UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Laboratoire d'Innovation Thérapeutique – UMR7200

THÈSE présentée par :

Célien JACQUEMARD

soutenue le : **14 janvier 2021**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Chimie informatique et théorique

Étude par modélisation moléculaire de la reconnaissance du corécepteur CCR5 du VIH-1 par la glycoprotéine virale gp120

THÈSE dirigée par :

Mme KELLENBERGER Esther	Professeur, Université de Strasbourg

RAPPORTEURS :

Mr XHAARD Henri	Docteur, Université de Helsinki
Mme SOPKOVA-DE OLIVEIRA SANTOS Jana	Professeur, Université de Caen

---

AUTRES MEMBRES DU JURY :

Mr MARCOU Gilles	Maître de conférences, Université de Strasbourg

Remerciements

Tellement de personnes à remercier. Mais le temps me fait défaut et je témoignerai ma gratitude à toutes les personnes qui ont fait partie de ma vie, m'ont aidé pendant cette thèse comme il se doit dans mon manuscrit final.

Néanmoins, je tiens à exprimer mes remerciements aux membres du jury qui me font l'honneur de juger mes travaux, à savoir le Dr. Henry XHAARD de l'Université d'Helsinki, le Pr. Jana SOPKOVA-DE OLIVEIRA SANTOS de l'Université de Caen et le Dr. Gilles MARCOU de l'Université de Strasbourg. Et bien entendu, je tiens à montrer tout mon respect au Pr. Esther KELLENBERGER qui m'a encadré avec un dévouement sans pareil depuis mon arrivée au sein du laboratoire.

Direction de la **recherche**
et de la **valorisation**
Collège **doctoral**
Université de Strasbourg

Déclaration sur l'honneur

Declaration of Honour

J'affirme être informé que le plagiat est une faute grave susceptible de mener à des sanctions administratives et disciplinaires pouvant aller jusqu'au renvoi de l'Université de Strasbourg et passible de poursuites devant les tribunaux de la République Française.

Je suis conscient(e) que l'absence de citation claire et transparente d'une source empruntée à un tiers (texte, idée, raisonnement ou autre création) est constitutive de plagiat.

Au vu de ce qui précède, **j'atteste sur l'honneur que le travail décrit dans mon manuscrit de thèse est un travail original et que je n'ai pas eu recours au plagiat ou à toute autre forme de fraude.**

I affirm that I am aware that plagiarism is a serious misconduct that may lead to administrative and disciplinary sanctions up to dismissal from the University of Strasbourg and liable to prosecution in the courts of the French Republic.

I am aware that the absence of a clear and transparent citation of a source borrowed from a third party (text, idea, reasoning or other creation) is constitutive of plagiarism.

In view of the foregoing, ***I hereby certify that the work described in my thesis manuscript is original work and that I have not resorted to plagiarism or any other form of fraud.***

Nom : Jacquemard **Prénom : Célien**

Ecole doctorale : ED222

Laboratoire : Laboratoire d'Innovation Thérapeutique (UMR7200)

Date : 16/06/2021

Signature :

Table des matières

Table des figures

Liste des tableaux

Abbréviations

[5P7]CCL5 chimiokine (C-C motif) ligand 5 modifiée.

3D 3 dimensions.

7TM Domaine à sept hélices transmembranaire.

ADN Acide désoxyribonucléique.

ADRB2 Récepteur β2-adrénergique.

ALDR Aldo-keto reductase.

ARN Acide ribonucléique.

ARNm Acide ribonucléique messager.

AUC Area under the curve.

AZT Zidovudine.

BACE1 Beta-secretase 1.

bNAb Anticorps neutralisant à spectre large.

CAH2 Carbonic anhydrase 2.

CCL3 Chimiokine (C-C motif) ligand 3.

CCR5 Récepteur à C-C chimiokine de type 5.

CD4 Cluster de différenciation 4.

CDK2 Cyclin-dependant kinase 2.

CHK1 Protein kinase Chk1.

CHL Cholestérol.

cryo-EM Cryo-microscopie électronique.

CXCR4 Récepteur à C-X-C chimiokine de type 4.

Cα Carbone en position α.

DUD-E Database of Useful (Docking) Decoys – Enhanced.

ECL Boucle extracellulaire.

EM Microscopie électronique.

ESR1 Estrogen receptor alpha.

FDA Food Drug Administration.

gp120 Glycoprotéine 120.

gp41 Glycoprotéine 41.

GRIA4 Glutamate receptor.

HSP Heat shock protein.

HYD Hydrophobicity.

Hydrogen bond.

HYES epoxide hydrolase 2.

ICL Boucle intracellulaire.

IFP Interaction fingerprint.

IHME Institute for Health Metrics and Evaluation.

IN2P3 Institut national de physique nucléaire et de physique des particules.

IPA Interaction pseudo-atom.

LKHA4 Leukotriene A-4 hydrolase.

MMP12 Macrophage metalloelastase.

MVC Maraviroc.

OMS Organisation mondiale de la Santé.

ONUSIDA Programme commun des Nations Unies sur le VIH/SIDA.

OPM Orientation of proteins in membrane database.

PDB Protein data bank.

PDE10 Phosphodiesterase 10A.

POPC Phosphatidylcholine.

POPE Phosphatidyléthanolamine.

RCPG Récepteur couplé aux protéines G.

RMN Résonance magnétique nucléaire.

RMSD Écart quadratique moyen.

ROC Receiver operating characteristic.

RX Cristallographie aux rayons X.

SIDA Syndrome de l'immunodéficience acquise.

tat Protéine activatrice de transcription.

TCD4 Lymphocyte T CD4+.

TI Transcriptase inverse.

TM Transmembranaire.

TM1 Hélice transmembranaire 1.

TM2 Hélice transmembranaire 2.

TM3 Hélice transmembranaire 3.

TM4 Hélice transmembranaire 4.

TM5 Hélice transmembranaire 5.

TM6 Hélice transmembranaire 6.

TM7 Hélice transmembranaire 7.

Tys sulfotyrosine.

V1 Boucle variable 1.

V2 Boucle variable 2.

V3 Boucle variable 3.

V4 Boucle variable 4.

V5 Boucle variable 5.

VIH Virus de l'immunodéficience humaine.

VIH-1 Virus de l'immunodéficience humaine de type 1.

VIH-2 Virus de l'immunodéficience humaine de type 2.

VIS Virus de l'immunodéficience simienne.

Chapitre 1

Introduction générale

Les virus, ces entités microscopiques font frémir l'humanité depuis des siècles. Bien que la plupart d'entre eux soient inoffensifs ou tout du moins non mortels pour une majorité de personnes, des pathogènes virulents ont émergé dans l'histoire récente, provoquant des pandémies et causant des millions de morts. Parmi les virus les plus tristement célèbres, on peut citer les orthopoxvirus avec le virus de la variole ou les virus influenza responsables entre autre de la grippe espagnole en 1918. Les virus ayant la capacité de muter facilement, de nouvelles souches virales hautement pathogènes continuent à apparaître au cours du XXI^{ème} siècle. Cependant, les progrès amorcés au siècle précédent dans le domaine de la médecine et plus particulièrement en vaccination permettent de contenir les pandémies voire d'éradiquer certains virus. Mais un virus découvert dans les années 80 continue de décimer les populations et fait l'objet de recherches intensives : le virus de l'immunodéficience humaine (VIH).

1.1 Le syndrome de l'immunodéficience acquise

1.1.1 Généralités

Le VIH est responsable du syndrome de l'immunodéficience acquise (SIDA). Le virus s'attaque aux cellules immunitaires de l'organisme dont la quantité diminue au cours du temps au fur et mesure que le nombre de copies du virus augmente (FIGURE 1.1). Quand le nombre de cellules immunitaires devient trop bas pour assurer leur rôle dans la défense de l'organisme, l'infection atteint le stade SIDA. Dès lors, des maladies opportunistes font leur apparition et conduisent irrémédiablement au décès de la personne suite à des complications telles qu'une pneumonie. Ce stade s'étale sur une durée d'environ un an en l'absence de traitement antirétroviral (taux de survie médian : 300 jours en 1985) [1].

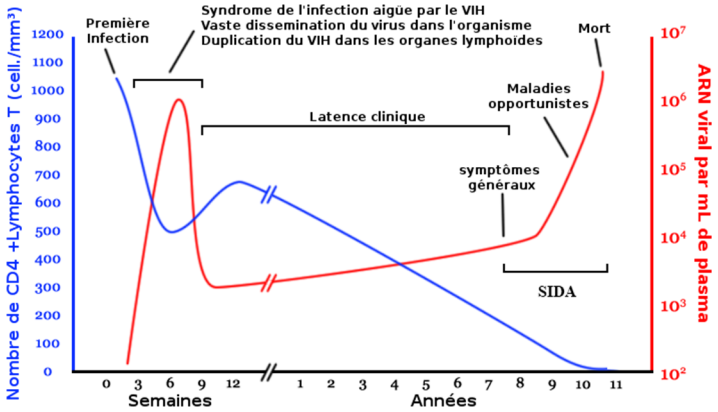

FIGURE 1.1 – Évolution de l'infection par le VIH au cours du temps. Image créée par SANAO sous licence CC BY-SA 3.0.

À ce jour, l'infection par le VIH constitue toujours une pandémie, touchant tous les pays du globe avec une sur-représentation en Afrique sub-saharienne et en Inde (FIGURE 1.2). En 2020, le site du programme commun des Nations Unies sur le VIH/SIDA (ONUSIDA) estime à 38 millions le nombre de personnes vivant avec le VIH et à 700 000 le nombre de personnes décédées en 2019 des suites de maladies contractées au stade SIDA. On compte 1,7 millions d'individus nouvellement infectés chaque année. Les politiques publiques mondiales, coordonnées par l'Organisation mondiale de la Santé (OMS), multiplient les efforts pour que le nombre

de morts et d'individus nouvellement infectés diminue d'année en année (objectifs 90-90-90 de l'OMS).

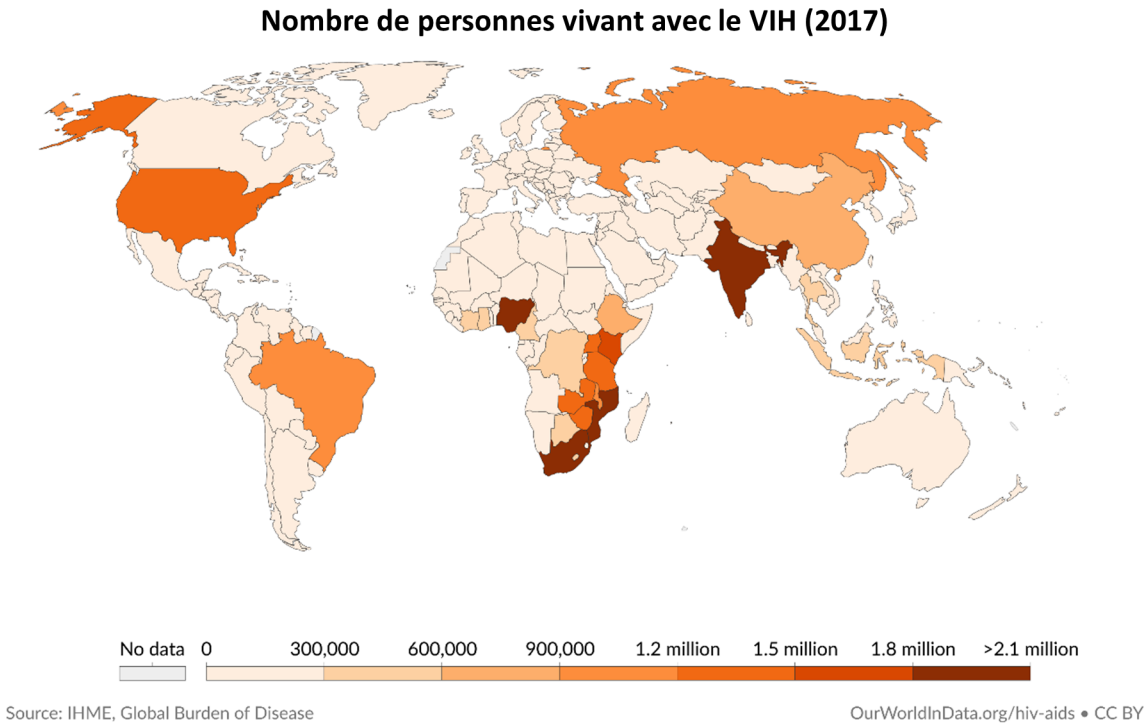

FIGURE 1.2 – Répartition des personnes infectées par le VIH dans le monde. L'image est issue du site `ourworldindata.org` et basée sur les chiffres de l'Institute for Health Metrics and Evaluation (IHME).

1.1.2 Traitements antiviraux

Lors de la découverte du VIH en 1983, aucun traitement n'était disponible pour les malades du SIDA, laissant aux personnes ayant été infectées par ce virus une espérance d'une dizaine d'années à partir de la primo-infection [2].

Le premier médicament antirétroviral approuvé par la Food and Drug Administration (FDA) fut la Zidovudine (AZT) en 1987 [3]. Malgré de sévères effets secondaires tels que l'anémie, l'AZT a quasiment triplé le taux de survie des patients ayant développé les symptômes du SIDA [4]. Vers le milieu des années 90, la découverte d'autres médicaments a permis de mettre en place la tri-thérapie. Cette combinaison de trois médicaments différents a augmenté considérablement l'espérance de vie des personnes infectées, offrant de plus une possibilité de traitement avant le stade SIDA. De nos jours, les antirétroviraux confèrent aux patients traités

précocement une espérance de vie proche de la normale, reléguant l'infection par VIH au rang de maladie chronique [5]. De plus, les thérapies anti-VIH maintiennent la charge virale des personnes traitées à un niveau tellement bas que les chances de transmission à un autre individu sont très faibles [6, 7]. Cependant, le traitement doit être pris quotidiennement, avec une observance stricte, et à vie. Il peut aussi entraîner des effets secondaires indésirables, et représente un coût non négligeable pour le patient et les organismes de santé.

Si les traitements anti-VIH empêchent le développement du SIDA, ils ne guérissent pas de l'infection par le VIH, car ils n'éradiquent pas complètement le virus de l'organisme. Le VIH se cache dans des réservoirs, et recommence à se multiplier dès l'arrêt du traitement. Il existe néanmoins deux cas de guérison complète : le patient de Berlin en 2012 et le patient de Londres en 2020 [8, 9]. Dans les deux cas, une greffe de moelle osseuse a permis la production de cellules immunitaires non reconnues par le virus, à cause d'une mutation génétique chez le donneur provoquant le défaut de présentation du corécepteur à la surface de ces cellules. Cette méthode n'est cependant pas applicable à grande échelle, d'une part à cause de la difficulté de trouver un donneur compatible portant cette mutation particulière, et d'autre part à cause de la lourdeur de l'intervention chirurgicale.

Depuis les années 80, les chercheurs ont également tenté de développer des vaccins protégeant les populations contre le VIH. Cependant, aucune stratégie efficace n'est disponible actuellement. Une explication à cet échec réside dans la capacité de mutation exceptionnelle du VIH, la meilleure parmi toutes les entités biologiques [10]. Ainsi, il existe une population de VIH très diverse au sein d'un même individu rendant la vaccination classique inopérante [11]. Récemment, une nouvelle approche basée sur des anticorps a été proposée pour cibler soit des régions conservées du virus par des anticorps neutralisants à spectre large (broadly neutralizing antibody en anglais, bNAb) soit directement les cellules hôtes pour bloquer l'entrée virale [12, 13]. Des essais cliniques ont montré l'efficacité des anticorps dans la prévention de ou la lutte contre une infection par VIH. D'autres essais sont actuellement en cours afin de vérifier si un vaccin peut induire la production de ces anticorps.

Cette capacité du VIH à muter très rapidement complique le développement de thérapies efficaces chez tous les patients et dans la durée. C'est pour cette raison qu'un cocktail de mé-

dicaments est toujours employé, afin de limiter la sélection de souches résistantes à un principe actif antirétroviral, mais celles-ci finissent néanmoins toujours par apparaître tôt au tard.

1.1.3 Le virus de l'immunodéficience humaine

Classification

Comme écrit précédemment, le VIH mute rapidement lui conférant une variabilité de séquence importante. Une classification a été établie sur la base de l'arbre phylogénétique comprenant également les virus de l'immunodéficience simienne (VIS) qui est à l'origine du VIH [14]. Il existe deux types majeurs du VIH : le type 1 (VIH-1) répandu sur toute la surface du globe et responsable de la pandémie, et le type 2 (VIH-2) localisé essentiellement en Afrique de l'ouest et moins virulent que son homologue de type 1 [15]. Dans ce manuscrit, je ne vais faire référence qu'au type 1. Le VIH-1 est divisé en quatre groupes : M (Major), N (non-M et non-O), O (Outlier) et P. Le groupe M est lui-même divisé en douze sous-types nommés par une lettre allant de A à L. La prévalence de chaque sous-type dépend de la zone géographique (FIGURE 1.3). De plus, il est possible que lors de double infection par des groupes différents, les matériels génétiques se combinent pour donner naissance à de nouvelles lignées [16, 17].

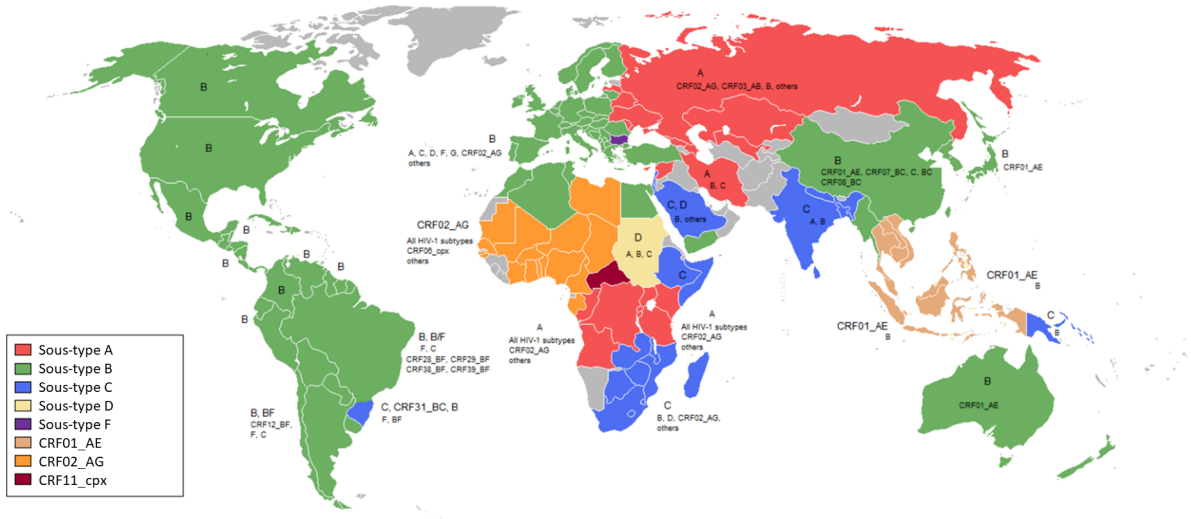

FIGURE 1.3 – Répartition géographique des sous-types majoritaires de VIH-1 dans le monde. Image issue de la référence [18].

Structure du VIH-1

Malgré la grande variabilité dans la séquence de son matériel génétique, le virus garde toujours la même structure. Il est de forme sphérique avec un diamètre moyen de 145 nm. Le VIH-1 est un virus enveloppé, caractérisé par une bicouche lipidique issue de la cellule hôte où a eu lieu la réplication et qui contient des protéines humaines embarquées et un complexe de glycoprotéines virales nommé *env* (FIGURE 1.4). Le complexe *env* est impliqué dans l'attachement du virus à la cellule hôte et dans la pénétration de la cellule qui s'en suit. À l'intérieur du virus, la capside enferme le matériel génétique viral, à savoir deux exemplaires d'acide ribonucléique (ARN) simple brin. D'autres protéines virales sont également présentes : la transcriptase inverse (TI), l'intégrase, la protéase ou encore la protéine activatrice de transcription (tat), chacune jouant un rôle particulier dans le cycle de réplication du virus.

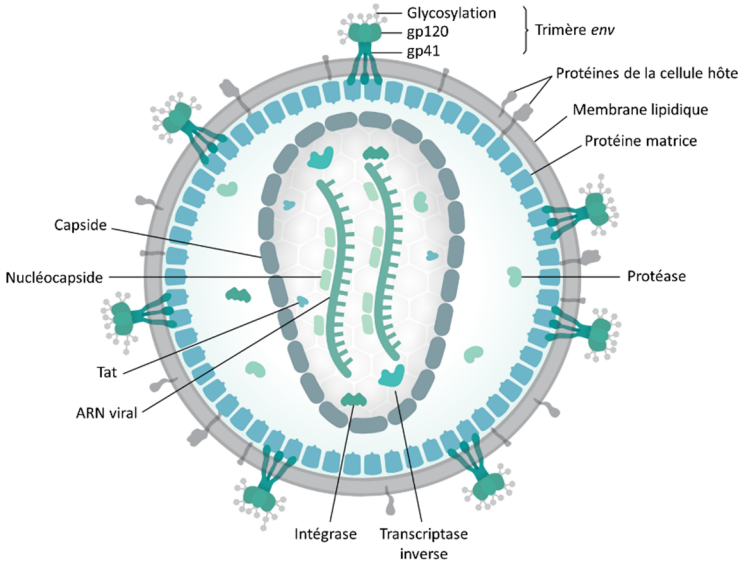

FIGURE 1.4 – Structure du VIH-1. Image créée par Thomas SPLETTSTOESSER (`www.scistyle.com`) sous licence CC BY-SA 4.0. L'image a été revue.

Cycle de réplication du VIH-1

Une fois le virus en circulation dans le système sanguin, il lui faut tout d'abord reconnaître la cellule hôte. Pour cela, il se sert des glycoprotéines à sa surface qui reconnaissent un récepteur et un corécepteur cellulaire. Quand ces protéines interagissent, la fusion des membranes est

déclenchée et s'en suit le déversement du contenu du virus à l'intérieur de la cellule. Dans la cellule, l'ARN viral est retrotranscrit par la TI pour produire l'acide désoxyribonucléique (ADN) double-brin proviral. C'est lors de cette étape que de nouvelles mutations apparaissent dans la séquence, la TI faisant beaucoup d'erreurs car elle est dépourvue d'activité de relecture. La TI est une cible de choix, utilisée par plusieurs des médicaments antirétroviraux tel que l'AZT. Une fois l'ADN proviral produit, l'intégrase se charge de l'incorporer dans le matériel génétique de la cellule à l'intérieur du noyau cellulaire. Par la suite, la machinerie cellulaire pourra transcrire les gènes proviraux et traduire les acides ribonucléiques messagers (ARNm) en polyprotéines virales. La protéase virale coupe ces chaînes polypeptidiques produisant les protéines virales matures, incorporées dans les nouveaux virions formés à la membrane. Les virions sont relargués dans le système sanguin détruisant au passage la cellule hôte.

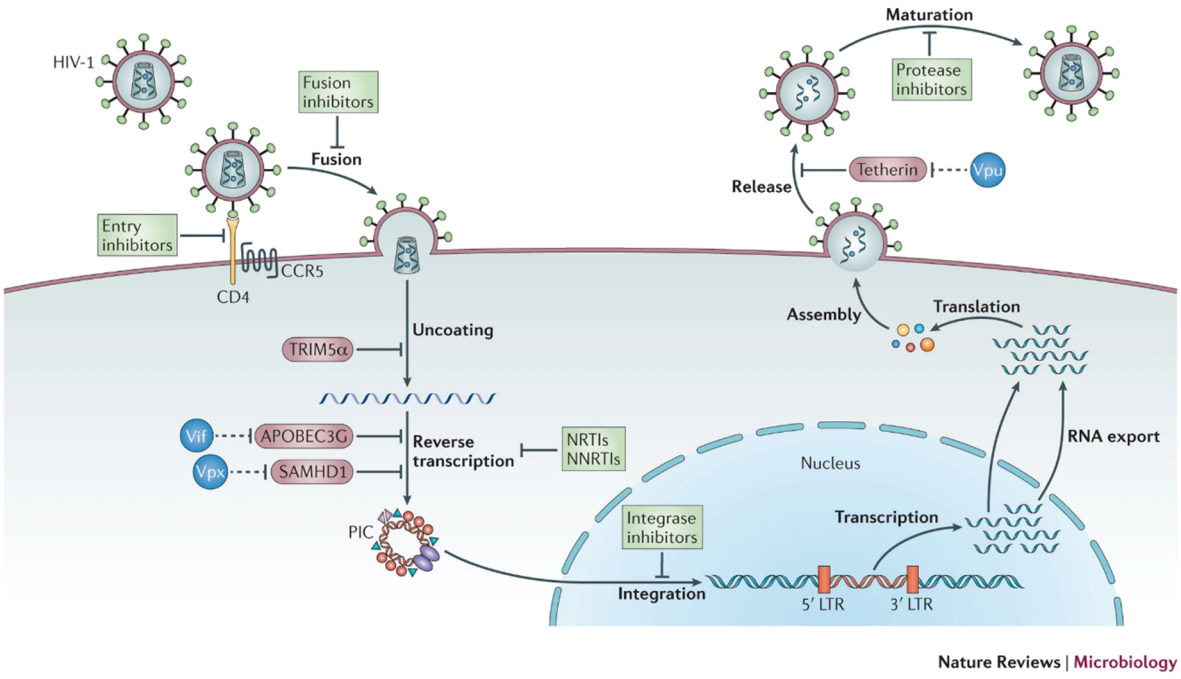

FIGURE 1.5 – Cycle de réplication du VIH-1. Image issue de la référence [19].

1.2 L'entrée virale

1.2.1 Mécanisme

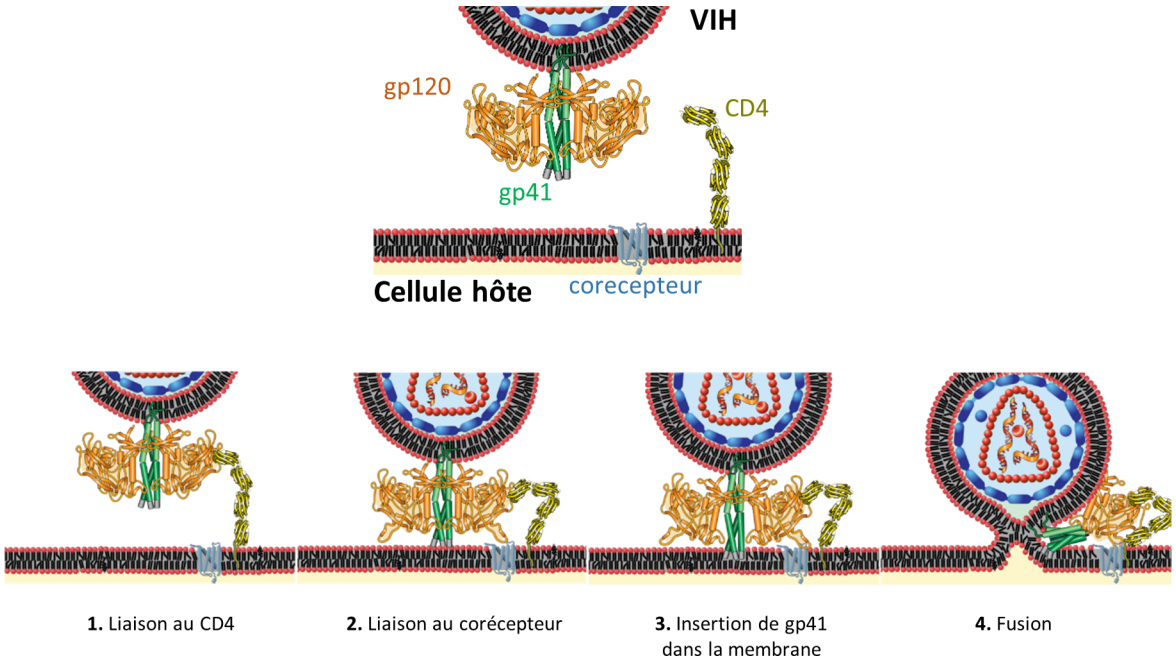

FIGURE 1.6 – Mécanisme de l'entrée de virale. Image créée par Mike JONES sous licence CC BY-SA 3.0. L'image a été revue.

L'entrée du VIH-1 dans la cellule hôte est un processus complexe et dynamique (FIGURE 1.6). Pour cela, le virus dispose d'un outil remarquable à sa surface, les protéines de l'enveloppe virale *env*. L'*env* est un trimère symétrique en forme de spicule : trois protéines appelées glycoprotéine 41 (gp41) forment une tige centrale entourée de trois protéines appelées glycoprotéine 120 (gp120). La base du trimère de gp41 permet d'ancrer l'*env* dans la membrane virale et la gp120 est exposée pour la reconnaissance séquentielle et spécifique de deux protéines membranaires de la cellule hôte, le cluster de différenciation 4 (CD4) et un corécepteur. Le corécepteur majoritaire est le récepteur à C-C chimiokine de type 5 (CCR5). La liaison de la gp120 au CD4 et au corécepteur attache le virus à la cellule hôte, rapprochant les membranes, permettant leur fusion et l'introduction du noyau viral dans le cytoplasme [20].

La structure tridimensionnelle (3D) de l'*env* est très flexible. Dans l'état de base avant la fusion, l'*env* adopte principalement une forme fermée [21]. Lorsque la gp120 est liée au CD4, sa structure 3D s'ouvre à différents degrés [22, 23, 24]. Lors de la liaison au CD4, un mouvement concerté des boucles variables 1 et 2 (V1, V2) de la gp120 découvre la boucle variable 3 (V3) qui est initialement repliée. La libération de la V3 conditionne la liaison de la gp120 au corécepteur [25, 26, 27]. Le mode de liaison de la gp120 au CCR5 implique deux sites, imitant ainsi le mode de liaison de la chimiokine endogène du récepteur. La V3 de la gp120 s'insère dans le domaine transmembranaire du corécepteur et le feuillet de pontage ou bridging sheet de la gp120 est en contact avec le domaine N-terminal du corécepteur.

1.2.2 Structure de l'*env*

La microscopie électronique (EM), la résonance magnétique nucléaire (RMN) et la cristallographie aux rayons X (RX) ont fourni une quantité importante d'informations sur l'organisation de l'*env* à l'échelle moléculaire et atomique, libre et liée à son récepteur et son corécepteur. La base de données publique des structures 3D, la Protein Data Bank (PDB), affiche actuellement plus d'une centaine de structures décrivant l'*env*, montrant des conformations allant de la forme ouverte à la forme fermée du trimère [28, 29]. Toutes les structures 3D de l'*env*, à une exception près, correspondent à des protéines virales qui ont été mutées pour augmenter la stabilité du trimère. La construction la plus couramment rencontrée est appelée SOSIP, pour les deux mutations "SOS" créant des ponts disulfure entre les protomères (A501C dans la gp120 et T605C dans la gp41) et la mutation "IP" limitant la dynamique de la tige (I559P dand la gp41). La seule structure de l'*env* native décrit une protéine immature, c'est-à-dire que la gp160 n'est pas clivée en la gp120 et la gp41 (entrée PDB : 6PWU) [30]. Il faut également noter que la majorité des structures de l'*env* ont été obtenues en présence d'un anticorps bloquant la conformation de l'Env afin de faciliter l'acquisition des données structurales.

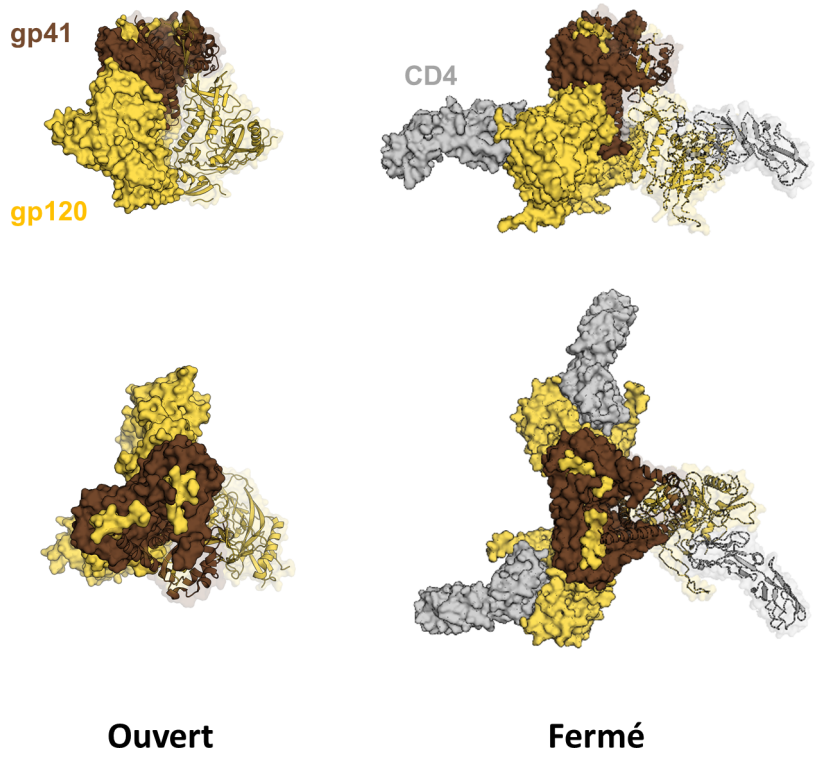

FIGURE 1.7 – Structure 3D du complexe *env*. La forme fermée (entrée PDB : 5CJX, à gauche) et la forme ouverte (entrée PDB : 5VN3, à droite) sont représentées par des surfaces ou des rubans pour un complexe gp120–gp41 ou gp41–gp120–CD4 (gp41 en marron, gp120 en jaune et CD4 en gris) vue de côté (en haut) et du dessus (en bas).

1.2.3 Généralités sur la glycoprotéine virale 120

La structure 3D de la gp120 a également été largement étudiée sous forme monomérique, libre ou liée à CD4 ou à des anticorps. Dans tous ces complexes, la structure du coeur de la gp120 est extrêmement bien conservée. Par exemple, la structure de la gp120 soluble, monomérique et liée au CD4 se superpose parfaitement à celle de la gp120 liée au CD4 dans le trimère Env (entrées PDB respectifs : 5VN3, 2B4C) [22, 24]. L'alignement 3D des coordonnées des carbones en position α (Cα) produit un écart quadratique moyen (RMSD) égal à 0,82 Å. Cette valeur ne prend pas en compte des parties flexibles de la gp120 que sont des boucles variables, car celles-ci ne sont pas décrites dans ces deux structures 3D. La position des V1 et V2 peut en fait être considérée comme une marque de la forme ouverte ou fermée de l'*env*. De même, la V3 adopte des caractéristiques structurales distinctes dans les deux forme de l'*env* : la V3 est repliée si l'*env* est sous forme fermée et elle est exposée si l'*env* est sous forme ouverte (FIGURE 1.8).

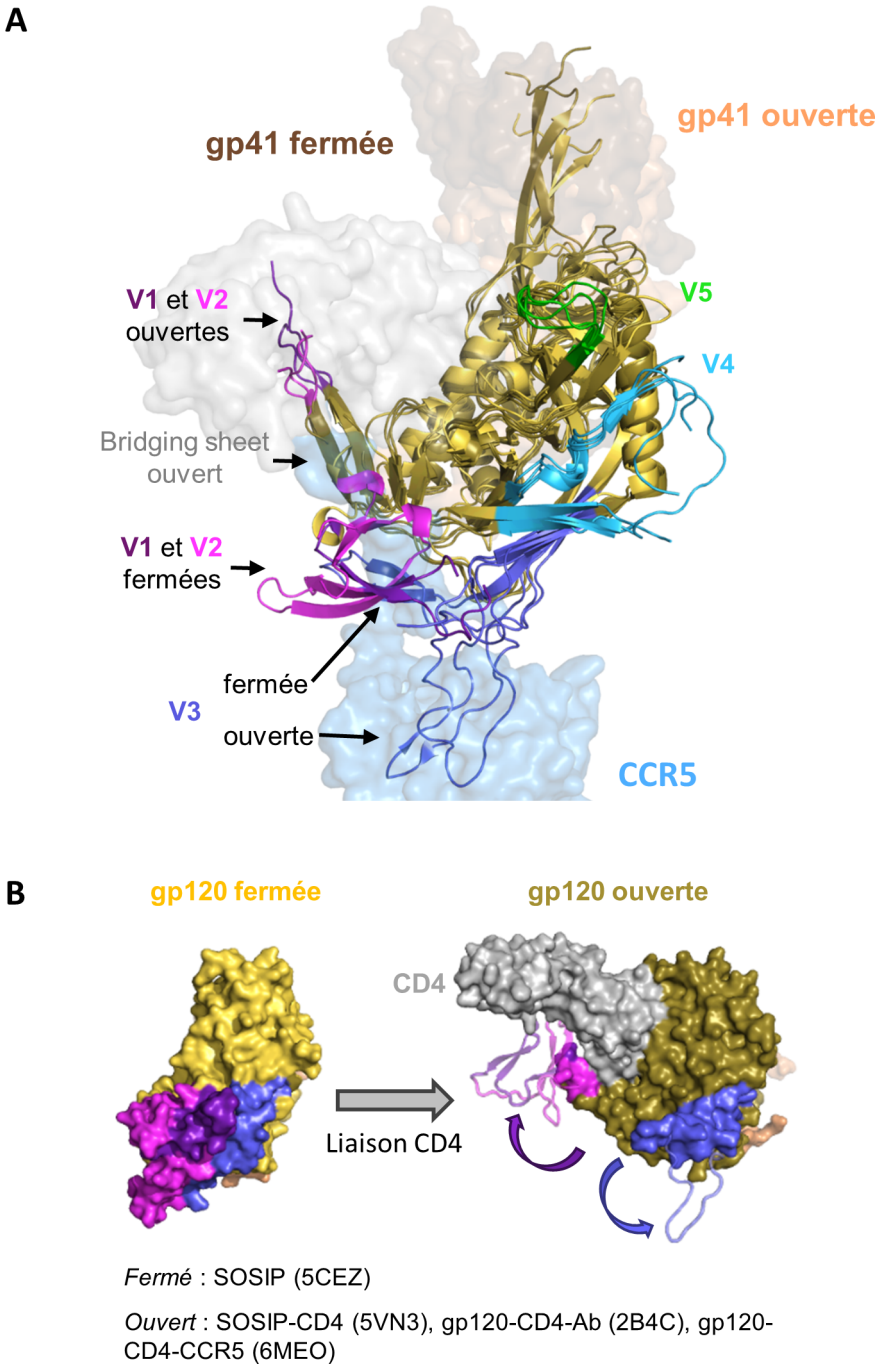

FIGURE 1.8 – Structure 3D de la gp120. (A) Structures superposées de la gp120 sous forme fermée (ruban dorée, entrée PDB : 5CEZ) et sous forme ouverte (ruban ocre, entrées PDB : 2B4C, 5VN3 et 6MEO) en complexe avec la gp41 (forme ouverte : surface transparente colorée en brun clair, entrée PDB : 5CEZ ; forme fermée : surface transparente colorée en brun foncé, entrée PDB : 5VN3) et le CCR5 (surface transparente colorée en bleue, entrée PDB : 6MEO). Les boucles variables 1 à 5 (V1, V2, V3, V4, V5) de la gp120 sont mises en évidence par des couleurs spécifiques : la V1 en violet, la V2 en mauve, la V3 en bleu, la V4 en cyan et la V5 en vert. (B) Représentation de la surface de la gp120 sous forme fermée (à gauche) et ouverte (à droite). Les structures des V1/V2 et de la V3 dans la forme ouverte de la gp120 ont été extraites des entrées 3J70 et 2B4C de la PDB, respectivement.

1.2.4 Corécepteurs de la glycoprotéine virale 120

La gp120 utilise un corécepteur lors de l'étape l'entrée du VIH-1. Si CCR5 est le corécepteur principal, certaines souches virale ciblent le récepteur à C-X-C chimiokine de type 4 (CXCR4). Le corécepteur ciblé définit le tropisme viral. Ainsi les virus sont de tropisme R5 s'ils utilisent CCR5 ou de tropisme X4 s'ils utilisent CXCR4. Il existe des virus capables d'utiliser les deux corécepteurs. Ils sont de tropisme R5/X4.

Au début de l'infection, la population virale est constituée majoritairement de virus à tropisme R5, et elle reste généralement ainsi quand l'infection évolue chez les patients non traités [31]. Par contre, la proportion de virus à tropisme X4 au stade SIDA est d'environ 50 % lorsque le patient a été traité par des antirétroviraux qui ne ciblent pourtant pas l'entrée virale [31].

Le CCR5 et le CXCR4 sont des récepteurs couplés aux protéines G (RCPG) et plus spécifiquement des récepteurs aux chimiokines. Les chimiokines jouent un rôle important dans le système immunitaire permettant de contrôler la migration et le positionnement des cellules exprimant les récepteurs correspondants. Ces corécepteurs sont exprimés par les lymphocytes T CD4+ (TCD4) et aussi par les macrophages en ce qui concerne le CCR5 [32, 33].

La sélection d'un récepteur est conditionnée par la séquence de la V3 de la gp120 [34]. La V3 possède ainsi plus de résidus chargés positivement dans les virus à tropisme X4, convenant d'avantage au site de liaison chargé négativement du CXCR4 [35, 36]. De plus, la cavité transmembranaire du CXCR4 est moins ouverte dans sa partie profonde que celle du CCR5 obligeant la V3 des virus X4 à privilégier des interactions avec des résidus de régions extracellulaires du corécepteur (FIGURE 1.9).

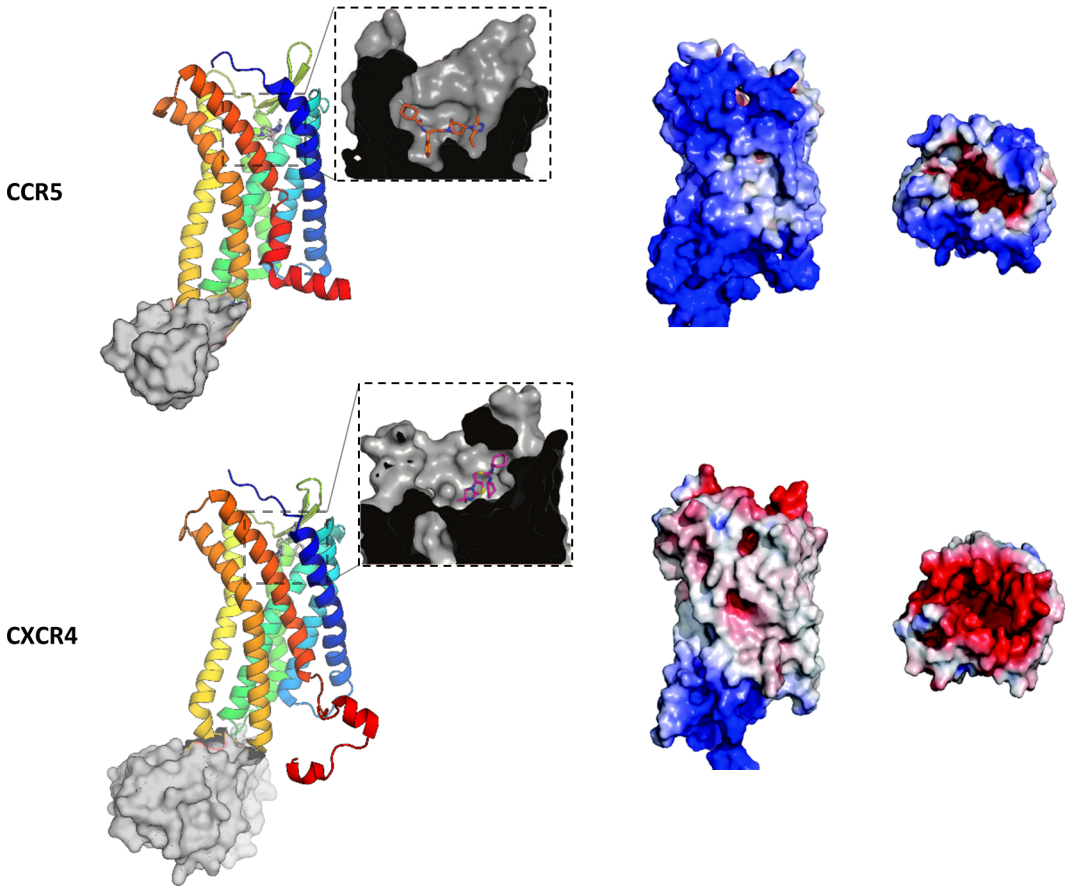

FIGURE 1.9 – Structure 3D des corécepteurs de la gp120. À gauche les 1ère structures résolues par RX du CCR5 [35] (entrée PDB : 4MBS) et du CXCR4 [36] (entrée PDB : 3ODU) représentées par des rubans dont la couleur est fonction du numéro de résidus (bleu 1er résidu et rouge dernier résidu dans la séquence d'acides aminés) avec leur protéine de fusion (surface gris clair). La cavité de chaque protéine est montrée en surface de couleur grise et les ligands en bâtonnets oranges dans CCR5 (code HET : MRV) et magenta dans CXCR4 (code HET : ITD). À droite, les récepteurs sont représentés par leur surface colorée selon leur potentiel électrostatique (en bleu pour les régions chargée positivement, en blanc pour les régions électriquement neutres et en rouge pour les régions chargées négativement). Les vues choisies sont perpendiculaire et parallèle (côté extracellulaire) au plan de la bicouche lipidique [37].

1.3 Le récepteur à C-C chimiokine de type 5

1.3.1 Le CCR5 comme cible thérapeutique

Dès la découverte de leur implication dans l'entrée virale, les corécepteurs ont constitué une cible de choix pour développer des nouvelles thérapies antivirales capables de bloquer la pénétration du virus dans les cellules. La preuve de concept a été apportée par la génétique à la fin des années 90. La présence d'un polymorphisme particulier chez 1 % de la population caucasienne a été associée à une immunité partielle voir totale vis-à-vis du VIH-1. Il s'agit de la délétion $\Delta 32$ qui aboutit à un CCR5 tronqué, qui n'est plus adressé à la surface des cellules [38]. Il n'a fallu qu'une petite dizaine d'années après la validation du CCR5 comme cible thérapeutique pour avoir un médicament antiviral ciblant le CCR5. Son principe actif est le maraviroc (MVC, FIGURE 1.10), développé par le laboratoire Pfizer et approuvé par la FDA en 2007 [39]. L'utilisation du maraviroc en tant qu'antirétroviral reste assez rare. Cette molécule est employée quand les autres thérapies ont échoué et son utilisation peut aboutir à l'émergence rapide de virus à tropisme X4 [40]. Le maraviroc est un agoniste inverse partiel du CCR5 qui se lie à la cavité transmembranaire du récepteur avec une haute affinité, de l'ordre du nanomolaire ($K_d = 0.69 \pm 0{,}26\,\text{nM}$ et $\text{IC}_{50} = 1.05 \pm 0{,}61\,\text{nM}$, déterminée par mesure directe de la radioactivité d'un maraviroc tritié) [41, 42, 35].

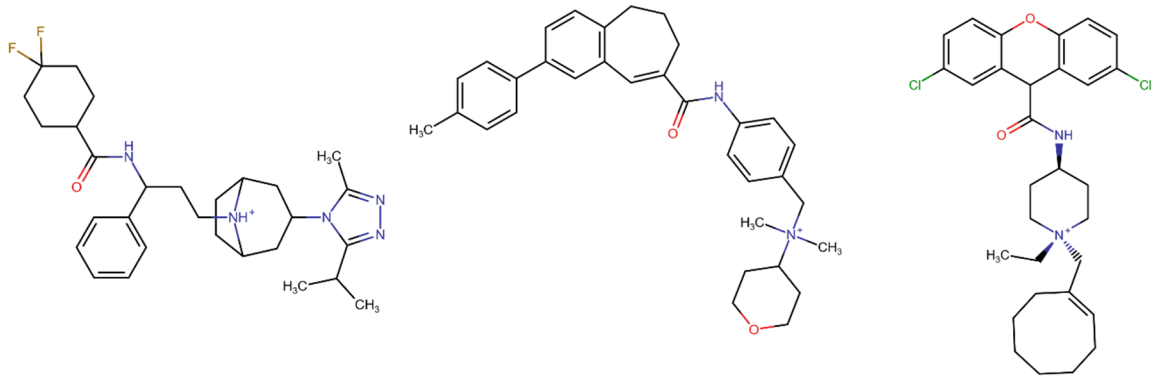

FIGURE 1.10 – Structures chimiques du maraviroc, TAK-779 et UCB35625.

1.3.2 La signalisation par le CCR5 implique différentes populations

Comme la plupart des membres de la famille des RCPG à laquelle il appartient, CCR5 existe à la surface des cellules sous différentes populations qui dépendent du contexte cellulaire.

Plusieurs facteurs sont susceptibles de favoriser une population particulière du CCR5 : les molécules de faible poids moléculaire comme le maraviroc [42, 43], les ligands endogènes comme les chimiokines [44], les anticorps monoclonaux ou encore la composition en lipides de la membrane [45]. Les populations du CCR5 influent sur le couplage à un effecteur, qui est une protéine G ou une β-arrestine, enclenchant la transduction du signal à l'intérieur de la cellule ou l'internalisation du récepteur par endocytose [46]. Du point de vue structural, la capacité du récepteur à lier, dans sa partie intracellulaire, l'effecteur dépend de sa conformation, qui porte en elle des caractéristiques de son état d'activation. Dans la transduction du signal, le CCR5 privilégiera un état d'activation en fonction des propriétés pharmacologiques du ligand lié dans sa cavité transmembranaire. Par exemple le TAK-779 (FIGURE 1.10), qui est un agoniste inverse, bloque le CCR5 dans une forme totalement inactivée, empêchant toute signalisation [42]. Le maraviroc, qui est un agoniste inverse mais partiel, induit une faible signalisation. À l'inverse, la chimiokine (motif C-C) ligand 3 (CCL3) est un agoniste qui active le récepteur [47]. Depuis une vingtaine d'année, des chercheurs ont mis en évidence la capacité de certains ligands à activer une voie de signalisation particulière dans la cellule, en recrutant un effecteur intracellulaire particulier. On parle alors d'agonisme biaisé. Par exemple, le ligand UCB35625 (FIGURE 1.10) est capable de recruter les protéines G mais pas la β-arrestine alors que la chimiokine permet les deux [48]. Dans sa forme libre ou en présence d'un antagoniste, le CCR5 est à un niveau d'activité basal et reste capable d'activer des voix de signalisation mais à un niveau inférieur qu'avec un agoniste. De plus, la mutation de résidus dans le CCR5 peut privilégier la forme active (T82P [49] et G286F [50]) ou inactive (P84A [51] et R126N [52]) du récepteur.

1.3.3 La gp120, un ligand qui questionne les structures expérimentales

La gp120 est décrite comme un antagoniste [42], elle ne bloque pas la signalisation du niveau basal du CCR5 comme le fait maraviroc. Dans les deux cas, la liaison de la chimiokine CCL3 au CCR5 est compromise, mais selon deux mécanismes différents. La gp120 et la CCL3 sont en compétition l'une avec l'autre pour la liaison au CCR5, alors que le TAK-779 et le maraviroc inhibent la liaison de CCL3 de manière allostérique. De manière similaire l'inhibition par maraviroc de la liaison de la gp120 au CCR5 est allostérique d'après les données pharmacologiques. Ces résultats semblent en contradiction avec les structure expérimentales décrivant les complexes entre le CCR5 avec ses ligands et qui montrent que les sites de liaison de la chimiokine (C-C motif) ligand 5 modifiée ([5P7]CCL5), du maraviroc et de la gp120 se recouvrent partiellement (FIGURE 1.11) [53]. La chimiokine [5P7]CCL5 est une chimiokine artificielle, conçue à partir de la chimiokine CCL5 et dont les résidus de la partie N-terminale ont été modifiés afin d'augmenter l'affinité de la chimiokine pour le récepteur. Cette modification rend la [5P7]CCL5 compétitive de la gp120 et antagoniste du CCR5 [54].

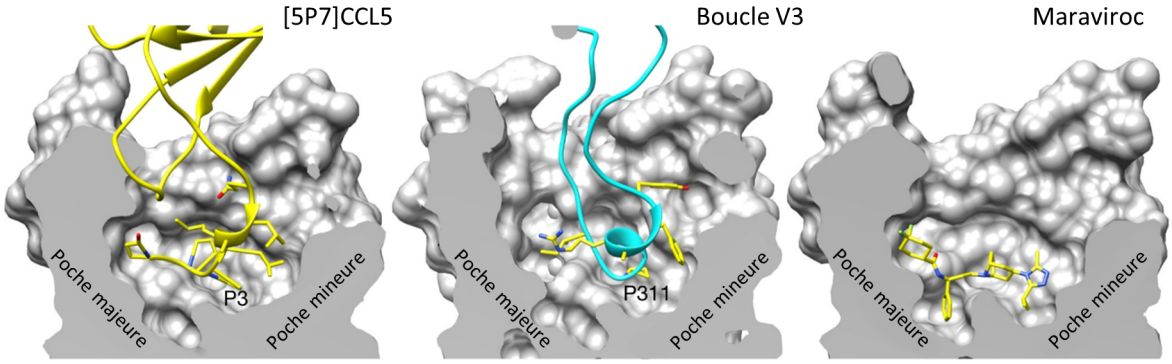

FIGURE 1.11 – Cavité du CCR5 avec trois ligands différents. Image issue de la référence [53]. Les RMSD sur les positions des Cα des hélices transmembranaires sont de 0,81 Å (CCR5–CCL3 vs CCR5–gp120), 0,89 Å (CCR5–maraviroc vs CCR5–[5P7]CCL5) et 1,16 Å (CCR5–maraviroc vs CCR5–gp120).

Dans les complexes entre le CCR5 et le maraviroc, la [5P7]CCL5 et la gp120, la chaîne principale de la partie transmembranaire, qui accueille le ligand, dévie peu d'une structure à l'autre (FIGURE 1.11), ce qui suggère que les trois ligands ciblent des conformations du récepteur proches, que l'on pourrait assimiler à une seule et même population. Cependant la nature allostérique de l'inhibition de la liaison de la chimiokine CCL3 invalide cette hypothèse. Il est

donc raisonnable de concevoir que les populations de CCR5 diffèrent par des changements de faible amplitude au niveau de la partie transmembranaire.

Le CCR5 est capable de former des homodimères, qui contribuent à la diversité des populations prises par le récepteur à la surface des cellules. L'organisation des deux protomères au sein du dimère est versatile, avec au moins trois interfaces différentes [55]. Les formes monomérique et dimérique du récepteur coexistent, et il s'avère que le maraviroc augmente la proportion de dimères tout en favorisant une interface particulière [55]. Dans la structure cristallographique du complexe CCR5–maraviroc, le récepteur est modifié par l'ajout de la rubredoxine qui compromet la formation de dimère [56]. On peut alors supposer que la conformation du protomère qui lie le maraviroc dans le dimère de CCR5 est différente de celle qui est reconnue par la gp120. Ainsi, les structures cristallographiques du CCR5 lié au maraviroc et à la [5P7]CCL5 ne donnent qu'un aperçu des conformations adoptées par CCR5 et que ce dernier peut être biaisé par les conditions expérimentales (mutations, protéine de fusion). La structure du CCR5 en complexe avec la gp120, obtenue par cryo-microscopie électronique (cryo-EM), décrit des protéines natives, mais son interprétation doit aussi considérer des limites expérimentales, en particulier au niveau de la résolution. Ainsi, la carte de densité électronique observée pour le CCR5 a une résolution supérieure à 4 Å au niveau des hélices, voire autour de 5 Å dans leur moitié inférieure [53]. On suppose alors qu'il existe un ensemble disjoint de populations de CCR5 dont ne rendent pas compte les structures expérimentales.

1.3.4 Étude des conformations du CCR5 par dynamique moléculaire

Depuis une dizaine d'années et surtout depuis la publication de la structure cristallographique du récepteur en 2013, des simulations par dynamique moléculaire de CCR5 dans différentes configurations ont été produites. Une des premières et des plus notables concerne la simulation par dynamique moléculaire dite classique (sans facteur d'accélération et non biaisée) de tous les atomes du CCR5 natif (du résidu Pro29 au Gln313) inséré dans une bicouche de 1,2-dipalmitoylphosphatidylcholine, sans ligand d'une part, et avec le maraviroc dans la cavité transmembranaire d'autre part [56]. Cette simulation d'une durée de l'ordre de la microseconde

montre que la rubredoxine contraint le cinquième domaine transmembranaire (TM5). En absence de protéine de fusion, les coordonnées atomiques du TM5 dévient largement de celles dans la structure cristallographique lors des simulations. De plus, la simulation entraîne un léger glissement dans la cavité du maraviroc, qui maintient néanmoins son ancrage par le pont ionique avec le Glu283, résidu indispensable pour la stabilisation du complexe. Une autre simulation par dynamique moléculaire accélérée du complexe CCR5–maraviroc montre que la dissociation du maraviroc s'effectue via un chemin empruntant deux sous-sites du CCR5, chargés négativement [57]. Avec la structure du CCR5 comme base de départ, des chercheurs ont proposé des simulations avec d'autres ligands, en particulier la gp120. La première simulation, d'une durée de $19 \times 20\,\text{ns}$, a modélisé la liaison du CCR5 avec la V3 de la gp120 issue d'un virus à tropisme R5/X4 [58]. L'étude identifie les interactions clés de la pointe de la V3 et suggère une inhibition du maraviroc par compétition directe.

En 2018, ZHANG et ses collaborateurs ont tenté d'évaluer les caractéristiques d'activation du récepteur en simulant le CCR5 natif, le CCR5 portant la mutation G286F et étant constitutivement actif, et le CCR5 portant la mutation R126N et étant constitutivement inactif [59]. Ils relèvent que la communication entre l'hélice transmembranaire 6 (TM6) et l'hélice transmembranaire 7 (TM7) est renforcée par la mutation R126N tandis que la mutation G286F induit un comportement particulier dans la région NPxxY. L'approche computationelle a également permis de mettre en avant des résidus pouvant être impliqués dans l'homodimérisation du CCR5 [60, 55]. Ces études ont montré la stabilité de plusieurs interfaces de dimérisation, impliquant l'hélice transmembranaire 4 (TM4), les TM5 et/ou TM6. Dans le dimère dont l'interface implique les hélices transmembranaires 1 (TM1), 2 (TM2), 3 (TM3) et 4, les deux protomères présentent une liaison asymétrique au maraviroc.

Dans notre laboratoire, des simulations du CCR5 dans différentes configurations à savoir natif sans ligand, constitutivement actif sans ligand, constitutivement inactif sans ligand, natif lié à la CCL3 et natif lié au maraviroc ont été produites, afin de constituer un inventaire de conformations du CCR5 dans des états fonctionnels différents.

1.3.5 La gp120 exploite la variabilité conformationnelle du CCR5

L'hétérogénéité structurale du CCR5 a aussi été liée à la diversité des gp120 observée chez les individus infectés [61, 62, 63, 64, 65, 66, 67]. Bernard LAGANE et son équipe ont étudié les propriétés de liaison au CCR5 de gp120 de souches variées, adaptées en laboratoire et souches primaires extraites de phases asymptomatiques ou SIDA. Ils ont ainsi démontré que les gp120 différentes utilisent des formes différentes du corécepteur [68]. L'effet de la mutation CCR5-L196K, affectant la capacité du CCR5 à dimériser [55], suggère par ailleurs que les gp120 peuvent former avec les dimères de CCR5 des complexes de stœchiométries différentes. Deux variantes de la gp120 aux comportements extrêmes ont été plus particulièrement caractérisées du point de vue phénotypique par Céline Galès et son équipe. Ces deux gp120, issues de virus isolés des cellules mononuclées du sang périphérique (PBMC) d'un individu de la cohorte d'Amsterdam, sont désignées par #25 et #34. Elles ont des tropismes cellulaires différents (TCD4 pour #25, TCD4 et macrophage pour #34) et une préférence marquée pour des stœchiométries différentes dans les complexes avec les dimères de CCR5 (1:2 pour #25 et 2:2 pour #34). Les données préliminaires montrent que #25 et #34 se comportent comme des agonistes biaisés d'un point de vue pharmacologique, avec un profil de signalisation distinct de la chimiokine (C-C motif) ligand 4.

1.4 Objectifs de la thèse

Mon projet de thèse s'inscrit dans l'étude de la sélectivité fonctionnelle des variantes de la gp120 des virus du groupe M de VIH-1 afin d'établir quelles en sont les bases moléculaires et de comprendre leur incidence sur les propriétés phénotypiques des virus et le développement de l'infection. Ce projet, financé par France Recherche Nord & Sud Sida-hiv Hépatites (ANRS) associe notre laboratoire à celui de Bernard LAGANE (CPTP, Toulouse) et de Céline GALÈS (I2MC, Toulouse). Ma contribution porte sur la caractérisation, à l'échelle atomique, des conformations du CCR5 qui sont reconnues par quatre variantes de séquences de gp120, pour répondre aux questions suivantes :

- Quelle est la diversité des conformations du CCR5 reconnues par les gp120 ?
- Est-ce que le mode de reconnaissance du CCR5 par les gp120 varie ?
- Est-il possible d'identifier, dans la séquence de gp120, les acides aminés déterminants pour la sélection des conformations du CCR5 ?
- Est-ce que les conformations du CCR5 lié aux gp120 possèdent des caractéristiques de l'activation d'un RCPG ?

Les trois premières questions sont abordées dans la première partie de ma thèse. Ainsi, le chapitre 2 décrit la simulation par dynamique moléculaire du complexe entre la gp120, le CCR5 et le CD4. La comparaison des conformations obtenues pour quatre variantes de la gp120 montre que les différentes gp120 ciblent bien des conformations distinctes du CCR5. Elle révèle également que la gp120 ne s'accommode pas d'une conformation rigide du CCR5 mais qu'elle s'adapte à un continuum de conformations similaires. Enfin nous observons que le mode d'interaction entre le CCR5 et la gp120 est caractéristique de la souche virale. En particulier, des interactions discriminantes impliquent la V3 de la gp120 et la cavité transmembranaire du CCR5.

La dernière question sera discutée dans une deuxième partie qui proposera pour cela une nouvelle méthode de représentation des états d'activation des RCPG. Ainsi, le chapitre 3 combine les aspects méthodologiques, avec le développement et la validation de cette méthode nommée ATOLL, et l'interprétation des résultats de son application au CCR5. Nos données sug-

gèrent que les différentes gp120 sont associées à des états différents du CCR5, qui sont distincts de ceux observés pour le récepteur lié au maraviroc ou à la CCL3.

L'objectif ultime du projet entrepris par les trois laboratoires partenaires est la conception rationnelle de ligands biaisés du CCR5, c'est-à-dire de molécules capables de cibler une sous-population donnée du récepteur, pour produire un effet fonctionnel particulier. Pour atteindre cet objectif, l'approche proposée au laboratoire est le criblage virtuel de chimiothèques commerciales de molécules de type candidats médicament ("druglike") par docking moléculaire à haut débit. Le docking est une méthode de choix pour prédire la pose d'un ligand dans son site de liaison de la protéine ciblée, néanmoins, les fonctions de score associées échouent souvent à identifier la bonne pose parmi celles possibles, et aussi à trier différents ligands par affinité pour la protéine cible. De nombreux travaux, dont ceux menés au laboratoire depuis une quinzaine d'années, ont montré que la sélection des poses par comparaison des modes d'interaction prédits aux modes d'interaction de structures de référence pouvait pallier aux défaut des fonctions de score de docking [69]. Ces approches encodent les modes de liaison à l'aide de modèles simples et les comparent deux par deux. Dans le cas de la conception de ligands biaisés du CCR5, nous serons confrontés à de très nombreuses comparaisons, puisque les structures de référence se comptent par milliers car issues des longues trajectoires de dynamique, qui plus est avec des modes de liaison proches, donc avec une incertitude sur la manière de traiter les scores. Des développements méthodologiques ont donc été nécessaires pour l'adaptation de cette méthode appelée LID détaillés dans le chapitre 4.

1.5 Références

[1] Richard D. Moore, Julia Hidalgo, Barbara W. Sugland, and Richard E. Chaisson. Zidovudine and the Natural History of the Acquired Immunodeficiency Syndrome. *New England Journal of Medicine*, 324(20) :1412–1416, Mai 1991.

[2] Kholoud Porter, Anne M. Johnson, Andrew N. Phillips, and Janet H. Darbyshire. The practical significance of potential biases in estimates of the AIDS incubation period distribution _in the UK Register of HIV Seroconverters. *AIDS*, 13(14) :1943–1951, Octobre 1999.

[3] Margaret A. Fischl, Douglas D. Richman, Michael H. Grieco, Michael S. Gottlieb, Paul A. Volberding, Oscar L. Laskin, John M. Leedom, Jerome E. Groopman, Donna Mildvan, Robert T. Schooley, George G. Jackson, David T. Durack, and Dannie King. The Efficacy of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex. *New England Journal of Medicine*, 317(4) :185–191, Juillet 1987.

[4] Neil M.H. Graham, Scott L. Zeger, Lawrence P. Park, Sten H. Vermund, Roger Detels, Charles R. Rinaldo, and John P. Phair. The Effects on Survival of Early Treatment of Human Immunodeficiency Virus Infection. *New England Journal of Medicine*, 326(16) : 1037–1042, Avril 1992.

[5] Adam Trickey, Margaret T May, Jorg-Janne Vehreschild, Niels Obel, M John Gill, Heidi M Crane, Christoph Boesecke, Sophie Patterson, Sophie Grabar, Charles Cazanave, Matthias Cavassini, Leah Shepherd, Antonella d'Arminio Monforte, Ard van Sighem, Mike Saag, Fiona Lampe, Vicky Hernando, Marta Montero, Robert Zangerle, Amy C Justice, Timothy Sterling, Suzanne M Ingle, and Jonathan A C Sterne. Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013 : a collaborative analysis of cohort studies. *The Lancet HIV*, 4(8) :e349–e356, Août 2017.

[6] Deborah Donnell, Jared M Baeten, James Kiarie, Katherine K Thomas, Wendy Stevens, Craig R Cohen, James McIntyre, Jairam R Lingappa, and Connie Celum. Heterosexual HIV-1 transmission after initiation of antiretroviral therapy : a prospective cohort analysis. *The Lancet*, 375(9731) :2092–2098, Juin 2010.

[7] Claire L. Townsend, Laura Byrne, Mario Cortina-Borja, Claire Thorne, Annemiek de Ruiter, Hermione Lyall, Graham P. Taylor, Catherine S. Peckham, and Pat A. Tookey. Earlier initiation of ART and further decline in mother-to-child HIV transmission rates, 2000–2011. *AIDS*, 28(7) :1049–1057, Avril 2014.

[8] Gero Hütter, Daniel Nowak, Maximilian Mossner, Susanne Ganepola, Arne Müßig, Kristina Allers, Thomas Schneider, Jörg Hofmann, Claudia Kücherer, Olga Blau, Igor W. Blau, Wolf K. Hofmann, and Eckhard Thiel. Long-Term Control of HIV by CCR5 Delta32/Delta32 Stem-Cell Transplantation. *New England Journal of Medicine*, 360(7) : 692–698, Février 2009.

[9] Ravindra K. Gupta, Sultan Abdul-Jawad, Laura E. McCoy, Hoi Ping Mok, Dimitra Peppa, Maria Salgado, Javier Martinez-Picado, Monique Nijhuis, Annemarie M. J. Wensing, Helen Lee, Paul Grant, Eleni Nastouli, Jonathan Lambert, Matthew Pace, Fanny Salasc, Christopher Monit, Andrew J. Innes, Luke Muir, Laura Waters, John Frater, Andrew M. L. Lever, Simon G. Edwards, Ian H. Gabriel, and Eduardo Olavarria. HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell transplantation. *Nature*, 568(7751) :244–248, Avril 2019.

[10] José M. Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely High Mutation Rate of HIV-1 In Vivo. *PLOS Biology*, 13(9) :e1002251, Septembre 2015.

[11] Stuart Z. Shapiro. Lessons for general vaccinology research from attempts to develop an HIV vaccine. *Vaccine*, 37(26) :3400–3408, Juin 2019.

[12] Dennis R. Burton. Advancing an HIV vaccine ; advancing vaccinology. *Nature Reviews Immunology*, 19(2) :77–78, Février 2019.

[13] Wanwisa Promsote, Megan E. DeMouth, Cassandra G. Almasri, and Amarendra Pegu. Anti-HIV-1 Antibodies : An Update. *BioDrugs*, 34(2) :121–132, Avril 2020.

[14] Paul M. Sharp and Beatrice H. Hahn. Origins of HIV and the AIDS Pandemic. *Cold Spring Harbor Perspectives in Medicine*, 1(1) :a006841, Janvier 2011.

[15] Omobolaji T. Campbell-Yesufu and Rajesh T. Gandhi. Update on Human Immunodeficiency Virus (HIV)-2 Infection. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, 52(6) :780–787, Mars 2011.

[16] D. S. Burke. Recombination in HIV : an important viral evolutionary strategy. *Emerging Infectious Diseases*, 3(3) :253–259, 1997.

[17] Denis M Tebit and Eric J Arts. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *The Lancet Infectious Diseases*, 11(1) : 45–56, Janvier 2011.

[18] André F. Santos and Marcelo A. Soares. HIV Genetic Diversity and Drug Resistance. *Viruses*, 2(2) :503–531, Février 2010.

[19] Françoise Barré-Sinoussi, Anna Laura Ross, and Jean-François Delfraissy. Past, present and future : 30 years of HIV research. *Nature Reviews Microbiology*, 11(12) :877–883, Décembre 2013.

[20] Stephen C. Harrison. Viral membrane fusion. *Virology*, 479-480 :498–507, Mai 2015.

[21] Marie Pancera, Tongqing Zhou, Aliaksandr Druz, Ivelin S. Georgiev, Cinque Soto, Jason Gorman, Jinghe Huang, Priyamvada Acharya, Gwo-Yu Chuang, Gilad Ofek, Guillaume B. E. Stewart-Jones, Jonathan Stuckey, Robert T. Bailer, M. Gordon Joyce, Mark K. Louder, Nancy Tumba, Yongping Yang, Baoshan Zhang, Myron S. Cohen, Barton F. Haynes, John R. Mascola, Lynn Morris, James B. Munro, Scott C. Blanchard, Walther Mothes, Mark Connors, and Peter D. Kwong. Structure and immune recognition of trimeric prefusion HIV-1 Env. *Nature*, 514(7523) :455–461, Octobre 2014.

[22] Peter D. Kwong, Richard Wyatt, James Robinson, Raymond W. Sweet, Joseph Sodroski, and Wayne A. Hendrickson. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393(6686) :648–659, Juin 1998.

[23] Haoqing Wang, Alexander A. Cohen, Rachel P. Galimidi, Harry B. Gristick, Grant J. Jensen, and Pamela J. Bjorkman. Cryo-EM structure of a CD4-bound open HIV-1 envelope

trimer reveals structural rearrangements of the gp120 V1V2 loop. *Proceedings of the National Academy of Sciences*, 113(46) :E7151–E7158, Novembre 2016.

[24] Gabriel Ozorowski, Jesper Pallesen, Natalia de Val, Dmitry Lyumkis, Christopher A. Cottrell, Jonathan L. Torres, Jeffrey Copps, Robyn L. Stanfield, Albert Cupo, Pavel Pugach, John P. Moore, Ian A. Wilson, and Andrew B. Ward. Open and closed structures reveal allostery and pliability in the HIV-1 envelope spike. *Nature*, 547(7663) :360–363, Juillet 2017.

[25] Lijun Wu, Norma P. Gerard, Richard Wyatt, Hyeryun Choe, Cristina Parolin, Nancy Ruffing, Alessândra Borsetti, Angelo A. Cardoso, Elizabeth Desjardin, Walter Newman, Craig Gerard, and Joseph Sodroski. CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. *Nature*, 384(6605) :179–183, Novembre 1996.

[26] Carlo D. Rizzuto, Richard Wyatt, Nivia Hernández-Ramos, Ying Sun, Peter D. Kwong, Wayne A. Hendrickson, and Joseph Sodroski. A Conserved HIV gp120 Glycoprotein Structure Involved in Chemokine Receptor Binding. *Science*, 280(5371) :1949–1953, Juin 1998.

[27] Jun Liu, Alberto Bartesaghi, Mario J. Borgnia, Guillermo Sapiro, and Sriram Subramaniam. Molecular architecture of native HIV-1 gp120 trimers. *Nature*, 455(7209) :109–113, Septembre 2008.

[28] Andrew B. Ward and Ian A. Wilson. The HIV-1 envelope glycoprotein structure : nailing down a moving target. *Immunological Reviews*, 275(1) :21–32, 2017.

[29] Zhi Yang, Haoqing Wang, Albert Z. Liu, Harry B. Gristick, and Pamela J. Bjorkman. Asymmetric opening of HIV-1 Env bound to CD4 and a coreceptor-mimicking antibody. *Nature Structural & Molecular Biology*, 26(12) :1167–1175, Décembre 2019.

[30] Junhua Pan, Hanqin Peng, Bing Chen, and Stephen C. Harrison. Cryo-EM Structure of Full-length HIV-1 Env Bound With the Fab of Antibody PG16. *Journal of Molecular Biology*, 432(4) :1158–1168, Février 2020.

[31] Peter W. Hunt, P. Richard Harrigan, Wei Huang, Michael Bates, David W. Williamson, Joseph M. McCune, Richard W. Price, Serena S. Spudich, Harry Lampiris, Rebecca Hoh, Teri Leigler, Jeffrey N. Martin, and Steven G. Deeks. Prevalence of CXCR4 Tropism among Antiretroviral-Treated HIV-1–Infected Patients with Detectable Viremia. *The Journal of Infectious Diseases*, 194(7) :926–930, Octobre 2006.

[32] Conrad C. Bleul, Lijun Wu, James A. Hoxie, Timothy A. Springer, and Charles R. Mackay. The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. *Proceedings of the National Academy of Sciences*, 94(5) :1925–1930, Mars 1997.

[33] Cédric Blanpain, Frédérick Libert, Gilbert Vassart, and Marc Parmentier. CCR5 and HIV Infection. *Receptors and Channels*, 8(1) :19–31, Janvier 2002.

[34] Stéphanie Raymond, Pierre Delobel, Maud Mavigner, Michelle Cazabat, Corinne Souyris, Karine Sandres-Sauné, Lise Cuzin, Bruno Marchou, Patrice Massip, and Jacques Izopet. Correlation between genotypic predictions based on V3 sequences and phenotypic determination of HIV-1 tropism. *AIDS*, 22(14) :F11, Septembre 2008.

[35] Qiuxiang Tan, Ya Zhu, Jian Li, Zhuxi Chen, Gye Won Han, Irina Kufareva, Tingting Li, Limin Ma, Gustavo Fenalti, Jing Li, Wenru Zhang, Xin Xie, Huaiyu Yang, Hualiang Jiang, Vadim Cherezov, Hong Liu, Raymond C. Stevens, Qiang Zhao, and Beili Wu. Structure of the CCR5 Chemokine Receptor–HIV Entry Inhibitor Maraviroc Complex. *Science*, 341 (6152) :1387–1390, Septembre 2013.

[36] Beili Wu, Ellen Y. T. Chien, Clifford D. Mol, Gustavo Fenalti, Wei Liu, Vsevolod Katritch, Ruben Abagyan, Alexei Brooun, Peter Wells, F. Christopher Bi, Damon J. Hamel, Peter Kuhn, Tracy M. Handel, Vadim Cherezov, and Raymond C. Stevens. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science*, 330(6007) :1066–1071, Novembre 2010.

[37] Olga V. Kalinina, Nico Pfeifer, and Thomas Lengauer. Modelling binding between CCR5 and CXCR4 receptors and their ligands suggests the surface electrostatic potential of the

co-receptor to be a key player in the HIV-1 tropism. *Retrovirology*, 10(1) :130, Novembre 2013.

[38] Michel Samson, Frédérick Libert, Benjamin J. Doranz, Joseph Rucker, Corinne Liesnard, Claire-Michèle Farber, Sentob Saragosti, Claudine Lapouméroulie, Jacqueline Cognaux, Christine Forceille, Gaetan Muyldermans, Chris Verhofstede, Guy Burtonboy, Michel Georges, Tsuneo Imai, Shalini Rana, Yanji Yi, Robert J. Smyth, Ronald G. Collman, Robert W. Doms, Gilbert Vassart, and Marc Parmentier. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature*, 382(6593) :722–725, Août 1996.

[39] Gerd Fätkenheuer, Anton L. Pozniak, Margaret A. Johnson, Andreas Plettenberg, Schlomo Staszewski, Andy I. M. Hoepelman, Michael S. Saag, Frank D. Goebel, Jürgen K. Rockstroh, Bruce J. Dezube, Tim M. Jenkins, Christine Medhurst, John F. Sullivan, Caroline Ridgway, Samantha Abel, Ian T. James, Mike Youle, and Elna van der Ryst. Efficacy of short-term monotherapy with maraviroc, a new CCR5 antagonist, in patients infected with HIV-1. *Nature Medicine*, 11(11) :1170–1172, Novembre 2005.

[40] Shawna M Woollard and Georgette D Kanmogne. Maraviroc : a review of its use in HIV infection and beyond. *Drug Design, Development and Therapy*, 9 :5447–5468, Octobre 2015.

[41] Javier Garcia-Perez, Patricia Rueda, Jose Alcami, Didier Rognan, Fernando Arenzana-Seisdedos, Bernard Lagane, and Esther Kellenberger. Allosteric Model of Maraviroc Binding to CC Chemokine Receptor 5 (CCR5). *Journal of Biological Chemistry*, 286(38) : 33409–33421, Septembre 2011.

[42] Javier Garcia-Perez, Patricia Rueda, Isabelle Staropoli, Esther Kellenberger, Jose Alcami, Fernando Arenzana-Seisdedos, and Bernard Lagane. New Insights into the Mechanisms whereby Low Molecular Weight CCR5 Ligands Inhibit HIV-1 Infection. *Journal of Biological Chemistry*, 286(7) :4978–4990, Février 2011.

[43] Sudarshan Rajagopal, Daniel L. Bassoni, James J. Campbell, Norma P. Gerard, Craig Gerard, and Tom S. Wehrman. Biased Agonism as a Mechanism for Differential Signaling

by Chemokine Receptors. *Journal of Biological Chemistry*, 288(49) :35039–35048, Juin 2013.

[44] Laura Tarancón Díez, Claudia Bönsch, Sebastian Malkusch, Zinnia Truan, Mihaela Munteanu, Mike Heilemann, Oliver Hartley, Ulrike Endesfelder, and Alexandre Fürstenberg. Coordinate-based co-localization-mediated analysis of arrestin clustering upon stimulation of the C–C chemokine receptor 5 with RANTES/CCL5 analogues. *Histochemistry and Cell Biology*, 142(1) :69–77, Juillet 2014.

[45] Dzung H. Nguyen and Dennis Taub. Cholesterol is essential for macrophage inflammatory protein 1β binding and conformational integrity of CC chemokine receptor 5. *Blood*, 99 (12) :4298–4306, Juin 2002.

[46] Martin Oppermann. Chemokine receptor CCR5 : insights into structure, function, and regulation. *Cellular Signalling*, 16(11) :1201–1210, Novembre 2004.

[47] Jenny Corbisier, Céline Galès, Alexandre Huszagh, Marc Parmentier, and Jean-Yves Springael. Biased Signaling at Chemokine Receptors. *Journal of Biological Chemistry*, 290 (15) :9542–9554, Octobre 2015.

[48] Jenny Corbisier, Alexandre Huszagh, Céline Galés, Marc Parmentier, and Jean-Yves Springael. Partial Agonist and Biased Signaling Properties of the Synthetic Enantiomers J113863/UCB35625 at Chemokine Receptors CCR2 and CCR5. *Journal of Biological Chemistry*, 292(2) :575–584, Janvier 2017.

[49] Diana Alvarez Arias, Jean-Marc Navenot, Wen-bo Zhang, James Broach, and Stephen C. Peiper. Constitutive Activation of CCR5 and CCR2 Induced by Conformational Changes in the Conserved TXP Motif in Transmembrane Helix 2. *Journal of Biological Chemistry*, 278(38) :36513–36521, Septembre 2003.

[50] Anne Steen, Stefanie Thiele, Dong Guo, Lærke S. Hansen, Thomas M. Frimurer, and Mette M. Rosenkilde. Biased and Constitutive Signaling in the CC-chemokine Receptor CCR5 by Manipulating the Interface between Transmembrane Helices 6 and 7. *Journal of Biological Chemistry*, 288(18) :12511–12521, Mars 2013.

[51] Cédric Govaerts, Cédric Blanpain, Xavier Deupi, Sébastien Ballet, Juan A. Ballesteros, Shoshana J. Wodak, Gilbert Vassart, Leonardo Pardo, and Marc Parmentier. The TXP Motif in the Second Transmembrane Helix of CCR5 A STRUCTURAL DETERMINANT OF CHEMOKINE-INDUCED ACTIVATION. *Journal of Biological Chemistry*, 276(16) : 13217–13225, Avril 2001.

[52] Bernard Lagane, Sébastien Ballet, Thierry Planchenault, Karl Balabanian, Emmanuel Le Poul, Cédric Blanpain, Yann Percherancier, Isabelle Staropoli, Gilbert Vassart, Martin Oppermann, Marc Parmentier, and Françoise Bachelerie. Mutation of the DRY Motif Reveals Different Structural Requirements for the CC Chemokine Receptor 5-Mediated Signaling and Receptor Endocytosis. *Molecular Pharmacology*, 67(6) :1966–1976, Juin 2005.

[53] Md Munan Shaik, Hanqin Peng, Jianming Lu, Sophia Rits-Volloch, Chen Xu, Maofu Liao, and Bing Chen. Structural basis of coreceptor recognition by HIV-1 envelope spike. *Nature*, 565(7739) :318–323, Janvier 2019.

[54] Yi Zheng, Gye Won Han, Ruben Abagyan, Beili Wu, Raymond C. Stevens, Vadim Cherezov, Irina Kufareva, and Tracy M. Handel. Structure of CC Chemokine Receptor 5 with a Potent Chemokine Antagonist Reveals Mechanisms of Chemokine Recognition and Molecular Mimicry by HIV. *Immunity*, 46(6) :1005–1017.e5, Juin 2017.

[55] Jun Jin, Fanny Momboisse, Gaelle Boncompain, Florian Koensgen, Zhicheng Zhou, Nelia Cordeiro, Fernando Arenzana-Seisdedos, Franck Perez, Bernard Lagane, Esther Kellenberger, and Anne Brelot. CCR5 adopts three homodimeric conformations that control cell surface delivery. *Science Signaling*, 11(529), Mai 2018.

[56] Ramin Ekhteiari Salmas, Mine Yurtsever, and Serdar Durdagi. Investigation of Inhibition Mechanism of Chemokine Receptor CCR5 by Micro-second Molecular Dynamics Simulations. *Scientific Reports*, 5(1) :13180, Août 2015.

[57] Qifeng Bai, Yang Zhang, Xiaomeng Li, Wenbo Chen, Huanxiang Liu, and Xiaojun Yao. Computational study on the interaction between CCR5 and HIV-1 entry inhibitor maraviroc : insight from accelerated molecular dynamics simulation and free energy calculation. *Physical Chemistry Chemical Physics*, 16(44) :24332–24338, Octobre 2014.

[58] Phanourios Tamamis and Christodoulos A. Floudas. Molecular Recognition of CCR5 by an HIV-1 gp120 V3 Loop. *PLOS ONE*, 9(4) :e95767, Avril 2014.

[59] Fuhui Zhang, Yuan Yuan, Haiyan Li, Liting Shen, Yanzhi Guo, Zhining Wen, and Xuemei Pu. Using accelerated molecular dynamics simulation to shed light on the mechanism of activation/deactivation upon mutations for CCR5. *RSC Advances*, 8(66) :37855–37865, Novembre 2018.

[60] Fuhui Zhang, Yuan Yuan, Minghui Xiang, Yanzhi Guo, Menglong Li, Yijing Liu, and Xuemei Pu. Molecular Mechanism Regarding Allosteric Modulation of Ligand Binding and the Impact of Mutations on Dimerization for CCR5 Homodimer. *Journal of Chemical Information and Modeling*, 59(5) :1965–1976, Mai 2019.

[61] Rebecca M. Lynch, Tongye Shen, S. Gnanakaran, and Cynthia A. Derdeyn. Appreciating HIV Type 1 Diversity : Subtype Differences in Env. *AIDS Research and Human Retroviruses*, 25(3) :237–248, Mars 2009.

[62] Joris Hemelaar. Implications of HIV diversity for the HIV-1 pandemic. *Journal of Infection*, 66(5) :391–400, Mai 2013.

[63] Leonardo Augusto Luvison Araújo and Sabrina E. M. Almeida. HIV-1 Diversity in the Envelope Glycoproteins : Implications for Viral Entry Inhibition. *Viruses*, 5(2) :595–604, Février 2013.

[64] James M. Fox, Richard Kasprowicz, Oliver Hartley, and Nathalie Signoret. CCR5 susceptibility to ligand-mediated down-modulation differs between human T lymphocytes and myeloid cells. *Journal of Leukocyte Biology*, 98(1) :59–71, 2015.

[65] Philippe Colin, Yann Bénureau, Isabelle Staropoli, Yongjin Wang, Nuria Gonzalez, Jose Alcami, Oliver Hartley, Anne Brelot, Fernando Arenzana-Seisdedos, and Bernard Lagane. HIV-1 exploits CCR5 conformational heterogeneity to escape inhibition by chemokines. *Proceedings of the National Academy of Sciences*, 110(23) :9475–9480, Juin 2013.

[66] Reem Berro, Per Johan Klasse, Danny Lascano, Ayanna Flegler, Kirsten A. Nagashima, Rogier W. Sanders, Thomas P. Sakmar, Thomas J. Hope, and John P. Moore. Multiple

CCR5 Conformations on the Cell Surface Are Used Differentially by Human Immunodeficiency Viruses Resistant or Sensitive to CCR5 Inhibitors. *Journal of Virology*, 85(16) : 8227–8240, Août 2011.

[67] Ravinder Abrol, Bartosz Trzaskowski, William A. Goddard, Alexandre Nesterov, Ivan Olave, and Christopher Irons. Ligand- and mutation-induced conformational selection in the CCR5 chemokine G protein-coupled receptor. *Proceedings of the National Academy of Sciences*, 111(36) :13040–13045, Septembre 2014.

[68] Philippe Colin, Zhicheng Zhou, Isabelle Staropoli, Javier Garcia-Perez, Romain Gasser, Marie Armani-Tourret, Yann Benureau, Nuria Gonzalez, Jun Jin, Bridgette J. Connell, Stéphanie Raymond, Pierre Delobel, Jacques Izopet, Hugues Lortat-Jacob, Jose Alcami, Fernando Arenzana-Seisdedos, Anne Brelot, and Bernard Lagane. CCR5 structural plasticity shapes HIV-1 phenotypic properties. *PLOS Pathogens*, 14(12) :e1007432, Décembre 2018.

[69] Inna Slynko, Franck Da Silva, Guillaume Bret, and Didier Rognan. Docking pose selection by interaction pattern graph similarity : application to the D3R grand challenge 2015. *Journal of Computer-Aided Molecular Design*, 30(9) :669–683, Septembre 2016.

Chapitre 2

Signatures structurales du CCR5 lié à quatre variantes de la gp120

2.1 Introduction

Le CCR5 est une protéine flexible qui existe dans de multiples conformations [1, 2]. La reconnaissance spécifique de sous-populations du CCR5 par la gp120 permet d'expliquer pourquoi le virus parvient à échapper à l'inhibition de l'entrée par les chimiokines [3, 4]. L'hétérogénéité conformationnelle du CCR5 explique également la variabilité du potentiel infectieux des souches R5 du VIH-1, comme l'a montré Bernard LAGANE et son équipe par l'évaluation minutieuse de la liaison au corécepteur de plusieurs gp120 dérivées de virus isolés de cellules sanguines d'individus séropositifs [5]. Dans cette étude, la quantité de CCR5 utilisée par la gp120 varie en fonction de l'origine de la protéine virale, du type de cellules qui expriment le CCR5, et des conformations du CCR5, telles qu'elles sont distinguées par des anticorps monoclonaux. L'occupation du CCR5 par les gp120 est indépendante de la liaison au récepteur CD4. Par ailleurs, il est intéressant de noter que le panel étudié comporte des gp120 issues de deux virus isolés chez le même individu, l'un pendant la phase chronique de l'infection et l'autre au moment du diagnostic du SIDA. Les deux gp120 présentent une occupation minimale et maximale du corécepteur, tout en ayant la même affinité pour le ce dernier. Ces deux gp120 ont été nommées respectivement #25 et #34, d'après le numéro attribué à la souche virale primaire correspondante.

Ce chapitre présente la modélisation à l'échelle atomique de la reconnaissance du CCR5 par des variantes de la gp120. Cette étude a pour objectif d'identifier les caractéristiques structurales propres aux différentes gp120, pour une meilleure compréhension du phénotype viral. Nous avons considéré la gp120 de quatre souches virales à tropisme R5 : les deux souches primaires #25 et #34, et deux souches adaptées en laboratoire et largement utilisées pour l'étude *in vitro* du VIH-1, à savoir JF-RL et Bx08. Les quatre gp120 ont des séquences très similaires et des mutations, insertions et délétions se produisant principalement dans des boucles variables (FIGURE 2.1). La simulation de la dynamique moléculaire des quatre trimères CCR5–gp120–CD4 correspondants a été analysée sous plusieurs angles : l'arrangement relatif du CCR5 et de la gp120, la conformation de la gp120, la conformation du CCR5 et le mode d'interaction entre la gp120 et le CCR5.

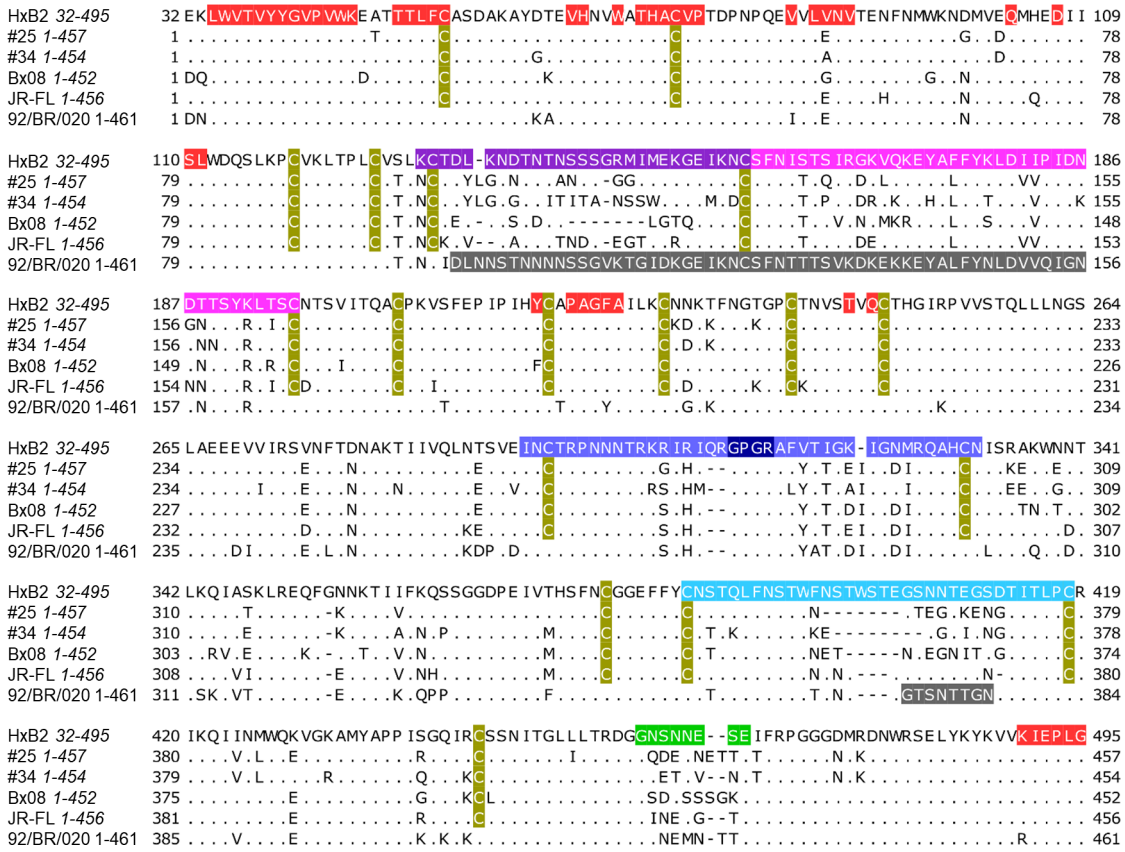

FIGURE 2.1 – Alignement de séquences multiples de gp120. La première séquence HxB2 est utilisée comme référence d'un point de vue de la séquence en acide aminés et de la numérotation des résidus de la gp120. Seules l'insertion, la suppression et les variations sont indiquées pour les autres séquences de gp120, à l'exception des résidus de cystéine engagés dans des ponts disulfure (fond jaune foncé). Les séquences étudiées ici sont les séquences #25, #34, Bx08 et JR-FL. La dernière séquence 92/BR/020 est tirée du complexe crsitallographique avec CCR5 et CD4 (entrée PDB : 6MEO) [6]. Les boucles variables de la gp120 sont mises en évidence dans la première séquence (V1 en violet, V2 en mauve, V3 en bleu, V4 en cyan et V5 en vert. Notez que la pointe de la V3 est indiquée par une couleur plus foncée). Le site de liaison de la gp41, tel qu'identifié dans le complexe *env* fermé (entrée PDB : 5CEZ) est mis en évidence en rouge.

2.2 Matériel et méthodes

2.2.1 Séquences de la gp120

Obtention des séquences

Dans ce chapitre, six séquences de gp120 issues de VIH-1 à tropisme R5 ont été considérées : HxB2, #25, #34, Bx08, JR-FL et 92/BR/020. La première HxB2 a été utilisée comme référence afin d'obtenir une numérotation homogène entre les différentes variantes de gp120 après alignement des cinq autres [7]. La variante HxB2 est souvent utilisée comme tel dans de nombreux travaux et est bien documenté avec notamment une définition standard des différentes régions (V1, V2, ...). La séquence des protéines *env* de la variante HxB2 a été obtenue sur le site Uniprot (`www.uniprot.org`, numéro d'accession : P04578). Les séquences #25 et #34 sont issues de deux souches de VIH-1 d'un même patient clinique avec #25 prélevée lors de la phase précoce de l'infection (25 mois après la séroconversion) et #34 prélevée lors de la phase du SIDA (128 mois après la séroconversion) [5]. Les séquences de la gp120 des variantes Bx08 et JR-FL ont été obtenues depuis la base de données des séquences du VIH du laboratoire national de Los Alamos [8] (`www.hiv.lanl.gov`, numéro d'accession : AY713411 et U63632 pour Bx08 et JR-FL respectivement). La dernière 92/BR/020 correspond à la variante de la structure cryo-EM décrivant le complexe CCR5–gp120–CD4 (entrée PDB : 6MEO) [6].

Alignement des séquences de la gp120

Les six séquences ont été alignées avec l'outil *Protein Align/Superpose* du logiciel MOE 2018.01 avec les réglages par défaut et la séquence de HxB2 comme référence. Les séquences ont été tronquées dans les parties N-terminale et C-terminale afin de correspondre à la séquence en acide aminés de la protéine décrite dans la structure 6MEO à savoir les résidus Glu32 à Gly495 (numérotation HxB2).

2.2.2 Définition des régions du CCR5 et de la gp120

CCR5

Dans cette étude, le CCR5 est constitué de 313 résidus. Sa séquence est native et complète jusqu'à l'hélice 8, la partie C-terminale étant manquante à partir du résidu Lys314. CCR5 faisant partie des RCPG de classe A, il est caractérisé par 7 hélices transmembranaires (TM) reliées entre elles par des boucles extracellulaire (ECL) et intracellulaire (ICL). La définition des hélices et des boucles correspond à celle du site GPCRdb [9] (`gpcrdb.org/protein/ccr5\_human/`, N-terminal : 1 à 21 ; TM1 : 22 à 58 ; ICL1 : 59 à 62 ; TM2 : 63 à 92 ; ECL1 : 93 à 96 ; TM3 : 97 à 132 ; ICL2 : 133 à 140 ; TM4 : 141 à 167 ; ECL2 : 168 à 185 ; TM5 : 186 à 224 ; ICL3 : 225 à 227 ; TM6 : 228 à 265 ; ICL3 : 266 à 267 ; TM7 : 268 à 300 ; H8 : 301 à 313). Une définition plus stricte des hélices a été rajoutée, *TM CCR5*, et correspond aux résidus insérés dans la membrane, TM1 : 31 à 57 ; TM2 : 66 à 86 ; TM3 : 99 à 129 ; TM4 : 145 à 162 ; TM5 : 190 à 219 ; TM6 : 237 à 254 ; TM7 : 277 à 300. Ces résidus sont dans les parties les plus rigides du CCR5 et cette définition est utilisée pour les tous les alignements structuraux de CCR5.

gp120

Les annotations des domaines variables de gp120 sont issues de la base de données des séquences du VIH du laboratoire national de Los Alamos [8] (`www.hiv.lanl.gov`) pour la souche HxB2, V1 : 130 à 157 ; V2 : 156 à 196 ; V3 : 294 à 332 ; V4 : 385 à 418 ; V5 : 459 à 466. Les parties de la séquence les plus conservées dans la gp120 définissent le cœur de la protéine (*gp120 core*), qui exclut les boucles variables et les régions N-terminale et C-terminale (C1 : 74 à 129 ; C2 : 197 à 293 ; C3 : 333 à 384 ; C4 : 419 à 458 ; C5 : 467 à 490). Ces parties sont aussi les plus rigides de gp120 et cette définition est par conséquent utilisée pour les alignements structuraux de gp120.

2.2.3 Construction des modèles

Les modèles du complexe CCR5–gp120–CD4 ont été construits à partir de la structure décrite dans la publication de Md Munan SHAIK et ses collaborateurs (entrée PDB : 6MEO) [6].

Les séquences du CCR5 et du CD4 n'ont pas été modifiées car natives toutes les deux dans la structure PDB. La structure du CCR5 (numéro d'accession Uniprot : P51681) est complète de l'extrémité N-terminale jusqu'à la fin de l'hélice 8 et comporte ainsi les résidus Met1 à Gln313, inclus. Les résidus tyrosine en position 10 et 14 sont représentés sous forme de sulfotyrosine dans la structure PDB, et ont été laissées tels quels. Pour le CD4 (numéro d'accession Uniprot : P01730), seul les domaines 1 et 2 sont présents et une partie est manquante au niveau du domaine N-terminal. La structure décrit le CD4 du résidus Lys26 à Val201.

La structure PDB décrit une gp120 glycosylée. Dans nos modèles, tous les oses ont été omis. Dans la structure PDB, la gp120 est par ailleurs incomplète. Les parties manquantes sont les V1, V2 et V4. Leur structure a donc été modélisée à partir d'une autre structure de la gp120, dans laquelle elles sont présentes (entrée PDB : 3J70). Cette entrée décrit la structure du trimère de gp120–gp41, dans sa forme ouverte stabilisée par un anticorps et le CD4. En pratique, les coordonnées des résidus Val127 à Ser195 (V1/V2) et Asn386 à Pro417 (V4) ont été extraites de la chaîne D du fichier PDB 3J70 avec le logiciel MOE 2018.01 puis greffées dans le modèle de la gp120 construit à partir du fichier PDB 6MEO avec l'outil *Loop Grafting* du logiciel MOE 2018.01, permettant la reconstruction des fragments de séquence Val126 à Ser195 (V1/V2) et Asn384 à Pro412 (V4). À partir de ce modèle complet de la gp120, quatre autres ont été construits par homologie à partir des séquences de la gp120 #25, #34, Bx08 et JRFL avec l'outil *Homology Model* (options par défaut, meilleur modèle) du logiciel MOE 2018.01. L'état de protonation des systèmes a été corrigé avec l'outil *Protonate3D* (options par défaut) du logiciel MOE 2018.01.

Les structures des quatre modèles du complexe CCR5–gp120–CD4 ont été alignées sur une structure de référence au laboratoire pour le CCR5 libre, et ce afin d'avoir un jeu de coordonnées

équivalent pour tous les modèles et qui convient à la construction des systèmes simulés par dynamique moléculaire.

2.2.4 Préparation des systèmes pour la simulation par dynamique moléculaire

Les quatre modèles (CCR5–gp120[#25]–CD4, CCR5–gp120[#34]–CD4, CCR5–gp120[bx08]–CD4 et CCR5–gp120[JR-FL]–CD4) ont été préparés pour la simulation par dynamique moléculaire avec le webservice CHARMM-GUI version 2.1 (#25 et #34) et version 3.0 (Bx08 et JRFL) [10, 11]. Pour mimer un environnement biologique, nous avons inséré la partie TM du CCR5 dans une bicouche lipidique avec l'outil *Bilayer builder*. La bicouche choisie est composée de phosphatidylcholine (POPC), de phosphatidyléthanolamine (POPE) et de cholestérol (CHL) en proportion 2:2:1 dans ses deux parties, supérieure et inférieure. Le CCR5 a été inséré dans la bicouche en utilisant l'option "Use PDB Orientation", les structures du CCR5 ayant été préalablement alignées sur une référence dans laquelle l'hélice 8 se positionne dans le plan de l'interface lipide/eau. Les noms des résidus ont été changés afin de correspondre à la convention de nommage de CHARMM (CYX en CYS, TYS en TYR, HID en HSD et HIE en HSE). La boîte est de forme rectangulaire avec des côtés de longueur $a = b = 200\,\text{Å}$. La couche d'hydratation a une épaisseur de 22,5 Å de part et d'autre de la bicouche sur l'axe z aboutissant à une hauteur $c = 175\,\text{Å}$ environ (légèrement différent en fonction de la variantes de la gp120). Des ions K^+ et Cl^- ont été ajoutés afin d'obtenir une concentration finale égale à 0,15 M. Ils ont été placés par la méthode de Monte-Carlo.

CHARMM-GUI positionne les protéines de telle sorte que les axes principaux soient alignés avec ceux de la boîte. Le complexe CCR5–gp120–CD4 possédant une forme globale en T, un volume important de la boîte ne comportait que de l'eau. La taille du système a donc été réduit en alignant le 2ème axe principal sur la petite diagonale de la boîte avec un script python "maison" `shrinkbox.py` (voir l'annexe à la fin du chapitre). Les longueurs sont ainsi réduites à $a = b = 120\,\text{Å}$ environ (dépendante de la variante). Le système réduit a été ensuite traité par le script `charmmlipid2amber.py` version 2.0.3 (disponible sur le site de CHARMM-

GUI) afin d'obtenir un nommage des résidus des lipides, de l'eau et des ions adéquat pour AMBER. Les noms des résidus histidine ont été renommés par un script python "maison" `split_assembly.py` (voir l'annexe en fin de chapitre) qui en plus sépare les différents objets du système en quatre fichiers PDB distincts en fonction de la nature moléculaires : protéines, membrane, eau et ions. Les résidus Tyr10 et Tyr14 de CCR5 ont été modifiés en sulfotyrosine (CHARMM-GUI modifie les TYS en TYR) avec l'outil *Protein Builder* du logiciel MOE version 2018.01. Le nom de résidu de paires de cystéines a été modifié en CYX dans les fichiers PDB et les atomes d'hydrogène des groupes sulfhydryle supprimés si la distance S – S est inférieure à 2,5 Å avec un script "maison" `fix_SSbridge.py` (voir l'annexe à la fin du chapitre). À ce stade, une inspection visuelle des structures a été réalisée avec le logiciel MOE, afin de détecter d'éventuelles interpénétrations de cycles (consignées dans les fichiers de sortie de CHARMM-GUI), et de modifier en conséquence l'angle de torsion des résidus incriminés. Le fichier PDB contenant les protéines a ensuite été scindé en autant de fichiers que de protéines (CCR5, gp120 et CD4) par un script python "maison" `split_proteins.py` (voir l'annexe à la fin du chapitre), qui ajoute une balise TER à la fin du bloc de définition des atomes. Les fichiers de topologie et de coordonnées, qui sont les entrées requises pour la simulation par dynamique moléculaire, ont été générés avec le programme tleap de la suite AMBER16 [12] et en utilisant les paramètres des champ de force *ff14SB* pour les protéines, *lipid14* pour la bicouche lipidique et TIP3 pour les molécules d'eau [13, 14]. Les paramètres pour le résidu sulfotyrosine ont été déterminés dans une précédente étude [1]. Les charges des systèmes ont également été neutralisées si besoin par ajout de contre-ions K^+ ou Cl^-.

2.2.5 Simulations par dynamique moléculaire

Les simulations ont été effectuées au centre de calcul haute performance de l'institut national de physique nucléaire et de physique des particules (IN2P3) sur une ferme de processeurs graphiques (GPU Nvidia® Tesla K80) avec le programme *pmemd.cuda* de la suite d'AMBER16 et la bibliothèque CUDA 8.0 [15, 16]. Les systèmes CCR5–gp120[#25]–CD4, CCR5–gp120[#34]–CD4, CCR5–gp120[Bx08]–CD4 et CCR5–gp120[JR-FL]–CD4 ont été d'abord minimisés en 15000 étapes en utilisant la méthode de la plus forte pente, ou steepest-descent, pour

les 10000 premières étapes et la méthode des gradients conjugués pour les 5000 suivantes. Ensuite, les systèmes ont été chauffés de 0 à 300 K en appliquant le thermostat de Langevin sur une durée de 300 ps à volume constant en contraignant les coordonnées atomiques. Les contraintes ont été progressivement relachées pendant 800 ps à volume constant. Les systèmes ont été équilibrés pendant 10 ns à pression constante. Pour l'étape de production, chaque système est répliqué trois fois sur une durée de 10 étapes $\times$ 10 ns en changeant la graine du thermostat de Langevin à chaque étape.

2.2.6 Analyse des trajectoires

Déviation des coordonnées atomiques de la chaîne principale des protéines

Pour chaque variante (#25, #34, Bx08 et JR-FL), la déviation des coordonnées atomiques au cours des périodes d'équilibrage et de production a été estimée par le RMSD entre les structures instantanées de la trajectoire et la structure minimisée de départ. Ces distances sont calculées avec la commande *rmsd* du logiciel CPPTRAJ version 17.00. [17] pour les atomes Cα, C, O et N de la chaîne principale, après l'alignement des deux structures comparées pour la meilleure superposition de la protéine ou du domaine considéré. Au total, six alignements différents des structures instantanées de la trajectoire ont été réalisés : sur tous les résidus du complexe protéique ("All") ; sur les 313 résidus du CCR5 ("CCR5") ; sur les résidus du CCR5 insérés dans la bicouche ("CCR5 TM", voir plus haut dans la rubrique "Définition des régions de CCR5 et gp120") ; sur tous les résidus de gp120 ("gp120") ; sur le cœur de la gp120 ("gp120 core", voir plus haut dans la rubrique "Définition des régions de CCR5 et gp120") ; et sur tous les résidus du CD4 ("CD4"). Les valeurs de RMSD en fonction du temps ont été tracées avec un script Python version 3.7.3 et le module `Matplotlib` version 3.1.1 (FIGURE 2.4).

Fluctuation des coordonnées atomiques par résidu

Pour chaque variante (#25, #34, Bx08 et JR-FL), la fluctuation atomique au cours de la période de production a été calculée par une moyenne quadratique, par résidu, des distances

entre l'atome Cα dans les structures instantanées de la trajectoire et l'atome Cα d'une structure moyenne, après l'alignement des régions rigides des structures instantanées de la trajectoire sur celle de la structure moyenne (ces régions rigides du complexe ont été définies plus haut, "CCR5 TM" et "gp120 core"). La structure moyenne (une par variante) correspond aux coordonnées moyennes des atomes Cα de toutes les structures instantanées de production préalablement alignées sur la structure minimisée de départ, en considérant aussi les régions le plus rigides du complexe. Toute la procédure a été effectuée avec un script Python version 3.7.1 et le module `MDAnalysis` version 0.20.1. La figure associée (FIGURE 2.5) a été générée avec un script Python v3.7.3 et le module `Matplotlib` version 3.1.1.

Partitionnement des dynamiques

Les structures des dynamiques de production ont été partitionnées par un algorithme hiérarchique ascendant en utilisant le RMSD comme mesure de distance en ayant préalablement aligné tous les atomes, sauf hydrogène, des parties rigides du CCR5 ("CCR5 TM", voir plus haut) des structures instantanées sur celles de la première d'entre elle avec la commande *cluster* de CPPTRAJ version 17.00, afin de produire cinq partitions par répliqua (15 partitions par variante de la gp120).

Projection des extrémités des hélices extracellulaires du CCR5

Les structures ont d'abord toutes été alignées sur les parties rigides du CCR5 ("CCR5 TM", voir plus haut) d'une unique structure du CCR5. Ne sont ensuite considérés que les résidus des extrémités extracellulaire et intracellulaire des domaines transmembranaires, à condition qu'ils soient dans une conformation d'hélice-α dans 50 % des structures instantanées (*dssp* dans CPPTRAJ) et le bout libre de la V3 (résidus Gly312, Pro313, Gly314 et Arg315). Les coordonnées *x* et *y* de l'atome Cα d'un résidu sélectionné sont tracées pour obtenir un nuage de points, ou scatter plot, dont la forme est déterminée par une estimation par noyau suivant une loi normale où les contours délimitent 90 % de la densité. La méthode de projection est détaillée dans le chapitre suivant.

Cartographie des fréquences d'interactions non covalentes intermoléculaires

Toute la procédure a été effectuée sur les structures instantanées issues de la phase de production avec un pas de 10 (1 structure sur 10) par des scripts Python v3.7.3 et le module `MDAnalysis` version 0.20.1. La matrice de distance atome/atome inter-résidu est calculée pour chaque structure instantanée sélectionnée. En fonction des propriétés des atomes du couple considéré (accepteur/donneur d'hydrogène, ionique positif/négatif ou hydrophobe), on établit qu'une interaction est présente si des critères de distance et d'angle (liaison hydrogène) sont respectés [18]. Les cartes d'interaction sont générées avec un script Python version 3.7.3 et le module `Matplotlib` version 3.1.1.

Carte des corrélations croisées des fluctuations atomiques

L'analyse a été effectuée sur l'intégralité des structures de l'étape de production des simulations. Les structures ont été préalablement alignées sur les Cα des résidus rigides de la gp120 ("gp120 core", voir plus haut) de la structure minimisée. Les valeurs de corrélation ont ensuite été calculées avec le commande *diagmatrix* du programme CPPTRAJ V17.00 avec les options *vecs 1024 nmwiz nmwizvecs 100*. Les cartes de corrélation de mouvement sont générées avec un script Python version 3.7.3 et le module `Matplotlib` version 3.1.1.

2.3 Résultats

2.3.1 Description générale de la modélisation et des simulations par dynamique moléculaire

La structure du trimère CCR5–gp120–CD4 a été déterminée récemment par cryo-EM à une résolution de 3,9 Å (FIGURE 2.2).

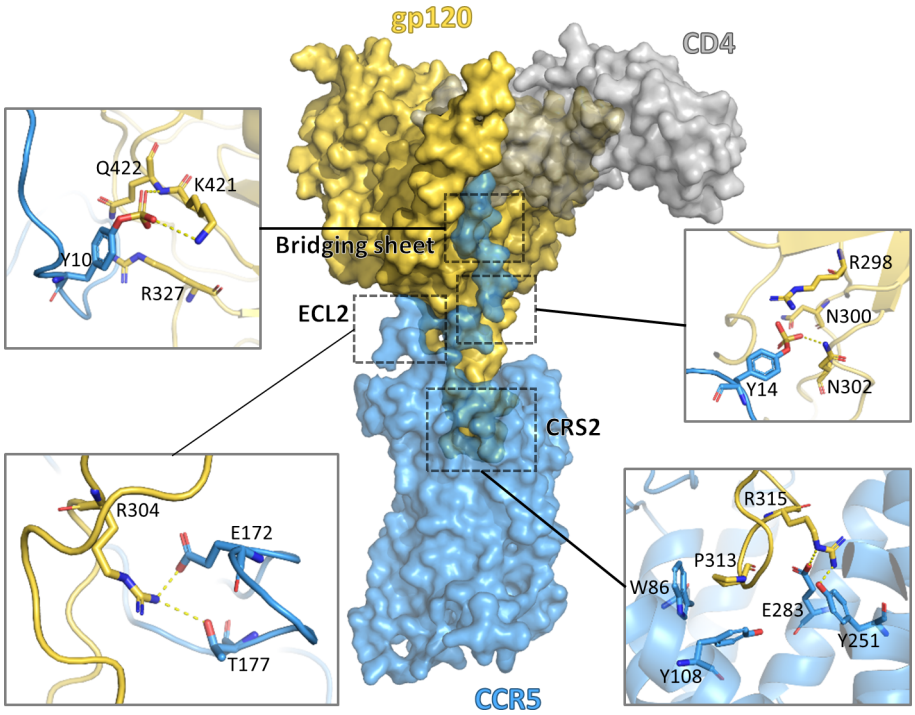

FIGURE 2.2 – Structure du complexe CCR5–gp120–CD4 de l'entrée PDB 6MEO.

Elle rend compte à l'échelle atomique du mode de liaison de la gp120 au CD4 et au CCR5 (entrée PDB : 6MEO). Point important, ni la séquence du CCR5 ni celle du CD4 n'ont été modifiées pour les besoins de l'étude structurale. La séquence du CCR5 est donc native, avec un domaine N-terminal complet et qui contient deux modifications post-traductionnelles fonctionnellement importantes, à savoir la sulfatation des résidus tyrosine sur les positions 10 et 14. La partie C-terminale de CCR5 située dans la région intracellulaire est absente à partir du résidu 314. Le CD4 a été tronqué à sa forme soluble, avec seulement les deux premiers domaines de type immunoglobuline, dont seul le premier est en contact avec la gp120. La gp120 est issue

de la souche adaptée en laboratoire 92/BR/020. La comparaison de sa séquence avec celles des quatre gp120 de cette étude indique 78 à 84 % d'identité de séquence. La glycosylation de la gp120 a été décrite dans la structure cryo-EM, mais n'a pas été transférée dans les modèles en raison de l'incertitude quant à la conservation des sites.

Sur la base de ces données expérimentales, nous avons construit les modèles de structure tridimentionnelle du complexe ternaire CCR5–gp120–CD4 pour les souches virales #25, #34, bx08 et JR-FL, en ne modifiant que la séquence gp120. Comme les boucles V1, V2 et V4 de la gp120 sont absentes dans la structure servant de gabarit (6MEO), les parties de la protéine ont été construites sur la base d'une autre structure expérimentale [19]. Le CCR5 a été incorporé dans une bicouche de lipides, puis le complexe ternaire CCR5–gp120–CD4 a été placé dans une boîte d'eau contenant des ions potassium et des ions chlorures. Les quatre systèmes, correspondant aux quatre variantes de la gp120, ont chacun été soumis à trois simulations indépendantes par dynamique moléculaire. Le temps d'une simulation a été limité à 100 ns car au delà, la gp120 ou la CD4 venait au contact avec la bicouche lipidique dans la plupart des simulations. Cet événement n'est pas biologiquement pertinent, d'autant que le système ne comprend pas de glycosylation et que le CD4 est incomplet sans ancrage à la membrane. Par souci de simplicité, les quatre systèmes modélisés seront dorénavant nommés en fonction de la souche virale de la gp120 dans le trimère : #25, #34, Bx08 et JR-FL.

2.3.2 Tous les modèles simulés du complexe CCR5–gp120–CD4 s'écartent de la structure résolue par cryo-EM

Dans les trois simulations des quatre modèles, la forme globale du trimère CCR5–gp120–CD4 a évolué de manière significative, avec une réorientation des trois protéines les unes par rapport aux autres. Plus précisément, la gp120, qui était initialement alignée en ligne droite avec le domaine transmembranaire de CCR5, a basculé vers la membrane (FIGURE 2.3).

Le mouvement implique la flexion d'une région charnière qui englobe la base de la boucle V3 de la gp120 et la partie de CCR5 qui s'y lie, à savoir le domaine N-terminal. Globalement,

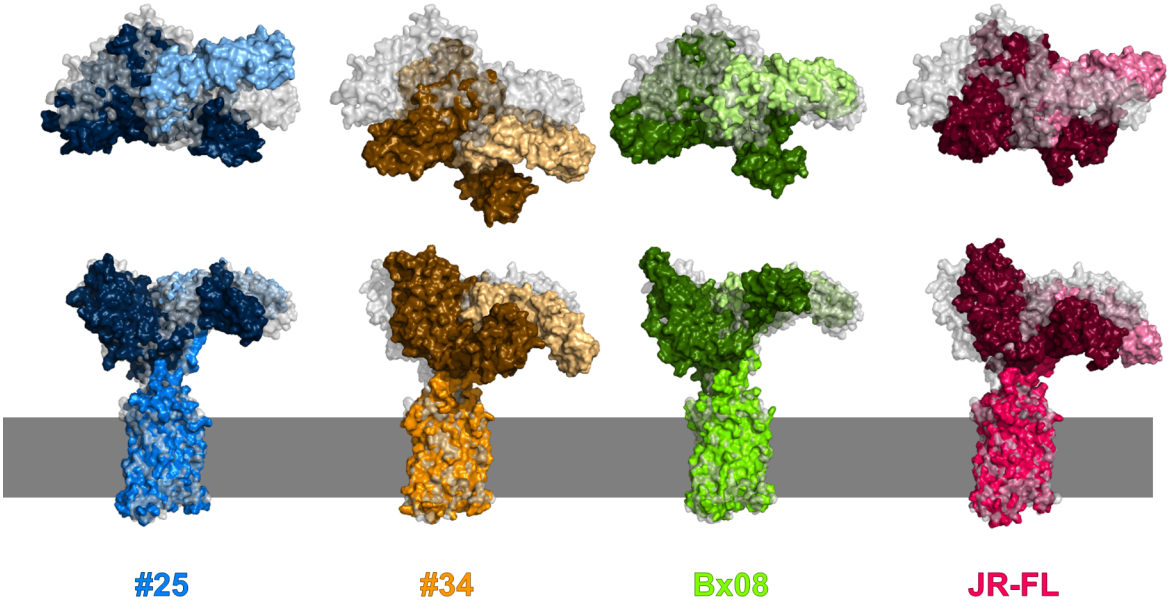

FIGURE 2.3 – Dernière structure des trajectoires du complexe CCR5–gp120–CD4. Les variantes de la gp120 sont colorées comme suit : #25 en bleu, #34 en orange, Bx08 en vert et JR-FL en rouge. CCR5, gp120 et CD4 sont représentés par des surfaces moléculaires, colorées dans des tons normaux, foncés et clairs respectivement. À titre de comparaison, la première structure de la trajectoire est également représentée, en gris (entrée PDB : 6MEO). Un point de vue du dessus de la membrane lipidique est montré en premier, un point de vue dans le plan de la membrane lipidique, qui est représenté schématiquement par une bande grise, est montré ensuite.

le mouvement est plus important dans le complexe #34 que dans les trois autres complexes. Cependant, quel que soit le système, le cœur de la gp120 n'a pas de position privilégiée par rapport au domaine à sept hélices transmembranaires (7TM) de CCR5. En effet, les trois simulations effectuées pour le même complexe, partant exactement de la même conformation, aboutissent à des structures finales différentes, alors que l'orientation du 7TM du CCR5 dans la membrane est conservée. La déviation des coordonnées de atomes Cα du 7TM par rapport à celle de la structure cryo-EM est faible, avec un RMSD de $1{,}1 \pm 0{,}1$ Å.

2.3.3 Les parties les plus flexibles du complexe CCR5–gp120–CD4 sont les boucles variables

Une fois la structure du complexe stabilisée, le CD4 ne subit d'autres changements structuraux importants, la fluctuation moyenne des atomes Cα dans la phase de production étant inférieure à 1,1 Å (FIGURE 2.4).

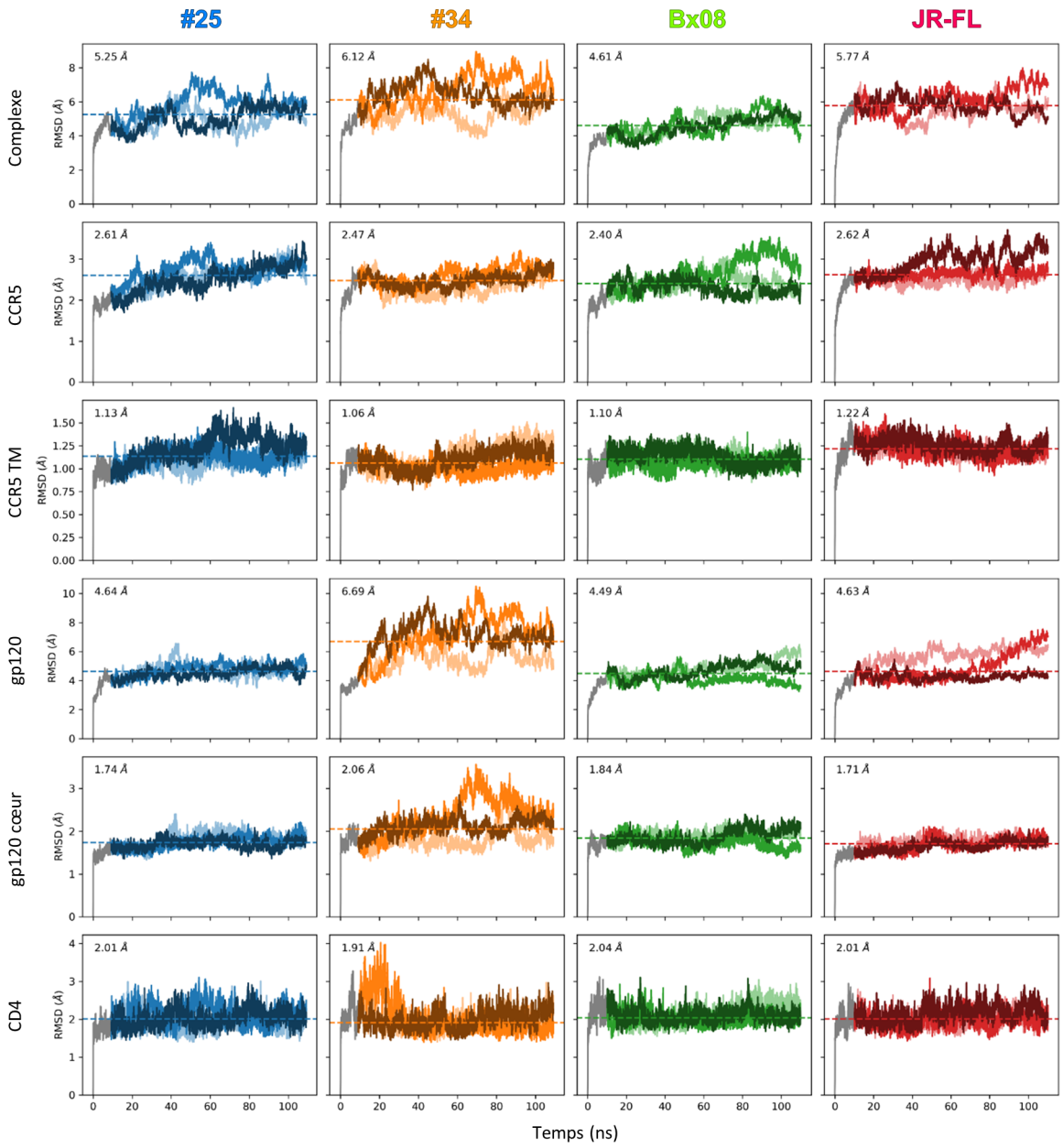

FIGURE 2.4 – Déviation des cordonnées atomiques au cours des dynamiques. Chaque ligne correspond à une région d'alignement différente. Chaque colonne correspond à un système différent. La structure de référence utilisée pour l'alignement correspond à la structure après minimisation. La partie grisée correspond à la phase d'équilibrage. La partie coloriée correspond à la phase de production des 3 réplicas. La valeur indiquée dans le coin supérieur gauche de chaque cellule correspond à la déviation médiane de la phase de production tous réplicas confondus.

La même observation est faite pour les hélices transmembranaires du CCR5, dont les mouvements ont été fortement contraints par la bicouche lipidique, et pour le cœur de gp120, dont la conformation est extrêmement bien conservée dans les nombreuses structures résolues expérimentalement par cristallographie aux rayon X ou par cryo-EM. Dans les quatre systèmes #25, #34, Bx08 et JR-FL, la structure des cinq boucles de la gp120 change au cours du temps (FIGURE 2.5).

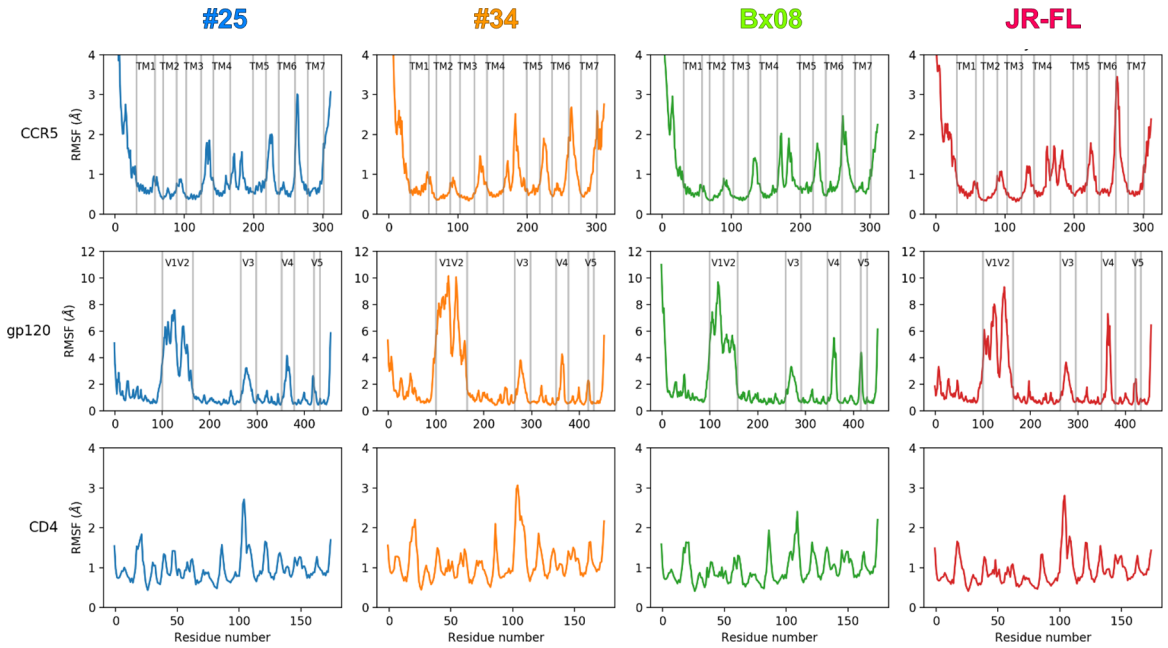

FIGURE 2.5 – Fluctuation des atomes Cα du CCR5, de la gp120 et du CD4 pendant la phase de production. Les structures ont été alignées sur la structure après minimisation.

Les fluctuations atomiques sont nettement plus importantes dans V1 et V2 que dans les autres boucles. À l'instar de sa plus grande réorganisation au sein du complexe (voir ci-dessus), la variante #34 présente les plus grandes fluctuations des boucles V1 et V2 (FIGURE 2.6).

Dans CCR5, les boucles intracellulaires et extracellulaires sont également plus flexibles que la partie du récepteur insérée dans la membrane (TM), mais dans une moindre mesure que les boucles de gp120. À titre d'exemple, la fluctuation maximale par résidu est d'environ 3,5 Å dans la ECL3 du CCR5 dans le système JF-RL, et elle est supérieure à 10 Å dans dans la V1 de la gp120 du système #34. Il est important de noter que pour le CCR5 et comme pour la gp120, le profil défini par les fluctuations atomiques le long de la séquence est spécifique au système,

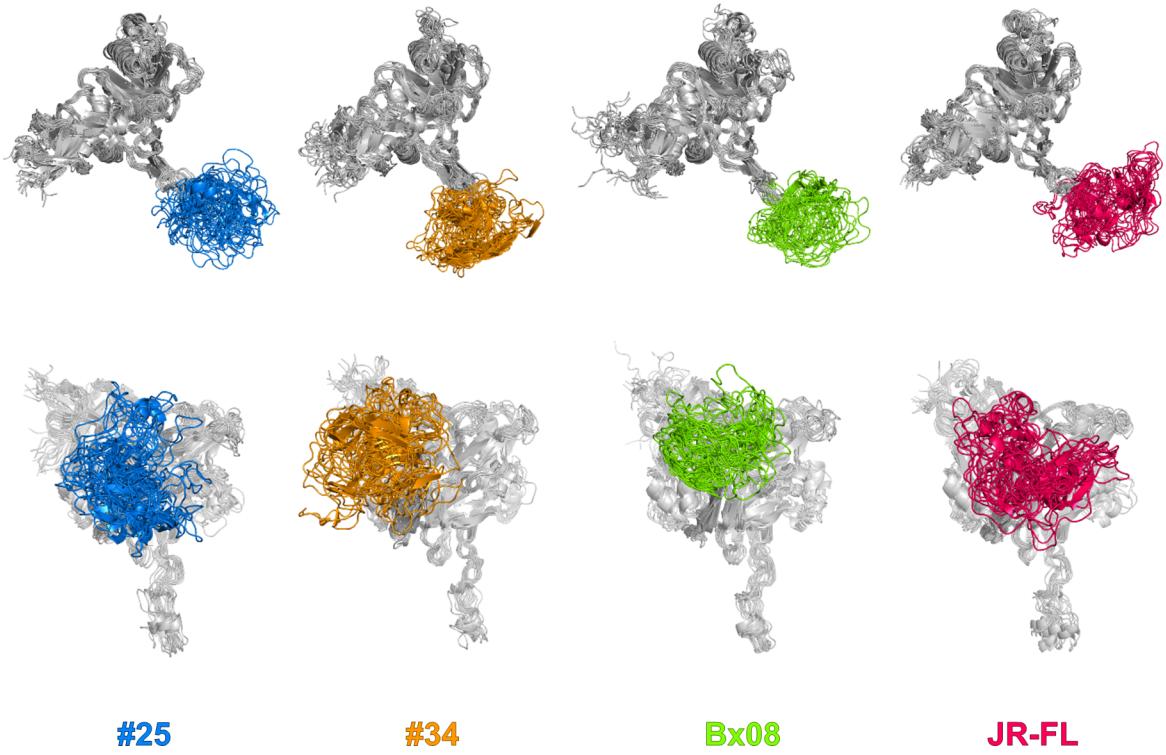

FIGURE 2.6 – Échantillonnage des structures de la gp120 tout au long de la trajectoire. Les variantes de la gp120 sont colorées comme suit : #25 en bleu, #34 en orange, Bx08 en vert et JR-FL en rouge. Un total de 15 structures représentatives a été sélectionné pour chaque variante. CCR5 et CD4 n'ont pas été représentés. La chaîne principale de la gp120 est représentée par des rubans gris, à l'exception des boucles variables V1 et V2 qui sont colorées selon les variantes. Un point de vue du dessus de la membrane lipidique est montré en premier, un point de vue dans le plan de la membrane lipidique est montré ensuite.

#25, #34, Bx08 ou JR-FL, suggérant ainsi que les changements dans la séquence de la gp120 ont effectivement un effet sur le comportement du CCR5.

2.3.4 La conformation du CCR5 s'adapte aux séquences de la gp120

Les boucles extracellulaires forment la partie du récepteur qui dépasse de la membrane. Les plus exposées, la deuxième boucle ECL2 et la troisième boucle ECL3 peuvent notamment moduler l'ouverture de la cavité transmembranaire du CCR5. Dans nos simulations, nous avons observé que la structure caractéristique en l'épingle à cheveux de l'ECL2 est préservée dans les systèmes #25, #34, Bx08 et JR-FL, mais que son inclinaison vers la cavité transmembranaire du récepteur varie. La différence de structure entre les quatre systèmes est encore plus marquée

pour l'ECL3, dont l'extrémité est orientée vers les lipides de la bicouche pour les systèmes #34 et Bx08 alors qu'elle tend à se situer au-dessus du récepteur dans le système JF-RL et d'avantage dans le système #25. Par conséquent, nos simulations ont montré que les conformations des boucles extracellulaires du CCR5 dépendent de la séquence de la gp120 liée. Quatre distances clés sont suffisantes pour distinguer les récepteurs dans les systèmes #25, #34, Bx08 et JR-FL : une première entre ECL2 et TM1, une seconde entre ECL2 et TM7, et deux autres entre ECL3-TM3 (FIGURE 2.7).

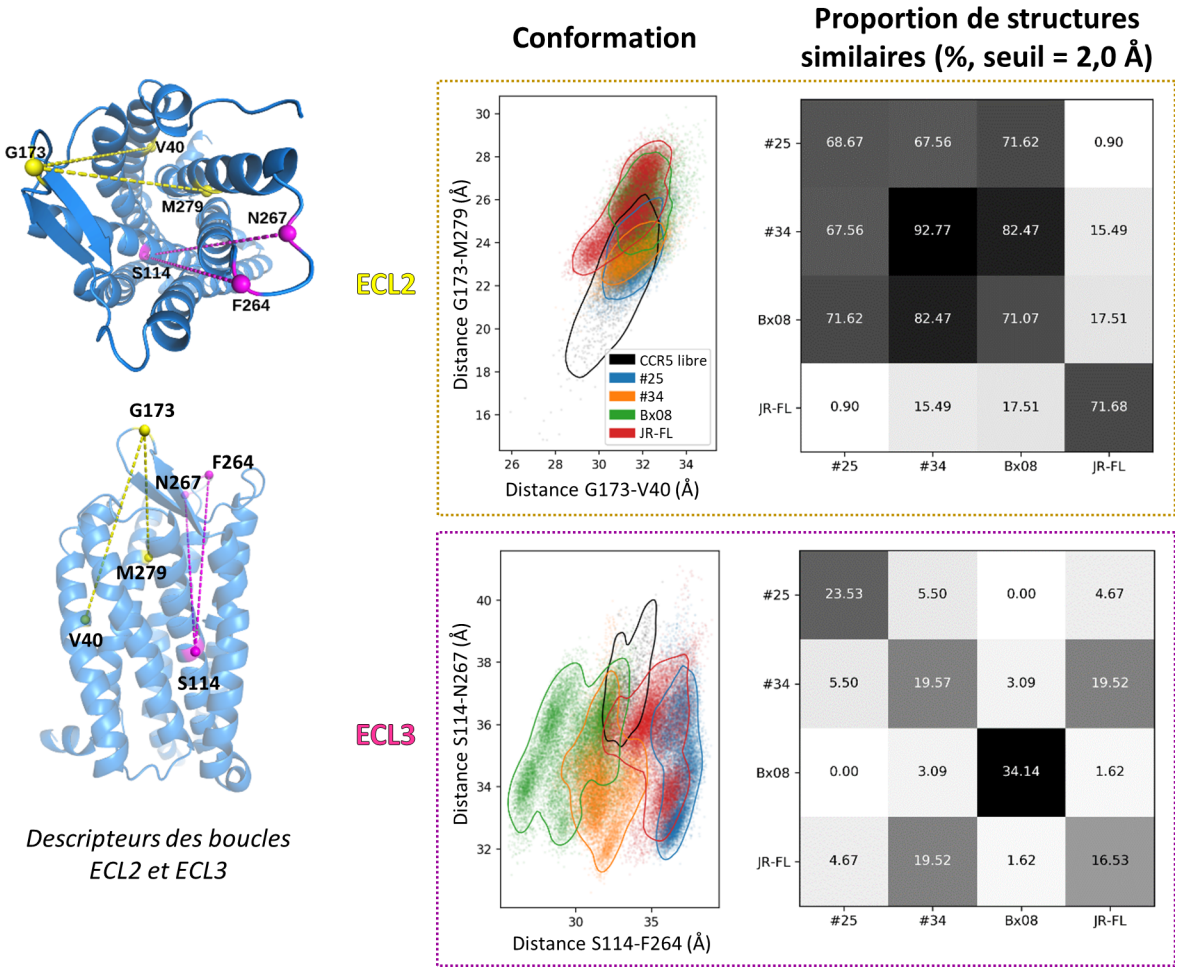

FIGURE 2.7 – Conformations des boucles ECL2 et ECL3 du CCR5 pendant les simulations. À gauche le CCR5 est représenté sous forme de ruban de couleur bleu. Les descripteurs caractérisant la boucle ECL2 (jaune) et ECL3 (magenta) sont représentés par des sphères et des lignes discontinues. Les graphiques sous le label "conformation" correspond aux valeurs des descripteurs des boucles ECL2 et ECL3. Les cartes de fréquences (heatmap en anglais) relatives à la "proportion de structures similaires" correspond au pourcentage de structures ayant un RMSD inférieur à 2,0 Å en ne considérant que les résidus de la boucle ECL2 (168 à 180) et ceux de la boucle ECL3 (264 à 269).

Dans nos simulations, les changements structuraux sont également observables dans le 7TM du récepteur, bien que, en moyenne, la chaîne principale dans le 7TM écarte peu de celle de la structure résolue par cryo-EM (environ 1,54 Å, voir ci-dessus). Là encore, les conformations explorées par le CCR5 sont différentes dans les quatre systèmes modélisés, en particulier dans l'extrémité extracellulaire des TM (FIGURE 2.8 à gauche).

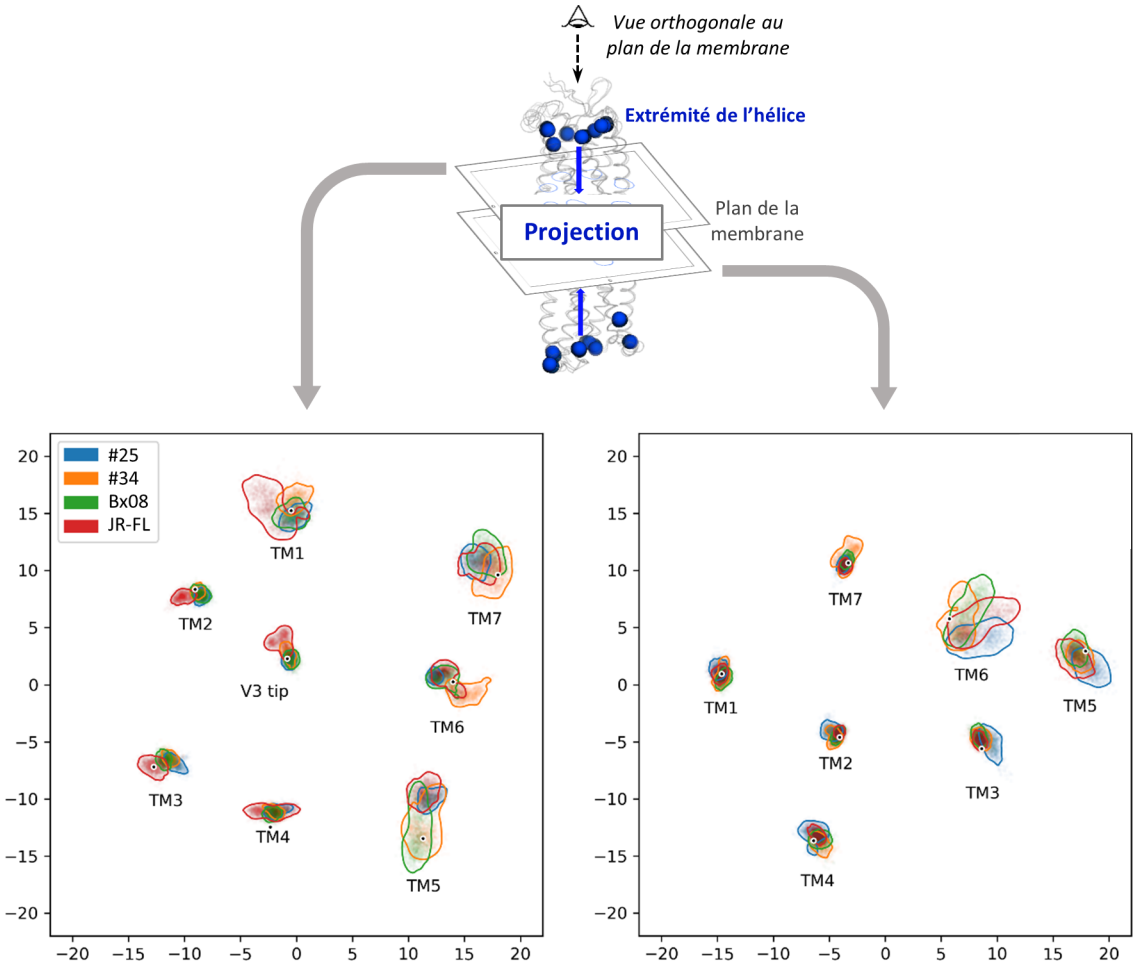

FIGURE 2.8 – Projection des positions des extrémités des hélices du CCR5 à partir des domaines extracellulaire (gauche) et intracellulaire (droite) au cours des simulations.

Dans le système JF-RL, les extrémités extracellulaires des TM1, TM2, TM3 et TM4 ont tendance à s'incliner vers les lipides, conduisant à l'élargissement de la cavité transmembranaire du récepteur. La position du TM1, qui est dans la continuité du domaine N-terminal, oscille pendant les simulations, tandis que les six autres hélices ont une position bien définie. Dans le système #25, les extrémités des sept hélices du 7TM sont bien définies, leur position variant très peu pendant les simulations. Dans le système #34, les extrémités extracellulaires des TM1,

TM5, TM6 et TM7 s'écartent du centre de CCR5 en se penchant vers les lipides membranaires, élargissant la partie opposée de la cavité transmembranaire, par rapport à ce qui a été observé dans le système JR-FL. Dans le système Bx08, seules les extrémités extracellulaires des TM5 et TM7 ont un tel comportement. Dans la partie intracellulaire du domaine transmembranaire, les différences sont moins marquées et concernent principalement le TM6 (FIGURE 2.8 à droite).

2.3.5 Les modes de liaison de la gp120 au CCR5 sont similaires mais néanmoins différents pour les quatre systèmes modélisés

Comme mentionné précédemment, la liaison de la gp120 à son corécepteur peut être considérée comme la reconnaissance mutuelle de deux bras flexibles attachés à deux parties rigides (FIGURE 2.2).

L'un des bras flexibles est constitué des 20 acides aminés N-terminaux du CCR5. Sa séquence alterne des résidus hydrophobes, polaires et chargés négativement. La reconnaissance de la gp120 implique principalement des contacts hydrophobes et des liaisons ioniques, en particulier avec les résidus sulfotyrosine sur les positions 10 et 14 [20]. Dans la structure du complexe CCR5–gp120–CD4 résolue par cryo-EM, les 30 acides aminés de la gp120 au voisinage de la partie N-terminale du CCR5 sont principalement situés à proximité de la partie appelée bridging sheet de la gp120, et également de la base de la boucle V3 de la gp120 (FIGURE 2.9).

Les résidus de la gp120 qui sont dans l'interface du dimère CCR5+-gp120 sont très bien conservés entre #25, #34, Bx08 et JR-FL, avec 26 positions strictement identiques. Les trois positions les plus variables sont 194, 322 et 440. Le résidu 194 est proche du résidu Met1 du CCR5, mais une interaction directe entre les deux résidus n'est observée que dans le système #25. Le residu 440 est proche des résidus sulfotyrosine (Tys) en positions 10 et 14 du CCR5. Cependant, dans l'ensemble, l'environnement local de Tys10 et de Tys14 est préservé dans les systèmes #25, #34, Bx08 et JR-FL (FIGURE 2.10). Dans toutes les simulations, des liaisons hydrogène sont formées fréquemment entre les deux résidus sulfotyrosine du CCR5 et les atomes de la chaîne principale des résidus Gln422, Ile423 et Asn302 de la gp120.

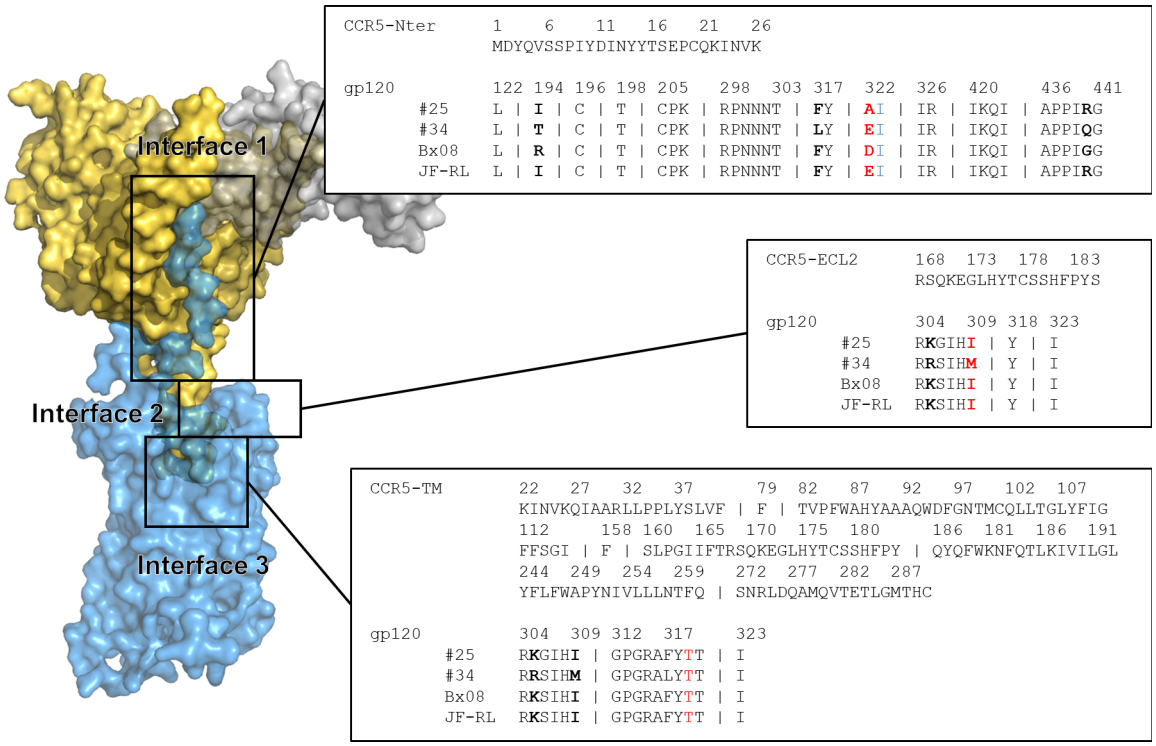

FIGURE 2.9 – Vue d'ensemble des 3 interfaces du CCR5 de la gp120 avec la séquence d'acides aminés correspondante. Les interfaces ont été déterminées sur la structure du modèle (entré PDB : 6MEO), en considérant les domaines de CCR5 et en sélectionnant les résidus gp120 proches (distance maximale = 4,5 Å). L'ECL2 dans CCR5 a été défini selon les annotations de la GPCRdb [21]. La cavité TM dans CCR5 a été détectée automatiquement (*Site Finder* dans MOE 2019.01). Les résidus en rouge sont en contact avec CCR5 pendant la simulation mais pas dans la structure cristallographique 6MEO. Les résidus en bleu sont numérotés 322A (insertion dans la séquence de référence HxB2). Les résidus en gras spécifient les positions où la mutation caractérise une variante de la gp120.

Globalement, les motifs définis par les liaisons hydrogène formées entre les résidus du domaine N-terminal de CCR5 et ceux de la gp120 sont similaires dans les systèmes #25, #34, Bx08 et JR-FL (FIGURE 2.11). Une caractéristique distincte du système #34 est néanmoins observée : deux liaisons hydrogène entre le résidu Tys14 de CCR5 et le résidu Arg298 de la gp120 persistent pendant les simulations. Ces deux liaisons hydrogène s'ajoutent à la liaison ionique formée entre ces deux mêmes résidus, liaisons présentes dans les quatre systèmes #25, #34, Bx08 et JR-FL (FIGURE 2.11).

La variation de séquence à la position 322 de la gp120 se traduit par des différences structurales notables. Ce résidu est de type Glu dans #25 et JR-FL, Ala dans #34 et un Asp dans Bx08. Une interaction ionique assistée par liaison hydrogène est formée entre le résidu Glu322 de la gp120 et la Lys26 du CCR5 dans le système JF-RL pendant plus de la moitié du temps

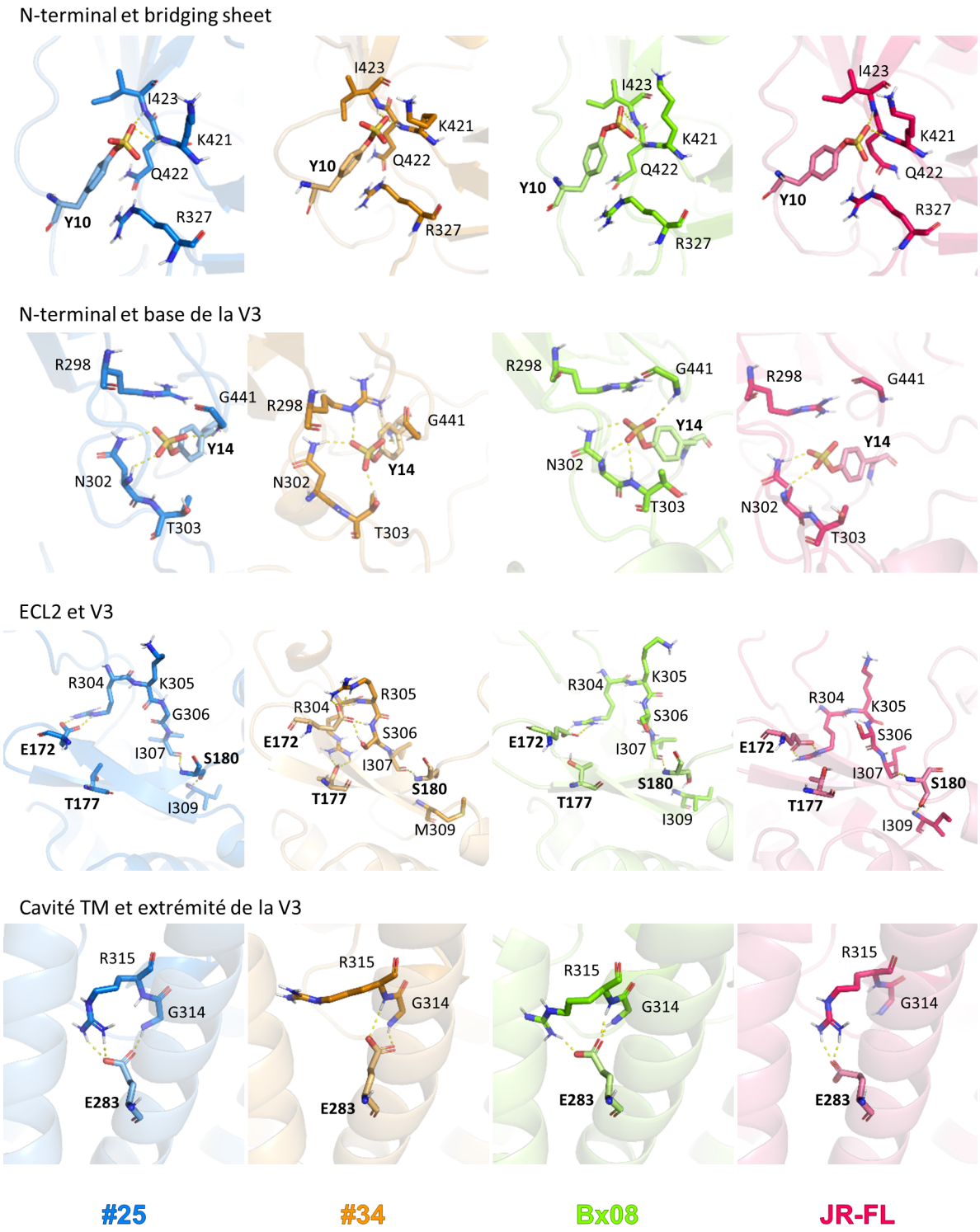

FIGURE 2.10 – Modes de liaison entre le CCR5 et la gp120. Les structures ont été sélectionnées à partir d'un partitionnement hiérarchique ascendant pour chaque système.

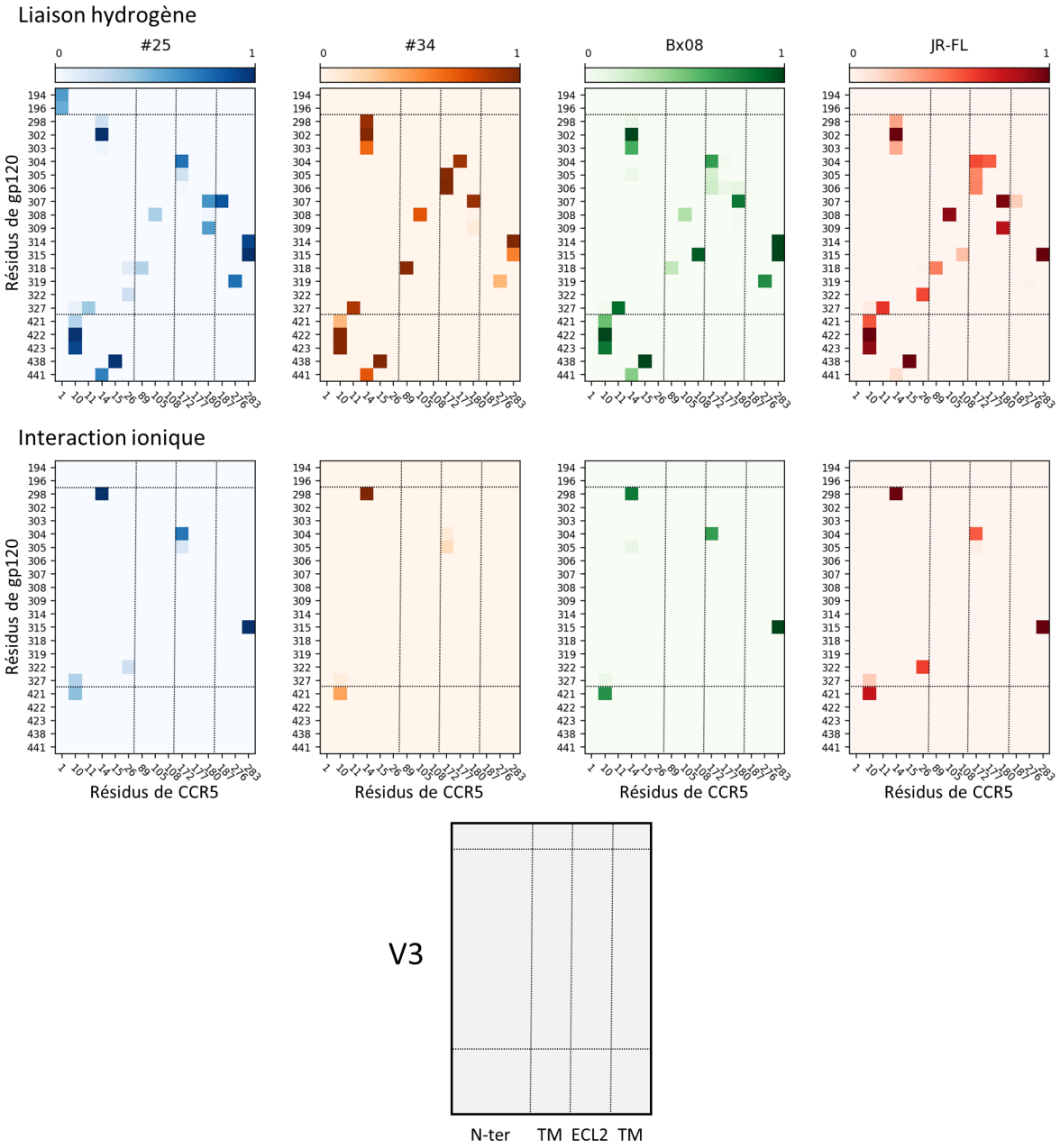

FIGURE 2.11 – Carte des fréquences d'interaction entre CCR5 et gp120. Les fréquences faibles sont représentées par des couleurs claires et les fréquences élevées par des couleurs foncées.

de simulation, et dans le système #25 pendant environ 20 % du temps de simulation. Dans ces deux systèmes, aucune interaction directionnelle n'est présente dans la moitié de l'interface qui est à proximité de la bicouche lipidique. Dans l'autre moitié de l'interface, aucune interaction directionnelle n'est formée entre le résidu CCR5 Met1 et le bridging sheet de la gp120, sauf pour le système #25 dans lequel des liaisons hydrogène engagent les atomes de la chaîne principale des résidus Ile194 et Cys196 de gp120 (voir ci-dessus).

Le deuxième bras impliqué dans la reconnaissance entre le CCR5 et la gp120 est la boucle V3 de la gp120, qui s'écarte du cœur de la protéine et s'étend pour atteindre le fond de la cavité du corécepteur tout en couvrant ECL2. Là encore, les résidus de gp120 exposés dans ces deux interfaces sont bien conservés dans #25, #34, Bx08 et JR-FL, avec 13 résidus identiques sur 17 et aucune variation modifiant radicalement la taille, la charge ou la polarité du résidu. Néanmoins, les simulations ont révélé des schémas de liaison distincts (FIGURE 2.11). Une liaison ionique a été observée dans chacune des deux interfaces dans les systèmes #25, Bx08 et JR-FL mais pas dans le système #34 malgré une conservation stricte des deux résidus chargés positivement correspondants (Arg304 lié à Glu172 de CCR5–ECL2 et Arg315 lié à Glu283 de CCR5–TM7, FIGURE 2.11). En outre, les motifs des liaisons hydrogène diffèrent dans les quatre systèmes. Dans JR-FL, la pointe de la V3 de gp120 est ancrée au fond de la cavité du corécepteur par une seule paire de résidus, qui sont les résidus Arg315 dans la gp120 et Glu283 dans le CCR5. Dans les systèmes #25, #34 et Bx08, jusqu'à trois paires supplémentaires de résidus sont engagées dans une liaison hydrogène. En fait, le mode de liaison de l'Arg315 de la gp120 au Glu283 de CCR5 est caractéristique du système : il implique soit uniquement la chaîne latérale de Arg315 (JR-FL), soit seulement sa chaîne principale (#34) ou les deux (#25 et Bx08) (FIGURE 2.11). Un examen plus approfondi des interactions de la V3 de la gp120 avec la ECL2 du CCR5 distingue la variante #34 des autres. Dans cette interface, le résidu Glu172 du CCR5 interagit fermement avec la chaîne latérale et la chaîne principale des résidus Arg305 et Ser306 de la gp120 dans le système #34 tandis qu'il forme uniquement une interaction ionique couplée à une liaison hydrogène avec le résidu Arg304 dans les systèmes #25, Bx08 et JR-FL. À Noter que le résidu en position 304 est une Arg dans les gp120 des systèmes #25, #34, JR-FL et Bx08. Tous les résidus impliqués dans les liaisons ioniques et les liaisons hydrogène mentionnées dans ce paragraphe, sont conservés, à l'exception du résidu à la position 305 de gp120 qui est une Arg

dans le système #34 et une Lys dans les trois autres systèmes. En résumé, les différences dans les modes de liaison ne sont pas liés aux mutations des résidus qui établissent des interactions clés avec le récepteur, mais à celles de résidus adjacents qui remodèlent localement la structure de la pointe de la V3 de la gp120. Ainsi, il est tentant d'émettre l'hypothèse que la nature des résidus aux positions 305, 306, 309 et/ou 317 dans la gp120 est susceptible de biaiser la reconnaissance de la gp120 pour des sous-populations de corécepteurs. La variante #25 présente une seule variation distinctive (Gly306 au lieu de Ser306) tandis que la variante #34 en présente trois (Arg305 au lieu de Lys, Met309 au lieu de Ile et Leu317 au lieu de Phe). La substitution Lys en Arg en position 305 est clairement liée au mode de liaison particulier de la V3 de la variante #34 de la gp120 à la ECL2 du CCR5.

2.3.6 Connexion entre des interfaces distantes

Pour déterminer l'influence de la variation de la séquence de la gp120 sur la direction du mouvement des domaines flexibles, nous avons calculé les corrélations croisées des fluctuations atomiques (FIGURE 2.12). Il est frappant de constater que le couplage de la V1 et de la V2, les boucles les plus flexibles de la gp120, avec les autres boucles variables de la gp120 est différent dans les quatre systèmes étudiés. Dans le système #25, la V1 et la V2 se déplacent de manière anti-corrélée avec la V3 et la V4. Dans JR-FL, on observe des anti-corrélations plus faibles, avec la V4 seulement. Dans Bx08, La V1 et la V2 ont un couplage modéré avec la V3 et la V5. Dans #35, il n'y a pas de couplage de la V1 et la V2 avec les autres boucles de variables de la gp120.

L'étude du mouvement de la gp120 par rapport au CCR5 distingue également les quatre systèmes (FIGURE 2.12). Dans les systèmes #34 et JR-FL, la corrélation des mouvements entre les domaines de la gp120 et du CCR5 est beaucoup plus forte à l'interface n°3 qu'à l'interface n°1. En revanche, dans les systèmes #25 et Bx08, la corrélation la plus forte est plus faible et concerne l'interface n°1 (Bx08) ou les interfaces n°1 et n°3 (#25). Les interfaces n°1 et n°3 étant reliées par l'interface n°2 (FIGURE 2.12), on peut supposer un couplage entre les éléments distants de la gp120 et de CCR5. Le mouvement de la V3 de gp120 a en effet été couplé, de faiblement à modérément, au mouvement du domaine N-terminal de CCR5 dans les quatre

systèmes. Le mouvement des résidus de CCR5 dans l'interface n°3 n'a cependant pas été corrélé avec le mouvement des résidus distants de la gp120, sauf pour la variante JR-FL. En bref, le mouvement des interfaces distantes ne semble pas fortement couplé.

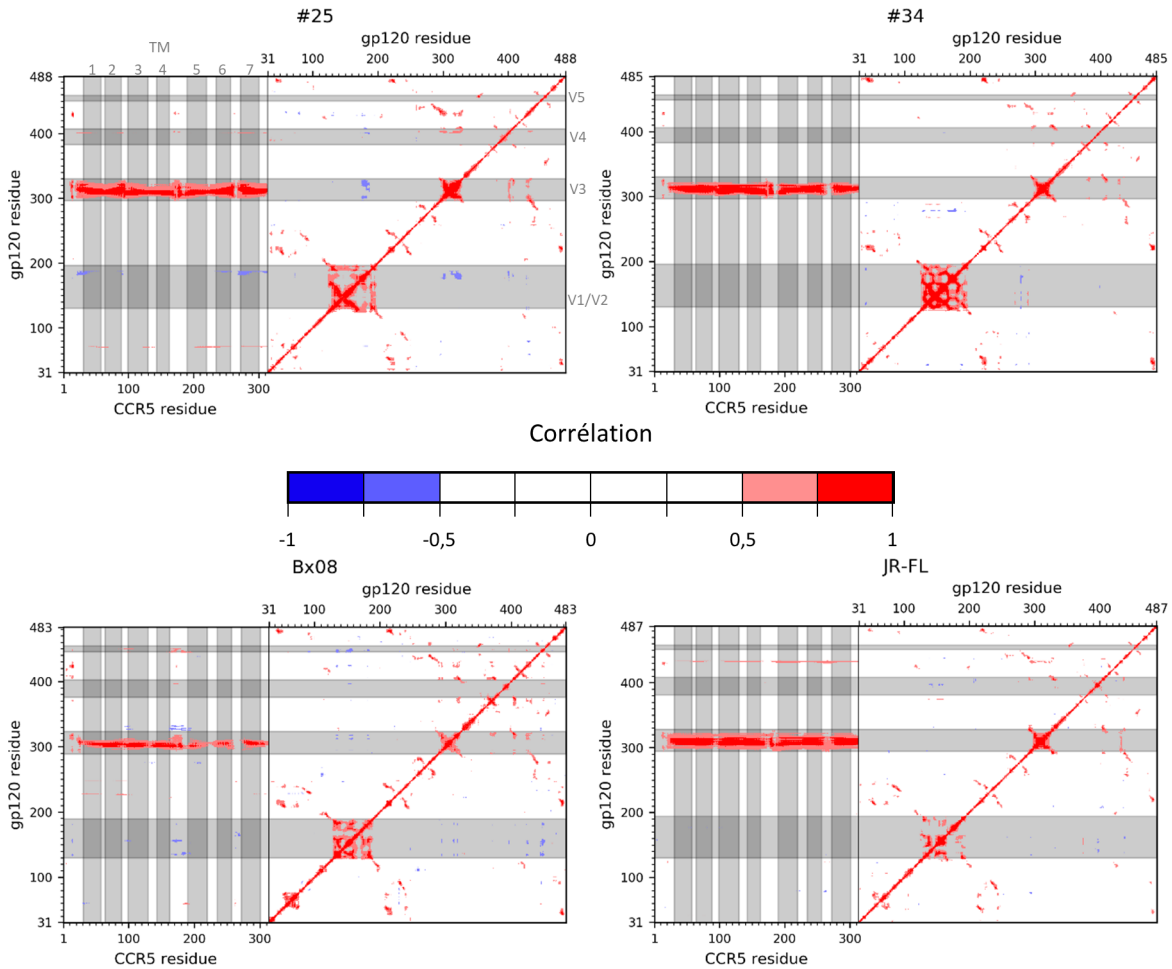

FIGURE 2.12 – Cartes des corrélations croisées des fluctuations atomiques des 4 variantes de la gp120 et du CCR5. Les structures ont été préalablement alignées sur les résidus de "cœur" de la structure minimisée de la gp120 ("gp120 core"). Les valeurs de corrélation des mouvements des résidus vont de 1 (couleur rouge), pour des mouvements corrélés, à -1 (couleur bleue) pour des mouvements anti-corrélés.

2.4 Discussion

2.4.1 Le domaine extracellulaire CCR5 est polymorphe mais subtilement

D'après les simulations de dynamique moléculaire, les quatre variantes de la gp120 ciblent différentes conformations de CCR5. Dans l'ensemble, les variations conformationnelles du co-récepteur sont beaucoup plus prononcées dans la partie extracellulaire que dans la région du 7TM. En particulier, l'ECL3 adopte des positions caractéristiques pour #25, #34, Bx08 et JR-FL, modifiant l'ouverture de la cavité transmembranaire (FIGURE 2.13).

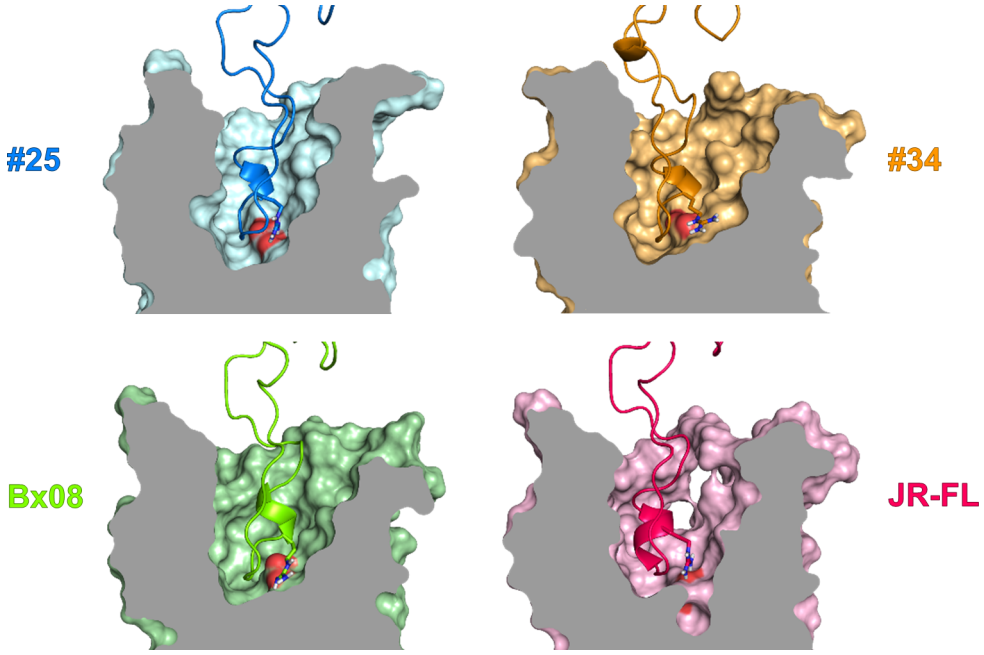

FIGURE 2.13 – Vue 3D en coupe de la cavité du CCR5, liée à la boucle V3 des 4 variantes de la gp120. CCR5 est représenté par des surfaces et les gp120 par des rubans (bleu pour #25, orange pour #34, vert pour Bx08 et rouge pour JR-FL). Le résidu Arg315 de la gp120 est représenté par des bâtonnets. La position des atomes d'oxygène de la chaîne latérale du résidu Glu283 du CCR5 est mise en évidence en rouge sur la surface du corécepteur.

La position de l'ECL2 est moins discriminante, et ne distingue pas la variante #25 du #34 par exemple. Il est important de noter que dans les quatre complexes simulés, l'ECL2 et l'ECL3 restent mobiles, avec un mouvement de pendule de grande amplitude, allant jusqu'à 6 Å (FIGURE 2.7). Par conséquent, une variante de la gp120 donnée ne reconnaît pas une conformation statique du CCR5, mais un continuum de conformations proches.

On peut alors se demander si la variabilité au sein de la sous-population du CCR5 reconnue par une variante de la gp120 est inférieure à la variabilité entre les sous-populations reconnues par les différentes variantes de la gp120. Pour répondre à cette question, nous avons d'abord aligné les structures des trajectoires de la dynamique moléculaire sur le 7TM rigide afin d'obtenir le meilleur ajustement des chaînes principales et de calculer le RMSD sur les atomes Cα. Comme on peut s'y attendre d'après le profil de fluctuation atomique (FIGURE 2.5), les proportions de structures similaires (seuil fixé à 2 Å) pour l'ECL3 au sein d'une sous-population (intra) reconnue par une même variante de la gp120 sont assez faibles (16,5 % à 34,1 %). Les proportions de structures similaires pour l'ECL3 entre les sous-populations (inter) reconnues par les différentes variantes sont encore plus faibles voir nulles. En revanche, les proportions de structures similaires inter et intra pour la boucle ECL2 de #25, #34 et Bx08 se situent toutes dans la même gamme (67,6 % à 92,8 %, FIGURE 2.7). La boucle ECL2 de JR-FL est également relativement bien définie avec une proportion de structures similaires inter égale à 71,7 %, mais les conformations adoptées s'écartent de celles observées pour #25, #34 et Bx08 (proportion de structures similaires $>$17,5 %). En résumé, bien que l'ECL3 soit le domaine le plus flexible du corécepteur, cette boucle tend à adopter des conformations qui sont spécifiques de la variante de la gp120.

Les quatre variantes de la gp120 ont-elles exploré des conformations du CCR5 très similaires ? La comparaison par paire des structures de dynamique moléculaire, en se concentrant sur les trois boucles extracellulaires prises dans leur ensemble, suggère un recouvrement faible. Environ 20 % des conformations sont communes entre #34/Bx08 et entre #25/#34 (FIGURE 2.14).

Toutes les autres paires partagent moins de 6 % des conformations similaires. À noter que la position de la région N-terminale du CCR5 fluctue plus largement que le 7TM insérée dans la bicouche lipidique dans toutes les simulations et pour les quatre systèmes #25, #34, Bx08 et JR-FL, ce qui accentue la diversité des conformations ciblées par les différentes gp120. En outre, aucune position commune de la région N-terminale de CCR5 n'est trouvée entre les sous-populations reconnues par les différentes variantes de la gp120 (FIGURE 2.14).

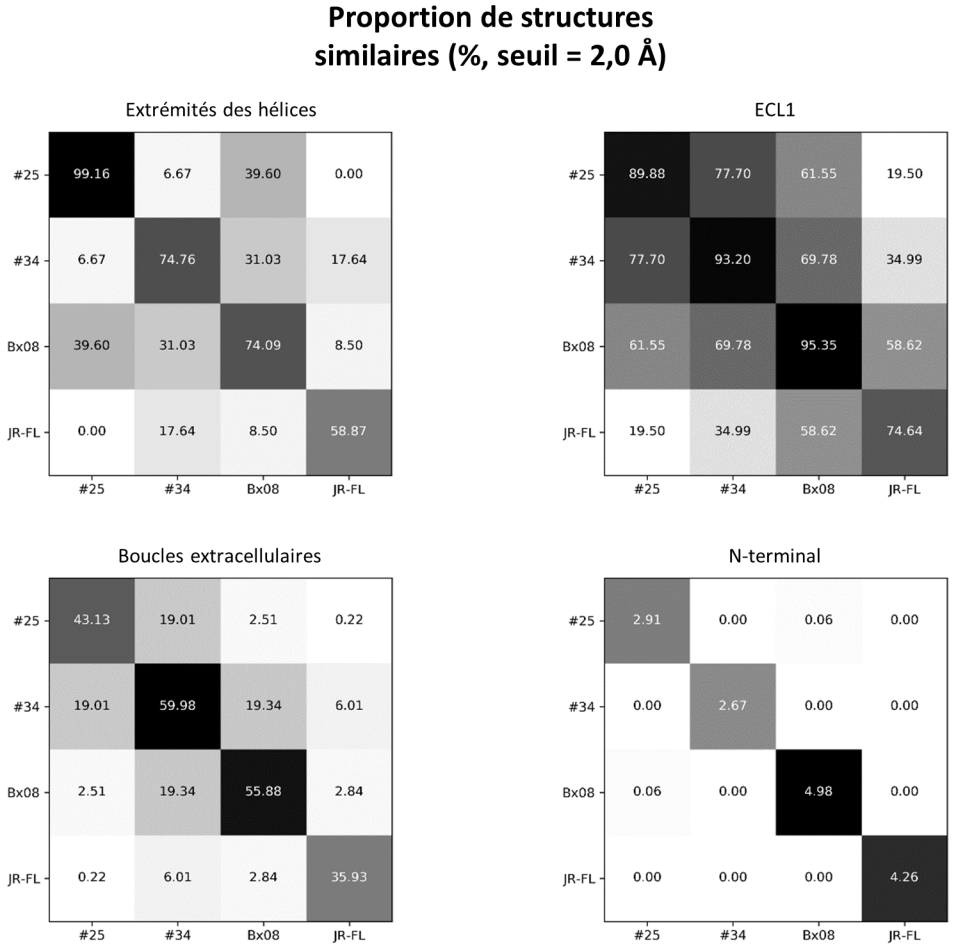

FIGURE 2.14 – Similarité des structures de CCR5 inter et intra variantes. Les extrémités des hélices comprennent seulement celles de la région extracellulaire.

2.4.2 Le mode de liaison #34 se distingue des autres variantes

Dans la structure cryo-EM de la gp120 liée au CCR5 [6], deux liaisons ioniques caractéristiques permet d'attacher l'extrémité de la V3 à l'ECL2 d'une part et à la région TM de CCR5 d'autre part (entre Arg304/Glu172 et entre Arg315/Glu283, respectivement). Comme mentionné ci-dessus, les deux interactions sont préservées lors des simulations de #25, Bx08 et JR-FL, mais sont manquantes dans les structures simulées de #34. La perte de liaisons ioniques dans la structure #34 est compensée par un gain de liaisons hydrogène (FIGURE 2.11). Quatre liaisons intermoléculaires persistantes contribuent à l'interaction stable des boucles ECL2 et V3, ce qui est une caractéristique distinctive de #34. D'autre part, ce mode de liaison particulier li-

bère la chaîne latérale de Arg315 de la pointe de la V3, et par conséquent pousse les résidus hydrophobes de la partie extracellulaire de TM6 vers la membrane.

2.5 Conclusion

Quatre simulations par dynamique moléculaire du complexe CCR5–gp120–CD4 correspondant aux quatre variantes #25, #34, Bx08 et JR-FL de la gp120 ont été produites. Elles décrivent des complexes stables et révèlent des différences dans les conformations adoptées par le CCR5 au cours du temps. Ces différences sont localisées à proximité de l'interface entre le CCR5 et la gp120, notamment au niveau des boucles ECL2, ECL3 et, en moindre mesure, dans les hélices de la cavité transmembranaire. Ainsi, on observe une conservation globale de la structure du CCR5 pour toutes les variantes de la gp120, et des variations locales qui permettent de discriminer ces variantes les unes des autres. Les conformations de la ECL3 distinguent les populations de récepteurs reconnues par les quatre gp120. La flexibilité du CCR5 permet d'accommoder les variations de séquence de la gp120 tout en maintenant le même mode de reconnaissance à deux sous-sites : le domaine N-terminal du CCR5 avec le bridging sheet de la gp120, et la cavité transmembranaire du CCR5 avec la V3 de la gp120. Les modes de liaison sont globalement similaires pour les quatre gp120, mais localement, les liaisons intermoléculaires varient, en particulier pour la V3. La variante #34 est remarquable, car elle privilégie les liaisons hydrogène au détriment des interactions ioniques, en particulier avec la ECL2. Ce mode de liaison spécifique peut être lié à la mutation ponctuelle du résidu 304 dans la pointe de la V3. Cette modification a un impact jusque dans la cavité transmembranaire où la chaîne latérale du résidu Glu283 du CCR5, connue depuis longtemps comme critique pour la liaison à la gp120, n'interagit plus avec la Arg315 de la V3 mais est exposé au solvant. À ce stade, il est intéressant de rappeler que la variante #34 est issue d'une souche présente lors du stade SIDA, et est donc très agressive. L'implication de mutations dans la V3 a déjà été pointée dans des souches virales résistantes au maraviroc, comme étant à l'origine de propriétés électrostatiques particulières modifiant sa liaison au CCR5 [22, 23, 24]. Les modes de liaison observés lors des simulations ont donné lieu à la proposition de mutants susceptibles de moduler sélectivement la liaison des différentes gp120. Ces mutants sont : T177A, Y187F, D276A et E283Q. Ils ont été construits et sont actuellement testés par Bernard LAGANE et son équipe.

Les gp120[#25] et gp120[#34], qui sont toutes les deux issues de virus du même patient, ont la même affinité pour le CCR5 ($K_D \approx 7\,\mathrm{nM}$), mais n'occupent pas les mêmes quantités de CCR5 ($B_{\max}(\#25) = 0.6 \pm 0{,}1\,\mathrm{pmol\,mg^{-1}}$, $B_{\max}(\#34) = 1.3 \pm 0{,}2\,\mathrm{pmol\,mg^{-1}}$) [5]. Le facteur 2 en faveur du #34 a été interprété par la liaison d'homodimère de CCR5 dans des stœchiométries différentes gp120 (1:2 pour #25 et 2:2 pour#34). Cette hypothèse a été confortée par l'étude de la liaison des gp120 à un mutant de CCR5 inhibant la formation de dimère de CCR5 (L196K). Nos modèles montrent des différences structurales dans la ECL2, la ECL3 et dans les parties extracellullaires des TM5, TM6 et TM7. La comparaison des structures du CCR5 en complexe avec les différentes gp120 avec les structures du CCR5 dans les modèles d'homodimères pourra fournir des éléments structuraux pour associer les populations du récepteur ciblées par les gp120 et leur capacité à former des dimères.

Dans ce chapitre, nous avons également montré que les populations du CCR5 liées aux différentes variantes de gp120 se distinguent par des positions moyennes et une direction de mouvement de l'extrémité intracellulaire du TM6. Ces données structurales laissent à penser que la gp120, qui est un antagoniste du CCR5 et donc ne modifie pas le niveau de signalisation basal du récepteur [25], peut néanmoins modifier subtilement l'engagement du CCR5 dans les voies de signalisation. Nous avons donc poursuivi la caractérisation de la partie intracellulaire du 7TM du CCR5 lié à différentes gp120, afin d'évaluer si celle-ci présentait des caractéristiques de l'activation du récepteur. Le prochain chapitre présente la méthode développée dans ce but, et l'applique à comparer les structures obtenues dans ce chapitre pour le CCR5 lié aux gp120 avec des structures de référence, de RCPG dans différents états d'activation et du CCR5 lié à d'autres ligands.

2.6 Références

[1] Jun Jin, Fanny Momboisse, Gaelle Boncompain, Florian Koensgen, Zhicheng Zhou, Nelia Cordeiro, Fernando Arenzana-Seisdedos, Franck Perez, Bernard Lagane, Esther Kellenberger, and Anne Brelot. CCR5 adopts three homodimeric conformations that control cell surface delivery. *Science Signaling*, 11(529), Mai 2018.

[2] Daniel Hilger, Matthieu Masureel, and Brian K. Kobilka. Structure and dynamics of GPCR signaling complexes. *Nature Structural & Molecular Biology*, 25(1) :4–12, Janvier 2018.

[3] Philippe Colin, Yann Bénureau, Isabelle Staropoli, Yongjin Wang, Nuria Gonzalez, Jose Alcami, Oliver Hartley, Anne Brelot, Fernando Arenzana-Seisdedos, and Bernard Lagane. HIV-1 exploits CCR5 conformational heterogeneity to escape inhibition by chemokines. *Proceedings of the National Academy of Sciences*, 110(23) :9475–9480, Juin 2013.

[4] James M. Fox, Richard Kasprowicz, Oliver Hartley, and Nathalie Signoret. CCR5 susceptibility to ligand-mediated down-modulation differs between human T lymphocytes and myeloid cells. *Journal of Leukocyte Biology*, 98(1) :59–71, 2015.

[5] Philippe Colin, Zhicheng Zhou, Isabelle Staropoli, Javier Garcia-Perez, Romain Gasser, Marie Armani-Tourret, Yann Benureau, Nuria Gonzalez, Jun Jin, Bridgette J. Connell, Stéphanie Raymond, Pierre Delobel, Jacques Izopet, Hugues Lortat-Jacob, Jose Alcami, Fernando Arenzana-Seisdedos, Anne Brelot, and Bernard Lagane. CCR5 structural plasticity shapes HIV-1 phenotypic properties. *PLOS Pathogens*, 14(12) :e1007432, Décembre 2018.

[6] Md Munan Shaik, Hanqin Peng, Jianming Lu, Sophia Rits-Volloch, Chen Xu, Maofu Liao, and Bing Chen. Structural basis of coreceptor recognition by HIV-1 envelope spike. *Nature*, 565(7739) :318–323, Janvier 2019.

[7] Bette Korber and Carla Kuiken. Sequence Quality Control. Technical report, Los Alamos National Lab. (LANL), Los Alamos, NM (United States), Décembre 1998.

[8] Brian Thomas Foley, Bette Tina Marie Korber, Thomas Kenneth Leitner, Cristian Apetrei, Beatrice Hahn, Ilene Mizrachi, James Mullins, Andrew Rambaut, and Steven Wolinsky. HIV Sequence Compendium 2018. Technical Report LA-UR-18-25673, Los Alamos National Lab. (LANL), Los Alamos, NM (United States), Juin 2018.

[9] Gáspár Pándy-Szekeres, Christian Munk, Tsonko M. Tsonkov, Stefan Mordalski, Kasper Harpsøe, Alexander S. Hauser, Andrzej J. Bojarski, and David E. Gloriam. GPCRdb in 2018 : adding GPCR structure models and ligands. *Nucleic Acids Research*, 46(D1) : D440–D446, Janvier 2018.

[10] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI : A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11) :1859–1865, 2008.

[11] Emilia L. Wu, Xi Cheng, Sunhwan Jo, Huan Rui, Kevin C. Song, Eder M. Dávila-Contreras, Yifei Qi, Jumin Lee, Viviana Monje-Galvan, Richard M. Venable, Jeffery B. Klauda, and Wonpil Im. CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *Journal of Computational Chemistry*, 35(27) :1997–2004, 2014.

[12] David A. Case and Peter A. Kollman. AMBER16. 2016.

[13] Jay W. Ponder and David A. Case. Force Fields for Protein Simulations. In *Advances in Protein Chemistry*, volume 66 of *Protein Simulations*, pages 27–85. Academic Press, Janvier 2003.

[14] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB : Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11 (8) :3696–3713, Août 2015.

[15] Andreas W. Götz, Mark J. Williamson, Dong Xu, Duncan Poole, Scott Le Grand, and Ross C. Walker. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation*, 8(5) :1542–1555, Mai 2012.

[16] Romelia Salomon-Ferrer, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation*, 9 (9) :3878–3888, Septembre 2013.

[17] Daniel R. Roe and Thomas E. Cheatham. PTRAJ and CPPTRAJ : Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, 9(7) :3084–3095, Juillet 2013.

[18] Gilles Marcou and Didier Rognan. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *Journal of Chemical Information and Modeling*, 47 (1) :195–207, Janvier 2007.

[19] Muhibur Rasheed, Radhakrishna Bettadapura, and Chandrajit Bajaj. Computational Refinement and Validation Protocol for Proteins with Large Variable Regions Applied to Model HIV Env Spike in CD4 and 17b Bound State. *Structure*, 23(6) :1138–1149, Juin 2015.

[20] Chih-chin Huang, Son N. Lam, Priyamvada Acharya, Min Tang, Shi-Hua Xiang, Syed Shahzad-ul Hussan, Robyn L. Stanfield, James Robinson, Joseph Sodroski, Ian A. Wilson, Richard Wyatt, Carole A. Bewley, and Peter D. Kwong. Structures of the CCR5 N Terminus and of a Tyrosine-Sulfated Antibody with HIV-1 gp120 and CD4. *Science*, 317(5846) : 1930–1934, Septembre 2007.

[21] Gáspár Pándy-Szekeres, Christian Munk, Tsonko M. Tsonkov, Stefan Mordalski, Kasper Harpsøe, Alexander S. Hauser, Andrzej J. Bojarski, and David E. Gloriam. GPCRdb in 2018 : adding GPCR structure models and ligands. *Nucleic Acids Research*, 46(D1) : D440–D446, Janvier 2018.

[22] Michael Roche, Martin R. Jakobsen, Jasminka Sterjovski, Anne Ellett, Filippo Posta, Benhur Lee, Becky Jubb, Mike Westby, Sharon R. Lewin, Paul A. Ramsland, Melissa J. Churchill, and Paul R. Gorry. HIV-1 Escape from the CCR5 Antagonist Maraviroc Associated with an Altered and Less-Efficient Mechanism of gp120-CCR5 Engagement That Attenuates Macrophage Tropism. *Journal of Virology*, 85(9) :4330–4342, Mai 2011.

[23] Reem Berro, Per Johan Klasse, John P. Moore, and Rogier W. Sanders. V3 determinants of HIV-1 escape from the CCR5 inhibitors Maraviroc and Vicriviroc. *Virology*, 427(2) : 158–165, Juin 2012.

[24] Javier Garcia-Perez, Isabelle Staropoli, Stéphane Azoulay, Jean-Thomas Heinrich, Almudena Cascajero, Philippe Colin, Hugues Lortat-Jacob, Fernando Arenzana-Seisdedos, Jose Alcami, Esther Kellenberger, and Bernard Lagane. A single-residue change in the HIV-1 V3 loop associated with maraviroc resistance impairs CCR5 binding affinity while increasing replicative capacity. *Retrovirology*, 12(1) :50, Juin 2015.

[25] Javier Garcia-Perez, Patricia Rueda, Isabelle Staropoli, Esther Kellenberger, Jose Alcami, Fernando Arenzana-Seisdedos, and Bernard Lagane. New Insights into the Mechanisms whereby Low Molecular Weight CCR5 Ligands Inhibit HIV-1 Infection. *Journal of Biological Chemistry*, 286(7) :4978–4990, Février 2011.

2.7 Annexes

2.7.1 Code source

Les codes sources des scripts utilisés pour l'étape de préparation des simulations sont présentés dans cette section. Certains scripts ne sont pas présent car ils sont trop long ou ont leur équivalent dans d'autres programmes comme CPPTRAJ.

shrinkbox.py

Code source

```
import sys
import os
import argparse
import pdb

import numpy as np

# Run with Python 2.7

# Pymol init
moddir='/softs/pymol/pymol1.8.6.1/lib64/python'
sys.path.insert(0, moddir)
os.environ['PYMOL_PATH'] = os.path.join(moddir, 'pymol/pymol_path')

import pymol
pymol.pymol_argv = ['pymol','-qc']
stdout = sys.stdout
stderr = sys.stderr
pymol.finish_launching()
sys.stdout = stdout
sys.stderr = stderr

class ObjectMatrix(object):
    def __init__(self, selection):
        self.selection = selection

    def get_coords(self):
        if not hasattr(self, '_coords'):
            coords = []
            model = pymol.cmd.get_model(self.selection)
            for atom in model.atom:
                coords.append(atom.coord)
            self._coords = np.array(coords)
```

```
        return self._coords

    def get_centroid(self):
        if not hasattr(self, '_centroid'):
            self._centroid = np.sum(self.get_coords(), axis=0) \
                             / len(self.get_coords()) \

        return self._centroid

    def get_box(self):
        if not hasattr(self, '_box'):
            self._box = np.array(
                [
                    np.min(self.get_coords(), axis=0),
                    np.max(self.get_coords(), axis=0)
                ]
            )

        return self._box

    def get_box_center(self):
        if not hasattr(self, '_box_center'):
            self._box_center = np.sum(self.get_box(), axis=0) / 2.0
        return self._box_center

    def get_box_length(self):
        if not hasattr(self, '_box_length'):
            min_box, max_box = self.get_box()
            self._box_length = max_box - min_box

        return self._box_length

class Renumbering(object):
    def __init__(self, system_groups):
        self.system_groups = system_groups
        self.current_atomid = 0
        self.current_resi = 0

    def get_next_atomid(self):
        if self.current_atomid >= 99999:
            return '*****'
        else:
            self.current_atomid += 1
            return self.current_atomid

    def default(self, lines):
        self.current_resi = 0
        previous_resn = None
        previous_resi = None
        for i, line in enumerate(lines):
            atomid = self.get_next_atomid()
            line_resn = line[17:22].split()[0]
            line_resi = int(line[22:28])
```

```
            if line_resn != previous_resn or line_resi != previous_resi:
                self.current_resi += 1

            resi = self.current_resi

            new_line = 'ATOM  '+'{:>5} '.format(atomid) \
                       +line[12:17]+'{:<4}'.format(line_resn) \
                       +' {:>5}'.format(resi)+line[27:]

            lines[i] = new_line

            previous_resn = line_resn
            previous_resi = line_resi

    def water(self):
        lines = self.system_groups['water']
        self.current_resi = 0
        previous_resn = None
        previous_resi = None
        new_lines = []

        # Sort lines according to coordinates
        lines.sort(
            key=lambda l: (float(l[30:38]), float(l[38:46]), float(l[46:54]))
        )

        # Split oxygen and hydrogens into separate lists
        oxygens = []
        hydrogens = []
        for i, line in enumerate(lines):
            atomname = line[12:17].strip()

            if atomname == 'OH2':
                oxygens.append(line)
            elif atomname in ('H1', 'H2'):
                hydrogens.append(line)

        # Retrieve corresponding oxygens and hydrogens
        for oxygen_line in oxygens:
            self.current_resi += 1
            resi = self.current_resi
            ox = float(oxygen_line[30:38])
            oy = float(oxygen_line[38:46])
            oz = float(oxygen_line[46:54])
            o_coord = np.array([ox,oy,oz])

            bound_hydrogen_indices = [None, None]
            for i, hydrogen_line in enumerate(hydrogens):
                hx = float(hydrogen_line[30:38])
                hy = float(hydrogen_line[38:46])
                hz = float(hydrogen_line[46:54])

                distance = np.linalg.norm(o_coord - np.array([hx,hy,hz]))
                if distance < 1.1:
```

```
                    hydrogen_atomname = hydrogen_line[12:17].strip()
                    if hydrogen_atomname == 'H1':
                        bound_hydrogen_indices[0] = i
                    elif hydrogen_atomname == 'H2':
                        bound_hydrogen_indices[1] = i

                if all([k is not None for k in bound_hydrogen_indices]):
                    if bound_hydrogen_indices[1] > bound_hydrogen_indices[0]:
                        h2_line = hydrogens.pop(bound_hydrogen_indices[1])
                        h1_line = hydrogens.pop(bound_hydrogen_indices[0])
                    else:
                        h1_line = hydrogens.pop(bound_hydrogen_indices[0])
                        h2_line = hydrogens.pop(bound_hydrogen_indices[1])

                    new_lines.extend([
                        'ATOM  '+'{:>5} '.format(self.get_next_atomid()) \
                                +oxygen_line[12:17]+'{:<4}'.format('TIP3') \
                                +' {:>5}'.format(resi)+oxygen_line[27:],

                        'ATOM  '+'{:>5} '.format(self.get_next_atomid()) \
                                +h1_line[12:17]+'{:<4}'.format('TIP3') \
                                +' {:>5}'.format(resi)+h1_line[27:],

                        'ATOM  '+'{:>5} '.format(self.get_next_atomid()) \
                                +h2_line[12:17]+'{:<4}'.format('TIP3') \
                                +' {:>5}'.format(resi)+h2_line[27:],
                    ])
                    break
            else:
                print 'Fuck. Missing hydrogens!'
                pdb.set_trace()

        self.system_groups['water'] = new_lines


AMINO_ACIDS = set([
    'ALA', 'ARG', 'ASN', 'ASP', 'CYS', 'GLN', 'GLU', 'GLY', 'HIS', 'ILE',
    'LEU', 'LYS', 'MET', 'PHE', 'PRO', 'PYL', 'SER', 'SEC', 'THR', 'TRP',
    'TYR', 'VAL',

    'TYS', 'CYX', 'HSD', 'HSE', 'HIE', 'HID'
])

LIPIDS = set([
    'POPE', 'POPC', 'CHL1'
])

WATERS = set([
    'TIP3'
])

IONS = set([
    'POT', 'CLA'
])
```

```
GROUPS = {
    'proteins': AMINO_ACIDS,
    'lipids': LIPIDS,
    'water': WATERS,
    'ions': IONS
}

def split_system_lines(pdb_lines):
    system_lines = {
        'proteins': [],
        'lipids': [],
        'water': [],
        'ions': []
    }
    for line in pdb_lines:
        for group_name in GROUPS:
            group_resnames = GROUPS[group_name]
            if line.startswith('ATOM'):
                atomid = int(line[6:12])
                resn = line[17:22].split()[0]
                resi = int(line[22:28])
                if resn in group_resnames:
                    system_lines[group_name].append(line)
                    break
    return system_lines

def main(args):
    # Check parameters
    if not os.path.isfile(args.input):
        print('Input file does not exist.')
        return 1

    output_dir = os.path.dirname(args.output)
    if output_dir and not os.path.isdir(output_dir):
        print('Output dir does not exist.')
        return 1

    print 'Shrinking box'

    # Init system
    water_thickness = args.thickness

    if args.new is not None:
        pymol.cmd.load(args.new, 'new_protein')
        pymol.cmd.load(args.input, 'inputs')
        pymol.cmd.remove('new_protein and (solvent or inorganic)')
        pymol.fitting.cealign(
            'inputs and polymer and name CA',
            'new_protein and polymer and name CA'
        )
        pymol.cmd.remove('inputs and polymer')
    else:
        pymol.cmd.load(args.input, 'inputs')
```

```
pymol.cmd.select('system', 'all')


pymol.cmd.select('proteins', 'polymer')
pymol.cmd.select('lipids', 'organic')
pymol.cmd.select('water', 'solvent')
pymol.cmd.select('ions', 'inorganic')

# Proceed
pymol.cmd.rotate('z', 45, 'system')
proteins_matrix = ObjectMatrix('proteins')

max_box_length = proteins_matrix.get_box_length() + water_thickness * 2

min_borders = proteins_matrix.get_box_center() - max_box_length[0] / 2.0
max_borders = proteins_matrix.get_box_center() + max_box_length[0] / 2.0
print 'New box length {}'.format(max_box_length[0])

pymol.cmd.select(
    'out_water',
    'water and not ((x>{} and y>{} and x<{} and y<{}) extend 2)'.format(
        min_borders[0],
        min_borders[1],
        max_borders[0],
        max_borders[1]
))

pymol.cmd.select(
    'out_lipids',
    'organic and not byres (x>{} and y>{} and x<{} and y<{})'.format(
        min_borders[0],
        min_borders[1],
        max_borders[0],
        max_borders[1]
))

pymol.cmd.select(
    'out_ions',
    'ions and not byres (x>{} and y>{} and x<{} and y<{})'.format(
        min_borders[0],
        min_borders[1],
        max_borders[0],
        max_borders[1]
))

pymol.cmd.remove('out_water')
pymol.cmd.remove('out_lipids')
pymol.cmd.remove('out_ions')

lipid_coords = {}
for atom in pymol.cmd.get_model('lipids').atom:
    try:
        lipid_coords[atom.resi].append(list(atom.coord))
    except KeyError:
```

```
            lipid_coords[atom.resi] = [list(atom.coord)]

    for resi in lipid_coords:
        res_coords = np.array(lipid_coords[resi])
        x, y, z = 1.0 / len(res_coords) * np.sum(res_coords, axis=0)
        if not (x > min_borders[0] and y > min_borders[1]
                and x < max_borders[0] and y < max_borders[1]):
            pymol.cmd.remove('organic and resi {}'.format(resi))

    # Rearrange
    system_matrix = ObjectMatrix('system')
    system_box_center = system_matrix.get_box_center()
    pymol.cmd.translate(
        [-system_box_center[0], -system_box_center[1], 0.0], 'system'
    )

    # Remove close waters and ions
    pymol.cmd.select(
        'contact_solvent',
        '(solvent near_to 1.0 of polymer) extend 2'
    )
    pymol.cmd.remove('contact_solvent')

    pymol.cmd.select('contact_ion', 'inorganic near_to 1.0 of polymer')
    pymol.cmd.remove('contact_ion')

    # Renumber residues
    print 'Renumbering'
    file_lines = pymol.cmd.get_pdbstr('system').split('\n')
    system_lines = split_system_lines(file_lines)
    renumber = Renumbering(system_lines)
    renumber.default(system_lines['proteins'])
    renumber.default(system_lines['lipids'])
    renumber.water()
    renumber.default(system_lines['ions'])


    with open(args.output, 'w') as f:
        for group_name in ('proteins', 'lipids', 'water', 'ions'):
            group_lines = system_lines[group_name]
            for line in group_lines:
                f.write(line+'\n')

if __name__ == '__main__':
    argparser = argparse.ArgumentParser()
    argparser.add_argument('--input', '-i', required=True,
        help='PDB file with proteins, solvent and others.'
        )

    argparser.add_argument('--new', '-n', default=None,
        help='Protein to replace.'
        )

    argparser.add_argument('--output', '-o', required=True,
```

```
        help='Output file with shrunk box'
        )

    argparser.add_argument('--thickness', '-t', required=True,
        help='Solvent shell at the top and the bottom (angstrom).'
        )

    argparser.set_defaults(func=main)

    args = argparser.parse_args()

    # Run
    status = args.func(args)
    sys.exit(status)
```

split_assembly.py

Code source

```
"""
Be careful! Charmm-gui does not respect PDB specifications on ATOM records.
"""

import sys
import os
import argparse
import pdb

# Run with Python 3

# Residue names considered
AMINO_ACIDS = set([
    'ALA', 'ARG', 'ASN', 'ASP', 'CYS', 'GLN', 'GLU', 'GLY', 'HIS', 'ILE',
    'LEU', 'LYS', 'MET', 'PHE', 'PRO', 'PYL', 'SER', 'SEC', 'THR', 'TRP',
    'TYR', 'VAL',

    'TYS', 'CYX', 'HSD', 'HSE', 'HIE', 'HID'
])

LIPIDS = set([
    'PA', 'PE', 'OL', 'PC', 'CHL'
])

WATERS = set([
    'WAT'
])

IONS = set([
    'K+', 'Cl-'
```

```
])

GROUPS = {
    'proteins': AMINO_ACIDS,
    'lipids': LIPIDS,
    'water': WATERS,
    'ions': IONS
}

MOD_RESNAMES = {
    'proteins': {
        'HSE': 'HIE',
        'HSD': 'HID'
    },

}

MOD_ATOMNAMES = {
    'proteins': {
        'OT1': 'O  ',
        'OT2': 'OXT'
    }
}


def main(args):
    def modify_text(group_name, mod_def, text):
        try:
            group_mod_def = mod_def[group_name]
            return group_mod_def[text]
        except KeyError:
            return text

    # Check parameters
    if not os.path.isfile(args.input):
        raise IOError('Input file does not exist.')

    if not os.path.isdir(args.output):
        raise IOError('Output directory does not exist.')

    group_lines = {k: [] for k in GROUPS.keys()}

    # Read assembly file
    previous_group = None
    with open(args.input) as f:
        for line in f:
            if line.startswith('ATOM'):
                atomname = line[12:16].strip()
                resname = line[17:21].strip()

                for group_name in GROUPS:
                    group_resnames = GROUPS[group_name]
                    if resname in group_resnames:
                        new_atomname = modify_text(
```

```
                            group_name,
                            MOD_ATOMNAMES,
                            atomname
                        )

                        new_resname = modify_text(
                            group_name,
                            MOD_RESNAMES,
                            resname
                        )

                        if atomname != new_atomname:
                            line = line.replace(atomname, new_atomname)
                        if resname != new_atomname:
                            line = line.replace(resname, new_resname)

                        group_lines[group_name].append(line)
                        previous_group = group_name
                        break
            elif line.startswith('TER'):
                group_lines[previous_group].append(line)

    # Write into separated files
    for group_name in group_lines:
        lines = group_lines[group_name]
        output_file = os.path.join(args.output, 'step5_'+group_name+'.pdb')
        with open(output_file, 'w') as f:
            for line in lines:
                f.write(line)

if __name__ == '__main__':
    argparser = argparse.ArgumentParser()
    argparser.add_argument('--input', '-i', required=True,
        help='PDB file given by charmm-gui.')
    argparser.add_argument('--output', '-o', required=True,
        help='The directory where output files will be written.'
        )

    argparser.set_defaults(func=main)

    args = argparser.parse_args()

    # Run
    status = args.func(args)
    sys.exit(status)
```

fix_SSbridge.py

Code source

```
"""
Detect disulfide bridges and rename residue according to Amber convention. Remove
hydrogens on sulfur atom if it's necessary
"""

# Run with Python 3

import sys
import os
import argparse
import itertools
import pdb

import numpy as np

def main(args):
    # Check parameters
    if not os.path.isfile(args.input):
        print('Input file does not exist.')
        return 1

    output_dir = os.path.dirname(args.output)
    if output_dir and not os.path.isdir(output_dir):
        print('Output dir does not exist.')
        return 1

    cystein_atoms = set([])
    file_lines = []
    is_ter = False

    # Read input file
    with open(args.input) as f:
        for line in f:
            if line.startswith('ATOM'):
                atomname = line[12:16].strip()
                resn = line[17:22].strip()
                resi = int(line[22:28])
                if resn == 'CYS' or resn == 'CYX':
                    atomid = int(line[6:11])
                    atomname = line[12:16].strip()
                    x = float(line[30:38])
                    y = float(line[38:46])
                    z = float(line[46:54])

                    cystein_atoms.add((atomid, atomname, resi, x, y, z))

                if atomname == 'OXT':
                    is_ter = True

                if is_ter and previous_resi != resi:
                    file_lines.append('TER\n')
                    is_ter = False

                previous_resi = resi
```

```
                file_lines.append(line)


    # Detect disulfure bridges
    sulfur_atoms = set([a for a in cystein_atoms if a[1] == 'SG'])
    sulfurbridge_resis = set([])
    for atom1, atom2 in itertools.combinations(sulfur_atoms, 2):
        pos1 = np.array(atom1[3:])
        pos2 = np.array(atom2[3:])
        distance = np.linalg.norm(pos1 - pos2)

        if distance < 2.5:
            sulfurbridge_resis.add(atom1[2])
            sulfurbridge_resis.add(atom2[2])
            print('bond <name>.{}.SG <name>.{}.SG'.format(atom1[2], atom2[2]))

    # Write new file
    with open(args.output, 'w') as f:
        for line in file_lines:
            if line == 'TER\n':
                f.write(line)
            else:
                resn = line[17:22].strip()
                resi = int(line[22:28])

                if resn != 'CYS':
                    f.write(line)
                else:
                    if resi in sulfurbridge_resis:
                        atomname = line[12:17].strip()
                        if atomname != 'HG':
                            line = line.replace('CYS', 'CYX')
                            f.write(line)
                    else:
                        f.write(line)


if __name__ == '__main__':
    argparser = argparse.ArgumentParser()
    argparser.add_argument('--input', '-i', required=True,
        help='PDB file of proteins.')
    argparser.add_argument('--output', '-o', required=True,
        help='Output file with modified cystein residues.'
        )

    argparser.set_defaults(func=main)

    args = argparser.parse_args()

    # Run
    status = args.func(args)
    sys.exit(status)
```

split_proteins.py

Code source

```
#!/opt/anaconda/envs/gpython3/bin/python

"""
Split proteins from one file into multiple files. C-ter must be tag OXT on
oxygen atom!
"""

import sys
import os
import argparse
import pdb

def main(args):
    # Check parameters
    if not os.path.isfile(args.input):
        raise IOError('Input file does not exist.')

    if not os.path.isdir(args.output):
        raise IOError('Output directory does not exist.')

    proteins = [[]]
    i = 0
    previous_resi = None
    is_ter = False

    # Read poteins file
    with open(args.input) as f:
        for line in f:
            if line.startswith('ATOM'):
                atomname = line[12:17].strip()
                resi = int(line[22:28].strip())

                if atomname == 'OXT':
                    is_ter = True

                if is_ter and previous_resi != resi:
                    proteins[i].append('TER\n')
                    is_ter = False
                    i += 1
                    proteins.append([])

                previous_resi = resi
                proteins[i].append(line)

    # Write into separated files
```

```
    for i, lines in enumerate(proteins):
        output_file = os.path.join(args.output, 'step5_prot'+str(i+1)+'.pdb')
        with open(output_file, 'w') as f:
            for line in lines:
                f.write(line)

if __name__ == '__main__':
    argparser = argparse.ArgumentParser()
    argparser.add_argument('--input', '-i', required=True,
        help='PDB file proteins.')
    argparser.add_argument('--output', '-o', required=True,
        help='The directory where output files will be write.'
        )

    argparser.set_defaults(func=main)

    args = argparser.parse_args()

    # Run
    status = args.func(args)
    sys.exit(status)
```

Chapitre 3

Projection du domaine transmembranaire du CCR5 et des RCPG de classe A

3.1 Introduction

Les RCPG jouent le rôle de régulation du fonctionnement des cellules par la transduction d'un signal à travers la membrane cellulaire [1, 2]. Ce phénomène est déclenché en réponse à la stimulation de la région extracellulaire par la fixation d'un ligand et le couplage d'un effecteur intracellulaire, qui généralement est une protéine G ou une β-arrestine. Les RCPG constituent une famille majeure de cibles thérapeutiques. Environ 30 % des médicaments approuvés par la FDA en 2017 ciblent des RCPG et environ 124 candidats médicaments développés pour un RCGP entre 2015 et 2019 sont en phase d'essais cliniques [3, 4]. Les RCPG sont caractérisées par sept segments flexibles adoptant une structure secondaire en hélice-α traversant la membrane lipidique et reliés entre eux par des boucles exposées à l'extérieur ou à l'intérieur de la cellule. Jusqu'à la fin des années 90, la disposition relative des hélices était inconnue et des modélisations de structure ont été proposées avec plus ou moins de précision. En 1997, une carte de densité électronique obtenue par cryo-EM à une résolution de 9 Å permis d'avoir un aperçu de l'organisation du 7TM [5]. En 2000, une première structure de haute résolution (2,8 Å), obtenue par diffraction aux rayons X, a permis une avancée majeure dans la compréhension des propriétés structurales à l'échelle atomique des RCPG [6]. Depuis, environ 370 structures ont été publiées avec une sur-représentation des RCPG de classe A, liées ou non à un effecteur intracellulaire [4].

Ces structures ont mis en lumière les différences structurales entre les récepteurs couplés et ceux non-couplés, définissant ainsi des états d'activation ou de pré-activation ainsi que des règles générales pour caractériser ces états. La signature principale de l'activation des RCPG de classe A est l'éloignement des extrémités intracellulaires des hélices 3 et 6 du centre du 7TM. Dans l'état inactif, les deux hélices sont maintenues par un verrou ionique (ionic lock en anglais), c'est-à-dire une liaison ionique efficace entre l'arginine en position 3.50 du motif conservé D(E)RY et l'acide glutamique en position 6.30 (numérotation de BALLESTEROS-WEINSTEIN [7]). Lors de la liaison d'un ligand dans la partie extracellulaire du RCPG, la réorganisation structurale perturbe le pont salin et les hélices 3 et 6 s'éloignent l'une de l'autre. Cet éloignement se traduit par l'ouverture de la cavité intracellulaire afin de permettre la liaison de

la protéine G. À noter que l'amplitude de la déviation sera dépendante du type de protéine G engagée (Gα-s, Gα-i/o) mais aussi de la nature des résidus dans le 7TM [8, 9, 10, 11, 12]. Un autre descripteur caractéristique, cette fois, de l'état actif concerne le motif NPxxY de l'hélice 7. Lors de l'activation, l'hélice 7 se déplace légèrement vers l'intérieur du récepteur afin d'établir un verrou aqueux (water lock en anglais), c'est-à-dire des liaisons hydrogène reliant la tyrosine 7.53 et la tyrosine 5.58 via une molécule d'eau. Bien que les mouvements décrits précédemment lors de l'activation s'appliquent à l'ensemble des RCPG de classe A, la variation des séquences au sein de la classe implique que les motifs d'interaction caractéristiques des états sont différents d'un récepteur à un autre. Par exemple, le verrou ionique est absent des récepteurs aux chimiokines de types C-C et C-X-C, et ceci est dû à la substitution du résidu Glu en position 6.30 par un résidu basique Arg/Lys. L'étude des interactions intramoléculaires au sein du 7TM des RCPG de classe A suggère qu'il n'y a pas de signature d'interaction caractéristique des états actif/inactif communs à tous les récepteurs de cette classe [13].

De ce fait, il peut être difficile de déterminer de manière absolue l'état d'activation d'un RCPG de classe A par l'analyse structurale d'un seul état conformationnel de ce récepteur. Lors de l'étude d'un récepteur par dynamique moléculaire par exemple, il est nécessaire de comparer les structures échantillonnées par la trajectoire avec des structures de référence dont les états d'activation sont connus. Nous proposons une méthode appelée ATOLL qui est basée sur la projection des extrémités des hélices transmembranaires afin de visualiser leur disposition et les comparer à des structures de référence afin d'en déduire leur état d'activation. Afin de vérifier son efficacité, la méthode a été testée sur une simulation de désactivation du récepteur β2-adrénergique (ADRB2) par dynamique moléculaire et sur 213 structures cristallographiques de RCPG de classe A. Puis la méthode a été appliquée aux simulations par dynamique moléculaire des quatre complexes CCR5–gp120–CD4 avec les variantes #25, #34, Bx08 et JR-FL de la gp120. Les structures simulées ont été comparées à des structures cristallographiques de CCR5 et de récepteurs aux chimiokines ainsi qu'à des structures simulées de CCR5 libre, lié à la CCL3 et lié au maraviroc.

3.2 Méthodes

3.2.1 Description générale d'ATOLL

ATOLL repose sur l'hypothèse que l'état d'activation d'un RCPG peut être caractérisé par la simple disposition des hélices transmembranaires. Le principe d'ATOLL est de projeter les positions des extrémités extracellulaire et intracellulaire des hélices transmembranaires sur le plan de la membrane (FIGURE 3.1). Préalablement, les structures comparées doivent être placées dans le même référentiel de coordonnées. Ainsi, deux graphiques sont générés par structure de récepteur, un pour la partie extracellulaire et un pour la partie intracellulaire, permettant de visualiser la disposition des hélices tout en révélant la variabilité de cette disposition (FIGURE 3.1).

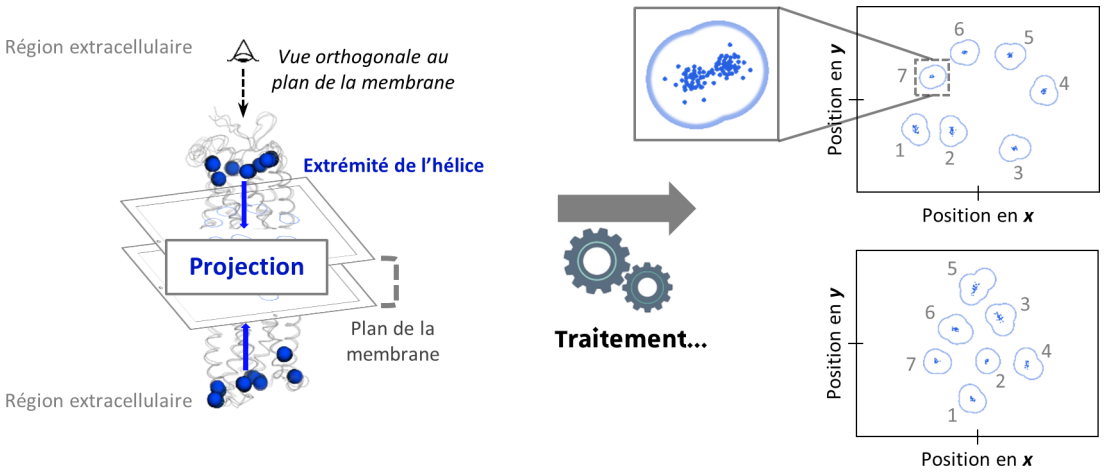

FIGURE 3.1 – Principe général d'ATOLL.

La procédure est entièrement codée avec le language Python version 3.7.3 et repose sur un large panel de bibliothèques tierces de haut niveau facilitant la programmation. Les modules utilisés sont énumérés dans le TABLEAU 3.1. Le programme est composé d'environ 2000 lignes de code et est toujours en développement.

Le programme est lancé via un interpréteur de commande sur une plateforme Windows ou Linux (MacOS n'a pas été testé). Il est possible de spécifier des options afin de définir les chemins d'accès des fichiers à traiter et aussi la manière dont le programme les traitera.

TABLEAU 3.1 – Liste des modules Python utilisés dans le programme ATOLL.

Nom	Version	Descriptif
NumPy	1.17.2	Manipulation des vecteurs et des matrices
SciPy	1.4.1	Calculs statistiques
MDAnalysis	0.20.1	Lecture/écriture de fichiers moléculaires de trajectoire de dynamique moléculaire. Alignement des structures. Manipulation des objets moléculaires.
BioPython	1.74	Manipulation et alignement de séquences
Matplotlib	3.1.1	Génération de graphique (nuage de points, courbes, histogramme, etc.)
YAML	5.3.1	Manipulation de fichier yaml. Sérialisation des données.

3.2.2 Fichiers d'entrées

Pour le bon déroulement du traitement, il sera nécessaire de fournir des fichiers obligatoires et dans certains cas des fichiers optionnels au programme ATOLL. Le caractère optionnel dépend du contexte. Les différents types de fichier pris en entrée par ATOLL sont :

- Fichiers obligatoires
  - un fichier de structure de référence
  - un ou plusieurs fichiers de structure à analyser (les entrées)
  - un fichier de définition des domaines de la structure
- Fichiers optionnels
  - un fichier d'alignement de séquence (obligatoire si les structures à analyser décrivent des protéines différentes)
  - un fichier d'annotation de chaque entrée

Fichiers de référence

La structure de référence joue un rôle crucial dans la procédure ATOLL. En effet, c'est sur cette dernière que sont superposées les structures à analyser. De plus, la définition des domaines est basée sur les résidus de la structure de référence. Par conséquent, la protéine doit être identique à celle des structures à analyser ou proche tout du moins. Il est également nécessaire que la structure soit placée dans un référentiel de coordonnées adapté aux projections. Actuellement, les positions des extrémités sont projetées sur le plan *xy*. Prochainement une routine sera ajoutée afin de placer la structure dans le référentiel adaptée en indiquant le plan de la membrane par des atomes. Il existe deux solutions afin d'avoir le référentiel adéquat. La première est que l'utilisateur place lui-même la protéine avec un logiciel tel que MOE ou Maestro. La deuxième possibilité est de télécharger la structure sur la base de données structurales des orientations des protéines dans la membrane (Orientation of proteins in membrane database, OPM, `opm.phar.umich.edu`) [14]. Cette base construite par l'Université du Michigan propose des structures de protéines membranaires processées issues de la PDB. Leur protocole consiste à déterminer les résidus insérés dans la membrane par un algorithme sophistiqué basé sur le calcul d'énergie de transfert ΔG_{transf} de chaque résidu d'un milieu aqueux à une bicouche lipidique [15]. Une fois les résidus identifiés, l'algorithme définit la borne supérieure (extracellulaire) et inférieure (intracellulaire) de la membrane. La protéine est ensuite centrée sur le centre géométrique de la membrane et la normale de la membrane est alignée sur l'axe *z*. C'est cette option qui est à privilégier. À noter que le programme ATOLL ne prend en compte que le premier conformère pour la structure de référence, les suivants étant ignorés. Les formats de fichier supportés sont le *.pdb* (conseillé) et le *.mol2*.

Fichiers de structure à analyser

Ces fichiers comportent toutes les structures qui seront analysées par le programme ATOLL. Ils peuvent décrire une ou plusieurs protéines, dont chacune peut être représentée par une ou plusieurs structures comme pour les simulations par dynamique moléculaire. Ces fichiers n'impliquent pas de dispositions particulières en terme de préparation. Il est possible d'utiliser

des structures issues de la PDB telles quelles, de même pour les trajectoires de dynamique moléculaire. Cependant, il est intéressant d'enlever tous les objets non-essentiels des fichiers comme les molécules d'eau qui vont augmenter le temps de lecture des trajectoires. ATOLL est capable de traiter à la fois des trajectoires et des structures statiques dans la même analyse. Pour les trajectoires de dynamique moléculaire, le fichier de topologie et le ou les fichiers de coordonnées de chaque entrée doivent être placés dans un répertoire dédié. Le programme effectue un scan du répertoire afin de retrouver les fichiers de topologie et de coordonnées selon l'extension de ces derniers. Le nom de fichier ou de répertoire définit le nom de l'entrée.

Les formats de fichier supportés sont le *.pdb* (conseillé) et le *.mol2* pour les structures statiques. Concernant les trajectoires, les formats supportés du fichier de topologie sont le *.prmtop* (Amber, conseillé), le *.parm7* (Amber) le *.psf* (CHARMM) et le *.pdb*. Pour les fichiers de coordonnées, les formats sont *.inpcrd* (Amber), *.rstrt* (Amber), *.nc* (Amber), *.ncdf* (Amber), *.dcd* (CHARMM). Il est possible de fournir des fichiers multi-pdb ou multi-mol2 bien que ce choix soit déconseillé pour des raisons d'encombrement de l'espace de disque.

Fichier d'alignement de séquences

Ce fichier est indispensable si les séquences des protéines décrites dans les structures sont différentes. Le programme n'incorpore pas de routine capable d'effectuer des alignements de séquences multiples. Par conséquent, il doit être fait par des logiciels tiers comme MOE capable d'intégrer l'information structurale lors de l'alignement ou bien Clustal Omega via le webservice (`www.ebi.ac.uk/Tools/msa/clustalo/`) [16]. Pour les RCPG, des alignements de séquences sont déjà proposés et accessibles en ligne, comme dans la GPCRdb [17].

Le seul format supporté par ATOLL est le format Stockholm (*.sto*, *.stk*) utilisé par exemple dans la base de donnée Pfam [18]. Ce format a l'avantage d'offrir la possibilité à l'utilisateur d'y insérer des annotations par le biais de fonctionnalités préexistantes ou personnalisées. Un exemple d'un fichier d'alignement est donné dans la FIGURE 3.2. Chaque séquence possède une étiquette composée le plus souvent du nom de la protéine ainsi que de l'intervalle

de séquence représenté. Dans le fichier, la séquence de référence est identifiée par la balise `"\#=GS label RE reference"` et sa numérotation sera utilisée afin de définir les domaines.

Fichier de définition des domaines

Durant la procédure d'alignement et de projection, le programme ATOLL se base sur la définition de deux domaines de la protéine. Le premier est le domaine d'alignement et le second le domaine de projection. Chaque domaine est composé d'un ou plusieurs groupes par exemple TM1, TM2, TM3, etc. Chaque groupe correspond à un ensemble de résidus défini par le premier et le dernier résidu de la séquence. La numérotation du résidu à utiliser est définie par l'utilisateur. La numérotation à utiliser est soit celle de la séquence de la protéine de référence ou bien celle correspondant aux positions des résidus dans le fichier d'alignement des séquences (FIGURE 3.2). Les données sont représentées dans le format YAML.

Fichier d'annotation des entrées

Le fichier d'annotation des entrées permet de consigner des informations pour les-dites entrées. Les données sont représentées sous forme d'un tableau dont les valeurs sont séparées par une tabulation. Pour l'heure, le tableau est composé de trois colonnes : le nom des entrées extrait du nom des fichiers ou des répertoires ; le nom de la protéine de l'entrée permettant de faire le lien entre l'entrée et la séquence ; la classe qui est définie par l'utilisateur, permettant une annotation personnalisée.

3.2.3 Description de la procédure

La procédure du programme ATOLL peut être décomposée en trois parties (FIGURE 3.3) :

— le pré-traitement qui consiste à charger les structures, les annotations et l'alignement des séquences dans le programme.

Fichier de séquences (*sequences.sto*)

```
# STOCKHOLM 1.0
#=GF SQ 5
#=GS ccr5_human/1-352 RE reference

Position             1   5    10   15   20   25   30   35   40   45   50   55   65   70
ccr5_human/19-352    PCQKINVKQIAARLLPPLYSLVFIFGFVGNMLVILILINCKRLKSMTDIYLLNLAISDLFFLLTVPFWAH
ccr2_human/34-374    KFD---VKQIGAQLLPPLYSLVFIFGFVGNMLVVLILINCKKLKCLTDIYLLNLAISDLLFLITLPLWAH
ccr6_human/36-374    CSL-QEVRQFSRLFVPIAYSLICVFGLLGNILVVITFAFYKKARSMTDVYLLNMAIADILFVLTLPFWAV
ccr7_human/52-378    DVR-----NFKAWFLPIMYSIICFVGLLGNGLVVLTYIYFKRLKTMTDTYLLNLAVADILFLLTLPFWAY
ccr9_human/40-369    KNN---VRQFASHFLPPLYWLVFIVGALGNSLVILVYWYCTRVKTMTDMFLLNLAIADLLFLVTLPFWAI
//
```

TM1 (*alignment*)

TM1 (*projection*)

Fichier de définition des domaines (*domains.yml*)

```
numbering: "position"      # Utilise la position des résidus afin de définir les domaines

alignment:                 # Nom du domaine
    TM1:                   # Nom du groupe
        start: 12          # Début du groupe
        end: 31            # Fin du groupe

projection:
    TM1:
        start: 9
        end: 39
```

Fichier de définition des entrées (*anno.tsv*)

```
Entry   Protein name   Class          # En-tête
4mbs    ccr5_human     inactive
6akx    ccr5_human     inactive
6aky    ccr5_human     inactive
5t1a    ccr2_human     inactive
6gps    ccr2_human     inactive
6gpx    ccr2_human     inactive
6wwz    ccr6_human     active
6qzh    ccr7_human     inactive
5lwe    ccr9_human     inactive
```

Nom des entrées

Classe définie par l'utilisateur

Nom des protéines dans le fichier d'alignement de séquence

FIGURE 3.2 – Formatage des fichiers d'alignement des séquences, de définition des domaines et d'annotation utilisés par le programme ATOLL.

— le traitement des structures qui sont alignées sur la référence. Les coordonnées des extrémités des hélices sont extraites à ce moment là.

— le post-traitement qui génère le graphique.

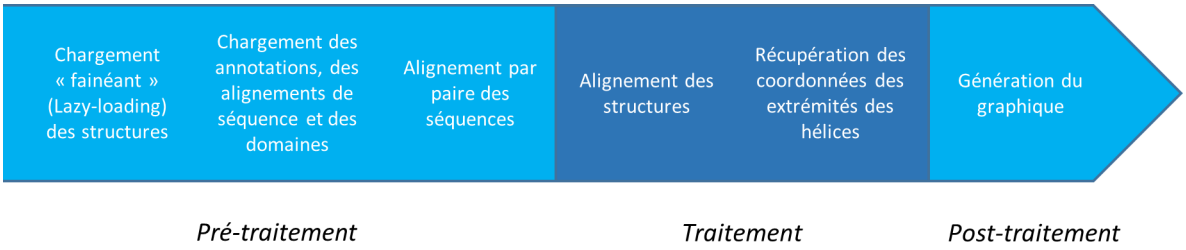

FIGURE 3.3 – Procédure simplifiée du programme ATOLL.

Pré-traitement

Les structures sont chargées de manière passive ou "fainéante" (lazy-loading en anglais) c'est-à-dire que les fichiers ne sont pas intégralement lus et chargés dans la mémoire à cette étape. Cela permet une faible consommation de la mémoire vive même pour des simulations conséquentes. Toutes les données d'annotations, des domaines et des séquences sont chargées et assignées aux entrées correspondantes. De plus, seuls les objets correspondant aux protéines sont sélectionnés. Les molécules d'eau, les molécules organiques et les ions sont ignorés.

Durant cette étape, le programme établi également les correspondances entre la numérotation des résidus dans la structure et celle des séquences dans le fichier d'alignement, et assigne par la suite les domaines pour toutes les entrées. Il est possible que dans une structure, la protéine ne soit pas complète, que la numérotation dans le fichier ne corresponde pas à celle dans la séquence Uniprot ou bien que des mutations soient présentes. Le programme effectue un alignement par paire entre la séquence de la structure et celle du fichier d'alignement afin d'établir le lien entre résidus (FIGURE 3.4).

L'algorithme d'alignement utilisé est celui de NEEDLEMAN–WUNSCH qui est un alignement dit global [19] adapté pour des séquences similaires. Pendant l'alignement, l'algorithme applique un score à chaque paire d'acides aminés. L'objectif de l'algorithme est de trouver un appariement optimal des résidus par maximisation du score. Si les deux résidus sont identiques alors l'algorithme assigne une valeur positive à cette paire. Par contre, si les résidus sont dif-

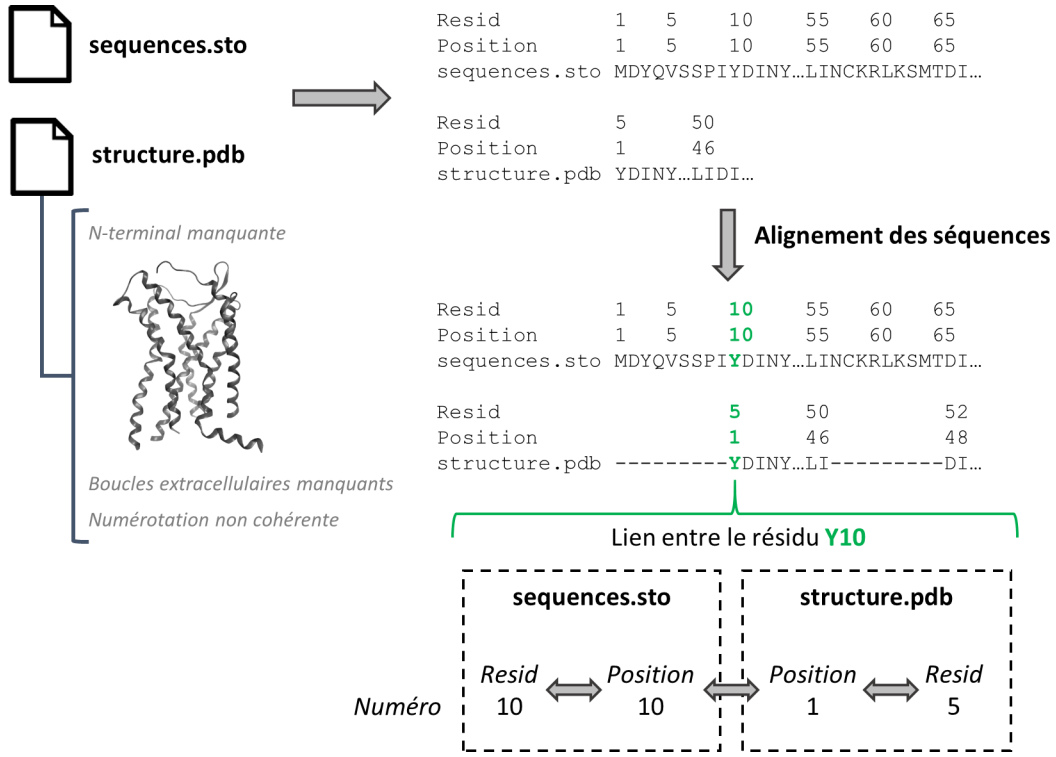

FIGURE 3.4 – Principe de l'alignement des séquences du fichier de la structure et du fichier d'alignement.

férents, l'algorithme applique une valeur de score en fonction de la nature des résidus. Par exemple, une paire arginine/phénylalanine sera fortement pénalisée par une valeur de score négative alors qu'une paire arginine/lysine donnera un score positif mais plus faible par rapport à une correspondance parfaite. Toutes les combinaisons d'acides aminés associés à leur valeur sont consignées dans la matrice BLOSUM62 [20] (TABLEAU 3.2).

L'algorithme est capable d'insérer des lacunes (gap en anglais) à un endroit dans les séquences afin que la correspondance des résidus par la suite soit meilleur. L'ajout d'un gap induit un score négatif (-2) et l'extension du gap est moins pénalisé (-1). Cela permet de favoriser le regroupement des gaps en un seul bloc.

Traitement

Le programme ATOLL traite les structures une par une. Par exemple pour une trajectoire de dynamique, ATOLL charge la première structure instantanée, effectue toutes les opérations,

TABLEAU 3.2 – Matrice de substitution BLOSUM62. Les résidus sont indiqués par leur notation à une lettre.

	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R		5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N			6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D				6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C					9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q						5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E							5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G								6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H									8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I										4	2	-3	1	0	-3	-2	-1	-3	-1	3
L											4	-2	2	0	-3	-2	-1	-2	-1	1
K												5	-1	-3	-1	0	-1	-3	-2	-2
M													5	0	-2	-1	-1	-1	-1	1
F														6	-4	-2	-2	1	3	-1
P															7	-1	-1	-4	-3	-2
S																4	1	-3	-2	-2
T																	5	-2	-2	0
W																		11	2	-3
Y																			7	-1
V																				4

puis passe à la structure instantanée suivante, etc. Cela permet de ne lire qu'une seule fois chaque structure, optimisant ainsi la vitesse d'exécution. Les opérations effectuées sont :

— Lecture des coordonnées atomiques

— Alignement sur la structure de référence

— Extraction des positions des extrémités des hélices

— Écriture des données sur le disque

Alignement des structures

Les structures à analyser sont alignées sur la structure de référence en considérant le atomes Cα des résidus définis par le domaine *alignment*. Le programme utilise la fonction

`align.rotation_matrix` du module `MDAnalysis` afin de calculer le RMSD et la matrice de rotation optimale des moindres carrés [21, 22].

Extraction des positions des extrémités des hélices

Le programme utilise la définition du domaine *projection* afin de sélectionner et d'extraire les coordonnées des atomes Cα des extrémités des hélices. Par défaut, les trois premiers résidus d'une extrémité sont sélectionnés et le centre géométrique des positions Cα est calculé (FIGURE 3.5).

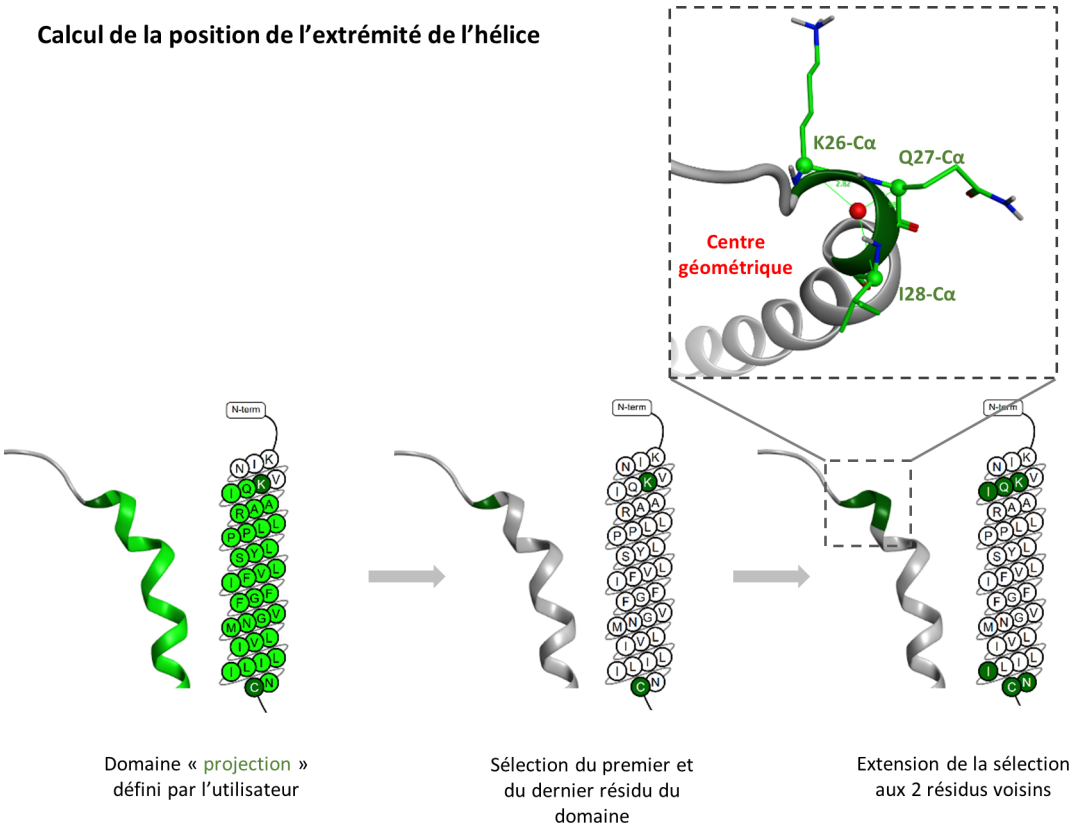

FIGURE 3.5 – Sélection des résidus et calcul de la position de l'extrémité de l'hélice.

Écriture des données sur le disque

Certaines données vues précédemment sont enregistrées sur le disque afin de garder une trace des résultats. Il est également possible pour l'utilisateur de les traiter lui-même par la suite

ou de vérifier qu'aucun problème ne soit survenu pendant les opérations. Les valeurs enregistrées sont le RMSD, la matrice de rotation et les positions des extrémités des hélices. De plus les structures alignées sont également sauvegardées (seulement les atomes Cα).

3.2.4 Post-traitement

Cette étape consiste à exploiter les données récupérées dans l'étape précédente et de générer le graphique des projections. Les valeurs des positions des hélices sont d'abord séparées en deux groupes : les positions extracellulaires et les positions intracellulaires (FIGURE 3.6).

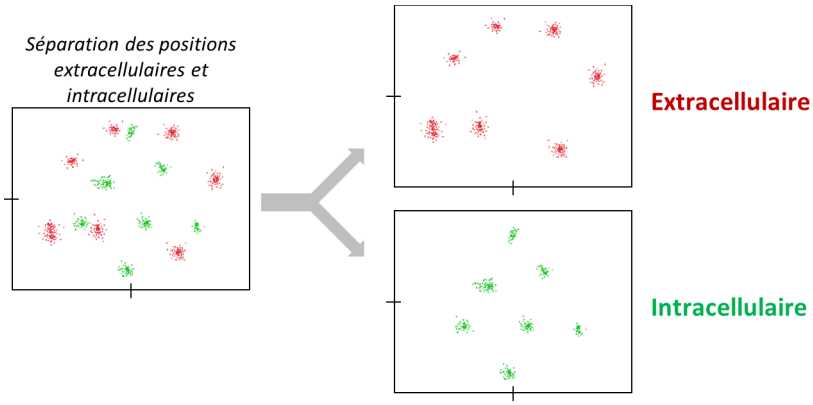

FIGURE 3.6 – Sélection des positions projetées extracellulaire et intracellulaire.

Ensuite, pour chacune des deux régions, les données sont regroupées et coloriées selon un schéma défini par l'utilisateur : par entrée, par protéine ou par classe (FIGURE 3.7).

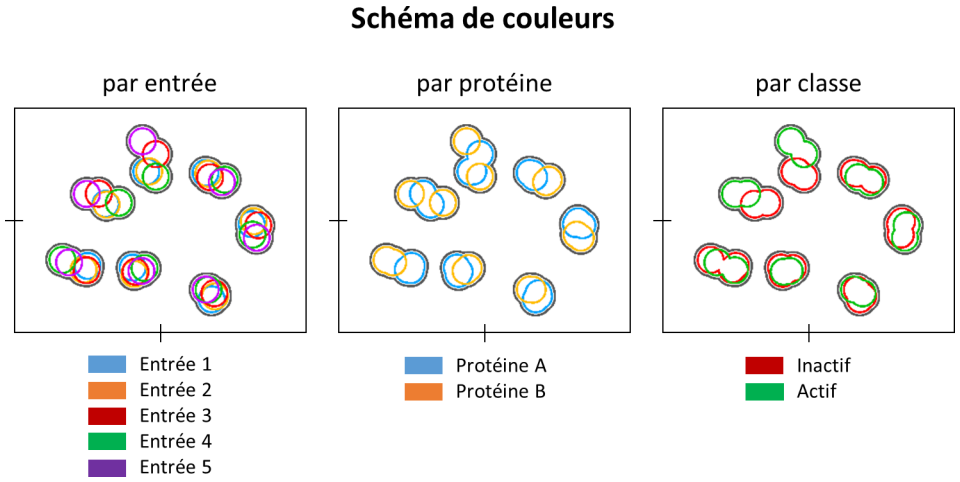

FIGURE 3.7 – Schéma de couleur utilisable dans le programme ATOLL.

Pour finir, les positions x et y de chaque hélice sont tracées sur le plan. La représentation des positions peut être choisie par l'utilisateur : un point ou un cercle du diamètre d'une hélice-α (FIGURE 3.8). Il est également possible d'appliquer un contour autour des hélices identiques (TM1, TM2, etc., FIGURE 3.7) afin de mieux les discerner les unes des autres. Un autre contour peut-être appliqué afin de visualiser les positions les plus peuplées ou bien l'entendu des positions quand ces dernières sont représentées sous forme de points en utilisant la fonction `spatial.ConvexHull` du module `SciPy`. Cela est pratique quand le nombre de structures est très grand, par exemple dans le cas des trajectoires de dynamique moléculaire.

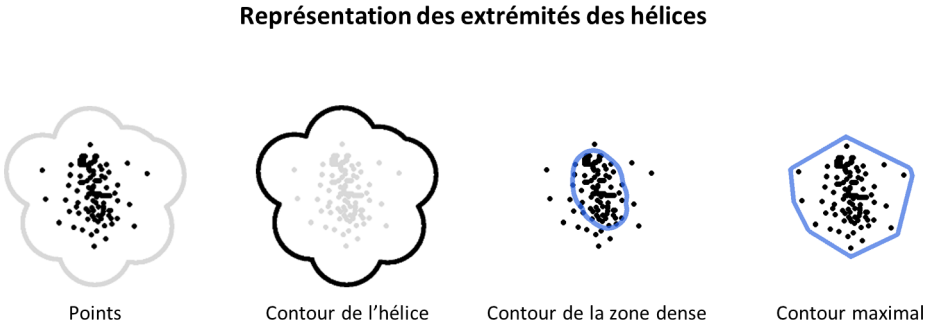

FIGURE 3.8 – Représentations des positions projetées dans le programme ATOLL.

Pour le contour en fonction de la densité, une estimation par noyau (Kernel Density Estimation en anglais) employant une loi normale est utilisée via la fonction `stats.gaussian_kde` de `SciPy`. À l'inverse, la représentation par des cercles est plus adaptée quand peu de structures sont à visualiser. Le programme génère un fichier image dans divers formats supportés par le module `Matplotlib` (exemples : *.png*, *pdf*, *.tif*, *.eps*, *.svg*, etc.). Le format *.jpg* n'est pas supporté. De toute manière, il constituerait un très mauvais choix pour encoder un graphique.

3.3 Matériel

3.3.1 Dynamique moléculaire de la désactivation du récepteur β2-adrenergique

Préparation du système de dynamique moléculaire

Les résultats de la simulation du ADRB2 ont été publiés en 2011 par Ron DROR et ses collaborateurs [23]. La dynamique d'une durée d'environ 11 µs, a été générée par le superordinateur ANTON caractérisé par ses circuits intégrés spécialisés reliés entre eux par un réseau torique afin d'accélérer la production des dynamiques [24]. La trajectoire sélectionnée dans ce chapitre correspond au réplica numéro 11 dans la publication [23]. La structure initiale correspond à la structure cristallographique du ADRB2 dans l'état actif lié à l'agoniste BI-167107 dans le domaine extracellulaire et à l'anticorps Nb80 dans le domaine intracellulaire (entrée PDB : 3P0G [25]). L'anticorps Nb80 a été préalablement retiré de la structure avant simulation. La protéine de fusion a été retirée laissant l'extrémité du TM5 et celle du TM6 libres en l'absence de la boucle ICL3.

Du fait du grand nombre de structures générées lors de la trajectoire (55556 structures), seules les sections décrivant la transition d'un état d'activation à un autre ont été sélectionnées : la section allant de 0 à 720 ns (actif à intermédiaire) et la section allant de 5580 à 6300 ns (intermédiaire à inactif).

Les portions de trajectoires sélectionnées ont été alignées avec ATOLL sur une structure de référence décrivant le ADRB2 à l'état actif, lié à l'agoniste BI-167107 et à l'anticorps à une résolution de 2,79 Å [26]. La structure orientée dans le référentiel adapté à l'analyse par ATOLL a été téléchargée depuis la base de données OPM (entrée : 4LDE) [14].

Pour permettre l'alignement des structures avec ATOLL, la structure de référence a été préparée comme suit. Les éléments non-nécessaires à la projection, à savoir l'anticorps, le ligand,

les lipides, les molécules d'eau et les ions, ont été supprimés. La protéine de fusion remplaçant la partie N-terminale du ADRB2 a été retirée, laissant un récepteur tronqué jusqu'au résidu His22. La structure a été corrigée avec l'outil *Structure Preparation* du logiciel MOE 2019.01. Elle a ensuite été éditée avec MOE 2019.01 pour coïncider avec la séquence native de la protéine (entrée Uniprot : P07550). Les positions mutées sont : D23W, V24D, T25A, Q26Y, E27A, R28A, M96T, M98T, N187E, C265A. La ICL3 (Lys232 à Lys263) n'a pas été reconstruite.

Projection des extrémités des TM du ADRB2

Les résidus sélectionnés pour l'alignement avec ATOLL des structures instantanées du ADRB2 sur la structure de référence du récepteur sont les résidus insérés dans la bicouche lipidique et qui fluctuent peu lors de la simulation par dynamique moléculaire. Dans la numérotation de la séquence native du ADRB2, il s'agit de, TM1 : 36 à 54 ; TM2 : 74 à 90 ; TM3 : 109 à 126 ; TM4 : 155 à 169 ; TM5 : 200 à 217 ; TM6 : 278 à 292 ; TM7 : 312 à 325 (FIGURE 3.9).

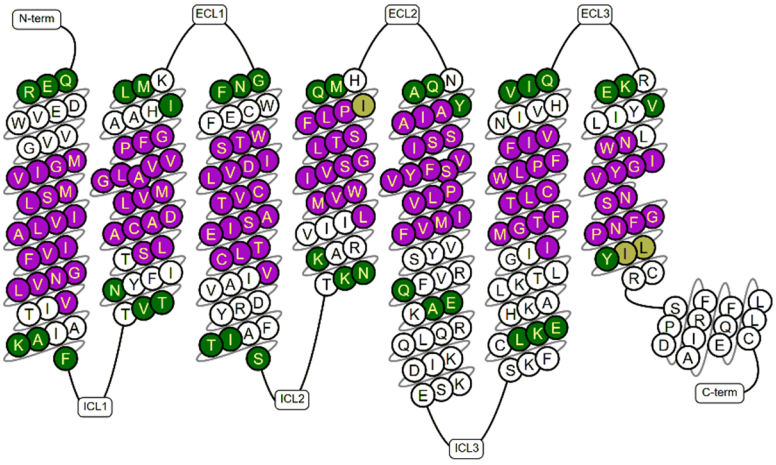

FIGURE 3.9 – Résidus sélectionnés lors de l'alignement (violet) et de la projection des hélices (vert foncé) du ADRB2. Les résidus en jaune foncé correspondent à des résidus impliqués à la fois dans l'alignement et la projection.

Les résidus sélectionnés pour la projection avec ATOLL correspondent à la définition des TM du ADRB2 dans la base de données GPCRdb. Nous avons vérifié que ces résidus ont une conformation en hélice-α dans la structure de référence du ADRB2, en déterminant les structures secondaires avec MOE 2019.01. Chaque extrémité de TM est définie par un ensemble de trois résidus consécutifs. Elle sera représentée par le centroïde des Cα de ces trois résidus. Les résidus

projetés sont énumérés dans le TABLEAU 3.3. Leur position dans la séquence du récepteur est montrée dans la FIGURE 3.9.

TABLEAU 3.3 – Résidus sélectionnés lors la projection des hélices du ADRB2. La numérotation correspond à celle de la séquence du ADRB2. Les deux bornes des intervalles sont inclusives.

Résidu	Extracellulaire	Intracellulaire
TM1	26–28	59–61
TM2	94–96	67–69
TM3	102–104	135–137
TM4	147–149	169–171
TM5	197–199	224–226
TM6	266–268	297–299
TM7	305–307	324–326

3.3.2 États d'activation de structures 3D de RCPG de classe A

Le travail de sélection, annotation, collection et préparation des structures a été préalablement réalisé par Florian KOENSGEN en thèse au laboratoire [27]. En ce qui concerne la préparation, les structures PDB ont été complétées pour ajouter les atomes d'hydrogène manquants. Les éléments non-constitutifs du récepteur comme les molécules d'eau, les lipides, les ligands, les ions, les protéines de fusion et les effecteurs intracellulaires (protéines G, anticorps, etc.) ont été retirés. Un point important de la préparation est l'identification des TM et la renumérotation des acides aminés avec le système proposé par BALLESTEROS et WEINSTEIN [7]. La définition des TM basée sur la numérotation de BALLESTEROS-WEINSTEIN est montrée sur la structure du récepteur de la rhodopsine bovine (entrée PDB : 1F88, TABLEAU 3.4, FIGURE 3.10).

Pour ATOLL, il est nécessaire de préparer un alignement multiple des séquences des RCPG dont les structures sont comparées. Cet alignement a été préparé en utilisant les modèles de structure des TM de 366 RCPG non olfactifs humains disponibles au laboratoire [28, 29]. Les modèles, au format *.mol2*, ont ainsi servi à extraire les séquences en numération BALLESTEROS-WEINSTEIN, sauvegardées au format Stockholm en utilisant le script `mol2seq.py` (en annexe

TABLEAU 3.4 – Définition des TM des RCPG de classe A. La numérotation est basée sur celle proposée par BALLESTEROS et WEINSTEIN [7]. Les deux bornes des intervalles sont inclusives.

Hélice	Résidus
TM1	1.30–1.59
TM2	2.38–2.66
TM3	3.22–3.54
TM4	4.40–4.62
TM5	5.36–5.60
TM6	6.30–6.55
TM7	7.32–7.55

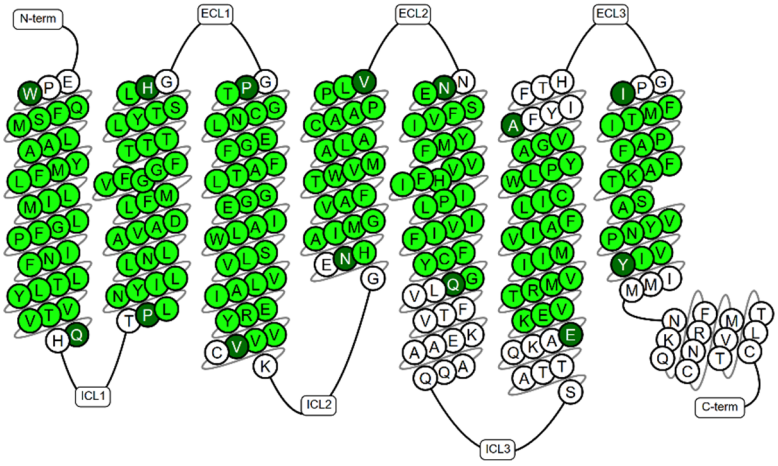

FIGURE 3.10 – Définition des TM des RCPG de class A basée sur le structure de la rhodopsine bovine (entrée PDB : 1F88).

de ce chapitre). Les résidus sélectionnés pour une première projection avec ATOLL sont ceux qui définissent les bornes des TM, comme définies précédemment (TABLEAU 3.4). Cependant, certaines structures ne sont pas résolues dans ces régions et les bornes ont été révisées pour que toutes les structures puissent être traitées. Les résidus finalement projetés sont énumérés dans le TABLEAU 3.5 et représentés sur la séquence de la rhodopsine bovine sur la FIGURE 3.11. Les résidus sélectionnés pour l'alignement des structures sont les suivants, TM1 : 1.36 à 1.53 ; TM2 : 2.44 à 2.60 ; TM3 : 3.28 à 3.48 ; TM4 : 4.46 à 4.56 ; TM5 : 5.42 à 5.54 ; TM6 : 6.37 à 6.49 ; TM7 : 7.38 à 7.62 (Numérotation de BALLESTEROS-WEINSTEIN, FIGURE 3.11).

Le récepteur utilisé comme référence est la rhodopsine bovine. La structure a été téléchargée depuis la base de donnée OPM (entrée : 1F88). Les éléments non-nécessaires à savoir le ligand, les lipides, les molécules d'eau, les ions et la représentation de la membrane ont été supprimés. La structure décrit une protéine native avec un gap au niveau de l'ECL3 qui n'a pas été corrigé.

TABLEAU 3.5 – Résidus sélectionnés lors la projection des hélices des RCPG de classe A. La numérotation est basée sur celle proposée par BALLESTEROS et WEINSTEIN [7]. Les deux bornes des intervalles sont inclusives.

Résidu	Extracellulaire	Intracellulaire
TM1	1.33–1.35	1.54–1.56
TM2	2.61–2.63	2.41–2.43
TM3	3.25–3.27	3.49–3.51
TM4	4.57–4.59	4.43–4.45
TM5	5.39–5.41	5.55–5.57
TM6	6.50–6.52	6.34–6.36
TM7	7.35–7.37	7.53–7.55

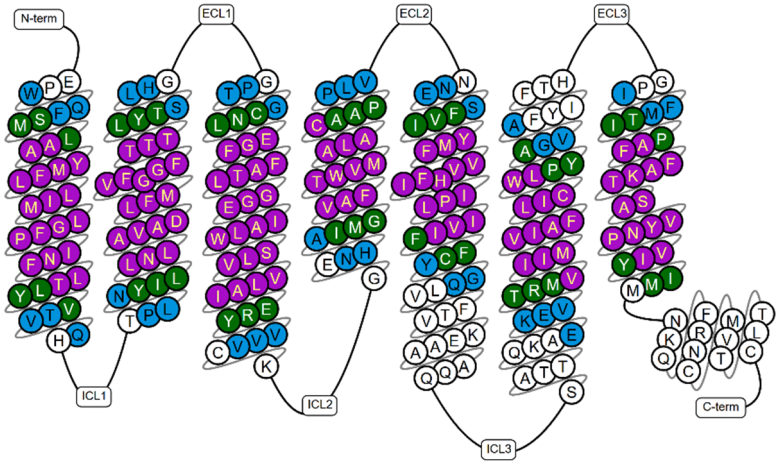

FIGURE 3.11 – Résidus sélectionnés lors de l'alignement (violet) et de la projection des hélices (vert foncé) des RCPG de classe A. Les résidus en bleu correspondent aux résidus définissant les TM décrit précédemment.

3.3.3 Dynamique moléculaire du complexe CCR5–gp120–CD4

Les trajectoires de dynamique moléculaire du complexe CCR5–gp120–CD4 analysées ont été décrites dans le chapitre précédent. Les trajectoires de dynamiques moléculaires de CCR5 libre, lié à CCL3 et lié au MVC ont été préparées par Florian KOENSGEN selon un protocole identique à celui décrit dans le chapitre précédent [27]. D'une durée de 5×60 ns, les dynamiques ont été allongées d'une durée de 5×40 ns, afin d'obtenir une durée totale de 500 ns pour chaque système.

Les structures cristallographiques des récepteurs aux chimiokines (CCR2, CCR5, CCR6, CCR7 et CCR9) ont été téléchargées sur le site de la PDB. Toutes les structures ont été préparées avec le même protocole. Tout d'abord, les séquences issues des structures ont été alignées sur celles recherchées dans la base de donnée Uniprot pour les protéines natives (Code d'accession : CCR2 : P41597 ; CCR5 : P51681 ; CCR6 : P51684 ; CCR7 : P32248 ; CCR9 : P51686) avec le logiciel MOE 2018.01 (outil *Protein Align/Superpose* avec les options par défaut). Les structures ont été éditées manuellement pour corriger les mutations et retirer les protéines de fusion au niveau des boucles intracellulaires. Les boucles manquantes dans les récepteurs n'ont pas été modélisées.

L'alignement des séquences complètes des récepteurs aux chimiokines a été téléchargé sur le site de la GPCRdb au format Fasta. Le fichier a été converti au format Stockholm avec le logiciel Jalview version 2.11.1.3. Les résidus utilisés pour l'alignement avec ATOLL correspondent aux positions en violet et jaune foncé de la FIGURE 3.12. Deux niveaux de projection ont été choisis qui se distinguent par leur proximité par rapport au centre de la membrane. Le premier est celui qui correspond aux résidus les plus éloignés du centre de la membrane (TABLEAU 3.6, couche 1). Le deuxième correspond aux bornes définies pour l'alignement et donc à des résidus insérés dans la bicouche (TABLEAU 3.6, couche 2).

TABLEAU 3.6 – Résidus sélectionnés lors de la projection des hélices de CCR5 et des récepteurs aux chimiokines. La numérotation des résidus appliquée correspond à celle de CCR5 et celle de la GPCRdb (BALLESTEROS-WEINSTEIN en exposant). Les deux bornes des intervalles sont inclusives.

	Couche 1		**Couche 2**	
	Extracellulaire	Intracellulaire	Extracellulaire	Intracellulaire
TM1	$26^{1.28}$–$28^{1.30}$	$55^{1.57}$–$57^{1.59}$	$31^{1.33}$–$33^{1.35}$	$55^{1.57}$–$57^{1.59}$
TM2	$87^{2.61}$–$89^{2.63}$	$64^{2.38}$–$66^{2.40}$	$86^{2.60}$–$88^{2.62}$	$64^{2.38}$–$66^{2.40}$
TM3	$98^{3.22}$–$100^{3.24}$	$129^{3.53}$–$131^{3.55}$	$99^{3.23}$–$101^{3.25}$	$127^{3.51}$–$129^{3.53}$
TM4	$163^{4.61}$–$165^{4.63}$	$142^{4.39}$–$144^{4.41}$	$162^{4.60}$–$164^{4.62}$	$143^{4.40}$–$145^{4.42}$
TM5	$187^{5.32}$–$189^{5.34}$	$221^{5.65}$–$223^{5.67}$	$190^{5.35}$–$192^{5.37}$	$217^{5.61}$–$219^{5.63}$
TM6	$257^{6.57}$–$259^{6.59}$	$228^{6.28}$–$230^{6.30}$	$254^{6.54}$–$256^{6.56}$	$235^{6.35}$–$237^{6.37}$
TM7	$269^{7.24}$–$271^{7.26}$	$298^{7.54}$–$300^{7.56}$	$277^{7.32}$–$279^{7.34}$	$298^{7.54}$–$300^{7.56}$

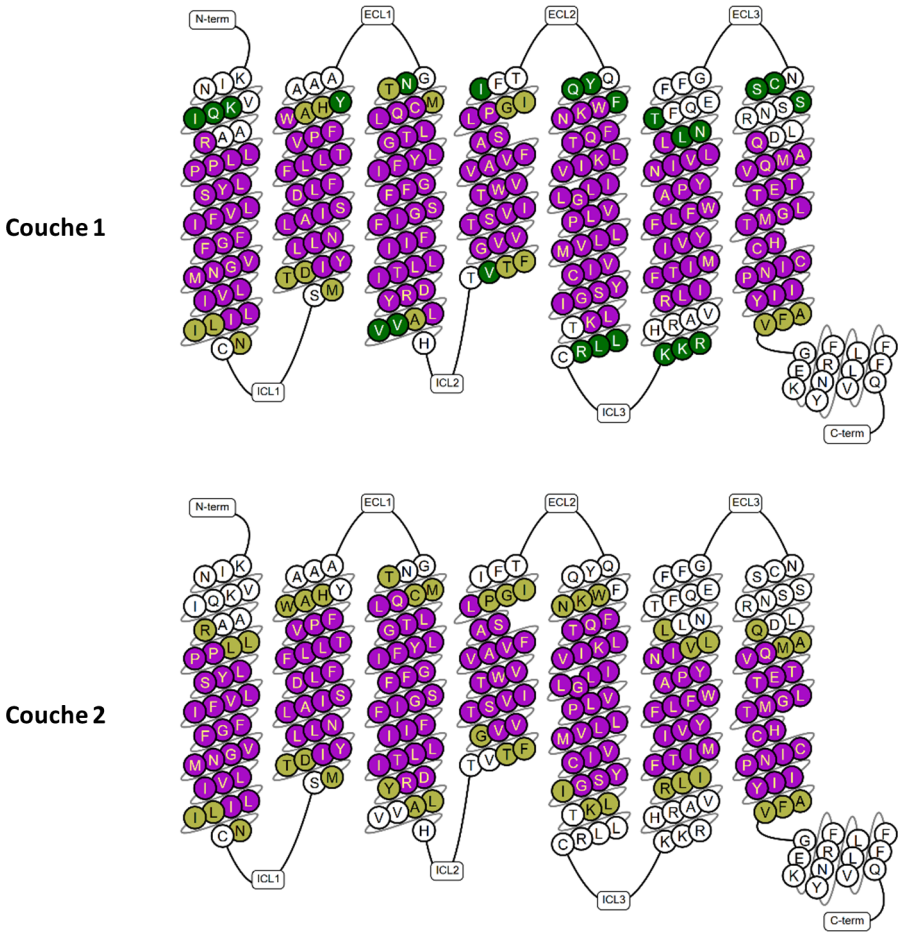

FIGURE 3.12 – Résidus sélectionnés lors de l'alignement (violet) et de la projection des hélices (vert foncé) du CCR5. Les résidus en jaune foncé correspondent à des résidus impliqués à la fois dans l'alignement et la projection.

3.4 Résultats et discussion

3.4.1 Dynamique moléculaire de la désactivation du récepteur β2-adrenergique

La trajectoire de dynamique moléculaire du récepteur ADRB2 publiée par Ron DROR et ses collaborateurs [23] simule la transition entre les états inactif et actif. Ces deux états ont été caractérisés dans les structures cristallographiques du ADRB2 en complexe avec l'agoniste inverse partiel carazolol (PDB : 2RH1) et avec l'agoniste BI-167107 et un nanocorps mimétique de la protéine G (entrée PDB : 3P0G). Les auteurs ont choisi un jeu de descripteurs structuraux afin de diviser la simulation en trois parties, dont chacune correspond à un état différent du récepteur : actif, intermédiaire et inactif. Le premier descripteur est la distance entre les atomes Cα des résidus 3.50 et 6.34 (numérotation BALLESTEROS-WEINSTEIN) qui caractérise la transition entre l'état intermédiaire et inactif (FIGURE 3.13A). Le second descripteur est la déviation des coordonnées atomiques (RMSD) calculée sur motif NPxxY du TM7 par rapport à la structure inactive, qui distingue l'état actif de l'état intermédiaire (FIGURE 3.13A). En 2018, Florian KOENSGEN a développé au laboratoire une méthode utilisant les motifs d'interaction intramoléculaires dans le RCPG afin de classer automatiquement et sans *a priori* les différents états structuraux de ADRB2. Il a ainsi montré que l'état intermédiaire est en fait composé d'au moins trois populations homogènes qui ne sont pas révélées par le jeu de descripteurs utilisés par Ron DROR et ses collaborateurs. (FIGURE 3.13B).

Les trois parties de cette même trajectoire du ADRB2, comprenant l'état actif (350 ns), les états intermédiaires (800 ns) et l'état inactif (200 ns) ont été traitées par ATOLL. Les structures tridimensionnelles du récepteur ont été alignées sur une structure de référence à l'état actif (base de données OPM : 4LDE) en considérant uniquement les résidus des parties des hélices incluses dans la bicouche lipidique pour la simulation par dynamique moléculaire. Ainsi, on obtient une valeur moyenne de RMSD égale à $0,79 + -0{,}680$ Å pour les structures actives, à $1,88 + -0{,}248$ Å pour les structures intermédiaires et à $2,07 + -0{,}130$ Å pour les structures inactives.

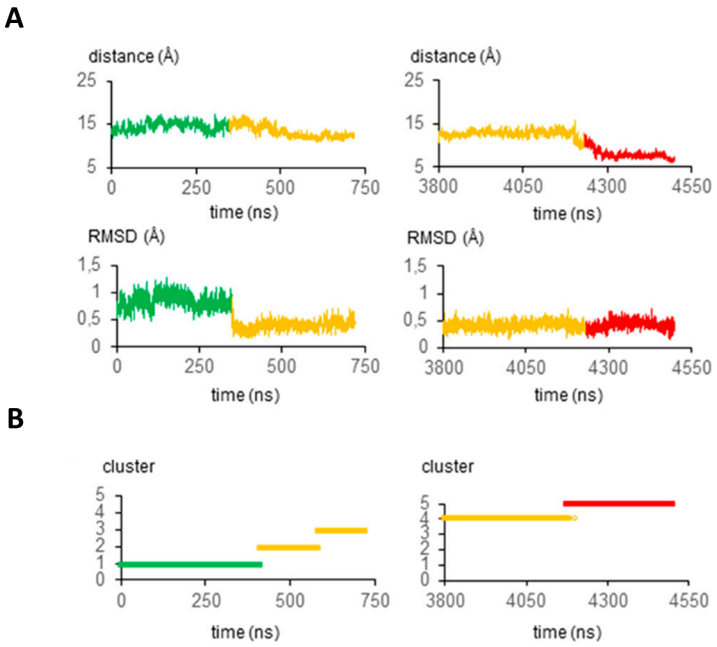

FIGURE 3.13 – Distinction des états conformationnels du récepteur β2-adrénergique humain simulé par dynamique moléculaire. (A) Classement en états actif (vert), inactif (rouge) et intermédiaire (jaune) basé sur la distance entre les Cα des résidus Arg131$^{3.50}$ et Leu272$^{6.34}$ et l'écart quadratique moyen par rapport aux coordonnées d'une structure de référence à l'état inactif (entrée PDB : 2RH1) dans le motif NP$^{7.50}$xxY [23]. (B) Regroupement des structures échantillonnées par dynamique moléculaire par ressemblance des motifs d'interaction intramoléculaires construits à partir de la structure du 7TM et des boucles [13]. Figure issue de la référence [13].

La première chose qui a été vérifiée est si les mouvements liés à l'activation du récepteur et captés par les descripteurs structuraux sont biens identifiés par la méthode ATOLL (FIGURE 3.14). L'observation de la projection de l'extrémité intracellulaire du TM7 montre la déviation vers l'extérieur lors de la transition de l'état actif à intermédiaire, et témoigne ainsi du ré-arrangement du motif NPxxY . De plus, les états actif et inactif sont complètement distincts, avec aucun recouvrement de leur projection du TM7. Ce descripteur peut donc constituer un choix judicieux afin de discerner ces deux états. Comme montré dans la FIGURE 3.14, les positions du TM7 dans l'état intermédiaire recouvrent celles de l'état inactif. Par ailleurs, l'hélice 7 est d'avantage fluctuante dans l'état intermédiaire suggérant un rôle des hélices 1 et/ou 6 voisines dans la stabilisation de l'état inactif. Concernant le marqueur structural principal de l'activation, à savoir TM6, on remarque un mouvement important de l'extrémité intracellulaire de cette hélice vers l'intérieur comme indiqué par le descripteur structural calculé par Ron DROR et ses collaborateurs. De plus, la projection générée par ATOLL confirme que le récepteur adopte trois

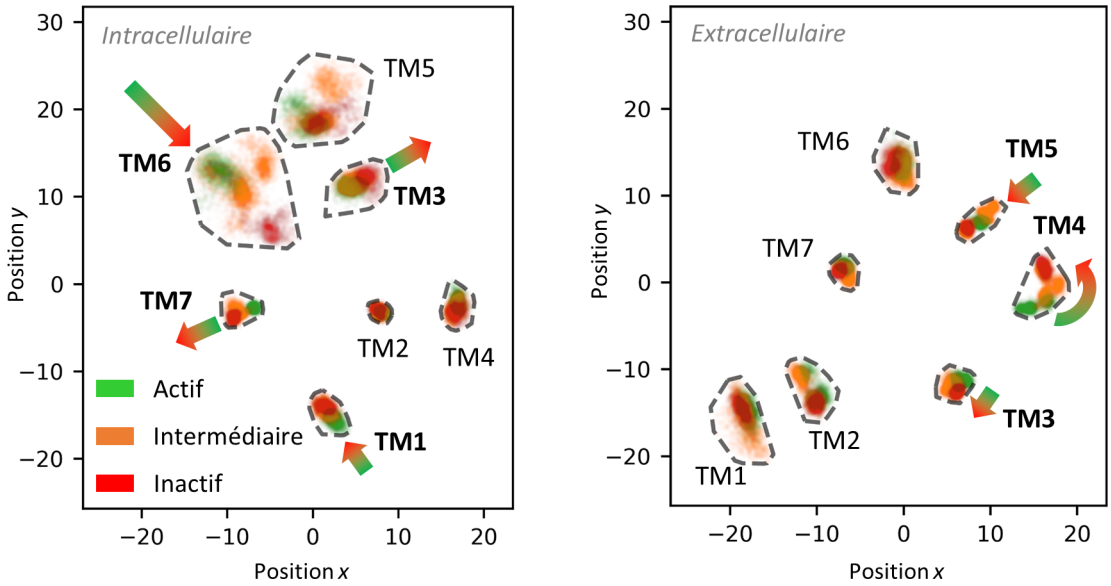

FIGURE 3.14 – Densité des positions des extrémités des hélices intracellulaire (à gauche) et extracellulaire (à droite).

conformations différentes dans l'état intermédiaire, comme suggéré par Florian Koensgen dans son analyse des interactions intramoléculaires. Pour une de trois conformations, on observe un recouvrement des positions projetées du TM6 avec celles de l'état actif, ce qui pose les limites de la méthode visuelle si elle ne considère qu'un seul point de projection pour l'extrémité du TM.

La vue globale apportée par ATOLL permet également de visualiser le mouvement des autres hélices dans la partie intracellulaire. Par exemple, le TM5 ne semble pas adopter de positions bien définies. On remarque que les positions des états actif et inactif sont proches mais toutes deux très différentes des conformations intermédiaires. Cela suggère que le TM5 passe d'une position lors de la transition actif/intermédiaire pour revenir à une position proche de celle initiale lors de la transition intermédiaire/inactif. Une autre chose intéressante est que la position du TM2 intracellulaire ne semble pas être influencée par les hélices voisines, et reste à la même position pendant toute la phase de désactivation.

Dans la partie extracellulaire du 7TM, on peut clairement observer l'influence de l'absence d'un ligand sur la forme de la cavité engendrant des mouvements des hélices pendant la désactivation. Parmi les mouvements les plus notables, on peut citer celui de l'extrémité extracellulaire du TM4 qui décrit un mouvement en arc de cercle pendant la désactivation. D'autant plus que

Ron DROR et ses collaborateurs ne font pas mention de ce phénomène qui n'est pas non plus observé par comparaison des structures cristallographiques du récepteur à l'état actif et à l'état inactif.

3.4.2 États d'activation de structures 3D de RCPG de classe A

Pour évaluer le domaine d'application d'ATOLL, nous avons aussi comparé des structures expérimentales résolues par cristallographie aux rayons X. Cette analyse a également pour but de mieux comprendre la variabilité de positionnement des hélices TM pour différents RCPG dans un même état. Nous nous sommes ainsi penchés sur la caractérisation des états d'activation des RCPG de classe A disponibles dans la PDB en 2018. L'analyse porte sur 53 RCPG décrits par 213 structures tridimensionnelles. Les structures ont été divisées en trois classes correspondant aux trois états d'activation. Une structure est considérée comme active si elle est liée à un partenaire intracellulaire (protéine G, nanocorps synthétique, anticorps ou β-arrestine). Une structure décrivant un récepteur non couplé mais en présence d'un agoniste dans le site de liaison orthostérique est considérée comme intermédiaire. Pour finir, une structure sans couplage et en présence d'un agoniste inverse ou d'un antagoniste est considérée comme inactive.

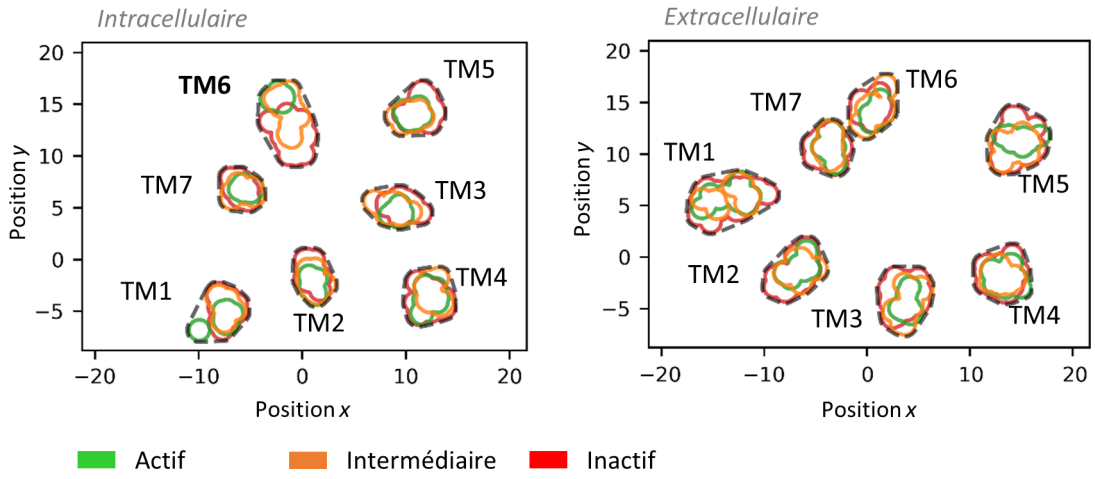

FIGURE 3.15 – Position des extrémités intracellulaire (à gauche) et extracellulaire (à droite) des hélices des structures cristallographiques de RCPG de classe A.

L'alignement de toutes les structures cristallographiques sur les Cα donne un RMSD moyen de $2,20 + -0,45$ Å.

Globalement, toutes les extrémités intracellulaires et extracellulaires des hélices adoptent des positions diverses et ce pour les trois états d'activation, contrairement à ce qui a été observé pour la trajectoire de dynamiques du ADRB2 pour laquelle certaines hélices restaient rigides. Dans la partie extracellulaire, aucune tendance claire ne peut être dégagée, sinon que les TM adoptent des positions variées, et qui se recouvrent pour les trois états d'activation. La position relative des sept hélices est probablement liée à la nature du récepteur. En effet, étant donné la diversité des ligands, de la molécule de faible poids moléculaire à la protéine, la forme de la cavité transmembranaire doit être adaptée à leur liaison.

Si on considère maintenant la partie intracellulaire du 7TM, on remarque que la position du TM6 s'éloigne du centre de la cavité lors de l'activation. De plus, pour cette hélice, on n'observe pas de recouvrement entre les positions dans l'état actif et celles dans l'état inactif. Par contre, la diversité des positions entre ces deux états n'est pas équivalente. Là où les structures actives adoptent un ensemble restreint de positions, l'extrémité du TM6 peut se placer à différents endroits dans les structures inactives. À noter également que les extrémités des TM3 et 7 voisins ne semblent pas aussi fluctuantes que celles du TM6. Il est envisageable que dans l'état inactif, le TM6 s'écarte plus du centre du 7TM en fonction de la nature des résidus qui encombrent plus ou moins le coeur du bouquet d'hélices.

Pour toutes les hélices, on observe la même tendance concernant la position de l'extrémité intracellulaire, qui est donc plus variable pour les RCPG à l'état inactif que pour les RCPG à l'état actif. Il est cependant important de noter que l'échantillon des structures inactives (154) est plus grand que celui des structures actives (29).

Enfin, le mode de visualisation proposé dans ATOLL ne distingue pas les états intermédiaires ni des états actifs et ni de ceux inactifs, ce qui est en accord avec la grande variabilité de ces conformations, pour le même RCPG comme dans l'exemple du ADRB2, ou pour des RCPG différents.

Les fluctuations dans la région intracellulaire pour les états intermédiaires et inactifs ont plusieurs explications possibles. La première est la différence dans la séquence en acides aminés. Pour certains récepteurs, un verrou structural comme le verrou ionique est absent, pouvant

engendrer des différences de positionnement des hélices impliquées. Une deuxième explication pourrait être la définition des états utilisés dans l'étude. En effet, les structures sont considérées comme inactives si un ligand agoniste inverse ou antagoniste est lié. Ces types pharmacologiques distincts ont des implications fonctionnelles différentes, et peuvent correspondre à des conformations différentes du récepteur, en particulier au niveau des hélices intracellulaires. Un autre point qui peut être pris en considération concerne les conditions d'obtention des structures expérimentales, qui sont loin d'être homogènes pour les 213 structures analysées.

Ainsi les projections réalisées avec la méthode ATOLL sont difficiles à interpréter finement pour la comparaison de multiples de RCPG de séquences très différentes. Néanmoins, le mouvement de grande amplitude du TM6 est bien mis en évidence.

3.4.3 Dynamique moléculaire du complexe CCR5–gp120–CD4

La méthode ATOLL a été appliquée aux quatre simulations de CCR5 lié aux gp120 issues des souches #25, #34, Bx08 et JR-FL, décrites dans le chapitre précédent, afin de caractériser la position des hélices transmembranaires dans leur partie intracellulaire.

Comparaison des complexes CCR5–gp120 avec les structures cristallographiques du CCR5

La disposition des hélices dans les structures instantanées issues des trajectoires de dynamique moléculaire a été comparée à celle dans les structures cristallographiques du CCR5 lié à l'agoniste inverse maraviroc (entrée PDB : 4MBS) et la chimiokine antagoniste modifié [5P7]CCL5 (entrée PDB : 5UIW) (FIGURE 3.16). À noter que ces deux structures sont dans l'état inactif sans partenaire intracellulaire. La protéine de fusion, la rubredoxine, utilisée afin de faciliter la formation de cristaux, stabilise les récepteurs dans une conformation inactive [30].

Dans la partie extracellulaire, les structures cristallographiques du CCR5 lié à maraviroc et à la [5P7]CCL5 adoptent des dispositions d'hélices qui recouvrent globalement les zones peuplées du CCR5 dans les structures des complexes CCR5–gp120–CD4 (FIGURE 3.17). Les

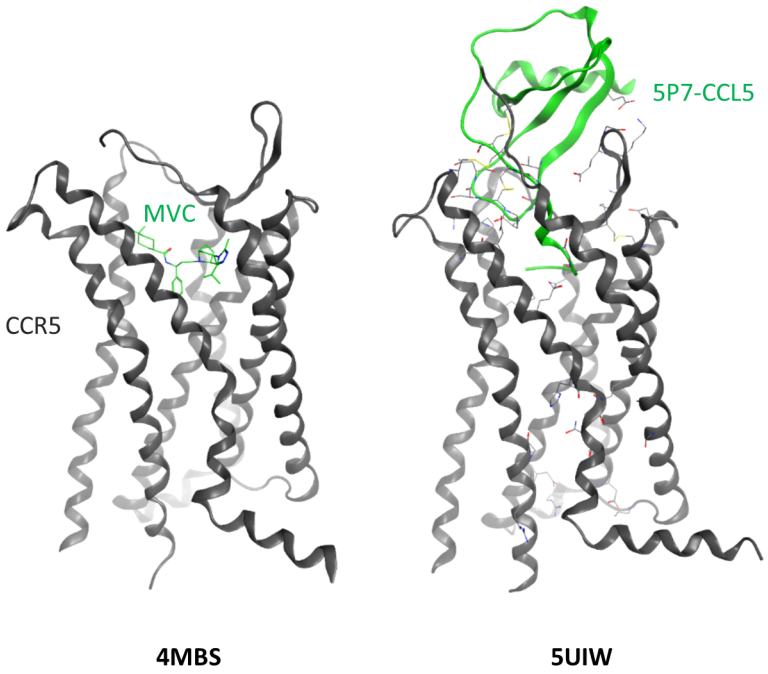

FIGURE 3.16 – Structures cristallographiques du CCR5 lié à maraviroc (entrée PDB : 4MBS) et à la chimiokine modifiée [5P7]CCL5 (entrée PDB : 5UIW). CCR5 et [5P7]CCL5 sont représentés sous forme de rubans gris foncé et vert respectivement. Maraviroc est représenté sous forme de bâtonnets verts.

exceptions sont les pointes du TM4 et du TM7 de CCR5–[5P7]CCL5 qui sont orientées légèrement vers le centre de la cavité (FIGURE 3.17A). Dans la structure CCR5–[5P7]CCL5, la cavité est d'avantage fermée au niveau des TM1, TM2, TM3 et TM7 que dans la structure avec le maraviroc, les TM étant plus orientés vers l'intérieur. Cela dit, les déviations restent faibles. Ce phénomène peut être imputé aux interactions entre la chimiokine et la partie N-terminale et la ECL2 du CCR5, interactions absentes avec maraviroc.

Dans la partie intracellulaire, les extrémités des TM se superposent aux zones peuplées du CCR5 dans les structures des complexes CCR5–gp120–CD4 et peu de différences sont à noter entre CCR5–maraviroc et CCR5–[5P7]CCL5. Seule la pointe du TM6 se distingue, car elle ferme la cavité dans la structure du complexe CCR5–maraviroc, alors qu'elle est orientée d'avantage vers l'extérieur dans la structure du complexe CCR5–[5P7]CCL5 (FIGURE 3.17A, FIGURE 3.18). Ces données sont en accord avec les données pharmacologiques, puisque la liaison du maraviroc au CCR5 exclut celle de la protéine G au récepteur, alors que la [5P7]CCL5 est décrite comme antagoniste et que la gp120 est capable de liaison au CCR5 libre ou lié à une

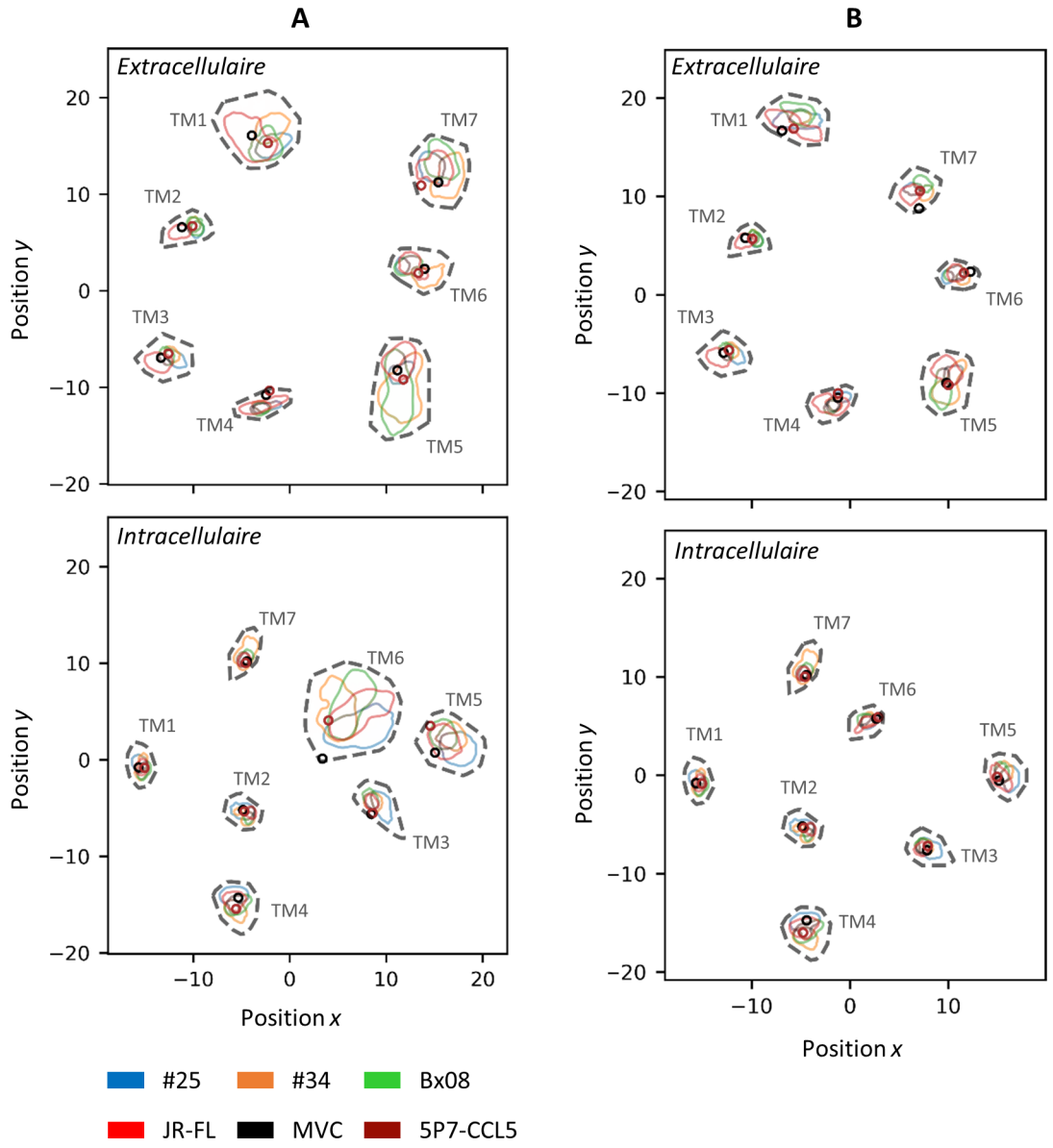

FIGURE 3.17 – Positions des extrémités intracellulaire et extracellulaire des hélices du CCR5 issues des quatre simulations de CCR5–gp120–CD4 et des structures cristallographiques de CCR5–maraviroc et de CCR5–[5P7]CCL5. Les structures des simulations des complexes CCR5–gp120–CD4 sont représentées par des contours délimitant les positions densément peuplées. Les structures expérimentales sont représentées par des cercles. (A) Projection au niveau de la couche 1. (B) Projection au niveau de la couche 2.

protéine G. Cependant, la moitié inférieure du TM6 reste dans une conformation identique dans toutes les structures, caractéristique d'un état inactif ou intermédiaire (FIGURE 3.17B).

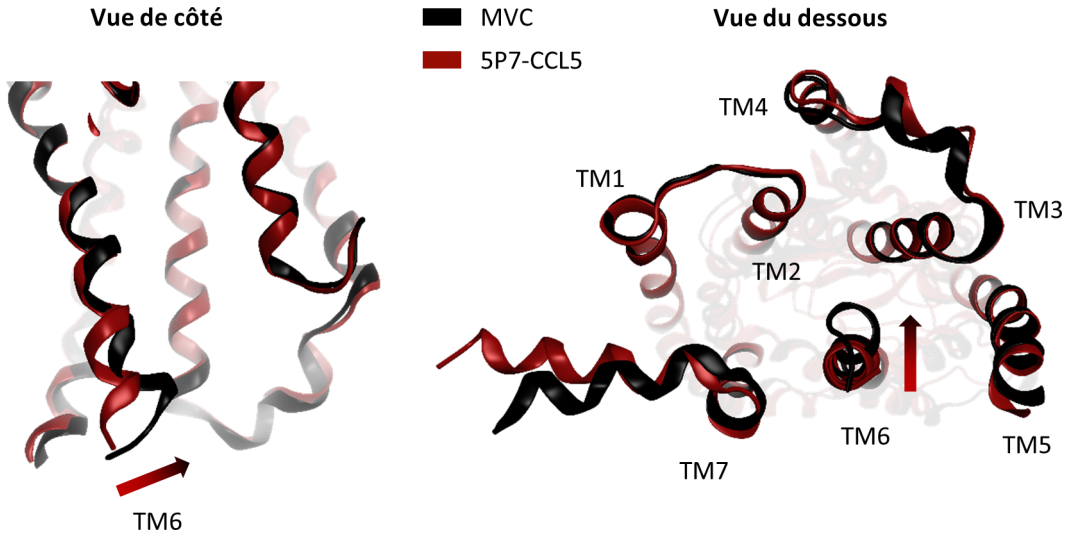

FIGURE 3.18 – Visualisation de la position du TM6 intracellulaire pour les structures des complexes CCR5–maraviroc et CCR5–[5P7]CCL5.

Comparaison des complexes CCR5–gp120–CD4 avec les structures cristallographiques de récepteurs aux chimiokines

La disposition des hélices dans les structures instantanées issues des trajectoires de dynamique moléculaire a aussi été comparée à celle de récepteurs à C-C chimiokine dans leurs structures cristallographiques (FIGURE 3.19). Dans la PDB, les structures disponibles à ce jour décrivent les récepteurs CCR2 (entrées PDB : 5T1A, 6GPS et 6GPX), CCR6 (entrée PDB : 6WWZ), CCR7 (entrée PDB : 6QZH) et CCR9 (entrée PDB : 5LWE).

Seule la structure du CCR6 montre le récepteur dans l'état actif. Le CCR6 y est lié à la protéine G(o) (hétérotrimère avec les sous-unités Gα, Gβ et Gγ) dans la cavité intracellulaire, et à la chimiokine agoniste CCL20 dans la cavité transmembranaire. L'état actif est stabilisé par la liaison d'un fragment d'anticorps scFv16 aux sous-unités Gα et Gβ. Toutes les autres structures montrent des récepteurs liés à des antagonistes, dans un état présumé inactif. Pour les structures 6GPS et 6GPX du CCR2, l'antagoniste est lié dans la cavité transmembranaire, du côté extracellulaire, et une partie du ligand est inséré entre le TM1 et le TM7. Dans la structures

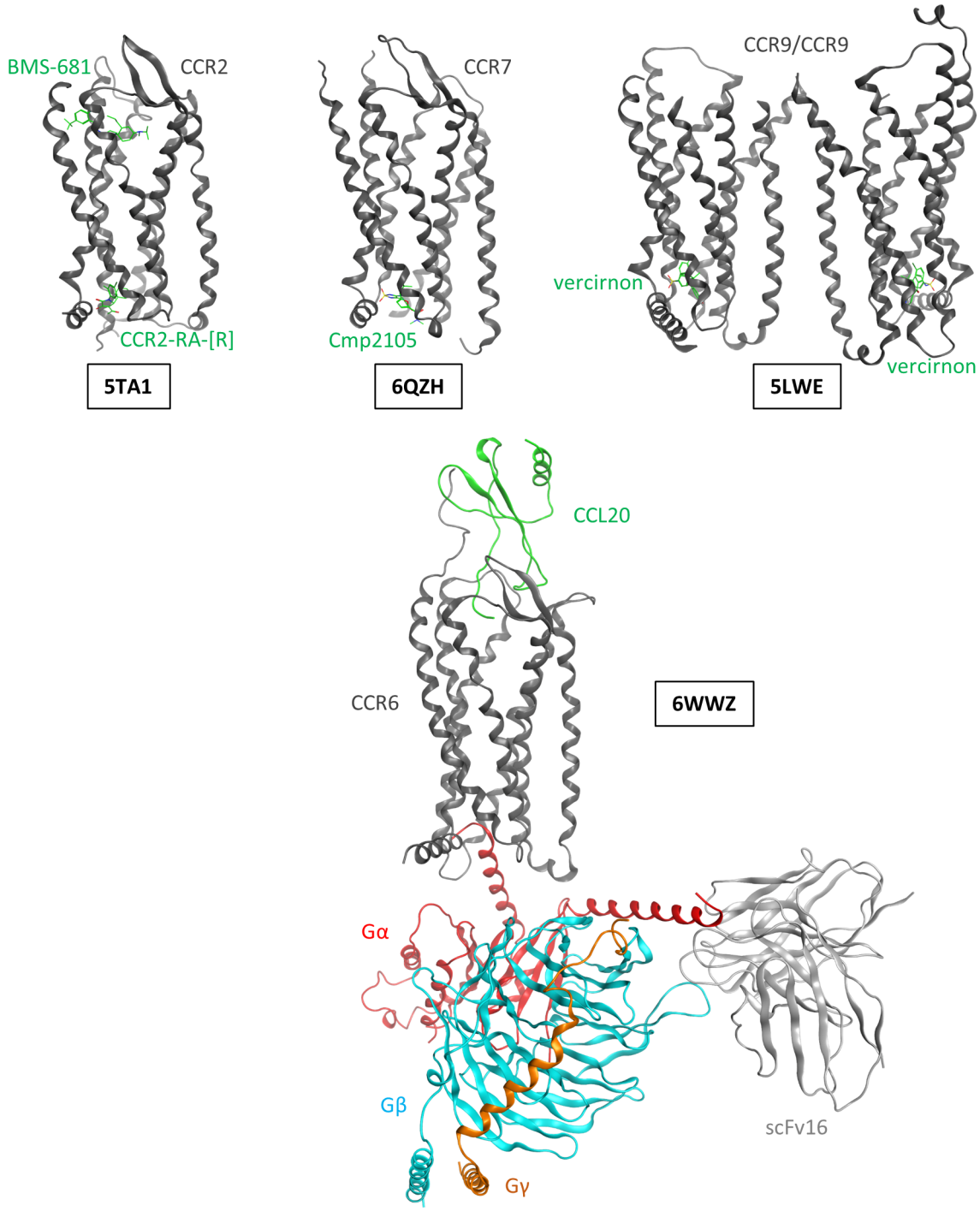

FIGURE 3.19 – Structures cristallographiques des CCR2, CCR6, CCR7 et CCR9. Les récepteurs aux chimiokines sont représentés sous forme de rubans gris foncés. Les ligands sont sous la forme de bâtonnets verts pour les molécules organiques ou de rubans verts pour la chimiokine. Dans la structure du CCR6, la protéine G est représentée par des rubans (Gα : rouges, Gβ : cyans, Gγ : oranges) ainsi que le fragment d'anticorps scFv16 par des rubans gris clairs. Les codes PDB de chaque structure sont encadrés.

5LWE du CCR9 et 6QZH du CCR7, l'antagoniste est lié dans une autre cavité du 7TM, plus près du côté intracellulaire. L'entrée 5T1A du CCR2 présente une autre particularité : deux antagonistes sont présents dans le 7TM, un dans la cavité donnant sur le côté extracellulaire et l'autre du côté intracellulaire. Les trois structures de CCR2 sont proches (RMSD < 1 Å). En conséquence seule la structure 5T1A a été utilisée pour la projection des TM, du fait de sa meilleur résolution (2,81 Å). Il est important de préciser que la structure du CCR9 (5LWE) est la seule qui décrit un homodimère dont les protomètres sont très similaires (RMSD $= 0{,}83$ Å).

Du point de vue des séquences, trois groupes de récepteurs peuvent être établis : CCR5/CCR2, CCR7/CCR9 et CCR6 (FIGURE 3.20). CCR5 et CCR2 sont très proches, avec 82,5 % de résidus identiques dans les TM (FIGURE 3.21) et assez distants du CCR6, CCR7 et CCR9 ($\geq$ 40,1 % sur les TM). Quant à eux, CCR7 et CCR9 ont des séquences similaires en considérant la nature des acides aminés mais nettement moins que CCR2/CCR5 (similarité d'environ 70 % comparée à 90 %). CCR6 est plus proche des récepteurs CCR7 et CCR9 que de CCR2 et CCR5 mais reste somme toute un homologue distant.

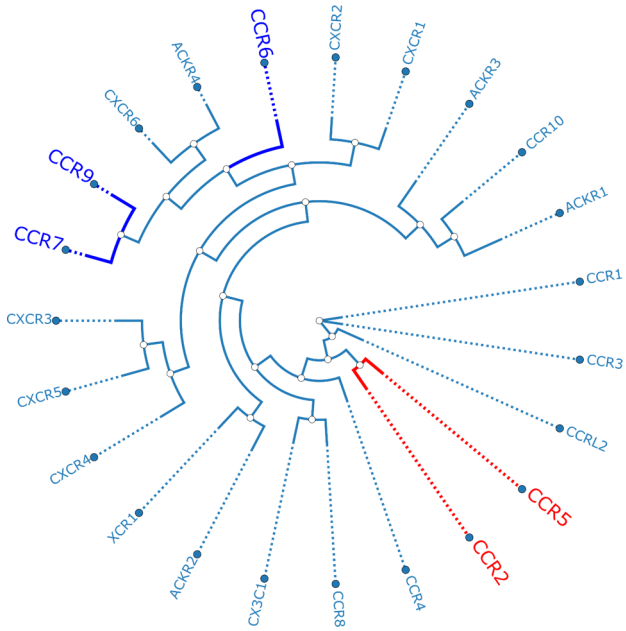

FIGURE 3.20 – Arbre phylogénétique des récepteurs aux chimiokines basé sur la similarité des séquences. Image générée depuis le site de la GPCRdb.

Cela dit, il existe des motifs communs à tous ces récepteurs : T(2.56)xP, résidu Arg/Lys en 6.30, P(6.50)xN, N(7.49)PxxY. À noter la présence d'un résidu Glu en position 7.39 dans CCR2,

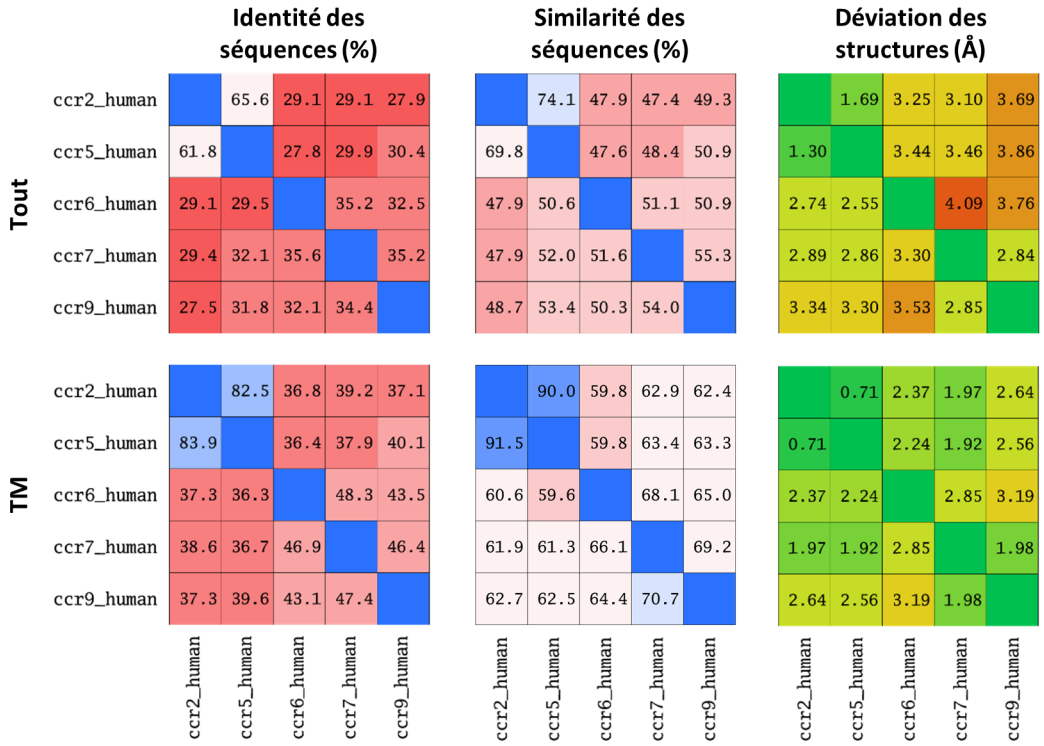

FIGURE 3.21 – Identité, similarité des séquences et déviation des structures des CCR2, CCR5, CCR6, CCR7 et CCR9. Les valeurs peuvent ne pas être identiques dans le triangle supérieur et dans le triangle inférieur car la taille des objets est différente.

CCR5, et CCR6, responsable d'une interaction forte entre le maraviroc ou la gp120 avec CCR5 (Glu283). La nature de la cavité est similaire pour les cinq récepteurs avec une sous-poche 1 hydrophobe et une sous-poche 2 plutôt polaire mais peu conservée [31].

Dans la région extracellulaire, les extrémités des hélices du CCR2 et du CCR6 ont une organisation qui se rapproche de celle observée pour les complexes CCR5–gp120–CD4, à l'exception du TM7 qui est légèrement plus proche du centre (FIGURE 3.22). À l'inverse, les structures du CCR7 et du CCR9 s'écartent le plus des structures du CCR5, notamment au niveau du TM3, qui dévie largement vers l'intérieur du récepteur dans le CCR9, et du TM5, qui s'éloigne franchement du centre du récepteur, aussi dans le CCR9.

Au niveau intracellulaire, on remarque que la position du TM6 du CCR6 est caractéristique de l'état actif (FIGURE 3.22). En se focalisant sur l'extrémité du TM6 plus proche de la ECL3, cette partie est orientée vers l'extérieur pour le CCR2, le CCR6 et le CCR9 contrairement à ce qui a été observé dans le complexe CCR5–maraviroc. Il faut noter que le CCR7 n'est pas

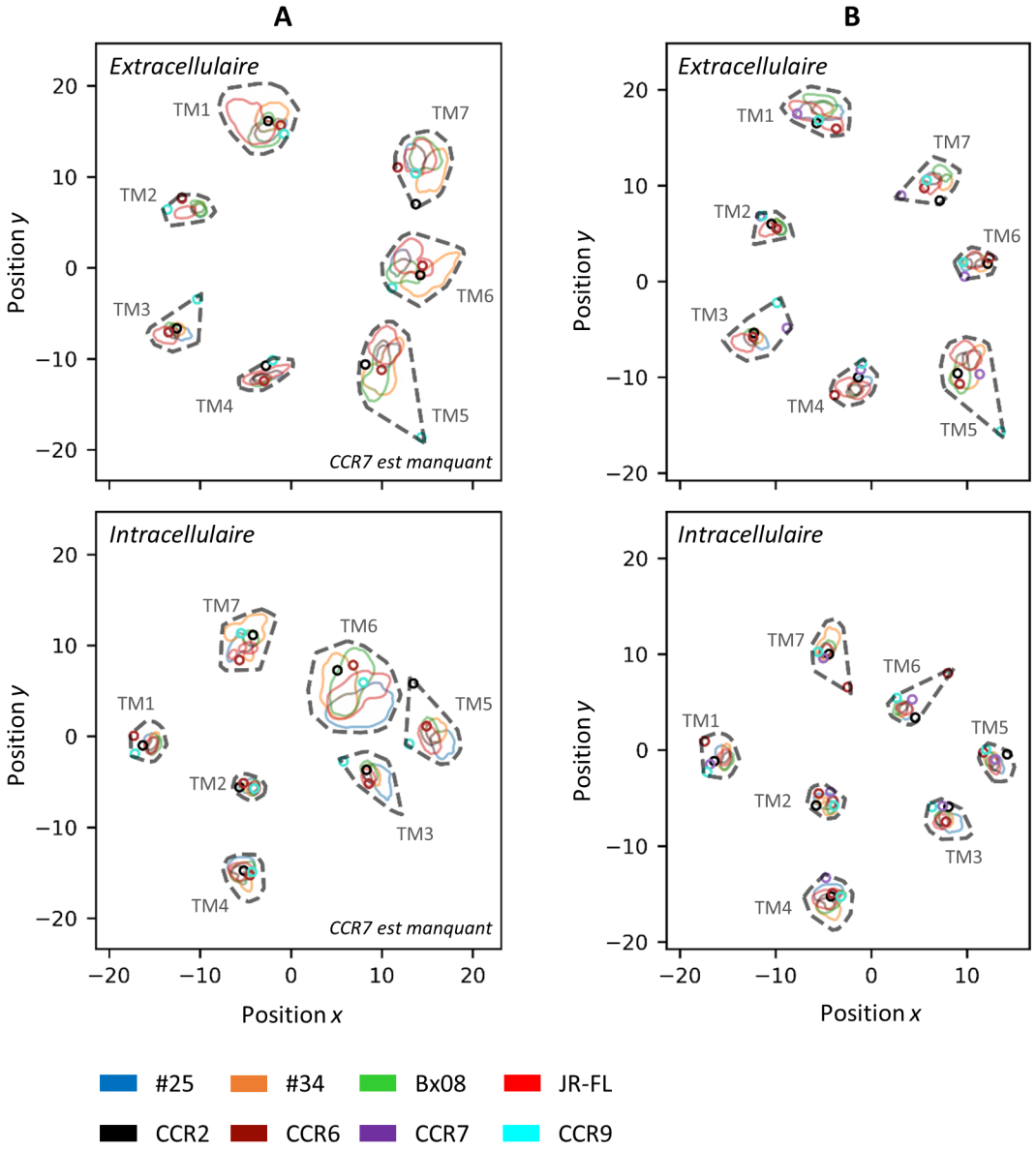

FIGURE 3.22 – Position des extrémités intracellulaire et extracellulaire des hélices du CCR5 issue des 4 simulations de CCR5–gp120–CD4 et des structures cristallographiques de CCR2, CCR6, CCR7 et CCR9. Les structures des simulations des complexes CCR5–gp120 sont représentées par des contours délimitant les positions densément peuplées. Les structures expérimentales sont représentées par des cercles. (A) Projection au niveau de la couche 1 (CCR7 n'est pas visible car des résidus ne sont pas résolus à cet endroit). (B) Projection au niveau de la couche 2.

représenté car les acides aminées sont mal définis dans cette région. Une autre particularité se situe sur le TM5 du CCR2 qui dévie largement par rapport aux positions des autres structures. Cependant, cette déviation n'est pas pertinente car elle est due à la protéine de fusion qui déforme l'extrémité de l'hélice.

Comparaison des complexes CCR5–gp120 avec le CCR5 libre, lié à CCL3 et lié au maraviroc

Pour finir, les trajectoires de dynamiques des complexes CCR5–gp120–CD4 ont été comparées à celle du CCR5 libre, et des complexes CCR5–CCL3 et CCR5–maraviroc. Toutes les simulations ont été réalisées en utilisant strictement les mêmes conditions pour une durée totale de 500 ns par système.

Les projections dans la partie extracellulaire indiquent des dispositions légèrement différentes des hélices entre les systèmes CCR5 libre, CCR5–CCL3 et CCR5–maraviroc et celles des complexes CCR5–gp120–CD4. Le TM1 se démarque clairement dans les systèmes du CCR5 libre, lié à la CCL3 ou au maraviroc par rapport aux complexes de CCR5–gp120–CD4, dans lesquels il s'oriente vers l'extérieur de la cavité, du fait de la contrainte imposée par la liaison au coeur de la gp120 (FIGURE 3.23). Les TM2, TM3 et TM4 distinguent les structures de CCR5–CCL3 de toutes les autres. Enfin, la variabilité de la position du TM5 distingue les structures CCR5–gp120–CD4 pour les variantes #34 et Bx08 de la gp120.

Dans la partie intracellulaire, la zone la plus intéressante est encore une fois celle de l'extrémité du TM6. En effet pour les trois systèmes CCR5 libre, CCR5–CCL3 et CCR5–maraviroc, on remarque trois profils distincts, tous différents de ceux observés pour les complexes CCR5–gp120–CD4. Le résultat le plus probant concerne le système CCR5–maraviroc pour lequel l'extrémité du TM6 ferme la cavité intracellulaire sur toute la durée de la trajectoire contrairement aux systèmes libres du CCR5 et lié à la CCL3.

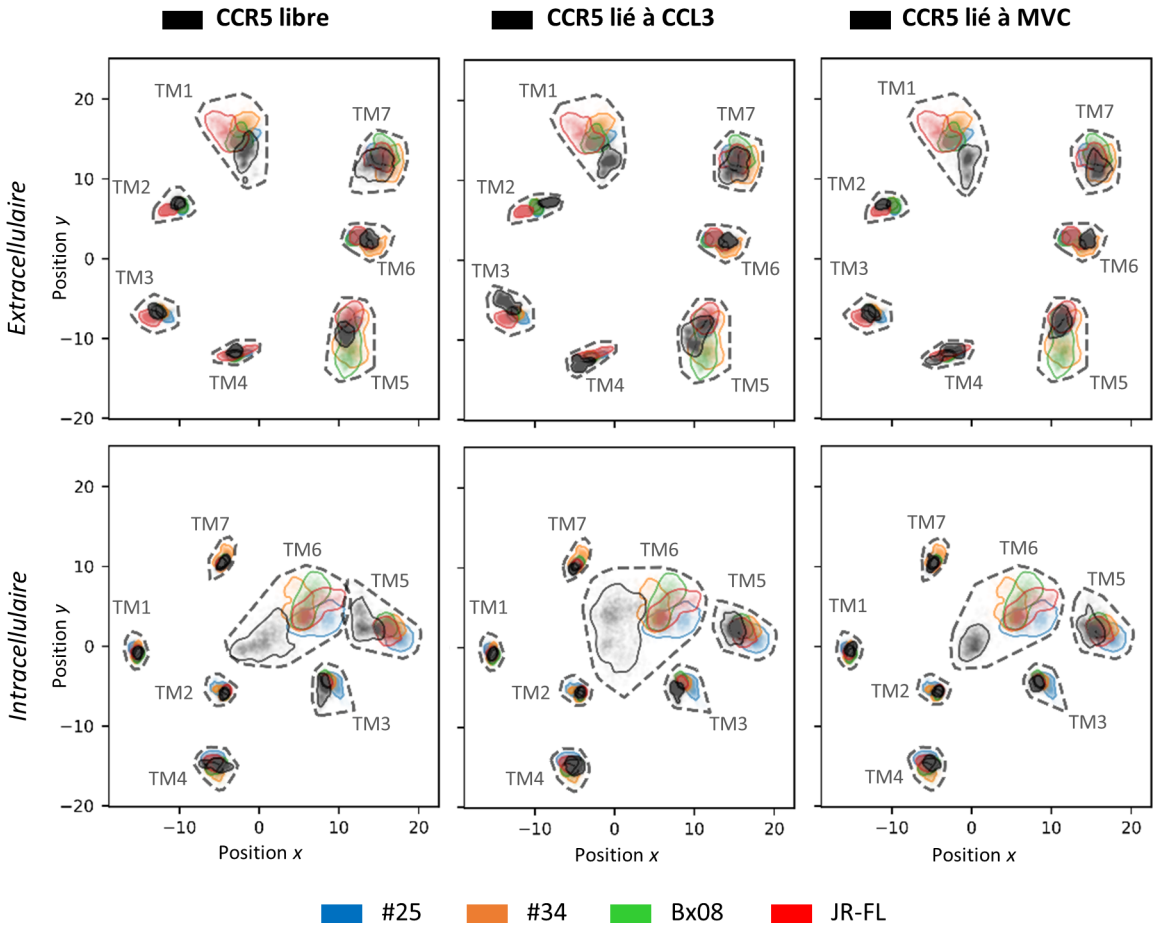

FIGURE 3.23 – Positions des extrémités intracellulaire et extracellulaire des hélices du CCR5 sans ligand (libre), CCR5–CCL3, CCR5–maraviroc et CCR5–gp120–CD4 issues des simulations par dynamique moléculaire. Seule la couche 1 est projetée.

3.5 Conclusion

La méthode ATOLL présentée dans ce chapitre permet de caractériser les différentes populations d'un récepteur. ATOLL permet de discerner les trois états d'activation du récepteur ADRB2 lors de la simulation de sa désactivation par dynamique moléculaire. La méthode permet également d'observer des sous-populations de l'état intermédiaire qui n'ont pas été relevées par les auteurs de l'étude [23]. La méthode appliquée à un large ensemble de structures cristallographiques de RCPG de classe A met en évidence la diversité structurale des récepteurs étudiés, y compris pour un même état d'activation. Elle révèle un seul descripteur caractéristique de l'activation des RCPG, à savoir le mouvement du TM6 dans la partie intracellulaire.

Une fois la méthode validée, ATOLL a été appliquée aux simulations par dynamique moléculaires du complexe CCR5–gp120–CD4 avec les variantes #25, #34, Bx08 et JR-FL de la gp120. Ces structures simulées ont été comparées à des structures cristallographiques du CCR5, à celles des récepteurs aux chimiokines CCR2, CCR6, CCR7 et CCR9, ainsi qu'à celles obtenues par simulations du CCR5 libre (sans ligand), lié à la CCL3 et lié au maraviroc.

Comme attendu, la partie fournissant le plus d'informations se situe au niveau de l'extrémité du TM6 proche de la ICL3. Celle-ci se place différemment en fonction du ligand. Elle ferme la cavité transmembranaire dans les structures RX et simulées du complexe CCR5–maraviroc, conformément au caractère agoniste inverse du ligand. Les gp120 et la chimiokine antagoniste [5P7]CCL5 induisent des placements du TM6 différents de celui induit par maraviroc ou observé dans le récepteur libre. Ces placements se distinguent néanmoins chacun l'un de l'autre, suggérant une possible signalisation biaisée de ces ligands. Dans le complexe de CCR5 impliquant un agoniste, CCL3, le positionnement du TM6 se distingue de tous les autres sans pour autant correspondre à celui d'un récepteur pleinement activé, comme celui décrit dans la structure expérimentale du CCR6. Nous pouvons cependant faire l'hypothèse que la population simulée montre un état intermédiaire entre des états inactif et actif, comme lors de la simulation du ADRB2. Dans CCR5, les mouvements observés ne concernent que l'extrémité du TM6, qui se plie tout en gardant les résidus insérés dans la membrane dans des positions caractéristiques

d'un état inactif. En comparaison, dans la structure à l'état actif du CCR6, le TM6 imprime un mouvement sur toute la moitié inférieure de l'hélice. Cela dit, nos résultats suggèrent que la seule considération de l'extrémité du TM6 permet de prédire les propriétés fonctionnelles du ligand, et ce même si les structures analysées sont dans un état qui n'est pas actif. Cette méthode est donc utilisable sur des simulations relativement courtes, et est donc d'autant plus intéressante qu'il est compliqué de modéliser l'activation d'un RCPG par dynamique moléculaire. Très peu de travaux publiés décrivent un tel phénomène. Au laboratoire, Florian KOENSGEN a tenté de simuler l'activation du CCR5 par dynamique moléculaire accélérée, mais sans succès.

Enfin, ATOLL ne s'applique pas qu'aux RCPG mais peut être étendu à d'autres familles de protéines transmembranaires afin de comparer l'organisation de leur domaines transmembranaires ou leurs états conformationnels. C'est notamment le cas des canaux ioniques qui sont aussi une cible de choix pour le développement de médicaments.

Maintenant que nous avons montré que les complexes CCR5–gp120–CD4 portent des signatures distinctes dans la partie intracellulaire, et que celles-ci peuvent être interprétées en terme fonctionnel, nous allons explorer la possibilité d'identifier des petites molécules capables de mimer les modes de liaison des différentes gp120 et ainsi induire la même signature structurale au niveau intracellulaire. Pour cela, nous avons développé une méthode capable d'intégrer toutes les interactions observées lors d'une simulation par dynamique moléculaire pour les utiliser ensuite dans une campagne de criblage virtuel.

3.6 Références

[1] Daniel Hilger, Matthieu Masureel, and Brian K. Kobilka. Structure and dynamics of GPCR signaling complexes. *Nature Structural & Molecular Biology*, 25(1) :4–12, Janvier 2018.

[2] Malin C. Lagerström and Helgi B. Schiöth. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nature Reviews Drug Discovery*, 7(4) :339–357, Avril 2008.

[3] Rita Santos, Oleg Ursu, Anna Gaulton, A. Patrícia Bento, Ramesh S. Donadi, Cristian G. Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I. Oprea, and John P. Overington. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1) :19–34, Janvier 2017.

[4] Miles Congreve, Chris de Graaf, Nigel A. Swain, and Christopher G. Tate. Impact of GPCR Structures on Drug Discovery. *Cell*, 181(1) :81–91, Avril 2020.

[5] Vinzenz M. Unger, Paul A. Hargrave, Joyce M. Baldwin, and Gebhard F. X. Schertler. Arrangement of rhodopsin transmembrane α-helices. *Nature*, 389(6647) :203–206, Septembre 1997.

[6] Krzysztof Palczewski, Takashi Kumasaka, Tetsuya Hori, Craig A. Behnke, Hiroyuki Motoshima, Brian A. Fox, Isolde Le Trong, David C. Teller, Tetsuji Okada, Ronald E. Stenkamp, Masaki Yamamoto, and Masashi Miyano. Crystal Structure of Rhodopsin : A G Protein-Coupled Receptor. *Science*, 289(5480) :739–745, Août 2000.

[7] Juan A. Ballesteros and Harel Weinstein. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In Stuart C. Sealfon, editor, *Methods in Neurosciences*, volume 25 of *Receptor Molecular Biology*, pages 366–428. Academic Press, Janvier 1995.

[8] Michael J. Capper and Daniel Wacker. How the ubiquitous GPCR receptor family selectively activates signalling pathways. *Nature*, 558(7711) :529–530, Juin 2018.

[9] Yanyong Kang, Oleg Kuybeda, Parker W. de Waal, Somnath Mukherjee, Ned Van Eps, Przemyslaw Dutka, X. Edward Zhou, Alberto Bartesaghi, Satchal Erramilli, Takefumi Morizumi, Xin Gu, Yanting Yin, Ping Liu, Yi Jiang, Xing Meng, Gongpu Zhao, Karsten Melcher, Oliver P. Ernst, Anthony A. Kossiakoff, Sriram Subramaniam, and H. Eric Xu. Cryo-EM structure of human rhodopsin bound to an inhibitory G protein. *Nature*, 558 (7711) :553–558, Juin 2018.

[10] Javier García-Nafría, Rony Nehmé, Patricia C. Edwards, and Christopher G. Tate. Cryo-EM structure of the serotonin 5-HT 1B receptor coupled to heterotrimeric G o. *Nature*, 558(7711) :620–623, Juin 2018.

[11] Christopher J. Draper-Joyce, Maryam Khoshouei, David M. Thal, Yi-Lynn Liang, Anh T. N. Nguyen, Sebastian G. B. Furness, Hariprasad Venugopal, Jo-Anne Baltos, Jürgen M. Plitzko, Radostin Danev, Wolfgang Baumeister, Lauren T. May, Denise Wootten, Patrick M. Sexton, Alisa Glukhova, and Arthur Christopoulos. Structure of the adenosine-bound human adenosine A 1 receptor–G i complex. *Nature*, 558(7711) :559–563, Juin 2018.

[12] Antoine Koehl, Hongli Hu, Shoji Maeda, Yan Zhang, Qianhui Qu, Joseph M. Paggi, Naomi R. Latorraca, Daniel Hilger, Roger Dawson, Hugues Matile, Gebhard F. X. Schertler, Sebastien Granier, William I. Weis, Ron O. Dror, Aashish Manglik, Georgios Skiniotis, and Brian K. Kobilka. Structure of the μ-opioid receptor–G i protein complex. *Nature*, 558 (7711) :547–552, Juin 2018.

[13] Florian Koensgen, Franck Da Silva, Didier Rognan, and Esther Kellenberger. Unsupervised Classification of G-Protein Coupled Receptors and Their Conformational States Using IChem Intramolecular Interaction Patterns. *Journal of Chemical Information and Modeling*, 59(9) :3611–3618, Septembre 2019.

[14] Mikhail A. Lomize, Irina D. Pogozheva, Hyeon Joo, Henry I. Mosberg, and Andrei L. Lomize. OPM database and PPM web server : resources for positioning of proteins in membranes. *Nucleic Acids Research*, 40(D1) :D370–D376, Janvier 2012.

[15] Andrei L. Lomize, Irina D. Pogozheva, and Henry I Mosberg. Anisotropic Solvent Model of the Lipid Bilayer. 2. Energetics of Insertion of Small Molecules, Peptides, and Proteins in Membranes. *Journal of Chemical Information and Modeling*, 51(4) :930–946, Avril 2011.

[16] Fábio Madeira, Young mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R. N. Tivey, Simon C. Potter, Robert D. Finn, and Rodrigo Lopez. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, 47(W1) :W636–W641, Juillet 2019.

[17] Gáspár Pándy-Szekeres, Christian Munk, Tsonko M. Tsonkov, Stefan Mordalski, Kasper Harpsøe, Alexander S. Hauser, Andrzej J. Bojarski, and David E. Gloriam. GPCRdb in 2018 : adding GPCR structure models and ligands. *Nucleic Acids Research*, 46(D1) : D440–D446, Janvier 2018.

[18] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Studholme, Corin Yeats, and Sean R. Eddy. The Pfam protein families database. *Nucleic Acids Research*, 32(suppl_1) :D138–D141, Janvier 2004.

[19] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443–453, Mars 1970.

[20] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22) :10915–10919, Novembre 1992.

[21] D. L. Theobald. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallographica Section A : Foundations of Crystallography*, 61(4) : 478–480, Juillet 2005.

[22] Pu Liu, Dimitris K. Agrafiotis, and Douglas L. Theobald. Fast determination of the optimal rotational matrix for macromolecular superpositions. *Journal of Computational Chemistry*, 31(7) :1561–1563, 2010.

[23] Ron O. Dror, Daniel H. Arlow, Paul Maragakis, Thomas J. Mildorf, Albert C. Pan, Huafeng Xu, David W. Borhani, and David E. Shaw. Activation mechanism of the β2-adrenergic receptor. *Proceedings of the National Academy of Sciences*, 108(46) :18684–18689, Novembre 2011.

[24] David E. Shaw, Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *ACM SIGARCH Computer Architecture News*, 35(2) :1–12, Juin 2007.

[25] Søren G. F. Rasmussen, Hee-Jung Choi, Juan Jose Fung, Els Pardon, Paola Casarosa, Pil Seok Chae, Brian T. DeVree, Daniel M. Rosenbaum, Foon Sun Thian, Tong Sun Kobilka, Andreas Schnapp, Ingo Konetzki, Roger K. Sunahara, Samuel H. Gellman, Alexander Pautsch, Jan Steyaert, William I. Weis, and Brian K. Kobilka. Structure of a nanobody-stabilized active state of the β 2 adrenoceptor. *Nature*, 469(7329) :175–180, Janvier 2011.

[26] Aaron M. Ring, Aashish Manglik, Andrew C. Kruse, Michael D. Enos, William I. Weis, K. Christopher Garcia, and Brian K. Kobilka. Adrenaline-activated structure of β 2 - adrenoceptor stabilized by an engineered nanobody. *Nature*, 502(7472) :575–579, Octobre 2013.

[27] Florian Koensgen. *Modélisation du récepteur aux chimiokines C-C de type 5 : caractérisation des états conformationnels et conception rationnelle de modulateurs de la dimérisation*. These de doctorat, Strasbourg, Octobre 2018.

[28] Caterina Bissantz, Antoine Logean, and Didier Rognan. High-Throughput Modeling of Human G-Protein Coupled Receptors : Amino Acid Sequence Alignment, Three-Dimensional Model Building, and Receptor Library Screening. *Journal of Chemical Information and Computer Sciences*, 44(3) :1162–1176, Mai 2004.

[29] Chris de Graaf and Didier Rognan. Customizing G Protein-Coupled Receptor Models for Structure-Based Virtual Screening. *Current Pharmaceutical Design*, 15(35) :4026–4048, Décembre 2009.

[30] Ramin Ekhteiari Salmas, Mine Yurtsever, and Serdar Durdagi. Investigation of Inhibition Mechanism of Chemokine Receptor CCR5 by Micro-second Molecular Dynamics Simulations. *Scientific Reports*, 5(1) :13180, Août 2015.

[31] Jean-Sebastien Surgand, Jordi Rodrigo, Esther Kellenberger, and Didier Rognan. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins : Structure, Function, and Bioinformatics*, 62(2) :509–538, 2006.

3.7 Annexes

3.7.1 Code source

mol2seq.py

Code source

```
import sys
import os
import pdb
from glob import glob

# Run with Python 3

"""
    Notes:
        * This script does not handle special amino acid notations
          (engineered, d-aa, ...). Add entry in the 'convert_3l_to_1l' dict.
"""

_CONVERT_3L_TO_1L = {
    'ARG': 'R', 'HIS': 'H', 'HIE': 'H', 'HID': 'H', 'LYS': 'K', 'ASP': 'D',
    'ASN': 'N', 'GLU': 'E', 'GLN': 'Q', 'SER': 'S', 'THR': 'T', 'CYS': 'C',
    'SEC': 'U', 'GLY': 'G', 'PRO': 'P', 'ALA': 'A', 'VAL': 'V', 'ILE': 'I',
    'LEU': 'L', 'MET': 'M', 'PHE': 'F', 'TYR': 'Y', 'TRP': 'W'}

class ProteinSequence:
    def __init__(self):
        self.name = None
        self.resids = None
        self.sequence = None
        self.extended_resids = None
        self.extended_sequence = None

    @property
    def start(self):
        return self.resids[0]

    @property
    def end(self):
        return self.resids[-1]

    def load_mol2(self, mol2_filepath):
        """
        Extract the amino acid sequence (1 letter) from à MOL2 file.

        Parameters:
```

```
            mol2_filepath (str): The path of MOL2 file.
        """
        resids = []
        sequence = []

        with open(mol2_filepath) as f:
            # Parse protein name
            while next(f) != '@<TRIPOS>MOLECULE\n':
                pass
            self.name = next(f).strip()

            # Parse residues
            while next(f) != '@<TRIPOS>SUBSTRUCTURE\n':
                pass

            for line in f:
                if line[0] == '@':
                    break

                items = line.split()
                try:
                    residue = items[1]
                except IndexError as e:
                    if not items:
                        continue
                    else:
                        raise IndexError(e)

                resname = residue[:3]
                try:
                    resid = int(residue[3:])
                except ValueError as e:
                    if items[0] == '#':
                        continue
                    else:
                        raise ValueError(e)

                resids.append(resid)
                sequence.append(_CONVERT_3L_TO_1L[resname])

            self.resids = resids
            self.sequence = sequence

    def add_gap(self, gap_char='-'):
        previous_resid = 0
        extended_resids = []
        extended_sequence = []

        for resid, resname in zip(self.resids, self.sequence):
            shift = resid - previous_resid - 1

            # If gap in sequence numbering is detected, add gap character
            for i in range(shift):
                extended_resids.append(previous_resid+i)
```

```
                extended_sequence.append(gap_char)

            extended_resids.append(resid)
            extended_sequence.append(resname)

            previous_resid = resid

        self.extended_resids = extended_resids
        self.extended_sequence = extended_sequence

    def to_stockholm(self):
        entry_name = f'{self.name}/{self.start}-{self.end}'
        entry = '{:<30}{}'.format(entry_name, ''.join(self.extended_sequence))

        return entry

def stockholm_writer(sequences, filepath):
    n_sequences = len(sequences)

    with open(filepath, 'w') as f:
        f.write('# STOCKHOLM 1.0\n')
        f.write(f'#=GF SQ {n_sequences}\n')

        f.write('\n'.join([sequence.to_stockholm() for sequence in sequences]))

        f.write('\n//')


def main(args):
    mol2_files = glob(os.path.join(args.input, '*mol2'))
    mol2_files.sort()
    sequences = []
    for mol2_file in mol2_files:
        sequence = ProteinSequence()
        sequence.load_mol2(mol2_file)
        sequence.add_gap()
        sequences.append(sequence)

    stockholm_writer(sequences, args.output)

if __name__ == '__main__':
    import argparse

    parser = argparse.ArgumentParser()
    parser.add_argument('-i', '--input')
    parser.add_argument('-o', '--output')

    parser.set_defaults(func=main)
    args = parser.parse_args()

    # Run
    status = args.func(args)
    sys.exit(status)
```

Chapitre 4

LID : une cartographie des motifs d'interaction pour évaluer les poses de docking

4.1 Préambule du chapitre

Ce chapitre présente le développement et la validation d'une méthode conçue pour le tri de poses de docking appelée Local Interaction Density (LID). Tout le contenu du chapitre est issu de l'article publiée en 2019 dans le journal *Molecule* de l'éditeur *MDPI* [1] et est en accès libre. La méthode est basée sur la comparaison des modes de liaison de poses de docking avec des modes de liaison dits de référence, issus de structures cristallographiques par exemple. Contrairement à d'autres méthodes comparant les modes de liaison deux à deux, LID regroupe les références en une seule représentation et calcule une carte de densité en fonction du type d'interaction (liaison hydrogène, contact hydrophobe, etc.). Les poses de docking pour lesquelles les interactions sont situées dans les zones denses des cartes sont ainsi retenues en priorité. Initialement, la méthode a été conçue pour fusionner les modes de liaison de très petites molécules (fragment ou additifs de cristallisation) avec des modes de liaison diverses afin de prédire le mode de liaison de molécules plus grandes.

Du fait de sa rapidité d'exécution et de la manière dont sont représentées les interactions, la méthode peut tout à fait s'appliquer à une trajectoire de dynamique moléculaire dont les différentes structures constitueraient le jeu de référence. La méthode serait alors utilisée afin d'identifier des petites molécules capables d'imiter les modes de liaison des variantes #25, #34, Bx08 et JR-FL de la gp120 avec le CCR5. La procédure consistera à docker dans des structures représentatives du CCR5 les millions de molécules d'une chimiothèque construite dans notre laboratoire à partir de catalogues commerciaux. Les poses de docking seront réévaluées avec la méthode LID en se basant sur les cartes de densité produites à partir des simulations des complexes CCR5–gp120–CD4. La première étape consistera à détecter les interactions (liaison hydrogène, interaction ionique, contact hydrophobe, empilement-π) entre la ECL2 et la cavité TM du CCR5 et la V3 des différentes gp120. Les quatre jeux de référence correspondant aux quatre variantes comporteront chacun 1500 modes de liaison (1 structure sur 10 dans la trajectoire). Leurs cartes de densité en interactions seront calculées avec le module *intgrid* de LID. Les structures du CCR5 utilisées pour le docking seront des structures représentatives de chaque trajectoire, sélectionnées par une classification ascendante hiérarchique (voir le chapitre 2). Chaque

système comportera 15 structures représentatives soit un total de 45 structures de la protéine dans lesquelles toutes les molécules de la chimiothèque seront dockées. Le docking sera effectué par le programme PLANTS, qui a déjà fait ses preuves pour sa capacité à générer des poses de docking pertinentes. Un total de 10 poses seront générées par ligand. Toutes les poses seront réévaluées avec le module *scoring* de LID pour identifier des touches virtuelles dont les modes de liaison reproduisent les motifs d'interaction au CCR5 de chacune des variantes de gp120. Chaque liste de touches comportera une centaine des molécules qui seront par la suite évaluées expérimentalement.

4.2 Introduction

Predicting how a ligand binds to a protein is one of the challenges of structural bioinformatics. In the early eighties, KUNTZ et al. proposed a geometric model of molecular recognition [2]. Since then, a plethora of programs have been developed to dock a ligand at its protein site based on both shape and electrostatic (or pharmacophoric) complementary. The protein is mostly treated as rigid body and ligand conformations are sampled by varying torsion angles. This approach allows for rapid prediction, so that docking has established itself as a method of choice for high throughput applications such as virtual screening. The literature reports many cases of identification of bioactive compounds by serial docking to a target protein [3]. However, examples confirming experimentally the predicted poses are rarer [4]. Benchmarking studies show that the quality of prediction is variable, although significant progress has been made over the past 15 years. The weakness of docking lies mainly in the scoring functions [5, 6]. The widely used docking programs are generally able to predict a ligand/protein three-dimensional (3D) structure similar to that observed by X-ray crystallography, but their scoring function does not necessarily reward it as the best. Logically, scoring functions are not very effective in more difficult exercises, such as distinguishing between active and inactive molecules on a target protein and ranking active molecules by binding affinity [7, 8]. A recent review by GUEDES et al. provides a good overview of empirical functions and their recent and future developments, while discussing the evolution in the design of test datasets, the contribution of learning machines and challenging topics [9].

Docking performance in pose prediction and virtual screening can be improved by post-processing the docking poses based on the analysis of the interactions formed between the ligand and the protein. The underlying assumption is that the binding mode of a relevant pose shares similarities with experimentally validated binding modes. The suggestion of DENG et al. in 2004 to convert interactions in a numerical fingerprint naturally led to the design of several simple and fast methods to compare two binding modes [10]. Similarity is evaluated according to the presence/absence of interactions in the two interaction fingerprints [11]. Interaction fingerprint has become a useful tool for drug discovery [12]. In 2013, we proposed to encode the binding

mode by an interaction pattern graph, so that the similarity score also takes into account the spatial relationships of interacting atoms. This method, called GRIM, is overall more efficient than our in-house interaction fingerprint (IFP) in pose prediction and in virtual screening, but is also more costly in computation time [13, 14]. The advantages of GRIM rescoring with respect to standard energy-based scoring functions have notably been acknowledged for various targets in two recent international docking contests [15, 16].

The success of rescoring with GRIM or IFP depends on the experimental reference 3D structures, which must include a relevant binding mode (FIGURE 4.1).

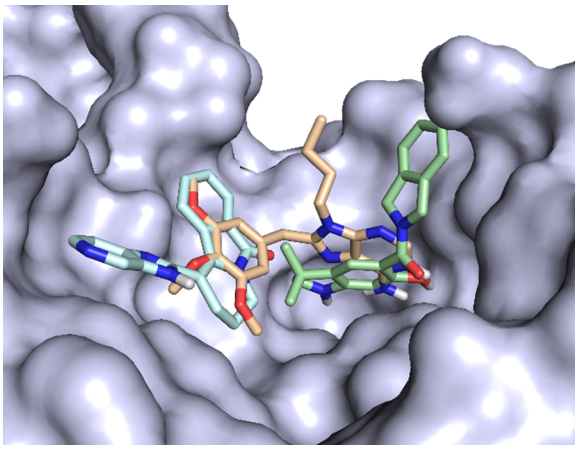

Figure 4.1 – Different binding modes to the heat shock protein HSP 90-alpha. The protein surface is colored in grey, the ligand carbon atoms in cyan (HET code: YJX, PDB code: 2YJX), light orange (HET code: PU3, PDB code: 1UY6) and pale green (HET code: L81, PDB code: 2YJX).

In addition, the approach is not applicable to proteins that have not been crystallized in the presence of a drug-like molecule, thus neglecting all the information provided by ions, solvent molecules or any other additive present in the binding site [17]. In this study, we propose to combine the multiple experimental binding modes into a single reference, in order to make the rescoring approach more robust, faster and applicable to a larger number of cases.

In this article, we describe the method, called Local Interaction Density or LID. LID and GRIM being based on the same representation of the binding mode, LID's performance in pose prediction and virtual screening are compared to that of GRIM.

The pose prediction was based on a high quality dataset [18]. Each protein was described by at least 20 3D structures containing diverse drug-like ligands. This allowed us to quantify

the minimum number of 3D structures needed to create a single useful reference. For three of the proteins, free protein 3D structures with crystallization additives in the binding site were available. We have therefore assessed whether crystallization additives alone are sufficient to create a useful reference.

Virtual screening was performed on eight target sets of the DUD-E dataset [19] meeting the LID requirements, i.e., those described by several experimental reference 3D structures. For one of the proteins, we also performed a virtual screening on a more challenging dataset which was created from the results of an experimental screening [20].

4.3 Results and discussion

4.3.1 Description of the LID Method

The LID method consists of two steps: (i) the creation of maps describing all the binding modes of a protein, and (ii) the calculation of a score for posing a ligand.

Creation of LID Maps

All protein 3D structures were superposed onto a single 3D structure whose binding site was representative of the structural ensemble (FIGURE 4.2A).

In each 3D structure, non-covalent interactions were detected using IChem [21]. We considered five interaction types: hydrogen bond (Hydrogen), ionic bond, metal chelation, π-stacking and hydrophobic contacts. In addition, IChem distinguished HB donor and HB acceptor subtypes, whether the donor was a ligand or a protein atom. Similarly, it distinguished cationic or anionic subtypes of ionic bond whether the cation was a ligand or a protein atom. In total, there were seven IChem interaction types. An interaction was represented by a triplet of interaction pseudo-atoms (IPA) (FIGURE 4.2B), the first was positioned on the ligand atom (IPA_{ligand}), the second on the protein atom ($IPA_{protein}$) and the third at mid-distance between the first two (IPA_{center}) (FIGURE 4.2B). Consequently, there were 21 pseudo-atom types (7 interaction types $\times$ 3 position tags).

The triplets of IPAs generated from all reference 3D structures were then fused (FIGURE 4.2C), and placed in a cubic grid with an edge length equal to 0.1 Å. The grid was built in the frame of the representative reference 3D structure. The grid boundaries were fixed by the two furthest IPAs which represented the diagonal and defined the center of voxels. To avoid edge effects, each voxel was assigned a density score which was the sum of the number of IPAs it contained and the number of IPAs contained in the adjacent voxel based on a Manhattan distance of 0.5 Å (FIGURE 4.2D). Each IPA type was considered separately, yielding 21 maps.

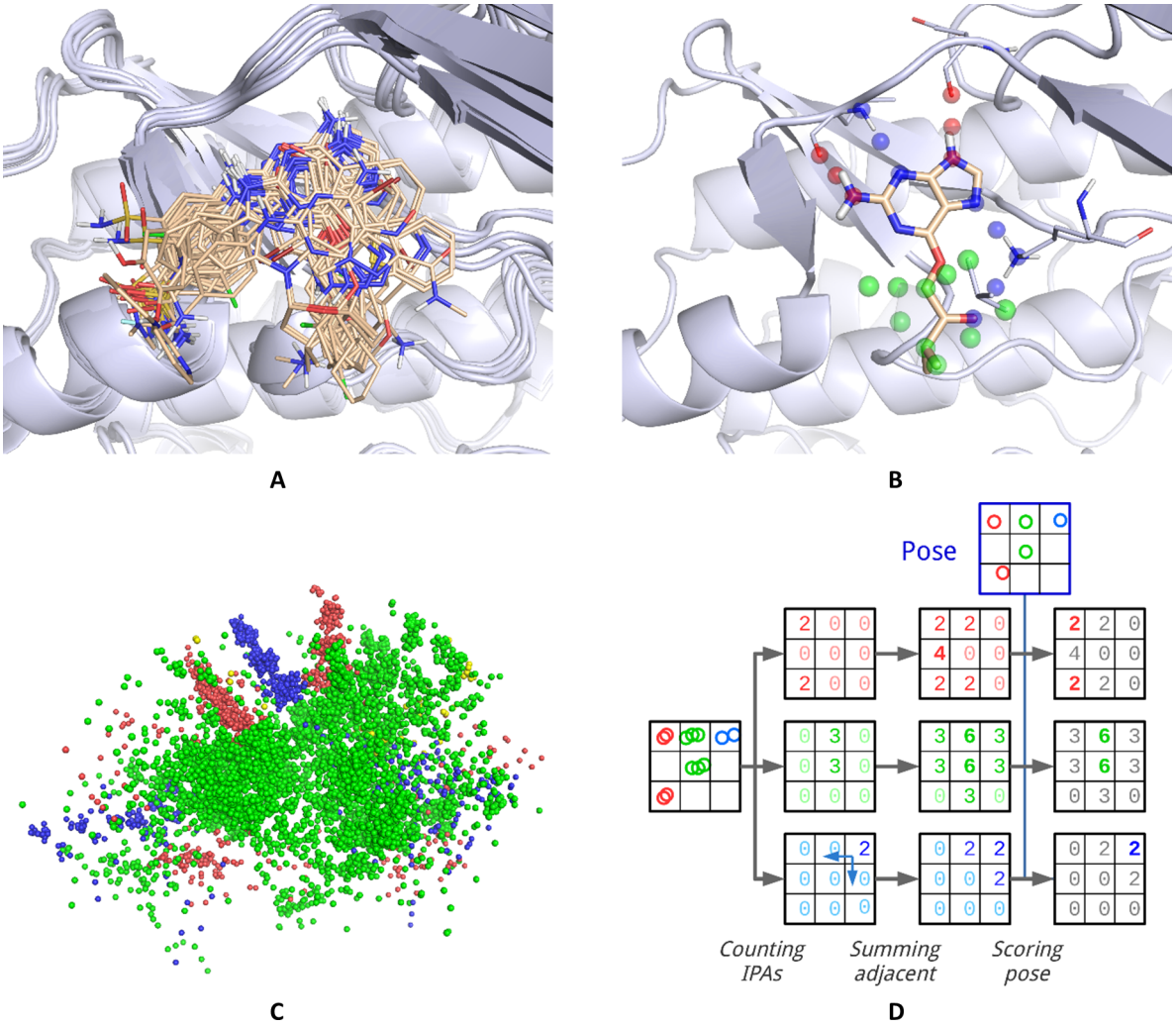

Figure 4.2 – LID method. A: superposition of the reference 3D structures onto the representative 3D structure of the protein binding site; example of CDK2 (PDB code: 2B53); B: detection of interactions between the protein and its ligand. Hydrogen bonds (blue and red interaction pseudo-atoms) and hydrophobic contacts (green interaction pseudo-atoms) between CDK2 and its ligand (PDB code: 1GZ8, HET code: MBP); C: ensemble of superimposed interaction pseudo-atoms; D: LID maps: generation from the merged interaction pseudo-atoms and used in rescoring. For the sake of clarity, only 3 of the 21 maps are represented. In A and B, proteins are represented by light grey ribbons and ligands by sticks (carbon atoms in pale orange). In A, B and C, interaction pseudo-atoms are represented by spheres, colored according to the corresponding bond type. The images in A, B and C show the same view.

Calculation of the LID Score

The LID score was obtained by comparing the docking pose IPAs with the LID maps (FIGURE 4.2D). For each IPA type, i.e., for each of the 21 LID maps, the density scores were summed across all voxels and the resulting sum was divided by the number of docking pose IPAs. The LID score was the sum of individual scores calculated from the 21 maps. A high LID score means that a high proportion of the docking pose interactions are found in the reference complexes and that the docking pose interactions are observed in a large number of reference complexes. The calculation of the LID score is formalized by the following equation:

$$\mathrm{LID}_{\mathrm{score}} = \sum_{i=1}^{N_{\mathrm{IPA}}} \frac{1}{N(M_i, T_i)} G(x_i, y_i, z_i, M_i, T_i) \tag{4.1}$$

where G represents the interaction grid, x_i, y_i, and z_i are the three cartesian coordinates of IPA_i, M_i is the IPA_i mode (i.e., $\mathrm{IPA}_{\mathrm{ligand}}$, $\mathrm{IPA}_{\mathrm{center}}$ or $\mathrm{IPA}_{\mathrm{protein}}$), T_i is the IPA_i interaction type (e.g., HB) and $N(M_i, T_i)$ is the number of IPAs with the same mode and type.

For comparison, the GRIM score quantifies the similarity between two IPA patterns, that of the docking pose and that of a reference structure. It was empirically determined by fitting six parameters to a shape-based similarity score on 1800 pairs of protein–ligand complexes (900 similar and 900 dissimilar) [13]. The six parameters take into account: the number of matched $\mathrm{IPA}_{\mathrm{ligand}}$; the number of matched $\mathrm{IPA}_{\mathrm{center}}$; the number of matched $\mathrm{IPA}_{\mathrm{protein}}$; the proportion of matched IPAs weighted by interaction type; the quality of the superimposition of matched IPAs; and the difference in the total number of IPAs between the docked pose and the reference pose.

The LID and GRIM scores are both additive scores. However, the LID score is obtained exclusively from positive contributions, i.e., elements common to the binding modes compared, while the GRIM score penalizes differences in geometry and the size of the compared binding modes. In addition, unlike the GRIM score, the LID score does not weight the contribution of interactions according to their types. Finally, the LID score, being determined by a set of 3D reference structures, is therefore customized for a particular site. On the contrary, the GRIM score was designed to be universal.

4.3.2 LID's Performance in Pose Prediction

Is LID able to recognize, among the docking poses selected by the scoring function, those that are close to the crystallographic structure of the ligand-protein complex? To answer this question, we sought to reproduce the crystallographic structures of 1382 ligand-protein complexes. These 3D structures have been carefully selected in the Protein Data Bank (PDB) to meet strict quality criteria (e.g., no mutation in the site, agreement between atomic coordinates and electron density), while being globally adapted to the docking approach (drug-like ligand, ligandable site).

The test dataset, called the LID dataset, describes 19 proteins (TABLE 4.3 in the supplementary material section). On average, a protein is described by about fifty different 3D structures. There are almost as many different ligands in complex with a protein as there are 3D structures of that protein (there can be more than one structure of the same complex). Four of the proteins are represented by more than 100 3D structures: cyclin-dependent kinase 2 (CDK2) with 156 ligands, carbonic anhydrase 2 (CAH2) with 155 ligands, beta-secretase 1 (BACE1) with 152 ligands, and heat shock protein 90-alpha (HSP90A) with 106 ligands. TABLE 4.1 gives the number of ligands per protein. Protein descriptors indicate whether the binding site is rather large or small, hydrophilic or hydrophobic, rigid or flexible. All cases are encountered in the dataset. In four proteins, the binding site contains a metal cation. The 19 proteins exhibit a variable proportion of interaction types with bound ligands (TABLE 4.4 in the supplementary material section).

The evaluation of the LID rescoring approach in pose prediction was performed as follows. The ligands' input 3D structures were generated from their SMILES codes. Up to 20 3D structures were obtained for the same SMILES code if several conformations were possible for a cyclic compound (e.g., cyclohexane in the chair or the boat conformation, substituent in the axial or the equatorial position). All the ligands of a protein were docked into the same site, using all 3D structures of the protein, except that of the self crystallographic complex. In other words, we did non-native or cross-docking. For each docking job, the 10 poses with the highest docking

TABLE 4.1 – LID dataset description: binding site (a), pose prediction (b) and virtual screening (c). Proteins are sorted by the number of 3D structures in descending order (HET codes). Site HYD is the relative hydrophobicity of the binding site (ratio of polar to apolar cavity describing atoms; see Material and methods). For each descriptor/count, the nine largest values are written in bold. Median LID RMSD is the median RMSD of docked poses after LID rescoring. TP% at 5% decoys is the true positives proportion, in percent, at a constant 5% decoys rate retrieval, in a virtual screening experiment using a single protein structure.

Uniprot ID	**HET codes**	**RMSD[(a)]** (Å)		**$N_{residues}$[(a)]**	**Site HYD[(a)]** (%)	**Metal ions[(a)]**	**Median LID RMSD[(b)]** (Å)	**TP% at 5% decoys[(c)]**		
		Cα	All					LID	GRIM	ChemPLP
CDK2_HUMAN	156	1.07	1.62	48	45.2		1.87	45.5	21.3	16.7
CAH2_HUMAN	155	0.20	0.49	37	27.4	Zn	1.91	85.2	85.7	5.08
BACE1_HUMAN	152	0.86	1.27	60	26.2		1.79	49.5	31.5	45.2
HS90A_HUMAN	106	1.45	1.77	48	42.8		2.70	15.9	23.9	4.55
PIM1_HUMAN	47	0.31	0.63	42	53.5		3.21			
TNKS2_HUMAN	39	0.96	1.62	45	36.4		0.55			
PK3CG_HUMAN	38	0.77	1.40	43	46.3		1.76			
PDE10_HUMAN	36	0.28	0.46	44	34.4	Mg,Zn	2.74			
PNMT_HUMAN	32	0.24	0.40	42	27.4		1.98			
ESR1_HUMAN	29	0.44	0.82	48	63.1		1.72	44.9	52.2	6.79
CHK1_HUMAN	29	0.48	0.57	46	50.4		3.24			
HYES_HUMAN	27	0.61	1.01	64	43.6		7.32			
PDPK1_HUMAN	27	0.41	0.94	49	47.7		1.70			
BRD4_HUMAN	26	0.34	0.63	25	51.8		3.23			
ALDR_HUMAN	23	0.33	0.56	36	16.9		1.61	44.0	41.5	37.1
GRIA2_RAT	23	1.13	1.27	51	36.1		1.7	6.33	46.8	18.4
LKHA4_HUMAN	21	0.15	0.29	57	24.6	Zn	3.83	73.1	71.4	73.7
MMP12_HUMAN	21	0.23	0.56	35	36.3	Zn	3.02			
TGT_ZYMMO	20	0.61	0.89	45	37.2		0.74			

scores were evaluated with LID. Up to 45,200 poses were compared for the same complex. The docking poses were also evaluated with GRIM.

The docking was carried out with the PLANTS software, which has the advantage of being freely distributed, and which, in our hands, reproduces the diversity of poses obtained with the GOLD program (using "enable the generate diverse solutions" and disabling "allow early termination") [22]. Tested on the LID dataset, PLANTS found a pose close to the crystallographic structure in 1313 out of the 1382 complexes (FIGURE 4.2A, see "Max" labelled bar). Here, we evaluated the similarity between the docked pose and the crystallographic structure using the

RMSD calculated on the non-hydrogen atoms of the ligand, and we considered that the docked pose is correct whether the RMSD is below 2 Å. The PLANTS default scoring function, namely ChemPLP, placed a correct pose in the first position in only 42 % of the cases (FIGURE 4.2A, see "ChemPLP" labelled bar), while GRIM did so in 60 % (FIGURE 4.2A, see "GRIM" labelled bar). LID score performed almost as well as GRIM, placing at the top position a correct pose in 54 % of cases (FIGURE 4.2A, see "LID" labelled bar). In particular, LID is less effective than GRIM in the selection of ligand poses for four proteins: the phosphodiesterase 10A (PDE10), the protein kinase Chk1 (CHK1), the epoxide hydrolase 2 (HYES) and the leukotriene A-4 hydrolase (LKHA4) (TABLE 4.5 in the supplementary material section).

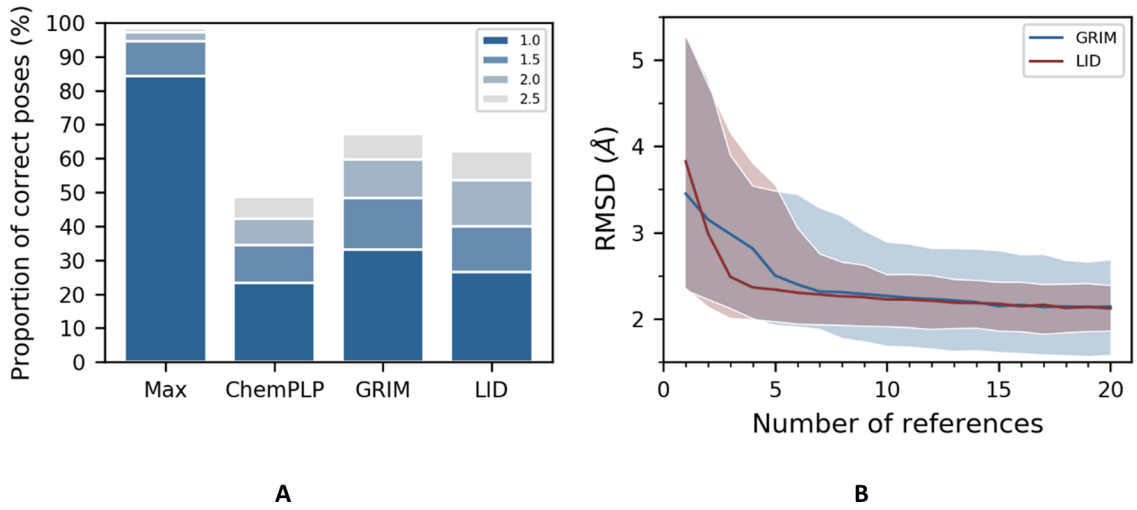

Figure 4.3 – LID's performance in pose prediction. (a) comparison with GRIM and ChemPLP. The legend indicates the threshold in Å under which poses are considered as correct. "Max" represents the proportion of correct poses according to the best possible RMSD criterion; (b) as a function of the number of reference 3D structures used to build the map. The distributions of median RMSDs obtained for the 19 proteins in the dataset using LID and GRIM are shown in red and blue, respectively. A line is drawn at the median value calculated on the 19 proteins, and the colored area delimits the first and ninth deciles of the distribution.

4.3.3 LID's Advantages and Limitations

As mentioned above, GRIM and LID are both based on IChem tools for the detection of ligand-protein interactions, and their encoding in IPAs. They, however, differ in two aspects: firstly, GRIM comparison of binding modes is based on clique detection and therefore does not require the pre-alignment of the compared 3D structures, unlike LID; secondly, GRIM rescoring

considers the reference 3D structures individually, while LID considers the reference 3D structures as a whole. As a consequence, we expect LID to be less sensitive than GRIM to the number of reference 3D structures, but more sensitive to structural variations in protein coordinates.

We assessed the extent to which GRIM and LID depended on the number of reference 3D structures by repeating the rescoring with a growing number of reference 3D structures. For each protein in the dataset, we tested all combinations of n reference 3D structures, with n ranging from 1 to 20 (the maximal number of tested combinations is equal to 1000). We thus generated up to 1000×21 new LID maps per protein. GRIM and LID best performance was observed for 10 or more reference 3D structures (FIGURE 4.2B). The decrease of the median RMSD however differed between GRIM and LID. GRIM curve shows a regular downward slope while LID curve shows a steep initial slope that becomes more gradual from three reference 3D structures on (FIGURE 4.2B). This suggests that the binding modes of three randomly chosen ligands may provide sufficient information to guide docking using LID. Moreover, from seven reference 3D structures on, LID achieves its best performance level for the majority of the tested cases. For comparison, GRIM showed larger deviations, even considering a larger number of reference 3D structures.

4.3.4 Application of LID to "apo" Proteins

LID and GRIM have been designed for already well characterized protein structures, for which binding mode information is available for at least one ligand. Is it possible to use LID for a protein whose structure has been resolved at the atomic level, in the absence of drug-like ligands but in the presence of diverse crystallization additives? To answer this question, we searched the LID dataset for proteins that are represented in the PDB in the form "apo" (i.e., in the absence of drug-like ligand) but whose site is not empty. Three proteins have at least three different additives in their binding site: CAH2, macrophage metalloelastase (MMP12) and glutamate receptor (GRIA4) (FIGURE 4.4A). Glycerol, sulfate, acetate, carbon dioxide, bicarbonate or cyanic acid are found in CAH2 (carbon dioxide and bicarbonate are ligands of the enzyme that have a key role in regulating cell pH). Acetic and acetohydroxamic acids and an azide ion are found in

MMP12. Sulphate, ethanediol and morpholinoethanesulphonic acid are found in GRIA2. Both CAH2 and MMP2 are metalloproteins, and in the 3D structures of the two proteins, we observed at least one anionic additive being a coordinating ligand of the metal cation (see the grey IPAs on FIGURE 4.4A). New LID maps were built using only the interactions detected between proteins and additives (note that in the case of CAH2 and MMP12, the three LID maps corresponding to ligand–metal interactions contain information). They yielded the identification of a correct pose for nearly half of the 155 CAH2 ligands (FIGURE 4.4B). For comparison, ChemPLP is three times less efficient. Opposite results were observed for GRIA2 and MMP12. We suspected the approach not to be suitable for large ligands, and therefore did the analysis again for fragments only (i.e., drug-like ligand with MW $\leq$ 300, number of non-hydrogen atoms $\leq$ 18). We confirmed that additives effectively helped to predict ligand placement in CAH2 and MMP12, but not in GRIA2.

Is it possible to predict whether the information provided by the additives alone is relevant for rescoring with LID? Distribution of IPAs suggested a positive answer. In the three study cases, additives revealed a motif of directional interactions, which was conserved in the ligands' binding modes (FIGURE 4.4A). However, the amino acids involved in these interactions were mostly rigid in CAH2 and MMP12 (all atom RMSD $\approx$ 0.5 Å) while they adopted different conformations in GRIA2 (all atom RMSD $\approx$ 1.7 Å). In summary, the LID approach using additives can be considered for fragment docking if the protein site is rigid. Since flexibility is not necessarily revealed by the 3D structures of "apo" proteins, it is nevertheless advisable to consider crystallographic structural factors or to perform a molecular dynamics simulation [23, 24].

4.3.5 LID's Performance in Virtual Screening

It is observed that ligand pose prediction performance by molecular docking was already improved with LID. At this point, another question arises as to whether LID is also capable of efficiently discriminating between true active compounds of a given protein and their chemically similar decoys or not. To answer this question, a retrospective virtual screening challenge using a set of DUD-E targets was carried out, with LID employed as docking-assistant tool.

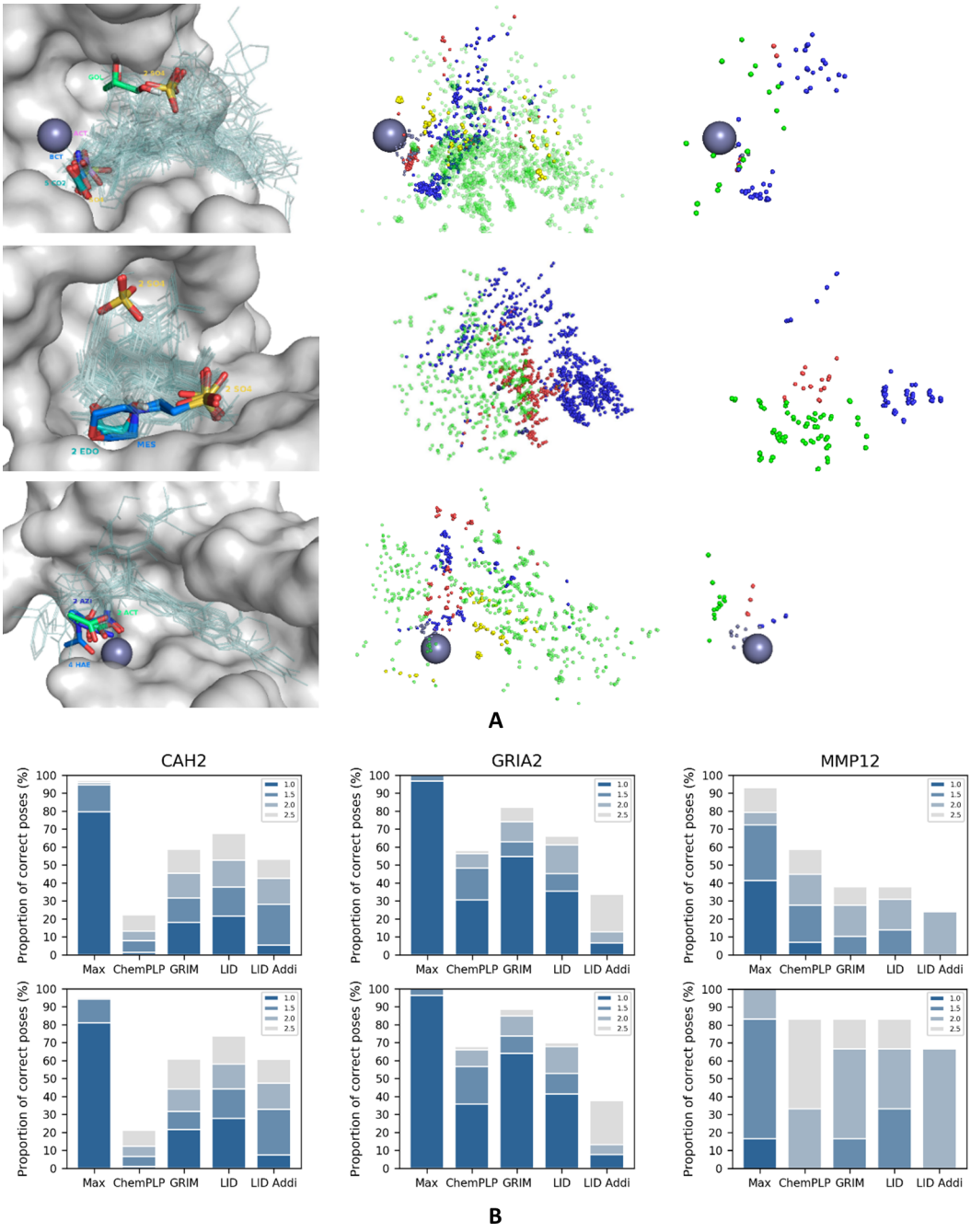

Figure 4.4 – Use of crystallization additives binding modes in LID. (a) additives crystallized in CAH2 (top), GRIA2 (middle) and MMP12 (bottom). Metal cation is represented with a grey sphere. On the left triad are shown the additives (thick sticks colored by HET code) and drug-like ligands (transparent lines) in the protein site (grey surface); on the middle triad are shown the drug-like ligands interaction pseudo-atoms, colored according to the corresponding bond; on the right triad are shown the additives interaction pseudo-atoms, colored according to the corresponding bond type (hydrogen bond in red and blue, π-stacking in yellow, hydrophobic contact in green and metal chelation in grey); (b) LID's performance in pose prediction of drug-like ligand (top) and fragment only (bottom). LID and LID Addi refer to the use of drug-like ligands and additives as reference, respectively.

Only eight protein targets included in both the DUD-E and the LID datasets as described above were investigated. These include: the aldo-keto reductase (ALDR), BACE1, CAH2, CDK2, the estrogen receptor alpha (ESR1), GRIA2, HSP90A, and LKHA4 (TABLE 4.1, TABLE 4.3 in the supplementary material section). A rigid docking protocol using a unique representative 3D structure for each target's binding site was implemented. True actives and decoys were already prepared by the DUD-E contributors and used as such. Each ligand corresponded to a maximum of eight stereoisomers, each of which issued 10 post-docking poses that were kept and analyzed.

LID's performance was evaluated according to the Receiver Operating Characteristic (ROC) curves that were obtained (FIGURE 4.5). It is observed that LID generally gave better performances than ChemPLP, with a mean area under the ROC curves (ROC AUCs) of 0.78, compared to an average value of 0.67 issued from the latter method (TABLE 4.6 in the supplementary material section). The enrichment factors were also significantly improved except for GRIA2 (TABLE 4.1). On the whole, the improvement brought by LID was quantitatively comparable to that by GRIM. However, the rescoring with LID took up a remarkably shorter amount of time (TABLE 4.2). This observation became more obvious when a docking process using multiple structures of a single protein was carried out. To be more specific, 10 structures were used as input for each of the two most flexible protein targets HS90A and GRIA2, with the aim of taking into account the most diverse structures of the ligand-protein binding site. A docking-based virtual screening process was conducted for each structure. The average recorded calculation time for GRIM rescoring was approximately 58 h per protein structure, while that with LID was only 10 min (noteworthy, GRIM was coded in C++, while LID was coded with Python). Nevertheless, the time required to align all protein structures for LID to function properly has to be considered. In terms of overall performances, the multitarget approach using LID was observed to give notably better results in HS90A and GRIA2, with an improvement of 9.30 and 19.0 in true positive percent at 5% decoys, respectively (TABLE 4.7 in the supplementary material section).

The DUD-E datasets have long been employed for a benchmarking of novel structure-based methods in computer-aided drug design. However, the real value of these datasets has been subject to much debate, primarily due to serious drawbacks in compound selection, e.g.,

TABLE 4.2 – Elapsed time for pose rescoring of DUD-E dataset with GRIM and LID with a single protein structure. The number of references may differ from the TABLE 4.1 because multiple copies are considered.

Uniprot ID	**Poses**	**References**	**Elapsed time** (min)	
			GRIM	LID
ALDR_HUMAN	93,500	41	220	5
BACE1_HUMAN	187,060	227	3000	13
CAH2_HUMAN	325,450	156	2200	13
CDK2_HUMAN	291,260	195	2500	13
ESR1_HUMAN	214,450	37	720	17
GRIA2_HUMAN	123,580	62	260	4
HS90A_HUMAN	50,670	139	590	4
LKHA4_HUMAN	97,210	26	180	7

the structural biases that led to artificial enrichment and an overestimation of virtual screening performance, or the fact that the potency of the decoys always remained unknown [25]. We therefore put LID into another retrospective virtual screening challenge using a dataset whose active and inactive compounds along with their potency were already verified by confirmatory dose-response biological assays. This dataset, ESR1, was prepared from the results of a high-throughput screening of small molecule antagonists of the estrogen receptor alpha that can be accessed on the website of PubChem BioAssays (for more details concerning the compound selection, see [20]). The dataset employed in this study comprises 1589 compounds, 59 of which are actives with potency values ranging from 3.9 nM to 9.6984 µM; the rest (1530 compounds) were confirmed as inactives.

As already observed with DUD-E, the ChemPLP scoring function gave an extremely poor performance in the screening challenge with ESR1. LID and GRIM did not give good performances like in the cases of DUD-E datasets (the two approaches had more difficulty in distinguishing between true actives and true inactives than in separating true actives from their chemically similar decoys); however, both methods were capable of selecting true actives among the early

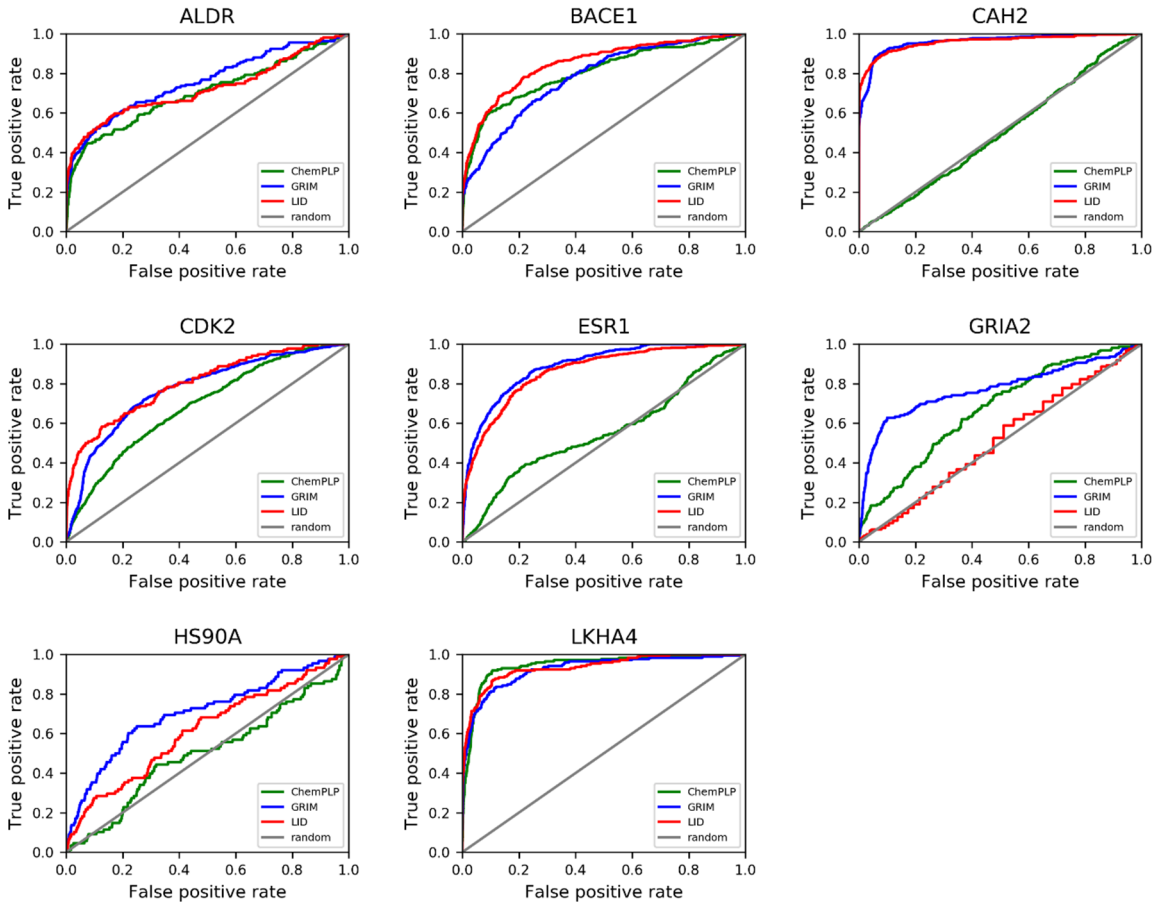

Figure 4.5 – LID's performances in retrospective virtual screening. All of the ROC curves show screening results with DUD-E. LID's performances are compared to those of GRIM and ChemPLP.

hits, with 18.6 % and 16.9 % of active compounds retrieved at a constant false positive rate of 5%, respectively (FIGURE 4.6B).

4.3.6 LID Cost in Calculation Time

The LID method has been designed to process a very large number of poses, typically in a virtual screening of large libraries using multiple 3D structures of the target protein. The tests performed as part of this study allowed us to estimate the relative time required for such an application (TABLE 4.2, TABLE 4.8 in the supplementary material section). The generation of LID maps, whose quantity is proportional to the number of reference 3D structures (N_{ref}), was made once. The generation time on a desktop computer was about $0.18 \times N_{\text{ref}}$ ($R^2 = 0.8$). LID score calculation lasted ≈ 0.003 s per pose ($R^2 = 0.6$) with an Intel® Xeon E3-1240 (3.70 GHz)

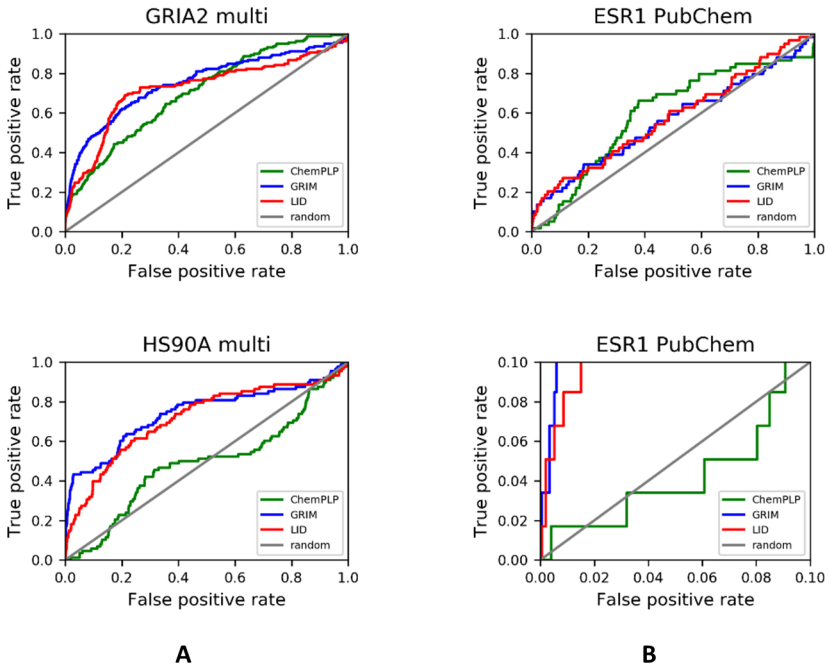

Figure 4.6 – LID's performances in retrospective virtual screening using 10 protein structures. (a) virtual screening with the DUD-E dataset; (b) virtual screening with the cleaned PubChem Bioassay dataset ESR1 (AID 743080). The bottom is an enlargement of the full ROC curve. LID's performances are compared to those of GRIM and ChemPLP.

processor. LID score calculation better correlated with the number of compared IPAs ($R^2 \geq 0.95$), yet this information is not known before rescoring. Nevertheless, the first approximation is valid since rescoring total time with LID is negligible compared to docking time. For the sake of comparison, GRIM rescoring time was longer and not proportional to the number of docked poses or reference 3D structures. Due to its clique detection step, GRIM calculation time highly depended on the number, types and distribution of IPAs (TABLE 4.9 in the supplementary material section). The difference between LID and GRIM calculation times in seconds was up to several orders of magnitude.

4.3.7 Comparison with Other Methods Using Aligned 3D Structures

The first strategies for using crystallographic structures of ligand-protein complexes in docking methods are almost 20 years old [26, 27]. Different ways of incorporating structural information have been proposed, for example by introducing spatial constraints targeting parts of the ligand that are common with a reference ligand, by guiding the placement of the ligand on

reference atom-centered functions or electron density, or by scoring pose by 3D pharmacophore matching similarity [28, 29, 30]. OKUNO et al. suggested to combine multiple ligand-protein crystallographic structure into a reference grid [31, 32]. The method, called VS-APPLE, aligned the ligand with the reference grid to predict the binding mode. Designed as a screening method, it was able to discriminate active compounds from decoys of the DUD dataset [33]. The LID and VS-APPLE methods are similar. However, LID is intended exclusively for scoring purposes and is not linked to a particular docking program. LID especially allows to take into account induced fit using a docking program that treats the protein as a rigid body. The method is indeed fast and efficient enough to sort the multiple poses of a ligand docked into an ensemble of 3D structures of its protein site. LID's performance assessment also indicated that the method can be successful using a limited amount of reference information. We estimated that less than ten different ligands are required for effective scoring. This is in line with our recent analysis of binding modes in PDB, which concluded that nine ligands achieve the coverage of interactions formed with the protein pocket [18]. Similarly, VS-APPLE good performance was also reached with an incomplete reference grid. The minimum information required on the 13 studied proteins of the DUD dataset was estimated at 30 percent of the total atomic coordinates.

4.4 Materials and methods

4.4.1 Dataset Preparation

Pose Prediction Challenge

The LID dataset was prepared as previously described [18]. The 3D structures selected from the PDB are all in high resolution and are completely described. There is no mutation in the binding site. Drug-like ligand complies to the Lipinski's rule of 5, with a maximum of one exception. The molecules of the LID dataset were protonated with the program Protoss v2.0 (University of Hamburg, Hamburg, Germany) [34]. Importantly, the cofactors and water molecules were removed from the protein structures.

Retrospective Virtual Screening Challenge

Eight protein targets common to the DUD-E and the LID datasets, namely ALDR, BACE1, CAH2, CDK2, ESR1, GRIA2, HSP90A, LKHA4, were used as input and processed as described above. The representative protein structure was determined by fitting and computing RMSD matrix on Cα carbons with the command cealign in PyMOL (The PyMOL Molecular Graphics System, Version 1.8.6.1, Schrödinger, LLC , New York, NY, USA). The structure with the minimal average value was selected. For the selection of 10 structures, an agglomerative hierarchical clustering of all structures was carried out with the scikit-learn v0.19.1 Python package, then the most representative of each cluster was kept. Another dataset, ESR1, comprising true actives and true inactives confirmed by a PubChem quantitative high-throughput screening assay of small molecule antagonists of the estrogen receptor alpha was directly employed after preparation steps described by Tran-Nguyen et al. [20].

Properties of Binding Sites

A binding site is defined as the consensus residues near the bound drug-like ligands. In each 3D structure, the protein residues with at least one non-hydrogen atom closer than 6.5 Å to any ligand non-hydrogen atom were identified. The binding site was defined as the ensemble of residues present in more than 10 % of the structures [18]. The RMSD values were computed either on Cα or all non-hydrogen atoms by fitting a structure to a reference (described in LID score subsection) with the command cealign in PyMOL v1.8.6.1. Cavity volumes and the percentage of hydrophobic interactions were determined with the Volsite module in IChem v5.2.9 (UMR7200 CNRS-University of Strasbourg, Illkirch, France).

4.4.2 Docking

Ligand Preparation

Each ligand in the SMILES file format was ionized with the program Filter of OpenEye (Filter 2.5.1.4 OpenEye Scientific Software, Santa Fe, NM, USA). The 3D structure was built with the program Corina 3.40 (Molecular Networks GmbH, Nürnberg, Germany) [35]. The option rc was used to generate multiple conformations of the rings (with a maximum of 100 stereoisomers per molecule). For CAH2 ligands with sulfonamide groups, the nitrogens linked to a sulfur atom were deprotonated (-1 charge).

Docking with PLANTS

A rigid docking procedure was performed with the PLANTS program v1.2 using the ChemPLP scoring function and the search speed set at 1 (highest accuracy) [36].

PLANTS is based on an ant colony algorithm to optimize the placement and the conformation of ligand as well as the positions of the protein hydrogen atoms that form hydrogen bonds with the ligand. PLANTS explores possible torsion angle values of the ligand but does not

modify the conformation of rings. In the case of metalloproteins, PLANTS considers geometry parameters related to metal atoms.

The cavity center of a protein site was defined from the centroid of all the ligands bound to this protein. The cavity radius was set as the maximum distance between the cavity center and the atoms of all the ligand crystallized in the binding site, plus 2 Å. On average, the radius was equal to 12 Å. Ten poses were saved per docking run.

4.4.3 Post-Processing of Docking

GRIM score

The GRIM score was computed with the GRIM module of the IChem program, v5.2.9, all options were kept as default [21].

LID score

The representative 3D structure of a protein was defined after multiple comparisons of possible binding sites using the Shaper v1.0 program (minimal average distance of the matrix) (UMR7200 CNRS-University of Strasbourg, Illkirch, France) [37]. Reference structures were aligned onto the representative with the CE v2003.03.13 program (RCSB Protein Data Bank, San Diego, CA, USA) [38]. The IPAs were obtained from the aligned reference structures with the ints module of IChem v5.2.9, the maximal distance to detect π–π stacking interactions was fixed at 5.0 Å (-D_Ar 5.0). All IPAs were merged into a mol2 file and were considered as a single pseudomolecule. This file was used for the creation of the LID maps, using the in-house intgrid program, v1.0. The LID program (v1.0) issued a LID score for each docking pose.

4.5 Conclusion

We propose LID, a novel docking-assistant tool that evaluates the relevance of protein–ligand binding modes by a comparison with those obtained from an accumulation of reference 3D structures. LID ameliorated pose prediction performance using ligands of 19 diverse proteins, and improved early enrichment of true active compounds for eight protein targets. LID's performance was comparable to that of GRIM, a former approach based on interaction pattern graph similarity. It is noteworthy that LID was faster and more robust than GRIM if few 3D structures were available to describe a particular target. LID was in particular more adapted to high-throughput applications, e.g., high-throughput virtual screening using multiple structures of the same protein. In addition, LID allowed a consideration of binding modes of additives used during the structure crystallization process.

4.6 References

[1] Célien Jacquemard, Viet-Khoa Tran-Nguyen, Malgorzata N. Drwal, Didier Rognan, and Esther Kellenberger. Local Interaction Density (LID), a Fast and Efficient Tool to Prioritize Docking Poses. *Molecules*, 24(14):2610, Janvier 2019.

[2] Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2):269–288, Octobre 1982.

[3] Peter Ripphausen, Britta Nisius, Lisa Peltason, and Jürgen Bajorath. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. *Journal of Medicinal Chemistry*, 53(24):8461–8467, Décembre 2010.

[4] Natalie J. Tatum, Fernanda Duarte, Shina C. L. Kamerlin, and Ehmke Pohl. Relative Binding Energies Predict Crystallographic Binding Modes of Ethionamide Booster Lead Compounds. *The Journal of Physical Chemistry Letters*, 10(9):2244–2249, Mai 2019.

[5] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18):12964–12975, Mai 2016.

[6] Ludovic Chaput and Liliane Mouawad. Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. *Journal of Cheminformatics*, 9(1):37, Juin 2017.

[7] James B. Dunbar, Richard D. Smith, Kelly L. Damm-Ganamet, Aqeel Ahmed, Emilio Xavier Esposito, James Delproposto, Krishnapriya Chinnaswamy, You-Na Kang, Ginger Kubish, Jason E. Gestwicki, Jeanne A. Stuckey, and Heather A. Carlson. CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *Journal of Chemical Information and Modeling*, 53(8):1842–1852, Août 2013.

[8] Zied Gaieb, Shuai Liu, Symon Gathiaka, Michael Chiu, Huanwang Yang, Chenghua Shao, Victoria A. Feher, W. Patrick Walters, Bernd Kuhn, Markus G. Rudolph, Stephen K. Burley, Michael K. Gilson, and Rommie E. Amaro. D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of Computer-Aided Molecular Design*, 32(1):1–20, Janvier 2018.

[9] Isabella A. Guedes, Felipe S. S. Pereira, and Laurent E. Dardenne. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Frontiers in Pharmacology*, 9, 2018.

[10] Zhan Deng, Claudio Chuaqui, and Juswinder Singh. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *Journal of Medicinal Chemistry*, 47(2):337–344, Janvier 2004.

[11] Anita Rácz, Dávid Bajusz, and Károly Héberger. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *Journal of Cheminformatics*, 10(1):48, Octobre 2018.

[12] Márton Vass, Albert J Kooistra, Tina Ritschel, Rob Leurs, Iwan JP de Esch, and Chris de Graaf. Molecular interaction fingerprint approaches for GPCR drug discovery. *Current Opinion in Pharmacology*, 30:59–68, Octobre 2016.

[13] Jérémy Desaphy, Eric Raimbaud, Pierre Ducrot, and Didier Rognan. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. *Journal of Chemical Information and Modeling*, 53(3):623–637, Mars 2013.

[14] Célien Jacquemard, Malgorzata N. Drwal, Jérémy Desaphy, and Esther Kellenberger. Binding mode information improves fragment docking. *Journal of Cheminformatics*, 11 (1):24, Mars 2019.

[15] Inna Slynko, Franck Da Silva, Guillaume Bret, and Didier Rognan. Docking pose selection by interaction pattern graph similarity: application to the D3R grand challenge 2015. *Journal of Computer-Aided Molecular Design*, 30(9):669–683, Septembre 2016.

[16] Priscila da Silva Figueiredo Celestino Gomes, Franck Da Silva, Guillaume Bret, and Didier Rognan. Ranking docking poses by graph matching of protein–ligand interactions: lessons learned from the D3R Grand Challenge 2. *Journal of Computer-Aided Molecular Design*, 32(1):75–87, Janvier 2018.

[17] Malgorzata N. Drwal, Célien Jacquemard, Carlos Perez, Jérémy Desaphy, and Esther Kellenberger. Do Fragments and Crystallization Additives Bind Similarly to Drug-like Ligands? *Journal of Chemical Information and Modeling*, 57(5):1197–1209, Mai 2017.

[18] Malgorzata N. Drwal, Guillaume Bret, Carlos Perez, Célien Jacquemard, Jérémy Desaphy, and Esther Kellenberger. Structural Insights on Fragment Binding Mode Conservation. *Journal of Medicinal Chemistry*, 61(14):5963–5973, Juillet 2018.

[19] Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, Juillet 2012.

[20] Viet-Khoa Tran-Nguyen, Franck Da Silva, Guillaume Bret, and Didier Rognan. All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception, and Virtual Screening. *Journal of Chemical Information and Modeling*, 59(1):573–585, Janvier 2019.

[21] Franck Da Silva, Jeremy Desaphy, and Didier Rognan. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. *ChemMedChem*, 13 (6):507–510, 2018.

[22] Marcel L. Verdonk, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, and Richard D. Taylor. Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 2003.

[23] Jordan J. Clark, Mark L. Benson, Richard D. Smith, and Heather A. Carlson. Inherent versus induced protein flexibility: Comparisons within and between apo and holo structures. *PLOS Computational Biology*, 15(1):e1006705, Janvier 2019.

[24] Cen Gao, Jeremy Desaphy, and Michal Vieth. Are induced fit protein conformational changes caused by ligand-binding predictable? A molecular dynamics investigation. *Journal of Computational Chemistry*, 38(15):1229–1237, 2017.

[25] Manon Réau, Florent Langenfeld, Jean-François Zagury, Nathalie Lagarde, and Matthieu Montes. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Frontiers in Pharmacology*, 9, 2018.

[26] Guosheng Wu and Michal Vieth. SDOCKER: A Method Utilizing Existing X-ray Structures To Improve Docking Accuracy. *Journal of Medicinal Chemistry*, 47(12):3142–3148, Juin 2004.

[27] Xavier Fradera, Ronald M. A. Knegtel, and Jordi Mestres. Similarity-driven flexible ligand docking. *Proteins: Structure, Function, and Bioinformatics*, 40(4):623–636, 2000.

[28] Cen Gao, Nels Thorsteinson, Ian Watson, Jibo Wang, and Michal Vieth. Knowledge-Based Strategy to Improve Ligand Pose Prediction Accuracy for Lead Optimization. *Journal of Chemical Information and Modeling*, 55(7):1460–1468, Juillet 2015.

[29] Brian P. Kelley, Scott P. Brown, Gregory L. Warren, and Steven W. Muchmore. POSIT: Flexible Shape-Guided Docking For Pose Prediction. *Journal of Chemical Information and Modeling*, 55(8):1771–1780, Août 2015.

[30] Lingling Jiang and Robert C. Rizzo. Pharmacophore-Based Similarity Scoring for DOCK. *The Journal of Physical Chemistry B*, 119(3):1083–1102, Janvier 2015.

[31] Tatsuya Okuno, Koya Kato, Tomoki P. Terada, Masaki Sasai, and George Chikenji. VS-APPLE: A Virtual Screening Algorithm Using Promiscuous Protein–Ligand Complexes. *Journal of Chemical Information and Modeling*, 55(6):1108–1119, Juin 2015.

[32] Tatsuya Okuno, Koya Kato, Shintaro Minami, Tomoki P. Terada, Masaki Sasai, and George Chikenji. Importance of consensus region of multiple-ligand templates in a virtual screening method. *Biophysics and Physicobiology*, 13:149–156, 2016.

[33] Niu Huang, Brian K. Shoichet, and John J. Irwin. Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801, Novembre 2006.

[34] Stefan Bietz, Sascha Urbaczek, Benjamin Schulz, and Matthias Rarey. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *Journal of Cheminformatics*, 6(1):12, Avril 2014.

[35] Jens Sadowski, Johann Gasteiger, and Gerhard Klebe. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *Journal of Chemical Information and Computer Sciences*, 34(4):1000–1008, Juillet 1994.

[36] Oliver Korb, Thomas Stützle, and Thomas E. Exner. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *Journal of Chemical Information and Modeling*, 49(1):84–96, Janvier 2009.

[37] Jérémy Desaphy, Karima Azdimousa, Esther Kellenberger, and Didier Rognan. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *Journal of Chemical Information and Modeling*, 52(8):2287–2299, Août 2012.

[38] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering, Design and Selection*, 11 (9):739–747, Septembre 1998.

4.7 Supplementary material

TABLE 4.3 – List of proteins used in pose prediction challenge.

Uniprot ID	**Name**	**Uniprot AC**	**Protein class**
CDK2_HUMAN	Cyclin-dependent kinase 2	P24941	Transferase
CAH2_HUMAN	Carbonic anhydrase 2	P00918	Lyase
BACE1_HUMAN	Beta-secretase 1	P56817	Hydrolase
HS90A_HUMAN	Heat shock protein HSP 90-alpha	P07900	Chaperone
PIM1_HUMAN	Serine/threonine-protein kinase pim-1	P11309	Transferase
TNKS2_HUMAN	Poly [ADP-ribose] polymerase tankyrase-2	Q9H2K2	Transferase
PK3CG_HUMAN	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunitgamma isoform	P48736	Transferase
PDE10_HUMAN	cAMP and cAMP-inhibited cGMP 3',5'-cyclic phosphodiesterase 10A	Q9Y233	Hydrolase
PNMT_HUMAN	Phenylethanolamine N-methyltransferase	P11086	Transferase
ESR1_HUMAN	Estrogen receptor	P03372	Activator
CHK1_HUMAN	Serine/threonine-protein kinase Chk1	O14757	Transferase
HYES_HUMAN	Bifunctional epoxide hydrolase 2	P34913	Hydrolase
PDPK1_HUMAN	3-phosphoinositide-dependent protein kinase 1	O15530	Transferase
BRD4_HUMAN	Bromodomain-containing protein 4	O60885	Chromatinregulator
ALDR_HUMAN	Aldo-keto reductase family 1 member B1	P15121	Oxidoreductase
GRIA2_RAT	Glutamate receptor 2	P19491	Ion channel
LKHA4_HUMAN	Leukotriene A-4 hydrolase	P09960	Hydrolase
MMP12_HUMAN	Macrophage metalloelastase	P39900	Hydrolase
TGT_ZYMMO	Queuine tRNA-ribosyltransferase	P28720	Transferase

TABLE 4.4 – Proportion of interaction points generated from all structures describing the 19 proteins. (a) From a protein atom point of view. HYD: hydrophobic contact; HBA: hydrogen-bond, acceptor on the protein; HBD: hydrogen-bond, donor on the protein.

Uniprot ID	HYD	π-stacking	HBA[(a)]	HBD[(a)]	Ionic +[(a)]	Ionic -[(a)]	Metal	Number of points
CDK2_HUMAN	72.54	0.62	13.5	12.28	0.51	0.56	0	7291
CAH2_HUMAN	64.26	2.77	8.1	14.98	0	0	9.88	4038
BACE1_HUMAN	68.11	1.41	18.17	7.34	0	4.96	0	12,255
HS90A_HUMAN	82.38	3.22	8.6	5.4	0.26	0.14	0	5720
PIM1_HUMAN	82.11	1.4	5.64	6.88	0.98	3	0	1934
TNKS2_HUMAN	65.86	10.07	8.7	15.37	0	0	0	3207
PK3CG_HUMAN	82.12	0	5.6	12.27	0	0	0	1695
PDE10_HUMAN	79.13	10.22	1.8	8.84	0	0	0	1663
PNMT_HUMAN	66.32	9.74	13.75	1.83	0.52	7.85	0	1746
ESR1_HUMAN	84.95	0.27	9.33	4.53	0	0.93	0	2252
CHK1_HUMAN	72.55	0.58	12.38	10.65	0.77	3.07	0	1042
HYES_HUMAN	77.05	6.81	7.65	7.56	0	0.93	0	1072
PDPK1_HUMAN	70.18	0	11.65	16.28	0	1.89	0	1425
BRD4_HUMAN	85.02	0.56	4.4	10.02	0	0	0	1068
ALDR_HUMAN	76.77	5.93	0.14	17.16	0	0	0	2075
GRIA2_RAT	25.36	1.36	15.64	38.49	10.17	8.97	0	2429
LKHA4_HUMAN	63.81	9.11	8.87	8.4	1.02	4.79	4	1274
MMP12_HUMAN	60.98	8.4	7.29	14.67	0	0	8.65	1179
TGT_ZYMMO	41.63	7.91	32.68	16.63	0	1.15	0	872

TABLE 4.5 – Performance of pose prediction for the 19 targets.

Uniprot ID	**Uniprot AC**	**Poses median RMSD** (Å)		
		ChemPLP	GRIM	LID
CDK2_HUMAN	P24941	2.35	1.56	1.87
CAH2_HUMAN	P00918	3.27	2.17	1.91
BACE1_HUMAN	P56817	1.76	1.41	1.79
HS90A_HUMAN	P07900	7.23	2.12	2.7
PIM1_HUMAN	P11309	4.08	3.12	3.21
TNKS2_HUMAN	Q9H2K2	0.59	0.55	0.55
PK3CG_HUMAN	P48736	1.85	1.48	1.76
PDE10_HUMAN	Q9Y233	3.97	1.35	2.74
PNMT_HUMAN	P11086	2.15	1.23	1.98
ESR1_HUMAN	P03372	0.87	1.64	1.72
CHK1_HUMAN	O14757	3.67	1.62	3.24
HYES_HUMAN	P34913	7.51	2.45	7.32
PDPK1_HUMAN	O15530	2.37	1.94	1.7
BRD4_HUMAN	O60885	4.77	2.62	3.23
ALDR_HUMAN	P15121	0.94	0.76	1.61
GRIA2_RAT	P19491	1.67	0.86	1.7
LKHA4_HUMAN	P09960	3.09	1.51	3.83
MMP12_HUMAN	P39900	2.16	2.86	3.02
TGT_ZYMMO	P28720	0.66	0.58	0.74

TABLE 4.6 – AUC and logAUC of ChemPLP, LID and GRIM in virtual screening challenge with the DUD-E and PubChem (in bold) dataset. $\lambda = 0.001$.

Uniprot ID	**Number of protein structures**	**AUC**			**logAUC**		
		Chem PLP	GRIM	LID	ChemPLP	GRIM	LID
ALDR_HUMAN	1	70.25	75.76	72.21	35.1	43.07	44.74
BACE1_HUMAN	1	80.39	77.93	85.07	43.39	39.82	49.43
CAH2_HUMAN	1	49.61	95.96	95.89	14.06	76.69	82.63
CDK2_HUMAN	1	68.7	77.34	80.26	24.96	31.6	43.7
ESR1_HUMAN	1	55.22	88.89	86.18	17.93	47.63	47.48
	10	60.02	56.16	57.72	18.2	22.49	22.61
GRIA2_RAT	1	66.9	77.23	51.35	25.2	38.84	14.84
	10	70.25	75.35	72.45	29.71	38.11	33.14
HS90A_HUMAN	1	50.07	70.65	61.48	14.7	29.58	22.63
	10	48.9	74.17	71.58	13.85	41.69	31.03
LKHA4_HUMAN	1	94.39	92.48	93.49	59.29	62.98	67.54

LogAUC was computed with the following formula:

$$\log_{\mathrm{AUC}_\lambda} = \frac{1}{\log_{10} \frac{1}{\lambda}} \times \sum_{i}^{\text{where } x_i \geq \lambda} (\log_{10} x_{i+1} - \log_{10} x_i) \times \left(\frac{y_{i+1} - y_i}{2} \right) \tag{4.2}$$

TABLE 4.7 – Enrichment factors and true positive percent at 5 % decoys of ChemPLP, LID and GRIM in virtual screening challenge with the DUD-E and PubChem (in bold) dataset.

Uniprot ID	**Number of protein structures**	**EF5%**			**EFmax**	**TP% at 5% decoys**		
		ChemPLP	GRIM	LID		Chem PLP	GRIM	LID
ALDR_HUMAN	1	8.06	9.2	9.71	56.5	37.1	41.5	44
BACE1_HUMAN	1	10.1	6.69	10.5	63.9	45.2	31.4	49.4
CAH2_HUMAN	1	1.02	19.7	21.9	63.2	5.08	85.7	85.1
CDK2_HUMAN	1	3.38	4.27	9.53	58.7	16.6	21.3	45.5
ESR1_HUMAN	1	1.38	12.2	9.96	53.9	6.79	52.2	44.9
	10	0.67	3.75	3.75	25.9	3.39	16.9	18.6
GRIA2_RAT	1	3.81	9.88	1.27	74.8	18.3	46.8	6.33
	10	4.36	7.88	5.36	74.8	20.8	37.9	25.9
HS90A_HUMAN	1	0.91	4.86	3.07	55.1	4.55	23.8	15.9
	10	0.225	10.1	4.87	55.1	1.13	43.1	25
LKHA4_HUMAN	1	17	17.2	18.8	55.3	73.6	71.3	73.1

EF5% was computed with the following formula:

$$\mathrm{EF}_{5\%} = \frac{\mathrm{Actives}_{5\%}/\mathrm{Compounds}_{5\%}}{\mathrm{Actives}_{\mathrm{all}}/\mathrm{Compounds}_{\mathrm{all}}} \tag{4.3}$$

TABLE 4.8 – Computation time of LID grid generation and scoring on the eight targets of the DUD-E dataset. Only active compounds are considered. The measures were done in triplicate. Std: standard deviation.

Uniprot ID	**Number of poses**	**Number of interaction points in docked poses**	**Number of references**	**Number of interaction points in reference**	**Grid generation time (s)**		**LID scoring time (s)**	
					Mean	Std	Mean	Std
ALDR_HUMAN	2200	82,425	41	4150	8.37	0.24	7.28	0.19
BACE1_HUMAN	4850	230,795	227	24,510	49.9	0.53	19.53	0.49
CAH2_HUMAN	8342	277,604	156	8076	14.6	0.53	20.9	0.69
CDK2_HUMAN	7978	278,549	195	14,582	33.77	0.91	20.43	0.15
ESR1_HUMAN	6270	466,733	37	4504	11.53	0.42	30.47	0.8
GRIA2_RAT	2970	87,154	62	5028	11.77	0.21	5.97	0.15
HS90A_HUMAN	1250	57,673	139	11,440	25.27	0.15	6.04	0.29
LKHA4_HUMAN	2440	124,538	26	2548	6.21	0.19	9.83	0.64

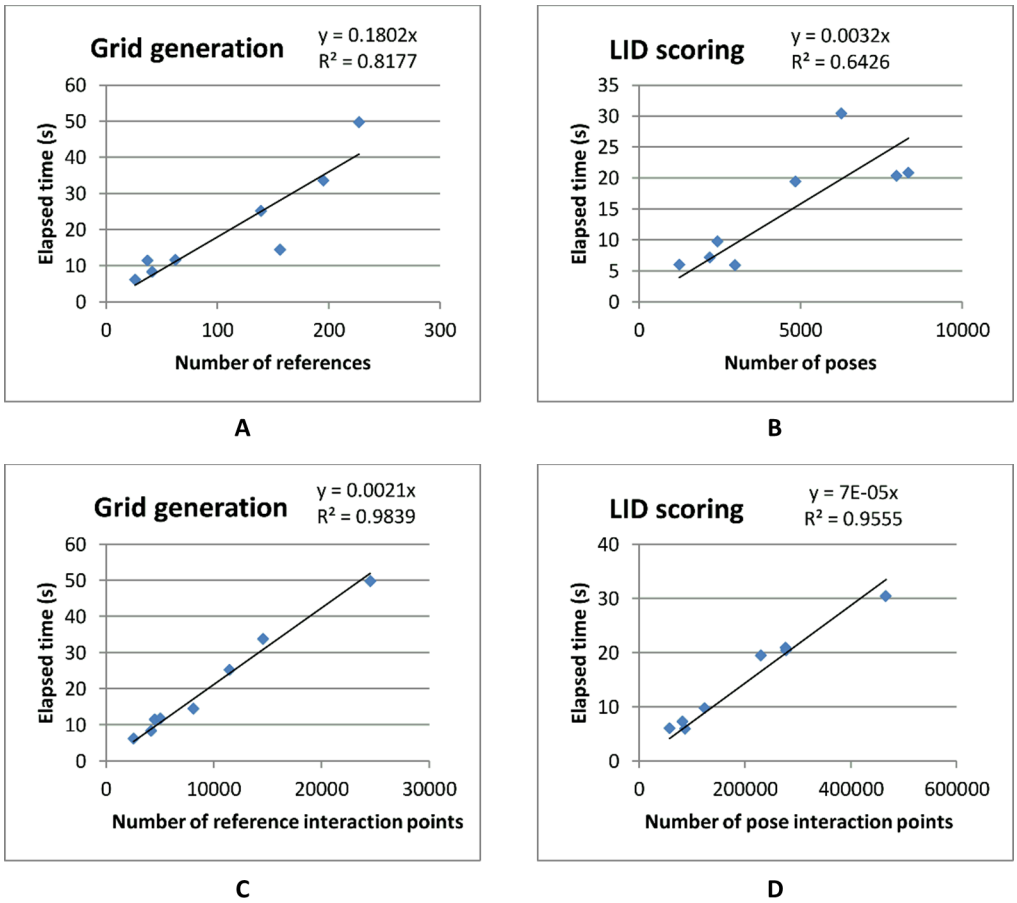

Figure 4.7 – Correlation between the number of features and the computation time.

TABLE 4.9 – Computation time of GRIM scoring on the eight targets of the DUD-E dataset. Only active compounds are considered. The measures were done in triplicate. Std: standard deviation.

Uniprot ID	**Number of poses**	**Number of interaction points in docked poses**	**Number of references**	**Number of interaction points in references**	**GRIM scoring time (s)**	
					Mean	Std
ALDR_HUMAN	2200	82,425	41	4150	810	30
BACE1_HUMAN	4850	230,795	227	24,510	11,990	887
CAH2_HUMAN	8342	277,604	156	8076	3931	88
CDK2_HUMAN	7978	278,549	195	14,582	7720	272
ESR1_HUMAN	6270	466,733	37	4504	13,430	540
GRIA2_RAT	2970	87,154	62	5028	470	17
HS90A_HUMAN	1250	57,673	139	11,440	1830	79
LKHA4_HUMAN	2440	124,538	26	2548		

Chapitre 5

Conclusion générale et perspectives

Le VIH-1 est un sujet d'étude qui est traité depuis bientôt 40 ans et qui continue à faire l'objet d'intenses recherches. La prévention et le développement d'antiviraux ont permis de contenir la pandémie et d'augmenter l'espérance de vie des personnes infectées. Mais le chemin à parcourir afin d'aboutir à une thérapie permettant son éradication semble encore long et devra mobiliser les compétences de scientifiques de diverses disciplines ainsi que des moyens des gouvernements du monde entier. Cette thèse fait part de ma contribution à l'avancée de la recherche contre le VIH-1 en ajoutant ma pierre à l'édifice.

J'ai abordé le problème de l'entrée virale avec ma vision de spécialiste de la modélisation, en apportant des réponses du point de vue de la structure à l'échelle atomique. Mes travaux ont été largement tournés vers le développement de méthodes. L'une d'elle, ATOLL, permet d'offrir une vision simple et globale des populations des RCPG et plus particulièrement du CCR5 en fonction du contexte biologique. À l'aide d'ATOLL et de simulations par dynamique moléculaire, j'ai pu identifier des populations différentes du CCR5 lié à quatre variantes #25, #34, Bx08 et JR-FL de la gp120. Ces populations sont caractérisées par des conformations proches, mais qui se distinguent par les propriétés structurales et dynamiques de la ECL3 et des parties engagées dans la liaison de la V3 de la gp120, à savoir la ECL2 et la cavité TM du CCR5. La modification d'un résidu dans la V3 est déterminant pour induire un comportement différent du CCR5 lors des simulations par dynamique moléculaire.

Un point commun entre le CCR5 et les autres membres de la famille des RCPG est que les particularités structurales de la partie extracellulaire du récepteur ont une influence sur la structure de sa partie intracellulaire, où le TM6 porte la signature de l'état fonctionnel du récepteur. J'ai montré par la méthode ATOLL que la fonction d'un ligand de CCR5 peut être évaluée par la simple projection de l'extrémité du TM6 de structures simulées par dynamique moléculaire. Les ligands bloquant toute signalisation du récepteur comme l'agoniste inverse maraviroc qui verrouille l'extrémité du TM6 dans une position refermant le site de liaison de l'effecteur intracellulaire. Les ligands antagonistes comme les gp120 ou agonistes comme les chimiokines restreignent moins le placement de l'extrémité du TM6 que l'agoniste inverse mais induisent somme toute un placement dans des zones caractéristiques de chaque ligand.

Ce constat m'a fait poser la question du caractère fonctionnalité biaisé des gp120 qui seraient capables de privilégier certaines voies de signalisation en fonction de leur séquence. Des résultats préliminaires de Bernard LAGANE et son équipe montrent en effet que certaines gp120 ont des propriétés de signalisation spécifiques. Il serait alors intéressant d'identifier des molécules capables d'imiter les liaisons des différents variantes de la gp120 et ainsi leur fonction. Le criblage virtuel par docking moléculaire de chimiothèques commerciales semble une technique de choix pour atteindre cet objectif, car ayant déjà fait ses preuves pour identifier de telles molécules. Cependant, la résolution de ce problème par une représentation statique des modèles comme cela est fait généralement aboutirait probablement à des résultats erronés. En effet, j'ai pu montrer que quand bien même les conformations du CCR5 sont proches, leur propriété en terme de liaison à un ligand peut être très différente. Il sera alors nécessaire de prendre en compte l'aspect dynamique du récepteur. De plus, les fonctions de score des programmes de docking font preuve d'une performance modérée et des méthodes basées sur des motifs d'interaction permettent d'écarter les mauvaises prédictions. En conséquence j'ai développé la méthode LID, qui a été validée en prédiction de pose et en criblage virtuel. Pour la recherche de ligands biaisés de CCR5, la méthode LID se baserait sur les motifs d'interaction au CCR5 de chaque variante de la gp120, tels qu'ils sont issus des simulations par dynamique moléculaire .

Publications et communications orales

5 articles de recherche dans des journaux internationaux à comité de lecture

1. Jacquemard C, Tran-Nguyen VK, Drwal MN, Rognan D, Kellenberger E. Local Interaction Density (LID), a Fast and Efficient Tool to Prioritize Docking Poses. *Molecules* **2019** *24*(14). pii : E2610. doi : 10.3390/molecules24142610.
2. Jacquemard C, Drwal MN, Desaphy J, Kellenberger E. Binding mode information improves fragment docking. *Journal of Chemoinformatics*. **2019** *11*(1):24. doi : 10.1186/s13321-019-0346-7.
3. Drwal MN, Jacquemard C, Perez C, Desaphy J, and Kellenberger E. Do Fragments and Crystallization Additives Bind Similarly to Drug-like Ligands ? *Journal of Chemical Information and Modeling* **2017** *57*(5), 1197-1209. doi : 10.1021/acs.jcim.6b00769.
4. Tran-Nguyen VK, Jacquemard C, Rognan D. LIT-PCBA : An Unbiased Data Set for Machine Learning and Virtual Screening. *Journal of Chemical Informatics and Modeling* **2020** *60*(9):4263-4273. doi : 10.1021/acs.jcim.0c00155.
5. Drwal MN, Bret G, Perez C, Jacquemard C, Desaphy J, Kellenberger E. Structural Insights on Fragment Binding Mode Conservation. *Journal of Medicinal Chemistry* **2018** *61*(14):5963-5973. doi : 10.1021/acs.jmedchem.8b00256.

Une revue dans un journal international à comité de lecture

1. Jacquemard C, Kellenberger E. A bright future for fragment-based drug discovery : what does it hold ? *Expert Opinion in Drug Discovery* **2019** *14*(5) :413-416. doi : 10.1080/17460441.2019.1583643.

4 présentations par affiche

1. Jacquemard C, Kellenberger E. ATOLL : A visualisation tool to compare transmembrane domains structures. *3rd international meeting of the European Research Network on Signal Transduction*, en ligne, 14 Septembre 2020.
2. Jacquemard C, Colin P, Brelot A, Lagane B, Kellenberger E. Modelling of the conformational response of CCR5 to the binding of HIV-1 gp120 variants. *6th annual meeting of the GDR 3545 RCPG-PhysioMed 2016*, Montpellier, Octobre 2019.
3. Jacquemard C, Tran-Nguyen VK, Drwal MN, Rognan D, Kellenberger E. Local Interaction Density (LID) a Fast and Efficient Tool to Prioritize Docking Poses. *9ème journées de la Société Française de Chémoinformatique*, Paris, 22 Novembre 2019.
4. Jacquemard C, Drwal MN, Perez P, Desaphy J, Kellenberger E. Fragment docking : Pose selection by consensus references binding mode. *Summer School on Chemoinformatics*, Strasbourg, Juin 2018 & *International Conference on Chemical Structure*, Noordwijkerhout (Pays Bas), Mai 2018.

Université de Strasbourg

Célien JACQUEMARD

ANRS France Recherche Nord & sud Sida-hiv Hépatites
Agence autonome de l'Inserm

Étude par modélisation moléculaire de la reconnaissance du corécepteur CCR5 du VIH-1 par la glycoprotéine virale gp120

Résumé

Depuis sa découverte en 1983, Le VIH-1 constitue un problème de santé publique majeur. A ce jour, malgré de nombreux traitements efficaces, les personnes infectées ne peuvent pas guérir, et le virus n'est toujours pas éradiqué. Pour prévenir l'émergence de souches résistances aux traitement ou plus agressives dans la progression de la maladie, il est essentiel de comprendre les interactions du virus avec la cellule qu'il infecte. L'entrée virale est pour cela une étape clé du cycle viral, avec la reconnaissance par le glycoprotéine virale gp120 du corécepteur, majoritairement CCR5. Il est désormais établi que le VIH-1 exploite la diversité des populations de récepteurs à la surface des cellules. Les travaux présentés dans cette thèse caractérisent, à l'échelle atomique, les différences de conformations correspondant aux populations ciblées par quatre virus différents. Nous avons utilisé la dynamique moléculaire pour simuler la dynamique du corécepteur CCR5 lié aux variantes de la gp120. L'analyse des trajectoires a conduit au développement d'une nouvelle méthode, ATOLL, qui permet de prédire les propriétés fonctionnelles du ligand, ici les variantes de la gp120, à partir de la position de extrémités intracellulaires des domaines transmembranaire d'un récepteur couplé à une protéine G, ici le CCR5. Nous avons aussi réussi à distinguer les populations de CCR5 liées aux gp120 par des motifs d'interactions intermoléculaires spécifiques. Dans le but d'exploiter ces motifs pour l'identification par criblage virtuel des petites molécules capables d'imiter une variante donnée de la gp120, nous avons développé LID, une méthode de score de docking capable de prendre en compte les multiples structures issues des simulations par dynamique moléculaire comme référence pour la sélection de poses.

Mots-clés : VIH – CCR5 – dynamique moléculaire – RCPG – docking – mode de liaison – agonisme biaisé

Résumé en anglais

Since its discovery in 1983, HIV-1 has been a major public health problem. Today, despite many effective treatments, the people infected are not cured, and the virus has still not been eradicated. To prevent the emergence of strains that are resistant to treatment or that are more aggressive in the progression of the disease, it is essential to understand the interactions of the virus with the cell it infects. Viral entry is therefore a key step in the viral cycle. It requires the recognition by the viral glycoprotein gp120 of the coreceptor, which is mainly CCR5. It is now established that HIV-1 exploits the diversity of receptor populations that exist on the surface of cells. The work presented in this thesis characterizes, at the atomic scale, the differences in conformations corresponding to the populations targeted by four different viruses. We used molecular dynamics to simulate the CCR5 coreceptor in complex with the gp120 variants. The trajectories analysis prompted us to the development of a new method, ATOLL, which is able to predict the functional properties of the ligand, here the variants of gp120, from the position of the intracellular tails of the transmembrane domains of a G-protein coupled receptor, here CCR5. We also observed that the CCR5 populations selected by the gp120 variants have specific intermolecular interaction patterns. In order to exploit these patterns for the identification by virtual screening of small molecules mimicking a given variant of gp120, we have developed a docking score method, named LID, which is able to take into account as a reference for the selection of poses the multiple structures issued by the simulations by molecular dynamics.

Keywords: HIV – CCR5 – molecular dynamics – GPCR – docking – binding mode – biaised agonisme