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

Discipline : Sciences Pharmaceutiques

Développement de nouvelles méthodes bioinformatiques pour l'étude des récepteurs couplés aux protéines G

Soutenue publiquement le 24 mai 2006
par Jean-Sébastien Surgand

Membres du jury

Directeur de thèse	Dr. Didier Rognan, Directeur de recherches, Strasbourg
Rapporteur interne	Pr. Mohamed Tajine, Professeur, Strasbourg
Rapporteur externe	Dr. Bernard Maigret, Directeur de recherches, Nancy
Rapporteur externe	Dr. Gilles Labesse, Chargé de recherches, Montpellier
Examinateur	Pr. Marcel Hibert, Professeur, Strasbourg

Remerciements

Je remercie tout d'abord mon directeur de thèse Didier Rognan de m'avoir accepté pour être son étudiant. Il m'a donné une excellente formation aux domaines de la bio- et chimioinformatique (que je ne connaissais pas) et m'a initié au monde de la recherche par sa compétence et sa patience. J'ai beaucoup appris grâce à lui et je lui dois beaucoup.

Je remercie également Marcel Hibert de m'avoir accueilli dans son laboratoire, ainsi que tous les membres de l'UMR 7175 pour avoir supporté mes exposés un peu inhabituels. Merci en particulier à Françoise Herth pour son aide administrative, à Bruno Didier et Nelly Maechler pour leur aide informatique.

Je remercie les rapporteurs pour avoir accepté de juger ce travail, et en particulier Mohamed Tajine qui m'a dirigé vers ce domaine et qui, dans ses cours, montre les mathématiques sous leur plus bel aspect.

Merci aux membres du groupe de Bioinformatique dans lequel j'ai évolué pendant 4 ans, et qui m'ont beaucoup aidé : Esther Kellenberger, Pascal Muller, Claire Schalon, Gilles Marcou, Nicolas Foata, Nathanaël Weill, Caterina Barillari, Bernard Coupez, Chris de Graaf. Et aussi les précédents membres que je n'oublie pas : Antoine Logean, Jordi Rodrigo de Losada, Guillaume Bret, Eui-Ki Kim, Patchreenard Saparpakorn, Sandrine Longuet, Michèle Mathis. Un grand merci en particulier pour Nicodème Paul, à qui j'ai

trop souvent essayé d'imposer mes points de vue, je lui dois beaucoup ; et aussi à Mireille Krier pour ses encouragements et sa bonne humeur.

Merci bien sûr à toute ma famille qui m'a toujours soutenu, durant ma thèse et avant.

Merci enfin à toutes les personnes que j'ai oubliées mais qui ont participé à ce travail.

Table des matières

Résumé

Les récepteurs couplés aux protéines G (RCPG) sont des protéines membranaires responsables de la transduction de signaux de l'extérieur vers l'intérieur de la cellule. Localisés partout dans l'organisme, ils sont impliqués dans de très nombreuses fonctions physiologiques comme la vision, l'olfaction, la croissance et l'adhésion cellulaire, la régulation hormonale, etc. Ils sont cibles d'une grande diversité de ligands : des photons, des ions, des amines biogènes, des hormones, des glycoprotéines, des molécules odorantes et gustatives, etc.

Un RCPG est formé de 7 hélices α transmembranaires reliées par des boucles intra- et extra-cellulaires. Ces 7 hélices α délimitent une cavité. Les ligands se fixent soit dans cette cavité soit au niveau des boucles extra-cellulaires, activant le récepteur qui va se lier à une protéine G à l'intérieur de la cellule, initiant une cascade de réactions chimiques. Un seul RCPG a été cristallisé jusqu'à présent, la rhodopsine bovine, apportant une information structurale précieuse.

Les RCPG présentent un grand intérêt pharmacologique. Leur diversité et les nombreuses fonctions qu'ils contrôlent les font intervenir dans de nombreuses pathologies. Plus de 30% des nouveaux médicaments mis sur le marché ciblent des RCPG. De plus, beaucoup de RCPG sont encore orphelins, c'est-à-dire sans ligand connu, et possèdent

donc un fort potentiel pharmacologique.

Les RCPG forment une super-famille de plus d'un millier de membres, dont la grande majorité sont responsables de la perception des molécules olfactives.

Le site de liaison de la plupart des RCPG est situé dans la cavité transmembranaire ; mais pour certains, le site est externe à la membrane et le ligand se fixe aux boucles extracellulaires. Mais même dans ces cas, les récepteurs possèdent une cavité transmembranaire et c'est sur elle que nous focalisons notre travail.

Plusieurs classifications des RCPG ont déjà été proposées : phylogénétiques, basées sur des automates statistiques, sur des empreintes physico-chimiques ou sur la composition en acides aminés. Mais aucune ne prend en compte précisément le point de vue pharmacologique, axé sur le ligand (le médicament).

C'est pourquoi nous proposons une nouvelle classification des RCPG orientée pharmacologie. Elle est basée sur l'étude de résidus supposés critiques de la cavité transmembranaire, qu'on pense interagir avec des ligands. Nous partons d'un jeu de données de 369 séquences de RCPG humains non-olfactifs, aussi « propre » que possible. Puis nous alignons les parties transmembranaires de façon automatique mais avec une étape de vérification et de raffinement manuels. Ensuite nous extrayons 30 résidus critiques en étudiant la cavité de la rhodopsine bovine, dont les parties transmembranaires ont une très forte identité avec celles de la rhodopsine humaine (94%). Nous supposons que ces 30 résidus sont critiques pour tous les RCPG, c'est-à-dire que la forme des cavités de tous les récepteurs est globalement conservée. Cette supposition est étayée par diverses publications. Enfin nous classifions ces séquences discontinues par une méthode de clus-

tering hiérarchique agglomératif (la méthode UPGMA). Les distances entre séquences sont simplement les scores d'identité entre elles. Une procédure de bootstrapping vient apporter un poids statistique à la classification qui aboutit à 22 clusters bien distincts.

Notre classification est en accord avec la classification GRAFS publiée récemment (analyse phylogénétique exhaustive de 342 RCPG humains non-olfactifs), c'est-à-dire que les clusters obtenus correspondent aux familles et sous-familles déjà identifiées, à quelques différences près.

On peut appliquer cette classification à la recherche de nouvelles cibles pour des ligands partageant des sous-structures communes (on parle de structures privilégiées). Partant de ligands dont on connait des récepteurs, on recherche parmi les 30 résidus critiques de ces récepteurs ceux qui interviennent dans la liaison. Ensuite on recherche d'autres récepteurs qui partagent les mêmes résidus. On peut proposer ces nouveaux récepteurs comme cibles potentielles pour les ligands de départ.

Une deuxième application immédiate de notre classification est la désorphanisation de récepteurs, c'est-à-dire la découverte d'un premier ligand pour un récepteur orphelin. Nous proposons comme point de départ pour un récepteur orphelin les ligands connus des récepteurs de la même classe.

La pertinence d'une classification basée sur les cavités transmembranaires pour des récepteurs dont le site de liaison n'est pas la cavité est discutable. Pourtant nous constatons que les familles pour lesquelles le ligand se fixe à l'extérieur de la membrane (famille de classe secrétine et glutamate) sont très bien identifiées et séparées les unes des autres.

Dans un deuxième temps, nous avons construit une autre classification des RCPG

en utilisant, non plus les séquences, mais des modèles 3D des récepteurs, construits par homologie. Pour schématiser, nous recherchons le point de vue qu'aurait un ligand placé dans la cavité. Nous plaçons donc une sphère conceptuelle, qui représente le ligand, au centre de gravité de la cavité. Cette sphère est découpée en 80 éléments triangulaires, de même taille et disposés uniformément sur la sphère. Ensuite nous projetons, à partir du carbone β des résidus, des informations physico-chimiques et géométriques de la cavité dans les triangles qui intersectent la demi-droite partant du centre de la sphère et allant vers le carbone β du résidu considéré. Enfin nous comparons les sphères deux à deux, chaque comparaison donnant un score utilisé pour générer une matrice de distances et finalement une classification par la même méthode que précédemment. La méthode est assez floue (discrétisation peu poussée, projection à partir des carbones β et non à partir de tous les atomes) pour masquer les erreurs dues à la modélisation.

Les modèles utilisés sont préalignés sur la structure qui a servi de point de départ, la structure cristallographique de la rhodopsine bovine. Malheureusement ces alignements ne sont pas suffisamment précis pour la génération d'une matrice de distances. Nous avons donc construit un outil d'alignement structural pour les raffiner. Cet outil est guidé par le score décrit ci-dessus pour trouver l'alignement entre deux cavités. Le meilleur score donnera le résultat de l'alignement.

La classification est homogène avec la précédente, mais le nombre de récepteurs non classés (singletons) est plus grand. Intéressant est le fait que le récepteur DUFFY, non classé par notre précédente classification, est classé ici avec les récepteurs des chimiokines, en accord avec la classification de la base Swiss-Prot.

Notre classification a en outre donné naissance à un outil d'alignement structural

de cavités adapté au travail sur des modèles, mais qui peut aussi aligner des structures cristallographiques.

Summary

G-protein-coupled receptors (GPCRs) are membrane proteins responsible for the transduction of signals from outside into the cell. Distributed in the whole body, they are implied in various physiological functions, like vision, olfaction, cellular growth and adhesion, etc. They are targeted by a tremendous diversity of possible ligands: photons, ions, biogenic amines, hormones, glycoproteins, olfactive and gustative molecules, etc.

A GPCR is composed of 7 transmembrane α-helices linked together by intra- and extracellular loops. These 7 α-helices delineate a cavity. Ligands bind either in this cavity, or on the extracellular loops, activating the receptor that will link to a G-protein inside the cell, initiating a cascade of secondary messengers. Up to now, the bovine rhodopsin is the only known crystallographic structure, bringing a valuable structural information.

GPCRs present a big pharmacological interest. Their diversity and the numerous functions they control get them involved in numerous pathologies. More than 30% of new commercialized drugs target GPCRs. Furthermore, many GPCRs are still orphan, without known ligand, and hence constitute potential pharmacological targets.

GPCRs form a superfamily of more than one thousand members. Three out of four are involved in the perception of olfactive molecules.

The binding site of most GPCRs is located in the transmembrane cavity; but for some

of them, it is located outside the membrane and the ligand binds an extracellular loop. But even for these cases, all receptors show a transmembrane cavity on which we will concentrate during this work.

Several classifications of GPCRs have been proposed: phylogenetic classifications, or based on statistical automata, or on physicochemical fingerprints, or based on their amino-acid composition. But none of them takes into account precisely the pharmacological point of view of the ligand (the drug).

That's why we propose a new classification of GPCRs which is pharmacologically-oriented. It is based on the study of some residues of the transmembrane cavity supposed to be critical and to interact with the ligand. We start from a dataset of 369 sequences of human nonolfactory GPCRs, as "clean" as possible. Then we align automatically the transmembrane parts, but with a manual check following. Then we extract 30 critical residues by studying the cavity of bovine rhodopsin, which transmembrane parts share a high identity score with the human rhodopsin (94%). We suppose that these 30 residues are critical for all GPCRs, i.e. the fold of all GPCRs is overall conserved. This hypothesis is supported by several publications. Eventually, we classify these sequences by an agglomerative hierarchical clustering algorithm (UPGMA). The distances between sequences are simply the identity scores between them. A bootstrap procedure brings a statistical support to the classification, that leads to 22 well-defined clusters.

Our classification agrees the recently published GRAFS classification (a phylogenetic analysis of 342 human non-olfactory GPCRs): resulting clusters correspond to already identified families and subfamilies, with slight differences.

This classification can be applied to find new targets to ligands that share some common substructures (called priviledged structures). We start from ligands with known receptors, we seek among the 30 critical residues of these receptors those that participate to the binding. Then we seek for other receptors that share the same residues. We can eventually propose these new receptors as putative targets for the ligands we started with.

A second straightforward application of the classification is the deorphanization of receptors, i.e. the discovery of a first ligand for an orphan receptor. We propose, as a starting point for an orphan receptor, the known ligands of the receptors of the same cluster.

The relevence of a classification based on the transmembrane cavities for receptors which binding site is not the this cavity is questionable. However we find that even the families for which the ligand bind outside the membrane (secretin and glutamate families) are well identified and well separated.

In a second time, we built another classification of GPCRs, based on 3D homology models rather than on sequences. To simplify, we try to take the point of view of a ligand inside a cavity. We put a conceptual sphere, that stands for the ligand, at the center of gravity of the cavity. This sphere is tessellated into 80 triangles, with same size, homogenously dispatched. Then we project from the β carbon of the residues some physicochemical and geometrical information into the triangles. Eventually we compare the spheres, each comparison gives a score used to build a distance matrix and a classification with the same method as for the previous one. This method is quite fuzzy (low resolution

discretization, projection from the β carbons and not from all atoms) to hide the errors due to modelling.

The models are prealigned on the cristal structure of bovine rhodopsin. But unfortunately these alignments were not precise enough to build a distance matrix. So we coded a structural alignment tool to refine the alignments. This tool is guided by the previsouly-described score to find an alignment between two cavities. The best score gives the output alignment.

The resulting clustering is homogenous with the previous one, but the number of non-classified receptors (singletons) is higher. Interestingly, the receptor DUFFY, not classified by our previous clustering, is classified here in the Chemokine cluster, in agreement with the Swiss-Prot database classification.

Our classification leads to a structural alignment tool adapted to work on models, but that can also work on crystal structures.

Introduction

L'ère post-génomique a produit une masse considérable de données lors de la transcription de nombreux génomes y compris du génome humain [1]. De plus, le nombre de protéines cristallisées ne cesse d'augmenter [2]. Toutes ces informations de séquences et de structures de protéines sont cruciales dans la compréhension des mécanismes moléculaires de la vie.

Parmi les protéines humaines, nous allons nous intéresser dans ce mémoire à une super-famille particulière : les récepteurs couplés aux protéines G (RCPG). Environ un millier, ces récepteurs interviennent dans de nombreuses fonctions physiologiques et donc dans d'aussi nombreuses pathologies.

Leur intérêt pharmaceutique est démontré par quelques chiffres : 30% des médicaments les plus vendus sur le marché ciblent les RCPG [3], ce nombre passe à 60% dans le cas des nouveaux médicaments mis sur le marché [4].

Pour créer un médicament ciblant une protéine, on doit étudier l'interaction du médicament avec la protéine. Pour cela il faut connaître la structure de la cible. Malheureusement dans le cas des RCPG, une seule structure cristallographique est résolue, celle de la Rhodopsine bovine [5], très proche de la Rhodopsine humaine.

Les RCPG formant une super-famille, certains récepteurs sont proches les uns des

autres, dans le sens où ils peuvent reconnaître des mêmes ligands. Or un médicament doit être sélectif d'une protéine (si on n'en vise qu'une). L'intérêt d'une classification des RCPG est donc d'intégrer cette recherche de sélectivité au début de la chaîne de création d'un médicament, évitant ainsi une perte de temps et d'argent dans le cas où il provoquerait des effets secondaires indésirables par manque de sélectivité.

Plusieurs classifications des RCPG existent déjà, mais leur principal inconvénient est de ne pas intégrer le point de vue du ligand (le médicament) : les classifications phylogénétiques considèrent les récepteurs dans leur ensemble, et non pas seulement le site actif où se fixera le ligand ; pour les autres, elles ne s'intéressent pas au classement des protéines humaines, mais plutôt au classement inter-espèces de tous les RCPG.

Nous nous proposons donc, dans ce travail de thèse, de construire une classification des RCPG intégrant le point de vue pharmacologique du ligand, et de ne travailler que sur des protéines humaines.

Bibliographie

[1] Venter J.C. (2001) The Sequence of the Human Genome. *Science* **291**:1304–1351.

[2] Berman H.M., Westbrook J., Feng Z., Gilliand G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**:235–242.

[3] Flower D.R. (1999) Modelling G-protein-coupled receptors for drug design. *Biochim. Biophys. Acta* **1422(3)**:207–234.

[4] Seifert R., Wieland T. (eds) (2005) G Protein-Coupled Receptors as Drug Targets: Analysis of Activation and Constitutive Activity. *John Wiley and Sons, Inc.*

[5] Palczewski K., Kumasaka T., Hori T., Behnke C.A., Motoshima H., Fox B.A., Le Trong I., Teller D.C., Okada T., Stenkamp R.E., Yamamoto M., Miyano M. (2000) Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **289**:739–745.

Chapitre 1

Classifications établies des récepteurs couplés aux protéines G

1.1 Introduction

Les récepteurs couplés aux protéines G (ou RCPG [1]) sont des protéines membranaires responsables de la transduction des signaux de l'extérieur vers l'intérieur des cellules. Ils ont une structure commune en 7 hélices transmembranaires reliées entre elles par des boucles intra- et extra-cellulaires. Ils sont activés par une très grande variété de ligands, comme la lumière [2], des molécules olfactives et gustatives, des ions, des peptides, des hormones, des amines biogènes, etc. De par cette diversité de ligands, et parce qu'ils sont présents à la surface de très nombreuses cellules, les RCPG interviennent dans un

1. Liste des abbréviations utilisées dans ce mémoire : RCPG (récepteur(s) couplé(s) aux protéines G), TM (transmembranaire ou partie(s) transmembranaire(s)).

2. Le ligand dans ce cas particulier est le rétinal, c'est lui qui, activé par la lumière, change de conformation et active à son tour le récepteur.

grand nombre de processus physiologiques, et donc de pathologies. Ils jouent ainsi un rôle crucial en pharmacologie.

Nous présentons dans ce chapitre les classifications déjà établies des RCPG. Elles sont assez nombreuses à partir des années 2000, tandis que de 1985 (date des premières bases de données de séquences de protéines) à 2000, seule la classification par ligands connus existait.

1.2 Les récepteurs couplés aux protéines G (RCPG)

Cette section présente les récepteurs couplés aux protéines G (RCPG) du point de vue biologique, chimique et pharmacologique.

1.2.1 Aspects biologiques

Les RCPG forment une super-famille de récepteurs présents sur les membranes cellulaires. Leur fonction générique est la transduction de signaux de l'extérieur vers l'intérieur de la cellule [1]. Ils sont présents chez toutes les espèces animales ainsi que chez les bactéries et les champignons [2], mais dans ce mémoire nous ne nous intéressons qu'aux récepteurs humains. Les RCPG interviennent dans de nombreux processus physiologiques, comme la vision, l'olfaction, la régulation hormonale, la croissance cellulaire [1]. L'analyse des résultats du séquençage du génôme humain prédit l'existence de quelques 1000 RCPG [3], dont 400 sont des récepteurs non-olfactifs [4] — qui ne sont pas liés au mécanisme de l'olfaction. Sur 30 000 protéines prédites, les RCPG représentent 3%, pour cette seule super-famille.

La structure des RCPG est caractérisée par un domaine transmembranaire formé de 7 hélices α hydrophobes reliées entre elles par des boucles intra- et extra-cellulaires (figure 1-1). Les longueurs de ces boucles sont très variables en fonction des récepteurs, allant de quelques résidus à plusieurs centaines. On abrège souvent le nom de ce domaine par 7TM, et on utilise TM pour parler d'une des hélices α. Le domaine transmembranaire forme une cavité qui est le site actif d'une grande partie des RCPG, notamment ceux dont les boucles extra-cellulaires sont relativement courtes. Pour ceux chez qui ces boucles sont longues, ces dernières peuvent servir de site actif.

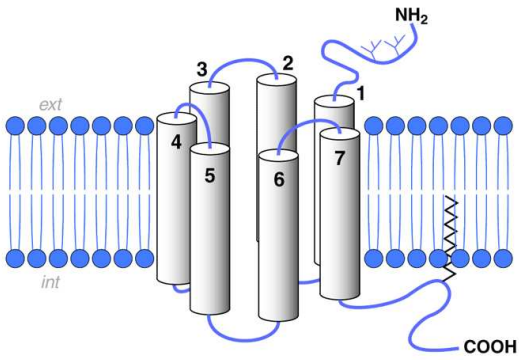


Figure 1-1. Vue schématique d'un RCPG.

Seule la rhodopsine bovine a été résolue par cristallographie à l'heure actuelle, par l'équipe de Palczewski en 2000 [5a/b], grâce à la diffraction par rayons X (figure 1-2). Le récepteur est complexé avec son ligand (le rétinal) par liaison covalente. Cette rare information structurale en est d'autant plus précieuse et sert de point de départ à plusieurs

modélisation par homologie des RCPG [2,6-8].

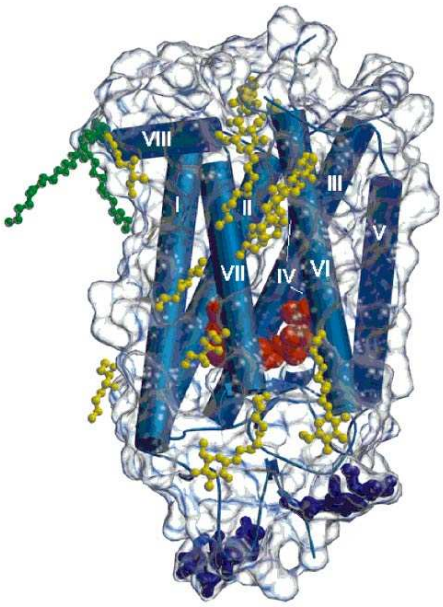


Figure 1-2. Structure cristallographique de la rhodopsine bovine (d'après [5]).

Les RCPG sont aussi nombreux que diversifiés. Cette diversité apparaît dans les processus physiologiques qu'ils régulent, mais aussi dans les ligands qu'ils reconnaissent. Ainsi, leurs classifications reflètent cette diversité. Nous détaillerons ces classifications dans la section suivante, mais globalement, la super-famille des RCPG est divisée hiérarchiquement en familles et sous-familles d'après des homologies de leurs séquences primaires [9-16].

1.2.2 Aspects chimiques

À partir d'une même structure de base, les RCPG peuvent être activés par une grande diversité de ligands : ions de faible masse moléculaire (Ca^{2+}), amines biogènes (dopamine, sérotonine), nucléoside et nucléotides (adénosine, ATP), peptides et hormones peptidiques (chimiokines, glucagon), lipides (sphingolipides, prostaglandines), mais aussi molécules

olfactives et gustatives exogènes, et enfin le rétinal, cas particulier car lié de façon covalente à son récepteur, et activé par la lumière [1,17]. Le schéma de la figure 1-3 présente quelques sites de fixation pour ces ligands, souvent dans la cavité transmembranaire, mais aussi au niveau des boucles extracellulaires.

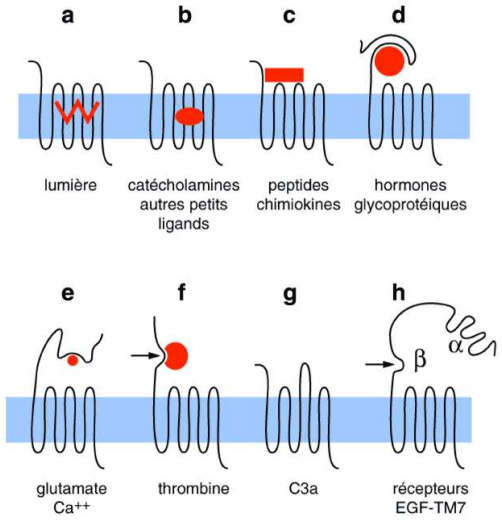


Figure 1-3. Schéma de quelques sites de fixation de ligands (d'après [17]).

Lors d'une liaison entre un ligand agoniste et un récepteur, ce dernier est activé par la liaison, c'est-à-dire qu'il va changer de conformation [16] par rapport à sa forme non liée au ligand. Ce changement conformationnel va impliquer une interaction avec une protéine G, liée au récepteur au niveau de son interface intracellulaire. Plus précisément, la protéine G, un hétérotrimère, va être scindée en deux parties et va en libérer une dans la cellule, débutant une cascade complexe d'événements produisant de nombreux messagers secondaires [1]. La rhodopsine est un récepteur un peu particulier, car son ligand, le rétinal, lui est lié de façon covalente. L'activation de la rhodopsine provient d'un changement conformationnel du rétinal qui, lorsqu'il est frappé par un photon, passe d'une forme *cis* à une forme *trans* [2]. Le changement de conformation des RCPG

de type rhodopsine (figure 1-4) fait intervenir un déplacement des hélices 3, 5, 6 et 7. [18]

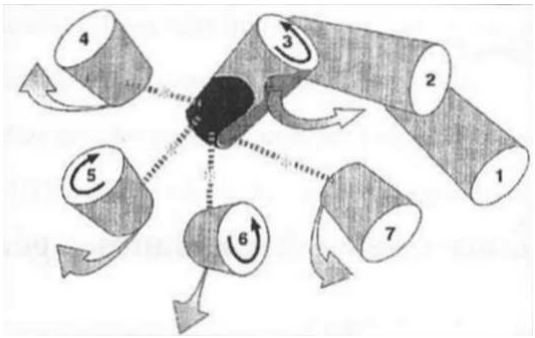


Figure 1-4. Changement conformationnel des RCPG de type rhodopsine lors de l'activation du rétinal (d'après [18]).

Les RCPG peuvent être présents sous forme non activée, sous forme activée avec un ligand, ou encore pour certains d'entre eux sous forme activée sans ligand (on parle alors d'activité intrinsèque ou constitutive). Toutes ces formes sont en équilibre et l'activité globale dépend de la proportion de ces formes en plus de l'activité des récepteurs, qui peut encore être modifiée par des effecteurs allostériques ou par multimérisation des récepteurs [19].

1.2.3 Aspects pharmacologiques

Les RCPG sont distribués sur la surface de nombreuses cellules et régulent un grand nombre de processus physiologiques, comme la vision (sous-famille des opsines), l'olfaction (sous-famille des olfactifs), la croissance et la prolifération cellulaire (famille frizzled), l'adhésion cellulaire (famille adhésion), la relaxation et la contraction des muscles lisses (sous-famille des prostanoïdes), la vasoconstriction (sous-famille des vasopeptides), et aussi la neurotransmission, les mécanismes hormonaux, des mécanismes inflammatoires, etc.

Chacun de ces processus physiologiques peut fonctionner de façon anormale. Les RCPG qui régulent ces processus interviennent donc dans la pathologie, soit parce qu'ils sont sur-exprimés ou sous-exprimés, soit qu'ils sont sur-activés ou bloqués. Dans le cas d'une sur-expression ou d'une sur-activité, il faut essayer de bloquer l'activité des récepteurs grâce à des ligands antagonistes ou agonistes inverses. Les ligands antagonistes se logent dans le site actif sans provoquer l'activation du récepteur, mais en prenant la place des ligands naturels ainsi empêchés ; l'activité globale est diminuée. Les ligands agonistes inverses sont utiles dans le cas où le récepteur possède une activité constitutive sans être complexé avec un ligand. En se liant avec le récepteur (comme un ligand agoniste), ils vont bloquer son activité, d'où le qualificatif d'inverse. L'activité globale est encore une fois diminuée. Dans le cas d'une sous-expression ou d'une sous-activité, il faut stimuler l'activité par des ligands agonistes, qui vont provoquer l'activation des récepteurs, ou supprimer une molécule inhibitrice.

Les RCPG forment donc une classe de cibles très attractives pour les interventions thérapeutiques. Voici quelques chiffres pour donner un ordre d'idée de leur succès auprès des industries pharmaceutiques : 30% des médicaments les plus vendus modulent l'actvité des RCPG [20] ; d'autres sources vont jusqu'à 60% des nouveaux médicaments mis sur le marché [21]. Leur intérêt pharmacologique est triple : tout d'abord il y a les récepteurs déjà ciblés mais pour lesquels on peut améliorer la sélectivité et l'activité des médicaments ; ensuite il y a les récepteurs dont on connaît des ligands endogènes, mais qui ne sont pas encore pris pour cible par des médicaments, et ils représentent la grande majorité (à l'heure actuelle, seuls 40 RCPG sont ciblés par des médicaments, sur les 1000 prévus par l'analyse du génôme humain) ; enfin il existe environ 100 RCPG orphelins, dont on ne

connait aucun ligand, et qui présentent un intérêt potentiel à plus long terme [22].

1.3 Classifications des RCPG

Les RCPG comprennent un millier de membres. Il est donc nécessaire et naturel des les classifier. Nécessaire, pour avoir une vue plus synthétique de la super-famille, trouver les points communs, les différences. Naturel, puisqu'ils se différencient suffisamment les uns des autres, par exemple par les ligands qu'ils reconnaissent. Une telle classification est importante en pharmacologie pour deux raisons. La première fait intervenir la sélectivité d'un ligand par rapport à un récepteur particulier : si on connait une classification, on peut prendre en compte cette sélectivité relativement tôt dans le processus de création d'un nouveau médicament, en vérifiant que le candidat médicament ne cible pas également les récepteurs voisins du récepteur visé. Il est naturel de penser qu'un ligand pour un récepteur donné puisse cibler des récepteurs voisins structurellement. La seconde raison est qu'il y a parmi les RCPG encore un centaine de récepteurs orphelins ; d'après le même raisonnement, si on parvient à trouver des récepteurs dont on connait des ligands voisins de ces récepteurs orphelins, on peut penser que ces ligands seront reconnus par les récepteurs orphelins. La classification constitue donc un moyen simple de trouver des ligands pour des récepteurs orphelins, cibles thérapeutiques potentielles.

Nous allons présenter dans le reste de ce chapitre certaines des classifications de RCPG existantes. Nous essaierons d'être critiques et allons noter les points forts et les lacunes de chacune, pour arriver à argumenter de l'intérêt des classifications que nous avons construites dans le cadre de cette thèse.

1.3.1 Classification historique par ligands

Les RCPG ont commencé à être regroupés systématiquement dans des bases de données dans la fin des années 1980 et le début des années 1990. Les deux principales bases étaient et sont toujours la Swiss-Prot et la GPCRDb. La base Swiss-Prot [23] (maintenant UniProt, voir chapitre suivant) regroupe des séquences de protéines de tous organismes ; elle a été mise en place en 1988 et sa vocation est de fournir une information curée des séquences de protéines avec, pour chaque entrée, de nombreux liens bibliographiques et des liens hyper-textes vers d'autres bases de données. La GPCRDb (*G-Protein-Coupled Receptors Database*) [2], mise en place en 1994, vise les récepteurs à 7 parties transmembranaires, et en particulier les RCPG. Elle regroupe de nombreuses informations, comme bien sûr les séquences, mais aussi des modèles 3D, une classification basée sur des similarités de séquences, et aussi des liens vers d'autres bases de données.

Les RCPG ont tout d'abord été classés, au début des années 1990, en trois classes ou familles, suivant la longueur des parties N-terminales. La nomenclature utilisée par la GPCRDb utilise le mot de *classe* et numérote ces classes de A à F. Au début, seules les trois premières classes (A-C) étaient présentes, mais actuellement (2006) il existe six classes, et d'autres sont proposées. Ces classes sont : (A) les récepteurs ressemblant à la Rhodopsine, qui ont un petit domaine N-terminal, (B) les récepteurs ressemblant à celui de la Secrétine, qui ont un grand domaine N-terminal structuré de façon complexe, (C) les récepteurs ressemblant à celui du Glutamate, qui ont un grand domaine N-terminal sous forme de dimère, (D) des récepteurs de phéromones fongiques (non humains), (E) des récepteurs de cAMP (non humains), et enfin (non numéroté) des récepteurs dénommés en anglais Frizzled et Smoothened, qui possèdent de très longs domaines N-terminaux.

Les noms des classes, en majuscule, provient soit du nom du récepteur, comme dans le cas de la Rhodopsine et Frizzled, soit du nom de ligands, comme dans le cas des classes Secrétine, Glutamate. D'un autre côté, la base de données Swiss-Prot utilise le mot de *famille* pour classer les récepteurs, et numérote les familles de 1 à 5 qui correspondent aux classes A à E, ainsi que la famille Frizzled/Smoothened.

Puis les deux classifications partagent la famille de la Rhodopsine en sous-familles ou sous-classes, d'après la nature chimique des ligands, des spécificités pour les agonistes et antagonistes connus et des associations pharmacologiques avec activation ou inhibition. Le tableau de la figure 1-5 donne la classification de la Swiss-Prot en familles et sous-familles des RCPG, basée sur les ligands connus. La classification actuelle de la GPCRDb est basée sur des similarités de séquences, non plus sur des ligands connus. Nous donnons pour chaque récepteur son ligand (ou classe de ligands). Ce tableau est en langue anglaise, car il est difficile de trouver une traduction pour chacun des récepteurs. Dans la rare littérature en langue française que nous avons consulté (par exemple [24]), la plupart des récepteurs ne trouvent pas de traduction... Nous avons donc préféré donner toute la classification dans une seule langue.

```
RHODOPSIN-like receptors
  Acetylcholine (muscarinic) receptors
    ACM[1-5]           muscarinic acetylcholine
  Adenosine and adenine nucletide receptors
    AA(1,2[AB],3)R     adenosine
    P2RY[1-689],
    P2Y1[0-4]          P2Y purine
    LT4R[12]           leukotriene
  Adrenergic receptors
    ADA(1[ABC],2[ABC]) alpha adrenergic
    ADRB[1-3]          beta adrenergic
  Adrenomedullin receptor
    ADMR               adrenomedullin
  Angiotensin receptors
```

```
  AGTR[12]            angiotensin II
Apelin receptor
  APJ                 apelin
Bile acid receptor
  GPBAR               bile acid
Bombesin receptors
  NMBR                neuromedin-B
  GRPR                gastrin-releasing peptide
  BRS3                bombesin
Bradykinin receptors
  BKRB[12]            bradykinin
Cannabinoid receptors
  CNR[12]             cannabinoid
Chemokines and chemoattractants factors receptors
  C[35]AR,C5ARL       anaphylatoxin
  FPR1                fMet-Leu-Phe
  FPRL[12]            FMLP-related
  CXCR[12]            high affinity interleukin-8
  CXCR[3-6]           C-X-C chemokine
  CCR[1-9],CCR10,
  CCRL1               C-C chemokine
  CCBP2               chemokine-binding protein
  XCR1                chemokine XC
  CX3C1               CX3C chemokine
  DUFFY               Duffy antigen/chemokine
Cholecystokinin/gastrin receptors
  CCKAR               cholecystokinin type A
  GASR                gastrin/cholecystokinin type B
Cysteinyl leukotriene receptors
  CLTR[12]            cysteinyl leukotriene
Dopamine receptors
  DRD[1-5]            dopamine
Eicosanoid receptors
  OXER1               oxoeicosanoid
Endothelin receptors
  EDNR[AB]            endothelin
Free fatty acid receptors
  FFAR[1-3]           free fatty acid
Glycoprotein hormones receptors
  FSHR                follicle-stimulating hormone
  LSHR                lutropin-choriogonadotropic hormone
  TSHR                thyrotropin
Histamine receptors
  HRH[1-4]            histamine
Kisspeptins receptors
  KISSR               KiSS-1
Krebs cycle intermediates receptors
  OXGR1               2-oxoglutarate
  SUCR1               succinate
Lysolipids receptors
  EDG[1358],GPR6      sphingosine 1-phosphate (S1P)
  EDG[247]            lysophosphatidic acid (LPA)
  G2A                 lysophosphatidylcholine (LPC)
  SPR1                sphingosylphospohrylcholine (SPC)
  PSYR                psychosine
```

```
Melanin-concentrating hormone receptors
  MCHR[12]           melanin-concentrating hormone
Melanocortins receptors
  ACTHR              adrenocorticotropic hormone
  MSHR               melanocyte-stimulating hormone
  MC[3-5]R           melanocortin
Melatonin receptors
  MTRA[ABL]          melatonin
Neuromedin U receptors
  NMUR[12]           neuromedin U
Neuropeptides B/W receptors
  NPBW[12]           neuropeptides B/W
Neuropeptide FF receptors
  NPFF[12]           neuropeptide FF
Neuropeptide S receptor
  GP154              neuropeptide S
Neuropeptide Y receptors
  NPY[1245]R         neuropeptide Y
Neurotensin receptors
  NTR[12]            neurotensin
Nicotinic acid receptors
  G109[AB]           nicotinic acid
Odorant/olfactory and gustatory receptors
  OLxxx, OLFxx
  GUxxx
Opioid peptides receptors
  OPR[DKM]           delta/kappa/mu-type opioid
  OPRX               nociceptin
Opsins
  OPSD               rhodopsin
  OPS[BGR]           blue/green/red-sensitive opsin
  OPN3               encephalopsin
  OPN4               melanopsin
  OPN5               neuropsin
  OPSX               peropsin
  RGR                RPE-retinal
Orexins receptors
  OX[12]R            orexin
Pheromone receptors
  VN1R[1-5]          vomeronasal type-1
Platelet-activating factor receptors
  PTAFR              platelet-activating factor
Prokineticin receptors
  PKR[12]            prokineticin
Prostanoids receptors
  P[EF]2R,PE2R[1-4]  prostaglandins
  PI2R               prostacyclin
  TA2R               thromboxan
Proteinase-activated receptors
  PAR[1-5]           proteinase
Relaxin receptors
  LGR8,RL3[12],RXFP1 relaxins
Releasing hormones receptors
  GNRHR,GNRR2        gonadotropin-releasing hormone
  TRFR               thyrotropin-releasing hormone
```

```
    GHSR                growth hormone secretagogue
    MTLR                motilin
  Serotonin receptors
    5HT(1[ABD],2[ABC],
    [467]R,5[AB])       5-hydroxytryptamin
  Somatostatin and urotensin receptors
    SSR[1-5]            somatostatin
    UR2R                urotensin
  Tachykinin receptors
    NK1R                substance-P
    NK2R                substance-K
    NK3R                neuromedin K
  Trace amine receptors
    TAAR[1235689]       trace amine
  Vasopressin/oxytocin receptors
    V1[AB]R,V2R         vasopressin
    OXYR                oxytocin
SECRETIN-like receptors
    CALCR,CALRL         calcitonin
    CRFR[12]            corticotropin-releasing factor
    GIPR                gastric inhibitory peptide
    GLR,GLP[12]R        glucagon
    GHRHR               growth hormone-releasing hormone
    PTHR[12]            parathyroid hormone
    PACR                pituitary adenylate cyclase-activity polypeptide
    SCTR                secretin
    VIPR[12]            vasoactive intestinal polypeptide
    BAI[1-3]            brain-specific angiogenesis inhibitor
    CD97                CD97 antigen
    ELTD1,EMR[1-4]      EGF
    LPHN[1-3]           latrophilin
    CELR[1-3]           cadherin
GLUTAMATE-like receptors
    MGR[1-8]            metabotropic glutamate
    CASR                extracellular calcium
    GABR1               GABA
    RAI3                retinoic acid-induced protein 3
```

Figure 1-5. Classification des RCPG par ligands. Les ligands sont donnés systématiquement.

Le point fort de cette classification, c'est son aspect directement pharmacologique, puisque basée sur le ligand. Elle ne tient toutefois pas compte de la permissivité de certains récepteurs (ex : récepteurs adrénergiques) pour divers ligands endogènes (ex : épinéphrine, norépinéphrine, dopamine). Elle servira de point de comparaison avec les

classifications que nous allons construire dans les prochains chapitres, et notamment avec la dernière classification (chapitre 3).

À partir des années 2000, de nombreuses méthodes de classification hiérarchiques en familles et sous-familles ont vu le jour. Nous en présentons maintenant quelques-unes.

1.3.2 Classifications phylogénétiques

Deux travaux proposent des classifications phylogénétiques des RCPG. Le premier et antérieur est celui de Joost et Methner [9] (2002) qui présente une classification de 277 RCPG humains non-olfactifs. La méthode utilisée est une méthode de distance phylogénétique entre séquences baptisée *neighbor-joining* (joindre les voisins) [25]. Le jeu de données utilisé comportait 196 récepteurs non-orphelins, et 81 récepteurs orphelins. La seconde classification est établie par Fredriksson et collaborateurs [4] (2003). Plus complète dans le jeu de données car plus récente, de méthode semblable, elle présente une révision de la classification des RCPG répartissant 342 récepteurs humains non-olfactifs selon 5 familles. Elle est baptisée GRAFS, acronyme des noms des 5 familles (Glutamate, Rhodopsin, Adhesion, Frizzled/taste2, Secretin).

La méthode utilisée, neighbor-joining, consiste à minimiser les différences entre les longueurs des branches de l'arbre construit et les « vraies » distances. On commence par un arbre enraciné, avec toutes les séquences reliées à cette racine. Puis on cherche la combinaison de deux séquences qui réduit le plus la somme des longueurs du nouvel arbre, en testant toutes les combinaisons possibles. Et ainsi de suite jusqu'à ce que l'arbre soit binaire (au début, l'arbre est N-aire si N est le nombre de séquences). Une procédure de validation statistique, le *bootstrap*, est appliquée pour attester de la pertinence de la

classification construite. Ainsi, on construit 1000 arbres (dans les deux cas, ces paramètres sont les mêmes) pour lesquels il manque des séquences, puis on regroupe le tout dans un arbre consensus. Plus les *valeurs de bootstrap* sont élevées, plus la classification est pertinente et insensible à des perturbations du jeu de données. Nous expliquerons plus longuement cette procédure dans le chapitre 2, car nous l'avons également utilisée.

Le résultat de la classification de Joost et Methner comprend 19 sous-familles de la famille Rhodopsine, et les deux familles Secrétine et Glutamate. Les valeurs de bootstrap sont élevées, indiquant une bonne séparation des familles. Le résultat de la classification GRAFS comprend, comme caché dans son appellation, 5 familles et une quinzaine de sous-familles de la Rhodopsine.

Ces deux classifications sont très homogènes entre elles, et surtout cohérentes avec la classification par ligands que nous avons présentée ci-dessus.

Le problème de ces deux classifications est qu'elles ne prennent pas en compte l'aspect pharmacologique : le ligand ne se lie qu'à une petite partie du récepteur ; mais l'information utilisée est la séquence complète sans discrimination d'un quelconque site actif. Dans la classification de Joost et Methner, on ne trouve que peu de groupes, et quelques récepteurs sont mal classés (comme un récepteur de la mélatonine dans le groupe des récepteurs de peptides, un récepteur des mélanocortines classé avec les récepteurs de lipides). La classification GRAFS comporte elle aussi quelques problèmes, notamment une région mal définie qui comprend pêle-mêle une opsine (RGR), un récepteur purinergique (P2RY11), un récepteur d'hormones (TRHR).

Finalement, ces deux classifications phylogénétiques basées sur une méthode de distance entre séquences, sont homogènes, mais les problèmes de classements proviennent

du fait que ce n'est pas l'aspect pharmacologique qui était recherché.

1.3.3 Classifications par automates statistiques

Plusieurs travaux utilisent des automates statistiques pour classifier des protéines. C'est le cas notamment de Qian et collaborateurs [11] et de Lui et Califano [12], qui utilisent des modèles de Markov cachés [26] pour classifier les RCPG.

Les modèles de Markov cachés (en anglais *Hidden Markov Models*, abrégé en HMMs) sont des automates statistiques qui modélisent un processus de Markov, c'est-à-dire un processus basé sur une fonction aléatoire. Ils ont été introduits vers 1965, puis ont été utilisés vers 1975 pour la reconnaissance de la parole [26], et enfin, à partir des années 1985, ils ont été utilisés en bioinformatique pour l'analyse des séquences biologiques, comme c'est le cas ici. Techniquement, ce sont des automates à états finis associés à une distribution de probabilités. Les transitions entre états sont dirigés par cette distribution, et ces transitions peuvent donner lieu à des « observables ». Les états sont cachés, mais ce qu'on aperçoit du comportement de l'automate, ce sont ces observables. Ainsi, étant donnée un ensemble d'états et une distribution de probabilités, un modèle de Markov caché peut générer une suite d'observables. Les modèles de Markov cachés permettent de résoudre trois genres de problèmes : 1) le problème de l'évaluation : étant donné une distribution de probabilité et une suite d'observables, quelle est la probabilité pour cette suite soit générée par un modèle particulier ? 2) le problème du décodage : avec les mêmes données, quelle est la séquence de transitions qui a produit l'observable ? et enfin 3) le problème de l'apprentissage : avec les mêmes données, comment ajuster les paramètres du modèle de façon à maximiser la probabilité que ce modèle donne une suite particulière

d'observable ? L'utilisation typique en bioinformatique sera d'entraîner le modèle (c'est la résolution du problème de l'apprentissage) de façon à ce qu'il reconnaisse (c'est la résolution du problème de l'évaluation) une famille de séquences. On aura donc un modèle de Markov par famille et sous-familles, chacun reconnaissant les membres de sa (sous-) famille seulement.

Le jeu de données des deux classifications étudiées sont extraits de la GPCRDb, et comprennent des protéines de toutes les espèces présentes dans la base. Les modèles sont entraînés à l'aide de la moitié des données, puis évalués avec l'autre moitié. Le nombre de mauvaise attribution est faible, inférieur au pourcent. C'est donc une bonne méthode pour classifier les RCPG.

Le point fort de ces méthodes est qu'elles sont capables par apprentissage à partir de données initiales apprises de trouver des relations pour des objets où on n'a peu ou pas de données expérimentales. Dans notre cas, à partir de récepteurs dont les ligands sont connus, ces modèles sont capables d'assigner les récepteurs orphelins à des classes de façon pertinente.

On pourra regretter les jeux de données, toujours peu appliqués à la pharmacologie : toutes les espèces sont représentées, et on ne trouve donc relativement que peu de protéines humaines.

1.3.4 Classification par propriétés des acides aminés

Deux travaux basent la classification des RCPG sur une analyse des propriétés des acides aminés, il s'agit de Lapinsh et collaborateurs [15] et d'Otaki et collaborateurs [16]. La description de départ des acides aminés est la même, mais les deux méthodes

diffèrent ensuite dans le traitement de cette description : par une analyse statistique pour [15] ou par un réseau neuronal pour [16]. Ces classifications sont indépendantes de tout alignement préalable des séquences.

Chacun des récepteurs est décrit par un vecteur de 26 descripteurs physico-chimiques (comme le poids moléculaire, le volume de Van der Waals, la surface moléculaire totale, la surface moléculaire polaire et non polaire, le nombre de donneurs et d'accepteurs de liaisons hydrogène, etc.) [15] Ensuite on applique soit une méthode d'analyse qui extrait de l'information de la matrice de N récepteurs sur M variables (les 26 descripteurs), et qui trouve un modèle de dimension moindre qui approxime la structure de données en entrée [15], soit on applique la matrice NxM à l'entrée d'un réseau neuronal. Le seul apport du réseau neuronal est sa capacité de détection de relations non-linéaires.

Voici la définition du *réseau neuronal* donné dans le Journal Officiel (10 octobre 1998) : c'est un « ensemble de neurones artificiels inter-connectés permettant la résolution non-algorithmique[3] et presque certaine de problèmes complexes tels que la reconnaissance de formes ou le traitement du langage naturel grâce à l'ajustement de paramètres dans une phase d'apprentissage ». Nous avons choisi cette définition, bien qu'il existe des sources plus mathématiques, car nous la trouvons claire, concise et suffisante dans le cadre de ce mémoire.

À partir de jeux de données tirés de la GPCRDb, Lapinsh et collaborateurs classent 929 RCPG de la famille Rhodopsine en sous-familles clairement séparées. De plus, 90 récepteurs orphelins sur 165 que comprenait le jeu ont été associés à des sous-familles de façon non ambiguë. Pour Otaki et collaborateurs, après une phase d'apprentissage du

3. Penrose discute sur le terme de non-algorithmique [27].

réseau neuronal, 4116 RCPG de famille Rhodopsine ont été assignés à 15 sous-familles avec une précision de 97,8%.

L'avantage de ces deux méthode réside dans l'indépendance par rapport aux alignements. S'il est relativement naturel de vouloir déterminer un alignement des parties transmembranaires des RCPG, ce que nous allons faire dans le prochain chapitre, il est en revanche difficile de trouver des alignements des boucles intra- et extracellulaires, dont les tailles peuvent varier de quelques résidus à plusieurs milliers.

Le point faible, comme pour les autres méthodes décrites ci-dessus, est de n'avoir qu'un jeu de données comprenant relativement peu de protéines humaines, et un nombre faible et fixé de classes. Il est également difficile de relier similarité et structure du récepteur. Enfin, ces deux travaux ne s'intéressent qu'aux récepteurs de la famille Rhodopsine et n'essaient pas d'insérer dans le jeu de données des récepteurs des autres familles, ce qui aurait pu être intéressant.

1.3.5 Classification par empreintes de résidus

L'avant-dernière classification des RCPG que nous présentons est basée sur des motifs spécifiques utilisés pour différentier des sous-familles [13]. Ces motifs, une fois générés en étudiant une classification existante, permettent d'assigner des récepteurs orphelins à des sous-familles. Nous signalons cette méthode car l'algorithme d'alignement de séquences [6] utilisé pour construire notre classification (chapitre 2) fonctionne également grâce à l'identification de motifs conservés dans les parties transmembranaires des RCPG.

1.3.6 Classification par composition en acides aminés

Pour terminer ce tour d'horizon des méthodes de classification des RCPG, en voici une dernière qui utilise la composition en acides aminés des récepteurs pour leur assigner une sous-famille parmi 4 (récepteurs de l'acétylcholine, adrénorécepteurs, récepteurs de dopamine et récepteurs de sérotonine) de la famille Rhodopsine [14]. Cette classification ne permet pas d'avoir toutes les sous-familles à cause du manque de données statistiques (certaines sous-familles sont réduites à quelques récepteurs...) mais c'est une approche intéressante puisqu'elle est également indépendante de tout alignement.

1.3.7 Discussion

Les méthodes que nous avons discutées au long de ce chapitre sont diverses, de la classique phylogénie aux automates statistiques, en passant par des motifs conservés ou la composition en acide aminés des récepteurs.

Deux objectifs sont recherchés : le premier est l'obtention d'une classification complète, le second est la recherche de la meilleure classe pour un récepteur orphelin. Le choix de la méthode est déterminé par l'objectif ; ainsi on ne cherche pas à construire une classification complète basée sur la composition en acides aminés des récepteurs.

Le point faible de beaucoup de classifications, si on recherche un point de vue pharmacologique, consiste en un jeu de données inadapté car contenant des protéines de toutes espèces, et donc relativement peu de protéines humaines. De plus, souvent les classifications cherchent à placer tous les récepteurs dans un nombre fixé de classes, plutôt que de créer un certain nombre de classes en fonction de la diversité des récepteurs.

Nous allons donc essayer de construire une nouvelle classification des RCPG d'un

point de vue pharmacologique. Pour cela, nous allons nous focaliser sur les parties transmembranaires qui concentrent l'information de liaison de la majorité des ligands. Ainsi, on se libère du difficile problème d'aligner les séquences complètes des récepteurs. De plus, nous allons tenter de constituer un jeu de données aussi complet que possible, mais ne comportant que des protéines humaines. Le nombre de classes ne doit pas fixé en avance, mais on devra pouvoir cerner une classe de façon non ambiguë. Nous verrons que tous ces critères sont réalisés dans les classifications que nous proposons dans les chapitres 2 et 3.

1.4 Conclusion

Dans ce chapitre, nous avons présenté brièvement les RCPG, ainsi que quelques-unes de leurs classifications, notamment la classification par ligands, qui sert de référence puisqu'elle n'a été construite qu'avec des récepteurs dont les ligands sont connus. De l'analyse de ces classifications, nous voyons qu'elles présentent toutes des défauts si on les regarde d'un point de vue de pharmacologue, malgré l'intérêt de leurs méthodes. Mais ces défauts peuvent être surmontés, ce que nous allons faire dans les prochains chapitres.

Bibliographie

[1] Hamm H.E. (1998) The Many Faces of G Protein Signaling. *J. Biol. Chem.* **273(2)**: 669–672.

[2] Horn F., Weare J., Beukers M.W., Hörsch S., Bairoch A., Chen W., Edvardsen Ø., Campagne F., Vriend G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **26(1)**:275–279.

[3] Venter J.C. (2001) The Sequence of the Human Genome. *Science* **291**:1304–1351.

[4] Fredriksson R., Lagerström M.C., Lundin L.-G., Schiöth H.B. (2003) The G-Protein-Coupled Receptors in the Human Genome From Five Main Families Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Mol. Pharmacol.* **63(6)**:1256–1272.

[5a] Palczewski K., Kumasaka T., Hori T., Behnke C.A., Motoshima H., Fox B.A., Le Trong I., Teller D.C., Okada T., Stenkamp R.E., Yamamoto M., Miyano M. (2000) Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **289**:739–745.

[5b] Filipek S., Teller D.C., Palczewksi K., Stenkamp R. (2003) The crystallographic model of thodopsin and its use in studies of other G protein-coupled receptors. *Annu. Rev. Biophys. Biomol. Struct.* **32**:375–397.

[6] Bissantz C., Logean A., Rognan D. (2004) High-Throughput Modelling of Human G-Protein Coupled Receptors: Amino Acid Sequence Alignment, Tree-Dimensional Model Building, and Receptor Library Screening. *J. Chem. Inf. Comput. Sci.* **44(3)**: 1162–1176.

[7] Becker O.M. *et al.* (2003) Modeling the 3D structure of GPCRs: advances and aplication to drug discovery. *Curr. Opin. Drug Discovery Dev.***6(4)**:353–361.

[8] Evers A., Klebe G. (2004) Ligand-Supported Homology Modeling of G-Protein-Coupled Receptors Sites Sufficient for Successful Virtual Screening. *Angew. Chem. Int. Ed. Engl.* **43(2)**:248–251.

[9] Joost P., Methner A. (2002) Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. *Genome Biology* **3(11)**:1–16.

[10] Karchin R., Karplus K., Haussler D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **18(1)**:147–159.

[11] Qian B., Soyer O.S., Neubig R.R., Goldstein R.A. (2003) Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. *FEBS Lett.* **554**:95–99.

[12] Liu A.H., Califano A. (2003) CASTOR: Clustering Algorithm for Sequence Taxonomical Organization and Relationship. *J. Comput. Biol.* **10(1)**:21–45.

[13] Attwood T.K. (2001) A compendium of specific motifs for diagnosing GPCR subtypes. *Trends Pharmacol. Sci.* **22**:162–165.

[14] Elrod D.W., Chou K.C. (2002) A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Eng.* **15**:713–715.

[15] Lapinsh M., Gutcaits A., Prusis P., Post C., Lundstedt T., Wikberg J.E. (2002) Classification of G-protein coupled receptors by alignment-independent extraction

of principal chemical properities of primary amino acid sequences. *Protein Sci.* **11**:795–805.

[16] Otaki J.M., Mori A., Itoh Y., Nakayama T., Yamamoto H. (2006) Alignment-Free Classification of G-Protein-Coupled Receptors Using Self-Organizing Maps *J. Chem. Inf. Model.* (in press).

[17] Bockaert J., Pin J.P. (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J.* **18(7)**:1723–1729.

[18] Hulme E.C. *et al.* (1999) The conformational switch in 7-transmembrane receptors: the muscarinic receptor paradigm. *European J. of Parmacology* **375**:247–260.

[19] George S.R., O'Dowd B.F., Lee S.P. (2002) G-protein-coupled receptor oligomerization and its potential for drug discovery. *Nat. Rev. Drug Discovery* **1**:808–820.

[20] Flower D.R. (1999) Modelling G-protein-coupled receptors for drug design. *Biochim. Biophys. Acta* **1422(3)**:207–234.

[21] Seifert R. (ed) *et al.* (2005) G Protein-Coupled Receptors as Drug Targets: Analysis of Activation and Constitutive Activity. *Wiley.*

[22] Shaaban S., Benton B. (2001) Orphan G-protein-coupled receptors: from DNA to drug targets. *Curr. Opin. Drug Discovery Dev.* **4**:535–547.

[23] Bairoch A., Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28(1)**:45–48.

[24] Botto J.-M. Une classification des Récepteurs Couplés aux Protéines-G <URL:http://www.123bio.net/revues/rcpg/>.

[25] Saitou N., Nei M. (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* **4(4)**:406–425.

[26] Rabiner L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **72(2)**:257–286.

[27] Penrose R. (1997) Les ombres de l'esprit. *InterÉdition*.

Chapitre 2

Classification des RCPG basée sur des séquences de résidus critiques des cavités transmembranaires

Ce chapitre fait partie d'un article publié dans *PROTEINS: Structure, Function, and Bioinformatics* **62(2)** 509–538 (2006).

Résumé

Les classifications des RCPG établies jusqu'ici sont toutes basées sur des méthodologies très diverses et intéressantes, mais assez éloignées des applications pharmacologiques, soit dans l'objectif (comme par exemple la validation de méthodes de classification), soit dans le jeu de données inadapté (très incomplet ou comportant de nombreuses autres espèces que l'homme). Nous proposons ici une nouvelle classification des RCPG dont le

point de vue sera pharmacologique par l'étude de la cavité transmembranaire, un jeu de données aussi complet que possible et limité à des protéines humaines.

Pour cela, nous partons de 369 séquences de RCPG humains non-olfactifs extraites de la base de données UniProt. Pour chacune, nous extrayons 30 acides aminés critiques sélectionnés parmi les résidus de la cavité transmembranaire pointant vers l'intérieur de la cavité. Ces résidus forment, pour chaque récepteur, une séquence discontinue de 30 acides aminés. Puis nous effectuons un clustering hiérarchique agglomératif de ces séquences discontinues. Le résultat est un arbre dont la sémantique est proche de celle d'un arbre phylogénétique. Cette méthode suppose un alignement précis des parties transmembranaires.

Nous retrouvons les 5 familles de la classification GRAFS [1] publiée récemment, ainsi que beaucoup de sous-familles. Des applications de cette classification, comme la désorphanisation de récepteurs, sont immédiates : nous proposons des ligands potentiels pour certains récepteurs orphelins.

2.1 Introduction

Nous avons présenté dans le premier chapitre de ce mémoire plusieurs classifications des RCPG [1-8]. Elles sont très variées dans leurs méthodes :

- Fredriksson *et al.* [1] et Joost *et* Methner [2] proposent chacun une classification phylogénétique construite grâce à une méthode de distance entre séquences,
- Karchin *et al.* [3], Qian *et al.* [4] et Liu *et* Califano [5] utilisent des automates statistiques,

– Attwood [6] compare des empruntes de protéines,

– Elrod *et* Chou [7] utilisent les différences de composition en acides aminés entre des familles de protéines,

– enfin Lapinsh *et al.* [8] extraient des propriétés physico-chimiques des séquences, sans considération d'alignement.

Ces méthodes ont des objectifs divers, qui sont à l'origine des choix effectués. Par exemple, la classification générale d'une famille ou super-famille de protéines est souvent donnée par une classification phylogénétique. Mais la classification établie par Liu *et* Califano [5] ne sert que de base d'évaluation d'une méthode de classification taxonomique.

Malgré leurs succès et leurs intérêts, toutes ces méthodes posent des problèmes si on étudie la classification des RCPG d'un point de vue pharmcologique :

– Le jeu de données est souvent limité à quelques sous-familles (par exemple seulement 4 pour [7]) sur la vingtaine que proposent [1] et [2] ; dans ce cas les critères de discrimination sont spécifiques à ces familles plutôt qu'à un jeu plus général.

– De nombreuses espèces sont représentées, et en corollaire relativement peu de protéines humaines ; dans ce cas on va classifier en premier lieu les protéines homologues d'espèces différentes, et seulement ensuite on déterminera les sous-familles.

– Des informations peu pertinentes du point de vue pharmacologique sont utilisées, comme par exemple la séquence entière, alors qu'un ligand n'interagit qu'avec une partie seulement des résidus de la protéine.

C'est pourquoi nous proposons ici une nouvelle classification des RCPG, orientée vers des applications en pharmacologie. Elle est basée sur un clustering hiérarchique agglomératif de séquences discontinues formées par 30 résidus critiques de la cavité transmembranaire des récepteurs.

Cette nouvelle classification présente un avantage de simplicité dans la méthode, mais surtout elle se focalise sur l'information concentrée dans la cavité transmembranaire, importante dans l'interaction avec la majorité des ligands.

Mais notre classification doit sa pertinence à la qualité des alignements. La méthode est sensible car nous ne sélectionnons que 30 résidus dans une séquence de longueur variable allant de 300 à 6000 résidus. Un décalage dans un alignement sélectionnera peut-être des résidus qui n'interviennent pas dans une interaction potentielle (car pointant vers l'extérieur de la cavité par exemple). Nous sommes pourtant confiants dans la qualité de nos alignements, vérifiés à la main.

Nous n'avons pas inclus les récepteurs olfactifs dans notre classification pour deux raisons. D'abord ils ne sont pas intéressants du point de vue pharmacologique car ils n'interviennent que dans le mécanisme de l'olfaction [1]. Ensuite, Fredriksson *et al.* classent ces récepteurs dans une sous-famille de la famille de la rhodopsine, bien séparée des autres [1]. Nous avions inclu ces récepteurs dans un premier jeu de données pour notre classification, et il ressortait qu'ils formaient deux clusters homogènes bien séparés des autres. Nous avons donc supprimé ces récepteurs de notre jeu de données, sans que la classification n'en soit perturbée.

2.2 Méthode

La séquence d'opérations pour construire la classification est la suivante (figure 2-1) :

1. extraction du jeu de données de la base UniProt,

2. détection et alignement des parties transmembranaires,

3. sélection des résidus critiques à partir des parties transmembranaires,

4. construction de la matrice de distances entre séquences,

5. construction de 1000 matrices avec 10% de recopie de séquences (bootstrap),

6. construction de 1000 arbres par clustering hiérarchique en utilisant ces matrices,

7. normalisation des arbres (pour avoir dans chaque arbre les mêmes entrées),

8. enfin, construction de l'arbre consensus.

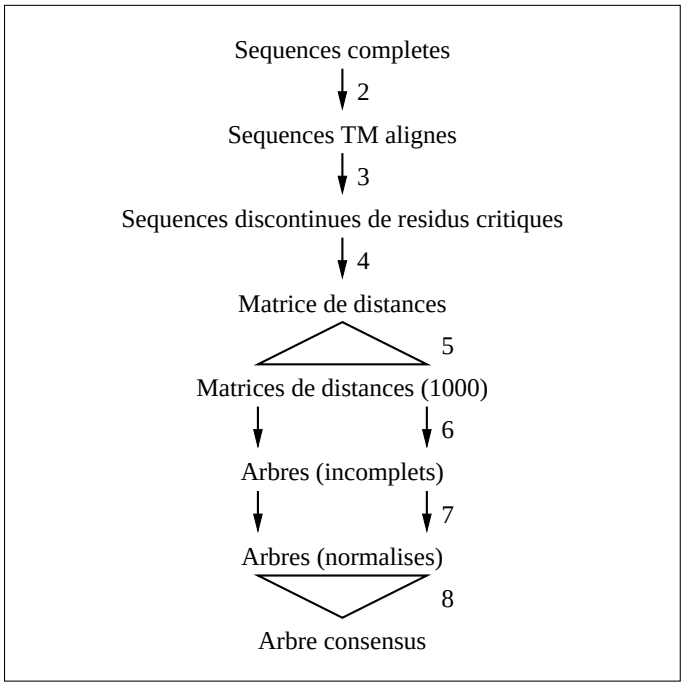


Figure 2-1. Procédure de classification. Les nombres correspondent au numéro de l'énumération précédente, ainsi qu'à la sous-section décrivant l'opération.

Les sous-sections suivantes détaillent chacune de ces opérations.

2.2.1 Les séquences des RCPG en entrée

Les séquences des RCPG sont tirées de la base de données de protéines UniProt (*the Universal Protein Resource*) [9]. Celle-ci est le résultat de la fusion de plusieurs bases d'instituts différents :

- les bases Swiss-Prot et TrEMBL [10], de l'EBI (*European Bioinformatics Institute*, localisé en Angleterre) et du SIB (*Swiss Institute of Bioinformatics*, Suisse),
- et la base PIR-PSD [11] (*Protein Sequence Database*) du PIR (*Protein Information Resource*, États-Unis).

Pour des raisons de continuité, les dénominations Swiss-Prot et TrEMBL sont restées, malgré la fusion des bases dans UniProt. La base Swiss-Prot est une base annotée et curée manuellement, alors que la base TrEMBL (*Translated EMBL Nucleotide Sequence Data Library*) est une base annotée automatiquement, et dont les entrées sont vouées à passer dans la Swiss-Prot au fur et à mesure de leur vérification manuelle. En 2002, les trois organismes cités se sont regroupés dans le consortium UniProt et ont fusionné leurs bases de données. La dernière version, UnitProt 6.0 datée du 13 septembre 2005, contient 194 317 protéines dans la partie Swiss-Prot et 2 105 517 protéines dans la partie TrEMBL. Ces bases sont accessibles en version hyper-texte ou téléchargeables dans leur entier (voir `www.uniprot.org`).

Le contenu d'une entrée de protéine comporte plusieurs parties : une description de la protéine, des références bibliographiques, des liens vers d'autres bases, les fonctions reconnues, et enfin la séquence d'acides aminés. Pour la version texte brut (que nous avons utilisée plutôt que la contrepartie XML, plus lourde), un exemple d'entrée est

donné à la figure 2-2. On y aperçoit le code d'identification de la protéine (champ ID), des dates d'entrée ou de modification (champ DT), une description (champ DE), des références de publications traitant de cette protéine (champs RP, RX, RA, RT, RL), des commentaires (champs CC), des liens vers d'autres bases de données (champ DR), des mots-clés (champ KW), la fonction de certains résidus (champs FT, à noter qu'on trouve ici des variantes de la séquence, toutes citées dans les références), et enfin la séquence des acides aminés (champ SQ).

```
ID   5HT1A_HUMAN    STANDARD;      PRT;   422 AA.
AC   P08908; Q6LAE7;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-APR-1993 (Rel. 25, Last sequence update)
DT   24-JAN-2006 (Rel. 49, Last annotation update)
DE   5-hydroxytryptamine 1A receptor (5-HT-1A) (Serotonin receptor 1A)
DE   (5-HT1A) (G-21).
GN   Name=HTR1A;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae;
OC   Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   MEDLINE=87315369; PubMed=3041227; DOI=10.1038/329075a0;
RA   Kobilka B.K., Frielle T., Collins S., Yang-Feng T.L., Kobilka T.S.,
RA   Francke U., Lefkowitz R.J., Caron M.G.;
RT   "An intronless gene encoding a potential member of the family of
RT   receptors coupled to guanine nucleotide regulatory proteins.";
RL   Nature 329:75-79(1987).
...
CC   -----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use as long as its content is in no way modified and this statement is not
CC   removed.
CC   -----------------------------------------------------------------------
DR   EMBL; M28269; AAA36440.1; -; Genomic_DNA.
...
KW   G-protein coupled receptor; Glycoprotein; Membrane; Multigene family;
KW   Polymorphism; Receptor; Transducer; Transmembrane.
FT   TOPO_DOM      1     36       Extracellular (Potential).
FT   TRANSMEM     37     62       1 (Potential).
```

```
FT   TOPO_DOM     63     73       Cytoplasmic (Potential).
FT   TRANSMEM     74     98       2 (Potential).
...
FT   VARIANT      16     16       P -> L (in dbSNP:1800041).
FT                                /FTId=VAR_003446.
...
SQ   SEQUENCE   422 AA;  46107 MW;  762664FCF62CFD8F CRC64;
     MDVLSPGQGN NTTSPPAPFE TGGNTTGISD VTVSYQVITS LLLGTLIFCA VLGNACVVAA
     IALERSLQNV ANYLIGSLAV TDLMVSVLVL PMAALYQVLN KWTLGQVTCD LFIALDVLCC
...
     RQ
//
```

Figure 2-2. Extrait de l'entrée 5HT1A_HUMAN de la Swiss-Prot.

Nous sommes partis de la version complètement téléchargée, et nous avons appliqué plusieurs filtres :

- ne conserver que les protéines humaines : le champ `OS` (pour "organism") doit comporter la valeur `Homo sapiens (Human)` ;
- ne conserver que les RCPG : le champ `KW` ("keyword") doit comporter la valeur `G-protein coupled receptor`, entre autre. Il faut noter que ce critère n'élimine aucun RCPG, même supposé, car même pour ces derniers la valeur indiquant un RCPG apparait ;
- ne conserver que les RCPG non olfactifs : le champ `KW` ne doit pas comporter la valeur `Olfaction`. Nous avons parlé dans l'introduction à ce chapitre d'une première classification que nous avons faite incluant ces récepteurs et dans laquelle ils formaient deux clusters homogènes. Nous accordons donc notre confiance dans le mot-clé `Olfaction` ;

– ne pas conserver les fragments (séquences incomplètes) : le champ `DE` ("description") ne doit pas comporter la chaîne de caractères `fragment` (en majuscules ou minuscules) ;

– enfin, ne pas conserver les séquences dont les parties transmembranaires sont redondantes avec une autre séquence (identité complète des parties transmembranaires) : cela a été fait après l'alignement à l'aide d'un script simple. Ces duplicata sont des entrées TrEMBL non encore annotées, soit des isoformes de protéines, soit des protéines mutées.

D'autres RCPG ont été ajoutés au fur et à mesure de leur entrée dans la base UniProt, en vérifiant bien qu'ils passaient tous les filtres précités. Les données de cette base sont en effet actualisées assez souvent, ainsi que l'architecture même de la base. Par exemple, entre le début et la fin de ce travail de thèse, la nomenclature des protéines a changé : la longueur autorisée pour les identifiants d'entrées (sans le suffixe de l'organisme) est passée de 4 à 5 caractères, à cause du nombre croissant d'entrées. La base elle-même a fait l'objet de la fusion de plusieurs bases antérieures (de Swiss-Prot à UniProt). De nombreuses séquences, numéros d'accès et annotations ont changé. Des entrées ont bien sûr été ajoutées, mais certaines ont été supprimées (certains fragments ont été complétés, on a donc supprimé le fragment au profit de la séquence complète). En dernier exemple, le format XML a été introduit afin de stocker les données dans le format d'un standard reconnu, et d'utiliser tous les outils associés à ce format. Pour donner un ordre d'idée sur la vitesse d'évolution, en moyenne 1 nouvelle version (majeure) est éditée tous les 4 mois, sans compter les version mineures qui peuvent n'être espacées que de 2 semaines. Cela donne une idée sur la bonne santé de ce champ de recherche.

Au final, nous avons donc 369 séquences de RCPG humains non-olfactifs, dont les parties transmembranaires sont non redondantes. La liste de ces RCPG est donnée dans la section consacrée aux résultats.

2.2.2 Alignements des parties transmembranaires

Toutes ces séquences doivent ensuite être alignées. Nous ne nous intéressons qu'à l'alignement des parties transmembranaires puisque nous nous focalisons sur la cavité transmembranaire.

À partir d'ici, nous allons émettre l'hypothèse que le repliement des parties transmembranaire a été conservé durant l'évolution. Alors les 7 hélices transmembranaires ont la même longueur et la même inclinaison dans la membrane, ainsi la longueur des parties transmembranaires est la même pour tous les RCPG. Deux indices vont dans ce sens, au moins pour les récepteurs de classe Rhodopsine :

- certains résidus sont très conservés malgré une identité de séquence faible (25–30%) entre les récepteurs de classe Rhodopsine, suggérant une architecture commune de la poche [12] ;
- Ballesteros *et al.* [13] suggèrent qu'un même repliement serait partagé entre la rhodopsine et les récepteurs d'amines, malgré une identité entre les séquences des parties transmembranaires d'à peine 20 à 25% avec la rhodopsine humaine.

Pour aligner les parties transmembranaires, nous avons utilisé le programme GPCRalign (du paquetage GPCRmod [14]) développé au laboratoire. En peu de mots, ce programme prédit d'abord grossièrement la localisation des parties transmembranaires grâce

à l'algorithme TMHMM (*TransMembrane Hidden Markov Model*) [15] [16] (je suggère la référence [17] pour une description mathématique claire des modèles de Markov cachés), puis raffine les alignements, TM par TM, en détectant des empreintes spécifiques de chaque famille et de chaque TM [14].

Pour les cas où GPCRalign ne peut pas déterminer l'alignement d'une partie TM, on utilisera les informations des champs FT de la Swiss-Prot (figure 2-2), ou d'autres suggestions d'alignements de la base GPCRDb [18] pour aligner à la main le récepteur avec d'autres déjà correctement alignés entre eux.

Après l'application de GPCRalign, nous avons raffiné les alignements à la main, pour vérifier que les résidus massivement conservés (notés en position 50 dans la nomenclature de Ballesteros [20]) soient correctement alignés. Cette vérification est nécessaire : c'est là un point important car de la qualité des alignements dépend de façon critique la pertinence de notre classification.

2.2.3 Sélection de 30 acides aminés critiques

Les 7 hélices transmembranaires forment une cavité. Pour trouver les résidus qui en tapissent l'intérieur et sélectionner parmi eux des résidus critiques dans la liaison avec le ligand, nous étudions la structure cristallographique de la rhodopsine bovine [19], seule structure cristallographique de RCPG résolue actuellement. On définit la cavité par 81 résidus compris dans une sphère de 10 Å autour du ligand. La surface accessible au solvant est ensuite calculée. Finalement, nous sélectionnons 30 résidus qui présentent une surface accessible au solvant de plus de 25% et dont la chaîne latérale pointe vers l'intérieur de la cavité (figure 2-3). Ce choix dérive d'une étude statique du récepteur, et ne tient pas

compte des aspects dynamiques.

L'hypothèse selon laquelle la cavité a conservé sa forme au cours de l'évolution est encore nécessaire. Les indices donnés précédemment qui confortent cette hypothèse indiqueraient que ces 30 résidus joueraient fréquemment un rôle dans les interactions avec les ligands.

Le tableau de la figure 2-4 présente la liste des 30 résidus sélectionnés, dans la notation de Ballesteros (de la forme N.MM où N est le numéro de l'hélice transmembranaire, et MM est le numéro de résidu par rapport à l'hélice, où 50 est le numéro du résidu le plus conservé de ce TM [20]).

À ce stade, nous avons un jeu de données de 369 séquences discontinues de 30 acides aminés.

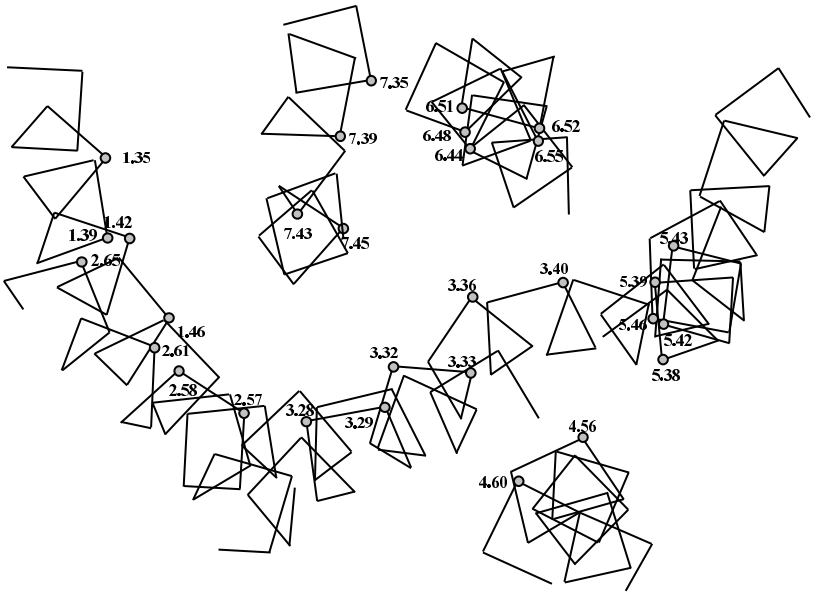


Figure 2-3. Vue du dessus des 30 résidus critiques, en notation de Ballesteros [20].

TM	résidu/TM	notation de Ballesteros	TM	résidu/TM	notation de Ballesteros
1	6	1.35	4	21	4.60
1	10	1.39	5	4	5.38
1	13	1.42	5	5	5.39
1	17	1.46	5	8	5.42
2	20	2.57	5	9	5.43
2	21	2.58	5	12	5.46
2	24	2.61	6	15	6.44
2	28	2.65	6	19	6.48
3	7	3.28	6	22	6.51
3	8	3.29	6	23	6.52
3	11	3.32	6	26	6.55
3	12	3.33	7	3	7.35
3	15	3.36	7	7	7.39
3	19	3.40	7	11	7.43
4	17	4.56	7	13	7.45

Figure 2-4. Résidus sélectionnés. La colonne « résidu/TM » donne le numéro du résidu par rapport au début du TM (1 est le 1^{er} résidu) ; la colonne « notation de Ballesteros » donne le même résidu en notation de Ballesteros.

Il est à noter que la totalité de ces 30 résidus critiques sont des points d'ancrage de

ligands connus (agonistes inverses et antagonistes) validés par mutagenèse dirigée [18].

2.2.4 Méthodes de classification

Il existe bien des méthodes de classification. Parmi elles, Jain *et al.* [21] passent en revue quelques techniques de *clustering* (nous en avons rencontré plusieurs dans la première partie de ce mémoire).

Mais tout d'abord, notons un point de vocabulaire particulier : la différence entre les mots *classification* et *clustering* (ce dernier terme pourrait être traduit par *agrégation*). Dans la langue française, seul le mot classification est utilisé, mais dans la langue anglaise, les deux mots ont des sens différents.

- La *classification* est la « répartition systématique en classes, en catégories, d'êtres, de choses ou de notions ayant des caractères communs, notamment afin d'un faciliter l'étude » (Trésor de la Langue Française). On voit que les objets peuvent être vagues (des notions) et rien n'est précisé sur la façon dont la répartition se fait.

- D'un autre côté, le *clustering* (de données) est « une technique d'analyse de données [...] qui consiste dans le partitionnement d'un jeu de données en sous-ensembles (*clusters*), tel que les données de chaque sous-ensemble partagent un trait commun, souvent la similarité ou la proximité liées à une mesure de distance » (traduction française de l'encyclopédie Wikipedia anglaise).

La différence est que le clustering est formel et basé sur des critères et une méthode précis, alors que la classification est plus générale et pas nécessairement formelle. Un homme pourra par exemple classer des tableaux en deux ou trois ensembles : ceux qui lui

plaisent, ceux qui lui déplaisent, et ceux qui le laissent indifférent. C'est un critère non formel. Le clustering est un cas particulier de classification.

Finalement, je choisis d'utiliser dans la suite de ce mémoire les termes anglais de *clustering* et de *cluster*. Ces termes n'ont pas, à mon avis, d'équivalents français satisfaisants, et les termes anglais sont tout à fait explicites.

Parmi les différentes méthodes de clustering, on distingue les méthodes hiérarchiques des méthodes partitionnelles. Les premières utilisent les clusters précédemment établis pour construire un nouveau cluster, récursivement. Les secondes déterminent tous les clusters en même temps.

Parmi les méthodes hiérarchiques, on distingue encore des méthodes agglomératives et des divisives. Les premières partent d'objets et vont les agglomérer au fur et à mesure, les plus proches d'abord, jusqu'à agglomération complète. On parle aussi de méthodes descendantes car, si on imagine un arbre racine en bas et feuilles en haut, on part des feuilles et on progresse jusqu'à la racine. Les secondes méthodes vont diviser le jeu complet d'objets en unités plus petites, récursivement, jusqu'à atteindre les objets eux-mêmes. On parle encore de méthodes ascendantes : on part de la racine (cluster qui contient tous les objets de départ) et on progresse vers les feuilles. Le vocabulaire ascendant/descendant est purement conventionnel. On peut aussi trouver des représentations d'arbres où la racine est en haut et les feuilles en bas... Comme on lit de haut en bas, la représentation choisie dépendra de ce qu'on veut montrer d'abord, la racine ou bien les feuilles.

Le résultat d'un clustering est un arbre (c'est la structure de données abstraite), qu'on visualisera sous la forme d'un dendrogramme (c'est la représentation graphique de l'arbre) (figure 2-5). Il schématise l'imbrication des clusters les uns dans les autres, ainsi

que la distance entre les clusters.

Nous justifions maintenant le choix de la méthode utilisée.

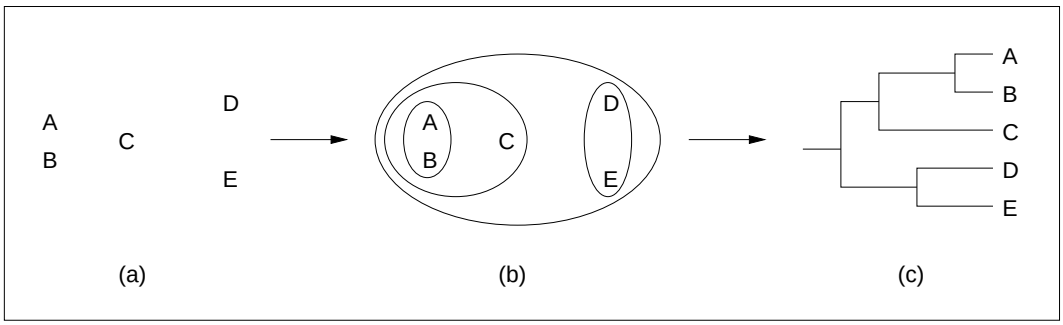


Figure 2-5. Clustering de 5 objets et dendrogramme associé. (a) Les 5 objets de départ. (b) Clustering de ces objets. Les contours schématisent les clusters imbriqués. (c) Dendrogramme qui représente le clustering.

2.2.5 Choix d'une méthode

Nous avons opté pour une méthode hiérarchique plutôt que partitionnelle car les relations entre protéines sont de nature hiérarchique par le mécanisme de l'évolution.

Nous avons choisi une méthode agglomérative pour construire notre classification. Le critère d'agglomération était assez immédiat : nous prenons les deux récepteurs les plus proches (concernant les 30 résidus sélectionnés) et formons un cluster qui les contient. Un tel critère pour une méthode divisive n'est pas aussi évident : comment diviser un ensemble de récepteurs ? Par exemple sur le résidu le plus conservé du cluster : le premier sous-cluster comprendra les récepteurs qui contiennent ce résidu, l'autre cluster prendra les récepteurs qui ne contiennent pas ce résidu. Ce critère donne bien sûr lieu à un clustering, que nous avons généré, mais nous n'avons pas pu lui trouvé de sens.

Dans le cas où nous aurions plusieurs protéines de plusieurs espèces, nous aurions plu-

sieurs possibilités de classification. Par exemple, on pourrait mélanger toutes les séquences et en faire une classification globale. Alors on peut détecter les protéines similaires entre plusieurs organismes et déduire qu'elles ont peut-être la même fonction, surtout si les organismes sont proches. On pourrait aussi classer d'abord par organisme et ensuite par familles de protéines, et déduire de la superposition des classifications des informations similaires, mais plus précises : on voit immédiatement quelles familles sont présentes ou non dans différents organismes, ce qui n'était pas évident avec la classifcation précédente. Cette classification à 2 dimensions est plus informative.

Comme dans notre cas nous ne considérons que des protéines humaines, le problème précédent ne se pose pas et nous pouvons utiliser une méthode simple et un classement mono-dimensionnel. Nous avons choisi la méthode UPGMA [22] (*Unweighted Pair Group Method with Arithmetic mean*), qui est hiérarchique et agglomérative. D'autres méthodes sont décrites dans [23].

2.2.6 La méthode UPGMA

Le fonctionnement est simple : on utilise un ensemble de clusters dont chacun ne contient au départ qu'une seule séquence. Puis on forme un nouveau cluster en regroupant les deux clusters les plus proches. Ces derniers sont remplacés dans l'ensemble par le nouveau cluster formé. On itère ainsi jusqu'à ce que l'ensemble ne contienne plus qu'un seul cluster. Les liens d'inclusion entre les clusters servent à construire un arbre. Si un cluster X est inclu dans un cluster Y, alors X sera un fils du nœud Y dans l'arbre. Par exemple, sur la figure 2-5, le cluster formé par A et B, qu'on note (AB), est inclu dans le cluster (ABC), donc (AB) est un fils de (ABC) dans l'arbre. Le dernier cluster restant

dans l'ensemble sera la racine de l'arbre.

Soient deux clusters X et Y. La distance $d(X,Y)$ entre eux est un paramètre de l'algorithme. Elle est définie à partir de la distance $d(x,y)$ entre les éléments (voir sous-section suivante) par :

$$d(X,Y) = \frac{1}{\mid X \mid\mid Y \mid} \sum_{x \in X, y \in Y} d(x,y)$$

Cette distance possède l'avantage de se calculer rapidement par récursion : elle permet en effet de supprimer la sommation systématique pour un calcul plus rapide. Soient I, J et K trois clusters, et si $A = (I,J)$ est le résultat de l'agglomération de I et J, alors :

$$d(A,K) = \frac{\mid I \mid d(I,K) + \mid J \mid d(J,K)}{\mid I \mid + \mid J \mid}$$

car (application de la définition à $d(A,K)$) :

$$d(A,K) = \frac{1}{\mid A \mid\mid K \mid} \sum_{a \in A, k \in K} d(a,k)$$

Mais $A = I \cup J$ (où I et J sont disjoints), donc

$$d(A,K) = \frac{1}{(\mid I \mid + \mid J \mid) \mid K \mid} \sum_{i \in I, j \in J, k \in K} (d(i,k) + d(j,k))$$

En appliquant à nouveau la définition, mais dans l'autre sens :

$$d(A,K) = \frac{1}{(\mid I \mid + \mid J \mid) \mid K \mid} (\mid I \mid\mid K \mid d(I,K) + \mid J \mid\mid K \mid d(J,K))$$

et en simplifiant, on retrouve la formule sans sommation. À partir de cette formule, on voit que si l'on conserve les distances entre tous les clusters dans une matrice (la « matrice de distances »), le clustering sera rapide : on ne fait qu'appliquer la formule sans somme.

La méthode UPGMA est très simple, mais une subtilité est à noter : le fait de devoir faire un choix dans une étape de l'algorithme. Quand on veut agglomérer les clusters les

plus proches, si plusieurs paires de clusters sont également proche et de distance minimale, alors il faut choisir par quelle paire commencer. Ce choix, qui peut influencer le résultat final, est fait de façon aléatoire.

2.2.7 Distance entre séquences

La méthode UPGMA nécessite une distance euclidienne entre éléments. Nullement nécessaire lors de la description de la méthode elle-même, elle le devient lors de l'application de la méthode à une classification particulière.

La distance $d(X, Y)$ entre deux clusters qui ne contiennent chacun qu'un seul élément ($X = \{x\}$, $Y = \{y\}$) est la distance $d(x, y)$ entre les deux séquences discontinues de résidus. Ces deux distances sont bien distinctes, bien que notées d toutes les deux ; la première est une distance entre ensembles (les clusters) alors que la seconde est une distance entre séquences.

Les deux séquences discontinues de résidus ont la même longueur L ($= 30$). La distance entre ces séquences est définie par :

$$d(x, y) = L - id(x, y)$$

où $id(x, y)$ est le score d'identité entre les deux séquences.

2.2.8 Le bootstrapping

Le *bootstrapping* est une technique de validation statistique du clustering. La procédure que nous avons appliquée est la suivante : au lieu de générer directement un arbre à partir de la matrice de distances, nous générons d'abord un grand nombre (1000) de matrices

intermédiaires dans lesquelles nous supprimons certaines entrées au hasard (10% des entrées). Puis nous construisons les arbres à partir de ces matrices. Nous complétons ensuite chacun des arbres avec les entrées manquantes que nous ajoutons au plus proche (la séquence la plus proche sera transformée en un nœud dont les deux fils seront cette même séquence et l'entrée ajoutée). Enfin nous construisons l'arbre consensus entre tous ces arbres.

Le consensus est effectué par le programme CONSENSE du paquetage PHYILP [24] développé par Joseph Felsenstein. Cette suite de programmes, gratuite, est toujours maintenue par son auteur.

La construction de l'arbre consensus apporte deux informations : la connectivité de l'arbre (comment lier les différentes entrées pour mimer au mieux les 1000 arbres intermédiaires) ainsi que des poids statistiques pour chacun des nœuds de l'arbre. Pour le choix des sous-arbres qui doivent apparaître dans l'arbre consensus, le programme CONSENSE applique une règle de majorité : tous les sous-arbres qui apparaissent dans plus de 50% des arbres sont inclus. Bien sûr, toutes les entrées sont présentes dans l'arbre consensus. Les poids statistiques comptent le nombre d'occurences de chaque sous-arbre dans chacun des arbres de départ. À un nœud donné, plus le poids est important, plus le sous-arbre dont la racine est ce nœud sera présent de manière stricte dans les arbres de départ.

Enfin, l'arbre consensus est visualisé à l'aide du programme TreeView [25]. Ce programme permet de visualiser les distances entre deux séquences, mais aussi les poids statistiques sur tous les nœuds de l'arbre.

Le format des arbres utilisé par tous ces programmes est un « standard » *de facto* :

le format Newick (aussi dénommé *New Hampshire standard form*). C'est un format très naturel pour les arbres utilisé par un grand nombre de programmes de phylogénie. On pourra regretter l'absence de texte de référence sur ce standard, mais on trouvera une description à l'URL suivante :

`<URL:http://evolution.genetics.washington.edu/phylip/newicktree.html>`

Pour terminer, remarquons qu'en plus de la validation statistique, la procédure de bootstrapping permet de prendre en compte les différents choix possibles dans le cas où une plus petite distance apparait plusieurs fois dans la matrice de distances, puisqu'on effectue un grand nombre de clusterings avec un choix aléatoire pour chacun des cas.

2.3 Résultats et discussion

Nous avons généré une classification des RCPG basée sur le clustering hiérarchique agglomératif de séquences discontinues de résidus critiques des cavités transmembranaires. Nous allons maintenant décrire et analyser le résultat obtenu et donner des applications de cette classification.

2.3.1 Description de l'arbre obtenu

Les 369 RCPG humains non olfactifs sont répartis dans l'arbre consensus en 22 clusters, dont 18 pour la famille Rhodopsine. Ces 18 clusters décrivent des sous-familles de la famille Rhodopsine et les 4 clusters restants décrivent les autres familles (Adhésion, Frizzled, Glutamate et Secrétine). La figure 2-6 présente l'arbre des clusters, et le tableau de la figure 2-7 détaille le contenu de chaque cluster. Nous omettons le suffixe "_HUMAN" pour les entrées de la Swiss-Prot pour ne pas alourdir la notation : nous ne considérons que des protéines humaines. L'arbre complet est donné en Annexe A.

Les clusters ont été déterminés en regardant la valeur de bootstrapping des nœuds. Si un nœud donné possède une valeur de bootstrapping beaucoup plus élevée que le nœud père, alors nous « coupons » l'arbre à ce nœud, et le sous-arbre associé à ce nœud formera un cluster. Ce critère est purement subjectif, mais permet un découpage. Nous insistons sur le fait que les clusters n'ont pas été formés pour qu'ils correspondent aux sous-familles, mais ont été créés d'abord, puis nous avons étudié leur correspondance avec les familles des RCPG et les sous-familles de la Rhodopsine.

Parmi tous les récepteurs, 32 n'ont pu être attribués à des clusters car ils se situent

proche de la racine de l'arbre (voir Annexe A et figure 2-7), avec des valeurs de bootstrapping très faibles qui indiquent une difficulté de placement dans l'arbre consensus. Plutôt que de créer de minuscules clusters ne contenant que de 1 à 3 récepteurs, nous avons préféré les laisser non attribués. Ces récepteurs non assignés, ou singletons, ne sont pas représentés sur la figure 2-6 pour des raisons de lisibilité.

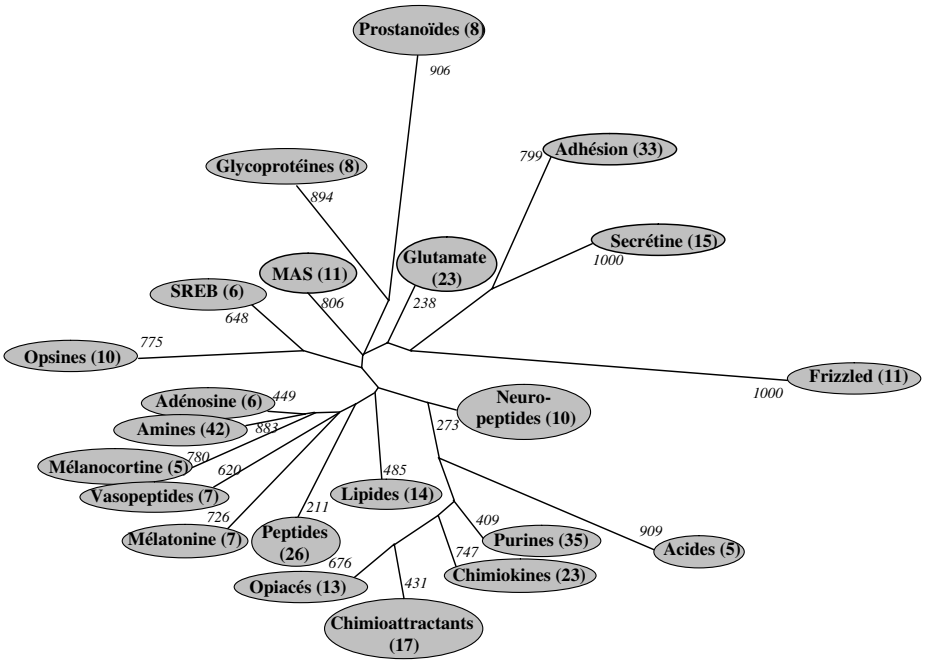


Figure 2-6. Clustering hiérarchique agglomératif des RCPG. Les nombres entre parenthèses indiquent le nombre d'entrées pour chaque cluster. Les nombres en italique sont les valeurs de bootstrapping des clusters. Les récepteurs non classés (singletons) ne sont pas affichés pour des raisons de lisibilité.

```
G-protein-coupled receptors
  Rhodopsin
    Acids
      G109B           G protein-coupled receptor HM74
      GPR31           Probable G protein-coupled receptor GPR31
      GPR81           FKSG80 protein
      Q8TDS4          G protein-coupled receptor HM74a
      Q8TDS5          Putative G-protein coupled receptor
    Adenosine
      AA(1,2A,2B,3)R  Adenosine receptors
      GNRHR           Gonadotropin-releasing hormone receptor
      GNRR2           GnRH-II-R
    Amines
      5HT(1[ABDEF],2[ABC],[467]R,5A)  5-hydroxytryptamines
      ACM[1-5]        Muscarinic acetylcholine receptors
      ADA(1[ABD],2[ABC])  Alpha adrenergic receptors
      ADRB[1-3]       Beta adrenergic receptors
      DRD[1-5]        Dopamine receptors
      GPR61           Probable G protein-coupled receptor GPR61
      GPR62           hGPCR8
      HRH[1-4]        Histamine receptors
      O14804          Putative neurotransmitter receptor
      TAR0[1345]      Trace amine receptors
    Brain-gut peptides
      GHSR            Growth hormone secretagogue receptor type 1
      GPR39           Putative G protein-coupled receptor GPR39
      MCHR1           Probable G protein-coupled receptor GPR24
      MCHR2           G protein-coupled receptor
      MTLR            Motilin receptor
      NTR[12]         Neurotensin receptors
      Q9GZQ4          Neuromedin U receptor 2
      Q9HB89          Neuromedin U receptor 1
      TRFR            Thyrotropin-releasing hormone receptor
    Chemoattractants
      AG2[2RS]        Angiotensin II receptors
      APJ             Apelin receptor
      BKRB[12]        Bradykinin receptors
      C[35]AR,C5ARL   Anaphylatoxin chemotactic receptors
      CML1            Chemokine receptor-like 1
      FPR1            fMet-Leu-Phe receptor
      FPRL[12]        FMLP-related receptors
      GPR1            Probable G protein-coupled receptor GPR1
      GPR15           G protein-coupled receptor GPR15 (BOB)
      GPR25           Probable G protein-coupled receptor GPR25
      GPR44           Putative G protein-coupled receptor GPR44
    Chemokines
      ADMR            Adrenomedullin receptor
      C3X1            CX3C chemokine receptor 1
      CCBP2           Chemokine binding protein 2
      CCR[1-9],CCR10,CCRL1  Chemokine receptors
      CXCR[12]        High affinity interleukin-8 receptors
      CXCR[3-6]       C-X-C chemokine receptors
      O75307          Putative chemokine receptor
```

```
  RDC1            G protein-coupled receptor RDC1 homolog
  XCR1            Chemokine XC receptor 1
Glycoproteins
  FSHR            Follicle stimulating hormone receptor
  LGR[4-6]        Leucine-rich repeat-containing G protein-coupled receptors
  LGR[78]         Relaxin receptors
  LSHR            Lutropin-choriogonadotropic hormone receptor
  TSHR            Thyrotropin receptor
Lipids
  CNR[12]         Cannabinoid receptors
  EDG1            Probable G protein-coupled receptor EDG-1
  EDG[247]        Lysophosphatidic acid receptors
  EDG[35]         Lysosphingolipid receptors
  EDG6            Putative G-protein coupled receptor, EDG6
  EDG8            Sphingosine 1-phosphate receptor Edg-8
  GP119           G protein-coupled receptor 119
  GPR12           Probable G protein-coupled receptor GPR12
  GPR3            Probable G protein-coupled receptor GPR3
  GPR6            Probable G protein-coupled receptor GPR6
MAS
  MAS             MAS proto-oncogene.
  MAS1L,MRGR[DEF] Mas-related G protein-coupled receptors
  MRGX[1-3]       G protein-coupled receptors
  MRGX4           G protein-coupled receptor SNSR6
  SNSR[25]        G protein-coupled receptors
Melanocortin
  ACTHR           Adrenocorticotropic hormone receptor
  MC[3-5]R        Melanocortin receptors
  MSHR            Melanocyte stimulating hormone receptor
Melatonin
  GPR22           Probable G protein-coupled receptor GPR22
  GPR45           Probable G protein-coupled receptor GPR45
  GPR63           Probable G protein-coupled receptor GPR63
  MTR1[ABL]       Melatonin receptors
  O43898          High-affinity lysophosphatidic acid receptor homolog
Opiates
  GPR7            Probable G protein-coupled receptor GPR7
  GPR8            Probable G protein-coupled receptor GPR8
  OPRD            Delta-type opioid receptor
  OPRK            Kappa-type opioid receptor
  OPRM            Mu-type opioid receptor
  OPRX            Nociceptin receptor
  R3R1            Somatostatin- and angiogenin-like peptide receptor
  R3R2            Relaxin 3 receptor 2
  SSR[1-5]        Somatostatin receptors: SSR1-5
Opsins
  OPN3            Opsin 3
  OPN4            Opsin 4
  OPSB            Blue-sensitive opsin
  OPSD            Rhodopsin
  OPSG            Green-sensitive opsin
  OPSR            Red-sensitive opsin
  OPSX            Visual pigment-like receptor peropsin
  Q6U736          Neuropsin
  Q9UQS0          Photopigment apoprotein
```

```
  RGR             RPE-retinal G protein-coupled receptor
Peptides
  BRS3            Bombesin receptor subtype-3
  CCKAR           Cholecystokinin type A receptor
  EDNR[AB]        Endothelin receptors
  GALR[1-3]       Galanin receptors
  GASR            Gastrin/cholecystokinin type B receptor
  GPR10           Prolactin-releasing peptide receptor
  GPR19           Probable G protein-coupled receptor GPR19
  GPR83           Probable G protein-coupled receptor GPR83
  GRPR            Gastrin-releasing peptide receptor
  KISSR           G protein-coupled receptor
  NK1R            Substance-P receptor
  NK2R            Substance-K receptor
  NK3R            Neuromedin K receptor
  NMBR            Neuromedin-B receptor
  NPFF[12]        Neuropeptide FF receptors
  NPY[1245]R      Neuropeptide Y receptors
  OX[12]R         Orexin receptors
  QRFPR           Orexigenic neuropeptide QRFP receptor
Prostanoids
  PD2R            Prostaglandin D2 receptor
  PE2R[1-4]       Prostagrandin E2 receptors
  PF2R            Prostaglandin F2-alpha receptor
  PI2R            Prostacyclin receptor
  TA2R            Thromboxane A2 receptor
Purines
  CLTR[12]        Cysteinyl leukotriene receptors
  EBI2            EBV-induced G protein-coupled receptor 2
  G2A             G protein-coupled receptor
  GP171           Probable G protein-coupled receptor 171
  GP174           Putative P2Y purinoceptor FKSG79
  GPR(17,34,4,4[0-3],8[067],9[12])   Probable G protein-coupled receptors
  P2RY[124569]    P2Y purinoceptors
  P2Y10           Putative P2Y purinoceptor 10
  P2Y1[12]        P2Y purinoceptors
  P2Y14           UDP-glucose receptor
  PAR[1-4]        Proteinase activated receptors
  PSYR            T cell-death associated protein (GPR65)
  PTAFR           Platelet activating factor receptor
  SPR1            Probable G protein-coupled receptor GPR68
SREBs
  GP1(01,61,73)   G protein-coupled receptors
  GPR(27,85)      Probable G protein-coupled receptors
  Q14439          G protein-coupled receptor
Vasopeptides
  OXYR            Oxytocin receptor
  PKR[12]         Prokineticin receptors
  Q6W5P4          GPRA isoform A
  V(1[AB],2)R     Vasopressin receptors
Singletons
  CML2            Chemokine receptor-like 2
  DUFFY           Duffy antigen/chemokine receptor
  ETBR2           Endothelin B receptor-like protein-2
  GP1(20,35,41,42,5[012],60)    G protein-coupled receptors
```

```
    GPR(18,2[016],3[57],5[25],82)   Probable G protein-coupled receptors
    GPR78           UNQ5925/PRO19818
    LT4R[12]        Leukotriene B4 receptors
    O95800          G-protein coupled receptor
    Q16538          Protein A-2
    Q6DWJ6          G protein-coupled receptor 139
    Q6NV75          G protein-coupled receptor 153
    Q8TDU6          G-protein coupled bile acid receptor BG37
    Q8TDU8          Putative G-protein coupled receptor
    Q9H2L2          GPR18-iso
    Q9NQS5          Inflammation-related G protein-coupled receptor EX33
  Special case
    GPR88           Striatum-specific G potein-coupled receptor GPR88
Adhesion
  BAI[1-3]          Brain-specific angiogenesis inhibitors
  CD97              Leucocyte antigen CD97
  CELR[1-3]         Cadherin EGF LAG seven-pass G-type receptors
  ELTD1             latrophilin and seven transmembrane domain containing protein 1
  EMR1              Cell surface glycoprotein EMR1
  EMR2              EGF-like module EMR2
  EMR[34]           EGF-like module containing mucin-like hormone receptor-like 3
  GP1(1[0-6],2[3-68],33,44)   G protein coupled receptors
  GPR56             G protein-coupled receptor 56
  GPR64             Epididymis-specific protein 6
  GPR97             G protein-coupled receptor 97
  LPHN1             Lectomedin-2
  LPHN2             Lectomedin-1 beta
  LPHN3             Lectomedin-3
  Q8WXG9            Very large G protein-coupled receptor 1b
  Singletons
    GP143           G protein-coupled receptor 143
Frizzled
  FZ(10,D[1-9])     Frizzled
  SMO               Smoothened homolog
Glutamate
  CASR              Extracellular calcium-sensing receptor
  GABR[12]          Gamma-aminobutyric acid type B receptor, subunits
  GPC5B             A-69G12.1
  GPC5C             RAIG-3
  GPC5D             G protein-coupled receptor family C group 5 member D
  MGR[1-8]          Metabotropic glutamate receptors
  Q6QR81            G protein coupled receptor 158
  Q8NFN8            GABAB-related G-protein coupled receptor
  Q8NHZ9            GPCRC6A
  Q8TDU1            PREDICTED: similar to seven transmembrane helix receptor
  RAI3              Orphan G protein-coupling receptor PEIG-1
  TS1R[1-3]         Taste receptors
Secretine
  CALCR             Calcitonin receptor
  CALRL             Calcitonin gene-related peptide type 1 receptor
  CRFR[12]          Corticotropin releasing factor receptors
  GHRHR             Growth hormone-releasing hormone receptor
  GIPR              Gastric inhibitory polypeptide receptor
  GLP[12]R          Glucagon-like peptide receptors
  GLR               Glucagon receptor
```

```
PACR            Pituitary adenylate cyclase activating polypeptide type I receptor
PTHR1           Parathyroid hormone/parathyroid hormone-related peptide receptor
PTHR2           Parathyroid hormone receptor
SCTR            Secretin receptor
VIPR[12]        Vasoactive intestinal polypeptide receptors
```

Figure 2-7. Liste complète des récepteurs pour chaque cluster, ainsi que des singletons. Chaque récepteur contient une courte description tirée de la base Swiss-Prot. Cette liste est en anglais, pour éviter de traduire certains noms, en anglais même dans les textes français que nous avons rencontrés. La notation crochet indique une liste de caractères (par exemple `[a-df]` dénote `a`, `b`, `c`, `d` et `f`). La notation virgule indique aussi une liste, mais de chaîne de caractères. La virgule peut être accompagnée de parenthèses pour clarifier sa portée. Ces notations sont inspirées des expressions régulières utilisées en informatique [26].

2.3.2 Comparaison avec la classification GRAFS

Nous comparons dans cette sous-section les résultats de notre classification avec ceux de la classification GRAFS [1], avant-dernière en date et la plus complète parmi les huit présentées dans le chapitre 1. C'est aussi la classification la plus proche de la notre dans la méthode et le jeu de données : les deux méthodes utilisent des distances entre récepteurs et construisent un arbre binaire ; les deux jeux de données sont formés de RCPG humains non-olfactifs et sont les plus exhaustifs possibles.

Comme nous l'avons déjà remarqué, les familles de RCPG sont clairement retrouvées et bien séparées les unes des autres. Les valeurs de bootstrapping sont hautes pour les familles Adhésion, Frizzled et Secrétine ; pour ces deux dernières familles, la valeur de

bootstrapping est de 1000 sur 1000, indiquant une séparation complètement nette des autres familles. La valeur de bootstrapping pour la famille Glutamate est faible (238), indiquant une séparation moins nette d'avec la famille Rhodopsine. La seule exception d'un récepteur placé dans le cluster d'une famille à laquelle il n'appartient pas est GPR88.

GPR88	26,7%	NTR1	Brain-gut peptides
	23,3%	SSR5	Opiates
	23,3%	GABR1	Glutamate
	23,3%	GPR39	Brain-gut peptides
	23,3%	MGR7	Glutamate
	23,3%	PAR4	Purine
	23,3%	Q9GZQ4	Brain-gut peptides
	23,3%	MRGRE	MAS

Figure 2-8. Les récepteurs les plus proches de GPR88 (identité sur les séquences de résidus critiques). Pour chacun sont donnés le pourcentage d'identité et le cluster d'appartenance.

Le récepteur GPR88, détecté par GPCRalign [14] comme étant de la famille Rhodopsine, et déterminé comme tel dans l'entrée Swiss-Prot associée [9], est classé ici dans le cluster des récepteurs du glutamate. Le tableau de la figure 2-8 présente les 8 récepteurs les plus proches de GPR88 (en identité des chaînes de 30 résidus critiques). Pourquoi 8 et pas un chiffre rond comme 10? Les 11 récepteurs suivants ont tous le même score d'identité, on ne peut pas n'en sélectionner que 2 parmi eux. Ces 8 récepteurs possèdent

un score d'identité sur les cavités avec GPR88 strictement supérieur à 20%. Parmi eux, 3 récepteurs appartiennent à un cluster de peptides (*Brain-gut peptides*), 2 au cluster des récepteurs au glutamate, et 3 à d'autres clusters de la famille Rhodopsine. L'ambiguïté du classement est compréhensible : l'algorithme de clustering dispose de plusieurs possibilités de placement de ce récepteur dans des clusters. Mais la valeur de bootstrapping du nœud père de GPR88, qui est 735, indique que dans plus de 73% des classements, le récepteur est placé parmi les récepteurs du glutamate.

La composition de chaque cluster ainsi que les empreintes spécifiques en acides aminés sont discutées en détail dans l'article joint en Annexe B à ce document.

2.3.3 Discussion

Nous discutons ici deux points.

Le premier point est celui de la pertinence de notre classification pour les récepteurs dont le site actif orthostérique n'est pas situé dans la cavité transmembranaire, ce qui est le cas des récepteurs des familles autres que Rhodopsine. Bien que nous nous soyons focalisés sur cette cavité, les récepteurs de ces familles sont quand même regroupés et les familles sont bien séparées les unes des autres, et séparées de la famille Rhodopsine. Des modèles de RCPG construits par homologie avec le programme GPCRgen du paquetage GPCRmod [14] présentent bien une cavité transmembranaire. On peut supposer qu'une pression de sélection a été appliquée au domaine transmembanaire pour maintenir la cavité.

Pour apporter un poids à la validité du clustering impliquant les familles dont le site orthostérique est sité en-dehors de la cavité transmembranaire, nous avons construit un

arbre en utilisant la méthodologie décrite dans ce chapitre, mais des résidus différents : le nombre de résidus est le même que pour la classification précédente (30), les résidus sont toujours choisis parmi ceux de la cavité transmembranaire, la proportion de résidus par partie transmembranaire est la même, mais les résidus sont choisis aléatoirement parmi ceux qui ne pointent pas vers l'intérieur de la cavité. Les résultats (non présentés ici) font clairement apparaître un mélange dans les sous-familles de la famille Rhodopsine, et surtout un mélange entre les différentes familles de RCPG. Seul le cluster Frizzled est retrouvé intact dans cette nouvelle classification, car dans l'ensemble ses parties transmembranaires présentent un très haut degré d'identité entre elles (plus de 55% d'identité entre les récepteurs FZD1 à FZ10), et un score d'identité faible entre un récepteur Frizzled et un récépteur non-Frizzled (29% pour SMO, et 17% pour d'autres récepteurs). Ainsi, les résidus choisis n'ont pas un rôle fonctionnel clair.

Le deuxième point de discussion est le choix que nous avons fait de la distance entre séquences. Nous avons utilisé le score d'identité entre séquences. Un premier brouillon de la classification était basée sur une distance différente : le score de similarité. Le score de similarité était calculé à l'aide des matrices BLOSUM [27]. Ce score est toujours supérieur ou égal au score d'identité, puisque deux acides aminés peuvent être différents mais avoir des propriétés assez similaires. Par exemple, une arginine et une lysine sont plus similaires qu'une arginine et une glycine : arginine et lysine ont de grosses chaînes latérales chargées positivement, tandis que la glycine n'a pas de chaîne latérale.

Le clustering calculé avec cette distance est pourtant moins précis, plus diffus, que celui qui utilise l'identité entre séquences. L'explication que nous donnons est en deux arguments : le premier est lié aux matrices BLOSUM, qui ont été calculées pour des

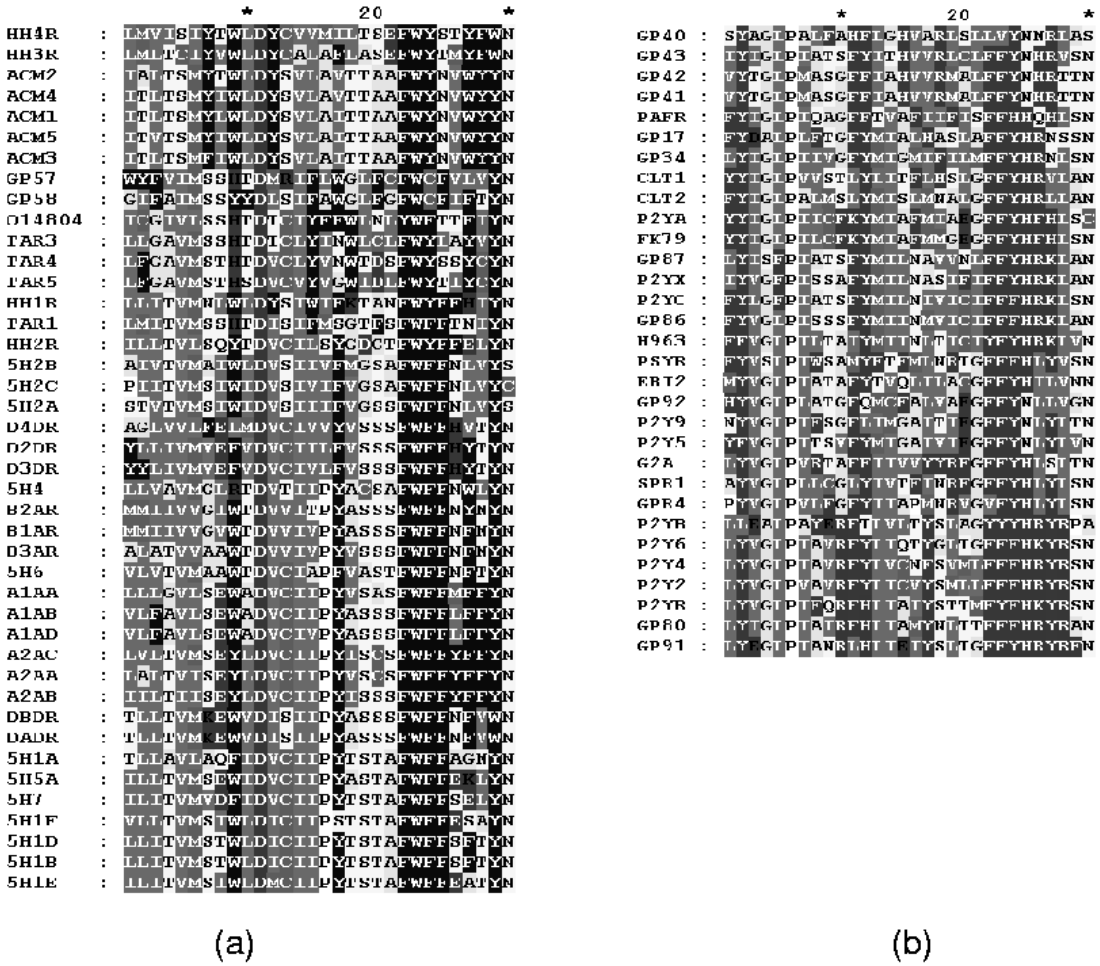
protéines globulaires, et les RCPG ne sont pas globulaires mais membranaires (remarquons que les alignements générés pour ce travail permettraient de construire une matrice propre aux RCPG) ; le second suppose une forte pression de sélection sur la cavité transmembranaire, imposant une plus grande précision dans le choix des acides aminés : une mutation d'un acide aminé vers un autre même similaire aura des conséquences plus importantes dans la cavité transmembranaire qu'ailleurs. Ainsi, utiliser un score d'identité semble plus approprié pour cette étude, et donne un résultat moins diffus.

2.3.4 Applications

La première application de notre classification est l'analyse chimioinformatique de la cavité transmembranaire. Nous pouvons ainsi étudier, par cluster, quels sont les résidus conservés, caractéristiques du cluster, et en déduire quelles sont les propriétés attendues pour les ligands des récepteurs de ce cluster. Cette analyse est détaillée dans l'Annexe B. Par exemple, le cluster des récepteurs des amines (figure 2-9a) laisse apparaître deux sous-sites autour du résidu Asp$^{3.32}$: le premier est plutôt composé de résidus aliphatiques (Val/Ile$^{1.35}$, Leu/Ile/Met/Thr$^{1.39}$, Val/Ile/Leu/Gly$^{1.42}$, Met/Leu/Val/Ile$^{2.58}$) tandis que le second sous-site est clairement aromatique (Phe/Tyr$^{6.44}$, Trp$^{6.48}$, Phe/Tyr$^{6.51}$). En revanche, le cluster des récepteurs purinergiques (figure 2-9b) est caractérisé par plusieurs résidus basiques (Arg, Lys) par cavité dont les charges positives peuvent neutraliser des ligands portant des charges négatives.

Une deuxième application de la classification, ou plutôt de l'alignement qui a servi à construire la classification, est l'établissement de liens entre récepteurs. En effet, certains ligands partageant des sous-structures communes (on parle de « structures privilégiées »)

sont connus pour se lier à plusieurs RCPG [12]. Il est naturel de supposer que les groupements conservés de ces ligands se lient à des parties conservées des récepteurs [12]. On recherche donc une correspondance entre la structure privilégiée et certaines parties conservées des récepteurs, et ceci indépendamment de toute analyse 3D. En cherchant parmi tous les autres RCPG ceux qui possèdent les mêmes propriétés, on peut proposer ces derniers comme des récepteurs potentiels pour les ligands qui partagent la structure privilégiée.



(a) (b)

Figure 2-9. Comparaison entre le cluster des récepteurs d'amines (a) et le cluster des

récepteurs purinergiques (b). Le code couleur reflète les propriétés physico-chimiques des acides aminés.

Par exemple, les biphenyltétrazoles et les acides biphenylcarboxiliques se lient à au moins 6 RCPG : AG22, AG2R, AG2S [28], GHSR [29], LT4R1, LT4R2 [30] (figure 2-10). Mais, grâce à l'analyse des résidus conservés parmi les 30 résidus critiques de notre classification (Phe$^{6.44}$, Trp$^{6.48}$, Phe/Tyr/His$^{6.51}$, Lys$^{5.42}$/Arg$^{6.55}$/Arg$^{7.35}$, His/Gln$^{6.52}$), nous proposons une liste d'autres RCPG qui partagent ces résidus et pourraient également lier ces molécules : APJ, C3AR, C5AR, C5ARL, CML1, FPR1, FPRL1, GPR15, GPR44, GPR1, MTLR, NTR1, Q9GZQ4, G2A, SPR1, GALR1, GALR2. Cette liste contient des récepteurs déjà identifiés par Bondensgaard *et al.* [12] (APJ, NTR1 entre autres) comme cibles potentielles de ces molécules, mais la validation de notre approche est fournie par Frimurer *et al.* [32] qui ont identifié des ligands d'AG2R comme étant également des ligands de GPR44. Il s'agit du candesartan (IC_{50} de 2,1 μM pour GPR44) et de l'indometacine (IC_{50}= 9,0 μM). Nous classons en effet les récepteurs AG2R et GPR44 dans le même cluster.

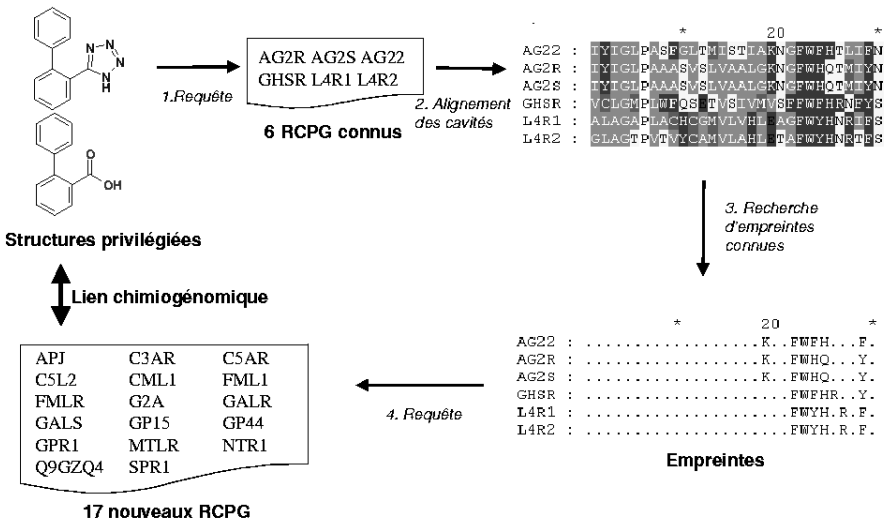


Figure 2-10. Proposition de ligands d'une famille de récepteurs pour des récepteurs d'une autre famille. (1) On recherche les récepteurs de ligands partageant des structures privilégiées. (2) Les cavités de ces récepteurs sont alignées. (3) La recherche d'empreintes communes sert de base à (4) une recherche de récepteurs qui partagent ces empreintes : ils sont proposés pour lier également les ligands de départ.

Enfin, la troisième application de notre classification, immédiate, est la désorphanisation de récepteurs, ou la proposition de ligands potentiels pour des récepteurs orphelins. Pour cela, nous proposons de tester tous les ligands connus des récepteurs d'un cluster sur les récepteurs orphelins du même cluster. Cette approche est très naturelle : à récepteurs similaires, ligands similaires. Mais elle n'est effective que pour les récepteurs orphelins qui appartiennent à des clusters [33]. Le tableau de la figure 2-11 propose des sources de ligands pour plusieurs récepteurs orphelins.

RCPG orphelin	Cluster	Source de ligands
GPR88, Q9NFN8	Glutamate	Ligands allostériques GABA-B
Q8NHZ9, Q8TDU1	Glutamate	Ligands allostériques CaSR
LRG4, LRG5, LRG6	Glycoprotéines	Ligands non-peptidiques LH/FSH
GP119	Lipides	Ligands des récepteurs de cannabinoïde
GPR19, GPR83	Peptides	Ligands des récepteurs de tachykinine
KISSR	Peptides	Ligands des récepteurs de galanine
Q6W5P4, PKR1, PKR2	Vasopeptides	Ligands des récepteurs d'oxytocine/ vasopressine
O14804	Amines	Ligands des récepteurs d'amines biogènes
GPR39	Neuropeptides	Ligands des récepteurs de neuromedin U
O75307, RDC1	Chimiokines	Ligands des récepteurs de chimiokines
GPR7, GPR8	Opiacés	Ligands des récepteurs de somatostatine
GPR15, GPR25, GPR44, GPR1	Chimioattractants	Ligands des récepteurs d'angiotensin II
EBI2, GPR92, P2RY5	Purines	Ligands de récepteurs de LPC/SPC
GP171, GPR87	Purines	Ligands de récepteurs purinergiques
GPR17, GPR34, GP174	Purines	Ligands de récepteur de leukotrienes

Figure 2-11. Sources possibles de ligands pour certains récepteurs orphelins.

2.4 Conclusion

Nous avons dans ce chapitre construit une classification des RCPG basée sur des résidus critiques des cavités transmembranaires des récepteurs. Nous avons extrait les séquences de la base UniProt, puis les avons filtrées pour ne conserver qu'un jeu de données sûr de 369 récepteurs. Ensuite nous avons aligné les parties transmembranaires. Puis nous avons sélectionné 30 résidus critiques tapissant la cavité transmembranaire, et ceci en étudiant la structure de la Rhodopsine bovine, seul RCPG cristallisé à ce jour. Enfin nous avons classifié les récepteurs par un clustering hiérarchique agglomératif des séquences discontinues des 30 résidus critiques.

Le résultat est un arbre formé de 22 clusters et 32 singletons. Les familles des RCPG sont retrouvées identiques à celles données par la classification GRAFS, à une exception près (GPR88). La pertinence de la classification pour les récepteurs dont le site de liaison est situé en dehors de la cavité transmembranaire est discutée. Parmi les applications de la classification et des données qui ont permis sa construction, on peut citer l'étude des propriétés des résidus spécifiques de la cavité pour chaque clusters et la déduction d'informations sur les ligands qu'elles peuvent accueillir. Nous pouvons également proposer de nouveaux ligands supposés en utilisant deux approches : la première utilise des structures privilégiées pour essayer de faire le lien entre des récepteurs grâce à des ligands qui partagent une même sous-structure ; la seconde est immédiate et propose, pour tous les récepteurs orphelins d'un cluster, les ligands connus des autres récepteurs du même cluster.

Notre classification a l'avantage de la simplicité dans sa méthode de construction. De plus, elle est indépendante de tout modèle de récepteurs car on ne considère que leurs

séquences primaires. Enfin, elle est bien appliquée à la pharmacologie car on se focalise sur la cavité transmembranaire, site d'interaction de la plupart des ligands non peptidiques.

Bibliographie

[1] Fredriksson R., Lagerström M.C., Lundin L.-G., Schiöth H.B. (2003) The G-Protein-Coupled Receptors in the Human Genome From Five Main Families Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Mol. Pharmacol.* **63(6)**:1256–1272.

[2] Joost P., Methner A. (2002) Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. *Genome Biology* **3(11)**:1–16.

[3] Karchin R., Karplus K., Haussler D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **18(1)**:147–159.

[4] Qian B., Soyer O.S., Neubig R.R., Goldstein R.A. (2003) Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. *FEBS Lett.* **554**:95–99.

[5] Liu A.H., Califano A. (2003) CASTOR: Clustering Algorithm for Sequence Taxonomical Organization and Relationship. *J. Comput. Biol.* **10(1)**:21–45.

[6] Attwood T.K. (2001) A compendium of specific motifs for diagnosing GPCR subtypes. *Trends Pharmacol. Sci.* **22**:162–165.

[7] Elrod D.W., Chou K.C. (2002) A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Eng.* **15**:713–715.

[8] Lapinsh M., Gutcaits A., Prusis P., Post C., Lundstedt T., Wikberg J.E. (2002) Classification of G-protein coupled receptors by alignment-independent extraction

of principal chemical properities of primary amino acid sequences. *Protein Sci.* **11**:795–805.

[9] Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N., Yeh L.S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**:D154-159.

[10] Bairoch A., Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28(1)**:45–48.

[11] Barker W.C., Garavelli J.S., Haft D.H., Hunt L.T., Marzec C.R., Orcutt B.C., Srinivasarao G.Y., Yeh, L.-S.L., Ledley R.S., Mewes H.-W., Pfeiffer F., Tsugita A. (1998) The PIR-International Protein Sequence Database. *Nucleic Acids Res.* **26(1)**:27–32.

[12] Bondensgaard K.,Ankersen M., Thgersen H., Hansen B.S., Wulff B.S., Bywater R.P. (2004) Recognition of Privileged Structures by G-Protein Coupled Receptors. *J. Med. Chem.* **47(4)**:888–899

[13] Ballesteros J., Shi L., Javitch J. (2001) Structural Mimicry in G Protein-Coupled Receptors: Implications of a High-Resolution Structure of Rhodospin for Structure-Function Analysis of Rhodopsin-Like Receptors. *Mol. Pharmcol.* **60(1)**:1–19.

[14] Bissantz C., Logean A., Rognan D. (2004) High-Throughput Modelling of Human G-Protein Coupled Receptors: Amino Acid Sequence Alignment, Tree-Dimensional Model Building, and Receptor Library Screening. *J. Chem. Inf. Comput. Sci.* **44(3)**:

1162–1176.

[15] Krogh A., Larsson B., von Heijne G., Sonnhammer E.L.L. (2001) Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* **305(3)**:567–580.

[16] Sonnhammer E.L.L., von Heijne G., Krogh A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *in J. Glasgow et al., eds., Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology* 175–182.

[17] Rabiner L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **72(2)**:257–286.

[18] Horn F., Weare J., Beukers M.W., Hörsch S., Bairoch A., Chen W., Edvardsen Ø., Campagne F., Vriend G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **26(1)**:275–279.

[19] Palczewski K., Kumasaka T., Hori T., Behnke C.A., Motoshima H., Fox B.A., Le Trong I., Teller D.C., Okada T., Stenkamp R.E., Yamamoto M., Miyano M. (2000) Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **289**:739–745.

[20] Ballesteros J., Palczewski K. (2001) G protein-coupled receptor drug discovery: Implication from the crystal structure of Rhodopsin. *Curr. Opin. in Drug Discov. & Devel.* **4(5)**:4561–4574.

[21] Jain A.K., Murly M.N., Flynn P.J. (1999) Data Clustering: A Review. *ACM Computing Surveys* **31(3)**:264–323.

[22] Sneath P.H., Sokal R.R. (1962) Numerical taxonomy. *Nature* **193**:855–860.

[23] Mount D.W. (2001) Bioinformatics: Sequence and Genome Analysis. *Cold Spring Harbor Laboratory Press.*

[24] Felsenstein J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**:164–166.

[25] Page R.D.M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**:357–358.

[26] Friedl J.E.F. (1997) Mastering Regular Expressions. *O'Reilly.*

[27] Henikoff S., Henikoff J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**(22):10915–10919.

[28] Ji H., Leung M., Zhang Y., Catt K.J., Sandberg K. (1994) Differential structural requirements for specific binding of nonpeptide and peptide antagonists to the AT1 angiotensin receptor. Identification of amino acid residues that determine binding of the antihypertensive drug losartan. *J. Biol. Chem.* **269**:16533–16536.

[29] Smith R.G., Cheng K., Schoen W.R., Pong S.S., Hickey G., Jacks T., Butler B., *et al.* (1993) A nonpeptidyl growth hormone secretagogue. *Science* **260**:1640–1643.

[30] Reiter L.A., Koch K., Piscopio A.D., Showell H.J., Alpert R., Biggers M.S., Chambers R.J., *et al.* (1998) trans-3-Benzyl-4-hydroxy-7-chromanylbenzoic acid derivatives as antagonists of the leukotriene B4 (LTB4) receptor. *Bioorg. Med. Chem. Lett.* **8**:1781–1786.

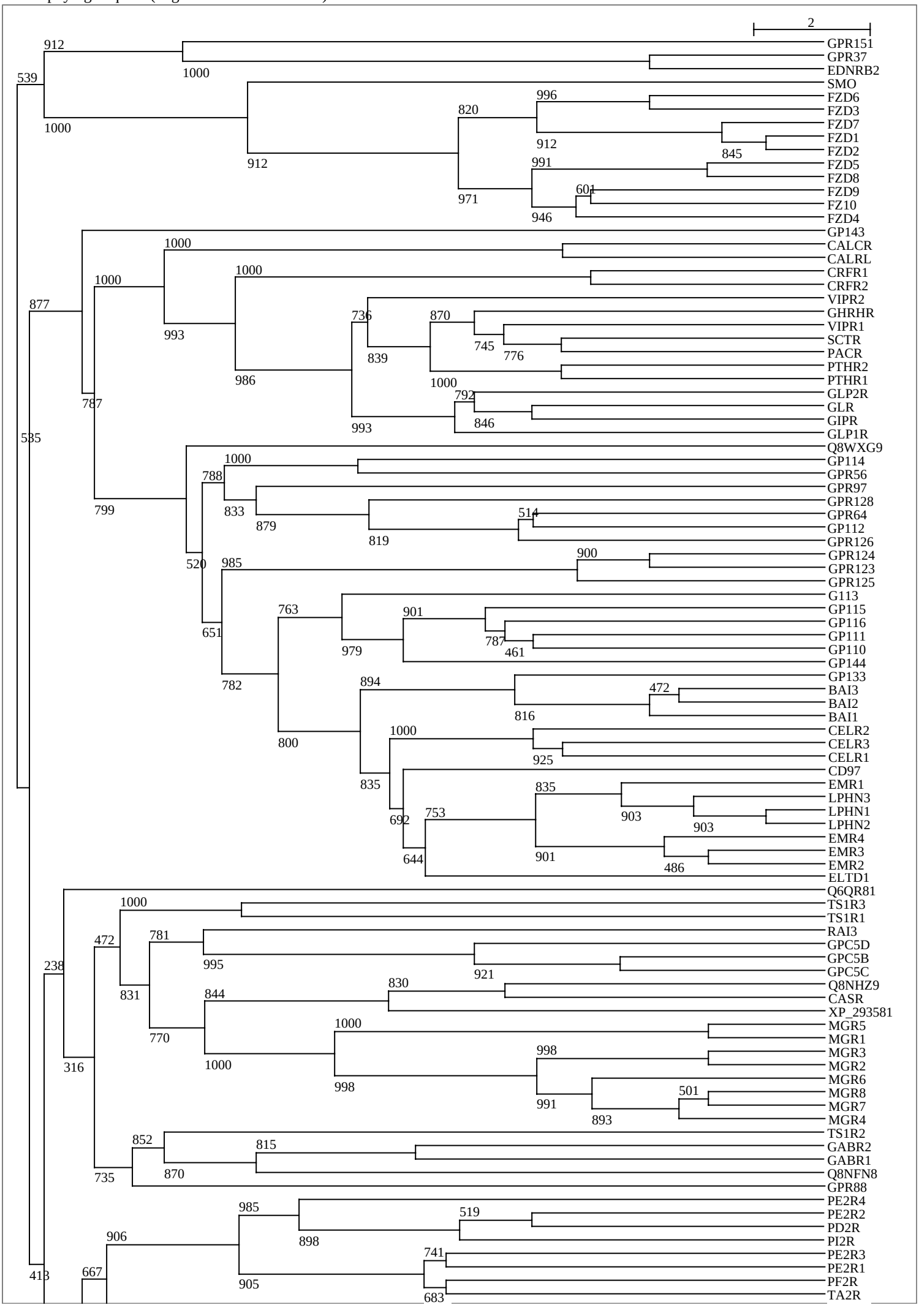
[31] Bywater R.P. Location and nature of the residues important for ligand recognition in G-protein coupled receptors. J. Mol. Recognit. 2005 18:60–72

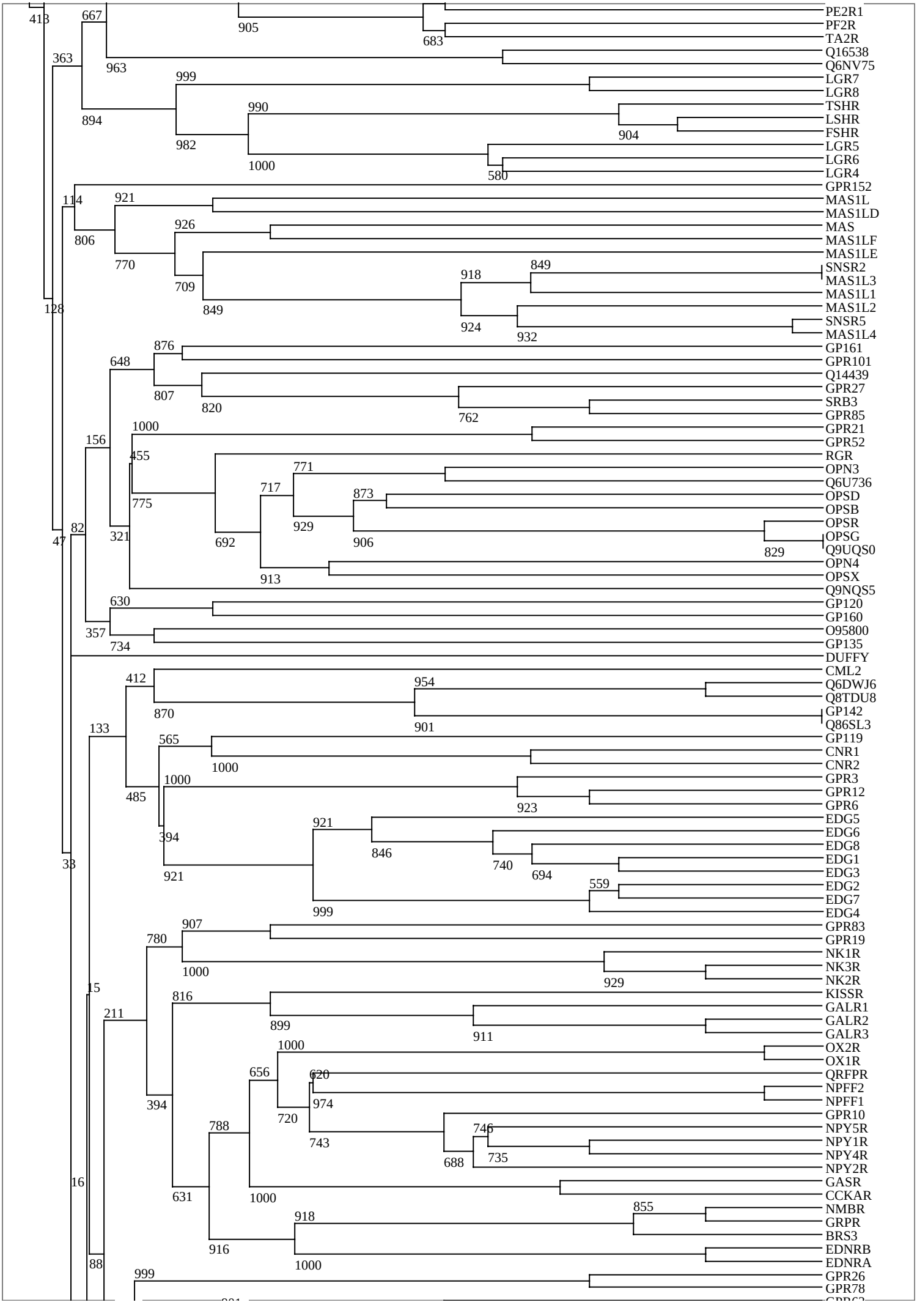
[32] Frimurer T., Ulven T., Elling C.E., Gerlach L.O., Kostenis E., Högberg T. (2005) A physicogenetic method to assign ligand-binding relationships between 7TM receptors. *Bioorg. Med. Chem. Lett.* **15**:3707–3712.

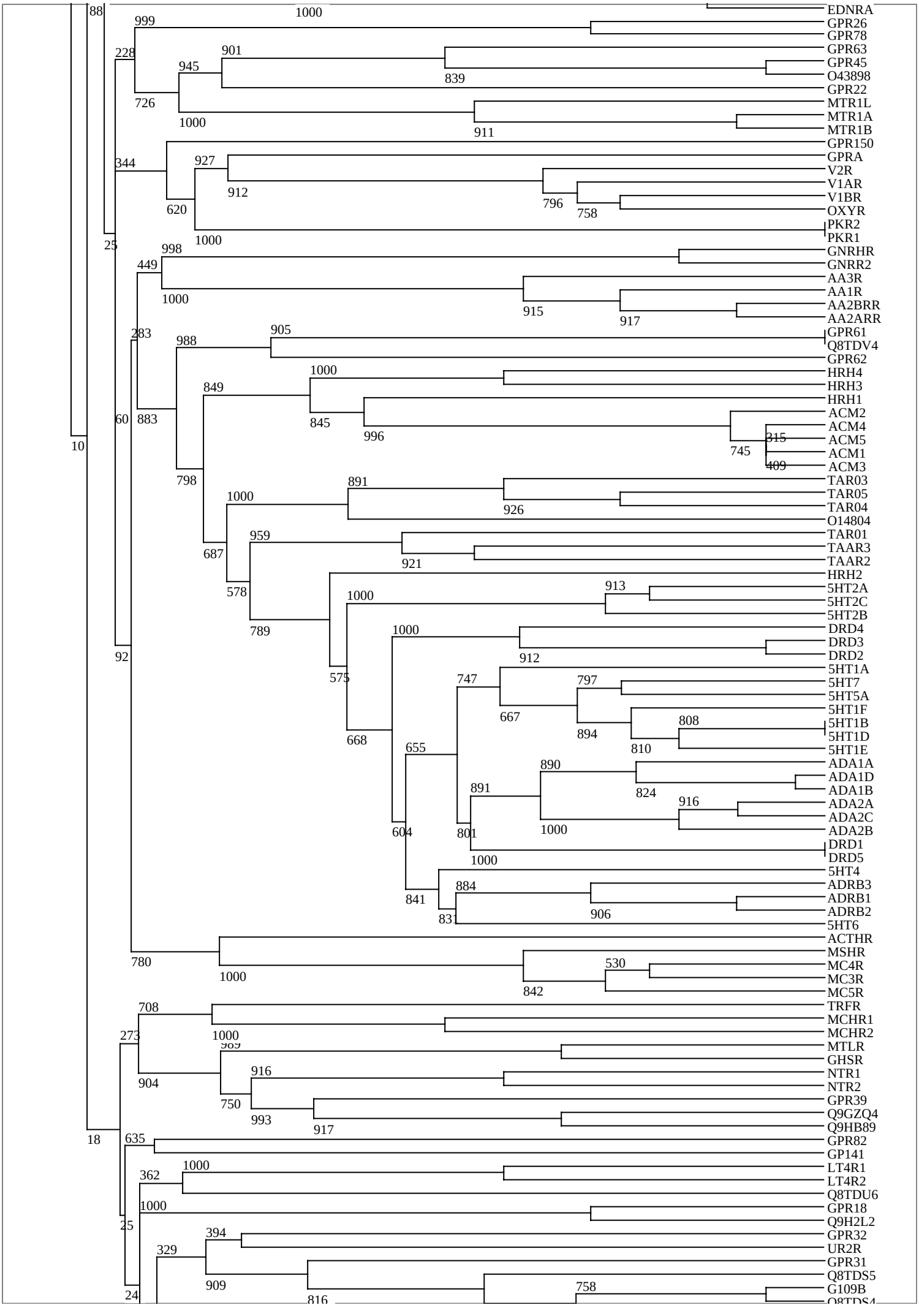
[33] Huang E.S. (2005) Predicting ligands for orphan GPCRs. *Drug Discovery Today*, **10(1)**:69–73.

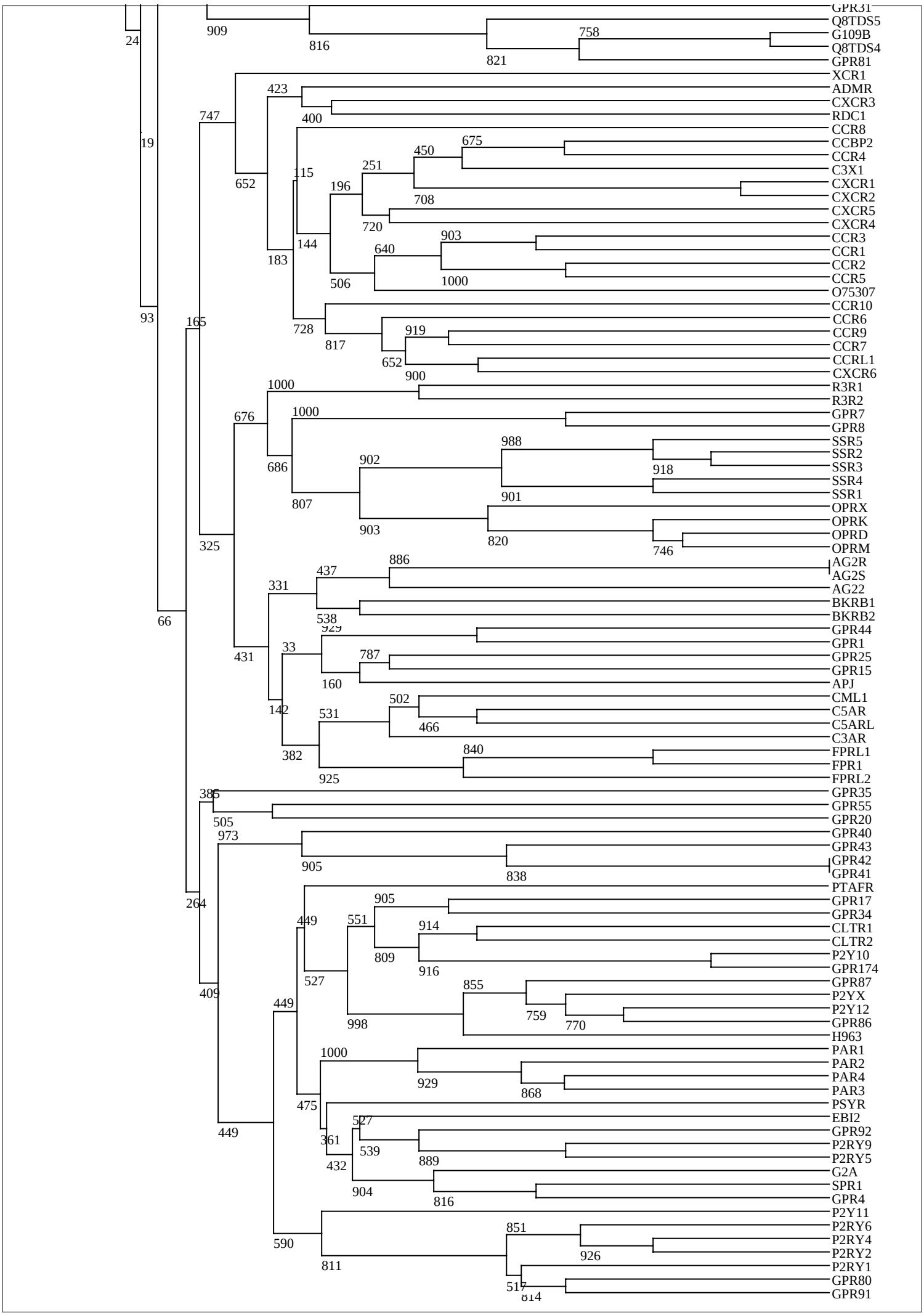
Annexe A

Arbre phylogénique 1 (alignement 1-D de cavités)




PE2R1
PF2R
TA2R
Q16538
Q6NV75
LGR7
LGR8
TSHR
LSHR
FSHR
LGR5
LGR6
LGR4
GPR152
MAS1L
MAS1LD
MAS
MAS1LF
MAS1LE
SNSR2
MAS1L3
MAS1L1
MAS1L2
SNSR5
MAS1L4
GP161
GPR101
Q14439
GPR27
SRB3
GPR85
GPR21
GPR52
RGR
OPN3
Q6U736
OPSD
OPSB
OPSR
OPSG
Q9UQS0
OPN4
OPSX
Q9NQS5
GP120
GP160
O95800
GP135
DUFFY
CML2
Q6DWJ6
Q8TDU8
GP142
Q86SL3
GP119
CNR1
CNR2
GPR3
GPR12
GPR6
EDG5
EDG6
EDG8
EDG1
EDG3
EDG2
EDG7
EDG4
GPR83
GPR19
NK1R
NK3R
NK2R
KISSR
GALR1
GALR2
GALR3
OX2R
OX1R
QRFPR
NPFF2
NPFF1
GPR10
NPY5R
NPY1R
NPY4R
NPY2R
GASR
CCKAR
NMBR
GRPR
BRS3
EDNRB
EDNRA
GPR26
GPR78
413
667
905
683
363
963
999
894
990
982
904
1000
580
114
921
926
806
770
709
849
918
924
932
128
876
648
807
820
762
1000
156
455
771
717
775
873
929
906
692
829
47
82
321
913
630
357
734
412
954
870
901
133
565
1000
485
923
394
921
846
740
694
559
999
33
907
780
1000
929
15
816
211
899
911
1000
656
620
974
720
394
788
746
743
688
735
631
1000
16
855
918
916
1000
88
999


EDNRA
GPR26
GPR78
GPR63
GPR45
O43898
GPR22
MTR1L
MTR1A
MTR1B
GPR150
GPRA
V2R
V1AR
V1BR
OXYR
PKR2
PKR1
GNRHR
GNRR2
AA3R
AA1R
AA2BRR
AA2ARR
GPR61
Q8TDV4
GPR62
HRH4
HRH3
HRH1
ACM2
ACM4
ACM5
ACM1
ACM3
TAR03
TAR05
TAR04
O14804
TAR01
TAAR3
TAAR2
HRH2
5HT2A
5HT2C
5HT2B
DRD4
DRD3
DRD2
5HT1A
5HT7
5HT5A
5HT1F
5HT1B
5HT1D
5HT1E
ADA1A
ADA1D
ADA1B
ADA2A
ADA2C
ADA2B
DRD1
DRD5
5HT4
ADRB3
ADRB1
ADRB2
5HT6
ACTHR
MSHR
MC4R
MC3R
MC5R
TRFR
MCHR1
MCHR2
MTLR
GHSR
NTR1
NTR2
GPR39
Q9GZQ4
Q9HB89
GPR82
GP141
LT4R1
LT4R2
Q8TDU6
GPR18
Q9H2L2
GPR32
UR2R
GPR31
Q8TDS5
G109B
Q8TDS4


GPR31
Q8TDS5
G109B
Q8TDS4
GPR81
XCR1
ADMR
CXCR3
RDC1
CCR8
CCBP2
CCR4
C3X1
CXCR1
CXCR2
CXCR5
CXCR4
CCR3
CCR1
CCR2
CCR5
O75307
CCR10
CCR6
CCR9
CCR7
CCRL1
CXCR6
R3R1
R3R2
GPR7
GPR8
SSR5
SSR2
SSR3
SSR4
SSR1
OPRX
OPRK
OPRD
OPRM
AG2R
AG2S
AG22
BKRB1
BKRB2
GPR44
GPR1
GPR25
GPR15
APJ
CML1
C5AR
C5ARL
C3AR
FPRL1
FPR1
FPRL2
GPR35
GPR55
GPR20
GPR40
GPR43
GPR42
GPR41
PTAFR
GPR17
GPR34
CLTR1
CLTR2
P2Y10
GPR174
GPR87
P2YX
P2Y12
GPR86
H963
PAR1
PAR2
PAR4
PAR3
PSYR
EBI2
GPR92
P2RY9
P2RY5
G2A
SPR1
GPR4
P2Y11
P2RY6
P2RY4
P2RY2
P2RY1
GPR80
GPR91
909
816
758
821
24
423
400
747
19
675
450
251
652
115
196
708
720
144
903
640
183
506
1000
93
165
728
817
919
652
900
1000
676
1000
988
902
686
918
807
901
903
820
746
325
886
437
331
538
66
929
33
787
431
160
502
142
531
466
382
840
925
385
505
973
905
838
264
905
449
551
914
809
916
527
855
409
449
759
770
998
1000
929
868
475
527
361
539
889
432
904
816
449
590
851
926
811
517
814

Annexe B

Chapitre 3

Classification des RCPG basée sur les structures des cavités transmembranaires

Résumé

Après avoir classifié les RCPG d'après des séquences discontinues de résidus critiques de leur cavité transmembranaire, nous allons chercher à étudier cette cavité d'un point de vue structural. En effet, les propriétés physico-chimiques sont bien prises en compte dans la classification du chapitre précédent, mais pas leur distribution dans l'espace, ni la géométrie de la cavité. Inclure cette géométrie apportera donc un supplément d'informations. Comme nous ne disposons que d'une seule structure cristallographique (celle de la rhodopsine bovine [1]), nous allons travailler sur des modèles construits par homologie [2] pour arriver à une classification des RCPG basée sur la similarité structurale des cavités

transmembranaires.

Pour cela, nous avons créé un outil d'alignement structural de cavités (non de protéines complètes), adapté au travail sur des modèles plutôt que sur des structures cristallographiques. Pour comparer entre elles deux cavités, nous plaçons dans chacune d'elles une sphère discrétisée en triangles, et nous y projetons des informations physico-chimiques et géométriques qui décrivent les cavités. Les sphères sont ensuite comparées entre elles pour aboutir à un score de similarité. Si une des sphères est mobile dans sa cavité, on peut lui faire essayer toutes les positions possibles et ne retenir que la configuration donnant le meilleur score. On aura alors trouvé le meilleur alignement entre les deux cavités. Ce score sera également utilisé pour construire une matrice de distances dans le cas d'un alignement multiple, et de là une classification.

Nous construisons, à l'aide de cet outil, une classification basée sur l'alignement structural de modèles de la cavité transmembanaire de 360 RCPG humains non olfactifs. Cette classification est très similaire à celle du chapitre précédent. Nous avons également appliqué l'outil d'alignement sur les sites actifs de la trypsine et la thrombine, deux enzymes de la famille des sérine protéases, ainsi que les sites actifs de l'acétylcholinestérase et de la butyrylcholinestérase.

3.1 Introduction

Dans le chapitre précédent, nous avons classifié les RCPG d'après certains résidus de leur séquence primaire, qui font partie de la cavité transmembranaire. Cette classification présente l'avantage d'être indépendante de tout modèle, car travailler sur des modèles

comporte une part de risque ; or on ne dispose actuellement que d'une seule structure cristallographique, celle de la rhodopsine bovine [1].

Mais la structure qui émerge[1] de la séquence primaire contient des informations différentes de celles du niveau de la séquence, et plus utiles aux pharmacologues. Il est plus naturel de parler d'un complexe *structural* ligand—récepteur que de lier la *séquence* d'un récepteur avec un ligand. C'est pourquoi, malgré l'absence de structures cristallographiques, non allons quand même tenter un travail avec des modèles de RCPG. Ce travail va nous permettre de mettre à jour des similarités structurales qui auraient été manquées dans la classification du chapitre précédent.

Nous aimerions donc utiliser l'information structurale des cavités pour construire une nouvelle classification des RCPG. Cette information structurale comporte des informations géométriques (la forme des cavités) et des informations physico-chimiques (comme par exemple la distribution des charges dans la cavité).

Pour cela nous avons construit un outil d'alignement structural des cavités de protéines qui permet de trouver la meilleure correspondance, quantifiée par un score, entre les propriétés structurales de deux cavités. La classification construite utilisera ce score comme distance entre deux récepteurs.

L'alignement structural de deux protéines est un problème difficile. Plus précisément, un algorithme qui résoud ce problème dans le cas général est NP-complet [5], c'est-à-dire qu'il lui faut un temps non polynomial (en fonction de la taille des données en entrée) pour

1. Je ne veux pas entrer dans une discussion sur le concept d'émergence. Le lecteur intéressé pourra consulter Morin [3] et Hofstadter [4], deux ouvrages d'une richesse exceptionnelle, qui traitent en partie de ce sujet.

trouver une solution, mais un temps polynomial pour vérifier qu'une solution proposée en est bien une.

Les approches actuelles se contentent de la meilleure solution lors d'un parcours heuristique de l'espace des solutions [6]. Citons par exemple la méthode CE [7] qui recherche des similarités structurales locales en partant de paires de fragments similaires alignées ; cette méthode n'est pas utilisable pour aligner des cavités, qui ne présentent généralement pas de telles paires de fragments similaires. Un autre exemple est SiteEngine [8] qui aligne des sites en représentant leurs points chimiquement importants sur des surfaces moyennement discrétisées (environ 10 points par résidu) ; cette méthode est adaptée à la comparaison de structures cristallographiques, à cause de la résolution de la discrétisation, mais pas à des modèles pour lesquels on chercherait plutôt un certain flou. Bien d'autres méthodes d'alignement structural existent ; nous renvoyons à [6] comme point d'entrée dans ce domaine.

Les objectifs que nous nous sommes fixés pour la construction de notre outil d'alignement structural sont les suivants :

- comme nous travaillons sur des modèles, nous avons décidé de ne pas rechercher une trop grande précision dans la description des cavités ;
- les cavités les plus pertinentes sont souvent difficiles à détecter de façon automatique [9], ainsi il doit être possible de sélectionner manuellement les résidus à comparer pour les deux cavités ;
- les scores de comparaison doivent être normalisés, pour pouvoir quantifier la similarité entre plusieurs sites.

Ces critères sont facilement satisfaisables et débouchent sur une méthode d'alignement structural relativement simple.

Nous allons construire un score de similarité entre deux sphères discrétisées chargées de descripteurs géométriques et physico-chimiques (nous appelons ces sphères chargées d'information des « cartes »). L'alignement sera dirigé par ce score, c'est-à-dire que, dans la recherche du meilleur alignement, nous suivrons un chemin de score croissant. Ce principe est dérivé des algorithmes génétiques [10], dirigés par une fonction « fitness ». Nous construirons enfin une classification des RCPG basée sur ce score de similarité.

Nous allons présenter dans la section 2 l'outil d'alignement structural des cavités. Dans la section 3 nous discuterons de ses performances, ainsi que de la classification des RCPG qui en résulte.

3.2 Méthode

Pour construire une classification en partant d'un ensemble de structures, il nous faut construire une matrice de distances puis appliquer la méthode de clustering décrite dans le chapitre précédent. Dans cette section, nous allons décrire la méthode d'obtention d'une distance entre deux structures. Pour construire la matrice complète, il suffira d'itérer le calcul sur toutes les paires de structures.

Le processus de calcul d'un score entre deux structures est schématisé par la figure

3-1. Les différentes étapes sont décrites en détail dans les sous-sections suivantes.

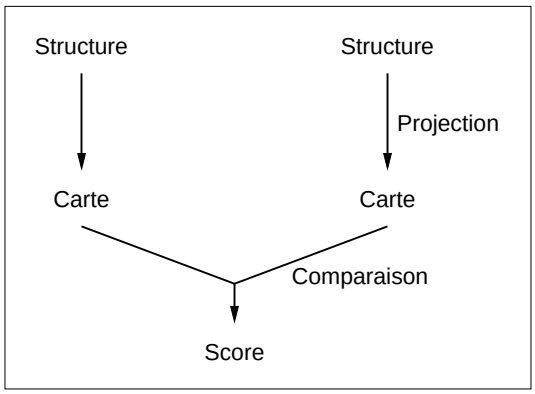


Figure 3-1. Calcul d'un score entre deux structures.

3.2.1 Les modèles utilisés

Les structures des RCPG utilisées pour la classification sont des modèles construits par homologie à l'aide du programme GPCRgen du paquetage GPCRmod [2]. À partir des 369 séquences de RCPG déjà utilisées dans le chapitre précédent, 360 modèles ont été produits (9 modèles n'ont pu être construit à cause d'une bogue dans le programme GPCRgen...)

Très succintement, voici la façon de construire un modèle. À partir d'un ensemble de modèles de RCPG validés par des expériences de mutagénèse dirigée, on reconstruit les hélices transmembranaires indépendamment les unes des autres en se basant sur les hélices les plus similaires des modèles validés. La chaîne principale est conservé et on ajoute les chaînes latérales à partir de deux librairies de rotamères. Enfin, on place la structure complète dans un champ de force pour minimiser son énergie interne et supprimer les conflits stériques apparus lors de l'ajout des chaînes latérales.

Il existe de nombreux modèles de RCPG dans la littérature (par exemple [21]) et sur

le web (voir par exemple <URL:http://www.gpcr.org/7tm/models/index.html>). Nous avons choisis les modèles de GPCRgen parce qu'ils ont été construits au laboratoire.

3.2.2 Les descripteurs calculés

Dans la cavité des RCPG nous plaçons une petite[2] sphère sur laquelle nous allons projeter diverses propriétés quantifiées par des descripteurs. Nous verrons dans les sous-sections suivantes comment la sphère est discrétisée et la façon dont les informations sont projetées dans les différentes parties de la sphère. Dans un objectif de souplesse, nous pouvons sélectionner manuellement les résidus de la cavité pour lesquels des descripteurs seront calculés. Nous avons sélectionné les mêmes résidus que ceux du chapitre précédent.

Dans cette sous-section nous présentons les différents descripteurs calculés pour chacun des résidus. Ils se classent en deux familles : des descripteurs géométriques et des descripteurs physico-chimiques.

Les descripteurs géométriques

Nous calculons 3 descripteurs géométriques par résidu sélectionné.

Le premier est la distance entre le carbone β du résidu et le centre de la sphère. Cette distance est discrétisée par pas de 0,5 Å(valeur arbitraire), d'après notre objectif de léger flou lié au travail sur des modèles. Dans le cas de la glycine, nous créons un pseudo-atome de carbone β à la même position que le carbone α.

Le deuxième descripteur est l'orientation de la chaîne latérale par rapport à la sphère.

2. La taille n'a pas d'importance. Nous parlons de *petite* sphère pour qu'on puisse l'imaginer à l'intérieur d'une cavité. Ce qui importe est le degré de discrétisation.

La chaîne latérale est dirigée soit vers la sphère ($d(C_\beta, O) < d(C_\alpha, O)$), soit elle fuit la sphère ($d(C_\beta, O) \geqslant d(C_\alpha, O)$).

Enfin, le troisième descripteur géométrique calculé est la taille de la chaîne latérale, qu'on divise en petite, moyenne et grande en fonction du nombre d'atomes lourds de cette chaîne [11]. Ce descripteur quantifie le volume d'un acide aminé pointant vers l'intérieur de la cavité. La distance seule ne donne qu'une information sur la position du carbone β par rapport au centre de la sphère. Le tableau de la figure 3-2 donne la répartition des acides aminés en classes de taille de chaîne latérale.

Taille de la chaîne latérale	Résidus dans cette catégorie de taille
petite (0–3 atomes lourds)	Alanine, Cystéine, Glycine, Proline, Serine, Thréonine, Valine
moyenne (4–6 atomes lourds)	Acide aspartique, Acide glutamique, Histidine, Isoleucine, Lysine, Leucine, Méthionine, Asparagine, Glutamine
grosse (7–10 atomes lourds)	Phénylalanine, Arginine, Tyrosine, Tryptophane

Figure 3-2. Résidus classés en fonction de la taille de la chaîne latérale. Les nombres entre parenthèses donnent le nombre d'atomes lourds de la chaîne latérale.

Les descripteurs physico-chimiques

Les descripteurs physico-chimiques sont au nombre de 5. Ce sont, pour chaque résidu sélectionné, le nombre d'accepteurs de liaisons hydrogène, le nombre de donneurs de liaisons hydrogène, le nombre d'interactions aromatiques possibles, le nombre de chaînes aliphatiques, enfin la charge portée par des groupements terminaux du résidu. Dans les travaux de Schmitt *et al.* [12], des descripteurs similaires sont placés au centre de masse des atomes qui portent les différentes propriétés, mais nous projetons ces descripteurs dans le triangle correspondant à l'intersection du centre de la sphère avec le carbone β du résidu, toujours pour la même raison de flou recherché. Le tableau de la figure 3-3 donne la liste des propriétés des résidus qui donnent lieu à une description.

Acide aminé	Aliphatique	Donneur	Accepteur	Aromatique	Charge
Acide aspartique			2		-1
Acide glutamique			2		-1
Alanine	1				
Arginine	1	3			+1
Asparagine		1	1		
Cystéine	1				
Glutamine		1	1		
Glycine					
Histidine		1	1	1	
Isoleucine	1				
Leucine	1				
Lysine	1	1			+1
Méthionine	1				
Phénylalanine				1	
Proline	1				
Sérine		1	1		
Thréonine	1	1	1		
Tryptophane		1		1	
Tyrosine		1	1	1	
Valine	1				

Figure 3-3. Liste des propriétés physico-chimiques des acides aminés.

La représentation des descripteurs

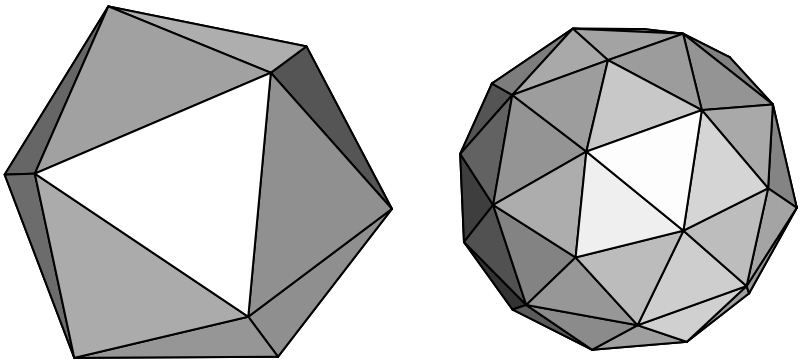
Jusqu'ici nous avons présenté les descripteurs sous une forme abstraite. Nous allons maintenant décrire précisément leur représentation. C'est la même distinction qu'entre un nombre et sa représentation : le nombre 5 peut être représenté par $5_{10} = 101_2 = 11_4$, respectivement en base 10, 2 et 4.

Nous avons choisi de représenter tous les descripteurs à l'aide de nombres entiers, représentation relativement naturelle. La distance du résidu au centre de la sphère aurait posé problème si elle n'avait pas été discrétisée. L'orientation de la chaîne latérale est représentée par deux valeurs (1 si elle est dirigée vers la sphère, 2 si elle la fuit). La taille de la chaîne latérale est représentée par trois valeurs (1 pour les petites chaînes, 2 pour les chaînes de taille moyenne, 3 pour les chaînes de grande taille). Nous verrons dans la section consacrée au calcul d'un score à partir de ces descripteurs que tous ces choix entraînent une simplicité du calcul du score. Les descripteurs physico-chimiques sont naturels : ils comptent des propriétés. Les charges sont représentées directement par des entiers relatifs.

Un autre type de représentation des descripteurs, que nous avons écarté, utilise des empruntes binaires. Les opérations associées sont essentiellement des opérations de comptage de bits utilisées pour le calcul d'un score. Nous justifierons dans la sous-section expliquant la comparaison des cartes pourquoi nous n'avons pas retenu la représentation par empruntes binaires.

3.2.3 La discrétisation de la sphère

Pour une structure donnée, les différents descripteurs sont projetés sur une sphère discrétisée placée au « centre » de la cavité. Ce « centre » est le centre de gravité des carbones α des résidus sélectionnés. Nous verrons plus loin que, pour aligner deux structures, une certaine souplesse dans le placement de la sphère est requise. Nous avons choisi le centre de gravité des carbones α plutôt que le centre de masse des résidus sélectionnés pour privilégier la chaîne principale par rapport aux chaînes latérales, parce que nous travaillons sur des modèles.



(a) (b)

Figure 3-4. (a) Icosaèdre et (b) discrétisation d'une sphère basée sur un icosaèdre.

La sphère est discrétisée en 80 triangles répartis presque uniformément sur sa surface (figure 3-4b). Nous partons d'une discrétisation en icosaèdre. Un icosaèdre (figure 3-4a)

est un solide platonicien de 20 faces triangulaires, 12 sommets et 30 arêtes. Les faces sont des triangles équilatéraux ; tous ont même aire. Puis nous divisons chaque triangle en 4 triangles de la manière suivante (figure 3-5) : nous coupons chaque arête en son milieu, et relions les milieux entre eux pour former de nouvelles arêtes ; les arêtes du triangle initial sont ainsi découpées, et forment chacune 2 nouvelles arêtes. Mais, si tous les sommets des triangles initiaux sont sur une même sphère, les nouveaux sommets (les milieux) ne sont pas sur cette sphère. Nous les y projetons, ce qui fait que les triangles dont les sommets sont tous des milieux sont légèrement plus grands que les autres. Nous quantifierons cette différence de tailles à la fin de cette sous-section.

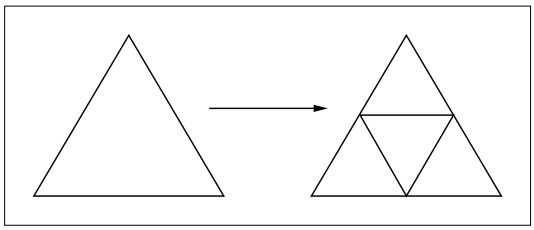


Figure 3-5. Découpage d'un triangle en 4 triangles

Nous pensons que 80 triangles sont suffisants pour décrire une cavité. En effet, pour les RCPG, nous avons vu dans le chapitre précédent que 30 résidus décrivent l'intérieur de la cavité. La prochaine étape dans la discrétisation serait de recommencer à découper chacun des triangles en 4, pour aboutir à 320 triangles. Étant donné que les informations d'un résidu ne sont projetées que dans un seul triangle, elles seraient trop dispersées dans 320 triangles.

Les solides platoniciens, au nombre de 5 (tétraèdre, cube, octaèdre, icosaèdre, dodécaèdre), sont les seuls polyèdres réguliers dont la surface forme une discrétisation uniforme de la sphère.

Nous allons commencer par calculer les coordonnées des 12 sommets d'un icosaèdre. Ces 12 points peuvent être divisés en 3 ensembles de 4 points, qui sont les sommets de 3 rectangles mutuellement orthogonaux de même dimension (figure 3-6). Si on appelle L et l respectivement les longueurs et largeurs des rectangles, et qu'on centre l'icosaèdre en $(0,0,0)$, alors les 12 sommets auront pour coordonnées : $(\pm\frac{l}{2}, \pm\frac{L}{2}, 0)$, $(0, \pm\frac{l}{2}, \pm\frac{L}{2})$, $(\pm\frac{L}{2}, 0, \pm\frac{l}{2})$.

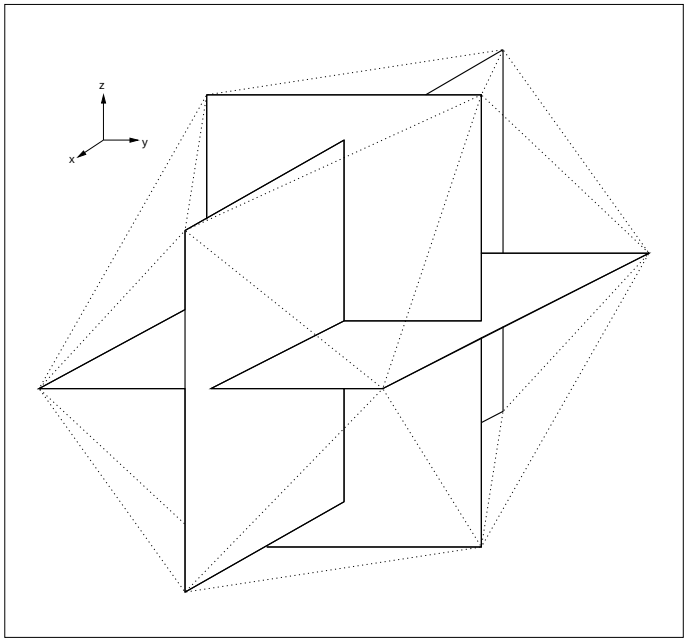


Figure 3-6. Les sommets d'un icosaèdre sont les sommets de 3 rectangles mutuellement orthogonaux

Il suffit manintenant d'exprimer L en fonction de l. Pour cela, prenons une coupe suivant le plan $y0z$ (avec $x = 0$) (figure 3-7) où :

- r est le rayon de la sphère dans laquelle l'icosaèdre est inscrit ;
- a est la longueur d'une arête $(a = l)$;
- h est la hauteur d'une face triangulaire ($h = \frac{\sqrt{3}}{2}a$ puisque le triangle est équilatéral) ;

– enfin, q est la distance entre le centre de l'icosaèdre et le milieu d'une arête ($q = \frac{L}{2}$).

Nous avons aussi les coordonnées des deux points P et Q de la même figure :

– $P = (0, \frac{a}{2}, q) = (0, \frac{l}{2}, \frac{L}{2})$;

– $Q = (0, q, 0) = (0, \frac{L}{2}, 0)$.

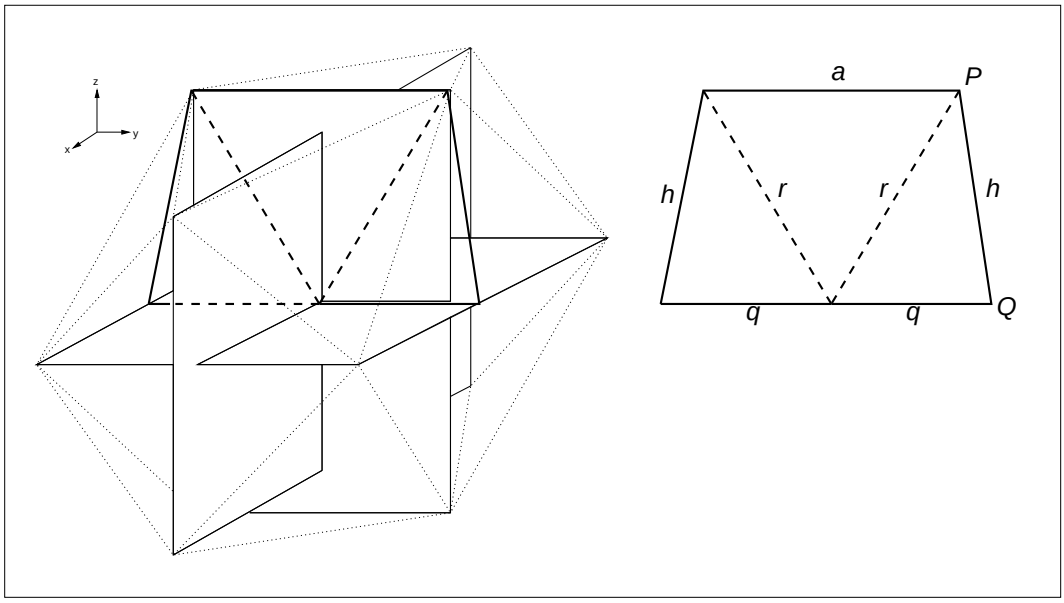


Figure 3-7. Demi-coupe suivant yOz, permettant de calculer la relation entre L et l.

Calculer la relation entre L et l revient à calculer la relation entre a et q. Celle-ci nous est donnée par la distance entre P et Q :

$$d(P, Q) = h$$

$$(\frac{a}{2} - q)^2 + q^2 = h^2 = \frac{3}{4}a^2$$

$$2q^2 - aq = \frac{a^2}{2}$$

$$q^2 - \frac{aq}{2} - \frac{a^2}{4} = 0$$

$$(q - \frac{a}{4})^2 - \frac{a^2}{16} - \frac{a^2}{4} = 0$$

$$(q - \frac{a}{4})^2 = \frac{5}{16}a^2$$
$$q - \frac{a}{4} = \frac{\sqrt{5}}{4}a$$

d'où finalement :

$$q = \frac{\sqrt{5}+1}{4}a = \frac{\varphi}{2}a$$

où φ est le nombre d'or $\frac{\sqrt{5}+1}{2}$. Comme $l = a$ et $\frac{L}{2} = q$, on a finalement $L = \varphi l$. Les coordonnées deviennent alors, si en plus on pose $l = 2$: $(\pm 1, \pm\varphi, 0)$, $(0, \pm 1, \pm\varphi)$, $(\pm\varphi, 0, \pm 1)$. Le rayon de la sphère qui contient tous ces points est $r = \sqrt{1 + \varphi^2}$.

Il nous faut maintenant partager chaque triangle en 4 comme décrit précédemment (figure 3-5). Nous calculons simplement les coordonnées des milieux des arêtes de l'icosaèdre. Ces milieux ne sont plus sur la sphère qui inscrit l'icosaèdre. Nous les y projetons de la façon suivante : si $P'(x, y, z)$ est un point à projeter sur la sphère de rayon r, alors soit $r' = \sqrt{x^2 + y^2 + z^2}$, et ainsi la projection P de P' sur la sphère de rayon r est :

$$P(x\frac{r}{r'}, y\frac{r}{r'}, z\frac{r}{r'})$$

Cette procédure a été automatisée et conduit à 80 triangles dont tous les sommets sont sur une sphère de rayon r.

Ainsi, la discrétisation construite comporte :

- 80 faces (20 faces de l'icosaèdre divisées en 4) ;
- 42 sommets (12 sommets pour l'icosaèdre + 30 milieux projetés, 1 par arête de l'icosaèdre) ;
- 120 arêtes (par arête de l'icosaèdre : 3 arêtes reliant les milieux, 6 demi-arêtes de l'icosaèdre, chacune partagée entre 2 triangles).

Pour terminer cette partie consacrée à la discrétisation de la sphère, nous quantifions son uniformité. Il y a en effet des triangles plus grands que d'autres, comme expliqué précédemment. Ainsi, 20 triangles sont légèrement plus grands que les 60 restants.

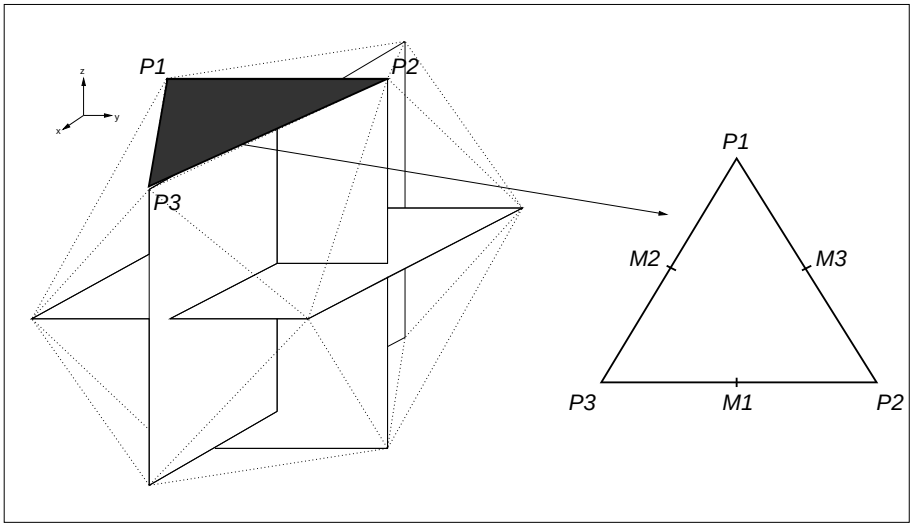


Figure 3-8. Triangle utilisé pour le calcul du rapport entre les aires d'un « grand » et d'un « petit » triangle.

Nous allons calculer explicitement l'aire d'un « grand » triangle et celle d'un « petit ». Pour cela nous choisissons la face $P_1P_2P_3$ où $P_1 = (0, -1, +\varphi)$, $P_2 = (0, +1, +\varphi)$ et $P_3 = (+\varphi, 0, +1)$ (figure 3-8). Les milieux M'_1, M'_2 et M'_3 des segments P_2P_3, P_1P_3 et P_1P_2 respectivement ont pour coordonnées :

- $M'_1 = (\frac{\varphi}{2}, +\frac{1}{2}, \frac{\varphi+1}{2})$,
- $M'_2 = (\frac{\varphi}{2}, -\frac{1}{2}, \frac{\varphi+1}{2})$,
- $M'_3 = (0, 0, \varphi)$.

Les projections sur la sphère (de rayon $r = \sqrt{1+\varphi^2}$) de M'_1, M'_2 et M'_3 sont M_1, M_2 et M_3 respectivement. Ces points ont pour coordonnées :

- $M_1 = r(\frac{1}{2}, +\frac{\varphi-1}{2}, \frac{\varphi}{2})$,

– $M_2 = r(\frac{1}{2}, -\frac{\varphi-1}{2}, \frac{\varphi}{2})$,

– $M_3 = r(0, 0, 1)$.

Nous allons calculer les aires des triangles avec la formule de Horner, qui donne l'aire A d'un triangle en fonction des longueurs (a, b et c) de ses côtés :

$$A = \sqrt{p(p-a)(p-b)(p-c)}$$

où p est le demi-périmètre du triangle ($p = \frac{a+b+c}{2}$). Il nous faut donc calculer les longueurs P_1M_2 et M_1M_2.

– $P_1M_2 = P_1M_3 = \sqrt{4 + 2\varphi - 2r\varphi}$,

– $M_1M_2 = M_1M_3 = M_2M_3 = r(\varphi - 1)$.

La formule explicite de l'aire de la face $M_1M_2M_3$ se calcule simplement par la formule de Horner avec $a = b = c = r(\varphi - 1)$ et donc $p = \frac{3}{2}r(\varphi - 1)$:

$$M_1M_2M_3 = \sqrt{\frac{3}{2}r(\varphi-1)\left[\frac{1}{2}r(\varphi-1)\right]^3} = \frac{\sqrt{3}}{4}r^2(\varphi-1)^2 = \frac{\sqrt{3}}{4}(3-\varphi)$$

Numériquement, on obtient $M_1M_2M_3 \approx 0,598$. Le calcul de la face $P_1M_2M_3$ donne une formule explicite plus compliquée :

$$P_1M_2M_3 = \sqrt{\frac{8+4\varphi-3r\varphi-r}{2}\left[\frac{r(\varphi-1)}{2}\right]^2\frac{8+4\varphi-5r\varphi+r}{2}}$$

Numériquement, on obtient $P_1M_2M_3 \approx 0,533$. Le rapport entre les aires d'une grande face et d'une petite est de $0,891$. Ce rapport nous satisfait puisque nous ne recherchons pas une grande précision.

3.2.4 La projection des descripteurs sur la sphère

À ce stade, nous avons calculé des descripteurs que nous devons projeter sur la sphère discrétisée. Les descripteurs sont calculés pour un certain nombre de résidus sélectionnés. Cette sous-section décrit comment le triangle qui recevra le jeu de descripteurs d'un résidu est sélectionné.

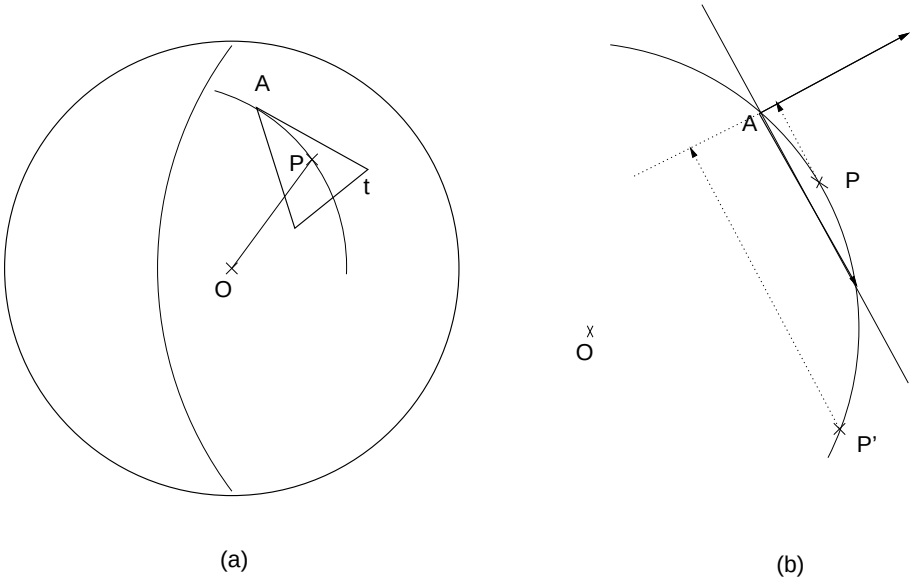


Figure 3-9. Projection d'un jeu de descripteurs dans un triangle. (a) Vue en 3D. (b) Projection en 2D : P « tombe » dans t, mais pas P'.

Soit un point P sur une sphère de rayon r et de centre O ($OP = r$) (figures 3-9a et 3-9b). Nous discrétisons cette sphère comme décrit dans la sous-section précédente. Trouver dans quel triangle « tombe » le point P revient à trouver un triangle t qui contient un point de la demi-droite $[OP)$ ($t \cap [OP) \neq \varnothing$). Notez qu'on ne parle pas d'un triangle curviligne qui épouse la sphère (un triangle sur la surface de la sphère), mais bien d'un triangle plan dans l'espace tridimensionnel. Comme le polyèdre dont la surface est formée par les 80 triangles qui discrétisent la sphère est convexe, nous avons un moyen simple

de déterminer si, oui ou non, P tombe dans t : le plan qui contient le triangle t partage l'espace en 2 demi-espaces, il suffit de savoir si P est dans le même demi-espace que O (dans ce cas la réponse à la question précédente sera « non, P ne tombe pas dans t ») ou si P est dans le demi-espace qui ne contient pas O (et alors la réponse sera « oui, P tombe dans t »).

Rappelons que le produit mixte de 3 vecteurs $\vec{u}$, $\vec{v}$ et $\vec{w}$, noté $[\vec{u}, \vec{v}, \vec{w}]$, est défini par $\vec{u} \cdot (\vec{v} \wedge \vec{w})$ où $\vec{u} \cdot \vec{v}$ est le produit scalaire entre $\vec{u}$ et $\vec{v}$, et $\vec{u} \wedge \vec{v}$ est le produit vectoriel entre $\vec{u}$ et $\vec{v}$. Soit le point P dont on veut déterminer s'il tombe dans le triangle $t = ABC$. Le produit vectoriel $\vec{AB} \wedge \vec{AC}$ est un vecteur orthogonal au triangle t.

Un critère mathématique pour ce test est de déterminer si le produit mixte entre $\vec{AO}$, $\vec{AB}$ et $\vec{AC}$ a le signe opposé de celui du produit mixte entre $\vec{AP}$, $\vec{AB}$ et $\vec{AC}$, ou, de façon équivalente, si le produit de ces deux produits mixtes est négatif. Dans ce cas, P tombe dans t. Si le produit est positif, P tombe dans un autre triangle que t. Pour résumer, P tombe dans $t \Leftrightarrow [\vec{AO}, \vec{AB}, \vec{AC}][\vec{AP}, \vec{AB}, \vec{AC}] \leqslant 0$.

Si nous classons les sommets de chaque triangle dans un certain ordre, nous pouvons simplifier le critère. Ce choix est possible grâce à une structure de donnée adéquate : un triangle est décrit par un tableau (donc ordonné) de ses 3 sommets. L'ordre choisi est tel que, pour tout triangle $t_i = A_iB_iC_i$ (tous les 80 triangles de la discrétisation), $[\vec{A_iO}, \vec{A_iB_i}, \vec{A_iC_i}] < 0$. Ainsi, le critère devient : P tombe dans $t \Leftrightarrow [\vec{AP}, \vec{AB}, \vec{AC}] \geqslant 0$.

Un cas spécial arrive quand le point P tombe sur une arête d'un triangle, ou pire sur un sommet. Dans ces cas, 2 ou 3 triangles sont candidats pour recevoir les descripteurs. Expérimentalement, nous trouvons extrêmement peu de cas où il y a ambiguïté sur une arête (1 pour 10 000), encore moins sur un sommet. Dans ce cas nous abandonnons les

descripteurs, sans conséquences.

Rappelons pour terminer que nous projetons les descripteurs à partir du carbone β, c'est-à-dire que le point P de cette sous-section est la projection du carbone β d'un résidu sur la sphère (continue).

3.2.5 Comparaison entre deux cartes

À ce stade, nous avons deux cartes remplies de descripteurs. L'étape suivante consiste à calculer un score entre ces deux cartes. Pour cela, les deux sphères sont superposées de manière à ce que leurs sommets respectifs coïncident.

Nous présentons d'abord la notation utilisée dans cette partie. Elle va du plus précis (un descripteur) au plus général (le score entre cartes). Nous disposons de 2 sphères discrétisées chacune en 80 triangles qui comportent chacun 8 descripteurs. Les indices commencent tous à 1 et courent jusqu'au nombre d'objets considérés.

- $v_{t,c}^{(d)}$ est la valeur du descripteur d dans le triangle t de la carte c.
- $v_{t,c} = (v_{t,c}^{(1)}, ..., v_{t,c}^{(8)})$ est l'ensemble des descripteurs dans le triangle t de la carte c. Si le triangle t de la carte c ne contient aucune information de résidu, on écrit $v_{t,c} = 0$.
- $s_t^{(d)}$ est le score entre $v_{t,1}^{(d)}$ et $v_{t,2}^{(d)}$ (score pour le descripteur d du triangle t entre les deux cartes).
- s_t est le score pour le triangle t entre les deux cartes.
- S_1 et S_2 sont deux scores entre cartes.

Il est simple de généraliser à plus de descripteurs ou plus de triangles.

Un critère que doit vérifier le score est sa normalisation. Nous devons être capable de comparer entre elles plus de deux cavités, pour pouvoir générer une matrice de distances, dans le but de construire une classification. Nous allons faire en sorte de vérifier ce critère.

Comme pour la notation, la présentation suivante va du plus précis (score entre deux descripteurs) au plus général (score entre deux cartes).

Scores entre descripteurs géométriques

Les descripteurs géométriques se comparent par simple différence. Tous les scores suivants sont calculés pour le triangle t.

Pour la distance, nous définissons $s_t^{(1)}$ par :

$$d = \mid v_{t,1}^{(1)} - v_{t,2}^{(1)} \mid$$

$$s_t^{(1)} = \begin{cases} 1 - \frac{d}{30} & \text{si } d \leqslant 30, \\ 0 & \text{sinon.} \end{cases}$$

Les cavités ne sont généralement pas trop profondes, d'où la valeur 30 (en demi-Å) choisie pour normaliser la différence.

La comparaison de deux orientations de la chaîne latérale se calcule simplement :

$$s_t^{(2)} = \begin{cases} 1 & \text{si } v_{t,1}^{(2)} = v_{t,2}^{(2)}, \\ 0 & \text{sinon.} \end{cases}$$

La comparaison entre deux tailles de chaînes latérales est définie par :

$$s_t^{(3)} = 1 - \frac{\mid v_{t,1}^{(3)} - v_{t,2}^{(3)} \mid}{2}$$

Nous rappelons que $v_{t,i}^{(3)}$ varie de 1 à 3. La différence $\mid v_{t,1}^{(3)} - v_{t,2}^{(3)} \mid$=0 à 2. Le tableau 3-10 donne le score $s_t^{(3)}$ en fonction des tailles des chaînes.

	petite	moyenne	grande
petite	1	0,5	0
moyenne	0,5	1	0,5
grande	0	0,5	1

Figure 3-10. Score $s_t^{(3)}$ en fonction des tailles relatives des chaînes latérales.

Scores entre descripteurs physico-chimiques

Pour ces descripteurs, il suffit de connaître la différence maximale entre deux descripteurs similaires. Appelons cette différence $M^{(i)}$ pour des paires de descripteurs. Alors le score pour cette paire sera :

$$s_t^{(i)} = 1 - \frac{\mid v_{t,1}^{(i)} - v_{t,2}^{(i)} \mid}{M^{(i)}}$$

Ces scores sont tous normalisés si les différentes valeurs de $M^{(i)}$ sont les mêmes quelque soient les comparaisons. Le tableau 3-11 donne ces valeurs pour les descripteurs physico-chimiques.

descripteur	valeur du descripteur	valeur de la différence ($M^{(i)}$)
charge	-1 à +1	$M^{(4)} = 2$
aliphatique	0 à 1	$M^{(5)} = 1$
donneurs	0 à 3	$M^{(6)} = 3$
accepteurs	0 à 2	$M^{(7)} = 2$
aromatique	0 à 1	$M^{(8)} = 1$

Figure 3-11. Intervalle de valeur des descripteurs physico-chimiques, et valeurs maximales de la différence entre ces valeurs.

Scores entre deux cartes

Le score entre deux triangles est défini comme la combinaison linéaire non pondérée des scores de comparaison entre les différents descripteurs pour toutes les propriétés dans ces triangles. S'ils sont tous normalisés, alors leur combinaison linéaire le sera aussi. Le score s_t entre deux triangles t est défini par :

$$s_t = \frac{1}{8} \sum_{d=1}^{8} s_t^{(d)}$$

Nous définissons deux scores entre cartes. Le premier, S_1, fait appel à toutes les paires de triangles (un pour chaque carte) dont l'un au moins des triangles comporte un jeu de descripteurs non nul (les triangles ne sont pas tous chargés d'information, et dans ce cas les descripteurs ont tous une valeur nulle). Il est défini par la combinaison linéaire non pondérée des scores entre ces triangles :

$$S_1 = \frac{1}{N_1} \sum_{t \text{ tel que } v_{t,1} \neq 0 \text{ ou } v_{t,2} \neq 0} s_t$$

N_1 est le nombre de paires de triangles dont l'un au moins comporte une information non nulle. C'est ce score qui dirige l'alignement.

Le deuxième score fait appel à toutes les paires de triangles dont les deux comportent un jeu de descripteurs non nuls. Il est défini de la même façon :

$$S_2 = \frac{1}{N_2} \sum_{t \text{ tel que } v_{t,1} \neq 0 \text{ et } v_{t,2} \neq 0} s_t$$

N_2 est le nombre de paires de triangles dont les deux comportent une information non nulle. Ce deuxième score a été introduit pour pouvoir comparer des cavités de tailles différentes. Il permet alors de ne comparer que les parties des cavités qui sont communes. Utiliser toutes les parties ferait baisser artificiellement le score ; c'est ce qui se passe pour

S_1. Mais il faut nous garder d'utiliser S_2 pour guider l'alignement, puisqu'il ne prend justement pas en compte toute l'information disponible.

Finalement, S_1 est utilisé pour aligner les cavités, tandis que S_2 sera utilisé pour comparer des cavités alignées.

Pourquoi n'avons-nous pas choisi les empruntes binaires ?

Les indices de similarités calculés grâce à des empruntes binaires sont souvent utilisés [13]. Le calcul est rapide car les opérations sont simples et peu nombreuses. Mais deux difficultés (que nous avons annoncées à la section 3.2.2) nous ont dissuadé de les utiliser.

Rappelons par exemple la définition du score de Tanimoto. Soient deux séquences de bits de même longueur L, qu'on note A et B. Soit a le nombre de bits à valeur 1 dans A, b le nombre de bits à valeur 1 dans B, et c le nombre de bits simultanément à 1 dans A *et* dans B (par simultanément, nous considérons le bit de même position dans les deux séquences). Alors une définition du score de Tanimoto [13] T est :

$$T = \frac{c}{a + b - c}$$

Ce score est compris entre les valeurs 0 et 1. Le principal intérêt de ce score est qu'il ne considère pas les parties simultanément à 0 dans les deux séquences. Il ne prend en compte que les parties « intéressantes », ou qui comportent de l'information.

La première difficulté est le codage des nombres entiers. Il faut trouver une représentation qui mime une différence entière avec le coefficient de Tanimoto. Il est évident que la simple représentation binaire d'un nombre entier n'est pas adaptée. Le code de Gray [14] est une alternative intéressante dans le sens où il n'y a qu'un bit de différence entre la représentation d'un nombre et la représentation du nombre voisin. Mais ceci n'est plus vrai

pour des voisins non immédiats. La représentation simple suivante permet de retrouver la différence entre deux nombre dans la différence entre le nombre de bits identiques de leurs représentations :

$0 \rightarrow 0000000$

$1 \rightarrow 0000001$

$2 \rightarrow 0000011$

$3 \rightarrow 0000111$

$4 \rightarrow 0001111$

$5 \rightarrow 0011111$

$6 \rightarrow 0111111$

$7 \rightarrow 1111111$

Mais cette représentation demande un grand nombre de bits. De plus le score de Tanimoto n'est pas normalisé pour elle : si les deux nombres à comparer sont $n1$ et $n2$, alors

$$T = \frac{\min(n1, n2)}{\max(n2, n2)}$$

plutôt que

$$T = \frac{\max(n1, n2) - \min(n1, n2)}{N}$$

La seconde difficulté provient de la différence de taille des représentation des différents descripteurs. La distance du carbone β au centre de la sphère, par exemple, demanderait 31 bits. En revanche, l'orientation de la chaîne latérale ne demande que deux bits pour être représentée. Pour donner un poids identique aux deux, il faudrait étendre la plus petite des représentation, en répliquant les bits, jusqu'à ce qu'elle ait la même taille que la grande. On peut ainsi jouer sur les poids des descripteurs, mais au prix d'un allongement

significatif des empruntes.

3.2.6 Alignement de deux structures

Lorsque l'on dispose d'une fonction de calcul d'un score entre deux cartes, il est possible d'utiliser ce score pour déterminer un alignement entre les deux structures représentées par ces cartes. Pour cela, nous déplaçons une des deux sphères dans sa cavité (l'autre sphère restant fixe) de manière à parcourir tout l'espace de la cavité, et à chaque étape nous complétons la sphère avec des descripteurs recalculés pour la nouvelle position. Suivant la position de la sphère, les descripteurs n'auront pas la même valeur, et ils seront disposés différemment dans les triangles.

Nous faisons donc parcourir à une des sphères l'espace des solution, d'abord de façon grossière. Les trois meilleurs scores sont mémorisés, puis la même sphère parcours alors plus finement l'espace des solutions autour de ces trois positions, et mémorise le meilleur score.

La démarche est similaire à celle des algorithmes génétiques, qui utilisent une fonction « fitness » pour caractériser une position dans l'espace des solutions. La recherche de la solution est dirigée par cette fonction : on chemine de solution en solution de façon à augmenter la valeur de la fonction « fitness ». Les algorithmes génétiques sont des algorithmes heuristiques qui ne parcourent pas complètement l'espace des solutions, bien trop vaste.

La précision du parcours de l'espace par la sphère est paramétrable. D'abord placée au centre de gravité de la cavité, elle effectue typiquement une vingtaine de rotations dans chaque direction de l'espace, et quelques fractions d'angström en translation autour

du centre de gravité.

Le meilleur alignement est défini comme celui ayant le meilleur score (le score maximal). Ainsi, le résultat de notre procédure d'alignement donnera d'une part un score d'alignement quantifiant la similarité des sites actifs comparés, et d'autre part un jeu de paramètres géométriques permettant de disposer dans l'espace une des cavités par rapport à l'autre, en d'autres termes, l'alignement des deux cavités. Ce jeu contient 6 paramètres : 3 angles et un vecteur. Ils sont donnés par la position et l'orientation d'une des sphères par rapport à l'autre, dans la configuration du meilleur score.

Le temps de calcul de cette procédure ne dépend que de la précision de la discrétisation de l'espace des solutions. Mais cette dépendance est forte : AR^3T^3 où A est le temps d'un calcul de score, R est le nombre de rotations dans une direction de l'espace et T est le nombre de translations dans une direction de l'espace. Le temps de calcul d'un alignement est typiquement de 13 minutes sur une machine de processeur Intel (de la famille Pentium) 3 GHz, pour laquelle le programme a été optimisé lors de la compilation (utilisation du compilateur du constructeur du processeur), avec 20 rotations par degré de liberté et 5 translations par degré de liberté axial autour du centre de gravité.

Ce programme peut se distribuer facilement sur une machine parallèle ou autre : il suffit de partager l'espace de recherche en sous-espaces, et de donner à chaque processeur la recherche du meilleur score sur ce sous-espace.

3.2.7 Alignements de plusieurs structures entre elles

La généralisation à l'alignement de plusieurs structures entre elles est immédiate. Nous pouvons en effet construire une matrice de scores en itérant l'alignement de deux

structures de façon à calculer un score pour chaque paire de structures.

Le temps de calcul est borné par TN^2 si T est la borne supérieure du temps de calcul d'un alignement, et N est le nombre de protéines à aligner deux à deux.

Ce type d'alignement se porte naturellement sur une machine vectorielle, puisque les calculs à faire sont identiques pour tous les alignements, et ont tous le même nombre d'étapes. En fait il est possible de distribuer de façon très souple ces calculs sur de nombreuses architectures parallèles : de la machine vectorielle à une machine parallèle dont le système est capable de distribuer les alignements en fonction des besoin, en passant par une architecture faiblement parallèle (un cluster de machines, peut-être hétérogènes, formant un réseau) dont le système distribue également les calculs sur les différentes machines.

3.3 Résultats et discussion

Nous allons dans cette section présenter la classification des RCPG obtenue grâce au score entre deux cavités transmembranaires, après alignement structural des cavités. Nous discuterons également des limitations de la méthode d'alignement structural. Enfin nous présenterons deux autres alignements réalisés à l'aide de cet outil.

3.3.1 La classification des RCPG obtenue

Les 360 modèles de RCPG ont donné lieu à une classification du point de vue de la structure de la cavité transmembranaire. Comme ces modèles ont été construits par homologie, ils étaient tous préalignés sur un très petit nombre de structures (les 8 modèles

validés de [2]) elles-mêmes préalignées sur la structure cristallographique de la rhodopsine bovine. Malheureusement, cet alignement n'est pas suffisamment précis pour construire une classification d'après la fonction de score définie précédemment. Nous avons donc raffiné ces alignements grâce à notre outil, construit à cet effet, en utilisant des paramètres pour ne parcourir l'espace qu'autour des solutions préalignées, soit sur un angle de 0,1 radian pour chaque degré de liberté angulaire, en 8 fois, et une distance de 0,5 Å par degré de liberté en translation, en 3 fois. Avec ces paramètres, la matrice complète a été calculée en 5 jours sur la machine 3 GHz déjà présentée. Nous avons vérifié au préalable que ces paramètres sont suffisants : plus grands, ils n'améliorent pas les scores car les résultats des alignements (angle et vecteur de déplacement d'une sphère par rapport à l'autre) sont compris dans les limites des paramètres.

Nous allons maintenant décrire la nouvelle classification obtenue, la comparer avec celle des deux chapitres précédents, et discuter des similarités et des différences. L'arbre est donnée en Annexe C à la fin de ce chapitre. Nous avons choisi la classification du chapitre précédent comme base de comparaison puisqu'elle est construite grâce aux mêmes résidus et qu'elle utilise le même protocole de clustering, et celle du premier chapitre puisque c'est une classification par ligands de tous les RCPG non orphelins.

Certains clusters sont retrouvés à l'identique dans les deux classifications. Il s'agit des clusters suivant (nous utilisons la même notation en expressions régulières que celle de la figure 2-7) :

– les récepteurs des acides (G109B, GPR(31,81), Q8TDS[45]),

– les récepteurs de la famille frizzled (FZD[1-9], FZ10, SMO),

– les récepteurs de la sous-famille MAS (MAS, MAS1L, MRGR[DEF], MRGX[1-4], SNSR[25]),

– les récepteurs des glycoprotéines ([FLT]SHR, LGR[4-8]),

– les récepteurs des protanoïdes (P[DFI]2R, PE2R[1-4], TA2R),

Ce sont tous de petits clusters très homogènes.

Dans cette classification apparaissent cependant plusieurs occurences d'un artefact de la méthode de clustering, soient des résultats qui peuvent induire des erreurs d'interprétation.

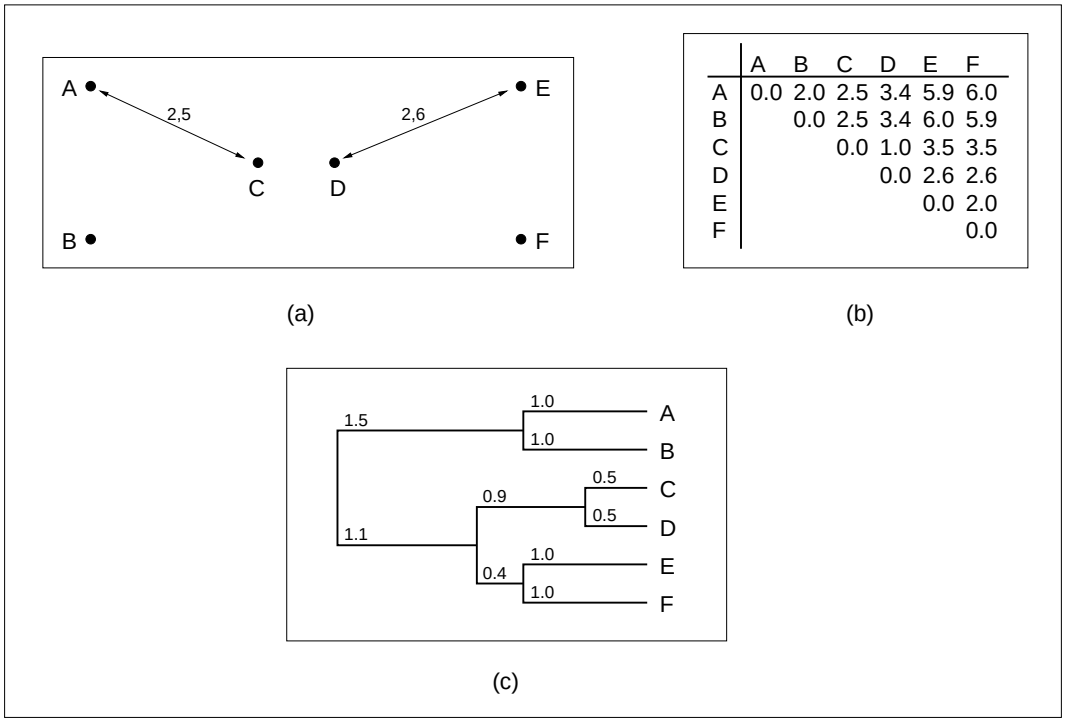


Figure 3-12. Illustration d'un artefact de clustering. (a) Distribution de points dans l'espace à regrouper en clusters. (b) Matrice de distances entre ces points. (c) Arbre obtenu par la méthode UPGMA.

Prenons l'exemple suivant (figure 3-12) : le point D est proche de C, puis de E et F dans l'ordre de distances croissantes ; mais il sera attiré vers A et B par le point C ; dans l'arbre, résultat du clustering hiérarchique agglomératif construit par la méthode UPGMA (figure 3-12c), les distances AC et AD sont identiques. Ainsi, l'arbre ne reflète pas la distribution des points dans l'espace.

Quelques entrée dans la classification sont dans le cas ci-dessus, « tirées » dans un cluster par une entrée proche. Nous avons systématiquement vérifié les cas ambigus en étudiant la matrice de distances qui a conduit à l'arbre. Ces cas sont signalés dans la suite par les termes d'« artefact de clustering ».

Nous allons maintenant décrire et analyser les autres clusters produits, par rapports à ceux de la classification du chapitre précédent. Nous commençons par les clusters décrivant les familles Adhésion, Glutamate et Secrétine (Frizzled n'a pas changé). Puis nous étudierons les clusters des sous-familles de la famille Rhodopsine. Les singletons de la classification du chapitre 2 ne sont toujours pas classés ici, sauf DUFFY (voir le cluster des Chimiokines).

La famille Adhésion reste presque identique (avec BAI[1-3], CD97, CELR[1-3], EMR[1-4], GP11[0-5], GP12[3-68], GP133, GP144, GPR56, GPR97, LPHN[123]), mais GP116, GP143 et Q8WXG9 (le très long récepteur composé de 6307 résidus) deviennent singletons. Pour ce dernier récepteur, c'est un artefact du clustering. Les deux récepteurs ELTD1 et GPR64 font partie des 9 récepteurs non modélisés.

La famille Glutamate (valeur de bootstrap : 108) (avec CASR, GPC5[ABC], MGR[1-8], Q8NHZ9, RAI3) perd les récepteurs du goût (TSR[1-3]) qui deviennent singletons, ainsi que Q8TDU1 et Q6QR81. Mais GPR88 est toujours présent. Nous remarquons

globalement que cette classification est plus sensible que la précédente, au sens où de nombreux clusters perdent des entrées qui deviennent soit des singletons soit de très petits clusters de 2 ou 3 entrées, et ne gagnent pas d'anciens singletons. Le fait que GPR88 soit toujours présent dans ce cluster nous conforte dans l'idée que ce n'est pas par hasard. La proposition de ligands potentiels pour les récepteurs orphelins que nous avons faite dans le chapitre précédent (figure 2-9) est toujours recommandée.

La famille Secrétine (valeur de bootstrap : 795) (avec CRFR[12], GHRHR, GIPR, GLP[12]R, GLR, PACR, PTHR[12], SCTR, VIPR[12]) a perdu CAL(CR,RL) devenus singletons. Dans la classification précédente, les deux familles Adhésion et Secrétine étaient proches dans le sens de leur distance inter-clusters, mais ne le sont plus maintenant.

Les clusters représentant les sous-familles de la famille Rhodopsine sont présentées ci-dessous.

Le cluster des récepteurs de l'Adénosine (AA(1,2[AB],3)R) perd les récepteurs GNR(HR,R2). Ceci est en accord avec la classification basée sur les ligands donnée dans le premier chapitre, où GNH(HR,H2) ont pour ligands l'hormone de libération de la gonadotropine. Mais ces deux récepteurs ne sont pas placés avec les autres récepteurs d'hormones.

Le cluster des récepteurs d'Amines (5HT(1[ABDEF],2[ABC],[467]R,5A), ACM[1-5], ADA(1[ABD],2[ABC]), ADRB[1-3], DRD[1-5], GPR6[12], HRH[2-4], O14804, TAR0[13-5]) perd, par artefact de clustering, les récepteurs HRH1 et gagne, toujours par artefact, GPR44.

Le cluster des récepteurs de Chimioattractants (AG22, BKRB2, C[35]AR, C5ARL,

CML1, FPR(1,L[12]), GPR1, GPR[12]5) perd par artefact de clustering GPR44. Elle perd également le récepteur de l'apéline APJ, classé dans une sous-famille à part dans la Swiss-Prot (chapitre 1). Les récepteurs BKRB1, AG2[RS] sont perdus au profit du cluster des Chimiokines, globalement très proche.

Le cluster des récepteurs de Chimiokines (C3X1, CCBP2, CCR[1-689], CCRL1, CXCR[13-6], XCR1) perd par artefact de clustering (et c'est facheux) les récepteurs CCR7, CCR10, CXCR2 et O75307. Elle perd également ADMR1, le récepteur de l'adrénomédulline, classé seul dans le chapitre 1. Par contre, elle gagne le récepteur DUFFY, et ceci est très étonnant et satisfaisant. Étonnant car notre précédente classification (chapitre 2) n'a pu le classer alors qu'il est visiblement proche des récepteurs des chimiokines dans la matrice (score de 89,5% avec CCR5, 89,2% avec CCBP2, 88,8% avec CCRL1, 87,9% avec CXCR4, 87,5% avec CXCR5, 87,3% avec CCR6, 87,1% avec CCR8, 87,0% avec CCR4, et ensuite vient seulement un récepteur d'une autre sous-famille !) Satisfisant car la classification par ligand exposée dans le chapitre 1 le classe également parmi les récepteurs des chimiokines.

Le cluster des récepteurs de Lipides (EDG[1-8], GP119) perd les deux récepteurs des cannabioïdes, classés à part dans le chapitre 1. Plus fâcheux, les trois récepteurs GPR3, GPR6 et GPR12 forment un cluster distant, alors qu'il a été déterminé qu'ils partageaient des ligands avec les récepteurs EDGx [15].

Le cluster des récepteurs de Mélanocortine (MC[3-5]R, MSHR) perd le récepteur ACTRH, qui fait partie des récepteurs non modélisés.

Le cluster des récepteurs de Mélatonine (MTRA[ABL]) perd GPR(45,63) et O43898, non modélisés. Elle perd également GPR22 par artefact de clustering.

Le cluster des récepteurs de Neuro-peptides (*Brain-gut peptides*) (GHSR, GPR39, MTLR, NTR[12], Q9GZQ4, Q9HB89) perd MCHR[12] et TRFR, trois récepteurs d'hormone. Les récepteurs MCHR[12] forment une classe à part dans la classification de la Swiss-Prot (chapitre 1).

Le cluster des récepteurs des Opiacés (GPR[78], OPR[DKMX], SSR[1-5]) perd par artefact de cluster les récepteurs R3R[12].

Le cluster des Opsines (OPN[34], OPS[BGRX], Q6U736, Q9UQS0, RGR) gagne HRH1, « attiré » car proche de RGR, mais HRH1 est bien sûr plus proche des amines dans l'ensemble. GPR22 est également gagné, globalement plus proche des opsines dans la matrice de distances, mais en l'absence d'information sur les ligands, nous ne pouvons rien en déduire sinon qu'il serait intéressant de tester des ligands des opsines sur ce récepteur.

Le cluster des récepteurs de Peptides est séparé en trois dans cette nouvelle classification : une partie neuro-peptides (GPR10, NPFF[12], NPY[1245]R, OX[12]R, QRFPR), une partie endothéline/bombésine (EDNR[AB], GALR[1-3], NMBR, GRPR, BRS3) et une partie cholécystokinine/gastrine/kisspeptine (CCKAR, KISSR, GASR). GPR19 n'a pas été modélisé, et les NK[123]R forment un petit cluster séparé.

Le cluster des récepteurs de Purines est étonnamment bien conservé vu sa taille. Il est partagé en deux clusters dans cette classification : la première regroupe les récepteurs de l'adrénomédulline (ADMR), de l'apéline (APJ), de leukotriènes (CLT[12]), de protéinases (PAR[1-4]) et des récepteurs purinergiques (P2RY[1246], P2Y10) avec trois récepteurs de lipides (G2A, PSYR et SPR1) et des récepteurs orphelins (RDC1, GPR4, GPR(35,55,92, 20,91,34,4[123]), GP174) Le second comprend des récepteurs purinergiques (P2RY5, P2Y-1[24]), PTAFR et des récepteurs orphelins (GPR40, GPR8[067], GP171).

Le cluster des récepteurs super-conservés exprimés dans le cerveau (*SREB* dans le chapitre 2) est complètement séparé en singletons.

Enfin, le cluster des récepteurs de Vasopeptides (avec OXYR, Q6W5P4, V(1[AB],2)R) a perdu les récepteurs de prokinéticine PKR[12], formant une classe à part dans la classification de la Swiss-Prot (chapitre 1).

Pour conclure cette présentation de la classification des RCPG basée sur la similarité de structure de la cavité transmembranaire, disons qu'elle est moins « nette » que celle du chapitre précédent basée sur des identités de séquences, dans le sens où nous avons vu apparaître plusieurs artefacts liés à la méthode de clustering. L'explication provient de la différence de distribution des scores.

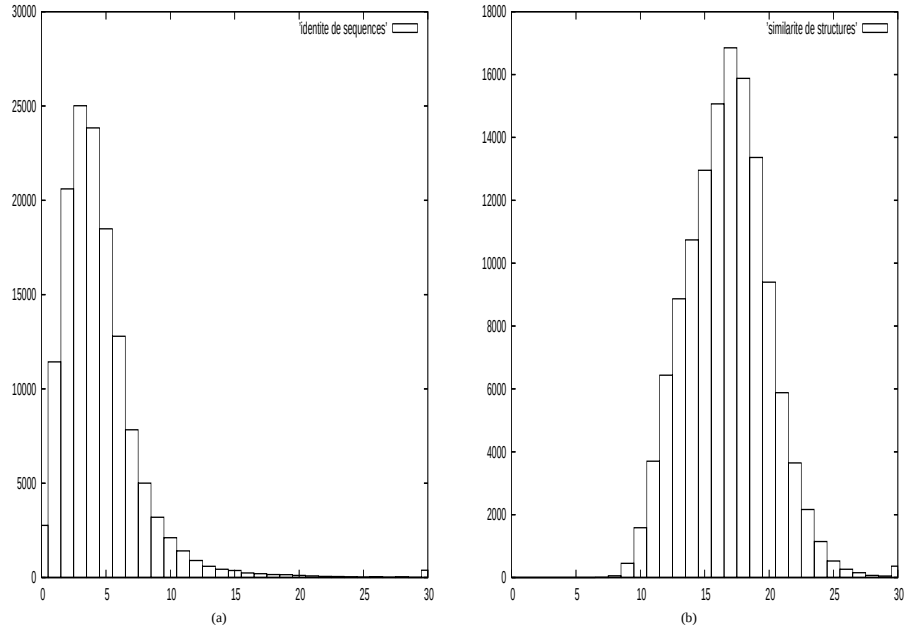


Figure 3-13. Distribution des scores (de 0 à 30) pour (a) la classification du précédent chapitre et (b) du présent chapitre.

La figure 3-13a présente la distribution des scores pour la classification du chapitre

précédent, tandis que la figure 3-13b présente celle du présent chapitre. On explique la moindre précision de la présente classification par un trop grand nombre de « bons » scores, qui en plus provoquent les artefacts de clustering. Le cas de la figure 3-13a est plus adapté à un clustering : un faible nombre de très bons scores qui n'impliquent pas d'ambiguïtés, et un grand nombre de faibles scores (entre récepteurs distants). bien que très homogène avec celle du chapitre précédent, n'apporte pourtant pas d'informations supplémentaires. Les clusters sont plus petits, comme dans la classification par ligand (chapitre 1), et seul le récepteur DUFFY, singleton dans le chapitre 2, semble avoir gagné une bonne classification ici.

3.3.2 Discussion sur l'outil d'alignement structural

La classification basée sur les structures de cavités des RCPG n'apporte pas d'information nouvelle, et présente plutôt quelques défauts dus à la fonction de score de l'outil d'alignement structural. Ce score est en effet composé en partie de descripteurs physico-chimiques et en partie de descripteurs géométriques. Ces derniers, pour les RCPG, sont assez similaires entre eux car les cavités se ressemblent beaucoup [2]. Ainsi l'information géométrique, plutôt que d'aider à discriminer les protéines, aura tendance à « diluer » l'information physico-chimique, qui elle est bien différente pour les récepteurs. Nous ne voulons toutefois pas éliminer l'information géométrique, qui pourra devenir nécessaire lors de l'alignement de cavités de protéines de différentes familles ou super-familles.

En revanche, la fonction d'alignement elle-même produit un alignement dont nous verrons deux résultats dans la prochaine sous-section.

L'outil d'alignement structural comporte quelques limitations inhérentes à la méthode.

La première limitation est celle du temps de calcul ; si un alignement de deux cavité dure environ 10 minutes, construire une matrice de distances complète (même symétrique) prendra par exemple 7h30 pour 10 cavités, 3,5 jours pour 100 cavités, 1 année pour 1000 cavités... Le deuxième limitation concerne le score : il ne permet malheureusement pas de différentier un bon alignement d'un mauvais. Bien sûr, les cas extrêmes sont facilement distingables : un excellent score est signe d'un alignement réussi, et un mauvais score est signe d'un échec de l'alignement. Mais dans le cas d'un score moyen, il n'y a aucun critère formel pour décider de la qualité de l'alignement qu'il faut étudier manuellement. Enfin, une troisième limitation demande une forme de cavité marquée ; le site ne doit pas se situer à la surface d'une protéine, ni avoir une forme complexe. De plus, les cavités devraient être de forme assez similaires. Si toutes ces limitations sont présentes à l'esprit lors de l'utilisation de l'outil, l'alignement produit et son score devraient être pertinents.

3.3.3 Applications

Il est possible d'appliquer l'outil d'alignement structural à d'autres protéines que les RCPG. Il suffit pour cela qu'elles présentent des cavités marquées. Nous avons aligné deux jeux de structures cristallographiques, pour montrer que notre outil, bien que conçu pour des modèles, est capable d'aligner également ce type de structures. Elles sont tirées de la PDB (*Protein Data Bank*) [16], une banque de données de structures cristallographiques de protéines. Les images des alignements ont été produites par le logiciel Sybyl [17]. L'homologie de forme n'est pas aussi importante que pour les RCPG. Les deux sites sont enfouis dans des cavités qui ont été extraites de la base sc-PDB [18], une base de données de sites de liaison de protéines d'intérêt thérapeutique.

Le premier exemple que nous donnons est l'alignement des sites actifs de la trypsine bovine (identifiant PDB : 1aq7) et de l'α-thrombine humaine (identifiant PDB : 1c5o), deux enzymes de la super-famille des sérine-protéases dont la fonction générale est la catalyse de l'hydrolyse de liaisons peptidiques covalentes [19]. Ces deux protéines partagent 55% d'identité entre les résidus des cavités, et le score de similarité structurale est de 0,90.

Figure 3-14. Sites actifs alignés de la thrombine α humaine (en vert) et de la trypsine bovine (en rouge).

Le second exemple d'application est le cas des sites actifs de l'acétylcholinestérase (identifiant PDB : 1odc) et de la butyrylcholinestérase (identifiant PDB : 1p0p) (figure 3-15). Ces deux enzymes sont des hydrolases de l'acétylcholine (un neurotransmetteur) et de la butyrylcholine (un composé synthétique très proche de l'acétylcholine) [20]. Le

score de similarité structural associé à l'alignement est 0,96. Il faut noter que cet outil d'alignement est capable d'aligner des sites de tailles différentes.

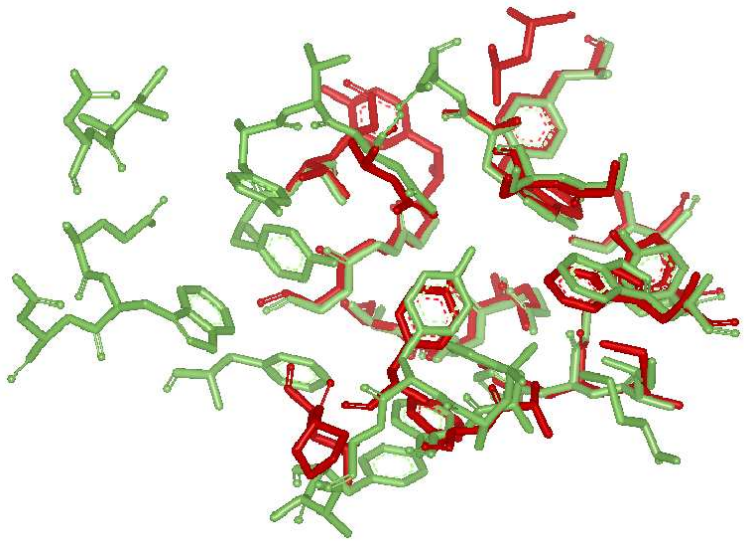


Figure 3-15. Sites actifs alignés de l'acétylcholinestérase (en vert) et de la butyrylcholinestérase (en rouge).

3.4 Conclusion

Nous avons dans ce chapitre classifié les RCPG d'après la structure tridimensionnelle de leurs cavités transmembranaires. Des descripteurs physico-chimiques et géométriques ont été projetés sur deux sphères dans les cavités à comparer, puis un score de similarité a été calculé entre les sphères chargées d'information. Les structures des RCPG, construites par modélisation par homologie, étaient préalignées sur la structure de base qui a permis la modélisation : la rhodopsine bovine. Mais ces alignements n'étaient pas suffisamment précis pour construire une classification. Nous avons donc créé un outil d'alignement structural qui, guidé par le score entre les deux sphères dans les cavités, cherche le meilleur alignement entre les deux cavités en cherchant à maximiser le score.

Ainsi, après raffinement des alignements, nous avons construit une nouvelle classification des RCPG. Celle-ci est très homogène avec la classification basée sur des séquences de résidus critiques des cavités transmembranaires (chapitre 2). En général, les clusters construits possèdent moins de récepteurs que ceux du chapitre 2, et il y a donc plus de singletons. Il semble que seul le singleton DUFFY de la classification précédente ait pu être classé ici dans le cluster des chimiokines.

Il est difficile d'interpréter le score donné par notre outil d'alignement. Ce score est normalisé donc permet de construire une classification, mais il ne permet pas, en revanche, de déterminer à lui seul si l'alignement a réussi ou échoué. Toutefois, les alignements réussi sont visibles « à l'œil », comme les deux exemples que nous donnons en fin de chapitre.

Bibliographie

[1] Palczewski K., Kumasaka T., Hori T., Behnke C.A., Motoshima H., Fox B.A., Le Trong I., Teller D.C., Okada T., Stenkamp R.E., Yamamoto M., Miyano M. (2000) Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **289**:739–745.

[2] Bissantz C., Logean A., Rognan D. (2004) High-Throughput Modelling of Human G-Protein Coupled Receptors: Amino Acid Sequence Alignment, Tree-Dimensional Model Building, and Receptor Library Screening. *J. Chem. Inf. Comput. Sci.* **44(3)**: 1162–1176.

[3] Morin E. (1981) La méthode 1. La Nature de la Nature. *Seuil*.

[4] Hofstadter D. (2000) Gödel, Escher, Bach. Les Brins d'une Guirlande Eternelle. *Dunod*.

[5] Lathrop R.H. (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* **7**:1059–1068.

[6] Godzik A. (1996) The structural alignment between two proteins: Is there a unique solution? *Protein Sci.* **5(7)**:1325–1338.

[7] Shindyalov I.N., Bourne P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of optimal path. *Protein Eng.* **11(9)**:739–747

[8] Shulman-Peleg A., Nussinov R., Wolfson H.J. (2004) Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **339**:607–633.

[9] Liang J., Edelsbrunner H., Woodward C. (1998) Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **7**:1884–1897.

[10] Goldberg D.E. (1991) Genetic Algorithms. *Addison Wesley.*

[11] Berg J.M., Tymoczko J.L., Stryer L. (2002) Biochemistry (5th edition). *W. H. Freeman and Company.*

[12] Schmitt S., Kuhn D., Klebe G. (2002) A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **323**:387–406.

[13] Holliday J.D., Salim N., Whittle M., Willett P. (2003) Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.* **43**:819–828.

[14] Black P.E. " Gray code " *in* Dictionary of Algorithms and Data Structures. `<URL:http://www.nist.gov/dads/>`.

[15] Kostenis E. (2004) Novel clusters of receptors for sphingosine-1-phosphate, sphingosylphosphorylcholine, and (lyso)-phosphatic acid: new receptors for "old" ligands. *J. Cell. Biochem.* **92(5)**:923–936.

[16] Berman H.M., Westbrook J., Feng Z., Gilliand G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**:235–242.

[17] SYBYL 7.1. TRIPOS, Assoc. Inc., St-Louis, MO.

[18] Kellenberger E., Muller P., Schalon C., Bret G., Foata N., Rognan D. (2006) sc-PDB: an Annotated Database of Druggable Binding Sites from the Protein Data Bank. *J. Chem. Inf. Model.* **46**:717–727.

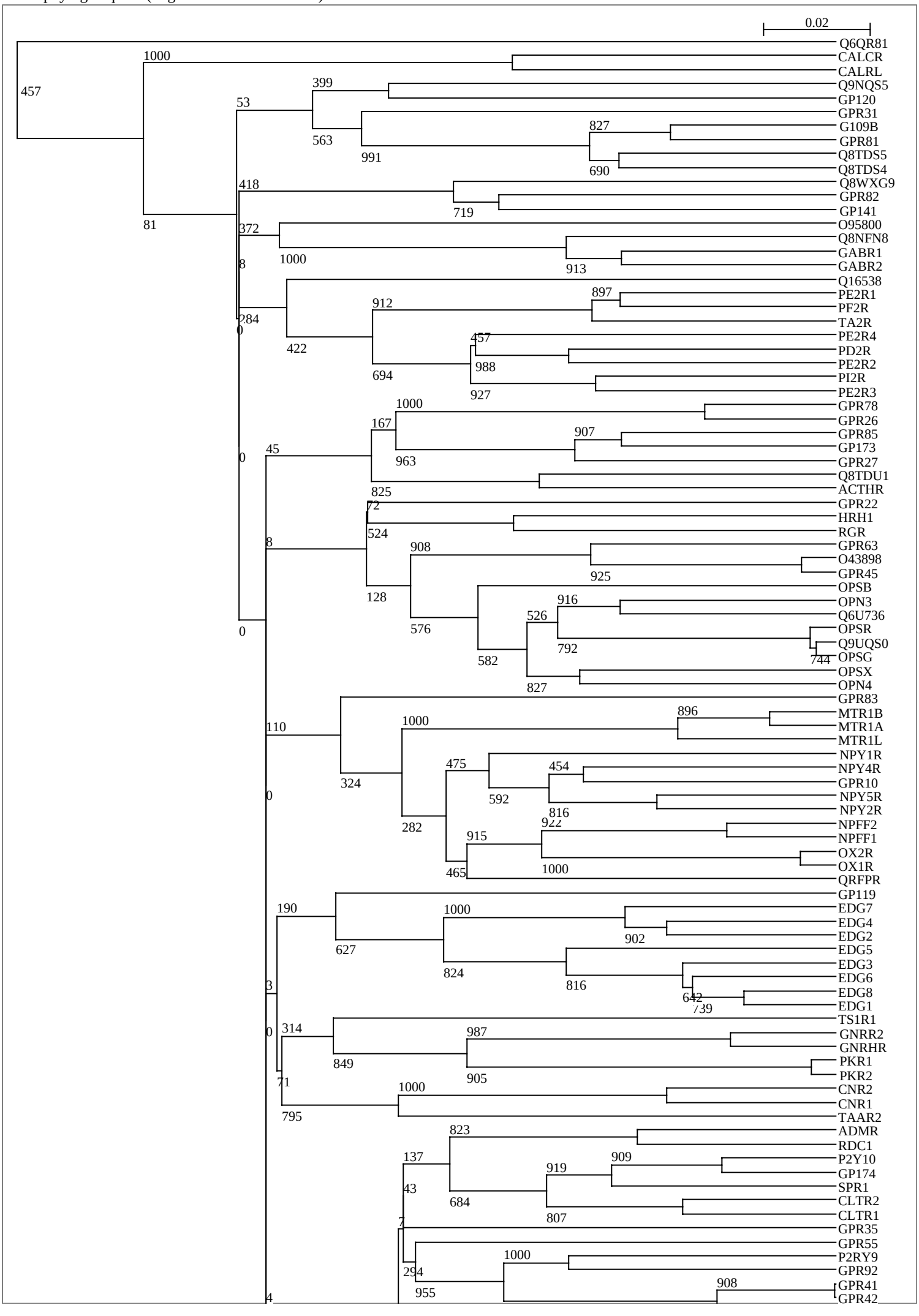
[19] Rawlings N.D., Barrett A.J. (1994) Families of serine peptidases. *Meth. Enzymol.* **244**:19–61.

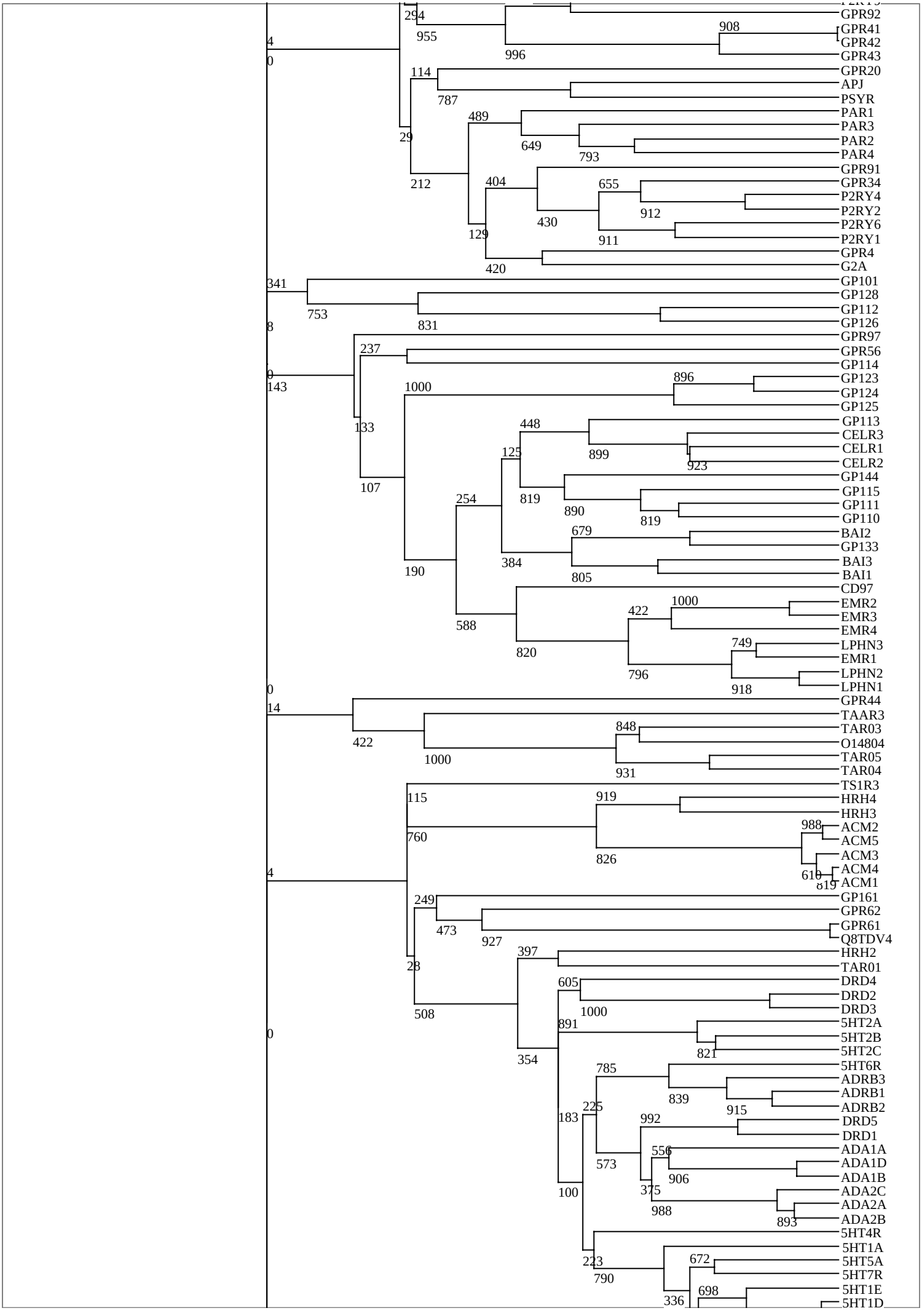
[20] Potocka J., Kuca K., Jun D. (2004) Acetylcholinesterase and butyrylcholinesterase – important enzymes of human body. *Acta Medica* **47(5)**:215–228.

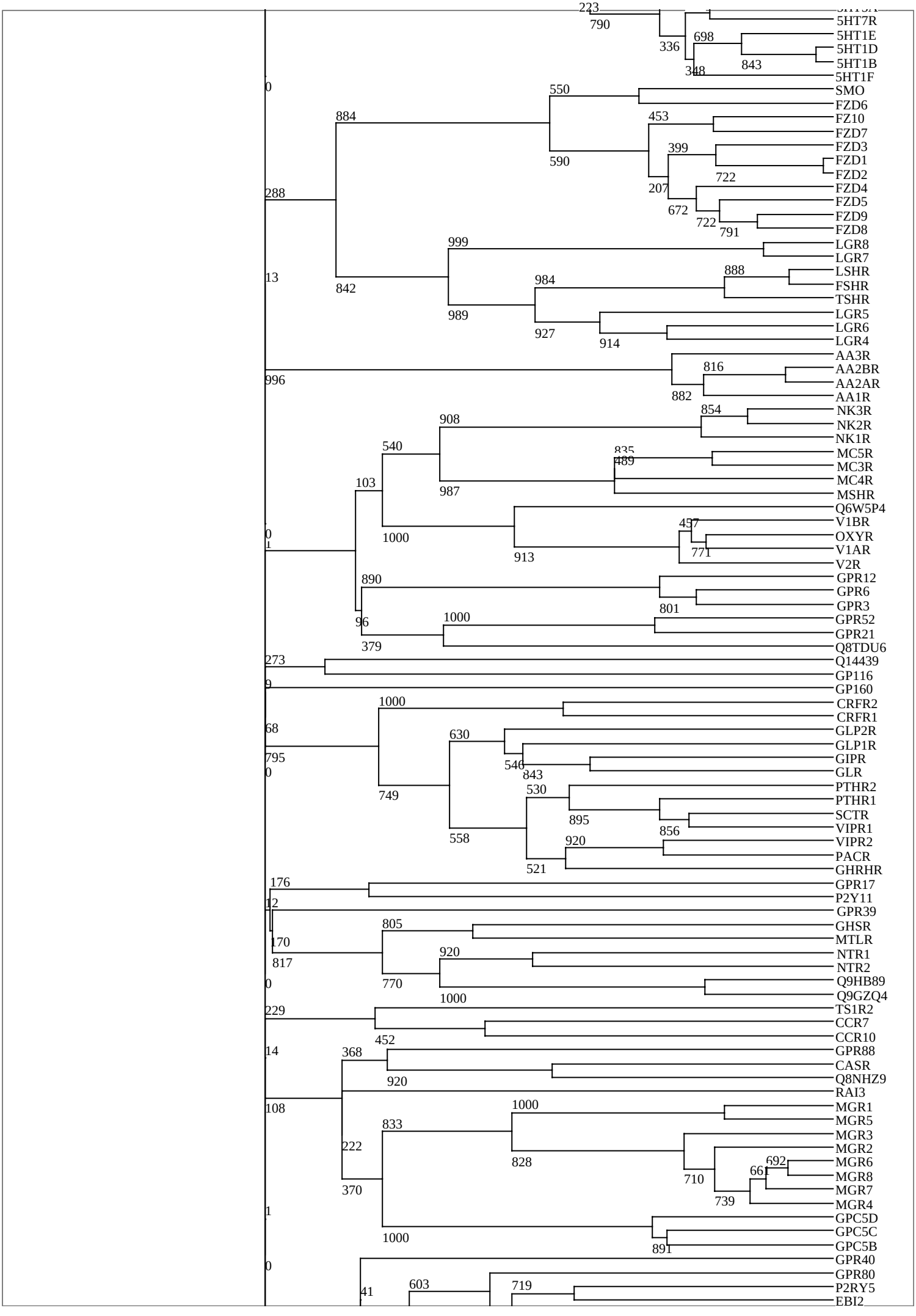
[21] Evers A., Klabunde T. (2005) Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J. Med. Chem.* **48(4)**:1088–1097.

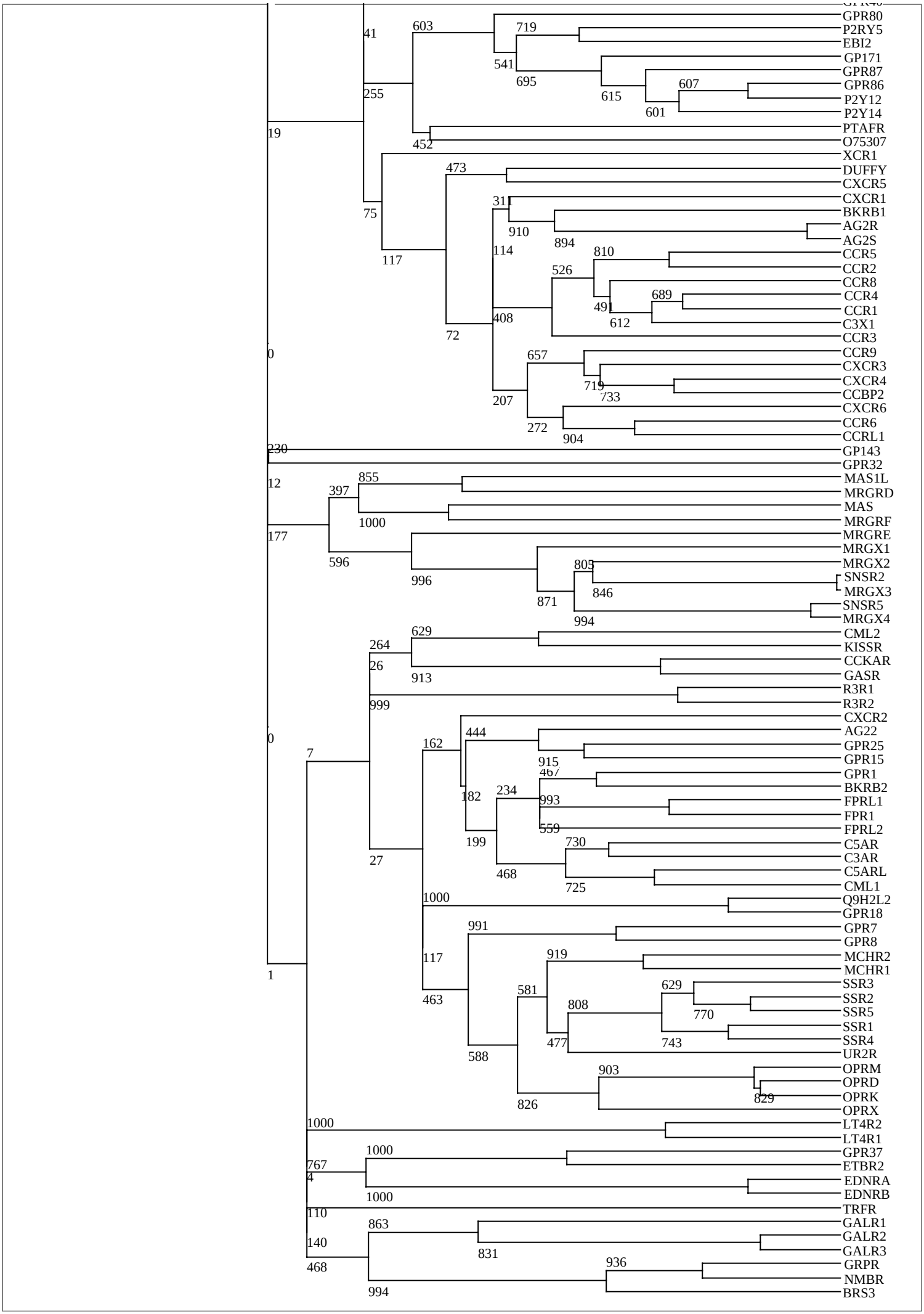
Annexe C

Arbre phylogénique 2 (alignement 3-D de cavités)




GPR92
GPR41
GPR42
GPR43
GPR20
APJ
PSYR
PAR1
PAR3
PAR2
PAR4
GPR91
GPR34
P2RY4
P2RY2
P2RY6
P2RY1
GPR4
G2A
GP101
GP128
GP112
GP126
GPR97
GPR56
GP114
GP123
GP124
GP125
GP113
CELR3
CELR1
CELR2
GP144
GP115
GP111
GP110
BAI2
GP133
BAI3
BAI1
CD97
EMR2
EMR3
EMR4
LPHN3
EMR1
LPHN2
LPHN1
GPR44
TAAR3
TAR03
O14804
TAR05
TAR04
TS1R3
HRH4
HRH3
ACM2
ACM5
ACM3
ACM4
ACM1
GP161
GPR62
GPR61
Q8TDV4
HRH2
TAR01
DRD4
DRD2
DRD3
5HT2A
5HT2B
5HT2C
5HT6R
ADRB3
ADRB1
ADRB2
DRD5
DRD1
ADA1A
ADA1D
ADA1B
ADA2C
ADA2A
ADA2B
5HT4R
5HT1A
5HT5A
5HT7R
5HT1E
5HT1D


5HT7R
5HT1E
5HT1D
5HT1B
5HT1F
SMO
FZD6
FZ10
FZD7
FZD3
FZD1
FZD2
FZD4
FZD5
FZD9
FZD8
LGR8
LGR7
LSHR
FSHR
TSHR
LGR5
LGR6
LGR4
AA3R
AA2BR
AA2AR
AA1R
NK3R
NK2R
NK1R
MC5R
MC3R
MC4R
MSHR
Q6W5P4
V1BR
OXYR
V1AR
V2R
GPR12
GPR6
GPR3
GPR52
GPR21
Q8TDU6
Q14439
GP116
GP160
CRFR2
CRFR1
GLP2R
GLP1R
GIPR
GLR
PTHR2
PTHR1
SCTR
VIPR1
VIPR2
PACR
GHRHR
GPR17
P2Y11
GPR39
GHSR
MTLR
NTR1
NTR2
Q9HB89
Q9GZQ4
TS1R2
CCR7
CCR10
GPR88
CASR
Q8NHZ9
RAI3
MGR1
MGR5
MGR3
MGR2
MGR6
MGR8
MGR7
MGR4
GPC5D
GPC5C
GPC5B
GPR40
GPR80
P2RY5
EBI2


GPR80
P2RY5
EBI2
GP171
GPR87
GPR86
P2Y12
P2Y14
PTAFR
O75307
XCR1
DUFFY
CXCR5
CXCR1
BKRB1
AG2R
AG2S
CCR5
CCR2
CCR8
CCR4
CCR1
C3X1
CCR3
CCR9
CXCR3
CXCR4
CCBP2
CXCR6
CCR6
CCRL1
GP143
GPR32
MAS1L
MRGRD
MAS
MRGRF
MRGRE
MRGX1
MRGX2
SNSR2
MRGX3
SNSR5
MRGX4
CML2
KISSR
CCKAR
GASR
R3R1
R3R2
CXCR2
AG22
GPR25
GPR15
GPR1
BKRB2
FPRL1
FPR1
FPRL2
C5AR
C3AR
C5ARL
CML1
Q9H2L2
GPR18
GPR7
GPR8
MCHR2
MCHR1
SSR3
SSR2
SSR5
SSR1
SSR4
UR2R
OPRM
OPRD
OPRK
OPRX
LT4R2
LT4R1
GPR37
ETBR2
EDNRA
EDNRB
TRFR
GALR1
GALR2
GALR3
GRPR
NMBR
BRS3

Conclusion

Dans ce travail de thèse, nous avons construit deux classifications des RCPG. La première est basée sur des séquences discontinues de 30 résidus critiques de la cavité transmembranaire des récepteurs. La seconde est basée sur des propriétés physico-chimiques et géométriques des structures tridimensionnelles de ces mêmes cavités. Ces classifications sont homogènes avec la classification basée sur les ligands connus des récepteurs, c'est-à-dire que nous retrouvons classés ensemble les récepteurs qui reconnaissent les mêmes ligands. De plus, nous avons proposé une affectation à certains récepteurs orphelins. Enfin, la seconde classification nous a conduit à construire un outil d'alignement structural simple, souple dans le choix des résidus à considérer dans l'alignement, et dont le code source est connu et modifiable (ce qui n'est pas le cas d'un code propriétaire).

Il serait intéressant de tester biologiquement l'affinité des ligands que nous proposons pour certains récepteurs orphelins (voir le tableau 2-11 du chapitre 2). Au fur et à mesure de la déorphanisation des récepteurs orphelins, les propositions émises pourront être confirmées ou infirmées, et la classification en sera affinée. Ce ne sont pas des classifications figées que nous avons proposées, mais des classifications dont le but est de faire progresser la recherche de nouveaux médicaments ciblant les RCPG.