Stratégies de docking-scoring assistées par analyse de données. Application au criblage virtuel des cibles thérapeutiques COX-2 et PPAR gamma

Alban Arrault

THESE PRESENTEE A L'UNIVERSITE D'ORLEANS

POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITE D'ORLEANS

Discipline: Modélisation Moléculaire et Chemoinformatique

Par

Alban ARRAULT

Stratégies de docking-scoring assistées par analyse de données.

Application au criblage virtuel des cibles thérapeutiques COX-2 et PPAR gamma

Soutenue le 30 novembre 2007

R. BUREAU	Rapporteur, professeur, CERMN, UPRES EA 3915, Université de Caen
J. VAMECQ	Rapporteur, chargé de recherche INSERM, Faculté de Médecine de Lille
D. GENEST	Président, directeur de recherche CNRS, CBM
M. NEUWELS	Docteur, Chimie du soin, L'OREAL
L. MORIN-ALLORY	Professeur, ICOA, UMR CNRS 6005, Université d'Orléans
C. MAROT	Docteur, HDR, ICOA, UMR CNRS 6005, Université d'Orléans

REMERCIEMENTS

Ce travail a été réalisé sous la direction du Professeur Luc MORIN-ALLORY et du Docteur Christophe MAROT à l'Université d'Orléans au sein de l'Institut de Chimie Organique et Analytique, UMR CNRS 6005, dirigé par le Professeur Olivier MARTIN.

Je tiens à exprimer ma plus profonde reconnaissance au Professeur Luc MORIN-ALLORY, au Docteur Christophe MAROT, aux Professeurs Gérald GUILLAUMET et Olivier MARTIN.

J'exprime également ma plus sincère sympathie aux Docteurs Monique et Daniel GENEST et Franck SUZENET avec qui j'ai étroitement collaboré durant ce travail.

Mes remerciements s'adressent également au Docteur Joseph VAMECQ, au Docteur Michel NEUWELS, au Docteur Daniel GENEST et au Professeur Ronan BUREAU qui ont accepté de juger ce travail et de siéger parmi les membres du jury.

Je remercie également toute l'équipe de Modélisation Moléculaire pour leur sympathie et leurs encouragements, en particulier mes collègues Dr. Aurélien MONGE, Dr. Maryline BOUROTTE et Laurent ROBIN.

A mes parents

A Karine

A Xavier et Axelle

A Mamie

Je vous dédie cet ouvrage

SOMMAIRE

INTRODUCTION GENERALE **11**
GENERALITES: Le processus de criblage virtuel **17**
A. Technique basée sur la structure de la protéine: le docking 18
1. La structure des protéines 19
a) La résolution 21
b) Le facteur R 22
c) Choix d'une structure cristallographique 22
d) Facteur d'agitation thermique 22
e) Détermination du site actif 22
f) Les interactions entre le ligand et le récepteur 23
2. Le processus de docking 24
a) La recherche conformationnelle 24
(1) L'approche «classique» de recherche conformationnelle 25
(2) L'approche « rationnelle» de recherche conformationnelle 25
b) La flexibilité de la protéine 27
c) Simulation de la présence d'eau 27
3. Le processus de scoring 28
a) Principe 28
b) Familles de fonctions de scoring 29
(1) Les fonctions de scoring empiriques 29
(2) Les fonctions de type knowledge-based 30
c) Description des fonctions de scoring utilisées 30
(1) La fonction de scoring DOCK 30
(2) La fonction de scoring GOLD 31
(3) La fonction de scoring FlexX 31
(4) La fonction de scoring Chemscore 32
(5) La fonction de scoring PMF 32
d) Types d'interactions 33
(1) La lipophilie 33
(2) Les interactions électrostatiques 33
(3) Les termes de solvatation et d'entropie 34
(4) Conclusion sur l'utilisation des fonctions avec COX-2 et PPARγ 34
4. Interprétation des fonctions de scoring 34
B. Technique basée sur la structure des ligands: le pharmacophore 35
C. Dynamique moléculaire 36
1. Principe 36
2. Intérêt de la dynamique moléculaire en amont du docking 36
D. Les chimiothèques dédiées au criblage virtuel 37
1. Généralités 37
2. Chimiothèque locale 38
3. Filtres physico-chimiques 38
4. Préparation des molécules 38
a) Etats de protonation 38
b) Raffinement de la structure des molécules 39
CHAPITRE 1.Elaboration d'un protocole de docking-scoring pour le criblage virtuel. Application à l'enzyme COX-2 **40**
A. Introduction 41

1. Fonction de la cyclooxygénase 42
2. Anti-inflammatoires non stéroïdiens (AINS) et agents coxibs 44
3. Site actif et interaction 47
4. Structures cristallographiques 48
B. Matériel et méthodes 50
1. Constitution de chimiothèques dédiées à évaluer les modèles 50
a) Constitution de l'«*ensemble total*» 50
(1) Généralités 50
(2) Molécules actives 50
(3) Molécules a priori inactives ou leurres 51
(4) Ensemble total 52
b) Constitution de l'*ensemble d'entraînement* et de test 52
2. Etape de docking 52
a) Généralités 52
b) Le processus de docking avec FlexX 53
(1) Choix du fragment de base 53
(2) Reconstruction sur le fragment de base 54
(3) Choix des meilleurs placements 55
(4) Test de recouvrement protéine/ligand 55
(5) Exploration des paramètres fondamentaux 56
c) Orientation des groupements donneurs /accepteurs 56
3. Modèle pharmacophorique 58
4. Les fonctions de scoring 59
a) Choix de la première pose 59
b) Re-scoring 60
(1) Consensus de fonctions de scoring 60
(2) Méthodes statistiques 64
c) Performance des fonctions de scoring 66
(1) La matrice de confusion 66
(2) Enrichissement 67
(3) Molécules rejetées par le modèle 68
(4) Courbes d'enrichissement 69
(5) Le R_s 71
C. Résultats et discussion 75
1. Comparaison des structures cristallographiques de cyclooxygénase de type 2 75
a) Etude de la 1CX2 d'origine 75
b) Optimisation de la structure 1CX2 80
2. Comparaison des fonctions de scoring dans un processus de re-scoring 81
a) Stratégies de consensus 83
(1) Les méthodologies «Rank by rank», «Rank by number» et «Rank by best» 83
(2) La méthodologie «Rank by vote» 85
b) Stratégie d'analyse factorielle discriminante 86
c) Comparaison de l'AFD et des méthodes de consensus 87
3. Evaluation des 7 structures cristallographiques 88
4. Validation du modèle 1CX2 optimisé 91
5. Evaluation du modèle vis-à-vis de molécules sans SO_2 93
6. Temps de calcul 97
7. Modèle pharmacophorique 98
a) FlexX-Pharm 98
(1) Généralités 98

(2) Principe 98
b) Application à la cyclooxygénase de type 2 99
8. Criblage virtuel 102
a) Constitution des chimiothèques dédiées au criblage virtuel 102
(1) Chimiothèque commerciale 102
(2) Chimiothèque nationale et ICOA 103
b) Protocole du criblage virtuel 103
c) Choix des produits 104
d) Tests biologiques 104
D. Conclusion et perspectives 108
CHAPITRE 2. Elaboration d'un modèle prédictif des récepteurs PPARγ, Implication de la flexibilité du site actif 110
A. Introduction 111
1. Le diabète 111
a) Généralités 111
(1) Le diabète de type I ou insulino-dépendant 112
(2) Le diabète de type II ou non-insulino-dépendant 113
b) La glycorégulation 113
c) Les sucres 114
d) Les lipides 115
e) Les protéines 115
f) Conséquences et complications du diabète 115
(1) Les complications aiguës 115
(2) Les complications mécaniques 115
2. Traitements disponibles 116
a) Les sulfonylurées 116
b) Les glinides 117
c) Les biguanides 117
d) Les thiazolidinediones 118
e) Autres traitements 118
f) En pratique 119
3. Les PPARs: Peroxysome Proliferator-Activated Receptors 120
a) Généralités 120
b) Mécanisme d'action 121
c) Homologie de séquence des trois PPARs 122
4. Ligands naturels et synthétiques de PPARα, δ et γ 123
a) Ligands PPARα 123
(1) Agonistes naturels 123
(2) Agonistes de synthèse 123
b) Ligands PPARδ 124
(1) Agonistes naturels 124
(2) Agonistes de synthèse 124
c) Ligands PPARγ 125
(1) Agonistes naturels 125
(2) Agonistes de synthèse 125
(3) Agoniste partiel de synthèse 126
(4) Antagonistes de synthèse 126
d) Co-agonistes PPARα/γ 127
5. Implication des différents isoformes 128
a) PPARα 128

(1) Métabolisme lipidique 128
(2) L'inflammation 129
b) PPARδ 129
c) PPARγ 129
6. Structure cristallographique de PPARγ 130
a) Généralités 130
7. Structure tertiaire des PPARγ 134
a) Variabilité spatiale 134
(1) Description 134
(2) Mécanismes de trans-conformation 135
b) Variabilité interactionnelle 137
B. Matériel et méthodes 139
1. Constitution de l'*ensemble total* 140
a) Généralités 140
b) Molécules actives 140
c) Molécules inactives (*leurres*) 141
2. Dynamique moléculaire 141
a) Détermination des résidus du site actif 142
b) Minimisations et simulations de dynamique moléculaire 142
c) Protocole des dynamiques moléculaires 142
d) Calcul du RMSD 143
e) Cartes RMSD 144
3. Couplage de la dynamique moléculaire au docking 145
4. Evaluation du mode de docking 146
a) Problématique 146
b) Le RMSD 147
c) Fitting-Score 147
(1) Calcul 147
(2) Courbes de fréquence relative cumulée 148
d) Sphères de présence 149
C. Résultats et discussions 150
1. Comparaison et choix des structures cristallographiques 150
a) Volume du site actif des PPARγ 150
b) Evaluation du docking 151
(1) Fréquences relatives cumulées 151
(2) Superposition des molécules actives dans deux cas extrêmes selon Fitting-Score 152
c) Evaluation des scores 154
(1) Test de similarité des fonctions de scoring 154
(2) Choix de la première pose 155
(3) Etape de re-scoring 156
(4) Consensus des fonctions de scoring 162
(5) Analyse factorielle discriminante 164
(6) Comparaison des consensus avec l'analyse factorielle discriminante 165
(7) Validation de l'AFD 166
2. Dynamique moléculaire 167
a) Simulation de 4PRG_a 167
(1) Carte RMSD 168
(2) Calcul du volume du site actif 169
b) Simulation de 2PRG 170

(1) Carte RMSD 170
(2) Evaluation du Fitting-Score 171
(3) Calcul du R_s et du R_a 172
(4) Courbes d'enrichissement des meilleurs conformères 173
(5) Consensus de conformères et de structures cristallographiques 173
(6) Evaluation des stratégies de consensus et d'AFD 174
3. Estimation des temps de calcul 177
4. Criblage virtuel 177
D. Conclusion et perspectives 180
CONCLUSION GENERALE 181
GLOSSAIRE 187

INTRODUCTION GENERALE

Le processus de découverte d'un médicament est long et onéreux. Entre 12 et 15 ans et près d'un milliard de dollars sont nécessaires à la mise sur le marché d'un médicament (Figure 1).[1,2]

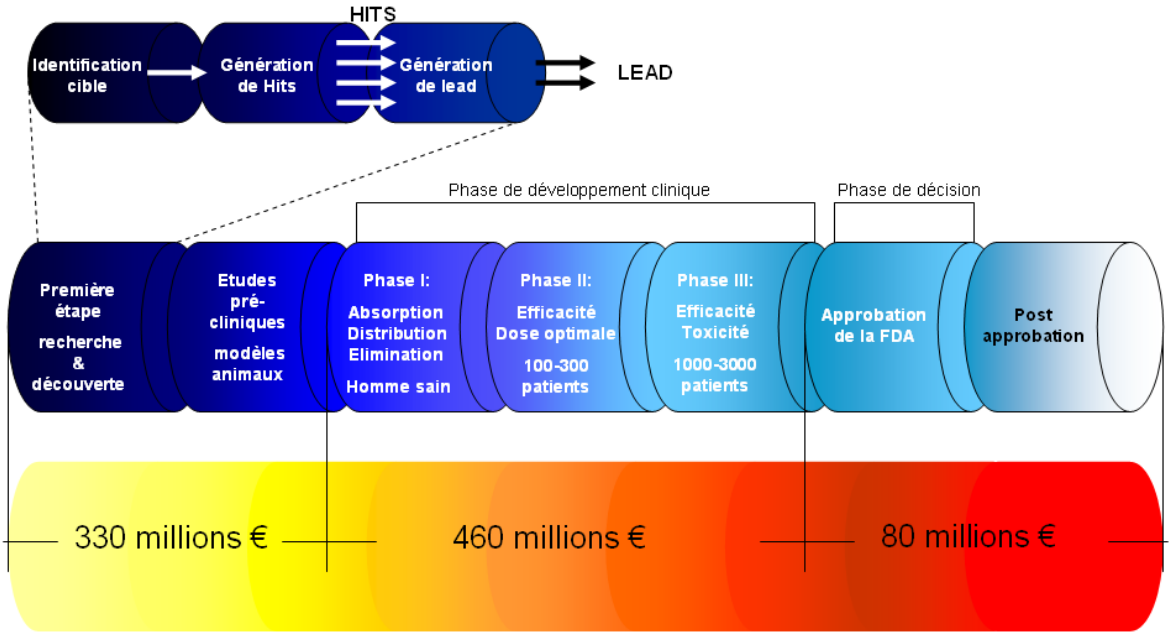


Figure 1. Etapes du processus de découverte d'un médicament

Depuis les trente dernières années, le processus de recherche et développement (R&D) au sein de l'industrie pharmaceutique a subit des changements majeurs. Plus particulièrement, depuis 1984, le «Waxman-Hatch act» encourage l'innovation dans le domaine du médicament tout en facilitant l'entrée des génériques. Les industries sont confrontées à une nouvelle ère dans laquelle les produits ont un cycle de vie limité. Dans ce contexte, accroître la quantité de molécules innovantes est indispensable pour répondre au mieux aux besoins grandissants d'alternatives thérapeutiques. Actuellement, d'un point de vue externe, il semblerait que la R&D de nouveaux composés soit en perte de vitesse. Ceci est lié à la fois à des contraintes toujours plus drastiques ainsi que des coûts de développement toujours plus élevés.

L'aspect «innovation» est également un point sur lequel nous devons insister. En effet, un projet thérapeutique est un enchaînement d'étapes complexes qui doivent être optimisées puis validées. Amener un nouveau produit sur le marché implique que chacune de ces étapes soit novatrice. De l'initiation d'un projet (aussi bien académique qu'industriel) jusqu'aux tests

[1] Schmid, E.F.; Smith, D.A. Keynote review: Is declining innovation in the pharmaceutical industry a myth? *Drug Discov. Today* **2005**, *10*, 1031-1039.
[2] O'Driscoll, C. A virtual space odyssey, *Horizon Symposia* **2004**, 1-4.

cliniques, un renouveau doit être apporté en terme de structure chimique, de formulation mais également d'un point de vue pharmacologique. L'impact des nouvelles sciences et de technologies toujours plus performantes est considérable sur la découverte de molécules à visée thérapeutique. De manière plus pragmatique, la conception rationnelle par les technologies informatiques occupe une position grandissante dans le processus de découverte. L'avancée de la biologie moléculaire est également un élément jouant en la faveur de l'industrie pharmaceutique, avec le décryptage du génome humain ainsi que la compréhension du fonctionnement de récepteurs capables de réguler l'expression des gènes.

Les contraintes économiques sont en partie dues aux investissements liés aux technologies mais surtout aux phases de test clinique, toujours plus onéreuses mais nécessaires à la commercialisation d'entités chimiques. Les produits qui échouent en fin de tests cliniques et n'accédant donc pas au marché du médicament entraînent des pertes financières colossales. Dans le pire des cas, des médicaments peuvent être retirés du marché après avoir obtenu l'approbation de l'AFSSA (agence nationale de sécurité sanitaire des aliments).[3] C'est généralement une toxicité observée lors de la phase de pharmaco-vigilance (phase 4 du développement clinique) qui est la cause du retrait. Cette situation conduit à des pertes économiques majeures car l'industrie impliquée doit généralement mener des études complémentaires, afin de prouver l'innocuité du produit mis en cause. Aujourd'hui, les scientifiques impliqués dans la recherche de médicaments n'ont jamais autant eu à considérer le marché, la commercialisation ainsi que la toxicité d'une molécule, très tôt dans le processus de R&D.

Des méthodes virtuelles capables de prédire l'affinité de produits sont de plus en plus utilisées. D'ailleurs, un certain nombre de médicaments mis sur le marché proviennent d'une conception rationnelle basée sur des stratégies de criblage virtuel comme par exemple:[4,5]

-Des inhibiteurs de l'adose réductase: par recherche dans des bases de données de composés

-Un inhibiteur d'un élément de réponse de la transactivation de la HIV-1 RNA: par docking rigide et recherche dans des bases de données de composés

-Un inhibiteur de la thrombine: par docking sur des chimiothèques combinatoires et par des méthodes *de novo*

[3] http://www.afssa.fr/

[4] Stahl, M. Structure-based library design. In Virtual Screening for Bioactive Molecules *Wiley-VCH* **2000**, 229-264.

[5] Schneider, G.; Böhm, H.-J. Virtual screening and fast automated docking methods *Drug Discov. Today* **2002**, *7*, 64-69.

-Un inhibiteur de la glyceraldéhyde-3-phosphate dehydrogenase: par docking sur des chimiothèques combinatoires

L'avantage majeur de ces méthodes est d'aider à la prédiction de molécules en un temps limité et surtout parfois sans avoir à synthétiser les composés (lorsque ceux-ci proviennent de chimiothèques commerciales). Le criblage virtuel s'apparente à une succession de techniques dont le point de départ est le pré-filtrage des chimiothèques à cribler. Souvent, la sélection des composés avant criblage virtuel suit des règles spécifiques aux médicaments déjà connus et mis sur le marché (règles «drug-like»).[6,7] Par exemple, la règle de Lipinski ou encore «règle des 5» peut être utilisée pour la sélection de composés « drug-like »:

- masse molaire $\leq$ 500 g/mol
- Log P $\leq$ 5
- Accepteurs de liaisons H $\leq$ 10
- Donneurs de liaisons H $\leq$ 5

Tous les composés ne validant pas au moins trois de ces conditions sont susceptibles de poser des problèmes d'absorption par voie orale.

Des adaptations peuvent être faites afin de rendre la «règle des 5» plus stricte. C'est la notion de «lead-like». La masse molaire requise est diminuée (Masse molaire $\leq$ 460 g/mol). De même, le Log P doit être compris entre 4 et 4,2. Les composés chef de file «lead-like» peuvent ainsi subir des optimisations ou des ajouts de fragments induisant une augmentation de la masse molaire ainsi que du Log P tout en restant dans des limites compatibles avec les règles «drug like». Les composés respectant les critères «lead-like» ne sont pas pour autant certains de devenir des médicaments. En effet, beaucoup de molécules actives sont abandonnées du fait qu'elles possèdent une pharmacocinétique en dehors des normes. Afin d'évaluer le métabolisme d'une molécule, des méthodes de prédiction *in silico* basées sur la modélisation des cytochromes P450 ont vu le jour.[8] De manière plus générale, l'ADME-Tox (Absorption, Distribution, Métabolisation, Elimination - Toxicité) des médicaments est un facteur à prendre en compte dans l'optimisation de candidats prometteurs. Nous avons vu précédemment que des filtres (par exemple, les règles de Lipinski) peuvent être utilisés afin

[6] Zheng, S.; Luo, X.; Chen, G.; Zhu, W.; Shen, J.; Chen, K.; Jiang, H. A New Rapid and Effective Chemistry Space Filter in Recognizing a Druglike Database *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 856-862.
[7] Lipinski, C.A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today* **2004**, *1*, 337-341.
[8] http://www.moldiscovery.com/soft_metasite.php

de réduire le nombre de molécules à analyser virtuellement. Il serait également judicieux d'élaborer des filtres basés sur les propriétés ADME mais également sur la toxicité potentielle, éliminant ainsi tous les composés qui induiraient des effets secondaires. L'élaboration de tels filtres est aujourd'hui encore très délicate car les phénomènes de toxicité sont difficiles à modéliser, du fait des multiples paramètres à prendre en compte. Après avoir choisi un ensemble de composés à analyser, une méthode de prédiction d'activité ou d'affinité doit être sélectionnée. Deux cas de figure sont à envisager:[9]

-Si la structure tridimensionnelle de la protéine est connue (structure-based design), l'arrimage (encore appelé docking) peut être utilisé pour prédire l'affinité des ligands pour récepteur donné.

-Si aucune donnée des protéines n'est disponible, une méthode alternative peut être envisagée.

Le modèle de prédiction est élaboré à partir de une ou plusieurs familles de molécules connues pour une activité biologique. Ce sont les méthodes basées sur les propriétés du ligand (ligand-based design). Le représentant de cette stratégie est le QSAR (Quantitative Structure Activity Relationship ou relation structure activité quantitative).[10] Il s'agit d'établir une corrélation entre les variables de propriétés moléculaires, encodées par des descripteurs moléculaires, et de l'activité biologique (ou l'affinité) pour un ensemble de composés.

Au cours de notre étude, nous nous sommes focalisés sur les méthodes basées sur la structure de la protéine. L'enrichissement toujours croissant en structures tridimensionnelles dont la qualité en terme de résolution est en nette amélioration, nous a guidés vers l'approche du docking. Des fonctions mathématiques (appelées fonctions de scoring) peuvent être ajoutées après docking afin d'évaluer numériquement le degré d'interaction et d'affinité entre les deux entités. Le docking peut être interprété de manière qualitative (par observation de l'entité «ligand» dans la cavité de la protéine) mais également de manière quantitative par traitement des données provenant des fonctions de scoring.

L'objectif de ce travail est d'explorer ces stratégies d'analyse impliquant les protéines mais également de développer des protocoles de criblage virtuel faisant participer plusieurs outils prédictifs. L'objectif est également de mettre en évidence le choix d'une méthodologie plutôt qu'une autre, dans une combinaison complexe d'algorithmes mathématiques et statistiques. Les étapes du processus ont été clairement identifiées, comparées puis sélectionnées dans un souci d'efficacité, tout en tenant compte du risque d'erreurs associées

[9] Jenkins, J.L.; Kao, R.Y.T.; Shapiro, R. Virtual Screening to Enrich Hit Lists From High-Throughput Screening: A Case Study on Small-Molecule Inhibitors of Angiogenin *Proteins* **2003**, *50*, 81-93.

[10] Walters, W.P.; Stahl, M.T.; Murcko, M.A.; Virtual screening-an overview *Drug Discov. Today* **1998**, *3*, 160-178.

au couplage de techniques différentes. Il n'y a pas de méthodologie de criblage virtuel prête à l'emploi, d'où la nécessité d'évaluer chaque algorithme.[11]

En introduction, nous positionnerons les différentes techniques employées dans un contexte de criblage virtuel et dont une des principales contraintes est le temps de calcul. Nous évoquerons également les outils utiles à la prédiction ainsi que la manière d'exploiter au mieux les données provenant des modèles élaborés. Nous avons aussi porté un intérêt tout particulier au couplage de techniques jusque là encore peu explorées avec le docking, tels que des pharmacophores ou la prise en compte de la flexibilité par dynamique moléculaire.

Dans une première partie, nous avons étudié le cas d'une cible impliquée dans les phénomènes de l'inflammation appelée la cyclooxygénase (COX). Nous évoquerons les filtres utilisés avec notamment l'utilisation de modèles pharmacophoriques, afin de réduire les temps de calculs engendrés par les algorithmes de docking. Nous montrerons également l'intérêt de comparer et de choisir les algorithmes avec, comme critère d'évaluation, le pouvoir prédictif du modèle en construction. Enfin, nous avons montré l'intérêt de traiter les ensembles de ligands de manière rigoureuse et plus précisément manipuler, avec précaution, l'aspect tridimensionnel des structures avant docking.

Dans la deuxième partie de ce travail, nous avons travaillé sur une cible thérapeutique différente d'un point de vue fonctionnel et structural. Les récepteurs au facteur activé de prolifération des péroxysomes (plus communément appelés PPAR, Peroxysome Proliferator Activated Receptor) sont une famille de protéines nucléaires, majoritairement responsable du métabolisme lipidique au sein de l'organisme. Une méthodologie permettant de prendre en compte les mouvements de la structure générale de ces récepteurs a été mise au point. Cette information a été exploitée dans un protocole de criblage virtuel dont les paramètres de docking et de fonctions de scoring ont été optimisés.

[11] Rester, U. Dock around the clock-Current status of small molecule docking and scoring *QSAR Comb. Sci.* **2006**, *25*, 605-615.

GENERALITES: Le processus de criblage virtuel

A. Technique basée sur la structure de la protéine: le docking

Le processus de criblage virtuel n'est généralement pas un enchaînement d'algorithmes défini par avance. Le choix d'un projet (type de cible, famille d'inhibiteurs ou agonistes) va conditionner une suite logique de programme dans lequel l'utilisateur va naviguer et en explorer les performances. Concrètement, la comparaison des programmes disponibles est nécessaire si l'on souhaite optimiser chaque étape d'un protocole aussi complexe qu'est le criblage virtuel. La conception rationnelle de médicament passe par l'identification d'un profil pharmacologique. Deux situations bien distinctes peuvent alors être décrites.

Le premier cas de figure est le projet dans lequel la cible thérapeutique n'a pas été identifiée pour plusieurs raisons (car trop instable ou impossible à cristalliser). C'est notamment le cas des récepteurs couplés aux protéines G (RCPG) pour lesquels la cristallisation demeure une difficulté technique. Dans la majorité de ces situations, le chimiste médicinal a recours aux techniques de relation structure activité quantitative (QSAR). Globalement, ces techniques tentent de corréler l'activité des molécules avec leurs propriétés physico-chimiques (codées par des descripteurs moléculaires). Les algorithmes de corrélation sont de deux natures: linéaire et non linéaire. L'avantage majeur est que la prédiction se fait avec l'activité de la molécule (et non sur l'affinité), c'est-à-dire le résultat de l'action de la molécule sur un ou plusieurs récepteur(s) biologique(s). L'inconvénient de cette méthode est que le modèle, basé sur une série de molécules, ne permet de prédire que des produits similaires voire, dans le meilleur des cas, des molécules dont le squelette de base varie sensiblement. Dans certains cas, même si la structure du récepteur n'a pas été résolue, des méthodes d'homologie peuvent être appliquées, se basant alors sur des structures supposées proche de la structure protéique à imiter. Dans le cas des RCPGs, la séquence des acides aminés de la protéine «mime» est alignée sur une des seules structures de RCPGs cristallisée à ce jour, appelée la rhodopsine. Il est évident que la qualité prédictive de la protéine «mime» dépend, entre autre, de la qualité de l'alignement des deux entités.

Le deuxième cas de figure correspond à celui de nos deux études pharmacologiques de la «cyclooxygénase de type 2» et des «récepteurs au facteur activé de prolifération des péroxysomes». Dans chacun de ces projets, les structures 3D ont été définies par cristallographie par rayons X. La diversité des ligands cristallisés dans la protéine est une information précieuse qui nous a permis d'observer d'une part les résidus clé de l'interaction mais également les variabilités structurales 3D de la protéine d'une molécule à l'autre. Le

docking est capable de confronter deux entités telles que le ligand et le récepteur. L'avantage est de pouvoir prendre en compte un environnement complexe d'acides aminés dans lequel va évoluer le ligand. Toutefois, un point faible du docking est d'omettre la variabilité structurale de la protéine lors de la fixation du ligand. Cette approximation peut parfois être la cause d'erreurs de prédiction.

Dans son sens large, le criblage virtuel est l'équivalent *in silico* du criblage réel à haut débit (HTS, High-throughput screening) avec lequel de nombreuses molécules sont testées, avec comme objectif de discriminer les agents actifs des inactifs. Cette stratégie est capable d'identifier de nouvelles molécules qui serviront de point de départ pour les chimistes médicinaux.

1. La structure des protéines

Trois méthodes expérimentales permettent aujourd'hui de déterminer la structure des protéines : la résonance magnétique nucléaire (RMN), la microscopie électronique et la cristallographie par rayons X. Cette dernière technique est responsable de la majorité des structures issues d'une base de données de structures accessibles gratuitement appelée la «protein data bank» (PDB),[12] c'est-à-dire plus de 40 000 structures protéiques (juin 2007). La RMN fournit, quant à elle, 6 200 structures contre seulement 150 dans le cas de la microscopie électronique. Nous avons représenté le nombre de structures définies les 30 dernières années, toutes techniques confondues (Figure 2).

[12] http://www.rcsb.org/.

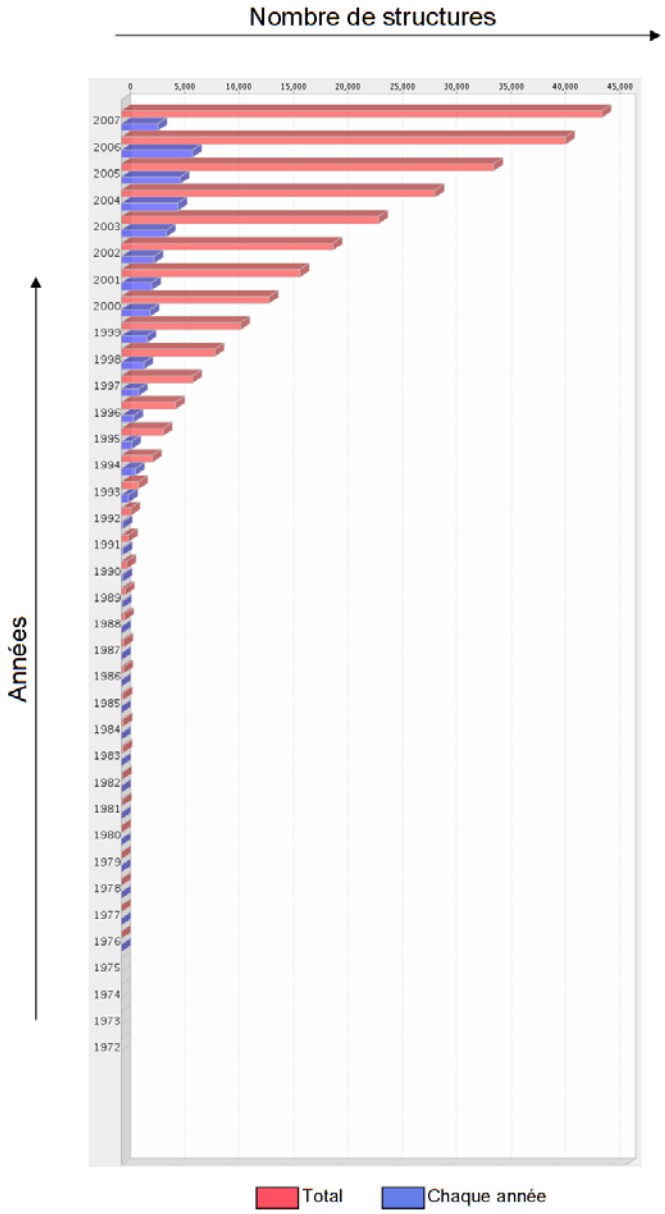


Figure 2. Evolution du nombre de structures protéiques disponibles dans la PDB

L'histogramme montre une évolution exponentielle du nombre de structures définies chaque année (en bleu). Le nombre de structures totales (en rouge) suit également une courbe exponentielle de croissance, prouvant la progression des techniques. Une autre preuve du succès de ces méthodes est la qualité des structures.[13]

[13] Morris, A.L.; MacArthur, M.W.; Hutchinson, E.G.; Thornton, J.M. Stereochemical quality of protein structure coordinates *Proteins* **1992**, *12*, 345-364.

a) La résolution

La résolution en angström de la protéine est une des données reflétant la qualité des structures ayant permis de construire le modèle cristallographique. Généralement, la résolution est limitée par la manière dont les cristaux diffractent, le temps nécessaire pour collecter des données de meilleures résolutions, la dynamique de la protéine, la qualité de l'appareillage et la température du système. Une résolution proche de 1 Å permet de distinguer tous les atomes y compris les hydrogènes. Une résolution de l'ordre de 6 Å permet seulement de distinguer que des structures de types «hélice α» ou «feuillet β» par exemple (Figure 3).

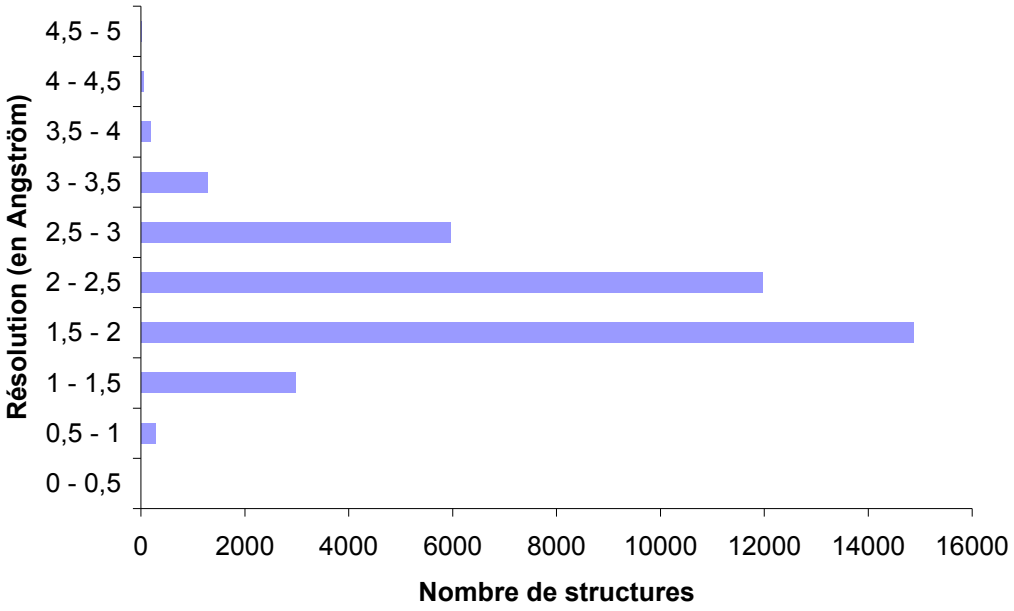


Figure 3. Nombre de structures protéiques par intervalle de résolution donné

L'histogramme précédent illustre parfaitement la répartition de la qualité des structures au sein de la PDB. La majorité des structures ont une résolution comprise entre 1,5 et 2,5 Å. Cet intervalle de résolution prouve que les structures sont de bonne qualité.

b) Le facteur R

Un autre indicateur est le facteur R qui est une grandeur indicatrice de l'écart entre les facteurs de structures observés et calculés. Le facteur R est compris entre 0 et 1 (plus le facteur R est proche de 0 et plus la prédiction est juste).

c) Choix d'une structure cristallographique

La combinaison des deux facteurs évoqués précédemment permet de réaliser un choix pertinent de structure cristallographique. Par exemple, une structure de protéine dont le facteur R est proche de 0 mais possédant une mauvaise résolution n'a aucune signification. Typiquement, une structure dont le facteur R avoisine la valeur de 0,2 et possédant une résolution correcte (inférieure ou égale à 2 Å) correspond la plupart du temps à une structure précise et bien définie. Il existe à ce jour des outils capables d'évaluer la qualité des structures cristallographiques (REDUCE[14]).[15]

d) Facteur d'agitation thermique

Les atomes du cristal bougent autour d'une position atomique moyenne. Les rayonnements ne voient pas les atomes exactement à la même position dans les différentes mailles du cristal. Ce facteur permet d'observer la flexibilité des résidus.

e) Détermination du site actif

Une étape fondamentale dans la stratégie de criblage virtuel est l'identification des résidus d'acides aminés pouvant intervenir dans le processus de reconnaissance du ligand. Un ligand cristallisé avec le récepteur est susceptible de renseigner sur l'emplacement du site actif. Malgré tout, cette information doit être interprétée avec prudence car, pour un même récepteur, le site actif peut être différent selon le profil pharmacologique de la molécule (agoniste, antagoniste, agoniste inverse).

Dans le cas où aucun ligand n'a été cristallisé avec le récepteur, des méthodes capables de détecter les cavités sont une bonne alternative. Nous avons utilisé CASTp pour mesurer le

[14] http://kinemage.biochem.duke.edu/software/reduce.php
[15] Word, J.M.; Lovell, S.C.; Richardson, J.S.; Richardson, D.C. Asparagine and glutamine: using hydrogen atom contacts in the choice of sidechain amide orientation *J. Mol. Biol*. **1999**, *285*, 1733-1745.

volume des sites actifs.[16] Cet algorithme, disponible en ligne, mesure les surfaces accessibles des cavités accessibles mais également inaccessibles. Les méthodes utilisées consistent à évaluer la surface accessible aux solvants (surface Richards[17]) et la surface moléculaire (surface Connolly[18]). Enfin, CASTp permet de mesurer le nombre d'ouvertures, leur taille ainsi que leur circonférence. Généralement, ce type d'algorithme identifie plusieurs régions possibles. Il revient alors à l'utilisateur de tester tous ces sites de fixation potentiels et accessibles vis-à-vis du profil pharmacologique de molécules connues.

f) Les interactions entre le ligand et le récepteur

Les interactions entre une protéine et un ligand sont en général de nature non covalente. L'énergie libre de Gibbs se calcule selon l'Équation 1.

$$\Delta G = -RT \ln K_i = \Delta H - T\Delta S$$

Équation 1. Energie libre de Gibbs

R: constante des gaz (8,314 J. K^{-1} mol^{-1})
T: Température (K)
K_i: constante d'inhibition

La grandeur du ΔG en solution aqueuse est comprise entre -10 et -70 KJ/mol. Les composantes de l'énergie libre sont:
- la complémentarité géométrique, stérique et de surface entre le ligand et le récepteur
- les contacts présents entre deux régions lipophiles des deux entités.
- les liaisons hydrogène

Les molécules d'eau dans les cavités des protéines peuvent parfois être un élément fondamental.[19] Elles sont capables d'assurer le relais entre le récepteur et le ligand et ainsi créer des réseaux de liaisons hydrogène.

[16] http://sts.bioengr.uic.edu/castp/
[17] Lee, B.; Richards, F.M. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **1971**, 55, 379-400.
[18] Connolly,M. L. Solvent-accessible surfaces of proteins and nucleic-acids *Science* **1983**, *221*, 709-713.
[19] Maréchal, Y. The hydrogen bond and the water molecule *Elsevier Science & Technology* **2007**.

2. Le processus de docking

Le processus de docking consiste à faire interagir une petite molécule organique avec le récepteur, généralement de nature protéique. Des études ont montré que certains algorithmes de docking sont plus fiables que d'autres pour reproduire le mode de fixation expérimentale de ligands (GLIDE,[20] GOLD[21]).[22] La contrepartie de ces techniques est généralement une hausse des temps de calcul. Par conséquent, il est nécessaire de se focaliser sur l'objectif du projet. Si le but est de cribler une chimiothèque d'une dizaine de composés, l'utilisation de ces méthodes lentes et précises est recommandée. A l'inverse, un projet impliquant le criblage virtuel de millions de produits ne pourra pas être accompli avec ce type d'algorithme mais plutôt des codes plus simples, dans lesquels les approximations engendrent un gain de temps de calcul (LigandFit,[23] FlexX[24,25]). Malgré tout, une étude de Warren *et al* ont montré que les algorithmes de docking précédemment évoqués sont capables de générer une conformation similaire à celle déterminée expérimentalement par cristallographie.[26]

a) La recherche conformationnelle

La principale préoccupation de l'opération de docking est de prendre en compte la flexibilité des deux entités. Tout algorithme de docking requiert en amont une recherche conformationnelle la plus exhaustive possible de la molécule organique. La génération et le choix de l'ensemble de conformères peuvent être effectués de plusieurs manières.

Lorsque l'on considère deux vecteurs A et B dans l'espace, ils peuvent subir translations et rotations dans les trois dimensions. Lorsque le nombre de liaisons simples (surtout de type acyclique) croît, le nombre théoriquement envisageable de conformations pour une molécule organique augmente. Une recherche conformationnelle exhaustive est donc impossible. Plus le degré de liberté de l'entité chimique augmente, plus l'espace

[20] Schrödinger, Portland, OR 97201
[21] Jones, G.; Willett, P.; Glen, R.-C.; Leach, A.-R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking *J. Mol. Biol.* **1997**, *267*, 727-748.
[22] Kontoyianni, M.; McClellan, L.; Sokol, G.S. Evaluation of docking performance: comparative data on docking algorithms J.Med.Chem. **2004**, *47*, 558-565.
[23] Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldan, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003***, 21,* 289-307.
[24] Rarey, M. ; Kramer, B. ; Lengauer, T. ; Klebe, G. A fast flexible docking method using an incremental construction algorithm *J. Mol. Biol.* **1996**, *261*, 470-489.
[25] http://www.biosolveit.de/FlexX/
[26] Warren, G.L.; Andrews, C.W.; Capelli, A.M.; Clarke, B.; LaLonde, J.; Lambert, M.; Lindvall, M.; Nevins, N.; Semus, S.F.; Senger, S.; Tedesco, G.; Wall, I.D.; Woolven, J.M.; Peishoff, C.E.; Head, M.S. A Critical Assessment of Docking Programs and Scoring Functions *J. Med. Chem.* **2006**, *49*, 5912-5931.

conformationnel s'accroît et plus l'approximation est importante. Par conséquent, il est raisonnable de travailler en présence de molécules dont le degré de liberté (nombre d'atomes et de liaisons sujettes à la rotation) est limité. Cette limite sera fixée en fonction de la puissance de calcul disponible. Pour cette raison, les molécules trop flexibles seront proscrites de nos études.

De multiples stratégies ont été élaborées pour déterminer les conformations de plus basse énergie. Nous avons développé deux principales approches qui nous ont paru intéressantes à confronter: l'approche classique et l'approche rationnelle.

(1) L'approche «classique» de recherche conformationnelle

L'approche dite «classique» regroupe les méthodes selon lesquelles un jeu de conformères est généré sans prendre en compte explicitement les acides aminés du site actif. Par contre, afin de mimer l'intérieur de la protéine, la molécule à analyser est explorée dans certaines conditions physico-chimiques d'hydrophilie et de lipophilie. La difficulté est de définir les proportions en caractère hydrophile ou lipophile à l'intérieur du site actif, qui varient selon la protéine étudiée. Ceci rend donc l'optimisation des paramètres délicate pour la recherche de conformères de plus basse énergie. Un autre élément à prendre en compte est le degré de flexibilité de la molécule étudiée. Plus celui-ci augmente, moins la recherche conformationnelle est exhaustive. Des algorithmes tels que DOCK[27] et FRED[28] utilisent cette approche «classique» de génération de conformères.

(2) L'approche « rationnelle» de recherche conformationnelle

Cette approche est un moyen adapté pour générer les conformères de plus basse énergie en tenant compte de la protéine. Plus précisément, ces techniques de construction incrémentale sont réalisées sous l'influence de la protéine. De ce fait, l'espace conformationnel se trouve réduit et, par la même occasion, les solutions finales sont moins nombreuses et parfois plus pertinentes. La limitation de ces méthodes est le choix du fragment de base servant d'amorce à la construction. L'emplacement choisi pour celui-ci conditionne la

[27] Kuntz, I.D.; Blaney, J.M.; Oatley, S.J.; Langridge, R.; Ferrin, T.E. A geometric approach to macromolecule-ligand interactions *J. Mol. Bio.* **1982**, *161*, 269-288.
[28] http://www.eyesopen.com/docs/html/fred/

direction du déplacement dans l'espace conformationnel qui va être entreprise. Chaque ajout doit satisfaire les contraintes imposées par le site de fixation de la protéine. Cette méthode rationnelle utilise généralement une fonction de scoring qui guide chaque étape de la construction du ligand. Une autre limitation de l'approche rationnelle est l'adaptation des fonctions de scoring qui sont généralement prévues pour une molécule entière et non pas pour des fragments moléculaires. LUDI utilise ce principe réutilisé, par la suite, par FlexX.[29]

De ces deux méthodologies, pas une n'est supérieure intrinsèquement que l'autre. Dans l'approche classique, la recherche conformationnelle se fait sans influence du site actif. Par conséquent, l'ensemble des conformations est théoriquement plus exhaustif.

Si maintenant nous utilisons une approche rationnelle, la recherche conformationnelle est plus ciblée et répond précisément aux critères nécessaires à la formation du complexe ligand/récepteur. L'espace conformationnel est alors plus restreint. Les deux approches sont, dans l'absolue, défendables. Toutefois, en gardant à l'esprit que le but de notre étude est de mimer le mode de fixation de ligands co-cristallisés dans un site actif, une approche rationnelle faisant intervenir explicitement les acides aminés nous a semblé plus adaptée.

D'autres méthodes sont également utilisées telles que les algorithmes stochastiques. Par exemple, la technique de Monte Carlo,[30] dans laquelle la conformation et le positionnement du ligand sont optimisés en même temps par perturbation aléatoire des variables définissant les degrés de liberté du complexe ligand/protéine, est utilisée dans le programme Autodock.[31] Des algorithmes qui ont également un avenir sont ceux qui se basent sur les règles de sélection génétique. Plus communément connus sous le nom d'algorithmes génétiques, ces programmes codent chaque type de degré de liberté dans un chromosome. Des mutations, crossovers ou encore migrations chromosomales sont nécessaires pour faire évoluer les populations vers des solutions au sein d'un espace complexe. GOLD utilise notamment cette approche.

Comme nous venons de le voir, de multiples stratégies peuvent être envisagées pour simuler l'espace conformationnel d'un ligand. Cet aspect est évidemment primordial. Cependant, pour être plus précis, la dynamique de la protéine doit être également considérée. C'est aujourd'hui un challenge que d'inclure la flexibilité d'une telle entité, dont les degrés de liberté sont à notre échelle infinis. Malgré tout, certains algorithmes ont développé ce volet

[29] Böhm, H.J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593-606.
[30] Young, D.C.; Computational chemistry: a practical guide for applying techniques to real-world problems, Weiley interscience, New York **2001**, 179-192.
[31] Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated Docking of Flexible Ligands: Applications of AutoDock *J. Mol. Recognition*, **1996**, *9*, 1-5.

dans leurs programmes de docking mais de manière ponctuelle et localisée sur certains acides aminés (FlexE,[32] GLIDE, GOLD).

b) La flexibilité de la protéine

Le concept ancien de la clé et de la serrure, figurant respectivement la molécule organique et la protéine, est aujourd'hui obsolète. Une image plus adaptée d'un tel complexe est celui d'une main dans un gant (respectivement la molécule organique dans la protéine). Chacun peut modifier sa forme, favoriser la complémentarité afin de limiter les interactions gênantes et augmenter le confort procuré par le contact entre les deux entités. La difficulté dans la majorité des protéines est qu'il existe toujours des zones très flexibles à proximité de zones rigides. Les zones flexibles correspondent généralement à des boucles verrouillant le ligand. Quant aux zones rigides, elles correspondent majoritairement à des sites catalytiques.

La principale faiblesse des processus de docking est de ne pas considérer les mouvements inhérents aux protéines. La dynamique d'acides aminés effectuée par FlexE, GLIDE et GOLD, par exemple, n'est qu'une approximation de ce qui peut avoir lieu en réalité. Des méthodes plus rigoureuses impliquant la dynamique moléculaire permettent d'appréhender la flexibilité des protéines. Nous verrons dans cette étude l'intérêt de combiner, à l'aide de stratégies de consensus, les conformères générés par dynamique moléculaire. La principale préoccupation de cette technique est le temps de calcul qui limite l'utilisation de telles méthodes lors de criblage virtuel de millions de produits chimiques.

c) Simulation de la présence d'eau

Certains algorithmes sont capables de simuler la présence de molécules d'eau dans les cavités des protéines. Par exemple, FlexX laisse l'opportunité de simuler l'existence de molécules d'eau dans l'environnement du site actif[33] (concept de «particle»). La présence d'eau est parfois primordiale pour assurer un relais entre le ligand et le site actif.[34] Dans l'algorithme GOLD, les molécules d'eau peuvent être activées ou non et sont soumises à un

[32] http://www.biosolveit.de/FlexE/
[33] Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: Placing discrete water molecules during protein-ligand docking predictions *Drug Discov. Today* **1999**, *34*, 17-28.
[34] Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006**, *11*, 580-594.

mouvement de rotation autour du ligand.[35] Nous n'avons pas mis en évidence dans nos deux projets thérapeutiques l'intérêt de considérer ce paramètre.

3. Le processus de scoring

a) Principe

Le score est une donnée numérique utile pour quantifier le degré avec lequel un ligand se complexe à un récepteur. C'est globalement une approximation de l'énergie libre résultant du passage de la forme libre de la protéine et du ligand à l'association sous forme de complexe. Le principe thermodynamique est le suivant:[36]

$$\Delta G = \Delta G_{complexe} - \Delta G_{ligand} - \Delta G_{protéine}$$

Concrètement, le score est une estimation de l'affinité entre la macromolécule et la petite molécule organique. Un score ne prédit donc en rien une activité mais bien une affinité. Il n'est donc pas réaliste de corréler une activité mesurée avec la valeur d'un score. Il est plus judicieux d'établir une corrélation des constantes d'inhibition K_i avec les scores *in silico*. Une mauvaise corrélation entre les affinités expérimentales et les scores ne signifie pas pour autant que la fonction de scoring est mauvaise. En effet, les molécules dont on connaît les affinités mesurées expérimentalement ne couvrent qu'un très faible intervalle de score comparé au large intervalle proposé par une fonction de scoring.[37]

L'utilisation des fonctions de scoring est double. Tout d'abord, elles permettent de déterminer la conformation qui représentera au mieux le ligand concerné. Cette conformation est appelée «première pose». L'autre utilisation des scores est de pouvoir classer les premières poses de chaque ligand afin d'établir un classement final des molécules les plus prometteuses. Malgré tout, il subsiste encore beaucoup d'inconnues dans le mécanisme de reconnaissance lors de la formation du complexe tels que la formation de liaisons hydrogène, les termes entropiques ainsi que le rôle à jouer des molécules d'eau dans le processus de solvatation-

[35] Verdonk, M.L.; Chessari, G.; Cole, J.C.; Hartshorn, M.J.; Murray, C.W.; Nissink, J.W.M.; Taylor, R.D.; Taylor, R. Modeling Water Molecules in Protein-Ligand Docking Using GOLD *J. Med. Chem.* **2005**, *48*, 6504-6515.

[36] Kollman, P.A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Domini, O.; Cieplak, P.; Srinivasan, J.; Case, D.A.; Cheatham, T.E.; Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models *Acc. Chem. Res.* **2000**, *33*, 889-897.

[37] Seifert, M. H. J. Assessing the Discriminatory Power of Scoring Functions for Virtual Screening *J. Chem. Inf. Model.* **2006**, *46*, 1456-1465.

désolvatation. Nous avons montré que le docking est accompagné d'une multitude d'approximations. Il en va de même pour les fonctions de scoring pour lesquelles il est impossible d'évaluer toutes les interactions intra et interatomiques. Bien que ces approximations existent, ces méthodes sont utiles pour nos applications de criblage virtuel. Nous avons utilisé dans cette étude principalement les fonctions de scoring provenant du module C-Score. Celui-ci compte cinq fonctions de scoring dont FlexX Score, Chemscore, PMF score, Dock Score et Gold Score.

b) Familles de fonctions de scoring

Les fonctions de scoring peuvent être classées dans deux grandes catégories: les fonctions empiriques et les fonctions basées sur la connaissance (encore appelées «knowledge-based»). Ces deux classes de fonctions sont basées sur un ensemble de complexes ligand-protéine.[38] Elles ont donc la faiblesse de ne pouvoir prédire que des interactions qui ne dévient pas trop de celles ayant été répertoriées dans les complexes étudiés.

(1) Les fonctions de scoring empiriques

Les termes représentant chacun des types d'interactions connues dans le complexe (liaison hydrogène, contact hydrophobe, interaction ionique, surface de contact ligand-protéine et parfois les contributions entropiques) sont additionnées. La pondération de chaque terme est déterminée par des méthodes de régression multivariées appliquées à un *ensemble d'apprentissage*. Les scores provenant de FlexX, DOCK, GOLD et Chemscore sont issus de fonctions de type empirique. Cinq principaux critères justifient le choix d'une fonction:

-L'exactitude, c'est-à-dire qu'une fonction de scoring doit proposer une précision acceptable dans un contexte défini.

-Le domaine d'application: il doit être assez étendu pour pouvoir permettre de prédire des composés par interpolation et extrapolation. Le modèle prédictif ne doit pas être trop restrictif. Des méthodes linéaires permettent généralement d'obtenir un modèle capable d'extrapoler et d'interpoler. Les modèles issus de méthodes non linéaires (réseaux de neurones par exemple)

[38] Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening *J. Med. Chem.* **2001**, *44*, 1035-1042.

ont parfois l'inconvénient de trop suivre les points ayant servi à leur élaboration et sont donc incapables d'interpoler et d'extrapoler.

-La robustesse de la méthode: dans des applications de QSAR, un problème majeur existe lorsque l'on applique les équations à des molécules différentes de l'*ensemble d'apprentissage*. C'est précisément ce qui peut se passer dans le cas des fonctions de scoring puisque l'on ne possède généralement que la structure tridimensionnelle des molécules les plus affines, dans un contexte où l'on souhaite appliquer les fonctions de scoring à des molécules toutes différentes les unes des autres. Toutefois, cette faiblesse peut être améliorée en élargissant les bases de données de complexes récepteur-ligand mais également en utilisant des méthodes de validation croisée afin de s'assurer que les coefficients de corrélation sont stables.

-La vitesse à des fins de criblage virtuel: les fonctions de scoring doivent être suffisamment rapides pour être appliquées au criblage virtuel. Toutefois, le facteur limitant n'est pas les fonctions de scoring mais le processus de docking.

-L'interprétation physique: il est souhaitable que le nombre de termes employés dans la régression soit le plus faible possible. L'ajout d'un nouveau terme doit faire la preuve d'un véritable bénéfice. Ceci est nécessaire pour obtenir une équation simple, facilement interprétable de manière physique.

(2) Les fonctions de type knowledge-based

Ces fonctions proviennent de l'analyse des structures tridimensionnelles de complexes ligand-protéine déterminés de manière expérimentale. Des règles définissant la géométrie préférentielle des interactions sont déduites de ces structures grâce à des moyens statistiques. Cette alternative aux fonctions empiriques est plus tolérante quant aux interactions présentes au sein du complexe. Leurs expressions sont moins strictes que dans le cas des fonctions empiriques. La fonction PMF fait partie de cette classe de fonction.

c) Description des fonctions de scoring utilisées

(1) La fonction de scoring DOCK

Dock-Score, élaborée par Kunst,[27] dérive directement du champ de force AMBER[39]. Cette fonction est composée de deux termes électrostatique et stérique.

[39] Case, D.A.; Pearlman, D.A.; Caldwell, J.W.; Cheatham, T.E.; Wang, J.; Ross, W.S.; Simmering, C.L.; Darder,

$$\text{Dock-Score} = E_{\text{électrostatique}}(d) + E_{\text{stérique}}(d)$$

où E est l'énergie dépendante de la distance (d) entre deux points susceptibles d'interagir.

Le terme électrostatique est de type potentiel de Coulomb dont les principales variables sont la charge et la distance les séparant. La constante diélectrique est fixée à 4, mimant ainsi la tendance hydrophobe d'un site actif. Le terme stérique est de type Lennard-Jones 6-12 (facteurs attractifs et répulsifs respectivement).[40]

(2) La fonction de scoring GOLD

Gold-Score est issue des travaux de Willet.[21] Cette fonction est la somme des énergies de stabilisation provenant des liaisons hydrogène, de l'énergie interne de Van der Waals pour la conformation du ligand en question ainsi que la force des interactions stériques entre le ligand et le récepteur.

$$\text{Gold-Score} = E_{\text{liaison H}}(d,\alpha) + E_{\text{stérique}}(d) + E_{\text{Van der Waals}}$$

Les deux premiers termes (énergie de liaison hydrogène et stérique) sont pondérés par la distance entre les deux entités appartenant au ligand et au récepteur. L'énergie de liaison hydrogène est également pondérée par l'angle entre les deux entités. Des pénalités sont apportées lorsque la liaison ou le contact stérique sont en dehors d'un certain angle ou d'une distance fixée.

(3) La fonction de scoring FlexX

Cette fonction dérive de l'équation de Böhm qui est bien plus complexe que les deux fonctions de scoring décrites auparavant.[41]

T.A.; Merz, K.M.; Stanton, R.V.; Cheng, A.L.; Vincent, J.J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R.J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G.L.; Singh, U.C.; Weiner, P.K.; Kollman, P.A. AMBER 7. University of California, San Fransisco, **2002**.

[40] Lennard-Jones, J. E. Cohesion. *Proceedings of the Physical Society* **1931**, *43*, 461-482.

[41] Böhm, H-J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure *J. Comput. Aided Mol. Des.* **1994**, *8*, 243-256.

$$\text{F-Score} = \Delta G_0 + \Delta G_{rot}\, N_{rot} + \Delta G_{liaison\ H}\,(\Delta R,\Delta\alpha) + \Delta G_{liaison\ ionique}\,(\Delta R,\Delta\alpha) + \Delta G_{aromatique}\,(\Delta R,\Delta\alpha) + \Delta G_{lipophilie}\,(\Delta d)$$

Cette équation considère la baisse d'entropie lors de la fixation du ligand dans le site actif. Les interactions sont prises en compte dans cette formule (liaisons hydrogène, liaisons ioniques, contacts aromatiques). Un terme prend également en compte la lipophilie. Les écarts d'angle et de distance sont impliqués dans une fonction pénalisant les termes décrits ci-dessus. Plus l'écart entre la géométrie de la liaison observée et la géométrie idéale est important, plus le terme est pénalisé et moins son énergie ΔG sera importante. Le terme lipophilique évalue les contacts atomiques entre la protéine et le ligand (interactions favorables hydrophobes et contacts déstabilisants).

(4) La fonction de scoring Chemscore

Chemscore a été élaborée par le groupe de Elbridge *et al*.[42] Elle prend également en considération des termes similaires aux autres fonctions jusque là décrites. Elle prend en compte l'énergie de liaison hydrogène, les interactions métal-ligand, les contacts lipophiles et enfin un terme prenant en considération l'entropie en fonction de la flexibilité de la molécule:

$$\text{Chemscore} = \Delta G_0 + \Delta G_{liaison}\ H\,(d,\alpha) + \Delta G_{métal}\,(d) + \Delta G_{lipophilie} + \Delta G_{rot} H_{rot}$$

L'énergie de liaison hydrogène est pondérée par la distance et l'angle entre les entités «donneur» et «accepteur». Le terme correspondant à la liaison du ligand avec un métal est pondéré par la distance entre l'atome du ligand et le métal.

(5) La fonction de scoring PMF

PMF (Potential Mean Force) est une fonction basée sur la connaissance (knowledge-based) mise au point par Muegge.[43,44] Cette approche exploite l'information structurale

[42] Eldridge, M.D.; Murray, C.W.; Auton, T.R.; Paolini, G.V.; Mee, R.P. Empirical scoring functions: The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes *J. Comput. Aided Mol. Des.* **1997**, *11*, 425-445.
[43] Muegge, I.; Martin, Y.C. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach *J. Med. Chem.* **1999**, *42*, 791-804.
[44] Muegge, I. PMF Scoring Revisited *J. Med. Chem.* **2005**, *49*, 5895-5902.

extraite de la Protein Data Bank et la convertit en énergie libre d'interaction entre paires d'atomes protéine-ligand. Les effets entropiques, enthalpiques et de solvatation sont traités de manière implicite. Une étude comparative sur les algorithmes de docking et les fonctions de scoring met en évidence que PMF est une fonction qui fonctionne bien dans le cas de cavité très polaires. Nous verrons plus tard que ceci peut être une explication de la défaillance de PMF dans nos deux projets, dont le site actif est majoritairement à tendance lipophile.[45]

d) Types d'interactions

Dans les deux projets COX-2 et PPARγ que nous exposons plus tard, les cavités sont majoritairement de nature hydrophobe, avec peu d'interactions directionnelles. Nous souhaitons prédire le comportement des fonctions de scoring décrites précédemment face à ces deux cibles dont la tendance physico-chimique est similaire.

(1) La lipophilie

La plupart des fonctions de scoring développées jusqu'à présent sont déficientes en termes capables d'évaluer les contacts lipophiles présents au sein d'un complexe. Toutefois, des fonctions comme FlexX-Score ou Chemscore incorporent dans leur formule un terme traitant ce genre d'interaction. Dock inclue également un terme lipophilique. Ces contacts n'étant pas localisés (comme pourraient l'être les interactions électrostatiques) mais dispersés entre le ligand et les acides aminés de la cavité, leur évaluation est délicate.

(2) Les interactions électrostatiques

Certaines fonctions de scoring sont connues pour surévaluer ces interactions. Cette tendance a été mise en évidence avec FlexX-Score.[38] L'inconvénient est qu'une molécule possédant un nombre de groupements réactionnels important est capable de récolter un score acceptable alors qu'elle n'est intrinsèquement pas affine. De même, la fonction Gold Score est essentiellement basée sur les paires donneur-accepteur de liaisons hydrogène. Chemscore et Dock Score possèdent également un terme évaluant les interactions électrostatiques au sein du complexe.

[45] Bissantz, C. ; Folkers, G. ; Rognan, D. Protein-based virtual screening of chemical databases. Evaluation of different docking/scoring combinations *J. Med. Chem.* **2000**, *43*, 4759-4767.

(3) Les termes de solvatation et d'entropie

Deux facteurs encore mal pris en compte par les fonctions de scoring sont l'entropie et la solvatation. Certaines fonctions telles que Chemscore et FlexX-Score considèrent l'aspect entropique. Ce type de phénomène est censé être mieux traité avec les fonctions de type «knowledge-based» comme PMF. Cette fonction considère également l'effet de solvatation.

(4) Conclusion sur l'utilisation des fonctions avec COX-2 et PPARγ

FlexX-Score et Chemscore réunissent les caractéristiques requises à l'analyse de complexes avec COX-2 et PPARγ. Chacune incorpore un terme mesurant la lipophile, les interactions électrostatiques et l'entropie.

4. Interprétation des fonctions de scoring

Comme nous l'avons montré précédemment, toute fonction de scoring est optimisée sur un ensemble de complexes d'apprentissage. Il n'existe donc *a priori* pas de fonction qui pourrait être efficace dans un large spectre de protéine. Chaque fonction a sa spécificité, son domaine d'action et traite plus ou moins bien les différents termes spécifiques à la reconnaissance des deux entités (liaison hydrogène, contact hydrophobe, phénomène de solvatation-désolvatation). Il est donc indispensable d'utiliser plusieurs fonctions de scoring, si possible ne provenant pas d'un même *ensemble d'apprentissage* ou n'étant pas dérivées d'une fonction déjà utilisée. Enfin, il est également préférable d'avoir au moins une fonction provenant de chacune des deux catégories de fonction de scoring (fonction empirique et fonction basée sur la connaissance). Dans notre étude, nous avons testé toutes les fonctions de scoring utilisées afin de déterminer les plus performantes, avec comme but de les inclure dans le modèle final prédictif.

Une stratégie qui s'est développée depuis quelques années consiste à regrouper l'information de plusieurs fonctions de scoring. Une méthode très intéressante appelée consensus, consistant à combiner des fonctions de scoring, a été décrite en 1999 par Charifson *et al*.[46] Ils la décrivent comme capable d'une meilleure discrimination entre des composés

[46] Charifson, P.S.; Corkery, J.J.; Murcko, M.A.; Walters, W.P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins *J. Med. Chem*. **1999**, *42*, 5100-5109.

actifs et inactifs en présence de plusieurs fonctions de scoring. Ces études comparant l'utilisation de fonctions seules avec des méthodes de consensus ont mis en évidence significativement la réduction du nombre de faux positifs par consensus de fonctions. Cette stratégie permet de combiner les avantages de fonctions différemment optimisées et élaborées.[47] Les faiblesses sont également cumulées. Ainsi, un consensus n'a de sens que lorsqu'il comporte un nombre optimisé de fonctions de scoring (celles dont les défauts sont trop importants seront éliminées du modèle). Il vise à utiliser l'information des fonctions les plus performantes pour un projet donné. Nous avons également introduit dans notre étude une technique encore peu utilisée dans le domaine: l'analyse factorielle discriminante (AFD).[48] Nous verrons comment elle permet d'exploiter l'information contenue dans la matrice multidimensionnelle de résultats. Le but de ces deux méthodes (consensus et AFD) est de limiter le nombre de molécules étant prédites actives alors qu'elles ne le sont pas. Ces composés indésirables sont encore appelés «faux positifs».

B. Technique basée sur la structure des ligands: le pharmacophore

Un pharmacophore est un ensemble de points caractéristiques que doit posséder une molécule pour être active. Un modèle pharmacophorique est un arrangement spatial de points résultant de l'alignement de molécules dont le mode de liaison à un récepteur est connu ou supposé. A chaque point de ce modèle correspond des coordonnées 3D, un volume (généralement sphérique) ainsi que des propriétés physico-chimiques (lipophilie, donneur/accepteur de liaisons hydrogène). Son élaboration passe tout d'abord par la sélection de composés actifs pour la cible en question. Des conformations sont ensuite générées et le choix hypothétique de l'une d'elle est effectué. Un alignement des poses sélectionnées permet par la suite de déduire les points significatifs responsables de l'affinité avec le récepteur.

L'intérêt de coupler cette technique au docking est d'une part de pouvoir éliminer directement les molécules indésirables et donc d'économiser du temps de calcul. Par ailleurs, une telle stratégie permet de guider la construction incrémentale du ligand au sein du site actif.

[47] Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1422-1426.
[48] Lebart, L.; Morineau, A.; Piron, M. *Statistique Exploratoire Multidimensionnelle*, Dunod, **2000**.

C. Dynamique moléculaire

1. Principe

Les simulations par dynamique moléculaire sont majoritairement utilisées pour étudier l'espace conformationnel des macromolécules biologiques. Elles sont valables pour mieux appréhender le comportement des protéines dans une période de temps donnée. Il est également possible d'étudier l'effet de molécules de solvant explicites sur la structure protéique. Ces méthodes ne sont pas uniquement utilisées pour rationaliser l'utilisation de structures issues de mesures expérimentales, mais sont également appliquées pour raffiner la plupart des structures issues de la cristallographie par rayons X et de la RMN. L'augmentation de la puissance informatique a permis d'allonger le temps de simulation qui est passé en une vingtaine d'années de la pico seconde à la nano seconde et parfois même à la micro seconde. Des progrès également dans les interfaces (membrane simulée, solvant explicite) permettent de mimer l'évolution du système dans un environnement plus complexe que le vide. De meilleurs champs de force ont par ailleurs vu le jour, impliquant un meilleur traitement des interactions électrostatiques longue distance. Toutefois, certains problèmes peuvent gêner l'utilisation de ces méthodes, en particulier le piégeage du complexe dans un minimum local qui ne serait pas représentatif de l'espace conformationn el dans lequel évolue l'ensemble. Les programmes couramment utilisés dans la simulation des biomolécules par dynamique moléculaire sont AMBER,[39] CHARMM,[49] GROMOS[50] et NAMD[51].

2. Intérêt de la dynamique moléculaire en amont du docking

La faiblesse majeure des algorithmes de docking est l'absence ou le peu de prise en compte de la flexibilité de la protéine lors de l'opération d'arrimage entre le ligand et la protéine.[52] Ceci ne permet donc pas une co-adaptation optimale du ligand et du récepteur. La dynamique moléculaire est capable de traiter la flexibilité de la protéine. Par conséquent, le

[49] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B.D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations *J. Comp. Chem.* **1983**, *4*: 187-217.

[50] van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. Biomolecular Simulation: The GROMOS96 Manual and User Guide; vdf Hochschulverlag AG an der ETH Zürich and BIOMOS b.v.: Zürich, Groningen, **1996**.

[51] Nelson, M.; Humphrey, W.; Gursoy, A.; Dalke, A.; Kale, L.; Skeel, R. D.; Schulten , K. NAMD - a Parallel, Object-Oriented Molecular Dynamics Program *International Journal Supercomputing Applications and High Performance Computing* **1996***, 10.*

[52] Rester, R. Dock around the Clock – Current Status of Small Molecule Docking and Scoring *QSAR Comb. Sci.* **2006**, *25*, 605-615.

couplage des deux puissantes techniques que sont le docking et la dynamique moléculaire doit théoriquement augmenter le pouvoir prédictif de notre modèle.

D. Les chimiothèques dédiées au criblage virtuel

1. Généralités

Le criblage virtuel est un processus permettant la découverte de molécules nouvelles et novatrices. Le choix des composés à cribler doit donc être pertinent. Hann et Oprea ont défini quatre types d'espaces chimiques:[53]

- La chimiothèque réelle: elle provient d'un organisme public ou privé (limite supérieure actuelle de l'ordre de 10 millions de composés)[54]
- La chimiothèque globale: elle regroupe tous les composés de chimiothèques réelles existantes (estimée comme supérieure à 80 millions de composés)
- La chimiothèque tangible: elle regroupe des composés faciles à synthétiser ou à obtenir auprès des fournisseurs de produits chimiques (supérieure à 10^{20} produits chimiques)[55]
- La chimiothèque virtuelle: elle est constituée de molécules théoriquement synthétisables (estimation à 10^{60}).

Le criblage virtuel sur chimiothèques réelles ou tangibles est celui qui présente le plus d'intérêt puisque son approvisionnement en composés est, la plupart du temps, envisageable. Des bases prêtes à l'emploi sont également disponibles, préalablement filtrées et parfois même optimisées pour le docking. C'est notamment le cas de la base ZINC qui propose des structures destinées au docking.[56] Toutefois, même si ces chimiothèques prêtes à l'utilisation peuvent paraître confortables et standardisées, il n'en est pas moins qu'elles présentent certaines interrogations et parfois même quelques réserves. En effet, quels sont les types de filtres appliqués pour les obtenir? Les molécules sont-elles disponibles chez le fournisseur de produits chimiques? La sélection des composés inclut-elle un algorithme de diversité? Les critères de sélection de ces composés sont-ils en accord avec le projet à traiter ? Nous verrons

[53] Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* **2004**, *8*, 255-263.
[54] Schuffenhauer, A.; Popov, M.; Schopfer, U.; Acklin, P.; Stanek, J. Jacoby, E. Molecular Diversity Management Strategies for Building and Enhancement of Diverse and Focused Lead Discovery Compound Screening Collections. *Comb Chem High Throughput Screen* **2004**, *7*, 771-781.
[55] Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374-380.
[56] http://zinc.docking.org/

dans cette étude que de générer une chimiothèque selon les règles de Lipinski (règle des 5)[57] n'a parfois aucun sens. Ce filtrage peut mener à bien des pertes de composés qui ont sans doute un potentiel intéressant vis-à-vis d'un récepteur donné. C'est particulièrement le cas d'une des cibles pharmacologiques exposées dans nos travaux.

2. Chimiothèque locale

Nous avons à disposition dans le laboratoire une chimiothèque tangible d'environ 3 millions de molécules provenant de 38 fournisseurs de produits chimiques. Notre base compte des molécules identifiées comme uniques par le code «InChi»[58]. Tous les composés sont accessibles *via* un ou plusieurs fournisseur(s) par produits. L'origine des composés est variée (hémi-synthétique, synthétique, naturelle). D'autres compilations de bases de molécules sont disponibles sur le marché dont une des plus importantes est celle de Chemnavigator[59] avec plus de 37 millions de molécules.

3. Filtres physico-chimiques

Les composés considérés indésirables pour un projet donné sont filtrés. Dans notre cas, ces filtres ont été élaborés de manière spécifique à chaque projet. Les propriétés moléculaires telles que la masse moléculaire, le LogP (qui regroupe la notion de solubilité aqueuse ainsi que la lipophilie), la proportion d'atomes donneurs et accepteurs, la présence/absence d'une fonctionnalité organique et le nombre de liaisons sujettes à la rotation sont les principaux critères de choix. Bien sûr, cette liste n'est pas exhaustive.

4. Préparation des molécules

a) Etats de protonation

L'état de protonation est un paramètre difficile à évaluer.[60] En effet, de multiples états sont parfois envisageables pour une même molécule au pH physiologique (pH=7,4). Il n'y a

[57] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **2001**, *46*, 3-26.
[58] Stein, S.E.; Heller, S.R.; Tchekhovskoi, D. An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier *International Chemical Information Conference (Nimes), Infonortics* **2003**, 131-143.
[59] http://www.chemnavigator.com/
[60] Mohan, V.; Gibbs, A.C.; Cummings, M.D.; Jaeger, E.P.; DesJarlais, R.L. Docking: Successes and Challenges *Curr. Pharm. Des.* **2005**, *11*, 323-333.

pas à ce jour de méthodes fiables permettant d'estimer l'état de protonation d'une molécule à un pH donné. De plus, le pH au sein d'une protéine peut varier d'une région à l'autre du site actif et donc peut différer du pH mesuré dans les conditions d'utilisation des médicaments. Par conséquent, nous avons choisi de déprotoner les acides et de protoner les bases, ce qui semble être une bonne moyenne de ce qui peut se passer dans le site de fixation des ligands.

b) Raffinement de la structure des molécules

Pour toutes les techniques décrites précédemment (aussi bien les méthodes basées sur la structure que celles basées sur les ligands), la structure tridimensionnelle est un élément essentiel. Plus précisément, la conversion d'une molécule dans un format bidimensionnel vers une structure tridimensionnelle doit être réalisée avec précaution. Plusieurs algorithmes sont disponibles pour accomplir cette transition tels que Corina[61] et Concord[62] pour les plus connus. Une fois le format tridimensionnel obtenu, une minimisation est nécessaire pour faire parvenir l'état énergétique de la molécule au plus bas possible. Dans notre étude, nous avons utilisé MMFF94s.[63]

[61] http://www.molecular-networks.com/software/corina/index.html.
[62] www.tripos.com/index.php?family=modules,SimplePage,sybyl_concord.
[63] Halgren, T.A. MMFF VI. MMFF94s Option for Energy Minimization Studies *J. Comput. Chem.* **1999**, *20*, 720-729.

CHAPITRE 1.

Elaboration d'un protocole de docking-scoring pour le criblage virtuel.

Application à l'enzyme COX-2.

Dans ce projet, nous avons exploité l'information structurale disponible sur l'enzyme cyclooxygénase de type 2. Le but de ce travail est de développer un modèle prédictif capable d'identifier des candidats médicaments pour la cible pharmacologique COX-2 en s'appuyant sur la dernière lignée d'inhibiteur conçue ces cinq dernières années. Nous avons développé une stratégie de docking-scoring assisté par différents outils permettant une réduction du temps de calcul (modèle pharmacophorique) avec un souci de transparence quant aux méthodes d'interprétation de données de scoring. En effet, un ensemble de cinq fonctions de scoring a été déployé afin d'augmenter les qualités prédictives du modèle.

Tout d'abord, nous avons focalisé nos travaux sur la chimiothèque à utiliser pour réaliser le criblage virtuel. Un total de 32 fournisseurs de produits chimiques a permis d'obtenir 3,2 millions de molécules. Nous avons souhaité identifier des molécules innovantes ne possédant pas de groupement sulfonyle, majoritairement présent dans les récentes molécules spécifiques à COX-2. Des filtres élaborés sur les critères des molécules inhibitrices COX-2 ont été appliqués, réduisant à 1 million le nombre de produits convenables pour le criblage virtuel. Finalement, après un protocole de docking-scoring nous avons choisi de récupérer 20 molécules à fort potentiel inhibiteur et de les tester afin de valider notre méthodologie.

E. Introduction

Les prostaglandines sont d'importants médiateurs de l'inflammation dans l'organisme. L'enzyme prostaglandine synthase (encore appelée cyclooxygénase) est chargée de convertir l'acide arachidonique en plusieurs produits dont les prostaglandines PGG_2.[64] Deux isoformes ont été principalement identifiés et décrits. La cyclooxygénase de type 1 (COX-1) est exprimée de manière constitutive dans l'organisme c'est-à-dire de façon permanente. Son inhibition conduit généralement à des évènements tels que des dérégulations rénales et des dysfonctionnements de la fonction ulcéro-gastrique. La cyclooxygénase de type 2 (COX-2) est un isoforme exprimé par l'action d'agents inducteurs. Cette forme de cyclooxygénase est responsable de la production élevée en prostaglandine lors du processus inflammatoire induisant fièvre, douleur et inflammation.[65,66] Un intérêt majeur dans le traitement de phénomènes inflammatoires est d'inhiber spécifiquement l'isoforme 2.

[64] van Ryn, J.; Pairet, M. Clinical experience with cyclooxygenase-2 inhibitors. *Inflamm. Res.* **1999**, *48*, 247-254.
[65] Dubois, R.N.; Abramson, S.B.; Crofford, L.; Gupta, R.A.; Simon, L.S.; Van De Putte, L.B.; Lipsky, P.E. Cyclooxygenase in biology and disease *fasebj* **1998**, *12*, 1063-1073.

Différentes familles d'inhibiteurs ont été élaborées. Ils sont appelés AINS (anti-inflammatoires non stéroïdiens). La première génération de ces composés n'offre pas de caractère spécifique et inhibe aussi bien COX-1 que COX-2 (naproxen, ibuprofen, diclofenac, indométhacine, nabumétone et aspirine). A faible dose, l'aspirine inhibe préférentiellement l'isoforme COX-1. Ceci a comme conséquence de réduire la production de thromboxane (agent vasoconstricteur et inducteur d'agrégation plaquettaire) et par conséquent de diminuer les risques de thrombose. En concentrations réduites, l'aspirine ne diminue pas la production de prostacycline (agent inhibant l'agrégation plaquettaire et inducteur de vasodilatation) provenant du COX-2. Ceci explique les bénéfices d'une administration à long terme de faibles doses d'aspirine chez les patients à haut risque, réduisant la fréquence d'accidents cardiovasculaires et d'infarctus.

Une nouvelle génération d'AINS a été mise sur le marché (Rofecoxib «Vioxx» et Celecoxib «Celebrex»). Ces composés sont décrits comme fortement sélectifs de l'isoforme COX-2. Cependant, le retrait du Vioxx (01/10/04) a positionné les nombreuses séries de dérivés du Vioxx dans une situation délicate. La théorie selon laquelle le COX-1 est le bon isoforme et le COX-2 le mauvais a été reconsidérée. Il est reproché aux composés trop sélectifs de l'isoforme 2 de supprimer totalement la concentration de prostacycline, agent vasodilatateur et inhibiteur de l'agrégation plaquettaire.

Dans cette étude, nous nous sommes intéressés à la recherche de molécules capables d'inhiber COX-2 sans prendre en compte la composante COX-1. La recherche de nouvelles entités chimiques inhibitrices de l'enzyme COX-2 est plus que jamais d'actualité.

1. Fonction de la cyclooxygénase

L'acide arachidonique est un acide gras endogène provenant de la dégradation des phospholipides *via* la phospholipase A2. La prostaglandine endoperoxide synthase (PGHS ou encore COX) est une enzyme bifonctionnelle capable de convertir des acides gras tels que l'acide arachidonique en prostaglandine G_2 (PGG_2) par voie de dioxygénation. Un site de péroxydation transforme la PGG_2 en PGH_2, précurseur de séries de prostaglandines participant, à dose normale, à la protection du tissu gastro-intestinal. Elle produit également des prostacyclines, inhibiteur des fonctions plaquettaires ainsi que des thromboxanes. Tout ceci participe, à dose élevée, au phénomène inflammatoire (Figure 4).

---

[66] van Ryn, J.; Pairet, M. Clinical experience with cyclooxygenase-2 inhibitors. *Inflamm. Res.* **1999**, *48*, 247-254.

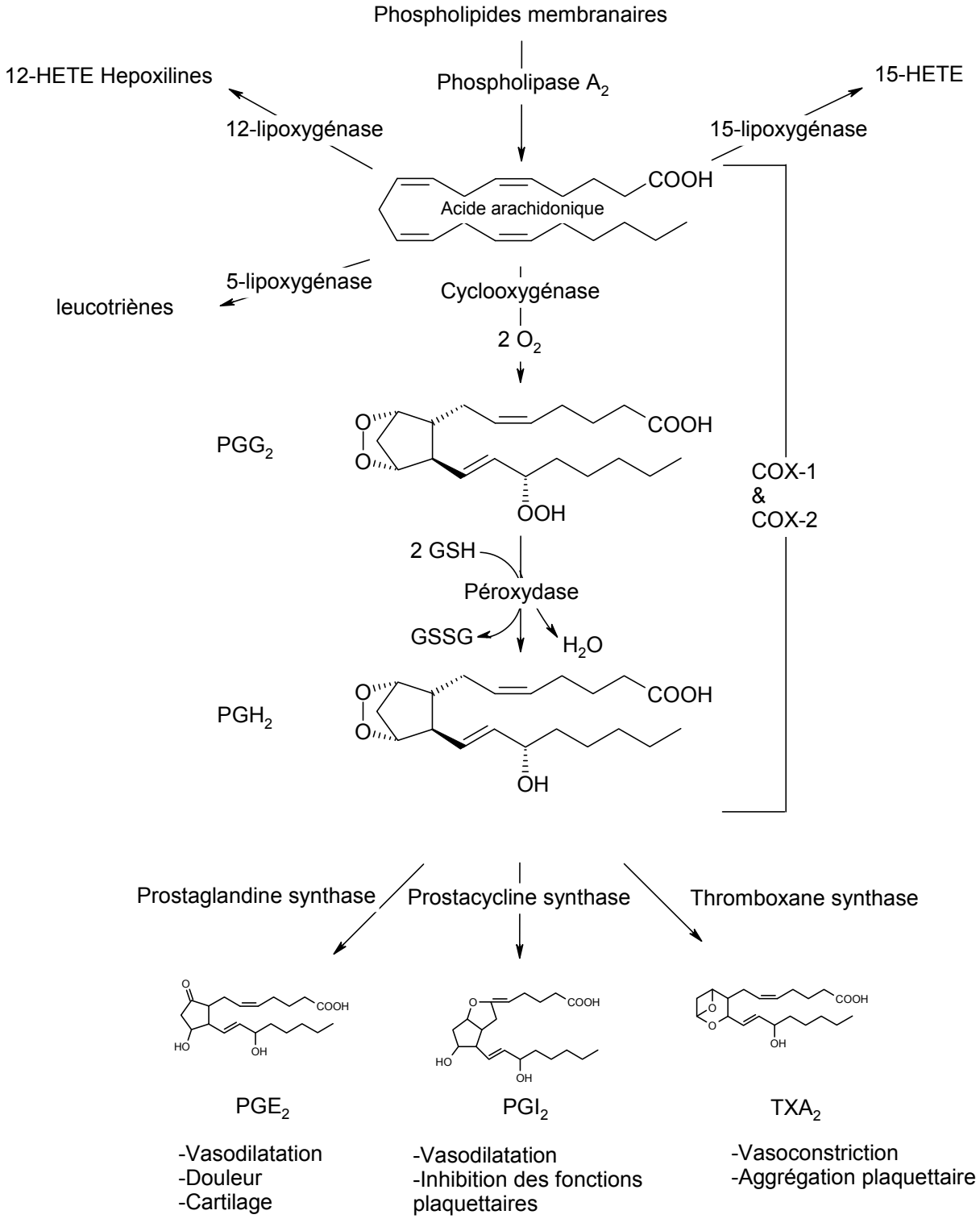


Figure 4. Cascade de l'acide arachidonique

La présence élevée de l'isoforme COX-2 est synonyme d'inflammation dans le tissu concerné. Son expression est généralement induite par des lipopolysaccharides ainsi que des cytokines pro-inflammatoires (interleukines, interféron$_\gamma$). Des études ont également montré

l'existence d'un troisième isoforme, la COX-3. Cependant, celle-ci ne semble pas être capable de produire de prostaglandines ayant une activité dans les tissus humains.

Tout ces études mettent en avant que l'inhibition de la cyclooxygénase de type 2 réduit de manière significative la production de prostaglandine dans les tissus et par voie de conséquence exerce une action anti-inflammatoire. Des travaux ont montré la présence élevée de COX-2 dans les muscles lisses des vaisseaux sanguins. L'activité COX-2 dans ces vaisseaux peut rectifier la diminution de production de prostaglandine PGI_2 observée lorsque l'endothélium vasculaire est endommagé. Localement, cette perte de PGI_2 augmente l'adhésion plaquettaire.[67] D'autres études ont corrélé le haut niveau de présence de l'enzyme COX-2 avec un risque accru de cancer du colon.[68,69] Nous verrons d'ailleurs dans le chapitre 3 que la cible PPARγ peut avoir une action inhibitrice de la production de l'enzyme COX-2 et ainsi ouvrir une voie thérapeutique de prévention du cancer du colon. D'autres résultats ont corrélé les fortes concentrations de prostaglandine PGE_2 produite par COX-2 avec la stimulation de voies de synthèse d'œstrogène dans le cas du cancer du sein.[70] A l'inverse, une inhibition de la cyclooxygénase de type 1 exerce une action néfaste sur la régulation rénale et la fonction gastro-intestinale par diminution du taux normal de prostaglandine.

2. Anti-inflammatoires non stéroïdiens (AINS) et agents coxibs

Les traditionnels anti-inflammatoires non stéroïdiens (naproxen, ibuprofen, diclofenac, indométhacine, nabumétone et aspirine) inhibent les deux formes de cyclooxygénase (Figure 5).[71] Leur action est principalement de bloquer la production de prostaglandine. Ils possèdent également des propriétés anti-pyrétiques et ont des effets analgésiques. Du fait qu'ils agissent également sur l'isoforme COX-1, ils sont responsables d'effets secondaires gastro-intestinaux et rénaux. Ils ont également la capacité d'inhiber l'action coagulante des plaquettes de manière réversible voire parfois irréversible dans le cas de l'aspirine.

---

[67] Mitchell, J.-A. ; Evans, T.-W Cyclooxygenase-2 as a therapeutic target *Inflamm Res* **1998**, *47*, 88-92.
[68] Pommery, J.; Pommery, N.; Hénichart, J.P. Modification of eicosanoid profile in human blood treated by dual COX/LOX inhibitors *Prostaglandins, Leukotrienes and Essential Fatty Acids* **2005**, *73*, 411-417.
[69] Tsujii, M.; Kawano, S.; Dubois, R.N. Cyclooxygenase-2 expression in human colon cancer cells increases metastatic potential *Proc. Natl. Acad. Sci.* **1997**, *94*, 3336-3340.
[70] Brodie, A.M.H.; Lu, Q.; Longa, B.J.; Fulton, A.; Chena, T.; Macpherson, N.; DeJong, P.C.; Blankenstein, M.A.; Nortier, J.W.R.; Slee, P.H.T.J.; van de Ven, J.; van Gorpc, J.M.H.H.; Elbers, J.R.J.; Schipper, M.E.I.; Blijham, G.H.; Thijssen, J.H. Aromatase and COX-2 expression in human breast cancers *Journal of Steroid Biochemistry & Molecular Biology* **2001**, *79*, 41-47.
[71] Warner, T.D.; Mitchell, J.A. Cyclooxygenases: new forms, new inhibitors, and lessons from the clinic *fasebj* **2004**, *18*, 790-804.

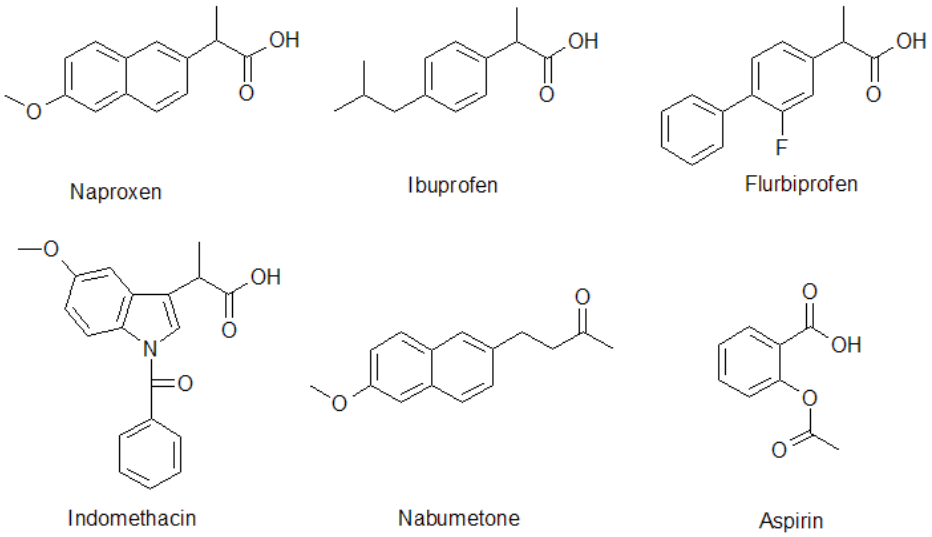


Figure 5. Structures chimiques de quelques anti-inflammatoires non-stéroïdiens (AINS)

De plus récentes molécules, appelées «coxibs» ont prouvé leur action sélective pour l'isoforme COX-2. Les représentants principaux de cette famille de composés sont le célécoxib, rofécoxib, valdécoxib et l'étoricoxib:

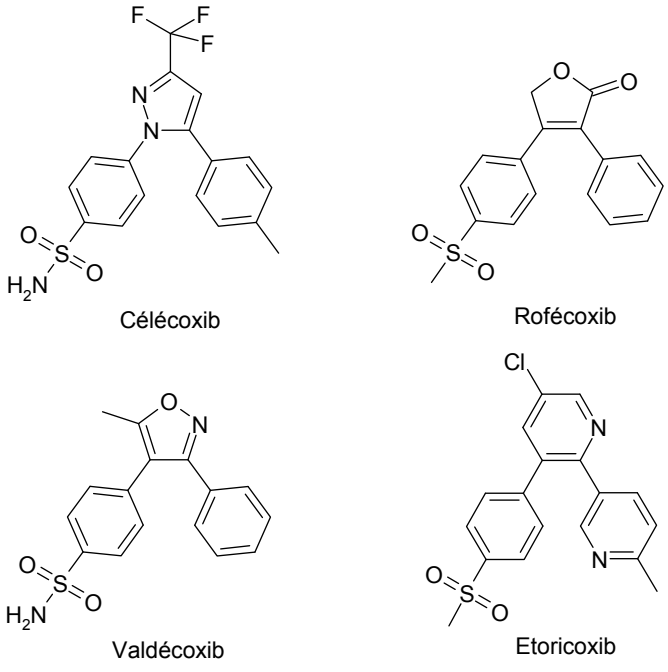
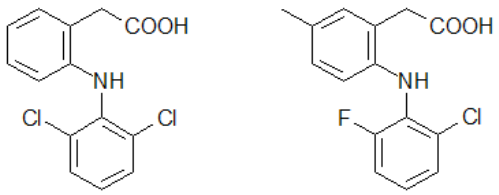


Figure 6. Structures chimiques d'anti-inflammatoires non-stéroïdiens de la famille des coxibs

Chacune de ces molécules possèdent le même squelette moléculaire: un hétérocycle disubstitué par un groupement aryle. Ces composés présentent tous ou presque une fonction

sulfonyle (sulfonamide ou méthyl-sulfone). La dénomination «coxib» n'est pas une terminologie attribuée à une classe structurale de molécule mais plutôt à une catégorie pharmacologique. C'est en effet le cas du «prexige» (lumiracoxib) dont la structure chimique est plus proche de celle d'un des AINS, le diclofenac. La figure suivante montre l'homologie entre deux molécules de classe pharmacologique différente (Figure 7).



Diclofenac Lumiracoxib

Figure 7. Homologie entre un AINS (diclofénac) et un coxib (lumiracoxib)

De nombreux essais cliniques ont permis de comparer les agents sélectifs COX-2 avec ceux de la classe AINS. La première étude rendue publique en mars 2000, appelée VIGOR (Vioxx Gastrointestinal Outcomes Research) avait au départ comme objectif d'évaluer les différences d'effets gastrointestinaux entre le naprosyn (naproxène, 500mg/jour) et vioxx (rofécoxib, 50mg/jour) chez des patients atteints d'arthrite rhumatoïde. Les résultats de l'étude VIGOR ont montré que les sujets du groupe traités au Vioxx présentent moins d'effets secondaires gastrointestinaux que ceux du groupe ayant été traités par le naproxène 500 mg.[72] Toutefois, des infarctus du myocarde ont été observés chez certains patients du groupe traité au Vioxx, entraînant son retrait immédiat du marché le 30 septembre 2004.[73] Une autre étude, menée par Merck, a également été faite sur ce médicament: APPROVe (Adenomatous Polyp Prevention on Vioxx) étudiait la prévention des polypes adénomateux avec Vioxx.[74] L'étude, qui est maintenant arrêtée, a été conçue pour évaluer l'efficacité de Vioxx 25 mg dans la prévention de la récurrence de polypes chez les patients ayant des antécédents d'adénomes colorectaux. L'étude a démontré qu'il existait un risque relatif accru d'incidents cardiovasculaires confirmés (crise cardiaque et accident cérébral) 18 mois après le début du

[72] Bombardier, C.; Laine, L.; Reicin, A.; Shapiro, D.; Burgos-Vargas, R.; Davis, B.; Day, R.; Bosi Ferraz, M.; Hawkey, C.J.; Hochberg, M.C.; Kvien, T.K.; Schnitzer, T.J. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. *N. Engl. J. Med.* **2000**, *343*, 1520-1528.
[73] Dogné, J.M.; Supuran, C.T.; Pratico, D. Adverse Cardiovascular Effects of the Coxibs *J. Med. Chem.* **2005**, *48*, 2251-2257.
[74] Bresalier, R.S.; Sandler, R.S.; Quan, H.; Bolognese, J.A.; Oxenius, B.; Horgan, K.; Lines, C.; Riddell, R.; Morton, D.; Lanas, A.; Konstam, M.A.; Baron, J.A. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *N Engl J Med* **2005**, *352*, 1092-102.

traitement par Vioxx, comparativement à un groupe de patients recevant un placebo. Les résultats pendant les premiers 18 mois de l'étude APPROVe n'ont démontré aucun risque accru d'incidents cardiovasculaires avec Vioxx. Une autre étude clinique, CLASS (Celecoxib Long term Arthrisis Safety Study) a comparé l'administration de «celebrex» (célécoxib) avec deux anti-inflammatoires non stéroïdiens, l'ibuprofène et le diclofénac chez des patients également atteints d'arthrite rhumatoïde.[75] Cette étude n'a pas remis en cause le composé «coxib» par quelque événement thrombotique que ce soit. La dernière étude que nous évoquerons est appelée TARGET (Therapeutic Arthritis Research and Gastrointestinal Event Trial).[76] Elle compare le «prexige» (lumiracoxib) avec le naproxène et l'ibuprofène. Cette étude n'a également pas permis d'impliquer de manière significative le prexige avec des risques accrus d'accidents thrombotiques comparables à ceux décrits avec le Vioxx.

3. Site actif et interaction

L'enzyme cyclooxygénase possède un site actif de nature majoritairement hydrophobe. Les deux isoformes ont une homologie de l'ordre de 61%. En règle générale, leur architecture est conservée. L'entrée du site de fixation dans les deux isoformes est délimitée par des résidus hydrophobes (Val116, Val349, Ser353, Tyr355, Leu359, Ala527 et Arg120). L'Arg120 est un résidu clé qui jouerait un rôle important dans la sélectivité COX-1/COX-2.[77] Des études ultérieures ont montré que des molécules d'eau pouvaient avoir un rôle à l'entrée du site actif lors de la reconnaissance avec le kétoprofène.[78] Dans nos études, nous n'avons pas mis en évidence l'intérêt d'introduire des molécules d'eau de manière implicite. La Tyr385 est un acide aminé commun aux deux isoformes responsables de la fixation irréversible de l'aspirine par une réaction d'acétylation. Des acides aminés spécifiques à l'isoforme 2 délimitent une poche adjacente, certainement responsable de la sélectivité entre les deux isoenzymes. La substitution de l'Ile523 (COX-1) par une Val523

[75] Silverstein,F.E.; Faich, G.; Goldstein, J.L.,; Simon L.S.; Pincus T.; Whelton, A. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis. The CLASS study: a randomized controlled trial. *JAMA* **2000**, *284*, 1247-1255.

[76] Schnitzer, T.J.; Burmester, G.D.; Mysler, E.; Hochberg, M.C.; Doherty, M.; Ehrsam, E.; Gitton, X.; Krammer, G.; Mellein, B.; Matchaba, P.; Gimona, A.; Hawkey, C.J. Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), reduction in ulcer complications: a randomized controlled trial. *Lancet*, **2004**, *364*, 665-674.

[77] Liu, H. ; Huang, X. ; Shen, J. ; Luo, X. ; Li, M. ; Xiong, B. ; Chen, G. ; Shen, J. ; Yang, Y. ; Jiang, H. ; Chen, K Inhibitory mode of 1,5-diarylpyrazole derivatives against cyclooxygenase-2 and cyclooxygenase-1 : molecular docking and 3D QSAR analyses *J. Med. Chem.* **2002**, *45*, 4816-4827.

[78] Palomer, A.; Pérez, J.J.; Navea, S.; Llorens, O.; Pascual, J.; Garcia, L.; Mauleon, D. Modeling cyclooxygenase inhibition. Implication of active site hydration on the selectivity of ketoprofen analogues *J.Med.Chem.* **2000**, 43, 2280-2284.

(COX-2) ouvre un accès à une poche complémentaire contenant des résidus polaires. Ceux-ci sont capables de former un réseau de liaisons hydrogène avec un inhibiteur (His90, Arg513).[79] Dans une étude de Soliva *et al* décrivant les interactions à l'intérieur du site actif, trois groupements structuraux des coxibs sont responsables de l'effet inhibiteur de ces composés (5MR: cycle central à 5 carbones, SR: benzène substitué par sulfone ou sulfonamide, BR: benzène non substitué).[80] Elles sont majoritairement de type électrostatique, Van der Waals et liaisons hydrogène. Nous avons regroupé ces interactions dans deux régions (Figure 8).

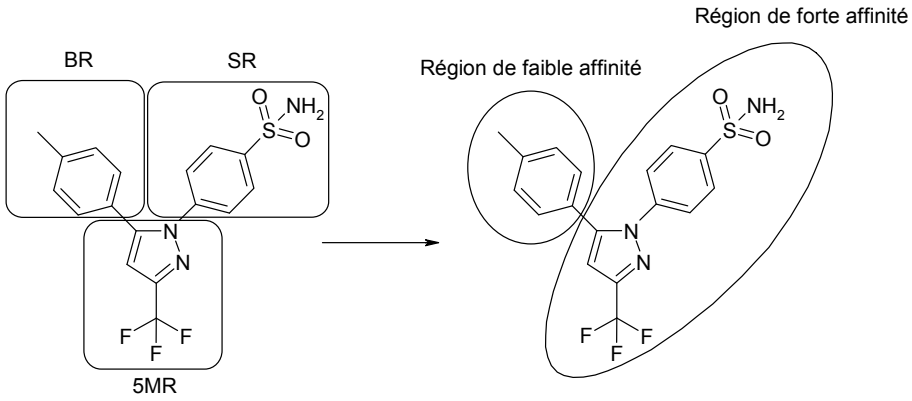


Figure 8. Carte d'interaction par région du SC-558 de la famille des coxibs

La partie centrale des coxibs ainsi que les groupements benzène substitués par une fonction sulfone ou sulfonamide sont impliqués dans de forts contacts de type Van der Waals mais également dans un réseau de liaisons hydrogène. C'est la «région de forte affinité». La partie restante de la molécule est moins affine pour le récepteur. Cette région est majoritairement de type hydrophobe et est nommée sur le schéma «région de faible affinité». Dans ces travaux, nous avons privilégié les inhibiteurs qui occupent de manière optimale la région 5MR tout en conservant une occupation de la poche adjacente SR. Par conséquent, tous les inhibiteurs que nous souhaitons extraire doivent occuper la «région de forte affinité».

4. Structures cristallographiques

Un total de 8 structures cristallographiques ont été extraites de la PDB (codes: 1CVU, 1DDX, 1PXX, 3PGH, 4COX, 5COX, 6COX et 1CX2) (Figure 9).

[79] Llorens, O.; Perez, J.-J; Palomer, A.; Mauleon,D. Different binding mode of diverse cyclooxygenase inhibitors *J Mol Graph Model.* **2002**, *20*, 359-371.
[80] Soliva, R. ; Almansa, C. ; Kalko, S.-G. ; Luque, F.-J. ; Orozco, M. Theoretical studies on the inhibition mechanism of cyclooxygenase-2. Is there a unique recognition site? *J. Med. Chem.* **2003**, *46*, 1372-1382.

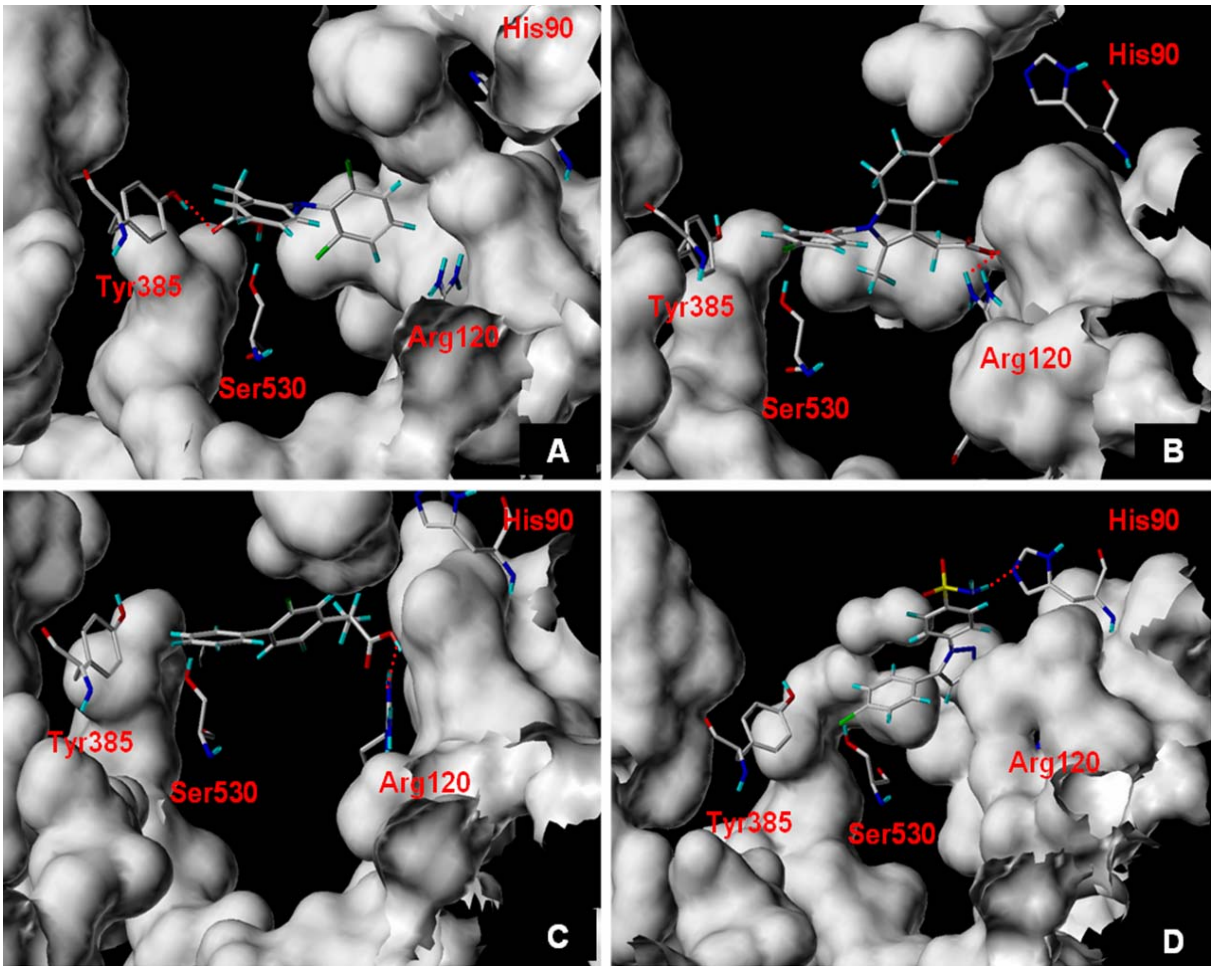


Figure 9. Mode d'interaction des différents inhibiteurs de la cyclooxygénase de type 2

Les structures 1CVU, 1DDX, 1PXX, 3PGH et 4COX sont respectivement cristallisées avec l'acide arachidonique, une prostaglandine, le diclofénac, le flurbiprofène et l'indométhacine. Les structures 1CX2 et 6COX sont cristallisées avec l'analogue du célébrex SC-558.

Le schéma A illustre le groupement carboxylique du diclofénac interagissant avec la Ser530 et la Tyr385. Une mutation de la Tyr385 conduit à une perte d'affinité du diclofénac. L'Arg120 n'a toutefois aucune interaction avec ce ligand. Enfin, le diclofénac occupe la région «BR» (Figure 8).

L'indométhacine existe sous deux conformations (cis et trans). La fonction carboxylique de l'indométhacine est à proximité de l'Arg120. Le groupement benzoyle entre en interaction avec la Ser530. Cette fonctionnalité est responsable de la transition de la forme cis à la forme trans. Le fragment p-chlorophenyl est à proximité de la Tyr385. Finalement, le carbohydrate forme une liaison saline avec l'Arg120. Ce ligand occupe la région «BR+5MR» (Schéma B).

Le flurbiprofène possède un groupement carboxylique qui fait face à l'Arg120 et la Tyr385. Le cycle phényl du flurbiprofène interagit avec la Ser530. Le flurbiprofène occupe la région «BR+5MR» (Schéma C). Cette molécule possède le même mode de fixation dans COX-1 et COX-2.[81]

Le SC-558 est le seul ligand occupant la poche adjacente contenant l'His90 et l'Arg513 certainement responsable de la sélectivité entre les deux isoformes (Schéma D). Les atomes d'oxygène de la fonction sulfonamide du SC-558 interagissent avec l'His90 et l'Arg513. Le groupement bromophényl est lié à la partie hydrophobe du site actif proche de la Tyr385 et la Ser530. Le SC-558 occupe la région «BR+5MR+SR».

La première étape de notre étude a consisté à comparer chaque structure tridimensionnelle de l'enzyme COX-2.

F. Matériel et méthodes

1. Constitution de chimiothèques dédiées à évaluer les modèles

a) Constitution de l'«*ensemble total*»

(1) Généralités

Tous les modèles ont été choisis et validés grâce à l'*ensemble total*. Cette librairie de molécules est utile pour se placer dans le contexte d'une chimiothèque de plusieurs millions de composés dans laquelle les molécules potentiellement actives sont en sous effectif. Nous souhaitons retrouver des molécules actives que nous avons mélangées avec des molécules *a priori* inactives, que nous appelons également *leurre*. Chaque modèle est jugé sur sa capacité à récupérer le plus de molécules actives possibles avec le moins de molécules inactives.

(2) Molécules actives

Des inhibiteurs COX-2 provenant de la littérature ont été extraits. Au total, 76 molécules de 7 familles différentes ont été regroupées (Figure 10).[82,83,84]

[81] Bayly, C.I.; Black, W.C.; Léger, S.; Ouimet, N.; Ouellet, M.; Percival, M.D. Structure-based design of COX-2 selectivity into flurbiprofen *Bioorg. Med. Chem. Lett.* **1999**, 307-312.
[82] Chen, H.; Li, Q.; Yao, X.; Fan, BT.; Yuan, S.; Panaye, A.; Doucet, J.P. CoMFA/CoMSIA/HQSAR and

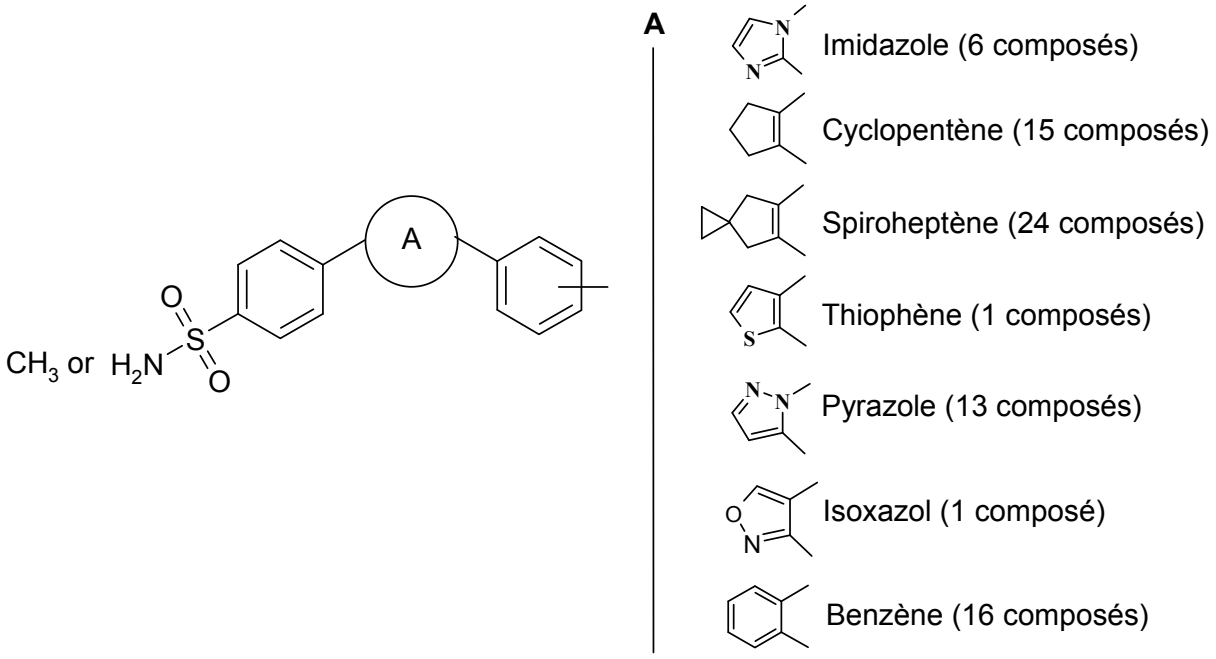


Figure 10. Structure générale des inhibiteurs de la cyclooxygénase de type 2 de la famille des coxibs

Seules les molécules dont la pIC_{50} est supérieure à 8,0 ont été sélectionnées.

(3) Molécules a priori inactives ou leurres

En parallèle, 230 leurres ont été sélectionnés dans notre chimiothèque locale. Les critères de choix sont tout d'abord que ces composés doivent posséder un groupement sulfonyle. Par ailleurs, leur masse molaire doit être comprise entre 240 et 530 g.mol-1. Le nombre maximum de liaisons sujettes à la rotation a été fixé à 15. Finalement, à diversité maximale, 230 molécules ont été sélectionnées. Il existe un grand nombre d'algorithmes de diversité décrits dans la littérature.[85] L'algorithme de diversité utilisé est basé sur les clés structurales «MACCS» mis au point par MDL.[86,87] Le coefficient de similarité de Tanimoto est ensuite appliqué pour mesurer la diversité au sein de l'ensemble.

---

docking study of the binding mode of selective cyclooxygenase (COX-2) inhibitors *QSAR Comb. Sci.* **2004**, *23*, 36-54.

[83] Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. Three-Dimensional Quantitative Structure-Activity Relationships of Cyclo-oxygenase-2 (COX-2) Inhibitors: A Comparative Molecular Field Analysis. *J. Med. Chem.* **2001**, *44*, 3223-3230.

[84] Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276-285.

[85] Agrafiotis, D. K.; Lobanov, V. S.; Rassokhin, D. N.; Izrailev, S. The Measurement of Molecular Diversity, in Virtual Screening for Bioactive Molecules **2000**, 265-300.

[86] Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for Use in Drug

(4) Ensemble total

Les 306 structures (76 molécules actives + 230 *leurres*), représentées sous forme bidimensionnelle, sont converties en 3D par Corina et ensuite minimisées par le champ de force MMFF94s implémenté dans Moe[88]. Nous avons appelé cet ensemble de molécules l'*ensemble total*. Cette librairie est très utile pour comparer des modèles entre eux.

b) Constitution de l'*ensemble d'entraînement* et de test

Nous avons utilisé l'*ensemble total* pour tester et comparer différents modèles entre eux. Une fois un modèle choisi, celui-ci doit être validé. L'obtention du modèle se fait principalement en deux étapes. Tout d'abord, le modèle est construit à partir de l'information contenue dans l'*ensemble d'entraînement* (encore appelé *ensemble d'apprentissage*). Finalement, il est testé et validé sur l'*ensemble de test*.

Plusieurs stratégies sont possibles pour envisager leur composition. Des algorithmes de diversité peuvent être employés afin d'introduire dans l'*ensemble d'entraînement* les molécules les plus diverses. Les composés restants les moins divers sont regroupés dans l'*ensemble de test*. Le principe de cette stratégie est de construire un modèle sur les molécules les plus diverses. *A priori*, un modèle provenant d'un ensemble divers doit être capable de prédire une plus grande variété de composés. L'inconvénient de cette méthodologie est que l'*ensemble de test* valide le modèle sur des molécules trop similaires entre elles. Pour cette raison, nous avons choisi de ne pas utiliser d'algorithme de diversité pour établir la constitution des ensembles et ainsi d'éviter le biais lors de la validation. Par conséquent, la composition a été faite au hasard. L'*ensemble d'entraînement* représente 70% de l'*ensemble total*. L'*ensemble de test* est constitué des 30% de molécules restantes.

2. Etape de docking

a) Généralités

FlexX est un algorithme de docking rapide capable d'arrimer des ligands de nature organique dans un environnement constitué d'acides aminés. La technique employée par

---

Discovery JCICS **2002**, *42*, 1273-1280.
[87] http://www.mdl.com/
[88] http://www.chemcomp.com/

FlexX est la construction incrémentale du ligand au sein du site actif. L'étape limitante de ce processus est le choix du fragment servant de base à la reconstruction. Son emplacement conditionne l'exploration de l'espace conformationnelle de la molécule. Le biais qu'impose donc ce choix peut parfois être fatal et pénaliser des molécules biologiquement actives. Le problème est moindre dans le cas d'une molécule inactive. En effet, même si un emplacement favorable est choisi pour le fragment de base, il reste peu probable qu'un ligand *a priori* inactif satisfasse toutes les étapes de reconstruction. Une fois l'emplacement choisi, les fragments restants sont ajoutés tour à tour. Les angles de torsion (provenant de la librairie d'angle MIMUMBA)[89] entre deux fragments sont explorés afin de déterminer la position optimale pour le dernier fragment ajouté. Cet algorithme permet de prendre en compte la liberté conformationnelle de la molécule et d'autre part d'induire un maximum d'interactions avec le site de fixation. FlexX combine de manière simultanée la détection d'un éventuel recouvrement entre la protéine et le ligand durant le processus de reconstruction mais également la recherche d'interactions toujours plus fortes en prenant en compte la flexibilité des conformations. Nous avons choisi cet algorithme en raison de sa rapidité à traiter une molécule (environ 1 molécule toutes les 30 secondes) avec une bonne précision. Les paramètres standard de FlexX ont été utilisés par défaut dans cette étude.

b) Le processus de docking avec FlexX

(1) Choix du fragment de base

La première étape, après découpe du ligand, est le choix d'un fragment rigide capable de répondre simultanément à un nombre d'interactions suffisant avec l'environnement de la protéine.[90] Le positionnement du fragment s'effectue en trois étapes. Tout d'abord, plusieurs points, susceptibles de former des interactions avec le fragment, sont générés. Ensuite, des jeux de trois points significatifs du site actif sont extraits, formant un triangle. Finalement, le fragment est superposé à chaque triangle. S'il y a concordance, l'étape de reconstruction peut avoir lieu. Sinon, un autre fragment de base est sélectionné (Figure 11).

[89] Klebe, G.; Mietzner, T.; Weber, F. Methodological developments and strategies for a fast flexible superposition of drug-size molecules. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 35-49.
[90] Rarey, M.; Kramer, B.; Lengauer, T. Multiple automatic base selection: protein-ligand docking based on incremental construction without manual intervention *J. Comput.-Aided Mol. Des.* **1997**, *11*, 369-384.

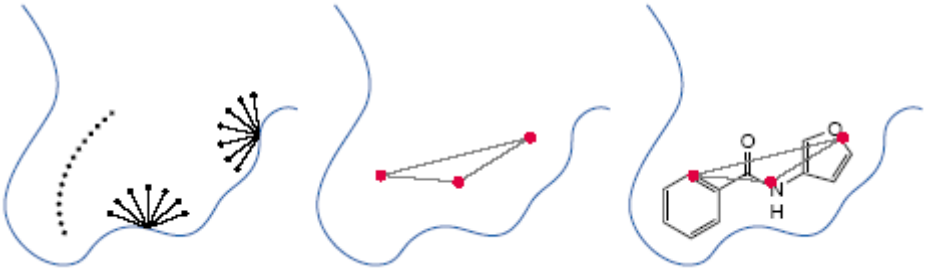


Figure 11. Superposition des différents points du site actif avec le ligand

Dans un premier temps, l'algorithme va vérifier les contraintes angulaires entre les deux triangles. Seules les interactions compatibles sont conservées. Celles dont les distances et/ou les angles entre deux triangles sont trop importants sont directement éliminées. Les déviations sont calculées par RMSD (Root Mean Square Deviation). Les solutions proches en termes de RMSD sont regroupées ensemble dans une même famille. Les placements retenus sont soumis à la fonction de Böhm calculant les énergies entre le triangle *site actif* et le triangle *fragment*. L'avantage de FlexX est de pouvoir sélectionner un fragment de base sur des critères de liaisons directionnelles mais également selon des contacts lipophiles.[91] Cette option est très utile dans le cadre de criblage virtuel sur chimiothèques provenant de techniques *de novo*.

Une restriction est à observer dans le cas de fragments trop petits pour lesquels la génération d'un triangle est impossible. Dans ce cas, FlexX utilise des lignes à la place de triangles.

(2) Reconstruction sur le fragment de base

Les groupes de placement du fragment de base servent de point de départ à la reconstruction du ligand. Ainsi, la construction incrémentale peut s'apparenter à un arbre de décision dans lequel le premier niveau correspond au nombre de solutions trouvées lors de la recherche d'un premier fragment (Figure 12).

[91] Rarey, M.;Kramer, B.;Lengauer, T. Docking of hydrophobic ligands with interaction-based matching algorithms *Bioinformatics* **1999**, *15*, 243-250.

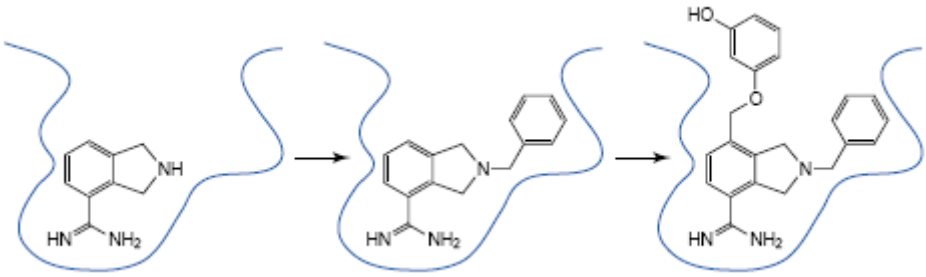


Figure 12. Construction incrémentale d'un ligand au sein du site actif

Lors de la recherche du fragment de base, si FlexX isole trois placements, alors la première ligne de l'arbre comptera trois noeuds. Un noeud représente une conformation (ou morceau de conformation). Ensuite, la reconstruction des fragments restants se fait sur chaque noeud. A chaque itération, l'arbre croît exponentiellement en nombre de solutions.

(3) Choix des meilleurs placements

Les fragments formant des liaisons hydrogène ou des liaisons ioniques sont conservés puisque ces interactions sont fortement directionnelles. Les *k* meilleurs placements sont déterminés comme étant les plus énergétiquement favorables et sont conservés pour la suite de la reconstruction. Le nombre *k* permet de prévenir les minimums locaux. Les placements proches en terme de distance (calcul par RMSD) et donc similaires sont enlevés de la liste des *k* meilleurs placements. Chaque fragment ajouté doit respecter les limites de recouvrement avec le récepteur (au delà de 4,5 $Å^3$, l'algorithme élimine le placement). A chaque ajout d'un fragment, l'algorithme vérifie si de nouvelles interactions sont trouvées ou même si le recouvrement total entre le récepteur et le ligand ne devient pas trop important.

(4) Test de recouvrement protéine/ligand

Ce test est réalisé grâce à la superposition de points triplets: (l_i, r_i, w_i) avec l_i un point du ligand, r_i un point du récepteur et w_i un poids. L'objectif est de minimiser la somme des carrés des écarts entre l_i et r_i, le tout pondéré par le poids (qui est la contribution énergétique de l'interaction entre l et r). Toutes les interactions présentant des violations de distances sont éliminées. Une superposition est à nouveau effectuée.

(5) Exploration des paramètres fondamentaux

Des paramètres faisant intervenir les valeurs d'incrément d'angle MIMUMBA ont pu être modifiés. Plus l'incrément est faible, plus la molécule a de chance de rencontrer des interactions favorables, mais plus le temps de calcul augmente. Nous n'avons pas mis en évidence l'intérêt de diminuer la valeur d'incrément car le gain de prédiction obtenu ne justifie pas une telle consommation en temps de calcul (parfois plus de 180 secondes par molécule soit trois fois plus de temps de calcul qu'en utilisant le paramètre standard). Un autre paramètre ayant également une place importante dans ce type d'algorithme incrémental est le nombre de fragmentations maximums admises. Plus ce nombre augmente, plus l'aspect flexible de la molécule est pris en compte et donc plus l'espace conformationnel est couvert. En contrepartie, le fait d'augmenter le nombre de fragmentations conduit inévitablement à des conformations trop artificielles, épousant de manière irréaliste le site actif de la protéine. Par conséquent, une molécule *a priori* inactive augmente ses chances d'obtenir une conformation la faisant passer pour active. Nous avons limité à 4 le nombre maximum de fragmentations possibles avec, pour chaque fragment, une limite de 30 conformations possibles. Une option intéressante est de pouvoir définir un fragment commun pour chaque molécule. Nous n'avons pas utilisé cette option qui est spécialement dédiée à une librairie provenant d'une stratégie *de novo design,* dans laquelle un fragment commun a été identifié. Cette option offre un intérêt lorsque l'on connaît l'emplacement du fragment généralement porteur d'une fonctionnalité au sein du site actif. Toutefois, du fait que nous souhaitons isoler des composés ne possédant pas la fonctionnalité sulfonyle, nous n'avons pas appliqué cette méthodologie à notre étude.

c) Orientation des groupements donneurs /accepteurs

Des acides aminés tels que la tyrosine, la sérine, la thréonine possèdent la fonctionnalité hydroxyle, facilement orientable et sujette à la rotation. Afin de favoriser la création de liaisons hydrogène, il est parfois nécessaire de moduler l'angle dièdre formé avec le reste du résidu et ainsi de choisir l'angle le plus favorable (Figure 13).

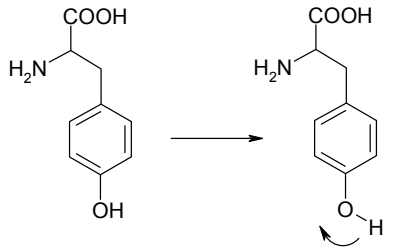


Figure 13. Variations de l'angle dièdre du résidu tyrosine

De même, les liaisons des carbonyles des fonctions «acide carboxylique» des chaînes latérales doivent être modifiés en liaisons délocalisées. C'est particulièrement le cas de l'acide aspartique et de l'acide glutamique (Figure 14).

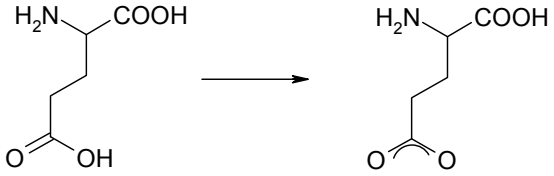


Figure 14. Délocalisation de la fonction acide du résidu acide glutamique

Dans la représentation de droite, l'acide glutamique est capable, tour à tour, de présenter ces oxygènes, aussi bien comme donneur que accepteur.

Enfin, l'histidine est un résidu dont la chaîne latérale peut facilement subir une rotation, exposant ainsi les deux azotes intra-cycliques à des zones favorables du ligand pour former des liens hydrogène (Figure 15).

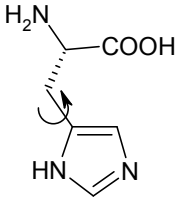


Figure 15. Rotation de la chaîne latérale du résidu histidine

Toutes ces optimisations conduisent à favoriser des réseaux de liaisons hydrogène au sein du complexe ligand-protéine.

3. Modèle pharmacophorique

Un filtre pharmacophorique a été envisagé dans nos études permettant de sélectionner les composés répondant à certains critères spatiaux et interactionnels particuliers. FlexX-Pharm est un module de FlexX permettant le couplage d'un modèle pharmacophorique avec l'algorithme de docking FlexX.[92] Deux types de contraintes peuvent être paramétrés.

Tout d'abord, le module *Pharm* propose de prendre en compte l'aspect interactionnel présent dans le site actif. Plus précisément, cette option appelée «contrainte interactionnelle» vérifie si les interactions connues ou supposées favorables sont présentes lors de l'association du ligand et du site actif. Une molécule, pour être admise au protocole de docking, doit satisfaire ces contraintes. Par ailleurs, l'option *Pharm* offre l'opportunité de définir des sphères dans lesquelles un certain type d'atome du ligand doit être présent dans ce volume prédéfini. Ce type de contrainte est dénommé «contrainte spatiale».

Ces deux types de contraintes ont pour but de guider la recherche conformationnelle et ainsi de réduire l'espace conformationnel accessible pour une molécule donnée. Afin d'ajouter de la souplesse au modèle, les deux types de contraintes décrites précédemment peuvent être nuancées. Dans le cas le plus strict, la contrainte peut être paramétrée comme «essentielle», c'est-à-dire que pour qu'une molécule passe le filtre pharmacophorique, elle doit absolument satisfaire cette contrainte. Un degré de souplesse peut être ajouté en définissant une contrainte comme «optionnelle». Dans ce cas, une molécule ne satisfaisant pas cette contrainte ne sera pas forcément rejetée. Prenons l'exemple d'un modèle pharmacophorique comprenant deux contraintes interactionnelles (dont une définie en «essentielle» et l'autre en «optionnelle») et trois contraintes spatiales (dont une définie comme «essentielle» et deux comme «optionnelle»). Une molécule réussira l'étape pharmacophorique que si elle satisfait les deux contraintes «essentielles» de type interactionnel et spatial. Par ailleurs, la molécule devra également répondre au minimum à une des trois contraintes définies comme «optionnelle» (une de type interactionnelle et deux de type spatiale). En résumé, plus le modèle possède de contraintes optionnelles, moins celui-ci est contraignant.

Nous avons élaboré le pharmacophore en utilisant le ligand co-cristallisé comme modèle d'alignement. Seules les 10 structures les plus diverses (calculées par le coefficient de Tanimoto sur la base des clés structurales MACCS) de notre ensemble de 76 molécules

[92] Hindle, S.-A.; Rarey, M.; Buning, C.; Lengauer, T. Flexible docking under pharmacophore type constraints *J. Comput. Aided Mol. Des.* **2002**, *16*, 129-149.

actives ont été alignées sur le SC-558, conservé rigide. Nous en avons déduit un modèle constitué uniquement de contraintes spatiales. En effet, nous n'avons pas désiré ajouter la composante interactionnelle proposée par *Pharm*, du fait que nous souhaitons récupérer des composés dont les groupements chimiques interagissent d'une manière différente de celle décrite avec les composés «coxibs». Une optimisation de ces sphères spatiales (diamètre et distance inter-sphères) a été nécessaire. Sa conception est basée sur le rejet ou non des molécules de l'*ensemble total*. Un modèle pharmacophorique qui élimine le plus de *leurres* tout en conservant un maximum de composés actifs sera sélectionné.

4. Les fonctions de scoring

Dans un processus de criblage virtuel, les fonctions de scoring ont une double utilité: choisir la première pose (parmi les 30 proposées par défaut dans le cas de FlexX) et établir le classement de la première pose de toutes les molécules. C'est l'étape de re-scoring

a) Choix de la première pose

Le choix de la première pose se fait en prenant la valeur prédite d'énergie la plus faible provenant de la fonction de scoring préalablement choisie. Nous avons récupéré les premières poses selon le score X, selon Y et selon Z (Tableau 1).

Molécule-pose	Score X	Score Y	Score Z
a-1	-25	-50	-3
a-2	-15	-210	-12
a-3	-32	-80	-20
b-1	-12	-300	-2
b-2	-42	-250	-10
b-3	-6	-410	-5
c-1	-21	-60	-12
c-2	-3	-325	-8
c-3	-40	-210	-5
d-1	-50	-230	-8
d-2	-6	-420	-3
d-3	-28	-256	-12

Choix de la 1ère pose

1ère pose selon X	1ère pose selon Y	1ère pose selon Z
a-3	a-2	a-3
b-2	b-3	b-2
c-3	c-2	c-1
d-1	d-2	d-3

Tableau 1. Choix de la première pose selon les fonctions X, Y, Z prises individuellement

Dans le tableau ci-dessus, chaque molécule (a, b, c et d) existe sous trois conformations différentes (1, 2 et 3). Choisir la première pose consiste à déterminer pour a, b,

c et d la meilleure conformation à conserver pour les études ultérieures (c'est-à-dire 1, 2 ou 3). On observe des variations de choix de première pose en fonction du score choisi (X, Y, ou Z). Par exemple, le choix de la première pose pour la molécule «d» apparaît controversé puisque chaque fonction trouve une solution différente (X choisit d-1, Y choisit d-2 et Z choisit d-3). En admettant que le Score X soit choisi comme fonction de récupération de la première pose, on aurait:

Poses	Score X	Score Y	Score Z
a-1	-25	-50	-3
a-2	-15	-210	-12
a-3	-32	-80	-20
b-1	-12	-300	-2
b-2	-42	-250	-10
b-3	-6	-410	-5
c-1	-21	-60	-12
c-2	-3	-325	-8
c-3	-40	-210	-5
d-1	-50	-230	-8
d-2	-6	-420	-3
d-3	-28	-256	-12

1ère pose selon Score X

1ère pose	Score X	Score Y	Score Z
a-3	-32	-80	-20
b-2	-42	-250	-10
c-3	-40	-210	-5
d-1	-50	-230	-8

Tableau 2. Récupération des scores pour la première pose de chaque molécule selon le Score X

Après récupération selon «Score X» des premières poses, chaque molécule se trouve décrite par les trois scores X, Y et Z. Par conséquent, un classement des premières poses de chaque molécule peut être effectué selon les trois fonctions X, Y ou Z.

b) Re-scoring

Le re-scoring de la première pose est une opération qui génère une matrice multidimensionnelle. Le traitement mathématique de ces données a été fait en utilisant des méthodes de consensus et des stratégies d'analyse de données multivariées.

(1) Consensus de fonctions de scoring

Charifson *et al*, en 1999, prouvent qu'une méthode intéressante consiste à combiner les fonctions de scoring ensemble.[93] Chaque fonction de scoring a ses avantages et désavantages. Il n'y a pas *de facto* de fonction standard fonctionnant dans tous les profils

[93] Charifson, P.S.; Corkery, J.J.; Murcko, M.A.; Walters, W.P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins *J. Med. Chem.* **1999**, *42*, 5100-5109.

pharmacologiques. Des méthodes de combinaisons de fonctions de scoring, appelées consensus, ont été décrites dans la littérature. Ces stratégies ont pour but d'augmenter le taux de bonnes prédictions tout en réduisant le nombre de faux positifs. Dans notre étude, nous avons mis en évidence l'intérêt de choisir les fonctions de scoring lors de l'utilisation d'un consensus. L'introduction de fonctions de scoring peu performantes voire inefficaces, entraîne irrémédiablement une augmentation du niveau d'erreur et donc une augmentation du taux de faux positifs dans l'ensemble final prédit.

Lors de la manipulation de plusieurs fonctions de scoring dans un même modèle, trois hypothèses doivent être vérifiées. Tout d'abord, le positionnement des molécules doit être observé pour s'assurer qu'un bon score ne soit pas attribué à une molécule en dehors du site actif. La deuxième hypothèse est que l'erreur de chaque fonction de scoring doit suivre une distribution normale centrée sur 0. Enfin, de toutes les fonctions de scoring utilisées, aucune ne doit dériver d'une autre par une simple transformation ou simplification d'une fonction déjà utilisée dans le modèle.

Trois stratégies de consensus ont été étudiées. Nous avons conservé la même terminologie que celle utilisée dans la littérature.[94,95]

(a) Harmonisation des scores

Afin de comparer les fonctions de scoring «f» entre elles nous les avons normalisées et centrées selon la formule suivante:

$$Score_{harmonisé} = \frac{f - f_{\min}}{f_{\max} - f_{\min}}$$

Équation 2. Harmonisation des scores des fonctions de scoring

1ère pose	Score X	Score Y	Score Z		1ère pose	Score X	Score Y	Score Z
a-3	-32	-80	-20		a-3	1,000	1,000	0,000
b-2	-42	-250	-10	Harmonisation →	b-2	0,444	0,000	0,667
c-3	-40	-210	-5		c-3	0,556	0,235	1,000
d-1	-50	-230	-8		d-1	0,000	0,118	0,800

Tableau 3. Harmonisation des scores X, Y et Z

[94] Yang, J.M.; Chen, Y.F.; Shen, T.W.; Kristal, B.S.; Hsu, D.F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening *J. Chem. Inf. Model.* **2005**, *45*, 1134-1146.
[95] Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions *J. Mol. Graph. Model.* **2002**, *20*, 281-295.

(b) La stratégie «Rank by rank»

Cette stratégie est obtenue en faisant la moyenne des rangs correspondant à chaque fonction de scoring. Le classement final est effectué par ordre croissant de la moyenne des rangs ($rang_{moyen}$) (Tableau 4).

Molécule	Score X	Score Y	Score Z		Score A	Score B	Score C		$Rang_{moyen}$
a	0.525	0.564	0.671		4	4	4		4.0
b	0.434	0.000	0.911	Classement	3	1	5	Classement	3.0
c	0.000	0.232	0.000	→	1	2	1	→	1.3
d	1.000	1.000	1.000	selon A, B, C	6	6	6		6.0
e	0.566	0.777	0.620		5	5	3		4.3
f	0.303	0.412	0.405		2	3	2		2.3

Tableau 4. Illustration de la méthode «Rank by rank», classement des molécules

(c) La stratégie «Rank by number»

La méthode «Rank by number» [96] consiste, après avoir redimensionné les scores, à calculer la moyenne des scores ($score_{moyen}$) pour chaque molécule (Tableau 5).

Les composés sont finalement classés selon le nouveau $score_{moyen}$. Le classement est effectué par ordre croissant de $score_{moyen}$ (un score proche de 0 correspond à une molécule prédite active, alors qu'un score voisin de 1 est associé à une molécule prédite inactive).

Molécule	Score X	Score Y	Score Z		$Score_{moyen}$		Rang
a	0.525	0.564	0.671		0.587		4
b	0.434	0.000	0.911		0.448		3
c	0.000	0.232	0.000	A + B + C →	0.077	Classement →	1
d	1.000	1.000	1.000		1.000		6
e	0.566	0.777	0.620		0.654		5
f	0.303	0.412	0.405		0.373		2

Tableau 5. Illustration de la méthode «Rank by number», classement des molécules

(d) La stratégie «Rank by best»

Cette méthode de consensus permet d'extraire, pour chacune des molécules, le meilleur des trois scores (Tableau 6). Le classement s'effectue classiquement par ordre croissant des meilleurs scores ($score_{meilleur}$).

[96] Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes *J. Chem. Inf. Model.* **2006**, *46*, 380-391.

Molécule	Score X	Score Y	Score Z		$Score_{meilleur}$		Rang
a	0.525	0.564	0.671		0.525		4
b	0.434	0.000	0.911	Choix du	0.000	Classement	2
c	0.000	0.232	0.000	→	0.000	→	1
d	1.000	1.000	1.000	meilleur score	1.000		6
e	0.566	0.777	0.620		0.566		5
f	0.303	0.412	0.405		0.303		3

Tableau 6. Illustration de la méthode «Rank by best», classement des molécules.

(e) La stratégie «Rank by vote»

La stratégie «Rank by vote» est plus complexe. Cette méthode consiste à faire voter chaque fonction de scoring. Les molécules sont les candidats à élire. Pour chaque fonction de scoring, une limite doit être fixée pour déterminer si une molécule est prédite active ou non.

Dans le cas où l'on a trois fonctions de scoring X, Y, Z, une molécule ne pourra récolter au maximum que 3 votes. Pour définir si une molécule est prédite active ou non, une autre limite à fixer est le nombre de votes que récoltera une molécule. Dans le cas où une molécule ne récolte aucun vote, elle sera certainement à considérer comme inactive. Dans les cas intermédiaires où la molécule récolte un total de 1 ou 2 votes, celle-ci pourra aussi bien être considérée active ou inactive. Enfin, dans le cas où une molécule récolte trois votes, elle sera certainement traitée comme active (Tableau 7).

Molécule	Score X	Score Y	Score Z		Nombre de vote		Election		Rang
a	0.525	0.564	0.671		0		non		4,5,6
b	**0.434**	**0.000**	0.911	Seuil de vote	2	Seuil de sélection	oui		3
c	**0.000**	**0.232**	**0.000**	→	3	→	oui	→	1.2
d	1.000	1.000	1.000	0,5	0	2	non		4,5,6
e	0.566	0.777	0.620		0		non		4,5,6
f	**0.303**	**0.412**	**0.405**		3		oui		1.2

Tableau 7. Illustration de la méthode «Rank by vote», classement des molécules.

Cette méthode impose deux choix:

- la limite de score pour attribuer un vote
- le nombre de votes nécessaire pour élire une molécule.

Le classement final est moins pertinent que les autres méthodes du fait des différentes approximations introduites dans le modèle. Par exemple, la molécule « a » peut être classée en position *ex aequo* 4, 5 et 6 tandis que la molécule «c» peut être associée au rang 1 ou 2.

La combinaison des données issues des fonctions de scoring peut également être obtenue en utilisant des méthodes statistiques.

(2) Méthodes statistiques

Les analyses factorielles sont des techniques capables de réduire la dimensionnalité au sein d'un environnement constitué de plusieurs variables. L'analyse factorielle discriminante, encore appelée AFD,[97] est une méthode statistique multivariée qui corrèle une variable qualitative (dans notre cas, une des deux classes active ou inactive) avec plusieurs variables quantitatives. Chaque observation est caractérisée par plusieurs variables (les différents scores dans notre cas).[98] Dans l'*ensemble d'entraînement*, la classe qualitative (active ou inactive) est connue. Une fois le modèle obtenu, il est possible de déduire la classe par rapport aux valeurs quantitatives des points. Le traitement mathématique est similaire à celui employé dans l'analyse en composantes principales (ACP).[99] Toutefois, les deux méthodes ne maximisent pas la même grandeur. L'AFD maximise la différence entre les valeurs moyennes des groupes. Le premier axe, constitué d'une combinaison linéaire des variables initiales, tente de maximiser la séparation entre les deux nuages crées par les deux différentes classes. Le second axe est perpendiculaire au premier et a le même objectif que le premier. Ce processus est répété jusqu'au dernier axe. Dans notre cas, il y a uniquement deux classes. Par conséquent, nous avons une fonction discriminante qui est le premier axe F_1. La figure 16 montre un exemple de discrimination obtenue selon l'axe F_1:

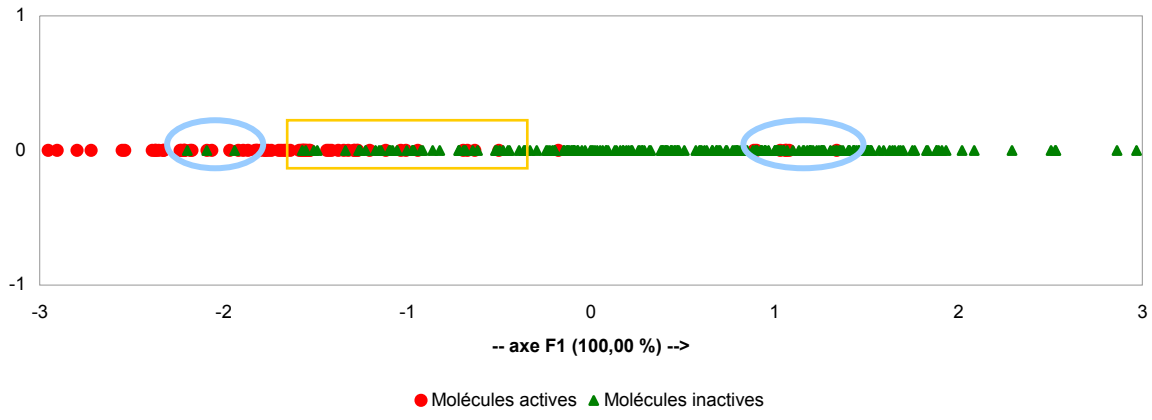


Figure 16. Représentation des individus selon l'axe factoriel F_1 exprimant la plus grande variance

[97] Terp, G.E.; Johansen, B.N.; Christensen, I.T.; Jørgensen, F.S. A New Concept for Multidimensional Selection of Ligand Conformations (MultiSelect) and Multidimensional Scoring (MultiScore) of Protein-Ligand Binding Affinities *J. Med. Chem.* **2001**, *44*, 2333-2343.
[98] Jacobsson, M.; Lidén, P.; Stjernschantz, E.; Boström, H.; Norinder, U. Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data *J. Med. Chem.* **2003**, *46*, 5781-5789.
[99] Morineau, A.; Aluja-Banet, T. Analyse en composantes principales *Centre international de statistique et d'informatique appliquées.*

Les coordonnées des molécules sont représentées sur l'axe F_1. Dans cet exemple, une bonne discrimination est obtenue. Les molécules actives (symbolisées par des points rouges) sont sur la partie gauche de l'axe tandis que les molécules inactives (représentées en vert) occupent la partie droite de F_1. Toutefois, des molécules actives sont observées en dehors de leur catégorie d'origine. Il en est de même pour quelques *leurres* qui se situent au milieu de la catégorie d'actifs (entourés par une ellipse bleue). Il est également possible d'observer une région dans laquelle les deux catégories se superposent. C'est la zone d'incertitude au sein de laquelle il est difficile de prédire la catégorie des molécules (entourée par un rectangle jaune). Une limite raisonnable pour discriminer correctement les deux catégories serait de se placer au point d'abscisse -1 par exemple. Afin de simplifier la représentation graphique, nous avons transformé l'unité de l'abscisse. Dans la figure suivante, la probabilité d'appartenance à une catégorie est la nouvelle unité d'abscisse comprise en 0 et 1. Plus la molécule se rapproche de 0, plus elle a de chance d'être inactive. Inversement, plus la probabilité attribuée à une molécule tend vers 1, plus celle-ci aura de chance d'être active. Une limite raisonnable dans ce genre de cas est de fixer une limite à 0,5 (Figure 17).

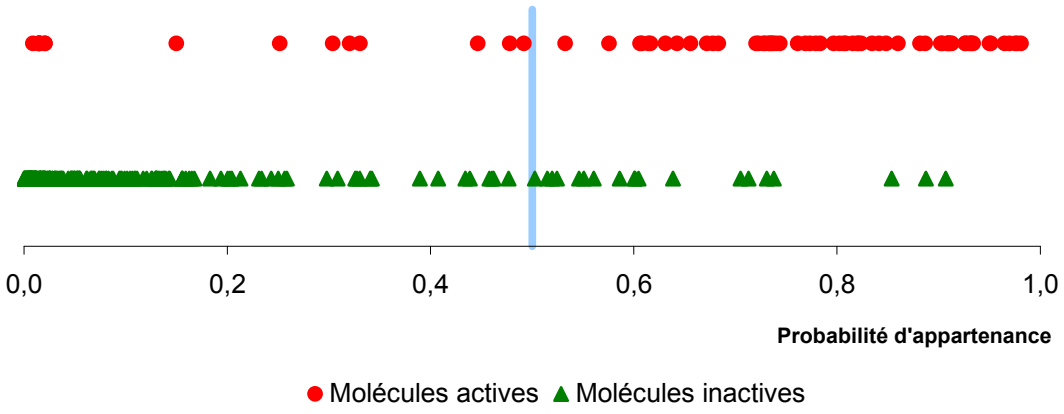


Figure 17. Représentation des individus selon leur probabilité d'appartenance au groupe des actifs

XlStat[100] est une extension de Excel qui nous a permis d'effectuer les calculs.

[100] http://www.xlstat.com

c) Performance des fonctions de scoring

(1) La matrice de confusion

(a) Principe

La qualité d'un modèle peut être évaluée de plusieurs manières possibles. Afin de mieux appréhender la notion de vrais/faux positifs/négatifs, nous avons représenté une matrice de confusion figurant les répartitions entre l'activité biologique des molécules et la catégorie dans laquelle le modèle les a prédites (Tableau 8). BA et BI sont respectivement le nombre de molécules biologiquement actives et biologiquement inactives provenant de l'*ensemble total*. PA et PI sont le nombre de molécules respectivement prédites actives et prédites inactives. Les VP et VN représentent le nombre de molécules bien prédites dans les deux catégories, encore appelées vrais positifs et vrais négatifs.

	Prédite Active	Prédite Inactive	Total des molécules Biologiquement connues
Biologiquement active (BA)	VP	FN	BA = VP + FN
Biologiquement inactive (BI)	FP	VN	BI = FP + VN
Total des molécules prédites	PA = VP + FP	PI = FN + VN	T = PA + PI = BA + BI

Tableau 8. Matrice de confusion résumant les différents ratios

(b) Calculs des pourcentages

Il est possible de calculer chaque valeur sous forme de pourcentage, afin de gagner en lisibilité (Équation 3).

$$\%_{VP} = \frac{VP}{PA} \times 100\,, \quad \%_{FP} = \frac{FP}{PA} \times 100\,, \quad \%_{FN} = \frac{FN}{PI} \times 100\,, \quad \%_{VN} = \frac{VN}{PI} \times 100$$

Équation 3. Vrais/faux positifs/négatifs

En résumé:

-Un vrai positif (VP) est une molécule «biologiquement active» prédite «*in silico* active»

-Un faux négatif (FN) est une molécule «biologiquement active» prédite «*in silico* inactive»

-Un faux positif (FP) est une molécule «biologiquement inactive» prédite «*in silico* active»

-Un vrai négatif (VN) est une molécule «biologiquement inactive» prédite «*in silico* inactive»

Ces ratios, qui caractérisent réellement les qualités du modèle, sont cependant difficilement utilisables lors de la phase d'optimisation de ce modèle. En effet, ils sont trop liés à la composition de l'ensemble de produits étudiés. Par conséquent, il est nécessaire de calculer, à partir de ces ratios, des indicateurs plus explicites.

(2) Enrichissement

(a) Principe

Une autre méthode permettant de mesurer la qualité d'un modèle est d'utiliser le facteur d'enrichissement «E_r». Celui-ci reflète le gain en composés actifs dans l'ensemble final créé par le modèle comparativement à un ensemble qui aurait été choisi au hasard. Une valeur plus faible que 1 signifie que le modèle a de moins bonnes performances qu'un modèle choisissant les molécules au hasard. Une valeur de 1 indique que le modèle a des performances similaires à un modèle qui choisirait les molécules au hasard. Enfin, dans le cas où le facteur d'enrichissement dépasse 1, cela signifie que le modèle obtenu surpasse les résultats d'une simple sélection au hasard des molécules.

(b) Calcul de E_r et E_r^{Max}

L'Équation 4 montre la manière dont le facteur d'enrichissement est calculé:

$$E_r = \frac{\left(\frac{VP}{PA}\right)}{\left(\frac{BA}{T}\right)}$$

Équation 4. Facteur d'enrichissement

Cette valeur ne doit être utilisée que pour comparer des modèles élaborés à partir d'un même *ensemble total*. En effet, le ratio E_r est fortement dépendant de la composition en

molécules actives et inactives de l'ensemble considéré. Par conséquent, un enrichissement maximal (E_r^{MAX}) doit être calculé si l'on utilise deux ensembles différents, cette valeur servant de référentiel de comparaison du E_r. La formule suivante illustre le calcul de l'enrichissement maximum obtenu au sein d'un modèle.

$$E_r^{MAX} = \frac{T}{BA}$$

Équation 5. Facteur d'enrichissement maximum

Avec un ensemble de produits ne contenant qu'une faible proportion de composés actifs, il est possible d'avoir une forte valeur de E_r, même avec un modèle de qualité moyenne. A l'inverse, un ensemble contenant une forte proportion de molécules actives n'aura jamais un facteur d'enrichissement E_r élevé. Un autre paramètre qui a une influence non négligeable dans l'interprétation du facteur d'enrichissement est la limite choisie pour séparer les molécules prédites actives de celles prédites inactives. Si nous choisissons une limite qui conduit à une très faible proportion de molécules prédites actives, le ratio VP/PA devient très haut (jusqu'à 1 si les premiers produits sont bien prédits) mais dans ce cas présent, le nombre de faux négatifs sera très élevé. Enfin, la comparaison de deux E_r ne peut être établie qu'en respectant deux règles fondamentales:

-Avoir la même composition de l'*ensemble total* entre les deux modèles à comparer.

-Avoir un nombre similaire de molécules prédites actives (PA, dans la matrice de confusion).

Dans notre étude, nous verrons que la majorité des modèles COX-2 vérifient les deux conditions qui permettent l'utilisation de cette métrique. Ce n'est pas le cas de PPARγ dans laquelle la deuxième règle n'est pas respectée. Nous n'avons donc pas toujours utilisé le E_r dans ce projet pour comparer les modèles.

Nous avons montré que la construction et la validation d'un modèle fiable, basée uniquement sur l'interprétation du E_r, reste limitée. Beaucoup de faux négatifs peuvent être récupérés avec ce type de validation.

(3) Molécules rejetées par le modèle

(a) Description

Le ratio R_a (R_a: Rejected active) est calculé sur la base du nombre de molécules actives de départ BA. Cette information est intéressante dans le cas de projets dans lesquels

des molécules actives ont des difficultés à pénétrer dans le site actif et qui sont donc directement éliminées.

(b) Calcul du R_a

Les performances d'un modèle, peuvent également être évaluées par l'estimation des actifs qui seraient perdus si l'on suivait strictement ce modèle lors du criblage réel (et donc si l'on ne testait réellement que les seuls produits prédits actifs par le modèle). Le nombre de molécules actives perdues par le modèle est calculé (Équation 6).

$$R_a = \frac{FN}{BA} \times 100$$

Équation 6. Nombre de molécules actives rejetées par le modèle.

Un modèle idéal aurait un nombre de faux négatifs nul et donc R_a serait alors égal à 0%. Un modèle qui rejetterait l'ensemble des positifs aurait un R_a à 100%. On cherchera donc un modèle ayant un R_a minimum. Comme nous le verrons dans le chapitre dédié à la cible PPARγ, le nombre de molécules actives rejetées par les modèles étant important, l'utilisation de cette métrique s'est révélée indispensable. Ce n'est pas le cas des modèles issus de la cible COX-2.

Mais cet indicateur n'est pas suffisant pour qualifier un modèle. En effet, si on ne se base que sur lui, il suffit d'augmenter le nombre de molécules prédites actives (en baissant le seuil de sélection) pour mécaniquement diminuer le R_a. En poussant ce raisonnement au maximum, un modèle qui prédirait l'ensemble des molécules proposées comme actives aurait un R_a idéal (0) mais serait à l'évidence totalement inutile.

(4) Courbes d'enrichissement

Dans un processus de scoring, il est possible de classer les composés selon leurs scores (ou selon une valeur composite provenant de la combinaison de plusieurs fonctions de scoring). A partir de ces données, des courbes d'enrichissement (encore appelées courbes d'accumulation) sont déduites, représentant le nombre de molécules biologiquement actives

récupérées durant le criblage de l'*ensemble total*. Les courbes d'enrichissement de deux modèles A et B ont été représentées (Figure 18).

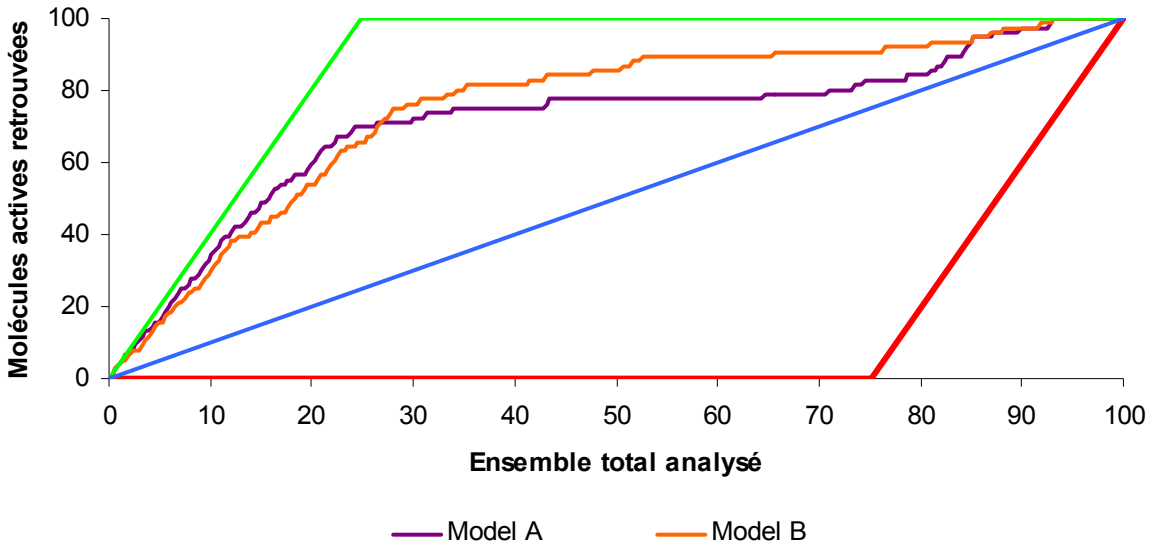


Figure 18. Exemples de courbes d'enrichissement de deux modèles

Ces courbes s'inscrivent dans un parallélogramme dont les côtés sont les limites d'un modèle. D'ailleurs, un modèle dont la courbe est proche de la courbe verte signifie que celui-ci prédit bien les composés actifs et les classe dans les premières portions du classement. Un modèle dont la courbe se rapproche de la ligne bleue a un pouvoir prédictif égal à celui que l'on aurait en prenant au hasard des composés dans l'*ensemble total*. Enfin, lorsque la courbe d'enrichissement est proche de la courbe rouge, cela signifie que le modèle classe les molécules inactives dans les premières portions du classement et les composés actifs à la fin du classement. Pour être utile, ce modèle doit être inversé. La discrimination des composés avec ce dernier modèle est aussi bonne que dans le cas de la courbe verte.

L'estimation qualitative de ces courbes permet d'en tirer des conclusions fondamentales quant à la puissance du modèle. Malgré tout, l'optimisation de modèles ne peut passer par cette évaluation visuelle. Seule une valeur numérique permet de détecter des problèmes majeurs d'enrichissement.

(5) Le R_s

(a) Principe

Ces courbes d'enrichissement sont à comparer aux courbes ROC,[101,102,103,104] utilisées dans la théorie de la détection de signal. En 1997, Bradley présente l'utilisation des courbes ROC (AUC, Area under curve) dans l'évaluation d'algorithmes d'apprentissage.[105] Les courbes ROC tracent les VP contre les FP tout au long du classement des composés. Dans cette représentation, la courbe idéale est constituée d'un premier fragment vertical figurant la récupération de 100% des actifs, suivie d'une ligne horizontale générée par les inactifs. Nos courbes d'enrichissement ne sont pas exactement les mêmes (les limites ne sont plus un rectangle mais un parallélogramme) mais le problème est similaire. Dans une publication récente, Truchon *et al* discutent de système métrique à utiliser dans le cas d'un criblage virtuel.[106] Les auteurs ont comparé l'aire sous la courbe ROC à l'aire sous la courbe d'enrichissement. Leur principal objectif a été d'analyser l'intérêt de ces métriques dans un problème de reconnaissance précoce des agents actifs, encore appelé «early recognition».

(b) Approche graphique du Rs

Nous avons utilisé une approche légèrement différente pour caractériser la qualité d'une courbe d'enrichissement. Nous avons pour cela calculé un ratio faisant intervenir l'aire entre la courbe du modèle et la diagonale ainsi que la moitié de l'aire du parallélogramme. Ceci est un excellent descripteur de la capacité du modèle à classer les molécules actives. Si cette valeur est proche de 1, alors le modèle est excellent et se superpose quasiment avec le côté supérieur du parallélogramme. Si la valeur est proche de 0, la courbe se rapproche de la diagonale du parallélogramme. Dans ce cas de figure, le modèle est aussi efficace qu'une sélection de composés faite au hasard. Dans le but de distinguer les cas où la courbe passe au dessus ou en dessous de la diagonale, nous avons fait varier le ratio de -1 à +1 comme dans le cas de coefficients de corrélation. Les limites extrêmes sont +1 (si la courbe est proche du

[101] Therrien, C.W. Decision and Classification: an introduction to pattern recognition and related topics *Wiley*, New York **1989.**
[102] www.jrocfit.org.
[103] www.medcalc.be/manual/roc.php.
[104] Azé, J.; Roche, M.; Kodratoff, Y.; Sebag, M. Learning to order terms: supervised interestingness measures in terminology extraction *IJCIO* **2004**, *1*, 1304-4508.
[105] Bradley, A.P. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms *Pattern recognition* **1997**, *30*, 1145-1159.
[106] Truchon, J.F.; Bayly, C.I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem *J. Chem. Inf. Model.* **2007**, *47*, 488-508.

côté du parallélogramme situé en haut à gauche), 0 (si la courbe est proche de la diagonale) et -1 (si la courbe est proche du côté du parallélogramme situé en bas à droite). Nous avons noté ce ratio «R_s». Le R_s peut être directement utilisé dans un processus d'optimisation automatique, prenant en compte sa valeur élevée au carré. Le processus d'optimisation va guider la valeur du R_s soit vers +1 ou -1. Le signe permet de choisir l'ordre global des produits.

Le calcul de l'aire sous la courbe (AUC) dans le cas des courbes ROC n'est pas aisé. La méthode la plus utilisée est la méthode des trapèzes, ne fournissant qu'une approximation de cette aire. Dans le cas du calcul de l'aire sous la courbe d'enrichissement (AUAC), Truchon *et al* proposent deux méthodes de calcul qui donnent des résultats similaires: lorsque l'ensemble est petit, les différences sont observables mais dans le cas où la taille des ensembles augmente, les deux valeurs AUC et AUAC convergent. Le maximum de ces valeurs (AUC et AUAC) dépend fortement de la composition en actifs dans l'ensemble. Dans notre approche, nous n'avons pas essayé d'approximer une aire mais plutôt de calculer un ratio entre deux aires. L'avantage de notre méthode est que cette valeur est univoque et facilement calculable comparée aux autres méthodes décrites précédemment.

Pour plus de compréhension, nous allons utiliser l'exemple d'un ensemble de taille limitée. Les côtés du parallélogramme sont délimités par des courbes en forme d'escalier (Figure 19).

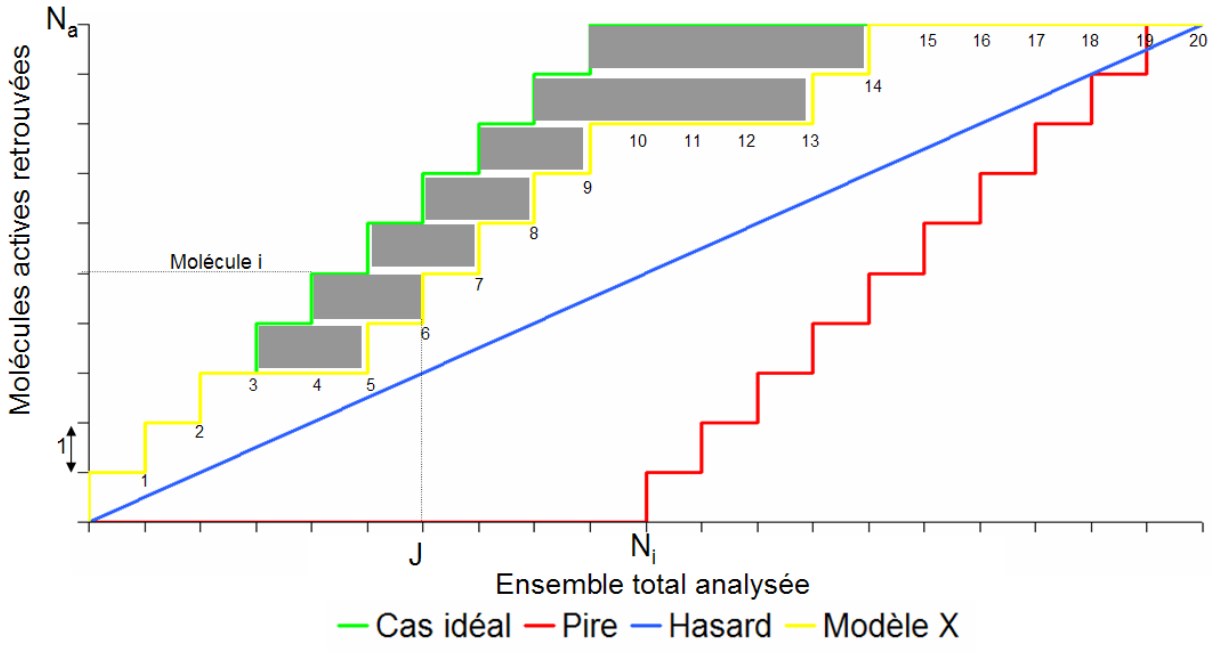


Figure 19. Principe détaillé des courbes d'enrichissement et mode de calcul du R_s

Dans cet exemple, un ensemble constitué de 10 molécules actives et 10 molécules inactives a été évalué. Les produits actifs ont été classés dans les positions n°1/2/3, 6/7/8/9/10, 14 et 15. La courbe jaune est ainsi créée à partir de ces valeurs.

(c) Calcul du R_s

La clé du calcul repose sur l'évaluation de l'aire figurée en gris. Nous avons noté S_R cette aire et S_P la surface du parallélogramme. La formule du R_s est alors:

$$R_s = 1 - \frac{S_R}{\frac{1}{2} \times S_P}$$

Équation 7. Expression du R_s

En considérant R(i) le rang de la $i^{ème}$ molécule active et en utilisant les notations précédentes (BA le nombre de molécules biologiquement actives et BI le nombre de molécules biologiquement inactives), la surface d'un des rectangles S_r (de longueur R(i) – I et de largeur 1) est:

s_r = (R(i) – i) x 1

La surface de la zone grise SR est

$$S_R = \sum_{i=1}^{BA} (R(i) - i) = \sum_{i=1}^{BA} R(i) - \sum_{i=1}^{BA} i = \sum_{i=1}^{BA} R(i) - \frac{(BA \times (BA+1))}{2}$$

La surface du parallélogramme S_p est la somme des rectangles BA (de longueur BI et de largeur 1):

S_p = BA x BI x 1 = BA x BI

Par conséquent:

$$R_s = 1 - \frac{\sum_{i=1}^{BA} R(i) - \frac{(BA \times (\mathrm{BA}+1))}{2}}{\frac{1}{2} \times BA \times BI} = 1 - \frac{2 \times \sum_{i=1}^{BA} R(i)}{\mathrm{BA} \times BI} + \frac{\mathrm{BA}+1}{BI} = \frac{BI + BA + 1}{BI} - \frac{2 \times \sum_{i=1}^{BA} R(i)}{BA \times BI}$$

D'où:

$$R_s = \frac{BI + BA + 1}{BI} - \frac{2 \times \sum_{i=1}^{BA} R(i)}{BA \times BI}$$

Dans l'exemple d'enrichissement décrit précédemment, S_P=10x10=100. L'aire de la surface grise est S_R=(0+0+0+2+2+2+2+2+5+5)=20 et donc R_s=0,6. Dans la figure 18, les deux modèles ont respectivement un R_s de 0,6 et 0,7.

Il peut être utile de comparer le R_s avec la deuxième forme de AUAC. En utilisant nos annotations, la formule de l'AUAC serait:

$$AUAC = 1 - \frac{\sum_{i=1}^{N_a} R(i)}{BA \times (BA + BI)}$$

Équation 8. Aire sous la courbe d'enrichissement (AUAC)

(d) Comparaison de R_s avec les indicateurs des courbes ROC

On remarque que ces deux formules ont une structure similaire. La fraction présente la somme des rangs des actives au numérateur et une combinaison du nombre de molécules actives et inactives au dénominateur. Il demeure néanmoins que les résultats proposés par ces deux méthodologies de calcul sont légèrement différents. Les limites du R_s sont constantes (intervalle de -1 à 1) et sont indépendantes du ratio de molécules positives dans l'*ensemble total*. Dans l'exemple précédemment évoqué (10 actives, 10 inactives), AUAC=0,625 or 0,65 (selon la formule utilisée). Ces valeurs sont à comparer à la valeur maximale 0,75. Avec le R_s, nous avons la possibilité de mesurer de manière plus simple que AUAC la tendance de

l'enrichissement au sein d'un modèle. En outre, le R_s permet de s'affranchir d'une valeur maximale utilisée en tant que référentiel.

Notre indicateur peut être utilisé pour optimiser les protocoles de criblage virtuel mais possède les mêmes faiblesses que AUC et AUAC pour différencier trois cas basiques mais très théoriques proposés par Truchon *et al*:

-la moitié des actives est récupérée au tout début de la liste des rangs et la deuxième partie est retrouvée à la fin.

-les molécules actives sont distribuées aléatoirement tout au long des rangs

-toutes les molécules actives sont retrouvées dans le milieu de la liste.

Dans chacun de ces trois cas, le R_s est égal à 0. Durant la phase de paramétrage du modèle, chacun de ces trois cas est considéré comme mauvais résultat et il est très délicat de privilégier un des cas plutôt qu'un autre.

G. Résultats et discussion

Le travail ici réalisé s'articule en trois parties. La première partie traite de la comparaison des différentes composantes du modèle (structures cristallographiques, fonctions de scoring). Une deuxième partie expose la validation du modèle. Une troisième et dernière partie montre l'intérêt d'un modèle pharmacophorique en amont du protocole de docking-scoring en vue d'une application au criblage virtuel.

1. Comparaison des structures cristallographiques de cyclooxygénase de type 2

a) Etude de la 1CX2 d'origine

Les premières structures de COX-2 ont été proposées par Kurumbail *et al*.[107] Plus particulièrement, une structure a été extraite de la PDB (code: 1CX2). Cette forme cristallographique est résolue avec le SC-558, molécule de la famille des coxibs. L'*ensemble total* décrit précédemment a été utilisé pour évaluer la capacité prédictive de cette structure tridimensionnelle. FlexX score (F-Score) étant la fonction utilisée pour diriger la reconstruction du ligand au sein du site actif, nous l'avons appliqué tout au long de notre

[107] Kurumbail, R.-G. ; Stevens, A.-M. ; Gierse, J.-K. ; McDonald, J.-J. ; Stegeman, R.-A. ; Pak, J.-Y. ; Gildehaus, D. ; Miyashiro, J.-M. ; Penning, T.-D. ; Seibert, K. ; Isakson, P.-C. ; Stallings, W.-C Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents *Nature* **1996**, *384*, 644-648.

étude pour déterminer la première pose. Dans un premier temps, aucune fonction de re-scoring n'a été utilisée. L'enrichissement en molécules actives a été représenté dans la figure 20.

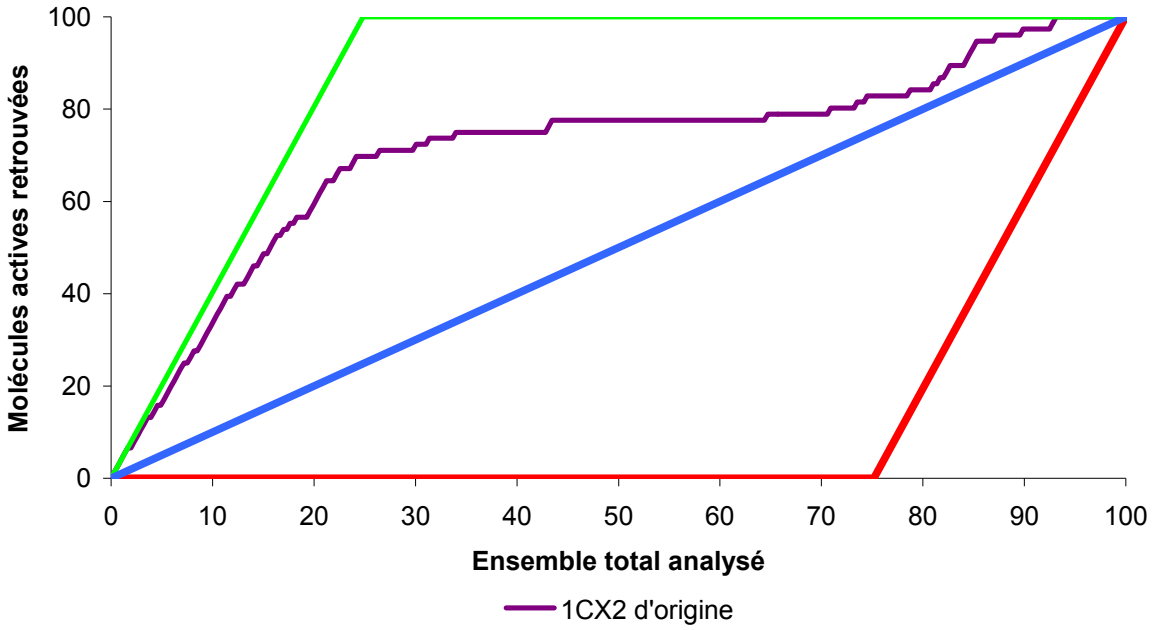


Figure 20. Courbe d'enrichissement de la structure 1CX2

Visuellement, l'allure de la courbe d'enrichissement montre un bon classement des composés actifs dans les premiers pourcentages de l'*ensemble total* analysés. Malgré tout, l'aire entre le cas idéal (courbe verte) et le modèle 1CX2 reste assez importante. De plus, une inflexion de la courbe d'enrichissement s'étendant de 75 à 100% laisse imaginer qu'une ou plusieurs famille(s) de molécules actives sont prédites de manière incorrecte par le modèle. Nous avons calculé les ratios décrits précédemment:

E_r=2,8; E_r^{max}=4,0; R_s=0,56.

Ces valeurs sont acceptables puisque le $E_r^{25\%}$ (le facteur d'enrichissement a toujours été estimé à 25% de l'*ensemble total* dans cette étude de la cyclooxygénase) calculé est supérieur à 1,0 et est plus proche de E_r^{max}. Le R_s est également supérieur à la limite pour lequel un modèle est défini comme mauvais, c'est-à-dire 0. Le comportement atypique de la courbe d'enrichissement dans les dernières portions de l'*ensemble total* nous a conduits à analyser les ligands mis en cause dans ce phénomène. Il s'agit de molécules possédant

majoritairement une fonctionnalité spiro sur l'hétérocycle «A». L'activité biologique de ces composés étant les plus fortes des 76 molécules actives, nous avons souhaité comprendre le mécanisme par lequel elles étaient prédites inactives. Au total, 24 composés ont été identifiés comme présentant le groupement spiro. Nous avons tracé la courbe d'enrichissement de ces 24 composés auxquels ont été ajoutés les 230 composés actifs (Figure 21).

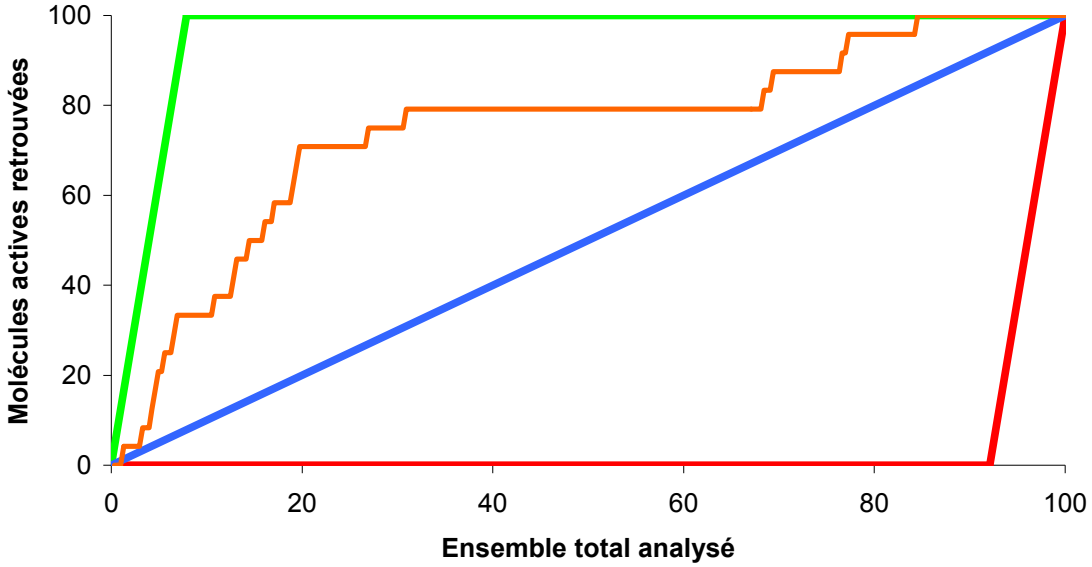


Figure 21. Courbe d'enrichissement des composés coxibs possédant une fonctionnalité spiro

Cette figure met en évidence la défaillance de l'algorithme de docking-scoring dans lequel les 24 molécules possédant un groupement spiro ont un comportement irrégulier. Nous avons tenté de déterminer les raisons de ce phénomène. La première hypothèse est qu'il est quelquefois difficile de prédire des molécules possédant des cycles tendus du fait de leur énergie interne parfois trop élevée. Dans d'autres projets, nous avons eu des problèmes similaires dans lesquels ces types de molécules sont mal traités. Une autre explication est l'encombrement stérique de ce type de groupement mais également et surtout la flexibilité très limitée d'une telle fonctionnalité. Du fait de cette rigidité d'ensemble, des phénomènes de recouvrement entre la protéine et le ligand peuvent avoir lieu et par conséquent pénaliser les molécules possédant ces groupements. Dans la cyclooxygénase de type 2, les spiros sont à proximité de l'Arg120. Ce résidu clé des COX-2 se localise à l'extrémité d'une région à tendance lipophile dans laquelle vient normalement se loger le sommet du cycle A (correspondant au groupement trifluorométhyle). Des variations de la chaîne latérale de

l'Arg120 sont possibles. Dans certaines positions, ces mouvements peuvent induire des gênes stériques. Afin d'observer de telles variations structurales, nous avons superposé deux structures cristallisées avec des ligands différents. Dans un environnement proche de l'Arg120, la structure 1CX2 (cristallisée avec un coxib, le SC-558) a été superposée à la structure 3PGH (cristallisée avec un non-coxib). Dans la figure suivante, la position de l'Arg120 dans la 3PGH est représentée en violet tandis que l'Arg120 de la 1CX2 est symbolisée en rouge (Figure 22).

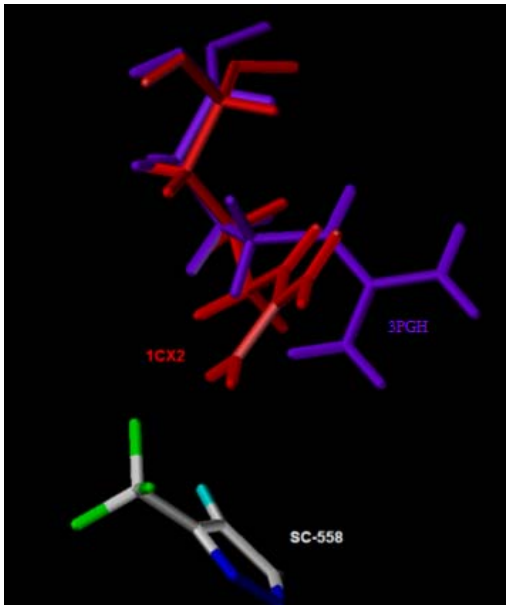


Figure 22. Variations de la chaîne latérale du résidu Arg120 de la structure 1CX2 à 3PGH

L'Arg120 est animée d'un mouvement lorsque l'on passe d'une structure cristallisée avec un coxib à une structure cristallisée avec un non-coxib. Il apparaît clairement que la chaîne latérale de l'Arg120 est à proximité du groupement CF_3 du SC-558. Par contre, l'Arg120 provenant de la 3PGH laisse un espace supplémentaire du côté du groupement trifluorométhyle. Afin d'augmenter le volume accessible dans cette région lipophile du site actif, nous avons déplacé la chaîne latérale de l'Arg120 provenant de 1CX2 dans une position similaire à celle de 3PGH. Le nouveau modèle tridimensionnel ainsi créé a été évalué par le protocole de docking-scoring. Les courbes d'enrichissement suivantes permettent de comparer le pouvoir prédictif du modèle d'origine 1CX2 et le modèle modifié 1CX2 (Figure 23).

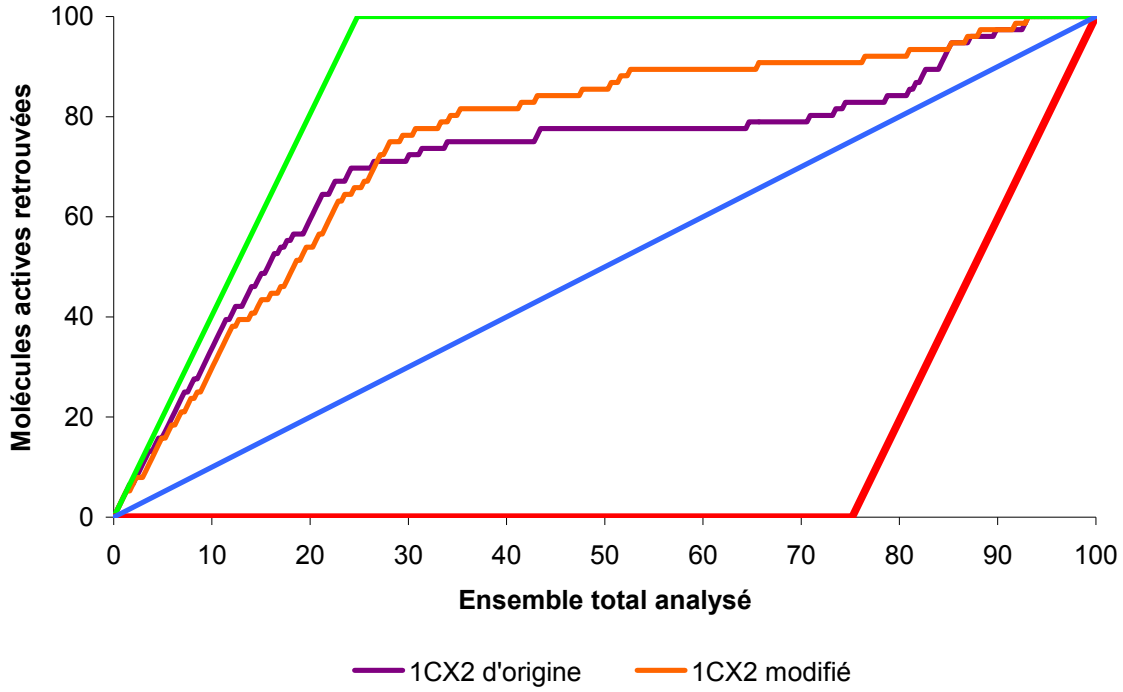


Figure 23. Courbes d'enrichissement des deux modèles d'1CX2 original et modifié

Les courbes d'enrichissement sont en faveur du modèle 1CX2 modifié dans le dernier intervalle de l'*ensemble total* (entre 75 et 100%). Cette démonstration illustre l'incidence de l'Arg120 sur le docking d'une famille de ligands particulière. Grâce au déplacement de l'Arg120 de 1CX2 dans une position proche de celle de la 3PGH, nous améliorons la prédiction des composés porteurs du groupe spiro. Toutefois, le 1CX2 modifié possède de moins bonnes performances que le 1CX2 d'origine dans les premiers 25% de l'*ensemble total*. Ce comportement s'explique facilement par le fait qu'en modifiant la position de l'Arg120, nous augmentons l'espace accessible dans cette région. En contrepartie, nous occasionnons également la perte d'interactions potentielles avec Arg120 qui devient trop éloignée des ligands pour interagir. Afin de mieux caractériser le pouvoir prédictif des deux modèles, nous avons calculé le facteur d'enrichissement E_r ainsi que le R_s:

$E_r^{\text{1CX2 d'origine}}$=2,8; $E_r^{\text{1CX2 modifié}}$=2,6; E_r^{max}=4,0; $R_s^{\text{1CX2 d'origine}}$=0,56 et $R_s^{\text{1CX2 modifié}}$=0,67.

Le choix d'un modèle destiné au criblage virtuel doit se faire en accord avec ces différents indicateurs. Plus précisément, le R_s montre que nous avons amélioré le classement d'une famille de molécules au dépend de l'enrichissement E_r. Le facteur E_r informe plus sur la prédiction de l'ensemble final après un processus de criblage virtuel que ne le fait le R_s. Par

conséquent, le modèle 1CX2 d'origine est le plus performant pour un processus de criblage virtuel. Mais ceci se fera au prix de la perte d'une famille de composés actifs.

b) Optimisation de la structure 1CX2

Nous avons conservé le modèle 1CX2 d'origine pour nos études ultérieures. Les états de protonation de différents résidus ont été vérifiés et optimisés. Différents états de protonation de l'His90, Glu520&524, Ser353&530 et l'Asp515 ont été explorés et testés. L'histidine est un résidu particulier puisqu'il est capable d'exister sous différents états de protonation (Figure 24). Ce résidu est capable aussi bien d'agir en tant que donneur d'hydrogène (position ε protonée, c'est-à-dire les cas de Hie et Hip) mais également comme accepteur d'hydrogène (position ε déprotonée, c'est-à-dire Hie).

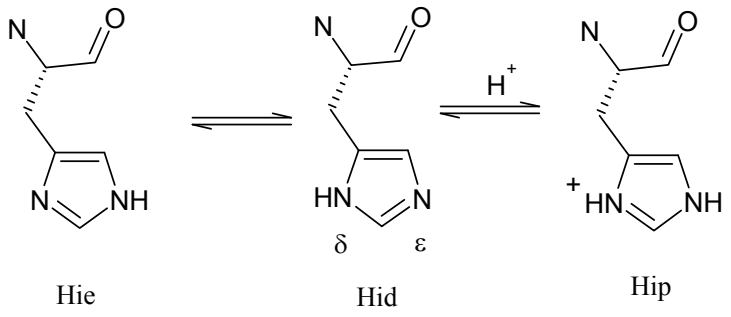


Figure 24. Etats de protonation possibles du résidu histidine

On retrouve ces deux cas de figure où l'histidine 90 est soit donneur d'hydrogène (cas de 6COX), soit accepteur d'hydrogène (cas de 1CX2). Dans la structure 6COX, l'His90 interagit avec un oxygène du groupement sulfonyle tandis que l'His90 de 1CX2 interagit avec le groupement NH_2 de la sulfonamide. Il est difficile de déterminer le pK_a de chacune de ces formes. Nous avons donc testé les deux états Hie et Hid:

E_r^{Hid}=2,8; E_r^{Hie}=2,4; E_r^{max}=4,0; R_s^{Hid}=0,56 et R_s^{Hie}=0,57

Les deux états de protonation Hid et Hie permettent d'avoir des résultats semblables avec toutefois une préférence pour la forme Hid présente dans 1CX2. Nous avons gardé l'état de protonation Hid pour le reste de l'étude. L'angle de torsion du groupe hydroxyle de la Tyr385 a été positionné à 160° afin de favoriser le potentiel donneur d'hydrogène.

L'hydroxyle de la Ser530 a également été modifié pour obtenir un angle de torsion de 180°. L'optimisation des différents états de protonation ainsi que la réorientation des différents groupements hydroxyle susceptibles d'entrer en liaison hydrogène ont nettement amélioré les résultats de prédiction:

$E_r^{1CX2\ d'origine}$=2,8; $E_r^{1CX2\ optimisé}$=2,8; E_r^{max}=4,0; $R_s^{1CX2\ d'origine}$=0,56 et $R_s^{1CX2\ optimisé}$=0,59

La forme optimisée de 1CX2 a un pouvoir prédictif supérieur à ce que l'on pouvait espérer avec le 1CX2 d'origine. Les résultats du modèle 1CX2 optimisé prouvent l'intérêt de prendre en compte l'état de protonation de certains résidus ainsi que l'orientation de groupements capables de créer des interactions hydrogène. Par conséquent, nous avons conservé cette structure tridimensionnelle 1CX2 optimisée.

Nous avons souhaité appliquer des fonctions de scoring à la première pose sélectionnée selon F-Score. Pour cela, les fonctions de scoring contenues dans le module C-Score ont été déployées.

2. Comparaison des fonctions de scoring dans un processus de re-scoring

Le premier test dans la comparaison de fonctions de scoring est d'observer le degré de corrélation entre elles. Des fonctions trop proches de 1,00 prouvent qu'elles sont trop semblables et ne sont donc pas utilisables dans les analyses par consensus que nous développerons ultérieurement. Le tableau suivant montre ces corrélations:

	F-Score	D-Score	PMF	G-Score	Chemscore
F-Score	1,00	0,23	0,56	0,48	0,10
D-Score		1,00	0,26	0,85	0,78
PMF			1,00	0,65	0,42
G-Score				1,00	0,41
Chemscore					1,00

Tableau 9. Matrice de corrélation entre les différentes fonctions

Le tableau 9 met bien en évidence que quasiment toutes les fonctions de scoring sont peu corrélées. D-Score et G-Score sont les seules à montrer un coefficient plus élevé que les

autres. Nous nous attendons donc à avoir des résultats similaires entre ces deux fonctions de scoring qui ont pourtant été optimisées différemment.

Afin d'optimiser l'utilisation des fonctions de scoring dans le processus de consensus, nous avons souhaité comparer les fonctions de scoring les unes entre les autres et de manière individuelle. Ces comparaisons ont été effectuées en utilisant la forme cristallographique 1CX2 optimisée. La première pose, choisie selon F-Score a été re-scorée par les 4 autres fonctions de scoring. Les courbes d'enrichissement ont été générées pour chaque fonction de re-scoring traitée individuellement (Figure 25).

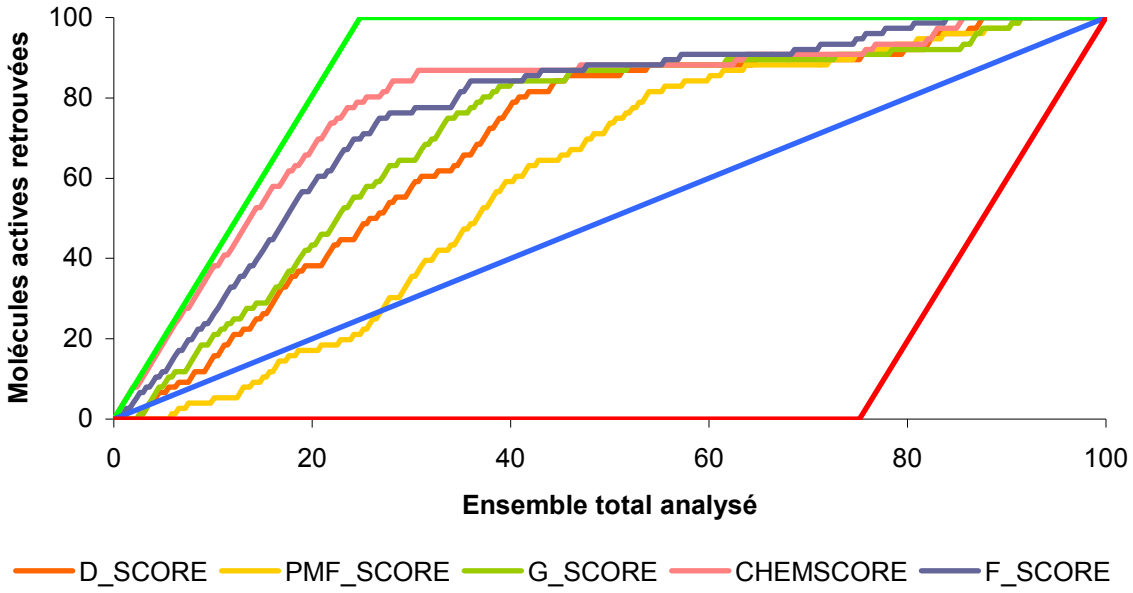


Figure 25. Courbes d'enrichissement de chaque fonction de scoring dans un processus de re-scoring

Les courbes d'enrichissement montrent des disparités d'une fonction à une autre, prouvant leur variabilité vis-à-vis d'un même profil pharmacologique. Chemscore s'approche le plus de la courbe du cas idéal et est donc considérée comme la fonction la plus performante d'un point de vue prédictif. A l'inverse, la fonction PMF se comporte de manière proche à la ligne de prédiction bleue correspondant à une sélection effectuée au hasard. Le point commun à chaque courbe d'enrichissement est le point d'inflexion entre 70 et 80% de l'*ensemble total* analysé. Ceci montre que la défaillance des molécules possédant la fonctionnalité spiro provient bien du docking (en particulier de l'emplacement de la molécule à proximité de Arg120) et non d'une fonction de scoring (problèmes de cycles tendus). Nous avons quantifié ces courbes par le E_r et le R_s (Figure 26).

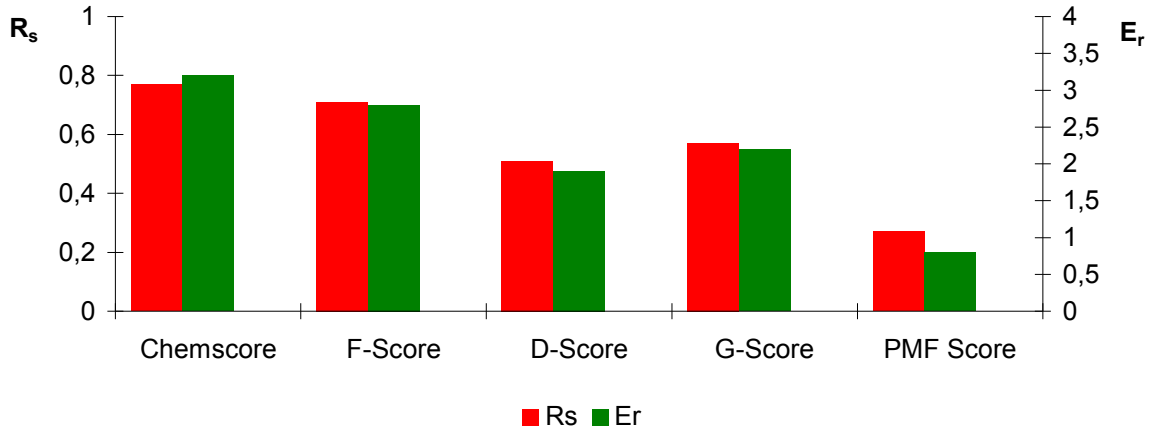


Figure 26. Comparaison des fonctions de scoring selon les ratios E_r et R_s

Dans cet histogramme, nous avons représenté sur l'ordonnée de gauche l'échelle du R_s, avec comme valeur maximum 1,0. Sur l'ordonnée de droite est figurée l'échelle de E_r dont le maximum est 4,0. Ce graphique met en évidence la force des fonctions de scoring Chemscore, F-Score ou même G-Score, tandis que PMF est la seule fonction dont le E_r est inférieur à 1,0 (E_r^{PMF}=0,8). Cette comparaison montre que cette fonction de scoring n'est pas adaptée au scoring d'inhibiteurs de la cyclooxygénase. Par ailleurs, il semble déraisonnable de n'utiliser qu'une seule fonction de scoring pour re-scorer les premières poses de nos inhibiteurs COX-2. Nous avons donc souhaité étudier l'influence des fonctions de scoring combinées entre elles dans différentes stratégies de consensus.

a) Stratégies de consensus

(1) Les méthodologies «Rank by rank», «Rank by number» et «Rank by best»

Les différentes stratégies de consensus envisagées dans cette étude ont été décrites précédemment. Les trois méthodologies qui nous ont paru intéressantes sont «Rank by rank», «Rank by number» et «Rank by best». Les fonctions de scoring sont combinées une à une, les moins performantes étant ajoutées le plus tard possible. Aux vues des résultats précédents, il est logique de commencer par la fonction de scoring chemscore car c'est elle qui fournit les meilleurs résultats en termes de E_r et R_s. Seul le R_s a été utilisé pour comparer les consensus

puisque ce ratio est capable de comparer en profondeur l'évolution de l'enrichissement durant l'analyse de l'*ensemble total* (Figure 27).

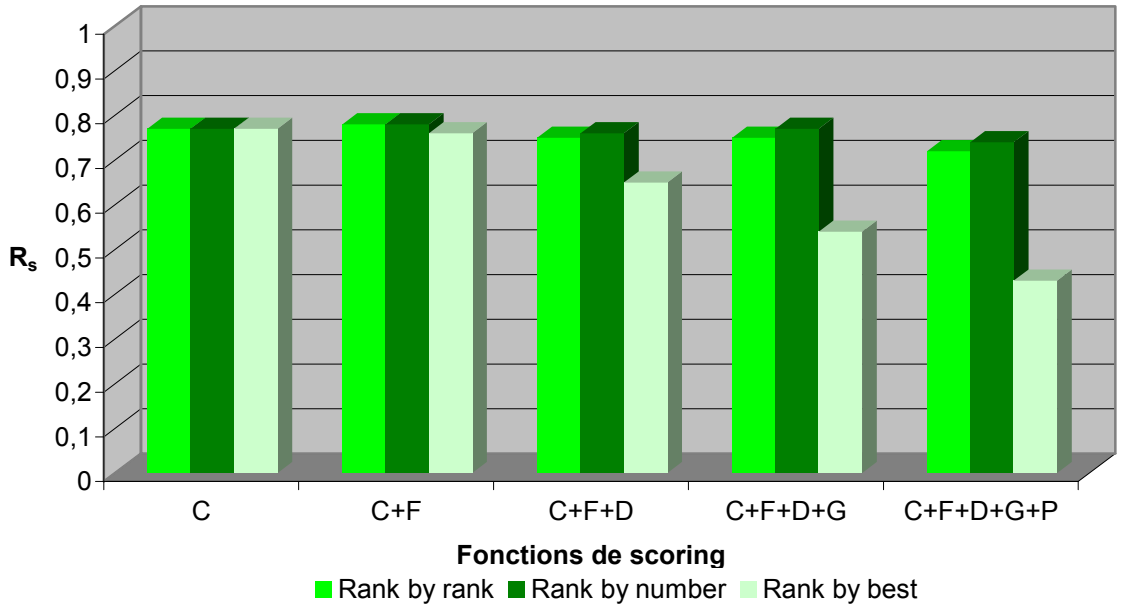


Figure 27. Comparaison de stratégies de consensus selon la composition en fonctions de scoring

C (Chemscore), F (F-Score), D (D-Score), G (G-Score) et P (PMF)

Nous avons débuté cette étude par l'ajout des deux fonctions de scoring les plus performantes individuellement (Chemscore et F-Score). Nous avons ensuite ajouté une à une les fonctions qui apportaient une information prédictive décroissante. Nous pouvons apercevoir qu'à partir de l'addition de D-Score, le modèle devient moins prédictif, jusqu'à une dégradation significative lors de l'ajout de PMF. La méthode «Rank by best» est la méthode la plus sensible aux fonctions de scoring peu performantes. En effet, le R_s passe de 0,8 (dans le cas C+F) à 0,3 (pour C+F+D+G+P). Contrairement aux résultats proposés par l'histogramme E_r/R_s comparant les fonctions de scoring individuellement, G-Score dégrade bien plus la stratégie «Rank by best» que ne le fait D-Score. Pourtant, G-Score seul était plus prédictif que ne l'était D-Score. En règle générale, dans le cas des méthodes «Rank by rank» et «Rank by number» la dégradation est moins importante. Pour ces deux techniques, le meilleur des cas (C+F), correspond à un R_s de 0,8 contre 0,7 dans le pire des cas (C+F+D+G+P). Les méthodes «Rank by rank» et «Rank by number» sont donc bien moins sensibles aux fonctions de scoring peu performantes que la technique «Rank by best». De

manière pragmatique, si nous avions à utiliser une stratégie, nous choisirions aussi bien «Rank by rank» que «Rank by number», avec comme composition en fonctions de scoring, Chemscore et F-Score.

(2) La méthodologie «Rank by vote»

La méthodologie «Rank by vote», majoritairement qualitative, ne convient pas à une analyse par le R_s qui nécessite des données quantitatives (Figure 28). Pour cette stratégie, nous avons associé les deux meilleures fonctions de scoring (Chemscore et F-Score).

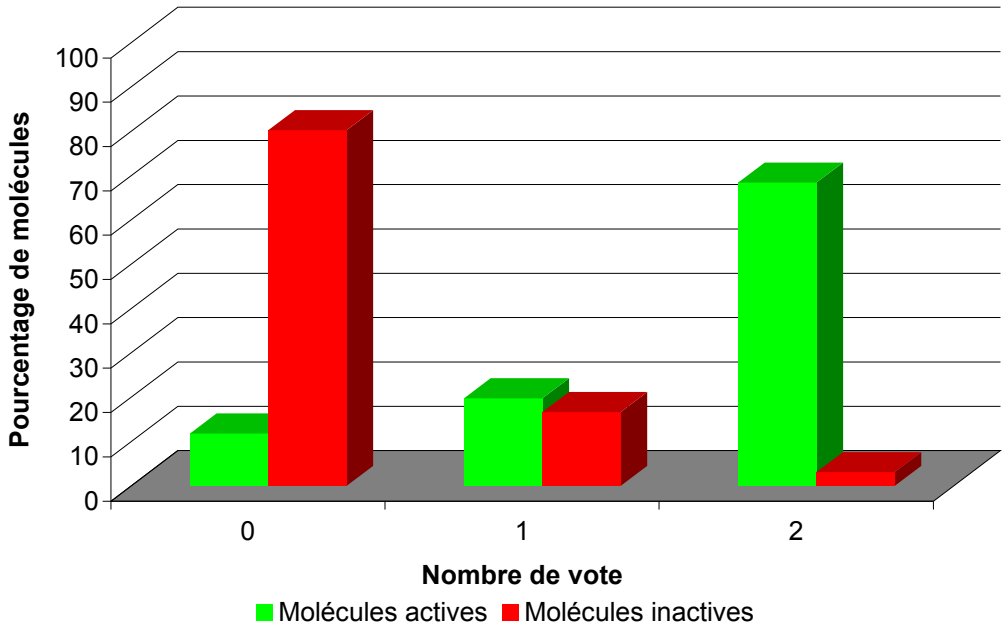


Figure 28. Evaluation de la stratégie «Rank by vote» avec Chemscore et F-Score

Pour chaque fonction de scoring, une limite a été déterminée afin d'attribuer à chaque molécule la classe active ou inactive. Cette valeur, permettant d'attribuer un vote à une molécule, a été optimisée pour chaque fonction de scoring. La fréquence des molécules est représentée en ordonnée et le nombre de votes reçus par chaque molécule en abscisse. Grâce à cet histogramme, nous pouvons visualiser distinctement les trois cas de figures:

-69% des molécules actives contre 3% des molécules inactives totalisent 2 votes

-20% des molécules actives contre 17% des molécules inactives totalisent 1 vote

-11% des molécules actives contre 81% des molécules inactives totalisent 0 vote

En résumé, lorsqu'une molécule a 2 votes, elle a de fortes chances d'être active. Dans le cas opposé, si une molécule n'a pas de vote, elle a quasiment toutes les chances d'être inactive. La zone d'incertitude dans laquelle une molécule n'a qu'une seule fonction de scoring ayant voté pour elle ne permet pas de déterminer la catégorie de la molécule. L'aspect discriminatoire de cette stratégie est quelque peu différent des deux autres méthodes de consensus. Elle présente un aspect plus qualitatif que les deux autres stratégies («Rank by rank» et «Rank by number») dont l'aspect est purement quantitatif. C'est la raison pour laquelle nous ne l'avons pas utilisée dans le criblage virtuel qui est un protocole nécessitant des données chiffrées. Par ailleurs, elle impose trop de choix qui sont parfois arbitraires (comme le choix de la limite pour attribuer un vote ou encore le nombre de votes nécessaires pour déterminer la catégorie de la molécule). La stratégie «Rank by vote» demeure toutefois très intéressante et discriminante.

b) Stratégie d'analyse factorielle discriminante

Nous avons également souhaité utiliser une méthodologie d'analyse de données autre que les méthodes de consensus «classiques». Les analyses de données multivariées telles que l'analyse factorielle discriminante permettent de combiner les informations issues des fonctions de scoring. Les courbes d'enrichissement, le E_r et le R_s ont été utilisés pour évaluer les modèles. Les variables dans ce genre de méthodes d'analyse correspondent aux fonctions de scoring. Les individus sont les molécules. Enfin, les variables qualitatives sont les catégories possibles pour les molécules (actif ou inactif). La courbe d'enrichissement ci-dessous montre la qualité prédictive obtenue en utilisant l'AFD pour interpréter les résultats de re-scoring. Contrairement aux méthodes de consensus nécessitant une sélection préalable des fonctions de scoring, l'analyse factorielle discriminante a été réalisée en présence des cinq fonctions de scoring (Figure 29).

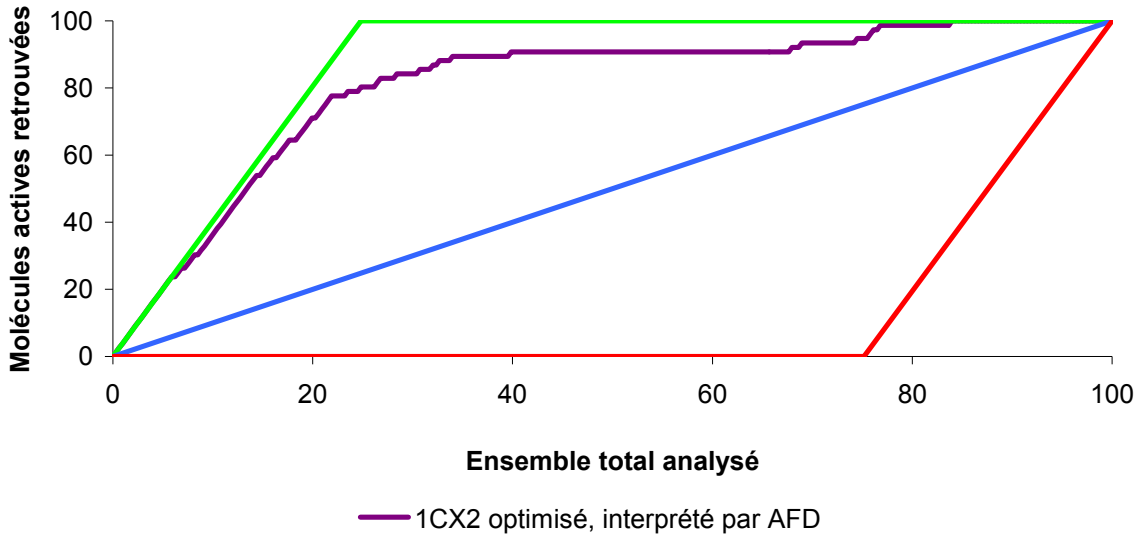


Figure 29. Courbe d'enrichissement du modèle 1CX2 optimisé

L'utilisation de l'analyse factorielle discriminante permet d'améliorer la tendance de la courbe d'enrichissement qui tend fortement vers la courbe verte de l'enrichissement idéal. Le R_s passe de 0,77 (par simple interprétation de la meilleure fonction de scoring Chemscore) à 0,82 (par interprétation des scores selon l'analyse factorielle discriminante). La courbe d'enrichissement montre une inflexion entre 75 et 80% de l'*ensemble total* analysé, due aux molécules possédant la fonctionnalité spiro. Ceci confirme la défaillance de cette famille de molécules qui provient uniquement de leur mode de docking (encombrement stérique à proximité de l'Arg120) et non des fonctions de scoring. Nous avons comparé les résultats obtenus par AFD avec ceux provenant de la meilleure combinaison de fonctions de scoring par consensus.

c) Comparaison de l'AFD et des méthodes de consensus

Nous avons comparé l'analyse factorielle discriminante avec les méthodes de consensus «Rank by rank» et «Rank by number», toutes deux avec Chemscore et F-Score:

$$R_s^{AFD}=0,82;\ R_s^{\text{Rank by rank}}=0,72,\ R_s^{\text{Rank by number}}=0,74$$

Le gain apporté par l'analyse factorielle discriminante est donc significatif. Par conséquent, nous l'avons utilisé comme méthode de référence pour le criblage virtuel.

Maintenant, nous allons évaluer d'autres structures cristallographiques de la PDB à l'aide de la même méthodologie.

3. Evaluation des 7 structures cristallographiques

Afin d'établir une comparaison équitable, les états de protonation ainsi que l'orientation des groupements hydroxyle de 1CX2 ont été attribués aux six structures (1CVU, 1DDX, 1PXX, 3PGH, 4COX et 6COX). Le protocole de docking décrit précédemment a été utilisé. Les cinq fonctions de scoring (F-Score, D-Score, Chemsore, G-Score, PMF) ont été utilisées dans une stratégie de re-scoring et combinées par AFD. Dans un premier temps, nous avons voulu observer la répartition des vrais/faux positifs/négatifs au sein de chaque modèle (Figure 30).

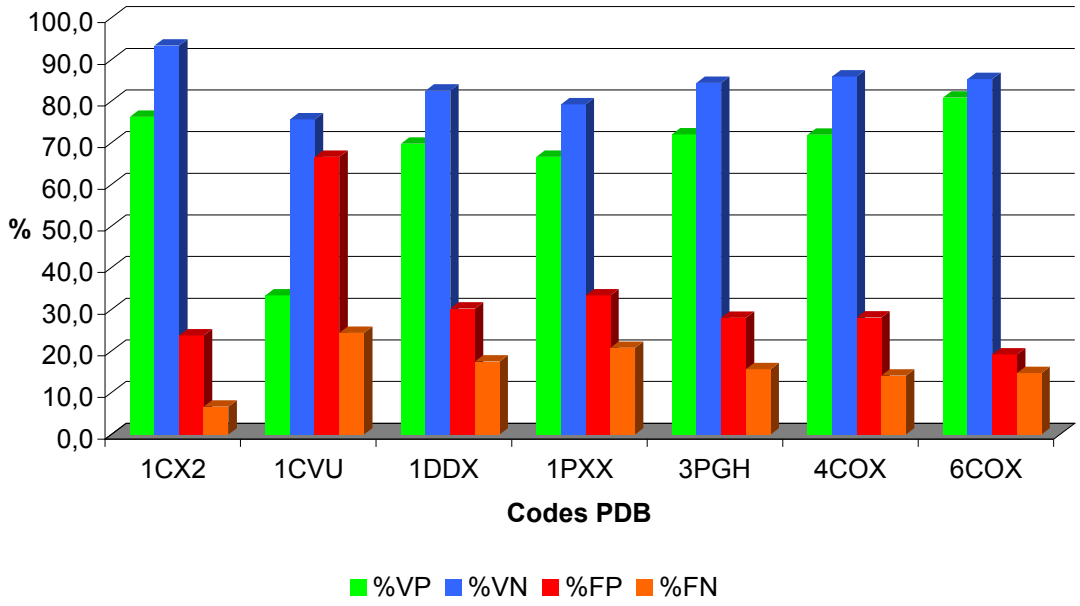


Figure 30. Distribution des vrais/faux positifs/négatifs dans chaque forme cristallographique

Bien que le site actif des COX-2 soit majoritairement considéré comme rigide, des changements importants de répartition des vrais/faux positifs/négatifs mettent en évidence la variabilité du site de fixation des ligands. En considérant les taux de vrais positifs/négatifs, il apparaît que les formes 1CX2 et 6COX sont les meilleures à prédire l'*ensemble total*. Cette observation est logique du fait que l'*ensemble total* est entièrement constitué de molécules de la famille des coxibs et que ces deux structures cristallographiques sont cristallisées avec des

molécules coxib. Les différences sont spectaculaires lorsque l'on compare par exemple 1CX2 avec 1CVU. Le ratio de vrais positifs passe respectivement de 76 à 33%. Les disparités en terme de vrais négatifs sont moins extrêmes (ce ratio passe de 76% dans le pire des cas 1CVU à 93% dans le meilleur modèle cristallographique 1CX2). Selon les ligands co-cristallisés, des différences peuvent opérer, prouvant que le site actif est susceptible de subir des changements. D'ailleurs, Filizola *et al* ont étudié ce phénomène par dynamique moléculaire et ont calculé les distances entre les différents ligands et les résidus responsables des interactions. Ils concluent sur la nature flexible du site actif de la cyclooxygénase.[108]

Malgré tout, ces ratios sont peu significatifs de la qualité prédictive d'un modèle et il est nécessaire d'utiliser le E_r et R_s pour mieux caractériser les sept structures cristallographiques et les comparer (Figure 31).

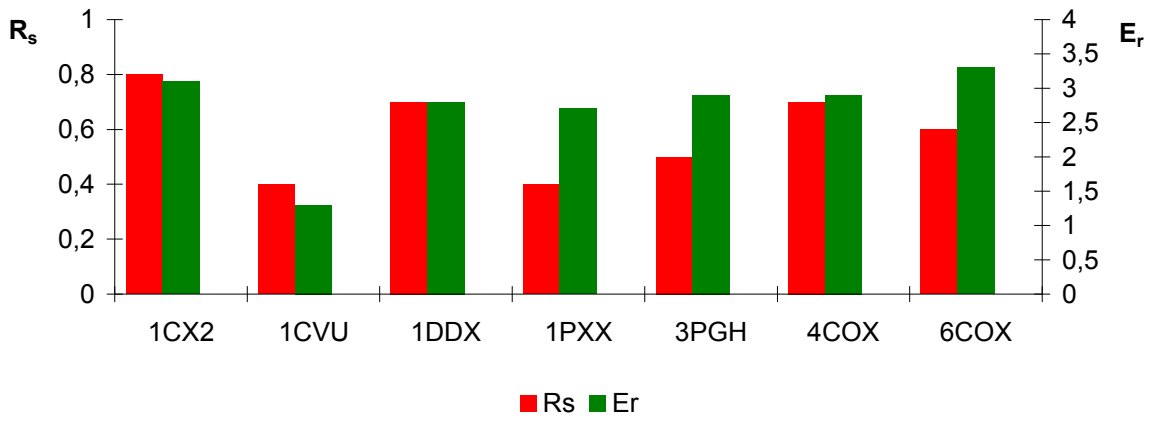


Figure 31. Evaluation des structures cristallographiques selon E_r et R_s

Les ratios E_r et R_s sont majoritairement en faveur de 1CX2. D'autres structures telles que 1DDX, 4COX montrent également de bonnes performances. Dans les cas où le E_r est proche de 3,0 mais avec un R_s faible (de l'ordre de 0,4), cela signifie que le modèle détecte bien une famille de molécules actives dès le début du classement, mais perd toutes les autres molécules actives restantes dans la fin du classement. C'est notamment le cas de 1PXX cristallisée avec un AINS, le diclofénac. Le plus mauvais des cas est le 1CVU. Une des hypothèses d'un résultat aussi médiocre est certainement l'origine de son ligand cristallisé (acide arachidonique) dont la structure aliphatique favorise, lors de la cristallisation, le

[108] Filizola, M. ; Perez, J.-J. ; Palomer, A. ; Mauleón, D. Comparative molecular modeling study of the three-dimensional structures of prostaglandin endoperoxide H2 synthase 1 and 2 (COX-1 and COX-2) *J. Mol. Graph. Model.* **1997**, *15*, 290-300.

rétrécissement du site actif. L'entrée de molécules plus encombrantes telles que les coxibs pose alors problème. Nous avons représenté les courbes d'enrichissement des deux formes cristallisées avec des composés coxibs (Figure 32):

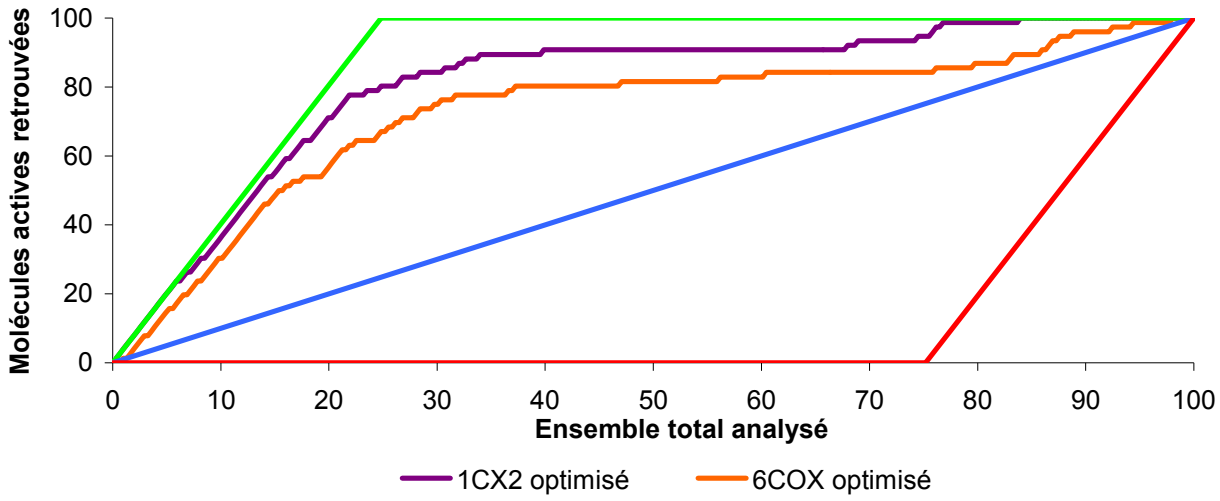


Figure 32. Courbes d'enrichissement des structures cristallographiques 1CX2 et 6COX

Le R_s a été utilisé pour quantifier ces courbes:

$$R_s^{1CX2}=0,82;\ R_s^{6COX}=0,59$$

Ces valeurs sont bien en accord avec la tendance générale des courbes. Le calcul du E_r, dans ces deux modèles, est en désaccord avec ce à quoi nous pourrions nous attendre par observation de la courbe:

$$E_r^{1CX2}=3,1,\ E_r^{6COX}=3,3 \text{ et } E_r^{max}=4,0$$

L'enrichissement que nous calculons depuis le début de cette étude est un enrichissement à 25% de l'*ensemble total* analysé. Par conséquent, en se plaçant à une abscisse de 25%, nous pourrions espérer voir un facteur d'enrichissement de 6COX inférieur à celui de 1CX2. Hors, le calcul des E_r précédent montre que $E_r^{1CX2} < E_r^{6COX}$. Nous sommes dans le cas de figure où la deuxième règle (énoncée dans la partie relatant le E_r) n'est pas respectée. Nous ne pouvons donc pas comparer les E_r entre eux. En effet, le nombre de molécules prédites actives diffère. La forme 1CX2 prédit 80 molécules comme étant actives

contre 47 dans le cas de 6COX. La structure 6COX propose moins de molécules actives dans le premier quart. Ainsi, plus le nombre de molécules prédites actives est faible, moins nous aurons de chance d'introduire de faux positifs. Dans le cas présent, si nous avions choisi notre modèle cristallographique sur l'unique base du E_r, nous aurions obtenu un modèle qui élimine trop de molécules actives de la sélection finale. Dans ce type de comparaison, le R_s a toute son importance par rapport au E_r, dont l'information est ici biaisée. Au cours de nos précédentes comparaisons (comparaisons de fonctions de scoring, comparaisons de structures cristallines originales et optimisées), nous avons parfois favorisé plutôt le E_r au R_s car dans ces situations, les deux règles d'utilisation du E_r étaient respectées.

En résumé, les performances des six structures cristallographiques sont inférieures à celles de 1CX2 optimisé. Cette structure sera utilisée pour nos études ultérieures.

Jusqu'à maintenant, aucune validation du modèle statistique d'AFD n'a été réalisée. En effet, nos calculs ont porté sur l'*ensemble total*. Cet ensemble est très utile pour tester et comparer tous les modèles entre eux, du fait que le nombre de molécules passées dans le modèle est maximal. La validation d'une méthode non supervisée de type consensus n'est pas nécessaire. Seules les courbes d'enrichissement, le E_r et le R_s permettent d'affirmer la validité du classement. Ce n'est pas le cas de l'AFD, qui est une méthode dite supervisée, nécessitant une validation. Nous avons validé l'AFD sur les *ensembles d'entraînement* et de *test* décrits précédemment.

4. Validation du modèle 1CX2 optimisé

L'inconvénient majeur d'une méthode utilisant la diversité pour constituer les deux ensembles est que toute l'information est concentrée dans l'*ensemble d'apprentissage*. Par conséquent, l'information structurale de l'*ensemble de test* est faible. Nous n'avons pas choisi la diversité pour définir la constitution de nos ensembles. Nous avons privilégié un algorithme aléatoire pour la répartition des molécules dans les deux ensembles. Nous avons constitué 10 paires d'*ensembles d'entraînement* et de *test*. Ce protocole permet d'établir une moyenne des erreurs sur chacun des ensembles. Les formules suivantes, dont les termes sont issus de la matrice de confusion, permettent d'évaluer les erreurs «E» présentes dans chacun des ensembles:

$$E_{entraînement} = \frac{FP_{entraînement}}{T_{entraînement}} + \frac{FN_{entraînement}}{T_{entraînement}} \quad , E_{test} = \frac{FP_{test}}{T_{test}} + \frac{FN_{test}}{T_{test}}$$

Équation 9. Erreurs estimées sur chaque ensemble

Les dix jeux ont été évalués par le protocole de docking-scoring décrit ci-dessus à partir de la structure cristallographique 1CX2 (Figure 33).

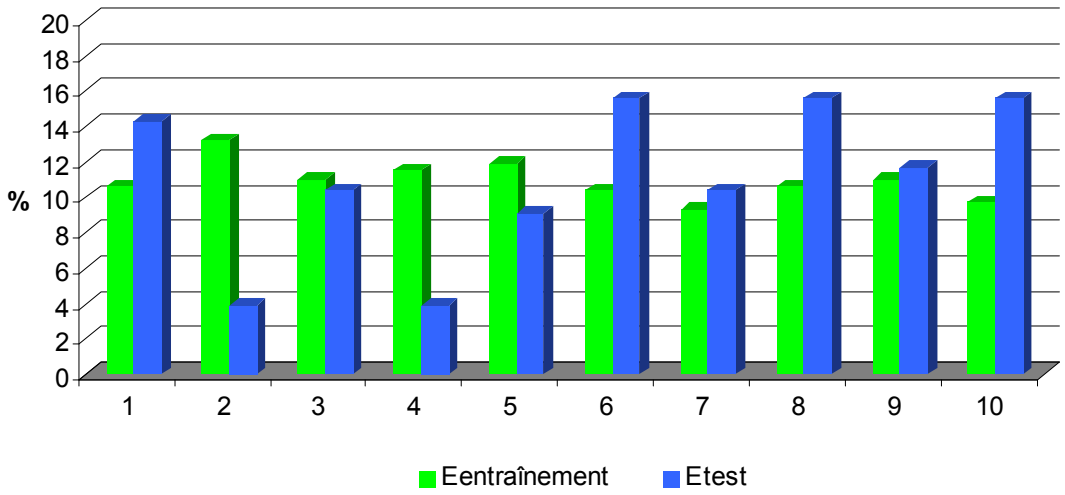


Figure 33. Evaluation du taux d'erreur sur l'*ensemble d'entraînement* et l'*ensemble de test*

Des ensembles choisis au hasard donnent des résultats différents. Ceci montre l'importance de la composition de l'*ensemble d'apprentissage* et de *test*. Les variations extrêmes de l'erreur sur l'*ensemble d'entraînement* varient de 9 à 13%, soit en moyenne une erreur de prédiction sur l'*ensemble d'entraînement* de 11%. L'écart type de cet ensemble est de l'ordre de 1%. Les erreurs de l'*ensemble de test* varient de 4 à 15%, soit en moyenne 11%. L'écart type de l'erreur sur l'*ensemble de test* est de 4%. Une observation intéressante est la répartition des erreurs entre les deux ensembles. En effet, parfois l'erreur est plus importante sur l'*ensemble de l'entraînement* que sur l'*ensemble de test* (modèles 2, 3, 4 et 5) et inversement (modèles 1, 6, 7, 8, 9 et 10). Ces variations sont dues à la répartition des familles de molécules susceptibles de poser problème (dans notre étude, les molécules possédant la fonctionnalité spiro). Le modèle 2 est celui qui génère le moins d'erreurs sur l'*ensemble de test*. Par conséquent, nous l'avons utilisé pour le criblage virtuel. C'est celui pour lequel la validation est la plus satisfaisante. Les individus de chaque ensemble ont été replacés sur l'axe composite issu de l'analyse factorielle discriminante (Figure 34).

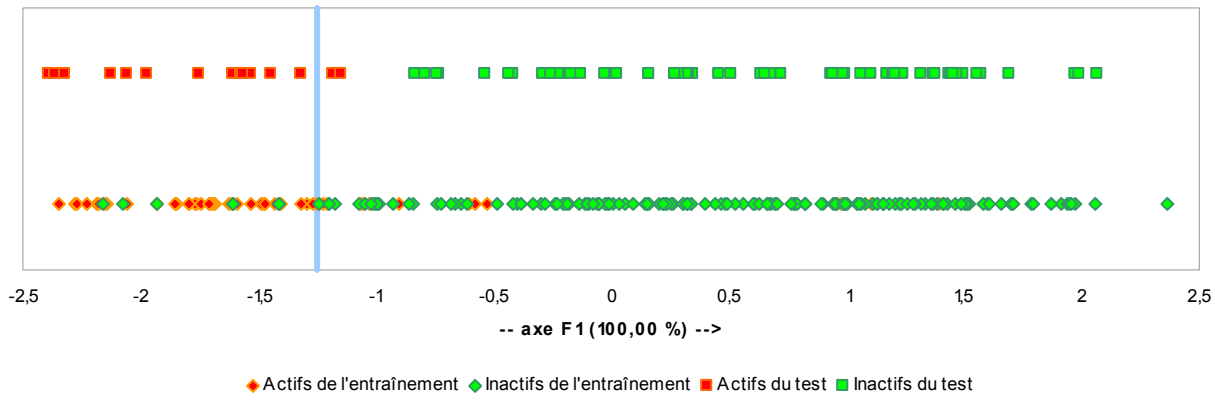


Figure 34. Répartition des molécules des deux ensembles sur l'axe F_1 composite de l'AFD

Les lignes représentent le domaine dans lequel se situent les molécules de l'*ensemble d'entraînement*. On observe la limite de discrimination entre les deux catégories de molécules de cet ensemble entre -1 et -1,5. Par ailleurs, les individus de l'*ensemble de test* sont représentés sous forme de carré. Il apparaît clairement que les catégories (actif et inactif) de l'*ensemble de test* se superposent parfaitement avec les catégories de l'*ensemble d'entraînement*. Par conséquent, le modèle basé sur les lignes (vertes et rouges) permet de ré attribuer les molécules de l'*ensemble de test* aux catégories correspondantes. Ceci valide le modèle.

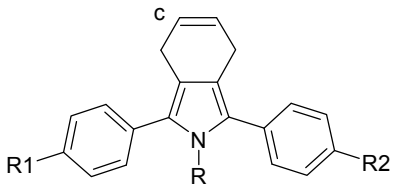
Cette validation est basée sur des composés actifs de la famille des coxibs (possédant tous des groupements sulfonamide et méthylsulfone) mélangés à des *leurres*. La prochaine étape consiste à évaluer le modèle pour des molécules ne possédant plus la fonctionnalité sulfonyle. Ce type de test permet d'évaluer la robustesse du modèle, c'est-à-dire son pouvoir d'extrapolation.

5. Evaluation du modèle vis-à-vis de molécules sans SO_2

Le protocole de criblage virtuel est maintenant optimisé et validé. Mais il reste une interrogation. L'objectif de ce travail est de cribler des produits sans SO_2. Puisque le modèle a été élaboré à partir de composés porteurs de la fonctionnalité sulfonyle, nous avons souhaité savoir si des composés actifs ne présentant pas ce type de fonction réactive ont des chances d'être considérés comme vrais positifs. Pour cela, nous avons extrait de la littérature des composés de la famille des isoindoles, dont l'activité sur COX-2 a été démontrée ($pIC_{50} \geq$

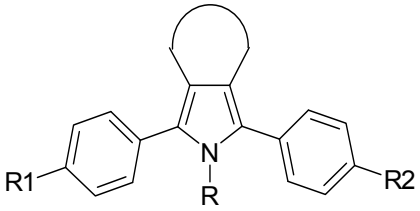
8,0).[109] Le docking avec FlexX de ces composés a déjà été décrit, mettant en évidence une corrélation entre l'activité biologique et le score de la fonction de scoring FlexX-Score (r^2=0,534).[110]

Du fait que ces composés sont issus de tests biologiques différents des molécules de notre *ensemble total*, nous les avons traitées comme des individus étrangers (ou supplémentaires) au modèle (ne rentrant en compte ni dans l'apprentissage ni dans la phase de test). La probabilité (provenant de l'AFD) d'appartenir au groupe des actifs a été évaluée pour chacun de ces dérivés 3-diaryl-4,5,6,7-tetrahydro-2H-isoindole (Tableau 10).



Composés	R	R_1	R_2	pIC_{50} (nM)	Probabilité
1	H	H	H	8.8	**0.6**
2[c]	H	H	H	8.5	**0.6**
3	NH_2	H	H	8.7	0.1
4	H	F	F	8.8	0.5
5	H	Cl	Cl	8.3	**0.8**

[109] Portevin, B.; Tordjman, C.; Pastoureau,P.; Bonnet,J.; De Nanteuil , G. 1,3-diaryl-4,5,6,7-tetrahydro-2H-isoindole derivatives: a new series of potent and selective cox-2 inhibitors in which a sulfonyl group is not a structural requisite *J.Med.Chem* **2000**, *43*, 4582-4593.
[110] Chakraborti, A.-K. ; Thilagavathi, R. Computer-aided design of selective cox-2 inhibitors: molecular docking of structurally diverse cycloxygenase-2 inhibitors using flexx method *Biochempress* **2003**, *2*, 1-2.



Composés	Cycle	R	R_1	R_2	pIC_{50} (nM)[a]	Probabilité[b]
6		H	H	H	8.5	**0.8**
7		H	H	H	9.2	**0.7**
8		NH_2	H	H	8.5	**0.7**
9		H	F	F	8.6	**0.7**
10		H	Cl	Cl	9.0	0.5
11		H	F	F	8.3	0.5
12		H	H	H	8.8	0.1
13		H	F	F	9.2	0

Tableau 10. Molécules de la famille des isoindoles

[a]pIC_{50}:valeurs logarithmiques de la concentration en composé nécessaire pour inhiber 50% des COX-2 du macrophage chez la souris, [b]Probabilités d'appartenance à la classe pharmacologique active (1) et inactive (0). [c] Double liaison en position 6,7. Les vrais positifs sont surlignés en rouge

Les dérivés isoindoles ont été replacés dans un repère à une dimension dont l'abscisse représente la probabilité de chaque composé (Figure 35).

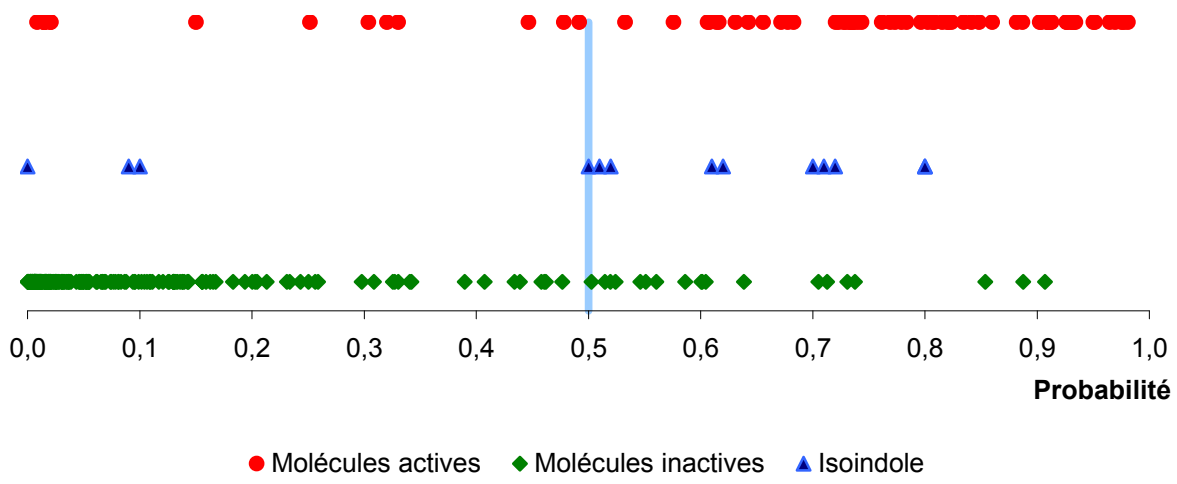


Figure 35. Répartition des dérivés isoindoles au milieu des deux catégories des molécules provenant de l'ensemble total

Le graphique montre que les isoindoles se répartissent majoritairement du côté des molécules actives. Seules 3 molécules sont traitées comme faux négatifs tandis que 3 molécules sont à la limite de la catégorie des actifs et des inactifs. Pour le restant des dérivés, leurs probabilités montrent catégoriquement qu'ils sont traités comme vrais positifs. Une autre représentation consiste à replacer ces 13 individus au sein d'une courbe d'enrichissement afin d'évaluer le pourcentage de l'*ensemble total* analysé, à partir duquel nous pourrions escompter les retrouver dans un processus de criblage virtuel (Figure 36).

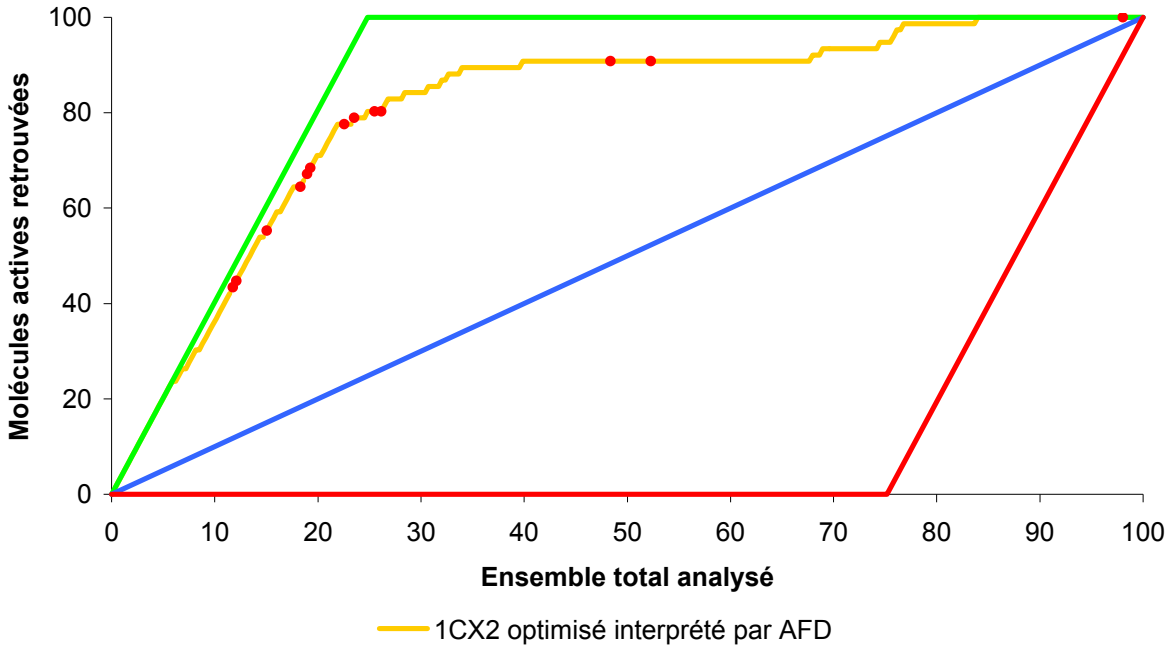


Figure 36. Représentation des dérivés isoindoles sur la courbe d'enrichissement du modèle 1CX2 optimisé

Au total, 10 composés sur 13 sortent dans les 30 premiers pour-cent de l'*ensemble total* analysé. Seuls trois composés sont détectés tardivement dans la courbe d'enrichissement. Les deux représentations graphiques précédentes suggèrent que notre modèle est capable de détecter des composés différents de ceux ayant servi à l'apprentissage et au test. C'est une qualité considérable des méthodes basées sur la structure tridimensionnelle de la protéine.

En résumé, nous avons construit un modèle prédictif de l'affinité de ligands vis-à-vis des cyclooxygénases de type 2. L'utilisation de l'analyse factorielle discriminante nous a permis d'une part de considérer toutes les fonctions de scoring, même la moins performante (PMF) qui apporte sa part d'information. Cette stratégie nous a également donné l'opportunité d'augmenter le pouvoir prédictif vis-à-vis d'une même classe pharmacologique mais également d'extrapoler les prédictions de composés différents de ceux ayant servi à la construction et à la validation du modèle. Le protocole est donc bien établi. Toutefois, il reste un point fondamental qui conditionne son application au criblage virtuel: le temps de calcul.

6. Temps de calcul

En moyenne, la phase de docking est le facteur limitant du processus. Cette opération peut consommer en moyenne 30 secondes par molécule (sur un nœud AMD 3800+). En se

basant sur une chimiothèque de un million de composés, un nœud mettrait 350 jours à terminer le processus. Du fait des besoins du laboratoire, ce temps de calcul doit être réduit. D'une part, la diminution du nombre de molécules en entrée est capable d'avoir un impact sur la réduction du temps de calcul. D'autre part, la modification de paramètres de docking (en particulier l'incrément donné à l'angle lors de l'ajout d'un fragment moléculaire) peut également avoir une forte incidence sur le temps de calcul. Par contre, ces modifications ont comme conséquences de négliger certaines interactions favorables à cause d'une incrémentation trop grossière. Nous avons donc opté pour la première solution qui est de filtrer les composés à l'entrée du protocole. L'ajout d'un filtre pharmacophorique en amont du docking est une solution pour éliminer directement les composés stériquement et électroniquement non favorables.

7. Modèle pharmacophorique

a) FlexX-Pharm

(1) Généralités

FlexX-Pharm est une extension de l'outil de docking FlexX, capable d'incorporer l'information provenant de l'association de ligands affins avec un récepteur donné. Les contraintes sont déterminées en favorisant dans les solutions finales les interactions «clé» ou même des volumes d'inclusion, dans lesquels il est essentiel de retrouver les poses. Cet outil permet de guider la construction de la molécule dans le site actif, éliminant rapidement les conformations incapables de respecter le modèle. Des améliorations nettes de RMSD (calculé entre le ligand co-cristallisé et la première pose de ce même ligand lors d'un processus de re-docking), lors du passage de FlexX à FlexX-Pharm, ont été observées. Ce module permet donc une réduction du RMSD ainsi que du temps de calcul.

(2) Principe

Dans FlexX-Pharm, deux types de contraintes peuvent être définies dans le site actif de la protéine: des contraintes interactionnelles et des contraintes spatiales. Pour le premier type d'interaction, il est possible de spécifier la surface d'interaction à satisfaire dans le site actif. Pour le second type d'interaction, il s'agit plus spécialement de sphères d'inclusion.

(a) Les contraintes interactionnelles

Ce type de contraintes est attribué à un groupement fonctionnel. Il est associé à un type d'interaction particulier. FlexX-Pharm, lors de l'ajout d'un fragment, s'assure que la zone d'interaction définie dans le site actif est en face d'une fonctionnalité favorable du ligand. Les interactions peuvent être de type donneur/accepteur d'hydrogène, accepteur de d'atome métallique (toutes ces interactions sont directionnelles). Les contacts hydrophobes directionnels (entre cycles et chaînes aliphatiques) sont également considérés. Enfin, des interactions moins directionnelles de type hydrophobe sont traitées (entre groupement méthyle, entre atomes de soufre).

(b) Les contraintes spatiales

Ce type de contraintes est utilisé pour forcer le ligand à se positionner dans des positions précises du site actif. Elles sont également définies par un type atomique, imposant un type d'atome du ligand dans ces sphères.

L'intérêt d'ajouter un pharmacophore est majeur car il permet de réduire l'espace conformationnel de la molécule et de le focaliser sur des critères précis. L'inconvénient de cette méthodologie est qu'elle impose un biais dans la recherche conformationnelle, certainement moins exhaustive. L'avantage majeur est le gain de temps qu'elle procure en éliminant des composés, qui de toutes manières, n'auraient eu aucune chance dans le site actif. L'autre point positif est la qualité de l'alignement des conformations sur le ligand co-cristallisé. Cet outil est adapté au criblage virtuel.

b) Application à la cyclooxygénase de type 2

La forme générale des coxibs nous a conduits à élaborer des contraintes de type spatial uniquement. Les 20 molécules les plus diverses de l'*ensemble total* ont été alignées sur le ligand co-cristallisé SC-558.

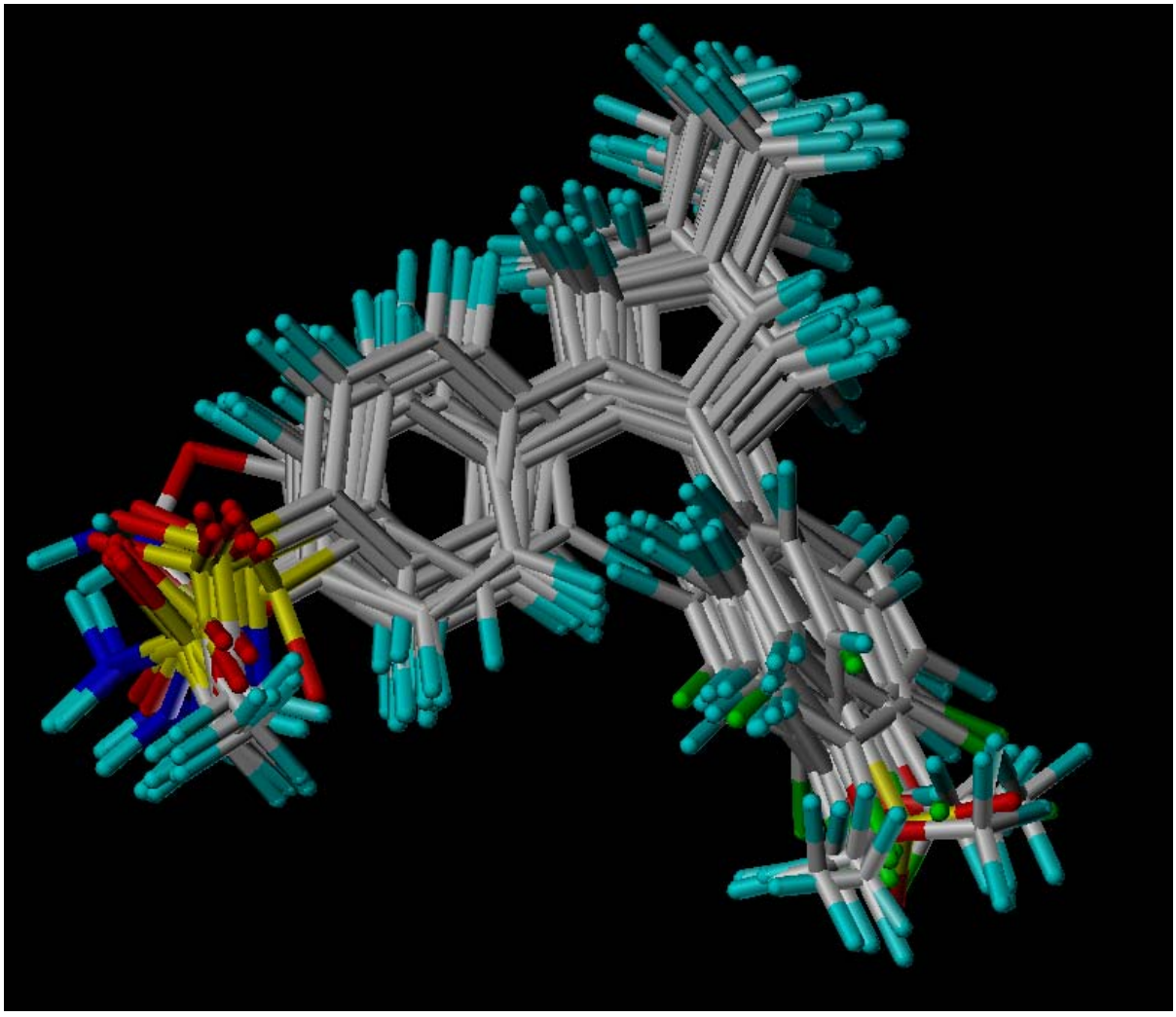


Figure 37. Alignement des 20 molécules coxib les plus diverses

De cet alignement ont été déduites trois sphères. Les distances inter-sphères ainsi que le diamètre des sphères ont été optimisés afin de filtrer au mieux les composés inactifs, tout en minimisant le nombre de composés actifs éliminés par le modèle pharmacophorique. Nous avons vu précédemment que la force de chaque contrainte peut être nuancée en «essentielle» ou en «optionnelle». La sphère superposée au cycle porteur de la fonctionnalité sulfonyle a été définie comme essentielle (notée «E» sur le schéma). Cette approche est bien sûr non pénalisante pour les composés ne possédant pas de fonctionnalité SO_2. Cette nuance a été attribuée du fait que l'on requiert des molécules qui pénètrent dans la poche adjacente contenant l'His90 et l'Arg513. Les deux autres sphères ont été définies comme «optionnelles» du fait que l'on attend toutes les molécules dans une des deux sphères restantes ou même dans les deux simultanément (notées «O»).

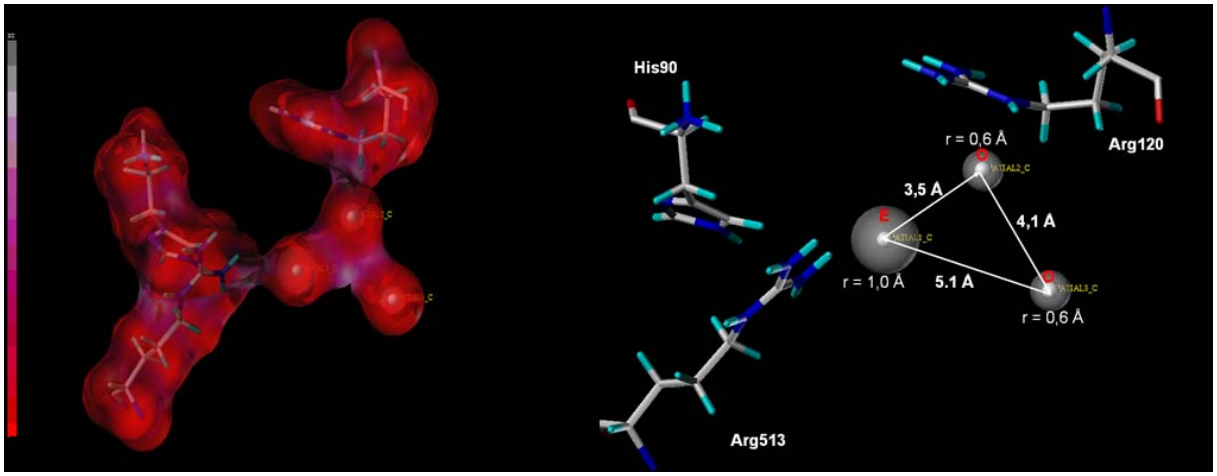


Figure 38. Modèle pharmacophorique

Nous avons évalué le pharmacophore par le protocole de docking-scoring élaboré précédemment en utilisant la structure cristallographique 1CX2 optimisée. Nous l'avons comparé aux résultats obtenus sans modèle pharmacophorique:

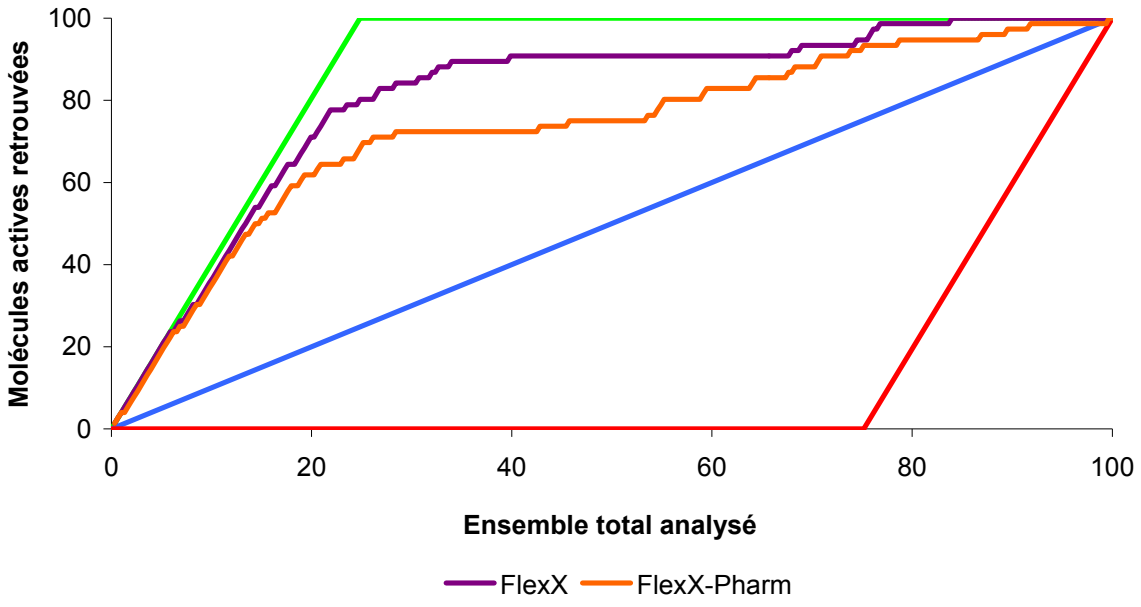


Figure 39. Courbes d'enrichissement obtenues avec FlexX seul et FlexX-Pharm

La courbe d'enrichissement de FlexX-Pharm est très similaire à celle de FlexX dans les 15 premiers pour-cent de l'*ensemble total* analysé. Par contre, la courbe de FlexX-Pharm se dégrade par la suite. Nous avons calculé les R_s pour estimer ces différences de comportement. Afin de rendre les courbes d'enrichissement comparables, les 27% de faux

négatifs (molécules ayant été éliminées directement par le pharmacophore) ont été ajoutés dans les dernières positions de l'*ensemble total* analysé.

$$R_s^{FlexX}=0,82;\ R_s^{FlexX\text{-}Pharm}=0,53$$

En présence de FlexX-Pharm, le R_s reflète la faible détection des molécules actives à partir de 15% de l'*ensemble total.* Au total, 68% des 230 *leurres* de l'*ensemble total* sont supprimées par FlexX-Pharm (traitées comme vrais négatifs) contre 27% des 76 molécules actives (traitées comme faux négatifs). C'est la raison pour laquelle on note un R_s aussi faible. Le facteur d'enrichissement a également été évalué:

$$E_r^{FlexX}=3,1;\ E_r^{FlexX\text{-}Pharm}=3,0;\ E_r^{Max}=4,0$$

Les résultats du $E_r^{FlexX\text{-}Pharm}$ sont prometteurs pour le criblage virtuel. Le gain en vitesse de calcul est considérable car on peut s'attendre à diminuer de plus de 25% le temps de calcul. Nous avons fait le choix d'utiliser FlexX-Pharm pour améliorer la vitesse de calcul au dépend des 27% de molécules actives qui seront éliminées par le pharmacophore. Nous avons compensé cette perte en sélectionnant les 10 000 meilleures molécules, par la suite re-dockée à l'aide de FlexX. Sans problème du temps de calcul limité, nous n'aurions pas utilisé FlexX-Pharm mais dans notre approche de criblage virtuel, ce module est indispensable.

8. Criblage virtuel

a) Constitution des chimiothèques dédiées au criblage virtuel

(1) Chimiothèque commerciale

Nous aurions pu tester une chimiothèque globale mais pour des raisons de valorisation de notre chimiothèque locale de l'ICOA ainsi que la chimiothèque nationale, nous avons constitué trois sous chimiothèques de taille différente. Une combinaison de 32 fournisseurs de produits chimiques nous a permis de regrouper au total 3,8 millions de molécules. Environ 2,6 millions ont été identifiés comme uniques, environ 1,9 millions respectent des critères «drug-

like» et 900 000 sont des composés «lead-like».[111] Nous avons utilisé des critères durcis «drug-like»:

- 3≤LogP≤8

- Pas de groupement NO_2

- Taille maximum des cycles: 7

- Nombre maximum de liaisons sujettes à la rotation: 11

- 100≤Poids moléculaire≤600

- Pas de chaîne perfluorée

- Pas de frequent hitter (composés qui montrent de l'affinité pour bon nombre de cibles thérapeutiques)

La sélection des composés a été faite avec «Screening assistant»[112] qui a été développé en interne dans notre laboratoire et distribué en open source.

(2) Chimiothèque nationale et ICOA

Nous avons souhaité valoriser les molécules de la chimiothèque nationale. Cette librairie de produits est issue d'un groupement de services (CNRS-Université). Ce groupement a pour mission de fédérer les collections de produits de synthèse et d'origine naturelle provenant des laboratoires publics français et d'en promouvoir la valorisation scientifique et industrielle.

b) Protocole du criblage virtuel

La première étape du criblage virtuelle est l'utilisation des filtres physico-chimiques décrits précédemment pour réduire la taille de la chimiothèque de 3,8 millions de composés. FlexX-Pharm est ensuite appliqué à l'ensemble pré-filtré. Pour chaque chimiothèque, les

[111] Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers *Mol. Divers.* **2006**, *10*, 389-403.
[112] http://www.univ-orleans.fr/icoa/screeningassistant/

produits ayant le meilleur score et dont le mode de fixation dans le site actif seront retenus pour les tests biologiques.

c) Choix des produits

La motivation première est de choisir les meilleures molécules de chacune des trois chimiothèques. L'objectif, en plus d'isoler des produits commerciaux potentiellement inhibiteurs des la COX-2, est d'apporter de la valeur ajoutée à des molécules issues de notre laboratoire mais également celles du domaine public.

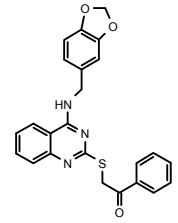
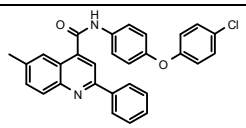
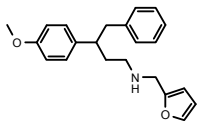
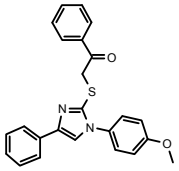
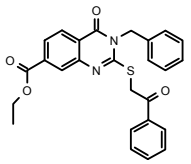
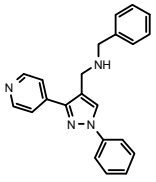
Le deuxième argument est l'approvisionnement aisé de ces composés *via* les fournisseurs de produits chimiques. Toutefois, une telle sélection pénalise l'ensemble final choisi. En effet, les meilleurs scores de la chimiothèque ICOA de 3000 molécules sont peut être dans les scores moyens de la chimiothèque nationale. Au risque de pénaliser notre sélection finale, nous avons quand même souhaité valoriser ces quelques molécules issues du domaine public. La sélection des composés (au sein de chacune des trois chimiothèques prises individuellement) s'est faite en considérant les composés ayant le meilleur score. La fixation dans le site actif est également un critère sur lequel nous avons choisi ces composés. Plus précisément, chaque produit occupe la poche adjacente, dans laquelle se trouvent His90 et Arg513, décrite comme responsable de la sélectivité. Chacun d'entre eux occupe la cavité qui contient l'Arg120. Une fois isolés, les composés ont été testés biologiquement.

d) Tests biologiques

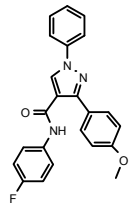
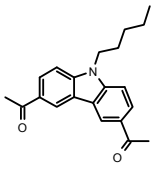
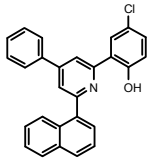
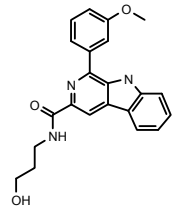
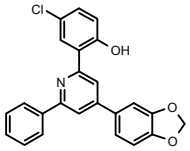
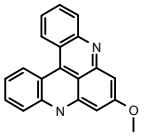
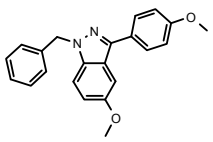
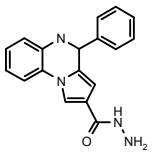
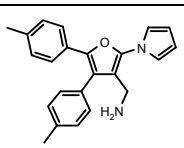
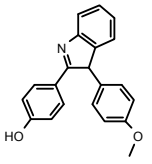
Notre budget étant limité, nous avons choisi de sélectionner seulement 20 molécules qui ont été isolées selon le protocole précédemment décrit. Cet ensemble final est constitué de 12 meilleures molécules de la chimiothèque commerciale, 4 meilleures de la chimiothèque nationale et enfin 4 meilleures de la chimiothèque de l'ICOA.

La totalité de ces molécules ne possèdent pas le groupement sulfonyle, fonctionnalité spécifique aux coxibs. Par ailleurs, l'algorithme de diversité basé sur les fingerprints structurales MACCS nous a permis d'obtenir une variété de familles pour cette sélection de 20 molécules.

Ces molécules ont été testées en collaboration avec la société GreenPharma.[113] Ce test consiste à mesurer le pouvoir inhibiteur des 20 molécules sur l'enzyme recombinante humaine à une concentration de 1µM (Tableau 11). Il s'agit d'un test EIA (enzyme immunoassay) qui détermine quantitativement la quantité de prostaglandines F, E, D et de thromboxane produite lors de la réaction de cyclooxygénation. Un anti-corps spécifique se fixant sur tous les produits de la réaction est utilisé. Ce test est approprié au criblage virtuel et à la recherche d'inhibiteurs. Il mène à un faible nombre de faux positifs par rapport aux autres méthodes.

Molécule	Structure	$\%_{inhibition}$	Chimiothèque
1		17,5	Commerciale
2		18,5	Commerciale
3		37,5	Commerciale
4		36,7	Commerciale
5		42,2	Commerciale
6		16,2	Commerciale

[113] http://www.greenpharma.com/

7		13,2	Commerciale
8		12,6	Commerciale
9		36,0	Commerciale
10		0,0	Commerciale
11		0,0	Commerciale
12		24,9	Nationale
13		32,0	Nationale
14		45,5	Nationale
15		3,7	Nationale
16		24,0	ICOA

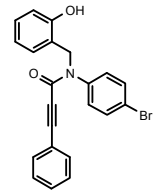
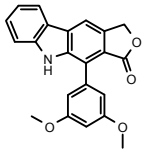
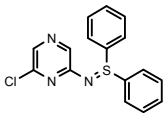
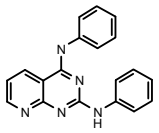
17		23,7	ICOA
18		0,0	ICOA
19		0,0	ICOA
20		0,0	ICOA

Tableau 11. Structures des composés actifs sur COX-2

Sur les 20 composés testés, deux ont un pourcentage d'inhibition proche de 50% d'inhibition à la concentration utilisée (1 µM) (un de la chimiothèque commerciale et un de la chimiothèque nationale). Une inhibition comprise entre 30 et 40% à une concentration micro molaire est obtenue pour 4 molécules (3 molécules de la chimiothèque commerciale et une pour la chimiothèque nationale). Enfin, 7 molécules ne présentent aucun pouvoir inhibiteur.

H. Conclusion et perspectives

Nous avons montré l'enjeu d'étudier l'enzyme cyclooxygénase de type 2 dans le traitement de l'inflammation. Il est plus que jamais fondamental d'innover, en terme de structures organiques, dans ce créneau aujourd'hui très délicat que sont les inhibiteurs de la cyclooxygénase. Dans ce premier chapitre, nous avons utilisé l'information structurale disponible dans la PDB à propos de la cyclooxygénase de type 2. Aucune étude de sélectivité de l'isoforme 2 par rapport à l'isoforme 1 n'est traitée ici. Nous avons occulté cet aspect pharmacologique du fait que la trop grande spécificité de certains composés (tel que le Vioxx) conduit à d'importants effets adverses, de type cardiovasculaire. Par ailleurs, l'information provenant de molécules organiques (agents coxibs) à fort potentiel inhibiteur a également été exploitée dans cette étude. Malgré tout, notre volonté a été d'isoler des composés dont l'affinité est inférieure à celle des coxibs mais sans le groupement fonctionnel sulfonyle très caractéristique de cette famille. Nous avons prouvé, qu'au travers de techniques basées sur la structure des protéines, il est possible d'identifier des composés inhibiteurs différents de ceux ayant servi à élaborer le modèle de docking. De ce fait, les composés ayant été sélectionnés possèdent des structures éloignées des molécules inhibitrices décrites dans la littérature.

L'approche selon laquelle l'utilisateur choisit ses molécules selon une fonction de scoring est désuète. Elle conduit généralement à un échec de prédiction. Notre stratégie de re-scoring assistée par un traitement statistique montre ici sa capacité à proposer des composés originaux tout en minimisant le nombre de composés faux positifs. L'utilisation des méthodes discriminantes comme l'AFD dans notre étude doit être un point de départ à de nouvelles approches de consensus de fonctions de scoring. Malgré tout efficace, les principales méthodologies de consensus décrites dans nos travaux ont une puissance limitée au choix des meilleures fonctions de scoring.

Les résultats obtenus à partir de l'*ensemble d'entraînement* et de test valident notre modèle statistique. Par ailleurs, nous avons également été capables de tester des composés différents de ceux des deux ensembles ayant permis de valider le modèle. Ces molécules ont majoritairement été traitées comme vraies positives.

Le protocole de docking-scoring a également été couplé à un modèle pharmacophorique dont l'efficacité, à des fins de criblage virtuel, est évidente. Ce modèle nous a permis de filtrer, avant l'étape de docking, tous les composés ne répondant pas aux critères d'occupation du site actif.

Les résultats biologiques nous permettent de valider ce protocole impliquant différentes stratégies et sont même encourageants, puisque tous les composés présentant une activité pour COX-2 et n'avaient, jusque là, pas été décrits dans la littérature.

Toutefois, nous avons mis en évidence que d'une forme cristallographique de COX-2 à l'autre, les résultats varient. Ceci prouve que le site actif est capable d'adapter partiellement sa géométrie et son volume. Par conséquent, il serait peut-être raisonnable de considérer cette information primordiale qui met en avant une certaine flexibilité locale du site de fixation. Nous aborderons ce problème sur une autre cible dans le prochain chapitre.

CHAPITRE 2.

Elaboration d'un modèle prédictif des récepteurs PPARγ. Implication de la flexibilité du site actif

I. Introduction

1. Le diabète

a) Généralités

Le diabète est une pathologie dont les origines sont multiples. Des facteurs environnementaux (habitudes alimentaires par exemple) ou encore des prédispositions génétiques peuvent être les initiateurs de la maladie. L'insulinorésistance liée à l'obésité et à la sédentarité est la première étape de la maladie. Pour contrer cette insulinorésistance, le pancréas, et plus particulièrement les cellules β-pancréatiques de Langerhans, sécrète plus d'insuline. Toutefois, cette compensation n'est que provisoire et on observe vite une augmentation de la glycémie en période post-prandiale. Le diabète est ainsi suspecté lorsque les taux de glucose s'élèvent dans le sang. Plusieurs critères sont actuellement utilisés pour diagnostiquer le diabète:

1° Une mesure de la glycémie à jeun dépassant les 1,26g/L.

2° Une mesure de glycémie prise au hasard supérieure à 2,00g/L accompagnée de symptômes de type polyurie, polydipsie, polyphagie.

3° Une glycémie supérieure à 2,00g/L deux heures après surcharge orale de glucose. Ce test d'hyperglycémie provoquée par voie orale n'est pas recommandé en pratique clinique.

Le diabète peut avoir une évolution lente sans présenter aucun symptôme. Ceci en fait une pathologie sournoise, silencieuse qui, une fois déclarée, expose à des morbidités sévères (25% des infarctus du myocarde, 10% des accidents cardiovasculaires cérébraux, une majeure partie des cécités et 80% des amputations majeures sont des diabétiques). Les différentes étapes précédant la déclaration de la maladie se passent souvent sans manifestation (Figure 40).

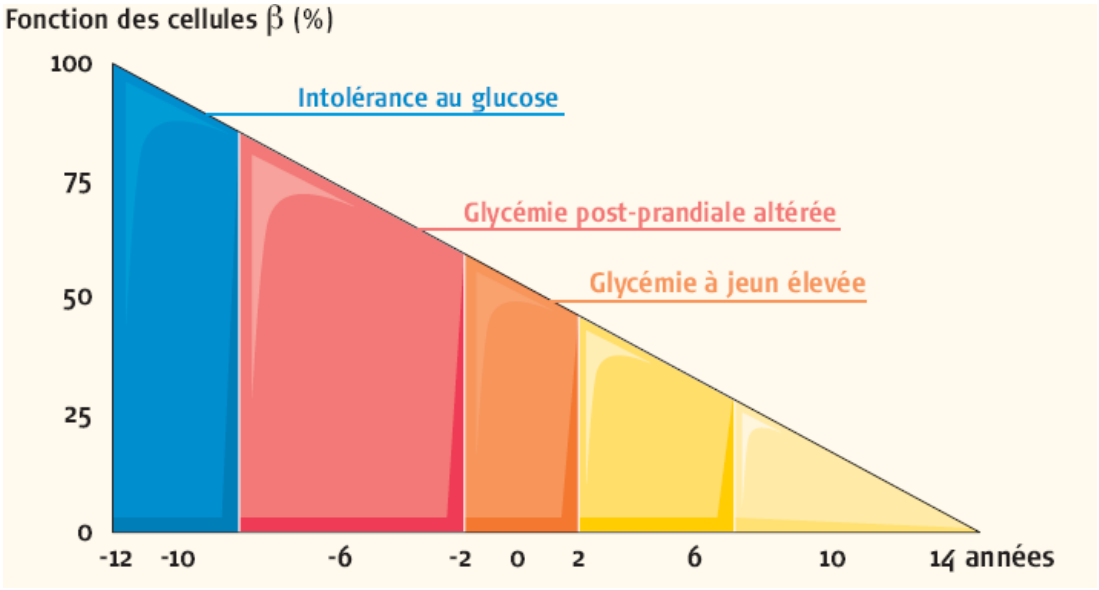


Figure 40. Etapes d'évolution du diabète

Plusieurs types de diabètes peuvent aujourd'hui être décrits:

-Le diabète primitif qui représente 90 à 95% des diabètes dont le diabète de type I et II

-Le diabète gestationnel qui est un cas de diabète transitoire lors de la grossesse.

-Les diabètes secondaires pouvant être causés par une altération du pancréas ou du foie, une sécrétion d'hormones antagonistes de l'insuline ou même des lésions au sein des centres glycorégulateurs.

-Le diabète nutritionnel provenant soit d'une carence protéique ou d'une pancréatite calcifiante.

(1) Le diabète de type I ou insulino-dépendant

Le diabète de type 1 est dû à une disparition quasiment complète de la sécrétion d'insuline par les îlots β de Langerhans du pancréas par un mécanisme auto-immun. Cette maladie touche généralement des sujets âgés de moins de 35 ans. Le traitement du diabète insulinodépendant se fait par insulinothérapie (administration sous-cutanée d'insuline car étant inactive par voie orale). Toutefois, des effets secondaires sont observés tels que des situations d'hypoglycémie, d'hypokaliémie, de lipodystrophie ainsi que des problèmes immunologiques (supprimés avec les insulines humaines). De nouvelles thérapies ont vu le jour, telles que des greffes de pancréas (malgré son taux de réussite très faible), d'îlots encapsulés (avec quelques réussites) et enfin des traitements immuno-suppresseurs.

(2) Le diabète de type II ou non-insulino-dépendant

Ce type de diabète est le cas le plus fréquent. Il se révèle par une diminution progressive de la sécrétion d'insuline et une résistance à l'insuline. Il représente à lui seul 90 à 95% de tous les cas de diabète. L'obésité ainsi que des déséquilibres alimentaires sont généralement la cause du déclenchement, à partir de l'âge de 40 ans, de la pathologie. Une particularité de ce type de diabète est sa forte augmentation au sein des populations des pays industrialisés. En effet, la sédentarisation ainsi qu'une alimentation parfois trop riche sont des éléments favorisant son importante croissance. La prévalence, tout type de diabète confondu, au niveau mondial en 2000 est de 150 millions de personnes. Une augmentation allant jusqu'à 220 millions de personnes atteintes est attendue pour 2010. En France, on estime à près de 2 millions de personnes touchées par cette pathologie.[114] Le diabète de type 2 se développe en moyenne après l'âge de 45 ans. On note également une augmentation du diabète de type 2 chez l'enfant. En résumé, le diabète est une pathologie métabolique qui regroupe disglycémie, obésité, dyslipidémie, élévation de la pression sanguine et disfonctionnements de l'endothélium vasculaire. Les traitements actuels sont regroupés dans deux grandes classes: les sulfonylurées, ayant des propriétés hypoglycémiantes et les biguanides dont l'action est normoglycémiante.

b) La glycorégulation

La régulation des taux de glucose plasmatique est régie par deux hormones:

L'insuline (découverte en 1922 par Banting et Mc Leod, hormone polypeptidique synthétisée par les cellules β de Langerhans du tissu endocrine du pancréas): elle est libérée en cas d'augmentation du taux de glucose dans la circulation sanguine. Ses actions sont multiples:

-Elle accroît la pénétration du glucose au sein de la cellule ;

-Elle favorise la formation de glycogène ;

-Elle accélère le stockage des glucides sous forme de lipides (lipogénèse).

Le glucagon (synthétisé par les cellules α de Langerhans): il est sécrété lors d'une hypoglycémie. Une fois libéré, il va agir en diminuant la quantité de glucose plasmatique. C'est le phénomène de la glycogénolyse. Cette hormone est capable d'agir contre les propriétés hypoglycémiantes de l'insuline (Figure 41).

[114] Skyler, J.S. Diabetes Mellitus: Pathogenesis and Treatment Strategies *J. Med. Chem.* **2004**, *47*, 4113-4117.

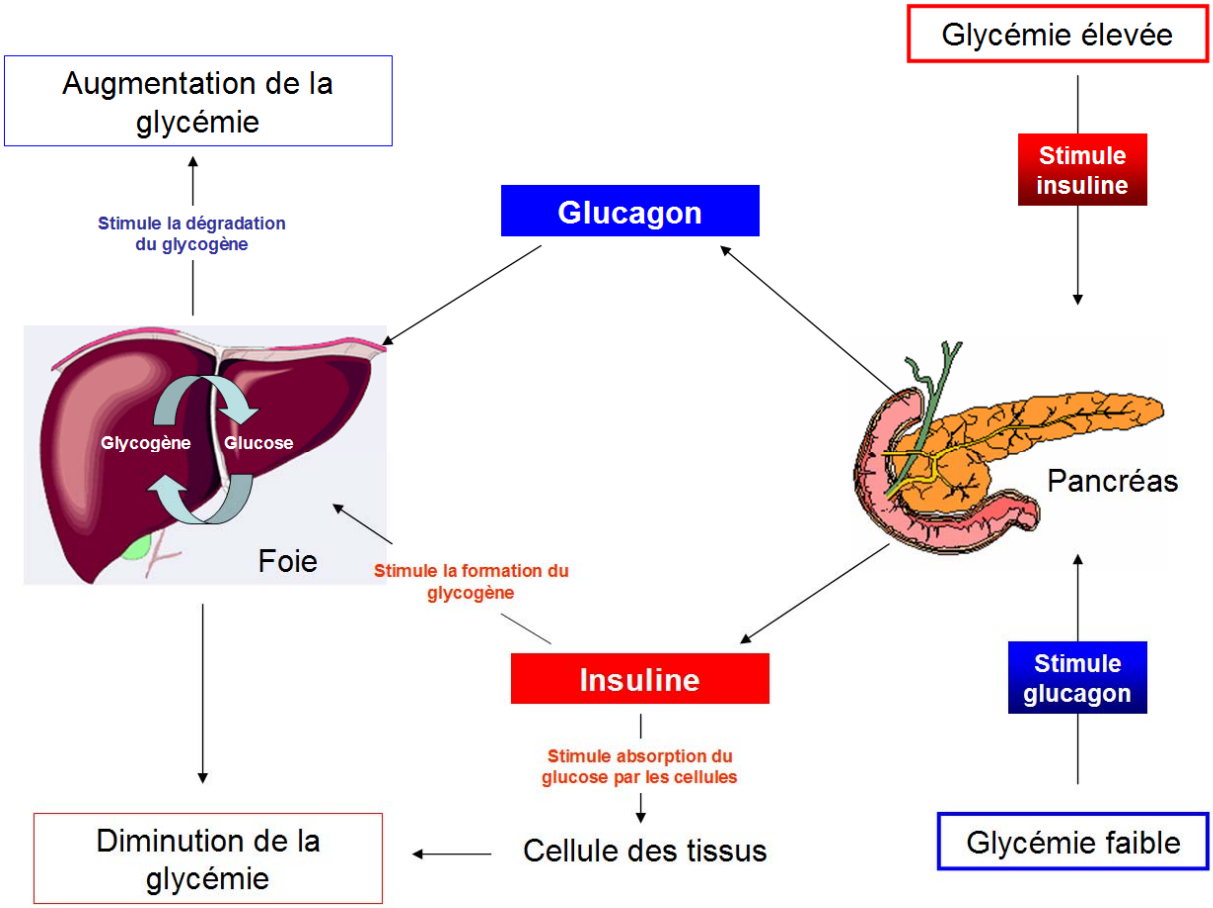


Figure 41. Processus de la glycorégulation

Les effets biologiques de l'insuline sont multiples. En effet, le métabolisme de trois catégories de macromolécules fluctue selon les besoins de l'organisme: les glucides, les lipides et les protéines.

c) Les sucres

Au niveau du foie, l'insuline diminue la glycogénolyse (diminution de production de glucose), augmente la glycogénèse (augmentation de la synthèse de glycogène) et diminue la gluconéogénèse (synthèse de sucre à partir d'acides aminés).

Au niveau du muscle, l'insuline favorise le transport membranaire du glucose ainsi que la conversion du glucose en glycogène. Elle favorise également les voies métaboliques d'Embden Meyerhof et du cycle de Krebs.

d) Les lipides

L’insuline a un effet antilipolytique, c'est-à-dire qu’elle diminue la libération des acides gras libres et du glycérol. Par ailleurs, elle stimule l’activité de la lipoprotéine lipase par augmentation des triglycérides. Enfin, elle favorise la synthèse des triglycérides.

e) Les protéines

L’insuline exerce une action anti-catabolique au niveau du muscle et du foie. Elle augmente la synthèse protéique et diminue le catabolisme protéique (diminution de la production d’urée). Enfin, la gluconéogénèse à partir d’acides aminés est ralentie.

f) Conséquences et complications du diabète

Les conséquences de cette maladie à long terme sont multiples mais en particulier on observe l'altération des fonctions rénale, nerveuse, circulatoire et cardiaque. On observe deux catégories de complications: aiguës et dégénératives.

(1) Les complications aiguës

Les complications aiguës sont les mieux prises en compte avec les insulinothérapies depuis l’identification de l’insuline. Ces défaillances du métabolisme glucidique peuvent conduire à une lipolyse accrue de l’organisme (consommation des graisses pour fournir de l’énergie) avec production de corps cétoniques et de dérivés acides. On parle alors d’acidocétose diabétique. Ces complications peuvent également être d’ordre nerveux avec une déshydratation des cellules du système nerveux. Les conséquences au niveau rénal sont également non négligeables puisqu’une insuffisance rénale peut aller jusqu’à causer un coma hyperosmolaire.

(2) Les complications mécaniques

Les complications mécaniques peuvent également avoir lieu telles que des angiopathies et des neuropathies.

L'altération des parois des artères est appelée macroangiopathie. Environ 50% des diabétiques de type II présentent déjà des micro et macroangiopathies au moment du diagnostic. L'augmentation de la perméabilité vasculaire et de la pression de perfusion conduisent à des situations ischémiques. Ces microangiopathies peuvent entraîner des pathologies telles que des rétinopathies. Elles sont la première cause de cécité entre 30 et 70 ans. De même, des atteintes rénales (néphropathies) vont aboutir à une insuffisance rénale chronique et à une protéinurie. Elles sont la première cause d'hémodialyse. Enfin, des lésions nerveuses auront pour conséquence la diminution de sensibilité des membres inférieurs (neuropathie).

2. Traitements disponibles

Les quatre familles d'antidiabétiques sont les sulfonylurées, les glinides, les biguanides et les thiazolidinediones.

a) Les sulfonylurées

Les sulfonylurées hypoglycémiantes sont des dérivés des sulfamides antibactériens (Figure 42).

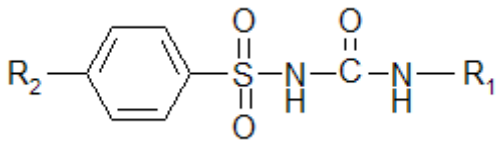


Figure 42. Structure générale de sulfonylurée anti-diabétique.

La cible pharmacologique de cette famille de composés se situe à la surface des cellules β de Langerhans. Ces agents sont impliqués dans la fermeture de canaux potassiques ATP dépendants qui sont capables d'induire une dépolarisation au niveau des cellules de Langerhans. L'augmentation de la concentration intracellulaire de calcium induit la sécrétion d'insuline. Elles ont également une action post-pancréatique. Plus particulièrement, les sulfonylurées potentialisent l'action périphérique de l'insuline. Malgré tout, ces composés présentent des risques d'hypoglycémie et un phénomène d'échappement caractérisé par

l'épuisement de l'activité au delà d'une dizaine d'années. Une incompatibilité a également été observée avec les anti-inflammatoires non stéroïdiens (évoqués au chapitre précédent).

b) Les glinides

Les glinides sont des composés dont le mode d'action est très similaire à celui des sulfonylurées. Ils interviennent auprès des canaux potassiques ATP dépendant (Figure 43).

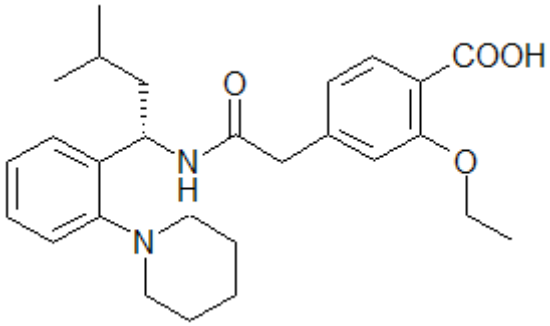


Figure 43. Structure de Repaglinide (Novonorm)

Des agents autres que ceux de la famille des sulfonylurées ont été développés et en particulier les incrétines qui ont pour but de stimuler l'insulinosécrétion pancréatique et l'uptake musculaire du glucose.

c) Les biguanides

La troisième famille de composés anti-diabétiques est appelée biguanide. Ces dérivés de la guanidine ne sont pas hypoglycémiants mais normoglycémiants. Plus précisément, ils diminuent la gluconéogénèse hépatique et améliorent l'utilisation périphérique du glucose (augmentation des transporteurs GLUT-4). Ils réduisent également l'absorption intestinale de glucose. Le médicament phare est la metformine (Glucophage, figure 44).

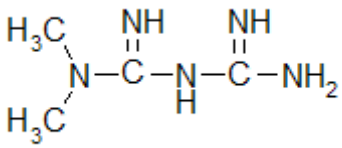


Figure 44. Structure de la metformine (Glucophage)

La metformine est capable, *via* l'activation de l'AMP-activated protein Kinase (AMPK), d'augmenter l'uptake musculaire de glucose en cas d'exercice physique, d'inhiber la production hépatique de glucose, de stimuler l'oxydation des acides gras et enfin de diminuer l'activité de l'acétyl-CoA Carboxylase (ACC). Toutefois, des effets indésirables tels que des troubles digestifs ou une acidose lactique peuvent parfois forcer à modifier le traitement.

d) Les thiazolidinediones

Cette famille de composés normalise la tolérance au glucose sans agir sur une libération excessive d'insuline. Les thiazolidinediones sont des activateurs des PPARγ (Peroxysome Proliferation Activated Receptors) (Figure 45). L'activation des PPARγ conduit à une augmentation des transporteurs de glucose GLUT-4, une diminution de l'expression des TNFα et une augmentation en lipoprotéine lipase au niveau adipocytaire. Tout ceci contribue à accélérer l'uptake des triglycérides. C'est donc un atout majeur que pourraient posséder ces molécules puisqu'elles éviteraient la perte inéluctable de la fonction insulinosecrétrice propre au pancréas.

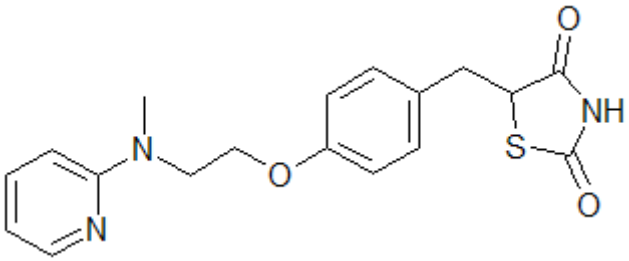


Figure 45. Structure de la rosiglitazone (Avandia)

e) Autres traitements

D'autres traitements consistent à diminuer l'apport en dérivés hydrocarbonés. La première manière de réduire ces apports est d'appliquer un régime alimentaire hypoglucidique. Une autre voie est l'inhibition de certaines enzymes impliquées dans l'hydrolyse intestinale des polysaccharides (amidon, sucrose). En effet, la dégradation de l'amidon en monosaccharide est indispensable pour permettre son absorption. L'inhibition de

deux enzymes responsables de la dégradation est envisageable et plus particulièrement les α-amylases et les α-glucosidases. Seuls les inhibiteurs d'α-glucosidase se sont révélés intéressants dans le traitement du diabète.

Nous ne traiterons pas des thérapies ciblant les complications telles que les inhibiteurs de polyols (inhibiteurs d'aldose réductase, inhibiteur d'aldose déshydrogénase) ainsi que les inhibiteurs de glycosylation.

f) En pratique

Dans les traitements actuels du diabète de type 2, on ne privilégie pas forcément les molécules de dernière génération (thiazolidinesdiones par exemple) par rapport à la metformine ou aux sulfonamides. L'administration chez les personnes à fort risque cardio-vasculaire de glitazone est également envisageable dans un but préventif.[115] Les traitements reposent principalement, en première intention, sur le respect d'un régime alimentaire, l'exercice physique et une variété d'agents pharmacologiques tels que l'insuline (ainsi que des analogues), les glinides, les biguanides, les sulfonylurées ainsi que les thiazolidinediones. Les premières générations d'anti-diabétiques possèdent des effets secondaires tels que la prise de poids (insuline, sulfonylurée, glinides, glitazones), hypoglycémie (insuline, sulfonylurée, glinides) ainsi que l'apparition d'oedèmes (Tableau 12). Le traitement en monothérapie échoue progressivement obligeant à combiner plusieurs thérapies (Glucovance, Metaglip, Avandamet).[116]

[115] Chou, K.C. Molecular Therapeutic Target for Type-2 Diabetes *Journal of proteome research* **2004**, *3*, 1284-1288.
[116] Rotella, D.P. Novel "Second-Generation" Approaches for the Control of Type 2 Diabetes *J. Med. Chem.* **2004**, *47*, 4111-4112.

Famille des sulfonylurées	
tolbutamide	Orinase
chlorpropamide	Diabinèse
tolazamide	Tolinase
acétohexamide	Dymélor
glipizide	Glucotrol
glyburide	Diabeta, Micronase
glimépiride	Amaryl
Famille des glinides (Meglitinides)	
repaglinide	Prandin
nateglinide	Starlix
Famille des biguanides	
metformine	Glucophage
Famille des glitazones (Thiazolidinediones)	
rosiglitazone	Avandia
pioglitazone	Actos
Famille des inhibiteurs d'α-glucosidase	
acarbose	Precose
miglitol	Glyset
Combinaisons antidiabétiques-metformine	
glyburide-metformine	Glucovance
glipizide-metformine	Metaglip
rosiglitazone-metformine	Avandamet

Tableau 12. Différentes familles de composés anti-diabétiques

Des molécules de la famille des thiazolidinediones, telles que la pioglitazone ou encore la rosiglitazone, ont prouvé leur efficacité dans l'activation des PPARγ. L'isoforme PPARα est, quant à lui, activé par la famille des fibrates (fénofibrate, bézafibrate). Un intérêt plus particulier est porté actuellement sur l'activation des deux isoformes PPARα et PPARγ.[117] Nous allons présenter ces récepteurs dans le paragraphe suivant.

3. Les PPARs: Peroxysome Proliferator-Activated Receptors

a) Généralités

Les PPARs (Peroxysome Proliferator-Activated Receptors) sont des récepteurs nucléaires découverts en 1990 par Isseman et Green[118] qui identifient chez le rongeur un récepteur activé par des proliférateurs de péroxysomes (pesticides, phtalates) appelé PPARα. Deux autres récepteurs ont également été découverts: PPARδ et γ. Les PPARs appartiennent à la superfamille des récepteurs nucléaires qui compte également les récepteurs X du foie (LXR

[117] Henke, B.R. Peroxisome Proliferator-Activated Receptor alpha/gamma Dual Agonists for the Treatment of Type 2 Diabetes *J. Med. Chem.* **2004**, *47*, 4118-4127.
[118] Isseman, I.; Green, S. Activation of a member of the steroid hormone receptor superfamily by peroxisome proliferators *Nature* **1990**, *347*, 645-650.

dont les hormones sont les oxystérols), les récepteurs à l'androgène (AR), à l'œstrogène (ER), aux xénobiotiques (PXR), de l'acide rétinoïque (RXR), de l'hormone thyroïdienne (RT) et enfin de la vitamine D (VDR). La localisation de chaque isoforme est diverse au sein de l'organisme:

-PPARα: exprimée dans les tissus dont la fonction métabolique est importante et plus précisément le foie, le cœur, les reins, le pancréas et le tissu adipeux brun.

-PPARδ: protéine ubiquitaire surtout exprimée dans le muscle, le cerveau, le tissu adipeux et la peau.

-PPARγ: trois isoformes de PPARγ sont présents chez l'homme (provenant de trois ARN_m différents). Le $PPAR\gamma_1$ existe au niveau du tissu adipeux, des reins et du foie, le $PPAR\gamma_2$ est présent dans le tissu adipeux et le $PPAR\gamma_3$ est principalement exprimé dans le macrophage, l'intestin et le tissu adipeux.

b) Mécanisme d'action

Les PPARs forment des hétérodimères avec les récepteurs à l'acide 9-cis rétinoïque. L'hétérodimère est essentiel pour se lier à la séquence hexanucléotidique AGGTCA. Cet enchaînement est appelé PPRE (PPAR Response Element). Les PPARs jouent le rôle de facteurs de transcription. Dès lors que le complexe RXR/PPAR/PPRE est formé, l'ADN peut être transcrit en ARN_m. Ce type d'activité est appelé «transactivation». La fixation d'un co-répresseur sur le complexe RXR/PPAR/PPRE n'induira pas de transactivation (Figure 46). Un tel blocage peut être détourné en faisant interagir un agoniste spécifique à PPARγ. L'association PPAR/co-répresseur n'est alors plus valide et se dissocie. Le changement conformationnel du récepteur PPAR après réception de son agoniste aide au recrutement d'un co-activateur, nécessaire à la transcription des gènes situés sur l'ADN. Ces petits facteurs activateurs ou répresseurs sont donc d'une extrême importance.[119]

[119] Nolte, R.T., Wisely, B., Westin, S., Cobb, J.E., Lambert, M.H., Kurokawa, R., Rosenfeld, M.G., Willson, T.M., Glass, C.K., Milburn, M.V. Ligand binding and co-activator assembly of the peroxisome proliferators-activated receptor-γ *Nature* **1998**, *395*, 137-143.

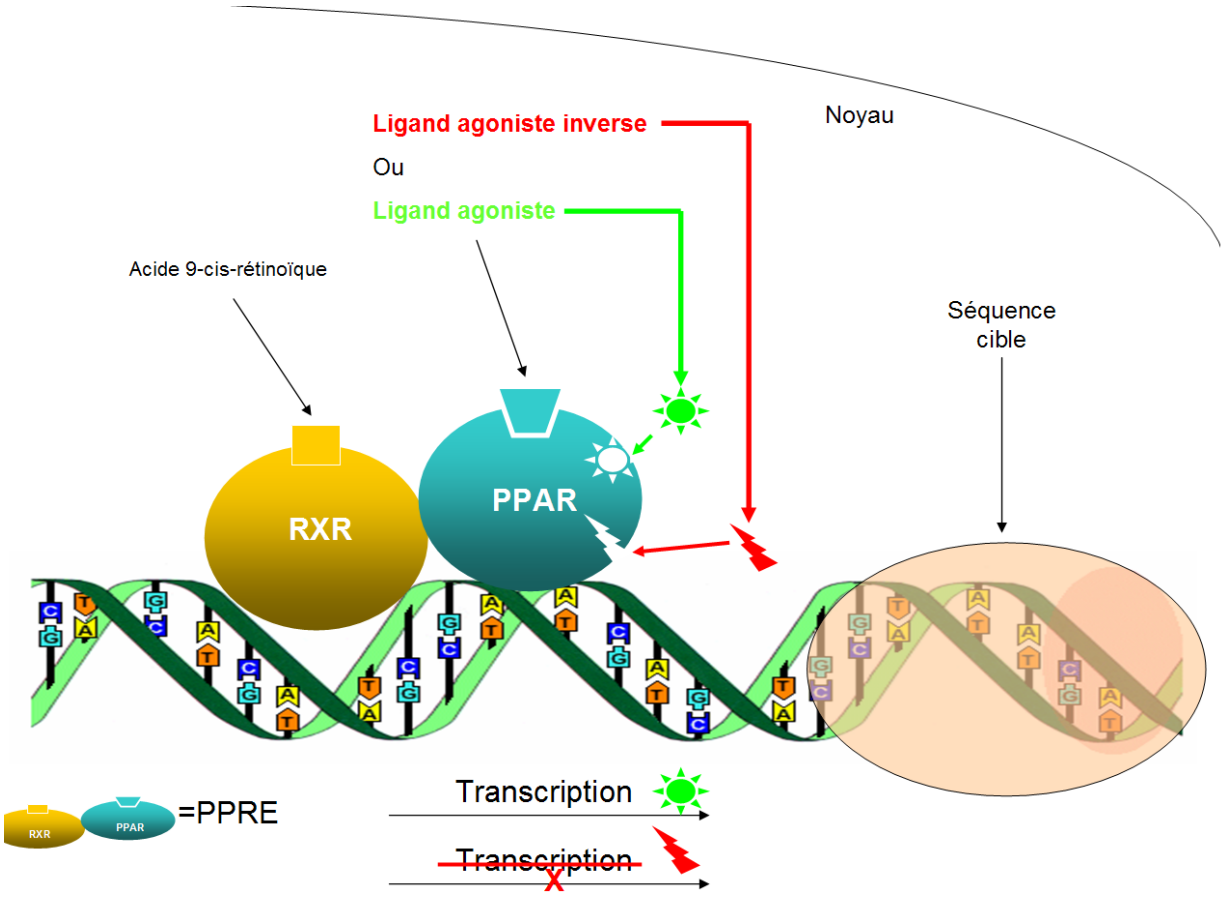


Figure 46. Mécanisme de reconnaissance de l'ADN par le complexe RXR/PPAR

c) Homologie de séquence des trois PPARs

La zone A-B est encore mal connue. Elle pourrait expliquer les différences d'activité biologique observée chez chacun des trois isoformes (Figure 47).[120]

La zone C est la région de fixation du complexe à l'ADN. L'arrimage entre les deux entités se fait par le biais d'une structure en doigt de zinc. Cette région hautement conservée d'un isoforme à l'autre (homologie de l'ordre de 80%) est dénommée «DNA Binding Domain» (DBD).

La zone E/F correspond au site de fixation des ligands PPAR (encore appelé «Ligand Binding Domain», LBD). Cette zone est moins conservée que ne l'est le DBD.[117]

[120] Willson, T.M.; Brown, P.J.; Sternbach, D.D.; Henke, B.R. The PPARs: From Orphan Receptors to Drug Discovery *J. Med. Chem.* **2000**, *43*, 527-550.

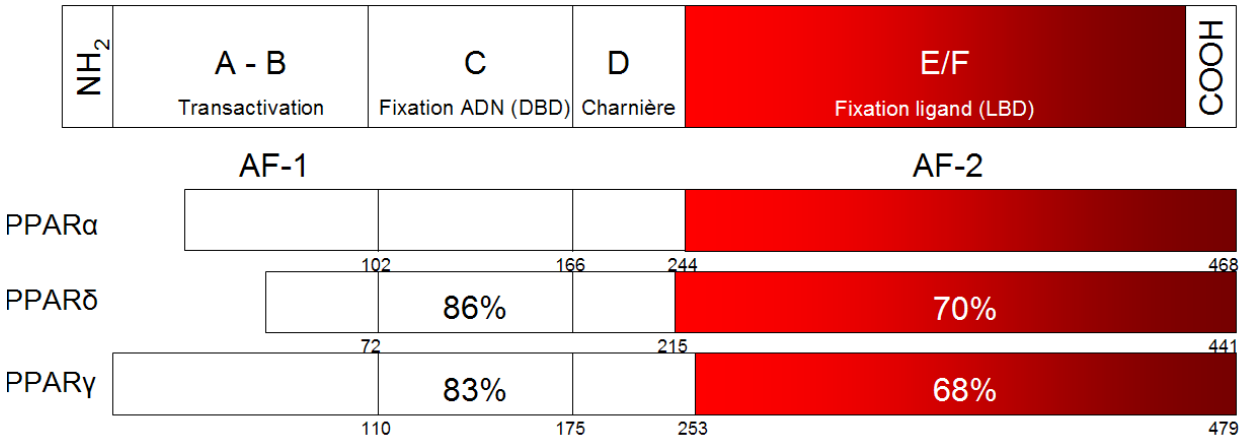


Figure 47. Structure générale et pourcentage de conservation des trois isoformes

La séquence peptidique du LBD entre les trois isoformes n'est conservée qu'à 69% tandis que le DBD a une homologie de séquence de l'ordre de 84,5%.

4. Ligands naturels et synthétiques de PPARα, δ et γ

a) Ligands PPARα

(1) Agonistes naturels

Les acides gras insaturés ou saturés tels que l'acide palmitique, l'acide oléique, l'acide linoléique et l'acide arachidonique sont des ligands naturels de PPARα. La lipoxygénase est également une source de ligands puissants de la famille des eicosanoïdes (provenant du métabolisme de l'acide arachidonique et l'acide linoléique) tels que l'acide 8(S)-hydroxyeicosatétranoïque (8(S)-HETE) et le leukotriène B4 (LTB4).

(2) Agonistes de synthèse

Des agents de synthèse de la famille des fibrates (fénofibrate) ont des propriétés hypolipidémiantes (Figure 48). Ils activent spécifiquement le PPARα. Le bézafibrate a la particularité d'activer les trois types de PPARα, δ, γ. D'autres ligands synthétiques spécifiques de PPARα ont été mis au point (GW9578).[121]

[121] Brown, P.J.; Winegar, D.A.; Plunket, K.D.; Moore, L.B.; Lewis, M.C.; Wilson, J.G.; Sundseth, S.S.; Koble, C.S.; Wu, Z.; Chapman, J.M.; Lehmann, J.M.; Kliewer, S.A.; Willson, T.M. A Ureido-Thioisobutyric Acid (GW9578) Is a Subtype-Selective PPARalpha Agonist with Potent Lipid-Lowering Activity *J. Med. Chem.* **1999**, *42*, 3785-3788.

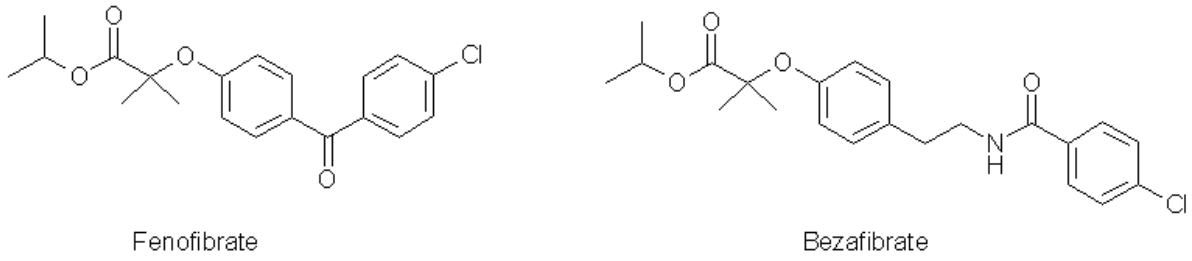


Figure 48. Structures d'agonistes de PPARα

b) Ligands PPARδ

(1) Agonistes naturels

L'isoforme PPARδ interagit, comme dans le cas de PPARα, avec des acides gras saturés et insaturés (acide arachidonique et acide eicosapenténoique).[122]

(2) Agonistes de synthèse

Il n'existe pas d'agoniste PPARδ sur le marché du médicament. Toutefois, afin d'étudier la pharmacologie de cette cible encore mal connue, un puissant activateur, le GW501516, a été testé (Figure 49).[123]

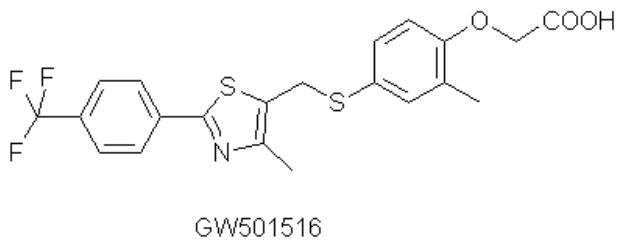


Figure 49. Structure d'un agoniste synthétique de PPARδ.

[122] Xu, H.E., Lambert, M.H., Montana, V.G., Parks, D.J., Blanchard, S.G., Brown, P.J., Sternbach, D.D., Lehmann, J.M., Wisely, G.B., Willson, T.M., Kliewer, S.A., Milburn, M.V. Molecular recognitionof fatty acids by Peroxysome Proliferator-Activated Receptors *Mol. Cell.* **1999**, *3*, 397-403.

[123] Szaidman, M.L.; Haffner, C.D.; Maloney, P.R.; Fivush, A.; Chao, E.; Goreham, D.; Sierra, M.L.; LeGrumelec, C.; Xu, H.E.; Montana, V.G.; Lambert, M.H.; Willson, T.M.; Olivier Jr, W.R.; Sternbach, D.D. Novel small molecules agonists for peroxysome proliferators-activated receptor delta: synthesis and biological activity *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1517-1521.

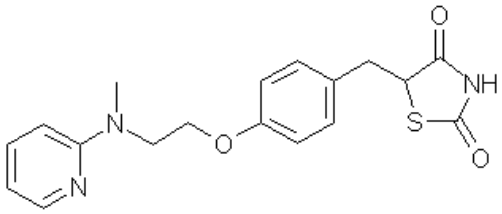
c) Ligands PPARγ

(1) Agonistes naturels

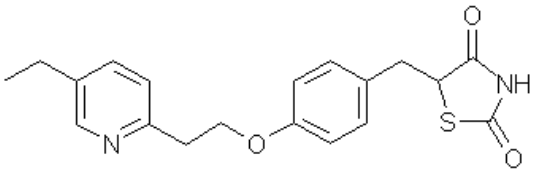
Les acides gras et les eicosanoïdes activent le PPARγ. Le 13-HODE (provenant du métabolisme de l'acide linoléique *via* la lipoxygénase) ainsi que certaines prostaglandines (15-deoxy-Δ 12,14-prostaglandines J2) sont des ligands naturels de PPARγ.

(2) Agonistes de synthèse

Des analogues des fibrates, les thiazolidinediones, ont été les premiers agonistes spécifiques du PPARγ (rosiglitazone,[124] pioglitazone,[119] figure 50). Ils agissent en régulant le taux de glucose dans le sang, par augmentation de la sensibilité à l'insuline.



Rosiglitazone



Pioglitazone

Figure 50. Structures d'agonistes synthétiques de PPARγ

D'autres ligands «non thiazolidinedione» plus prometteurs sont les analogues de la tyrosine. Ils appartiennent à la famille des glitazars (farglitazar, figure 51).[125,126] Enfin, des N-sulfonyl-2-indole carboxamides ont été décrits comme étant capables de se fixer sur les récepteurs PPARγ et d'induire une réponse dans le traitement de l'ostéoporose.[127]

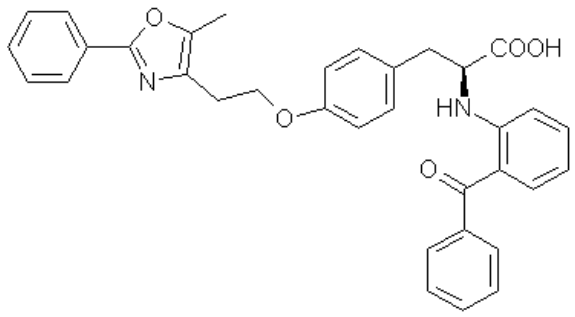
---

[124] Haffner, C.D.; Lenhard, J.M.; Miller, A.B.; McDougald, D.L.; Dwornik, K.; Ittoop, O.R.; Gampe Jr., R.T.; Xu, H.E.; Blanchard, S.; Montana, V.G.; Consler, T.G.; Bledsoe, R.K.; Ayscue, A.; Croom, D. Structure-based design of potent retinoid X receptor alpha agonists *J. Med. Chem.* **2004**, *47*, 2010-2029.

[125] Gampe Jr., R.T.; Montana, V.G.; Lambert, M.H.; Miller, A.B.; Bledsoe, R.K.; Milburn, M.V.; Kliewer, S.A.; Willson, T.M.; Xu, H.E. Asymmetry in the PPARgamma/RXRalpha crystal structure reveals the molecular basis of heterodimerization among nuclear receptors *Mol. Cell.* **2000**, *5*, 545-555

[126] Lehmann, J.M.; Lenhard, J.M.; Orband-Miller, L.A.; Miller, J.F.; Mook, R.A.; Noble, S.A.; Oliver, W.; Parks, D.J.; Plunket,K.D.; Szewczyk, J.R.; Willson, T.M. N-(2-Benzoylphenyl)-L-tyrosine PPAR gamma Agonists. 1. Discovery of a Novel Series of Potent Antihyperglycemic and Antihyperlipidemic Agents *J. Med. Chem.* **1998**, *41*, 5020-5036.

[127] Hopkins, C.R.; O'Neil, S.V.; Laufersweiler, M.C.; Wang, Y.; Pokross, M.; Mekel, M.; Evdokimov, A.; Walter, R.; Kontoyianni, M.; Petrey, M.E.; Sabatakos, G.; Roesgen, J.T.; Richardson, E.; Demuth, T.P. Design and synthesis of novel N-sulfonyl-2-indole carboxamides as potent PPAR-gamma binding agents with potential application to the treatment of osteoporosis *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5659-5663.

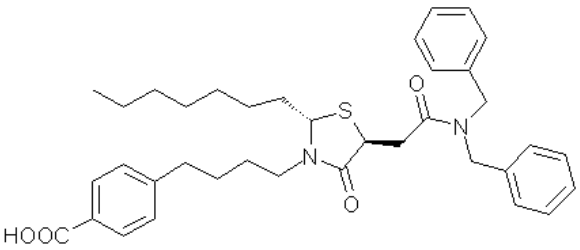


Farglitazar

Figure 51. Structure d'un agoniste synthétique non « TZD » de PPARγ

(3) Agoniste partiel de synthèse

Le GW0072 (Figure 52) se présente comme un agoniste partiel de PPARγ induisant une transactivation équivalente à 20% de l'efficacité de la rosiglitazone.[128]



GW0072

Figure 52. Structure d'un agoniste partiel synthétique de PPARγ

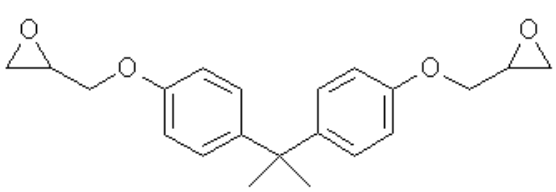
(4) Antagonistes de synthèse

Le GW9662 (Figure 53) est un antagoniste du PPARγ.[129] Le PD068235 ou encore le BADGE ont également un potentiel puissant capable de bloquer l'action de la rosiglitazone et de l'insuline (blocage de l'adipogénèse).

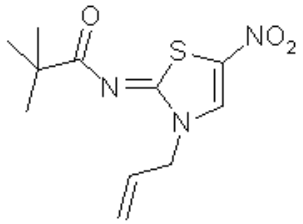
---

[128] Oberfield, J.L.; Collins, J.L.; Holmes, C.P.; Goreham, D.M.; Cooper, J.P.; Cobb, J.E.; Lenhard, J.M.; Hull-Ryde, E.A.; Mohr, C.P.; Blanchard, S.G.; Parks, D.J.; Moore, L.B.; Lehmann, J.M.; Plunket, K.; Miller, A.B.; Milburn, M.V.; Kliewer, S.A.; Willson, T.M. A peroxisome proliferator-activated receptor gamma ligand inhibits adipocyte differentiation *Biochemistry* **1999**, *96*, 6102-6106.

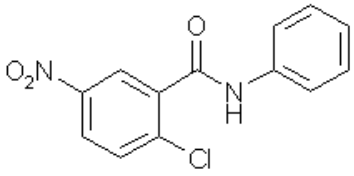
[129] Lee, G.; Elwood, F.; McNally, J.; Weiszmann, J.; Lindstrom, M.; Amaral, K.; Nakamura, M.; Miao, S.; Cao, P.; Learned, R.M.; Chen, J.L.; Li, Y. T0070907, a Selective Ligand for Peroxisome Proliferator-activated Receptor gamma, Functions as an Antagonist of Biochemical and Cellular Activities *J Biol Chem.* **2002**, *277*, 19649-19657.



BADGE



PD068235



GW9662

Figure 53. Structures d'antagonistes synthétiques de PPARγ

d) Co-agonistes PPARα/γ

Un grand nombre d'entités chimiques possédant des activités PPARα et γ sont actuellement en études pré-cliniques et cliniques.[130,131,132,133] Deux composés de la famille des glitazar, le ragaglitazar[134] et le tesaglitazar induisent une diminution des taux de triglycérides circulants et de glucose plasmatique (ragaglitazar) ainsi qu'une amélioration de l'efflux de cholestérol des macrophages vers les HDL (Figure 54).

---

[130] Koyama, H.; Miller, D.J.; Boueres, J.K.; Desai, R.C.; Jones, A.B.; Berger, J.P; MacNaul, K.L.; Kelly, L.J.; Doebber, T.W.; Wu, M.S.; Zhou,G.; Wang,P.R.; Ippolito, M.C.; Chao, Y.S.; Agrawal, A.K.; Franklin, R.; Heck, J.V.; Wright, S.D.; Moller, D.E.; Sahoo, S.P. (2R)-2-Ethylchromane-2-carboxylic Acids: Discovery of Novel PPARalpha/gamma Dual Agonists as Antihyperglycemic and Hypolipidemic Agents *J. Med. Chem.* **2004**, *47*, 3255-3263.

[131] Sauerberg, P.; Pettersson, I.; Jeppesen, L.; Bury, P.S.; Mogensen, J.P.; Wassermann, K.; Brand, C.L.; Sturis, J.; Wöldike, H.F.; Fleckner, J.; Andersen, A.S.T.; Mortensen, S.B.; Svensson, L.A.; Rasmussen, H.B.; Lehmann, S.V.; Polivka, Z.; Sindelar, K.; Panajotova, V.; Ynddal, L.; Wulff, E.M. Novel Tricyclic-α-lkyloxyphenylpropionic Acids: Dual PPARα/γ Agonists with Hypolipidemic and Antidiabetic Activity. *J. Med. Chem.* **2002**, *45*, 789-804.

[132] Shi, G.Q.; Dropinski, J.F.; McKeever, B.M.; Xu, S.; Becker, J.W.; Berger, J.P.; MacNaul, K.L.; Elbrecht, A.; Zhou, G.; Doebber,T.W.; Wang, P.; Chao, Y.S.; Forrest, M.; Heck, J.V.; Moller, D.E.; Jones, A.B. Design and Synthesis of r-Aryloxyphenylacetic Acid Derivatives: A Novel Class of PPARgamma/alpha Dual Agonists with Potent Antihyperglycemic and Lipid Modulating Activity *J. Med. Chem.* **2005**, *48*, 4457-4468.

[133] Pinelli, A.; Godio, C.; Laghezza, A.; Mitro, N.; Fracchiolla, G.; Tortorella, V.; Lavecchia, A.; Novellino, E.; Fruchart,J.C.; Staels, B.; Crestani, M.; Loiodice, F. Synthesis, Biological Evaluation, and Molecular Modeling Investigation of New Chiral Fibrates with PPAR gamma and PPAR alpha Agonist Activity *J. Med. Chem.* **2005**, *48*, 5509-5519.

[134] Ebdrup, S.; Pettersson, I.; Rasmussen, H.B.; Deussen, H.J.; Frost Jensen, A.; Mortensen, S.B.; Fleckner, J.; Pridal, L.; Nygaard, L.; Sauerberg, P. Synthesis and biological and structural characterization of the dual-acting peroxisome proliferator-activated receptor alpha/gamma agonist ragaglitazar *J. Med. Chem.* **2003**, *46*, 1306-1317.

Ils ont également la capacité de réduire la perte en îlots pancréatiques spécifique au diabète.

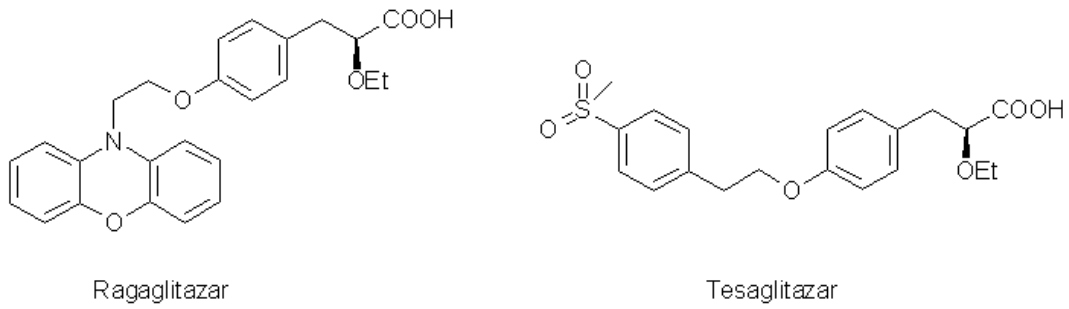


Figure 54. Structures d'agonistes synthétiques PPARα/γ

5. Implication des différents isoformes

a) PPARα

La dyslipidémie est une cause de l'athérosclérose. Ce phénomène peut être diagnostiqué par l'augmentation des concentrations plasmatiques des triglycérides (TG) et de la lipoprotéine LDL (riches en lipides). Il est également caractérisé par une diminution plasmatique de la lipoprotéine HDL (impliquée dans le transport du cholestérol vers le foie). Des processus inflammatoires peuvent également être à l'origine de l'athérosclérose. Les agonistes de l'isoforme PPARα constituent un traitement efficace de ces évènements métaboliques et inflammatoires.[120]

(1) Métabolisme lipidique

Le métabolisme des acides gras est effectué par la β-oxydation au sein des mitochondries et des peroxysomes dans le foie. Le récepteur PPARα intervient à ce niveau puisqu'il contrôle la transcription de gènes codant pour des protéines impliquées dans le métabolisme des acides gras et plus particulièrement leurs passages par transporteur membranaire (« Fatty Acid Transport Protein » ainsi que la « Fatty Acid Translocase »). La β-oxydation mitochondriale des acides gras est également médiée par une protéine transcrite par activation de PPARα. C'est l'Acyl CoA Oxydase.

Les lipoprotéines LDL et HDL jouent également un rôle prépondérant dans la dyslipidémie. Plus particulièrement, l'activation de PPARα induit l'expression de l'apolipoprotéine C-III et de la lipoprotéine lipase permettant l'hydrolyse des triglycérides

constituant les LDL. Enfin, l'activation de PPARα conduit à l'expression de deux apolipoprotéines (apo-AI et apo-AII) entraînant l'augmentation de la concentration plasmatique de HDL.

(2) L'inflammation

Les agonistes des PPARα ont été désignés comme étant des inhibiteurs d'un facteur de transcription de médiateurs de l'inflammation appelé NF-κB. Les formations athérosclérotiques nécessitent le recrutement de monocytes dans les artères *via* l'expression de molécules d'adhésion synthétisées par les cellules endothéliales. Ces agents sont appelés «Vascular Cell Adhesion Molecule» (VCAM-1) et sont exprimés par la voie NF-κB.

L'inhibition de NF-κB conduit également à la diminution de l'expression de cyclooxygénases associées aux processus inflammatoires.

Ainsi, les agonistes PPARα ont un double potentiel en tant que régulateurs du métabolisme lipidique et en tant qu'inhibiteurs du phénomène inflammatoire associé au risque de rupture de la plaque athérosclérotique.

b) PPARδ

Des études ont montré que l'activation de PPARδ est impliquée dans l'augmentation du cholestérol HDL plasmatique et dans la diminution des triglycérides et LDL basse densité.[135] La cible PPARδ pourrait alors être privilégiée pour le traitement de l'hyperlipidémie. Enfin, PPARδ pourrait être responsable d'une décroissance de la tumorogénèse des cellules cancéreuses du colon humain. Il serait alors une cible pour le traitement du cancer colorectal.

c) PPARγ

Les récepteurs PPARγ sont impliqués dans le contrôle de la différenciation adipocytaire. Leur activation conduit à favoriser la lipolyse et le stockage des triglycérides

[135] Oliver, W.R.; Jr., Shenk, J.L.; Snaith, M.R.; Russell, C.S.; Plunket, K.D.; Bodkin, N.L.; Lewis, M.C.; Winegar, D.A.; Sznaidman, M.L.; Lambert, M.H. *Proc. Natl. Acad. Sci.* **2001**, *98*, 5306-5311.

dans les adipocytes. Ceci contribuerait à diminuer les quantités d'acide gras et triglycéride circulantes.

Le récepteur PPARγ joue un rôle dans la diminution de la pression artérielle. Le processus selon lequel les thiazolidinediones agissent sur la contraction vasculaire demeure encore mal connu. PPARγ est également impliqué dans des processus liés à l'inflammation. La réduction du niveau de cytokine proinflammatoire TNFα dans les tissus adipeux a été mise en évidence. Enfin, d'autres études ont montré une implication des PPARγ dans le cancer.[136] Plus précisément, PPARγ est hautement exprimé dans les lignées cellulaires de cancer du sein, du colon, de la prostate, de la vessie et de l'estomac. Des agonistes de PPARγ (TZD) ont montré un intérêt dans l'induction de l'apoptose de plusieurs lignées cellulaires.

Enfin, un axe très prometteur dans la prévention et le traitement du cancer du colon est celui de la diminution de l'expression de l'enzyme COX-2, très présent dans ce type de cancer. Il a été démontré que l'activation des PPARγ (par traitement à la ciglitazone) diminue la biosynthèse de la cyclooxygénase de type 2.[137,138]

6. Structure cristallographique de PPARγ

a) Généralités

La structure apo de PPARγ révèle un site actif en forme de Y et dont le volume est d'approximativement 1300 $Å^3$ (Figure 55). Cette région s'étend de l'hélice α C-terminale au feuillet β situé entre l'hélice 3 et 6.[139]

[136] Mora, F.D.; Jones,D.K.; Desai,P.V.; Patny,A.; Avery,M.A.; Feller,D.R.; Smillie,T.; Zhou,Y.D.; Nagle, D.G. Bioassay for the Identification of Natural Product-Based Activators of Peroxisome Proliferator-Activated Receptor-gamma (PPARgamma): The Marine Sponge Metabolite Psammaplin A Activates PPARgamma and Induces Apoptosis in Human Breast Tumor Cells *J. Nat. Prod* **2006**, *69*, 574-552.
[137] Yang, W.-L.; Frucht, H. Activation of the PPAR patway induces apoptosis and COX-2 inhibition in HT-29 human colon cancer cells *Carcinogenesis* **2001**, *22*, 1379-1383.
[138] Han, S.; Inoue, H.; Flowers, L.C.; Sidell, N. Control of COX-2 gene expression through peroxisome proliferator-activated receptor gamma in human cervical cancer cells *Clin. Canc. Res.* **2003**, *9*, 4627-4635.
[139] Sheu, S.H.; Kaya, T.; Waxman, D.J.; Vajda, S. Exploring the Binding Site Structure of the PPAR gamma Ligand-Binding Domain by Computational Solvent Mapping *Biochemistry* **2005**, *44*, 1193-1209.

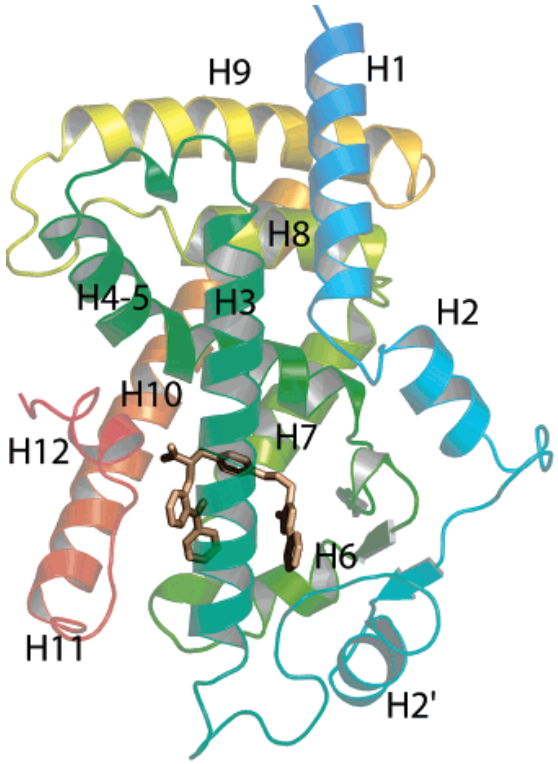


Figure 55. Structure tertiaire du récepteur PPARγ

Les hélices 3, 7 et 10 confèrent la forme de Y au site actif. La rosiglitazone, dont la conformation bioactive dans la protéine est en forme de U, n'occupe que 40% du site actif. Celle-ci forme des liaisons de type hydrogène avec l'His449, Tyr473, His323, Ser289 ainsi que Gln286 présumées essentielles pour les qualités d'un ligand agoniste.[140] Plusieurs formes de PPARγ ont été cristallisées avec des agonistes, agonistes partiels et des ligands naturels (acides gras, prostaglandines).

Les structures disponibles dans la Protein Data Bank (PDB) sont: 2HFP,[127] 1WMO,[141] 2ATH,[142] 2GOG,[143] 1ZEO,[132] 2GOH,[143] 1I7I,[144] 2F4B,[142] 1ZGY,[145] 2GTK,[146] 2PRG,[119]

---

[140] Iwata, Y.; Miyamoto, S.; Takamura, M.; Yanagisawa, H.; Kasuya, A. Interaction between peroxisome proliferator-activated receptor gamma and its agonists: Docking study of oximes having 5-benzyl-2,4-thiazolidinedione *J. Mol. Graph. Model.* **2001**, *19*, 536-542.

[141] Ostberg, T.; Svensson, S.; Selén, G.; Uppenberg, J.; Thor, M.; Sundbom, M.; Sydow-Bäckman, M.; Gustavsson, A.L.; Jendeberg, L. A New Class of Peroxisome Proliferator-activated Receptor Agonists with a Novel Binding Epitope Shows Antidiabetic Effects. *J. Biol. Chem.* **2004,** *279*, 41124-41130.

[142] Mahindroo, N.; Huang, C.F.; Peng, Y.H.; Wang, C.C.; Liao, C.C.; Lien, T.W.; Chittimalla, S.K.; Huang, W.J.; Chai, C.H.; Prakash, E.; Chen, C.P.; Hsu, T.A.; Peng, C.H.; Lu, I.L.; Lee, L.H.; Chang, Y.W.; Chen, W.C.; Chou, Y.C.; Chen, C.-T.; Goparaju, C.M.V.; Chen, Y.S.; Lan, S.J.; Yu, M.C.; Chen, X.; Chao, Y.S.; Wu, S.Y.; Hsieh, H.P. Novel indole-based peroxisome proliferator-activated receptor agonists: design, SAR, structural biology, and biological activities *J. Med. Chem.* **2005**, *48*, 8194-8208.

[143] Lu, I.L.; Huang, C.F., Peng, Y.H., Lin, Y.T., Hsieh, H.P.; Chen, C.T.; Lien, T.W.; Lee, H.J.; Mahindroo, N.; Prakash, E.; Yueh, A.; Chen, H.Y.; Goparaju, C.M.; Chen, X.; Liao, C.C.; Chao, Y.S.; Hsu, J.T.; Wu, S.Y. Structure-Based Drug Design of a Novel Family of PPARgamma Partial Agonists: Virtual Screening, X-ray Crystallography, and in Vitro/in Vivo Biological Activities *J. Med. Chem.* **2006**, *49*, 2703-2712.

3PRG,[147] 4PRG,[128] 1PRG,[119] 1RDT,[124] 1KNU,[131] 1NYX,[134] 1FM9,[125] 1K74[148] et 1FM6.[125] Les résolutions s'échelonnent entre 1,80 et 2,90 Å. Par ailleurs, chaque structure est issue de la cristallographie par rayons X sur des récepteurs de source humaine exprimés par la bactérie Escherichia coli.

Structure	Année	Résolution	R-Value	R-Free	Numéro de chaine	Ligand
2HFP	2006	2,00	0,209	0,285	A	N-sulfonyl-2-indole carboxamide
2ATH	2006	2,28	0,216	0,288	A,B	Dérivé acide indolyl-éthanoïque
1WMO	2004	2,90	0,195	0,295	X	2-BABA
2GOG	2006	2,54	0,216	0,264	A,B	Benzenesulfonamide
1ZEO	2006	2,50	0,222	0,280	A,B	Dérivé acide alpha-aryloxyphenylacétique
2GOH	2006	2,30	0,238	0,293	A,B	Benzenesulfonamide
1I7I	2002	2,35	0,238	0,284	Actif: A - Inactif: B	Dérivé methylsulfone (AZ242)
2F4B	2006	2,07	0,222	0,288	A,B	Dérivé acide indolyl-acétique
1ZGY	2005	1,80	0,231	0,258	A	Rosiglitazone
2GTK	2006	2,10	0,198	0,261	A	Dérivé indolique
2PRG	1999	2,30	0,207	0,264	A,B	Rosiglitazone
4PRG	1999	2,90	0,240	0,283	Actif: A,C - Inactif: B,D	GW0072
3PRG	1999	2,90	0,209	0,271	A	Aucun
1PRG	2001	2,20	0,246	0,318	Actif: A - Inactif: B	Aucun
1RDT	2004	2,40	0,221	0,259	D	Dérivé acide naphtalenyl-benzofuranyl propionique
1KNU	2002	2,50	0,224	0,264	A,B	Analogue de ragaglitazar
1NYX	2003	2,65	0,240	0,306	A,B	Ragaglitazar
1FM9	2001	2,10	0,239	0,268	D	Farglitazar
1K74	2001	2,30	0,238	0,279	D	GW409544
1FM6	2001	2,10	0,250	0,292	D,X	Rosiglitazone

Tableau 13. Description des différentes structures cristallographiques des PPARγ

En vert foncé sont figurées les agonistes des PPARγ, le vert clair représente les agonistes partiels et en blanc les structures Apo.

Certaines structures sont sous forme de dimère, l'une correspond à une structure de PPARγ active, l'autre est une forme inactive. C'est le cas par exemple de 4PRG, 1PRG et 1I7I. Les structures de chaque ligand co-cristallisé sont représentées sur la figure suivante:

---

[144] Cronet, P.; Petersen, J.F.W.; Folmer, R.; Blomberg, N.; Sjöblom, K.; Karisson, U.; Lindstedt, E.L.; Bamberg, K. Structure of the PPAR α and γ Ligand Binding Domain in Complex with AZ 242; Ligand Selectivity and Agonist Activation in the PPAR Family. *Structure* **2001**, *9*, 699-706.

[145] Li, Y.; Choi, M.; Suino, K.; Kovach, A.; Daugherty, J.; Kliewer, S.A.; Xu, H.E. Structural and biochemical basis for selective repression of the orphan nuclear receptor liver receptor homolog 1 by small heterodimer partner *Proc. Natl. Acad. sci.* **2005**, *102*, 9505-9510.

[146] Kuhn, B.; Hilpert, H.; Benz, J.; Binggeli, A.; Grether, U.; Humm, R.; Meyer, M.; Mohr, P. Structure-based design of indole propionic acids as novel PPARalpha/gamma co-agonists *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4016-4020.

[147] Uppenberg, J.; Svensson, C.; Jaki, M.; Bertilsson, G.; Jendeberg, L.; Berkenstam, A. Crystal structure of the ligand binding domain of the human nuclear receptor PPARgamma *J. Biol. Chem.* **1998**, *273*, 31108-31112.

[148] Xu, H.E.; Lambert, M.H.; Montana, V.G.; Plunket, K.D.; Moore, L.B.; Collins, J.L.; Oplinger, J.A.; Kliewer, S.A.; Gampe, R.T.; McKee, D.D.; Moore, J.T.; Willson, T.M. Structural determinants of ligand binding selectivity between the peroxisome proliferator-activated receptors *Proc. Natl. Acad. Sci.* **2001**, *98*, 13919-13924.

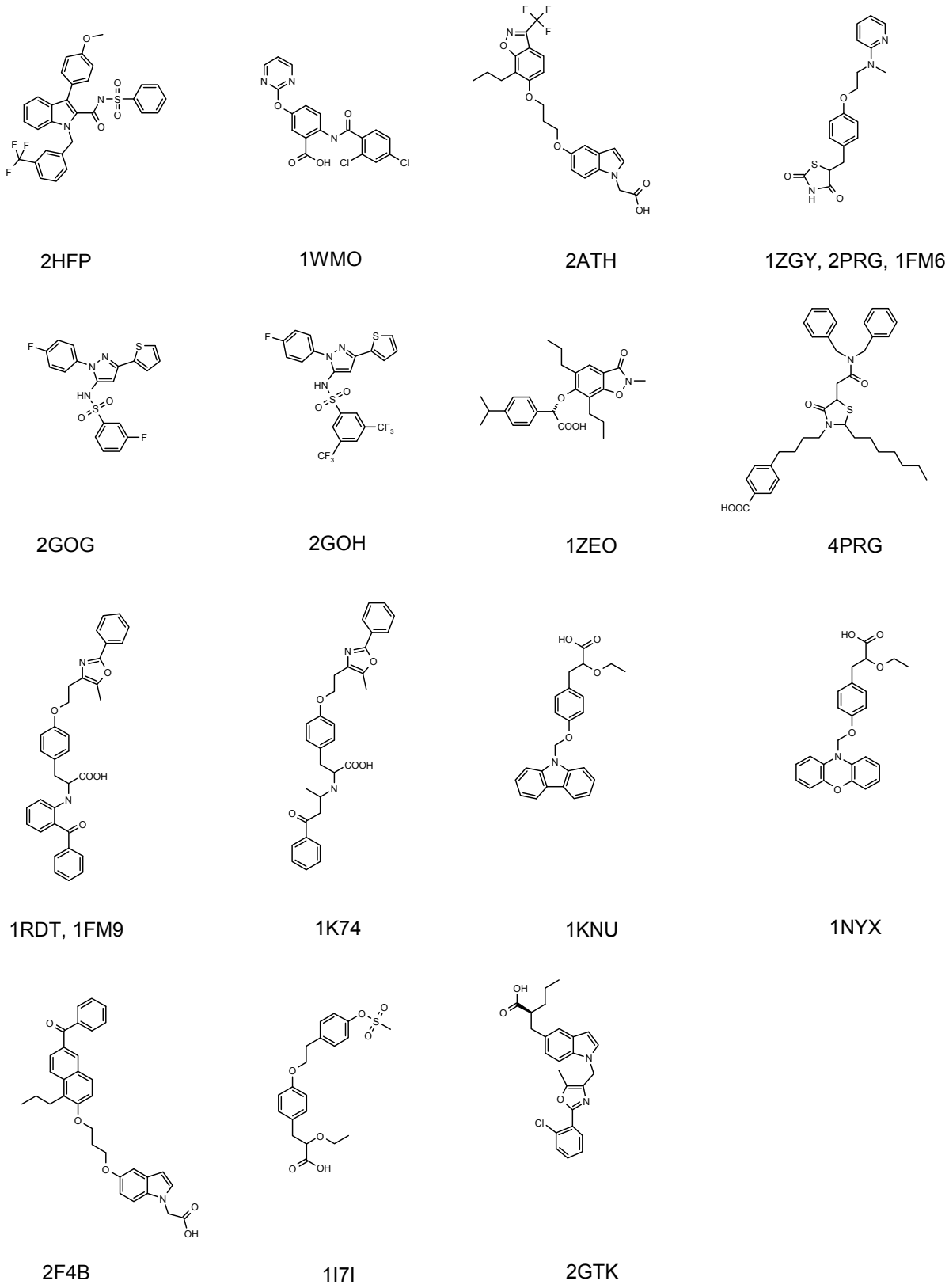


Figure 56. Structures des ligands co-cristallisés dans PPARγ

La figure 56 montre la diversité des ligands capables d'interagir avec la cavité de PPARγ.

7. Structure tertiaire des PPARγ

a) Variabilité spatiale

(1) Description

La structure tertiaire des PPARγ compte 12 hélices et un feuillet β de 4 brins. Les 11 acides aminés compris entre l'hélice 2b et 3 ont une haute mobilité (Figure 57).

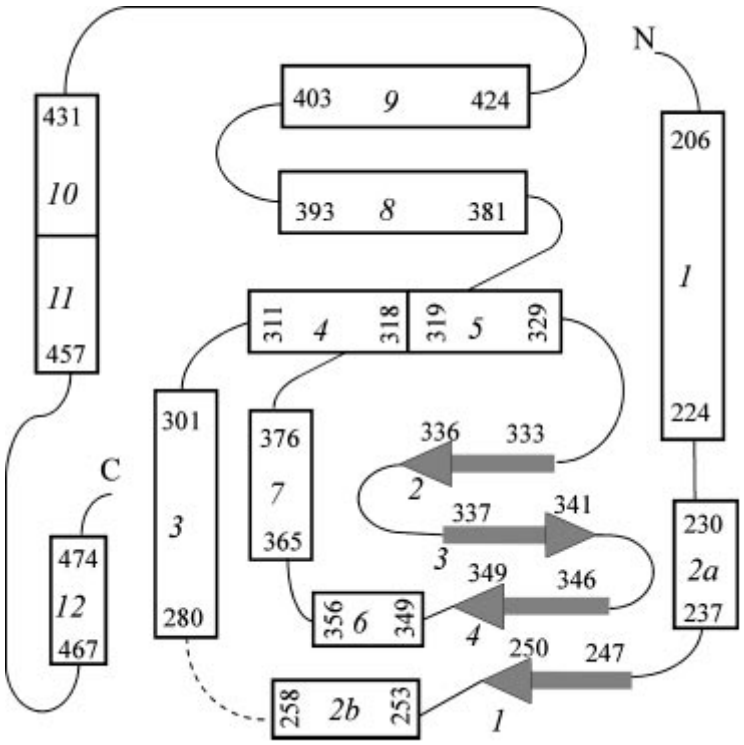


Figure 57. Structure secondaire de la protéine PPARγ

La cavité est majoritairement de nature hydrophobe. Ceci est en accord avec le type de ligands naturels qui s'y fixent: acides gras polyinsaturés, prostaglandines. La zone inférieure du site actif est constituée de résidus d'acides aminés appartenant aux hélices 3, 5, 10, 11 et 12 formant une surface polaire (His323, Tyr327, Lys367, His449, Tyr473).[147] Une seconde cavité s'étend de l'hélice 1 jusqu'au feuillet β. Cette zone est délimitée par les résidus Phe266, Pro227, Ile296 et Met329.

(2) Mécanismes de trans-conformation

Les PPARs sont dotés d'un mécanisme d'adaptabilité et de flexibilité permettant d'accueillir une grande diversité de molécules dans le site actif.

Des remaniements sont observés au niveau de H3 et H11. De même, l'hélice H12 est sujette à un déplacement de grande amplitude, lors du passage de la forme sans ligand (apo-structure) à la forme avec ligand (holo-structure). Elle agit ainsi comme un «piège à souris»[119]. L'hélice H12 module l'accessibilité d'une région «interface» (formée par H3, H4 et H12) sur laquelle peut se fixer un co-activateur hydrophobe. On appellera la conformation agoniste celle qui laisse la zone «interface» accessible au co-activateur. A l'inverse, la conformation antagoniste est une forme dans laquelle H12 gène le recrutement du co-activateur. Il y a compétition entre H12 et le co-activateur. Un cas intermédiaire est la fixation d'un agoniste partiel. Dans ce cas, le GW0072 (agoniste partiel des PPARγ) présente une transactivation équivalente à 20% de celle obtenue avec la rosiglitazone.

Afin de différencier deux conformations, nous avons superposé une partie du site actif de la structure 2PRG (en vert, cristallisée avec la rosiglitazone) et de 4PRG (en rouge, cristallisée avec GW0072) (Figure 58). Seuls les résidus impliqués dans un réseau de liaisons hydrogène avec les agonistes PPARγ ont été représentés (His323, His449 et Tyr473). Les His323 et His449 occupent un emplacement constant tandis que la Tyr473, portée par H12, varie significativement d'une structure à l'autre.

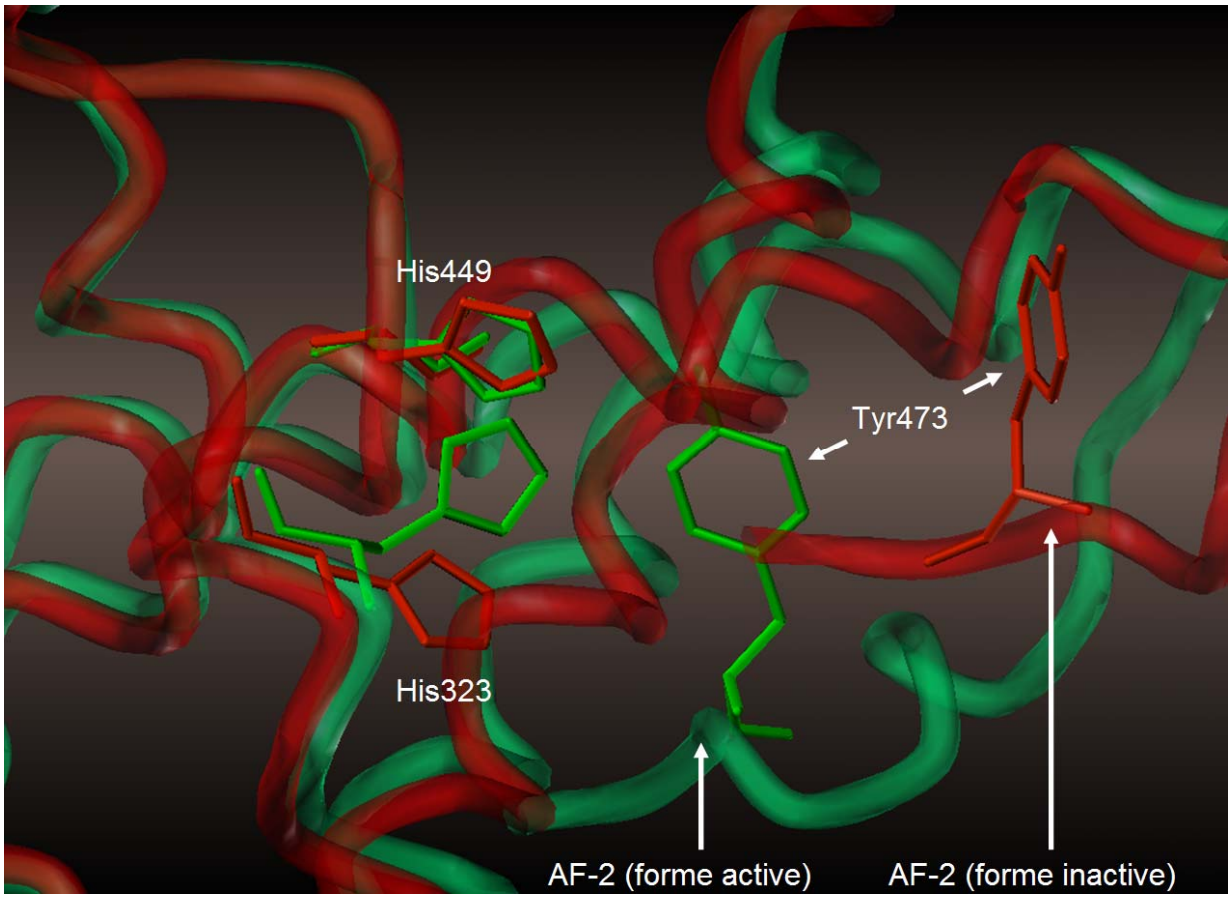


Figure 58. Superposition de 2PRG et de 4PRG

Dans le cas de la forme «active» ou conformation agoniste (structure verte), la fonction hydroxyle de la Tyr473 pointe vers l'intérieur du site actif et interagit ave la tête polaire des agonistes. Cette association permet de verrouiller H12.

A l'inverse, la forme «inactive» ou conformation agoniste inverse (structure rouge) rejette la Tyr473 vers l'extérieur, supprimant ainsi toute interaction hydrogène qui aurait stabilisé H12.

On peut alors corréler le degré de transactivation d'une molécule avec son potentiel à interagir avec H12, rendant accessible la zone «interface» à un co-activateur. Le complexe est alors stabilisé.

Les récepteurs aux oestrogènes (de la famille des récepteurs nucléaires) ont été décrits comme étant similaires structuralement aux PPARγ.[149] D'ailleurs, l'hélice H12 des récepteurs oestrogéniques est sensible au profil pharmacologique du ligand (selon qu'il est de nature agoniste ou antagoniste).

[149] Brzozowski, A.M.; Pike, A.C.W.; Dauter.Z.; Hubbard, R.E.; Bonn, T.; Engström, O.; Öhman, L.; Greene, G.L.; Gustafsson, J.A.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor *Nature* **1997**, *389*, 753-758.

b) Variabilité interactionnelle

En plus de la liberté conformationelle de H3, H11 et H12, des variations importantes des résidus «clé» du site actif ont été mises en évidence par des méthodes de cartographie. En effet, ces techniques utilisant des sondes moléculaires mimant des solvants (acétonitrile, acétone, tert-butanol, phenol, méthanol, 2-propanol et urée) ont mis en évidence des différences en termes d'interaction entre 8 structures cristallographiques et leur ligand co-cristallisé. La Figure 59 illustre les interactions non-liées (partie A) ainsi que les liaisons hydrogène (partie B) entre 10 structures cristallographiques issues de la PDB et leur ligand co-cristallisé:[139]

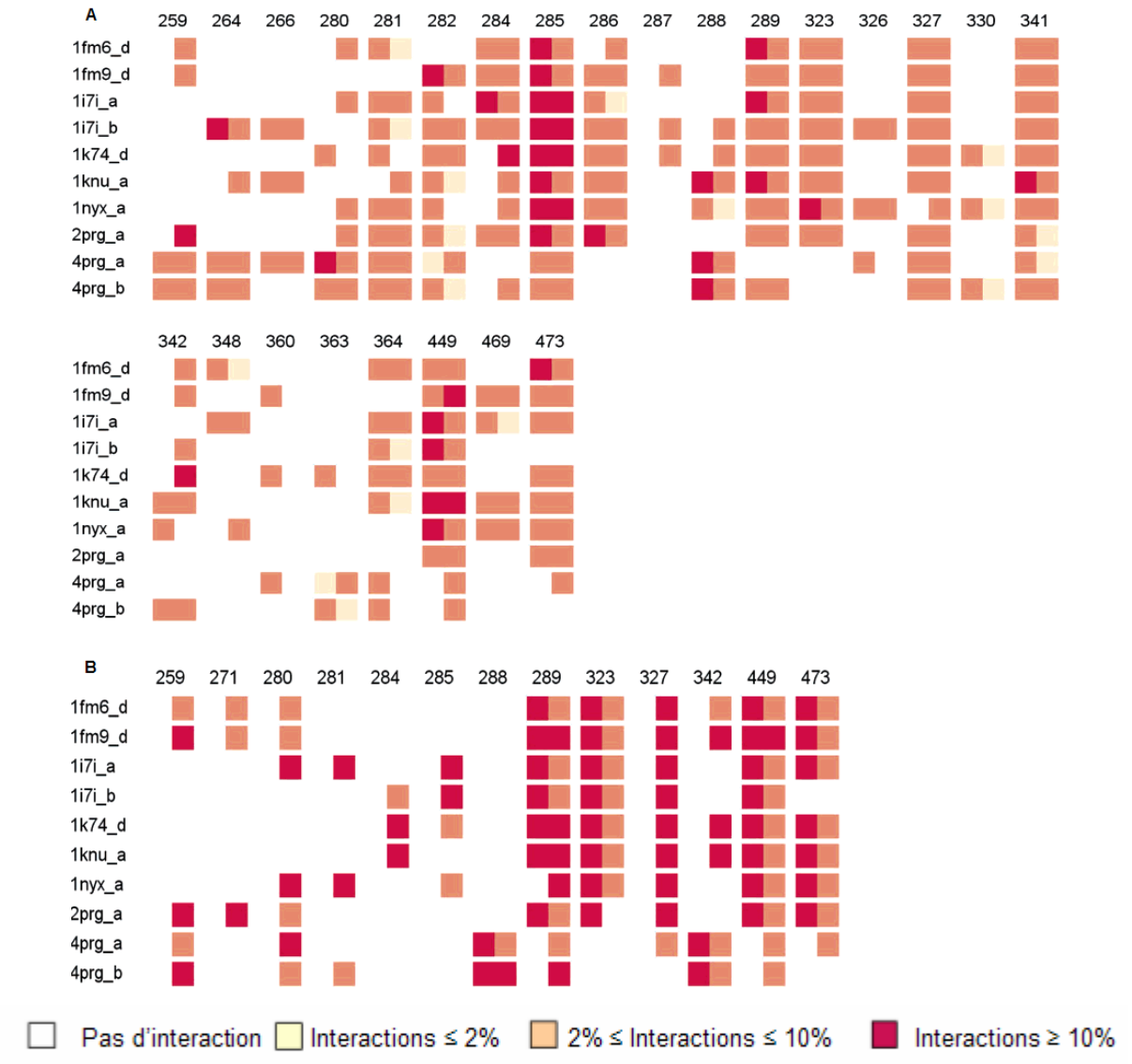


Figure 59. Distribution des interactions intermoléculaires entre ligands et résidus du LBD de la protéine PPARγ

Les interactions ont été déterminées par deux méthodes: la première est l'utilisation du programme HBPLUS[150] (partie gauche de chaque rectangle) alors que la deuxième est le calcul d'interaction par cartographie *via* les sondes décrites précédemment.

Sheu *et al* ont également déterminé différents sites au sein du LBD de PPARγ (Figure 60). Plus précisément, le site P1 (Phe282, Cys285, Gln286, Ser289, His323, Tyr327, Phe363, His449, Lys469, Tyr473) est présent lorsque le récepteur est co-cristallisé avec un agoniste possédant des fonctions capables d'interagir par liaisons hydrogène (fonction thiazolidinedione) avec l'hélice 12.

Le site P2, localisé entre H3 et H5, (Cys285, Arg288, Ser289, Ile326, Tyr327, Lys330, Phe363, Met364) est présent dans le cas de la structure « apo » de PPARγ (structure cristallisée seule). La structure « holo » cristallisée avec l'agoniste partiel GW0072 fait apparaître le site P2 au lieu du site P1. Ce ligand est particulier puisqu'il est le seul à atteindre le site P3 coincé entre l'hélice H2 et H3 (Arg288, Glu291, Ala292, Asp295, Met329, Asp343). P4 est un morceau de la poche à tendance hydrophobe.

La zone B est comprise entre H7 et H10/11 tandis que F se situe entre H3 et H12. C1 représente une surface entre H10 et H12 qui contribue certainement à l'arrimage du cofacteur. C2 est une zone qui permet le recrutement du co-activateur SRC-1. E1 est définie par la partie terminale des hélices H3, H7 et H10. Enfin, E2 figure l'accès des ligands entre H2' et le feuillet β.

[150] www.biochem.ucl.ac.uk/bsm/hbplus/home.html.

Figure 60. Répartition des différents sites de liaison aux ligands.

L'arrivée des ligands jusqu'à P1 est nécessaire pour que le PPARγ se mette sous la forme active. Les interactions avec des résidus tels que His323, His449 et Tyr473 contribuent à stabiliser H12. Ceci conduit au recrutement d'un co-activateur. La confirmation de la Figure 60 est faite par Ebdrup *et al* qui montre que la Tyr473, dans la forme inactive du récepteur, n'est pas dans une position propice à l'interaction avec un ligand.[134]

J. Matériel et méthodes

L'objectif de ce travail est d'isoler des composés affins pour la cible PPARγ. Tous les modèles ont été élaborés de manière identique au chapitre précédent. L'algorithme de docking FlexX a été utilisé ainsi que le module C-Score contenant les cinq fonctions de scoring F-Score, D-Score, PMF, G-Score et Chemscore. Enfin, l'analyse de données a été réalisée par consensus et analyse factorielle discriminante. L'*ensemble total* a été élaboré sur une base d'agonistes ce qui signifie que le modèle devra être capable d'isoler ce type d'agent pharmacologique. Après avoir mis en évidence la variabilité de volume du site actif, nous

avons choisi de coupler la dynamique moléculaire en amont du protocole de docking-scoring afin de prendre en compte la flexibilité du site de fixation des récepteurs PPARγ.

1. Constitution de l'*ensemble total*

a) Généralités

L'*ensemble total* est constitué de molécules actives (20% de l'*ensemble total*) et de molécules inactives (représentant 80% de l'*ensemble total*) provenant de plusieurs laboratoires avec qui nous sommes en relation. Cet ensemble a été constitué exclusivement pour la construction d'un modèle PPARγ. Du fait que l'étude ne porte que sur PPARγ, aucune donnée concernant la composante PPARα n'a été ajoutée.

b) Molécules actives

Des données d'affinité (K_i) nous ont permis de récupérer les molécules les plus affines (limite fixée à un K_i=50nM) provenant de différents laboratoires. Les molécules utilisées dans cette étude font l'objet de confidentialité. Nous avons représenté les structures de manière générale (Figure 61).

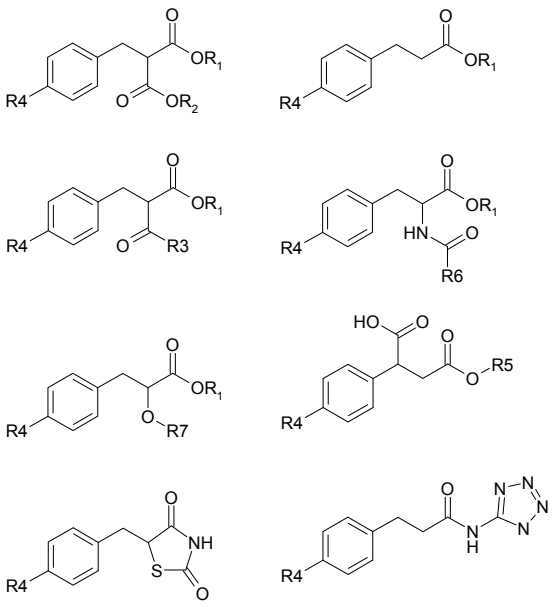


Figure 61. Structures des molécules actives

Nous avons également ajouté des molécules «référence» agonistes des PPARγ telles que le GW409544 (produit en phase III, EC_{50} = 0,28 nM), le tesaglitazar (produit en phase III, EC_{50} = 1300 nM), le ragaglitazar (EC_{50} = 600 nM), le GW1929 (EC_{50} = 6,2 nM) et le farglitazar (EC_{50} = 0,34 nM).[126] Au total, 70 molécules actives ont été regroupées.

c) Molécules inactives (*leurres*)

Nous avons complété ce jeu de données par l'ajout de 280 *leurres*, provenant d'une chimiothèque virtuelle d'environ 3,8 millions de composés. Nous avons focalisé la sélection sur des critères propres aux molécules actives sur PPARγ. Le calcul de descripteurs sur les 70 molécules actives nous a permis de déduire des critères «PPAR-like»:

$-6 \leq$ Nombre de liaisons «Rotable» ≤ 22
$-2 \leq$ Nombre d'accepteurs de liaisons hydrogène ≤ 11
$-1 \leq$ Nombre de petits cycles ≤ 6
$-70 \leq$ surface polaire topologique (TPSA) ≤ 150
-Taille des grands cycles ≤ 7
$-290 \leq$ Masse moléculaire ≤ 670
-Log P ≤ 8
$-NO_2 = 0$

Au total, 1 063 653 molécules ont été sélectionnées par Screening Assistant.[112] Par la suite, un algorithme de diversité a été appliqué à cet ensemble pour finalement sélectionner les 280 molécules à diversité maximum.[151]

2. Dynamique moléculaire

Tout le travail de dynamique moléculaire ainsi que la génération de structure a été réalisé par l'équipe du Dr Daniel Genest, dans le cadre de la collaboration entre le CBM (Centre de Biophysique moléculaire) et l'ICOA. L'utilisation de cette technique en amont du protocole de docking-scoring nous a permis d'identifier des formes préférentielles de PPARγ et ainsi de réaliser les opérations de docking sur des conformations préalablement choisies.

[151] Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512-524.

a) Détermination des résidus du site actif

Seuls les résidus du site actif de la protéine ont été considérés pour la réalisation de la dynamique moléculaire. La détermination des acides aminés constituant la poche de fixation du ligand a été menée grâce au programme CASTp (Computed Atlas of Surface Topography of proteins). CASTp calcule toutes les poches présentes au sein de la protéine, dont le site de fixation du ligand.[152] Le programme renseigne sur le volume, la surface et la constitution atomique des différentes cavités. Le principe est d'utiliser la sphère de Connolly et Richards comme objet capable d'accéder à toutes les surfaces de la protéine.

b) Minimisations et simulations de dynamique moléculaire

Les études de minimisation d'énergie ainsi que la dynamique moléculaire ont été effectuées par le programme PMEMD[153] basé sur le programme AMBER. Une boite parallélépipédique d'eau a été choisie comme solvant explicite dans les différents systèmes élaborés. De même, le champ de force parm98 a servi à simuler le site de liaison du ligand des PPARγ. La neutralité du système a été établie par l'ajout de cinq ions sodium. L'algorithme SHAKE[154] a été utilisé afin de pouvoir négliger les fluctuations des liaisons impliquant un atome d'hydrogène. La méthode PME[155,156] (Particle Mesh Ewald, méthode d'approximation de la sommation d'Ewald) a été appliquée lors de toutes les dynamiques moléculaires afin de limiter le temps de calcul des interactions électrostatiques et approximer au mieux leur valeur.

c) Protocole des dynamiques moléculaires

Les dynamiques moléculaires ont toutes été réalisées selon le protocole suivant:

---

[152] Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design *Protein Science* **1998**, *7*, 1884-1897.
[153] Duke, R.E.; Pedersen, L.G. PMEMD 3, University of North Carolina-Chapel Hill, **2003**.
[154] Ryckaert, J.P.; Cicootti, G.; Berendsen, H.J.C. Numerical integration of the Cartesian equation of motion of a system with constraints: molecular dynamics of N-alkanes *J. Comput. Phys.* **1977**, *23*, 327-341.
[155] Darden, T.; Toukmaji, A.; Pedersen, L.G. Long-range electrostatic effects in biomolecular simulations *J. Chim. Phys.* **1997**, *94*, 1346-1364.
[156] Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

i) Minimisation d'énergie de 2000 cycles. La méthode de la plus grande pente («steepest descent») a été utilisée pour effectuer les 100 premiers cycles, avant que la méthode du gradient conjugué ne soit ensuite utilisée.
ii) Chauffage par palier de 50K entre 0K et 250K avec de fortes contraintes sur la position des ions et des atomes de la protéine. Chaque étape a une durée de 3ps.
iii) Trois étapes à 250K pour une diminution progressive de la contrainte appliquée aux ions. Chacune de ces étapes a duré 3ps.
iiii) Une étape à 275K de 6ps.
iiiii) Sept étapes à 300K. Les deux premières étapes ont une durée de 12ps, et les cinq suivantes une durée de 6ps. Les contraintes appliquées aux atomes de la protéine sont progressivement réduites jusqu'à être nulles à la fin de cette période d'équilibration.
iiiiii) Production de 10 à 12ns de dynamique moléculaire sans contrainte.

Les étapes 2 à 5 de ce protocole ont été réalisées en utilisant l'ensemble thermodynamique canonique (NVT, Nombre d'atomes, Volume et Température constant). L'étape 6 de production a été effectuée dans l'ensemble NPT (Nombre d'atomes, Pression et Température constants).

d) Calcul du RMSD

Les différentes conformations issues de la dynamique sont analysées et comparées entre elles par des superpositions et des calculs de déviations (Root Mean Square Deviation: RMSD) (Equation 10).

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\overrightarrow{r_{i1}} - \overrightarrow{r_{i2}}\right)^2}$$

Équation 10. Expression du RMSD

Avec:

N le nombre d'atomes du système étudié

$\overrightarrow{r_{i1}}$ le vecteur position de l'atome i dans la structure 1

$\overrightarrow{r_{i2}}$ le vecteur position de l'atome i dans la structure 2

Plus le RMSD est important, plus les structures 1 et 2 ont des géométries spatiales éloignées.

e) Cartes RMSD

Des cartes de RMSD ont été générées afin de visualiser les régions les plus stables de la protéine (Figure 62). Afin de simplifier le contexte, nous avons établi la carte RMSD issue de la superposition de 20 conformères:

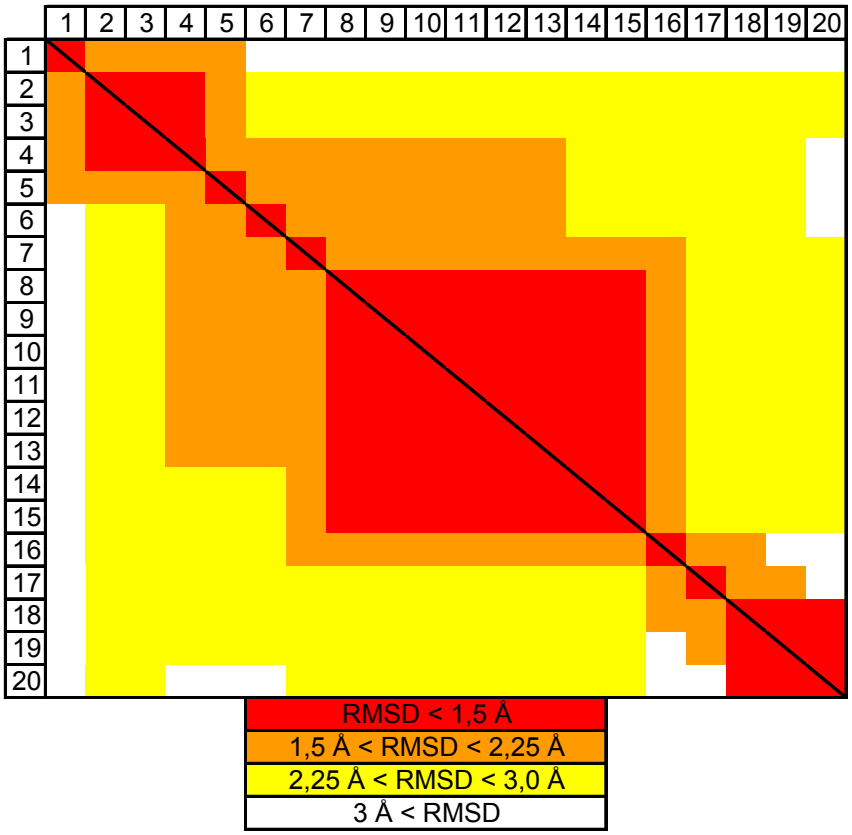


Figure 62. Exemple de carte RMSD pour 20 conformères

Différentes régions apparaissent sur cette carte. En simplifiant, il y a trois grandes régions (ou clusters en rouge) dans lesquelles le RMSD est inférieur à 1,5 Å. Cela signifie que la protéine est dans une zone de stabilité dans laquelle la structure protéique varie peu. C'est particulièrement le cas des conformères 2 à 4, 8 à 15 et enfin 18 à 20. Entre ces clusters, des zones de transition sont également observables. Les conformères 1, 5, 6, 7, 16 et 17 sont des états de transition de la protéine. L'échantillonnage d'un individu par cluster permet d'être représentatif de la mobilité de la protéine. Nous pourrions sélectionner arbitrairement les

conformères 3 et 19 pour les deux petits clusters et 8 et 15 comme deux individus opposés du grand cluster du milieu.

3. Couplage de la dynamique moléculaire au docking

Prendre en considération la dynamique de la protéine avant ou pendant l'étape de docking est devenu essentiel, surtout pour des entités sujettes à la flexibilité.[157] Dans notre étude, le principe du couplage docking/dynamique moléculaire est d'utiliser une structure «X», issue de l'expérience (par exemple par cristallographie) pour docker une chimiothèque (par exemple, 1 million de produits) dans le site actif. Une fois réduite en taille, la chimiothèque à analyser devient alors «X-like» (100 000 meilleurs produits par exemple), c'est-à-dire que l'on sait que toutes les molécules sont capables d'entrer dans le site actif de «X». La deuxième étape consiste à appliquer un protocole de dynamique moléculaire à la structure cristallographique «X» et d'en déduire un jeu de conformères «X_1, X_2, …, X_{20}» (selon l'exemple pris de la carte RMSD). La dernière étape est l'utilisation des conformères les plus prometteurs, énergétiquement viables mais également les plus divers possible. Les 100 000 molécules dites «X-like» sont alors arrimées dans chaque conformère sélectionné (par exemple X_3, X_9, X_{14}, X_{19}). D'une part, nous avons conservé les conformères provenant d'une méthode *in silico*. Nous en avons fait de même pour les structures issues de l'expérience (Y et Z). Ces deux types d'information (*in silico* et expérience) sont différents et sont donc susceptibles d'apporter une information plus variée que si nous avions simplement considéré la structure cristallographique «X» (Figure 63).[158]

[157] Alonso, H.; Bliznyuk, A.A.; Gready, J.E. Combining docking and molecular dynamic simulations in drug design Med. Res. Rev. **2006**, *26*, 531-568.
[158] Wong, C.F.; Kua, J.; Zhang, Y.; Straatsma, T.P.; McCammon, J.A. Molecular Docking of Balanol to Dynamics Snapshots of Protein Kinase A *Proteins* **2005**, *61*, 850-858.

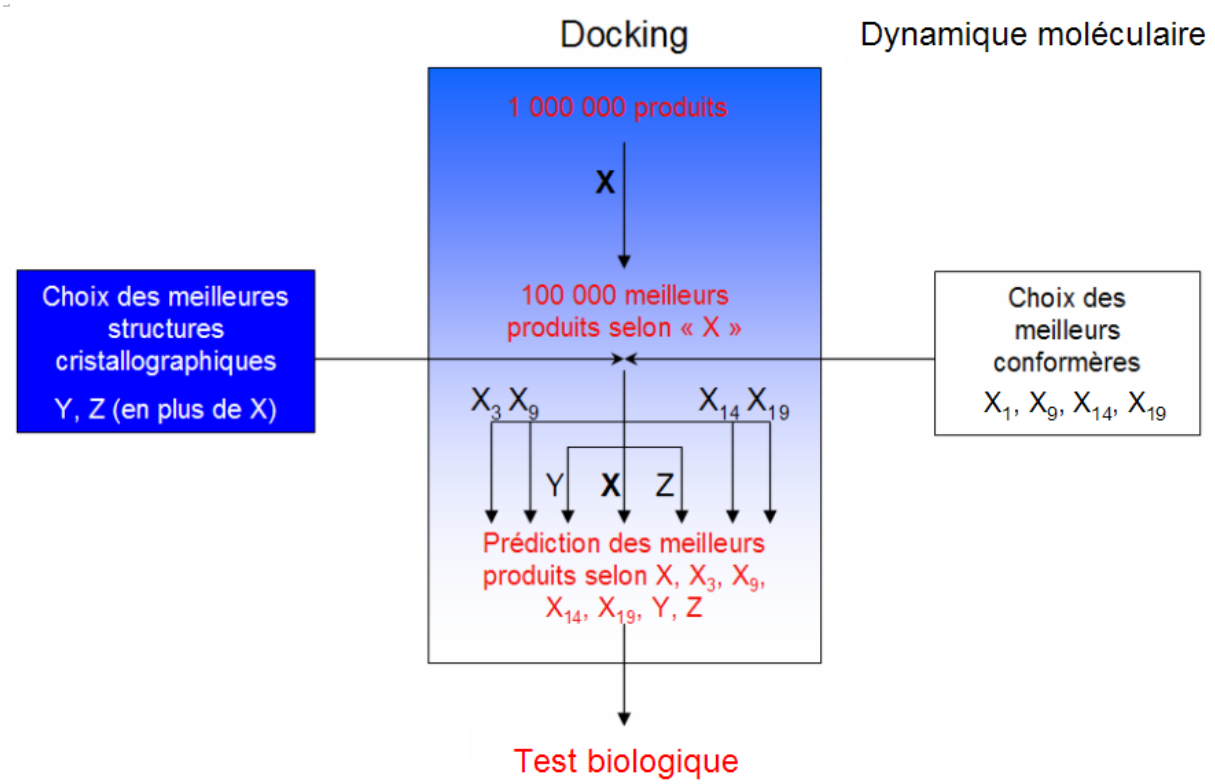


Figure 63. Protocole de couplage de la dynamique moléculaire avec le docking

Par cette méthodologie, nous sommes ainsi capables de mesurer l'affinité des molécules les plus prometteuses sur des conformations diverses du récepteur et ainsi de prendre en compte partiellement les mouvements de la protéine, en un temps de calcul réduit.

4. Evaluation du mode de docking

a) Problématique

La dynamique moléculaire offre la possibilité de générer des milliers de conformères issus d'une structure cristallographique. Du fait que le ligand est retiré avant la simulation de dynamique moléculaire, le site actif de la protéine est susceptible de subir des changements de volume. Plus précisément, le site de fixation de certains conformères se verra diminué et donc plus restrictif d'un point de vue stérique. Par conséquent, une bonne partie des conformères de la protéine ne seront pas exploitables, car les molécules de l'*ensemble total* auront des difficultés à pénétrer dans la cavité. Dans un tel contexte, il est important de vérifier le mode de docking des produits de l'*ensemble total* et de s'assurer qu'une molécule active scorée

corresponde à une molécule dockée. Il peut arriver qu'un composé récolte un bon score alors qu'il n'est pas fixé dans le site actif. Il est indispensable d'éliminer ce cas de figure.

b) Le RMSD

La manière dont les ligands sont arrimés dans le site actif peut être généralement décrite numériquement par le RMSD entre le ligand co-cristallisé et la molécule organique dockée. Il est ainsi aisé de déterminer si un ligand se superpose avec le ligand co-cristallisé. Plus la valeur de RMSD se rapproche de 0, plus la molécule organique se superpose avec la référence co-cristallisée.

Toutefois, dans notre protocole de dynamique moléculaire, le ligand co-cristallisé a été retiré du site actif avant simulation. Nous ne pouvons donc pas utiliser les RMSD pour évaluer de manière quantitative le positionnement des premières poses de la molécule dockée dans la protéine.

c) Fitting-Score

Nous avons développé un outil, que nous avons appelé «Fitting-Score», capable de nous renseigner sur la manière dont le ligand est docké. Le programme est basé sur le calcul du centre de géométrie du site actif et de celui de chaque molécule analysée. La distance entre les deux points est alors évaluée. Si la distance inter-point est trop importante, les données issues des fonctions de scoring ne seront pas exploitables (cela signifierait que l'on interprèterait des scores de molécules qui ne sont pas dockées dans le site actif).

(1) Calcul

(a) Centre de masse

Le calcul du centre de masse de coordonnées X, Y, Z se fait en calculant la moyenne des x_i, y_i et z_i de chaque atome «i» de masse «m_i» du site actif «A» dont la masse molaire totale est M (Equation 11).

$$X = \frac{1}{M} \times \sum_i (m_i \times x_i)$$
$$Y = \frac{1}{M} \times \sum_i (m_i \times y_i)$$
$$Z = \frac{1}{M} \times \sum_i (m_i \times z_i)$$

Équation 11. Coordonnées du centre de masse

Avec cette méthode, il est possible de pondérer toute la structure ou simplement quelques points interactionnels du ligand et du site actif. Cette stratégie permet de déplacer les centres de masse dans des régions de plus fortes interactions. En plus de la position, l'orientation du ligand au sein du site actif peut être prise en compte.

(b) Centre de géométrie

Le calcul du centre de géométrie se fait en appliquant la formule du centre de masse et en remplaçant «m_i» par 1 et «M» par le nombre d'atomes présents dans l'entité.

(c) Distance inter-centres

La distance «d» doit alors être mesurée entre deux centres du même type (géométrie ou masse) (Equation 12).

$$d_{AB} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2 + (Z_A - Z_B)^2}$$

Équation 12. Distance entre le centre de géométrie du site actif «A» et celui de la molécule «B»

(2) Courbes de fréquence relative cumulée

Afin de visualiser la répartition des distances entre centres de géométrie, nous avons utilisé les fréquences relatives cumulées. L'abscisse représente la distance entre centres de géométrie et l'ordonnée la fréquence de ces distances. Nous avons estimé les fréquences de distance uniquement pour les molécules actives (Figure 64).

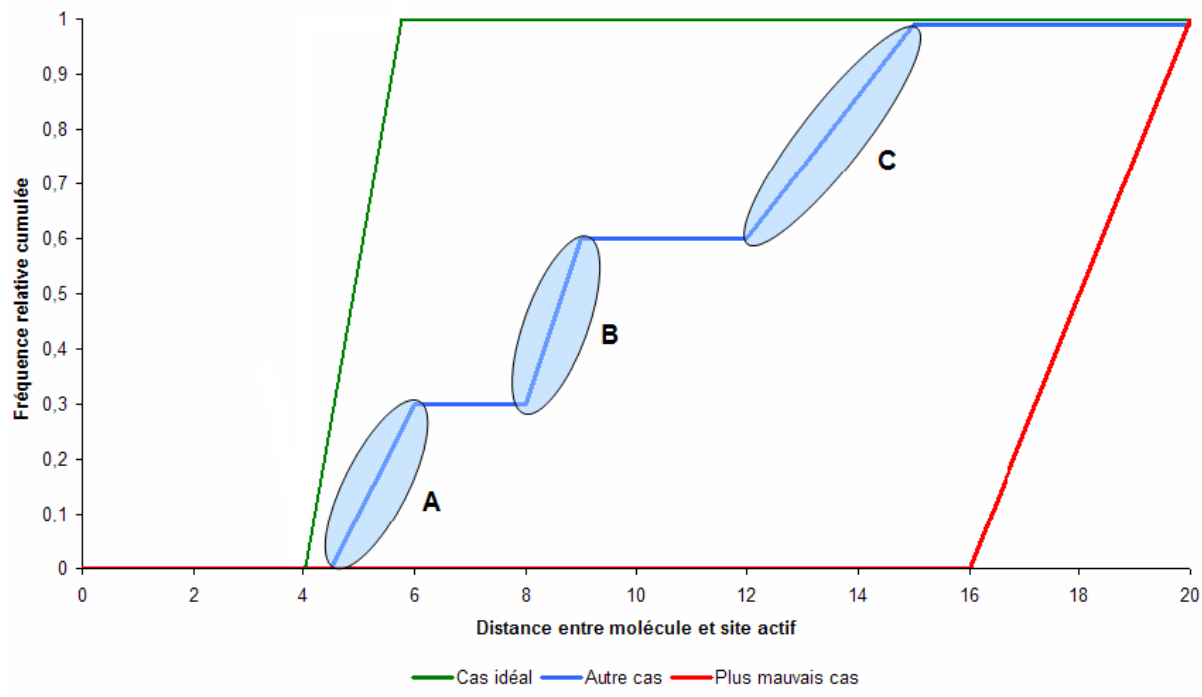


Figure 64. Illustration du fonctionnement des courbes de fréquence relative cumulée

Du fait que le centre de masse ne coïncide pas forcément avec le centre de masse du ligand, nous avons construit la courbe du cas idéal (en vert) à partir de 4 Å. Dans notre cas, cette distance est celle que l'on observe entre le centre de masse du site actif et celui du ligand co-cristallisé. Cette courbe représente le cas où toutes les molécules actives sont au même emplacement que le ligand co-cristallisé. La courbe rouge symbolise le plus mauvais cas de docking: toutes les molécules actives sont dockées hors du site actif. En effet, on peut imaginer qu'une distance supérieure à 10 Å entre le centre de la molécule et celui du site actif correspond à des molécules en dehors de la poche de fixation. Enfin, la courbe bleue montre un cas dans lequel des composés se fixent dans plusieurs régions (A et B) de la poche. Dans d'autres cas, le modèle est incapable de positionner les molécules dans la poche (C). On qualifiera d'«étalement» la tendance d'une courbe dont les produits actifs se fixent loin ou en dehors de la poche (courbe bleue).

d) Sphères de présence

Une sphère de présence est définie par un centre qui correspond au centre de masse ou de symétrie du site actif. Le rayon de la sphère est égal à la distance mesurée entre le centre de géométrie ou de masse d'une molécule de référence et le site actif.

K. Résultats et discussions

1. Comparaison et choix des structures cristallographiques

Afin d'évaluer la variabilité spatiale du site actif, nous avons procédé au calcul des volumes accessibles dans les différentes structures cristallographiques disponibles dans la PDB.

a) Volume du site actif des PPARγ

Les volumes des sites actifs des formes cristallographiques ont été mesurés. Au maximum, cinq sites ont été identifiés dans certaines formes de PPARγ (Figure 65).

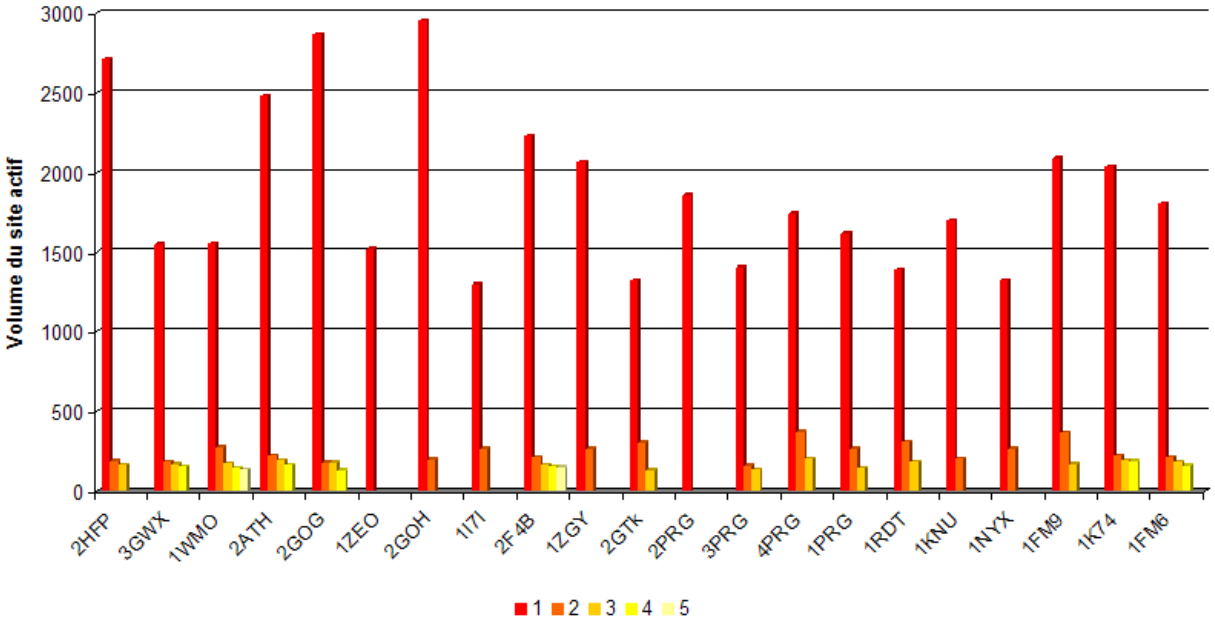


Figure 65. Volumes des sites actifs des différentes structures cristallographiques de PPARγ

Le volume moyen mesuré des structures cristallographiques est de 1880 $Å^3$. Il apparaît clairement une variabilité du nombre (de 1 à 5) et des volumes (de 1300 à 3000 $Å^3$) du site actif des PPARγ. Une adaptation de la cavité en fonction du volume du ligand co-cristallisé est alors à envisager.

b) Evaluation du docking

(1) Fréquences relatives cumulées

Ici encore, seules les molécules actives sont concernées par cette analyse. En effet, nous ne nous occupons pas des molécules inactives dans ce type d'analyse puisque même si une molécule inactive est bien dockée, nous espérons la discriminer de la catégorie des actives lors de l'étape de scoring.

Nous avons mesuré la distance entre chaque molécule active positionnée dans le site actif. Nous avons calculé le Fitting-Score pour chaque molécule active dans chacune des structures cristallographiques. Nous en avons déduit les courbes de fréquences relatives cumulées qui illustrent la distance (en Angström) entre le centre de géométrie de chacune des molécules actives et celui du site actif. Cette étude a été faite en dockant l'ensemble des produits actifs sur chaque structure cristallographique (21 formes) (Figure 66).

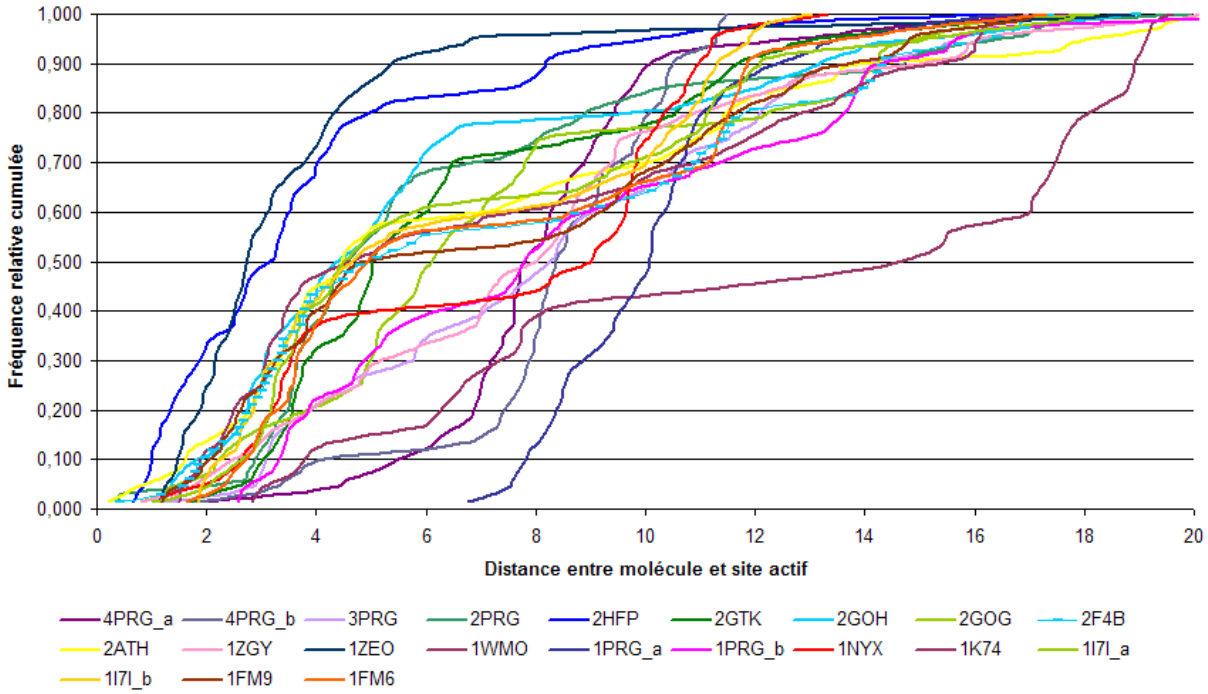


Figure 66. Fréquence relative cumulée pour chaque structure cristallisée

Les courbes de fréquence relative cumulée ci-dessus représentent les distances entre chaque molécule active dockée et le site actif de chaque structure cristallisée.

Les courbes correspondant aux structures 1WMO et 1PRG_a montrent l'existence d'une divergence entre le centre de géométrie des deux entités (molécule et site actif). Par ailleurs, la tendance de la courbe de la structure 1WMO à s'étaler montre que les molécules

actives se fixent un peu n'importe où, dans et à l'extérieur du site actif (elle est répartie entre 2 et 19 Å). Par conséquent, on peut imaginer que le degré de signification des scores correspondant à ces structures sera faible et que les résultats provenant de celles-ci seront erronés. Les autres structures ont un comportement quasiment similaire: elles possèdent toutes un point d'inflexion d'ordonnée 0,5 compris entre 2 et 6 Å d'abscisse. Nous avons isolé les structures cristallographiques pour lesquelles le point d'inflexion s'étale de 1 à 6 Å (Figure 67).

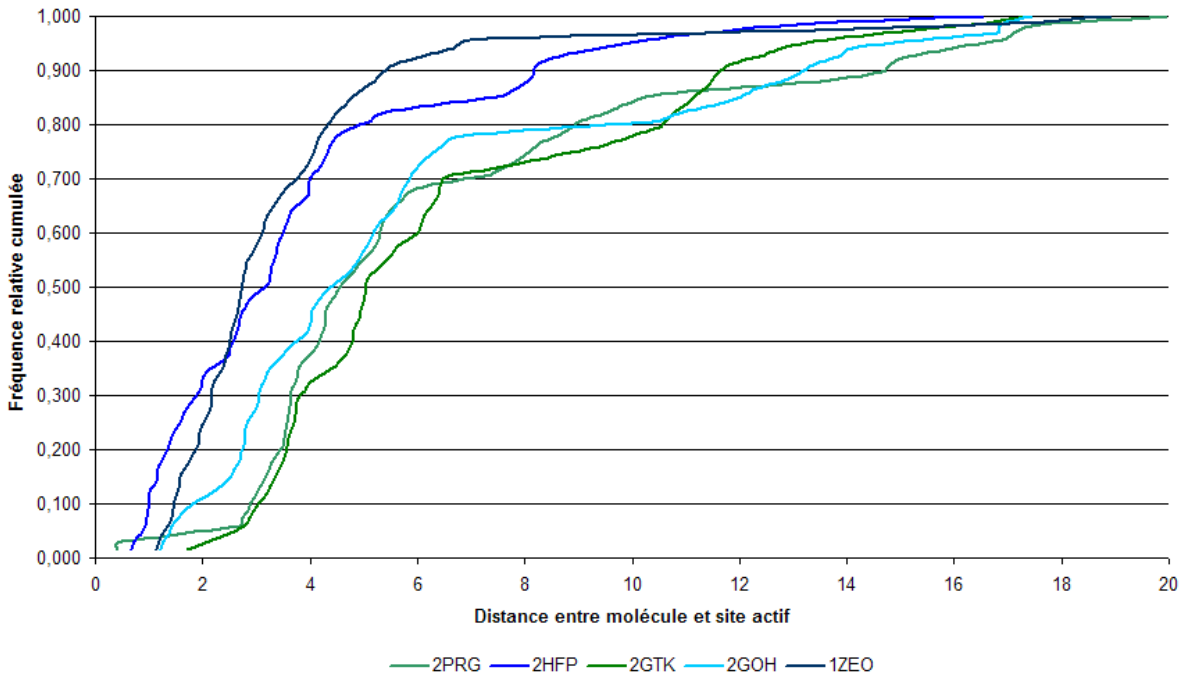


Figure 67. Comparaison du docking des molécules actives dans les meilleures structures cristallographiques

La tendance de chacune de ces courbes montre que le docking des composés actifs se fait aux alentours de 4 à 6 Å (en accord avec la sphère de présence de rayon 3,2 ± 2 Å).

(2) Superposition des molécules actives dans deux cas extrêmes selon Fitting-Score

Nous avons superposé toutes les molécules actives après docking dans la forme cristallographique 2PRG afin d'observer le taux d'occupation du site actif (Figure 68).

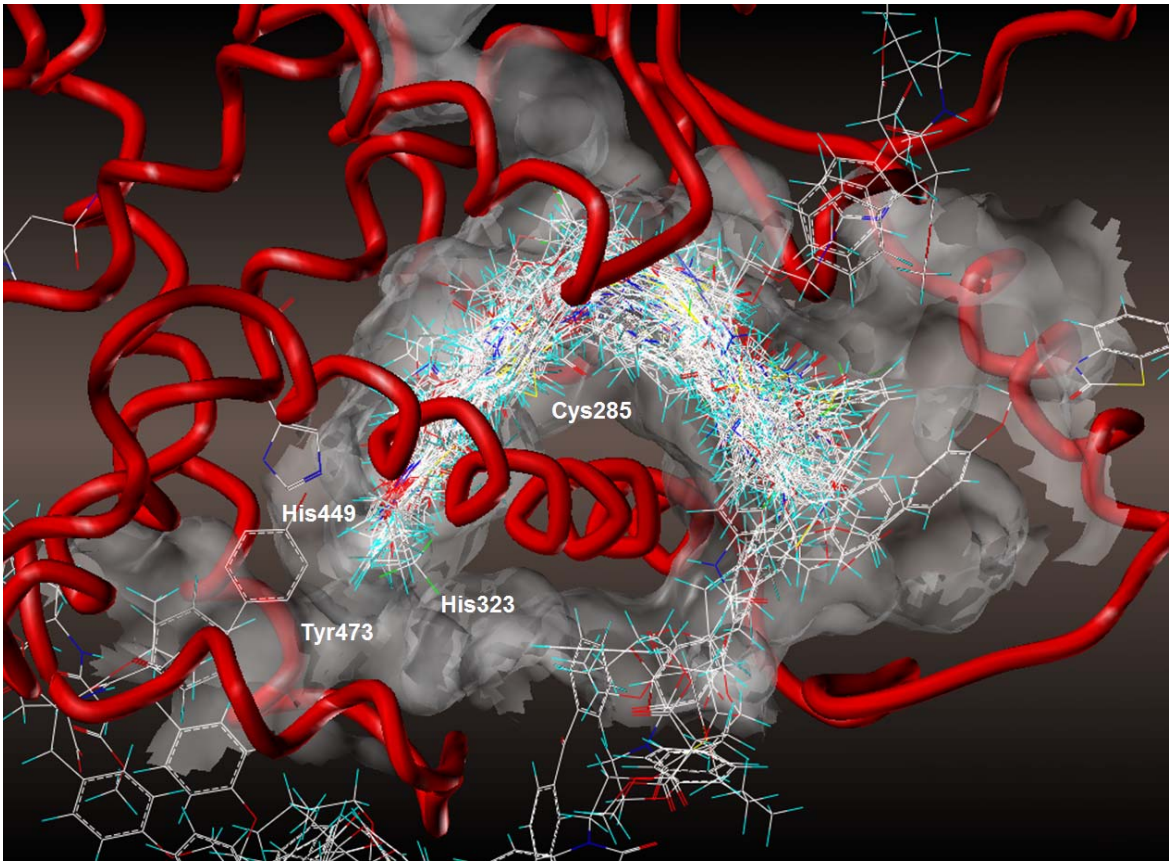


Figure 68. Superposition des molécules actives dans le site actif de 2PRG

Dans la structure 2PRG, les molécules se répartissent majoritairement dans le site actif. Seules quelques molécules sont dockées hors de la poche. Ces composés apparaissent à la traîne de la courbe de fréquence relative cumulée entre 8 et 16 Å.

Nous avons réalisé la même superposition pour la forme cristallographique la moins performante (1WMO) en termes d'étalement le long de la courbe de fréquence relative cumulée (Figure 69).

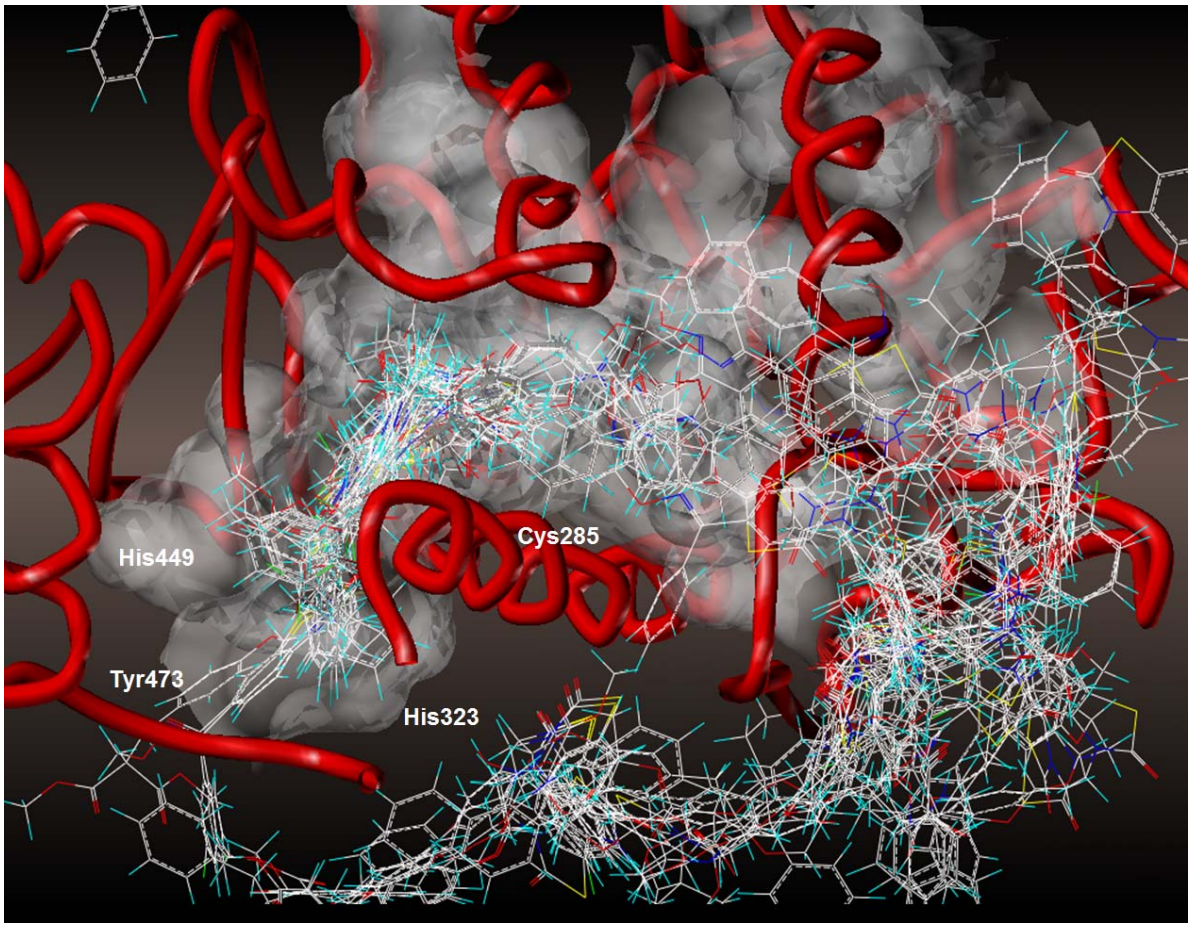


Figure 69. Superposition des molécules actives dans le site actif de 1WMO

Dans le cas 1WMO, la courbe des fréquences relatives cumulées est en accord avec la répartition des superpositions de molécules actives dans le site actif. En conclusion, la distribution du Fitting-Score est en accord avec les réalités du docking et nous permet de juger de la viabilité des résultats proposés par les scores.

c) Evaluation des scores

(1) Test de similarité des fonctions de scoring

Les cinq fonctions de scoring ont été soumises au test de similarité permettant de déterminer si les données provenant des fonctions de scoring sont multicolinéaires. Pour cela, nous avons calculé la matrice de similarité par les coefficients de corrélation au sein des données issues de la structure cristallographique 2PRG (Tableau 14).

	Fitting score	F-Score	D-Score	PMF	G-Score	Chemscore
Fitting score	1,00	0,35	0,57	0,48	0,45	0,58
F-Score		1,00	0,37	0,22	0,21	0,51
D-Score			1,00	0,42	0,80	0,89
PMF				1,00	0,35	0,45
G-Score					1,00	0,73
Chemscore						1,00

Tableau 14. Matrice de corrélation entre les différentes fonctions et aussi Fitting-Score

Les deux fonctions de scoring les plus corrélées entre elles sont D-Score et Chemscore avec une corrélation de 0,89. Même si ces deux fonctions ont des tendances similaires, chacune apporte une information spécifique.

(2) Choix de la première pose

Comme nous l'avons montré précédemment, le choix de la première pose est primordial. Une première étude a été menée avec l'*ensemble total*, en utilisant F-Score pour choisir la première pose. Finalement, celle-ci a été re-scorée par les quatre fonctions restantes de C-Score. Nous avons appliqué ce processus à toutes les structures cristallographiques disponibles dans la PDB que nous avons évaluées par le R_s (Figure 70).

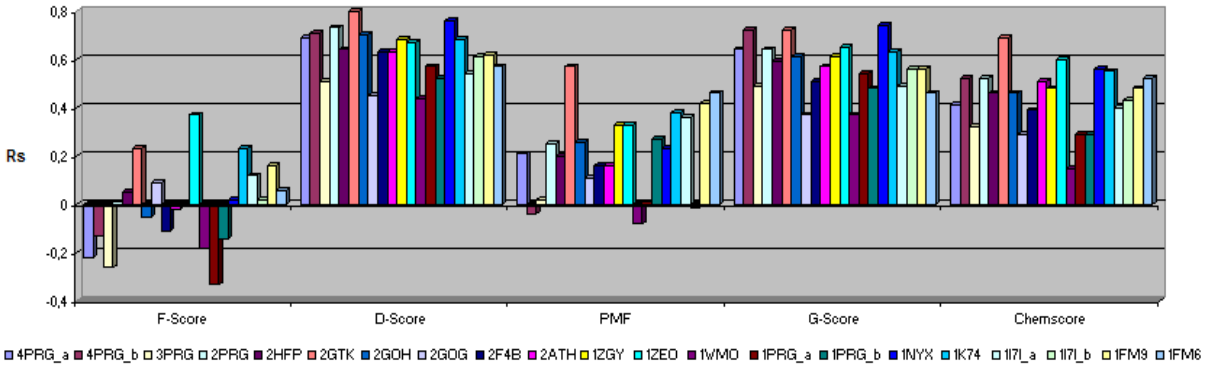


Figure 70. Re-scoring de la première pose choisie par F-Score

La figure ci-dessus illustre les disparités de performance entre les cinq fonctions de re-scoring (dont F-Score, utilisée pour le choix de la première pose). Nous observons également que chaque structure cristallographique se comporte de manière irrégulière vis-à-vis des molécules de l'*ensemble total*. D'une manière générale, la fonction F-Score est celle dont le

R_s est le plus faible. La fonction D-Score est celle qui fournit, *a priori*, les meilleurs résultats pour la majorité des structures cristallographiques.

Nous avons donc effectué la même procédure de re-scoring mais en utilisant D-Score pour choisir la première pose. Les résultats sont indiqués dans la Figure 71.

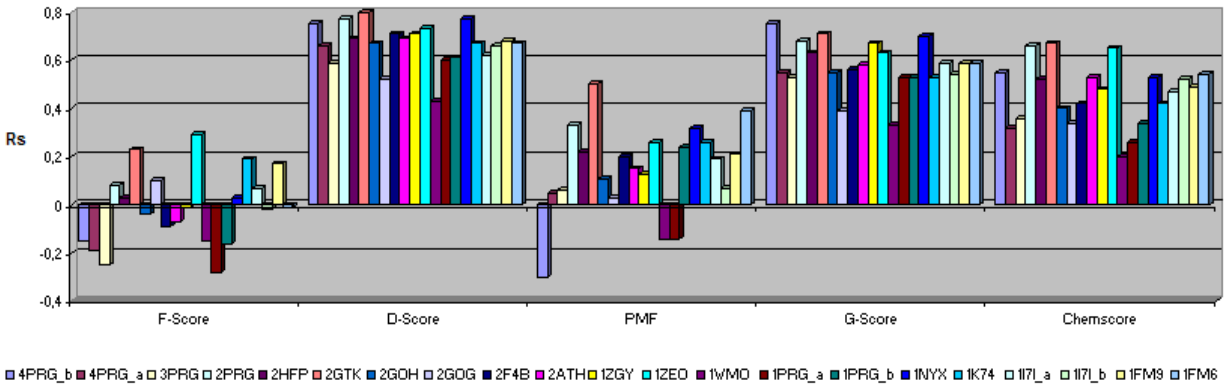


Figure 71. Re-scoring de la première pose choisie par D-Score

On ne note pas de différence fondamentale entre les résultats des deux approches. Les deux fonctions de scoring (F-Score et D-Score) possèdent des performances similaires pour choisir la première pose. Toutefois, du fait que F-Score est utilisée lors du processus de docking, nous avons préféré conserver cette fonction pour le choix de la première pose. Les figures 70 et 71 montrent également la puissance de trois fonctions de scoring dans le processus de re-scoring: D-Score, G-Score et Chemscore. Nous verrons dans la partie traitant du consensus de fonctions de scoring que ces fonctions offrent de bons résultats.

(3) Etape de re-scoring

(a) Calcul des vrais/faux positifs/négatifs

Comme décrit précédemment, les cinq fonctions de scoring de C-Score ont été appliquées à la première pose, préalablement choisie selon F-Score. Nous avons montré dans le chapitre traitant de la cible COX-2 que l'analyse factorielle discriminante (AFD) est la méthode de choix pour interpréter les données multidimensionnelles. Elle a donc été appliquée à la matrice dont les lignes correspondent aux molécules et les colonnes aux fonctions de scoring. La répartition des vrais/faux positifs/négatifs en a été déduite (Figure 72).

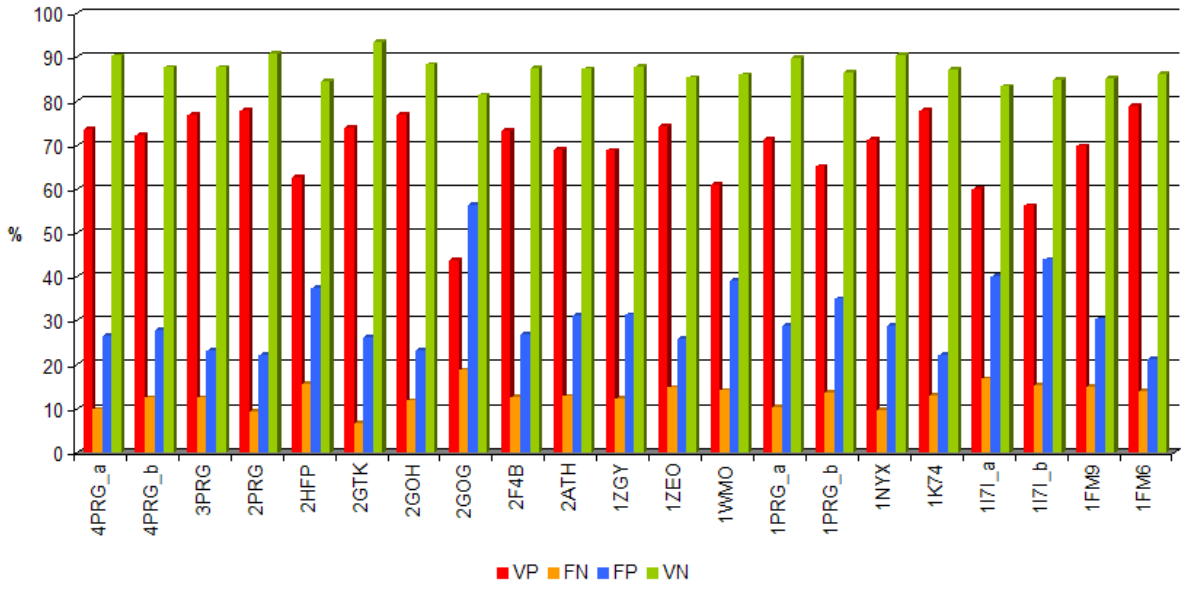


Figure 72. Evaluation de chacune des structures cristallographiques par les ratios VP, FN, FP, VN

Majoritairement, les vrais négatifs sont constants d'un modèle à l'autre et sont distribués autour de 85%. De même, les faux négatifs sont limités à une moyenne d'environ 15%. Le taux de vrais positifs est soumis à des variations plus amples (échelonnées de 42% pour 2GOG jusqu'à 79% pour 1FM6).

On peut choisir un modèle plutôt qu'un autre selon le ratio utilisé. En considérant le taux des vrais positifs, on choisirait 1FM6. En utilisant le taux de vrais négatifs (modèle capable de bien éliminer les composés biologiquement inactifs lors du criblage virtuel), la structure cristallographique 2GTK serait sélectionnée. Aux vues de la difficulté de choisir un modèle plutôt qu'un autre, il est essentiel de considérer d'autres métriques.

(b) Calcul du R_s

Nous avons ainsi utilisé le R_s afin d'évaluer le classement obtenu avec les différents modèles. Nous avons également évalué l'ajout de Fitting-Score, en tant que variable, dans le modèle multivarié (Figure 73).

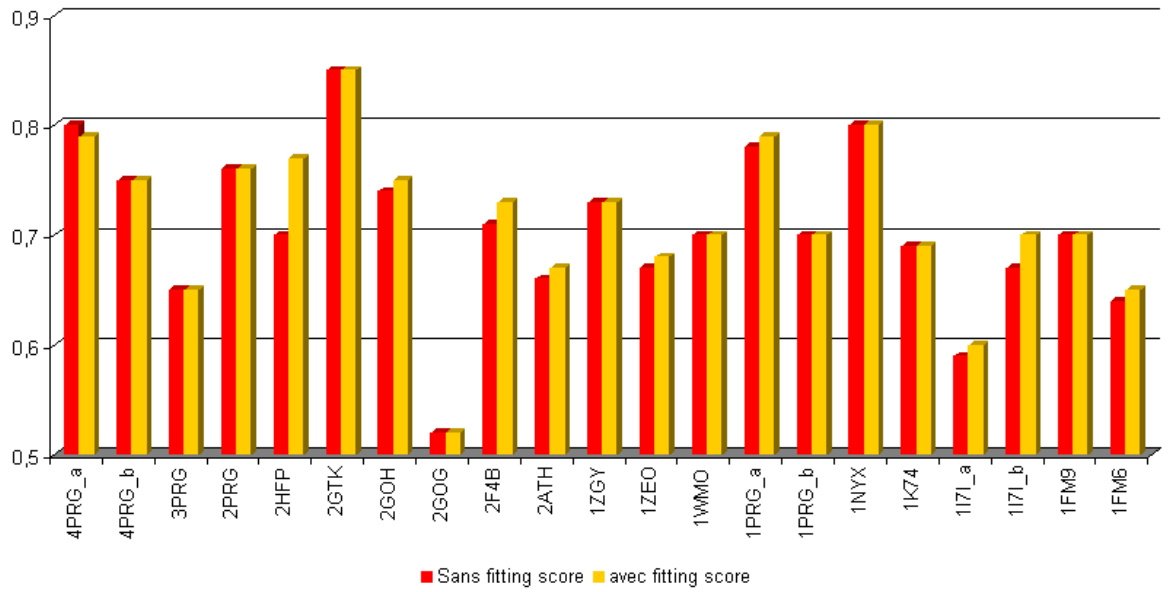


Figure 73. Evaluation de chacune des structures cristallographiques par le ratio R_s avec et sans Fitting-Score

L'histogramme d'évaluation des différentes structures cristallographiques montre des disparités. Les deux extrêmes sont obtenus pour les structures 2GOG et 2GTK. Les formes fournissant les meilleures réponses sont 2GTK, 4PRG_a, 1NYX, 1PRG_a et 2PRG. D'autre part, des variations sont obtenues entre des formes actives et inactives de structures (entre 4PRG_a et 4PRG_b, entre 1PRG_a et 1PRG_b, puis entre 1I7I_a et 1I7I_b). Dans le cas de 4PRG et 1PRG, la forme active des PPARγ répond mieux que la forme inactive. A l'inverse, l'état actif de la forme 1I7I donne de moins bons résultats que la forme inactive. Il est donc clairement difficile d'établir un lien entre l'état (actif ou inactif) du PPARγ et sa capacité à classer les molécules actives de l'*ensemble total*. La forme inactive offre un accès facilité au ligand (par déplacement de l'hélice H12) mais supprime un éventuel lien hydrogène entre un groupement accepteur du ligand et la fonction hydroxyle de Tyr473. Dans l'état actif de PPARγ, l'accès du ligand dans le site actif semble plus restreint que la forme inactive. Par contre, elle rend accessible au ligand la Tyr473.

Enfin, une amélioration du R_s dans la moitié des structures cristallographiques peut être obtenue en incluant les données provenant de Fitting-Score dans l'analyse factorielle discriminante.

Le nombre de molécules actives rejetées est également un critère fondamental de choix du modèle: il permet de ne garder que les modèles qui limitent la perte de composés actifs.

(c) Calcul du R_a

L'évaluation du nombre de molécules rejetées par le modèle a également été faite par le calcul du R_a avant et après ajout de Fitting-Score dans le modèle multivarié et est représentée sur l'histogramme de la figure 74.

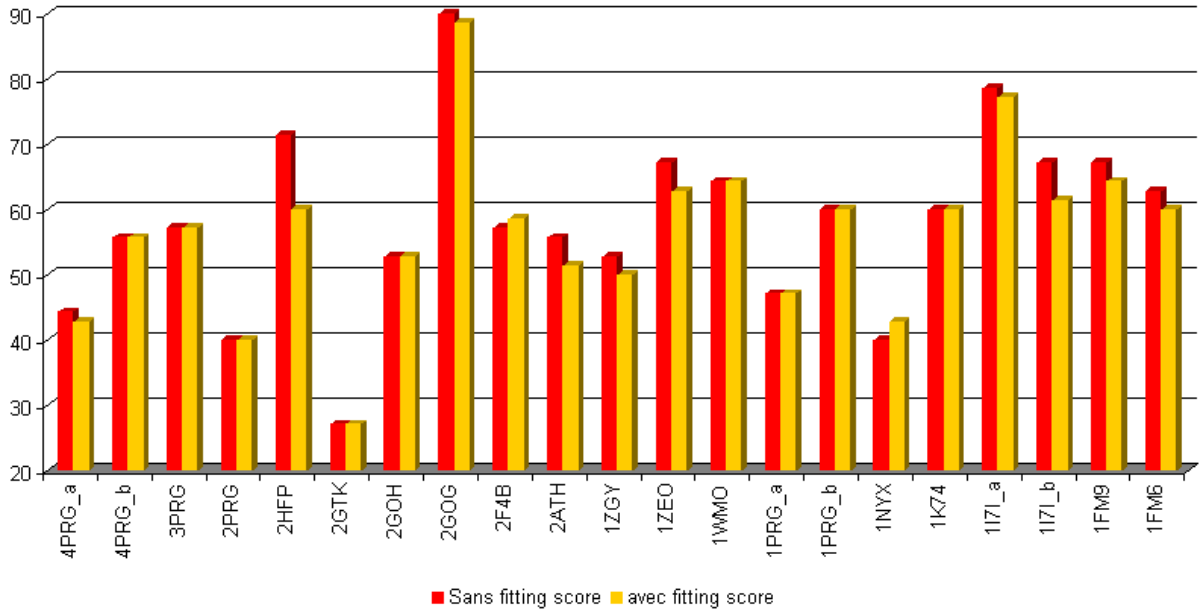


Figure 74. Evaluation de la proportion de molécules actives rejetées R_a par chacune des structures cristallographiques

La proportion de molécules rejetées est minimale dans le cas de 2GTK, 2PRG et 1NYX. La structure 2GOG est celle qui élimine le plus de composés actifs. L'introduction de Fitting-Score dans la matrice multidimensionnelle améliore nettement le taux de molécules éliminées par les modèles (particulièrement dans le cas de 2HFP). Ces résultats prouvent que Fitting-Score a une influence positive sur l'identification des molécules actives

(d) Calcul du E_r

L'histogramme suivant illustre la qualité de l'enrichissement pour chaque structure cristallographique (le maximum d'enrichissement se situe à 5,0) (Figure 75).

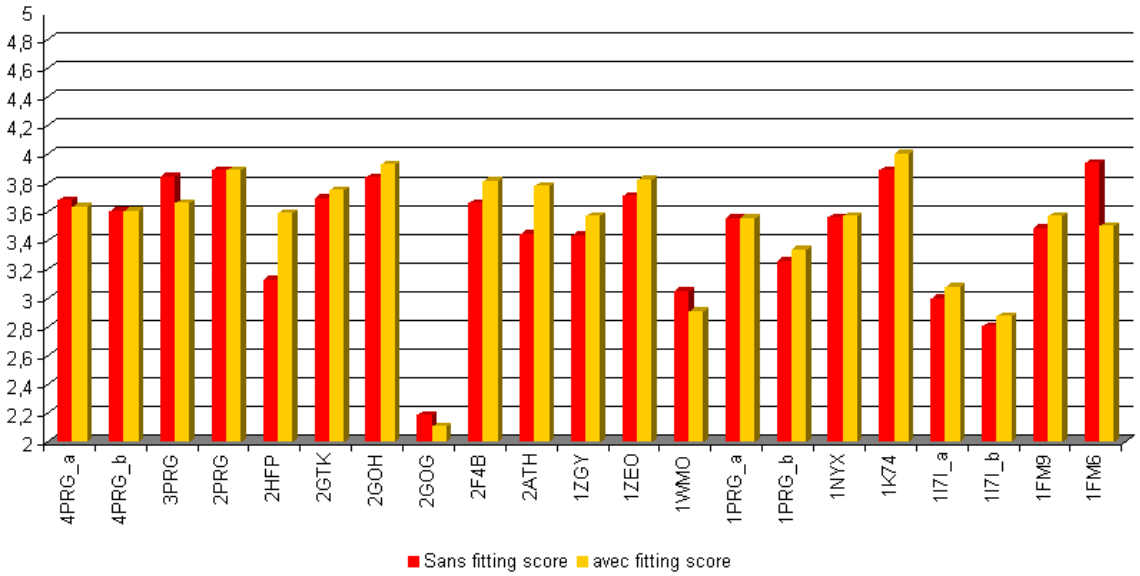


Figure 75. Facteur d'enrichissement E_r de chaque structure cristallographique

Ces résultats montrent que les ratios calculés précédemment (VP, VN, FP, FN, R_s et R_a) ne sont pas forcément en adéquation avec l'enrichissement E_r. En effet, alors que 2GTK récolte le meilleur R_s, il est classé en dessous des structures 1K74, 2PRG, et 2GOH. Un point commun à tous les indicateurs prédictifs est la faiblesse de prédiction de la structure 2GOG. Finalement, on observe une nette amélioration par l'ajout de Fitting-Score dans la majorité des cas.

(e) Conclusion des différents outils prédictifs

Les résultats des différents indicateurs privilégient certaines structures cristallographiques répertoriées dans le tableau 15. Elles seront utilisées lors du criblage virtuel.

Indicateurs	Limite	Structures sélectionnées
VP	≥ 75%	3PRG, 2PRG, 2GOH, 1K74, 1FM6
VN	≥ 90%	4PRG_a, 2PRG, 2GTK, 1PRG_a, 1NYX
R_s	≥ 0,75	4PRG_a, 4PRG_b, 2PRG, 2GTK, 1PRG_a, 1NYX
R_a	≤ 25%	2GTK
E_r	≥ 3,8	3PRG, 2PRG, 2GOH, 1K74, 1FM6

Tableau 15. Structures privilégiées après recoupement des différents indicateurs

Selon les indicateurs choisis, les structures sélectionnées ne sont pas toujours les mêmes. Malgré tout, des structures cristallographiques sont récurrentes telles que 2PRG,

2GTK (respectivement rosiglitazone et l'acide (2S)-3-(1-{[2-(2-chlorophenyl)-5-methyl-1,3-oxazol-4-yl]methyl}-1H-indol-5-yl)-2-ethanypropanoique). Ces deux ligands sont des agonistes. Nous avons décidé de conserver ces deux structures de par leurs performances prouvées par les différentes métriques employées. Nous avons également décidé d'ajouter 4PRG à ces deux structures et plus précisément 4PRG_a et 4PRG_b (respectivement la forme active et inactive de PPARγ). En effet, 4PRG est cristallisée avec un agoniste partiel (contrairement à la majorité des structures qui sont cristallisées avec un agoniste). L'intérêt est que ces structures apportent une information complémentaire, du fait qu'elles soient co-cristallisées avec un agent pharmacologiquement différent (GW0072), même si notre objectif est de travailler sur les agonistes.

(f) Courbes d'enrichissement

Même si le R_s fournit une bonne approximation de la tendance de la courbe d'enrichissement, nous avons souhaité observer graphiquement la manière dont les 70 composés actifs de l'*ensemble total* sont classés. Les courbes d'enrichissement des quatre structures cristallographiques 2PRG, 2GTK, 4PRG_a et 4PRG_b ont été représentées. La stratégie retenue pour ces représentations graphiques est l'analyse factorielle discriminante en présence des cinq fonctions de scoring de C-Score (Figure 76).

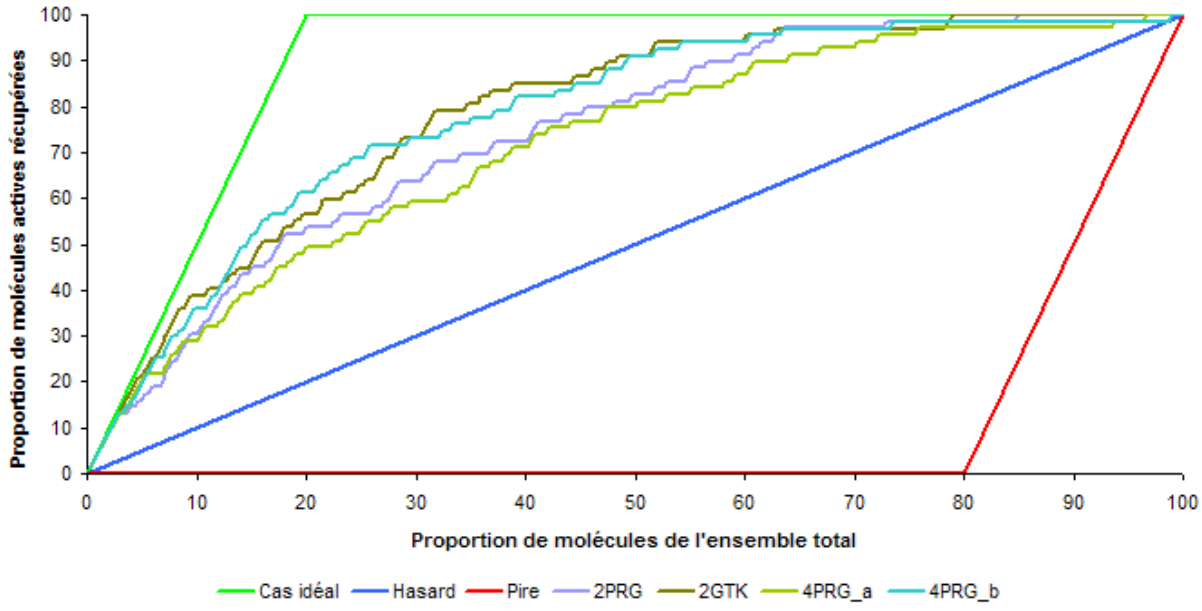


Figure 76. Courbes d'enrichissement des structures 2PRG, 2GTK, 4PRG_a et 4PRG_b

Les courbes d'enrichissement prouvent que le R_s est efficace pour évaluer quantitativement les courbes. Nous avons classé par ordre décroissant de performance: 2GTK > 4PRG_a > 2PRG > 4PRG_b. En effectuant un classement qualitatif (par inspection de la tendance générale des courbes) de ces structures on obtient la même tendance, ce qui prouve la concordance de deux méthodes l'une quantitative, l'autre qualitative.

(4) Consensus des fonctions de scoring

Bien que l'analyse factorielle discriminante soit une méthode puissante, nous avons quand même souhaité la comparer avec les stratégies de consensus précédemment évoquées. Le consensus des fonctions de scoring les plus performantes a été réalisé en tenant compte de l'amélioration des résultats après ajout des fonctions les unes après les autres (D: D-Score, G: G-Score, C: Chemscore, P: PMF, F: F-Score). Les deux fonctions les plus performantes (D-Score et G-Score) (Figure 70) ont été combinées au départ dans chaque stratégie de consensus. Les autres fonctions les moins pénalisantes ont été ajoutées au modèle. Différentes stratégies ont été utilisées pour déterminer les fonctions qui seront choisies pour le consensus.

(a) Stratégie «Rank by number»

La stratégie de consensus «Rank by number» a été testée sur l'*ensemble total*. A chaque ajout de fonctions de scoring, nous avons calculé le R_s (Figure 77):

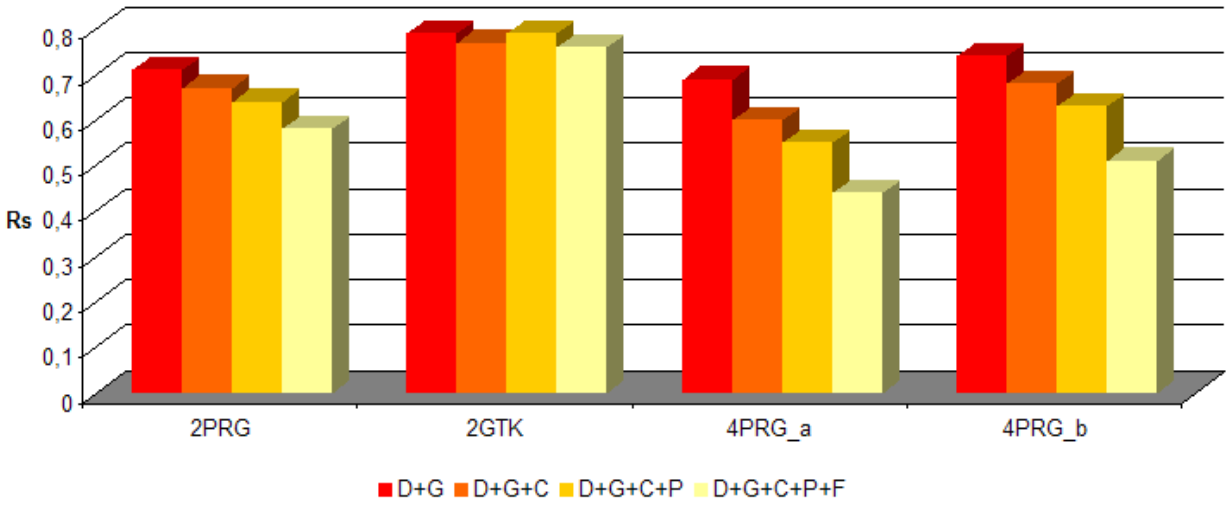


Figure 77. Résultats de la stratégie de consensus «Rank by Number»

La stratégie «Rank by number» est efficace dans le cas où le consensus est composé des deux fonctions de scoring D-Score et G-Score. L'ajout de chemscore, PMF puis F-Score dégrade le R_s. Une exception est observée dans le cas de la structure cristallographique 2GTK pour laquelle l'ajout de C dégrade très légèrement le R_s.

D-Score et G-Score fournissent la meilleure prédiction de notre *ensemble total*. La méthode «Rank by number» fonctionne particulièrement bien pour les quatre structures.

(b) Stratégie «Rank by best»

Un autre consensus appelé «Rank by best» a été appliqué à notre *ensemble total* (Figure 78).

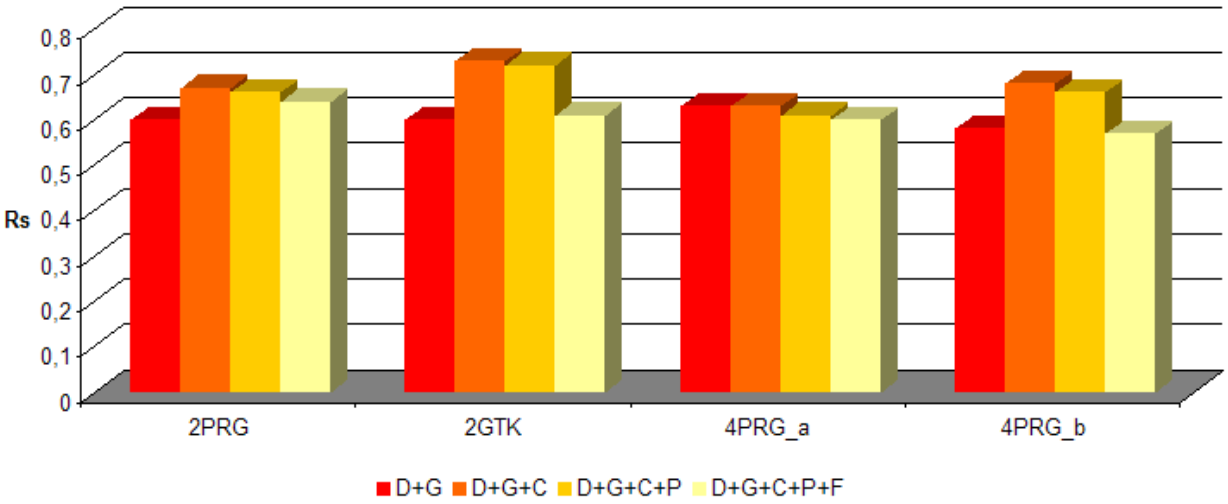


Figure 78. Résultats de la stratégie de consensus «Rank by best»

La stratégie «Rank by best» est moins performante que la technique de consensus «Rank by number». L'ajout de chemscore améliore considérablement le R_s dans le cas des structures 2PRG, 2GTK et 4PRG_b.

(c) Stratégie «Rank by rank»

La stratégie de consensus «Rank by rank» a également été envisagée (Figure 79).

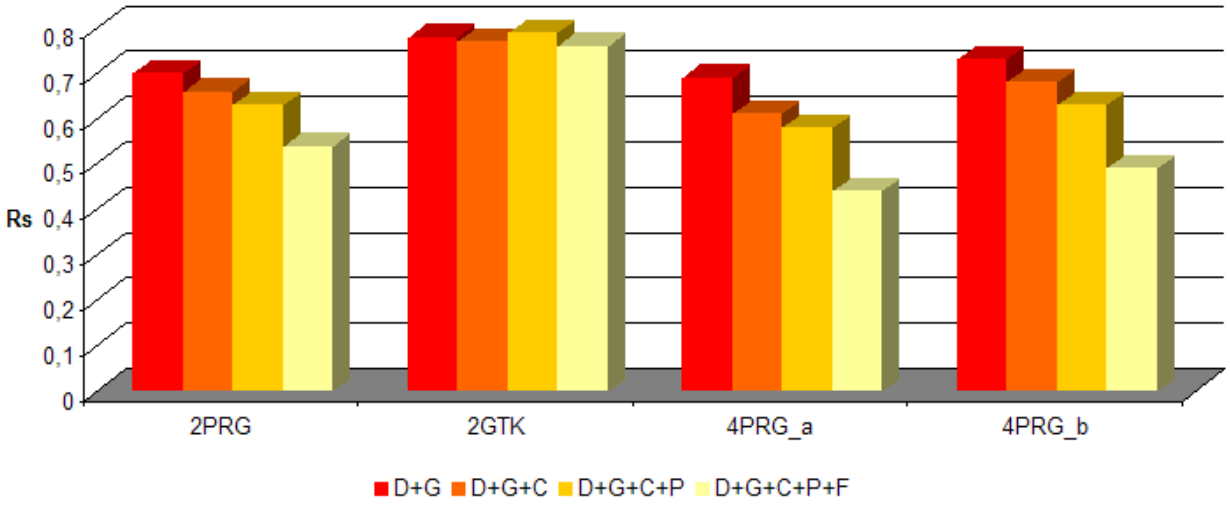


Figure 79. Résultats de la stratégie de consensus «Rank by rank»

Le comportement de la stratégie «Rank by rank» et «Rank by number» est très similaire. Malgré tout, nous préférerons utiliser la méthode «Rank by number» du fait qu'elle est basée sur le calcul d'une moyenne de score et non d'une moyenne de rang, susceptible d'introduire des approximations.

(5) Analyse factorielle discriminante

Les méthodes de consensus précédemment décrites sont évidemment applicables à une logique de criblage virtuel au sein de laquelle la récupération rapide de composés actifs est au cœur des préoccupations. Le calcul du R_s, selon la technique d'analyse factorielle discriminante, a déjà été effectué précédemment. Nous avons décomposé l'ajout des fonctions une à une pour évaluer et comparer l'incidence de certaines fonctions sur le modèle d'AFD (Figure 80).

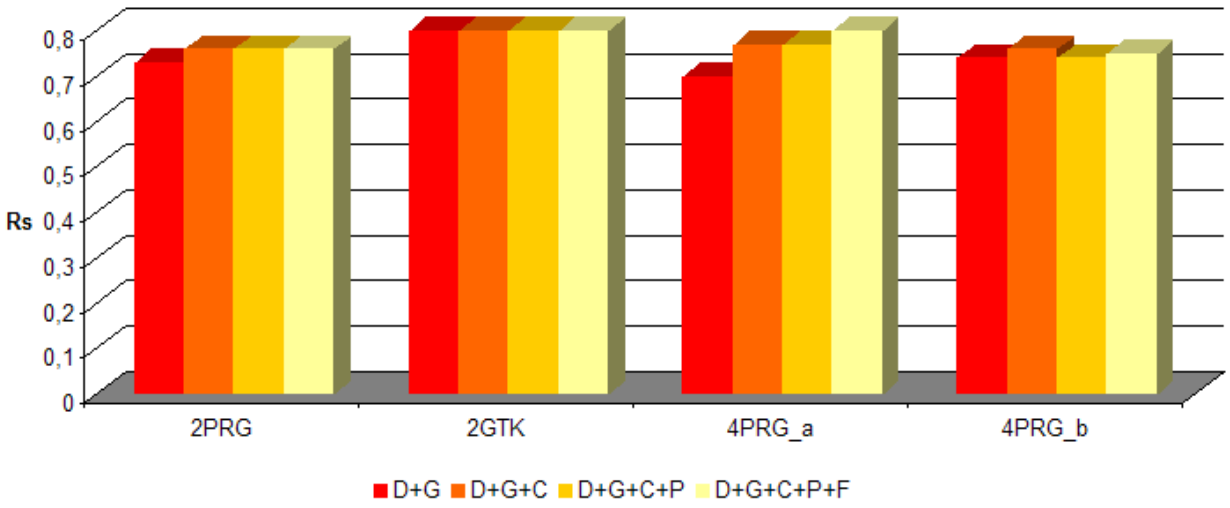


Figure 80. Résultats de l'utilisation de l'analyse factorielle discriminante «pas à pas»

L'ajout successif des différentes fonctions de scoring montre moins de variabilité que dans le cas des méthodes de consensus. En effet, la variation du R_s lors de l'ajout successif des fonctions de scoring est faible. Pour la forme 4PRG_a, nous prouvons que même les deux fonctions les plus performantes (D-Score+G-Score) ont un moins bon rendement que cinq réunies (D-Score+G-Score+Chemscore+PMF+F-Score). Le choix du nombre de fonctions de scoring dans le modèle de l'analyse factorielle discriminante semble être moins important que dans le cas des consensus.

L'AFD a l'énorme avantage de pouvoir exploiter toutes les fonctions de scoring sans se préoccuper de leurs performances.

(6) Comparaison des consensus avec l'analyse factorielle discriminante

Les meilleures stratégies de consensus ont été comparées aux résultats fournis par l'analyse factorielle discriminante (Figure 81).

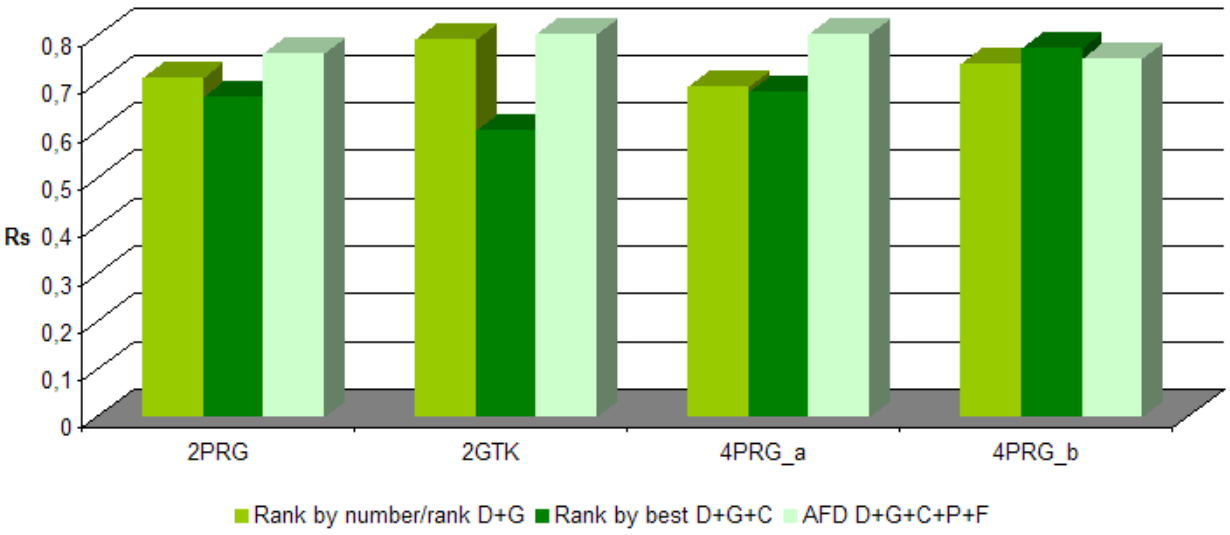


Figure 81. Comparaison des différentes méthodes de consensus à l'analyse factorielle discriminante

L'analyse factorielle discriminante est la méthode qui fournit les meilleurs R_s. A l'inverse, même si la méthode «Rank by number» présente de bons résultats, sa composition en fonctions de scoring doit être optimisée afin de parvenir à des résultats proches de l'analyse factorielle discriminante.

(7) Validation de l'AFD

Un point clé dans l'utilisation de l'analyse factorielle discriminante est la validation de l'expérience par l'utilisation d'un *ensemble de test*. En effet, le modèle est préalablement optimisé et évalué sur un *ensemble d'entraînement*. Les taux de substitution sur les données ($E_{entraînement}$ et E_{test}) ont été calculés.

Nous avons employé la même stratégie que dans la partie traitant de la cible COX-2. Au total, 10 jeux *entraînement/test* ont été élaborés et testés. Nous n'avons conservé que les meilleures itérations pour chaque structure cristallographique (Figure 82).

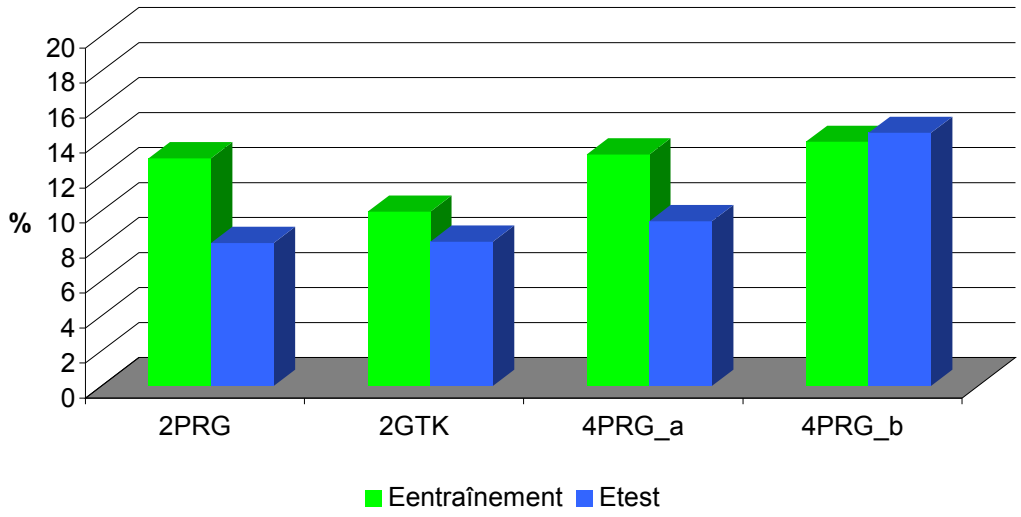


Figure 82. Taux d'erreur sur les échantillons d'apprentissage et de test

Les écarts entre l'*ensemble d'entraînement* et l'*ensemble de test* sont faibles ce qui prouve la robustesse du modèle. Les trois structures 2PRG, 2GTK et 4PRG_a sont validées avec un taux d'erreur estimé moyen de 8,5%. La structure 4PRG_b possède un taux d'erreur estimé de 14,5% qui est supérieur au taux d'erreur apparent. Nous ne validerons pas cette structure et ne sera donc pas utilisée dans le processus de criblage virtuel.

La prochaine étape de ces travaux est l'utilisation de la dynamique moléculaire pour prendre en compte ponctuellement la flexibilité de la protéine PPARγ.

2. Dynamique moléculaire

a) Simulation de 4PRG_a

La molécule GW0072 est le ligand le plus volumineux étudié dans le cas des PPARγ. Sa caractéristique principale est qu'il est doté d'une flexibilité importante. Ce modèle a donc été utilisé dans un premier temps pour évaluer les variations de RMSD et du volume du site actif. En l'absence du ligand, la poche peut se rétracter, engendrant une forte diminution de volume. La dynamique a été réalisée sur le site de liaison du PPARγ seul pendant 11,52 ns.

(1) Carte RMSD

Un total de 12 000 conformations a été généré par dynamique moléculaire. Cette simulation représente 12 ns (une structure générée par picoseconde). Des cartes de RMSD ont été générées afin de visualiser les régions les plus stables de la protéine. Seules 1 200 structures ont été conservées pour dresser ces cartes (une structure sur dix).

La carte des RMSD de 4PRG_a (sans ligand) a été dressée. Le RMSD moyen des éléments structuraux de la poche (hélice α, feuillet β) s'élève à 2,15 Å tandis que le RMSD moyen des résidus de la poche de fixation est de 2,43 Å (Figure 83).

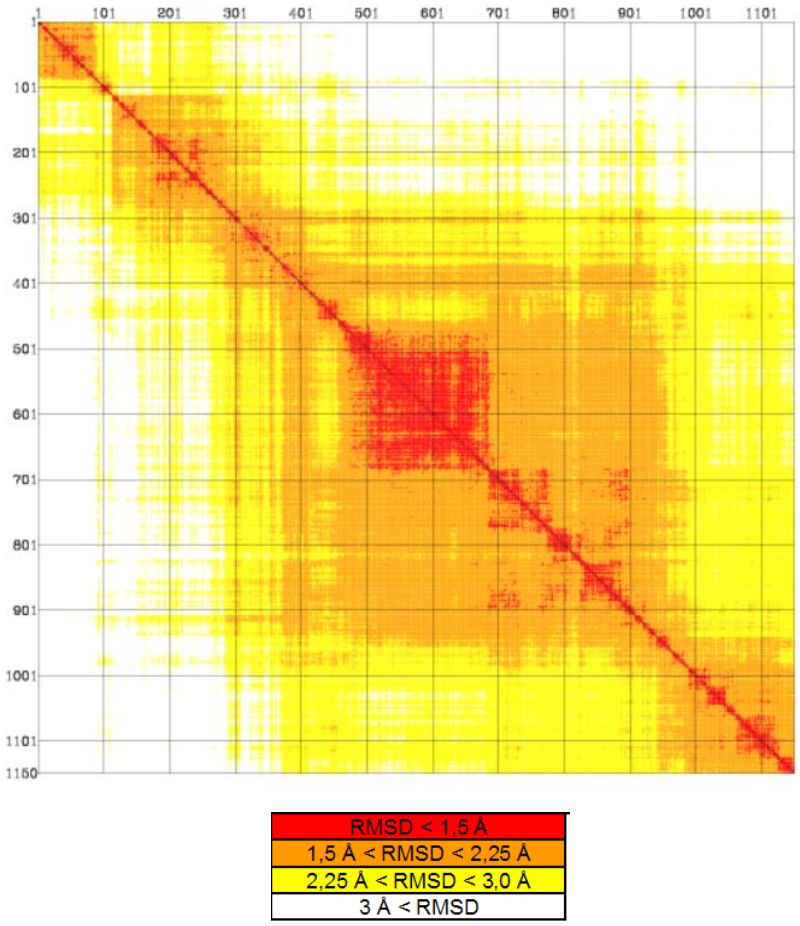


Figure 83. Carte RMSD des conformations de 4PRG sans son ligand

La figure ci-dessus montre des zones de stabilité (en rouge) au sein desquelles certaines structures conformationnelles stagnent. Nous avons étudié le contenu de chacune de ces régions. Nous en avons extrait des conformations significatives. Une étude par un protocole de docking-scoring du pouvoir prédictif de chacune des structures pourrait

permettre d'envisager l'utilisation de ces conformations en complément des structures cristallographiques déjà existantes.

(2) Calcul du volume du site actif

Le volume de la poche a également été déterminé au cours de cette dynamique toutes les 640ps. Le volume initial de la poche était d'environ 2200 $Å^3$ au départ de la simulation et diminue entre 790 et 2100 $Å^3$. Ainsi, même si la structure varie peu dans sa globalité, le volume subit de grandes variations (Figure 84).

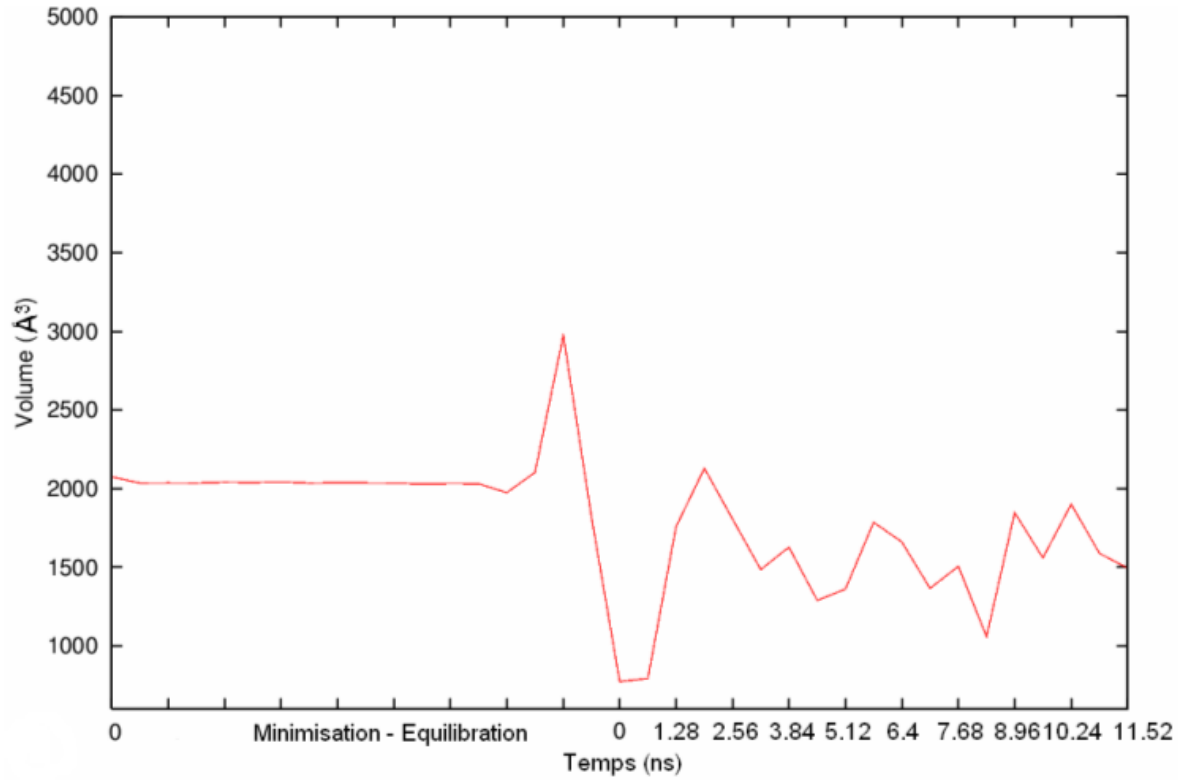


Figure 84. Variation du volume de 4PRG après retrait du ligand GW0072

Après retrait du ligand, la poche fait preuve d'adaptabilité par variation de son volume interne. A un moment de la dynamique, une conformation de la protéine est capable de fixer des molécules qui n'auraient eu aucune chance dans la structure cristallographique. A l'inverse, la structure cristalline aura peut-être la capacité de lier une famille de molécule qui n'aura aucune affinité pour les structures issues de la dynamique.

L'utilisation de la structure 4PRG dans cette étude a permis de mettre en évidence la flexibilité de certaines zones du site actif de PPARγ. Malgré tout, nous n'avons pas utilisé les conformères issus de 4PRG_a.

b) Simulation de 2PRG

Nous avons étudié une structure cristallographique co-cristallisée avec la rosiglitazone, agoniste des PPARγ (code PDB: 2PRG). Cette structure a subi le même protocole de simulation que 4PRG_a. Nous aurions également pu utiliser la structure 2GTK qui possède un ligand intéressant pour nos études. Toutefois, dans le fichier PDB, la structure secondaire interrompue à certains endroits du récepteur peut gêner la simulation par dynamique moléculaire.

(1) Carte RMSD

Le retrait de la rosiglitazone a été effectué avant d'initier la dynamique moléculaire. Au total, 12 000 structures ont été extraites, représentant un total de 12 ns. La carte des RMSD a été générée à partir de 1 200 structures (une sur dix a été récupérée) et des points d'extraction de conformations ont été repérés (A, B, C,…,H) (Figure 85).

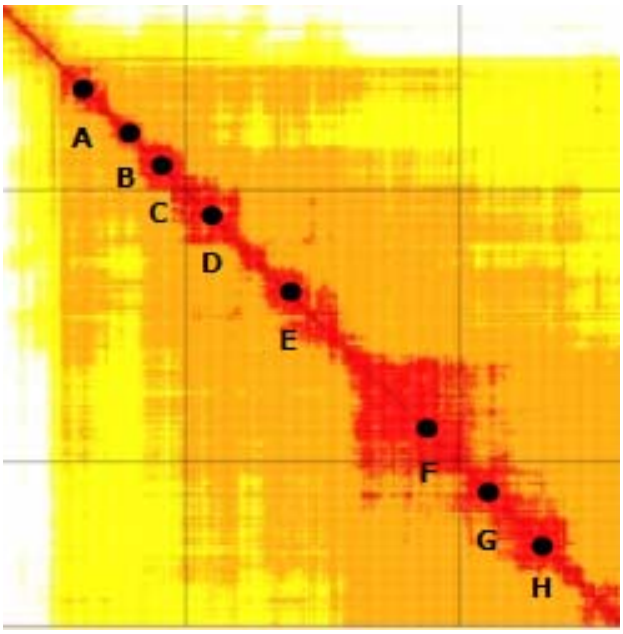


Figure 85. Carte RMSD des conformations de 2PRG

Chaque point correspond à la récupération d'une structure issue de la dynamique moléculaire. Les conformations seront évaluées par notre protocole de docking-scoring en choisissant Fitting-Score pour évaluer la qualité des modes de docking. L'analyse factorielle discriminante a été appliquée pour combiner les fonctions de scoring.

(2) Evaluation du Fitting-Score

Nous avons souhaité évaluer les modes de docking de ces structures issues de la dynamique. Le Fitting-Score a été appliqué à chacune des molécules positionnées dans le récepteur. Les courbes de fréquence relative cumulée ont été évaluées à partir de ces données (Figure 86).

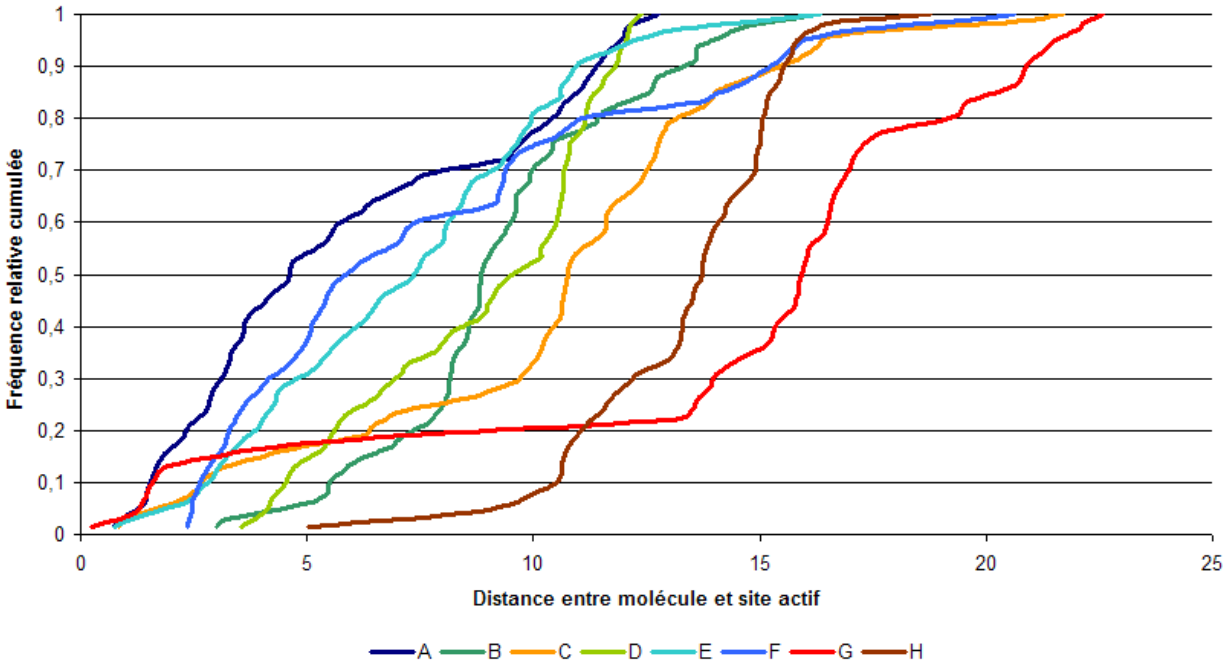


Figure 86. Courbes de fréquence relative cumulée de Fitting-Score selon les différentes conformations

Les courbes de fréquence relative cumulée des différentes conformations provenant de la dynamique moléculaire montrent des irrégularités. En effet, les conformations C, G et plus particulièrement H ont tendance à mal docker les composés actifs avec des distance moyennes de l'ordre de 10 à 15 Å. Par conséquent, toutes les structures issues de la dynamique moléculaire ne seront pas viables pour une utilisation dans un modèle prédictif d'affinité. Les

autres structures ne posent *a priori* pas de problème de docking. Ainsi, les conformations A, B, D, E et F sont exploitables pour les tests de scoring.

(3) Calcul du R_s et du R_a

Nous avons calculé le R_s et le R_a pour chaque conformation (Figure 87).

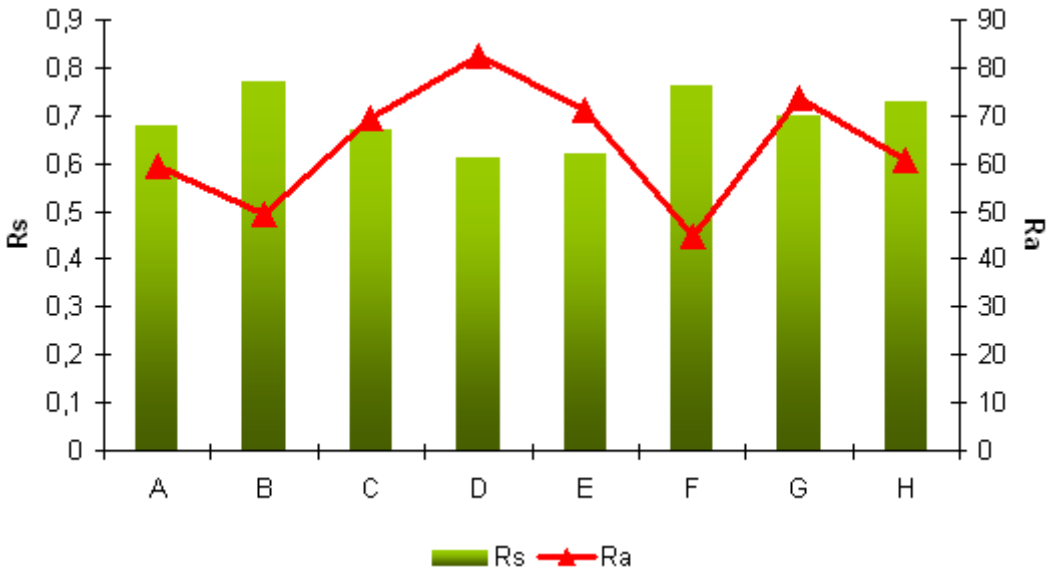


Figure 87. Répartition du R_s et du R_a par conformation de la structure 2PRG

Toutes les conformations génèrent des R_s s'échelonnant de 0,6 à 0,8. Toutefois, du fait que C, G et H ne répondent pas aux critères de Fitting-Score relaté sur les courbes de fréquence relative cumulée, elles ont été retirées du lot de conformères exploitables.

Nous souhaitons récupérer les conformères pour lesquels le $R_s \geq 0,6$ et $R_a \leq 70\%$. Le R_a met en évidence un nombre important de molécules actives rejetées par tous les modèles. Par contre, le R_s prouve les bonnes qualités de chaque modèle à classer les molécules actives.

Nous n'avons conservé que les structures A, B, E et F car leurs ratio R_s et R_a sont en accord avec la limite que nous avons fixé préalablement (D n'est pas sélectionné car il possède un trop fort R_a). Les fortes variations du R_s et du R_a mettent en évidence le fait que chaque conformère traite l'*ensemble total* d'une manière différente. Cela signifie que le volume de la cavité et le positionnement des acides aminés changent. Ainsi, certaines familles de l'*ensemble total* sont dockées différemment d'une conformation de la protéine à l'autre. Il se peut qu'une famille soit bien dockée avec le conformère A et ne le soit pas avec le

conformère E. Cette diversité de mode de fixation est une excellente manière de considérer la flexibilité de la cavité.

(4) Courbes d'enrichissement des meilleurs conformères

L'enrichissement des structures issues de la dynamique moléculaire a été évalué qualitativement par les courbes d'enrichissement (Figure 88).

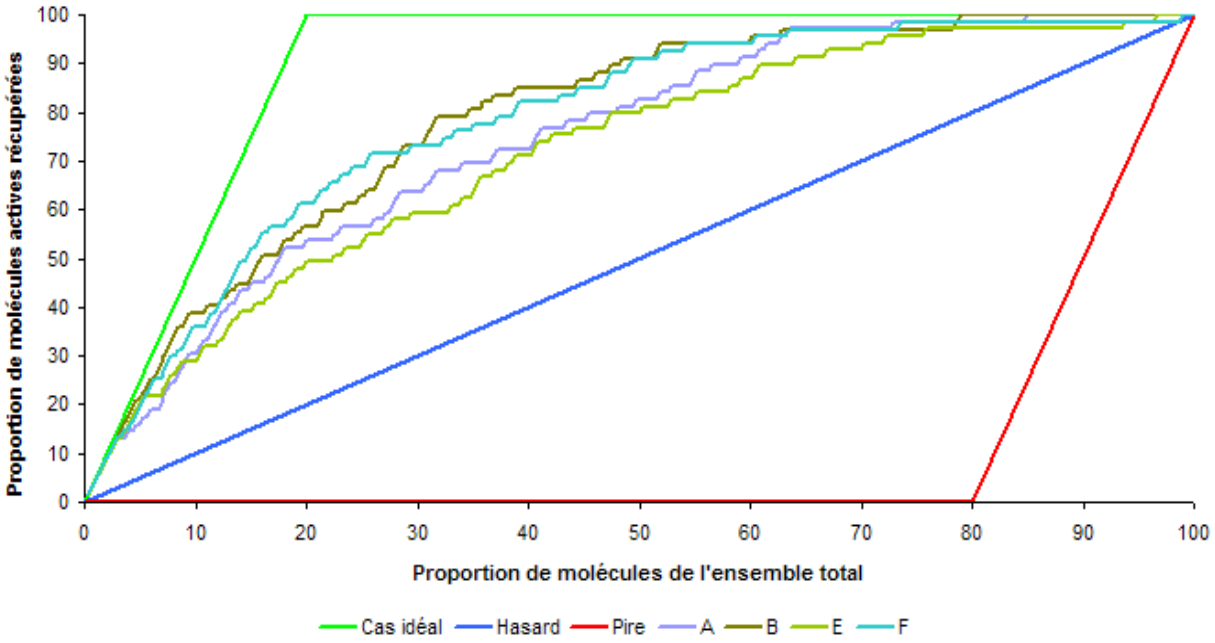


Figure 88. Courbes d'enrichissement des conformères A, B, E et F

Comme dans le cas des structures cristallographiques, les courbes d'enrichissement sont en accord avec le R_s précédemment calculé qui classe par ordre décroissant de performance les conformères: B > F > A > E.

(5) Consensus de conformères et de structures cristallographiques

Une nouvelle approche, dans l'utilisation des données des quatre conformations issues de la dynamique moléculaire, consiste à combiner les structures A, B, E et F auxquelles nous avons ajouté 2GTK, 2PRG ainsi que 4PRG_a (donc 7 structures au total). Ce regroupement de structures PPARγ (issues de la dynamique moléculaire et de la cristallographie) peut être

assimilé à un consensus. Celui-ci a été supervisé par des méthodes de consensus «classique» et d'analyse factorielle discriminante. Les lignes de la matrice de données sont les composés de l'*ensemble total* et les colonnes sont les scores provenant des quatre conformations sélectionnées.

(6) Evaluation des stratégies de consensus et d'AFD

(a) Par le R_s

Nous avons choisi d'utiliser les techniques «Rank by number», «Rank by best» et l'analyse factorielle discriminante pour évaluer les variations de R_s au sein de chaque modèle (Figure 89).

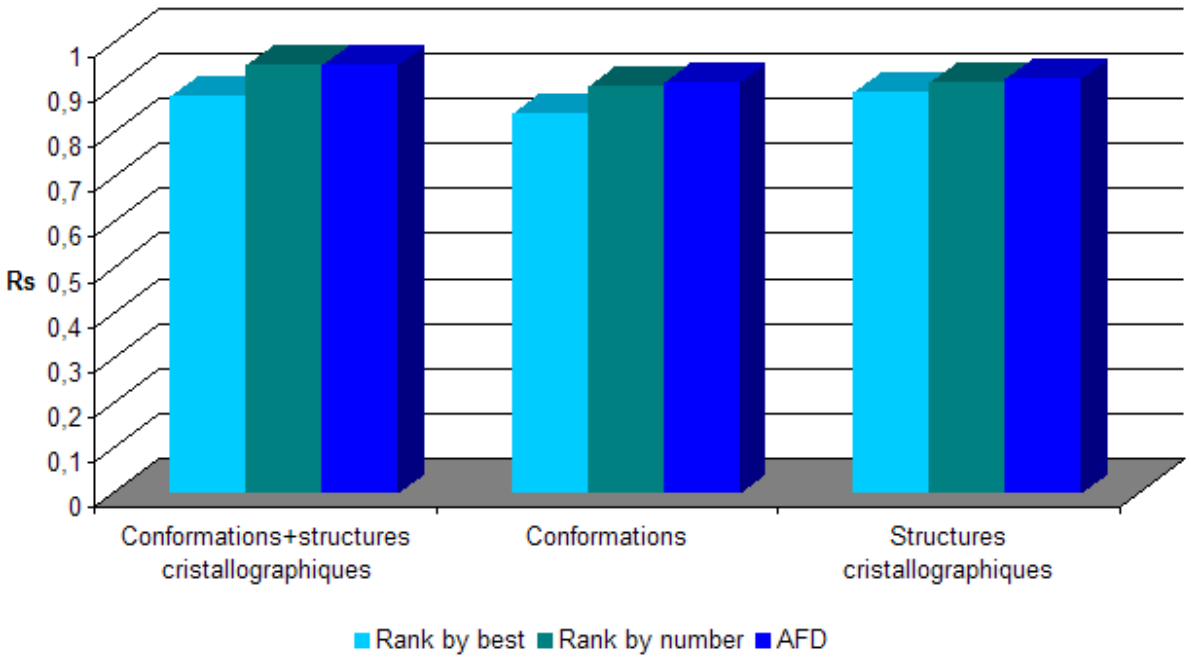


Figure 89. R_s selon les stratégies de consensus et analyses factorielles appliquées aux conformations issues de dynamique moléculaire et des structures cristallographiques

La variation du R_s est en faveur des techniques d'analyse factorielle discriminante ainsi que de la méthode de consensus «Rank by number». La stratégie «Rank by best» demeure moins performante que «Rank by number» et l'AFD, mais reste quand même une méthode capable d'améliorer les résultats. En effet, comparativement aux structures des conformations et des structures cristallographiques prises individuellement, le R_s est

fortement augmenté par l'utilisation du consensus et des méthodes d'analyse de données multivariées.

Par ailleurs, le pouvoir de récupération des composés actifs présents dans l'*ensemble total* peut varier en fonction des structures utilisées. Si l'on utilise les conformations d'un côté et les structures cristallographiques de l'autre, le R_s sera toujours inférieur à celui correspondant à toutes les structures (conformations + structures cristallographiques) regroupées ensemble. Il y a donc un intérêt à garder les deux types d'information: celle provenant de la dynamique moléculaire et celle de la cristallographie.

(b) Par le R_a

Le R_a a également été calculé afin de mesurer le nombre de composés actifs perdus par les différents modèles (Figure 90).

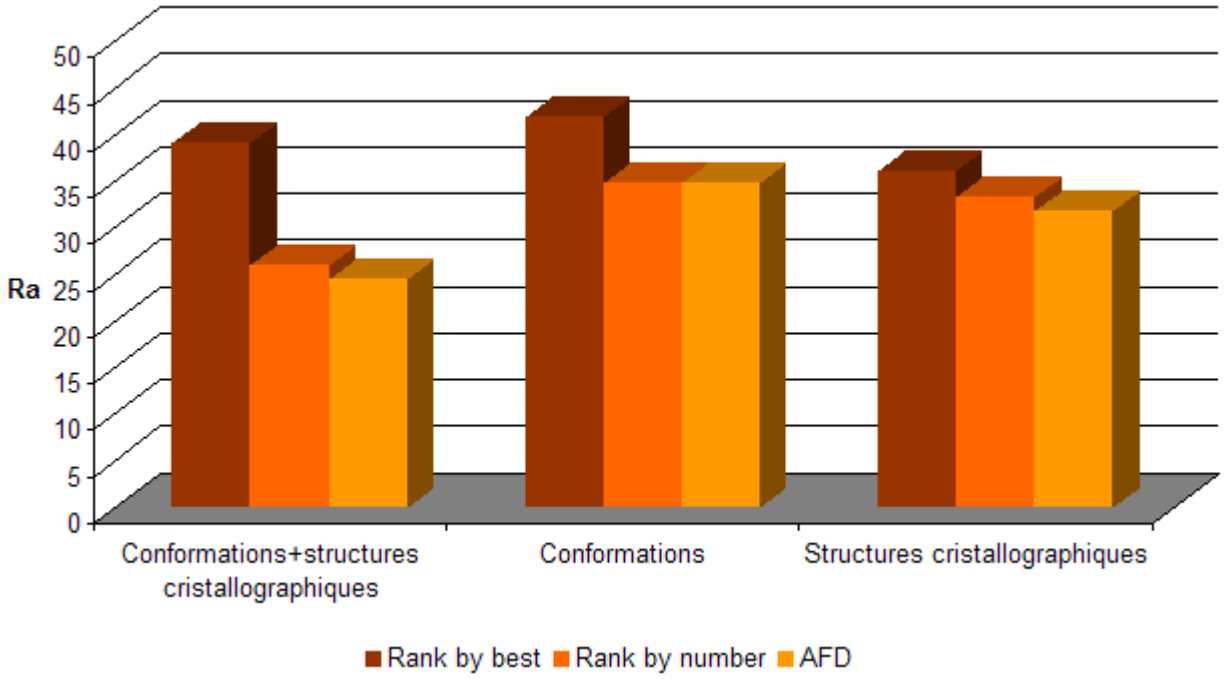


Figure 90. R_a selon les stratégies de consensus et analyses factorielles appliquées aux conformations issues de dynamique moléculaire et des structures cristallographiques

Les variations du R_a confirment les résultats précédents du R_s. Le minimum de molécules rejetées est obtenu grâce à l'utilisation de toutes les structures avec la stratégie d'analyse factorielle discriminante.

Nous avons montré que la combinaison des différentes structures de PPARγ augmente considérablement la récupération de composés actifs lors du criblage de l'*ensemble total*.

(c) Représentation graphique

L'analyse factorielle discriminante est la méthode de choix pour analyser ces données multivariées. Les représentations suivantes montrent la discrimination des deux familles de molécules (actif et inactif) dans un repère à un axe factoriel F_1 (Figure 91).

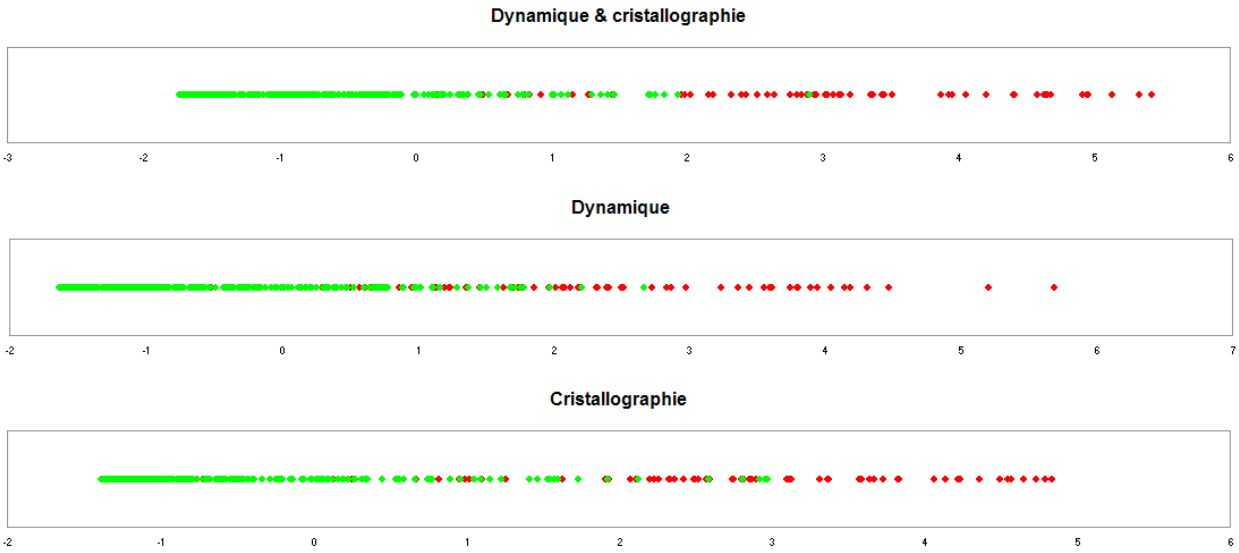


Figure 91. Distribution des molécules de l'ensemble total sur l'axe F_1 factoriel

Le graphique ci-dessus représente la répartition des molécules constituant l'*ensemble total*. Les molécules actives sont représentées en rouge et les inactives en vert. Les trois représentations mettent en évidence la discrimination des deux catégories de l'*ensemble total*. La stratégie de consensus sur les trois types de données est efficace puisque l'on observe dans chaque situation la séparation des composés de l'*ensemble total*. En réalité, la simple observation de ces graphiques ne suffit pas à conclure sur un modèle meilleur que les autres. Nous avons donc évalué le taux d'erreur associé à chaque stratégie:

$E_{\text{Dynamique \& cristallographie}}$=6,1%, $E_{\text{Dynamique}}$=9,8% et $E_{\text{Cristallographie}}$=8,3%

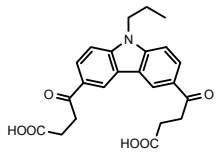
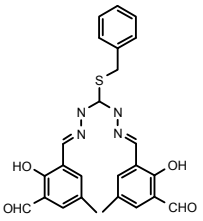
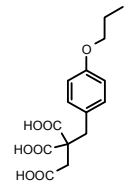
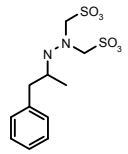
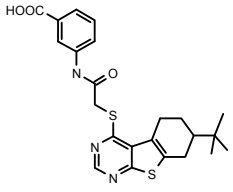
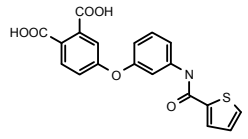
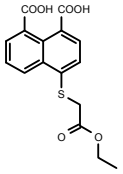
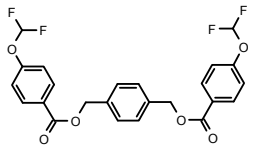
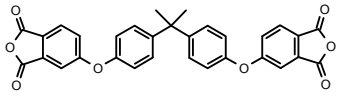
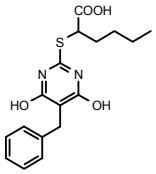
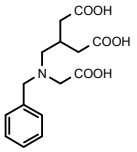
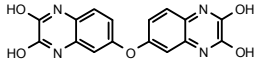
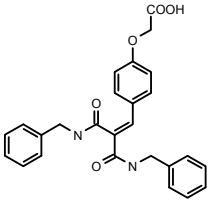
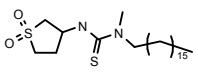
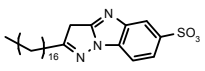
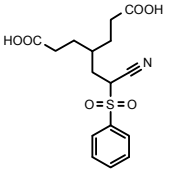
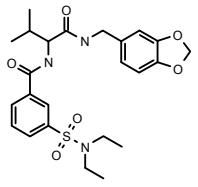
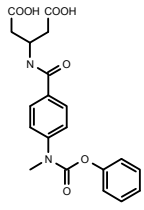
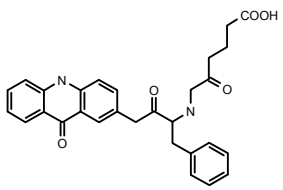
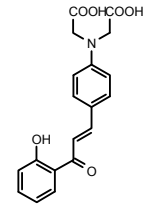
L'erreur associée à la stratégie utilisant les données issues de la dynamique moléculaire est la plus faible. Nous avons donc utilisé cette méthodologie pour le criblage virtuel.

3. Estimation des temps de calcul

Une préoccupation majeure lors de l'utilisation de plusieurs structures tridimensionnelles de protéine est la consommation en temps de calcul. En effet, les processus de criblage virtuel manipulant plusieurs millions de composés mènent à des temps de calcul importants. En particulier, en considérant une chimiothèque de 1 000 000 de composés, l'estimation du temps de calcul est d'environ de 40 jours sur un cluster de 10 PC AMD 64 3800+ pour une structure de la protéine. Il est donc difficile d'appliquer en série le docking de ces composés sur les 7 formes sélectionnées (2GTK, 2PRG, 4PRG_a, A, B, E et F).

4. Criblage virtuel

Nous avons choisi de cribler la chimiothèque dans un premier temps sur la structure 2PRG (car étant à l'origine des conformations A, B, E, F). Les 50 000 meilleures molécules ont été extraites après criblage. Les 6 structures restantes ont été appliquées sur les 50 000 composés intermédiaires (voir Figure 67). Le temps de calcul nécessaire à cette étape est de 12 jours. Un total de 50 jours est nécessaire pour cribler 1 million de composés et introduire ainsi ponctuellement la notion de flexibilité du site actif. Nous avons sélectionné au total 34 composés qui répondent aux 7 structures 2GTK, 2PRG, 4PRG_a, A, B, E et F (Figure 92).

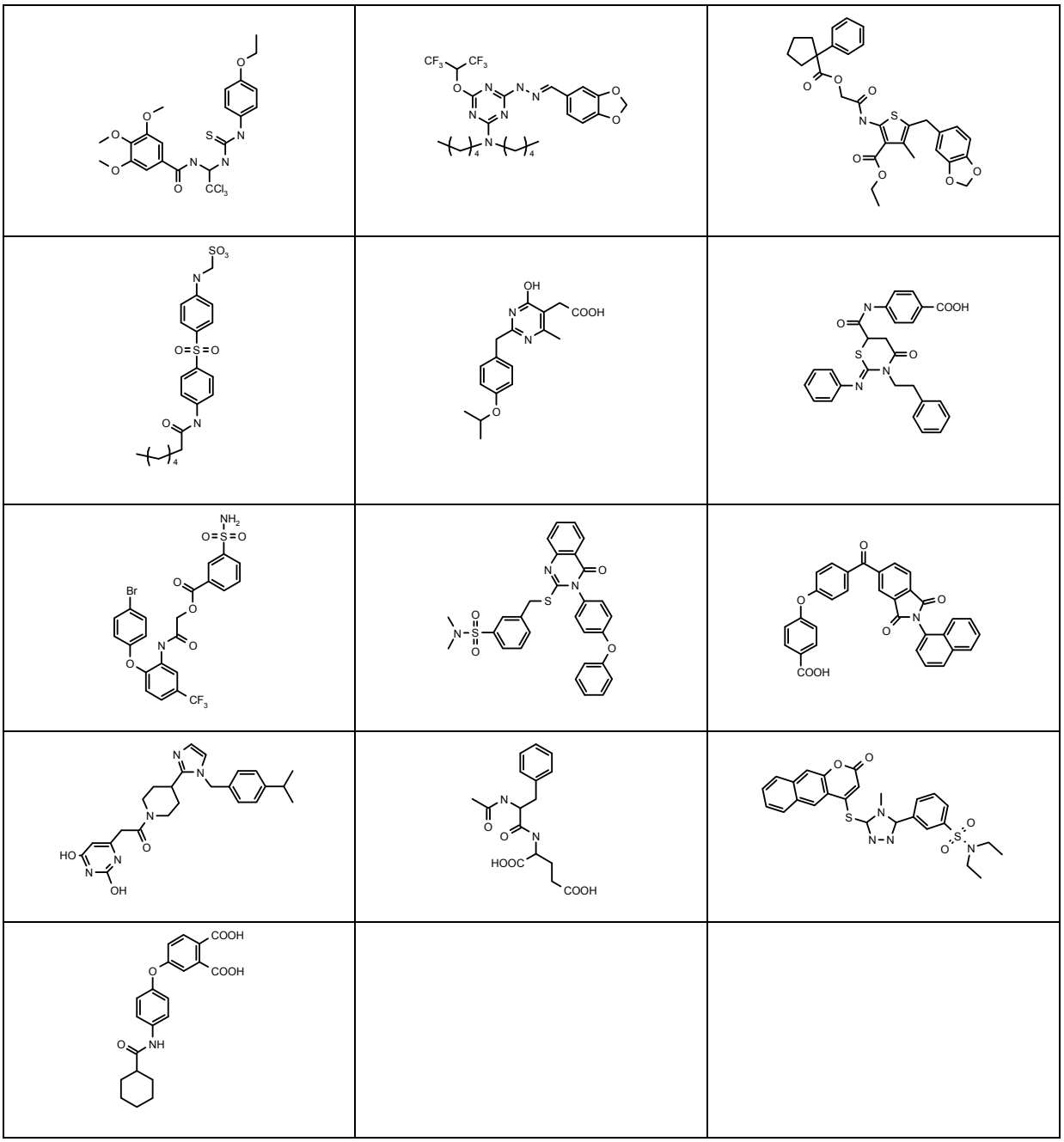


Figure 92 Sélection des 34 composés selon 2GTK, 2PRG, 4PRG_a, A, B, E et F

La figure 92 montre une diversité structurale des ligands sélectionnés pour les test biologiques (actuellement en cours). Une fonctionnalité récurrente est l'acide carboxylique, capable d'interagir avec les résidus His323, His449 et Tyr473. Des fonctions sulfonyle ont également été sélectionnées.

L. Conclusion et perspectives

Le but de cette étude était d'élaborer un protocole applicable à un processus de criblage virtuel sur PPARγ. Pour ce faire, nous avons dans un premier temps évalué toutes les données cristallographiques disponibles dans la PDB. Nous avons ainsi identifié les espèces tridimensionnelles de PPARγ possédant les meilleurs potentiels prédictifs vis-à-vis de notre *ensemble total*.

Nous avons aussi introduit la notion de flexibilité grâce à l'utilisation de la dynamique moléculaire par échantillonnage de conformères produits lors des 10ns de simulation. L'utilisation de plusieurs structures tridimensionnelles (structures cristallographiques et issues de la dynamique moléculaire) a prouvé un réel gain en terme de prédiction. Nous avons également testé, à chaque étape de l'étude, la validité des données de scoring par l'utilisation de Fitting-Score. Les courbes de fréquence relative cumulée nous ont permis de visualiser immédiatement le mode de docking des molécules actives. Nous avons mis en évidence que, dans certains cas, les données de Fitting-Score n'étaient pas forcément en adéquation avec les scores.

Une de nos préoccupations a également été d'évaluer le temps nécessaire à l'application d'une telle stratégie à un ensemble de molécules à cribler d'environ 1 million de composés.

Le choix des conformères issus de la dynamique moléculaire reste difficile à justifier. Nous avons ici choisi des conformères provenant des îlots de stabilité. Nous envisageons de tester l'utilisation de logiciels de classement des structures générées par dynamique moléculaire. Ceci apporterait une autre information pour ce choix mais cette méthode devra être validée.

CONCLUSION GENERALE

Ce travail nous a donné l'opportunité d'utiliser différentes stratégies applicables au criblage virtuel. Nous en avons testé les performances sur deux projets pharmacologiquement différents: l'inflammation et le diabète de type 2.

Dans la première partie de ce rapport, nous avons décrit des méthodes utilisables pour analyser l'affinité d'une molécule organique pour une cible de nature protéique. En particulier, nous avons montré l'intérêt du docking pour lequel il est primordial de considérer la flexibilité du ligand. Nous avons également introduit la notion de dynamique moléculaire en amont du docking afin d'introduire l'information de variabilité structurale de la cavité. Par ailleurs, les fonctions de scoring ont été décrites ainsi que la manière dont l'information doit en être interprétée. Plus précisément, plusieurs méthodes de classement ont été exposées telles que le consensus mais aussi l'analyse factorielle discriminante. Nous avons appliqué des métriques telles que le facteur d'enrichissement, les taux de vraies/fausses prédictions et le R_s dans l'objectif d'exploiter au mieux l'information numérique des fonctions de scoring. Leurs limites d'utilisation ont été démontrées.

La deuxième partie est consacrée au développement de protocoles de docking-scoring pour la cible « cyclooxygénase de type 2 », impliquée dans les phénomènes de l'inflammation. Après avoir optimisé les paramètres structuraux de la protéine, nous avons comparé les performances de chaque fonction de scoring prise de manière individuelle. Par la suite, la combinaison par consensus des scores a montré un gain en terme de prédiction mais également une amélioration du classement des composés actifs après criblage de notre *ensemble total*. De plus, nous avons analysé les multiples structures cristallographiques de COX-2 disponibles dans la PDB. Cette comparaison a prouvé la variabilité du site actif qui pourrait laisser envisager l'utilisation de méthodes capables de prendre en compte les adaptations de la cavité. Enfin, l'utilisation d'un pharmacophore comme filtre avant docking (FlexX-Pharm) nous a permis d'envisager le criblage virtuel sur une chimiothèque de un million de composés.

La troisième et dernière partie de nos travaux concerne le récepteur nucléaire PPARγ, lié au diabète de type 2. Dans un premier temps, nous avons prouvé, par docking de l'*ensemble total* dans les différentes structures provenant de la PDB, que la cavité était capable de fortement s'adapter. De même, les calculs de volume du site actif confirment l'aspect flexible et adaptatif de cette cible polymorphe.

Ainsi, nous avons couplé au docking la dynamique moléculaire, susceptible de fournir une information supplémentaire sur les mouvements de la cavité et ainsi d'améliorer les résultats de prédiction. Enfin, les techniques de consensus normalement décrites dans le cadre

des fonctions de scoring ont été ici appliquées aux données provenant des différents conformères générés par dynamique moléculaire mais aussi par la voie expérimentale (cristallographie par rayons X).

Enfin, des études ont montré que la cible PPARγ régule de manière négative l'expression de la cyclooxygénase de type 2. Par conséquent, des agonistes de la cible PPARγ peuvent être considérés comme inhibiteurs indirects des COX-2. Ainsi, une perspective intéressante serait d'inclure, dans les tests biologiques sur PPARγ, la composante COX-2.

Ces travaux ainsi que les projets connexes auxquels j'ai participé ont donné lieu à différentes communications et trois publications dont une soumise. La liste des travaux est indiquée ci-dessous:

Communications

Arrault, A. ; Monge, A. ; Violas, S. ; Genest, M. ; Genest, D. ; Garnier, N. ; Morin-Allory, L. ; Marot, C. *Molecular dynamic assisted docking. Application to the screening on PPAR gamma* 10th Anniversary of MipTec – The Leading European Event for Drug Discovery Technologies 05-2007 - Basel (Switzerland).

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. *Docking-scoring assisté par des stratégies de consensus et d'analyses de données multivariées appliqué à la cible COX-2*. Groupe de Graphisme et Modélisation Moléculaire (GGMM'2007) 05-2007 - Autrans.

Arrault, Al. ; Monge, A. ; Genest, M. ; Genest, D. ; Garnier, N. ; Morin-Allory, L. ; Marot, C. *Docking assisté par dynamique moléculaire. Application au criblage virtuel sur la cible PPAR gamma* Groupe de Graphisme et Modélisation Moléculaire (GGMM'2007) 05-2007 - Autrans.

Violas, S. ; Arrault, Al. ; Genest, M. ; Garnier, N. ; Morin-Allory, L. ; Marot, C. ; Genest, D. *Pathways for a ligand to escape from its binding site in human PPAR-gamma as probed by molecular dynamics simulation* 20ème congrès de la Société Française de Biophysique 10-2006 - Anglet.

Arrault, Al. ; Monge, A. ; Violas, S. ; Genest, M. ; Genest, D. ; Garnier, N. ; Morin-Allory, L. ; Marot, C. *Molecular dynamic assisted docking. Application to the screening on PPAR gamma* 16th European Symposium on Quantitative Structure-Activity Relationships & Molecular Modelling (Euro QSAR) 09-2006 - Civitavecchia (Italia).

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *ScreeningAssistant: a free software for managing chemical databases* 16th European Symposium on Quantitative Structure-Activity Relationships & Molecular Modelling (Euro QSAR) 09-2006 - Civitavecchia (Italia).

Arrault, Al. ; Monge, A. ; Morin-Allory, L. ; Marot, C. *La rencontre d'un médicament avec sa cible thérapeutique, un mécanisme simulé par méthode informatique* Sciences en Sologne 2006 06-2006 - Orléans.

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *De la chimiothèque au criblage virtuel* 4ème journée du projet CASCIMODOT 06-2006 - Tours.

Arrault, Al. ; Monge, A. ; Morin-Allory, L. ; Marot, C. *Optimisation d'un protocole de docking-scoring* XXèmes Journées Franco-Belges de Pharmacochimie 05-2006 - Lille.

Arrault, Al. ; Monge, A. ; Morin-Allory, L. ; Marot, C. *Optimization of a docking-scoring protocol*Chemoinformatics in Europe: Research and Teaching 05-2006 - Obernai.

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Plate selection: application to screening library design* 12th International Workshop on Quantitative Structure-ActivityRelationships in Environmental Toxicology (QSAR 2006) 05-2006 - Lyon.

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *ScreeningAssistant : A free platform for managing huge chemical databases and selecting compounds for screening* Chemoinformatics in Europe: Research and Teaching 05-2006 - Obernai.

Arrault, Al. ; Monge, A. ; Morin-Allory, L. ; Marot, C. *Optimization of a docking-scoring protocol* Molecular Modelling 2006 04-2006 - Perth (Australia).

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Analysis of a virtual database of five million commercially available compounds* Molecular Modelling 2006 04-2006 - Perth (Australia).

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. *Protocole de docking-scoring appliqué au criblage virtuel sur la cible thérapeutique COX-2* 41èmes Rencontres Internationales de Chimie Thérapeutique 07-2005 - Paris.

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Utilisation de "ScreeningAssistant" pour la conception d'une base de criblage optimisée à partir de 4 millions de produits* 41èmes Rencontres Internationales de Chimie Thérapeutique 07-2005 - Paris.

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. *Conception de molécules à visée thérapeutique par modélisation moléculaire* Sciences en Sologne 2005 06-2005 - Orléans.

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *La chimiothèque : un élèment clé de la découverte de nouveaux médicaments* Sciences en Sologne 2005 06-2005 - Orléans.

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. *Docking/scoring couplé à un pré-filtrage pharmacophorique. Application au criblage virtuel sur la cible COX-2* XIVème Colloque du Groupe de Graphisme et Modélisation Moléculaire (GGMM'2005) 05-2005 - îles des Embiez.

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. *Elaboration d'un modèle de docking/scoring. Application au criblage virtuel sur les cibles PPAR alpha et gamma* XIVème Colloque du Groupe de Graphisme et Modélisation Moléculaire (GGMM'2005) 05-2005 - îles des Embiez.

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Analyse de l'espace chimique de plus de 3 millions de molécules destinées au criblage virtuel ou au criblage à haut débit* XIVème Colloque du Groupe de Graphisme et Modélisation Moléculaire (GGMM'2005) 05-2005 - îles des Embiez.

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Screening Assistant : un logiciel de gestion de chimiothèques* XIVème Colloque du Groupe de Graphisme et Modélisation Moléculaire (GGMM'2005) 05-2005 - îles des Embiez.

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. *Computer-aided design of original COX2- inhibitors: docking protocols and virtual screening under pharmacophoric constraints* 10th Electronic Computational Chemistry Conference (ECCC10) 04-2005 - Internet and World Wide Web.

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Analysis of a set of 2.6 million unique compounds gathered from the libraries of 32 chemical providers* 10th Electronic Computational Chemistry Conference (ECCC10) 04-2005 - Internet and World Wide Web.

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. *Virtual screening as an alternative and a complement to high throughput screening* PharmaSolution Expo 03-2005 - London (Angleterrre).

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Conception of a huge chemical database and test set design for virtual and high throughput screening* PharmaSolution Expo 03-2005 - London (Angleterrre).

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. *Design of selective non-steroidal anti-inflammatory inhibitors to cyclooxygenase 2 under docking methods* 5th International EMBL PhD students symposium 12-2004 - Heidelberg (Allemagne).

Monge, A. ; Arnoult, E. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Development of a chemical databases management software* The 15th European Symposium on Quantitative Structure-Activity Relationships & Molecular Modeling (QSAR 2004) 09-2004 - Istanbul (Turkey).

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. *Conception d'un logiciel de gestion de grandes chimiothèques destinées au screening virtuel* Sciences en Sologne 2004 05-2004 - Orléans.

Publications

Monge, A. ; Arrault, Al. ; Marot, C. ; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers *Molecular Diversity* 2006, *10*, 389-403.

Genest, D.; Garnier, N.; Arrault, A.; Marot, C.; Morin-Allory, L. Ligand-escape pathways from the ligand binding domain of PPAR gamma receptor as probed by molecular dynamics simulations Acceptée à *J. Mol. Biol.*

Arrault, Al. ; Monge, A. ; Marot, C. ; Morin-Allory, L. A study on novel non-steroidal Anti-inflammatory inhibitors (NSAIDs) to Cyclooxygenase 2 using docking methods involved in a virtual screening process. Soumise à *J. Med. Chem.*.

GLOSSAIRE

ADME	critères (absorption, distribution, métabolisme et excrétion) selon lesquelles il est possible de définir la pharmacocinétique et la pharmacologie d'un composé candidat-médicament.
Analyse en composante principale (ACP)	méthode de compression de données provenant d'une matrice multidimensionnelle et permettant l'observation des variables et des individus dans un repère de moindre dimensionnalité.
Analyse factorielle discriminante (AFD)	méthode statistique multilinéaire similaire à l'analyse en composantes principales permettant l'étude, dans notre cas, de la discrimination entre deux groupes (actif et inactif) d'individus.
Centre de géométrie	centre d'une molécule dont les coordonnées ne sont pas pondérées par la masse des atomes.
Centre de masse	centre d'une molécule dont les coordonnées sont pondérées par la masse des atomes.
Consensus	approche consistant à combiner plusieurs méthodes pour tirer profit de leur complémentarité.
Criblage virtuel	processus qui consiste à tester *in silico* l'affinité de molécules vis-à-vis d'une cible donnée.
Docking	étape de positionnement d'un ligand dans le site de fixation de l'enzyme.
Dynamique moléculaire	méthode de simulation combinant un champ de force de mécanique moléculaire et les équations classiques de la cinétique pour prédire le mouvement des atomes dans le temps.

Ensemble d'apprentissage (training set)	ensemble de molécules permettant l'élaboration d'un modèle. Il correspond en proportion à 2/3 des molécules actives et 2/3 des molécules inactives de l'*ensemble total*.
Ensemble de test (test set)	ensemble de molécules permettant de tester le modèle élaboré à partir de l'*ensemble d'apprentissage*. Il comprend le 1/3 des molécules actives restantes et le 1/3 des molécules inactives restantes de l'*ensemble total*.
Ensemble total	ensemble constitué de molécules actives et inactives pour une cible donnée (70 molécules actives, 280 molécules inactives dans notre cas).
Facteur d'enrichissement (E_r)	facteur d'augmentation de la fraction de composés affins dans un sous-ensemble de ligands comparé à la base d'origine.
Faux négatif (FN)	molécule *in vitro et/ou in vivo* active prédite *in silico* inactive.
Faux positif (FP)	molécule *in vitro et/ou in vivo* inactive prédite *in silico* active.
Fitting-Score	score capable d'évaluer le mode de docking d'une molécule au sein du site actif.
Fonction de scoring	algorithme mathématique capable de prédire l'énergie d'un complexe protéine/ligand.
Lead	molécule possédant une activité intéressante vis-à-vis d'une cible et dont les propriétés d'ADME et de solubilité aqueuse doivent être optimisées afin de mener ce candidat médicament aux phases de développement.
Première pose	conformation la plus représentative d'un jeu de conformères sélectionnée selon des critères de choix basés sur les fonctions de

	scoring.
Rejected active (R_a)	indicateur non biaisé par le surnombre des molécules inactives (comme dans le cas du calcul du %FN) traduisant la proportion de molécules actives perdues par le modèle.
RMSD	« root mean square deviation », représente l'écart entre deux conformations.
R_s	indicateur capable de décrire quantitativement la qualité d'une courbe d'enrichissement.
Score/scoring	valeur numérique traduisant l'évaluation des énergies d'interaction entre un ligand et une protéine.
Sphère de présence	volume dans lequel le centre de géométrie ou de masse d'un ligand doit se situer pour le considérer positionné dans le site actif.
Vrai négatif (VN)	molécule *in vitro et/ou in vivo* inactive prédite *in silico* inactive.
Vrai positif (VP)	molécule *in vitro et/ou in vivo* active prédite *in silico* active.

Stratégies de docking-scoring assistées par analyse de données. Application au criblage virtuel des cibles thérapeutiques COX-2 et PPAR gamma.

Le criblage virtuel est une technique permettant d'extraire, d'une chimiothèque donnée, des produits actifs ou affin pour une cible ou un profil pharmacologique donné. Nous avons développé une méthodologie impliquant les données tridimensionnelles des protéines COX2 et PPARγ. Tout d'abord, nous avons comparé les différentes structures entre elles mais également les fonctions de scoring utilisées pour prédire l'affinité de molécules pour ces cibles. Par ailleurs, nous avons étudié des méthodes de consensus et d'analyse de données multivariée pour interpréter les fonctions de scoring. De plus, l'incorporation de techniques originales au protocole de docking-scoring a été testée. Plus précisément, un modèle pharmacophore, agissant comme filtre de composés indésirables, a été évalué pour diminuer les temps de calcul mais également pour améliorer le choix de la première pose. Par ailleurs, le couplage de la dynamique moléculaire, en amont du docking, nous a permis de prendre en compte la flexibilité du site actif. Nous avons montré l'utilité d'une telle stratégie pour améliorer les prédictions. Enfin, nous avons appliqué les méthodes de consensus et d'analyse de données multivariées (normalement employées pour les fonctions de scoring) aux données provenant des conformères issus de la dynamique moléculaire.

MOTS-CLES : Arrimage, fonctions de scoring, fouille de données, criblage virtuel.

Docking-scoring strategies assisted by data analysis. Application to virtual screening of COX-2 and PPAR gamma therapeutical targets.

Virtual screening is a strategy able to pick up high affinity or activity compounds for a given pharmacological target. We have developed a methodology which involves three dimensional data of COX-2 and PPARγ receptors. First, we have compared the available structures but we have also studied the different scoring functions (able to predict the binding of a molecule within a protein). Moreover, we have tested consensus techniques but also multivariate data analysis methodologies to treat the information from the scoring functions. We have also studied the incorporation of pharmacophoric constraints prior to docking, acting as a filter to remove undesirable compounds. This pharmacophoric model has also improved the choice of the first pose. Another investigation has been to assist the docking procedure with molecular dynamic. The aim of this task has been to take into consideration the flexibility of the active site of the protein. We have shown that a gain can be expected with such a strategy. Finally, consensus and multivariate data analysis has been applied to the data generated by all the conformers.

KEY WORDS : Docking, scoring functions, data mining, virtual screening.