
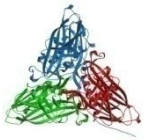
Structure primaire	correspond à la succession d'acides aminés (séquence).
Structure secondaire 	correspond aux repliements locaux du squelette dus aux liaisons hydrogènes entre les groupements δ^+ N-H et δ^- C=0. Ces structures secondaires sont généralement classifiées en hélices, feuillets bêta et boucles.
Structure tertiaire	correspond au repliement global du polypeptide et de ses structures secondaires. Il est toutefois important de noter qu'une même protéine ou une même chaîne polypeptidique peut avoir plusieurs structures tertiaires dépendantes des conditions de l'environnement et des partenaires en association.
Structure quaternaire 	correspond à l'association de plusieurs chaînes polypeptidiques.

Figure 15. Les différentes échelles de structures des protéines

a. Acides aminés

Les acides aminés réguliers sont au nombre de 20 (Tableau 2). Formés à partir d'un squelette chimique commun contenant un groupement amine (-NH2 / $-NH3^+$) et un groupement carboxylique (-COOH / $-COO^-$), les acides aminés se différencient par leur chaîne latérale R portée par le carbone alpha (Cα). Cette chaîne latérale confère à chaque acide aminé des propriétés physico-chimiques remarquables (hydrophobie, acidité, flexibilité, volume, …). Bien que la majorité des acides aminés contenus dans les protéines soient sous forme Lévogyre pour permettre la formation de structures secondaires stables, il n'est pas rare de retrouver des formes Dextrogyres (D-Alanine ou D-Lysine) dans les parois bactériennes (Figure 16). On notera qu'il existe des enzymes capables de dégrader les acides aminés sous formes Dextrogyres.

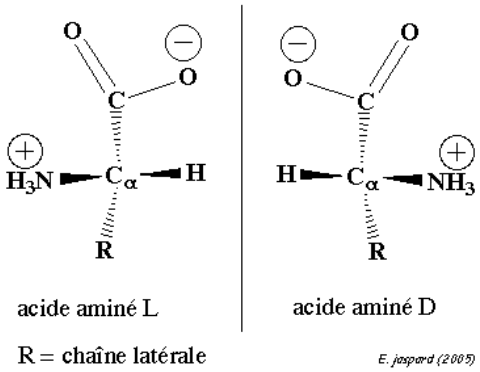


Figure 16. Les deux énantiomères L (Lévogyre) et D (Dextrogyre) des acides aminés avec leur carbone alpha (Cα) et leur chaîne latérale R.

Pour certains acides aminés, la chaîne latérale R possède un groupement chimique capable de s'ioniser en fonction du pH de l'environnement : c'est le cas notamment de la tyrosine (pKa de 10.1), de l'arginine (pKa de 12.1), de la lysine (pKa de 10.67), de l'histidine (pKa 6.04), de l'acide aspartique (pKa 3.71), ou encore de l'acide glutamique (pKa 4.15). La facilité à retirer ou ajouter un proton sur ces groupements dépend toutefois également de l'environnement dans lequel ils sont plongés, et est fonction de la polarité des résidus voisins. Un grand nombre d'approches existent pour prédire ou calculer le pKa des chaînes latérales à partir des données structurales [53-55].

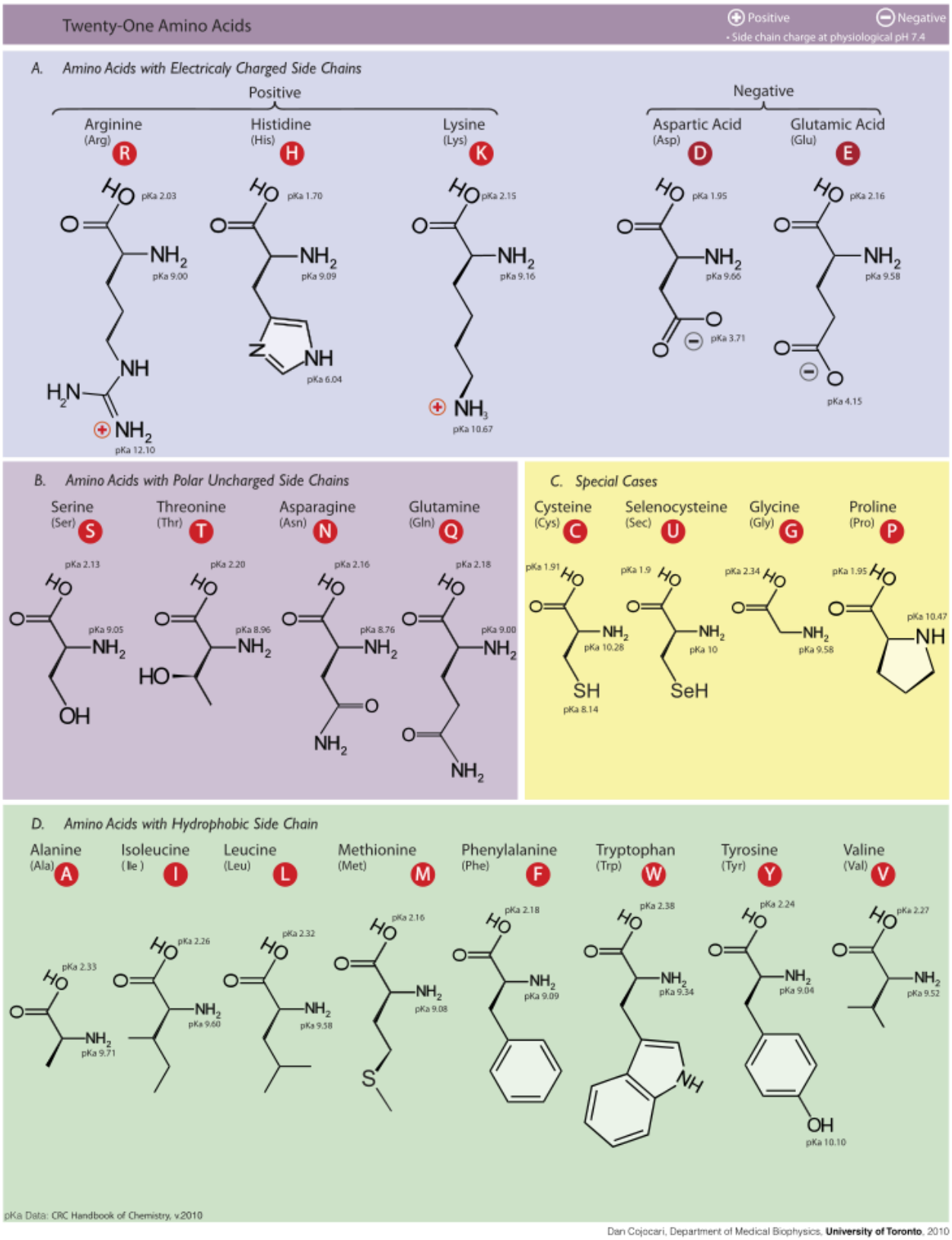


Tableau 2. Les 20 acides aminés réguliers et la sélénocystéine. *Lorsque la chaîne latérale est ionisable, le pKa est précisé.*

De très nombreuses analyses décrivent en détails les propriétés physico-chimiques de ces acides aminés [56, 57], toutefois il est particulièrement difficile de les classifier en raison de la diversité chimique qu'ils représentent. Si l'on considère uniquement les charges, 5 groupes sont généralement admis :

- **Apolaires** : *glycine, alanine, valine, leucine, isoleucine, méthionine, phénylalanine, proline, tryptophane*
- **Polaires sans fonction acide/base** : *serine, thréonine, glutamine, asparagine*
- **Polaires avec fonction acide/base anionique à pH7 :** *acide aspartique, acide glutamique*
- **Polaires avec fonction acide/base neutre à pH7** : *histidine, cystéine, tyrosine*
- **Polaires avec fonction acide/base cationique à pH7** : *lysine, arginine*

Toutefois, il est possible d'argumenter sur cette classification que la glycine, avec pour seule chaîne latérale un hydrogène, pourrait être polaire. De la même façon, si la lysine et l'arginine possèdent un groupement cationique $NH3^+$ à pH 7, leur chaîne latérale est longue et hydrophobe. Le pKa de 6 de l'histidine lui vaut également d'être soit neutre soit cationique en fonction du pH local où elle se trouve. D'autres classifications sont possibles, notamment en fonction de la taille de l'acide aminé (impliquant soit son encombrement stérique, soit sa masse en Da), ou bien encore sa flexibilité (capacité à adopter d'autres conformations). Les aromatiques sont par exemple des résidus peu flexibles et avec un fort encombrement stérique, alors que la leucine est un résidu relativement petit et flexible.

Le diagramme de Venn a ainsi été utilisé par William Taylor [58] afin de proposer une représentation simple des distances entre acides aminés en fonction de leurs propriétés physico-chimiques et évolutives (Figure 17). Chaque sous-groupe représente des acides aminés ayant des propriétés similaires.

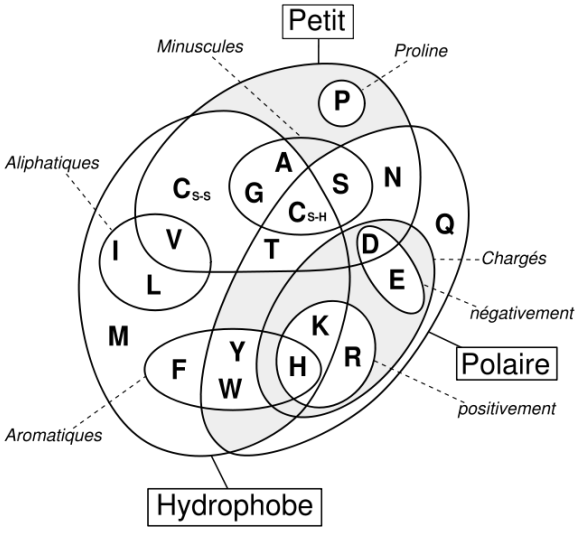


Figure 17. Diagramme de Venn des 20 acides aminés réguliers. *Ce diagramme regroupe les acides aminés en fonction de leurs tailles et polarités.*

Une classification que nous utiliserons souvent au cours de ce travail est fonction des potentialités d'interactions entre résidus :

- **Hydrophobe** : *glycine, alanine, valine, leucine, isoleucine, méthionine, proline, cystéine*
- **Aromatique** : *phénylalanine, tyrosine, tryptophane, (histidine)*
- **Polaire négatif** : *sérine, thréonine*
- **Polaire polyvalent** : *asparagine, glutamine*
- **Anionique** : *acide aspartique, acide glutamique*
- **Cationique** : *lysine, arginine, (histidine)*

En effet, les hydrophobes et aromatiques sont ici séparés en raison de la polyvalence des fonctions des aromatiques : les effets mésomères de leurs cycles permettent de participer à un grand nombre de types d'interactions, formant notamment des empilements avec d'autres aromatiques, ou capables d'interagir avec des charges anioniques ou cationiques. De la même façon, l'histidine peut jouer un double rôle, d'aromatique et de cationique. L'asparagine et la glutamine ont quant à elles la particularité de présenter et un groupement cétone =O (δ^-), et un groupement amine -NH2 (δ^+).

Ces dernières décennies ont également permis de démontrer la présence de résidus d'acides aminés modifiés dans les protéines, notamment la sélénocystéine et l'hydroxyproline.

b. Liaison peptidique et squelette polypeptidique

Lors de la traduction de l'ARNm en protéine par le ribosome, une chaîne polypeptidique se forme par la condensation du groupement carboxyle d'un premier résidu avec le groupement amine d'un second résidu. L'ajout d'un nouvel acide aminé se fait par l'arrivée d'un aminoacyl-ARNt au site A du ribosome, puis par la formation d'une liaison peptidique entre ce résidu et la chaîne en formation. La liaison peptidique est stabilisée par mésomérie et provoque une rigidité du squelette en limitant la libre rotation autour de son axe. Les atomes C, O, N et H de la liaison peptidique sont coplanaires, et deux acides aminés adjacents forment alors un plan dièdre avec un angle de torsion phi (Φ) sur la liaison N-Cα, et un angle de torsion psi (ψ) sur la liaison Cα-C (Figure 18).

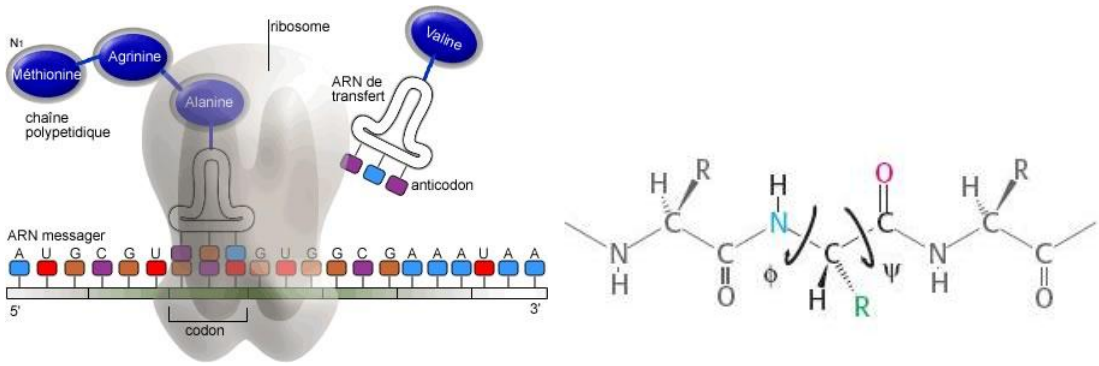


Figure 18. Représentation simplifiée d'une chaîne polypeptidique en formation, d'une liaison peptidique et des angles de torsion phi (Φ) et psi (ψ).

c. Structure primaire

La structure primaire d'un polypeptide correspond à l'enchaînement des acides aminés le constituant. La détermination des séquences de ces polypeptides, le plus souvent à partir du séquençage haut-débit des génomes [59-62], est devenue une étape essentielle permettant leurs identifications et caractérisations fonctionnelles. La séquence du polypeptide peut ensuite être utilisée par des approches de bioinformatique pour prédire des structures secondaires [63, 64], tertiaires [65], et dans certains cas, des sites de liaison [66, 67]. Toutefois la détermination précise de la séquence primaire d'un polypeptide après sa formation dans la cellule (post-traductionnelle) reste difficile en raison de la présence de modifications de certains acides aminés, notamment des glycosylations qui peuvent radicalement changer sa fonction.

En 1998, les protéines à motifs répétés (Figure 19) représentaient 14% de toutes les protéines contenues dans la SwissProt [68] ; une étude démontra qu'elles étaient davantage représentées chez les eucaryotes que chez les procaryotes [69]. Quel que soit le taxon considéré, il semble qu'en dessous de 500 acides aminés, la taille des protéines n'ait pas d'incidence sur la fréquence et le nombre des motifs répétés. Ces répétitions de motifs peuvent être regroupées en deux classes : des motifs avec une faible complexité (une succession d'un seul acide aminé ou d'un petit nombre d'acides aminés similaires), et des motifs avec une forte complexité (généralement plus longs et contenant une plus forte diversité d'acides aminés). Seules 4% des protéines à motifs répétés se retrouvent à la fois chez les eucaryotes et procaryotes, ce qui suggérerait que la plupart des motifs répétés ont émergé après la divergence eucaryote-procaryote. Le nombre et la présence des motifs répétés peuvent avoir une incidence sur des maladies telles que la maladie de Huntington [70].

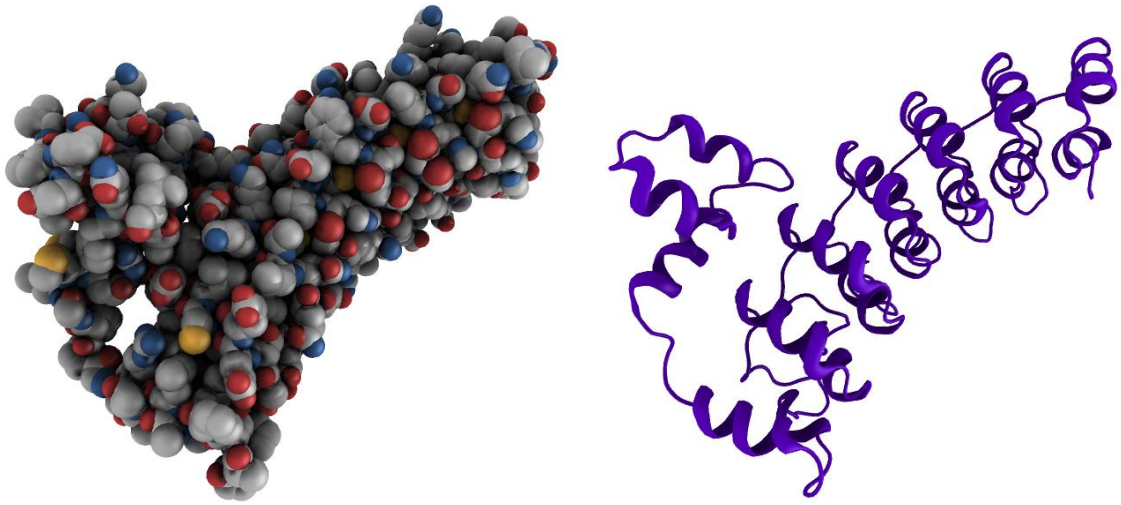


Figure 19. Observations de motifs répétés au niveau de structures secondaires. Ici, le motif répété de l'ankyrine (PDB : 2xai).

d. Structure secondaire

Ce sont les travaux de Linus Pauling et de Robert Corey, qui démontrèrent en 1951 la présence de motifs répétés et réguliers dans les structures protéiques, qu'ils appelèrent hélices alpha et feuillets bêta [71]. Les liaisons hydrogènes entre les groupements C=O (groupe accepteur, δ^-) et N-H (groupe donneur, δ^+) du squelette permettent en effet trois grandes classes de repliements locaux du squelette : les hélices, les feuillets bêta et les boucles (Figure 20). Ces liaisons hydrogènes sont de faibles énergies (10-20kj.mol^{-1}, soit près de 10-20 fois moins énergétiques qu'une liaison covalente) et expliquent en partie la flexibilité conformationnelle des protéines et de leurs squelettes. Dans le cœur protéique, la présence d'hélices et de feuillets bêta permet de neutraliser les charges partielles δ^+ et δ^- du squelette, et donc de favoriser la constitution d'un cœur hydrophobe stable.

La taille physique des atomes limite les conformations que peut prendre un polypeptide, et donc les valeurs autorisées des deux angles de torsion (Φ, ψ) du squelette. Le diagramme de Ramachandran [72] permet à la fois de délimiter les couples de valeurs (Φ, ψ) autorisés (ne provoquant pas de gène stérique), mais aussi de localiser les structures secondaires remarquables, en particulier les hélices alpha, les feuillets bêta parallèles et les feuillets bêta anti-parallèles. Contrairement aux boucles, les hélices et feuillets bêta forment des motifs réguliers et répétitifs du squelette qui peuvent être observés et décrits par les valeurs des angles de torsion Φ et ψ (Figure 21). L'aspect périodique et régulier des hélices alpha peut notamment permettre de former des hélices amphipathiques avec une face polaire et une face apolaire. Le pas d'une hélice pour faire 360° étant en moyenne de 3.5 résidus, il suffit

d'avoir un résidu apolaire tous les 3 à 4 résidus pour obtenir une hélice avec une face apolaire. L'hélice alpha est stable, même isolée.

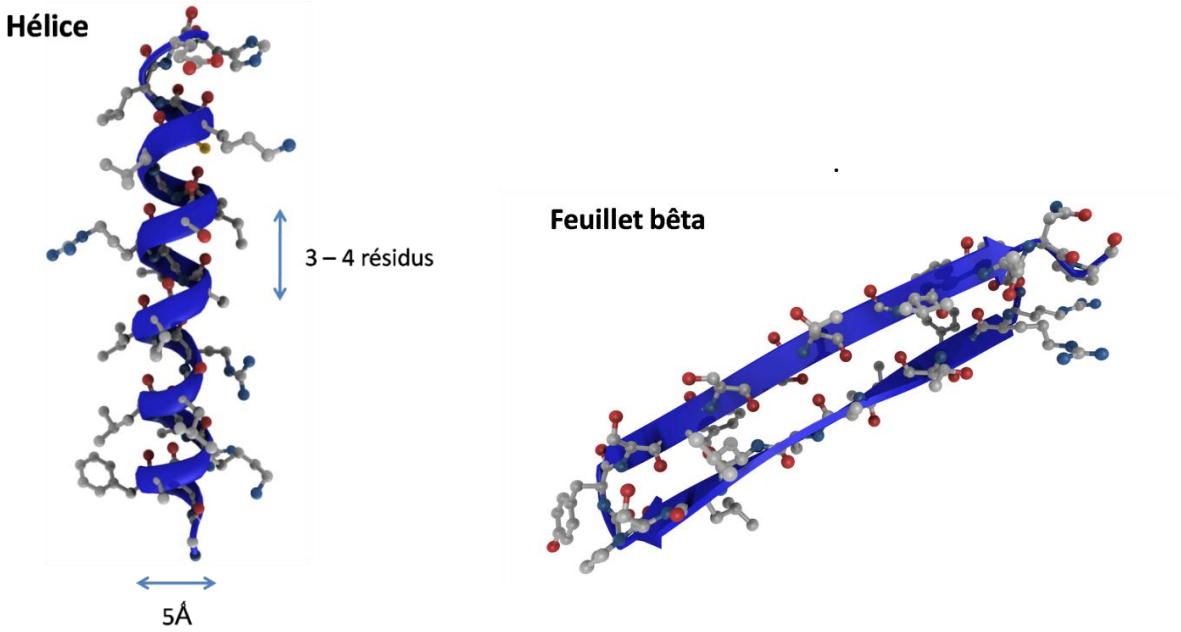


Figure 20. Représentations des structures secondaires d'hélice alpha et de feuillet bêta.

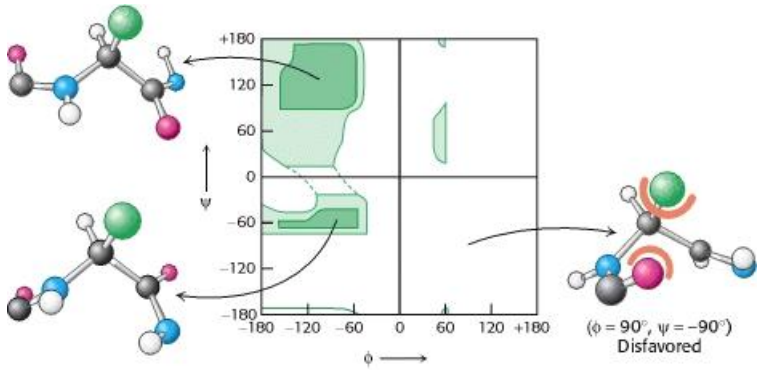


Figure 21. Diagramme de Ramachandran et exemples de conformations de deux résidus en fonction de leurs angles de torsion Φ et ψ. *A gauche, un dipeptide subit une rotation de 180° sur la liaison Cα-C (angle de torsion ψ). A droite, un exemple de couple (Φ, ψ) provoquant une gêne stérique.*

La détermination des structures secondaires est une étape importante de l'analyse des protéines. Elles peuvent donner des indices renseignant sur la structure tertiaire de la protéine, et sur sa localisation cellulaire (les protéines en baril sont ainsi exclusivement localisées dans les membranes). Un grand nombre d'approches ont été développées pour les

prédire soit à partir des données de séquence avec une précision de l'ordre de 75% [63, 73-75] soit à partir des coordonnées atomiques des structures [76]. La détermination des débuts et fins de ces structures secondaires reste cependant problématique. De plus, des structures secondaires peuvent se former lors de l'assemblage de deux protéines (Figure 22).

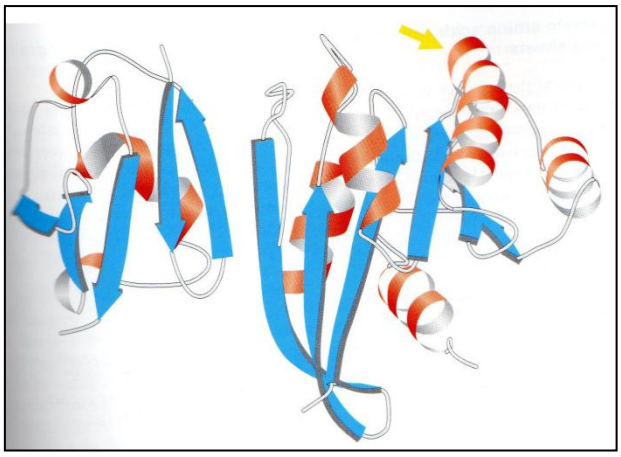


Figure 22. Exemple du complexe Rap-Raf formant un feuillet bêta à l'interface. *(PDB : 1gua, source : Petsko et coll. [49]).*

En complément des structures secondaires décrites ci-dessus, des alphabets structuraux [77-79] ont été établis afin de fournir une description plus détaillée des conformations des squelettes polypeptidiques. En particulier, ces alphabets structuraux ont été utilisés pour discrétiser la structure tertiaire d'une protéine en une séquence de lettres, ainsi que pour aider à la comparaison et à la prédiction *ab initio* des structures tertiaires de protéines [78, 80]. En utilisant les algorithmes d'alignement de séquences, les alphabets structuraux peuvent également servir à aligner les structures très rapidement.

L'agencement et la succession des structures secondaires sur la séquence protéique sont si importants qu'ils permettent souvent d'identifier l'appartenance de la protéine à une famille, à une super-famille et à un repliement [81].

e. Structure tertiaire

Le repliement global d'une chaîne polypeptidique et de ses structures secondaires correspond à la structure tertiaire. La détermination de ces structures peut notamment se faire via des expériences de cristallographie aux rayons X et de RMN.

En 1960, Christian Anfinsen et ses collègues posèrent la question des relations entre la structure primaire, secondaire et tertiaire des protéines. Par une série d'expériences, ils

montrèrent que des protéines après dénaturation pouvaient ré-adopter seules leurs structures tertiaires en solution : « l'hypothèse thermodynamique » [82]. En d'autres termes, l'information contenue dans la séquence primaire serait alors suffisante pour retrouver la structure tertiaire de la protéine. Toutefois, on connaît à présent l'existence de protéines qui nécessitent des chaperonnes afin d'adopter leur conformation native [83-85]. Les forces gouvernant le repliement des protéines sont essentiellement, les interactions hydrophobes de Van der Waals, mais aussi les interactions polaires (liaisons hydrogènes) et chargées (ponts salins). Comme pour toute interaction, ces forces dépendent de l'environnement où elles sont plongées. Aussi le repliement d'un polypeptide est-il influencé par la présence d'eau dans le milieu qui pousse les résidus hydrophobes à se rassembler et à s'enfouir dans un cœur hydrophobe pour minimiser les contacts entre atomes polaires et apolaires. La structure tertiaire d'un polypeptide devrait de plus être étudiée en fonction de sa taille :

Petit peptide : les polypeptides de moins de 20 acides aminés sont trop petits pour former un cœur et peuvent jouer le rôle de ligand venant se loger dans de petites crevasses ou cavités de protéines plus grosses. Bien que fréquents dans les structures moléculaires, les petits peptides sont en réalité très souvent des fragments de polypeptides plus gros.

Peptide : les polypeptides de plus de 20 acides aminés, mais de moins de 60 ou 80 acides aminés, sont également souvent trop petits pour disposer d'un cœur protéique hydrophobe stabilisant, toutefois leur taille conséquente leur permet d'adopter plusieurs conformations différentes. Notamment, c'est dans cette catégorie que l'on retrouve un certain nombre de protéines désordonnées.

Protéine : les polypeptides de plus de 60 ou 80 acides aminés, comportant presque toujours un cœur protéique hydrophobe stabilisant la structure complète. Ces grandes structures disposent de vastes surfaces permettant de diversifier leurs modes d'action notamment en autorisant la liaison de partenaires de différents types moléculaires.

En termes prédictifs, il est généralement possible d'établir un bon modèle 3D d'une protéine par des méthodes de modélisation par homologie, si l'on dispose d'une empreinte (*template*) ayant au moins 30% d'identité de séquence [86-88]. Toutefois, il n'est pas rare d'observer des protéines ayant des séquences partageant moins de 30% d'identité de séquence mais adoptant un repliement similaire ; c'est le cas notamment des immunoglobulines [89]. Aussi, la comparaison des structures tertiaires des protéines ont été utilisées pour déterminer de nouveaux liens de parenté évolutifs. Une étude récente tend à montrer que la structure tertiaire serait en effet de 3 à 10 fois plus conservée que la structure primaire [90]. D'autres méthodes sont possibles, notamment le *threading* ou les prédictions *ab initio / de novo* qui ne nécessitent pas de structures homologues mais sont

plus difficiles à automatiser. A ce jour, les approches *ab initio / de novo* ne sont réellement applicables qu'aux polypeptides ayant une taille inférieure à 100 ou 150 acides aminés.

Les structures tertiaires sont généralement comparées par l'agencement dans l'espace de leurs structures secondaires par des algorithmes tels que VAST [91] ou DALI, ou par l'alignement de fragments structuraux par des approches combinatoires telles que CE [92]. L'approche CE a notamment été utilisée pour calculer les cartographies moléculaires présentées dans l'article de M-ORBIS [1] ; les alignements des chaînes structurales réalisés par CE ont permis de vérifier si les résidus de sites de liaison homologues partageaient des positions similaires dans l'espace. D'autres approches pour aligner ces structures sont toutefois possibles, notamment en les discrétisant à l'aide d'alphabets structuraux et en les alignant à partir de ces encodages. Plus récemment, l'approche développée par Lee Sael et coll. permet ainsi d'aligner rapidement les structures tertiaires de protéines en alignant leurs surfaces grâce aux harmoniques sphériques et aux descripteurs 3D de Zernike [93].

La comparaison par paire des structures tertiaires des protéines contenues dans la Protein Data Bank [94] a permis leur classification hiérarchique dès les années 1995, avec la création de bases de données dédiées, telles que SCOP [95] ou CATH [96]. Ces deux bases de données regroupent les protéines en fonction de la ressemblance de leurs structures tertiaires et plus précisément de l'agencement de leurs structures secondaires (Figure 23).

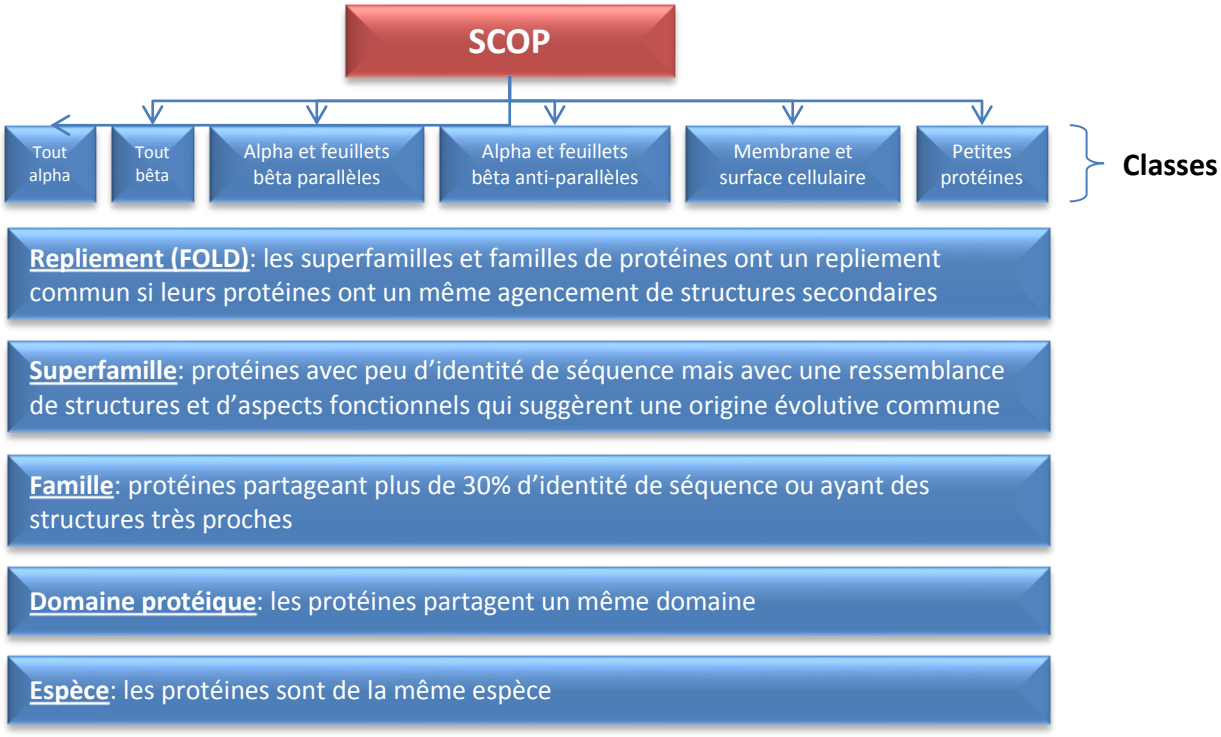


Figure 23. Classification hiérarchique des structures protéiques dans SCOP. *La base de données SCOP regroupe les structures moléculaires en fonction de leurs identités de séquences, et de leurs structures secondaires et tertiaires.*

Pour comprendre la valeur des données structurales, il est important de noter qu'une même protéine peut être représentée parfois par plusieurs centaines de structures décrivant ses fonctions et sa dynamique dans différents contextes environnementaux (chapitre Analyse haut-débit des interactions moléculaires). L'étude des structures tertiaires permet donc de mieux comprendre et décrire à la fois les fonctions et la dynamique d'une molécule, mais également de mieux appréhender la diversité des mécanismes d'interaction.

La perte de la stabilité d'une structure tertiaire de protéines, que ce soit par mutation ou par un changement de l'environnement, est l'un des facteurs importants des maladies monogéniques [97].

f. Structure quaternaire

Les chaînes polypeptidiques peuvent s'assembler pour former un assemblage qui constitue alors la structure quaternaire d'une protéine. Un grand nombre de protéines n'ont de fonction/d'activité dans la cellule que lorsqu'elles adoptent une structure quaternaire en liant une ou plusieurs autres chaînes peptidiques [98]. La structure quaternaire a donc également un rôle dans la régulation des fonctions cellulaires (Figure 24).

En effet, la fonction d'une protéine dans un contexte cellulaire peut être vue comme la somme des interactions auxquelles elle participe, ce qui, comme nous l'avons vu précédemment, indique que la fonction d'une molécule dépend de l'environnement dans lequel elle évolue. La perte ou le gain d'une capacité à interagir avec un partenaire X peut provoquer des maladies [99, 100] ou au contraire avoir des effets bénéfiques. Par exemple, la concentration des enzymes de dégradations dans les lysosomes permet de limiter leurs interactions avec d'autres molécules, et donc, de les dégrader. Le rôle des interactions et des structures quaternaires est donc fondamental à la fois pour la compréhension des mécanismes moléculaires menant à une fonction cellulaire, mais également dans le développement de candidats médicaments, notamment pour comprendre et agir sur le développement d'une pathologie [101].

Ces assemblages sont bien souvent transitoires (des protéines déjà repliées se rencontrent pour former une structure quaternaire), mais il existe également des assemblages permanents, notamment pour les protéines oligomériques [102]. Dans certains cas, les assemblages permanents se font en même temps que le repliement des chaînes polypeptidiques qui les constituent.

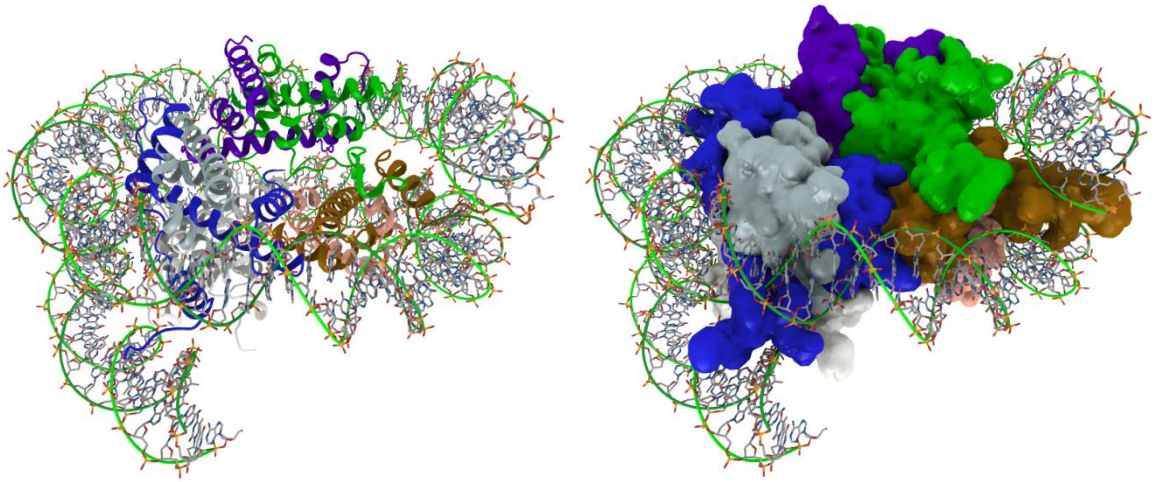


Figure 24. Le nucléosome, une structure quaternaire constituée de 8 chaînes polypeptidiques d'histone. *Le nucléosome a une structure quaternaire (PDB : 3afa) qui permet à l'ADN de s'enrouler et de compacter la chromatine.*

D'autres protéines dites « désordonnées » n'ont pas de structure stable en solution, toutefois lorsqu'elles s'associent avec une autre protéine pour former un assemblage quaternaire, elles peuvent alors adopter une structure stable [103-105]. La formation d'un assemblage moléculaire peut donc induire des changements conformationnels plus ou moins locaux, et peut drastiquement changer la stabilité des partenaires associés.

La prédiction des structures quaternaires de protéines est d'autant plus difficile qu'il faut également prendre en compte la présence de multiples contacts cristallins [106] dans les structures issues des expériences de cristallographies (Chapitre « Structure, Surface et interactions moléculaires ») (Tableau 3).

De nos jours, l'importance des interactions entre molécules pour comprendre leur rôle dans le fonctionnement de la cellule ou d'un système biologique précis, a pris tant d'importance que de nombreuses approches expérimentales ou prédictives ont été créées pour les déterminer. Des méthodes telles que le TAP-Tag [107], le double hybride [108], l'anisotropie de fluorescence [109] ou bien encore le FRET/BRET [110-112] permettent de déterminer des interactions entre protéines avec une précision décroissante, fonction du haut-débit de la méthode. *In silico*, les données structurales ont été utilisées pour prédire les sites de liaisons sur les protéines [67, 113-121], mais aussi pour prédire les possibles conformations d'assemblages protéiques par les méthodes d'amarrages moléculaires ou « *Docking* » [122-128]. Plus récemment, ces méthodes se sont diversifiées pour permettre la prédiction d'assemblages protéine-acide nucléique [127] et protéine-ligand [129-131].

PDB	http://www.rcsb.org/pdb/	Contient les données expérimentales des structures de protéines, d'acides nucléiques et d'assemblages moléculaires.
PQS	http://pqs.ebi.ac.uk/	Prédiction des Structures Quaternaires de protéines basée sur l'aire enfouie des interfaces et un score d'énergie.
PISA	http://www.ebi.ac.uk/msd-srv/prot_int/	Outil intéractif pour l'exploration des interfaces de macromolécules, comprenant la prédiction des structures quaternaires et la possibilité de rechercher des assemblages similaires.
PITA	http://www.ebi.ac.uk/thornton-srv/databases/pita/	Base de données donnant accès à l'unité biologique vraisemblable d'une structure cristallographique, basée sur les opérateurs de symétrie du cristal et le calcul d'un score d'énergie.
ProtBuD	http://dunbrack.fccc.edu/ProtBuD/	Base de données comparative des unités asymétriques (ASU) et des unités biologiques (BU) de la Protein Data Bank (PDB) et de la Protein Quaternary Server (PQS)
VIPERdb	http://viperdb.scripps.edu/	Base de données des structures de capsides virales icosaédriques
PiQSi	http://www.piqsi.org/	Facilite l'exploration et la curation des structures quaternaires

Tableau 3. Bases de données structurales différenciant les contacts cristallins des contacts biologiques. *(source : Janin et coll., 2008 [98])*

III. Biologie Structurale

« If you want to understand function, study structure »
Francis Crick, 1988

Dans la section précédente, nous avons décrit certaines des caractéristiques chimiques et fonctionnelles des protéines et acides nucléiques présents dans une cellule. Cette troisième partie a pour but de présenter les différentes méthodes expérimentales permettant l'obtention des structures de ces molécules, ainsi que de préciser l'étendue des données structurales disponibles en 2010. Nous concluons cette partie sur l'existence de bases de données structurales spécialisées, permettant d'analyser les interactions moléculaires, les phénomènes de phosphorylations, ou bien encore, certaines des variabilités génétiques liées à des maladies monogéniques. Le lien entre le génotype, le phénotype et les structures moléculaires sera davantage abordé dans l'article de SM2PH-db (Structural Mutation to Pathology Phenotypes in Human Database) présent dans la section « Publication connexe » de cette thèse.

1. Les méthodes expérimentales

La Biologie Structurale est la Science qui permet de déterminer expérimentalement les structures de molécules à partir de techniques telles que la Résonance Magnétique Nucléaire (RMN), la Cristallographie aux rayons X (X-Ray), ou encore la Cryo-Microscopie Electronique (CME). La discipline englobe également un certain nombre d'analyses combinant les données structurales et de séquences afin d'élucider des mécanismes moléculaires et des fonctions biologiques sur des bases géométriques, physico-chimiques et évolutives. La vision moléculaire des pathologies introduite par Linus Pauling, ainsi que la compréhension des mécanismes de réplication et de translation par observations de la structure d'ADN, ont largement contribué à démontrer l'importance des structures dans l'étude du vivant.

Les méthodes de RMN présentent l'avantage de pouvoir étudier la structure d'une molécule en solution, ce qui facilite l'étude des molécules instables et permet d'avoir une idée de leurs dynamiques conformationnelles. Les structures issues des expériences de RMN représentent 12,9% des structures disponibles dans la PDB (http://www.rcsb.org/pdb/statistics/holdings.do). Cette méthodologie présente cependant deux inconvénients : une résolution assez faible, et ne pouvoir déterminer que de « petites » molécules, généralement moins de 5000 à 10 000 atomes [132] pour les protéines.

La Cristallographie à rayons X (86.4%) permet quant à elle de résoudre des structures de molécules plus conséquentes, pouvant dépasser le million d'atomes pour des capsides virales. Les structures cristallographiques bénéficient d'une meilleure résolution permettant de localiser les chaînes latérales (R < 3 Á), et dans certains cas, de renseigner sur l'emplacement des hydrogènes et des molécules d'eau (R < 2 Á). Toutefois, cette approche nécessite l'obtention d'un cristal, et donc la stabilisation des molécules par de multiples jeux de concentrations ioniques. L'une des difficultés et expertise du cristallographe consiste donc dans les stratégies et les choix de solvant et/ou de partenaires pour stabiliser la ou les molécules, et ainsi former un cristal qui diffracte.

Bien que ne représentant que 0,5% des structures de la PDB, la Cryo-Microscopie Electronique (CEM) est une discipline récente permettant l'observation de macro-assemblages dans leur environnement natif (les molécules restent dans leurs conditions physiologiques), mais à de faibles résolutions (de 7 à 50 Á). Ainsi, le macro-assemblage TFIIH essentiel pour la transcription, est composé de 9 chaînes polypeptidiques pour un poids moléculaire total de 460kDa (Figure 25). Sa structure a pu être déterminée par CEM à la résolution de 38 Á, et il a été possible de localiser par immunomarquage les protéines au sein du macrocomplexe (p44, cdk7, cyclin H, MAT1 et les hélicases XPB XPD [133]).

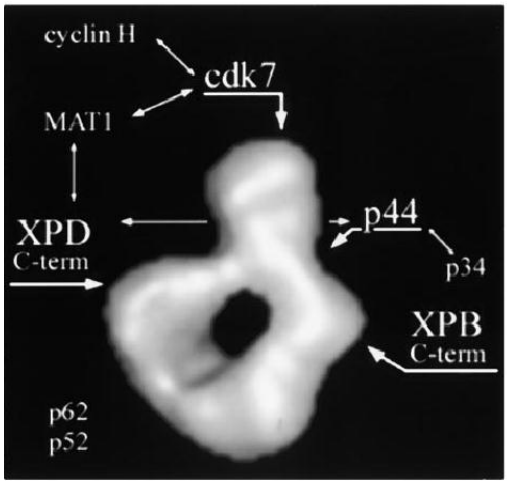


Figure 25. Complexe multi-protéique TFIIH obtenu par cryo-microscopie électronique. *(source : Schultz et coll., 2000 [133])*

L'une des évolutions attendues de la Biologie Structurale sera de fournir les structures de macro-assemblages fonctionnels à des résolutions suffisantes permettant la localisation de leurs sous-unités et l'analyse de leurs interactions. Les autres évolutions concerneront l'obtention d'un plus grand nombre de structures de protéines membranaires, ainsi que d'ADN et d'ARN. L'un des défis des dix à vingt prochaines années consistera de plus à

explorer et prendre en compte la dynamique conformationnelle des molécules et des assemblages.

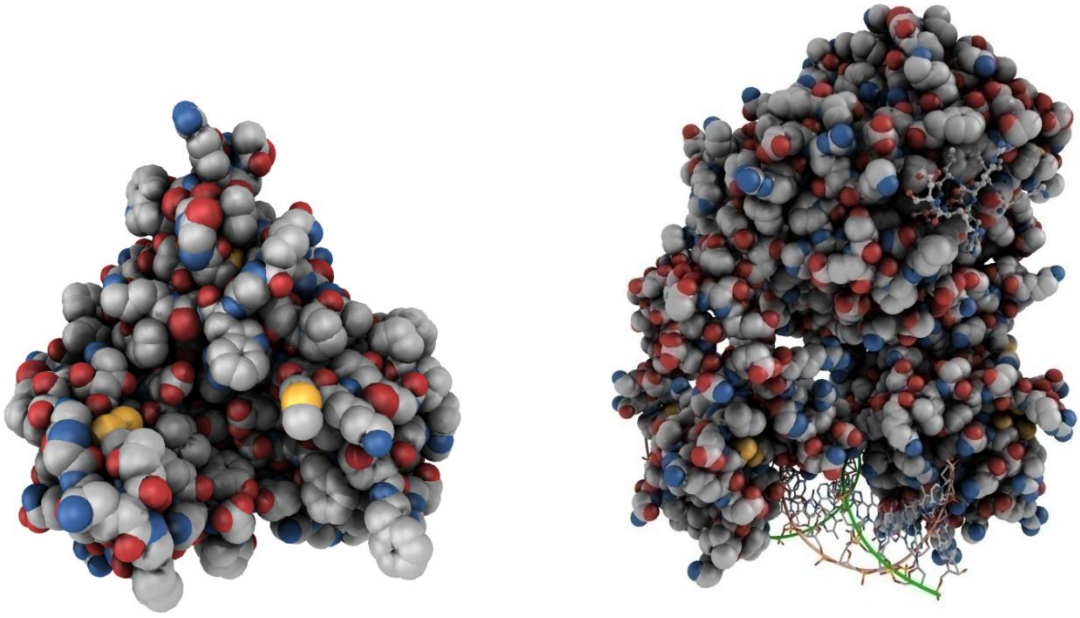
2. Les projets de Génomique Structurale

L'accès aux séquences de plus de 1000 génomes complets de différents organismes (880 bactéries, 65 archea, 114 eucaryotes) a permis d'automatiser l'identification d'un très grand nombre de séquences protéiques (http://www.genomesonline.org). Les projets de génomiques structurales initiés dans les années 2000 reposent sur ces données de séquence et ont pour ambition de fournir une structure à chacune des protéines connues, en tentant de maximiser la découverte de nouveaux repliements [134-136]. En effet, le nombre de repliements (*fold*) des protéines a été estimé entre 1000 et 5000 par Brenner et coll. en 1997 [137] ; et il faudrait déterminer approximativement 10 000 structures cibles pour pouvoir couvrir la quasi-totalité des repliements existants [138]. La sélection de ces cibles a par conséquent provoqué de nombreux débats, certains souhaitant focaliser l'attention sur des protéines propres à leurs systèmes, d'autres souhaitant étudier davantage les structures de protéines à fort potentiel thérapeutique. L'une des directives ayant contribué au succès de ces projets est de donner un libre accès à ces données.

Pour déterminer les structures de chaque protéine connue, un procédé en quatre étapes fut initié : (i) réunir les protéines en familles (ii) sélectionner les protéines les plus représentatives de familles, (iii) résoudre expérimentalement les structures de ces protéines modèles, (iv) modéliser par homologie les structures des autres représentants de chaque famille.

En 2005, près de la moitié des structures mettant à jour un nouveau repliement ont été déterminées par des instituts participant aux projets de génomiques structurales [136]. Plus généralement, ces projets ont permis de diminuer les coûts et temps permettant de déterminer les structures de biomolécules (acides nucléiques compris), de modéliser un grand nombre de structures et de les conserver dans la base de données ModBase [139], d'observer des changements conformationnels, et d'identifier de nouvelles évolutions communes entre protéines [134, 136]. L'une des conséquences « néfastes » de ces projets de grande envergure a été de produire des structures « non fonctionnelles » de protéines, c'est-à-dire des structures ne renseignant pas sur de possibles partenaires ou ligands. A contrario, nous appellerons « structures fonctionnelles », les structures montrant des aspects fonctionnels de la protéine (Figure 26). L'analyse de ces structures « non fonctionnelles » et leur comparaison avec d'autres structures « fonctionnelles » revêtent donc une grande importance afin d'assigner de façon sûre et robuste des fonctions à ces

protéines. Pour y parvenir, il est nécessaire de concevoir et développer des algorithmes performants, rapides et traçables (de sorte que l'on puisse comprendre comment l'inférence d'une fonction est faite et l'invalider/l'améliorer si besoin est) pour analyser et comparer les structures moléculaires. Cette problématique sera détaillée dans le chapitre « Criblage haut-débit des régions moléculaires ».



A) PDB : 1aun **B) PDB : 3dzy**

Figure 26. Structure isolée et structure « fonctionnelle ». *En A), une structure « isolée », ne comportant qu'une seule biomolécule (Osmotin-like protein) entourée de solvant. B) Structure « fonctionnelle » comportant plusieurs biomolécules (RXR, PPAR, coactivators, DNA) qui permettent de caractériser plusieurs de leurs aspects fonctionnels (ex : site de liaison, partenaire).*

En Europe, les projets SPINE (*Structural Proteomics IN Europe*) rassemblent quelques-uns des plus grands laboratoires de biologie structurale d'Europe dont l'IGBMC, afin de développer de nouvelles méthodes et technologies permettant une future cristallographie « haut débit » des molécules. Le projet achevé en 2006 a permis l'obtention de 308 nouvelles structures et l'élaboration de nouveaux standards pour la cristallographie aux rayons X. Une seconde phase du projet, SPINE2 – complexes, a débuté et s'oriente à présent vers la détermination « haut débit » de structures « fonctionnelles », c'est-à-dire, d'assemblages moléculaires (http://www.spine2.eu/).

3. Les données de la Biologie Structurale en 2010

La Biologie Structurale a connu un essor remarquable au cours des dix dernières années, suite aux progrès réalisés lors des différents projets de génomiques structurales. Ceux-ci ont en effet conduit à développer de nouveaux robots et instruments permettant de diminuer la quantité de produits nécessaires et d'accélérer la vitesse d'obtention des cristaux. Des améliorations notables de la résolution cristallographique des structures font que la plupart des structures obtenues en 2010 ont une résolution meilleure que 3 Á. La totalité des structures publiques est stockée sous forme de fichiers texte ou XML, dans la Protein Data Bank (PDB), membre du consortium international WorldWide Protein Data Bank (wwPDB). PDBsum propose de nombreux liens pour étudier différents aspects de chaque entrée PDB, notamment son appartenance à un repliement (*fold*), et des données de conservations (HSSP).

PDB [140]	http://www.rcsb.org/pdb/	Chercher et parcourir plus de 66 000 structures PDB
NDB [141]	http://ndbserver.rutgers.edu/	Structures d'acides nucléiques, certaines n'étant pas présentes dans la PDB
PDB-Ligand [142]	http://www.idrtech.com/PDB-Ligand/	6466 structures de ligands associés majoritairement à des protéines
PDB-TM [143]	http://pdbtm.enzim.hu/	1229 structures des protéines transmembranaires
MSD	http://www.ebi.ac.uk/msd/	Ressources alternatives décrivant les entrées de la PDB
ModBase [139]	http://modbase.compbio.ucsf.edu/	Collection de modèles par homologie de protéines
PDBsum [144]	http://www.ebi.ac.uk/pdbsum/	Collection d'informations relative à chaque entrée de la PDB
Phospho3D [145]	http://cbm.bio.uniroma2.it/phospho3d/	Collection de structures 3D mettant en évidence des phosphorylations
MutDB [146]	http://www.mutdb.org/	Analyse structurale de mutations
SM2PH-db [5]	http://decrypthon.igbmc.fr/sm2ph/	Rassemble les protéines impliquées dans des maladies monogéniques humaines, et fournit des données évolutives et structurales pour étudier les mutations

Tableau 4. Bases de données structurales spécialisées.

Les structures d'acides nucléiques sont également rassemblées dans la Nucleic Acid Database [147, 148]. Il est également possible d'accéder à des modèles par homologie pré-construits de protéines via la base de données ModBase [139, 149, 150]. Enfin, les bases de données MutDB [146, 151, 152] et SM2PH-db [5] rassemblent des données structurales sur les mutations (Tableau 4). Dans ce qui va suivre, nous nous limiterons à la description de la PDB, de ModBase, de MutDB et d'un bref aperçu de SM2PH-db.

a. Protein Data Bank

Seul un petit nombre d'organismes sont à l'heure actuelle représentés dans la PDB. On retrouve en particulier *Escherichia coli*, une bactérie modèle pour l'expression des protéines (17% des structures); *Thermus thermophilus*, archaea modèle pour les systèmes hyperthermophiles (13,4%) ; *Homo sapiens* (10,1%), faisant suite au Human Genome Project, notamment pour des raisons thérapeutiques ; *Saccharomyces cerevisiae*, organisme modèle pour l'étude de nombreux comportements génétiques (7,7%) ; *Mus musculus*, organisme modèle proche de l'Homme souvent utilisé pour tester les effets *in vivo* des candidats médicaments et permettant d'étudier le développement de nombreuses maladies (5,4%).

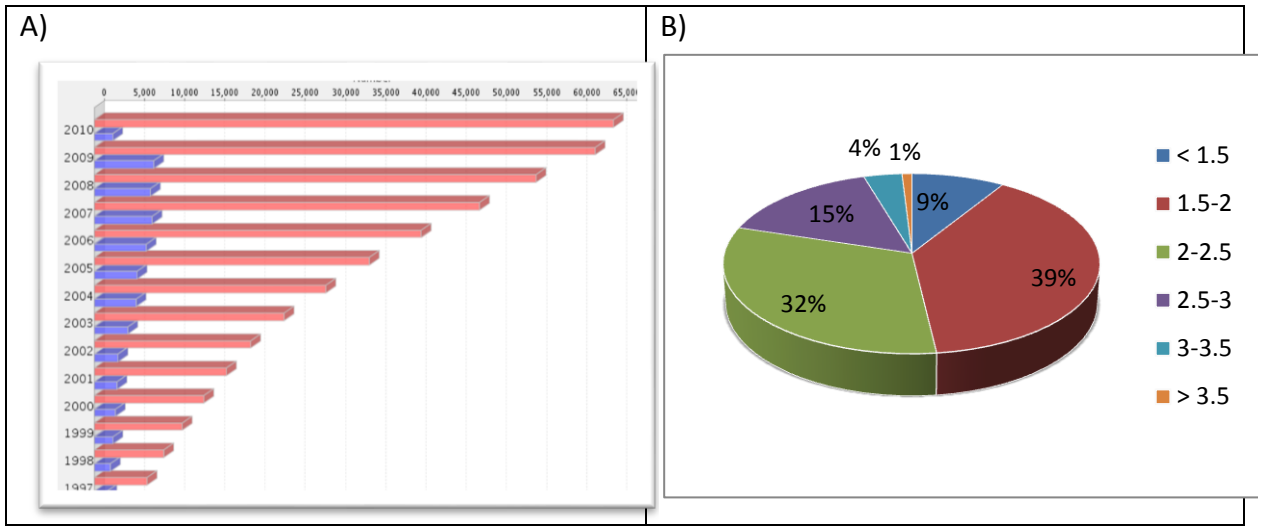


Figure 27. Croissance des structures et répartition par résolution. *A) le nombre total de structures disponibles (rouge) dans la Protein Data Bank, ainsi que la croissance annuelle (bleu). B) distribution de ces structures en fonction de la résolution en Angstrom. Données PDB au 19 avril 2010.*

A partir de 7 Á de résolution, il devient possible de localiser les structures secondaires dans les cartes de densités ; à partir de 3 Á, ce sont les chaînes latérales qui peuvent être localisées. En dessous de 2 Á, l'eau peut être observée avec une résolution suffisante, tandis

que pour localiser les hydrogènes, il faudra compter sur des résolutions de moins de 1.5 Á. Aussi, l'étude des interactions moléculaires se fait-elle le plus souvent sur les structures ayant des résolutions inférieures à 3 Á afin de pouvoir définir la position des chaînes latérales et définir des contacts entre atomes et résidus d'une interface [153-156].

La PDB est une base de données très redondante où il est possible de retrouver un très grand nombre de structures mettant en jeu une même molécule. Entre 1996 et 2006, on estime que près de 70% des structures déterminées correspondaient à des structures déjà déterminées mais présentant soit des mutations clés, soit des partenaires différents [136]. C'est le cas notamment des kinases, famille protéique la plus répandue dans l'espèce humaine (~ 3% des gènes codants) et pour lesquelles plus de 1000 structures sont disponibles ; c'est aussi le cas des récepteurs nucléaires pour lesquels plus de 200 structures sont disponibles. A ce jour, la PDB contient 66 000 structures dont plus de 54 000 contiennent des protéines, et plus de 1200 contiennent des acides nucléiques (http://www.rcsb.org/pdb/statistics/holdings.do). Une analyse que nous avons menée sur la PDB montre qu'elle contient plus de 10 800 ligands différents, et plus de 2800 petits peptides (suivant notre définition d'un polypeptide ayant moins de 20 acides aminés). Une courte liste des ligands les plus fréquents est présentée dans le tableau 5.

Cette redondance des structures est à la fois une richesse et un handicap pour l'analyse structurale : une richesse car elle permet d'étudier une molécule dans une très grande variété de conditions différentes ; un handicap parce que nous ne disposons à ce jour que d'une couverture modérée des protéines exprimées dans différents organismes. On estime par exemple que seules 20% des protéines humaines ont au moins une structure décrite dans la PDB. Nous utilisons la redondance des structures de la PDB dans notre approche de cartographie moléculaire pour donner une vision dynamique de la molécule en fonction de son environnement, et mieux caractériser ses fonctions [1].

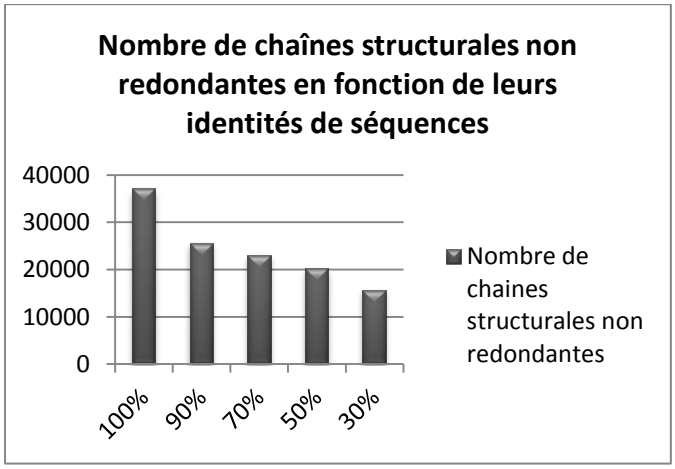


Figure 28. Redondance des chaînes structurales présentes dans la PDB en fonction de leurs identités de séquences. *Il y a 15 000 chaînes structurales ayant un pourcentage d'identité de séquences inférieur à 30%.*

Ainsi que nous le verrons par la suite, la Bioinformatique Structurale joue un rôle majeur à la fois dans la détermination *in silico* de structures moléculaires, mais aussi dans leur analyse approfondie. Alors que le séquençage des génomes et exomes (partie du génome formée par les exons) devient de plus en plus facile et rentable, la prédiction *in silico* des structures protéiques continue de gagner en importance en permettant de modéliser efficacement les petites variations qui différencient chaque individu, patient ou population.

Nom (3 lettres)	Nom complet	Nombre d'atomes du ligand	Formule du composé	Structures contenant le ligand
GOL	*GLYCEROL*	14	C3 H8 O3	4441
NAG	*N-ACETYL-D-GLUCOSAMINE*	30	C8 H15 N1 O6	2343
HEM	*PROTOPORPHYRIN IX CONTAINING FE*	75	C34 H32 Fe1 N4 O4	2226
EDO	*1,2-ETHANEDIOL*	10	C2 H6 O2	1703
FAD	*FLAVIN-ADENINE DINUCLEOTIDE*	86	C27 H33 N9 O15 P2	927
ADP	*ADENOSINE-5'-DIPHOSPHATE*	42	C10 H15 N5 O10 P2	862
MAN	*ALPHA-D-MANNOSE*	24	C6 H12 O6	832
NAD	*NICOTINAMIDE-ADENINE-DINUCLEOTIDE*	71	C21 H27 N7 O14 P2	659
BMA	*BETA-D-MANNOSE*	24	C6 H12 O6	632
BME	*BETA-MERCAPTOETHANOL*	10	C2 H6 O1 S1	600
PLP	*PYRIDOXAL-5'-PHOSPHATE*	26	C8 H10 N1 O6 P1	568
NAP	*NADP NICOTINAMIDE-ADENINE-DINUCLEOTIDE PHOSPHATE*	76	C21 H28 N7 O17 P3	552
MPD	*(4S)-2-METHYL-2,4-PENTANEDIOL*	22	C6 H14 O2	529
FMN	*FLAVIN MONONUCLEOTIDE*	52	C17 H21 N4 O9 P1	492
ATP	*ADENOSINE-5'-TRIPHOSPHATE*	47	C10 H16 N5 O13 P3	488

Tableau 5. Ligands les plus représentés dans la Protein Data Bank. *(données calculées en Août 2010)*

b. ModBase

ModBase est une base de données de modèles par homologie de séquences et de domaines protéiques, maintenue par le groupe d'Andrej Sali pour un grand nombre d'organismes [139, 149]. Elle comporte plus de 5 millions de modèles de domaines et plus de 1,5 millions de modèles de domaines non redondants. Cette modélisation par homologie systématique des protéines accompagne les efforts et la démarche initiée dans les

programmes de génomiques structurales où l'on détermine expérimentalement un représentant de chaque famille protéique avant de modéliser les autres membres de ces familles (Tableau 6).

La modélisation repose sur le programme Modeller [65] qui, à partir de l'alignement d'une protéine à modéliser avec une ou plusieurs empreintes (*templates*), et de base de données structurales de boucles, permet de reconstituer et d'optimiser les torsions du squelette protéique et de positionner les chaînes latérales.

Organisme	**Nombre de transcrits**	**Nombre de séquences partiellement modélisées**	**Pourcentage de transcrits modélisés**	**Source de séquences**
Pyrococcus furiosus	2113	1524	72%	NCBI
Bacillus subtilis	4105	3374	82%	NCBI
Escherichia coli	4206	3150	75%	NCBI
Staphylococcus aureus	2635	1184	45%	NCBI
Streptococcus pyogenes	1691	1440	85%	NCBI
Arabidopsis thaliana	30 707	23 807	78%	ENSEMBL
Drosophila melanogaster	17 104	9381	55%	NCBI
Homo sapiens	32 010	21 270	66%	OrthoMCL-DB
Mus musculus	30 133	25 338	84%	NCBI
Saccharomyces cerevisiae	6600	3035	46%	NCBI
Xenopus laevis	27 952	25 457	91%	NCBI
Uniprot 2005	1 742 816	1 025 196	59%	Uniprot

Tableau 6 : Contenu de ModBase en 2009. *(adapté à partir de U.Pieper et coll.* [139]*).*

c. SNP et MutDB

Les SNP (*Single Nucleotide Polymorphism*) sont la forme la plus commune de variations génétique chez l'homme, représentant plus de 90% des différences de séquences [157] et survenant avec une fréquence de l'ordre de 1 mutation pour 1000 bases [158]. L'étude de

ces variations de séquence (plusieurs millions de SNP ont été recensés) représente un véritable défi pour comprendre comment, et par quels mécanismes, elles peuvent faciliter ou induire des maladies. L'un des enjeux actuels est d'arriver à lier ces informations aux structures afin de donner une explication fonctionnelle, voire mécanistique, du lien entre une maladie et l'occurrence de une ou plusieurs SNP.

La base de données MutDB a été développée dans cette optique, afin d'identifier la conséquence moléculaire probable d'une mutation SNP (qu'elle soit neutre ou associée à des maladies héréditaires), et de la visualiser sur la protéine correspondante. MutDB présente l'avantage de donner un accès direct aux phénotypes dans le cas des SNP associés à des maladies héréditaires, mais ne permet de localiser ces SNP que sur les structures déterminées expérimentalement.

d. SM2PH-db

La base de données SM2PH-db (*Structural Mutation to Pathology Phenotypes in Human Database*), donne accès à un grand nombre d'informations concernant les relations entre génotypes et phénotypes, et facilite la visualisation des SNP et des données évolutives sur des structures expérimentales mais aussi modélisées. SM2PH-db se concentre principalement sur les maladies mendéliennes où l'émergence d'un phénotype pathologique est liée à l'altération d'un seul gène. Les mutations qui affectent les protéines altèrent principalement la solubilité et la stabilité de leurs structures [97, 159], mais peuvent aussi altérer localement la géométrie et la répartition spécifique des charges d'un site de liaison bloquant ou créant ainsi de nouvelles interactions. Les données structurales et de SNP sont de plus liées aux données évolutives issues du programme de gestion des alignements multiples de séquence MACSIMS [160], développé dans notre laboratoire.

Une description plus détaillée de SM2PH-db est disponible dans le chapitre « Publication connexe » de cette thèse.

IV. Bioinformatique

Cette quatrième partie illustre quelques-unes des applications de la bioinformatique pour le traitement massif des données biologiques. Plus précisément, on décrit le processus général de la bioinformatique qui consiste dans la comparaison des données de séquences ou de structures à des fins d'analyses et de prédictions.

La Bioinformatique est la Science qui étudie les données biologiques à l'aide de la capacité de traitement des ordinateurs. Elle débute souvent par une phase d'analyse statistique permettant de révéler les grandes tendances et les éléments importants d'un système biologique (ex : les biais de composition, de conservation dans une famille protéique ou nucléique). Souvent, en se basant sur cette analyse, elle débouche alors sur des programmes prédictifs permettant de renseigner sur des fonctions d'objets biologiques (ex : la prédiction de structure secondaire et tertiaire ; la prédiction de sites de liaisons ; la prédiction de la structure d'un assemblage biologique – ou *docking* -).

1. Bioinformatique des Séquences

« When I started my work, I never imagined that the comprehensive data that we have now would be so readily available. It's phenomenal »

David Lipman, 2009

a. De la comparaison par paire des séquences à l'inférence de fonction

Historiquement, les débuts de la bioinformatique pourraient se situés avec l'apparition des matrices de score PAM [161] et BLOSUM [162], permettant l'alignement de séquences par les algorithmes tels de Needleman & Wunch [163] ou Smith & Waterman [164].

```
# Aligned_sequences: 2
# 1: RXRA_MOUSE
# 2: RARA_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 233
# Identity:      64/233 (27.5%)
# Similarity:   108/233 (46.4%)
# Gaps:          17/233 ( 7.3%)
# Score: 242.5
#
#
#=======================================

RXRA_MOUSE         1 ------------SANEDMP----VEKILEAELAVEPPNDPVTNICQAADK     34
                               :..|..|    :.|...:|..|....|......:.:.|
RARA_HUMAN         1 PEVGELIEKVRKAHQETFPALCQLGKYTTSEQRVSLDIDLWDKFSELSTK     50

RXRA_MOUSE        35 QLFTLVEWAKRIPHFSELPLDDQVILLRAGWNELLIASASHRSIAVKDGI     84
                     .:...||:||::|.|:.|.:.||:.||:|...::||.....|....:|.:
RARA_HUMAN        51 CIIKTVEFAKQLPGFTTLTIADQITLLKAACLDILILRICTRYTPEQDTM    100

RXRA_MOUSE        85 LLATGLHVHRNSAHSAGVGAIFDRVLTELVSKMRDMQMDKTELGCLRAIV    134
                     ..:.||.::|...|:||.|.:.|.|.. ..:::.:::||..|.|.|.||.
RARA_HUMAN       101 TFSDGLTLNRTQMHNAGFGPLTDLVFA-FANQLLPLEMDDAETGLLSAIC    149

RXRA_MOUSE       135 LFNPDSKGLSNPAEVEALREKVYASLEAYCKHKYPEQPGRFAKLLLRLPA    184
                     |...|.:.|..|..|:.|:|.:...|:.|.:.:.|.:|..|.|:|:::..
RARA_HUMAN       150 LICGDRQDLEQPDRVDMLQEPLLEALKVYVRKRRPSRPHMFPKMLMKITD    199

RXRA_MOUSE       185 LRSIGLKCLEHLFFFKLIGDTPIDTFLMEMLEA    217
                     ||||..|..|.:...|:.....:...:.||||.
RARA_HUMAN       200 LRSISAKGAERVITLKMEIPGSMPPLIQEMLEN    232


#---------------------------------------
```

Figure 29. Exemple d'alignement par paire des séquences de RXRα et de RARα. *Les deux séquences mesurent respectivement 217 et 232 acides aminés, et elles partagent 27,5% d'identité de séquence et 46,4% de similarité de séquence. Une identité est représentée par un trait vertical, et une similarité par deux points verticaux. L'alignement a été réalisé à partir du programme Align de l'EBI (http://www.ebi.ac.uk/Tools/emboss/align/)*

Cette étape fondamentale de la bioinformatique a permis pour la première fois de comparer des macromolécules telles que les protéines et acides nucléiques, à l'aide de programmes informatiques (Figure 29) et de leur assigner une appartenance commune. Si l'accumulation des données expérimentales sur des protéines modèles a été une étape nécessaire qui a permis de mieux connaître leurs fonctions, la capacité à comparer ces molécules sur la base de leurs séquences a permis d'assigner par similarité, des fonctions aux molécules nouvellement identifiées et produites. La base de données Uniprot [165, 166] répertorie ainsi plus de 500 000 fiches informatives sur des protéines annotées manuellement, et près de 12 millions de séquences annotées en automatique par de telles approches comparatives.

Au cours des années 1990-2000, les programmes bioinformatiques se sont diversifiés et améliorés pour permettre la recherche systématique et rapide de zones fonctionnelles plus petites. De telles zones peuvent représenter des domaines de tailles variées pour les protéines (exemple : le domaine SH2 est composé d'approximativement 100 acides aminés, alors que le domaine EF-hand n'en contient que 36), ou bien des sites de liaisons sur l'ADN pour des facteurs de transcription tels que la TBP ou les récepteurs nucléaires. La base de données Pfam ne recense que des domaines alors que la base de données Interpro recense également des motifs et sites importants des protéines [167-169].

b. De la comparaison multiple de séquences à l'annotation des protéines

L'alignement multiple de plusieurs séquences [170-173] permet d'aller plus loin dans l'inférence de fonctions des molécules et de caractériser des événements évolutifs au sein d'une famille de protéines, tels que la perte ou le gain d'un domaine, et la conservation ou la coévolution de résidus (Figure 30). Plus généralement, les alignements multiples permettent d'établir un profil des domaines, fonctions, mutations et résidus importants (ex : résidus catalytiques) et de partager ces propriétés aux membres d'une même famille protéique [160]. C'est notamment par l'analyse des alignements multiples de séquences nucléiques qu'il a été possible de localiser les SNP et d'étudier leurs variations. Les alignements multiples de séquences servent également à annoter des génomes complets. Au laboratoire, le programme MACSIMS (*Multiple Alignment of Complete Sequences System Information Management System*) [160] a ainsi été utilisé pour l'annotation des protéines des génomes de l'archaea *Pyrococcus abyssi* ou de la bactérie *Mycobacterium smegmatis* (http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gms) et des banques de cDNA de l'annélide *Alvinella pompejana,* et a permis d'assigner au moins une fonction à plus de 80% des protéines (données non publiées).

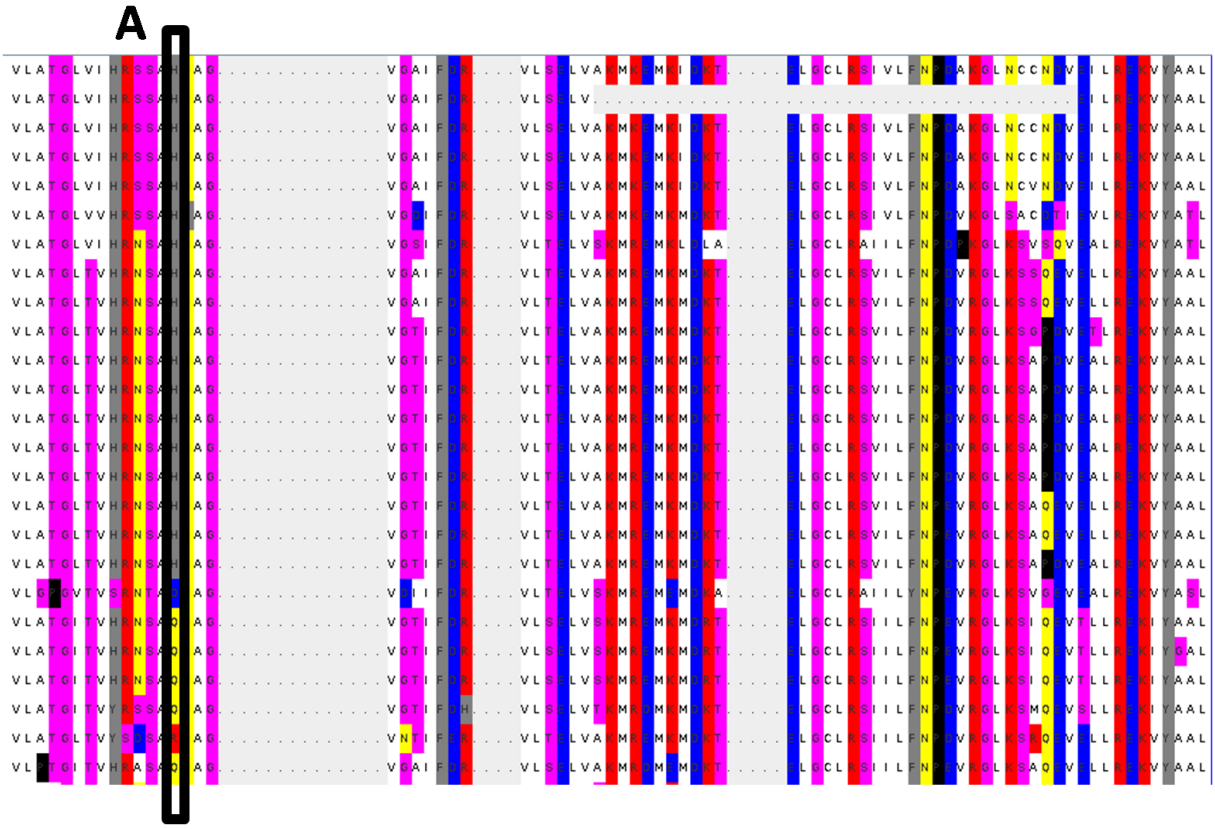


Figure 30. Exemple d'alignement multiple de séquences de récepteurs nucléaires. *Pour chaque résidu d'une molécule étudiée, il est possible d'observer les variations dans les séquences alignées. Ainsi, à la position A, on peut voir que l'histidine (gris) est le résidu le plus fréquent, suivi de l'asparagine (jaune), de l'arginine (rouge) et de l'acide aspartique (bleu).*

En particulier, puisque l'alignement multiple de séquences donne accès à la variabilité des résidus en chaque position, il est possible de quantifier la conservation du résidu, notamment par l'entropie de Shannon issue de la théorie de l'information, et qui a été utilisée pour identifier que les résidus faisant partie du cœur de l'interface sont plus conservés que le restant de la surface [279]. Les notions d'interface et de cœur d'interface seront abordées dans le chapitre « Structure, Surface et interactions moléculaires ».

$$s(i) = -\sum_{k=1}^{20} p(k).\ln(p(k))$$

Où s(i) est l'entropie ou la conservation du résidu (ou d'une classe de résidu) en position i, et où p(k) est la fréquence observée dans l'alignement multiple du résidu de type k. Plus s(i) est faible et plus le résidu est conservé, c'est-à-dire, plus il a été l'objet d'une pression de sélection.

Cette capacité à comparer les objets biologiques rapidement, et à quantifier leurs similarités, a également ouvert la porte aux études systémiques qui tendent à comprendre comment et pourquoi un phénomène biologique se produit. La forte croissance des données structurales a permis à la bioinformatique de se diversifier avec la branche de la Bioinformatique Structurale, spécialisée dans le traitement des structures.

2. Bioinformatique Structurale

La Bioinformatique Structurale est la Science qui étudie les données structurales provenant des expérimentations de la Biologie Structurale. Elle voit ses débuts avec l'apparition de programmes de visualisation moléculaire tels que RasMol [174], ainsi qu'avec les études d'une poignée de pionniers sur deux problématiques essentielles ou « Saint Graal »: (1) quels sont les facteurs permettant à une protéine de se replier, et peut-on prédire une structure à partir de sa seule séquence (modélisation) ? (2) quels sont les facteurs permettant à deux molécules d'interagir de façon spécifique, et peut-on prédire la structure probable de cet assemblage (*docking*) ?

a. Prédiction des structures protéiques

La question est sans doute plus difficile qu'il n'y paraît. Dans un premier temps, les travaux de Anfinsen ont permis de montrer qu'au moins une fraction des protéines étaient capables, après dénaturation, de retrouver leurs formes natives sans aide extérieure [82]. Cette première expérience suggérait alors que toute l'information nécessaire au repliement de la protéine est contenue dans la séquence, et que par conséquent, il devrait être possible de prédire cette structure tertiaire à partir de la séquence.

Toutefois, la découverte des chaperonnes et des *Heat Shock Proteins* (HSP) telles que GroEL a démontré leur rôle fondamental dans le repliement de plusieurs protéines [175]. Une troisième classe de protéines, découverte à la fin des années 90, correspond aux protéines désordonnées, c'est-à-dire les protéines qui n'ont pas de forme 3D native (tertiaire) mais qui peuvent adopter une conformation précise dès lors qu'elles interagissent avec un partenaire.

On différencie les approches de modélisation par homologie telles que Modeler [65], qui ont besoin d'une structure présentant au moins 30% d'identité de séquence avec la séquence de la protéine à modéliser, des approches de *threading* et de modélisation *ab initio* telles que Rosetta [176].

b. *Prédiction des assemblages protéiques*

Depuis les années 1970, de nombreux chercheurs ont caractérisé les interfaces moléculaires et les sites de liaisons associés [45, 153-155, 177-181]. Ces études ont permis de mettre en évidence des différences géométriques et physico-chimiques entre différentes classes d'interfaces : les interfaces homodimériques sont par exemple connues pour être davantage hydrophobes et plus planes que les interfaces hétérodimériques [181] ; les interfaces protéine-acide nucléique sont plus chargées afin de compenser les charges négatives des groupements phosphates de l'ADN ou de l'ARN [45, 179]. Plus important peut-être, il a pu être démontré que les sites de liaisons et interfaces d'assemblages biologiques présentaient des différences notables permettant de les distinguer du restant de la surface non-interagissante et des interfaces cristallines [106, 156, 182, 183]. Ces résultats ont permis d'envisager dès 1975, des programmes informatiques d'amarrage protéine-protéine (ou *docking*) permettant de prédire la conformation d'un assemblage protéique à partir des seules structures des sous-unités [178]. Ces méthodes se sont grandement complexifiées et améliorées avec le temps, permettant par exemple de prédire la conformation d'assemblages protéine-ligand, ou même protéine-acide nucléique. L'initiative de créer le concours international CAPRI (*Critical Assessment of PRediction of Interactions*) de prédiction d'assemblages biologiques, a permis de suivre les progrès réalisés dans ce domaine, ainsi que la diversité des solutions employées. On notera par exemple la possibilité d'utiliser des données biologiques pour guider l'amarrage moléculaire avec le programme HADDOCK [126, 184], ou bien la possibilité de procéder à des amarrages très rapidement grâce à des techniques de hachage géométrique [185].

Des études de 2002 et 2004 indiquent que les approches d'amarrages moléculaires parviennent à identifier correctement 70 à 80% des assemblages protéine-ligand en faisant varier la conformation du ligand [186-188].

c. *Perspectives et apports pour la biologie et biomédecine*

La Bioinformatique Structurale a pu se développer progressivement en raison de ces deux champs d'applications majeurs : (1) pouvoir prédire la structure d'une molécule et (2) pouvoir prédire les interactions entre molécules à partir de leurs structures.

Ces applications jouent en effet un rôle majeur dans le développement de candidats médicaments (*drug discovery*) et ont donné naissance à une nouvelle sous-discipline : le développement de candidats médicaments à partir de structures (*structure-based drug design*) [189-193]. Ainsi que relaté par H. Kubinyi de l'Université de Heidelberg [194], le premier succès de cette discipline fut le développement de la *Captopril*, un médicament utilisé dans le traitement de l'hypertension artérielle : pour la première fois, la structure du

médicament était déterminée de façon rationnelle à partir d'un modèle structural de la cible et d'un site de liaison souhaité. Ce qui semblait alors n'être qu'un rêve inatteignable commença à prendre forme : l'augmentation croissante du nombre de structures biologiques disponibles encouragea le développement « dirigé par les données structurales» de nouveaux médicaments tels que la *Dorzolamide* (antiglaucoma agent), le *Nelfinavir* et l'*Amprenavir* (inhibiteurs de la protéase HIV) ou même, le *Zanamivir*, inhibiteur de la neuraminidase du H1N1, grippe que beaucoup redoutèrent en raison de son niveau de son pouvoir de contagion [186, 194].

V. La régulation de la transcription eucaryote

Cette cinquième et dernière partie de l'introduction générale illustre le phénomène de régulation de l'expression de l'information génétique à partir d'un signal extérieur qui sera perçu par des récepteurs et provoquera une réponse directe ou indirecte entraînant l'activation ou l'inhibition de gènes. Le déchiffrage de certains des mécanismes de régulation des récepteurs nucléaires est illustré à partir de la détermination et de l'analyse des données de séquences et de structures.

Pour survivre et se développer, une cellule doit percevoir les signaux de son environnement et y répondre de façon adéquate. Un très grand nombre de protéines, correspondant à plus de 20% des gènes chez l'homme, est impliqué dans ces mécanismes de communication et de transduction du signal depuis la réception du signal par des récepteurs, jusqu'à sa propagation et à son intégration dans la cellule : plus d'un cinquième des gènes de l'Homme est impliqué directement ou indirectement dans ces mécanismes de détection et d'intégration du signal [195]. On différenciera notamment deux types de récepteurs :

- **Les récepteurs membranaires (RM)** où la molécule signal ne pénètre pas dans la cellule mais induit des changements sur le récepteur qui va l'activer et lui permettre de médier un signal vers l'intérieur de la cellule via des facteurs de transcription tels que NF-kB.

- **Les récepteurs nucléaires (RN)** où la molécule signal pénètre dans la cellule et active ces récepteurs pour réguler directement la transcription des gènes en se positionnant sur les sites cis-régulateurs. Pour cette raison, les RN se comportent également comme des facteurs de transcription.

1. Famille des récepteurs nucléaires

Les RN sont présents chez tous les métazoaires et 48 (resp. 49) gènes ont été identifiés chez l'homme (resp. la souris) [196] (Figure 32). 24 d'entre eux, dont Nurr1 [197] ou NGFI-B [198] sont considérés comme étant des récepteurs orphelins, c'est-à-dire, des récepteurs pour lesquels aucun ligand n'a été identifié (Tableau 7). L'étude comparative des gènes et protéines de RN a montré qu'ils pouvaient être divisés en 6 régions numérotées de A à F à partir de l'extrémité N-terminale (Figure 31). Les domaines C et E sont les plus conservés [199] et correspondent respectivement au domaine de liaison à l'ADN (DBD) et au domaine de liaison au ligand (LBD). L'aspect modulaire en domaines des RN fut redémontré en 1987 par des protéines chimères constituées du LBD de ER (liant l'oestradiol) et du DBD de GR [200] (Tableau 7). Cette protéine chimérique pouvait effectivement lier l'oestradiol, mais ne pouvait pas activer les gènes régulés par ER. En revanche, la protéine chimérique comportant le DBD de GR, était capable d'activer les gènes normalement régulés par GR.

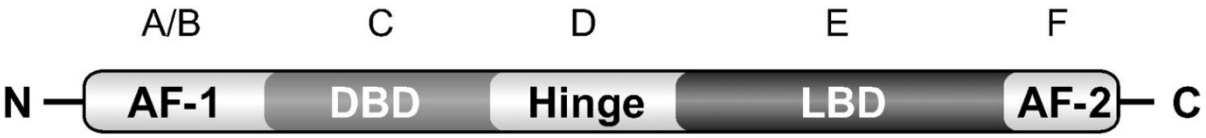


Figure 31. Organisation des 6 domaines des récepteurs nucléaires. *Le DBD et le LBD sont les deux domaines les plus conservés, et le hinge est une région adaptatrice instable dont la taille varie grandement d'un récepteur à un autre.*

L'activation ou la désactivation de facteurs de transcription peut se faire par de nombreux mécanismes moléculaires différents tels que la phosphorylation [201-204], la glycosylation, ou encore par des changements allostériques induits par la liaison d'un partenaire ou d'un ligand, comme c'est le cas des RN [205-209]. Les RN sont capables d'agir sur la régulation de plusieurs systèmes biologiques, dont l'embryogénèse, la différenciation ou encore l'apoptose [210].

Nom	Isoforme	Nomenclature	Ligand	Structure
TR Thyroid Hormone Receptor	α	NR1A1	Thyroid hormone (T3)	3ilz : A (*)
	β	NR1A2	Thyroid hormone (T3)	3gws : X
RAR Retinoic acid receptor	α	NR1B1	Retinoic acid	3kmr : A (*)
	β	NR1B2	Retinoic acid	1xap : A
	γ	NR1B3	Retinoic acid	4lbd : A
PPAR Peroxisome proliferator activated receptor	α	NR1C1	Fat acids	2npa : A
	δ	NR1C2	Fat acids	2znq : A
	γ	NR1C3	Fat acids, prostaglandine J2	2q61 : A
Rev-erb Reverse ErbA	α	NR1D1	-	3n00 : A
	β	NR1D2	-	2vc7 : A
ROR RAR-related orphan receptor	α	NR1F1	-	1s0x : A
	β	NR1F2	-	1nq7 : A (*)
	γ	NR1F3	-	3kyt : A
LXR Liver X receptor	α	NR1H3	Oxysterol	2acl :F
	β	NR1H2	Oxysterol	1pq9 : B
FXR Farnesoid receptor	α	NR1H4	Bile acid, feraxamine	3fxv : A
VDR Vitamin D receptor		NR1I1	Vitamin D	3a2i : A
PXR Pregnane X receptor		NR1I2	Xenobiotics	1m13 : A
CAR Constitutive androstan receptor		NR1I3	Xenobiotics, phenobarbital	1xnx : A (*)
HNF4 Human nuclear factor 4	α	NR2A1	-	1m7w : B
	β	NR2A3	-	
	γ	NR3A2	-	1lv2 : A
RXR Retinoid X receptor	α	NR2B1	9-cis retinoic acid	1g5y : C
	β	NR2B2	9-cis retinoic acid	1h9u : D
	γ	NR3B3	9-cis retinoic acid	2gl8 : C
TR2 Testis receptor		NR2C1	-	
TR4 Testis receptor		NR2C2	-	
TLL Tailess		NR2E2	-	
PNR Photoreceptor-specific nuclear receptor		NR2E3	-	
COUP-TF Chicken ovalbumin upstream promoter-transcription factor	I	NR2F1	-	
	II	NR2F2	-	3cjw : A
EAR2 ErbA2-related-gene-2		NR2F6	-	
ER Estrogen receptor	α	NR3A1	Oestradiol-17 β, tamoxifene, raloxifene	3dt3 : A
	β	NR3A2	Oestradiol-17 β	2giu: A
ERR	α	NR3B1	-	3d24: C

Estrogen receptor-related	β	NR3B2	-	
	γ	NR3B3	-	2zbs: A
GR Glucocorticoid receptor		NR3C1	Cortisol, dexamethasone, RU486	1m2z : D
MR Mineralocorticoid receptor		NR3C2	Progesterone, medroxyprogesterone acetate, RU486	2aa2 : R
PR Progesterone receptor		NR3C3	Testosteron, flutamide	3g8o : A
AR Androgen receptor		NR3C4	-	1t5z : A
NGFI-B NGF-induced factor B		NR4A1	-	2qw4: B
NURR1 Nur-related factor 1		NR4A2	-	1ovl : F
NOR1 Neuron-derived orphan receptor 1		NR4A3	-	
SF1 Steroidogenic factor 1		NR5A1	-	
LRH1 Liver receptor homologous protein 1		NR5A2	-	1zdu : A
GCNF Germ cell nuclear receptor		NR6A1	-	
DAX-1 DSS-AHC critical region on the chromosome, gene 1		NR0B1	-	
SHP Short heterodimeric receptor		NR0B2	-	

Tableau 7 : Les 48 gènes des récepteurs nucléaires humains. *A l'exception des structures indiquées avec un (*) qui viennent d'autres espèces, les PDB représentent les structures humaines du domaine de liaison au ligand (LBD) des récepteurs nucléaires. (adapté à partir de Brelivet et coll., 2008 [211]).*

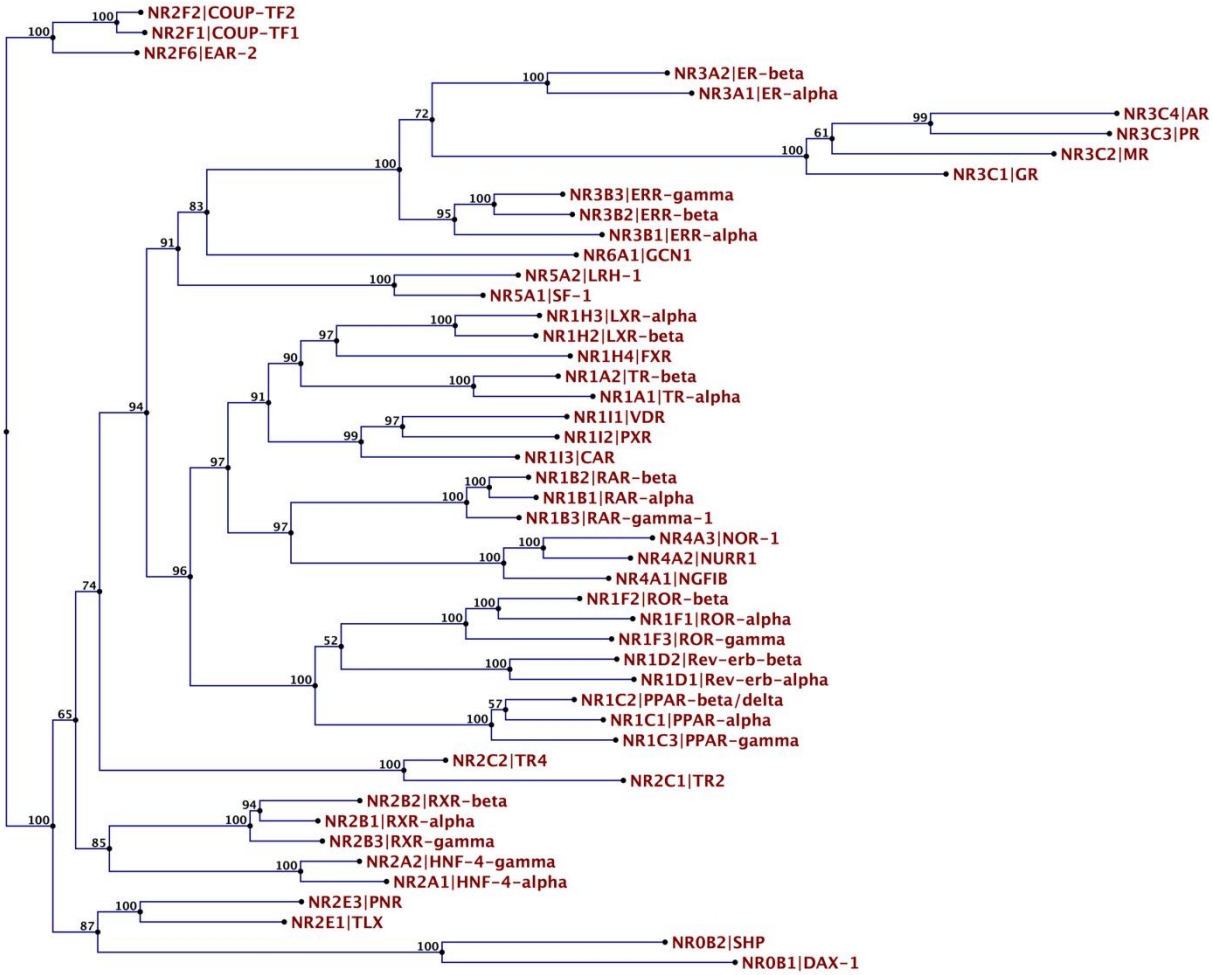


Figure 32. Phylogénie des 48 récepteurs nucléaires de l'Homme. *(source : http://upload.wikimedia.org/wikipedia/commons/7/7f/Nr_alignment_tree.jpg)*

2. Structures et oligomérisation des récepteurs nucléaires

Comme nous pouvons l'observer dans le Tableau 7, 36 des 48 récepteurs nucléaires humains ont à présent une structure cristallographique de leurs LBD. L'utilisation conjointe des structures de LBD et des séquences de ces RN a permis de les regrouper en deux classes, en fonction de leurs aptitudes à former des homodimères ou hétérodimères [212]. Ainsi, la classe I regroupe tous les RN capables de s'homodimériser, (tels que HNF, RXR, USP, ER, ERR…) et la classe II regroupe tous les RN capables de former des hétérodimères avec RXR ou USP (tels que TR, RAR, PPAR, LXR, VDR, PXR, …). La classe I est décrite à partir de la conservation de deux ponts salins reliant l'hélice du site de liaison au cofacteur au site de dimérisation (Figure 33). La classe II quant à elle est décrite à partir de la conservation d'un pont salin dans la boucle H8, H9. Dans ce cadre, RXR et USP ont un rôle particulier puisqu'ils sont capables de former à la fois des homodimères, mais aussi des hétérodimères.

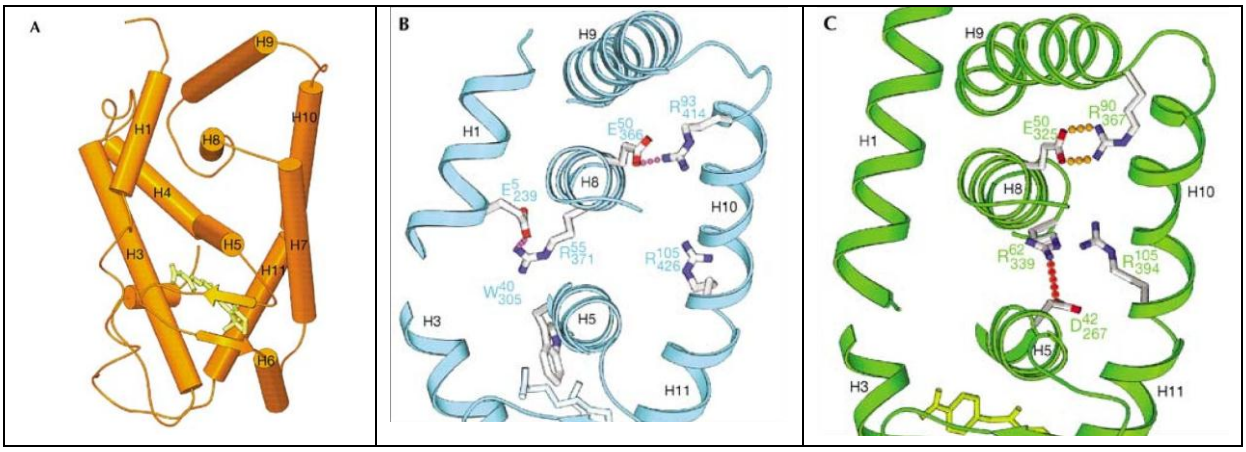


Figure 33. les classes I et II des récepteurs nucléaires définis à partir des ponts salins entre le site de dimérisation, le site de liaison au cofacteur et la boucle H8-H9. *A) Structure générale du LBD d'un récepteur nucléaire. B) classe I identifiée à partir de ponts salins entre H1 et H8 et H8 et H10. C) classe II identifiée à partir du pont salin entre H8 et H9. (source : Brelivet et al, 2004 [212]).*

Les sites de dimérisation et de liaison au cofacteur seront ré-examinés par l'apport du criblage de régions moléculaires dans le cinquième et dernier chapitre de cette thèse. En particulier, deux « arbres fonctionnels » ont été construits à partir de ces deux régions et renseignent sur la similarité fonctionnelle de ces régions (Chapitre « Criblage haut-débit des régions moléculaires »). Par ailleurs, une liste complète des 832 structures de récepteurs nucléaires est disponible en annexe 1 de cette thèse, et précise la présence de ligands et de peptides. Dans le chapitre « Criblage virtuel des régions moléculaires », le Tableau 26 précise également quelques statistiques sur ces structures, montrant notamment que seules 53 structures ont pu être déterminées sans ligand ni peptide, et que 92% des structures de récepteurs nucléaires ont été cristallisées avec un ligand.

3. Structures moléculaires et Génétique

Plus généralement, la détermination des structures de molécules a eu un impact important dans tous les domaines de la biologie, y compris la génétique, en révélant les mécanismes chimiques derrière les fonctions biologiques, en explicitant l'impact structural d'une mutation, et en contribuant à détailler la spécificité de la reconnaissance des partenaires (Figure 34). En particulier, il existe un grand nombre de protéines pour lesquelles des structures de la forme sauvage et des formes mutées sont disponibles. La conclusion de ces expériences est que la plupart du temps, une mutation n'a pas d'effet et n'entraîne que de petits remaniements des chaînes latérales avoisinantes et du squelette. D'autres

mutations ont des effets beaucoup plus délétères et près de 73 000 d'entre elles réparties sur 2800 gènes associés à des maladies monogéniques sont recensées notamment dans la base de données HGMD [213, 214].

L'étude structurale des mécanismes de l'épigénétique (changements héréditaires de phénotypes ou d'expression de gènes sans altération des séquences de l'ADN) a également été rendue possible par la détermination des octamères d'histone, et la microscopie électronique a permis l'observation directe d'une succession de nucléosomes, chacun séparé d'environ 200 paires de bases [215].

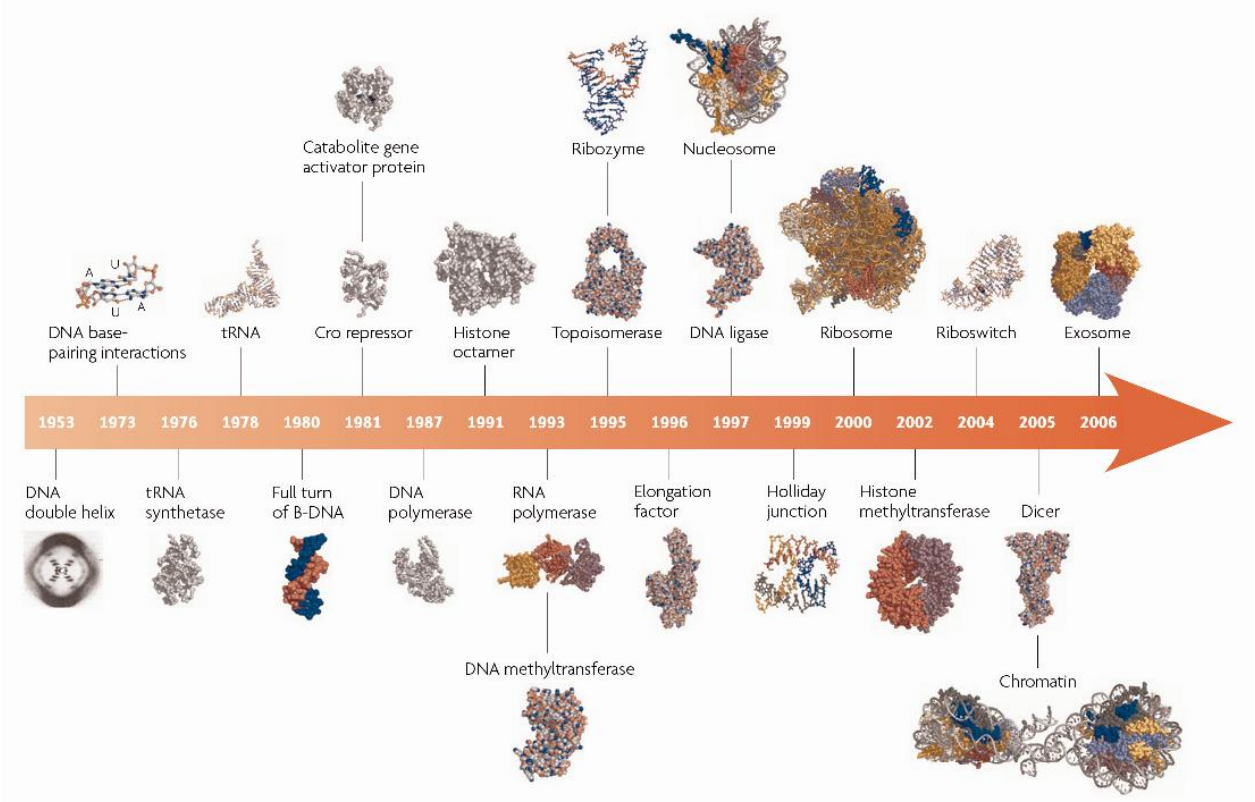


Figure 34. Découverte de structures impliquées dans la régulation de l'expression des gènes et des produits des gènes. *(source : Laskowski et coll., 2008 [216])*

Structures, surfaces et interactions moléculaires

I. La modélisation des structures moléculaires

Dans cette première partie, nous aborderons les problèmes rencontrés lors de la modélisation des structures moléculaires. Le choix du rayon des atomes, délimitant des sphères d'influences relatives à la propriété de l'atome est particulièrement important pour étudier à la fois la surface moléculaire mais aussi les interfaces d'assemblages moléculaires. Nous introduisons ensuite le modèle géométrique des tesselations de Voronoï et certaines de ses applications aux structures biologiques, notamment pour évaluer le volume des atomes et pour étudier les contacts inter-atomiques de façon rapide et non ambigüe. Nous présentons ensuite les complexes et formes alphas, des dérivées des tesselations de Voronoï, que nous avons utilisés dans l'article de 2008 pour décrire les surfaces moléculaires, former des régions contiguës, et calculer un score de courbure non biaisé par la présence de cavités sous la surface. Avec d'autres auteurs, nous proposons ces représentations géométriques simples et robustes pour l'analyse systématique et rapide des surfaces et des interactions moléculaires. Ces modèles présentent l'avantage de ne pas altérer la position des atomes de la structure, de décrire les contacts inter-atomiques de façon non ambigüe, et d'être plus simples d'un facteur 10 que les représentations usuelles de la surface de Connolly et de Lee de Richards.

Une molécule, indépendamment de sa fonction peut être considérée comme un simple assemblage d'atomes discrétisés dans des sphères de tailles et de charges variées. Le plus généralement, les molécules présentes dans notre organisme ne se composent que de Carbone (C), d'Azote (N), d'Oxygène (O), d'Hydrogène (H), de Soufre (S) et de Phosphore (P). Toutefois, des molécules telles que l'hémoglobine ou la calmoduline ont besoin, pour fonctionner correctement, de lier des ions tels que le fer (Fe) et le calcium (Ca2+).

La modélisation de structures moléculaires à partir des coordonnées des atomes d'une molécule nécessite donc la connaissance de leurs rayons ou « sphères d'influences ». Chaque type d'atome a ainsi plusieurs rayons décrivant chacun différents aspects de ses potentialités. Le choix de ces rayons pour modéliser l'atome et la molécule a donc une importance considérable, cependant peu de programmes prennent encore pleinement compte de ces particularités.

1. L'atome, ses rayons, ses volumes

Afin de modéliser la surface ou le volume d'une molécule, il est nécessaire de définir les rayons (ou sphères d'influences) applicables à chaque atome la composant. De tels rayons indiquant la taille approximative des atomes, ont d'abord été théorisés par le Prix Nobel de physique Johannes D. Van der Waals en 1910, puis qualitativement déterminés par Linus Pauling dès 1939 dans *The nature of the chemical bond* [217]. Par définition, le rayon de Van der Waals est égal à la moitié de la distance minimale à laquelle peuvent s'approcher deux atomes d'un même élément sans se lier [218] (Figure 35).

Les rayons ioniques correspondent quant à eux à la distance minimale à laquelle peuvent s'approcher deux atomes chargés. En particulier, les anions qui disposent d'un ou plusieurs électrons supplémentaires ont des rayons ioniques supérieurs à leurs rayons de Van der Waals (forme neutre de l'atome). A l'inverse, les cations qui ont moins d'électrons que la forme neutre de l'atome ont des rayons ioniques inférieurs à ceux de Van der Waals.

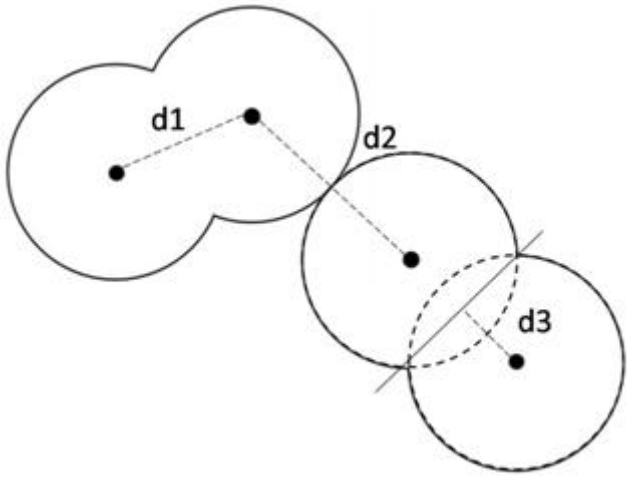


Figure 35. Les différents rayons décrivant les atomes. *d2 est la distance minimale de contact entre les noyaux de deux atomes du même élément lorsque ceux-ci ne sont pas liés de façon covalentes ; d2/2 correspond au rayon de Van der Waals. d1 est la distance de contact entre les noyaux de deux atomes d'un même élément lorsque ceux-ci sont liés de façon covalente ; d1/2 ou encore d3 sont donc les rayons de covalence.*

Le rayon de chaque atome, déterminé à partir de la proximité de ses voisins, n'est donc pas le même dans tous les environnements. Il a donc été nécessaire de définir plusieurs rayons et volumes pour un même atome en fonction de cet environnement [219, 220]. En particulier, les analyses réalisées sur un grand nombre de structures ont permis d'identifier les différentes variations de ces rayons et volumes et de proposer des échelles comportant les rayons et volumes des atomes C, N, O, S dans plus de 24 environnements [219-221]

(Tableau 8) pouvant dépendre entre autre de la structure secondaire à laquelle appartient l'acide aminé [222]. De plus, la position des atomes d'hydrogène n'étant que rarement déterminée, on considère le plus souvent des groupes chimiques dont les rayons des atomes centraux (C,N,O,S) sont augmentés empiriquement pour tenir compte de ce biais dans les données structurales. En se basant sur les données de la Cambridge Structural Database (CSD), J. Tsai et coll. déterminent les rayons de 13 de ces groupes chimiques [219].

Groupe atomique	Formule chimique	Valence	Rayons (a)	Rayons (b)	Rayons (c)
N3H0	$>N^-$	3	1,64	1,5	1,7
N3H1	>NH	3	1,64	1,65	1,7
N3H2	$-NH_2$	3	1,64	1,65	1,6
N4H3	$-NH_3^+$	4	1,64	1,5	2
O1H0	=O ou $-O^-$	1	1,42	1,4	1,4
O2H1	-OH	2	1,46	1,4	1,6
C3H0	-C<	3	1,61	1,76	1,7
C3H1	-CH-	3	1,76	1,76	1,7
C4H1	-CH<	4	1,88	1,87	2
C4H2	-CH2-	4	1,88	1,87	2
C4H3	-CH3	4	1,88	1,87	2
S2H0	-S-	2	1,77	1,85	1,8
S2H1	-SH	2	1,77	1,85	-

Tableau 8. Différentes évaluations des rayons de Van der Waals pour différents groupes atomiques. *(adapté à partir de J. Tsai et coll. [219])*

Notons que lorsque le rayon de Van der Waals est utilisé pour décrire des liaisons polaires favorables entre deux atomes, les rayons de Van der Waals de ces deux atomes se recouvrent. En effet, les interactions polaires favorables forment un empilement plus serré. A.J. Li et R. Nussinov ont ainsi introduit en 1998 le concept de rayons de Coulomb pour des atomes participant à des interactions polaires favorables [223]. Cette distinction est importante pour la modélisation des surfaces moléculaires et également, pour le positionnement de molécules de solvant autour de la molécule. En effet, si le solvant est polaire, cas de l'eau, et que l'on considère un atome polaire de surface, alors le centre de la molécule d'eau doit être à une distance de Coulomb et non de Van der Waals de cet atome (Tableau 9). Inversement, si le solvant est apolaire, ou si l'atome de surface est apolaire, alors le centre de la molécule d'eau doit être à une distance de Van der Waals de cet atome [223].

Symbole	$R_{Coulomb}$	R_{VdW}	$R_{Coulomb} - R_{VdW}$
N	1.49	1.71	-0.22
NH	1.55	1.66	-0.11
NH+	1.41	1.65	-0.24
NH2	1.49	1.62	-0.13
NH2+	1.49	1.67	-0.18
NH3+	1.4	1.67	-0.27
O	1.41	1.49	-0.08
O_{co}	1.41	1.52	-0.11
O_{coo}	1.35	1.49	-0.14
OH	1.35	1.54	-0.19
H_2O	1.37	1.68	-0.31

Tableau 9. Différences entre les rayons de Coulomb et les rayons de Van der Waals pour les atomes polaires. *(adapté à partir de J. Tsai et coll. [219])*

Il est important de comprendre que la bonne assignation de rayons aux atomes permet non seulement une meilleure modélisation des structures moléculaires, et donc une meilleure détermination des contacts inter-atomiques, mais permet également d'avoir des points de référence pour évaluer par exemple le bon ou mauvais repliement de protéines, et pour comparer la densité des régions. En utilisant ces données, il a par exemple été possible de montrer que la densité atomique à l'intérieur des protéines est relativement homogène, à l'exception des cavités. Toutefois ces cavités sont le plus souvent hydratées, ce qui contribue à homogénéiser la densité en ces régions particulières des molécules [219].

J. Liang et K. Dill suggèrent une certaine plasticité des molécules, précisant que le cœur protéique n'est peut-être pas aussi dense que dans certains solides, leur conférant une certaine tolérance aux mutations [224]. Parallèlement H. Nakamura envisage cette plasticité des molécules comme une innovation de l'Évolution pour leur conférer une dynamique conformationnelle essentielle pour l'allostérie et la régulation de nombreuses interactions, comme on l'observe pour les kinases [225].

2. Surface moléculaire et surface accessible

Alors que la surface moléculaire de Connolly est définie comme l'enveloppe dessinée par la surface de contact entre la molécule étudiée et une sonde parcourant sa surface [2], la surface accessible de Lee et Richards correspond quant à elle à l'enveloppe dessinée par le centre de la sonde en parcourant sa surface [3].

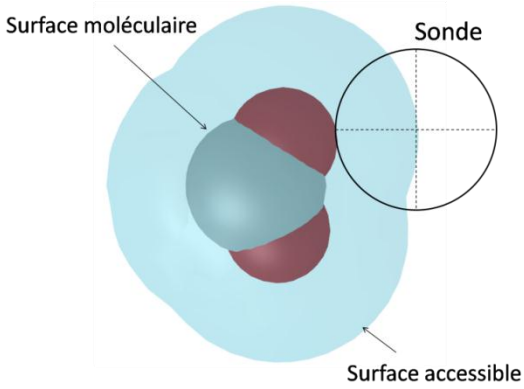


Figure 36. Surface moléculaire et surface accessible.

La modélisation de la surface accessible de Lee et Richards et la modélisation de la surface moléculaire de Connolly dépendent donc toutes deux de la position des molécules de solvant autour de la molécule étudiée (Figure 37). Par conséquent, modéliser le rayon des atomes, soit par leurs rayons de Coulomb soit par leurs rayons de Van der Waals en fonction de l'environnement et de la nature polaire ou apolaire de la sonde employée, permet de définir des surfaces moléculaires plus précises. En effet, la modélisation de ces surfaces dépend également de la forme, de la taille et de la nature de la sonde [226] : la surface modélisée est la surface perçue par une autre molécule ; elle peut être très différente en fonction de la sonde qui la dessine.

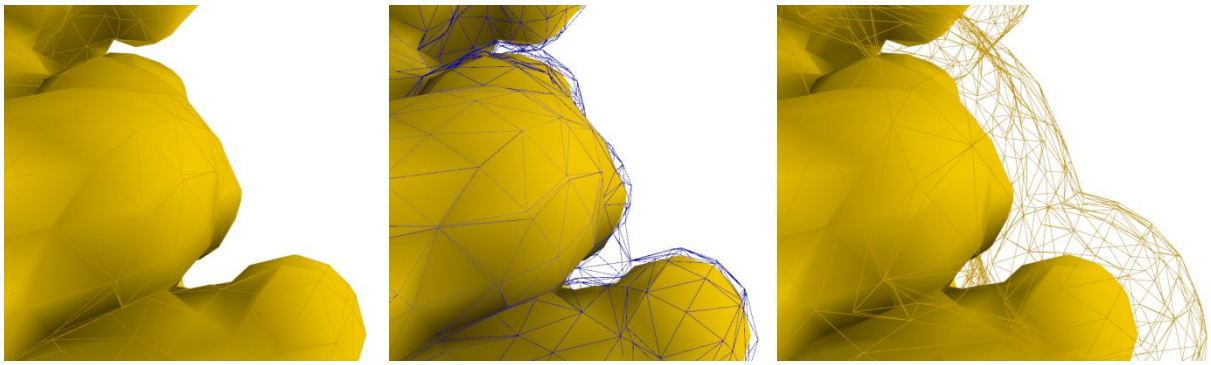


Figure 37. Surface moléculaire de Connolly en fonction de la polarité du solvant. *Dans les trois figures A, B, C, la surface moléculaire de Connolly (jaune) d'une région locale de RXRα (1dkf :A) a été calculée pour un solvant polaire et une sonde de taille 1.4 Á correspondant à une molécule d'eau. B) on représente par un maillage la même surface moléculaire de Connolly calculée pour un solvant apolaire et une même taille de sonde. C) la surface accessible au solvant de Lee et Richards pour un solvant polaire et une sonde de taille 1.4 Á est indiquée par le maillage.*

De plus, la surface d'une même molécule peut être modélisée à différentes résolutions, et rendre ainsi compte d'un nombre plus ou moins important de détails, notamment aux interfaces moléculaires [227]. Plus la résolution est élevée et plus on observera de détails ; pour des résolutions inférieures à la taille de la sonde, il est notamment possible d'observer des micro-crevasses qui ne lui sont pas accessibles (Figure 38). Inversement, plus la résolution est faible et plus la surface est simplifiée, cachant ou approximant des motifs locaux importants.

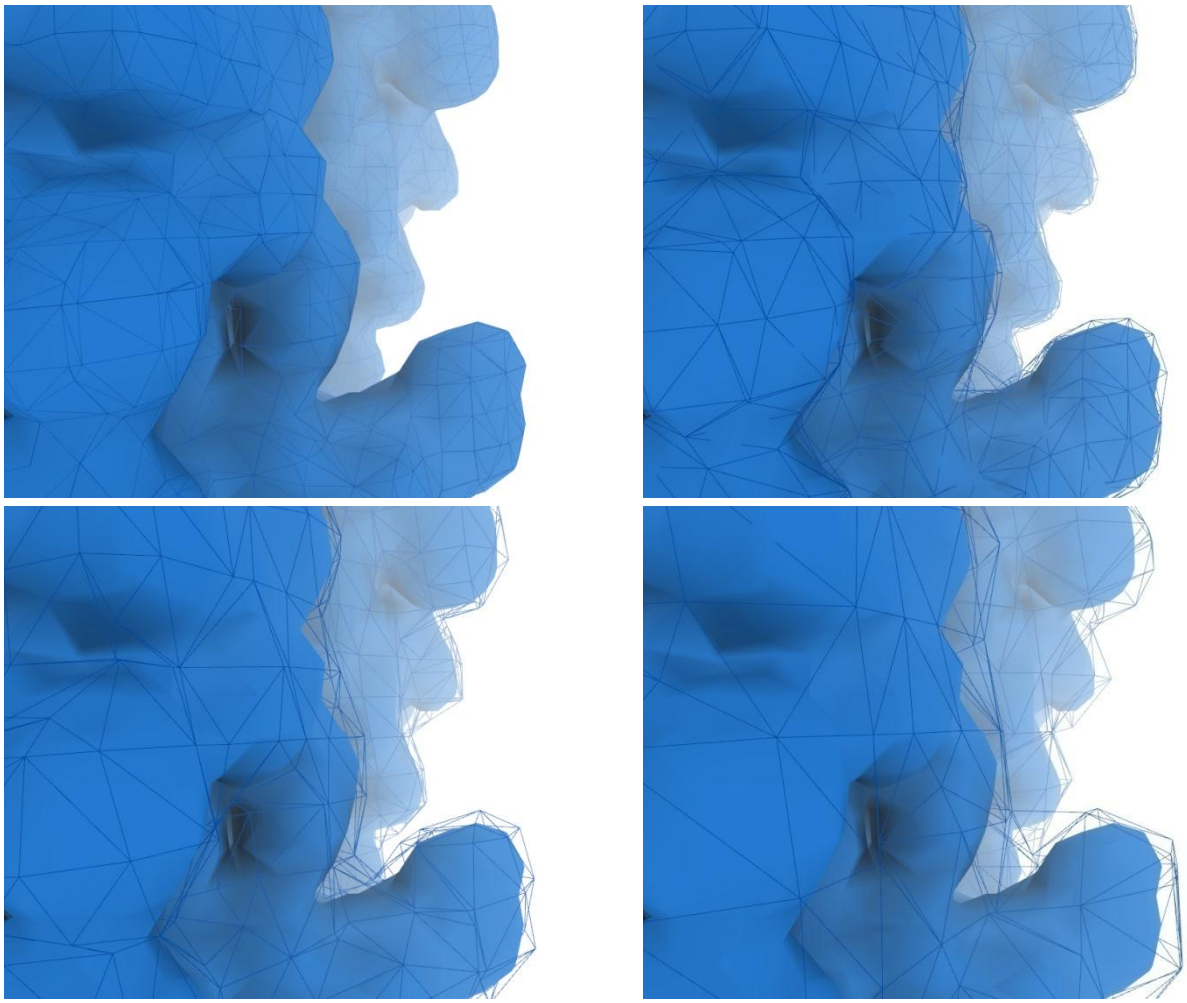


Figure 38. Surface moléculaire à différentes résolutions. *En haut à gauche, la surface moléculaire pour une résolution de grille de 1.2 Å. Puis de haut en bas à droite, pour des résolutions de grille respective de 1.6 Å, 2.2 Å et 3.2 Å.*

Le calcul de ces surfaces a permis de quantifier l'aire de surface accessible au solvant (ASA), de calculer des aires accessibles de référence pour les résidus, et de distinguer les résidus de surface des résidus du cœur protéique [228, 229]. La distinction de ces régions essentielles des protéines s'est vue vérifiée par l'analyse de leurs compositions en acides

aminés, ce qui en retour a permis de créer des programmes de prédiction des résidus de surface et du cœur protéique à partir des seules séquences (Tableau 10). Plus récemment, une analyse plus poussée tente de prédire jusqu'à l'accessibilité relative des résidus à partir des seules données de séquence [230, 231]. De telles approches peuvent aider à modéliser la structure tertiaire des protéines, et dans certains cas, à localiser des sites de liaisons [119].

Résidu	Composition de la surface	Composition du cœur protéique
Ala	7,9	11
Arg	4	0,4
Asn	6,3	2
Asp	7,4	2,2
Cys	1,8	5,4
Gln	4,5	1,3
Glu	5,2	1
Gly	8,8	9,7
His	2,2	2,4
Ile	3	10,5
Leu	4,3	12,8
Lys	8,9	0,3
Met	0,9	3
Phe	2,5	7,7
Pro	4,7	2,2
Ser	8,9	5
Thr	7,1	4,6
Trp	1,3	2,7
Tyr	4,8	3,3
Val	4,6	12,7
Hydrophobe	34,2	61,9
Aromatique	10,8	16,1
Polaire	26,8	12,9
Chargé	25,5	3,9

Tableau 10. Composition en acides aminés de la surface et du cœur protéique. *(adapté à partir de Janin et coll., 1988 [228]).*

Dans notre article de 2008, nous précisons que les résidus des surfaces protéiques sont en moyenne 40% moins hydrophobes, 65% plus flexibles et 75% moins conservés que les résidus appartenant aux cœurs protéiques [4]. La distinction des résidus de surface avec les

résidus du cœur protéique, ainsi que l'analyse des aires accessibles, nous ont également permis de démontrer que les résidus interagissants sont en moyenne 15% plus exposés que les résidus non-interagissants (connus pour n'interagir qu'avec le solvant ou des ions), que ceux-ci se situent dans des crevasses ou dans des bosses.

Comme nous le verrons par la suite, la modélisation des surfaces moléculaires a eu également un impact décisif pour l'étude des sites de liaisons et des interfaces moléculaires, notamment en permettant le calcul des aires d'interface, et en déterminant la contribution de chaque résidu à cette aire d'interface.

3. La tesselation de Voronoï

Les surfaces de Connolly [2] et de Richards [3] sont toutefois relativement complexes et leurs analyses et comparaisons nécessitent des calculs relativement intensifs, ce qui les rend difficilement applicables aux traitements à hauts débits des surfaces et des interactions moléculaires. Pour appréhender cette complexité, un simple calcul permet d'observer que la surface de Connolly du récepteur nucléaire RXRα est composée de près de 20 000 facettes, alors qu'une surface moléculaire simplifiée, telle que la forme alpha, n'est composée que de 1700 facettes. Le développement et l'utilisation de représentations géométriques simplifiées des molécules telles que la forme alpha, permet donc de diminuer leur complexité d'un facteur 10 et accélère considérablement leurs comparaisons et analyses. La description et l'utilisation des formes alphas pour l'étude des surfaces et interactions moléculaires seront davantage détaillées par la suite, notamment dans notre article de 2008 [4] et dans le chapitre « criblage haut débit des régions moléculaires ».

De plus, les surfaces de Connolly et de Richards ne fournissent aucune description des parties internes de la molécule. Elles ne permettent donc pas d'évaluer le volume des atomes et des résidus, ni d'étudier des régions internes de la molécule.

La tesselation de Voronoï divise un espace en autant de régions qu'il y a de points. Chacune de ces régions, encore appelée « cellule de Voronoï », est un polyèdre dont chaque côté est partagé par très exactement deux points de l'espace (Figure 39). Dans le cas des molécules, chaque cellule de Voronoï définit la zone d'influence d'un atome en fonction du milieu et de ses voisins atomiques. Ainsi, à l'intérieur d'une cellule de Voronoï $V(p_i)$ formée autour d'un atome p_i, tout point p_j est plus proche de l'atome p_i que de tout autre atome de l'espace R^d:

$$V(p_i) = \{\, x \in R^d : \|x - p_i\| \leq \|x - p_j\|, \forall\, j \leq n \,\}$$

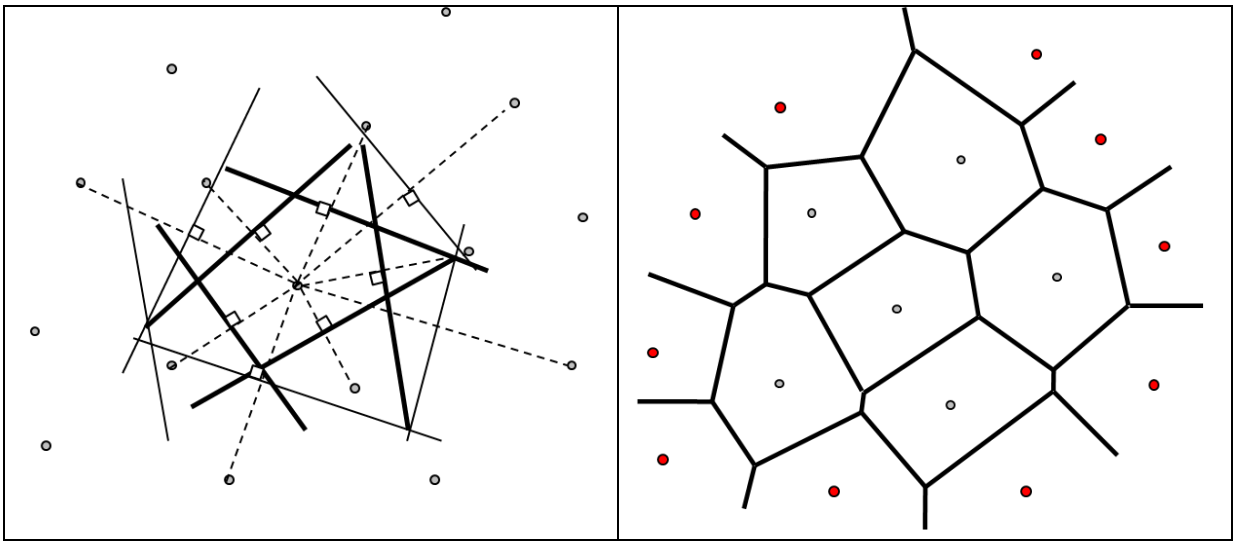


Figure 39. Construction de la tesselation de Voronoï d'un espace à deux dimensions. *A gauche, pour un point donné, il est possible de tracer l'ensemble des bissectrices (tirets) entre chaque couple de points. L'ensemble des bissectrices les plus proches du point donné forme alors un polyèdre aussi appelé cellule de Voronoï. A droite, en itérant sur chacun des points de l'espace, il est possible d'obtenir la tesselation de Voronoï de cet espace. Notons que les points rouges ont des cellules de Voronoï de très grande taille, voire infinie ; cette propriété permet donc de déterminer la liste des atomes de surface.*

Bien que la tesselation de Voronoï présente l'avantage de diviser l'espace en régions d'influences des atomes et fournisse une notion non ambigüe du voisinage de chaque atome, cette représentation présente deux inconvénients limitant pour l'étude des molécules : (1) les cellules de Voronoï sont définies grâce aux voisinages atomiques, or les atomes de surface ont moins de voisins et sont modélisés par des cellules démesurément grandes ; (2) les cellules de Voronoï sont formées sans tenir compte de la diversité physico-chimique des atomes, diversité qui se reflète notamment par des rayons d'influences différents.

Pour pallier au premier problème, Richards proposa dès 1974 d'ajouter des molécules d'eau autour d'une protéine étudiée, puis de relaxer leurs positions [220]. Cette méthode a ensuite été améliorée à la fin des années 90 par Tsai, Gerstein et Levitt [232, 233]. D'autres approches tentent de résoudre ce premier problème en utilisant une représentation d'unions de sphères [234]. Dans un espace en trois dimensions, la tesselation de Voronoï permet dès lors de calculer le volume des atomes et des résidus dans les structures moléculaires, que ceux-ci soient enfouis, en surface, ou à l'interface avec d'autres molécules [220, 235].

Une autre construction, le diagramme de puissance (ou diagramme de Laguerre) permet de répondre simultanément aux deux problèmes en pondérant et en limitant la taille des cellules de Voronoï par un facteur de « puissance » représentant l'influence d'un point dans l'espace. La puissance d'un point x et pour un atome p de centre c et de rayon r est calculée comme suit :

$$power_p(x) = (x - c)(x - c) - r^2$$

Il devient alors également possible de générer un diagramme pondéré de Voronoï en définissant les limites des cellules de Voronoï par leurs puissances :

$$V_{pondéré}(p_i) = \{ x \in R^d : power_{pi}(x) \leq power_{pj}(x), \forall j \leq n \}$$

Le diagramme de Laguerre permet donc de fournir une tesselation de l'espace prenant en compte les rayons respectifs des atomes. En 1992, Edelsbrunner et Shah ont développé un algorithme incrémental permettant la construction rapide de ces diagrammes de puissance, plus tard repris pour l'analyse des cavités moléculaires [236-239].

4. Le complexe de Delaunay

Le complexe de Delaunay est une forme duale de la tesselation de Voronoï en ce sens qu'il est possible de passer directement de l'une à l'autre des représentations. En effet, le complexe de Delaunay peut notamment être obtenu en reliant les couples de points qui partagent une facette dans la tesselation de Voronoï. Inversement, il est possible de redéfinir les cellules de Voronoï en traçant les lignes bissectrices des arêtes du complexe de Delaunay.

Tout comme la tesselation de Voronoï, le complexe de Delaunay présente l'avantage de définir une notion de voisinage non ambigüe : pour chaque atome, on définit l'ensemble de ses atomes voisins les plus proches dans chacune des directions de l'espace (Figure 40). En particulier, et comme nous le verrons dans l'article sur la modélisation, les structures du complexe de Delaunay et de la forme alpha (présentée ci-après) se prêtent particulièrement bien à une transposition dans un Graphe où chaque nœud représente par exemple un atome et où chaque arête représente par exemple une interaction entre voisins atomiques. Cette transposition dans la structure de Graphe donne accès à de nombreux algorithmes performants développés depuis les années 1950, pour l'analyse et la comparaison des structures moléculaires.

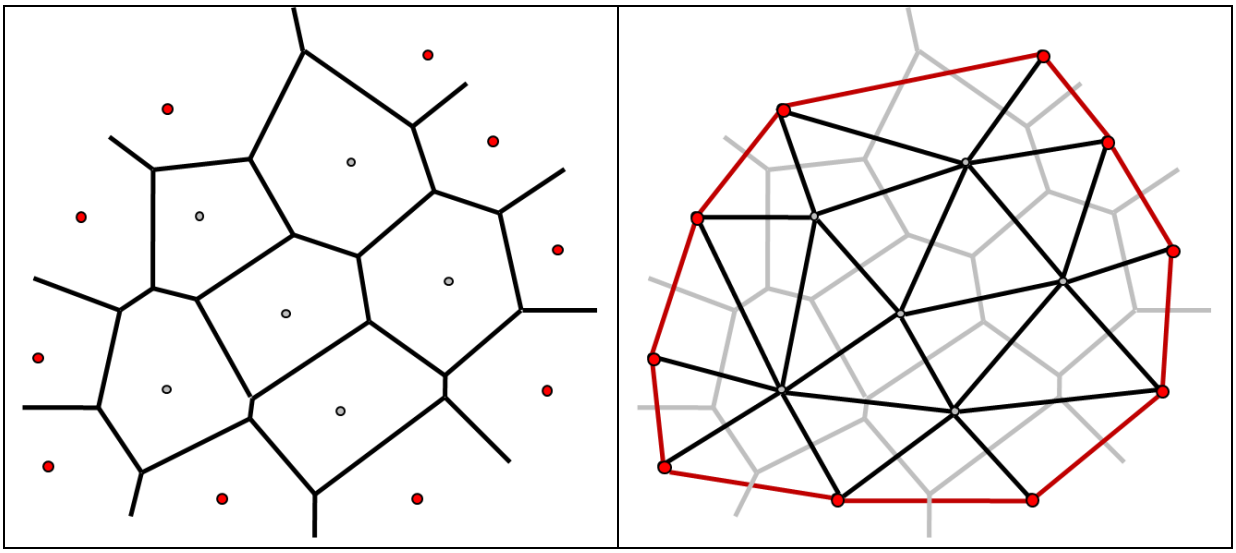


Figure 40. Tesselation de Voronoï et complexe de Delaunay d'un espace à deux dimensions. *A gauche, la tesselation de Voronoï définissant un ensemble de cellules divisant l'intégralité de l'espace. A droite, le complexe de Delaunay défini à partir de la tesselation de Voronoï en reliant les couples de points qui partagent une facette commune. La notion d'atome de surface est conservée et il est même possible de définir une surface contour en rouge.*

5. Le complexe alpha et la forme alpha

Le complexe alpha est directement obtenu à partir du complexe de Delaunay en retirant les arêtes ayant une distance supérieure à un seuil [239]. Alors que le complexe de Delaunay définit les voisins atomiques les plus proches d'un atome (quelle que soit la distance les séparant), le complexe alpha ne définit que les voisins atomiques suffisamment proches (en fonction du seuil), représentant par exemple les interactions entre atomes (Figure 41).

Les complexes et formes alphas, ainsi que les complexes de Delaunay, ont été utilisés depuis les années 1995 par Liang et Edelsbrunner pour localiser et décrire les crevasses et cavités des molécules [237, 238].

Les surfaces définies par les points et arêtes du contour du complexe de Delaunay ou du complexe alpha fournissent une représentation tridimensionnelle très simplifiée des molécules. En particulier, la forme alpha (contour du complexe alpha) corrèle précisément avec la surface moléculaire de Connolly (Figure 42). Ainsi que nous le verrons dans le chapitre « Modélisation des structures moléculaires », les formes alphas peuvent être utilisées pour définir efficacement les atomes et résidus de surface d'une molécule, ainsi que pour définir des régions contiguës et déterminer des courbures non biaisées par la présence de cavités sous la surface [4].

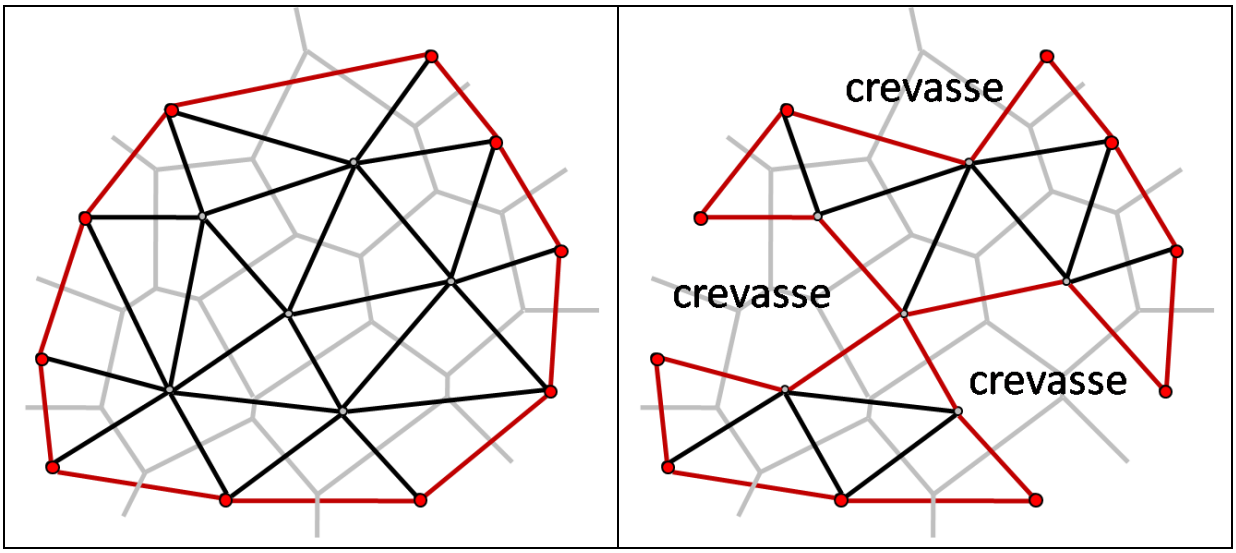


***Figure 41. Du complexe de Delaunay au complexe alpha et à la forme alpha**. En retirant les arêtes du complexe de Delaunay (à gauche) qui sont plus longues qu'une distance seuil, on obtient le complexe alpha (à droite) où chaque arête peut notamment représenter une interaction entre atomes. On note en particulier que les deux représentations n'ont pas le même contour (en rouge), et que la soustraction du complexe de Delaunay au complexe alpha permet de retrouver les atomes qui appartiennent soit à une crevasse (en surface), soit à une cavité (en interne). La forme alpha est obtenue en ne conservant que les atomes et arêtes du contour du complexe alpha.*

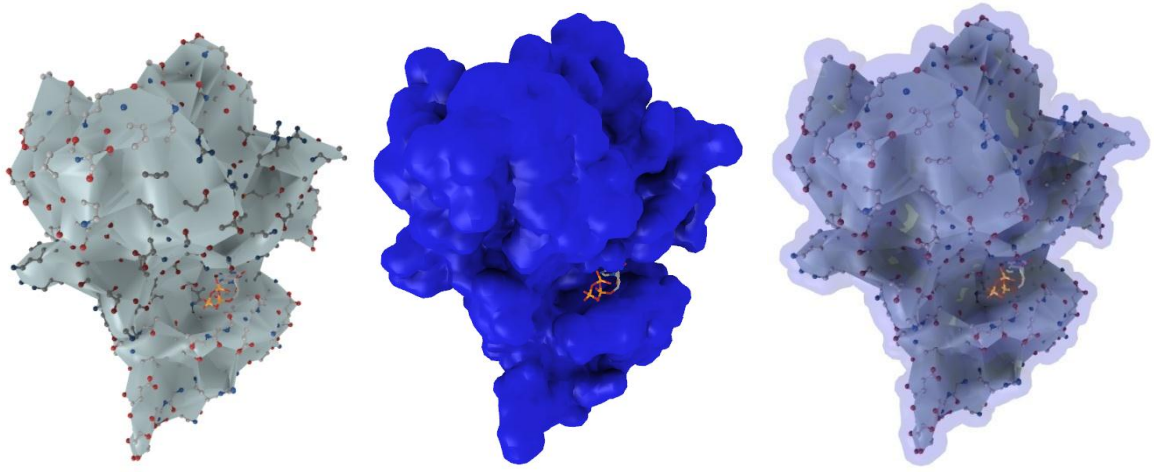


***Figure 42. Correspondances entre la surface moléculaire de Connolly et la forme alpha.** A gauche, la kinase cycline dépendante 2 est représentée par sa forme alpha en gris ainsi que par le modèle boules et bâtons pour visualiser la décomposition de la surface fournie par la forme alpha (1fin :A). Au milieu, la représentation conventionnelle de la surface moléculaire de la kinase. A droite, la superposition de ces deux surfaces montrant leur ressemblance.*

La librairie CGAL [240] donne un accès relativement simple à des algorithmes robustes et performants permettant de générer ces différentes représentations simplifiées d'objets tridimensionnels et de structures moléculaires.

6. Distinction entre la surface et le cœur des protéines

La distinction entre les atomes et résidus de surface et les atomes et résidus du cœur protéique revêt une importance particulière notamment pour l'étude des interactions moléculaires. Le but est alors de définir les atomes et résidus de la protéine susceptibles d'interagir avec d'autres partenaires. L'étude des atomes et résidus de surface permet également d'étudier différents phénomènes comme la solvatation, ou de différencier la flexibilité des chaînes latérales des résidus du cœur et de la surface.

Une fois la surface de la molécule modélisée, il devient possible de distinguer deux catégories d'atomes : les atomes de surface, et les atomes enfouis ou atomes du cœur protéique. Pour ce faire, on calcule l'accessibilité des atomes à partir de la surface accessible de Lee et Richards en identifiant les facettes de surface appartenant à chaque atome et en calculant leur aire. Le programme Naccess a été utilisé dans plusieurs études sur la surface et les interactions des molécules [1, 155, 156] et sert à présent d'outil de référence pour calculer ces aires accessibles [241]. Un atome est généralement considéré comme étant de surface s'il a au moins 1 Å² d'aire accessible ; les autres atomes étant considérés comme enfouis. D'autres définitions sont toutefois possibles ; notamment les atomes de surface correspondent également à l'enveloppe formée par la forme alpha, ainsi qu'il en est fait mention dans notre article de 2008 [4].

Il peut s'avérer également pratique de différencier les résidus de surface des résidus du cœur protéique. Pour ce faire, deux principales méthodes sont envisageables : la première, basée sur l'accessibilité ; la seconde basée sur la profondeur à laquelle se trouve enfoui le résidu dans la structure :

- Pour la première méthode, il a été nécessaire de calculer les accessibilités maximales des résidus, par exemple à partir des conformations de molécules Gly – X – Gly. Dès lors, on considérera qu'un résidu est de surface s'il a au moins 5% de son accessibilité maximale [228].

- Pour la seconde méthode, il est nécessaire de définir une notion de profondeur qui, si elle est relativement intuitive pour des molécules globulaires, s'avère toutefois complexe pour un certain nombre de molécules aux formes variées. Les méthodes couramment utilisées consistent généralement à calculer la distance de l'un des atomes du résidu, ou de

son centroïde (centre de gravité) à l'atome ou résidu de surface le plus proche. Une alternative consiste également à positionner des molécules d'eau autour de la surface de la molécule, en effectuant éventuellement une optimisation par dynamique ou mécanique moléculaire, et de calculer cette distance par rapport à la molécule d'eau la plus proche [242].

En perspectives du chapitre « Modélisation des structures moléculaires », nous présenterons une méthode permettant de calculer rapidement la profondeur d'atomes et de résidus à partir du complexe alpha.

II. Les méthodes d'analyses des interactions moléculaires

Dans cette seconde partie, nous décrivons les différentes méthodes d'analyses des interactions moléculaires, et portons l'accent sur les analyses nouvelles des interfaces par les tesselations de Voronoï et dérivées. L'analyse des interactions moléculaires nécessite la détermination des atomes et résidus interagissants, ainsi que de quantifier leurs contributions à l'interaction.

1. Détection des résidus interagissants par distance de contact

Etant donné que le rayon moyen d'un atome est de 1,5 Á, et que l'analyse moléculaire porte principalement sur des structures de résolution cristallographique de l'ordre de 2 à 3 Á, il est généralement d'usage de considérer que deux atomes interagissent s'ils sont à moins de 4,5 Á (deux fois le rayon moyen, auquel on ajoute un facteur d'incertitude sur la localisation dépendant de la résolution et de la dynamique de la structure).

En variante, il est possible d'utiliser les rayons de Van der Waals et de Coulomb de deux atomes plutôt qu'un rayon moyen pour évaluer une interaction entre ces deux atomes [1].

La détection d'une interaction entre deux résidus peut se faire sur un critère de distance qui va dépendre soit de la distance entre les atomes des deux résidus, soit de la distance entre les centroïdes (centres de gravité) de ces résidus. En pratique, voici les définitions généralement utilisées dans la littérature :

- sur la base des contacts inter-atomiques : pour interagir, il est nécessaire qu'au moins deux atomes de deux résidus soient en contact (la distance de contact atome-atome étant alors de l'ordre de 4 à 4.5 Á). Une variante plus stringente consiste soit à diminuer la distance seuil identifiant un contact inter-atomique, soit à augmenter le nombre nécessaire de contacts atomiques inter-résidus pour définir une interaction résidu-résidu.

- sur la base des centroïdes : les centroïdes des résidus ou de leurs chaînes latérales doivent être à moins de 6 à 8 Å. Cette notion de contact est plus floue que la précédente, mais elle peut être utilisée lorsque des atomes sont manquants dans les structures ou lorsque leurs positions sont incertaines.

Dans le cas particulier de l'étude des interactions entre molécules d'eau et résidus d'une protéine ou d'un acide nucléique, des critères plus astringents sont utilisés pour filtrer les contacts non-spécifiques : les atomes polaires d'un résidu (O, N) doivent être à moins de 3.5 Å de la molécule d'eau [180].

2. Détection des résidus interagissants par analyse des surfaces accessibles

Une autre méthode consiste à calculer l'accessibilité des résidus pour chacune des chaînes protéiques considérées séparément (forme libre), et de recalculer l'accessibilité de ces mêmes résidus lorsque les chaînes protéiques interagissent (forme liée) :

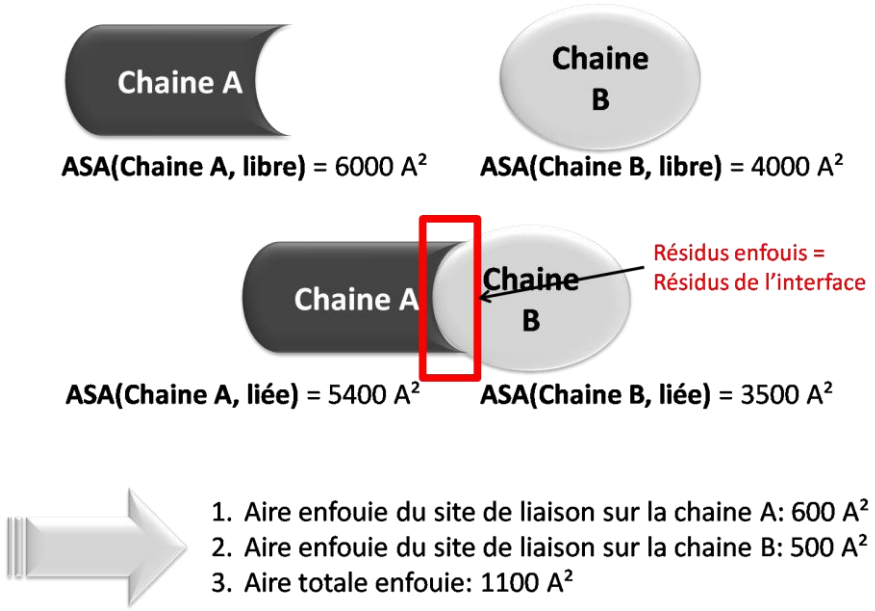


Figure 43. Détection des résidus interagissants et de l'aire d'interface à partir du calcul d'aire accessible. *On calcule l'Aire de Surface Accessible (ASA) de la chaîne A et de la chaîne B, à la fois sous forme libre (lorsqu'elles sont prises séparément), mais aussi sous forme liée (lorsqu'elles interagissent). Les résidus interagissants sont les résidus qui perdent de l'accessibilité entre la forme libre et la forme liée. L'aire d'interface est la somme des aires enfouies pour chaque site de liaison.*

Si B_{AB} est l'aire totale enfouie dans l'assemblage moléculaire formé des chaînes A et B, c'est-à-dire l'aire de l'interface des chaînes A et B, alors :

$$B_{AB} = ASA_{chaîneA} + ASA_{chaîneB} - ASA_{chaînesAB}$$

Où $ASA_{chaîneA}$ (resp. $ASA_{chaîneB}$) est l'aire accessible au solvant de la chaîne A (resp. de la chaîne B) et où $ASA_{chaîneAB}$ est l'aire accessible au solvant de l'assemblage moléculaire formé des chaînes A et B [177].

Dès lors, les résidus interagissants correspondent aux résidus qui ont perdu de l'accessibilité entre leur forme libre et leur forme liée [153-155, 177]. Il est de plus possible de calculer l'aire enfouie pour chacun des résidus interagissants en soustrayant leur accessibilité sous forme libre à leur accessibilité sous forme liée.

Une étude de 1998, basée sur des mutants ponctuels, suggère cependant que l'aire enfouie des résidus ne corrèle que modérément (r=0.44) avec leur contribution à l'énergie libre de liaison $\Delta\Delta G$ [243]. Les deux approches soulignent cependant l'importance des résidus aromatiques dans la formation des interfaces. Toutefois, faisons la remarque que si l'interface formée par les sites de liaison des chaînes A et B n'est pas parfaitement complémentaire, il peut arriver qu'une crevasse se retrouve enfouie dans l'assemblage A-B et induise la détection non justifiée de plusieurs résidus interagissants [244]. A notre connaissance, cette analyse de la corrélation entre les résidus contribuant beaucoup à l'aire d'interface, et les résidus contribuant beaucoup à l'énergie libre de liaison (*hot spot*), n'a pas été renouvelée en considérant le possible biais introduit par la présence de cavités à l'interface qui tend à détecter davantage de résidus interagissants. De plus, précisons que la contribution à l'énergie libre de liaison est évaluée par l'approche d'alanine scanning qui consiste à remplacer les résidus de l'interface par une alanine ; par conséquent, cette mesure de l'énergie libre doit également refléter dans une certaine mesure le remplacement d'un résidu par une alanine, et non seulement la contribution énergétique d'un résidu à l'interface.

3. Détection des résidus interagissants par les tesselations de Voronoï et dérivées

La tesselation de Voronoï fournit un ensemble de cellules qui décrit de façon non ambigüe le volume et le voisinage de chaque atome. Par conséquent, la représentation de Voronoï permet d'accéder directement aux contacts entre atomes et de déterminer les résidus interagissants d'une interface sans procéder au calcul exhaustif des distances inter-atomiques, et sans nécessiter l'élaboration d'une distance seuil [235, 245, 246] (Figure 44).

Ainsi, F.Cazals et coll. ont montré que si tous les atomes interagissants perdant de l'accessibilité lors de la formation d'un assemblage sont bien détectés par la construction de Voronoï, cette nouvelle notion de contact, basée sur le voisinage des cellules, permet également d'identifier des interactions entre atomes enfouis dans les chaînes interagissantes. Le nombre moyen de contacts inter-atomiques par atome à l'interface est ainsi de 3.5 [245].

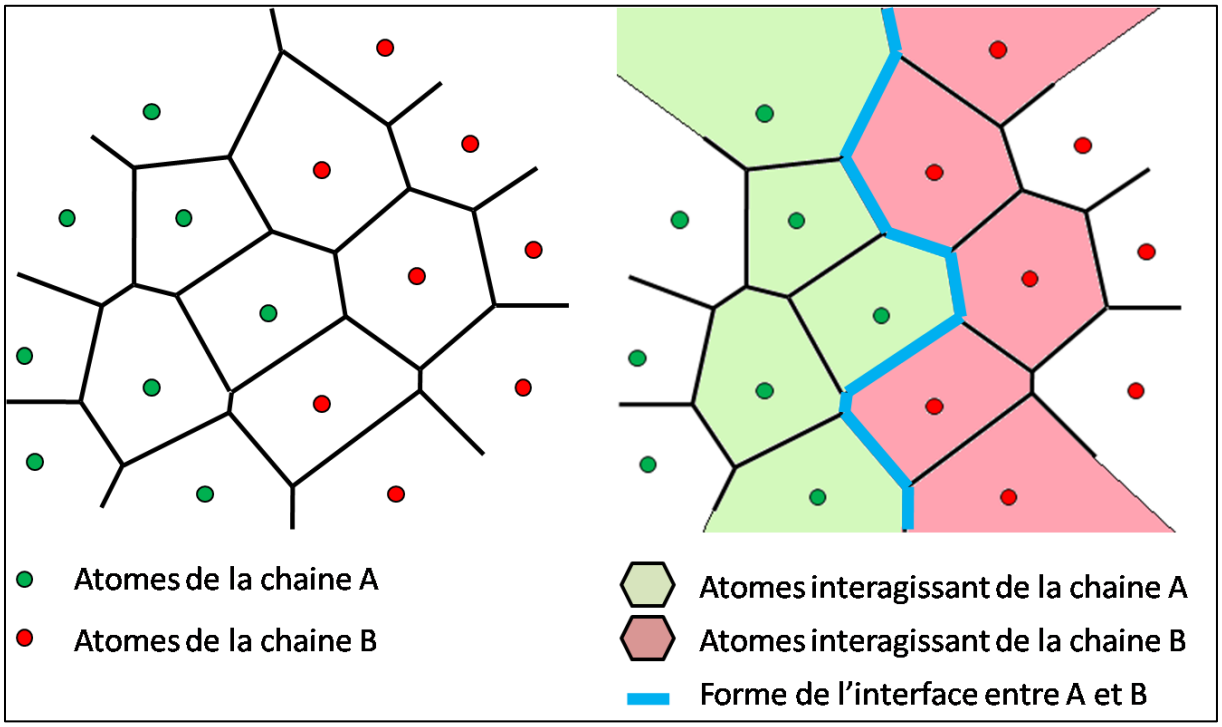


Figure 44. Tesselation de Voronoï d'une interface et détection des atomes et résidus interagissants. *Les atomes des chaînes A et B ont été ici représentés respectivement en vert et rouge, et l'on a construit la tesselation de Voronoï pour cet ensemble d'atomes. Dans ce diagramme, un atome de A interagit avec un atome de B si les cellules de ces deux atomes partagent une arête en commun. Le diagramme de droite précise pour cette tesselation, quels sont les atomes interagissants de la chaîne A, et ceux de la chaîne B.*

Le complexe alpha que nous avons utilisé dans ce travail de thèse est une forme duale de la tesselation de Voronoï et nous a également permis d'explorer et d'exploiter les contacts inter-atomiques et inter-résidus, notamment pour générer des régions moléculaires contiguës et pour étudier les propriétés des sites de liaisons [4]. La description des molécules en fonction de leurs régions permet également de comparer ces molécules indépendamment de leurs séquences ou de leurs structures secondaires et tertiaires. Ce point sera davantage abordé dans la partie « Criblage virtuel haut-débit des régions moléculaires ».

L'utilisation du complexe alpha et de la forme alpha plutôt, que la tesselation de Voronoï permet de plus de faciliter l'observation et l'étude de la complémentarité des interfaces (Figures 44 et 45). La complémentarité observée à l'interface par les formes alpha permettra également de procéder au criblage des régions complémentaires (Chapitre « Criblage haut-débit des régions moléculaires ») afin de procéder à une variante accélérée de *docking rigide* (logique de *double hybride in silico*).

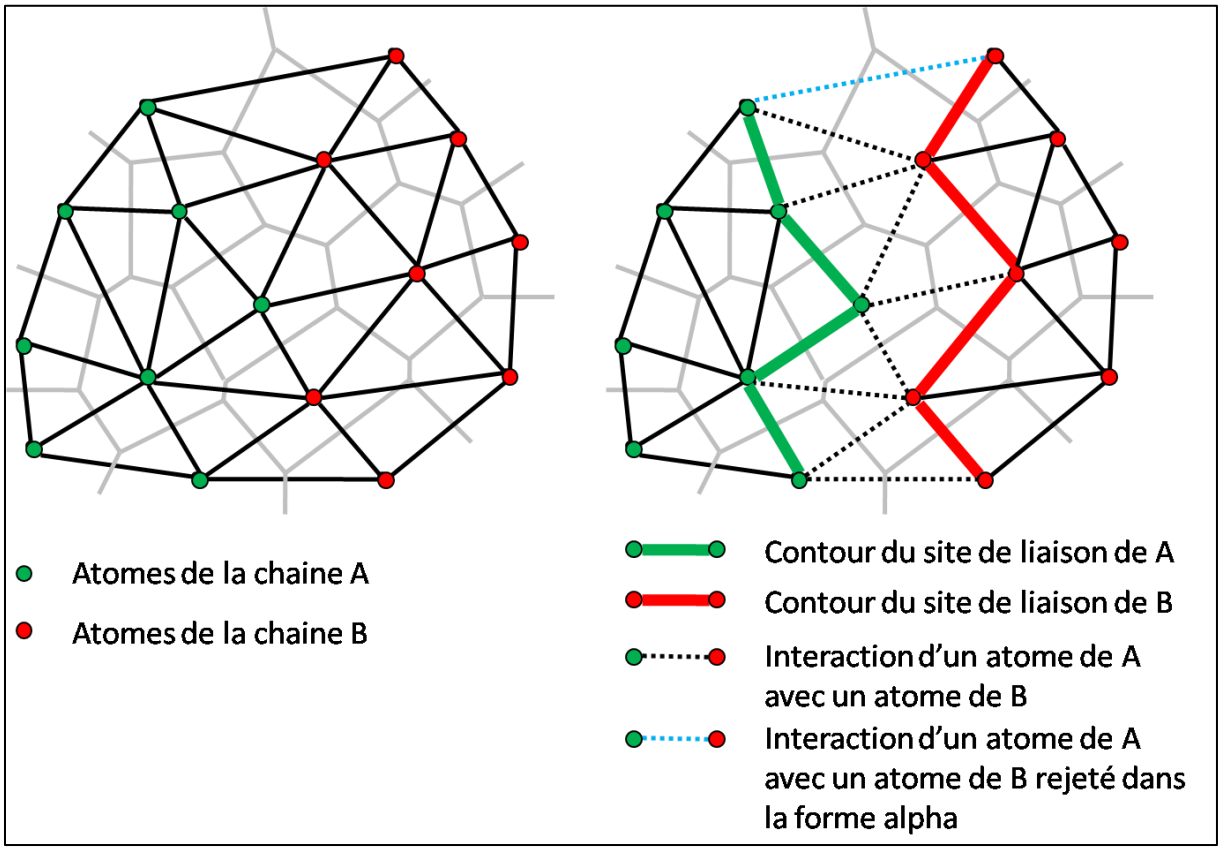


Figure 45. Complexe de Delaunay et complexe alpha d'une interface et détection des atomes et résidus interagissants. *Le complexe de Delaunay, tout comme la tesselation de Voronoï, renseigne sur le volume des atomes et fournit une notion non ambigüe du voisinage permettant de définir les atomes et résidus de l'interface. Alors que la tesselation de Voronoï va renseigner sur la forme de l'interface, le complexe de Delaunay renseigne sur la forme des sites de liaisons des chaînes A et B, et permet également d'étudier leur complémentarité.*

4. Amarrage moléculaire

Le *docking*, ou « amarrage moléculaire », est une approche généralement en deux temps : dans un premier temps, on génère l'ensemble des conformations définissables entre deux structures correspondant aux deux partenaires ; puis dans un second temps, on évalue par des scores d'énergies quelles sont les conformations les plus stables. Le premier

algorithme inventé par S. Wodak et J. Janin en 1978 [178] consiste à déplacer un ligand autour d'un récepteur en faisant varier 6 degrés de liberté (5 rotations et une translation) de sorte qu'ils soient en contact et génèrent des conformations d'assemblages (Figure 46). L'approche fut reprise en 1991 en améliorant la minimisation de l'énergie [247]. En 1992, Katchalski-Katzir et coll. proposèrent la première méthode de *docking* par transformée de Fourier rapide [125]. Bien que plus rapide, cette approche reste cependant trop longue pour une application haut-débit (une protéine contre toutes les structures disponibles). De nombreuses autres méthodes de *docking* ont été développées depuis et certaines sont référencées dans le Tableau 11.

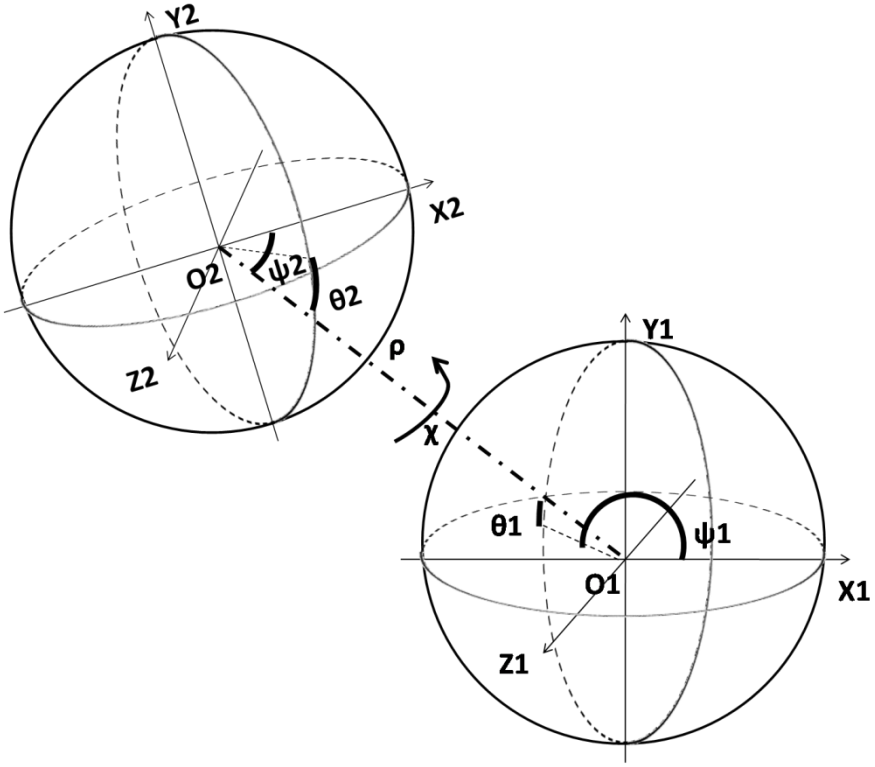


Figure 46. Les six degrés de liberté dans la recherche de la conformation d'un assemblage moléculaire, tels qu'utilisés dans le programme DOCK. *ϑ1 et ψ1 (resp. ϑ2 et ψ2) sont les latitudes et longitudes de la molécule 1 centrée en O1 (resp. de la molécule 2 centrée en O2). ρ est la distance séparant les centres O1 et O2 des molécules, et χ est une libre rotation sur l'axe O1O2. La position de la molécule 2 centrée en O2 peut donc être retrouvée et modulée à partir du repère (X1, Y1, Z1) et de ces 6 paramètres.*

L'élaboration du score d'énergie permettant de différencier les interfaces biologiques des « fausses » interfaces (ou leurres) nécessite des jeux d'interfaces vérifiées à partir de données expérimentales sûres. La base de données Dockground propose par exemple des jeux de données pour ces deux types d'interfaces (biologiques et fausses) entre protéines [248, 249]. La littérature propose également un jeu de données d'assemblages protéiques

qui précisent à la fois les formes libres et les formes liées des structures, de sorte qu'il soit possible de prédire la conformation d'un assemblage à partir des seules formes libres [250, 251]. Le concours international CAPRI [252] recense les succès et les échecs d'un grand nombre de méthodes de *docking* existantes (Tableau 11). CAPRI a permis une évaluation critique des performances de ces méthodes en précisant leurs points forts et leurs points faibles, c'est-à-dire les conditions dans lesquelles il est raisonnable de faire confiance à une prédiction. Au cours des années, CAPRI a permis d'observer l'évolution positive des prédictions : si la prédiction des assemblages « rigides » (ceux n'induisant pas ou peu de changements conformationnels dans les sous-unités intéragissantes) donne de bons résultats, le véritable défi est dans la prédiction « flexible » des assemblages (ceux induisant des changements de conformations dans les sous-unités intéragissantes), sans pour autant générer de nouvelles conformations déraisonnables.

ClusPro	http://cluspro.bu.edu/	Amarrage protéine-protéine
GRAMM-X	http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/	Amarrage protéine-protéine
HADDOCK	http://haddock.chem.uu.nl/	Amarrage protéine-protéine
PatchDock	http://bioinfo3d.cs.tau.ac.il/PatchDock	Amarrage basé sur la complémentarité des formes
PRISM	http://prism.ccbb.ku.edu.tr/prism/	Intéraction protéique par reconnaissance structurale (*PRotein Interactions by Structural Matching*)
RosettaDock	http://rosettadock.graylab.jhu.edu/	Amarrage protéine-protéine
SKE-Dock	http://www.pharm.kitasato-u.ac.jp/biomoleculardesign/files/ske_dock.htm	Amarrage protéine-protéine
SmoothDock	http://structures.pitt.edu/servers/smoothdock/	Amarrage protéine désordonnée-peptides
SymmDock	http://bioinfo3d.cs.tau.ac.il/SymmDock/	Prédiction des assemblages de symétrie Cn
Dockground	http://dockground.bioinformatics.ku.edu/	Jeux d'assemblages biologiques et d'assemblages non-spécifiques (*decoys*)
CAPRI	http://capri.ebi.ac.uk/	Evaluation critique des prédictions d'interactions (*Critical Assessment of PRediction of Interactions*)

Tableau 11. Différentes approches d'amarrage protéine-protéine existantes sur internet. *(source : Janin et coll., 2008 [98])*

Les méthodes d'amarrages moléculaires sont à l'origine de la découverte de nombreux ligands capables de lier des cibles protéiques [253]. Dans l'industrie pharmaceutique, de très nombreux amarrages sont effectués entre une cible d'intérêt et une collection de petites molécules (ligands). Seule une fraction de ces petites molécules (*hits*) ayant les meilleurs scores d'énergie est ensuite testée expérimentalement pour définir notamment leur constance d'association/dissociation. De telles études ont par exemple permis de trouver des ligands agonistes capables de lier les récepteurs à l'acide rétinoïque avec une affinité de l'ordre du micromolaire et une efficacité vérifiée en culture cellulaire [254].

III. Les caractéristiques géométriques et physico-chimiques des interfaces

Dans cette troisième partie, nous reprenons les méthodologies décrites précédemment pour caractériser les différents types d'interactions moléculaires. Les travaux sur les interfaces cristallines, des interfaces non-spécifiques et par définition non représentatives d'assemblages biologiques, seront détaillés, ainsi que les méthodes permettant de les distinguer des interfaces spécifiques/biologiques.

Nous préciserons les concepts de cœur et de bord d'interface, ainsi que de régions (ou patch) d'interaction, avant de présenter quelques méthodes de prédiction des sites de liaison, et de conclure sur leurs limitations actuelles.

1. Les différents types d'interfaces moléculaires

a. *Présentation générale*

Les premières études pionnières sur les interactions moléculaires remontent aux années 1970, alors même que moins d'une centaine de structures étaient disponibles dans la Protein Data Bank [255]. Ainsi, en 1975 et 1976, C. Chothia et J. Janin proposèrent les premières analyses « systématiques » de plusieurs assemblages protéiques [177, 256, 257].

En 1988, une étude spécialisée dans les **homo-oligomères** (association d'au moins deux protéines identiques) montre que l'aire d'interface varie de 700 Å² à plus de 10 000 Å² pour la catalase.

Les études de 1995 [258] et 1999 [153] sur les **hétérodimères** (association de deux protéines différentes) démontrent, sur davantage de données structurales, que les assemblages ayant une aire d'interface de l'ordre de 1600 Å² ne subissent que peu de changements conformationnels en comparaison des interfaces plus grandes (>2000 Å²).

Suivront les études des interactions **protéine-peptide** dont la diversité des mécanismes de reconnaissances fut décrite par R. Stanfield en 1995 [259]; les interactions **protéine-ADN** dès 1999 [179, 260] et **protéine-ARN** dès 2001 [45, 261]. La figure 46 résume les différents types d'interactions représentées dans la Protein Data Bank.

Le **rôle médiateur de l'eau** dans les interactions moléculaires ne sera étudié que plus tard par Rodier et coll. en 2005 [180], montrant l'existence d'interface « sèche » (*dry interface*) où l'eau encercle l'interface, et d'interface « humide » (*wet interface*) où l'eau sert de médiateur pour quelques interactions entre les sites de liaisons de l'interface :

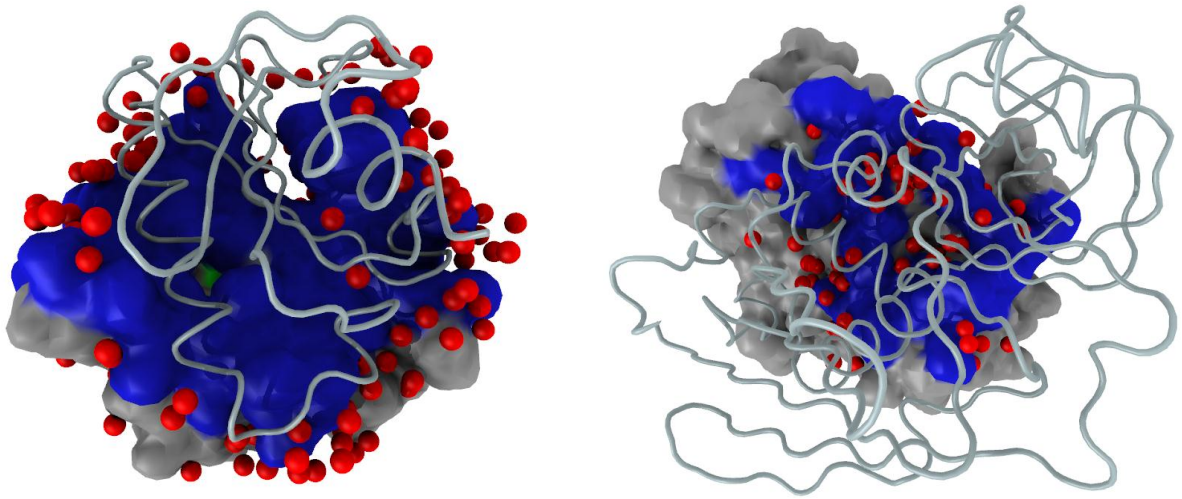


Figure 47. Interfaces sèches et rôle médiateur de l'eau dans les interfaces humides. *A gauche, l'interface sèche de l'homodimère de la protéine kinase C (1kpf). A droite, l'interface humide de l'homodimère de l'entérotoxine de type A (1i4g). Les sites de liaison protéine-protéine sont indiqués en bleu ; pour plus de clarté, seules les molécules d'eau (en rouge) à moins de 10 Å des atomes de l'interface ont été représentées.*

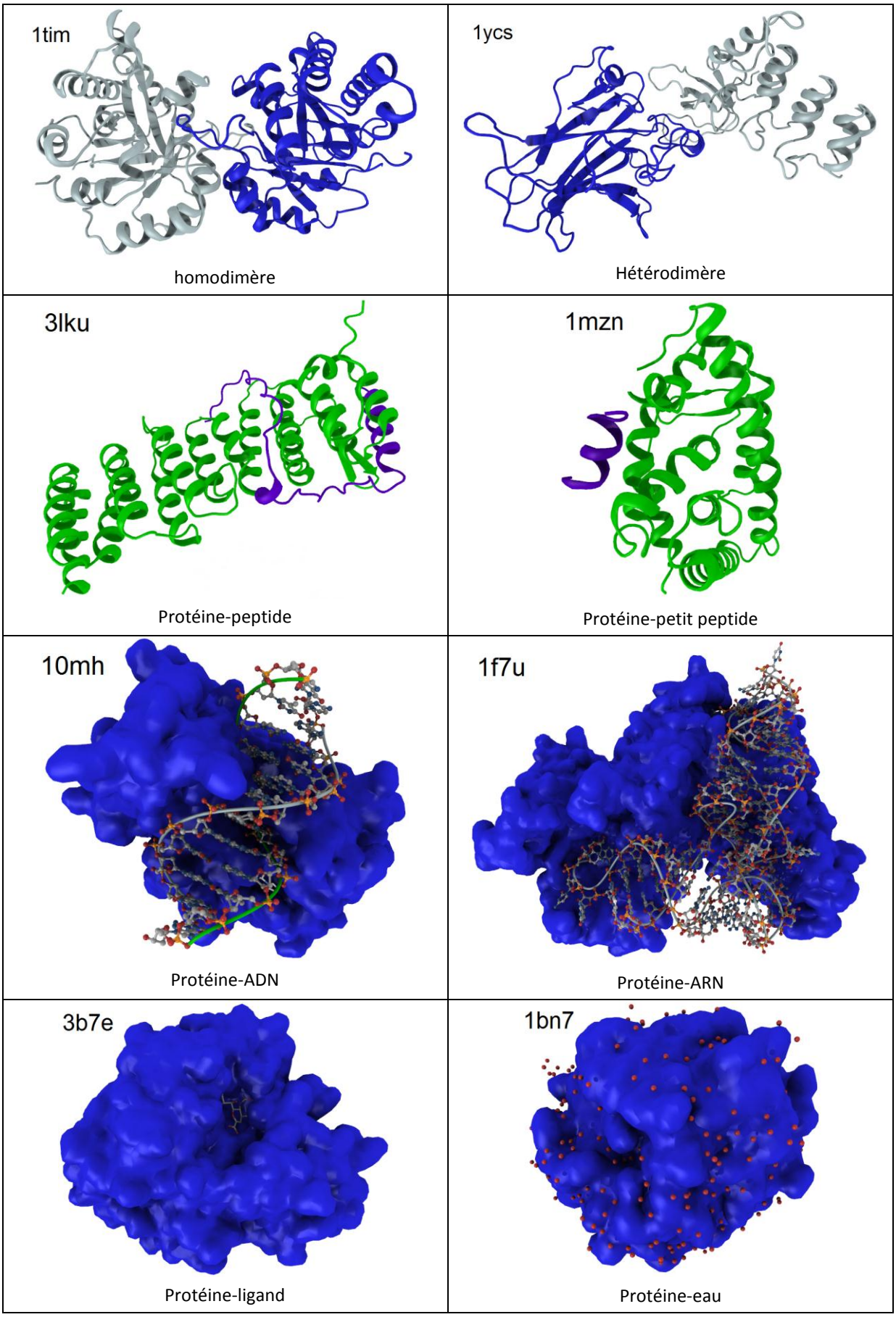


Figure 48. Les différents types d'interactions moléculaires.

Parallèlement à ces recherches, l'essor des données structurales et cristallographiques a permis de mettre en évidence, dès 1995 [106, 182], la fréquence et l'importance des **contacts cristallins** dans les structures biologiques. Depuis, de multiples outils de prédiction [156, 183, 246, 262] et des bases de données, telles que PQS et PISA [263, 264] ont vu le jour pour différencier les contacts biologiques (spécifiques) des contacts cristallins (non-spécifiques) à partir des données cristallographiques.

b. Problématiques et objectifs

Comme nous le verrons par la suite, la quasi-totalité des études réalisées sur les données structurales se focalise sur l'analyse d'un type d'interaction, et compare les propriétés observées pour ces interactions au restant de la surface non-interagissante. Toutefois, dans ces études, la description de cette surface non-interagissante est assez limitée puisqu'elle est formée de l'ensemble des résidus de surface qui ne sont pas vus comme interagissants dans la structure étudiée.

Une première problématique tient donc dans l'utilisation et l'intégration des données structurales pour mieux décrire l'ensemble des interactions connues pour une protéine étudiée, et sans se limiter aux seules données d'une structure.

Une seconde problématique, résultant de la première, est de mieux décrire la surface interagissante des protéines, c'est-à-dire la surface capable d'interagir avec des molécules autres que celles du solvant. Notamment, quel est le pourcentage moyen de la surface capable de participer à de telles interactions.

Une troisième problématique immédiate est de vérifier si les différents sites de liaisons capables de participer à différents types d'interactions, sont localisés en des endroits distincts de la surface protéique, ou bien s'ils se recouvrent.

Si les différents sites de liaisons peuvent se recouvrir, existe-t-il des sites de liaisons polyvalents, c'est-à-dire, capables de lier plusieurs types de molécules, et des sites de liaisons spécifiques, c'est-à-dire, capables de lier uniquement un type moléculaire ?

Enfin, en raison de la reconnaissance tardive de l'importance des molécules d'eau pour la stabilité des protéines et pour médier les interactions, leur détermination dans les structures biologiques a été pendant longtemps des plus partielles. La détection de ces molécules d'eau dans les cartes de densité électronique dépend également de la résolution de la structure, si bien que pour ces deux raisons, il est difficile de déterminer les molécules d'eau qui lient la protéine de façon spécifique. Toutefois, la Protein Data Bank présente un inconvénient qui dans ce cas précis s'avère un atout : cette base de données est très redondante et bien

souvent lorsqu'il existe la structure d'une protéine, on en retrouve une dizaine, voire plusieurs centaines de variantes. Aussi, la cinquième et dernière problématique consiste dans la détermination des molécules d'eau liant spécifiquement les protéines, et dans la détermination du pourcentage moyen des résidus de surface hydratés.

2. Caractéristiques des interfaces moléculaires

a. Oligomères et homodimères

En 1988, J. Janin et coll. étudièrent les différences entre la surface, le cœur protéique et les interfaces de 23 oligomères protéiques, montrant que la composition en acides aminés différait entre ces différentes régions [228]. Ils observèrent également que les interfaces sont plus hydrophobes que le restant de la surface, mais présentent toutefois des résidus chargés. Les aires d'interface pour ces 23 oligomères varient de 700 Á² à 10 000 Á², montrant une importante diversité chimique et fonctionnelle impliquée dans la reconnaissance des partenaires.

En 2003, R. Bahadur et coll. étudient les interfaces de 122 homodimères et identifient par des techniques de regroupement (*clustering*) sur les atomes interagissants [154], les différentes régions (ou *patchs*) formant chaque interface. Ils montrent ainsi que l'aire moyenne d'interface par chaîne interagissante est de l'ordre de 1940 $\mp$ 1100 Å², et que l'aire d'interface par région (patch) est de l'ordre de 1370 $\mp$ 620 Å². Les interfaces étudiées comprennent de 1 à 6 régions distinctes d'interaction (ou encore, sites de liaison). Sur la base de l'accessibilité des surfaces, on observe que les résidus hydrophobes (resp. aromatiques, chargés) contribuent 1,4 fois plus (resp. 2,1 fois plus, 1,7 fois moins) à l'aire d'interface qu'à l'aire de surface (Tableau 12).

La notion introduite par Lo Conte [153] distingue trois classes d'atomes d'interface en fonction de leurs accessibilités dans les formes libres et liées des assemblages, et en fonction de s'ils sont à des distances de Van der Waals d'atomes du site de liaison opposé.

Résidu	Composition de la surface	Composition de l'interface	Contribution à l'aire de surface	Contribution à l'aire d'interface
Ala	7,6	7,4	5,2	5,2
Arg	6,5	6,4	9,8	9,7
Asn	5,0	4,7	5,1	3,8
Asp	7,5	5,5	8,2	4,3
Cys	1,0	1,4	0,2	1,2
Gln	4,7	4,0	6,1	4,1
Glu	8,8	6,2	11,9	5,6
Gly	7,5	6,1	4,4	3,5
His	2,6	3,2	2,6	3,3
Ile	3,5	4,9	2,4	5,5
Leu	6,1	9,0	4,4	10,4
Lys	8,1	5,4	13,2	5,4
Met	1,6	2,9	1,5	4,0
Phe	2,8	4,6	2,0	6,3
Pro	5,7	5,2	5,9	4,9
Ser	5,9	5,6	4,7	4,0
Thr	5,7	5,4	4,8	4,3
Trp	1,2	1,7	1,1	2,6
Tyr	3,3	4,6	2,9	6,3
Val	4,8	6,0	3,3	5,6
Hydrophobe	36,8	41,5	27,1	39,1
Aromatique	9,9	14,1	8,6	18,5
Polaire	21,3	19,7	20,7	16,2
Chargé	30,9	23,5	43,1	25

Tableau 12. Une composition différente pour la surface et les interfaces d'homodimères. *(adapté à partir de Bahadur et coll., 2003 [155]).*

A partir de cette notion, Chakrabarti et coll. proposèrent de distinguer les résidus du cœur de l'interface (*core residues*) et du bord de l'interface (*rim residues*) sur la base de l'accessibilité de leurs atomes : les résidus du cœur de l'interface possèdent au moins un atome complètement enfoui, alors que les résidus du bord de l'interface ne possèdent que des atomes au moins partiellement accessibles (Tableau 13). Les résidus du cœur constituent en moyenne 77% de l'interface, les résidus du bord formant les 23% restants.

Résidu	Contribution à l'aire d'interface	Contribution à l'aire du cœur de l'interface	Contribution à l'aire du bord de l'interface
Ala	5,2	5,2	5,0
Arg	9,7	9,8	9,4
Asn	3,8	3,5	4,8
Asp	4,3	3,8	6,1
Cys	1,2	1,2	1,5
Gln	4,1	3,9	4,7
Glu	5,6	5,2	6,7
Gly	3,5	3,0	5,4
His	3,3	3,5	2,5
Ile	5,5	5,9	3,9
Leu	10,4	11,2	7,9
Lys	5,4	4,1	9,7
Met	4,0	4,1	3,8
Phe	6,3	7,5	2,4
Pro	4,9	4,4	6,7
Ser	4,0	3,7	4,9
Thr	4,3	4,1	4,8
Trp	2,6	3,0	1,3
Tyr	6,3	7,3	3,1
Val	5,6	5,6	5,5
Hydrophobe	39,1	39,4	38,2
Aromatique	18,5	21,3	9,3
Polaire	16,2	15,2	19,2
Chargé	25	22,9	31,9

Tableau 13. Une composition différente pour le cœur et le bord des interfaces homodimériques. *(adapté à partir de Bahadur et coll., 2003 [155]).*

Ainsi, si la composition des interfaces d'homodimères diffère de la composition du restant de la surface, la composition du cœur de l'interface ressemble davantage au cœur protéique et les résidus du bord de l'interface ressemblent davantage au restant de la surface. En particulier, il y a deux fois plus de résidus aromatiques et 1,4 fois moins de résidus chargés dans les résidus du cœur de l'interface homodimérique que dans les résidus du bord de l'interface.

b. Hétérodimères

En 1999, Lo Conte et coll. étudièrent un jeu de 75 hétérodimères (aussi appelés complexes), montrant que les changements conformationnels induits par la formation d'un assemblage sont corrélés à l'aire d'interface de cet assemblage [153]. Ainsi, 52 de ces 75 hétérodimères ont une aire d'interface moyenne de 1600 Å² ∓ 400Å², qui se traduit uniquement par de petits changements conformationnels. 20 autres de ces hétérodimères ont de grandes interfaces (variant de 2000 à 4660 Å²) et leurs formations impliquent des changements conformationnels importants. Ces remaniements structuraux sont généralement de trois types : (1) des remaniements des chaînes latérales, (2) des remaniements du squelette peptidique et, (3) stabilisation de tout ou partie d'une molécule désordonnée, la formation de l'assemblage conférant à cette molécule (ou à une région de la molécule) une structure ordonnée. La partie désordonnée C-terminale de l'hirudine, se réorganise lors de l'association avec la thrombine pour former une structure stable. Les régions « *hinge* » des récepteurs nucléaires qui relient les domaines de liaison au ligand (LBD) et de liaison à l'ADN (DBD) sont elles aussi des régions désordonnées de 15 à plus de 100 acides aminés dont la dynamique sert à accommoder différentes conformations [265, 266]. La détermination en 2008 de la structure complète de l'hétérodimère de PPARγ et RXRα met en évidence des interactions qui stabilisent cette région « hinge » et permettent son élucidation [266, 267].

A la différence des homodimères, les hétérodimères ont une composition non polaire proche du restant de la surface, ce qui complique leur détection par des approches prédictives (Tableau 14). Bien qu'elles comportent généralement moins de résidus chargés que le restant de la surface, les interfaces d'hétérodimères sont plus hétérogènes que celles des homodimères et leur composition, aussi bien que leur taille, peut varier de façon significative. En règle générale, les interfaces d'homodimères sont 1,3 fois plus hydrophobes que les interfaces d'hétérodimères, et 1,2 fois moins chargées que les interfaces d'hétérodimères.

Résidu	Contribution à l'aire d'interface (HoD)	Contribution à l'aire d'interface (HeD)
Ala	5,2	3,2
Arg	9,7	9,3
Asn	3,8	5,1
Asp	4,3	6,1
Cys	1,2	1,1
Gln	4,1	4,3
Glu	5,6	7,2
Gly	3,5	3,4
His	3,3	4,4
Ile	5,5	4,0
Leu	10,4	5,6
Lys	5,4	7,6
Met	4,0	4,2
Phe	6,3	5,0
Pro	4,9	4,2
Ser	4,0	4,6
Thr	4,3	4,8
Trp	2,6	3,4
Tyr	6,3	7,2
Val	5,6	4,7
Hydrophobe	39,1	29,3
Aromatique	18,5	20
Polaire	16,2	18,8
Chargé	25	30,2

Tableau 14. Compositions des interfaces d'homodimères et d'hétérodimères. *(adapté à partir de Lo Conte et coll. pour les hétérodimères (HeD) [153] et de Bahadur et coll. pour les homodimères (HoD) [156])*

En 2002, Chakrabarti et coll. montrent qu'il peut être important d'étudier les interactions moléculaires sur la base de « *patch* » ou régions locales de reconnaissance [154]. Dans cette étude, les régions sont formées à partir d'un regroupement spatial des atomes d'interface les plus proches (Figure 49). Alors que l'aire d'interface des 70 hétérodimères étudiés varie de 900 à plus de 5000 Å² avec une valeur moyenne de 1906 $\mp$ 756 Å², les auteurs montrent que les aires d'interfaces formées d'une seule région de reconnaissance sont quant à elles de l'ordre de 1560 Å², avec un écart type plus réduit de 340 Å². Ainsi, 43 des 46 interfaces

composées d'une seule région de reconnaissance mesurent moins de 2000 Å². Cela signifie que par molécule, une région de reconnaissance mesure environ 800 Å², soit environ 85 atomes.

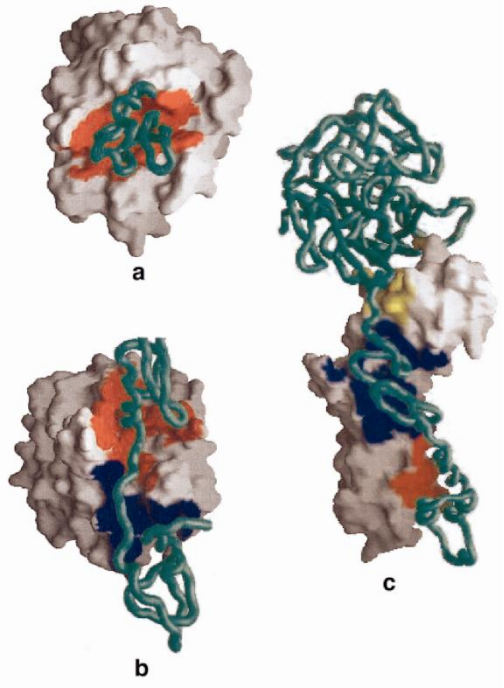


Figure 49. Régions de reconnaissance dans des complexes protéase-inhibiteur. *(source : Chakrabarti et coll., 2002 [154]). Les différentes sous-régions de chaque interface sont indiquées en rouge, bleu et jaune. En a) l'interface de la trypsine et de PTI ne comprend qu'une région (2ptc). En b) l'interface de l'omithodorine avec la thrombine a deux sous-régions (1toc). En c) l'interface du facteur VIIA de coagulation sanguine comporte trois sous-régions de reconnaissance avec le facteur de tissu soluble (1dan).*

La distinction des résidus du cœur et du bord de l'interface montre que le cœur est 2,5 fois plus aromatique, 1,5 fois moins chargé et 1,2 fois moins polaire que le bord de l'interface (Tableau 15). La composition en résidus hydrophobes ne varie pas entre le cœur et le bord de l'interface, et le restant de la surface non-interagissante (Tableau 14).

La comparaison du cœur des interfaces montre que les homodimères sont en moyenne 1,4 fois plus hydrophobes, 1,2 fois moins aromatiques et 1,3 fois moins polaires que les hétérodimères (Tableau 16). La différence d'hydrophobicité, qui se retrouve également dans le bord de l'interface, est donc bien une propriété propre des homodimères, corrélée avec une diminution des résidus chargés et polaires.

Résidu	Contribution à l'aire d'interface	Contribution à l'aire du cœur de l'interface	Contribution à l'aire du bord de l'interface
Ala	2,8	2,7	3,1
Arg	10,1	10,1	9,9
Asn	5,7	5,4	6,4
Asp	5,1	4,5	6,6
Cys	1,7	1,9	1,3
Gln	4,3	4,3	4,2
Glu	6,0	4,4	10,0
Gly	4,8	4,2	6,4
His	3,8	4,4	2,4
Ile	4,6	4,9	3,5
Leu	5,7	5,8	5,3
Lys	6,5	5,2	9,7
Met	3,2	3,7	2,0
Phe	4,1	5,5	1,1
Pro	3,6	3,5	4,1
Ser	5,4	4,8	7,3
Thr	5,0	4,7	5,9
Trp	4,2	5,3	1,6
Tyr	9,4	10,9	5,3
Val	3,8	3,8	3,9
Hydrophobe	28,5	28,6	28,3
Aromatique	21,5	26,1	10,4
Polaire	20,4	19,2	23,8
Chargé	27,7	24,2	36,2

Tableau 15. Une composition différente pour le cœur et le bord des interfaces hétérodimériques. *(adapté à partir de Chakrabarti et coll., 2002 [154]).*

La différence essentielle entre les résidus du cœur et du bord de l'interface semble être la diminution drastique (2,5 fois) du nombre de résidus aromatiques dans le bord de l'interface. L'une des raisons de cette observation tient peut-être dans la taille des résidus aromatiques qui pourrait provoquer une gêne stérique entre les sites de liaisons interagissants, suffisante pour en diminuer la vitesse d'association.

Type de résidu	Cœur (HoD)	Cœur (HeD)	Bord (HoD)	Bord (HeD)
Hydrophobe	39,4	28,6	38,2	28,3
Aromatique	21,3	26,1	9,3	10,4
Polaire	15,2	19,2	19,2	23,8
Chargé	22,9	24,2	31,9	36,2

Tableau 16. Ressemblances et différences entre les homodimères et hétérodimères au niveau de leurs cœurs et bords d'interfaces. *HoD : homodimère, HeD : hétérodimère.*

De plus, comme nous le verrons par la suite dans la distinction des homodimères, hétérodimères et contacts cristallins, les homodimères forment des interfaces plus complémentaires que les hétérodimères [153, 156, 181].

c. Contacts et assemblages cristallins

Par définition, un assemblage cristallin est un assemblage non spécifique de molécules qui ne joue aucun rôle fonctionnel et qui par conséquent n'est pas sujet à la pression de sélection [106]. Dans les analyses systématiques d'interactions moléculaires, il est donc important et nécessaire de différencier ces assemblages cristallins non-spécifiques des assemblages biologiques spécifiques, qui eux ont un rôle fonctionnel.

L'étude de différentes formes cristallines de la ribonucléase A (cristallisée dans différents groupes d'espaces) montre que certaines interfaces cristallines recouvrent jusqu'à 40% de la surface. Si l'on considère toutes les formes possibles d'assemblages cristallins, la quasi-totalité de la surface semble pouvoir être impliquée dans ce genre de contact non-spécifique [268].

Bien que les contacts cristallins impliquent de mêmes forces physico-chimiques, le fait qu'ils n'aient pas de rôle fonctionnel et qu'ils ne soient pas soumis à la pression de sélection, a permis de dégager un certain nombre de propriétés remarquables permettant de les différencier des assemblages biologiques. Ainsi, il a été remarqué très tôt que l'aire d'interface des assemblages cristallins (évaluée à ~570 Å² en 1995 [182]) est dans l'ensemble inférieure à l'aire d'interface des assemblages biologiques (évaluée à ~1900 Å² pour les hétérodimères en 2002 [154] et à ~3900 Å² pour les homodimères en 2003 [155]) (Figure 50).

L'aspect « aléatoire » des interactions cristallines a de plus été renforcé par des expériences d'amarrages (*docking*) de petites protéines dans des conformations aléatoires qui montrent que leur aire moyenne d'interface est similaire à celle des interfaces cristallines.

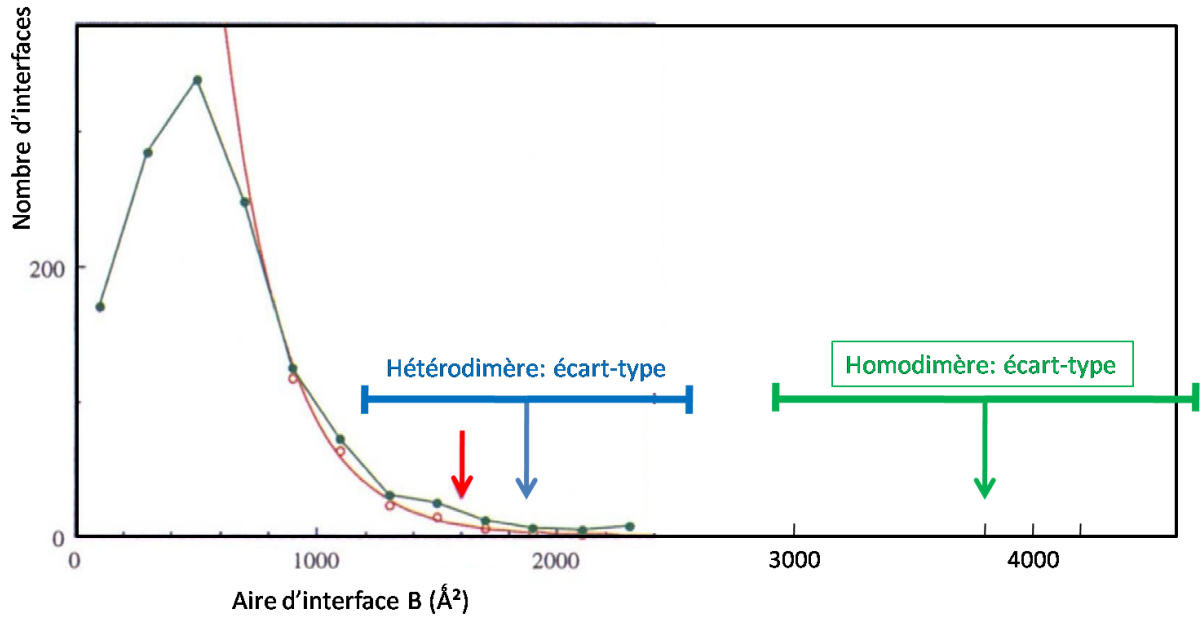


Figure 50. Distribution des aires d'interfaces cristallines à partir de 1320 couples d'interfaces. *La flèche rouge indique l'aire moyenne d'interface des complexes anticorps-antigènes (1600 Å²) ; les flèches bleu et verte indiquent l'aire moyenne des interfaces d'hétérodimères (1906* Å²) et d'homodimères (3880 Å²)*. (adaptée à partir de Janin et coll., 1997 [106])*

Si la distribution des aires d'interfaces cristallines représentée sur la figure 50 est représentative de l'ensemble des interfaces cristallines contenues dans les structures, alors les auteurs proposent une formule décrivant la probabilité qu'une interface soit non-spécifique, étant donné son aire d'interface :

$$P(B) = 4.2 \exp(-B/260)$$

Ainsi, une interface de 800 Å² a 19.4% de chance d'être une interface non-spécifique, alors que des interfaces de 1000, 1500 et 2000 Å² ont respectivement 9%, 1,3% et 0,2% de chances d'être non-spécifiques.

Ponstingl et coll. arrivent ainsi à différencier avec 85% de précision les monomères et dimères biologiques [183] en fixant un seuil d'aire d'interface de 856 Å². Notons à cet égard

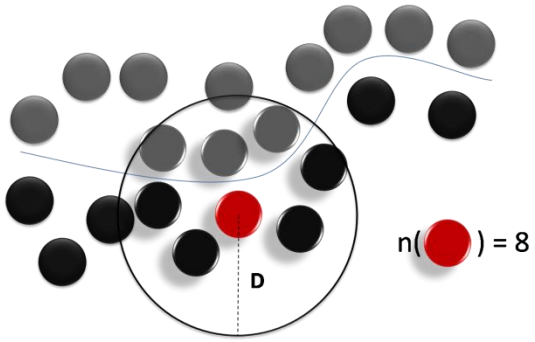
que la formule de P(B) prévoyait une erreur de 15.6%, soit très proche de ce qui fut observé par Ponstingl.

Toutefois, ainsi que noté dès 1988, il existe néanmoins des assemblages cristallins dont les aires d'interface sont du même ordre de grandeur que les aires d'interfaces biologiques [268], comme pour la phosphoglycérate kinase C (2760 Å², PDB : 13pk) ou encore la phosphoénolpyruvate carboxykinase (1848 Å², PDB : 1ayl). De plus, Il existe également des homodimères dont l'aire d'interface est relativement faible, avec l'exemple de la Carboxylestérase (1332 Å², PDB : 1auo).

Aussi, les assemblages ayant une faible aire d'interface ne sont pas tous des assemblages cristallins et les assemblages ayant des aires d'interface importantes ne sont pas tous des assemblages biologiques.

Ainsi, en 2004, Bahadur et coll. proposent de se concentrer sur l'analyse des assemblages cristallins dont l'aire d'interface est supérieure à 800 Å² afin de les comparer et différencier des interfaces biologiques. Les auteurs constituent ainsi un jeu de 188 assemblages cristallins dont l'aire moyenne d'interface est de 1660 Å², soit proche des aires moyennes d'interface des hétérodimères (encore appelés complexes), et de taille comparable à des homodimères pour lesquels il était précisément difficile de les distinguer des assemblages cristallins. Citons parmi ces homodimères, l'homodimère de RXRα chez l'homme, avec une interface de 2064 Å² (PDB : 1mzn) qui présente la particularité d'avoir un site de liaison polyvalent (voir article 2 [1]) qui lui permet de former à la fois des homodimères et des hétérodimères. Sur ce nouveau jeu d'assemblages cristallins combinés à 70 hétérodimères [154] et 122 homodimères [155], l'approche de Ponstingl et coll. a un taux d'erreur de 18% pour la prédiction des homodimères, mais de 25% pour la prédiction des monomères.

La distinction des assemblages cristallins et biologiques nécessite d'étudier en particulier la complémentarité et l'hydrophobicité des interfaces. Ainsi, la complémentarité d'une interface peut être évaluée en calculant la fraction d'atomes complètement enfouis (inaccessibles au solvant dû à la complémentarité des interfaces), et/ou en calculant la densité locale moyenne (LD) pour chaque atome de l'interface :



$$LD = \frac{1}{N}\sum n_i \text{ [156]}$$

Où N est le nombre d'atomes de l'interface, et n_i représente le nombre d'atomes de l'interface à l'intérieur d'une sphère de rayon D centrée sur l'atome i.

La fraction non polaire de l'interface peut être également calculée en mesurant notamment l'aire enfouie des atomes de carbone (non polaire). Il est également possible de décrire la tendance d'un résidu à faire partie d'une interface biologique spécifique, ou d'une interface cristalline non-spécifique en utilisant la formule suivante :

$$RP = \ln(f_i/f_i^0) \text{ [156]}$$

Où f_i est le nombre ou la contribution à l'aire d'interface du résidu de type i, et où f_i^0 est le nombre ou la contribution à l'aire accessible de surface du résidu de type i. Lorsque $f_i = f_i^0$, *RP* vaut 0, ce qui signifie que le résidu n'a pas de préférence pour la surface ou l'interface. Lorsque *RP* a une valeur positive, cela indique que le résidu est préférentiellement retrouvé à l'interface, et inversement, sur le restant de la surface non-interagissante.

	Dimères cristallins	**Homodimères**	**Complexes**
Aire d'interface (Å²)	1660 (570)	3880 (2200)	1910 (760)
Fnp (%)	58 (8)	65 (7)	58 (6)
Fbu (%)	22 (9)	36 (9)	34 (9)
No de liaisons hydrogènes	6 (4)	18,5 (15,2)	10 (4,8)
Densité Locale (LD)	32 (6)	45 (8)	42 (6)
Score de tendance du résidu (RP)	-0,8 (2,8)	4,3 (4,9)	0,9 (2,3)
Score d'interaction hydrophobique (HI)	87 (32)	223 (125)	106 (41)

Tableau 17. Comparaison de propriétés géométriques et physico-chimiques des interfaces spécifiques et non-spécifiques. *Les nombres entre parenthèses représentent les déviations standards. (adapté à partir de Bahadur et coll., 2004 [156]).*

En combinant l'aire enfouie contribuée par les atomes non polaires (*fnp*) à la fraction des atomes enfouis à l'interface (*fbu*), ainsi que le score de tendance des acides aminés à faire partie d'une interface biologique ou cristalline, Bahadur et coll. parviennent à obtenir un taux de prédiction de l'ordre de 93 à 95% (Figure 51).

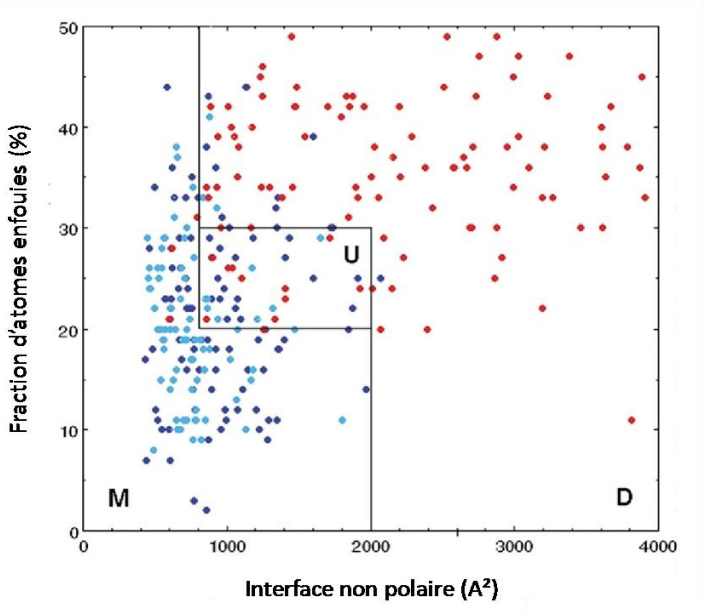


Figure 51. Distribution des interfaces biologiques et cristallines en fonction de l'aire non polaire et de la fraction des atomes enfouis. *La zone M correspond majoritairement à des monomères (donc des assemblages cristallins), la zone D à des dimères et la zone U aux assemblages pour lesquels il est difficile de distinguer entre interfaces biologiques et cristallines. (source : Bahadur et coll., 2004 [156]).*

Lors de ce travail de thèse, nous avons par ailleurs montré que les sites de liaison polyvalents homodimère-hétérodimère corrélaient davantage avec les contacts cristallins (0,62) que ne le font les sites de liaison spécifiques aux homodimères (corrélation de 0,1). De la même façon, les sites de liaison polyvalents homodimère-solvant sont beaucoup plus proches des contacts cristallins (0,95). Ce sujet sera davantage approfondi dans l'article sur la cartographie moléculaire, ainsi que dans les perspectives du chapitre correspondant à l'article [1].

d. Protéine-petit peptide

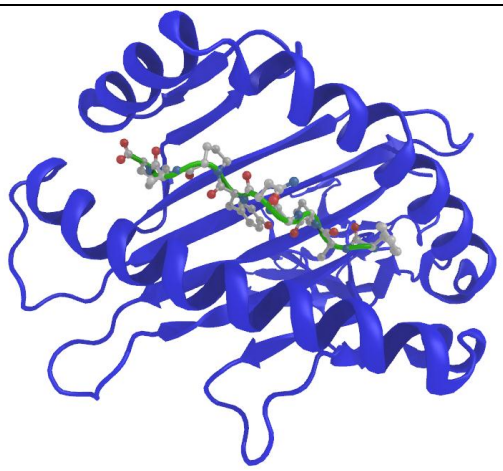
A la différence des peptides dont la taille varie en moyenne de 20 à 60-100 acides aminés, les petits peptides sont constitués d'après les observations de quelques auteurs [259, 269,

270] et de notre classification, de moins de 20 acides aminés en moyenne (Figure 52). La différenciation de ces deux types de polypeptide tient d'une part, dans la capacité des petits peptides à simuler le rôle et les comportements des ligands (y compris dans leur mode d'interaction en se liant par exemple dans des crevasses de la surface protéique), et d'autre part, dans la multiplicité des conformations que peuvent prendre un petit peptide et un peptide. De plus amples différences entre ces deux types de polypeptide sont abordées dans notre article de cartographie moléculaire [1].

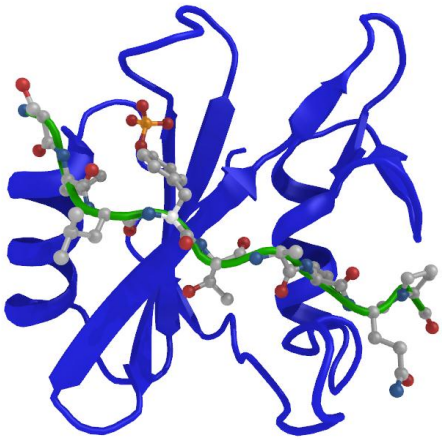
Une analyse que nous avons menée à partir des données de la Protein Data Bank montre qu'à la rédaction de ce manuscrit en 2010, 2888 différents petits peptides ont au moins une structure (polypeptide de moins de 20 acides aminés). En comparaison, 10 857 ligands ont au moins une structure dans la Protein Data Bank. Parmi les petits peptides les plus représentés, notons l'alpha thrombine, les variants d'huridine, un fragment de la chymotrypsine, les peptides interagissant avec les récepteurs glucocorticoïdes. L'analyse des 2888 petits peptides montre que seul 11,5% de leurs résidus forment des hélices, le restant formant soit des feuillets bêta à l'interface d'autres protéines (stabilisant ainsi le peptide), soit restant sous forme de boucle. En raison de leur faible taille, les petits peptides n'ont pas de cœur enfoui favorisant et stabilisant un repliement particulier.

Bien qu'à notre connaissance, il n'y ait pas eu d'étude systématique des interactions protéine-petit peptide, Russel et coll. décrivent que 15 à 40% des interactions dans une cellule pourraient être médiées par les petits peptides [271]. Les petits peptides seuls en solution n'ont généralement pas de conformations stables, mais ils adoptent une conformation stable lors de la liaison d'un partenaire [259, 269]. Récemment, Vanhee et coll. rapportent la construction de la base de données pepX qui contient 1431 assemblages protéine-petit peptide regroupés en 505 interfaces uniques [270].

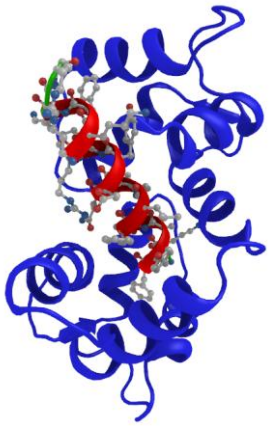
Malgré leur faible taille, les petits peptides peuvent néanmoins enfouir de 370 à plus de 1400 Å² de la surface de leur cible protéique. C'est le cas notamment du phosphopeptide B (1shb) qui enfouit 383 Å² de la tyrosine kinase, du peptide PDGFR-740 qui enfouit 374 Å² de la tyrosine phosphatase (1ayc), ou encore du peptide CaMKII qui enfouit 1128 Å² de la calmoduline [259].



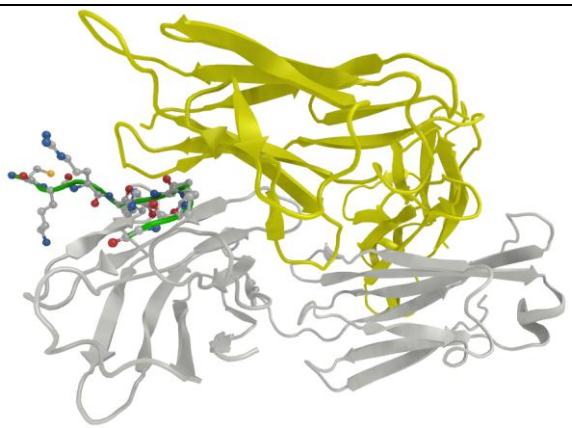
Protéine MHC liée à un fragment peptidique du virus sendai (2vab)



Domaine SH2 liant le peptide PDGFR-2009 (1aya)



Calmoduline liée à un fragment peptidique du canal calcium voltage dépendant (2f3y)



Fab 50.1 liée au peptide KRIHIGPG de HIV-1 (1ggi)

Figure 52. Quelques exemples d'interactions protéine-petit peptide.

e. Protéine-ADN

Bien qu'essentiels à l'étude du vivant et de la régulation de l'expression des gènes, les assemblages protéines-ADN sont encore peu représentés dans la Protein Data Bank et dans la Nucleic Acid Bank en raison des difficultés techniques (stabilisation) devant être relevées pour obtenir ces structures.

En 1999, Jones et coll. étudient un jeu de 26 assemblages protéine-ADN excluant des protéines redondantes et ne comprenant que des ADN doubles brins [260]. Avec des aires d'interfaces variant de 618 à 2832 Å², les auteurs montrèrent que les sites de liaison pour les ADN recouvrent de 5 à 33% de la surface protéique cible. Les résidus des sites de liaison pour les ADN sont beaucoup plus polaires et chargés que les autres sites de liaisons connus, ce qui se traduit notamment par un plus grand nombre de liaisons hydrogènes à l'interface.

Cette augmentation des résidus polaires aux interfaces protéine-ADN se traduit également par une augmentation du nombre de molécules d'eau médiant les interactions polaires entre les atomes polaires de l'interface. Une analyse plus poussée de la composition sera présentée en résumé de cette partie sur les interactions moléculaires.

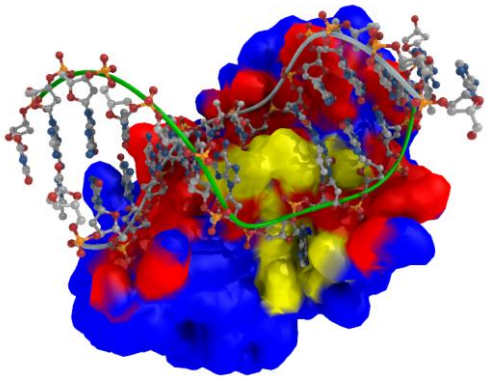
Les protéines peuvent lier l'ADN soit au niveau du petit sillon ou du grand sillon, soit au niveau du petit sillon et du grand sillon. Si le jeu de 26 assemblages protéine-ADN de Jones et coll. est représentatif de la réalité biologique, alors on peut en conclure que les protéines reconnaissent préférentiellement les grands sillons (15/26), ainsi que l'ensemble petit sillon et grand sillon (8/26). Le petit sillon semble ne pas suffire à la reconnaissance protéine-ADN sauf dans de quelques rares exceptions comme la protéine de liaison à la boîte TATA (1ytb) où le sillon est plus ouvert et permet un meilleur accès aux bases azotées.

Propriété	**Protéine-ADN**	**Hétérodimères permanents**	**Hétérodimères transitoires**
Aire d'interface enfouie sur l'une des chaînes protéiques (Å²)	1528	1722	804
Liaisons hydrogènes / 100 Å²	1,35	0,7	1.1
Molécules d'eau médiant des interactions / 100 Å²	0.7	0.3	0.4
Pourcentage d'atomes polaires	46,7	35,1	41,2

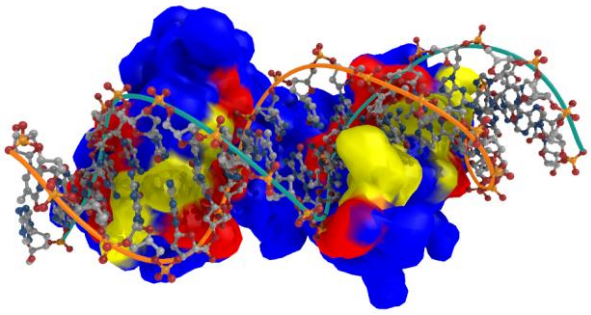
Tableau 18. Différences entre les interfaces protéine-ADN et protéine-protéine. *(source : S.Jones et coll., 1999 [260]).*

En considérant que la spécificité de reconnaissance d'un ADN par une protéine se joue au niveau des bases azotées de l'ADN, trois grands modes de liaisons peuvent être définis : (1) les résidus de la protéine qui interagissent avec les bases de l'ADN sont regroupés dans un *cluster* ; (2) les résidus de la protéine qui interagissent avec les bases de l'ADN sont regroupés dans deux *clusters* ; (3) la protéine enveloppe l'ADN créant ainsi un micro-environnement propice notamment aux réactions enzymatiques (Figure 53).

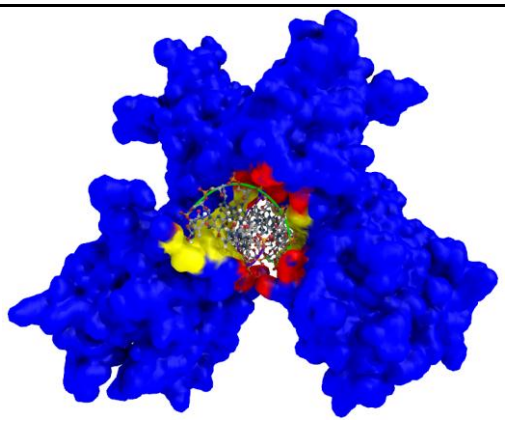
Si les résidus servant à la reconnaissance des sucres et des phosphates de l'ADN sont vraisemblablement très importants pour la stabilité de l'assemblage, il est probable qu'ils contribuent moins à la spécificité de la reconnaissance protéine-ADN que ne le font les résidus liant directement les bases azotées (Figure 54, Tableau 19).



Une région de reconnaissance des bases de l'ADN (1vas)



Deux régions de reconnaissance de l'ADN (1lmb)



La protéine enveloppe l'ADN et crée un micro-environnement (1nfk)

Figure 53. Trois grands modes de liaison de l'ADN par les protéines. *En rouge sont indiqués les atomes des protéines qui sont en contact avec les sucres ou phosphates de l'ADN. En jaune sont indiqués les atomes des protéines qui sont en contact avec les bases nucléiques.*

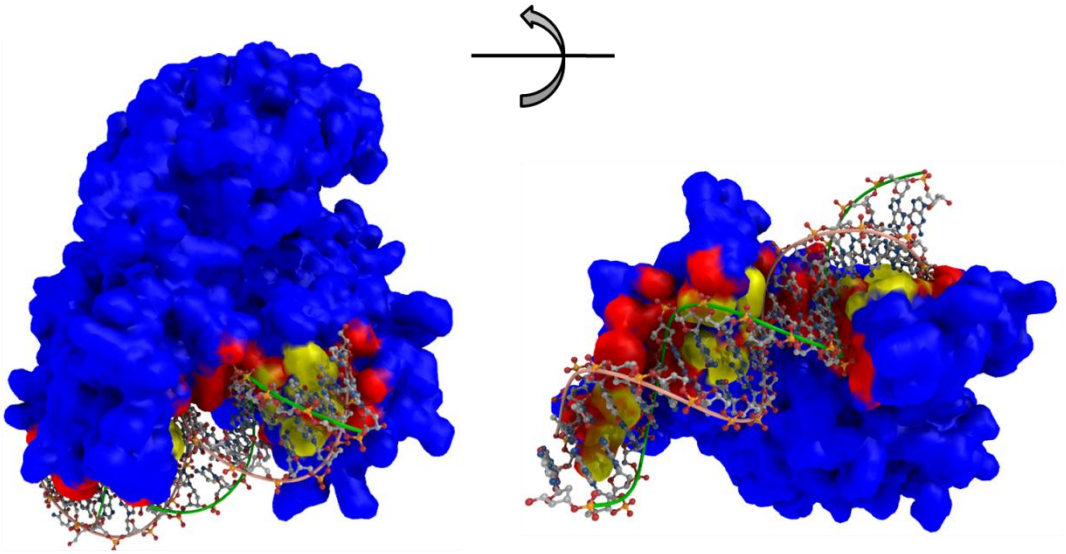


Figure 54. Détection des résidus protéiques impliqués dans la reconnaissance spécifique d'un segment d'ADN. *La structure complète de l'hétérodimère de PPARγ et de RXRα comporte un domaine de liaison à l'ADN (DBD) représenté ici en rouge pour les résidus interagissant avec les sucres et phosphates, et en jaune, pour les résidus interagissant avec les bases azotées.*

RXRα	ADN chaîne C	ADN chaîne F
Glu153		A4004, C4005
Gly154		G4003
Lys156	A3014, G3015	
Gly157		G4003
Lys160	G3016, T3017	
Arg161		T4002, G4003

PPARγ	ADN chaîne C	ADN chaîne F
Glu129		A4011, C4012
Gly130		G4010
Lys132	A3007, G3008	C4012
Arg136	G3008, G3009, T3010	
Arg137		T4009, G4010
Phe182		G4016, T4017
Gly183	A3004, C3005	T4017, T4018
Arg184	A3002, A3003, A3004	T4018, T4019

Tableau 19. Liste des résidus de RXRα et de PPARγ impliqués dans la reconnaissance des bases azotées de l'ADN.

Du point de vue conformationnelle, la liaison d'une protéine peut provoquer des remaniements de l'ADN, notamment au niveau des petits et grands sillons [260]. Ces changements conformationnels peuvent permettre un meilleur accès aux bases azotées de l'ADN et jouer un rôle important dans la reconnaissance et la régulation spécifique de l'expression des gènes.

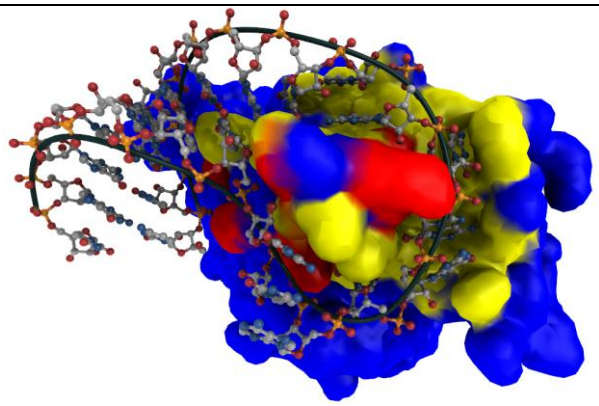
f. Protéine-ARN

L'ARN est par nature une molécule plus dynamique et plus instable que l'ADN, notamment en raison du groupe 2'OH du ribose absent dans le désoxyribose. En 2000, seules 200 structures étaient disponibles, et en 2010, la PDB n'en recense qu'approximativement 800. En revanche, l'interaction avec des protéines stabilise généralement les acides nucléiques, ce qui explique la présence de 2700 structures d'assemblages protéine-acide nucléique.

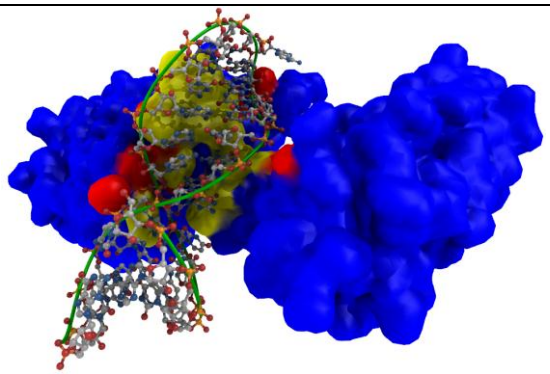
En 2008, Bahadur et coll. constituent un jeu non redondant de 81 assemblages transitoires protéine-ARN [45]. Par transitoire, il faut comprendre que les protéines et ARN interagissants sont capables de s'associer et de se dissocier dans le temps. Les molécules d'ARN peuvent également être différenciées selon qu'elles forment ou non une structure de double hélice. Une structure de double hélice peut être obtenue soit par l'appariement de

deux brins d'ARN, soit par le repliement de l'ARN sur lui-même. Appelons un « duplex d'ARN », un tel ARN capable de mimer la structure d'une double hélice (Figure 55).

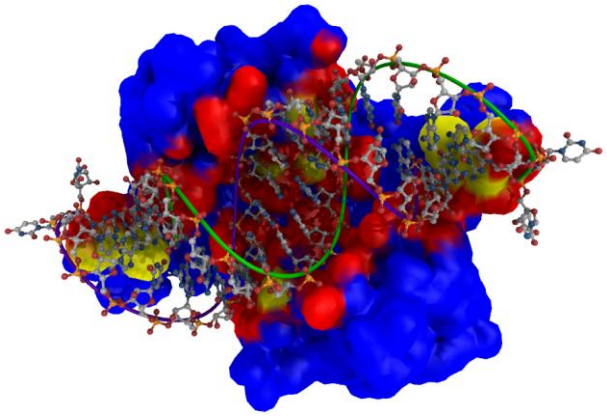
Les ARN simple brin (ou monocaténaire) forment des interfaces plus petites (1890 ∓ 880 Å²) que celles formées par les duplexes d'ARN (2630 ∓ 1270 Å²). En revanche, les nucléotides de l'ARN contribuent en moyenne davantage à l'aire d'interface (106 Å² / nucléotide) que les nucléotides des duplexes d'ARN (75 Å² / nucléotide). De plus, les interfaces protéine-ARN sont plus densément hydratées (12.6 molécules d'eau / 1000 Å²) que les interfaces protéine-ADN (6.7 molécules d'eau / 1000 Å²). L'étude de la fraction des atomes enfouis (*fbu*) indique également que les interfaces protéine-acide nucléique sont dans l'ensemble moins compactes que dans les interfaces protéine-protéine. Notons cependant que les interfaces protéine-acide nucléique participent à davantage d'interactions électrostatiques longues distances, ce qui pourrait biaiser le calcul de *fbu*.



A) Assemblage protéine-ARN simple brin (1a9n)



B) Assemblage protéine-ARN simple brin formant une hélice (1ooa)



C) Assemblage protéine-ARN double brin (1rpu)

Figure 55. Trois types d'assemblages protéine-ARN. *Lorsque l'ARN est simple brin, il peut adopter une très grande variété de conformations. En A), l'ARN simple brin s'enroule autour de la protéine. En B), l'ARN simple brin s'enroule sur lui-même, formant une structure qui mime la double hélice. En C), l'ARN double brin mime également la structure de double hélice.*
Comme pour la précédente figure sur les interactions protéine-ADN, la surface en rouge représente les résidus protéiques en contact avec les phosphates ou le sucre de l'ARN, alors qu'en jaune sont indiqués les résidus en contact avec les bases azotées.

De façon notable, l'aire d'interface des assemblages protéine-acide nucléique est dans l'ensemble davantage contribuée par les ARN (1320 Å²) que par les protéines (1210 Å²), une raison possible étant la surreprésentation de formes convexes dans l'ARN (Tableau 20).

L'étude comparative des interactions protéine-ADN et protéine-ARN montre que les groupements phosphates de l'ARN contribuent moins à l'interface (26%) que ceux des ADN (43%). Le ribose de l'ARN contribue en revanche davantage (39%) aux interfaces que le désoxyribose de l'ADN (29%), et son groupement 2'OH contribue à lui seul à 25% des liaisons hydrogènes avec les partenaires protéiques, et à 29% des contacts avec les molécules d'eau. Egalement, les bases azotées participent dans l'ensemble davantage aux interfaces protéine-ARN (35%) qu'aux interfaces protéine-ADN 27%) [45].

	Protéine - ARN	**Protéine - ARN double brins**	**Protéine - ARN simple brin**	**Protéine - ADN**	**Hétérodimère protéique**
Aire d'interface (Å²)	2530	2630	2890	3100	1910
Aire d'interface sur les protéines (Å²)	1210	1270	880	1540	-
Aire d'interface sur les acides nucléiques (Å²)	1320	1360	1010	1560	-
Aire d'interface par nucléotide (Å²)	75	75	106	72	-
Fraction d'atomes enfouis (fbu)	29	41	32	24	34
Molécules d'eau par 1000 Å² d'interface	12,6	-	-	6,7	10

Tableau 20. Comparaison des ARN simple et double brins avec l'ADN et les hétérodimères protéiques. *(adapté à partir de Bahadur et coll., 2008 [45]).*

Etant donné l'importance grandissante des ARN dans l'étude du vivant (découvertes des différentes variétés d'ARN), de leurs rôles enzymatiques (ribosymes) et de leurs rôles en tant que facteurs de régulation de l'expression des gènes (siARN), une meilleure description géométrique et physico-chimique de ces molécules sera un véritable défi à relever. Une connaissance plus approfondie de leurs structures, de leurs dynamiques conformationnelles et de leurs mécanismes d'interaction permettra de mieux modéliser l'environnement cellulaire.

g. Résumé des caractéristiques des différents types d'interactions moléculaires

Sans nul doute, les interfaces les plus chargées sont celles entre les protéines et acides nucléiques, avec en moyenne 38% de leurs aires composées de résidus cationiques (en raison des interactions avec les groupements phosphates des acides nucléiques), et 6% constituées de résidus anioniques (légèrement défavorisés également en raison des phosphates chargés négativement des acides nucléiques). Suivent les interfaces cristallines 34,6%), les interfaces d'hétérodimères transitoires (31,9%) et les interfaces d'homodimères (24,2%) (Tableau 21).

Les interfaces les plus hydrophobes sont celles des homodimères (39,1%), d'autant plus si l'on considère les interfaces spécifiques à la formation d'homodimères où le pourcentage de résidus hydrophobes impliqués dans l'aire d'interface monte à 66,5% [1]. Les interfaces protéine-acide nucléique sont quant à elles moins hydrophobes en raison de la présence de nombreuses charges totales (cationique et anionique) et partielles (polaires). Les variations de l'aromaticité des interfaces sont plus sensibles à analyser. Ainsi, si les différences d'aromaticité entre les interfaces cristallines (12,7%) et les interfaces d'hétérodimères (21,5%) et d'homodimères (18,5%) sont très marquées, ces différences sont à peine notables pour les assemblages protéine-ADN (11,5%), et protéine-ARN (14,7%). Aussi, l'aromaticité peut être considérée comme un marqueur des interfaces d'hétérodimères et des homodimères, mais pas nécessairement des autres interfaces. De la même façon, l'hydrophobicité (resp. la présence de charges cationiques) peut être considérée comme un marqueur des interfaces d'homodimères (resp. des interfaces protéine-acide nucléique), mais pas nécessairement des autres interfaces.

Notons également que les résidus du cœur et du bord de l'interface diffèrent de par leur composition ; notamment, le cœur de l'interface est plus aromatique et moins chargé que le bord, que ce soit pour les homodimères ou les hétérodimères (Tableau 22).

Résidu	Interface cristalline	Interface Protéine-ARN	Interface Protéine-ADN	Interface d'hétérodimère	Interface d'homodimère
Ala	4,1	3,1	3,4	2,8	5,2
Arg	9,9	20,6	23,8	10,1	9,7
Asn	5,5	6,0	6,3	5,7	3,8
Asp	6,0	3,8	1,6	5,1	4,3
Cys	0,5	0,3	0,8	1,7	1,2
Gln	6,7	4,7	5,1	4,3	4,1
Glu	9,7	4,2	2,5	6	5,6
Gly	4,5	4,7	3,6	4,8	3,5
His	2,9	4,2	3,8	3,8	3,3
Ile	3,3	2,9	2,8	4,6	5,5
Leu	6,4	3,9	2,4	5,7	10,4
Lys	9,0	14,0	17,5	6,5	5,4
Met	2,2	1,9	1,2	3,2	4
Phe	2,7	3,6	3,8	4,1	6,3
Pro	5,5	3,6	2,2	3,6	4,9
Ser	4,9	4,6	6,3	5,4	4
Thr	4,9	4,2	6,7	5	4,3
Trp	1,5	1,9	0,5	4,2	2,6
Tyr	5,6	5,0	3,4	9,4	6,3
Val	4,3	2,7	2,4	3,8	5,6
Hydrophobe	30,3	22,8	18,0	28,5	39,1
Aromatique	12,7	14,7	11,5	21,5	18,5
Polaire	22,0	19,5	24,4	20,4	16,2
Chargé	34,6	42,6	45,4	27,7	25,0
Cationique	18,9	34,6	41,3	16,6	15,1
Anionique	15,7	8,0	4,1	11,1	9,9

Tableau 21. Récapitulatif des caractéristiques de différents types d'interfaces moléculaires. *La contribution à l'aire d'interface des assemblages cristallins est dérivée de Bahadur et coll., 2004 [156]; celles des interfaces protéine-ARN et protéine-ADN dérivent respectivement de Bahadur et coll., 2008 [45] et de Nadassy et coll., 1999 [179]. Enfin, les contributions des résidus à l'aire d'interface des hétérodimères et des homodimères dérivent des études de Chakrabarti et coll., 2002 [154] et de Bahadur et coll., 2003 [155].*

	Hétérodimère		Homodimère	
	Coeur	*Bord*	*Cœur*	*Bord*
Hydrophobe	28,6	28,3	39,4	38,2
Aromatique	26,1	10,4	21,3	9,3
Polaire	19,2	23,8	15,2	19,2
Chargé	24,2	36,2	22,9	31,9

Tableau 22. Différences entre le cœur et le bord de l'interface pour les homo et hétérodimères. *Le cœur des interfaces d'homodimères est plus hydrophobe et moins aromatique et polaire que celui des hétérodimères.*

Nous verrons au cours de nos deux articles sur la surface [4] et la cartographie moléculaire [1] que les compositions des interfaces peuvent de plus varier en fonction de leur localisation sur la surface (dans une région concave ou convexe par exemple), mais également en fonction de la spécificité de reconnaissance du type de site de liaison. Ainsi, il existe des sites de liaison spécialisés dans la reconnaissance et la formation d'homodimères, qui diffèrent des sites de liaison polyvalents, capables de reconnaître différents types moléculaires. Notons également que l'étude par similarité des sites de liaison permet d'observer un nouveau paramètre qui représente la fréquence de contact ($f_{contact}$) d'un résidu à une interface donnée (Figure 56). Pour N interfaces homologues alignées, il est possible de déterminer la fréquence à laquelle un résidu à une position donnée est vu interagissant avec un partenaire molécule d'un type donné. Cette mesure renseigne donc sur la fréquence à laquelle un résidu est sollicité pour permettre un type d'interaction donné. Ainsi, il est possible de calculer la fréquence de contact d'un résidu pour former un homodimère, un hétérodimère, une interaction protéine-ligand, etc. Le fcontact est particulièrement utile comme nous le verrons dans la partie cartographie moléculaire, pour identifier les résidus hydratés de façon spécifiques ou transitoires.

Il est ainsi possible de distinguer les résidus qui participent toujours à un type d'interaction (exemple : homodimère, hétérodimère, protéine-ADN, …) et qui ne sont pas spécifiques du partenaire considéré. Inversement, il devient également possible de quantifier et de repérer les résidus qui servent à moduler la spécificité de la reconnaissance du partenaire, en observant les résidus qui sont moins fréquemment en contact pour un type d'interaction donné. Dans le cas des interactions protéine-acide nucléique, les résidus fréquemment interagissants pourraient par exemple corréler avec les résidus qui interagissent avec les groupements phosphates et les sucres, afin de répondre à une contrainte structurale et physico-chimique, indépendamment de la spécificité de reconnaissance de l'assemblage.

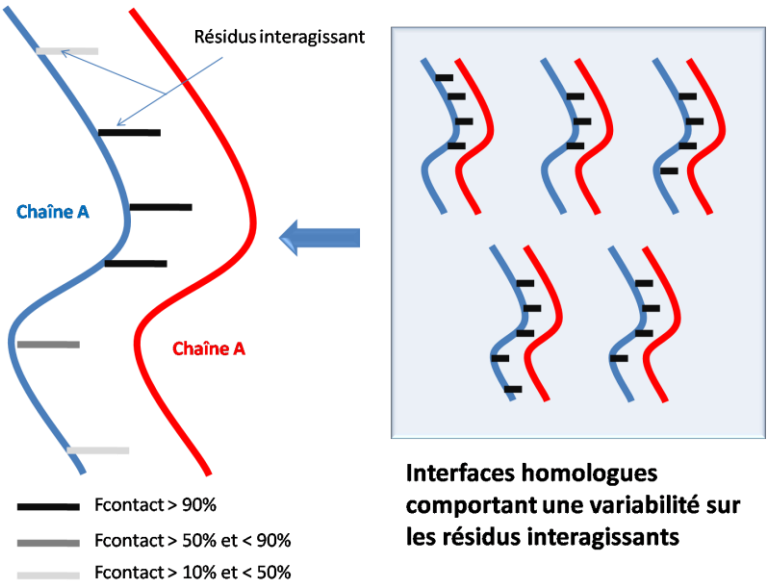


Figure 56. Représentation schématique du calcul et de la signification de la fréquence de contact $f_{contact}$ d'un résidu. *A gauche, deux sites de liaisons interagissants sont représentés par des lignes courbes en rouge et en bleu. Les barres représentent des résidus interagissants de la chaîne A. Plus la barre est noire et plus la fréquence de contact de ce résidu est importante ; inversement, plus elle est claire et plus la fréquence de contact du résidu est faible. La détermination de ces fréquences de contact pour chaque résidu se fait à partir de l'analyse des interactions de ce résidu dans des structures contenant des interfaces homologues (schématisées à droite).*

S'il existe des différences physico-chimiques claires entre les sites de liaison d'assemblages protéine-protéine et protéine-acide nucléique, il reste à déterminer si les différences observées entre les homodimères, hétérodimères, protéine-peptide et protéine-petit peptide sont dues à des contraintes différentes pour chacun de ces types d'interactions, ou bien si elles sont dues à l'aspect transitoire/permanent de ces interactions.

3. Prédiction des sites de liaisons

Ces nombreuses analyses des différents types d'interactions moléculaires ont déclenché un vif intérêt dans la communauté scientifique pour tenter de prédire *ab initio* les sites de liaisons à la surface des protéines. Ces prédictions se servent des différences géométriques et physico-chimiques observées entre, d'une part les sites de liaisons que l'on cherche à identifier, et d'autre part, la surface non-interagissante. Un premier problème tient donc dans l'identification de cette surface non-interagissante qui, à défaut de posséder davantage

de données fonctionnelles sur les protéines, correspond à l'ensemble des résidus de surface qui ne sont pas connus pour interagir. Un second problème tient dans le fait que pour que la prédiction soit pertinente, il est nécessaire que l'apprentissage se fasse sur les formes libres (structure de la protéine seule en solution, ou participant au plus à des contacts cristallins) et non sur les formes liées (structure de l'assemblage, avec de possibles remaniements du squelette et des chaînes latérales) des protéines [121].

La comparaison des compositions en acides aminés de la surface moléculaire permet de construire un diagramme renseignant sur la tendance de chaque acide aminé à faire partie des interfaces, plutôt que de la surface non-interagissante. Cette tendance est calculée comme le score RP présenté précédemment, c'est-à-dire, comme le logarithme de la fréquence du résidu aux interfaces sur la fréquence du résidu à la surface non-interagissante (Figure 57). Cette propriété est discriminante pour différencier les résidus de l'interface des résidus de la surface non-interagissante.

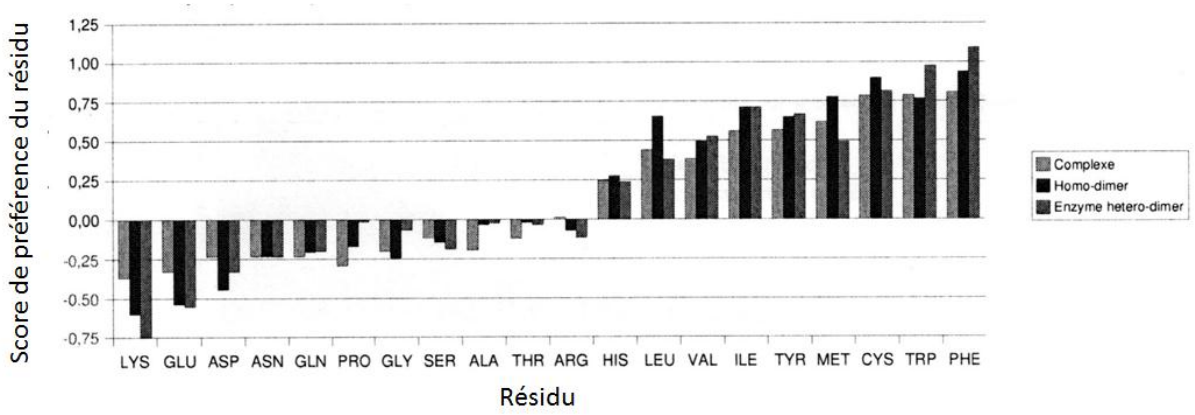


Figure 57. Tendance des acides aminés à faire partie des interfaces protéiques plutôt que de la surface non-interagissante. *Evalué sur trois jeux de données représentant des hétérodimères transitoires (complexe), des homodimères, et des hétérodimères enzymatiques (interactions permanentes), le diagramme montre que les acides aminés aromatiques et hydrophobes ont davantage tendance à faire partie du cœur de l'interface.*

a. Trois approches pour prédire les sites de liaison

Il existe principalement trois approches pour prédire les sites de liaisons : (1) les approches qui n'utilisent que les données de séquences, cherchant des biais de composition dans de petites sous-séquences [67, 114] ; (2) les approches qui utilisent les données de chaque structure prise séparément pour identifier les sites de liaison, par exemple par la formation de région ou *patch* de surface [113, 117, 272]; (3) les approches plus récentes qui tentent d'intégrer les données de plusieurs structures caractérisant une même protéine. Ces

dernières approches utilisent par exemple la présence de résidus structuralement conservés dans une famille [273], ou résolvent partiellement le problème de la surface non-interagissante en repérant par homologie de possibles interactions [119, 274]. Dans l'ensemble, ces approches ont un taux de précision de l'ordre de 75%, étant admis qu'au regard de l'imprécision commise en définissant la surface non-intéragissante, il est difficile de préciser l'exactitude de cette prédiction.

b. Description et analyse des surfaces moléculaires par les régions

L'analyse des structures moléculaires et la description des différences entre les différentes zones de la molécule requièrent la génération de régions tridimensionnelles qui peuvent être de deux types : (a) des régions qui reflètent directement une propriété telle que l'hydrophobicité [275], la conservation des acides aminés [276] ou la courbure de surface (pour détecter des crevasses) [237, 277] ; (b) des régions qui sont générées de façon arbitraire et exhaustive par des critères de tailles et de formes avec pour optique finale, la prédiction d'une propriété biologique.

Les approches de prédiction des sites de liaison de types 2 et 3 (définis dans la sous-section précédente) reposent sur la deuxième forme de réalisation (b), c'est-à-dire sur la génération arbitraire et exhaustive de régions tridimensionnelles. Par l'analyse de propriétés moyennées sur ces régions, on tente alors de distinguer les régions qui sont des sites de liaison, des régions appartenant à la surface non-interagissante. Par exemple, nous avons vu que les sites de liaison correspondant aux interactions homodimériques étaient plus hydrophobes que le restant de la surface, cela doit donc se refléter localement dans ces régions. Inversement, nous avons vu que les sites de liaison correspondant à des interactions protéine-acide nucléique sont plus cationiques que le restant de la surface non-interagissante ; la détection de tels sites de liaisons peut donc se faire par la détection de régions ayant localement une forte densité de résidus cationiques.

Ainsi, en 1997, Jones et coll. proposent une méthode de génération de régions tridimensionnelles basée sur le regroupement de N résidus de surface en fonction de vecteurs pointant vers le solvant (milieu extérieur à la protéine). Un résidu est alors défini comme faisant partie de la surface si son accessibilité est d'au moins 5% de son accessibilité maximale [229]. Pour chaque résidu de surface, on crée alors une région (*patch*) contenant les N plus proches résidus de surface discrétisés par leurs carbones alphas. Pour tenter de former des régions contiguës (s'assurer de la formation d'une région unique et non de plusieurs sous-régions qui ne se toucheraient pas nécessairement), les auteurs proposent de calculer pour chaque carbone alpha un vecteur pointant vers le solvant, et de ne

sélectionner que les résidus de surface pour lesquels l'angle entre les vecteurs solvants d'un résidu central et d'un quelconque résidu de la région doit être inférieur à 110°. Nous verrons par la suite que cette définition n'assure cependant pas la notion de contigüité, en particulier dans les régions en forme de crevasses, qui sont pourtant fonctionnellement importantes, notamment pour la liaison de ligands. Bien que les auteurs admettent que cette définition de région soit assez simpliste, elle fut reprise dans la plupart des études visant à prédire les sites de liaisons [113, 117] Une variante plus rapide et plus simple pour former une région, mais n'assurant pas non plus la notion de contigüité, consiste à sélectionner tous les atomes et/ou résidus de surface à l'intérieur d'une sphère centrée sur un résidu de surface ou un point de la surface accessible [121].

L'analyse de ces régions formées de façon arbitraire le long de la surface a toutefois permis de montrer que celles-ci pouvaient assez bien recouvrir les sites de liaisons : c'est-à-dire que l'une des régions générées de façon automatique pouvait comporter tout ou partie des atomes et résidus d'un site de liaison. En moyenne, Jones et coll. estiment que 15% des régions générées ainsi de façon systématique recouvrent plus de 70% d'un site de liaison. Dans notre étude de 2008, la génération de régions contenant 20 résidus montre que l'approche de Jones et coll. permet de trouver une région qui recouvre en moyenne 56% d'un site de liaison pour une structure sous forme libre (pour la protéine sans partenaire). En comparaison, notre approche de génération de régions avec de mêmes critères permet d'obtenir une région recouvrant en moyenne 62 à 66% d'un site de liaison [4]. Le fait de parvenir à générer en automatique des régions capables de recouvrir un site de liaison est une étape nécessaire pour assurer la détection d'un signal permettant d'identifier les propriétés remarquables qui permettent à une région de lier une molécule.

Parmi les propriétés qui peuvent être utilisées pour différencier les sites de liaisons des autres régions de surface, notons également un accroissement de 10 à 15% de l'accessibilité des résidus interagissants, observables dans les formes libres [4, 119], la présence de résidus structuralement conservés aux interfaces [273, 278], une conservation accrue des résidus interagissants en raison de la pression de sélection qui s'exerce sur eux [276, 279, 280]. Plus récemment, nous avons également mis en évidence que les résidus formant un site de liaison coévoluent davantage avec le cœur protéique qu'avec le restant de la surface. Une possible raison serait les phénomènes d'allostérie où la liaison d'un partenaire à un site de liaison de la protéine se traduit par des changements conformationnels d'autres régions éloignées de la protéine [281]. C'est le cas notamment de l'interface de dimérisation des récepteurs nucléaires avec la cavité liant les ligands qui a rapidement été illustrée à la figure 31.

c. *Conservation et coévolution : exemple de RXRα*

Les études basées sur la conservation ou la coévolution des résidus dépendent de la sélection de séquences partageant localement de mêmes fonctions (exemple : la capacité à lier une molécule précise). Cette sélection est une étape particulièrement difficile à atteindre, d'autant plus que nous observons les variations d'une protéine non pas au sein d'une même espèce (où l'on s'assurerait qu'elle partagerait de mêmes fonctions), mais au sein de plusieurs espèces plus ou moins éloignées du système étudié. Avec les nouvelles méthodes de séquençage et la volonté de mieux observer et comprendre la variabilité des individus d'une même espèce, de nouvelles données devraient permettre dans les années à venir de fournir des résultats plus fiables permettant de mieux observer à l'échelle du résidu, la pression de sélection et donc, de possibles fonctions.

En utilisant la formule de l'entropie de Shannon présentée dans la partie « Bioinformatique des séquences », il est possible de calculer un score de conservation pour chaque résidu d'une protéine ou d'un acide nucléique. La Figure 58 corrèle la conservation des résidus de surface avec la localisation du site de dimérisation de RXRα.

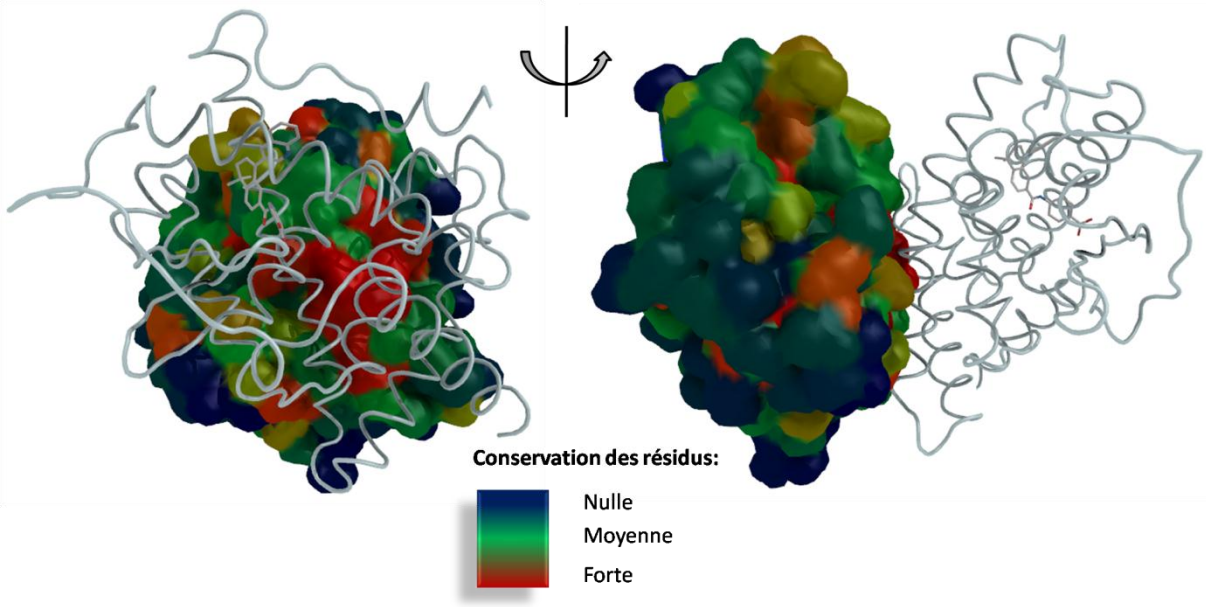


Figure 58. Conservation des résidus de surface du récepteur RXRα. *Un alignement multiple de séquences a été réalisé grâce au programme PipeAlign de l'IGBMC [282]. A partir de cet alignement multiple de séquences, qui aligne les résidus correspondants dans plusieurs protéines homologues, il est possible d'observer et de quantifier les variations pour établir un score de conservation représenté ici par un code couleur. On remarque que le site de liaison au récepteur RAR contient effectivement des résidus très conservés.*

Résidu à l'interface de RXR-RAR	Conservation	Accessibilité (Å²)	Accessibilité relative (%) (a)	Aire enfouie à l'interface (b)	Ratio (b) / (a)
ARG:353	0,43	53	22	22	0,99
THR:356	0,22	78	54	4	0,07
GLU:357	**0,95**	**51**	**28**	**44**	**1,60**
LYS:361	0,91	89	42	55	1,30
ASP:364	0,13	110	73	1	0,02
ASP:384	0,64	120	80	46	0,58
GLU:395	0,55	78	42	31	0,72
ARG:398	0,50	48	20	9	0,44
GLU:399	0,87	87	47	55	1,16
TYR:402	**0,40**	**67**	**29**	**62**	**2,11**
GLU:406	0,48	58	32	40	1,28
LYS:410	0,37	119	56	13	0,23
PRO:417	0,21	115	81	33	0,41
GLY:418	0,37	40	47	22	0,48
PHE:420	**0,95**	**31**	**14**	**27**	**1,88**
ALA:421	0,47	54	48	50	1,04
LYS:422	0,41	105	50	51	1,03
LEU:424	**0,95**	**43**	**24**	**43**	**1,80**
LEU:425	**1,00**	**109**	**61**	**90**	**1,49**
ARG:426	1,00	37	15	16	1,03
LEU:427	0,95	33	19	22	1,16
PRO:428	1,00	79	55	76	1,39
ALA:429	0,46	16	14	8	0,56
ARG:431	**0,90**	**104**	**43**	**64**	**1,48**
SER:432	0,45	46	37	23	0,62
LEU:435	0,46	96	53	37	0,70
GLU:439	0,55	113	62	11	0,18
Conservation moyenne	0,62		**Aire d'interface pour RXRα (Å²)**	955	

Tableau 23 : Accessibilités, aires enfouies et conservations des résidus de RXRα à l'interface RXRα - RAR. *Le score de conservation déterminé à partir de l'entropie de Shannon est normalisé entre 0 et 1, plus la valeur est proche de 0 (resp. 1) et moins le résidu est conservé (resp. plus le résidu est conservé). L'accessibilité relative du résidu correspond au pourcentage de son accessibilité maximale ; elle indique si le résidu est ou non très accessible à la surface. Le ratio (b) / (a) indique quant à lui l'implication d'un résidu à l'interface en fonction de son accessibilité relative. Notons que la quasi-totalité de ces résidus sont effectivement très conservés.*

De plus, il est important de noter que la pression de sélection que l'on observe au travers de la conservation des résidus reflète non seulement la fonctionnalité de certaines régions, mais également les résidus nécessaires à la stabilité de la protéine [283]. Si les résidus impliqués dans la stabilité de la protéine font majoritairement partie du cœur protéique, il en est toutefois qui font partie de la surface et doivent être différenciés des résidus conservés en raison de la présence d'un site de liaison.

En générant des régions de la surface moléculaire et en calculant la conservation relative des résidus formant chaque région, ainsi que l'exposition relative des chaînes latérales des acides aminés, nous avons conçu un arbre de décisions simple qui permet de différencier les résidus interagissants des résidus non-interagissants. La conservation des résidus a été calculée à partir de l'entropie de Shannon détaillée précédemment (prenant en compte les 20 acides aminés), et à partir d'alignements multiples de séquences déterminés par le programme PipeAlign de l'IGBMC [282]. Evalué par une méthode de 10-validation croisée sur 12 000 résidus (6000 résidus identifiés comme faisant partie du cœur d'interfaces, et 6000 résidus choisis aléatoirement parmi les résidus de surface supposés non-interagissants), nous parvenons à prédire ces deux classes de résidus avec une précision de 74,9% (Figure 59). Bien qu'il eût été possible de définir d'autres arbres de décisions offrant une précision quelque peu plus avantageuse (quelques pourcents), nous avons volontairement choisi l'arbre le plus simple, c'est-à-dire, utilisant le moins de propriétés possibles, de sorte que sa valeur prédictive soit plus grande.

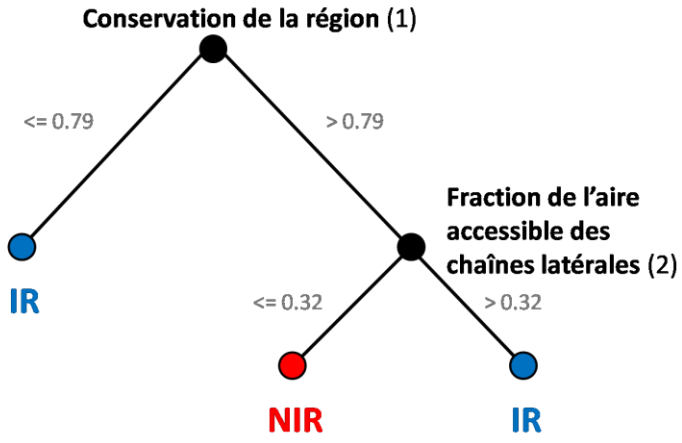


Figure 59. Arbre de décisions permettant de prédire les sites de liaison. *Le score de conservation de la région (2) correspond à l'entropie de Shannon moyennée par résidu ; la fraction de l'aire accessible des chaînes latérales (1) donne une indication de l'exposition relative des chaînes latérales au solvant. IR signifie « Interacting Residue » et NIR « Non Interacting Residue ». Cet arbre de décision très simple permet de prédire avec une précision de 74,9% les résidus du cœur de l'interface.*

De la même façon, les coévolutions observées entre deux résidus d'une même protéine peuvent avoir une signification fonctionnelle (comme la coopérativité entre sites, l'allostérie), ou bien peuvent refléter une interaction nécessaire au repliement de la protéine. C'est l'une des raisons, avec la dynamique conformationnelle des protéines, pour laquelle deux résidus éloignés dans l'espace peuvent être détectés comme coévoluant :

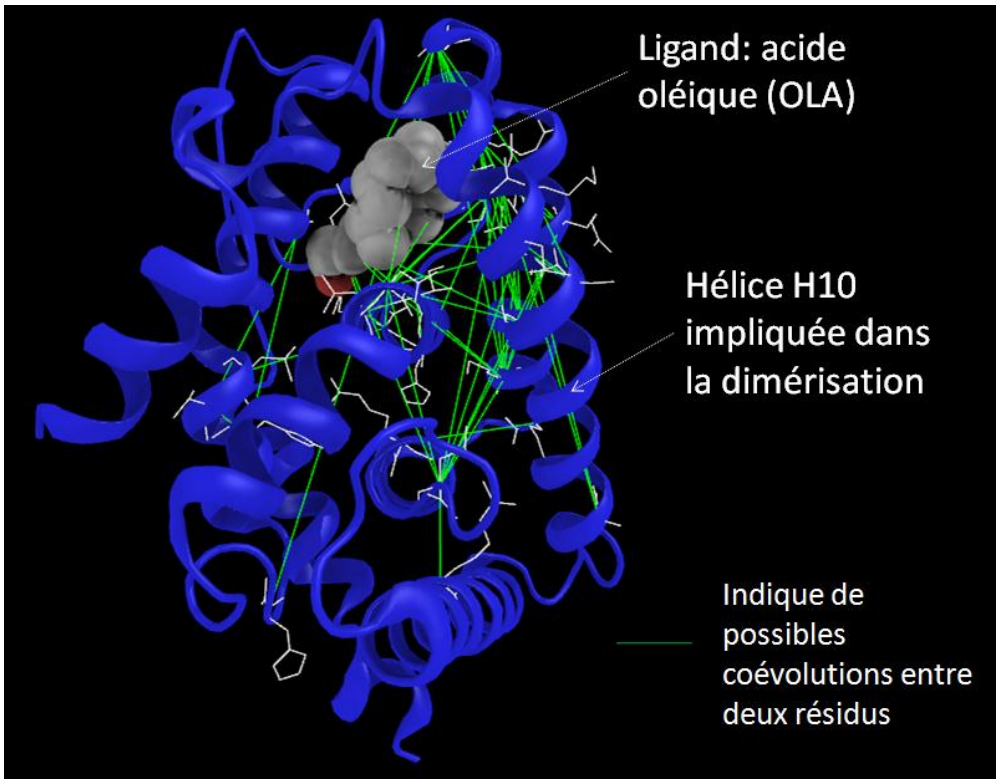


Figure 60. De possibles coévolutions entre les résidus impliqués dans la liaison du ligand et les résidus impliqués dans la dimérisation de RXRα. *Les coévolutions indiquées par des lignes vertes représentent des couples de résidus pour lesquels lorsque l'un des deux résidus mute, on observe fréquemment une mutation corrélée à une autre position.*

d. Problèmes liés à la génération de régions par les méthodes usuelles

Toutefois, cette définition présente deux problèmes : (1) les vecteurs « solvants » ne pointent pas nécessairement vers le milieu extérieur, en fonction de la forme locale de la surface ; (2) le critère visant à former des régions contiguës peut malgré tout générer des régions discontinues (formées de plusieurs sous-régions séparées dans l'espace).

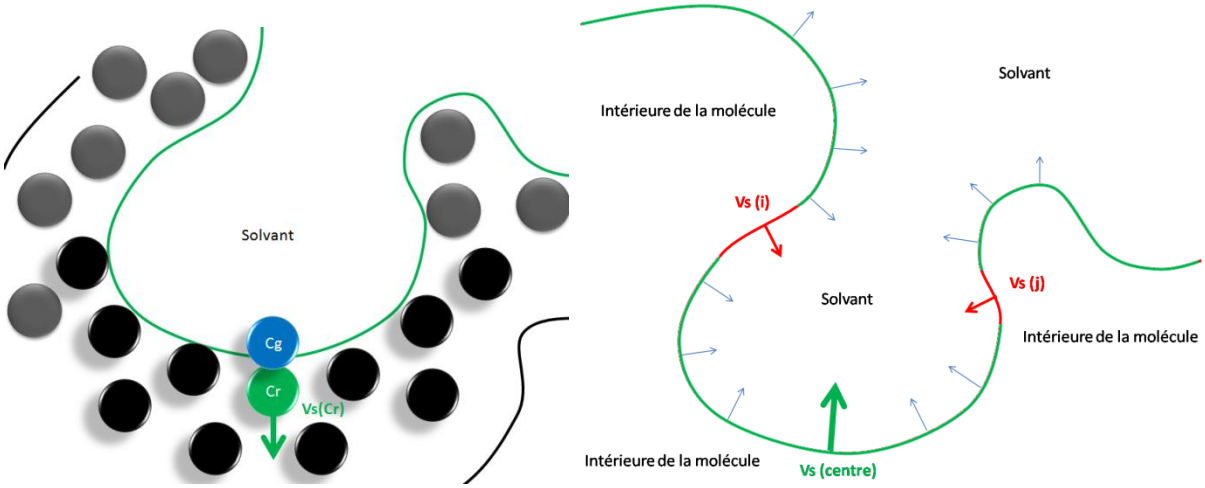


Figure 61. Problèmes liés aux approches usuelles de génération de régions. *A gauche, on cherche à déterminer le vecteur solvant Vs du résidu de surface Cr qui appartient à la surface représentée en vert. Pour cela, l'approche de Jones et coll. propose de sélectionner les 10 carbones alphas les plus proches (ici en noir foncé) et de déterminer leur centre de gravité Cg (ici en bleu). Le vecteur solvant (Cg, Cr) devrait alors pointer vers le solvant, toutefois dans notre exemple, on observe qu'il est possible qu'il pointe vers l'intérieur de la molécule. A droite, l'approche de Jones et coll. consiste alors à sélectionner les N carbones alphas les plus proches tels que l'angle formé par leurs vecteurs solvants Vs soit inférieur à 110°. Dans cet exemple, les vecteurs solvants des carbones alphas i et j forment des angles avec le vecteur solvant du centre de la région en formation, supérieur à 110° : ils ne seront donc pas sélectionnés dans la région, même s'ils sont contigus et proches du centre de la région. Ainsi, on a représenté en vert la surface pouvant être sélectionnée dans la région à partir du centre, et en rouge, la surface (et les atomes/résidus) ne pouvant être sélectionnée.*

Cette problématique liée à la génération de régions contiguës est traitée dans l'article de 2008 où nous proposons une méthode capable de générer des régions contiguës de façon rigoureuse et rapide, par extension de la région depuis un point de surface.

e. Problèmes liés à la prédiction des sites de liaison

Toutefois, et comme nous le verrons dans notre article sur la cartographie moléculaire [1], les approches de prédiction des sites de liaisons souffrent de deux problèmes techniques et d'un problème théorique:

(1) les approches les plus récentes qui réalisent une analyse intégrative des données structurales ne différencient pas les différents types d'interactions et sites de

liaisons. Le plus souvent, ils ne prédisent que les sites de liaisons pour les interactions protéine-protéine.

(2) de plus, le critère d'homologie basé sur la comparaison des structures complètes n'assure pas que les sites de liaisons (des régions locales de ces structures) soient fonctionnellement identiques. Le changement d'un seul résidu a été rapporté dans la littérature comme pouvant empêcher des interactions : par exemple, la mutation de R426A du récepteur nucléaire RXRα affecte son homodimérisation et la mutation de Y402A affecte son hétérodimérisation avec RAR [212].

(3) si, comme nous l'avons montré, la surface interagissante des protéines (en considérant tous les types d'interactions) est bien de l'ordre de 44%, l'utilité de telles prédictions est réduite, que ce soit pour guider les expériences de mutagénèse ou d'amarrage moléculaire. De plus, la prédiction des sites de liaison ne donne aucune information sur de possibles partenaires moléculaires. Aussi, il semble raisonnable d'envisager d'autres approches i) pour apporter davantage de précisions (par exemple, la fréquence de contact des résidus interagissants, ou la détection des points chauds ou « hot spots »), ii) pour préciser également les partenaires moléculaires.

IV. Le criblage virtuel : vers une nouvelle ère du développement de médicaments

« Le microbe n'est rien. Le terrain est tout »
Louis Pasteur, XIX^{ème} siècle

Dans cette première partie, nous discuterons de la révolution qui s'opère lentement dans la discipline de recherche de médicaments, et qui consiste notamment à intégrer l'analyse des données structurales pour concevoir des médicaments. Nous parlerons de l'apport de la bioinformatique et de la chemoinformatique, à la fois pour modéliser les cibles thérapeutiques, et caractériser leurs interactions moléculaires, mais aussi pour déterminer des composés capables de lier ces cibles. Dans un troisième temps, nous ferons une ouverture sur les dernières révolutions de la discipline qui tentent d'identifier de nouvelles cibles cellulaires à des composés par l'analyse de leurs effets secondaires et/ou des similarités de structures des composés. Pour conclure cette partie, nous présenterons succintement la nouvelle forme de criblage virtuel que nous avons développée, basée sur la comparaison des régions moléculaires, et préciserons certaines de ses applications pour la biologie et le développement de médicaments.

---

1. Le développement de médicaments, une science interdisciplinaire en révolution

La recherche de médicaments (*drug research*) est une science interdisciplinaire et industrielle qui n'a pas plus d'un siècle. Elle repose sur les travaux et interprétations de scientifiques tels que L.C. Virchow, L. Pasteur, C. Bernard, A. Fleming, P. Ehrlich, J.N. Langley et L. Pauling, qui contribuèrent à réunir la médecine clinique, la biologie, la chimie, la pharmacologie et la biologie moléculaire. Ainsi, si Virchow et Pasteur démontrèrent respectivement l'existence de pathologies cellulaires, et les phénomènes d'infections, Fleming découvrit la pénicilline en 1929, antibiotique de prédilection en raison de son efficacité et de sa faible toxicité. Les travaux de Ehrlich puis de Langley en 1905 menèrent à postuler l'existence de « chemorécepteurs », des protéines qui reçoivent et transmettent des signaux spécifiques dans la cellule, et dont l'activité peut être modulée par de petites molécules antagonistes ou agonistes [284]. Les travaux de L. Pauling, démontrèrent quant à eux, l'origine moléculaire de certaines maladies en se fondant sur l'anémie falciforme [20].

La synergie de ces différentes disciplines a eu pour effet, à partir de la seconde moitié du XX^{ème} siècle, de ne plus seulement créer des médicaments à partir des observations empiriques limitées et de l'extraction (isolation et purification) des principes actifs par la chimie analytique (comme l'identification de la morphine par W. Sertürner en 1815), mais d'apporter un dialogue entre chimistes et biologistes permettant de mieux comprendre les mécanismes moléculaires des maladies. Cette meilleure compréhension des maladies, notamment par la génomique, le séquençage haut débit, l'identification de gènes impliqués dans des pathologies et les essais cellulaires, a profondément influencé le développement de médicaments en aidant à la détermination de cibles cellulaires capables de réguler différentes pathologies. On notera par exemple le rôle des récepteurs à la leptine pour le traitement de l'obésité découvert par Haffner et coll. en 1998 [285, 286], ou des interleukines 4 pour le traitement d'allergies [287].

La « crédibilité » et la pertinence de telles cibles sont cependant difficiles à établir en raison de la diversité et de la complexité du fonctionnement d'une cellule, et encore plus d'un organisme multicellulaire. Aussi, la détermination de ces cibles capables de réguler le développement de pathologies reste un problème ouvert, notamment parce que : (1) les produits des gènes liés à des maladies ne sont pas forcément les bonnes cibles thérapeutiques (elles doivent présenter un site de liaison capable de lier une molécule administrable, et cette liaison doit pouvoir moduler les fonctions de la protéine) [288-291] ; (2) il existe à l'heure actuelle un véritable fossé entre l'action d'une petite molécule *in vitro* et *in vivo* (dans une cellule). Encore de nos jours, on crée le plus souvent un candidat

médicament pour une cible, sans considérer les interférences qu'il pourrait avoir avec d'autres systèmes biologiques ; et (3) il existe un certain polymorphisme de ces cibles pour les individus d'une même espèce, ce qui signifie qu'il est nécessaire de repérer quelles sont les formes alléliques et épissées les plus représentées, et de vérifier si ces protéines sont effectivement capables de moduler le développement de la pathologie [19].

La détermination des structures de ces cibles par la biologie structurale et les nouvelles capacités de traitements informatiques et bioinformatiques, plus particulièrement la modélisation de cibles, l'amarrage moléculaire et le criblage virtuel, constituent une nouvelle révolution dans la conception et le développement de médicaments. Ensemble, ces deux disciplines ont pour but d'une part, de fournir les caractéristiques géométriques et physico-chimiques essentielles de la cible, et d'autre part, de modéliser et visualiser les sites de liaisons, et de prédire les différentes interactions possibles entre différents ligands et une cible (amarrage moléculaire et criblage virtuel) [192, 193, 292-295].

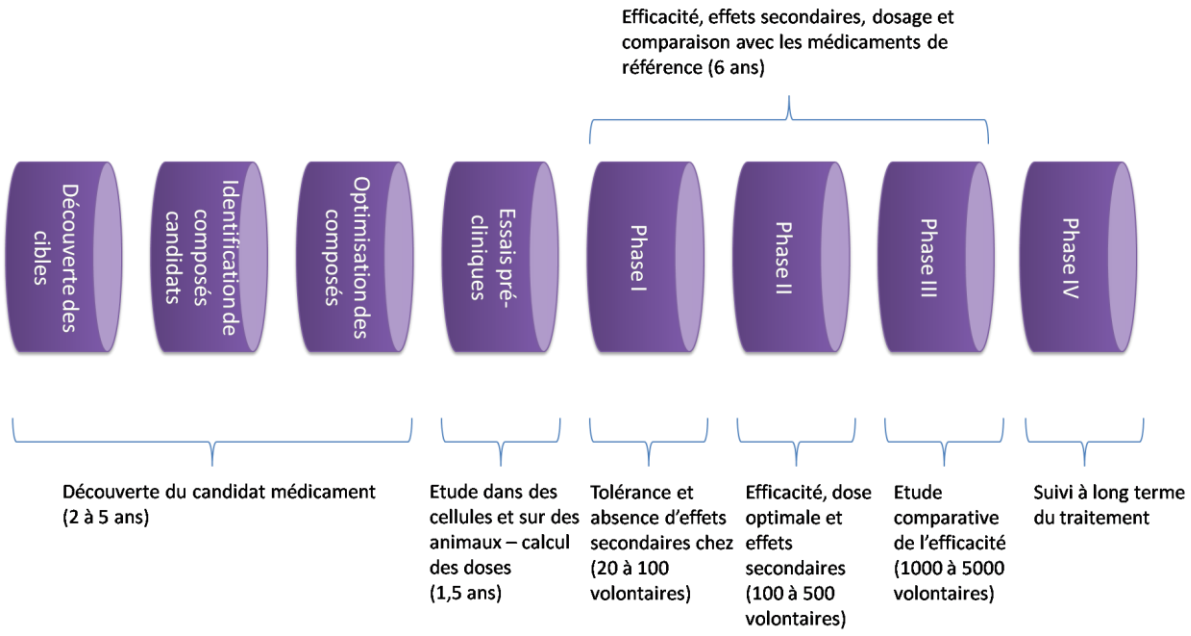


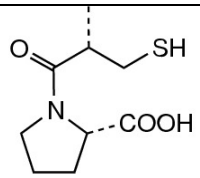
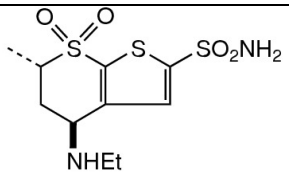
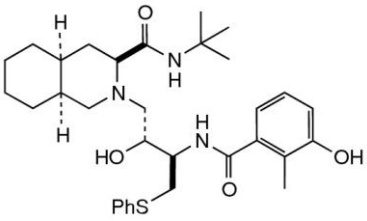
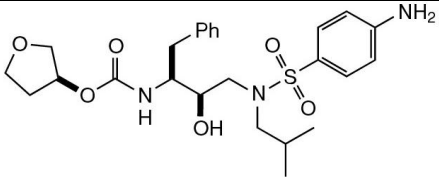
Figure 62. Les différentes étapes de la découverte et du développement de médicaments. *Les étapes auxquelles nous nous intéresserons concernent principalement les trois premières, c'est-à-dire, la découverte des cibles et de composés spécifiques. Toutefois, comme nous le verrons par la suite, les données structurales et le criblage virtuel peuvent être utilisés pour aider à l'évaluation de l'efficacité et des effets secondaires dans la cellule (essais pré-cliniques).*

2. Les données structurales dans le développement de médicaments

« Nowadays, molecular modelling plays a most important role in rational drug research. Structure-based and computer-aided design methods are applied to perform rational lead discovery and structural optimization with the least effort. »

Kubinyi, Success Stories in Computer-Aided Design, 2006

La biologie structurale a permis d'élucider la structure tridimensionnelle d'un grand nombre de cibles thérapeutiques, y compris de protéines membranaires telles que les GPCR [296, 297]. Grâce aux efforts internationaux dans les projets de génomiques structurales, certaines étapes dans l'expression et la cristallisation ont pu être automatisées, comme les robots nano gouttes (*drops*) qui ne requièrent que 25 à 100nl de protéine pour tester une condition de cristallisation [298]. Un bon exemple de structure difficile à obtenir est celle du cytochrome P450s, une protéine associée à la membrane dont l'isoforme P450 3A4 est responsable du métabolisme primaire de près de 50% des médicaments dans le corps humain [186]. Sa structure a été résolue dans les années 2000 [299] puis ré-étudiée avec différents ligands [300, 301].

Nom	Structure	Maladie	Cible	Date
Captopril (capoten)		Hypertension	Angiotensin-converting enzyme (1uzf)	1981
Dorzolamide (trusopt)		Glaucome	Anhydrase carbonique (3f3w)	1985
Nelfinavir (viracept)		HIV	Protéase HIV (3ekx)	1999
Amprenavir (agenarase)		HIV	Protéase HIV (3ekv)	1999

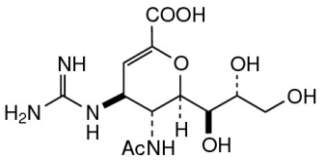
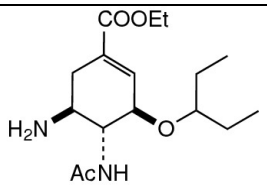
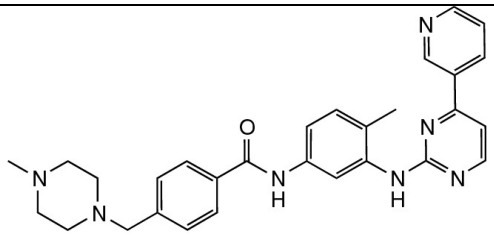
Zanamivir (Relanza)		Influenza	Neuraminidase (3b7e)	1999
Oseltamivir (tamiflu)		Influenza	Neuraminidase (2hu0)	1999
Imatinib (gleevec)		Leucémie myeloïde chronique	BRC-ABL (3gvu)	2001

Tableau 24. Sept médicaments déterminés par l'analyse des structures de cibles thérapeutiques : le structure-based drug design. *Pour chaque composé, une structure PDB de référence est indiquée. (adapté à partir de Congreve et coll., 2005 [186]).*

En 2003, plus de 40 composés entrés en phase clinique ont été développés à partir d'approches bioinformatiques et chemoinformatiques travaillant sur les structures de cibles thérapeutiques [186].

3. La bioinformatique et la chemoinformatique dans le développement de médicaments

a. Modéliser les structures de cibles

Bien que la biologie structurale ait progressé de façon significative depuis ces dix dernières années grâce aux initiatives internationales et aux projets de génomiques structurales, il reste difficilement envisageable de déterminer la structure de l'ensemble des protéines connues. Comme nous l'avons vu dans le chapitre sur les données de la biologie structurale, il a été décidé que seules les structures des représentants de grandes familles protéiques seraient déterminées [134, 302]. A partir de ces structures modèles, il serait alors possible de modéliser la quasi-totalité des protéines humaines avec une qualité suffisante pour aider à la caractérisation fonctionnelle des protéines [134, 139]. La modélisation des structures de cibles thérapeutiques à partir de structures homologues a commencé à être

exploitée dès 1980 [303, 304] avec l'exemple de la modélisation de la rénine (une cible pour les antihypertenseurs) à partir des structures de protéases aspartiques [304].

La détermination des structures de cibles thérapeutiques peut alors aider à concevoir ou optimiser un composé à partir de la complémentarité de sa surface avec la surface de son site de liaison [292]. On notera par exemple *l'Agenarase* et le *Viracept*, deux médicaments développés à partir de la structure de la protéase HIV en 1989 [305, 306], ou bien encore le *Relanza* développé à partir de la structure de la neuraminidase en 1999 [307, 308].

b. *Identifier les sites de liaisons et protéines druggables*

Nous avons présenté dans le chapitre sur les interactions moléculaires, quelques exemples d'approches prédictives des sites de liaisons, basées sur les biais de composition des interfaces [113, 121, 309], les différences d'accessibilité des résidus de surface [4, 119], la conservation structurale de résidus [273, 278], ou bien encore sur la conservation évolutive de ces résidus [276, 279, 280]. L'identification des sites de liaisons d'une cible thérapeutique renseigne sur sa capacité à lier principalement d'autres protéines ou ADN. L'approche que nous avons développée avec la cartographie moléculaire permet d'aller plus loin en précisant la localisation des différents types de sites de liaisons, en indiquant les partenaires associés, en précisant la fréquence de contact des résidus interagissants ainsi que leur niveau d'hydratation [1]. Toutefois, et comme nous le verrons dans la partie « Criblage haut-débit des régions moléculaires », les inférences réalisées lors de ces cartographies nécessitent de récupérer des structures proches, présentant des similarités suffisantes sur la quasi-totalité de la protéine modélisée. En d'autres termes, les inférences sont limitées par le nombre de structures homologues de la protéine étudiée. En allant plus loin dans la comparaison des structures moléculaires, il est toutefois possible d'inférer des fonctions par une « homologie » ou similarité structurale et physico-chimique de régions plutôt que de structures complètes. Dès lors, quelles que soient la protéine et sa structure étudiée, et que l'on ait ou non d'autres structures homologues, il est possible d'identifier des sites de liaisons et des partenaires associés à partir de la comparaison des régions de cette structure, avec l'ensemble des régions définies par les structures disponibles. Ce point sera davantage abordé par la suite et constitue, avec ses autres applications, le résultat majeur de cette thèse.

L'étude et la comparaison des structures biologiques permettent par ailleurs d'aider à l'identification et à la sélection des cibles « druggables », c'est-à-dire de cibles susceptibles de pouvoir lier des composés administrables [288, 289, 291]. En effet, l'accumulation des données de réussite et d'échec de conception de médicaments pour des protéines a conduit

la communauté scientifique à déterminer que certaines cibles cellulaires sont de mauvaises cibles thérapeutiques [186]. Cela peut être dû soit notamment parce que ces cibles ne possèdent pas une crevasse capable de lier le composé, soit parce que les crevasses capables de moduler le développement de la pathologie sont inaccessibles.

4. Quels autres développements ?

Il est nécessaire à ce stade d'insister sur le fait que, jusqu'à présent, la quasi-totalité des approches de criblage virtuel ne permettent que deux applications : (1) identifier un ligand capable de lier une cible [294, 295, 310] ; et (2) créer un ligand capable de lier une cible, par exemple par les méthodes de *fragment-based drug design* [311-315]. Dans les parties qui vont suivre, d'autres exemples d'applications de criblage virtuel seront présentés. Le criblage virtuel et de nouvelles applications seront davantage détaillés dans le chapitre « Criblage haut-débit des régions moléculaires ».

a. Des bases de données sur les effets secondaires et la toxicité de composés

La comparaison des structures moléculaires dans le cadre du développement de candidats médicaments prend une importance grandissante, d'autant plus que les données concernant les structures, les composés et leurs effets dans la cellule deviennent plus accessibles.

Ainsi, la base de données DrugBank (http://www.drugbank.ca) recense environ 4900 médicaments et détaille plus de 1500 cibles thérapeutiques. Pour chacun de ces médicaments, DrugBank donne accès à une centaine de champs renseignant sur des données chimiques, pharmacologiques, pharmacogénomiques et des données de biologie moléculaire [316-318].

D'autres bases de données comme SIDER (http://sideeffects.embl.de/) ou encore SuperToxic (http://sideeffects.embl.de/), répertorient les effets secondaires [319] et les réponses toxiques [320] de plusieurs milliers de composés.

b. L'importance du contexte cellulaire

Au cours du premier chapitre d'introduction de cette thèse et lors de cette brève description du développement de candidats médicaments, nous avons détaillé deux points : (1) il existe différents environnements cellulaires dans lesquels seule une fraction du génome est exprimée ; nous redécouvrons depuis quelques années que des médicaments

connus pour traiter une maladie (donc dans un tissu donné) peuvent avoir des effets bénéfiques dans d'autres maladies touchant d'autres tissus ; (2) jusqu'à présent, le développement de candidats médicaments implique la détermination d'une cible et la création d'un composé liant avec une très forte affinité cette cible ; toutefois, on ne tient pas compte des autres cibles du composé dans la cellule, ce qui explique 49% des échecs de ces candidats par des effets secondaires/toxiques ou par un manque d'efficacité.

En effet, l'efficacité d'un médicament dans un contexte cellulaire dépend non seulement de son affinité pour sa cible, mais également de ses affinités pour ses autres cibles. Moins le composé sera « spécifique » de sa cible souhaitée, c'est-à-dire, plus il pourra interagir avec d'autres cibles cellulaires, et plus on augmentera les risques d'effets secondaires/toxiques par interférences avec d'autres systèmes, et plus on diminuera implicitement l'efficacité du composé. Dans la partie concernant le « Criblage virtuel haut-débit des régions moléculaires », nous parlerons de ces aspects et présenterons quelques cas précis.

c. Vers une nouvelle utilisation des données structurales

« *… It can be argued that structural biology could play a very important role in systems biology… Structures of macromolecules, especially molecular machines, could provide quantitative parameters, help to elucidate functional networks or enable rational designed perturbation experiments for reverse engineering* »
Luis Serrano, Structures in System Biology, 2007

Le suivi des patients dès la première phase clinique a permis d'observer et de répertorier un certain nombre d'effets secondaires. Nous avons parlé précédemment de la base de données SIDER qui répertorie environ 1500 effets secondaires. En 2008, M. Campillos et coll. s'interrogèrent sur la possibilité de déduire que deux médicaments partagent une partie de leurs cibles cellulaires si ceux-ci induisent de mêmes effets secondaires [321]. Le raisonnement alors tenu consiste à penser que si deux composés induisent de mêmes effets secondaires, alors il est possible qu'ils perturbent effectivement de mêmes cibles dans la cellule.

En testant cette hypothèse sur 502 médicaments, dont des cibles cellulaires sont connues, les auteurs observèrent une forte corrélation entre la similarité de leurs effets secondaires et la vraisemblance qu'ils partagent une même cible. Ils constatèrent également le phénomène connu qui consiste à observer que des ligands aux structures similaires

partagent également de mêmes cibles [322, 323]. En travaillant sur 746 médicaments commercialisés dont 244 pour lesquels aucune cible n'a été identifiée, M. Campillos et coll. identifièrent 2903 paires de médicaments ayant plus de 25% de probabilité de partager une cible (Figure 63). Parmi ces 2903 paires, 956 étaient connues pour partager une même cible. En retirant les prédictions qui pourraient être obtenues à partir de la comparaison des composés, les auteurs ont ainsi pu prédire que 754 paires de médicaments pourraient partager une cible. Des validations expérimentales menées sur 20 de ces paires de médicaments montrent que 13 paires partagent effectivement une cible prédite.

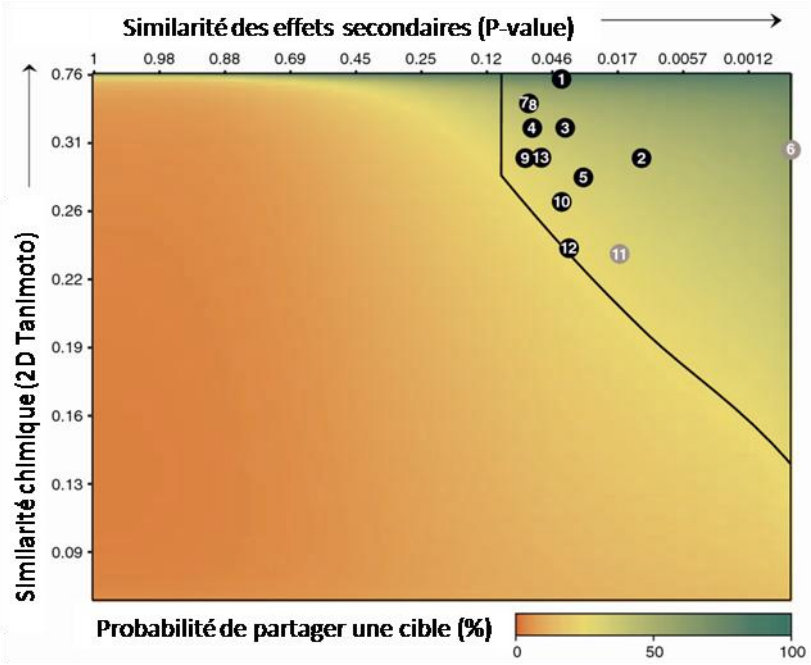


Figure 63. Probabilité de deux médicaments de partager une cible en fonction de la similarité de leurs effets secondaires et de leur similarité chimique. *La partie supérieure droite du schéma délimitée par le trait noir indique les paires de médicaments qui ont une probabilité supérieure à 25% de partager une cible. Les 13 paires qui ont été vérifiées expérimentalement comme partageant une cible cellulaire sont indiquées par des ronds noirs. (source : M. Campillos et coll., 2008 [321]).*

Ainsi, bien que les médicaments soient conçus pour être les plus spécifiques possibles, ils peuvent néanmoins lier d'autres cibles cellulaires, expliquant ainsi de possibles problèmes d'effets secondaires, de toxicité et de manque d'efficacité. Récemment, fin 2009, Keiser et coll. proposent de déterminer de nouvelles cibles de 3665 médicaments approuvés par la FDA (*Food and Drug Administration*) à partir de leurs similarités chimiques, et des cibles qui leur sont déjà connues [324]. Comme pour l'étude de M. Campillos, cette étude révéla des milliers de nouvelles cibles prédites pour ces médicaments. Testés expérimentalement sur

30 couples médicament-cible, 23 nouvelles associations médicament-cible s'avérèrent correctes, cinq d'entre elles ayant des affinités inférieures à 100nM.

L'élucidation de ces nouvelles cibles de composés et de médicaments est importante sinon fondamentale puisqu'elles renseignent sur les modes d'actions des médicaments et permettent de mieux comprendre comment ceux-ci régulent le développement de pathologies. La détermination des nouvelles cibles de médicaments est une étape cruciale dans le développement de médicaments et permettra d'expliquer pour une part les problèmes d'efficacité, d'effets secondaires et de toxicité qui représentent 49% des échecs dans le développement de médicaments. Toutefois, les méthodes existantes dans la littérature nécessitent soit de connaître des effets secondaires de nos composés (ce qui nécessite des études relativement poussées), soit de comparer le composé d'intérêt aux banques de médicaments déjà connus pour identifier des cibles par similarité sur le composé.

La partie « Criblage virtuel haut-débit des régions moléculaires » présente une méthodologie capable d'identifier des cibles de composés à partir de la similarité de leurs sites de liaisons, ou à partir de la complémentarité des régions du composé, avec des régions de cibles moléculaires. L'approche n'est donc pas limitée par l'état des connaissances actuelles sur les médicaments pour inférer de nouvelles cibles. En particulier, et comme nous le verrons par la suite, l'utilisation des structures pour modéliser un environnement cellulaire permet de fournir la première évaluation de la spécificité (ou fréquence) d'une région moléculaire. Elle permet par conséquent de décrire le paysage moléculaire d'une cellule avec finesse en précisant quelles sont les régions les plus représentées et les moins représentées. L'étude de la spécificité d'une région est particulièrement importante pour le développement de candidats médicaments : il est en effet possible de cartographier sur une structure moléculaire l'ensemble de ces régions et de préciser quelles sont celles qui sont fréquentes dans l'environnement (que l'on retrouve sur d'autres molécules), et quelles sont celles qui sont rares, voire qui ne se retrouvent que sur la cible étudiée. Dès lors, cette information supplémentaire peut être utilisée lors du développement d'un candidat médicament en choisissant de créer un composé qui cible de préférence les régions spécifiques, c'est-à-dire les régions propres de la cible cellulaire. Ce faisant, on limite très en amont du développement de médicaments, de possibles risques d'interférences avec d'autres molécules ; ces interférences étant susceptibles d'engendrer des effets secondaires et des réponses toxiques, ainsi que de diminuer l'efficacité générale du composé par équilibre thermodynamique avec d'autres cibles.

De plus, pour réellement évaluer les effets secondaires, toxiques et les manques d'efficacité, il est nécessaire de pouvoir modéliser correctement un environnement cellulaire donné, c'est-à-dire de déterminer un ensemble représentatif des molécules en présence dans cet environnement, en précisant leurs fréquences. Dans de futurs développements, nous comptons inclure des données de génomiques fonctionnelles et plus particulièrement de transcriptomique (données d'expression des gènes), afin de pouvoir quantifier cette fréquence des molécules dans l'environnement.

La modélisation de ces phénomènes est pour nous un point capital et je pense qu'une évaluation correcte de ces problèmes permettra de révolutionner la recherche de nouveaux médicaments, en évaluant dès les premières étapes du développement, de possibles risques d'effets secondaires, de toxicité ou de manque d'efficacité. L'importance que revêtent pour moi ce domaine et ces questions est telle qu'elles ont motivé la création de l'entreprise de biotechnologies BIONEXT en avril 2009. Evaluée par un jury national, BIONEXT est lauréat du Concours National 2010 de Création d'Entreprises Innovantes. L'entreprise a permis de donner les moyens techniques de réaliser ces développements et de fournir la puissance de calcul nécessaire pour ces criblages.

Modélisation des structures moléculaires

I. Problématique : vers une modélisation rapide et générique des surfaces moléculaires

Nous avons vu au cours des chapitres précédents que l'analyse des surfaces moléculaires et la prédiction de sites de liaisons sont deux étapes importantes pour caractériser fonctionnellement les molécules. La croissance du nombre et de la taille des structures moléculaires est telle qu'il est nécessaire de développer des approches bioinformatiques capables de modéliser et d'analyser rapidement et précisément ces molécules, quel que soit leur type moléculaire. L'une des façons privilégiées d'analyser les surfaces moléculaires consiste à générer des régions, de tailles plus ou moins grandes, afin de repérer des propriétés remarquables et définir, identifier et regrouper les molécules par ces propriétés.

Le travail présenté dans cet article est donc une première étape méthodologique expliquant comment modéliser ces surfaces et régions moléculaires, en précisant notamment que ces régions peuvent correspondre à des sites de liaisons déjà identifiés. En particulier, par l'analyse de ces régions, il a été possible de formuler un score de courbure local non biaisé permettant d'analyser la composition en acides aminés des bosses et des creux de surface. Cette analyse a mis en évidence des différences notables entre ces régions, et notamment pour les résidus interagissants.

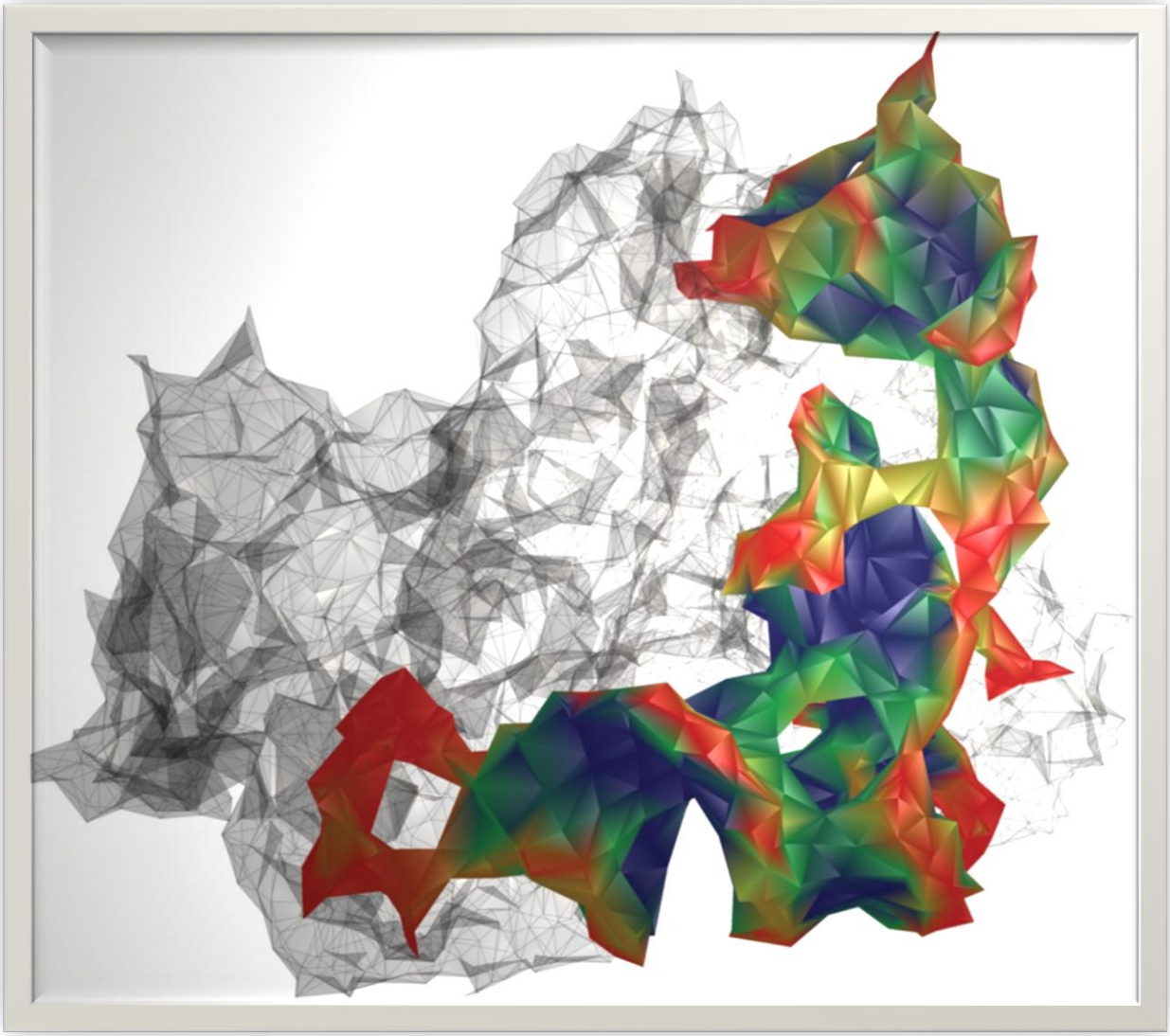


Figure 64. Etude et caractérisation des surfaces moléculaires par les formes alphas. *En arrière plan se trouve la structure de l'argynile ARNt synthétase modélisée par les formes alphas en transparent (PDB : 1f7u). Au premier plan, l'ARNt(Arg) est représenté par cette même forme alpha et colorée en fonction de la nouvelle méthode de calcul des courbures locales (crevasses en bleu, plats en vert, bosses en rouge). La figure a fait la couverture du journal Proteins : Structure, Function and Bioinformatics en Juillet 2009 [4].*

II. Article : Defining and characterizing protein surfaces using alpha shapes

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Defining and characterizing protein surface using alpha shapes

Laurent-Philippe Albou,[1†] Benjamin Schwarz,[1,2†] Olivier Poch,[1*] Jean Marie Wurtz,[1] and Dino Moras[1]

[1] Department of Biology and Structural Genomics, IGBMC, CNRS, INSERM, ULP, Illkirch, France
[2] LSIIT UMR 7005 CNRS, Université de Strasbourg, Strasbourg, France

ABSTRACT

The alpha shape of a molecule is a geometrical representation that provides a unique surface decomposition and a means to filter atomic contacts. We used it to revisit and unify the definition and computation of surface residues, contiguous patches, and curvature. These descriptors are evaluated and compared with former approaches on 85 proteins for which both bound and unbound forms are available. Based on the local density of interactions, the detection of surface residues shows a sensibility of 98%, whereas preserving a well-formed protein core. A novel conception of surface patch is defined by traveling along the surface from a central residue or atom. By construction, all surface patches are contiguous and, therefore, allows to cope with common problems of wrong and nonselection of neighbors. In the case of protein-binding site prediction, this new definition has improved the signal-to-noise ratio by 2.6 times compared with a widely used approach. With most common approaches, the computation of surface curvature can be locally biased by the presence of subsurface cavities and local variations of atomic densities. A novel notion of surface curvature is specifically developed to avoid such bias and is parametrizable to emphasize either local or global features. It defines a molecular landscape composed on average of 38% knobs and 62% clefts where interacting residues (IR) are 30% more frequent in knobs. A statistical analysis shows that residues in knobs are more charged, less hydrophobic and less aromatic than residues in clefts. IR in knobs are, however, much more hydrophobic and aromatic and less charged than noninteracting residues (non-IR) in knobs. Furthermore, IR are shown to be more accessible than non-IR both in clefts and knobs. The use of the alpha shape as a unifying framework allows for formal definitions, and fast and robust computations desirable in large-scale projects. This swiftness is not achieved to the detriment of quality, as proven by valid improvements compared with former approaches. In addition, our approach is general enough to be applied on nucleic acids and any other biomolecules.

Proteins 2009; 76:1–12.
© 2008 Wiley-Liss, Inc.

Key words: surface; alpha shape; patch; curvature; binding site; interaction; knob; cleft; structural bioinformatics; computational biology.

INTRODUCTION

The biological function of a protein essentially relies on its interactions with solvent and other biomolecules. Chemical and structural diversity observed at molecular surfaces allow for the wide variety of interactions necessary for cellular life. To decipher biological processes, it is thus crucial to accurately define the nature and shape of these surfaces. The determination of the surface in terms of atoms, residues, and surface patches has already allowed to conduct numerous studies in the protein–protein interaction fields[1–3] as well as to develop several prediction algorithms for the detection of binding sites and the modeling of complexes.[4–7] More detailed characterization of the surface in terms of clefts and knobs (respectively concavities and convexities) was also used for the study of interface complementarity and docking of molecules.[8–11]

Amongst the methodologies used for the description of the surface of a molecule, the alpha shape theory[12] is probably one of the most promising. The alpha shape model of a molecule is a polyhedral representation that uniquely decomposes the space occupied by its atoms and retains interesting characteristics such as the shape of the molecule and a notion of interatom neighborhood. Despite the relative complexity of the theory, alpha shapes have been used to address a wide variety of problems in structural biology, such as the computation of protein surface and volume[13] as well as their derivatives,[14] the detection of pockets in known structures,[15–17] the construction of molecular surface meshes,[18,19] the validation of structures,[20,21] or the study of interfaces.[22,23]

Additional Supporting Information may be found in the online version of this article.
Grant sponsors: The Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé et de la Recherche Médicale (INSERM), Structural Proteomics in Europe (SPINE2-Complexes, CEE FP7 LSHG-CT-2006-031220), the Décrypthon program initiated by the Association Française contre les Myopathies (AFM, CAMI 12727), IBM, the Ligue Nationale contre le Cancer, comité du Haut-Rhin and the Université Louis Pasteur de Strasbourg (ULP).
[†]Laurent-Philippe Albou and Benjamin Schwarz contributed equally to this work.
*Correspondence to: Olivier Poch, 1 rue Laurent Fries, BP 10142, 67404 Illkirch CEDEX, France. E-mail: poch@igbmc.fr
Received 26 June 2008; Revised 30 September 2008; Accepted 1 October 2008
Published online 21 October 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22301

In this article, the alpha shape theory is used as a unifying framework to compute various properties depicting the surface of a biomolecule. The definition of *surface atoms* is straightforwardly provided by the alpha shape model. For the definition of *surface residues* a novel notion is introduced, the *valence Vr* representing the density of surface interactions around an accessible residue. By radiating on the surface around a surface residue, we give a novel and intuitive definition of *contiguous surface patches*. Curvature computations based on solid angle approaches[8,24] usually do not differentiate the empty space due to subsurface cavities from the empty space above the surface; as a result they detect more protrusions than they should. To tackle this problem, we define the *exposure* of an atom as a modified solid angle computed locally above the alpha shape surface. This exposure is then smoothed in a surrounding region to define its local surface curvature. Based on this latest notion, clefts and knobs are detected on the surface and characterized in terms of accessibility and composition.

The biological relevance of these definitions is validated on a dataset of 85 proteins involved in transient heterodimeric interactions for which both the bound and unbound forms are available. We consider a protein chain to be in an unbound form if it participates only to crystal packing contacts.[25] As some conformational changes can occur during an assembly formation, it has been proposed to predict protein-binding sites using only these unbound forms.[5] In our dataset where small conformational changes are observed, we verified that most of the interacting residues (IR) seen in bound forms are also found on the surface in their respective unbound forms. In the context of protein-binding site analysis, we show that our surface patches have a better overlap with known binding sites and a better signal-to-noise ratio than the commonly used approach of Jones and Thornton.[26] In addition, our conception of local surface curvature correlates well with visual inspection and is compared with a former approach.[24] It allows a fast detection of clefts and knobs, dividing the surface in 38% knobs, the remaining being clefts. Knob residues are found to interact 30% more with partner proteins than cleft residues. IR are also shown to be, respectively, 15 and 18% more accessible than noninteracting residues (non-IR) in knobs and clefts. A more detailed analysis of these accessibilities reveal that hydrophobic and aromatic IR have 54% more accessibility in clefts than in knobs with respect to non-IR. The IR in knobs are indeed found to be more charged and less hydrophobic and less aromatic than those in clefts.

Our implementations benefit from fast and robust algorithms recently developed in Computational Geometry and provided by the CGAL library.[27] The geometric representation of alpha shapes combined with our geometric descriptors is generic and applicable to any molecular structure such as proteins, nucleic acids, or lipids. Furthermore, these tools are fast enough for use in large-scale projects such as interactomics.

METHODS

Alpha shapes

As depicted in Figure 1(A), molecules are generally described as a union of balls representing either van der Waals (VdW) or solvent accessible (SA) models.[28] Another common model is provided by the molecular surface,[29] defined as the limit of space around the molecule that a rolling probe sphere can actually touch.

Molecules can also be modeled with their *Delaunay complex*[30] [Fig. 1(B)] which is a unique partition of the three-dimensional (3D) space in nonoverlapping tetrahedra whose vertices are atom centers. This construction bears information on the atom neighborhood: a Delaunay edge links the two nearest atoms in the direction of that edge. Such edges can be arbitrarily long, covering for instance a surface cavity [Fig. 1(C)], segment [ab]). By trimming the "largest" Delaunay edges, triangles, and tetrahedra, it is possible to distinguish between the voids surrounding the molecule and the actual molecular object [Fig. 1(C), gray colored triangles]. This task is achieved through the *alpha complex*,[12] a filtration of Delaunay edges, facets, and tetrahedra based on the growth of soft balls virtually placed on all atom centers of the molecule. The size of these so called α-balls increases with alpha, and the alpha complex registers contacts between alpha balls: when two (respectively three or four) alpha balls touch each other, the corresponding Delaunay edge (respectively facet or tetrahedron) belongs to the alpha complex for this specific alpha value. In the present study we restrict our use to an alpha parameter of 0 [Fig. 1(C)]. This particular construction (also referred as *the dual complex* in the literature) corresponds to the case, where the radius of a ball modeling an atom measure the van der Waals radius of this atom raised by a probe sphere radius (generally 1.4 Å corresponding to a water molecule). In the following descriptions and discussions, we will assume a value of 0 for every occurrence of alpha. As demonstrated by Edelsbrunner[31] this construction is a unique and precise representation of the molecule.

The *alpha shape* (with alpha equal to 0, also known as *dual shape*) of a molecule is the border of its alpha complex [Fig. 1(C), bold segments]. It is a polyhedron with triangular facets that precisely depicts the surface of a molecule (see Fig. 2). A facet in the alpha shape links a triplet of surface atoms blocking a probe sphere, whereas an edge links two atoms that allow the same probe sphere to roll from one blocking position to another. The vertices of the alpha shape are exactly the atoms with a strictly positive accessible surface area (ASA $>$ 0 Å).

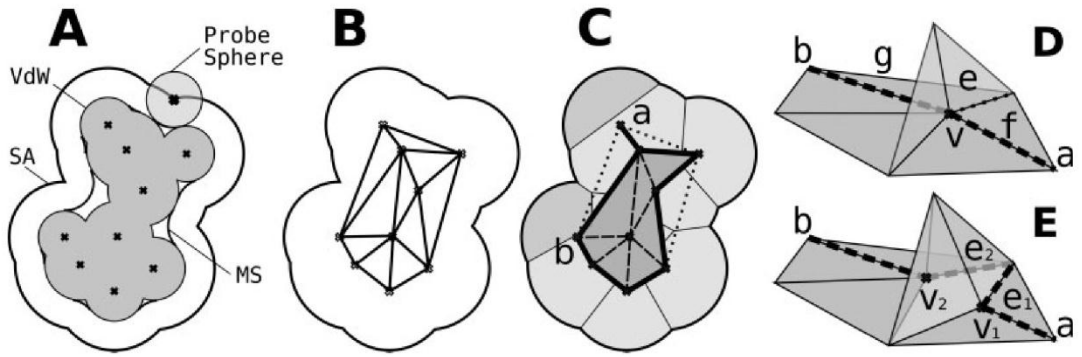


Figure 1

Molecular surface representations. (**A**) 2D example of van der Waals (VdW) representation of the molecule modeled as a union of balls with VdW radii. By rolling a probe sphere on the VdW surface one defines the surface accessible (SA) and Molecular Surface (MS) models. (**B**) In 2D, the Delaunay complex of a molecule (here represented in its SA model) is composed of straight lines between atom centers and the triangles they delineate. In 3D this "triangulation" contains also tetrahedra. (**C**) The molecule's 0-balls (matching the atoms in their SA representation) have been represented and their contacts emphasized by thin straight lines separating them. The alpha complex ($\alpha = 0$) comprises all Delaunay edges except the three dotted ones (for instance, the edge between vertices a and b is stripped because the corresponding grayed balls do not touch each other). For similar reasons three Delaunay triangles are stripped and only the seven grayed triangles belong to the alpha complex. The alpha shape is the border of the alpha complex, it is pictured here with bold edges. (**D**) A case of 3D surface ambiguity, for clarity the upper facets are depicted transparent. Traveling on the surface from vertex a to b with edges f and g allows one to cross the upper facets through vertex v. (**E**) To forbid such surface crossings, v is split in two, as well as the transparent facets and its two basis edges. Traveling from a to b now necessitates to visit edges e1 and e2.

The surface of the alpha shape may present ambiguities (nonmanifoldness) in cases where an atom (vertex) is shared by two sides of the surface [Fig. 1(D)]. Ambiguous vertices, edges and triangles are virtually split to prohibit surface crossing of facets [Fig. 1(E)]. The resulting surface is stored in a half-edge data structure. Ulti-

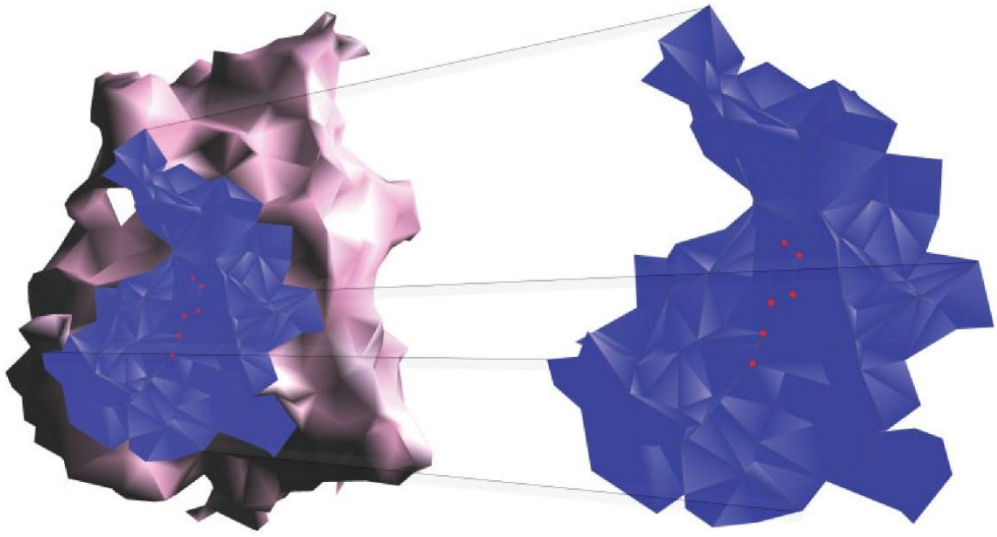


Figure 2

Alpha shape and a surface patch. Alpha shape of the protein RAR α (1dkfB) in purple. In blue, a contiguous surface patch generated by our approach. For better visualization, the right part of the figure presents a focus on the surface patch.

mately, the vertices and edges of this modified alpha shape provide a *graph depicting the neighborhood of atoms* on the protein surface. Only atoms that share a surface intersection in the SA model will be connected by an edge in this graph. A *graph of surface residue* neighbors is also constructed by connecting residues that share at least one edge in their atom neighborhood graph.

To compute alpha shapes, we rely on the library CGAL,[24] and use as parameters an alpha value of 0, a probe sphere radius of 1.4 Å, and common van der Waals radius for atoms. For a deeper insight into the alpha shape theory and for its relationship with molecular models please refer to other articles from Edelsbrunner and Mucke[12] and Edelsbrunner.[31] An introduction to these models is also provided by A. Poupon.[32]

Dataset of bound and unbound protein structures

To evaluate our surface descriptors, we have built a dataset of 85 proteins for which the structures of both their bound and unbound forms are available. Bound forms correspond to the structure of the protein extracted from the structure of an assembly, whereas the unbound forms correspond to the structure of the protein that participates only to crystal contacts.[25] Each of these proteins is involved in transient heterodimers (following the definition of Nooren and Thornton[33]), and therefore both the bound and unbound forms have a biological meaning.

As a first step, structures of protein assemblies with resolution better than 3 Å are extracted both from the Protein Data Bank (PDB)[34] and already published datasets.[3] Then, a non-redundant dataset of 225 transient heterodimers is built with a maximum sequence identity of 30%. Antigen-Antibody structures and assemblies with fragmented proteins are also removed. The Average length of protein chains is 240 amino acids and no protein chains have less than 50 amino acids or more than 576 amino acids.

The transient state is inferred *in silico* by checking in the PDB if known IR detected in a structure assembly are found to interact with at least one different partner in another assembly. However, these contacts may result from crystal packing and therefore do not necessarily occur *in vivo*. For this reason, all our assemblies and their transient aspects have been manually verified by consulting experiments described in the literature.

The unbound forms are then retrieved using the following approaches:

(a) Protein structures that are described as monomers in PISA[35] are retrieved.
(b) Each protein chain is compared with the monomers of PISA, using BLAST.[36] To lower the occurrence of conformational changes due to key mutations and emphasize only those due to the assembly formation, only structural candidates with at least 95% residue identity over 95% of the sequence length are retrieved.
(c). If several unbound structures remain, the one with the best resolution and b-factor is selected.

As a result of this process, 85 proteins are obtained for which both their bound and unbound forms are available (Supp. Info. Table I).

IR are then detected on bound forms by a change of accessibility of at least 1 Å^2 during the assembly formation.[3] The ASA values of atoms and residues are computed using the Naccess program[37] with default parameters. IR are then mapped on the corresponding unbound forms, using a pairwise alignment.

Surface residues

A *surface atom* is defined as an *accessible atom* (ASA > 0 Å^2). As previously stated, accessible atoms correspond exactly to the vertices of the alpha shape. The *valence Va of an accessible atom* is defined as the number of its accessible atom neighbors (the number of its edge connected atoms in the alpha shape). The *valence Vr of an accessible residue* (ASA > 0 Å^2) is defined as the number of edges connecting atoms from that residue to atoms of other accessible residues. An accessible residue is then considered as a surface residue by combining its number of surface atoms Nr, and its valence Vr (see "Results and Discussion" section).

In the approach of Miller *et al.*,[38] surface residues are defined as those having an observed ASA of at least 5% of their reference ASA. The reference ASA of a residue X is the ASA of the residue in a polypeptide extended-state Gly-X-Gly.

More recently Chakravarty *et al.*,[39] have proposed a novel way of defining surface residues by computing a notion of depth for every atom and residue of a protein structure. To compare this approach with ours, we implemented this notion of depth using the surface atoms as reference, as proposed by Pintar *et al.*[40]

To optimize our definition of surface residues, a measure of sensitivity is assessed by considering the fraction of known IR that are described as being part of the surface in bound forms. Residues that are not described as surface residues are considered part of the protein core. A measure of specificity is then evaluated by keeping trace of the fraction of amino acids that constitutes the protein core.

We perform an optimization of our surface residue detection by varying simultaneously Nr and Vr. During this optimization, the best definition of surface residues is attained when almost all IR are detected as being surface residues, whereas the protein core contains the biggest fraction of amino acids. For comparisons with the Miller *et al.*[38] and Pintar *et al.*[40] approaches, we vary respectively the percentage of accessibility and the thresh-

old of depth used to detect surface residues and evaluate the fraction of core residues.

Surface patches

Two kinds of surface patches are generally considered in the literature: surface patches of variable size that correspond to subregions of an interface assembly[3] and surface patches of a given size that are generated evenly over the surface of a single protein.[26] Although the first approach is aimed at better characterizing a known interface, the second approach is commonly used to average properties on a specific region to predict a biologically relevant fact such as protein or nucleic acid binding sites. We focus on this second definition.

To construct an atom surface patch around a surface atom, we gather the nearest surface atoms that are reachable over a continuous surface from that center (see Fig. 2). This is achieved by computing minimal distances from the central atom to every other atom in the graph of surface atoms introduced in the Alpha shapes section. This computation relies on the Dijsktra shortest distance path algorithm[41] where edges linking two surface atoms are weighted according to the euclidian distance separating them. Essentially, the distance over the surface computed for any two atoms is the sum of the edge lengths forming the shortest path between these two atoms.

Residue surface patches are computed in the same way and the weight of an edge linking two accessible residues correspond to the minimal euclidian distance between any of their atoms. This is achieved by assigning a distance of 0 to each atom of the central residue in the Dijsktra algorithm.

By construction, our surface patches are edge connected. This means that any atom of the patch is reachable from any other atom of the patch through a list of atomic intersections over the surface. In the following evaluation, this property has been chosen for the study of *surface patch contiguity*.

In the commonly used definition of surface patches of Jones and Thornton in 1997,[26] surface residues are characterized by their C_α and a solvent vector pointing toward the solvent. To select only surface residues that are on the same side of the surface, surface residues are added to the patch if the angle between their solvent vectors and the solvent vector of the central residue is less than 110°.

Every surface patch of 20 residues[5,7] was generated over the surface of every protein of our dataset of bound forms with both approaches. These surface patches are mapped into sub-graphs, taking as reference our graph of surface atoms. Then, several measures are analyzed:

1. For each protein chain, the maximum overlap (in terms of residues) between the known binding site and any of the surface patches.
2. The number of contiguous subregions that compose a surface patch, that is the number of connex composants of the surface patch graph. By construction, our surface patches always define a unique region.

To further understand the differences observed between these approaches, we evaluated the number of surface atoms and residues, that are contiguous to a central surface residue in our method and that are not present in the corresponding patch obtained by the former approach.

For the prediction of protein-binding sites, where the interacting potential of a residue is determined by the analysis of properties in the surrounding region, it is necessary to generate automatically the surface patch that will best overlap with a binding site while having the lowest number of non-IR. This problem consists in finding the patch size N that will optimize the *signal-to-noise* ratio Q [Eq. (1)].

$$Q_{(P,N)} = \frac{O_{(P,N)}}{NIR_{(P,N)}} \quad (1)$$

$O_{(P,N)}$ is the best observed percentage of overlap between a surface patch of size N (expressed in terms of number of atoms, number of residues or distance) and the known binding site of the protein P. $NIR_{(P,N)}$ is the percentage of non-IR inside that patch.

Surface curvature

The *relative exposure* Ω of a surface atom a is defined as the fraction of a tiny sphere centered on a, that lies outside the alpha shape. In the present two-dimensional (2D) example [Fig. 3(A)], this value corresponds to the sum of normalized angles ω_1, ω_2, ω_3 of « empty » Delaunay triangles at atom a. In 3D this generalizes to a sum of tetrahedra solid angles [Fig. 3(B)]. To better differentiate the values corresponding to clefts and knobs, Ω was normalized to define values ranging from −1 (cleft) to 1 (knob), with 0 defining a flat region. The *relative exposure of a residue* is defined as the mean of its atom exposures and as such, follows the same rules of normalization.

In some "degenerate" cases, the surface of an atom might be scattered in disconnected atomic components as illustrated in the 2D example [Fig. 3(C)]. The initial split of vertices in the alpha shape (presented in "Material and Methods" section) allows us to maintain a distinct value for each atomic component. A per atom value is obtained by summing up all component of the atom with the exception of cavities [Fig. 3(C), b2].

To define the *local surface curvature* for an atom, relative exposures are smoothed on a surrounding concentric region. Starting from a central atom, a *smoothing region* is determined by considering all surface atoms

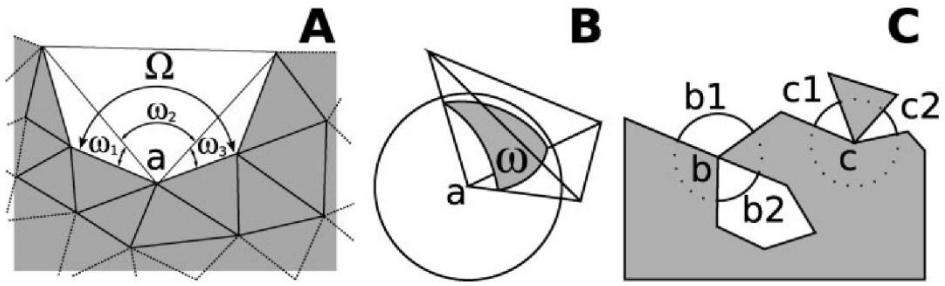


Figure 3

Relative exposure. (**A**) In 2D the relative exposure Ω of vertex a corresponds to an "empty angle" around this vertex. It is computed as a sum $\omega_1 + \omega_2 + \omega_3$ of solid angles. (**B**) The solid angle ω at vertex a of a tetrahedron is the portion of surface area lying inside the tetrahedron of a tiny sphere centred on a. (**C**) A 2D example where a vertex may have more than one component. The atom b has two components, the component b1 is on the surface of the protein and the component b2 is in a cavity. Atom c has two components, both on the surface of the protein.

accessible through a maximum of *s* edges, where *s* is a size parameter.

To emphasize local features, more importance is given to atoms near the patch center than to remote ones:

$$C(a) = \sum_{i \in \text{surfacepatch}} \frac{\Omega(i)}{d(a,i)} \qquad (2)$$

where $C(a)$ is the local surface curvature of atom *a*, $d(a,i)$ is the distance over the surface (in the graph) between atom *a* and atom *i*, and $\Omega(i)$ is the relative exposure of atom *i*. The local surface curvature of a residue is computed as the mean of its atom values.

To validate our definition of local surface curvature, we compared our values with those computed by the CX program,[27] an approach similar to common solid angle approaches.[8–11] To assess the curvature, CX approximates the amount of space filled by atoms within a sphere of 10 Å. We further used our definition of local surface curvature to define clefts and knobs over the surface and characterized them in terms of accessibility and composition.

RESULTS AND DISCUSSION

Defining the protein surface

Surface residues

Following the definition of sensitivity and specificity proposed in "Material and Methods" section, the best definition of a surface residue is found to be an accessible residue that either possesses five surface atoms or has a valence Vr higher than 10. With these parameters, 98% of the IR were detected as being part of the surface, whereas 20% of the residues were assigned to the protein core (see Fig. 4).

Similar results are observed for the approach of Miller *et al.*[38] with the default parameter of 5% accessibility, leading to the detection of 96% of IR as being part of the surface, while keeping a protein core formed on average of 24% of the residues.

The best result that could be achieved with the residue depth approach[40] was obtained with a depth of 2.3 Å. Although 98% of IR are detected as part of the surface, the protein core is less well defined with only 15% of the residues. Other depth-based approaches that use water molecules as surface referents might perform better, but are far more time consuming due to the placement of these referents either by Molecular Dynamics or Monte Carlo approaches.

We confirm the good distinction between protein surface and core by analyzing several physicochemical properties known to differentiate these two structurally different regions: the hydrophobicity and flexibility are computed with the amino acid scales of Argos (1982) and Creamer (2000), and the amino acid conservation is computed as a Shannon Entropy[42] derived from a multiple alignment generated by PipeAlign.[43] These three approaches that divide residues into surface and protein core emphasize the same differences in physicochemical and evolutive properties: the protein surface is on average (1) 40% less hydrophobic, (2) 65% more flexible, and (3) 75% less conserved than the protein core (Supp. Info. Table II).

Our approach to define surface residues is comparable with the one of Miller *et al.*,[38] which is widely used to differentiate protein surface and core. This suggests the importance of a novel parameter introduced in this study, the valence Vr, which represents the density of interactions at the surface. Interestingly, IR are detected as being part of the surface equivalently in both bound and unbound forms (data not shown). This remark further supports the possibility to predict protein-binding

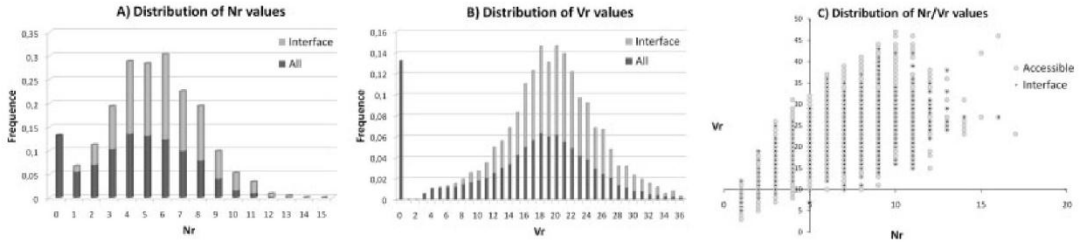


Figure 4

Selection of parameters to define surface residues. The distribution of Nr values for both Interface and All residues is shown. To detect most of the interface as being part of the surface, it seems reasonable to select an Nr superior or equals to 2 if the parameter is used alone. For Vr distribution, a threshold of 10 or 11 can be chosen to detect by itself most of the interface. Finally, the plot of Nr/Vr leads to select Nr = 5 and Vr = 10 to detect more than 98% of the interface as being part of the protein surface.

sites using unbound forms where only small or moderate conformational changes occurred during the molecular assembly formation.

Surface patches

All surface patches were generated on the dataset of 85 bound forms, with a patch size similar to previous studies (N = 20 residues).[4–7] Binding sites in our dataset are composed of 27 amino acids on average. For each protein chain, our best overlapping surface patch contains on average 15–16 IR, corresponding to an average overlap of 62.3% with known binding sites, while retaining only four to five (22%) non-IR. The signal-to-noise-ratio Q was thus improved by 63.6%, ranging from 2.2 for the Jones and Thornton approach, to 3.6 for our method (Table I).

Generating residue surface patches of 20 residues on each surface atom rather than on each surface residue further increases the signal-to-noise ratio to 4.9. For such a definition, the best overlap with a known binding site is 65.9% on average, while retaining only 17.4% of non-IR, thus improving the signal quality by more than two times compared with Jones and Thornton.

Finally, to obtain the best signal-to-noise ratio with our approach, an optimization was performed by varying the size of the patch. Experiments revealed a peak for 15 residues, with respectively Q = 5.8 for bound forms and Q = 5.1 for unbound forms. With this patch size, the average best overlap of a surface patch with a binding site is respectively 55.4% in bound forms and 55.8% in unbound forms, while the percentage of non-IR inside the patch represents no more than 9.3% in bound forms and 11.8% in unbound forms. Compared with the approach of Jones and Thornton, the use of residue surface patches generated on a per atom basis combined with a patch size of N = 15 residues thus improved the signal-to-noise ratio by a factor of 2.6.

Although the per atom approach generates about 10 times more patches than the per residue approach, it is fast enough to be applied in large-scale projects thanks to the combination of two fast computational tools : alpha shapes and Dijkstra graph travel.

As a further refinement, our surface patches can be extended to define core and rim patches. Further experiments will be conducted to explore a potential correlation between these definitions and the notions of interface core and interface rim.[2,3]

Characterizing the protein surface

Our definition of relative exposure shares similitudes with the ASA. This statement was verified on our dataset

Table I
Comparisons of the Best Overlapping Patches with Binding Sites

	Alpha-shape[a]	Alpha-shape[b]	Jones and Thornton
Maximum overlap ($O_{(P,M)}$)	62.3%	65.9%	56%
Fraction of non interacting residues ($NIR_{(P,M)}$)	22%	17.4%	25.5%
Signal-to-noise ($Q_{(P,M)}$)	3.6	4.9	2.2
Number of missed contiguous atoms^{+}	—	—	18
Number of missed contiguous residues^{+}	—	—	3.9
Number of subregions selected per patch^{+}	—	—	1.5

Surface patches are generated using a size parameter of 20 residues on the dataset of bound forms. Values of missed atoms and residues as well as the number of subregions are not applicable to our approach.
Values for parameters noted with a ($^{+}$) have not been computed for a and b as alpha shape is taken as reference.
[a]Residue surface patches generated for each surface residues.
[b]Residue surface patches generated for all surface atoms.

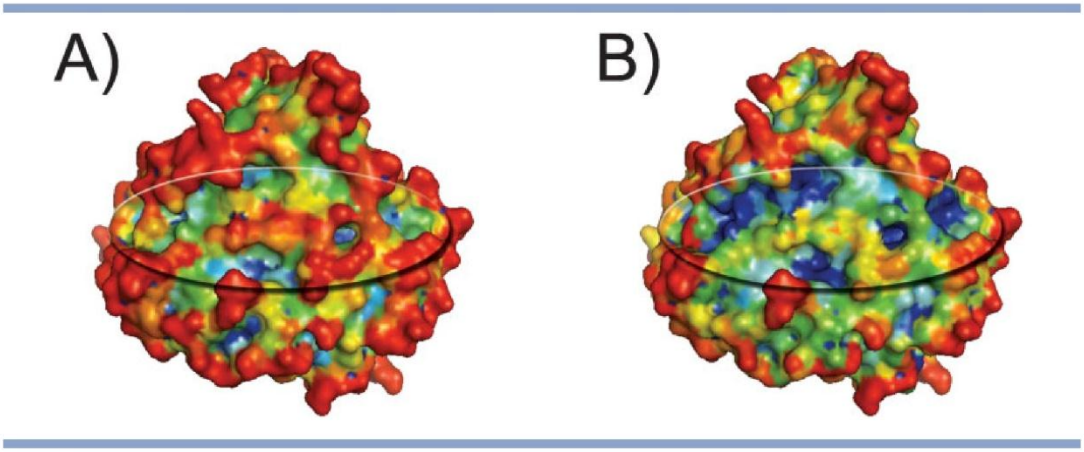


Figure 5

Protein structure of Cytochrome P450 (1eup:A) visualized with PyMol.[44] CX values are represented in (**A**), our local surface curvature index in (**B**). To allow comparisons between (A) and (B), values have been normalized using 90% of all atomic values and using a threshold of 1 to differentiate between knobs and clefts for CX values and 0 for our solid angle approach. Blue indicates clefts (values near −1), green indicates flat surfaces (values near 0) and red indicates knobs (values near 1). The cavity of 855 $Å^3$ (SA model) below the ellipsoid region increases the global amount of empty space and bias the curvature of CX toward the detection of flat and knob regions.

by a strong Pearson product-moment correlation coefficient between these two notions (respectively 0.86 for atoms and 0.89 for residues).

Like the ASA values, the relative exposure values are subject to great variations in the neighborhood of an atom. This property is explicitly depicted by the lacunary nature of the alpha shape surface. To reflect a local trend of the surface around an atom, we introduced the notion of local surface curvature by smoothing the relative exposures in the atom neighborhood. The size of the smoothing region is a critical parameter used to define the level of details to be observed on the surface. When this smoothing region is small, the accent is placed on local details, whereas when it is larger global features are emphasized. A visual inspection was performed and a reasonable balance between local and global surface features was achieved for a size of smoothing region $s = 2$ [Fig. 5(B)].

Several approaches have already been proposed to address the determination of protein surface curvature. Because of the intuitiveness of this notion, no quantitative evaluation exists to differentiate poor and good approaches. Nevertheless, we compared our results with CX,[27] an existing method similar to common solid angle approaches.[8–11] In both cases, curvature values were computed for all surface atoms and only a moderate correlation (0.64) was observed. For a more detailed understanding of this low correlation, we proceeded to a quantitative comparison to verify if both approaches were able to detect the same protruding regions. Considering the 10% of atoms with higher values, an overlap of 66% is observed between both approaches. The main differences were observed for regions detected either as flat (near 0) or cleft (near −1) by our local surface curvature score. This low overlap can be explained by the difference in methodology behind the two approaches. CX being based on local atomic densities can be biased by the presence of subsurface cavities, a problem common to most solid angle approaches.[8–11] In the extreme case, when influenced by local variations of densities or by the presence of cavities, even clefts can be detected as protruding [Fig. 5(A)]. In contrast, our method allows to distinguish between the empty space above and below the surface and only considers the empty space above the surface to reflect the real local curvature.

To study the protein surface topography, we defined *knob residues* as surface residues with a local surface curvature greater than 0, and *cleft residues* as those with a local surface curvature smaller than 0. Following this definition the surface is composed on average of 62% of clefts and 38% of knobs. Furthermore, knob IR contributes to 6.7% of the surface of the protein whereas cleft IR contribute to 8.6% of this surface. Therefore, the bayesian probability of having an interacting residue in a knob is 0.174 and 0.139 in a cleft. IR are thus 30% more frequent in knobs than in clefts, a result that correlates with the fact that IR are known to be, on average, relatively accessible (Table II).

To further understand the differences between cleft and knob regions, we proceeded to a detailed analysis of

Table II
Amino Acid Accessibility and Composition of Protein Surfaces and Interfaces

	Accessibility[a]							Area[b]						Propensities[c]			
	Surface			Interface			Interface/surface[d]	Surface			Interface			Ln (Knob/cleft) (1)		Interface/surface (2)	
Residue	All	Knob	Cleft	All	Knob	Cleft	Knob	Cleft	All	Knob	Cleft	All	Knob	Cleft	Surface	Interface	Ln (Knob/knob)	Ln (Cleft/cleft)
PHE	41	67	29	76	108	40	61	38	1.8	1.2	3.4	4.9	4.8	4.9	−1.01	−0.00	1.36	0.36
TYR	58	94	38	80	113	49	20	29	3.3	2.5	5.2	7.4	7.0	8.3	−0.73	−0.17	1.03	0.47
TRP	51	79	37	71	93	59	18	59	1.1	0.8	2.0	2.1	1.3	4.3	−0.95	−1.22	0.47	0.74
ALA	44	56	28	47	58	30	4	7	4.6	4.7	4.6	3.9	3.9	3.8	0.02	0.04	−0.18	−0.20
VAL	45	66	30	52	73	35	11	17	3.5	2.8	5.3	4.0	3.7	4.9	−0.65	−0.30	0.27	−0.08
LEU	47	77	33	66	88	43	14	30	4.8	3.5	8.4	6.5	5.9	7.9	−0.89	−0.30	0.53	−0.06
ILE	45	74	33	63	97	41	31	24	2.4	1.6	4.5	5.3	4.1	8.3	−1.03	−0.71	0.92	0.61
MET	59	93	33	85	110	51	18	55	1.4	1.3	1.7	2.8	2.9	2.4	−0.23	0.19	0.79	0.37
ASP	66	80	42	70	87	39	9	−7	7.7	8.1	6.5	4.8	5.2	3.9	0.21	0.29	−0.44	−0.52
GLU	83	99	49	84	107	50	8	2	10.8	12.1	7.2	7.1	7.3	6.7	0.52	0.09	−0.51	−0.08
LYS	99	116	62	110	125	65	8	5	12.9	14.3	9.2	7.7	8.9	4.8	0.45	0.62	−0.48	−0.66
ARG	94	120	58	114	143	69	19	19	8.6	8.8	7.9	10.9	11.6	9.1	0.11	0.25	0.27	0.13
SER	53	66	31	58	74	31	12	0	6.1	6.6	4.8	5.8	6.1	5.2	0.30	0.17	−0.07	0.07
THR	54	69	36	63	79	40	14	11	5.5	5.3	6.3	5.4	5.0	6.1	−0.17	−0.19	−0.05	−0.03
ASN	68	86	43	72	90	45	5	5	6.5	6.6	6.2	5.0	5.0	5.1	0.07	−0.03	−0.29	−0.19
GLN	79	97	50	82	108	50	11	0	6.1	6.3	5.4	4.1	3.9	4.7	0.16	−0.19	−0.50	−0.15
CYS	25	40	19	30	38	26	−5	37	0.5	0.3	1.0	0.7	0.4	1.4	−1.20	−1.39	0.17	0.37
HIS	65	88	36	72	103	34	17	−6	2.5	2.6	2.3	3.2	3.5	2.4	0.15	0.39	0.29	0.04
PRO	63	80	34	74	87	36	9	6	5.8	6.3	4.3	4.3	5.3	1.9	0.40	1.01	−0.17	−0.79
GLY	35	45	21	42	51	28	13	33	4.1	4.2	3.9	4.3	4.3	4.2	0.08	0.03	0.02	0.07
AVG	59A^2	80A^2	37A^2	71A^2	92A^2	43A^2	15%	18%										

Accessibility and composition have been computed on the bound dataset of 85 protein structures. Surface residues are those detected by our method and show a 0.97 correlation with previously published amino acid scale. As for interface residues, our amino acid scale has a 0.9 correlation with the amino acid scale of Chakrabarti.[3]
[a]Accessibility of amino acids for either the surface or the interface and decomposed following knob and cleft regions.
[b]Amino acid contributions to the surface area or the interface area.
[c]Propensity for a residue (1) to be part of a knob rather than a cleft or (2) to be part of an interacting knob/cleft rather than a noninteracting knob/cleft.
[d]Indicates the raise of residue accessibility (in percentage) between interacting and noninteracting residues.

amino acid accessibility and composition for both non-IR and IR. First, our scale of amino acid contribution to the surface area (Table IIb: Surface: All) shows a strong correlation of 0.97 with a previously published amino acid scale[3]; a strong correlation of 0.9 for the composition of IR was also observed. Then this amino acid contribution to the surface area was decomposed into knob surface area and cleft surface area. Phe, Tyr, and Trp constitute the aromatic cluster, Ala, Val, Leu, Ile, Met constitute the hydrophobic cluster, Asp and Glu the anionic cluster and Lys, Arg the cationic cluster. Whereas knob surfaces are composed of only 4.5% aromatics and 13.9% hydrophobic residues, cleft surfaces are composed on average of twice as many aromatics (10.7%) and hydrophobic residues (24.5%). For interacting surfaces, aromatics represent 13.2% of knob regions (compared with 4.5% for noninteracting surfaces) and 17.5% of cleft regions (compared with 10.6%). These interacting surfaces show also more hydrophobic residues than noninteracting surfaces, in particular for knobs (20.5% of the surface area versus 13.9% for noninteracting surfaces) and somewhat less for clefts that are already very hydrophobic (27.3% against 24.5%). Furthermore, cationic and anionic residues are shown to be more present in knob than in cleft regions for both IR and non-IR, although knob IR are less anionic (12.5%) than knob non-IR (20.2%). Cleft non-IR are also shown to be more charged (30.8% for both anionic and cationic residues) than cleft IR (24.3%).

To conclude, knob regions are shown to be more charged, less hydrophobic and less aromatic than cleft regions. Furthermore, greater differences of surface contribution are seen between knob IR and knob non-IR than for cleft IR and cleft non-IR. Knob IR resemble cleft non-IR and are shown to be less charged, more hydrophobic and more aromatic than knob non-IR. It is, therefore, easier to distinguish knob IR from knob non-IR than to distinguish cleft IR from cleft non-IR.

To sum up these conclusions, euclidian metrics were computed (see Fig. 6), as proposed by Chakrabarti and Janin,[3] by computing the distance Δf between two compositions f_i and f'_i:

$$(\Delta f)^2 = 1/19 \sum_{i=1}^{20} (f_i - f'_i)^2$$

Finally, this finer description of the surface should help to improve the prediction of protein-binding sites by allowing to compare separately knob and cleft regions.

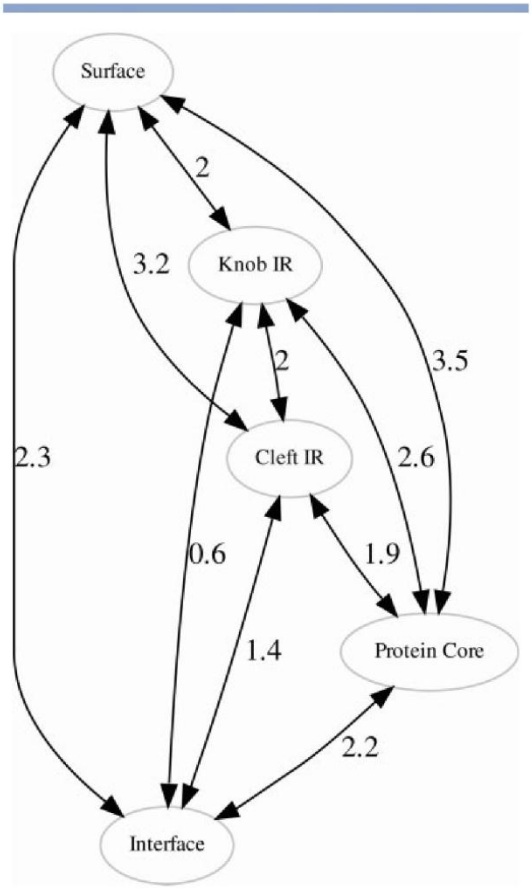


Figure 6

Distances between amino acid compositions. Distances are those defined in the text as Δf and are expressed as percentages. The area-based composition of each region (except for the protein core) is listed in Table II. Area-based composition for the protein core is taken from Lo Conte *et al.*[2] The noninteracting surface is more distant from Cleft IR than Knob IR, and Cleft IR are more similar to the protein core than the noninteracting surface.

Nucleic acids and other biomolecules

With the exception of the definition of surface residues, our approaches can be directly used to analyze and characterize the surface of other biomolecules such as DNA, RNA, or lipids. As defined for protein structures, the surface atoms of these biomolecules still correspond to the accessible atoms extracted from the alpha shape. By construction, our local surface curvature depends only on a continuous surface and is influenced neither by variations of density nor by the presence of cavities. Therefore, this surface descriptor can also be applied to these biomolecules [Fig. 7(A)]. Finally, our approach also allows for the generation of surface patches along nucleic acid surfaces [Fig. 7(B)].

By unifying and facilitating the analysis and comparison of molecular surfaces, we hope that this geometrical approach will benefit the emergent structural studies both on current and newly characterized biomolecules.

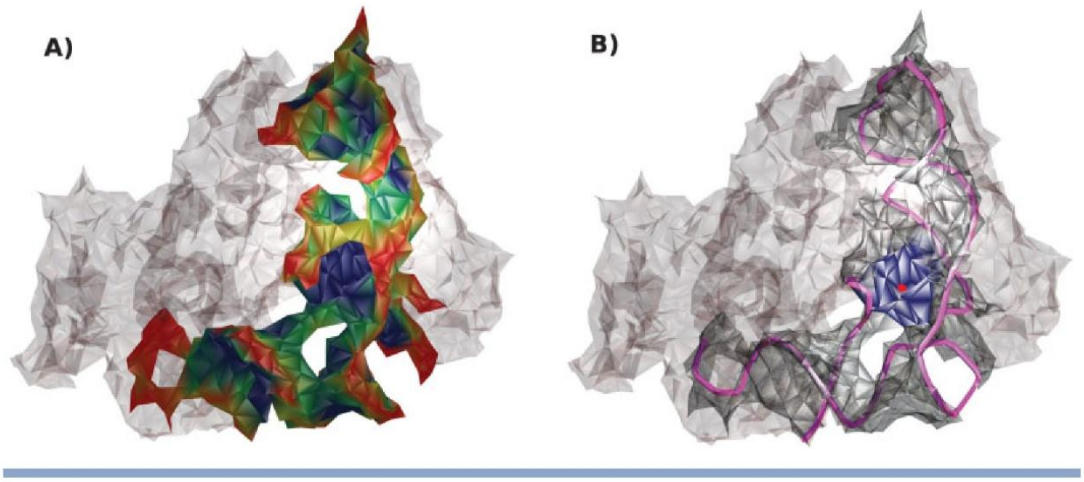


Figure 7

Structure of the Arginyl-tRNA synthetase complexed with the tRNA(Arg) (1f7u). The Arginyl-tRNA synthetase is transparent in the background to precise the orientation of the tRNA. (**A**) tRNA mapped with the local surface curvature following the same color code as in Figure 5. (**B**) The backbone of the tRNA is in mauve, the alpha shape of the tRNA in transparent gray and the edges connecting surface atoms are represented with black lines. A surface patch around O5052 (red point) of U908 is delineated and pictured blue.

CONCLUSIONS

In this study, alpha shapes have been used to model and study properties of protein surfaces that are relevant to the description of the surface and the analysis of molecular interactions. Using this framework, we were able to define surface atoms and residues, as well as to generate contiguous surface patches. Using the field of protein-binding site prediction to evaluate the relevance of our definitions, we achieved a significant improvement in the determination of surface patches, where the signal-to-noise ratio in the definition of the interacting potential of a residue is increased by 2.6 times with respect to a previous approach.

The alpha shape framework was further used to define a conception of surface curvature that is biased neither by the variation of atomic density nor by the presence of cavities below the surface. In the characterization of the molecular surface topography, this conception revealed a landscape composed on average of 38% knobs (the remaining being clefts), where IR are 30% more frequent in knobs than in clefts. This distinction is important for IR as demonstrated by the differences in accessibility and composition between these two regions. These results remain true when considering unbound forms, where only small conformational changes occurred during the assembly formation.

The robust geometric framework of alpha shapes has allowed us to unify the computation of several properties relevant to the analysis and comparison of any molecular surfaces, with a proven improvement compared with former approaches. Our algorithms are fast enough to be used in large-scale projects such as interactomics, and will be applied in the future for the analysis of protein and nucleic acid interactions.

ACKNOWLEDGMENTS

The authors thank Mme Dominique Bechmann (IGG-LSIIT) for cosupervision of Benjamin Schwarz as well as Professor Joël Janin (IBBMC) for his invaluable comments and encouragement.

REFERENCES

1. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
2. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites J Mol Biol 1999;285:2177–2198.
3. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins 2002;47:334–343.
4. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
5. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004;338:181–199.
6. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. Nucleic Acids Res 2006;34:3698–3707.
7. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins 2007;66:630–645.
8. Connolly ML. Measurement of protein surface shape by solid angles. J Mol Graph 1986;4:3–6.
9. Cazals F, Chazal F, Lewiner T. Molecular shape analysis based upon the morse-smale complex and the connolly function. In: Proceed-

ings of the 19th Annual Symposium on Computational Geometry (SCG) 2003. pp 351–360.
10. Norel R, Wolfson HJ, Nussinov R. Small molecule recognition: solid angles surface representation and molecular shape complementarity. Comb Chem High Throughput Screen 1999;2:177–191.
11. Norel R, Lin SL, Wolfson HJ, Nussinov R. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. J Mol Biol 1995;252:263–273.
12. Edelsbrunner H, Mucke EP. Three-dimensional alpha shapes. ACM Trans Graph 1994;13:43–72.
13. Liang J, Edelsbrunner H, Fu P, Sudharkar PV, Subramaniam S. Analytic shape computation of macromolecules I: molecular area and volume through alpha shape. Proteins 1998;33:1–17.
14. Edelsbrunner H, Koehl P. The weighted-volume derivative of a space-filling diagram. Proc Natl Acad Sci USA 2003;100:2203–2208.
15. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
16. Liang J, Edelsbrunner H, Fu P, Sudharkar PV, Subramaniam S. Analytic shape computation of macromolecules II: inaccessible cavities in proteins. Proteins 1998;33:18–29.
17. Edelsbrunner H, Facello MA, Liang J. On the definition and the construction of pockets in macromolecules. Discrete Appl Math 1998;88:83–102.
18. Edelsbrunner H. Deformable smooth surface design. Discrete Comput Geom 1999;21:87–115.
19. Bajaj CL, Lee HY, Merkert R, Pascucci V. NURBS based B-rep models for macromolecules and their properties. In: Proceedings of symposium on Solid Modeling and Applications. 1997. pp 217–228.
20. Zomorodian A, Guibas L, Koehl P. Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials. Comput Aided Geomet Des 2006;23:531–544.
21. Li X, Hu C, Liang J. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins 2003;53:792–805.
22. Cazals F, Proust F, Bahadur RP, Janin J. Revisiting the Voronoi description of protein-protein interfaces, Protein Sci 2006;15:2082–2092.
23. Ban YA, Edelsbrunner H, Rudolph J. Interface surfaces for protein-protein complexes. J ACM 2006;53:361–378.
24. Pintar A, Carugo O, Pongor S. CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics 2002;18:980–984.
25. Janin J. Specific versus non-specific contacts in protein crystals. Nat Struct Biol 1997;12:973–974.
26. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
27. CGAL, Computational Geometry Algorithms Library, http://www.cgal.org.
28. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
29. Connolly ML. Analytical molecular surface calculation. J Appl Crystallogr 1983;16:548–558.
30. Delaunay B. Sur la sphère vide. Izvestia Akademii Nauk SSSR Otdelenie Matematicheskii i Estestvennyka Nauk 1934;7:793–800.
31. Edelsbrunner H. The union of balls and its dual shape. Discrete Comput Geom 1995;13:415–440.
32. Poupon A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol 2004;14:233–241.
33. Nooren IMA, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 2003;325:991–1018.
34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
35. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372:774–797.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
37. Hubbard SJ, Thornton JM. NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College London 1992.
38. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–656.
39. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure 1999;7:723–732.
40. Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. Biophys J 2003;84:2553–2561.
41. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math 1959;1:269–271.
42. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci USA 2005;102:15447–15452.
43. Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry JC, Thompson JD, Wicker N, Poch O. PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Res 2003;31:3829–3832.
44. DeLano WL. The PyMOL Molecular Graphics System. San Carlos, CA, USA: DeLano Scientific; 2002.

III. Conclusions et Perspectives

1. Empreintes structurales

Dans cet article, nous avons exploré la possibilité de générer des régions de tailles variées, mais de formes généralement circulaires puisque les atomes d'une même région sont réunis en fonction d'une distance à un centre. Il est toutefois possible de générer des régions non plus sur un seul critère de distance spatiale, mais en fonction de différentes propriétés géométriques, physico-chimiques ou évolutives :

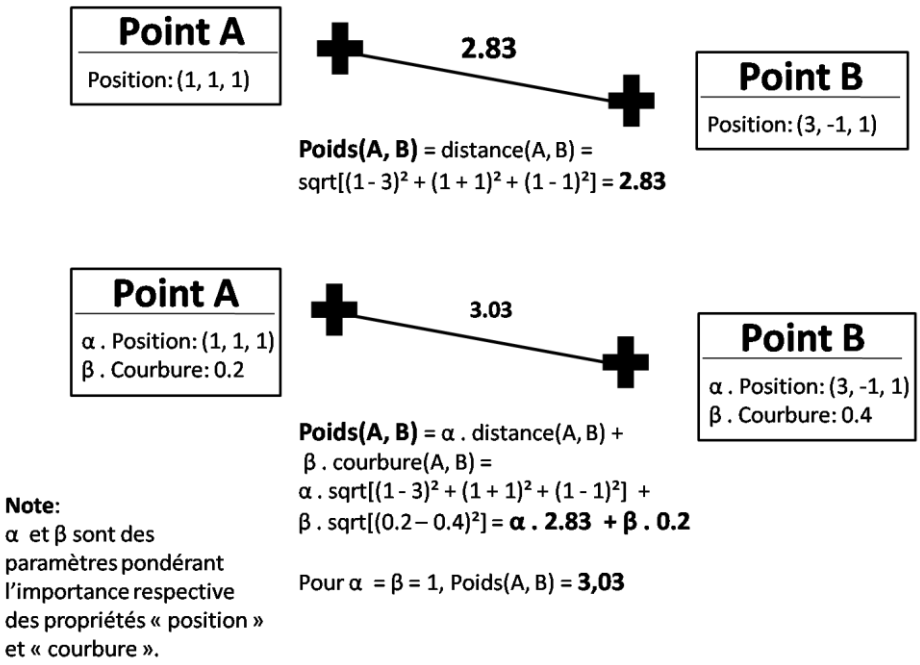


Figure 65. Calcul d'une distance entre deux points, deux atomes ou deux résidus en fonction de N propriétés. *En haut, la distance entre les points A et B est calculée en fonction de leurs seules positions ; il s'agit donc de la distance spatiale les séparant. En bas, nous calculons la distance séparant ces deux points en fonction de leurs positions, mais également de leurs courbures respectives. Cette distance entre A et B tient alors compte de cette propriété pour étendre la région.*

S'il est possible de calculer des distances entre atomes et résidus d'une molécule non plus seulement en fonction de leurs positions dans l'espace, mais en fonction d'autres propriétés telles que la courbure, les charges électrostatiques, ou encore la conservation, alors il devient possible de générer des régions remarquables de la molécule à partir d'un algorithme proche de celui employé dans l'article (Figure 66). Ces régions remarquables caractérisent une ou plusieurs propriétés bien précises (ex : les régions concaves (creuses) et

hydrophobes ; les régions convexes (bosses) et cationiques ou conservées, etc) et sont par conséquent appelées « empreintes structurales » de la molécule.

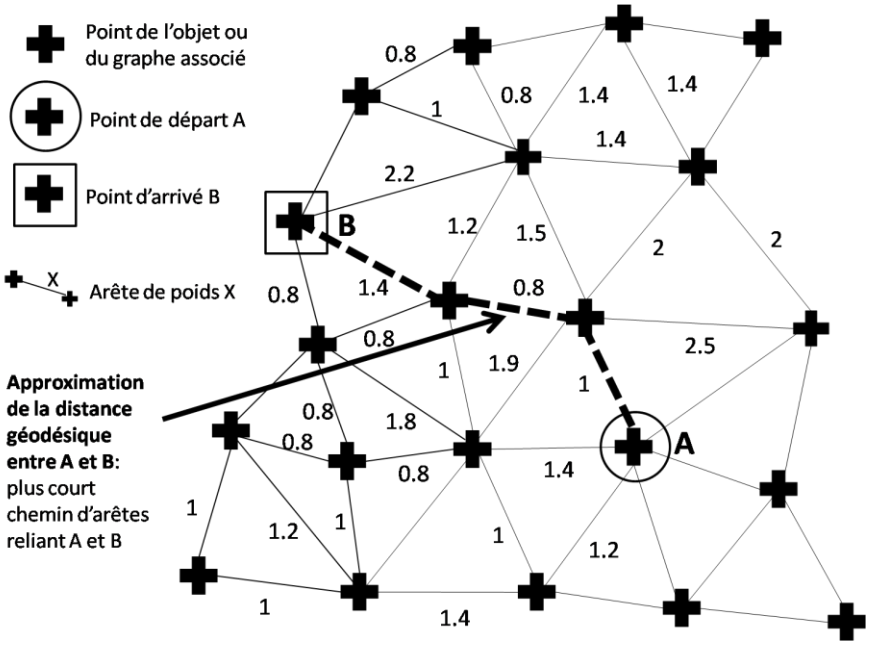


Figure 66. Calcul d'une distance géodésique séparant les points A et B par le plus court chemin d'arêtes. *L'approximation que nous faisons de la distance géodésique (sur la surface) correspond à la somme des distances de chacune des arêtes formant le parcours entre A et B. Ici cette distance géodésique correspond à 3,2. Cette distance peut représenter une distance en termes de propriétés telles que la courbure, des charges, la flexibilité, ou bien encore la conservation.*

Puisqu'elles sont générées sans a priori, ces empreintes structurales sont de tailles et de formes variées. Souvent, elles peuvent mettre en évidence des régions partagées par une famille ou une sous-famille de molécules. Elles peuvent donc notamment servir à la classification des molécules.

2. Régions intermédiaires et intérieures

Cet article était dédié à l'analyse des surfaces moléculaires, aussi nous n'avons traité que la génération des régions de surface. Toutefois, en utilisant le complexe alpha (voir chapitre sur la modélisation des surfaces), il est possible de modéliser les parties internes (inaccessibles au solvant) de la molécule. Par un algorithme similaire à celui développé dans cet article, il devient possible de générer des régions internes (ne contenant que des atomes

inaccessibles au solvant), et des régions intermédiaires (des régions contenant des atomes de surfaces et des atomes inaccessibles au solvant).

Cette description des parties internes et intermédiaires des molécules revêt un intérêt particulier pour l'étude des sites de liaison enfouis, tels que les sites de liaison aux ligands de certains récepteurs nucléaires. Ces régions permettent également de mesurer l'empilement et le volume des atomes et des résidus afin de détecter d'éventuelles zones de faibles ou de fortes densités. Plus généralement, elles permettront, comme pour l'étude des surfaces, de moyenner des propriétés géométriques, physico-chimiques et évolutives afin de mieux comprendre notamment les phénomènes d'allostérie entre les surfaces et le cœur enfoui des molécules [281]. En particulier, des travaux préliminaires semblent indiquer que les sites de liaison de la surface moléculaire semblent coévoluer davantage avec le cœur protéique. En découpant le cœur en régions internes, il sera peut-être possible d'affiner davantage ce message et d'observer des chemins de communication allostérique partant d'une région de surface, traversant une région du cœur, pour finalement atteindre une région distante.

Cette notion de profondeur [242] est une alternative à la distinction des résidus de surface et du cœur d'une molécule qui ne requiert pas le calcul de l'accessibilité relative des résidus. La modélisation des parties internes de la molécule par le complexe alpha, rend également possible le calcul de la distance du plus court chemin d'arêtes séparant un atome enfoui de n'importe quel atome de surface. En d'autres termes, à partir d'un graphe représentant les atomes et leurs voisinages (Figure 66), il est possible de déterminer la profondeur de tout atome comme étant le plus court chemin le reliant à un atome de surface.

3. Régions sous contraintes de formes

A l'exception des empreintes structurales, les régions de ces molécules sont donc majoritairement circulaires, ce qui ne reflète pas l'ensemble de la diversité géométrique des sites de liaison observés.

Afin de mieux refléter cette diversité biologique, il est possible de générer des régions de tailles et de formes variées en appliquant des vecteurs de contraintes favorisant ou pénalisant la croissance d'une région dans une ou plusieurs directions (Figure 67).

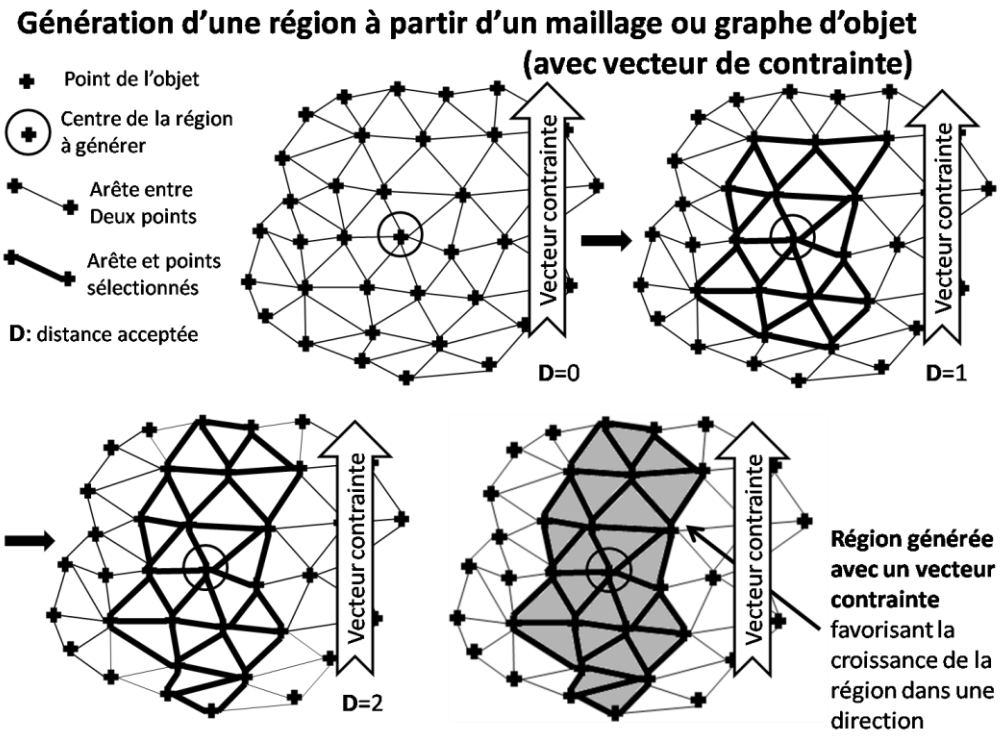


Figure 67. Principe de la génération de régions sous contraintes de formes. *Ici, une région est générée à partir d'un point central indiqué par un cercle et sous la contrainte d'un vecteur direction indiquée par une flèche. La distance séparant deux points (ou atomes) de cette représentation est alors augmentée ou diminuée par l'angle formé entre le vecteur contrainte et le vecteur reliant ces deux points. La diminution de la distance des arêtes formant un petit angle avec le vecteur contrainte permettra de générer des régions étendues dans le sens ou la direction du vecteur contrainte.*

4. La complémentarité des sites de liaison de l'interface

Dans ces travaux, nous avons utilisé la forme alpha pour modéliser la surface moléculaire, générer des régions, calculer des courbures et faire ressortir les tendances des acides aminés pour ces régions.

Toutefois, bien que moins attrayante que la surface moléculaire de Connolly, la surface dessinée par les formes alphas présente un intérêt considérable dans l'étude de la complémentarité des sites de liaison d'une interface. En effet, le fait que la surface dessinée par les formes alphas passe par le centre des atomes, plutôt que le long de leurs sphères de Van der Waals, permet d'observer de façon unique cette complémentarité de l'interface et d'observer plus facilement les contacts inter-atomiques entre chaînes interagissantes (Figures 68 et 69).

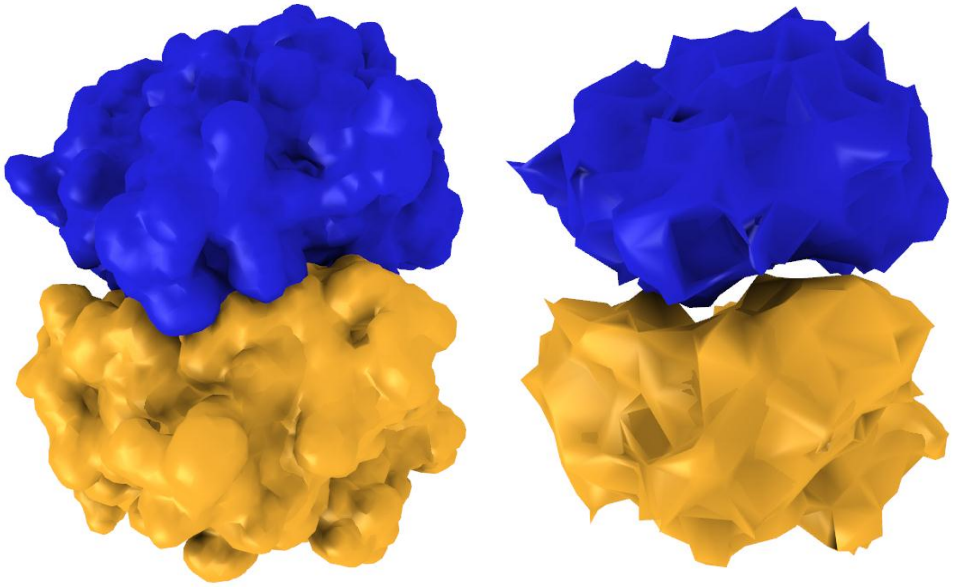


Figure 68. Complémentarité de l'interface de RXRα et RARβ vue par surface moléculaire et forme alpha. *RXRα est ici représentée en bleu et RARβ en jaune. A gauche, la surface moléculaire permet de deviner mais non d'observer la complémentarité de l'interface. A droite, la forme alpha permet d'observer directement la complémentarité de l'interface.*

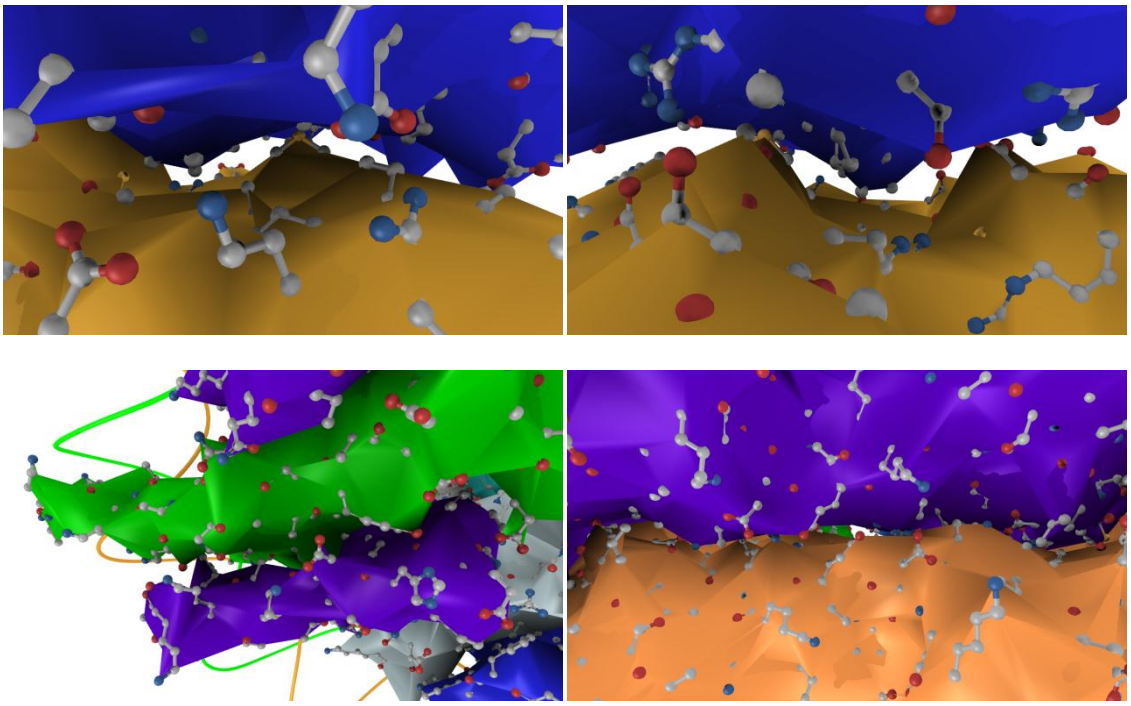


Figure 69. Quatre observations rapprochées d'interfaces moléculaires représentées par forme alpha. *Les chaînes interagissantes ont ici été représentées avec chacune des couleurs différentes de sorte que l'on puisse observer plus facilement la complémentarité des sites de liaisons à l'interface. Le modèle boules et bâtons a également été rajouté afin que l'on puisse directement observer quels sont les résidus participant à une interface.*

Analyse haut débit des interactions moléculaires

I. Problématique : identification des différents types de molécules et d'interactions, et inférence de fonctions par homologie

L'une des composantes importantes de cette thèse a consisté à développer des outils de modélisation et d'analyses capables d'identifier, de différencier et de traiter l'ensemble des différents types d'interactions moléculaires détaillées précédemment. Cette analyse intégrative des données structurales permet de :

- classifier les structures en fonction de l'environnement qu'elles représentent, de sorte que pour une protéine donnée, il soit possible d'observer son comportement moyen dans un environnement donné, et de quantifier les changements qui s'opèrent sur la structure entre le passage d'un environnement à un autre (voir « contexte moléculaire » dans l'article M-ORBIS [1]).

- procéder à l'analyse comparative des structures afin d'identifier rapidement tous les types d'interfaces moléculaires, et de mieux caractériser la surface interagissante et la surface non-interagissante [1, 4]. C'est ainsi que nous avons pu mettre en évidence l'existence de sites de liaisons polyvalents (capables de lier plusieurs types différents de molécules), et des sites spécifiques (capables de ne lier qu'un seul type moléculaire).

- déterminer que la limite inférieure du pourcentage moyen de la surface interagissante est de 44%. Ces 44% correspondent à la surface protéique impliquée dans l'ensemble des interactions protéine-protéine, protéine-peptide, protéine-petit peptide, protéine-ADN, protéine-ARN et protéine-ligand. Sur des jeux d'assemblages moléculaires ne présentant qu'un seul de ces types d'interaction, la surface interagissante était auparavant évaluée à 20 ou 25%.

- déterminer que plus de 86% de la surface protéique peut être hydratée de façon transitoire, mais qu'en moyenne, seule 15% de la surface semble interagir de façon spécifique avec les molécules d'eau.

Type d'interaction	Pourcentage de la surface protéique
Homodimère	26,5
Héterodimère	28,2
Protéine-peptide	16,9
Protéine-petit peptide	10,7
Protéine-ADN	20,6
Protéine-ARN	19,8
Protéine-ligand	14,7
Protéine-eau	15,1
Surface interagissante	43,9

Tableau 25. Les différents types d'interactions et leurs représentations sur la surface protéique. *(adapté à partir de Albou et coll., 2010 [1])*

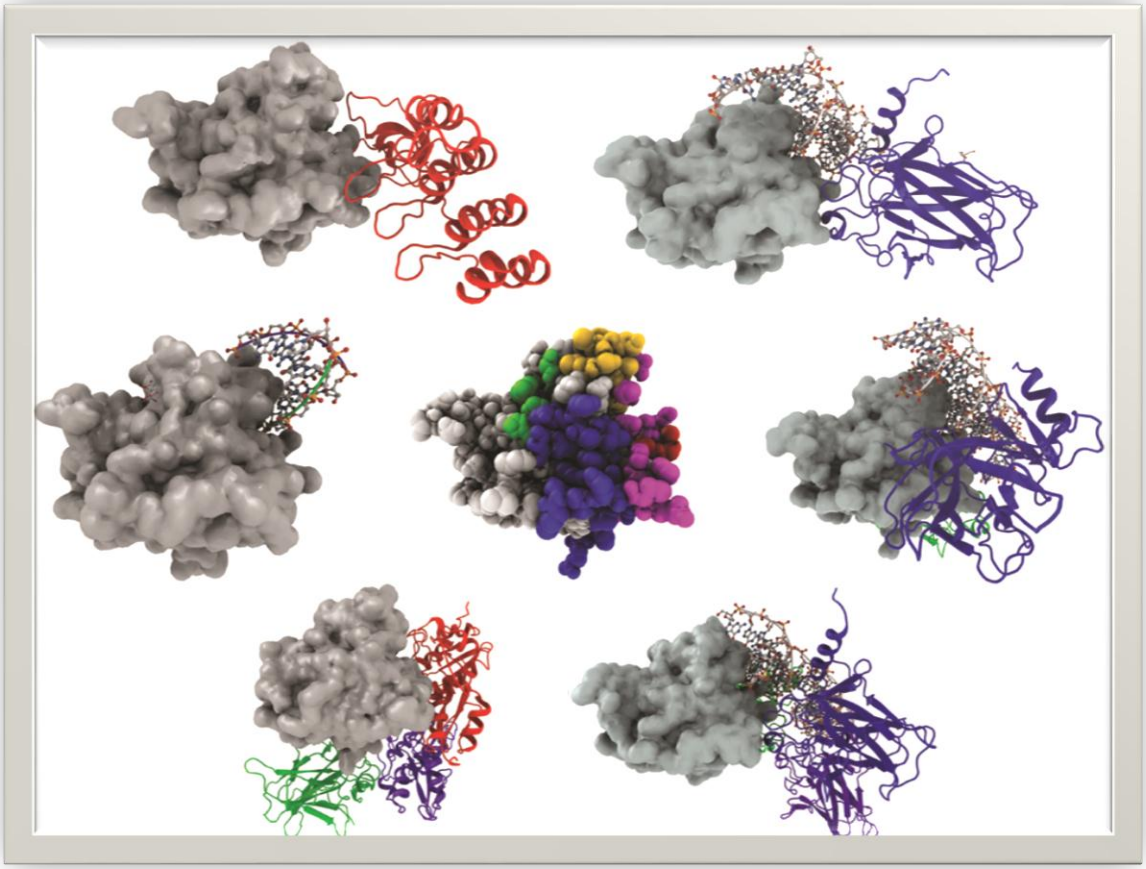


Figure 70. Principe de la cartographie moléculaire des sites de liaisons illustrée avec la protéine p53. *En bleu et vert sont représentés respectivement les sites de liaisons pour la formation d'homodimères et d'assemblages protéine-ligands. En jaune, les sites de liaisons pour la formation d'assemblages protéine-acide nucléique, et en mauve, les sites de liaisons polyvalents. La figure a été proposée comme couverture au journal Nucleic Acid Research en 2010.*

II. Article : M-ORBIS : Mapping of mOleculaR Binding sItes and Surfaces

30–43 Nucleic Acids Research, 2011, Vol. 39, No. 1 *Published online 2 September 2010*
doi:10.1093/nar/gkq736

M-ORBIS: Mapping of mOleculaR Binding sItes and Surfaces

Laurent-Philippe Albou[1,2], Olivier Poch[1,*] and Dino Moras[1]

[1]Department of Biology and Structural Genomics, IGBMC, Illkirch, 67404 and [2]Department of Structural Bioinformatics, BIONEXT, Boulogne Billancourt, 92100, France

Received May 3, 2010; Revised July 31, 2010; Accepted August 3, 2010

ABSTRACT

M-ORBIS is a Molecular Cartography approach that performs integrative high-throughput analysis of structural data to localize all types of binding sites and associated partners by homology and to characterize their properties and behaviors in a systemic way. The robustness of our binding site inferences was compared to four curated datasets corresponding to protein heterodimers and homodimers and protein–DNA/RNA assemblies. The Molecular Cartographies of structurally well-detailed proteins shows that 44% of their surfaces interact with non-solvent partners. Residue contact frequencies with water suggest that ~86% of their surfaces are transiently solvated, whereas only 15% are specifically solvated. Our analysis also reveals the existence of two major binding site families: specific binding sites which can only bind one type of molecule (protein, DNA, RNA, etc.) and polyvalent binding sites that can bind several distinct types of molecule. Specific homodimer binding sites are for instance nearly twice as hydrophobic than previously described and more closely resemble the protein core, while polyvalent binding sites able to form homo and heterodimers more closely resemble the surfaces involved in crystal packing. Similarly, the regions able to bind DNA and to alternatively form homodimers, are more hydrophobic and less polar than previously described DNA binding sites.

INTRODUCTION

A widely used approach in Biology/Bioinformatics is to detect patterns, identify their functions and to use these patterns to gain knowledge on unknown systems. Approaches such as BLAST, or the PROSITE or Pfam databases (1–3), are now commonly used to infer and annotate molecular functions based on sequence comparisons. Similarly, the comparison of protein structures reported and summarized in databases such as CATH and SCOP (4–6) have also been widely used to classify proteins into families and subfamilies, to infer functions and to give insights into the landscape of macromolecular folds available in the cell. The detection of common patterns can also serve other purposes such as the modeling of molecular structures by homology (7) that requires structural templates to function correctly. The last decades of research in experimental and computational structural biology have been mainly devoted to the analysis, characterization and prediction of protein structures and protein assemblies. With structural genomic projects and the work realized by structural biologists, the trend is moving increasingly towards the structural characterization of proteins and nucleic acids as functional and dynamic objects by predicting protein, DNA or peptide binding sites (1–4), by studying intrinsic variability (5) or by studying local differences between unbound and bound forms (6,7). To help in the prediction of protein–protein binding sites Porollo *et al.* (3) for instance proposed to retrieve the homologous structural chains. Some time before, Chung *et al.* (8) also proposed to predict binding sites by localizing the residues which were structurally conserved in several homologous structures. More recently, a Japanese group developed a database of interaction sites also based on the inference of binding sites by homology (9).

Nevertheless, the robust and systematic retrieval of homologous structures in specific 'molecular contexts' (structures sharing a same set of interaction types, such as protein–protein, protein–DNA, protein–ligand, etc.) as well as the identification and mapping by homology of all the different types of binding sites onto a single protein has not yet been investigated. It allows to characterize both the molecule and its binding sites in an integrative and systemic way by extracting their properties, dynamics and functions from the different sets of structures each reflecting a 'molecular context'. In particular, it allows to evaluate if a region of a molecule is able to bind several

*To whom correspondence should be addressed. Tel: +33 3 88 65 32 94; Fax: +33 3 88 65 32 76; Email: poch@igbmc.fr

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

distinct partners of similar or different molecular types. Several major issues have to be considered carefully: the first problem is to avoid the systematic selection of non-specific binding sites due to crystal packing (10) as they can represent as much as 50% of all interactions detailed in structural databases (11). Indeed, although several protein quaternary structure databases exist (12,13) and some methods differentiate very well between biologically meaningful interfaces and crystal packing interfaces (14–17), it is still difficult to have access to a robust and automated process that tests each interface and gives full and easy access to the structures of biological assemblies. The second issue is the automatic identification of the different molecular types present in a structure file, and the distinction between each interface type, including the differentiation of true heterodimers (different molecules interacting), and false heterodimers (interaction between fragments of a same molecule). The work of scientists such as J. Janin, J.M. Thornton, S. Jones or R. Bahadur (18–23) has clearly shown the existence of distinct interface families, which suggests distinct binding site families. For instance, homodimeric binding sites are usually shown to be more hydrophobic and less planar than heterodimeric binding sites (18). As for DNA and RNA binding sites which necessarily reflect the negatively charged phosphate groups of nucleic acids, they are far more cationic than any other known binding sites (19,22). The last-but not least-issue to be considered, is that a structural homology at a global scale does not necessarily imply the same function at a local scale. For instance, even if two molecules share a global shape with a very low RMSD or a very high sequence identity, the mutation of a single residue at a binding site can still drastically change its functions (24–26).

The aims of this work are first, to propose a fully automated, yet robust and coherent approach named M-ORBIS (for Mapping of mOleculaR Binding sItes and Surfaces) to extensively describe a molecule in specific 'molecular contexts' from the scattered structural data; and second, to give some insights into the general properties and behaviors of molecular surfaces and interactions. The global and extensive mapping of a molecule in specific 'molecular contexts' (describing a precise set of interaction types), has been named 'Molecular Cartography', as it gives a very detailed functional and dynamic representation of molecules. This definition of 'molecular contexts' is used to classify the retrieved structures into groups illustrating—like snapshots—some of the dynamic events of the studied molecule. As each interaction contained in each of the structural homologues is analyzed and stored, M-ORBIS also allows to transfer the binding site locations and corresponding partners onto the studied molecule. This inference is based on several sequence and structural criteria to ensure sufficient similarities at the global and local scales. Furthermore, as M-ORBIS exploits both sequence and structure alignments, it allows to quantitatively and qualitatively analyze the change of conformations between any two molecular contexts.

The M-ORBIS approach has been validated on four curated datasets (18,19,22,27) describing different interface types and demonstrates sensitivity and specificity >90%. Next, it has been used to demonstrate the existence of binding sites specific to a particular molecular type, and polyvalent binding sites which can bind two or more different molecular types. Interestingly, polyvalent binding sites exhibit amino acid compositions that are intermediate between the specific binding sites they represent. Specific homodimer binding sites are nearly twice as hydrophobic and are two times less charged than polyvalent homodimer binding sites. Our results reveal that at least 44% of the protein surface is designed to interact with non-solvent/ion partners. The characterization of protein–water interactions at different contact frequencies (observed in homologous structures) also suggests that ~86% of the surface can be transiently solvated, whereas only 15% appears to be specifically solvated.

MATERIALS AND METHODS

Structural databases and crystal contacts

Structural data are deposited in the Protein Data Bank (28) and are both easily accessible and retrievable. For some proteins, generally of therapeutic or cosmetic interest, there exists more than 500 structures of the same molecule (e.g. kinases), corresponding to specific environments (specific sets of interactions with different partners of different molecular types).

For the identification and mapping of non-specific binding sites due to crystal packing, another structural database, the Protein Quaternary Structure (PQS) database is used (12). PQS is based on the PDB but attempts to differentiate between specific and crystal packing interfaces based mainly on the buried ASA (Accessible Surface Area) observed at each interface and a solvation energy term. The classification error rate of PQS for the prediction of the oligomeric state of proteins is ~16% (11,29).

Summary of the workflow

In order to compute a Molecular Cartography, the M-ORBIS process requires only the selection of a structural chain as input data. This can be done either using the M-ORBIS command line and providing a PDB file and both the molecule and chain to be treated, or using the MSVM (molecular structure visualization and mapping) graphical-user interface. Once the structural chain is selected, M-ORBIS will process the data in a 7-step workflow described in Figure 1. Finally, M-ORBIS generates an output matrix file which is a specialized multiple alignment of structures storing several physico-chemical and geometrical properties for each analyzed residue. Scripts in Java to analyze the M-ORBIS output matrices are available upon request. Each of these steps is described in more detail in the following subsections.

Alignment and retrieval of homologous structural chains

A protein chain is used as input to a local version of the PipeAlign platform (30), in order to automatically construct a high quality hierarchical multiple alignment

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

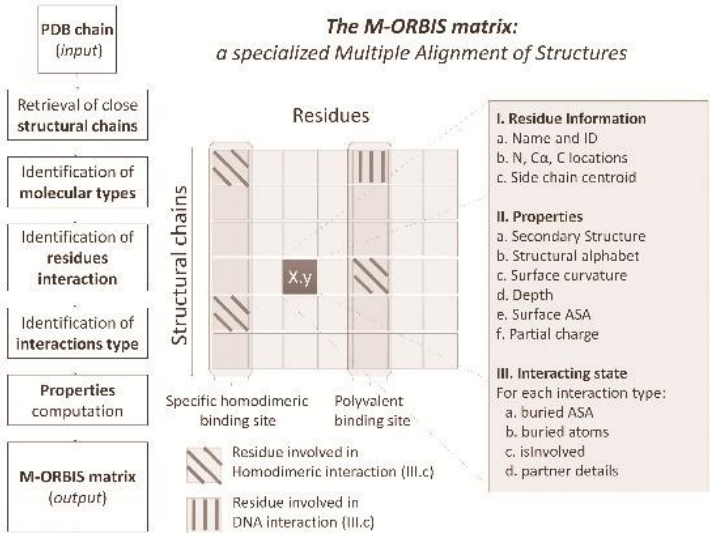


Figure 1. The M-ORBIS matrix as a specialized multiple alignment of structures. The M-ORBIS output matrix is obtained via a 7-step workflow illustrated in the left part of the figure. As in traditional multiple alignments of sequences or structures, the M-ORBIS matrix gives access to every homologous chain and residue. Nevertheless, to perform the complex task of Molecular Cartography, M-ORBIS also stores information relative to the residue locations (field I.), to the physico-chemical and geometrical properties (field II.) and to the interacting state (field III.) of each analyzed residue. A cell of the M-ORBIS matrix is illustated in the right-hand table and gives insights into some of the properties that are stored. For instance, field III.c indicates that the first residue is involved in only homodimeric interaction in two of six related structural chains. A polyvalent binding site is detected when for a given residue (column), the corresponding residues of related structural chains are involved in at least two different binding site types (residue described in column 5). A specific binding site is inferred when only one binding site type is detected at that position (residue described in column 1).

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

of sequences related to the query. To enhance the alignment quality, a maximum of 50 non-structure sequences that share <95% sequence identity (31) with the protein chain are introduced; all other sequences in the multiple alignment correspond to structure sequences derived from PQS. Once the alignment is obtained, only structures of resolution 3 Å or better are kept to ensure a high-quality analysis. Also, structures with <30% sequence identity or 75% residues aligned with the query are excluded. For each sequence alignment between the query and a retrieved structure, a structural alignment is also performed, using the CE algorithm (32) with default parameters. Together, the sequence and structure alignments combined are used in the detection of conformational moves and in the inference of functional binding sites by homology. In this study, a necessary but not sufficient criterion for two related residues to share the same interacting states is an observed distance of <1.5 Å between their two aligned Cα.

Further selection of structural chains is then achieved by M-ORBIS, by retrieving a subset of these stored chains using user criteria such as percentage of sequence identity, the presence of interaction types (depicting a molecular context), the presence of solvent molecules, etc. Figure 1 illustrates the properties stored in the M-ORBIS output matrix, for each residue of each analyzed chain, in particular their interacting status. The scripts developed in Java parse and analyze this output and allows for instance to retrieve the chains involved in protein–DNA interactions by searching for the chains that contains residues involved in protein–DNA interaction (here, first line). Similarly, it allows to retrieve the chains involved in at least homodimeric interaction (here, lines 1, 4 and 5), or exclusively homodimeric interaction (here, lines 4 and 5).

Differentiating molecular types

The MSVM research platform (http://www.bio-next.com) allows to read PDB files and to automatically differentiate between protein, peptide, small peptide, DNA, RNA, ligand, ion and solvent molecules. Each molecular type is automatically defined on the basis of written conventions defined by the PDB and the IUPAC code and is hierarchically managed via MSVM.

DNA residues are detected using both the old and new written conventions (+/D). Nevertheless, some differences in conventions are still observed in structural databases and may lead to some errors in the definition of molecular type (e.g. in the PDB 1AIS, DNA are coded with DG, DC, DT, DA, whereas in PQS, the same residues are coded with G, C, T, A). When using jointly PDB and PQS files, this problem is handled by defining the molecular types based solely on the PDB file and by further mapping this definition onto PQS chains.

Proteins, peptides and small peptides are differentiated on the basis of their size: chains composed of more than 60 amino acids are defined as protein, between 20 and

60 amino acids as peptide and as small peptides otherwise. This differentiation is important since small peptides can act as regulating keys (e.g. co-factor) of biological processes, and peptides which do not usually possess a stabilizing protein core may have several different conformations.

Ligand molecules are then defined as being the remaining unknown residues in PDB files. They are further characterized by using the database of 'monomers' provided by the PDB.

Residue interaction detection

The interactions between protein, peptide, small peptide, DNA and RNA residues are detected based on a change of ASA. A residue is considered to be interacting if it loses at least 1 $Å^2$ ASA during the assembly formation (18,20,21,27). ASA values were computed with the NACCESS program (33) and default parameters. For a few of these macromolecules (like the 70S ribosome subunit: 1vp0), NACCESS was not able to compute an ASA; therefore, the detection of interacting residues was inferred by distance criteria as described below.

For ligand, solvent and ion interactions, which may involve buried residues of the protein core, residue interactions are detected on a distance criteria basis. A protein–ligand interaction is observed if at least two atoms of the ligand and two atoms of the protein are nearer than their sum of van der Waals weighted by an uncertainty factor of 1.4. For instance, a C–O contact is observed if the distance separating these two atoms is less than $(1.7+1.5)\times 1.4$, i.e. 4.48 Å. Nevertheless, and as previously described (34), this cutoff may not be restrictive enough to detect specific protein–water interactions. Therefore, these particular interactions are identified if two polar atoms (N, O, S) are within 3.5 Å.

Differentiating interface types

A first level of distinction for molecular interactions is between a homodimer (assembly of identical molecules) and a heterodimer (assembly of different molecules). Protein homodimers are detected if the chains involved in the interaction share the same Uniprot ID (DBREF field of PDB file), while heterodimers are detected if the chains involved have distinct Uniprot ID. For nucleic acids and in the cases where no Uniprot ID are available, chains involved in homodimeric interactions must have >90% sequence identity and chains involved in heterodimeric interactions must have <40% sequence identity. Other cases are considered uncertain and are discarded by M-ORBIS. As PQS does not conserve the DBREF section of PDB, all chains present in a PQS file structure are assigned a chain in the PDB file structure in order to have access to this information.

The interactions can then be classified, according to the types of molecules involved, into either protein–protein, protein–peptide, protein–small peptide, protein–DNA, protein–RNA, protein–ligand, protein–ion, protein–solvent, protein–crystal packing and/or protein–fragment. As each molecule has a molecular type identified automatically, each of the detected interactions is also assigned a type automatically. True heterodimers and interactions of the same fragmented molecule are differentiated when the UniprotID information is available: if two interacting proteins have the same Uniprot ID, but are non-overlapping fragments of the full sequence protein, then they represent a protein–fragment interaction.

To take into account the diversity of interaction types considered and to further differentiate between crystal packing and biological interfaces, a few simple criteria are added: the minimum buried ASA for homodimeric and heterodimeric binding sites are set to 450 $Å^2$ and 350 $Å^2$, respectively, while for peptide, DNA and RNA binding sites, we mainly discard artifacts by removing interfaces of <50 $Å^2$. Furthermore, each of the binding sites considered for homodimer, heterodimer, peptide binding sites must contain at least 10 residues, while binding sites for small peptides and ligands are simply required to have more than one interacting residue.

Phosphorylation or glycosylation sites are also mapped on the structure, using either the written conventions in PDB that indicate a phosphorylation, or using Uniprot information (35). For instance, phosphorylated serines (SEP), threonines (TPO) and phosphotyrosines (PTR) are identified from the PDB files and are used to locate the phosphorylation sites.

Molecular contexts

Each structural chain is involved in different types of interaction (protein–protein, protein–DNA, protein–RNA, protein–peptide, protein–small peptide, protein–ligand, protein–ion, protein–water) and can be assigned to a precise molecular context that can be further analyzed and compared to other molecular contexts in order to average, characterize and understand the dynamic events between two sets of contexts. More precisely, a specific molecular context is defined by selecting the structures with a requested set of interaction types and using different logical operators (Figure 1): at least one of the interaction type must be present (OR); at most one interaction type must be present (exclusive OR); all selected interaction types must be present but other are accepted (AND); only but all the selected interaction types must be present (exclusive AND); all interactions are accepted except the selected ones (NOT).

As a consequence, it is possible to examine several geometrical and physico-chemical parameters either for the structures of a same molecular context (to observe intrinsic variability) or for two sets of structures reflecting two different molecular contexts.

Contact frequencies and solvated surfaces

The residue contact frequency $f_{contact}$ is a general measure that describes the fraction of interacting residues (for a specific interaction type) at a precise residue position and for a given molecular context (several related chains). Figure 1 illustrates how the $f_{contact}$ is derived from the M-ORBIS matrix output file. The first residue which is seen involved in homodimeric interaction in two related structures has a $f_{contact}$ of 2/6 for this particular

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

type of interaction, whereas the fifth residue which is seen involved in one homodimeric interaction and one protein–DNA interaction in all related structures has a 1/6 $f_{contact}$ for homodimer interaction and a 1/6 $f_{contact}$ for DNA interaction. Thus, the $f_{contact}$ can be used to indicate: (i) whether the residue is involved in a specific or polyvalent binding site and (ii) if the residue is frequently involved in a given type of interaction. In this context, the SWD_{10} and SWD_{90} parameters shown in Table 2 and computed on the SWD dataset (see below), describe the percentage of surface residues always involved ($f_{contact}$ of 90%) or occasionally involved ($f_{contact}$ of 10%) in a given type of interaction. As a consequence, the ratio SWD_{90}/SWD_{10}, indicates the fraction of interacting residues that is always seen interacting for a given type of interaction.

When a sufficient number of homologues is available, the $f_{contact}$ measure indicates whether the residue is specific or not for the given interaction type. In this study, unless otherwise stated, the $f_{contact}$ measures were computed on the SWD dataset, which is composed of proteins that are each described by at least 50 different structural chains. The solvating state of a residue is inferred from this residue contact frequency, although some more stringent criteria for the selection of related structural chains are added: (i) as the average number of water molecules observed per structure is correlated with the crystallographic resolution, only structures with a resolution between 1.5 and 2.5 Å were considered, (ii) structures containing the related chain must contain at least five water molecules to ensure they have been at least partially considered by the experimentalist. A residue is then considered as transiently solvated if it is in contact with water in at least 10% of the related structural chains. A residue that interacts with water in >90% of the related structural chains is considered to be specifically solvated.

As the contact frequency is dependent of a molecular context, it is possible to describe the variation of these contact frequencies for different sets of contexts. For solvation, this proves to be important as it allows to observe the change of residue solvation upon different assembly formation (protein–protein, protein–peptide, etc.).

Inference of the interacting state

Given a chain, it is possible to infer the interacting states of each of its residues by observing the interacting states of its aligned residues on related chains. M-ORBIS uses four main criteria to infer the interacting states of a residue from homologues: (i) the percentage of sequence identity between the studied chain and the related chain; (ii) the percentage of sequence identity for all the residues involved in the given interaction; (iii) a $f_{contact}$ value for the given interaction types; (iv) the distance between the Cα of aligned residues. These four parameters are highly dependent on the molecular context as it determines the chains selected.

As for the other criteria described previously, the selection of chains also depends on resolution criteria (e.g. the selection of structures with resolutions between [R – deviation] Å and [R + deviation] Å). In some cases, as for the study of protein–water contacts, it is preferable to select only the structures with at least five water molecules and to discard the structures in which the water was not or only partially resolved. These selections are available to users via the MSVM research platform and the M-ORBIS module.

Structural datasets

Four published non-redundant datasets (18,19,22,27) representing different types of interaction as well as the protein–protein docking benchmark 3.0 (7) are used throughout this study. These datasets are composed of structures extracted from the PDB and PQS using several criteria and are further curated by checking for crystal contacts, biological units and in some cases the literature. The docking benchmark dataset differs from other datasets by describing both a bound and unbound form for each protein.

A structurally well-defined dataset (SWD) of 102 proteins has been constructed by merging the four curated datasets and keeping only the protein chains with more than 100 residues that have at least 50 known structural homologues in the PDB. The resulting dataset is composed of proteins having length varying from 107 to 796 amino acids.

RESULTS

In the following, we use the term 'non-interacting surface' to refer to the surface involved only in crystal packing, solvent or ion interactions, while the 'interacting surface' will refer to the different binding sites. Interacting residues will be referred to as IR and surface residues as SR.

Validation of the interaction analysis

The M-ORBIS Molecular Cartography of a molecule includes several annotations for each residue of each related chain analyzed (Figure 1). In particular, the interacting state of a residue, as well as its ASA values describing both its accessibility to solvent and its buried ASA are stored. The annotations present in the curated datasets (ASA and interacting state) were then compared to those provided by M-ORBIS (Supplementary Table S1). For each of the three tested datasets, the sensitivity and specificity of SR and IR detection were ~100%. The slight differences observed in IR sensitivity and specificity could be explained either by the choice of minimal buried ASA to detect IR or by the rare but wrong assignment of a modified nucleic acid residue to another molecular type. Surface ASA and Buried ASA values for both curated analysis and M-ORBIS also have a near perfect correlation ($correl^1$ columns). With the exception of protein–RNA assemblies, $correl^2$ indicates that both surface ASA and buried ASA can be accurately predicted from the analysis of structural homologues. It also suggests that when a residue is involved in a particular interaction type, its buried ASA is globally conserved over its family.

Molecular context assignments

For each structural chain analyzed in M-ORBIS, the set of interacting residues and partners can be easily retrieved

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

and it is possible to define a molecular context according to the types of interaction the chain is involved in (see 'Materials and Methods' section). The molecular context assignment was evaluated on the four curated datasets (18,19,22,27), and results show that with the exception of the DNA dataset, 100% of the interactions described manually were correctly characterized by M-ORBIS. In the case of the DNA dataset, six structures (1asy; 1gtr; 1zdi; 1urn; 1ttt; 1ser) were identified as participating in protein–RNA interactions rather than protein–DNA interactions but the consultation of the structures proves M-ORBIS to be right, where the nucleic acids were mainly tRNA (1asy; 1gtr; 1zdi; 1ser).

The ability to correctly define a molecular context from a structure was further tested on the docking benchmark 3.0 dataset (7) where for each protein chain, both bound and unbound forms are described. Among the 124 assemblies, 191 partners were described as single protein chains (not as a group of chains). Starting from these 191 single protein chains in bound forms, M-ORBIS was able to find (by searching for homologous chains not involved in protein–protein, protein–peptide, protein–DNA or protein–RNA interactions) 155 of the corresponding unbound chains described in the article (81% accuracy). Another 11 chains from the benchmark 3.0 were described by M-ORBIS as participating in either protein–protein or protein–peptide interactions and were therefore not considered as unbound forms. Nevertheless, these interactions were present in these 11 structures and the M-ORBIS analysis was correct according to our unbound definition, thus raising our accuracy to 87%. The remaining 25 unbound chains described in the benchmark 3.0 but not found by M-ORBIS were due to three problems: (i) a change of PDB name, (ii) a change of chain name between the PDB and PQS files (due to the adding or removal of chains needed in PQS) and (iii) the non retrieval of the PDB chain by M-ORBIS.

Binding site inferences

In a previous section, we demonstrated that M-ORBIS stores and retrieves the correct mapping for each residue of each related chain. Here, we are interested in the inference of binding sites by homology. The inference of binding sites and putative partners by M-ORBIS is influenced mainly by four parameters described in 'Materials and Methods' section. In the following study, the minimal percentages of sequence identity for related chains and interacting residues are set to 50%; the minimal contact frequency $f_{contact}$ is set to 10%, while the accepted distance variation (in Å) between the Cα of the studied chain and the Cα of the related chains is set to 1.5 Å.

Table 1 illustrates the high sensitivity and specificity of the M-ORBIS binding site inference for each of the four interaction types considered. As the M-ORBIS annotations contains on average all the interacting residues detailed in the curated datasets (Supplementary Table S1), the Molecular Cartography of these datasets (which rely also on other homologous structures to infer binding sites) always describes a larger fraction of the surface as interacting. The fraction of surface involved in a binding site type ($f_{contact}$) was computed by considering only the structures with this type of binding site. By considering the residues described as interacting by M-ORBIS but not by the curated analysis as putative false positives, a lower bound of the binding site inference specificity is determined: 68.3% for heterodimers, and 86, 85.7% for homodimers and RNA, respectively. However, the Pearson product-moment correlations between the amino acid scales extracted from the curated and M-ORBIS analyses indicate that these new interacting residues respect the observed composition bias for each of these interaction types, suggesting they are not false positives. For instance, homodimer interacting residues are still shown to be more hydrophobic and aromatic, whereas DNA and RNA interacting residues are much more polar and cationic. To ensure the selection of only homologous chains and strengthen the inference of binding sites, a more drastic selection of related chains was performed with a minimal percentage of identity set to 90% and similar results with slightly smaller percentages of interacting surface were obtained (data not shown).

The landscape of binding sites and interacting surfaces

The fractions of protein surface occupied by binding sites and solvent were first evaluated on the four manually

Table 1. Comparisons of binding site annotations provided by the curated datasets and M-ORBIS

Datasets	Sensitivity (IR)	Lower bound specificity (IR)	Pearson correlation[a]	Manually curated interface[b]	M-ORBIS interface[c]	Interacting surface[d]
Homodimer (18)	99.7	86.4	0.98	26.8 (12.8)	31.8 (15.4)	33.8 (16)
Heterodimer (27)	96.5	68.3	0.97	19.1 (9)	32.8 (23.6)	39.9 (18.5)
RNA (19)	98.2	85.7	0.99	18.4 (8.3)	21.4 (9.7)	28.2 (16)
DNA (22)	–	–	0.99	–	26.8 (11)	40.3 (26.5)

The sensitivity corresponds to the fraction of residues that are inferred as interacting by M-ORBIS, among all the interacting residues described in each curated dataset. The lower bound specificity is the precision at which a residue described as interacting by M-ORBIS was annotated as such in the curated datasets. Numbers in parenthesis are standard deviations.
[a]A Pearson correlation indicates the resemblance between the IR compositions calculated from the curated and M-ORBIS analysis.
[b]The fractions of interacting surface described by the curated analysis.
[c]The fractions of interacting surface described by M-ORBIS
[d]The interacting surface corresponds to the surface involved in at least one non-solvent interaction.

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

curated datasets. With a $f_{contact}$ of 10%, the interacting surface is shown to vary from 28.2% for the RNA dataset to 40.3% for the DNA dataset (Table 1). However, this first evaluation of the interacting surface suffers from the heterogeneity of the proteins studied: the M-ORBIS approach reveals that some of these proteins have less than ten homologous chains (1kq2:A, 1ser:A, etc.), thus leading to an incomplete mapping of their binding sites, whereas other proteins have more than 200 homologous chains (1us1:A, 1ajs:A, 1amk:A, etc.), leading to a more complete mapping of their binding sites and functions.

To cope with this problem and to strengthen our results, we used the SWD dataset where each protein has at least 50 structural homologous chains (see 'Materials and Methods' section). The results are summarized in Table 2 for different contact frequencies $f_{contact}$; interestingly, the binding site fractions of protein surface are not additive which suggests the existence of some overlap between binding sites (see specific and polyvalent binding sites). For a $f_{contact}$ of 10%, the average interacting surface is 43.9% with the larger fraction of binding sites occupied by heterodimers (28.2%) and homodimers (26.5%), followed by DNA and RNA binding sites with 20.6 and 19.8%, respectively. Small peptide and ligand binding sites occupy the smallest fractions of the protein surface with 10.7 and 14.7%, respectively. We also verified that the average interacting surface described here in terms of residues (43.9%) was indeed similar to the average interacting surface in terms of accessible surface area (45%).

By increasing the minimal contact frequency to a more stringent value, the mapping can be set to emphasize the invariant interacting residues. At 50% minimal contact frequency, 31.9% of the surface is then seen as interacting, while at 90%, only 23.6% of the protein surface is described as interacting. The SWD_{90}/SWD_{10} ratio indicates that 69 and 61% of RNA and DNA binding sites, respectively, are composed of residues that are seen interacting in >90% of related structures, whereas the ligand binding site is composed of only 31% of these frequently interacting residues (see 'Discussion' section).

The non-interacting surface landscape: the transient and specific protein–water contacts

The analysis of protein–water contact frequencies by M-ORBIS indicates that 86.5% of the surface residues are solvated in at least 10% of the structural homologues, whereas with a $f_{contact}$ of 50% and 90%, the fractions of solvated residues decrease to 58.4% and 15.1%, respectively (Figure 2, Table 2). Those residues that are in contact with water molecules in at least 90% of the related structural homologues are considered to specifically bind water molecules.

The analysis of the amino acid compositions involved in protein–water contacts on the SWD dataset shows, for a $f_{contact}$ of 10%, a high correlation of 0.95/0.94 with the non-interacting surfaces previously described in the literature (18,21,27). The correlation with a study dedicated to the hydratation of protein surface and interface (34) is also good, with a 0.88 Pearson correlation. Interestingly, for a $f_{contact}$ of 90%, the first two correlations are decreased to 0.76/0.75, respectively, indicating some differences in the preferences of their amino acids to form contacts with water molecules. If Gly, Ala, Val, Leu, Ile, Pro and Met are considered as hydrophobic residues, and Asp, Glu, Lys and Arg as charged residues, a comparison with the previously published amino acid scales suggests that residues involved in specific water contacts are less hydrophobic (23.5% against 35.6%), and more charged (36.5% against 29.3%) than previously observed. Furthermore, the manual visualization of the SWD suggests that residues involved in specific water contacts are often co-localized with ligand binding sites. This result was partially verified in the following study concerning the polyvalent binding sites and is illustrated in Figure 2.

Specific and polyvalent binding sites

It has been observed in a previous section, that the binding site fractions of the surfaces add up to more than the global interacting surface which suggests some overlap between binding sites. Using the M-ORBIS approach, it was possible to automatically locate the residues that had the ability to alternatively participate in at least two distinct interaction types (Figure 1). More precisely, two binding sites of different types (e.g. homodimer and heterodimer) are said to be polyvalent if at least 10 of their residues overlap and participate in both interaction types. In the case of overlap with ligand or small peptide binding sites which are smaller, only two residues were required.

Two questions are investigated: first, does a specific binding site (a binding site that is seen to be involved in

Table 2. Average functional landscape of protein structures

Dataset	Homodimer	Heterodimer	Peptide	S.Peptide	DNA	RNA	Ligand	Solvant	Int.Surface
SWD_{10}	26.5 (13.5)	28.2 (14)	16.9 (5.9)	10.7 (6.5)	20.6 (8.2)	19.8 (9.4)	14.7 (8.6)	86.5 (11.7)	43.9 (16.1)
SWD_{25}	23.4 (10.9)	21.7 (9.1)	16.6 (5.7)	9.2 (5.1)	18.7 (7.6)	18 (8.3)	9.9 (4.8)	78.5 (14.5)	37.6 (12.8)
SWD_{50}	19.2 (8.5)	17.4 (8)	15.3 (5.4)	8.1 (3.8)	17.5 (7.6)	16.8 (7.6)	7.1 (4.2)	58.4 (17.9)	31.9 (11.1)
SWD_{75}	16.6 (9.2)	14.1 (6.9)	11.9 (5.7)	6.9 (2.8)	15 (6.7)	15 (7)	5.1 (4.3)	30 (16.9)	26.7 (9.8)
SWD_{90}	15.3 (8.9)	11.4 (6.5)	11.2 (5.8)	5.6 (2.5)	12.6 (5.9)	13.6 (5.4)	4.6 (4.4)	15.1 (13.3)	23.6 (9.6)
SWD_{90}/SWD_{10} (%)	58	40	66	52	61	69	31	17	54

Each line described the average fractions of protein surface involved in each interaction type for a given $f_{contact}$. These fractions have been computed on the SWD dataset for $f_{contact}$ of 10, 25, 50, 75 and 90%, respectively. The contribution of a structure to each fraction of interacting surface was considered only when it presented the given interaction type. The ratio SWD_{90}/SWD_{10} indicates the percentage of frequently interacting residues for a given binding site type. Numbers in parenthesis are standard deviations.

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

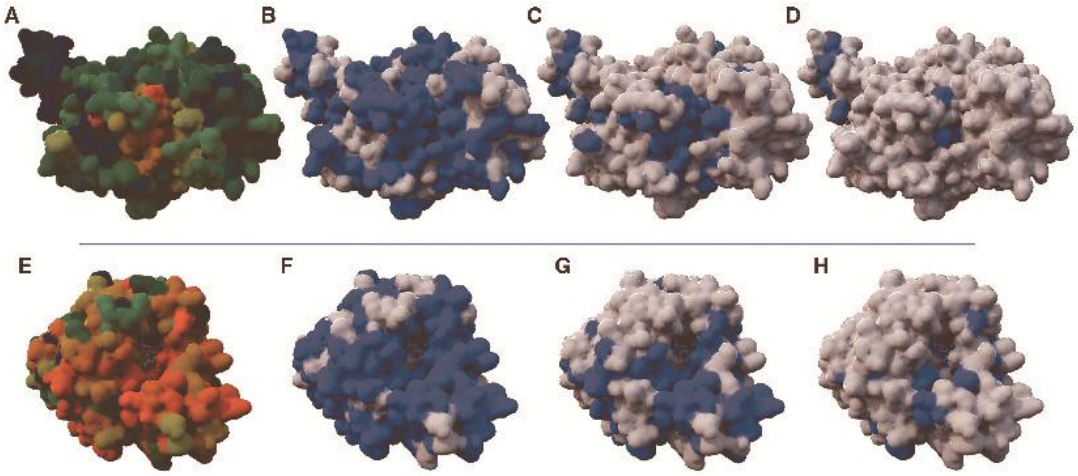


Figure 2. Protein–water contacts at different contact frequencies. (**A**, **B**, **C** and **D**) Structures of the guanine nucleotide-binding protein G(i) (PDB: 1BOF), while (**E**), (**F**), (**G**) and (**H**) are structures of the H1N1 Neuraminidase (PDB: 3b7e). In (A) and (E), the contact frequencies $f_{contact}$ with water molecules are shown, as calculated by M-ORBIS with 49 and 169 related chains, respectively; non-solvated regions are indicated in dark blue ($f_{contact}$ ~0%), while partially solvated regions are in green ($f_{contact}$ ~50%). Red surfaces correspond to the regions that specifically bind water molecules with a $f_{contact}$ ~100%. The surface in contact with water molecules is shown in light blue in (B), (C), (D), (F), (G), (H) with $f_{contact}$ of 25, 50, 75, 50, 75 and 90%, respectively. For these two molecules, the central solvated regions are ligand binding sites.

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

only one interaction type in all homologues) have the same physico-chemical properties as a polyvalent binding site (a binding site that is seen to be involved in at least two different interaction types) (Table 3); second, does a binding site observed to participate in an interaction type have some preference for participation in other interaction types (Table 4).

Differences between specific and polyvalent binding sites. The analysis of observed amino acid compositions between specific homodimer binding sites and polyvalent homodimer binding sites emphasizes several important differences. The most remarkable is the hydrophobic composition which is increased from 39.7% (for previously defined homodimer binding sites) to 66.5% for specific homodimer binding sites. Charged (Asp, Glu, Lys, Arg) and polar compositions (Ser, Thr, Asn, Gln) vary accordingly, being divided by a factor of 2. As a consequence, specific homodimer binding sites are shown to be much more correlated with the composition of the protein core than polyvalent homodimer binding sites and far less correlated with the composition of crystal packing interfaces. Indeed, homodimer interfaces had already been described as resembling the protein core (18,21). This hypothesis is further supported and detailed by our results on specific homodimer binding sites.

Concerning nucleic acids, as expected, both polyvalent homodimer/DNA and specific DNA binding sites are shown to be much more cationic than both polyvalent and specific homodimer binding sites. In addition, specific DNA binding sites tends to be less hydrophobic (31.3%) than polyvalent homodimer/DNA binding sites (39.1%) and more polar (30.3% against 26.7%).

As suspected, other homodimer polyvalent binding sites have amino acid compositions closer to what was previously described, explaining the observed differences between specific and polyvalent homodimer binding sites. For instance, polyvalent homodimer/heterodimer binding sites are more correlated to heterodimer (0.77) than specific homodimer binding sites (0.17). Furthermore, polyvalent homodimer/heterodimer and homodimer/solvent binding sites are seen to be more charged and polar which results in poor correlations with the protein core composition and higher correlations with the crystal packing binding sites.

Favorable and unfavorable polyvalent binding sites. Each binding site has different frequencies in the SWD dataset, the most frequent being the homodimer (17%), followed by ligand (15%) and heterodimer (11%) binding sites. Therefore, to perform unbiased observations of the likelihood of a binding site type A to be co-localized (polyvalent) with a binding site type B, we used the well established methodology described by Henikoff (36). A total of 525 co-localizations of binding site types were observed. Only 19 RNA binding sites were present; therefore, log-odd ratios concerning RNA should be considered as first approximations. The results are summarized in Table 4 and show that different polyvalent binding sites may be either favorable (as for the homodimer/heterodimer or heterodimer/peptide pairs), or unfavorable (homodimer/peptide or homodimer/RNA pairs). As expected, heterodimer binding sites can easily bind peptides (log odd: 0.79) and small peptides (log odd: 0.63), whereas a homodimer binding site are less able to bind peptides (−0.35) and small peptides (−0.3).

Table 3. Amino acid composition/contribution of the specific versus polyvalent binding sites

Residue	HoD *(A)*	HoD,HoD *(B)*	Log(B/A)	HoD,HeD *(C)*	Log(C/B)	HoD,Solvent *(D)*	Log(D/B)	DNA	HoD,DNA
GLY	6.3	7.2	0.05	5.4	−0.12	6.2	−0.07	5.3	2.6
ALA	6.5	9.2	0.15	4.4	−0.32	5.8	−0.20	4.2	2.1
VAL	5.5	10.5	0.28	5.2	−0.31	4.1	−0.41	6.0	11.3
LEU	8.8	17.1	0.29	7.5	−0.36	6.3	−0.43	3.5	6.7
ILE	4.8	10.2	0.33	2.8	−0.56	3.2	−0.51	3.9	4.6
PRO	5.1	8.0	0.20	3.4	−0.38	4.3	−0.27	3.5	3.1
MET	2.6	4.2	0.22	2.8	−0.18	2.1	−0.31	4.9	8.7
PHE	4.4	6.9	0.20	8.0	0.07	3.5	−0.29	0.9	0.0
TRP	1.6	1.1	−0.16	1.3	0.07	1.7	0.18	0.4	0.0
SER	5.7	3.3	−0.24	8.0	0.38	6.4	0.29	9.7	6.2
THR	5.6	4.0	−0.14	6.7	0.22	6.1	0.18	8.8	8.7
ASN	4.2	1.5	−0.45	2.8	0.27	5.0	0.52	7.2	6.2
GLN	4.4	1.9	−0.37	3.4	0.25	5.2	0.44	4.6	5.6
ASP	5.8	1.9	−0.49	5.2	0.43	7.0	0.57	4.6	2.6
GLU	6.8	2.3	−0.47	6.7	0.47	8.1	0.55	4.1	5.1
CYS	1.2	1.5	0.09	0.5	−0.47	1.1	−0.13	0.9	0.0
TYR	4.2	2.6	−0.22	7.7	0.48	4.7	0.26	2.8	1.5
HIS	3.3	1.5	−0.33	2.3	0.18	3.8	0.39	2.7	3.1
LYS	6.1	3.1	−0.30	10.6	0.54	7.1	0.37	12.7	11.3
ARG	6.9	1.9	−0.56	5.4	0.45	8.4	0.65	9.2	10.8
Hydrophob	*39.7*	***66.5***		*31.4*		*31.9*		***31.3***	***39.1***
Anionic	*12.6*	***4.2***		*11.9*		*15.1*		*8.7*	*7.7*
Cationic	*13*	***5***		*16*		*15.5*		*21.9*	*22.1*
Polar	*19.9*	*10.7*		*20.9*		*22.7*		***30.3***	***26.7***
C(HoD)	0.97	0.64		0.61		0.76		0.45	0.49
C(HeD)	0.86	0.17		0.77		0.90		0.77	0.6
C(Pcore)	0.43	**0.93**		0.17		**−0.07**		−0.11	0.17
C(Crystal)	0.84	**0.10**		0.62		**0.95**		0.67	0.5

C(X) is a Pearson product–moment correlation between the amino acid scale described in column and the published amino acid scale X. When X is HoD (homodimer), the amino acid scale is from Bahadur *et al.* (18); when X is HeD (Heterodimer) or Pcore (protein core), the amino acid scales are from Albou *et al.* (27); when X is Crystal (crystal packing), it is from Bahadur *et al.* (15). HoD,X (such as HoD,HeD) indicates an homodimer/X polyvalent binding site.

Table 4. Average log-odd ratios for two binding sites to be co-localized

	Homodimer	Heterodimer	Peptide	S.Peptide	DNA	RNA	Ligand	Solvent
Homodimer	0.12	0.13	−0.35	−0.30	−0.02	−0.60	−0.09	0.02
Heterodimer	0.13		0.79	0.63	0.09	−0.85	0.12	−0.06
Peptide	−0.35	0.79	−0.02	0.84			−0.09	−0.23
S.Peptide	−0.30	0.63	0.84				0.30	−0.15
DNA	−0.02	0.09			0.34		0.63	−0.12
RNA	−0.60	−0.85				2.37	0.43	0.04
Ligand	−0.09	0.12	−0.09	0.30	0.63	0.43	−1.56	0.15
Solvent	0.02	−0.06	−0.23	−0.15	−0.12	0.04	0.15	−0.01

Each cell corresponds to the log-odd ratio of the observed probability of having a binding site type A overlapping a binding site type B over the expected probability of having A and B overlapping. Positive values indicate favorable overlapping/polyvalence between the binding site types A and B, whereas negative values indicate unfavorable overlapping/polyvalence. Empty cells are unsolved values due to an insufficient number of overlaps between the two considered binding sites; however, they should indicate low log-odd values.

Interestingly, small peptides and ligands which both serve as key regulating factors of protein activities can preferentially share a same binding site (log odds: 0.3). As for ligand binding sites, they are the most polyvalent binding sites of all, being preferentially co-localized with all other binding site types considered, with the exception of homodimer and peptide where no clear preference can be observed. Concerning water binding sites, only the ligand binding sites seem to be preferentially co-localized with a moderate log odds ratio of 0.15; nevertheless, this result agrees with the manual visualization of the SWD Molecular Cartographies (example in Figure 2).

DISCUSSION

This work was dedicated to the analysis of protein surfaces and binding sites. In addition to describing an automated approach capable of performing fast comparative and integrated analysis of structures, we have demonstrated the existence of two major families of binding sites: (i) specific binding sites that are only able to bind a

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

specific type of molecule, (ii) polyvalent binding sites that can bind different types of molecules. Our analysis suggests that some of these specific and polyvalent binding sites can be distinguished based solely on their amino acid composition. Additionally, polyvalent binding sites often highlight an intermediate composition between the specific binding sites they represent, e.g. a polyvalent DNA/homodimer binding site exhibits properties both from specific DNA binding sites and homodimer binding sites.

The question of whether specific binding sites could demonstrate stronger or even permanent interaction remains to be answered. In the case of specific homodimer binding sites which are relatively well correlated to the core of proteins (Pearson correlation of 0.93), one should nevertheless emphasize that such localized hydrophobic regions (on average 66%) are unlikely to remain free in the cell and would indeed suggest stronger interaction. Similarly, the polyvalent binding sites which are capable of binding at least two different types of molecules are necessarily involved in weaker interactions since they need to dissociate and associate with different partners.

Three examples of polyvalent binding sites are reported in Figure 3. Among them, an important and well documented example is the Retinoid X Receptor (RXR) which belongs to the large and important family of nuclear receptors. Indeed, RXR differentially regulates gene expression by forming either homodimers or heterodimers with other nuclear receptors such as RAR, VDR or PPAR (25,37,38). As for the transcription factor TATA binding protein (TBP) involved in DNA melting, its homodimer binding site is also co-localized with its DNA binding site.

The examples above describe binary polyvalent binding sites, but the Molecular Cartography of the pancreatic α-amylase also suggests that a single binding site may accept the binding of more than three different molecular types.

Previous results based on the analysis of a single type of interaction identified on average 20% of the protein surface as interacting. Using our integrative approach, we show that by considering all types of binding sites, the average interacting surface is re-evaluated at 44% (in terms of surface residues) or 45% (in terms of accessible surface area). However, this average interacting surface should be considered as a lower approximation since it is probable that not all the biological binding sites of studied molecules were described in the PDB. Similarly, it is well known that the number of solved water molecules not only depends on the crystallographic resolution, but suffers from partial determination since their importance both in molecular stability and interaction mediation were recognized only recently (34,39,40). The averaging proposed by M-ORBIS reduces these effects and allows a mapping of protein–water contact frequencies (Figure 2). Although the specifically solvated residues can be regarded as an accurate result (criteria to detect residue–water contacts are very stringent and the contacts are seen in at least 50 structures), the transiently solvated surface should be considered as a lower approximation due to the possible omission of water molecules in structures.

Interestingly, the observation of the interacting surfaces at different minimal contact frequencies (Table 2), suggests that about half of the residues composing a binding site are not specific to the partner and always participate to a given type of interaction (ratio SWD_{90}/SWD_{10}). This also implies that the other half of the residues could modulate the recognition specificity of distinct partners. In particular, DNA and RNA binding sites seem to have more physico-chemical constraints since 69 and 61%, respectively, of their binding sites are composed of residues that are always seen interacting in related structures. Inversely, ligand binding sites which can bind different ligands with different affinities are shown to be composed of only 31% of those frequently interacting residues, so the remaining 69% of the residues could serve to modulate the recognition specificity of different compounds. From these observations, we would like to propose the notion of a 'core' binding site defined by the residues that are always seen interacting for a given type of interaction. This core binding site would meet some of the requirements required for a given type of interaction, whereas the remaining residues of the binding site would be more fuzzy, thus serving the purpose of recognition specificity. If this core binding site is indeed used to meet the requirements for a given type of interaction, it may be more evolutionarily conserved than the remainder of the binding site. This notion of core binding site may also share similarities with the existing notion of'core'and 'rim'residues composing a binding site which were defined according to the presence of fully buried atoms at the interface (20).

These results may have consequences for protein binding site predictions, for example to redirect attention towards the integrative prediction of the different types of binding sites (a harder computational problem) or to focus efforts towards the prediction and characterization of the core binding sites and hot spots (24). Such integrative mapping of the different types of binding sites (Molecular Cartography) of molecules will help to accelerate docking approaches by giving fast and easy access to existing structural data (Figure 5). In particular, since the docking problems now include the distinction between protein–protein, protein–nucleic acid and protein–ligand interactions, M-ORBIS should prove to be a tool of choice as it differentiates and locates all the known types of binding sites and identifies the frequently interacting residues and the frequently solvated residues.

Additionally, the likelihood of polyvalent binding sites indicates that small peptides and ligands are more co-localized than peptides and ligands or proteins and ligands. The logic behind the distinction of protein, peptide and small peptide relies on the observation that proteins (defined here as a polypeptide of more than 60 residues) often have a stabilizing protein core which may reflect conformational flexibility behaviors different from those of peptides, which generally do not have a well formed protein core. Also, small peptides are separated from peptides due to their small size and the intuitive idea that they can mimic ligands (small chemical compounds). Although arbitrary, previous studies of Stanfield (41) or Petsalaki (42) also suggested

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

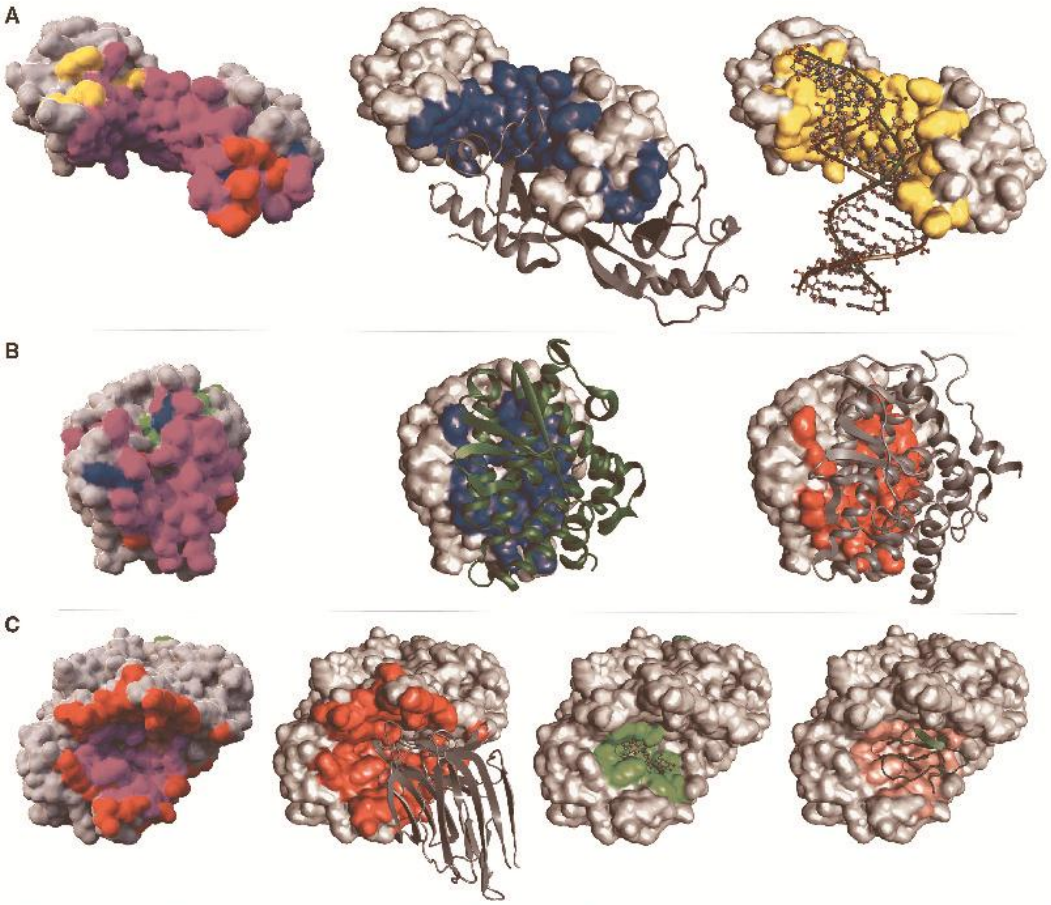


Figure 3. Examples of proteins exhibiting polyvalent binding sites. Polyvalent binding sites are a general phenomena in molecular structures. (**A**) The TATA-binding protein (TBP) processed by the M-ORBIS Molecular Cartography approach starting from structure 1ais:A. (**B**) The Retinoid X Receptor-Alpha (RXR) cartography from the structure 1dkf:A. (**C**) The Pancreatic Alpha-Amylase (PAA) cartography from the structure 1dhk:A. Binding site types are represented in different colors: blue for homodimer, red for heterodimer, yellow for DNA, green for ligand and salmon for peptide. For TBP and RXR, the homodimer partner shown is extracted from structures 1d3u and 1dkf, respectively. For PAA, the heterodimeric and peptide partners are extracted from structures 1dhk and 1clv, respectively, whereas the ligand partner was extracted from 1g9h.

differentiating small peptides from other polypeptides using a size limit near 20 residues. A clearer distinction between proteins, peptides and small peptides could be achieved either by studying the percentage of residues constituting the polypeptide core, or by studying the intrinsic conformational flexibilities, or by optimizing the likelihood of co-localization of small peptide and ligand binding sites.

As M-ORBIS collects and classifies structures according to their molecular contexts, it is possible to analyze the conformational flexibility for a given molecular context or upon the change from a context to another. For instance, the CDK is a well studied family of enzymes which catalyses the transfer of a phosphate group from ATP onto the hydroxyl group of a serine or threonine. They play a crucial role in cell cycle regulation and are activated by their binding to different cyclins. The unbound (monomeric) form is known to be inactive due to several structural constraints (43). By defining two molecular contexts (the first corresponding to the structures of CDK in interaction only with water and ligands, the second corresponding to the structures of CDK involved only in heterodimeric, solvent and ligand interactions), the M-ORBIS cartography approach was able to automatically detect, average and map the changes of conformation between these two states (Figure 4). They correspond, with some minor differences, to what was previously observed in the literature (44), with the shifting of the T coil towards the heterodimeric partner (here the cyclin), and the displacement of the PSTAIRE helix in the

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

Nucleic Acids Research, 2011, Vol. 39, No. 1 41

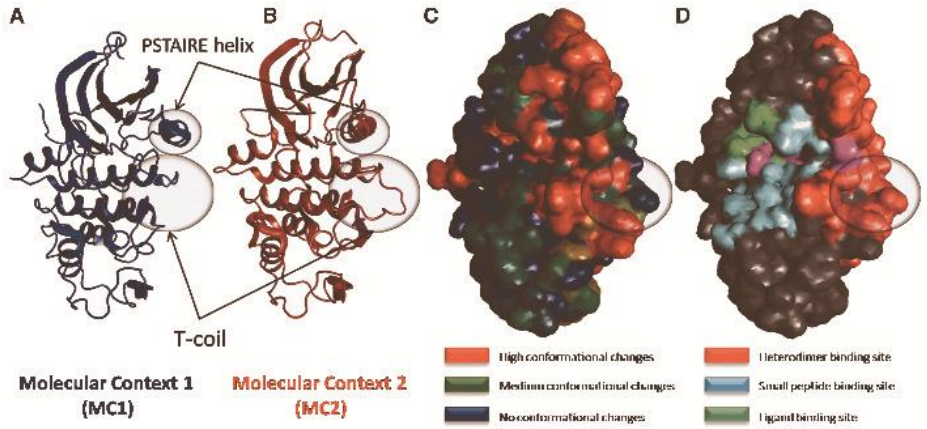


Figure 4. Change of conformation between molecular contexts: the case of CDK. Molecular Cartography obtained from the structure of the CDK2 (1fin:A). Two molecular contexts are defined: MC1 corresponds to the CDK2 in an environment where it interacts with water and ligand only, whereas MC2 corresponds to an environment where it also interacts with another protein to form a heterodimer. For each of these molecular contexts, the corresponding sets of structures has been automatically detected by M-ORBIS and an averaged backbone is computed and shown in (**A** and **B**). Two main conformational changes are involved between MC1 and MC2: the T-Coil helix moves towards the heterodimer binding site, while the PSTAIRE-helix is pushed in the opposite direction. The amount of conformational change has been mapped onto the protein surface in (**C**); blue, no change; green, small change; red, important change. The molecular cartography shown in (**D**) allows to correlate these conformational changes with the location of each binding site type.

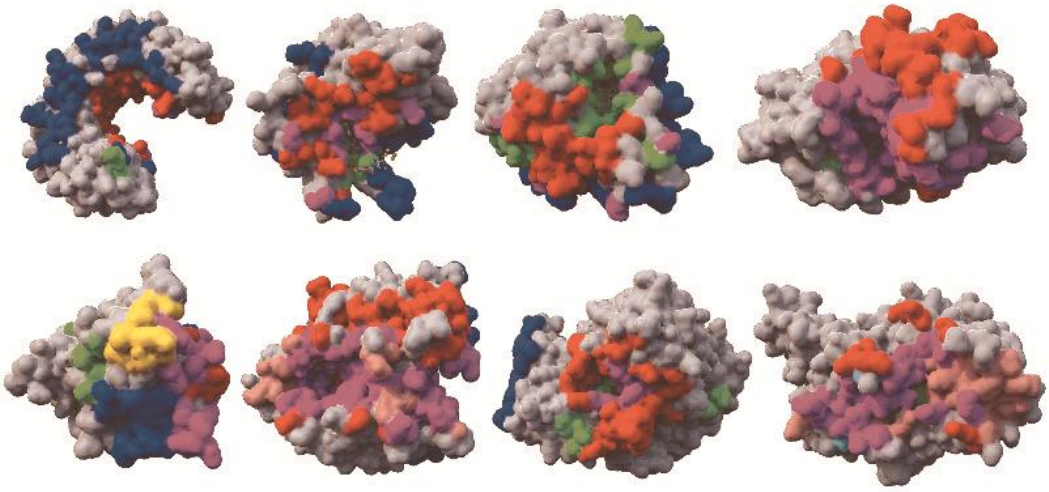


Figure 5. Molecular Cartography of binding sites. Heterodimeric binding sites are represented in red, homodimeric and ligand binding sites in blue and green, respectively and DNA and peptide binding sites in yellow and salmon, respectively. Polyvalent binding sites are indicated in purple. From top left to bottom right, the cartography of proteins: ribonuclease inhibitor (1dfj), ferrodoxin-nadp reductase (1ewy); neuraminidase (3b7e); cAMP-dependent kinase (1ydr); p53 tumor suppressor (1tsr); neurotoxin bont/A (1xtg); acetylcholinesterase (1fss); guanine nucleotide-binding protein G(i) (1bof).

opposite direction. It was also possible to define other molecular contexts, for instance representing (i) only solvent interactions, (ii) only heterodimer and solvent interactions or (iii) heterodimer, ligand, small peptide and solvent interactions. The comparison of these other molecular contexts demonstrates that the T coil displacement results only from the heterodimer formation and not from the binding of the ligand or the small peptide.

Most of the programs used to generate an M-ORBIS Molecular Cartography are generic and should be soon applicable to other molecules such as RNA and DNA. Furthermore, if our approach is aimed at describing the functional and dynamic behaviors of a single molecular chain, it has been observed that an assembly (group of chains) can be required to perform an interaction with another molecular partner (7,20). As a consequence, it

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

should be possible to search not only for the structures containing a specific single chain, but also for structures containing a specific assembly. For instance, the retrieval of related structural chains is currently achieved using a sequence-based comparison engine (PipeAlign), but recent advances in structural comparisons using spherical harmonics (45) could be used to retrieve structures containing a specific assembly. Such structural comparisons will also enhance both the rapidity and sensitivity of the molecular cartographies as some homologies are easier to detect on structures than on sequences. Other comparison methods are also being investigated. In further developments, our automated analysis will benefits from other databases such as PiSA (13) to help discriminate between biological and crystal packing interfaces.

As a conclusion, starting from a molecular structure with no or little functional knowledge, the ultimate goal of Molecular Cartography is to provide an extensive description and characterization of the dynamic functions and behaviors of a molecule by integrating the analysis of related structural data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank J. Janin for his teaching and passion for molecular interactions, R. Bahadur for sharing his data on molecular interactions to help in the validation of M-ORBIS, J. Thompson for her critical reading of the article and F. Gros for his ongoing encouragement and passion for molecular biology. All images were generated by the MSVM platform.

FUNDING

This work was supported by funds from the French Fondation 'Louis D, Institut de France', Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale and the Université de Strasbourg. Funding for open access charge: Centre Européen de Recherche en Biologie et en Médecine (Institut de Génétique et de Biologie Moléculaire et Cellulaire).

Conflict of interest statement. None declared.

REFERENCES

1. Bartlett,G.J., Todd,A.E. and Thornton,J.M. (2003) Inferring protein function from structure. *Methods Biochem. Anal.*, **44**, 387–407.
2. Brylinski,M., Kochanczyk,M., Broniatowska,E. and Roterman,I. (2007) Localization of ligand binding site in proteins identified in silico. *J. Mol. Model*, **13**, 665–675.
3. Porollo,A. and Meller,J. (2007) Prediction-based fingerprints of protein-protein interactions. *Proteins*, **66**, 630–645.
4. Jones,S. and Thornton,J.M. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, **8**, 3–7.
5. Domingues,F.S., Rahnenfuhrer,J. and Lengauer,T. (2007) Conformational analysis of alternative protein structures. *Bioinformatics*, **23**, 3131–3138.
6. Brylinski,M. and Skolnick,J. (2008) What is the relationship between the global structures of apo and holo proteins? *Proteins*, **70**, 363–377.
7. Hwang,H., Pierce,B., Mintseris,J., Janin,J. and Weng,Z. (2008) Protein-protein docking benchmark version 3.0. *Proteins*, **73**, 705–709.
8. Chung,J.L., Wang,W. and Bourne,P.E. (2006) Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins*, **62**, 630–640.
9. Higurashi,M., Ishida,T. and Kinoshita,K. (2009) PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.*, **37**, D360–D364.
10. Janin,J. (1997) Specific versus non-specific contacts in protein crystals. *Nat. Struct. Biol.*, **4**, 973–974.
11. Janin,J., Bahadur,R.P. and Chakrabarti,P. (2008) Protein-protein interaction and quaternary structure. *Q. Rev. Biophys.*, **41**, 133–180.
12. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
13. Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.
14. Ponstingl,H., Henrick,K. and Thornton,J.M. (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57.
15. Bahadur,R.P., Chakrabarti,P., Rodier,F. and Janin,J. (2004) A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.*, **336**, 943–955.
16. Bernauer,J., Bahadur,R.P., Rodier,F., Janin,J. and Poupon,A. (2008) DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, **24**, 652–658.
17. Zhu,H., Domingues,F.S., Sommer,I. and Lengauer,T. (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, **7**, 27.
18. Bahadur,R.P., Chakrabarti,P., Rodier,F. and Janin,J. (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708–719.
19. Bahadur,R.P., Zacharias,M. and Janin,J. (2008) Dissecting protein-RNA recognition sites. *Nucleic Acids Res.*, **36**, 2705–2716.
20. Chakrabarti,P. and Janin,J. (2002) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343.
21. Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.
22. Nadassy,K., Wodak,S.J. and Janin,J. (1999) Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
23. Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
24. Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.
25. Brelivet,Y., Kammerer,S., Rochel,N., Poch,O. and Moras,D. (2004) Signature of the oligomeric behaviour of nuclear receptors at the sequence and structural level. *EMBO Rep.*, **5**, 423–429.
26. Clackson,T., Ultsch,M.H., Wells,J.A. and de Vos,A.M. (1998) Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.*, **277**, 1111–1128.
27. Albou,L.P., Schwarz,B., Poch,O., Wurtz,J.M. and Moras,D. (2009) Defining and characterizing protein surface using alpha shapes. *Proteins*, **76**, 1–12.
28. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
29. Ponstingl,H., Kabir,T. and Thornton,J.M. (2003) Automatic inference of protein quaternary structure from crystals. *J. Appl. Cryst.*, **36**, 1116–1122.
30. Plewniak,F., Bianchetti,L., Brelivet,Y., Carles,A., Chalmel,F., Lecompte,O., Mochel,T., Moulinier,L., Muller,A., Muller,J. *et al.* (2003) PipeAlign: a new toolkit for protein family analysis. *Nucleic Acids Res.*, **31**, 3829–3832.

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

31. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
32. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
33. Hubbard,S.J. and Thornton,J.M. (1992) *NACCESS Computer Program*. Departement of Biochemestry and Molecular Biology.
34. Rodier,F., Bahadur,R.P., Chakrabarti,P. and Janin,J. (2005) Hydration of protein-protein interfaces. *Proteins*, **60**, 36–45.
35. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
36. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
37. Ijpenberg,A., Tan,N.S., Gelman,L., Kersten,S., Seydoux,J., Xu,J., Metzger,D., Canaple,L., Chambon,P., Wahli,W. *et al.* (2004) In vivo activation of PPAR target genes by RXR homodimers. *EMBO J.*, **23**, 2083–2091.
38. Bourguet,W., Vivat,V., Wurtz,J.M., Chambon,P., Gronemeyer,H. and Moras,D. (2000) Crystal structure of a heterodimeric complex of RAR and RXR ligand-binding domains. *Mol. Cell*, **5**, 289–298.
39. Knight,J.D., Hamelberg,D., McCammon,J.A. and Kothary,R. (2009) The role of conserved water molecules in the catalytic domain of protein kinases. *Proteins*, **76**, 527–535.
40. Raschke,T.M. (2006) Water structure and interactions with protein surfaces. *Curr. Opin. Struct. Biol.*, **16**, 152–159.
41. Stanfield,R.L. and Wilson,I.A. (1995) Protein-peptide interactions. *Curr. Opin. Struct. Biol.*, **5**, 103–113.
42. Petsalaki,E., Stark,A., Garcia-Urdiales,E. and Russell,R.B. (2009) Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput. Biol.*, **5**, e1000335.
43. De Bondt,H.L., Rosenblatt,J., Jancarik,J., Jones,H.D., Morgan,D.O. and Kim,S.H. (1993) Crystal structure of cyclin-dependent kinase 2. *Nature*, **363**, 595–602.
44. Jeffrey,P.D., Russo,A.A., Polyak,K., Gibbs,E., Hurwitz,J., Massague,J. and Pavletich,N.P. (1995) Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature*, **376**, 313–320.
45. Sael,L., Li,B., La,D., Fang,Y., Ramani,K., Rustamov,R. and Kihara,D. (2008) Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins*, **72**, 1259–1273.

Downloaded from nar.oxfordjournals.org by guest on May 23, 2011

III. Conclusions et Perspectives

1. Etude de la surface non-interagissante

Cette étude met en évidence qu'en moyenne, plus de 44% de la surface des protéines est capable d'interagir avec des molécules autres que du solvant et des ions. Il est alors raisonnable de se poser la question s'il existe ou non une surface non-interagissante, c'est-à-dire une surface présentant des propriétés géométriques et physico-chimiques telles qu'elle ne permet pas de former des contacts suffisamment spécifiques pour former un assemblage durable.

Un premier élément de réponse est de considérer qu'effectivement, il existe un biais de composition entre les différents sites de liaison et le restant de la surface moléculaire considérée comme non-interagissante. Cette différence est encore plus importante si l'on considère les cœurs des sites de liaison (généralement plus hydrophobes et aromatiques que le restant des sites de liaison), ou bien si l'on considère les sites de liaison spécifiques qui dénotent des différences plus marquées avec le restant de la surface (exemple des sites de liaison spécifique à la formation d'homodimères ayant en moyenne 66% de résidus hydrophobes).

Si une telle surface non-interagissante existait, elle pourrait avoir un rôle dans la cinétique de rencontre et de formation des assemblages spécifiques. En effet, l'un des mystères importants qu'il reste à élucider dans l'étude de la formation des assemblages biologiques est comment, dans un milieu cellulaire très dense, deux molécules faites pour interagir sont capables de se rencontrer suffisamment rapidement pour provoquer la réponse cellulaire escomptée. Lorsque l'on prend l'exemple de médicaments ou de poisons qui agissent en l'espace de quelques secondes, on comprend qu'il est fondamental de déterminer quels sont ces mécanismes dans la cellule, ou les propriétés, qui favorisent cette rencontre rapide des molécules « devant » ou « pouvant » interagir.

Une telle surface non-interagissante pourrait permettre de limiter la recherche conformationnelle de l'assemblage optimal de deux molécules A et B lors de leur rencontre, en évitant de retenir A et B liés trop longtemps lorsque ceux-ci sont dans une conformation non-spécifique. En d'autres termes, dans cette hypothèse, la surface non-interagissante servirait à accélérer la cinétique de rencontre et de formation de l'assemblage optimal de deux molécules capables d'interagir, en accélérant la cinétique de dissociation des molécules qui ne peuvent pas interagir (Figure 71).

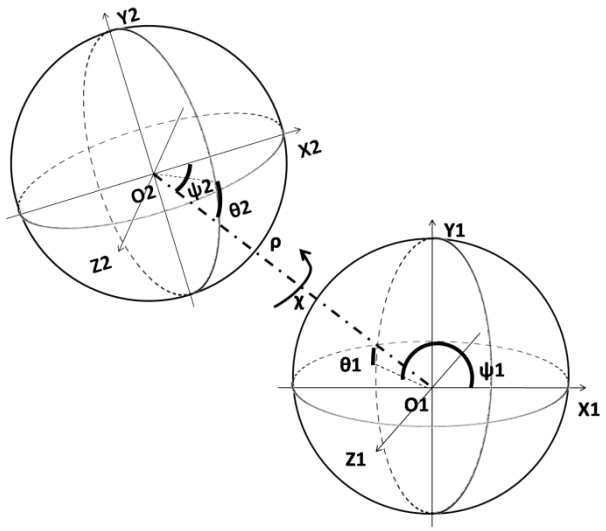


Figure 71. Recherche conformationnelle de l'assemblage spécifique de deux molécules. *En effet, si l'on imagine que la recherche de l'assemblage optimal de deux molécules dans la cellule peut être assimilée au modèle d'amarrage moléculaire (docking), alors on peut comprendre que la surface non-interagissante permettrait d'éviter de tester un ensemble de conformations « inutiles » (ne pouvant produire d'interactions viables), et donc d'accélérer considérablement la cinétique de dissociation des partenaires non-spécifiques, et par conséquent la cinétique d'association des partenaires spécifiques.*

Nous développons un modèle de simulation de rencontres moléculaires dans un environnement 3D pour analyser cette cinétique de rencontre des partenaires spécifiques dans un environnement plus ou moins dense. En particulier, nous souhaitons identifier l'impact de la présence d'une surface non-interagissante, libérant plus rapidement deux molécules ne pouvant interagir, sur cette cinétique. Bien que cette simulation soit relativement simple, elle reste conceptuellement difficile à réaliser en ce sens qu'il faut éviter d'introduire des paramètres « à priori » qui pourraient influencer cette dynamique de rencontre. Par exemple, la surface non-interagissante peut-elle être placée n'importe où sur ces molécules générées aléatoirement ? Ou bien, doit-on la faire correspondre à des formes particulières, comme des crevasses ? En fait, la surface non-interagissante n'est-elle pas codée directement par la forme géométrique ? En effet, bien que notre modèle de rencontre et d'association de molécules soit simple, il a déjà mis en évidence trois observations : (1) lorsque deux molécules rentrent en contact par des régions qui ne sont pas planes, celles-ci restent plus longtemps en contact dans cette conformation parce que la complémentarité géométrique de leurs interfaces diminue la liberté avec laquelle elles peuvent se déplacer dans le milieu. (2) la vitesse d'association de molécules augmente en présence de « courants cellulaires » qui permettent d'une part de dissocier plus facilement les assemblages, et d'autre part, de faire parcourir plus rapidement de longues distances aux molécules. Toutefois, cela nécessite la présence de courants inverses dans la cellule, c'est-à-dire, de courants amenant les molécules d'un point A à un point B, et de courants les amenant d'un point B à un point A. (3) la création de sous-compartiments dans l'environnement permet de concentrer des molécules précises dans ces environnements et

de diminuer l'espace à parcourir pour rencontrer d'autres molécules. Ces sous-compartiments, représentatifs par exemple du noyau, de l'appareil de golgi ou encore des mitochondries, permettent donc non seulement de limiter les possibles interactions entre molécules, mais aussi de créer des espaces plus petits où certaines molécules sont plus concentrées et se rencontrent plus rapidement. Ce rôle cinétique de la compartimentation des cellules eucaryotes pourrait expliquer les observations qui tendent à montrer que les procaryotes ont des réseaux d'interactions beaucoup plus limités que ceux des eucaryotes.

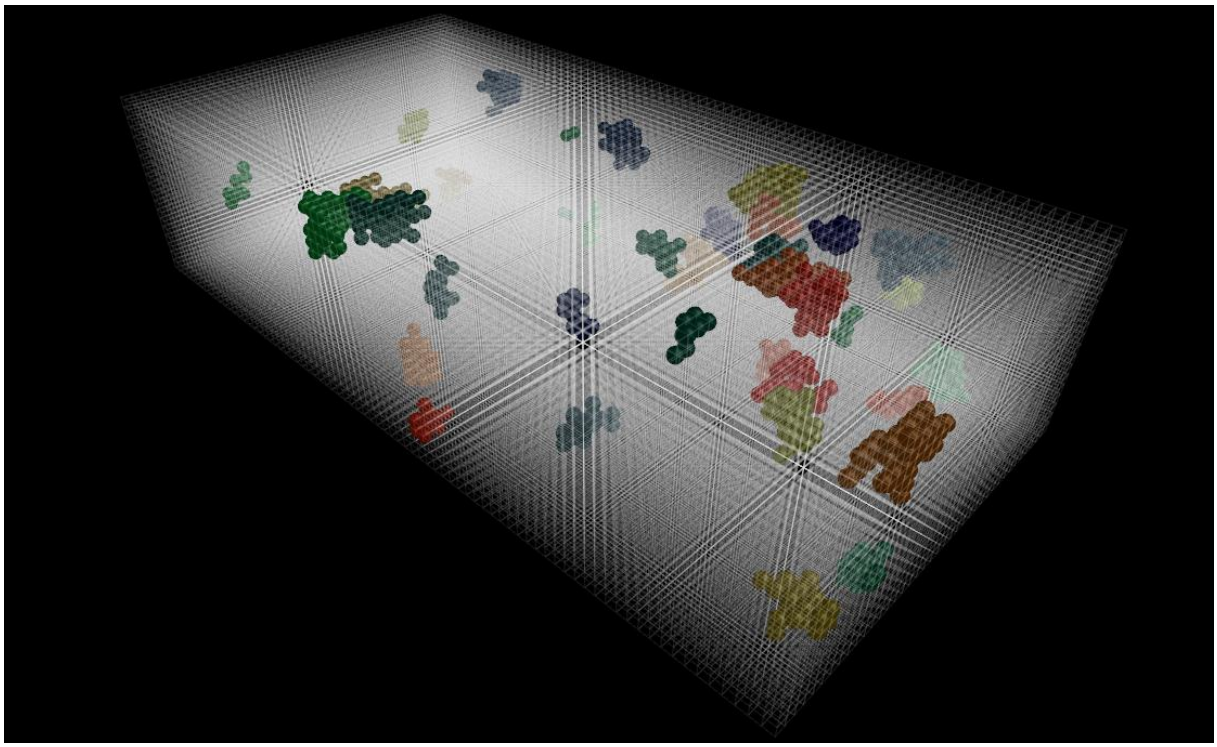


Figure 72. La rencontre des molécules ou le « Jeu de la Cellule ». *En référence à la simulation informatique du « Jeu de la Vie », nous développons une simulation « Jeu de la Cellule » qui permettrait d'analyser la vitesse de rencontre de deux molécules « devant » ou « pouvant » interagir dans un environnement donné. En particulier, nous tentons de voir s'il existe un équilibre entre la complexité des interactions nécessaires à la vie cellulaire, et un pourcentage de surface non-interagissante qui permettrait d'accroître la cinétique de dissociation des assemblages non spécifiques, et donc, d'accroître la cinétique d'association des assemblages spécifiques. Dans cette représentation de l'environnement, chaque molécule est coloriée de façon différente pour pouvoir les distinguer des autres.*

2. La Cartographie Moléculaire par le criblage de régions fonctionnelles

Le principe de cartographie moléculaire que nous avons développé permet de mieux caractériser les fonctions et comportements dynamiques d'une molécule en regroupant ses structures en fonction des environnements qu'elles représentent. Toutefois, pour l'inférence des sites de liaison et de partenaires associés, la cartographie moléculaire nécessite l'existence de plusieurs structures caractérisant cette molécule.

Comme nous allons le voir dans la partie « Criblage virtuel haut-débit des régions moléculaires », il est toutefois possible d'inférer des fonctions aux différentes régions de la molécule à partir de la comparaison (du criblage) de ses régions, avec les régions fonctionnelles déjà identifiées ou identifiables dans les structures.

Dès lors, l'inférence d'une fonction d'une molécule à partir de sa seule structure peut se faire indépendamment de toute ressemblance de séquence, de structures secondaires ou tertiaires avec la structure d'une autre molécule. L'inférence de fonctions se fait par la similarité ou la complémentarité des régions structurales de leurs molécules.

Criblage haut débit des régions moléculaires

« Les conceptions les plus hardies, les spéculations les plus légitimes, ne prennent un corps et une âme que le jour où elles sont consacrées par l'observation et l'expérience »

Louis Pasteur, 1868

Dans ce dernier chapitre, nous présentons le problème général de la comparaison de molécules par des approches de bioinformatique, et mettons l'accent sur l'état de l'art des approches comparatives des données structurales. L'approche nouvelle de criblage virtuel que nous avons développée sera ensuite détaillée en précisant son état de développement et les futurs développements nécessaires à la meilleure modélisation d'un environnement cellulaire. Dans un premier temps, nous illustrerons la capacité de l'approche par l'élaboration de deux arbres fonctionnels des récepteurs nucléaires, basés respectivement sur le site de dimérisation et le site de liaison au coactivateur. Dans un second temps, nous décrirons de nouvelles interactions protéine-protéine, protéine-petit peptide et protéine-ligand, prédites par l'approche. Nous conclurons cette partie sur les interfaces utilisateurs développées qui permettent de visualiser les alignements de régions et d'évaluer la pertinence de l'approche et les problèmes conceptuels qu'il reste à traiter.

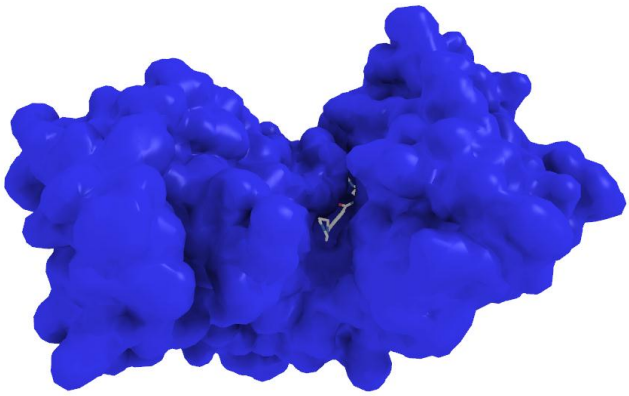
I. Problématique

La biologie et l'étude des fonctions des molécules ont été révolutionnées par la comparaison rapide de leurs séquences par des programmes tels que Fasta, BLAST [325] ou Clustalw [172]. Ces comparaisons rapides des séquences de molécules ont permis par exemple d'identifier les gènes communs aux espèces, de partager des fonctions expérimentalement vérifiées sur une protéine à l'ensemble des protéines présentant de mêmes séquences, ou encore de révéler les liens évolutifs entre espèces.

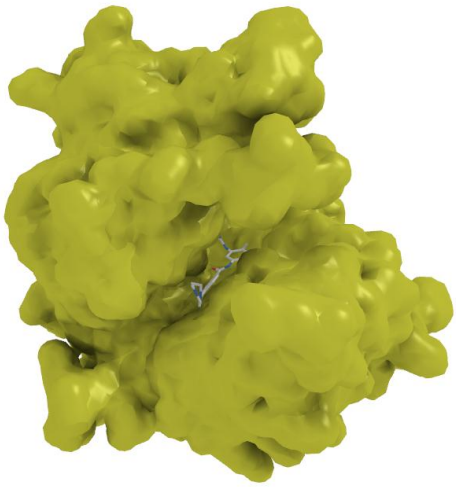
La comparaison des structures secondaires et tertiaires des protéines a permis de révéler des liens de parenté plus distants et difficilement observables sur les seules séquences. Dans l'approche que nous avons nommée « cartographie moléculaire », la comparaison des structures tertiaires d'une même protéine permet de mieux observer sa variabilité conformationnelle, d'explorer ses différents sites de liaisons et partenaires associés, ainsi que de déterminer les résidus qui sont hydratés de façon spécifique.

Toutefois, la comparaison des structures secondaires et tertiaires des protéines est limitée en ce sens que deux régions de surface (telles que des sites de liaison) peuvent être

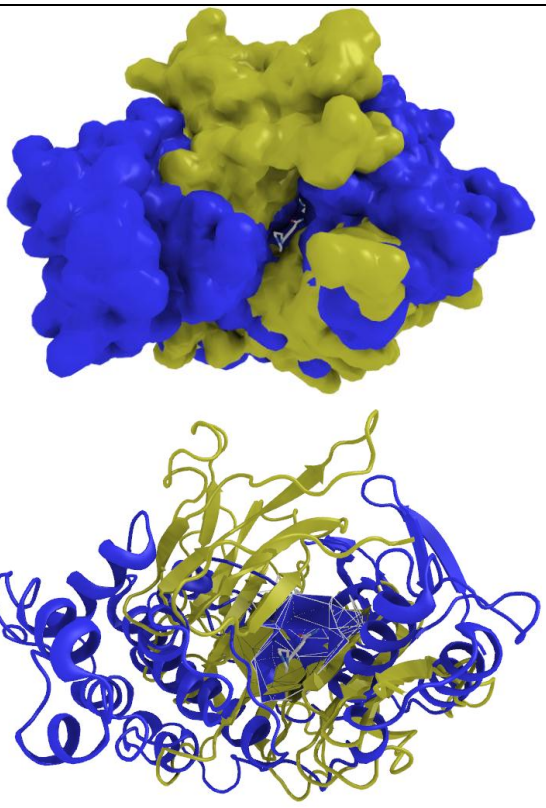
identiques même si les protéines qui les portent ne possèdent aucune identité de séquences, de structures secondaires ou tertiaires (Figure 73).



A) **Mitogen Activated Protein kinase 14** (***MAPK14***) (3hec : A)

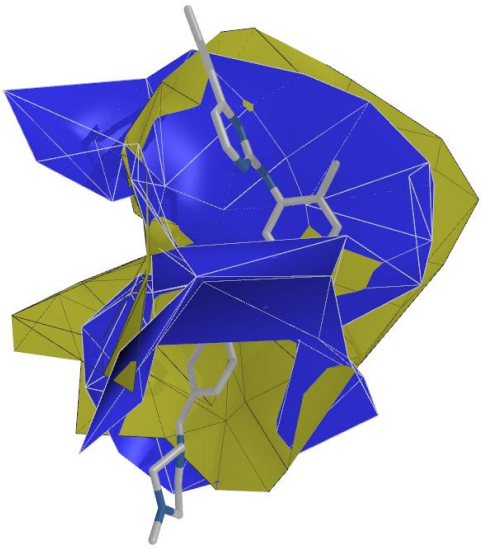


B) **Nucléoporine SEH1** (3f3p : A)



C) **Alignement des deux protéines** à partir de de leurs régions devant lier l'imatinib

D) **Alignement de la région de la MAPK14 liant l'imatinib** à une région retrouvée de la nucléoporine SEH1



57% de similarité structurale et physico-chimique des régions. **11,3%** d'identité de séquence de MAPK14 et N.SEH1

Figure 73. Deux molécules avec des séquences et structures secondaires et tertiaires différentes peuvent lier un même partenaire. *La MAPK14 (en bleu), aussi appelée p38, est connue pour lier l'Imatinib (Gleevec), un médicament contre certains types de cancer. B) Le criblage du site de liaison de l'Imatinib a permis de retrouver plusieurs cibles, dont la nucléoporine SH1 de S.cerevisae.*

Par conséquent, la comparaison des structures moléculaires sur la base de leurs régions va d'une part, nous permettre d'identifier davantage de ressemblances et de liens fonctionnels entre des molécules distantes, et d'autre part, nous permettre de traiter un certain nombre de problèmes encore inexplorés ou peu explorés. Citons pour exemple la détection de nouvelles cibles cellulaires de médicaments [324], et l'évaluation de l'efficacité et des effets secondaires/toxiques qui lui sont liés. De plus, en criblant les sites de liaison de protéines ou d'autres molécules, notre approche de criblage peut être utilisée afin de retrouver également des partenaires protéiques (logique de *double hybride in silico*, et de *docking rigide*).

II. L'importance de la notion de contiguïté pour le criblage de régions

Au cours des 7 dernières années, plusieurs méthodes ont été développées et permettent le criblage de petits motifs structuraux [326, 327]. Toutefois, ces approches n'assurent pas la notion de contiguïté de la région, pourtant importante pour accroître la spécificité du criblage en diminuant l'apparition de faux positifs (Figure 74).

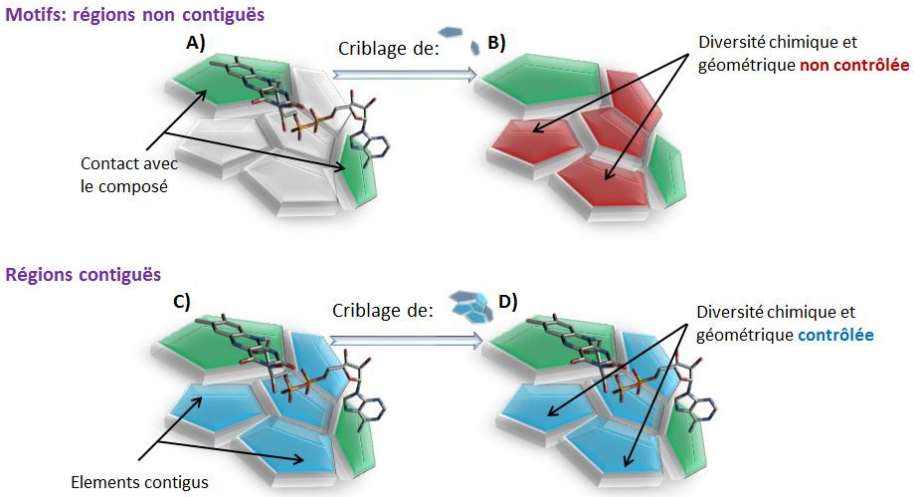


Figure 74. Importance de la sélection de régions contiguës pour le criblage virtuel. *En A et C est représentée une même région hypothétique capable de lier un composé, ici le FAD. En A, on crible un motif composé des atomes et/ou résidus suffisamment proches de FAD (éléments verts). En B, on crible la région contiguë contenant à la fois le motif d'interaction (éléments verts) mais aussi les éléments contigus les reliant (bleus clairs). Le criblage du motif A entraîne la récupération de régions contenant effectivement ces motifs, mais présentant une variabilité non contrôlée au niveau des éléments rouges qui n'ont pas été considérés lors du criblage. Cette variabilité peut s'exprimer soit par des changements de forme (occasionnant de possibles gênes stériques avec le FAD), soit par des changements d'électrostatiques (favorisant ou défavorisant la liaison du ligand). Le criblage de la région contiguë C permet quant à elle de récupérer systématiquement des régions contiguës en y contrôlant la variabilité géométrique et physico-chimique de tous ses éléments.*

D'autres approches, notamment celles basées sur les harmoniques sphériques [328], permettent de cribler des régions moléculaires, mais n'assurent pas nécessairement cette notion de contiguïté. La rapidité du criblage obtenu par les harmoniques sphériques a cependant un prix : la surface modélisée est relativement simple et ne reflète pas nécessairement les détails importants des sites de liaison. En effet, les harmoniques sphériques sont connues pour n'être applicables qu'à des objets en forme d'étoile, c'est-à-dire, ne subissant pas plusieurs repliements successifs dans une même direction. Des développements intéressants de cette méthodologie seront à suivre dans la littérature au cours des prochaines années. Plus généralement, les méthodes existantes de criblages de motifs passent par une discrétisation plus ou moins fidèle des structures moléculaires et de leurs surfaces, mais qui ne préservent pas la position des atomes. En comparaison, la représentation des surfaces moléculaires par les formes alphas conserve la position des atomes, et il est alors possible de traiter la variabilité conformationnelle des régions lors de l'étape de comparaison, et non dans la modélisation.

III. Le criblage des régions moléculaires

1. Définition de la liste exhaustive des régions contiguës d'une molécule

A partir des atomes de la surface moléculaire, par exemple représentée par les formes alpha, il est possible de définir la liste exhaustive des régions contiguës d'une molécule en générant des régions de tailles et de formes variées. La diversité des tailles et des formes des régions peut être obtenue notamment par la génération sous contrainte de formes, ou par la génération des empreintes structurales.

Pour générer la liste exhaustive et non redondante des régions de la molécule, il suffit alors de ne conserver que les régions qui ne se recouvrent pas davantage qu'un seuil X, c'est-à-dire, qu'aucun couple de régions R_1 et R_2 de cette liste L ne peut posséder plus de X% d'atomes en commun :

$$L: \{ R_i \in L, \quad R_j \in L : recouvrement(R_i\, R_j) < X, \forall\, i, j \}$$

Où *L* est la liste exhaustive et non redondante de régions à *X%*, et où recouvrement(R_i, R_j) retourne le pourcentage d'atomes en commun des régions R_i et R_j.

Ainsi, pour un seuil de 90%, il est possible de générer de 100 à 2000 régions selon la taille de la structure moléculaire considérée.

2. Comparaison et criblage de régions moléculaires

La première forme de comparaison « naïve » des régions moléculaires consiste à aligner les deux régions à l'origine du repère O, c'est-à-dire que les centroïdes (centres de gravités) de ces régions coïncident. Pour les régions de surface, une seconde étape consiste à faire pivoter 360/α fois la région d'un pas d'angle α autour de l'axe Y qui doit alors coïncider avec les normales à la surface des deux régions à aligner. On obtient alors un premier alignement optimal qui peut être optimisé en appliquant des rotations et translations modérées autour (ou le long) des axes OX et OZ.

La recherche de l'alignement optimal nécessite la création d'une fonction d'énergie permettant de quantifier la ressemblance ou au contraire la dissimilitude de deux régions. Le score d'énergie noté $S_{global}(R_1, R_2)$ de deux régions R1 et R2 est la somme des scores d'énergie locaux $S_{local}(S_1, S_2)$ correspondant à l'alignement, ou au non alignement, de deux points de R_1 et de R_2, caractérisés par n propriétés P_i (telle que la localisation spatiale, la courbure, les charges, etc) ;

$$Score_{local}(S_1, S_2) = \sum_{i=1}^{n} \alpha_i Score_{P_i}(S_1, S_2)$$

Où α_i est un paramètre permettant de pondérer l'importance de la propriété P_i par rapport aux autres propriétés. Le score $P_i(S_1, S_2)$ correspond alors à la « distance » séparant les points S_1 et S_2 en fonction de la propriété P_i. Une fonction logistique permet de renvoyer un score normalisé entre [-1, 1] et renseignant sur l'écart des états de la propriété P_i aux points S_1 et S_2. Pour suivre les conventions sur les scores d'énergies, lorsque les états des points S_1 et S_2 sont proches (resp. éloignés), alors le score $P_i(S_1, S_2)$ renvoie une valeur proche de -1 (resp. proche de 1). Il est important de noter que l'utilisation de cette équation permet de mesurer les ressemblances et différences de deux régions en fonction de une ou de plusieurs propriétés. En d'autres termes, il est par exemple possible de comparer les régions en fonction de leur géométrie, en fonction de leur charge, en fonction de leur géométrie et de leur charge, etc.

Le score global correspondant à l'alignement de deux régions R1 et R2 est alors :

$$Score_{global}(R_1, R_2) = \sum_{S_i \subset R_1} Score_{local}\left[S_i, Eq_{R_2}(S_i)\right]$$

Où $Eq_{R2}(Si)$ correspond au point S_j de R2 qui est le plus proche du point S_i de R1.

Plus le score d'énergie a une valeur négative (resp. positive) et meilleur (resp.pire) sera l'alignement des deux régions. Toutefois, notons que la valeur de ce score dépend du nombre de points formant la région R_1. Par conséquent, et afin de pouvoir comparer les

alignements entre eux, il a été nécessaire de développer un score d'énergie normalisé renvoyant une valeur par exemple entre 0 et 100%, indiquant la similarité des régions. Pour ce faire, il suffit de calculer le score d'énergie global correspondant à l'alignement de la région à cribler avec elle-même, de sorte d'obtenir le score d'énergie maximal atteignable. Dès lors, tous les scores d'énergies issus de la comparaison de R1 avec une collection d'autres régions sont normalisés par ce score d'énergie maximal.

Le criblage des molécules sur la base de leurs régions structurales correspond en quelque sorte à un « BLAST structural » qui permet de récupérer l'ensemble des molécules portant des régions similaires ou complémentaires et de quantifier la ressemblance des régions retrouvées avec la région requête. Tout comme dans un BLAST, il est possible d'effectuer des criblages itératifs, c'est-à-dire, de récupérer les meilleurs résultats (triés en fonction du score d'énergie normalisé), et de les recribler sur une même base de données, représentative d'un environnement. Le criblage itératif est l'une des techniques permettant d'accroître la sensibilité de la recherche en tenant compte de variations conformationnelles observées dans les structures.

Le criblage de sites de liaison de ligands fréquemment rencontrés dans la PDB (Tableau 5) a permis d'évaluer à la fois la sensibilité mais aussi la borne inférieure de la spécificité de l'approche.

La sensibilité (TP / (TP + FN)) : il a été possible de concevoir une base de données ne contenant que les régions contiguës capables de lier une molécule (exemple : ATP), et de regarder la fraction de ces régions, retrouvée par criblage de l'une d'entre elle. La sensibilité moyenne est de l'ordre de 70% mais dépend du nombre de sites différents qui sont criblés. Lorsqu'une majorité des conformations d'un site de liaison est criblée, la sensibilité peut atteindre 100% dans certains cas.

La spécificité (TP / (TP + FP)) : en raison du manque d'annotations biologiques, seule la borne inférieure de la spécificité peut être évaluée (FP = une région retrouvée qui n'est pas connue pour lier une molécule, ce qui ne signifie pas qu'elle soit véritablement incapable de lier ladite molécule). Mesuré sur l'ATP, le FAD et le MAN, la borne inférieure de la spécificité varie entre 50 (MAN) et 95% (ATP).

3. Filtres pour améliorer les performances du criblage et GRID computing

L'algorithme de comparaison des régions moléculaires peut toutefois se révéler assez lent lorsque l'on considère que près de 90 millions de régions ont pu être générées à partir des

66 000 structures disponibles dans la Protein Data Bank. Il a donc été nécessaire de développer un certain nombre de filtres permettant d'éviter des comparaisons inutiles.

Différents filtres ont ainsi été développés, basés sur la composition et la distribution des états de propriété des régions (exemple : la composition en acides aminés, en résidus cationiques, anioniques ; la présence de sous-régions concaves, convexes, etc). Nous ne détaillerons ici qu'un seul filtre particulièrement puissant qui renseigne rapidement sur la forme générale des régions, et notamment sur leur degré de plissements (Figure 75).

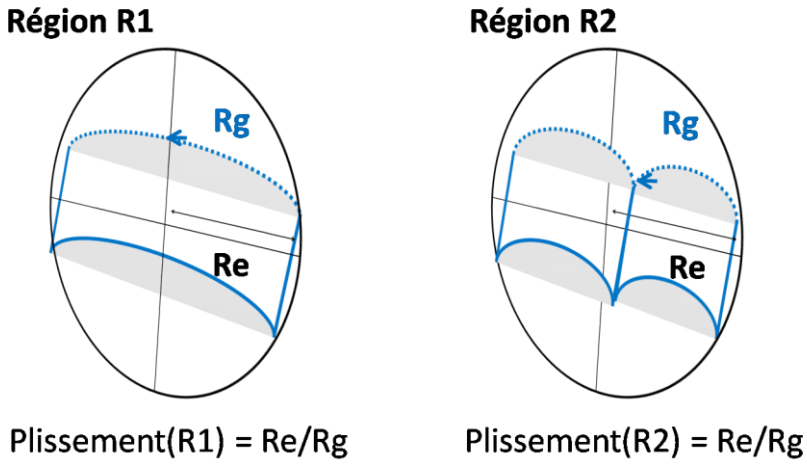


Figure 75. Filtre pour la comparaison de régions, basé sur leurs degrés de plissements. *Re correspond au rayon euclidien du centre de la région à son contour moyenné, alors que Rg correspond au rayon géodésique, c'est-à-dire à la distance qu'il faut parcourir le long de la surface pour relier le centre au contour moyen. Par conséquent, le rapport Re/Rg ou Rg/Re renseigne sur le degré de plissements des régions et permet d'éliminer les régions qui ne peuvent se ressembler. A noter toutefois que deux régions peuvent avoir un même degré de plissement mais être très différentes.*

Malgré ces filtres et malgré l'utilisation de modèles géométriques permettant d'accélérer la comparaison, le criblage de 90 millions de régions moléculaires peut prendre un temps relativement important (quelques heures). Par conséquent, nous avons développé une architecture réseau de calcul partagé (*GRID computing*) permettant de répartir le criblage de régions sur N nœuds de calcul. A présent, le criblage d'une région moléculaire sur les 90 millions de régions prend quelques minutes. De nouveaux développements sont en cours pour diminuer ce temps de calcul d'un facteur 10 à 100.

Notons toutefois que la vitesse de criblage d'une région moléculaire dépend de la base de données de régions sur laquelle on souhaite la cribler. Ainsi, le criblage d'une région sur l'ensemble des régions fonctionnelles contenu dans la Protein Data Bank et décrit par nos

approches, ne prendra que quelques minutes sur un simple ordinateur de bureau (quelques secondes sur une architecture de calcul). Il est donc possible de cribler l'intégralité des régions d'une molécule sur de telles bases de données, notamment afin d'identifier des sites de liaison et partenaires associés qui pourront ensuite être utilisés pour la cartographie moléculaire.

IV. Arbre fonctionnel d'une région

Dans cette section, nous allons illustrer la possibilité de créer des « arbres fonctionnels » qui renseignent non pas sur les évolutions communes et liens de parenté des protéines d'une même famille, mais qui renseignent sur la similarité de chacune des régions fonctionnelles composant ces protéines. De tels arbres sont formés à partir de la comparaison croisée (du criblage croisé) d'une région fonctionnelle portée par une protéine, avec les régions fonctionnelles équivalentes portées par les autres protéines de la famille. Puisque ces arbres fonctionnels décrivent et regroupent les protéines à partir de la similarité de l'une de leurs fonctions, il est possible, lorsque la protéine étudiée a plusieurs régions fonctionnelles distinctes, d'établir plusieurs arbres fonctionnels de cette protéine. Nous illustrerons cette approche par la création de deux arbres fonctionnels des récepteurs nucléaires (RN) correspondant (1) au site de dimérisation et (2) au site de liaison au cofacteur/coactivateur.

La comparaison des molécules en fonction de leurs régions structurales (et notamment fonctionnelles) permet de les classer en regroupant celles qui partagent une même fonction « codée » dans une région locale (comme la capacité à lier un partenaire). La comparaison croisée d'une région fonctionnelle portée par N molécules permet de remplir une matrice NxN renseignant sur les ressemblances de cette région entre les N molécules. Une telle matrice de distances, quantifiant les ressemblances de cette région pour les N molécules, peut ensuite être utilisée par des approches de types UPGMA, pour construire un arbre fonctionnel qui rassemble les molécules dans des branches en fonction de la similarité de ces régions. Des molécules ainsi rassemblées sur une même branche sont supposées partager une même région fonctionnelle, c'est-à-dire, capable de lier de mêmes partenaires moléculaires pour peu que ces molécules évoluent dans un même environnement.

Nous appellerons donc « arbre fonctionnel », un arbre regroupant les molécules selon la ressemblance fonctionnelle de l'une de leurs régions. Dans le cas des récepteurs nucléaires, il est possible d'identifier 4 grandes régions fonctionnelles : le site de dimérisation du LBD, le site de liaison au cofacteur du LBD, le site de liaison au ligand également dans le LBD, et le site de liaison à l'ADN dans le DBD. Il est donc possible de construire quatre arbres

fonctionnels regroupant les récepteurs nucléaires selon la similarité fonctionnelle de chacune de ces quatre régions. Dans ce qui va suivre, nous ne décrirons que les arbres fonctionnels des RN décrivant les ressemblances des sites de dimérisation d'une part, et des sites de liaison au cofacteur d'autre part.

1. Récupération des structures de récepteurs nucléaires

Le tableau 6 de la partie « Régulation de l'expression des gènes eucaryotes », résume les 48 récepteurs nucléaires de l'Homme [196] et référence, lorsque cela est possible, une structure de ce récepteur. Pour chacun des 36 récepteurs possédant une structure, nous avons récupéré l'ensemble des structures de récepteurs homologues en effectuant un BLAST [329] de sa séquence, sur une base de données ne contenant que les séquences des structures de la Protein Data Bank [140]. Seules les structures présentant plus de 30% d'identité de séquence avec au moins l'un des récepteurs recherchés ont été conservées. La plupart des structures étant liées à la base de données Uniprot [330] par un identifiant unique dans le champ DBREF des fichiers PDB, il est possible d'avoir un accès direct à plusieurs informations importantes relatives à la structure du récepteur nucléaire, notamment s'il s'agit d'un récepteur de l'Homme. A partir des 1387 structures récupérées par la recherche par BLAST, nous avons pu retirer par une analyse semi-automatique l'ensemble des protéines qui n'étaient pas des récepteurs nucléaires, comme le cytochrome C6 (PDB : 2v08), ou bien des protéines « probables », « hypothétiques » ou de « fusion », et obtenu ainsi un jeu de 832 structures. Afin de n'observer que les différences fonctionnelles entre récepteurs nucléaires, nous limiterons cette analyse aux 669 structures de protéines humaines (Tableau 26).

Organisme	Nombre de structures	Structures sans ligand et petit peptide	Structures avec un ligand	Structures avec un petit peptide	Structures avec un ligand et un petit peptide
Tous	832	53	769	306	296
Humain	669	37	625	239	232

Tableau 26. Données structurales des récepteurs nucléaires. *(données calculées en Août 2010 à partir de la PDB).*

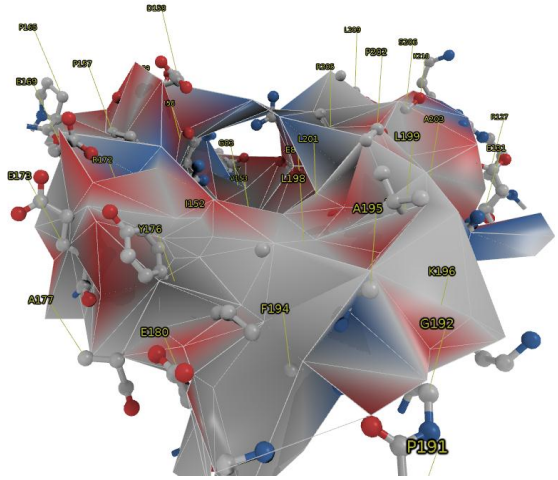
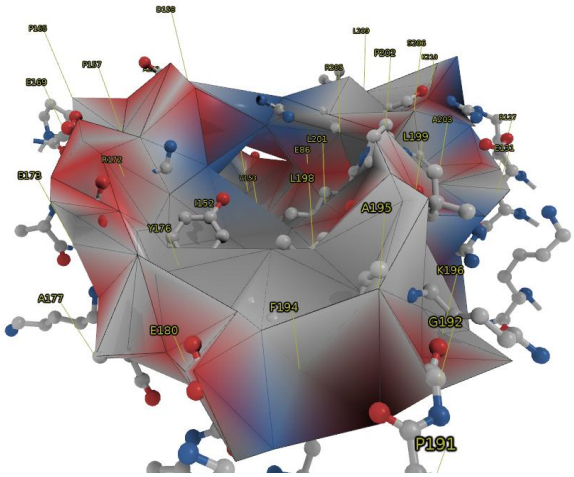
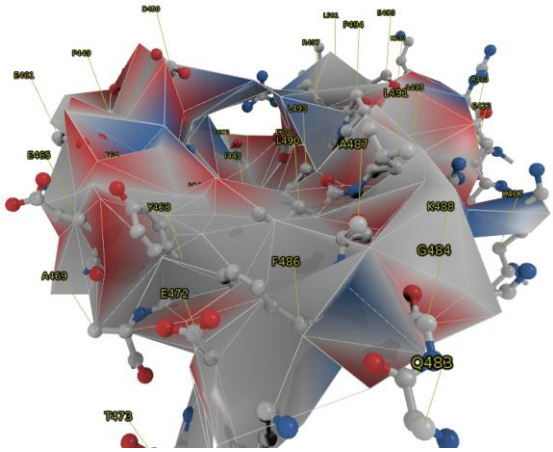
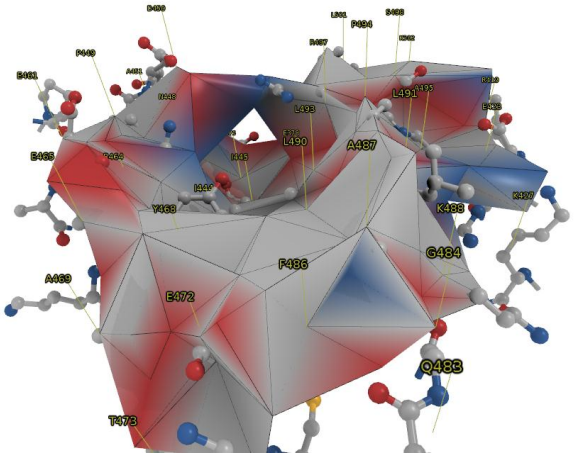
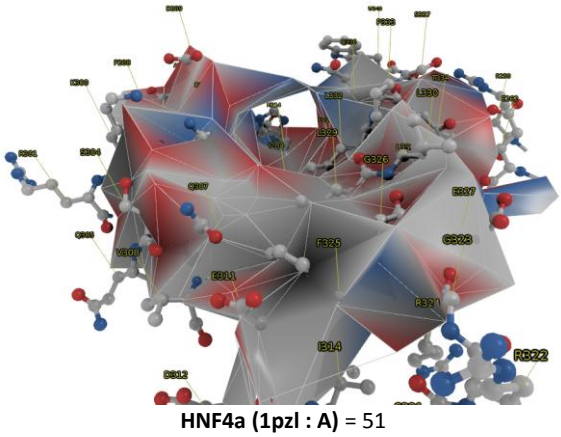
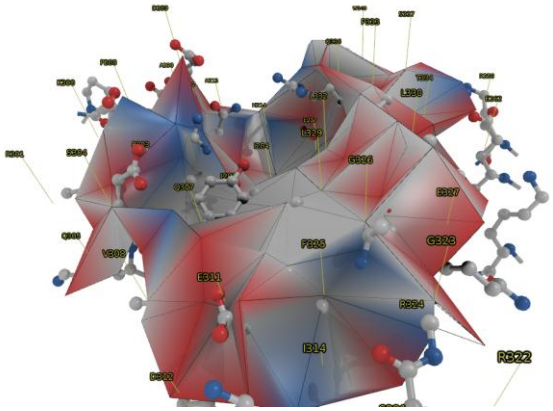
A partir de 669 structures, il a été possible de générer 597 087 régions contiguës de taille 15 Å (taille approximative du site de dimérisation) et 624 852 régions de taille 10 Å (taille approximative du site de liaison au cofacteur). Par la suite, nous appellerons « B_{RN} », la Base

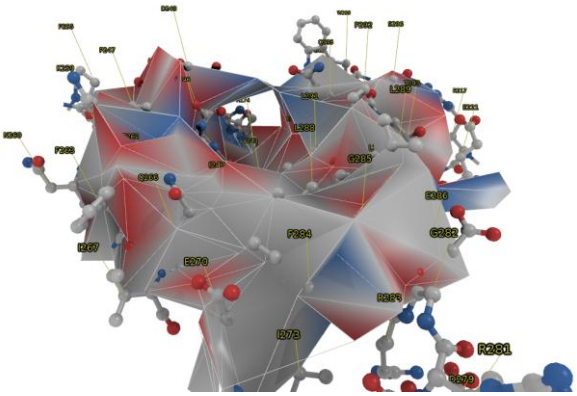
de données des régions moléculaires des Récepteurs Nucléaires. Pour donner une échelle de temps, la génération de la B_{RN} prend deux heures sur un portable. La génération d'une base de données contenant les régions de toutes les structures moléculaires disponibles prendrait quant à elle de l'ordre de 3 à 4 jours. Cette génération comprend le pré-calcul de propriétés géométriques et physico-chimiques telles que la courbure, la charge partielle, l'orientation des normales, afin de permettre un accès rapide lors du criblage.

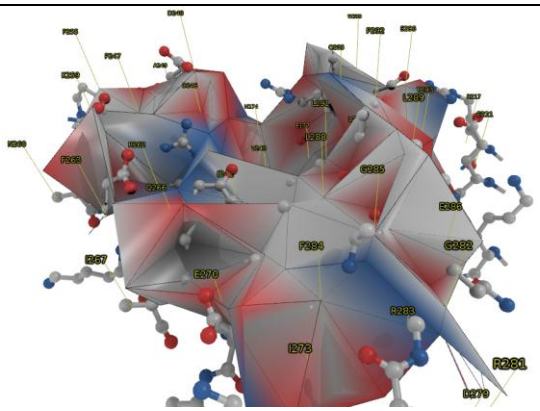
2. Criblage du site de dimérisation de RXRα sur la famille des récepteurs nucléaires

Le site de dimérisation de RXRα humain a été identifié à partir de la structure PDB 1fby:A. Le criblage du site de dimérisation de RXRα humain sur la B_{RN} permet de récupérer l'ensemble des RN portant le site de dimérisation sous une conformation proche. Il est alors possible d'observer les différences fines entre le site de dimérisation de RXRα et de tous les autres RN humains (Tableau 27). Nous avons par ailleurs vérifié manuellement que chaque région ainsi repérée sur une structure, correspondait bien à la localisation théorique du site de dimérisation. Tous les sites de dimérisation partagent un motif structural commun qui correspond à une sorte de cuvette centrale dont les bords sont formés par les résidus portés d'un côté par les hélices H10 et H11, et de l'autre, par les hélices H5 et H8. Une caractéristique importante de RXR qui est propre à ses isoformes, est la création d'un pont salin entre les résidus R426 et (N377, D379) qui relie ces deux groupes d'hélices. Par ailleurs, plus au centre du site de dimérisation, la tyrosine 397 semble être un élément majeur de la spécificité de reconnaissance de RXR, ainsi que nous le montrent les différents alignements par paire des sites de dimérisation du tableau 27. De plus, on observe sur ces alignements une très grande diversité dans la distribution des charges atomiques partielles. Certaines régions comme celles de VDR ou de ESR1 sont sensiblement plus chargées.

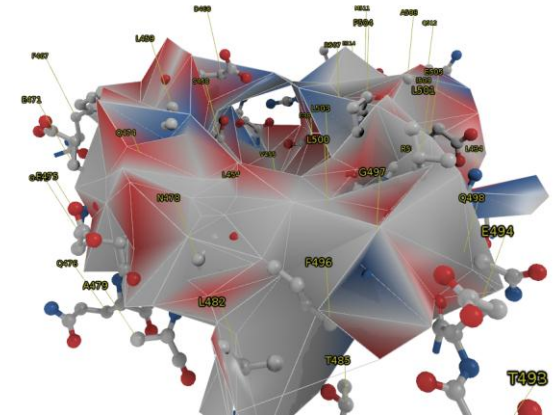
Le tableau qui suit illustre des différences fines entre le site de dimérisation de RXRα et les sites de dimérisation de tous les autres récepteurs nucléaires. Chaque figure montre l'alignement de deux sites de dimérisation, l'un représenté par la forme alpha, l'autre par le modèle boules et bâtons (BB). Le dépassement de résidus par rapport à la forme alpha permet de repérer les résidus qui diffèrent le plus. La coloration rend compte des charges.

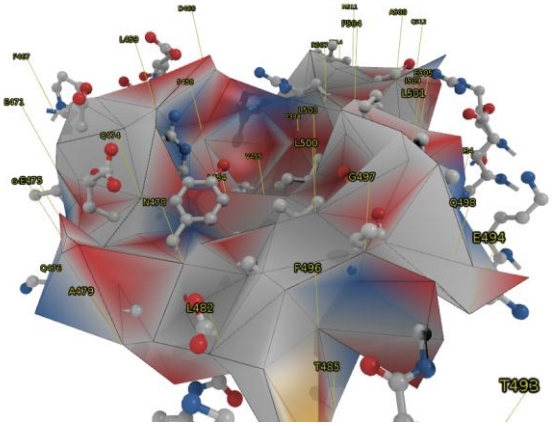
Site de dimérisation de RXRα (surface de forme alpha) aligné avec le site théorique d'un récepteur nucléaire (modèle boules et bâtons)	**Site de dimérisation de RXRα (modèle boules et bâtons)** aligné avec le site théorique d'un récepteur nucléaire (surface de forme alpha)
 **RXRg (2gl8 : B)** = 59,7	
 **RXRb (1h9u : B)** = 54,3	
 **HNF4a (1pzl : A)** = 51	



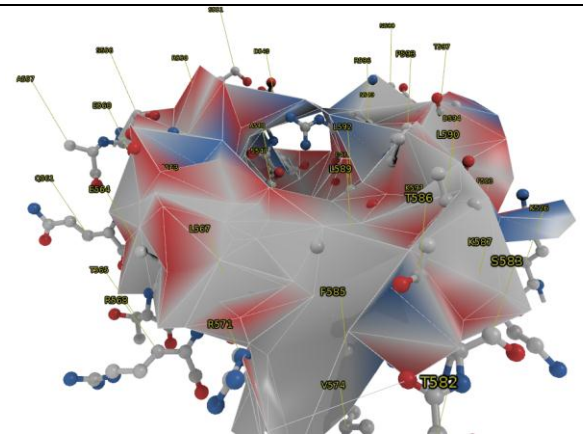


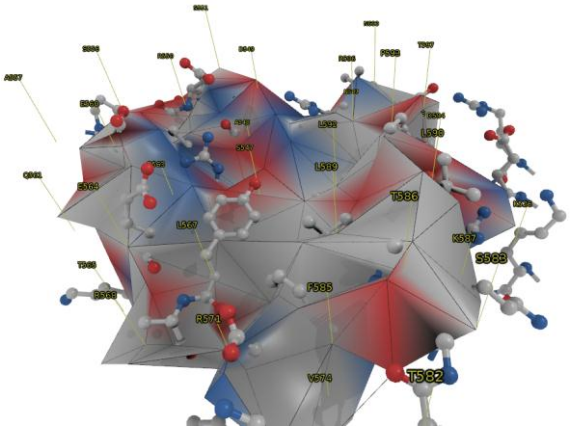
HNF4g (1lv2 : A) = 50,9



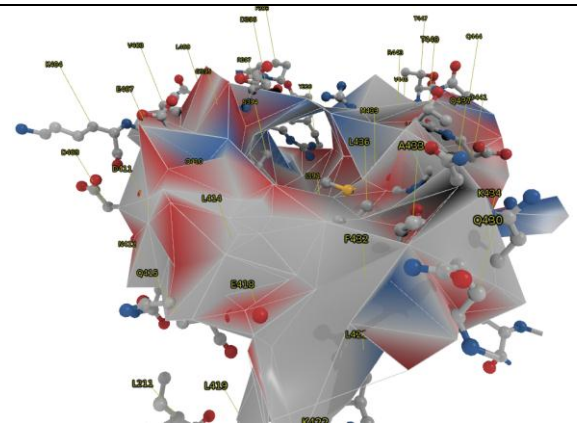


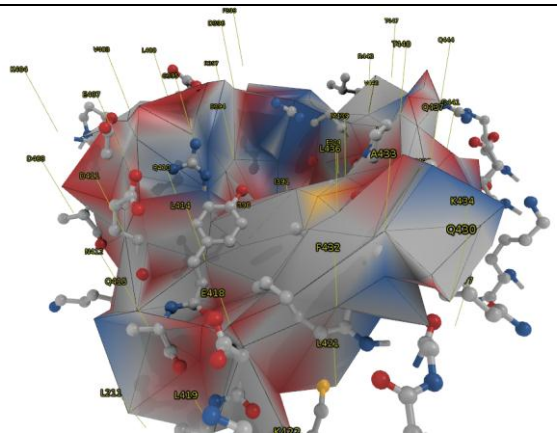
NR5A2(1yuc : A) = 49,9



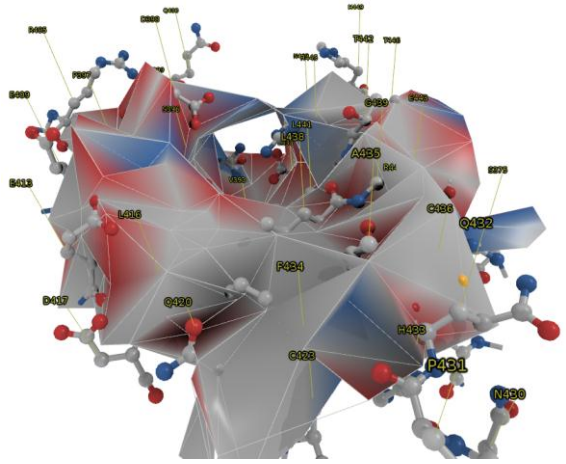


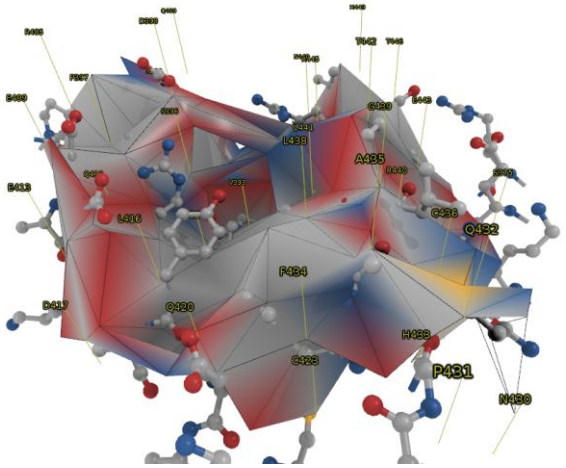
REV-ERBAa (3n00 : A) = 44,7



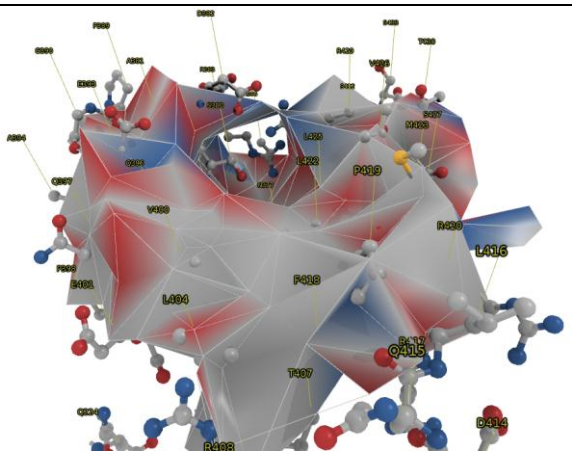


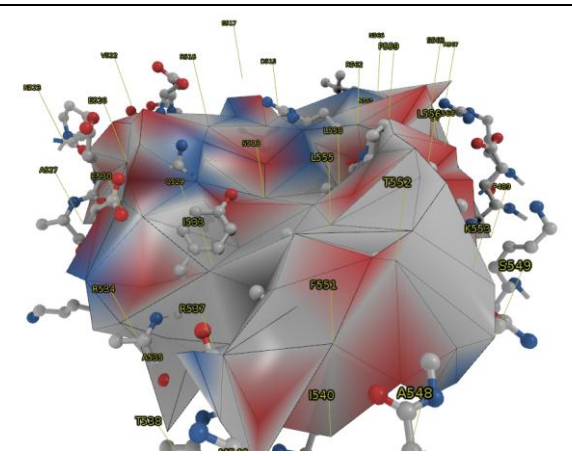
PPAT (4prg : C) = 41,14



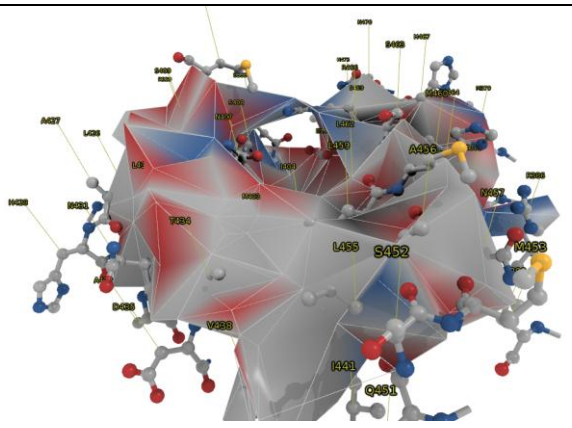


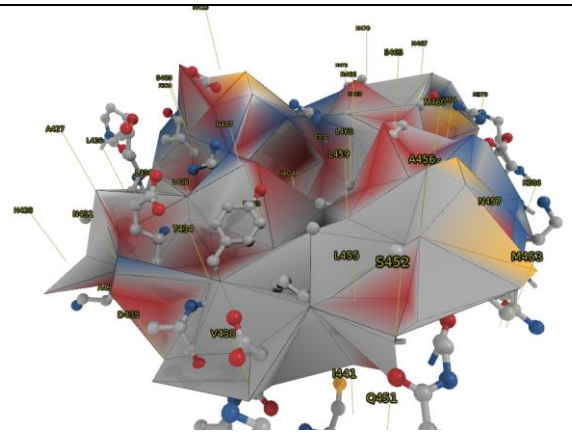
NR1H4 (1osh : A) = 40,85



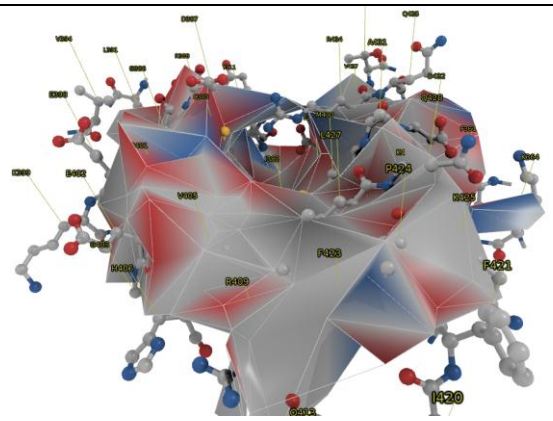


NR1H2 (1pq6 : A) = 39,34

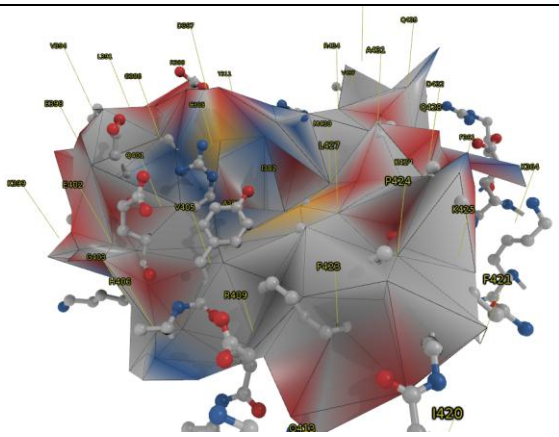


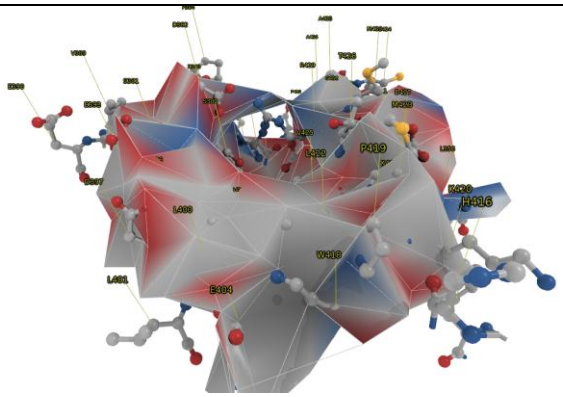


ESR2 (1yy4 : A) = 39,14

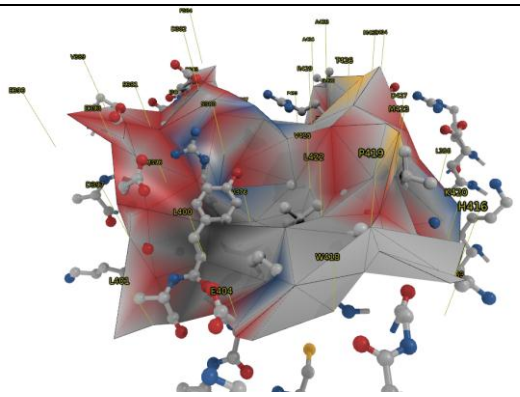


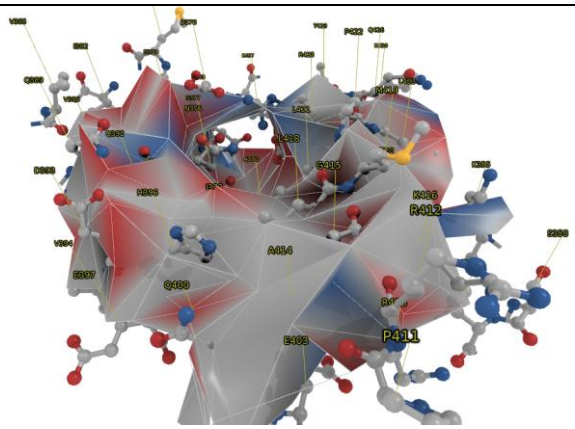
PPARa (1i7g : A) = 36,82



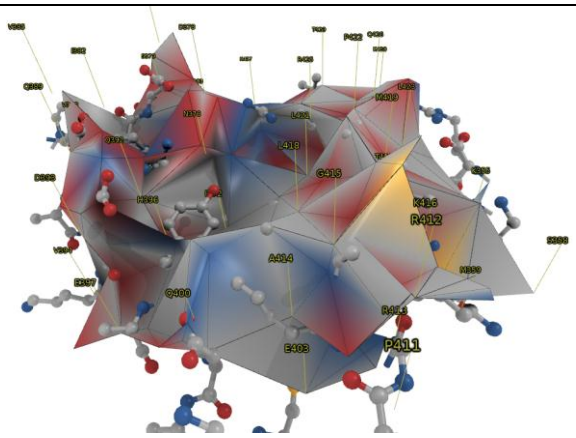


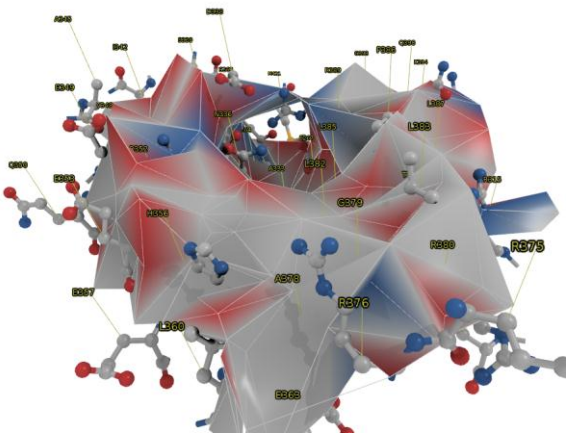
THRB-1 (1n46 : A) = 36,14



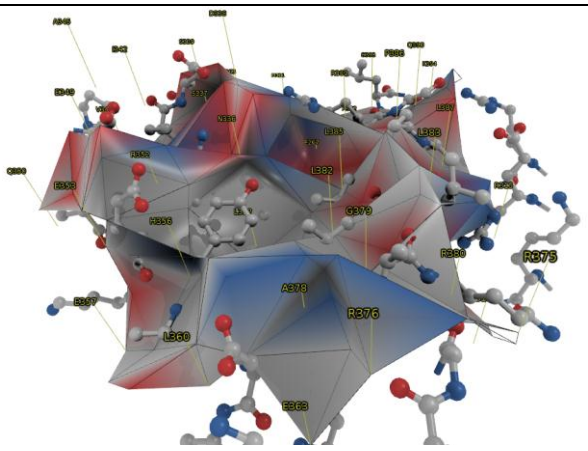


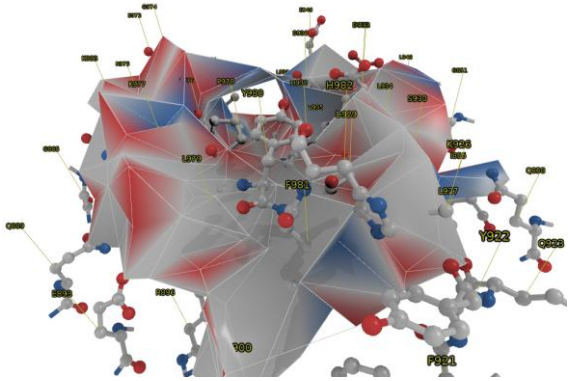
ERR3 (2gp7 : A) = 36,02



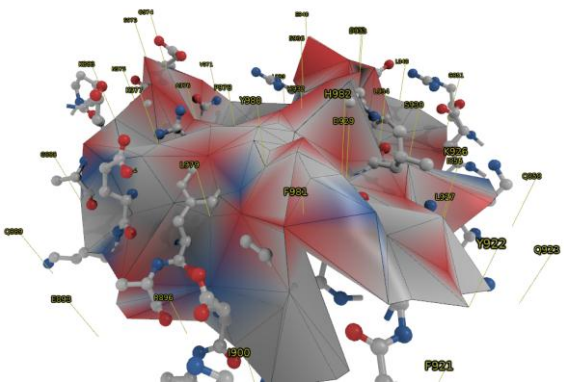


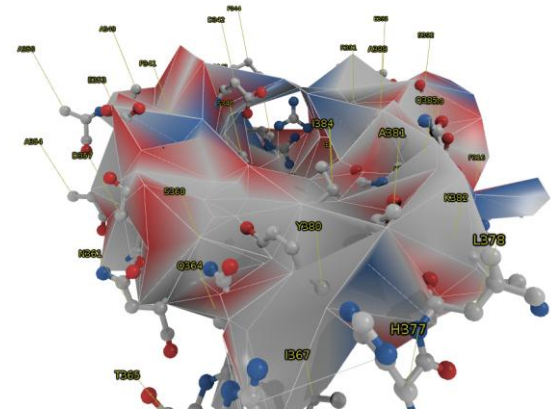
ERR1 (3d24 : C) = 35,73



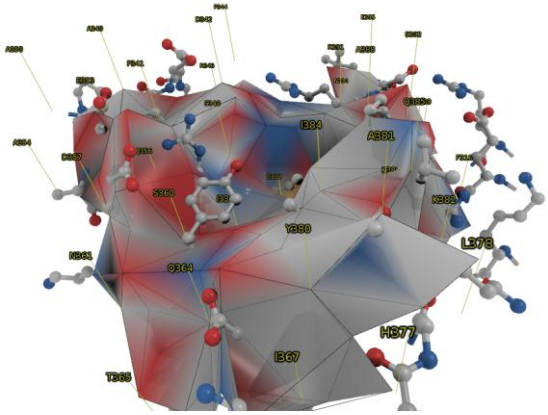


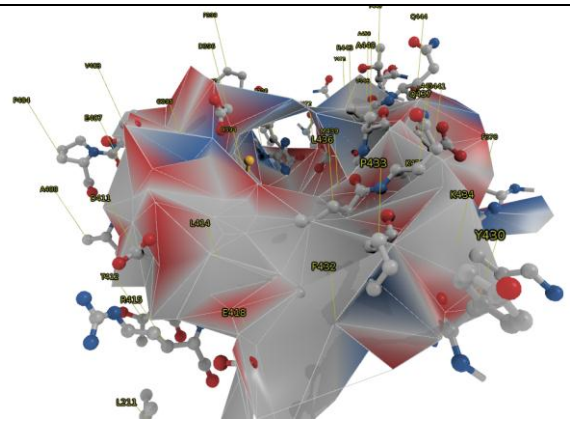
MCR (2aa : X) = 35,66



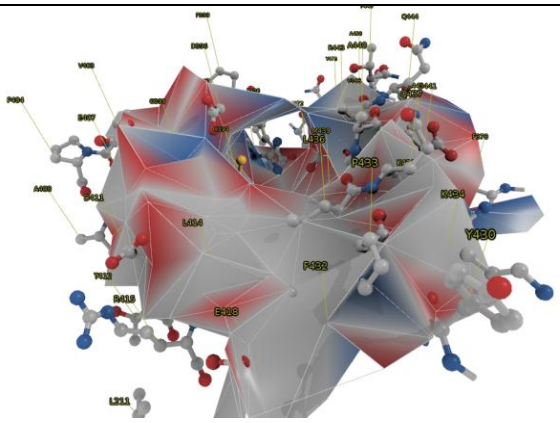


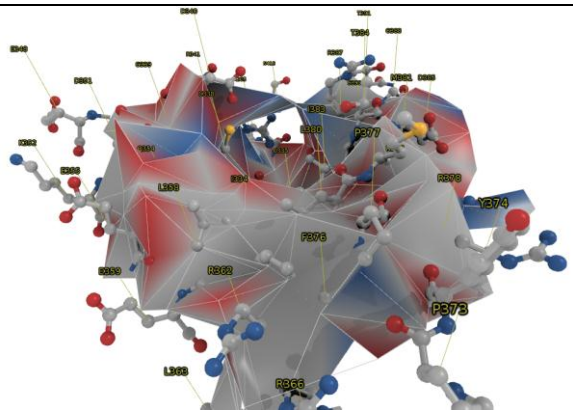
VDR (2hb8 : A) = 34,9



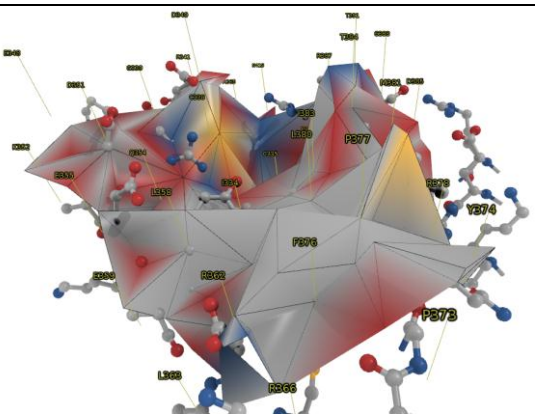


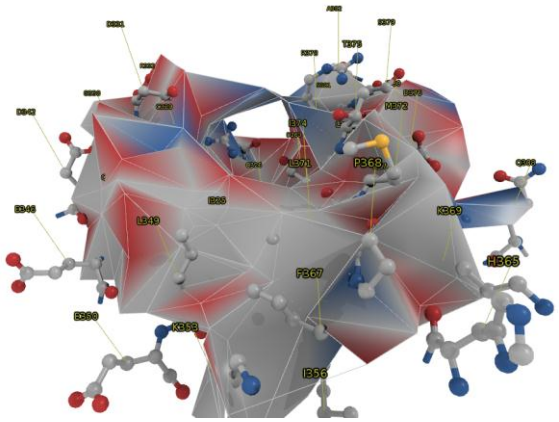
PPARD (3dy6 : B) = 34,52

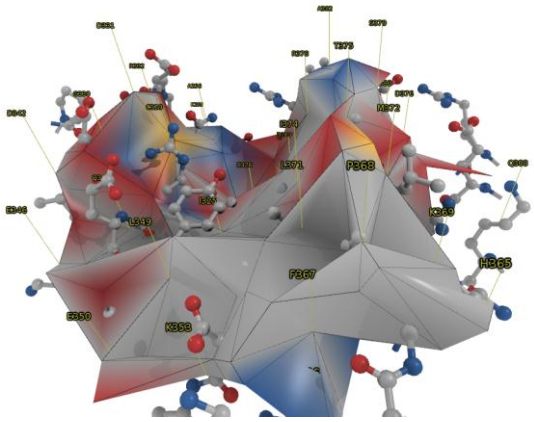




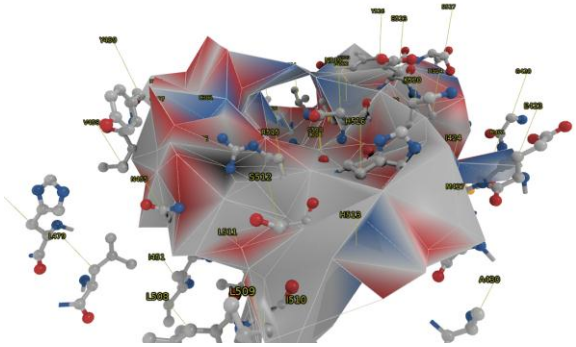
RARg (3lbd) = 33,9

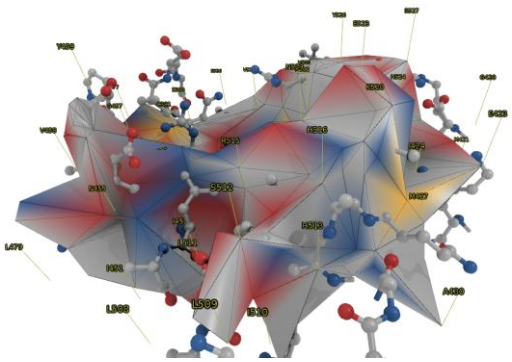




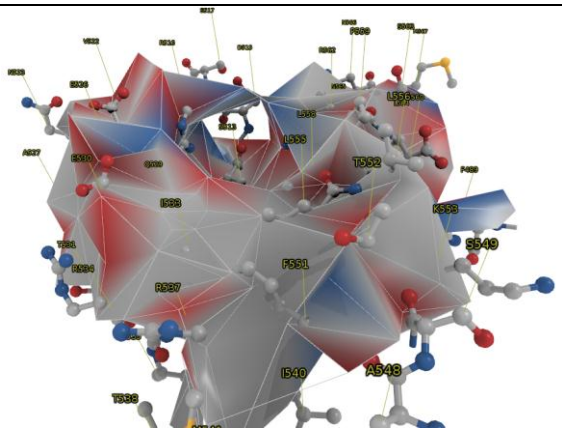


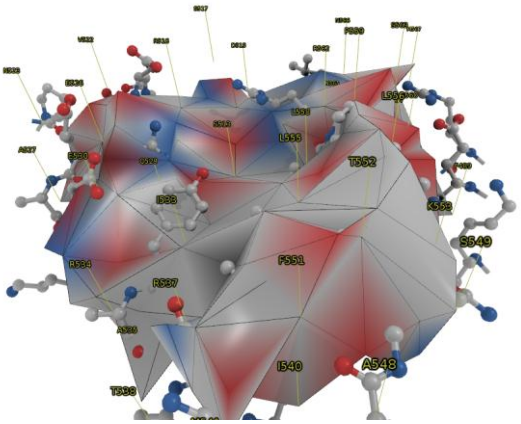
RARb (1xap : A) = 33,61



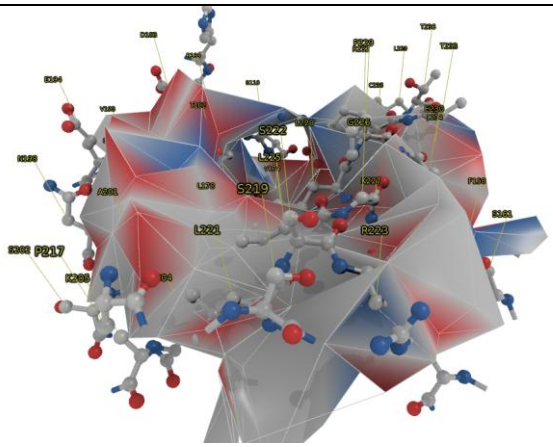


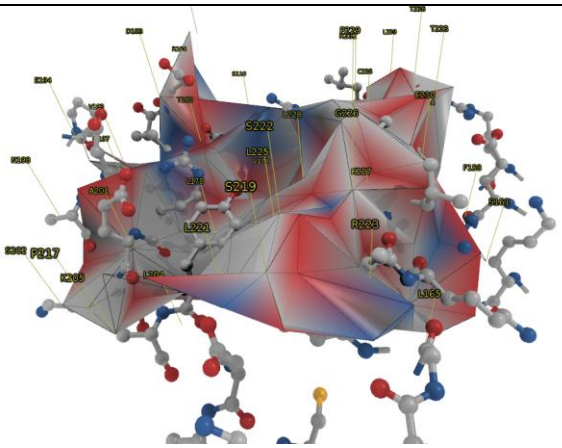
ESR1 (2qxs : B) = 33,06



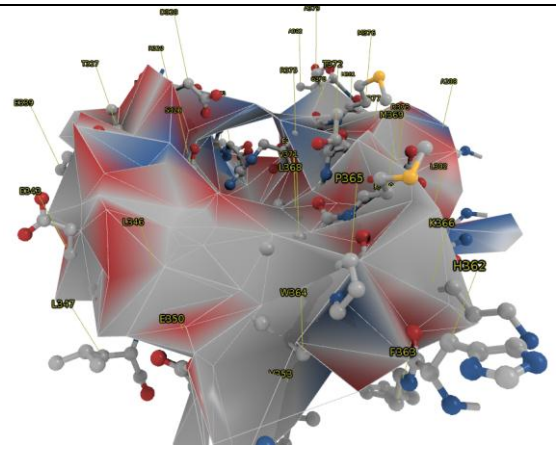


NR1D2 (2v0v : B) = 32,72

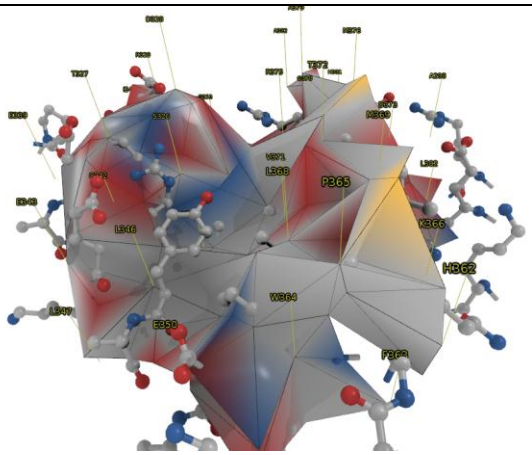


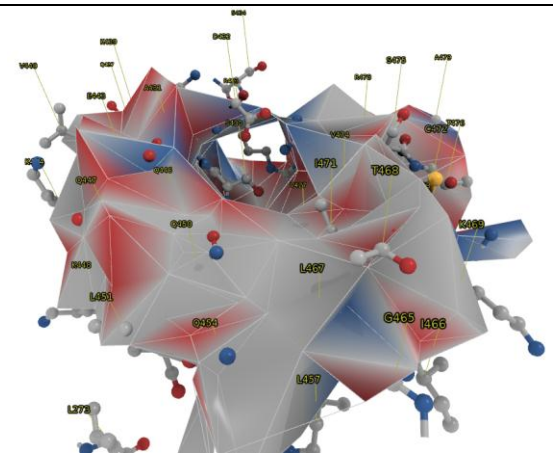


NR4A1 (2qw4 : A) =32,68

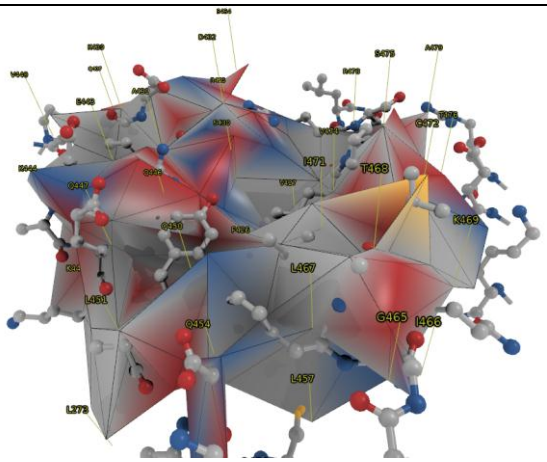


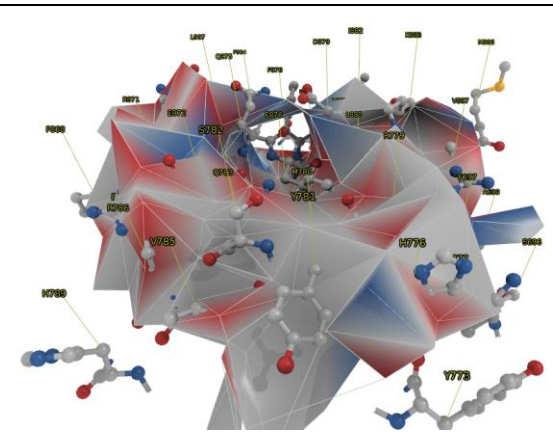
THRA (2h77 : A) = 32,05



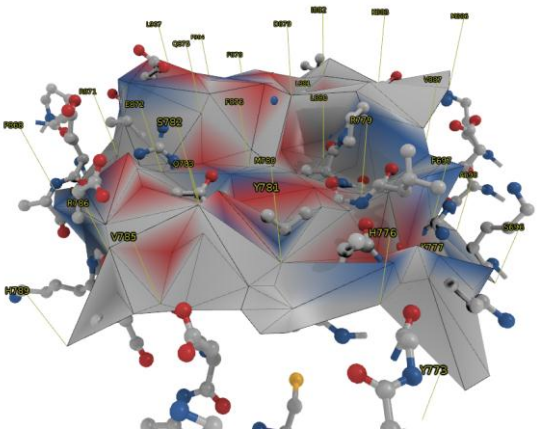


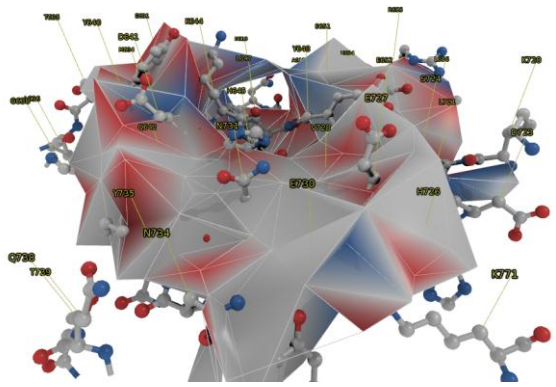
RORa (1s0x : A) = 32,04

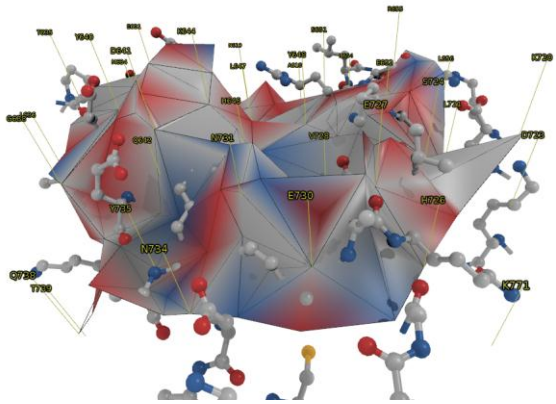




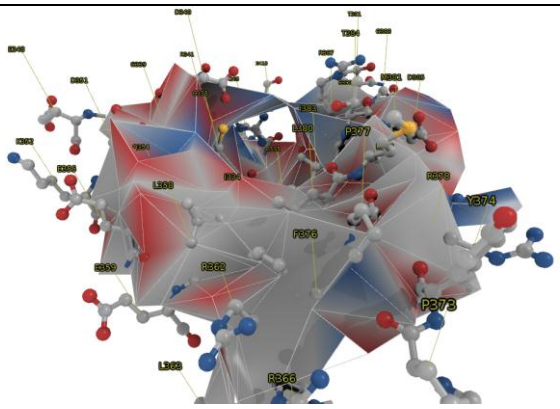
AR (2piq : A) = 31,99

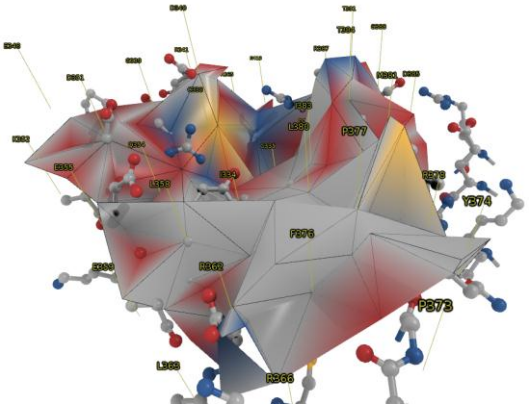




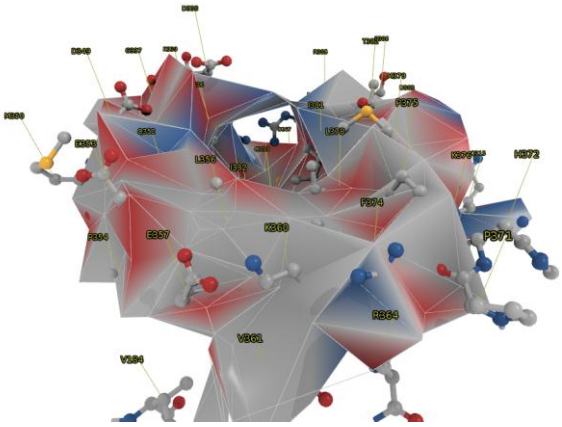


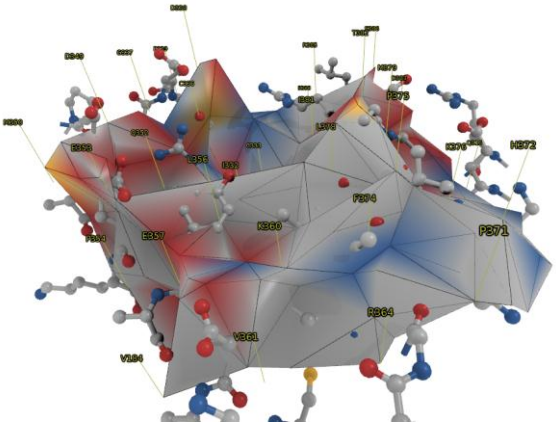
GR (3k23 : B) = 31,63



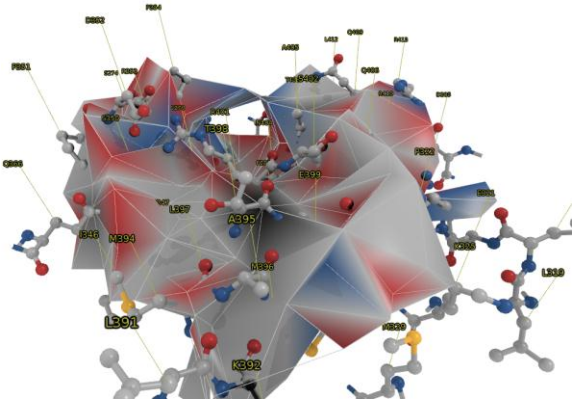


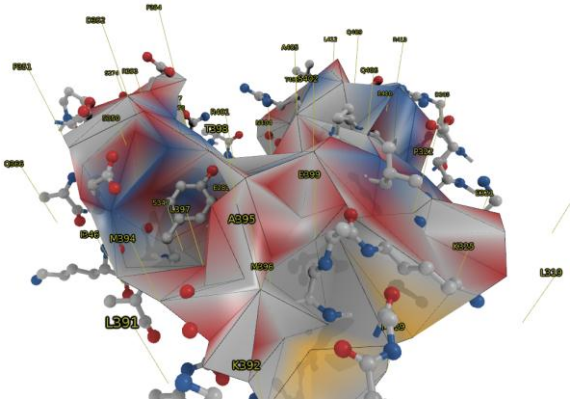
RARg-1 (1fcz : A) = 31,45



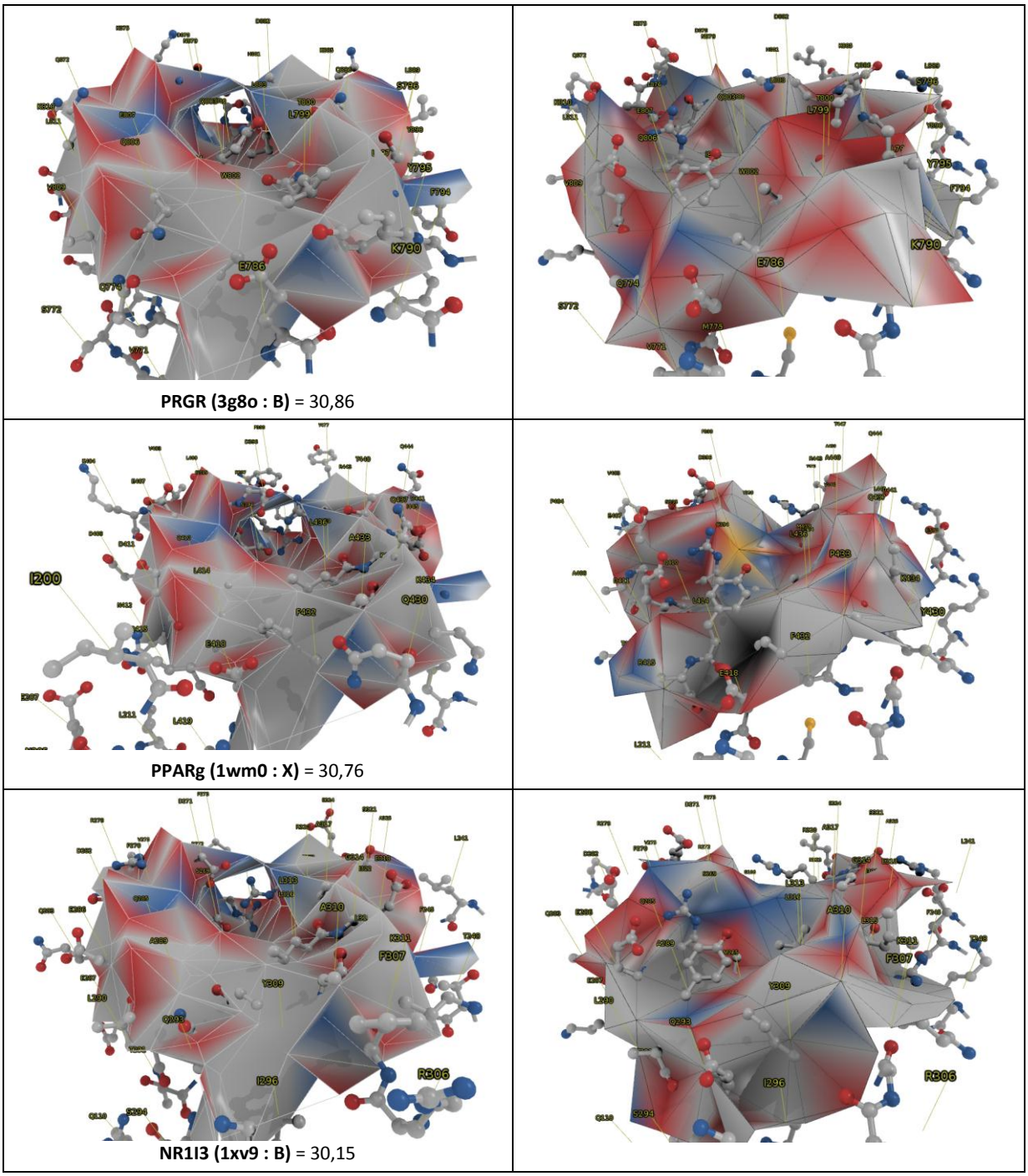


RARa (1dkf : B) = 30,98





PXR (3hvl : B) = 30,9



PRGR (3g8o : B) = 30,86

PPARg (1wm0 : X) = 30,76

NR1I3 (1xv9 : B) = 30,15

Tableau 27. Alignement par paires du site de dimérisation de RXRα avec les sites de dimérisation théoriques d'autres RN. *La colonne de gauche représente le site de dimérisation de RXRα sous forme alpha, aligné avec le site de dimérisation d'un récepteur nucléaire représenté avec le modèle boules et bâtons. Inversement, la colonne de droite représente le site de dimérisation de RXRα avec le modèle boules et bâtons, aligné avec le site de dimérisation d'un récepteur nucléaire représenté par les formes alpha. Le code couleur pour les atomes est celui couramment usité (carbone : gris, azote : bleu, oxygène : rouge). Les récepteurs nucléaires sont ordonnés en fonction de leurs ressemblances au site de dimérisation de RXRα. Le code couleur pour la forme alpha est basé sur la charge partielle*

des atomes : gris pour une charge nulle (hydrophobe), rosé pour une charge partielle négative, rouge pour une charge négative, bleu ciel pour une charge partielle positive, et bleu foncé pour une charge positive. L'observation des formes alphas de la colonne de droite permet d'observer la variabilité de formes et de charges des sites de dimérisation. Les deux scores indiqués représentent respectivement la ressemblance géométrique, et la ressemblance des charges partielles.

Il est également intéressant de noter que les récepteurs AR, GR, PRGR, NURR1 et MCR ont une extrémité C-terminale plus longue que celles des autres RN. Cette extrémité vient interférer avec la forme et la distribution des charges électrostatiques du site de dimérisation, ce qui précise un comportement différent de tous les autres RN vis-à-vis de la dimérisation.

Afin de mieux appréhender ce phénomène de reconnaissance moléculaire, nous avons calculé la distribution du score d'énergie issu de nos comparaisons de régions, et avons comparé cette distribution au pourcentage d'identité des récepteurs nucléaires avec la séquence de RXRα (Figure 76).

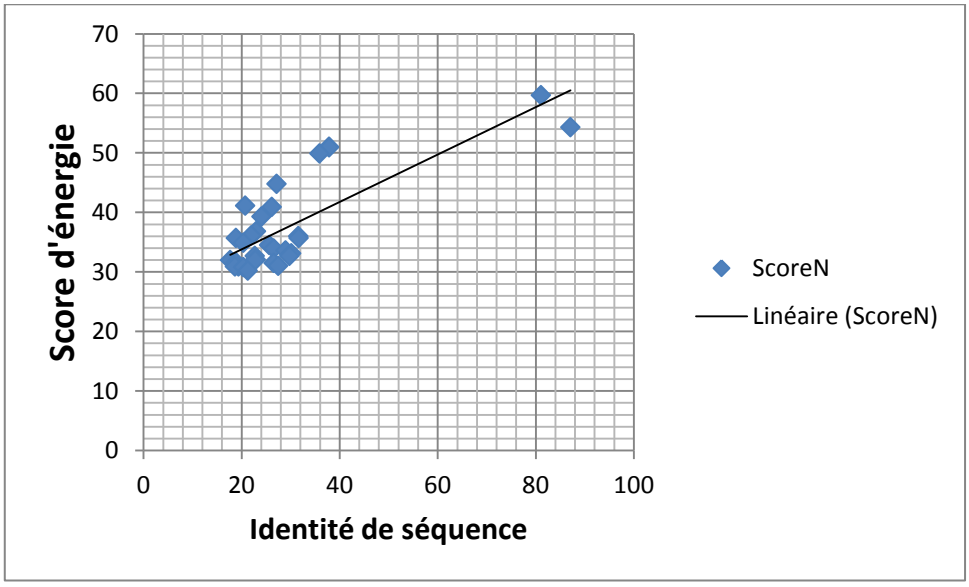


Figure 76. Corrélation entre le score d'énergie et l'identité de séquence pour la famille des récepteurs nucléaires. *Le score d'énergie correspond au score d'énergie normalisé (ScoreN) entre 0 et 100 et renseigne sur la similarité géométrique et physico-chimique des sites de dimérisation. L'identité de séquence renseigne sur la ressemblance des séquences des LBD portant ces sites de dimérisation.*

Il est de plus possible d'étudier la corrélation entre la ressemblance en termes de séquence des récepteurs nucléaires avec les différentes composantes de notre score d'énergie. Notamment, quatre composantes renseignent sur : la position des atomes, leur normale à la surface, la courbure locale, et la charge partielle (Figure 77).

Ces résultats signifient que pour l'étude des fonctions d'une protéine au sein d'une même famille, l'identité de séquence peut être utilisée comme première approximation sur la ressemblance du site de dimérisation (corrélation de Pearson de 0,77). Par ailleurs, l'information contenue dans la séquence corrèle davantage avec la position spatiale des résidus et leurs charges (types), qu'avec la courbure locale et l'orientation des résidus (normal) de la région.

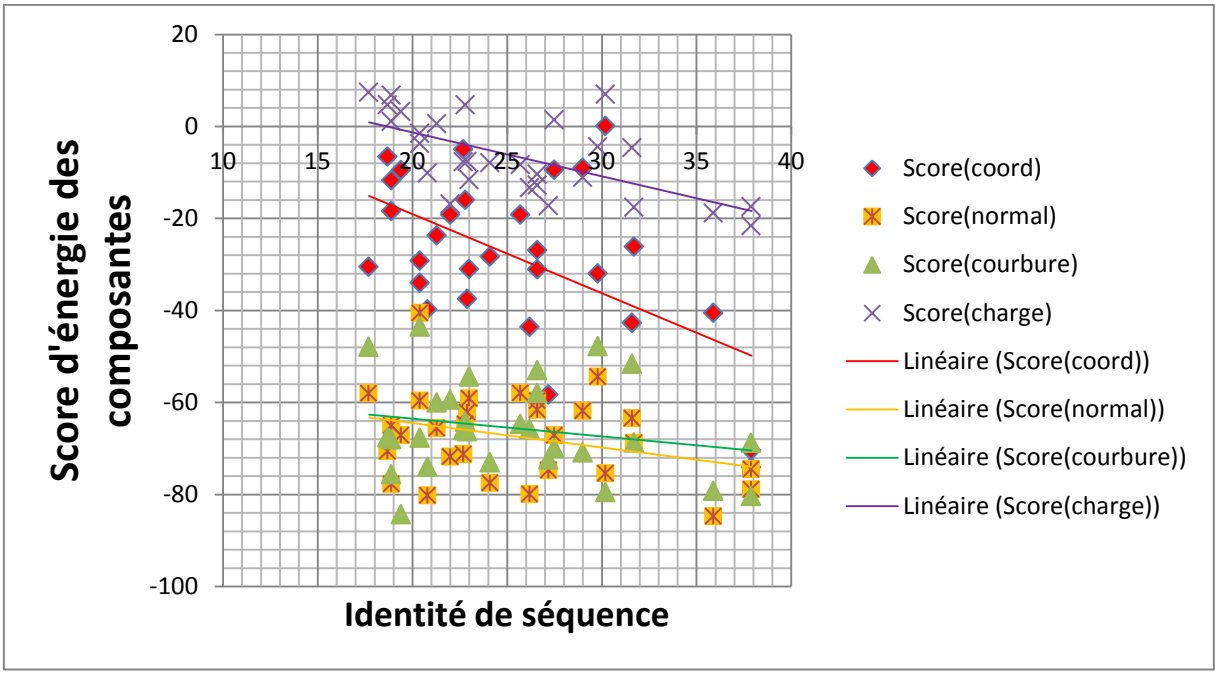


Figure 77. Corrélation des composantes du score d'énergie et l'identité de séquence pour la famille des récepteurs nucléaires. *Les composantes de coordonnées (coord) et de charge corrèlent le mieux avec l'identité de séquence, avec des coefficients de Pearson respectivement de 0,65 et 0,77. Les composantes de courbure et de normal corrèlent moins avec l'identité de séquence, avec des coefficients de Pearson respectifs de 0,22 et 0,44.*

Ces corrélations n'ont toutefois de sens que lorsque nous limitons l'étude à une seule famille de protéines présentant un même repliement commun, comme pour les récepteurs nucléaires. Dans ces conditions, il n'est en effet pas possible de trouver une protéine très différente en termes de séquence et de structure qui présenterait pourtant une même région. Par la suite, nous décrirons notamment le criblage du Dorzolamide, un médicament

contre l'hypertension intraoculaire, qui semble pouvoir lier de nombreuses autres cibles que sa cible souhaitée, l'anhydrase carbonique. Testée sur 76 sites de liaison ayant plus de 68% de similarité structurale et physico-chimique avec le site de liaison au Dorzolamide de l'anhydrase carbonique, aucune corrélation n'est observable entre identité de séquence des cibles, et similarité de ces sites de liaisons (Figure 78). Par ailleurs, ces autres cibles ont pour la plupart moins de 20% d'identité de séquence avec l'anhydrase carbonique. Il serait par conséquent difficile sinon impossible de les identifier par comparaison de séquences ; la plupart de ces cibles ne partagent par ailleurs pas un même repliement.

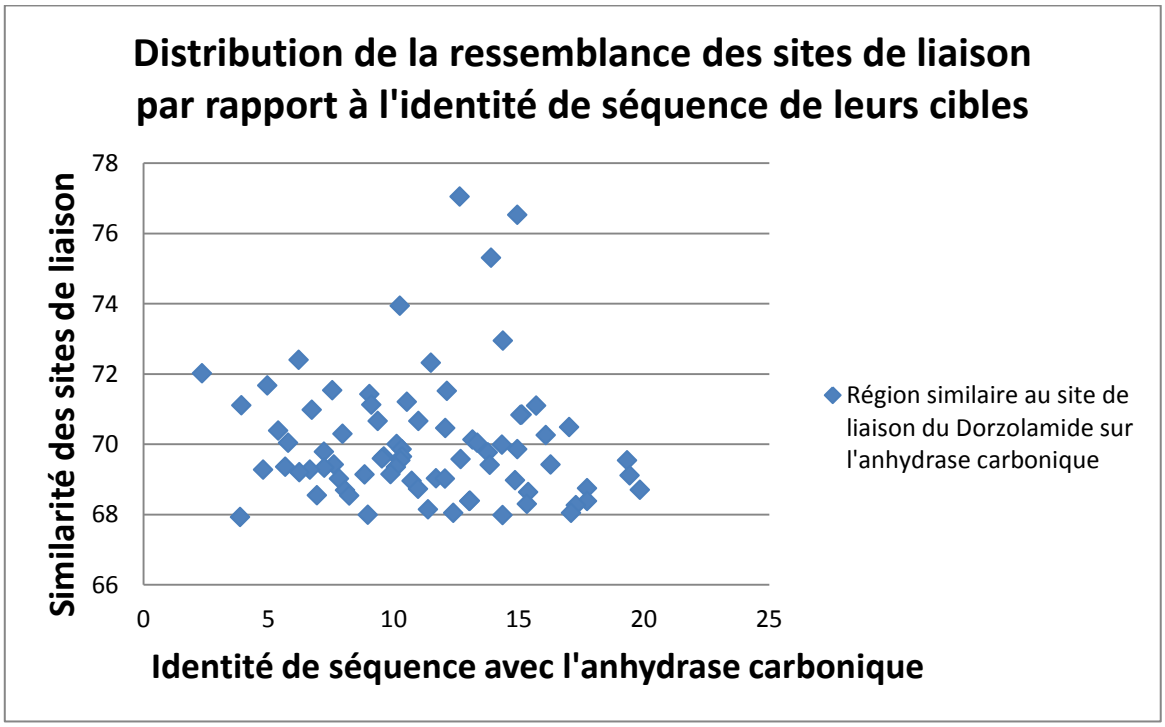


Figure 78. En dehors des familles protéiques, aucune corrélation entre l'identité de séquence des cibles et la ressemblance de leurs sites de liaison n'est observable. *Si l'on considère des cibles moléculaires ne faisant pas partie de la famille des anhydrases carboniques, alors aucune corrélation ne peut être observée entre la ressemblance des séquences des cibles, et la ressemblance de leurs sites de liaison : corrélation de Pearson de -0,06.*

3. Arbre fonctionnel du site de dimérisation

La Figure 79 représente l'arbre fonctionnel des RN pour le site de dimérisation. La construction de cet arbre a reposé d'une part sur le calcul des similarités entre tous les sites de dimérisation des RN retrouvés par notre approche de criblage de régions, et d'autre part, sur la méthode de regroupement UPGMA. Les RN regroupés dans des branches reliées par un nœud commun sont attendus de partager des capacités de dimérisation proches.

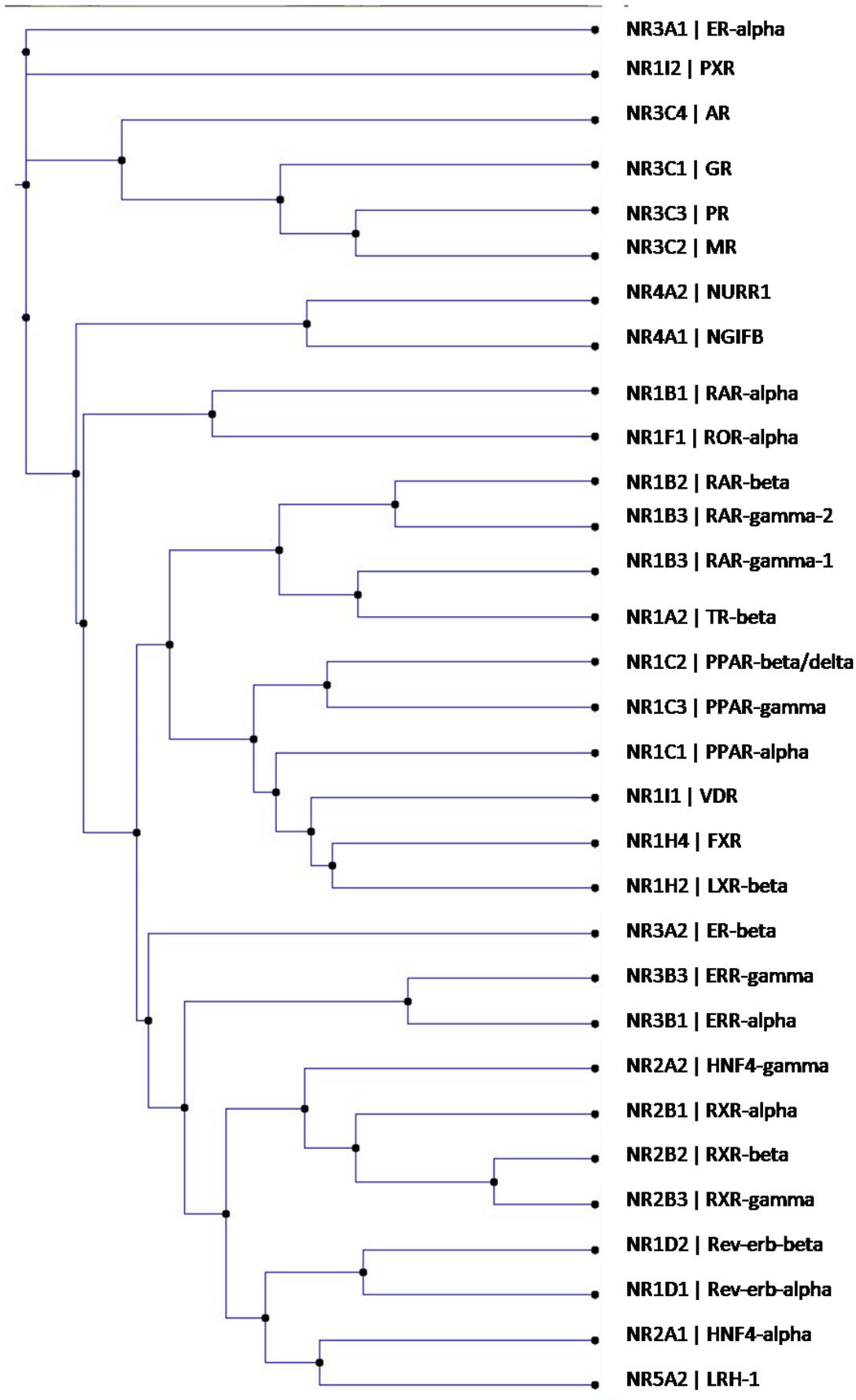


Figure 79. Arbre fonctionnel du site de dimérisation des récepteurs nucléaires humains.

L'arbre fonctionnel des sites de dimérisation de 31 RN peut être comparé à l'arbre évolutif des RN renseigné à la Figure 32. Les deux arbres comportent des similitudes (la réunion de la sous-famille RXR avec la sous-famille HNF4), mais aussi des différences notables (RXRβ et RXRγ vus plus proches que RXRα et RXRβ).

Bien que les récepteurs LXR-alpha (NR1H3, PDB : 1uhl : B), ROR-gamma (NR1F3, PDB : 3kyt : A), TR-alpha (NR1A1, PDB : 1nav : A), COUP-TFII (NR2F2, PDB : 3cjw : A) et CAR (NR1I3, PDB : 1xv9 : B) soient présents dans la B_{RN}, et aient effectivement été retrouvés lors des criblages, nous prions les lecteurs de bien vouloir excuser leurs omissions dans les Figures 79 et 80. Ces résultats seront toutefois présentés à la soutenance.

4. Arbre fonctionnel du site de liaison au cofacteur

Il est possible de réaliser l'arbre fonctionnel du site de liaison au cofacteur des RN par un procédé similaire à celui utilisé pour la création de l'arbre fonctionnel du site de dimérisation (Figure 80). Les RN réunis sur des branches avec un nœud commun devraient pouvoir lier des cofacteurs relativement similaires. Ainsi par exemple, AR et LXR-beta sont capables de lier des petits peptides quasiment identiques (ENALLRYLLDKDD et KENALLRYLLDK), de même pour RXR-alpha, ER-alpha et PXR (KHKILHRLLQD, KILHRLLQD, SLTERHKILHRLLQE).

Si l'arbre fonctionnel reflète suffisamment la spécificité de reconnaissance des RN pour leurs cofacteurs, alors il implique par exemple que PPAR-beta/delta puisse lier un cofacteur proche de celui de PPAR-gamma (ERHKILHRLLQEGSPS). En termes de régulation de l'expression des gènes, cet arbre fonctionnel indique (1) que les RN réunis sous un même nœud sont susceptibles de lier des cofacteurs relativement proches, et que (2) ces RN puissent être régulés simultanément par la présence de ces cofacteurs.

La comparaison par paires des régions de liaison au cofacteur des RN permet d'observer des différences au niveau atomique et d'expliquer de possibles variations d'affinité d'un RN pour certains petits peptides.

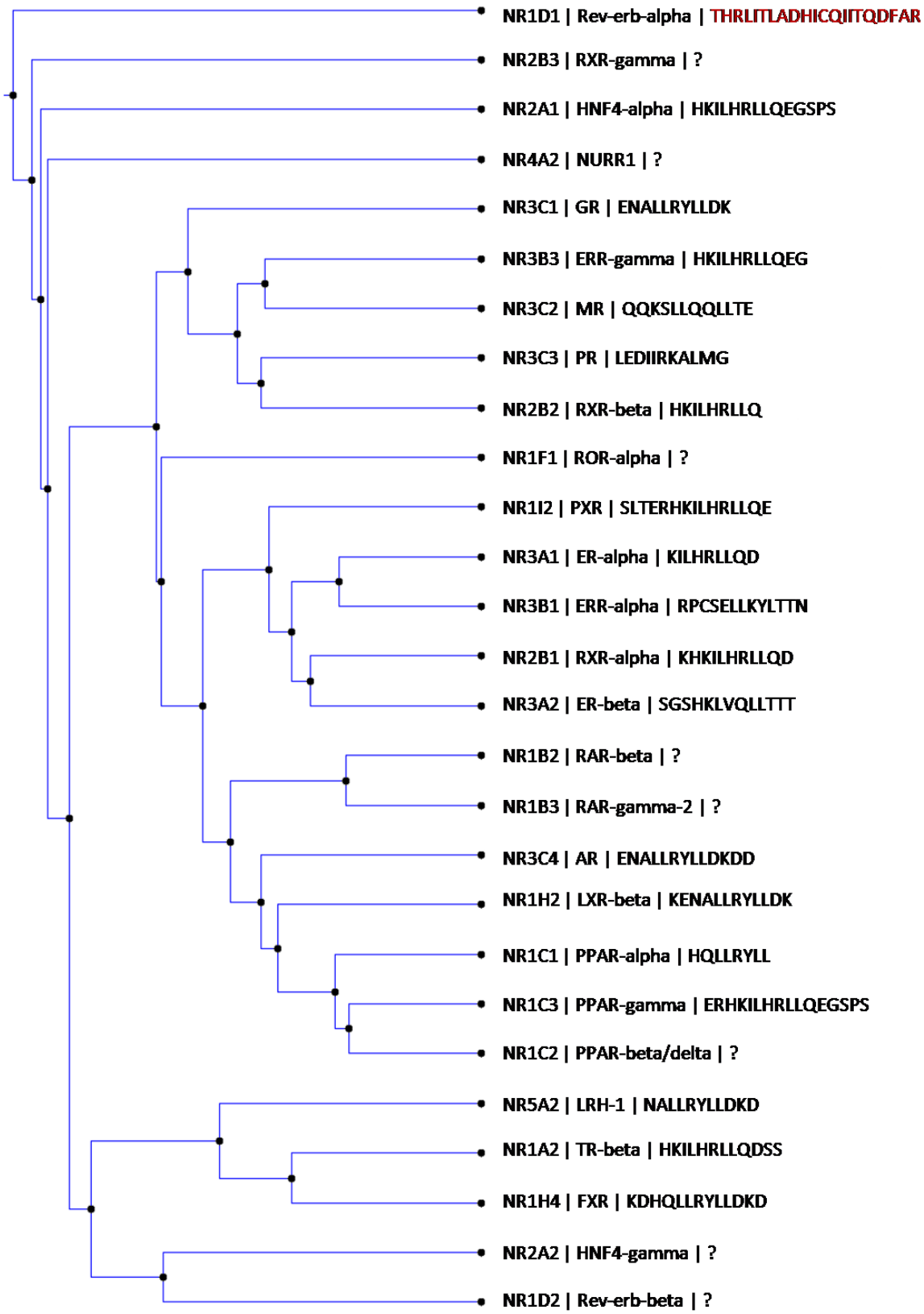


Figure 80. Arbre fonctionnel du site de liaison au cofacteur des récepteurs nucléaires humains. *Les séquences des cofacteurs structuralement connus pour lier les RN sont indiquées en fin de ligne. Un « ? » signifie que structuralement, aucun cofacteur n'a été identifié. Un co-répresseur est indiqué en rouge pour NR1D1.*

Il est intéressant de s'imaginer qu'il est possible de réaliser un travail comparable sur le site de liaison aux ligands des RN (afin de comparer leurs aptitudes à lier des ligands plus ou moins proches), ou bien sur le site de liaison à l'ATP de la famille des kinases qui contient plus de 1000 structures dans la Protein Data Bank. Dès lors, ces arbres fonctionnels vont renseigner sur les capacités communes des protéines étudiées, à accomplir une fonction ; que ce soit lier un partenaire, catalyser une réaction ou phosphoryler un résidu.

5. La famille des GR, MR, PR, AR et NURR1

Cette sous-famille de récepteurs nucléaires est relativement spéciale en ce sens que leurs extrémités C-terminales affectent directement la conformation du site de dimérisation et partiellement le site de liaison au cofacteur (Figure 81). Les criblages réalisés sur cette famille montrent en effet qu'il est parfois difficile de détecter ces régions : les régions les plus proches des sites habituels de dimérisation et de fixation de cofacteur diffèrent parfois dans leur localisation spatiale sur ces récepteurs. En l'absence de changements conformationnels impliquant le déplacement de cette extrémité C-terminale, il est en effet probable que soit ces récepteurs ne puissent dimériser, soit que leurs dimérisations impliquent des remaniements importants de l'extrémité C-terminale, soit que des régions d'interactions, différentes des sites de liaison usuels pour la famille des RN, sont impliquées.

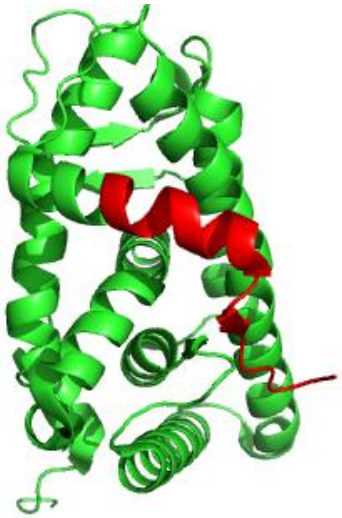


Figure 81. Extrémité C-terminale prolongée d'une sous-famille de RN modulant la forme et la distribution des charges du site usuel de dimérisation. *La structure du récepteur au glucocorticoïde (GR) a été choisie pour illustrer ce phénomène (PDB : 3E7c). L'extrémité C-terminale modifiant le site de dimérisation est indiquée en rouge.*

V. Criblage de régions fonctionnelles et détection d'interactions

Dans la section précédente, nous avons effectué le criblage de deux régions fonctionnelles dans un environnement spécifique ne comportant que les structures de récepteurs nucléaires. Ces criblages nous ont permis d'évaluer les ressemblances et différences fonctionnelles de ces récepteurs en fonction de ces deux régions. Toutefois, en considérant des environnements plus grands, soit représentatifs de l'ensemble des molécules contenues dans la Protein Data Bank, soit représentatifs par exemple des molécules humaines (en se limitant aux seules structures humaines), il est possible de découvrir des similarités fonctionnelles entre des récepteurs nucléaires et des molécules éloignées ne présentant aucune similarité de séquence, de structure secondaire ou tertiaire. En particulier, l'homologie à l'échelle des régions structurales peut être utilisée pour inférer des fonctions (exemple : capacité à lier une protéine, un acide nucléique, un ligand, etc..) à des structures moléculaires pour lesquelles peu ou pas de données biologiques sont disponibles.

Le procédé de criblage de régions moléculaires peut être appliqué dans deux sens :

- soit en criblant une région pour laquelle une fonction est déjà connue : on recherche alors l'ensemble des régions similaires pour leur inférer cette fonction. C'est le cas notamment du criblage des sites de liaison de macromolécules ou de molécules, qui permet de retrouver plusieurs cibles cellulaires capables de lier ces mêmes macromolécules ou molécules. Les exemples de l'ATP et du Dorzolamide sont donnés en annexe de cette thèse.

- soit en criblant une région pour laquelle aucune fonction n'est connue : on recherche alors si, parmi l'ensemble des régions similaires récupérées, l'une d'entre elles a une fonction déjà identifiée. Le procédé de caractérisation d'une macromolécule consiste donc à la découper en régions contiguës, puis de cribler chacune de ces régions pour identifier des régions similaires pour lesquelles une fonction est connue. Les fonctions des régions similaires sont ensuite rapportées sur la molécule à caractériser. Cette application du criblage virtuel de régions moléculaires constitue l'une des évolutions futures de la cartographie moléculaire M-ORBIS et permettra d'identifier des fonctions à partir d'une similarité de régions, et non d'une similarité de structure tertiaire.

1. Modélisation de l'environnement par les structures moléculaires

Plus généralement, le criblage de régions moléculaires nécessite la modélisation d'un environnement cellulaire (ou autre) à partir de la sélection des molécules en présence. Dans

l'état des développements présents, il est facile de créer des environnements correspondant aux protéines des différents organismes recensés dans les bases de données, pour peu qu'une structure existe ou soit modélisable. Dans les mois qui suivront la soutenance de cette thèse, nous concentrerons nos efforts sur la génomique fonctionnelle afin de bénéficier des données de transcriptomique. En effet, pour modéliser plus précisément un environnement cellulaire, nous aurons besoin de connaître la présence et la fréquence d'une molécule dans l'environnement. Ce point est particulièrement important lorsque l'on cherche à mesurer l'efficacité ou les effets secondaires/toxiques que peut avoir une molécule dans l'environnement. En effet, si l'on considère qu'un composé est par exemple capable de lier un très grand nombre de cibles cellulaires, mais que celles-ci ne sont pas ou peu exprimées dans le type cellulaire considéré, il sera nécessaire d'en tenir compte dans l'évaluation de l'efficacité. De la même façon, lorsque nous chercherons à évaluer la spécificité (la fréquence) de régions composant le « paysage moléculaire » d'une cellule, il sera impératif de connaître la fréquence des molécules portant ces régions.

On comprend également que si deux environnements diffèrent par leurs molécules en présence, alors l'introduction d'un composé pourra ne pas provoquer les mêmes réponses cellulaires, puisqu'il ne participera pas nécessairement aux mêmes interactions moléculaires. En d'autres termes, la fonction d'une molécule, tout comme la fonction d'un gène, n'est pas définissable lorsque ces entités sont considérées seules. En revanche, il est possible de définir les fonctions d'une molécule ou d'un gène, à partir de ses interactions avec l'environnement, et de la pression de sélection qu'exerce l'environnement sur ces entités. La recherche de médicaments se concentre depuis quelques années sur l'étude des différentes réponses que peut provoquer un composé dans différents types cellulaires. La modélisation suffisamment « proche » du contenu des différents types cellulaires nous permettra également d'évaluer les interactions entre molécules pour ces types cellulaires, ainsi que de mieux évaluer les spécificités des régions, et les problèmes d'efficacité et effets secondaires/toxiques des molécules.

2. Perspectives sur la modélisation des environnements cellulaires à partir des structures moléculaires

Nous avons souhaité, au cours du premier chapitre d'introduction, remémorer et resituer quelques notions importantes sur la cellule, sur la variabilité génétique et sur l'environnement qui détermine les fonctions d'une molécule. En effet, il est conceptuellement facile d'imaginer que, si ce sont les jeux d'interactions auxquels participent une molécule, qui définissent sa fonction, alors une même molécule n'a pas

nécessairement une fonction identique dans deux environnements différents car elle ne participe pas nécessairement aux interactions dans ces deux environnements.

Lors de nos criblages, nous avons modélisé uniquement l'environnement en fonction des molécules présentes chez l'homme ou un organisme. L'information du type cellulaire ou de la localisation sous-cellulaire (organite) n'a pas encore été prise en compte. Toutefois, des développements futurs sont prévus pour les prendre en compte, utilisant des données de génomiques fonctionnelles, par exemple issu de la base de données GeneAtlas maintenue à Paris [331, 332].

De plus, avec l'apparition des exomes (plus facile à séquencer qu'un génome complet) et les dernières innovations de séquençage, il devient facile, rapide et peu coûteux de déterminer l'ensemble des protéines exprimées d'un ou plusieurs individus représentatifs d'une population. Cela signifie que nous entrons dans une ère où il sera possible de déterminer les variations génétiques différenciant des groupes de population, ou bien différenciant des groupes d'individus présentant un phénotype « normal » de ceux présentant des phénotypes pathologiques.

L'un des apports majeurs des approches de modélisation de structures par homologie sera sans nul doute de permettre de modéliser rapidement et efficacement les structures des protéines de ces groupes de population, en tenant compte de leurs variabilités génétiques, et ce, sans avoir à redéterminer expérimentalement les structures de ces protéines présentant des variations. En effet, on considère que, au-delà de 90% d'identité de séquence avec une structure homologue, les approches de modélisation de structures par homologie atteignent une qualité de résolution cristallographique [132]. En d'autres termes, il sera possible de modéliser les environnements cellulaires de groupes de population à partir de leurs protéines exprimées et des structures moléculaires correspondantes. Les deux ombres qui se cachent derrière ces modélisations sont : d'une part la dynamique conformationnelle des molécules, puisqu'il ne faut pas oublier que les structures disponibles sont des « instantanés » des molécules qui ne reflètent pas cette dynamique, et d'autre part, qu'il est pour le moment impossible de modéliser la plupart des structures d'ARN dont les rôles dans la cellule et le nombre sont pourtant importants.

La méthode de criblage virtuel haut-débit des régions moléculaires que nous avons développée pourrait s'intégrer ensuite dans ces démarches afin de permettre d'identifier des différences fines dans les jeux d'interactions de deux groupes d'individus, soit représentatifs de deux populations, soit représentatifs de phénotypes « sauvages » et de phénotypes « pathologiques ».

3. Spécifité d'une région moléculaire

Le criblage de nombreuses régions moléculaires sur les 66 000 structures biologiques connues, montre une très grande disparité dans leurs fréquences. Certaines régions sont ainsi très représentées, indépendamment du problème de redondance de la Protein Data Bank, la fréquence d'une région étant normalisée par le nombre de molécules différentes portant la région et non par le nombre de structures. Le site de liaison à l'ATP est par exemple sur-représenté, alors que le site de dimérisation des récepteurs nucléaires est sous-représenté. Un des développements futurs consistera à mieux détailler la spécificité (fréquence) des régions, et de tenter de corréler cette spécificité à la localisation de sites fonctionnels.

Conclusions générales

Suite aux efforts importants investis dans les projets de Génomiques Structurales, et alors que de plus en plus de structures moléculaires sont disponibles, il apparaît nécessaire de développer des approches haut-débit permettant d'une part leurs analyses approfondies, et d'autre part, de prédire des fonctions biologiques. En effet, la disponibilité de plus de 66 000 structures biologiques ne permet plus aux scientifiques de connaître intimement chacune des structures d'un système d'intérêt. Il est pourtant nécessaire de connaître les différentes informations pouvant être extraites de ces structures, comme la localisation des sites de liaison et la description des partenaires associés, ou bien encore, des sites de phosphorylation, l'état de solvatation des résidus ou leur conservation évolutive, etc.

Dans un premier temps, nous avons développé l'approche de « Cartographie Moléculaire » permettant de procéder à l'analyse intégrative des données structurales en réunissant l'ensemble des informations pertinentes pouvant être extraites des structures homologues, et en les résumant sur une seule et même structure. La description des surfaces moléculaires implique par exemple la distinction des surfaces interagissantes (les atomes et résidus de surface pouvant interagir avec des molécules autres que du solvant ou des ions), et des surfaces non-interagissantes (ceux qui ne peuvent interagir qu'avec le solvant ou des ions). Leur distinction est loin d'être triviale, et par manque de données, les approches développées depuis plus de dix ans, considèrent la surface interagissante comme n'étant que les atomes et résidus vus interagissant dans une seule et unique structure, la surface non-interagissante étant alors constituée des autres atomes et résidus de surface. Nous avons montré que cette ancienne définition menait à une surface interagissante recouvrant de l'ordre de 20% des surfaces protéiques. Profitant de la redondance des structures de la Protein Data Bank qui mettent en scène une même molécule dans différentes conditions (avec différents partenaires), nous avons développé l'approche M-ORBIS consistant, à partir d'une molécule initiale, à récupérer l'ensemble de ses structures homologues, à analyser chacune de leurs interactions, et à inférer les sites de liaison et partenaires associés à cette molécule initiale. Cette inférence des fonctions, par l'analyse intégrative des données structurales, permet de mieux caractériser les régions fonctionnelles d'une molécule, ainsi que sa dynamique conformationnelle.

Ainsi, la description par M-ORBIS de plus d'une centaine de protéines, a conduit à ré-évaluer le pourcentage de la surface interagissante à 44% au lieu de 20%. Le regroupement de structures homologues a également permis d'observer des fréquences de contact, soit avec d'autres macromolécules, soit avec des molécules d'eau. Cette notion de fréquence de contact des résidus interagissants est également importante puisqu'elle permet de

distinguer, pour un type d'interaction donnée, les résidus qui sont toujours vus interagissant, de ceux qui ne le sont que pour certains partenaires. Ainsi, la fréquence de contact devrait permettre dans un avenir proche de mieux décrire les résidus impliqués dans la stabilisation d'un type d'interaction (qui nécessite de remplir certaines contraintes physico-chimiques, comme pour les charges anioniques des acides nucléiques), et ceux impliqués dans la spécificité de reconnaissance d'un partenaire. Par ailleurs, la fréquence de contact d'un résidu avec des molécules d'eau a permis de montrer que si plus de 86% de la surface des protéines peut être hydratée de façon transitoire, seules 16% de ces surfaces peuvent l'être de façon spécifique (avec une fréquence de contact supérieure à 90%). Plus généralement, la Cartographie Moléculaire d'une protéine par M-ORBIS, permet d'identifier les résidus qui sont hydratés de façon spécifique et qui peuvent avoir un rôle fonctionnel, soit dans la stabilisation de la protéine, soit pour médier des interactions moléculaires.

De plus, l'analyse intégrative des données structurales par M-ORBIS a permis de mieux décrire le comportement des sites de liaison, et en particulier d'observer si un même site de liaison est capable de lier plusieurs types différents de molécules (nous les appellerons des sites de liaison « polyvalents »), ou s'il n'est capable de lier qu'un seul type de molécule (nous les appellerons des sites de liaison « spécifiques »). Ces deux grandes familles de sites de liaison diffèrent grandement, y compris dans leurs compositions en acides aminés. En particulier, les sites de liaison spécifiques à la formation d'homodimères sont près de deux fois plus hydrophobes (66%) que les sites de liaison polyvalents d'homodimères. Les sites de liaison polyvalents homodimère/hétérodimère et homodimère/solvant sont quant à eux plus polaires et chargés, ressemblant ainsi davantage à la surface non-interagissante des protéines. Les notions de sites de liaison polyvalents et spécifiques devront toutefois être corrélées le plus tôt possible aux aspects transitoires et permanents des interactions, les données thermodynamiques manquant pour le moment. Notons cependant que pour qu'un site de liaison soit polyvalent, cela implique qu'il soit capable de se dissocier pour se réassocier avec d'autres partenaires moléculaires, ce qui implique des interactions transitoires. Dans le cas des sites de liaison spécifiques à la formation d'homodimères, une telle sur-exposition de résidus hydrophobes dans une région de la surface moléculaire, tend à penser qu'elle ne peut rester ainsi sans partenaire, allant par conséquent dans le sens d'interactions plus fortes. Le site de dimérisation des récepteurs nucléaires est un bon exemple de site de liaison polyvalent, permettant de former à la fois des homodimères et des hétérodimères, chacun d'eux engendrant des mécanismes différents de régulation de l'expression des gènes.

Dans un second temps, nous avons montré que pour l'analyse haut-débit des structures, la représentation géométrique des surfaces moléculaires par les formes alphas, présente de nombreux avantages par rapport aux représentations usuelles des surfaces : en premier lieu,

elle permet de définir une notion de voisinage non ambigüe des atomes en filtrant les contacts inter-atomiques sur des critères géométriques et physico-chimiques ; en second lieu, elle n'approxime pas ni ne change la position des atomes de la molécule, comme le font la plupart des représentations simplifiées ; en troisième lieu, elle simplifie d'un facteur 10 à 20 la représentation des structures moléculaires, ce qui facilite considérablement leurs analyses et comparaisons par des approches bioinformatiques.

Nous avons ainsi proposé une définition rigoureuse des régions contiguës de structures moléculaires, sur la base d'une extension depuis un point de la structure. La définition de régions est une étape nécessaire à l'analyse des structures et a par exemple servi à la prédiction des sites de liaison depuis plus de dix ans, en permettant de moyenner localement les états de propriétés discriminant la surface interagissante de la surface non-interagissante. En particulier, si l'on considère qu'il existe un « Code Structural » permettant de renseigner sur les fonctions d'une molécule dans un environnement cellulaire, alors il est nécessaire de pouvoir définir des régions contiguës de tailles et de formes variées, c'est-à-dire, des blocs indivisibles capables d'accomplir à eux seuls une fonction (lier un partenaire, catalyser une réaction, etc). L'information contenue dans une seule structure peut alors être comparée à ce « Code Structural » pour identifier rapidement et efficacement de telles fonctions. L'identification d'un tel « Code Structural » peut se faire notamment par le criblage haut-débit des régions moléculaires.

En effet, l'analyse intégrative des données structurales réalisée par M-ORBIS présente l'inconvénient de ne pouvoir traiter que les protéines pour lesquelles plusieurs structures homologues sont disponibles : pour inférer des sites de liaisons et des partenaires, M-ORBIS a besoin à la fois d'une similarité locale, mais aussi et surtout d'une homologie sur les structures complètes. Notre approche, permettant la génération de régions contiguës des structures moléculaires, ouvre la voie à la comparaison des structures sur la base de leurs régions locales. Dès lors, s'il est possible d'inférer des fonctions par ressemblance des régions locales (Code Structural), alors il devient possible d'identifier des fonctions à des structures moléculaires pour lesquelles aucune interaction ou fonction n'avaient été détaillées auparavant.

Dans un premier temps, nous avons montré que le criblage haut débit des régions moléculaires permettait effectivement de récupérer les régions attendues, notamment sur la famille des récepteurs nucléaires avec les sites de dimérisation et de liaison au cofacteur peptidique. Nous avons ensuite pu montrer que si l'on considère les protéines appartenant à une même famille, et donc présentant un même repliement, l'identité des séquences des protéines portant ces régions était corrélée à 0,77, à la ressemblance des régions. En revanche, le criblage d'une région moléculaire sur une base de données contenant toutes les

structures moléculaires a montré qu'il n'existait aucune corrélation (-0.06) entre l'identité de séquence d'une cible protéique, et sa ressemblance à un site de liaison donné (cas du *Dorzolamide*).

Validée sur des sites de liaison de ligands, très nombreux dans la Protein Data Bank, notre approche montre une spécificité de l'ordre de 70 à 80% pour identifier des sites de liaison fonctionnels sur des protéines ne partageant aucune identité de séquence, ni de structure secondaire ou tertiaire. Aussi, le criblage haut débit des régions moléculaires constitue une nouvelle approche à part entière pour la comparaison des structures moléculaires et comporte de nombreuses perspectives intéressantes. En particulier, partant du site de liaison d'une molécule (protéine, ADN, ARN, composé …), il devient possible d'évaluer l'ensemble des cibles cellulaires de cette molécule, dès lors que ces cibles sont représentées par des structures moléculaires. Le problème de la dynamique conformationnelle des molécules reste cependant très important et de nouveaux développements seront nécessaires pour en tenir davantage compte.

Le criblage de régions, plutôt que la comparaison de séquences ou de structures secondaires, a également permis de réaliser des « arbres fonctionnels », renseignant sur la similarité de régions fonctionnelles de molécule. Dans ce manuscrit, nous proposons deux arbres fonctionnels des récepteurs nucléaires ; le premier correspondant au site de dimérisation, le second correspondant au site de liaison au cofacteur. Dans une certaine mesure, les deux arbres fonctionnels sont corrélés à l'arbre phylogénétique qui renseigne sur l'évolution commune des protéines. Ainsi, le lien de parenté entre protéines peut également se voir à l'échelle plus petite de leurs régions structurales et fonctionnelles. Toutefois, les arbres fonctionnels détaillent davantage les capacités communes de régions, à accomplir une fonction donnée (lier un partenaire, catalyser une réaction, phosphoryler un résidu, etc), et il est possible d'observer quelques différences significatives qui séparent des récepteurs nucléaires faisant partie d'une même famille évolutive, et qui rapprochent des récepteurs faisant partie de différentes familles évolutives.

Le fait de pouvoir évaluer les cibles cellulaires de molécules, notamment de médicaments, permet de mieux comprendre d'une part leurs mécanismes d'action, et d'autre part, de mieux évaluer les problèmes d'efficacité et d'effets secondaires/toxiques. En effet, nous considérons dans cette thèse que l'efficacité d'un médicament est la résultante de son affinité pour sa cible souhaitée, moins les affinités pour ses autres cibles détectées. Des développements importants liés à la génomique fonctionnelle seront toutefois nécessaires pour permettre d'évaluer la fréquence des molécules dans un environnement cellulaire donné. Cette fréquence des molécules permettra alors de pondérer la représentativité des régions moléculaires capables de lier ces composés. De la même façon,

nous considérons que l'apparition des effets secondaires/toxiques d'un médicament croît (en moyenne) avec le nombre de ses cibles cellulaires, en risquant de provoquer des interférences avec des systèmes non souhaités. De plus, le fait d'identifier les cibles cellulaires sur la base de structures moléculaires, permet d'identifier la localisation de ces sites de liaison et de vérifier notamment s'ils recouvrent des interfaces biologiques connues.

Ainsi, le criblage des régions moléculaires apporte une connaissance supplémentaire permettant de mieux comprendre les effets secondaires/toxiques en permettant de corréler la liaison d'un composé à une région donnée d'une cible, avec les éventuelles fonctions connues de cette région de la cible.

Outre la caractérisation fonctionnelle des structures moléculaires par le criblage de leurs régions moléculaires, ainsi que l'identification de cibles cellulaires de composés, cette approche pourra avoir des répercussions intéressantes sur l'identification de biomarqueurs d'effets secondaires et toxiques, voire de biomarqueurs « d'efficacité » qui renseigneraient sur le nombre moyen de cibles du composé dans un environnement cellulaire donné.

Publication connexe : SM2PH-db

INFORMATICS

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

SM2PH-db: An Interactive System for the Integrated Analysis of Phenotypic Consequences of Missense Mutations in Proteins Involved in Human Genetic Diseases

Anne Friedrich,[1†] Nicolas Garnier,[2†] Nicolas Gagnière,[1] Hoan Nguyen,[1] Laurent-Philippe Albou,[1] Valérie Biancalana,[3] Emmanuel Bettler,[2] Gilbert Deléage,[2] Odile Lecompte,[1] Jean Muller,[1] Dino Moras,[1] Jean-Louis Mandel,[3,4] Thierry Toursel,[5] Luc Moulinier,[1] and Olivier Poch[1*]

[1]*Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), Centre National de la Recherche Scientifique/Institut National de la Santé et de la Recherche Médicale/Université de Strasbourg, Illkirch, France;* [2]*Institut de Biologie et Chimie des Protéines (UMR 5086); Centre National de la Recherche Scientifique/Université de Lyon, Lyon, France;* [3]*Laboratoire de Diagnostic Génétique, CHRU, Faculté de Médecine et laboratoire de Génétique Médicale EA3949, Université de Strasbourg, Strasbourg, France;* [4]*Département de Neurobiologie et Génétique, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), Centre National de la Recherche Scientifique/Institut National de la Santé et de la Recherche Médicale/Université de Strasbourg, Illkirch, France;* [5]*Association Franç aise contre les Myopathies, Evry, France*

Communicated by Mauno Vihinen

Received 6 May 2009; accepted revised manuscript 2 November 2009.
Published online 17 November 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21155

ABSTRACT: Understanding how genetic alterations affect gene products at the molecular level represents a first step in the elucidation of the complex relationships between genotypic and phenotypic variations, and is thus a major challenge in the postgenomic era. Here, we present SM2PH-db (http://decrypthon.igbmc.fr/sm2ph), a new database designed to investigate structural and functional impacts of missense mutations and their phenotypic effects in the context of human genetic diseases. A wealth of up-to-date interconnected information is provided for each of the 2,249 disease-related entry proteins (August 2009), including data retrieved from biological databases and data generated from a Sequence–Structure–Evolution Inference in Systems-based approach, such as multiple alignments, three-dimensional structural models, and multidimensional (physicochemical, functional, structural, and evolutionary) characterizations of mutations. SM2PH-db provides a robust infrastructure associated with interactive analysis tools supporting in-depth study and interpretation of the molecular consequences of mutations, with the more long-term goal of elucidating the chain of events leading from a molecular defect to its pathology. The entire content of SM2PH-db is regularly and automatically updated thanks to a computational grid data federation facilities provided in the context of the Decrypthon program.
Hum Mutat 31:127–135, 2010. © 2009 Wiley-Liss, Inc.

KEY WORDS: SM2PH-db; human genetic disease; mutation impact; genotype–phenotype relationship; structural homology model

Introduction

The completion of the human genome has provided a large volume of data that represents the basis for the characterization of all human genes [International Human Genome Sequencing Consortium, 2004]. It has also paved the way to the systematic study of its variability [Ring et al., 2006; The International HapMap Consortium, 2003], which is related to the evolution of the human species, required for its constant development and adaptation, but also to the emergence of human diseases. Human variability is naturally expressed through, for example, genetic shuffling during meiosis, but also by all sorts of accidental DNA changes, ranging from the substitution of a single nucleic acid to major chromosomal rearrangements.

A major source of interindividual human variation results from the substitution of one residue by another one, called single nucleotide polymorphism (SNP). SNPs are highly abundant and distributed throughout the genome [Stranger et al., 2007], and in addition, SNPs are the mutations that are the most related to human diseases [Antonarakis et al., 2000]. SNPs can be linked to the emergence of or to the predisposition to disease and influence its severity, progression, as well as its drug sensitivity. Deleterious SNPs occur in both coding and noncoding regions. In noncoding regions, the variation mostly affects gene expression by disrupting functional sites at the transcriptional level (e.g., transcription factor binding sites) [Kim et al., 2008], or result in splicing defects

Additional Supporting Information may be found in the online version of this article.

[†]The first two authors contributed equally to this article.

Present address for Nicolas Garnier: Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India.

Present address for Jean Muller: Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.

*Correspondence to: Olivier Poch, Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), 67 404 Illkirch, France. E-mail: poch@igbmc.fr

© 2009 WILEY-LISS, INC.

[Krawczak et al., 2007]. In coding regions, and in particular protein-coding genes, deleterious SNPs are mainly non synonymous SNPs (nsSNPs), also called missense mutations, which result in the modification of the amino acid sequence of the encoded protein. nsSNPs have been linked to a wide variety of diseases, for example by affecting protein function, by reducing protein solubility or by destabilizing protein structure [Chasman and Adams, 2001]. All these perturbations can be considered as the primary molecular phenotype associated with the missense mutation, having a cascade of consequences and finally leading to the emergence of a genetic disease.

The elucidation of the complex relationships between genotypic and phenotypic variations is a major challenge in the postgenomic era. Indeed, this step is crucial to a better understanding of gene functions and networks, aimed at revealing the mechanism of diseases and the development of specific therapeutic solutions.

With the current amount of information available in various biological databases, including sequences, structures, functions, pathways, interactions, variations, etc. [Galperin and Cochrane, 2009], and the subsequent development of in silico analysis tools, it is now possible to better understand and/or predict the correlation between a missense mutation and its associated molecular phenotypes. However, to gain further insight into the mechanisms of diseases, these molecular phenotypes have to be linked to several levels of phenotypic consequences, from the cell to the organism, ideally on the basis of ontologies. Unfortunately, the access to genotypic data with precise phenotypic descriptions represents a bottleneck in the elucidation of these complex relationships. Briefly, two main classes of databases incorporate mutations and the description of their consequences [Horaitis and Cotton, 1999]: (1) central databases, such as OMIM [McKusick, 2007], HGMD [Stenson et al., 2008], and UniProtKB/Swissprot [Yip et al., 2008], that include mutations related to a wide range of genes with limited genotypic/phenotypic descriptions; (2) locus-specific databases (LSDBs), which are specific to a gene (or gene family) and typically contain much more precise genotypic and phenotypic descriptions. Around 700 LSDBs are listed on the Human Genome Variation Society Web site [Horaitis et al., 2007] as being currently available on the Web. However, the LSDBs have very different data formats, although some efforts are being undertaken toward their homogenization, in particular the development of generic tools to facilitate LSDB implementation, such as the Universal Mutation Database (UMD) [Beroud et al., 2005], the Leiden Open (source) Variation Database (LOVD) [Fokkema et al., 2005], and MUTbase [Riikonen and Vihinen, 1999].

In this context, significant efforts have been devoted to providing links between human SNPs and their molecular effects, and have led to the development of novel systems combining data storage and analysis tools, both of which are indispensable for the characterization of the consequences of SNPs [Tavtigian et al., 2008]. Systems such as coliSNP [Kono et al., 2008], LS-SNP [Karchin et al., 2005], MutDB [Singh et al., 2008], SAAPdb [Hurst et al., 2009], SNPs3D [Yue et al., 2006], and topoSNP [Stitziel et al., 2004] all present structural information related to the mutated protein, which has been shown to be essential [Chasman and Adams, 2001; Wang and Moult, 2001]. Among these, however, LS-SNP is the only one that provides computationally generated comparative protein structure models when no experimentally determined structure is available. Concerning the phenotypic effects associated to the SNPs, only MutDB provides a direct access to the observed phenotype, although this is limited to the disease name. SNPs3D and LS-SNP give access to pathogenicity prediction such as the SIFT score [Ng and Henikoff, 2003], as well as in-house developed tools, but the absence of phenotypic descriptions clearly represents an obstacle in the interpretation of the molecular consequences of a nsSNP and the related disease mechanisms. Furthermore, the postgenomic era is characterized by a torrent of biological information flooding the databases: a major requirement for any newly developed database is hence to be based on a solid computing infrastructure and to ensure its regular and automated update [Philippi and Kohler, 2006]. ColiSNP and SAAPdb are the only systems that are regularly updated and consequently respond to these requirements.

To address these limitations, we have developed SM2PH-db, which stands for "from Structural Mutation to Pathology Phenotypes in Human database." SM2PH-db provides access to a wide range of up-to-date interconnected information related to the relationship between genotype and phenotype, with particular attention being focused on structural information via 3D structure or comparative models. A detailed multidimensional (physicochemical, functional, structural, and evolutionary) characterization of the mutations based on a Sequence–Structure–Evolution Inference in Systems (SStEISy), as well as an interactive analysis platform, ensure the investigation of structural and functional impacts of missense mutations toward a better understanding of their potential molecular effects, with regard to their individual phenotypic effects.

SM2PH-db deals mainly with proteins involved in human monogenic diseases, also known as Mendelian diseases. These disorders are particularly studied, because the emergence of a pathological phenotype is linked to the alteration of a single gene. Phenotypic diversity in monogenic diseases primarily reflects mutation heterogeneity, although the action of gene modifiers and epigenetic and environmental factors must also be considered [Jirtle and Skinner, 2007; Weatherall, 1998]. Although these so-called genetic, epigenetic, and environmental backgrounds are not integrated in our system, the accurate multidimensional characterization of missense mutations will provide an initial insight into the molecular basis of monogenic diseases that could open the way in the future, to a significant improvement in our understanding of the molecular and genetic basis of common, complex diseases [Antonarakis and Beckmann, 2006].

The establishment of a suitable infrastructure for SM2PH-db required the development of automated procedures to allow the integration of information from heterogeneous sources as well as regular updates. This prompted us to perform our developments in the context of the Decrypthon program (http://www.decrypthon.fr/english/), which provides access to a computational grid as well as data storage and federation facilities, via the Decrypthon Data Center [Nguyen et al., 2008].

Currently (August 2009), SM2PH-db holds a total of 2,249 human proteins related to genetic diseases and is publicly accessible online at http://decrypthon.igbmc.fr/sm2ph.

SM2PH-db Content

The information associated with each disease-related protein can be classified into two main categories: data retrieved from existing databases and information produced for SStEISy purposes.

Data retrieved from Existing Databases

The retrieved information is organized into two classes:

1. General information such as the gene and protein names and synonyms, as well as the known splicing variants and

sequences, are extracted from the UniProtKB database [UniProt Consortium, 2008]. Cytogenetic band information is downloaded from the Genecards database [Safran et al., 2003], gene ontology annotations are extracted from the GO database [Harris et al., 2004] and the associated disease names are obtained from the Online Mendelian Inheritance in Man (OMIM) database [McKusick, 2007].

2. Missense mutations related to the disease proteins linked to their associated phenotypes are extracted from the index of "Human polymorphisms and disease mutations" from UniProtKB/Swissprot (http://www.uniprot.org/docs/humsavar.txt) and from two LSDBs (UMD-MTM1 [Biancalana et al., in preparation] and the Tissue Nonspecific Alkaline Phosphatase Gene Mutations Database, http://www.sesep.uvsq.fr/database_hypo/Mutation.html). Each disease-causing missense mutation is linked to the name of its related disease, supplemented by a severity description when available. Nonpathogenic missense mutations are associated with the "polymorphism" term, in accordance with the nomenclature used in the UniProtKB database.

At the time of this writing (August 2009), 27,884 missense mutations are recorded in SM2PH-db, among which 20,252 are considered as disease-causing and 7,632 as nonpathogenic.

Information Produced through SStEISy Approaches

To establish an appropriate SStEISy workbench, data related to sequence, structure, and evolution are processed for each disease-related protein and missense mutations are then characterized in this context. The information resulting from the processing of these data is listed below:

1. Evolutionary background information network: structural and functional annotations are linked to each disease-related protein thanks to MACSIMS (Multiple Alignment of Complete Sequences Information Management System) [Thompson et al., 2006], an information management system that combines knowledge-based methods with complementary ab initio sequence-based predictions. MACSIMS takes advantage of the multiple alignment ontology MAO [Thompson et al., 2005] to integrate several types of data in the framework of Multiple Alignments of Complete Sequences (MACS). Indeed, subfamily characterization based on MACS allows to highlight some discriminative aspects of sequence information, such as distinct conservation/variability patterns or domain organization between different phylogenetic levels [Lecompte et al., 2001]. For each disease-related protein, two MACS are computed with a modified version of the PipeAlign suite of programs [Plewniak et al., 2003], which integrates several steps ranging from homolog searches in protein sequence and structure databases to the definition of the hierarchical relationships between subfamilies. The first MACS is composed of the closest eukaryotic sequences and is used to identify evolutionary constraints at particular sequence positions that are characteristic of the protein family or subfamily. The second MACS is constructed with a sampling strategy to significantly reduce the number of aligned sequences, while at the same time maintaining the potential structural and functional information in the alignment [Friedrich et al., 2007].
2. Structural information network: the availability of a 3D structure or model of the protein is essential to gain insight into the structural impact of a mutation. The best source of protein structural information is the PDB [Berman et al., 2000], which stores almost all the experimentally resolved crystallographic structures. However, only 574 proteins out of our 2,249 human disease-related proteins are currently represented in the PDB. To enhance the available experimental data, 3D models of the wild-type proteins are automatically constructed by homology, using Modeller [Eswar et al., 2008]. The models are built by inferring the structure of a protein (the target) from the structure of another putatively homologous protein (i.e., a sequence sharing at least 30% identity) solved by experimental methods (the template). The selection of a suitable template is based on BLAST similarity searches in the PDB: templates covering the full protein are preferred, but shorter domains can be modeled when a full template is lacking, in which case several 3D models may be associated with a single protein. The pairwise alignment of the target and template proteins is extracted from the sampled MACS and is used as input to Modeller. Five homology models are constructed and the one with the best normalized DOPE score [Eramian et al., 2008] is integrated in SM2PH-db. Currently, 1,551 structures and 3D models related to wild-type proteins are stored in the database, concerning 1,370 different proteins.
3. Missense mutation information network: missense mutations are characterized according to 31 parameters (Supp. Table S1), which can be classified into three main levels of information: (a) physicochemical changes induced by the amino acid substitution: modifications in size, charge, polarity, and hydrophobicity are independently described [Taylor, 1986]. A global score reflecting the degree of modification induced by the substitution is also assigned. This score corresponds to the residues interdistance (Supp. Fig. S1) based on a vector representation of the amino acids [French and Robson, 1983] and normalized on a scale from 0 to 100, with larger distances implying less conservative substitutions; (b) functional and structural features related to the substituted position: these features include MACSIMS annotations, descriptions of the 3D context (e.g., residue relative accessibility, in contact with an annotated site, etc.) and a conservation ranking. This ranking is calculated using an in-house developed method based on a three-step process: (i) two independent conservation scores are computed for each column of the sampled MACS, namely, the free energy score [Lockless and Ranganathan, 1999] and a score based on the two-dimensional vector representation of the amino acids (Supp. Fig. S2); (ii) these scores are classified with a Dirichlet mixture algorithm [Sjolander et al., 1996] to define "groups of conservation;" (iii) the groups are then ranked. This process is reiterated for each MACS subfamily. Thus, two global conservation classes (rank 1 and rank 2) and one subfamily conservation class per subfamily are finally defined; (c) structural modifications induced by the amino acid substitution, based on the mutant 3D models. These are automatically constructed with Modeller for missense mutations that can be mapped onto a wild-type 3D model sharing more than 50% identity with its PDB template: at the time of this writing (August 2009), 9,435 3D mutant models are available in SM2PH-db (7,863 for disease-causing mutants and 1,572 for mutants considered as nonpathogenic). The change in protein relative stability upon single-site mutation ($\Delta\Delta G$ value) [Casadio et al., 1995] is predicted with I-Mutant2.0 [Capriotti et al., 2005]: a positive $\Delta\Delta G$ value implies a protein stability increase, whereas a negative $\Delta\Delta G$ value suggests a destabilizing mutation. The wild-type residue contacts, computed with the CSU software [Sobolev et al., 1999], are

compared to the mutated residue ones in the 3D model and any change induced by the substitution is stored.

SM2PH-db Data Generation and Update

A fully automated workflow has been developed for the generation and regular update of the entire database contents (Fig. 1), which can be divided into five main processes:

1. Protein entry list update, based on the OMIM database. The list of gene entries with disease-causing mutations is obtained as described in Amberger et al. [2009]. Based on this list, a file containing all the selected human entry sequences in FASTA format is created.
2. Mining of heterogeneous databases (OMIM, UniProtKB, Genecards, GO, several LSDBs) through the Decrypthon Data Center, based on the protein entry list. The entire set of retrieved data associated with each disease-related protein is constituted during this process.
3. Construction of the SStEISy workbench. First, structural templates are determined based on similarity searches in the PDB and two MACS are constructed and annotated, including the template sequences. Second, a new wild-type 3D model is generated if the structural template differs from the one used in the previous version of SM2PH-db or if a new 3D template is available.
4. Multidimensional characterization of the mutants. This process initiates with the generation of mutant 3D models not available in the previous version of SM2PH-db (e.g., for novel proteins, newly generated wild-type 3D models, or new mutations). Physicochemical changes and structural modifications induced by the substitution as well as functional and structural features related to the mutated position are then listed.
5. Finally, the entire content of the SM2PH-db is upgraded, by integrating all this information. Particular attention has been paid to the automation and optimization of this workflow. We have optimized communication and synchronization between the processes by analyzing the data dependencies. To speed up the entire workflow, processes 2 and 3 are launched in parallel. Moreover, as SM2PH-db includes about 2,250 entry proteins, the update of all the alignments and their annotation represented a potential bottleneck in terms of computing time and memory. To overcome these limitations, we have implemented PipeAlign and MACSIMS on the Decrypthon computation grid and the required databases have been integrated in the Decrypthon Data Center. The construction and annotation of all SM2PH-db alignments takes around two days using the Decrypthon infrastructure, which is constituted of 120 computational nodes, whereas the same tasks took about 1 month to run on our internal server (four processor SUN Enterprise V40z). Currently, the complete SM2PH-db update procedure is launched every 2 months, guaranteeing that the user is working with up-to-date information.

Software Implementation

SM2PH-db is implemented in a relational database management system (PostgreSQL), has a schema with 26 tables and runs on a Linux server. It takes advantage of a molecular model database management system called Modeome3D [Garnier et al., in preparation]. The generation of the Web interface is programmed in Python and uses the AJAX (Asynchronous JavaScript and XML) methodology for dynamic updating of the pages.

SM2PH-db Web Interface

Database Search

SM2PH-db can be queried using textual search forms (Fig. 2A), based on a combination of keywords (e.g., protein name, disease name) or via a BLAST search [Altschut et al., 1997] of the database entries.

The search results consist of a summary of the data related to each protein that matches the required criteria (Fig. 2B). *Protein*

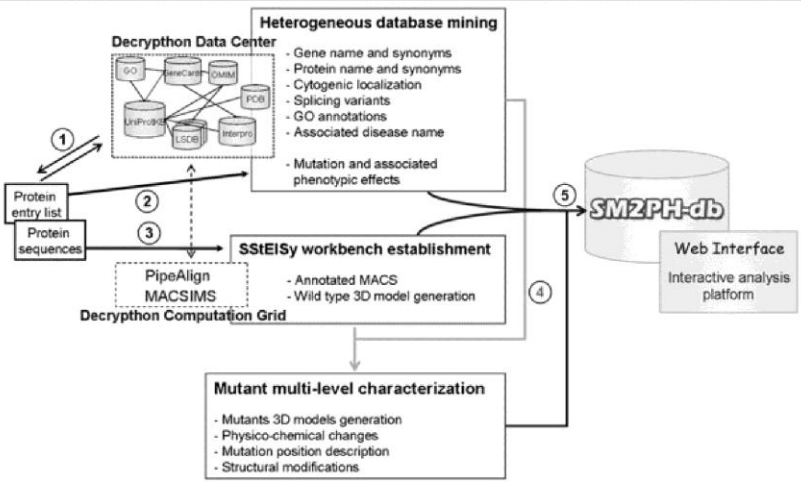


Figure 1. Schematic representation of SM2PH-db automated workflow for data generation and integration: (1) protein entry list update; (2) heterogeneous data mining via the Decrypthon Data Center; (3) SStEISy workbench construction; (4) mutant multidimensional characterization; (5) SM2PH-db data integration.

details pages, that assemble all the retrieved data, can be accessed by clicking on the protein identifiers. When a 3D model is available, a diagram that schematizes the modeled region(s) of the sequence is presented and an interactive analysis interface can be accessed that allows the visualization and manipulation of the processed information.

Figure 2. SM2PH-db search interface. **A:** The textual search form allows querying of the database with a combination of keywords. A list of keywords matching the first letters entered by the user is automatically proposed to facilitate the search. **B:** The result page displays a table containing the protein identifier, its name, the associated disease name and information related to the 3D model if one has been constructed. Links to the entry protein details and to the interactive analysis interface are provided. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Protein Details Page

The *protein details* page can be considered to be a protein "identity card." It displays data retrieved from existing databases; links to external databases are provided and a visual summary of the processed information. The retrieved data are divided into three main sections: *General Information*, *Crossreferences*, and *Mutations* (Fig. 3A). The latter includes a list of missense mutations associated with a short description of their phenotypic consequences.

The processed information summary mainly concerns the functional domain annotations assigned to the protein by MACSIMS, details of the 3D model if one has been built and its secondary structure composition. The schematic view reveals all these features in a simple, linear representation (Fig. 3B).

Interactive Graphical Interface

The SM2PH-db interactive interface (Fig. 4A) is based on the MAGOS Web server [Garnier et al., 2006]. It has been upgraded to meet the SStEISy requirements and enhanced to allow integrative missense mutation analyses. This graphical interface is divided into three frames, each related to a produced information network:

1. Frame 1: the right-hand part of the interface shows the annotated MACS. The sequence of the human disease-related protein is always the first in the alignment and the modeled region is underlined. By default, one sequence of each subfamily of the MACS is displayed; subfamily sequences can be expanded with the "Uncollapse all" button. Functional and structural annotations such as active sites, PFAM domains, residues conservation during evolution, transmembrane regions, etc., are available through the *Features* menu: these are mapped on the MACS when selected.

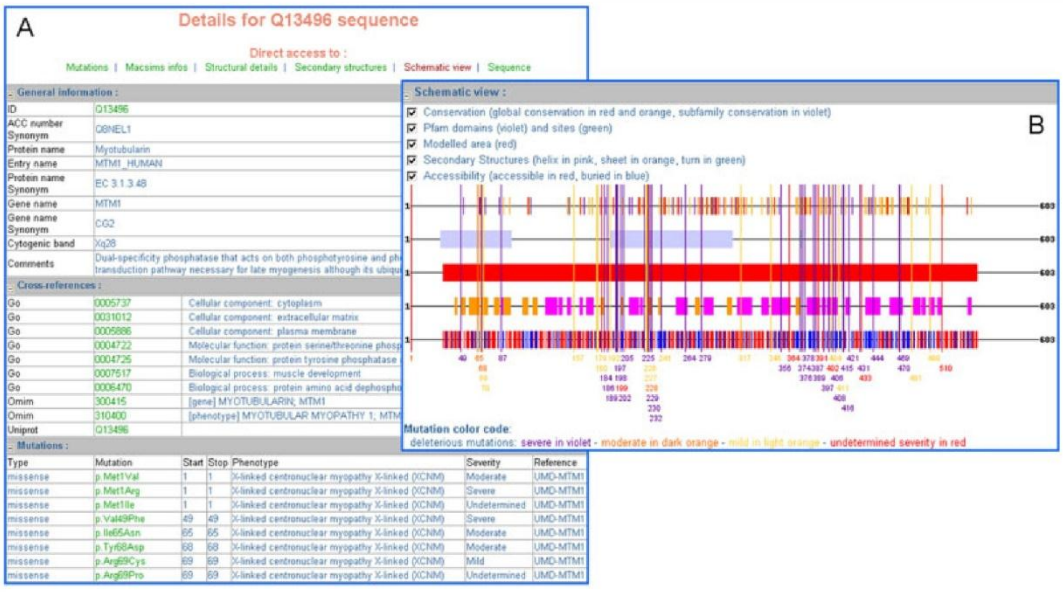


Figure 3. Screenshots of the Myotubularin details page. **A:** The retrieved data are displayed, divided into several sections. **B:** The schematic view associated with Myotubularin summarizes, in a linear representation, the main protein functional and structural features as well as the mutant positions. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

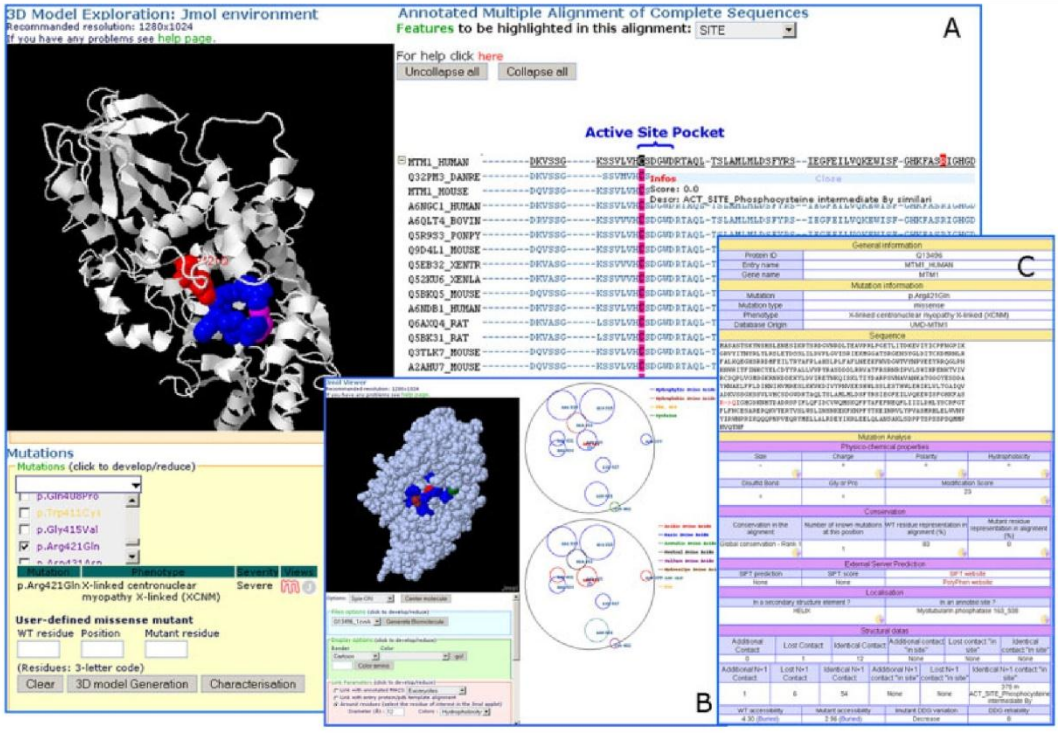


Figure 4. SM2PH-db interactive graphical interface for Myotubularin. **A:** The dynamic Web page is divided into three interconnected frames: the right-hand frame gives access to the annotated MACS interconnected with the 3D model/structure displayed in the upper left frame. All the features mined and predicted by MACSIMS can be displayed both on the query sequence in the context of the MACS and on its structural representation. The lower left frame shows the missense mutations that can be positioned and analysed within the structural and functional context of the protein. **B:** Residues surrounding a position of interest can be highlighted with the *Around residues* option. The diameter of investigation can be modified by the user. **C:** The p.Arg421Gln mutation characterization page, which provides descriptions concerning the modifications induced by the substitution as well as information related to the conservation of the mutated residue, its position relative to functional features and within the 3D model, etc. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

2. Frame 2: the upper part of the left-hand side of the interface displays the 3D model in a Jmol environment (http://jmol.sourceforge.net/) and allows its in-depth exploration. The *Display options* section is dedicated to a basic manipulation of the 3D model by the intermediary of predefined rendering types (e.g., cartoon, spacefill, wireframe) and residue coloring (e.g., according to polarity, accessibility), whereas the *Console* section allows a high-level 3D model manipulation using the Jmol/Rasmol command line. The visualization of the residues positioned in a structural environment close to a residue of interest (i.e., within a distance defined by the user) can be performed with the *Around residues* option found in the *Link Parameters* section (Fig. 4B).
3. Frame 3: the *Mutations* section, in the lower part of the left-hand side of the page, is devoted to the missense mutation informational network. The missense mutations mapped on the 3D model can be accessed via the top menu of this section. A table is shown under mutation selection, in which the associated phenotypes and two links are furnished. The first link provides access to the mutation characterization page (Fig. 4C), that includes the 31 parameters describing mutations (i.e., physicochemical changes induced by the substitution, information related to the substituted position such as conservation, functional, and structural features). The second link allows the replacement of the wild-type 3D model by the mutant model in the Jmol window.

Although these frames represent suitable analysis tools on their own, the main advantage of SM2PH-db graphical interface relies on their interconnection, which allows high-level integrative missense mutation analysis to be performed. First, the disease-related protein sequence in the MACS is connected to its 3D model: the annotations assigned to the underlined part of the sequence (i.e., the modeled part) are simultaneously displayed on the 3D model using the same color code. The selection of a residue in the structure will also highlight this residue within the sequence and vice versa. Furthermore, the mutation frame is connected to the other two frames: the selection of a missense mutation leads to the simultaneously mapping of the wild-type residue within the wild-type 3D model and in the MACS. This allows the study of the mutated position in a SStEISy environment.

Thanks to this graphical interface, in addition to providing insights into the disease mechanisms, SM2PH-db represents an ideal workbench for an in-depth analysis of novel user's mutations through an interactive approach. Indeed, the user can study the substituted position in terms of structural/functional features, ask

for the construction of a 3D model for his mutant of interest, and explore this model through the Jmol interface (*Generate* button in the *Mutations* section). Moreover, to support specific user investigations, the complete mutation characterization page, summarizing the main information related to this substitution and its position can be generated "on the fly."

SM2PH-db Statistics Page

General statistics related to the mutations stored in SM2PH-db can be accessed via the *Statistics* link, in the main menu. These statistics concern 10 parameters (out of the 31 computed) that characterize either the mutation substitutions or their substituted positions. For each of these parameters, two graphs representing the distribution of the associated values are provided: one for disease-causing mutations, the other for nonpathogenic mutations.

Each graph represents one single parameter and cannot be used in an independent manner to define a threshold to discriminate between disease-causing and nonpathogenic mutations. However, coupled with human expertise, these statistics could facilitate the interpretation of the molecular consequences of a given mutation. Therefore, these graphs can be viewed directly in a pop-up window from the mutation characterization pages, by clicking on the statistics icon located in the lower right corner of the concerned parameter cell.

Discussion

SM2PH-db has been organized to give the user an easy access to a wealth of information relevant to the study of genotype/phenotype relationships in the context of human genetic diseases. Our database regroups up-to-date heterogeneous interconnected information, ranging from sequence to structure and including evolutionary and functional features, thus providing a high-level SStEISy workbench. This workbench, in combination with the interactive analysis platform, allows the investigation of known missense mutation molecular impacts with regard to their phenotypic effects as well as the in-depth exploration of novel missense mutations to infer their potential molecular and phenotypic effects.

At time of writing (August 2009), a total of 27,884 missense mutations are recorded in the database, among which 20,252 (73%) are disease-causing and 7,632 (27%) are considered as nonpathogenic. 9,435 3D mutant models have been created and are available via the *Interactive graphical interface*, of which 7,863 are disease-causing and 1,572 nonpathogenic.

SM2PH-db: Insight into Mutation Effects

The molecular consequences of missense mutations are related to the functional and structural contexts of the affected position, as well as to the physicochemical characteristics of the substitution [Saunders and Baker, 2002; Terp et al., 2002]. All these types of information are represented in SM2PH-db for the stored missense mutations.

The integration capabilities of SM2PH-db can be illustrated by the analysis of the molecular consequences of a selected missense mutation. Here, we consider the p.Arg421Gln missense mutation that affects Myotubularin, which is associated with a severe deleterious phenotype.

Myotubularin can be searched by querying the *Protein name* field in the textual search form. This search indicates that four proteins whose names contain the term myotubularin are stored in the database, including three myotubularin-related proteins (Fig. 2). It should be noted that a 3D model of Myotubularin has been constructed with a template that shares 69% of identity, suggesting that this model is of good quality (Fig. 2B).

Before any further analysis, the protein details page should be visualized (Fig. 3). Here, the domain and active site localizations are of particular interest when trying to interpret mutation consequences. These can be viewed in the *Macsims infos* section, as well as in the linear schematic view of this entry, where structural and evolutionary information are also provided.

The interactive graphical interface can then be accessed via the *View in Jmol* link. The Arginine in position 421 can be mapped onto the 3D structure under selection in the *Mutations* section (Fig. 4A). A quick visual inspection shows that this residue is part of an alpha-helix and seems to be buried. The mutation characterization page (Fig. 4C) can then be accessed via the link provided in the table in the *Mutations* section. The physicochemical modifications linked to this substitution are not drastic: the modified score is 23 and the main change to be noted is the decrease of the residue size. The substituted position is located in the *Myotubularin phosphatase* domain and is well conserved during evolution: arginine represents more than 95% of this column in the alignment and has been classified as a rank 1 conserved position. In this example, an important feature can be observed, concerning the wild-type and mutated residue contacts: a residue that is in contact with our substituted residue is in direct contact with the Myotubularin active site, at position 375. This contact, although predicted as maintained, can obviously be modified by the substitution by a smaller residue. As a consequence, one might hypothesize that the p.Arg421Gln substitution destabilizes the active site pocket, which could explain the severe phenotype associated.

In conclusion, this example shows that the SM2PH-db infrastructure is suitable for in-depth investigations of mutations and can support the formulation of hypotheses related to the molecular consequences of known or newly discovered mutations.

Phenotypic Information

The availability of phenotypic information is central to SM2PH-db, because this information should help the scientist in the understanding of the disease pathogenesis. The retrieval of missense mutations from the UniProtKB/Swissprot database ensures the integration of phenotypic data for almost all the disease-related proteins, but our goal is to provide access to more precise information. Our system has consequently been organized to store information from LSDBs, excluding any patient data, although their integration has to be envisaged in a step-by-step manner because of the highly variable data formats. In the future, this task should become simpler, thanks to current standardization efforts in the context of the Gen2Phen European initiative (http://www.gen2phen.org) or the Human Variome Project [Ring et al., 2006], devised to collect and curate all genetic variation, its phenotypes and associated diseases, that will probably speed up the homogenization process. However, apart from the format aspects, another limiting factor for the integration of LSDB data is the reluctance of some clinicians to provide access to allelic variations and related information until after publication. To address this problem, we provide restricted access to private data on demand, based on the use of personal logins and passwords.

The Decrypthon Data Center is already capable of managing the UMD and LOVD formats and SM2PH-db stores data mined for example, from the UMD–MTM1 database. This database is dedicated to the Myotubularin protein (UniProt/Swiss-Prot: Q13496), involved in Myotubular Myopathy (MIM# 310400), and includes data related to 68 different missense variants, of which 61 have a corresponding three-level degree of severity.

We are currently contacting LSDB curators, seeking their consent to access and possibly diffuse their nonconfidential data, to provide users with precise genotypic/phenotypic information for a large number of disease-related proteins.

Conclusions and Perspectives

SM2PH-db represents an initial effort toward an effective and automated integration of mutation and phenotypic data involved in human genetic diseases. The sustainability of the database is guaranteed by the robust Decrypthon infrastructure on which it is based. Moreover, this latter also ensures that the biological data in SM2PH-db is always up to date.

In its current state, SM2PH-db and the numerous features offered by this novel system facilitates efficient study and interpretation of the molecular consequences of a given mutation.

In the future, a tool dedicated to the automated discrimination of disease-causing from nonpathogenic mutations will be integrated in our server. These predictions will be presented as propositions and complemented with all individual characterizing parameters, to give the user access to all the information required to judge the prediction accuracy. This complete system will hopefully contribute to the elucidation of the chain of events leading from a molecular defect to the pathology.

To achieve this, we intend to further enhance the available data by including, not only more detailed genotypic and phenotypic information, but also interactomic data such as functional and physical interactions mined from the STRING [Jensen et al., 2009] and IntAct [Kerrien et al., 2007] databases, as well as structural surface topology descriptions and interacting interface predictions [Albou et al., 2009]. We also plan to create *variant description* pages "on the fly" for missense variants of interest to the user. This will allow us to provide a simple access to customized information, which will aid the interpretation of the potential molecular effects of the mutation on the user's protein.

Acknowledgments

The authors are grateful to Raymond Ripp for his assistance during this work and Julie Thompson for a critical reading of the manuscript. This work was funded by the French Association against Myopathy via the Decrypthon program (AFM12727-13836), EVI-GENORET LSHG-CT-2005-512036, EvolHHuPro ANR-07-BLAN-0054-02, as well as institute funds from the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, and the Université de Strasbourg.

References

Albou LP, Schwarz B, Poch O, Wurtz JM, Moras D. 2009. Defining and characterizing protein surface using alpha shapes. Proteins 76:1–12.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37:D793–D796.

Antonarakis SE, Beckmann JS. 2006. Mendelian disorders deserve more attention. Nat Rev Genet 7:277–282.

Antonarakis SE, Krawczak M, Cooper DN. 2000. Disease-causing mutations in the human genome. Eur J Pediatr 159(Suppl 3):S173–S178.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28:235–242.

Beroud C, Hamroun D, Collod-Beroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum Mutat 26:184–191.

Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33:W306–W310.

Casadio R, Compiani M, Fariselli P, Vivarelli F. 1995. Predicting free energy contributions to the conformational stability of folded proteins from the residue sequence with radial basis function networks. Proc Int Conf Intell Syst Mol Biol 3:81–88.

Chasman D, Adams RM. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307:683–706.

Dayhoff MO, Eck RV, Park CM. 1972. A model of evolutionary change in proteins. In: Foundation NBR, editor. Atlas of protein sequence and structure. Washington, DC. p 89–99.

Eramian D, Eswar N, Shen MY, Sali A. 2008. How well can the accuracy of comparative protein structure models be predicted? Protein Sci 17:1881–1893.

Eswar N, Eramian D, Webb B, Shen MY, Sali A. 2008. Protein structure modeling with MODELLER. Methods Mol Biol 426:145–159.

Fokkema IF, den Dunnen JT, Taschner PE. 2005. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26:63–68.

French S, Robson B. 1983. What is a conservative substitution? J Mol Evolut 19:171–175.

Friedrich A, Ripp R, Garnier N, Bettler E, Deleage G, Poch O, Moulinier L. 2007. Blast sampling for structural and functional analyses. BMC Bioinformatics 8:62.

Galperin MY, Cochrane GR. 2009. Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. Nucleic Acids Res 37:D1–D4.

Garnier N, Friedrich A, Bolze R, Bettler E, Moulinier L, Geourjon C, Thompson JD, Deleage G, Poch O. 2006. MAGOS: multiple alignment and modelling server. Bioinformatics 22:2164–2165.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261.

Horaitis O, Cotton RG. 1999. 6th International HUGO Mutation Database Meeting, March 27, 1999, Brisbane, Australia. Hum Mutat 14:183–185.

Horaitis O, Talbot Jr CC, Phommarinh M, Phillips KM, Cotton RG. 2007. A database of locus-specific databases. Nat Genet 39:425.

Hurst JM, McMillan LE, Porter CT, Allen J, Fakorede A, Martin AC. 2009. The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum Mutat 30:616–624.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–D416.

Jirtle RL, Skinner MK. 2007. Environmental epigenomics and disease susceptibility. Nat Rev Genet 8:253–262.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. 2005. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21:2814–2820.

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. 2007. IntAct—open source resource for molecular interaction data. Nucleic Acids Res 35: D561–D565.

Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J. 2008. SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. BMC Bioinformatics 9(Suppl 1):S2.

Kono H, Yuasa T, Nishiue S, Yura K. 2008. coliSNP database server mapping nsSNPs on protein structures. Nucleic Acids Res 36:D409–D413.

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. 2007. Single base-pair substitutions in exon–intron junctions of human genes:

nature, distribution, and consequences for mRNA splicing. Hum Mutat 28:150–158.
Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. 2001. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene 270:17–30.
Lockless SW, Ranganathan R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286:295–299.
McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80:588–604.
Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.
Nguyen H, Friedrich A, Berthommier G, Poidevin L, Moulinier L, Ripp R, Poch O. 2008. Introduction du nouveau Centre de Données Biomedicales Décrypthon. Tregastel: CORIA.
Philippi S, Kohler J. 2006. Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet 7:482–488.
Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry JC, Thompson JD, Wicker N, Poch O. 2003. PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Res 31:3829–3832.
Riikonen P, Vihinen M. 1999. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15:852–859.
Ring HZ, Kwok PY, Cotton RG. 2006. Human Variome Project: an international collaboration to catalogue human genetic variation. Pharmacogenomics 7:969–972.
Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E, Adato A, Peter I, Khen M, Atarot T, Groner Y, Lancet D. 2003. Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. Nucleic Acids Res 31:142–146.
Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322:891–901.
Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, Heiland R, Mooney SD. 2008. MutDB: update on development of tools for the biochemical analysis of genetic variation. Nucleic Acids Res 36:D815–D819.
Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12:327–345.
Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. 1999. Automated analysis of interatomic contacts in proteins. Bioinformatics 15:327–332.
Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. 2008. Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet 45:124–126.
Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. 2004. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32:D520–D522.
Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853.
Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. 2008. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29:1327–1336.
Taylor WR. 1986. The classification of amino acid conservation. J Theor Biol 119:205–218.
Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M. 2002. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. Hum Mutat 20:98–109.
The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.
Thompson JD, Holbrook SR, Katoh K, Koehl P, Moras D, Westhof E, Poch O. 2005. MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Res 33:4164–4171.
Thompson JD, Muller A, Waterhouse A, Procter J, Barton GJ, Plewniak F, Poch O. 2006. MACSIMS: multiple alignment of complete sequences information management system. BMC Bioinformatics 7:318.
UniProt Consortium. 2008. The universal protein resource (UniProt). Nucleic Acids Res 36:D190–D195.
Wang Z, Moult J. 2001. SNPs, protein structure, and disease. Hum Mutat 17: 263–270.
Weatherall DJ. 1998. The phenotypic diversity of monogenic disease: lessons from the thalassemias. Harvey Lect 94:1–20.
Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A. 2008. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat 29:361–366.
Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166.

Annexes

I. Annexe 1 : Liste des 832 structures de récepteurs nucléaires regroupées par famille

ANDR_HUMAN (43)

1e3g	A	+		h
1gs4	A	+		h
1t5z	A	+	*	h
1t63	A	+	*	h
1t65	A	+	*	h
1xj7	A	+	*	h
1xow	A	+	*	h
1xq3	A	+		h
1z95	A	+		h
2am9	A	+		h
2ama	A	+		h
2amb	A	+		h
2ao6	A	+	*	h
2ax6	A	+		h
2ax7	A	+		h
2ax8	A	+		h
2ax9	A	+		h
2axa	A	+		h
2hvc	A	+		h
2oz7	A	+		h
2pio	A	+		h
2pip	L	+		h
2piq	A	+		h
2pir	A	+		h
2pit	A	+		h
2piu	A	+		h
2piv	A	+		h
2piw	A	+		h
2pix	A	+		h
2pkl	A	+	*	h
2pnu	A	+		h
2q7i	A	+	*	h
2q7j	A	+	*	h
2q7k	A	+	*	h
2q7l	A	+	*	h
2z4j	A	+	*	h
3b5r	A	+		h
3b65	A	+		h
3b66	A	+		h
3b67	A	+		h
3b68	A	+		h
3l3x	A	+	*	h
3l3z	A	+	*	h

ANDR_PANTR(8)

1t73	A	+	*
1t74	A	+	*
1t76	A	+	*
1t79	A	+	*
1t7f	A	+	*
1t7m	A	+	*
1t7r	A	+	*
1t7t	A	+	

ANDR_RAT (6)

1i37	A	+
1i38	A	+
1xnn	A	+
2ihq	A	+
2nw4	A	+
3g0w	A	+

ERR1_HUMAN (5)

1xb7	A	+	*	h
2pjl	A	+		h
2pjl	B	+		h
3d24	A		*	h
3d24	C		*	h

ERR3_HUMAN (30)

1kv6	A		*	h
1kv6	B		*	h
1tfc	A		*	h
1tfc	B		*	h
2e2r	A	+		h
2ewp	A	+		h
2ewp	B	+		h
2ewp	C	+		h
2ewp	D	+		h
2ewp	E	+		h
2gp7	A			h
2gp7	B			h
2gp7	C			h
2gp7	D			h
2gpo	A		*	h
2gpp	A	+	*	h
2gpp	B	+	*	h
2gpu	A	+		h
2gpv	A	+	*	h
2gpv	B	+	*	h
2gpv	C	+	*	h
2gpv	D	+	*	h
2gpv	E	+	*	h
2gpv	F	+	*	h
2p7a	A	+		h
2p7g	A	+		h
2p7z	A	+		h
2zas	A	+		h
2zbs	A			h
2zkc	A	+		h

ERR3_MOUSE (8)

1s9p	A	+
1s9p	B	+
1s9p	C	+
1s9p	D	+
1s9q	A	+
1s9q	B	+
1vjb	A	+
1vjb	B	+

ESR1_HUMAN (111)

1a52	A	+		h
1a52	B	+		h
1ere	A	+		h
1ere	B	+		h
1ere	C	+		h
1ere	D	+		h
1ere	E	+		h
1ere	F	+		h
1err	A	+		h
1err	B	+		h
1g50	A	+		h
1g50	B	+		h
1g50	C	+		h
1gwq	A	+	*	h
1gwq	B	+	*	h
1gwr	A	+	*	h
1gwr	B	+	*	h
1l2i	A	+	*	h
1l2i	B	+	*	h
1pcg	A	+	*	h
1pcg	B	+	*	h
1qkt	A	+		h
1qku	A	+		h
1qku	B	+		h
1qku	C	+		h
1r5k	A	+		h
1r5k	B	+		h
1r5k	C	+		h
1sj0	A	+		h
1x7e	A	+	*	h
1x7e	B	+	*	h
1x7r	A	+	*	h
1xp1	A	+		h
1xp6	A	+		h
1xp9	A	+		h
1xpc	A	+		h
1xqc	A	+		h
1xqc	B	+		h
1xqc	C	+		h
1xqc	D	+		h
1yim	A	+		h
1yin	A	+		h
1zky	A	+	*	h
1zky	B	+	*	h
2ayr	A	+		h
2b1v	A	+	*	h
2b1v	B	+	*	h
2b1z	A	+	*	h
2b1z	B	+	*	h
2b23	A	+	*	h
2b23	B	+	*	h
2fai	A	+	*	h
2fai	B	+	*	h
2g44	A	+	*	h
2g44	B	+	*	h
2g5o	A	+	*	h
2g5o	B	+	*	h
2i0j	A	+		h
2i0j	B	+		h
2i0j	C	+		h
2i0j	D	+		h
2iog	A	+		h
2iok	A	+		h
2iok	B	+		h
2ocf	A	+		h
2ouz	A	+		h
2p15	A	+	*	h
2p15	B	+	*	h
2pog	A	+		h
2pog	B	+		h
2q6j	A	+	*	h
2q6j	B	+	*	h
2q70	A	+		h
2q70	B	+		h
2qa6	A	+	*	h
2qa6	B	+	*	h
2qa8	A	+	*	h
2qa8	B	+	*	h
2qab	A	+	*	h
2qab	B	+	*	h
2qe4	A	+		h
2qe4	B	+		h

Notes :

\+ : présence d'un ligand

* : présence d'un petit peptide

h : protéine humaine

ESR1_HUMAN (suite)

2qgt	A	+	*	h
2qgt	B	+	*	h
2qgw	A	+	*	h
2qgw	B	+	*	h
2qh6	A	+	*	h
2qh6	B	+	*	h
2qr9	A	+	*	h
2qr9	B	+	*	h
2qse	A	+	*	h
2qse	B	+	*	h
2qxm	A	+	*	h
2qxm	B	+	*	h
2qxs	A	+		h
2qxs	B	+		h
2qzo	A	+	*	h
2qzo	B	+	*	h
2r6w	A	+		h
2r6w	B	+		h
2r6y	A	+		h
2r6y	B	+		h
3dt3	A	+		h
3dt3	B	+		h
3erd	A	+	*	h
3erd	B	+	*	h
3ert	A	+		h
3hlv	A	+	*	h
3hlv	B	+	*	h
3hm1	A	+	*	h
3hm1	B	+	*	h

ESR2_HUMAN (39)

1l2j	A	+		h
1l2j	B	+		h
1nde	A	+		h
1qkm	A	+		h
1u3q	A	+		h
1u3q	B	+		h
1u3q	C	+		h
1u3q	D	+		h
1u3r	A	+	*	h
1u3r	B	+	*	h
1u3s	A	+	*	h
1u3s	B	+	*	h
1u9e	A	+	*	h
1u9e	B	+	*	h
1x76	A	+	*	h
1x76	B	+	*	h
1x78	A	+	*	h
1x78	B	+	*	h
1x7b	A	+	*	h
1x7b	B	+	*	h
1x7j	A	+	*	h
1x7j	B	+	*	h
1yy4	A	+	*	h
1yy4	B	+	*	h
1yye	A	+	*	h
1yye	B	+	*	h
1zaf	A	+	*	h
1zaf	B	+	*	h
2fsz	A	+		h
2fsz	B	+		h
2giu	A	+		h
2i0g	A	+		h
2i0g	B	+		h
2nv7	A	+	*	h
2nv7	B	+	*	h
2qtu	A	+		h
2qtu	B	+		h
2z4b	A	+		h
2z4b	B	+		h

ESR2_RAT(4)

1hj1	A	+		
1qkn	A	+		
2j7x	A	+	*	
2j7y	A	+	*	

GCR_HUMAN (19)

1m2z	A	+		h
1m2z	D	+		h
1nhz	A	+		h
1p93	A	+	*	h
1p93	B	+	*	h
1p93	C	+	*	h
1p93	D	+	*	h
3bqd	A	+	*	h
3cld	A	+	*	h
3cld	B	+	*	h
3e7c	A	+	*	h
3e7c	B	+	*	h
3h52	A	+	*	h
3h52	B	+	*	h
3h52	C	+	*	h
3h52	D	+	*	h
3k23	A	+	*	h
3k23	B	+	*	h
3k23	C	+	*	h

HNF4A_RAT(4)

1m7w	A	+		
1m7w	B	+		
1m7w	C	+		
1m7w	D	+		

MCR_HUMAN (25)

1y9r	A	+		h
1y9r	B	+		h
1ya3	A	+		h
1ya3	B	+		h
1ya3	C	+		h
2a3i	A	+	*	h
2aa2	A	+		h
2aa5	A	+		h
2aa5	B	+		h
2aa6	A	+		h
2aa6	B	+		h
2aa7	A	+		h
2aax	A	+		h
2aax	B	+		h
2ab2	A	+		h
2ab2	B	+		h
2abi	A	+		h
2abi	B	+		h
2abi	C	+		h
2oax	A	+		h
2oax	B	+		h
2oax	C	+		h
2oax	D	+		h
2oax	E	+		h
2oax	F	+		h

NR1D2_HUMAN (7)

2v0v	A			h
2v0v	B			h
2v0v	C			h
2v0v	D			h
2v7c	A			h
2v7c	B			h
3cqv	A	+		h

NR1H2_HUMAN(19)

1p8d	A	+	*	h
1p8d	B	+	*	h
1pq6	A	+		h
1pq6	B	+		h
1pq6	C	+		h
1pq6	D	+		h
1pq9	A	+		h
1pq9	B	+		h
1pq9	C	+		h
1pq9	D	+		h
1pqc	A	+		h
1pqc	B	+		h
1pqc	C	+		h
1pqc	D	+		h
3kfc	A	+		h
3kfc	B	+		h
3kfc	C	+		h
3kfc	D	+		h
3l0e	A	+	*	h

NR1H3_HUMAN (6)

1uhl	B	+	*	h
3ipq	A	+	*	h
3ips	A	+	*	h
3ips	B	+	*	h
3ipu	A	+	*	h
3ipu	B	+	*	h

NR1H3_MOUSE (6)

2acl	B	+		
2acl	D	+		
2acl	F	+		
2acl	H	+		
3fal	B	+		
3fal	D	+		

NR1H4_HUMAN (9)

1osh	A	+		h
3bej	A	+	*	h
3bej	B	+	*	h
3dct	A	+	*	h
3dcu	A	+	*	h
3fli	A	+		h
3gd2	A	+	*	h
3hc5	A	+	*	h
3hc6	A	+	*	h

NR1H4_RAT(4)

1osv	A	+	*	
1osv	B	+	*	
1ot7	A	+	*	
1ot7	B	+	*	

NR1I2_HUMAN(4)

3ctb	A			h
3ctb	B			h
3hvl	A	+		h
3hvl	B	+		h

NR1I3_HUMAN(4)

1xv9	B	+	*	h
1xv9	D	+	*	h
1xvp	B	+	*	h
1xvp	D	+	*	h

NR1I3_MOUSE (6)

1xls	E	+	*	
1xls	F	+	*	
1xls	G	+	*	
1xls	H	+	*	
1xnx	A	+		
1xnx	B	+		

NR4A1_HUMAN(4)

2qw4	A			h
2qw4	B			h
2qw4	C			h
2qw4	D			h

NR4A2_HUMAN (6)

1ovl	A	+		h
1ovl	B	+		h
1ovl	C	+		h
1ovl	D	+		h
1ovl	E	+		h
1ovl	F	+		h

NR5A2_HUMAN(4)

1yok	A	+	*	h
1yuc	A	+	*	h
1yuc	B	+	*	h
1zdu	A	+	*	h

NR5A2_MOUSE(5)

1pk5	A			
1pk5	B			
1zh7	A		*	
1zh7	B		*	
3f5c	A			

PPARA_HUMAN (14)

1i7g	A	+		h
2npa	A	+	*	h
2npa	C	+	*	h
2p54	A	+	*	h
2rew	A	+		h
2znn	A	+		h
3et1	A	+	*	h
3et1	B	+	*	h
3fei	A	+	*	h
3g8i	A	+	*	h
3kdt	A	+		h
3kdt	B	+		h
3kdu	A	+		h
3kdu	B	+		h

PPARD_HUMAN (13)

2q5g	A	+		h
2q5g	B	+		h
2znp	A	+		h
2znp	B	+		h
2znq	A	+		h
2znq	B	+		h
3d5f	A	+		h
3d5f	B	+		h
3dy6	A	+		h
3dy6	B	+		h
3et2	A	+		h
3et2	B	+		h
3gz9	A	+		h

PPARG_HUMAN (105)

1fm6	D	+	*	h
1fm6	X	+	*	h
1fm9	D	+	*	h
1k74	D	+	*	h
1zeo	A	+		h
1zeo	B	+		h
2ath	A	+		h
2ath	B	+		h
2f4b	A	+		h
2f4b	B	+		h
2fvj	A	+	*	h
2g0g	A	+		h
2g0g	B	+		h
2g0h	A	+		h
2g0h	B	+		h
2gtk	A	+	*	h
2hfp	A	+	*	h
2hwq	A	+		h
2hwq	B	+		h
2hwr	A	+		h
2hwr	B	+		h
2i4j	A	+		h
2i4j	B	+		h
2i4p	A	+		h
2i4p	B	+		h
2i4z	A	+		h
2i4z	B	+		h
2om9	A	+		h
2om9	B	+		h
2om9	C	+		h
2om9	D	+		h
2p4y	A	+		h
2p4y	B	+		h
2pob	A	+		h
2pob	B	+		h
2prg	A	+		h
2prg	B	+		h
2q59	A	+		h
2q59	B	+		h
2q5p	A	+		h
2q5p	B	+		h
2q5s	A	+		h
2q5s	B	+		h
2q61	A	+		h
2q61	B	+		h
2q6r	A	+		h
2q6r	B	+		h
2q6s	A	+		h
2q6s	B	+		h
2q8s	A	+		h
2q8s	B	+		h
2qmv	A			h
2zk0	A			h
2zk0	B			h

PPARG_HUMAN (suite)

2zk1	A	+		h
2zk1	B	+		h
2zk2	A	+		h
2zk2	B	+		h
2zk3	A	+		h
2zk3	B	+		h
2zk4	A	+		h
2zk4	B	+		h
2zk5	A	+		h
2zk5	B	+		h
2zk6	A	+		h
2zk6	B	+		h
2zno	A	+		h
2zno	B	+		h
2zvt	A	+		h
2zvt	B	+		h
3b3k	A	+		h
3b3k	B	+		h
3bc5	A	+		h
3cdp	A	+		h
3cdp	B	+		h
3cds	A	+		h
3cds	B	+		h
3cs8	A	+	*	h
3cwd	A	+	*	h
3cwd	B	+	*	h
3d6d	A	+		h
3d6d	B	+		h
3dzu	D	+	*	h
3dzy	D	+	*	h
3e00	D	+	*	h
3et0	A	+		h
3et0	B	+		h
3et3	A	+	*	h
3fej	A	+	*	h
3fur	A	+	*	h
3g9e	A	+	*	h
3gbk	A	+		h
3gbk	B	+		h
3h0a	D	+	*	h
3ho0	A	+		h
3ho0	B	+		h
3hod	A	+		h
3hod	B	+		h
3ia6	A	+		h
3ia6	B	+		h
3k8s	A	+		h
3k8s	B	+		h
3kmg	A	+	*	h
3kmg	D	+	*	h
3lmp	A	+	*	h

PPAR_HUMAN (8)

1k7l	A	+	*	h
1k7l	C	+	*	h
1k7l	E	+	*	h
1k7l	G	+	*	h
1kkq	A	+	*	h
1kkq	B	+	*	h
1kkq	C	+	*	h
1kkq	D	+	*	h

PPAS_HUMAN (14)

1gwx	A	+		h
1gwx	B	+		h
1y0s	A	+		h
1y0s	B	+		h
2awh	A	+		h
2awh	B	+		h
2b50	A	+		h
2b50	B	+		h
2baw	A	+		h
2baw	B	+		h
2gwx	A			h
2gwx	B			h
3gwx	A	+		h
3gwx	B	+		h

PPAT_HUMAN (16)

1i7i	A	+		h
1i7i	B	+		h
1knu	A	+		h
1knu	B	+		h
1nyx	A	+		h
1nyx	B	+		h
1prg	A			h
1prg	B			h
1rdt	D	+	*	h
1wm0	X	+	*	h
1zgy	A	+	*	h
3prg	A			h
4prg	A	+		h
4prg	B	+		h
4prg	C	+		h
4prg	D	+		h

PRGR_HUMAN (22)

1a28	A	+		h
1a28	B	+		h
1e3k	A	+		h
1e3k	B	+		h
1sqn	A	+		h
1sqn	B	+		h
1sr7	A	+		h
1sr7	B	+		h
1zuc	A	+		h
1zuc	B	+		h
2ovh	A	+	*	h
2ovm	A	+	*	h
3d90	A	+		h
3d90	B	+		h
3g8n	A	+		h
3g8n	B	+		h
3g8o	A	+		h
3g8o	B	+		h
3hq5	A	+		h
3hq5	B	+		h
3kba	A	+		h
3kba	B	+		h

PXR_HUMAN (9)

1ilg	A			h
1ilh	A	+		h
1m13	A	+		h
1nrl	A	+	*	h
1nrl	B	+	*	h
1skx	A	+		h
2o9i	A	+	*	h
2o9i	B	+	*	h
2qnv	A	+		h

RARA_HUMAN(4)

1dkf	B	+		h
3kmr	A	+	*	h
3kmz	A	+	*	h
3kmz	B	+	*	h

RARG1_HUMAN(4)

1fcx	A	+		h
1fcy	A	+		h
1fcz	A	+		h
1fd0	A	+		h

RARG2_HUMAN(4)

1exa	A	+		h
1exx	A	+		h
3lbd	A	+		h
4lbd	A	+		h

RXRA_HUMAN (51)

1fby	A	+		h
1fby	B	+		h
1fm6	A	+	*	h
1fm6	U	+	*	h
1fm9	A	+	*	h
1g1u	A			h
1g1u	B			h
1g1u	C			h
1g1u	D			h
1g5y	A	+		h
1g5y	B	+		h
1g5y	C	+		h
1g5y	D	+		h
1k74	A	+	*	h
1lbd	A			h
1mv9	A	+	*	h
1mvc	A	+	*	h
1mzn	A	+	*	h
1mzn	C	+	*	h
1mzn	E	+	*	h
1mzn	G	+	*	h
1rdt	A	+	*	h
1xls	A	+	*	h
1xls	B	+	*	h
1xls	C	+	*	h
1xls	D	+	*	h
1xv9	A	+	*	h
1xv9	C	+	*	h
1xvp	A	+	*	h
1xvp	C	+	*	h
2acl	A	+		h
2acl	C	+		h
2acl	E	+		h
2acl	G	+		h
2p1t	A	+	*	h
2p1u	A	+	*	h
2p1v	A	+	*	h
2zxz	A	+	*	h
2zy0	A	+	*	h
2zy0	C	+	*	h
3dzu	A	+	*	h
3dzy	A	+	*	h
3e00	A	+	*	h
3e94	A	+	*	h
3fal	A	+		h
3fal	C	+		h
3fc6	A	+		h
3fc6	C	+		h
3fug	A	+	*	h
3h0a	A	+	*	h
3kwy	A	+	*	h

RXRB_HUMAN (5)

1h9u	A	+		h
1h9u	B	+		h
1h9u	C	+		h
1h9u	D	+		h
1uhl	A	+	*	h

RXRG_HUMAN(4)

2g18	A			h
2g18	B			h
2g18	C			h
2g18	D			h

THB1_HUMAN (16)

1bsx	A	+	*	h
1bsx	B	+	*	h
1n46	A	+		h
1n46	B	+		h
1nax	A	+		h
1nq0	A	+		h
1nq1	A	+		h
1nq2	A	+		h
1nuo	A	+		h
1q4x	A	+		h
1r6g	A	+		h
1xzx	X	+		h
1y0x	X	+		h
2j4a	A	+		h
2pin	A	+		h
2pin	B	+		h

THB_HUMAN (5)

3d57	A	+	h
3d57	B	+	h
3gws	X	+	h
3imy	A	+	h
3jzc	A	+	h

USP_DROME (6)

1hg4	A	+
1hg4	B	+
1hg4	C	+
1hg4	D	+
1hg4	E	+

VDR_HUMAN (18)

1db1	A	+	h
1ie8	A	+	h
1ie9	A	+	h
1s0z	A	+	h
1s19	A	+	h
1txi	A	+	h
2ham	A	+	h
2har	A	+	h
2has	A	+	h
2hb7	A	+	h
2hb8	A	+	h
3a2i	A	+	h
3a2j	A	+	h
3a3z	X	+	h
3a40	X	+	h
3cs4	A	+	h
3cs6	A	+	h
3m7r	A	+	h

VDR_RAT (17)

1rjk	A	+	*
1rk3	A	+	*
1rkg	A	+	*
1rkh	A	+	*
2o4j	A	+	*
2o4r	A	+	*
2zfx	A	+	*
2zl9	A	+	*
2zla	A	+	*
2zlc	A	+	*
2zmh	A	+	*
2zmi	A	+	*
2zmj	A	+	*
2zxm	A	+	*
2zxn	A	+	*
3a2h	A	+	*
3afr	A	+	*

Divers (111)

1dkf	A	+		
1k4w	A	+	*	
1lv2	A	+		h
1g2n	A	+		
1n4h	A	+	*	
1n83	A	+		h
1nav	A	+		h
1nq7	A	+	*	
1pdu	A			
1pdu	B			
1pzl	A	+	*	h
1r1k	A	+		
1r1k	D	+		
1r20	A	+		
1r20	D	+		
1s0x	A	+		h
1uom	A	+		
1upv	A	+		
1upw	A	+		
1xap	A	+		h
1xdk	A	+	*	
1xdk	B	+	*	
1xdk	E	+	*	
1xdk	F	+	*	
1xiu	A	+	*	
1xiu	B	+	*	
1yje	A			
1ymt	A	+	*	
1yow	A	+	*	h
1yp0	A	+	*	
1z5x	E	+		
1z5x	U	+		
1zdt	A	+	*	h
1zdt	B	+	*	h
2bj4	A	+	*	
2bj4	B	+	*	
2h77	A	+		h
2h79	A	+		h
2hbh	A	+	*	
2hc4	A	+	*	
2hcd	A	+	*	
2iz2	A		*	
2j14	A	+		
2j14	B	+		
2jf9	A	+	*	
2jf9	B	+	*	
2jf9	C	+	*	
2jfa	A	+	*	
2jfa	B	+	*	
2jj3	A	+		
2jj3	B	+		
2lbd	A	+		h
2nxx	A	+		
2nxx	B	+		
2nxx	C	+		
2nxx	D	+		
2nxx	E	+		
2nxx	F	+		
2nxx	G	+		
2nxx	H	+		
2q1h	A	+		
2q1v	A	+		
2q60	A			
2q60	B			
2q60	C			
2q60	D			
2qpy	A	+	*	
2r40	A	+		
2r40	D	+		
2vsr	A	+		
2vsr	B	+		
2vst	A	+		
2vst	B	+		
2vv0	A	+		
2vv0	B	+		
2vv1	A	+		
2vv1	B	+		
2vv2	A	+		
2vv2	B	+		
2vv3	A	+		
2vv3	B	+		
2vv4	A	+		
2vv4	B	+		
2w8y	A	+		
2w8y	B	+		
3cjw	A			h
3dr1	A	+	*	

Divers (Suite)

3f5c	B			
3f5c	C			
3f7d	A	+	*	
3fc6	B	+		
3fc6	D	+		
3fs1	A	+	*	h
3fxv	A	+	*	h
3gn8	A	+	*	
3gn8	B	+	*	
3gyt	A	+	*	
3gyu	A	+	*	
3hzf	A	+		h
3ilz	A	+		h
3jzb	A	+		h
3kyt	A	+	*	h
3l0j	A	+	*	h
3l0l	A	+	*	
3l0l	B	+	*	
3l1b	A	+		h
3eyb	A			
3eyb	B			
3eyb	C			
3eyb	D			
3n00	A			h

II. Annexe 2: Quelques résultats du criblage de l'ATP sur toutes les structures moléculaires

Compound	UniprotAC	UniprotID	Structure	Score	Geometry	Charge	Aligned
CELL DIVISION PROTEIN KINASE 2	CDK2_HUMAN	P24941	3f5x:protein:C	92.4	94.0	81.5	1.0
PERIPHERAL PLASMA MEMBRANE PROTEIN CASK	CSKP_HUMAN	O14936	3mfu:protein:A	72.8	74.7	60.5	1.0
TYROSINE-PROTEIN KINASE LYN	LYN_MOUSE	P25911	2zv7:protein:A	72.6	73.4	67.6	1.0
G PROTEIN-COUPLED RECEPTOR KINASE 6	GRK6_HUMAN	P43250	2acx:protein:B	71.5	76.2	40.7	1.0
SR PROTEIN KINASE	SKY1_YEAST	Q03656	1q97:protein:A	71.5	73.9	55.8	1.0
FIBROBLAST GROWTH FACTOR RECEPTOR 2	FGFR2_HUMAN	P21802	2pwl:protein:A	70.6	74.8	43.2	1.0
DUAL SPECIFICITY TYROSINE-PHOSPHORYLATION-	DYRK2_HUMAN	Q92630	3kvw:protein:A	70.4	71.2	65.1	1.0
PROTO-ONCOGENE TYROSINE-PROTEIN KINASE LCK	LCK_HUMAN	P06239	3acj:protein:A	69.0	71.5	52.9	1.0
GLYCOGEN SYNTHASE KINASE-3 BETA	GSK3B_HUMAN	P49841	1q3d:protein:B	69.0	73.0	42.4	1.0
RHODOPSIN KINASE	RK_BOVIN	P28327	3c50:protein:B	68.8	72.8	42.4	0.96
SERINE/THREONINE KINASE 24 (STE20 HOMOLOG	Q6P0Y1_HUMAN	Q6P0Y1	3a7h:protein:A	68.0	70.5	51.6	1.0
SERINE/THREONINE-PROTEIN KINASE CHK1	CHK1_HUMAN	O14757	2hy0:protein:A	67.7	71.1	45.3	1.0
PROTO-ONCOGENE TYROSINE-PROTEIN KINASE ABL1	ABL1_HUMAN	P00519	2hz4:protein:A	67.6	74.8	20.5	1.0
PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC	SRC_CHICK	P00523	3g5d:protein:A	67.4	70.9	44.1	1.0
SERINE/THREONINE-PROTEIN KINASE 17B	ST17B_HUMAN	O94768	3lm0:protein:A	67.2	68.4	59.2	1.0
SERINE/THREONINE-PROTEIN KINASE HASPIN	HASP_HUMAN	Q8TF76	3f2n:protein:A	66.7	66.8	66.3	1.0
SERINE/THREONINE-PROTEIN KINASE PAK 1	PAK1_HUMAN	Q13153	3fxz:protein:A	66.2	69.7	43.2	1.0
ACTIVIN RECEPTOR TYPE-1	ACVR1_HUMAN	Q04771	3mtf:protein:A	65.7	69.4	41.5	0.98
DMPK PROTEIN	Unavailable	Unavailable	2vd5:protein:B	65.6	69.4	40.3	1.0
CALCIUM/CALMODULIN DEPENDENT PROTEIN KINASE II	Q9U6Q0_CAEEL	Q9U6Q0	3kk9:protein:A	65.3	67.6	50.0	1.0
CASEIN KINASE II ALPHA CHAIN	CSK21_HUMAN	P68400	1jwh:protein:A	65.3	65.3	65.1	0.98
INTERLEUKIN-1 RECEPTOR-ASSOCIATED KINASE 4	IRAK4_HUMAN	Q9NWZ3	2nry:protein:D	65.2	68.2	45.7	1.0
TYROSINE-PROTEIN KINASE CSK	CSK_HUMAN	P41240	3d7u:protein:C	64.8	67.7	46.2	1.0
ACTIVATED CDC42 KINASE 1	ACK1_HUMAN	Q07912	1u54:protein:B	64.4	67.7	43.2	0.98
DUAL SPECIFICITY PROTEIN KINASE CLK3	CLK3_HUMAN	P49761	2eu9:protein:A	64.1	64.4	62.1	1.0
DEATH-ASSOCIATED PROTEIN KINASE 1	DAPK1_HUMAN	P53355	3dgk:protein:A	64.0	67.0	44.5	0.98
MACROPHAGE COLONY-STIMULATING FACTOR 1 RECEPTOR	CSF1R_HUMAN	P07333	3lcd:protein:A	63.8	67.5	39.4	1.0
PROTEIN KINASE C ALPHA TYPE	KPCA_HUMAN	P17252	3iw4:protein:B	63.6	66.6	44.1	1.0
EPHRIN TYPE-A RECEPTOR 2	EPHA2_HUMAN	P29317	1mqb:protein:B	63.4	67.4	36.9	0.96
ACETYLCHOLINESTERASE	ACES_HUMAN	P22303	1e3q:protein:A	63.4	66.0	46.2	0.98
3-PHOSPHOINOSITIDE DEPENDENT PROTEIN KINASE-1	PDPK_HUMAN	O15530	1uvr:protein:A	63.3	64.8	53.3	1.0

Tous les scores sont normalisés entre 0 et 100. La colonne « score » indique le score global normalisé entre 0 et 100 ; la colonne « geometry » indique le score normalisé des propriétés géométriques des régions ; la colonne « charge » indique le score normalisé des propriétés électrostatiques des régions ; la colonne « aligned » indique le pourcentage d'atomes alignés de la région à comparer avec la région cible.

III. Annexe 3: Quelques résultats du criblage du Dorzolamide sur toutes les structures moléculaires

No	Compound	UniprotAC	UniprotID	Structure	Score	Geometry	Charge	Aligned
1	CARBONIC ANHYDRASE 4	CAH4_MOUSE	Q64444	3f7b:protein:B	100.	100.	95.7	1.0
2	CARBONIC ANHYDRASE II	CAH2_HUMAN	P00918	1g4j:protein:A	88.5	88.2	91.5	0.97
3	CARBONIC ANHYDRASE IV	CAH4_MOUSE	Q64444	3znc:protein:A	84.5	85.3	77.9	1.0
4	CARBONIC ANHYDRASE 13	CAH13_HUMAN	Q8N1Q1	3d0n:protein:B	83.3	83.5	81.3	1.0
5	CARBONIC ANHYDRASE V	CAH5A_MOUSE	P23589	1urt:protein:A	82.3	82.8	78.8	1.0
6	CARBONIC ANHYDRASE XII	CAH12_HUMAN	O43570	1jcz:protein:B	80.3	80.2	80.5	0.94
7	CARBONIC ANHYDRASE III	CAH3_HUMAN	P07451	1z97:protein:A	79.6	78.8	86.4	1.0
8	PHOTOPROTEIN BEROVIN	Unavailable	Unavailable	2hpk:protein:A	77.0	76.3	83.0	1.0
9	HEMOGLOBIN (FERROUS CARBONMONOXY) (ALPHA CHAIN)	HBB_MUSGR	Q9YGW1	1coh:protein:C	76.5	77.0	72.0	1.0
10	GLOBIN III EXTRACELLULAR	GLB3_LUMTE	P11069	1x9f:protein:K	75.3	72.4	100.	1.0
11	CARBONIC ANHYDRASE 9	CAH9_HUMAN	Q16790	3iai:protein:C	74.6	73.7	82.2	0.91
12	CARBONIC ANHYDRASE I	CAH1_HUMAN	P00915	1j9w:protein:A	73.9	73.1	80.5	1.0
13	L-RHAMNOSE ISOMERASE	Q75WH8_PSEST	Q75WH8	3iuh:protein:A	73.9	76.8	49.1	1.0
14	PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR GAMMA	PPARG_HUMAN	P37231	2fvj:protein:A	72.9	74.1	62.7	1.0
15	RESINIFERATOXIN-BINDING PHOSPHOTRIESTERASE-	Q3IVY4_RHOS4	Q3IVY4	3k2g:protein:C	72.4	76.9	33.8	1.0
16	STEROID DELTA-ISOMERASE	SDIS_COMTE	P00947	8cho:protein:A	72.3	75.2	47.4	1.0
17	AMMONIA CHANNEL	AMTB_ECOLI	P69681	2nmr:protein:A	72.0	76.8	30.5	1.0
18	SQUALENE--HOPENE CYCLASE	SQHC_ALIAC	P33247	1gsz:protein:C	71.6	72.9	61.0	1.0
19	DIPEPTIDYL PEPTIDASE 4 (DIPEPTIDYL PEPTIDASE IV)	AAO43404	2856618	2oqv:protein:B	71.5	74.3	47.4	1.0
20	THYMIDINE KINASE	KITH_HHV11	P03176	1p75:protein:D	71.5	78.2	14.4	1.0
21	ALPHA-1-ANTITRYPSIN	A1AT_HUMAN	P01009	1ezx:protein:A	71.4	75.9	33.0	1.0
22	BETA-GLUCOSIDASE	O58104_PYRHO	O58104	1e55:protein:B	71.2	77.7	15.2	1.0
23	PUTATIVE GALACTONATE DEHYDRATASE	Q8XE07_ECO57	Q8XE07	3d47:protein:H	71.1	73.1	54.2	1.0

IV. Annexe 4: Quelques résultats du criblage du coactivateur 2 des récepteurs nucléaires sur toutes les structures moléculaires (KHKILHRLLQD)

No	Compound	UniprotAC	UniprotID	Structure	Score	Geometry	Charge	Aligned
1	RETINOIC ACID RECEPTOR RXR-ALPHA	RXRA_HUMAN	P19793	3e94:protein:A	83.6	81.6	97.2	1.0
2	RETINOIC ACID RECEPTOR RXR-ALPHA	RXRA_MOUSE	P28700	1xdk:protein:E	65.2	64.4	70.8	1.0
3	RETINOIC ACID RECEPTOR RXR-BETA	RXRB_HUMAN	P28702	1uhl:protein:A	61.1	59.9	69.2	1.0
4	STEROID HORMONE RECEPTOR ERR1	ERR1_HUMAN	P11474	3d24:protein:C	58.1	60.0	44.7	1.0
5	PROTEIN (THYROID HORMONE RECEPTOR BETA)	THB1_HUMAN	P10828	1bsx:protein:B	57.8	61.2	34.6	0.98
6	ORPHAN NUCLEAR RECEPTOR PXR	PXR_HUMAN	O75469	2o9i:protein:B	56.9	59.7	37.7	1.0
7	ESTROGEN RECEPTOR BETA	ESR2_HUMAN	Q92731	2nv7:protein:A	55.8	60.0	27.6	1.0
8	ORPHAN NUCLEAR RECEPTOR NR5A2	NR5A2_HUMAN	O00482	1yok:protein:A	55.8	54.4	64.9	0.98
9	PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR ALPHA	PPARA_HUMAN	Q07869	3g8i:protein:A	55.0	57.1	40.8	1.0
10	ESTROGEN RECEPTOR	ESR1_HUMAN	P03372	2g44:protein:B	54.8	57.4	36.9	1.0
11	FARNESOID X RECEPTOR	B7ZM06_HUMAN	B7ZM06	3l1b:protein:A	54.1	57.9	28.4	1.0
12	RXR-LIKE PROTEIN	AAL86461	1938646	1xiu:protein:A	53.7	54.7	47.4	1.0
13	MINERALOCORTICOID RECEPTOR	MCR_HUMAN	P08235	2aa7:protein:A	52.1	56.3	23.7	1.0
14	ESTROGEN RECEPTOR BETA	ESR2_HUMAN	Q9UEV6	1yy4:protein:B	51.4	54.0	33.4	1.0
15	PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR GAMMA	PPARG_HUMAN	P37231	2q6s:protein:A	51.3	53.4	36.9	1.0
16	VITAMIN D3 RECEPTOR	VDR_RAT	P13053	2zmj:protein:A	51.0	52.0	43.9	0.97
17	NR1H4 PROTEIN	A1L4K5_HUMAN	A1L4K5	3fxv:protein:A	50.9	52.4	40.8	1.0
18	VITAMIN D RECEPTOR	Q9PTN2_BRARE	Q9PTN2	2hcd:protein:A	50.9	53.3	34.6	1.0
19	NUCLEAR RECEPTOR ROR-GAMMA	RORG_HUMAN	P51449	3kyt:protein:A	50.7	52.0	42.0	0.96
20	STEROIDOGENIC FACTOR 1	STF1_HUMAN	Q13285	1zdt:protein:A	50.3	52.9	32.6	1.0
21	OXYSTEROLS RECEPTOR LXR-BETA	NR1H2_HUMAN	P55055	3l0e:protein:A	50.1	52.6	33.0	1.0
22	GLUCOCORTICOID RECEPTOR	GCR_HUMAN	P04150	3e7c:protein:A	49.2	51.2	35.4	1.0
23	PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR DELTA	PPARD_HUMAN	Q03181	3dy6:protein:A	49.1	51.6	32.6	1.0
24	NUCLEAR HORMONE RECEPTOR FTZ-F1	Unavailable	Unavailable	2iz2:protein:A	49.1	50.1	42.4	1.0
25	HYDROLASE HALOACID DEHALOGENASE-LIKE FAMILY	Q97NG6_STRPN	Q97NG6	2go7:protein:C	48.9	49.1	47.0	1.0
26	INORGANIC PYROPHOSPHATASE	IPYR_YEAST	P00817	2ik7:protein:B	48.9	52.9	21.4	1.0
27	OUTER MEMBRANE PROTEIN II	Unavailable	Unavailable	3a2u:protein:X	48.8	52.4	24.5	0.96
28	HYPOTHETICAL PROTEIN PH1688	O59312_PYRHO	O59312	2z2v:protein:A	48.7	52.7	21.4	1.0
29	IGG1 MCG INTACT ANTIBODY (HEAVY CHAIN)	AAB21181	243866	1mco:protein:H	48.5	50.0	38.1	1.0
30	PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR GAMMA	Unavailable	Unavailable	2vv1:protein:A	48.2	51.9	23.3	1.0
31	NUCLEAR RECEPTOR ROR-BETA	RORB_RAT	P45446	1nq7:protein:A	48.1	50.1	34.6	1.0
32	IG VH ANTI-LYSOZYME	Unavailable	Unavailable	3a67:protein:H	47.9	54.7	1.55	1.0
33	PROLYL ENDOPEPTIDASE	PPCE_PIG	P23687	1h2x:protein:A	47.9	52.0	20.2	1.0
34	IGG1-KAPPA D44 1 FAB (HEAVY CHAIN)	PC4202	PC4202	1mlb:protein:B	47.9	50.7	28.4	1.0
35	BIFUNCTIONAL 3'-PHOSPHOADENOSINE 5'-	PAPS2_HUMAN	O95340	2ax4:protein:C	47.8	51.4	23.3	1.0
36	ALPHA1 3-FUCOSYLTRANSFERASE	O30511_HELPY	O30511	2nzy:protein:C	47.7	52.2	17.1	1.0
37	PROTEIN (FERRIC HYDROXAMATE UPTAKE RECEPTOR)	FHUA_ECOLI	P06971	1fcp:protein:A	47.6	48.3	43.1	1.0
38	GLUCOSAMINE 6-PHOSPHATE SYNTHASE	GLMS_ECOLI	P17169	1mor:protein:A	47.3	51.4	19.4	1.0
39	AMINOPEPTIDASE N	AMPN_ECOLI	P04825	3b2x:protein:A	47.2	52.0	15.1	0.96
40	LMAJ004091AAA	Q4Q7M2_LEIMA	Q4Q7M2	1xtp:protein:A	47.1	50.5	24.1	1.0
41	HISTONE DEACETYLASE-LIKE AMIDOHYDROLASE	Unavailable	Unavailable	2vcg:protein:D	47.0	51.3	18.2	1.0
42	EXODEOXYRIBONUCLEASE I	EX1_ECOLI	P04995	3c95:protein:A	47.0	50.0	26.4	1.0
43	VIRAL PROTEIN F93	Q6Q0J9_9VIRU	Q6Q0J9	2co5:protein:A	47.0	53.0	6.22	1.0
44	MN20B9 34 ANTI-P1 4 ANTIBODY FAB HEAVY CHAIN	Unavailable	Unavailable	2brr:protein:H	46.8	47.3	43.5	1.0

Références

1. Albou, L.P., O. Poch, and D. Moras, *M-ORBIS: Mapping of mOleculaR Binding sItes and Surfaces.* Nucleic Acids Res, 2010.
2. Connolly, M.L., *Computation of molecular volume.* J Am Chem Soc, 1985. **107**(5): p. 1118-1124.
3. Lee, B. and F.M. Richards, *The interpretation of protein structures: estimation of static accessibility.* J Mol Biol, 1971. **55**(3): p. 379-400.
4. Albou, L.P., et al., *Defining and characterizing protein surface using alpha shapes.* Proteins, 2009. **76**(1): p. 1-12.
5. Friedrich, A., et al., *SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases.* Hum Mutat, 2010. **31**(2): p. 127-35.
6. Hooke, R., *Micrographia: or, Some physiological descriptions of minute bodies made by magnifying glasses.* London: J. Martyn and J. Allestry, 1665.
7. Egerton, F.N., *Van Leeuwenhoek as a founder of animal demography.* Journal of the History of Biology, 1967.
8. Schleiden, M.J., *Contributions to Phytogenesis (Beiträge zur Phytogenesis).* 1838.
9. Schwann, T., *Recherches microscopiques sur la conformité de structure et de croissance des animaux et des plantes (Mikroskopische Untersuchungen über die Übereinstimmung in der Struktur und dem Wachstume der Tiere und Pflanzen).* 1839.
10. Sapp, J., *The prokaryote-eukaryote dichotomy: meanings and mythology.* Microbiol Mol Biol Rev, 2005. **69**(2): p. 292-305.
11. Woese, C.R. and G.E. Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms.* Proc Natl Acad Sci U S A, 1977. **74**(11): p. 5088-90.
12. Woese, C.R., L.J. Magrum, and G.E. Fox, *Archaebacteria.* J Mol Evol, 1978. **11**(3): p. 245-51.
13. Mendel, J.G., *Experiments in Plant Hybridization.* Proceedings of the Natural History Society of Brünn, 1866.
14. Darwin, C., *On the Origin of Species by Means of Natural Selection.* John Murray, 1859.
15. Franzen, J.L., et al., *Complete primate skeleton from the Middle Eocene of Messel in Germany: morphology and paleobiology.* PLoS One, 2009. **4**(5): p. e5723.
16. Hopkins, A.L., *Network pharmacology: the next paradigm in drug discovery.* Nat Chem Biol, 2008. **4**(11): p. 682-90.
17. Delfosse, V., et al., *Structure of the archaeal pab87 peptidase reveals a novel self-compartmentalizing protease family.* PLoS One, 2009. **4**(3): p. e4712.
18. Virchow, R.L.C., *Die Cellular-Pathology.* 1858.

19. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-4.

20. Pauling, L., H.A. Itano, and et al., *Sickle cell anemia a molecular disease.* Science, 1949. **110**(2865): p. 543-8.

21. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.

22. Perutz, M.F., *Structure of hemoglobin.* Brookhaven Symp Biol, 1960. **13**: p. 165-83.

23. Perutz, M.F., et al., *Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis.* Nature, 1960. **185**(4711): p. 416-22.

24. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.

25. Gros, F., et al., *Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli.*. Nature, 1961. **190**: p. 581-5.

26. Crick, F.H., et al., *General nature of the genetic code for proteins.* Nature, 1961. **192**: p. 1227-32.

27. Nirenberg, M.W. and J.H. Matthaei, *The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides.* Proc Natl Acad Sci U S A, 1961. **47**: p. 1588-602.

28. Pestka, S., R. Marshall, and M. Nirenberg, *Rna Codewords and Protein Synthesis. V. Effect of Streptomycin on the Formation of Ribosome-Srna Complexes.* Proc Natl Acad Sci U S A, 1965. **53**: p. 639-46.

29. Trupin, J.S., et al., *Rna Codewords and Protein Synthesis, Vi. On the Nucleotide Sequences of Degenerate Codeword Sets for Isoleucine, Tyrosine, Asparagine, and Lysine.* Proc Natl Acad Sci U S A, 1965. **53**: p. 807-11.

30. Brown, H., F. Sanger, and R. Kitai, *The structure of pig and sheep insulins.* Biochem J, 1955. **60**(4): p. 556-65.

31. Ryle, A.P., et al., *The disulphide bonds of insulin.* Biochem J, 1955. **60**(4): p. 541-56.

32. Sanger, F., E.O. Thompson, and R. Kitai, *The amide groups of insulin.* Biochem J, 1955. **59**(3): p. 509-18.

33. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.

34. Gibson, D.G., et al., *Creation of a bacterial cell controlled by a chemically synthesized genome.* Science. **329**(5987): p. 52-6.

35. Boesen, J.J., et al., *How variable is a spontaneous mutation rate in cultured mammalian cells?* Mutat Res, 1994. **307**(1): p. 121-9.

36. Drake, J.W., et al., *Rates of spontaneous mutation.* Genetics, 1998. **148**(4): p. 1667-86.

37. Elmore, E. and J.C. Barrett, *Measurement of spontaneous mutation rates at the Na+/K+ ATPase locus (ouabain resistance) of human fibroblasts using improved growth conditions.* Mutat Res, 1982. **97**(5): p. 393-404.

38. Glass, B. and R.K. Ritterhoff, *Spontaneous-mutation rates at specific loci in Drosophila males and females.* Science, 1956. **124**(3216): p. 314-5.

39. Meyers, L.A., *Constraints on Variation from Genotype through Phenotype to Fitness.* B.H., K.HB editors - Variation, a central concept in biology, 2005: p. 87-111.

40. Casali, P., et al., *DNA repair in antibody somatic hypermutation.* Trends Immunol, 2006. **27**(7): p. 313-21.

41. Kunkel, T.A., *DNA replication fidelity.* J Biol Chem, 2004. **279**(17): p. 16895-8.

42. Stumpf, M.P., et al., *Estimating the size of the human interactome.* Proc Natl Acad Sci U S A, 2008. **105**(19): p. 6959-64.

43. Jukes, T.H., et al., *Evolution of anticodons: variations in the genetic code.* Cold Spring Harb Symp Quant Biol, 1987. **52**: p. 769-76.

44. Cech, T.R., *Crawling out of the RNA world.* Cell, 2009. **136**(4): p. 599-602.

45. Bahadur, R.P., M. Zacharias, and J. Janin, *Dissecting protein-RNA recognition sites.* Nucleic Acids Res, 2008. **36**(8): p. 2705-16.

46. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.

47. Knudsen, B. and J. Hein, *Pfold: RNA secondary structure prediction using stochastic context-free grammars.* Nucleic Acids Res, 2003. **31**(13): p. 3423-8.

48. Avery, O.T., C.M. Macleod, and M. McCarty, *Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii.* J Exp Med, 1944. **79**(2): p. 137-58.

49. Petsko G.A. and R. D., *Protein Structure and Function.* Primers in Biology, 2004.

50. Jeffery, C.J., *Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins.* Curr Opin Struct Biol, 2004. **14**(6): p. 663-8.

51. Ostrovsky de Spicer, P. and S. Maloy, *PutA protein, a membrane-associated flavin dehydrogenase, acts as a redox-dependent transcriptional regulator.* Proc Natl Acad Sci U S A, 1993. **90**(9): p. 4295-8.

52. Linderstrom-Lang, K., *Proteins and Enzymes.* Stanford University Press, 1952. **Lane Memorial Lectures**.

53. Huang, R.B., et al., *A fast and accurate method for predicting pKa of residues in proteins.* Protein Eng Des Sel. **23**(1): p. 35-42.

54. Li, H., A.D. Robertson, and J.H. Jensen, *Very fast empirical prediction and rationalization of protein pKa values.* Proteins, 2005. **61**(4): p. 704-21.

55. Lee, A.C. and G.M. Crippen, *Predicting pKa.* J Chem Inf Model, 2009. **49**(9): p. 2013-33.

56. Kawashima, S., H. Ogata, and M. Kanehisa, *AAindex: Amino Acid Index Database.* Nucleic Acids Res, 1999. **27**(1): p. 368-9.

57. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D202-5.

58. Taylor, W.R., *The classification of amino acid conservation.* J Theor Biol, 1986. **119**(2): p. 205-18.

59. Kircher, M. and J. Kelso, *High-throughput DNA sequencing--concepts and limitations.* Bioessays. **32**(6): p. 524-36.

60. Smith, S., et al., *High throughput DNA sequencing using an automated electrophoresis analysis system and a novel sequence assembly program.* Biotechniques, 1993. **14**(6): p. 1014-8.

61. Huang, G.M., *High-throughput DNA sequencing: a genomic data manufacturing process.* DNA Seq, 1999. **10**(3): p. 149-53.

62. Miller, N.A., et al., *Management of High-Throughput DNA Sequencing Projects: Alpheus.* J Comput Sci Syst Biol, 2008. **1**: p. 132.

63. Sen, T.Z., et al., *GOR V server for protein secondary structure prediction.* Bioinformatics, 2005. **21**(11): p. 2787-8.

64. Rost, B., C. Sander, and R. Schneider, *PHD--an automatic mail server for protein secondary structure prediction.* Comput Appl Biosci, 1994. **10**(1): p. 53-60.

65. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.

66. Hoskins, J., S. Lovell, and T.L. Blundell, *An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements.* Protein Sci, 2006. **15**(5): p. 1017-29.

67. Ofran, Y. and B. Rost, *Predicted protein-protein interaction sites from local sequence information.* FEBS Lett, 2003. **544**(1-3): p. 236-9.

68. Bairoch, A. and B. Boeckmann, *The SWISS-PROT protein sequence data bank.* Nucleic Acids Res, 1991. **19 Suppl**: p. 2247-9.

69. Marcotte, E.M., et al., *A census of protein repeats.* J Mol Biol, 1999. **293**(1): p. 151-60.

70. Djian, P., *Evolution of simple repeats in DNA and their relation to human disease.* Cell, 1998. **94**(2): p. 155-60.

71. Pauling, L., R.B. Corey, and H.R. Branson, *The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.* Proc Natl Acad Sci U S A, 1951. **37**(4): p. 205-11.

72. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, *Stereochemistry of polypeptide chain configurations.* J Mol Biol, 1963. **7**: p. 95-9.

73. Rost, B., *PHD: predicting one-dimensional protein structure by profile-based neural networks.* Methods Enzymol, 1996. **266**: p. 525-39.

74. Frishman, D. and P. Argos, *Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence.* Protein Eng, 1996. **9**(2): p. 133-42.

75. Pirovano, W. and J. Heringa, *Protein secondary structure prediction.* Methods Mol Biol. **609**: p. 327-48.

76. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

77. de Brevern, A.G., *New assessment of a structural alphabet.* In Silico Biol, 2005. **5**(3): p. 283-9.

78. Etchebest, C., et al., *A structural alphabet for local protein structures: improved prediction methods.* Proteins, 2005. **59**(4): p. 810-27.

79. Li, S.C., et al., *Designing succinct structural alphabets.* Bioinformatics, 2008. **24**(13): p. i182-9.

80. de Brevern, A.G., et al., *Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship.* Protein Sci, 2002. **11**(12): p. 2871-86.

81. Gong, H. and G.D. Rose, *Does secondary structure determine tertiary structure in proteins?* Proteins, 2005. **61**(2): p. 338-43.

82. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-30.

83. Makino, Y., et al., *Chaperonin-mediated folding of green fluorescent protein.* J Biol Chem, 1997. **272**(19): p. 12468-74.

84. Thirumalai, D. and G.H. Lorimer, *Chaperonin-mediated protein folding.* Annu Rev Biophys Biomol Struct, 2001. **30**: p. 245-69.

85. Fenton, W.A. and A.L. Horwich, *Chaperonin-mediated protein folding: fate of substrate polypeptide.* Q Rev Biophys, 2003. **36**(2): p. 229-56.

86. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J, 1986. **5**(4): p. 823-6.

87. Chothia, C. and M. Gerstein, *Protein evolution. How far can sequences diverge?* Nature, 1997. **385**(6617): p. 579, 581.

88. Sander, C. and R. Schneider, *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins, 1991. **9**(1): p. 56-68.

89. Halaby, D.M. and J.P. Mornon, *The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity.* J Mol Evol, 1998. **46**(4): p. 389-400.

90. Illergard, K., D.H. Ardell, and A. Elofsson, *Structure is three to ten times more conserved than sequence--a study of structural response in protein cores.* Proteins, 2009. **77**(3): p. 499-508.

91. Gibrat, J.F., T. Madej, and S.H. Bryant, *Surprising similarities in structure comparison.* Curr Opin Struct Biol, 1996. **6**(3): p. 377-85.

92. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Eng, 1998. **11**(9): p. 739-47.

93. Sael, L., et al., *Fast protein tertiary structure retrieval based on global surface shape similarity.* Proteins, 2008. **72**(4): p. 1259-73.

94. Berman, H., K. Henrick, and H. Nakamura, *Announcing the worldwide Protein Data Bank.* Nat Struct Biol, 2003. **10**(12): p. 980.

95. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.

96. Knudsen, M. and C. Wiuf, *The CATH database.* Hum Genomics. **4**(3): p. 207-12.

97. Yue, P., Z. Li, and J. Moult, *Loss of protein structure stability as a major causative factor in monogenic disease.* J Mol Biol, 2005. **353**(2): p. 459-73.

98. Janin, J., R.P. Bahadur, and P. Chakrabarti, *Protein-protein interaction and quaternary structure.* Q Rev Biophys, 2008. **41**(2): p. 133-80.

99. Alberts, B., *The cell as a collection of protein machines: preparing the next generation of molecular biologists.* Cell, 1998. **92**(3): p. 291-4.

100. Eisenberg, D., et al., *Protein function in the post-genomic era.* Nature, 2000. **405**(6788): p. 823-6.

101. Wells, J.A. and C.L. McClendon, *Reaching for high-hanging fruit in drug discovery at protein-protein interfaces.* Nature, 2007. **450**(7172): p. 1001-9.

102. Janin, J., et al., *Macromolecular recognition in the Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2007. **63**(Pt 1): p. 1-8.

103. Dunker, A.K., et al., *Function and structure of inherently disordered proteins.* Curr Opin Struct Biol, 2008. **18**(6): p. 756-64.

104. Liu, J., J.R. Faeder, and C.J. Camacho, *Toward a quantitative theory of intrinsically disordered proteins and their function.* Proc Natl Acad Sci U S A, 2009. **106**(47): p. 19819-23.

105. Romero, P., Z. Obradovic, and A.K. Dunker, *Natively disordered proteins: functions and predictions.* Appl Bioinformatics, 2004. **3**(2-3): p. 105-13.

106. Janin, J., *Specific versus non-specific contacts in protein crystals.* Nat Struct Biol, 1997. **4**(12): p. 973-4.

107. Gavin, A.C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-7.

108. Ito, T., et al., *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.* Proc Natl Acad Sci U S A, 2000. **97**(3): p. 1143-7.

109. Dagher, R., et al., *Use of a fluorescent polarization based high throughput assay to identify new calmodulin ligands.* Biochim Biophys Acta, 2006. **1763**(11): p. 1250-5.

110. Boute, N., R. Jockers, and T. Issad, *The use of resonance energy transfer in high-throughput screening: BRET versus FRET.* Trends Pharmacol Sci, 2002. **23**(8): p. 351-4.

111. Trugnan, G., et al., *[FRAP, FLIP, FRET, BRET, FLIM, PRIM...new techniques for a colourful life].* Med Sci (Paris), 2004. **20**(11): p. 1027-34.

112. Carriba, P., et al., *Detection of heteromerization of more than two proteins by sequential BRET-FRET.* Nat Methods, 2008. **5**(8): p. 727-33.

113. Jones, S. and J.M. Thornton, *Prediction of protein-protein interaction sites using patch analysis.* J Mol Biol, 1997. **272**(1): p. 133-43.

114. Gallet, X., et al., *A fast method to predict protein interaction sites from sequences.* J Mol Biol, 2000. **302**(4): p. 917-26.

115. Jones, S., et al., *Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins.* Nucleic Acids Res, 2003. **31**(24): p. 7189-98.

116. Reiss, D.J. and B. Schwikowski, *Predicting protein-peptide interactions via a network-based motif sampler.* Bioinformatics, 2004. **20 Suppl 1**: p. i274-82.

117. Liang, S., et al., *Prediction of the interaction site on the surface of an isolated protein structure by analysis of side chain energy scores.* Proteins, 2004. **57**(3): p. 548-57.

118. Murakami, Y. and S. Jones, *SHARP2: protein-protein interaction predictions using patch analysis.* Bioinformatics, 2006. **22**(14): p. 1794-5.

119. Porollo, A. and J. Meller, *Prediction-based fingerprints of protein-protein interactions.* Proteins, 2007. **66**(3): p. 630-45.

120. Petsalaki, E., et al., *Accurate prediction of peptide binding sites on protein surfaces.* PLoS Comput Biol, 2009. **5**(3): p. e1000335.

121. Neuvirth, H., R. Raz, and G. Schreiber, *ProMate: a structure based prediction program to identify the location of protein-protein binding sites.* J Mol Biol, 2004. **338**(1): p. 181-99.

122. Chen, R., L. Li, and Z. Weng, *ZDOCK: an initial-stage protein-docking algorithm.* Proteins, 2003. **52**(1): p. 80-7.

123. Li, L., R. Chen, and Z. Weng, *RDOCK: refinement of rigid-body protein docking predictions.* Proteins, 2003. **53**(3): p. 693-707.

124. Pierce, B., W. Tong, and Z. Weng, *M-ZDOCK: a grid-based approach for Cn symmetric multimer docking.* Bioinformatics, 2005. **21**(8): p. 1472-8.

125. Katchalski-Katzir, E. et al., *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc Natl Acad Sci U S A, 1992. **89**(6): p. 2195-9.

126. Dominguez, C., R. Boelens, and A.M. Bonvin, *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information.* J Am Chem Soc, 2003. **125**(7): p. 1731-7.

127. van Dijk, M., et al., *Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility.* Nucleic Acids Res, 2006. **34**(11): p. 3317-25.

128. Lyskov, S. and J.J. Gray, *The RosettaDock server for local protein-protein docking.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W233-8.

129. Martin, L., V. Catherinot, and G. Labesse, *kinDOCK: a tool for comparative docking of protein kinase ligands.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W325-9.

130. Diago, L.A., et al., *Setting up a large set of protein-ligand PDB complexes for the development and validation of knowledge-based docking algorithms.* BMC Bioinformatics, 2007. **8**: p. 310.

131. Mobley, D.L., et al., *Predicting absolute ligand binding free energies to a simple model site.* J Mol Biol, 2007. **371**(4): p. 1118-34.

132. Bourne, P.E. and H. Weissig, *Structural Bioinformatics.* John Wiley & Sons, 2003.

133. Schultz, P., et al., *Molecular structure of human TFIIH.* Cell, 2000. **102**(5): p. 599-607.

134. Burley, S.K., et al., *Structural genomics: beyond the human genome project.* Nat Genet, 1999. **23**(2): p. 151-7.

135. Thornton, J., *Structural genomics takes off.* Trends Biochem Sci, 2001. **26**(2): p. 88-9.

136. Chandonia, J.M. and S.E. Brenner, *The impact of structural genomics: expectations and outcomes.* Science, 2006. **311**(5759): p. 347-51.

137. Brenner, S.E., C. Chothia, and T.J. Hubbard, *Population statistics of protein structures: lessons from structural classifications.* Curr Opin Struct Biol, 1997. **7**(3): p. 369-76.

138. Sanchez, R. and A. Sali, *Large-scale protein structure modeling of the Saccharomyces cerevisiae genome.* Proc Natl Acad Sci U S A, 1998. **95**(23): p. 13597-602.

139. Pieper, U., et al., *MODBASE, a database of annotated comparative protein structure models and associated resources.* Nucleic Acids Res, 2009. **37**(Database issue): p. D347-54.

140. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.

141. Berman, H.M., et al., *The Nucleic Acid Database.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 889-98.

142. Shin, J.M. and D.H. Cho, *PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures.* Nucleic Acids Res, 2005. **33**(Database issue): p. D238-41.

143. Tusnady, G.E., Z. Dosztanyi, and I. Simon, *PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank.* Nucleic Acids Res, 2005. **33**(Database issue): p. D275-8.

144. Laskowski, R.A., V.V. Chistyakov, and J.M. Thornton, *PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.* Nucleic Acids Res, 2005. **33**(Database issue): p. D266-8.

145. Zanzoni, A., et al., *Phospho3D: a database of three-dimensional structures of protein phosphorylation sites.* Nucleic Acids Res, 2007. **35**(Database issue): p. D229-31.

146. Singh, A., et al., *MutDB: update on development of tools for the biochemical analysis of genetic variation.* Nucleic Acids Res, 2008. **36**(Database issue): p. D815-9.

147. Dayhoff, M.O., et al., *Nucleic acid sequence database.* DNA, 1981. **1**(1): p. 51-8.

148. Berman, H.M., et al., *The nucleic acid database.* Methods Biochem Anal, 2003. **44**: p. 199-216.

149. Sanchez, R. and A. Sali, *ModBase: a database of comparative protein structure models.* Bioinformatics, 1999. **15**(12): p. 1060-1.

150. Sanchez, R., et al., *MODBASE, a database of annotated comparative protein structure models.* Nucleic Acids Res, 2000. **28**(1): p. 250-3.

151. Mooney, S.D. and R.B. Altman, *MutDB: annotating human variation with functionally relevant data.* Bioinformatics, 2003. **19**(14): p. 1858-60.

152. Dantzer, J., et al., *MutDB services: interactive structural analysis of mutation data.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W311-4.

153. Lo Conte, L., C. Chothia, and J. Janin, *The atomic structure of protein-protein recognition sites.* J Mol Biol, 1999. **285**(5): p. 2177-98.

154. Chakrabarti, P. and J. Janin, *Dissecting protein-protein recognition sites.* Proteins, 2002. **47**(3): p. 334-43.

155. Bahadur, R.P., et al., *Dissecting subunit interfaces in homodimeric proteins.* Proteins, 2003. **53**(3): p. 708-19.

156. Bahadur, R.P., et al., *A dissection of specific and non-specific protein-protein interfaces.* J Mol Biol, 2004. **336**(4): p. 943-55.

157. Collins, F.S., M.S. Guyer, and A. Charkravarti, *Variations on a theme: cataloging human DNA sequence variation.* Science, 1997. **278**(5343): p. 1580-1.

158. Taillon-Miller, P., et al., *Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms.* Genome Res, 1998. **8**(7): p. 748-54.

159. Chasman, D. and R.M. Adams, *Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation.* J Mol Biol, 2001. **307**(2): p. 683-706.

160. Thompson, J.D., et al., *MACSIMS: multiple alignment of complete sequences information management system.* BMC Bioinformatics, 2006. **7**: p. 318.

161. Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt, *A model of Evolutionary Change in Proteins.* National Biomedical Research Foundation, 1978.

162. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.

163. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.

164. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

165. Leinonen, R., et al., *UniProt archive.* Bioinformatics, 2004. **20**(17): p. 3236-7.

166. *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2007. **35**(Database issue): p. D193-7.

167. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2000. **28**(1): p. 263-6.

168. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2009.

169. Mulder, N.J., et al., *InterPro: an integrated documentation resource for protein families, domains and functional sites.* Brief Bioinform, 2002. **3**(3): p. 225-35.

170. Notredame, C., *Computing multiple sequence/structure alignments with the T-coffee package.* Curr Protoc Bioinformatics. **Chapter 3**: p. Unit 3 8 1-25.

171. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

172. Thompson, J.D., et al., *DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches.* Nucleic Acids Res, 2000. **28**(15): p. 2919-26.

173. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

174. Sayle, R.A. and E.J. Milner-White, *RASMOL: biomolecular graphics for all.* Trends Biochem Sci, 1995. **20**(9): p. 374.

175. Wang, J.D., M.D. Michelitsch, and J.S. Weissman, *GroEL-GroES-mediated protein folding requires an intact central cavity.* Proc Natl Acad Sci U S A, 1998. **95**(21): p. 12163-8.

176. Bradley, P., et al., *Free modeling with Rosetta in CASP6.* Proteins, 2005. **61 Suppl 7**: p. 128-34.

177. Chothia, C. and J. Janin, *Principles of protein-protein recognition.* Nature, 1975. **256**(5520): p. 705-8.

178. Wodak, S.J. and J. Janin, *Computer analysis of protein-protein interaction.* J Mol Biol, 1978. **124**(2): p. 323-42.

179. Nadassy, K., S.J. Wodak, and J. Janin, *Structural features of protein-nucleic acid recognition sites.* Biochemistry, 1999. **38**(7): p. 1999-2017.

180. Rodier, F., et al., *Hydration of protein-protein interfaces.* Proteins, 2005. **60**(1): p. 36-45.

181. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proc Natl Acad Sci U S A, 1996. **93**(1): p. 13-20.

182. Janin, J. and F. Rodier, *Protein-protein interaction at crystal contacts.* Proteins, 1995. **23**(4): p. 580-7.

183. Ponstingl, H., K. Henrick, and J.M. Thornton, *Discriminating between homodimeric and monomeric proteins in the crystalline state.* Proteins, 2000. **41**(1): p. 47-57.

184. de Vries, S.J., M. van Dijk, and A.M. Bonvin, *The HADDOCK web server for data-driven biomolecular docking.* Nat Protoc. **5**(5): p. 883-97.

185. Wolfson, H.J. and R. Nussinov, *Geometrical docking algorithms. A practical approach.* Methods Mol Biol, 2000. **143**: p. 377-97.

186. Congreve, M., C.W. Murray, and T.L. Blundell, *Structural biology and drug discovery.* Drug Discov Today, 2005. **10**(13): p. 895-907.

187. Nissink, J.W., et al., *A new test set for validating predictions of protein-ligand interaction.* Proteins, 2002. **49**(4): p. 457-71.

188. Halgren, T.A., et al., *Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening.* J Med Chem, 2004. **47**(7): p. 1750-9.

189. Kubinyi, H., *Combinatorial and computational approaches in structure-based drug design.* Curr Opin Drug Discov Devel, 1998. **1**(1): p. 16-27.

190. Ivanov, A.S., et al., *[Computer drug design based on analysis of a target macromolecule structure. I. Search and description of a ligand binding site in a target molecule].* Vopr Med Khim, 2002. **48**(3): p. 304-15.

191. Erlanson, D.A., R.S. McDowell, and T. O'Brien, *Fragment-based drug discovery.* J Med Chem, 2004. **47**(14): p. 3463-82.

192. Ghosh, S., et al., *Structure-based virtual screening of chemical libraries for drug discovery.* Curr Opin Chem Biol, 2006. **10**(3): p. 194-202.

193. Schneider, G. and H.J. Bohm, *Virtual screening and fast automated docking methods.* Drug Discov Today, 2002. **7**(1): p. 64-70.

194. Kubinyi, H., *Success stories in Computer-Aided Design.* Computer Applications in Pharmaceutical Research and Development, 2006.

195. Copley, R.R., I. Letunic, and P. Bork, *Genome and protein evolution in eukaryotes.* Curr Opin Chem Biol, 2002. **6**(1): p. 39-45.

196. Chambon, P., *The nuclear receptor superfamily: a personal retrospect on the first two decades.* Mol Endocrinol, 2005. **19**(6): p. 1418-28.

197. Wang, Z., et al., *Structure and function of Nurr1 identifies a class of ligand-independent nuclear receptors.* Nature, 2003. **423**(6939): p. 555-60.

198. Mangelsdorf, D.J., et al., *The nuclear receptor superfamily: the second decade.* Cell, 1995. **83**(6): p. 835-9.

199. Krust, A., et al., *The chicken oestrogen receptor sequence: homology with v-erbA and the human oestrogen and glucocorticoid receptors.* EMBO J, 1986. **5**(5): p. 891-7.

200. Green, S. and P. Chambon, *Oestradiol induction of a glucocorticoid-responsive gene by a chimaeric receptor.* Nature, 1987. **325**(6099): p. 75-8.

201. Kaffman, A., et al., *Phosphorylation of the transcription factor PHO4 by a cyclin-CDK complex, PHO80-PHO85.* Science, 1994. **263**(5150): p. 1153-6.

202. Takaki, T., et al., *Cdk-mediated phosphorylation of pRB regulates HDAC binding in vitro.* Biochem Biophys Res Commun, 2004. **316**(1): p. 252-5.

203. Kjaerulff, S., et al., *Cdk phosphorylation of the Ste11 transcription factor constrains differentiation-specific transcription to G1.* Genes Dev, 2007. **21**(3): p. 347-59.

204. Honey, S. and B. Futcher, *Roles of the CDK phosphorylation sites of yeast Cdc6 in chromatin binding and rereplication.* Mol Biol Cell, 2007. **18**(4): p. 1324-36.

205. Szondy, Z., U. Reichert, and L. Fesus, *Retinoic acids regulate apoptosis of T lymphocytes through an interplay between RAR and RXR receptors.* Cell Death Differ, 1998. **5**(1): p. 4-10.

206. Monden, T., et al., *Unliganded RXR acts as an inhibitory factor on troglitazone-induced activation.* Life Sci, 2004. **76**(7): p. 731-41.

207. Field, F.J., E. Born, and S.N. Mathur, *LXR/RXR ligand activation enhances basolateral efflux of beta-sitosterol in CaCo-2 cells.* J Lipid Res, 2004. **45**(5): p. 905-13.

208. Matsusue, K., et al., *Ligand-activated PPARbeta efficiently represses the induction of LXR-dependent promoter activity through competition with RXR.* Mol Cell Endocrinol, 2006. **256**(1-2): p. 23-33.

209. Wang, K., et al., *Retinoids activate the RXR/SXR-mediated pathway and induce the endogenous CYP3A4 activity in Huh7 human hepatoma cells.* Toxicol Sci, 2006. **92**(1): p. 51-60.

210. Ribeiro, R.C., P.J. Kushner, and J.D. Baxter, *The nuclear hormone receptor gene superfamily.* Annu Rev Med, 1995. **46**: p. 443-53.

211. Brelivet, Y., *Etude de voies de communication. Application aux Récepteurs Nucléaires.* Thèse de l'Université Louis Pasteur, Strasbourg 1, 2008.

212. Brelivet, Y., et al., *Signature of the oligomeric behaviour of nuclear receptors at the sequence and structural level.* EMBO Rep, 2004. **5**(4): p. 423-9.

213. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.

214. Cooper, D.N., P.D. Stenson, and N.A. Chuzhanova, *The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms.* Curr Protoc Bioinformatics, 2006. **Chapter 1**: p. Unit 1 13.

215. Kornberg, R. and Y. Lorch, *Chromatine rules.* Nat Struct Biol, 2007. **14**.

216. Laskowski, R.A. and J.M. Thornton, *Understanding the molecular machinery of genetics through 3D structures.* Nat Rev Genet, 2008. **9**(2): p. 141-51.

217. Pauling, L., *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry.* 1939.

218. Arnaud, P., *Chimie-Physique.* Dunod.

219. Tsai, J., et al., *The packing density in proteins: standard radii and volumes.* J Mol Biol, 1999. **290**(1): p. 253-66.

220. Richards, F.M., *The interpretation of protein structures: total volume, group volume distributions and packing density.* J Mol Biol, 1974. **82**(1): p. 1-14.

221. Tsai, J., N. Voss, and M. Gerstein, *Determining the minimum number of types necessary to represent the sizes of protein atoms.* Bioinformatics, 2001. **17**(10): p. 949-56.

222. Fleming, P.J. and F.M. Richards, *Protein packing: dependence on protein size, secondary structure and amino acid composition.* J Mol Biol, 2000. **299**(2): p. 487-98.

223. Li, A.J. and R. Nussinov, *A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking.* Proteins, 1998. **32**(1): p. 111-27.

224. Liang, J. and K.A. Dill, *Are proteins well-packed?* Biophys J, 2001. **81**(2): p. 751-66.

225. Pellicena, P. and J. Kuriyan, *Protein-protein interactions in the allosteric regulation of protein kinases.* Curr Opin Struct Biol, 2006. **16**(6): p. 702-9.

226. Bhat, S. and E.O. Purisima, *Molecular surface generation using a variable-radius solvent probe.* Proteins, 2006. **62**(1): p. 244-61.

227. Zhang, Q., M. Sanner, and A.J. Olson, *Shape complementarity of protein-protein complexes at multiple resolutions.* Proteins, 2009. **75**(2): p. 453-67.

228. Janin, J., S. Miller, and C. Chothia, *Surface, subunit interfaces and interior of oligomeric proteins.* J Mol Biol, 1988. **204**(1): p. 155-64.

229. Miller, S., et al., *Interior and surface of monomeric proteins.* J Mol Biol, 1987. **196**(3): p. 641-56.

230. Nguyen, M.N. and J.C. Rajapakse, *Prediction of protein relative solvent accessibility with a two-stage SVM approach.* Proteins, 2005. **59**(1): p. 30-7.

231. Chen, K., M. Kurgan, and L. Kurgan, *Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values.* J Biomedical Science and Engineering, 2008. **1**: p. 1-9.

232. Gerstein, M., J. Tsai, and M. Levitt, *The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra.* J Mol Biol, 1995. **249**(5): p. 955-66.

233. Tsai, J. and M. Gerstein, *Calculations of protein volumes: sensitivity analysis and parameter database.* Bioinformatics, 2002. **18**(7): p. 985-95.

234. McConkey, B.J., V. Sobolev, and M. Edelman, *Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure.* Bioinformatics, 2002. **18**(10): p. 1365-73.

235. Poupon, A., *Voronoi and Voronoi-related tessellations in studies of protein structure and interaction.* Curr Opin Struct Biol, 2004. **14**(2): p. 233-41.

236. Edelsbrunner, H., M. Facello, and J. Liang, *On the definition and the construction of pockets in macromolecules.* Pac Symp Biocomput, 1996: p. 272-87.

237. Liang, J., H. Edelsbrunner, and C. Woodward, *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.* Protein Sci, 1998. **7**(9): p. 1884-97.

238. Liang, J., et al., *Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins.* Proteins, 1998. **33**(1): p. 18-29.

239. Liang, J., et al., *Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape.* Proteins, 1998. **33**(1): p. 1-17.

240. *CGAL, Computational Geometry Algorithms Library, http://www.cgal.org.*

241. Hubbard, S. and J.M. Thornton, *Naccess.* Department of Biochemistry and Molecular Biology, 1993. **University College London**.

242. Chakravarty, S. and R. Varadarajan, *Residue depth: a novel parameter for the analysis of protein structure and stability.* Structure, 1999. **7**(7): p. 723-32.

243. Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces.* J Mol Biol, 1998. **280**(1): p. 1-9.

244. Hubbard, S.J. and P. Argos, *Cavities and packing at protein interfaces.* Protein Sci, 1994. **3**(12): p. 2194-206.

245. Cazals, F., et al., *Revisiting the Voronoi description of protein-protein interfaces.* Protein Sci, 2006. **15**(9): p. 2082-92.

246. Bernauer, J., et al., *DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions.* Bioinformatics, 2008. **24**(5): p. 652-8.

247. Cherfils, J., S. Duquerroy, and J. Janin, *Protein-protein recognition analyzed by docking simulation.* Proteins, 1991. **11**(4): p. 271-80.

248. Douguet, D., et al., *DOCKGROUND resource for studying protein-protein interfaces.* Bioinformatics, 2006. **22**(21): p. 2612-8.

249. Liu, S., Y. Gao, and I.A. Vakser, *DOCKGROUND protein-protein docking decoy set.* Bioinformatics, 2008. **24**(22): p. 2634-5.

250. Chen, R. and Z. Weng, *Docking unbound proteins using shape complementarity, desolvation, and electrostatics.* Proteins, 2002. **47**(3): p. 281-94.

251. Hwang, H., et al., *Protein-protein docking benchmark version 3.0.* Proteins, 2008. **73**(3): p. 705-9.

252. Janin, J. and S. Wodak, *The third CAPRI assessment meeting Toronto, Canada, April 20-21, 2007.* Structure, 2007. **15**(7): p. 755-9.

253. Shoichet, B.K., et al., *Lead discovery using molecular docking.* Curr Opin Chem Biol, 2002. **6**(4): p. 439-46.

254. Schapira, M., et al., *In silico discovery of novel retinoic acid receptor agonist structures.* BMC Struct Biol, 2001. **1**: p. 1.

255. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. **112**(3): p. 535-42.

256. Chothia, C., *The nature of the accessible and buried surfaces in proteins.* J Mol Biol, 1976. **105**(1): p. 1-12.

257. Janin, J. and C. Chothia, *Stability and specificity of protein-protein interactions: the case of the trypsin-trypsin inhibitor complexes.* J Mol Biol, 1976. **100**(2): p. 197-211.

258. Janin, J., *Principles of protein-protein recognition from structure to thermodynamics.* Biochimie, 1995. **77**(7-8): p. 497-505.

259. Stanfield, R.L. and I.A. Wilson, *Protein-peptide interactions.* Curr Opin Struct Biol, 1995. **5**(1): p. 103-13.

260. Jones, S., et al., *Protein-DNA interactions: A structural analysis.* J Mol Biol, 1999. **287**(5): p. 877-96.

261. Jones, S., et al., *Protein-RNA interactions: a structural analysis.* Nucleic Acids Res, 2001. **29**(4): p. 943-54.

262. Ponsting, H., T. Kabir, and J.M. Thornton, *Automatic inference of protein quaternary structure from crystals.* J Appl Cryst, 2003. **36**: p. 1116-1122

263. Henrick, K. and J.M. Thornton, *PQS: a protein quaternary structure file server.* Trends Biochem Sci, 1998. **23**(9): p. 358-61.

264. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* J Mol Biol, 2007. **372**(3): p. 774-97.

265. Miyamoto, T., et al., *The role of hinge domain in heterodimerization and specific DNA recognition by nuclear receptors.* Mol Cell Endocrinol, 2001. **181**(1-2): p. 229-38.

266. Moras, D., *Structure of full-length PPARgamma-RXRalpha: a snapshot of a functional complex?* Cell Metab, 2009. **9**(1): p. 8-10.

267. Chandra, V., et al., *Structure of the intact PPAR-gamma-RXR-alpha nuclear receptor complex on DNA.* Nature, 2008: p. 350-356.

268. Crosio, M.P., J. Janin, and M. Jullien, *Crystal packing in six crystal forms of pancreatic ribonuclease.* J Mol Biol, 1992. **228**(1): p. 243-51.

269. Vanhee, P., et al., *Protein-peptide interactions adopt the same structural motifs as monomeric protein folds.* Structure, 2009. **17**(8): p. 1128-36.

270. Vanhee, P., et al., *PepX: a structural database of non-redundant protein-peptide complexes.* Nucleic Acids Res. **38**(Database issue): p. D545-51.

271. Neduva, V., et al., *Systematic discovery of new recognition peptides mediating protein interaction networks.* PLoS Biol, 2005. **3**(12): p. e405.

272. Liang, S., et al., *Protein binding site prediction using an empirical scoring function.* Nucleic Acids Res, 2006. **34**(13): p. 3698-707.

273. Chung, J.L., W. Wang, and P.E. Bourne, *Exploiting sequence and structure homologs to identify protein-protein binding sites.* Proteins, 2006. **62**(3): p. 630-40.

274. Higurashi, M., T. Ishida, and K. Kinoshita, *PiSite: a database of protein interaction sites using multiple binding states in the PDB.* Nucleic Acids Res, 2009. **37**(Database issue): p. D360-4.

275. Lijnzaad, P., H.J. Berendsen, and P. Argos, *Hydrophobic patches on the surfaces of protein structures.* Proteins, 1996. **25**(3): p. 389-97.

276. Landau, M., et al., *ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W299-302.

277. Tseng, Y.Y., et al., *SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W384-9.

278. Rajamani, D., et al., *Anchor residues in protein-protein interactions.* Proc Natl Acad Sci U S A, 2004. **101**(31): p. 11287-92.

279. Guharoy, M. and P. Chakrabarti, *Conservation and relative importance of residues across protein-protein interfaces.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15447-52.

280. Raviscioni, M., et al., *Evolutionary identification of a subtype specific functional site in the ligand binding domain of steroid receptors.* Proteins, 2006. **64**(4): p. 1046-57.

281. Kern, D. and E.R. Zuiderweg, *The role of dynamics in allosteric regulation.* Curr Opin Struct Biol, 2003. **13**(6): p. 748-57.

282. Plewniak, F., et al., *PipeAlign: A new toolkit for protein family analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3829-32.

283. Chelliah, V., et al., *Distinguishing structural and functional restraints in evolution in order to identify interaction sites.* J Mol Biol, 2004. **342**(5): p. 1487-504.

284. Langley, J.N., *On the reaction of cells and of nerve-endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curari.* J Physiol, 1905. **33**(4-5): p. 374-413.

285. Filozof, C.M., et al., *Low plasma leptin concentration and low rates of fat oxidation in weight-stable post-obese subjects.* Obes Res, 2000. **8**(3): p. 205-10.

286. Aronne, L.J. and Z.D. Thornton-Jones, *New targets for obesity pharmacotherapy.* Clin Pharmacol Ther, 2007. **81**(5): p. 748-52.

287. Chouchane, L., et al., *A repeat polymorphism in interleukin-4 gene is highly associated with specific clinical phenotypes of asthma.* Int Arch Allergy Immunol, 1999. **120**(1): p. 50-5.

288. Kellenberger, E., et al., *sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank.* J Chem Inf Model, 2006. **46**(2): p. 717-27.

289. Han, L.Y., et al., *Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness.* Drug Discov Today, 2007. **12**(7-8): p. 304-13.

290. An, J., M. Totrov, and R. Abagyan, *Comprehensive identification of "druggable" protein ligand binding sites.* Genome Inform, 2004. **15**(2): p. 31-41.

291. Hajduk, P.J., J.R. Huth, and C. Tse, *Predicting protein druggability.* Drug Discov Today, 2005. **10**(23-24): p. 1675-82.

292. Campbell, S.F., *Science, art and drug discovery: a personal perspective.* Clin Sci (Lond), 2000. **99**(4): p. 255-60.

293. Mizutani, M.Y. and A. Itai, *Efficient method for high-throughput virtual screening based on flexible docking: discovery of novel acetylcholinesterase inhibitors.* J Med Chem, 2004. **47**(20): p. 4818-28.

294. Shoichet, B.K., *Virtual screening of chemical libraries.* Nature, 2004. **432**(7019): p. 862-5.

295. Klebe, G., *Virtual ligand screening: strategies, perspectives and limitations.* Drug Discov Today, 2006. **11**(13-14): p. 580-94.

296. Lundstrom, K., *An overview on GPCRs and drug discovery: structure-based drug design and structural biology on GPCRs.* Methods Mol Biol, 2009. **552**: p. 51-66.

297. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor.* Science, 2000. **289**(5480): p. 739-45.

298. Bodenstaff, E.R., et al., *The prospects of protein nanocrystallography.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 11): p. 1901-6.

299. Williams, P.A., et al., *Mammalian microsomal cytochrome P450 monooxygenase: structural adaptations for membrane binding and functional diversity.* Mol Cell, 2000. **5**(1): p. 121-31.

300. Williams, P.A., et al., *Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone.* Science, 2004. **305**(5684): p. 683-6.

301. Yano, J.K., et al., *The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-A resolution.* J Biol Chem, 2004. **279**(37): p. 38091-4.

302. Gaasterland, T., *Structural genomics taking shape.* Trends Genet, 1998. **14**(4): p. 135.

303. Blundell, T.L., *Structure-based drug design.* Nature, 1996. **384**(6604 Suppl): p. 23-6.

304. Blundell, T., B.L. Sibanda, and L. Pearl, *Three-dimensional structure, specificity and catalytic mechanism of renin.* Nature, 1983. **304**(5923): p. 273-5.

305. Miller, M., et al., *Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution.* Science, 1989. **246**(4934): p. 1149-52.

306. Lapatto, R., et al., *X-ray analysis of HIV-1 proteinase at 2.7 A resolution confirms structural homology among retroviral enzymes.* Nature, 1989. **342**(6247): p. 299-302.

307. Kubinyi, H., *Structure-based design of enzyme inhibitors and receptor ligands.* Curr Opin Drug Discov Devel, 1998. **1**(1): p. 4-15.

308. Kim, C.U., X. Chen, and D.B. Mendel, *Neuraminidase inhibitors as anti-influenza virus agents.* Antivir Chem Chemother, 1999. **10**(4): p. 141-54.

309. Jones, S. and J.M. Thornton, *Analysis of protein-protein interaction sites using surface patches.* J Mol Biol, 1997. **272**(1): p. 121-32.

310. Naylor, E., et al., *Identification of a chemical probe for NAADP by virtual screening.* Nat Chem Biol, 2009. **5**(4): p. 220-6.

311. Carr, R.A., et al., *Fragment-based lead discovery: leads by design.* Drug Discov Today, 2005. **10**(14): p. 987-92.

312. Jhoti, H., *Fragment-based drug discovery using rational design.* Ernst Schering Found Symp Proc, 2007(3): p. 169-85.

313. Congreve, M., et al., *Recent developments in fragment-based drug discovery.* J Med Chem, 2008. **51**(13): p. 3661-80.

314. Warr, W.A., *Fragment-based drug discovery.* J Comput Aided Mol Des, 2009.

315. Fischer, M. and R.E. Hubbard, *Fragment-based ligand discovery.* Mol Interv, 2009. **9**(1): p. 22-30.

316. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.

317. Wishart, D.S., *DrugBank and its relevance to pharmacogenomics.* Pharmacogenomics, 2008. **9**(8): p. 1155-62.

318. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets.* Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.

319. Kuhn, M., et al., *A side effect resource to capture phenotypic effects of drugs.* Mol Syst Biol. **6**: p. 343.

320. Schmidt, U., et al., *SuperToxic: a comprehensive database of toxic compounds.* Nucleic Acids Res, 2009. **37**(Database issue): p. D295-9.

321. Campillos, M., et al., *Drug target identification using side-effect similarity.* Science, 2008. **321**(5886): p. 263-6.

322. Fliri, A.F., et al., *Analysis of drug-induced effect patterns to link structure and side effects of medicines.* Nat Chem Biol, 2005. **1**(7): p. 389-97.

323. Schuffenhauer, A., et al., *Similarity metrics for ligands reflecting the similarity of the target proteins.* J Chem Inf Comput Sci, 2003. **43**(2): p. 391-405.

324. Keiser, M.J., et al., *Predicting new molecular targets for known drugs.* Nature, 2009. **462**(7270): p. 175-81.

325. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

326. Jambon, M., et al., *A new bioinformatic approach to detect common 3D sites in protein structures.* Proteins, 2003. **52**(2): p. 137-45.

327. Schalon, C., et al., *A simple and fuzzy method to align and compare druggable ligand-binding sites.* Proteins, 2008. **71**(4): p. 1755-78.

328. Mavridis, L., B.D. Hudson, and D.W. Ritchie, *Toward high throughput 3D virtual screening using spherical harmonic surface representations.* J Chem Inf Model, 2007. **47**(5): p. 1787-96.

329. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

330. Bairoch, A., et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33**(Database issue): p. D154-9.

331. Frezal, J., *Genatlas database, genes and development defects.* C R Acad Sci III, 1998. **321**(10): p. 805-17.

332. Roux-Rouquie, M., et al., *Human genes involved in chromatin remodeling in transcription initiation, and associated diseases: An overview using the GENATLAS database.* Mol Genet Metab, 1999. **67**(4): p. 261-77.