



**HAL**  
open science

# Techniques de Modélisation Moléculaire appliquées à l'Etude et à l'Optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance.

Antoine Fortuné

► **To cite this version:**

Antoine Fortuné. Techniques de Modélisation Moléculaire appliquées à l'Etude et à l'Optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance.. Médicaments. Université Joseph-Fourier - Grenoble I, 2006. Français. NNT: . tel-00289767

**HAL Id: tel-00289767**

**<https://theses.hal.science/tel-00289767>**

Submitted on 23 Jun 2008

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**UNIVERSITE JOSEPH FOURIER - GRENOBLE I**

Année 2006

Thèse N°

**THESE**

Pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITE JOSEPH FOURIER**

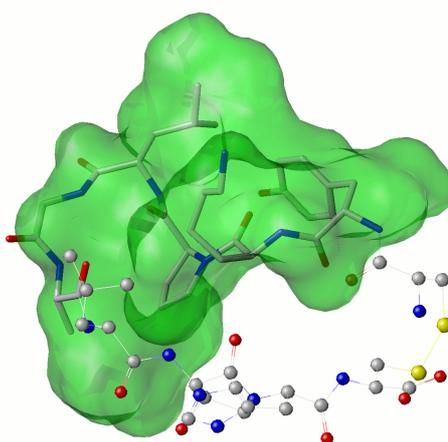
Discipline : **SCIENCES DU MEDICAMENT**

Présentée et soutenue publiquement le 21 décembre 2006

par

**Antoine FORTUNÉ**

**Techniques de Modélisation Moléculaire Appliquées à  
l'Etude et à l'Optimisation de Molécules Immunogènes  
et de Modulateurs de la Chimiorésistance**



Composition du jury :

**Dr. A. IMBERTY**

**Dr. A. BOUMENDJEL**

**Prof. P.-A. CARRUPT**

**Dr. L. CHICHE**

**Dr. S. CROUZY**

**Prof. J.-L. DECOUT**

Directeur de Recherche, CNRS - Grenoble

Maître de Conférence, Université Grenoble 1

Professeur, Université de Genève

Directeur de Recherche, CNRS - Montpellier

Chercheur, CEA - Grenoble

Professeur, Université Grenoble 1

Directeur de thèse

Directeur de thèse

Rapporteur

Rapporteur

Examineur

Examineur

## REMERCIEMENTS

Les travaux présentés dans ce mémoire ont été réalisés au Département de Pharmacochimie Moléculaire, UMR5063 UJF - CNRS et au Centre de Recherche sur les Macromolécules Végétales, UPR 5301 CNRS.

Je remercie Monsieur le Professeur Jean-Luc Décout, directeur du DPM, de m'avoir permis de faire cette thèse parallèlement à mes activités d'ingénieur d'étude et de m'avoir permis d'aller régulièrement travailler au CERMAV pendant près d'un an. Je remercie Monsieur Serge Pérez, directeur du CERMAV, de m'avoir accueilli dans son unité et encouragé dans cette entreprise. Merci Messieurs de m'avoir fait l'honneur d'examiner ce travail.

Je suis très honoré que Monsieur Pierre-Alain Carrupt et Monsieur Laurent Chiche aient accepté de juger cette thèse. Qu'ils trouvent ici l'expression de ma plus vive considération et de ma sincère gratitude pour avoir participé à ce jury.

J'adresse ma profonde reconnaissance à Anne Imberty qui m'a accepté dans son équipe et à co-dirigé ce travail. Un **grand merci** pour m'avoir formé à la modélisation moléculaire et en particulier à la modélisation des polysaccharides et des protéines. J'ai particulièrement apprécié ta disponibilité, ton enthousiasme, ta confiance et tes conseils, nombreux et avisés.

Un autre **grand merci**, à Ahcène Boumendjel, co-directeur de ma thèse, pour ta confiance, ton soutien et ton intérêt constants pour ce travail. Ces années ont été riches en enseignements et je suis heureux que notre collaboration se poursuive, sur ce sujet et sur d'autres.

Mes très sincères remerciements à Luc pour avoir pris le temps de répondre à mes questions et pour tes précieux conseils, particulièrement en statistiques.

Mes remerciements au CECIC (Centre Experimental de Calcul Intensif en Chimie) grace auquel la majeure partie des calculs ont été réalisables et réalisés. Merci à Alain, Jean-Daniel, Pierre et Sébastien pour votre sympathie, pour votre aide en informatique, et pour vos interventions express.

Je tiens à exprimer ma sincère gratitude à tous ceux qui ont contribué, à cette thèse par leurs conseils, leur disponibilité, leur bonne humeur et leur convivialité : Anna, Yung-Sing, Karim, Marie-Claire, Madeleine, Jérôme, Marie-Carmen, Isabelle, Eric, Annabelle, Chantal, Denis, Anne-Marie, Gilbert, Christine, Monsieur Bakri, Nawel, Béatrice, Marine, Benjamin.

Je remercie l'ensemble de mes collègues du DPM qui savent composer avec mes multiples activités.

Je souhaite également exprimer mon profond respect, ma reconnaissance et mon admiration pour Monsieur le Professeur Philippe-Jean Coulomb de l'Université d'Avignon et Monsieur Henri Garreta de la Faculté de Luminy, dont j'ai eu la chance et l'honneur d'être un élève durant mes études universitaires. Par la qualité de leur enseignement et leur passion pour la Science, ils m'ont fait découvrir l'émerveillement que peuvent procurer la contemplation et l'étude de notre Univers. Si j'en suis là aujourd'hui, c'est aussi grâce à eux.

**À mes parents, mon frère et ma soeur,**

*pour m'avoir donné les moyens de réussir et avoir cru en moi.*

**À ma grande et chère famille,**

*source intarissable de réconfort, de joies et d'émulation.*

**À Maud, mon amie de 30 ans (!), à François, à Elodie et à Manu,**

*mes compagnons d'aventures, merci pour tous ces moments inoubliables.*

*Votre amitié m'est précieuse.*

**À Valérie, ma compagne,**

*pour ton soutien de tous les jours et ton aide dans les moments difficiles, pour ta gentillesse,*

*ta patience et ton courage ... et tellement plus ...*

# SOMMAIRE

|  |           |
|--|-----------|
| <b>REMERCIEMENTS</b>   | <b>3</b>  |
| <b>SOMMAIRE</b>  | <b>5</b>  |
| <b>TABLEAUX ET FIGURES</b>   | <b>8</b>  |
| <b>INTRODUCTION</b>  | <b>11</b> |
| <b>CONSTRUCTION DE PROTEINES PAR HOMOLOGIE ET AMARRAGE MOLECULAIRE</b>   | <b>13</b> |
| <b>A - INTRODUCTION</b>  | <b>14</b> |
| <b>B - PRINCIPES ET METHODES</b>   | <b>15</b> |
| <b>I - Construction de protéine par homologie</b>  | <b>15</b> |
| 1. Recherche et identification des protéines homologues  | 15        |
| 2. Identification des régions conservées, communes aux structures de référence sélectionnées   | 18        |
| 3. Alignement spatial des structures et alignement de la séquence cible  | 19        |
| 4. Construction des régions structurellement conservées  | 19        |
| 5. Construction des régions variables  | 20        |
| 6. Affinement du modèle  | 22        |
| 7. Validation partielle du modèle  | 22        |
| <b>II - Amarrage avec AUTODOCK</b>   | <b>23</b> |
| 1. Modélisation du potentiel énergétique   | 23        |
| 2. Calcul du potentiel énergétique   | 26        |
| 3. Exploration de l'environnement spatial et conformationnel.  | 27        |
| a. <i>Les Algorithmes Génétiques (AG).</i>   | 27        |
| b. <i>L'Algorithme Génétique Lamarckien (AGL).</i>   | 30        |
| <b>III - La procédure AUTODOCK</b>   | <b>32</b> |
| 1. Préparation des fichiers.   | 32        |
| 2. Définition des pivots flexibles du ligand.  | 33        |
| 3. Calcul des grilles de potentiels.   | 33        |
| 4. Recherche des solutions d'amarrage.   | 34        |
| 5. Analyse des résultats.  | 36        |
| <b>C - APPLICATION : ETUDE DU MECANISME DE RECONNAISSANCE D'UN POLYOSIDE BACTERIEN ET DE MIMES PEPTIDIQUES PAR L'ANTICORPS SPECIFIQUE IGA-I3</b> | <b>37</b> |
| <b>I - Introduction</b>  | <b>37</b> |
| 1. La shigellose   | 37        |
| 2. Les stratégies de vaccination   | 38        |
| 3. Les mimes peptidiques   | 40        |
| <b>II - Modèle de l'anticorps</b>  | <b>41</b> |
| 1. Recherche de structures tridimensionnelles de séquence homologue.   | 43        |
| a. <i>Recherche d'homologues de la chaîne légère.</i>  | 43        |
| b. <i>Recherche d'homologues de la chaîne lourde.</i>  | 45        |

|  |            |
|--|------------|
| 2. Construction de la chaîne légère.   | 47         |
| 3. Validation de la chaîne légère.   | 47         |
| 4. Construction de la chaîne lourde.   | 48         |
| 5. Validation de la chaîne lourde.   | 49         |
| 6. Assemblage des chaînes lourdes et légères.  | 49         |
| <b>III - Modèles du polysaccharide.</b>  | <b>49</b>  |
| <b>IV - Modèles du mime peptidique.</b>  | <b>51</b>  |
| <b>V - Amarrage des antigènes à l'anticorps.</b>                                       | <b>52</b>  |
| 1. Définition de la boîte d'amarrage   | 52         |
| 2. Amarrage du polysaccharide sur IgA I3   | 53         |
| <i>a. Amarrage de la forme condensée T1.</i>   | 53         |
| <i>b. Amarrage de la forme étendue T4.</i>   | 55         |
| 3. Amarrage du peptide p100c sur IgA I3  | 57         |
| <b>VI - Comparaison des modes d'interaction du polysaccharide et du peptide.</b>       | <b>64</b>  |
| <b>VII - Conclusion de L'étude</b>   | <b>65</b>  |
| <b>D - CONCLUSION</b>  | <b>66</b>  |
| <b>E - BIBLIOGRAPHIE</b>   | <b>667</b> |
| <b>F - ANNEXES</b>   | <b>72</b>  |
| Annexe I : Abréviations et codage alphabétique des acides aminés selon la règle FASTA. | 72         |
| Annexe II : Document de prise en main rapide d'Autodock3.                              | 73         |
| Annexe III : Article du <i>Journal of Biological Chemistry</i> .                       | 78         |
| <br>   |            |
| <b>ANALYSE QUANTITATIVE EN TROIS DIMENSIONS DES RELATIONS STRUCTURE-<br/>ACTIVITE</b>  | <b>95</b>  |
| <b>A - INTRODUCTION</b>  | <b>96</b>  |
| <b>B - PRINCIPES ET METHODES DE MODELISATION STATISTIQUE</b>                           | <b>98</b>  |
| <b>I - Principes généraux de statistiques</b>  | <b>98</b>  |
| 1 - Homogénéité des données.   | 98         |
| <i>a. Source des données</i>   | 98         |
| <i>b. Traitement des données brutes</i>  | 98         |
| <i>c. Echantillonnage de l'espace chimique</i>   | 100        |
| 2 - Les descripteurs chimiques   | 100        |
| <i>a. Descripteurs 2D</i>  | 100        |
| <i>b. Descripteurs 3D</i>  | 101        |
| 3 - Les modèles statistiques   | 101        |
| 4 - Les outils de validation des modèles   | 101        |
| <i>a. L'ajustement</i>   | 103        |
| <i>b. La précision de l'ajustement</i>   | 103        |
| <i>c. Le pouvoir de prévision interne</i>  | 104        |
| <i>d. Le test de Fischer</i>   | 104        |
| <i>e. L'auto-corrélation</i>   | 106        |

|   |            |
|---|------------|
| <i>f . Le pouvoir de prévision externe</i>  | 107        |
| <i>g . Le biais d'ajustement, modèles centrés</i>   | 107        |
| <i>h . Les critères de validation d'un modèle</i>   | 108        |
| <i>i . L'interprétation du <math>q^2</math></i>   | 108        |
| 5 - Constitution des jeux d'apprentissage et de test  | 109        |
| 6 - Stratégie globale d'une étude QSPR  | 110        |
| <b>II - Adaptation des outils statistiques aux données pharmaco-chimiques : QSAR-3D basés sur des Champs d'Interactions ou d'indices de similarité Moléculaire.</b> | <b>111</b> |
| 1 - Comparative Molecular Field Analysis (CoMFA)  | 112        |
| 2 - Comparative Molecular Similarity Indices Analysis (CoMSIA)  | 113        |
| 3 - Alignement des structures   | 113        |
| 4 - Calcul des champs d'interaction moléculaire   | 114        |
| 5 - Séquence et résultats des analyses CoMFA et CoMSIA  | 116        |
| <i>a . Construction d'une table de molécules</i>  | 116        |
| <i>b . Calcul des descripteurs</i>  | 116        |
| <i>c . Séquences des analyses PLS</i>   | 117        |
| 6 - Visualisation graphique des modèles   | 118        |
| 7 - Prévision et extrapolation  | 118        |
| <b>C – APPLICATION : QSAR-3D AVEC LES DESCRIPTEURS COMSIA D'UNE SERIE DE FLAVONOIDES ET DE BOERHAVINONES MODULATEURS DU TRANSPORTEUR BCRP.</b>                      | <b>119</b> |
| <b>I - Introduction</b>   | <b>119</b> |
| <b>II - Données structurales et biologiques</b>   | <b>121</b> |
| <b>III - Conformations et alignement</b>  | <b>124</b> |
| <b>IV - Construction et validations des modèles</b>   | <b>125</b> |
| 1 - Modèles CoMSIA du jeu de données complet.   | 125        |
| <i>a . Construction des modèles :</i>   | 125        |
| <i>b . Analyse des modèles :</i>  | 125        |
| 2 - Validation Externe.   | 127        |
| <i>a . Constitution des jeux d'apprentissage et construction des modèles.</i>   | 127        |
| <i>b . Estimation du pouvoir de prévision externe :</i>   | 129        |
| <i>c . Analyse du pouvoir de prévision de la combinaison «S,E,D»</i>  | 130        |
| 3 - Pouvoir de prévision du modèle «S,E,D»  | 132        |
| 4 - Test de robustesse du modèle « S,E,D »  | 133        |
| 5 - Conclusion des tests de validation  | 134        |
| <b>V - Relations structure-activité</b>   | <b>134</b> |
| 1 - Explication des relations structure-activité par le modèle « S,E,D »  | 134        |
| <i>a . Le champ stérique</i>  | 134        |
| <i>b . Le champ électrostatique</i>   | 136        |
| <i>c . Le champ donneur de liaison hydrogène</i>  | 137        |
| 2 - Evaluation des modèles de qualité voisine   | 138        |
| <i>a . Les modèles « E,H,D » et « S,E,H,D »</i>   | 138        |
| <i>b . Les modèles « S,E,D,A » et « S,E,H,D,A »</i>   | 139        |
| 3 - Pouvoir explicatif des modèles.   | 140        |
| 4 - Synthèse des caractéristiques qui influencent l'activité  | 141        |
| <b>VI - Domaine et Limites du modèle</b>  | <b>142</b> |

|   |            |
|---|------------|
| 1 - Echantillonnage de l'espace chimique                              | 142        |
| 2 - Corrélation de champs   | 142        |
| 3 - Extension de l'espace chimique                                    | 142        |
| 4 - Limitation de la transformation Logit                             | 143        |
| 5 - Limitations du test biologique                                    | 143        |
| <b>VII - Consolidation du modèle et optimisation des inhibiteurs.</b> | <b>143</b> |
| <b>VIII - Conclusion de l'étude</b>                                   | <b>146</b> |
| <b>D - CONCLUSION</b>   | <b>147</b> |
| <b>E – PARTIE EXPERIMENTALE</b>                                       | <b>148</b> |
| <b>I - Calcul des descripteurs COMSIA</b>                             | <b>148</b> |
| <b>II - Construction et la validation des modèles</b>                 | <b>148</b> |
| <b>III - Test de hasardisation</b>                                    | <b>150</b> |
| <b>F - BIBLIOGRAPHIE</b>  | <b>151</b> |
| <b>G - ANNEXES</b>  | <b>156</b> |
| Annexe I : Table de Fischer pour un risque de 5 %.                    | 156        |
| Annexe II : Code du script QBF1.                                      | 157        |
| Annexe III : Exemple de fichier de sortie du script QBF1.             | 161        |
| Annexe IV : Code du script QYR.                                       | 162        |
| <b>CONCLUSION</b>   | <b>165</b> |

## TABLEAUX ET FIGURES

### PARTIE CONSTRUCTION DE PROTEINES PAR HOMOLOGIE ET AMARRAGE MOLECULAIRE

#### Liste des Tableaux :

|  |    |
|--|----|
| Tableau 1 : Liaisons hydrogène entre T1 et IgA I3 dans les modes de liaison parallèle (à gauche) et perpendiculaire (à droite). Les résidus de la boucle hypervariable H3 sont notés en rouge.   | 55 |
| Tableau 2 : Listes des interactions entre l'anticorps et l'oligosaccharides (conformation étendue) associés selon deux modes de liaison différents.  | 57 |
| Tableau 3 : Résultats d'amarrage des 20 conformères du peptide P100C. Relevé des interactions du peptide avec la boucle H3 de l'anticorps (H3), et des contacts de la Tyr2 (T2) et du segment Pro-Leu-Gly-Ala (PLGA) du peptide avec la protéine. L'énergie de la meilleure solution acceptable est donnée avec son rang et sa population ainsi que la plus basse énergie proposée pour chaque conformère. | 58 |
| Tableau 4 : Relevé des contacts et liaisons hydrogène entre les solutions d'amarrage des conformations 15, 16 et 19 du peptides P100C en interaction avec IgAI3.   | 61 |

#### Liste des Figures :

|   |    |
|---|----|
| Figure 1 : Schéma d'une grille de potentiel englobant le site actif d'une macromolécule.  | 26 |
| Figure 2 : Comportement des algorithmes génétiques, principes de Darwin (à droite) et de Lamarck (à gauche).  | 30 |
| Figure 3 : Formule de l'Ag-O de <i>S. flexneri</i>  | 41 |
| Figure 4 : Schéma de la structure d'une immunoglobuline de type IgA. Les domaines de reconnaissance spécifiques sont situés à l'extrémité des chaînes.  | 42 |
| Figure 5 : Résultat de la procédure BLAST cherchant les séquences homologues à la chaîne légère d'IgAI3 dans la Protein Data Bank.  | 44 |
| Figure 6 : Résultat de la procédure BLAST cherchant les séquences homologues à la chaîne lourde d'IgAI3 dans la Protein Data Bank.  | 45 |
| Figure 7 : Alignement des séquences de IgA I3 (haut) et 1A4J (bas) (numérotation PDB). Les résidus surlignés d'une étoile sont invariants, ceux surlignés par un tiret sont dans une boucle hypervariable [64].   | 46 |
| Figure 8 : Modèle de l'anticorps IgA I3. Représentation de la chaîne principale en tube (à gauche) et représentation de la surface de Conolly, surface accessible au solvant, du site de reconnaissance (à droite). Les boucles hypervariables sont en vert pour la chaîne légère, en violet pour la chaîne lourde. | 49 |
| Figure 9 : Fragment du O-SP. Conformations condensée (en haut) et étendue (en bas).   | 50 |
| Figure 10 : Structures superposées de 20 conformères du peptide P100C   | 51 |
| Figure 11 : Position et dimensions de la boîte d'amarrage autour du site de reconnaissance de l'anticorps IgA I3 (chaîne lourde en rouge, chaîne légère en vert).   | 52 |
| Figure 12 : Mode de liaison parallèle de T1 sur l'anticorps (à gauche) et superposition du polysaccharide sur T1 (à droite).  | 54 |
| Figure 13 : Mode de liaison perpendiculaire de T1 (à gauche) et superposition du polysaccharide sur T1 (à droite).  | 54 |

|  |    |
|--|----|
| Figure 14 : Mode de liaison parallèle de T4 sur l'anticorps (à gauche) et superposition du polysaccharide sur T4 (à droite).   | 56 |
| Figure 15 : Mode de liaison perpendiculaire de T4 sur l'anticorps (à gauche) et superposition du polysaccharide sur T4 (à droite).   | 56 |
| Figure 16 : Meilleure solution d'amarrage du conformère 15 de P100C sur IgAI3. La chaîne légère de l'anticorps est représentée en vert, la chaîne lourde en violet.  | 62 |
| Figure 17 : Position et orientation du conformère 15 de P100C amarré à IgAI3. La chaîne principale des boucles hypervariables de la chaîne légère est représenté en tubes verts, les boucles de la chaîne lourde en tubes violets. Les chaînes latérales de l'anticorps en interaction avec le peptide sont représentées en bâtonnets. | 62 |
| Figure 18 : Solution d'amarrage du conformère 10 de P100C sur IgAI3.   | 63 |
| Figure 19 : Détail des interactions du conformère 10 de P100C avec IgAI3 .   | 63 |
| Figure 20 : Superposition du volume occupé par les résidus DA(E)BC du polysaccharide T4 en orientation parallèle (vert transparent) et du conformère 15 de P100C. Les résidus de P100C en contact avec l'anticorps sont représentés en bâtonnets, les atomes des résidus sans contact sont représentés par des balles.                 | 64 |

## **PARTIE ANALYSE QUANTITATIVE 3D DES RELATIONS STRUCTURE-ACTIVITE**

### **Liste des Tableaux :**

|  |     |
|--|-----|
| Tableau 1 : Structure et activité des composés étudiés.  | 123 |
| Tableau 2 : Statistiques de PLS des modèles utilisant toutes les combinaisons de descripteurs CoMSIA et dérivés de l'ensemble des molécules.               | 126 |
| Tableau 3 : Constitution des jeux d'apprentissage (1) et de test (0)   | 128 |
| Tableau 4 : Statistiques des modèles CoMSIA : valeurs minimales (min), maximales (max) et moyennes (moy) sur l'ensemble des 9 jeux d'apprentissage / test. | 129 |
| Tableau 5 : Proposition de structures visant à confirmer et à enrichir le modèle.  | 144 |

### **Liste des Figures :**

|   |     |
|---|-----|
| Figure 1 : Procédure de validation des modèles [25].  | 111 |
| Figure 2 : Forme générale des potentiels classiques utilisés par CoMFA (traits fins) et du potentiel utilisé par CoMSIA (trait gras).   | 122 |
| Figure 3 : Influence de différentes transformation sur la distribution des données.   | 117 |
| Figure 4 : Structure du noyau Chromone commun à tous nos composés.  | 124 |
| Figure 5 : Alignement des 27 composés utilisés pour l'analyse CoMSIA.   | 124 |
| Figure 6 : Graphes de corrélation du modèle «S,E,D» calculé sur le jeu numéro 3 : (A) corrélation interne sur le jeu d'apprentissage, (B) validation croisée du jeu d'apprentissage et (C) corrélation des prévisions du jeu de test. | 131 |
| Figure 7 : Graphes de corrélation du modèle «S,E,D» sur le jeu de données complet. Corrélation interne à gauche, validation croisée à droite.   | 133 |
| Figure 8 : Le champ stérique occupé par les composés 12 à gauche (95 % d'inhibition) et 20 à droite (94 % d'inhibition). Régions favorables en vert, défavorables en orange.  | 135 |
| Figure 9 : Le composé 18 (25 % d'inhibition) dans le champ stérique. Régions favorables à l'encombrement en vert, régions défavorables en orange.   | 136 |

- Figure 10 : Composés 12 à gauche (95 % d'inhibition) et 20 à droite (94 % d'inhibition) dans le champ électrostatique. Régions électronégatives en rouge, régions électropositives en bleu. 137
- Figure 11 : Composés 20 à gauche (95 % d'inhibition) et 26 à droite (2 % d'inhibition), dans le champ donneur de liaisons hydrogène. Régions favorables en bleu, régions défavorables en violet. 138
- Figure 12 : Superposition des champs stérique du modèle «S,E,D» et hydrophobe du modèle « E,H,D ». Les régions stériques favorables sont en vert opaque, les régions défavorables sont en orange opaque. Les régions jaunes transparentes et bleue transparentes sont respectivement hydrophobes et hydrophiles. 139
- Figure 13: Diagramme de relation structure-activité des flavonoïdes orientées vers l'inhibition de BCRP. 141
- Figure 14 : Diagramme des relations structure-activité des boerhavinones orientées vers l'inhibition de BCRP. 141

# INTRODUCTION

**L**'objet de ce travail est de faire une étude détaillée de méthodes de modélisation appliquées à l'analyse des mécanismes de reconnaissance moléculaire et à la conception de nouveaux composés bioactifs. La faisabilité et l'efficacité de ces méthodes seront démontrées par leur application à des cas concrets issus des activités des laboratoires où s'effectue ce travail. Enfin, les connaissances et le savoir-faire acquis devront être pérennisés par leur diffusion et leur réutilisation dans ces laboratoires.

Les méthodes présentées sont représentatives des deux grands axes de conception de molécules par modélisation : la conception basée sur la structure des récepteurs et la conception basée sur la structure des ligands. Chacune sera illustrée par une application concrète sélectionnée parmi les travaux auxquels j'ai participé en tant qu'ingénieur d'étude en modélisation moléculaire du DPM.

Dans le cadre du premier axe, nous détaillerons une méthode de construction de protéine par homologie et une méthode d'amarrage de petites molécules organiques sur cette protéine. Cette méthode sera appliquée à l'étude des mécanismes de reconnaissance moléculaires d'un antigène polysaccharidique et de mimes peptidiques de cet antigène par un anticorps protecteurs. L'objectif de cette étude est de contribuer à établir les règles de conception de nouveaux vaccins basés sur des mimotopes peptidiques.

Dans le cadre du second axe, nous détaillerons les outils statistiques et les descripteurs chimiques permettant de construire des modèles tridimensionnels quantitatifs des relations structure-activité de composés biologiquement actifs. Cette méthode sera appliquée à l'étude d'une série d'analogues de composés naturels, de type flavonoïde, inhibiteurs du mécanisme de résistance multiple aux anticancéreux développé par les cellules tumorales.

# **CONSTRUCTION DE PROTEINES PAR HOMOLOGIE ET AMARRAGE MOLECULAIRE**

## A - INTRODUCTION

La première grande voie d'étude et de conception de molécules bioactives par modélisation moléculaire est celle qui se fonde sur la structure des récepteurs. Cette approche est basée sur l'exploitation de la structure moléculaire tridimensionnelle de la protéine cible. Les coordonnées atomiques sont principalement issues d'analyses structurales par diffraction des rayons X. Quelques 37.000 structures de protéines sont aujourd'hui disponibles dans la *Protein Data Bank* [1] mais de nombreuses familles ne sont pas représentées. Cependant, lorsque une structure fait défaut, il est parfois possible d'en construire un modèle, grâce aux techniques de construction par homologie. Par ailleurs, les études structurales par résonance magnétique nucléaire fournissent un grand nombre d'informations complémentaires, notamment sur le comportement dynamique des complexes ligand – récepteur en solution.

Lorsque l'on dispose d'un modèle tridimensionnel d'une protéine cible, il est alors possible d'étudier les interactions de ce récepteur avec de petites molécules organiques, telles que le substrat ou le ligand naturel, des activateurs ou des inhibiteurs. Des logiciels d'amarrage recherchent les positions et orientations les plus probables de ces petites molécules en interaction avec la cible et évaluent l'énergie d'interaction de chaque complexe. Ces modèles aident à comprendre le mode d'action des molécules bioactives et à concevoir de nouvelles structures capables d'exploiter au mieux la spécificité du site d'amarrage et le potentiel d'interaction existant.

Nous allons détailler la méthode de construction de protéines par homologie du module COMPOSER du logiciel SYBYL et la méthode d'amarrage du logiciel AUTODOCK3. Dans le cadre d'une collaboration entre l'Institut Pasteur de Paris et le CERMAV de Grenoble, ces méthodes ont été appliquées à l'étude des mécanismes de reconnaissance d'un antigène polysaccharidique et d'un mime peptidique de cet antigène par un anticorps protecteur.

## **B - PRINCIPES ET METHODES**

### ***I - CONSTRUCTION DE PROTEINE PAR HOMOLOGIE***

La construction d'une protéine par homologie est un processus que l'on peut décomposer en 7 étapes majeures :

- Recherche et identification de protéines de structure connue et de séquences homologues à celle qui nous intéresse (le modèle).
- Identification des régions conservées, communes aux structures de référence sélectionnées.
- Alignement des structures par les régions conservées et alignement de la séquence cible.
- Construction des régions structurellement conservées.
- Construction des régions variables.
- Affinement du modèle (optimisation de la géométrie).
- Contrôle de la conformation des liaisons de la chaîne principale.

Ce chapitre présente une méthode de recherche des séquences homologues avec les logiciels BLAST (*Basic Local Alignment Search Tool*), la construction d'un modèle de protéine par la méthode de Blundell [2-5] implémentée dans le module COMPOSER de SYBYL, l'affinement du modèle par un champ de forces (autres modules de SYBYL) et une première validation de la géométrie du modèle par PROCHECK.

#### **1. RECHERCHE ET IDENTIFICATION DES PROTEINES HOMOLOGUES**

La *Protein Data Bank* (PDB) est un répertoire mondial de dépôt d'informations sur la structure tridimensionnelle des protéines et des acides nucléiques [1]. Ces molécules proviennent de l'ensemble des règnes biologiques. Les structures tridimensionnelles sont issues principalement d'analyses par diffraction des rayons X, les autres d'analyses par résonance magnétique nucléaire (RMN) ou de modélisations moléculaires. La PDB est gratuitement accessible par Internet et contient un grand nombre d'informations complémentaires comme la séquence ou la phylogénie des macromolécules. Les séquences d'acides aminés sont codées selon une procédure

qui fait correspondre une lettre de l'alphabet à chaque acide aminé (Annexe I p.72). La recherche de protéines homologues est basée sur la comparaison des séquences d'acides aminés des protéines de structures connues avec la séquence d'acides aminés de la protéine que l'on veut construire. Cette recherche peut se faire par différentes méthodes telles que celle de Needleman & Wunsch [6], FASTA (*Fast Alignment*) [7] ou BLAST (*Basic Local Alignment Search Tool*) [8].

Les algorithmes FASTA et BLAST permettent de comparer en un temps très court une ou plusieurs séquences cibles aux milliers de séquences contenues dans les bases de données telles que la PDB. Ils procèdent par comparaisons de paires d'acides aminés. Chaque comparaison obtient un score qui reflète l'identité ou le degré de similarité entre les séquences comparées. Plus le score est élevé, plus le degré de similarité est important. Le résultat final se présente sous la forme d'un alignement des séquences trouvées accompagnées de scores d'identité et de similarité. L'alignement peut être global ou local selon le paramétrage utilisé. L'alignement global est l'alignement optimal qui englobe tous les caractères de chaque séquence alors que l'alignement local est l'alignement optimal des régions les mieux conservées.

Les programmes issus de l'algorithme BLAST comparent les séquences par des alignements locaux. BLAST découpe la séquence cible et les séquences de la librairie en fragments appelés « mots » et commence par aligner ces mots entre eux. Le premier alignement est réalisé avec un mot de taille « M » ayant un score minimal « T » pour une matrice de substitution donnée. Les mots obtenant les meilleurs scores sont alors étendus dans les deux directions pour obtenir un nouveau score. L'extension se poursuit tant que le score résultant reste au dessus d'une valeur seuil. On obtient ainsi un alignement local des séquences de la librairie avec la séquence cible.

Le score des alignements est calculé position par position. Lorsque les acides aminés présents sur une position sont identiques, la paire obtient le score maximum. Lorsqu'ils sont différents, le score diminue selon les valeurs d'une matrice de substitution. Pour chaque substitution possible, cette matrice contient une valeur qui reflète la probabilité que l'acide aminé de la séquence analysée puisse muter vers celui de la séquence cible. Il existe différentes matrices de substitution (PAM, BLOSUM ou PSSM ...). Le score d'alignement de deux séquences est la somme des scores de chaque position.

Ces matrices empiriques sont construites à partir d'un échantillon étendu et varié de paires d'acides aminés provenant d'alignements validés. Lorsque le volume du jeu de données devient statistiquement significatif, l'échantillon devient représentatif de la population et la matrice reflète, de façon fiable, la probabilité qu'une mutation donnée survienne au cours d'une période

d'évolution. Le choix d'une matrice de substitution repose en grande partie sur l'expérience de l'utilisateur. Cependant la matrice BLOSUM62 [9] est l'une des plus polyvalentes et des plus couramment utilisées.

Il arrive, lors d'un alignement, qu'à un ou plusieurs acides aminés d'une chaîne corresponde un espace vide dans l'autre. Cet écart est appelé un « gap ». C'est le résultat d'une mutation génétique qui a entraîné l'insertion ou la délétion d'un nombre variable d'acides aminés. Ce type de mutation est généralement le signe d'un écart important entre deux séquences du point de vue phylogénique. Cela implique que la seule présence d'un *gap* a plus de signification que sa longueur. Par conséquent, la présence d'un gap entre deux séquences pénalisera plus fortement leur score d'alignement que le malus attribué à chaque acide aminé surnuméraire. Par contre il n'y a pas de consensus quand au coût à attribuer à cette pénalité qui est calibré différemment selon les valeurs de la matrice de substitution et selon les développeurs.

L'algorithme PSI-BLAST (*Position Specific Iterative BLAST*) donne la possibilité de relancer itérativement BLAST avec les séquences résultats. A chaque nouvelle itération, celles-ci sont transformées en « profils » qui sont recherchés à leur tour dans la banque de séquences [10]. Les itérations s'arrêtent lorsqu'il y a convergence, c'est à dire lorsque les séquences résultats de l'itération  $n$  sont identiques à celles de l'itération  $n-1$ .

Ces algorithmes d'alignement partiel peuvent malheureusement produire des artefacts c'est à dire attribuer de bons scores à des séquences qui présentent en réalité peu de similarité avec la cible. Pour évaluer la probabilité qu'un résultat soit un artefact, le score est accompagné de deux valeurs : la valeur attendue (*Expected value*) et la valeur prévisionnelle (*Predictive value*). La valeur attendue est calculée avec des séquences aléatoires, de même longueur que la séquence comparée à la cible et de composition similaire. C'est la probabilité qu'une de ces séquences aléatoires ait un score supérieur ou égal au score de l'alignement étudié. Plus la valeur E est faible, plus l'alignement retenu est significatif. La valeur prévisionnelle est la probabilité qu'il existe un alignement ayant un meilleur score que l'alignement étudié [11, 12]. Plus cette valeur est proche de 0 (zéro), plus l'alignement retenu est significatif.

Un autre facteur peut induire la fonction de score en erreur : les séquences de faible complexité. Ce sont des zones où l'on rencontre la répétition d'un même acide aminé ou d'un motif peu complexe. Si l'on compare deux de ces régions, dopées par les mêmes acides aminés, on a de fortes chances d'obtenir un bon score, quel que soit l'alignement proposé. Ce score ne sera pas dû à la similarité de séquences précises mais au hasard et à la similarité de composition des

segments [13, 14]. Les algorithmes de BLAST comportent donc un filtre qui permet d'écartier ces segments de faible complexité [15].

Les méthodes d'alignement local ont l'avantage d'être plus sensibles que celles d'alignement global. En effet, elles permettent de mettre en évidence l'homologie de séquences dont le pourcentage d'identité est faible. Les régions structurellement conservées sont bien identifiées et alignées. Dans les zones moins bien conservées, les matrices de substitution permettent de quantifier les écarts et donc de quantifier l'homologie ou la « parenté » entre deux séquences.

Une fois qu'une famille d'homologues est identifiée, pouvant ne comporter qu'un seul membre, les fichiers de coordonnées atomiques de ces protéines sont insérés dans le module PRODAT de SYBYL. PRODAT est la banque de structures locale du logiciel à partir de laquelle le module COMPOSER nous permet de construire des modèles.

## **2. IDENTIFICATION DES REGIONS CONSERVEES, COMMUNES AUX STRUCTURES DE REFERENCE SELECTIONNEES**

Après une recherche de la séquence cible sur la banque PRODAT, COMPOSER affiche les séquences alignées des protéines présentant les meilleurs scores d'homologie. Ce premier alignement est réalisé par la méthode de Needleman & Wunsch [6]. Les séquences de faible homologie ou ayant une mauvaise résolution atomique au niveau de leur chaîne principale sont écartées. Les séquences présentant moins de 30% d'identité sont difficilement exploitables.

Une fois qu'une famille de protéines homologues de la cible est identifiée, il faut localiser les régions structurellement conservées de cette famille (*Structurally Conserved Regions - SCRs*). Le module COMPOSER demande la localisation d'au moins trois SCRs. Chaque SCR comporte un ou plusieurs acides aminés identiques sur des positions équivalentes dans chaque protéine. La présence de gaps dans ces zones est très fortement pénalisée et doit absolument être évitée. COMPOSER refuse d'ailleurs l'utilisation de SCR comportant des gaps. Le logiciel aligne les séquences de la famille deux à deux et sélectionne le couple de séquences possédant le plus fort pourcentage d'identité. Il définit chaque région ne comportant aucun gap comme un SCR initial. COMPOSER aligne ensuite les autres séquences avec la même méthode. Si certaines séquences présentent des gaps au niveau des SCRs, ces SCRs sont redéfinis de façon à exclure tous les gaps. Les acides aminés surnuméraires, c'est à dire présents sur des positions qui n'existent pas dans d'autres séquences, sont également pénalisés. L'utilisateur a la possibilité d'ajuster cette pénalité.

Lorsque toutes les séquences sont alignées, on sélectionne les résidus qui sont identiques à travers l'ensemble de la famille et un poids relatif est affecté à chaque structure. Ce poids est proportionnel au carré du pourcentage d'identité d'une séquence par rapport à la séquence cible. Il va déterminer la contribution relative de chaque structure dans la construction de l'armature du modèle.

### **3. ALIGNEMENT SPATIAL DES STRUCTURES ET ALIGNEMENT DE LA SEQUENCE CIBLE**

COMPOSER prend un premier jeu de résidus topologiquement équivalents pour aligner tous les homologues de la famille. Il localise les SRCs et calcule la géométrie moyenne de leur chaîne principale pour constituer l'armature des SCR du modèle. Seuls les carbones  $\alpha$  sont utilisés pour cette étape. COMPOSER est donc très tolérant vis-à-vis des imperfections présentes dans ce premier jeu de résidus. La construction de cette armature est un processus itératif qui minimise les écarts entre chaque carbone  $\alpha$  de l'armature et les carbones  $\alpha$  correspondants dans les structures de référence. Ce processus prend en compte le poids relatif de chaque structure calculé à l'étape précédente. Il exclut progressivement des SRCs les résidus dont la position chez certains homologues est trop éloignée de l'armature composite. Après un affinement des SRCs visant à ne conserver que les résidus identiques, un second affinement est opéré sur les bases géométriques des structures homologues.

Une fois que l'on connaît les SRCs de la famille d'homologues, il faut localiser les régions correspondantes dans la séquence cible. Pour y parvenir, on aligne la séquence cible sur les SRCs des homologues. La présence de gaps au niveau des SRCs dans la séquence cible est fortement pénalisée et entraîne une redéfinition des SRCs. Les outils d'alignement automatique étant perfectibles, un contrôle visuel des alignements est nécessaire. Pour chaque SCR, la structure présentant le meilleur pourcentage d'identité avec la cible est sélectionné. En cas d'égalité entre plusieurs structures, les scores d'alignement sont comparés.

### **4. CONSTRUCTION DES REGIONS STRUCTURELLEMENT CONSERVEES**

La construction proprement dite des SRCs peut alors commencer. L'armature composite construite précédemment doit être affinée car elle correspond à une position moyenne des carbones  $\alpha$ . COMPOSER ajuste sur chaque SCR composite, le fragment de chaîne primaire du meilleur homologue de ce SCR. Les carbones  $\alpha$  du fragment sont superposés à l'armature composite du SCR par la méthode d'ajustement des moindres carrés (RMS fit). Lors de

l'ajustement, il est possible de pondérer chaque carbone en fonction de la conservation du résidu à la position qu'il occupe sur l'ensemble de la famille d'homologue.

Les chaînes latérales des acides aminés sont ensuite ajoutées. Leur conformation est déterminée à partir d'une base de données tenant compte de la structure secondaire de la chaîne principale et de la conformation des chaînes des résidus sur la position équivalente des homologues.

COMPOSER procède en trois étapes. Il localise les résidus correspondants dans chaque homologue de la famille et calcule pour chacun un score de substitution qui reflète le nombre d'atomes compatibles entre ce résidu homologue et le résidu cible. S'il n'y a qu'un seul meilleur score, il est sélectionné. S'il y en a deux, l'un des deux est pris au hasard. S'il y en a plus de deux, COMPOSER fait une analyse médiane des conformations des chaînes latérales et sélectionne le résidu le plus proche de la médiane comme étant le plus représentatif du groupe.

Le résidu sélectionné comme source est alors ajusté sur l'armature du modèle par les 3 atomes N, C $\alpha$ , CO.

Les coordonnées des atomes de la chaîne latérale compatibles avec le résidu cible sont copiées telles quelles. Les coordonnées des atomes restant sont issues d'une base de conformations canoniques. Cette base contient des conformères de tous les résidus dans différentes structures secondaires.

## 5. CONSTRUCTION DES REGIONS VARIABLES

Les régions variables sont les domaines compris entre les régions conservées. Ce sont généralement des boucles. Pour construire chaque boucle, il faut trouver un fragment compatible avec le reste du modèle et dont la séquence est la plus proche de la cible. D'une façon générale, le premier critère de sélection est la taille de la boucle. Tout homologue possédant une boucle de même taille que la cible fournit un fragment intéressant dans la mesure où ce fragment est soumis à des contraintes d'environnement similaires dans l'homologue et dans le modèle. S'il n'existe pas de fragments de même longueur mais que la boucle cible est supposée appartenir à une classe de boucles définie, il faut sélectionner un fragment appartenant cette classe. Finalement, si l'on ignore à quelle classe appartient la boucle cible, il faut chercher des fragments dont la géométrie est compatible avec celle des régions conservées qui bordent la boucle. La géométrie et la séquence des fragments candidats seront les seuls guides pour faire un choix.

Cette méthode utilisant des fragments de boucles connues pour construire les boucles d'un modèle à été décrite par Jones et Thirup [16] et développée par Claessens [17].

Lorsque l'on parcourt la base de données à la recherche de fragments de boucles, COMPOSER utilise comme critère de recherche les coordonnées des C $\alpha$  des résidus du modèle qui bordent la boucle. Ces résidus constituent les zones d'ancrage de la boucle aux SCRs. Ils sont au nombre de trois à chaque extrémité ce qui permet d'appréhender l'environnement généré par les structures secondaires qui encadrent la boucle. Les coordonnées des atomes d'ancrage peuvent provenir du modèle ou de l'armature composite. Cette dernière solution permet d'adoucir les écarts entre les extrémités des SCRs des structures homologues.

Lorsque que l'on insère les fragments de boucle au modèle, les coordonnées des résidus d'ancrage peuvent, eux aussi, provenir de deux sources : le modèle ou la base de données de fragments. Ces sources ne sont pas identiques et il y a deux façons de procéder pour ajuster les coordonnées de ces zones de chevauchement. La première consiste à calculer les coordonnées des résidus d'ancrage par une moyenne pondérée des coordonnées du fragment et du modèle. Le poids des coordonnées du fragment augmente lorsqu'on se rapproche de la boucle. La seconde, sans doute préférable, consiste à faire des ajustements des angles de torsion des résidus du fragment de la base pour superposer les résidus d'ancrage.

L'algorithme de construction est le suivant :

- 1 - Sélectionner les fragments de même longueur que la cible (étape optionnelle). S'il y en a, aller à l'étape 3.
- 2 - Chercher les fragments dont la longueur se rapproche le plus de celle de la boucle cible. Sélectionner les fragments dont la distance entre les extrémités est acceptable (tolérance modifiable par l'utilisateur).

Calculer pour chaque fragment la distance entre les C $\alpha$  des extrémités des SCRs et les comparer à la distance correspondante dans le modèle. Les fragments dont le RMS dépasse le seuil de tolérance sont écartés.

Ajuster les fragments aux SCRs par les régions d'ancrage et écarter les fragments mal ajustés ou incompatibles avec la construction ultérieure des ponts disulfures nécessaires.

- 3 - Ajuster les coordonnées des atomes d'ancrage par la méthode de transition ou par ajustement des angles de torsion du fragment.
- 4 - Construire les chaînes latérales suivant la même méthode que pour les régions conservées.
- 5 - Calculer un score d'homologie de séquence avec la boucle cible pour chaque fragment. Classer les boucles par score d'homologie puis par RMS d'ajustement. La présence de gaps dans les boucles entraîne des pénalités dans les scores d'alignement mais ces

pénalités sont beaucoup moins fortes que dans les régions conservées. On peut alors sélectionner le fragment ayant le meilleur score pour construire la boucle.

Le modèle brut est terminé et COMPOSER génère le fichier des coordonnées atomiques et de connectivité du modèle. Ce fichier est au format standard de SYBYL : le format *mol2*.

## **6. AFFINEMENT DU MODELE**

Le modèle brut est encore incomplet. Il reste à construire les ponts disulfures connus de la structure, à ajouter les hydrogènes et à calculer les charges partielles des atomes.

L'ajout des ponts disulfures peut se faire soit dans le module COMPOSER soit dans le module BIOPOLYMER de SYBYL. L'ajout des hydrogènes et des charges nécessite le module BIOPOLYMER.

Une fois que tous les hydrogènes sont en place, on peut optimiser la géométrie du modèle par minimisation de son énergie. Une simple évaluation de l'énergie permet de repérer visuellement les zones de contact impropres grâce à une coloration de la structure en fonction de l'énergie de ses éléments. Une minimisation locale de l'énergie des chaînes latérales en contact permet d'améliorer leurs conformations. Cette minimisation ne porte que sur les interactions de Van der Waals.

Les charges sont ensuite ajoutées. Pour les protéines, les jeux de charges généralement utilisés sont les jeux Kollman-UNI et Kollman-ALL [18].

## **7. VALIDATION PARTIELLE DU MODELE**

La stéréochimie et la géométrie du modèle construit doivent être contrôlées. Le logiciel PROCHECK [19] effectue un contrôle global à partir de paramètres expérimentaux [20] et indique les acides aminés ayant une géométrie improbable. Les principaux paramètres contrôlés sont les longueurs des liaisons, les angles de valence et les angles dièdres de la chaîne principale, la chiralité des centres asymétriques et la planéité des groupements aromatiques ou conjugués. Ce logiciel produit différents graphiques simples et faciles à interpréter.

## II - AMARRAGE AVEC AUTODOCK

Il existe de nombreux outils d'amarrage automatique ou *docking*, qui se répartissent en deux grandes catégories selon la méthode utilisée : l'ajustement de fragments (*matching*) ou la simulation de trajectoire. Les méthodes par ajustement de fragment commencent par construire un modèle « en négatif » du site actif. Ce modèle est constitué par le volume accessible du site et les points d'interaction tels que les sites de liaison hydrogène, les charges ou les sites lipophiles. Les ligands sont décomposés en fragments puis reconstruits dans le modèle négatif du site, en essayant de faire correspondre les géométries ainsi que les fonctions chimiques. Des logiciels tels que DOCK [21, 22] ou SURFLEX [23] utilisent cette approche qui permet notamment un criblage rapide de vastes bibliothèques de composés.

La seconde approche, par trajectoire, est plus précise : à partir d'une position initiale aléatoire, à l'extérieur du site actif, le ligand explore le site étudié par la répétition successive de mouvements et d'évaluations de l'interaction ligand-récepteur. Les mouvements sont effectués par des opérations de translation, de rotation et de changement de conformation. L'énergie d'interaction est calculée par une fonction énergétique. Les mouvements du cycle à venir sont guidés par les variations d'énergie induites par les mouvements des cycles précédents. L'algorithme s'arrête lorsqu'il a trouvé la position idéale du ligand dans le récepteur. Ces techniques sont plus lentes que celles par *matching* mais prennent mieux en compte la flexibilité du ligand et permettent l'exploration de régions plus vastes. AUTODOCK [24, 25] fait parti de la seconde catégorie. Il combine une méthode d'évaluation de l'énergie à partir de grilles de potentiels et différentes méthodes d'exploration allant de la dynamique moléculaire (méthode Métropolis : recuit simulé Monte Carlo - RSMC) aux algorithmes génétiques. Nous allons détailler les principes et le fonctionnement de la dernière version de ce logiciel : AUTODOCK 3.0.5

### 1. MODELISATION DU POTENTIEL ENERGETIQUE

Pour calculer l'énergie libre du complexe ligand-récepteur, AUTODOCK utilise les termes d'un champ de force traditionnel auxquels sont ajoutés deux termes liés à l'entropie. L'énergie est donnée par l'Équation 1.

Équation 1: 
$$\Delta G = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{tor} + \Delta G_{sol}$$

Les trois premiers termes sont des termes classiques de mécanique moléculaire ; il s'agit respectivement des énergies : dispersion / répulsion des atomes, liaisons hydrogène et interactions électrostatiques.  $\Delta G_{tor}$  est un terme qui traduit l'augmentation d'énergie du système due à la restriction des rotors libres du ligand et à la restriction des rotations et translations du ligand lors de la complexation au récepteur. Cette perte de degrés de liberté est une perte d'entropie.  $\Delta G_{sol}$  est un autre terme lié à l'entropie qui décrit les variations d'énergie du système lors de la désolvation du ligand au moment de la complexation au récepteur. C'est une modélisation partielle de ce que l'on appelle l'effet hydrophobe dû aux variations de l'entropie du solvant aux interfaces solvant – soluté.

L'énergie libre calculée par l'intermédiaire de champs de forces ne traduit malheureusement pas directement la stabilité du complexe. Comparer les énergies libres des complexes d'un récepteur avec différents ligands devient hasardeuse dès que les ligands diffèrent de quelques atomes. Le passage par une relation empirique de type QSAR est nécessaire pour relier la structure des complexes et l'énergie libre de liaison. Le modèle empirique utilisé dans AUTODOCK est une régression linéaire multiple des différents termes de l'équation d'énergie libre. Chaque terme est ainsi pondéré par un coefficient dérivé d'un jeu étendu de complexes protéine–inhibiteur pour lesquels la constante d'inhibition  $K_i$  est connue.

La relation entre la constante d'inhibition et l'énergie libre de liaison est donnée par l'Équation 2.

Équation 2: 
$$\Delta G_{obs} = RT \ln K_i$$

ou R est la constante des gaz parfaits ( $1,987 \text{ cal.K}^{-1}.\text{mol}^{-1}$ ) et T la température absolue.

Pour calculer la valeur des différents termes de l'Équation 1, AUTODOCK utilise l'approche de Wesson et Eisenberg [26] qui décompose l'énergie libre de liaison de la façon suivante :

Équation 3:

$$\begin{aligned} \Delta G = & C_{vdw} \cdot \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) \\ & + C_{hbond} \cdot \sum_{i,j} E(t) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond} \right) \\ & + C_{elec} \cdot \sum_{i,j} \frac{Q_i Q_j}{\epsilon(r_{ij}) r_{ij}} \\ & + C_{tor} \cdot N_{tor} \\ & + C_{sol} \cdot \sum_{i,j} S_i V_j e^{(r_{ij}^2 / 2\sigma^2)} \end{aligned}$$

Les cinq termes  $C_{xxx}$  sont les coefficients empiriques déterminés par l'analyse en régression linéaire d'un jeu de complexes protéine–ligand de constantes de liaison connues. Les sommes sont faites sur l'ensemble des couples formés d'un atome du ligand  $i$  et d'un atome de la protéine  $j$  et sur toutes les paires d'atomes du ligand séparés par au moins 3 liaisons covalentes.

Le terme de dispersion / répulsion est modélisé par un potentiel de Lennard-Jones 12-6.

Les liaisons hydrogènes sont modélisées par un potentiel directionnel de Lennard-Jones en 12-10 auquel on ajoute une constante de désolvatation ( $E_{\text{hbond}}$ ). Ce potentiel est pondéré en fonction de l'angle  $t$  formé entre la direction du doublet libre de l'accepteur et la liaison entre l'hydrogène et l'atome polaire qui le porte. L'angle optimum est de  $180^\circ$  alors que pour un angle inférieur à  $90^\circ$  la liaison devient impossible et la fonction  $E(t)$  devient nulle. La constante ( $E_{\text{hbond}}$ ) représente l'énergie moyenne estimée d'une liaison hydrogène entre une molécule d'eau du solvant et un atome polaire du ligand afin de prendre en compte l'énergie de désolvatation des atomes polaires lorsqu'ils entrent en contact avec la protéine. Elle pénalise en fait les atomes polaires du ligand qui ne reforment pas de liaisons hydrogène avec le récepteur, favorisant ainsi les solutions exploitant au mieux le potentiel de liaisons hydrogène du site actif.

Les interactions électrostatiques sont données par le potentiel de Coulomb où  $Q_i$  et  $Q_j$  sont les charges partielles des atomes d'une paire,  $r_{ij}$  est la distance entre les charges et  $\epsilon(r_{ij})$  est une fonction diélectrique qui modélise l'effet d'écran généré par la polarité du milieu environnant (le solvant).

La perte d'entropie du ligand lorsqu'il se lie au récepteur, est prise en compte dans la fonction de l'énergie. Ce terme est proportionnel au nombre de liaisons  $sp^3$  dans le ligand  $N_{\text{tor}}$  [27]. Il traduit la restriction des degrés de liberté conformationnelle.

L'énergie nécessaire à la désolvatation du ligand lors de l'association au récepteur est rapportée par le dernier terme de l'équation. Ce terme est calculé par une variante de la méthode de Stouten [28]. C'est une méthode, basée sur les volumes, permettant de calculer la contribution de chaque atome à l'énergie totale de désolvatation du ligand. Pour chaque atome du ligand, on calcule le volume partiel des atomes du récepteur qui l'entourent et l'on pondère ce volume par une fonction exponentielle. La somme de ces volumes partiels donne la proportion  $V_j$  du volume autour de l'atome du ligand qui est occupé par des atomes du récepteur. Ce pourcentage est finalement pondéré par le paramètre de solvation atomique  $S_i$  de l'atome pour donner l'énergie de solvation. La mise au point de ce terme a conduit les développeurs d'AUTODOCK à n'évaluer que les carbones du ligand, en faisant la distinction entre les carbones aliphatiques et les carbones aromatiques. L'évaluation des atomes polaires (N et O) du ligand, qui forment des

liaisons hydrogènes avec le solvant, est intégrée au calcul de l'énergie de formation des liaisons hydrogène (constante  $E_{\text{hbond}}$ ).

## 2. CALCUL DU POTENTIEL ENERGETIQUE

AUTODOCK3 est un logiciel de simulation d'amarrage de ligand flexible sur un récepteur rigide. Quelle que soit la technique de simulation employée, un très grand nombre d'évaluations d'énergie est nécessaire tout au long de la simulation. Le calcul de l'énergie d'interaction doit donc être extrêmement rapide si l'on veut que la simulation se déroule en un temps acceptable avec des moyens de calcul courants. Le récepteur étant rigide, son champ d'interaction moléculaire est constant. D'autre part, la nature du potentiel énergétique utilisé (cf § 1 p.23) permet un découpage du calcul selon les différents types d'atomes présents. Ainsi, l'évaluation rapide de l'énergie peut se faire par l'intermédiaire de grilles de potentiel d'affinité atomique précalculées pour chaque type d'atome [29]. Une grille est une matrice tridimensionnelle qui englobe l'intégralité ou une région intéressante du récepteur étudié (Figure 1). L'espacement entre les points peut aller de 0,2 Å à 1,0 Å. Chaque point de la grille enregistre le potentiel d'interaction entre une sonde, constituée d'un ou plusieurs atomes, et l'ensemble des atomes du récepteur. Le calcul de ces grilles est effectué par AUTOGRID, un programme d'AUTODOCK, qui calcule automatiquement les grilles nécessaires parmi les types C, N, O et H. Une grille supplémentaire est calculée pour le potentiel d'interaction électrostatique avec pour sonde une charge ponctuelle de +1 eV.

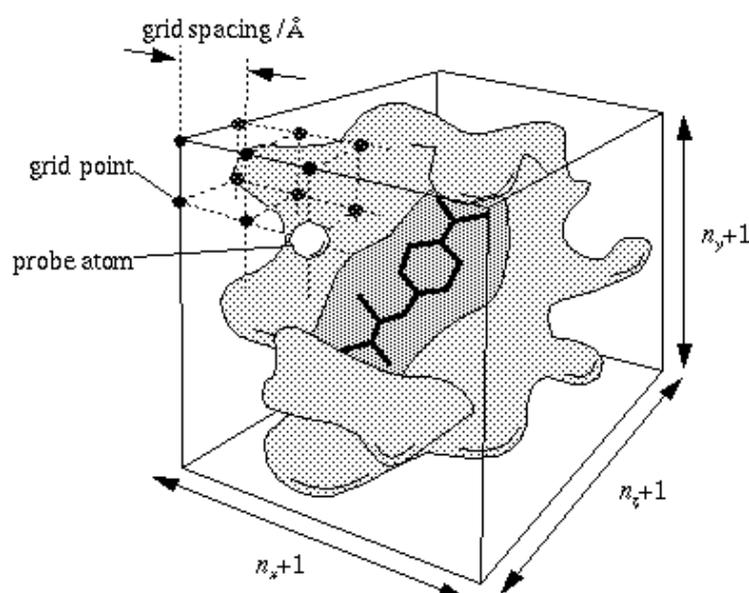


Figure 1: Schéma d'une grille de potentiel englobant le site actif d'une macromolécule.

Les dimensions de la matrice définissent une boîte qui sera la zone à explorer par le ligand et AUTODOCK donnera la configuration la plus stable comme solution d'amarrage dans ce volume. La boîte doit englober le site d'intérêt mais le temps de calcul pour l'explorer sera fonction de ses dimensions. Le volume explorable est donc limité par le temps et la puissance de calcul dont on dispose.

L'énergie d'interaction du ligand dans une configuration particulière est évaluée en faisant une interpolation tri-linéaire des valeurs d'affinité des huit points de grille qui entourent chaque atome du ligand. Le temps de calcul de l'énergie dépend uniquement du nombre d'atomes dans le ligand et est indépendant de la taille de la macromolécule.

### **3. EXPLORATION DE L'ENVIRONNEMENT SPATIAL ET CONFORMATIONNEL.**

#### ***a. Les Algorithmes Génétiques (AG).***

Les algorithmes génétiques reprennent des mécanismes et la terminologie de la génétique naturelle et de l'évolution biologique. L'organisation d'un ligand en complexe avec un récepteur peut être définie par un jeu de paramètres décrivant la position, l'orientation et la conformation du ligand par rapport au récepteur. Ces paramètres sont les « variables d'état » et dans un AG, chaque variable d'état correspond à un gène. La valeur de ces variables correspond au génotype et les coordonnées atomiques associées correspondent au phénotype. Chaque état défini du ligand correspond à un individu. Pour faire évoluer le ligand dans le champ d'interaction du récepteur, on emploie un certain nombre d'opérateurs qui vont agir sur le génotype. Le phénotype qui en découle sera amélioré par une méthode de sélection. Des paires aléatoires d'individus sont combinées selon le principe du croisement (*crossover*) pour donner des individus fils qui héritent de gènes provenant de leurs deux parents. D'autre part, certains enfants peuvent être le résultat de mutations où un gène est modifié de façon aléatoire. La sélection des individus fils constituant la nouvelle génération est basée sur la qualité de leur interaction avec le récepteur : les solutions qui s'ajustent mieux au récepteur que leurs parents persistent alors que les autres disparaissent. Le critère d'évaluation de la qualité d'une solution d'amarrage est l'énergie d'interaction totale du système ligand – récepteur.

Pour l'implémentation de l'AG dans AUTODOCK, le chromosome est composé de gènes de valeur réelle :

- trois coordonnées cartésiennes pour la position du ligand,
- trois coordonnées cartésiennes pour définir un point sur l'axe principal de la molécule ;
- une valeur pour l'angle de rotation du ligand autour de l'axe principal ;
- une valeur d'angle pour chaque pivot en libre rotation dans le ligand.

Les trois coordonnées définissant l'axe principal et l'angle de rotation autour de cet axe constituent le « quatérior » d'orientation. L'ordre des gènes qui encodent les angles de torsion est défini par un arbre de torsion créé par AUTOTORS, un programme de paramétrage inclus dans AUTODOCK. AUTOTORS permet de sélectionner les rotors flexibles du ligand. L'ensemble de ces variables ou gènes constitue le chromosome du ligand. L'implantation de valeurs dans ces gènes donne un état du ligand appelé individu.

L'algorithme génétique commence par créer une population aléatoire d'individus, dont la taille c'est à dire le nombre d'individus, est définie par l'utilisateur. Les différents gènes de chaque individu reçoivent des valeurs aléatoires comprises dans les limites de l'espace à explorer : les coordonnées de position sont localisées à l'intérieur de ce volume, les coordonnées du quatérior permettent toute orientation du ligand dans ce volume et les pivots peuvent prendre n'importe quelle valeur d'angle entre  $-180^\circ$  et  $+180^\circ$ . Tout individu dont le génotype se traduit par la présence d'atomes hors du volume à explorer est éliminé. La définition de ce volume survient au moment du calcul des grilles de potentiel décrit dans le paragraphe 2 p. 26.

Après la création aléatoire de la première population, le cycle de génération est répété jusqu'à ce que soit atteint le nombre maximum de générations ou le nombre maximum d'évaluations de l'énergie. Le cycle de génération se décompose en cinq étapes : transcription du génotype en phénotype (*mapping*) avec évaluation de l'énergie d'interaction ligand – récepteur, sélection, croisement, mutation et sélection élitiste. Chaque étape s'applique à l'ensemble des individus de la population. Pour les algorithmes génétiques lamarckiens (AGL), ce cycle est suivi d'une optimisation locale dont les détails sont l'objet du paragraphe b p.30.

La transcription est la phase de lecture du génotype, de sa traduction en coordonnées atomiques (le phénotype) et d'enregistrement. Elle s'applique à tous les individus de la population et permet l'évaluation de l'« adaptation » des individus au récepteur. Cette évaluation se fait sur la base de l'énergie interne du ligand et de son énergie d'interaction avec le récepteur. Plus l'énergie est basse, plus l'interaction est forte et stable. La nature physicochimique de la fonction d'évaluation de l'énergie est décrite au paragraphe 1. Chaque fois que l'énergie d'un individu est calculée, que ce soit au cours de la recherche globale ou de l'optimisation locale (voir AG Lamarckiens, § b p.30), le compteur du nombre total d'évaluations de l'énergie est incrémenté. Ce compteur est

l'un des paramètres d'arrêt conditionnel de l'exploration.

La phase de sélection permet de déterminer le nombre d'enfants qu'aura chaque individu dans la génération suivante. Ainsi les individus qui auront une meilleure interaction que la moyenne avec le récepteur, auront proportionnellement plus d'enfants. Le nombre d'enfants est donné par l'Équation 4 :

$$\text{Équation 4:} \quad n_0 = \frac{E_p - E_i}{E_p - E_m} \quad E_p \neq E_m$$

où  $n_0$  est le nombre entier d'enfants qu'aura l'individu  $i$ ,  $E_i$  est l'énergie d'interaction de l'individu  $i$  considéré,  $E_p$  est l'énergie d'interaction de l'individu le plus mal ajusté au récepteur parmi les  $N$  dernières générations et  $E_m$  est l'énergie d'interaction moyenne de la génération en cours.  $N$  est défini par l'utilisateur et vaut 10 par défaut. L'énergie de l'individu le plus mal ajusté étant toujours supérieure à l'énergie moyenne et à l'énergie de l'individu en cours de sélection, les individus ayant un meilleur ajustement que la moyenne auront au moins un enfant. AUTODOCK attend l'égalité  $E_p = E_m$  pour considérer que la population a convergé vers la meilleure solution ; c'est l'une des conditions d'arrêt de l'algorithme.

Les croisements et mutations s'opèrent sur un nombre aléatoire d'individus de la population selon un taux de croisement et de mutation défini par l'utilisateur. Les croisements ont lieu en premier. Ils sont effectués par deux points de coupure sur des positions identiques des chromosomes parents suivit de l'échange de fragments. Ainsi le chromosome de chaque parent est découpé en trois fragments contenant un ou plusieurs gènes, par exemple : ABC pour l'un des parents et abc pour l'autre. Les chromosomes des enfants après un croisement en deux points seront : AbC et aBc. Ces enfants remplacent alors leurs parents dans la population pour garder une taille de population constante.

Les mutations sont obtenues par l'ajout, à la valeur d'un gène, d'une grandeur réelle dont la probabilité suit la loi de distribution de Cauchy ou loi de Lorentz :

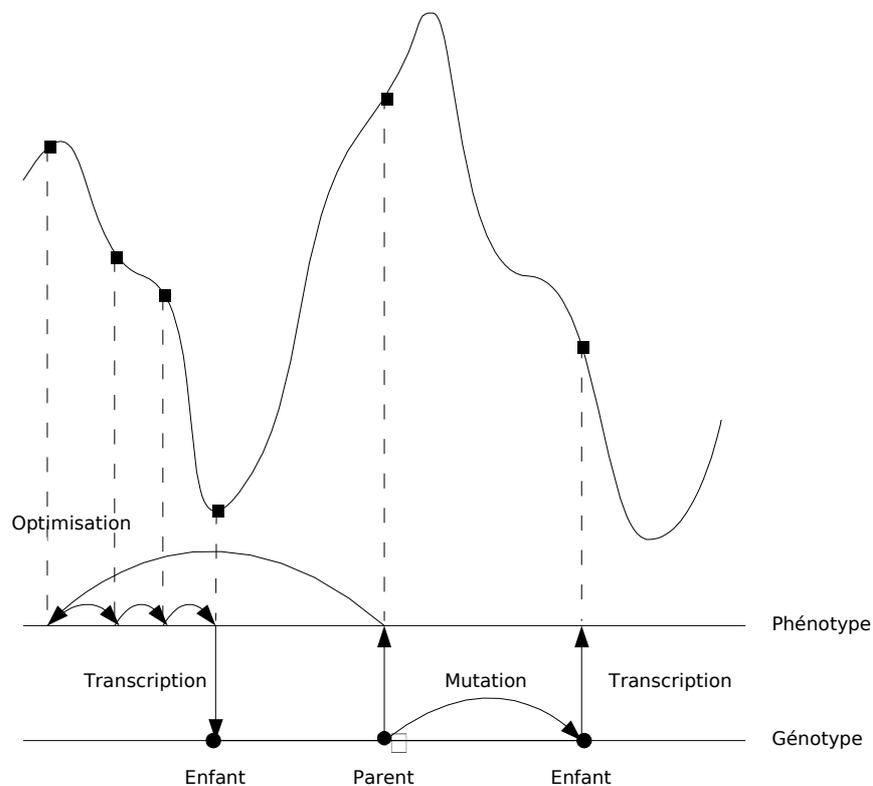
$$\text{Équation 5:} \quad C(\alpha, \beta, x) = \frac{\beta}{\pi(\beta^2(x + \alpha)^2)}$$

où  $\alpha$  et  $\beta$  sont des paramètres affectant la moyenne et la variance de la distribution. Cette loi de distribution favorise les petites déviations, centrées sur la moyenne mais permet des variations de grande amplitude avec plus de probabilité que n'en donne une loi Normale. La distribution de Cauchy ressemble à une distribution gaussienne très « aplatie ».

La sélection élitiste est un paramètre défini par l'utilisateur qui indique combien des meilleurs individus survivent automatiquement à la génération suivante. Par défaut, cette valeur est 1 : le meilleur uniquement.

### ***b. L'Algorithme Génétique Lamarckien (AGL).***

La plupart des algorithmes génétiques reproduisent le comportement de l'évolution darwinienne en appliquant le principe de la génétique de Mendel c'est-à-dire le transfert à sens unique des informations du génotype vers le phénotype. Ce comportement est illustré par la partie droite de la Figure 2. Par contre, dans le cas où il existe un mécanisme de transcription inverse, un génotype peut être induit par un phénotype. Il est alors possible, pour un individu, d'acquérir de nouveaux caractères génétiques en fonction de son environnement. Les enfants pourront hériter, à leur tour, de ces caractères acquis durant la vie de leur parent. Dans le cadre de l'interaction ligand-recepteur, on peut ainsi effectuer une optimisation locale du ligand par rapport au récepteur et remonter les informations du phénotype optimisé vers le génotype de l'individu. Ce comportement est illustré par la partie gauche de la Figure 2. C'est ce qu'on appelle un algorithme génétique lamarckien par analogie avec la théorie, aujourd'hui discréditée, de Jean Baptiste de Lamarck selon laquelle les caractéristiques acquises par un individu durant sa vie pouvaient devenir héréditaires [30].



*Figure 2. Comportement des algorithmes génétiques, principes de Darwin (à droite) et de Lamarck (à gauche).*

AUTODOCK dispose d'un AGL dont la phase d'optimisation locale est particulière. Elle utilise une variante de la méthode de Solis et Wets [31] où l'opérateur travaille sur l'espace génotypique du ligand pour minimiser son énergie alors que la plupart des algorithmes classiques travaillent sur l'espace phénotypique. L'intégration d'une fonction de traduction inverse n'est donc pas nécessaire mais cette combinaison de recherche globale et de recherche locale reste de type lamarckien. En effet, toutes les adaptations environnementales du ligand acquises pendant l'optimisation locale seront transmises à ses enfants, s'il en a. La méthode de Solis – Wets procède avec un opérateur semblable à l'opérateur de mutation de l'AG mais le gène affecté subit une modification qui n'est pas aléatoire. Au contraire, l'opérateur modifie les gènes affectés, un par un, par pas réguliers et identiques pour chaque gène. De plus la méthode est adaptative. Elle ajuste la taille du pas en fonction de l'historique énergétique des optimisations : après un nombre déterminé de hausses consécutives de l'énergie, la taille du pas est doublée. A l'inverse, après un nombre déterminé de baisses consécutives de l'énergie, la taille du pas est divisée par deux. La méthode d'optimisation locale implémentée dans AUTODOCK est une variante de Solis – Wets dans laquelle la taille du pas est différente pour chaque type de gène : une variation de 1 Å dans un gène de position aura beaucoup plus d'impact qu'une variation de 1° dans un gène d'orientation ou de torsion. Aussi la taille du pas pour un gène de position est paramétrée par défaut à 1,0 Å alors que la taille du pas pour un gène d'orientation ou de torsion est par défaut de 50°.

A chaque génération, il est possible de faire une optimisation locale sur une fraction de la population. L'efficacité de l'amarrage est sensiblement améliorée avec une fréquence d'optimisation locale de seulement 6 % alors que le gain supplémentaire pour une fréquence de 100 % est très faible.

AUTODOCK permet d'explorer l'espace conformationnel avec, au choix, un algorithme génétique darwinien, un algorithme génétique lamarckien ou par recuit simulé par la méthode de Monté Carlo (RSMC). Des trois méthodes, RSMC est la première implémentée et était la seule disponible dans les premières versions d'AUTODOCK. Les algorithmes génétiques sont apparus dans la version 3 d'AUTODOCK et c'est l'algorithme génétique lamarckien qui donne les résultats les plus fiables et les temps de calcul les plus courts [25].

### **III - LA PROCEDURE AUTODOCK**

AUTODOCK est constitué d'un ensemble de programmes et d'utilitaires exécutés en ligne de commandes sous environnement Unix. Nous avons rédigé un document de prise en main rapide du logiciel présentant la séquence de commandes à exécuter pour effectuer un calcul d'amarrage. Ce guide à l'usage des utilisateurs débutant avec AUTODOCK, permet de lancer rapidement les premiers calculs à partir des fichiers bruts du ligand et du récepteur. Des indications sur la nature et l'influence des différents outils et paramètres jalonnent la procédure ainsi que quelques recommandations. L'objectif de ce document, joint en Annexe II p.73, est de laisser une trace du savoir-faire acquis au cours de ce travail.

La procédure AUTODOCK se décompose en cinq étapes :

- préparation des fichiers de coordonnées atomiques du ligand et de la macromolécule ;
- définition des pivots flexibles du ligand ;
- calcul des grilles de potentiel d'interaction du récepteur ;
- recherche des solutions d'amarrage ;
- analyse des résultats.

#### **1. PREPARATION DES FICHIERS.**

La macromolécule doit être représentée avec ses hydrogènes polaires ainsi que les charges partielles de tous ses atomes. Cette opération peut se faire avec le logiciel SYBYL et son module BIOPOLYMER. A une structure PDB ou au résultat d'une construction par homologie avec COMPOSER, il ne faut ajouter que les hydrogènes dits « essentiels » c'est à dire polaires et assigner les charges *Kollman United Atoms* (KOLLUA)[18]. Pour les charges KOLLUA la charge des hydrogènes manquants est reportée sur les atomes qui les portent, évitant ainsi un déficit de charges. On peut également supprimer tous les doublets libres de la macromolécule. La protéine ainsi préparée est enregistrée dans un fichier au format *mol2* (SYBYL) puis converti, avec l'utilitaire CNVMOL2TOPDBQ (fourni avec AUTODOCK), au format *pdbq* qui reprend le format *pdb* (*Protein Data Bank*) en y ajoutant une colonne pour les charges. On intègre ensuite les paramètres de solvation des atomes de la macromolécule avec l'utilitaire ADDSOL qui génère un fichier *pdbs* utilisable par AUTOGRID.

Le ligand doit comporter tous ses hydrogènes et toutes ses charges partielles. On peut calculer ces charges avec un logiciel implémentant une méthode empirique telle que celle de Gasteiger ou une méthode semi empirique utilisant l'hamiltonien AM1 par exemple. Le ligand est d'abord enregistré au format *mol2*.

## 2. DEFINITION DES PIVOTS FLEXIBLES DU LIGAND.

Les pivots qui seront en libre rotation durant la simulation doivent être désignés et les carbones aromatiques doivent être renommés dans le fichier de structure. L'utilitaire DEFTORS exécute cette dernière opération automatiquement et permet à l'utilisateur de sélectionner les pivots libres parmi la liste des pivots existant. DEFTORS prend le fichier de coordonnées atomiques du ligand au format *mol2* et enregistre le résultat de son traitement au format *pdbq*.

## 3. CALCUL DES GRILLES DE POTENTIELS.

Les grilles de potentiels sont calculées par le programme AUTOGRID selon les instructions du fichier de paramétrage qu'on lui fournit. Le fichier de paramétrage est un fichier texte dont l'extension est *gpf* (*Grid Parameters File*). Les paramètres, codés par des mots clé, indiquent :

- les noms de fichiers du ligand et de la macro, aux formats *pdbq* et *pdbqs* respectivement ;
- la position et les dimensions de la boîte ainsi que l'espacement entre les points de la grille (0,4 Å par défaut) ;
- les noms des types d'atomes présents dans le ligand et pour lesquels il faut calculer une grille ;
- la valeur des constantes des différents termes à calculer.

Chaque grille calculée est enregistrée dans un fichier dont l'extension est *map*. Le fichier de paramètres peut être généré automatiquement, par l'utilitaire MKGPF3, à partir des fichiers du ligand et du récepteur. La boîte est alors centrée sur le ligand et ses dimensions sont proportionnelles à la taille de ce ligand. Par conséquent, il est nécessaire d'éditer ce fichier généré pour modifier les paramètres de position et de dimension de la boîte afin de l'ajuster au mieux à la région du récepteur que l'on étudie. L'utilitaire MKBOX génère un fichier *pdb* des coordonnées de la boîte d'après les paramètres du fichier *gpf* ce qui permet de contrôler visuellement sa position par rapport au récepteur.

Les grilles ainsi calculées peuvent être réutilisées pour tout ligand ne comportant pas de nouveau type d'atome et devant explorer la même région de la macromolécule.

#### 4. RECHERCHE DES SOLUTIONS D'AMARRAGE.

C'est le programme AUTODOCK qui va faire cette recherche en fonction des paramètres qu'on lui transmet par l'intermédiaire d'un fichier *dpr* (*Dock Parameters File*). Les paramètres d'un calcul d'AUTODOCK sont :

- les noms des fichiers contenant le ligand et les grilles de potentiels à utiliser ;
- l'état initial du ligand (position, orientation et conformation aléatoire ou précisée) ;
- la méthode de recherche à utiliser (RSMC, AG ou AGL) avec les paramètres associés et les paramètres de l'algorithme d'optimisation locale pour la méthode AGL.

Les paramètres de l'algorithme génétique sont :

- La taille de la population. La valeur standard est de 50 mais peut être modifiée en fonction du nombre de degrés de liberté et de l'étendue de la boîte à explorer.
- Le nombre maximum d'évaluations d'énergie. C'est un paramètre d'arrêt de l'amarrage utile lorsque le processus met trop de temps, ou ne parvient pas, à converger.
- Le nombre maximum de générations qui est un autre paramètre d'arrêt. Ces deux derniers paramètres sont à adapter conjointement en fonction de la taille de la population, du nombre de degrés de liberté, de l'étendue de la boîte à explorer et du degré de convergence à atteindre. Le degré de convergence est corrélé au nombre de solutions distinctes, ou classes, que l'on veut obtenir.
- Le niveau de sélection élitiste : le nombre des meilleurs individus qui survivent automatiquement à la génération suivante. La valeur standard est 1.
- Le taux de croisements. Le croisement est l'opérateur d'exploration globale, sa probabilité doit être élevée pour que l'AG ait ce caractère global. La valeur standard est 0,8.
- Le taux de mutations. Cet opérateur joue un rôle d'optimisateur local pour les AGs purs (darwinien) en opérant des variations de faible amplitude sur les gènes, chose que le croisement fait très difficilement. Avec l'opérateur d'optimisation locale explicite des AGL, ce rôle devient inutile et la mutation sert seulement à réintroduire des allèles

éliminés par la sélection. Sa probabilité doit donc être faible dans un AGL. La distribution de Cauchy donne un compromis entre des variations radicales et une exploration trop détaillée de la topographie conformationnelle. La valeur standard est 0,02.

- Le nombre de générations précédentes  $N$  dans lesquelles on recherche l'individu le plus mal adapté pour calculer le facteur de sélection proportionnelle. Si cette valeur est trop grande, les plus anciennes générations qui sont peu optimisées, vont perturber la sélection en gommant les écarts entre les nouvelles solutions. Les meilleurs individus de la génération en cours n'auront pas beaucoup plus d'enfants que les moins bons individus et la population convergera très lentement. Si  $N$  est trop petit, la population convergera rapidement vers le minimum local le plus proche. On peut faire l'analogie entre ce paramètre et à un facteur d'échelle : si on est trop près du relief énergétique on n'en voit qu'une partie et on trouve une solution locale, si on est trop loin on ne distingue plus le relief. La valeur optimale est de 10 générations.
- Les valeurs de  $\alpha$  et  $\beta$ , assimilables à la moyenne et la variance de la distribution de Cauchy pour la mutation des gènes. Les valeurs standard sont 0 et 1 respectivement.
- Le nombre de cycles à exécuter c'est-à-dire le nombre de solutions souhaitées.

Dans le cadre d'une recherche par AGL les paramètres de l'optimisateur local (Solis Wets) sont :

- Le nombre maximal d'itérations d'optimisation, 300 pour une petite molécule organique ;
- Le nombre de succès consécutifs de l'optimisation avant de modifier la taille du pas de l'opérateur. La valeur standard est 4.
- Le nombre d'échecs consécutifs de l'optimisation avant de modifier la taille du pas de l'opérateur. La valeur standard est 4.
- La taille du pas de l'opérateur sur les gènes de position : entre 0,2 Å et 1 Å.
- La taille du pas de l'opérateur sur les gènes d'orientation : entre 5° et 50°. Les multiples de 360 sont à écarter si l'on veut éviter de re-tester périodiquement les mêmes valeurs d'angle. Il est également souhaitable que ce pas soit un réel ou un entier grand, supérieur à 36, pour assurer l'exploration fine de ces gènes. Les pas de translation et de rotation sont des pas de recherche locale.
- La taille initiale de « l'espace de recherche locale ». C'est le facteur qui est doublé ou divisé par deux après quatre optimisations consécutives de même tendance et qui pondère

les pas de recherche locale. Sa valeur initiale est de 1 et ne devrait pas être modifiée par l'utilisateur.

- La taille limite inférieure de l'espace de recherche locale. C'est un paramètre d'arrêt de l'optimisation locale. Il indique la « finesse » d'optimisation en dessous de laquelle on ne descend pas. La valeur par défaut est de 0,01. A titre d'exemple, il faut au minimum 4x7 itérations d'optimisation pour que la taille de l'espace de recherche passe de 1 à 0,01.
- Le nombre maximum de cycles d'optimisation.

Un cycle d'amarrage s'arrête soit :

- lorsque le nombre maximal de générations est atteint ;
- lorsque le nombre maximal d'évaluation d'énergie est atteint ;
- lorsque la population a convergé vers une unique solution.

A la fin de chaque cycle d'exploration, AUTODOCK enregistre la meilleure solution d'amarrage du cycle c'est à dire le phénotype de l'individu de la dernière génération ayant la meilleure interaction avec le ligand. Le fichier de résultats contiendra donc autant de solutions que de cycles exécutés. Ce fichier textuel porte l'extension *dlg* (Dock LoG).

## 5. ANALYSE DES RESULTATS.

AUTODOCK peut faire une première analyse des résultats en regroupant les solutions en classes (clusters) en fonction de leur proximité spatiale. La mesure de la proximité entre deux solutions est calculée par la racine de la moyenne des carrés des écarts (*Root Mean Square Deviation – RMSD*) de leurs coordonnées atomiques. Si le RMSD entre deux molécules est inférieure à une distance seuil, ces deux solutions sont dans la même classe. Le seuil de distance est appelé « tolérance de classe » et sa valeur par défaut, pour AUTODOCK, est de 0,5 Å. Ce paramètre est transmis à AUTODOCK par le fichier de paramétrage (*dpr*) avant le lancement de l'amarrage. Le résultat de l'analyse par classes figure sous forme d'histogramme à la fin du fichier de résultats.

La meilleure solution est celle de plus basse énergie et une convergence de toutes les solutions vers la même classe paraît souhaitable. Malheureusement, un ligand peut adopter différents modes de liaison avec le récepteur et le modèle de l'énergie n'est pas parfait. Il est donc intéressant d'obtenir plusieurs classes sur une gamme d'une dizaine de kcal.mol<sup>-1</sup>. Elles peuvent révéler ces différents modes de liaison. Le nombre de classes obtenu dépend du degré de convergence de la recherche et c'est en jouant sur le nombre d'évaluations de l'énergie et le nombre de générations qu'on module le nombre de classes obtenues.

# C - APPLICATION : ETUDE DU MECANISME DE RECONNAISSANCE D'UN POLYOSIDE BACTERIEN ET DE MIMES PEPTIDIQUES PAR L'ANTICORPS SPECIFIQUE IGA-13

## I - INTRODUCTION

### 1. LA SHIGELLOSE

La shigellose, ou dysenterie bacillaire, sévit surtout dans les régions tropicales, où elle est endémique toute l'année, avec des poussées épidémiques à certaines saisons ou lors de désastres humanitaires. La shigellose n'est pas la plus fréquente des maladies diarrhéiques mais, dans sa forme typique dysentérique, elle est sans aucun doute la plus sévère. Chaque année, elle tue en effet entre 600 000 et 1 million de personnes dans le monde, pour l'essentiel des enfants de moins de 5 ans.

Les espèces bactériennes les plus fréquentes dans les pays en voie de développement, et responsables des symptômes les plus sévères, sont *Shigella flexneri*, causant la forme endémique de la maladie et *Shigella dysenteriae* sérotype I qui cause des épidémies brutales. Une autre espèce, *Shigella sonnei*, est prévalente dans les pays industrialisés.

La shigellose est par excellence une maladie due à l'insuffisance d'hygiène. Les shigelles sont transmises par voie féco-orale. Elles sont extrêmement infectieuses puisque 10 à 100 bacilles suffisent à induire la maladie. L'homme est le seul réservoir des shigelles. Le plus souvent, la transmission est directe, du malade à son entourage. L'eau et les aliments souillés par des déjections contenant *Shigella* peuvent également transmettre la maladie. Du fait de ses conditions de survenue, la maladie touche donc essentiellement les enfants vivant dans les régions pauvres et surpeuplées de la planète où les infrastructures sanitaires et l'hygiène individuelle sont insuffisantes [32]. Elle peut toucher aussi les militaires en opération dans ces régions, les personnels humanitaires et les touristes.

Les shigelles envahissent les cellules épithéliales intestinales puis le tissu constituant la muqueuse recto-colique. Ce processus aboutit à une inflammation aiguë et une destruction massive des tissus [33].

La forme dysentérique aiguë typique chez l'adulte débute brusquement, après une incubation brève. Elle se caractérise par des douleurs abdominales, des vomissements, des épreintes, l'émission permanente de selles parfois hémorragiques et d'une fièvre élevée. Les malades guérissent souvent spontanément en quelques jours mais le risque de complications est élevé entraînant les formes graves voire mortelles de la maladie. Ces complications, généralement associées à l'hypoglycémie et à la déshydratation, touchent principalement les nourrissons et les jeunes enfants.

A la différence des autres maladies diarrhéiques, la shigellose ne peut être traitée par la seule réhydratation du fait de l'inflammation et de la destruction des tissus infectés. Les antibiotiques permettent généralement une guérison rapide. Cependant, l'émergence de souches multi résistantes compliquent le traitement et nécessitent le recours à des antibiotiques plus rares et bien plus chers. La prophylaxie, reposant sur l'amélioration des conditions de vie et surtout d'hygiène, reste illusoire dans de nombreuses régions du monde.

Le développement d'un vaccin est donc urgent et de première importance.

## **2. LES STRATEGIES DE VACCINATION**

Les bactéries gram négatives, telles que les shigelles, sont enveloppées par une sorte de manteau moléculaire constitué de longues chaînes de sucres. Ces sucres, appelés antigène O (Ag-O), sont arrimés à la membrane externe par une ancre lipophile constituée de chaînes grasses. Ces macromolécules appelées lipopolysaccharides (LPS) sont des facteurs de virulence bien connus de ce type de bactéries [34]. En revanche, le mécanisme de reconnaissance spécifique de leur sérotype n'est pas clairement établi mais la partie osidique y joue un rôle majeur. Elle est la cible de la réponse immunitaire développée par l'hôte, conduisant à une immunité protectrice contre la réinfection.

Plusieurs vaccins polysaccharidiques dirigés contre des bactéries telles que *Streptococcus pneumoniae*, *Neisseria meningitidis* ou *Salmonella typhi* ont démontré leur efficacité chez l'adulte et sont commercialisés. Malheureusement, les LPS utilisés comme vaccins, induisent chez l'adulte comme chez l'enfant une réponse médiée essentiellement par les lymphocytes B qui ont une faible mémoire. De plus, le système immunitaire immature des nouveau-nés répond mal aux antigènes non présentés par les lymphocytes T. Par conséquent les vaccins polysaccharidiques ne sont pas ou peu immunogènes chez les nourrissons. Or, l'incidence de nombreuses maladies bactériennes comme les infections par les pneumocoques ou les shigelles, est très élevée durant les deux premières années de la vie. Cette difficulté a pu être contournée

par la combinaison des polysaccharides immunogènes avec des peptides qui jouent un rôle de vecteur et améliorent la présentation de l'anticorps au système immunitaire. Ainsi une protection efficace y compris chez l'enfant, a pu être induite contre des infections du groupe C, dont celles provoquées par *Haemophilus influenzae* b, *S. pneumoniae* et *N. meningitidis* [35].

Une alternative possible consiste à ne plus utiliser les polysides bactériens comme inducteur de la réponse immunitaire mais des mimes de l'antigène, capables d'induire une réponse identique. La recherche de ces mimes s'est orientée vers deux types de structures : les premiers sont des oligosides de synthèse reprenant la structure de la partie osidique reconnue par les anticorps protecteurs (l'épitope) ; les seconds sont des molécules non osidiques, en l'occurrence des peptides, mimant cet épitope reconnu par les anticorps protecteurs, anticorps monoclonaux anti-sucres (mAb).

Les glycoconjugués comportant des mimes oligosidiques d'antigènes polysaccharidiques bactériens se sont révélés très immunogènes chez la souris [36-38]. La validité du concept a ensuite été démontrée chez l'homme par l'efficacité de ces glycoconjugués semi-synthétiques à conférer une protection contre l'infection par *H. influenzae* b [39]. Parallèlement, les recherches se sont orientées vers l'exploration du potentiel de mimétisme des peptides [40-42]. Les mimotopes peptidiques sont les mimes peptidiques capables d'induire une réponse immunitaire par des anticorps anti-sucres. Ils ont été proposés comme antigènes de substitution potentiels dans le développement de nouveaux vaccins. Leurs avantages par rapport aux oligosides obtenus par cultures bactériennes ou par synthèses multi-étapes, sont la facilité de production et la présence de propriétés immunogènes intrinsèques. Cependant, parmi tous les mimes peptidiques d'antigènes polysaccharidiques, très peu ont des propriétés mimotopiques et les vaccins basés sur ces mimotopes peptidiques sont encore peu nombreux [43-48].

Il semble donc qu'une meilleure compréhension des bases moléculaires du phénomène de mimétisme peptide-sucres soit nécessaire pour le développement rationnel de vaccins basés sur ces mimotopes. Il s'agit en particulier de distinguer les composantes structurales et fonctionnelles impliquées dans ce mécanisme. Les données cristallographiques sur les complexes oligosaccharide-protéine et mime peptidique-protéine correspondant, tout comme les données thermodynamiques de l'interaction peptide-protéine, sont très rares [49-51]. A ce jour, les données disponibles proviennent principalement d'études de RMN et de modélisations moléculaires [52].

### 3. LES MIMES PEPTIDIQUES

Dans le but d'obtenir une protection contre les infections par *Shigella flexneri*, des laboratoires de l'Institut Pasteur de Paris s'intéressent au développement de mimes synthétiques et protecteurs des antigènes majeurs de ce pathogène.

La réponse immunitaire induite par l'infection est spécifique du sérotype bactérien [33] et est principalement dirigée contre la partie polysaccharidique O-spécifique (O-SP) du LPS bactérien. L'Institut a développé des glycoconjugués du LPS détoxifié de *S. flexneri* 2a, sérotype majoritaire chez l'homme, qui ont montré leur non toxicité et leur pouvoir immunogène à la fois chez l'adulte et chez l'enfant [53]. Plus récemment, ces équipes ont développé des glycoconjugués entièrement synthétiques ainsi que des néoglycoprotéines présentant des fragments saccharidiques de synthèse. Ces composés sont capables de mimer le O-SP de *S. flexneri* 2a ce qui en fait des vaccins potentiels prometteurs contre les infections homologues [54-56]. Parallèlement, les recherches de mimotopes peptidiques ont apporté le premier exemple de peptides mimétiques de polysaccharides ayant des propriétés immunogènes. Ces peptides ont été obtenus par criblage d'une banque de nonapeptides présentés à deux anticorps monoclonaux protecteurs, d'isotype A (mIgA) et spécifiques de *S. flexneri* sérotype 5a : mIgA C5 et mIgA I3 [57]. Parmi les 19 séquences peptidiques sélectionnées par criblage, p100c (-CYKPLGALTHC-) sélectionné par mIgA I3 et p115 (KVPPWARTA) sélectionné par mIgA C5, sont les seuls à induire, chez la souris, la production d'anticorps anti-O-SP. Plus intéressant encore, ces deux mimotopes ne partagent aucune séquence commune et n'interfèrent pas ensemble. Leur seule similitude est qu'ils comportent, comme on l'observe souvent [58, 59], des acides aminés aromatiques et hydrophobes ainsi que des acides aminés à chaîne latérale cyclique dont au moins une proline.

Le O-SP 5a de *S. flexneri* est constitué par la répétition d'un pentasaccharide ramifié de motif A(E)BCD (Figure 3)[60]. Grâce à des données de RMN, un modèle conformationnel de cet O-SP a pu être construit par modélisation moléculaire. C'est lui qui est reconnu par les immunoglobulines dont IgA I3 et IgA C5, au niveau du tractus digestif.

Une étude systématique de conformation et d'antigénicité menée sur les 4 pentasaccharides de synthèse correspondant aux 4 « découpages » possibles le long de la chaîne, indique que c'est la séquence DA(E)BC qui mime le mieux l'antigène O-SP bactérien lors de la reconnaissance spécifique par les anticorps. Plus récemment, les propriétés mimotopiques de ce pentasaccharide ont été constatées : il est capable d'induire une réponse immunitaire et confère une protection contre l'infection (L. Mulard et A. Phalipon - résultats non publiés).

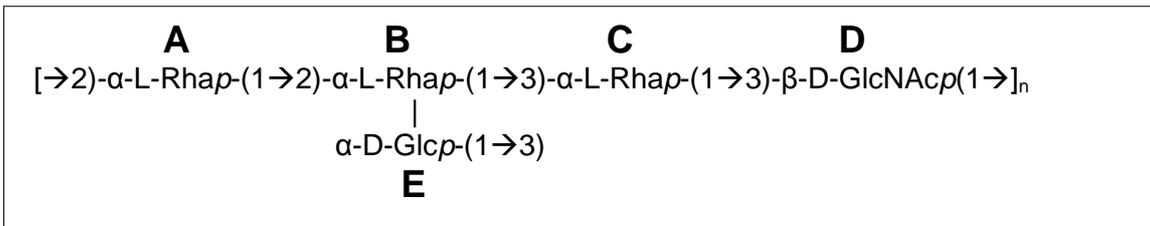


Figure 3. Formule de l'Ag-O de *S. flexneri*

Des analyses RMN de transfert de NOE (trNOE) et de différence de transfert de saturation (STD) ont permis d'établir les conformations de différents antigènes (O-SP 5a, p100c, p115 et p22) à la fois dans leurs formes libres et dans leurs formes liées à un anticorps. Une cartographie de l'épitope c'est à dire la localisation des interactions de liaison anticorps-antigène, est également issue de cette étude [61, 62].

L'objet de cette nouvelle étude est de construire, à partir des données précédentes, un modèle théorique de la reconnaissance du O-SP 5a de *S. flexneri* par mIgA I3. Ce modèle contribuera à l'analyse des éléments de reconnaissance encore mal connus, d'un polysaccharide et d'un mime peptidique de ce polysaccharide par un anticorps monoclonal.

## **II - MODELE DE L'ANTICORPS**

Les immunoglobulines A (mIgA) sont les anticorps excrétés dans les liquides biologiques : salive, sucs digestifs, sécrétions bronchiques, larmes, colostrum et lait. Elles comportent cinq types de sous-unités : des chaînes légères [κ] ou [λ], des chaînes lourdes de type [α] et deux petites sous-unités, les pièces J et S. Les chaînes lourdes sont groupées deux par deux, associées de façon parallèle sur une partie de leur longueur, donnant une forme en 'Y' à l'ensemble. Sur la partie non liée de chaque chaîne lourde est associée une chaîne légère. Cet ensemble, composé de 2 chaînes légères et de 2 chaînes lourdes, constitue un monomère d'immunoglobuline (Figure 4). Les IgA sécrétoires sont constituées d'au moins 2 unités monomériques liées entre elles par leur base grâce à une sous unité de jonction (pièce J). Une dernière sous unité, nécessaire au mécanisme de sécrétion (pièce S), complète l'ensemble. Cette dernière sous unité assure la bonne localisation tissulaire grâce à des résidus osidiques permettant l'ancrage de l'anticorps au mucus recouvrant l'épithélium [63]. Toutes les associations de chaînes sont réalisées au moyen de ponts disulfures.

Alors que la séquence de ces chaînes est fortement conservée dans chaque classe d'anticorps, on observe à chaque extrémité apicale un domaine dit hypervariable (Fab)[64]. Il est le siège de nombreuses mutations, insertions et délétions. C'est au niveau de ce domaine qu'a lieu le

mécanisme de reconnaissance moléculaire de l'antigène par l'anticorps. Ainsi chaque IgA porte quatre sites de reconnaissance de l'antigène.

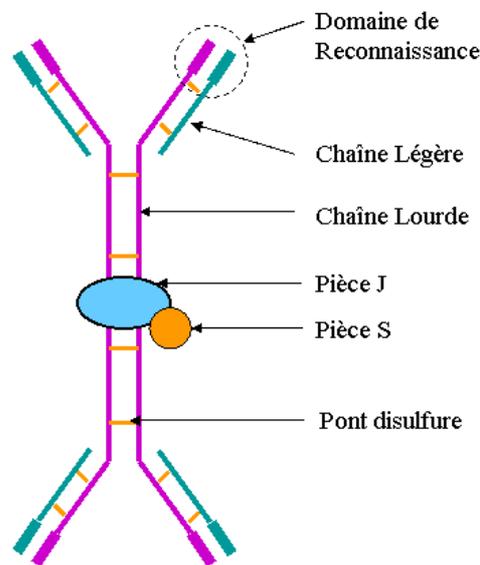


Figure 4 : Schéma de la structure d'une immunoglobuline de type IgA. Les domaines de reconnaissance spécifiques sont situés à l'extrémité des chaînes.

La variabilité du domaine est répartie sur chacune des deux chaînes constituant le Fab (Vl et Vh). On dénombre 3 zones hypervariables dénommées L1, L2 et L3 sur la chaîne légère et 3 zones hypervariables H1, H2 et H3 sur la chaîne lourde. Ces 6 zones hypervariables ont toutes une structure secondaire en boucle et ces boucles sont spatialement rapprochées. Elles se trouvent au coeur du site de reconnaissance de l'anticorps et constituent le domaine de reconnaissance complexe (CDR) de l'antigène. La boucle H3 est la plus variable de toutes, celle qui présente le plus de singularités.

Ne disposant pas de formes cristallines d'IgA I3, nous ne connaissons pas la structure tridimensionnelle du Fab. Cependant, grâce à la forte conservation structurale de chaque classe d'immunoglobuline et grâce à l'abondance de structures cristallographiques d'homologues, il est possible de construire un modèle théorique d'IgA I3 par homologie. Nous allons construire un modèle du Fab d'IgA I3 selon une procédure établie, basée sur ces homologues de structures [3-5]. Chaque chaîne est construite indépendamment. Après avoir sélectionné les structures homologues, nous identifions les régions structurellement conservées qui serviront d'armature au modèle puis nous construirons les boucles variables à partir de fragments homologues. Les modèles de chaque chaîne seront ensuite assemblés par superposition aux chaînes de la structure de référence présentant la plus forte homologie et la meilleure résolution.

## **1. RECHERCHE DE STRUCTURES TRIDIMENSIONNELLES DE SEQUENCE HOMOLOGUE.**

La structure tridimensionnelle d'un grand nombre d'anticorps a pu être déterminée. Leurs coordonnées atomiques sont conservées et diffusées par la PDB. Ces structures sont déterminées expérimentalement par des techniques de diffraction des rayons X ou de RMN. Pour chaque structure, les coordonnées atomiques et un certain nombre d'informations connexes (auteur, méthode d'acquisition, résolution, connectivité...) sont rassemblées dans un fichier portant un nom formaté composé de 4 caractères et d'une extension « pdb ». Pour des raisons pratiques, on dénomme très souvent une structure issue de la PDB par son code d'enregistrement dans cette base.

Il est possible d'effectuer des comparaisons entre la séquence d'acides aminés de ces structures connues et une ou plusieurs séquences cibles grâce à des algorithmes spécialisés tels que BLAST [10].

Pour trouver des structures homologues d'IgA13, nous avons utilisé l'algorithme PSI-BLAST avec la matrice de substitution BLOSUM62. Nous avons donné au programme les séquences de la chaîne légère puis de la chaîne lourde et nous avons effectué une recherche sur la librairie des structures de la PDB. Le seuil maximal de « E value » (valeur attendue) est de 0,01 pour les résultats retenus.

L'extrémité C terminale des structures d'immunoglobulines sélectionnées est tronquée et seul le domaine aligné à la séquence du Fab d'IgA I3 est représenté. La partie tronquée correspond à un domaine riche en feuillet beta, située en arrière du Fab et dont le rôle est essentiellement structural. Ce domaine n'est pas impliqué dans le mécanisme de reconnaissance moléculaire et n'est pas modélisé pour cette étude.

Sur les figures d'alignement, les acides aminés sont représentés en rouge lorsqu'ils diffèrent de la séquence recherchée. La table de codage des acides aminés est indiquée en Annexe I p.72. Les résultats de PSI-BLAST sont présentés sur la Figure 5 et la Figure 6 pour la chaîne légère et pour la chaîne lourde respectivement. La séquence cible est notée « Query », les séquences trouvées sont annotées par leur code pdb.

### ***a. Recherche d'homologues de la chaîne légère.***

La séquence de la chaîne légère d'IgA I3 compte 115 acides aminés. Les différences de séquence entre les homologues trouvés et la cible sont peu nombreuses : 91% d'identité minimum. Ces différences surviennent presque toujours sur les mêmes positions et les acides aminés variants

sont des homologues positifs c'est à dire dont la mutation est probable d'après la matrice de substitution. Si on les prend en compte, on atteint 95 % de similarité. Les structures sélectionnées pour construire la chaîne légère sont les chaînes légères de 1MRC, 1MRD, 1A4J, 1A4K et 1CBV. Toutes ces structures ont été déterminées expérimentalement par diffraction des rayons X.

1MRC et 1MRD contiennent la structure du Fab d'une immunoglobuline anti ARN de souris, en complexe avec l'inosine-5'-diphosphate pour 1MRD et libre pour 1MRC [65]. La séquence de cet anticorps présente 93 % d'identité avec la séquence d'IgA I3. 1MRD a une résolution de 2,3 Å et celle de 1MRC a une résolution de 2,4 Å.

1A4J et 1A4K contiennent la structure du Fab d'anticorps de souris modifié pour la catalyse de réactions de Diels-Alders libre [66]. 1A4J a une résolution 2,1 Å et sa séquence présente 92 % d'identité avec la séquence d'IgA I3. 1A4K a une résolution de 2,4 Å.

1CBV est un anticorps spécifique d'un ADN simple brin complexé avec un trinuéclotide TTT et sa résolution est de 2,66 Å [67].

Les séquences homologues divergent de la cible au niveau des positions 26 à 34, 94, 102 et 107. 1MRD présente une mutation en position 98 (A<sup>98</sup>-T<sup>98</sup>) qui se trouve dans une boucle hypervariable (CDR-L3) alors que 1A4J est identique à la cible sur cette position. Les structures 1A4J et 1A4K présentent des mutations aux positions 110 et 111 (AD-TV) mais ces acides aminés ne sont pas localisés dans une boucle hypervariable.

Ces cinq structures sont ajoutées à la base locale PRODAT de SYBYL.

| IgA I3 – Chaîne légère                                       |  |
|--|--|
| Query  | : 1 MTQTPLSLPVS LGDQASISCRSSQSLIDSKGNIYLHWYLQKPGQSPKLLIYKVS NRF 57     |
| 1MRD   | : 4 MTQTPLSLPVS LGDQASISCRSSQSLVHSNGNTYLHWYLQKPGQSPKLLIYKVS NRF 60     |
| 1A4J   | : 4 MTQTPLSLPVS LGDQASISCRSSQSLVHSNGNTYLHWYLQKPGQSPKLLIYKVS NRF 60     |
| 1CBV   | : 4 MTQTPLSLPVS LGDQASISCRSSQSLVHSNGNTYLHWYLQKPGQSPKLLIYKVS NRF 60     |
| 1A4K   | : 4 MTQTPLSLPVS LGDQASISCRSSQSLVHSNGNTYLHWYLQKPGQSPKLLIYKVS NRF 60     |
| Query  | : 58 SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSAHVPPTFGSGTKLKIKRADAAP 115    |
| 1MRD   | : 61 SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHVPRTFGGGTKLEIKRADAAP 118    |
| 1A4J   | : 61 SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHVPPTFGGGTKLEIKRTVAAP 118    |
| 1CBV   | : 61 SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHVP LTFGAGTKLELEIKRADAAP 118 |
| 1A4K   | : 61 SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQVTHVPPTFGGGTKLEIKRTVAAP 118    |
| 1MDR_L Identities = 107/115 (93%), Positives = 109/115 (94%) |  |
| 1A4J_L Identities = 106/115 (92%), Positives = 108/115 (94%) |  |
| 1CBV_L Identities = 106/135 (92%), Positives = 110/115 (95%) |  |
| 1A4K_L Identities = 105/135 (91%), Positives = 108/115 (94%) |  |

Figure 5 :Résultat de la procédure BLAST cherchant les séquences homologues à la chaîne légère d'IgAI3 dans la Protein Data Bank.

### ***b. Recherche d'homologues de la chaîne lourde.***

L'identité de séquence est légèrement inférieure avec 82 % à 85 % d'identité, mais reste néanmoins très élevée. Si l'on tient compte des homologies positives, on atteint 88 % à 91 % de similarité. Les principales divergences se trouvent dans la zone comprise entre les acides aminés 89 et 96 de notre séquence cible. D'un homologue à l'autre, on constate que toutes les séquences n'ont pas la même longueur : certaines comportent des *gaps*. Cela se traduit par des boucles de plus ou moins grande taille. Ces différences apparaissent dans l'une des zones hypervariables de la chaîne lourde, la plus variable de toutes, notée CDR-H3. Pour construire cette chaîne de l'anticorps, nous devons combiner les structures de plusieurs protéines homologues.

Les structures homologues retenues sont les chaînes lourdes de trois immunoglobulines : 1A4J, 1NCA et 1TET.

A l'exception de la zone H3, la séquence qui présente le plus d'homologie avec la chaîne lourde d'IgA I3 est la chaîne lourde de 1A4J, avec 85 % d'identité. La structure 1NCA est un complexe de Fab d'une immunoglobuline de souris spécifique de la neuraminidase NC41 du virus Influenza N9 [68]. Sa résolution est de 2,5 Å.

| Iga I3 - Chaîne lourde |   |
|------------------------|---|
| Query: 14              | KLLGFTFTIYGMNWVKQAPGKGLKWMGWINTYTGEPTYADDFKGRFAFSLETSASTAFLQ 73     |
| 1A4J : 23              | KASGYTFTNYGMNWVKQAPGKGLKWMGWINTYTGEPTYADDFKGRFAFSLETSASTAYLQ 82     |
| 1TET : 23              | KASGYTFTTYGMSWVKQTPGKGFKWMGWINTYSGVPTYADDFKGRFAFSLETSASTAYLQ 82     |
| 1IAI : 23              | KASGYTFTNYGMNWVKQAPGKGLKWMWAWINTYTGEPTYADDFKGRFAFSLETSASTAYLQ 82    |
| 1NCA : 23              | KASGYTFTNYGMNWVKQAPGKGLKWMGWINTNTGEPTYGEEFKGRFAFSLETSASTANLQ 82     |
|                        |   |
| Query: 74              | INNLKNEEDTATYFCARADDY-----FDYWGQGTTLTVSS 107                        |
| 1A4J : 83              | INNLKNEEDTATYFCVQAERLRRT--FDYWGAGTTVTVSS 119                        |
| 1TET : 83              | INNLKNEEDTATYFCARRSWY-----FDVWGTGTTVTVSS 116                        |
| 1IAI : 83              | INNLKNEEDTATYFCARDGYENYAMDYWGQGTSTVTVSS 121                         |
| 1NCA : 83              | INNLKNEEDTATFFCARGEDNFGSL--SDYWGQGTTVTVSS 120                       |
|                        |   |
| 1a4j_H                 | Identities = 83/97 (85%), Positives = 89/97 (91%), Gaps = 3/97 (3%) |
| 1tet_H                 | Identities = 79/94 (84%), Positives = 85/94 (90%)                   |
| 1iai_H                 | Identities = 83/99 (83%), Positives = 88/99 (88%), Gaps = 5/99 (5%) |
| 1nca_H                 | Identities = 81/98 (82%), Positives = 88/98 (89%), Gaps = 4/98 (4%) |

*Figure 6 : Résultat de la procédure BLAST cherchant les séquences homologues à la chaîne lourde d'IgA I3 dans la Protein Data Bank.*

Pour la boucle du CDR-H3 par contre nous utiliserons la structure du CDR-H3 de la protéine 1TET [69] dont le principal avantage est de comporter le même nombre d'acides aminés et une résolution de 2,3Å comparable à celle de 1A4J. Si les séquences pour cette boucle divergent un peu avec « seulement » 84 % d'identité et 90 % de similarité, les chaînes principales conservent vraisemblablement la même géométrie dans l'espace. En effet elles sont soumises à un

environnement de contraintes similaire dans les deux structures. De plus, la petite taille de cette boucle lui confère peu de mobilité.

La structure 1IAI [70] présente comme séquence homologue n'a pas été sélectionnée bien qu'elle possède une forte homologie avec la cible. Sa résolution de 2,9 Å est jugée insuffisante et sa classe d'anticorps, idiotype anti-idiotype, semble trop éloignée de celle d'IgA I3.

L'alignement des chaînes d'IgA I3 avec 1A4J est présenté dans la Figure 7 où sont également délimitées les zones hypervariables correspondant aux CDRs selon la description faite par Chothia et al [64].

|   |            |            |                 |             |               |            |   |  |   |
|---|------------|------------|-----------------|-------------|---------------|------------|---|--|---|
| <b>Chaîne légère</b> : 106 résidus identiques sur 115                             |            |            |                 |             |               |            |   |  |   |
|   |            |            | *---*--L1       | -----*-     |               | *-L2--     |   |  | * |
| 10  | 20         | 30         | 40              | 50          | 60            | 70         |   |  |   |
| MTQTPLSLPV  | SLGDQASISC | RSSQSLIDSK | GNIYLHWYLQ      | KPGQSPKLLI  | YKVSNRFSGV    | PDRFSGSGSG |   |  |   |
| MTQTPLSLPV  | SLGDQASISC | RSSQSLVHSN | GNTYLHWYLQ      | KPGQSPKLLI  | YKVSNRFSGV    | PDRFSGSGSG |   |  |   |
| SCR1  |            |            |                 |             | SCR2          |            |   |  |   |
|   |            |            | *-L3-*          | --          |               |            |   |  |   |
| 80  | 90         | 100        | 110             | 120         |               |            |   |  |   |
| TDFTLKISRV  | EAEDLGVYFC | SQSAHVPPTF | GSGTKLKIKR      | ADAAP       |               |            |   |  |   |
| TDFTLKISRV  | EAEDLGVYFC | SQSTHVPPTF | GGTKLEIKR       | TVAAP       |               |            |   |  |   |
| SCR2  |            |            |                 |             |               |            |   |  |   |
| <b>Chaîne lourde</b> : 96 résidus identiques sur 110 et une délétion de 3 résidus |            |            |                 |             |               |            |   |  |   |
|   |            |            | *-----H1-----** |             | *-----H2----- | *          | * |  |   |
| 10  | 20         | 30         | 40              | 50          | 60            | 70         |   |  |   |
| QLKRPGETVR  | ISCKLLGFTF | TIYGMNWVKQ | APGKGLKWMG      | WINTYTGTEPT | YADDFKGRFA    | FSLETSASTA |   |  |   |
| ELKKPGETVK  | ISCKASGYTF | TNYGMNWVKQ | APGKGLKWMG      | WINTYTGTEPT | YADDFKGRFA    | FSLETSASTA |   |  |   |
| SCR1  |            |            |                 |             | SCR2          |            |   |  |   |
|   |            |            | *-H3-----*      | *           | ***           |            |   |  |   |
| 80  | 90         | 100        | 110             |             |               |            |   |  |   |
| FLQINNLKNE  | DTATYFCARA | DDY---FDYW | GQGTTLTVSS      |             |               |            |   |  |   |
| YLQINNLKNE  | DTATYFCVQA | ERLRRTFDYW | GAGTTVTVSS      |             |               |            |   |  |   |
| SCR3  |            |            |                 |             |               |            |   |  |   |

Figure 7 : Alignement des séquences de IgA I3 (haut) et 1A4J (bas) (numérotation PDB). Les résidus surlignés d'une étoile sont invariants, ceux surlignés par un tiret sont dans une boucle hypervariable [64].

Les fichiers de coordonnées atomiques 1NCA et 1TET ont été ajoutés à la base PRODAT de SYBYL, pour compléter le jeu de structures de référence du module COMPOSER que nous allons utiliser pour la construction des modèles.

## 2. CONSTRUCTION DE LA CHAÎNE LÉGÈRE.

Les structures sélectionnées sont 1MRC, 1MRD, 1A4J, 1A4K et 1CBV.

Les régions structurellement conservées (SCR) sont :

- SCR1 : Met<sup>1</sup> à Leu<sup>26</sup> ;
- SCR2 : Tyr<sup>34</sup> à Pro<sup>115</sup>.

COMPOSER aligne les SCRs des structures tridimensionnelles sélectionnées, par leur chaîne principale (*backbone*). Il construit un squelette composite à partir des positions moyennes des carbones  $\alpha$  pour ces segments conservés. Il utilise ensuite les coordonnées de la *chaîne principale* de la chaîne légère de 1MRD (1MRD-L) pour construire les chaînes principales des SCRs du modèle puis intègre les chaînes latérales. Seules quatre positions portent des acides aminés différents sur 1MRD-L et sur la séquence cible.

Pour la boucle entre SCR1 et SCR2, des fragments sont sélectionnés dans la base de structures pour le bon ajustement de leurs résidus d'ancrage aux résidus d'ancrage du squelette composite.

La boucle Ile<sup>27</sup> – Ile<sup>33</sup> est ajoutée à partir du fragment correspondant de 1MRD-L qui est de même taille. COMPOSER reprend la chaîne principale du fragment et l'ajuste sur le squelette composite. Les trois résidus d'ancrage, de part et d'autre de la boucle du fragment, sont superposés aux résidus correspondants des modèles des SCRs précédemment construits. L'ajustement des coordonnées est fait par modification des angles de torsion de la chaîne principale. Les chaînes latérales sont ensuite construites soit par reproduction, lorsque les résidus sont identiques, soit par la méthode de reconstruction décrite dans la partie méthodologique.

Un pont disulfure est établi entre les cystéines 20 et 90.

Une première série d'optimisations locales du modèle brut permet de minimiser les interactions de Vander Waals entre chaînes latérales.

Les hydrogènes sont ajoutés avec le module BIOPOLYMER. Une nouvelle optimisation est appliquée mais cette fois aux hydrogènes seulement, les atomes lourds étant figés.

Les charges partielles sont alors calculées par la méthode Kollman-ALL.

## 3. VALIDATION DE LA CHAÎNE LÉGÈRE.

Une première validation du modèle est effectuée avec le logiciel PROCHECK.

Celui-ci nous indique que seule la valine 53 est dans une conformation défavorable. 88,5 % des résidus, autres que glycine et proline, se trouvent dans les régions les plus favorables et 10,4 %

se trouvent dans les régions autorisées supplémentaires. Les concepteurs de PROCHECK indiquent qu'un modèle de bonne qualité doit comporter 90 % de ses résidus dans les régions favorables. La définition de ces zones et du taux minimum de résidus corrects ont été établis sur la base d'une analyse de 118 protéines de résolution inférieure à 2,0 Å et de *R-Factor* inférieur à 20 %. Or, la structure la plus précise dont nous disposons, 1A4J, est moins précise que les structures de référence avec une résolution de 2.1Å et un *R-factor* de 22,9 %. De plus notre séquence ne comporte que 115 résidus correspondant à la moitié seulement de la chaîne. Par conséquent, le poids relatif de chaque résidu "incorrect" sur le pourcentage final est plus important que dans la base de référence utilisée pour établir ces seuils statistiques.

Nous considérons donc qu'avec 88,5 % des résidus « corrects », ce modèle a une chaîne principale de conformation acceptable.

#### 4. CONSTRUCTION DE LA CHAÎNE LOURDE.

Les structures sélectionnées sont : 1A4J, 1A4K, 1NCA et 1TET.

Les régions structurellement conservées sont :

- SCR1 : Gln<sup>1</sup> à Leu<sup>15</sup>
- SCR2 : Asn<sup>26</sup> à Arg<sup>89</sup>
- SCR3 : Trp<sup>100</sup> à Ser<sup>110</sup>

Après dérivation du squelette composite des SCRs à partir des structures sélectionnées, la structure de 1A4J est utilisée pour construire la chaîne principale et les chaînes latérales des SCRs du modèle.

Les fragments de boucles sont sélectionnés dans la base de structures par la méthode d'ajustement des résidus d'ancrage sur leurs coordonnées dans le squelette composite. La boucle Leu<sup>16</sup> – Met<sup>25</sup> est construite à partir du fragment correspondants provenant de 1A4J et la boucle Ala<sup>90</sup> – Tyr<sup>99</sup> est construite à partir du fragment correspondants provenant de 1TET. L'ajustement des résidus d'ancrage est réalisé par la méthode du gradient des coordonnées c'est à dire par transition des coordonnées du modèle vers celles du fragment. La méthode par ajustement des angles de torsion ne convergeant pas est abandonnée.

Les cystéines 13 et 87 sont reliées par un pont disulfure.

Les interactions de van der Waals sont réduites par minimisation locale des chaînes latérales en conflit. Les hydrogènes sont ajoutés puis optimisés comme précédemment. Les charges atomiques partielles sont calculées par la méthode de Kollman- ALL.

## 5. VALIDATION DE LA CHAÎNE LOURDE.

Le contrôle du modèle par PROCHECK indique qu'aucun résidu n'est dans une conformation interdite et que plus de 90 % d'entre eux sont dans des conformations d'énergie optimale. Le modèle de la chaîne lourde a donc une géométrie acceptable.

## 6. ASSEMBLAGE DES CHAINES LOURDES ET LEGERES.

Les modèles de la chaîne lourde et la chaîne légère sont assemblés par ajustement (« RMS fit ») de leurs carbones  $\alpha$  des régions SCRs aux carbones correspondants de la structure 1A4J. La valeur des RMS résultants est inférieure à 0,6 Å ce qui indique un excellent ajustement.

Le modèle définitif du Fab d'IgA I3 est obtenu après une dernière série de minimisations locales des résidus en conflit stérique. Ce modèle est présenté sur la Figure 8.

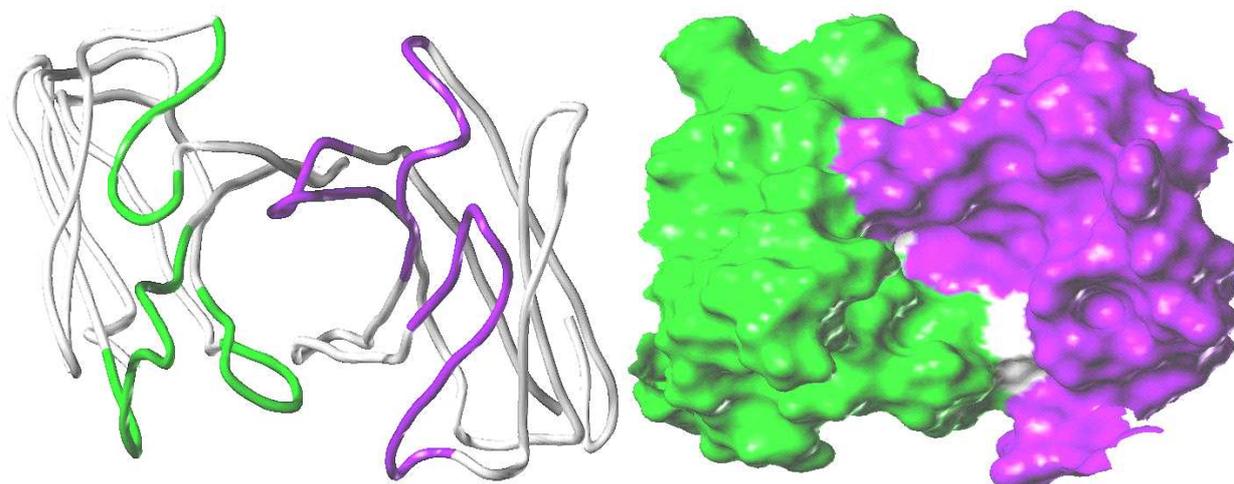


Figure 8 : Modèle de l'anticorps IgA I3. Représentation de la chaîne principale en tube (à gauche) et représentation de la surface de Conolly, surface accessible au solvant, du site de reconnaissance (à droite). Les boucles hypervariables sont en vert pour la chaîne légère, en violet pour la chaîne lourde.

## III - MODELES DU POLYSACCHARIDE.

Une précédente étude a montré que le O-SP de *S. flexneri* 5a en solution adoptait une conformation hélicoïdale [71]. Cette structure en hélice droite, oscille entre deux positions extrêmes par élongation/contraction le long de l'axe longitudinal de l'hélice. Les deux conformations extrêmes ont un pas d'hélice de 19,4 Å et de 23,2 Å respectivement. Au cours de cette même étude, un modèle moléculaire de chaque conformère a été construit d'après les contraintes géométriques fournies par les analyses RMN (Figure 9). Les analyses de transfert

d'aimantation nous apprennent par ailleurs que la zone de contact du polysaccharide avec l'anticorps implique particulièrement les résidus A, E et B. Cette information nous permettra de valider les solutions d'amarrage. Celles ne présentant pas d'interaction par ces résidus seront écartées.

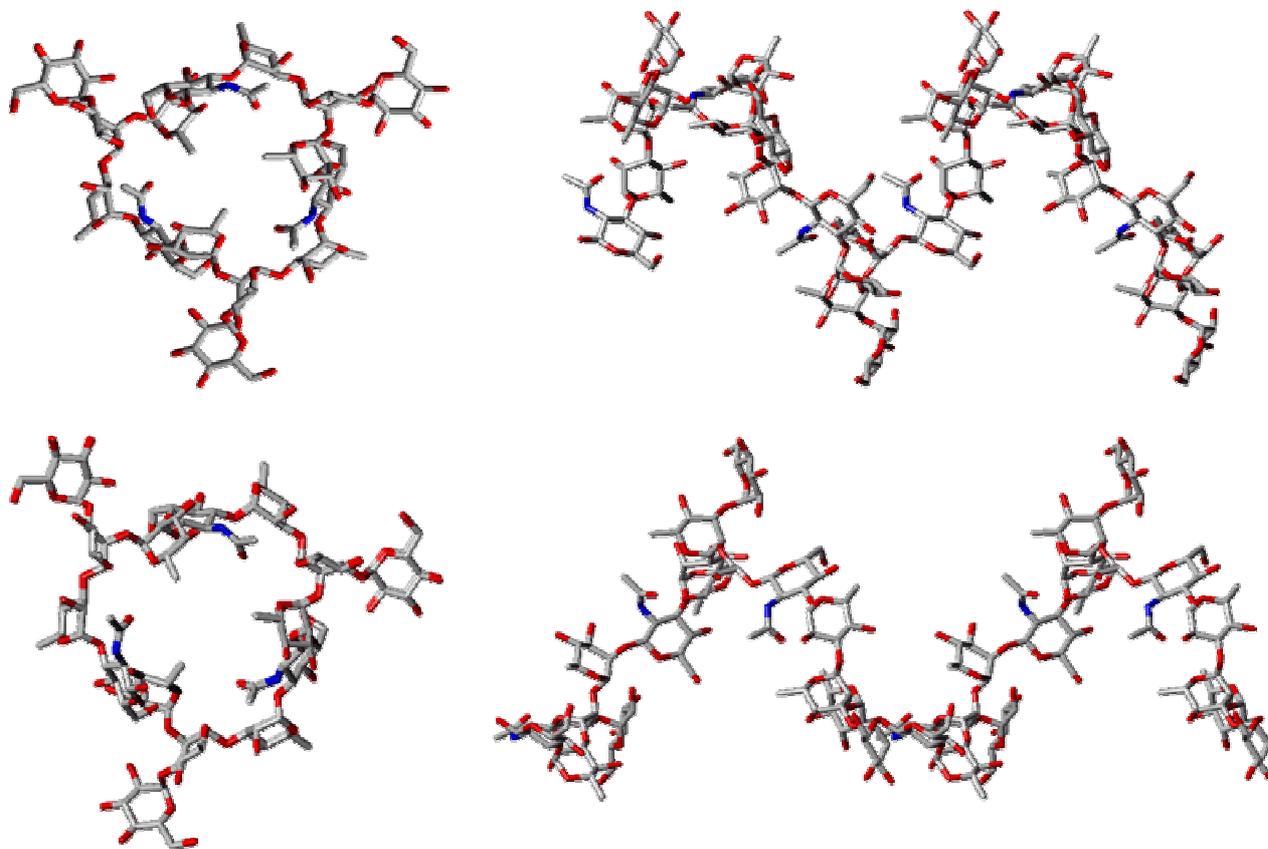


Figure 9 : Fragment du O-SP. Conformations condensée (en haut) et étendue (en bas).

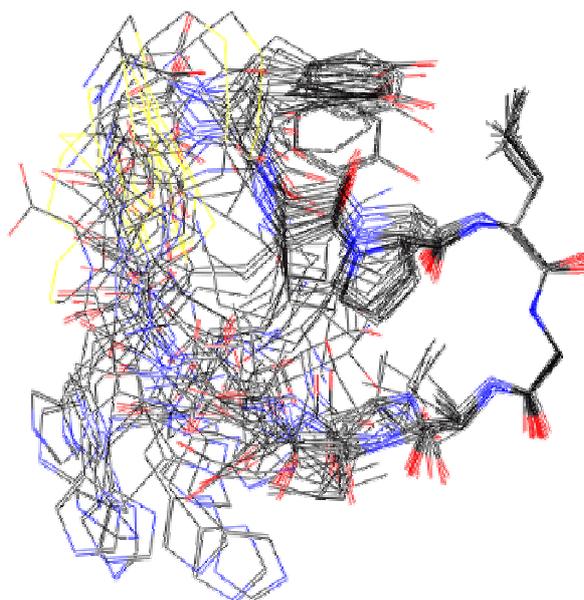
Pour réaliser l'amarrage du polysaccharide à l'anticorps avec un logiciel d'amarrage, nous devons travailler sur un fragment de O-SP. La conformation générale du polysaccharide et celle des liaisons glycosidiques en particulier sont connues grâce à la RMN. Le squelette du fragment glycosidique sera arrimé de façon rigide. Par contre, les substituants présents sur ce squelette, les groupements hydroxyles et N-acétyles, seront flexibles. AUTODOCK ne prend en compte que 32 pivots au maximum. Du fait de cette limitation, le plus gros fragment partiellement flexible que nous pouvons traiter est un nonasaccharide. A partir du fragment DA(E)BC qui a le meilleur pouvoir antigénique vis-à-vis d'IgA13 et par extension de deux résidus à chaque extrémité, nous obtenons le fragment BCDA(E)BCDA (1/2 pas d'hélice). Deux nonasaccharides, dénommés T1 et T4, sont respectivement extraits des conformations condensée et étendue du polysaccharide O-SP.

#### ***IV - MODELES DU MIME PEPTIDIQUE.***

Nous avons recherché les modes de liaison possible du peptide P100C. Ce peptide de séquence CYKPLGALTHC, est cyclisé par un pont disulfure entre les cystéines 1 et 11. L'analyse par RMN en solution du complexe IgA I3 avec ce peptide nous donne des conformations possibles du peptide complexé. Ne disposant pas d'informations équivalentes concernant P115 (KVPPWARTA) en interaction avec IgA I3, aucun calcul d'amarrage n'a été effectué pour ce dernier. En effet, ce nonapeptide linéaire comporte 28 pivots ; il est donc extrêmement flexible et nous ne disposons d'aucun élément permettant de valider les solutions proposées.

AUTODOCK ne peut pas traiter les liaisons simples intracycliques comme des pivots bien que celles-ci disposent d'une liberté partielle de rotation. La flexibilité du cycle ne peut donc pas être directement prise en compte par le logiciel et pour contourner ce problème, nous devons arrimer différents conformères rigides du peptide.

Les distances et constantes de couplage dérivées des analyses RMN du peptide P100C en interaction avec IgA I3, ont servi de contraintes au logiciel DYANA pour générer une famille de conformères de ce peptide. Les vingt meilleures structures, en terme de respect des contraintes et d'énergie, ont été sélectionnées pour être arrimées à IgA I3. Ces conformères, superposés par la chaîne principale du segment Pro<sup>4</sup> – Ala<sup>7</sup>, sont présentés sur la Figure 10. Les contraintes internes issues de la RMN conditionnent aussi bien la conformation du cycle que celle des chaînes latérales. Les conformères pourront donc être arrimés de façon entièrement rigide puisqu'ils ont géométries réalistes.



*Figure 10 : Structures superposées de 20 conformères du peptide P100C*

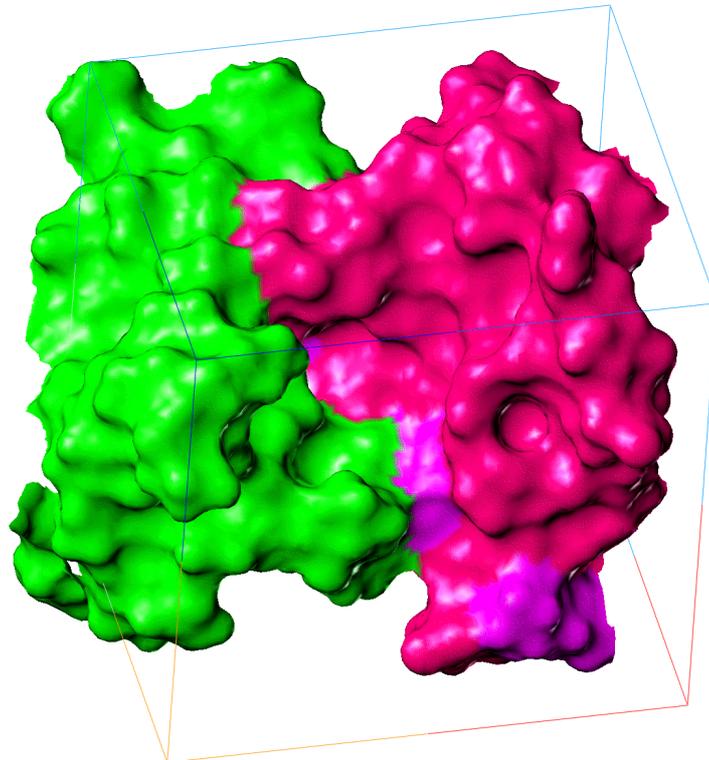
Cette superposition des conformations probables suggère que le segment Pro<sup>4</sup>-Ala<sup>7</sup>, peu « mobile », est en interaction avec l'anticorps. Le reste du peptide est plus flexible donc peu ou pas en interaction avec IgA I3.

## ***V - AMARRAGE DES ANTIGENES A L'ANTICORPS.***

### **1. DEFINITION DE LA BOITE D'AMARRAGE**

Nous avons défini une boîte d'amarrage unique, englobant largement toutes les boucles hypervariables et permettant la libre orientation des tous les ligands dans cet espace. Les grilles de potentiels sont calculées dans ce volume pour les types d'atome C, H, O, N, S, et pour le potentiel électrostatique. Tous les conformères semi-rigides des ligands, nonasaccharides et peptide P100C, sont arrimés dans cet espace représenté sur la Figure 11.

Les dimensions de cette boîte sont 30x26x30 Å<sup>3</sup> (23 625 Å<sup>3</sup>) avec 80x70x80 points par dimension et 0,375 Å entre chaque point (448 000 points par grille).



*Figure 11 : Position et dimensions de la boîte d'amarrage autour du site de reconnaissance de l'anticorps IgA I3 (chaîne lourde en rouge, chaîne légère en vert).*

## 2. AMARRAGE DU POLYSACCHARIDE SUR IGA I3

Le volume à explorer est important et les nonasaccharides, bien que partiellement rigides, comportent 28 pivots mobiles. Le nombre total de degrés de liberté est donc porté à 35.

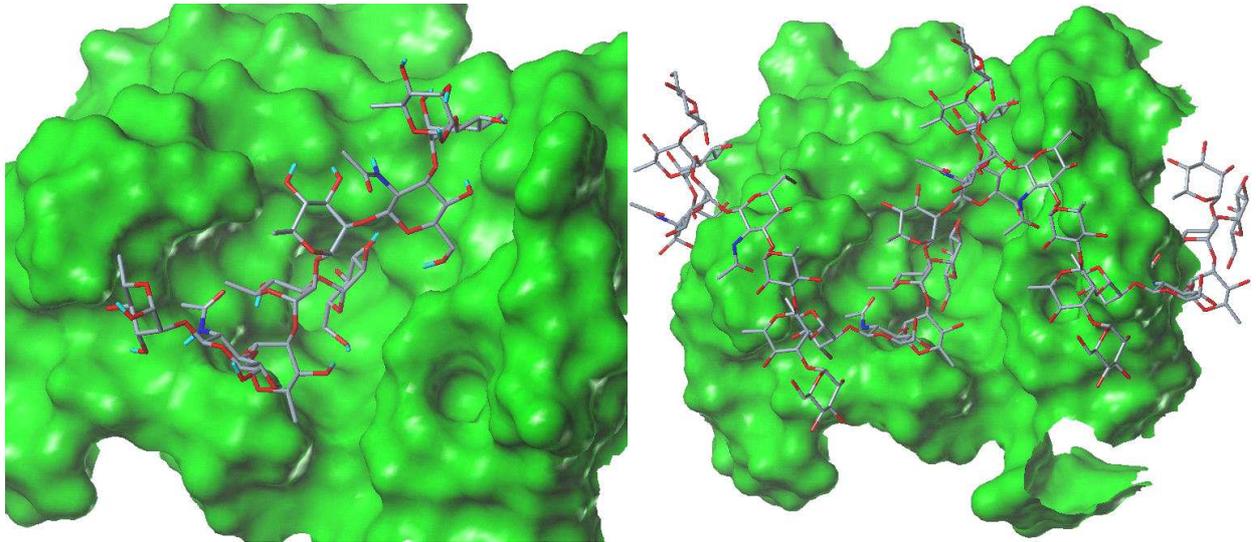
Afin d'explorer largement l'espace disponible et de permettre la convergence des solutions, nous avons réalisé l'amarrage avec une population de 125 individus. Les paramètres d'arrêt ont été repoussés à 10.000.000 d'évaluations d'énergie et 27.000 générations. Ces paramètres nous permettent d'obtenir des solutions dont les écarts d'énergie de liaison ne dépassent pas 5 kcal.mol<sup>-1</sup> entre la meilleure solution et la moins bonne.

D'autre part, pour avoir un échantillon de solutions statistiquement intéressant, nous avons calculé 240 solutions d'amarrage pour chaque conformère.

Les temps de calcul étant relativement longs, le travail a été distribué sur différentes machines par paquets de 20 solutions. Le logiciel AUTODOCK est par ailleurs limité à 140 solutions par exécution. Pour les machines SGI (système IRIX), les temps de calcul vont de 50 à 200 heures pour 20 solutions alors que sur PC (système Linux) les temps de calcul vont de 40 à 50 heures. Pour effectuer ces calculs nous disposions de plusieurs machines SGI et d'un seul PC Linux. D'autre part, nous n'étions pas certains d'obtenir exactement les mêmes résultats sur chaque plateforme. Afin de garantir l'homogénéité des résultats, l'ensemble des calculs a été effectué sur SGI. L'obtention de 240 solutions représente plus de 1000 heures de temps CPU pour chaque conformère.

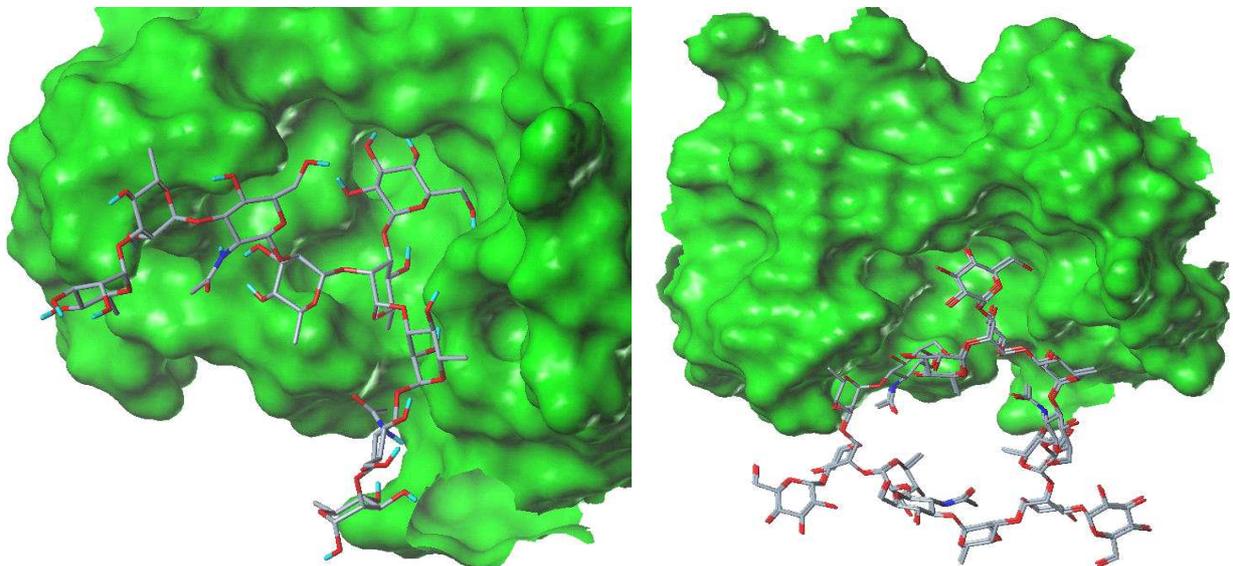
### *a. Amarrage de la forme condensée T1.*

Le nonasaccharide n'étant qu'un fragment du polysaccharide, il peut adopter des modes d'interaction absolument incompatibles avec le comportement du polysaccharide. Cela est dû en particulier à leurs différences de taille et d'encombrement stérique. Ainsi, lorsqu'on superpose le polysaccharide sur le fragment arrimé, celui-ci entre souvent en collision avec l'anticorps. L'analyse visuelle des solutions montre que 77 % d'entre elles (184/240) sont à rejeter pour cette raison. Les 56 solutions compatibles avec la superposition du polysaccharide se répartissent selon deux orientations distinctes c'est à dire deux modes de liaison.



*Figure 12 : Mode de liaison parallèle de T1 sur l'anticorps (à gauche) et superposition du polysaccharide sur T1 (à droite).*

La première orientation place T1 dans un axe parallèle au sillon de l'anticorps avec le résidu glucosyle E enfoui au fond de ce sillon (Figure 12). Le second mode d'amarrage place T1 perpendiculairement au sillon, le résidu E également enfoui dans la cavité de l'anticorps (Figure 13).



*Figure 13 : Mode de liaison perpendiculaire de T1 (à gauche) et superposition du polysaccharide sur T1 (à droite).*

Le relevé des liaisons hydrogène entre T1 et IgA I3 est donné sur le Tableau 1. Les données sont en colonnes avec, dans l'ordre de gauche à droite, le nom des résidus glycosidiques impliqués dans une liaison, le nom des fonctions du résidu sucre impliquées dans une liaison, le nom des

fonctions de l'anticorps à l'autre extrémité de la liaison, le nom du résidu portant la fonction, la boucle à laquelle appartient ce résidu et une indication de position de la fonction : « bb » si elle est sur la chaîne principale.

On remarque que dans les deux cas, T1 forme des interactions avec des résidus de la chaîne hypervariable H3, la plus spécifique de l'anticorps. D'autre part, pour chaque mode de liaison, le résidu E est le plus profondément enfoui dans le site de reconnaissance spécifique de l'anticorps. Parmi les solutions acceptables, celle de plus basse énergie de liaison correspond à l'orientation parallèle de T1. Cependant, la solution perpendiculaire de basse plus énergie n'a que 0,2 kcal.mol<sup>-1</sup> de plus.

*Tableau 1 : Liaisons hydrogène entre T1 et IgA I3 dans les modes de liaison parallèle (à gauche) et perpendiculaire (à droite). Les résidus de la boucle hypervariable H3 sont notés en rouge.*

| --- Oligosaccharide nonaT1 --- |   |                             |   |
|--------------------------------|---|-----------------------------|---|
| -- Solution parallèle          |   | -- Solution perpendiculaire |   |
| B:                             |   | B:                          |   |
| C:                             |   | C:                          | O2-H ... OH      SER <sup>29</sup> (L1)     |
| D:                             | HO6 ... H-N      GLY <sup>24</sup> (H1) bb  | D:                          |   |
| A:                             |   | A:                          |   |
| E:                             | O2-H ... O=C      ASP <sup>91</sup> (H3) bb | E:                          | O3-H ... O=C      ASP <sup>91</sup> (H3) bb |
|                                | O3-H ... O=C      ASP <sup>91</sup> (H3) bb |                             | O4-H ... O=C      ASP <sup>91</sup> (H3) bb |
|                                | HO3 ... H-N      TYR <sup>93</sup> (H3) bb  |                             |   |
| B:                             | O4-H ... O=C      SER <sup>93</sup> (L3) bb | B:                          |   |
| C:                             |   | C:                          | HO2 ... H-Nc      TRP <sup>41</sup> (H2)    |
|                                |   |                             | HO2 ... H-O      THR <sup>50</sup> (H2)     |
|                                |   |                             | c-O ... H-O      THR <sup>50</sup> (H2)     |
| D:                             |   | D:                          |   |
| A:                             | HO4 ... H-O      SER <sup>29</sup> (L1)     | A:                          |   |

### ***b. Amarrage de la forme étendue T4.***

L'amarrage du nonasaccharide T4, extrait de la forme étendue du polysaccharide, donne 102 solutions acceptables sur 240 solutions (42,5 %). Ces solutions se répartissent selon deux modes de liaison comme précédemment : un mode parallèle au sillon du site de reconnaissance et un mode perpendiculaire. Le résidu glucosyle est toujours enfoui au cœur du domaine. Là encore, la solution de plus basse énergie de liaison correspond à une orientation parallèle au sillon. L'écart d'énergie entre la meilleure solution parallèle et la meilleure solution perpendiculaire est de 2,0 kcal.mol<sup>-1</sup>. La première position regroupe 63 solutions et la seconde en regroupe 39. Le mode de liaison parallèle est représenté sur la Figure 14, le mode perpendiculaire est représenté sur la Figure 15.

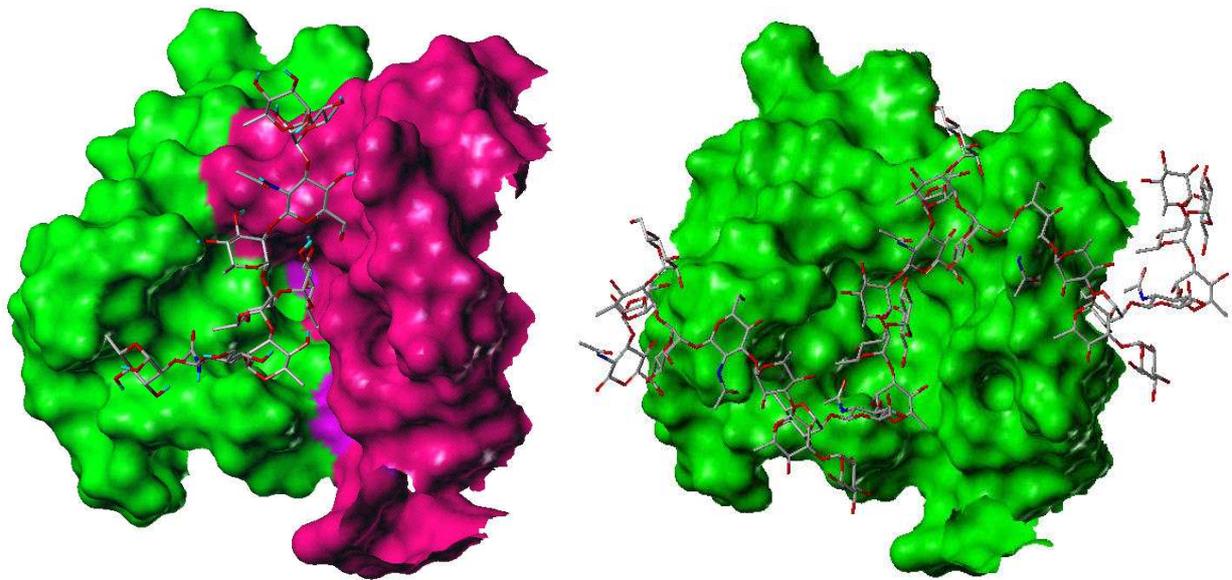


Figure 14 : Mode de liaison parallèle de T4 sur l'anticorps (à gauche) et superposition du polysaccharide sur T4 (à droite).

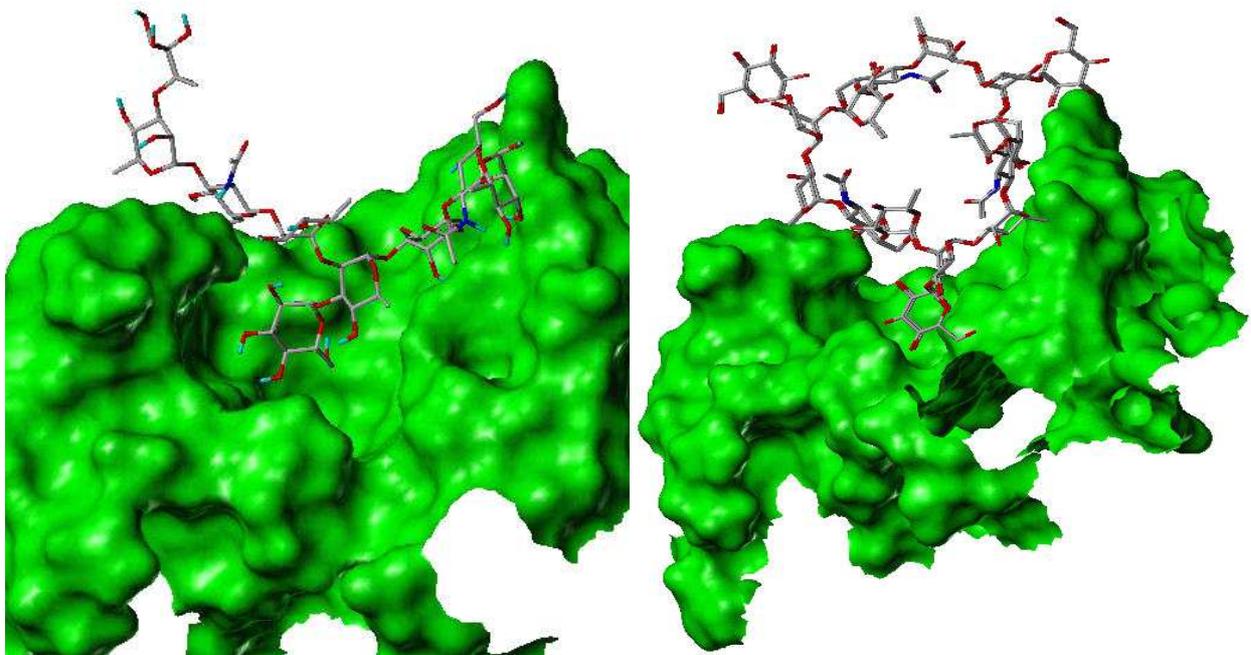


Figure 15 : Mode de liaison perpendiculaire de T4 sur l'anticorps (à gauche) et superposition du polysaccharide sur T4 (à droite).

Le relevé des liaisons hydrogène entre T4 et IgA I3 est donné sur le Tableau 2. Les données sont formatées de la même façon que dans le Tableau 1.

On remarque que là encore, T4 forme dans les deux cas des interactions avec des résidus de la chaîne hypervariable H3. D'autre part, pour chaque mode de liaison, le résidu E est enfoui dans la cavité du site de reconnaissance de l'anticorps. Parmi les solutions acceptables, celle de plus

basse énergie de liaison correspond à l'orientation parallèle de T4. La solution perpendiculaire de plus basse énergie a 2,0 kcal.mol<sup>-1</sup> de plus que la solution parallèle.

Tableau 2 : Listes des interactions entre l'anticorps et l'oligosaccharides (conformation étendue) associés selon 2 modes de liaison différents.

| --- Oligosaccharide nonaT4 --- |  |                             |   |
|--------------------------------|--|-----------------------------|---|
| -- Solution parallèle          |  | -- Solution perpendiculaire |   |
| B:                             |  | B:                          |   |
| C:                             |  | C:                          |   |
| D:                             | O4-H ... O=C ILE <sup>22</sup> (H1) bb | D:                          | Oc ... H-N(3+) LYS <sup>30</sup> (L)    |
|                                | O6-H ... O=C ALA <sup>90</sup> (H3) bb |                             | O6-H ... O=CO ASP <sup>92</sup> (H3)    |
|                                | HO6 ... H-N GLY <sup>24</sup> (H1) bb  |                             |   |
| A:                             | O3-H ... O=CO ASP <sup>92</sup> (H3)   | A:                          |   |
|                                | HO4 ... H-N(3+) LYS <sup>30</sup> (L1) |                             |   |
| E:                             | O3-H ... O=CO ASP <sup>91</sup> (H3)   | E:                          | O4-H ... OH SER <sup>93</sup> (L)       |
|                                | HO3 ... H-N TYR <sup>93</sup> (H3) bb  |                             | O6-H ... O=CO ASP <sup>91</sup> (H3)    |
|                                |  |                             | HO6 ... H-N TYR <sup>93</sup> (H3) bb   |
| B:                             | O4-H ... O=C SER <sup>93</sup> (L3) bb | B:                          |   |
| C:                             |  | C:                          | O6-H ... O=C-NH2 ASN <sup>43</sup> (H2) |
|                                |  |                             | Oc ... H-NH-CO ASN <sup>43</sup> (H2)   |
| D:                             |  | D:                          | O6-H ... OH TYR <sup>45</sup> (H2)      |
| A:                             |  | A:                          |   |

### 3. AMARRAGE DU PEPTIDE P100C SUR IGA I3

AUTODOCK3 a démontré sa capacité à proposer des solutions d'amarrage, de peptides dans un anticorps, très proches des observations cristallographiques [51].

Nous disposons de 20 conformations probables du peptide P100C. Chacun des conformères est arrimé à IgA I3 de façon entièrement rigide. Le faible nombre de degrés de liberté du ligand rigide nous permet de diminuer le temps de calcul en modifiant certains paramètres. Le logiciel a ainsi calculé 100 solutions pour chaque conformère avec une population de 75 individus et un nombre maximum de 750 000 évaluations de l'énergie. Ces paramètres permettent à l'algorithme de converger vers des solutions de basse énergie réparties sur une gamme de 8 kcal.mol<sup>-1</sup> et regroupées en une quinzaine de familles.

Lors de l'examen visuel des solutions, nous avons écarté toutes celles ne répondant pas aux critères suivants :

- contact de la Tyr<sup>2</sup> avec l'anticorps ;
- contact du segment Pro<sup>4</sup>-Leu<sup>5</sup>-Gly<sup>6</sup>-Ala<sup>7</sup> (PLGA) avec l'anticorps,;
- interactions avec la boucle H3.

Le Tableau 3 résume les résultats d'amarrage des 20 conformères sur IgA13. La colonne « H3 » indique s'il y a interaction avec le CDR-H3 et précise si l'interaction est établie par une fonction des chaînes latérales (CL) ou de la chaîne principale (bb). La colonne « Y<sup>2</sup> » concerne les interactions de la tyrosine<sup>2</sup> du peptide, et la colonne « PLGA » concerne les interactions du fragment PLGA du peptide. La meilleure solution acceptable est celle de plus basse énergie qui remplisse les critères de validité. Le rang d'une famille est attribué en fonction de l'énergie de sa meilleure solution ; la famille de rang 1 est celle de plus basse énergie mais elle n'est pas toujours acceptable. La colonne « Pop. Famille » indique le nombre de solutions regroupées dans la première famille acceptable. L'énergie de la famille de rang 1 est donnée pour indiquer l'écart entre la solution de plus basse énergie proposée par le logiciel et la meilleure de celles qui sont acceptables.

*Tableau 3 : Résultats d'amarrage des 20 conformères du peptide P100C. Relevé des interactions du peptide avec la boucle H3 de l'anticorps (H3), et des contacts de la Tyr<sup>2</sup> (T<sup>2</sup>) et du segment Pro-Leu-Gly-Ala (PLGA) du peptide avec la protéine. L'énergie de la meilleure solution acceptable est donnée avec son rang et sa population ainsi que la plus basse énergie proposée pour chaque conformère.*

| Conformère P100C | H3 | Y <sup>2</sup> | PLGA | Energie de la meilleure solution acceptable (kcal.mol <sup>-1</sup> ) | Rang de la première famille acceptable | Pop. Famille (/100) | Energie de la famille de rang 1 |
|------------------|----|----------------|------|---|--|---------------------|---------------------------------|
| 1                |    |                |      |   |  |                     | -9.95                           |
| 2                |    |                |      |   |  |                     | -9.23                           |
| 3                | CL |                | CL   | -10.5   | 1                                      | 31                  | -10.5                           |
| 4                |    |                |      |   |  |                     | -9.32                           |
| 5                |    |                |      |   |  |                     | -11.1                           |
| 6                |    |                |      |   |  |                     | -11.19                          |
| 7                | bb |                | CL   | -7.7  | 5                                      | 3                   | -8.54                           |
| 8                | CL |                | CL   | -7.48   | 9                                      | 1                   | -12.7                           |
| 9                |    |                |      | -7.84   | 4                                      | 15                  | -13.1                           |
| 10               | CL |                | CL   | -9.83   | 1                                      | 54                  | -9.83                           |
| 11               |    |                | CL   | -10.7   | 1                                      | 81                  | -10.7                           |
| 12               | CL |                |      | -8.05   | 3                                      | 3                   | -9.79                           |
| 13               | CL |                |      | -8.54   | 2                                      | 14                  | -9.86                           |
| 14               |    |                |      |   |  |                     | -10.1                           |
| 15               | CL | CL             | CL   | -9.26   | 1                                      | 27                  | -9.26                           |
| 16               | CL | CL             | CL   | -6.76   | 10                                     | 3                   | -11.2                           |
| 17               |    |                |      |   |  |                     | -9.54                           |
| 18               | CL |                | CL   | -7.34   | 10                                     | 2                   | -9.55                           |
| 19               | bb | CL             | CL   | -9.01   | 6                                      | 5                   | -10.6                           |
| 20               |    |                |      |   |  |                     | -9.51                           |

Cette fois ci, très peu des solutions proposées sont acceptables. Pour beaucoup d'entre elles, le pont salin entre les extrémités chargées N et C terminales du peptide, situé à l'opposé du segment PLGA, se trouve en contact avec l'anticorps et enfoui dans le sillon du site de reconnaissance. Or les études de transfert d'aimantation en RMN indiquent qu'elle n'est pas ou peu en interaction avec l'anticorps. De plus, cette portion du peptide est très hydrophile donc vraisemblablement en contact avec le solvant à l'extérieur de l'anticorps. Les résultats du logiciel peuvent s'expliquer par le fait que le solvant, en l'occurrence l'eau, n'est pas modélisé de façon explicite par la présence de molécules d'eau mobiles mais par des approximations implicites au travers de la fonction diélectrique et du terme entropique de désolvatation. Par conséquent, on peut supposer que la fonction énergétique sous-estime les effets hydrophobes. A l'inverse, les interactions électrostatiques peuvent être surestimées. Cette approximation est peu gênante dans le cas d'un site actif de petites dimensions et très concave où les interactions avec le solvant sont peu nombreuses. Dans le cas de cette étude, où le site est vaste et très ouvert, l'influence de molécules d'eau dans le mécanisme de reconnaissance est fort probable. L'implication de l'eau a notamment été constatée dans la structure cristalloraphique du fragment Fab de SYA/J6, un anticorps spécifique de *shigella flexneri Y*, en complexe avec un octapeptide [51]. Il serait possible de contourner cette limitation par l'intégration explicite de molécules d'eau liantes dans le modèle de l'anticorps mais nous n'avons aucune information quant à leur nombre et à leur position.

Néanmoins, les conformères **15**, **16** et **19** présentent chacun une solution d'amarrage proche des critères établis par l'analyse RMN (Tableau 3). Un relevé détaillé des interactions de ces solutions est donné dans le Tableau 4.

Pour le conformère **15**, la solution acceptable de plus basse énergie ( $-9.26 \text{ kcal.mol}^{-1}$ ) fait parti de la famille de rang 1 et représente 27 % des solutions trouvées.

Pour le conformère **16**, la solution acceptable de plus basse énergie ( $-6.76 \text{ kcal.mol}^{-1}$ ) est dans la famille de rang 10. Elle présente une moins bonne interaction avec l'anticorps que la solution du conformère **15** qui a une énergie inférieure. Sa famille ne regroupe que 3 % des solutions proposées.

Pour le conformère **19**, la solution acceptable de plus basse énergie ( $-9.01 \text{ kcal.mol}^{-1}$ ) est dans la famille de rang 6 qui regroupe 5 % des solutions proposées. Son énergie, bien que supérieure, est très proche de celle de la solution du conformère 15. Une interaction avec la boucle H3 est présente mais elle est peu spécifique car il s'agit d'une liaison hydrogène entre deux groupements des chaînes principales (LEU<sup>5</sup> bb N-H ... O=C bb ASP<sup>91</sup>) alors que dans la

solution du conformère **15**, l'interaction a lieu au niveau des chaînes latérales. De plus l'alanine du segment PLGA n'est pas en contact avec l'IgA, ce qui n'est pas en accord avec les données de RMN.

Parmi les 4800 solutions calculées, la meilleure solution du conformère **15** représente un des modes de liaison les plus probables de l'interaction entre ce peptide cyclique et l'immunoglobuline IgA I3. Cette solution est représentée sur la Figure 16 et sur la Figure 17.

Tableau 4 : Relevé des contacts et liaisons hydrogène entre les solutions d'amarrage des conformations 15, 16 et 19 du peptides P100C en interaction avec IgA13.

|                             |                         |           |
|-----------------------------|-------------------------|-----------|
| <b>Conformère 15</b>        |                         |           |
| Liaisons hydrogènes         |                         |           |
| <b>Tyr</b> <sup>2</sup> .NH | Ile <sup>22</sup> .O    | H1        |
| <b>Tyr</b> <sup>2</sup> .OH | Asn <sup>26</sup> .NH2  | H1        |
| Lys <sup>3</sup> .NH3+      | Asp <sup>92</sup> .COO- | <b>H3</b> |
| Contacts                    |                         |           |
| Tyr <sup>2</sup>            | Asp <sup>91</sup>       | H3        |
| Tyr <sup>2</sup>            | Trp <sup>41</sup>       | H2        |
| <b>Pro</b> <sup>4</sup>     | Trp <sup>41</sup>       | H2        |
| <b>Pro</b> <sup>4</sup>     | Pro <sup>98</sup>       | H3        |
| <b>Leu</b> <sup>5</sup>     | Phe <sup>94</sup>       | H3        |
| <b>Leu</b> <sup>5</sup>     | Ala <sup>90</sup>       | H3        |
| <b>Gly</b> <sup>6</sup>     | Val <sup>96</sup>       | L3 (2.3A) |
| <b>Gly</b> <sup>6</sup>     | Ala <sup>94</sup>       | L3        |
| <b>Ala</b> <sup>7</sup>     | Val <sup>96</sup>       | L3        |
| Leu <sup>8</sup>            | Lys <sup>30</sup>       | L1        |
| Cys <sup>11</sup>           | Asn <sup>43</sup>       |           |
| <b>Conformère 16</b>        |                         |           |
| Liaisons hydrogènes         |                         |           |
| <b>Tyr</b> <sup>2</sup> .OH | Ser <sup>93</sup> .OH   | L3        |
| Leu <sup>5</sup> .NH        | Asp <sup>91</sup> .CO   | <b>H3</b> |
| Contacts                    |                         |           |
| Tyr <sup>2</sup>            | Ser <sup>93</sup>       | L3        |
| <b>Pro</b> <sup>4</sup>     | Asp <sup>91</sup>       | H3        |
| <b>Pro</b> <sup>4</sup>     | Asp <sup>92</sup>       | H3        |
| Leu <sup>5</sup>            | Ala <sup>90</sup>       | H3        |
| Gly <sup>6</sup>            | Gly <sup>24</sup>       | H1        |
| <b>Conformère 19</b>        |                         |           |
| Liaisons hydrogènes         |                         |           |
| <b>Tyr</b> <sup>2</sup> .NH | Asp <sup>28</sup> .COO- | L1        |
| Leu <sup>5</sup> .NH        | ASP <sup>91</sup> .CO   | <b>H3</b> |
| Leu <sup>5</sup> .CO        | Gly <sup>24</sup> .NH   | H1        |
| Contacts                    |                         |           |
| Cis <sup>1</sup>            | Asp <sup>28</sup>       | L1        |
| Tyr <sup>2</sup>            | Asp <sup>28</sup>       | L1        |
| Tyr <sup>2</sup>            | Ser <sup>93</sup>       | L3        |
| Tyr <sup>2</sup>            | Asp <sup>92</sup>       | H3        |
| Lys <sup>3</sup>            | Lys <sup>30</sup>       | L1        |
| <b>Pro</b> <sup>4</sup>     | Asp <sup>91</sup>       | H3        |
| <b>Pro</b> <sup>4</sup>     | Asp <sup>92</sup>       | H3        |
| Leu <sup>5</sup>            | Ala <sup>90</sup>       | H3        |
| Leu <sup>5</sup>            | Gly <sup>24</sup>       | H1        |
| Gly <sup>6</sup>            | Gly <sup>24</sup>       | H1        |
| Gly <sup>6</sup>            | Trp <sup>41</sup>       | H2        |

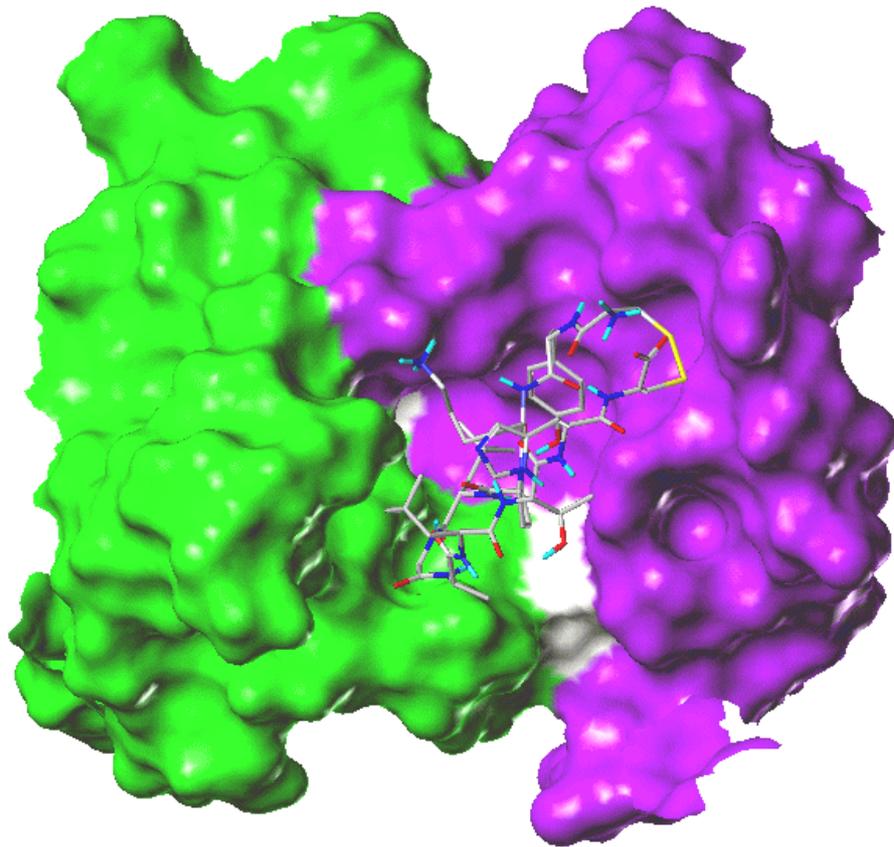


Figure 16 : Meilleure solution d'amarrage du conformère **15** de P100C sur IgA13. La chaîne légère de l'anticorps est représentée en vert, la chaîne lourde en violet.

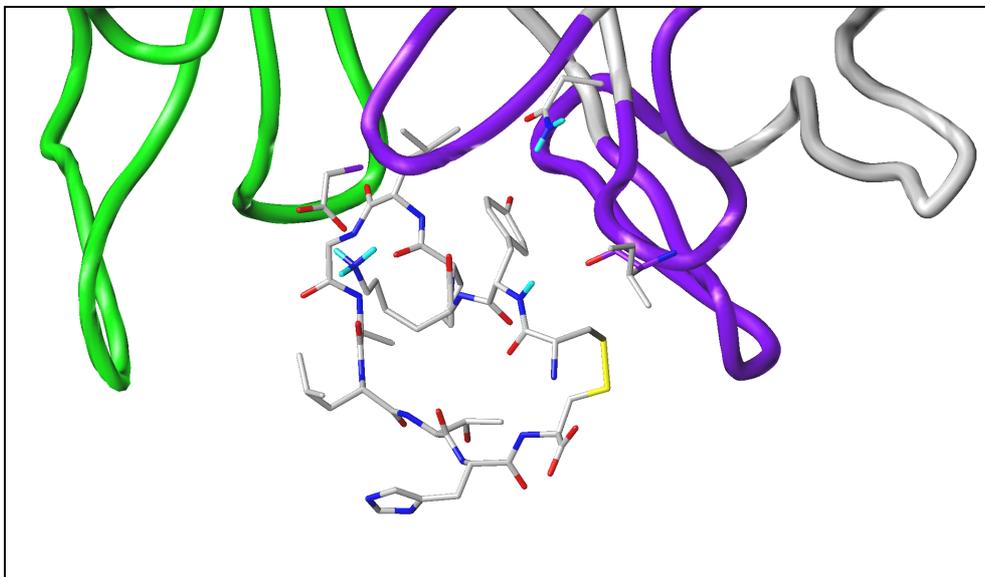
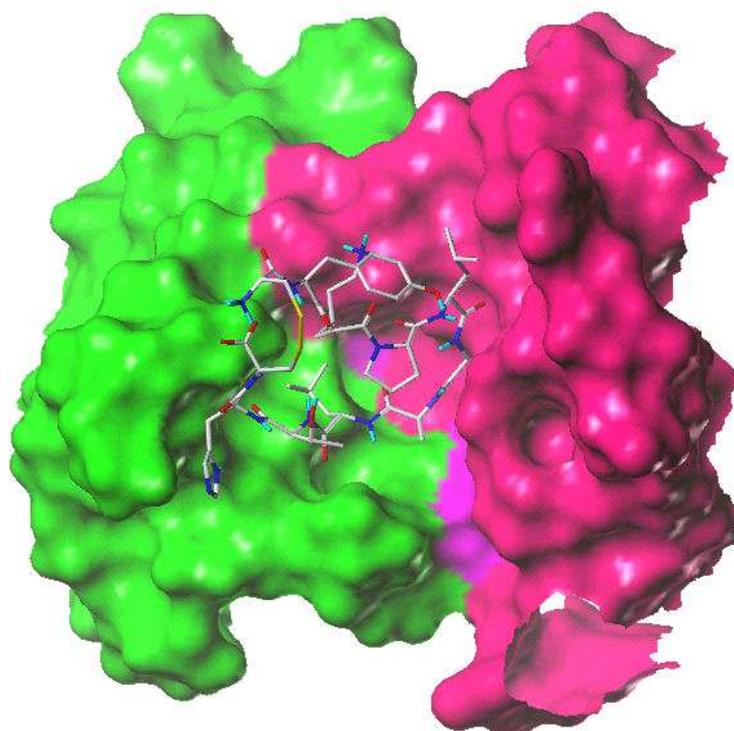
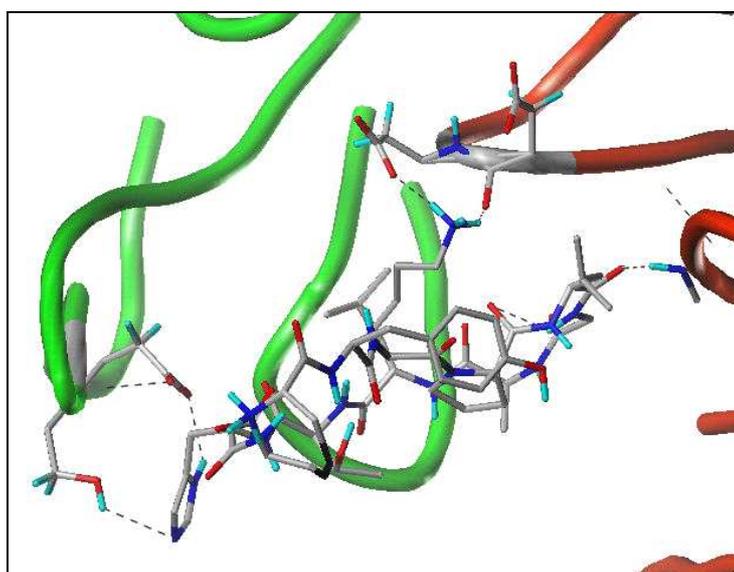


Figure 17 : Position et orientation du conformère **15** de P100C amarré à IgA13. La chaîne principale des boucles hypervariables de la chaîne légère est représenté en tubes verts, les boucles de la chaîne lourde en tubes violets. Les chaînes latérales de l'anticorps en interaction avec le peptide sont représentées en bâtonnets.

Les conformères **3** et **10** ont également fourni des solutions intéressantes. La position et l'orientation de ces solutions sont très proches. Bien que nous n'ayons pas relevé d'interactions entre la tyrosine 2 du peptide avec l'anticorps, les solutions de rang 1 forment des interactions avec le CDRH3 par les chaînes latérales. Le peptide est bien orienté et le segment PLGA est en interaction avec l'anticorps (Tableau 3). En outre, ces solutions forment les interactions les plus fortes avec  $-10,5 \text{ kcal.mol}^{-1}$  pour le conformère 3 et  $-9,83 \text{ kcal.mol}^{-1}$  pour le conformère **10**. La solution du conformère 10 est présentée sur la Figure 18 et la Figure 19.



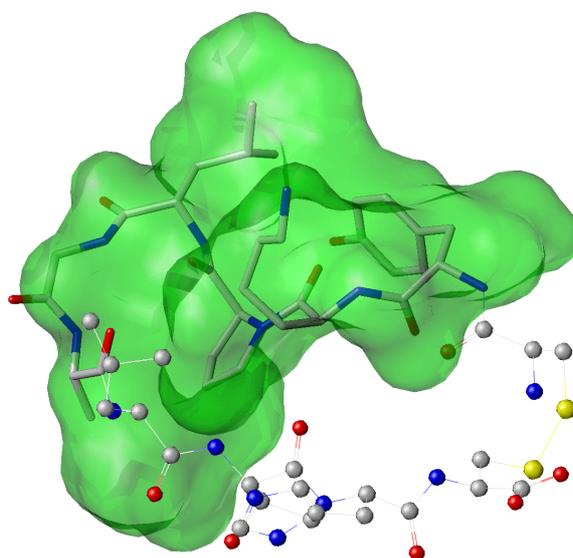
*Figure 18 : Solution d'amarrage du conformère 10 de P100C sur IgA13.*



*Figure 19 : Détail des interactions du conformère 10 de P100C avec IgA13 .*

## ***VI - COMPARAISON DES MODES D'INTERACTION DU POLYSACCHARIDE ET DU PEPTIDE.***

Si l'on compare les modes d'interactions du polysaccharide étendu (T4), en orientation parallèle, avec le mode d'interaction du conformère **15** de P100C, les points communs sont des liaisons hydrogène avec l'isoleucine 22 du CDRH1 et avec l'aspartate 92 du CDRH3. Le polysaccharide forme des liaisons hydrogène supplémentaires avec les résidus Ala<sup>90</sup> et Asp<sup>91</sup> du CDRH3 et Lys<sup>30</sup> du CDRL1. Le peptide est en contact avec ces mêmes résidus mais ne forme pas de liaisons hydrogène avec eux. Par contre il établit une liaison hydrogène avec le résidu Asn<sup>26</sup> du CDRH1.



*Figure 20 : Superposition du volume occupé par les résidus DA(E)BC du polysaccharide T4 en orientation parallèle (vert transparent) et du conformère **15** de P100C. Les résidus de P100C en contact avec l'anticorps sont représentés en bâtonnets, les atomes des résidus sans contact sont représentés par des balles.*

Si l'on compare les positions relatives de ces deux ligands, on constate que certains résidus du peptide occupent le même espace que les résidus DA(E)BC du polysaccharide. Ces résidus sont justement ceux que la RMN a indiqué comme étant en contact avec l'anticorps. A l'exception de l'extrémité du cycle aromatique du résidu Tyr<sup>2</sup> et de sa fonction phénol, les résidus Tyr<sup>2</sup> à Ala<sup>7</sup> du peptide, sont compris dans le volume occupé par le polysaccharide.

Une représentation en transparence du volume de l'unité minimale antigénique DA(E)BC est donnée sur la Figure 20. La solution du conformère **15** de P100C est également représentée sur

cette figure. Les résidus Tyr<sup>2</sup> à Ala<sup>7</sup> du peptide sont représentés en bâtonnets et les atomes des résidus non impliqués dans l'interaction sont représentés sous forme de balles.

Il semble que le premier aspect du mimétisme du peptide pour le polysaccharide repose sur une similarité de forme. Le peptide et le polysaccharide occupent le même espace dans l'anticorps.

Par contre, les points d'interaction sont reproduits de façon moins systématique. Hormis l'interaction avec le CDRH3, les liaisons hydrogène du peptide semblent n'avoir qu'un rôle de fixation au site de reconnaissance de l'anticorps. La spécificité de ces liaisons pourrait n'être que géométrique, sans assumer de rôle fonctionnel dans le processus immunogène.

En modifiant la nature des acides aminés du peptide, de façon à renforcer les interactions soit en mimant d'avantage les interactions du polysaccharide soit en exploitant d'autres sites de fixation de l'anticorps, nous améliorerons l'affinité du ligand et peut être son pouvoir immunogène.

## ***VII - CONCLUSION DE L'ETUDE***

Bien qu'aucune structure cristallographique de l'immunoglobuline IgA13 ne soit disponible, les structures précises d'anticorps homologues nous ont permis de construire un modèle fiable d'IgA13. Le logiciel AUTODOCK3 nous a permis de calculer des positions d'amarrage de l'antigène polysaccharidique dans cet anticorps ainsi que des positions probables pour le mime peptidique P100C de cet antigène. Les modèles ainsi obtenus apportent de nouvelles informations et contribuent à la compréhension des mécanismes moléculaires du mimétisme d'antigènes polysaccharidiques par des mimotopes peptidiques. Ils participent à la compréhension des règles nécessaires à la conception future de nouveaux vaccins efficaces basés sur des mimotopes. Les résultats de ce travail ont fait l'objet d'un article dans un journal scientifique à comité de lecture [72]. Cet article est fourni en Annexe III p.78.

## D - CONCLUSION

Le fait de disposer de la structure tridimensionnelle de la cible biologique est un atout majeur pour l'analyse et la conception de molécules bioactives. Les techniques de construction par homologie sont donc extrêmement importantes pour accéder à ces structures qui restent rares et font souvent défaut.

Grâce à elles, une analyse fine des interactions ligand-recepteur à l'échelle moléculaire est possible. La construction par homologie ne permet pas, cependant, de déterminer le nombre et la position des molécules d'eau liantes. Ces molécules d'eau ne sont pas liées de façon covalente mais sont néanmoins constitutives des protéines. Elles occupent des positions précises à leur surface et sont indispensables à leur fonctionnement. Seules les analyses précises aux rayons X permettent de les localiser. Lorsque cette information est disponible, il est possible de l'intégrer aux modèles, ce qui accroît souvent leur performances de façon significative [51]. Pouvoir modéliser systématiquement ces molécules, avant analyse expérimentale, constituerait une avancée importante pour ce type d'étude.

Une autre amélioration des modèles d'amarrage passe par l'intégration du polymorphisme dynamique des protéines. En effet, les logiciels considèrent encore les macromolécules comme des objets rigides ou très peu flexibles alors qu'ils sont en réalité très flexibles. Lors de l'amarrage du ligand au récepteur, le ligand s'adapte à la forme du récepteur mais le récepteur aussi se déforme pour « capturer » son ligand. Les chaînes latérales ainsi que les boucles ont des degrés de liberté importants. Nos modèles devront être capable de reproduire ce phénomène pour nous aider à concevoir des structures « parfaitement » adaptées à leur cible.

## E - BIBLIOGRAPHIE

1. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *The Protein Data Bank*. Nucleic Acids Research, 2000. **28**(1): p. 235-242.
2. Blundell, T.L., G. Elliott, S.P. Gardner, T. Hubbard, S. Islam, M. Johnson, D. Mantafoinis, P. Murray-Rust, J. Overington, and et al., *Protein engineering and design*. Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 1989. **324**(1224): p. 447-60.
3. Blundell, T., D. Carney, S. Gardner, F. Hayes, B. Howlin, T. Hubbard, J. Overington, D.A. Singh, B.L. Sibanda, and M. Sutcliffe, *Knowledge-based protein modeling and design*. European Journal of Biochemistry, 1988. **172**(3): p. 513-20.
4. Sutcliffe, M.J., F.R.F. Hayes, and T.L. Blundell, *Knowledge based modeling of homologous proteins. Part II: Rules for the conformations of substituted side chains*. Protein Engineering, 1987. **1**(5): p. 385-92.
5. Sutcliffe, M.J., I. Haneef, D. Carney, and T.L. Blundell, *Knowledge-based modeling of homologous proteins. Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures*. Protein Engineering, 1987. **1**(5): p. 377-84.
6. Needleman, S.B. and C.D. Wunsch, *General method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 1970. **48**(3): p. 443-53.
7. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison*. Proceedings of the National Academy of Sciences of the United States of America, 1988. **85**(8): p. 2444-8.
8. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool*. Journal of Molecular Biology, 1990. **215**(3): p. 403-10.
9. Pearson, W.R., *Effective protein sequence comparison*. Methods in enzymology, 1996. **266**: p. 227-58.
10. Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.
11. Karlin, S. and S.F. Altschul, *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes*. Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(6): p. 2264-8.
12. Karlin, S. and S.F. Altschul, *Applications and statistics for multiple high-scoring segments in molecular sequences*. Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(12): p. 5873-7.
13. Claverie, J.M. and D.J. States, *Information enhancement methods for large scale sequence analysis*. Computers & Chemistry (Oxford, United Kingdom), 1993. **17**(2): p. 191-201.
14. Altschul, S.F., M.S. Boguski, W. Gish, and J.C. Wootton, *Issues in searching molecular sequence databases*. Nature Genetics, 1994. **6**(2): p. 119-29.
15. Wootton, J.C. and S. Federhen, *Statistics of local complexity in amino acid sequences and sequence databases*. Computers & Chemistry (Oxford, United Kingdom), 1993. **17**(2): p. 149-63.
16. Jones, T.A. and S. Thirup, *Using known substructures in protein model building and crystallography*. EMBO Journal, 1986. **5**(4): p. 819-22.

17. Claessens, M., E. Van Cutsem, I. Lasters, and S. Wodak, *Modeling the polypeptide backbone with 'spare parts' from known protein structures*. Protein Engineering, 1989. **2**(5): p. 335-45.
18. Singh, U.C. and P.A. Kollman, *An approach to computing electrostatic charges for molecules*. Journal of Computational Chemistry, 1984. **5**(2): p. 129-45.
19. Laskowski, R.A., M.W. MacArthur, D.S. Moss, and J.M. Thornton, *PROCHECK: a program to check the stereochemical quality of protein structures*. Journal of Applied Crystallography, 1993. **26**(2): p. 283-91.
20. Morris, A.L., M.W. MacArthur, E.G. Hutchinson, and J.M. Thornton, *Stereochemical quality of protein structure coordinates*. Proteins 1992(12): p. 20.
21. Kuntz, I.D., J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin, *A geometric approach to macromolecule-ligand interactions*. Journal of Molecular Biology, 1982. **161**(2): p. 269-88.
22. Shoichet, B.K. and I.D. Kuntz, *Matching chemistry and shape in molecular docking*. Protein Engineering, 1993. **6**(7): p. 723-32.
23. Jain, A.N., *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine*. Journal of Medicinal Chemistry, 2003. **46**(4): p. 499-511.
24. Morris, G.M., D.S. Goodsell, R. Huey, and A.J. Olson, *Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4*. Journal of Computer-Aided Molecular Design, 1996. **10**(4): p. 293-304.
25. Morris, G.M., D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson, *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function*. Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.
26. Wesson, L. and D. Eisenberg, *Atomic solvation parameters applied to molecular dynamics of proteins in solution*. Protein science : a publication of the Protein Society, 1992. **1**(2): p. 227-35.
27. Bohm, 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*. Journal of computer-aided molecular design, 1994. **8**(3): p. 243-56.
28. Stouten, P.F.W., C. Froemmel, H. Nakamura, and C. Sander, *An effective solvation term based on atomic occupancies for use in protein simulations*. Molecular Simulation, 1993. **10**(2-6): p. 97-120.
29. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*. Journal of Medicinal Chemistry, 1985. **28**(7): p. 849-57.
30. Lamarck, J.B. Zoological Philosophie. 1914, London: Macmillan.
31. Solis F. J. and W.R. J.-B, Math. Oper. Res., 1981(6).
32. Kotloff, K.L., J.P. Winickoff, B. Ivanoff, J.D. Clemens, D.L. Swerdlow, P.J. Sansonetti, G.K. Adak, and M.M. Levine, *Global burden of Shigella infections: implications for vaccine development and implementation of control strategies*. Bulletin of the World Health Organization, 1999. **77**(8): p. 651-66.
33. Phalipon, A. and P.J. Sansonetti, *Shigellosis: innate mechanisms of inflammatory destruction of the intestinal epithelium, adaptive immune response, and vaccine development*. Critical reviews in immunology, 2003. **23**(5-6): p. 371-401.
34. MacLeod, C.M., R.G. Hodges, M. Heidelberg, and W.G. Bernhard, J. Exp. Med. , 1945(82): p. 61.
35. Roy, R., *New trends in carbohydrate-based vaccines*. Drug Discovery Today: Technologies, 2004. **1**(3): p. 327-336.

36. Goebel, H.H., K. Ikeda, F. Schulz, U. Burck, and A. Kohlschutter, *Fingerprint profiles in lymphocytic vacuoles of mucopolysaccharidoses I-H, II, III-A, and III-B*. *Acta neuropathologica*, 1981. **55**(3): p. 247-9.
37. Pozsgay, V., C. Chu, L. Pannell, J. Wolfe, J.B. Robbins, and R. Schneerson, *Protein conjugates of synthetic saccharides elicit higher levels of serum IgG lipopolysaccharide antibodies in mice than do those of the O-specific polysaccharide from Shigella dysenteriae type I*. *Proceedings of the National Academy of Sciences of the United States of America*, 1999. **96**(9): p. 5194-5197.
38. Peeters, C.C.A.M., D. Evenberg, P. Hoogerhout, H. Kayhty, L. Saarinen, C.A.A. Van Boeckel, G.A. Van der Marel, J.H. Van Boom, and J.T. Poolman, *Synthetic trimer and tetramer of 3-b-D-ribose-(1,1)-D-ribitol-5-phosphate conjugated to protein induce antibody responses to Haemophilus influenzae type b capsular polysaccharide in mice and monkeys*. *Infection and Immunity*, 1992. **60**(5): p. 1826-33.
39. Verez-Bencomo, V., V. Fernandez-Santana, E. Hardy, M.E. Toledo, M.C. Rodriguez, L. Heynngnezz, A. Rodriguez, A. Baly, L. Herrera, M. Izquierdo, A. Villar, Y. Valdes, K. Cosme, M.L. Deler, M. Montane, E. Garcia, A. Ramos, A. Aguilar, E. Medina, G. Torano, I. Sosa, I. Hernandez, R. Martinez, A. Muzachio, A. Carmenates, L. Costa, F. Cardoso, C. Campa, M. Diaz, and R. Roy, *A Synthetic Conjugate Polysaccharide Vaccine Against Haemophilus influenzae Type b*. *Science (Washington, DC, United States)*, 2004. **305**(5683): p. 522-525.
40. Pirofski, L.-a., *Polysaccharides, mimotopes and vaccines for fungal and encapsulated pathogens*. *Trends in Microbiology*, 2001. **9**(9): p. 445-451.
41. Johnson, M.A., A. Rotondo, and B.M. Pinto, *NMR Studies of the Antibody-Bound Conformation of a Carbohydrate-Mimetic Peptide*. *Biochemistry*, 2002. **41**(7): p. 2149-2157.
42. Monzavi-Karbassi, B., G. Cunto-Amesty, P. Luo, and T. Kieber-Emmons, *Peptide mimotopes as surrogate antigens of carbohydrates in vaccine discovery*. *Trends in Biotechnology*, 2002. **20**(5): p. 207-214.
43. Luo, P., M. Agadjanyan, J. Qiu, M.A.J. Westerink, Z. Steplewski, and T. Kieber-Emmons, *Antigenic and immunological mimicry of peptide mimotopes of Lewis carbohydrate antigens*. *Molecular Immunology*, 1998. **35**(13): p. 865-879.
44. Cunto-Amesty, G., T.K. Dam, P. Luo, B. Monzavi-Karbassi, C.F. Brewer, T.C. Van Cott, and T. Kieber-Emmons, *Directing the immune response to carbohydrate antigens. [Erratum to document cited in CA135:287167]*. *Journal of Biological Chemistry*, 2001. **276**(44): p. 41526.
45. Fleuridor, R., A. Lees, and L.-A. Pirofski, *A cryptococcal capsular polysaccharide mimotope prolongs the survival of mice with Cryptococcus neoformans infection*. *Journal of Immunology*, 2001. **166**(2): p. 1087-1096.
46. Lesinski, G.B. and M.A.J. Westerink, *Novel vaccine strategies to T-independent antigens*. *Journal of Microbiological Methods*, 2001. **47**(2): p. 135-149.
47. Maitta, R.W., K. Datta, A. Lees, S.S. Belouski, and L.-a. Pirofski, *Immunogenicity and efficacy of Cryptococcus neoformans capsular polysaccharide glucuronoxylomannan peptide mimotope-protein conjugates in human immunoglobulin transgenic mice*. *Infection and Immunity*, 2004. **72**(1): p. 196-208.
48. Buchwald, U.K., A. Lees, M. Steinitz, and L.-a. Pirofski, *A peptide mimotope of type 8 pneumococcal capsular polysaccharide induces a protective immune response in mice*. *Infection and Immunity*, 2005. **73**(1): p. 325-333.
49. Harris, S.L., L. Craig, J.S. Mehroke, M. Rashed, M.B. Zwick, K. Kenar, E.J. Toone, N. Greenspan, F.I. Auzanneau, J.R. Marino-Albernas, B.M. Pinto, and J.K. Scott, *Exploring the basis of peptide-carbohydrate crossreactivity: evidence for discrimination by peptides*

- between closely related anti-carbohydrate antibodies.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(6): p. 2454-9.
50. Young, A.C.M., P. Valadon, A. Casadevall, M.D. Scharff, and J.C. Sacchettini, *The three-dimensional structures of a polysaccharide binding antibody to Cryptococcus neoformans and its complex with a peptide from a phage display library: implications for the identification of peptide mimotopes.* Journal of Molecular Biology, 1997. **274**(4): p. 622-634.
  51. Vyas, N.K., M.N. Vyas, M.C. Chervenak, D.R. Bundle, B.M. Pinto, and F.A. Quiocho, *Structural basis of peptide-carbohydrate mimicry in an antibody-combining site.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(25): p. 15023-15028.
  52. Johnson, M.A. and B.M. Pinto, *Saturation-transfer difference NMR studies for the epitope mapping of a carbohydrate-mimetic peptide recognized by an anti-carbohydrate antibody.* Bioorganic & Medicinal Chemistry, 2004. **12**(1): p. 295-300.
  53. Ashkenazi, S., J.H. Passwell, E. Harlev, D. Miron, R. Dagan, N. Farzan, R. Ramon, F. Majadly, D.A. Bryla, A.B. Karpas, J.B. Robbins, and R. Schneerson, *Safety and immunogenicity of Shigella sonnei and Shigella flexneri 2a O-specific polysaccharide conjugates in children.* The Journal of infectious diseases, 1999. **179**(6): p. 1565-8.
  54. Wright, K., C. Guerreiro, I. Laurent, F. Baleux, and L.A. Mulard, *Preparation of synthetic glycoconjugates as potential vaccines against Shigella flexneri serotype 2a disease.* Organic & Biomolecular Chemistry, 2004. **2**(10): p. 1518-1527.
  55. Belot, F., C. Guerreiro, F. Baleux, and L.A. Mulard, *Synthesis of two linear PADRE conjugates bearing a deca- or pentadecasaccharide B epitope as potential synthetic vaccines against Shigella flexneri serotype 2a infection.* Chemistry--A European Journal, 2005. **11**(5): p. 1625-1635.
  56. Phalipon, A., C. Costachel, C. Grandjean, A. Thuizat, C. Guerreiro, M. Tanguy, F. Nato, B. Vulliez-Le Normand, F. Belot, K. Wright, V. Marcel-Peyre, P.J. Sansonetti, and L.A. Mulard, *Characterization of Functional Oligosaccharide Mimics of the Shigella flexneri Serotype 2a O-Antigen: Implications for the Development of a Chemically Defined Glycoconjugate Vaccine.* Journal of Immunology, 2006. **176**(3): p. 1686-1694.
  57. Phalipon, A., A. Folgori, J. Arondel, G. Sgaramella, P. Fortugno, R. Cortese, P.J. Sansonetti, and F. Felici, *Induction of anti-carbohydrate antibodies by phage library-selected peptide mimics.* European Journal of Immunology, 1997. **27**(10): p. 2620-2625.
  58. Pincus, S.H., S.R. Lepage, R.F. Jung, J.G. Massey, and M. Jaseja, *Initial studies of the molecular basis of peptide mimicry of group B streptococcal type III capsular polysaccharide.* International Reviews of Immunology, 2001. **20**(2): p. 221-227.
  59. Prinz Deborah, M., S.L. Smithson, and M.A.J. Westerink, *Two different methods result in the selection of peptides that induce a protective antibody response to Neisseria meningitidis serogroup C.* Journal of immunological methods, 2004. **285**(1): p. 1-14.
  60. Lindberg, A.A., P.D. Cam, N. Chan, L.K. Phu, D.D. Trach, G. Lindberg, K. Karlsson, A. Karnell, and E. Ekwall, *Shigellosis in Vietnam: seroepidemiologic studies with use of lipopolysaccharide antigens in enzyme immunoassays.* Reviews of infectious diseases, 1991. **13 Suppl 4**: p. S231-7.
  61. Clore, G.M. and A.M. Gronenborn, *Theory and applications of the transferred nuclear Overhauser effect to the study of the conformations of small ligands bound to proteins.* Journal of Magnetic Resonance (1969-1992), 1982. **48**(3): p. 402-17.
  62. Clore, G.M. and A.M. Gronenborn, *Theory of the time-dependent transferred nuclear Overhauser effect: applications to structural analysis of ligand-protein complexes in solution.* Journal of Magnetic Resonance (1969-1992), 1983. **53**(3): p. 423-42.

63. Phalipon, A., A. Cardona, J.-P. Kraehenbuhl, L. Edelman, P.J. Sansonetti, and B. Corthesy, *Secretory component: A new role in secretory IgA-mediated immune exclusion in vivo*. *Immunity*, 2002. **17**(1): p. 107-115.
64. Chothia, C., A.M. Lesk, A. Tramontano, M. Levitt, S.J. Smith-Gill, G. Air, S. Sheriff, E.A. Padlan, D. Davies, and W.R. Tulip, *Conformations of immunoglobulin hypervariable regions*. *Nature*, 1989. **342**(6252): p. 877-83.
65. Pokkuluri, P.R., F. Bouthillier, Y. Li, A. Kuderova, J. Lee, and M. Cygler, *Preparation, characterization and crystallization of an antibody Fab fragment that recognizes RNA. Crystal structures of native Fab and three Fab-mononucleotide complexes*. *Journal of Molecular Biology*, 1994. **243**(2): p. 283-97.
66. Romesberg, F.E., B. Spiller, P.G. Schultz, and R.C. Stevens, *Immunological origins of binding and catalysis in a Diels-Alderase antibody*. *Science (Washington, D. C.)*, 1998. **279**(5358): p. 1929-1933.
67. Herron, J.N., X.M. He, D.W. Ballard, P.R. Blier, P.E. Pace, A.L.M. Bothwell, E.W. Voss, Jr., and A.B. Edmundson, *An autoantibody to single-stranded DNA: comparison of the three-dimensional structures of the unliganded Fab and a deoxynucleotide-Fab complex*. *Proteins: Structure, Function, and Genetics*, 1991. **11**(3): p. 159-75.
68. Tulip, W.R., J.N. Varghese, W.G. Laver, R.G. Webster, and P.M. Colman, *Refined crystal structure of the influenza virus N9 neuraminidase-NC41 Fab complex*. *Journal of Molecular Biology*, 1992. **227**(1): p. 122-48.
69. Shoham, M., *Crystal structure of an anticholera toxin peptide complex at 2.3 Å*. *Journal of Molecular Biology*, 1993. **232**(4): p. 1169-75.
70. Ban, N., C. Escobar, R. Garcia, K. Hasel, J. Day, A. Greenwood, and A. McPherson, *Crystal structure of an idiotype-anti-idiotype Fab complex*. *Proceedings of the National Academy of Sciences of the United States of America*, 1994. **91**(5): p. 1604-8.
71. Clement, M.-J., A. Imberty, A. Phalipon, S. Perez, C. Simenel, L.A. Mulard, and M. Delepierre, *Conformational Studies of the O-specific Polysaccharide of Shigella flexneri 5a and of Four Related Synthetic Pentasaccharide Fragments Using NMR and Molecular Modeling*. *Journal of Biological Chemistry*, 2003. **278**(48): p. 47928-47936.
72. Clement, M.-J., A. Fortune, A. Phalipon, V. Marcel-Peyre, C. Simenel, A. Imberty, M. Delepierre, and L.A. Mulard, *Toward a Better Understanding of the Basis of the Molecular Mimicry of Polysaccharide Antigens by Peptides using Shigella flexneri 5a*. *Journal of Biological Chemistry*, 2006. **281**(4): p. 2317-2332.

## F - ANNEXES

### Annexe I : Abréviations et codage alphabétique des acides aminés selon la règle FASTA.

|               |     |          |
|---------------|-----|----------|
| Alanine       | Ala | <b>A</b> |
| Arginine      | Arg | <b>R</b> |
| Asparagine    | Asn | <b>N</b> |
| Aspartic Acid | Asp | <b>D</b> |
| Cysteine      | Cys | <b>C</b> |
| Glutamine     | Gln | <b>Q</b> |
| Glutamic Acid | Glu | <b>E</b> |
| Glycine       | Gly | <b>G</b> |
| Histidine     | His | <b>H</b> |
| Isoleucine    | Ile | <b>I</b> |
| Leucine       | Leu | <b>L</b> |
| Lysine        | Lys | <b>K</b> |
| Methionine    | Met | <b>M</b> |
| Phenylalanine | Phe | <b>F</b> |
| Proline       | Pro | <b>P</b> |
| Serine        | Ser | <b>S</b> |
| Threonine     | Thr | <b>T</b> |
| Tryptophan    | Trp | <b>W</b> |
| Tyrosine      | Tyr | <b>Y</b> |
| Valine        | Val | <b>V</b> |

## Annexe II : Document de prise en main rapide d'Autodock3.

### AUTODOCK 3.05

### MARCHE A SUIVRE

#### I Introduction:

Autodock est une suite d'exécutables et de scripts permettant d'obtenir des solutions d'amarrage d'un **ligand flexible** sur une **cible rigide**. Autodock est particulièrement adapté à l'étude de petits ligands (comprenant 28 pivots au maximum) en interaction avec des protéines.

#### II Principes:

On définit une zone (boîte) à la périphérie de la macromolécule dans laquelle on souhaite chercher des solutions d'amarrage pour un ligand.

Autogrid calcule alors une matrice de points régulièrement espacés dans le volume de cette boîte. Pour chacun des types d'atome présent dans le ligand, Autogrid calcule l'énergie d'interaction entre un atome de ce type placé en un point de la matrice et l'ensemble de la macromolécule. Ainsi, si le ligand est constitué d'atomes de carbone, d'oxygène, d'hydrogène et d'azote, il y aura une matrice pour le carbone, une pour l'oxygène, une pour l'hydrogène et une pour l'azote. Autogrid calcule également une grille de potentiel électrostatique correspondant à l'interaction d'une charge +1eV avec l'ensemble de la macromolécule et ce pour chacun des points de la matrice.

Autodock fait ensuite varier les différents degrés de liberté du ligand (axes de translation, de rotation et angles dièdres) à l'intérieur de la boîte. Il utilise les grilles de potentiels pour calculer l'énergie d'interaction entre un conformère du ligand en une position p et la macromolécule.

Cette méthode permet de calculer très rapidement l'énergie d'interaction ligand - récepteur car les potentiels sont pré-calculés. Autodock propose plusieurs méthodes de recherche conformationnelle dont la dynamique moléculaire par un algorithme de Monte-Carlo et des algorithmes génétiques.

### III Procédure Standard

Préparer les sources

Générer le fichier de la macromolécule.

Générer le fichier du ligand.

Générer le fichier de paramétrage d'Autogrid.

Générer le fichier de paramétrage d'Autodock.

Calcul des grilles de potentiels avec Autogrid.

Recherche et proposition de solution d'amarrage avec Autodock.

Analyser et visualiser les résultats.

### IV Procédure Détaillée :

Préparer un fichier au format mol2 pour le ligand et pour la macromolécule (macro) :

Pour le ligand, vérifier que :

- les types atomiques sont corrects,
- tous les atomes d'hydrogène sont présents,
- les valences sont complètes.
- Assigner les charges partielles par la méthode qui convient (MOPAC, Gasteiger ...) et sauvegarder au format mol2.

Pour la macro:

- supprimer tous les hydrogènes,
- ajouter les hydrogènes essentiels avec le module BIOPOLYMER de Sybyl,
- assigner les charges partielles avec BIOPOLYMER de Sybyl (type KOLL UNI),
- supprimer les paires libres,
- sauvegarder au format mol2.

Créer un répertoire et y placer les fichiers *macro.mol2* et *ligand.mol2*.

Générer le fichier *macro.pdbqs* :

- %>**cnvmol2topdbq macro.mol2 > macro.pdbq**

crée le fichier *macro.pdbq* à partir du fichier *macro.mol2*

- %>**rem-lp macro.pdbq**

supprime les doublets libres ;

- %>**addsol macro.pdbq macro.pdbqs**

ajoute les paramètres de solvatation et crée le fichier *macro.pdbqs*

Générer le fichier *ligand.pdbq* :

- %>**deftors ligand.mol2**

deftors permet de définir la partie fixe par rapport à laquelle les pivots tournent et la liste des pivots qui peuvent tourner (32 max).

Générer le fichier des paramètres d'Autogrid :

- %>**mkgpf3 ligand.pdbq macro.pdbqs**

crée le fichier *macro.gpf* (Grid Parameter File) ;

- %>**mkbox macro.gpf >! macro.gpf.box.pdb**

générer un fichier contenant la position et les dimensions de la boîte au format pdb ;

- visualiser la boîte avec Sybyl pour vérifier que sa position et ses dimensions englobent bien le site de la macro à étudier ;

- éditer le fichier *macro.gpf* et corriger les paramètres de la boîte à sa convenance ;

- régénérer le fichier *macro.gpf.box.pdb* comme précédemment et recommencer jusqu'à obtention de la boîte désirée;

Générer le fichier des paramètres d'Autodock :

**%>mkdpf3 ligand.pdbq macro.pdbqs**

crée le fichier *ligand.macro.dpf* (Dock Parameter File)

- éditer le fichier *ligand.macro.dpf* et ajuster les paramètres (choix de l'algorithme de recherche, précision de recherche, nombre de cycles, etc.).

Calculer les matrices de potentiels :

**- %>autogrid3 -p macro.gpf -l macro.glg &**

crée le fichier *macro.glg* (Grid LoG) et les fichiers contenant les matrices.

Lancer le docking :

**- %>autodock3 -p ligand.macro.dpf -l ligand.macro.dlg &**

crée le fichier *ligand.macro.dlg* qui contient tous les résultats.

Analyser les résultats :

- éditer le fichier *ligand.macro.dlg*. L'avant dernière partie du fichier contient un histogramme regroupant les solutions proposées par similitude et les classe par énergie croissante. Ces histogrammes donnent quelques statistiques et permettent d'identifier la meilleure solution proposée.

Visualiser les résultats :

**- %>get-docked ligand.macro.dlg**

extrait les coordonnées des solutions souhaitées et les rassemble toutes dans le fichier *ligand.macro.dlg.pdb* au format pdb. Visualiser les solutions dans Sybyl.

## V Techniques, Trucs et Astuces :

### **Préparation des fichiers sources (ligand.mol2, macro.mol2) :**

Positionner le ligand à quelques angström du site récepteur à étudier avant de sauvegarder (créer les fichiers ligand.mol2 et macro.mol2). Le script mkgpf3 positionne le centre de la boîte sur le centre du ligand et dimensionne celle-ci en fonction de la taille du ligand. Il est donc important que le ligand et le récepteur aient des coordonnées relatives proches mais qu'ils ne se superposent pas.

### **Deftors :**

Deftors édite la liste des pivots présents dans le ligand, les numérote et donne les numéros des atomes des angles dièdres. Sélectionner les pivots qui seront mobiles au cours du calcul.

Le nombre de pivots est limité à 32.

Pour chaque pivot défini, autodock recalculera toutes les interactions non liantes internes au ligand après chaque modification de conformation. Ce nombre d'interactions non liantes est limité à 2096 au total.

### **Modification de la boîte d'amarrage (fichier GPF) :**

La position et les dimensions de la boîte sont calculées par rapport au ligand.

- Pour modifier les **dimensions**, il faut jouer sur le nombre de points de matrice sur chaque axe défini par le paramètre **npts x y z**. On peut également modifier le paramètre **spacing** qui définit l'espacement entre chaque point.
- Pour modifier la **position** de la boîte, il faut modifier les coordonnées de son centre défini par le paramètre **gridcenter X Y Z**.
- Utiliser la commande **mkbox** pour visualiser la boîte. Le coin orange est l'origine des coordonnées, en rouge l'axe **X**, en blanc l'axe **Y**, en bleu l'axe **Z**.

### **Paramétrage d'Autodock (fichier DPF):**

L'algorithme d'exploration conformationnelle le plus efficace est de loin l'Algorithme Génétique Lamarckien (LGA). Il comprend 1 optimiseur global (GA) + 1 optimiseur local (Solis & Wets). Utiliser donc le fichier DPF généré par défaut par mkdpf3 qui est pré-paramétré pour lancer un LGA.

Convergence :

La convergence dépend des conditions d'arrêt qui sont :

- Nombre d'évaluations de l'énergie = GA\_NUM\_EVALS;
- Nombre maximum de générations = GA\_NUM\_GENERATIONS;

Le fichier DLG indique pour chaque run le nombre de générations calculées. Il faut atteindre au moins 1500 générations pour avoir une convergence correcte. Plus on a de générations plus les résultats convergent vers la meilleure solution.

Exploration de l'espace conformationnel :

Plus il y a de degrés de liberté dans le ligand, plus l'espace conformationnel est vaste (augmentation exponentielle !). Pour augmenter l'étendue de l'exploration il faut augmenter le nombre de générations calculées (repousser les conditions d'arrêt) et le nombre de directions

d'exploration c'est à dire le nombre d'individus dans la population. Cette variable est définie par le paramètre GA\_POP\_SIZE.

Nombre de cycles (runs) :

Chaque cycle de LGA donne par défaut (GA\_ELITISM) 1 solution (la meilleure trouvée).

Il faut répéter l'opération jusqu'à avoir un échantillon de solutions représentatif. La taille de cet échantillon dépend du nombre de degrés de liberté du ligand, des dimensions de la boîte à explorer et de la dispersion des résultats. 50 solutions est un minimum.

Le nombre de cycles à exécuter est défini par la variable GA\_RUN.

Il vaut mieux faire plusieurs job en même temps avec un petit nombre de cycles qu'un seul job avec beaucoup de cycle ! Un cycle peut facilement durer 2h ... même au CECIC ! Le travail va plus vite et en cas d'erreur on perd beaucoup moins de temps.

### **Programmation des jobs :**

Règle à suivre : 1 seul job par CPU !

Il est plus rapide de faire 2 jobs l'un après l'autre sur une même CPU que 2 jobs en même temps (si si c'est prouvé !). Il faut donc programmer le système pour qu'il lance les jobs les uns après les autres. Utiliser pour cela la commande **at** :

```
at -f nomfichier -t HH:MM [mois JJ]
```

#### Exemple :

Créer un fichier commande.txt contenant la ligne "autodock3 -p *ligand.macro.dpf* -l *ligand.macro.dlg* &" (sans les caractères ").

```
%> at -f commande.txt -t 01:00 avr 07
```

lance la ligne de commande contenue dans le fichier *commande.txt* à 1h00 du matin le 07 avril. L'heure est obligatoire, la date est facultative.

Il faut donc bien évaluer la dure d'un job pour savoir quand doit partir le suivant. Avec Autodock, la dure d'un job augmente de façon linéaire avec le nombre de cycles. Faire un test sur 1 cycle (ga\_run 1) pour évaluer la durée totale du job ... Il ne reste plus qu'à prier pour que personne ne vienne poser un job supplémentaire sur la CPU utilisée. Une chapelle ardente a été dressée à cet effet : 1ere porte à gauche en sortant de la salle info du 2eme étage.

Autre règle à suivre : **lire la doc !...**

**Annexe III : Article du *Journal of Biological Chemistry*.**

# Toward a Better Understanding of the Basis of the Molecular Mimicry of Polysaccharide Antigens by Peptides

## THE EXAMPLE OF *SHIGELLA FLEXNERI* 5A\*<sup>[5]</sup>

Received for publication, September 15, 2005, and in revised form, October 25, 2005. Published, JBC Papers in Press, October 26, 2005, DOI 10.1074/jbc.M510172200

Marie-Jeanne Clément<sup>‡</sup>, Antoine Fortuné<sup>§</sup>, Armelle Phalipon<sup>¶</sup>, Véronique Marcel-Peyre<sup>¶</sup>, Catherine Simenel<sup>‡</sup>, Anne Imberty<sup>||</sup>, Muriel Delepierre<sup>‡1</sup>, and Laurence A. Mulard<sup>\*\*</sup>

From the <sup>‡</sup>Unité de RMN des Biomolécules, URA CNRS 2185, Institut Pasteur, <sup>\*\*</sup>Unité de Chimie Organique, URA CNRS 2128, Institut Pasteur, <sup>¶</sup>Unité de Pathogénie Microbienne Moléculaire, Institut Pasteur, 28 Rue du Dr. Roux, 75724 Paris Cedex 15, <sup>§</sup>DPM UMR5063 UJF/CNRS, 5 Avenue de Verdun 38240 Meylan, France, and the <sup>||</sup>CERMAV-CNRS (affiliated with Université Joseph Fourier), 38041 Grenoble BP53, Cedex 09, France

Protein conjugates of oligosaccharides or peptides that mimic complex bacterial polysaccharide antigens represent alternatives to the classical polysaccharide-based conjugate vaccines developed so far. Hence, a better understanding of the molecular basis ensuring appropriate mimicry is required in order to design efficient carbohydrate mimic-based vaccines. This study focuses on the following two unrelated sets of mimics of the *Shigella flexneri* 5a O-specific polysaccharide (O-SP): (i) a synthetic branched pentasaccharide known to mimic the average solution conformation of *S. flexneri* 5a O-SP, and (ii) three nonapeptides selected upon screening of phage-displayed peptide libraries with two protective murine monoclonal antibodies (mAbs) of the A isotype specific for *S. flexneri* 5a O-SP. By inducing anti-O-SP antibodies upon immunization in mice when appropriately presented to the immune system, the pentasaccharide and peptides p100c and p115, but not peptide p22, were qualified as mimotopes of the native antigen. NMR studies based on transferred NOE (trNOE) experiments revealed that both kinds of mimotopes had an average conformation when bound to the mAbs that was close to that of their free form. Most interestingly, saturation transfer difference (STD) experiments showed that the characteristic turn conformations adopted by the major conformers of p100c and p115, as well as of p22, are clearly involved in mAb binding. These latter experiments also showed that the branched glucose residue of the pentasaccharide was a key part of the determinant recognized by the protective mAbs. Finally, by using NMR-derived pentasaccharide and peptide conformations coupled to STD information, models of antigen-antibody interaction were obtained. Most interestingly, only one model was found compatible with experimental data when large O-SP fragments were docked into one of the mIgA-binding sites. This newly made available system provides a new contribution to the understanding of the molecular mimicry of complex polysaccharides by peptides and short oligosaccharides.

Bacterial capsular polysaccharides (CPS)<sup>2</sup> and lipopolysaccharides (LPS) are known to be important virulence factors and major targets of the protective immune response of the host (1). Several polysaccharide vaccines such as those targeting *Streptococcus pneumoniae*, *Neisseria meningitidis*, or *Salmonella typhi* were proven efficient in adults and are thus commercially available. Their ineffectiveness in infants has been successfully circumvented with the licensing of polysaccharide:protein conjugates such as those targeting *Haemophilus influenzae* b, *S. pneumoniae*, and *N. meningitidis* group C infections (2). A possible alternative may derive from the use of accurate synthetic mimics of the bacterial polysaccharide antigens. This innovative approach has been mostly developed along two lines, including the use of either synthetic oligosaccharides or peptides mimicking the carbohydrate determinants recognized by anti-carbohydrate monoclonal antibodies (mAb) conferring protection in experimental models of infection. Indeed, semi-synthetic glycoconjugates incorporating oligosaccharides mimicking fragments of bacterial polysaccharide antigens were shown to be highly immunogenic in mice (3–5). The “proof of concept” was recently demonstrated in humans with the efficacy of such a semi-synthetic glycoconjugate in protecting against *H. influenzae* b infection (6).

However, access to the required carbohydrate haptens is often a roadblock. Therefore, besides the investigation of anti-idiotypic antibody (7), expanding the concept of mimicry led in the recent past to extensive exploring of the potential mimicking of polysaccharide and/or complex oligosaccharide antigens by peptides (8–10). These peptide mimotopes, *i.e.* peptide mimics inducing an anti-carbohydrate antibody response upon immunization, have been proposed as potential surrogate antigens of carbohydrates in vaccine development (10). Indeed, because of their ease of manufacture and their intrinsic immunogenic properties, peptide mimotopes may have greater advantage over complex carbohydrate haptens issued from bacterial cell cultures or low yielding multi-step syntheses. However, not all peptide mimics of carbohydrate antigens behave as mimotopes. Despite the large number of known peptide mimics, only few peptide mimotope-based experimental vaccines have been reported so far (11–16).

It is believed that a better understanding of the molecular basis of peptide-carbohydrate mimicry could help the rational design of potent peptide mimotope-based vaccines. In particular, whether mimicry is

\* This work was supported by MENRT (Programme de Recherche Fondamentale en Microbiologie et Maladies Infectieuses et Parasitaires), Délégation Générale pour l'Armement Contract 99 34 029, and the CNRS Program Physique et Chimie du Vivant. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

<sup>[5]</sup> The on-line version of this article (available at <http://www.jbc.org>) contains Figs. 15–45.

<sup>1</sup> To whom correspondence should be addressed: Unité de RMN des Biomolécules, Institut Pasteur, 28 Rue du Dr. Roux, 75724 Paris Cedex 15, France. Tel.: 33-1-45-68-88-71; Fax: 33-1-45-68-89-29; E-mail: [muriel@pasteur.fr](mailto:muriel@pasteur.fr).

<sup>2</sup> The abbreviations used are: CPS, capsular polysaccharides; NOE, nuclear Overhauser effect; tr, transferred; ROESY, rotating frame nuclear Overhauser enhancement spectroscopy; NOESY, nuclear Overhauser effect spectroscopy; ELISA, enzyme-linked immunosorbent assay; Ab, antibody; mAb, monoclonal Ab; STD, saturation transfer difference; O-SP, O-specific polysaccharide; LPS, lipopolysaccharides; PBS, phosphate-buffered saline; BSA, bovine serum albumin; TOCSY, total correlation spectroscopy.



and ROESY experiments were acquired using a mixing time of 80 and 400 ms, respectively. Water suppression was performed using the WATERGATE pulse sequence (45). All NMR spectra were collected in the phase-sensitive mode using the States-Haberhorn method (46).

**Ligand-Antibody Interactions**—Shigemi tubes were used for all samples. In order to prepare NMR samples of pentasaccharide DA(E)BC-OMe in the presence of the antibodies mIgA C5 and mIgA I3, mAbs were concentrated after repeated cycles of exchange with D<sub>2</sub>O buffer (50 mM deuterated sodium phosphate, 100 mM NaCl, pH 6.5) in Amicon Centriprep-10 concentrators. trNOE experiments (34, 35, 47) performed on different pentasaccharide:binding site ratios (5:1, 10:1, 15:1, 20:1, and 30:1) showed that the most favorable ratio for trNOE was 20:1. So the final samples were prepared with 3.75  $\mu$ M antibody and 0.3 mM pentasaccharide in 380  $\mu$ l of the above mentioned D<sub>2</sub>O buffer. trNOE, trROE, and STD experiments (36) on pentasaccharide DA(E)BC-OMe in the presence of mIgA C5 and mIgA I3 were recorded at 500 and 600 MHz, respectively. trNOE experiments were performed with mixing times of 100, 150, 250, 300, and 400 ms at 303 K to obtain build-up curves and trROE with a mixing time of 400 ms.

The conformation of the free peptides was studied at pH 5. However, with this pH value being close to the isoelectric points of the mIgAs, a study of peptides in their bound conformation was performed at pH 6.5 to avoid precipitation of the mAb. Similarly to the DA(E)BC-mIgA complexes, a peptide:antibody-binding site ratio of 20:1 was used (0.3 mM:3.75  $\mu$ M). trNOE experiments were performed with mixing times of 100, 150, 250, 300, and 400 ms. To be sure that the observed negative cross-peaks were real trNOEs, NOESY spectra were recorded under the same pH, temperature, and concentration values with the peptides alone. Furthermore, to discard any impact on NOE effects of viscosity increase as a result of the mAb presence, a NOESY spectrum ( $\tau_m = 200$  ms) of p115 was registered in the presence of BSA at the same concentration ratio as that used with the mIgAs. Because no negative NOE cross-peaks were observed in either case, it was assumed that the negative NOEs observed in the presence of mIgA were trNOEs.

Selective saturation of antibody resonances were performed for all STD-NMR experiments at 0.3 ppm (30 ppm for reference spectra) using a series of 40 gaussian-shaped pulses (50- and 10-ms delay between pulses, excitation width  $\gamma B_1/2\pi$ , approximately 50 Hz) for a total saturation time of 2.4 s. The one-dimensional STD spectra were recorded with 4096 scans at 288 and 298 K for the pentasaccharide and the peptides, respectively. Subtraction of saturated spectra from reference spectra was obtained by phase cycling (36). For DA(E)BC-OMe, two STD-TOCSY experiments (48) were recorded with selective saturation at 0.3 and 30 ppm, respectively. Differences between the two spectra were performed using the VNMR software. No attempt here was made to quantify STD-NMR intensities, as it is known that these exhibit a complex dependence on relaxation times, correlation times, exchange rates, and on binding site proton density. Indeed, only when short saturation times are used, *i.e.* less than 1 s, can intensities reflect ligand proton-protein proton distances (37). Here the saturation time of 2.4 s prevented us from quantitative analysis.

**Distance and Angle Constraints**—The cross-peak volumes from trNOESY and trROESY experiments of the pentasaccharide in the presence of mIgAs were measured with the VNMR software. Distances between neighboring protons were calculated by the usual  $1/r^6$  NOE/distance relationship (49). NOE-derived and trROE distances were obtained from initial NOE build-up rates, which were calculated by NOE volumes fitting during different mixing times. The intra-residue distance of 2.52 Å between the H-1 and H-2 protons of the  $\alpha$ -L-rhamnopyranosyl unit B was used as a reference for distance calibration.

Distance constraints of free peptides were obtained from the ROESY spectrum run at 298 K with a 400-ms mixing time. For peptides in the presence of mIgA, distance constraints were obtained from the trNOESY spectra run at 298 K with a 200-ms mixing time. NOE intensities were evaluated from the height of the cross-peaks. For structure calculations, upper limit distances of 2.8, 3.5, and 5 Å were used for strong, medium, and weak NOEs, respectively (50). The  $^3J_{\text{NH-H}\alpha}$  values were used to restrain  $\Phi$  angles as follows: for  $^3J > 9$  Hz,  $-155^\circ < \Phi < -85^\circ$ ; for 8 Hz  $< ^3J < 9$  Hz,  $-175^\circ < \Phi < -65^\circ$ ; for 5 Hz  $< ^3J < 7$  Hz,  $-105^\circ < \Phi < -55^\circ$ ; for  $^3J < 5$  Hz,  $-90^\circ < \Phi < -40^\circ$  (51).

**Structure Calculations**—Structure calculations of free and bound peptides were run on a Silicon Graphics work station using the standard protocol of the DYANA program (52). A total of 100 structures were calculated using the torsion angle dynamics protocol. The structures were sorted according to the final value of the target function, and the 20 best structures were analyzed in terms of distance and angle violations. Of these 20 structures, the 10 best structures were visualized by using MOLMOL (53).

**Homology Modeling of the IgA I3 Fab Fragment and Docking**—The search for structures with sequences similarities was performed with Blast (54) on sequences of all proteins with known three-dimensional structure in the Protein Data Bank (55). Five structures of interest were downloaded and used as template by the Composer program for the building of VL and VH chains of IgA I3 (56).

The Tripos force field (57) option of the Sybyl program (SYBYL) was used to minimize the energy of the resulting model whose stereochemical features were validated with the PROCHECK program (58).

The Autodock3 program (59) was used for docking oligosaccharides and peptides in the binding site of modeled IgA I3 Fab. Because the goal was to model the behavior of the O-SP, calculations were performed on the largest possible fragment compatible with the limitations of the software, in that case a nonasaccharide. The 9-carbohydrate residue fragment was thus chosen as BCDA(E)BCDA in which the key pentasaccharide DA(E)BC is flanked by two residues on each side. The two conformations that were shown previously to correspond to helical shapes of the O-SP (33) were used as starting models. Hydroxyl and *N*-acetyl bonds were considered as flexible, whereas glycosidic bonds were considered as rigid to keep the helical conformation, resulting in 28 degrees of freedom. AMBER force field charges were assigned to all protein atoms, and partial charges were assigned to the atoms according to the PIM force field (60). Grids of probe atom interaction energies and electrostatic potential were generated around the whole protein by the AutoGrid program present in Autodock3 with a spacing of 0.5 Å. All probes were placed arbitrarily at a distance of 10 Å from the protein surface, and their exocyclic torsion angles were allowed to rotate freely. For each monosaccharide, one job of 240 docking runs was performed using a population of 100 individuals and an energy evaluation number of  $10 \times 10^6$ . Clustering of solutions was done by root mean square fitting ( $< 1$  Å). The best solution of each cluster was used to propagate the helices to 20 residues while keeping the conformations determined previously (33). Twenty different conformers of the p100c peptide were also docked in the mIgA I3 Fab-binding site using the rigid body approach of the Autodock3 program. For each of them, one docking run was performed using a population of 100 individuals and an energy evaluation number of  $0.75 \times 10^6$ .

## RESULTS

**Antigenicity of the Ligands Used in the Study**—The binding of the synthetic nonapeptides p100c, p115, and p22 (28) and synthetic pentasaccharide DA(E)BC-OMe (32) to mIgA I3 and mIgA C5 was evalu-

## Peptide Mimics of *S. flexneri* 5a Polysaccharide Antigen

ated by inhibition ELISA to determine the concentration of ligands inhibiting 50% of mIgA binding to LPS ( $IC_{50}$  value). Because of the multivalency of both the mIgAs (dimeric mAb, thus four binding sites) and LPS, the  $IC_{50}$  value does not reflect the true binding affinity but allows relative comparison of the ligand recognition by the mIgAs. In agreement with the low affinity of mAb for carbohydrate antigens, an  $IC_{50}$  value close to 25  $\mu M$  was observed for the interaction of DA(E)BC-OME with each of the mIgAs.

$IC_{50}$  values for recognition of the mimotopes by mIgAs revealed that p100c was better recognized by mIgA I3 ( $IC_{50} = 75 \pm 29 \mu M$ ) than by

mIgA C5 ( $IC_{50} > 1000 \mu M$ ). In contrast, p115 was better recognized by mIgA C5 ( $IC_{50} = 197 \pm 39 \mu M$ ) than by mIgA I3 ( $IC_{50} > 1000 \mu M$ ). Most interestingly, p22 exhibited a higher  $IC_{50}$  value for both mIgAs ( $70 \pm 11 \mu M$  and  $0.03 \pm 0.01 \mu M$  for mIgA I3 and mIgA C5, respectively) than those measured for p100c and p115.

**NMR Parameters for the Free Peptide**—Peptide proton chemical shifts were assigned following standard procedures (50). The peptide conformations were probed through analysis of proton chemical shifts, three bond  $^3J_{NH-H\alpha}$  coupling constants, and proton-proton dipolar interactions observed in the ROESY spectra. Dihedral angles and distance constraints deduced from these data were then used to model the averaged solution structure of each peptide analyzed with the DYANA program (52).

**Peptide 115 (KVPPWARTA)**—The NMR spectrum revealed that p115 displayed four different conformers as a result of the *cis-trans* isomerization of the amide bonds involving the two sequential prolines, Val<sup>2</sup>–Pro<sup>3</sup> and Pro<sup>3</sup>–Pro<sup>4</sup>, respectively. Based on signal intensities, it was estimated that the major conformer represented 80% of the different species, whereas the three other forms altogether made up for the remaining 20%. Because no information was available for the conformer recognized upon selection from the phage displayed peptide library, structural analysis was conducted for this major conformer only. Chemical shifts and three bond  $^3J_{NH-H\alpha}$  coupling constants are reported in Table 1. Significant deviations from random coil values are only observed for the H- $\alpha$  protons of Pro<sup>3</sup> and Pro<sup>4</sup>, whereas all three bond  $^3J_{NH-H\alpha}$  coupling constants are those expected for flexible peptides. Inter-residue dipolar interactions observed in the ROESY experiment between the H- $\alpha$  of the residue preceding a proline and the H- $\delta$  proton of the proline indicates that both Pro<sup>3</sup> and Pro<sup>4</sup> adopt a *trans*-conformation in the peptide major conformer. In addition to standard sequential interactions, several medium range interactions were also observed between side chain protons of Val<sup>2</sup>, Pro<sup>3</sup>, and Pro<sup>4</sup> and the CH<sub>3</sub>- $\beta$  protons of Ala<sup>6</sup> (Fig. 1). These ROE connectivities were used as distance constraints to model the conformation of the p115 major conformer using the DYANA program. The 10 best structures, *i.e.* with the lowest

**TABLE 1**

**<sup>1</sup>H chemical shifts of the *trans-trans*-isomer of p115 in H<sub>2</sub>O/D<sub>2</sub>O (90/10), pH 5.1 and 298 K**

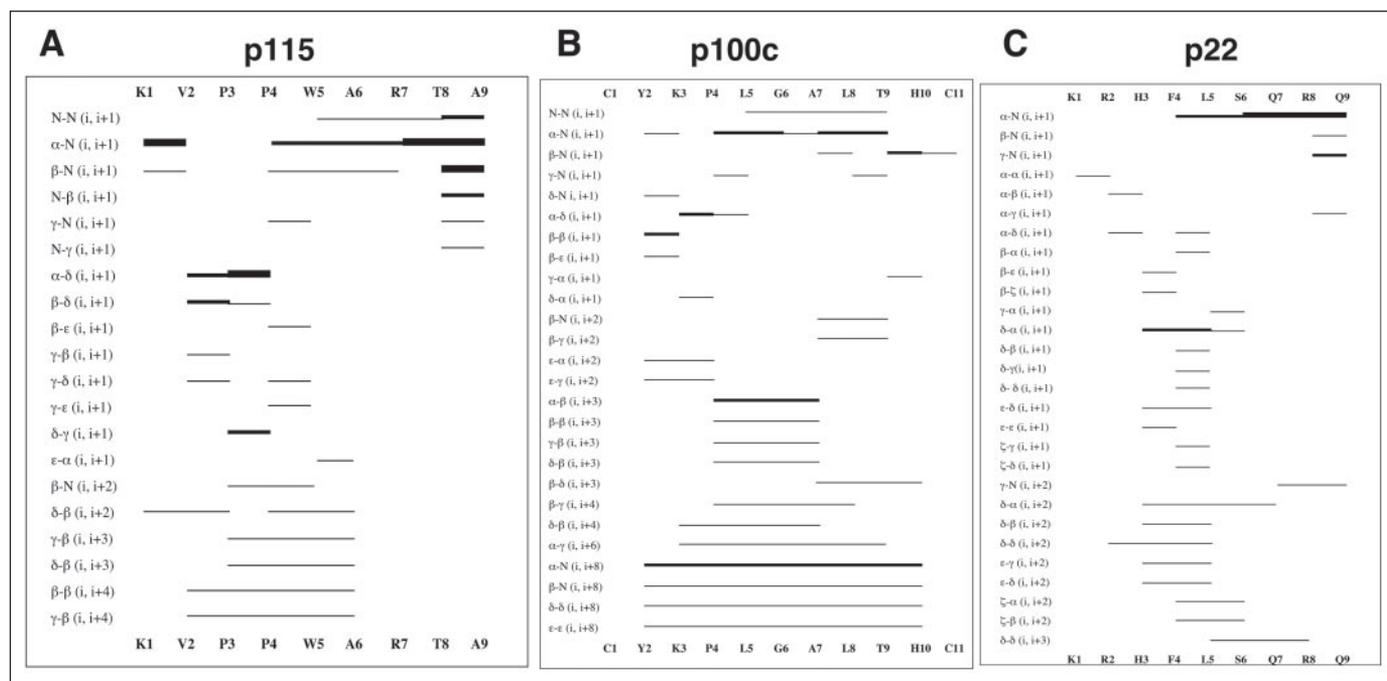
Chemical shifts measured in ppm with an accuracy of  $\pm 0.01$  ppm are referenced to external 4,4-dimethyl-4-silapentane sodium sulfonate ( $\delta_{H1}$  0.00).

| Residue          | H <sub>N</sub>           | H <sub><math>\alpha</math></sub> | H <sub><math>\beta</math></sub> | Others  |
|------------------|--------------------------|----------------------------------|---------------------------------|---|
| Lys <sup>1</sup> | ND <sup>a</sup>          | 4.02<br>–0.30 <sup>b</sup>       | 1.84                            | H- $\gamma$ 1.38<br>H- $\delta$ 1.66<br>H- $\epsilon$ 2.96<br>H- $\zeta$ ND     |
| Val <sup>2</sup> | 8.54 (ND)                | 4.41                             | 2.02                            | H- $\gamma$ 0.94–0.89   |
| Pro <sup>3</sup> |                          | 4.52<br>–0.21                    | 1.97–1.05                       | H- $\gamma$ 1.71–1.85<br>H $\delta$ 3.81–3.51                                   |
| Pro <sup>4</sup> |                          | 4.28<br>–0.14                    | 2.24–1.91                       | H- $\gamma$ 1.99<br>H- $\delta$ 3.69–3.44                                       |
| Trp <sup>5</sup> | 7.31 (6.80) <sup>c</sup> | 4.61<br>–0.05                    | 3.35–3.26                       | H- $\delta$ 17.24<br>H- $\zeta$ 37.16<br>H- $\epsilon$ 110.26                   |
| Ala <sup>6</sup> | 7.64 (6.60)              | 4.27<br>–0.05                    | 1.15                            |   |
| Arg <sup>7</sup> | 8.00 (6.70)              | 4.28<br>–0.06                    | 1.82–1.60                       | H- $\gamma$ 1.72<br>H- $\delta$ 3.17<br>H- $\eta$ 7.16–6.65<br>H- $\gamma$ 1.17 |
| Thr <sup>8</sup> | 8.17 (8.10)              | 4.30<br>–0.05                    | 4.23                            |   |
| Ala <sup>9</sup> | 7.97 (6.60)              | 4.11                             | 1.32                            |   |

<sup>a</sup> ND indicates not determined.

<sup>b</sup> Data in italics are the difference between the H $\alpha$  chemical shifts of the residues of the peptide and those of the same residues in nonstructured peptides GGXAGG or GGXPPG (90).

<sup>c</sup> Data in parentheses are  $^3J_{NH,H\alpha}$  coupling constants (in Hz  $\pm$  0.2 Hz) measured from the one-dimensional spectrum.



**FIGURE 1. <sup>1</sup>H-<sup>1</sup>H NOE connectivities observed for the major conformers of peptides. A, p115 *trans-trans*-isomer; B, p100c *trans*-isomer; C, p22. The intensity of the NOE cross-peak is indicated by the thickness of the lines (weak, –; medium, —; strong ■).**

energy function, showed that p115 adopts a rather organized conformation comprising residues Pro<sup>3</sup> to Arg<sup>7</sup>, whereas the N- and C-terminal ends are quite flexible (Fig. 2) as expected for peptides of this size. Based on C- $\alpha_i$ -C- $\alpha_{i+3}$  distances as well as on  $\Phi$  and  $\Psi$  angle values (Table 2), the conformation of the Pro<sup>3</sup>-Arg<sup>7</sup> fragment can be described as two sequential  $\beta$ -turns, a nonclassified one for PPWA and a type I  $\beta$ -turn for PWAR (61, 62).

**Peptide 100c (CYKPLGALTHC)**—Selected from a library displaying nonapeptides flanked by two cysteines (pVIII-9aa Cys) (28), synthetic p100c was chemically converted into its cyclic form. NMR data showed that in solution p100c existed as a 9:1 equilibrium between two conformers resulting from the *cis-trans*-isomerization of the amide bond between Lys<sup>3</sup> and Pro<sup>4</sup>. Again, despite the lack of information on the p100c-mIgA I3 recognition, only the major conformer was analyzed (Table 3). Significant deviation from standard chemical shift values was observed for the H- $\alpha$  protons of residues Tyr<sup>2</sup>, Lys<sup>3</sup>, Pro<sup>4</sup>, Leu<sup>5</sup>, and Gly<sup>6</sup> suggesting some restricted flexibility along the Tyr<sup>2</sup>-Gly<sup>6</sup> sequence. Furthermore, the three bond  $^3J_{\text{NH-H}\alpha}$  coupling constant values for residues Leu<sup>5</sup>, Gly<sup>6</sup>, and Ala<sup>7</sup> are slightly smaller than those measured for the other residues, 5 Hz versus 7–8 Hz (Table 3), strengthening the hypothesis of a probable structuring of the Lys<sup>3</sup>-Ala<sup>7</sup> segment. Inter-residues dipolar interactions observed in the ROESY experiment

between H- $\alpha$  of Lys<sup>3</sup> and H- $\delta$  of Pro<sup>4</sup> indicate that Pro<sup>4</sup> adopts a *trans*-conformation in the major conformer of p100c. In addition to standard sequential interactions, several medium range interactions were also observed between side chain protons of residues Pro<sup>4</sup> to Ala<sup>7</sup> as for example between all protons of Pro<sup>4</sup> and the methyl group of Ala<sup>7</sup> (Fig. 1). Moreover, four long range ROE connectivities were observed between Tyr<sup>2</sup> and His<sup>10</sup> protons, confirming the cyclic nature of the peptide (Fig. 1). ROE derived distances and coupling constants were used as constraints to generate a family of structures for p100c using DYANA. The 10 best structures indicate that the Pro<sup>4</sup>-Pro<sup>7</sup> fragment of p100c is conformationally organized into a type I  $\beta$ -turn (Fig. 2 and Table 2) (61, 62). Because both the cyclic form and Pro<sup>4</sup> can induce this type of conformational behavior, the structural analysis was extended to reduced p100c (data not shown). The type I  $\beta$ -turn remained, suggesting that Pro<sup>4</sup> alone is responsible for its formation, although the cyclic structure might contribute to its stabilization.

**Peptide 22 (KRHFLSQRQ)**—Available data (Table 4) suggest that p22 is very flexible. Indeed, except for the slight deviation observed for the Ser<sup>6</sup> H- $\alpha$  proton, chemical shifts do not significantly deviate from standard values. Meanwhile, none of the coupling constants of internal residues could be measured because of extensive signals overlaps. Nevertheless, medium range ROE connectivities were observed between side chains protons of residues His<sup>3</sup> and Leu<sup>5</sup> as well as between those of Phe<sup>4</sup> and Ser<sup>6</sup> (Fig. 1). The 10 structures of lowest energy matching those distance constraints show that fragment His<sup>3</sup> to Ser<sup>6</sup> of p22 is organized into a nonclassified  $\beta$ -turn (Table 2), although the peptide N- and C-terminal ends remain quite flexible (Fig. 2). That available chemical shift and coupling constant values do not reflect such an organized conformation in solution suggests a weaker stability of the  $\beta$ -turn.

**Ligand Interaction with the Protective mIgAs**—Investigation of the molecular pattern of the interactions involved in the peptide- and pentasaccharide-mIgA complexes relied on two complementary methodologies, namely trNOE and STD NMR experiments, whose combination was found to model accurately mAb-ligand interactions (38, 47). Indeed, the former technique provides key information on the conformation of the bound ligand, whereas the latter allows epitope mapping via magnetization transfer from the protein to the residues of the ligand that are in close contact with the protein. To carry out these experiments a few requirements have to be fulfilled as follows: (i) have an important contribution from the bound state to the NOEs, and (ii) have an exchange rate that is fast enough compared with the free ligand longitudinal relaxation. Because here the  $IC_{50}$  was the sole information available in terms of binding parameters for the complexes, for each system the best ligand:mIgA ratio was first evaluated from titration experiments according to the method of try and assay.

**Binding of DA(E)BC-OME to mIgA I3 and mIgA C5**—Bound pentasaccharide conformation. Independently of the mIgA tested, the best

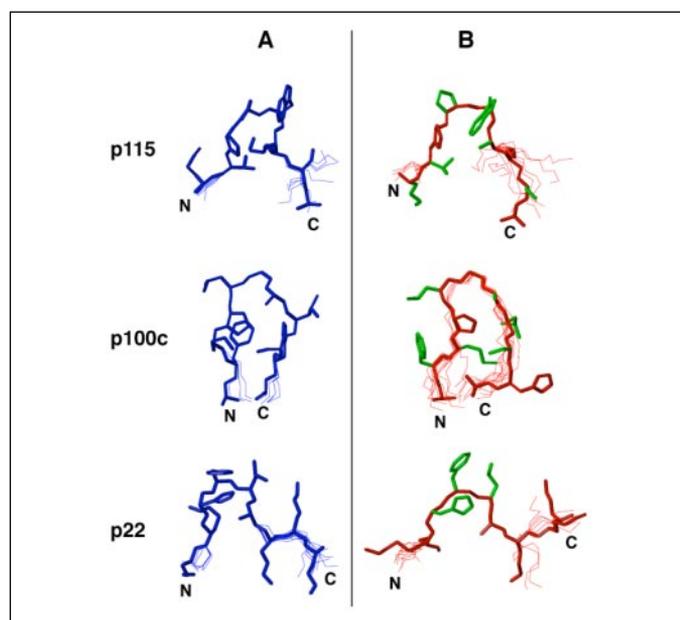


FIGURE 2. Structures of the major conformers of peptides in their free (A) and mIgA-bound (B) forms. p115 (on top), p100c (in the middle), and p22 (in the bottom) are shown. For each peptide, the backbone of the 10 best calculated structures are superimposed over the residues in  $\beta$ -turn conformation. B, the peptides side chains in contact with mIgAs, according to STD experiments, are shown in green.

**TABLE 2**  
Characterization of  $\beta$ -turn types of peptides p115, p100c, and p22 free and bound to protective mIgAs

| Peptides      | Turns <sup>a</sup> | Position $i + 1$ |                  | Position $i + 2$  |                  | $\beta$ -turn types                    |                                     |
|---------------|--------------------|------------------|------------------|-------------------|------------------|--|-------------------------------------|
|               |                    | $\phi$ (°)       | $\psi$ (°)       | $\phi$ (°)        | $\psi$ (°)       | Ramachandran nomenclature <sup>b</sup> | Classical nomenclature <sup>c</sup> |
| p115 free     | PPWA               | $-75.0 \pm 0.1$  | $-32.8 \pm 0.2$  | $-50.7 \pm 0.4$   | $-15.0 \pm 0.3$  | $\beta_p\alpha$                        | I                                   |
|               | PWAR               | $-50.7 \pm 0.4$  | $-15.0 \pm 0.3$  | $-46.6 \pm 0.1$   | $-20.4 \pm 0.2$  | $\alpha\alpha$                         | I                                   |
| p115 mIgA C5  | PPWA               | $-75.0 \pm 0.1$  | $140.0 \pm 0.4$  | $97.5 \pm 0.1$    | $20.7 \pm 0.1$   | $\beta_p\gamma$                        | II                                  |
| p100c free    | PLGA               | $-88.8 \pm 27.2$ | $-66.6 \pm 19.0$ | $-93.3 \pm 27.7$  | $6.1 \pm 12.1$   | $\alpha\alpha$                         | I                                   |
| p100c mIgA I3 | PLGA               | $-95.0 \pm 7.2$  | $64.2 \pm 23.0$  | $163.9 \pm 23.1$  | $-41.3 \pm 13.0$ | $\beta\gamma\alpha\alpha\gamma$        | $\approx$ II                        |
| p22 free      | HFLS               | $159.8 \pm 5.8$  | $-30.3 \pm 19.4$ | $-153.3 \pm 21.1$ | $-18.3 \pm 13.1$ | $\gamma\alpha$                         | I                                   |
| p22 mIgA I3   | HFLS               | $-121.6 \pm 2.3$ | $122.5 \pm 0.8$  | $55.7 \pm 0.1$    | $76.7 \pm 0.5$   | $\beta_E\gamma$                        | $\approx$ II                        |

<sup>a</sup> Residues in turn conformation.

<sup>b</sup>  $\beta$ -Turn types as defined by Wilmot and Thornton (91).

<sup>c</sup> Richardson classification system (92).

## Peptide Mimics of *S. flexneri* 5a Polysaccharide Antigen

**TABLE 3**

<sup>1</sup>H chemical shifts of the *trans*-isomer of p100c in H<sub>2</sub>O/D<sub>2</sub>O (90/10), pH 5.1 and 298 K

Chemical shifts measured in ppm with an accuracy of ±0.01 ppm are referenced to external 4,4-dimethyl-4-silapentane sodium sulfonate (δ<sub>H</sub> 0.00).

| Residue           | H <sub>N</sub>           | H <sub>α</sub>             | H <sub>β</sub> | Others                                       |
|-------------------|--------------------------|----------------------------|----------------|--|
| Cys <sup>1</sup>  | ND <sup>a</sup>          | ND                         | 3.29–3.22      |  |
| Tyr <sup>2</sup>  | 8.87 (7.20) <sup>b</sup> | 4.65<br>–0.10 <sup>c</sup> | 3.04–2.91      | H-δ 7.12<br>H-ε 6.80                         |
| Lys <sup>3</sup>  | 8.05 (8.00)              | 4.47<br>–0.13              | 1.55           | H-γ 1.23<br>H-δ 1.63<br>H-ε 2.90<br>H-ζ 7.45 |
| Pro <sup>4</sup>  |                          | 4.30<br>–0.12              | 2.23–1.84      | H-γ 1.88<br>H-δ 3.39                         |
| Leu <sup>5</sup>  | 8.22 (5.30)              | 4.23<br>–0.11              | 1.61           | H-γ 1.61<br>H-δ 0.92–0.88                    |
| Gly <sup>6</sup>  | 8.46 (5.40)              | 4.05–3.74<br>–0.09/–0.22   |                |  |
| Ala <sup>7</sup>  | 8.02 (5.40)              | 4.31<br>–0.01              | 1.38           |  |
| Leu <sup>8</sup>  | 8.32 (7.20)              | 4.37<br>0.03               | 1.69           | H-γ 1.62<br>H-δ 0.92–0.85                    |
| Thr <sup>9</sup>  | 7.69 (7.80)              | 4.30<br>–0.05              | 4.17           | H-γ 1.15                                     |
| His <sup>10</sup> | 8.30 (7.80)              | 4.38<br>–0.35              | 3.36           | H-ε1 8.58<br>H-δ2 7.31                       |
| Cys <sup>11</sup> | 8.49 (7.40)              | 4.51<br>–0.20              | 3.28–3.05      |  |

<sup>a</sup> ND indicates not determined.

<sup>b</sup> Data in parentheses are <sup>3</sup>J<sub>H<sub>N</sub>,H<sub>α</sub></sub> coupling constants (in Hz ± 0.2 Hz) measured from the one-dimensional spectrum.

<sup>c</sup> Data in italics are the difference between the H<sub>α</sub> chemical shifts of the residues of the peptide and those of the same residues in nonstructured peptides GGXAGG or GGXPGG (90).

DA(E)BC-OMe:mIgA ratio to observe trNOEs was shown to be 20:1 in binding sites. Because the highest attainable mIgA concentration was 3.75 μM, a 0.3 mM concentration of DA(E)BC-OMe was used to fulfill the 20:1 ratio requirement. Because NOE intensities depend, among other parameters, on the correlation time for reorientation and therefore on temperature, the later parameter was optimized so that ωτ<sub>c</sub> equals 1 for the free pentasaccharide, thus allowing us to distinguish NOE from trNOE connectivities. Indeed, the NOESY spectrum of the free pentasaccharide at 30 °C displayed only few positive and weak NOE connectivities characteristic of small molecules, whereas negative NOE connectivities were observed in the trNOESY spectrum of DA(E)BC-OMe when interacting with either mIgA C5 or mIgA I3. Because these effects did not result from the increased solution viscosity as probed by recording NOESY spectra of the pentasaccharide in the presence of BSA at the same w/v concentration (37), it was assumed that they corresponded to trNOE connectivities for the mIgA-bound DA(E)BC-OMe.

trNOESY spectra obtained with several mixing times, ranging from 100 to 400 ms, allowed us to trace the build-up curves (trNOEs intensities versus τ<sub>m</sub>) from which the distance information was extracted. In addition, inter-residue <sup>1</sup>H–<sup>1</sup>H distances were also calculated from a trROESY spectrum obtained with a mixing time of 400 ms to take spin diffusion into account, if any. Comparison of these distances with those measured for unbound DA(E)BC-OMe (33) suggested that the pentasaccharide conformation was not significantly modified upon binding to the mIgAs (Table 5).

**Epitope Characterization**—The key elements involved in DA(E)BC-OMe binding to the mIgAs were then characterized based on STD experiments. As for the trNOE experiments, an oligosaccharide to mAb ratio of 20:1 in binding site was used. To decrease the exchange rate, the temperature was set at 15 °C. The one-dimensional STD spectrum of DA(E)BC-OMe interacting with mIgA C5 shows that protons H1 and H2 and H6 (methyl group) of rhamnosides A and B, as well as all protons belonging to glucose (E), are in close contact with the mIgA C5-binding

**TABLE 4**

<sup>1</sup>H chemical shifts of the *trans*-isomer of p22 in H<sub>2</sub>O/D<sub>2</sub>O (90/10), pH 5.1 and 298 K

Chemical shifts measured in ppm with an accuracy of ±0.01 ppm are referenced to external 4,4-dimethyl-4-silapentane sodium sulfonate (δ<sub>H</sub> 0.00).

| Residue          | H <sub>N</sub>   | H <sub>α</sub> | H <sub>β</sub> | Others                                       |
|------------------|------------------|----------------|----------------|--|
| Lys <sup>1</sup> | ND <sup>a</sup>  | 3.88<br>–0.41  | 1.80           | H-γ 1.36<br>H-δ 1.67<br>H-ε 2.96<br>H-ζ ND   |
| Arg <sup>2</sup> | ND               | 4.28<br>–0.06  | 1.68           | H-γ 1.52<br>H-δ 3.14<br>H-η ND               |
| His <sup>3</sup> | ND               | 4.76<br>–0.03  | 3.18–3.00      | H-ε1 7.86<br>H-δ2 6.98                       |
| Phe <sup>4</sup> | 8.22 (7.20)      | 4.59<br>–0.03  | 3.09–2.97      | H-δ 7.20<br>H-ε 7.29<br>H-ζ 7.33<br>H-γ 1.50 |
| Leu <sup>5</sup> | 8.33 (ND)        | 4.32<br>–0.02  | 1.57           | H-δ 0.89–0.84                                |
| Ser <sup>6</sup> | 8.22 (ND) (6.70) | 4.38<br>–0.09  | 3.85           |  |
| Gln <sup>7</sup> | 8.35 (ND)        | 4.35<br>0.01   | 2.11–1.96      | H-γ 2.34<br>H-δ 6.82–7.53                    |
| Arg <sup>8</sup> | 8.33 (ND)        | 4.32<br>–0.02  | 1.86–1.74      | H-γ 1.62<br>H-δ 3.18<br>H-η ND               |
| Gln <sup>9</sup> | 8.04 (7.70)      | 4.15<br>–0.19  | 2.09–1.90      | H-γ 2.28<br>H-δ 6.78–7.50                    |

<sup>a</sup> ND indicates not determined.

<sup>b</sup> Data in italics are the difference between the H<sub>α</sub> chemical shifts of the residues of the peptide and those of the same residues in nonstructured peptides GGXAGG or GGXPGG (90).

<sup>c</sup> Data in parentheses are <sup>3</sup>J<sub>H<sub>N</sub>,H<sub>α</sub></sub> coupling constants (in Hz ± 0.2 Hz) measured from the one-dimensional spectrum.

site (Fig. 3). Indeed, these interacting protons were fully identified in the corresponding two-dimensional STD-TOCSY spectrum (Fig. 4). Similar results were obtained for DA(E)BC-OMe binding to the mIgA I3 (data not shown). Furthermore, protons from the glucose (E) and the methyl group of residue B give the strongest signal enhancements, suggesting that they are in closest contact with both mIgAs (Fig. 5) and play a crucial role in the oligosaccharide-mAb interaction.

**Binding of the Peptide Mimics to mIgAs**—Based on available IC<sub>50</sub> values, analysis was run on the p115-mIgA C5, p100c-mIgA I3, p22-mIgA C5, and p22-mIgA I3 complexes.

**Interaction of Peptide 115 (KVPPWARTA) with mIgA C5**—trNOE experiments recorded for the p115-mIgA C5 complex (see supplemental Fig. 4S) showed new NOE connectivities when compared with those observed for the free peptide. These additional cross-peaks, such as those observed between residues Trp<sup>5</sup> and Ala<sup>6</sup>, were clearly identified as representative of the p115-bound form. Interestingly, most of the NOE connectivities involving amide protons of the free p115 were no longer observed in bound p115 with the exception of the Trp<sup>5</sup>, Ala<sup>6</sup> amide proton connectivity (see supplemental Fig. 1S). The pH increase from 5 in the free peptide to 6.5 in the peptide:mIgA solution might account for such experimental observations since amide protons exchange faster at higher pH. Nevertheless, the medium range NOE connectivities observed between the side chain proton of residues Val<sup>2</sup> and Pro<sup>3</sup>, and the CH<sub>3</sub>-β of Ala<sup>6</sup> remained (see supplemental Fig. 1S). Distance constraints derived from trNOE intensities were used to establish the conformation of p115 when bound to mIgA C5. Superimposition of the 10 lowest energy backbone conformations of free p115 to those of mAb-bound p115 showed that only the turn involving residues Pro<sup>3</sup> to Ala<sup>6</sup>, observed in the free form, is maintained in the bound form. Based on C-α<sub>i</sub>–C-α<sub>i+3</sub> distances as well as on Φ and Ψ angle values (Table 2) the type I β-turn observed between residues Pro<sup>4</sup> and Arg<sup>7</sup> for the free peptide is no longer present. Whereas in the free form fragment

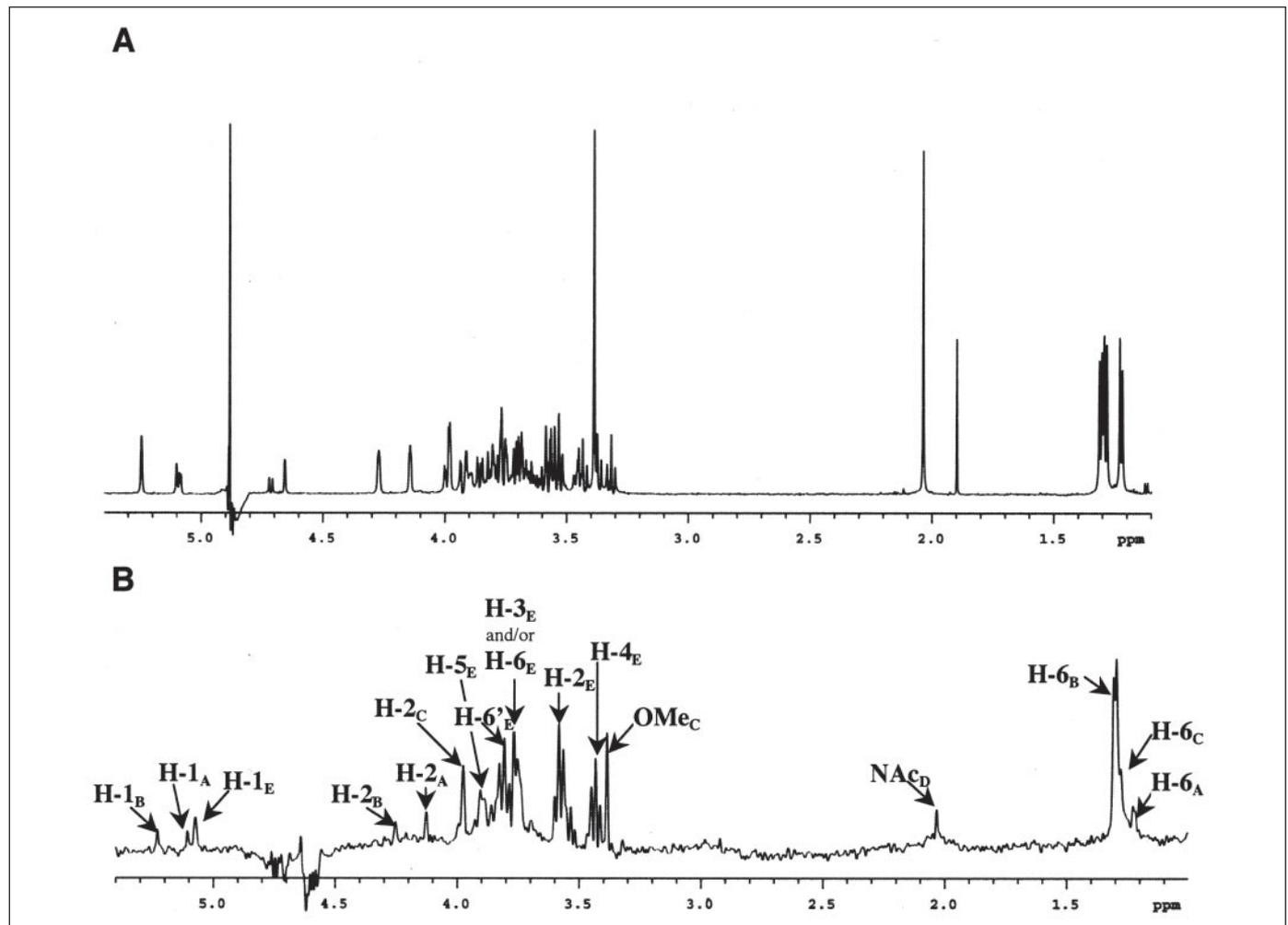
**TABLE 5**

<sup>1</sup>H-<sup>1</sup>H inter-residue distances (Å) extracted from dipolar interactions observed in ROESY and NOESY spectra of DA(E)BC-OMe pentasaccharide free and in interaction with mIgA C5 and mIgA I3

The two values correspond to distances extracted from ROESY (400 ms) (left) and NOESY (right), respectively (accuracy, ±10%). ND indicates not determined.

| Proton pairs <sup>a</sup> | DA(E)BC-OMe | DA(E)BC-OMe/IgA C5 | DA(E)BC-OMe/IgA I3 |
|---------------------------|-------------|--------------------|--------------------|
| A-1/B-1                   | 3.3/ND      | ND/3.0             | ND/2.9             |
| A-1/B-2                   | 2.2/ND      | ND/2.4             | 2.1/2.2            |
| A-5/B-1                   | 2.4/2.5     | ND/2.7             | 2.4/2.5            |
| A-6/B-1                   | 3.4/3.4     | ND/3.2             | ND/3.2             |
| B-1/C-3                   | 2.3/2.3     | ND/2.3             | 2.3/2.2            |
| B-2/E-1                   | 2.3/ND      | ND/2.5             | 2.2/2.3            |
| B-3/E-1                   | 2.5/2.5     | ND/3.2             | 2.8/2.8            |
| B-3/E-5                   | 3.3/3.2     | ND/2.7             | ND/2.7             |
| B-6/C-2                   | 3.6/3.5     | ND                 | ND/3.5             |

<sup>a</sup> A-1 corresponds to proton 1 of rhamnose A.



**FIGURE 3. mAb binding epitope of the pentasaccharide DA(E)BC-OMe.** *A*, one-dimensional <sup>1</sup>H reference spectrum of pentasaccharide DA(E)BC-OMe in the presence of mIgA C5 (20:1 ratio in site). *B*, one-dimensional STD-NMR of DA(E)BC-OMe in the presence of mIgA C5 with selective saturation of antibody resonances at 0.3 ppm. Protons of DA(E)BC-OMe affected by the selective saturation of mIgA C5 and so in contact with the mAb are labeled.

Pro<sup>3</sup>-Ala<sup>6</sup> adopts a nonclassified type of β-turn, in the bound form a well defined type II β-turn is clearly present (61, 62) (Fig. 2 and Table 2).

STD experiments (Fig. 6), run under the experimental conditions used for trNOE experiments, showed that the p115 protons in close contact with mIgA C5 are the side chain protons of Lys<sup>1</sup> and Pro<sup>4</sup>, all the methyl groups thus implicating Val<sup>2</sup>, Ala<sup>6</sup>, and Ala<sup>9</sup>, and all protons of the Trp<sup>5</sup> aromatic ring. Clearly, in addition to side interactions involving the N- and C-terminal ends, major contacts involve residues from the Pro<sup>4</sup>-Ala<sup>6</sup> segment, suggesting that the structured Pro<sup>3</sup>-Ala<sup>6</sup> type II

β-turn is crucial for p115:mIgA C5 recognition. It is worth noting that the major form of p115 was that recognized by mIgA C5.

*Interaction of Peptide p100c (CYKPLGALTHC) with mIgA I3*—As compared with data corresponding to the free form, the trNOESY spectrum of p100c in interaction with mIgA I3 displayed new data specific for the bound form. These include new sequential NOEs such as those observed between Lys<sup>3</sup> and Pro<sup>4</sup> or between Pro<sup>4</sup> and Leu<sup>5</sup>, whereas sequential NOEs between Ala<sup>7</sup> and Leu<sup>8</sup>, or between Leu<sup>8</sup> and Thr<sup>9</sup>, were no longer visible. However, medium range NOE connectivities

FIGURE 4. The branched glucose residue E of the pentasaccharide DA(E)BC-OMe constitutes a key element in mAb recognition. *A*, two-dimensional STD-TOCSY of DA(E)BC-OMe in the presence of mIgA C5 (20:1 ratio in site) with selective saturation of antibody resonances at 0.3 ppm. *B*, a zooming of the two-dimensional STD-TOCSY in the 3.4–4 ppm region emphasizing the glucose E proton connectivities.

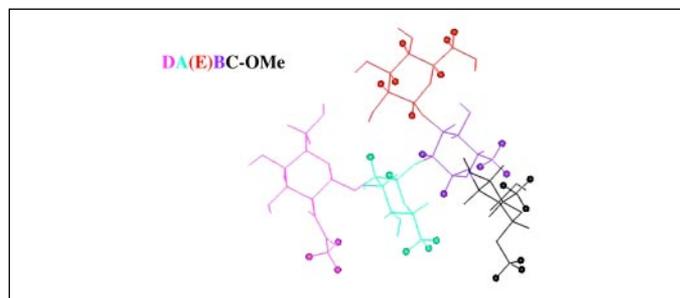
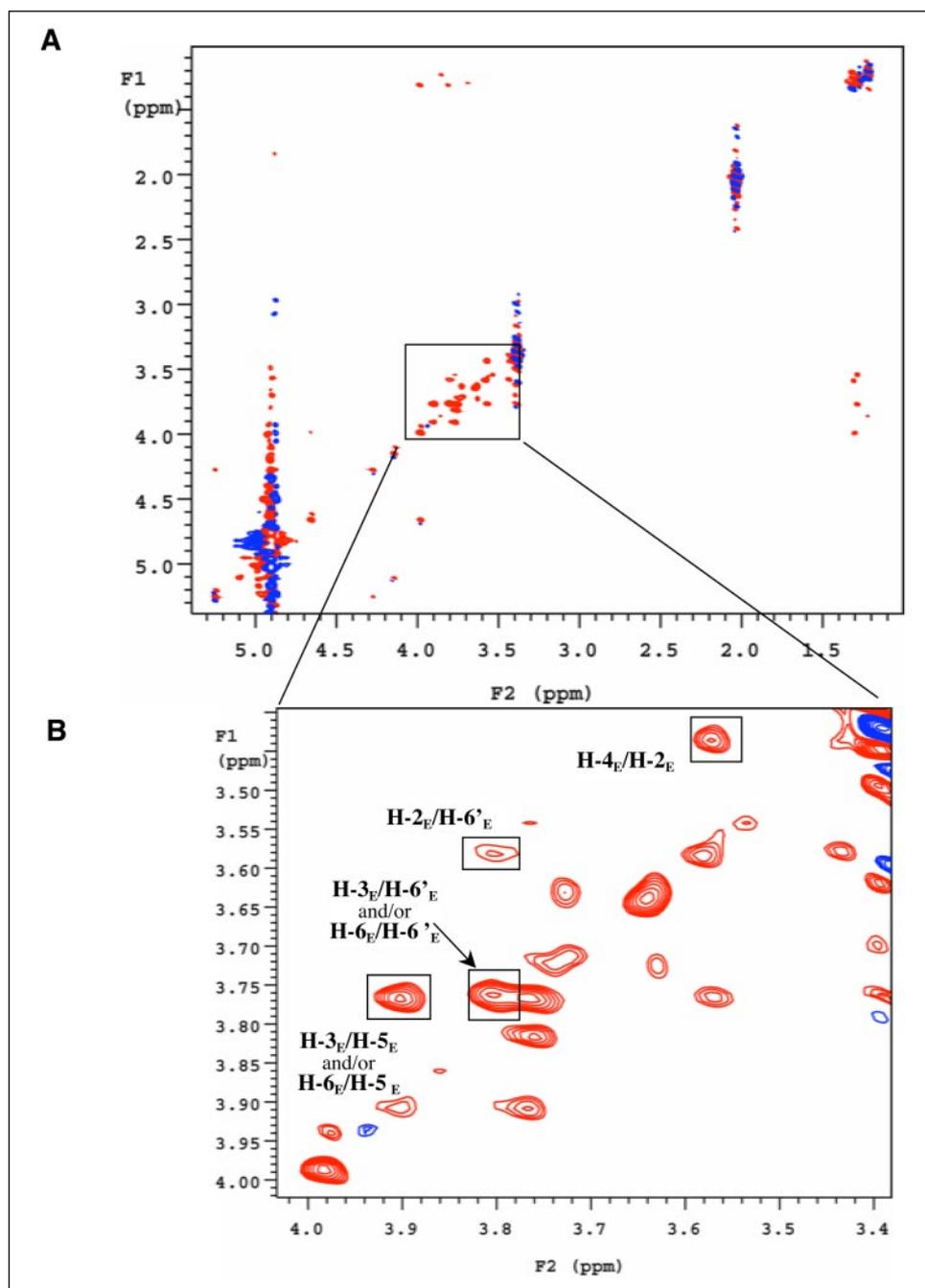


FIGURE 5. Structure of the pentasaccharide DA(E)BC-OMe. The representation of a lowest energy conformation of DA(E)BC-OMe as determined with the CICADA method and in agreement with NMR data (33). The *small spheres* indicate the protons that are in contact with the protective mAbs, mIgA C5, and mIgA I3, in agreement with the STD-NMR experiments.

observed between Pro<sup>4</sup> and Ala<sup>7</sup> in the free form remained. In addition, medium range interactions specific to the bound form were observed, such as those involving side chain protons of Tyr<sup>2</sup> and Leu<sup>5</sup>, Lys<sup>3</sup> and Leu<sup>8</sup>, as well as Tyr<sup>2</sup> and His<sup>10</sup> (see supplemental Fig. 2S). Superimposition of the backbone (Pro<sup>4</sup> to Ala<sup>7</sup> fragment) of the 10 lowest energy conformations of free and mIgA I3-bound p100c matching the distance constraints showed that, as observed for p115, the turn observed in the free form was maintained in the bound form. Furthermore, data pointed to a switch from a type I  $\beta$ -turn in the free form to a type II  $\beta$ -turn in the bound form (Table 2) (61, 62). Additional significant rearrangements were observed for the rest of the backbone (Fig. 2).

More detailed epitope identification was derived from the one-dimensional STD experiment. All methyl group protons of p100c, thus involving residues Leu<sup>5</sup>, Ala<sup>7</sup>, and Leu<sup>8</sup> as well as the Tyr<sup>2</sup> side chain

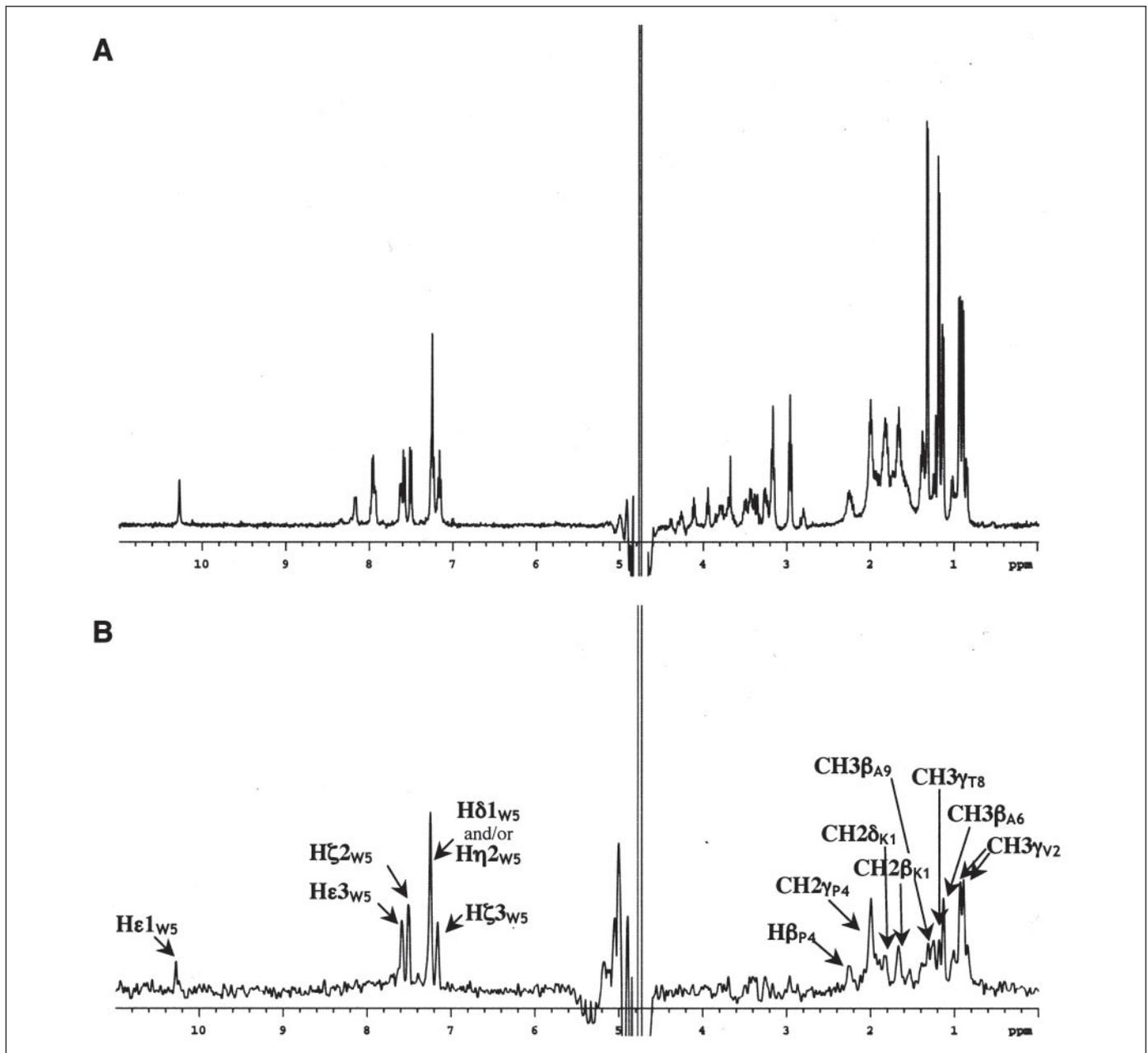


FIGURE 6. **mAb binding epitope of the peptide p115.** *A*, one-dimensional reference spectrum of p115 in the presence of mIgA C5 (20:1 ratio in binding site). *B*, one-dimensional STD-NMR of p115 in presence of mIgA C5 with selective saturation of mAb at 0.3 ppm. Protons of p115 affected by the selective saturation and so in interaction with mIgA C5 are labeled.

protons and the Lys<sup>3</sup> CH<sub>2</sub>-γ protons, were in contact with the mIgA I3 (Fig. 7). Within the turn, only Leu<sup>5</sup> and Ala<sup>7</sup> methyl groups contacted the mAb. Similarly to p115, the methyl groups of the hydrophobic residues and the aromatic residue of p100c were involved in the interaction with the mIgA I3.

**Interaction of Peptide p22 (KRHFLSQRQ) with mIgA I3 and mIgA C5**—Comparison with the spectrum of the free peptide shows that the trNOE connectivities observed when p22 is bound to mIgA I3 differ for medium range interactions. Indeed, dipolar interactions observed between Phe<sup>4</sup> and Ser<sup>6</sup> in free p22 disappeared to the benefit of new interactions between His<sup>3</sup> and Ser<sup>6</sup> in the bound form. However, dipolar interactions between His<sup>3</sup> and Leu<sup>5</sup> were observed both in the free and the bound forms (see supplemental Fig. 3S). Comparison of the lowest energy conformations adopted by p22 in the free and bound forms showed that the nonclassified β-turn

involving the His<sup>3</sup>–Ser<sup>6</sup> segment in the free form changed to a type II β-turn in the bound form, as observed for p115 and p100c. Additional rearrangements were observed for the rest of the backbone (Fig. 2). trNOE experiments for p22 bound to the mIgA C5 were unsuccessful because sparse negative NOEs were observed (data not shown). The high affinity of mIgA C5 ( $IC_{50}$  of 0.03 μM), most probably associated with an equilibrium constant for dissociation ( $K_d$ ) below  $10^{-6}$  M and thus not compatible with trNOE observations ( $10^{-3} > K_d > 10^{-6}$  M), might be responsible for this effect. Despite this drawback, epitope mapping by STD experiments was successfully undertaken as the lower limit for exchange was less stringent in terms of  $K_d$  ( $10^{-8}$  M). Epitope identification for p22 bound to mIgA I3 and mIgA C5, respectively, was obtained from the one-dimensional STD experiments (Fig. 8). Whether mIgA I3 or IgA C5 was concerned, p22 residues in direct contact with the mAbs were identical. They included the His<sup>3</sup> imid-

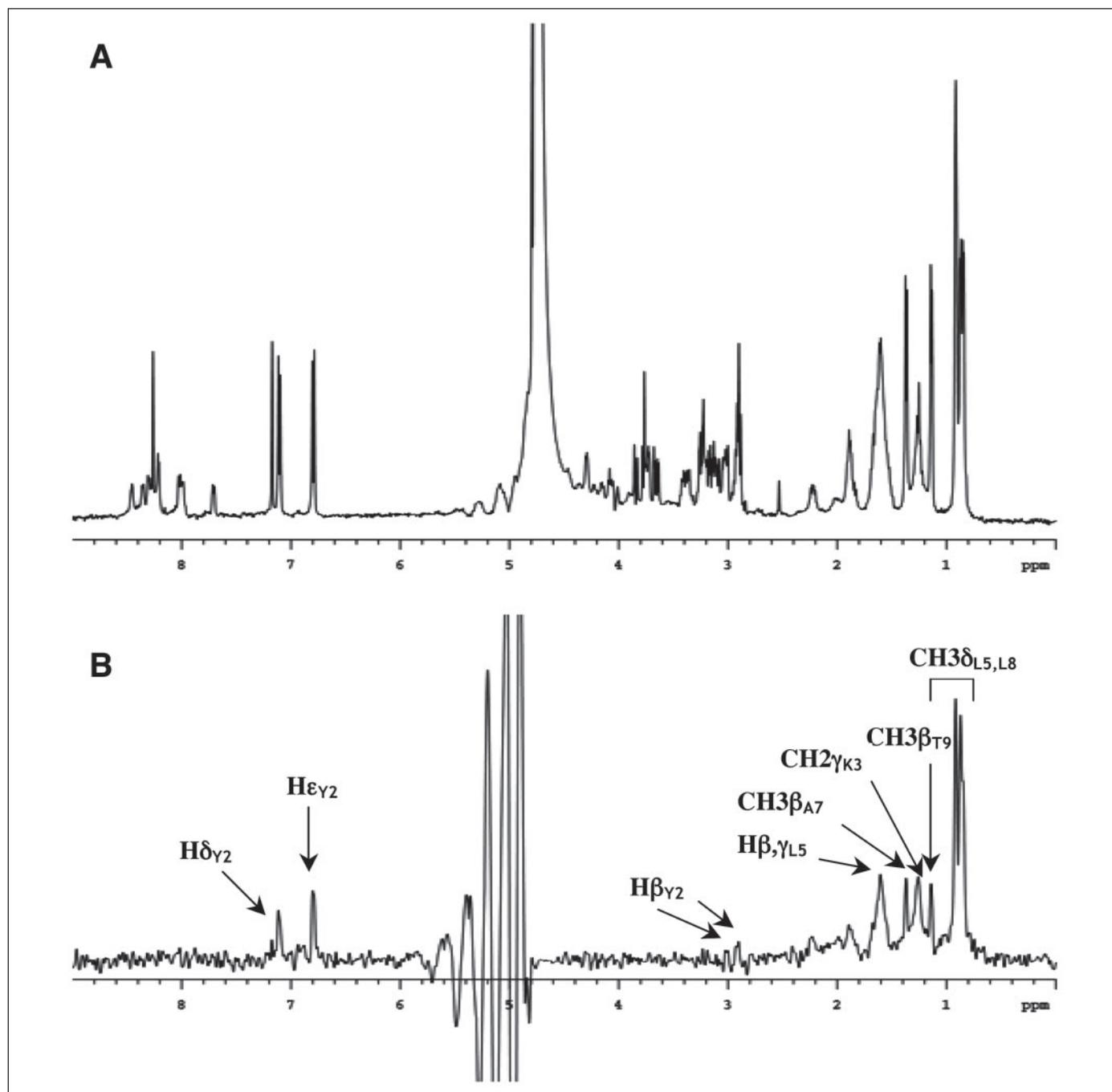


FIGURE 7. **mAb binding epitope of the peptide p100c.** *A*, one-dimensional reference spectrum of p100c in presence of mIgA I3 (20:1 ratio in binding site). *B*, one-dimensional STD-NMR of p100c in presence of mIgA I3 with selective saturation of mAb resonances at 0.3 ppm. Protons of p100c affected by the selective saturation and so in interaction with mIgA I3 are labeled.

azole protons, Phe<sup>4</sup> aromatic protons, as well as Leu<sup>5</sup> methyl protons, all corresponding to amino acids taking part in the turn observed in free p22. Although the conformation of mIgA C5-bound p22 remained undisclosed, available data suggest that the nonclassified turn naturally adopted by p22 contributed to both mIgA I3 and mIgA C5 recognition.

**Modeling of the Fab Domain of mIgA I3**—A blast search in the Protein Data Bank (55) allowed us to identify three mAb chains with high sequence similarity to either the VL or the VH chain of mIgA I3. Sequence alignments are displayed in Fig. 9 together with Protein Data Bank code for the structures of interest that include anti-RNA mAb (code 1MRD) (63), the anti-DNA mAb (code 1CBV) (64), the catalytic mAb (code 1A4J) (65), anti-influenza neuraminidase (code 1NCA) (66), and anti-cholera toxin mAb

(code 1TET) (67). Each chain was built by homology modeling using the standard procedure of the composer program (56). The murine Fab fragment with Diels-Alder catalytic properties (65) displayed high similarity for both chains and was used as a template for assembling the two chains. As a general feature for mAbs, the H3 loop of CRD is known as the most variable one. Among the structures with high homology, the H3 loop of the anti-cholera toxin Fab (67) was selected as a template because it displays the same number of amino acids as the target. After optimization of the side chain conformation, the binding site of mIgA I3 appeared to have a distinct “groove” character located between the variable loops with a deep central pocket. The sides of the groove were flanked by the CDRs, mostly H2 and H3 of the heavy chain and L1 in the light chain.

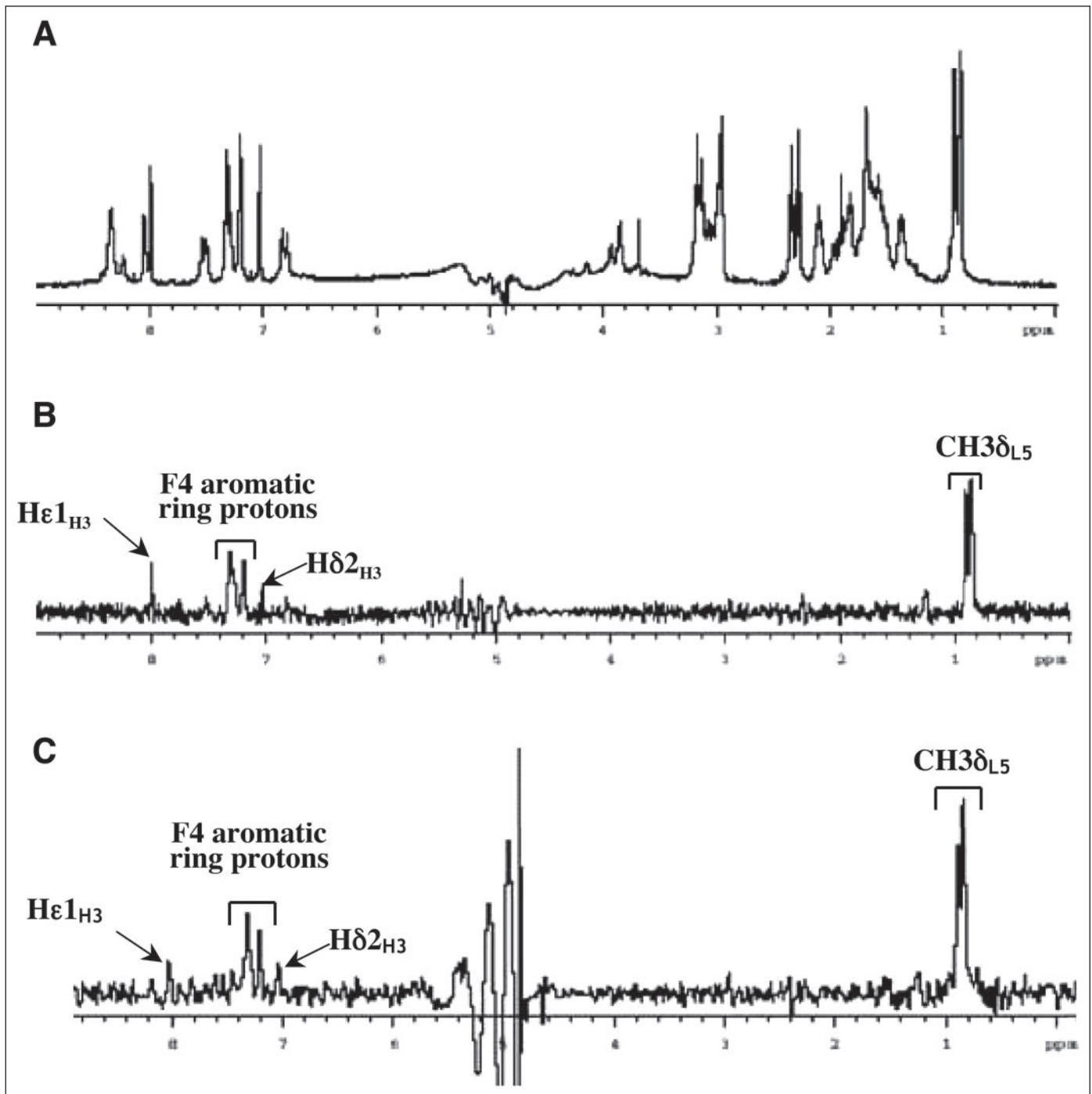


FIGURE 8. **mAb binding epitope of the peptide p22.** *A*, one-dimensional reference spectrum of p22 in presence of mIgA I5 (20:1 ratio in binding site). *B*, one-dimensional STD-NMR of p22 in presence of mIgA I3. *C*, one-dimensional STD-NMR of p22 in presence of mIgA C5. The selective saturation of mAb resonances was done at 0.3 ppm. Protons of p22 affected by the selective saturation and so in contact with the mAbs are labeled.

*Docking of Oligosaccharides and Modeling of Complexes with mIgA I3*—Previous studies identified two different helical conformations of *S. flexneri* 5a O-SP as being the most stable ones in solution (33). Both can be described as right-handed 3-fold helices, but one is more extended (E) than the other (O) with helical repeats of 23.2 and 19.4 Å, respectively. The nonasaccharide BCDA(E)BCDA was selected for docking studies as the largest O-SP fragment that can be treated as a flexible ligand. When the nonasaccharide BCDA(E)BCDA fixed in both conformations is docked into the Fab-binding site, four solutions can be identified, two for the E conformation and two for the O one. In both cases, the nonasaccharide can

fit either with a parallel orientation to the binding site groove or with a perpendicular one. In any case, the branched  $\alpha$ -D-glucopyranose residue E is deeply buried into the central pocket of the groove. In all of the four solutions, the mAb features involved in carbohydrate recognition are the three loops from the heavy chain and the L1 and L3 loops from the light chain. The variable loop H3 plays the major role, with its two Asp residues (Asp<sup>91</sup> and Asp<sup>92</sup>) involved in recognition for most binding modes.

When the nonasaccharide-bound conformations were propagated into polysaccharide structures comprising four repeating units, only two docking modes appeared to be stable with additional contacts cre-

## Peptide Mimics of *S. flexneri* 5a Polysaccharide Antigen

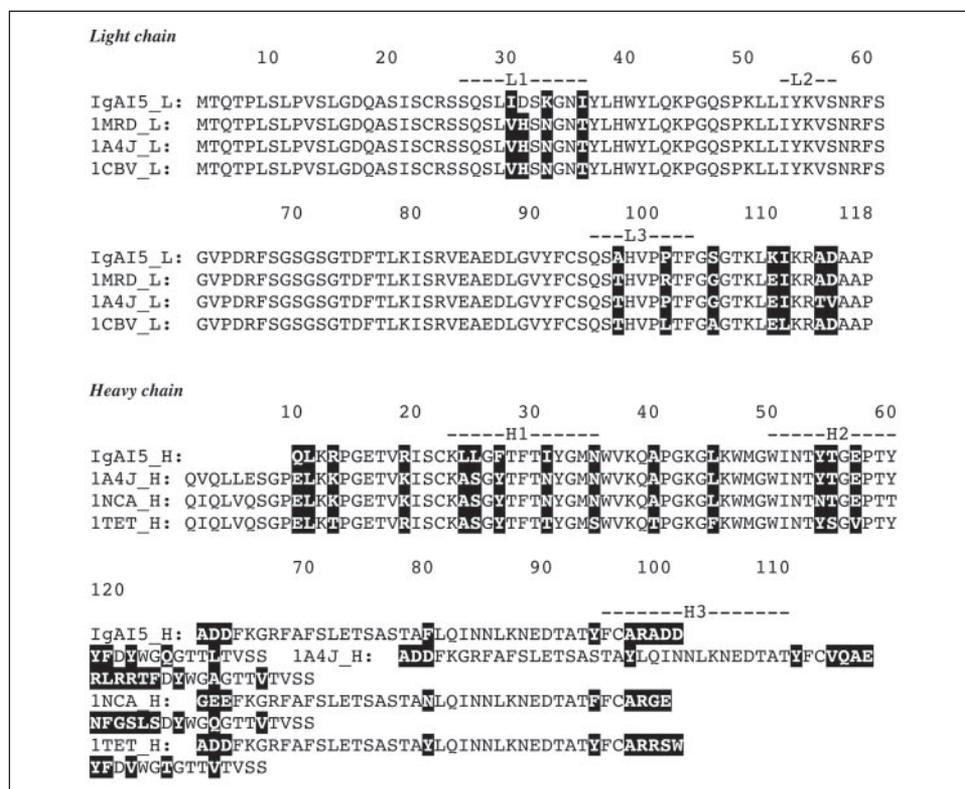


FIGURE 9. Alignment of mIgA I3 sequence with related sequences taken from the Protein Data Bank. Sequence differences are highlighted by displaying amino acid code using white letter on black background.

ated on both sides of the binding site. In each case, this ability corresponded to the parallel arrangement of BCDA(E)BCDA. Therefore, only these two solutions, *i.e.* docking of E and O conformations in parallel mode, were considered as possible mimics of O-SP binding to mIgA I3. The two possible docking modes of BCDA(E)BCDA and the polysaccharide of DP4 are displayed in Fig. 10 (A–D) and the contacts of interest are listed in Table 6. For both conformations of each ligand, the central trisaccharide A(E)B makes most of the binding contribution, and the additional contacts established by the GlcNAc (D) residue are minor.

**Docking of Peptide Mimics**—Docking of peptides to mAbs was performed on one example in order to rationalize the protective effect. The cyclic peptide p100c is conformationally constrained and was therefore selected for the modeling studies with the mAb that displays the highest affinity, *i.e.* mIgA I3. For the two lowest energy docking solutions, the p100c conformation displayed good shape complementarity with the binding site central pocket of mIgA I3. In the first solution, a strong interaction (three H-bonds and two salt bridges) appeared between the peptide and the mAb, mainly located on CDRH3 domain involving Asp<sup>91</sup> and Asp<sup>92</sup> (Table 6). The second docking solution led to identical main interactions between p100c and CDRH3 (Table 6). In both cases, the Leu<sup>5</sup>–Gly<sup>6</sup>–Ala<sup>7</sup> motif of p100c, seen as constrained by NMR, was driven into the central pocket (van der Waals interactions). Both docking modes allowed for a strong interaction between the Lys<sup>3</sup> of the peptide and the protein Asp<sup>92</sup>. Nevertheless, the Tyr<sup>2</sup> residue of the peptide was buried in the second solution and established strong van der Waals contacts with the aromatic side chain of Trp<sup>41</sup> and Phe<sup>94</sup> of mIgA I3. Therefore, this model displayed in Fig. 10 (E and F) was that in best agreement with NMR data.

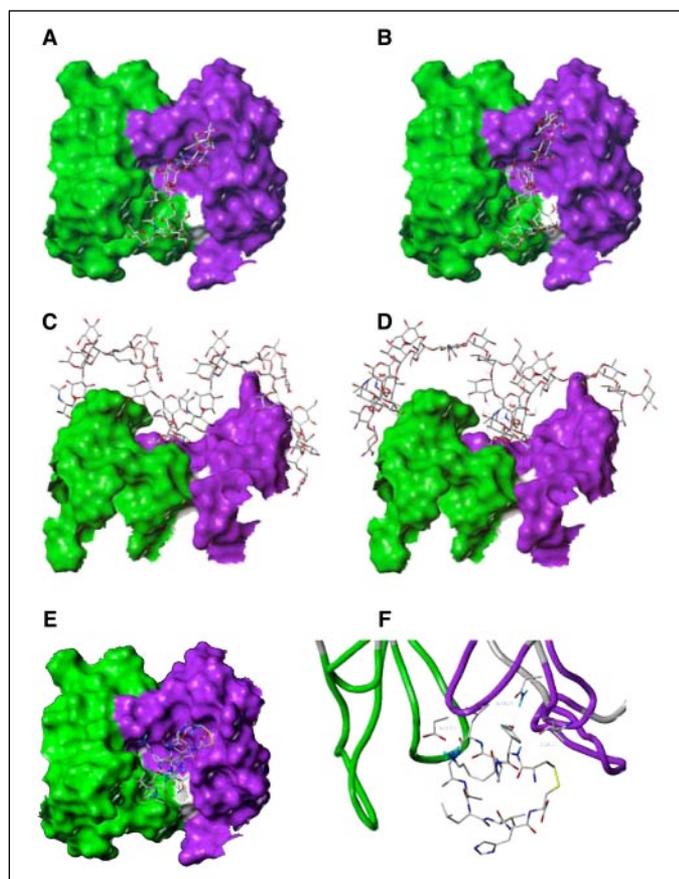
## DISCUSSION

As a novel strategy to improve vaccine design, molecular mimicry has gained a growing interest in the recent past. For mimicry of polysaccharides, the mimics can be of the same molecular class as the natural

antigen, *i.e.* oligosaccharides, or different, as for instance the peptide mimics. The nature of peptide carbohydrate mimicry has not yet been deciphered, and detailed structural studies of both oligosaccharide-mAb complexes and carbohydrate-mimicking peptides-mAb complexes are still needed to expand the thus far limited structural data base available in the series. Peptide-carbohydrate mimicry is either structural, functional, or both. Structural mimicry resides in mimicry of specific chemical groups of the carbohydrate by chemical groups of the peptide, thus both ligands contact the same residues in the mAb-binding site (68). Mimicry is termed functional when the mimic differs structurally from the natural antigen. Both the antigen and the mimic cross-react specifically with the mAb used for selection, although the protein residues involved in recognition differ (17, 21, 69).

Here, by aiming at designing new vaccine strategies against *Shigella* infection, we developed synthetic mimics, carbohydrates and peptides, of *S. flexneri* 5a O-SP. Previous NMR and molecular modeling studies from our laboratory have shown that, among the four possible frame-shifted pentasaccharides representative of the O-SP, DA(E)BC-OME best mimics the conformational features of *S. flexneri* 5a O-SP (33). More importantly, the trNOE data reported here indicate that the conformation of DA(E)BC-OME when bound to O-SP-specific mIgA I3 and mIg C5 does not differ from its conformation when free in solution. Noteworthy, selection of free solution conformers often predominates in mAb-carbohydrate recognition processes (70). However, it is not always so as exemplified with mAb Se155-4 binding to a trisaccharide antigenic determinant of the *Salmonella paratyphi* B branched O-SP (71).

Another interesting example of such induced conformational change is the mAb SYA/J6 binding to the pentasaccharide ABCDA' fragment of the linear O-SP defining *S. flexneri* serotype Y (72). Modeling of the linear heteropolysaccharide has shown that it is structured into a left-handed helical chain of three ABCD repeating units (73, 74). Most interestingly, extension of the modeling study to *S. flexneri* 5a O-SP



**FIGURE 10. Graphical representation of the different models of antibody mIgA I3 (light chain in green and heavy chain in violet) with docked oligosaccharide and cyclic peptide p100c.** A and B, two possible docking modes for nonasaccharide in mIgA I3. C and D, corresponding interaction with the polysaccharide (four repeating units) after propagation of the nonasaccharide conformation along the 20-residue chain. E and F, two different views of the docking mode of p100c in mIgA I3-binding site displaying the best agreement with NMR data.

suggested that residue E, which is associated with serotype specificity, crucially impacts the overall O-SP conformation. In fact, the branched heteropolysaccharide, whose repeating unit (I) bears the E side chains, behaves as a right-handed 3-fold helix with residue E protruding outwardly (33). Along this line, the interaction of DA(E)BC-OMe with the serotype-specific mIgAs is mainly driven by the branched E residue as evidenced by the large number of E-specific signals enhanced in the STD spectra of the pentasaccharide in complex with mIgA I3. Furthermore, NMR experiments showed that all rhamnose methyl groups are also in close contact with the mAb, with the methyl group of rhamnose B on which E is branched being the major contributor. The *N*-acetyl group of residue D gives only weak contacts with the mIgA protons, indicating that it probably lies at the surface of the mAbs. These data are supported by the inhibition ELISA results showing that all frame-shifted tri-, tetra-, and pentasaccharides, bearing A(E)B-branched trisaccharide characteristic of *S. flexneri* 5a serotype, are recognized by a protective serotype 5a-specific mIgG (33). Further insights on the central role played by the branched  $\alpha$ -D-glucopyranosyl residue E in mAb recognition derives from docking of the nonasaccharide BCDA(E)BCDA and fragments of the O-SP comprising four repeating units in the mIgA I3 Fab-binding site. The latter appears to have a distinct groove character with a deep central pocket, a type of binding site often encountered for mAbs binding internal polysaccharide sequences (75). Most interestingly, this finding is identical to that observed for Strep 9, a mouse mAb of the IgG3 subclass directed against the cell wall polysaccharide of

**TABLE 6**

**Contact between ligand and antibody in the different model for docking oligosaccharide and cyclic peptide p100c**

| Ligand atom                                    | Protein atom                                    | A $\beta$ chain |
|--|---|-----------------|
| <b>Oligosaccharide model 1 (extended)</b>      |   |                 |
| D:GlcNAc.O4-H                                  | Ile <sup>22</sup> .O                            | H1              |
| D:GlcNAc.O6-H                                  | Ala <sup>90</sup> .O                            | H3              |
| D:GlcNAc.O6                                    | Gly <sup>24</sup> .NH                           | H1              |
| A:Rha.O3-H                                     | Asp <sup>92</sup> .COO <sup>-</sup>             | H3              |
| A:Rha.O4                                       | Lys <sup>30</sup> .NH <sub>3</sub> <sup>+</sup> | H1              |
| E:Glc.O3-H                                     | Asp <sup>91</sup> .COO <sup>-</sup>             | L3              |
| E:Glc.O3                                       | Tyr <sup>93</sup> .NH                           | H3              |
| B:Rha.O4-H                                     | Ser <sup>93</sup> .O                            | L3              |
| <b>Oligosaccharide model 2 (extended)</b>      |   |                 |
| D:GlcNAc.O6                                    | Gly <sup>24</sup> .NH                           | H1              |
| E:Glc.O <sub>2</sub> -H                        | Asp <sup>91</sup> .O                            | H3              |
| E:Glc.O3-H                                     | Asp <sup>91</sup> .O                            | H3              |
| E:Glc.O3                                       | Tyr <sup>93</sup> .NH                           | H3              |
| B:Rha.O4-H                                     | Ser <sup>93</sup> .O                            | L3              |
| A:Rha.O4                                       | Ser <sup>29</sup> .OH                           | L1              |
| <b>Peptide p100c model 1</b>                   |   |                 |
| H-bonds  |   |                 |
| Lys <sup>3</sup> .NH <sub>3</sub> <sup>+</sup> | Asp <sup>91</sup> .COO <sup>-</sup>             | H3              |
| Lys <sup>3</sup> .NH <sub>3</sub> <sup>+</sup> | Asp <sup>92</sup> .COO <sup>-</sup>             | H3              |
| Leu <sup>5</sup> .O                            | Gly <sup>24</sup> .NH                           | H1              |
| His <sup>10</sup> .NH <sup>+</sup>             | Asp <sup>28</sup> .COO <sup>-</sup>             | L3              |
| His <sup>10</sup> .NH <sup>+</sup>             | Ser <sup>29</sup> .OH                           | L3              |
| Hydrophobic interactions                       |   |                 |
| Pro <sup>4</sup>                               | Trp <sup>41</sup>                               | L2              |
| Ala <sup>7</sup>                               | Trp <sup>41</sup>                               | L2              |
| <b>Peptide p100c model 2</b>                   |   |                 |
| H-bonds  |   |                 |
| Tyr <sup>2</sup> .NH                           | Ile <sup>22</sup> .O                            | H1              |
| Tyr <sup>2</sup> .OH                           | Asn <sup>26</sup> .NH <sub>2</sub>              | H1              |
| Lys <sup>3</sup> .NH <sub>3</sub> <sup>+</sup> | Asp <sup>92</sup> .COO <sup>-</sup>             | H3              |
| Hydrophobic interactions                       |   |                 |
| Tyr <sup>2</sup>                               | Trp <sup>41</sup>                               | L2              |
| Pro <sup>4</sup>                               | Trp <sup>41</sup>                               | L2              |
| Pro <sup>4</sup>                               | Phe <sup>94</sup>                               | H3              |
| Pro <sup>4</sup>                               | Pro <sup>98</sup>                               | H3              |

group A *Streptococcus* (76) made of repeats comprising a branched  $\beta$ -*N*-acetyl-D-glucosamine residue linked to a linear di-rhamnopyranosyl backbone. As found earlier, sides of the mIgA I3 groove are flanked by CDR2 of the heavy chain and CDR1 of the light chain. Moreover, aromatic residues such as Tyr<sup>45</sup> of the heavy chain and Tyr<sup>34</sup> of the light chain define the pocket region, pointing once more to the importance of such amino acids in carbohydrate recognition (77, 78). Indeed, whatever the orientation of nonasaccharide BCDA(E)BCDA, parallel or perpendicular relative to the groove binding site, the glucose residue E was always deeply buried in the central pocket of the groove and was poorly accessible to solvent. Most interestingly, relying on molecular modeling only, a heptasaccharide related to *Brucella abortus* O-SP exemplifies another O-SP-mAb interaction for which the mAb-binding site identified as a groove bearing a pocket in its center could also accommodate two binding modes of an O-SP fragment (79). As shown here, docking of *S. flexneri* 5a O-SP large fragments in the mIgA I3-binding site pointed to only one possible binding mode of the O-SP, namely the parallel mode, independently of the length of the helical repeat taken into account. Thus, in addition to the branched glucopyranosyl residue E behaving as an anchor and to the trisaccharide A(E)B providing the critical epitope exposed on the O-SP, chain elongation also takes part in O-SP:mIgA recognition, highlighting the essential contribution of some kind of conformational epitope or presentation in an extended surface. However, we are aware that small changes at the V<sub>L</sub>:V<sub>H</sub> interface of the mAb may result in significant alteration of the binding mode, which cannot be predicted at this stage (80). Thus, data provided here are only meant to provide a model of *S. flexneri* 5a O-SP binding to a homologous protective mIgA, which needs to be further assessed based on crystallographic data.

## Peptide Mimics of *S. flexneri* 5a Polysaccharide Antigen

Peptides cross-reacting with *S. flexneri* 5a O-SP have been identified by screening phage-displayed peptide libraries with protective mIgA C5 and mIgA I3 (28). Nonconstrained and constrained peptide libraries were screened. Indeed, it is expected that constraining a peptide limits its flexibility and therefore may improve its affinity for mAb binding, and consequently, may allow the selection of better mimics of the natural antigen (81). All selected peptides, whether mimotopes or mimics only, exhibited an  $IC_{50}$  value ranging from micromolar to submicromolar (this work and Footnote 5). They were better recognized than the pentasaccharide best mimicking *S. flexneri* 5a O-Ag ( $IC_{50}$  in the millimolar range) by at least one of the mIgAs used for selection. However, as outlined previously for other systems (17, 82, 83) and observed here for p115 and p100c, but not for p22, most selected peptides could discriminate between the two mIgAs used for selection. This is in agreement with the assumption that anti-polysaccharide mAbs may not necessarily recognize a single antigen topography. In line with previous work (83), it was thus hypothesized that peptide mimics reacting with a panel of anti-O-SP-specific mAbs would have a better potential to act as mimotopes than those with a strong discriminating potential. However, as observed by others (18), our data do not support this hypothesis. Indeed, among the three sequences selected for the study, discriminating p115 and p100c behave as mimotopes, whereas p22, recognized by both mIgAs, is only a mimic of *S. flexneri* 5a O-SP. Besides, considering the high affinity of p22 for mIgA C5 ( $IC_{50} \sim 30$  nM), our model fits to others, such as that on *Cryptococcus neoformans* (18, 84) and that on *N. meningococcus* C (30), suggesting that commonly used parameters for selecting peptide mimicking polysaccharide antigens, such as high-affinity binding to mAb, are not predictive of the ability of the selected peptides to act as mimotopes. Indeed, based on x-ray analysis, several lines of evidence support the idea that peptide binding to an anti-polysaccharide mAb may differ significantly from that of the natural antigen. On one hand, data on the dodecapeptide PA1 mimicking *C. neoformans* CPS suggest poor steric complementarity between PA1 and the heavy chain of the mAb used for selection, which may explain why PA1 acts only as a partial mimotope (20). On the other hand, an octapeptide functional mimic of *S. flexneri* serotype Y O-SP was found to complement the shape of the groove-type binding site of mAb SYA/J6, used for selection, much better than the ABCDA' pentasaccharide fragment of the O-SP (21). However, the peptide does not fully complement the deep pocket located in the center of the groove and occupied by rhamnose C upon binding of ABCDA'. This may explain, at least in part, the poor ability of the octapeptide to behave as an immunogenic mimic (9).

Not surprisingly, although the three peptides do not share any consensus sequence, the conformations they adopt in their free form encompass many rapidly interconverting conformers with short internal sequences spending long lifetimes organized in turn like motifs. Although p22 was found much more flexible than p115 or p100c, all three peptides adopted  $\beta$ -turn conformations, either of nonclassified type or of type I. This appears to be a rather common conformational feature for short peptides representative of antigenic regions of proteins (85) or polysaccharide antigens such as group A *Streptococcus* CP (9) and group B *Streptococcus* CP (38). Indeed, it has been suggested that a  $\beta$ -turn allows appropriate exposure of side chain residues for optimal fit within the mAb combining site. Most interestingly, in the later example, peptide FDTGAFDPDWP, a molecular mimotope of the CPS, was earlier thought to adopt a nonrandom coil conformation in aqueous solution assimilated to a nascent helix that could potentially mimic the extended helical form of the natural carbohydrate epitope (29). This

discrepancy underscores the high complexity of conformational studies dealing with short peptides. The relative heterogeneity of  $\beta$ -turn types adopted by p115, p100c, and p22 in the free form is completely lost in the bound form, as the three peptides adopt a type II  $\beta$ -turn conformation, which appears to be crucial for binding independently of the involved mAb. Thus, the lack of consensus sequence among the selected peptides seems to be compensated by structural consensus induced upon fitting to the mAb combining sites. Moreover, the type II  $\beta$ -turn structure starts from a proline (Pro<sup>3</sup> and Pro<sup>4</sup>, respectively) and ends with an alanine (Ala<sup>6</sup> and Ala<sup>7</sup>, respectively) for both p115 and p100c, underscoring the partial structural resemblance between the two peptides. Major contributions of the p115-turn to binding involve aromatic Trp<sup>5</sup> and cyclic Pro<sup>4</sup>, whereas hydrophobic Leu<sup>5</sup> was the residue most involved in mAb binding to the p100c-turn. p22 differs notably from p115 and p100c because it has no proline. In this case, aromatic His<sup>3</sup> and Phe<sup>4</sup> together with hydrophobic Leu<sup>5</sup> are the major turn components contributing to binding independently of the mAb. Noteworthy, additional residues do not seem to be engaged in mAb recognition, which may explain the ability of p22 to bind the two mIgAs. On the contrary, going from His<sup>1</sup> to Ala<sup>9</sup>, p115 binding to mIgA C5 necessitates that most residues along the peptide chain, especially those at the N terminus, make specific contacts with the combining site. Similarly, residues at the N terminus of p100c appear critical for peptide binding to mIgA I3. In particular, the two docking models obtained for p100c interacting with mIgA I3 reveal a salt bridge formed between the peptide Lys<sup>3</sup> residue and the residue Asp<sup>92</sup> within the CDR H3 loop, similarly to those observed between rhamnose A or glucose E and Asp<sup>92</sup> or Asp<sup>91</sup> of the mIgA CDR H3 loop, respectively, upon DA(E)BC binding. Although not probed at this stage, analogous ionic contributions to peptide-mAb interactions may be anticipated because all selected peptides share basic residues. However, available data suggest that independently of the peptide mimic under study, all mIgA-peptide interactions derived mostly from the direct contact of peptide aromatic residues and methyl groups with the mAb-binding site, suggesting that recognition was basically driven by hydrophobic and van der Waals contacts. In that matter, data provided for the *S. flexneri* 5a system fully support previous observations made for other models implicating peptides mimicking polysaccharide antigens in complex with specific mAbs (11).

In addition, all data reported here strongly emphasize the crucial role played by the type II  $\beta$ -turn topology in governing molecular mimicry of *S. flexneri* 5a O-antigen. Most interestingly, superimposition of family of conformations obtained for bound p22 with this obtained for bound DA(E)BC-OMe shows that the type II  $\beta$ -turn in the peptide seems to mimic the nascent helicoidal shape of the oligosaccharide main chain with the aromatic ring of Phe<sup>4</sup> having the same orientation as the crucial branched glucose E (Fig. 11). This is in agreement with previous observations showing that aromatic amino acids are considered as ideal residues for mimicking glycan side chain structures (69, 86) and that  $\beta$ -turn/extended structures may be accurate conformational mimics of helices (87, 88). Yet, except for the number of residues involved, discriminating between the binding modes of the three peptides remains difficult. Thus, based on binding complementarity only, rules governing the selection of potent *S. flexneri* 5a O-Ag mimotopes remain undisclosed. Moreover, our data emphasize that the rational design of peptides mimicking the immunological properties of polysaccharides remains a challenge (89). Because investigating the peptide binding features left several unanswered questions, the behavior of the free peptides in solution was analyzed more closely. None of the free peptides adopt the required type II  $\beta$ -turn conformation fitting in the mAb combining sites. However, they are somehow predisposed to conformational reorganization for binding. More importantly, p115 bearing Pro<sup>3</sup> and Pro<sup>4</sup> both

<sup>5</sup> V. Marcel-Peyre and A. Phalipon, unpublished data.

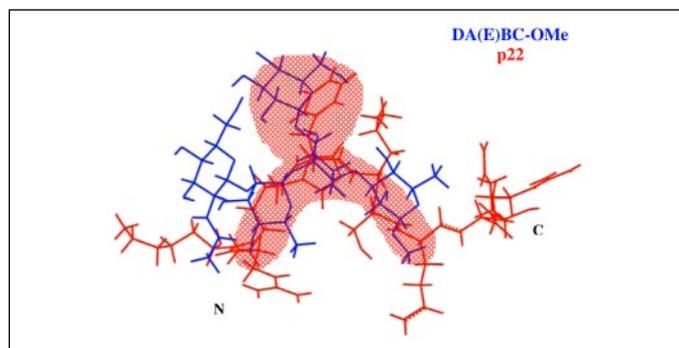


FIGURE 11. Potential conformational mimetism of the peptide p22 in relation to the pentasaccharide DA(E)BC-OMe. The figure shows superimposition of the p22 lowest energy structure with that of the DA(E)BC-OMe. Structural elements showing the potential conformational mimetism of the peptide in relation to the oligosaccharide are shaded in pink.

involved in a turn and p100c being cyclic are both highly constrained, whereas p22 adopting a poorly stabilized nonclassified turn is much more flexible. Thus, our data suggest that the pre-organized conformation adopted by the free peptide is critical for efficiency as a mimotope. It is hypothesized that, besides the number of contacts with the mAb involved in binding, it may be especially important that the free peptides have a stable structure closely resembling that of the mAb-bound conformation (20, 89). Along this line as exemplified with p22, one of the possible explanations for the failure of some peptide mimics to induce a potent anti-LPS immune response is the limited ability of small free peptides to adopt the stable conformations necessary for functional mimicry. As a consequence, the immune response induced is broadened resulting in the induction of low titers of Abs exhibiting the required specificity. To our knowledge, a crucial issue for the development of mimotope-based vaccines remains unanswered, that is: what is the acceptable ability of polysaccharide-peptide mimics to induce cross-reactive antibodies? In other words, what is an acceptable conformational flexibility for the free peptide mimics? Ideally if a monovalent vaccine is the target, the range of accessible conformations should match those of the native antigen only. In that case, induction of cross-reactive Ab would be seen as a disadvantage, because as mentioned above the level of specific Ab is consequently lowered. More importantly, the risk of inducing an Ab response directed against self-antigens may be increased. However, by allowing cross-protection against different serotypes of a given bacterial species, cross-reactivity related to controlled flexibility may constitute an advantage if the development of multivalent vaccines is envisioned. This issue has to be further investigated.

In conclusion, our work brings new data that contribute to a better understanding of molecular mimicry of polysaccharide antigens by mimotopes. Encouraging data demonstrating the feasibility of using mimotopes as surrogates of native antigens for the induction of protective immune responses have been provided in the last 10 years. However, major breakthroughs are still needed to further establish the rules governing the molecular mimicry of polysaccharide antigens by peptides. Those should help the future design of efficient mimotope-based vaccines.

**Acknowledgments**—We warmly thank Françoise Baleux (Unité de Chimie Organique, Institut Pasteur, France) for the expertise, advice, and helpful discussion on peptide synthesis and purification. We also thank Audrey Thuizat (Unité de Pathogénie Microbienne Moléculaire, Institut Pasteur, Paris) and Monique Reinhardt (Institut de Biochimie, ISREC, Epalinges, Switzerland) for their technical support in the production and sequencing of mIgA, respectively. The 600-MHz NMR spectrometer was funded by the Région Ile de France and the Institut Pasteur (Paris, France).

## REFERENCES

- MacLeod, C. M., Hodges, R. G., Heidelberg, M., and Bernhard, W. G. (1945) *J. Exp. Med.* **82**, 445–465
- Roy, R. (2004) *Drug Discovery Today: Technologies* **1**, 327–336
- Goebel, H. H., Ikeda, K., Schulz, F., Burck, U., and Kohlschütter, A. (1981) *Acta Neuropathol.* **55**, 247–249
- Pozsgay, V., Chu, C., Pannell, L., Wolfe, J., Robbins, J. B., and Schneerson, R. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96**, 5194–5197
- Peeters, C. C., Evenberg, D., Hoogerhout, P., Kayhty, H., Saarinen, L., van Boeckel, C. A., van der Marel, G. A., van Boom, J. H., and Poolman, J. T. (1992) *Infect. Immun.* **60**, 1826–1833
- Verez-Bencomo, V., Fernandez-Santana, V., Hardy, E., Toledo, M. E., Rodriguez, M. C., Heynngnezz, L., Rodriguez, A., Baly, A., Herrera, L., Izquierdo, M., Villar, A., Valdes, Y., Cosme, K., Deler, M. L., Montane, M., Garcia, E., Ramos, A., Aguilar, A., Medina, E., Torano, G., Sosa, I., Hernandez, I., Martinez, R., Muzachio, A., Carmenates, A., Costa, L., Cardoso, F., Campa, C., Diaz, M., and Roy, R. (2004) *Science* **305**, 522–525
- Stein, H., Gatter, K. C., Heryet, A., and Mason, D. Y. (1984) *Lancet* **2**, 71–73
- Pirofski, L. A. (2001) *Trends Microbiol.* **9**, 445–451
- Johnson, M. A., Rotondo, A., and Pinto, B. M. (2002) *Biochemistry* **41**, 2149–2157
- Monzavi-Karbassi, B., Cunto-Amesty, G., Luo, P., and Kieber-Emmons, T. (2002) *Trends Biotechnol.* **20**, 207–214
- Luo, P., Agadjanyan, M., Qiu, J., Westerink, M. A., Steplewski, Z., and Kieber-Emmons, T. (1998) *Mol. Immunol.* **35**, 865–879
- Cunto-Amesty, G., Dam, T. K., Luo, P., Monzavi-Karbassi, B., Brewer, C. F., Van Cott, T. C., and Kieber-Emmons, T. (2001) *J. Biol. Chem.* **276**, 30490–30498
- Fleuridor, R., Lees, A., and Pirofski, L. (2001) *J. Immunol.* **166**, 1087–1096
- Lesinski, G. B., and Westerink, M. A. (2001) *J. Microbiol. Methods* **47**, 135–149
- Maitta, R. W., Datta, K., Lees, A., Belouski, S. S., and Pirofski, L. A. (2004) *Infect. Immun.* **72**, 196–208
- Buchwald, U. K., Lees, A., Steinitz, M., and Pirofski, L. A. (2005) *Infect. Immun.* **73**, 325–333
- Harris, S. L., Craig, L., Mehroke, J. S., Rashed, M., Zwick, M. B., Kenar, K., Toone, E. J., Greenspan, N., Auzanneau, F. I., Marino-Albernas, J. R., Pinto, B. M., and Scott, J. K. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94**, 2454–2459
- Valadon, P., Nussbaum, G., Oh, J., and Scharff, M. D. (1998) *J. Immunol.* **161**, 1829–1836
- Cunto-Amesty, G., Luo, P., Monzavi-Karbassi, B., Lees, A., and Kieber-Emmons, T. (2001) *Vaccine* **19**, 2361–2368
- Young, A. C., Valadon, P., Casadevall, A., Scharff, M. D., and Sacchettini, J. C. (1997) *J. Mol. Biol.* **274**, 622–634
- Vyas, N. K., Vyas, M. N., Chervenak, M. C., Bundle, D. R., Pinto, B. M., and Quiocho, F. A. (2003) *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15023–15028
- Johnson, M. A., and Pinto, B. M. (2004) *Carbohydr. Res.* **339**, 907–928
- Kotloff, K. L., Winickoff, J. P., Ivanoff, B., Clemens, J. D., Swerdlow, D. L., Sansonetti, P. J., Adak, G. K., and Levine, M. M. (1999) *Bull. W.H.O.* **77**, 651–666
- Phalipon, A., and Sansonetti, P. J. (2003) *Crit. Rev. Immunol.* **23**, 371–401
- Ashkenazi, S., Passwell, J. H., Harlev, E., Miron, D., Dagan, R., Farzan, N., Ramon, R., Majadly, F., Bryla, D. A., Karpas, A. B., Robbins, J. B., and Schneerson, R. (1999) *J. Infect. Dis.* **179**, 1565–1568
- Wright, K., Guerreiro, C., Laurent, I., Baleux, F., and Mulard, L. A. (2004) *Org. Biomol. Chem.* **2**, 1518–1527
- Belot, F., Guerreiro, C., Baleux, F., and Mulard, L. A. (2005) *Chemistry* **11**, 1625–1635
- Phalipon, A., Folgori, A., Arondel, J., Sgarrella, G., Fortugno, P., Cortese, R., Sansonetti, P. J., and Felici, F. (1997) *Eur. J. Immunol.* **27**, 2620–2625
- Pincus, S. H., Lepage, S. R., Jung, R. F., Massey, J. G., and Jaseja, M. (2001) *Int. Rev. Immunol.* **20**, 221–227
- Prinz, D. M., Smithson, S. L., and Westerink, M. A. (2004) *J. Immunol. Methods* **285**, 1–14
- Lindberg, A. A., Cam, P. D., Chan, N., Phu, L. K., Trach, D. D., Lindberg, G., Karlsson, K., Karnell, A., and Ekwall, E. (1991) *Rev. Infect. Dis.* **13**, Suppl. 4, 231–237
- Mulard, L. A., and Ughetto-Monfrin, J. (2000) *J. Carbohydr. Chem.* **19**, 193–220
- Clement, M. J., Imbert, A., Phalipon, A., Perez, S., Simenel, C., Mulard, L. A., and Delepierre, M. (2003) *J. Biol. Chem.* **278**, 47928–47936
- Clare, G. M., and Gronenborn, A. M. (1982) *J. Magn. Reson.* **48**, 402–417
- Clare, G. M., and Gronenborn, A. M. (1983) *J. Magn. Reson.* **53**, 423–442
- Mayer, M., and Meyer, B. (1999) *Angew. Chem. Int. Ed.* **38**, 1784–1788
- Johnson, M. A., and Pinto, B. M. (2002) *J. Am. Chem. Soc.* **124**, 15368–15374
- Johnson, M. A., Jaseja, M., Zou, W., Jennings, H. J., Copie, V., Pinto, B. M., and Pincus, S. H. (2003) *J. Biol. Chem.* **278**, 24740–24752
- Johnson, M. A., and Pinto, B. M. (2004) *Bioorg. Med. Chem.* **12**, 295–300
- Mulard, L. A., Clément, M.-J., Segat-Dioury, F., and Delepierre, M. (2002) *Tetrahedron* **58**, 2593–2604
- Phalipon, A., Kaufmann, M., Michetti, P., Cavaillon, J. M., Huerre, M., Sansonetti, P., and Kraehenbuhl, J. P. (1995) *J. Exp. Med.* **182**, 769–778

## Peptide Mimics of *S. flexneri* 5a Polysaccharide Antigen

42. Rance, M., Sorensen, O. W., Bodenhausen, G., Wagner, G., Ernst, R. R., and Wuthrich, K. (1983) *Biochem. Biophys. Res. Commun.* **117**, 479–485
43. Griesinger, C., Otting, G., Wüthrich, K., and Ernst, R. R. (1988) *J. Am. Chem. Soc.* **110**, 7870–7872
44. Kessler, H., Griesinger, C., Kerssebaum, R., Wagner, K., and Ernst, R. R. (1987) *J. Am. Chem. Soc.* **109**, 607–609
45. Piotto, M., Saudek, V., and Sklenar, V. (1992) *J. Biomol. NMR* **2**, 661–665
46. States, D. J., Haberkorn, R. A., and Ruben, D. J. (1982) *J. Magn. Reson.* **48**, 286–292
47. Jimenez-Barbero, J., and Peters, T. (eds) (2002) *NMR Spectroscopy of Glycoconjugates*, pp. 289–307, Wiley-VCH, New York
48. Herfurth, L., Weimar, T., and Peters, T. (2000) *Angew. Chem. Int. Ed. Engl.* **39**, 2097–2099
49. Baleja, J., Moul, J., and Sykes, B. D. (1990) *J. Magn. Reson.* **87**, 375–384
50. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, Inc., New York
51. Pardi, A., Billeter, M., and Wüthrich, K. (1984) *J. Mol. Biol.* **180**, 741–751
52. Guntert, P., Mumenthaler, C., and Wüthrich, K. (1997) *J. Mol. Biol.* **273**, 283–298
53. Koradi, R., Billeter, M., and Wüthrich, K. (1996) *J. Mol. Graphics* **14**, 51–55
54. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402
55. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res.* **28**, 235–242
56. Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L., and Sutcliffe, M. (1988) *Eur. J. Biochem.* **172**, 513–520
57. Clark, M., Cramer, R. D. I., and van den Opdenbosch, N. (1989) *J. Comput. Chem.* **10**, 982–1012
58. Laskowski, R., MacArthur, M., Moss, D., and Thornton, J. (1993) *J. Appl. Crystallogr.* **26**, 283–291
59. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998) *J. Comput. Chem.* **19**, 1639–1662
60. Imberty, A., Bettler, E., Karababa, M., Mazeau, K., Petrova, P., and Pérez, S. (1999) in *Perspectives in Structural Biology* (Vijayan, M., Yathindra, N., and Kolaskar, A. S., eds) pp. 392–409, Indian Academy of Sciences and Universities Press, Hyderabad
61. Rose, G. D., Gierasch, L. M., and Smith, J. A. (1985) *Adv. Protein Chem.* **37**, 1–109
62. Wilmot, C. M., and Thornton, J. M. (1988) *J. Mol. Biol.* **203**, 221–232
63. Pokkuluri, P. R., Bouthillier, F., Li, Y., Kuderova, A., Lee, J., and Cygler, M. (1994) *J. Mol. Biol.* **243**, 283–297
64. Herron, J. N., He, X. M., Ballard, D. W., Blier, P. R., Pace, P. E., Bothwell, A. L., Voss, E. W., Jr., and Edmundson, A. B. (1991) *Proteins* **11**, 159–175
65. Romesberg, F. E., Spiller, B., Schultz, P. G., and Stevens, R. C. (1998) *Science* **279**, 1929–1933
66. Tulip, W. R., Harley, V. R., Webster, R. G., and Novotny, J. (1994) *Biochemistry* **33**, 7986–7997
67. Shoham, M. (1993) *J. Mol. Biol.* **232**, 1169–1175
68. Agadjanyan, M., Luo, P., Westerink, M. A., Carey, L. A., Hutchins, W., Steplewski, Z., Weiner, D. B., and Kieber-Emmons, T. (1997) *Nat. Biotechnol.* **15**, 547–551
69. Dinglasan, R. R., Porter-Kelley, J. M., Alam, U., and Azad, A. F. (2005) *Vaccine* **23**, 2717–2724
70. Bernardi, A., Potenza, D., Capelli, A. M., Garcia-Herrero, A., Canada, F. J., and Jimenez-Barbero, J. (2002) *Chemistry* **8**, 4597–4612
71. Bundle, D. R., Eichler, E., Gidney, M. A., Meldal, M., Ragauskas, A., Sigurskjold, B. W., Sinnott, B., Watson, D. C., Yaguchi, M., and Young, N. M. (1994) *Biochemistry* **33**, 5172–5182
72. Vyas, N. K., Vyas, M. N., Chervenak, M. C., Johnson, M. A., Pinto, B. M., Bundle, D. R., and Quiocho, F. A. (2002) *Biochemistry* **41**, 13575–13586
73. Bock, K., Josephson, S., and Bundle, D. R. (1982) *J. Chem. Soc. Perkins Trans.* **2**, 59–70
74. Kreis, U. (1997) *J. Mol. Struct. (Theochem)* **395–396**, 389–409
75. Cisar, J., Kabat, E. A., Dorner, M. M., and Liao, J. (1975) *J. Exp. Med.* **142**, 435–459
76. Pitner, J. B., Beyer, W. F., Venetta, T. M., Nycz, C., Mitchell, M. J., Harris, S. L., Marino-Albernas, J. R., Auzanneau, F. I., Forooghian, F., and Pinto, B. M. (2000) *Carbohydr. Res.* **324**, 17–29
77. Padlan, E. A. (1990) *Proteins* **7**, 112–124
78. Bernardi, A., Arosio, D., Potenza, D., Sanchez-Medina, I., Mari, S., Canada, F. J., and Jimenez-Barbero, J. (2004) *Chemistry* **10**, 4395–4406
79. Oomen, R. P., Young, N. M., and Bundle, D. R. (1991) *Protein Eng.* **4**, 427–433
80. Rose, D. R., Przybylska, M., To, R. J., Kayden, C. S., Oomen, R. P., Vorberg, E., Young, N. M., and Bundle, D. R. (1993) *Protein Sci.* **2**, 1106–1113
81. Lauvrak, V., Berntzen, G., Heggelund, U., Herstad, T. K., Sandin, R. H., Dalseg, R., Rosenqvist, E., Sandlie, I., and Michaelsen, T. E. (2004) *Scand. J. Immunol.* **59**, 373–384
82. Valadon, P., Nussbaum, G., Boyd, L. F., Margulies, D. H., and Scharff, M. D. (1996) *J. Mol. Biol.* **261**, 11–22
83. Pincus, S. H., Smith, M. J., Jennings, H. J., Burritt, J. B., and Glee, P. M. (1998) *J. Immunol.* **160**, 293–298
84. Beenhouwer, D. O., May, R. J., Valadon, P., and Scharff, M. D. (2002) *J. Immunol.* **169**, 6992–6999
85. Dyson, H. J., Rance, M., Houghten, R. A., Lerner, R. A., and Wright, P. E. (1988) *J. Mol. Biol.* **201**, 161–200
86. Monzavi-Karbassi, B., Shamloo, S., Kieber-Emmons, M., Jousheghany, F., Luo, P., Lin, K. Y., Cunto-Amesty, G., Weiner, D. B., and Kieber-Emmons, T. (2003) *Vaccine* **21**, 753–760
87. Monfardini, C., Kieber-Emmons, T., VonFeldt, J. M., Godillot, A. P., Voet, D., Weiner, D. B., and Williams, W. V. (1996) *Proc. Assoc. Am. Physicians* **108**, 420–431
88. Mer, G., Kellenberger, E., and Lefevre, J. F. (1998) *J. Mol. Biol.* **281**, 235–240
89. Johnson, M. A., Eniade, A. A., and Pinto, B. M. (2003) *Bioorg. Med. Chem.* **11**, 781–788
90. Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D. (1995) *J. Biomol. NMR* **5**, 67–81
91. Wilmot, C. M., and Thornton, J. M. (1990) *Protein Eng.* **3**, 479–493
92. Richardson, J. S. (1981) *Adv. Protein Chem.* **34**, 167–339

**ANALYSE QUANTITATIVE EN TROIS  
DIMENSIONS DES RELATIONS  
STRUCTURE-ACTIVITE**

## A - INTRODUCTION

Les coordonnées atomiques des protéines sont obtenues par des analyses structurales telles que la diffraction des rayons X ou la RMN. Ces techniques nécessitent des échantillons purs, en quantités importantes et, pour les rayons X, sous forme cristalline. Il est malheureusement fréquent que de tels échantillons ne soient pas disponibles, en particulier lorsque la cible est une protéine membranaire. Outre les difficultés de production, extraire une protéine de sa membrane est problématique. La membrane assure en partie le maintien de ces structures dans leur forme active et le passage dans le milieu d'extraction entraîne généralement un « effondrement » de la partie incluse dans la membrane. Cette modification de la conformation induit généralement une perte partielle ou totale de l'activité. La dénaturation des protéines membranaires lors de l'extraction reste l'un des défis majeurs de la biologie structurale.

Faute de données sur la structure tridimensionnelle d'une protéine, la découverte et l'optimisation de nouvelles molécules actives sur cette cible devront se fonder sur l'étude des relations structure – activité de ses ligands. D'une façon générale, il est bien connu que dans une série de molécules analogues, il existe une forte corrélation entre les structures et les propriétés observées. Un exemple simple qui illustre ce phénomène, est la corrélation entre le nombre de carbones des alcanes et leur point d'ébullition. Il y a une relation forte entre l'augmentation de la température d'ébullition et l'augmentation du nombre de carbones. On peut utiliser cette relation pour prévoir le point d'ébullition des alcanes à partir des mesures relatives à quelques uns.

On peut ainsi étudier nombre de propriétés physico-chimiques et biologiques à partir de descripteurs moléculaires. Dans le champ de la chimie pharmaceutique, on s'intéressera aux propriétés pharmacocinétiques telles que solubilité, capacité à franchir des membranes, dégradation, élimination, toxicité et aux propriétés pharmacodynamiques comme le pouvoir inhibiteur ou activateur et la sélectivité.

A partir de mesures quantitatives des propriétés étudiées, des méthodes statistiques permettent de construire des modèles qui décrivent les observations de façon rationnelle. Les modèles validés par une série de contrôles permettent ensuite de prévoir l'activité de nouvelles molécules et d'établir un ordre de priorité avant leur synthèse. Le processus cyclique, de construction de modèles, prévisions, synthèses et tests, se répète jusqu'à l'obtention d'une molécule ayant la combinaison de propriétés désirée. Cette méthode très générale compte un très grand nombre de variantes regroupées sous l'acronyme QSPR (*Quantitative Structure Property Relationship*) et

lorsqu'elle s'applique à l'étude d'activités biologiques on parle alors de QSAR (*Quantitative Structure Activity Relationship*).

Historiquement, cette technique a été inventée et développée par des chimistes pour répondre à une question simple : « Quelle est la prochaine molécule à synthétiser ? ». En d'autres termes, comment interpréter les résultats déjà obtenus et identifier les tendances pour optimiser les structures ? Les premiers modèles exploitables remontent aux années 1960 avec la publication de l'équation de Hansch [1]. Le nombre d'algorithmes pour les analyses QSAR s'est considérablement accru depuis et la méthode a produit quantité de résultats pertinents.

Par ailleurs, les résultats de ces analyses intéressent autant les chimistes qui s'en servent pour orienter leurs synthèses, que les biologistes qui en tirent des informations structurales et fonctionnelles sur les protéines ciblées.

Dans cette seconde partie nous allons décrire les principes et outils statistiques standards utilisés en QSAR, les méthodes de validation des modèles et l'adaptation de ces méthodes à l'étude de propriétés pharmacochimiques. L'application de cette technique à l'analyse quantitative en trois dimensions des relations structure-activité d'une série de flavonoïdes et de boerhavinones inhibiteurs de la protéine BCRP illustrera sa capacité à décrire les interactions mises en jeu et son intérêt pour la conception de nouveaux composés actifs.

## **B - PRINCIPES ET METHODES DE MODELISATION STATISTIQUE**

### ***I - PRINCIPES GENERAUX DE STATISTIQUES***

#### **1 - HOMOGENEITE DES DONNEES.**

##### **a . Source des données**

L'homogénéité des données biologiques est fondamentale. Si l'on veut comparer l'activité biologique d'une série de molécules, il faut s'assurer que cette activité est le résultat de leur interaction avec une seule et même cible et plus précisément avec le même site actif.

L'activité doit être mesurée par un seul et même test, avec des conditions expérimentales identiques pour chaque molécule. L'obtention de résultats complets et homogènes sur toute la série est souvent difficile. La mise au point des tests, leur réalisation ainsi que la synthèse des composés demandent beaucoup de temps et de moyens.

Les composés testés ont deux origines possibles : ce sont soit des produits de synthèse soit des produits d'extraction à partir de matériel biologique et de plantes en particulier. Quelle que soit son origine, il arrive qu'un échantillon ne soit pas pur mais corresponde à un mélange racémique. Le résultat du test d'un tel échantillon pose alors problème : il est impossible de savoir quelle est la contribution de chaque énantiomère dans l'activité observée. Il est donc exclu d'incorporer des structures dont la propriété étudiée est mesurée sur un mélange racémique.

##### **b . Traitement des données brutes**

L'homogénéité de la distribution des valeurs mesurées doit être contrôlée. En effet, les analyses statistiques telles que les régressions multiples et plus encore les indices paramétriques qui en découlent, reposent sur l'hypothèse que la distribution des valeurs observées suit une loi Normale. Il est donc nécessaire de contrôler la normalité de cette distribution. Il existe pour cela des tests statistiques de normalité mais la simple représentation des données sur un histogramme de distribution permet d'évaluer cette caractéristique.

Dans le cas défavorable, des transformations mathématiques permettent, parfois, de retrouver une distribution normale sans que l'information contenue dans le jeu de données ne soit modifiée. Le test de Box-Cox [2] nous aide à trouver la transformation adéquate en cherchant un

paramètre  $\lambda$  tel que  $X^\lambda$  suive une loi normale. Le paramètre  $\lambda$  est optimisé par la méthode du maximum de vraisemblance et l'on peut transformer les données brutes avec la relation générale suivante :

$$\text{Équation 1 : } f(X) = \frac{(X^\lambda - 1)}{\lambda}$$

En fonction de la valeur de  $\lambda$ , la relation peut être simplifiée :

- Si  $\lambda \geq 2$ , prendre le carré des valeurs brutes ;
- Si  $\lambda = 1$  ou si la valeur 1 est dans l'intervalle de confiance de  $\lambda$ , aucune transformation n'est nécessaire.
- Si  $\lambda = 0$  ou si la valeur 0 est dans l'intervalle de confiance de  $\lambda$ , utiliser un logarithme ;

La normalisation de la distribution des données est un préalable indispensable à l'emploi d'outils de régressions multiples et le choix d'une transformation dépend de la distribution des données sur la gamme d'activité.

Une autre transformation est possible lorsque la variable mesurée est bornée et que de nombreuses mesures sont très proches voire identiques. C'est le cas par exemple lorsque le test biologique mesure un pourcentage d'inhibition. La réponse du test varie entre 0 et 100 et la variable ne peut pas prendre de valeurs inférieures à 0 ou supérieures à 100. Les molécules non actives auront toutes une valeur d'activité  $x$  nulle ou proche de 0. A l'inverse les molécules très actives auront toutes une activité  $x$  proche ou égale à 100 %. La variable est bornée à gauche et à droite.

En divisant la variable d'activité  $x$  par  $(100-x)$ , la borne de droite est annulée car lorsque  $x$  tend vers 100,  $x/(100-x)$  tend vers  $+\infty$ . Si  $x$  tend vers 0 alors  $x/(1-x)$  tend aussi vers 0. On applique alors une seconde transformation, de type logarithmique de sorte que, lorsque  $x$  tend vers 0, la transformation tend vers  $-\infty$ . Cette double transformation est appelée le Logit de  $x$  [3].

$$\text{Équation 2 : } \text{Logit} = \log\left(\frac{x}{100-x}\right)$$

Les valeurs proches de 50 seront peu affectées par la transformation alors que les valeurs proches des bornes seront transformées de façon exponentielle avec l'éloignement de la médiane.

### **c . Echantillonnage de l'espace chimique**

L'homogénéité des structures et leur diversité sont également des facteurs importants dans la qualité des modèles construits. La diversité définit l'espace chimique que l'analyse va couvrir et l'homogénéité traduit la régularité de l'échantillonnage de cet espace. Il existe une multitude de descripteurs pour exprimer ces deux notions et malheureusement aucun ne permet de décrire toutes les situations. Ces descripteurs chimiques sont généralement les mêmes que ceux utilisés par l'analyse QSAR.

## **2 - LES DESCRIPTEURS CHIMIQUES**

Il existe des centaines de descripteurs pour représenter les molécules et leurs fragments. Chaque descripteur quantifie une caractéristique. Le choix des descripteurs conditionne la qualité du modèle puisque l'on cherche les caractéristiques liées à l'activité. C'est l'analyse statistique qui validera ce choix. Un descripteur est un jeu de valeurs en relation avec une structure chimique en 2 ou en 3 dimensions (descripteur 2D et 3D respectivement). Chaque descripteur peut lui-même comporter une ou plusieurs dimensions.

### **a . Descripteurs 2D**

Les descripteurs 2D sont des propriétés numériques qui peuvent être calculées à partir de la table de connectivité d'une molécule ou d'une représentation planaire (2D) de la structure. Ils sont basés sur les éléments présents, les charges partielles, la nature des liaisons, etc mais n'exploitent pas les coordonnées atomiques spatiales. On peut calculer ainsi :

- des propriétés physiques : polarisabilité, charge totale, réfractivité moléculaire, masse, densité, coefficient de partage eau/octanol (logP) ...
- des approximations d'aires de surfaces : surface de Van der Waals, surface accessible au solvant de chaque atome en relation éventuellement avec une autre propriété atomique comme la réfractivité ou la contribution au logP ;
- le dénombrement des atomes, liaisons et pivots ;
- des indices de formes et de connectivité : indices de Kier et Hall [4] ;
- des descripteurs topologiques : matrices de distance et de connectivité ;
- les dénombrements de fonctions pharmacophoriques : donneurs, accepteurs de liaison hydrogène, polaire (donneur et accepteur), positif, négatif, hydrophobe et autres ... ;

- les descripteurs de charges partielles : charge partielle positive (négative) totale, aire de la surface de Van der Waals positive (négative), aire de la surface de Van der Waals polaire (hydrophobe), ... ;
- les indices de Hansch :  $\rho$ ,  $\sigma$  et  $\pi$  [1] ;
- etc.

### **b . Descripteurs 3D**

Les descripteurs 3D décrivent des objets tridimensionnels et se répartissent en 2 groupes : ceux qui ne dépendent que des coordonnées internes de la molécule et ceux qui dépendent de son orientation absolue. Parmi tous ces descripteurs on peut citer :

- les descripteurs de l'énergie potentielle : valeur de l'énergie potentielle et composantes de cette énergie : Van der Waals, électrostatiques, atomes « hors-du-plan », torsion, solvation, etc. ;
- des descripteurs de formes et de volumes : aire et volume de Van der Waals, surface accessible au solvant, moment d'inertie, globularité (molécule sphérique, plane ou linéaire) ;
- des descripteurs du moment dipolaire : orientation et intensité ;
- les champs de potentiel d'interaction moléculaire : stérique, électrostatique, lipophile, donneur ou accepteur d'hydrogène [5-7];
- les descripteurs électroniques : densité électronique, distribution de charges ;
- les descripteurs quantiques basés sur les orbitales moléculaires : TQSI [8], MQSM [8], QS-SM [9] ;
- les descripteurs de spectres énergétiques : IR, RAMAN, RMN [10-12].
- etc.

Le choix des descripteurs dépend des outils dont on dispose, de la nature des composés décrits et de la propriété ciblée. L'expérience du modélisateur est, ici, mise à contribution.

## **3 - LES MODELES STATISTIQUES**

Toutes les caractéristiques d'une molécule ne sont pas liées à son activité biologique. Les descripteurs chimiques sont souvent corrélés entre eux et donnent la même information.

L'objectif de l'analyse statistique est justement de « démêler » ces descripteurs et d'identifier ceux qui sont corrélés à la variable cible, qui produisent du signal, de ceux qui ne le sont pas, qui produisent du bruit. L'analyse statistique permet également d'identifier les descripteurs qui sont corrélés entre eux pour ne garder que les principaux et réduire ainsi la redondance d'informations.

Dans la terminologie des statistiques, la propriété étudiée est appelée variable cible ou variable dépendante ou encore variable Y. C'est la caractéristique que l'on cherche à expliquer et à optimiser.

Les descripteurs, en l'occurrence chimiques, sont appelés variables explicatives, variables indépendantes ou encore variables X.

L'analyse statistique détermine et quantifie les corrélations entre les descripteurs et la variable cible. Elle indique également la contribution relative de chaque descripteur dans l'explication globale de l'activité. Le modèle statistique est une équation donnant la valeur de la variable cible en fonction de la somme des valeurs pondérées des descripteurs. Les principaux outils statistiques pour obtenir un modèle sont :

- la régression linéaire multivariée (*Multivariate Linear Regression* - MLR) ;
- la régression en composante principale (*Principale Component Regression* - PCR) ;
- la régression des moindres carrés partiels (*Partial Least Squares* - PLS) [13] ;
- les réseaux de neurones artificiels (*Artificial Neural Network* - ANN) [14, 15].

Le choix de la méthode dépend principalement de la question qui est posée et de la nature des données à traiter.

La MLR est d'un usage assez restrictif et peu adaptée au QSAR [16] : elle requiert un jeu de données très complet où toutes les combinaisons de substituants ont été testées. En d'autres termes, il faut faire autant d'expériences que de variations possibles, ce qui est rarement le cas en pratique.

La PLS est une variante de la MLR, plus souple et plus robuste. Elle permet d'étudier des jeux de données où l'on a plus de variables explicatives que d'expériences. C'est la méthode la plus adaptée aux études QSAR.

La PCR est un outil de simplification qui permet d'identifier les dimensions essentielles d'un problème multivarié et de transposer les données dans ces dimensions principales.

Les réseaux neuronaux ont un fonctionnement particulier. L'un de leurs principes fondamentaux est l'approximation fonctionnelle : ils apprennent une fonction en regardant des exemples de la

dite fonction. Le réseau neuronal peut être considéré comme une boîte noire qui utilise les mêmes variables en entrée et en sortie que la fonction à imiter.

#### 4 - LES OUTILS DE VALIDATION DES MODELES

La validation des modèles est une autre étape sensible. Le modèle QSAR étant le résultat d'une analyse statistique, son interprétation et son exploitation doivent se faire dans le cadre très précis du domaine couvert par l'analyse. Toute extrapolation hors de ce cadre exige beaucoup de précautions et est d'autant plus hasardeuse qu'on s'éloigne du cadre. Pour éviter les erreurs, tant au moment de la validation qu'au moment de l'exploitation, les limites du modèle doivent être clairement établies : le modèle est-il robuste, quel est son pouvoir de prévision et dans quel espace chimique ?

##### a . L'ajustement

Une méthode statistique de type PLS donne un premier indice de qualité du modèle qui est le coefficient de corrélation interne  $r^2$ . Cet indice exprime l'écart entre les valeurs observées et les valeurs calculées par le modèle pour les structures du jeu d'apprentissage. Lorsque sa valeur approche de 1, le modèle est bien ajusté aux données expérimentales, il est capable de les reproduire. Lorsque  $r^2$  s'approche de 0, le modèle n'est pas corrélé aux observations. Les détails du calcul de  $r^2$  sont précisés au paragraphe d .

Ce paramètre est important mais n'est pas suffisant. On a, entre autre, besoin de connaître le nombre optimal de composantes, également appelées tendances, à intégrer au modèle. Il est également souhaitable de connaître l'intervalle de confiance des coefficients affectés à chaque descripteur. Ces informations s'obtiennent par les méthodes de re-échantillonnage (*bootstapping*) et de validation croisée (*crossvalidation*) respectivement.

##### b . La précision de l'ajustement

Le *bootstrapping* consiste à dériver un modèle sur un échantillon du jeu d'apprentissage, de même taille que le jeu d'apprentissage et constitué à partir d'éléments tirés au hasard, certains éléments pouvant être piochés plusieurs fois. L'opération est répétée pour constituer plusieurs jeux d'apprentissage. L'écart entre la moyenne des coefficients ainsi obtenus et les coefficients calculés sur le jeu d'apprentissage complet donne une idée de la variabilité de ces coefficients. Cette procédure permet d'évaluer les intervalles de confiances des coefficients du modèle.

### **c . Le pouvoir de prévision interne**

La validation croisée nous renseigne sur le pouvoir de prévision d'un modèle. Ce pouvoir de prévision est dit « interne » car il est calculé à partir des structures utilisées pour construire ce modèle. La validation croisée est particulièrement utile en PLS car elle permet aussi d'établir le nombre de composantes qui optimise le rapport signal/bruit. Lorsque l'on compare des modèles de pouvoirs prévisionnels équivalents, le modèle ayant le moins de composantes sera privilégié. En effet, moins il y a de composantes, moins le modèle est complexe, plus il est robuste.

La validation croisée consiste à re-dériver un modèle en oubliant un ou plusieurs éléments du jeu d'apprentissage et à prévoir l'activité de ces éléments oubliés avec le modèle ainsi dérivé. L'opération est répétée jusqu'à ce que tous les éléments du jeu aient été oubliés une fois.

Dans la plupart des logiciels c'est la procédure LOO (*Leave One Out*) qui est implémentée. Un seul composé est oublié avant chaque dérivation et son activité est calculée par le modèle correspondant. Le coefficient de corrélation  $q^2$  entre les activités ainsi calculées et les activités observées exprime le pouvoir de prévision interne du modèle.

Plus la valeur de  $q^2$  se rapproche de l'unité, meilleur est le pouvoir de prévision interne du modèle. A l'inverse, un  $q^2$  proche de 0 (zéro) voire négatif invalide le modèle de façon certaine. Pour être acceptable, le pouvoir de prévision interne doit être supérieur à 0,6.

La variante qui consiste à oublier plusieurs composés en même temps (*Leave Many Out - LMO*) est plus robuste que LOO. Cependant, telle qu'elle est implémentée dans SYBYL, elle donne des valeurs de  $q^2$  qui ne sont pas reproductibles car la sélection étant aléatoire, on ne sait pas quelles sont les molécules écartées à chaque cycle.

### **d . Le test de Fischer**

Le test F de Fischer permet de savoir si un modèle de régression est globalement significatif ou non. Le rappel de quelques notions de statistiques de base est préalablement nécessaire pour expliquer ce test.

La somme des carrés des écarts entre les valeurs mesurées ( $y$ ) et la valeur de l'activité moyenne ( $\bar{Y}$ ) est appelée somme des écarts totaux et notée  $ss_{tot}$ . L'activité moyenne que l'on veut utiliser ici est la moyenne de toutes les molécules actives qui existent. En langage statistique cette moyenne est appelée « espérance de la population » ( $\bar{Y}$ ). Elle peut être estimée par la moyenne d'un échantillon de la population et plus cet échantillon sera grand plus il sera représentatif de la population. Dans le cadre d'une étude QSAR, la valeur de l'activité moyenne est estimée par la

moyenne de toutes les observations disponibles ( $\bar{y}$ ). Le calcul de  $ss_{tot}$  est décrit par l'Équation 3.

$$\text{Équation 3 : } \quad ss_{tot} = \sum (y - \bar{y})^2$$

L'activité observée est le résultat de l'addition de l'activité réelle et de l'erreur expérimentale appelée également incertitude de la mesure.

Lorsque l'on a construit un modèle qui s'efforce d'expliquer et de reproduire une variable cible comme l'activité, on observe là encore des différences entre les valeurs calculées ( $\hat{y}$ ) et les valeurs mesurées. Cet écart est appelé « résidu ». Le résidu est dû d'une part à l'incertitude de la mesure expérimentale et d'autre part à l'imperfection du modèle.

La somme des carrés totaux est la somme de la somme des carrés du modèle ( $ss_{mod}$ ) et de la somme des carrés résiduels ( $ss_{res}$ ) (Équation 4). Le calcul de la somme des carrés résiduels est décrit par l'Équation 5. La somme des carrés du modèle est toujours calculée par la différence des deux autres (Équation 6).

$$\text{Équation 4 : } \quad ss_{tot} = ss_{mod} + ss_{res}$$

$$\text{Équation 5 : } \quad ss_{res} = \sum (y - \hat{y})^2$$

$$\text{Équation 6 : } \quad ss_{mod} = ss_{tot} - ss_{res} = \sum (y - \bar{y})^2 - \sum (y - \hat{y})^2$$

La concordance entre la prévision et la mesure est évaluée par un indice qui le rapport de la somme des carrés du modèle sur la somme des carrés totaux (Équation 7). Ce coefficient s'appelle le coefficient de corrélation interne ( $r^2$ ). C'est la proportion de l'activité observée, que le modèle est capable d'expliquer. Puisqu'il s'agit d'un rapport, sa magnitude n'est pas affectée par l'échelle de mesure.

$$\text{Équation 7 : } \quad r^2 = \frac{ss_{mod}}{ss_{tot}} = \frac{ss_{tot} - ss_{res}}{ss_{tot}} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

La variance est définie comme le rapport de la somme des carrés d'une variable sur le nombre de degrés de liberté de cette variable. Dans le cadre d'une PLS sur des descripteurs à 3 dimensions comme ceux de CoMFA et CoMSIA, le nombre de degrés de liberté du modèle ( $dll_{mod}$ ) est le nombre de ses composantes ( $c$ ). Le nombre de degrés de liberté total ( $dll_{tot}$ ) est égal au nombre de composés inclus dans le modèle moins 1. Le nombre de degrés de liberté du résidu est calculé par différence. Là encore, le nombre de degrés de liberté total est la somme des degrés de liberté du modèle et des degrés de liberté du résidu.

Le test de Fisher permet de dire si deux variances sont équivalentes ou bien significativement différentes. Nous voulons savoir si la variance du modèle c'est-à-dire ce qu'il explique, est significativement différente de la variance du résidu c'est-à-dire ce qu'il n'explique pas : l'erreur. On calcule pour cela la valeur F qui est le rapport de la variance du modèle sur la variance résiduelle (Équation 8). Par analogie, on peut dire que c'est le rapport signal/bruit du modèle.

$$\text{Équation 8 : } F = \frac{\frac{SS_{\text{mod}}}{dll_{\text{mod}}}}{\frac{SS_{\text{res}}}{dll_{\text{res}}}} = \frac{\sum (y - \bar{y})^2 - \sum (y - \hat{y})^2}{\sum (y - \hat{y})^2} \times \frac{n - c - 1}{c}$$

La table de Fischer, donnée pour un risque de 5%, nous indique la valeur de F critique ( $F_{\text{crit}}$ ) en fonction du nombre de degrés de liberté du modèle et du nombre de degrés de liberté du résidu (Équation 9). Cette table est fournie en Annexe I p.156.

$$\text{Équation 9 : } F_{\text{crit}} = F(0,05, dll_{\text{mod}}, dll_{\text{res}})$$

On formule alors l'hypothèse suivante :

$H_0$  : les 2 variances sont égales.

Si  $F < F_{\text{crit}}$ , on ne peut pas réfuter l'hypothèse  $H_0$ , en d'autres termes : « on ne peut pas dire que les variances ne sont pas égales ». On interprète la réponse du test de la façon suivante : les variances sont du même ordre de grandeur.

Si  $F > F_{\text{crit}}$ , on réfute l'hypothèse  $H_0$ , ce qui se traduit par : « on ne peut pas dire que les variances sont égales ». On interprète cette réponse en disant que d'un point de vue statistique, les variances sont significativement différentes.

Un modèle sera donc significatif si F est plus grand que  $F_{\text{crit}}$ . L'erreur commise par le modèle est alors significativement plus petite que la part de l'activité observée qu'il explique.

### e . L'auto-corrélation

Le pouvoir de prévision interne  $q^2$  a malheureusement tendance à être surestimé par la procédure LOO. Une valeur élevée de cet indice peut résulter d'une corrélation due au hasard [17] ou résulter de la redondance des structures lorsque les différences entre les composés du jeu d'apprentissage sont minimales. Ce mécanisme est appelé auto-corrélation.

L'approche la plus couramment utilisée pour contrôler la robustesse d'un modèle est le test de hasardisation des réponses (*Y-randomization test*) [18, 19]. Les valeurs de la variable cible sont redistribuées de façon aléatoire sur l'ensemble du jeu d'apprentissage et un nouveau modèle est

dérivé. L'opération est répétée plusieurs fois et si la moyenne des indices  $r^2$  et  $q^2$  reste élevée, on peut en conclure qu'aucun modèle QSAR acceptable ne peut être obtenu par cette méthode statistique sur ce jeu de données.

### **f . Le pouvoir de prévision externe**

Des valeurs élevées pour les indices internes  $r^2$  et  $q^2$  sont donc nécessaires mais sont encore insuffisantes pour valider la qualité d'un modèle. Une véritable validation ne peut être obtenue que par l'évaluation d'un jeu de données externes, n'ayant pas servi à construire le modèle. On peut alors calculer l'indice de corrélation qui décrit le pouvoir de prévision externe :  $r_{pred}^2$ .

Équation 10: 
$$r_{pred}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - (PRESS / SD)$$

avec  $y_i$  et  $\hat{y}_i$  les valeurs observées et calculées par validation externe sur le jeu de test et  $\bar{y}$  la valeur moyenne de l'activité estimée sur l'ensemble des observations disponibles donc sur l'ensemble des jeux de test et d'apprentissage. PRESS et SD sont les abréviations couramment utilisées dans la littérature anglo-saxonne pour la somme des carrés résiduels et la somme des carrés totaux.

Il a été démontré sur plusieurs jeux de données qu'il n'existe aucune corrélation entre  $q^2$  et  $r_{pred}^2$  [20-22]. La validation par un jeu de test externe est donc une étape essentielle de la validation d'un QSAR.

### **g . Le biais d'ajustement, modèles centrés**

L'indice de corrélation  $r^2$  caractérise un modèle ajusté afin de produire des prévisions corrélées aux données expérimentales. Pour optimiser l'ajustement, le modèle ajoute une constante. Cette constante représente une part de l'activité qui n'est pas expliquée par les descripteurs et constitue un biais d'ajustement. Le modèle est de la forme «  $y = Ax+B$  ». Or ce que nous cherchons en définitive est un modèle capable de reproduire l'expérience plutôt que les valeurs d'un échantillon. Pour estimer ce biais on peut calculer un modèle dit centré dont l'ordonnée à l'origine est nulle c'est-à-dire un modèle sans constante. Le coefficient de corrélation d'un tel modèle est appelé coefficient de détermination, noté  $r_0^2$ . Plus le coefficient de détermination diffère du coefficient de corrélation interne, plus le modèle a de biais.

Sur le graphe de corrélation (Y calculé contre Y observé) d'un modèle centré peu biaisé, la régression linéaire du nuage de points passe près de l'origine (0,0) et a une pente A proche de 1.

Un autre moyen de s'affranchir de telles constantes additives, en particulier en cas d'écart entre

les moyennes des jeux de test et d'apprentissage, est de calculer l'indice  $p^2$ . Cet indice de pouvoir de prévision est basé sur les différences entre activité prévue et activité observée pour chaque paire du jeu de test :

$$\text{Équation 11 : } p^2 = 1 - \frac{\sum_{i,j} [(\hat{y}_i - \hat{y}_j) - (y_i - y_j)]^2}{\sum_{i,j} (y_i - y_j)^2}$$

#### **h . Les critères de validation d'un modèle**

Pour Golbraik et al [21] un modèle a un pouvoir de prévision acceptable s'il remplit les critères suivants :

- $q^2 > 0,5$
- $r^2 > 0,6$
- $\frac{(r^2 - r_0^2)}{r^2} < 0,1$
- $0,85 < A < 1,15$
- une probabilité de corrélation due au hasard très faible.

#### **i . L'interprétation du $q^2$**

Une valeur de  $q^2$  supérieure à 0,5 est un indicateur favorable du pouvoir de prévision d'un modèle. En revanche lorsqu'il est inférieur voire négatif, le modèle n'est pas prédictif. Les raisons d'un mauvais score sont multiples mais la première cause possible à envisager dans le cadre d'un QSAR3D de type CoMFA ou CoMSIA, est la présence dans le jeu d'apprentissage d'une ou de plusieurs structures mal alignées. En visualisant le graphe de corrélation, on remarquera ces composés très éloignés de la diagonale. Dans cette hypothèse, il faudra revoir l'alignement des structures.

Une autre raison expliquant la mauvaise prévision de l'activité d'une structure peut être sa trop grande dissemblance par rapport au reste des structures du jeu d'apprentissage. Le modèle aura le plus grand mal à prévoir l'activité d'un composé ayant des valeurs de descripteurs peu ou pas représentées dans le jeu d'apprentissage. Ce composé, s'il fait parti du jeu d'apprentissage, pourra en être écarté. S'il fait parti d'un jeu de test ou d'un jeu de nouvelles structures à évaluer, la prévision sera peu fiable car la structure du composé est hors de l'espace chimique couvert par l'analyse.

Une mauvaise prévision d'activité peut encore venir d'une mesure erronée de la valeur d'activité observable.

## 5 - CONSTITUTION DES JEUX D'APPRENTISSAGE ET DE TEST

Le modélisateur dispose généralement d'un seul ensemble de structures avec les activités biologiques associées. Cet ensemble doit donc être divisé en un jeu d'apprentissage et un jeu de test afin d'évaluer le pouvoir de prévision du modèle. Les contraintes essentielles pour le jeu de test sont :

- qu'il compte au moins 5 composés ;
- que ces composés couvrent la gamme des structures et des activités du jeu d'apprentissage [22, 23] ;
- que chaque composé du jeu d'apprentissage soit proche d'un composé du jeu de test [21].

La méthode la plus simple consiste à faire appel au hasard en prélevant un nombre déterminé de composés de façon aléatoire dans l'ensemble de départ. Les composés restant constituent le jeu d'apprentissage et l'opération est répétée pour obtenir différents couples de jeux de tests et d'apprentissage [24].

On peut améliorer la représentativité du jeu de test par une sélection plus rationnelle. Les méthodes de partage sont nombreuses, trois sont présentées ici.

Après avoir trié les composés par activité, on divise l'ensemble en classes d'intervalles d'activité constants. On sélectionne un certain nombre d'éléments, au hasard ou par rang, dans chaque classe pour constituer le jeu de test ; on prend le dernier de chaque classe par exemple. Les composés restants forment le jeu d'apprentissage. Le jeu de test ainsi obtenu sera plus représentatif de la gamme d'activité du jeu d'apprentissage. Le nombre optimal de classes dépend du nombre d'éléments dans l'ensemble. On peut en faire une approximation par la racine carrée du nombre total d'éléments. Un trop petit nombre de classes conduit à une perte d'informations alors qu'un trop grand nombre de classes conduit à des classes vides ou sous peuplées, ce qui génère du bruit et de l'incohérence.

On peut aussi constituer des classes d'effectif constant et décider d'un rapport de taille entre le futur jeu d'apprentissage et le futur jeu de test. Les meilleurs composés de chaque classe vont dans le jeu d'apprentissage et les autres dans le jeu de test, en respectant le rapport fixé. En faisant varier ce rapport et le nombre de classes on obtient différents jeux de test et d'apprentissage.

Une sélection plus rationnelle encore utilise la méthode des sphères d'exclusions basée sur le nombre de descripteurs et des niveaux de dissimilarité [23]. Pour chaque couple de molécules, un indice de similarité est calculé à partir des descripteurs. Si cet indice est supérieur à une valeur

seuil, préalablement définie, alors la molécule la moins active des deux est exclue du jeu d'apprentissage et va dans le jeu de test. Cette méthode permet de construire des jeux de test plus représentatifs de l'espace chimique du jeu d'apprentissage. Chaque molécule du jeu de test ressemble à une molécule du jeu d'apprentissage.

Ces différentes méthodes ont été comparées dans leur capacité à produire de bons jeux de tests et toutes présentent des avantages et des inconvénients. La méthode par exclusion de sphères produit cependant des jeux plus fiables et les plus représentatifs alors que l'introduction du hasard diminue cette représentativité [21].

## **6 - STRATEGIE GLOBALE D'UNE ETUDE QSPR**

La stratégie de développement d'un QSAR et plus généralement d'un QSPR va donc s'articuler autour de 5 points :

- Constituer une base de données Structure - Propriété à partir de mesures quantitatives, fiables et normalisées de la propriété cible, pour chaque composé. Sélectionner des descripteurs chimiques en relation avec la propriété cible.
- Diviser ce jeu de données en un jeu d'apprentissage et un jeu de test ;
- Construire des modèles à partir de jeu d'apprentissage avec les outils statistiques souhaités. Caractériser ces modèles par leurs indices de validation internes et vérifier leur robustesse par un test de hasardisation ;
- Valider les modèles avec le jeu de test et calculer leur indice de corrélation externe. Répéter l'opération de division pour obtenir d'autres jeux d'apprentissage et de test. La division optimale donne le plus petit jeu d'apprentissage capable de bonnes prévisions pour le plus grand jeu de test ;
- Explorer et exploiter les modèles validés pour comprendre les mécanismes possibles et faire des prévisions.

L'ensemble de la procédure de validation est représenté sur la Figure 1.

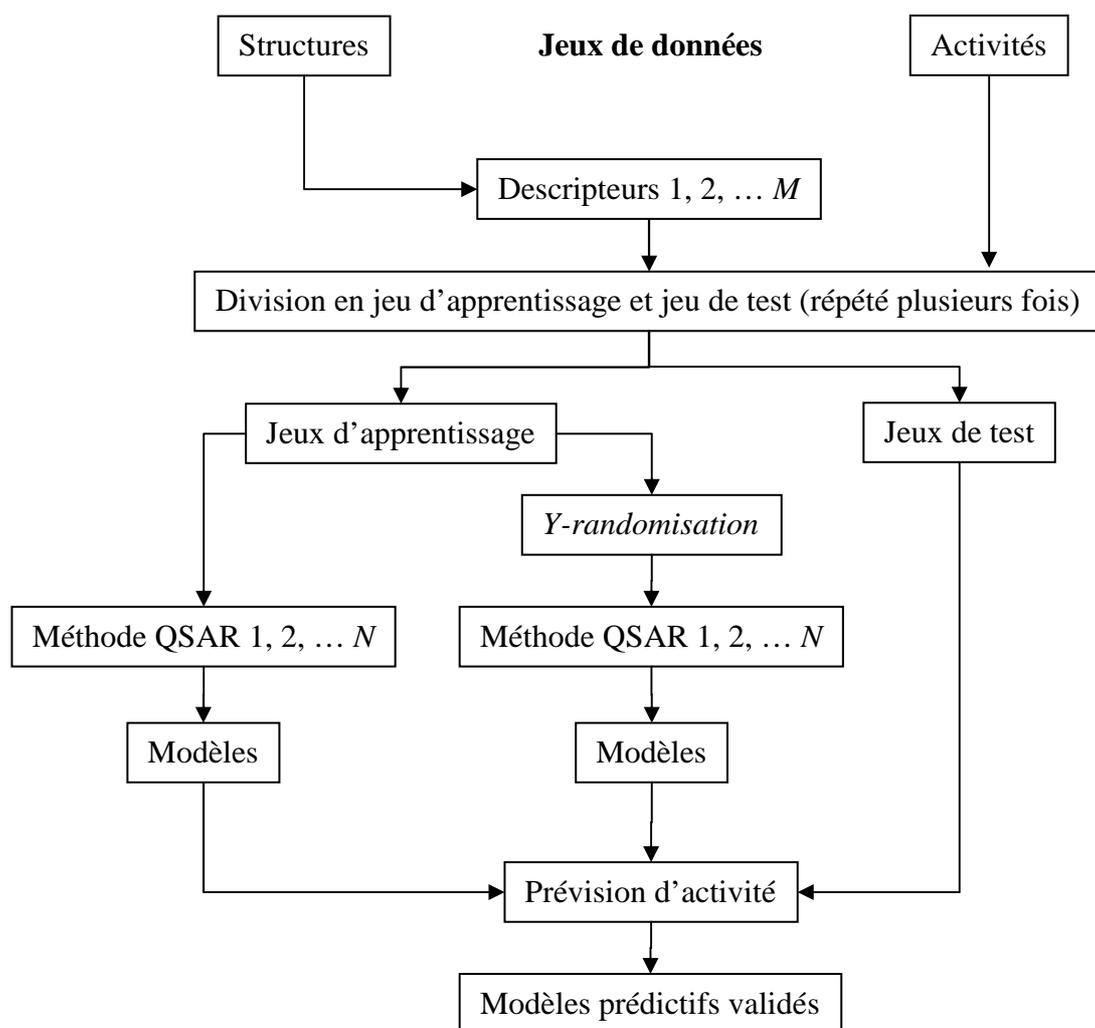


Figure 1 : Procédure de validation des modèles [25].

## ***II - ADAPTATION DES OUTILS STATISTIQUES AUX DONNEES PHARMACO-CHIMIQUES : QSAR-3D BASES SUR DES CHAMPS D'INTERACTIONS OU D'INDICES DE SIMILARITE MOLECULAIRE.***

Les QSAR classiques reposent sur des descripteurs 2D qui négligent par nature un grand nombre d'informations stéréochimiques. A la fin des années 80, le QSAR a franchi un seuil décisif tant du point de vue théorique que pratique avec l'apparition des techniques CoMFA [6] et GRID [5], puis CoMSIA[7]. Ces techniques sont basées sur des descripteurs des structures tridimensionnelles, ce qui règle certaines des déficiences majeures inhérentes aux techniques classiques. Pour être plus exact, ce ne sont plus les structures elles-mêmes qui sont décrites mais les différents potentiels d'interactions qu'elles génèrent et qui les entourent. Ces potentiels sont directement liés à la géométrie dans l'espace des molécules et à la nature des atomes qui les

composent. Les potentiels sont décrits par leur position et leur expansion dans l'espace ainsi que par leur intensité. Les descripteurs employés portant sur des objets tridimensionnels, ce type d'étude QSAR est couramment appelé QSAR-3D.

L'analyse statistique vise ensuite à établir les corrélations entre les variations de champs et les variations d'activité entre les composés.

Ces techniques peuvent donner des modèles de très grande qualité et ont donc été rapidement et très largement utilisées. Elles souffrent cependant d'un inconvénient important qui est la nécessité d'un alignement préalable des structures dans leur conformation bioactive. Or cette conformation bioactive est généralement inconnue pour la majorité des molécules.

Les techniques CoMFA et CoMSIA sont des composantes du module QSAR du logiciel SYBYL.

## **1 - COMPARATIVE MOLECULAR FIELD ANALYSIS (COMFA)**

L'idée fondatrice de CoMFA est que les différences d'activité entre les molécules d'une série sont liées à des différences de forme des champs d'interaction moléculaire (Molecular Interaction Field - MIF) non covalents. La méthode CoMFA analyse les champs d'interactions stérique et électrostatique.

La forme des champs d'interaction est décrite par l'échantillonnage, à intervalle régulier, de leur magnitude, dans l'espace environnant les molécules. Cet échantillonnage produit donc une matrice tridimensionnelle. Une fois décrits pour chaque molécule, ces champs sont comparés par une analyse statistique du type PLS. Cette analyse permet d'établir les corrélations entre les variations de position et d'intensité des potentiels et les variations d'activité associées. On accède ainsi aux tendances qui sont favorables ou défavorables à la propriété cible.

Pour comparer ces objets 3D, il est cependant nécessaire de les aligner dans un référentiel commun ce qui implique un alignement des molécules.

D'une façon indirecte, le modèle décrit aussi le site actif de la cible mais ceci n'est vrai que si le modèle est construit avec la position relative et surtout la conformation active des composés. L'alignement correct des molécules dans leur position relative et dans leur conformation active est donc un pré-requis indispensable à la réalisation de modèles CoMFA. Malheureusement, cette information est difficile à obtenir et fait souvent défaut. Il en découle une multiplication des hypothèses qui rendent cette étape du choix des conformères et de leur alignement, l'étape la plus longue et la plus délicate de ce type de QSAR.

La construction des modèles et leur validation suivent ensuite les principes généraux des QSPR énoncés précédemment.

## **2 - COMPARATIVE MOLECULAR SIMILARITY INDICES ANALYSIS (COMSIA)**

La méthode CoMSIA est une extension de CoMFA qui utilise, en plus des champs d'interaction stérique et électrostatique, un champ d'interaction lipophile, un champ « accepteur de liaison hydrogène » et un champ « donneur de liaison hydrogène ». Les deux techniques diffèrent également dans la manière dont sont implantés les champs d'interaction moléculaires. Elles donnent généralement des résultats comparables mais les modèles CoMSIA sont souvent plus riches et plus faciles à interpréter.

## **3 - ALIGNEMENT DES STRUCTURES**

Pour réaliser un QSAR3D de type CoMFA ou CoMSIA, il est d'abord nécessaire d'aligner les structures dans leur conformation active. La situation idéale est donc celle où l'on dispose de la structure du récepteur co-cristallisé avec un ou plusieurs ligands de structures proches de la série étudiée. Le ligand co-cristallisé sert de référence sur laquelle sont alignées les structures de la série. Disposer de la structure du récepteur permet également de réaliser un docking des composés dont le résultat pourra constituer un alignement vraisemblable.

Malheureusement, dans la majorité des cas, la structure du récepteur n'est pas disponible. L'alignement devra donc se faire sur la seule base des structures de la série et reposera en grande partie sur l'expérience du modélisateur. Des méthodes rationnelles et automatisées ont été développées mais elles restent difficiles d'accès :

- soit l'exécution des algorithmes fait appel à différents programmes et scripts indépendants, la séquence des opérations est peu ou pas automatisée et de nombreuses interventions de l'utilisateur sont nécessaires ;
- soit l'ensemble de la procédure est intégrée dans un seul et même logiciel mais celui-ci est généralement commercial.

Les outils d'alignement se répartissent selon deux approches :

- l'alignement basé sur des points topologiques ;
- l'alignement basé sur des propriétés ;

Dans l'alignement par points, des paires d'atomes ou de pharmacophores sont alignées par une

méthode d'ajustement par les moindres carrés. L'inconvénient majeur de cette technique est qu'elle implique une définition préalable des points d'ancrage pour chaque structure. Dans le cas de structures très différentes, la détection automatique de ces points reste problématique malgré la multitude d'algorithmes proposés.

Les algorithmes basés sur des propriétés offrent un plus large choix de descripteurs qui incluent des propriétés moléculaires variées telles que la forme et le volume, la densité électronique ou la distribution de charges. Cependant, toutes ces techniques ont besoin d'un référentiel, généralement appelé indice de similarité qui décrit le degré de recouvrement entre les structures superposées. La superposition consiste alors à optimiser cet indice par différentes méthodes mathématiques.

Nombre de ces méthodes ont été passées en revue par Lemmen et Lengauer [26] et des contributions plus récentes [27-29] complètent l'inventaire.

Reste la méthode de base : l'alignement manuel. Les structures sont alignées sur une structure de référence qui est généralement le conformère le plus stable du composé le plus actif et le plus rigide de la série. Il est cependant rare qu'un composé combinant toutes ces qualités soit disponible.

Quelle que soit la méthode utilisée, l'expérience et l'intuition du modélisateur influent largement sur la qualité du résultat.

#### **4 - CALCUL DES CHAMPS D'INTERACTION MOLECULAIRES**

Pour CoMFA comme pour CoMSIA, les champs d'interaction moléculaire sont calculés par l'évaluation du potentiel d'interaction entre une sonde et la molécule étudiée. Cette évaluation est faite sur des points régulièrement espacés d'une matrice englobant la molécule.

CoMFA calcule le champ stérique avec un potentiel de Lennard-Jones et le champ électrostatique avec un potentiel de Coulomb [6]. Cette approche, bien que largement acceptée et très efficace, pose quelques problèmes. En particulier, les deux fonctions de potentiels ont une pente très forte à proximité de la surface de Van der Waals de la molécule (Figure 2). Cela provoque des changements brusques dans la description des surfaces et conduit à l'emploi de valeurs seuils pour éviter de calculer les potentiels à l'intérieur de la molécule. De plus un facteur d'échelle est appliqué au champ stérique pour qu'il soit comparable et utilisable avec le champ électrostatique dans la même PLS. Pour finir, si l'on change l'orientation de la matrice de mesure par rapport au jeu de molécules alignées, on peut observer des changements significatifs dans les résultats de l'analyse [30]. Ces différences sont probablement due à l'emploi de valeurs

seuils strictes.

Dans CoMSIA, ce sont 5 champs de similarité distincts qui sont calculés. Ce sont les champs stérique, électrostatique, hydrophobe, donneur de liaison hydrogène et accepteur de liaison hydrogène. Ces champs couvrent les types majeurs d'interaction impliqués dans la liaison ligand récepteur [31]. Les indices de similarité sont calculés dans une matrice 3D comparable à celle de CoMFA. Les potentiels dépendant de la distance entre la sonde et la molécule sont modélisés par une fonction gaussienne (Figure 2). La forme de cette fonction est différente de celle des fonctions de potentiels classiques et permet de calculer les indices de similarité de tous les points de la matrice, à l'intérieur comme à l'extérieur de la surface de Van der Waals.

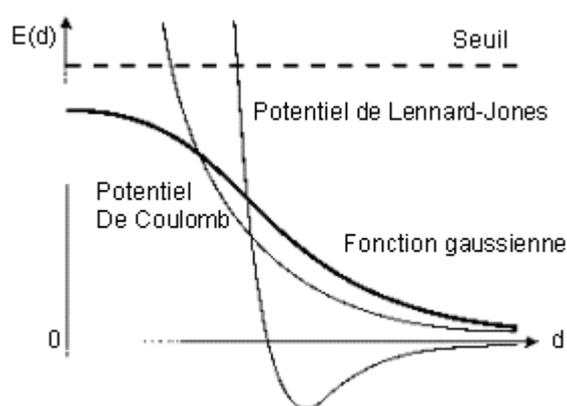


Figure 2 : Forme générale des potentiels classiques utilisés par CoMFA (traits fins) et du potentiel utilisé par CoMSIA (trait gras).

L'indice de similarité A en un point q de la matrice F est calculé par la somme des interactions de type F de tous les atomes j de la molécule i. La fonction d'indice de similarité est décrite par l'Équation 12.

Équation 12 :

$$A_{F,k}^q(j) = \sum_i w_{sonde,k} w_{ik} e^{-\alpha r_{i,q}^2}$$

$w_{sonde,k}$  est la sonde, de rayon 1Å, de charge +1, d'hydrophobicité +1, donneur d'1 liaison hydrogène et accepteur d'1 liaison hydrogène.

$w_{i,k}$  est la valeur de la propriété physico-chimique k de l'atome i.

$r_{i,q}$  est la distance entre la sonde au point q et l'atome i de la molécule.

$\alpha$  est le facteur d'atténuation dont la valeur par défaut est de 0,3 mais qui peut varier de 0,2 à 0,4 [32].

Les valeurs des propriétés physico-chimiques proviennent de différentes sources.

Pour le champ stérique, CoMSIA utilise la table de Van der Waals implémentée dans SYBYL. Cette table est codée et n'est pas modifiable.

Pour le champ électrostatique, les charges partielles atomiques sont directement utilisées.

Pour le champ hydrophobe, CoMSIA utilise une table d'hydrophobicité dont les valeurs sont basées sur les travaux de Viswanadhan *et al.* [33].

Pour les champs donneur et accepteur de liaison hydrogène, CoMSIA crée des atomes fantômes (*dummy atoms*) sur les sites donneurs et accepteurs. Les mêmes atomes fantômes sont utilisés pour les deux champs. Le champ accepteur contient les informations sur les positions de possibles groupements donneurs sur le récepteur. Le champ donneur indique les régions où des groupements accepteurs du récepteur pourraient se trouver.

## **5 - SEQUENCE ET RESULTATS DES ANALYSES COMFA ET COMSIA**

### **a . Construction d'une table de molécules**

Les molécules de la série à analyser sont construites et optimisées avec leurs charges partielles puis alignées selon les techniques adéquates. Les structures sont enregistrées dans une base de données de SYBYL.

Les analyses QSAR sont réalisées sur une table de molécules (*Molecular Spreadsheet*<sup>TM</sup> – MSS) construite à partir de la base de données. Sur cette table, chaque ligne correspond à une molécule de la base et chaque colonne peut contenir un descripteur ou un indicateur.

Une de ces colonnes doit contenir la valeur cible à exploiter. Les valeurs sont saisies à la main ou importées selon la procédure proposée par le MSS.

### **b . Calcul des descripteurs**

Le MSS permet ensuite de calculer les différents descripteurs CoMFA et CoMSIA de son choix. Lors du calcul d'un premier champ d'interaction, le logiciel crée une matrice dont la position, l'orientation et les dimensions sont automatiquement fixées de façon à englober largement toutes les molécules de la série. Pour le calcul des autres champs, il est souhaitable d'utiliser la même matrice pour obtenir une série de champs homogènes et comparables. Cette méthodologie est proposée par le logiciel mais n'est pas automatique. Il est possible de modifier l'espacement entre les points de la matrice et le facteur d'atténuation  $\alpha$ , grâce à des variables d'environnement du logiciel (*Tailor Values*). Chaque descripteur calculé occupe une colonne et la valeur affichée

est l'écart-type du champs. Une fois les descripteurs calculés, la table est prête pour l'analyse PLS.

Afin de limiter le bruit généré par les faibles variations de champ, il est possible de filtrer les valeurs des descripteurs (*Column Filtering*). La PLS ne tiendra pas compte des variations inférieures à un certain seuil. Ce seuil est défini par l'utilisateur avant de lancer un calcul et est généralement fixé à  $1,0 \text{ kcal.mol}^{-1}$ . Le filtrage permet également de raccourcir considérablement les temps de calcul.

### **c . Séquences des analyses PLS**

La première analyse à réaliser est une PLS avec validation croisée par la procédure LOO et un nombre de composantes allant de 1 à  $C_{\max}$ , le nombre maximal de composantes. Pour que le modèle puisse être significatif, il est souhaitable que  $C_{\max}$  soit inférieur ou égal au tiers du nombre de molécules dans le jeu d'apprentissage. Cette première analyse nous donne le  $q^2$  du modèle et son erreur standard de prévision ainsi que le nombre optimal de composantes  $C_{\text{opt}}$ .

On dérive alors un second modèle, sans validation croisée mais avec le nombre optimal de composantes indiqué par l'analyse précédente. On obtient l'indice de corrélation  $r^2$ , l'erreur standard d'estimation et la valeur F avec les degrés de liberté du modèle et du résidu. Le poids relatif de chaque descripteur utilisé dans le modèle est également donné à chaque analyse.

On peut ensuite répéter cette séquence en calculant les modèles centrés afin d'estimer le biais d'ajustement des modèles non centrés.

Il est possible que certains descripteurs soient corrélés entre eux et donc redondants. L'emploi simultané de ces descripteurs risque de générer plus de bruit que de signal. Il se peut aussi que certains descripteurs ne soient pas corrélés à la propriété cible. Leur intégration dans le modèle ne générera que du bruit et alourdira inutilement les calculs. Evidemment, il est difficile de prévoir quelle est la meilleure association de descripteurs. Il faudrait les tester toutes pour le savoir ...

Nous avons donc écrit un script, dans le langage de programmation de SYBYL (*Sybyl Programming Language – SPL*) qui permet de dériver un modèle pour toutes les combinaisons de descripteurs et pour tous les jeux d'apprentissage souhaités. Il collecte également les indices descriptifs de chaque modèle dans un fichier de sortie. Ce script a été nommé QBF1.spl (*Qsar Best Fit*). L'exécution de ce script doit être paramétrée par l'utilisateur. Son fonctionnement est décrit dans la partie expérimentale et son code est donné dans l'Annexe II p.157. Un exemple de fichier de sortie est fourni dans l'Annexe III p.161.

Les fichiers « pls » sont codés en binaires et ne sont lisibles que par l'intermédiaire du logiciel. Ils permettent d'accéder à tous les paramètres de la PLS et au modèle calculé. C'est à partir de ce fichier que l'on pourra visualiser un modèle et l'utiliser pour faire des prévisions d'activité.

## **6 - VISUALISATION GRAPHIQUE DES MODELES**

Le premier intérêt des modèles CoMFA ou CoMSIA est qu'ils permettent de construire des graphiques de visualisation spatiale des relations structure-activité. On a ainsi un accès direct et intuitif aux explications du modèle et ce à l'échelle moléculaire. Ces graphiques sont basés sur le produit de la variance des points d'un champ par leur coefficient dans le modèle ( $SD \times \text{Coefficient}$ ). Ils représentent des surfaces de contour passant par les points de même valeur. Pour que ces contours soient comparables d'un champ à l'autre, les valeurs sont préalablement transposées sur une échelle de 0 à 100. Les contours sont des pourcentiles de la gamme de valeurs présente dans le champ. L'utilisateur indique au logiciel quels pourcentiles il souhaite visualiser. Les niveaux 20% et 80% sont généralement utilisés pour représenter respectivement les régions du champ défavorables et favorables à l'activité. Ces graphes de contour indiquent à quel endroit les variations de champ expliquent les variations d'activité.

## **7 - PREVISION ET EXTRAPOLATION**

Le second intérêt des modèles CoMFA et CoMSIA est qu'ils permettent de faire des prévisions pour de nouvelles structures. Cependant, un modèle performant ne l'est que dans le domaine qu'il couvre. Les prévisions d'activité sont plus fiables pour les composés dont les valeurs des descripteurs sont similaires à celles du jeu d'apprentissage. Le logiciel compare les valeurs des descripteurs des composés estimés avec la gamme de valeurs des composés du jeu d'apprentissage. Il indique le nombre de ces valeurs « hors-jeu » qu'il a dû extrapoler ainsi que la contribution de ces points extrapolés dans la prévision. Ces deux indices nous donnent une estimation du degré de similarité des composés évalués par rapport au jeu d'apprentissage.

## C - APPLICATION : QSAR-3D AVEC LES DESCRIPTEURS COMSIA D'UNE SERIE DE FLAVONOIDES ET DE BOERHAVINONES MODULATEURS DU TRANSPORTEUR BCRP.

### I - INTRODUCTION

La principale raison de l'échec des chimiothérapies anticancéreuses est due au phénomène de résistance multiple des cellules cancéreuses (MDR – *Multi-Drug Resistance*). Ces cellules ont ou acquièrent un système d'évacuation des agents thérapeutiques par transport actif. Ce mécanisme de détoxification est présent dans toutes les cellules mais est souvent surexprimé par les cellules tumorales. Les responsables de ce phénomène sont des protéines appartenant à la superfamille des transporteurs ABC (ATP Binding Casette). Les plus connues sont :

- la Glycoprotéine P (PgP ou MDR1 ou ABCB1) [34]
- la *Multidrug Resistance Protein* (MRP1 ou ABCC1) [35]
- la *Breast Cancer Resistance Protein* (BCRP, ABCG2, MXR ou ABCP) [36-38].

Ces transporteurs sont localisés au niveau de la membrane cytoplasmique de cellules. Ils évacuent de manière active une multitude de composés, grâce à l'énergie issue de l'hydrolyse de l'ATP.

BCRP est la plus récemment découverte et est considérée comme un demi transporteur ABC. Elle ne possède que 6 hélices transmembranaires et un site de fixation de l'ATP alors que le reste de la famille compte 12 domaines transmembranaires et deux sites de fixation de l'ATP. Pour être active, elle doit cependant être sous forme d'homodimère [39] mais elle a également été détectée sous forme de tétramère et sous des degrés d'oligomérisation plus élevés encore [40].

Il s'avère que de nombreux principes actifs de première importance sont les substrats de BCRP. On recense notamment le flavopiridol, le méthotrexate (MTX), la mitoxantrone (MX), des inhibiteurs de la transcriptase inverse du VIH et de la topoisomérase 1 comme les anthracyclines [41]. BCRP a été détectée dans de nombreuses tumeurs humaines [42-45] et doit contribuer, par un mécanisme propre, au phénomène de résistance multiple [46, 47]. De nombreux substrats sont communs à BCRP et à PgP, mais ce n'est pas le cas de leurs inhibiteurs, ce qui explique l'échec des premières tentatives de modulation de la MDR par les seuls inhibiteurs de PgP [48].

BCRP est largement répandu dans l'organisme. On la trouve dans les organes et les tissus excréteurs comme les membranes des canaux hépatiques, mais également au niveau des barrières physiologiques, à la surface des muqueuses de la lumière intestinale, au niveau de la barrière hémato-encéphalique ou du placenta [49-51]. Ce profil de distribution indique que BCRP doit jouer un rôle important dans la perméabilité des barrières pharmacocinétiques majeures [41]. Ce rôle a été mis en évidence lors de la co-administration de GF120918, un inhibiteur de BCRP, et du topotécan, un substrat de BCRP. On a constaté notamment une augmentation de la biodisponibilité du topotécan et une diminution de son excrétion par la bile, à la fois chez la souris et chez l'homme [52, 53]. Par conséquent, des inhibiteurs efficaces de BCRP présentent un grand intérêt comme agent améliorant les capacités pharmacocinétiques de principes actifs substrats de BCRP, en augmentant leur biodisponibilité orale et plasmatique ou en augmentant leur pénétration dans les régions cérébrale et foétale.

Parmi les inhibiteurs connus des transporteurs ABC, se trouvent de nombreux composés naturels ainsi que des analogues de synthèse. On compte notamment de nombreux inhibiteurs appartenant à la classe des flavonoïdes.

Les flavonoïdes sont des composés naturels formant une vaste classe de molécules polyphénoliques. Ils sont très largement répandus dans le règne végétal si bien qu'on les retrouve en abondance dans notre alimentation, dans les légumes, les fruits et autres préparations à base de plantes. Plus de 6500 flavonoïdes ont été décrits et leur consommation moyenne en occident est estimée entre 200 et 1000 mg par jour et par personne [54]. Les flavonoïdes joueraient un rôle bénéfique dans la prévention de cancers, de maladies cardio-vasculaires, de l'ostéoporose et de nombreuses autres maladies liées à l'âge [55-59]. Ils auraient également des propriétés antivirales et anti-inflammatoires. D'autre part, ces composés ont généralement une faible toxicité [58-61]. Leur usage dans le cadre de « médecines douces » ou comme compléments diététiques est en pleine expansion sans qu'aucun contrôle des autorités sanitaires ne s'exerce. Les interactions potentielles avec des traitements médicamenteux conventionnels ne sont pas bien évaluées si bien que leur usage n'est pas tout à fait sans risque. Certaines interactions pharmacocinétiques, parfois sévères, ont d'ailleurs été observées chez l'animal et chez l'homme [62-64]. Mieux connaître les interactions des flavonoïdes avec les mécanismes de biodistribution des principes actifs, en particulier avec les transporteurs membranaires et les enzymes du métabolisme, ainsi que les conséquences cliniques de ces interactions renforcent l'intérêt des chercheurs pour ces molécules.

Notre laboratoire contribue, parmi d'autres, à l'étude des flavonoïdes comme modulateurs de la

résistance aux anticancéreux. Il sont évalués en tant qu'inhibiteurs de PgP [65-69], de MRP [48, 70-72] et depuis peu de BCRP.

La définition de relations structure-activité entre les flavonoïdes ou leurs dérivés et l'inhibition de l'efflux de substrats par BCRP a débuté avec la publication des premiers modèles non quantitatifs [73] et quantitatifs [74]. Notre objectif dans cette étude est d'établir un modèle 3D quantitatif des relations structure-activité de l'inhibition du type sauvage de BCRP par une série de flavonoïdes.

## ***II - DONNEES STRUCTURALES ET BIOLOGIQUES***

Notre série est constituée de 18 flavonoïdes synthétisés au laboratoire et de 9 dérivés de flavonoïdes, appelés boerhavinones.

Les données d'activité biologique nous ont été fournies par le laboratoire des Protéines de Résistance aux Agents Chimiothérapeutiques (PRAC) de l'Institut de Biologie et de Chimie des Protéines (IBCP) de Lyon.

Le test biologique évalue l'effet de ces composés sur des cellules cancéreuses sauvages humaines surexprimant BCRP (lignée R482). Il mesure l'efflux d'un substrat fluorescent de BCRP, Hoeschst33342, en présence d'un inhibiteur. Les mesures sont exprimées en pourcentage d'inhibition (%I) avec 10  $\mu$ M d'inhibiteur. Le Tableau 1 regroupe les structures et les activités des composés étudiés.

Les données brutes, ne suivent pas exactement une loi de distribution normale. On peut s'en rendre compte par un graphe de distribution comme celui de la Figure 3. Sur ce graphe, la gamme d'activité est divisée en 7 classes de largeurs égales sur la gamme de 0 à 100%. La distribution des données brutes est décrite par la courbe violette. Par rapport à une distribution Normale théorique, la classe numéro 2 est surpeuplée, la troisième est vide et les deux dernières sont également sur-représentées

Le test de Box-Cox permet de déterminer quelle transformation est nécessaire pour normaliser au mieux la distribution des données. Nous avons effectué ce test sur nos données avec le logiciel mROC [75]. Il nous donne une valeur de  $\lambda = 1,25$  avec l'intervalle de confiance [0,83 ; 1,78]. Le paramètre  $\lambda$  est compris entre 1 et 2 avec la valeur 1 comprise dans l'intervalle de confiance. Ceci indique qu'il n'est pas indispensable de transformer cette variable. Nous pouvons utiliser les données brutes ou appliquer une transformation de Box-Cox avec le paramètre  $\lambda = 1,25$ . Si on applique cette transformation, la distribution des valeurs (courbe jaune) a la même allure que

celle des données brutes et accentue même le déséquilibre de la distribution dans les trois premières classes.

Notons que la valeur 0 (zéro) n'est pas comprise dans l'intervalle de confiance de  $\lambda$ . Ceci indique qu'une transformation par un logarithme ne normalisera pas la distribution. Effectivement, une transformation logarithmique déséquilibre complètement la distribution comme le montre la courbe cyan de la Figure 3. La moitié des données se trouve alors dans la dernière classe.

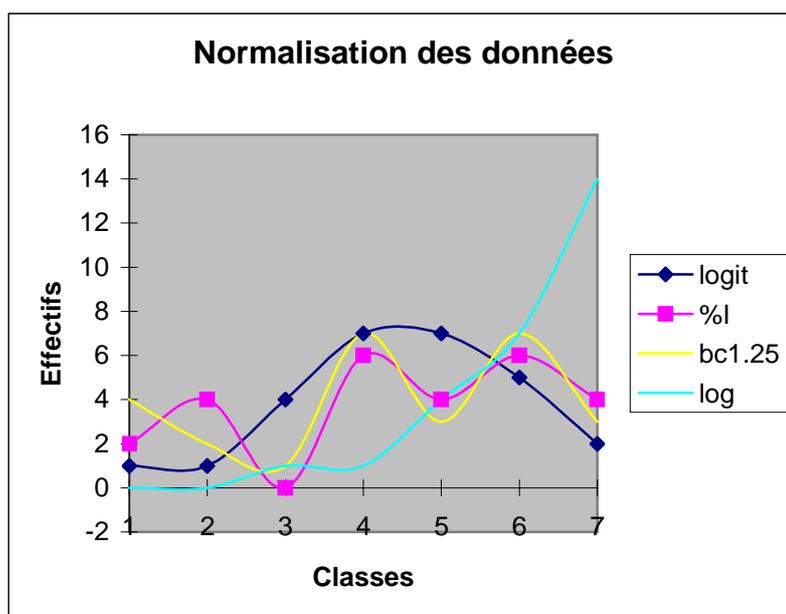
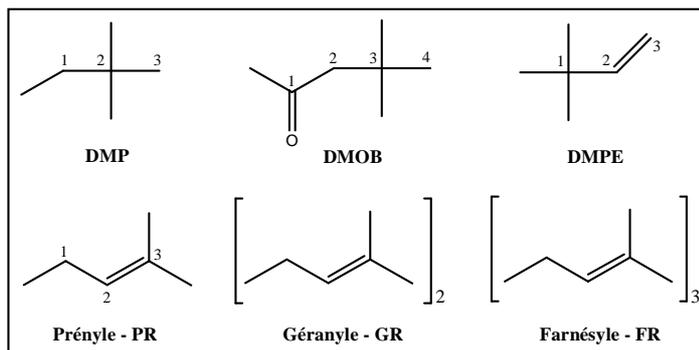
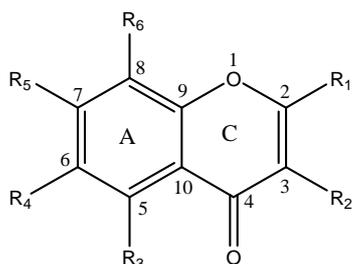


Figure 3 : Influence de différentes transformation sur la distribution des données.

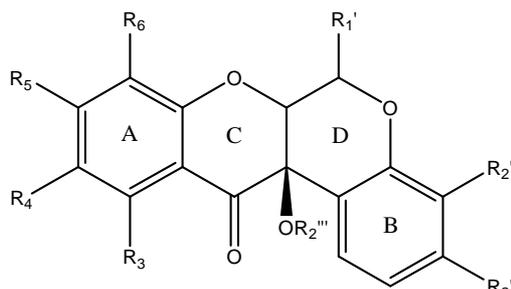
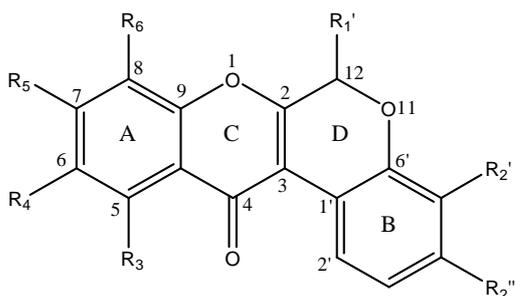
La nature du test biologique pose une contrainte supplémentaire. Le pourcentage d'inhibition est une variable bornée. Sa valeur ne peut pas dépasser 100 ou être inférieure à 0. Tous les composés sont testés à la concentration de 10  $\mu\text{M}$  donc tous ceux qui ont un fort pouvoir inhibiteur à cette concentration auront un pourcentage d'inhibition proche ou égal à 100%. On aura donc une très faible variation d'activité entre ces composés actifs alors que grandes différences structurales pourront exister entre eux. Il en va de même pour les composés peu actifs dont l'activité observée est proche de 0 (zéro). Une transformation Logit des données brutes est donc indiquée afin de s'affranchir de ces bornes et de « ré-étaler » les valeurs proches des limites. Sur notre jeu de données, cette transformation présente l'avantage subsidiaire de normaliser la distribution (courbe bleu foncé Figure 3).

Tableau 1 : Structure et activité des composés étudiés.

**Flavonoïdes**



**Boerhavinones**



| Flavonoïdes | R1  | R2   | R3 | R4   | R5   | R6   | %I | LOGIT(%I) |
|-------------|-----|------|----|------|------|------|----|-----------|
| 1           | phy | H    | H  | H    | H    | H    | 80 | 0.602     |
| 2           | phy | OH   | H  | H    | H    | H    | 55 | 0.087     |
| 3           | phy | H    | H  | H    | OH   | H    | 58 | 0.140     |
| 4           | DMP | DMOP | OH | H    | H    | H    | 90 | 0.954     |
| 5           | phy | H    | OH | H    | OH   | H    | 70 | 0.368     |
| 6           | phy | H    | OH | H    | OMe  | H    | 70 | 0.368     |
| 7           | phy | H    | OH | H    | OEt  | H    | 65 | 0.269     |
| 8           | phy | H    | OH | H    | OPr  | H    | 61 | 0.194     |
| 9           | phy | H    | OH | H    | OiPr | H    | 58 | 0.140     |
| 10          | phy | H    | OH | H    | Obut | H    | 50 | 0.000     |
| 11          | phy | H    | OH | H    | OMe  | OMe  | 75 | 0.477     |
| 12          | phy | H    | OH | PR   | OH   | H    | 95 | 1.279     |
| 13          | phy | H    | OH | DMPE | OH   | H    | 81 | 0.630     |
| 14          | phy | H    | OH | H    | OH   | PR   | 88 | 0.865     |
| 15          | phy | H    | OH | H    | OH   | DMPE | 83 | 0.689     |
| 16          | phy | H    | OH | GR   | OH   | H    | 86 | 0.788     |
| 17          | phy | H    | OH | GR   | OH   | GR   | 80 | 0.602     |
| 18          | phy | H    | OH | FR   | OH   | H    | 25 | -0.477    |

| Boerhavinones | R1' | R2' | R2'' | R2''' | R3 | R4  | R5  | R6 | %I | LOGIT(%I) |
|---------------|-----|-----|------|-------|----|-----|-----|----|----|-----------|
| 19            | OH  | H   | H    | -     | OH | H   | OH  | H  | 29 | -0.389    |
| 20            | OMe | OH  | H    | -     | OH | H   | OMe | H  | 94 | 1.195     |
| 21            | OH  | H   | H    | -     | OH | Me  | OH  | H  | 27 | -0.432    |
| 22            | OMe | H   | H    | -     | OH | Me  | OH  | H  | 55 | 0.087     |
| 23            | H   | H   | H    | OH    | OH | OMe | OMe | H  | 15 | -0.753    |
| 24            | H   | OH  | H    | OH    | OH | Me  | OMe | H  | 31 | -0.347    |
| 25            | OMe | OH  | H    | -     | OH | Me  | OMe | H  | 86 | 0.788     |
| 26            | OH  | H   | H    | -     | OH | Me  | OMe | OH | 2  | -1.690    |
| 27            | OH  | H   | OH   | -     | OH | Me  | OH  | H  | 56 | 0.105     |

### III - CONFORMATIONS ET ALIGNEMENT

Cette série comporte des molécules très rigides et à l'exception de quelques chaînes aliphatiques, ces structures ne présentent pas de complexité conformationnelle importante. Tous les composés ont en commun un noyau chromone (benzopyran-4-one) qui est planaire (Figure 4). La recherche des conformations les plus stables pour les chaînes flexibles a été réalisée avec un algorithme génétique d'analyse conformationnelle (GA Conf. Search). La méthode semi empirique AM1 de MOPAC (QCPE #455) a ensuite été appliquée pour optimiser les géométries et calculer les charges partielles.

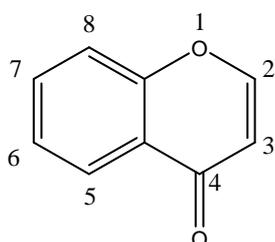


Figure 4 : Structure du noyau Chromone commun à tous nos composés.

L'alignement des structures optimisée a été réalisé par simple superposition des 11 atomes du noyau chromone. Le résultat de l'alignement est présenté sur la Figure 5.

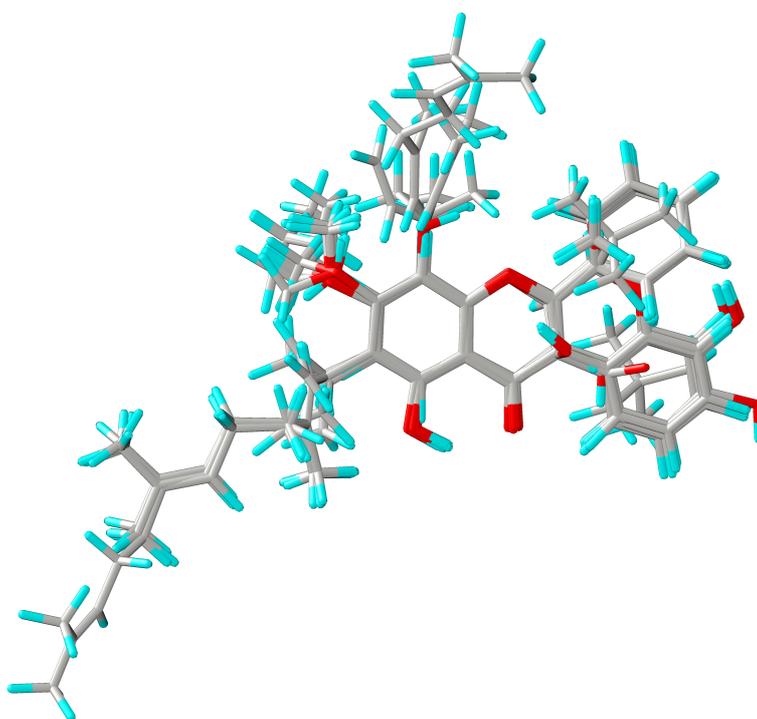


Figure 5 : Alignement des 27 composés utilisés pour l'analyse CoMSIA.

## ***IV - CONSTRUCTION ET VALIDATIONS DES MODELES***

### **1 - MODELES COMSIA DU JEU DE DONNEES COMPLET.**

#### **a . Construction des modèles :**

Les modèles 3D-QSAR ont été réalisés par une analyse PLS des descripteurs de type CoMSIA. Les descripteurs utilisés sont les champs d'interaction stérique (S), électrostatique (E), hydrophobe (H), donneur de liaison hydrogène (D) et accepteur de liaison hydrogène (A). Les 31 combinaisons possibles de ces 5 descripteurs ont été étudiées et les modèles ont été calculés avec le pourcentage d'inhibition transformé par logit comme variable cible. Pour chaque combinaison, deux modèles ont été dérivés : l'un avec constante, l'autre sans (modèle centré). L'ensemble de ces calculs ont été exécuté grâce au script QBF1.

Tous les modèles obtenus sont décrits par le nombre de leurs composantes ( $n_c$ ), leur coefficient de corrélation interne ( $r^2$ ), l'erreur standard d'estimation associée ( $see$ ) ainsi que l'ordonnée à l'origine ( $y_0$ ) du modèle non centré et la valeur F du test de Fischer. Le pouvoir de prévision interne ( $q^2_{cv}$ ) des modèles et l'erreur standard de prévision ( $sep$ ) sont établis par le test de validation croisée en procédure « leave-one-out » (LOO). Le biais introduit par la constante des modèles pour optimiser l'ajustement aux données est évalué par le rapport  $(r^2 - r_0^2)/r^2$  où  $r_0^2$  est le coefficient de corrélation interne du modèle centré (coefficient de détermination). L'ensemble des résultats est présenté dans le Tableau 2.

La table de Fischer pour un risque de 5 % est jointe en Annexe I p.156. Pour une série de 27 composés et des modèles comptant entre 3 et 8 composantes, la lecture de la table donne les valeurs de  $F_{crit}$  suivantes :

$$F(0,05 ; n_1=3 ; n_2=23) = 3,03$$

$$F(0,05 ; n_1=8 ; n_2=18) = 2,51$$

Pour des modèles ayant entre 3 et 8 composantes, la valeur de  $F_{crit}$  varie entre 3,03 et 2,51. Les modèles dont la valeur de F est supérieure à  $F_{crit}$  sont statistiquement significatifs.

#### **b . Analyse des modèles :**

Les modèles produits avec un seul descripteur ont un coefficient de corrélation interne peu élevé. Les trois meilleurs ont des pouvoirs de prévision proches de 0,5 et pour les deux derniers ce pouvoir est presque nul.

Le champ accepteur produit le modèle simple ayant le meilleur pouvoir de prévision interne mais

ce modèle présente un biais d'ajustement important. Le champ hydrophobe produit le modèle simple le moins corrélé à l'activité puisque son pouvoir de prévision interne est pratiquement nul. Le champ électrostatique produit le modèle qui est globalement le plus intéressant avec un pouvoir de prévision interne proche du meilleur, une excellente corrélation interne et un biais d'ajustement très faible. Aucun n'est cependant capable d'expliquer à lui seul les relations structure-activité. Ces modèles comportent tous un nombre de composantes élevé, de 3 à 6.

Tableau 2 : Statistiques de PLS des modèles utilisant toutes les combinaisons de descripteurs CoMSIA et dérivés de l'ensemble des molécules.

| Descripteurs | nc | $q^2_{cv}$ | sep   | $r^2$ | see   | y0     | F       | $(r^2-r_0^2)/r^2$ |
|--------------|----|------------|-------|-------|-------|--------|---------|-------------------|
| S            | 4  | 0.147      | 0.646 | 0.657 | 0.410 | -0.486 | 10.541  | 0.51              |
| E            | 6  | 0.463      | 0.538 | 0.942 | 0.177 | 0.037  | 54.081  | 0.04              |
| H            | 3  | 0.018      | 0.678 | 0.629 | 0.417 | 0.196  | 12.996  | 0.19              |
| D            | 3  | 0.447      | 0.509 | 0.713 | 0.367 | 0.258  | 19.023  | -0.09             |
| A            | 4  | 0.479      | 0.505 | 0.747 | 0.352 | 0.033  | 16.213  | -0.16             |
| S,E          | 6  | 0.494      | 0.522 | 0.938 | 0.182 | -0.298 | 50.636  | 0.06              |
| S,H          | 3  | 0.077      | 0.657 | 0.588 | 0.439 | -0.042 | 10.929  | 0.67              |
| S,D          | 8  | 0.520      | 0.535 | 0.987 | 0.089 | 0.304  | 166.492 | 0.00              |
| S,A          | 6  | 0.503      | 0.517 | 0.963 | 0.142 | 0.397  | 85.865  | 0.00              |
| E,H          | 6  | 0.515      | 0.511 | 0.949 | 0.166 | 0.298  | 61.662  | 0.04              |
| E,D          | 8  | 0.624      | 0.474 | 0.987 | 0.087 | 0.118  | 176.407 | 0.02              |
| E,A          | 6  | 0.533      | 0.501 | 0.962 | 0.144 | -0.049 | 83.665  | 0.00              |
| H,D          | 8  | 0.521      | 0.535 | 0.987 | 0.089 | 0.198  | 169.283 | 0.04              |
| H,A          | 6  | 0.452      | 0.543 | 0.966 | 0.134 | 0.475  | 95.838  | -0.26             |
| D,A          | 3  | 0.465      | 0.501 | 0.751 | 0.341 | 0.215  | 23.118  | -0.24             |
| S,E,H        | 6  | 0.493      | 0.523 | 0.944 | 0.173 | 0.096  | 56.571  | 0.05              |
| S,E,D        | 7  | 0.637      | 0.453 | 0.988 | 0.081 | 0.082  | 230.152 | 0.00              |
| S,E,A        | 6  | 0.553      | 0.491 | 0.972 | 0.123 | -0.041 | 115.232 | 0.02              |
| S,H,D        | 7  | 0.491      | 0.537 | 0.984 | 0.096 | 0.230  | 162.499 | 0.04              |
| S,H,A        | 6  | 0.471      | 0.533 | 0.960 | 0.146 | 0.370  | 80.797  | 0.03              |
| S,D,A        | 8  | 0.502      | 0.546 | 0.984 | 0.096 | 0.420  | 142.780 | 0.09              |
| E,H,D        | 8  | 0.636      | 0.466 | 0.992 | 0.070 | 0.086  | 273.891 | -0.05             |
| E,H,A        | 8  | 0.573      | 0.505 | 0.991 | 0.074 | 0.214  | 244.352 | 0.01              |
| E,D,A        | 8  | 0.580      | 0.501 | 0.986 | 0.091 | 0.048  | 159.365 | -0.23             |
| H,D,A        | 8  | 0.496      | 0.549 | 0.989 | 0.082 | 0.360  | 195.953 | 0.09              |
| S,E,H,D      | 8  | 0.638      | 0.466 | 0.992 | 0.069 | 0.123  | 277.045 | 0.00              |
| S,E,H,A      | 8  | 0.570      | 0.507 | 0.988 | 0.083 | 0.183  | 191.699 | 0.01              |
| S,E,D,A      | 8  | 0.609      | 0.484 | 0.991 | 0.073 | 0.105  | 247.611 | 0.01              |
| S,H,D,A      | 8  | 0.488      | 0.554 | 0.988 | 0.085 | 0.383  | 185.172 | -0.02             |
| E,H,D,A      | 8  | 0.615      | 0.480 | 0.993 | 0.066 | 0.097  | 302.628 | -0.05             |
| S,E,H,D,A    | 8  | 0.615      | 0.480 | 0.992 | 0.068 | 0.133  | 291.676 | 0.00              |

Les variations d'activité entre les molécules de la série sont donc le résultat d'un mécanisme d'interaction ligand récepteur complexe faisant intervenir différents types d'interactions

combinées dans plusieurs régions de l'espace.

Parmi les modèles combinés, les meilleurs  $q^2_{cv}$  sont obtenus en combinant les champs stérique, électrostatique et donneur («S,E,D» :  $q^2_{cv}=0,637$  ;  $nc=7$ ), électrostatique, hydrophobe et donneur (E,H,D :  $q^2_{cv}=0,636$  ;  $nc=8$ ) et stérique, électrostatique, hydrophobe et donneur (S,E,H,D :  $q^2_{cv}=0,638$  ;  $nc=8$ ). Le meilleur de ces 3 modèles est le premier car il utilise une composante de moins à pouvoir de prévision interne égal.

Les modèles les moins performants sont ceux utilisant les champs S et H simultanément et ceux utilisant le descripteur A. Les champs S et H sont individuellement les moins corrélés à l'activité. Leur combinaison ne peut pas apporter plus que ce que chacun contient. Les champs D et A peuvent être corrélés entre eux puisque leur combinaison « D,A », n'est pas plus efficace que chaque champ individuellement. Cependant, le champ D donne de meilleurs résultats lorsqu'on l'associe à d'autres. Les modèles utilisant le descripteur A à la place du champ D comptent généralement moins de composantes mais ce au prix d'une perte de pouvoir de prévision supérieure à 5 points. Le modèle « S,E,A » par exemple n'utilise que 6 composantes mais son pouvoir de prévision interne est inférieur de 8,4 points par rapport au modèle « S,E,D ».

Le pouvoir de prévision interne n'est cependant pas suffisant pour évaluer la qualité des modèles. Les tests de validation externe nous donneront des informations supplémentaires.

## **2 - VALIDATION EXTERNE.**

### **a . Constitution des jeux d'apprentissage et construction des modèles.**

La procédure de validation croisée LOO ayant tendance à surestimer le pouvoir de prévision des modèles, l'évaluation de ce paramètre a été approfondie. Pour ce faire, le jeu de données initial de 27 composés a été partagé en différents jeux d'apprentissage et de test.

Nous avons effectué le partage du jeu initial suivant deux méthodes différentes. Les jeux numéros 2 à 5 sont obtenus par une méthode rationnelle, les jeux numéro 6 à 10 sont obtenus par la sélection de 6 molécules au hasard. Les jeux constitués sont décrits dans le Tableau 3. La valeur 0 indique que la molécule fait partie du jeu de test.

La méthode de sélection rationnelle a consisté à trier les molécules par ordre d'activité décroissante puis à les diviser en classes d'effectif constant. Nous avons sélectionné la dernière, l'avant dernière ou les deux dernières de chaque classe. Nous avons obtenu un échantillonnage de 5 à 7 composés sur toute la gamme d'activité de la série pour les jeux de test. Tous les composés appartiennent au moins une fois à un jeu de test. Tous les jeux de test comportent au

moins 5 composés.

Tableau 3 : Constitution des jeux d'apprentissage (1) et de test (0)

| Composé | Logit  | Jeu01 | Jeu02 | Jeu03 | Jeu04 | Jeu05 | Jeu06 | Jeu07 | Jeu08 | Jeu09 | Jeu10 | Σ |
|---------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---|
| 12      | 1.279  | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 1     | 1     | 1     | 9 |
| 20      | 1.195  | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 9 |
| 4       | 0.954  | 1     | 1     | 1     | 1     | 0     | 1     | 1     | 1     | 1     | 1     | 9 |
| 14      | 0.865  | 1     | 1     | 0     | 1     | 1     | 0     | 1     | 1     | 1     | 1     | 8 |
| 23      | 0.788  | 1     | 0     | 1     | 1     | 1     | 0     | 1     | 1     | 1     | 1     | 8 |
| 16      | 0.788  | 1     | 1     | 1     | 0     | 1     | 1     | 1     | 0     | 1     | 1     | 8 |
| 15      | 0.689  | 1     | 1     | 1     | 0     | 1     | 1     | 1     | 1     | 1     | 1     | 9 |
| 13      | 0.630  | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 9 |
| 17      | 0.602  | 1     | 1     | 0     | 1     | 0     | 1     | 1     | 0     | 1     | 1     | 7 |
| 1       | 0.602  | 1     | 0     | 1     | 1     | 1     | 0     | 1     | 1     | 0     | 1     | 7 |
| 11      | 0.477  | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 1     | 1     | 0     | 8 |
| 5       | 0.368  | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 1     | 1     | 9 |
| 6       | 0.368  | 1     | 1     | 1     | 0     | 1     | 1     | 0     | 1     | 1     | 1     | 8 |
| 7       | 0.269  | 1     | 1     | 0     | 0     | 1     | 1     | 1     | 1     | 0     | 1     | 7 |
| 8       | 0.194  | 1     | 0     | 1     | 1     | 0     | 1     | 1     | 1     | 1     | 1     | 8 |
| 3       | 0.140  | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 0     | 8 |
| 9       | 0.140  | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 0     | 1     | 1     | 8 |
| 25      | 0.105  | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 1     | 9 |
| 22      | 0.087  | 1     | 1     | 0     | 1     | 1     | 0     | 1     | 1     | 1     | 1     | 8 |
| 2       | 0.087  | 1     | 0     | 1     | 0     | 1     | 1     | 1     | 1     | 1     | 1     | 8 |
| 10      | 0.000  | 1     | 1     | 1     | 0     | 0     | 1     | 1     | 1     | 0     | 1     | 7 |
| 27      | -0.347 | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 0     | 1     | 1     | 8 |
| 19      | -0.389 | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 9 |
| 21      | -0.432 | 1     | 1     | 0     | 1     | 1     | 1     | 0     | 0     | 1     | 1     | 7 |
| 18      | -0.477 | 1     | 0     | 1     | 1     | 1     | 0     | 1     | 1     | 1     | 1     | 8 |
| 26      | -0.753 | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 0     | 9 |
| 24      | -1.690 | 1     | 1     | 1     | 0     | 0     | 0     | 1     | 1     | 0     | 1     | 6 |

Les jeux d'apprentissage servent à dériver de nouveaux modèles dont le pouvoir de prévision externe évalue leur capacité à prévoir correctement l'activité des composés du jeu de test correspondant. Toutes les combinaisons de descripteurs sont testées et pour chaque combinaison deux modèles sont dérivés : l'un avec constante, l'autre sans. Nous avons donc calculé 558 modèles supplémentaires avec l'aide du script QBF1. L'ensemble des résultats est résumé dans le tableau Tableau 4, avec les valeurs minimales (min), maximales (max) et moyennes (moy) de chaque paramètre pour chaque combinaison de descripteurs. Les modèles centrés ont également été dérivés.

Ici encore les modèles présentant les valeurs moyennes de  $r^2$  et  $q^2$  les plus élevées utilisent les champs E et D. Nous retrouvons les modèles «S,E,D» ( $q^2_{cv}=0,556$  ;  $nc=7.4$ ), « S,E,H,D » ( $q^2_{cv}=0,528$  ;  $nc=6.9$ ), « E,D » ( $q^2_{cv}=0,548$  ;  $nc=7.4$ ) et « E,H,D » ( $q^2_{cv}=0,530$  ;  $nc=7.6$ ). Leur

pouvoir de prévision interne a baissé d'environ 10 points en raison de la perte d'informations entraînée par le retrait de 5 à 7 molécules de leur jeu d'apprentissage. Ces composés écartés vont nous permettre d'évaluer le pouvoir de prévision externe des modèles.

Tableau 4 : Statistiques des modèles CoMSIA : valeurs minimales (min), maximales (max) et moyennes (moy) sur l'ensemble des 9 jeux d'apprentissage / test.

| Descripteurs | nc<br>gamme/moy | $q^2_{cv}$ |       |       | sep<br>moy | $r^2$ |       |       | $r^2_{pred}$ |       |        |
|--------------|-----------------|------------|-------|-------|------------|-------|-------|-------|--------------|-------|--------|
|              |                 | min        | max   | moy   |            | min   | max   | moy   | min          | max   | moy    |
| S            | 2-8/4.6         | -0.159     | 0.468 | 0.153 | 0.647      | 0.419 | 0.981 | 0.726 | -2.068       | 0.663 | -0.023 |
| E            | 3-8/6           | 0.213      | 0.481 | 0.383 | 0.581      | 0.782 | 0.991 | 0.944 | -2.412       | 0.624 | 0.166  |
| H            | 1-8/4.6         | -0.116     | 0.269 | 0.091 | 0.675      | 0.229 | 0.993 | 0.700 | -1.831       | 0.709 | 0.105  |
| D            | 2-5/3.2         | 0.242      | 0.671 | 0.387 | 0.524      | 0.594 | 0.901 | 0.716 | -1.488       | 0.772 | 0.384  |
| A            | 3-8/4.1         | 0.334      | 0.551 | 0.431 | 0.523      | 0.597 | 0.928 | 0.748 | -0.793       | 0.851 | 0.373  |
| S,E          | 3-8/6           | 0.187      | 0.531 | 0.407 | 0.568      | 0.746 | 0.993 | 0.928 | -2.097       | 0.832 | 0.351  |
| S,H          | 1-8/4.4         | -0.167     | 0.341 | 0.110 | 0.662      | 0.257 | 0.992 | 0.687 | -1.707       | 0.677 | 0.139  |
| S,D          | 2-8/6.2         | 0.235      | 0.696 | 0.446 | 0.550      | 0.831 | 0.994 | 0.957 | -1.353       | 0.765 | 0.372  |
| S,A          | 3-8/6.2         | 0.189      | 0.730 | 0.446 | 0.549      | 0.836 | 0.990 | 0.959 | -0.845       | 0.687 | 0.370  |
| E,H          | 5-8/6.5         | 0.174      | 0.515 | 0.401 | 0.581      | 0.904 | 0.994 | 0.968 | -1.745       | 0.733 | 0.285  |
| E,D          | 5-8/7.4         | 0.309      | 0.758 | 0.548 | 0.518      | 0.977 | 0.992 | 0.988 | -1.228       | 0.877 | 0.512  |
| E,A          | 4-8/6.3         | 0.375      | 0.634 | 0.475 | 0.538      | 0.958 | 0.992 | 0.974 | -0.948       | 0.931 | 0.474  |
| H,D          | 3-8/6.1         | 0.305      | 0.540 | 0.435 | 0.553      | 0.882 | 0.996 | 0.962 | -1.174       | 0.810 | 0.439  |
| H,A          | 3-8/5.9         | 0.266      | 0.561 | 0.431 | 0.554      | 0.897 | 0.993 | 0.966 | -0.854       | 0.729 | 0.372  |
| D,A          | 3-5/3.3         | 0.130      | 0.555 | 0.386 | 0.529      | 0.647 | 0.927 | 0.759 | -1.310       | 0.780 | 0.425  |
| S,E,H        | 3-8/6.7         | 0.164      | 0.531 | 0.391 | 0.592      | 0.729 | 0.997 | 0.944 | -1.864       | 0.695 | 0.264  |
| S,E,D        | 6-8/7.4         | 0.315      | 0.739 | 0.556 | 0.516      | 0.988 | 0.995 | 0.992 | -0.852       | 0.858 | 0.554  |
| S,E,A        | 5-8/6.6         | 0.335      | 0.626 | 0.490 | 0.538      | 0.966 | 0.994 | 0.982 | -0.963       | 0.952 | 0.458  |
| S,H,D        | 3-7/5.5         | 0.251      | 0.571 | 0.420 | 0.550      | 0.896 | 0.996 | 0.965 | -0.813       | 0.787 | 0.448  |
| S,H,A        | 3-8/6.1         | 0.118      | 0.603 | 0.404 | 0.571      | 0.816 | 0.996 | 0.963 | -0.622       | 0.679 | 0.399  |
| S,D,A        | 3-8/6.9         | 0.153      | 0.666 | 0.441 | 0.566      | 0.863 | 0.998 | 0.965 | -0.795       | 0.718 | 0.428  |
| E,H,D        | 6-8/7.6         | 0.298      | 0.712 | 0.530 | 0.533      | 0.990 | 0.997 | 0.994 | -0.874       | 0.888 | 0.566  |
| E,H,A        | 6-8/7.2         | 0.333      | 0.669 | 0.496 | 0.543      | 0.982 | 0.998 | 0.992 | -0.962       | 0.884 | 0.467  |
| E,D,A        | 3-8/7.1         | 0.234      | 0.684 | 0.502 | 0.539      | 0.871 | 0.990 | 0.975 | -1.129       | 0.856 | 0.495  |
| H,D,A        | 3-8/6.3         | 0.219      | 0.561 | 0.430 | 0.560      | 0.798 | 0.998 | 0.949 | -0.794       | 0.763 | 0.443  |
| S,E,H,D      | 3-8/6.9         | 0.288      | 0.694 | 0.528 | 0.518      | 0.947 | 0.997 | 0.988 | -0.806       | 0.890 | 0.537  |
| S,E,H,A      | 6-8/7.5         | 0.286      | 0.643 | 0.491 | 0.554      | 0.981 | 0.998 | 0.992 | -0.972       | 0.891 | 0.463  |
| S,E,D,A      | 3-8/7           | 0.249      | 0.698 | 0.526 | 0.522      | 0.922 | 0.994 | 0.984 | -0.960       | 0.839 | 0.504  |
| S,H,D,A      | 3-8/6.6         | 0.219      | 0.597 | 0.426 | 0.569      | 0.887 | 0.997 | 0.969 | -0.707       | 0.754 | 0.433  |
| E,H,D,A      | 3-8/6.5         | 0.266      | 0.683 | 0.515 | 0.518      | 0.904 | 0.995 | 0.975 | -0.998       | 0.858 | 0.509  |
| S,E,H,D,A    | 3-8/6.8         | 0.262      | 0.676 | 0.517 | 0.523      | 0.935 | 0.996 | 0.987 | -0.840       | 0.868 | 0.517  |

#### b . Estimation du pouvoir de prévision externe :

Le pouvoir de prévision externe est rapporté dans le Tableau 4 par le coefficient  $r^2_{pred}$ . Pour être tout à fait significatif, la validation externe devrait se faire avec un jeu de test d'effectif au moins égal au nombre de composantes du modèle testé. Malheureusement notre effectif total est limité à 27 composés et dériver des modèles fiables avec moins de 20 structures devient très difficile.

Ce test nous donne néanmoins une indication sur le pouvoir de prévision de nos modèles.

Les modèles ayant les meilleurs pouvoirs de prévision externe concordent avec les modèles de meilleurs pouvoirs de prévision interne. Les plus hautes valeurs de  $r^2_{\text{pred}}$  sont obtenues par combinaison des champs «E,H,D» ( $r^2_{\text{pred}}=0,566$ ), «S,E,D» ( $r^2_{\text{pred}}=0,554$ ), «S,E,H,D» ( $r^2_{\text{pred}}=0,537$ ) et «S,E,H,D,A» ( $r^2_{\text{pred}}=0,517$ ).

Le modèle «S,E,D» offre le meilleur compromis sur l'ensemble des indices. L'analyse détaillée de ces indices nous aidera à identifier les forces et les faiblesses du modèle.

### c . Analyse du pouvoir de prévision de la combinaison «S,E,D»

Les modèles combinant les descripteurs «S,E,D» ont le pouvoir de prévision interne moyen et le pouvoir de prévision externe moyen les plus élevés ( $q^2_{\text{cv}}$  moyen = 0,556 ;  $r^2_{\text{pred}}$  moyen = 0,554). La proximité de ces valeurs indique la stabilité globale de l'information apportée par ces 3 descripteurs. Le biais d'ajustement est très faible comme l'indique l'ordonnée à l'origine moyenne  $\bar{y}_0 = 0,073$  et la moyenne de l'indice  $(r^2 - r_0^2)/r^2$  inférieure à 0,01.

La Figure 6 représente les graphes de corrélation du modèle «S,E,D» dérivé du jeu numéro 3 ( $r^2 = 0,994$  ;  $q^2_{\text{cv}} = 0,631$  ;  $r^2_{\text{pred}} = 0,841$ ).

Le coefficient de corrélation interne sur le jeu d'apprentissage est excellent. Les points du graphe de corrélation interne (Figure 6 A) sont tous très proche de la droite idéale, de pente 1 et d'ordonnée à l'origine nulle. Cette droite est représentée en trait noir fin sur tous les graphes de corrélation. Ce modèle explique donc l'activité de chaque molécule du jeu d'apprentissage avec une erreur minimale.

Sur le graphe de corrélation de la validation croisée (Figure 6 B), la corrélation est un peu moins bonne. L'écart entre les points et la droite idéale est plus important mais reste néanmoins acceptable. A l'exception de deux points, le modèle est capable d'expliquer l'activité de composés inconnus avec une faible erreur. De même, la prévision d'activité des composés du jeu de test est bonne (Figure 6 C). Cela signifie que ce jeu de test est représentatif de l'espace chimique du jeu d'apprentissage : chaque structure du jeu de test est proche d'au moins une structure du jeu d'apprentissage.

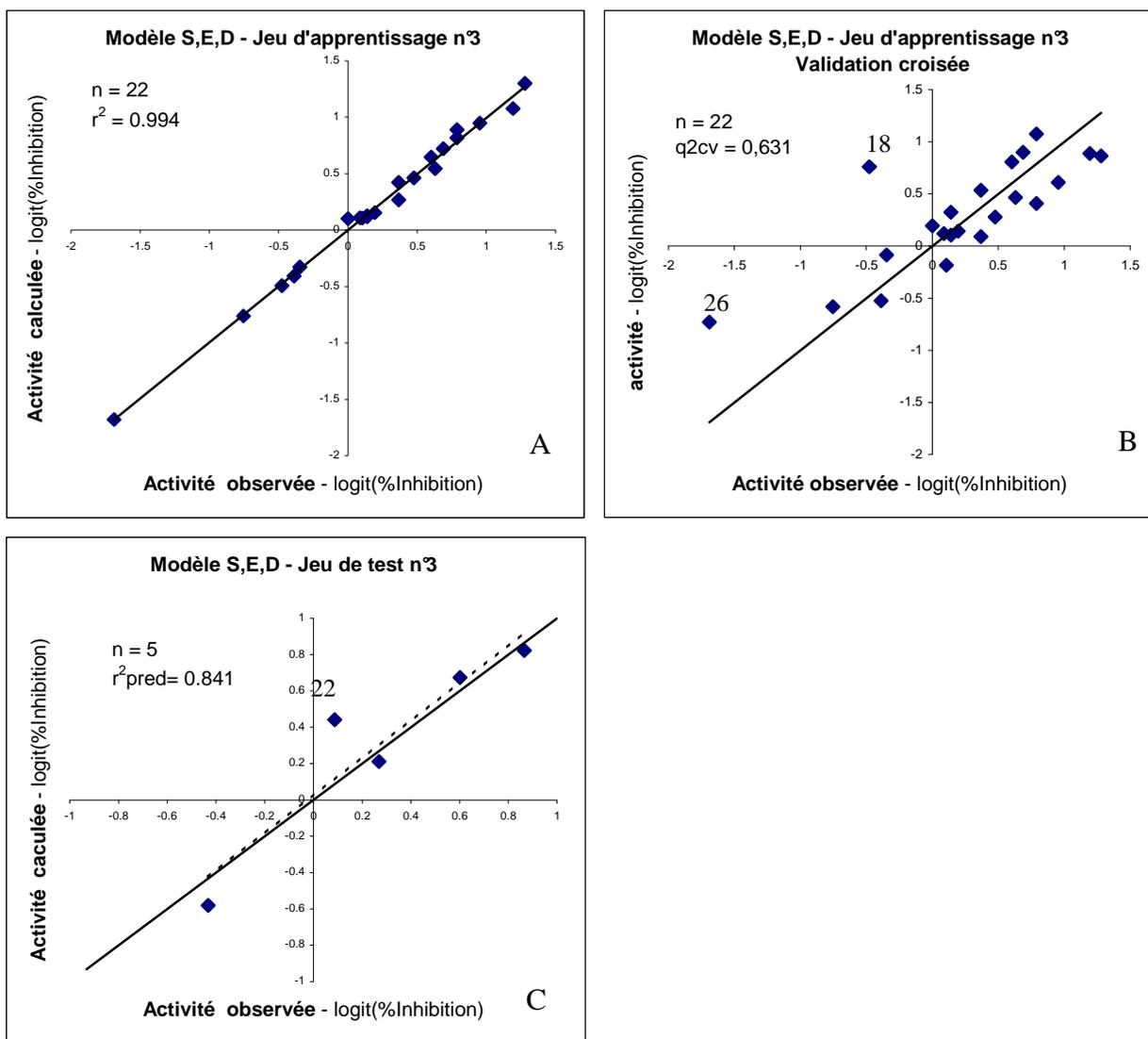


Figure 6 : Graphes de corrélation du modèle «S,E,D» calculé sur le jeu numéro 3 : (A) corrélation interne sur le jeu d'apprentissage, (B) validation croisée du jeu d'apprentissage et (C) corrélation des prévisions du jeu de test.

Lors du test de validation croisée, la prévision d'activité des composés **18** et **26** est surestimée.

Pour le composé **18**, cette surestimation est due au fait qu'il est le seul à posséder un groupement farnésyle comme substituant R4. Ce groupement est extrêmement étendu puisqu'il comporte une chaîne de 15 carbones. Lors de la validation croisée, lorsque ce composé est retiré du jeu d'apprentissage pour être testé en aveugle, le modèle n'a plus d'information dans la région qu'occupe l'extrémité du groupement farnésyle. Il ignore donc l'impact négatif de l'encombrement stérique à cette extrémité. L'activité calculée est donc comparable à celle du composé **16** (logit = 0,76) qui porte un substituant géranyle, l'homologue à 10 carbones du farnésyle, comme substituant R4.

Pour le composé **26**, la surestimation vient du fait que c'est le composé qui a la plus faible

activité (2%) et qu'il est isolé dans cette partie de la gamme d'activités. En effet, il n'y a que 5 composés entre 0 et 30% d'activité et l'écart entre le plus faible et le suivant est de 13 points. De plus, la transformation Logit accentue très fortement l'éloignement des valeurs extrêmes. Le modèle est donc dérivé sur une moyenne de valeurs cibles beaucoup plus élevées. D'autre part, le composé **26** est le seul à avoir un hydroxyle comme substituant R6 dont l'impact est très négatif sur l'activité. Lorsque le composé est retiré pour être testé en aveugle, le modèle ignore l'effet négatif de l'hydroxyle sur cette position. Ces deux facteurs se combinent pour conduire à une sous-estimation du rôle très négatif de l'hydroxyle en R6 et donc une surestimation de la valeur calculée.

Lors de l'évaluation du jeu de test, le composé **22** est légèrement surestimé. Il semble que le modèle surestime l'impact bénéfique du groupement méthoxyle sur la position R2'. En effet, si l'on compare les composés **21** et **22**, la seule différence entre les structures est la présence en R2', d'un groupement méthoxyle sur le composé **22** alors qu'il y a un hydroxyle sur le composé **21**. La différence d'activité entre ces composés permet donc une lecture directe de l'impact de cette variation sur l'activité. Or les composés **21** et **22** sont tous les deux dans le jeu de test et le jeu d'apprentissage ne contient plus de structures permettant une observation aussi précise.

Le modèle « S,E,D » a donc un pouvoir de prévision acceptable. Ce pouvoir pourrait être accru par l'ajout de structures complémentaires dans le jeu de données initial. Il faudrait notamment évaluer l'activité de molécules occupant la même région que la dernière unité prenyle du groupement farnésyle, et des composés possédant un groupement hydroxyle en position R6.

### 3 - POUVOIR DE PREVISION DU MODELE «S,E,D»

Le modèle «S,E,D», calculé avec l'ensemble des composés, possède les caractéristiques suivantes :  $r^2 = 0,988 > 0,6$  ;  $q^2_{cv} = 0,64 > 0,5$  ;  $(r^2 - r_0^2)/r^2 = 0,01 < 0,1$ . La Figure 7 représente les graphes de corrélation interne, à gauche et de corrélation lors de la validation croisée, à droite.

Sur le graphe de corrélation interne, on constate que tous les points sont très proches de la droite idéale, représentée en trait continu et fin. Sur le graphe de corrélation de la validation croisée, la dispersion des points est plus importante. L'erreur commise par le modèle est cependant acceptable puisque le coefficient de corrélation de validation croisée reste supérieur à 0,6. La droite de régression des points est représentée en pointillé.

Les points les plus éloignés de la droite idéale correspondent là encore aux composés **18** et **26** et ce pour les raisons évoquées au paragraphe précédent. Ce modèle satisfait les critères de validation du pouvoir de prévision. De plus, il présente l'avantage de n'utiliser que 7

composantes là où les modèles de qualité comparable (E,H,D ; S,E,H,D ; E,H,D,A) en utilisent 8. C'est donc celui que nous exploiterons pour faire les prévisions d'activités de nouvelles structures.

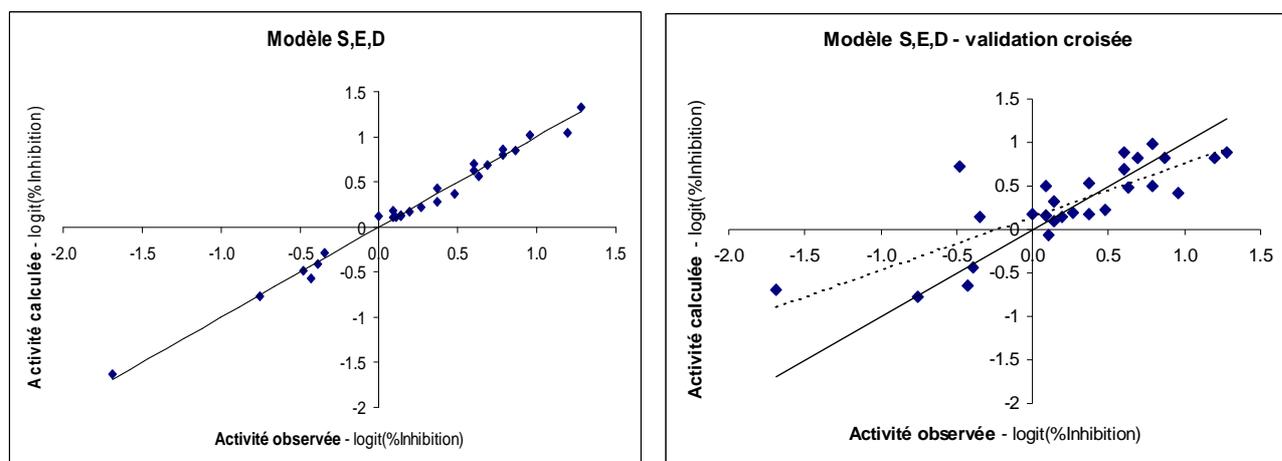


Figure 7 : Graphes de corrélation du modèle «S,E,D» sur le jeu de données complet. Corrélation interne à gauche, validation croisée à droite.

La PLS nous indique également la contribution relative de chaque champ dans le modèle :

- Champ stérique : 19,7%
- Champ électrostatique : 42,7%
- Champ donneur : 37,6%.

La corrélation entre les variables explicatives et la valeur cible est donc établie. Il faut maintenant vérifier que cette corrélation n'est pas due au hasard ou à un phénomène d'autocorrélation. Un test de robustesse par hasardisation des réponses nous permettra de contrôler ce point.

#### 4 - TEST DE ROBUSTESSE DU MODELE « S,E,D »

Lors de ce test, les valeurs d'activité des composés de la série sont mélangées et redistribuées de façon aléatoire. Le modèle «S,E,D» est redérivé et les différents paramètres sont relevés. L'opération est répétée 100 fois et les moyennes des coefficients de corrélation donnent  $r^2 = 0,41$  et  $q^2_{cv} = -0,249$ . Cette série de calculs a été effectuée grâce au script QYR. Le fonctionnement de ce script est détaillé dans la partie expérimentale et son code est fourni en Annexe IV p. 162.

Les valeurs moyennes de  $r^2$  et  $q^2_{cv}$  obtenues pour des valeurs cibles aléatoires sont nettement inférieures à celles du modèle «S,E,D» et aux seuils de validation du pouvoir de prévision. Ce

test indique que les corrélations décrites par le modèle entre les descripteurs et l'activité ne sont pas dues au hasard. Le modèle est robuste.

## **5 - CONCLUSION DES TESTS DE VALIDATION**

L'ensemble des tests et paramètres statistiques du modèle « S,E,D » convergent favorablement. Le modèle est robuste et présente une excellente corrélation interne. Son pouvoir de prévision interne et son pouvoir de prévision externe sont satisfaisants. Le biais d'ajustement est négligeable et le test de Fischer indique qu'il est statistiquement significatif c'est-à-dire que l'erreur qu'il commet est nettement inférieure à ce qu'il explique. Nous considérons donc que le modèle « S,E,D » est valide. Nous pouvons maintenant l'exploiter pour comprendre les relations structure-activité de notre série de composés et faire des prévisions d'activité pour de nouvelles structures.

## ***V - RELATIONS STRUCTURE-ACTIVITE***

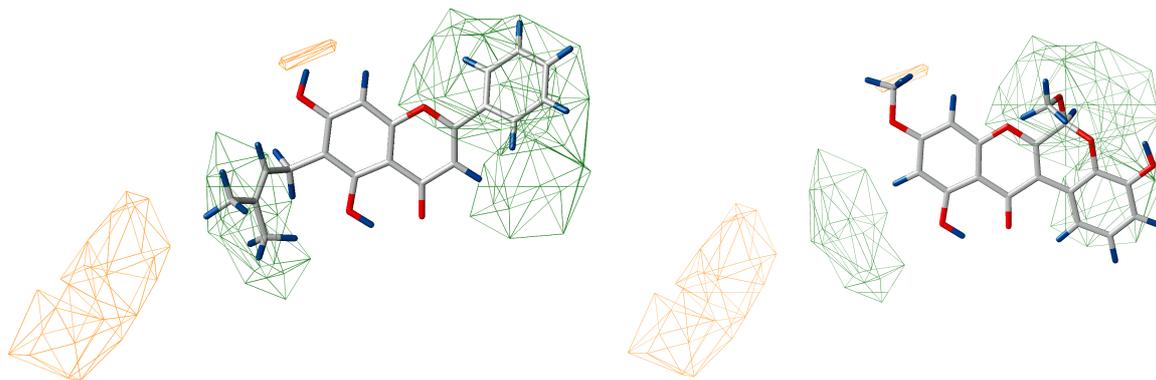
### **1 - EXPLICATION DES RELATIONS STRUCTURE-ACTIVITE PAR LE MODELE « S,E,D »**

Afin de mieux comprendre les relations structure-activité de notre série de composés nous allons analyser les graphes de contour SD\*Coefficient. Pour le champ stérique, nous allons visualiser le niveau 30 % qui correspond aux régions défavorables et le niveau 80 % qui correspond aux régions favorables. Pour les champs électrostatique et donneur de liaison hydrogène, nous utilisons les niveaux 20 % et 80 %. Ces niveaux sont représentés avec différents codes de couleurs.

#### **a . Le champ stérique**

Dans le champ stérique, les régions en vert sont favorables à l'encombrement et doivent être occupées alors que les régions oranges sont défavorables à l'encombrement et doivent rester vides. Les composés **12** et **20** sont les plus actifs avec 95% et 94% d'inhibition respectivement. Ils occupent largement les régions favorables et laissent les régions défavorables vides (Figure 8). La région favorable de droite correspond à l'espace occupé par les substituants R1 et R2 des flavonoïdes, les cycles B et D ainsi que les substituants R1' et R2' des boerhavinones. La région favorable de gauche est occupée par les substituants R4.

Sur la Figure 8, la petite région orange, défavorable à l'encombrement, est située à 4 Å derrière le plan de la molécule. Elle est due aux substituants R5 qui sont une série d'homologation (composés **5** à **10**) dont l'activité décroît progressivement avec l'allongement de la chaîne. C'est l'extrémité de la chaîne la plus longue qui est indiquée comme défavorable à l'encombrement stérique. Elle induit une perte sensible de 20 % d'activité par rapport au composé **5** portant la chaîne la plus courte.



*Figure 8 : Le champ stérique occupé par les composés **12** à gauche (95 % d'inhibition) et **20** à droite (94 % d'inhibition). Régions favorables en vert, défavorables en orange.*

La région défavorable de gauche peut être occupée par des substituants en R4. Parmi ces substituants, il y a une série d'homologues terpéniques (**12** prényle, **14** géranyle, **18** farnésyle) dont l'activité décroît également avec l'allongement. Le composé **18** qui porte le plus long groupement (farnésyle), est l'un des moins actifs de la série (25 % d'inhibition) alors que le composé **12** qui porte le plus petit (prényle), est le composé le plus actif (95 % d'inhibition). La région verte de gauche correspond à la première unité isoprène, identifiée comme très favorable (Figure 8 à gauche) alors que la région orange de gauche correspond à la troisième unité isoprène du farnésyle, très défavorable (Figure 9).

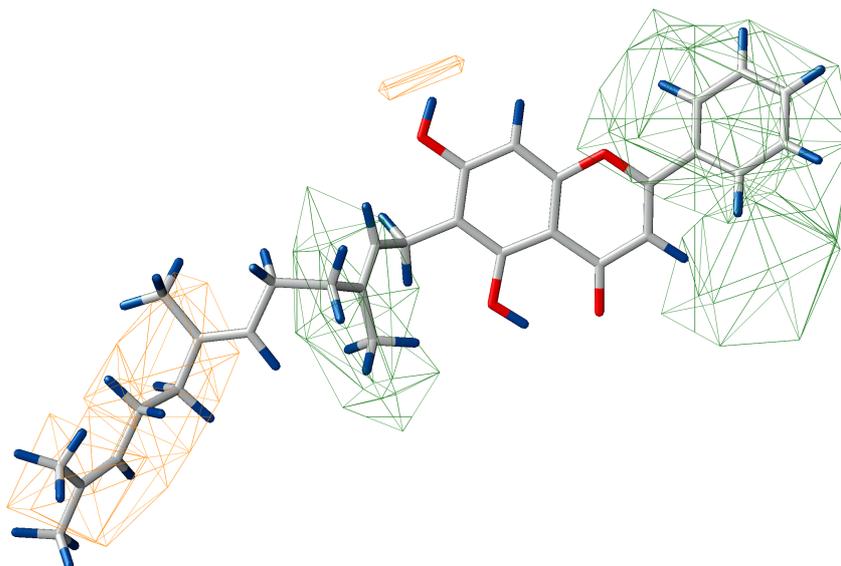


Figure 9 : Le composé **18** (25 % d'inhibition) dans le champ stérique. Régions favorables à l'encombrement en vert, régions défavorables en orange.

### **b . Le champ électrostatique**

Ce champ est le plus riche en information. C'est celui qui contribue le plus à expliquer l'activité avec un poids relatif de 42,7%.

Sur la Figure 10, les deux composés les plus actifs sont représentés avec les zones où les variations d'électronégativité ont le plus d'influence sur l'activité. Les zones favorables à l'augmentation de l'électronégativité sont représentées en rouge, celles favorables à la diminution de l'électronégativité sont en bleu.

La liaison entre les carbones 2 et 3 traverse une région rouge. Dans cette région, l'augmentation de la densité de charge négatives est favorable à l'activité. Pour nos structures cela se traduit par le fait qu'une double liaison est préférable à une liaison simple.

La région électronégative au niveau des substituants R1 et R1' indique que ces substituants renforcent l'activité s'ils sont riches en électrons. Le groupement phényle et les atomes électronégatifs comme l'oxygène sont bénéfiques dans cette région.

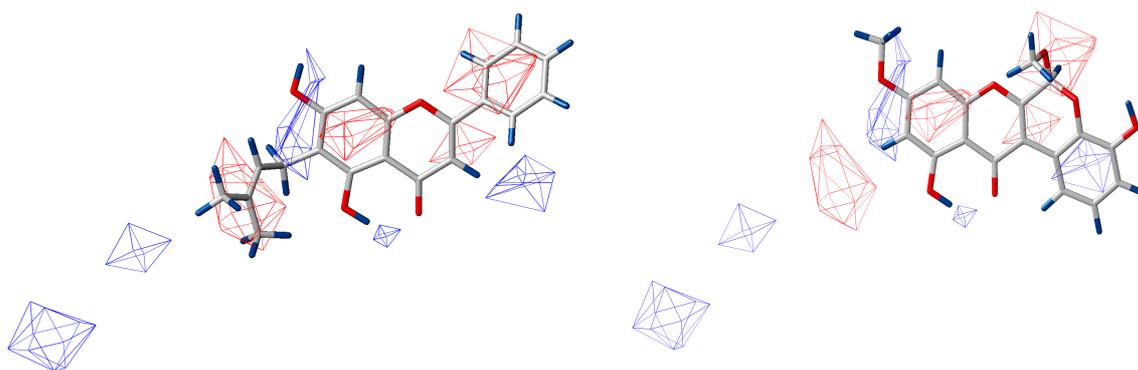


Figure 10 : Composés **12** à gauche (95 % d'inhibition) et **20** à droite (94 % d'inhibition) dans le champ électrostatique. Régions électronégatives en rouge, régions électropositives en bleu.

La position  $\alpha$  du substituant R4 n'est pas propice à un atome électronégatif. Il faut donc éviter les substituants hydroxyles ou éthers à cet endroit.

La région bleue la plus à droite est située au centre du cycle aromatique B des boerhavinones. Cela indique qu'il est préférable que ce cycle soit appauvri en électrons donc qu'il porte des substituants électro-attracteurs. Or les substituants R2' et R2'' branchés sur ce cycle sont soit des hydrogènes soit des hydroxyles. L'impact de la densité électronique sur l'activité est donc corrélée à l'influence de ces substituants qui interviennent fortement dans le champ donneur de liaison hydrogène (cf. ci-dessous).

Le même phénomène de corrélation s'observe pour la région rouge la plus à gauche où se situe le groupement prényle du composé le plus actif (composé **12**). La double liaison du prényle, riche en électrons, est corrélée au caractère très favorable de l'encombrement du groupement. Le modèle ne peut pas attribuer le gain d'activité à l'un ou l'autre de ces deux caractères. Il faudrait compléter la série de molécules avec des variations sur un seul des deux champs dans cette zone, des substituants R4 de type benzyle ou propen-2-one par exemple.

La région bleue la plus à gauche est située au niveau de la double liaison terminale du groupement farnésyle, défavorable à l'activité. L'encombrement stérique et l'électronégativité sont là encore corrélés. Des variations dans cette région, n'impliquant qu'un seul des deux champs, sont nécessaires pour quantifier la contribution de chaque champ à l'activité.

### c . Le champ donneur de liaison hydrogène

Sur les figures suivantes, les régions favorables à un donneur de liaison hydrogène sont en bleu et les régions défavorables sont en violet.

Les substituants R1, R1' et R6 ne conviennent pas à un donneur de liaison hydrogène. La faible

activité du composé **26** vient du fait qu'il porte des substituants donneurs en R1' et R6, orientés vers les zones violettes (Figure 11 à droite). Par contre les régions R2' et R2'' sont très favorables à un donneur comme l'illustre la Figure 11 de gauche où un composé actif possède un hydroxyle en position R2'.

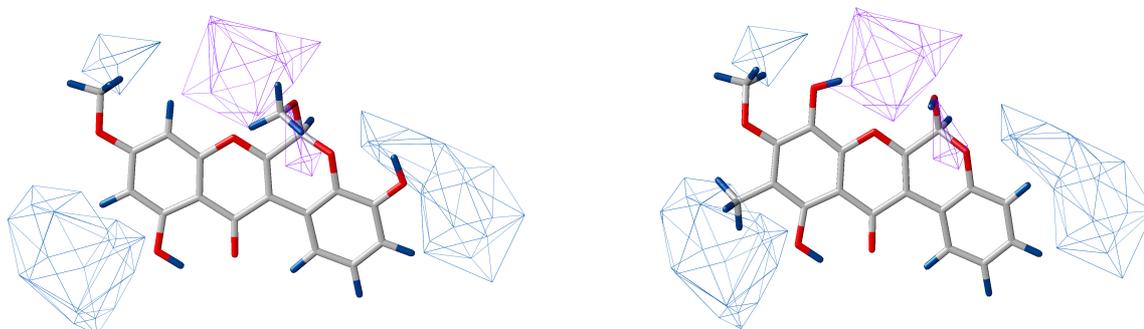


Figure 11 : Composés **20** à gauche (95 % d'inhibition) et **26** à droite (2 % d'inhibition), dans le champ donneur de liaisons hydrogène. Régions favorables en bleu, régions défavorables en violet.

Les substituants R3 et R5 peuvent être des donneurs tels que les groupements hydroxyle ou amine, orientés vers l'une des deux régions bleues voisines. En revanche, nous avons vu que pour R4, un atome électronégatif n'est pas souhaitable en position  $\alpha$ . Un donneur pourrait cependant orienter un hydrogène vers la grande zone bleue de gauche. Il faudrait évaluer l'activité d'un composé portant un donneur électropositif tel qu'une amine sur cette position.

Le modèle « S,E,D » est donc très riche en informations. Cependant, ne néglige-t-il pas certains éléments que d'autres champs nous auraient apportés ? Nous devons le comparer aux autres modèles, de qualité voisine.

## 2 - EVALUATION DES MODELES DE QUALITE VOISINE

### a . Les modèles « E,H,D » et « S,E,H,D »

Le modèle « E,H,D » ( $r^2 = 0,992$  ;  $q^2_{cv} = 0,636$  ;  $c = 8$  ;  $(r^2 - r_0^2)/r^2 = 0,001$  ;  $r^2_{pred} = 0,566$ ) présente des paramètres proches de ceux de « S,E,D ». Il respecte également les critères de validation du pouvoir de prévision. Cependant il utilise une composante de plus que le modèle « S,E,D » ce qui le rend moins robuste donc moins intéressant. Le champ stérique de « S,E,D » y est remplacé par le champ hydrophobe (H). Or, les variations importantes d'encombrement

stérique sont principalement dues à des chaînes aliphatiques hydrophobes. Les deux champs stérique et hydrophobe sont donc certainement corrélés. La superposition du graphe de contour stérique du modèle «S,E,D» au graphe de contour hydrophobe du modèle «E,H,D» va dans le sens cette hypothèse (Figure 12).

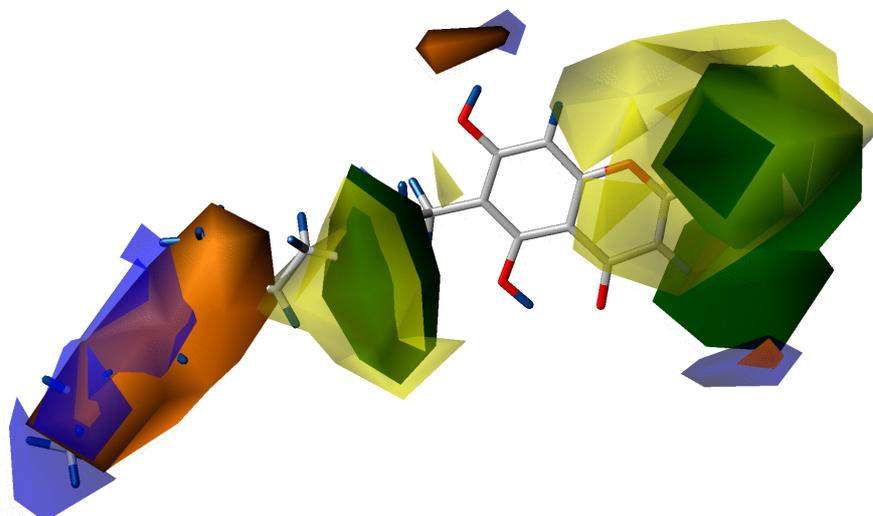


Figure 12 : Superposition des champs stérique du modèle «S,E,D» et hydrophobe du modèle «E,H,D». Les régions stériques favorables sont en vert opaque, les régions défavorables sont en orange opaque. Les régions jaunes transparentes et bleue transparentes sont respectivement hydrophobes et hydrophiles.

Les zones hydrophobes englobent les régions favorables à l'encombrement alors que les zones défavorables à l'encombrement sont marquées comme hydrophiles dans le modèle «E,H,D».

Les champs S et H sont pratiquement équivalents en terme d'information. L'utilisation de ces deux champs dans le même modèle ne présente donc aucun intérêt puisqu'il conduit plus à une redondance d'informations qu'à un gain. Les paramètres du modèle «S,E,H,D» le confirment : ils sont statistiquement équivalents à ceux de «S,E,D» mais avec une composante de plus ( $r^2 = 0,992$  ;  $q^2_{cv} = 0,638$  ;  $c = 8$  ;  $(r^2 - r_0^2)/r^2 = 0,001$  ;  $r^2_{pred\text{moyen}} = 0,537$ ).

#### **b . Les modèles «S,E,D,A» et «S,E,H,D,A»**

Les données du Tableau 2 et du Tableau 4 indiquent que l'emploi du champ accepteur de liaisons hydrogène n'apporte pas d'information supplémentaire significative. Les paramètres  $r^2$  et  $q^2_{cv}$  sont même légèrement inférieurs pour un nombre de composantes toujours égal à 8.

Le champ accepteur est probablement corrélé au champ donneur. En effet, les variations dans ces champs ont lieu principalement sur les positions R1', R2' et R2''. Ces variations sont apportées par la présence ou l'absence de substituants d'hydroxyles. Or, ces groupements sont à la fois

donneurs et accepteurs de liaisons hydrogène. Par conséquent, les champs donneur et accepteur suivent vraisemblablement les mêmes tendances.

D'autre part, les accepteurs de liaisons hydrogène sont principalement les hétéroatomes d'oxygène. Ces atomes sont électronégatifs et, dans notre série, sont les principaux responsables des variations d'électronégativité. Il est donc possible que les champs A et E soient corrélés et redondants.

### **3 - POUVOIR EXPLICATIF DES MODELES.**

Ces différentes comparaisons indiquent que l'association des champs S, E et D permet de rassembler l'essentiel de l'information contenue dans le jeu d'apprentissage. Le modèle « S,E,D » est le plus explicatif et l'absence des champs H et A n'entraîne qu'une perte minimale d'information. Nous pouvons maintenant faire le bilan de ces explications c'est-à-dire des caractéristiques favorables et défavorables à l'activité inhibitrice des composés.

#### 4 - SYNTHÈSE DES CARACTÉRISTIQUES QUI INFLUENCENT L'ACTIVITÉ

Les observations issues des graphes de contours concernant la nature et la position des caractéristiques influençant l'activité sont résumées sur les diagrammes suivants :

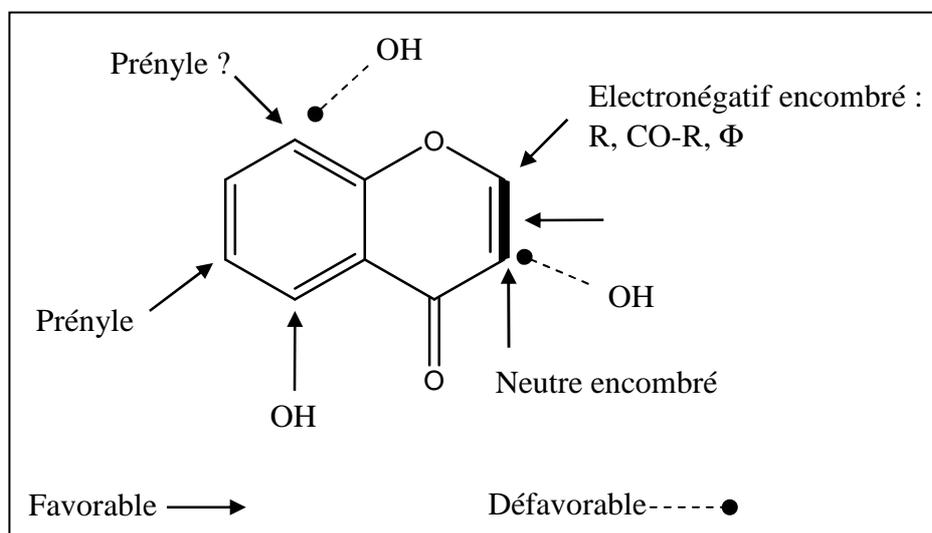


Figure 13: Diagramme de relation structure-activité des flavonoïdes orientés vers l'inhibition de BCRP.

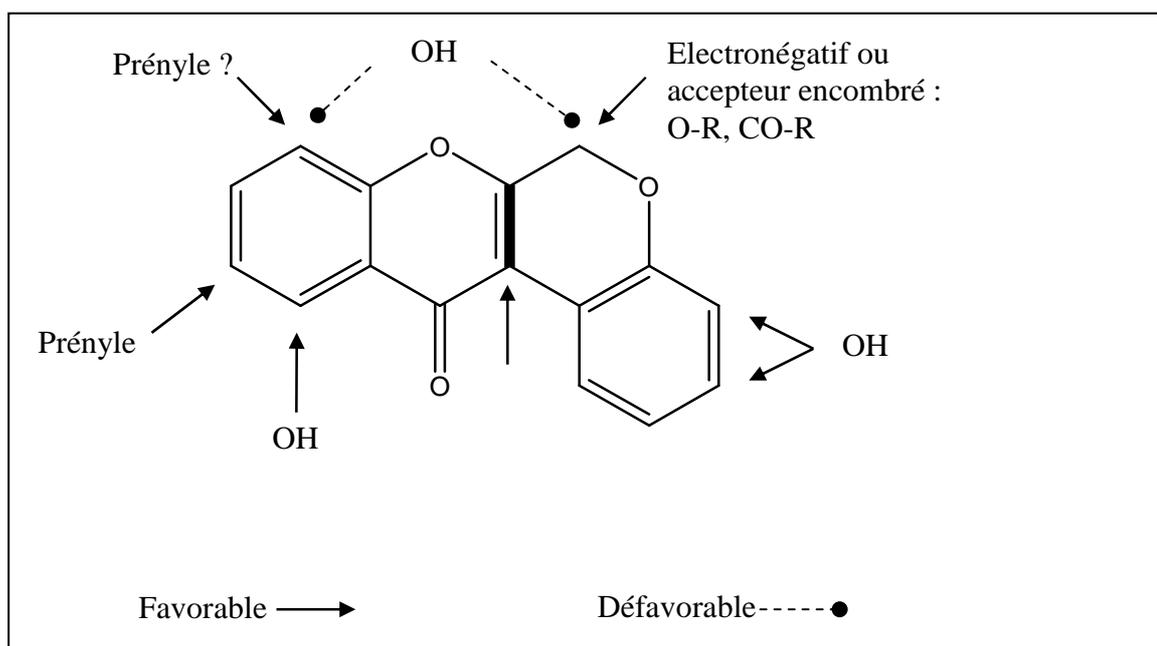


Figure 14 : Diagramme des relations structure-activité des boerhavinones orientés vers l'inhibition de BCRP.

## ***VI - DOMAINE ET LIMITES DU MODELE***

### **1 - ECHANTILLONAGE DE L'ESPACE CHIMIQUE**

Le pouvoir de prévision du modèle est acceptable mais perfectible. Le modèle a des difficultés à prévoir l'activité de certains composés des lors qu'ils ont absents du jeu d'apprentissage. Ils sont cependant riches en informations et le modèle doit être enrichi en composés analogues en terme de structure ou d'activité.

### **2 - CORRELATION DE CHAMPS**

Nous avons constaté que certains champs sont partiellement corrélés. L'ambiguïté entre les types d'interaction que produisent certains groupements, pourraient être corrigées par l'intégration de structures complémentaires. Les groupements hydroxyles, par exemple, sont simultanément électronégatifs, donneurs et accepteurs de liaisons hydrogènes. En les remplaçant par des groupements isostères ne revêtant qu'un seul de ces caractères, nous pourrions préciser la nature de l'interaction mise en jeu. On peut ainsi utiliser une amine pour son caractère exclusivement donneur de liaison hydrogène et un groupement trifluorométhyle pour son caractère exclusivement électronégatif.

### **3 - EXTENSION DE L'ESPACE CHIMIQUE**

Le domaine de l'analyse couvre les variations structurales d'un jeu de 27 molécules. Ce domaine est donc modeste et son extension passe par l'intégration de variations structurales originales dans le jeu d'apprentissage. Les nouvelles variations possibles sont infinies et c'est principalement l'accessibilité de ces nouvelles variantes qui orientera l'exploration.

On peut se demander, par exemple, quel est le rôle de la fonction cétone en position 4. Nous n'avons aucune information sur son influence puisque tous les composés de la série possèdent cette fonction à cette position. Est-elle impliquée dans l'activité des composés et donc en interaction avec la cible ? Si c'était le cas, est-ce son caractère électronégatif ou son caractère accepteur de liaison hydrogène qui est en jeu ? Nous pouvons remplacer cette cétone par une thiocétone ou une imine pour évaluer l'influence de ces deux caractères.

Les nouveaux substituants à employer doivent permettre d'explorer et de définir l'espace disponible autour du noyau et la nature des groupements utiles ou néfastes à l'activité dans les régions encore inexplorées.

#### **4 - LIMITATION DE LA TRANSFORMATION LOGIT**

La transformation Logit nous a permis de redistribuer les valeurs d'activité proches des limites et de les écarter de façon exponentielle. Le modèle fait donc ses prévisions en unité Logit. Dès lors, il est impossible d'atteindre les limites et donc les 100 % d'inhibition. Or, c'est expérimentalement possible. Les prévisions auront globalement tendance à sous-estimer les activités et ce d'autant plus que l'on se rapproche du maximum d'inhibition.

#### **5 - LIMITATIONS DU TEST BIOLOGIQUE**

Les composés très actifs à la concentration du test donnent tous la même réponse de 100 %. Il n'est alors plus possible de les comparer ; le test n'est plus quantitatif et revient à une réponse « oui /non ». Certains composés du jeu d'apprentissage ont déjà une activité de 95 %. Compte tenu de l'erreur expérimentale, estimée par les biologistes à 5 %, nous sommes aux limites de ce que ce test peut nous apprendre de façon quantitative.

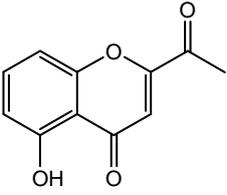
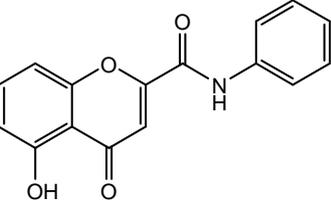
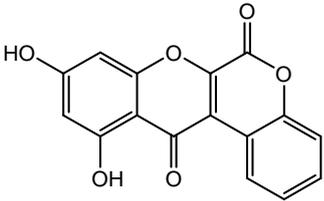
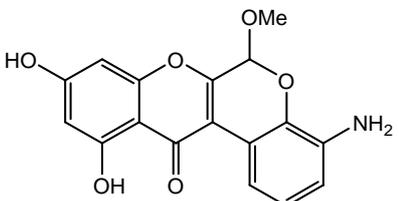
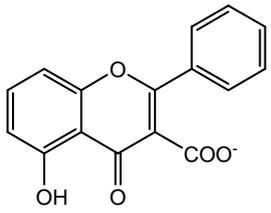
Le test mesurant la concentration de composé qui inhibe 50 % de l'activité de la cible (IC50) serait préférable. Sa réponse n'est pas bornée, nous évitant le recours au Logit. Les prévisions seraient toujours quantitatives et ce test devient incontournable si l'on veut poursuivre l'amélioration des inhibiteurs. L'IC50 des flavonoïdes a été mesurée et celle des boerhavinones est en cours.

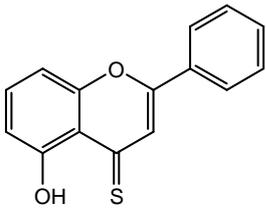
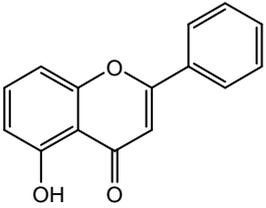
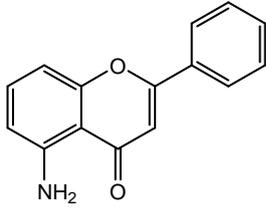
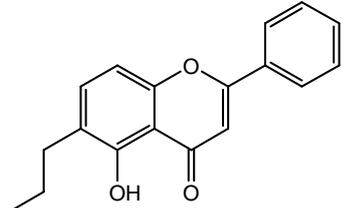
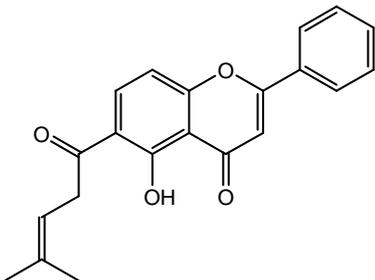
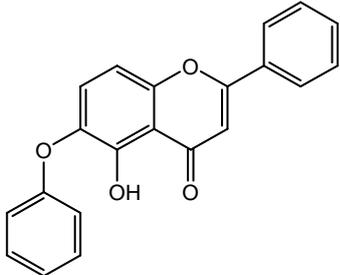
### ***VII - CONSOLIDATION DU MODELE ET OPTIMISATION DES INHIBITEURS.***

La finalité de cette étude est de guider la conception et la sélection de nouvelles structures pour aboutir à des composés plus actifs que ceux déjà connus. La précision des prévisions d'activité de ces nouveaux composés démontrera la validité du modèle. Un certain nombre de structures ont donc été proposées accompagnées d'une prévision d'activité. Quelques unes ont été rassemblées et présentées dans le Tableau 5. Les prévisions sont calculées en unité Logit puis converties en pourcentage d'inhibition. Pour chaque prévision, le nombre de points de matrice extrapolés est indiqué ainsi que la contribution de ces points à la prévision. Plus ces valeurs sont élevées, moins la prévision est fiable.

Certaines de ces structures ont déjà été testées et d'autres sont en cours de synthèse. Les résultats biologiques disponibles figurent également dans le Tableau 5.

Tableau 5: Proposition de structures visant à confirmer et à enrichir le modèle.

| #  | Structure   | Activité                 |              |       |              | Observation |
|----|---|--------------------------|--------------|-------|--------------|-------------|
|    |   | Prévision                |              |       | % Inhibition |             |
|    |   | Extrapolations<br>(/548) | Contribution | Logit |              |             |
| 28 |    | 58                       | 0,134        | 0,808 | 86 %         | ?           |
| 29 |   | 102                      | -0.3         | 0,45  | 74 %         | ?           |
| 30 |  | 50                       | 0,112        | 0,316 | 67 %         | ?           |
| 31 |  | 26                       | 1.225        | 1.659 | 98 %         | ?           |
| 32 |  | 58                       | -0,159       | 0,409 | 72 %         | < 100 %     |

|    |   |    |       |       |      |      |
|----|---|----|-------|-------|------|------|
| 33 |    | 17 | 0,243 | 1,154 | 93 % | 100% |
| 34 |    | 2  | 0,01  | 1,039 | 92 % | 94%  |
| 35 |   | 56 | 0,917 | 1,262 | 95 % | 100% |
| 36 |  | 48 | 0,217 | 1,197 | 94 % | ?    |
| 37 |  | 44 | 0,705 | 1,281 | 95 % | ?    |
| 38 |  | 69 | 0,403 | 0,759 | 85 % | ?    |

## ***VIII - CONCLUSION DE L'ETUDE***

Nous avons construit et validé un modèle quantitatif 3D des relations structure-activité d'un groupe de flavonoïdes et de boerhavinones inhibiteurs de BCRP. Ce modèle nous permet d'identifier la nature et la position des groupements qui améliorent l'activité, nous donnant des orientations pour concevoir et synthétiser de nouveaux composés actifs. Les tendances observées sont en accord et complètent les résultats déjà obtenus par d'autres études [73, 74]. De nouvelles structures dont les prévisions d'activité sont élevées ou qui nous permettront d'affiner le modèle, sont en cours de synthèse et de test. Un brevet va être déposé pour protéger certains des composés utilisés dans les modèles ainsi que ceux issus de cette étude qui se révéleront actifs.

Le test biologique à concentration constante et la transformation Logit, nécessaire pour traiter le jeu de données initial, s'avèrent pénalisants pour la construction et surtout l'exploitation des modèles QSAR. De nouveaux tests biologiques nous permettront d'éviter ce handicap et d'améliorer les prévisions.

## D - CONCLUSION

La qualité des modèles QSAR est sensible à de nombreux facteurs. La nature et la précision des tests biologiques, tout comme la résolution et la détermination structurale des composés testés conditionnent largement la réussite de ce type d'étude. Le traitement des données brutes est également important. Il est souhaitable de normaliser au mieux la distribution des données sans modifier les informations qu'elles contiennent.

Les QSAR-3D de type CoMFA ou CoMSIA exigent l'alignement préalable des structures dans une conformation particulière : celle qu'adopte chaque composé lorsqu'il forme un complexe actif avec la cible. Si l'on ne dispose pas de données cristallographiques de référence, de résultats de calculs d'arrimage ou d'une série très rigide, alors l'alignement des conformères actifs est très difficile. Nous avons pu effectuer cet alignement sans difficulté du fait de la rigidité des composés de notre série mais cette étape est souvent limitante. Le développement de méthodes robustes d'alignement permettrait d'inclure à nos modèles les composés plus flexibles ou plus originaux dont nous disposons.

La création de scripts automatisant la production des modèles et systématisant leur validation nous a permis de réduire considérablement les temps de calculs. Leur simplicité d'utilisation facilite la mise en œuvre de cette technique et permet de se concentrer sur l'analyse des résultats.

Enfin, l'utilisation de champs d'interaction supplémentaires, décrivant notamment les sites de fixation de molécules d'eau liantes, pourraient enrichir nos modèles.

## **E - PARTIE EXPERIMENTALE**

### ***I - CALCUL DES DESCRIPTEURS COMSIA***

Le champ d'indices de similarité moléculaire stérique (S) a été calculé dans une matrice dont la position, l'orientation et les dimensions ont été automatiquement ajustées par le logiciel. La boîte s'étend à 4 Å au delà de la surface des molécules superposées. L'espacement entre les points de la matrice est de 1 Å. Les champs électrostatique (E), hydrophobe (H), donneur de liaison hydrogène (D), accepteur de liaison hydrogène (A) ont ensuite été calculés avec la même matrice. Le facteur d'atténuation pour le calcul de chaque champ est de 0,3. La sonde employée a les caractéristiques suivantes : rayon 1 Å, charge de +1, hydrophobicité +1, donneur de liaison hydrogène +1 et accepteur +1.

### ***II - CONSTRUCTION ET LA VALIDATION DES MODELES***

L'ensemble des calculs de modèles sans validation, avec validation croisée et de modèles centrés ont été réalisés par l'exécution du script d'automatisation QBF1.

Les commentaires et indications présents dans le code sont rédigés en langue anglaise de sorte que le script soit utilisable par le plus grand nombre. QBF1 ne combine que des marqueurs de même origine c'est-à-dire soit des descripteurs CoMFA, soit des descripteurs CoMSIA. Selon la nature des descripteurs employés, QBF1 fait une mise à l'échelle de type CoMFA\_Standard pour les champs CoMFA et AUTOSCALE pour les champs CoMSIA. Ce script intègre l'ensemble de la procédure de validation en calculant :

- Un modèle de validation croisée du jeu d'apprentissage par LOO ;
- Un modèle sans validation du jeu d'apprentissage avec le nombre optimal de composantes ;
- Une prévision d'activité pour chaque composé du jeu de test correspondant.

Cette procédure est appliquée à toutes les combinaisons de descripteurs disponibles et pour tous les jeux d'apprentissage fournis.

Le script est paramétrable par l'utilisateur grâce à des variables figurant en tête du code. L'utilisateur doit :

- donner un nom à l'analyse ;
- indiquer la position des colonnes contenant la variable cible et les descripteurs ;
- indiquer le nombre maximal de composantes ;
- indiquer si les modèles doivent être centrés ou non ;
- indiquer le seuil de filtrage des colonnes ;
- définir les jeux d'apprentissage.

Le jeu d'apprentissage numéro 1 contient toutes les molécules. Il n'y a pas de jeu de test correspondant.

Le script extrait ou calcule les paramètres caractéristiques de chaque modèle et les enregistre dans un fichier de sortie. Le fichier de sortie est au format texte et son nom est formaté de la façon suivante :

```
qbf1_NOM DE L'ETUDE.log
```

Le nom de l'analyse est celui fourni par l'utilisateur en tête de script.

Par ailleurs, chaque calcul de PLS est enregistré par SYBYL dans un fichier dont l'extension est « pls ». Le nom du fichier doit être précisé par l'utilisateur. QBF1 génère automatiquement un nom de fichier formaté pour chaque PLS. Le format de ce nom est le suivant :

```
qbf1_NOM DE L'ETUDE_s xxmyy_[y0_]type.pls
```

Les variables *xx* et *yy* sont respectivement le numéro du jeu d'apprentissage et le numéro de la combinaison de descripteurs utilisés. Le numéro des combinaisons de descripteurs sont indiqués dans le fichier de sortie de QBF1. La chaîne « y0\_ » est insérée dans le nom du fichier s'il s'agit d'un modèle centré. La variable *type* est remplacée par la chaîne « nv » s'il s'agit d'une PLS sans validation croisée et par la chaîne « loo » s'il s'agit d'une validation croisée par LOO. Ce format de nom de fichier permet de retrouver facilement n'importe quel modèle calculé par QBF1. Ce point est important car le nombre de modèles calculés par le script peut être de plusieurs centaines.

Le code est donné dans l'Annexe II p.157. Un exemple de fichier de sortie est fourni dans Annexe III p. 161.

### ***III - TEST DE HASARDISATION***

Pour le test de hasardisation des réponses, aucune procédure automatique n'est fournie dans le module CoMFA\_STANDARD. Nous avons donc écrit un autre script SPL baptisé QYR.spl (*Qsar Y-Randomisation*) pour effectuer cette tâche fastidieuse et répétitive.

Ce second script crée une nouvelle colonne dans le MSS, dans laquelle il recopie les valeurs de la variable cible mais en les réaffectant de façon aléatoire aux composés. Le modèle de validation croisée et le modèle sans validation sont dérivés avec cette nouvelle variable cible, pour la combinaison de marqueurs désirée et pour le jeu d'apprentissage voulu. L'opération est répétée autant de fois que l'utilisateur l'indique et les paramètres caractéristiques de chaque modèle calculé sont enregistrés dans un fichier de sortie. Le format du nom de fichier de sortie est le suivant :

```
qyr_NOM DE L'ETUDE.log
```

Le début du script est modifiable par l'utilisateur de façon à paramétrer l'exécution. QYR ne conserve pas les modèles calculés. Les fichiers « pls » sont effacés au fur et à mesure de la progression des calculs. Seuls les paramètres descriptifs des modèles sont conservés dans le fichier de sortie. Le script de hasardisation est donné dans Annexe IV p.162.

## F - BIBLIOGRAPHIE

1. Hansch, C. and T. Fujita, *r-s-p Analysis; method for the correlation of biological activity and chemical structure*. Journal of the American Chemical Society, 1964. **86**(8): p. 1616-26.
2. Box, G.E.P. and C.D. R., *An analysis of distributions*. Journal of the royal statistical society, serie B, 1964. **26**: p. 211 - 243.
3. Armitage, P. and G. Berry, *Statistical Methods in Medical Research* 3rd edition ed. 1994: Blackwell.
4. Hall, L.H. and L.B. Kier, *The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling*. Reviews in Computational Chemistry, 1991. **2**: p. 367-422.
5. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*. Journal of Medicinal Chemistry, 1985. **28**(7): p. 849-57.
6. Cramer, R.D., III, D.E. Patterson, and J.D. Bunce, *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins*. Journal of the American Chemical Society, 1988. **110**(18): p. 5959-67.
7. Klebe, G., U. Abraham, and T. Mietzner, *Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity*. Journal of Medicinal Chemistry, 1994. **37**(24): p. 4130-46.
8. Robert, D., L. Amat, and R. Carbo-Dorca, *Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: prediction of the corticosteroid-binding globulin binding affinity for a steroid family*. Journal of chemical information and computer sciences FIELD Publication Date:1999, 1999. **39**(2): p. 333-44. FIELD Reference Number: FIELD Journal Code:7505012 FIELD Call Number:.
9. Amat, L., E. Besalu, R. Carbo-Dorca, and R. Ponec, *Identification of active molecular sites using quantum-self-similarity measures*. Journal of chemical information and computer sciences FIELD Publication Date:2001, 2001. **41**(4): p. 978-91. FIELD Reference Number: FIELD Journal Code:7505012 FIELD Call Number:.
10. Bursi, R., T. Dao, T. Van Wijk, M. de Gooyer, E. Kellenbach, and P. Verwer, *Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities*. Journal of Chemical Information and Computer Sciences, 1999. **39**(5): p. 861-867.
11. Turner, D.B., P. Willett, A.M. Ferguson, and T. Heritage, *Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application*. Journal of Computer-Aided Molecular Design, 1997. **11**(4): p. 409-422.
12. Tuppurainen, K., M. Viisas, R. Laatikainen, and M. Peraekylae, *Evaluation of a Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set*. Journal of Chemical Information and Computer Sciences, 2002. **42**(3): p. 607-613.
13. Norden, B., U. Edlund, D. Johnels, and S. Wold, *Simplified carbon-13 NMR parameters related to the carcinogenic potency of polycyclic aromatic hydrocarbons*. Quantitative Structure-Activity Relationships, 1983. **2**(2): p. 73-6.
14. Aoyama, T., Y. Suzuki, and H. Ichikawa, *Neural networks applied to quantitative structure-activity relationship analysis*. Journal of medicinal chemistry, 1990. **33**(9): p. 2583-90.

15. Andrea, T.A. and H. Kalayeh, *Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors*. Journal of Medicinal Chemistry, 1991. **34**(9): p. 2824-36.
16. Dunn, W.J., III, *Quantitative structure-activity relationships (QSAR)*. Chemometrics and Intelligent Laboratory Systems, 1989. **6**(3): p. 181-90.
17. Clark, M. and R.D. Cramer, III, *The probability of chance correlation using partial least squares (PLS)*. Quantitative Structure-Activity Relationships, 1993. **12**(2): p. 137-45.
18. Wold, S. and L. Eriksson, *Statistical validation of QSAR results. Validation tools*. Methods and Principles in Medicinal Chemistry, 1995. **2**: p. 309-18.
19. van der Voet, H., *Comparing the predictive accuracy of models using a simple randomization test*. Chemometrics and Intelligent Laboratory Systems, 1994. **25**(2): p. 313-23.
20. Peterson, S.D., W. Schaal, and A. Karlen, *Improved CoMFA Modeling by Optimization of Settings*. Journal of Chemical Information and Modeling, 2006. **46**(1): p. 355-364.
21. Golbraikh, A., M. Shen, Z. Xiao, Y.-D. Xiao, K.-H. Lee, and A. Tropsha, *Rational selection of training and test sets for the development of validated QSAR models*. Journal of Computer-Aided Molecular Design, 2003. **17**(2-4): p. 241-253.
22. Golbraikh, A. and A. Tropsha, *Beware of q<sup>2</sup>!* Journal of Molecular Graphics & Modelling, 2002. **20**(4): p. 269-276.
23. Golbraikh, A. and A. Tropsha, *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection*. Journal of Computer-Aided Molecular Design, 2002. **16**(5/6): p. 357-369.
24. Globisch, C., I.K. Pajeva, and M. Wiese, *Structure-activity relationships of a series of tariquidar analogs as multidrug resistance modulators*. Bioorganic & Medicinal Chemistry, 2006. **14**(5): p. 1588-1598.
25. Oprea, T., I., *Cheminformatics in Drug Discovery*. 2005: J. Wiley.
26. Lemmen, C. and T. Lengauer, *Computational methods for the structural alignment of molecules*. Journal of computer-aided molecular design FIELD Publication Date:2000, 2000. **14**(3): p. 215-32. FIELD Reference Number:142 FIELD Journal Code:8710425 FIELD Call Number:.
27. Melani, F., P. Gratteri, M. Adamo, and C. Bonaccini, *Field Interaction and Geometrical Overlap: A New Simplex and Experimental Design Based Computational Procedure for Superposing Small Ligand Molecules*. Journal of Medicinal Chemistry, 2003. **46**(8): p. 1359-1371.
28. Kroonenberg, P.M., W.J. Dunn, and J.J.F. Commandeur, *Consensus Molecular Alignment Based on Generalized Procrustes Analysis*. Journal of Chemical Information and Computer Sciences, 2003. **43**(6): p. 2025-2032.
29. Arakawa, M., K. Hasegawa, and K. Funatsu, *Novel Alignment Method of Small Molecules Using the Hopfield Neural Network*. Journal of Chemical Information and Computer Sciences, 2003. **43**(5): p. 1390-1395.
30. Cho, S.J. and A. Tropsha, *Cross-Validated R<sup>2</sup>-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method To Achieve Consistent Results*. Journal of Medicinal Chemistry, 1995. **38**(7): p. 1060-6.
31. Klebe, G. and U. Abraham, *Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries*. Journal of Computer-Aided Molecular Design, 1999. **13**(1): p. 1-10.
32. Bohm, M., J. St rzebecher, and G. Klebe, *Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa*. Journal of medicinal chemistry, 1999. **42**(3): p. 458-77.

33. Viswanadhan, V.N., A.K. Ghose, G.R. Revankar, and R.K. Robins, *Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics*. Journal of Chemical Information and Computer Sciences, 1989. **29**(3): p. 163-72.
34. Juliano, R.L. and V. Ling, *A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants*. Biochimica et Biophysica Acta, Biomembranes, 1976. **455**(1): p. 152-62.
35. Cole, S.P., G. Bhardwaj, J.H. Gerlach, J.E. Mackie, C.E. Grant, K.C. Almquist, A.J. Stewart, E.U. Kurz, A.M. Duncan, and R.G. Deeley, *Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line*. Science, 1992. **258**(5088): p. 1650-4.
36. Miyake, K., L. Mickley, T. Litman, Z. Zhan, R. Robey, B. Cristensen, M. Brangi, L. Greenberger, M. Dean, T. Fojo, and S.E. Bates, *Molecular cloning of cDNAs which are highly overexpressed in mitoxantrone-resistant cells: demonstration of homology to ABC transport genes*. Cancer research, 1999. **59**(1): p. 8-13.
37. Doyle, L.A., W. Yang, L.V. Abruzzo, T. Krogmann, Y. Gao, A.K. Rishi, and D.D. Ross, *A multidrug resistance transporter from human MCF-7 breast cancer cells*. Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(26): p. 15665-70.
38. Allikmets, R., L.M. Schriml, A. Hutchinson, V. Romano-Spica, and M. Dean, *A human placenta-specific ATP-binding cassette gene (ABCP) on chromosome 4q22 that is involved in multidrug resistance*. Cancer research, 1998. **58**(23): p. 5337-9.
39. Ozvegy, C., T. Litman, G. Szakacs, Z. Nagy, S. Bates, A. Varadi, and B. Sarkadi, *Functional characterization of the human multidrug transporter, ABCG2, expressed in insect cells*. Biochemical and biophysical research communications, 2001. **285**(1): p. 111-7.
40. Xu, J., Y. Liu, Y. Yang, S. Bates, and J.-T. Zhang, *Characterization of Oligomeric Human Half-ABC Transporter ATP-binding Cassette G2*. Journal of Biological Chemistry, 2004. **279**(19): p. 19781-19789.
41. Doyle, L.A. and D. Ross Douglas, *Multidrug resistance mediated by the breast cancer resistance protein BCRP (ABCG2)*. Oncogene, 2003. **22**(47): p. 7340-58.
42. Van der Kolk, D.M., E. Vellenga, G.L. Scheffer, M. Muller, S.E. Bates, R.J. Scheper, and E.G.E. De Vries, *Expression and activity of breast cancer resistance protein (BCRP) in de novo and relapsed acute myeloid leukemia*. Blood, 2002. **99**(10): p. 3763-3770.
43. Sargent, J.M., C.J. Williamson, M. Maliapaard, A.W. Elgie, R.J. Scheper, and C.G. Taylor, *Breast cancer resistance protein expression and resistance to daunorubicin in blast cells from patients with acute myeloid leukaemia*. British journal of haematology, 2001. **115**(2): p. 257-62.
44. Kanzaki, A., M. Toi, K. Nakayama, H. Bando, M. Mutoh, T. Uchida, M. Fukumoto, and Y. Takebayashi, *Expression of multidrug resistance-related transporters in human breast carcinoma*. Japanese journal of cancer research : Gann, 2001. **92**(4): p. 452-8.
45. Ross, D.D., J.E. Karp, T.T. Chen, and L.A. Doyle, *Expression of breast cancer resistance protein in blast cells from patients with acute leukemia*. Blood, 2000. **96**(1): p. 365-8.
46. Friedrich, R.E., C. Punke, and A. Reymann, *Expression of multi-drug resistance genes (mdr1, mrp1, bcrp) in primary oral squamous cell carcinoma*. In Vivo, 2004. **18**(2): p. 133-147.
47. Steinbach, D., W. Sell, A. Voigt, J. Hermann, F. Zintl, and A. Sauerbrey, *BCRP gene expression is associated with a poor response to remission induction therapy in childhood acute myeloid leukemia*. Leukemia : official journal of the Leukemia Society

- of America, Leukemia Research Fund, U.K, 2002. **16**(8): p. 1443-7.
48. Leslie, E.M., Q. Mao, C.J. Oleschuk, R.G. Deeley, and S.P.C. Cole, *Modulation of multidrug resistance protein 1 (MRP1/ABCC1) transport and ATPase activities by interaction with dietary flavonoids*. *Molecular Pharmacology*, 2001. **59**(5): p. 1171-1180.
  49. Cisternino, S., C. Mercier, F. Bourasset, F. Roux, and J.-M. Scherrmann, *Expression, Up-Regulation, and Transport Activity of the Multidrug-Resistance Protein Abcg2 at the Mouse Blood-Brain Barrier*. *Cancer Research*, 2004. **64**(9): p. 3296-3301.
  50. Cooray, H.C., C.G. Blackmore, L. Maskell, and M.A. Barrand, *Localisation of breast cancer resistance protein in microvessel endothelium of human brain*. *NeuroReport*, 2002. **13**(16): p. 2059-2063.
  51. Maliepaard, M., G.L. Scheffer, I.F. Faneyte, M.A. van Gastelen, A.C. Pijnenborg, A.H. Schinkel, M.J. van De Vijver, R.J. Scheper, and J.H. Schellens, *Subcellular localization and distribution of the breast cancer resistance protein transporter in normal human tissues*. *Cancer research*, 2001. **61**(8): p. 3458-64.
  52. Kruijtzter, C.M.F., J.H. Beijnen, H. Rosing, W.W. ten Bokkel Huinink, M. Schot, R.C. Jewell, E.M. Paul, and J.H.M. Schellens, *Increased oral bioavailability of topotecan in combination with the breast cancer resistance protein and P-glycoprotein inhibitor GF120918*. *Journal of Clinical Oncology*, 2002. **20**(13): p. 2943-2950.
  53. Jonker, J.W., J.W. Smit, R.F. Brinkhuis, M. Maliepaard, J.H. Beijnen, J.H. Schellens, and A.H. Schinkel, *Role of breast cancer resistance protein in the bioavailability and fetal penetration of topotecan*. *Journal of the National Cancer Institute*, 2000. **92**(20): p. 1651-6.
  54. Harborne, J.B. and C.A. Williams, *Advances in flavonoid research since 1992*. *Phytochemistry*, 2000. **55**(6): p. 481-504.
  55. Hertog, M.G., *Epidemiological evidence on potential health properties of flavonoids*. *The Proceedings of the Nutrition Society*, 1996. **55**(1B): p. 385-97.
  56. Hertog, M.G., E.J. Feskens, and D. Kromhout, *Antioxidant flavonols and coronary heart disease risk*. *Lancet*, 1997. **349**(9053): p. 699.
  57. Hertog, M.G., P.M. Sweetnam, A.M. Fehily, P.C. Elwood, and D. Kromhout, *Antioxidant flavonols and ischemic heart disease in a Welsh population of men: the Caerphilly Study*. *The American journal of clinical nutrition*, 1997. **65**(5): p. 1489-94.
  58. Lee, H.P., L. Gourley, S.W. Duffy, J. Esteve, J. Lee, and N.E. Day, *Dietary effects on breast-cancer risk in Singapore*. *Lancet*, 1991. **337**(8751): p. 1197-200.
  59. Potter, S.M., J.A. Baum, H. Teng, R.J. Stillman, N.F. Shay, and J.W. Erdman, Jr., *Soy protein and isoflavones: their effects on blood lipids and bone density in postmenopausal women*. *The American journal of clinical nutrition*, 1998. **68**(6 Suppl): p. 1375S-1379S.
  60. Matsuo, M., N. Sasaki, K. Saga, and T. Kaneko, *Cytotoxicity of flavonoids toward cultured normal human cells*. *Biological & Pharmaceutical Bulletin*, 2005. **28**(2): p. 253-259.
  61. Middleton, E., Jr., C. Kandaswami, and T.C. Theoharides, *The effects of plant flavonoids on mammalian cells: implications for inflammation, heart disease, and cancer*. *Pharmacological reviews*, 2000. **52**(4): p. 673-751.
  62. Hsiu, S.-L., Y.-C. Hou, Y.-H. Wang, C.-W. Tsao, S.-F. Su, and P.-D.L. Chao, *Quercetin significantly decreased cyclosporin oral bioavailability in pigs and rats*. *Life Sciences*, 2002. **72**(3): p. 227-235.
  63. Dresser, G.K., D.G. Bailey, B.F. Leake, U.I. Schwarz, P.A. Dawson, D.J. Freeman, and R.B. Kim, *Fruit juices inhibit organic anion transporting polypeptide-mediated drug uptake to decrease the oral availability of fexofenadine*. *Clinical Pharmacology & Therapeutics* (St. Louis, MO, United States), 2002. **71**(1): p. 11-20.
  64. Ruschitzka, F., P.J. Meier, M. Turina, T.F. Luscher, and G. Noll, *Acute heart transplant rejection due to Saint John's wort*. 2000: ENGLAND: United Kingdom. p. 548-9.

65. Di Pietro, A., G. Conseil, J.M. Perez-Victoria, G. Dayan, H. Baubichon-Cortay, D. Trompier, E. Steinfels, J.M. Jault, H. De Wet, M. Maitrejean, G. Comte, A. Boumendjel, A.M. Mariotte, C. Dumontet, D.B. McIntosh, A. Goffeau, S. Castanys, F. Gamarro, and D. Barron, *Modulation by flavonoids of cell multidrug resistance mediated by P-glycoprotein and related ABC transporters*. Cellular and Molecular Life Sciences, 2002. **59**(2): p. 307-322.
66. Conseil, G., H. Baubichon-Cortay, G. Dayan, J.M. Jault, D. Barron, and A. Di Pietro, *Flavonoids: a class of modulators with bifunctional interactions at vicinal ATP- and steroid-binding sites on mouse P-glycoprotein*. Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(17): p. 9831-6.
67. Bois, F., C. Beney, A. Boumendjel, A.M. Mariotte, G. Conseil, and A. Di Pietro, *Halogenated chalcones with high-affinity binding to P-glycoprotein: potential modulators of multidrug resistance*. Journal of medicinal chemistry, 1998. **41**(21): p. 4161-4.
68. Shapiro, A.B. and V. Ling, *Effect of quercetin on Hoechst 33342 transport by purified and reconstituted P-glycoprotein*. Biochemical pharmacology, 1997. **53**(4): p. 587-96.
69. Castro, A.F. and G.A. Altenberg, *Inhibition of drug transport by genistein in multidrug-resistant cells expressing P-glycoprotein*. Biochemical pharmacology, 1997. **53**(1): p. 89-93.
70. Wu, C.-P., A.M. Calcagno, S.B. Hladky, S.V. Ambudkar, and M.A. Barrand, *Modulatory effects of plant phenols on human multidrug-resistance proteins 1, 4 and 5 (ABCC1, 4 and 5)*. FEBS Journal, 2005. **272**(18): p. 4725-4740.
71. Trompier, D., H. Baubichon-cortay, X.B. Chang, M. Maitrejean, D. Barron, J.R. Riordan, and A. Di Pietro, *Multiple flavonoid-binding sites within multidrug resistance protein MRP1*. Cellular and Molecular Life Sciences, 2003. **60**(10): p. 2164-2177.
72. Nguyen, H., S. Zhang, and M.E. Morris, *Effect of flavonoids on MRP1-mediated transport in Panc-1 cells*. Journal of Pharmaceutical Sciences, 2003. **92**(2): p. 250-257.
73. Ahmed-Belkacem, A., A. Pozza, F. Munoz-Martinez, S.E. Bates, S. Castanys, F. Gamarro, A. Di Pietro, and J.M. Perez-Victoria, *Flavonoid Structure-Activity Studies Identify 6-Prenylchrysin and Tectochrysin as Potent and Specific Inhibitors of Breast Cancer Resistance Protein ABCG2*. Cancer Research, 2005. **65**(11): p. 4852-4860.
74. Zhang, S., X. Yang, R.A. Coburn, and M.E. Morris, *Structure activity relationships and quantitative structure activity relationships for the flavonoid-mediated inhibition of breast cancer resistance protein*. Biochemical Pharmacology, 2005. **70**(4): p. 627-639.
75. Kramar, A., D. Faraggi, A. Fortune, and B. Reiser, *mROC: a computer program for combining tumour markers in predicting disease states*. Computer methods and programs in biomedicine, 2001. **66**(2-3): p. 199-207.

## **G - ANNEXES**

**Annexe I: Table de Fischer pour un risque de 5 %.**

## Annexe II : Code du script QBF1.

Script SPL calculant les modèles, centrés ou non, de validation croisée LOO et sans validation pour toutes les combinaisons de descripteurs et pour tous les jeux d'apprentissage souhaités.

Nom du fichier de script : QBF1.spl

```
# script : QsarBestFit.spl
# Version: 0.9
# Author : Antoine FORTUNE - DPM UMR5063 UJF/CNRS
# Date : 09/2006
# Task : Generate PLS models for CoMFA or CoMSIA
# with both nv and loo procedures
# for any training sets and all descriptors
# combinations.
# loo= leave on out, nv= no validation
# Can force models with Y-intercept=0
# Input :
# Output : RUN_NAME.log file

#####
# USER DEFINED PARAMETERS SECTION
#####
#
# Run name (for outputs filenames)
SETVAR RUN_NAME S_fb02logit
# MSS table [path]filename.tbl
SETVAR TABLE ../fb2.tbl
# Y-values column
SETVAR COL_Y 2
# First descriptor column
SETVAR FST_DESC 4
# Nb of descriptors. MIN 1 MAX 5
SETVAR NB_DESC 5
# Max number of components (MAX : nb molecules / 3).
SETVAR COMPMAX 8
# Force Y0=0 i.e. models without constant. YES or NO
SETVAR center NO
# Column filtering level (default 1.0 kcal)
SETVAR col_filter 1.0
# Debug level : 0 quiet 1 verbose
SETVAR DEBUG 0
#
# Definition of training and test sets
# -----
#
# Use %array_set($a_TRsets SET_NUMBER
"ROWS_EXPRESSION")
#
# SET_NUMBER : incremental integer
# set #1 already defined as "all lines" (*)
#
# ROW_EXPRESSION :
# Comma separated lines number. a:b = range a to b
# ex : 1,2,5:7 for 1,2,5,6,7
#
# Define only training sets.
# Corresponding test sets deduced automatically
#
# !! DO NOT MODIFY NEXT 2 LINES !!
#
setvar a_TRsets %array_create()
%array_set($a_TRsets 1 "*")
#
# Ex : %array_set($a_TRsets 2 "1,2,5:7,9:15,17:19")
#
%array_set($a_TRsets 2 "3:14,16,18,20:27")

%array_set($a_TRsets 3 "1:3,5:8,10:21,23:25")
%array_set($a_TRsets 4 "2,4:7,9,11:20,23,25:27")
%array_set($a_TRsets 5 "1:3,5:9,11:17,20,22:27")
%array_set($a_TRsets 6 "1,3:8,11:14,16,18:25,27")
%array_set($a_TRsets 7 "1:5,7:10,12:19,21:23,26")
%array_set($a_TRsets 8
"1,2,5:10,12:15,17:19,21:26")
%array_set($a_TRsets 9
"1:6,8,9,11,12,14:16,18:20,23:27")
%array_set($a_TRsets 10 "1:4,8:11,13,15:22,24:27")
#
# END USER SECTION
#####
##
#
# Do not modify the following code
# unless you know what you do
#
# #####
# INTERNAL VARIABLES SETUP
# #####
#
setvar qbf_version 0.9
IF %STREQ(%TABLE_NAME() "")
TABLE OPEN $TABLE
ENDIF
TABLE DEFAULT %TABLE_NAME()
setvar nb_mol %table(* row count)
setvar desc_type %table($FST_DESC column
column_type)
setvar a_logmodel %array_create()
setvar nb_sets %array_size($a_TRsets)

# Results log filenames
SETVAR LOGFILE %cat(qbf1_ $RUN_NAME .log)
SETVAR SUMLOGFILE %cat(qbf1_sum_ $RUN_NAME .log)

# SCALING METHOD
switch $desc_type
case COMFA)
tailor set PLS SCALING_METHOD COMFA_STD ||
;;
case COMSIA)
tailor set PLS SCALING_METHOD AUTOSCALE ||
;;
endswitch

# CENTERED MODELS FILENAME EXTENSION
switch $center
case YES)
setvar str_y0 _y0_
;;
case NO)
setvar str_y0 _
;;
endswitch

# COLUMN FILTERING THRESHOLD
```

```

TAILOR SET QSAR_ANALYSIS MINIMUM_SIGMA $col_filter
||
#
# Generation of descriptors combinations
# #####

setvar colmax  $("%FST_DESC" + "$NB_DESC" -1)$ 
setvar a_coldesc  $\%array\_create()$ 
setvar combi

for i in  $\%range($FST_DESC $colmax)$ 
  setvar jmin  $("%FST_DESC" + 1)$ 
  if  $\%gt("$jmin" "$colmax")$ 
    setvar jmin $colmax
  endif
  for j in 0  $\%range($jmin $colmax)$ 
    if  $\%and("\%neq("$j" 0)" "\%lteq("$j" "$i)")$ 
      continue
    endif
    setvar kmin  $("%FST_DESC" + 2)$ 
    if  $\%gt("$kmin" "$colmax")$ 
      setvar kmin $colmax
    endif
    for k in 0  $\%range($kmin $colmax)$ 
      if  $\%and("\%neq("$k" 0)" "\%or("\%eq("$j" 0)"$ 
 $\%lteq("$k" "$j)")$ )"
        continue
      endif
      setvar lmin  $("%FST_DESC" + 3)$ 
      if  $\%gt("$lmin" "$colmax")$ 
        setvar lmin $colmax
      endif
      for l in 0  $\%range($lmin $colmax)$ 
        if  $\%and("\%neq("$l" 0)" "\%or("\%eq("$k" 0)"$ 
 $\%lteq("$l" "$k)")$ )"
          continue
        endif
        for m in 0 $colmax
          if  $\%and("\%neq("$m" 0)" "\%or("\%eq("$l" 0)"$ 
 $\%lteq("$m" "$l)")$ )"
            continue
          endif
          setvar col 0

          setvar combi $i
          if  $\%neq("$j" 0)$ 
            setvar combi  $\%cat($combi , $j)$ 
          endif
          if  $\%neq("$k" 0)$ 
            setvar combi  $\%cat($combi , $k)$ 
          endif
          if  $\%neq("$l" 0)$ 
            setvar combi  $\%cat($combi , $l)$ 
          endif
          if  $\%neq("$m" 0)$ 
            setvar combi  $\%cat($combi , $m)$ 
          endif
           $\%array\_push($a\_coldesc "$combi")$ 
        endfor
      endfor
    endfor
  endfor
endfor

#
# SETUP END
# #####
#
# #####
# LOGGING RUN PARAMETERS IN LOGFILE
#

echo # QsarBestFit.spl  $\%cat(v. $qbf\_version)$  logfile
> $LOGFILE
echo # Date :  $\%time()$  >> $LOGFILE
echo >> $LOGFILE

echo Job name : $RUN_NAME >> $LOGFILE
echo Used table : $TABLE >> $LOGFILE
echo Dependent variable column : $COL_Y >>
$LOGFILE
echo Independents variables columns : $FST_DESC -
\  

 $\%math("%FST_DESC" + "$NB_DESC" -1 )$  >> $LOGFILE
echo Descriptors type $desc_type >> $LOGFILE
echo Centered : $center >> $LOGFILE
if  $\%streql($center YES)$ 
  echo Y0=0 : No constant in models.>> $LOGFILE
else
  echo Y0 <> 0 : constant used in models. >>
$LOGFILE
endif

echo >> $LOGFILE

#
# #####
# CONTROLS
# #####
#
setvar error 0
if  $\%or(\%lt($NB_DESC 1) \%gt($NB_DESC 5))$ 
  echo >> $LOGFILE
  echo PARAMETER ERROR ! >> $LOGFILE
  echo NB_DESC = $NB_DESC
  echo NB_DESC MUST be bitween 1 and 5
  echo >> $LOGFILE
  incr error
endif

if  $\%and(\%lteq($COL_Y "\%MATH("%FST_DESC" +$ 
 $\%NB_DESC" - 1)) \%gteq($COL_Y $FST_DESC))$ 
  echo >> $LOGFILE
  echo PARAMETER ERROR ! >> $LOGFILE
  echo NB_DESC = $NB_DESC
  echo >> $LOGFILE
  echo NB_DESC MUST be bitween 1 and 5
  echo >> $LOGFILE
  incr error
endif

if  $\%gt($NB_DESC 1)$ 
  for i in  $\%range(1 $NB_DESC)$ 
    setvar tmp  $\%table(\%math($FST_DESC + $i - 1)$ 
column column_type)
    if  $\%not(\%streql("$desc\_type" "$tmp"))$ 
      echo >> $LOGFILE
      echo PARAMETER ERROR ! >> $LOGFILE
      echo First descriptor type (col : $FST_DESC TYPE
: $desc\_type)\
differs $i th (col :  $\%math($FST_DESC + $i - 1)$ 
TYPE : $tmp) >> $LOGFILE
      echo All descriptors MUST have same type >>
$LOGFILE
      echo >> $LOGFILE
      incr error
    endif
  endfor
endif

if  $\%gt($error 0)$ 
  echo QBF script stopped ... >> $LOGFILE
  echo >> $LOGFILE
  echo
  echo ERROR ! ERROR ! ERROR ! ERROR ! ERROR !
  echo ERROR ! ERROR ! ERROR !
  echo
  echo All results following should not be
considered ! >> $LOGFILE
  goto end
  return
endif
#
# END CONTROLS
# #####

```

```

#
#
# #####
# LOGGING OTHER PARAMETERS IN LOG FILE
#
echo Training sets definition : >> $LOGFILE
echo "%array_print($a_TRsets)" >> $LOGFILE
echo # models : %array_size($a_coldesc) >> $LOGFILE

echo >> $LOGFILE
echo Results summary >> $LOGFILE
echo >> $LOGFILE

if %streql($center YES)
  echo ##### >> $LOGFILE
  echo !! CENTERED MODELS !! Forced Y0=0 >> $LOGFILE
  echo ##### >> $LOGFILE
  echo >> $LOGFILE
else
  echo Models NOT centered : constant used (Y0 <> 0).
  >> $LOGFILE
  echo >> $LOGFILE
endif

echo model_name descriptors nbTRmol opt_comp q2cv
sep R2 see y0 F \
q2ext sepext >> $LOGFILE
echo model_name descriptors nbTRmol opt_comp q2cv
sep R2 see F \
q2ext sepext >> $LOGSUMFILE

#
# #####
# STARTING COMPUTATION
# #####
#
#####
# Derive models for each combination of descriptors
# and each training set
#
#####

IF %eq($DEBUG 2)
  for s in %range(1 %array_size($a_TRsets))
    for i in %range(1 %array_size($a_coldesc))
      if %lt($i 10)
        setvar m %cat(0 $i)
      else
        setvar m $i
      endif
      setvar modele %cat(qbf_ $RUN_NAME s $s m $m _loo)
      qsar analysis delete pls $modele
      setvar modele %cat(qbf_ $RUN_NAME s $s m $m _nv)
      qsar analysis delete pls $modele
    endfor
  endfor
ENDIF

for s in %range(1 %array_size($a_TRsets))
  setvar row_sel %array_get($a_TRsets $s)
  table select rows $row_sel
  setvar nb_tr_mol %table({selected()}) row count)
  if %lt($s 10)
    setvar jeu %cat(0 $s)
  else
    setvar jeu $s
  endif

  for i in %range(1 %array_size($a_coldesc))
    echo modele : $i
    setvar col_sel %cat($COL_Y , %array_get($a_coldesc

```

```

$ i))
  if %lt($i 10)
    setvar m %cat(0 $i)
  else
    setvar m $i
  endif

#
# #####
# PLS WITH LOO VALIDATION
# #####
#
# PLS tailor parameters for LOO
#
TAILOR SET PLS \
CENTERING $center \
COMPONENTS $COMPMAX \
CROSSVALIDATION $nb_tr_mol ||
#
# QSAR tailor parameters
#
TAILOR SET QSAR_ANALYSIS MINIMUM_SIGMA 1.0 ||
#
# Execute LOO analysis
#
setvar modelname %cat(qbf1_ $RUN_NAME s $jeu m $m
$str_y0 loo)
echo $modelname

#echo QSAR ANALYSIS DO INTERACTIVE $row_sel
$col_sel PLS $COL_Y | $modelname
QSAR ANALYSIS DO INTERACTIVE $row_sel $col_sel PLS
$COL_Y | $modelname \
> $NULLDEV
#
# Get results
# !! fishing values with %arg in bugsfull
%QSAR_RETRIEVE function results !!
# sybyl version 7.0
#
setvar opt_comp %arg(5
%QSAR_RETRIEVE(OPTIMAL_COMPONENTS))
setvar q2cv %arg(%math(8 + "$opt_comp")
%qsar_retrieve(R_SQUARED))
setvar sep %arg(%math(8 + 2 * "$opt_comp")
%qsar_retrieve(STANDARD_ERRORS))
#
# #####
# PLS WITH NO VALIDATION
# #####
#
# PLS tailor parameters for NV
#
TAILOR SET PLS \
CENTERING $center \
COMPONENTS $opt_comp \
CROSSVALIDATION 0 ||
#
# Execute NV analysis
#
setvar modelname %cat(qbf1_ $RUN_NAME s $jeu m $m
$str_y0 nv)
echo $modelname

QSAR ANALYSIS DO INTERACTIVE $row_sel $col_sel \
PLS $COL_Y | $modelname \
> /dev/null
#
# Get results
#
setvar R2 %arg(7 %qsar_retrieve(R_SQUARED))
setvar see %arg(7 %qsar_retrieve(STANDARD_ERRORS))
setvar y0 %arg(7 %qsar_retrieve(OFFSET))

# F test
  echo n1= $opt_comp n2= %math($nb_tr_mol -
$opt_comp - 1)
  setvar d %math(%math($nb_tr_mol - $opt_comp - 1)

```

```

/ $opt_comp)
setvar f %math($R2 / %math(1 - $R2))
setvar FR2 %math($d * $f)
echo F = $FR2

#
# #####
# EXTERNAL VALIDATION
# #####
#
if %lt($nb_tr_mol $nb_mol)
echo $modelname EV
TABLE COLUMN_APPEND PREDICT %table_name()\
%qsar_ana_file(%cat($modelname .pls)) pred
TABLE EVALUATE NEW_ONLY ~($row_sel) pred
table update
QSAR ANALYSIS RSQUARED_PRED $COL_Y * pred
~($row_sel) > qbf.tmp
setvar f_tmp %open(qbf.tmp)
setvar tmp %read($f_tmp)
setvar tmp %read($f_tmp)
setvar q2ext %arg(4 $tmp)
setvar tmp %read($f_tmp)
setvar press_str %arg(8 $tmp)
#echo press_str $press_str
setvar press %split( " " $press_str)
#echo press $press
setvar v %math($press / %math($nb_tr_mol -
$opt_comp - 1))
#echo v $v
setvar sepept %sqrt($v)

%close($f_tmp)
setvar col_sel %cat($COL_Y , %array_get($a_coldesc
$i))
TABLE MODIFY COLUMN DELETE pred
else
setvar q2ext -
setvar sepept -
endif
echo
echo

#
# #####
# RECORD RESULTS SUMMARY
# #####
#
setvar combidesc %array_get($a_coldesc $i)
if %lt($s 10)
setvar set %cat(0 $s)
else
setvar set $s
endif
if %lt($i 10)
setvar mod %cat(0 $i)
else
setvar mod $i
endif

setvar tmp_str %cat(s $set m $mod " "

%array_get($a_coldesc $i) " "\
$nb_tr_mol " " $opt_comp " " $q2cv " " $sepept " "\
$R2 " " $see " " $y0 " " $FR2 " " $q2ext " "
$sepept)
%array_push($a_logmodel "$tmp_str")

# FIN BOUCLE i
endfor
# FIN BOUCLE s
endfor
#
# END COMPUTATION
# #####
#
# #####
# DUMP RESULTS IN LOG FILES
#
for i in %range(1 %array_size($a_logmodel))
echo %array_get($a_logmodel $i) >> $LOGFILE
echo %array_get($a_logmodel $i) >> $SUMLOGFILE
endfor
#
# #####
# FREE MEMORY AND TMP FILE
#
%array_free($a_coldesc)
%array_free($a_logmodel)
%file_delete(/qbf.tmp)

IF %eq($DEBUG 2)
wait 5
for s in %range(1 %array_size($a_TRsets))
for i in %range(1 %array_size($a_coldesc))
if %lt($i 10)
setvar m %cat(0 $i)
else
setvar m $i
endif
setvar modele %cat(qbf_ $RUN_NAME s $s m $m
_loo)
qsar analysis delete pls $modele
setvar modele %cat(qbf_ $RUN_NAME s $s m $m
_nv)
qsar analysis delete pls $modele
endif
endfor
ENDIF

end:
#
# #####
# CLOSE LOGFILE
#
echo >> $LOGFILE
echo %time() >> $LOGFILE
echo QsarBESTFit done >> $LOGFILE
echo
echo done !

```

## Annexe III : Exemple de fichier de sortie du script QBF1.

Les paramètres de calcul sont rappelés en début de fichier. Les paramètres descriptifs de chaque modèle sont rassemblés dans une table en fin de fichier. Dans cette table, les colonnes sont séparées par un espace. Ce format de table peut être facilement importé dans un tableur.

```
# QsarBestFit.spl v.0.9 logfile
# Date : Wed Oct 4 11:12:44 2006

Job name : S_fb02logit
Used table : ../fb2.tbl
Dependent variable column : 2
Independents variables columns : 4 - 8
Descriptors type COMSIA
Centered : NO
Y0 <> 0 : constant used in models.

Training sets definition :
1: 1,3,5:8,10:21,23:25

# models : 31

Results summary

Models NOT centered : constant used (Y0 <> 0).

model_name descriptors nbTRmol opt_comp q2cv sep R2 see y0 F q2ext sepext
s03m01 4 22 3 0.0495659 0.718005 0.574018 0.480686 -0.355736 8.085102 0.658 0.06164414 11.54386
s03m02 4,5 22 5 0.371137 0.61947 0.908446 0.236364 -0.465432 31.75205 0.810 0.048798053 13.642105
s03m03 4,5,6 22 6 0.396501 0.626752 0.955748 0.169717 0.120934 53.994622 0.693 0.064031242 5.6433225
s03m04 4,5,6,7 22 7 0.606064 0.524145 0.992833 0.0706956 0.0985079 277.05678 0.885 0.040443965 15.391304
s03m05 4,5,6,7,8 22 8 0.595392 0.551249 0.994037 0.066919 0.13908 270.88883 0.865 0.045573272 10.412037
s03m06 4,5,6,8 22 8 0.490777 0.618422 0.99495 0.061588 0.188194 320.15717 0.891 0.040950252 13.283257
s03m07 4,5,7 22 8 0.630951 0.526469 0.993524 0.0697392 0.104684 249.3015 0.841 0.049380781 8.5951258
s03m08 4,5,7,8 22 8 0.614052 0.538388 0.992211 0.0764834 0.122838 207.00256 0.813 0.053708616 7.0648396
s03m09 4,5,8 22 6 0.476017 0.584004 0.973334 0.131747 -0.101489 91.252345 0.937 0.028867513 37.18254
s03m10 4,6 22 3 0.0221562 0.728284 0.608585 0.460771 -0.08236 9.3289986 0.472 0.076739096 5.3636363
s03m11 4,6,7 22 6 0.482102 0.580602 0.978733 0.117654 0.227591 115.05302 0.699 0.06340347 5.8056478
s03m12 4,6,7,8 22 8 0.476255 0.627178 0.987039 0.0986639 0.388322 123.75113 0.689 0.069170914 3.6000804
s03m13 4,6,8 22 6 0.418663 0.615136 0.967754 0.144875 0.385606 75.028995 0.764 0.056154548 8.0932202
s03m14 4,7 22 7 0.517343 0.580173 0.981863 0.112465 0.285642 108.27182 0.710 0.064420494 4.8965518
s03m15 4,7,8 22 8 0.517149 0.602195 0.987115 0.0983716 0.466176 124.49064 0.692 0.068836484 3.650974
s03m16 4,8 22 8 0.465212 0.633756 0.984202 0.108925 0.553209 101.23612 0.650 0.073379939 3.0178572
s03m17 5 22 6 0.300986 0.674528 0.957622 0.166084 0.0691322 56.492873 0.570 0.075718778 3.3139535
s03m18 5,6 22 6 0.42032 0.614258 0.96322 0.154726 0.384587 65.471725 0.630 0.070285134 4.2567568
s03m19 5,6,7 22 8 0.609022 0.541885 0.994638 0.0634584 0.119233 301.43357 0.886 0.041970686 12.629386
s03m20 5,6,7,8 22 8 0.605911 0.544036 0.993447 0.0701553 0.105359 246.35302 0.855 0.047312382 9.5818965
s03m21 5,6,8 22 8 0.498094 0.613963 0.994782 0.0625998 0.225996 309.797 0.883 0.042426407 12.263889
s03m22 5,7 22 8 0.640009 0.519968 0.989519 0.0887218 0.124749 153.41746 0.796 0.056021973 6.3406863
s03m23 5,7,8 22 8 0.603568 0.545652 0.988013 0.094883 0.070556 133.93853 0.779 0.058375443 5.7279412
s03m24 5,8 22 6 0.452801 0.596801 0.965267 0.150358 -0.0517501 69.477658 0.916 0.033466401 27.261905
s03m25 6 22 6 0.0427173 0.789364 0.926041 0.219408 1.19379 31.302513 0.135 0.10739336 0.3901734
s03m26 6,7 22 7 0.496936 0.592311 0.983207 0.108218 0.277123 117.09724 0.735 0.061528159 5.5471698
s03m27 6,7,8 22 7 0.503134 0.588652 0.984916 0.102566 0.360357 130.59082 0.680 0.067665142 4.25
s03m28 6,8 22 8 0.422261 0.658714 0.989389 0.0892699 0.554155 151.51797 0.572 0.081145643 2.171729
s03m29 7 22 4 0.376471 0.59842 0.744114 0.383355 0.393998 12.358959 0.644 0.064762007 7.6882024
s03m30 7,8 22 3 0.436988 0.552618 0.75854 0.3619 0.193064 18.848836 0.640 0.063245553 10.666667
s03m31 8 22 4 0.376424 0.598443 0.740405 0.386123 0.0787162 12.121656 0.851 0.041797832 24.27349

Wed Oct 4 11:15:46 2006
QsarBEstFit done
```

## Annexe IV : Code du script QYR.

Le script QYR effectue le test de hasardisation des réponses. Il copie la colonne de la variable dans une nouvelle colonne mais en distribuant les valeurs de façon aléatoire sur l'ensemble des lignes. Un modèle de validation croisée par LOO puis un modèle sans validation sont calculés avec cette nouvelle colonne comme variable cible. Les paramètres  $C_{opt}$ ,  $q^2$ , sep puis  $r^2$ , see et F sont collectés pour chaque modèle et enregistrés dans un fichier de sortie. L'opération est répétée autant de fois qu'indiqué dans la variable NB\_TEST.

```
# script : QsarYRandomisation
# Version: 0.1
# Author : Antoine FORTUNE - DPM UMR5063 UJF/CNRS
# Date   : 10/2006
# Task   : Generate PLS models for CoMFA or CoMSIA
#         with both nv and loo procedures
#         for n randomised Y-value
#         loo= leave on out, nv= no validation
#         Can force models with Y-intercept=0
# Input  :
# Output : see doc at the end of this file

# #####
# USER DEFINED PARAMETERS SECTION
# #####
#
# Run name (for outputs filenames)
SETVAR RUN_NAME fb02_SED2
# MSS table [path]filename.tbl
SETVAR TABLE fb2.tbl
# Y-value column
SETVAR COL_Y 2
# Nb of Y-randomised tests
SETVAR NB_TESTS 100
# Descriptors columns to use
# Ex : %array_set($a_descset 1 "4,5,7")
# DO NOT MODIFY NEXT LINE
setvar a_descset %array_create()
%array_set($a_descset 1 "4,5,7")
# Max number of components (MAX : nb molecules / 3).
SETVAR COMPMAX 8
# Force Y0=0 i.e. no constant in models. YES or NO
SETVAR center NO
# Column filtering level (default 1.0 kcal)
SETVAR col_filter 1.0
# Keep analysis after the run (YES / NO)
setvar KEEP_ANA NO
# Debug level : 0 quiet 1 verbose
SETVAR DEBUG 0
#
# END USER SECTION
#
#####
#
# #####
# INTERNAL VARIABLES SETUP
# #####
#
setvar qyr_version 0.1
IF %STREQ("TABLE_NAME") ""
  TABLE OPEN $TABLE
ENDIF
TABLE DEFAULT %TABLE_NAME()
setvar nb_mol %table(* row count)

setvar fst_desc %split(, %array_get($a_descset
1))
setvar desc_type %table($fst_desc column
column_type)
setvar a_logmodel %array_create()
switch $desc_type
  case COMFA)
    tailor set PLS SCALING_METHOD COMFA_STD ||
    ;;
  case COMSIA)
    tailor set PLS SCALING_METHOD AUTOSCALE ||
    ;;
endswitch
switch $center
  case YES)
    setvar str_y0 _y0_
    ;;
  case NO)
    setvar str_y0 _
    ;;
endswitch

TAILOR SET QSAR_ANALYSIS MINIMUM_SIGMA
$col_filter ||

# Results log filenames
SETVAR LOGFILE %cat(qyr_ $RUN_NAME .log)
#
# SETUP END
# #####
#
# #####
# STARTING COMPUTATION
# #####
#
echo # QsarYRandomisation.spl %cat(v.
$qyr_version) logfile > $LOGFILE
echo # Date : %time() >> $LOGFILE
echo >> $LOGFILE
echo Job name : $RUN_NAME >> $LOGFILE
echo Used table : $TABLE >> $LOGFILE
echo Dependent variable columns : $COL_Y >>
$LOGFILE
echo Independents variables columns :
%array_get($a_descset 1) >> $LOGFILE
echo Descriptors type $desc_type >> $LOGFILE
echo Centered : $center >> $LOGFILE
if %streql($center YES)
  echo Y0=0 : No constant in models.>> $LOGFILE
else
  echo Y0 <> 0 : constant used in models. >>
$LOGFILE
endif
echo >> $LOGFILE
```

```

if %streql($center YES)
  echo ##### >> $LOGFILE
  echo !! CENTERED MODELS !! Forced Y0=0 >> $LOGFILE
  echo ##### >> $LOGFILE
  echo >> $LOGFILE
else
  echo Models NOT centered : constant used (Y0 <> 0).
>> $LOGFILE
  echo >> $LOGFILE
endif

echo run opt_comp q2cv sep R2 see y0 F >> $LOGFILE

#
# #####
# STARTING COMPUTATION
# #####
#
#####
# Derive models for each Y-Randomisation #
#####
#
table select rows *
setvar row_sel *
TABLE COLUMN_APPEND FLOAT YR
setvar col_sel %cat(yr, %array_get($a_descset 1))

for r in %range(1 $NB_TESTS)
  setvar a_yr %array_create()

  for i in %range(1 $nb_mol)
    %array_push($a_yr %rcell($i $COL_Y PRECISE))
  endfor

  for i in %range(1 $nb_mol)
    setvar nbleft %math($nb_mol - $i + 1)
    setvar who %ceil(%math(%rand() * $nbleft))
    setvar rdata %array_get($a_yr $who)
    TABLE MODIFY DATA EXPLICITLY_ENTER $i yr $rdata >>
$NULLDEV
    %array_delete_element($a_yr $who)
  endfor

  TABLE UPDATE
#wait 5

  if %lt($r 10)
    setvar run_num %cat(0 $r)
  else
    setvar run_num $r
  endif
echo -- run $run_num --
#
# #####
# PLS with LOO validation
# #####
#
# PLS tailor parameters for LOO
#
TAILOR SET PLS \
CENTERING $center \
COMPONENTS $COMPMAX \
CROSSVALIDATION $nb_mol ||
#
# QSAR tailor parameters
#
TAILOR SET QSAR_ANALYSIS MINIMUM_SIGMA $col_filter ||
#
# Execute LOO analysis
#
setvar modelname %cat(qyr_ $RUN_NAME yr $run_num
$str_y0 loo)
echo $modelname

QSAR ANALYSIS DO INTERACTIVE $row_sel $col_sel PLS yr
| $modelname \
> $NULLDEV
#
# Get results
# !! fishing values with %arg in bugsfull
%QSAR_RETRIEVE function results !!
# sybyl version 7.0
#
setvar opt_comp %arg(5
%QSAR_RETRIEVE(OPTIMAL_COMPONENTS))
setvar q2cv %arg(%math(8 + "$opt_comp")
%qsar_retrieve(R_SQUARED))
setvar sep %arg(%math(8 + 2 * "$opt_comp")
%qsar_retrieve(STANDARD_ERRORS))
echo c = $opt_comp q2cv = $q2cv
echo

#
# #####
# PLS with no validation
# #####
#
# PLS tailor parameters for NV
#
TAILOR SET PLS \
CENTERING $center \
COMPONENTS $opt_comp \
CROSSVALIDATION 0 ||
#
# Execute NV analysis
#
setvar modelname %cat(qyr_ $RUN_NAME yr
$run_num $str_y0 nv)
echo $modelname

QSAR ANALYSIS DO INTERACTIVE $row_sel $col_sel
\
PLS yr | $modelname \
> /dev/null
#
# Get results
#
setvar R2 %arg(7 %qsar_retrieve(R_SQUARED))
setvar see %arg(7
%qsar_retrieve(STANDARD_ERRORS))
setvar y0 %arg(7 %qsar_retrieve(OFFSET))
# F value
# echo nl= $opt_comp n2= %math($nb_mol -
$opt_comp - 1)
setvar d %math(%math($nb_mol - $opt_comp - 1)
/ $opt_comp)
setvar f %math($R2 / %math(1 - $R2))
setvar FR2 %math($d * $f)
echo r2 = $R2 F = $FR2

#
# #####
# RECORD RESULTS SUMMARY
# #####
#
setvar tmp_str %cat(run $run_num " "\
$opt_comp " " $q2cv " " $sep " "\
$R2 " " $see " " $y0 " " $FR2 )
%array_push($a_logmodel "$tmp_str")

# FIN BOUCLE r
endfor
#
# Dump results in logfiles
#
for i in %range(1 $NB_TESTS)
  echo %array_get($a_logmodel $i) >> $LOGFILE
endfor
#
# Free memory and tmp column
#

```

```

IF %streql($KEEP_ANA NO)
echo deleting ...
wait 2
echo
for i in %range(1 $NB_TESTS)
  if %lt($i 10)
    setvar run_num %cat(0 $i)
  else
    setvar run_num $i
  endif
  setvar modele %cat(qyr_ $RUN_NAME yr $run_num
$str_y0 loo)
  qsar analysis delete pls $modele
  setvar modele %cat(qyr_ $RUN_NAME yr $run_num
$str_y0 nv)
  qsar analysis delete pls $modele

                                endfor
ENDIF

%array_free($a_logmodel)
%array_free($a_yr)
TABLE MODIFY COLUMN DELETE yr

#
# Close logfile
#
echo >> $LOGFILE
echo %time() >> $LOGFILE
echo QsaryRandomisation done >> $LOGFILE
echo
echo done !

```

# CONCLUSION

L'analyse et l'optimisation des molécules bioactives par modélisation moléculaire est une discipline à l'interface des nombreuses autres. Elle permet, grâce à l'informatique, de rassembler des données d'analyses physicochimiques, des données d'analyses biologiques et des structures chimiques dans une unité de lieu et de temps. Les modèles moléculaires qui en découlent font la synthèse des informations de ces différentes sources et participent ainsi à la compréhension des mécanismes d'interaction entre les petites molécules bioactives et leur cible biologique. La modélisation moléculaire fournit également des outils efficaces d'aide à la conception et à la sélection de nouvelles structures par la simulation de leurs interactions et par des prévisions quantitatives d'activité.

Ce travail a permis d'introduire dans nos laboratoires des techniques variées de modélisation moléculaire, adaptées à la conception et l'optimisation de ligands dans les deux principaux cas de figure rencontrés : celui dans lequel la structure tridimensionnelle de la cible est connue et celui dans lequel elle ne l'est pas. Dans les deux cas, les techniques employées ont produit des modèles fiables et riches en informations de qualité. Les résultats de la première étude ont fait l'objet d'un article dans une revue scientifique et un brevet va être déposé sur certains des composés issus de la seconde étude.

La qualité de tout modèle repose essentiellement sur la nature, la quantité et la qualité des données analysées. Cependant, l'amélioration des performances et des méthodes d'exploitation des outils conduiront à des modèles plus efficaces. Dans le domaine des techniques d'amarrage, ces améliorations passent par la prise en compte de la flexibilité du récepteur, tant au niveau des chaînes latérales des résidus que des chaînes principales des boucles. La modélisation explicite des molécules d'eau doit également se généraliser. Dans le domaine des techniques de QSAR-3D, les progrès passent par l'amélioration des champs d'interaction moléculaire et surtout par le développement de techniques d'alignement plus robustes et plus fiables. La difficulté que pose cette étape est telle que certains chercheurs essaient de s'en affranchir, par la mise au point de descripteurs 3D indépendants de l'orientation et de la conformation des structures.

La modélisation moléculaire est une étape clé de rationalisation qui s'insère logiquement dans le cycle « synthèse organique – test biologique ». La maîtrise et les progrès de cette discipline contribueront à la généraliser.

# **Techniques de Modélisation Moléculaire appliquées à l'Etude et à l'Optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance.**

**Auteur :** Antoine FORTUNÉ – Thèse soutenue le 21 décembre 2006.

## **Résumé :**

L'objet de ce travail est de présenter de façon détaillée des méthodes de modélisation appliquées à l'analyse des mécanismes de reconnaissance moléculaire et à la conception de nouveaux composés bioactifs selon deux approches : la conception basée sur la structure des récepteurs et la conception basée sur la structure des ligands.

Dans le cadre du premier axe, la méthode de construction de protéines par homologie de Blundell, implémentée dans le module COMPOSER de SYBYL et la méthode d'amarrage de Morris, implémentée dans le logiciel AUTODOCK3, sont décrites et appliquées à la modélisation et à l'étude des mécanismes de reconnaissance moléculaire d'un antigène polysaccharidique de la bactérie *Shigella flexneri* 5a et de mimes peptidiques immunogènes par un anticorps humain protecteur : IgA I3.

Dans le cadre du second axe, l'analyse statistique de descripteurs de champs d'interaction moléculaire de type CoMSIA et les méthodes de validation des modèles qu'elle génère sont présentées et appliquées à l'étude des relations structure activité en trois dimensions d'une série de 27 analogues de flavonoïdes modulateurs du transporteur ABCG2 (BCRP), impliqué dans le mécanisme de résistance multiple aux anticancéreux que développent les cellules tumorales. La production de modèles statistiquement fiables et performants a permis de concevoir de nouveaux composés biologiquement actifs.

**Mots Clé :** modélisation moléculaire, amarrage, Autodock, anticorps, polysaccharides, mimes peptidiques, QSAR3D, CoMSIA, flavonoïdes, résistance multiple (MDR), BCRP – ABCG2.

## **Study and Optimisation of Immunogenic Compounds and Multidrug Resistance Modulators by Molecular Modeling Technics**

**Author :** Antoine FORTUNÉ – PhD defended december 21<sup>th</sup>, 2006.

## **Abstract :**

The aim of this work is to present molecular modeling methods applied to molecular recognition analysis and new drug design by two approaches: structure base design or ligand based design.

In the first part, knowledge-based homology modeling of protein structure by Blundell, implemented in Sybyl-Composer, and docking method by Morris, implemented in Autodock3, are described and applied to study molecular recognition mechanisms of a polysaccharidic antigen of *Shigella flexneri* 5a and peptides mimics by human protective antibody : IgA I3.

In the second part, statistical analysis of CoMSIA molecular interaction field descriptors and validation methods of resulting models are described and used to build three dimensional quantitative structure activity relationship models of 27 flavonoid compounds modulators of ABCG2 transporter (BCRP), a protein involved in multidrug resistance capability developed by tumour cells. Designing new bioactive drugs was possible using the most statistically reliable and performing models built.

**Keywords :** molecular modelling, docking, Autodock, antibody, polysaccharides, peptide mimics, 3DQSAR, CoMSIA, flavonoïdes, multidrug resistance (MDR), BCRP, - ABCG2.