UNIVERSITÉ DE STRASBOURG

Institut de Génétique et de Biologie Moléculaire et Cellulaire

Département de Biologie et de Génomique Structurale

Thèse présentée pour obtenir le grade de
Docteur de l'Université de Strasbourg

Discipline : Sciences du vivant
Spécialité : Aspects mollaires et cellulaires de la biologie

par Benjamin SCHWARZ

Application de la théorie des formes-α pour la caractérisation de la surface et des poches de macromolécules biologiques

Soutenue publiquement le 7 septembre 2009

Membres du jury

Directeur : M. Jean-Marie Wurtz, Professeur
Université de Strasbourg
Rapporteur interne : M. Marcel HIBERT, Professeur
Université de Strasbourg
Rapporteurs externes : M. Patrice Koehl, Professeur
University of California, Davis
M. Frédéric Cazals, Directeur de Recherche
INRIA Sophia-AntipolisUniversity of California
Examinateur : Michel Souchet, chef de projet
LORIA Espace Transfert
Co-encadrants (membres invités) : Mme. Dominique Bechmann, Professeur
Université de Strasbourg

Toutes les représentations moléculaires produites dans ce document ont été réalisées avec *PovRay* [Povray 96] au travers du logiciel de visualisation de molécules *VMD* [Humphrey 96].
Les surfaces moléculaires ont systématiquement été générées avec *msms* [Sanner 95].
Les figures et schémas ont été réalisées avec *Inkscape* (`http://www.inkscape.org/`), et *TheGimp*(`http://www.gimp.org/`).

Ce document a été composé avec LaTeX.

A Aniline

Si on résume, finalement, c'est quoi une thèse ? Une période plus ou moins longue dans la vie d'un individu, où l'attention est entièrement captée par un objectif parfois (souvent ?) élusif. Une période où s'alternent les moments de doutes et ceux de certitudes, ponctués d'instants de stress ou d'exaltation. C'est aussi l'occasion d'apprendre énormément, et d'apporter un peu ; et tout cela se résume enfin aux quelques pages d'un mémoire de thèse. Arrivé au terme de ce voyage je me retourne et je constate le chemin parcouru. Un chemin dont le cours doit pour beaucoup à de nombreuses rencontres que j'ai plaisir à saluer ici.

En tout premier lieu, je tiens à remercier ceux sans qui cette aventure n'aurait tout simplement pu débuter. Le Pr. Moras qui m'a accueilli au sein de son laboratoire et à qui je dois le financement des deux premières années ; et bien entendu, Jean-Marie et Dominique qui ont accepté de m'encadrer en dépit des difficultés et de l'investissement en temps que représentent un sujet "à l'interface".

Au cours de ces quelques années, une alternance d'amis ont partagé et construit mon quotidien de laboratoire. Souphatta qui n'a pas épargné de son temps de rédaction pour me donner les rudiments de biologie qui me faisaient défaut à mon arrivée. Olivier l'homme à la thyroïde d'acier. Christophe qui a veillé à ma bonne oxygénation, ces promenades me manqueront. Pierre et Nathalie ; et enfin Yann, arrivé tardivement et que j'aurais aimé connaître mieux. La fine équipe d'"en face" : Luc (je ne sais pas si ce sont les discussions ou la bière qui me manqueront le plus), Nico qui ne m'a jamais fait payer les fameux "un-euro", Dave... Et bien entendu Fred et sa présence temporisante. Et enfin, Basile et Nathanaël pour toutes ces discussions qui permettent d'approfondir ou de changer le point de vue.

Je veux aussi remercier ceux qui n'ont jamais épargné de leur temps pour répondre à mes questions. James Kanze qui sur le *newsgroup* `fr.comp.lang.c++` répond systématiquement et de manière circonstanciée aux questions même les plus naïves. La *mailing-list* `CGAL` et les participants des `JGA`. Enfin, Pierre pour sa présence rassurante à l'arrière de ma debian.

Merci à Arnaud (Miaou !) et Julien, qui, le temps de la rédaction étant venu, ont perpétué la tradition du passage de fichier template LaTeX ; la forme de ce document leur doit beaucoup. Merci à Claire pour ces quelques mois de rédaction partagés et pour avoir remis de l'eau dans ma théière lorsque je fumais trop. Je ne remercierai jamais assez ceux qui ont donné de leur temps pour la correction de ce manuscrit. A commencer par Mathieu, chantre de LaTeX et typographe pinailleur qui a passablement influé tant sur la forme que sur le fond de ce document. Merci aussi à Vincent, Justyna et Bénédicte pour le marathon de la correction des derniers jours avant impression, ainsi que pour leur présence au quotidien ; et tout particulièrement à Bénédicte pour m'avoir toujours supporté.

Enfin, je tiens à remercier les membres du Jury pour avoir accepté de juger mon travail et tout particulièrement les rapporteurs pour leurs commentaires nourris et constructifs.

Table des matières

Liste des figures

Abréviations et notations

A	molécule
a_i	atome (d'une molécule A)
b_i, b_{a_i}	boule $(a_i, r_i) = \{x \mid d(x, a_i) \leq r_i\}$ associée à l'atome a_i
s_i, s_{a_i}	sphère $(a_i, r_i) = \{x \mid d(x, a_i) = r_i\}$ associée au bord de l'atome a_i
$\mathcal{D}$	triangulation de Delaunay d'un ensemble de points
$\mathcal{V}$	diagramme de Voronoï d'un ensemble de points
$\mathcal{V}_i$	cellule de Voronoï associée à l'atome a_i
$\mathcal{K}$	complexe dual d'un ensemble de points
$\mathcal{K}_\alpha$	complexe-α d'un ensemble de points
$\mathcal{H}$	enveloppe convexe d'un ensemble de points
$\mathcal{H}_p$	enveloppe convexe de la poche p
k_p	indice de convexité de la poche p
$\mathcal{P}_p$	représentation polyédrique de la poche p
Ω	angle solide
$\alpha(v)$	exposition-α du sommet v
$lc(v)$, $lc_N(v)$, $lc_N^\mu(v)$	courbure locale du sommet v obtenue avec la "fonction de moyenne" μ sur le voisinage $V_N(v)$
$lc_{N:m:M}^\mu(v)$	courbure locale du sommet v pour la méthode de lissage μ, la taille de voisinage N, et un seuillage dans l'intervalle $[m, M]$
$lc_{N:\chi\%}^\mu$, $lc_{N:\chi\%:m:M}^\mu$	courbure locale du sommet v pour la méthode de lissage μ, la taille de voisinage N, et un seuillage à $\chi\%$ des valeurs extrêmes.
$\mathcal{D}_f$	domaine d'application d'une fonction f
$d(a, b)$	distance euclidienne du point a au point b
$\pi_{\mathcal{C}_c^r}(x)$	puissance du point x à la sphère $\mathcal{C}_x^r$, de centre c et de rayon r. $\pi(x, \mathcal{C}_c^r) = d(x, c)^2 - r^2$
V_v^N	voisinage de taille N autour de v
C_v^N	$N^{\text{ième}}$ corolle de v
`VdW`	Van der Waals
`SA`	surface accessible (modèle, ou surface)
`MS`	surface moléculaire (aussi surface de Connolly)
ASA	aire de la surface accessible (au solvant) (accesible surface area)

Glossaire

adaptation induite	Phénomène dans lequel les intervenants moléculaires d'une interaction modifient leur forme pour maximiser leur complémentarité. Ce modèle complète le modèle clé-serrure proposé par Fischer au 19^{e} siècle.
adjacent	Deux simplexes de même dimension, partageant une partie de leur bord sont dits adjacents
affinité	Mesure la force de la liaison entre deux partenaires.
boule	En mathématique on parle de la boule centrée sur un point x et de rayon r, pour désigner l'ensemble des points situés à une distance au plus r du point x ; et ce quelle que soit la dimension considérée. En deux dimensions, une boule est appelée un disque.
cellule (combinatoire)	De manière générale, une k-cellule est un ensemble homéomorphe à une k-boule. Dans l'idiome `CGAL`, une cellule est plus spécifiquement un 3-simplexe
clé-serrure (modèle)	Modèle proposé par Fisher au 19^{e} siècle pour expliquer la reconnaissance moléculaire. Dans ce modèle, on suppose que la forme (et la nature physico-chimique) des deux intervenants moléculaires que sont le ligand et l'enzyme (ou plus généralement la protéine) sont complémentaires. La réalité expérimentale montre que dans certains cas les protagonistes doivent modifier légèrement leur conformation afin d'accomoder leur forme. Ce phénomène porte le nom d'adaptation induite.
coface	Dans une structure combinatoire, la notion de coface est duale de celle de face : τ est une coface de σ signifie exactement que σ est une face de τ
corolle (d'un sommet)	La $n^{\text{ième}}$ corolle d'un sommet v d'un maillage de surface V_v^N, est l'ensemble des sommets qu'on peut atteindre en n arêtes et pas en moins.
domaine	Élément structural d'une protéine généralement clairement décelable dans la structure tertiaire, et isolément fonctionnel.
étoile, *star*	Dans un complexe simpliciel, l'étoile d'un simplexe σ est l'ensemble des simplexes incidents à σ.
face	Les faces d'un k-simplexe σ sont tous les l-simplexes ($l < k$) situés sur le bord de σ

faisceau (d'ombrelle)	Désigne l'espace vide au dessus de l'ombrelle. Un faisceau peut comporter une composante infinie (en dehors de l'enveloppe convexe du nuage de points) et une composante finie (interne à l'enveloppe convexe).
faisceau (de tétraèdres)	l'ensemble des tétraèdres des $\mathcal{D}$ participant à une ombrelle (i.e. recouvrant le faisceau de cette ombrelle).
incident	Deux simplexes dont l'un est sur le bord de l'autre sont dits incidents.
link	Dans une structure combinatoire (plus formellement dans un complexe simpliciel), le link $\mathrm{lk}(\sigma)$ d'un simplexe σ est constitué de tous les simplexes τ n'ayant aucune intersection commune avec σ, mais dont l'union avec σ constitue un simplexe.
maillage (de surface)	Représentation informatique d'une surface basée sur une décomposition en sommets, en arêtes, et souvent aussi en facettes. Ces facettes peuvent être de taille et de forme variées.
monomère	Structure d'une molécule isolée (c'est-à-dire non impliquée dans un complexe).
ombrelle	Sur la surface duale ou une variété combinatoire, l'ombrelle autour d'un sommet v désigne la structure restreinte aux éléments incidents à ce sommet (i.e. son étoile.). Dans le complexe dual, on peut définir des ombrelles autour d'un sommet à partir du link de ce sommet.
PDB	Protein Data Bank, une (la) base de données autorisant à un vaste ensemble de structures moléculaires produites par les chercheurs du monde entier. Par extension, l'acronyme *pdb* désigne aussi un format de fichiers pour le stockage de l'information de structure.
protomère	Structure d'une molécule impliquée dans un complexe mais considérée isolément.
remplissage de forme diagramme à	En anglais, *spacefilling diagram*. Modèle dans lequel chaque atome d'une molécule se voit attribuer une partie de la boule qui le représente.
reconnaissance (moléculaire)	Combinaison d'affinité et de spécificité permettant à une molécule de se lier précisément à une autre molécule ou à un certain type de molécules.
simplexe	Dans ce document ce sont les éléments d'une triangulation : sommet, arête, facette triangulaire ou cellule tétraédrique.
sous-unité	Désigne une macromolécule spécifique dans un assemblage moléculaire.
spécificité	Mesure la capacité d'une molécule à reconnaître un partenaire précis.
sphère	Le bord d'une boule, l'ensemble des points à égale distance d'un point central. Par extension, un cercle est une sphère, on en précise la dimension en parlant de d-sphère.

traduction (de l'ADN)	La "transformation" du message de l'ADN en protéine. Ce processus a lieu à l'intérieur du complexe ribosomal : des triplets de nucléotides (appelés codons) sont "lus" sur un brin d'ARN messager (ARNm) et appariés à leur anticodon porté par un ARN de transfert (ARNt). Les acides aminés portés par les ARNt sont séquentiellement assemblés en une chaîne peptidique, résultant en protéine.
transcription (de l'ADN)	L'extraction et la dissémination d'une partie de l'information contenue sur l'ADN. L'ADN est lu, et son information transférée sous la forme d'ARN messager.
transduction	Passage d'un "signal" ou d'une "information". Généralement, le terme de transduction fait référence à un changement de support de cette information.
variété	Classe d'objets permettant de définir formellement l'intuition de surface idéale. Les surfaces variété sont présentées à la page 70.
voxel	Élément volumique dans les approches discrètes (grilles de voxel) ; équivalent tridimensionnel du pixel en deux dimension.

Avant-propos, comment lire ce document ?

Nos travaux ont essentiellement consisté à proposer de nouveaux outils dédiés à l'analyse de la structure des macromolécules ; une courte présentation des implications d'une telle étude pourra être trouvée dans l'introduction à la page 1, et une annexe sur la structure des macromolécules (annexe A page 165) fournit les connaissances de bases en biologie pour autoriser l'accès à ce document.

Notre étude repose en grande partie sur des modèles et des formalismes issus de la géométrie algorithmique, une compréhension minimale de ces modèles s'avère donc nécessaire. Une section introductive (1.2 page 10) donne une présentation générale de ces modèles avec des exemples d'utilisation en bioinformatique structurale ; les modèles que nous avons plus particulièrement utilisés sont décrits et définis plus en détail dans la partie Méthodologie au chapitre 6 page 59. Enfin, nous avons repoussé en annexe B (page 177) un certain nombre de propriétés et de détails concernant ces modèles qui n'étaient pas d'une importance capitale pour notre travail, mais qui s'avèrent nécessaires pour une bonne compréhension de ces modèles.

Nous nous sommes plus particulièrement intéressés à la topographie de surface des molécules et à la détection des poches dans celles-ci. Les intérêts de ces deux problématiques sont succinctement présentés dans l'introduction générale (page 1), et repris plus en détail dans l'étude bibliographique, respectivement au chapitre 3 (page 25) et au chapitre 4 (page 37).

Les apports de notre travail sont à trouver :

- Au chapitre 7 de la partie méthodologie (page 75) où nous présentons la *surface duale*, une nouvelle construction que nous avons introduite pour faciliter le parcours sur la surface moléculaire.
- Dans les trois chapitres de la partie Résultats, où nous avons essentiellement (*i*) introduit la mesure de *courbure locale* dédiée à la caractérisation de la topographie de la surface moléculaire, (*ii*) appliqué nos formalismes et modèles à la détection des zones interagissantes à la surface des molécules, et (*iii*) proposé de nouveaux algorithmes pour la détection des poches dans les macromolécules.

D'une certaine manière, la synthèse sur les problématiques de la topographie et de la détection des poches, proposée dans la partie bibliographique constitue aussi un apport (a notre connaissance, rien de tel n'existe encore dans la littérature) ; de même que la présentation pratique des modèles issus de la théorie des formes-α que nous faisons au chapitre (6) de la partie méthodologie.

Introduction générale

Le travail présenté dans ce document s'inscrit dans le cadre de la **bioinformatique structurale** (en anglais *structural bioinformatics*, ou *computational biology*), une discipline scientifique dont l'objet est **l'analyse du lien existant entre la structure**[1] **et la fonction des macromolécules biologiques**, et dont les moyens sont informatiques, mathématiques, et très souvent géométriques. Les motivations de notre étude sont ainsi issues de la biologie, et les méthodes que nous avons proposées reposent sur des principes et des outils de géométrie algorithmique. La juxtaposition de la biologie et de la géométrie peut paraître surprenante au premier abord, elle trouve cependant une justification immédiate dans le cadre de la modélisation des objets étudiés : les macromolécules biologiques.

La fonction des macromolécules biologiques (protéines, ADN, ARN, ...) repose essentiellement sur leurs **interactions**, qu'on discrimine habituellement suivant la taille des intervenants. On distingue ainsi les interactions entre macromolécules de celles mettant en jeu une macromolécule et une molécule plus petite, un ligand. Dans ce dernier cas, le substrat de petite taille vient généralement se nicher dans des zones anfractuées à la surface de la macromolécule, et parfois même totalement encloses dans son intérieur. Ces espaces particuliers pouvant accueillir une petite molécule sont communément désignés par le terme de "**poche**". L'interaction de plusieurs macromolécules implique généralement une "**interface**", une zone souvent plus "plate" et plus large à la surface de chacune des molécules impliquées. La figure 1 met en évidence ces deux types de sites dans la structure d'un hétérodimère de récepteurs nucléaires[2].

Dans les deux cas (poches et interfaces), ces interactions reposent sur la présence de motifs **complémentaires** à la surface des intervenants, un aspect classiquement désigné comme le **modèle clef-serrure** [Fischer 94]. Dans ses travaux sur l'évaluation de la complémentarité des sous-unités $\alpha 1$ et $\beta 1$ de l'hémoglobine, M.L Connolly remarque que cette complémentarité repose sur deux aspects distincts, la complémentarité "physico-chimique" et la complémentarité de "forme" ; deux aspects qui peuvent être décorellés pour être traités séparément [Connolly 86c]. En raison essentiellement des empêchements stériques[3], la géométrie apparaît comme plus discriminante et est ainsi souvent considérée comme un premier filtre, une première étape pour l'étude de la structure des macromolécules.

[1] Le terme de structure désigne l'agencement d'une ou plusieurs molécules dans l'espace, c'est-à-dire la donnée des coordonnées spatiales des atomes composant ces molécules ainsi que la nature de ces atomes.

[2] Ce même exemple est repris et détaillé en annexe A.4 page 174. Le lecteur non averti trouvera dans la même annexe quelques bases concernant la structure des protéines, nécessaires à la compréhension de ce mémoire. Il y trouvera en particulier la description des modèles usuellement utilisés pour représenter les molécules et dont nous faisons usage dans nos illustrations moléculaires.

[3] En chimie, le qualificatif stérique désigne la position relative des atomes, et plus spécifiquement leur occupation spatiale. Par extension, les empêchements stériques sont relatifs à l'impossibilité de superposer l'espace alloué à deux atomes de deux molécules distinctes.

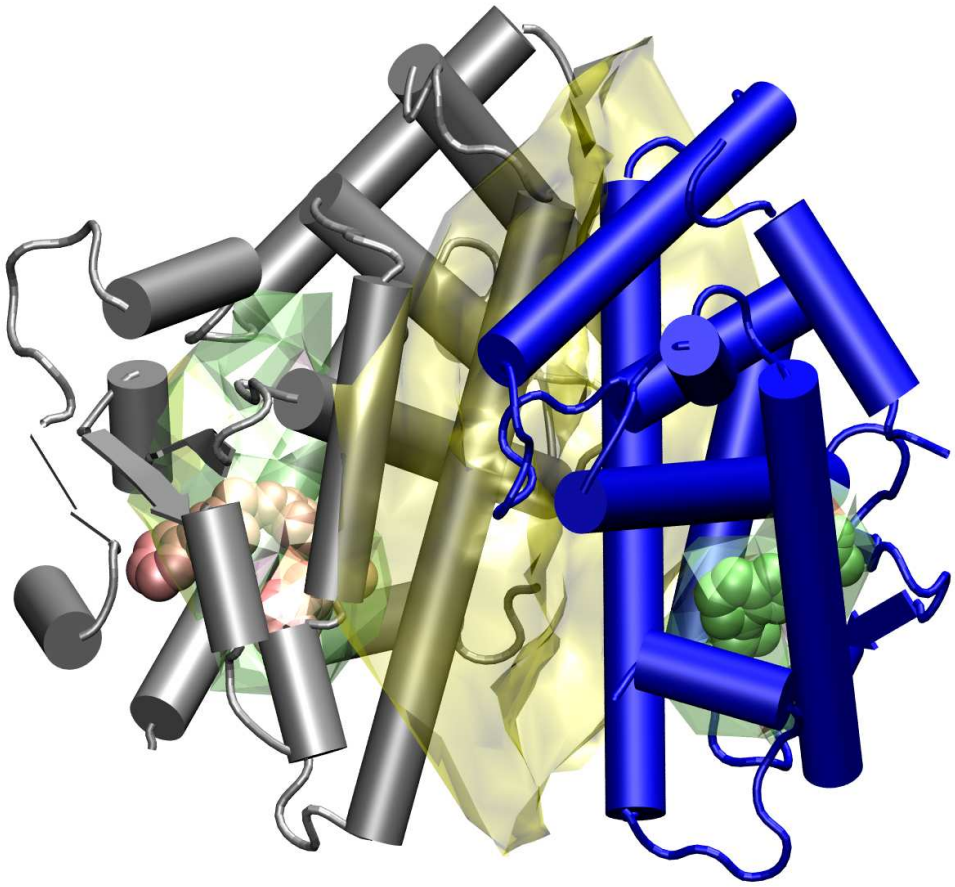


Figure 1: Exemple de poches et d'interfaces dans un complexe d'hétérodimère des domaines de liaison au ligand des deux récepteurs nucléaires PPARγ (en blanc) et RXRα (en bleu) (1RDT). Les deux récepteurs nucléaires sont représentés en modèles *cartoon*, leurs ligands respectifs en représentation de Van der Waals. L'interface (symbolisée par la surface mitoyenne en jaune transparent) a été calculée et affichée à l'aide d'*intersurf* [Ray 05] ; les deux poches de fixation du ligand (affichées en vert transparent) ont été calculées avec notre logiciel *Pck*.

De par leur lien avec les modèles moléculaires usuels[4], les structures issues de la géométrie algorithmique, et plus particulièrement celles issues de la théorie des formes-α[5] constituent un cadre adapté à la représentation des molécules, et à la caractérisation de leur "forme".

Les travaux présentés dans ce document concernent la caractérisation du paysage à la surface macromoléculaire ; plus spécifiquement nous avons différencié et traité les deux problématiques de la caractérisation topographique de la surface moléculaire en terme de "creux" et de "bosses", et de la recherche plus spécifique des "poches". Nous présenterons encore nos résultats dans le domaine de la prédiction des zones interagissantes sur la surface des protéines.

Le document est architecturé de la manière suivante.

Partie I - *Problématique* : nous présenterons ici le cadre de nos travaux au travers d'une étude bibliographique. Nous y présenterons les principaux modèles géométriques utilisés pour étudier les macromolécules biologiques, les différentes manières de définir la surface d'une macromolécule biologique, et d'en caractériser la topographie. Nous verrons encore

[4] Un rappel rapide sur le modèle de Van der Waals, le modèle Surface Accessible et le diagramme à remplissage de forme est présenté dans la première partie page 7.

[5] Cette théorie sera introduite dans la première partie (page 10), et précisée dans la seconde (page 59).

diverses motivations pour la détection des poches, proposerons des définitions géométriques pour préciser le sens de ce terme élusif, et étudierons les moyens mis en oeuvre par d'autres auteurs pour les détecter. Enfin, nous achèverons cette partie par une présentation détaillée des objectifs de notre travail.

Partie II - *Modèles géométriques* : nous présenterons ici plus formellement les modèles issus du domaine de la géométrie algorithmique en nous restreignant à ceux que nous avons effectivement utilisés dans notre étude. Nous introduirons la *surface duale*, un nouveau modèle que nous avons défini afin de faciliter le parcours à la surface des modèles moléculaires, et que nous avons utilisé dans nos autres développements pour la constitution de voisinages de résidus ou d'atomes à la surface de la molécule, ainsi que pour la définition de la courbure locale. Nous proposerons un algorithme pour la construction de cette surface polyédrique.

Partie III - *Applications et résultats* : dans cette partie nous introduirons la *courbure locale*, un nouvel indice attribué à chaque atome de surface d'une molécule et permettant d'en caractériser la topographie. Nous évaluerons l'intérêt de notre formalisme dans le cadre spécifique de la prédiction de résidus interagissants à la surface des molécules biologiques. Nous présenterons enfin notre contribution dans le domaine de la détection des poches.

Première partie

Problématique

Chapitre 1

Modèles utilisés en bioinformatique structurale

Ce chapitre est dévolu à la présentation de modèles couramment utilisés dans le cadre d'une étude en bioinformatique structurale, modèles sur lesquels nos propres travaux reposent. La première section concerne la présentation des modèles utilisés pour représenter les macromolécules biologiques. La seconde section sera consacrée à la présentation des principaux modèles géométriques utilisés pour la caractérisation et l'étude des macromolécules, et au rappel de leur utilisation dans des travaux antérieurs.

1.1 Modélisation géométrique des molécules

À certains égards, les modèles de Van der Waals, Surface Accessible et la surface de Connolly, étant définis à partir de boules, peuvent être considérés comme des modèles géométriques. La raison d'être de ces modèles est le besoin de définir l' "espace de la molécule", une zone occupée par les atomes de la molécule et inaccessible à tout autre atome extérieur à la molécule.

Ces modèles s'appuient sur une approximation : la modélisation de l'espace d'un atome par une boule centrée sur les coordonnées de l'atome (la position du noyau atomique dans l'espace) et à laquelle on attribue un rayon, généralement son rayon de Van der Waals. Cette approximation s'avère erronée dans les faits, Bondi [Bondi 64] par exemple remarque que la forme sphérique n'est en général pas physiquement fondée, mais qu'elle constitue une approximation suffisamment fiable[1] tout en étant pratique à manipuler. Le choix des rayons de Van der Waals à attribuer aux divers atomes constitue également une difficulté, et de nombreuses valeurs ont été proposées [Bondi 64, Lee 71, Richards 74, Tsai 02]. Ces différents ensembles de valeurs dépendent des approches utilisées pour réaliser les estimations, ainsi que du choix des structures sur lesquelles ont été mesurées les distances de contact entre atomes ou entre groupements. Le choix des rayons atomiques n'influe pas énormément sur la forme générale de la molécule ; il peut cependant induire des changements topologiques (comme illustré dans la figure 1.1), et influence les résultats volumétriques calculés sur les molécules [Tsai 02]. Une autre justification de la pertinence de ces modèles a été avancée par Richards [Richards 77], qui remarque que la pente drastique de la composante de répulsion dans l'équation de Lennard-Jones à elle-seule justifie l'épithète "solide" dans l'expression "modèle à sphères solides" couramment utilisée pour désigner ce type de modèles. Dans la suite du document nous amalgamerons souvent un atome a_i, ses coordonnées dans l'espace (x_i, y_i, z_i) et la boule $b_i = (a_i, r_i)$ qui le représente. On parlera ainsi indistinctement

[1] Bondi la caractérise plutôt "d'utile"

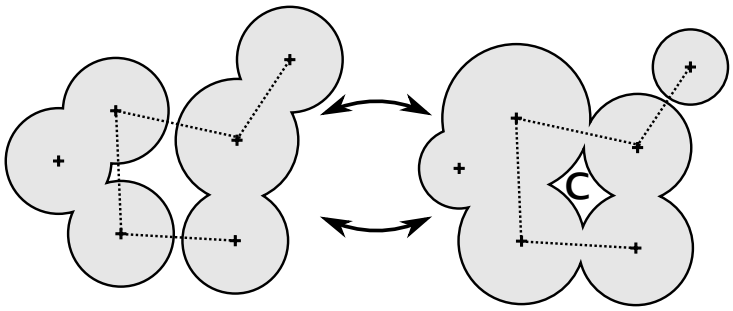


Figure 1.1: Influence des rayons atomiques sur les modèles moléculaires, un exemple en deux dimensions. Le même ensemble de six centres atomiques, considéré avec des rayons différents, induit des différences de connectivité entre les atomes, mis en évidence par des segments pointillés dans la figure. Ces variations induisent même des variations de topologie dans la molécule : la cavité c ne peut être observée que dans la figure de droite.

d'atome, de boule, ou de boule-atome ; nous parlerons plus spécifiquement de centre atomique pour désigner le centre de la boule.

1.1.1 Modèle de Van der Waals

Le *modèle de Van der Waals* (`VdW`) définit la molécule comme l'ensemble de ses atomes-boules, ou, lorsqu'on a besoin de matérialiser un volume dans l'espace, comme leur union. Le bord de cette union définit la *surface de Van der Waals*. S'il modélise effectivement l'espace propre à la molécule, ce modèle s'avère dans les faits assez peu utile pour identifier les espaces réellement interdits ou accessibles à une autre molécule ou à un solvant. Cet état de fait peut être observé dans la figure 1.2, où tous les espaces hors des boules grises ne sont en effet pas accessibles à la boule s modélisant une molécule de solvant. Pour cette raison, ce modèle peut être qualifié de lacunaire et se prête mal par exemple aux algorithmes de détection de poches (Kleywegt [Kleywegt 94] parle de *can of worms* pour désigner cet aspect).

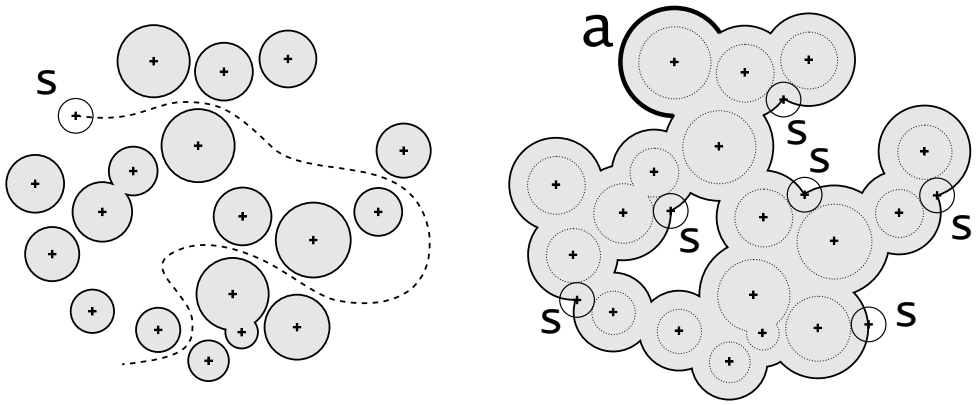


Figure 1.2: Modèle de Van der Waals et modèle Surface Accessible d'une molécule, un exemple en deux dimensions. Dans le modèle de Van der Waals (à gauche), les atomes sont représentés par des boules. Ce modèle contient de nombreuses lacunarités dans lesquelles une sphère-solvant (symbolysée par la sphère s) ne peut se tenir. Dans cette illustration, le trajet indiqué en traits discontinus est impossible à de nombreux endroits. À droite, la surface du modèle Surface Accessible est matérialisée par le centre d'une sphère solvant s en circulation sur la surface de Van der Waals. L'accessibilité d'un atome est mesurée comme l'aire a de sa contribution à la surface.

1.1.2 Modèle Surface Accessible

En 1971, Lee et Richards [Lee 71] ont introduit le modèle Surface Accessible (`SA`) pour représenter plus exactement l'espace accessible à une sphère de la taille d'une molécule de solvant. Ils ont utilisé ce modèle pour détecter les cavités dans les protéines et estimer leur accessibilité au solvant. Comme indiqué dans la figure 1.2, la Surface Accessible (ou parfois aussi *surface accessible au centre*) est définie comme le lieu parcouru par le centre d'une *sphère-test* (ou *sphère-solvant*) qui se déplacerait en restant en contact avec le modèle de Van der Waals. Il peut être alternativement défini en augmentant simultanément le rayon de tous les atomes de la molécule dans le modèle de Van der Waals d'une même constante r_p, égale au rayon de la sphère-solvant. Les invaginations observées dans le modèle Van der Waals sont ainsi closes, et les cavités sont naturellement définies comme des lieux emprisonnant une sphère-test, l'empêchant de rejoindre l'espace du solvant. L'accessibilité d'un atome est mesurée comme la portion d'aire de cet atome qui participe à la surface de l'union de boules ainsi définie. Le modèle Surface Accessible est défini comme la composante finie délimitée par la Surface Accessible (en grisé dans la figure).

1.1.3 Surface de Connolly

Le modèle Surface Accessible que nous venons de voir décrit les zones accessibles au *centre* d'une sphère de solvant virtuel, il ne donne donc pas exactement une vision de l'espace réellement accessible à cette sphère-solvant, ou inversement, de l'espace qui ne lui est pas accessible, c'est-à-dire de l'espace de la molécule. La *surface moléculaire* [Richards 77] — appelée aussi *surface de Connolly* du nom de M.L. Connolly qui a été le premier à en donner une formulation analytique [Connolly 83] ainsi qu'un algorithme pour la construire — adresse plus exactement cette problématique. La figure 1.3 explicite la définition de cette surface à partir d'une sphère-solvant roulant librement en contact avec les atomes du modèle Van der Waals.

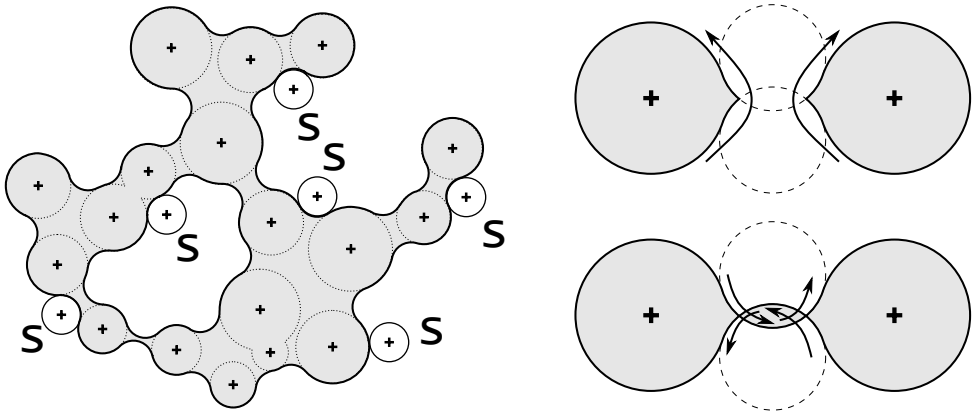


Figure 1.3: Modèle de Connolly d'une molécule, un exemple en deux dimensions. À gauche, le modèle de Connolly (en gris) est superposé au modèle de Van der Waals (surface en pointillés). Des sphères-solvant en contact avec cet dernier sont matérialisées par des disques blancs labellés s. À droite, dans le cas de l'auto-intersection d'une sphère-solvant (cercles en traits interrompus), deux choix sont possibles pour la surface de Connolly : en haut, la surface est déconnectée et deux pointes apparaissent ; en bas la surface s'auto-intersecte.

Ce troisième modèle convient généralement mieux à la définition de l'espace d'une molécule, il est aussi pratiquement "lisse", à l'inverse des trois précédents qui présentent des discontinuités de tangente aux endroits où deux atomes de la molécule se rencontrent. La figure 1.3 illustre en deux dimensions le phénomène qui induit des angles saillants sur la surface moléculaire : lorsque

deux sphères-solvant ont la possibilité de s'intersecter en deux endroits distincts de la molécule. En trois dimensions, ces cas de figure peuvent apparaître pour des sphères solvant en contact avec trois atomes de la molécule, générant un bord circulaire saillant (ou *cusp*), ou pour une sphère solvant en rotation libre autour de deux atomes, générant une pointe (ou *spindle*) à la surface de chacun des deux atomes.

1.1.4 Modèle à remplissage de forme

En complément des modèles à sphère solide (*hard sphere models*) que sont les modèles de Van der Waals et Surface Accessible, on a souvent besoin d'attribuer spécifiquement une partie de l'union des boules à chaque atome de la molécule. Ce découpage est généralement réalisé avec un modèle à remplissage de forme (*spacefilling diagram*) comme il sera précisé au chapitre 6.1.1 page 60. Un tel découpage peut être observé dans la figure 1.4.

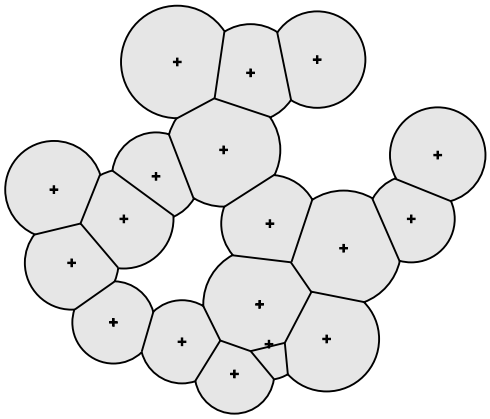


Figure 1.4: Diagramme à remplissage de forme d'une molécule, un exemple en deux dimensions. La molécule est partitionnée : chaque atome est représenté par une sphère tronquée.

1.1.5 Autres modèles moléculaires

D'autres modèles moléculaires existent, basés par exemple sur une approximation du nuage électronique à l'aide de fonctions gaussiennes, ils seront succinctement abordés dans la section suivante traitant de la modélisation d'une surface moléculaire, page 17. Pour la visualisation des molécules, des modèles particuliers ont été développés qui mettent par exemple en valeur les spécificités du type de la (macro)molécule observée. Les illustrations présentées dans ce mémoire font par exemple usage des modèles *cartoon* et fil de fer qui sont présentés en annexe A.3 page 169 à l'intention du non bioinformaticien.

1.2 Modèles issus de la géométrie algorithmique

Les modèles que nous présenterons dans cette sous-section ne sont pas à proprement parler des modèles moléculaires ; ils ne représentent pas les molécules, mais en donnent un encodage particulier, éclairant certains de leurs aspects. Le diagramme de Voronoï adresse la problématique de zone d'appartenance autour d'un atome, la triangulation de Delaunay celle de connectivité d'un atome, de voisinage entre les atomes. Le complexe dual raffine cette notion en ajoutant un critère de distance à ce voisinage.

Historiquement, le diagramme de Voronoï est le premier modèle issu de la géométrie algorithmique à avoir été appliqué à une étude en bioinformatique structurale. Ses applications sont de plus toujours nombreuses et il entretient un rapport étroit avec les deux autres modèles sur lesquels notre travail s'appuie. Pour toutes ces raisons, bien que nous ne l'ayons pas utilisé dans notre étude, nous en donnerons tout de même une description dans cette sous-section. Dans la suite de cette sous-section, nous présenterons succinctement chacun de ces trois modèles et mentionnerons quelques unes des études en bioinformatique structurale dans lesquelles ils ont trouvé une utilisation. Nous achèverons cette revue par un commentaire général sur l'intérêt de ces modèles dans le cadre d'une telle étude.

1.2.1 Diagramme de Voronoï

Étant donné un ensemble de points ou d'atomes a_i, le diagramme de Voronoï $\mathcal{V}$ constitue un découpage de l'espace en cellules de proximités (appelées aussi *cellules de Voronoï*) : à chaque atome a_i on attribue la cellule $\mathcal{V}_{a_i}$ des points de l'espace qui sont plus proches de a_i que de tous les autres atomes de la molécule. On attribue généralement la paternité de ce diagramme au mathématicien russe G.F. Voronoï [Voronoi 07, Voronoi 08, Voronoi 09], quoique des sources antérieures, telles que Dirichlet, Thiesen ou E. Laguerre soient aussi parfois avancées [Aurenhammer 91]. Un analogue "pondéré" de ce diagramme existe, qui tient compte du rayon des atomes ; un exemple en deux dimensions est visible dans la figure 1.5, et une définition plus rigoureuse pourra être trouvée en annexe B.1 page 177.

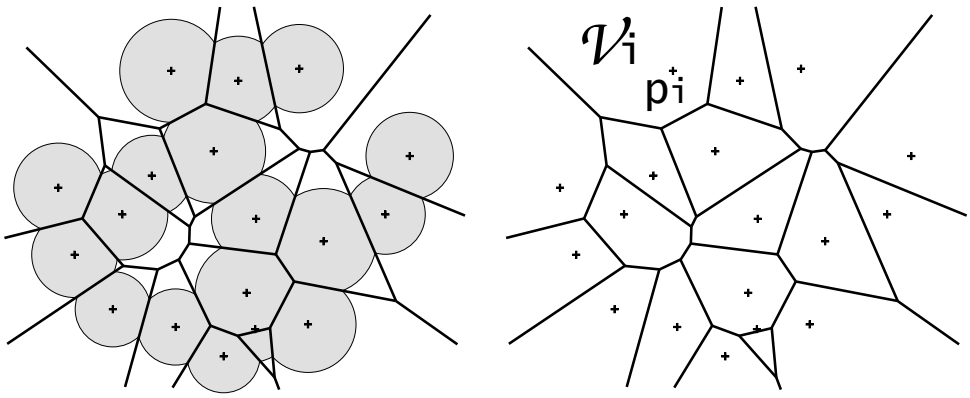


Figure 1.5: Diagramme de Voronoï d'une molécule, un exemple en deux dimensions. Le diagramme de Voronoï superposé à sa molécule dans la figure de gauche est isolé pour une meilleure appréhension visuelle dans la figure de droite. Le centre atomique p_i se voit attribuer la cellule $\mathcal{V}_i$ des points de l'espace "plus proches" de p_i que de tous les autres centres atomiques de la molécule.

Les diagrammes de Voronoï ont été intensivement utilisés en bioinformatique structurale, adaptés à des applications aussi variées que l'estimation de la densité atomique [Richards 74, Richards 77, Gellatly 82, Pontius 96, Tsai 99, Goede 97, Rother 03], le calcul volumétrique des poches dans les macromolécules [Chakravarty 02, Bhinge 04], la squelétisation des tunnels dans les structures tertiaires ou quaternaires pour en étudier le profil [Petřek 07], la détection des domaines d'une macromolécule [Wernisch 99], ou la matérialisation et l'étude des interfaces entre sous-unités dans un assemblage moléculaire [Ban 04, Cazals 06, Ban 06, Bernauer 08, J. Bernauer 05].

1.2.2 Triangulation de Delaunay

La triangulation de Delaunay d'une molécule (ou d'un ensemble de boules) se définit par dualité à partir de son diagramme de Voronoï ; elle décompose l'espace de la molécule (ou plus exactement, l'enveloppe convexe des centres atomiques de la molécule) en une union disjointe de tétraèdres dont les sommets sont des centres d'atomes de la molécule. Dans l'exemple de la figure 1.6, l'enveloppe convexe des centres atomiques est partitionnée par des triangles, l'équivalent en deux dimensions des tétraèdres. Cette construction porte le nom du mathématicien russe B. Delaunay qui a été le premier à en donner une définition mathématique et à en étudier les propriétés [Delaunay 34]. La donnée du premier algorithme pour la construire en trois dimensions [2] est bien plus tardive [Edelsbrunner 96].

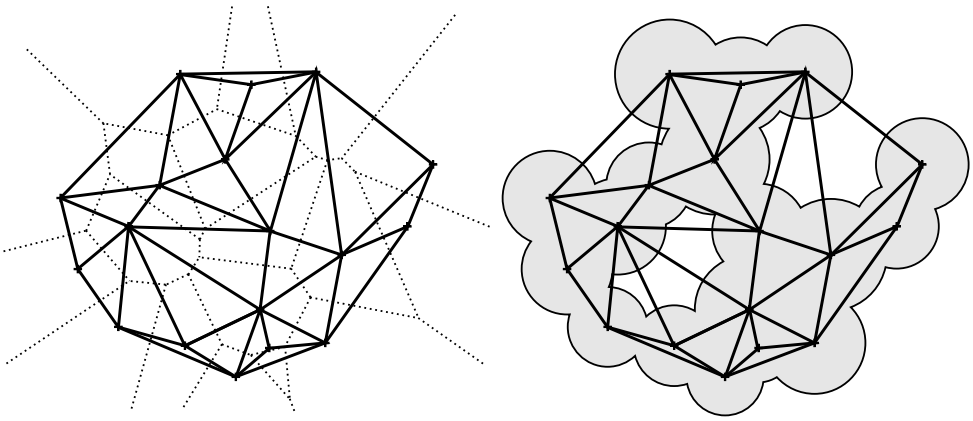


Figure 1.6: Triangulation de Delaunay d'une molécule, un exemple en deux dimensions. La triangulation de Delaunay construite à partir du diagramme de Voronoï dans la figure de gauche est superposée à sa molécule dans la figure de droite.

La triangulation de Delaunay est définie plus en détail dans la partie méthodologique aux chapitres 6.1.4 et 6.1.5 pages 66 et 69, et une section en annexe B présente certaines de ses propriétés les plus intéressantes pour une utilisation en bioinformatique structurale. Parmi ces propriétés, la définition intuitive d'un voisinage directionnel (explicitée à l'annexe B.4 page 183) justifie souvent à elle seule l'utilisation de ce modèle en bioinformatique structurale. L'unicité de la triangulation de Delaunay pour un ensemble d'atomes donné est aussi une propriété appréciable pour garantir la reproductibilité d'un calcul ou d'un algorithme. Un inconvénient de cette consrtuction réside dans son instabilité : de petites variations de position ou de rayon des atomes de la molécule peuvent induire une connectivité différente dans la triangulation.

Le graphe défini par les arêtes de la triangulation de Delaunay a par exemple été utilisé comme support pour détecter des domaines dans la structures des protéines [Taylor 06] ou pour matérialiser l'interface entre deux sous-unités [Ray 05]. La forme des tétraèdres et la composition de leurs sommets en terme de résidus ou d'atomes ont aussi été utilisées pour définir des scores probabilistes destinés à estimer la "qualité" [3] d'une structure [Singh 96, Munson 97, Vaisman 98, Krishnamoorthy 03].

[2] L'algorithme proposé par Edelsbrunner adresse en fait la construction de cette triangulation en dimension quelconque.

[3] De tels scores sont par exemple utilisés en dynamique moléculaire pour conserver des structures le plus "naturelles" possibles à chaque pas de simulation.

1.2.3 Complexe dual, forme duale, complexe-α, et forme-α

Un autre formalisme en relation avec les deux modèles précédents, bien qu'introduit plus récemment, a déjà fait l'objet de nombreuses applications en bioinformatique structurale ; souvent désigné sous l'appellation générale de *théorie des formes-α*, il regroupe quatre objets distincts : le complexe dual, la forme duale, le complexe-α, et la forme-α. Le complexe dual et le complexe-α [Edelsbrunner 94b, Edelsbrunner 92] sont obtenus à partir de la triangulation de Delaunay en filtrant les éléments la constituant (sommets, arêtes...) sur des critères de "taille". Le complexe dual d'une molécule conserve uniquement les éléments de la triangulation de Delaunay dont toutes les boules situées aux sommets ont une intersection commune[4] (voir la figure 1.7) ; il conserve ainsi la forme et les caractéristiques topologiques de la molécule. Le complexe-α enregistre l'ensemble des complexes duaux d'une molécule en fonction d'un paramètre α contrôlant la taille des atomes de cette molécule ; il permet ainsi d'observer la forme de la molécule à différents niveaux de détail. La forme duale et la forme-α correspondent respectivement à l'espace occupé par les éléments du complexe dual et à l'évolution de cet espace en fonction du paramètre *alpha*. Nos travaux reposant sur ces constructions, on en trouvera une définition et une étude plus détaillée dans la partie méthodologie aux chapitres 6.1.2 page 61 et 6.1.3 page 64. Le bord

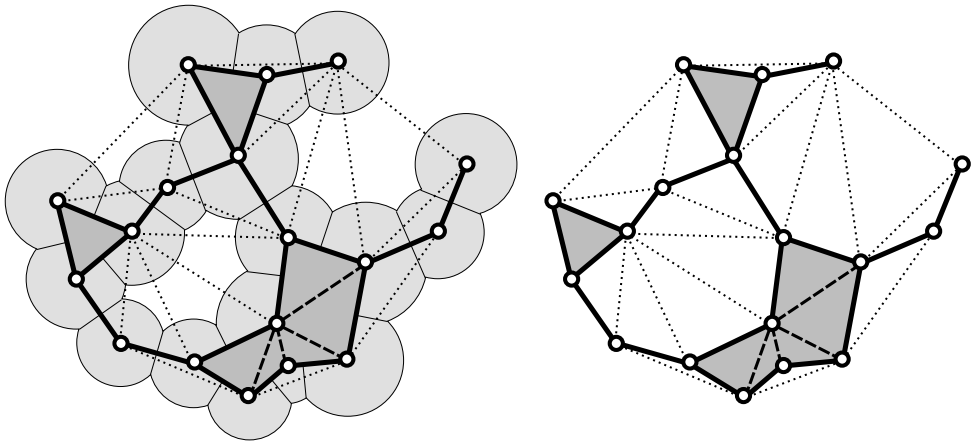


Figure 1.7: Complexe dual d'une molécule, un exemple en deux dimensions. À gauche, le complexe dual de la molécule (sommets blancs, arêtes épaisses et triangles gris) est superposé au diagramme à remplissage de forme de la molécule ; ses sommets sont des centres atomiques, ses arêtes correspondent à l'intersection de deux atomes, ses triangles à l'intersection de trois boules, et (en trois dimensions) ses tétraèdres à l'intersection de quatre boules. À droite, le complexe dual isolé permet de mieux appréhender la similitude de forme.

de ces constructions constituant un polyèdre, nous parlerons parfois de *modèle polyédrique* ou de *représentation polyédrique* d'une molécule pour désigner ces modèles.

Ces constructions, définies par H. Edelsbrunner au début des années 1990, étaient initialement destinées à l'étude de la "forme d'un nuage de point", par exemple à des visées de reconstruction d'objets à partir de données extraites d'un scanner numérique. Le rapport naturel avec les modèles biologiques dits en "sphère-solide" (*hard-sphere model*) exposés dans la partie précédente a été rapidement exposé [Edelsbrunner 95a], et exploité depuis dans de nombreuses applications. La théorie des forme-α a par exemple permis de proposer une nouvelle méthode de calcul, à la fois exacte et plus rapide, des volumes et des aires des modèles moléculaires [Edelsbrunner 95a, Attali 05], ainsi que des dérivées de ces valeurs en fonction de la position des atomes de la

[4] Pour être exact il faut considérer la restriction des boules à leur cellule de Voronoï, comme suggéré par les segments séparants les boules dans la figure 1.7. Cet aspect sera développé plus en détail au chapitre 6.1.2.

molécule [Edelsbrunner 03, Bryant 04]. La filtration naturelle des arêtes sur un critère de distance implicite a été utilisée pour constituer ou simplifier des potentiels destinés à l'estimation de la "qualité"[5] [Li 03, Zomorodian 06]. Le complexe dual a été employé à la détection de poches dans les macromolécules [Akkiraju 95, Peters 96, Edelsbrunner 98] (ces algorithmes seront plus amplement discutés dans la partie de l'état de l'art dédiée à l'étude de la détection des poches au chapitre 4.4 page 41), ou comme cadre pour la construction et la maintenance en temps réel de maillages de la surface de Connolly à base de NURBS[6] [Bajaj 03b, Bajaj 03a]. Le complexe dual a encore été employé pour modéliser la membrane lipidique des cellules et étudier leur mode de fusion [Kasson 07], ou comme représentation approximative des sous-unités en cryomicroscropie à des fins de comparaison et de caractérisation [De-Alarcon 02, Jimenez-Lozano 03]. La théorie des forme-α a également été employée à la caractérisation de sites d'interaction spécifiques à la surface des protéines [Tondel 02] et appliquée dans une procédure d'amarrage moléculaire [Tondel 06].

1.2.4 Intérêt des modèles issus de la géométrie algorithmique pour une analyse en bioinformatique structurale

L'emploi des modèles issus de la géométrie algorithmique dans le cadre d'une étude en bioinformatique structurale n'a cessé de croître depuis le début des années 1970, démontrant l'utilité de ces objets pour ce cadre d'application. Leur succès, en dépit du formalisme parfois ardu qui les caractérise, peut s'expliquer par :

(i) leur adéquation aux problématiques moléculaires ;

(ii) l'extrême simplification des modèles[7], dépouillés à l'extrême pour ne plus rendre apparentes que des propriétés simples (cellule de proximité, voisinage directionnel, intersection d'atomes...), et leur description géométrique propice aux définitions rigoureuses ainsi qu'à l'édification d'algorithmes simples et démontrables ;

(iii) le fait que ces modèles ont fait l'objet d'une attention soutenue de plusieurs communautés scientifiques, tant en mathématiques qu'en informatique, ayant abouti à une bonne connaissance théorique de ces modèles ainsi qu'à la donnée d'algorithmes pour les construire et les manipuler, et le fait qu'il existe maintenant des implémentations à la fois robustes et rapides de ces algorithmes.

Les approches basées sur des modèles issus de la géométrie algorithmique se caractérisent généralement par :

(i) Un grande rapidité d'exécution, inhérente pour une part à l'efficacité des algorithmes de construction des modèles, et d'autre part à la simplicité des modèles eux-mêmes, qui favorisent des algorithmes simples et rapides. Cet aspect est particulièrement appréciable pour les traitements en masse, par exemple pour un emploi en dynamique moléculaire, ou pour un traitement sur une base de structures.

(ii) Une reproductibilité des résultats, inhérente à l'unicité du modèle pour une molécule donnée. Cette qualité n'est pas toujours nécessaire, par exemple pour l'évaluation d'une donnée chiffrée, tant qu'on peut garantir une marge d'erreur (ou une différence) faible entre deux exécutions ; elle peut s'avérer plus problématique dans un contexte qualitatif (présence ou non d'un caractère recherché, comme par exemple l'existence ou non d'une

[5] Cet aspect a déjà été présenté dans la section précédente exposant les applications de la triangulation de Delaunay.

[6] Les ***N**on **U**niform **R**ational **B**eta **S**plines* permettent de donner une représentation paramétrique d'objets surfaciques, elles constituent une généralisation des splines ou des surfaces de Bézier.

[7] Si les théories sur lesquelles ils s'appuient peuvent être considérées comme compliquées, les objets, sont eux extrêmement simples et intuitifs.

bouche[8]). En particulier, ces approches offrent des garanties de reproductibilité au niveau topologique, ce que par exemple les approches discrètes ne permettent pas (à ce sujet, voir l'annexe E).

[8] c'est-à-dire l'ouverture d'une poche sur l'espace du solvant, comme il sera vu un peu plus loin dans la section 4 dédiée à l'étude des poches.

Chapitre 2

Définir la surface des macromolécules

La surface d'une macromolécule peut se définir comme la partie en contact avec l'extérieur de la molécule et libre d'interagir avec d'autres molécules. L'étude de cette zone particulière porte donc un éclairage sur le comportement de la molécule dans son environnement, sur sa fontion ; ce en quoi elle s'avère d'un intérêt fondamental. Mais la notion même de surface moléculaire reste floue et sujette à discussion, d'autant plus que le terme "surface" trouve une résonance intuitive dans l'esprit de chacun.

Intuitivement, le terme de "surface" fait référence à quelque chose de "continu", voire de "lisse" ; une limite physique interdisant un franchissement, comme la surface d'une table sur laquelle on peut poser une tasse sans craindre de la voir traverser cette dernière. C'est aussi la frontière entre deux milieux, auquel cas elle définit généralement un "intérieur" et un "extérieur" ; elle enclôt et délimite un "volume" dont elle devient "le bord", le lieu d'un contact possible avec un autre objet.

Dans le cas des macromolécules, ces notions intuitives font répons à une réalité physique : des énergies de répulsion interdisent l'interpénétration des atomes de deux molécules et autorisent la définition d'une surface limite, accessible pour le contact avec une autre molécule.

Une autre approche pour définir la surface d'une macromolécule, moins intuitive à première vue, consiste à dresser la liste des atomes (ou, dans le cas spécifique des protéines, la liste des résidus) qui constituent cette surface ; autrement dit il s'agit de la donnée ensembliste des éléments constitutifs de la molécule qui sont suceptibles d'intéragir avec "l'extérieur".

Il existe donc essentiellement deux manières de concevoir la surface d'une molécule : une approche continue, qui matérialise une limite et un volume dans l'espace, une enveloppe visualisable ; et une approche ensembliste basée sur la donnée de résidus ou d'atomes susceptibles d'intéragir. Dans la suite de ce chapitre nous détaillerons séparément ces deux interprétations, nous expliquerons plus avant les motivations qui mènent à l'une ou à l'autre de ces interprétations de la notion de surface, et donnerons des exemples commentés d'approches proposées dans la littérature pour les traiter.

2.1 Représentation continue de la surface

Le besoin d'une description continue peut trouver de nombreuses motivations.

L'utilisation d'une surface continue permet d'appréhender visuellement l'espace occupé par la molécule et en facilite l'analyse, par exemple avec la projection sur cette enveloppe de propriétés physico-chimiques telles que l'hydropathie, la charge, ou la nature des résidus sous la surface.

La matérialisation d'une enveloppe est aussi très souvent un prérequis pour la caractérisation de la forme et de la topographie d'une molécule, une problématique dont les moyens et l'intérêt seront discutés dans le chapitre suivant.

Enfin, la donnée d'une surface autorise la réalisation de calculs volumétriques, utiles par exemple pour évaluer des énergies de solvatation [Wesson 92] et modéliser le comportement d'une protéine dans un solvant pour étudier les mécanismes présidant à son repliement.

Il existe essentiellement trois approches pour définir une surface moléculaire continue, toutes basées sur la modélisation implicite ou explicite d'intéractions moléculaires, et plus spécifiquement axées sur des forces de répulsion.

2.1.1 Représentation du nuage électronique

Arguant de la répulsion induite par les charges négatives du nuage électroniques, de nombreuses approches [Duncan 93b, Grant 95, Laskowski 95, Gabdoulline 96, Maggiora 01] assimilent la surface de la molécule à son nuage electronique. Ces approches reposent très généralement sur la donnée d'une fonction d'energie dont un isocontour matérialise, modélise, ce nuage. Le cortège électronique autour de chaque noyau atomique est représenté par une fonction isotrope décroissante à mesure qu'on s'éloigne du noyau — généralement une expression exponentielle — et la fonction d'énergie associée à la molécule entière est définie comme un mélange — souvent une simple somme — de ces fonctions atomiques.

Ces procédés produisent une surface moléculaire complètement lisse, et donnent des définitions continuement dérivables de l'aire et du volume de la molécule [Grant 95], deux propriétés qu'on ne peut par exemple pas obtenir avec les modèles géométriques actuels.

La manipulation de fonctions exponentielles peut cependant s'avérer coûteuse et rendre malaisée la matérialisation de la surface. En particulier, le choix de la valeur pour l'isocontour est un premier choix arbitraire ; pour certaines isovaleurs on rencontre en outre des problèmes d'ambiguïté et d'auto-intersection de la surface moléculaire (voir figure 2.1). Enfin, ce type d'approche rend quasi obligé le recours à des techniques de discrétisation, dont les principes, qualités et limitations sont discutés en annexe E.

2.1.2 Représentation d'une surface d'interaction

Une autre approche pour définir la surface moléculaire consiste à évaluer les interactions que ferait une molécule-sonde avec la macromolécule, et à chercher les lieux favorables à une telle interaction [Jackson 02, An 04, Laurie 05]. Concrètement, à chaque position de l'espace on évalue la somme des énergies des intéractions non liées intervenant entre la sonde et les atomes de la molécule. Dans *DRUGSITE* [An 04] seules les énergie de Van der Waals sont considérées et représentées par un potentiel de Lennard-Jones modifié, dans *Q-SiteFinder* [Laurie 05] et *Q-fit* [Jackson 02] les liaisons hydrogène et les forces électrostatiques sont aussi considérées. La surface de la molécule est encore une fois définie comme l'isocontour d'une fonction d'énergie, mais cette fois ci la fonction implicite représente les zones les plus énergétiquement favorables à une intéraction. Encore une fois, elle tend vers zéro lorsqu'on s'éloigne de la molécule et elle est fortement positive à l'intérieur de la molécule, par contre, elle admet une série de minima locaux aux abords de la surface. Bien que relativement coûteuse en temps, et généralement abordée avec des méthodes discrètes, une telle définition est naturellement utilisée pour la recherche de sites de fixation d'un ligand.

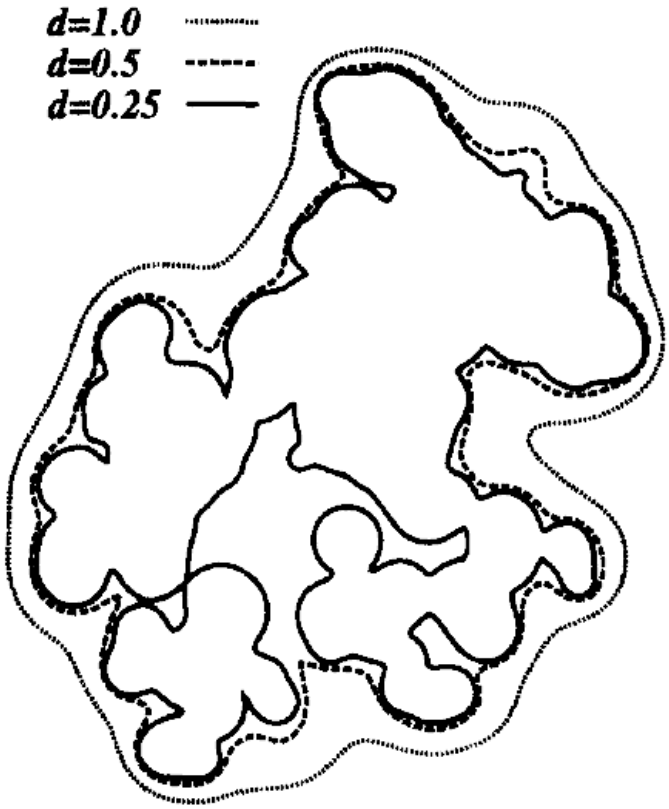


Figure 2.1: Représentation implicite de la surface moléculaire, figure extraite de la publication d'ADS [Gabdoulline 96] La figure montre trois isocontours obtenus pour trois isovaleurs distinctes sur un même mélange de noyaux gaussiens centrés sur les atomes d'une molécule. Pour des isovaleurs faibles la surface est plus détailée, mais elle peut aussi présenter des auto-intersections.

2.1.3 Représentation par des modèles géométriques

Une troisième approche consiste à modéliser la surface moléculaire à partir d'une définition purement géométrique. Nous avons vu les trois modèles moléculaires usuels que sont le modèle de Van der Waals, le modèle Surface Accessible et la surface de Connolly à la section précédente traitant des modèles moléculaires (page 7) ; leur validité et leurs limitations ont déjà été discutées, de même que leur relation avec les modèles issus de la géométrie algorithmique que sont la triangulation de Delaunay et les structures décrites dans la théorie des formes-α. Ces modélisations moléculaires géométriques constituent une approximation suffisante pour de nombreuses applications ; de par leur simplicité elles offrent en outre un cadre adapté à la description d'algorithmes. Des exemples de surfaces délimitées par ces modèles peuvent être observés dans la figure 2.2.

La surface de Connolly est souvent préférée aux deux autres car plus "lisse" ; elle constitue aussi une définition intuitivement plus "juste" de l'espace moléculaire en bouchant les invaginations du modèle de Van der Waals [Richards 77]. De nombreux travaux ont été menés pour représenter cette surface au travers d'un échantillonnage de points ou par un maillage triangulaire [Connolly 83, Connolly 85b, Varshney 93, Akkiraju 96, Totrov 96, Vorobjev 97, Sanner 95, Sanner 96, Bajaj 97, Laug 01, Laug 02, Laug 03], de même que pour quantifier l'aire et le volume délimités par cette surface [Richmond 84, Connolly 85a, Hayryan 05].

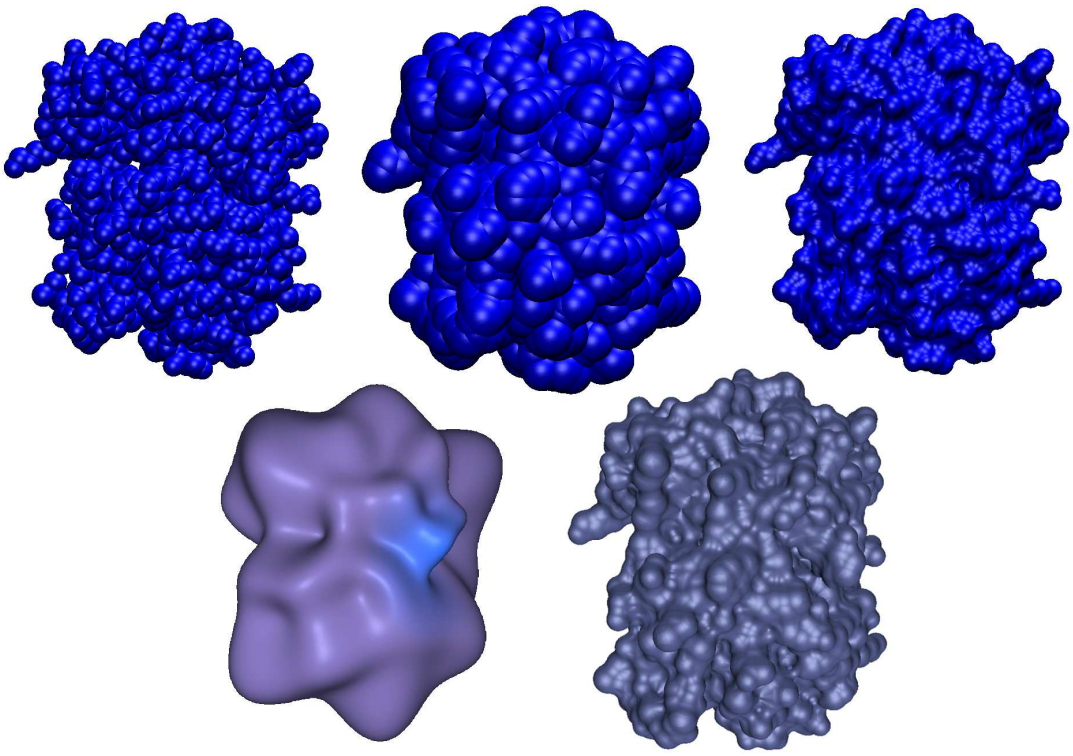
Il existe des définitions de surfaces moléculaires géométriques alternatives ; elles trouvent généralement leur motivation dans des propriétés ou des applications particulières, et leur pertinence dans une approximation "suffisamment bonne" de la surface de Connolly. Deux modèles surfaciques de ce type sont présentés dans la figure 2.2 où ils peuvent être comparés aux trois modèles précédents ; leur cadre théorique est succinctement explicité ci-après.

Les surfaces moléculaires basées sur une décomposition en harmoniques sphériques [Duncan 93a, Cai 98, Ritchie 99] produisent par exemple des définitions lisses et multi-échelles[1]. Ce type de description permet en outre une accélération des calculs de convolution, utiles par exemple pour évaluer rapidement une similarité [Ritchie 99] ou une complémentarité [Wensheng 02, Ritchie 00] de deux surfaces. La théorie des harmoniques sphériques constitue un équivalent de la théorie de Fourier en coordonnées sphériques ; elle autorise la description de fonctions ou de surfaces exprimables en coordonnées sphériques dans une base de fonctions dites harmoniques. De fait, ces descriptions ne sont applicables directement qu'aux convexes étoilés[2], mais peuvent être étendues à toutes les surfaces topologiquement équivalentes à une sphère au prix d'une première étape visant à donner un paramétrage sphérique de la surface [Duncan 93a, Cai 98] ; cependant, pour peu que la surface à représenter contienne au moins un tunnel, la représentation sera nécessairement erronnée. Le calcul de cette représentation en harmoniques sphériques est assez coûteux, surtout si l'on désire un niveau de détail élevé. Un des intérêts de cette description réside dans la possibilité d'évaluer rapidement une similarité ou une complémentarité pour toutes les conformations obtenues par rotation de l'une ou l'autre des deux molécules ; néanmoins, dès lors que l'on souhaite translater une des molécules il est nécessaire d'en recalculer la représentation en harmoniques sphériques.

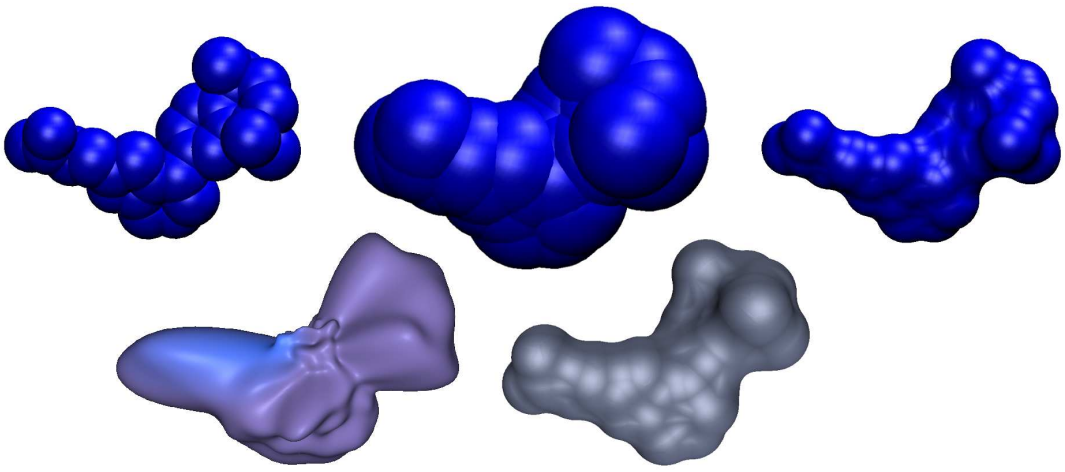
Les surfaces skins introduites plus récemment proposent une approximation basée sur la théorie des formes-α et une algèbre des sphères [Edelsbrunner 99] ; elles offrent un cadre théo-

[1] C'est-à-dire, avec un niveau de détail que l'utilisateur peut contrôler.

[2] Un convexe étoilé est un ensemble E dans lequel il existe au moins un point central c, directement relié à tous les points de E. Plus formellement, il existe c dans E tel que si x est dans E, alors le segment $[cx]$ est entièrement dans E.



(a) Le domaine de liaison au ligand du récepteur nucléaire à l'acide rétinoïque comprend 1622 atomes.



(b) sur son ligand composé de 28 atomes (en bas)

Figure 2.2: Cinq représentations de la surface moléculaire, une illustration sur le domaine de liaison au ligand d'un récepteur nucléaire à l'acide rétinoïque RXR-α (figure 2.2(a)), et sur son ligand (2.2(b))) (1RDT chaîne A). Pour chacune des deux molécules la ligne du haut présente les modèles géométriques, de gauche à droite : la surface de Van der Waals, la surface accessible et la surface de Connolly ; La ligne du bas présente les modèles alternatifs, toujours de gauche à droite : en harmonique sphérique (avec un niveau de résolution $L = 15$) et surface *skins* (avec deux niveaux de subdivisions pour la protéine, et trois pour le ligand.).

rique rigoureux pour la définition d'une surface moléculaire totalement lisse (au sens géométrique). Contrairement à la surface de Connolly, une surface skin ne présente aucun cas d'auto-intersection, elle est en outre topologiquement équivalente[3] à la surface accessible à laquelle elle est associée. Des algorithmes ont été proposés pour mailler ces surfaces [Cheng 01, Kruithof 07b] ou les représenter en lancer de rayons [Chavent 08], mais aucune expression analytique n'existe à ce jour pour le calcul du volume et de l'aire des surfaces skins.

2.2 Représentation ensembliste de la surface : détermination des résidus de surface

Lorsqu'il observe ou lorsqu'il modélise une interaction, plus qu'à une notion de surface continue, le bioinformaticien s'intéresse à la composition de cette surface en terme d'atomes, ou de groupement d'atomes, car c'est à ce niveau que se situent les modèles d'interaction qu'il manipule (liaisons hydrogène, ponts salins, forces hydrophobes,...). C'est en effet la nature des atomes, des groupements atomiques, et — dans le cas des protéines[4] — des résidus qui détermine le type et la force d'une interaction.

De manière plus générale, la donnée des résidus est d'une importance capitale : c'est la succession de ces résidus qui confère aux protéines leurs formes, leurs propriétés, leurs spécificités, et en définitive leurs fonctions ; la mutation d'un seul de ces résidus peut avoir une influence drastique sur la fonction d'une protéine, pouvant par exemple résulter dans une diminution de son activité, voire dans sa totale inhibition.

On oppose habituellement la surface et l'intérieur de la molécule ; aux résidus situés à l'intérieur (*résidus enfouis*) on attribue généralement un rôle "structurant" de la protéine, à ceux situés en bordure de la molécule (*résidus de surface*) on attribue un rôle dans les interactions avec les molécules de solvant ou d'autres types de molécules [Miller 87]. Dans le cas où cette interaction est avérée, par exemple lorsqu'elle est observée dans une structure, on parle même de *résidu interagissant*.

Les résidus de surface sont des résidus qui peuvent potentiellement interagir avec d'autres molécules, ou qui, en un sens, sont *accessibles* pour une intéraction avec une autre molécule. De nombreuses approches ont donc été développées pour proposer des définitions alternatives de la surface, elles relèvent chacune d'une interprétation du terme "accessibilité".

2.2.1 Résidus accessibles au solvant

Dans une première approche, Lee et Richards ont proposé de considérer comme faisant partie de la surface, les résidus qui seraient *exposés à l'espace du solvant* [Lee 71]. Les auteurs ont défini et évalué cette accessibilité comme la contribution d'un résidu à l'aire de la surface accessible au solvant ou ASA (pour *accessible surface area*), voir figure 2.3 A.

Miller *et al.* [Miller 87] ont modulé cette première définition en définissant l'*exposition relative* d'un résidu en terme de perte d'accessibilité par rapport à une valeur référence. Pour ce faire, ils calculent l'aire d'un résidu dans le modèle Surface Accessible de la molécule, et la comparent à une valeur calculée préalablement dans un modèle où le résidu est totalement exposé (concrètement, dans une chaîne Gly-Xxx-Gly[5]). Dans leur définition, un résidu est considéré accessible (et donc en surface) lorsque la perte de surface accessible au solvant (ASA) n'excède pas 95% de l'aire de référence.

[3] En particulier, les cavités et les tunnels se correspondent exactement dans les deux modèles.
[4] Une courte introduction à la structure des protéines pourra être trouvée en annexe A.
[5] La glycine (Gly) étant l'acide aminé le moins "encombrant".

2.2.2 Résidus accessibles à longue distance

Dans ses travaux Todd O. Yeates inverse cette première notion d'accessibilité : il ne s'intéresse pas à la contribution des atomes ou des résidus de la molécule, mais plutôt à la facilité qu'une molécule-sonde aura à les atteindre ; ce qu'il appelle l'*accessibilité à longue distance* (voir la figure 2.3 *B*) [Yeates 95]. Le *rayon maximal de contact* est défini par le rayon de la plus grande sphère pouvant entrer en contact avec l'atome considéré sans pénetrer le modèle de Van der Waals de la molécule. C'est aussi un indice sur la taille maximale de la molécule avec laquelle la protéine est susceptible d'interagir en un point de sa surface. L'*accessibilité à la diffusion*, mesure la propension d'un atome à entrer en collision avec une particule brownienne. Elle est calculée par des marches aléatoires : pour chaque atome de la molécule on décompte le nombre de marche dans lesquelles l'atome a été le premier collisionné. Les algorithmes proposés par l'auteur impliquent des temps de calcul élevés.

2.2.3 Résidus accessibles sous la surface

Dans les deux approches présentées jusqu'à maintenant, seuls les atomes contribuant à la Surface Accessible sont pris en compte. Ainsi, un résidu dont tous les atomes seraient enfouis sous cette surface ne sera jamais considéré comme en faisant partie ; or, dans un contexte dynamique, de tels résidus peuvent venir à s'exposer en surface et à interagir avec d'autres molécules. Arguant de ce que ces approches ne tiennent pas compte des aspects flexibles des molécules, des méthodes alternatives ont été proposées qui permettent de définir comme contribuant potentiellement à la surface, des résidus enfouis dans le modèle Surface Accessible. Ces techniques reposent généralement sur un critère de *distance à la surface* ; les définitions différant alors essentiellement sur la "matérialisation" de cette surface (voir figure 2.3 *C*). Dans les travaux de Pedersen *et al.* [Pedersen 91], comme dans ceux de Chakravarty *et al.* [Chakravarty 99], cette surface est matérialisée par des molécules d'eau réparties "virtuellement" autour de la molécule avec de coûteuses simulations. Dans *DPX* [Pintar 03b, Pintar 03a], c'est la plus petite distance au centre d'un des atomes contribuant à la Surface Accessible qui est calculée avec une approche à la fois simple et rapide. Chakravarty *et al.* [Chakravarty 99] ont aussi proposé et testé l'utilisation de points référents répartis sur la surface de Connolly. Ils ont conclu à l'inadaptation de ces valeurs pour leurs applications, essentiellement parce que dans les faits, les molécules d'eau ne peuvent occuper simultanément toutes les positions en contact de la molécule ; par exemple, on ne pourra avoir simultanément les référents w_1 et w_2 de la figure 2.3 *C*.

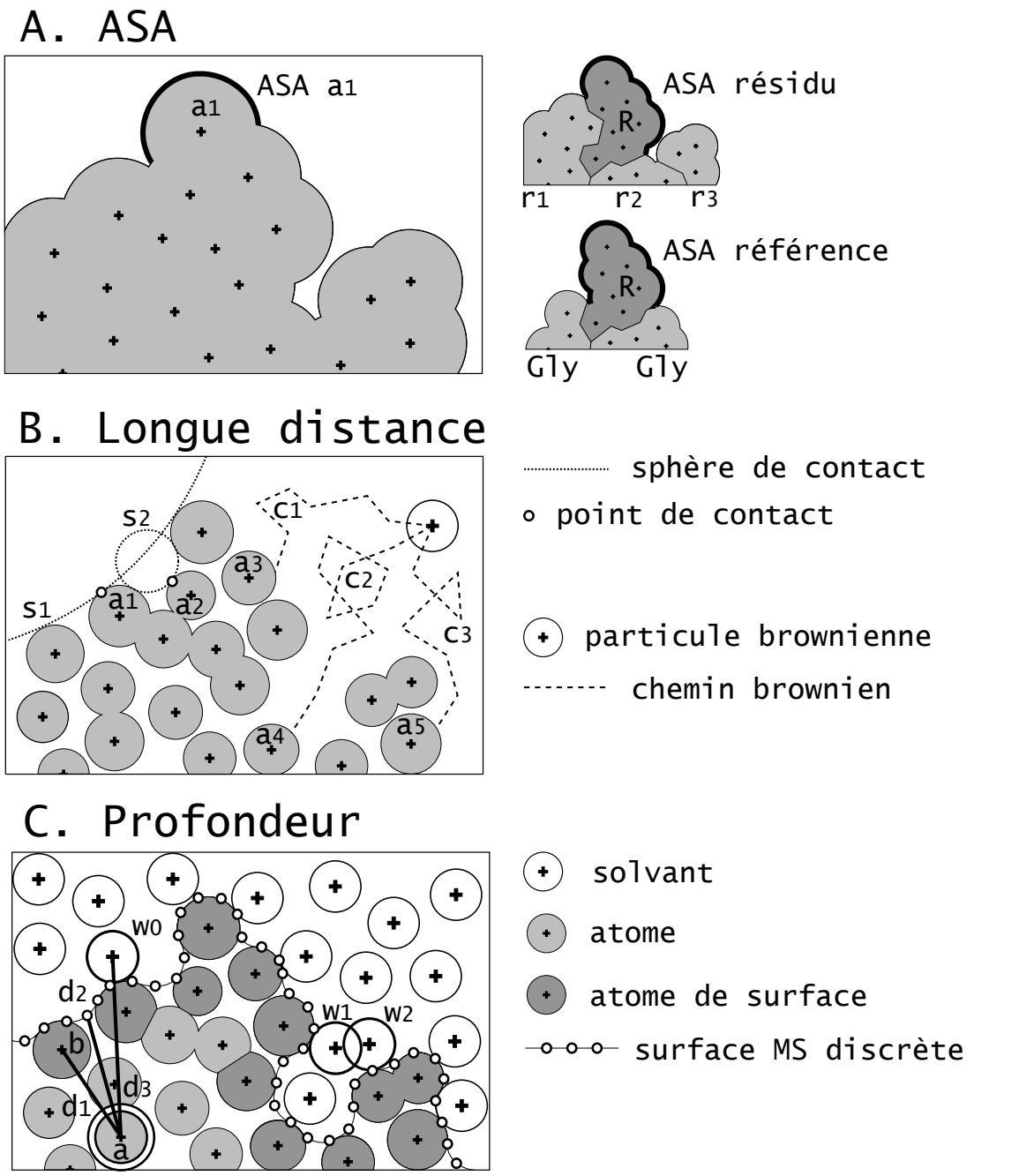


Figure 2.3: Définir les résidus à la surface de la molécule : différentes définitions de l'accessibilité.
A. L'ASA de l'atome a_1 est la contribution de cet atome à l'aire de la molécule (en trait gras) dans le modèle Surface Accessible. L'ASA du résidu en gris foncé est la somme des ASA de ses atomes, et son ASA relative est le rapport entre l'ASA du résidu dans une chaîne Gly-Xxx-Gly, et son ASA dans la protéine.
B. Accessibilité à longue distance : le rayon maximal de contact des atomes a_1 et a_2 sont respectivement les rayons de s_1 et s_2. L'accessibilité à la diffusion des atomes a_3, a_4 ou a_5 comptabilise le nombre de chemins aléatoires (c_1, c_2 et c_3) qui finissent leur course en impactant l'atome.
C. La profondeur d'un atome ou d'un résidu a est calculée par rapport à un référent de la surface de la molécule. Il peut s'agir de la plus courte distance d_1 à un atome de surface b (*i.e* un atome d'ASA strictement positive), de la plus courte distance d_2 à un point sur une discrétisation de la surface de Connolly ou de la plus courte distance d_3 à une molécule de solvant w_0 placée par optimisation.

Chapitre 3

Topographier la surface des macromolécules

Dans le langage courant, le terme de *topographie* désigne la problématique consistant à dresser la carte d'un lieu, bien souvent sur des critères de "hauteur". Une telle représentation appelle la caractérisation de la surface en terme de propriétés topographiques telles que les "creux" ou les "bosses". Cette intuition "d'élévation" s'appuie essentiellement sur l'expérience de la topographie terrestre, et plus particulièrement sur l'impression que nous avons localement d'une référence plane : le niveau de la mer, ou le niveau d'un sol idéalement plat (figure 3.1). En supposant la terre dans sa forme sphérique, une approche similaire consisterait à remplacer la "hauteur" par une "distance au centre de la terre". Dans le cas d'objets plus complexes, tels que la surface des macromolécules biologiques, une approche de ce type est beaucoup moins aisée. Dans ce document, le terme de *topographie* sera employé pour désigner la problématique consistant à caractériser la forme d'un objet en terme de creux et de bosses. Nous verrons en quoi cette problématique est importante dans le cadre de l'analyse structurale d'une macromolécule, ainsi que les définitions et les solutions qui ont été apportées jusqu'à présent. Un commentaire général sur la notion de topographie et sur les solutions proposées sera donné en fin de chapitre.

3.1 Motivations présidant à la topographie de la surface des macromolécules biologiques

L'étude de la relation entre la structure et la fonction des macromolécules biologiques constitue le grand thème de la bioinformatique structurale. Une des approches les plus courantes dans cette étude consiste à scinder l'aspect structural suivant ses deux aspects "géométrique" et "bio-chimique". La caractérisation de la forme des molécules (de leur géométrie) trouve essentiellement son sens dans la recherche de "*similarité*" ou de "*complémentarité*" entre agents biologiques, voire plus directement dans l'étude de "l'*accessibilité*" de diverses parties de leurs surfaces. Ces qualificatifs sont illustrés dans la figure 3.2 et plus amplement détaillés dans le texte ci-après.

Des motifs similaires à la surface de macromolécules sont en effet des indices d'une potentielle similarité de fonction [Via 00]. De telles ressemblances peuvent être observées à la surface de protéines partageant peu d'identité de séquence, ce pour quoi une approche structurale pourrait apporter des informations invisibles à partir d'une étude de séquences seule. La similarité d'éléments de surface peut encore être le signe d'une divergence ancienne au cours de l'évolution, où seuls des motifs nécessaires à une fonction particulière auraient été conservés ; elles pourraient

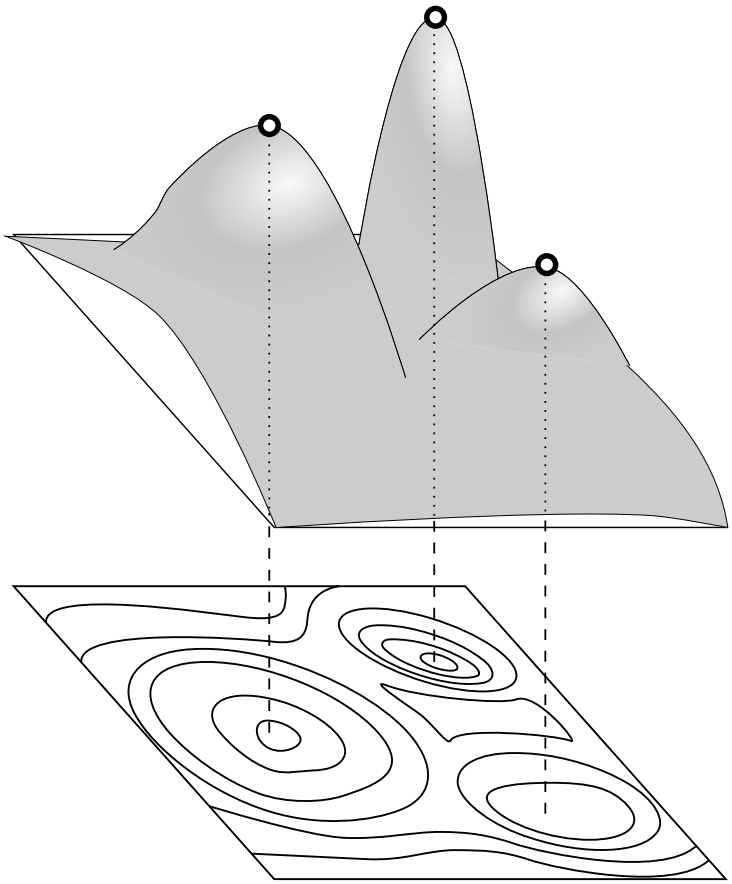


Figure 3.1: En topographie terrestre, la planète est supposée localement plane, et une notion de hauteur est définie à partir de ce plan de référence. Les sommets et les vallées correspondent aux extrema locaux de cette fonction-hauteur.

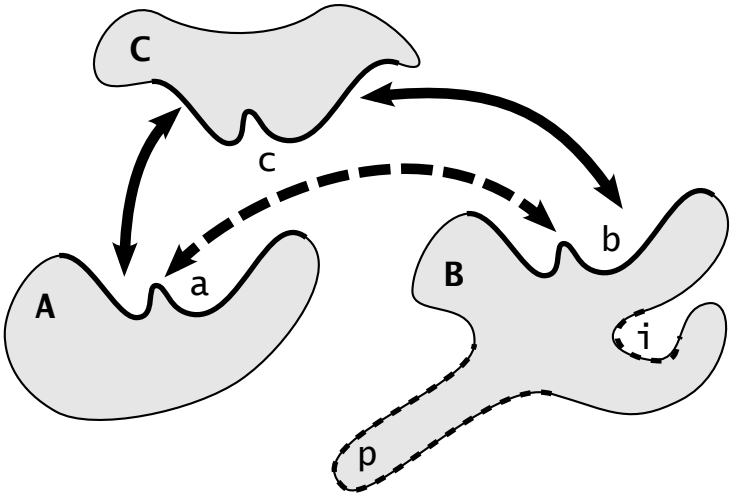


Figure 3.2: Similarité, complémentarité et accessibilité à la surface des macromolécules biologiques. Trois molécules A, B et C sont schématisées en deux dimensions. Les molécules A et B présentent des motifs similaires à leur surface, dénotés respectivement par a et b. La molécule C présente un motif c complémentaire de a et b. Sur la molécule B, les zones surfaciques p et i (mises en évidence par un trait interompus) dénotent respectivement une zone “proéminente” et une zone “anfractuée”.

également être expliquées par une convergence de fonction.

La présence de motifs complémentaires à la surface de macromolécules peut suggérer une interaction entre les deux macromolécules considérées. La recherche de motifs particuliers tels que les bosses et les creux de deux macromolécules connues est couramment employée pour fournir une représentation simplifiée de la molécule en vue de la recherche explicite de l'appariement de deux molécules (ou *molecular docking*). Dans ce genre d'approche, on tente de faire correspondre les "creux" et les "bosses" d'une molécule avec leurs pendants dans la seconde [Connolly 86c, Norel 94, Norel 95, Hendrix 98, Exner 02b]. Cette technique a également été adaptée à la recherche de l'amarrage d'une petite molécule à une protéine [Norel 99]. Les parties les plus "*proéminentes*" d'une macromolécule sont aussi les plus "*accessibles*", et de ce fait les plus susceptibles de participer à une interaction avec une autre macromolécule. La notion de proéminence a par exemple été explorée pour retrouver automatiquement les déterminants antigéniques dans les protéines [Novotny 86, Thornton 86]. De manière plus large, l'implication des résidus dans des interactions de complexes protéiques a été correllée à une mesure de leur accessibilité [**?**]. À l'inverse, les zones infractuées des macromolécules biologiques sont généralement considérées comme des sites maximisant les forces dans des interactions avec des ligands de petites tailles. De tels sites sont souvent impliqués dans des processus biologiques critiques tels que l'activité enzymatique des kinases [Hubbard 98] ou la modulation de la transcription par les récepteurs nucléaires [Laudet 01].

3.2 Approches pour topographier les macromolécules

Topographier la surface d'une macromolécule biologique consiste à mettre en évidence les "*bosses*" et les "*creux*" présents à sa surface[1]. Par analogie avec la topographie terrestre, on définit généralement une notion de "*hauteur*" ("d'*élévation*") sur la surface ; les extrema locaux de cette fonction définissent les propriétés topographiques recherchées. Par extension, les valeurs les plus hautes définissent les zones "élevées" ou "saillantes" (les "bosses") et les faibles valeurs les "vallées" (les "creux"). Déterminer une bonne fonction-hauteur et ses valeurs-seuil afin de définir les zones proéminentes et les vallées n'est pas un problème aisé. Il est parfois abordé dans la littérature sous l'appellation de problème de définition du "niveau de la mer" (*sea level problem*). La définition d'une bonne fonction-hauteur dépend de l'acception que l'on souhaite se donner des termes "bosses" et "creux", flous par nature. Ces termes peuvent être précisés diversement à l'aide de notions plus spécifiques —quoiqu'elles-aussi intuitives— telles que la *proéminence*, l'*accessibilité*, l'*enfouissement* ou encore l'*incurvation*. Ces différentes notions sont illustrées dans la figure 3.3. Des définitions plus précises seront données ci-après dans le texte et illustrées par des exemples issus de la littérature.

3.2.1 Notion de proéminence

Les notions de *protubérance* ou de *proéminence* adressent généralement la recherche des "excroissances" les plus "significatives" de la molécule. Il s'agit essentiellement de trouver les parties saillantes "significatives" de "grande taille", ou à un "haut niveau de résolution". Ces notions peuvent toutefois être déclinées à divers niveaux d'échelle. L'idée générale derrière la mise en œuvre de cette notion consiste à calculer la distance d'un point sur la surface de la molécule par rapport à une représentation grossière de la même molécule. D'une certaine manière, la notion d'"élévation" est ici transformée en "profondeur". Dans une première approche, les protéines

[1] Des qualités plus complexes de cette surface, telles que les méplats, les vallées, les crêtes ou les cols peuvent aussi être considérées.

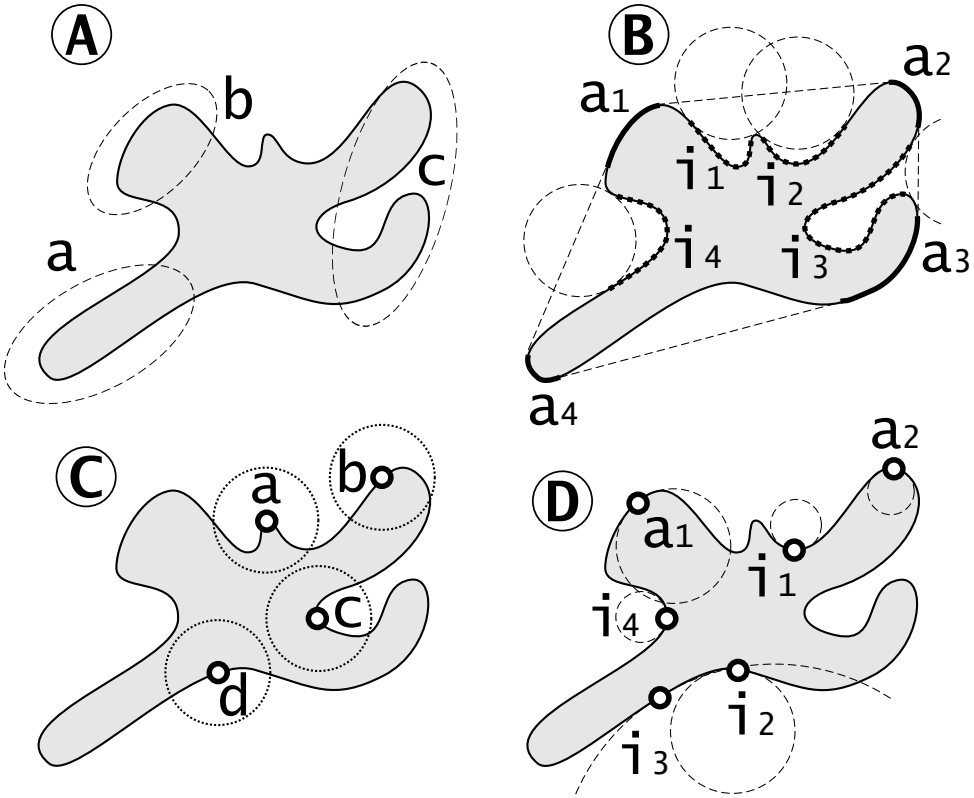


Figure 3.3: Différentes notions permettant de définir la topographie à la surface de macromolécules biologiques. La même molécule a été représentée dans quatre schémas en deux dimensions.
A : Les trois zones les plus "proéminentes" de la molécule sont mises en évidence par des ovales en traits intérompus.
B : Les quatre zones "accessibles" a_i peuvent être accédées par une sphère de rayon infini (un demi-plan), et les quatre zones invaginées i_i sont inaccessibles à une sphère d'un rayon r donné.
C : La notion d'enfouissement prend en compte le voisinage d'un point pour déterminer son degré d'exposition à la surface, ou d'enfouissement dans le corps de l'objet considéré. Le point a est situé dans une zone particulièrement exposée, à l'inverse du point c situé dans un creux ; les points b et d se situant sur des zones relativement planes.
D : Les mesures d'incurvation (de courbure) approximent localement la surface par un cercle. Les sommets a_1 et i_2 ont la même mesure d'incurvation, mais à l'inverse d'a_1, i_2 est "rentrant" ou "invaginé" car le cercle approximant est situé en dehors de la surface. Les points a_1 et a_2 sont tous deux situés sur une "bosse", mais le cercle approximant la surface est plus petit au voisinage de a_2, cette saillie est donc plus incurvée.

globulaires peuvent être assimilées à un ellipsoïde (voir figure 3.4 *A*). À partir d'un *ellipsoïde englobant* approximant au mieux la protéine[2], la "proéminence" (la "profondeur") peut être matérialisée par une famille d'ellipsoïdes concentriques, les ellipsoïdes les plus extérieurs représentant les couches atomiques les plus proéminentes. L'espace entre deux ellipsoïdes de cette famille concentrique peut être défini sur des critères de distance par rapport à l'ellipsoïde maximal ou en considérant la proportion de résidus (ou d'atomes) contenus dans une telle couche. Cette méthode simple a été utilisée avec succès pour retrouver les épitopes dans la structure des antigènes [Novotny 86, Thornton 86]. Plus récemment, une approche [Coleman 06] prenant l'enveloppe convexe d'une molécule pour représentation grossière a été proposée (voir figure 3.4 *B*). Avec *TravelDepth*, Coleman et Sharp proposent en outre une méthode pour calculer la plus courte distance à l'enveloppe convexe sans autoriser le franchissement de la surface de la molécule. Comme l'indique la figure 3.4 *C*, ce type d'approche prenant uniquement en compte une distance "à travers le vide" permet une meilleure mise en évidence des cavités.

L'emploi d'une forme "grossière" (figure 3.4 *C*) permet de rendre compte de propriétés topographiques générales de la molécule mais ignore les propriétés locales. Ces dernières peuvent être détectées par l'utilisation d'une forme approximante plus fine 3.4 *D*). Une telle approche est possible, en utilisant par exemple une forme-α comme représentation grossière, à la manière de ce qu'ont proposé Peters *et al.* [Peters 96] pour le logiciel *APROPOS*. Cette approche sera abordée plus en détail page 50 dans la section bibliographique dédiée à l'étude des poches (section 4).Ce dernier type d'approche s'apparente en outre à l'idée d'accessibilité, détaillée ci-après.

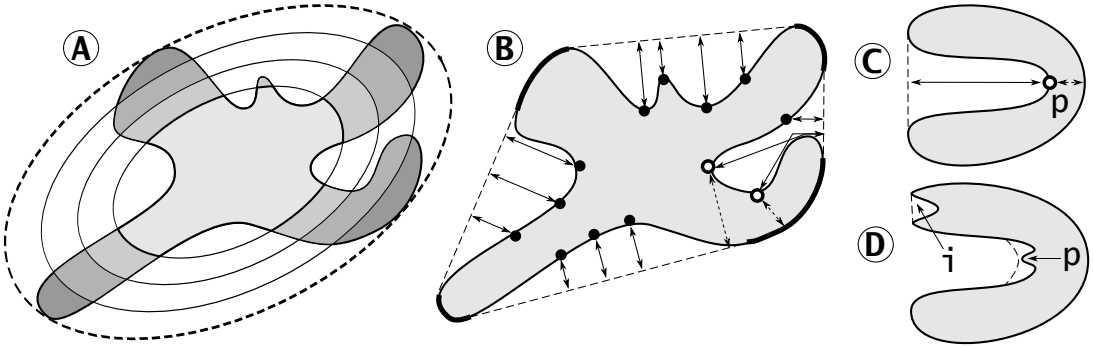


Figure 3.4: Notion de protrusion pour la caractérisation de la topographie des macromolécules. La macromolécule est comparée à une de ses représentations grossières, un ellipsoïde (figure *A*) ou son enveloppe convexe (figure *B*).
A : Une famille d'ellipsoïdes concentriques est dérivée d'un ellipsoïde englobant approximant la molécule. Les parties proéminentes de la molécule (en gris foncé) sont situées dans les ellipsoïdes les plus extérieurs.
B. : L'enveloppe convexe de la molécule comprend les parties épaissies de la surface moléculaire et les lignes pointillées. Une distance à l'enveloppe convexe peut être calculée pour chaque point de la surface moléculaire ou pour chaque centre atomique de la molécule (matérialisée ici par une flèche). Pour certains points (considérer les deux points blancs), le plus court chemin vers l'enveloppe convexe traverse la surface (flèches pointillées), on peut alors choisir le plus court chemin vers l'enveloppe convexe au travers du vide (flèches continues), qui garantit une meilleure définition des cavités.
C : Le point p est proche ou éloigné de l'enveloppe convexe selon qu'on s'autorise ou non à traverser l'intérieur de la molécule.
D : Avec la distance à l'enveloppe convexe, l'invagination i aurait une valeur de proéminence bien supérieure à celle de la petite saillie p ; une surface approximante moins grossière (en pointillés) permet d'étudier les détails à une échelle moindre.

[2]On prend généralement le plus petit ellipsoïde contenant tous les centres atomiques de la molécule

3.2.2 Notion d'accessibilité et d'exposition

La notion de proéminence est intrinsèque à l'objet étudié ; elle caractérise les propriétés topographiques de l'objet uniquement à partir de sa surface. À l'inverse, la notion "d'*accessibilité*" définit les propriétés topographiques de l'objet à partir d'une "sonde" extérieure interagissant avec l'objet étudié. Les parties de la surface que cette sonde peut atteindre sont dites "accessibles" et définissent les "bosses" ; les parties inaccessibles définissent les "creux" (voir la figure 3.5 A). Cette définition stricte peut être nuancée en considérant la taille de la plus grande sonde sphérique pouvant accéder à un endroit de la surface (figure 3.5 B). De cette manière, on définit une fonction d'accessibilité continue à la surface de la molécule, pouvant servir de fonction-hauteur ; cette approche a été utilisée dans le logiciel *Phecom* [Kawabata 07] pour la détection et la caractérisation des poches dans les macromolécules biologiques. Quoique continue, cette fonction peut présenter de fortes variations, comme au voisinage du point p_1 de la figure 3.5 B. Dans le cas des poches présentant une constriction (comme la cavité C de la figure 3.5 B), on peut alternativement s'intéresser à la sphère-sonde maximale pouvant localement entrer en contact avec un point de la surface. Cette sphère ne pouvant pas passer la constriction garantissant l'accès à l'espace de la poche, donne par sa taille une indication d'exposition locale à l'espace du solvant plutôt qu'une mesure d'accessibilité. Cette propriété a été utilisée par Kuhn *et al.* [Kuhn 92] pour démontrer la préférence des molécules d'eau pour les parties invaginées de la surface. Une autre méthode de calcul a été proposée par Todd.O. Yeates [Yeates 95] sous l'appellation de *Rayon maximal de contact* ; nous l'avons déjà évoquée dans la section précédente, page 23, en même temps qu'une autre mesure d'accessibilité basée sur la probabilité d'impact d'une sonde "neutre" lors d'une marche aléatoire, *l'accessibilité à la diffusion*. Une autre mesure d'"accessibilité" est aussi parfois employée : l'aire de surface accessible [Lee 71] (ASA) d'un atome (ou d'un résidu). Cette dernière propriété offre une mesure très locale de l'"exposition" d'un atome ou d'un résidu à l'espace du solvant. Les notions de proéminence et d'accessibilité se rejoignent en ce sens que les zones proéminentes sont aussi les plus "accessibles". Néanmoins, les mesures d'accessibilité conviennent généralement mieux à la caractérisation des anfractuosités.

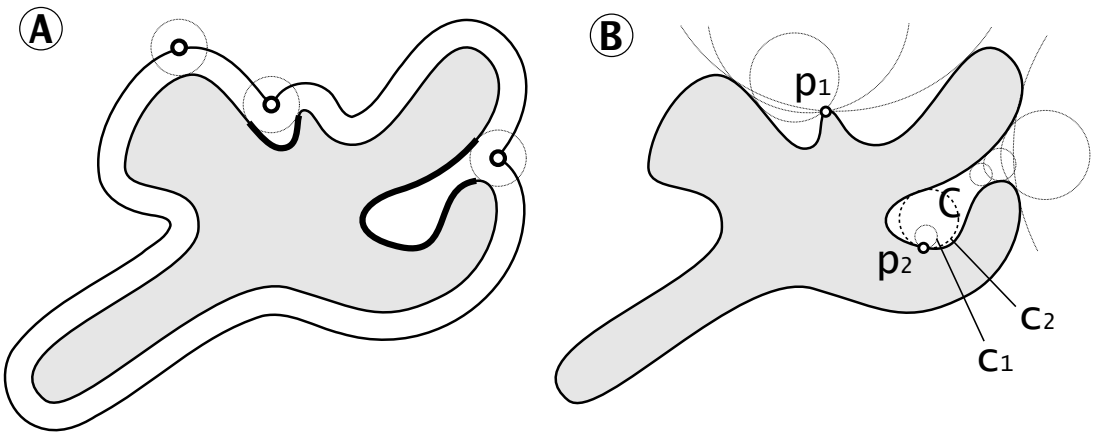


Figure 3.5: Notion d'accessibilité pour la caractérisation de la topographie des macromolécules.
A : Les "cavités" (contourées en lignes épaisses) sont inaccessibles à une sonde sphérique.
B : L'"accessibilité" d'un point à la surface de la molécule est définie par la taille de la plus grosse sonde pouvant entrer en contact avec ce point. Au point p_2 dans la cavité C on peut également considérer la taille de la sphère-sonde "localement maximale" C_2 (en pointillés) ; cette mesure s'apparente plus à une mesure d'exposition locale.

3.2.3 Mesures d'incurvation

Les notions d'*incurvation*, de *courbure*, de *torsion* ou de *fléchissement* font référence à des intuitions liées à la "forme" d'une surface, et plus spécifiquement à la manière dont cette surface dévie d'une surface plane. Ces notions autorisent généralement une palette descriptive plus large de la topographie ("points selles", "vallées", "crêtes"). En mathématiques, une définition usuelle de la courbure d'une surface lisse mesure les variations maximales de cette surface autour d'un de ses points ; elle est localement comparée à une forme quadratique dont on observe les caractéristiques (les axes de plus grandes pentes, et une mesure de cette pente) comme indiqué sur la figure 3.6. Les deux *courbures principales* k_1 et k_2 sont définies respectivement comme les courbures maximale et minimale autour d'un point. Par convention, les courbures principales sont signées positivement lorsque la surface est convexe et négativement lorsqu'elle est concave. La moyenne des courbures

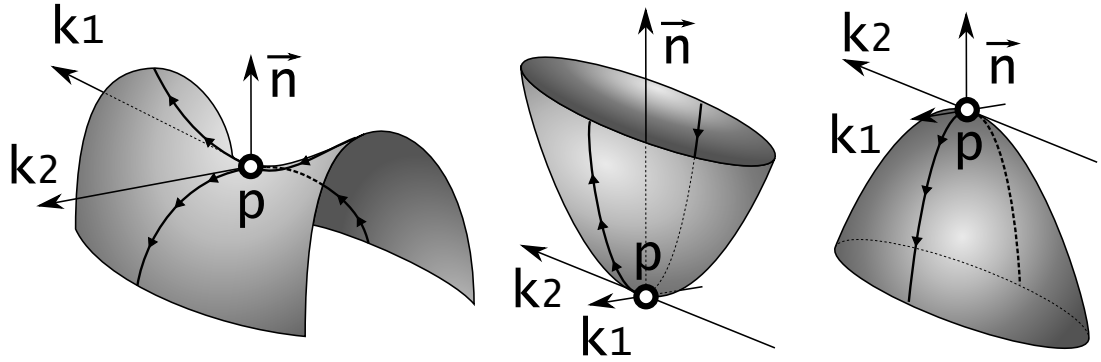


Figure 3.6: Définition mathématique de la courbure : autour d'un de ses points p, la surface est localement comparée à une forme quadratique. Elle peut alors prendre les trois formes décrites dans la figure, respectivement une selle de cheval, une concavité ellipsoïdale, et une convexité ellipsoïdale. Dans chaque figure, on a représenté un élément de surface autour de p et au dessus de l'objet considéré (le vecteur $\overrightarrow{n}$ normal à la surface en p pointe systématiquement en dehors de l'objet étudié.). Les directions des courbures maximales sont indiquées par des vecteurs libellés par la courbure correspondante. Sur la selle de cheval on a $k1 > 0$ et $k2 < 0$, dans la concavité, $k1 < k2 < 0$, et dans la convexité $k1 > k2 > 0$.

principales $H = \frac{1}{2}(k_1 + k_2)$ définit la *courbure moyenne* et leur produit $K = k_1 k_2$, la *courbure de Gauss*. La courbure ainsi définie est un indice très local mesurant l'écart de la surface par rapport au plan tangent. Cette définition théorique a été utilisée dès le début des années 1990 pour caractériser la forme des macromolécules biologiques : Duncan et Olson ont proposé une méthode de calcul pour le cas d'une surface moléculaire définie comme l'isocontour d'un mélange gaussien [Duncan 93b], ou pour une représentation en harmoniques sphériques [Duncan 93a]. Plus récemment, une solution a été proposée pour calculer ces valeurs pour une surface moléculaire définie analytiquement [Tsodikov 02]. Dans leurs travaux initiaux, Duncan et Olson ont proposé l'utilisation d'indices monodimensionnels dérivés des courbures principales : l'indice de forme (*shape index*) S et la *mesure d'incurvation* (ou *curvedness*) R.

$$S = \frac{2}{\pi} \arctan \frac{k_1 + k_2}{k_1 - k_2}$$

$$R = \frac{1}{2}\sqrt{k_1^2 + k_1^2}$$

S est à valeurs dans [-1 ; 1] et permet de discriminer les zones concaves (-1) des zones convexes (1), les points selles et les plats (0), ainsi que les vallées ($-\frac{1}{2}$) et les crêtes ($\frac{1}{2}$). R est une valeur positive mesurant l'amplitude de l'écart au plan tangent. Ces mesures étant par nature très locales,

Duncan et Olson ont proposé de les lisser localement autour de chaque point [Duncan 93b]. Heiden et Brickmann ont proposé un autre indice similaire basé sur les courbures principales. L'indice de topographie de surface (*Surface Topography Index* ou *STI*) [Exner 02a] prend ses valeurs dans $[0;4]$, à la manière du *shape index*. Une autre approche pour mesurer l'incurvation d'une surface consiste à l'interpoler localement par une primitive simple, comme une sphère, tel que le proposent Burr *et al.* [Burr 04, Coleman 05]. Leur idée, résumée dans la figure 3.7, consiste à considérer le rayon de la sphère approximant au mieux (c'est-à-dire au sens des moindres carrés) la surface dans le voisinage d'un point comme une mesure de la courbure dans ce voisinage.

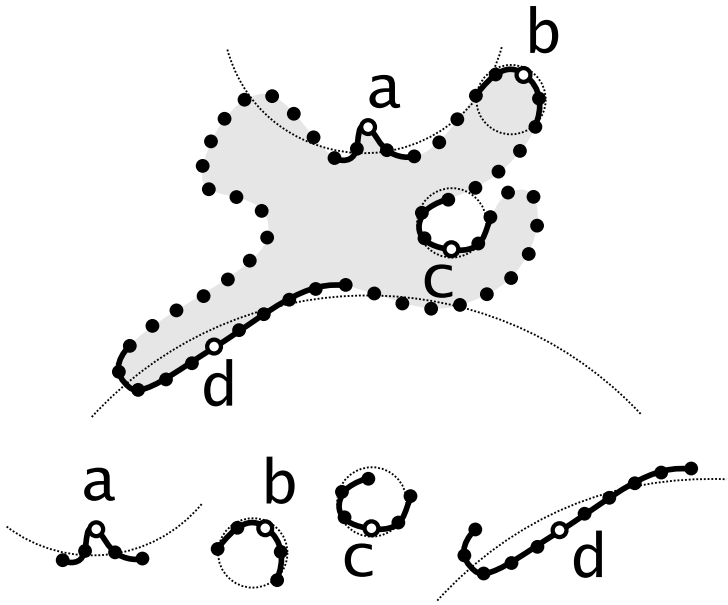


Figure 3.7: Incurvation par approximation à l'aide d'une sphère. La surface discrétisée d'une molécule est schématisée ici en deux dimensions par une succession de points noirs. La surface est approximée par des sphères aux voisinages des points a, b, c (pour des voisinages comprenant 5 points) et d (avec un voisinage de 11 points). Pour une meilleure lisibilité, les approximations aux voisinages de ces points sont répétées sous la figure. La courbure aux points b et c est élevée car les cercles approximants sont petits, à l'inverse du point d situé sur une zone plane.

3.2.4 Méthodes basées sur l'enfouissement

La notion d'*enfouissement* d'un point dans l'objet étudié rend compte de la position de ce point par rapport au reste de l'objet ou par rapport à son voisinage dans l'objet. Les zones de la surface les plus "enfouies" définissent les "creux", tandis que les zones les moins enfouies constituent des "bosses" à la surface de l'objet. La notion d'enfouissement peut être caractérisée par des mesures de *densité*, de *répartition*, ou de *visibilité* (voir figure 3.8). Évaluer la densité d'un objet autour d'un point p de sa surface consiste à mesurer la proportion de volume occupé par l'objet dans un voisinage défini par une sphère $\mathcal{S}$ centrée en p (figure 3.8 A). Cette mesure dépend d'un "paramètre d'échelle", le rayon r de la sphère $\mathcal{S}$ définissant le voisinage[3]. Une telle mesure constitue un indice de l'"enfouissement" dans l'objet à une échelle d'observation donnée. Les points de la surface ayant une forte densité sont enfouis dans l'objet ; une densité faible indique que le point considéré est situé sur une saillie hors de l'objet. Avec *CX*, Pintar *et al.* ont

[3]Pour les applications biologiques, cette valeur est généralement fixée entre 6 et 12 Å, des valeurs empiriques qui permettent d'obtenir une bonne description de la topographie de la surface moléculaire ([Pintar 02]).

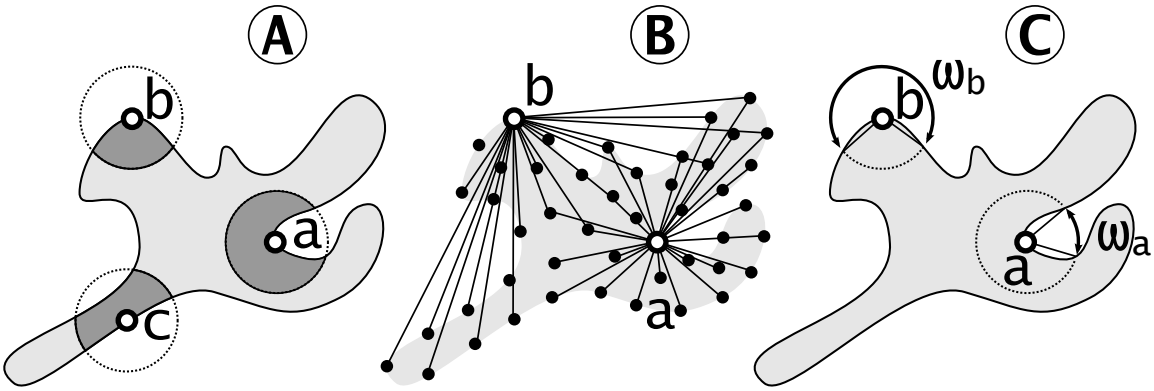


Figure 3.8: Caractérisation de l'enfouissement par des mesures de densité (A), de répartition (B) ou de visibilité (C).
A : le point a est enfoui car son voisinage est principalement occupé par la molécule, à l'inverse du point b essentiellement entouré de "vide". La mesure au point c prend en compte le vide de deux "côtés" de la molécule.
B : le point a est enfoui car il a de nombreux voisins à proximité; ceux-ci étant "harmonieusement" répartis autour de lui. À l'inverse, le point b a peu de proches voisins, tous situés du même côté.
C : le point a est enfoui, car son "horizon libre" (matérialisé par l'angle ω_a) est restreint, à l'inverse du point b dont l'horizon est dégagé.

proposé une méthode simple et rapide pour évaluer cette densité en se basant uniquement sur le dénombrement des atomes présents dans la sphère de voisinage [Pintar 02]. Pour caractériser les crevasses à la surface des protéines, Kuhn *et al.* [Kuhn 92] ont introduit l'indice fractal (ou *subfractal index*) de la surface en un de ses points. Cette mesure permet d'observer non seulement la densité autour du point considéré, mais également l'évolution de cette densité en fonction du rayon r de la sphère de voisinage. Comme le montre l'exemple du point c dans la figure 3.8 A, ce genre de mesure indépendante de la surface peut définir comme "exposées" des zones qui ne le sont pas. En particulier, ces mesures sont dépendantes de la densité atomique locale de la molécule ainsi que de la présence potentielle de cavités proches de la surface. Étant donné un nuage de points $A = \{a_i\}_i$ (par exemple des centres atomiques), une méthode pour évaluer l'enfouissement d'un point a_i dans le nuage A consiste à observer la "*répartition*" des a_j ($j \neq i$) autour de a_i (figure 3.8 B). Une première approche consiste à étudier les distances $d(a_i, a_j)$. Un a_i "enfoui" aura de nombreux proches voisins, soit beaucoup de courtes distances. Cette approche a été explorée pour caractériser la forme d'objets volumiques ou de surfaces quelconques par Heiden *et al.* avec l'indice *SEP* normé (*Surface Embededness Potential*) [Heiden 05].

$$SEP(a_i) = \frac{1}{N} \sum_{j \neq i} \frac{1}{1 + d(a_i, a_j)^2}$$

Une autre approche, proposée par Mihaly Mezei pour détecter les poches dans les macromolécules ainsi que les liaisons entre domaines (ou *linkers*), consiste à observer la répartition spatiale des a_j autour d'un a_i [Mezei 03]. L'indice d'"intériorité" (*insideness*) est ici défini par une mesure de variance circulaire CV.

$$CV = 1 - \frac{1}{n} \left| \sum_{j \neq i} \frac{a_i - a_j}{|a_i - a_j|} \right|$$

Une version pondérée CV^W de la variance circulaire permet en outre de donner plus d'importance aux atomes proches, tout en réduisant le temps de calcul en limitant le nombre d'opérations de

division.

$$CV^W = 1 - \frac{|\sum_{j \neq i} a_i - a_j|}{\sum_{j \neq i} |a_i - a_j|}$$

Une autre manière de caractériser l'enfouissement d'un point à la surface d'un objet consiste à observer le "point de vue" ou la "*visibilité*" qu'aurait un observateur placé en ce point (figure 3.8 *C*). D'une certaine manière, on inverse ici la problématique de la densité en ce sens qu'on cherche à mesurer une proportion d'espace vide autour du point. Une première approche de ce type a été proposée par M.L. Connolly pour topographier la surface des macromolécules [Connolly 86a]. L'indice initialement proposé par Connolly consistait en la mesure d'angle solide prise à une distance donnée autour d'un point p de la surface de la molécule. L'angle solide est un analogue tridimensionnel de la mesure d'angle ; il se définit en observant une sphère intersectant un objet et en mesurant l'aire de la partie de la sphère située en dehors de l'objet (voir figure 3.9). Comme le montre la figure 3.10, le rayon choisi pour la sphère influe grandement sur la valeur

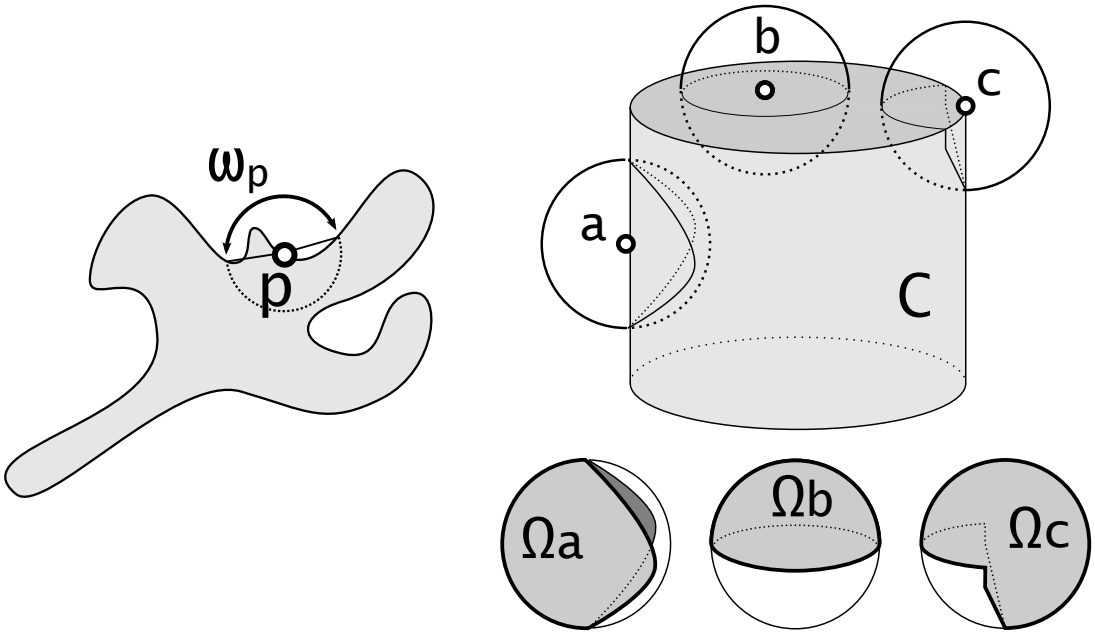


Figure 3.9: Définition de l'angle solide. Dans l'exemple en deux dimensions de gauche, la "visibilité" au point p est mesurée par l'angle ω_p matérialisé par l'intersection d'un cercle centré en p avec la surface considérée. L'équivalent en trois dimensions (exemple de droite) est l'angle solide Ω correspondant à une mesure d'aire sur une sphère. Les angles solides en a, b et c, respectivement au cylindre C, sont les aires Ω_a, Ω_b et Ω_c reportées sous la figure pour plus de visibilité.

d'angle solide, et cette notion donne une vision de la surface restreinte à une distance donnée. Pour pallier ce problème et tenir compte des modulations de la surface à l'intérieur de la sphère, M.L. Connolly a proposé d'intégrer ce premier indice en faisant varier le rayon de la sphère de voisinage [Connolly 86c]. La *fonction de Connolly* ainsi définie s'apparente à une mesure de densité.

3.3 Difficultés inhérentes à la caractérisation de la topographie

Caractériser la topographie de surface des macromolécules (ou d'objets quelconques) est une problématique difficile, en grande partie en raison du flou dans l'appréciation des définitions.

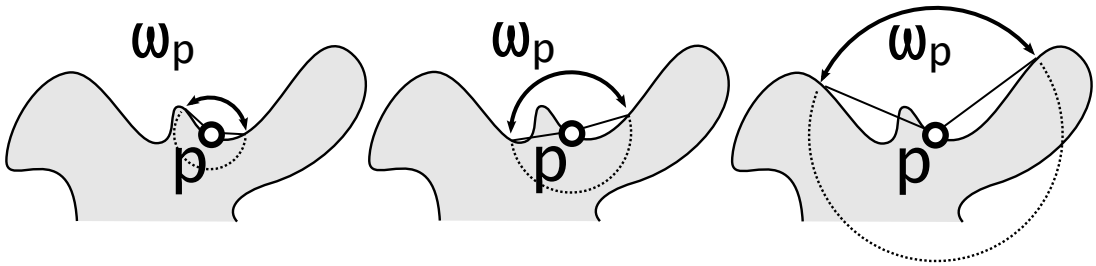


Figure 3.10: Fonction de Connolly. Dans cet exemple en deux dimensions, l'angle solide au point p est représenté par l'angle ω_p. Pour un petit rayon (à gauche) ou pour un grand rayon (à droite), l'angle solide est proche de $\frac{2\pi}{3}$, indiquant une anfractuosité. Pour un rayon intermédiaire, l'angle solide est proche de π, indiquant une zone plate. Bien qu'elle soit couverte par le cercle définissant l'angle solide, la petite saillie à gauche du point p n'est pas prise en compte dans le calcul de cette valeur.

Confronté à l'image d'une molécule, un opérateur humain pourra assez aisément détourer les "bosses" et les "creux" à sa surface sans pour autant avoir eu une définition préalable de ces termes. Deux opérateurs distincts détoureront néanmoins souvent des zones différentes, avec parfois des divergences marquées, généralement explicables par un domaine d'application spécifique visé par chacun. Afin de permettre une automatisation de l'annotation topographique, les définitions doivent donc être préalablement précisées par un spécialiste du domaine visé. Cependant, quel que soit le domaine d'application, certains cas "pathologiques" peuvent s'avérer indécidables même pour le spécialiste, et le cas des "limites" à partir desquelles un élément topographique devient un "creux" ou une "bosse" est toujours problématique.
Globalement, les difficultés rencontrées dans la création ou l'utilisation d'un indice topographique peuvent donc concerner des aspects qualitatifs (c'est-à-dire concernant les définitions "intuitives") ou quantitatifs (concernant leur mise en valeur).

La première difficulté réside dans l'existence d'intuitions différentes pouvant servir de définitions aux termes de "creux" et de "bosse" (figure 3.11 *a* à *c*) ; nous avons par exemple mentionné la *proéminence*, l'*accessibilité*, l'*exposition*, l'*incurvation* et l'*enfouissement*. Généralement, la problématique étudiée permet de lever l'ambiguïté des définitions.

Une autre problématique difficile — transversale à la majorité de ces approches — est la notion d'*échelle*, ou de *niveau de résolution* (figure 3.11 *d*) : l'utilisateur doit définir la taille des propriétés topographiques qu'il souhaite étudier. En particulier il est compliqué d'étudier simultanément différents niveaux de détail.

Une autre difficulté réside dans la mise en valeur de notions intuitives. Outre l'aspect souvent arbitraire d'une telle procédure (répartition de ces valeurs, choix de paramètres tels que la taille d'un voisinage...), l'implémentation de programmes informatiques peut s'avérer malaisée, en raison par exemple de problèmes numériques ou de temps de calcul. Une fois ces valeurs calculées, la détermination des "creux" et des "bosses", basée sur cet indice, constitue un autre processus arbitraire et complexe. La méthode la plus simple consiste à considérer comme faisant partie d'un "creux" ou d'une "bosse" les atomes ayant une valeur respectivement inférieure ou supérieure à des valeurs-seuils "arbitrairement" choisies. D'autres méthodes plus élaborées adressent ce problème en utilisant des outils de logique floue [Exner 02a] ou des diagrammes de persistance [Edelsbrunner 00].

Toutes ces difficultés rendent malaisé, et souvent subjectif, la comparaison ou le classement d'indices topographiques différents. Pour commencer, de tels indices n'ont pas nécessairement été créés pour répondre à des problèmes identiques. Ensuite, les plages et les répartitions de valeurs sont souvent dissimilaires entre deux indices, biaisant toute tentative d'établir une correlation. De

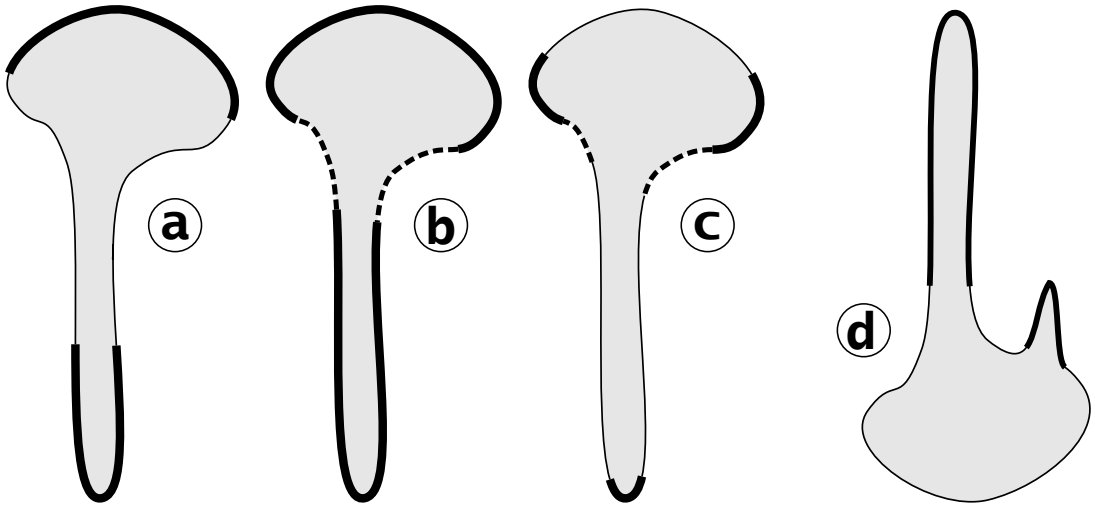


Figure 3.11: Les cas a à c montrent différentes définitions des termes "creux" (en pointillés) et "bosses" (en lignes épaisses). Les définitions du cas a seront accessibles au travers de mesures de proéminence, celles du cas b au travers de l'accessibilité à une sphère-sonde de taille idoine, le cas c pourra être obtenu par des mesure d'incurvations ou des mesures de densité.
L'objet d montre deux "bosses" à des résolutions différentes.

plus, les indices topographiques sont souvent dépendants de paramètres (par exemple la taille d'un voisinage), et pour comparer deux indices il faudra choisir des paramètres pour chacun d'eux ce qui constitue encore un point d'arbitraire. La comparaison est de ce fait bien souvent restreinte à une inspection visuelle dans laquelle on se contente de commenter les similarités et les différences entre les indices observés, ainsi que leurs adaptations à une problématique précise.

En définitive, si aucun indice topographique ne peut être qualifié d'universel, de nombreuses propositions existent pour définir et détecter les "creux" et les "bosses" d'une surface moléculaire, chacune plus ou moins adaptée à une problématique spécifique.

Après cette présentation générale de la caractérisation topographique de la surface moléculaire, nous nous intéressons à un problème à la fois proche et distinct : la détection et la caractérisation des "poches" dans les macromolécules.

Chapitre 4

Détecter et caractériser les poches dans les macromolécules

Dans le langage courant le terme “poche” désigne un contenant, une surface entourant un espace, et dans lequel on peut enclore (ou, est enclos) un objet. Cette notion suggère tout à la fois une délimitation de l'espace (le bord de la poche), l'espace qu'il enclos (son volume), et la potentialité d'un contenu.

Dans le contexte macromoléculaire, le terme de “poche” désigne habituellement une zone particulièrement anfractuée et généralement impliquée dans des interactions avec des substrats de petite taille ; par ce terme on peut encore désigner un “défaut de densité” (*packing defect*) potentiellement impliqué dans des aspects dynamiques de la molécule. Les poches dans les structures macromoléculaires — tout comme dans un objet quelconque — peuvent encore être définies sur des critères uniquement géométriques ; c'est là une des approches les plus communes pour les définir et leur détecter, et c'est celle pour laquelle nous avons opté dans nos travaux.

Dans une première section quelques exemples de processus biologiques impliquant des poches seront présentés, qui permettront dans la seconde section une première tentative de définition ou de discrimination du terme de “poche” en fonction des motivations présidant à leur détection. Dans la troisième section nous présenterons l'aspect géométrique du problème ainsi qu'une classification de ces “poches géométriques” selon leur “forme”. Un inventaire classé et commenté des nombreuses approches qui ont été proposées dans la littérature pour détecter ces poches est présenté dans l'avant dernière section. Dans la mesure où nos travaux sont à classer parmi ces types d'approches, nous y avons mis l'accent essentiellement sur les approches géométriques et plus spécifiquement sur les travaux d'H. Edelsbrunner [Edelsbrunner 98] et de Peters *et al.* [Peters 96]. Les difficultés inhérentes à la détection des poches seront résumées et discutées dans la dernière section.

4.1 Motivations biologiques de l'étude des poches

L'analyse de la structure tridimensionnelle des macromolécules biologiques révèle de nombreuses lacunarités de tailles et de formes variées. Ces sites enfouis dans le corps d'un agent moléculaire sont souvent impliqués dans la fonction de ce dernier.

La densité atomique dans les protéines a été étudiée très tôt comme un indice potentiel concernant le processus de repliement de ces macromolécules (*folding process*), ainsi que leur flexibilité [Richards 74]. Un antagonisme entre flexibilité et stabilité est en effet nécessaire pour

la réalisation de la fonction d'une protéine : les défauts de densité (ou *packing defects*) d'une structure permettraient aux éléments plus denses de s'agencer entre eux. Ogata *et al.* [Ogata 96] ont par exemple exhibé le cas de la protéine Myb dont la reconnaissance spécifique d'un partenaire est tributaire d'une cavité dans la structure.

Le transfert d'ions ou de substrat d'un milieu à un autre est souvent effectué au travers de tunnels, des poches présentant plusieurs bouches. Ces anfractuosités peuvent être matérialisées par l'assemblage de plusieurs sous-unités, comme dans le cas des canaux du ribosome [Petřek 07], ou à l'intérieur d'une seule molécule, comme dans le cas des canaux ioniques de la gramicidine-A [Smart 93].

Les poches peuvent encore être impliquées dans les réactions enzymatiques. La protéine CARM1 [Troffer-Charlier 07], par exemple, est responsable de la méthylation spécifique de certains histones. Le cofacteur qui fournit le groupement méthyle est enfoui dans une invagination particulièrement profonde de la protéine. Le site catalytique, bien que plus ouvert, est lui-aussi considérablement anfractué. Une étude plus détaillée de cette enzyme sera donnée page 110 lorsque nous en étudieront la topographie.

Enfin, les poches peuvent être impliquées dans la transduction d'un signal. Dans le cas des récepteurs nucléaires, par exemple, l'activité du récepteur est régulée par l'accomodation d'une petite molécule (ou ligand) dans une poche de taille conséquente enfouie sous la surface de la protéine. Un exemple simplifié du fonctionnement de cette famille de protéines est donné en annexe A.4, et une étude sur un cas particulier sera donnée plus loin lorsque nous en étudieront la topographie page 105.

4.2 Définir les poches dans les macromolécules

L'étude des poches dans les macromolécules biologiques constitue un pas vers la compréhension de leur fonctionnement. Dans cette optique on peut être amené à définir les contours et à caractériser une poche connue, soit pour l'étudier directement au moyen d'une analyse visuelle détaillée, soit pour réaliser une analyse automatisée, souvent massive et consistant à comparer un grand nombre de poches entres elles sur la base de propriétés variées (géométriques comme physico-chimiques). Les motivations d'une telle étude peuvent être par exemple d'inférer la fonction d'une protéine dont on vient de résoudre la structure en y retrouvant des motifs existant dans des structures déjà résolues et dont la fonction est connues [Binkowski 03a, Keil 04, Gold 06]. Cette motivation a d'ailleurs donné lieu à une proposition de taxonomie du *Pocketome* [An 05].

La connaissance d'une poche ou d'un sillon peut aussi être utilisée dans la recherche d'une molécule partenaire, dont la fixation dans la poche permettra d'activer ou d'inhiber la fonction de la macromolécule [Zhao 05].

La notion de "poche" dépend ainsi du contexte et des motivations ; en particulier, on peut distinguer les cas suivants :

- L'utilisateur sait déjà où se situe la poche — il possède par exemple une structure ligandée — et souhaite préciser son contour[1] ou ses caractéristiques pour l'étudier, l'analyser, la comparer, ou plus prosaïquement valider la détection de la poche du ligand par un autre algorithme auquel on n'aura pas fourni la position de ce dernier.

[1] Ce contour peut être donné par une surface lisse ou maillée, ou plus généralement par la liste des atomes ou des résidus participant au bord de la poche, c'est-à-dire ceux qui interagissent avec un ligand.

– L'utilisateur connaît la structure de plusieurs macromolécules et possède de l'information sur le ou les types de poches qu'il cherche. Dans ce genre de cas on a généralement une phase d'"apprentissage" des caractéristiques de ces poches suivie d'une phase de *prédiction* visant à déterminer sur une nouvelle structure l'emplacement de poches putatives.

– L'utilisateur connaît la structure d'une macromolécule biologique et celle d'un ligand putatif, il souhaite découvrir les endroits où le ligand est le plus susceptible de se fixer, et quelle conformation il y adopte. Cette approche est désignée sous l'appellation d'arrimage moléculaire (ou *(molecular) docking*). Une approche "inversée" consiste à tester non pas l'interaction potentielle d'un ensemble de ligands sur une structure de protéine, mais à chercher les cibles macromoléculaires potentielles d'un ligand donné. Ce genre d'approches sont communes dans les premières phases d'une recherche de nouveau candidat médicament (*drug design*).

– L'utilisateur connaît uniquement la structure d'une macromolécule biologique. Il peut n'avoir aucun présupposé quant à la position des poches dans la structure, ou bien il peut avoir une idée et vouloir guider sa *détection*. Dans ce genre de cas, la définition du terme "poche" est purement géométrique et consiste en la recherche d'une zone "vide" entourée par les atomes de la molécule.

C'est cette dernière acception que nous avons traitée dans nos travaux, aussi nous focaliserons nous plus spécifiquement sur cet aspect dans la suite du document. Nous continuerons donc d'employer le terme "poche" avec un sens générique pouvant servir à désigner tous les sites d'intérêt biologique évoqués :

> *Une poche d'une macromolécule est un espace à la fois proche de la molécule, enfoui dans celle-ci, et vide d'atomes de cette molécule.*

Les notions floues de "proximité" ou d'"enfouissement" dans la molécule restent à définir.

Par opposition, on parlera de *site de fixation au ligand* ou de *site d'interaction* pour désigner plus spécifiquement des "poches" d'intérêt biologique.

4.3 Définir géométriquement les poches

Une classification des poches sur des critères de forme est souvent faite dans la littérature, nous la reportons ci-après et l'illustrons dans la figure 4.1. Comme nous le verrons dans la section suivante, de nombreux algorithmes destinés à la détection des poches basent leurs recherches sur ces définitions ou sont plus adaptés à la détection de l'une ou l'autre de ces poches. On distingue généralement :

Les tunnels : des "poches" fines et allongées reliant deux espaces plus larges. Il pourra s'agir d'un conduit reliant l'espace du solvant à une grande cavité dans la structure étudiée, ou d'un canal traversant la molécule de part en part comme la poche a de la figure 4.1.

Les cavités : des poches totalement enfouies sous la surface de la molécule. Cette notion d'enfouissement fait généralement référence à un vide ou un ligand totalement isolé de l'espace du solvant (figure 4.1 poche c).

Les poches refermées : des poches présentant un rétrécissement avant leur ouverture vers l'espace du solvant (figure 4.1 poche c).

Les Poches ouvertes : des poches grandes ouvertes sur l'espace solvant (ne présentant pas de rétrécissement). Elles sont parfois subdivisées en puits et sillons suivant qu'elles s'étendent

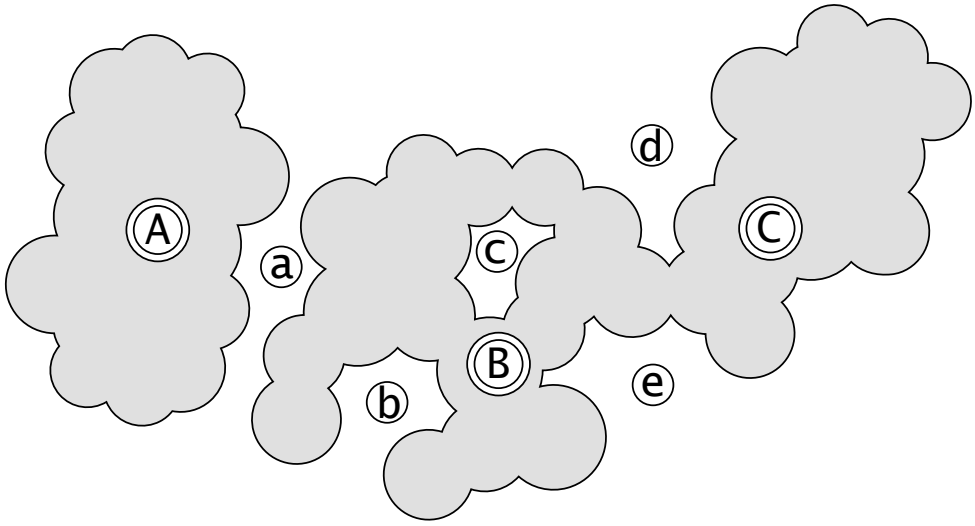


Figure 4.1: Classification des poches suivant leurs formes. Deux sous unités sont représentées, l'une constituée d'un unique domaine A, l'autre constituée des deux domaines B et C. La poche a est un tunnel, la poche b est *refermée*, c est une *cavité*, d et e sont *ouvertes*.
La poche a est matérialisée par la zone d'interface des deux sous-unités, les poches d et e sont situées à la jonction des deux domaines B et C, et les poches b et c appartiennent au domaine B.

en profondeur ou en longueur (figure 4.1 poche d et e).
On parle de *bouche* pour désigner la séparation entre l'espace attribué à la poche et l'espace du solvant. La définition de cette frontière n'est pas toujours aisée, elle s'avère particulièrement floue dans le cas des poches ouvertes où elle s'apparente au problème du "niveau de la mer" évoqué dans la partie précédente (page 27). Notons que dans le cas de poches refermées, le resserrement définit naturellement l'emplacement d'une bouche ; néanmoins, lorsque la constriction est peu marquée, l'emplacement de cette délimitation est sujet à caution et la poche peut s'avérer de forme fantaisiste, ou arbitrairement grande ; auquel cas d'autres critères sont nécessaires pour donner une frontière plus satisfaisante, tant d'un point de vue intuitif (ou géométrique) que pour assurer une pertinence biologique. Notons encore que certaines poches peuvent présenter des constrictions internes, complexifiant la définition de poche refermée : dans la figure 4.2, faut-il considérer chaque sous-poche comme une poche à part entière, ou faut-il ne considérer que l'une des deux ? Et qu'en est-il lorsque les deux sous poches sont de natures différentes, l'une refermée et l'autre ouverte ?

Comme esquissé dans la section 4.1, cette classification géométrique correspond à une réalité biologique telle qu'observée dans les sructures moléculaires. Nous avons vus des exemples de tunnels avec les canaux du ribosome et les canaux ioniques transmembranaires de la gramicidine-A. La "poche du ligand" d'un récepteur nucléaire, elle, est généralement soit une cavité, soit une poche refermée présentant une constriction très étroite. La poche du cofacteur dans CARM1 ou celle dans les protéines kinases sont, elles-aussi, des poches refermées, alors que le site actif présente une constriction peu marquée, voire inexistante. De manière plus générale, des études systématiques menées sur un grand nombre de structures ont montré que le ligand se nichait très généralement dans la plus grande proche d'une protéine et le plus souvent au moins dans l'une des trois plus grandes [Laskowski 96, Liang 98b, Guilloux 09].

Enfin, dans le cadre de l'analyse des structures tridimensionnelles, les spécificités d'un problème posé (par exemple, la recherche de sites ou de cibles particuliers), on peut considérer d'autres critères, comme des propriétés physico-chimiques (densité des liaisons hydrogène,

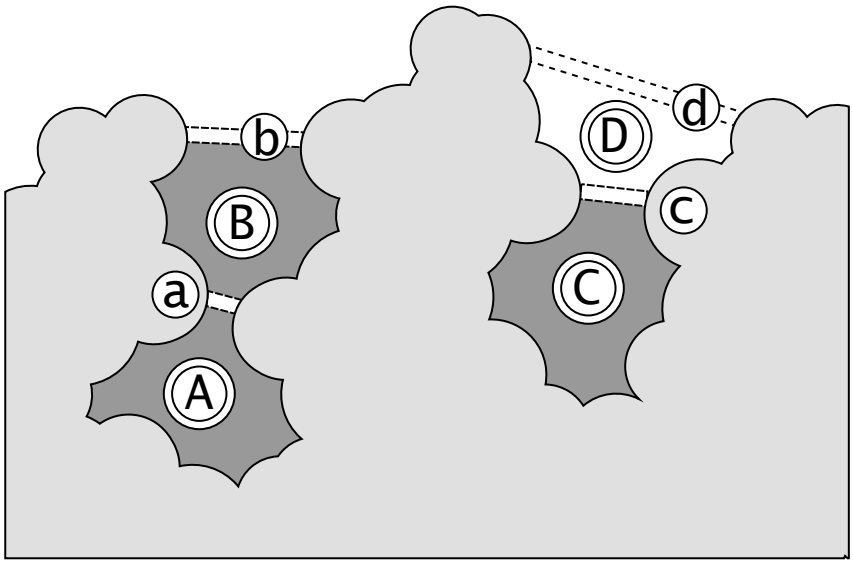


Figure 4.2: Bouches, constrictions et sous-poches, schéma représentant une coupe transversale dans une structure tridimensionnelle. Les trois sous-poches A, B et C sont bien délimitées par les resserrements a, b et c. La sous-poche D est grande ouverte sur l'espace du solvant ; sa limite d avec cet espace est floue. Suivant que l'on considère les sous-poches séparément, ou qu'on considère leur union, ces limites font ou non office de bouche.

charges, hydrophobicité) ou des informations concernant la séquence (nature des acides aminés, conservation de ces acides aminés), ou encore d'autres critères géométriques décrivant mieux la "forme" du site recherché. Comme nous le verrons dans la section suivante, de telles informations peuvent être insérées et utilisées directement dans la détection de sites d'interactions ; elles peuvent aussi être utilisées en seconde instance après une première détection purement géométrique. L'importance de la composante "forme" dans la reconnaissance moléculaire a été démontrée entre autre par Nayal et Hönig [Nayal 06], qui après avoir testé 408 propriétés différentes de nature physico-chimique comme topographique, en ont extrait 18 statistiquement pertinentes pour définir les sites de fixation d'un ligand actif, et remarqué que la majorité d'entre elles étaient de nature géométrique. De fait, de nombreuses approches en deux temps se trouvent justifiées ; la recherche des poches sur des critères géométriques constituant un premier filtre, affiné par une étude de ces poches géométriques sur des critères physico-chimiques.

4.4 Détecter les poches dans les macromolécules biologiques

De nombreuses approches ont été explorées pour détecter les poches dans la structure de macromolécules biologiques ; elles diffèrent essentiellement sur la stratégie mise en oeuvre. Par stratégie on désigne l'idée générale exploitée pour détecter les poches, qui dépend fortement du type de poche recherchée. Dans la suite de cette section nous proposons une classification des stratégies existantes dont un résumé est visible dans le diagramme de la figure 4.3. La distinction n'étant pas toujours claire, on présentera quelques travaux visant à chercher des *sites de fixation*, mais dans la mesure où elles sont plus proches des travaux que nous avons menés, nous développerons plus avant les méthodes visant à chercher des *poches géométriques*, et plus spécifiquement des *cavités* ainsi que des *poches refermées* ou des *poches ouvertes*.

Une première distinction majeure peut-être faite entre les stratégies de détection des poches,

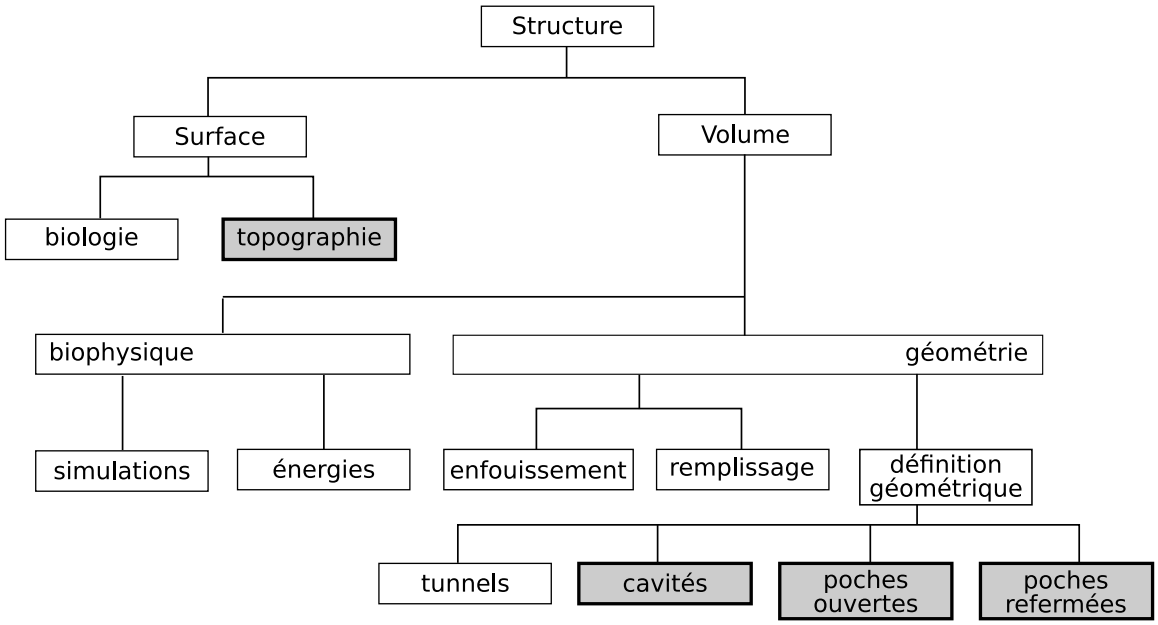


Figure 4.3: Classification des stratégies pour la détection des poches dans les macromolécules. En grisé, les approches que nous avons explorées.

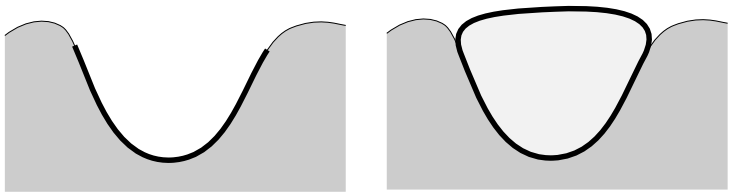


Figure 4.4: Deux paradigmes principaux pour la définition des poches, un schéma en deux dimensions. Une même poche dans un objet peut alternativement être comprise comme un élément de la surface de cet objet (matérialisé par un trait gras dans le schéma de gauche), ou comme un élément de volume dans le complémentaire de cet objet (matérialisé par une forme gris clair dans le schéma de droite).

suivant qu'on considère ces dernières comme une partie de la *surface* de la molécule étudiée, ou comme un élément de *volume* dans l'espace du solvant (voir la figure 4.4).

4.4.1 Stratégies basées surface

Deux catégories de stratégies basées sur une description surfacique peuvent être distiguées ; la première basée sur une segmentation purement géométrique (dans ce cas on parle plutôt de topographie), la seconde utilisant des informations biochimiques. Dans le premier cas on a généralement recours à des techniques de segmentation de surface basées sur des indices topographiques calculés à la surface de la molécule. De telles approches ont entre autre été proposées avec des indices basés sur un calcul d'angle solide tel que la fonction de Connolly [Connolly 86b], ou d'autres indices décrivant par exemple l'incurvation de la surface STI [Exner 02a] présentée au chapitre précédent (page 32) ou sa profondeur à une surface référence [Coleman 06].

Une segmentation de la surface sur d'autres critères est aussi envisageable. Par exemple, Keil *et al.* [Keil 04] commencent par une segmentation de la surface suivant la densité de liaisons hydrogène, puis caractérisent les zones ainsi détourées suivant six autres propriétés, dont trois géométriques (la courbure, la profondeur et le STI). Cette étude leur a permis d'établir des

motifs communs à divers types de zones interagissantes suivant la nature du partenaire : ADN, protéine, ou ligand de petite taille.

D'autres approches prédictives basées sur l'étude systématique des propriétés de la surface n'ont pas recours à une segmentation explicite sur un premier critère, mais reposent plutôt sur la notion de parcelle de surface (ou *surface patch*). Une parcelle de surface autour d'un résidu est constituée d'un ensemble connexe de résidus de surface répartis de manière approximativement circulaire autour de ce résidu le long de la surface. Pour chaque parcelle centrée sur un résidu de surface, des propriétés sont calculées, dépendant de la nature des atomes et des résidus qui la composent, ainsi que de leurs positions dans l'espace. Les parcelles sont ensuite fusionnées si elles ont les mêmes profils et partagent des résidus communs. Jones et Thornton [Jones 97a, Jones 97b] ont par exemple utilisé ce type d'approche pour caractériser et prédire les sites d'interaction entre deux protéines.

4.4.2 Stratégies basées volume

Dans les approches volumiques on cherche à matérialiser un espace vide entouré par la macromolécule ; cet espace alloué à la poche peut ensuite être utilisé pour en calculer le volume, ou simplement pour déterminer les atomes qui le bordent et participent ainsi à la surface de la poche. Encore une fois, on distingue les approches exclusivement géométriques des approches prenant en compte des informations physico-chimiques.

Dans ces dernières, on peut distinguer les approches fondées sur le placement de molécules d'eau de celles basées sur un calcul énergétique. Certaines méthodes mettent à profit l'observation que les molécules d'eau sont moins mobiles dans les zones anfractuées à la surface d'une macromolécule qu'elles ne le sont dans l'espace du solvant [Chakravarty 02, Bhinge 04]. Dans ces approches la macromolécule est virtuellement solvatée, puis le système est soumis à une simulation de dynamique moléculaire à l'issue duquel les poches sont identifiées comme les regroupements de molécules d'eau ayant montré peu de mobilité. D'autres approches se basent sur des calculs énergétiques similaires à ceux employés pour évaluer et optimiser le positionnement d'un ligand dans une protéine au cours d'un processus d'amarrage virtuel (*docking*). Dans *Q-SiteFinder* [Laurie 05] l'énergie non liée d'un groupement méthyle est évaluée en plusieurs points de l'espace, et *DrugSite* [An 04] utilise un potentiel de Lennard Jones modifié prenant en compte uniquement la position du ligand et non sa nature. Les calculs sont systématiquement réalisés sur une grille de voxels et les poches sont matérialisées par des agrégats de voxels énergétiquement favorables.

Les approches basées sur des considérations exclusivement géométriques cherchent à identifier explicitement une zone de l'espace du solvant enclavée dans la molécule.

Une première approche pragmatique consiste à observer que l'enclavement d'un point de l'espace solvant peut être évalué en mesurant le nombre de directions dans lesquelles la visibilité est bloquée par une paire d'atomes de la molécule (voir figure 4.5 *A*). Ce genre d'approche repose systématiquement sur une discrétisation de l'espace, avec un calcul d'enclavement pour les voxels de l'espace solvant et une matérialisation des poches comme agrégats de voxels partageant une mesure d'enclavement supérieure à un seuil donné. Dans le logiciel *POCKET* [Levitt 92], la grille de voxels est parcourue trois fois, une fois dans la direction de chaque axe. A chaque passage un compteur est mis à jour dans chaque voxel, qui enregistre la survenue d'un événements PSP, c'est-à-dire la succession d'un segment de voxels codant l'espace du solvant encadré par deux segments de voxels codant des atomes de la protéine. Le sigle PSP est d'ailleurs resté pour désigner globalement ce type d'approches. Le même algorithme a été utilisé dans le logiciel

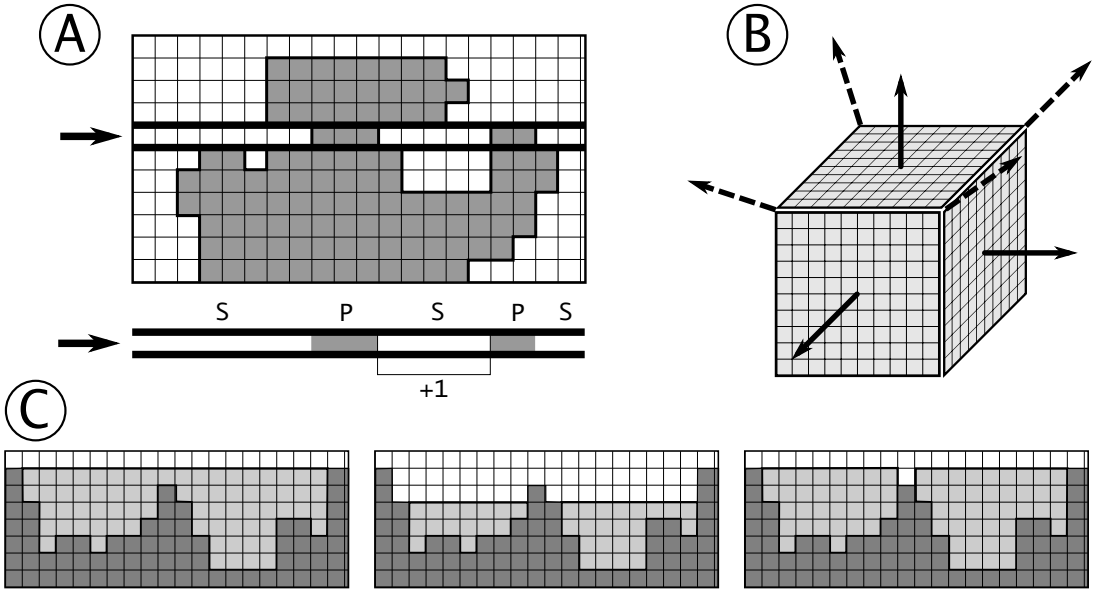


Figure 4.5: Detection des poches avec la stratégie PSP (protein-solvent-protein), une approche discrète.
A : dans cet exemple en deux dimensions la molécule est matérialisée par des pixels grisés. La ligne de la grille indiquée par une flèche est répétée sous la figure, et un événement PSP y est mis en exergue. Le compteur des pixels S encadrés par des pixels P est incrémenté.
B : une grille de voxels en 3-dimensions. Le logiciel *POCKET* cherche des événements PSP sur les trois axes principaux (flèches solides), *LIGSITE* ajoute les quatre diagonales représentées par les flèches en traits interrompus.
C : déconnexion de deux poches (pixels gris clair) dans une molécule (pixels gris foncé), un exemple en deux dimensions. De gauche à droite, la poche est érodée de deux pixels depuis le solvant, puis dilatée de deux pixels vers le solvant.

LIGSITE [Hendlich 97] où des mesures supplémentaires ont été prises dans les quatre diagonales du cube, essentiellement afin de moduler les mesures et de répondre aux problèmes d'instabilité des résultats en fonction de l'orientation de la molécule dans la grille de discrétisation (voir la figure 4.5 *B*). Ce même algorithme est aussi employé pour la détection de poches dans *SYBIL* [Exner 98], où une passe d'érosion-dilatation a été ajoutée pour supprimer les petites poches et les liaisons entre poches interconnectées par des isthmes non pertinents (figure 4.5 *C*).
Plus récemment, Weisel *et al.* ont proposé avec *PocketPicker* [Weisel 07] de mesurer "l'enfouissement" d'un voxel avec une approche monodirectionnelle : Depuis chaque centre de voxel, trente rayons cylindriques de 10Å de long et de 0.9Å de diamètre sont envoyés, et un compteur d'enfouissement retient le nombre de ces rayons rencontrant un voxel codant pour la molécule (voir figure 4.6). Dans *PocketDepth* [Kalidas 08] le problème du PSP est inversé : tous les voxels consti-

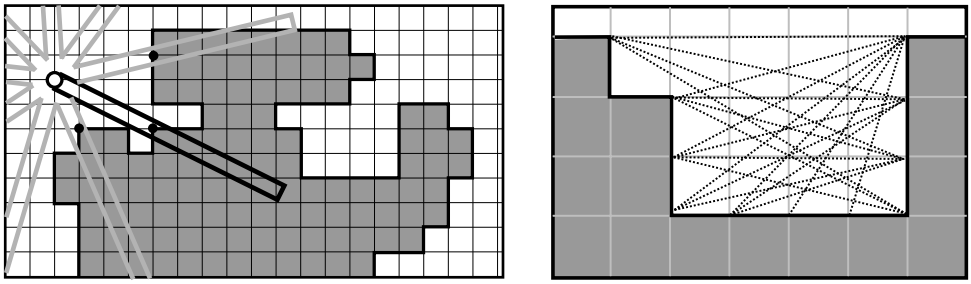


Figure 4.6: Détection des poches basée sur l'enfouissement, une illustration en deux dimensions. A gauche, la méthode de *PocketPicker* : trois des neuf rayons lancés depuis le pixel blanc rencontrent un pixel de la molécule, l'un d'entre eux est matérialisé en noir. A droite, la méthode de *PocketDepth* : le nombre de rayons joignant des pixels de surface est plus important au centre de la poche.

tuant la surface sont considérés, et des segments sont tracés entre chaque paire de tels voxels pour peu qu'ils soient situés entre 2 et 15Å l'un de l'autre et joignables au travers de l'espace du solvant[2]. L'enfouissement d'un voxel dans l'espace du solvant est mesuré par le nombre de segments traversant le voxel (voir figure 4.6). La notion de profondeur (le *depth* de *PocketDepth*) traitée ici, est elle aussi inversée : on ne s'intéresse pas à une profondeur à l'intérieur de la molécule en prenant sa surface pour référence, ni à une profondeur de cette surface par rapport à un référentiel comme dans *TravelDepth* [Coleman 06], mais on s'intéresse à la centralité d'un point à l'intérieur de la poche.

Une autre approche pragmatique repose sur la remarque qu'une enclave est "quelque chose que l'on peut remplir". L'idée générale consiste donc à proposer des solutions pour "remplir" explicitement les vides d'une molécule. Deux algorithmes différents ont été proposés exploitant cette vision : *SURFNET* [Laskowski 95] et *PASS* [Brady 00]. Dans *SURFNET*, toutes les paires d'atomes de la molécule sont considérées, et on cherche à placer une sphère-intersticielle (ou *gap sphere*) dans l'espace mitoyen. Seules les sphères-intersticielle de tailles adéquates (ni trop grandes, ni trop petites) sont conservées. Fusionnées entre elles, un ensemble de telles sphères présentant des intersections communes définissent le volume d'une poche (figure 4.7 en haut). *PASS* émule un processus itératif d'accrétion-élution d'un solvant virtuel (figure 4.7 en bas). A chaque étape de ce processus la phase d'accrétion consiste à répartir une nouvelle couche de sphères-solvant en contact géométrique autour du système, la phase d'élution consiste à retirer parmi les sphères nouvellement agrégées, celles qui ne sont pas assez enfouies (qui n'ont pas assez de voisines) parmi les sphères-atome de la molécules et les sphères-solvant.

Un autre type d'approche consiste à utiliser directement l'une ou l'autre des définitions géo-

[2] c'est-à-dire, que le segment n'intersecte pas d'atome de la molécule

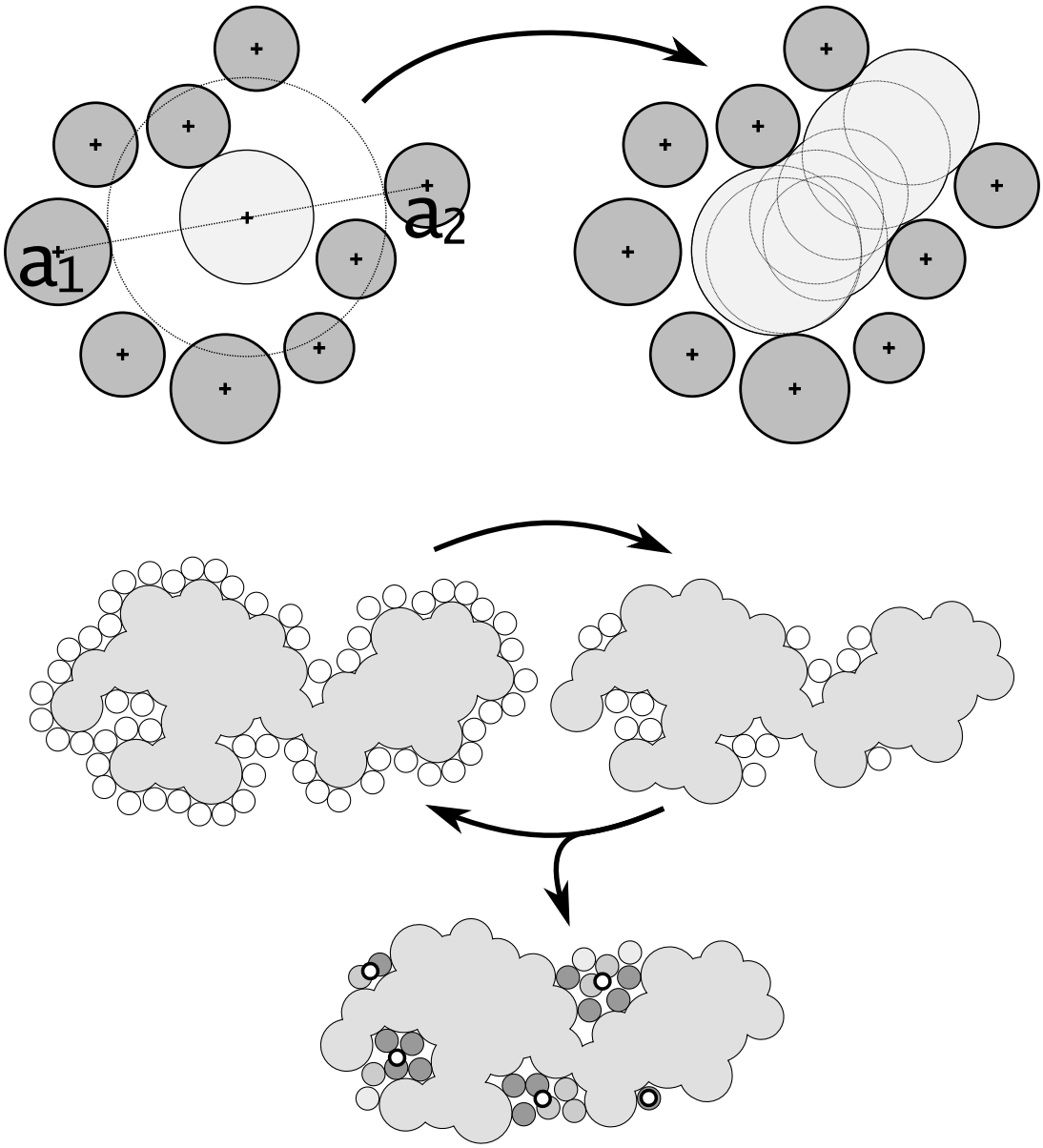


Figure 4.7: Détection des poches par des approches de remplissage.
En haut l'approche *SURFNET* : Les atomes de la molécule sont représentés par des disques gris foncé. Un cercle mitoyen aux atomes a_1 et a_2 intersecte six atomes de la molécule ; sa taille doit être réduite pour obtenir une sphère-intersticielle (en gris clair).
En bas, l'approche *PASS* : à gauche, des molécules de solvant virtuel sont réparties en contact avec la surface de la molécule. Seules les plus enfouies sont conservées (à droite), et les deux opération se répètent jusqu'à stabilité (en bas). Les poches sont matérialisées par un agglomérat de solvant virtuel ainsi acrété, ou par leur centre de masse (boule blanche cerclée de noir)

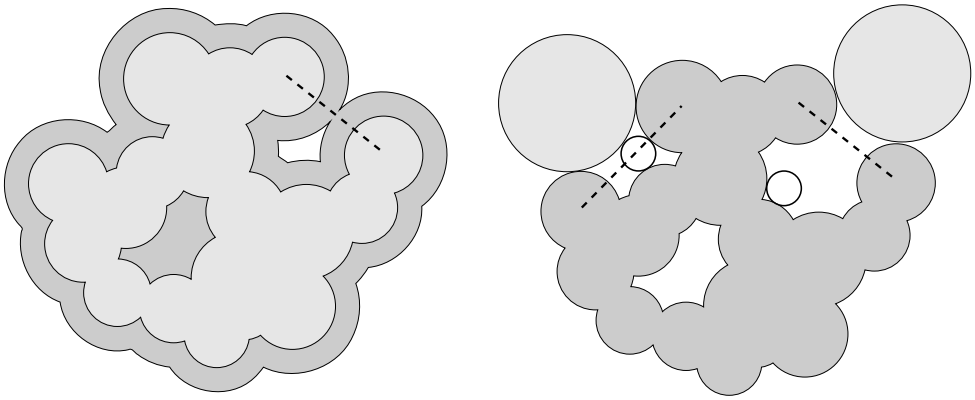


Figure 4.8: Detection des poches refermées et ouvertes, reformulation du problème sur un exemple en deux dimensions.
A gauche : En augmentant la taille des atomes de la molécule gris clair on obtient une autre union de boules (en gris foncé). La cavité dans la molécule foncée correspond à une poche refermée de la molécule gris claire. La bouche (segment pointillé) est matérialisée par le premier couple d'atomes qui, dans leur épaississement, closent la poche.
A droite : l'espace accessible à une petite sphère (blanche) et inaccessible à une grosse sphère (gris clair) permet de définir les poches ouvertes.

métriques strictes données dans la figure 4.1 de la section précédente.

La détection et la caractérisation des tunnels par exemple est généralement traitée par des techniques de "squeletisation" dans lesquelles on cherche le chenal le plus central à une poche. Une première approche a été proposée avec *HOLE* [Smart 93] pour la caractérisation des "pores" de la gramicidine-A. Elle consistait en une série de déplacements infinitésimaux basés sur des optimisations locales afin de maximiser la distance à un ensemble d'atomes. Une autre approche basée sur une discrétisation de l'espace du solvant a été proposée par Petrek *et al.* [Petrek 06] dans *CAVER*, où les voxels se voient attribuer une valeur dépendant de leur éloignement à la surface de la molécule, et le chenal central matérialisé par un algorithme de plus court chemin dans un graphe pondéré [Dijkstra 59]. Avec *MOLE*, les mêmes auteurs ont remarqué que le *1-squelette*[3] du diagramme de Voronoï constituait une bonne approximation du chemin central, et ont employé cette construction — associée encore une fois à l'algorithme de Dijkstra — à la caractérisation des chemins d'accès dans le ribosome [Petřek 07].

Les cavités totalement enfouies dans la molécule constituent un cas particulier souvent traité par effet de bord de nombreuses approches. Il a néanmoins fait l'objet d'une attention particulière de Liang *et al.* [Liang 98a] avec *CAST*, la première approche exploitant le rapport entre la forme duale d'une molécule et sa représentation Surface Accessible pour la détection des poches (voir section 6.1.2 pour une introduction à cette théorie, et la figure 6.3 page 63 où le lien avec les cavités est explicité.).

La détection des poches refermées a été effectuée de deux manières différentes, toutes deux basées sur la même remarque : supposer la présence d'une constriction dans la poche est équivalent à supposer qu'en augmentant virtuellement la taille des atomes composant la molécule, la poche, avant de se clore complètement deviendrait une cavité déconnectée de l'espace du solvant (voir figure 4.8). Dans *VOIDOO* [Kleywegt 94], des cycles de léger grossissement des atomes sont réalisés explicitement, et à chaque étape une recherche des cavités est effectuée dans une représentation discrète de la molécule. *CASTp* [Edelsbrunner 98, Binkowski 03b] et

[3] c'est-à-dire la structure composée des sommets et des arêtes de ce diagramme

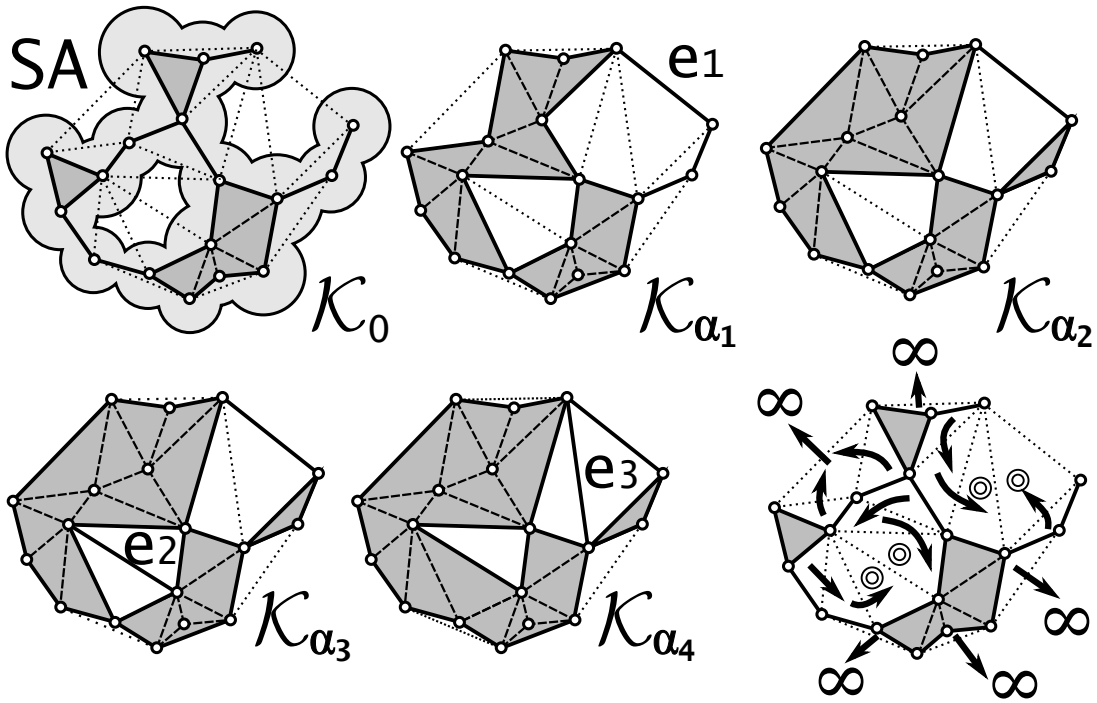


Figure 4.9: Détection des poches refermées avec le flux discret, un exemple en deux dimensions. En haut à gauche le complexe dual $\mathcal{K}_0$ superposé au modèle SA d'une molécule. Avec l'augmentation du paramètre α, le complexe-α décrit une famille de complexes duaux $\mathcal{K}_\alpha$ emboîtés. La bouche de la poche refermée, matérialisée par l'arête e_1 arrive assez rapidement dans le complexe. Dans $\mathcal{K}_{\alpha_3}$ l'arête e_2 déconnecte la cavité en deux sous cavités distinctes chacune composée d'un unique triangle. Chacun de ces triangles constitue le collecteur de la sous-poche qu'il représente. La même chose s'applique avec l'arrivée de e_3 dans $\mathcal{K}_{\alpha_4}$. De manière imagée, lorsqu'un triangle vide (*i.e* absent de $\mathcal{K}_0$) entre dans le complexe-α en même temps qu'un de ses côtés c, on dit qu'il s'écoule vers son voisin de l'autre côté de c. La dernière illustration reprend le complexe dual et matérialise l'écoulement (le flux discret) des triangles vides vers les quatre collecteurs marqués d'une pastille.

pocket [Edelsbrunner 03] utilisent le cadre de la théorie des formes-α (présentée succinctement à la page 13), et en particulier le principe de flux discret pour déterminer les bouches, c'est-à-dire ces constrictions qui finiront par déconnecter la poches de l'espace du solvant, ainsi que les tétraèdres collecteurs (ou *sinks*) qui seront les derniers à disparaître dans une sous-poche[4] (voir figure 4.9). Dans ce modèle, les sous-poches correspondent à des flux de tétraèdres (ou de triangles en deux dimensions) s'écoulant vers un collecteur. Concrètement, les poches et les cavités sont détectées en deux temps : en premier lieu, les tétraèdres vides s'écoulant dans la composante infinie sont itérativement détectés et supprimés, puis les poches et cavités sont définies comme les composantes connexes dans les tétraèdres vides restant. Le flux discret peut aussi être compris comme une restriction du flux défini par Giessen [Giesen 03, Dey 03] aux arêtes du diagramme de Voronoï duales d'un simplexe du complexe dual (voir figure 4.10). Le mode de grossissement des atomes est différent dans les deux approches ; avec *VOIDOO* il se fait proportionnellement à la taille initiale de l'atome, dans les approches basées sur la théorie des formes-α le grossissement se fait en puissance[5], les atomes les plus imposants grossissent un peu moins vite. De manière

[4]Concrètement, un collecteur est un tétraèdre qui entre dans le complexe-α pour une valeur de α strictement supérieur à celles pour lesquelles ses quatre facettes incidentes y entrent elles-mêmes.

[5]le rayon $r_{i,\alpha}$ d'un atome expensé est donné par $\sqrt{r_i^2+\alpha}$ comme il sera vu au chapitre 6 (page 64) ainsi qu'à l'annexe B.2 (page 178).

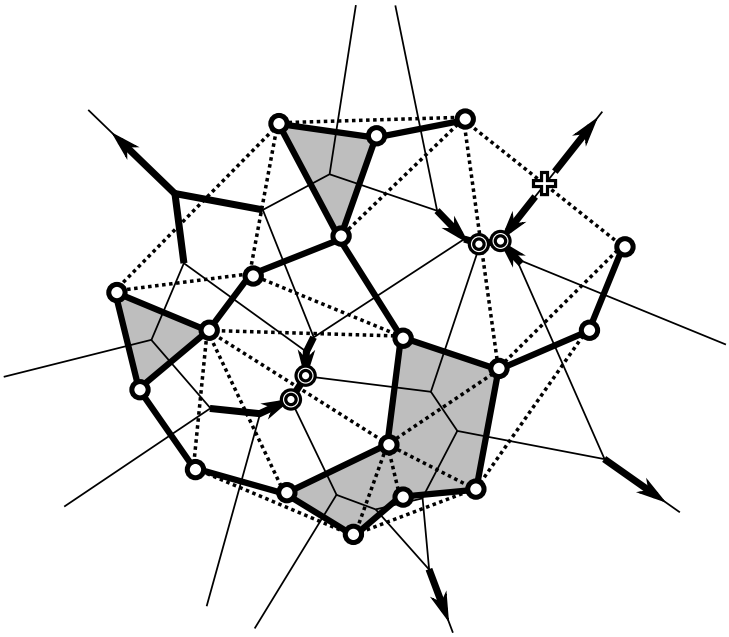


Figure 4.10: Flux sur les segments de Voronoï. Le diagramme de Voronoï de la molécule (arêtes fines) a été superposé à son complexe dual. Les arêtes de Voronoï indiquées en gras sont duales d'un simplexe extérieur au complexe dual. La fonction $\min_i \pi(p, a_i)$ qui a un point du plan associe sa plus petite puissance à une des sphères-atome de la molécule induit un flux sur les arêtes de Voronoï, représenté ici par des flèches. Les maxima locaux de ce flux sont les sommets de Voronoï duaux d'un collecteur, et le flux induit sur les arêtes correspond à la relation de flux discret présentée dans la figure 4.9

générale, le cadre théorique des formes-α permet une implémentation robuste et moins coûteuse en temps comme en espace mémoire. Un problème dont souffre *VOIDOO* et que la forme-α permet d'éviter, est la détection de deux sous cavités (voir figure 4.2). En effet, dans un processus de grossissement explicite des atomes il peut arriver qu'une première sous-cavité soit déconnectée avant une autre, plus proche du solvant.

De nombreuses approches pour la détection des poches ouvertes ont été proposées. Toutes se fondent sur une même formulation géométrique : une poche est un espace à la fois accessible à une sphère-solvant de petite taille, et inaccessible à une sphère-solvant de plus grande taille (voir figure 4.8). Une telle définition a déjà été évoquée concernant la caractérisation de la topographie d'une surface moléculaire sur des critères d'accessibilité (page 30). Cette formulation permet en fait la détection des poches ouvertes comme des poches refermées, et contrairement à la précédente elle dépend de deux paramètres : les tailles de la petite et de la grande sphère, rapprochées respectivement de la *taille* et de la *profondeur* de la poche par Kawabata *et al.* [Kawabata 07]. Masuya et Doi [Masuya 95] ont implémenté cette approche sur une grille discrète en remarquant que cette définition revenait à réaliser une différence entre deux surfaces de Connolly : celle générée par la grosse sphère-solvant, et celle générée par la petite, et qu'une surface de Connolly pouvait être obtenue par la succession d'une dilatation et d'une érosion (ces deux opérations ont été évoquées à la page 45 et explicitées dans la figure 4.5 *C* dans un contexte similaire). Dans un travail plus récent, Zhang et Bajaj [Zhang 07] ont la même approche, qu'ils accélèrent en mettant à profit un algorithme de *marching-front* basé sur une description de la surface en mélange de noyaux gaussiens. Kawabata *et al.* [Kawabata 07], eux, déterminent analytiquement le placement des sphères solvant. Enfin, les approches utilisées dans *SiteFinder* (un logiciel de détection de poche implémenté dans *MOE* mis à disposition par le *Chemical Computing Group* [CCG]) et

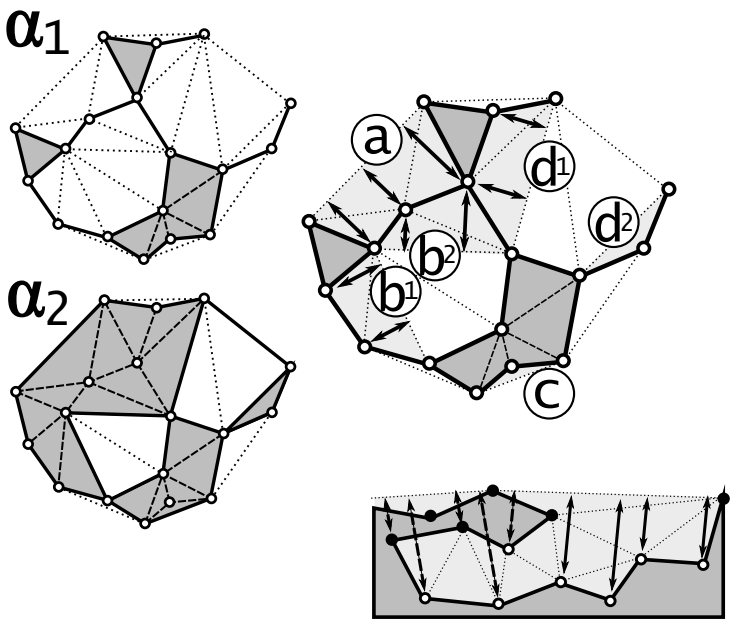


Figure 4.11: Détection des poches ouvertes dans *APROPOS*, un exemple en deux dimensions. A gauche, l'une au dessus de l'autre, deux formes-α d'une même molécule pour $\alpha_1 < \alpha_2$. A droite, en haut, les triangles gris clairs appartiennent à $\mathcal{A}_{\alpha_2}$ (le complexe-α pour α_2) mais pas à $\mathcal{A}_{\alpha_1}$ (le complexe-α pour α_1). Le bord de $\mathcal{A}_{\alpha_1}$ définit la surface de la molécule, en particulier ses sommets sont les atomes de surface. La distance d'un atome de $\mathcal{A}_{\alpha_1}$ à $(A)_{\alpha_2}$ est matérialisée par une flèche noire. Cinq poches sont ainsi découvertes dont deux dans la cavité, et deux dans la poche refermée. En bas à droite, un exemple montrant des mesures de distance à la surface effectuées au travers de la molécule (flèches pointillées). Les atomes de surface composant les poches sont représentés par des sommets blancs, les autres par des sommets noirs.

dans *Fpocket* [Guilloux 09] reposent sur la localisation de "sphères alpha" de taille acceptable (ni trop grandes ni trop petites), et la constitution d'agrégats de telles sphères.

Un problème communément rencontré avec cette reformulation du problème est le bruit généré par les trop nombreuses zones anfractuées de faible profondeur. En trois dimensions, la surface de la molécule est rapidement couverte de failles, puits et autres poches, interconnectés entre eux par un réseau de de petites vallées de faible profondeur. Une des problématiques consiste à déconnecter ces vallées. Dans *SCREEN*, Nayal et Honig [Nayal 06] utilisent une surface de Connolly construite avec une grosse sphère-solvant de 5Å comme "niveau de la mer" à partir de laquelle ils réalisent une déconnexion des différentes poches en ne considérant que les éléments atomiques à 2Å sous cette référence en hauteur. Cette dernière étape correspond en fait à une érosion. Dans *APROPOS* [Peters 96], la zone comprise entre deux tailles de sphères-solvant est interprétée comme une différence entre deux formes-α de la même molécule pour des valeurs $\alpha_1 < \alpha_2$ (voir figure 4.11). La déconnections des poches est réalisée sur des critères de distance entre les deux surfaces polyédriques : les poches sont définies comme l'ensemble des atomes du bord de $\mathcal{A}_{\alpha_1}$ (la forme-α pour α_1) qui sont suffisamment éloignés de $\mathcal{A}_{\alpha_2}$ (la forme-α pour α_2). Les poches sont définies en agglomérant les sommets de $\mathcal{A}_{\alpha_1}$ ainsi sélectionnés. Bien que l'approche initiale soit volumique, le résultat de cet algorithme donne uniquement une représentation surfacique des poches. En outre, le calcul des distances entre $\mathcal{A}_{\alpha_1}$ et $\mathcal{A}_{\alpha_2}$ n'est pas satisfaisant, et peut par exemple traverser la surface, induisant des erreurs dans la détection des atomes participant à des poches enfouies à proximité de la surface. Une approche plus récente [Kim 07] utilise la forme-β comme canevas pour gérer un grossissement linéaire des sphères. Cet analogue de la forme-α

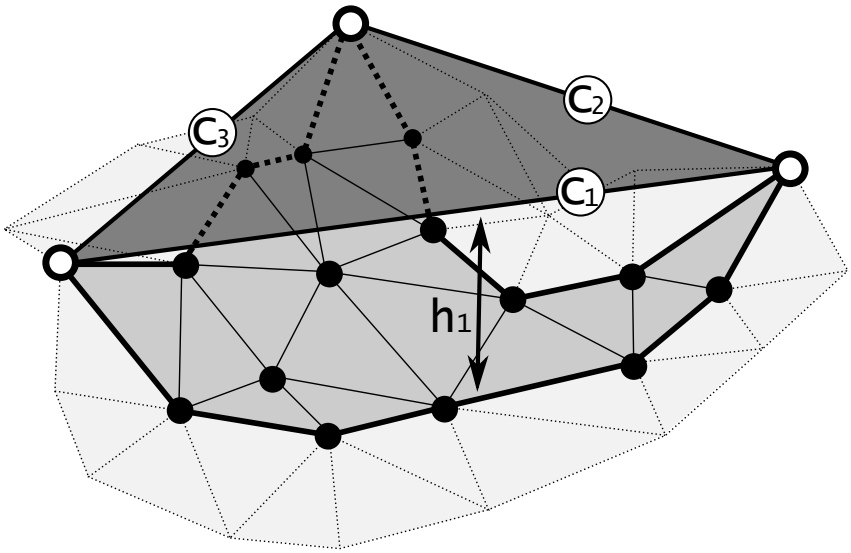


Figure 4.12: Détection de poche ouverte par projection d'un "niveau de la mer", un schéma en trois dimensions présentant l'approche proposée par Kim *et al.* [Kim 07]. Une forme-β_1 est représentée par des petits triangles : chaque triangle relie les centres de trois atomes qui bloquent une sphère-solvant de rayon β_1 à la surface de la molécule. Trois sommets de cette surface (marqués d'un point blanc) appartiennent aussi à une forme-β_2, ils sont reliés par un triangle (gris foncé) dans cette forme-β_2. Le plus court chemin d'un de ces sommets blancs à son voisin le long de la forme-β_1 (arêtes épaisses) décrit une vallée dont la hauteur moyenne au triangle de la forme-β_2 donne l'encaissement (h_1 pour le côté c_1). Les trois vallées encadrent une *primitive* composée de triangles de la forme-β_1 (en gris plus clair et sommets noirs). La vallée sous c_3 est trop "plate" et la primitive ne sera agglomérée qu'à ses deux primitives du côté de c_1 et de c_2.

pour les diagrammes d'Appolonius[6] permet d'étudier effectivement le lieu occupé par le centre des plus grandes sphère-sonde simultanément en contact avec 2, 3 ou 4 atomes de la molécule. A nouveau, les poches sont considérées comme l'espace emprisonné entre deux niveaux : une forme-β_1 modélisant la surface de la molécule et une forme-β_2 modélisant un "niveau de la mer" et correspondant exactement aux triplets d'atomes de surface bloquant la grosse sphère solvant (soit $\beta_1 < \beta_2$). Par opposition avec *APROPOS*, dans cette approche c'est la forme grossière qui est projetée sur la forme détaillée (voir figure 4.12). Pour chaque triplet d'atomes formant un triangle t sur la forme-β_2, on appelle *primitive* l'ensemble des triangles de la forme-β_1 enfouis sous t. Les trois bords d'une primitive forment des vallées, dont la profondeur moyenne est calculée en observant la mesure de hauteur à t. Les primitives sont ensuite agglomérées si elles partagent une vallée encaissée. Encore une fois, bien que reposant sur une expression volumique, cette approche aboutit à une description surfacique des poches. L'emploi de la forme-β est peut-être plus justifiable du point de vue du modèle, que celui de la forme-α ; le gain est cependant discutable en regard de la une chute conséquente des performances de calcul[7]. Enfin, la définition des poches par projection d'une enveloppe grossière sur une enveloppe plus fine n'est pas tout à fait aboutie. En particulier, l'existence d'un trou dans une primitive n'est pas traitée par les

[6]Le diagramme d'Appolonius [Aurenhammer 91, Boissonnat 98, Kim 06] est un diagramme de Voronoï d'un ensemble de sphères dans lequel on considère une "distance additive" : ses cellules décrivent le lieu des points qui sont plus proches d'une sphère (c'est-à-dire de son bord, et non de son centre) que d'une autre. Les frontières de ce diagramme sont des morceaux d'hyperbole.

[7]La complexité du calcul d'une forme β pour n atomes est de $O(n^3)$ [Kim 06, Kim 07], celui du complexe-α est de $O(n^2)$ [Edelsbrunner 94b].

auteurs.

4.4.3 Approches combinées

De nouvelles approches ont récemment vu le jour qui utilisent la détection de poches géométriques en combinaison avec d'autres techniques, soit pour affiner, trier ou catégoriser les poches géométriques détectées, soit pour détecter des poches dont l'existence pourrait n'être que transitoire en l'absence d'un ligand.

• Approches combinées géométrique, physico-chimique et conservation des résidus

Les approches géométriques pour la détection de poches génèrent souvent un nombre important de sites dont une partie peut être dégrossie sur des critères simples, comme la taille ; dans *SiteFinder* comme dans *Fpocket* c'est un score d'hydrophobicité qui est utilisé pour discriminer les agrégats de "sphères alpha" : dans *SiteFinder* les "sphères alpha" hydrophyles et éloignées de sphères hydrophobes sont retirées avant agrégation, dans *Fpocket* ce sont les agrégats trop chargés qui sont supprimés.

La taille de la poche peut aussi constituer un critère ; les poches dont le volume est inférieur à une centaine d'Ångström cube ne sont par exemple généralement pas considérées comme significatives pour la fixation d'un ligand de type médicament [An 04].

Dans d'autres cas les poches géométriques peuvent aussi s'avérer trop volumineuses pour être pertinentes dans le cadre d'une étude biologique, par exemple pour correspondre à un site de fixation du ligand. Dans ce genre de cas on peut avoir recours à une information complémentaire par exemple de nature physico-chimique pour restreindre le volume de la poche géométrique. Il a aussi été remarqué que le degré de conservation des résidus était plus élevé dans les sites d'interaction ; cette information est mise à profit dans *SURFNET-consurf* [Glaser 06] et *LIGSITE-csc* [Bingding 06] pour affiner les poches respectivement détectées par *SURFNET* et *LIGSITE*.

L'utilisation de propriétés physico-chimiques peut aussi être envisagée, non pour diminuer la taille d'une poche, mais pour la caractériser et ainsi inférer le type d'interaction et la nature du partenaire attendu (ligand, ADN, protéine,...) [Stahl 00].

• Approche combinée géométrique et dynamique moléculaire

Par nature, les approches géométriques sont statiques : elles permettent la détection de poches dans des structures figées, assimilables à un ensemble de boules maintenues dans une position donnée. Ces approches ne prennent pas en compte l'aspect dynamique des macromolécules biologiques, les phénomènes d'*adaptation induite*[8] ou plus simplement la "respiration" des macromolécules. De nouvelles approches ont vu le jour dernièrement, qui observent l'évolution des poches d'une protéine au cours d'une expérience de dynamique moléculaire[9]. Les premiers résultats obtenus avec de telles méthodes semblent indiquer que de nombreuses poches de taille conséquente ouvrent et ferment fréquemment des bouches dans des intervalles de temps courts (de l'ordre de la pico-seconde) [Eyrisch 07, Eyrisch 09]. De plus, certaines poches connues pour fixer un ligand semblent ne se former en tant que poche géométrique détectable[10] que de manière

[8] Phénomène dans lequel les molécules interagissante modifient leur forme pour maximiser leur complémentarité.

[9] La dynamique moléculaire est une technique permettant de simuler l'évolution de la structure d'une molécule (la position de ses atomes dans le temps) au travers du temps.

[10] Plus précisément, détectables par l'algorithme *PASS* utilisé pour mener ces investigations

transitoire. Cette constatation constitue une motivation certaine pour le développement de ce genre d'approches.

4.5 Difficultés inhérentes à la détection des poches et intérêt des méthodes géométriques

Le nombre croissant des solutions proposées au problème de la détection et de la caractérisation des poches dans les macromolécules biologiques au cours de la dernière décennie témoigne à la fois de l'importance et de la difficulté de cette problématique. Une des explications au nombre considérable des algorithmes destinés à la détection des poches réside certainement — comme pour le cas de la problématique de la caractérisation topographique de la surface moléculaire — dans l'impossibilité de proposer une solution universelle.

En premier lieu, la définition du terme de poche est en effet elle-même sujette à interprétation suivant la motivation finale de l'utilisateur (détection ou prédiction du site de fixation d'un ligand, échange de substrat, site catalytique, étude d'espaces vides pour inférer la plasticité d'une protéine...).

L'approche géométrique permet de préciser le terme ambigu de "poche" par des définitions à priori strictes (tunnels, cavités, poches refermées, poches ouvertes). Cependant, même dans ce cadre à première vue rigoureux, la notion de poche demeure élusive, en particulier dans le cas des poches présentant une ouverture sur l'espace du solvant, et pour lesquelles la question de la limite de l'espace alloué à la poche est très souvent arbitraire. Les modèles issus de la théorie des formes-α offrent un cadre adapté à la clôture des poches, autrement dit, à l'allocation d'un volume physique et à la définition de bouches[11].

[11] Historiquement, l'algorithme d'H. Edelsbrunner est le premier à avoir permis une définition stricte ainsi qu'une caractérisation des bouches [Edelsbrunner 98].

Chapitre 5

Motivations et objectifs de notre étude

Notre travail s'inscrit dans le cadre de la bioinformatique structurale, et consiste à proposer de nouveaux outils pour faciliter l'analyse des structures des macromolécules biologiques.

Dans ce domaine on peut distinguer deux grands types de besoins suivant que l'utilisateur souhaite examiner spécifiquement une structure dans le cadre d'une analyse visuelle détaillée, ou qu'il souhaite réaliser de nombreuses études simultanément (ou séquentiellement) sur un ensemble de structures, par exemple pour caractériser et comparer un grand nombre de structures entre elles. Dans le premier cas, l'accent est mis sur l'exploitation visuelle et l'interaction avec l'utilisateur. La vitesse d'exécution n'est pas d'une importance capitale tant qu'il n'est pas nécessaire de répéter souvent les calculs, mais elle constitue tout de même une qualité appréciable. Dans le second cas par contre, motivée par un nombre de plus en plus important de structures dans la PDB (par exemple dans des travaux impliquant la comparaison deux à deux de toutes les structures contenues dans un jeu de données exhaustif [An 05, Gold 06]), ou par un grand nombre de structures à analyser dans une étude en dynamique moléculaire [Eyrisch 07, Eyrisch 09], la vitesse d'exécution s'avère au contraire critique. Le développement de ce type d'approches requiert la prise en compte autant que possible de l'aspect vitesse dans les développements d'outils calculant des propriétés descriptives des structures macromoléculaires.

Évoquée dès l'introduction, deux problématiques s'avèrent d'un intérêt particulier dans une analyse structurale : la caractérisation de la forme d'une macromolécule, et la définition et l'étude de ses poches. Ces problématiques, quoique proches[1], sont souvent séparées dans la littérature, et l'on peut se hasarder ici à caractériser ce qui les distingue l'une de l'autre : de notre étude bibliographique il ressort que la topographie constitue généralement une caractéristique intrinsèque de la surface qui en décrit le "relief", et autorise la définition des notions idiosyncratiques de "creux" et de "bosses" ; quant à la notion de "poche", si elle répond elle-aussi à l'intuition de "creux", elle a généralement trait à la matérialisation d'un volume physique[2] à la fois extérieur à la molécule, et partageant tout ou partie de sa surface. Cette interprétation "volumique" du terme de poche est d'ailleurs indépendant de la construction effective de ce volume. Dans les faits, et pour faire le lien avec l'analyse topographique, le terme de "poche" désigne aussi des "creux" particulièrement anfractués, voire totalement enfouis dans le corps de la molécule.

Parmi les approches pour la conduite de ce type d'étude, la géométrie algorithmique apparaît comme un cadre particulièrement adapté, en témoignent les nombreux travaux dans lesquels elle a été employée, et la multiplication de ceux-ci ces dernières années. Le succès de ce type d'approche s'explique en premier lieu par le lien unissant ces modèles au diagramme à remplissage de forme

[1] Après tout, une poche n'est-elle pas une variété de creux à la périphérie de la molécule ?

[2] Eargle *et al.* parlent d'"espace dual" [Eargle 06].

et aux autres modèles moléculaires (notamment Van der Waals et Surface Accessible). Le cadre à la fois simple et rigoureux qu'ils offrent pour la description d'algorithmes, ainsi que l'existence d'implémentations robustes et rapides pour leur construction constitue un second intérêt de ces modèles. Un autre aspect intéressant réside dans l'unicité de ces constructions pour un ensemble d'atomes fixé ; cette propriété garantit la reproductibilité des résultats pour deux exécutions successives dans une configuration donnée, ce qui n'est pas forcément le cas par exemple avec les approches discrètes (sur ce sujet on poura consulter l'annexe E).

Dans ce cadre, les objectifs de notre travail de thèse ont consisté dans le développement de nouveaux outils théoriques et pratiques pour la caractérisation de la topographie de la surface des macromolécules biologiques (cet aspect sera abordé au chapitre 8), ainsi que pour la détection et la caractérisation des poches (qui sera abordée au chapitre 10). Nos algorithmes ont été implémentés et sont librement accessibles à la communauté sous la forme d'exécutables pour linux (version 32 bits et 64 bits) et pour windows (on trouvera une description des logiciels en annexe D). Nos implémentations sont rapides, et de ce fait utilisables dans le cadre de projets à grande échelle ; afin de faciliter l'analyse interactive, nous avons en outre fourni des greffons au logiciel de visualisation de molécules VMD [Humphrey 96].

Deuxième partie

Modèles et méthodologie

Chapitre 6

Les modèles géométriques et leurs implémentations

Nos travaux reposent sur des objets géométriques — succinctement exposés dans l'état de l'art du présent document (voir page 10) — et que nous présenterons plus en détail dans la première section de ce chapitre. Des algorithmes permettant la construction et la manipulation de ces objets ont été développés dans les années 90, mais leur implémentation reste encore une affaire de spécialiste ; nous présenterons rapidement dans la seconde section les difficultés inhérentes à ce domaine, et donnerons une liste d'implémentations proposées. Parmi ces efforts, la bibliothèque `CGAL` constitue la "boîte à outil" la plus aboutie tant en terme de conception que du nombre de problématiques adressées. Nos développements étant entièrement basés sur cette bibliothèque, nous en avons donné une présentation en annexe F page 195 où nous nous concentrons sur ses particularités de conception et sur les trois modules que nous avons plus spécifiquement utilisés.

6.1 Modèles géométriques

Cette section regroupe la présentation des modèles que nous avons utilisés dans notre étude. Nous avons souhaité mettre l'accent sur l'utilisation de ces modèles dans le cadre de la bioinformatique structurale, aussi avons-nous opté pour une présentation différente de celles faites habituellement dans la littérature, plus axée sur les aspects pratiques que sur les aspects théoriques. La première sous-section est ainsi dévolue à la présentation du diagramme à remplissage de forme, couramment utilisé pour attribuer un espace disjoint à chaque atome d'une molécule ; le complexe dual et la forme duale sont décrits dans la sous-section suivante à partir de ce premier modèle, et utilisés dans la troisième sous-section pour définir le complexe-α et la forme-α. Les notions de triangulation et de simplexe, plus théoriques, sont définies dans la quatrième sous-section ; la cinquième sous-section étant plus spécifiquement consacrée à l'introdution de la triangulation de Delaunay. Pour une vision plus formelle de ces objets on pourra par exemple se référer aux publications et rapports d'Edelsbrunner [Edelsbrunner 95a, Edelsbrunner 94b, Edelsbrunner 92, Edelsbrunner 96].

Enfin, la sous-section 6.1.6 présente la notion de surface variété orientable, qui autorise une définition formelle de l'intuition de surface "lisse". La structure de données de demi-arêtes — couramment utilisée pour représenter cette classe de surface, et que nous utiliserons au chapitre suivant pour construire la *surface duale* — y sera présentée.

Mis à part le diagramme à remplissage de forme, tous les modèles présentés ici sont des constructions combinatoires, c'est-à-dire des structures composées d'éléments (sommets, arêtes,

facettes,...) reliés entre eux par des relations d'incidence ou d'adjascence.

6.1.1 Diagramme à remplissage de forme

Dans les modèles Van der Waals et Surface Accessible, chaque atome a_i d'une molécule A est représenté par une boule $b_{a_i} = (a_i, r_i) = \{x \mid d(x, a_i) \leq r_i\}$, et l'union de ces boules modélise l'espace de la molécule. Comme schématisé dans la figure 6.1 A, de nombreuses intersections ont lieu entre les boules modélisant les atomes d'une molécule.

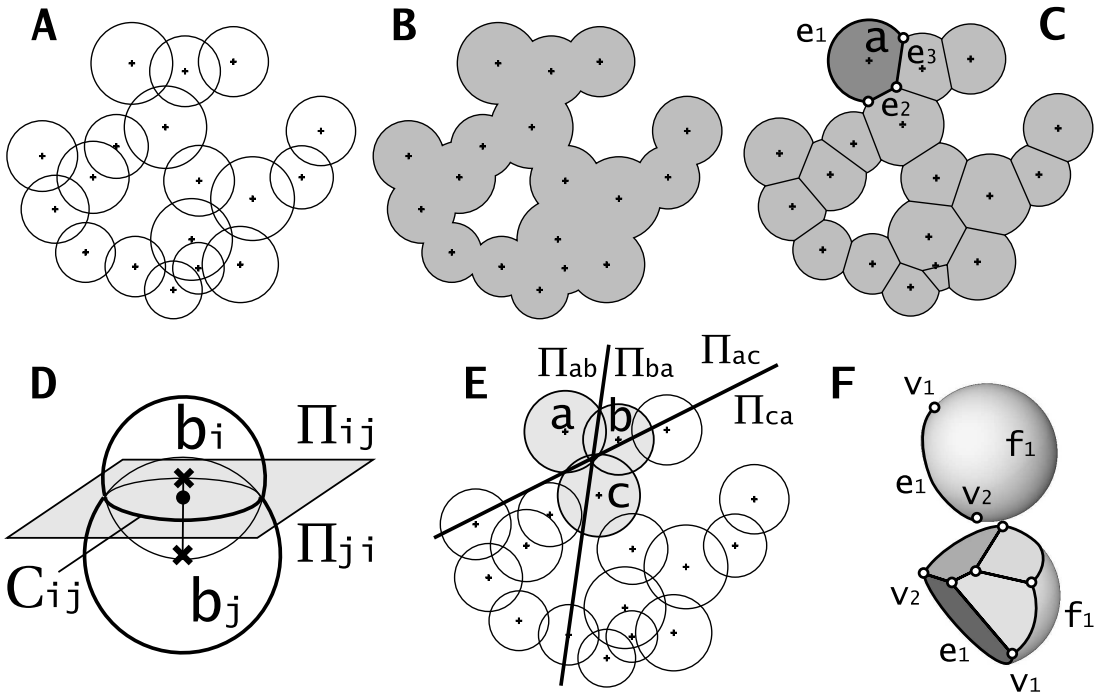


Figure 6.1: Union de boules et diagramme de remplissage de forme. Les schémas A, B, C et E sont en deux dimensions, les schémas D et F sont en trois dimensions. (A) Une molécule modélisée par un ensemble de boules présentant de nombreuses intersections communes. (B) L'union des boules-atomes modélise la molécule. (C) Diagramme en remplissage de forme de la molécule. (D) L'intersection des sphères constituant le bord des boules b_i et b_j forme le cercle $\mathcal{C}_{ij}$. Le plan portant ce cercle sépare $\mathbb{R}^3$ en deux demi-espaces Π_{ij} et Π_{ji}. (E) Construction de l'atome tronqué a dans le diagramme à remplissage de forme à partir de l'atome a de la molécule et de ses voisins b et c. (F) Deux vues tri-dimensionnelles d'un même atome tronqué sous deux angles différents.

Une décomposition en *diagramme à remplissage de forme* (aussi appelée *représentation compacte*, ou *spacefilling diagram*) réalise une union disjointe de l'union des boules-atome en attribuant à chaque atome le volume qui lui est propre ainsi qu'une partie du volume qu'il partage avec ses voisins (figure 6.1 C). Concrètement, lorsque deux boules b_i et b_j s'intersectent sans que l'une soit totalement incluse dans l'autre (voir figure 6.1 D et E), l'intersection de leur bord forme un cercle $\mathcal{C}_{ij}$, et le plan portant ce cercle scinde l'espace en deux demi-plans Π_{ij} et Π_{ji} utilisés pour attribuer l'espace appartenant à chaque boule : chaque boule est intersectée avec son demi-plan pour obtenir un *atome tronqué*. Les boules ainsi tronquées modélisent l'atome dans son contexte moléculaire ; leurs bords sont composés de facettes planes correspondant aux intersections avec d'autres atomes, et éventuellement de facettes sphériques participant à la surface de la molécule. Dans l'exemple en deux dimensions de la figure 6.1 C, l'atome tronqué a présente une partie sphérique e_1 correspondant à une contribution de l'atome à la surface de la

molécule, et des parties planes e_2, e_3 enfouies sous la surface de la molécule et correspondant aux intersections de l'atome a avec ses deux voisins. La figure 6.1 E montre la manière dont les frontières de l'atome tronqué a sont constituées : l'atome a est intersecté avec les deux demi-plans Π_{ab} et Π_{ac} obtenus respectivement à partir de l'intersection de a avec b et avec c. L'exemple en trois dimensions de la figure 6.1 F, montre un atome tronqué présentant une facette sphérique f_1 contribuant à la surface de la molécule, et quatre facettes planes enfouies et constituant la limite de cet atome avec ses quatre voisins.

Les frontières planes délimitées par le diagramme à remplissage de forme correspondent aux frontières des cellules de puissance, les cellules de Voronoï pour un analogue pondéré du diagramme de Voronoï présenté en annexe B.2 page 178.

6.1.2 Complexe dual et forme duale

Introduit au milieu des années 90 [Edelsbrunner 95a], le complexe dual (d'une molécule) offre une représentation moléculaire simplifiée. Il se définit à partir d'un ensemble de boules et encode leurs intersections observées, s'avèrant ainsi dual des modèles moléculaires usuels (Van der Waals, Surface Accessible et Surface Moléculaire). La forme duale (d'une molécule) — définie comme l'espace sous-jacent du complexe dual (de la même molécule) — consiste en un polyèdre conservant la topologie[1] et la forme de la molécule. Ses arêtes, ses sommets et ses facettes (triangulaires) enregistrent les intersections des boules constituant la surface de l'union de boules considérée. Lorsque l'on souhaitera mettre cet aspect en exergue, on parlera ainsi de *modèle polyédrique* ou de *représentation polyédrique* pour désigner ces modèles.

Définitions du complexe dual et de la forme duale

Le complexe dual est défini à partir du diagramme de remplissage de forme : à chaque atome tronqué on associe un sommet dans le complexe dual. Ces sommets sont reliés entre eux pour former des arêtes, des facettes (triangulaires) et des tétraèdres si les atomes tronqués correspondants dans le diagramme de remplissage de forme ont une frontière commune. La figure 6.2 A montre un exemple d'une telle construction en deux dimensions. Dans cet exemple, tous les centres atomiques (marqués d'un disque blanc) sont des sommets du complexe dual, c'est en particulier le cas des atomes a, b, c et d. Les arêtes ab, ac, ad, bc et cd appartiennent au complexe dual, de même que le triangle abc, qui correspond au sommet enfoui de l'atome a marqué d'une croix et partagé par les atomes b et c.

Les sommets, arêtes, facettes triangulaires et tétraèdres composant le complexe dual sont appelés des *simplexes*, et correspondent respectivement à un atome tronqué, une de ses facettes planes, une de ses arêtes droites ou un de ses sommets enfouis ; un support visuel en trois dimensions est fourni dans la figure 6.2 B. Les quatre facettes planes constituent chacune une frontière avec un autre atome, et correspondent ainsi chacune à un segment joignant cet atome à un autre dans le complexe dual. Les cinq arêtes droites (représentées en gras) sont enfouies sous la surface et sont communes à trois atomes ; elles correspondent ainsi chacune à une facette triangulaire reliant trois sommets dans le complexe dual. Les deux sommets enfouis (marqués d'une croix), sont communs à quatre atomes et correspondent donc chacun à un tétraèdre dans le complexe dual.

Les simplexes du bord du complexe dual peuvent être *réguliers* ou *singuliers* suivant qu'ils sont ou non sur le bord d'un autre simplexe. Les simplexes qui ne participent pas du bord sont

[1] Pour être exact, il faudrait plutôt parler de type d'homotopie.

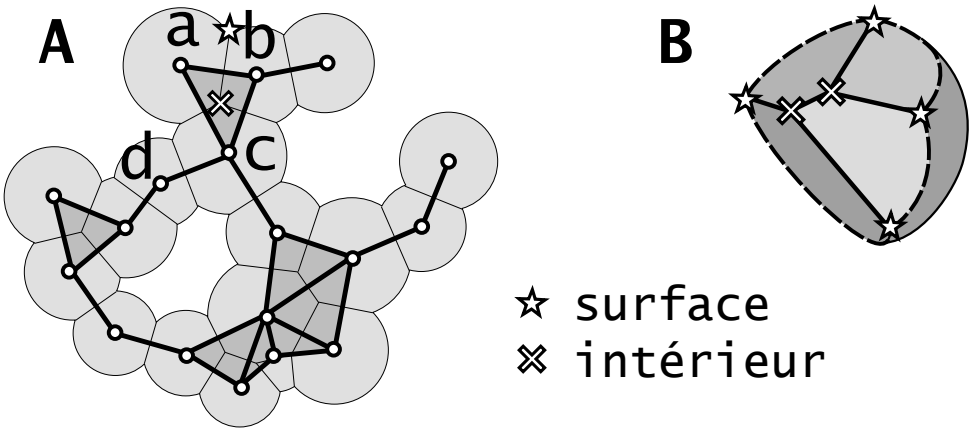


Figure 6.2: Complexe dual d'une molécule. (A) Le complexe dual d'une molécule est superposé à son diagramme en remplissage de forme. (B) Représentation d'un atome tronqué dans le diagramme à remplissage de forme. Les sommets, facettes et arêtes de cet atome sont duaux de simplexes du complexe dual (voir texte).

dits *intérieurs*. La figure 6.3 explicite cette classification des simplexes pour le cas de l'exemple en deux dimensions présenté précédemment (figure 6.2 A). Tous les exemples moléculaires de ce document adopteront le même code couleur : bleu pour les arêtes et les facettes régulières, jaune pour les facettes singulières (ou, dans les exemples en deux dimensions, pour les arêtes singulières), et en trois dimensions, rouge pour les arêtes singulières. La figure 6.4 montre deux exemples concrets sur les molécules déjà exposées dans la figure 2.2 au chapitre précédent page 21. Le complexe dual est une structure combinatoire, elle consiste en la donnée d'un ensemble de simplexes et de relation (incidence et adjascence) qui les unissent. La *forme duale* de la molécule désigne le volume physique occupé par les simplexe les simplexes du complexe dual dans l'espace, c'est donc l'union des simplexes du complexe dual.

Propriétés du complexe dual et de la forme duale

En même temps qu'il en a donné la définition, H. Edelsbrunner a démontré que la forme duale avait le même type d'homotopie[2] que l'union des boules à partir de laquelle elle est construite [Edelsbrunner 95a]. Dans le schéma en deux dimensions de la figure 6.3, la flèche en traits interrompus exhibe une telle correspondance entre une cavité dans les deux modèles.

Moins formellement, la forme duale conserve également — et de manière assez fine — la "forme" de la molécule. Dans la figure 6.3, les flèches en pointillés montrent la correspondance entre deux enclaves à la surface de la molécule et à la surface de la forme duale. Cette similarité de forme, peut aussi être observée sur les exemples moléculaires de la figure 6.4.

Par définition, le complexe dual constitue un encodage "simplifié" du diagramme à remplissage de forme : il contient l'information de connectivité entre les atomes tronqués ; son bord est un encodage de la surface de l'union de boules. Le bord du complexe dual est en effet composé d'arêtes et de facettes triangulaires comprenant exclusivement des sommets dont les atomes correspondants participent à la surface de la molécule dans le modèle à remplissage de forme. Cet aspect sera utilisé dans nos travaux pour décrire la surface de la molécule et sa forme, à commencer par la définition de la surface duale au chapitre 7.

[2]Ce qui revient à dire que les deux objets ont le même nombre de composantes connexes, de cavités enfouies sous la surface et de tunnels au travers de leur surface ; voire que ces propriétés se correspondent dans les deux modèles.

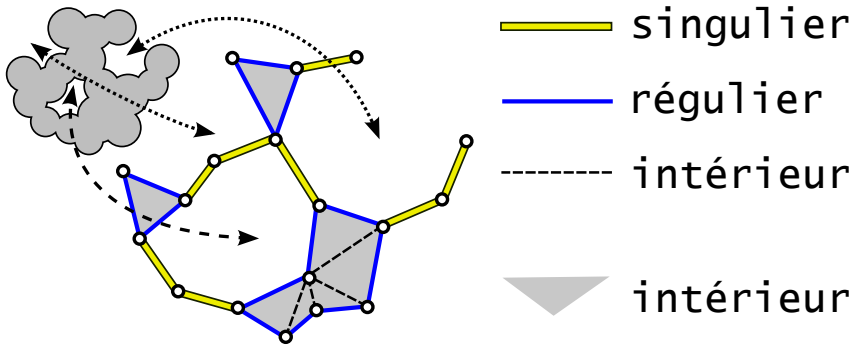


Figure 6.3: Correspondance de forme et de topologie entre une molécule et son complexe dual, un exemple en deux dimensions. L'union des boules et son complexe dual sont représentés séparément.
Les arêtes régulières sont colorées en bleu, les arêtes singulières en jaune, et les arêtes intérieures sont représentées en traits noirs interrompus. Les triangles gris appartiennent au complexe dual.
La correspondance topologique entre la cavité enfouie dans les deux modèles est matérialisée par une flèche en traits interrompue. Les flèches en pointillés indiquent des concavités révélées à la surface de chaque modèle et qui se correspondent grâce à l'analogie de forme des deux modèles.

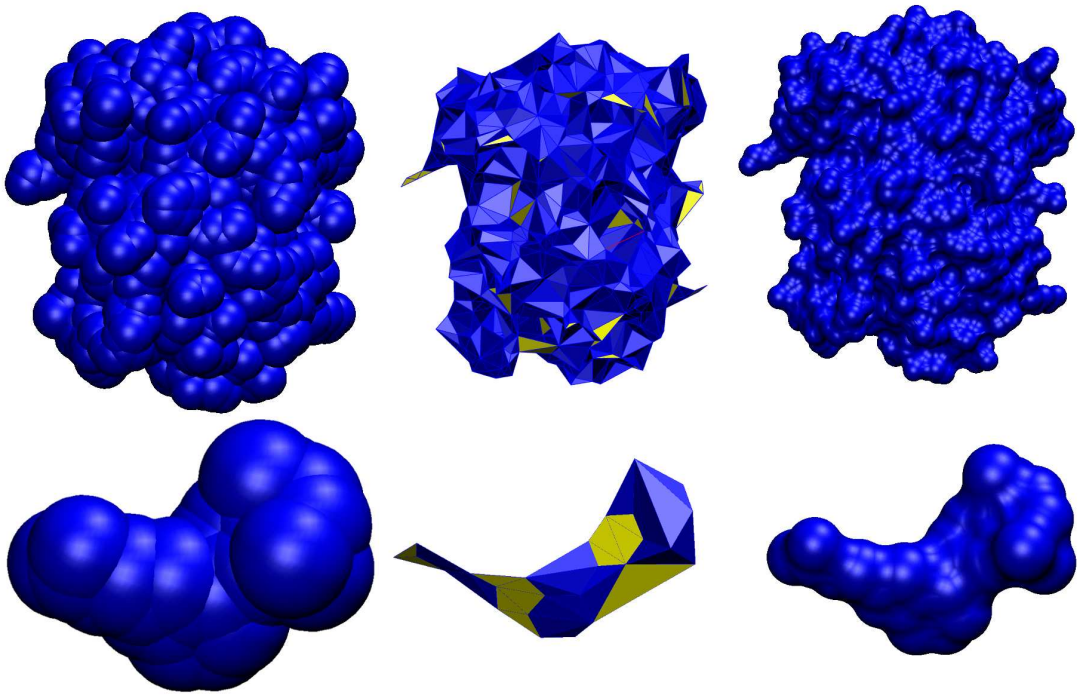


Figure 6.4: Complexe dual d'une molécule, deux exemples en trois dimensions. En haut le domaine de liaison au ligand du récepteur nucléaire RXR-α (1RDT), en bas le ligand de ce récepteur nucléaire dans la même structure. De gauche à droite, la représentation Surface Accessible, la forme duale associée, et la surface de Connolly.

Par souci de justesse, il convient de remarquer que deux aspects ont été éludés jusqu'à présent : le fait que certains atomes ne sont associés à aucun sommet du complexe dual, et le fait que le complexe dual est composé uniquement d'arêtes, de triangles et de tétraèdres (pas de carrés, de cubes...).

La première propriété est inhérente au fait que dans le diagramme en remplissage de forme, certains atomes sont tellement masqués par leurs voisins qu'ils disparaissent dans le processus de troncature ; ils sont appelés *points redondants* ou *atomes redondants* et n'apparaissent pas dans les sommets du complexe dual. En l'absence d'atomes d'hydrogène dans une structure (en utilisant un modèle intégré et des rayons en conséquence), ces cas sont anecdotiques même dans le modèle Surface Accessible. Avec des atomes d'hydrogène explicites, ces cas sont beaucoup plus fréquents et l'utilisateur doit être conscient de cet état de fait qui peut s'avérer problématique pour son cadre d'utilisation. Cet aspect est observé plus avant en annexe B.5 page 184.

Par définition du complexe dual, la seconde propriété revient à supposer que l'ensemble des boules est en *position générale*, c'est-à-dire qu'une même arête d'un atome tronqué ne peut-être partagée par plus de trois atomes, de même qu'un même sommet d'un atome tronqué ne peut être partagé par plus de quatre atomes tronqués. Cette propriété est bien entendu fausse dans les faits : par exemple, les six atomes d'un cycle aromatique s'intersectent sur un axe commun dont une partie composera une arête commune aux six atomes tronqués. La facette duale de cette arête devrait donc être un hexagone. Néanmoins, pour faciliter et rendre les modèles génériques, on suppose que ces cas — dits *dégénérés* — n'arrivent jamais, et des techniques existent pour ramener les cas dégénérés à des positions générales. Ces techniques reposent généralement sur l'utilisation d'une perturbation symbolique [Edelsbrunner 90] de l'ensemble de boules, c'est-à-dire la simulation d'un déplacement infinitésimal de ces boules pour se placer dans une position générale.

6.1.3 Complexe-α et forme-α

Le complexe-α et la forme-α permettent d'étudier l'évolution du complexe dual et de la forme duale pour un ensemble de boules dont la taille (les rayons des boules) varie en fonction d'un paramètre α. En particulier, ces constructions permettent d'appréhender la forme d'une molécule à différents niveaux de détail.

Définitions du complexe-α et de la forme-α

De manière évidente, deux ensembles de boules constitués à partir d'un même ensemble de centres atomiques mais pourvus de rayons différents ont généralement des complexes duaux différents. C'est par exemple le cas pour les représentations Van der Waals et Surface Accessible d'une même molécule, comme on peut le voir dans la figure 6.5. Cet exemple met aussi en évidence les propriétés du complexe dual de chacun de ces deux modèles : le complexe dual du modèle de Van der Waals reflète essentiellement les liaisons de covalences à l'intérieur de la molécule ; et s'avère beaucoup plus "lacunaire" que le complexe dual du modèle Surface Accessible. De fait, les facettes composant le bord de ce dernier représentent chacune le lieu où une sphère-solvant est bloquée par les trois atomes composant la facette.

Une méthode particulière de modification collective des atomes — dite "croissance en puissance" — permet toutefois de garantir une relation d'inclusion entre les complexes duaux de l'ensemble initial de boules, et de l'ensemble des boules modifiées. Étant donné un ensemble de boules $A = \{(a_i, r_i)\}_i$ et un paramètre réel α, on construit un nouvel ensemble de boules

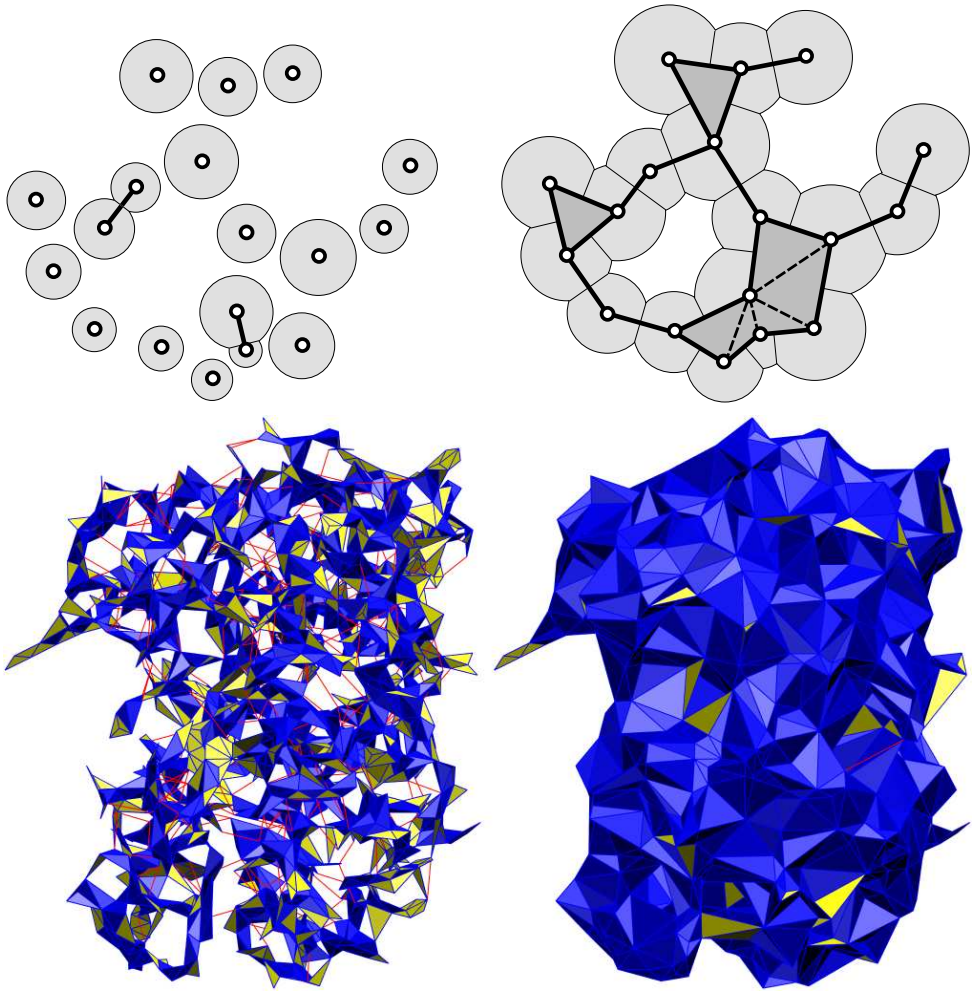


Figure 6.5: Complexes duaux des modèles Van der Waals (à gauche) et Surface Accessible (à droite). En haut, un exemple en deux dimensions, en bas un exemple en trois dimensions avec le domaine de liaison au ligand du récepteur nucléaire RXR (1RDT).

$A_\alpha = \{(a_i, r_{i,\alpha})\}$ en modifiant simultanément tous les rayons r_i en $r_{i,\alpha} = \sqrt{r_i^2 + \alpha}$. Comme illustré en deux dimensions dans la figure 6.6, les boules tronquées dans les représentations en remplissage de forme de A et de A_α partagent leurs frontières planes. Plus généralement, étant données deux valeurs $\alpha_1 < \alpha_2$, pour chaque atome a_i la boule tronquée associée à a_i dans A_{α_2} contient la boule tronquée associée à a_i dans A_{α_1}. De plus, les bords plans de la boule tronquée associée à a_i dans A_{α_1} sont inclus dans les bords plans de la boule tronquée associée à a_i dans A_{α_2}. Pour cette raison, le complexe dual $\mathcal{K}_{\alpha_2}$ contient le complexe dual $\mathcal{K}_{\alpha_1}$. Avec la croissance de α, les boules tronquées remplissent des polyèdres correspondant aux cellules du diagramme de Voronoï $\mathcal{V}$ associé à l'ensemble de sphères $A = A_O$. Ce diagramme de Voronoï est d'ailleurs partagé par tous les A_α. Pour une valeur N suffisamment grande du paramètre α, la forme duale remplit intégralement l'enveloppe convexe des centres atomiques a_i, tandis que le complexe dual $\mathcal{K}_N$ morcelle cet espace en une partition de tétraèdres bien particulière appelée triangulation de Delaunay (de l'ensemble d'atomes A).

La suite des différents complexes duaux obtenus avec des valeurs différentes de α est appelée *complexe-α* ; par abus de langage on parle aussi de complexe-α pour désigner le complexe dual correspondant à une valeur de α donnée. On désigne enfin par *forme-α* la suite des formes duales correspondant au complexe-α. Comme illustré plus concrètement dans la figure 6.7, ces constructions permettent d'appréhender la forme d'une molécule à différents niveaux de détail.

Pour finir, on notera le complexe dual d'une molécule n'est autre que son complexe-α pour la valeur $\alpha = 0$.

6.1.4 Triangulation

Étant donné un ensemble de points $A = \{a_i\}_i$, le terme de *triangulation* désigne une partition[3] décomposant une partie du plan ou de l'espace en un ensemble de triangles (lorsqu'on travaille dans le plan) ou de tétraèdres (lorsqu'on travaille dans l'espace) dont les sommets appartiennent à A. Pour que cette construction soit une triangulation il faut en outre que les triangles ou les tétraèdres qui la composent, ou bien ne s'intersectent pas, ou bien s'intersectent entièrement sur l'une de leurs faces[4] (voir la figure 6.8 A). Cette définition peut être étendue au cas d'un ensemble d'atomes assimilés à des points pondérés (ou des sphères) $A = \{(a_i, r_i)\}_i$, en ne considérant que les centres atomiques a_i. Pour un même ensemble de points ou d'atomes, il existe de nombreuses triangulations différentes ; la figure 6.8 B montre deux triangulations distinctes de l'enveloppe convexe[5] d'un même ensemble de centres atomiques.

Dans la suite du document, on considérera uniquement des triangulations en trois dimensions, même si pour des besoins de clarté, la plupart des schémas représenteront des triangulations dans le plan.

Notion de simplexe

Les divers "éléments" constituant une triangulation sont appelés des *simplexes*, et on parle plus spécifiquement de *$k-$simplexe* — avec $k = 0, 1, 2$ ou 3 — pour désigner respectivement les sommets, arêtes, facettes (triangulaires) et tétraèdres de la triangulation ; k est la *dimension* du simplexe. On parle de *faces* d'un simplexe σ pour désigner les l-simplexes composant le bord de

[3] C'est-à-dire une union disjointe recouvrant l'espace considéré.

[4] On prend ici le terme de "face" au sens combinatoire : comme il sera exposé un peu plus loin, les sommets, les arêtes et les facettes triangulaires composant un "tétraèdre" sont des faces de ce tétraèdre.

[5] Pour rappel, un ensemble est dit convexe si pour toute paire de points dans cet ensemble, le segment reliant ces deux points est aussi contenu dans cet ensemble. L'enveloppe convexe d'un ensemble E est le plus petit ensemble convexe contenant E.

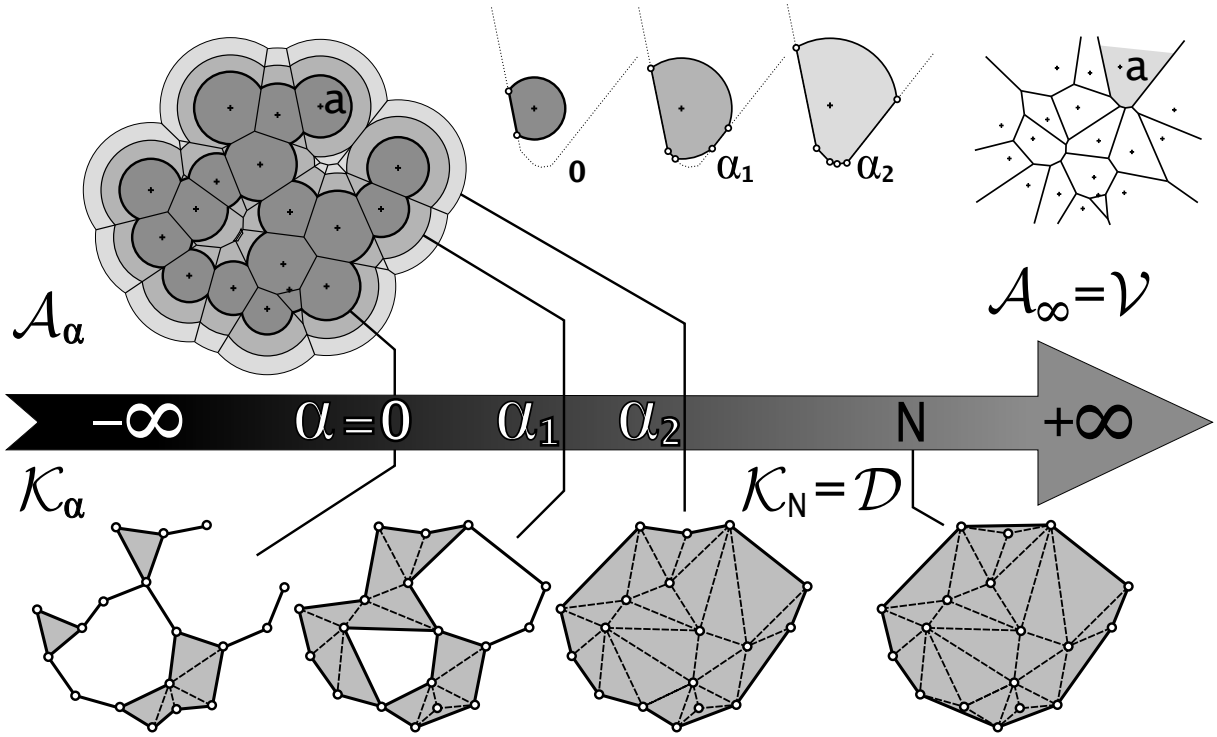


Figure 6.6: Croissance de la forme-α, un exemple en deux dimensions. Dans la partie supérieure, trois ensembles de boules A_α sont représentés en dégradé de gris et superposés dans leur diagramme à remplissage de forme ; le gris foncé correspondant aux valeurs faibles de α. L'évolution de la boule tronquée associée à l'atome a a été isolée en haut à doite de la superposition des diagrammes à remplissage de forme. Dans l'extrémité en haut à droite, une représentation du diagramme de Voronoï partagé par tous les ensembles A_α. Les complexes duaux K_α associés aux A_α sont représentés en bas, ordonnés suivant les valeurs croissantes de α.

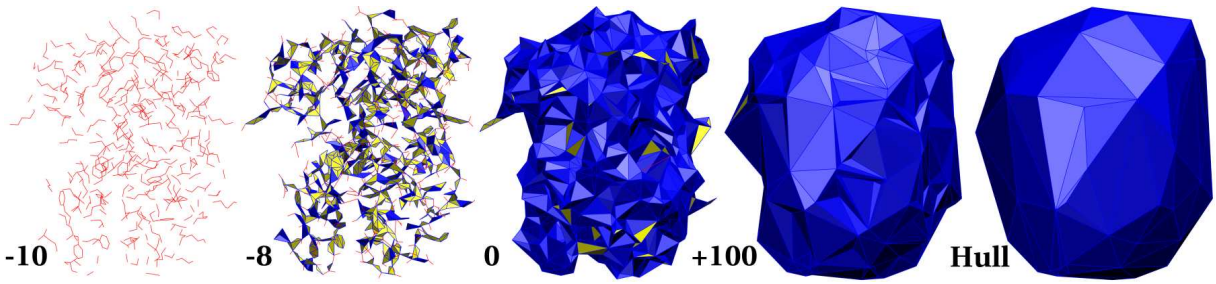


Figure 6.7: Un exemple en trois dimensions de croissance de la forme-α du modèle SA du domaine de liaison au ligand d'un récepteur nucléaire RXR-α (1RDT). De gauche à droite les valeurs de α valent -10, -8, 0, 100 et 100 000.

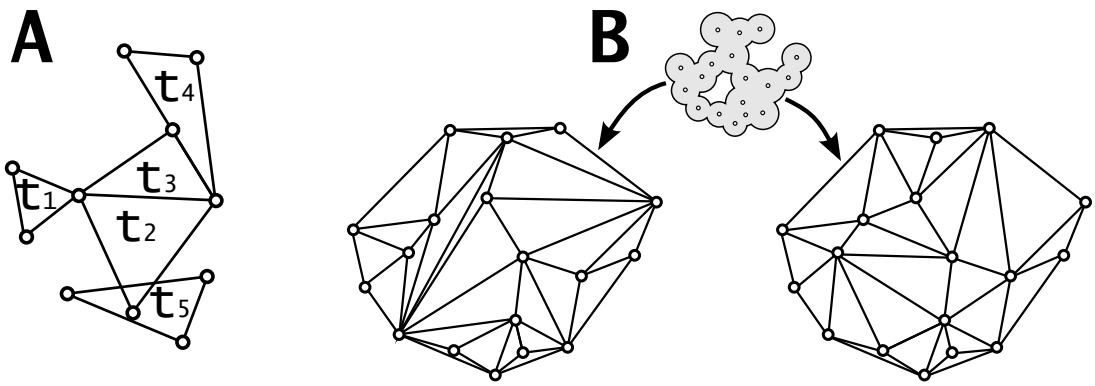


Figure 6.8: Triangulations, un exemple dans le plan. (A) Intersections autorisées et invalides dans une triangulation. Les triangles t_1 et t_2 partagent un sommet, les triangles t_2 et t_3 partagent une arête. En trois dimensions, on peut aussi avoir deux tétraèdres partageant une de leurs facettes, une de leurs arêtes ou un de leurs sommets. L'intersection imparfaite des triangles t_3 et t_4 sur une de leurs arêtes est invalide, de même que le cas de figure des triangles t_2 et t_5 partageant une partie de leurs intérieurs. (B) Deux triangulations différentes d'un même nuage de centre atomiques.

σ (nécessairement $l < k$), et de ses *cofaces* pour désigner les m-simplexes τ sur les bords desquels σ est situé (nécessairement $k < m$)[6].

Une triangulation est ainsi un objet à la fois combinatoire (c'est-à-dire composé d'un nombre fini d'éléments — ici, des simplexes — reliés entre eux par des relations d'incidence et d'adjacence) et géométrique (en ce sens qu'à chaque sommet on associe une position dans l'espace). On parle généralement de *topologie* pour désigner la partie combinatoire, et de *plongement* pour désigner le placement des sommets dans l'espace. Dans ce contexte, on fait aussi la distinction entre le *sommet* (ou 0-simplexe), qui est un objet combinatoire, et le *point* (assimilé à ses coordonnées), qui est un objet géométrique. De la même manière on peut être amené à faire la distinction entre un 3-simplexe, qui est un objet combinatoire, et le tétraèdre, qui est l'objet géométrique associé. Dans la suite, on utilisera souvent le terme de *cellule* [7] d'une triangulation pour désigner un de ses 3-simplexes.

Point à l'infini et triangulation de l'espace

Pour certaines applications, il peut-être intéressant de considérer une triangulation recouvrant l'espace entier. Pour ce faire, une technique consiste à insérer un point virtuel à l'infini. La figure 6.9 A illustre cette insertion d'un point de vue combinatoire. La triangulation de l'enveloppe convexe $\mathcal{C}$ d'un nuage de points est complétée en ajoutant un nouveau sommet v_∞ combinatoirement relié aux simplexes de l'enveloppe convexe. Dans la figure 6.9 A, ces liens consistent en de nouvelles arêtes (symbolisées par des lignes pointillées) ainsi que par les 2-simplexes (ou facettes) qu'elles délimitent ; concrètement, une nouvelle facette est crée par arête sise sur le bord de l'enveloppe convexe et des liens d'adjacence convenables sont maintenues entre elles. En particulier, ces nouvelles facettes sont exactement l'ensemble des facettes incidentes au sommet infini. En trois dimensions, la même construction s'applique, et une nouvelle cellule est ainsi créée pour chaque facette du bord de l'enveloppe convexe.

[6] *i.e* σ est une coface de τ si et seulement si τ est une face de σ.

[7] Le terme de cellule est ici emprunté à la terminologie en vigueur dans la librairie `CGAL` [Da 06] et ne doit pas être confondu avec une autre acception usuelle dans laquelle une cellule — élément d'un complexe cellulaire — peut-être comprise comme une généralisation de la notion de simplexe pour une dimension quelconque.

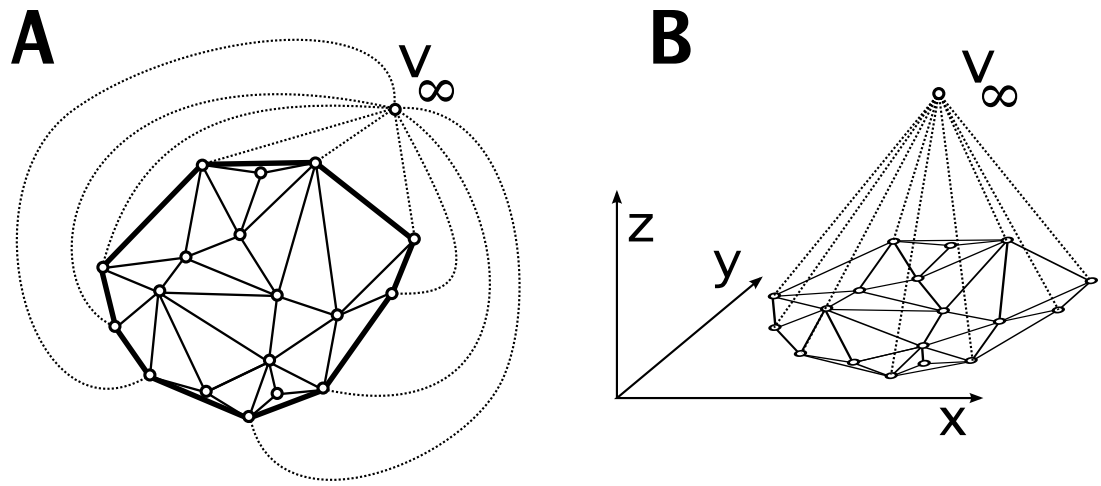


Figure 6.9: Adjonction d'un sommet infini v_∞ dans une triangulation, un exemple en deux dimensions.

Comme le suggère la figure 6.9 B, l'ajout du sommet infini peut aussi être comprise — en passant à la dimension supérieure — comme une triangulation de la sphère.

Une telle triangulation permet entre autres de paver l'espace entier au lieu de rester confiné à l'enveloppe convexe du nuage de points ; elle facilite aussi l'encodage et le parcours du bord de l'enveloppe convexe par une simple interrogation d'adjacence au sommet à l'infini.

6.1.5 Triangulation de Delaunay

La triangulation de Delaunay d'un ensemble d'atomes $A = \{(a_i, r_i)\}_i$ est une triangulation particulière caractérisée par un ensemble de propriétés relatives à la forme "régulière" des tétraèdres (ou, en deux dimensions, des triangles) qui la composent. Cette triangulation a été évoquée plus tôt dans ce même chapitre (page 66), comme le complexe-α $\mathcal{K}_N$ d'un ensemble de boules pour une valeur $\alpha = N$ suffisamment élevée ; une autre définition — plus classique — a été abordée dans l'introduction (page 12) par dualité du diagramme de Voronoï. Notons au passage que cette dualité implique que pour tout ensemble d'atome il existe une triangulation de Delaunay et que celle-ci est unique[8]. Le lecteur trouvera une description plus détaillée en annexe B page 177 où une présentation très générale de cette triangulation et de ses principales propriétés est proposée.

Parmi ces propriétés, nous avons déjà évoqué la possibilité d'utiliser les arêtes de la triangulation pour définir une notion intuitive de *voisinage directionnel* autour de chaque atome (cette propriété est explicitée plus en détail dans la section B.4 page 183). A la section 6.1.2 traitant du complexe-α (page 64), nous avons également mentionné une caractéristique potentiellement limitante de cette triangulation, la "disparition" occasionnelle d'atomes trop enfouis par leurs voisins ; le cas de ces atomes dits *redondants* est explicité en annexe B.5 page 184. Une troisième propriété — dite *propriété de la sphère vide* (définie plus avant dans la section B.3 page 181) — caractérise complètement la triangulation de Delaunay.

Concrètement, c'est cette dernière propriété qui est généralement mise à profit pour construire la triangulation de Delaunay d'un ensemble de points ou d'atomes [Edelsbrunner 96]. De fait, s'il est plus pratique de définir la triangulation de Delaunay à partir du diagramme de Voronoï, il est plus aisé de la construire à partir de la propriété de la sphère vide ; le diagramme de Voronoï est alors souvent lui-même construit par dualité à partir d'une triangulation de Delaunay ainsi

[8] Encore une fois, on suppose que les atomes ont en *position générale*.

construite. De même, il est plus pratique de construire le complexe-α comme une filtration des simplexes de la triangulation de Delaunay, c'est-à-dire comme une triangulation de Delaunay à laquelle on adjoint un moyen d'interroger les valeurs de α pour lesquelles un simplexe donné sera intérieur, extérieur, singulier ou régulier (les détails de cette classification ont été présentés page 61). Nous verrons page 199 que cette construction du complexe-α à partir de la triangulation de Delaunay a influé le schéma de conception des classes afférentes à ces problématiques dans la bibliothèque `CGAL`.

Pour la distinguer du cas non pondéré (lorsqu'on considère uniquement des centres atomiques et qu'on oublie leurs rayons), on l'appelle souvent *triangulation régulière*.

6.1.6 Structure de données de demi-arêtes

La structure de données de demi-arêtes [Kettner 99] n'est pas, à proprement parler, un objet géométrique ; il s'agit en fait d'une structure combinatoire permettant la description d'objets surfaciques en dimension 3. Cette structure de donnée est plus particulièrement (et exclusivement) adaptée à la représentation d'une classe de surfaces idéales[9], les variétés orientables. Les notions de variété et d'orientabilité sont respectivement explicitées dans les figures 6.10 et 6.11.

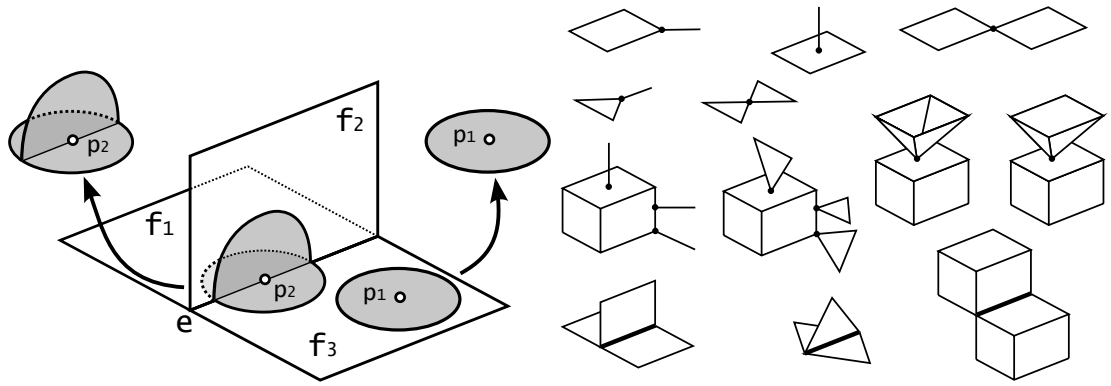


Figure 6.10: Définition d'une surface variété, et exemples de surfaces non variété. Une surface est dite variété lorsque chaque point qui la compose admet un voisinage (topologiquement) assimilable à un disque ouvert. Dans la figure de gauche, trois feuillets f_1, f_2 et f_3 sont assemblés le long d'une même arête e. Quel que soit le point pris sur un des feuillets et en dehors de cette arête commune, il est possible de trouver autour de ce point un petit voisinage assimilable à un disque ; un exemple est donné en gris autour du point p_1. À l'inverse, tout point pris sur e, tel p_2 par exemple, est sujet à une ambiguïté de voisinage. La figure de droite montre d'autres surfaces non variété. Les arêtes et les points sujets à problème sont épaissis.

Il convient de bien différencier la *demi-arête*, élément primitif servant à la construction des arêtes ainsi que du maillage, et la *structure de données de demi-arêtes*, qui contient un ensemble de demi-arêtes et éventuellement de sommets et de facettes, et permet de représenter et de manipuler un maillage de surface.

Comme représenté dans la figure 6.12 *A*, une demi-arête correspond intuitivement à une arête orientée, et une paire de demi-arêtes opposées l'une à l'autre constitue une arête dans le maillage. Chaque demi-arête contient un lien vers son opposée ainsi que vers celle qui la suit dans la facette. La nomenclature *opposite()* et *next()* utilisée dans la figure correspond à l'implémentation fournie dans la librairie `CGAL`. Dans une structure de données de demi-arêtes, une facette est entièrement

[9] au sens qu'elle constitue l'idée intuitive que tout un chacun se fait à l'évocation du mot "surface"

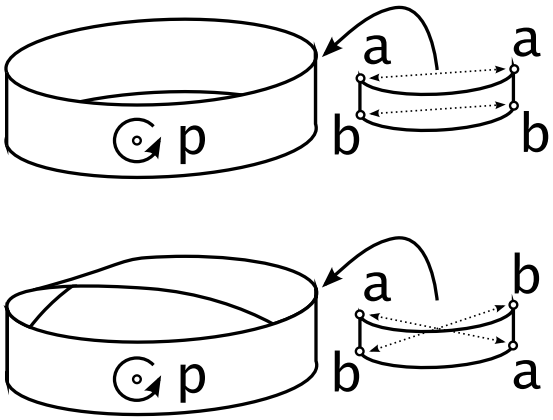


Figure 6.11: Exemple de variété orientable et de variété non-orientable. Pour chaque point p de la surface, on définit une orientation en choisissant un sens de parcours sur un chemin autour de ce point. Une surface variété est dite orientable s'il est possible d'orienter de manière consistante le voisinage de chaque point p de sa surface, c'est-à-dire de telle manière que deux points proches sur la surface aient leur voisinage orienté de la même manière. Le ruban obtenu en cousant bord à bord une bande de papier est orientable. À l'inverse, la bande de Möbius obtenue de la même manière mais en ayant préalablement fait subir une torsion à la bande de papier n'est pas orientable.

encodée par une succession de demi-arêtes ; de la même façon, un sommet correspond à un ensemble de demi-arêtes liées par une succession d'opérations simples (voir figure 6.12 B).

Pour ces raisons, l'utilisation de sommets et de facettes explicites n'est pas absolument nécessaire, mais pour les besoins d'une implémentation spécifique, on peut vouloir stocker de l'information dans l'un ou l'autre de ces objets (figure 6.12 C). Dans ce cas, chaque demi-arête est associée au sommet vers lequel elle pointe, ainsi qu'à sa facette incidente. Les facettes et sommets de la structure de données de demi-arêtes n'ont besoin que de retenir un lien vers une des demi-arêtes qui leurs sont incidentes, les autres pouvant être retrouvées en "orbitant" comme exprimé ci-avant. Enfin, pour accélérer certains traitements, il est possible d'ajouter aux demi-arêtes un pointeur vers la demi-arête précédente dans la face.

Pour finir, chaque facette d'une structure de données de demi-arêtes est naturellement orientée par la succession de ses facettes, et cette même orientation est partagée par l'ensemble des facettes de la structure (figure 6.12 D).

6.2 Généralités sur les implémentations d'algorithmes en géométrie algorithmique

Bien que certaines des constructions qu'elle étudie aient plus d'un siècle, la géométrie algorithmique en tant que discipline à part entière est un domaine relativement récent. Rattachée à l'informatique, son champ d'application couvre la définition d'objets géométriques et l'étude d'algorithmes pour leur construction et leur utilisation.

La problématique de fournir des implémentations "concrètes" de ces algorithmes "théoriques" est plus récente encore. L'implémentation des algorithmes théoriques proposés dans la littérature soulève des problèmes relatifs aux limitations du support informatique, tels que l'encombrement en mémoire, la stabilité numérique ou la robustesse quant à la gestion des cas particuliers. Dans le cadre spécifique de l'implémentation d'algorithmes géométriques, ces deux derniers problèmes sont particulièrement critiques. En effet, si pour la résolution de problèmes numériques des approximations sont souvent envisageables, les constructions géométriques, elles, reposent générale-

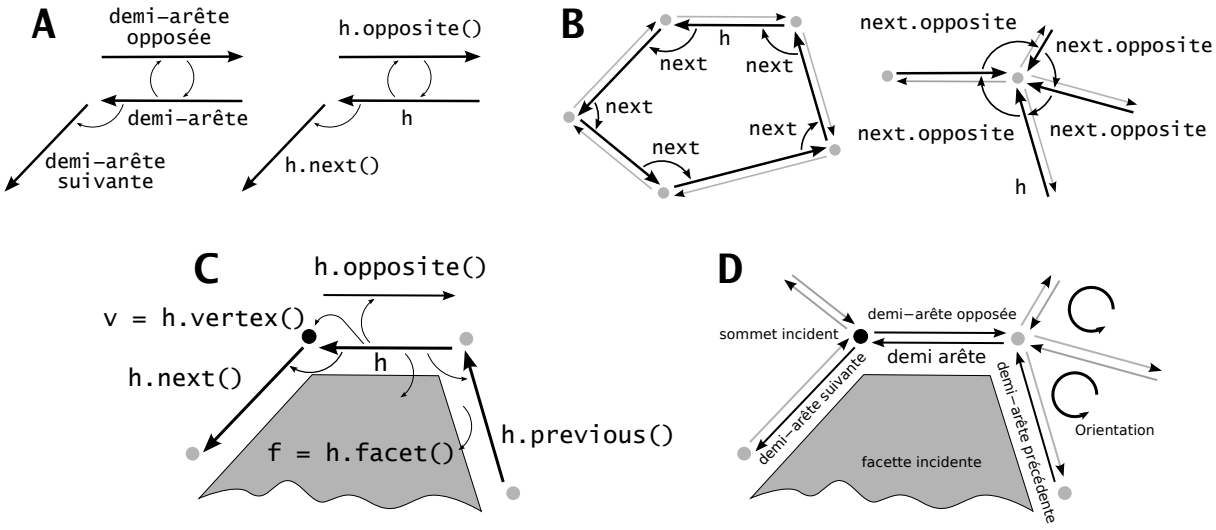


Figure 6.12: Structure de donnée de demi-arêtes.
A : Une demi-arête est reliée à deux autres demi-arêtes, son opposée *opposite()*, et sa suivante *next()*.
B : Une succession d'opérations *next()* délimite une facette en parcourant toutes les demi-arêtes qui la bordent. Les demi-arêtes pointant vers un même sommet se correspondent par une suite d'opérations *next()* et *opposite()* combinées.
C : Une structure de données de demi-arêtes minimale contient uniquement des demi-arêtes, il est néanmoins possible de l'étendre pour ajouter des sommets et des facettes explicites.
D : Par construction, une structure de données de demi-arêtes voit toutes ses facettes orientées dans le même sens, l'orientation étant implicitement donnée par le sens des demi-arêtes associées aux faces.

ment sur une succession d'étapes à chacune desquelles il faut pouvoir répondre de manière exacte à une question posée. La construction de la triangulation de Delaunay, par exemple, repose sur la capacité à déterminer la position d'un point par rapport à une sphère (à l'intérieur de la sphère, à l'extérieur ou sur son bord). Une réponse erronée à une seule étape pourra conduire à une construction erronée, ou à un algorithme bouclant indéfiniment.

L'implémentation d'algorithmes géométriques n'est donc pas une tâche aisée, même lorsque ceux-ci sont bien connus et font l'objet d'une description exhaustive dans la littérature. Heureusement, de nombreuses initiatives ont mené à des implémentations libres de droit sous la forme de programmes exécutables ou de bibliothèques de programmation. Les travaux les plus susceptibles de nous intéresser ont été répertoriés dans la liste ci-après.

- *Hull* [Clarkson 93] est un programme en langage `C-ANSI` dont les sources sont accessibles librement. Il permet de calculer le diagramme de Voronoï, la triangulation de Delaunay et la forme-α d'un ensemble de points, mais pas d'un ensemble de sphères.

- *Qhull* [Barber 96] est une suite de programmes écrits en langage `C-ANSI` pour le calcul des diagrammes de Voronoï, de l'enveloppe convexe et de la triangulation de Delaunay. Les sources ainsi que des versions précompilées sont accessibles à l'adresse `http://www.qhull.org`.

- Réalisée par Ernst Peter Mücke en collaboration avec Herbert Edelsbrunner, *alpha* [Mucke 93, Akkiraju 95] est la première implémentation des algorithmes permettant de construire une forme-α. Elle se présente sous la forme d'une suite de programmes permettant individuellement de calculer et de visualiser la triangulation de Delaunay d'un ensemble de sphères ainsi que sa filtration. Des routines sont aussi fournies pour le calcul vo-

lumétrique des diagrammes de remplissage de forme et de leurs vides. Le code source n'est pas disponible, mais les exécutables peuvent être librement téléchargés à l'adresse : `ftp://ftp.ncsa.uiuc.edu/Visualization/Alpha-shape/`

- *Pocket* [Edelsbrunner 03] est un programme dédié à la recherche et à la caractérisation des poches dans les macromolécules. Basé sur la théorie des formes-α, il fournit une implémentation (en `Fortran` et en langage `C`) pour la construction des triangulations de Delaunay et de la forme duale d'une molécule.

- *Vtk* [Schroeder 96] est une librairie `C++` offrant, entre autres, de nombreuses classes pour la manipulation et la visualisation d'objets géométriques. Les triangulations de Delaunay et formes-α de points y sont traitées mais pas le cas pondéré.

- La librairie `CGAL` [CGAL 09] fournit des structures de données et des algorithmes intégrant de nombreuses problématiques de géométrie algorithmique sous la forme de classes `C++`.

La bibliothèque `CGAL` est le projet le plus abouti dans le domaine, tant en terme de conception que de problématiques adressées ; nous avons donc naturellement opté pour son emploi. On trouvera une présentation détaillée de la librairie `CGAL` en annexe F page 195.

Dans le chapitre suivant nous utiliserons les modèles présentés ici pour proposer, à l'aide d'une structure de données de demi-arêtes, une méthode de parcours du bord du complexe dual d'une molécule.

Chapitre 7

Surface duale : parcourir le bord du complexe dual

DANS NOTRE étude, nous avons caractérisé la surface et la forme des macromolécules au travers du polyèdre défini par le bord du complexe dual. Cette surface n'est pas variété[1], ce qui la rend impropre pour un emploi dans des traitements où un parcours "le long de la surface" est nécessaire. Ces aspects seront présentés dans une première section. Pour pallier ces problèmes, nous avons défini une surface polyédrique similaire au bord du complexe dual et qui soit variété. Cette surface, que nous avons appelée *surface duale*, sera définie et caractérisée dans une seconde section et un algorithme pour la construire à partir du complexe dual sera présenté dans la troisième section. Les propriétés de cette surface seront présentées dans la quatrième section.

7.1 Contexte et motivations

La figure 7.1 présente des cas de non variété sur le bord du complexe dual. Ces exemples sont illustrés dans un cas concret dans la figure 7.2 où ils sont mis en évidence sur le complexe dual d'un récepteur nucléaire. Comme indiqué dans la figure 7.1, les occurrences de non variété peuvent mettre en jeu des arêtes ou des facettes *pendantes*, ou totalement *intégrées* au bord du complexe dual. On remarquera que les arêtes et facettes singulières (définies dans la section 6.1.2 page 61) induisent systématiquement des cas de non variété (comme dans les figures 7.1 A, B et F). La régularité des arêtes et facettes du complexe dual n'est cependant pas une condition suffisante pour avoir une surface variété, comme le montrent les cas C, D et E de la même figure. Un même sommet, une même arête ou une même facette du complexe dual peut "appartenir" simultanément à ce que l'on aimerait intuitivement considérer comme des "côtés" différents de la surface. Cette notion intuitive de "surface solide" peut être dégagée en considérant un objet posé sur cette surface, capable de s'y déplacer mais pas de la traverser (voir la figure 7.3). Une surface non variété contient des ambiguïtés dans la notion de voisinage et autorise des "franchissements illégaux" qui vont à l'encontre de cette intuition de "surface". En particulier, un parcours le long des arêtes du bord du complexe dual ne permettra pas de garantir qu'on reste bien du même côté de la surface de ce polyèdre. Cet aspect a été illustré en deux dimensions et en trois dimensions dans la figure 7.4.

[1] Le complexe dual a été défini au chapitre 6.1.2 page 61, et une définition du terme *variété* a été donnée dans la figure 6.10 page 70.

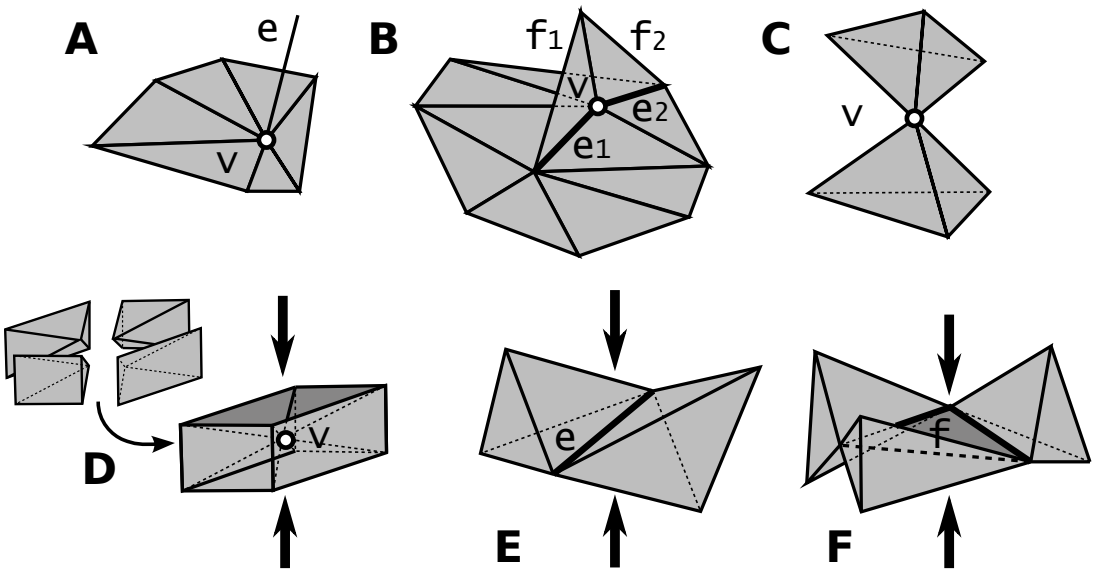


Figure 7.1: Cas de non variété du bord du complexe dual. Comme expliqué dans la figure 6.10 page 70, une surface est dite variété lorsque le voisinage de chacun de ses points est (topologiquement) assimilable à un disque. (A) L'arête e n'est attachée à la surface que par le point v ; elle est dite "pendante". Ne serait cette arête, le voisinage du point v serait assimilable à un disque, mais du fait de la présence de e, cette surface n'est pas variété. (B) Les facettes f_1 et f_2 sont à la fois singulières et pendantes. Elles ne sont rattachées à la surface que par les arêtes e_1 et e_2 dont chaque point n'admet pas de voisinage variété. (C) Le sommet v joint deux tétraèdres. La surface autour du point v est similaire à deux disques partageant uniquement un point central. La notion de "surface" autour de ce point v est floue ; faut-il considérer deux surfaces séparées pour chaque tétraèdre joint par v, ou bien l'information de jonction entre les deux surfaces est-elle importante ? (D) Cette figure est obtenue en assemblant quatre pyramides autour d'un sommet v de manière à ce que chaque pyramide (une pyramide n'est pas un simplexe, mais peut être obtenue par assemblage de deux tétraèdres) partage une face avec sa voisine. Le sommet v est aussi partagé entre deux "parties" de la surface. Dans cet exemple, contrairement au précédent, la notion intuitive de surface autour de v est assez claire, et on aimerait distinguer chacune des deux composantes indiquées par une flèche. De la même manière, dans les cas (E) et (F), c'est respectivement une arête et une facette qui sont partagées entre deux parties "distinctes" de la surface.

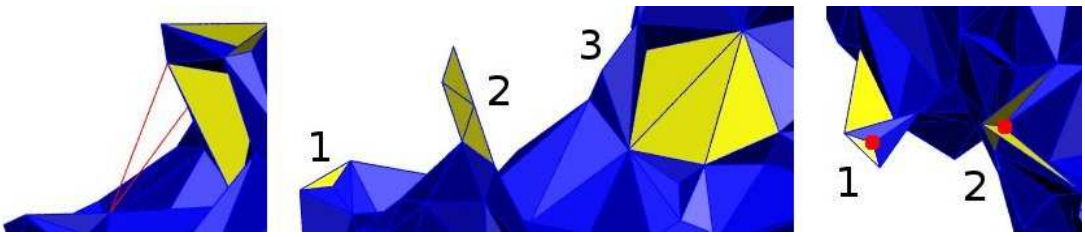


Figure 7.2: Le bord du complexe dual d'une molécule est rarement variété ; ici, trois détails pris sur le complexe dual du domaine de liaison au ligand du récepteur nucléaire à l'acide rétinoïque (1RDT). Les arêtes et facettes singulières sont colorées respectivement en rouge et en jaune, les arêtes et facettes régulières en bleu. Figure de gauche : deux arêtes singulières relient les résidus *Lys405* et *Pro412*. Dans la figure centrale, trois patchs de facettes singulières sont mis en évidence. Le premier au bout du résidu *Glu456* et le second sur la longueur de *Lys381* constituent des patchs pendants. Le troisième est un patch entièrement intégré à la surface, il sépare le "vide" de l'espace du solvant du vide d'une cavité située juste sous la surface. La figure de droite montre deux sommets (matérialisés par des points rouges) partagés entre deux (pour le sommet 1 représentant le carbone δ de la *Gln361*) et trois (pour le sommet 2 représentant le carbone ϵ de la *Lys407*) parties de la surface. Ces sommets correspondent à des atomes possédant, respectivement, deux et trois composantes de surface dans le modèle Surface Accessible.

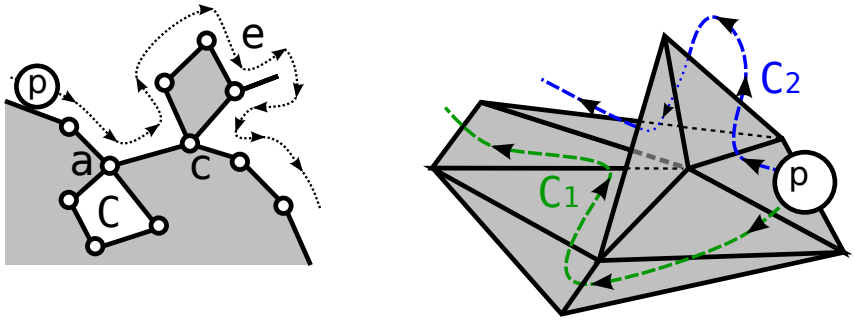


Figure 7.3: Parcours d'un objet sur le bord du complexe dual, un exemple en deux dimensions (à gauche) et en trois dimensions (à droite). Dans chaque exemple, un objet sphérique p a été posé sur la surface, et des chemins le long de cette surface ont été représentés en pointillés. Dans l'exemple en deux dimensions, la sphère p ne peut visiter la petite cavité C sous le point a. De même, elle ne peut traverser le point c, mais peut le visiter de manière différenciée de deux côtés. Il en va de même de l'arête e toute entière qui peut être visitée de deux côtés. Dans l'exemple en trois dimensions, deux chemins ont été représentés. Le chemin vert C_1 évite les facettes singulières. Le chemin bleu C_2 visite les deux côtés d'une des facettes singulières.

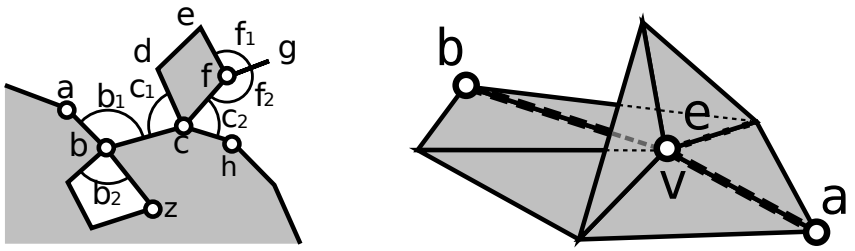


Figure 7.4: Ambiguité de parcours le long des arêtes sur le bord du complexe dual, un exemple en deux dimensions (à gauche) et en trois dimensions (à droite). Dans l'exemple en deux dimensions, le volume d'un objet polygonal a été rendu en gris. Son bord comprend des ambiguïtés aux sommets b, c, et f. Un déplacement du sommet a au sommet h "le long de la surface" ne devrait pas pouvoir passer dans la cavité sous b et visiter par exemple le sommet z. De même, on ne devrait pas pouvoir passer "au travers" du polyèdre autour du sommet c. Pour ce faire, il faudrait être capable de différencier deux composantes de surfaces autour des sommets b, c et f, virtuellement partagés de part et d'autre de la surface. De manière analogue, l'exemple en trois dimensions montre le cas de deux facettes pendantes. Un parcours du sommet a au sommet b en suivant les arêtes en surface du polyèdre ne devrait pas traverser les "murs" matérialisés par les facettes pendantes au sommet v, et devrait par exemple les contourner en empruntant deux fois l'arête e.

7.2 Définition de la surface duale

Pour pallier ce problème, il est nécessaire de différencier chaque "côté" de la surface, et donc de considérer chaque objet (sommet, arête, facette) suivant le côté de la surface d'où il est observé. Pour ce faire, chaque simplexe du bord du complexe dual appartenant à plusieurs côtés de la surface est dupliqué en autant d'exemplaires que de côtés de la surface auxquels il participe, et des relations d'incidence et d'adjacence cohérentes sont maintenues entre ces objets pour encoder de manière consistante le trajet d'un objet le long de la surface (voir l'exemple de la figure 7.5). La duplication des simplexes dont il est question ici est purement combinatoire : les sommets, arêtes et facettes ainsi "éclatés" sont superposés au modèle dont ils sont issus sur le bord du complexe dual et seule la connectivité est impactée. Dans cette construction, les arêtes singulières du complexe dual sont systématiquement supprimées (voir la figure 7.6 A) et les connexions "ambigües" de deux surfaces autour d'un sommet sont systématiquement déconnectées (figure 7.6 C et D). Cette nouvelle surface a été baptisée *surface duale* en raison de son rapport avec le complexe dual, et parce qu'à quelques détails près, elle est combinatoirement duale de la Surface Accessible, comme il sera précisé dans la section 7.4.3 page 85. Comme on peut l'observer dans l'exemple en deux dimensions de la figure 7.5, la surface duale peut comporter plusieurs composantes déconnectées les unes des autres. Certaines de ces composantes sont en contact avec la partie infinie de l'espace, et seront appelées *composantes extérieures*. Les autres composantes sont totalement enfouies à l'intérieur d'une composante extérieure ; elles constituent des cavités dans l'espace clos de cette composante extérieure, et seront appelées *composantes intérieures*.

7.3 Construction de la surface duale à partir du complexe dual

Nous présentons ici une méthode pour la construction de la surface duale à partir du complexe dual. Dans nos développements, nous avons choisi la structure de données de demi-arêtes[2] pour stocker la surface duale ainsi construite, aussi les algorithmes présentés dans cette section sont-ils adaptés à cette structure de donnée particulière. Le principe est néanmoins applicable à d'autres structures. Par convention, dans les descriptions d'algorithmes ci-après, nous utiliserons des lettres capitales pour désigner les objets sous-jacents à la surface duale (ou à la structure de données de demi-arêtes qui lui est associée), et des lettres minuscules pour désigner les objets relatifs au complexe dual.
Le principe général de notre construction s'appuie essentiellement sur les deux remarques préliminaires consignées ci-après et illustrées dans la figure 7.7 :

Remarque 7.1. *Deux facettes de la surface duale sont adjacentes lorsqu'il est possible de passer de l'une à l'autre au travers d'un espace vide dans le complexe dual.*

Remarque 7.2. *Une facette sur le bord du complexe dual peut-être vue comme une des faces d'une cellule extérieure au complexe dual. Lorsque la facette est singulière, deux cellules extérieures distinctes se partagent cette facette.*

La première remarque motive l'implémentation d'un algorithme incrémental par propagation successive autour d'une facette du complexe dual. La seconde remarque motive l'utilisation des cellules de la triangulation associée au complexe dual pour le stockage des facettes de la surface duale en cours de construction. Cet artifice permet entre autre de différencier des facettes

[2] La structure de données de demi-arêtes et son implémentation dans `CGAL` ont été présentées dans la section 6.1.6 page 70.

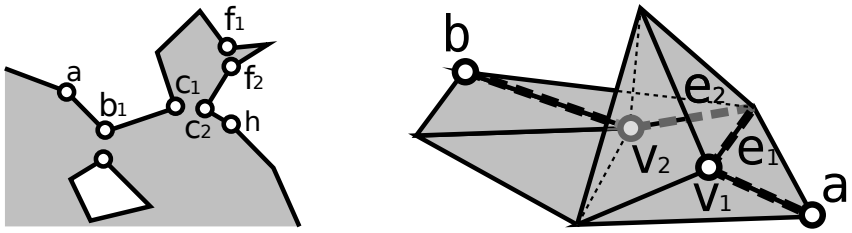


Figure 7.5: Désambiguïsation de parcours par la surface duale. Les exemples en deux dimensions (à gauche) et en trois dimensions (à droite) reprennent ceux de la figure 7.4. Dans l'exemple en deux dimensions, les sommets b, c et f ont été dupliqués, de même que l'arête reliant les sommets f et g. Le franchissement direct de b à h en passant par c n'est maintenant plus possible. Dans l'exemple en trois dimensions, les deux facettes pendantes sont dupliquées ainsi que trois arêtes et le sommet v. Pour aller du sommet a au sommet b, il est maintenant impossible de passer directement de v_1 à v_2, et l'on devra par exemple emprunter les arêtes e_1 et e_2. Dans cette figure, les sommets "éclatés" ont été séparés pour une meilleure visibilité. Dans les faits, ces sommets conservent les coordonnées du sommet initial et sont donc géométriquement superposés, tout comme les arêtes et facettes "éclatées".

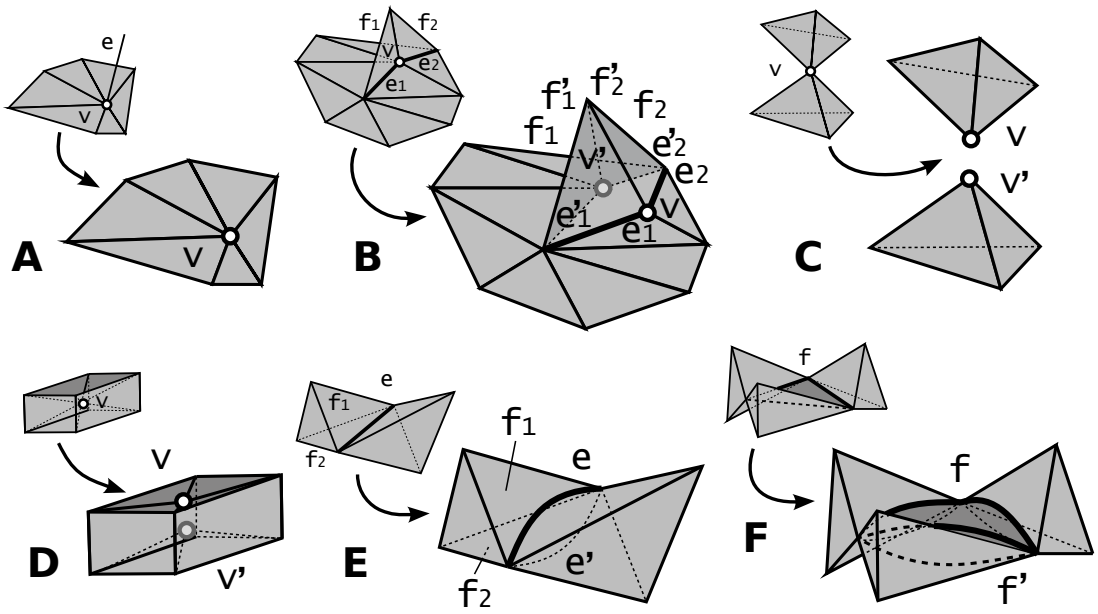


Figure 7.6: Passage du bord du complexe dual à la forme duale pour les exemples de la figure 7.1. (A) Les arêtes singulières sont systématiquement supprimées. (B) Le sommet v ainsi que les arêtes e_1 et e_2, et les facettes f_1 et f_2 sont chacuns dupliqués. (C) Le sommet v connectant de manière "ambigüe" deux parties de surface est dupliqué, et les deux parties déconnectées. (D) Le sommet v connectant deux côtés distincts de surface est dupliqué et les deux parties déconnectées. (E) L'arête e intégrée à la surface et partagée entre deux côtés distincts de cette surface est dupliquée. Les sommets incidents à e n'appartiennent qu'à un côté de cette surface, ils ne sont pas dupliqués. Les facettes f_1 et f_2, adjacentes sur le bord du complexe dual, ne le sont plus sur la surface duale. (F) La facette f partagée entre deux côtés de la surface est dupliquée. Ici, les trois arêtes incidentes sont elles-aussi dupliquées, mais les trois sommets incidents ne le sont pas.

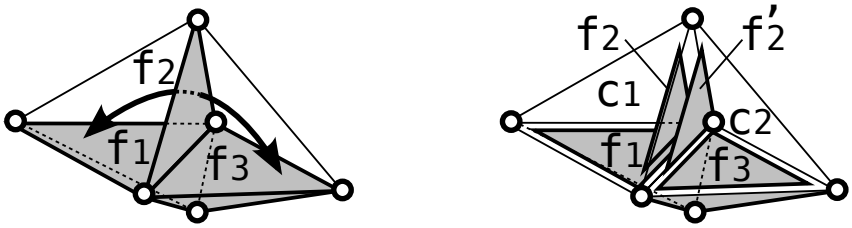


Figure 7.7: Remarques préalables à la construction de la surface duale à partir du complexe dual, un exemple simplifié comprenant six sommets. La figure de gauche montre le complexe dual des six sommets considérés. Les facettes f_1 et f_3 partagent une même arête dans le complexe dual, mais elles ne sont pas adjacentes dans la surface duale car pour passer de l'une à l'autre autour de l'arête qu'elles partagent il faudrait pouvoir passer au travers de la facette f_2 faisant partie du complexe dual. La figure de droite montre la même triangulation de Delaunay. Des quatre cellules de cette triangulation, seules c_1 et c_2 sont extérieures au complexe dual. Les facettes de la surface duale peuvent être considérées comme des facettes du complexe dual observées depuis une cellule extérieure incidente. La facette f_2 du complexe dual correspond aux facettes f_2 et f_2' suivant qu'elle est observée depuis c_1 ou c_2 respectivement.

distinctes de la surface duale issues d'une même facette du complexe dual (cas des facettes singulières).

7.3.1 Boucle principale pour la construction d'une composante de la surface duale

L'algorithme 7.1 construit la composante de la surface duale associée à une facette f sur le bord du complexe dual. Dans un premier temps, les arêtes et facettes de la composante de surface duale sont construites par une procédure récursive *processFacet* (ligne 2) dont l'algorithme est explicité au paragraphe suivant. Il n'est pas possible de construire les sommets de composante de surface duale avant d'en avoir construit toutes les arêtes et facettes. Pour cette raison, la procédure *processFacet* est chargée d'apparier chaque demi-arête qu'elle construit avec le sommet du complexe dual vers lequel la demi-arête pointe, et d'empiler ces paires sur une pile initialement vide (définie ligne 1). Les sommets peuvent alors être construits et insérés dans la structure de données de demi-arêtes de la composante de surface duale ; c'est la seconde phase de l'algorithme (lignes 3 à 12). La condition de la ligne 4 se résume à vérifier que le lien `H.vertex()` est `NULL`, auquel cas un nouveau sommet doit être créé et inséré dans la structure de données de demi-arêtes et attaché à la demi-arête H (respectivement lignes 5 et 7). Ce même sommet doit être attaché à toutes les demi-arêtes pointant vers le même sommet de la surface duale que H (Cette opération est réalisée dans la boucle des lignes 8 à 10, en itérant des opérations `H'=H'.next().opposite()` jusqu'à retomber sur H.). La ligne 6 correspond au plongement de la structure de données de demi-arêtes. Suivant l'application, on pourra directement assigner les coordonnées du sommet v au sommet V ou garder un lien de V vers v.

7.3.2 Processus itératif pour la construction des demi-arêtes et facettes d'une composante de la forme duale

L'algorithme 7.2 décrit la construction itérative des demi-arêtes et facettes d'une composante de la surface duale. Cet algorithme comprend deux phases. En premier lieu (lignes 2-11), les trois arêtes incidentes à la facette f passée en paramètre sont explorées à la recherche des facettes

Algorithme 7.1 : Algorithme pour la construction d'une composante de la surface duale à partir du complexe dual et d'une facette de cette composante.

Entrées :
k : un complexe dual
f : une facette sur le bord de k
D : une structure pour stocker la surface duale
Sorties :
La composante de la surface duale contenant f est ajoutée à D
▷ *Créer une liste pour associer chaque demi-arête à son sommet dans le complexe dual*
1 définir P, une pile vide
▷ *Construire les arêtes et les facettes de la composante de surface duale*
2 processFacet(k,f,P,D)
▷ *Insérer les sommets dans la surface duale*
3 **pour** *chaque paire (H,v) dans P* **faire**
4 | **si** *la demi-arête H n'a pas encore de sommet attaché* **alors**
5 | | créer un sommet V dans D
6 | | associer v à V
7 | | associer V à H
8 | | **pour** *chaque demi-arête H' de D pointant sur v* **faire**
9 | | | associer V à H'
10 | | **fin**
11 | **fin**
12 **fin**

adjacentes à f, les demi-arêtes correspondantes sont construites et associées aux arêtes de f et la procédure est récursivement appelée sur les facettes adjacentes. Dans un second temps (lignes 12-14), la facette F correspondant à f est effectivement construite dans la structure de données de demi-arêtes. Conformément à la remarque 7.2, la facette f sur le bord du complexe dual, ainsi que ses facettes adjacentes, sont désignées par une cellule extérieure au complexe dual et une facette dans cette cellule.
À la ligne 1, la facette f est marquée comme visitée pour garantir qu'elle ne sera traitée qu'une seule fois. Après quoi, dans la boucle des lignes 2 à 11, on recherche les facettes adjacentes à f du même côté de la surface. Concrètement, les trois arêtes e de f sont successivement observées (ligne 3). Si une demi-arête est déjà associée à e, cela signifie que la facette f_e adjacente à f et partageant e a déjà été visitée (les demi-arêtes sont toujours créées par paires opposées, voir ligne 5) ; le "traitement" de cette arête n'est donc pas nécessaire et on passe à la suivante. Dans le cas contraire, la facette adjacente est encore inconnue ; elle doit être recherchée et construite dans la structure de données de demi-arêtes. Ligne 4, la facette adjacente est recherchée dans le complexe dual. Cette recherche a été implémentée en circulant parmi les cellules adjacentes à c, la cellule extérieure associée à f, comme on peut dans la figure 7.8. Lorsque f_e a été trouvée, deux demi-arêtes opposées sont créées, insérées dans la structure de données de demi-arêtes, et associées à l'arrête e respectivement vue dans les faces f et f_e (ligne 5). Les deux demi-arêtes sont en outre appariées au sommet vers lequel elles pointent dans le complexe dual et empilées sur P (ligne 6) pour insertion ultérieure des sommets dans la structure de données de demi-arêtes (voir l'algorithme 7.1). Si la facette f_e n'a pas encore été visitée (ligne 7), la procédure doit lui être récursivement appliquée (ligne 8). Si par contre la facette f_e a été marquée comme visitée (ce qui correspond à la ligne 1 de cette procédure dans un contexte antérieur), cela signifie que le traitement de la facette a déjà débuté et a conduit à la visite de la facette f. Dans ce cas là, le traitement de f_e a débuté et sera achevé une fois tous les appels récursifs "dépilés". Une fois

toutes les arêtes de f traitées, on est assuré que les trois demi-arêtes correspondantes ainsi que leurs demi-arêtes opposées ont été construites. La facette correspondant à f dans la surface duale est matérialisée par la succession de ces trois demi-arêtes. L'opération est concrètement effectuée à la ligne 12 et consiste à lier chaque demi-arête H_i attachée à une arête e_i de f à sa suivante (*via* l'opération `H.next()`). Si la structure de données de demi-arêtes possède des facettes explicites, une facette F correspondant à f peut être construite et insérée dans la structure de données de demi-arêtes (ligne 13). Dans nos développements nous avons effectivement créé cette facette explicite, et nous l'avons utilisée pour stocker un lien vers la facette f lui correspondant sur le bord du complexe dual ; cette information s'avère en effet nécessaire pour nos développements ultérieurs concernant la caractérisation de la forme de la surface duale (chapitre 8). Pour finaliser la structure de données de demi-arêtes, F doit être liée à une des demi-arêtes incidentes, et chaque demi-arête incidente doit être liée à F (ligne 14).

Algorithme 7.2 : Algorithme **processFacet**, pour la construction des demi-arêtes et des facettes d'une composante de la surface duale à partir du complexe dual et d'une facette de cette composante.

Entrées :
k : un complexe dual
f : une facette sur le bord de k
P : une pile
D : une structure de données de demi-arêtes
Sorties :
D contient la composante de la surface duale associée à f
P contient des paires associant chaque demi-arête insérée dans D à son sommet respectif dans le complexe dual

```
1  Marquer la facette f comme visitée
   ▷ Exploration des facettes adjacentes à f dans le complexe dual
2  pour toutes les arêtes e de f faire
3  |  si e n'est pas encore attachée à une demi-arête alors
4  |  |  Chercher f_e, la prochaine facette de la surface duale autour de e
5  |  |  Créer H et H' dans D, les demi-arêtes opposées liant f à f_e dans la surface duale
6  |  |  Apparier H et H' à leurs sommets respectifs dans k, et empiler les paires sur P
7  |  |  si f_e n'est pas encore marquée comme visitée alors
   |  |  |  ▷ appel récursif de cette procédure sur f_e
8  |  |  |  processFacet(k, f_e, P, D)
9  |  |  fin
10 |  fin
11 fin
   ▷ Création de la face F dans la surface duale
12 Lier entre elles les trois demi-arêtes associées aux trois arêtes de f
13 Créer la facette F équivalente à f dans D, la lier à f et à une des demi-arêtes
14 Lier les trois demi-arêtes à leur facette incidente F
```

Afin d'en simplifier la lecture, l'algorithme présenté ici ne fait pas mention de l'orientation des demi-arêtes. La construction d'une structure de données de demi-arêtes valide nécessite cependant que les facettes adjacentes partagent une même orientation et que les demi-arêtes opposées pointent chacune sur un sommet différent de la même arête. Ces considérations impactent la ligne 12 pour la jonction cohérente des trois arêtes d'une même face, et la ligne 6 pour l'attribution du bon sommet à chacune des deux demi-arêtes opposées.

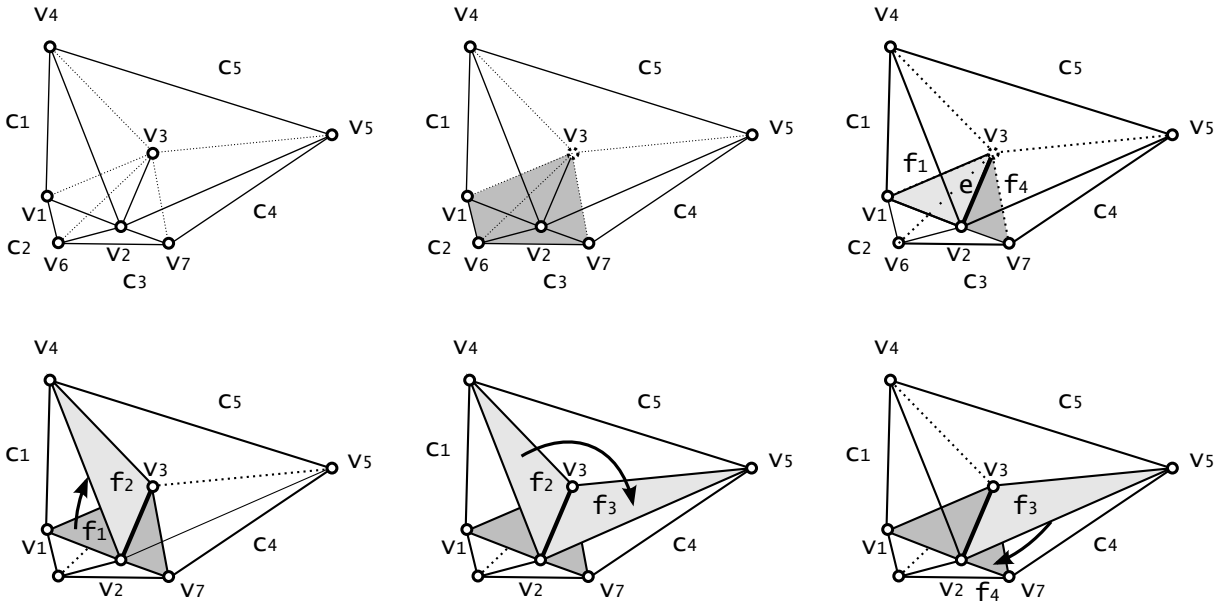


Figure 7.8: Construction de la surface duale : recherche d'une facette adjacente autour d'une arête dans le complexe dual, un exemple dans une triangulation en trois dimensions comprenant sept sommets v_i. En haut, de gauche à droite : la triangulation de Delaunay est représentée seule ; elle comprend cinq cellules c_i et seize arêtes. Le complexe dual est représenté en grisé, les cellules c_2 et c_3 sont intérieures au complexe dual, les trois autres (c_1, c_4 et c_5) sont extérieures. Les deux facettes f_1 et f_4 sont adjacentes dans le complexe dual, et partagent une arête e (en gras). La facette f_1 (en gris clair) est régulière, elle correspond à une unique facette de la surface duale et est associée à la cellule c_1. De même pour f_4 (en gris foncé) qui est associée à c_4. Les trois illustrations de la ligne du bas explicitent la recherche de la facette adjacente à (c_1, f_1) autour de l'arête e : les cellules adjacentes à c_1 opposées à v_1 (le sommet de f_1 n'appartenant pas à e) sont successivement explorées jusqu'à trouver une facette du complexe dual. Ici, (c_1, f_2), (c_5, f_3) puis (c_4, f_4) sont visitées. À l'issue de ce parcours, la facette adjacente de f_1 "vue depuis c_1" et partageant l'arête e est la facette f_4 "vue depuis c_4".

7.4 Propriétés de la surface duale

Pour observer ce qui se produit autour des sommets de la surface duale, on définit les notions d'*ombrelle* et de *multiplicité*.

7.4.1 Ombrelle et multiplicité d'un sommet de la surface duale

On désignera par *ombrelle* autour d'un sommet v de la surface duale, le cycle de facettes incidentes à v, et partageant deux à deux une arête (voir la figure 7.9 A). De par la construction

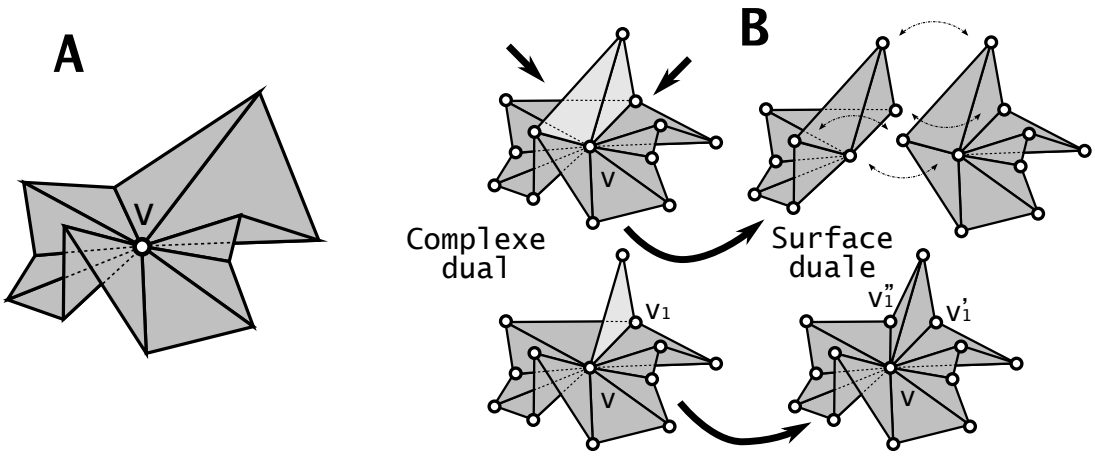


Figure 7.9: Définition d'une ombrelle autour d'un sommet, et séparation des ombrelles dans le complexe dual.
A : Une ombrelle autour d'un sommet v est un assemblage de facettes incidentes à v, topologiquement assimilable à un disque.
B : Deux exemples d'extraction d'ombrelles autour d'un sommet v du complexe dual (colonne de gauche) et leur retransposition dans la surface duale (colonne de droite). En haut, un sommet v de multiplicité 2 sur le bord du complexe dual peut être visualisé de deux "côtés" de la surface. Chaque côté définit une ombrelle autour de v. Dans la surface duale, les deux ombrelles sont explicitement séparées et les deux facettes singulières communes à ces deux ombrelles (en gris clair) sont dupliquées, de même que le sommet v. En bas, le sommet v est de multiplicité 1. La facette singulière (en gris clair) du complexe dual est éclatée, et les deux facettes résultantes intégrées à une unique ombrelle autour de v.

de la surface duale, une ombrelle est topologiquement assimilable à un disque. Comme explicité dans la figure 7.9 B, la définition de la surface duale isole des ombrelles autour de chaque sommet du bord du complexe dual. Une notion d'ombrelle plus formelle peut aussi être proposée pour les sommets du complexe dual : pour un sommet v du complexe dual, on considère $lk(v)$, le *link* de v, c'est-à-dire l'ensemble des simplexes τ du complexe dual qui ne sont pas eux-mêmes incidents à v, mais qui constituent des faces de simplexes incidents à v. Cette construction décrit un polygone dont chaque composante connexe du bord consiste en un cycle d'arêtes. En connectant ces arêtes à v, on définit des ombrelles autour de v dans le complexe dual, et chacune d'elles correspond à une ombrelle de la surface duale. On définit la *multiplicité* d'un sommet v du complexe dual comme le nombre de bords de $lk(v)$; c'est aussi le nombre d'ombrelles définies sur la surface duale autour de ce sommet ; ou encore le nombre de sommets de la surface duale correspondant à ce sommet du complexe dual. Par extension, on parlera de *multiplicité d'un atome* ou d'un sommet de la surface duale pour la multiplicité du sommet qui lui est associé dans le complexe dual.

7.4.2 La surface duale est une variété

La surface duale a été définie de manière à constituer une "variétisation" du bord du complexe dual. Pour s'en convaincre on observera que pour chaque point intérieur à une facette de la surface duale, on peut trouver un voisinage assimilable à un disque. De même, chaque arête de la surface duale reliant exactement deux facettes de cette surface, tout point intérieur à une arête admet un voisinage assimilable à un disque. La seule réelle difficulté concerne les sommets du complexe dual, et ceux-ci sont systématiquement entourés d'une ombrelle, assimilable à un disque, comme il a été remarqué précédemment.

7.4.3 "Quasi-dualité" de la surface accessible et de la surface duale

À la section 6.1.2 (page 61), il a été remarqué que les sommets du bord du complexe dual correspondent chacun à un atome différent sur la Surface Accessible. Comme on peut le voir dans la figure 7.10, certains atomes constituant la Surface Accessible peuvent néanmoins présenter plusieurs morceaux surfaciques déconnectés, et il n'est pas possible de distinguer ces différents morceaux avec le complexe dual seul. La surface duale, elle, permet cette différentiation : les morceaux surfaciques de la Surface Accessible correspondent exactement aux sommets de la surface duale, ainsi qu'à ses ombrelles. De fait, la multiplicité d'un atome constitue aussi le nombre de morceaux surfaciques déconnectés qui le constituent dans la Surface Accessible. La figure 7.11 expose ces différences d'encodage de la Surface Accessible par le complexe dual et par la surface duale. Plus formellement, la Surface Accessible peut être considérée de ma-

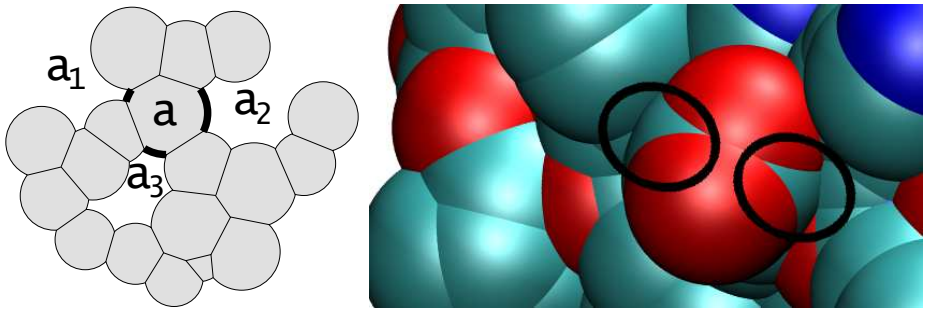


Figure 7.10: Composante de surface d'un atome. Dans le modèle Surface Accessible, certains atomes peuvent présenter plusieurs morceaux de surface déconnectés les uns des autres. Dans la représentation en deux dimensions (à gauche), c'est par exemple le cas de l'atome a qui comprend trois morceaux déconnectés à la surface de la molécule, dont deux (a_1 et a_2) sont situées sur la surface accessible au solvant, et la troisième à l'intérieur d'une cavité. L'exemple de droite montre un détail à la surface du domaine de liaison au ligand du récepteur nucléaire à l'acide rétinoïque X RXRα (1RDT). Le carbone δ de la *Glu434* présente deux morceaux de surface (mis en évidence par des ovales noirs) déconnectés l'un de l'autre.

nière combinatoire (voir la figure 7.12) : ses facettes correspondent aux morceaux surfaciques d'atomes évoqués ci-avant, ses arêtes aux arcs de cercles bordant ces facettes, et ses sommets à l'intersection de deux de ces arêtes. Une définition similaire, présentée dans la figure 7.12 présente ces mêmes éléments combinatoires respectivement comme un morceau surfacique, un arc de cercle correspondant à l'intersection de la surface de deux atomes de surface, et un point d'intersection de trois tels atomes. Les arêtes et les facettes de la surface duale correspondent respectivement à des arêtes et des sommets de la Surface Accessible, mais l'inverse n'est pas nécessairement vrai. En particulier, les arêtes singulières du complexe dual n'apparaissent pas dans

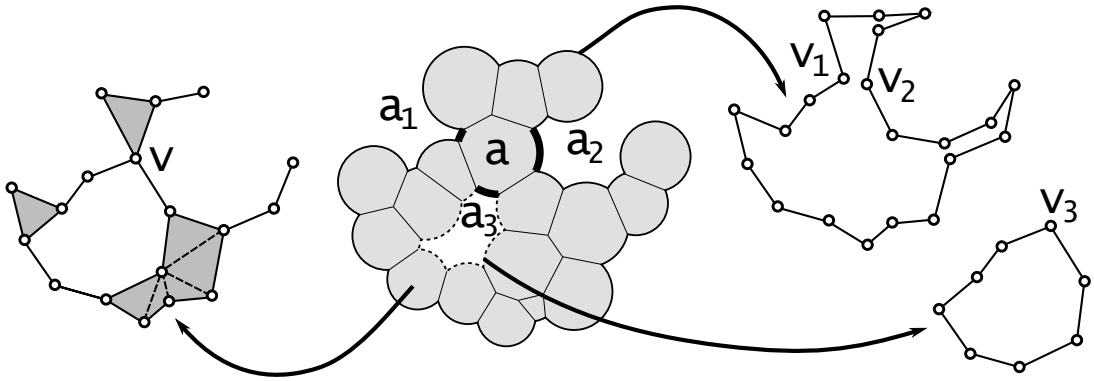


Figure 7.11: Différence d'encodage d'une molécule au travers du complexe dual et de la surface duale, un exemple en deux dimensions. Au centre, une molécule est représentée dans le modèle Surface Accessible, à gauche son complexe dual, et à droite sa surface duale. Comme évoqué dans la section 6.1.2, le complexe dual encode la combinatoire des atomes tronqués (c'est-à-dire leurs relations d'adjascence). En particulier, les atomes sur le bord du complexe dual correspondent exactement aux atomes de surface (c'est-à-dire ceux ayant une ASA strictement positive). La molécule présente deux composantes déconnectées l'une de l'autre : une exposée à l'espace du solvant, l'autre complètement enfouie. Chacune de ces surfaces se voit associer une composante connexe dans la surface duale. Dans le complexe dual les morceaux de surface d'un atome ne sont pas différenciés : à l'atome a correspond le sommet v. Dans la surface duale, chaque morceau de surface d'un atome est différencié : aux morceaux a_1, a_2 et a_3 correspondent respectivement les sommets v_1, v_2 et v_3.

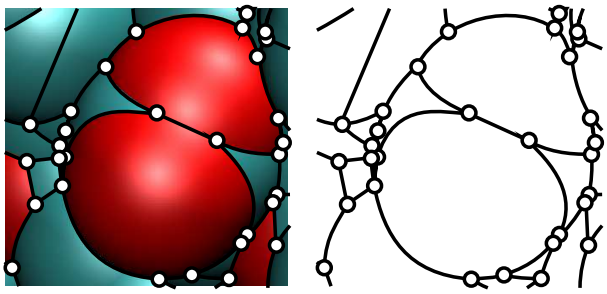


Figure 7.12: Encodage combinatoire de la Surface Accessible. Un détail de la Surface Accessible autour du carbone δ de la *Glu434* présentée dans les figures précédentes. À gauche, la structure combinatoire a été superposée à l'image. À droite, la structure combinatoire représentée seule. Chaque morceau surfacique d'atome constitue une facette de la Surface Accessible. Les arêtes de cette surface correspondent aux arcs de cercles résultants de l'intersection de la surface de deux atomes dans le modèle Surface Accessible. Les sommets correspondent à un point d'intersection de la surface de trois tels atomes.

la surface duale ; elles correspondent pourtant aux arêtes isolées[3] de la Surface Accessible. Dans ses travaux pour la constitution d'un maillage de la surface de Connolly, M.F. Sanner a défini la surface réduite [Sanner 92, Sanner 95, Sanner 96] (ou *reduced surface*) par dualité combinatoire de la Surface Accessible. La notion de dualité combinatoire est explicitée dans la figure 7.13 et illustrée dans le contexte moléculaire dans la figure 7.14. La surface duale constitue un analogue

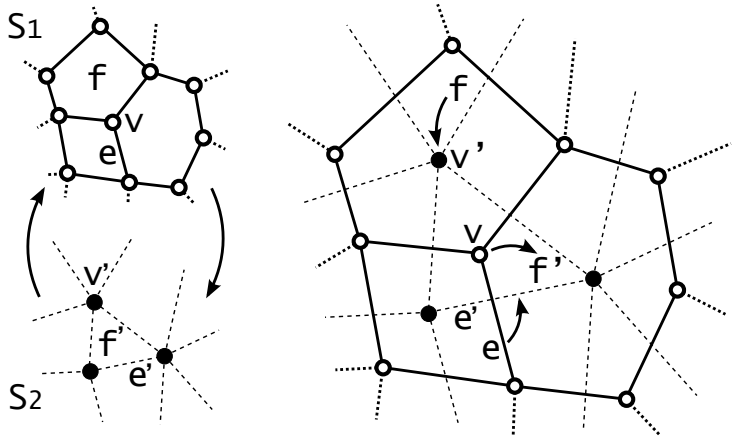


Figure 7.13: Deux structures combinatoires sont dites duales lorsque les facettes, arêtes et sommets de l'une sont en correspondance respectivement avec les sommets, arêtes et facettes de l'autre, et que les relations d'adjacence sont conservées. Les deux structures combinatoires S_1 et S_2 représentées l'une au dessus de l'autre à gauche et superposées dans la même figure à droite sont duales l'une de l'autre. Les facettes, arêtes et sommets f, e et v de S_1 correspondent respectivement aux sommets, arêtes et facettes v', e', et f' de S_2. Cette définition est généralisable en dimensions supérieures.

variété de la surface réduite : en dehors de certaines connectivités non variétés ambigües (cas A et C de la figure 7.6), les deux surfaces sont identiques. De fait, en dehors de ces cas, la surface duale est (combinatoirement) duale de la Surface Accessible. Cette *quasi-dualité* peut-être comprise au travers du trajet d'une sphère solvant virtuelle roulant le long de la surface de Van der Waals de la molécule ; on se rappellera en effet que le modèle Surface Accessible peut être défini comme le lieu des centres d'une telle sphère se déplaçant le long de la surface de Van der Waals. Dans son trajet, représenté dans la figure 7.15, cette sphère virtuelle rencontre trois types de configurations :

- Une position de blocage entre trois atomes de la molécule. Cette configuration correspond à un sommet de la Surface Accessible ainsi qu'à une facette de la surface duale.
- Une circulation en arc de cercle. Lorsque bloquée par deux atomes, elle se déplace d'une "position de blocage" à une autre. Cette configuration correspond à une arête de la Surface Accessible.
- Une circulation "libre", en contact avec un unique atome de la molécule. Cette configuration correspond à un morceau surfacique dans le modèle Surface Accessible, et à un sommet de la surface duale.

L'algorithme 7.2 émule le trajet de cette sphère solvant virtuelle, en se focalisant sur son déplacement d'une position bloquante à une autre, tout en restant systématiquement en contact avec deux atomes. En particulier, la recherche d'une facette voisine autour d'une arête e (à la ligne 4) correspond au passage d'une position de blocage à la suivante, en restant en contact avec les deux atomes dont les sommets dans la surface duale définissent l'arête e.

[3]Les arêtes de la Surface Accessible réalisant un cercle entier sont dites isolées, car déconnectées des autres arêtes de cette surface.

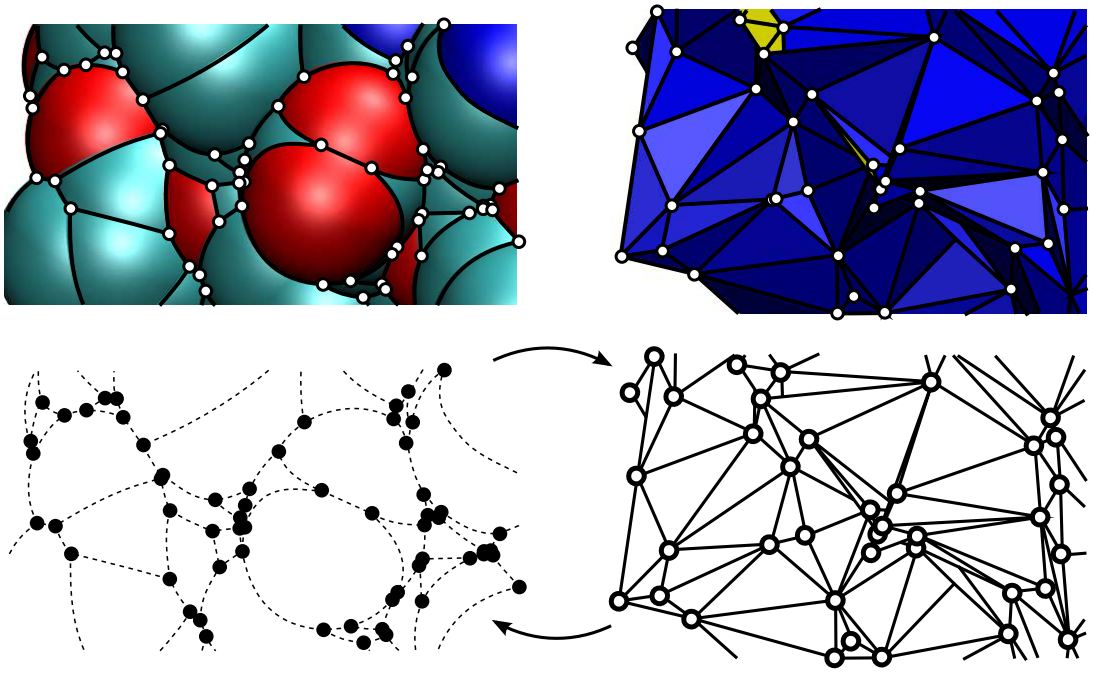


Figure 7.14: Dualité combinatoire de la Surface Accessible et de la surface réduite de M.F. Sanner. Dans le modèle Surface Accessible, la surface est composée de sommets correspondant à l'intersection de trois atomes en surface, d'arêtes (courbes) reliant ces sommets et correspondant à l'intersection de deux atomes en surface, et de facettes (courbes) correspondant à un morceau surfacique d'un atome. Ces éléments combinatoires correspondent respectivement à une facette, une arête et un sommet de la surface réduite.

7.5 Discussion

Nous avons défini la *surface duale*, une surface variété équivalente à la forme duale et qui autorise un parcours émulant le déplacement d'un objet sur les facettes de la forme duale. Cette "variétisation" a nécessité la déconnection (et parfois la supression) de certains sommets et arêtes singulières de la forme duale qui rendaient le parcours ambigu. La surface duale est donc une nouvelle construction adaptée aux contextes et algorithmes nécessitant un parcours le long du bord de la forme duale et pour lesquels les informations de connectivité entre atomes sont moins pertinentes que l'information de surface. Un exemple d'application nécessitant ce genre de données sera présenté dans le chapitre 8 avec l'introduction de nouveaux indices topographiques pour caractériser la surface des molécules.

Notons que dans le cas où l'information de connectivité non variété s'avérerait nécessaire, il serait possible d'ajouter une étape supplémentaire à l'algorithme présenté dans cette partie. Cette étape additionnelle prendrait en charge l'insertion séquentielle des arêtes pendantes (cas de la figure 7.6 *A*) et la reconnection des sommets déconnectés (cas de la figure 7.6 *C*).

Les algorithmes présentés ici ont initialement été développés pour les besoins spécifiques de notre étude en bioinformatique structurale ; des messages sur la liste de diffusion de la librairie `CGAL` faisant état du même besoin de la part d'autres utilisateurs de basculer le bord de la forme alpha dans une structure polyédrique ont motivé une implémentation générique sous la forme de classes template qui peuvent être téléchargées sur ma page web[4].

[4] `http://alnitak.u-strasbg.fr/~schwarz/CGAL-code.html`

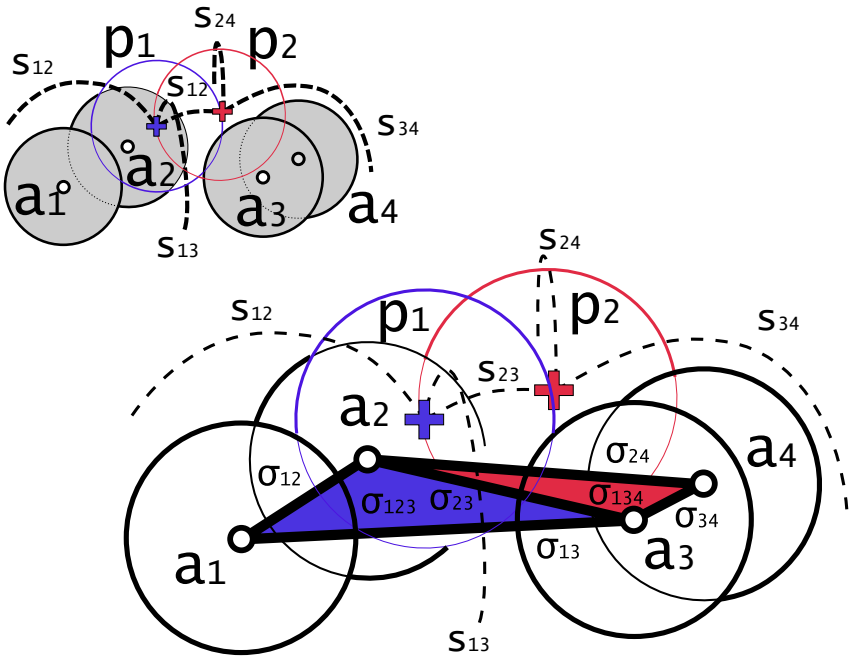


Figure 7.15: Quasi-dualité de la surface duale et de la Surface Accessible. Les quatre atomes a_i sont représentés par des sphères grises dans la petite figure en haut à gauche, et par des sphères transparentes dans la figure principale. Les arcs de cercle pointillés s_{ij} représentent le chemin emprunté par le centre d'une sphère-solvant virtuelle p roulant librement en contact entre les deux atomes a_i et a_j, c'est aussi un segment de la Surface Accessible. Les croix colorées correspondent au centre des sphères-solvants p_1 et p_2 de la même couleur ; ces deux sphères sont chacunes bloquées par trois atomes et leurs centres correspondent chacun à un sommet de la Surface Accessible. La figure principale montre aussi les sommets, arêtes et facettes de la surface duale. Les facettes σ_{ijk} de couleur bleue et rouge dans la surface duale correspondent respectivement aux sommets de la même couleur dans la Surface Accessible, et les arêtes σ_{ij} aux arêtes s_{ij}.

Troisième partie

Applications et résultats

Chapitre 8

Topographier la surface des molécules

La notion de *topographie* de la surface moléculaire a été introduite dans la section 3 page 25. Elle consiste à donner une description du relief de cette surface, essentiellement en termes de "bosses" et de "creux". Dans ce chapitre, nous présenterons la réponse que nous avons apportée à cette problématique. Notre approche exploite la similarité de forme existant entre une molécule et sa forme duale, et utilise la surface duale comme représentation surfacique intuitive de cette dernière[1]. L'idée générale consiste à définir un indice topographique sur la surface duale et à en reporter les valeurs sur les atomes correspondants dans la molécule. Un premier indice, l'exposition-α, est introduit dans la première section de ce chapitre ; il caractérise l'exposition d'un sommet dans la surface duale. Lissé dans le voisinage d'un sommet de la surface duale, cet indice permet de définir la *courbure locale* comme une "tendance à être exposé" dans le voisinage de ce sommet. Ce second indice est présenté dans la seconde section. La troisième section présentera l'application de ces deux indices pour la caractérisation topographique de la surface de macromolécules biologiques. Enfin, la quatrième section conclura ce chapitre et présentera des pistes pour améliorer l'existant.

8.1 Exposition-α, une mesure de l'exposition des sommets de la surface duale

8.1.1 Définition de l'exposition-α

Pour chaque sommet v de la surface duale, on définit l'exposition-α Ω_v comme l'angle solide de l'ombrelle au dessus de v (voir la figure 8.1). Nous utilisons des valeurs d'angle solide normalisées, correspondant à la proportion de l'aire laissée libre par l'ombrelle sur une sphère. De cette façon, un sommet dans une zone concave a une valeur d'exposition-α proche de 0, à l'inverse d'un sommet sur une zone saillante qui aura une valeur proche de 1. Cette définition est illustrée dans l'exemple en deux dimensions de la figure 8.2. On se rappellera que chaque sommet de la surface duale correspond à une ombrelle autour d'un sommet du bord du complexe dual. En deux dimensions, une telle ombrelle est composée d'un sommet et de deux arêtes qui lui sont incidentes. Dans le cas d'arêtes singulières, les deux arêtes de l'ombrelle peuvent être confondues ; c'est par exemple le cas pour l'ombrelle du sommet c de la figure ; complètement exposé à l'espace vide, il saura une valeur d'exposition-α de 1. Au sommet b du complexe dual correspond un unique sommet b dans la surface duale, à l'inverse des sommets a et d donnant chacun lieu à deux

[1] La forme duale a été définie au chapitre 6.1.2 page 61, et le chapitre 7 est entièrement dévolu à l'étude de la surface duale.

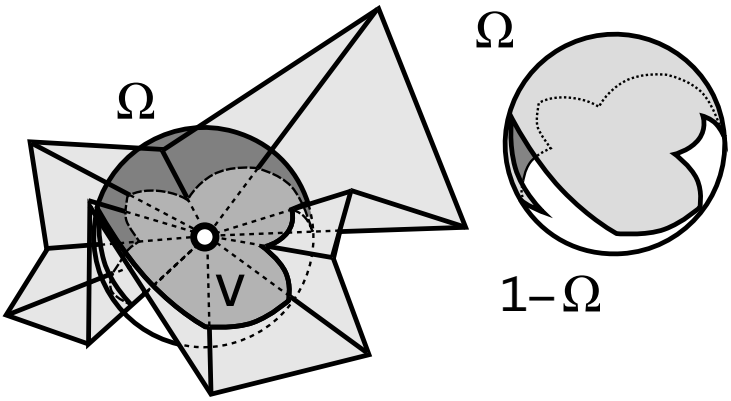


Figure 8.1: Définition de l'exposition-α d'un sommet v de la surface duale comme la "portion de vide" visible au dessus de ce sommet (à gauche). La figure de droite isole la trace laissée par l'ombrelle au dessus de v sur une sphère ; l'exposition-α du sommet v est donnée par Ω, la proportion d'aire grise rapportée à l'aire totale de la sphère.

ombrelles et à deux sommets dans la surface duale. Deux sommets de la surface duale issus d'un même sommet sur le bord du complexe dual auront potentiellement des valeurs d'exposition-α différentes. C'est par exemple le cas des sommets a_1 et a_2 issus du même sommet a, avec respectivement des valeurs approximatives de $\frac{1}{2}$ et $\frac{1}{4}$.

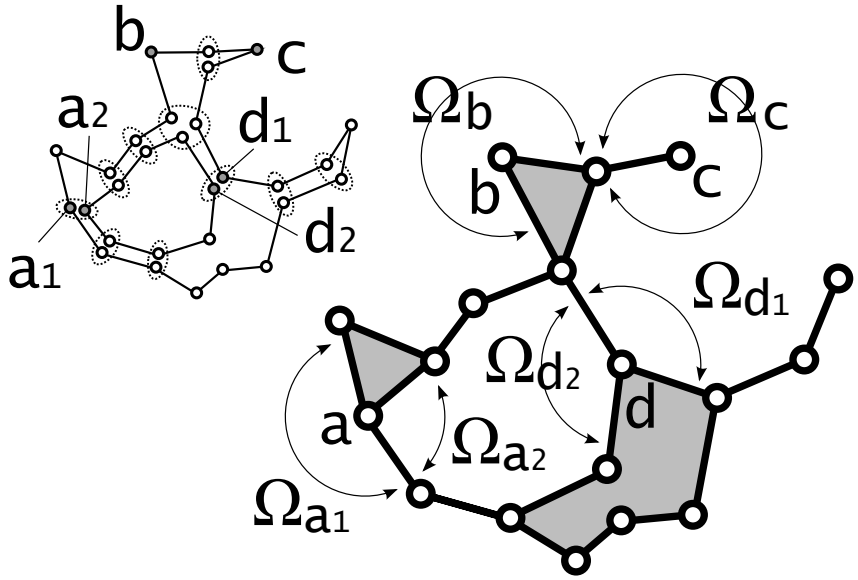


Figure 8.2: Définition de l'exposition-α sur un exemple en deux dimensions. Le bord d'un complexe dual a été représenté avec sa surface duale associée (en haut à gauche). Plusieurs sommets de la surface duale (entourés d'une ligne en pointillés) peuvent correspondre à un même sommet sur le bord du complexe dual ; ils isolent des ombrelles distinctes autour de ce sommet. L'exposition-α du sommet a_1 de la surface duale est donnée par l'angle Ω_{a_1} défini par l'ombrelle associée.

8.1.2 Calcul de l'exposition-α

Le calcul proposé ici de l'exposition-α pour un sommet v de la surface duale repose sur quelques remarques préliminaires concernant le remplissage de la partie supérieure d'une ombrelle.

Remarques préliminaires et notion de faisceau d'ombrelle

Dans un premier temps, on remarque que le voisinage d'un sommet peut être différencié selon son appartenance à l'enveloppe convexe des centres atomiques. Comme illustré dans la figure 8.3, le voisinage d'un sommet de l'enveloppe convexe comprend invariablement une partie en dehors de cette enveloppe (en rouge). À l'inverse, un sommet à l'intérieur de l'enveloppe convexe comprend un voisinage entièrement à l'intérieur de l'enveloppe convexe. Dans la suite,

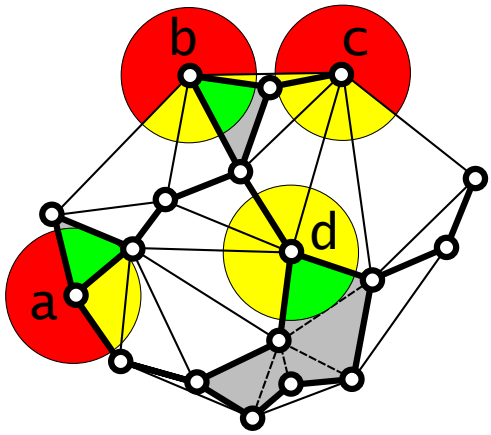


Figure 8.3: Décomposition du faisceau d'une ombrelle pour le calcul de l'exposition-α, un exemple en deux dimensions. Le complexe dual de la figure 8.2 a été superposé à la triangulation de Delaunay dont il est issu. Le sommet d, intérieur à l'enveloppe convexe des centres atomiques, est entièrement entouré de triangles de Delaunay ; il comporte ainsi un voisinage totalement inclus dans l'enveloppe convexe. Les parties jaunes dénotent un voisinage intersectant un triangle intérieur au complexe dual, en vert l'intersection avec des triangles intérieurs. Les sommets sur le bord de l'enveloppe convexe des centres atomiques, comme a, b et c, ont une partie de leur voisinage qui n'est pas recouverte par des triangles de Delaunay (en rouge), et qui se trouve par là même en dehors de l'enveloppe convexe.

on parlera de *faisceau* d'une ombrelle (ou de son sommet) pour désigner l'espace vide au dessus de celle-ci, et on différenciera deux parties d'un faisceau en fonction de son intersection ou non avec l'enveloppe convexe. On parlera de *composante infinie* d'un faisceau d'ombrelle pour la partie extérieure à l'enveloppe convexe (en rouge dans la figure 8.3), et de *composante finie* pour la partie intersectant l'enveloppe convexe (en jaune dans la même figure). De ces définitions et observations on extrait les remarques suivantes.

Remarque 8.1. *La composante finie d'un faisceau d'ombrelle correspond à un ensemble de tétraèdres de Delaunay extérieurs au complexe dual.*

Remarque 8.2. *Seuls les sommets de l'enveloppe convexe peuvent présenter une composante infinie.*

Remarque 8.3. *Si les centres atomiques ne sont pas tous coplanaires, un sommet du complexe dual ne peut avoir plus d'une ombrelle avec une composante infinie.*

Les deux premières remarques découlent des définitions d'ombrelle (données page 84) et de celles de faisceaux finis et infinis. La troisième remarque découle de la définition d'enveloppe convexe. En effet, si les centres atomiques ne sont pas tous coplanaires, l'enveloppe convexe est un polyèdre convexe. Au voisinage de tout point de sa surface, et en particulier de ses sommets, l'espace est déconnecté en deux composantes connexes : une partie interne à l'enveloppe convexe et une composante qui lui est extérieure, la *composante infinie*. Cette composante ne peut être

partagée entre plusieurs ombrelles car toutes les facettes d'une ombrelle de la surface duale sont contenues dans l'enveloppe convexe. On remarquera en outre que dans le cas d'une molécule complètement plate, l'enveloppe convexe est un polygone. Dans ce cas particulier, les seules valeurs possibles d'exposition-α sont 1 et $\frac{1}{2}$ suivant que l'atome considéré se trouve sur le bord du complexe dual ou dans son intérieur. Ces remarques suggèrent une méthode de calcul de l'exposition-α d'un sommet par séparation de ses composantes finies et infinies, calcul de leurs angles solides respectifs Ω° et Ω^∞, et somme de ces angles solides.

$$\Omega = \Omega^\circ + \Omega^\infty \tag{8.1}$$

Calcul de l'angle solide dans la composante finie d'un faisceau d'ombrelle

On définit $\circ_v$ le *faisceau de tétraèdres* d'une ombrelle (ou de son sommet v) comme l'ensemble des tétraèdres de Delaunay recouvrant la composante finie du faisceau de l'ombrelle. La remarque 8.1 autorise cette définition et permet de préciser que le faisceau de tétraèdres comprend uniquement des tétraèdres issus de 3-simplexes extérieurs au complexe dual. Cette remarque suggère également de calculer l'angle solide de la composante finie d'une ombrelle comme la somme des angles solides pris séparément dans chaque élément du faisceau de tétraèdres de cette ombrelle.

$$\Omega^\circ = \sum_{\tau \in \circ_v} \omega_\tau \tag{8.2}$$

Ce calcul est illustré en deux dimensions dans la figure 8.4. Le calcul de l'angle solide ω_i au

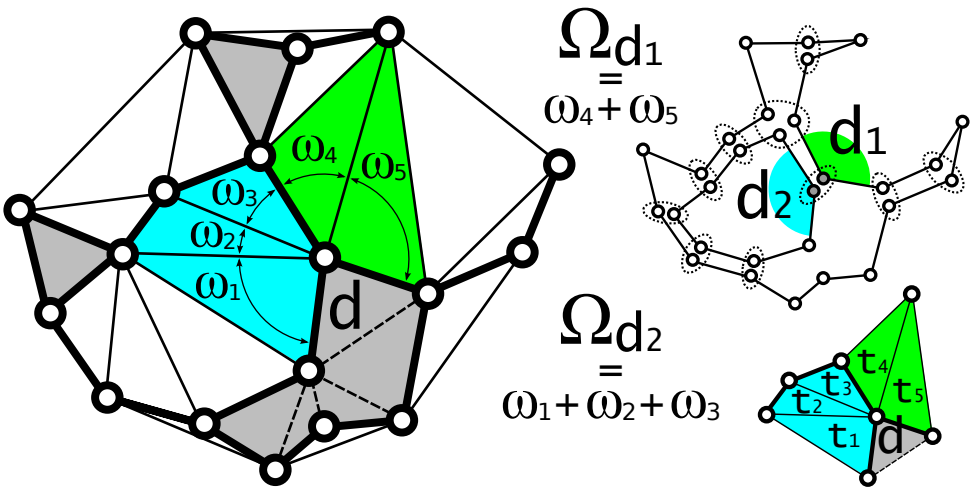


Figure 8.4: Calcul de l'angle solide d'une composante finie dans la triangulation de Delaunay, un exemple en deux dimensions. À gauche, le complexe dual et sa triangulation de Delaunay. En haut à droite, la surface duale associée. Les faisceaux associés aux deux ombrelles du sommet d ont été matérialisées par des demi-cercles colorés. Ces deux faisceaux sont finis. En bas à droite, un détail de la figure de gauche autour du sommet d. Les faisceaux des ombrelles associées à d_1 et d_2 sont recouverts respectivement par les faisceaux de triangles $\{t_4, t_5\}$ et $\{t_1, t_2, t_3\}$. Les valeurs d'exposition-α de ces deux sommets correspondent respectivement à la somme des angles ω_i de ces triangles pris au sommet d.

sommet i d'un tétraèdre de sommets i, j, k, l s'exprime en fonction de ses trois angles diédraux. Les notions d'angle solide et d'angle diédral sont illustrées dans la figure 8.5 et les formules rappelées ci-après. L'angle solide w_i pris au sommet i d'un tétraèdre est donné par

$$\omega_i = \frac{\varphi^i_{j,l,k} + \varphi^i_{k,j,l} + \varphi^i_{l,k,j}}{2} - \frac{1}{4}$$

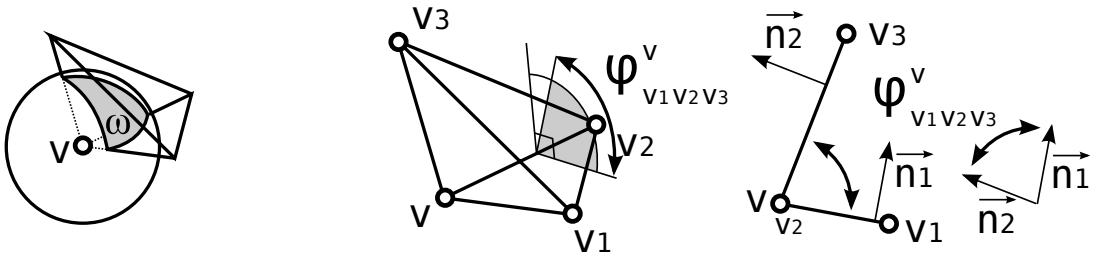


Figure 8.5: Angle solide et angle diédral dans un tétraèdre. La figure de gauche montre un tétraèdre et une sphère centrée sur v, un des quatre sommets de celui-ci. L'angle solide $\omega = \omega_v$ du tétraèdre en son sommet v est donné par la portion de surface (en gris) que le tétraèdre dessine sur la sphère. Les deux figures de droite illustrent géométriquement la notion d'angle diédral $\varphi^v_{v_1,v_2,v_3}$. Cette valeur mesure l'angle entre les deux triangles $\{v, v_2, v_1\}$ et $\{v, v_2, v_3\}$. La figure de droite montre les mêmes triangles vus de "profil". L'angle diédral $\varphi^v_{v_1,v_2,v_3}$ peut aussi être pris entre les deux normales $\overrightarrow{n_1}$ et $\overrightarrow{n_2}$ à ces triangles.

où $\varphi^i_{j,l,k}$ est l'angle diédral entre les triangles $\{i, l, k\}$ et $\{i, l, j\}$, qui se calcule par

$$\varphi^i_{j,l,k} = \frac{\arccos(\overrightarrow{n_1} \times \overrightarrow{n_2})}{2\pi}$$

avec $\overrightarrow{n_1}$ et $\overrightarrow{n_2}$ des vecteurs unitaires orthogonaux respectivement aux triangles précédemment cités.

Calcul de l'angle solide dans la composante infinie d'un faisceau d'ombrelle

Cette composante de faisceau n'étant pas explicitement recouverte par des tétraèdres, le calcul exposé dans la partie précédente n'est pas directement applicable. Grâce à la remarque 8.3, on peut affirmer que le complémentaire de la composante infinie du faisceau d'ombrelle de v est entièrement recouvert par des tétraèdres de Delaunay ; et plus précisément par l'ensemble des tétraèdres de Delaunay incidents au sommet v. De fait, l'angle solide de la composante infinie Ω^∞ peut être calculé de manière équivalente à celui de la composante finie par

$$\Omega^\infty = 1 - \Omega^\mathcal{D} \tag{8.3}$$

où $\Omega^\mathcal{D}$ désigne la somme des angles solides mesurés en v pour chaque tétraèdre de Delaunay incident au sommet v.

8.2 Courbure locale, une mesure de l'incurvation de la surface duale

La courbure locale d'un sommet v de la surface duale est calculée en "lissant" les valeurs d'exposition-α calculées dans la section précédente. Dans les deux sous-sections à suivre nous préciserons la notion de voisinage que nous avons utilisée ainsi que les types de lissages que nous avons implémentés. Dans la dernière sous-section nous présenterons le principe de re-échantillonage que nous avons implémenté pour augmenter la lisibilité des valeurs de courbure dans les applications visuelles. Nous y définirons des notations utilisées dans la suite du document.

8.2.1 Constitution d'un voisinage sur la surface duale

On définit V^N_v, le voisinage de taille N d'un sommet v de la surface duale, comme l'ensemble des sommets de la surface duale atteignables depuis le sommet v en parcourant au plus N arêtes

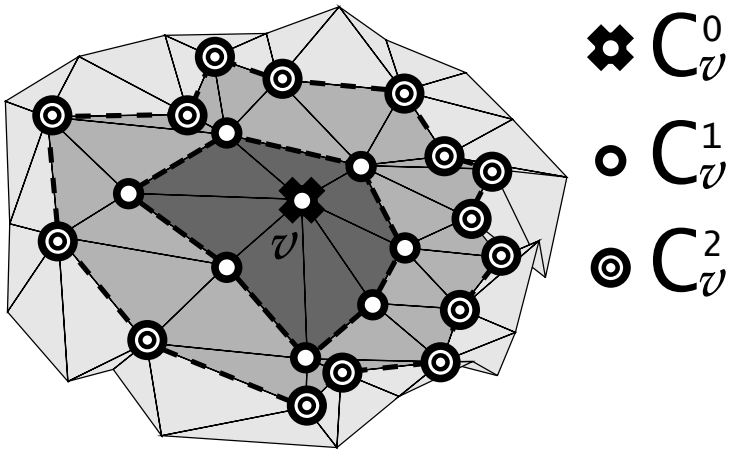


Figure 8.6: Voisinage le long de la surface duale pour le calcul de la courbure locale. Le voisinage de taille 0 du sommet v est réduit au sommet v lui même.

consécutives sur la surface duale (voir la figure 8.6). Dans la suite du document on parlera de la n-ième *corolle* d'un sommet v pour désigner l'ensemble des sommets de V_v^n n'appartenant pas à V_v^{n-1}. C'est l'ensemble des sommets que l'on peut atteindre depuis v en n arêtes consécutives, et que l'on ne peut atteindre en moins ; on notera C_v^n la n-ième *corolle* de v.

8.2.2 Lissage de l'exposition-α

La *courbure locale* $lc(v)$ d'un sommet v est obtenue en moyennant les valeurs d'exposition-α dans un voisinage de v. Trois types de moyennes sur un ensemble de sommets E ont été considérées et implémentées, où nous avons systématiquement pris le voisinage V_v^N comme "ensemble de lissage" E :

- La *moyenne arithmétique* :
$$\mu_E^a(v) = \frac{1}{|E|} \sum_{w \in E} \alpha(w)$$
- La *moyenne quadratique* :
$$\mu_E^q(v) = \frac{1}{|E|} \sqrt{\sum_{w \in E} \alpha(w)^2}$$
- Une *moyenne pondérée* :
$$\mu_E^w(v) = \sum_{w \in E} W_v(w)\alpha(v)$$
où $W_v(w)$ désigne une "fonction de poids" qui, à un sommet de la surface duale associe une valeur dans $\mathbb{R}^+$.

Comme pondération des sommets pour μ^q, nous avons utilisé la proximité au centre du voisinage : $W_v(w)\alpha(v) = \frac{N-i}{|E|}$, où i désigne l'indice de la corolle de v à laquelle le sommet w appartient[2]. D'autres types de pondérations pourraient aussi être étudiées, par exemple en prenant en compte une distance géodésique au centre le long de la surface duale, ou un noyeau gaussien basé sur cette dernière valeur. Un de nos objectifs étant de privilégier, nous avons considéré ce simple éloignement en nombre d'arêtes comme une approximation qui s'est avérée suffisante dans nos

[2] c'est-à-dire $i \mid w \in C_v^i$

applications.

La courbure locale $lc(v) = lc^{\mu}_{N}(v)$ d'un sommet de la surface duale est donc une valeur dépendant de deux paramètres : la taille du voisinage considéré et la méthode μ de lissage des valeurs d'exposition-α sur ce voisinage. Dans la suite du document, on notera lc^{a}_{N}, lc^{q}_{N} et lc^{w}_{N} les courbures locales obtenues respectivement avec les lissages arithmétiques, quadratiques et pondérés.

8.2.3 Re-échantillonnage des valeurs

De manière générale, un procédé de lissage semblable à celui décrit dans la section précédente peut être utilisé pour révéler les tendances locales d'une fonction ; le paramètre de "taille" permettant de contrôler la notion de localité. Cette opération a néanmoins pour effet "d'aplanir" les tendances de la fonction et de réduire sa plage de valeurs, ce qui peut s'avérer génant, au moins pour une inspection visuelle. Cet inconvénient peut être circonvenu par une opération de re-échantillonnage présentée ci-après.

La figure 8.7 montre un exemple de lissage de valeurs d'une courbe dans le plan, en moyenne arithmétique sur des voisinages de taille 0, 1, 4, 13 et 19. Ce type de lissage est aussi appelé *moyenne glissante* ou *moyenne mobile*. Dans cet exemple en une dimension, la présence d'un "bord" sur le support de la fonction f induit des difficultés pour la définition d'un voisinage de lissage, aussi le domaine[3] des fonctions lissées est-il réduit. Dans le cas d'un support sans bord, ces difficultés n'existent pas et le lissage ne réduit pas la taille du support. C'est par exemple le cas pour les fonctions cycliques, et plus généralement pour les fonctions définies sur les 2-variétés sans bord comme la surface duale. En modulant la taille du voisinage de lissage on peut explorer les tendances de la courbe à des échelles différentes. Comme illustré dans la figure 8.7, avec une taille de voisinage N croissante, la plage de valeurs des fonctions lissées f_N a tendance à diminuer pour "s'écraser" autour d'une valeur moyenne. Une méthode pour pallier ce problème consiste à re-échelonner les valeurs des fonctions lissées dans l'intervalle défini par l'image de la fonction f. Soit à définir une fonction

$$\lambda_N : Im(f_N) \rightarrow Im(f)$$

et des fonctions f'_N par composition avec les λ_N

$$f'_N = \lambda_N \circ f_N$$

Cette méthode a été illustrée en une dimension dans la figure 8.8 en reprenant le lissage de taille 4 de l'exemple précédent, et en choisissant pour fonction λ_4 l'application linéaire réalisant une bijection de $Im(f_4)$ sur $Im(f)$. Un tel re-échantillonnage linéaire n'est pas toujours suffisant pour accentuer les valeurs "pertinentes" d'une fonction lissée, en particulier si l'on est plus intéressé par les fortes "tendances" de la fonction que par ses valeurs effectives. Une méthode simple pour révéler cette information consiste à amplifier le re-échantillonnage linéaire et à "seuiller" les valeurs sortant de l'image de f. Comme indiqué dans la figure 8.9, un tel procédé peut aussi se concevoir comme un re-échantillonnage tronqué d'un sous intervalle $[m, M]$ de l'image de la fonction lissée. Dans cette figure, le dégradé de couleur sur les graphes permet une autre perception visuelle des valeurs des fonctions. Le même code couleur (bleu pour les valeurs élevées de la fonction f et rouge pour ses valeurs basses) sera repris ultérieurement dans les illustrations moléculaires où le même procédé de lissage – re-échantillonnage est appliqué. Dans la suite du document on indicera par $m:M$ pour désigner un tel lissage tronqué. En particulier $lc^{\mu}_{N:m:M}$ désignera la

[3]Le domaine (d'application) d'une fonction f est l'ensemble $\mathcal{D}_f$ sur lequel cette fonction est définie.

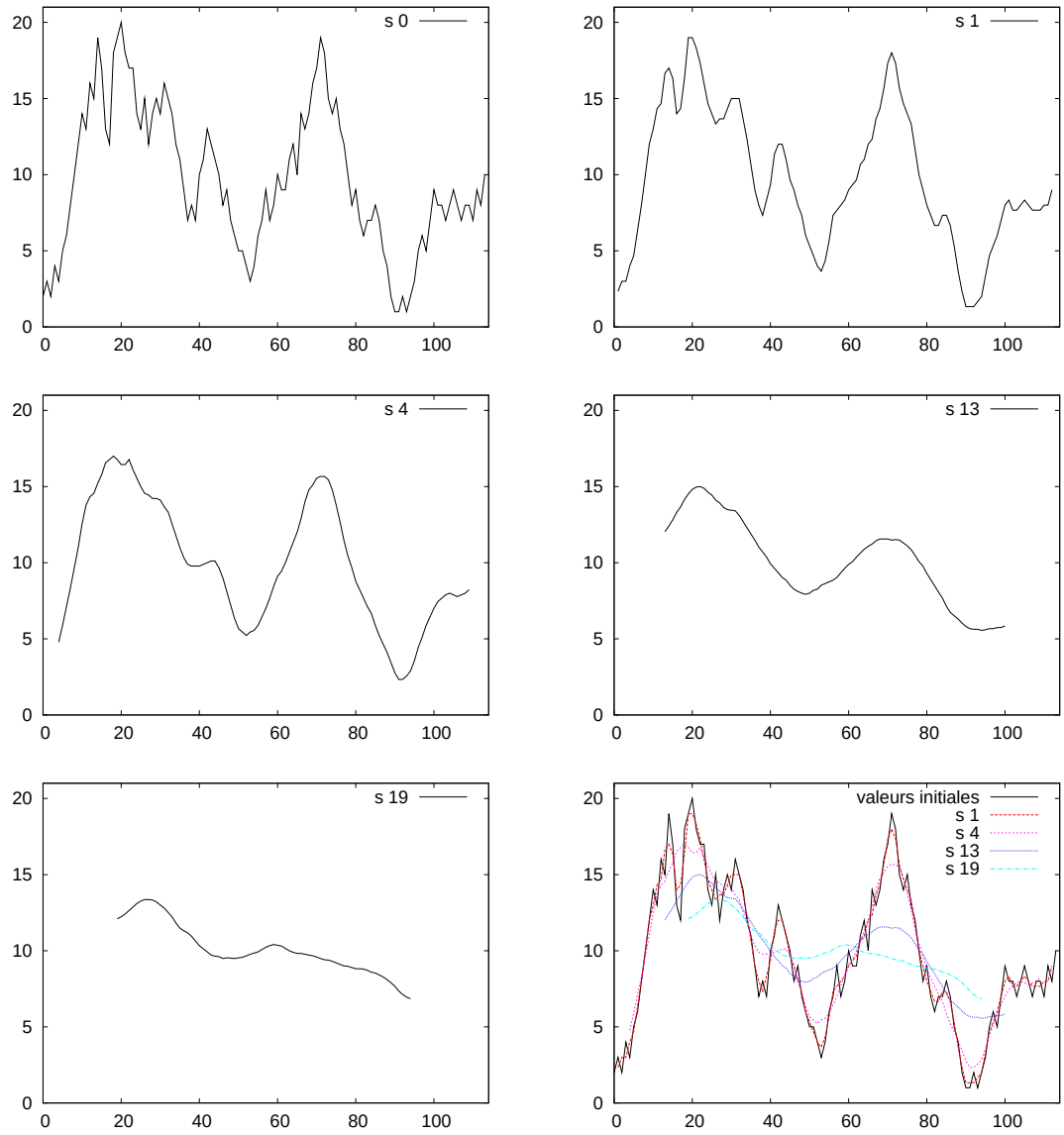


Figure 8.7: Lissage de valeurs discrètes. Le premier graphe (en haut à gauche) représente une fonction discrète f à valeurs réelles. Les quatre graphes suivants représentent les fonctions obtenues par moyenne mobile, respectivement pour les lissages de taille $s = 1$ (fenêtre de trois valeurs), 4 (neuf valeurs), 13 (vingt sept valeurs), 19 (trente neuf valeurs). Les fonctions lissées retiennent la "tendance" générale de la fonction qu'elles lissent ; l'utilisation d'une fenêtre de lissage "réduite" ou "large" permet d'explorer des tendances entre le local et le global. Le dernier schéma superpose le graphe de f et ses quatre lissages. L'augmentation de la taille de la fenêtre de lissage permet d'explorer l'allure générale de la courbe, mais réduit le domaine fonctionnel et "tasse" la fonction autour d'une valeur moyenne.

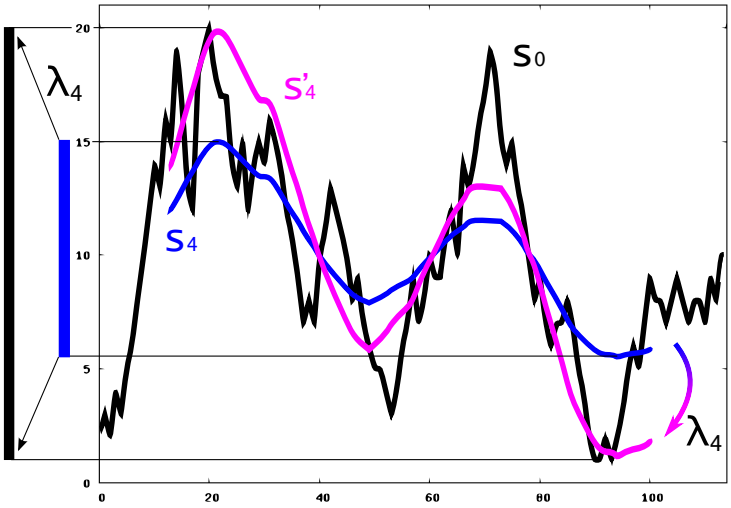


Figure 8.8: Re-échantillonnage linéaire de valeurs lissées pour la fonction f de la figure 8.7. Le graphe S_0 de la fonction f est représenté en noir, sa plage de valeurs $Im(f)$ est matérialisée par un rectangle noir à gauche de la figure. De la même manière, le graphe s_4 du lissage f_4 de taille 4 et sa plage de valeur ont été représentés en bleu. Les valeurs de ce lissage sont ensuite linéairement réparties dans l'image de la fonction f pour obtenir f'_4. Cet étirement correspond à une composition de la fonction de lissage par une application linéaire λ_4.

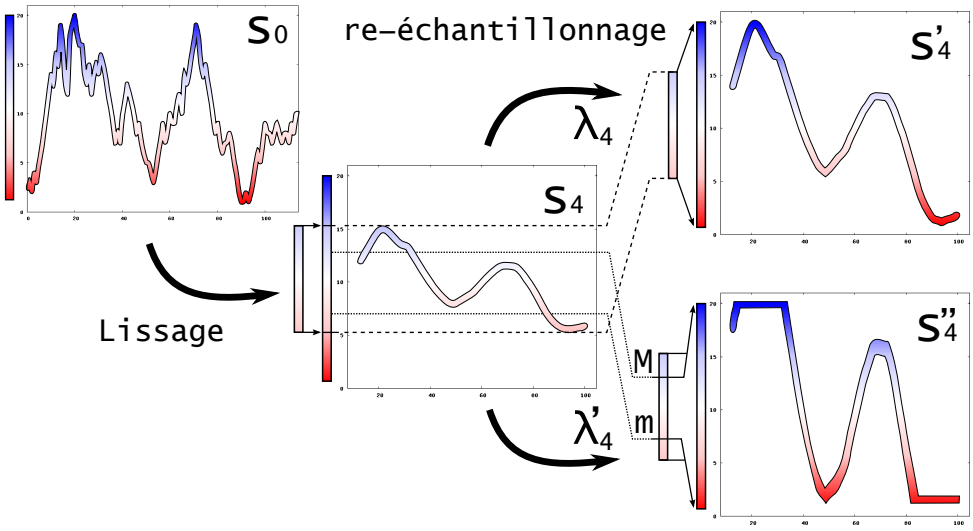


Figure 8.9: Re-échantillonnage linéaire tronqué de valeurs lissées pour la fonction f de la figure 8.7. Le graphe des fonctions est agrémenté d'un code couleur allant du rouge pour représenter les petites valeurs au bleu pour les valeurs importantes. Le graphe s_0 de la fonction f est représenté à gauche et le graphe s_4 de la fonction lissée f_4 au milieu. Les fonctions f'_4 et f''_4, de graphes respectifs s'_4 et s''_4, sont obtenues par composition, respectivement avec les applications λ_4 et λ'_4 où λ_4 est la bijection linéaire de la figure 8.8 et λ'_4 réalise une bijection linéaire d'un intervalle $[m, M]$ inclus dans l'image de f_4. Les valeurs de f_4 supérieures à M et inférieures à m sont respectivement envoyées sur les bornes supérieures et inférieures de l'image de f. f''_4 sera notée $f_{4:m:M}$.

courbure locale pour un lissage de taille N avec la méthode de lissage μ et un re-échantillonnage linéaire tronqué à l'intervalle $[m, M]$. La sélection automatique de "bonnes" bornes m et M est une problématique complexe. Une première méthode simple consiste à réduire le choix arbitraire de ces deux variables au choix, tout aussi arbitraire mais plus intuitif, d'une unique variable X contrôlant le pourcentage de valeurs extrêmes à seuiller. Dans notre contexte, la fonction f est la courbure locale, et son domaine d'application — l'ensemble des sommets de la surface duale — est fini. Ce seuillage automatisé peut donc être effectué rapidement en classant les sommets sur leur valeur de courbure locale, et en sélectionnant les bornes de re-échantillonnage de manière à attribuer respectivement les valeurs 0 et 1 aux X premiers pourcents et aux X derniers. Dans la suite du document, $lc^{\mu}_{N:X\%}$ désignera la courbure locale pour un lissage de taille N avec la méthode de lissage μ et un re-échantillonnage linéaire tronqué à $X\%$. On pourra aussi spécifier les valeurs de bornes m et M induites par le choix de X en écrivant explicitement $lc^{\mu}_{N:X\%:m:M}$.

8.2.4 Estimation du temps de calculs de la courbure locale

Les algorithmes de calcul de l'exposition-α et de la courbure locale ont été implémentés et testés sur un jeu de données non redondant de 2305 structures dont la constitution est précisée en annexe C page 187. Les temps de calculs présentés dans la figure 8.10 ont été mesurés sur une machine Sun x4200 M2 à 2 Opteron 2220 (2800 MHz - Dual Core). Ces temps mesurent

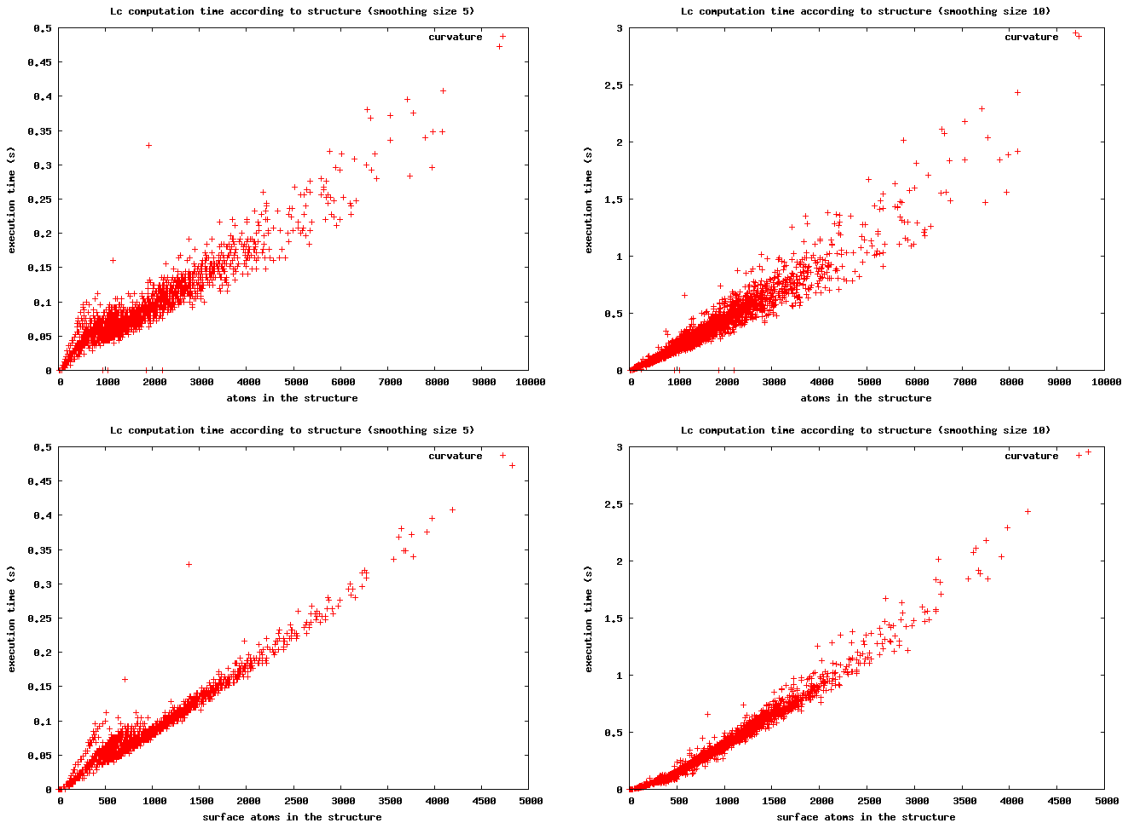


Figure 8.10: Temps de calcul de la courbure locale sur une base de 2305 structures. A chaque structure est attribuée un point dans le plan dont l'abscisse représente le nombre d'atomes N considérés dans la structure et l'ordonnée le temps de calcul des R premières valeurs de courbure locale. R vaut 5 dans la colonne de gauche et 10 dans celle de droite. Dans la ligne du haut N représente le nombre d'atomes dans la structure, et uniquement le nombre d'atomes de surface dans la ligne du bas.

uniquement le calcul de l'exposition-α et de la courbure locale pour des rayons de lissage de 1 à R arêtes avec $R = 5$ et $R = 10$. Pour traiter les problèmes de reproductibilité inhérents à tout environement multitâche, les mesures ont été prises sept fois pour chaque structure et nous avons utilisé la médiane de ces sept valeurs pour constituer nos graphes. Le temps s'échelonne linéairement en fonction du nombre d'atomes considéré, qu'il s'agisse du nombre total d'atomes dans la structure ou du nombre d'atomes constituant la surface (ASA>0) de la molécule. La tendance linéaire est plus franche lorsqu'on considère uniquement les atomes de surface, ce qui s'explique par le fait que les valeurs de courbure locale sont plus directement liées à ces atomes.

Le temps de calcul des valeurs de courbure locale dépend du rayon de lissage considéré, une zone de lissage plus large impliquant naturellement plus d'atomes. Pour une molécule de taille conséquente (10000 atomes), ce temps est d'une demi seconde si l'on s'arrête au calcul des cinq premiers lissages, un rayon observé suffisant pour la plupart des applications ; il passe à trois secondes si l'on calcule les dix premières valeurs.

8.3 Passage des valeurs d'indices de la surface duale à la surface moléculaire

Comme évoqué dans l'introduction de ce chapitre, les indices topographiques (exposition-α ou courbure locale) calculés pour les sommets de la surface duale sont reportés aux atomes correspondants dans la molécule, Comme on l'a vu au chapitre 7.4.3 (page 85), tous les atomes de la molécule ne sont pas en correspondence biunivoque avec un sommet de la surface duale ; on souhaite pourtant associer une valeur d'indice topographique à chacun d'eux.

Les atomes intérieurs[4] à la molécule, par exemple, ne sont associés à aucun sommet de la surface duale. Par convention, ces atomes "internes" sont affublés d'une valeur d'indice nulle, représentant leur "enfouissement" dans la molécule.

D'autres atomes, bien que présents à la surface de la molécule, ne sont également associés à aucun sommet de la surface duale ; ce sont les atomes situés au bout d'une arête pendante de la forme duale ou à la jonction de deux telles arêtes. Ces atomes étant, selon notre définition, particulièrement exposés, nous leur associeront une valeur d'indice de 1. Il est à noter que ce cas de figure est relativement rare à la surface d'une molécule et dénote généralement d'un résidu pointant hors de la molécule. Les valeurs d'indice topographique pour les atomes de la molécule peuvent donc êtres calculées par la succession de trois procédures

1. les valeurs d'indice des atomes de la molécule sont mises à 0 ;
2. les valeurs d'indice des atomes sur le bord de la forme duale sont mises à 1 ;
3. les valeurs d'indice sont calculées pour les sommets de la surface duale et reportées aux atomes correspondants dans la molécule.

Pour la réalisation de cette dernière procédure, il convient de remarquer le cas particulier des atomes associés à plusieurs sommets de la surface duale. Ces atomes présentent plusieurs "morceaux surfaciques" dans le modèle Surface Accessible et correspondent aux sommets de multiplicité supérieure à 1 dans la surface duale. Une valeur d'indice topographique a été calculée pour chaque sommet correspondant à un tel atome, autrement dit pour chacun de ses morceaux surfaciques, et il convient de choisir une manière de les amalgamer pour générer une valeur d'indice propre à l'atome entier. Quatre choix simples ont été implémentés et sont proposés dans le logiciel `Lc` : la moyenne des valeurs d'indice calculées pour chaque morceau surfacique, leur somme, la plus petite valeur calculée pour chaque morceau surfacique, ou encore la plus grande.

[4] On se rappelera ici que l'on considère comme intérieur à la molécule tout sommet d'ASA nulle.

Par défaut, dans la suite du document, c'est la moyenne des valeurs qui sera systématiquement choisie, car nous avons constaté que de toutes ces solutions, c'était celle qui perturbait le moins la preception visuelle.

8.4 Application de la courbure locale aux macromolécules biologiques

Différentes valeurs de courbure locale ont été calculées et reportées à la surface de diverses macromolécules biologiques. Dans une première section nous ferons quelques remarques générales en prenant pour exemple les valeurs d'exposition-α et de courbure locale calculées à la surface d'un récepteur nucléaire. Nous nous attacherons ensuite à montrer, par le biais de plusieurs exemples biologiques, l'intérêt de ces valeurs pour faire ressortir des informations pertinentes.

8.4.1 Topographier la surface de macromolécules biologiques avec la courbure locale

Comme on peut s'en rendre compte dans la figure 8.11, la répartition des valeurs d'exposition-α est assez disparate à la surface d'une molécule : que l'on observe une région proéminente ou une région invaginée, on trouve des atomes très exposés qui en côtoient d'autres très enfouis. Bien que

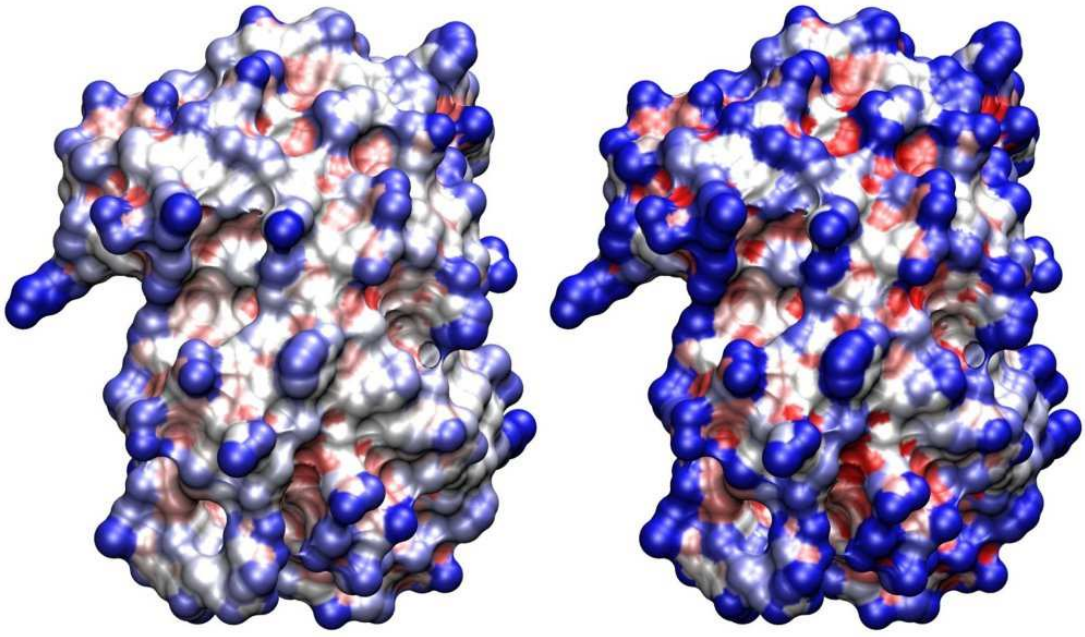


Figure 8.11: Exposition-α à la surface du domaine de liaison au ligand du récepteur nucléaire à l'acide rétinoïque RXR-α (1RDT chaîne A). À gauche, les valeurs d'exposition-α brutes, à droite les mêmes valeurs seuillées à 15% ($lc^{w}_{0:15\%:.217:.754}$).

la répartition de l'exposition soit chaotique, les valeurs sont homogénéisées dès le premier lissage de l'exposition-α (voir figure 8.12), ce qui révèle que le nombre de sommets de forte (ou faible) exposition dans un voisinage donné est révélateur de la protrusion de cette région. Cette figure révèle aussi que dès le premier lissage les valeurs de protrusion sont fortement rapprochées de la moyenne, et permettent difficilement une distinction visuelle des caractéristiques topographiques. Il est nécessaire de la re-échantillonner et de la seuiller pour accentuer les zones proéminentes et

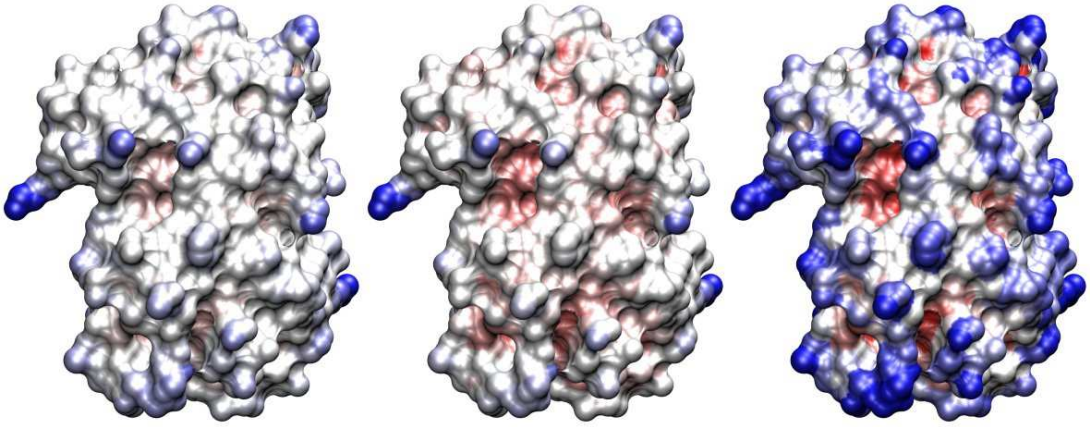


Figure 8.12: Courbure locale lissée à la surface du domaine de liaison au ligand du récepteur nucléaire à l'acide rétinoïque RXR-α (1RDT chaîne A). De gauche à droite, les valeurs lissées pondérées dans un voisinage de taille 1 (lc^w_1), les mêmes valeurs re-échantillonnées linéairement ($lc^w_{1:0\%:.16:.1}$), les mêmes valeurs seuillées à 5% ($lc^w_{1:5\%:.294:.671}$).

invaginées. La figure 8.13 montre la surface du même récepteur nucléaire pour un lissage pondéré, des tailles du voisinage de lissage de 0 à 5 et plusieurs niveaux de seuil. L'influence des variables sur la visualisation y est flagrante, de même que la capacité de la courbure locale à détecter les caractéristiques topographiques intuitives telles que les "creux" et les "bosses".

8.4.2 Topographie d'un récepteur nucléaire

Les récepteurs nucléaires constituent une famille de protéines partageant une organisation commune en 5 à 6 domaines conservés [Laudet 01]. Leur activité concerne l'initiation et la régulation de la transcription de l'ADN (un exemple simplifié de ce mécanisme est donné à l'annexe A.4.). Au sein des facteurs de transcription, les récepteurs nucléaires se distinguent notamment par leur habilité à lier de petites molécules hydrophobes (des ligands). La capacité à lier un ligand est relative à un des modules du récepteur nucléaire, le domaine de liaison au ligand (LBD pour *ligand binding domain*). La figure 8.14 montre une structure de ce domaine dans le cas de RXR-α, le récepteur nucléaire au rétinoide X. On peut y voir la structure caractéristique du LBD en "sandwich" d'hélices-α. Lorsqu'il est lié au récepteur, le ligand est enfoui dans une cavité à proximité de l'hélice $H12$. Dans le cas d'un ligand agoniste (c'est-à-dire un ligand favorisant la transcription), l'hélice $H12$ est repliée en contact contre le LBD dans une position plus compacte. Cet agencement laisse apparent un sillon spécifiquement reconnu par un cofacteur dont la présence est nécessaire au recrutement de la machinerie transcriptionnelle. Certains récepteurs nucléaires ne sont actifs que sous forme de dimères, auquel cas ils s'arriment systématiquement le long de leur hélice $H10$ respective (voir figure 8.15).

Calculée sur la surface d'un récepteur nucléaire, la courbure locale permet d'identifier les trois sites d'intérêt que sont la poche de fixation au ligand, le sillon du cofacteur, et le sillon à l'interface de dimérisation (figure 8.16).

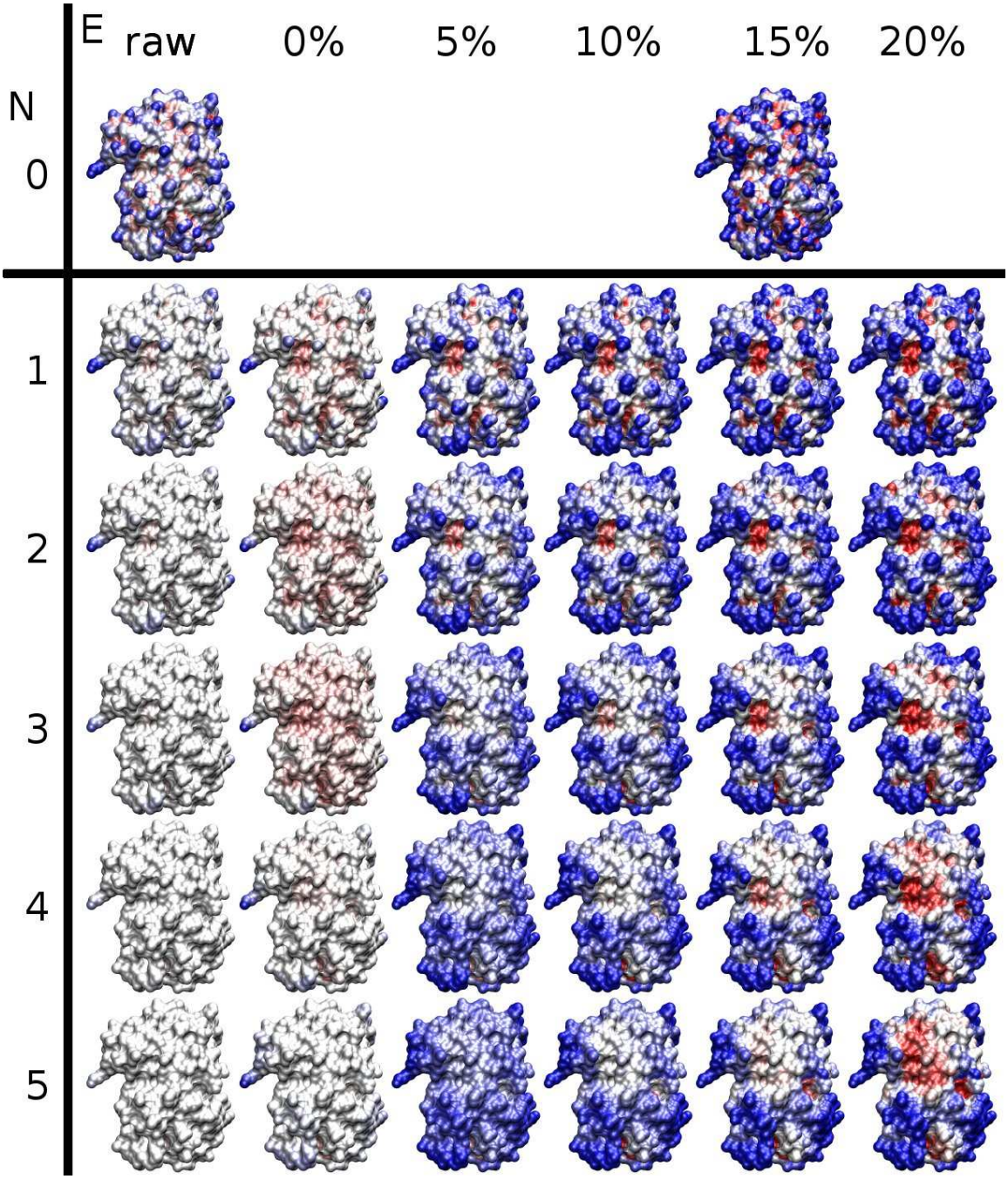


Figure 8.13: Courbures locales $lc^{w}_{N:E\%}$ à la surface du domaine de liaison au ligand du récepteur à l'acide rétinoïque RXR-α (1RDT chaîne A). En ordonnée, la taille N du voisinage de lissage. En abscisse, le seuil d'écrêtement E. La colonne *raw* désigne les valeurs brutes, sans écrêtement ni re-échantillonnage, la colonne 0% correspond à un re-échantillonnage linéaire sans écrêtement.

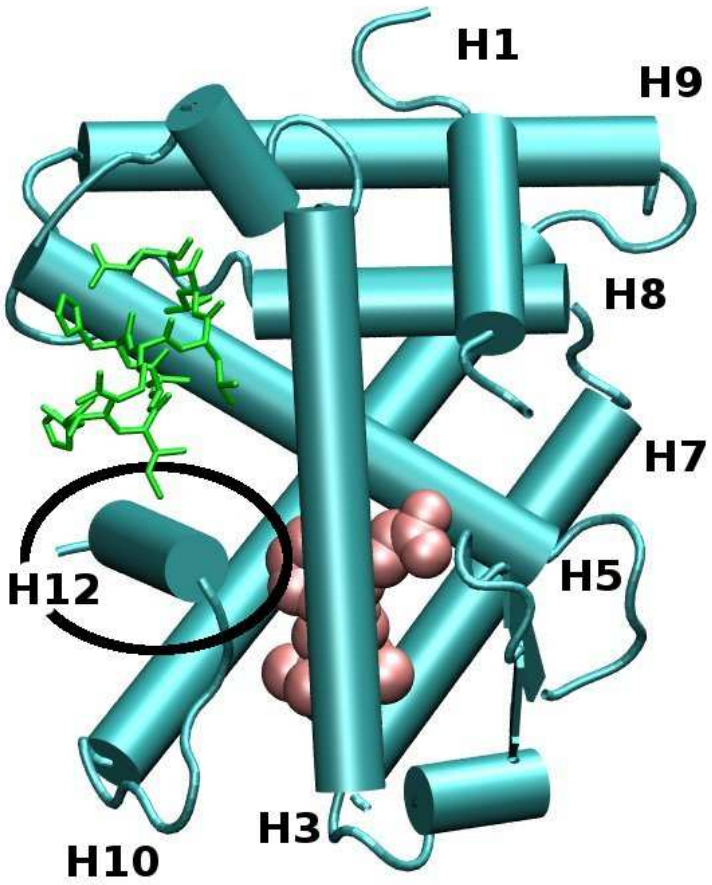


Figure 8.14: Structure du LBD d'un récepteur nucléaire. La représentation *cartoon* montre l'agencement des éléments secondaires du domaine de liaison au ligand du récepteur à l'acide rétinoïque RXRα (1RDT chaîne A). Dans cette structure, le récepteur a été cristallisé avec un cofacteur (en vert) et un ligand agoniste (en rose). L'agencement, commun à toute la famille des récepteurs nucléaires, comprend un feuillet-β et une douzaine d'hélices-α se succédant de $H1$ (à l'extrémité N-terminale) à $H12$ (à l'extrémité C-terminale). Les douze hélices forment un "sandwich", dans lequel $H1$ et $H3$ sont ici au premier plan, $H7$ à l'arrière avec l'hélice $H10$ intervenant dans la dimérisation du récepteur nucléaire. Dans cette configuration agoniste l'hélice $H12$ (entourée en noir) clos la poche de fixation du ligand et matérialise une crevasse dans laquelle le cofacteur vient se nicher.

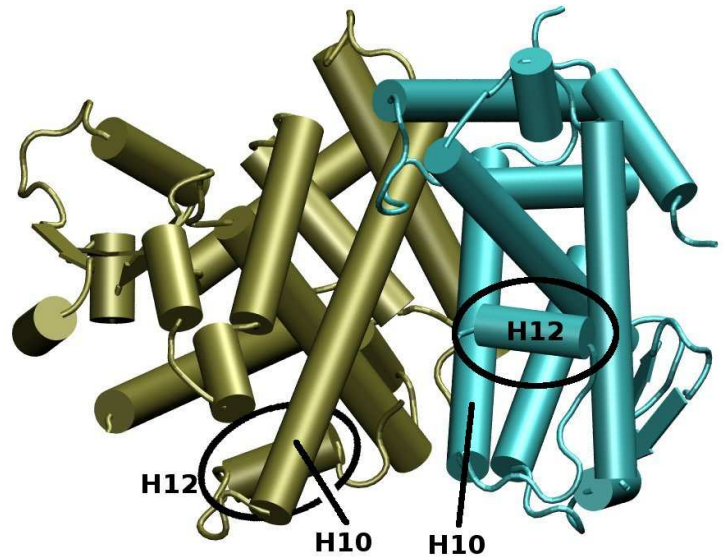


Figure 8.15: Dimérisation des récepteurs nucléaires RXRα et PPARγ (1RDT chaînes A et C). La structure a subi une rotation d'approximativement 90 degrés sur l'axe des z par rapport à la figure 8.14.

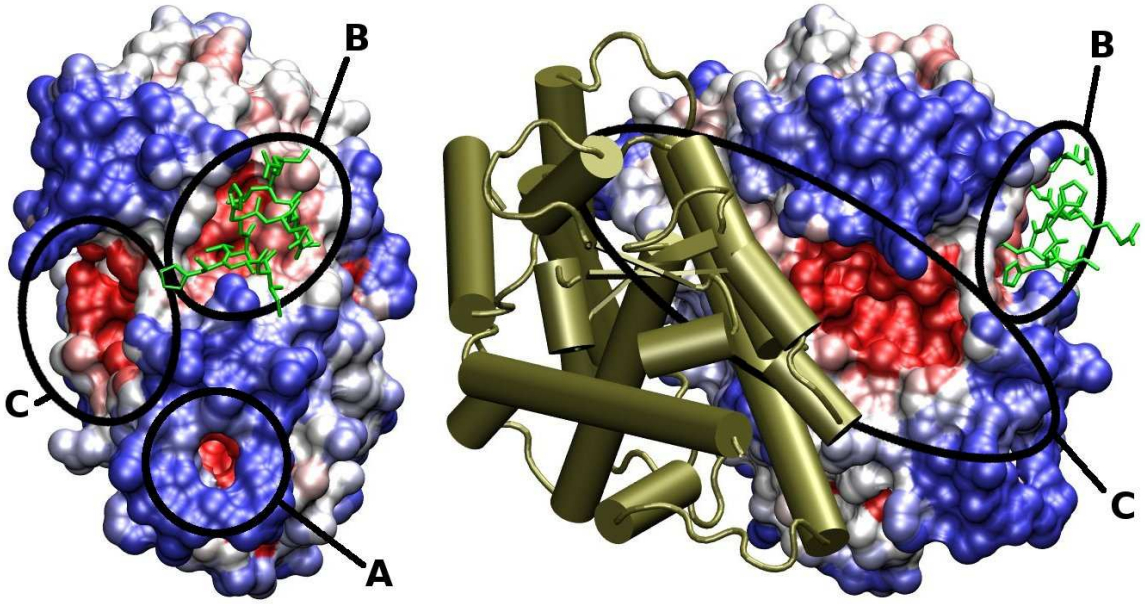


Figure 8.16: Topographie du domaine de liaison au ligand du récepteur nucléaire RXR-α (1RDT chaîne A). Les deux figures montrent la même protéine respectivement dans les configurations de la figure 8.14 et 8.15, avec un cofacteur en vert, et son partenaire PPARγ en représentation *cartoon* dans la figure de droite. La surface moléculaire de RXRα a été tracée avec une sphère-solvant de 1.3Å de rayon, et colorée avec une courbure locale $lc^{w}_{2:.43:.50}$. Le site de dimérisation (C) apparaît comme un long sillon longeant l'hélice $H10$, le sillon de recrutement du cofacteur (B) est lui aussi bien mis en valeur, de même que la cavité abritant le ligand qu'on aperçoit en rouge au delà de la bouche en (A).

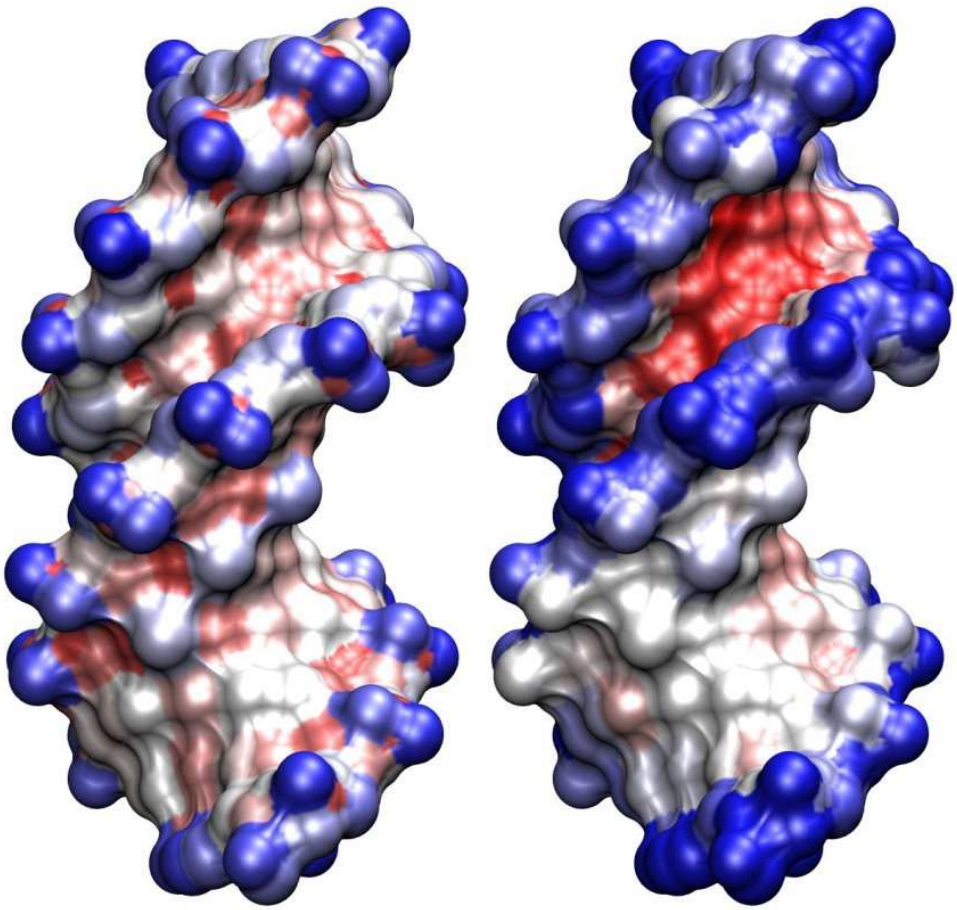


Figure 8.17: Surface moléculaire d'un duplex dodécamère d'ADN (1D65) en dégradé de couleur relatif aux valeurs d'exposition-α (lc_0) (à gauche), et de courbure locale ($lc^w_{1:.49:.52}$) (à droite). Figure de gauche : À l'exception du phosphore, les atomes les plus enfouis sont situés dans les sillons. Les atomes d'oxygène du groupement phosphate ne participant pas à la liaison entre deux nucléotides sont détectés comme les atomes les plus exposés de la molécule. Figure de droite : La forme caractéristique du double brin d'ADN est mise en évidence par la courbure locale. Le squelette ribose-phosphate de la molécule est matérialisée par un fin liseré bleu, son grand sillon par une spirale blanche, et le petit sillon par une spirale rouge.

8.4.3 Topographie d'un duplex dodécamère d'ADN

Comme illustré dans la figure 8.17, la courbure locale permet de retrouver les caractéristiques principales de la double hélice d'ADN. Les valeurs d'exposition-α (à gauche) expriment les propriétés d'exposition (très locales) des atomes de surface. Les atomes d'oxygène des groupements phosphate ce participant pas à la liaison entre deux nucléotides sont ainsi visuellement identifiés comme les plus proéminents de la molécule, en accord avec leur rôle dans la solvatation. À l'inverse, les atomes de phosphore, stériquement plus encombrés sont systématiquement identifiés comme enfouis. Comme on s'y attend, à l'exception notable du phosphore, les atomes les plus "masqués" sont toutefois plus fréquemment situés dans les sillons, et plus particulièrement dans le petit sillon. Avec un lissage de taille 1 (figure 8.17 à droite), la courbure locale met en valeur le squelette de l'ADN sous la forme d'une fine crête bleue isolant les deux sillons. Les valeurs de courbure locale sont d'ailleurs sensiblement différentes dans les deux sillons, permettant une identification visuelle.

8.4.4 Topographie de CARM1

La protéine CARM1 (pour *Co-Associated aRginine Méthyl transferase*) est une enzyme impliquée dans différents mécanismes biologiques ; elle agit dans la modulation de la transcription au travers de la méthylation de résidus arginines spécifiques des histones H3. Des indices expérimentaux tendent à prouver l'activité de cette protéine sous forme d'homodimère, où les domaines catalytiques de chaque protomère se fixent symétriquement l'un à l'autre par l'intermédiaire d'un bras composé de quatre hélices (deux hélices-α et deux hélices-$3-10$) [Troffer-Charlier 07]. Chaque protomère possède son propre site catalytique, orienté vers son partenaire. L'assemblage des deux domaines catalytiques prend la forme approximative d'un *doughnut* où les deux sites catalytiques sont accessibles dans le tunnel central. La figure 8.18 montre un homodimére de domaines catalytiques de CARM1. La symétrie axiale du dimère et le bras d'amarrage joignant chaque protomère sont mis en évidence dans la figure de gauche, la figure de droite montre le tunnel résultant de l'assemblage. La structure (3B3F) présentée dans ces figures montre le do-

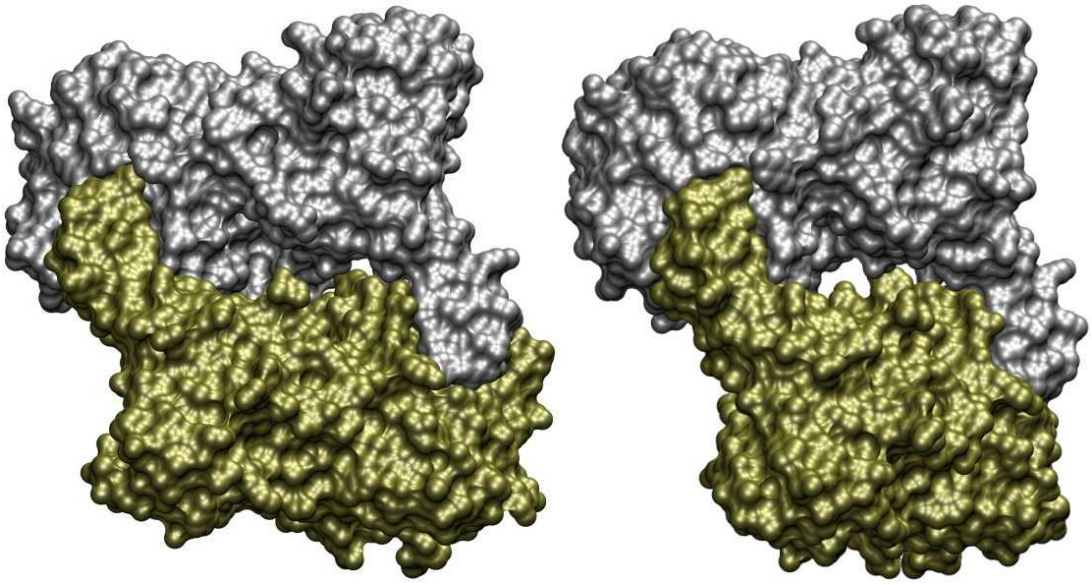


Figure 8.18: Homodimère du domaine catalytique de la protéine CARM1 (3B3F). Les surfaces moléculaires des chaînes A et B ont été représentées respectivement en argenté et doré. La figure de gauche permet d'observer la symétrie axiale du complexe, ainsi que le rôle particulier d'un bras dans l'assemblage des deux protéines. La figure de droite montre la même structure après une légère rotation. Elle met en évidence la présence d'un tunnel créé par l'assemblage des deux sites catalytiques.

maine catalytique de CARM1 co-cristallisé en complexe avec le substrat SAH fournissant le groupement méthyle. La molécule SAH est entièrement enfouie dans une cavité intérieure à la protéine, et invisible dans ces images. L'échange de méthyle a lieu au niveau du site catalytique, mis en valeur par la courbure locale dans la figure 8.19 ainsi que le bras d'amarrage et sa zone d'accrochage.

La figure 8.20 offre une autre visualisation du même dimère. Les valeurs de courbure locale sont calculées sur un des protomères (à gauche) ou sur le dimère entier (à droite) et affichées uniquement sur la surface moléculaire de la chaîne A. Le blanchiment de l'extrémité du bras de dimérisation dans la figure de droite indique que dans le complexe, cette surface résultante de l'arrimage est relativement "plate". Une bonne partie de la surface de la chaîne A est masquée par la chaîne B, ces atomes se voient automatiquement attribuer des valeurs d'exposition-α et de courbure locale nulle. De fait, dans la figure de droite, l'interface de dimérisation apparaît

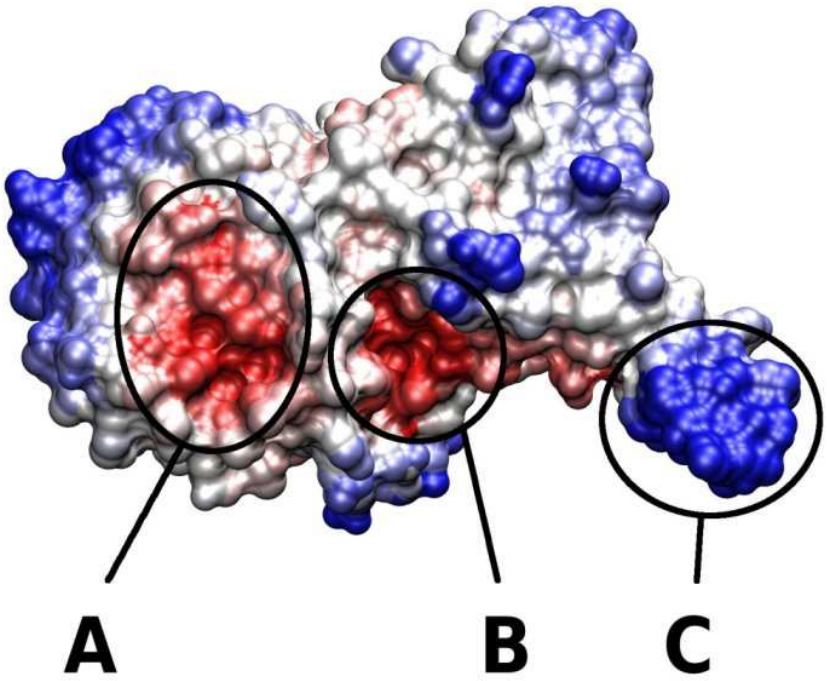


Figure 8.19: Courbure locale à la surface du domaine catalytique de l'enzyme CARM1 (3B3F chaîne A). La surface moléculaire est agrémentée d'un dégradé de couleurs relatif à une courbure locale pondérée avec un lissage de taille 4 et des seuils à 0,44 et 0,52 ($lc^{w}_{4:.44:.52}$). La courbure locale permet de mettre en évidence le site catalytique (B) comme une des zones les plus invaginées de la surface. La zone d'accroche de l'extrémité du bras de dimérisation (A) est, elle aussi, particulièrement bien détectée comme zone invaginée. L'extrémité du bras de dimérisation (C) est, lui, détecté comme une protrusion.

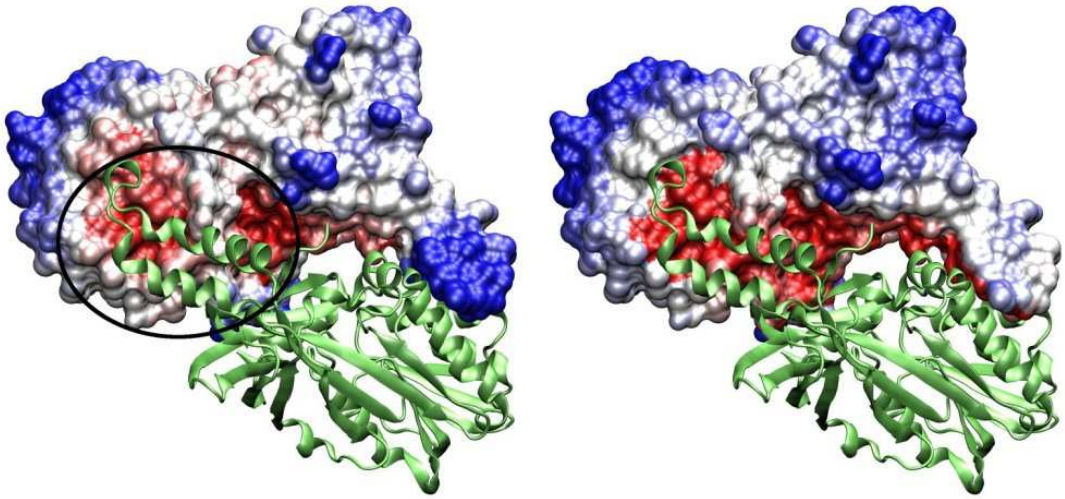


Figure 8.20: Homodimère du domaine catalytique de l'enzyme CARM1, même structure, même chaîne même code couleur que dans la figure précédente. La surface moléculaire de la chaîne A est colorée suivant ses valeurs de courbure locale, cependant que la chaîne B en représentation cartoon permet de visualiser l'arrangement "tête-bêche" du complexe. En particulier, on retrouve l'extrémité du bras de dimérisation de la chaîne B arrimée dans le réceptacle idoine de la chaîne A (cerclé en noir dans la figure de gauche). La figure de gauche montre les valeurs de courbure locale calculées sur le protomère de la chaîne A. Dans la figure de droite les valeurs de courbure locale ont été calculées sur le dimère complet (chaînes A et B) et affichées uniquement sur la chaîne A.

nettement en rouge et se prolonge en dégradé de rouge très prononcé au niveau du tunnel entre les deux unités.

8.4.5 Topographie de l'Aspartyle-ARNt synthétase

Les Aminoacyle-ARNt synthétases sont des agents clés dans la traduction de l'ADN. Leur rôle consiste à reconnaître et associer un acide aminé avec l'ARNt portant l'anticodon correspondant. La fonction de ces agents moléculaires requiert de nombreux mécanismes de reconnaissance moléculaire. En particulier, une forte sélectivité est à l'œuvre en deux sites distants à la surface de cette molécule : le "sillon" où l'acide aminé vient se nicher, et la "pince de lecture" de l'anticodon de l'ARNt. Ces deux sites ont été mis en évidence dans la figure 8.21 dans le cas particulier de l'aspartyle-ARNt synthétase. Cette figure montre en outre qu'un indice de

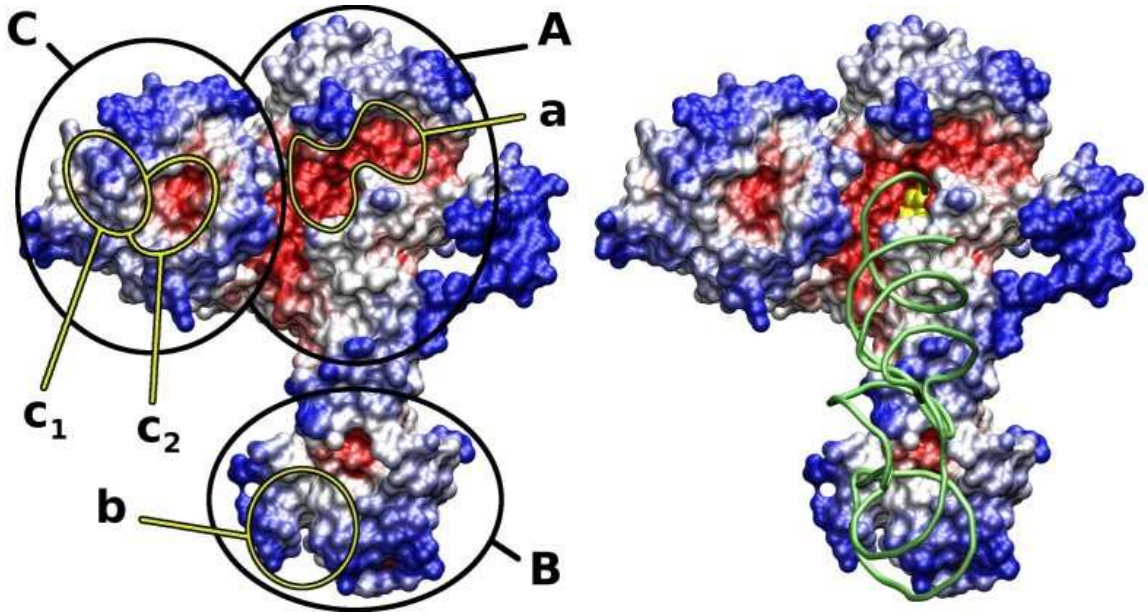


Figure 8.21: Structure de l'aspartyle-ARNt synthétase exprimée dans E. coli (1IL2 chaîne A), en complexe avec son ARNt et son ligand acide aminé. La surface moléculaire de la protéine est représentée en dégradé de couleur suivant les valeurs de courbure locale $lc^{w}_{5:0.43:0.51}$ Les deux figures sont identiques, en dehors des annotations (à gauche), et de l'adjonction des molécules complexées (à droite) : l'acide aminé aspartyle est représenté en jaune dans le modèle Van der Waals, et l'ARNt en vert dans une représentation *cartoon*. L'aspartyle-ARNt synthétase est composée de trois domaines : le domaine catalytique (A) contient le sillon majeur (a) dans lequel l'acide aminé vient se nicher ; le domaine anticodon (B) est composé d'un tonneau β et de deux hélices dont l'une constitue une pince (b) impliquée dans la reconnaissance de l'anticodon spécifique sur l'ARNt ; le rôle de l'extra-domaine (C) n'est pas encore connu, bien que la présence de deux séquences consécutives de quatre résidus conservés (c_1 et c_2) constitue un indice pour une fonction relative à une interaction.

courbure locale permet de retrouver clairement les deux sites d'importance biologique précédemment évoqués. L'Aspartyle-ARNt synthétase est flanquée d'un domaine supplémentaire dont la fonction n'a, à notre connaissance, pas encore été mise en évidence. Deux séquences de quatre résidus consécutifs dans cet extra-domaine sont particulièrement conservées entre protéines orthologues [Moulinier 97] ; Dans la figure 8.21, la courbure locale révèle qu'une de ces deux séquences compose le fond d'une petite crevasse bien marquée à la surface de ce domaine, cependant que la seconde séquence constitue une petite épine dans le prolongement de cette crevasse. La présence

intrigante de ces motifs conservés coïncidants à des caractères topographiques marqués constitue un indice pour d'une interaction potentielle avec une autre molécule.

8.5 Conclusions et perspectives

Nous avons défini deux indices topographiques permettant la caractérisation de la surface moléculaire en terme de "creux" et de "bosses". Intuitivement, l'exposition-α d'un atome définit une mesure de la "proportion de vide" entourant très localement un atome. Malgré la nature "chaotique"[5] de la surface duale, un simple lissage de ces valeurs permet de révéler une tendance différente de l'exposition-α dans les "creux" et dans les "bosses" dont nous avons observé la pertinence sur des exemples moléculaires. Bien qu'elle ne soit basée ni sur une mesure de la courbure au sens mathématique strict, ni sur une approximation locale par une sphère ou une forme quadratique, la courbure locale ainsi définie est à rapprocher des mesures d'incurvation de la surface telles que définies au chapitre 3. Cet indice n'est en effet défini et calculé qu'à partir d'une donnée surfacique, et produit des valeurs exprimant localement la forme de cette surface, son incurvation.

De par sa définition dépendant uniquement de la surface, notre indice s'abstrait de problèmes éventuels liés à la présence de cavités sous la surface ou de modulations de la densité atomique. Il est en outre indépendant de la position de la molécule dans l'espace et est, plus généralement, invariant à toute transformation solide[6]. Enfin, le cadre de la théorie des formes-α fournit une structure propice à l'édification d'algorithmes simples.

Comparée aux autres approches de calcul de l'incurvaison présentées au chapitre 3, notre approche ne nécessite pas la construction explicite d'une surface moléculaire (Surface Accessible ou Surface Moléculaire) ; elle se contente de l'approximation fournie par la surface duale, comprenant moins de sommets tout en retenant la forme de la molécule. Le faible nombre de sommets à considérer, la simplicité des algorithmes et le nombre restreint d'opérations autorisent des temps de calcul courts comparés aux autres approches présentées dans la section 3.2.3 de l'état de l'art dédiée aux mesure d'incurvation de surface. En outre, comme nous l'avons illustré sur des exemples biologiques, ces avantages ne sont pas obtenus au détriment de la pertinence.

La courbure locale présente toutefois quelques limitations qui méritent d'être discutées plus avant :

- l'attribution d'une valeur aux atomes de multiplicité supérieure à 1 ;
- la dépendance aux rayons de Van der Waals utilisés pour modéliser les atomes de la molécule ;
- le choix des variables pour le calcul de la courbure locale ;
- le cantonnement de l'étude à une résolution fixe.

L'attribution d'une valeur de courbure locale aux atomes présentant plusieurs morceaux surfaciques est nécessairement arbitraire ; ceux-ci participant à une partie distincte de la surface et étant parfois très éloignés les uns des autres le long de la surface. Une valeur de courbure locale étant calculée pour chacun de ces morceaux surfaciques, la solution la plus immédiate — et celle que nous avons adoptée — consiste alors à prendre la moyenne de ces valeurs.

Un autre inconvénient concernant les indices topographiques que nous avons calculés est leur sensibilité au choix des rayons de Van der Waals utilisés pour la modélisation de la molécule

[5] Le bord de la forme duale présente un aspect accidenté, constellé de légers pics et creux. De fait, les valeurs d'exposition-α sont très disparates au voisinage d'un même point.

[6] Composition de rotations, de translations et d'homothéties.

ainsi qu'au rayon de la sphère solvant utilisée pour la définition de la Surface Accessible. Ce comportement ne peut être évité car nos indices décrivent la forme de la surface, et que c'est la définition même de cette surface qui est tributaire de ces paramètres ; la plupart du temps, de faibles modifications dans ces paramètres n'auront que peu de conséquences sur la forme de la surface, et donc sur nos indices topographiques. Il faut toutefois être conscient que certaines variations minimes peuvent avoir des conséquences topologiques drastiques, comme l'ouverture d'une bouche (voir figure 8.22 *A*).

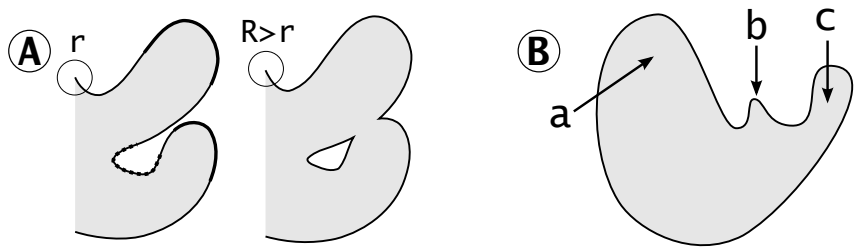


Figure 8.22: Dépendance de la courbure locale aux paramètres d'entrée, et limitation à un niveau de résolution : une illustration en deux dimensions.
A : Les deux figures montrent un même détail d'une molécule. Dans la figure de droite, la sphère solvant, un peu plus grosse, a clos la bouche étroite de la figure de gauche et généré une cavité. Les atomes qui participaient à l'une ou l'autre des deux proéminences (en gras dans la figure de gauche) participent maintenant à une zone relativement "plate".
B : Les bosses annotées *a*, *b* et *c* sont des propriétés topographiques visibles à des niveaux de résolutions différents. Elles peuvent chacune être observées avec la courbure locale pour une taille de voisinage de lissage adéquate. Elles ne seront cependant généralement pas simultanément observables.

Une difficulté — commune à d'autres approches que la notre — réside dans le choix de "bonnes" variables (taille du voisinage de lissage, choix d'une méthode de lissage, choix éventuel de valeurs seuil) pour paramétrer la courbure locale. Nous avons montré sur des exemples moléculaires la possibilité de trouver "à la main" des variables faisant apparaître les caractéristiques topographiques, mais l'établissement de critères automatiques reste à explorer. En particulier, l'histogramme des valeurs de courbure locale observées varie beaucoup d'une molécule à l'autre et devrait entrer en jeu dans l'établissement de ces valeurs.

Enfin, la courbure locale — tout comme bon nombre d'autres indices topographiques — exprime essentiellement une propriété observée à une échelle donnée, paramétrée par la taille du voisinage de lissage (voir figure 8.22 *B*). On pourrait néanmoins souhaiter étudier simultanément une information topographique à des niveaux de résolution différents. L'approche proposée par Lee *et al.* [Lee 05] pour décrire la "saillance" d'un maillage à partir de mesures de courbures pourrait par exemple être appliquée ici à la courbure locale. Une autre approche assez similaire consisterait à étudier les variations de l'histogramme des valeurs de courbure locale dans les voisinages et les corolles d'un point de la surface duale.

Depuis son introduction dans le cadre de la bioinformatique structurale [Edelsbrunner 95a], la forme duale a été appliquée à diverses problématiques pour son habilité à modéliser la forme des macromolécules. En premier lieu, elles ont été employées pour le calcul exact des valeurs volumétriques (volume et aire de la surface) des molécules [Edelsbrunner 95a], ainsi que pour la détection et les calculs volumétriques des cavités et des poches dans les macromolécules [Edelsbrunner 95a, Edelsbrunner 98, Peters 96]. Leur dualité avec le modèle SA a aussi été exploitée pour accélérer la construction d'une surface moléculaire discrète [Coleman 05], ou pour définir et construire une nouvelle surface lisse [Edelsbrunner 99]. Elles ont aussi été employées

comme une représentation simplifiée de la forme globale d'une macromolécule, partageant ses caractéristiques topologiques. De Alarcon *et al.* ont par exemple utilisé cette représentation pour extraire rapidement des caractéristiques globales (volume et nombre de "tunnels" et de "cavités") de cartes de densité de très basse résolution [De-Alarcon 02]. Kasson *et al.* ont utilisé ce modèle pour représenter les bicouches lipidiques et étudier leur mode de fusion [Kasson 07]. À notre connaissance, notre approche constitue la première utilisation de la forme duale pour la similarité de la forme de sa surface avec celle de la surface moléculaire.

Précisons enfin que les algorithmes présentés dans cette partie ont été développés et mis à disposition au travers du logiciel `Lc` dont on trouvera les principales caractéristiques en annexe D.

Chapitre 9

Descripteurs de surface pour la détection de sites interagissants, une application des formes-α

Nous avons appliqué les modèles de la théorie des formes-α pour revisiter trois descripteurs de la surface moléculaire communément employés dans les études en bioinformatique structurale : les notions de *résidus de surface*, de *parcelle de surface* ainsi que celle d'*accessibilité*. Ces trois notions ont été validées dans le contexte de la caractérisation des zones interagissantes à la surface des protéines, où elles ont montré une utilité et une pertinence au moins égales à celles observées dans les approches existantes. Ce travail a été réalisé en collaboration avec l'équipe d'O.Poch au département de génomique et de biologie structurale à l'IGBMC.

9.1 Cadre applicatif : la prédiction de zones interagissantes à la surface d'un monomère

La fonction d'une protéine est en très grande part liée aux interactions transitoires qu'elle réalise avec d'autres macromolécules, en particulier avec d'autres protéines. Une interaction est qualifiée de *transitoire* (*transient*)[1] lorsque les intervenants sont susceptibles de s'associer et de se dissocier aisément. Le même qualificatif s'applique plus généralement lorsque les intervenants peuvent être observés dans la cellule à l'état de monomère comme à l'état d'oligomère, ou encore lorsqu'au moins l'un des protagonistes est connu pour s'associer à une autre molécule sur le même site d'interaction. Ces dernières définitions sont en effet prises comme des preuves de la capacité des protagonistes à se dissocier. Cette habilité à s'associer et à se dissocier est un des moyens par lesquels une protéine réalise sa fonction. Dans notre étude, nous avons désigné par *résidus interagissants*, les résidus d'une protéine qu'on sait impliqués dans une interaction transitoire. Nous parlons aussi de *zone interagissante* pour un ensemble connecté de résidus interagissants.

La prédiction des zones interagissantes à la surface d'un monomère donné est donc une des pistes pour la compréhension de la fonction d'une macromolécule. Les moyens d'une telle étude reposent généralement sur *(i)* la caractérisation de zones interagissantes connues, *(ii)* l'apprentissage de ses éléments caractéristiques, et leur détection dans une structure où les zones interagissantes sont inconnues.

[1] Par opposition, on parle d'*interaction permanente* ou de *complexes permanents* (*obligate*) pour désigner des molécules connues pour former un complexe suffisamment stable pour ne jamais se dissocier.

9.2 Jeu de données et résidus interagissants

La constitution d'un jeu de données de structures de protéines impliquées dans des interactions transitoires est une première étape nécessaire à la réalisation d'une analyse des zones interagissantes. Au cours d'un appariement, les protéines sont sujettes à des modifications structurales plus ou moins drastiques ; afin de valider la possibilité d'utiliser des caractéristiques enregistrées à la surface d'intervenants impliqués dans un dimère pour prédire des caractéristiques à partir de la surface d'un monomère, il est nécessaire de constituer un jeu de données de protéines pour lesquelles on dispose à la fois d'une structure isolée et d'une structure où chaque protéine est impliquée dans une interaction. Nous avons appelé *forme liée* (ou *protomère* d'après Jones [Jones 97b]) une chaîne protéique extraite d'une structure où elle est présente à l'état de dimère avec une autre chaîne, et *forme non-liée* une chaîne protéique issue d'une structure où elle est présente à l'état de monomère. À l'issue du processus de sélection, 85 protéines ont été retenues où nous disposions à la fois d'une structure liée et d'une structure non-liée. La qualité des structures, ainsi que la redondance du jeu de données, ont été prises en compte dans le processus de sélection.

À partir de ce jeu de données, nous avons défini les résidus interagissants comme les résidus dont l'aire varie de plus d'un Ångström entre le monomère et le dimère.

9.3 Présentation de nos descripteurs

Dans cette étude nous nous sommes intéressés à la réalisation de deux outils descriptifs primordiaux dans le cadre d'une analyse des sites interagissants : la définition des résidus composant la surface, et une nouvelle définition de parcelle de surface. Nous avons en outre validé l'intérêt des indices topographiques que nous avons définis au chapitre 8 pour la discrimination des zones interagissantes à la surface d'une protéine.

9.3.1 Résidus de surface

Définir les résidus participant à la surface d'une protéine n'est pas une chose évidente. Comme nous l'avons vu à la section 2.2 (page 22, la surface d'une molécule peut être comprise comme l'ensemble de ses résidus en contact avec le solvant, ou susceptibles d'interagir avec d'autres (macro)molécules ; on l'oppose généralement au cœur de la macromolécule, auquel on associe un rôle structurant.

Dans notre approche, nous avons définis les résidus de surface à partir de la surface duale en faisant usage de deux paramètres : N_r, le nombre d'atomes du résidu r participant à la surface[2], , et sa valence v_r, un nouvel indice comptabilisant le nombre d'arêtes sur la surface duale joignant le résidu r à un autre résidu sur la surface de la molécule. Nous avons posé l'hypothèse raisonnable que les résidus interagissants devaient apparaître à la surface de la protéine, et avons ainsi utilisé la proportion de tels résidus définis comme résidus de surface comme une mesure de pertinence pour évaluer les définitions de la surface. Cette sélection des résidus interagissants ne devant pas se faire au détriment du cœur de la protéine, nous avons veillé à ce qu'une proportion suffisante des résidus composant la molécule ne soit pas attribuée à la surface. Les résultats obtenus en imposant qu'un résidu de surface possède au moins cinq atomes de surface ou une valence supérieure à dix ont permis de retrouver 98% des résidus interagissants tout en conservant un cœur composé de 20% de l'ensemble des résidus de la molécule.

[2]Rappelons que nous considérons comme appartenant à la surface tout atome dont la contribution d'aire dans le modèle Surface Accessible est strictement positive. Ces atomes sont associés à un sommet de la surface duale.

9.3.2 Parcelle de surface

Une *parcelle de surface* (ou *surface patch*) autour d'un résidu de surface consiste en la définition d'un voisinage de résidus à la surface de la molécule. Comme il a été évoqué dans le chapitre dédié à l'état de l'art page 119, de telles parcelles sont couramment utilisées pour caractériser localement la surface d'une molécule autour d'un résidu.

Dans notre étude, nous avons exploité l'équivalence entre la surface duale et la surface Surface Accessible (ce que nous avons appelé la quasi-dualité au chapitre 7.4.3 page 85) pour définir une notion de parcelle qui garantisse une sélection de résidus réellement voisins le long de la surface de la molécule et répartie le plus circulairement possible autour d'un résidu central.

Notre approche a été comparée à celle de Jones *et al.* [Jones 97b], une des définition les plus couramment utilisées dans les études en bioinformatique structurale. La figure 9.1 présente les

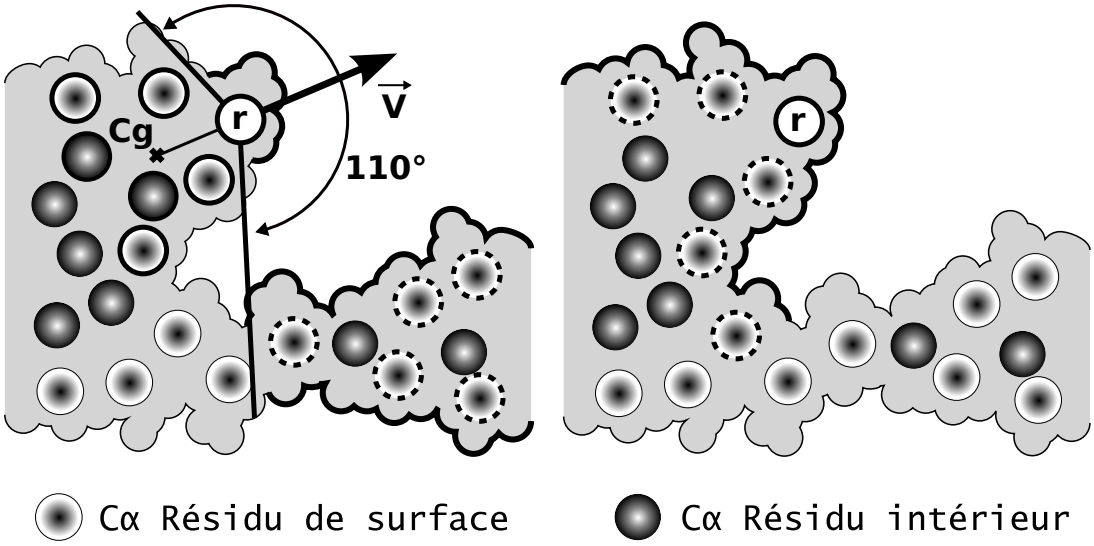


Figure 9.1: Deux approches pour la définition des parcelles de surface, un exemple en deux dimensions. Pour visualisation, les résidus sont assimilés à leur C_α dans les deux figures bien que cette modélisation des résidus ne soit concrètement utilisée que dans la méthode de Jones *et* Thornton [Jones 97b]. La parcelle centrée sur le résidu de surface r est matérialisée par un contour plus gras au niveau de la surface moléculaire, et ses résidus-C_α sont entourés d'une ligne interrompue.
Dans la méthode de Jones (à gauche), les résidus sont assimilés à leur C_α. Le centre de masse C_g de tous les résidus les plus proches de r (entourés d'un trait gras) est utilisé pour définir, avec r, un vecteur $\vec{v}$ indiquant la direction de l'espace du solvant. La parcelle de surface autour de r est constituée des $N = 20$ résidus de surface les plus proches de r et situés dans un cône de visibilité défini par $\vec{v}$.
Notre méthode (à droite) sélectionne une parcelle centrée sur r par propagation d'un voisinage le long de la surface moléculaire.

deux approches en mettant l'accent sur les défauts de la solution proposée par Jones *et al.* et la manière dont nous les avons circonvenus. Le cas présenté dans cette figure est volontairement caricatural, l'espacement des C_α ne permettant que rarement — pour ne pas dire "jamais" — des écarts aussi drastiques que ceux présentés ici. Cette illustration permet néanmoins de mettre en évidence des défauts de sélection pouvant effectivement apparaître dans la méthode de Jones *et al.* et que la nôtre interdit : la possibilité de sélectionner des résidus appartenant à des "côtés" différents de la surface, le mauvais centrage de la parcelle autour de son résidu central r, la sélection de résidus de surface potentiellement éloignés de r, et la non-connexité de la parcelle sélectionnée.

Nous avons en outre montré que notre définition de parcelle permettait un meilleur rapport

signal-bruit dans la récupération des résidus interagissants, et par là même validé l'intérêt de notre méthode dans le contexte de la prédiction de zones interagissantes à la surface des protéines.

9.3.3 Topographie

Nous avons utilisé une implémentation de la courbure locale pondérée dans un voisinage de taille 2 pour définir la topographie de surface des macromolécules, et distinguer les *résidus de zones creuses* (*clefts*) des *résidus de zones exposées* (*knobs*).

Comme on pouvait s'y attendre, une étude sur notre jeu de données a montré que les résidus interagissants sont plus présents dans les zones exposées que dans les creux. Des distinctions, en terme de composition en acides aminés et de propriétés physicochimiques, ont été observées entre les zones anfractuées et les zones exposées ainsi qu'entre résidus interagissants et non-interagissants. Une étude plus fine a révélé que les mêmes distinctions entre résidus interagissants et résidus non-interagissants restent valables lorsque l'on restreint l'observation à l'intérieur de chacune des deux classes de résidus : anfractuées ou exposées ; la distinction étant cependant plus marquée dans les zones exposées.

9.4 Conclusions et perspectives

Dans cette étude, nous avons utilisé avec succès les modèles issus de la théorie des formes-α pour définir trois descripteurs de la surface moléculaire. La pertinence de ces descripteurs a été établie dans le cadre de la caractérisation des zones interagissantes où nous avons obtenu des résultats au moins similaires à d'autres approches existantes, et parfois meilleurs. Ces observations motivent l'utilisation de nos outils à des fins de prédiction de zones interagissantes à la surface de monomères.

9.5 *Defining and characterizing protein surface using alpha shapes* (article)

Les pages suivantes sont composées de "*l'article Defining and characterizing protein surface using alpha shapes*", de Benjamin Schwarz*, Laurent-Philippe Albou*, Olivier Poch, Jean-Marie Wurtz, et Dino Moras, et publié en octobre 2008 dans *Proteins : structure, function and bioinformatics*

* ces auteurs ont contribué à part égale

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.

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 s*urface 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).

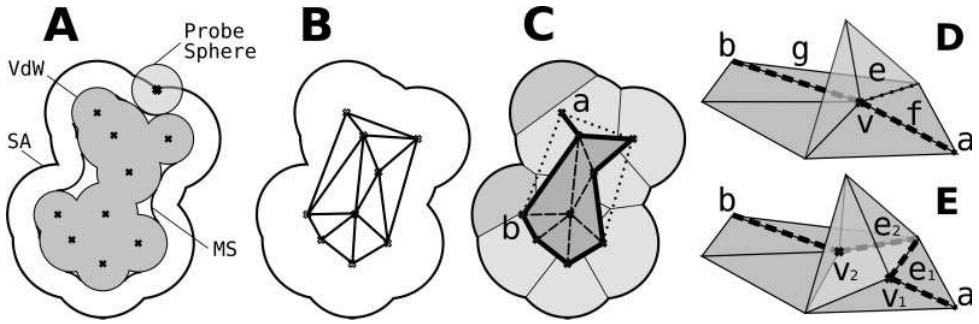


Figure 1

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-

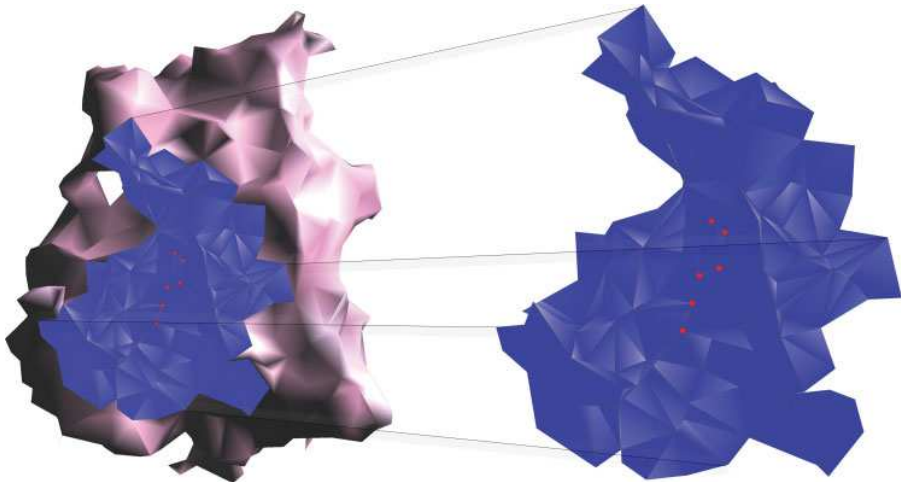


Figure 2

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

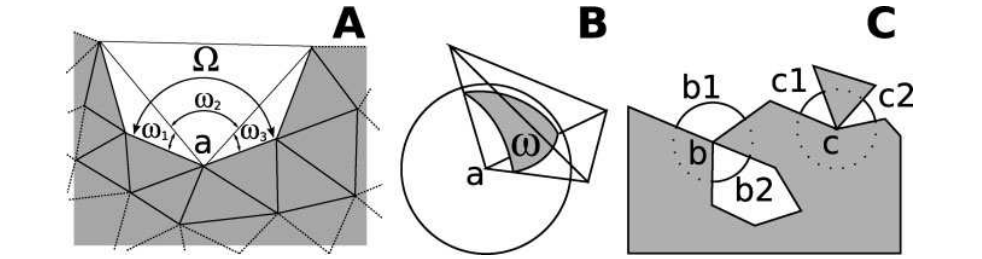


Figure 3

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:

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

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 Ω(*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

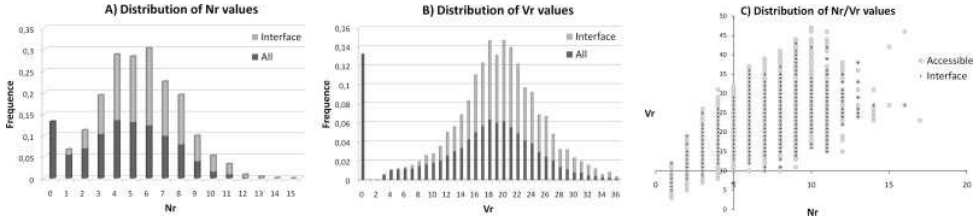


Figure 4

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.

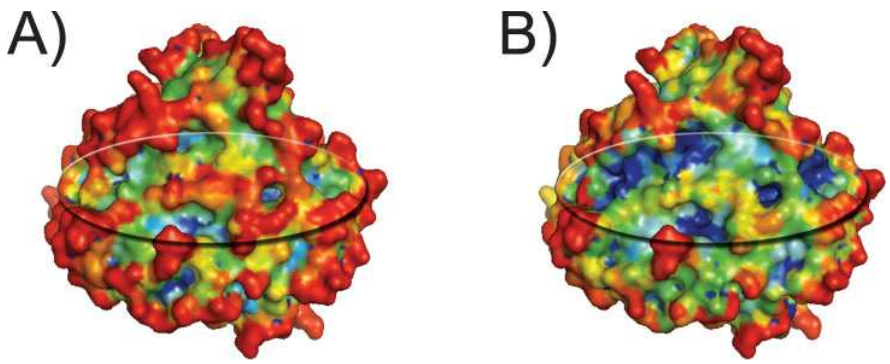


Figure 5

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.

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.

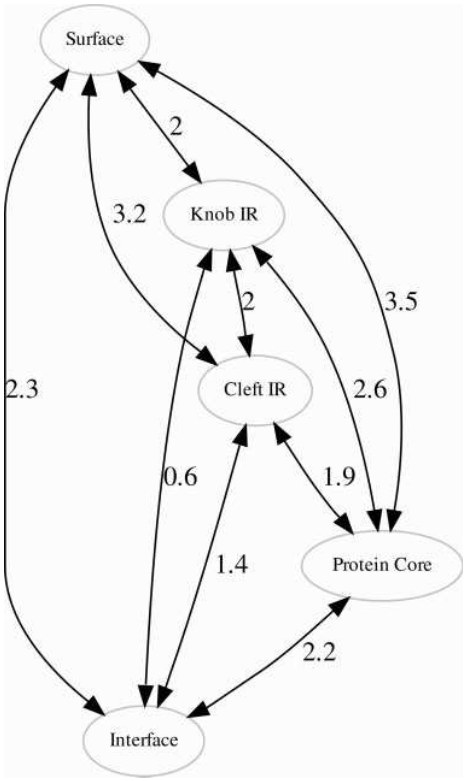


Figure 6

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.

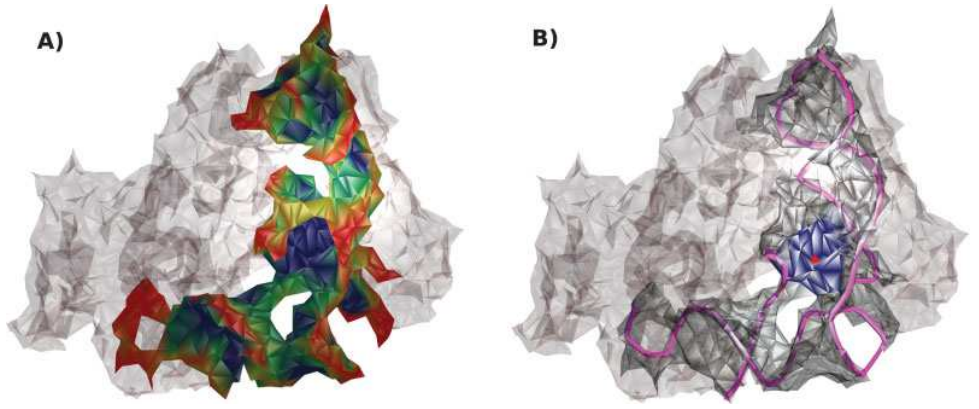


Figure 7

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.

Chapitre 10

Détection et caractérisation des poches dans les macromolécules biologiques

Dans ce chapitre nous présenterons les travaux que nous avons réalisés dans le domaine de la détection et de la caractérisation des poches dans les macromolécules. En dépit des efforts croissants concernant cette problématique, de nombreuses insatisfactions demeurent. D'un point de vue strictement pratique, de nombreux algorithmes proposés dans la littérature ne sont pas librement accessibles, ou bien le sont au travers de services réseau ou de programmes exécutables, rendant impossible la modification des propriétés calculées ou l'évolution des algorithmes proposés ; des modifications s'avèrent pourtant souvent nécessaires pour adapter les réponses apportées à un problème particulier. Ces raisons constituaient une motivation suffisante pour développer `Pck`, un logiciel de détection et de caractérisation interne au laboratoire[1].

Dans `Pck` nous proposons quatre algorithmes pour détecter trois types distincts de poches[2], avec des apports originaux dans deux d'entre eux. Ces algorithmes seront présentés dans la première section du chapitre. Dans la seconde section nous présenterons les propriétés proposées dans `Pck` pour caractériser les poches détectées, et dans la troisième nous observerons l'habilité de chaque algorithme à détecter un type particulier de poche. Le chapitre s'achèvera sur une discussion concernant nos apports, et nous présenterons des pistes pour des développements ultérieurs autour de `Pck`.

10.1 Présentation des algorithmes de détection des poches implémentés dans `Pck`

Nous avons implémenté et inclus dans `Pck` quatre algorithmes distincts pour la détection des poches ; pour chacun d'entre eux nous avons proposé la définition d'un volume de l'espace alloué à la poche. Les quatre approches reposent sur les mêmes bases issues de la théorie des formes-α rappelées dans une première sous-section. La description complète des algorithmes est donnée dans les sous-sections suivantes. Les tables 10.1 et 10.2 offrent une présentation synthétique de ces algorithmes et de leurs paramètres : *tight* constitue une implémentation de l'algorithme d'H.

[1] Nos logiciels sont fournis sous forme d'exécutables, et les sources sont accessibles sur demande, comme expliqué plus en détail dans l'annexe D.

[2] Comme il a été remarqué dans le chapitre bibliographique (page 37) traitant des poches, plusieurs définitions alternatives peuvent coexister pour le terme de "poche" en fonction par exemple des motivations ("poches de fixation d'un ligand", "canaux ioniques", "sites catalytiques") ou de caractères tenant à la forme ; on parle alors de "poches géométriques". Nos travaux sont exclusivement axés sur des définitions et des approches géométriques.

algorithme	principe
tight	Détection des poches refermées et des cavités (présence d'une constriction) ; implémentation de *CASTp* [Binkowski 03b].
wide	Détection des poches ouvertes (*inaccessibilité pour une sonde de taille* α_1).
groove-c	Détection des zones incurvées (*calcul de courbure locale*) orientée définition volumique (*agglomérat de cellules anfractuées*). L'abbréviation *groove-c* signifie ***groove(s)*** *detection based on Delaunay* ***cells***.
groove-fr	Détection des zones incurvées (*calcul de courbure locale*) orientée définition surfacique (*constitution de "fonds de poche" suivi d'une "volumisation"*). L'abbréviation *groove-fr* signifie ***groove(s)*** *detection based on Delaunay* ***facets*** *and* ***restricted*** *to a common bucket*.

Tableau 10.1: Récapitulatif des stratégies de détection de poches implémentés dans le logiciel `Pck`

algorithme	paramètres
tight	aucun
wide	α_1 **:(*inaccessibilité*)**, contrôle un "niveau de la mer" ou un niveau de détail ; rayon de la plus petite sphère-solvant n'accèdant pas à la poche. ϵ **:(*érosion*)**, contrôle à la fois le nombre de poches détectées et leur taille.
groove-c et *groove-fr*	s **:(*lissage*)** contrôle un niveau de détail, une résolution. Utilisé pour le calcul de courbure locale. t **:(*seuil*)** contrôle un "niveau de la mer", sommets considérés comme potentiellement impliqués dans une poche sur la surface duale.

Tableau 10.2: Présentation des paramètres des algorithmes de détection de poches implémentés dans le logiciel `Pck`

Edelsbrunner présent dans le logiciel *CASTp* [Binkowski 03b], *wide* détecte les zones accessibles[3] à une sphère-solvant de faible taille et inaccessibles à une sphère plus large, les deux algorithmes de type *groove* utilisent la courbure locale pour détecter les zones les plus anfractuées de la molécule (*groove-c* cherche des cellules dont les quatres sommets sont situés dans une telle partie de la surface, et *groove-fr* localise ces zones directement à la surface avant d'en donner une description volumique). Nos quatre algorithmes décrivent les poches détectées à la fois comme une liste d'atomes et comme un volume tri-dimensionnel.

10.1.1 Rappels sur la représentation polyédrique offerte par le complexe dual

Nous avons précédemment mentionné la triangulation de Delaunay et le complexe dual d'une molécule, et avons présenté leur relation avec la topologie et la forme de l'union de boules représentant la même molécule ; une introduction de ces modèles peut être trouvée aux pages 12 et 13 de la section 1.2, et des définitions plus formelles ont été apportées au chapitre 6 (pages 69 et 61). La figure 10.1 reprend les principales caractéristiques de ces modèles dans un exemple en deux dimensions, et précise leur emploi dans la détection des poches. On y voit une molécule dans sa représentation Surface Accessible superposée à sa triangulation de Delaunay et à son complexe dual ; ce dernier offre une *représentation simplicielle* de la molécule, ainsi que de ses poches : la molécule est matérialisée par les simplexes du complexe dual, et à l'inverse, les espaces vides entourant la molécule sont matérialisés par des simplexes de la triangulation de Delaunay ne faisant pas partie du complexe dual. En particulier, les poches constituent des ensembles connectés de triangles (ou de tétraèdres, en trois dimensions) de la triangulation de Delaunay extérieurs au complexe dual de la molécule. Dans la suite on parlera de *triangle vide* (ou de *tétraèdre vide*, leurs analogue en trois dimensions) pour désigner ces simplexes ; à l'inverse, on désignera par *triangle plein* (ou respectivement *tétraèdre plein*), ces mêmes simplexes lorsqu'ils sont inclus dans le complexe dual. Conformément à la nomenclature utilisée dans la librairie `CGAL`, on parlera aussi de cellules (vides ou pleines) pour désigner ces tétraèdres. Enfin, la surface duale de chaque poche ρ constitue un polyèdre $\mathcal{P}_\rho$ qu'on appellera *représentation polyédrique* de la poche.

Différentes approches peuvent être menées pour regrouper les tétraèdres vides afin de constituer des poches, et dans la suite nous présenterons celles que nous avons implémentées dans `Pck`.

10.1.2 Détection des cavités et des poches refermées : l'approche *tight*

Cette méthode de détection des poches est une implémentation de l'algorithme proposé en 1998 par H. Edelsbrunner [Edelsbrunner 98]. Cet algorithme bénéficie déjà de deux implémentations : le logiciel payant *CASTp* dont il existe aussi une version accessible en ligne au travers d'un service *web* [Binkowski 03b], et le logiciel libre (sous license LGPL) *Pocket* [Edelsbrunner 03] dont les sources en langage `Fortran` et en langage `C` sont peu exploitables dans le cadre d'une reprise du code pour modification ou évolution.

Cet algorithme permet la détection des cavités et des poches refermées telles que définies dans la section sur les poches de à la page 39 (respectivement les cas B et C de la figure 10.1). Les résultats de cet algorithme sont exploitables dans un très grand nombre de cas, et son implémentation dans `Pck` s'avérait un passage obligé ; elle nous a en outre permis de nous familiariser

[3] Cette notion d'accessibilité a été évoquée comme un moyen de caractériser la topographie d'une molécule à la section 3.2.2 page 30 de la partie bibliographie.

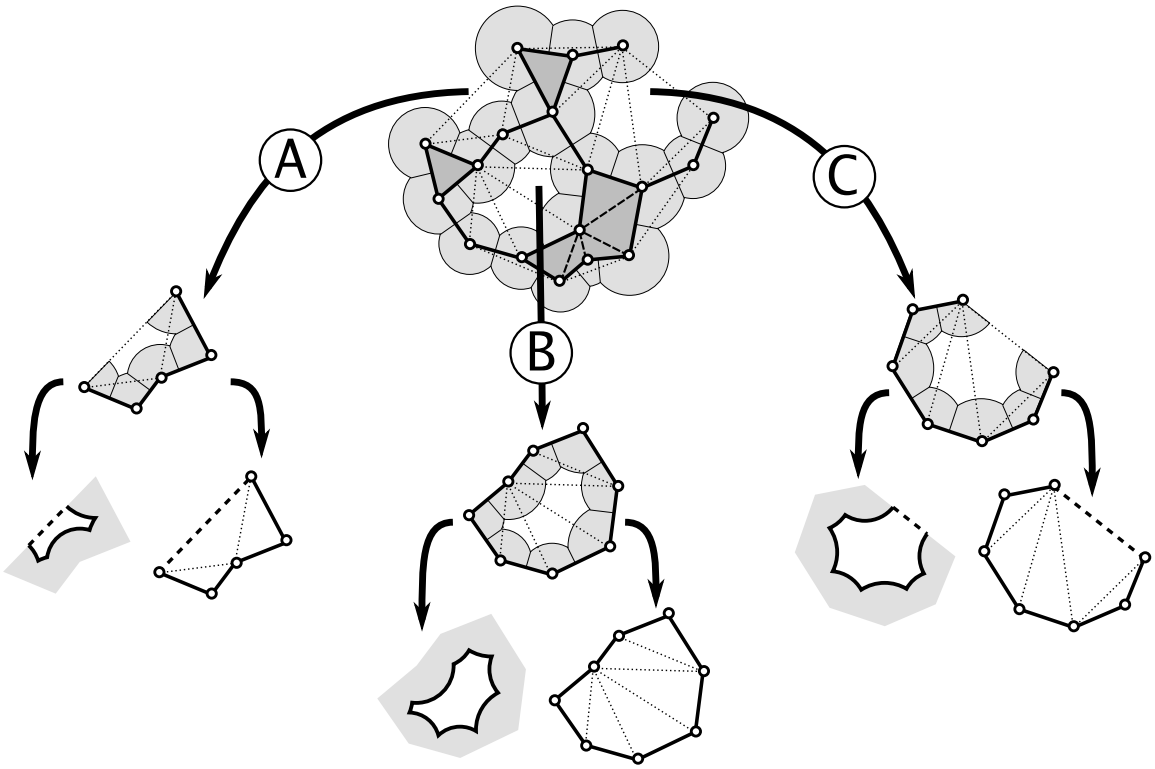


Figure 10.1: Représentation polyédrique des poches dans la forme duale, un exemple en deux dimensions. Le complexe dual d'une molécule (polyèdre en arêtes épaisses et triangles gris foncés) et sa triangulation de Delaunay (arêtes en pointillés) sont superposés au diagramme à remplissage de forme (boules tronquées grises). Trois poches (A, B et C) sont mises en exergue sous la figure ; leur volume, matérialisé par une ligne noire épaisse) est reproduit séparément dans leur représentation Surface Accessible (à gauche) et polyédrique (à droite). Les bouches sont matérialisées en traits interrompus.

avec le formalisme de la théorie des formes-α. Dans ce premier développement, les codes sources de *Pocket* nous ont été d'une grande aide en phase de débogage.

Une présentation de cet algorithme est donnée dans la section 4.4.2 page 48, et pour sa description complète on pourra se référer à la publication originale [Edelsbrunner 98].

10.1.3 Détection des poches ouvertes : l'approche *wide*

Nous avons exploré une méthode de détection des poches basée sur la différence d'accessibilité à la surface pour deux sphères-solvant de tailles différentes. Une première sphère de petite taille décrit la surface de la molécule, et une seconde, plus large, définit une surface plus grossière qu'on peut aussi considérer comme un "niveau de la mer"[4]. Les poches sont définies comme les espaces accessibles à la petite sphère et inaccesibles à la grande, ou de manière imagée, les espaces situés sous le "niveau de la mer". Ce genre d'approche pour la détection des poches dans les macromolécules a été évoqué dans l'état de l'art et plus particulièrement dans la figure 4.8 page 47.

Nous avons re-exprimé le problème dans le contexte de la théorie des formes-α, en considérant la différence entre deux complexes-α ; une poche est ainsi constituée de cellules de la triangulation de Delaunay qui appartiennent à un complexe-α $\mathcal{K}_{\alpha_2}$ mais n'appartiennent pas à un autre complexe-α $\mathcal{K}_{\alpha_1}$, avec des variables données $\alpha_2 < \alpha_1$ (voir figure 10.2). Dans nos travaux, nous

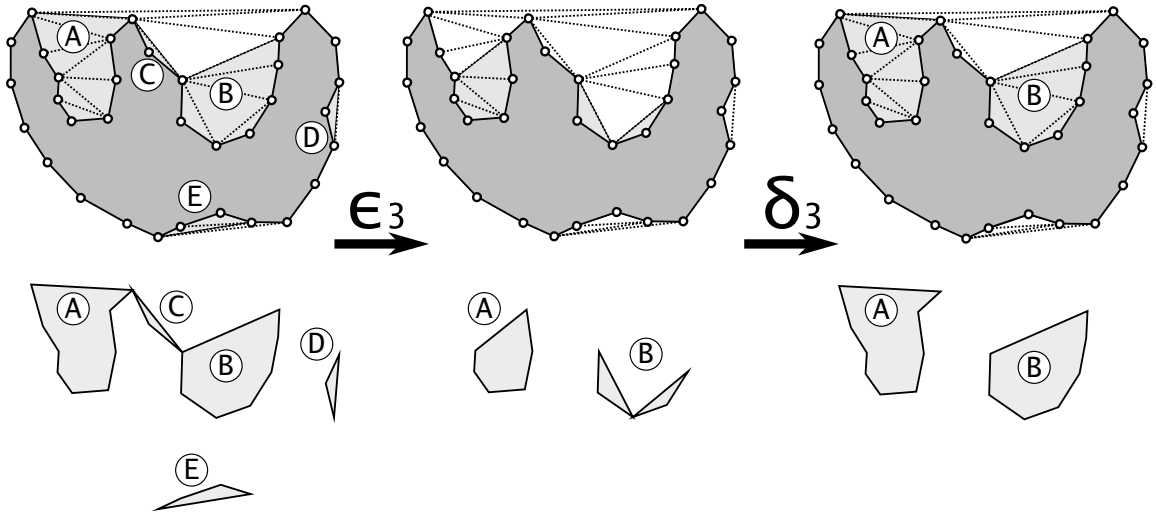


Figure 10.2: Principe général de l'algorithme de détection de poches *wide*, un exemple en deux dimensions. Sur la ligne du haut le complexe dual d'une molécule est représenté en gris foncé ; seuls les triangles de Delaunay extérieurs au complexe dual ont été représentés, matérialisés par des arêtes en pointillés. La ligne du bas reprend uniquement les agglomérats de triangles gris clairs représentant les poches dans la ligne du haut. Dans la figure de gauche, le complexe dual est superposé à un complexe-α $\mathcal{K}_{\alpha_1}$ pour une constante α_1 strictement positive. Une première étape d'érosion (ϵ_3) sur une profondeur de trois triangles "retire de la matière" et déconnecte les poches A et B ; les poches C, D et E disparaissent. Une phase de dilatation sur trois triangles (δ_3) matérialise l'espace des poches.

considérons la taille de la petite sphère comme implicitement donnée par la taille de la sphère-solvant utilisée pour la description du modèle Surface Accessible. Notre description "fine" de la surface correspond donc au complexe dual $\mathcal{K} = \mathcal{K}_0$ de la molécule dans sa représentation Surface Accessible, et nous fixons ainsi α_2 à 0.

[4] Nous avons évoqué cette intuition dans la partie bibliographique, au chapitre 3 présentant la topographie de la surface page 27.

Dans la pratique, la seule définition des poches au travers d'une appartenance différenciée à deux complexes-α n'est pas suffisante pour obtenir des réponses pertinentes. En effet, les poches ainsi définies sont encombrées de nombreuses "scories" telles que les petites poches D et E ou la connexion C entre les poches A et B dans la figure 10.2. En trois dimensions ce phénomène est encore plus flagrant, et ces "ébarbures" ont tendance à s'interconnecter pour ne faire plus qu'une unique et énorme poche recouvrant quasi intégralement la surface polyédrique de la molécule. Nous avons remédié à ce problème en utilisant un procédé d'érosion-dilatation — classique en informatique discrète, et que nous avons déjà mentionné au chapitre 4.4.2 page 45 — que nous avons adapté à notre contexte (voir les deux illustrations de droite dans la figure 10.2). En trois dimensions, ce processus d'érosion-dilatation peut parfois générer de petites aberrations, telles qu'une cellule vide au beau milieu d'une poche. Pour prévenir les occurrences de ce genre de "défaut de masse" dans les poches, nous avons recours à une dernière étape de *volumisation* dans laquelle nous ajoutons à une poche ρ toutes les cellules vides dont les quatre sommets sont incidents à une cellule de ρ.

10.1.4 Détection des sillons et des légères dépressions à la surface : l'approche *groove*

Les résultats visuels de la partie 8.4 suggéraient l'habilité de la courbure locale pour caractériser les espaces anfractués à la surface des macromolécules. Nous avons mis cette propriété à profit pour définir les poches sur un critère d'incurvation des atomes qui les composent, et proposé plusieurs stratégies pour définir un volume extérieur à la molécule à partir de ces zones légèrement enclavées à sa surface. Nous avons décliné ce principe général sur deux algorithmes : *groove-c*, orienté volume ; et *groove-fr*, orienté surface.

Les deux algorithmes partagent une première phase dans laquelle la courbure locale est calculée pour chaque sommet de la *composante extérieure* de la *surface duale* de la molécule ; de ce fait, ils partagent un premier paramètre s (pour *smooth*) qui permet de contrôler la taille du lissage utilisé pour la définition de la courbure locale. Un second paramètre partagé par ces deux algorithmes, t (pour *threshold*), définit le seuil de courbure locale au dessous duquel une zone à la surface de la molécule doit être considérée comme appartenant potentiellement à une poche.

Ces algorithmes reposent en partie sur une opération commune : la création d'agglomérats d'objets (cellules vides ou facettes sur la surface duale) en fonction d'un critère de voisinage (partage de facette ou d'arête). La réalisation de ce traitement est effectuée au travers d'une fonction générique *gatherConnectedParts(LL,L,P)* que nous avons implémentée à l'aide de la structure de données *union-find* [Cormen 01] dont on trouve une implémentation dans `CGAL` [Kettner 08]. Cette fonction prend en paramètres d'entrée une liste d'objets L et un prédicat P permettant de décider si deux objets sont voisins ; elle renvoie une liste LL dont chaque élément est un aggrégat d'objets présenté sous la forme d'une liste d'objets.

• *groove-c*, par agrégat de cellules anfractuées

Les principales étapes de *groove-c* (***groove(s)*** *detection based on Delaunay* ***cells***) sont résumées dans l'algorithme 10.1. Après une première étape où les valeurs de courbure locale sont calculées (ligne 1), les cellules vides du complexe dual sont parcourues, et seules celles dont les quatre sommets incidents ont une valeur de courbure inférieure à un seuil donné sont conservées (lignes 3 à 7). Parmi ces cellules, celles qui partagent une facette extérieure au complexe dual sont ensuite agglomérées, et chacun de ces aggrégats encode le volume d'une poche.

La ligne 4 de l'algorithme demande à ce qu'on puisse attribuer une valeur de courbure locale pour chaque sommet du complexe dual. Comme nous l'avons vu au chapitre 8.3 (page 103) ces

Algorithme 10.1 : détection de poche *groove-c*

Entrées :
K : une structure contenant le complexe dual et la triangulation de Delaunay sur laquelle il est construit
s : la taille du voisinage de lissage
t : une valeur seuil pour la courbure locale
Sorties :
$pckL$: Une liste de listes de cellules de K
▷ *Calculer des valeurs de courbure locale*
1 Constituer la composante extérieure de la surface duale, et calculer les valeurs de courbure locale.
▷ *Constituer une liste d'espaces anfractués*
2 Créer *groovyCells* une liste de cellules de K initialement vide
3 **pour** *chaque cellule vide c de K* **faire**
4 | **si** *les 4 sommets de c ont une valeur de courbure locale sous le seuil t* **alors**
5 | | insérer c dans groovyCells
6 | **fin**
7 **fin**
▷ *Agglomérer les cellules vides voisines*
8 $gatherConnectedParts(pckL, groovyCells, CellsAreConnected)$

valeurs sont calculées pour chaque sommet de la surface duale, et certains atomes étant associés à plusieurs de ces sommets, une stratégie doit être choisie pour l'attribution d'une valeur à un atome. Pour accélerer nos calculs nous avons choisi d'associer à un atome la valeur minimale de courbure calculée pour les sommets de la surface duale qui lui sont associés. Une implémentation plus précise mais moins efficace pourrait prendre en compte le côté de la surface où se situe la cellule considérée, et utiliser les valeurs de courbure locale associées aux sommets de la surface duale spécifiquement concernés.

• *groove-fr*, mise en volume d'un "fond de poche" défini par des facettes connectées sur la surface duale

Les étapes principales de *groove-fr* (***grooves*** *detection based on* ***facets****, and* ***restricted*** *to a common bucket*) sont synthétisées dans l'algorithme 10.2. Dans une première étape, l'ensemble des facettes de la surface duale sont traversées, et seules celles dont la valeur de courbure locale — calculée comme la moyenne des trois courbures locales mesurées aux sommets de la facette — sont conservées (lignes 3 à 8). Parmi ces facettes, celles qui partagent une arête sont ensuite agrégées pour constituer des fonds de poche (ou *buckets*) (ligne 10). Au cours d'une seconde étape (lignes 11 à 15), les cellules de la triangulation sont parcourues, et celles dont les quatre sommets appartiennent au même fond de poche sont retenues pour décrire le volume de la poche.

10.2 Caractérisation des poches

`Pck` propose trois types de propriétés pour caractériser les poches qu'il détecte :
- Les propriétés volumétriques (aire et volume des poches) sont calculables grâce à la description d'un volume physique pour chaque poche. Elles offrent une première idée des caractéristiques d'une poche.
- Un indice de convexité permet d'affiner les résultats volumétriques en procurant des indications sur la forme de la poche.

Algorithme 10.2 : principe de l'algorithme de détection de poche *groove-fr*

Entrées :
K : une structure contenant le complexe dual et la triangulation de Delaunay sur laquelle il est construit
s : la taille du voisinage de lissage
t : une valeur seuil pour la courbure locale
Sorties :
$pckL$: Une liste de listes de cellules de K
▷ *Calculer des valeurs de courbure locale*
Constituer S, la composante extérieure de la surface duale et calculer les valeurs de courbure locale.
▷ *Constituer des fonds de poche*
Créer $groovyFacets$ une liste de facettes initialement vide
pour *chaque facette* f *de* S **faire**
    Calculer l la moyenne des valeurs de courbure locale des trois sommets de f
    si l *est inférieure à* t **alors**
        empiler f sur $groovyFacets$
    fin
fin
Créer *bucket* une liste de listes de facettes de la surface duale initialement vide
$gatherConnectedParts(bucket, groovyFacets, FacetsShareEdge)$
▷ *Constituer le volume des poches à partir des fonds de poche*
pour *chaque cellule vide* c *de* K **faire**
    si *les quatre sommets de* c *sont dans le même fond de poche d'indice* b **alors**
        Insérer c dans pckL[b]
    fin
fin

algorithme	Volumétrie			type de poches détectées
	SA	VdW	Poly.	
molécule	o	o	o	
tight	o	e	o	cavités et poches refermées
wide	e	e	o	poches ouvertes, accessibilité différenciée
groove	e	e	o	poches ouvertes, anfractuées

Tableau 10.3: Résumé des algorithmes implémentés dans Pck. Les trois premières colonnes indiquent les valeurs volumétriques (volume et aire) calculées dans chacun des trois modèles, un *o* indique une valeur exacte, un *e* une valeur erronée qui peut toutefois être utilisée à titre indicatif.

– L'exploration de la proximité entre les poches permet d'inférer de potentiels recrutements d'espaces libres.

Ces propriétés sont détaillées ci-après.

10.2.1 Volumétrie des poches

L'aire et du volume sont des valeurs couramment employées pour caractériser une poche. Elles permettent par exemple d'estimer la taille d'un ligand potentiel, ou la pertinence de la poche observée pour une problématique donnée (trop petite ou trop grande) ; elle permettent aussi la comparaison de poches entre elles. Pck fournit des valeurs volumétriques pour la molécule comme pour les poches.

Edelsbrunner a proposé une méthode pour calculer les valeurs volumétriques d'une molécule de manière à la fois rapide et exacte [Edelsbrunner 95a] ; ces formules ont par la suite été adaptées au calcul de la volumétrie des cavités et des poches refermées [Edelsbrunner 95b, Edelsbrunner 94a, Edelsbrunner 98]. Appliquées aux poches, ces expressions permettent le calcul de l'aire et du volume d'une poche refermée dans le modèle Surface Accessible ; en guise d'approximation nous les avons aussi appliquées au calcul de la volumétrie de ces mêmes poches [5] dans le modèle Van der Waals, ainsi qu'au calcul volumétrique des poches détectées par nos autres algorithmes. Nous calculons aussi des valeurs volumétriques pour la représentation polyédrique de ces poches. Le tableau 10.3 résume la justesse de ces valeurs volumétriques calculées pour chaque type d'algorithme proposé par Pck.

10.2.2 Indice de convexité d'une poche

La donnée seule d'un volume ou d'une aire n'est pas suffisante pour se faire une représentation de la forme d'une poche ; pour affiner l'interprétation des valeurs volumétriques nous leur avons adjointe un *indice de convexité* qui permet de mesurer l'écart de la poche à son enveloppe convexe (voir figure 10.3).

Formellement, l'indice de convexité k_ρ d'une poche ρ est donné par le rapport entre le volume de la poche dans sa représentation polyédrique $\mathcal{P}_\rho$, et le volume de son enveloppe convexe $\mathcal{H}_\rho$.

$$k_\rho = \frac{Vol(\mathcal{P}_\rho)}{Vol(\mathcal{H}_\rho)}$$

Cet indice donne généralement une bonne mesure de la complexité de forme d'une poche, de ses "ramifications".

[5] C'est-à-dire de l'espace vide délimité par le même polyèdre.

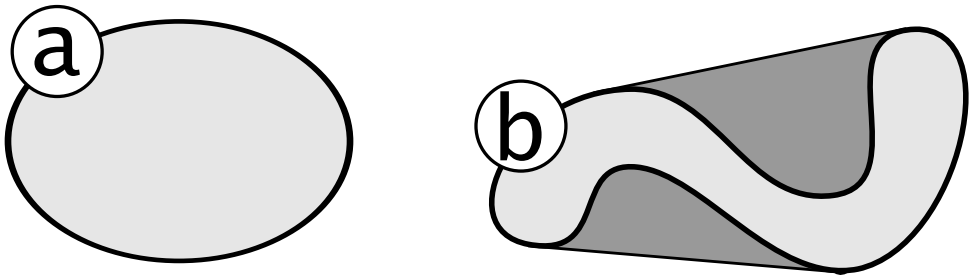


Figure 10.3: Indice de convexité, un exemple en deux dimensions. les deux objets gris clair a et b ont même volume, mais l'objet convexe a aura un indice de convexité de 1 alors que l'objet b, de forme plus complexe et d'épaisseur moindre, aura un indice proche de $\frac{1}{2}$.

10.2.3 Proximité entre poches

La proximité d'une poche à la surface de la molécule, ou la proximité de deux poches entre elles constitue une information exploitable par exemple dans le cadre de la conception d'un nouveau type de ligand ; elle indique une possible malléabilité de la molécule ainsi que des espaces potentiellement annexables, par exemple pour accommoder un groupement chimique supplémentaire à un ligand déjà connu.

`Pck` calcule une matrice contenant les distances minimales entre toutes les paires de poches détectées par un de nos algorithmes ; cette information est entre autres exploitée dans un outil d'exploration des poches voisines, dans le greffon VMD fourni avec `Pck`.

Nous avons encore mis à profit la classification[6] des facettes dans le complexe dual pour explorer plus directement le même type d'information et lui donner un support graphique intuitif (voir la figure 10.4). Dans le cadre particulier de l'étude des poches d'une macromolécule, la classification des facettes peut en effet s'interpréter simplement de la manière suivante (les indications en italique et entre parenthèses correspondent au formalisme employé dans `Pck`, et les couleurs font référence à celles employées dans la figure 10.4) :

- **Facettes régulières** (*pr*, en vert) : ces facettes sont communes à un tétraèdre vide et un tétraèdre plein dans le complexe dual. Elles symbolisent une zone du bord de la poche, infranchissable par une sphère solvant, et derrière laquelle on trouve des atomes de la molécule.

- **Facettes singulières** : bien qu'elles-mêmes appartiennent au complexe dual, ces facettes sont communes à deux tétraèdres vides. Elles symbolisent une zone du bord de la poche infranchissable par une sphère solvant, mais derrière laquelle on trouve du vide. Elles indiquent une zone potentiellement malléable. Dans `Pck` ces facettes sont discriminées en
  - **singulières dans la poche** (*pp*, en violet) : lorsque le vide de part et d'autre de la facette est situé dans la même poche.
  - **singulières entre deux poches** (*po*, en rouge) : lorsque le vide de part et d'autre de la facette est situé dans deux poches distinctes. Ce type de facette donne une indication sur un potentiel recrutement de poche.
  - **singulières sur l'infini** (*pi*, en bleu) : lorsque d'un des deux côtés de la facette le vide est situé dans l'espace du solvant.

- **Facettes extérieures** : ces facettes joignent deux tétraèdres vides dans le complexe dual.

[6] Cette classification a été introduite au chapitre 6.1.2 (page 61).

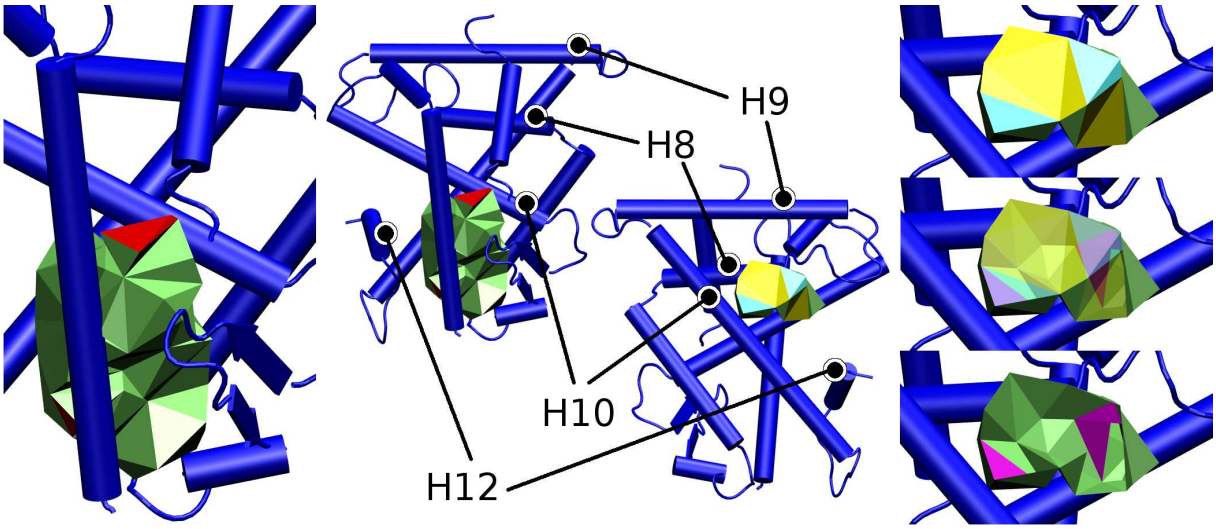


Figure 10.4: Classification des facettes dans `Pck`, un exemple sur deux poches dans la structure du domaine de fixation du ligand de RXRα (1RDT). Au centre, la structure du domaine est représentée en modèle *cartoon* dans deux orientations différentes, avec une poche distincte (détectées par *tight*) dans les deux cas : à gauche, la structure est présentée dans son orientation canonique avec la poche de fixation du ligand, à droite la structure a subi une rotation de 180° autour de l'axe des z de manière à montrer la poche dans le creux du sillon à l'interface de dimérisation le long de l'hélice H10. Les poches sont répétées de part et d'autre de cette figure. Les trois illustrations de droite montrent chacune la poche dans le creux du sillon, avec les facettes de la bouche (en haut) ou sans elles (en bas) ; l'image du milieu montre la même poche avec les facettes de la bouche en transparence.
Les facettes vertes sont régulières et pavent le fond de la poche, les rouges sont singulières et constituent un opercule partagé par deux poches, les violettes (uniquement dans la figure du bas de la poche du sillon) constituent le même genre de fermeture mais au sein du volume d'une poche. Les facettes jaunes et bleu (poche du sillon) constituent respectivement, les bouches d'une poche, et une fine membrane garantissant la poche du franchissement du solvant.

Lorsqu'elles sont situés sur le bord d'une poche, elles participent à ses bouches (*mo*, en jaune).

10.3 Détection des poches : une étude concrète sur quelques macromolécules

Dans cette section nous comparerons les poches détectées avec les différentes approches que nous avons implémentées dans `Pck`. Les structures moléculaires utilisées pour cette étude ont déjà été introduites au chapitre 8 où nous avons étudié la topographie de leur surface.

10.3.1 Détection des poches à l'interface d'un dimère de récepteurs nucléaires

La famille des récepteurs nucléaires a été introduite à la section 8.4.2 (page 105) lorsque nous avons étudié la topographie de la surface du domaine de liaison au ligand de RXRα ; nous y avons entre autres évoqué que les récepteurs nucléaires peuvent être fonctionnels sous forme de monomères, d'homodimères, ou d'hétérodimères. La figure 10.5 montre une telle association entre les domaines de liaison au ligand des deux récepteurs nuclaires PPARγ et RXRα. Nous avons

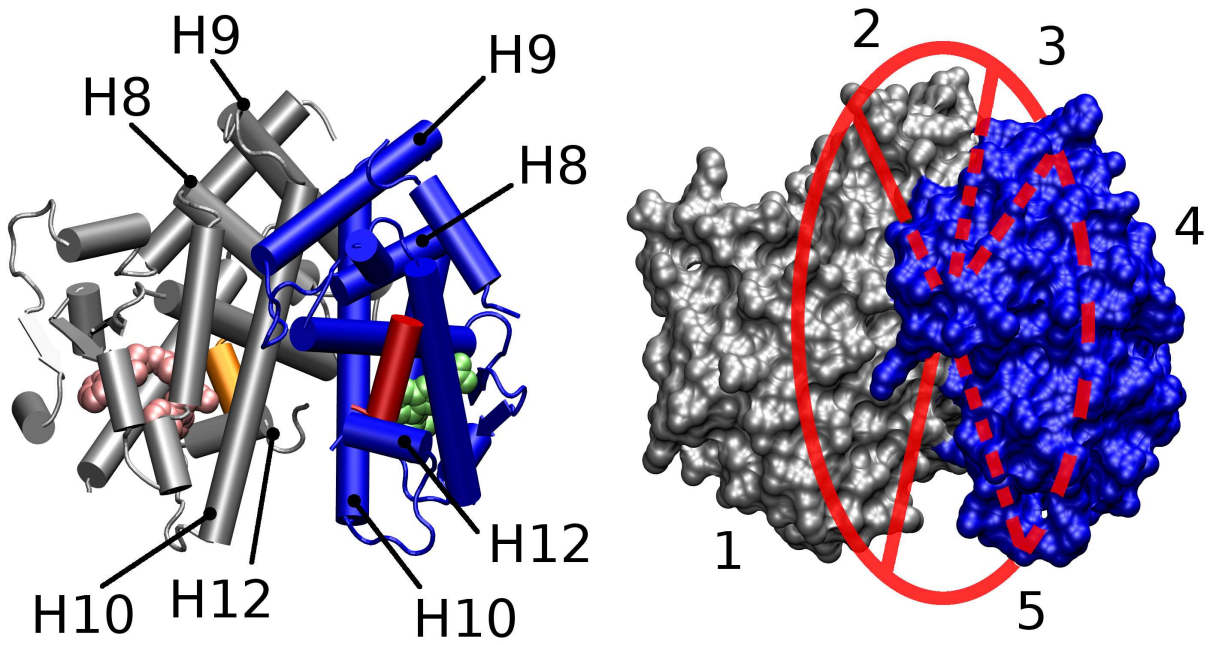


Figure 10.5: Structure du dimère impliquant les domaines de liaison au ligand des récepteurs nucléaires PPARγ et RXRα (à gauche), et découpage de la crevasse à l'interface (à droite). Dans la figure de gauche les deux domaines de liaison au ligand sont en représentation *cartoon* ; RXRα en bleu à l'avant, et PPARγ en gris à l'arrière. Dans cette structure, chacun des deux récepteurs nucléaires est complexé à un ligand (en représentation de Van der Waals et en vert pour RXRα, en rose pour PPARγ) ainsi qu'à un cofacteur (en représentation *cartoon* et en rouge pour RXRα, en orange à l'arrière pour PPARγ). La figure de droite montre l'assemblage des deux structures en représentation Surface Moléculaire avec les mêmes couleurs que dans l'illustration de droite ; l'interface est cernée d'une crevasse dans laquelle on peut distinguer cinq sections, représentées par des arcs de cercle rouges numérotés.

illustré l'utilisation de nos algorithmes de détection des poches sur cet assemblage ; les résultats observés sont analysés ci-après.

La caractéristique principale de cette structure — si l'on excepte les caractéristiques de chacun des deux domaines pris séparément — réside dans une crevasse particulièrement marquée à l'interface des deux domaines (figures 10.5, 10.7, et 10.6). Cette crevasse peut être découpée en quatre ou cinq sections. La plus marquée et la plus évasée de ces sections (section 1) se situe à l'avant de l'image, entre les hélices H9, H10 et la boucle joignant H8 à H9 dans RXRα, et les hélices H7 et H10 de PPARγ. Une section équivalente quoique moins marquée et plus anfractuée peut être observée à l'opposée de la première, mettant en jeu les mêmes éléments secondaires dans les domaines opposés. Cette crevasse s'achève sur un puit entre les deux structures, limité par la boucle joignant les hélices H8 à H9 de PPARγ. Deux autres sections de moindre taille peuvent être observées en haut et en bas de la structure. A peu de choses près, tous les algorithmes ont révélés les mêmes éléments de relief dans ce sillon, quoique de manière sensiblement différente, et pas toujours satisfaisante.

Comme on peut l'observer dans la figure 10.6, *tight* ne détecte que le haut du sillon principal ; tandis qu'avec des paramètres corrects, *wide* en permet une extraction plus satisfaisante. Avec *wide*, l'utilisation des paramètres permet en outre d'explorer les espaces accessibles à une sphère de taille différente, mettant en lumière un la "profondeur" de la crevasse : plus le paramètre α_1 est élevé, et plus la définition de la crevasse est volumineuse. Remarquons aussi qu'avec de mauvais paramètres (notamment avec trop peu d'érosion), on peut en venir à connecter toutes les petites crevasses en surface.

La figure 10.7 montre la même crevasse détectée par *groove-fr*. Dans cet exemple la sensibilité aux paramètres est flagrante, avec une valeur de seuil de 0.41 pour définir les facettes constituant les "fonds de poche", seul le fond du secteur avant de la crevasse est sélectionné, alors qu'avec une valeur de 0.43 ou plus, le sillon détecté fait quasiment le tour du complexe.

10.3.2 Détection des poches à la surface de récepteurs nucléaires

Dans cette sous-section l'étude des différences entre les poches détectées par nos algorithmes sera restreinte à la surface d'un récepteur nucléaire. La figure 10.8 montre isolement le domaine de liaison au ligand du récepteur nucléaire RXRα présenté précédemment en complexe avec PPARγ (figure 10.5). On y retrouve la structure caractéristique en "sandwich" d'hélices-α des récepteurs nucléaires évoquée à la section 8.4.2 page 105. L'image de gauche montre aussi la poche accomodant le ligand dans cette structure, il s'agit en l'occurrence d'une cavité (une poche totalement enfouie sous la surface). L'image de droite montre une autre poche, ouverte sur l'espace du solvant celle-ci, qui a été proposée comme une poche de liaison alternative dans le cas des récepteurs nucléaires à L'estrogène ERα et ERβ [van Hoorn 02]. Les deux facettes rouges en haut de la poche de fixation du ligand (figure 10.8, image de gauche) sont partagées entre les deux poches, suggérant la possibilité de connecter les deux espaces avec un réarrangement spatial moindre des résidus Ser312, Leu309 et Gln275, directement impliqués dans ces triangles, ou de leurs voisins.

Cette structure présente de nombreuses poches et concavités à sa surface ; dans la suite de cette sous-section nous nous intéresserons plus particulièrement à trois d'entre elles : le sillon de recrutement du cofacteur, le sillon à l'interface de dimérisation le long de l'hélice H10, et la poche alternative que nous venons de décrire. La position spatiale de ces trois zones caractéristiques à la surface de RXRα est synthétisée dans la figure 10.9, et les trois éléments peuvent être observés respectivement dans les figures 10.10, 10.11 et 10.12 où l'habilité des divers algorithmes implémentés dans `Pck` est observée.

La poche où se niche le cofacteur est assez peu marquée ; en particulier elle ne présente pas de constriction et s'ouvre largement sur l'espace du solvant (voir figure 10.10). De fait, *tight* ne

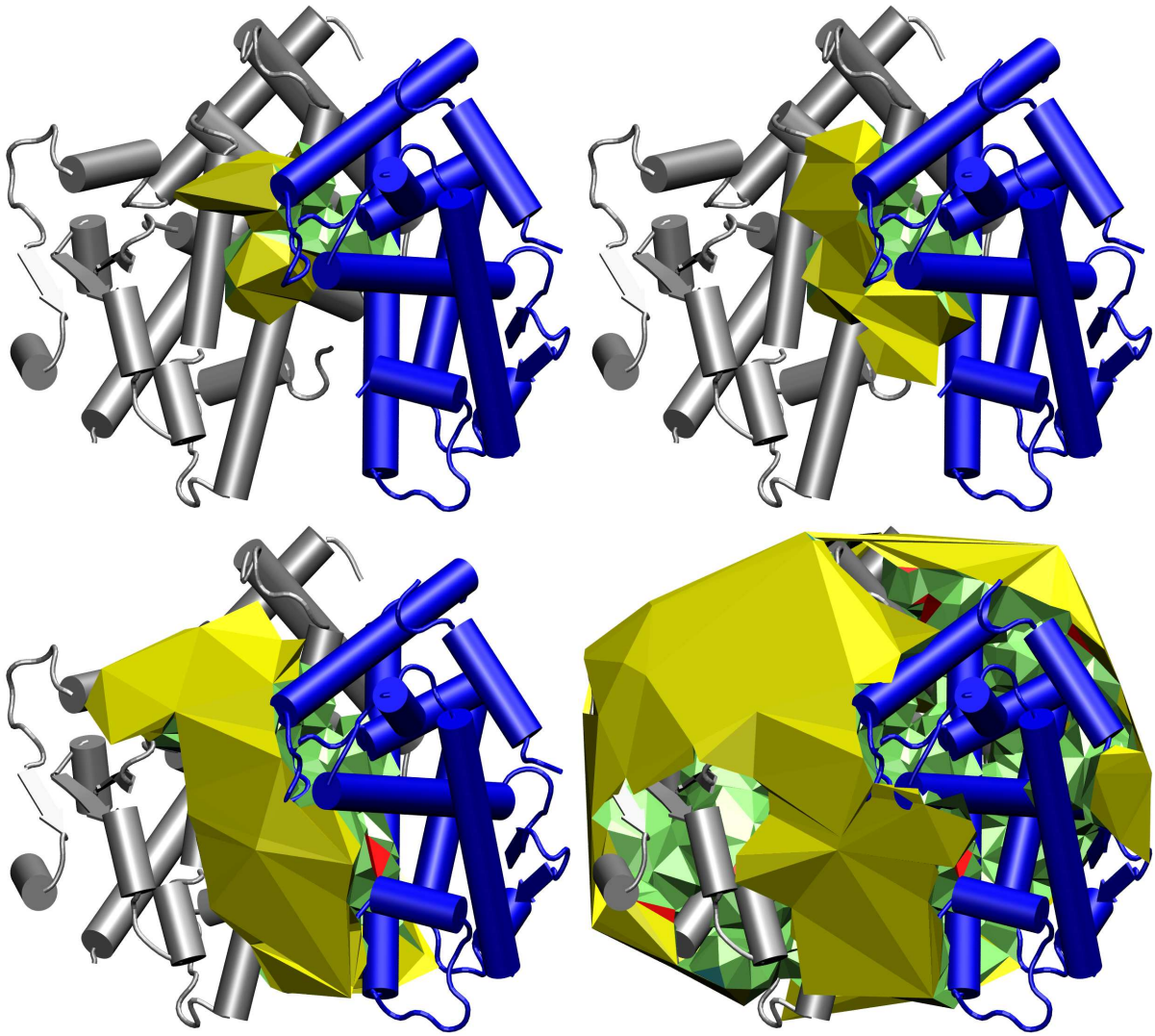


Figure 10.6: Détection de la crevasse principale à l'interface des domaines de liaison au ligand de PPARγ et RXRα (1RDT) ; comparaison de *tight* (en haut à gauche), et de *wide* avec les valeurs de paramètres respectivement de $\alpha_1 = 40$, $\epsilon = 10$ en haut à droite, $\alpha_1 = 80$, $\epsilon = 10$ en bas à gauche, $\alpha_1 = 80$, $\epsilon = 5$ en bas à droite.

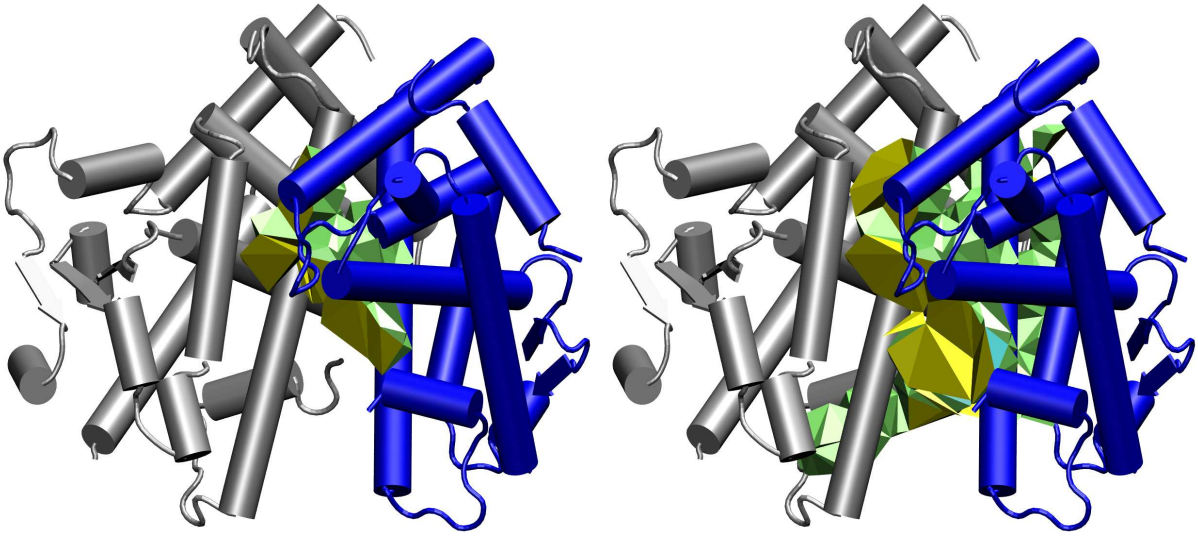


Figure 10.7: Détection de la crevasse principale à l'interface des domaines de liaison au ligand de PPARγ et RXRα (1RDT) avec *groove-fr*. A gauche, avec les paramètres $s = 3$, $t = .41$, à droite avec les paramètres $s = 3$, $t = .43$.

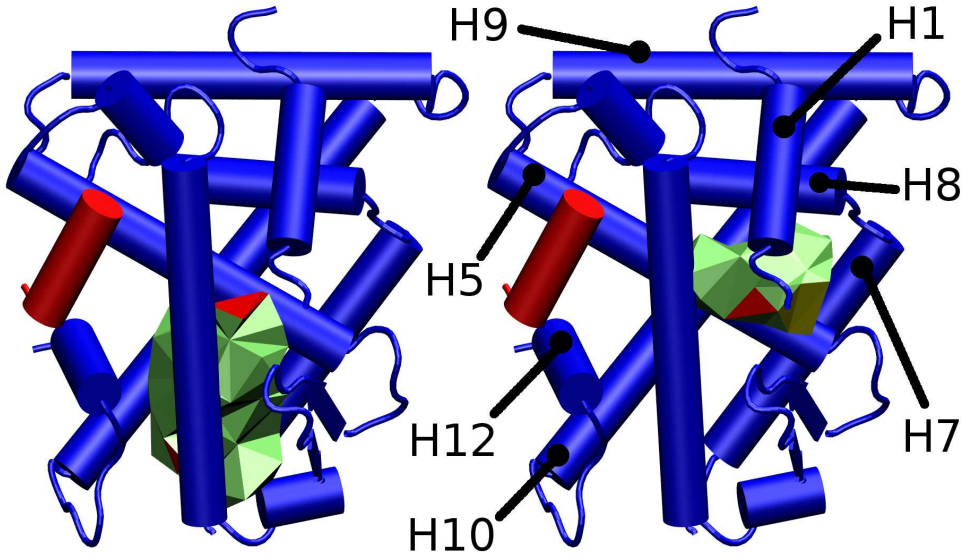


Figure 10.8: Structure du domaine de liaison au ligand du récepteur nucléaire RXRα (1RDT) et de deux de ses poches détectée par *tight*. L'image de gauche montre la poche du ligand, et celle de droite montre la poche alternative.

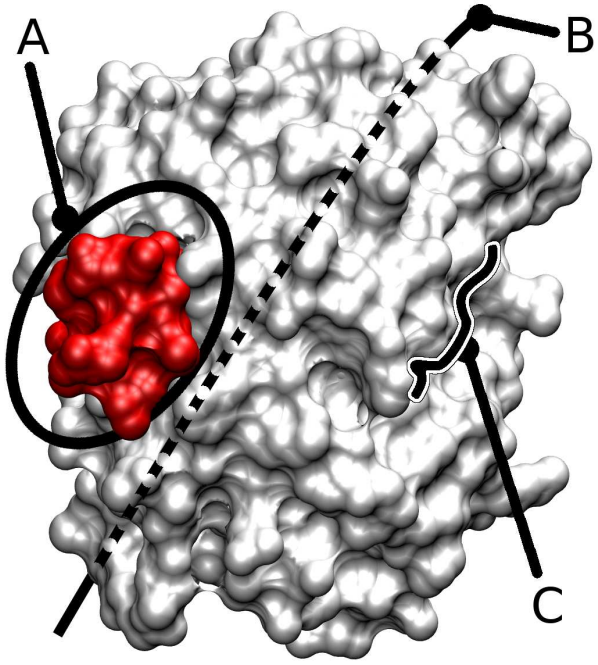


Figure 10.9: Localisation de trois poches caractéristiques à la surface du domaine de liaison au ligand de RXRα (1RDT). La protéine (en blanc) et son cofacteur (en rouge) sont représentés en modèle Surface Moléculaire, et on peut apprécier leur complémentarité de forme ; le sillon du cofacteur est indiqué par un ovale noir (A). Au niveau de l'interface de dimérisation, de l'autre côté de la structure, un sillon matérialisé par une courbe en pointillés court le long de l'hélice H10 (B). L'entrée de la poche alternative est marquée d'une ligne noire (C).

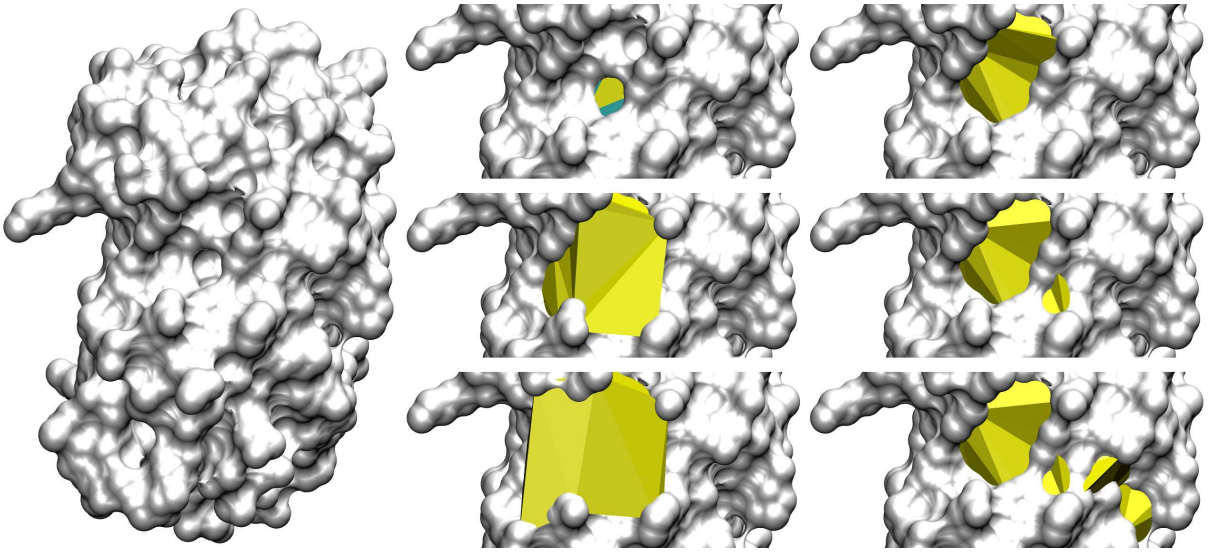


Figure 10.10: Détection du sillon du cofacteur à la surface du domaine de liaison au ligand de RXRα avec les algorithmes de `Pck`. A gauche, la structure de RXRα a subi une légère rotation autour de l'axe des z par rapport à sa présentation dans la figure 10.9 afin de présenter le sillon du cofacteur. A droite, le détail de la détection de ce sillon pour nos différents algorithmes. De haut en bas, dans la première colonne : *tight*, *wide*($\alpha_1 = 50$, $\epsilon = 5$), *wide*($\alpha_1 = 80$, $\epsilon = 8$). De haut en bas, dans la dernière colonne : *groove-c*($s = 2$, $t = .56$), *groove-fr*($s = 2$, $t = .47$), *groove-fr*($s = 2$, $t = .48$).

trouve que les six atomes les plus profondément enfouis dans le fond de ce sillon. Avec *wide* cette déclivité est correctement décrite avec une inaccessibilité et une érosion moyenne ($\alpha_1 = 50$, $\epsilon = 5$), et en augmentant l'inaccessibilité, le sillon prend un peu de volume et s'étend vers le sillon de l'interface de dimérisation ($\alpha_1 = 80$, $\epsilon = 8$). Notons que pour des facteurs d'érosion à peine plus élevés (($\alpha_1 = 50$, $\epsilon = 7$) et ($\alpha_1 = 80$, $\epsilon = 8$)), la poche n'est plus détectée. Les algorithmes *groove-c* et *groove-fr* permettent aussi une bonne reconnaissance du sillon du cofacteur. Pour *groove-fr*, on remarque toutefois que le passage des paramètres ($s = 2$, $t = .47$) à ($s = 2$, $t = .48$) révèle un sillon annexe s'échappant jusque sous l'hélice H1.

La figure 10.11 présente l'interface de dimérisation du domaine de liaison au ligand de RXRα. Comme dans le cas du sillon du cofacteur, le sillon courant le long de l'hélice H10 à l'interface

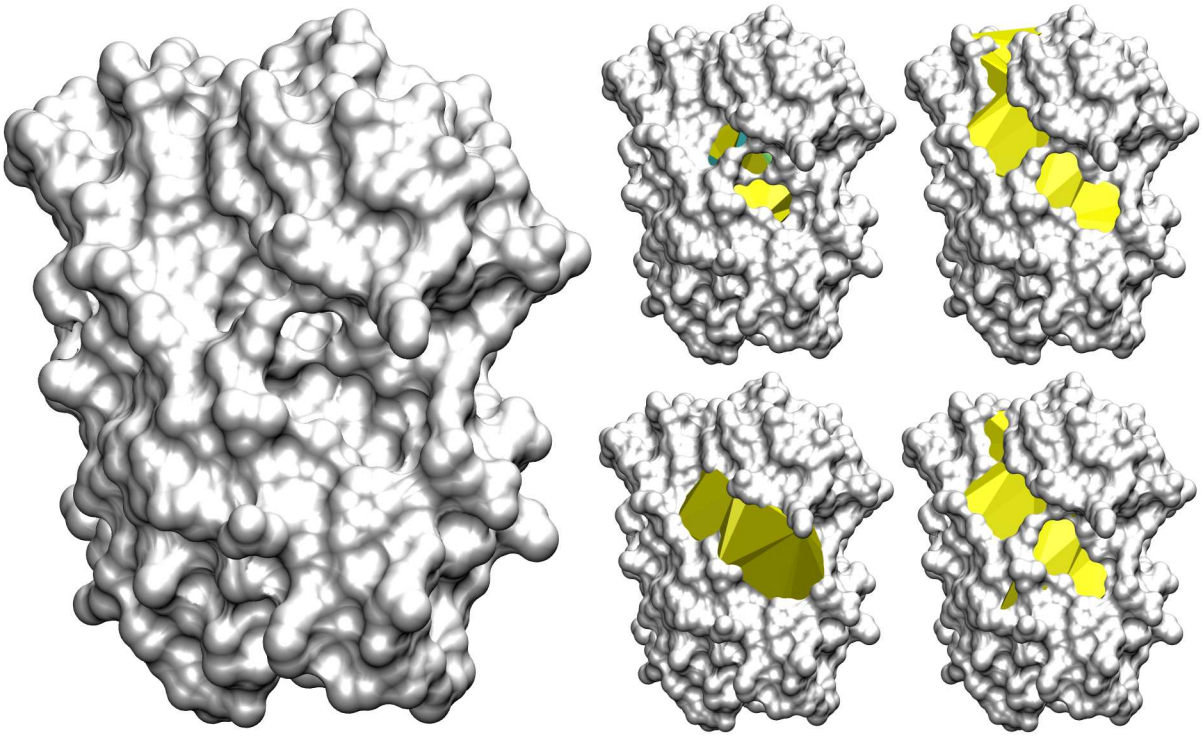


Figure 10.11: Détection du sillon courant le long de l'hélice H10 au niveau de l'interface de dimérisation du domaine de liaison au ligand de RXRα avec les algorithmes implémentés dans `Pck`. A gauche, la molécule a subi une rotation de près de 180° autour de l'axe des z par rapport à sa présentation dans la figure 10.9 afin de présenter l'interface de dimérisation et son sillon. A droite, de haut en bas dans la colonne du milieu : *tight*, *wide*($\alpha_1 = 80$, $\epsilon = 8$). De haut en bas, dans la colonne de droite : *groove-c*($s = 2$, $t = .56$), *groove-fr*($s = 2$, $t = .47$).

de dimérisation ne présente aucune constriction ; de fait, *tight* n'en détecte que les deux parties les plus centrale sous la forme de deux poches distinctes. *wide*, par contre, se comporte particulièrement bien, avec d'infimes variations même lorsque les paramètres varient entre ($\alpha_1 = 50$, $\epsilon = 6$) et ($\alpha_1 = 80$, $\epsilon = 10$) (l'image correspondante à ce second paramètre n'est pas présentée). *groove-c* permet aussi une bonne détection du fond du sillon ; sa définition est moins volumineuse mais plus allongée que celle de *wide* ; elle met en exergue la remontée du sillon au dessus de la structure, vers une petite "fracture" qu'elle révèle au milieu de l'hélice H9. *groove-fr* permet de détecter approximativement le même sillon ; il faut toutefois noter qu'en utilisant des paramètres sensiblement équivalents (($s = 2$, $t = .48$) au lieu de ($s = 2$, $t = .47$)), toute la zone d'interface est considérée comme une poche, indiquant que vu de loin, le sillon est à peine plus marqué

que le reste de la zone d'interface. De manière intéressante, à plus haute résolution et avec un seuil drastique ($s = 5$, $t = .43$), seules deux poches sont détectées (et bien formées) : le sillon de l'interface de dimérisaton et la poche alternative ; indiquant qu'à cette échelle, ces zones sont reconnues comme les plus anfractuées de la molécule.

La figure 10.12 montre dans la structure de RXRα, la poche évoquée plus haut comme poche alternative du ligand dans le cas de ERα et ERβ. Cette poche, présentant une légère constriction

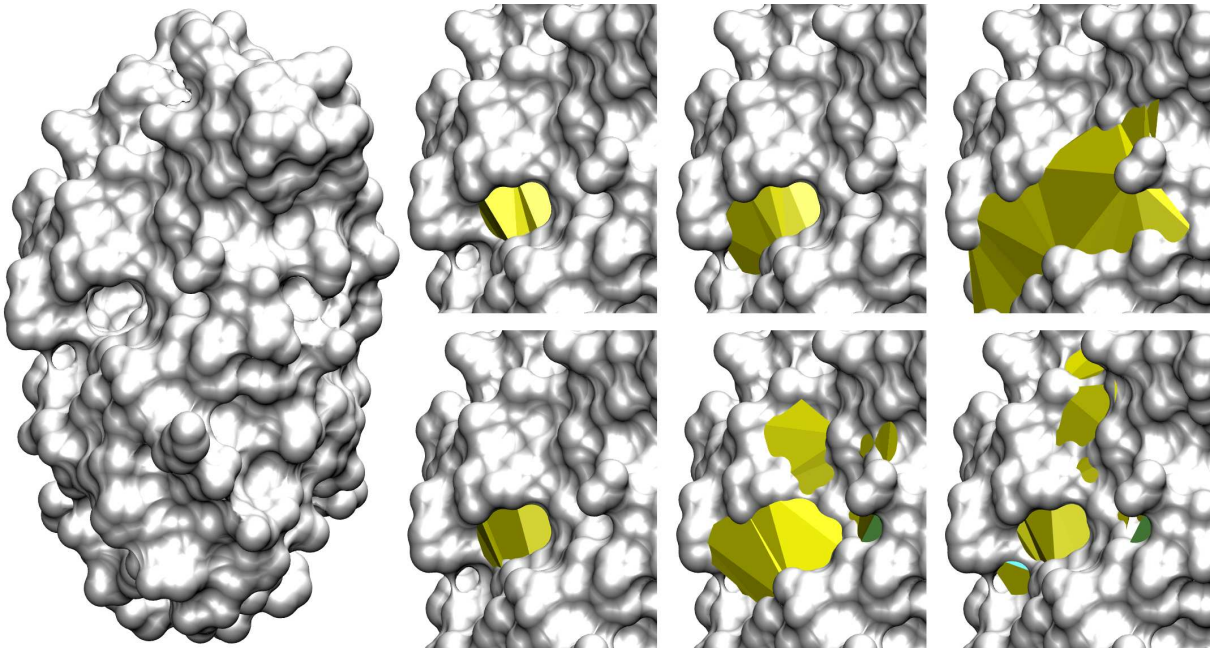


Figure 10.12: Détection de la poche alternative dans le domaine de liaison au ligand de RXRα avec les algorithmes de `Pck`. A gauche, la molécule a subi une rotation de près de 90° autour de l'axe des z par rapport à sa présentation dans la figure 10.9 afin de présenter la poche alternative. A droite : de gauche à droite dans la ligne du haut *tight*, *wide*($\alpha_1 = 50$, $\epsilon = 10$), *wide*($\alpha_1 = 50$, $\epsilon = 5$). De gauche à droite dans la ligne du bas : *groove-c*($s = 5$, $t = .50$), *groove-c*($s = 2$, $t = .56$), *groove-fr*($s = 2$, $t = .47$).

à sa sortie, est convenablement détectée avec *tight* (en seconde position si l'on se réfère au volume, et juste après la poche d'accomodation du ligand). Avec les paramètres ($\alpha_1 = 50$, $\epsilon = 10$), *wide* clos convenablement la poche et révèle dans le même temps la présence d'un sillon étroit sous la boucle de l'hélice H1. Ce sillon est en fait provoqué par l'abscence des résidus 245 à 263 dans la structure. Avec cinq érosions au lieu de dix, d'autres sillons sont recrutés qui mettent en lumière une crevasse à la sortie de la poche. A haute résolution avec un seuil drastique ($s = 5$, $t = .50$), *groove-c* délimite la poche de la même manière que *tight*. Avec une résolution plus locale et un seuil plus élevé, ($s = 5$, $t = .50$) le même algorithme révèle un sillon au dessus de l'entrée de la poche qui n'a pas été considéré par les algorithmes précédents. Le même sillon est détecté par *groove-fr* à un niveau de résolution moindre ($s = 2$, $t = .47$).

Nous nous sommes interrogés sur la capacité de nos algorithmes à déceler les mêmes éléments caractéristiques à la surface d'un autre récepteur nucléaire, PPARγ. Le domaine de liaison au ligand de ce récepteur nucléaire est plus volumineux que celui de RXRα (25500Å^3 et 21000Å^3 pour les volumes respectifs de leur modèles de Van der Waals). La poche de fixation du ligand de PPARγ est, elle aussi, beaucoup plus volumineuse que celle de RXRα ; elle s'ouvre sur l'espace du solvant au travers d'une crevasse allongée et profonde (pour comparaisons, différentes valeurs volumétriques sont rassemblées dans le tableau 10.4). La figure 10.13 montre la détection de cette

poche ainsi que celle du sillon du cofacteur par les quatre algorithmes implémentés dans `Pck`. Les

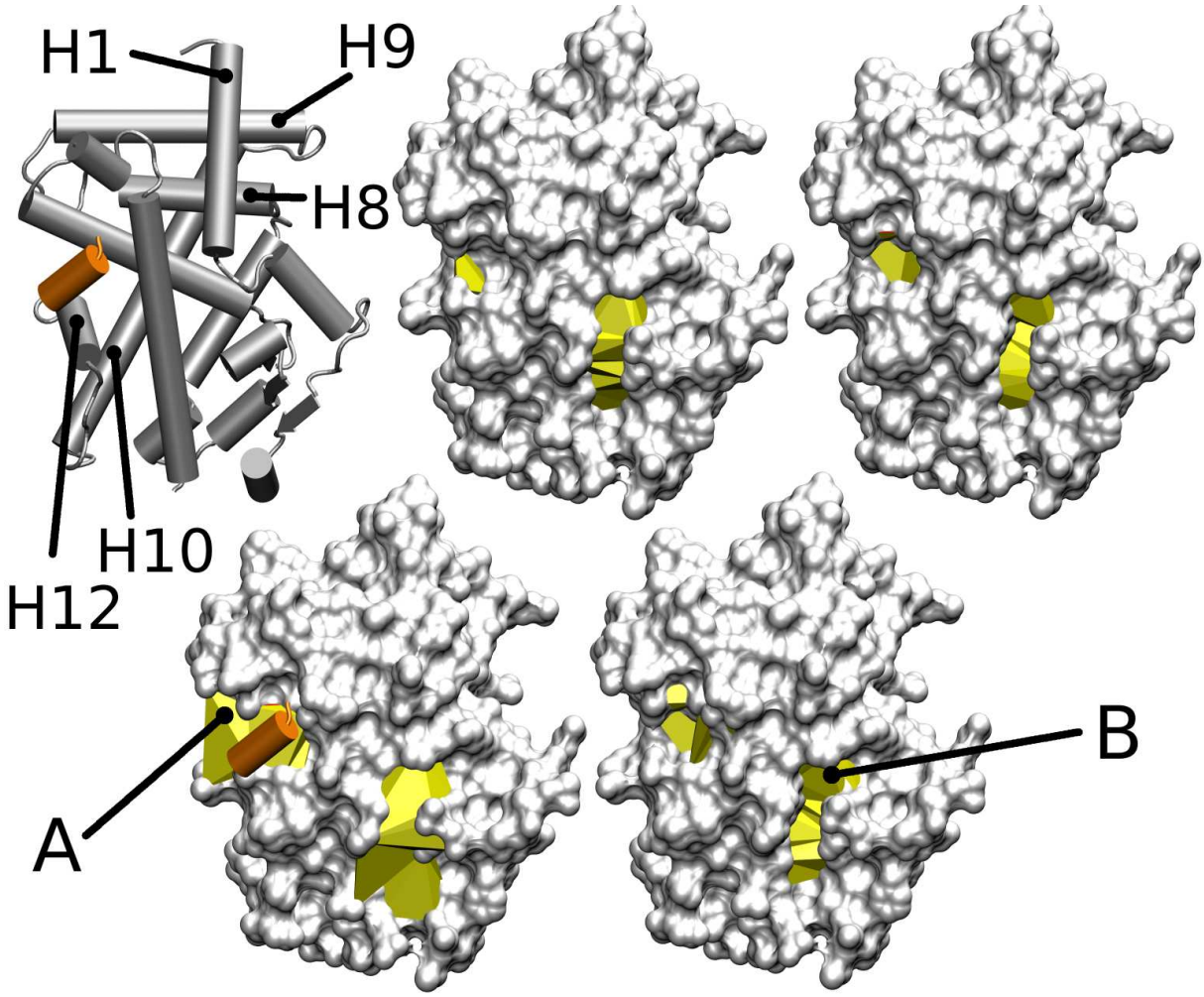


Figure 10.13: Structure du domaine de liaison au ligand de PPARγ (en haut à gauche en représentation *cartoon*, le cofacteur en orangé), et détection de la poche de fixation du ligand (B) et du sillon du cofacteur (A) avec les algorithmes de `Pck`. Les poches sont détectées avec les algorithmes suivants : Ligne du haut et de gauche à droite *tight* et *groove-c*($s = 3$, $t = .52$) ; ligne du bas de gauche à droite, *wide*($\alpha_1 = 40$, $\epsilon = 6$) et *groove-fr*($s = 3$, $t = .46$).

quatre algorithmes détectent de manière satisfaisante la poche du ligand, il n'en va pas de même pour le sillon du cofacteur. Comme dans le cas de RXRα, *tight* ne détecte qu'une infime partie du sillon du cofacteur, qui plus est excentrée. A l'inverse, les deux algorithmes *groove* détectent parfaitement la localisation de ce sillon. La zone autour de l'ouverture de la poche du ligand s'élargissant progressivement, *wide* lui accorde plus de volume que les autres algorithmes, et ce malgré un paramètre d'inaccesibilité $\alpha_1 = 40$ peu élevé. De même, *wide* accorde plus de volume au sillon du cofacteur.

La poche de fixation du ligand de PPARγ affleure presque à la surface de la molécule au dessus du feuillet β ; à l'endroit où l'on peut trouver la poche alternative dans RXRα. Nos algorithmes montrent que cette caractéristique de surface est aussi présente à la surface de PPARγ, quoique beaucoup moins marquée ; il s'agit ici plutôt d'une dépression à la surface que réellement d'une "poche" (voir figure 10.14). En raison encore une fois de l'aspect très évasé de cette zone de la surface, *tight* n'est pas très efficace et trouve uniquement quatre minuscules modules déconnectées.

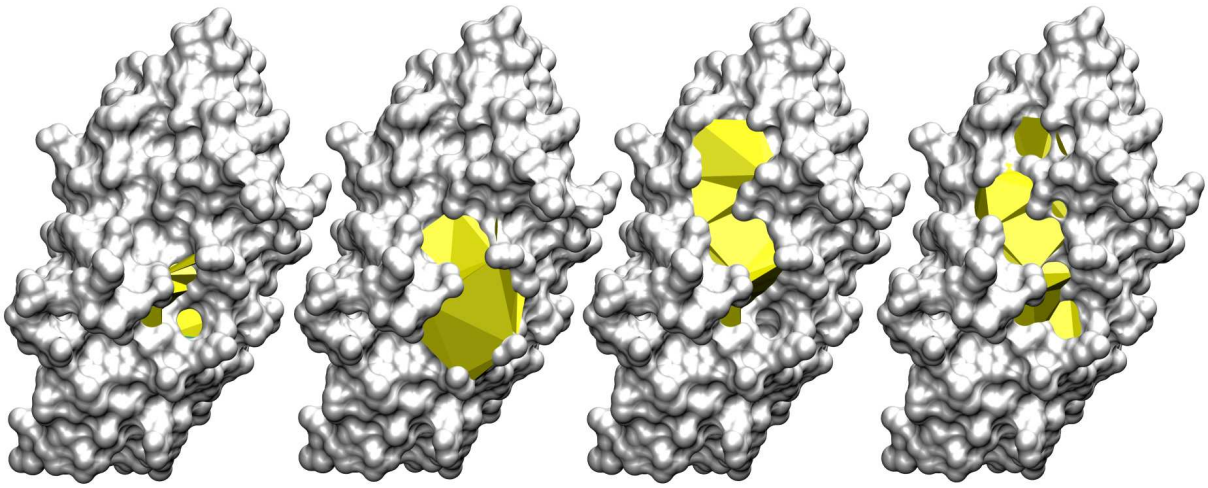


Figure 10.14: Détection de la poche alternative dans le domaine de liaison au ligand de PPARγ avec les algorithmes de `Pck`. De droite à gauche, les poches détectées avec les algorithmes *tight*, *wide*($\alpha_1 = 40$, $\epsilon = 6$), *groove-c*($s = 3$, $t = .51$), et *groove-fr*($s = 3$, $t = .45$).

poche	volume		aire		aire des bouches		k_ρ
	SA	Poly	SA	Poly	SA	Poly	
PPARγ (poche du ligand)	285	2620	687	1300	21	173	63.6
RXRα (poche du ligand)	143	1340	285	817	0	0	73.2
RXRα (poche alternative)	47	527	117	355	4	60	85.7
PPARγ-RXRα(puit à l'interface)	19	466	86	360	0,1	17	91.9

Tableau 10.4: Propriété calculées pour quelques poches détectées avec *tight* dans la structure de l'hétérodimère PPARγ-RXRα (1RDT).

A l'inverse, *wide* détecte parfaitement la poche alternative et lui accorde un volume satisfaisant. Les algorithmes *groove-c* et *groove-fr* retrouvent eux aussi la poche, mais lui accordent moins de volume ; en revanche, l'un comme l'autre mettent en lumière une extension du sillon vers l'hélice H9.

10.3.3 Détection des sillons de l'ADN

La figure 10.15 montre les différentes définitions des sillons que nous avons pu obtenir avec les algorithmes implémentés dans `Pck`. Le petit sillon présente suffisamment de constrictions pour être détecté par *tight*, ce n'est pas le cas du grand sillon. Les trois autres algorithmes détectent correctement les deux sillons. De nos quatre algorithmes, *groove-fr* permet de récupérer les deux sillons les plus complets, et est moins dépendant des paramètres d'entrée que ne le sont *wide* et *groove-c*.

10.4 Discussion et perspectives

Nous avons présenté nos développements dans le cadre de la détection et de la caractérisation des poches dans les macromolécules biologiques. Ces travaux ont concrètement abouti au logiciel `Pck` développé en langage `C++` et utilisant la bibliothèque `CGAL` [CGAL 09], ce qui a permis de réaliser des exécutables suffisamment rapides pour une utilisation dans des traitements par lot impliquant de nombreux traitements. Un outil d'analyse est aussi fourni sous la forme d'un

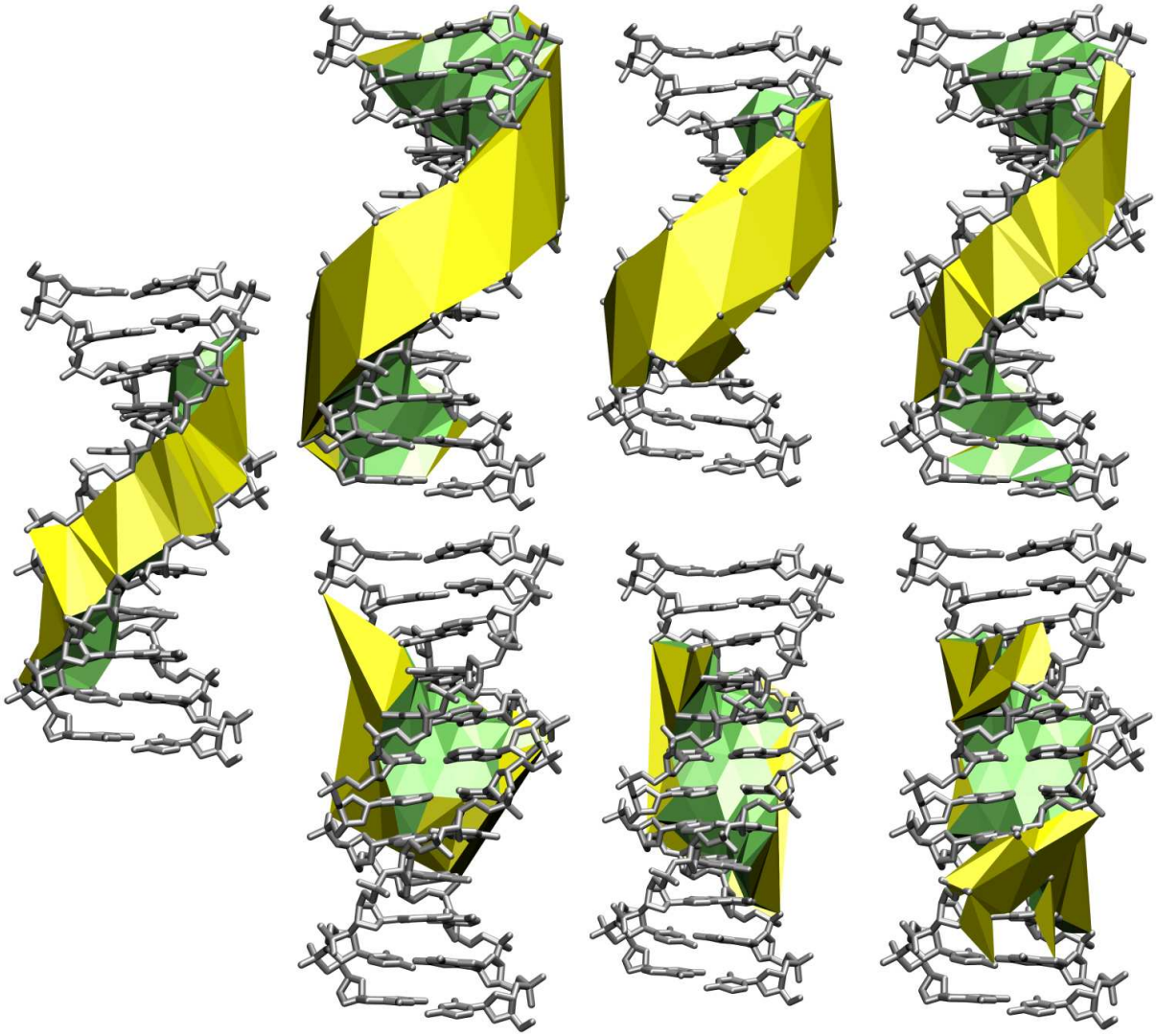


Figure 10.15: Détection des sillons sur un dodécamère d'ADN (1AD65) avec les algorithmes de `Pck`. A gauche le petit sillon détecté par *tight*, la colonne du milieu montre les deux sillons détectés par *wide* ($\alpha_1 = 40$, $\epsilon = 6$), et la colonne de droite montre les deux sillons détectés par *groove-fr* ($s = 2$, $t = .54$).

greffon `tclTk` pour le logiciel de visualisation VMD [Humphrey 96], comme il est précisé dans l'annexe D qui en résume les fonctionnalités.

10.4.1 Concernant la détection des poches

Dans le cadre de ce développement nous avons implémenté quatre algorithmes distincts autorisant la détection d'un éventail assez large de poches dans les macromolécules, et l'attribution systématique d'un volume physique à chacune d'entre elle. A notre connaissance, notre outil est le seul à proposer plusieurs méthodes de détection alternatives et complémentaires aux utilisateurs. La détection des cavités et des poches refermées[7] est assurée par *tight*, une implémentation de l'algorithme proposé par H. Edelsbrunner [Edelsbrunner 98]. Trois nouveaux algorithmes *wide*, *groove-c* et *groove-fr* ont été introduits, et leur habilité à détecter certains types de poches à la surface de macromolécules montrée sur des exemples particuliers. De manière générale, *wide*, basé sur une notion d'accessibilité différentiée détecte correctement les poches ouvertes et leur attribue un volume satisfaisant ; de par sa définition cependant, l'arrêt des poches qu'il détecte à la surface des molécules est tributaire de caractéres très locaux de cette surface, comme la présence d'infimes picots saillant hors d'un sillon (La figure 10.16 offre une interprétation imagée du phénomène). A l'inverse, les approches de type *groove*, basées sur des calculs de courbure

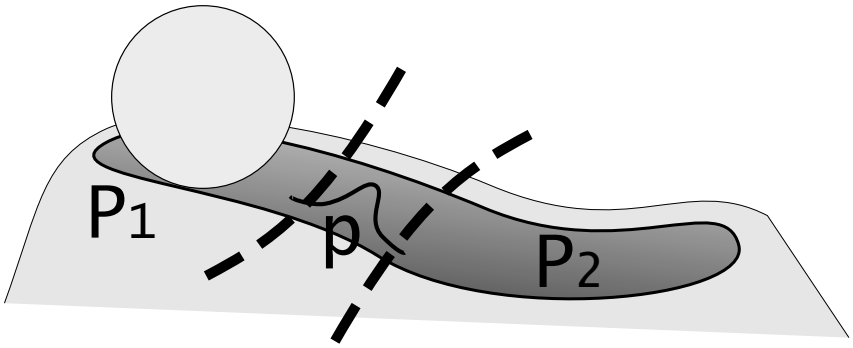


Figure 10.16: Influence de caractères sporadiques sur l'algorithme *wide*. La figure présente un cratère en forme de sillon étroit, pourvu d'un léger picot (p) en son centre qui pointe en dehors de la crevasse. Une sphère d'inaccesibilité roulant en contact avec les bords du cratère sera bloquée au dessus du sillon à cet endroit et *wide* identifiera deux poches, P_1 et P_2, au lieu d'une seule.

locale ne sont pas sensibles à ces considérations sporadiques et révèlent de réèlles tendances de la surface, telles les sillons, souvent mal détectés par les autres approches. Les algorithmes de type *groove* ont néanmoins plus de mal à définir un volume satisfaisant pour ces éléments particuliers ; *groove-c* fournit généralement de meilleures définitions des volumes des poches que *groove-fr*, mais il est souvent moins efficace pour leur détection.

En conclusion, il est important de remarquer qu'aucun de nos algorithmes de détection des poches ne peut être considéré comme "meilleur" que les autres dans l'absolu ; chacun d'entre eux trouvant son application dans un cadre d'étude particulier.

• Instabilité relative aux paramètres

En dehors de *tight*, les algorithmes que nous avons présentés sont dépendants de deux paramètres dont la variation même infime peut parfois entraîner des modifications drastiques dans les résultats (c'est surtout vrai pour les algorithmes de type *groove*). Cette instabilité en fonction

[7] Nous renvoyons le lecteur au chapitre 4 page 39 pour les définitions des poches géométriques employées ici.

des paramètres constitue un réel problème pour une analyse automatisée, et l'étude de procédé pour la mise au point automatique de ces valeurs en fonction de caractéristiques de la molécule (comme la taille, ou l'histogramme des valeurs de courbure locale) constitue une piste d'étude intéressante. Il s'agit là d'un problème complexe, transversal à d'autres domaines et dont le propos dépasse le cadre de notre travail ; il a été évoqué dans la conclusion de la partie 8 (page 114) présentant nos travaux sur la topographie de surface et sera discuté plus avant dans la conclusion générale.

Dans le contexte actuel, les algorithmes de type *groove* paraissent donc comme limités à une analyse avec un opérateur humain, ou pour un traitement par lot sur une famille spécifique de molécules avec étalonnage préalable des paramètres.

• Comparaison de `Pck`(*wide*) et de *APROPOS* [Peters 96]

Une description des poches basée sur une différence de complexes-α a déjà été proposée dans les travaux de Peters *et al.* [Peters 96] ; nous nous en distingons sous divers aspects. Tout d'abord, nous utilisons une triangulation régulière en place d'une triangulation de Delaunay conventionnelle ; de ce fait notre définition polyédrique de la surface moléculaire peut être considérée comme plus exacte car nous prenons en compte la taille des atomes. Ensuite, notre approche se veut intégralement "volumique", non seulement au sens que nous avons donné dans l'état de l'art page 42, mais aussi parce que le résultat final de notre algorithme décrit concrètement un volume physique alloué à la poche ; *APROPOS* (le logiciel implémentant les algorithmes proposés par Peters *et al.*) ne donne que les atomes participant à la poche. Enfin, dans leur approche, la déconnexion des "scories" était effectuée par un calcul de distance des atomes à la surface du complexe-α, calcul potentiellement effectué au travers du corps de la molécule, et induisant parfois la délétion d'atomes de fond de poche. Notre approche se soustrait à ces désagréments d'une certaine manière en restreignant le calcul de distance au travers de l'espace vide, ce que réalise explicitement Coleman [Coleman 06] au prix d'une coûteuse approche discrète basée sur un algorithme de plus court chemin de Dijkstra.

• Limitations des implémentations des algorithmes *groove*, et résolution possibles

◦ **Concernant *groove-c* :** en périphérie de la triangulation de Delaunay les tétraèdres sont particulièrement allongés ; de ce fait, des atomes spatialement très éloignés peuvent s'avérer incidents à une même cellule de Delaunay. Dans l'algorithme *groove-c*, des sommets localement anfractués mais distants les uns des autres peuvent ainsi être sélectionnés comme faisant partie de la même poche. En raison des propriétés de la triangulation de Delaunay, ce genre de cas de figure est peu fréquent, les régions anfractuées étant justement pavées de tétraèdres de petite taille. Des garanties que ce genre de cas de figure n'arrive jamais pourraient être apportées, par exemple en contrôlant la taille des cellules constituant les poches.

◦ **Concernant *groove-fr* :** avec *groove-fr* une poche peut être constituée d'amas de cellules déconnectées. En effet, lors de la phase de "volumisation" on considère les cellules vides dont tous les sommets appartiennent au même "fond de poche" ; rien ne garantit cependant qu'une facette donnée de ce "fond de poche" correspondra effectivement à une cellule en fin de compte. Cet état de fait n'est pas forcément problématique dans la mesure où les amas de cellules ainsi définis, provenant d'un même "fond de poche", sont généralement proches les uns des autres le long de la surface duale, et donc dans l'espace. Si le besoin de poches connexes est essentiel, on pourrait néanmoins adjoindre à peu de frais une phase de déconnection les amas cellulaires obtenus par notre algorithme.

Un autre inconvénient de cet algorithme concerne la présence occasionnelle de "trous" dans les "fonds de poche" : lorsque de petites zones de surface exhibent des valeurs légèrement supérieures à la valeur seuil t au milieu d'un ensemble de sommets dont les valeurs de courbure locale sont inférieures au seuil. Cet état de fait pourrait être circonvenu en détectant ces trous et en ajoutant les facettes qui les composent au "fond de poche" qui les entoure. Une telle opération n'est pas triviale ; il faut posséder un critère pour juger de la pertinence de pics supérieurs au seuil t avant de les inclure au fond de poche. Les diagrammes de persistence [Edelsbrunner 00] semblent une piste adaptée à ce genre de problématique pour le traitement d'un "bruit de fond".

• Décomposition en sous-poches

Une piste d'évolution de l'algorithme *tight* pourrait consister dans la décomposition d'une poche en sous-poches déconnectées les unes des autres par des constrictions internes comme il a été évoqué dans la figure 4.2 de l'introduction (page 41), et comme illustré dans la figure 10.17. Un algorithme alternatif pourrait être implémenté à moindre coût en considérant strictement le

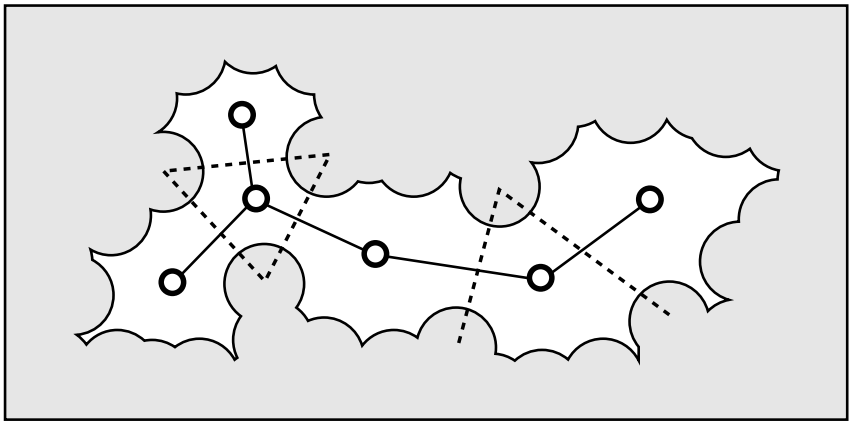


Figure 10.17: Définition des sous poches d'une poche, un exemple en deux dimensions. La cavité en blanc au centre comporte plusieurs constrictions mises en exergue par des segments en traits interrompus. Des pastilles matérialisent le centre des sous poches ainsi définies, et sont connectées par des segments illustrant leur connectivité.

flux discret présenté dans la figure 4.10 de l'introduction (page 4.10).

• Concernant l'espace alloué aux poches

On l'a vu, les algorithmes de type *groove* — et plus particulièrement *groove-fr* — donnent une description assez plate des sillons qu'ils détectent. C'est aussi le cas parfois de *wide*, par exemple dans le cas de la description du sillon d'accomodation du cofacteur d'un récepteur nucléaire, comme on a pu le voir par exemple dans la figure 10.13 page 151 (poche marquée A dans la vignette en bas à gauche).

De fait, la définition générale du volume proposée par nos algorithmes s'apparente à une notion d'opercule tendu au dessus de la surface, et qui ne correspond pas forcément aux attentes d'un utilisateur, souhaitant par exemple matérialiser le volume qu'occuperait un ligand à cet endroit de la macromolécule (voir figure 10.18). Une piste intéressante pour atteindre cet objectif serait de considérer l'union des sphères-α [CCG , Guilloux 09] duales des tétraèdres composant les poches détectées, ou plus empiriquement, de répartir des sphères d'une taille donnée au dessus des atomes de la molécule qui auront été détectés comme faisant partie d'une poche par nos algorithmes.

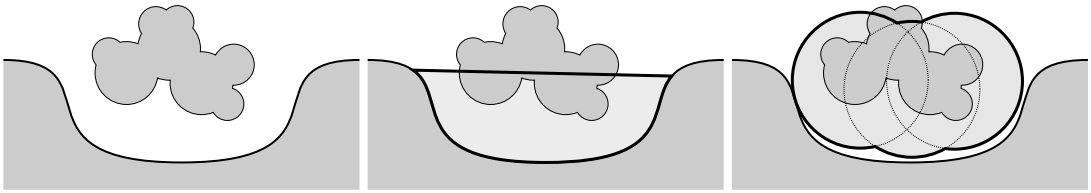


Figure 10.18: Alternatives pour l'attribution d'un volume à une poche, un exemple en deux dimensions. La déclivité à la surface de la molécule n'est pas assez marquée pour enfouir totalement le ligand (à gauche), une définition trop linéaire de la bouche (au centre) ; à droite, l'emploi de sphères réparties au dessus du fond de la déclivité donne une meilleure visualisation de ce volume.

10.4.2 Concernant la caractérisation des poches

• Concernant la caractérisation des facettes des poches

L'utilité avérée du simple coloriage des facettes suivant leur type dans le complexe-α suggère un autre coloriage plus fin révélant le "degré de vide" derrière une facette dans la poche. Une première approche pourrait associer à une facette bordant une poche, la distance à parcourir derrière elle (par exemple en suivant sa normale) dans la forme duale avant de rencontrer à nouveau du vide. On pourrait aussi considérer une mesure de densité (en nombre d'atomes dans une sphère centrée sur la facette), ou pour évaluer plus spécifiquement une maléabilité "derrière la facette", une mesure de densité "dirigée", par exemple dans un cône centré sur la normale à la facette[8].

• Concernant le calcul volumétrique des poches

Les valeurs volumétriques calculées sur le modèle polyédrique fournissent des indices utiles dans le cadre d'une comparaison entre poches, par exemple dans le cadre d'une simulation en dynamique moléculaire, pour suivre l'évolution d'une poche dans le temps. Dans d'autres types d'analyse, on préfère cependant le volume décrit par la surface de Connolly, ou le modèle de Van der Waals. Ces deux valeurs permettent en effet d'inférer plus finement le volume laissé libre dans la poche pour un ligand potentiel. La version actuelle de `Pck` ne permet pas le calcul de ces valeurs, mais des solutions ont été envisagées.

Une première implémentation d'un processus de Monte Carlo s'étant avérée prohibitive en temps de calcul, nous l'avons simplement désactivée de la version packagée.

Dans le contexte d'un calcul volumétrique — une fois la poche convenablement détectée et décrite dans le modèle polyédrique — une approche discrète pourrait aussi s'avérer bien adaptée ; elle se résumerait en le décompte des voxels intérieurs simultanément au polyèdre de la poche et d'un des atomes sis sur un de ses sommets. Les récents travaux concernant la théorie des level-set et leur application à la description discrète des molécules pourraient typiquement être appliqués à cette approche [Can 06].

[8] Une telle mesure pourrait par exemple être utilisée pour visualiser à moindre coût les zones que l'on pourra potentiellement ouvrir pour l'accomodation du groupement chimique d'un ligand particulier.

Conclusions et perspectives

Nos travaux ont porté sur la proposition de solutions théoriques et logicielles adaptées à l'étude de la structure des macromolécules biologiques et de leurs assemblages. Nous avons rappelé les intérêts et les moyens d'une telle analyse dans la première partie, dédiée à l'état de l'art, en nous focalisant essentiellement sur les deux aspects que nous avons traités : *(i)* la caractérisation topographique de la surface moléculaire, et *(ii)* la détection et la caractérisation des poches dans les structures tridimensionnelles. Nos contributions dans les deux problématiques ont été discutées dans les conclusions des chapitres 8 et 10, respectivement, aux pages 113 et 152, et des pistes de développements ultérieurs y ont été proposées ; nous les résumerons ci-après. Nos algorithmes reposent sur les modèles issus de la théorie des formes-α que nous avons présentés aux chapitres 1 (page 10) et 6 (page 59), où nous avons explicité leur adaptation au cadre de l'analyse structurale des macromolécules.

Avec la *courbure locale*, définie au chapitre 8 (page 93), nous avons proposé une nouvelle mesure de l'*incurvation* de la surface moléculaire, propice à une analyse de ses reliefs à différentes échelles, et dépendante uniquement de la forme de la surface. Nous avons visuellement observé la pertinence de cette propriété pour différencier les notions intuitives de "creux" et de "bosses" à la surface des molécules (chapitre 8.4 page 104), et montré que cette caractérisation du paysage à la surface moléculaire était discriminante dans le cadre d'une analyse des zones interagissantes (chapitre 9 page 117). Nous avons encore utilisé cette propriété dans le cadre de la détection des poches dans la structure des protéines (voir les algorithmes de type *groove* au chapitre 10 page 138). Nos algorithmes pour le calcul des valeurs de courbure locale ont été implémentés et sont accessibles librement au travers du logiciel `Lc`, présenté en annexe D page 189.

Au chapitre 10 (page 133), nous avons proposé trois nouveaux algorithmes de détection des poches spécifiquement adaptés à la localisation des *poches ouvertes* et des *sillons* à la surface des macromolécules. Le cadre de la théorie des formes-α nous a permis de délimiter des *bouches* et d'attribuer un volume dans l'espace pour chacune des poches détectées. Ce même modèle nous a en outre offert la possibilité de caractérisations simples concernant la proximité des poches ou la caractérisation de leur forme au travers d'un indice de convexité. Nos trois algorithmes pour la détection et la caractérisation des poches sont accessibles au travers du logiciel `Pck` (présentés à l'annexe D page 189), et une implémentation de l'algorithme de *CASTp* [Binkowski 03b] y a été adjointe pour la détection des *poches refermées* et des *cavités*.

Nos développement reposent en grande partie sur la *surface duale*, une construction que nous avons introduite au chapitre 7 (page 75) et qui facilite le parcours de la surface moléculaire. Cette représentation polyédrique de la molécule et de ses vides a été utilisée pour le calcul des valeurs de courbure locale (chapitre 8 page 97) ainsi que pour la proposition d'une notion de *parcelle de résidus voisins* à la surface d'une protéine ; évaluée dans le contexte de la prédiction des zones interagissantes, notre définition a donné de meilleurs résultats que l'approche habituellement utilisée (chapitre 9 page 119).

L'emploi de modèles de la théorie des formes-α permet, entre autres, des algorithmes simples résultant dans des implémentations suffisamment rapides pour une utilisation dans des applications à grande échelle ; afin de favoriser aussi l'analyse spécifique d'une structure par un utilisateur, nous avons adjoint à nos deux développements des greffons au logiciel de visualisation VMD [Humphrey 96] (voir la présentation des logiciels à l'annexe D page 189).

Perspectives autour de la courbure locale

Un inconvénient de la courbure locale, également présent dans la majorité des travaux analogues et auquel il pourrait être intéressant de circonvenir, réside dans l'existence d'un paramètre[9] qui peut paraître fastidieux à manipuler : la taille du voisinage de lissage, assimilée à une résolution, ou une échelle. Une bonne valeur pour ce paramètre, c'est-à-dire une valeur faisant apparaître les éléments caractéristiques de la macromolécule, dépendra de la taille des éléments qu'on peut ou veut y déceler, mais aussi de la taille de la molécule. Notons toutefois que pour des molécules de formes similaires, les valeurs de courbure locale seront globalement comparables ; ce qui autorise l'analyse d'un ensemble des protéines d'une même famille avec une même valeur de paramètre, par exemple pré-établie sur la base d'une observation sur un des membres de cette famille. La mise au point automatique de ce paramètre constituerait toutefois un gain appréciable. Cette question pourrait par exemple être traitée en calculant la courbure locale à diverses échelles et en observant l'histogramme des valeurs, le nombre d'extrema locaux, leur répartition à la surface, et les écarts de valeurs entre extrema avant de choisir la résolution qui porte le plus d'information ; l'utilisation des diagrammes de persistance permettrait en sus de trier les extrema pertinents[10] [Edelsbrunner 00]. Une approche alternative consisterait à proposer un nouvel indice, non plus topographique, mais de "*pertinence topographique*" qui retiendrait les résultats pertinents à chaque échelle ; dans leurs travaux sur la "saillance", Lee *et al.* proposent par exemple de réaliser une moyenne pondérée par la pertinence observée à chaque échelle [Lee 05] ; et il serait intéressant d'étudier l'adaptation de ce genre de technique à notre contexte.

Perspectives autour de la détection des poches

De nombreuses extensions et fonctionnalités pourraient être apportées à notre logiciel `Pck` pour la détection et la caractérisation des poches. Nous pourrions par exemple considérer l'emploi de filtres simples (sur des critères de taille des poches ou des bouches) permettant à un utilisateur de restreindre la liste des poches renvoyées par le logiciel ; une telle opération permettrait de limiter le nombre de poches sur lesquels les traitements ultérieurs sont effectués et diminuerait encore les temps de calcul.

Une caractérisation des poches en terme de profondeur pourrait aussi s'avérer utile pour une discrimination de celles-ci, et cette valeur pourrait être obtenue à partir des arêtes du diagramme de Voronoï, à la manière dont les tunnels sont caractérisés dans *MOLE* [Petřek 07].

Nous l'avons abordée dans la conclusion du chapitre 10 page 152, la projection sur les facettes d'une poche d'un degré de lacunarité derrière celles-ci pourrait être utile par exemple pour

[9] De manière générale, l'existence de paramètres peut être simultanément considérée comme un avantage parce qu'elle autorise un ajustement à une problématique donnée, et comme un inconvénient lorsqu'elle oblige cet ajustement à chaque problématique, souvent au travers de critères abscons et dont la mise au point requierre l'intervention d'un opérateur humain.

[10] Il est à prévoir que les valeurs de courbure locale présentent de nombreux extrema locaux situés très proches les uns des autres avec des valeurs très semblables. Les diagrammes de persistance constituent une méthode pour trier le "bruit" dans ce genre de données, et extraire les extrema pertinents.

visualiser la flexibilité de la région.

D'autres développements pourraient encore être facilités par la représentation polyédrique des poches offerte par la théorie des formes-α, comme la réalisation d'opérations booléennes (unions et intersections) sur la représentation polyédrique de plusieurs poches ; l'union de poches permettant par exemple d'étudier la forme d'une poche après l'annexion d'une de ses voisines, et l'intersection permettant de comparer entre elles deux poches distinctes.

Au chapitre précédent, nous avons encore évoqué la possibilité d'explorer un autre type de définition de l'espace alloué aux poches en utilisant l'union de sphères réparties au dessus des facettes appartenant à une poche ; une telle approche permettrait par exemple de définir des volumes plus proches de l'"enveloppe d'un ligand", et en tout cas plus volumineux que la définition proposée actuellement pour certains sillons détectés par les algorithmes de type *groove*.

Annexes

Annexe A

Structure des macromolécules biologiques

Cette annexe constitue une introduction succincte à la structure des macromolécules biologiques. Elle est essentiellement destinée aux non biologistes pour lesquels des bases seront peut-être nécessaires à la compréhension des travaux et des objectifs présentés dans ce document.

Cette annexe est architecturé en quatre sections. Les deux premières traitent des notions fondamentales concernant la structure des macromolécules, et plus spécifiquement — dans la seconde section — celle des protéines. Ces notes ont été élaborées essentiellement à partir des ouvrages de Weinmann et Méhul [Weinmann 00] de Horton *et al.* [Horton 06] et de Branden [Branden 98] que le lecteur est encouragé à compulser pour une vision plus détaillée du domaine. La troisième section présente les modèles couramment utilisés pour représenter les (macro)molécules et que nous avons intensivement employés dans les nombreuses illustrations de ce document. Un détail simplifié du mécanisme biologique dans lequel l'ADN est lu et transcrit en ARN fera l'objet de la quatrième et dernière section. Cet exemple met en exergue les éléments principaux qui caractérisent les interactions entre macromolécules et motivent une étude structurale : la reconnaissance moléculaire, le besoin de topographier la surface des macromolécules, et l'importance de l'étude des poches.

A.1 Macromolécules biologiques

Le fonctionnement cellulaire repose sur l'existence d'un grand nombre de molécules de taille et de nature diverses, parmi lesquelles les acides nucléiques (ADN et ARN) et les protéines jouent un rôle particulier et se distinguent notamment par leur grande taille.

Ces macromolécules[1], sont plus spécifiquement des (hétéro)polymères, c'est-à-dire des molécules constituées de motifs similaires répétés. Dans le cas des acides nucléiques, comme dans celui des protéines, ces motifs de base sont constitués d'une chaîne principale (*backbone* ou *main-chain* en anglais) identique pour tous les "maillons", et d'une chaîne latérale (*sidechain*) pouvant varier d'un maillon à l'autre. Si dans le cas de l'ADN ou de l'ARN il existe quatre motifs latéraux possibles, dans le cas des protéines il en existe vingt ; c'est certainement une des raisons qui expliquent la diversité de forme et de fonction de ces dernières.

Les polymères biologiques sont souvent considérés comme des rubans contenant une information séquentielle, un code. Dans le cas de l'ADN, on parle de code génétique pour désigner cette

[1] *i.e* des molécules comprenant un grand nombre d'atomes

succession de "lettres", et dans celui des protéines, de *code protéique*, de *séquence protéique*, ou bien encore de *structure primaire*.

A.2 Structure des protéines

La famille des protéines n'est qu'un des nombreux éléments du paysage cellulaire, mais un élément particulièrement présent et intéressant tant par le nombre de ses représentants (le nombre de protéines présentes dans la cellule, mais aussi le nombre de représentants différents dans la famille) que par leur diversité de formes et de fonctions. Les protéines constituent en effet la seule famille de molécules à remplir des rôles aussi divers : du maintien de la structure cellulaire à la locomotion, en passant par la transduction du signal, les protéines sont actrices dans quasiment toutes les activités de la cellule.

A.2.1 Structure primaire

Les "maillons" constituant les protéines sont des acides aminés ; la figure A.1 présente leur formule chimique générique. La structure d'un acide aminé est architecturée autour d'un atome

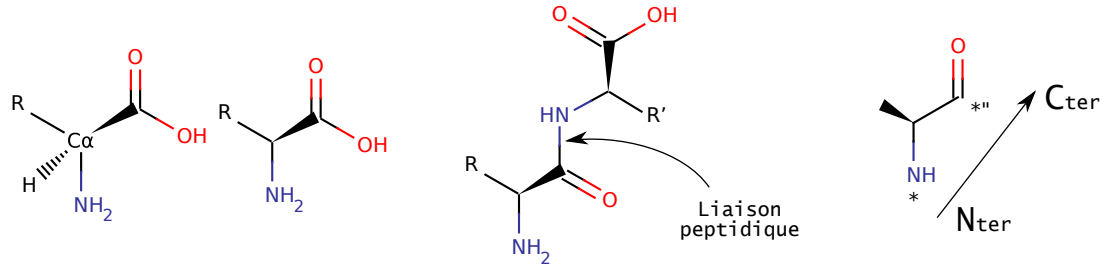
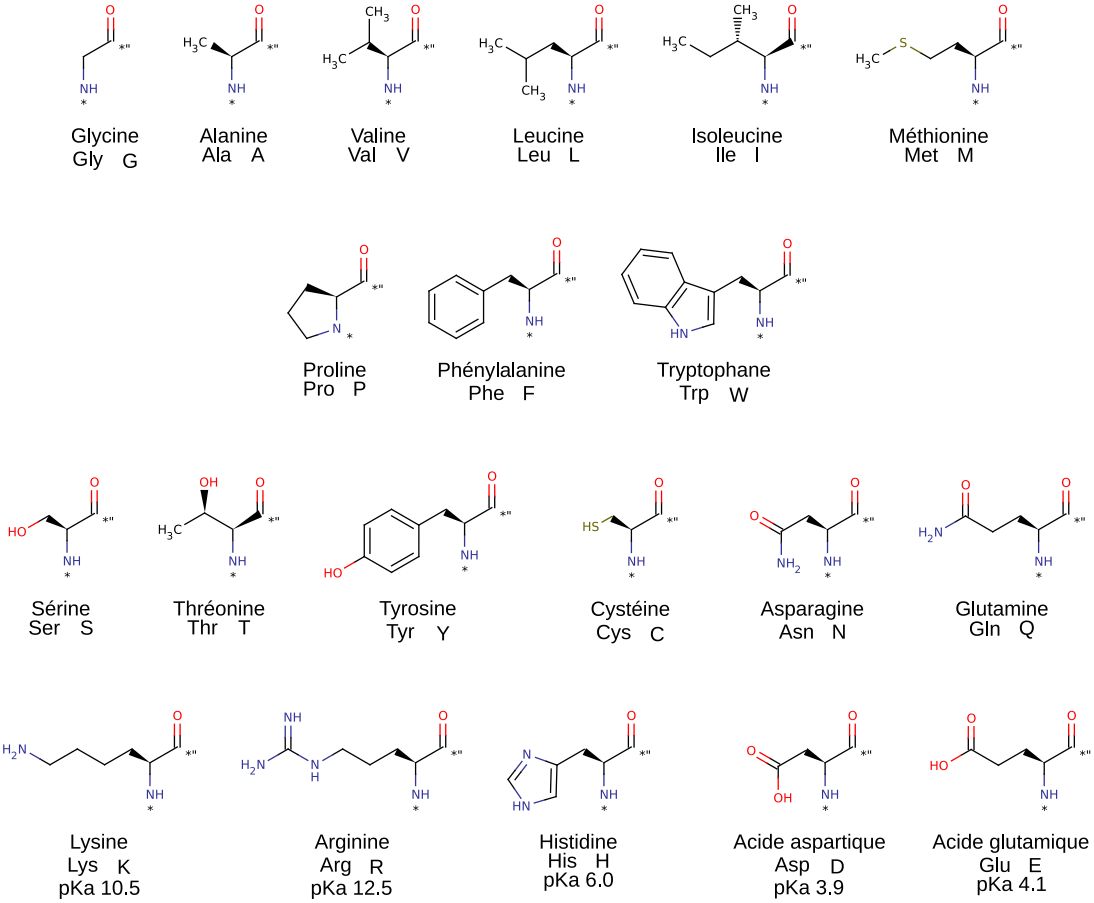


Figure A.1: Acide aminé et liaison peptidique. À gauche, deux représentations — dites de Fischer — d'un acide aminé générique. Dans la première représentation, le carbone α est indiqué explicitement, de même que l'atome d'hydrogène qui lui est associé, de manière à mettre en évidence la géométrie tétraédrique autour de cet atome de carbone. Par convention on ne représente pas explicitement les atomes de carbone et les hydrogènes qui leur sont liés de façon covalente, comme on peut le voir dans la seconde représentation.
Au milieu, deux acides aminés liés par une liaison peptidique. Cette liaison, répétée plusieurs centaines de fois, forme une protéine.
À droite, un acide aminé impliqué dans une chaîne peptidique est orienté par convention de l'atome d'azote de sa chaîne principale vers son atome de carbone.

de carbone désigné comme carbone α et noté C_α. Cet atome est lié de façon covalente à un atome d'hydrogène H, un groupement carboxyle $COOH$, un groupement aminé NH_2, et un *résidu* R composant la chaîne latérale. Les acides aminés s'assemblent au travers d'une liaison covalente pour former une chaîne polypeptidique, ou protéine. On parle aussi parfois de *peptide* pour désigner une protéine de petite taille (une cinquantaine d'acides aminés au maximum). Cette liaison induit une orientation de la structure primaire, qu'on prend par convention de l'atome d'azote (dit N_{ter} pour N terminal) vers l'atome de carbone (C_{ter} pour C terminal).

Une nomemclature associe à chacun des vingt acides aminés une première abbréviation composée de trois lettres et une seconde abbréviation composée d'une seule lettre ; la liste des vingt acides aminés et de leurs abbréviations respectives est produite dans la figure A.2. Dans la même figure on trouvera le diagramme de Venn présentant une classification des acides aminés suivant leurs propriétés physicochimiques.



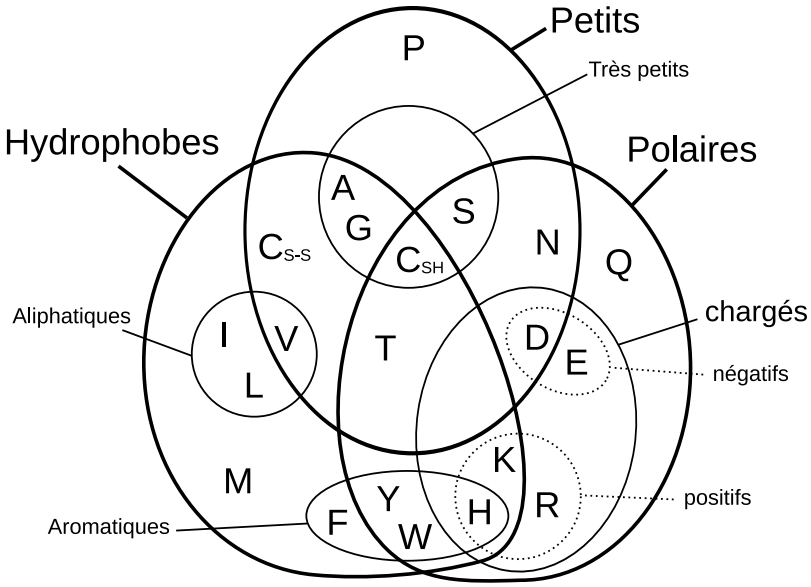


Figure A.2: En haut, la liste des vingt acides aminés dans une représentation de Cram, avec leurs Abbreviation sur trois lettres sur une lettre. En bas, le diagramme de Venn qui offre une classification des acides aminés sur critères de taille et de propriétés physicochimique.

A.2.2 Structure secondaire

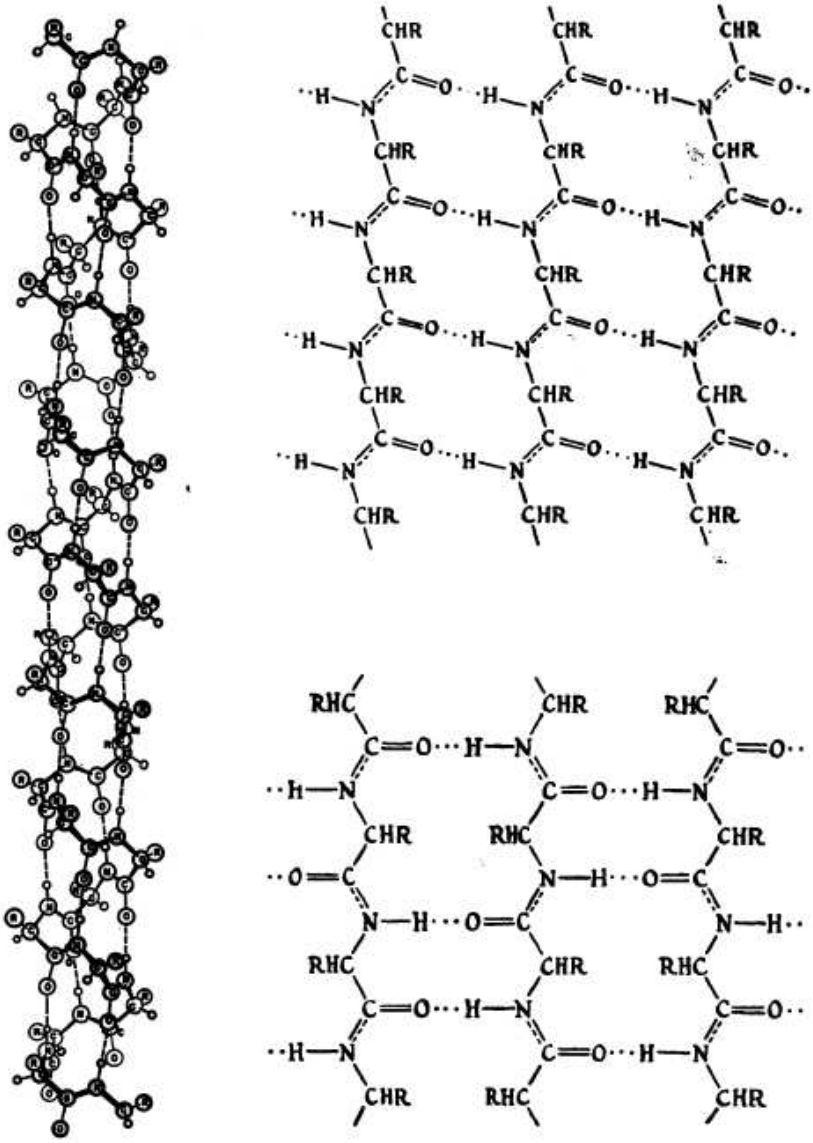


Figure A.3: Éléments de structure secondaire, les traits discontinus symbolisent des liaisons hydrogène. À gauche, l'hélice-α telle que présentée dans l'article de Pauling *et al.* [Pauling 51b]. À droite, les structures des feuillets-β parallèles et antiparallèles (respectivement en haut et en bas) telles que présentées dans l'article de Pauling *et al.* [Pauling 51a].

La nature des acides aminés se succèdant séquentiellement le long de la chaîne primaire induit localement une géométrie de la chaîne peptidique. Cette géométrie est influencée par différents paramètres comme l'occupation stérique [2] des résidus voisins et leurs polarités. Certaines successions d'acides aminés induisent des arrangements géométriques spécifiques et particulièrement stables qu'on retrouve très souvent dans les structures de protéines, c'est ce qu'on appelle la structure secondaire. Les deux motifs les plus courants sont l'hélice-α et le feuillet-β, que l'on peut observer dans la figure A.3. Outre les considérations stériques et polaires locales, ces deux

[2] On peut comprendre cette notion comme un encombrement volumique de chaque atome ou de chaque chaîne.

types de structures doivent leur grande stabilité à la maximisation du nombre de liaisons hydrogène qu'elles favorisent entre les groupements CO et NH de la chaîne principale. Dans une hélice-α ces liaisons hydrogène mettent en jeu les groupements de résidus espacés de quatre acides aminés dans la séquence ; dans un feuillet-β, ce sont des résidus successifs de plusieurs sections continues le long de la séquence (des brins-β) qui sont mis en jeu. On distingue en outre les feuillets parallèles des feuillets antiparallèles suivant que les brins-β concernés sont orientés dans le même sens ou dans le sens inverse.

A.2.3 Structure tertiaire

Le repliement de ces structures secondaires et leur agencement spatial dans une conformation stable définit La structure tertiaire d'une la protéine. La stabilité d'une telle conformation est essentiellement tributaire d'interactions à courte distance (liaisons hydrogène, ponts salins et contacts de Van der Waals) entre les acides aminés composant la protéine eux-mêmes, ainsi que celles qu'ils ont avec les molécules de solvant ou des cofacteurs.

Les protéines de taille conséquente forment généralement des sous-régions fonctionnelles ou *domaines* qui peuvent parfois être répétés au sein de la même protéine, ou qu'on peut retrouver à l'identique dans des protéines différentes. L'aspartyle-ARNt-synthétase présentée dans la figure A.4 et décrite plus en détail dans la section 8.4.5 page 112 présente par exemple trois domaines dont seuls deux ont une fonction connue. L'arrangement des domaines (A) et (B) est partagé entre tous les amino-acyles ARNt-synthétase, à l'inverse du domaine (C) qui est spécifique à l'aspartyle-ARNt-synthétase [Moulinier 97]. Les domaines (A) et (B) ont des fonctions spécifiques bien connues, transverses à tous les amino-acyles ARNt-synthétase : le domaine central (B) contient le site actif dans lequel l'acide aminé ASP est adjoint à l'ARNt, et le domaine (A) est responsable de la reconnaissance du codon correspondant à l'acide aminé. Aucune fonction n'est encore connue pour l'"extra-domaine" (C).

Un domaine ne correspond pas forcément à une sous-séquence connexe d'acides aminés dans la structure primaire, par exemple le domaine central de l'aspartyl-ARNt-synthétase est composé d'acides aminés de la partie N_{ter} (en rouge) et de la partie C_{ter} (en blanc) de la molécule.

A.2.4 Structure quaternaire

On parle de structure quaternaire pour désigner l'agencement des intervenants d'un assemblage protéique ; chaque protéine impliquée dans un tel *complexe* est désignée comme une *sous-unité* de l'assemblage. De nombreuses protéines ne sont fonctionnelles qu'en collaboration avec d'autres protéines ; les complexes qu'elles forment peuvent être transitoires ou stables, et impliquer un nombre élevé d'intervenants. Les filaments d'actine qui composent les muscles, par exemple, forment de longues chaînes quaternaires impliquant des interactions stables entre les sous-unités [Weinmann 00]. A l'inverse, d'autres assemblages comptent un nombre d'intervenants moindre, et leur fonction peut reposer sur la dissociation de l'un d'eux, comme par exemple dans le cas des membres de la famille des protéines-G, qui constituent un hétérotrimère de trois chaînes α, β et γ dont la fonction dans la transduction du signal repose sur la dissociation de la sous unité α [Digby 06, K 08].

A.3 Modélisation et visualisation des macromolécules biologiques

Différentes approches existent pour représenter les molécules ; elles peuvent dépendre de la taille des objets étudiés, de leur nature ou de ce que l'on souhaite y observer.

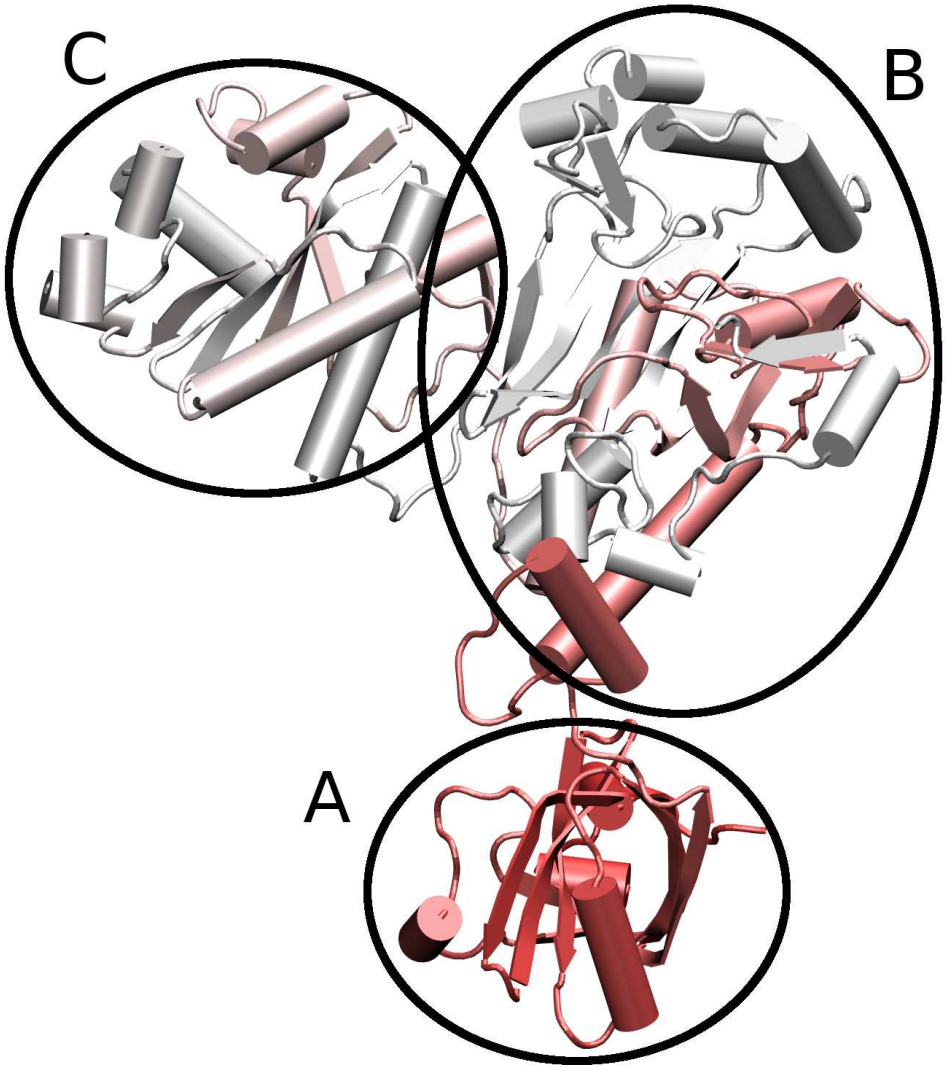


Figure A.4: Structure de l'aspartyl-ARNt-synthétase (1IL2 chaîne A) dans une représentation *cartoon*. Le dégradé de couleurs correspond à l'indexation des acides aminés se succédant dans la chaîne primaire (N_{ter} en rouge et C_{ter} en blanc). L'aspartyle-ARNt-synthétase possède trois domaines (A), (B) et (C) qu'on peut clairement distinguer sur des critères géométriques.

En première instance, les représentations de Fisher et de Cram (figure A.5) sont couramment employées en chimie. Elle mettent en évidence la composition atomique des molécules ainsi que leurs liaisons covalentes. Elle permettent aussi de se faire une idée de la localisation spatiale des atomes, de l'espace qu'ils occupent et de la place laissée vacante par ces atomes, donnant des indices sur la flexibilité des divers groupements.

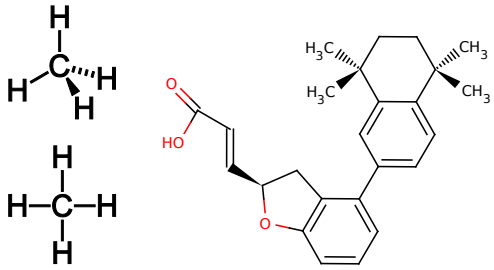


Figure A.5: Représentation des molécules en chimie. À gauche, une molécule de méthane dans sa représentation de Cram (en haut) et dans sa représentation de Fisher (en bas). À droite, la représentation de Cram d'un ligand synthétique du récepteur X à l'acide rétinoïque RXRα,
(S)-(2E)-3[4-(5,5,8,8-tetramethyl-5,6,7,8-tetrahydro-2-naphtalenyl)tetrahydro-1-benzofuran-2-YL]-2-propenoic acid
portant le code (S)-46a,b dans la publication où sa synthèse est décrite et où il est présenté en complexe avec RXRα dans la structure 1RDT [Haffner 04].

Lorsque les molécules comprennent un nombre élevé d'atomes (typiquement dans le cas des macromolécules biologiques), cees schémas de structure ne sont plus adaptés et on doit avoir recours à des représentations spatiales. Les premiers modèles mis au point étaient constitués de bois, de métal voire de plastique (figure A.6). Dans le modèle CPK[3] (figures de gauche), les atomes sont représentés par des (morceaux de) sphères ; leurs liaisons de covalence sont matérialisées par des mécanismes permettant de maintenir les sphères entre elles. Ce modèle permet une représentation spatiale de la molécule, et offre une approche intuitive de l'encombrement stérique[4] de chaque atome la composant. Les modèles du type de celui développé par Dreiding mettent eux l'accent sur les liaisons covalentes en les représentant par de petits segments mécaniquement joints aux centres des atomes (figure de droite).

De manière générale, ces modèles permettent d'appréhender la forme et la flexibilité de molécules de toutes tailles et de toutes complexités, mais leur construction peut s'avérer longue et onéreuse. L'émergence et la démocratisation de l'informatique ont favorisé le développement et l'utilisation de représentations moléculaires informatiques équivalentes aux modèles CPK et Dreiding (voir figure A.7). Utilisées en conjonction avec des méthodes de visualisation stéréographiques, elles permettent une appréhension tridimensionnelle de la structure des molécules étudiées. D'autres modèles ont aussi été développés pour mettre en relief les spécificités structurales des macromolécules biologiques et permettre de meilleures descriptions et appréhensions visuelles de leurs agencements spatiaux (on parle de *topologie* moléculaire). Les représentations les plus utilisées sont le tracé du squelette du polymère (tracé-α), ainsi que les représentations schématisées (*cartoon*) mettant en évidence les spécificités de la molécule, telles les éléments de structure secondaire dans les protéines (figure A.8) ou l'empilement des bases azotées (π-*stacking*) dans le cas des acides nucléiques (figure A.9).

Dans le cas des protéines, les représentations "bâton" et "boule et bâton" permettent d'étudier

[3] L'acronyme provient des initiales des trois chercheurs (Robert B. Corey, Linus Pauling et W.L. Koltun) qui ont proposé le premier modèle physique d'assemblage à base de boules pour représenter les molécules [Corey 53].

[4] Chaque atome occupe une certaine place dans l'espace qui ne peut être pénétrée par un autre atome ; on parle

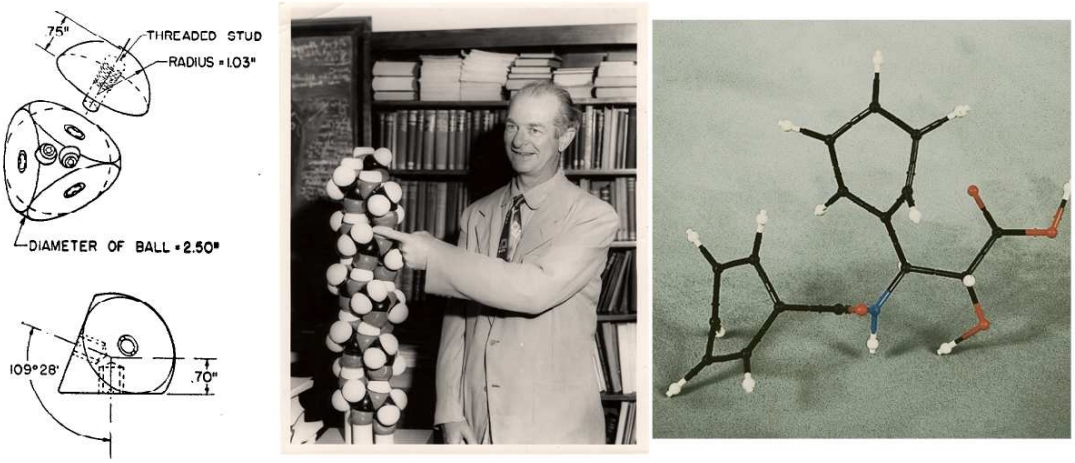


Figure A.6: Modèles moléculaires physiques. À gauche, deux schémas extraits de la description du modèle CPK tel que proposé par Corey et Pauling en 1953 [Corey 53]. Au centre, Linus Pauling à côté d'un modèle moléculaire CPK (image mise à disposition sur le site de la NLM, *National Library of Medicine*). À droite, un modèle moléculaire de Dreiding en plastique.

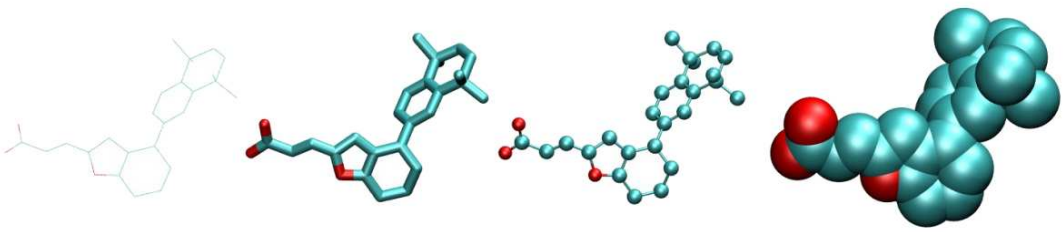


Figure A.7: Modèles moléculaires en visualisation informatique pour le ligand présenté dans la figure A.5. De gauche à droite, sa représentation en fil de fer (*lines*), bâton (*licorice*), boules et bâtons et Van der Waals (selon la nomenclature utilisée dans le logiciel de visualisation VMD [Humphrey 96]).

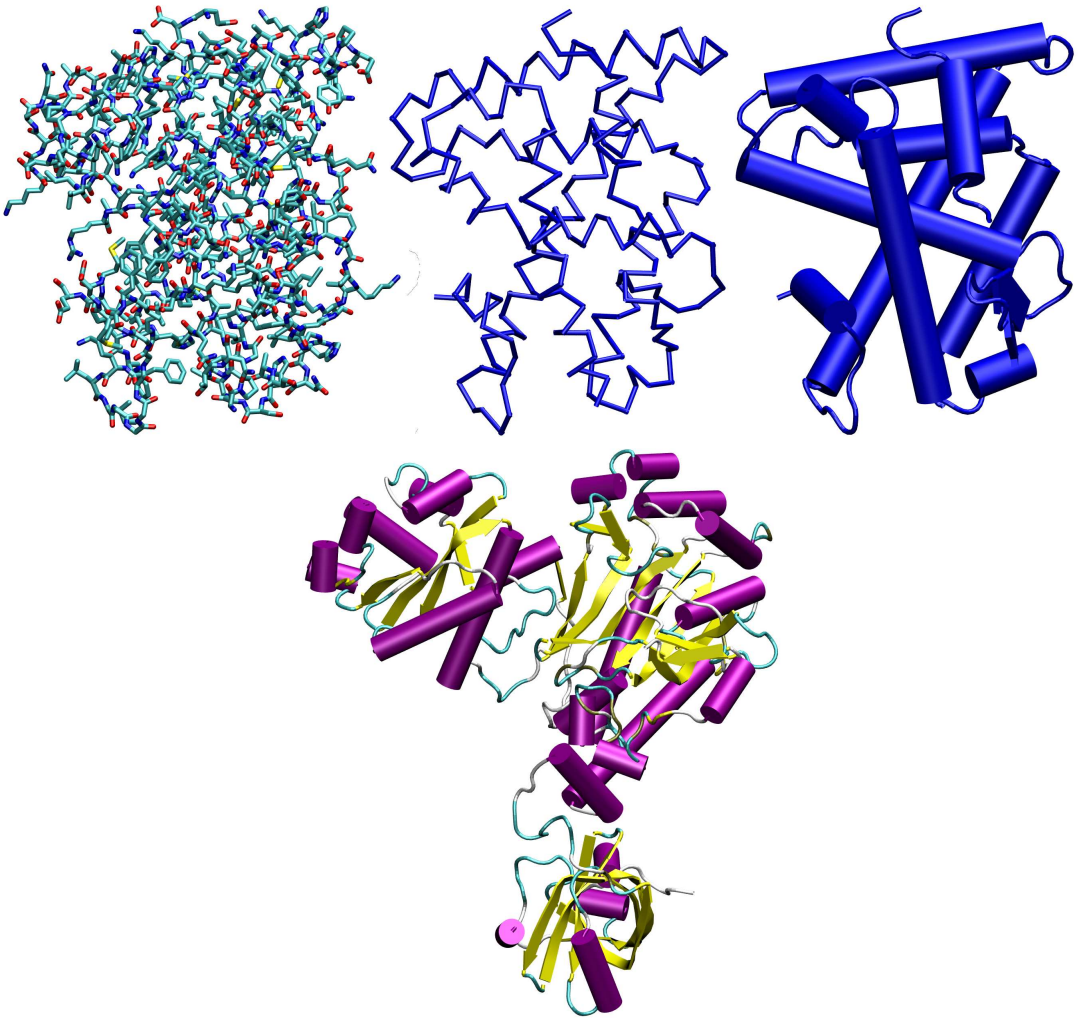


Figure A.8: Modèles moléculaires usuels pour la visualisation informatique des protéines. En haut, la structure du domaine de liaison au ligand d'un récepteur à l'acide rétinoïque RXR-α (1RDT). De gauche à droite, la représentation "bâton", le tracé α et la représentation *cartoon*. En bas, la représentation *cartoon* d'un aspartyl-ARNt-synthétase (1IL2). Les couleurs correspondent aux éléments de structure : violet pour les hélices et jaune pour les brins-β.

localement la stéréochimie de la molécule, mais noient l'œil sous les détails lorsqu'il s'agit de regarder globalement la structure ou la forme de la molécule. Le tracé α permet de comprendre l'enchaînement spatial des acides aminés et d'appréhender le repliement spatial de la protéine ; les éléments de la structure secondaire sont aussi plus apparents. Avec la représentation *cartoon*, les éléments de la structure secondaire sont explicitement représentés par des tubes (pour les hélices) et des flèches plates (pour les brins-β). Cette représentation est particulièrement bien adaptée pour appréhender la topologie de la protéine. Comme on peut l'observer dans la figure, le domaine de liaison au ligand d'un récepteur nucléaire est très structuré et constitué quasi exclusivement d'hélices-α, à l'inverse des aminoacyle-ARNt-synthétases qui se comprennent un nombre conséquent de feuillets.

La figure A.9 montre un dodécamère d'ADN (une double hélice d'ADN de douze paires de bases) dans les trois mêmes représentations. La représentation *cartoon* est ici adaptée pour mettre en évidence les spécificités structurales de l'ADN que sont l'appariement et l'empilement des paires de bases.

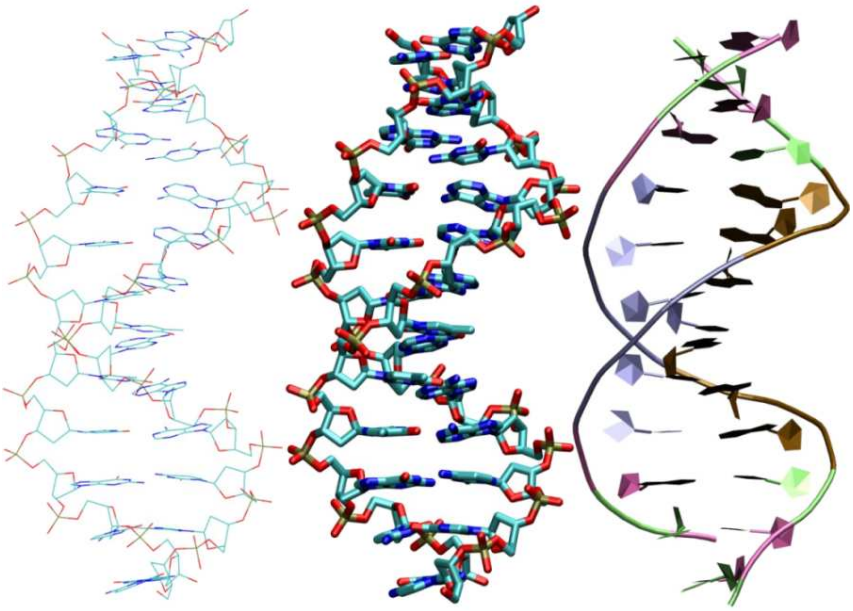


Figure A.9: Modèles moléculaires usuels pour la visualisation informatique d'un duplex d'ADN (1D65). De gauche à droite, les représentations en fil de fer, bâton et *cartoon*.

A.4 Modulation de la transcription par les récepteurs nucléaires, un exemple simplifié de mécanisme biologique

La transcription de l'ADN est le mécanisme dans lequel une petite partie de l'ADN est "lue" et "transcrite" en un brin d'ARN.

La représentation simplifiée de la figure A.10 présente les acteurs principaux de ce mécanisme et leurs interactions. L'ADN (A) est un long filament moléculaire composé d'une succession de maillons de quatre types différents représentés par les lettres A, T, G et C. L'information génétique est entièrement codée dans l'ADN par la succession de ces motifs. Afin que ce message puisse s'exprimer, une première phase consiste en la transcription de certaines sous-parties de l'ADN (les

d'encombrement stérique pour désigner ce volume.

gènes) en un code similaire porté par des brins d'ARN (B). Certains ARN sont fonctionnels en l'état, d'autres portent le message génétique vers une unité de décodage (le ribosome) où il sera interprété et traduit en protéine.

La transcription est régulée par un grand nombre d'acteurs moléculaires, dont les récepteurs nucléaires (C). Cette famille de protéines possède une structure architecturée en cinq à six domaines conservés dont on trouvera une description succincte au chapitre 8.4.2 page 105, ainsi qu'une étude plus détaillée dans la littérature [Laudet 01, Gronemeyer 04]. Ces protéines possèdent un domaine de liaison à l'ADN (DBD, pour *DNA binding Domain*) qui reconnaît et se fixe sur des motifs spécifiques de l'ADN. Certains membres de cette famille sont capables d'agir seuls, d'autres doivent s'assembler (dimériser) avec un autre récepteur nucléaire pour remplir leur fonction ; on parle alors d'homodimérisation ou d'hétérodimérisation suivant que cette dimérisation met en jeu deux macromolécules similaires ou différentes. L'activité des récepteurs nucléaires est très généralement modulée par une petite molécule (ligand, hormone, ...) (D) venant se fixer dans une poche interne à l'un de ses autres domaines, le LBD (pour *ligand binding domain*). L'arrimage de ce ligand engendre une modification structurale de la protéine (une variation de sa forme et de sa stabilité) qui favorise ou prohibe le recrutement d'autres protéines de la machinerie transcriptionnelle (E). La transcription de l'ADN en ARN pour un gène donné sera ainsi réprimée ou favorisée en fonction de l'absence ou de la présence d'un ligand, et de la nature de celui-ci. Les ligands dits activateurs (agonistes) favorisent l'arrimage de la machinerie transcriptionnelle (D), un complexe composé d'un grand nombre de protéines et dont l'activité permet la transcription, notamment par l'ouverture de la double hélice d'ADN. Les ligands dits inhibiteurs (antagonistes, ou répresseurs) répriment la transcription en gênant l'arrimage de la machinerie transcriptionnelle, en l'empêchant complètement, ou en favorisant l'arrimage d'un autre complexe, inhibiteur.

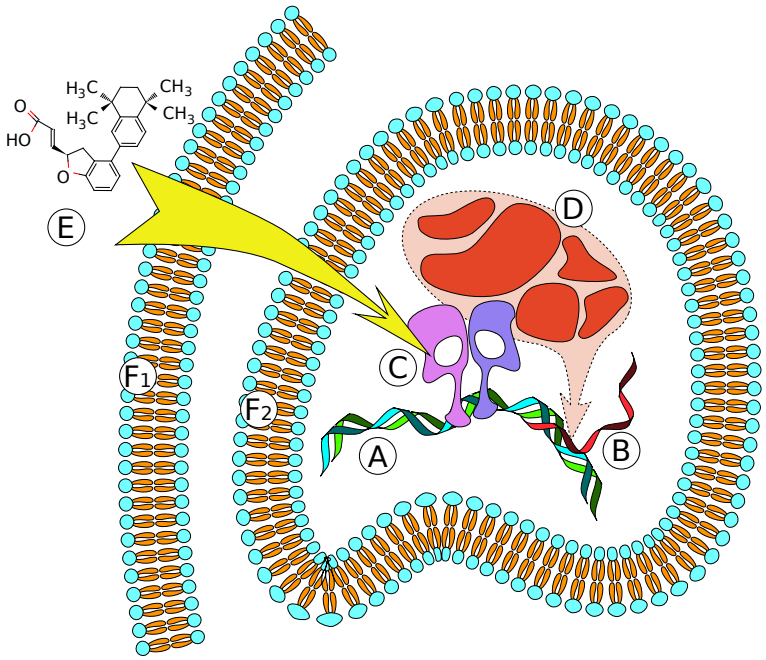


Figure A.10: Transcription de l'ADN, un schéma simplifié : un message extérieur à la cellule est relayé par le biais d'un ligand (E) traversant la membrane cellulaire (F1) et la membrane nucléaire (F2). L'information est transmise lorsqu'en se fixant dans la poche d'un récepteur nucléaire (C) le ligand en modifie la structure. L'arrimage de nouvelles protéines à la surface du récepteur nucléaire est ainsi modulé par la nature du ligand. Les ligands activateurs favorisent le recrutement de la machinerie transcriptionnelle (D), résultant dans la transcription de l'ADN (A) en ARN (B).

En résumé, dans le mécanisme présenté ici, une activité biologique intracellulaire (la transcription) est modulée par une information extérieure à la cellule, portée par le ligand. La *reconnaissance* spécifique de ce ligand par un récepteur nucléaire a lieu dans une *poche interne* à l'un des domaines de cette protéine. Cette liaison induit une modification de la structure du récepteur, menant à la création d'un *sillon* en surface de la protéine. L'action du ligand sur la transcription de l'ADN dépend de la *reconnaissance* de ce sillon par des co-activateurs ou des co-répresseurs de la transcription. Qui plus est, l'activité des récepteurs nucléaires est tributaire de leur habilité à reconnaître un site spécifique à la surface de l'ADN, et pour certains d'entre-eux de leur association spécifique avec un récepteur nucléaire partenaire. La caractérisation de la topographie de ces macromolécules s'avère ainsi une étape quasi-obligée pour déterminer une complémentarité de forme.

De manière plus générale, cet exemple permet de mettre en exergue le lien fondamental que joue la structure d'une protéine sur sa fonction, et l'importance d'une étude structurale pour déterminer ou comprendre son rôle.

Annexe B

Propriétés du diagramme de Voronoï et de la triangulation de Delaunay

Cette annexe regroupe des définitions, et des propriétés concernant la triangulation de Delaunay et — dans une moindre mesure — le diagramme de Voronoï. L'objectif de cette annexe est de présenter de manière plus détaillée certains aspects de ces constructions qui méritent une attention particulière pour la compréhension de leur utilisation comme modèles en bioinformatique structurale.

Ces notes ont été synthétisées à partir du livre de Boissonnat et Yvinec [Boissonnat 95] ainsi que de quelques articles de référence, tels que les revues d'Aurenhammer sur les diagrammes de Voronoï [Aurenhammer 91] (et celle plus spécifique sur les diagrammes de puissance [Aurenhammer 87]) et l'article de H. Edelsbrunner [Edelsbrunner 96] présentant une construction incrémentale de la triangulation régulière en dimension quelconque. Cette annexe ne se substitue pas à un ouvrage de référence et n'a d'autre vocation que de constituer une introduction intuitive à ces modèles issus de la géométrie algorithmique, à leur emploi et à leurs limites potentielles dans le cadre de la géométrie algorithmique.

Une première section reprend les définitions générales du diagramme de Voronoï et de la triangulation de Delaunay dans le cas non pondéré, et explicite la dualité de ces deux constructions. Les analogues pondérés de ces modèles sont présentés dans la seconde section ; ils constituent des généralisations du diagramme de Voronoï et de la triangulation de Delaunay pour un ensemble de boules (d'atomes, ou de boules-atomes comme on pourra les appeler). En particulier, la triangulation régulière que nous y présentons est la triangulation de Delaunay que nous avons utilisée dans nos travaux. Les trois dernières parties présentent des propriétés de la triangulation de Delaunay : la propriété de la sphère vide utilisée pour construire la triangulation de Delaunay, la définition d'un voisinage directionnel intuitif avec les arêtes de cette triangulation, et la disparition potentielle d'atomes dans la triangulation régulière.

B.1 Dualité du diagramme de Voronoï et de la triangulation de Delaunay

Comme esquissé dans l'introduction, le diagramme de Voronoï $\mathcal{V}$ d'un ensemble de points $P = \{p_i\}_i$ partitionne l'espace en autant de polyèdres convexes $\mathcal{V}_i$, où $\mathcal{V}_i$, la cellule de Voronoï associée au point p_i, est définie comme l'ensemble des points de l'espace qui sont plus proches de p_i que de tous les autres points de P.

$$V_i = \{x \in \mathbb{R}^3 \ \mid \ d(x, p_i) \leq d(x, p_j) \ \forall \ j \neq i\}$$

En conséquence de cette définition, la cellule de Voronoï du point p_i se définit comme l'intersection de tous les demi-espaces Π_{ij} définis comme plus proches de p_i que de p_j (voir l'illustration en deux dimensions de la figure B.1). La triangulation de Delaunay $\mathcal{D}$ de P peut être définie par

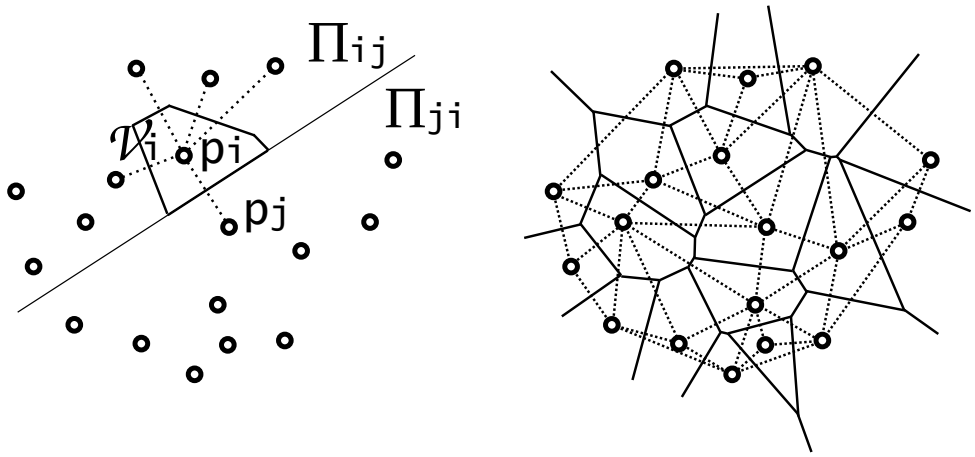


Figure B.1: Dualité du diagramme de Voronoï et de la triangulation de Delaunay, un exemple en deux dimensions dans le cas non pondéré, c'est-à-dire pour un nuage de points (matérialisés par des points blancs). À gauche, la cellule de Voronoï $\mathcal{V}_i$ de p_i est mise en évidence par des arêtes épaisses. La médiatrice des points p_i et p_j est indiquée en trait fin. À droite, les arêtes du diagramme de Voronoï (en traits continus) et leurs arêtes duales dans la triangulation de Delaunay (en traits pointillés).

dualité de $\mathcal{V}$ en observant les intersections des cellules de Voronoï. Ainsi, les sommets de $\mathcal{D}$ sont duaux des $\mathcal{V}_i$ et correspondent très exactement aux p_i. L'intersection de deux cellules $\mathcal{V}_i$ et $\mathcal{V}_j$, lorsqu'elle n'est pas vide, appartient au plan médian des points p_i et p_j ; son dual dans la triangulation de Delaunay est une arête joignant les sommets associés aux points p_i et p_j. L'intersection de trois cellules $\mathcal{V}_i$, $\mathcal{V}_j$ et $\mathcal{V}_k$, lorsqu'elle n'est pas vide, est portée par une droite, et son dual dans la triangulation de Delaunay est une facette triangulaire joignant les sommets associés à p_i, p_j et p_k. Enfin, l'intersection de quatre cellules $\mathcal{V}_i$, $\mathcal{V}_j$, $\mathcal{V}_k$ et $\mathcal{V}_l$, lorsqu'elle n'est pas vide, est un sommet de Voronoï, et son dual dans la triangulation de Delaunay est un tétraèdre joignant les sommets associés aux points p_i, p_j, p_k et p_l.

La dualité de ces deux constructions géométriques peut être observée en deux dimensions dans la figure B.1 ; elle est encore rappelée dans les tables B.1 et B.2 exposant la dualité entre les éléments des deux constructions, respectivement en deux et trois dimensions.

Diagramme de Voronoï		Triangulation de Delaunay
cellule	$\longleftrightarrow$	sommet
arête	$\longleftrightarrow$	arête
sommet	$\longleftrightarrow$	triangle

Tableau B.1: Dualités entre les éléments du diagramme de Voronoï et ceux de la triangulation de Delaunay en deux dimensions.

B.2 Triangulation régulière (Delaunay pondéré)

Pour une utilisation dans un cadre moléculaire, on souhaite généralement prendre en compte la taille des atomes, autrement dit le rayon des boules qui les représentent. Dans ce cas, dit pondéré, on utilise généralement un diagramme de Voronoï particulier, basé non pas sur une

notion de distance minimale aux centres atomiques, mais sur une notion de *puissance* minimale aux boules-atomes de la molécule. La puissance $\pi_{b_i}(x)$ d'un point x de $\mathbb{R}^3$ à une boule b_i est donnée par

$$\pi_{b_i}(x) = d(x, a_i)^2 - r_i^2$$

Comme dans le cas du diagramme de Voronoï non pondéré exposé dans la section précédente, le lieu des points à égale puissance de deux boules est un plan perpendiculaire au segment reliant les centres des deux boules (ou une droite, en deux dimensions) ; on l'appelle *plan radical* (respectivement *droite radicale*). Le lieu des points à égale puissance de trois boules est un axe (appelé axe radical) résultant de l'intersection de deux plans. La figure B.2 donne une illustration en deux dimensions du placement du plan radical de deux boules a et b. Lorsque les boules ont un rayon identique, le plan radical coïncide avec la médiatrice (colonne de gauche, figure du haut) ; si les rayons sont différents, le plan radical se déplace vers l'atome de plus faible rayon (figure du bas). La colonne de droite illustre le déplacement du plan radical en fonction du déplacement de la boule b. Lorsque les deux boules sont disjointes, le plan radical est situé entre elles ; lorsqu'elles s'intersectent sans que l'une soit incluse dans l'autre, le plan radical est situé à l'intersection des sphères bordant les boules. Lorsqu'une boule est totalement incluse dans l'autre, le plan radical s'éloigne rapidement "vers l'infini" ; et lorsque les centres atomiques sont superposés (cas non illustré dans les figures), le plan radical est inexistant, et la cellule de Voronoï de la plus petite des boules est vide. Dans le diagramme pondéré, la cellule de Voronoï d'une boule peut ainsi parfois ne contenir ni le centre de la boule, ni même la boule dans sa totalité. Il arrive même que certaines boules n'aient pas de cellule, on les appelle points redondants ; cette spécificité du diagramme pondéré est traitée un peu plus loin en page 184.

Pour les distinguer de leurs analogues non pondérés, ces diagrammes particuliers, ainsi que les triangulations qui leur sont associées par dualité, sont couramment appelés *diagrammes de puissances* et *triangulations régulières*. Un exemple en deux dimensions est proposé dans la figure B.3, où l'on peut observer un ensemble de quatre boules centrées sur quatre mêmes points et dont seuls les rayons varient d'une figure à l'autre. On peut y observer en particulier que les diagrammes de puissance et la triangulation régulière d'un ensemble de boules toutes de rayon nul sont respectivement équivalents aux diagrammes de Voronoï et à la triangulation de Delaunay des centres de ces boules. On observera en outre que la modification des rayons des boules influe systématiquement sur le diagramme de Voronoï, et que son incidence sur la triangulation de Delaunay est plus diffuse, car cette dernière ne dépend pas à proprement parler du diagramme de Voronoï mais de sa combinatoire (plus particulièrement des intersections observées entre ses cellules). La figure B.4 montre superposés les diagrammes de Voronoï pondéré et non pondéré d'une molécule. En raison des faibles variations de rayons entre les boules-atome, les différences sont assez infimes. On remarquera néanmoins que les deux atomes de taille conséquente a et b exhibent des cellules de Voronoï plus larges dans le diagramme de puissance, à l'inverse de l'atome c, plus petit, dont la cellule s'est amenuisée. En tout état de cause, dans cet exemple, la triangulation de Delaunay sera la même qu'on considère ou non les rayons atomiques. Ce n'est

Diagramme de Voronoï		Triangulation de Delaunay
cellule	$\longleftrightarrow$	sommet
facette	$\longleftrightarrow$	arête
arête	$\longleftrightarrow$	facette (triangulaire)
sommet	$\longleftrightarrow$	tétraèdre

Tableau B.2: Dualités entre les éléments du diagramme de Voronoï et ceux de la triangulation de Delaunay en trois dimensions.

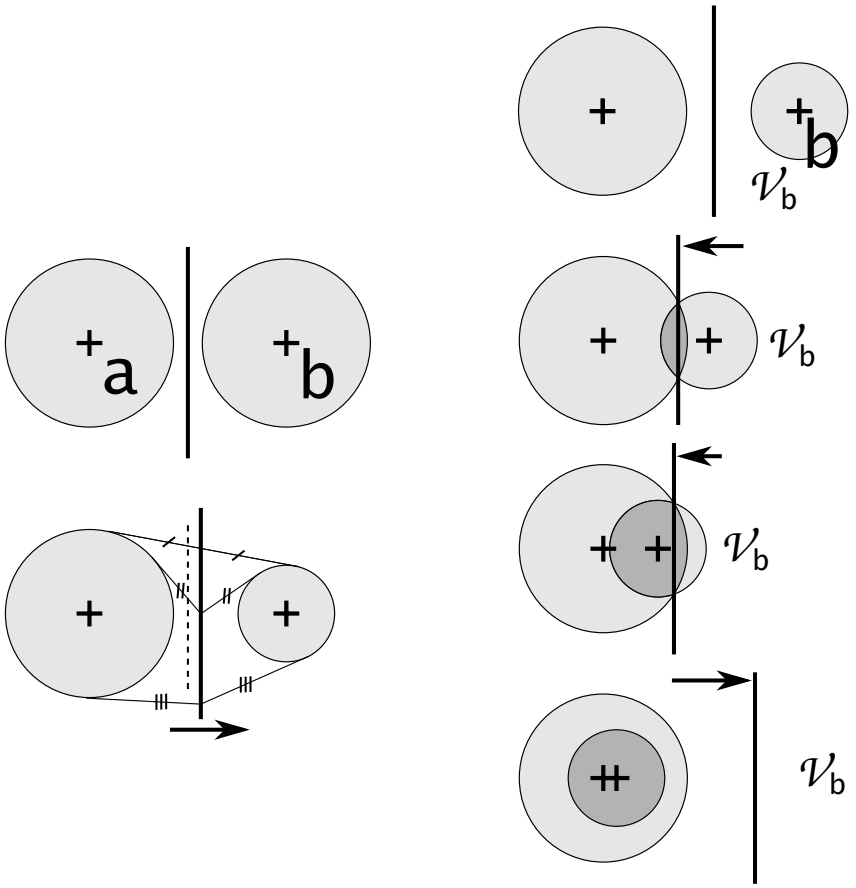


Figure B.2: Placement de la droite radicale à deux boules a et b dans le plan. Dans la colonne de gauche, les centres des boules sont situés à égale distance dans les deux cas de figure, mais le rayon de b varie. Dans la colonne de droite, les rayons des boules ne changent pas, mais la boule b se rapproche de a. Les flèches symbolisent le déplacement de la droite radicale depuis la configuration précédente (dans la figure du dessus) ; et la cellule de Voronoï $\mathcal{V}_b$ de la boule b est systématiquement indiquée.

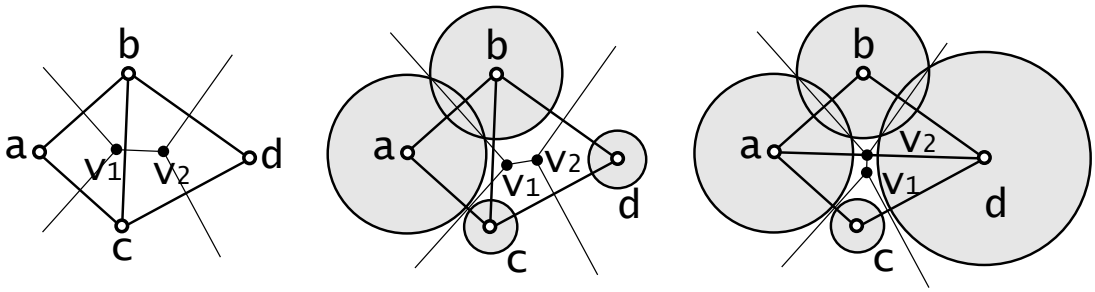


Figure B.3: Trois triangulations régulières d'un ensemble de boules centrées sur les mêmes points a, b, c et d. Le diagramme de Voronoï est apparent en lignes fines, la triangulation en lignes épaisses. À gauche, la triangulation régulière d'un ensemble de boules, toutes de rayon nul. Dans les autres figures, les boules sont toutes affublées de rayons positifs, et représentées en gris.

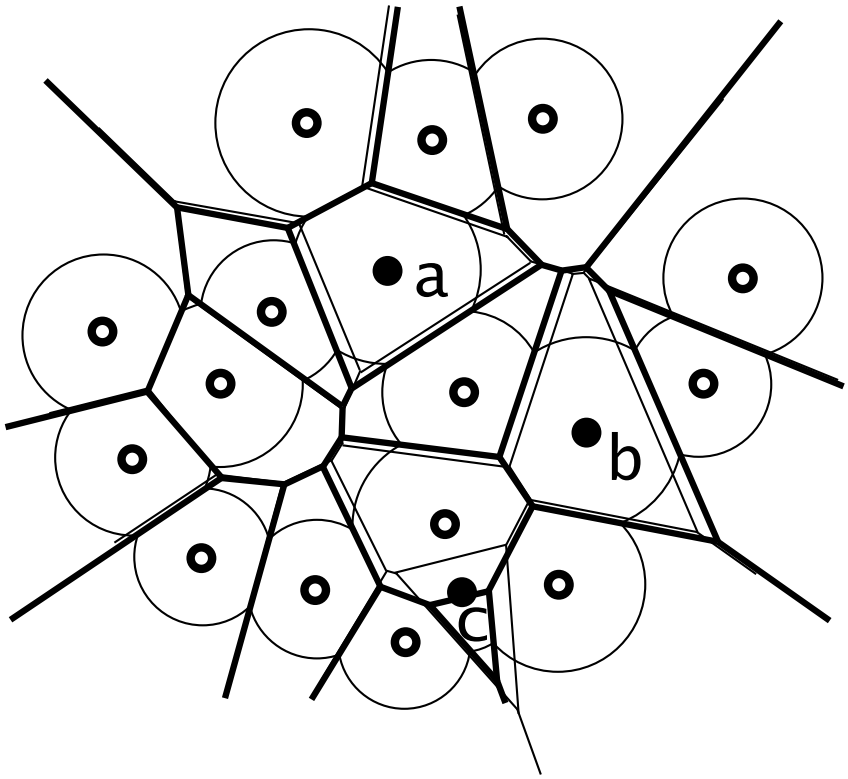


Figure B.4: Superposition du diagramme de Voronoï pondéré (en lignes épaisses) et non pondéré (en lignes fines) d'une molécule (suggérée par le contour de l'union de ses boule-atomes).

très généralement pas le cas pour des molécules de taille conséquente en trois dimensions.

On remarquera enfin que la définition des plans radicaux passant par les intersections du bord de deux atomes permet de répartir intuitivement le volume de deux atomes lorsque ceux-ci s'intersectent. Cette propriété a été remarquée et utilisée très tôt par Gellatly et Finney [Gellatly 82]. Nous l'avons implicitement utilisée au chapitre 6.1.1 page 60 pour définir le diagramme en remplissage de forme.

B.3 Propriété de la sphère vide

Dans le cas non pondéré, on dit qu'une triangulation *a la propriété de la sphère vide* [Delaunay 34] si, quel que soit le tétraèdre t considéré dans cette triangulation (ou en deux dimensions le triangle), la sphère circonscrite aux sommets de t ne contient aucun autre point de la triangulation ni sur son pourtour, ni en son intérieur. Cette propriété peut être observée dans l'exemple en deux dimensions de la figure B.5 A. Dans la triangulation de gauche les cercles $\mathcal{C}_1$ et $\mathcal{C}_2$ sont respectivement vides des points d et a. À l'inverse, la triangulation de droite ne possède pas la propriété de la sphère vide.

On peut montrer que la triangulation de Delaunay d'un ensemble de points possède la propriété de la sphère vide, et que parmi toutes les triangulations d'un même ensemble de points, c'est en fait la seule à posséder cette propriété. Comme illustré en deux dimensions dans la figure B.5 C, ce résultat découle directement de la dualité avec le diagramme de Voronoï. En effet, les sommets du diagramme de Voronoï sont par définition situés à égale distance de quatre sites (ou de trois en dimension deux), et plus proches de ces quatre sites que de tous les autres ; en

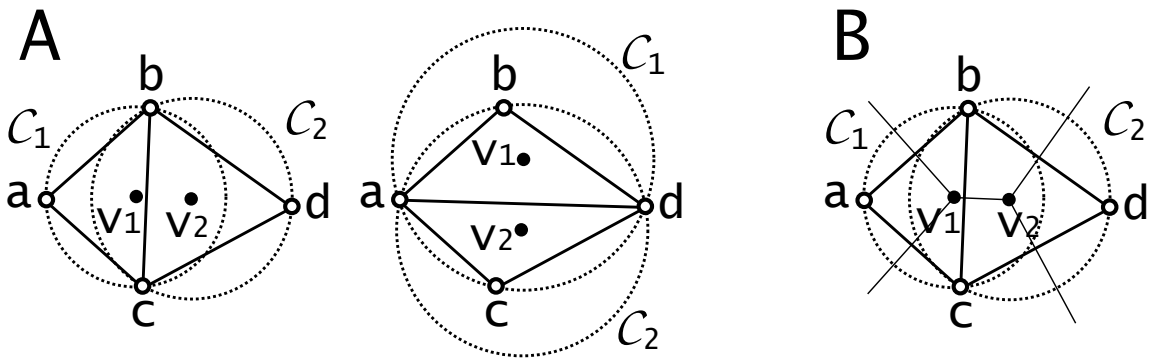


Figure B.5: Propriété de la sphère vide, un exemple en deux dimensions.
A : Deux triangulations d'un ensemble de quatre points a, b, c et d. Les sommets v_1 et v_2 sont les centres des cercles $\mathcal{C}_1$ et $\mathcal{C}_2$ respectivement circonscrits aux triangles abc et bcd dans la figure de gauche, et abd et acd dans la figure de droite.
B : Le diagramme de Voronoï des quatre points est apparent en lignes fines, la triangulation de Delaunay en lignes épaisses. Les centres des sphères circonscrites aux triangles correspondent aux sommets du diagramme de Voronoï.

d'autres termes, ce sont les centres des cercles circonscrits à un 3-simplexe (un 2-simplexe, en dimension deux) de la triangulation de Delaunay.

La propriété de la sphère vide est une propriété globale ; pour son utilisation dans le cadre d'une construction incrémentale, on la remplace par une propriété plus locale illustrée dans la figure B.6. On peut en effet montrer que demander à ce que chaque cercle circonscrit à un triangle

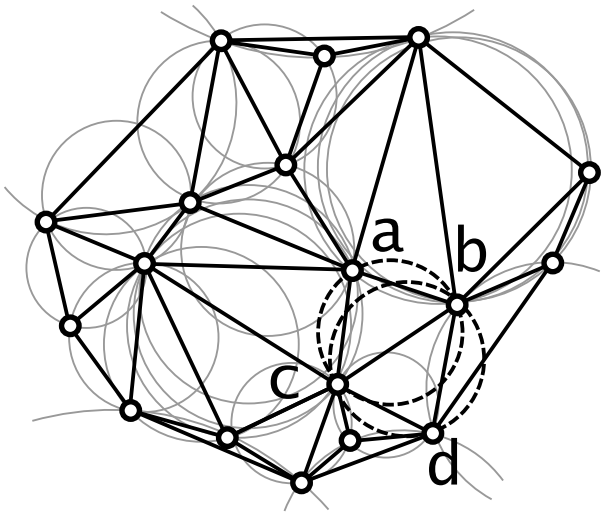


Figure B.6: Propriété de la sphère vide dans un exemple en deux dimensions ; aspect global et aspect local. Les cercles circonscrits à tous les triangles de la triangulation ont été représentés en gris, mis à part deux d'entre eux qui sont représentés en ligne interrompue. On pourra vérifier que chaque disque ainsi défini est vide d'autres points que les trois qui lui sont circonscrits.

de $\mathcal{D}$ soit vide d'autres sommets de la triangulation est équivalent à demander à ce que cette propriété soit respectée localement pour chaque paire de tétraèdres adjacents. Dans l'exemple en deux dimensions de la figure, cela revient à vérifier par exemple que le sommet a est exclu du disque défini par le cercle circonscrit à bcd, qu'il en est de même pour d et le disque adjoint à abc, et à faire cette vérification pour toutes les paires de triangles adjacents.

Cette propriété locale motive une construction incrémentale dans laquelle la triangulation n'aura besoin d'être modifiée qu'au voisinage du nouveau sommet inséré, les autres propriétés locales restant vérifiées.

Un analogue de cette ambivalence entre local et global existe pour les triangulations régulières en dimension quelconque (à ce sujet, voir par
exemple [Edelsbrunner 96]).

B.4 Voisinage directionnel défini par la triangulation de Delaunay

Les arêtes de la triangulation de Delaunay d'une molécule définissent une notion de *voisinage directionnel* pour chaque atome de la molécule. Intuitivement, la connexion de deux atomes par une telle arête signifie que ces deux atomes "se voient", que l'espace qui les sépare n'est pas encombré par d'autres atomes. Cette notion n'est pas équivalente à une proximité. Dans la figure B.7, on montre que deux atomes connectés par une telle arête peuvent être très éloignés l'un de l'autre, et qu'à l'inverse, des atomes très proches dans l'espace peuvent ne pas être connectés pour peu qu'ils soient séparés par un ou plusieurs autres atomes. Cette propriété de voisinage

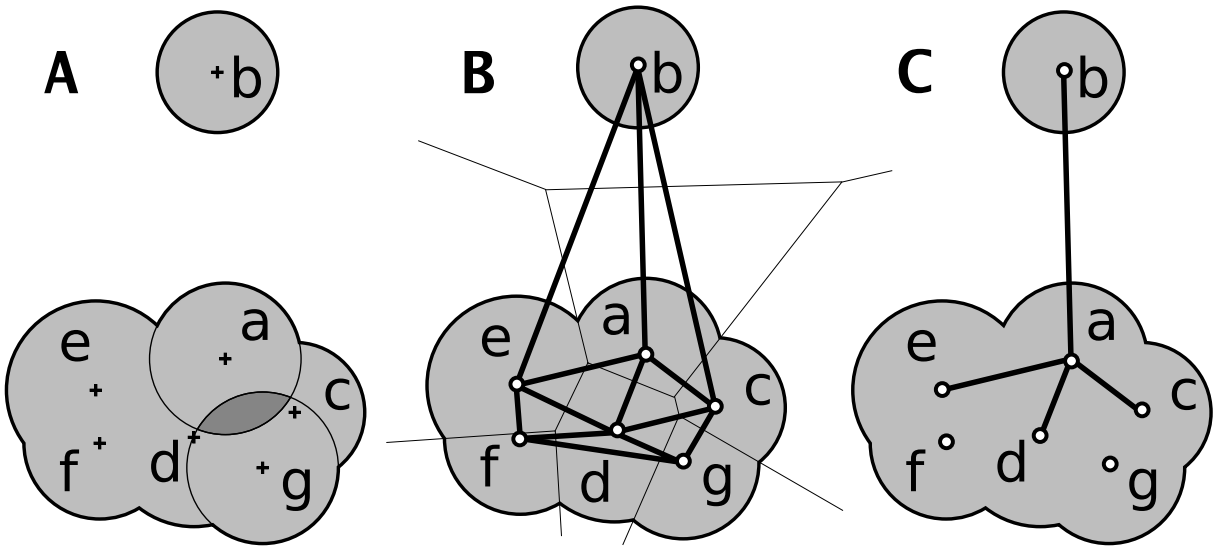


Figure B.7: Un exemple en deux dimensions montrant les différences entre la notion de *voisinage directionnel* encodée dans la triangulation de Delaunay d'une molécule et la notion intuitive de *proximité*. Les trois figures montrent un modèle Surface Accessible d'une même molécule composée de sept atomes. La figure A met en évidence une intersection entre les atomes a et g. La figure B superpose le diagramme de Voronoï (en lignes fines) et la triangulation de Delaunay (en lignes épaisses) de la molécule. Dans la figure C, seules les arêtes de la triangulation de Delaunay issues du sommet a ont été conservées, elles définissent un *voisinage directionnel* autour de l'atome a. L'atome b, bien qu'éloigné de a, est connecté à ce dernier ; c'est en effet le plus proche voisin de a "dans cette direction". À l'inverse, l'atome g n'est pas connecté à a bien qu'ils soit si proches que les boules qui modélisent les deux atomes s'intersectent. La raison de cet "oubli" est que l'atome g est "protégé" de a par les atomes c et d.

directionnel peut-être intuitivement comprise comme une notion de "plus proche voisin dans une direction donnée" tenant compte de l'encombrement stérique. Ce premier filtre de voisinage directionnel peut être affiné en écrémant les arêtes trop longues pour avoir une signification bio-

logique[1]. Le complexe dual trouve une application naturelle à ce genre de définition du voisinage directionnel, et a déjà été employé dans ce sens par Zomorodian *et al.* [Zomorodian 06] pour définir une fonction de score permettant d'évaluer la "fiabilité" d'une structure.

B.5 Points redondants

Les sommets de la triangulation régulière d'une molécule correspondent systématiquement à des centres d'atomes constituant la molécule, mais tous les atomes ne participent pas à cette construction. Certains atomes, masqués par des voisins trop encombrants, en sont en effet absents ; on les appelle *points redondants* ou *atomes redondants*. Ces "disparitions" ont essentiellement lieu dans le modèle Surface Accessible et concernent des atomes de faible rayon enfouis dans un environnement encombré. Il s'agit généralement d'atomes d'hydrogène mais d'autres types d'atomes peuvent être impactés, comme le carbone-ζ de l'arginine. Le complexe dual, ainsi que le complexe-α et la surface duale, étant construits sur une triangulation régulière, sont naturellement sujets aux mêmes disparitions.

Cette perte d'atome peut s'avérer problématique pour certaines applications bioinformatiques, aussi l'utilisateur de ces modèles doit-il avoir conscience de ce phénomène. La figure B.8 donne un exemple en deux dimensions explicitant la disparition potentielle d'un atome de faible poids dans une triangulation régulière. Dans chacune des trois lignes l'atome d est enfoui dans l'union des trois autres atomes. Dans le dernier cas cependant, lorsque d a un rayon suffisamment petit, sa cellule de puissance (obtenue par l'intersection des demi-plans Π_{da}, Π_{db} et Π_{dc}) est vide (colonne centrale). Dans ce cas, l'atome disparaît du diagramme en remplissage de forme (colonne de droite), comme de la triangulation régulière (colonne centrale).

[1] En considérant qu'à partir d'une certaine distance, l'interaction entre deux atomes est trop faible pour qu'ils soient considérés comme voisins.

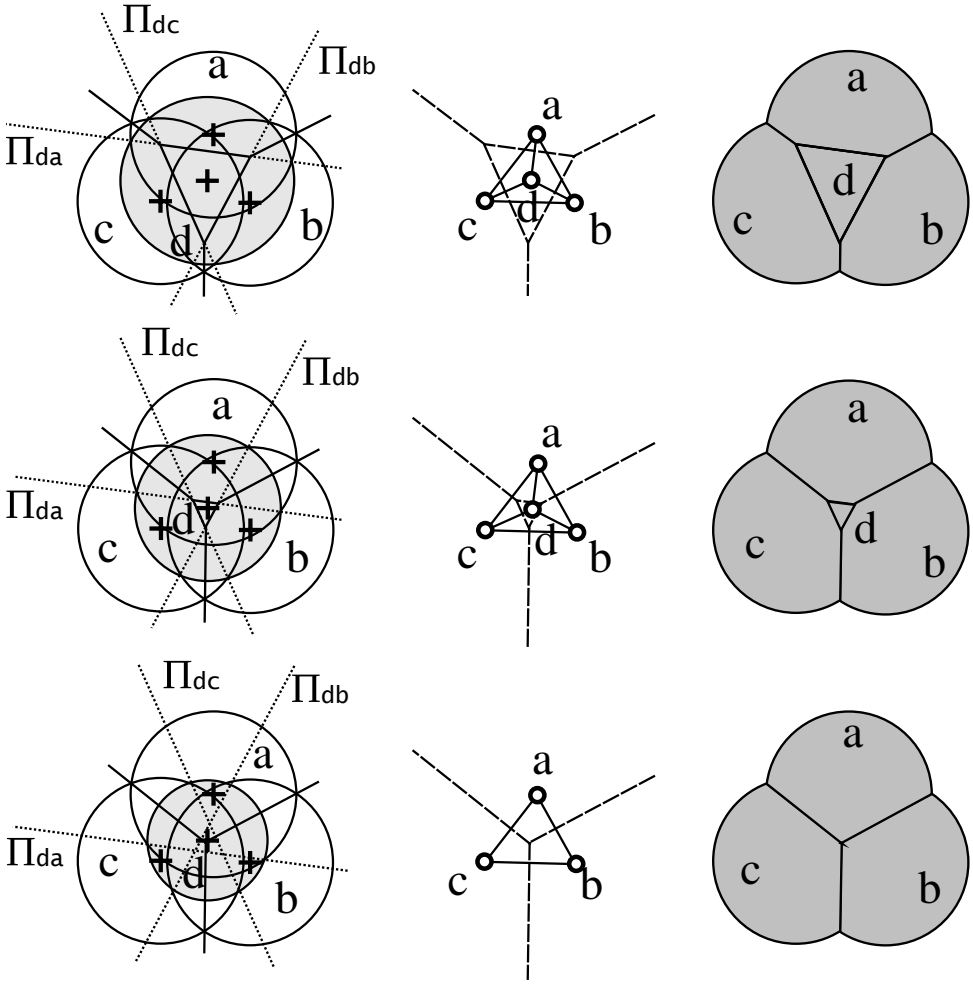


Figure B.8: Points redondants d'une triangulation régulière ou d'une molécule, un exemple en deux dimensions comprenant quatre atomes a, b, c et d. La colonne de gauche montre les quatre atomes représentés par quatre cercles ; l'atome d dont le rayon varie d'une ligne à l'autre est mis en évidence par un disque gris clair. Les droites radicales impliquant l'atome d sont indiquées en pointillés. Dans la colonne du milieu, la triangulation régulière des quatre atomes est matérialisée par ses sommets et ses arêtes (en traits pleins) ; les arêtes du diagramme de puissance sont représentées en traits interrompus. La dernière colonne montre le diagramme en remplissage de forme de la molécule.

Annexe C

Constitution d'un jeu de structures moléculaires

La constitution d'un jeu de données de structures est toujours une problématique compliquée ; elle consiste en effet à sélectionner un ensemble de structures à la fois pertinent pour répondre à une question donnée, et contenant le moins de répétition[1] possible pour éviter de biaiser les résultats obtenus.

Dans notre contexte, nous avions besoin d'un jeu de données de structures qui nous permette d'évaluer les performances de nos implémentations, et plus particulièrement celles de `Lc` (voir page 189 pour une description du logiciel, et page 93 pour la description des algorithmes). Il convenait donc de trouver un ensemble de structures représentatives des objets à l'étude desquels notre programme est susceptible d'être employé. Comme mentionné dans le chapitre 8, nos algorithmes sont adaptés à la caractérisation de la topographie de nimporte quel type de macromolécule, du peptide de quelques acides aminés à des structures complexes comprenant l'assemblage de plusieurs sous-unités. Dans un tel contexte la redondance peut s'exprimer en terme de similarité de structure, une problématique complexe fréquemment abordée au travers d'une similarité de séquence [2] ; c'est l'approche pour laquelle nous avons opté.

Constitution du jeu de données

Le site du ncbi[3] fournit un service permettant la constitution d'un jeu de structures non redondantes en séquence ; en septembre 2008, ce service était accessible à l'adresse

`http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html` Dans ce processus, toutes les séquences des protéines dont la structure est contenue dans la `PDB` [Berman 00] sont comparées entre elles et regroupées sur des critères de similarité[4]. Un représentant est ensuite choisi dans chaque groupe sur des critères de qualité de la structure[5].

Cette sélection de structure a résulté en la compilation de 4148 chaînes extraites de 3905 fichiers ; 31 de ces fichiers n'ayant pu être récupérés, les ressources pdb n'existant plus, nous

[1] Par répétition, on entend l'existence de structures "proches" ou "similaires"

[2] Bien que cette approximation ne soit pas toujours exacte [Alexander 07, Kosloff 08], il est couramment admis que des séquences similaires admettent généralement un repliement similaire.

[3] le *National Center for Biotechnology Information* est un organisme qui fournit de nombreux services et de nombreuses bases de données pour l'analyse du génôme

[4] Cette première étape est effectuée au moyen d'une analyse BLAST, paramétrée par une *p-value* que nous avons fixée à 10^{-7} recommandée par le ncbi afin d'obtenir un jeu le moins redondant possible

[5] Cette mesure de qualité prend en compte des critères tels que la certitude sur la nature des acides aminés composant la structure, la résolution de la structure, la taille de la chaîne...

taille	0	1	2	3	4	5	6	7	8	9
structures	635	887	483	176	65	35	13	8	2	1

Tableau C.1: Répartition du nombre de structures par taille (exprimée en nombre d'atomes) de notre jeu de données. Les valeurs inscrites dans la ligne du haut sont à prendre en millier d'atomes.

nous sommes restreint au contenu de 3874 fichiers. Nous avons encore restreint notre jeu de fichiers aux 3111 issus d'experiences aux rayons X , et avons conservé uniquement les 2305 ayant une résolution suffisante (inférieure à 2.4Å) et dont tous les atomes étaient connus (sur ce dernier critère, seul 7INSG a été retiré). Enfin, les atomes d'hydrogène ont été systématiquement retirés des structures, ainsi que l'eau et les ligands.

Annexe D

Logiciels réalisés

Les travaux présentés dans ce document ont donné lieu au développement des logiciels `Lc` et `Pck`, respectivement pour le calcul de la courbure locale sur la surface des macromolécules et la détection des poches dans les macromolécules. Les algorithmes ont été implémentés en language `C++` en utilisant la librairie `CGAL`, et interfacés avec le logiciel de visualisation moléculaire VMD [Humphrey 96]. Cette dernière partie a été réalisée au travers d'une interface graphique développée en langage `tclTk`. Les paquetages logiciels sont librement téléchargeables sur des sites web dédiés où l'on poura trouver une description complète ainsi qu'un manuel d'utilisation. Ils comprennent le programme `C++` compilé (des versions existent pour les architectures Windows XP et linux (32 bits et 64 bits)) ainsi que son greffon *VMD*. Les caractéristiques principales de chaque logiciel sont rappelées ci après.

Site web de `Lc` : `http://alnitak.u-strasbg.fr/Lc/`.
Le logiciel `Lc` exécuté en ligne de commande permet de calculer les valeurs de courbure locale d'une molécule (voir le chapitre 8 pour la description de cet indice et de son calcul). Pour faciliter son utilisation en favorisant un retour visuel direct, une interface au logiciel de visualisation *VMD* est fournie (figure D.1). La partie supérieure de cette interface autorise le calcul de valeurs de courbure locale : l'utilisateur choisit la sélection d'atomes à considérer ainsi que la taille du voisinage de lissage maximal. Les valeurs calculées par le programme sont chargées dans *VMD* ; il est ensuite possible de choisir le type de lissage, de re-échantillonner les valeurs et de les seuiller au travers de la seconde partie de l'interface.

Site web de `Pck` : `http://alnitak.u-strasbg.fr/Pck/`.
Exécuté en ligne de commande, le logiciel `Pck` autorise la détection et la caractérisation des poches d'une macromolécule. Une interface *VMD* est fournie pour visualiser les poches et interagir avec le programme exécutable (Voir la copie d'écran de la figure D.2).

Les principales fonctionnalités de `Pck` sont :

- Détection de poches géométriques :
  Dans sa version actuelle, `Pck` implémente deux algorithmes de détection des poches, tous deux basés sur la théorie des formes-α. Le premier consiste en une imlémentation de l'algorithme de flux discret proposé par Edelsbrunner [Edelsbrunner 98], et permet la détection et le calcul volumétrique des cavités et poches refermées. Le second algorithme est une variation d'APROPOS [Peters 96], et permet une détection des poches en surface de la molécule.
- Affichage des poches et interaction dans le logiciel de visualisation *VMD*.
- Affichage coloré des facettes bordant les poches, permettant d'appréhender la présence de

poches juxtaposées, ou d'ouvertures potentielles sur l'espace du solvant.
- Calcul du volume et de l'aire des molécules dans leur modèle Surface Accessible, Van der Waals ou Polyédrique.
- Calcul du volume et de l'aire des poches, ainsi que du nombre de bouches et de leurs aires respectives.
- Détection et sélection des atomes et résidus contribuant aux poches detectées
- Calcul d'une matrice de distance entre les poches ; cette information est entre autre utilisée pour permettre d'inférer un potentiel "recrutement", ou une plasticité dans les poches.
- Détection des molécules d'eau cristallographiques dans les poches.

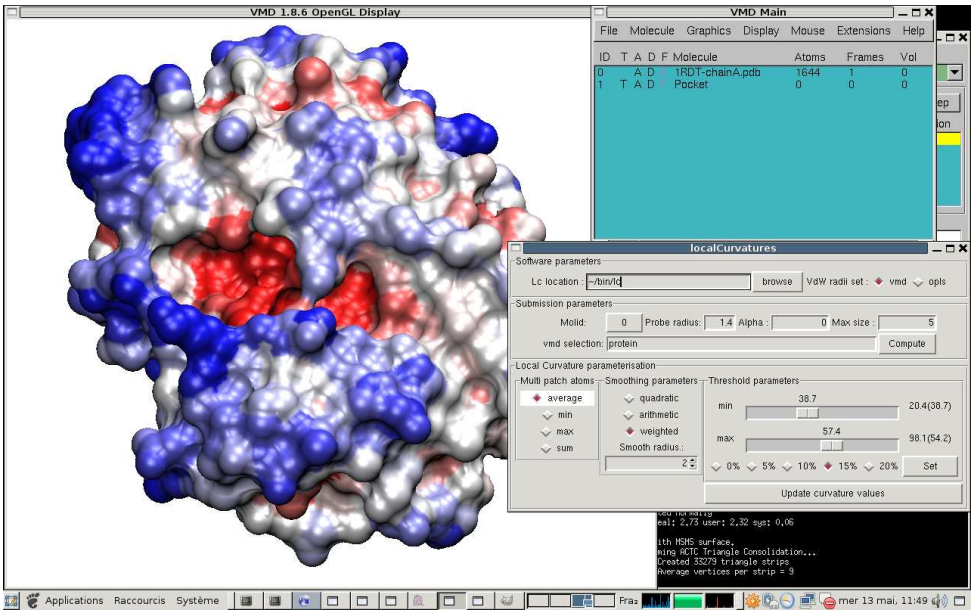


Figure D.1: Capture d'écran du logiciel `Lc` interfacé avec VMD. Visualisation d'un sillon à l'interface de dimérisation d'un récepteur nucléaire (1RDT).

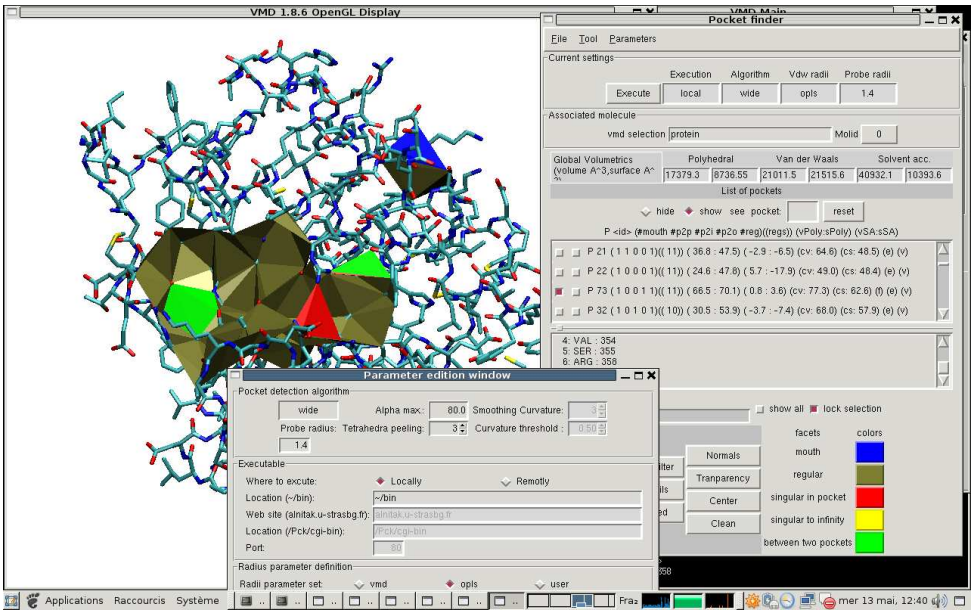


Figure D.2: Capture d'écran du logiciel `Pck` interfacé avec VMD.

Annexe E

Principes et défauts de la discrétisation

Discrétiser un objet ou une fonction consiste à plonger virtuellement cet objet ou cette fonction dans une grille régulière. Pour la discrétiser, la fonction sera évaluée à chaque nœud de la grille, et les valeurs calculées seront, par exemple, utilisées pour approximer (ou extrapoler) les valeurs de la fonction entre les nœuds de la grille. Pour discrétiser un objet (voir la figure E.1 A), on identifie les nœuds qui font partie de l'objet (comme le nœud a de la figure) et ceux qui sont en dehors (comme le nœud b). Ces nœuds sont fréquemment assimilés à un élément complet de la grille : en deux dimensions, ces éléments carrés sont nommés pixels, en trois dimensions, ces éléments sont cubiques et nommés voxels. Dans la figure, les nœuds sont par exemple assimilés au pixel dont ils constituent le coin inférieur gauche, comme on peut le voir pour le pixel gris foncé adjoint au nœud a.

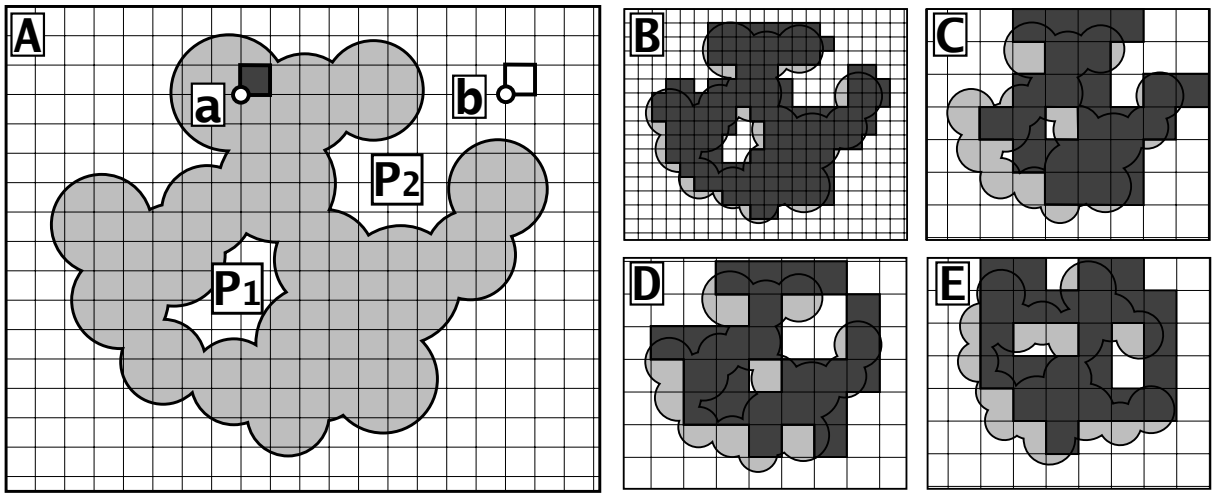


Figure E.1: Discrétisation d'une molécule, principes, qualités et défauts ; un exemple en deux dimensions. La figure (A) montre une molécule en modèle Surface Accessible plongée dans une grille. Les nœuds a et b sont respectivement à l'intérieur et à l'extérieur de cette molécule ; le pixel attaché à a est colorié en gris foncé pour symboliser le fait qu'il représente une partie de la molécule discrétisée. Les figures (B) à (E) constituent des discrétisations de la même molécule. La discrétisation (B) est plus fine que les trois autres ; dans la représentation (D), la molécule a subi une légère translation dans la grille, et dans la représentation (E), elle a subi une rotation.

Des techniques basées sur une discrétisation sont souvent mises en œuvre car simples à manipuler et rapides à implémenter. Elles souffrent néanmoins de désagréments rédhibitoires pour certaines applications. La discrétisation dépend par exemple de paramètres tels que la

résolution de la grille (schémas (B) et (C)), la position des objets ou leur orientation dans la grille (schémas (C), (D) et (E)). Ces paramètres peuvent aussi influer sur la topologie de l'objet discrétisé, comme on peut l'observer pour la poche P_2 qui devient une cavité dans le schéma (E). Corollaire de cette observation, une discrétisation ne garantit pas la conservation de la topologie de l'objet discrétisé. Ce genre d'instabilité implique l'impossibilité de garantir la reproductibilité des résultats calculés sur un changement même minime des paramètres. Enfin, les temps de calculs sont souvent élevés en raison du grand nombre de voxels à considérer. De manière générale, ces approches impliquent un compromis entre précision et ressources (mémoire et temps de calcul) ; la précision peut en effet être améliorée avec une grille plus fine, mais une grille d fois plus fine contient d^3 fois plus de voxels.

Annexe F

Présentation de la bibliothèque CGAL

Le projet CGAL(pour ***C**omputational **G**eometry **A**lgorithms **L**ibrary*) est le fruit d'une collaboration entre plusieurs sites de recherche en géométrie algorithmique, essentiellement européens. Le but de ce projet colaboratif est de fournir des structures de données et des algorithmes pour la construction et la manipulation du plus grand nombre d'objets du domaine sous la forme d'une bibliothèque *open source* de classes C++. Pour indication, la figure F.1 montre les paquetages principaux fournis par la bibliothèque et, en grisé, ceux que nous avons utilisés. Initié en

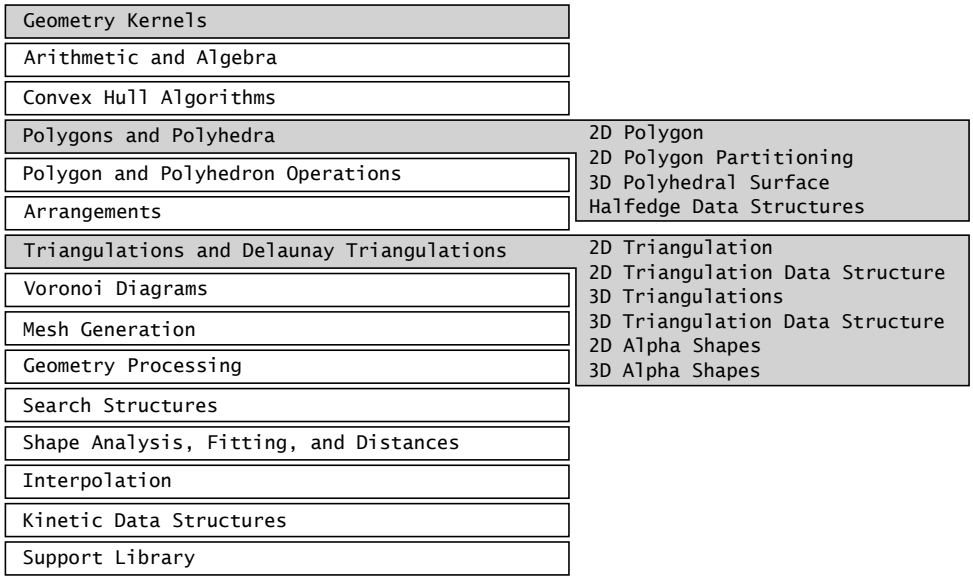


Figure F.1: Problématiques traitées par la version 3.3 de la bibliothèque CGAL. Les entrées grisées correspondent aux éléments que nous avons utilisés dans le cadre de notre travail.

1996, le nombre d'utilisateurs et de contributeurs a fortement augmenté, traitant de nouvelles problématiques et garantissant une robustesse aux erreurs informatiques et une bonne réactivité à la fois des développeurs et des utilisateurs au travers d'une liste de diffusion dynamique.

Dans la suite de cette section nous présenterons succinctement la conception générale de la librairie CGAL, puis celle des trois classes principales que nous avons utilisées dans nos développements.

F.1 Présentation générale de la bibliothèque CGAL

Destinée à la fois au monde universitaire et au monde industriel, une attention particulière a été accordée à la conception de la bibliothèque CGAL. Le choix du langage C++ s'est imposé à la fois pour des raisons de performance et pour l'utilisation du paradigme objet[1]. La conception de CGAL s'inspire de celle de la librairie STL[2], et des ponts sont proposés avec une autre librairie faisant office de standard dans le langage, entre autres pour la gestion des graphes : la librairie *boost*[3]. L'emploi des templates du C++[4] en combinaison avec des *techniques de traits*[5] [Myers 95] offre enfin un niveau de configuration permettant de décorréler les structures de données afférentes aux objets géométriques, les algorithmes permettant de les construire ou de les manipuler, et les types numériques utilisés pour leur construction (par exemple pour la définition et la manipulation des coordonnées des centres atomiques).

La bibliothèque CGAL est globalement architecturée sur trois axes, les *types numériques*, les *noyaux*, et les *problèmes géométriques*.

F.1.1 Types numériques CGAL

Les types numériques usuels du C++ (float, double, long double,...) peuvent être employés pour définir les objets CGAL, mais d'autres types sont fournis par la librairie, qui permettent par exemple de travailler dans des espaces courbes, ou de réaliser des calculs exacts ou filtrés[6].

F.1.2 Noyaux CGAL

Un noyau (ou *kernel*) contient les objets géométriques de base (tels que les points, les lignes ou les vecteurs), ainsi que les opérations et prédicats sur ces objets (additions, produits scalaires, test d'appartenance d'un point à une droite,...). Les noyaux CGAL sont paramétrés par un type numérique, et pour faciliter l'utilisation de la librairie, des noyaux pré-paramétrés sont fournis pour les utilisations les plus courantes.

Le noyau pré-paramétré *Exact_predicates_inexact_constructions_kernel* que nous avons systématiquement employé dans nos développements est défini sur un type numérique utilisant un filtre arithmétique qui permet un calcul exact des prédicats géométriques sans trop dégrader

[1] Un type de conception de programme facilitant la prise en main, l'extension et la réutilisabilité.

[2] La ***S**tandard **T**emplate **L**ibrary* est la bibliothèque standard du C++. Elle définit et fournit essentiellement des classes de conteneurs (listes, vecteurs, files, tables de hachage...). La conception de la STL permet de décorréler les conteneurs, les méthodes d'accès aux éléments contenus (itérateurs) et les algorithmes applicables aux conteneurs (fonction de tri ou de recherche d'un élément précis par exemple).

[3] `http://www.boost.org`

[4] Les templates du C++ permettent l'écriture de classes ou de fonctions génériques. Concrètement, la classe ou la fonction est définie avec un ou plusieurs paramètres "génériques" (ou *template*) qui seront instanciés au moment de la compilation. Il est ainsi possible de définir une classe `liste` contenant des objet d'un type générique `T`, qui sera automatiquement déclinée par exemple en liste d'entiers, liste de chaînes de caractères, ou liste d'acides nucléiques en fonction des besoins. Une telle classe générique s'écrit :
template <**class** T> **class** liste {...}

[5] Les arguments génériques (ou *templates*) peuvent être utilisés pour paramétrer une classe avec tout un contexte afférent à une problématique. Ce contexte est généralement passé sous la forme d'une classe rassemblant une information parfois volumineuse et composée de définitions de types ou de classes, de constantes, et d'algorithmes. Une telle classe est usuellement appelée *classe de trait*.

[6] Les filtres arithmétiques sont utilisés pour accélérer les calculs de prédicats. Ils consistent en la réalisation d'un premier calcul approximatif rapide, suivi d'un second calcul plus précis mais aussi plus lent, uniquement si le premier calcul ne permet pas de répondre de manière satisfaisante.

les performances. Ce noyau permet de construire sans erreur la triangulation de Delaunay mais pas le diagramme de Voronoï[7].

F.1.3 Les problèmes géométriques

La version 3.3 de la librairie CGAL fournit une infrastructure traitant un vaste éventail de problématiques géométriques dont : la gestion de maillages de surfaces (et plus spécifiquement des surfaces polyédriques[8]), les triangulations en 2D et en 3D (en particulier les triangulations de Delaunay, pondérées ou non), ainsi que les formes-α.

Ces problématiques sont implémentées sous la forme de classes C++, offrant une interface de haut niveau pour la construction et la manipulation de ces objets géométriques. Elles sont généralement paramétrées par une classe de trait définissant les types des objets géométriques de base (points, lignes...) et les prédicats nécessaires à la construction de l'objet géométrique.

F.2 Modèles géométriques implémentés dans les classes CGAL

Pour nos développements, nous avons essentiellement eu recours à trois classes de la libraire CGAL : la classe *Regular_triangulation_3* [Pion 07] pour la construction et la gestion des triangulations de Delaunay pondérées, la classe *Alpha_shape_3* [Da 06] pour la construction et la gestion des complexes-α et des formes-α, ainsi que la classe *CGAL_HALFEDGEDS_DEFAULT* [Kettner 06] pour la gestion d'un maillage de surface au travers d'une structure de données de demi-arêtes. Ces trois classes et leurs spécificités d'implémentation sont rapidement présentées ci-après.

F.2.1 Triangulation de Delaunay

L'infrastructure logicielle de CGAL concernant la gestion des triangulations est présentée dans la figure F.2. La classe *Triangulation_3* constitue une interface de haut niveau pour la gestion des triangulations en trois dimensions. Elle fournit entre autres des méthodes pour accéder aux divers simplexes d'une triangulation et la traverser en suivant les relations d'adjacence ou d'incidence entre simplexes. On peut ainsi visiter tour à tour tous les simplexes d'une certaine dimension au moyen d'itérateurs, ou parcourir tous les simplexes incidents à un autre simplexe à l'aide de circulateurs[9]. Pour des raisons d'optimisation, dans les triangulations fournies par CGAL, les seuls objets combinatoires stockés dans la structure de données sous-jascente sont '*cellules* (3-simplexes), et les sommets (0-simplexes). Chaque sommet d'une triangulation CGAL contient un lien vers une de ses cellules adjacentes, et chaque cellule contient des liens vers ses sommets incidents. Pour chaque cellule, les liens vers les sommets sont accessibles au travers d'un indice i à valeur dans $\{0, 1, 2, 3\}$, comme illustré dans l'exemple de la figure F.3 *A*. Un même sommet, partagé par deux cellules distinctes n'aura pas nécessairement le même indice dans chaque cellule. Dans notre exemple, le sommet v_2 a l'indice j dans la cellule c_1 — ce que nous avons écrit $v_2 = c_1.v(j)$ — et l'indice j' dans c_2, avec j potentiellement différent de j'. L'implémentation

[7] Avec le noyau *Exact_predicates_inexact_constructions_kernel*, la combinatoire du diagramme de Voronoï (le nombre de sommets et leur regroupement en arêtes, facettes et cellules) sera correcte, mais la position des sommets pourra être erronée.

[8] Des algorithmes de subdivision sont aussi fournis, ainsi qu'un support pour la génération de maillages de surfaces skins.

[9] Les itérateurs et les circulateurs constituent un type de classe particulier permettant d'accéder séquentiellement à tous les éléments d'un conteneur. Un itérateur émule un conteneur de type "liste" ; la séquence des éléments visités a un début et une fin. À l'inverse, le circulateur boucle à l'infinie sur sa séquence.

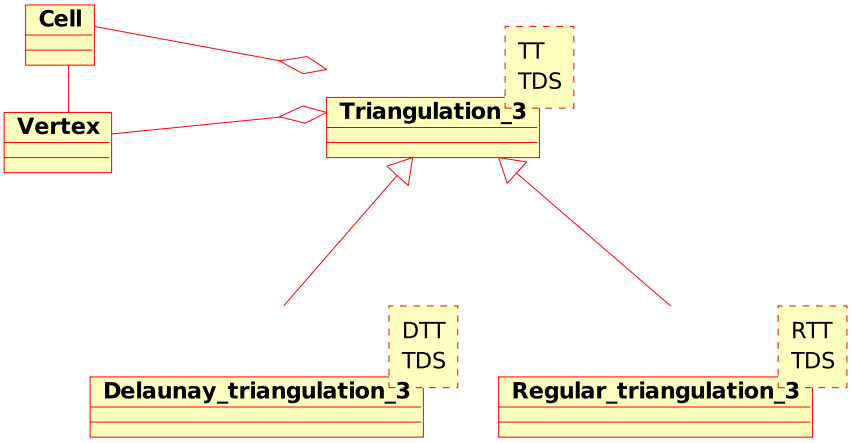


Figure F.2: La classe *Triangulation_3* fournit une interface de haut niveau pour la gestion des triangulations de points dans l'espace. Elle est paramétrée par deux arguments templates, respectivement une classe de trait *TT* (*Triangulation traits*) géométrique définissant les prédicats et objets géométriques (points, segments, lignes, triangles, tétraèdres...), et une classe *TDS* (*Triangulation data structure*) fournissant une structure de données pour le stockage et la gestion de la combinatoire de la triangulation. Cette dernière doit impérativement définir les classes *Cell* et *Vertex* qui seront utilisées pour construire et gérer les 3-simplexes et les 0-simplexes de la triangulation (les 1-simplexes et les 2-simplexes ne sont pas explicitement stockés). Il est possible par ce biais de créer ses propres classes de simplexes, contenant par exemple une information relative au problème que l'on souhaite traiter.
Les classes *Delaunay_triangulation_3* et *Regular_triangulation_3* sont dérivées de *Triangulation_3*, elles fournissent respectivement une interface pour la construction et la manipulation des triangulations de Delaunay d'un ensemble de points et d'un ensemble de sphères. Elles sont paramétrées par des classes de traits géométriques *DTT* (*Delaunay triangulation traits*) et *RTT* (*Regular triangulation traits*) fournissant les prédicats nécessaires à leur construction.

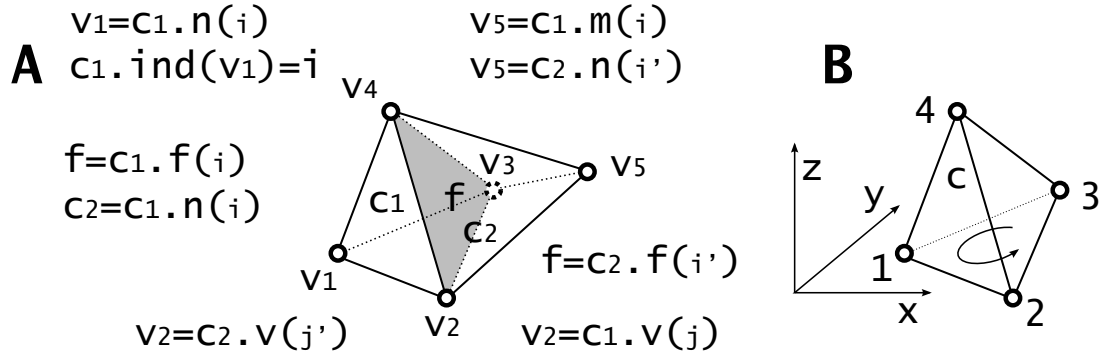


Figure F.3: Cellules et sommets d'une triangulation CGAL.
A : Un exemple de triangulation de cinq sommets v_i en trois dimensions. Cette triangulation comprend deux cellules c_1 et c_2 partageant une facette f. Les méthodes d'accès $v()$, $ind()$, $n()$, $m()$ et $f()$ symbolisent des opérations fournies dans l'implémentation CGAL et permettent respectivement d'obtenir un sommet dans une cellule, l'indice d'un sommet dans une cellule, une cellule adjacente, un sommet opposé et une facette.
B : Pris dans l'ordre croissant de leurs indices, les sommets d'une cellule sont orientés positivement.

proposée dans CGAL permet de récupérer l'indice d'un sommet dans une cellule ($c.ind(v_1) = i$), de visiter la cellule adjacente à une autre cellule ou un sommet opposé au travers d'une facette. Ces deux dernières opérations sont rendues possibles par l'utilisation du même indice que celui du sommet opposé à la facette partagée par les deux cellules mises en jeu ($c_2 = c_1.n(i)$) et ($v_5 = c.m(i)$).

Les facettes (2-simplexes) ne sont pas directement stockées dans la triangulation, mais représentées par une paire *(cellule, indice)* qui repère une cellule de la triangulation contenant la facette concernée, et l'indice dans cette cellule du sommet opposé à la facette. De même, les arêtes (1-simplexes) sont représentées par un triplet *(cellule, indice, indice)* repérant une cellule de la triangulation contenant cette arête, et les indices dans cette cellule des deux sommets composant l'arête. Comme indiqué dans la figure F.3 *B*, les sommets des cellules sont indexées de manière à ce que, pris dans l'ordre croissant, ils orientent la cellule positivement : un observateur placé sur le sommet d'indice 3 et regardant vers le plan formé par les trois autres sommets verra les trois sommets d'indice 0, 1 et 2 (pris dans cet ordre) orientés dans le sens trigonométrique.

La classe *Regular_ triangulation_ 3* permet la construction d'une triangulation de Delaunay à partir d'un ensemble de points pondérés (des boules). Dérivant de la classe *Triangulation_ 3*, elle en offre les mêmes fonctionnalités, en particulier concernant le parcours des simplexes.

F.2.2 Complexe-α et forme-α

La classe *Alpha_ shape_ 3* de la librairie CGAL permet la construction et la manipulation du complexe-α d'un ensemble de points ou de sphères. Comme indiqué dans la figure F.4, la classe *Alpha_ shape_ 3* étend une classe de triangulation, et fournit ainsi directement les méthodes d'accès aux simplexes de la triangulation sous-jacente. Elle y ajoute des constantes et des méthodes relatives au tri des simplexes et à la gestion de la valeur α. Ce choix d'implémentation est justifié par la possibilité de définir le complexe-α comme un tri (une *filtration*) des simplexes de la triangulation de Delaunay comme nous l'avons vu dans la section 6.1.5 page 70.

Comme vu précédemment (section 6.1.2 page 61), les simplexes d'une triangulation de Delaunay peuvent être classés en quatre types : *intérieur*, *extérieur*, *régulier* et *singulier* selon leur appartenance ou non au complexe-α et à son bord. Cette classification est reprise sous la forme

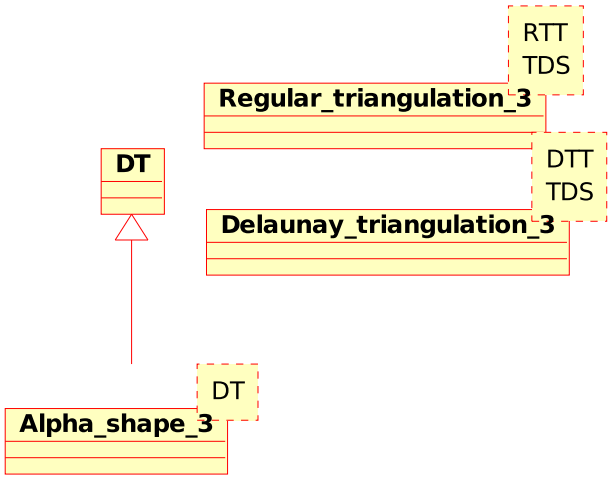


Figure F.4: La classe *Alpha_shape_3* de la librairie CGAL permet la construction et la gestion des complexes-α d'un ensemble de points ou d'atomes. Elle hérite d'une classe *DT* (*Delaunay triangulation*) passée en argument template et qui peut-être une classe de triangulation de Delaunay non pondérée (*Delaunay_triangulation_3*) ou pondérée (*Regular_triangulation_3*).

EXTERIOR	Pour les simplexes n'appartenant pas au complexe-α considéré.
SINGULAR	Pour les simplexes participant du bord au complexe-α, et dont aucun des simplexes de dimension supérieure incidents n'est dans le complexe-α. Ces simplexes sont soit des arêtes isolées, soit des facettes isolées.
REGULAR	Pour les simplexes participant du bord du complexe-α, et dont l'un des simplexes de dimension supérieure incidents appartient aussi au complexe-α.
INTERIOR	Pour les simplexes inclus dans le complexe-α et dont tous les simplexes de dimension supérieure sont aussi inclus dans le complexe-α.

Tableau F.1: Constantes définies par la classe *Alpha_shape_3* dans CGAL pour la classification des simplexes d'une triangulation de Delaunay.

de constantes définies dans la classe *Alpha_shape_3* et présentées dans la table F.1.

La classe *Alpha_shape_3* fournit des itérateurs pour parcourir l'ensemble des simplexes d'un type et d'une dimension donnés, ainsi que des méthodes permettant de connaître le type d'un simplexe donné.

F.2.3 Structure de demi-arête

Dans notre étude, nous avons caractérisé la surface et la forme des macromolécules au travers du polyèdre délimité par le bord du complexe dual. Plusieurs approches existent pour représenter et manipuler informatiquement ce genre d'objets géométriques au travers de structures combinatoires. Parmi celles-ci, les demi-arêtes [Kettner 99] sont adaptées à la construction de maillages de surfaces variété orientables (voir les figures 6.10 et 6.11), et les G-cartes [Bertrand 92] permettent de manipuler des variétés non orientables, voire des surfaces non-variétés. Nos besoins étant limités au cas des surfaces variété orientables, nous avons choisi d'utiliser une structure de demi-arête, dont la librairie CGAL contient une implémentation [Kettner 06].

La conception proposée par Lutz Kettner dans la librairie CGAL (présentée dans la figure F.5) prévoit une grande souplesse d'utilisation. Une classe d'*Items* fournit les objets de base de la structure (demi-arêtes, sommets, facettes). Sa redéfinition par l'utilisateur permet entre autres un contrôle de l'information contenue dans ces objets pour répondre à des problèmes spécifiques. Une classe *Halfedge_ data_ structure*, paramétrée par une classe d'*Items*, implémente la structure de données de demi-arêtes. Elle consiste essentiellement en un conteneur d'*Items*, et fournit des méthodes pour insérer ceux-ci et y accéder. Une classe *Polyhedron* fournit des méthodes de plus haut niveau pour la manipulation générique des polyèdres en dimension 3. Construite de manière transparente sur une classe *Halfedge_ data_ structure*, elle en partage les *Items* sans autoriser l'accès à la structure de données de demi-arêtes sous-jacente. Cette classe est utilisée dans d'autres modules CGAL où la connaissance de la structure de donnée sous-jacente au polyèdre n'est pas nécessaire, comme pour la constitution des surfaces skins [Edelsbrunner 99, Kruithof 07a] ou pour la génération de surfaces de subdivision [Shiue 07].

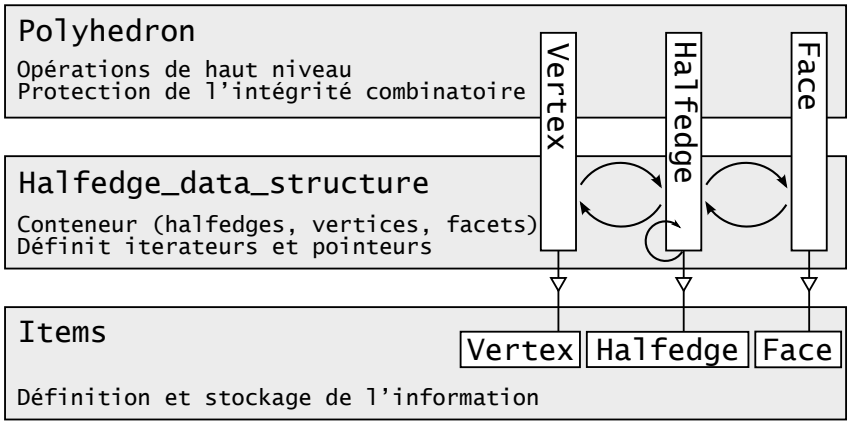


Figure F.5: Design CGAL pour la gestion des maillages de surface polygonales. La classe de haut niveau *Polyhedron* est construite sur une classe *Halfedge_ data_ structure* ; ces deux classes sont paramétrées par une classe d'*Items* contenant les types de sommets, de demi-arêtes et de facettes utilisés dans la structure de données de demi-arêtes.

Index

Bibliographie

[Akkiraju 95] N. Akkiraju, H. Edlesbrunner, M. Facello, P. Fu, E. Mucke & C. Varela. *Alpha Shapes : Definition and Software.* In Proceedings of the 1st International computational Geometry Software Workshop, pages 63–66, 1995.

[Akkiraju 96] Nataraj Akkiraju & Herbert Edelsbrunner. *Triangulating the Surface of a Molecule.* Discrete Applied Mathematics, vol. 71, no. 1-3, pages 5–22, 1996.

[Alexander 07] Patrick A Alexander, Yanan He, Yihong Chen, John Orban & Philip N Bryan. *The design and characterization of two proteins with 88sequence identity but different structure and function.* Proc Natl Acad Sci U S A, vol. 104, no. 29, pages 11963–8, 2007.

[An 04] Jianghong An, Maxim Totrov & Ruben Abagyan. *Comprehensive identification of "druggable" protein ligand binding sites.* Genome Inform Ser Workshop Genome Inform, vol. 15, no. 2, pages 31–41, 2004.

[An 05] Jianghong An, Maxim Totrov & Ruben Abagyan. *Pocketome via comprehensive identification and classification of ligand binding envelopes.* Mol Cell Proteomics, vol. 4, no. 6, pages 752–61, 2005.

[Attali 05] Dominique Attali & Herbert Edelsbrunner. *Inclusion-exclusion formulas from independent complexes.* In SCG '05 : Proceedings of the twenty-first annual symposium on Computational geometry, pages 247–254, New York, NY, USA, 2005. ACM Press.

[Aurenhammer 87] Franz Aurenhammer. *Power Diagrams : Properties, Algorithms and Applications.* SIAM J. Comput., vol. 16, no. 1, pages 78–96, 1987.

[Aurenhammer 91] Franz Aurenhammer. *Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure.* ACM Comput. Surv., vol. 23, no. 3, pages 345–405, 1991.

[Bajaj 97] Chandrajit L. Bajaj, H. Y. Lee, R. Merkert & Valerio Pascucci. *NURBS Based B-rep Models for Macromolecules and Their Properties.* In Symposium on Solid Modeling and Applications, pages 217–228, 1997.

[Bajaj 03a] Chandrajit L. Bajaj, Valerio Pascucci, Ariel Shamir, Robert J. Holt & Arun N. Netravali. *Dynamic maintenance and visualization of molecular surfaces.* Discrete Applied Mathematics, vol. 127, pages 23–51, April 2003.

[Bajaj 03b] Chandrajit L. Bajaj, Guoliang Xu, Robert J. Holt & Arun N. Netravali. *Nurbs Approximation of A-Splines and A-Patches.* Int. J. Comput. Geometry Appl., vol. 13, no. 5, pages 359–390, 2003.

[Ban 04] Y.-H. A. Ban, H. Edelsbrunner & J. Rudolph. *Interface surfaces for protein-protein complexes.* In Intl. Conf. Res. Comput. Mol. Bio., numéro 8th, pages 205–212, San Diego, California„ 2004.

[Ban 06] Yih-En Andrew Ban, Herbert Edelsbrunner & Johannes Rudolph. *Interface surfaces for protein-protein complexes.* J. ACM, vol. 53, no. 3, pages 361–378, 2006.

[Barber 96] C. Bradford Barber, David P. Dobkin & Hannu Huhdanpaa. *The quickhull algorithm for convex hulls.* ACM Trans. Math. Softw., vol. 22, no. 4, pages 469–483, 1996.

[Berman 00] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov & P E Bourne. *The Protein Data Bank.* Nucleic Acids Res, vol. 28, no. 1, pages 235–42, 2000.

[Bernauer 08] Julie Bernauer, Ranjit Prasad Bahadur, Francis Rodier, Joel Janin & Anne Poupon. *DiMoVo : a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions.* Bioinformatics, vol. 24, no. 5, pages 652–8, 2008.

[Bertrand 92] Y. Bertrand, J.-F. Dufourd, J. Françon & P. Lienhardt. *Modélisation volumique à base topologique.* In Proc. of MICAD'92, pages 59–74, 1992.

[Bhinge 04] Akshay Bhinge, Purbani Chakrabarti, Kavitha Uthanumallian, Kanika Bajaj, Kausik Chakraborty & Raghavan Varadarajan. *Accurate detection of protein :ligand binding sites using molecular dynamics simulations.* Structure (Camb), vol. 12, no. 11, pages 1989–99, 2004.

[Bingding 06] Huang Bingding & Schroeder Michael. *LIGSITEcsc : predicting ligand binding sites using the Connolly surface and degree of conservation.* BMC Struct Biol., vol. 6, no. 19, 2006.

[Binkowski 03a] T. Andrew Binkowski, Larisa Adamian & Jie Liang. *Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns.* J. Mol. Biol, vol. 332, pages 505–526, 2003.

[Binkowski 03b] T Andrew Binkowski, Shapor Naghibzadeh & Jie Liang. *CASTp : Computed Atlas of Surface Topography of proteins.* Nucleic Acids Res, vol. 31, no. 13, pages 3352–5, 2003.

[Boissonnat 95] Jean-Daniel Boissonnat & Mariette Yvinec. Géométrie algorithmique. Ediscience international, Paris, 1995.

[Boissonnat 98] Jean-Daniel Boissonnat & Mariette Yvinec. Algorithmic geometry. Cambridge University Press, New York, NY, USA, 1998.

[Bondi 64] A. Bondi. *Van der waals volumes and radii.* Journal of Physical Chemistry, vol. 68, pages 441–451, 1964.

[Brady 00] G P Jr Brady & P F Stouten. *Fast prediction and visualization of protein binding pockets with PASS.* J Comput Aided Mol Des, vol. 14, no. 4, pages 383–401, 2000.

[Branden 98] Iv Branden Carl. Introduction to structural biology. Routledge, 1998.

[Bryant 04] Robert Bryant, Herbert Edelsbrunner, Patrice Koehl & Michael Levitt. *The Area Derivative of a Space-Filling Diagram.* Discrete & Computational Geometry, vol. 32, no. 3, pages 293–308, 2004.

[Burr 04] Michael Burr, Alan Cheng, Ryan Coleman & Diane L. Souvaine. *Transformations and algorithms for least sum of squares hypersphere fitting.* In CCCG, pages 104–107, 2004.

[Cai 98] Wensheng Cai, Maosen Zhang & Bernard Maigret. *New approach for representation of molecular surface.* Journal of Computational Chemistry, vol. 19, no. 16, pages 1805–1815, 1998.

[Can 06] Tolga Can, Chao-I Chen & Yuan-Fang Wang. *Efficient molecular surface generation using level-set methods.* J Mol Graph Model, vol. 25, no. NIL, pages 442–454, 2006.

[Cazals 06] Frederic Cazals, Flavien Proust, Ranjit P Bahadur & Joel Janin. *Revisiting the Voronoi description of protein-protein interfaces.* Protein Sci, vol. 15, no. 9, pages 2082–92, 2006.

[CCG] CCG. *The Chemical Computing Group.* http ://www.chemcomp.com.

[CGAL 09] CGAL. Cgal, *Computational Geometry Algorithms Library*, 2009. http ://www.cgal.org.

[Chakravarty 99] S Chakravarty & R Varadarajan. *Residue depth : a novel parameter for the analysis of protein structure and stability.* Structure Fold Des, vol. 7, no. 7, pages 723–32, 1999.

[Chakravarty 02] Suvobrata Chakravarty, Akshay Bhinge & Raghavan Varadarajan. *A procedure for detection and quantitation of cavity volumes proteins. Application to measure the strength of the hydrophobic driving force in protein folding.* J Biol Chem, vol. 277, no. 35, pages 31345–53, 2002.

[Chavent 08] Matthieu Chavent, Bruno Levy & Bernard Maigret. *MetaMol : high-quality visualization of molecular skin surface.* J Mol Graph Model, vol. 27, no. 2, pages 209–16, 2008.

[Cheng 01] Ho-Lun Cheng, Tamal K. Dey, Herbert Edelsbrunner & John Sullivan. *Dynamic Skin Triangulation.* Discrete & Computational Geometry, vol. 25, no. 4, pages 525–568, 2001.

[Clarkson 93] Kenneth L. Clarkson, Kurt Mehlhorn & Raimund Seidel. *Four results on randomized incremental constructions.* Comp. Geom. : Theory and Applications, pages 185–121, 1993.

[Coleman 05] Ryan G Coleman, Michael A Burr, Diane L Souvaine & Alan C Cheng. *An intuitive approach to measuring protein surface curvature.* Proteins, vol. 61, no. 4, pages 1068–74, 2005.

[Coleman 06] Ryan G Coleman & Kim A Sharp. *Travel depth, a new shape descriptor for macromolecules : application to ligand binding.* J Mol Biol, vol. 362, no. 3, pages 441–58, 2006.

[Connolly 83] M. L. Connolly. *Analytical Molecular Surface Calculation.* Journal of Applied Crystallography, no. 16, pages 548–558, 1983.

[Connolly 85a] M. L. Connolly. *Computation of Molecular Volume.* J. Am. Chem. Soc., vol. 107, pages 1118–1124, 1985.

[Connolly 85b] M. L. Connolly. *Molecular Surface Triangulation.* J. Appl. Cryst, vol. 18, pages 499–505, 1985.

[Connolly 86a] M L Connolly. *Measurement of protein surface shape by solid angles.* J. Mol. Graph., vol. 4, no. 1, pages 3–6, 1986.

[Connolly 86b] M. L. Connolly. *Measurement of protein surface shape by solid angles.* J. Mol. Graphics, vol. 4, pages 3–6, 1986.

[Connolly 86c] M L Connolly. *Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface.* Biopolymers, vol. 25, no. 7, pages 1229–47, 1986.

[Corey 53] Robert B Corey & Linus Pauling. *Molecular Models of Amino Acids, Peptides, and Proteins.* Review of Scientific Instruments, 1953.

[Cormen 01] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest & Clifford Stein. Introduction to algorithms. MIT Press and McGraw-Hill, 2001.

[Da 06] Tran Kai Frank Da. *3D Alpha Shapes.* In CGAL Editorial Board, editeur, CGAL-3.2 User and Reference Manual. 2006.

[De-Alarcon 02] Pedro A De-Alarcon, Alberto Pascual-Montano, Amarnath Gupta & Jose M Carazo. *Modeling shape and topology of low-resolution density maps of biological macromolecules.* Biophys J, vol. 83, no. 2, pages 619–32, 2002.

[Delaunay 34] Boris Delaunay. *Sur la sphère vide.* Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk, vol. 7, pages 793–800, 1934.

[Dey 03] Tamal K. Dey, Joachim Giesen & Matthias John. *Alpha-shapes and flow shapes are homotopy equivalent.* In STOC, pages 493–502. ACM, 2003.

[Digby 06] Gregory J Digby, Robert M Lober, Pooja R Sethi & Nevin A Lambert. *Some G protein heterotrimers physically dissociate in living cells.* Proc Natl Acad Sci U S A, vol. 103, no. 47, pages 17789–94, 2006.

[Dijkstra 59] Edsger Wybe Dijkstra. *A note on two problems in connexion with graphs.* Numerische Mathematik, 1959.

[Duncan 93a] B S Duncan & A J Olson. *Approximation and characterization of molecular surfaces.* Biopolymers, vol. 33, no. 2, pages 219–29, 1993.

[Duncan 93b] B S Duncan & A J Olson. *Shape analysis of molecular surfaces.* Biopolymers, vol. 33, no. 2, pages 231–8, 1993.

[Eargle 06] John Eargle & Zaida Luthey-Schulten. *Visualizing the dual space of biological molecules.* Comput Biol Chem, vol. NIL, no. NIL, page NIL, 2006.

[Edelsbrunner 90] H. Edelsbrunner & E. P. Mucke. *Simulation of Simplicity : a technique to cope with degenerate cases in geometric algorithms.* ACM Trans. Graphics, 1990.

[Edelsbrunner 92] H. Edelsbrunner. *Weighted Alpha Shapes.* Report Computer Science UIUCDCS-R-92-1760, U of Illinois, Urbana, IL, 1992.

[Edelsbrunner 94a] H. Edelsbrunner & P. Fu. *Measuring space filling diagrams and voids.* Rapport technique Rept. UIUC-BI-MB-94-01, Molecular Biophysics Group, Beckman Inst. Univ. Illinois at Urbana-Champaign, 1994.

[Edelsbrunner 94b] H. Edelsbrunner & E. P. Mucke. *Three-dimensional alpha shapes.* ACM Trans. Graphics, 1994.

[Edelsbrunner 95a] H. Edelsbrunner. *The union of balls and its dual shape.* Discrete Computational Geometry, 1995.

[Edelsbrunner 95b] Herbert Edelsbrunner, Michael A. Facello, Ping Fu & Jie Liang. *Measuring proteins and voids in proteins.* In HICSS (5), pages 256–264, 1995.

[Edelsbrunner 96] Herbert Edelsbrunner & Nimish R. Shah. *Incremental Topological Flipping Works for Regular Triangulations.* Algorithmica, vol. 15, no. 3, pages 223–241, 1996.

[Edelsbrunner 98] Herbert Edelsbrunner, Michael A. Facello & Jie Liang. *On the Definition and the Construction of Pockets in Macromolecules.* Discrete Applied Mathematics, vol. 88, no. 1-3, pages 83–102, 1998.

[Edelsbrunner 99] Herbert Edelsbrunner. *Deformable Smooth Surface Design.* Discrete & Computational Geometry, vol. 21, no. 1, pages 87–115, 1999.

[Edelsbrunner 00] H. Edelsbrunner, D. Letscher & A. Zomorodian. *Topological persistence and simplification.* In FOCS '00 : Proceedings of the 41st Annual Symposium on Foundations of Computer Science, page 454, Washington, DC, USA, 2000. IEEE Computer Society.

[Edelsbrunner 03] Herbert Edelsbrunner & Patrice Koehl. *The weighted-volume derivative of a space-filling diagram.* Proc Natl Acad Sci U S A, vol. 100, no. 5, pages 2203–8, 2003.

[Exner 98] Th. Exner, M. Keil, G. Moeckel & J. Brickmann. *Identification of Substrate Channels and Protein Cavities.* J.Mol.Mod., vol. 4, pages 340–343, 1998. algo de Sybyl.

[Exner 02a] Thomas E. Exner, Matthias Keil & Jürgen Brickmann. *Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory.* Journal of Computational Chemistry, vol. 23, no. 12, pages 1176–1187, 2002.

[Exner 02b] Thomas E. Exner, Matthias Keil & Jürgen Brickmann. *Pattern recognition strategies for molecular surfaces. II. Surface complementarity.* Journal of Computational Chemistry, vol. 23, no. 12, pages 1188–1197, 2002.

[Eyrisch 07] S. Eyrisch & V. Helms. *Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction.* Journal of Medicinal Chemistry, vol. 50, no. 15, pages 3457–3464, 2007.

[Eyrisch 09] Susanne Eyrisch & Volkhard Helms. *What induces pocket openings on protein surface patches involved in protein-protein interactions ?* J Comput Aided Mol Des, vol. 23, no. 2, pages 73–86, 2009.

[Fischer 94] Emil Fischer. *Einfluss der Configuration auf die Wirkung der Enzyme.* Berichte der deutschen chemischen Gesellschaft, 1894.

[Gabdoulline 96] R R Gabdoulline & R C Wade. *Analytically defined surfaces to analyze molecular interaction properties.* J Mol Graph, vol. 14, no. 6, pages 341–53, 374–5, 1996.

[Gellatly 82] B J Gellatly & J L Finney. *Calculation of protein volumes : an alternative to the Voronoi procedure.* J Mol Biol, vol. 161, no. 2, pages 305–22, 1982.

[Giesen 03] Joachim Giesen & Matthias John. *Computing the Weighted Flow Complex.* In Thomas Ertl, editeur, VMV, pages 235–243. Aka GmbH, 2003.

[Glaser 06] Fabian Glaser, Richard J Morris, Rafael J Najmanovich, Roman A Laskowski & Janet M Thornton. *A method for localizing ligand binding pockets in protein structures.* Proteins, vol. 62, no. 2, pages 479–88, 2006.

[Goede 97] Andrean Goede, Robert Preissner & Cornelius Frömmel. *Voronoi cell : New method for allocation of space among atoms : Elimination of avoidable errors in calculation of atomic volume and density.* Journal of Computational Chemistry, vol. 18, no. 9, pages 1113–1123, 1997.

[Gold 06] Nicola D Gold & Richard M Jackson. *Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships.* J Mol Biol, vol. 355, no. 5, pages 1112–24, 2006.

[Grant 95] J. A. Grant & B. T. Pickup. *A Gaussian Description of Molecular Shape.* Journal of Physical Chemistry, vol. 99, no. 11, pages 3503–3510, 1995.

[Gronemeyer 04] Hinrich Gronemeyer, Jan-Ake Gustafsson & Vincent Laudet. *Principles for modulation of the nuclear receptor superfamily.* Nat Rev Drug Discov, vol. 3, no. 11, pages 950–64, 2004.

[Guilloux 09] Vincent Le Guilloux, Peter Schmidtke & Pierre Tuffery. *Fpocket : an open source platform for ligand pocket detection.* BMC Bioinformatics, vol. 10, no. NIL, page 168, 2009.

[Haffner 04] Curt D Haffner, James M Lenhard, Aaron B Miller, Darryl L McDougald, Kate Dwornik, Olivia R Ittoop, Robert T Jr Gampe, H Eric Xu, Steve Blanchard, Valerie G Montana, Tom G Consler, Randy K Bledsoe, Andrea Ayscue & Dallas Croom. *Structure-based design of potent retinoid X receptor alpha agonists.* J Med Chem, vol. 47, no. 8, pages 2010–29, 2004.

[Hayryan 05] Shura Hayryan, Chin-Kun Hu, Jaroslav Skrivanek, Edik Hayryane & Imrich Pokorny. *A new analytical method for computing solvent-accessible surface area of macromolecules and its gradients.* J Comput Chem, vol. 26, no. 4, pages 334–343, Mar 2005.

[Heiden 05] Wolfgang Heiden & Rita Cornely. *Representation and Analysis of 3D Shape.* In GraphiCon, volume 15, Novosibirsk (Russia), June 2005.

[Hendlich 97] M Hendlich, F Rippmann & G Barnickel. *LIGSITE : automatic and efficient detection of potential small molecule-binding sites in proteins.* J Mol Graph Model, vol. 15, no. 6, pages 359–63, 389, 1997.

[Hendrix 98] D K Hendrix & I D Kuntz. *Surface solid angle-based site points for molecular docking.* Pac Symp Biocomput, vol. NIL, no. NIL, pages 317–26, 1998.

[Horton 06] Robert Horton, A. Moran Laurence, Gray Scrimgeour, Marc Perry & David Rawn. principles of biochemistry horton moran scrimgeour 4th edition. Prentice Hall, 2006.

[Hubbard 98] S R Hubbard, M Mohammadi & J Schlessinger. *Autoregulatory mechanisms in protein-tyrosine kinases.* J Biol Chem, vol. 273, no. 20, pages 11987–90, 1998.

[Humphrey 96] W. Humphrey, A. Dalke & K. Schulten. *VMD : visual molecular dynamics.* J Mol Graph, vol. 14, no. 1, pages 33–8, 27–8, Feb 1996.

[J. Bernauer 05] J. Azé J. Janin J. Bernauer A. Poupon. *A docking analysis of the statistical physics of protein-protein recognition.* Phys Biol., vol. 2, no. 2, pages S17–23, 2005.

[Jackson 02] Richard M. Jackson. *Q-fit : A probabilistic method for docking molecular fragments by sampling low energy conformational space.* Journal of Computer-Aided Molecular Design, vol. 16, no. 1, pages 43–57, 2002.

[Jimenez-Lozano 03] N Jimenez-Lozano, M Chagoyen, J Cuenca-Alba & J M Carazo. *FEMME database : topologic and geometric information of macromolecules.* J Struct Biol, vol. 144, no. 1-2, pages 104–13, 2003.

[Jones 97a] S. Jones. *Prediction of Protein-Protein Interaction Sites using Patch Analysis.* J Mol Biol, pages 133–143, September 1997.

[Jones 97b] S. Jones & J. M. Thornton. *Analysis of protein-protein interaction sites using surface patches.* J Mol Biol, vol. 272, no. 1, pages 121–132, September 1997.

[K 08] Khafizov K, Lattanzi G & Carloni P. *G protein inactive and active forms investigated by simulation methods.* Proteins, vol. 75, no. 4, pages 919–930, 2008.

[Kalidas 08] Yeturu Kalidas & Nagasuma Chandra. *PocketDepth : A new depth based algorithm for identification of ligand binding sites in proteins.* Journal of Structural Biology, 2008.

[Kasson 07] Peter M Kasson, Afra Zomorodian, Sanghyun Park, Nina Singhal, Leonidas J Guibas & Vijay S Pande. *Persistent voids : a new structural metric for membrane fusion.* Bioinformatics, vol. 23, no. 14, pages 1753–9, 2007.

[Kawabata 07] Takeshi Kawabata & Nobuhiro Go. *Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites.* Proteins : Structure, Function, and Bioinformatics, vol. 68, no. 2, pages 516–529, April 2007.

[Keil 04] Matthias Keil, Thomas E. Exner & Jürgen Brickmann. *Pattern recognition strategies for molecular surfaces : III. Binding site prediction with a neural network.* Journal of Computational Chemistry, vol. 25, no. 6, pages 779–789, 2004.

[Kettner 99] Lutz Kettner. *Using generic programming for designing a data structure for polyhedral surfaces.* Comput. Geom. Theory Appl., vol. 13, no. 1, pages 65–90, 1999.

[Kettner 06] Lutz Kettner. *Halfedge Data Structures.* In CGAL Editorial Board, editeur, CGAL-3.2 User and Reference Manual. 2006.

[Kettner 08] Lutz Kettner, Sylvain Pion, & Michael Seel. *Profiling Tools Timers, Hash Map, Union-find, Modifiers.* In CGAL Editorial Board, editeur, CGAL User and Reference Manual. 3.4 edition, 2008.

[Kim 06] Deok-Soo Kim, Jeongyeon Seo, Donguk Kim, Joonghyun Ryu & Cheol-Hyung Cho. *Three-dimensional beta shapes.* Computer-Aided Design, 2006.

[Kim 07] Donguk Kim, Cheol-Hyung Cho, Youngsong Cho, Joonghyun Ryu, Jonghwa Bhak & Deok-Soo Kim. *Pocket extraction on proteins via the Voronoi diagram of spheres.* Journal of Molecular Graphics and Modelling, 2007.

[Kleywegt 94] G J Kleywegt & T A Jones. *Detection, delineation, measurement and display of cavities in macromolecular structures.* Acta Crystallogr D Biol Crystallogr, vol. 50, no. Pt 2, pages 178–85, 1994.

[Kosloff 08] Mickey Kosloff & Rachel Kolodny. *Sequence-similar, structure-dissimilar protein pairs in the PDB.* Proteins, vol. 71, no. 2, pages 891–902, 2008.

[Krishnamoorthy 03] B. Krishnamoorthy & A. Tropsha. *Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations.* Bioinformatics, vol. 19, no. 12, pages 1540–1548, August 2003.

[Kruithof 07a] Nico Kruithof. *3D Skin Surface Meshing.* In CGAL Editorial Board, editeur, CGAL User and Reference Manual. 3.3 edition, 2007.

[Kruithof 07b] Nico Kruithof & Gert Vegter. *Meshing skin surfaces with certified topology.* Comput. Geom., vol. 36, no. 3, pages 166–182, 2007.

[Kuhn 92] L A Kuhn, M A Siani, M E Pique, C L Fisher, E D Getzoff & J A Tainer. *The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures.* J Mol Biol, vol. 228, no. 1, pages 13–22, 1992.

[Laskowski 95] R A Laskowski. *SURFNET : a program for visualizing molecular surfaces, cavities, and intermolecular interactions.* J Mol Graph, vol. 13, no. 5, pages 323–30, 307–8, 1995.

[Laskowski 96] R A Laskowski, N M Luscombe, M B Swindells & J M Thornton. *Protein clefts in molecular recognition and function.* Protein Sci, vol. 5, no. 12, pages 2438–52, 1996.

[Laudet 01] Vincent Laudet & Hinrich Gronemeyer. The nuclear receptor factsbook. Academic Press, 2001.

[Laug 01] P. Laug & H. Borouchaki. *BLMOL, Molecular Surface Mesh Generator.* http ://www-rocq.inria.fr/Patrick.Laug/blmol/index.html, 2001.

[Laug 02] Patrick Laug & Houman Borouchaki. *Molecular Surface Modeling and Meshing.* Eng. Comput. (Lond.), vol. 18, no. 3, pages 199–210, 2002.

[Laug 03] P. Laug & H. Borouchaki. *Generation of finite element meshes on molecular surfaces.* International Journal of Quantum Chemistry, vol. 93, no. 2, pages 131–138, Feb 2003.

[Laurie 05] Alasdair T R Laurie & Richard M Jackson. *Q-SiteFinder : an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, vol. 21, no. 9, pages 1908–16, 2005.

[Lee 71] B Lee & F M Richards. *The interpretation of protein structures : estimation of static accessibility.* J Mol Biol, vol. 55, no. 3, pages 379–400, 1971.

[Lee 05] Chang H. Lee, Amitabh Varshney & David W. Jacobs. *Mesh saliency.* In SIGGRAPH '05 : ACM SIGGRAPH 2005 Papers, pages 659–666, New York, NY, USA, 2005. ACM.

[Levitt 92] D G Levitt & L J Banaszak. *POCKET : a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids.* J Mol Graph, vol. 10, no. 4, pages 229–34, 1992.

[Li 03] X. Li, C. Hu & J. Liang. *Simplicial edge representation of protein structures and alpha contact potential with confidence measure*, 2003.

[Liang 98a] J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudharkar & S. Subramaniam. *Analytic shape computation of macromolecules II : inaccessible cavities in proteins.* Proteins : Structure, Function, and Genetics, vol. 33, pages 18–29, 1998.

[Liang 98b] J Liang, H Edelsbrunner & C Woodward. *Anatomy of protein pockets and cavities : measurement of binding site geometry and implications for ligand design.* Protein Sci, vol. 7, no. 9, pages 1884–97, 1998.

[Maggiora 01] G M Maggiora, D C Rohrer & J Mestres. *Comparing protein structures : a Gaussian-based approach to the three-dimensional structural similarity of proteins.* J Mol Graph Model, vol. 19, no. 1, pages 168–78, 2001.

[Masuya 95] M Masuya & J Doi. *Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations.* J Mol Graph, vol. 13, no. 6, pages 331–6, 1995.

[Mezei 03] Mihaly Mezei. *A new method for mapping macromolecular topography.* J Mol Graph Model, vol. 21, no. 5, pages 463–72, 2003.

[Miller 87] S Miller, J Janin, A M Lesk & C Chothia. *Interior and surface of monomeric proteins.* J Mol Biol, vol. 196, no. 3, pages 641–56, 1987.

[Moulinier 97] Luc Moulinier. *Étude structurale d'un complexe hétérologue : aspartyl-ARNt synthétase d'E. coli-ARNtAsp de levure.* Thèse nouveau doctorat, Université de Strasbourg 1, 1997.

[Mucke 93] E. P. Mucke. *Shapes and Implementations in Three-Dimensional Geometry.* Dept. comput. sci., Univ. Illinois at Urbana-Champaign, 1993. UIUCDCS-R-93-1836.

[Munson 97] P J Munson & R K Singh. *Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment.* Protein Sci, vol. 6, no. 7, pages 1467–81, 1997.

[Myers 95] N. C. Myers. *Traits : a new and useful template technique.* C++ Report, 1995.

[Nayal 06] Murad Nayal & Barry Honig. *On the nature of cavities on protein surfaces : application to the identification of drug-binding sites.* Proteins, vol. 63, no. 4, pages 892–906, 2006.

[Norel 94] R Norel, S L Lin, H J Wolfson & R Nussinov. *Shape complementarity at protein-protein interfaces.* Biopolymers, vol. 34, no. 7, pages 933–40, 1994.

[Norel 95] R Norel, S L Lin, H J Wolfson & R Nussinov. *Molecular surface complementarity at protein-protein interfaces : the critical role played by surface normals at well placed, sparse, points in docking.* J Mol Biol, vol. 252, no. 2, pages 263–73, 1995.

[Norel 99] R Norel, H J Wolfson & R Nussinov. *Small molecule recognition : solid angles surface representation and molecular shape complementarity.* Comb Chem High Throughput Screen, vol. 2, no. 4, pages 223–37, 1999.

[Novotny 86] J Novotny, M Handschumacher, E Haber, R E Bruccoleri, W B Carlson, D W Fanning, J A Smith & G D Rose. *Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains).* Proc Natl Acad Sci U S A, vol. 83, no. 2, pages 226–30, 1986.

[Ogata 96] K Ogata, C Kanei-Ishii, M Sasaki, H Hatanaka, A Nagadoi, M Enari, H Nakamura, Y Nishimura, S Ishii & A Sarai. *The cavity in the hydrophobic core of Myb DNA-binding domain is reserved for DNA recognition and trans-activation.* Nat Struct Biol, vol. 3, no. 2, pages 178–87, 1996.

[Pauling 51a] L Pauling & R B Corey. *The pleated sheet, a new layer configuration of polypeptide chains.* Proc Natl Acad Sci U S A, vol. 37, no. 5, pages 251–6, 1951.

[Pauling 51b] L Pauling, R B Corey & H R Branson. *The structure of proteins ; two hydrogen-bonded helical configurations of the polypeptide chain.* Proc Natl Acad Sci U S A, vol. 37, no. 4, pages 205–11, 1951.

[Pedersen 91] T G Pedersen, B W Sigurskjold, K V Andersen, M Kjaer, F M Poulsen, C M Dobson & C Redfield. *A nuclear magnetic resonance study of the hydrogen-exchange behaviour of lysozyme in crystals and solution.* J Mol Biol, vol. 218, no. 2, pages 413–26, 1991.

[Peters 96] K P Peters, J Fauck & C Frommel. *The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria.* J Mol Biol, vol. 256, no. 1, pages 201–13, 1996.

[Petrek 06] Martin Petrek, Michal Otyepka, Pavel Banas, Pavlina Kosinova, Jaroslav Koca & Jiri Damborsky. *CAVER : a new tool to explore routes from protein clefts, pockets and cavities.* BMC Bioinformatics, vol. 7, no. NIL, page 316, 2006.

[Petřek 07] Martin Petřek, Pavlína Košinová, Jaroslav Koča & Michal Otyepka. *MOLE : A Voronoi Diagram-Based Explorer of Molecular Channels, Pores, and Tunnels.* Structure, vol. 15, no. 11, pages 1357–1363, November 2007.

[Pintar 02] Alessandro Pintar, Oliviero Carugo & Sandor Pongor. *CX, an algorithm that identifies protruding atoms in proteins.* Bioinformatics, vol. 18, no. 7, 2002.

[Pintar 03a] Alessandro Pintar, Oliviero Carugo & Sandor Pongor. *Atom depth in protein structure and function.* Trends Biochem Sci, vol. 28, no. 11, pages 593–7, 2003.

[Pintar 03b] Alessandro Pintar, Oliviero Carugo & Sandor Pongor. *DPX : for the analysis of the protein core.* Bioinformatics, vol. 19, no. 2, 2003.

[Pion 07] Sylvain Pion & Monique Teillaud. *3D Triangulations.* In CGAL Editorial Board, editeur, CGAL User and Reference Manual. 3.3 edition, 2007.

[Pontius 96] J. Pontius, J. Richelle & S.J. Wodak. *Deviations from standard atomic volumes as a quality measure for protein crystal structures.* Journal of Molecular Biology, vol. 264, pages 121–136, 1996.

[Povray 96] Povray. *Povray - The Persistence of Vision Ray Tracer*, 1996.

[Ray 05] Nicolas Ray, Xavier Cavin, Jean-Claude Paul & Bernard Maigret. *Intersurf : dynamic interface between proteins.* J Mol Graph Model, vol. 23, no. 4, pages 347–354, Jan 2005.

[Richards 74] F M Richards. *The interpretation of protein structures : total volume, group volume distributions and packing density.* J Mol Biol, vol. 82, no. 1, pages 1–14, 1974.

[Richards 77] F M Richards. *Areas, volumes, packing and protein structure.* Annu Rev Biophys Bioeng, vol. 6, no. NIL, pages 151–76, 1977.

[Richmond 84] T. J. Richmond. *Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect.* J Mol Biol, vol. 178, no. 1, pages 63–89, Sep 1984.

[Ritchie 99] David W. Ritchie & Graham J. L. Kemp. *Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces.* Journal of Computational Chemistry, vol. 20, no. 4, pages 383–395, February 1999.

[Ritchie 00] David W. Ritchie & Graham J. L. Kemp. *Protein docking using spherical polar Fourier correlations.* Proteins : Structure, Function, and Genetics, vol. 39, no. 2, pages 178–194, March 2000.

[Rother 03] Kristian Rother, Robert Preissner, Andrean Goede & Cornelius Frömmel. *Inhomogeneous molecular density : reference packing densities and distribution of cavities within proteins.* Bioinformatics, vol. 19, no. 16, pages 2112–2121, 2003.

[Sanner 92] Michel F. Sanner. *Sur la modélisation des surfaces moléculaires.* Informatique, Université de Mulhouse, 1992.

[Sanner 95] Michel F. Sanner, Arthur J. Olson & Jean-Claude Spehner. *Fast and Robust Computation of Molecular Surfaces.* In Symposium on Computational Geometry, pages C6–C7, 1995.

[Sanner 96] M. F. Sanner, A. J. Olson & J. C. Spehner. *Reduced surface : an efficient way to compute molecular surfaces.* Biopolymers, vol. 38, no. 3, pages 305–320, Mar 1996.

[Schroeder 96] William J. Schroeder, Kenneth M. Martin & William E. Lorensen. *The design and implementation of an object-oriented toolkit for 3D graphics and visualization.* In VIS '96 : Proceedings of the 7th conference on Visualization '96, pages 93–ff., Los Alamitos, CA, USA, 1996. IEEE Computer Society Press.

[Shiue 07] Le-Jeng Andy Shiue. *3D Surface Subdivision Methods.* In CGAL Editorial Board, editeur, CGAL User and Reference Manual. 3.3 edition, 2007.

[Singh 96] R. K. Singh, A. Tropsha & I. I. Vaisman. *Delaunay tessellation of proteins : four body nearest-neighbor propensities of amino acid residues.* J Comput Biol, vol. 3, no. 2, pages 213–221, 1996.

[Smart 93] O S Smart, J M Goodfellow & B A Wallace. *The pore dimensions of gramicidin A.* Biophys J, vol. 65, no. 6, pages 2455–60, 1993.

[Stahl 00] Martin Stahl, Chiara Taroni & Gisbert Schneider. *Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network.* protein Engineering, vol. 2, page 8, 2000.

[Taylor 06] Todd J. Taylor & Iosif I. Vaisman. *Protein Structural Domain Assignment with a Delaunay Tessellation Derived Lattice.* In ISVD '06 : Proceedings of the 3rd International Symposium on Voronoi Diagrams in Science and Engineering, pages 232–240, Washington, DC, USA, 2006. IEEE Computer Society.

[Thornton 86] J. M. Thornton, M. S. Edwards, W. R. Taylor & D. J. Barlow. *Location of 'continuous' antigenic determinants in the protruding regions of proteins.* EMBO J, vol. 5, no. 2, pages 409–413, February 1986.

[Tondel 02] Kristin Tondel, Endre Anderssen & Finn Drablos. *Protein Alpha Shape Similarity Analysis (PASSA) : a new method for mapping protein binding sites. Application in the design of a selective inhibitor of tyrosine kinase 2.* J Comput Aided Mol Des, vol. 16, no. 11, pages 831–40, 2002.

[Tondel 06] Kristin Tondel, Endre Anderssen & Finn Drablos. *Protein Alpha Shape (PAS) Dock : a new gaussian-based score function suitable for docking in homology modelled protein structures.* J Comput Aided Mol Des, vol. 20, no. 3, pages 131–44, 2006.

[Totrov 96] M. Totrov & R. Abagyan. *The contour-buildup algorithm to calculate the analytical molecular surface.* J Struct Biol, vol. 116, no. 1, pages 138–143, 1996.

[Troffer-Charlier 07] Nathalie Troffer-Charlier, Vincent Cura, Pierre Hassenboehler, Dino Moras & Jean Cavarelli. *Functional insights from structures of coactivator-associated arginine methyltransferase 1 domains.* EMBO J, vol. 26, no. 20, pages 4391–401, 2007.

[Tsai 99] J Tsai, R Taylor, C Chothia & M Gerstein. *The packing density in proteins : standard radii and volumes.* J Mol Biol, vol. 290, no. 1, pages 253–66, 1999.

[Tsai 02] Jerry Tsai & Mark Gerstein. *Calculations of protein volumes : sensitivity analysis and parameter database.* Bioinformatics, vol. 18, no. 7, pages 985–95, 2002.

[Tsodikov 02] Oleg V Tsodikov, M Thomas Jr Record & Yuri V Sergeev. *Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature.* J Comput Chem, vol. 23, no. 6, pages 600–9, 2002.

[Vaisman 98] Iosif I. Vaisman, Er Tropsha & Weifan Zheng. *Compositional preferences in quadruplets of nearest neighbor residues in protein structures : statistical geometry analysis.* In Proceedings of the IEEE Symposia on Intelligence and Systems, pages 163–168, 1998.

[van Hoorn 02] Willem P van Hoorn. *Identification of a second binding site in the estrogen receptor.* J Med Chem, vol. 45, no. 3, pages 584–9, 2002.

[Varshney 93] Amitabh Varshney & Frederick P. Brooks Jr. *Fast Analytical Computation of Richard's Smooth Molecular Surface.* In Gregory M. Nielson & R. Daniel Bergeron, editeurs, IEEE Visualization, pages 300–307. IEEE Computer Society, 1993.

[Via 00] A Via, F Ferre, B Brannetti & M Helmer-Citterich. *Protein surface similarities : a survey of methods to describe and compare protein surfaces.* Cell Mol Life Sci, vol. 57, no. 13-14, pages 1970–7, 2000.

[Vorobjev 97] Y N Vorobjev & J Hermans. *SIMS : computation of a smooth invariant molecular surface.* Biophys J, vol. 73, no. 2, pages 722–32, 1997.

[Voronoi 07] G. Voronoi. *Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier Mémoire : Sur quelques propriétés des formes quadratiques parfaites.* J. Reine Angew. Math., vol. 133, pages 97–178, 1907.

[Voronoi 08] G. Voronoi. *Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième Mémoire : Recherches sur le parallélloèdres primitifs.* J. Reine Angew. Math., vol. 134, pages 198–287, 1908.

[Voronoi 09] G. Voronoi. *Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième Mémoire : Recherches sur le parallélloèdres primitifs.* J. Reine Angew. Math., vol. 136, pages 67–181, 1909.

[Weinmann 00] Serge Weinmann & Pierre Méhul. Biochimie. structure et fonction des protéines. Sciences Sup. Dunod, 2000.

[Weisel 07] Martin Weisel, Ewgenij Proschak & Gisbert Schneider. *PocketPicker : analysis of ligand binding-sites with shape descriptors.* Chemistry Central Journal, vol. 1, page 7, 2007.

[Wensheng 02] Cai. Wensheng, Shao. Xueguang & Maigret. Bernard. *Protein-ligand recognition using spherical harmonic molecular surfaces : towards a fast and efficient filter for large virtual throughput screening.* Journal of molecular graphics and modelling, vol. 20, no. 4, pages 313–28, Jan 2002.

[Wernisch 99] L Wernisch, M Hunting & S J Wodak. *Identification of structural domains in proteins by a graph heuristic.* Proteins, vol. 35, no. 3, pages 338–52, 1999.

[Wesson 92] L Wesson & D Eisenberg. *Atomic solvation parameters applied to molecular dynamics of proteins in solution.* Protein Sci, vol. 1, no. 2, pages 227–35, 1992.

[Yeates 95] T O Yeates. *Algorithms for evaluating the long-range accessibility of protein surfaces.* J Mol Biol, vol. 249, no. 4, pages 804–15, 1995.

[Zhang 07] Xiaoyu Zhang & Chandrajit Bajaj. *Extraction, quantification and visualization of protein pockets.* Comput Syst Bioinformatics Conf, vol. 6, no. NIL, pages 275–86, 2007.

[Zhao 05] Lei Zhao & Jean Chmielewski. *Inhibiting protein-protein interactions using designed molecules.* Curr Opin Struct Biol, vol. 15, no. 1, pages 31–4, 2005.

[Zomorodian 06] Afra Zomorodian, Leonidas Guibas & Patrice Koehl. *Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials.* Computer Aided Geometric Design, vol. 23, pages 531–544, August 2006.

Résumé

NOTRE TRAVAIL s'inscrit dans le cadre le la bioinformatique structurale, et plus spécifiquement à l'interface entre biologie structurale et géométrie algorithmique, dont nous empruntons les constructions issues de la théorie des formes-α pour représenter et étudier les molécules. L'objectif global de notre travail est de proposer de nouveaux outils théoriques et pratiques destinés à favoriser l'étude de la relation structure—fonction des macromolécules biologiques ; et nous nous intéressons plus particulièrement à caractériser les lieux d'une intéraction possible à la surface de ces molécules.

Dans le cadre de cette étude nous proposons un nouveau modèle, la Surface duale, un encodage variété de la combinatoire de la Surface Accessible qui favorise le parcours de cette dernière et permettant notamment de constituer des voisinages d'atomes à la surface de la molécule. Nous avons utilisé ce modèle dans l'ensemble de nos travaux, qui peuvent être décomposés selon trois axes : (a) une caractérisation topographique de la surface moléculaire au travers d'une mesure d'incurvation permettant d'y définir des zones protrudentes et des zones anfractuées, (b) la définition de propriétés utilisables dans le cadre d'une caractérisation et d'une prédiction des surfaces d'interaction entre protéines, (c) la détection et la caractérisation des poches, crevasses et cavités dans les macromolécules biologiques.

Nos travaux ont été mis en pratiques dans deux logiciels mis à disposition de la communauté scientifique : `Lc` et `Pck`, respectivement pour la description topographique de la surface moléculaire et pour la détection et la caractérisation des poches dans les macromolécules biologiques.

Mots-clefs : bioinformatique structurale, géométrie algorithmique, formes-α, biologie structurale, poches, topographie, modèles moléculaires.

Abstract

OUR STUDY is concerned with structural bioinfomatics (aka computational biology), more specifically, we borrow models from the α-shape theory to represent and study molecules. Roughly, our aim is to provide new theoretical and practical tools to ease the study of structure-function relationship in biological molecules. We are more specifically interested in characterising the usual locations of a possible interaction at the surface of such molecules.

In this context we propose a novel model, the *dual surface*, that constitutes a manifold polyhedral surface encoding the Surface Accessible. This construction eases the the construction of continuous surface tracks at the surface of a molecule, and therefore allows notably, the construction of molecular surface patches. We adapted this model mainly to address three distinct problems : (a) the proposal of a novel index to describe the molecular surface landscape in terms of knobs and clefts, (b) the definition of surface descriptors that can be used to study interacting patches on a protein surface, (c) the detection and characterisation of cavities, pockets, clefts and crevices at the surface of macromolecules.

Two software tools were developped based on these works and are now freely accessible to the scientific community : `Lc` and `Pck`, respectively dedicated to the description of the molecular surface topography, and to the detection and characterisation of pockets in molecular structures.

Keywords: Structural bioinformatics, computational biology, α-shape, pockets, molecular surface topography, structural biology .