ANALYSE EVOLUTIVE DES RECEPTEURS COUPLES AUX PROTEINES G (RCPG)

Julien Pelé

université angers

ANNÉE 2010
N° d'ordre : 1114

ANALYSE EVOLUTIVE DES RECEPTEURS COUPLES AUX PROTEINES G (RCPG)

THÈSE DE DOCTORAT

Spécialité : Bioinformatique

ECOLE DOCTORALE BIOLOGIE/SANTÉ

Présentée et soutenue publiquement

le 20 décembre 2010

à Angers

par

Julien Pelé

Devant le jury ci-dessous :

Pr C. Delamarche (rapporteur), PU, UMR CNRS 6026, Rennes
Dr J.F. Gibrat (rapporteur), DR, INRA, Jouy-en-Josas
Pr H. Abdi (examinateur), PU, School of Behavioral and Brain Sciences, Richardson, USA
Dr C. Legros (examinateur), MCU, UPRES EA 2647, Angers
Dr M. Chabbert (examinateur), CR, UMR CNRS 6214 – INSERM 771, Angers

Directeur de thèse : Dr M. Chabbert, CR, UMR CNRS 6214 – INSERM 771, Angers

Laboratoire :
UMR CNRS 6214 – INSERM 771, Laboratoire de Biologie NeuroVasculaire Intégrée
UFR Sciences médicales
Rue Haute de Reculée
49045 ANGERS CEDEX 01

ED 502

ENGAGEMENT DE NON PLAGIAT

Je, soussigné (e) ……, déclare être pleinement conscient(e) que le plagiat de documents ou d'une partie d'un document publiés sur toutes formes de support, y compris l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée. En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour écrire ce rapport ou mémoire.

Signature :

Cet engagement de non plagiat doit être inséré en première page de tous les rapports, dossiers, mémoires.

Table des matières

Table des figures....5
Table des tableaux....5
Remerciements....7
Avant-propos....8
1. CONTEXTE....9
1.1 Présentation du laboratoire d'accueil....10
1.2 Présentation du projet....11
2. METHODES D'ETUDE DE L'EVOLUTION DES FAMILLES DE PROTEINES....13
2.1 Méthodes de phylogénie....14
2.2 Méthodes de réduction dimensionnelle....16
2.3 Principe de la coévolution....16
3. LES RECEPTEURS COUPLES AUX PROTEINES G....20
3.1 Généralités sur les RCPG....21
3.2 Structure et mécanisme d'activation....23
3.3 Classification....28
3.3.1 Les différentes familles de récepteurs humains....29
3.3.2 Les sous-familles de la classe A....32
3.4 Évolution de la classe A....36
3.5 Application des méthodes de corrélation aux RCPG....38
3.6 Travaux de l'équipe de bioinformatique....39
4. OBJECTIFS DETAILLES....42
5. OUTILS MIS EN PLACE....45
5.1 Jeu de données....46
5.2 Distances entre séquences....46
5.2.1 Calcul de la différence....47
5.2.2 Calcul de la dissimilarité....47
5.3 DISTATIS....49
5.4 MDS métrique....50
5.5 Projection d'éléments supplémentaires....52
5.6 Clustering....54
5.7 Analyse des groupes....55
5.7.1 Termes communs....55
5.7.2 Entropie....57
5.7.3 Différentes méthodes de corrélation entre groupes et séquences....57
5.7.3.1 Méthodes basées sur le chi-2....57
5.7.3.2 Méthodes basées sur l'information mutuelle....58
5.7.3.3 Optimisation de l'entropie combinatoire....59
5.7.3.4 Comparaisons....59
5.8 Méthodes d'analyse des mutations corrélées....60
5.8.1 Méthodes basées sur le chi-2....64
5.8.2 Méthodes basées sur l'information mutuelle....64
5.8.3 Méthodes asymétriques....65
5.8.3.1 Analyse de couplage statistique....66
5.8.3.2 Probabilité explicite de covariation de sous-ensembles....66
5.8.4 Méthodes basées sur des matrices de substitution....66
5.8.5 Implémentation....67

5.9 Base de données relationnelle....67
5.9.1 Modèle conceptuelle de données....68
5.9.2 Interface AJAX....70
5.10 Package R bio2mds....71
6. RESULTATS....75
6.1 L'analyse de l'évolution des RCPG de classe A par MDS métrique....76
6.1.1 Article : A Novel Multidimensional Scaling Technique Reveals the Main Evolutionary Pathways of Class A G-Protein-Coupled Receptors....76
6.1.2 Conclusion....98
6.2 Application des méthodes de mutations corrélées aux RCPG de classe A....102
6.2.1 Caractéristiques des trois jeux de données....102
6.2.2 Analyse par entropie....102
6.2.3 Comparaison des méthodes d'AMC....105
6.2.4 Alignement multiple de positions corrélées....108
6.2.5 Application d'OMES2 aux trois jeux de données....112
6.2.6 Visualisation sur la structure de CXCR4....113
6.2.7 Conclusion....115
7. CONCLUSIONS ET PERSPECTIVES....117
Liste des abréviations....122
Bibliographie....125
ANNEXES....133
A. Formule du calcul de dissimilarités....134
B. Formules des méthodes de corrélation entre groupes et séquences....135
B.1 Rappel des notations....135
B.2 Méthodes basées sur le chi-2....135
B.2.1 Définitions de Nobs et Nex....135
B.2.2 Démonstration d'OMES1....136
B.2.3 Démonstration d'OMES2....136
B.3 Méthodes basées sur l'information mutuelle....137
B.3.1 Démonstration de la MI classique....137
B.3.2 Démonstration de la MI normalisée....137
B.4 Optimisation de l'entropie combinatoire....138
C. Formules des méthodes d'AMC....139
C.1 Méthodes basées sur le chi-2....139
C.2 Méthodes basées sur l'information mutuelle....140
C.3 Méthodes asymétriques....142
C.4 Méthodes basées sur des matrices de substitution....143
D. Applications du package R bio2mds....146
D.1 Contenu....146
D.2 Structure....148
D.3 Documentation....148
D.4 Applications....150
D.4.1 Initialisation....150
D.4.2 Lecture d'un AMS....151
D.4.3 Calcul de distances....151
D.4.4 Comparaison des matrices de distances....152
D.4.5 MDS métrique en 2 dimensions....153
D.4.6 MDS métrique en 3 dimensions....157

Table des figures

Figure 1 : Décomposition de la coévolution....19
Figure 2 : Transduction du signal par les RCPG....22
Figure 3 : Structure schématique des RCPG....25
Figure 4 : Cycle d'activation des RCPG....27
Figure 5 : Relations phylogénétiques entre les RCPG du génome humain....31
Figure 6 : Relations phylogénétiques entre les RCPG de classe A du génome humain....33
Figure 7 : Arbre d'évolution des RCPG pour différents génomes....37
Figure 8 : Motifs proline sur les arbres phylogénétiques des RCPG de classe A....40
Figure 9 : Méthode DISTATIS....51
Figure 10 : Projection d'éléments supplémentaires dans le cadre de la MDS....53
Figure 11 : Analyse des corrélations groupes-séquences....56
Figure 12 : Comparaison des méthodes de corrélation groupes-séquences....61
Figure 13 : Analyse des mutations corrélées....62
Figure 14 : Modèle conceptuel des données amélioré....69
Figure 15 : Interface en AJAX....72
Figure 16 : Mesure d'entropies des HTM2, 4 et 6....104
Figure 17 : Comparaison des méthodes d'AMC....107
Figure 18 : Alignement multiple de positions corrélées....111
Figure 19 : Visualisation des réseaux de corrélation sur la structure cristalline de CXCR4....114
Figure 20 : Localisation des structures connues de RCPG sur l'analyse MDS....119
Figure 21 : Workflow du package R bio2mds....149
Figure 22 : Comparaison par DISTATIS des matrices de distances....154
Figure 23 : Graphique d'éboulis et projections par MDS métrique....156
Figure 24 : Visualisation sous Pymol de la projection de séquences....159

Table des tableaux

Tableau 1 : Structures résolues de RCPG de classe A....24
Tableau 2 : Matrices de substitution....48
Tableau 3 : Les différentes méthodes d'AMC....63

« Plus une découverte est originale, plus elle semble évidente par la suite »
Arthur Koestler

Remerciements

Tout d'abord, je tiens à remercier Mme le Dr. Marie Chabbert pour m'avoir accueilli dans l'équipe de bioinformatique et encadré tout au long de ce doctorat. Elle m'a fait confiance et toujours laissé une certaine liberté de travail que j'ai grandement appréciée. J'espère avoir répondu à ses attentes et contribué aux avancés dans le domaine de recherche du laboratoire.

Un grand merci à M. le Dr Daniel Henrion, Directeur de l'unité mixte CNRS 6214-INSERM U771, pour m'avoir accueilli durant 3 ans au sein de son laboratoire.

Je remercie le Conseil Général pour avoir assuré le soutien financier de ma thèse. Je remercie l'Université d'Angers, plus précisément le Collège Doctorale et l'Ecole Doctorale Biologie Santé pour leur disponibilité et leur professionnalisme.

Je remercie M. le Pr Delamarche et M. le Dr Gibrat pour avoir accepter d'être rapporteurs pour ma thèse de doctorat.

Je remercie M. le Pr Abdi et M. le Dr Legros pour avoir bien voulu accepter de participer au jury de thèse.

Je remercie M. le Pr Abdi pour sa méthode de projection d'éléments supplémentaires par MDS et ses conseils techniques pour l'appliquer à l'analyse évolutive des RCPG.

Je remercie tous les étudiants qui ont effectué un passage dans l'équipe de bioinformatique. J'espère avoir répondu le plus justement à leurs questions et qu'ils auront passé un bon moment en ma compagnie. Je leur souhaite à tous une bonne continuation.

Dans une moindre mesure, je remercie le Restaurant Universitaire de la Faculté de Médecine d'Angers pour m'avoir nourri, surtout à base de frites!

Je remercie mes parents (ils se reconnaîtront!) pour la logistique (on va dire) et pour m'avoir posé la devenue célèbre question quotidienne : « Alors ça avance l'article ? ».

Je remercie mon frère pour avoir toujours su squatter mon PC.

Je remercie également toutes les personnes que j'ai pu oublier (pour ne froisser personne!)

Pour finir, à titre personnel, je remercie les personnes qui ont participé, de près ou de loin, à la réalisation des logiciels libres, sans lesquels la réalisation de ce projet de thèse et la rédaction de ce présent document auraient été beaucoup plus ardues.

Avant-propos

Après un Master 1 en Sciences, Technologies et Ingénierie de la Santé, je souhaitais me diriger vers un autre domaine, pour acquérir une double compétence. J'avais déjà eu connaissance de la bioinformatique car j'avais été familiarisé à ce domaine à de nombreuses reprises lors de mon cursus. Ainsi, je m'étais orienté vers une formation purement informatique, un Master 2 Compétences Complémentaires en Informatique. Pour le stage d'informatique à accomplir, je me suis aiguillé d'emblée vers un projet alliant mes deux domaines de compétences. Le stage a été effectué dans l'équipe de bioinformatique du Dr Chabbert, du Laboratoire de Biologie NeuroVasculaire Intégrée de la Faculté de Médecine d'Angers. J'ai pu appliquer mes connaissances dans le contexte de la recherche scientifique. Le niveau d'initiative nécessaire à la mise en place des outils informatiques avait contribué à ma motivation. Le contexte biologique était très intéressant car il portait sur un sujet qui avait des implications dans la santé humaine. L'aspect statistique a été le plus délicat à aborder car c'était le seul domaine qui sortait de mes compétences initiales. Cependant, après avoir fait preuve de persévérance, ce stage a été une réelle expérience de travail et d'adaptation.

Sur cette lancée, le laboratoire et moi avions décidé de poursuivre cette aventure en doctorat. L'approche a été différente car le niveau d'exigence a été plus élevé, l'initiative davantage présente et la durée beaucoup plus longue. Il a fallu poursuivre les recherches à partir des acquis de l'équipe de bioinformatique. Tout au long de ces 3 ans (et 1 jour, pour être précis), mon souci a toujours été d'essayer d'apporter des applications nouvelles et de tenter d'avoir une approche novatrice sur le sujet. Avec cette thèse de doctorat, j'ai essayé d'apporter un certain compromis entre les trois domaines abordés (la biologie, les statistiques et l'informatique) et toujours prendre en considération les néophytes, en expliquant bien les concepts, et les spécialistes, en ne vulgarisant pas à outrance. Le projet m'a donné le sens de l'organisation, il m'a permis de pratiquer de nombreux langages informatiques. L'aspect le plus passionnant de ce projet a été d'associer plusieurs domaines et de savoir que ce projet participe, à sa manière, au progrès des connaissances dans le domaine des RCPG.

L'aspect le plus positif que je retiendrai de cette formation c'est la flexibilité qu'il faut toujours avoir à l'esprit pour s'adapter le mieux possible au défi que représente la recherche : apporter des connaissances nouvelles. Toute cette expérience que j'ai pu acquérir a été bénéfique et j'espère qu'elle me servira pour ma future vie professionnelle.

1. CONTEXTE

1.1 Présentation du laboratoire d'accueil

La thèse de doctorat s'est déroulée au Laboratoire de Biologie NeuroVasculaire Intégrée, UMR CNRS 6214 - INSERM U771, de la Faculté de Médecine d'Angers. L'axe de recherche du laboratoire est l'étude de la circulation sanguine dans les petits vaisseaux (*i.e.*, microcirculation) en lien avec les pathologies humaines associées. L'étude de la microcirculation est originale et prometteuse car elle joue un rôle crucial dans les maladies cardiovasculaires (hypertension et myopathie) et le diabète chez l'homme.

L'unité comprend une quarantaine de personnes et est constituée de plusieurs équipes qui étudient différents aspects de la microcirculation :

- l'équipe dirigée par D. Henrion étudie le rôle de la mécanotransduction et des systèmes neuro-humoraux locaux dans le remodelage vasculaire,

- l'équipe dirigée par L. Loufrani travaille sur des protéines impliquées dans le cytosquelette en lien avec la mécanotransduction,

- l'équipe dirigée par G. Leftheriotis porte son travail sur la dysfonction vasculaire dans le cadre de pathologies métaboliques,

- l'équipe dirigée par M.A. Custaud a pour objectif d'étudier l'influence des conditions environnementales et du déconditionnement sur le système cardiovasculaire,

- l'équipe dirigée par N. Guérineau se focalise sur l'action d'hormones médullosurrénaliennes en lien avec la physiopathologie du stress et les pathologies vasculaires,

- l'équipe dirigée par M. Chabbert travaille dans le domaine de la bioinformatique et s'intéresse davantage à la problématique structurale et fonctionnelle des récepteurs couplés aux protéines G (RCPG) impliqués dans la régulation cardio-vasculaire. Cet axe de recherche est basé sur l'analyse des séquences protéiques et la modélisation moléculaire

dans le but de mettre en évidence le rôle de résidus spécifiques et de comprendre les mécanismes d'activation des RCPG.

1.2 Présentation du projet

Le projet de thèse a consisté à s'appuyer sur les travaux déjà engagés dans l'équipe de bioinformatique, au milieu des années 2000, par Julie Devillé et Marie Chabbert (voir chapitre 3.6) sur l'analyse et l'évolution des motifs structuraux des RCPG de classe A. En particulier, l'équipe avait montré que la proline dans l'hélice transmembranaire 2 jouait un rôle important dans la structure des récepteurs d'intérêt pour le laboratoire, comme les récepteurs de l'angiotensine II (AT1 et AT2), et que ces motifs étaient susceptibles d'être de véritables marqueurs évolutifs.

Concernant mon projet, nous avons poursuivi l'analyse évolutive mais d'un point de vue des séquences uniquement. Nous avons étudié l'évolution des RCPG de classe A grâce à une nouvelle approche méthodologique et mis en évidence les motifs clés de chacune des sous-familles. Nous avons émis une hypothèse de schéma évolutif des RCPG de classe A et voulu la relier aux mécanismes d'activation de ces récepteurs. Pour cela, nous avons conduit une étude comparative des méthodes de coévolution pour en sélectionner la plus adaptée. Ensuite, nous l'avons appliquée à nos données pour tenter de mettre en évidence les liens qui unissent contraintes évolutives et fonctionnelles des RCPG de classe A.

Ces exigences ont demandé une certaine polyvalence de travail et ont mis à contribution trois domaines de compétences interdépendants :

- l'aspect biologique : le traitement d'alignement multiple de séquences (AMS) constitue le point de départ du projet. L'interprétation des résultats se fait en lien avec les connaissances actuelles sur la structure tridimensionnelle et les mécanismes d'activation des RCPG.

- l'aspect statistique : les méthodes employées font appel aux notions de réduction dimensionnelle, de clustering et de covariation. L'apport majeur fut l'application d'une méthode statistique inédite liée à la réduction dimensionnelle dans le but d'étudier l'évolution des RCPG.

- l'aspect informatique : les analyses furent conduites à l'aide de scripts informatiques (*i.e.*, programmes), écrits en différents langages informatiques, et une partie des méthodes statistiques est accessible par l'intermédiaire d'un package R (langage informatique spécialisé dans les statistiques) qui fut spécialement développé.

Dans un premier temps, le principe de la phylogénie sera expliqué en lien avec les différentes approches méthodologiques qui permettent d'analyser des familles de protéines. Le support biologique sera présenté pour mieux appréhender les enjeux du travail effectué et le contexte dans lequel le projet se situe. Les objectifs seront passés en revue avec la présence de détails techniques. Ensuite, la présentation des différents outils mis en place permettra d'appréhender les aspects statistiques et informatiques du projet. Puis, les résultats seront exposés en partie sous la forme d'un article scientifique. Une conclusion, ainsi que plusieurs perspectives de recherche concernant l'équipe de bioinformatique clôtureront cette thèse de doctorat. Des annexes viennent compléter les différents chapitres avec des détails concernant les formules mathématiques utilisées et une documentation utilisateur pour le package R.

2. METHODES D'ETUDE DE L'EVOLUTION DES FAMILLES DE PROTEINES

La phylogénie est l'étude de l'histoire évolutive des protéines en utilisant des représentations graphiques adaptées. D'un point de vue moléculaire, c'est l'étude des mutations qui s'opèrent au niveau de différentes positions de la séquence protéique permettant d'émettre des hypothèses sur les mécanismes d'évolution, qui ont permis d'obtenir la diversité de la famille protéique étudiée. La phylogénie est habituellement étudiée par des méthodes permettant d'obtenir des arbres évolutifs. Cependant, dans certaines circonstances, il est utile de compléter ces arbres. Par exemple, dans certaines familles protéiques, l'analyse d'orthologues (même protéine d'espèces différentes) est parfois complexe à cause d'un grand nombre de paralogues (protéines homologues dans une espèce donnée). Des méthodes alternatives telle la réduction dimensionnelle peuvent donc être utilisées. D'où l'existence de deux types de méthodes pour mettre en évidence les relations entre les protéines : les arbres phylogénétiques et la réduction dimensionnelle. Ces différentes méthodes graphiques permettent de visualiser facilement les relations entre séquences protéiques et étudier leurs mécanismes évolutifs.

2.1 Méthodes de phylogénie

Découvrir l'histoire évolutive d'une famille de protéines nécessite un certain nombre d'hypothèses. La première concerne les séquences protéiques utilisées pour la reconstruction phylogénétique qui doivent être homologues, c'est à dire qu'elles doivent partager une origine commune puis avoir divergé au cours de l'évolution. La divergence phylogénétique est supposée être bifurquée, à savoir que l'arbre phylogénétique donne naissance à deux branches filles pour n'importe quel nœud donné. La raison de la bifurcation des branches de l'arbre provient du processus évolutif de dichotomie. Parfois, les nœuds des arbres peuvent donner naissance à plusieurs branches filles et deviennent donc multifurqués. Cette polytomie peut résulter d'une résolution insuffisante de l'arbre (polytomie « artefact ») dans lequel l'ordre exact des bifurcations ne peut pas être déterminé précisément. Elle peut aussi être réelle (polytomie « dure »). Dans ce cas, elle est liée au mécanisme d'évolution appelé radiation [1,2]. Une autre hypothèse en phylogénétique est que chaque position dans une séquence évolue indépendamment des autres. Pour construire un arbre phylogénétique à partir d'un AMS, il faut suivre différentes étapes : choisir un modèle évolutif, déterminer la méthode de construction à utiliser et évaluer la fiabilité de l'arbre obtenu.

La première étape consiste à évaluer le degré de divergence entre deux séquences. Il peut être

calculé de manière simple en se basant sur le pourcentage de différence ou en utilisant des modèles d'évolution complexes lorsque les positions d'une séquence n'évoluent pas selon la même vitesse. Dans le cadre de séquences protéiques, ces modèles complexes se basent sur l'utilisation de matrices de substitution, telles que la Mutation Acceptée pour 100 acides aminés (PAM pour Point Accepted Mutation) ou celle de Jones, Taylor et Thornton (JTT pour Jones, Taylor, Thornton) et/ou différentes corrections, telles que la correction de Poisson ou Gamma. Des études empiriques ont montré que dans la reconstruction phylogénétique, les distances simples, telles que le pourcentage de différence, surpassent souvent les distances complexes (sauf dans le cas de divergence entre espèces) [3].

La seconde étape consiste à sélectionner une méthode de construction de l'arbre. On peut les diviser en deux catégories : les méthodes basées sur les caractères et celles basées sur la distance. L'hypothèse de base pour la première catégorie concerne les caractères des différentes positions qui doivent être homologues. De plus, les caractères doivent évoluer de manière indépendante. Cette catégorie inclut le Maximum de Parcimonie (MP pour Maximum Parsimony) et le Maximum de Vraisemblance (ML pour Maximum Likelihood). Le MP est une méthode lente mais tend à produire des arbres plus précis que ceux basés sur la distance lorsque la divergence est faible. Le ML est beaucoup plus robuste mais lorsque le nombre de taxa est élevé, il devient presque impossible de l'utiliser. Pour résoudre ce problème, de nouvelles approches ont été développées, telles que l'analyse bayésienne. La deuxième catégorie suppose que toutes les séquences impliquées soient homologues et que les branches de l'arbre soient additives, c'est à dire que la distance entre deux taxa est égale à la somme des longueurs de leurs branches respectives. Cette catégorie comprend l'UPGMA (UPGMA pour Unweighted Pair Group Method with Arithmetic Mean) et le Joindre les Voisins (NJ pour Neighbor Joining). L'avantage majeur de ces méthodes est leur rapidité et leur capacité à traiter de grands jeux de données.

La troisième étape concerne l'évaluation statistique de la fiabilité des arbres phylogénétiques qui est réalisée par l'intermédiaire de techniques statistiques spécifiques : le bootstrapping (*i.e.*, rééchantillonnage), par exemple. Cette technique répète la construction de l'arbre phylogénétique par échantillonnage répété du jeu de données. Les fluctuations qui sont provoquées dans le jeu de données produisent de multiples arbres légèrement différents. Leur comparaison permet de déterminer la robustesse et la reproductibilité de l'arbre phylogénétique par le calcul d'un degré de confiance dans la topologie de l'arbre.

La phylogénie est un outil fondamental pour comprendre l'évolution et les relations des séquences protéiques. Il est important de réaliser que la construction d'arbres phylogénétiques n'est pas triviale et nécessite de sélectionner une méthode adaptée à ses données. Pour minimiser les erreurs phylogénétiques, il est conseillé d'utiliser plusieurs méthodes pour vérifier la robustesse des résultats obtenus. Cependant, lorsqu'on se base sur une hypothèse de dépendance des positions, lorsque la taille du jeu de données devient conséquente et lorsqu'on souhaite comparer les relations qui existent entre des protéines d'une même famille mais de génomes différents, il est envisageable d'utiliser des méthodes alternatives complémentaires.

2.2 Méthodes de réduction dimensionnelle

La réduction dimensionnelle permet de passer de données en hautes dimensions, beaucoup trop détaillées, à des données en dimensions principales, que l'on peut désormais interprétées. De nombreuses méthodes de réduction dimensionnelle existent telles que l'Analyse en Composante Principale (PCA pour Principal Component Analysis), l'Échelonnement MultiDimensionnel (MDS pour MultiDimensional Scaling) ou encore la Carte Auto Adaptative (SOM pour Self Organising Map). Toutes ces méthodes permettent de mettre en évidence les caractéristiques principales et sont utilisées dans des conditions bien précises. La technique MDS est préférée car elle permet d'analyser une métrique arbitraire, contrairement à la PCA, et autorise la projection d'éléments supplémentaires (voir chapitre 5.5), contrairement à la SOM. La MDS est disponible sous forme métrique (ou classique) ou non-métrique. La MDS métrique est préférée parce que la forme non-métrique ne prend en compte que l'information ordinale et est plus adaptée à la notion de rang. L'analyse MDS, comme la méthode PCA, possède un atout majeur qui est absent des arbres phylogénétiques lorsque qu'il faut comparer des arbres de différents génomes : c'est la possibilité de projeter des éléments supplémentaires. Cette fonctionnalité n'a jamais encore été appliquée dans le domaine de l'analyse de l'évolution des protéines.

2.3 Principe de la coévolution

L'hypothèse selon laquelle l'évolution des positions serait indépendante n'est pas compatible avec la mise en évidence du mécanisme de coévolution, impliquant certaines positions qui évoluent

de façon corrélée au cours de l'évolution.

La séquence en acides aminés d'une protéine, c'est à dire sa structure primaire, détermine les spécificités structurales qui sont à l'origine du repliement protéique. La structure en trois dimensions d'une protéine joue un rôle clé dans le maintien de ses fonctions biologiques. Ces deux contraintes forcent la protéine à limiter le taux de variabilité de sa séquence en acides aminés et en particulier au niveau de sites spécifiques, plus cruciaux que d'autres pour la préservation de la stabilité de la protéine et de ses fonctions.

Cependant, à cause de l'indisponibilité fréquente des structures protéiques correspondantes, la plupart des études sur l'évolution protéique se base uniquement sur les séquences linéaires. Les interprétations ne sont pas totalement complètes du fait que ces études ignorent la dimension spatiale et s'affranchissent des dépendances distantes qui peuvent exister. Les méthodes se basant sur la notion de coévolution constituent un outil prometteur pour révéler des sites d'acides aminés dépendants, répondant ainsi à la problématique du manque d'informations tridimensionnelles [4].

La coévolution est intimement liée à la notion d'interdépendance. Lorsque des mutations aléatoires affectent le génome d'un organisme vivant, un acide aminé d'un site fonctionnel peut être modifié. Cette mutation est susceptible d'affecter une ou plusieurs fonctions de la protéine. L'organisme subissant la mutation a de plus faibles chances de survivre à cause de protéines pourvues de propriétés altérées. Pour garder l'intégrité et la viabilité de la protéine, cette mutation peut être compensée par une mutation supplémentaire d'un résidu complémentaire du site fonctionnel mis en jeu : ce sont les mutations corrélées. Ces compensations se retrouvent souvent entre les acides aminés qui sont proches physiquement, c'est à dire sur les acides aminés d'une paire qui sont éloignés dans la structure primaire mais proches au niveau de la structure repliée. Ce mécanisme est étroitement lié à l'évolution des protéines. En effet, au cours de l'évolution, de nouvelles sous-familles apparaissent et font intervenir des changements de séquences, qui impliquent très probablement les mutations corrélées. L'analyse des mutations corrélées pourrait donner des indices clés sur les origines de la diversité d'une famille protéique.

Les analyses des mutations corrélées sont traditionnellement employées pour l'identification de contacts entre résidus à l'intérieur ou entre différentes chaînes protéiques et la compréhension des mécanismes mis en jeu lors de l'activation de certaines protéines. La première approche pour détecter des résidus coévoluant dans un AMS fut proposée en 1994 [4]. Depuis, de nombreuses

autres méthodes furent développées mais la précision de prédiction pour les contacts structuraux dans les protéines globulaires excède rarement les 20%, et cela quelle que soit la méthode [5].

En effet, la coévolution ne se limite pas qu'à la notion de contacts mais concerne un plus large panel de raisons biologiques. La coévolution qui est observée entre deux sites *i* et *j* peut être décomposée en plusieurs composantes sous-jacentes (**Figure 1**). Elle s'exprime sous la forme d'un modèle linéaire C_{ij} [7] :

$$C_{ij} = C_{phylogénie} + C_{structure} + C_{fonction} + C_{interactions} + C_{stochastique}$$

- le terme $C_{phylogénie}$ représente l'histoire évolutive qui est commune aux résidus des différents sites. On s'attend à ce que les substitutions de compensation qui affectent une séquence ancestrale se transmettent aux séquences de la descendance [8].

- les termes $C_{structure}$, $C_{fonction}$ et $C_{interactions}$ expliquent la coévolution provenant des contraintes structurales, fonctionnelles et d'interactions avec d'autres atomes ou d'autres partenaires [6]. Le repliement protéique n'autorise que certains remplacements d'acides aminés et qu'à certains sites particuliers. Par exemple, si un acide aminé volumineux (leucine) du cœur de la protéine est remplacé par un acide aminé plus petit (alanine), ce changement peut déstabiliser la structure de la protéine [9]. De plus, ces substitutions sont sous l'influence des contraintes fonctionnelles, tel que le site de fixation d'un ligand spécifique.

- le terme $C_{stochastique}$ se réfère au bruit de fond qui ne peut s'expliquer par les raisons principales et leurs interactions. Par exemple, les effets aléatoires d'un échantillonnage incomplet de séquences pour l'analyse.

Il faut noter que les différentes composantes ne sont pas indépendantes et peuvent être difficiles à distinguer. Les méthodes d'analyse des mutations corrélées (AMC) permettent d'étudier ce mécanisme de coévolution et s'efforcent d'en distinguer les différentes composantes.

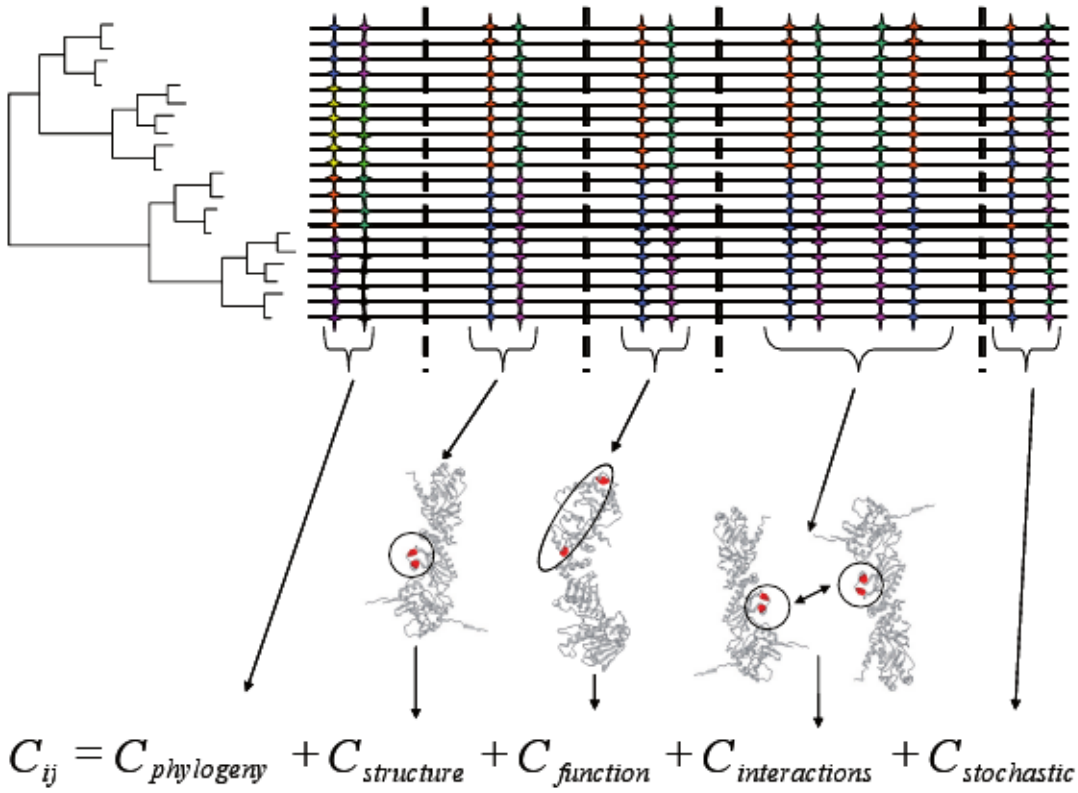

Figure 1 : Décomposition de la coévolution

La coévolution entre deux sites d'acides aminés peut être décomposée en coévolution phylogénétique $C_{phylogénie}$, structurale $C_{structure}$, fonctionnelle $C_{fonction}$, liée aux interactions entre atomes $C_{interactions}$ et stochastique $C_{stochastique}$. Les séquences de l'AMS sont représentées par des lignes horizontales et sont reliées par un arbre phylogénétique. Les étoiles de couleurs représentent les sites impliqués dans la coévolution. Les lignes verticales en pointillés séparent les composantes de la coévolution. Figure tirée de [6].

3. LES RECEPTEURS COUPLES AUX PROTEINES G

Tout d'abord, nous aborderons les caractéristiques principales des RCPG en lien avec leur rôle dans les pathologies humaines. Puis nous exposerons les études de classification et d'évolution qui ont été menées dans la littérature scientifique. Ensuite, nous présenterons les travaux sur les RCPG qui ont déjà été réalisés par l'équipe de bioinformatique.

3.1 Généralités sur les RCPG

Les RCPG constituent l'une des plus vastes familles de récepteurs protéiques transmembranaires présentes dans le corps humain [10]. Plusieurs centaines de RCPG humains différents ont été identifiés à ce jour [11]. Les RCPG montrent une conservation supérieure pour la structure tridimensionnelle que pour la séquence primaire. La diversité des RCPG participe à la difficulté de développer un système de classification globale [12]. Les RCPG sont localisés dans la membrane cytoplasmique et permettent la communication entre milieux intra et extracellulaires (**Figure 2**). Un signal qui provient de l'extérieur de la cellule est transmis par ces récepteurs pour déclencher toute une cascade de réactions intracellulaires. Ce signal est véhiculé par des ligands spécifiques d'un récepteur donné. Les différents ligands sont de tailles différentes et de natures extrêmement diverses. Ils peuvent être endogènes (peptides, lipides, nucléotides, ions et hormones) ou exogènes (photons, molécules odorantes et gustatives). Les ligands physiologiques sont activateurs. Leur liaison sur la région extracellulaire ou transmembranaire des RCPG engendre un changement de la structure tridimensionnelle du récepteur cible. La conséquence directe est le transfert du signal à des protéines intracellulaires, liées aux récepteurs, appelées protéines G (Protéines fixant le guanosine di ou triphosphate) qui vont activer des canaux ioniques transmembranaires et des enzymes intracellulaires (l'adénylate cyclase, la phospholipase C et des kinases, par exemple) et déclencher des réactions en cascade. Ces réactions chimiques gouvernent certains facteurs de transcription et par extension, le comportement cellulaire (prolifération cellulaire, par exemple) et à plus grande échelle, contribuent à la régulation de mécanismes physiologiques importants, tels que la perception sensorielle, la défense immunitaire, la communication cellulaire ou encore la neurotransmission. De par leur implication, à la fois, en tant que régulateurs de fonctions physiologiques et en tant que responsables de pathologies chez l'homme, ces protéines sont des cibles privilégiées pour développer des molécules médicamenteuses. Près de la moitié des médicaments actuels cible ces récepteurs [14] et certains de ces traitements sont retrouvés parmi les 100 produits pharmaceutiques les plus vendus au monde [15].

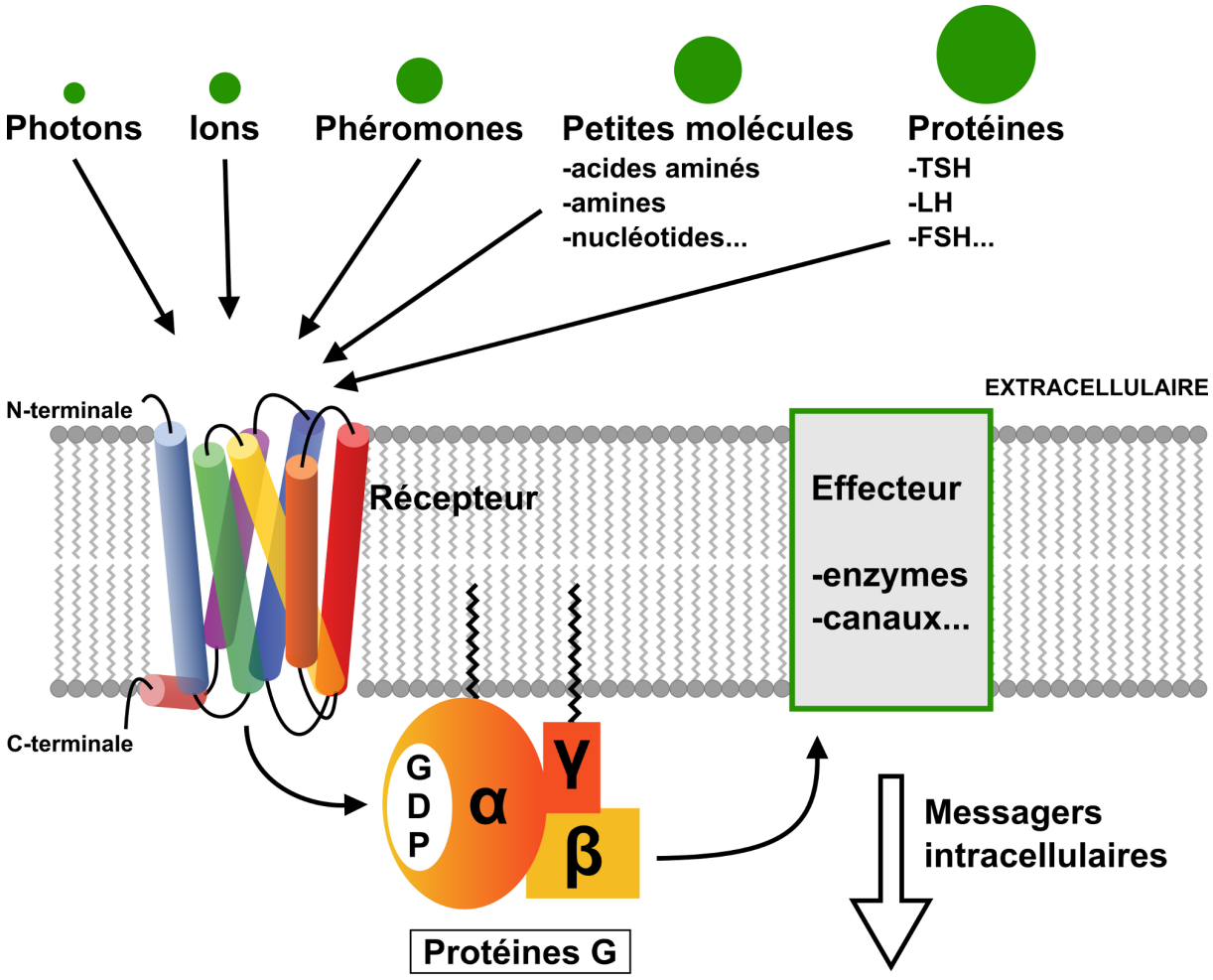

Figure 2 : Transduction du signal par les RCPG

Les ligands extracellulaires qui se fixent aux RCPG peuvent être de différentes natures : photons, ions (calcium), phéromones, petites molécules (acides aminés, amines, nucléotides...) et protéines (thyréotropine, hormone lutéinisante, hormone folliculo-stimulante...). Une cascade de réaction moléculaire est activée et implique les protéines G intracellulaires qui régulent l'activité de certains effecteurs (enzymes, canaux...). Des voies de signalisation sont stimulées et font intervenir des messagers intracellulaires pour déclencher certaines réponses cellulaires (prolifération, différenciation...). Figure inspirée de [13].

3.2 Structure et mécanisme d'activation

Les structures protéiques peuvent être résolues par diverses techniques : la diffraction par rayons X, la spectroscopie par résonance magnétique nucléaire (RMN), la microscopie électronique et l'association de plusieurs de ces techniques. D'après les statistiques actuelles de la Protein Data Bank (PDB), près de 90% des structures répertoriées sont résolues par rayons X et cette technique convient davantage aux protéines de grandes tailles, telles que les RCPG. Cependant, la disponibilité des structures des RCPG en haute résolution est encore rare. Cela est dû principalement aux difficultés expérimentales d'obtenir de grandes quantités de protéine suffisamment pure. De plus, les RCPG sont généralement instables dans les détergents [16].

Depuis longtemps, seule la structure en haute résolution de la rhodopsine bovine inactive était disponible [17]. La deuxième structure d'un RCPG a été résolue en octobre 2007 et concerne le récepteur β2 adrénergique [18]. Récemment, la résolution de nouvelles structures de RCPG s'est accélérée (**Tableau 1**) et les structures du récepteur β1 adrénergique [24], du récepteur de l'adénosine A2A [25] et, très récemment, du récepteur de chimiokines CXCR4 [26] ont été successivement résolues. Toutes ces structures possèdent certaines caractéristiques structurales qui sont communes aux RCPG (**Figure 3**). En effet, tous les RCPG possèdent une extrémité N-terminale extracellulaire, une structure de segments transmembranaires (TM) hydrophobes reliés par trois boucles extracellulaires (EC 1-3) et trois boucles intracellulaires (IC 1-3), ainsi qu'une extrémité C-terminale cytoplasmique [28]. Les EC peuvent être glycosylées et contiennent deux résidus cystéines conservés qui produisent des ponts disulfures pour stabiliser la structure protéique. Les régions les plus variables concernent l'extrémité N-terminale, la boucle intracellulaire IC3 et l'extrémité C-terminale. A l'inverse, les segments TM montrent une grande homologie entre les RCPG et les structures en basse et haute résolution montrent qu'ils sont formés de 7 hélices transmembranaires (HTM) α, de longueurs différentes et inclinées suivant différents angles, selon le plan de la membrane [29]. Ces hélices sont généralement notées HTM1 à HTM7 [30] et possèdent toutes un résidu d'ancrage extrêmement conservé parmi les RCPG de classe A : N pour l'HTM1, D pour l'HTM2, R pour l'HTM3, W pour l'HTM4, P pour les HTM5, 6 et 7. Ces acides aminés vont servir de point de repère pour chaque HTM et la position de chaque acide aminé pour chaque HTM est déterminée par le schéma de numérotation de Ballesteros-Weinstein [31,32]. Ce système est composé de deux chiffres, N1 et N2, séparés par un point : N1 pour le numéro d'HTM et N2 pour le numéro de la position par rapport au résidu de référence à qui est attribué la position 50.

Date	PDB	Espèce	Ligand	Résolution (Å)	Référence
Rhodopsine					
04/08/2000	1F88	*Bos taurus*	Rétinal	2,8	[17]
04/07/2001	1HZX	*Bos taurus*	Rétinal	2,8	[20]
24/06/2008	3CAP	*Bos taurus*	Sans ligand	2,9	[21]
06/05/2008	2ZIY	*Todarodes pacificus* (calmar)	Rétinal	3,7	[22]
13/05/2008	2Z73	*Todarodes pacificus* (calmar)	Rétinal	2,5	[23]
Récepteur β2 adrénergique					
30/10/2007	2RH1	*Homo sapiens*	Agoniste inverse partiel	2,4	[18]
Récepteur β1 adrénergique					
24/06/2008	2VT4	*Meleagris gallopavo* (dindon)	Antagoniste de haute affinité	2,7	[24]
Récepteur aux adénosines A2A					
14/10/2008	3EML	*Homo sapiens*	Antagoniste de haute affinité	2,6	[25]
Récepteur aux chimiokines CXCR4					
27/10/2010	3ODU	*Homo sapiens*	Petite molécule antagoniste	2,5	[26]

Tableau 1 : Structures résolues de RCPG de classe A

Le tableau affiche les étapes majeures dans la résolution des structures des RCPG de classe A. Les colonnes indiquent la date de sortie de la structure, le code d'identification PDB, l'espèce concernée et les caractéristiques principales de la structure, respectivement. Figure inspirée de [19].

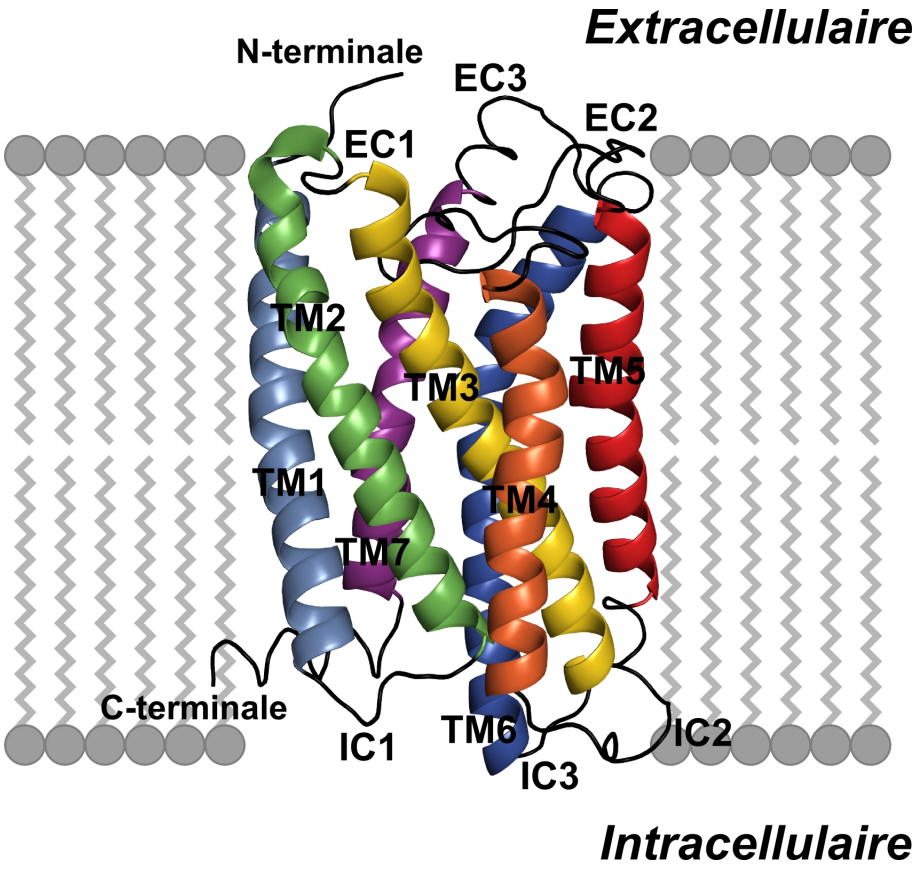

Figure 3 : Structure schématique des RCPG

Les hélices correspondent aux différentes HTM : bleu ciel pour l'HTM1, vert pour l'HTM2, jaune pour l'HTM3, orange pour l'HTM4, rouge pour l'HTM5, bleu pour l'HTM6 et magenta pour l'HTM7. Les termes EC1, EC2 et EC3 indiquent l'emplacement des boucles extracellulaires 1, 2 et 3, respectivement. Même chose pour IC1, IC2 et IC3 mais intracellulaires. Les termes N-terminale et C-terminale précisent les positions des 2 extrémités. Figure inspirée de [27].

La numérotation décroît vers l'extrémité N-terminale et croît vers l'extrémité C-terminale. Il est d'usage d'indiquer en amont de la notation le résidu auquel on veut faire référence. Par exemple, la position D2.50 indique l'acide aspartique de référence pour l'HTM2. Les résidus proline sont les résidus de référence de trois HTM (HTM5, 6, et 7) et, comme on le verra plus loin dans ce présent document, cette prépondérance n'est pas due au hasard.

La conservation relative de ces HTM est en contraste avec la diversité structurale des ligands naturels des RCPG [30]. La nature et la taille des ligands déterminent le site de fixation aux RCPG. Tandis que les petits agonistes organiques se fixent au niveau des segments TM, les hormones peptidiques et protéiques se lient souvent à l'extrémité N-terminale et aux boucles EC. L'activation du RCPG par la fixation de son ligand engendre un signal de transduction qui agit sur différents systèmes physiologiques de la cellule. Tout un schéma de régulation se met en place pour que le récepteur qui passe d'un état inactif à un état actif repasse ensuite à son état initial inactif (**Figure 4**). Le schéma de régulation se présente sous la forme d'un cycle qui est bien connu et qui se retrouve chez la majorité des RCPG. Il fait intervenir les protéines G qui se présentent sous la forme d'un hétérotrimére de trois sous-unités α, β et γ (Gα, Gβ et Gγ). Gα possède un site de fixation au guanosine triphosphate (GTP) et au guanosine disphosphate (GDP) et possède une activité GTPase intrinsèque [33]. Gβ et Gγ forment un dimère indissociable. Dans l'état inactif, Gα est fixé au complexe Gβγ et au GDP. Lorsque qu'une molécule agoniste se fixe sur le récepteur, le récepteur activé change de conformation et active le trimère Gαβγ qui échange le GDP par le GTP. Le dimère Gβγ se dissocie de la sous-unité α. Les protéines Gα et Gβγ ainsi activées se fixent à leur tour à de nombreux effecteurs pour transmettre le signal à différents types de messagers secondaires, tels que les canaux ioniques et des enzymes. Après la transduction du signal, l'activité GTPase de Gα rentre en jeu et convertit le GTP fixé en GDP et ainsi inactive la cascade des protéines G en re-associant la sous-unité α avec le dimère Gβγ. L'activité GTPase peut être régulée par des régulateurs de protéines G (RGS pour Regulator of G-protein Signaling) et de nombreux effecteurs. A ce stade, le GDP est de nouveau fixé à Gα et le cycle devient complet. Les RCPG jouent un véritable rôle d'interface entre le ligand et les protéines G et les HTM y jouent un rôle clé car les interactions entre hélices contribuent au repliement, à la stabilité, à la fixation du ligand et au changement conformationnel lors de l'activation des RCPG.

D'un point de vue structural, la fixation des différents ligands déclenche un ensemble de

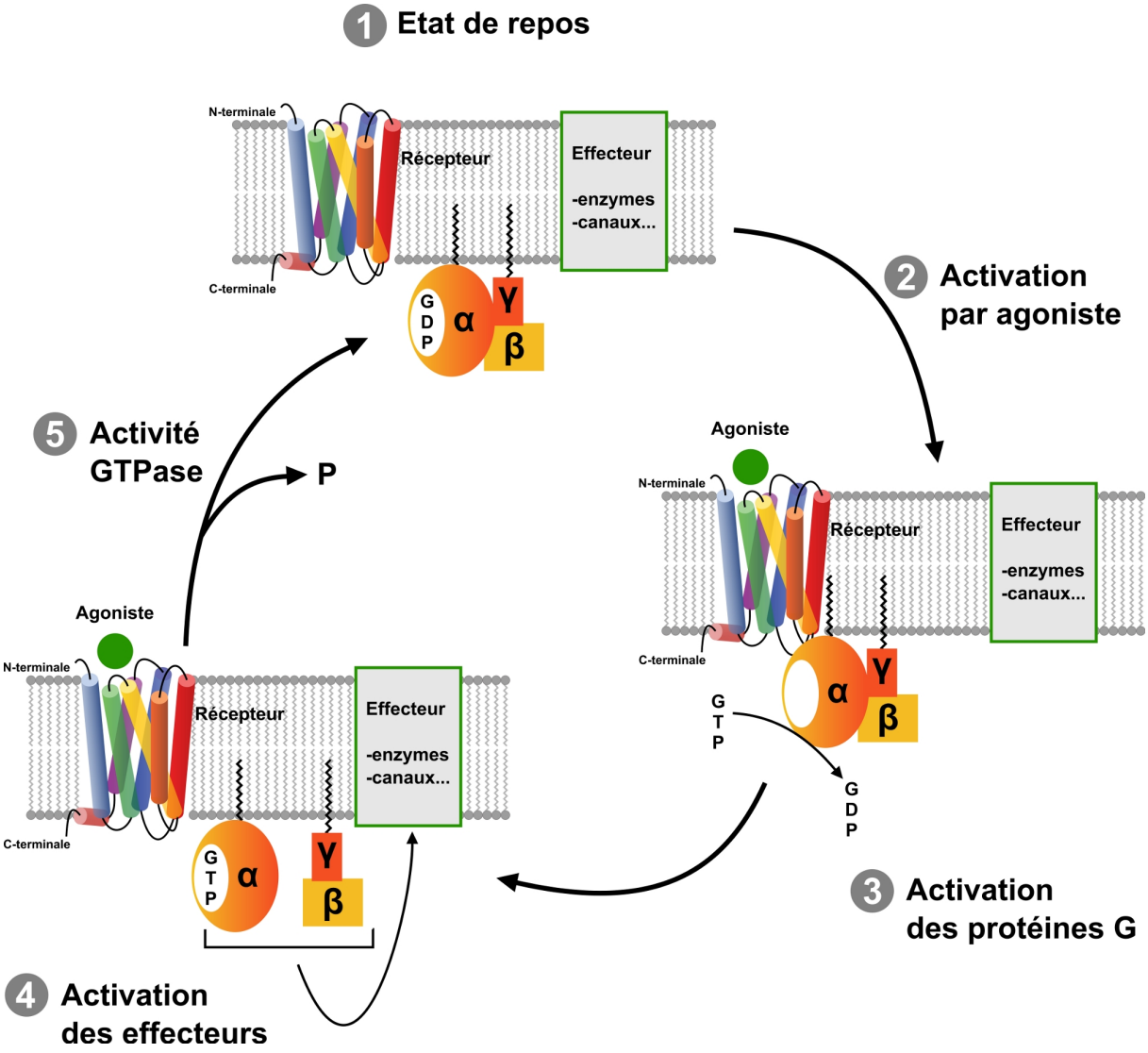

Figure 4 : Cycle d'activation des RCPG

Ce schéma est constitué de cinq étapes qui, ensemble, constituent un cycle. La fixation d'un agoniste sur les RCPG active les protéines G. Le GTP est remplacé par du GDP au niveau de la sous-unité α et les protéines G se dissocient. La sous-unité α et le dimère βγ activent certains effecteurs qui agissent sur la réponse cellulaire. L'activité GTPase de la sous-unité α permet de rétablir l'état de repos initial des RCPG. Figure inspirée de [33].

réarrangements conformationnels des segments TM proches des domaines de fixation des protéines G. La comparaison de la structure de la rhodopsine inactive [34] avec celle de l'opsine (sans ligand) [21] qui contient de nombreuses informations sur l'état actif présumé, montre que durant le processus d'activation des RCPG, le changement le plus important intervient au niveau de l'HTM6 dont la partie intracellulaire s'incline vers l'extérieur de 6-7 Å, l'HTM5 s'approche de l'HTM6 et le résidu R3.50 du motif DRY de l'HTM3 adopte une conformation étendue pointant vers le cœur de la protéine. Ces changements conformationnels perturbent l'interaction ionique entre le résidu R3.50 et les chaînes latérales chargées négativement des positions 3.49 et 6.30 et facilitent les interactions entre les résidus Y5.58 et K5.66 de l'HTM5, et R3.50 de l'HTM3 et E6.30 de l'HTM6, respectivement. Ces observations suggèrent fortement que l'HTM6 joue un rôle central dans le processus d'activation des RCPG [35].

Pendant longtemps, les RCPG ont longtemps été considérés que sous une forme monomérique. Depuis une dizaine d'années [36], il est reconnu que les RCPG peuvent interagir entre elles, et former ainsi des homo/hétérodimères ou des homo/hétéromultimères [37]. De telles interactions sont importantes pour la maturation protéique, la distribution des récepteurs à la surface cellulaire et la structure de ces complexes semble être centrale dans les mécanismes d'activation et de liaison aux protéines G [38]. Les HTM sont impliquées dans ces interactions et plus particulièrement les HTM4 à 6 dans le cadre des dimères [39]. Toutefois, les HTM impliquées semblent varier d'un récepteur à l'autre, ce qui pourrait s'expliquer par des schémas d'interactions avec des points de contact variés pour les différents RCPG [40]. Des protéines qui ne sont pas des RCPG peuvent être concernées, en tant que régulateurs, comme les β-arestines [41]. Ce nouvel aspect des RCPG est susceptible d'avoir un fort impact sur la découverte de médicaments et offre un champ d'applications élargi dans le cadre de cibles thérapeutiques.

Mieux comprendre les changements conformationnels complexes qui mènent à l'activation des RCPG est impératif pour décrypter le rôle de ces récepteurs dans les pathologies humaines [42]. L'interprétation de toutes ces informations peut être facilitée en les mettant en relation avec la classification des RCPG.

3.3 Classification

Certains systèmes de classification se basent sur l'emplacement de fixation du ligand, tandis que

d'autres utilisent simultanément les caractéristiques physiologiques et structurales [13]. Cependant, en 1994, l'une des premières classifications utilisée se base sur l'identité de séquences et sépare les RCPG en classes A à F [11], dont les trois premières sont présentes chez les humains. Ces classes sont établies à partir des RCPG connus des invertébrés et des vertébrés :

- la classe A comprend énormément de récepteurs, environ 300 non olfactifs et 400 olfactifs. Elle est de loin la famille la plus étudiée car elle comprend la rhodopsine et les récepteurs β-adrénergiques, et représente jusqu'à 90% des RCPG humains. L'homologie globale entre les récepteurs de cette classe est faible et limitée à un petit nombre de résidus hautement conservés, primordiaux pour l'intégrité structurale et fonctionnelle.

- la classe B contient environ 20 récepteurs spécifiques d'hormones peptidiques et des neuropeptides, tels que le peptide intestinal vasoactif (VIP), la calcitonine, la parathormone (PTH) et le glucagon. Excepté le pont disulfure entre boucles EC2 et EC3, la classe B ne montre pas d'autres caractéristiques structurales communes avec la classe A. De plus, le motif DRY est absent et les prolines conservées sont différentes de celles de la classe A. La particularité la plus apparente de la classe B se situe au niveau de l'extrémité N-terminale car elle est plus longue (environ 100 résidus) et contient de nombreuses cystéines formant un réseau de ponts disulfures.

- la classe C est composée des récepteurs métabotropes des neurotransmetteurs et caractérisée par une extrémité N-terminale extrêmement longue (500-600 résidus). Cette classe contient également le pont disulfure au niveau des boucles EC mais ne partagent aucun résidu conservé avec les classes B et C.

- les classes D et E sont rattachées aux récepteurs des phéromones de la levure. Les récepteurs à l'adénosine monophosphate cyclique (AMPc) constituent la classe F.

3.3.1 Les différentes familles de récepteurs humains

Du fait de la diversité et de la complexité de classer rigoureusement les RCPG, les chercheurs se

sont intéressés à la classification phylogénétique. La classification généralement utilisée est la classification GRAFS [43], acronyme des noms des 5 familles (Glutamate, Rhodopsin, Adhesion, Frizzled/taste2 et Secretin) et qui ne concerne que les RCPG humains. La classification GRAFS (**Figure 5**) a été réalisée par Neighbour Joining (NJ) et maximum de parcimonie (MP) et a permis de classer précisément les récepteurs humains :

- la famille Glutamate contient 15 membres : 8 récepteurs métabotropes du glutamate (GRM), 2 récepteurs de l'acide γ-aminobutyrique (GABA), 1 récepteur du calcium (CASR) et 5 récepteurs qui semblent être des récepteurs du goût (TAS1). Cette famille correspond à la classe C du système de classification A-F.

- la famille Adhesion contient 24 membres. Le nom de la famille est lié au fait que l'extrémité N-terminale de ces récepteurs contient des motifs susceptibles de participer à l'adhésion cellulaire [44]. Certains récepteurs proches des sécrétines ont auparavant été placés dans cette famille mais ils forment maintenant une famille à part.

- la famille Frizzled/taste2 contient 24 membres qui peuvent se diviser en deux groupes : les récepteurs Frizzled et les récepteurs du goût TAS2. Ces deux groupes se retrouvent dans la même famille car ils affichent des caractéristiques communes, qui ne sont pas retrouvées parmi les autres RCPG. Les récepteurs Frizzled contrôlent la différentiation, la prolifération et la polarité cellulaire durant le développement métazoaire et leur nom provient du caractère « bouclé » de leurs ligands.

- la famille Secretin contient 15 membres et correspond à la classe B du système de classification A-F. Le premier récepteur qui a été cloné est un récepteur de sécrétine, d'où le nom de la famille. Comme membres, on peut citer le récepteur de la calcitonine (CALCR) et le récepteur du glucagon (GLPR).

- la famille Rhodopsin est à part car elle contient environ 700 membres (300 récepteurs non-olfactifs et 400 récepteurs olfactifs) et correspond à la classe A du système A-F. Cette famille affiche certaines particularités : un motif NSxxNPxxY dans l'HTM7, le motif

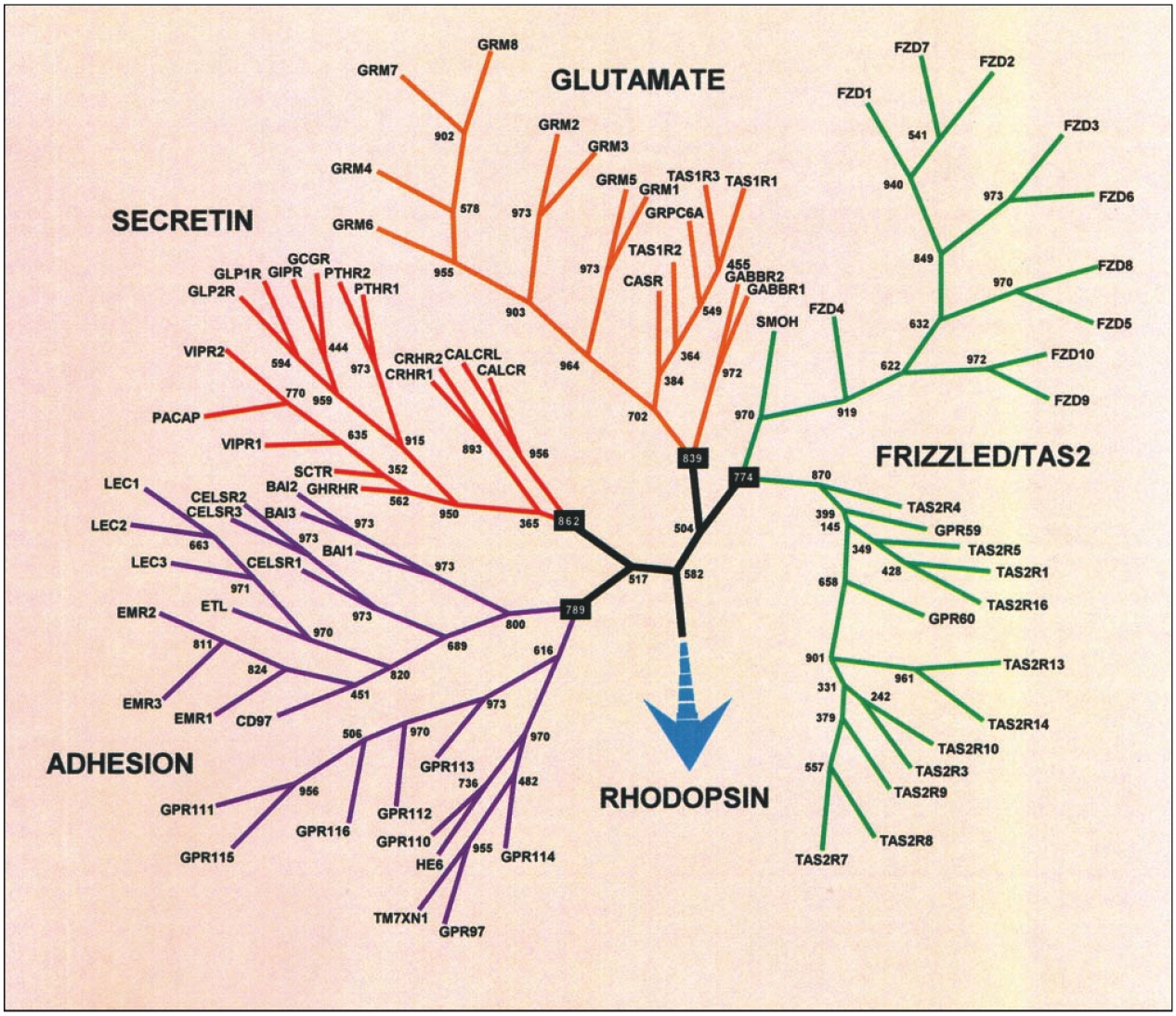

Figure 5 : Relations phylogénétiques entre les RCPG du génome humain

L'arbre a été calculé par maximum de parcimonie sur 1000 copies du jeu de données. Les numéros présents à l'intersection des branches indiquent les valeurs de bootstrap. Les branches principales en violet, rouge, orange et vert représentent les familles Adhesion, Secretin, Glutamate et Frizzled/taste2, respectivement. La famille Rhodopsin étant très diversifiée, elle est amplement détaillée sur la figure 6. Figure tirée de [43].

DRY ou D(E)-R-Y(F) entre l'HTM3 et la boucle IC2. Seules quelques récepteurs n'affichent pas ces motifs mais peuvent être affiliés à la famille Rhodopsin par d'autres marqueurs. La classe A est très étudiée parce qu'elle représente un intérêt pharmaceutique particulier pour les récepteurs des amines. Le nombre de médicaments pour les autres RCPG augmente, en particulier pour les récepteurs fixant les peptides. Le potentiel thérapeutique des autres sous-familles de classe A n'a pas encore été totalement exploité. Beaucoup de ces récepteurs restent orphelins, c'est à dire sans ligand connu.

La classe A affiche une grande diversité de récepteurs et nécessite une analyse détaillée du fait du faible degré de similarité entre ses membres (**Figure 6**). Les récepteurs de la classe A non-olfactifs peuvent être divisées en 12 sous-familles distinctes selon la classification GRAFS.

3.3.2 Les sous-familles de la classe A

La sous-famille des récepteurs des prostaglandines (PTGR) : avec 15 membres, elle contient 8 récepteurs des prostaglandines (PTGR) et 7 récepteurs orphelins. Les PTGR affichent entre 19 et 41% d'identité et partagent les motifs IxDPW et LxxTDxxG dans l'HTM7 et 1, respectivement.

La sous-famille des récepteurs des amines (AMIN) : avec 40 membres, elle compte les récepteurs de la sérotonine (HTR), de la dopamine (DRD), muscariniques (CHRM), des histamines (HRH), adrénergiques (ADR), des traces d'amines (TAR) et quelques récepteurs orphelins. Tous les ligands connus de cette branche ont une structure sous la forme d'une petite molécule amine avec un seul noyau aromatique. Le degré de conservation des séquences varie selon le type de récepteurs.

La sous-famille des récepteurs des opsines (OPN) : avec 9 membres, elle comprend les récepteurs visuels dans les bâtonnets (RHO), dans les trois types de cônes (OPN1SW, OPN1LW, OPN1MW), la péropsine (RRH), l'encéphalopsine (OPN3), la mélanopsine (OPN4), et le récepteur lié au rétinal (RGR). Les opsines sont les seuls RCPG à réagir à l'excitation lumineuse et font intervenir un changement de conformation du rétinal lors de l'activation.

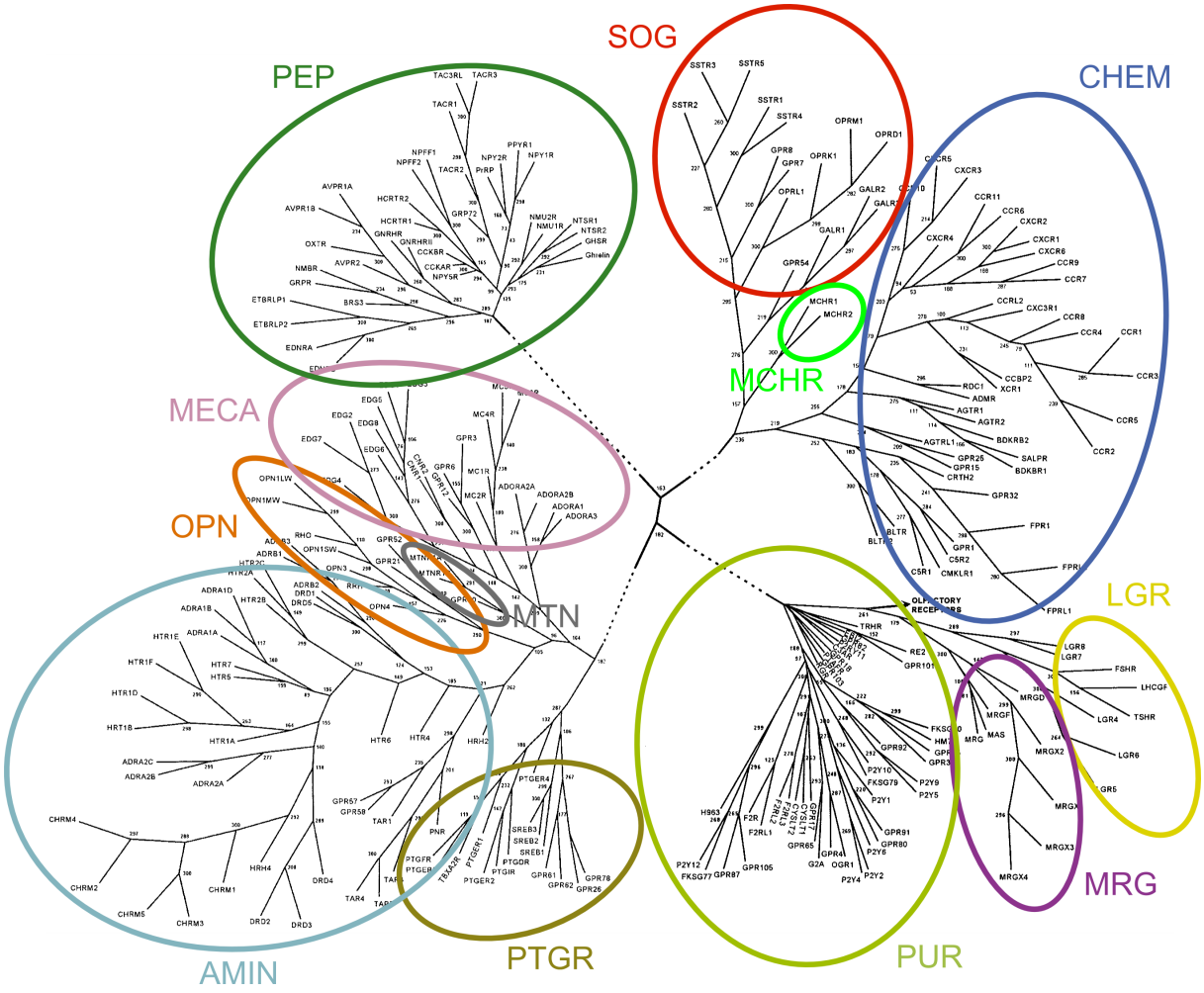

Figure 6 : Relations phylogénétiques entre les RCPG de classe A du génome humain

L'arbre a été calculé par maximum de parcimonie sur 300 copies du jeu de données. Les numéros présents à l'intersection des branches indiquent les valeurs de bootstrap. Les récepteurs OLF ne sont pas visibles et sont représentés par une flèche entre les sous-familles PUR, MRG et LGR. Figure modifiée à partir de [43].

La sous-famille des récepteurs des mélatonines (MTN) : avec 3 membres, qui peuvent être divisés en deux groupes : les récepteurs des mélatonines d'une part (MTNR1A, MTNR1B) et un récepteur orphelin d'une autre part (GPR50).

La sous-famille des mélanocortines-endoglines-cannabinoïdes-adénosines (MECA) : avec 22 membres, elle est représentée par les récepteurs des mélanocortines (MCR), de la différenciation endothéliale (EDGR), des cannabinoïdes (CNR), de fixation à l'adénosine (ADORA). Trois récepteurs orphelins appartiennent à cette sous-famille (GPR-3, -6 et -12). Il est intéressant de noter que ces récepteurs se fixent à des ligands de natures différentes : l'hormone de stimulation des mélanocytes (MCR), l'acide lysophosphatidique (EDGR), l'anandamide (CNR) et l'adénosine. Les récepteurs orphelins sont identiques à 55% entre eux et à peine identiques de 25% aux MCR.

La sous-famille des récepteurs des peptides (PEP) : avec 36 membres, elle inclut les récepteurs de l'hypocrétine (HCRTR), les récepteurs du neuropeptide FF (NPFF), les récepteurs de la tachykinine (TACR), les récepteurs de la cholécystokinine (CCK), les récepteurs du neuropeptide Y (NPYR), les récepteurs liés aux endothélines (EDNR et ETBRLP1/2), le récepteur du peptide de libération de la gastrine (GRPR), le récepteur de la neuromédine B (NMBR), le récepteur de l'uterinbombésine (BRS3), les récepteurs de la neurotensine (NTSR), le récepteur de l'hormone de croissance sécrétagogue (GHSR), les récepteurs de la neuromédine (NMUR), le récepteur de l'hormone de libération de la thyrotropine (TRHR), le récepteur de la ghréline (GHSR), les récepteurs de la vasopréssine (AVPR), les récepteurs de l'hormone de libération de la gonadotropine (GNRHR) , le récepteur de l'oxytocine (OXTR) et 1 récepteur orphelin.

La sous-famille des récepteurs des somatostatine/opioïde/galanine (SOG) : avec 15 membres, elle contient donc les récepteurs de la somatostatine (SSTR), des opioïdes (OPR), de la galanine (GALR) et de la kisspeptide (GPR54). Les récepteurs GPR7 et GPR8 se lient au neuropeptide W. Les ligands connus de cette sous-famille sont tous peptidiques mais ils ne partagent pas de similarités structurales.

La sous-famille des récepteurs de l'hormone de concentration de mélanine (MCHR) : avec

seulement 2 membres, le ligand est un neuropeptide cyclique de 19 acides aminés qui est impliqué dans la régulation du comportement alimentaire.

La sous-famille des récepteurs des chimiokines (CHEM) : avec 42 membres, elle inclut les récepteurs des chimiokines classiques (CCR et CXCR), de l'angiotensine (AGTR) et de la bradykinine (BDKR), des récepteurs de peptides vasoactifs (apéline, bradykinine) et des récepteurs orphelins. La plupart des ligands sont peptidiques (chimiokine, angiotensine, apéline et bradykinine).

La sous-famille des récepteurs liés à MAS (MRG) : avec 8 membres, elle contient le récepteur à l'oncogène MAS1 (MAS) et les récepteurs reliés à MAS (MRG et MRGX). Les MRGX possèdent une haute identité de séquences (environ 65%).

La sous-famille des récepteurs aux glycoprotéines (LGR) : avec 8 membres, elle est constituée des récepteurs des hormones glycoprotéiques classiques (FSHR, TSHR et LHCGR) et des répétitions riches en leucine (LGR). L'identité de séquences au sein des groupes est élevée (autour de 50%) mais faible entre les groupes (environ 20%).

La sous-famille des récepteurs purinergiques (PUR) : avec 42 membres, elle est liée aux récepteurs de peptides formylés (FPR), de nucléotides (P2Y) et un nombre non négligeable de récepteurs orphelins. Les ligands connus comprennent les nucléotides extracellulaires, les leucotriènes et les trombines.

Le laboratoire de bioinformatique a retrouvé les mêmes classifications des RCPG de classe A humains en sous-familles en utilisant également la technologie des arbres phylogénétiques par NJ [45]. Cependant, quelques différences sont observées :

- les récepteurs de la galanine de la sous-famille SOG sont rattachés à la sous-famille PEP. La sous-famille SOG devient la sous-famille SO.

- les récepteurs de la sous-famille MCHR sont regroupés avec la sous-famille SO.

- les récepteurs qui ne sont affiliés clairement à aucune sous-famille sont regroupés sous le dénominateur « non-classifiés » (UC).

Ces différences soulèvent la problématique de l'évolution des RCPG et sa résolution permettrait de s'assurer de l'affiliation de tel récepteur à telle sous-famille.

3.4 Évolution de la classe A

Les RCPG peuvent se retrouver dans presque tous les organismes eucaryotes, incluant aussi bien les insectes [46] que les plantes [47], indiquant vraisemblablement une origine ancienne de ces récepteurs. Il existe également une protéine à 7 HTM sensible à la lumière dans les bactéries, la rhodopsine bactérienne, mais on ne sait pas actuellement si cette protéine a une origine commune avec les RCPG des eucaryotes, à cause de l'absence des protéines G dans la transduction et le manque d'homologie avec les RCPG [48]. Déterminer les relations entre RCPG d'espèces relativement distantes et révéler l'origine et l'expression de chacun des groupes aideraient grandement à lever le voile sur l'origine des RCPG humains. Suite au développement de la classification GRAFS, l'équipe de Fredriksson a réalisé une analyse évolutive des RCPG (**Figure 7**). Pour cela, les RCPG de 12 génomes ont été sélectionnés : 2 levures (*Saccharomyces cerevisiae* et *Schizosaccharomyces pombe*), 1 protozoaire (*Plasmodium falciparum*), 2 plantes (*Arabidopsis thaliana* et *Oryza sativa*), 1 nématode (*Caenorhabditis elegans*), 2 insectes (*Anopheles gambiae* et *Drosophila melanogaster*), 1 ascidie (*Ciona intestinalis*), 2 poissons (*Danio rerio* et *Takifugu rubripes*) et 2 mammifères (*Mus musculus* et *Homo sapiens*). Ce jeu de données va permettre de suivre l'évolution des récepteurs, d'observer l'apparition de nouvelles sous-familles et d'analyser les différences au niveau des séquences qui surviennent au cours des importantes étapes évolutives, par exemple : l'apparition des chordés avec *Ciona intestinalis* et des vertébrés avec *Danio rerio*.

Les données concernant les RCPG de classe A humains, d'environ 700 membres, ont été reparties en 13 groupes distincts qui correspondent aux différentes sous-familles détaillées précédemment, plus les récepteurs olfactifs (OLF). Si des récepteurs ne répondent pas aux critères, ils sont groupés dans la catégorie «non-classifiés» (UC). Ces récepteurs humains ont ensuite été confrontés aux

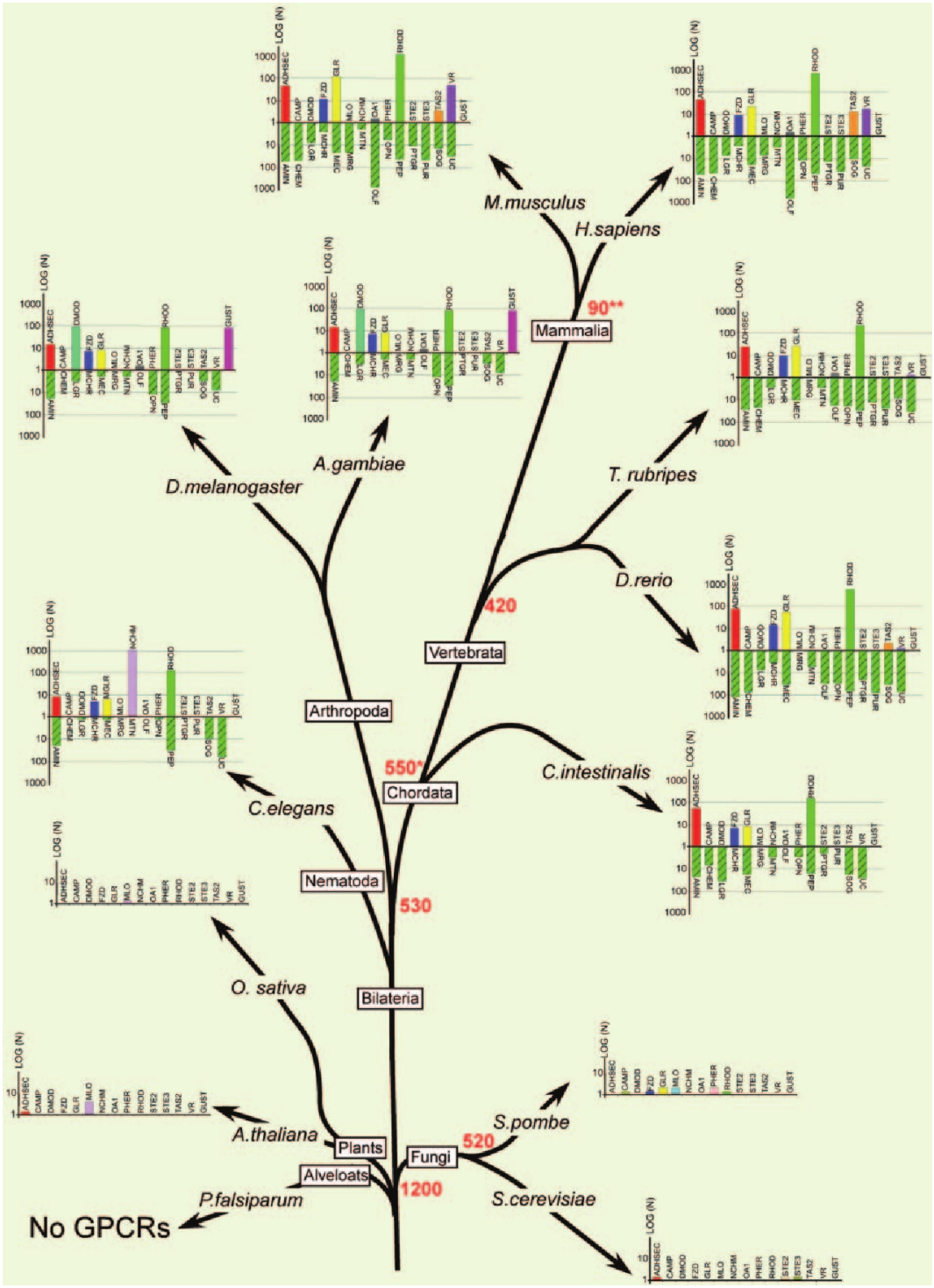

Figure 7 : Arbre d'évolution des RCPG pour différents génomes

Pour chaque espèce, un graphique des RCPG est affiché, avec au-dessus de l'axe des abscisses le logarithme du nombre de RCPG des classes principales et en-dessous celui des RCPG des différentes sous-familles de classe A. Les nombres en rouge indiquent, en millions d'années, à quelle période la séparation du nœud s'est effectuée. Figure tirée de [49].

RCPG de classe A des autres génomes pour attribuer une sous-famille à chaque récepteur.

Les RCPG de classe A se retrouvent dans toutes les espèces bilatérales. Certaines sous-familles sont présentes dans toutes les espèces (par exemple PEP, AMIN et SOG) alors que d'autres apparaissent chez des espèces plus proches de l'homme. C'est le cas par exemple des sous-familles MEC, CHEM et LGR qui apparaissent chez les chordés, des sous-familles PUR et OLF qui apparaissent chez les vertébrés et de la sous-famille MRG qui apparaît chez les mammifères. De plus, certaines sous-familles se développent particulièrement dans des espèces spécifiques. C'est le cas de la sous-famille LGR chez *C. intestinalis*. Les récepteurs AMIN et PEP sont fortement représentés dans toutes ces espèces. Il paraît clair que certaines sous-familles sont présentes uniquement chez les espèces proches des humains alors que d'autres sont présentes chez des espèces animales très éloignées des vertébrés [50].

3.5 Application des méthodes de corrélation aux RCPG

L'application des méthodes de corrélation de séquences aux RCPG de classe A est relativement récente. Les travaux de Gouldson [51] représentent la première application d'envergure des méthodes d'AMC sur les RCPG de classe A. La méthode utilisée est relativement proche [52] de la méthode MCBASC de par son algorithme et l'utilisation de la matrice de scores de McLachlan (voir chapitre 5.8.4). Comme la classification GRAFS n'avait pas encore été découverte, la méthode d'AMC n'avait été appliquée qu'à un jeu de données partiel, à savoir les récepteurs des chimiokines (CHEM), neurokinines (PEP), opioïdes (SO), somatostatines (SO) et TSH (LGR). Malgré tout, cette étude avait montré qu'une proportion importante des résidus corrélés est localisée au niveau des régions externes des hélices, orientées vers la partie lipidique de la membrane cellulaire. Cette constatation suggère que les résidus corrélés seraient impliqués dans les interactions entre protéines et interviendraient dans la formation de dimères. L'application sur un plus large jeu de données [53] de RCPG a permis d'identifier des réseaux de mutations corrélées constitués de positions impliquées dans les sites fonctionnels participant à la liaison du ligand, à l'activation du récepteur et au couplage covalent avec les protéines G. Également, les positions corrélées seraient plus vulnérables aux mutations ponctuelles associées à certaines pathologies, comparées aux autres positions [54]. L'entropie est une notion totalement primordiale pour révéler les résidus clés de la région

transmembranaire des RCPG de classe A [55]. Récemment, une méthode mettant à profit l'information mutuelle (voir chapitre 5.8.2) et la théorie des graphes a montré que les positions clés des RCPG de classe A sur les structures cristallines de la rhodopsine et des récepteurs β-adrénergiques encadrent de manière très proche les ligands co-cristallisés [56]. Cette observation est corroborée par les travaux de Surgand [57] qui en construisant des arbres phylogénétiques des RCPG, uniquement sur les positions clés proches du site du ligand, retrouvent les sous-familles classiques de Fredriksson. Très récemment, une étude a montré que les réseaux de 2 à 5 résidus seraient impliquées dans le repliement alors que les réseaux de plus grandes distances sont dus à la nature dynamique des RCPG, qui n'est pas visible directement sur une structure cristalline figée [58]. L'application des méthodes d'AMC peut parfois concerner d'autres domaines [59]. Toutes ces études mènent à penser que les méthodes d'AMC peuvent aider grandement à déceler les résidus cruciaux pour les RCPG, au-delà de leur proximité spatiale.

3.6 Travaux de l'équipe de bioinformatique

Les travaux de l'équipe de bioinformatique ont permis de relier l'évolution d'un motif structural de cassures d'hélices à l'évolution des RCPG de classe A [45]. Cette étude a été menée sur un jeu de données légèrement différent de celui de Fredriksson, à savoir sur les RCPG de class A non-olfactifs des espèces de *C. elegans*, *D. melanogaster*, *C. intestinalis*, *D. rerio* et *H. sapiens*. L'évolution du motif proline dans l'HTM2 a été plus spécifiquement étudiée, vu son rôle dans la structure de cette hélice (**Figure 8**).

Le motif proline de l'HTM2 peut être présent aux positions 2.58, 2.59 et 2.60. Concernant le motif P2.58, les espèces les plus récentes montrent une augmentation du pourcentage de récepteurs ayant ce motif. Le pourcentage est de 7% chez *C. elegans* et de 40% chez l'homme. Le nombre de récepteurs avec une proline à la position 2.59 ou sans proline est variable suivant les espèces et ne montre pas une tendance évolutive.

L'analyse phylogénétique, basée sur le NJ, montre que le motif P2.58 n'est conservé que dans certaines branches. Chez *C. elegans*, le motif n'est présent que dans une branche liée à la sous-famille SO à hauteur de 73%. Pour *D. melanogaster*, l'observation est particulière. Les récepteurs P2.58 correspondent à 2 récepteurs SO, 4 récepteurs OPN et 2 récepteurs qui appartiennent aux sous-familles PEP et AMIN, respectivement. Chez *C. intestinalis*, les récepteurs P2.58 sont

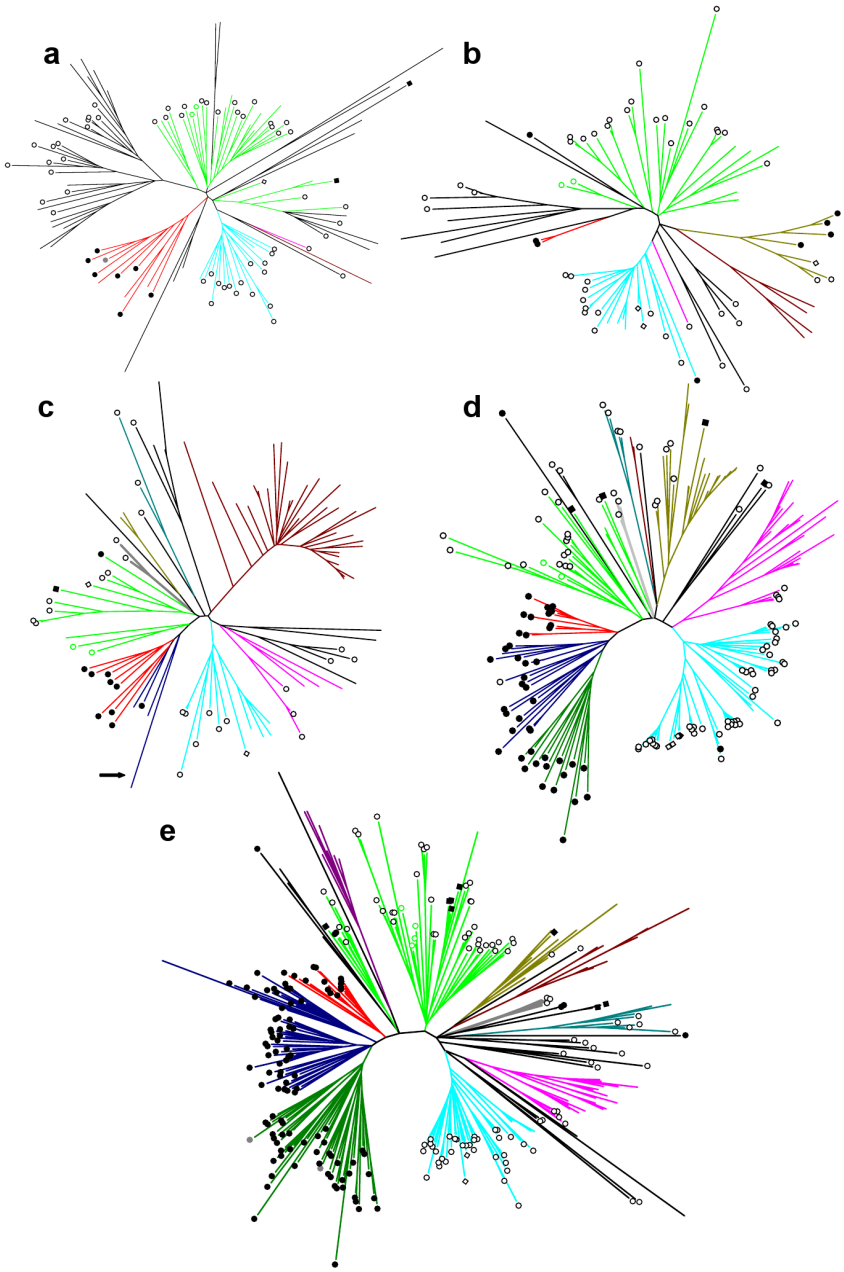

Figure 8 : Motifs proline sur les arbres phylogénétiques des RCPG de classe A

Les arbres ont été élaborés par NJ avec 500 copies et concernent les génomes suivants : *C. elegans* (a), *D. melanogaster* (b), *C. intestinalis* (c), *D. rerio* (d) et *H. sapiens* (e). Les symboles aux extrémités correspondent aux différents motifs proline rencontrés : P2.58 (cercle noir), P2.59 (cercle blanc), P2.60 (carré noir), P2.58P2.59 (cercle gris), P2.59P2.60 (diamant blanc). Les sous-familles sont visibles par un code couleur : PUR (vert foncé), CHEM (bleu), SO (rouge), PEP (vert clair), OPN (olive), LGR (marron), MECA (magenta), AMIN (cyan), MTN (gris), PTGR (cyan foncé), MRG (violet) et UC (noir). Figure tirée de [45].

regroupés dans la même branche et ses membres peuvent être assignés aussi bien aux récepteurs SO qu'aux récepteurs CHEM. Chez *D. rerio* et *H. sapiens*, excepté quelques cas, les récepteurs P2.58 se composent des récepteurs SO, CHEM et PUR. Concernant les autres motifs, les récepteurs AMIN et PEP privilégient une proline à la position 2.59. Cependant, quelques récepteurs ne possèdent aucune proline, comme les récepteurs à l'acétylcholine des chordés. Exceptionnellement, une proline en 2.58 ou en 2.60 peut être présente pour ces deux sous-familles. La sous-famille LGR est caractérisée par l'absence de proline pour l'ensemble des espèces étudiées. Les MECA de *C. elegans* à *D. melanogaster* sont proches des récepteurs des adénosines et possèdent un résidu proline en 2.59. Les autres membres des MECA ne possèdent pas de proline dans l'HTM2 et apparaissent avec les chordés. Tous les récepteurs de la sous-famille PTGR possèdent une proline à la position 2.59 chez *C. intestinalis* et *D. rerio* mais pas pour *H. sapiens*. Les prolines à la position 2.60 sont rares et se retrouvent dans quelques PEP et OPN. Des doublets P2.58P2.59 et P2.59P2.60 sont observés pour les récepteurs possédant une proline aux positions 2.58 et 2.59, respectivement. Il est à noter que la sous-famille OPN est spéciale car elle inclut des récepteurs avec des motifs P2.58, P2.59, P2.60 et sans proline. La plupart des récepteurs OPN des vertébrés ne possède pas de proline alors que celle des invertébrés en possède une, mais avec un motif variable. Par exemple, le motif P2.58 est spécifique des insectes et n'est pas observé chez les autres espèces d'invertébrés. Le motif proline doit être placé dans le cadre de contraintes structurales et l'étude exhaustive des structures des cassures d'hélices par l'équipe de bioinformatique a révélé que c'est un mécanisme d'indel (insertion/délétion) qui est l'origine de deux types de structures : la structure en renflement π des récepteurs P2.59 et P2.60 est reliée à la structure en coude des récepteurs P2.58 par un indel. Le motif proline de l'HTM2 peut être utilisé comme un véritable marqueur évolutif et permettre de suivre l'évolution des RCPG de classe A. Deux évènements majeurs de mutations surviennent au cours de l'évolution. Le premier indel se passe en amont de l'évolution des RCPG, au niveau de l'ancêtre bilatéral, avant la divergence entre protostomes et deutérostomes. Cet indel conduit à la séparation entre les SO P2.58 et les PEP P2.59. Le deuxième indel concerne les OPN d'insectes et correspond à une délétion. Les sous-familles avec une proline en 2.59 ou sans proline se développent précocement, alors que les récepteurs P2.60 demeurent marginaux. Les récepteurs P2.58 suivent une évolution rapide chez les vertébrés avec le développement des sous-familles CHEM et PUR à partir des SO. Ces découvertes doivent être mises en relation avec les récepteurs d'AT1 et AT2 qui sont étudiés par le laboratoire d'accueil et qui affichent une proline à la position 2.58.

4. OBJECTIFS DETAILLES

Pour rappel, l'objectif a consisté à poursuivre l'analyse des différentes sous-familles des RCPG de classe A pour en déterminer leurs caractéristiques et les relier aux mécanismes d'activation par analyse de séquences. Cette perspective ardue nous a amené à diviser le travail en différents sous-objectifs :

- visualiser la répartition des sous-familles des RCPG pour découvrir les relations qui existent entre elles.

  Les travaux de Fredriksson (2003) ont permis d'établir une classification des RCPG humains, et plus précisément ceux de classe A, sous la forme de différentes sous-familles [43]. Cependant, le travail de phylogénie d'une famille aussi dense que celle des RCPG est loin d'être facile. De plus, les arbres phylogénétiques impliquent forcement l'hypothèse selon laquelle les récepteurs s'ordonneraient suivant une structure hiérarchique bifurquée. Nous avons choisi d'utiliser une méthode alternative pour répondre à ces deux problématiques. Notre choix s'est porté sur la réduction dimensionnelle et plus précisément sur la méthode statistique multidimensional scaling (MDS). Cette méthode permet de s'affranchir de toute hypothèse de départ et demande un temps de calcul beaucoup plus réduit. Suite à des résultats prometteurs, à savoir le fait que les RCPG de classe A humains s'organiseraient en groupes, comprenant plusieurs sous-familles différentes déjà reconnues, nous avons continué dans cette voie pour répondre au second sous-objectif.

- étudier l'évolution des RCPG en comparant différentes espèces par rapport à l'espèce humaine.

  Bien que les travaux de Fredriksson (2005) aient permis de suivre l'évolution des RCPG à travers différentes espèces [49], les relations qui existent entre les différentes sous-familles des RCPG de classe A n'ont pas été clairement établies. Cette lacune semble rester inhérente aux arbres phylogénétiques [60]. Développer une méthode visuelle permettant de suivre la position des récepteurs d'une espèce par rapport à une autre aiderait grandement à la compréhension des origines des différentes sous-familles. Nous avons voulu valoriser la méthode MDS en la complétant d'une méthode qui permettrait de répondre à ce besoin. Le Pr Abdi, de l'Université du Texas a développé une méthode de projection d'éléments

supplémentaires dans le cadre de la MDS [79]. Notre équipe l'a appliquée dans le cadre bioinformatique pour permettre de projeter des RCPG de classe A des espèces plus ou moins lointaines sur l'espace des RCPG humains afin de suivre l'évolution des récepteurs.

- caractériser chaque groupe pour faire le lien entre particularités séquentielles et spécialisation des récepteurs.

  Les travaux précédents du laboratoire avaient mis en évidence le rôle clé de certains motifs dans les hélices en tant que marqueurs évolutifs. Nous avons développé différentes méthodes pour révéler les marqueurs liés à chacun des groupes, obtenus suite à la MDS. Elles sont toujours calculées suivant une comparaison d'un groupe et de son complémentaire. Ces méthodes peuvent être séparées en deux catégories. Premièrement, la différence d'entropies permet de mettre en évidence les différences de conservation entre groupes. Deuxièmement, des méthodes de corrélation vont déceler les motifs spécifiques pour chacun des groupes. Il faut bien distinguer ces deux principes car des résidus peuvent être spécifiques d'un groupe mais pas forcement y être totalement conservés. Suite à la découverte de résidus bien caractéristiques pour chacun des groupes, nous avons émis l'hypothèse d'un schéma évolutif des RCPG de classe A.

- déterminer les positions impliquées dans la covariation des résidus.

  L'activation des RCPG met en jeu des mouvements complexes de structures secondaires impliquées dans des réseaux d'interactions spécifiques. Nous avons choisi d'utiliser les méthodes d'analyse des mutations corrélées (AMC) pour mieux comprendre les résidus et les mécanismes mis en jeu lors de l'activation des RCPG. Les travaux de Ye (2006) ont montré que les résidus jouant des rôles clés sont extrêmement liés à la notion d'entropie [61]. C'est pourquoi nous avons comparé les méthodes d'AMC sur nos jeux de données avec l'entropie comme facteur discriminant. Les résultats nous ont permis de mettre en évidence une méthode d'AMC robuste pour nos jeux de données et nous envisageons de l’appliquer pour mettre en relation les résidus covariants avec l'évolution des RCPG de classe A.

5. OUTILS MIS EN PLACE

La plupart des outils mis en place a été implémentée sous forme de scripts dans le langage de programmation Perl, version 5.8.0, appelant parfois des scripts écrits en langage R, version 2.6.2.

5.1 Jeu de données

Les séquences des RCPG de classe A non-olfactifs ont été obtenues à partir de la base de données UniProt (http://www.uniprot.org/) avec le profil PROSITE PS50262 (RCPG de classe A). Les récepteurs olfactifs (profil InterPro IPR000725), des phéromones (profil InterPro IPR004072) et de goût (profil InterPro IPR007960) ont été exclus. Les séquences redondantes ont été supprimées avec le programme nrdb [62], avec 90% d'identité de séquences. L'AMS des séquences a été obtenu avec le programme ClustalX (http://www.clustal.org/) [63] et manuellement vérifié avec le programme Genedoc (http://www.nrbsc.org/). Les séquences alignées ont été utilisées pour produire un modèle caché de Markov des RCPG de classe A non-olfactifs en utilisant le programme HMMER (http://hmmer.janelia.org/) [64]. Le profil PS50262 et le modèle caché de Markov ont permis ensuite d'obtenir les séquences pour les génomes de *N. vectensis*, *C. elegans et D. rerio* à partir de la base UniProt. Les séquences de *C. intestinalis* ont été obtenues auprès de Fredriksson et Schioth (2005) [49]. Les séquences redondantes ont également été supprimées et les séquences ont ensuite été alignées à celle d'*H. Sapiens*. L'AMS obtenu contient 236 positions qui se basent sur les 7 HTM, la huitième hélice intracellulaire putative et des parties des boucles IC et EC. L'étude des RCPG par arbres phylogénétiques effectuée par l'équipe de bioinformatique [45] a permis d'assigner les récepteurs humains. Concernant les récepteurs non humains, l'assignation a été basée sur l'identité de séquences avec les cinq plus proches orthologues humains. Si au moins les quatre premiers orthologues sont de la même sous-famille, le récepteur non humain appartient à cette sous-famille. Dans le cas contraire, le récepteur appartient à la catégorie UC.

5.2 Distances entre séquences

Le calcul de distances entre séquences d'un AMS conduit à l'obtention d'une matrice de distances. Cette structure de données est extrêmement utile pour visualiser les divergences et constitue la base pour certaines méthodes d'arbres phylogénétiques et l'application de méthodes de réduction dimensionnelle. Ils existent différentes méthodes de calcul de proximité entre deux séquences protéiques. On peut les classer en deux groupes : celles basées sur la différence et celles

fondées sur la dissimilarité.

5.2.1 Calcul de la différence

Le principe du calcul de la différence est simple et se base sur l'inverse du pourcentage d'identité : le nombre de positions alignées qui n'affichent pas les mêmes acides aminés, divisé par un dénominateur (longueur de l'alignement, plus courte séquence...). Il n'existe pas encore de méthode standard pour le calcul de la différence [65].

Les études des différentes variantes de la différence n'ont pas montré clairement la prédominance d'un dénominateur par rapport à un autre. Des préférences peuvent toutefois exister dans des cas très particuliers [66]. Pour notre implémentation, un seul dénominateur est utilisé : nombre de positions alignées sans gap entre deux protéines. Ce choix permet de comparer les résultats obtenus avec le calcul de la dissimilarité (voir chapitre suivant), qui lui ne prend pas en compte les gaps. Les valeurs de cette différence sont comprises entre 0 (séquences identiques) et 1 (séquences totalement différentes).

5.2.2 Calcul de la dissimilarité

Le principe du calcul de dissimilarité se base sur celui de la similarité, qui nécessite l'utilisation de matrices de substitution. A chaque paire d'acides aminés correspond un score, positif si la substitution est favorisée et négatif si la substitution a moins de chance de se produire qu'aléatoirement [67]. C'est pourquoi cette méthode est plus appropriée aux séquences *a priori* distantes mais partageant des caractéristiques physico-chimiques. Le principe du calcul de similarité peut paraître simple : la somme des scores attribuées aux paires d'acides aminés, sans gap, de l'alignement. Cependant, la similarité totale peut être positive ou négative (également au niveau de la diagonale de la matrice obtenue) et n'est donc pas comprise entre 0 et 1. Une étude concernant la problématique du regroupement d'objets biologiques utilise une formule permettant de répondre à ce critère [68,69].

Cet algorithme peut utiliser différentes matrices de substitution (**Tableau 2**). Différentes matrices génériques Point Accepted Mutation (PAM) [70] et BLOcks SUbstitution Matrix (BLOSUM) [71] ont été sélectionnées pour permettre de tester le calcul de dissimilarité avec des séquences peu

Date	Acronyme	Définition	Note	Référence
Génériques				
1978	PAM	Point Accepted Mutation	40, 120, 250	[70]
15/10/1992	BLOSUM	BLOcks SUbstitution Matrices	45, 62, 80	[71]
Spécialisées				
21/02/1994	PAM250TM	TransMembrane PAM250	proche PAM	[72]
16/09/2000	PHAT	Predicted Hydrophobic And Transmembrane	proche BLOSUM	[73]
02/04/2001	SLIM	Scorematrix Leading to IntraMembrane domains	base de données	[74]
01/04/2003	BATMAS	BActerial Transmembrane MAtrix of Substitutions	procaryote	[75]

Tableau 2 : Matrices de substitution

Le tableau affiche les matrices de substitution utilisé pour le calcul de dissimilarité. Les matrices PAM40, PAM120, PAM250, BLOSUM 45, BLOSUM62 et BLOSUM80 représentent les matrices génériques. Les matrices spécialisées PAM250TM et PHAT sont proches des matrices PAM et BLOSUM, respectivement. Les matrices SLIM et BATMAS ne sont pas utilisées car elles sont employées dans des cas bien particuliers.

divergentes (PAM 40 ou BLOSUM 80) et très divergentes (PAM 250 ou BLOSUM 45). Cependant, l'analyse de protéines affichant des caractéristiques physico-chimiques et une composition en acides aminés non-standard (comme les RCPG) nécessiterait l'utilisation de matrices appropriées. Les RCPG sont des protéines transmembranaires. L'environnement hydrophobe des acides aminés localisés dans la membrane lipidique est très différent du milieu aqueux extracellulaire. Ainsi, une matrice spécialisée conviendrait bien mieux aux protéines transmembranaires que celles généralisées à partir de toutes les protéines. Les principales différences entre les matrices généralistes et spécialisées se situent au niveau de la proline qui est hautement conservée dans les segments transmembranaires, due à son rôle clé dans les cassures d'hélices, et au niveau des acides aminés hydrophobes qui tolèrent davantage de substitutions (isoleucine, méthionine et valine). La matrice TransMembrane PAM250 (PAM250TM) [72] diverge de la matrice PAM250 par la nature transmembranaire des segments utilisés lors de sa construction. Elle donne de bons résultats concernant l'alignement de protéines transmembranaires proches mais de mauvais pour la recherche d'informations. Le principe de l'élaboration de la matrice Predicted Hydrophobic And Transmembrane (PHAT) [73] est proche de celui des BLOSUM. Elle surpasse les matrices généralistes et transmembranaires (PAM250TM) concernant la recherche en base de données et peut aider à construite des alignements structuraux et des arbres phylogénétiques de protéines membranaires. La matrice Scorematrix Leading to IntraMembrane domains (SLIM) [74] est spécialisée dans la recherche en base de données et n'est pas symétrique. La matrice BActerial Transmembrane MAtrix of Substitutions (BATMAS) [75] est caractéristique des protéines procaryotes. Ces deux dernières matrices ne sont pas utilisées de par leur caractère très spécialisé. Pour avoir un aperçu des différences entre les matrices utilisées, veuillez vous référer à l'annexe concernant les applications du package R.

5.3 DISTATIS

Le choix d'une méthode de calcul de distances passe par la comparaison des degrés de différence et de dissimilarité. Lorsque l'une de ces méthodes est appliquée à un AMS, elle retourne une matrice symétrique contenant des distances calculées pour toutes les paires de séquences. Une méthode permettant d'analyser et de comparer visuellement des matrices de distances calculées à partir des mêmes objets (*i.e.*, séquences) existe : c'est la méthode DISTATIS [76,77].

Cette méthode est considérée comme une généralisation de la MDS métrique et son nom provient d'une autre méthode, appelée Structuration des Tableaux A Trois Indices de la Statistique (STATIS). La méthode DISTATIS analyse un ensemble de matrices de distances qui peuvent correspondre à des mesures opérées à différents moments, dans le cadre d'une cinétique, ou à des algorithmes de calcul différents, comme dans notre cas ; pourvu que les matrices soient calculées à partir du même jeu de données. Premièrement, cette méthode évalue la similarité entre les matrices de distances. Ensuite, elle calcule une matrice de compromis qui représente la meilleure cohésion des matrices d'origine. Pour finir, les matrices d'origine sont projetées dans l'espace de compromis. Concernant l'algorithme (**Figure 9**). La méthode DISTATIS permet de comparer et de visualiser plusieurs éléments :

- les méthodes grâce aux différentes coordonnées de la matrice de facteurs. Habituellement, la première composante explique la grande proportion de la variance qui coexiste entre les méthodes.

- les objets dans l'espace de compromis grâce à la matrice de compromis des facteurs. Des structures cachées entre les objets sont mises en évidence, communes à toutes les méthodes.

- les positions des matrices de projections sur l'espace de compromis. L'influence de chacune des méthodes sur chaque objet permet une meilleure compréhension des différentes sensibilités.

Une fois les différentes matrices comparées, nous avons souhaité analyser précisément les données contenues dans les matrices elles-mêmes.

5.4 MDS métrique

La MDS métrique est une méthode d'analyse statistique qui représente les objets dans un espace de faibles dimensions en étant aussi fidèle que possible à la configuration obtenue avec un espace à hautes dimensions. Elle met en évidence des structures cachées en tenant compte de la proximité entre objets. Ces derniers sont dits actifs car ils contribuent directement à la génération de l'espace. Concernant l'algorithme (**Figure 10**), N est le nombre d'éléments actifs, **D** la matrice active des

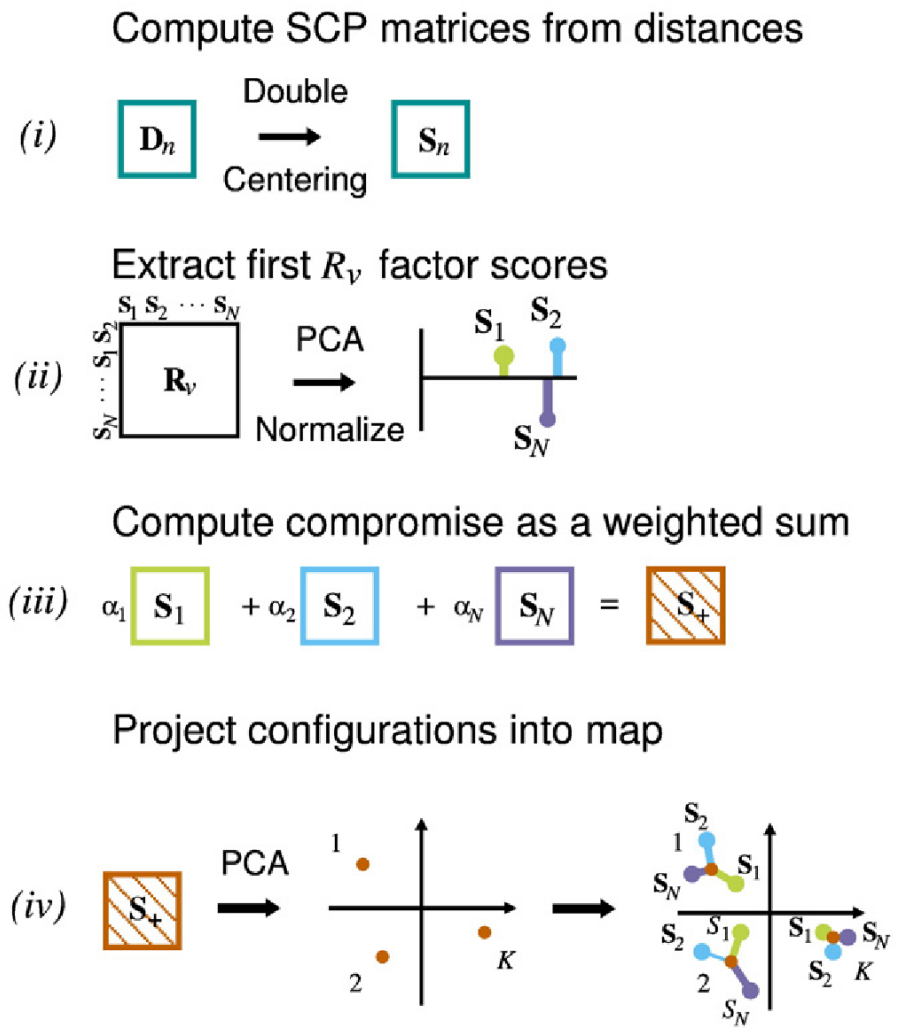

Figure 9 : Méthode DISTATIS

La première étape (*i*) consiste à obtenir les matrices de produit vectoriel (SCP pour Sum of Cross-Product) à partir des matrices de distances. Ensuite, la deuxième étape (*ii*) concerne le calcul des coefficients Rv qui permettent de comparer les méthodes entre elles. Le calcul de l'espace de compromis est réalisé lors de la troisième étape (*iii*). Pour finir, la quatrième étape (*iv*) permet de projeter des objets dans cette espace de compromis et d'évaluer l'influence de chacune des méthodes sur la position de chaque objet. Figure tirée de [78].

distances au carré, N par N. La MDS métrique comprend deux calculs principaux [79] :

- transformation de la matrice active **D** en matrice de produit vectoriel **S** :

$$\mathbf{S} = -0.5[\mathbf{I} - (\mathbf{1}/\mathrm{N})\mathbf{1}] \times \mathbf{D} \times [\mathbf{I}(\mathbf{1}/N)\mathbf{1}] \ , \tag{1}$$

  avec **I** une matrice d'identité N par N et **1** une matrice de valeurs 1, N par N.

- transformation de la matrice de produit vectoriel **S** en matrice de facteurs **F** :

$$\mathbf{F} = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}} \ , \tag{2}$$

  avec **U** la matrice de vecteurs propres et $\boldsymbol{\Lambda}$ la matrice diagonale des valeurs propres, telle que $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{T}$. La matrice **F** donne les coordonnées des éléments actifs pour la génération de l'espace actif.

La méthode PCA, assez proche de la méthode MDS, offre la possibilité de projeter des variables qualitatives et quantitatives supplémentaires sur l'espace actif. Récemment, le Pr. Abdi a développé une technique permettant d'appliquer ce principe de projection mais pour des éléments supplémentaires dans le cadre de la MDS métrique [79].

5.5 Projection d'éléments supplémentaires

Après avoir réalisé une MDS métrique, les éléments supplémentaires (ou illustratifs) qui sont projetés ne participent pas à la génération de l'espace actif et leurs positions dépendent uniquement des distances par rapport aux éléments actifs. Concernant l'algorithme (**Figure 10**), N_{sup} est le nombre d'éléments supplémentaires, $\mathbf{D}_{sup}$ la matrice supplémentaire des distances au carré, N_{sup} par N. La projection comprend deux calculs principaux [79] :

- transformation de la matrice supplémentaire $\mathbf{D}_{sup}$ en matrice de produit vectoriel $\mathbf{S}_{sup}$:

$$\mathbf{S}_{sup} = -0.5[\mathbf{I} - (\mathbf{1}/\mathrm{N})\mathbf{1}]\left(\mathbf{D}_{sup}^{\mathrm{T}} - (1/\mathrm{N})\mathbf{D}\,\mathbf{1}_{sup}\right) \ , \tag{3}$$

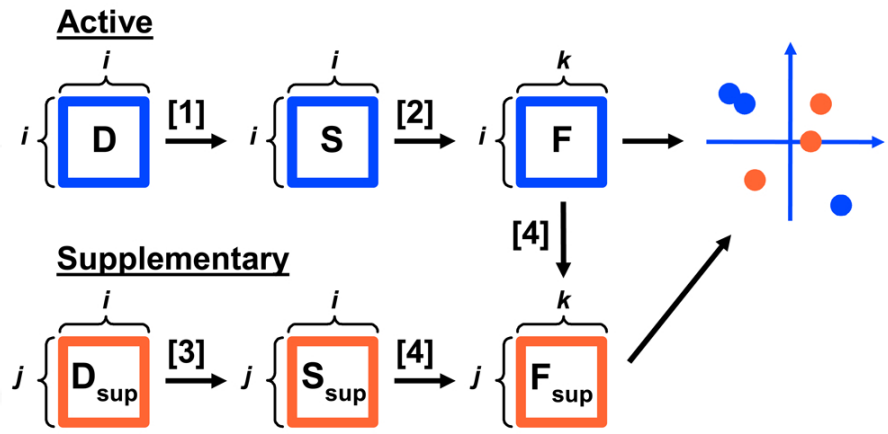

Figure 10 : Projection d'éléments supplémentaires dans le cadre de la MDS

Les symboles **D**, **S** et **F** représentent la matrice de distances au carré, de produit vectoriel et de facteurs pour les éléments actifs, respectivement. Les symboles $\mathbf{D}_{sup}$, $\mathbf{S}_{sup}$ et $\mathbf{F}_{sup}$ concernent les mêmes matrices respectivement mais pour les éléments supplémentaires. Le nombre d'éléments actifs, supplémentaires et le nombre de composantes sont indiqués par les lettres *i*, *j* et *k*, respectivement. Sur le repère orthogonal, les points bleus représentent les éléments actifs alors que les points oranges concernent les éléments supplémentaires. Les chiffres entre crochets correspondent à la numérotation des formules de la MDS métrique.

avec $\mathbf{1}_{sup}$ une matrice de valeurs 1, N_{sup} par N.

- transformation de la matrice de produit vectoriel $\mathbf{S}_{sup}$ en matrice de facteurs $\mathbf{F}_{sup}$:

$$\mathbf{F}_{sup} = \mathbf{S}_{sup}^{T} \mathbf{F} \boldsymbol{\Lambda}^{-1} , \quad (4)$$

avec $\mathbf{F}_{sup}$ donnant les coordonnées des éléments supplémentaires dans l'espace actif.

Cette projection permet d'identifier les caractéristiques qui sont partagées entre les éléments actifs et supplémentaires et, par exemple, de donner un début d'explication pour les éléments supplémentaires lorsque leurs caractéristiques sont inconnues.

5.6 Clustering

Suite à une MDS métrique, les données peuvent se regrouper sous forme de clusters ou groupes. La visualisation des groupes peut être confirmée par des méthodes de clustering dont l'objectif est d'associer en groupes des données telles que la proximité entre les objets d'un même groupe soit maximal et que celle entre les objets de groupes différents soit minimal. Le clustering par partitionnement se différencie du clustering hiérarchique car les données ne sont pas structurées sous forme d'un arbre hiérarchique. La méthode des K-moyennes (K-means) est une méthode par partitionnement largement utilisée et demande un temps de calcul faible (utile dans le cadre de bootstrapping).

Parfois, le nombre de groupes ne peut pas être déterminé visuellement et reste incertain du fait d'une distribution particulière des données. Le nombre de groupes est intimement lié aux notions de compactivité et de séparation. La compactivité et la séparation fonctionnent de manière antagoniste car la compactivité augmente avec le nombre de groupes mais fait diminuer la séparation. Plusieurs indices de validation permettent de tester la qualité des groupes tels que l'indice de Dunn, l'indice de Davies-Bouldin ou encore l'indice de silhouette. Étant donné le caractère bruité des données biologiques, le dernier indice est préféré car l'indice de Dunn est justement sensible aux données aberrantes et l'indice de Davies-Bouldin donne seulement une valeur globale par groupe [80]. L'indice de silhouette mesure le degré de confiance dans l'affectation du groupe pour un objet

particulière et les valeurs s'échelonnent de -1 à 1 [81]. Un objet se plaçant entre deux groupes affiche une valeur proche de 0 et un objet qui ne se place pas dans le bon groupe exprime une valeur négative. Le score de silhouette peut être calculé et représente la moyenne des indices. Un score élevé est recherché car il confirme le nombre de groupes qu'il est conseillé de prendre.

5.7 Analyse des groupes

La clusterisation permet de distinguer les séquences et de les regrouper mais pas d'expliquer les critères qui font qu'une séquence appartient à un groupe et pas à un autre. Les positions de l'AMS ne doivent pas être comparées pour chaque groupe indépendamment mais par confrontation d'un groupe par rapport à un autre (**Figure 11**). Pour identifier les positions propres, il a fallu développer une stratégie de calcul. Tout d'abord, chaque groupe est comparé à son complémentaire. Pour chaque combinaison et chaque position *i* de l'AMS, deux calculs sont réalisés :

- d'une part, le calcul de la différence de conservations entre deux groupes.

- d'une autre part, le calcul de la spécificité de résidus envers chacun des deux groupes.

Pour le premier point, on a fait appel à la différence d'entropies de Shannon. Pour le second point, il a fallu adapter différentes méthodes, majoritairement des méthodes d'analyse des mutations corrélées (voir chapitre 5.8), à notre stratégie de calcul.

5.7.1 Termes communs

Les méthodes qui ont été utilisées et développées possèdent des termes techniques en commun :

x : l'ensemble des acides aminés présents à la position i.
g : l'ensemble des groupes.
N : le nombre total de séquences.
$f(g)$: fréquence du groupe g.
$f_x(i)$: fréquence de l'acide aminé x à la position i.

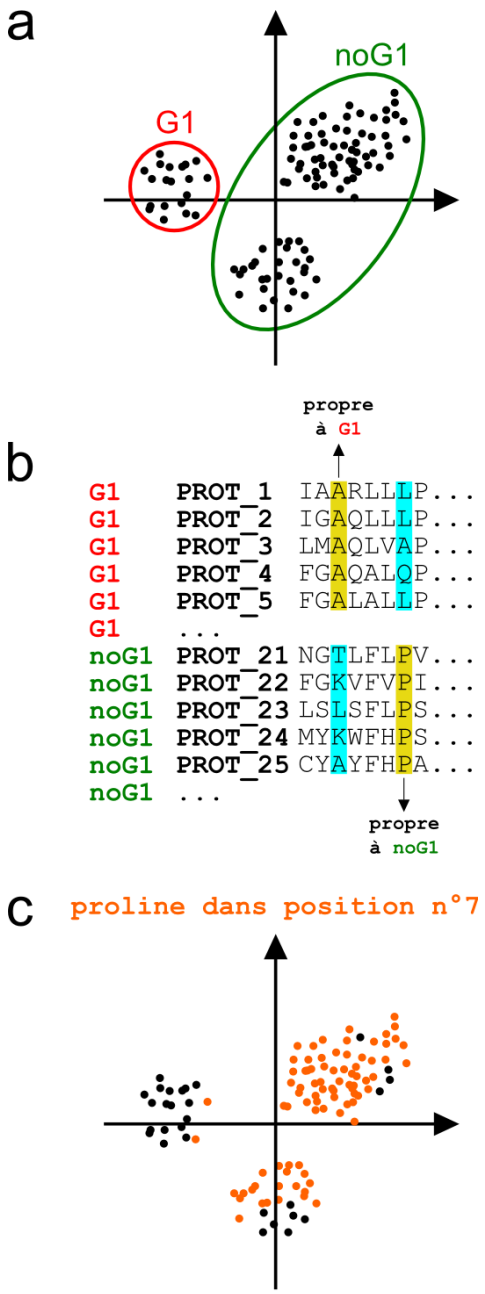

Figure 11 : Analyse des corrélations groupes-séquences

(a) Suite à l'analyse MDS et au clustering, les séquences sont réunies par leurs coordonnées en deux groupes, par exemple : G1 et son complémentaire, noG1. (b) L'AMS qui correspond à une partie des séquences de G1 (PROT_1-5) et de noG1 (PROT_21-25) est affiché. La position n°3 est propre à G1, avec un résidu alanine, mais pas à noG1 et inversement pour la position n°7, avec un résidu proline. Les méthodes de calcul des corrélations groupes-séquences vont révéler ces positions caractéristiques. (c) La présence du résidu proline dans la position n°7 est reportée et permet de visualiser la correspondance sur le premier graphique.

$f_x(i,g)$: fréquence de l'acide aminé *x* à la position *i* pour le groupe *g*.
$f_x(i^g)$: fréquence de l'acide aminé *x* présent dans le groupe g à la position *i* (à ne pas confondre avec la fréquence précédente car $f_x(i^g)$ se calcule par rapport au nombre de séquences totales).

5.7.2 Entropie

Le principe de l'entropie est lié à celui de la théorie de l'information. L'entropie représente le degré de conservation que l'on peut trouver au niveau d'une position *i* d'un AMS. L'entropie de Shannon est adaptée à la bioinformatique et devient l'entropie de séquence comme suit :

$$H_i = -\sum_x f_x(i) \times \ln f_x(i)$$

Le logarithme peut être népérien ou de base 20 (égal au nombre de type d'acides aminés). Dans les deux cas, si la position *i* contient un seul type d'acides aminés, l'entropie est de 0. Si la position *i* affiche une complète variabilité, avec la même proportion de chacun des 20 acides aminés, l'entropie est d'environ 3, pour le logarithme népérien, ou de 1, en base 20.

5.7.3 Différentes méthodes de corrélation entre groupes et séquences

Dans la littérature scientifique en bioinformatique, il n'existe pas de méthodes de calcul qui soient adaptées à ce que l'on souhaite mettre en évidence. Ainsi, nous avons décidé d'adapter certaines méthodes statistiques en modifiant leurs algorithmes pour réussir à mettre en évidence les liens qui peuvent exister entre des groupes et certaines positions d'un AMS. Nous avons utilisé des méthodes d'AMC (voir chapitre 5.8) et la méthode spécialisée l'optimisation combinatoire de l'entropie (CEO pour Combinatorial Entropy Optimization). Seules les méthodes d'AMC les plus fréquemment rencontrées dans la littérature ont été utilisées : les méthodes observé moins attendu au carré (OMES pour Observed Minus Expected Squared) et l'information mutuelle (MI pour Mutual Information). De plus, seule une partie de l'algorithme de la méthode CEO a été exploitée.

5.7.3.1 Méthodes basées sur le chi-2

Les deux versions de la méthode OMES on été utilisées : version de Kass et Horovitz (OMES1) [82] et version de Fodor et Aldrich (OMES2) [5]. Les algorithmes correspondent à ceux des

mutations corrélées, mais ils ont été modifiés pour correspondre aux calculs des corrélations entre groupes et séquences (voir annexes). La formule de Kass et Horovitz devient :

$$OMES1_i = N \sum_g \sum_x \frac{f(g) \times \left(f_x(i,g) - f_x(i)\right)^2}{f_x(i)}$$

S'il existe une corrélation parfaite entre groupes et une position i, la formule donne N. S'il n'existe aucune corrélation, la formule donne la valeur 0 car fx(i,g) = fx(i). Cette méthode OMES modifiée peut être normalisée de façon à ce que ses valeurs s'échelonnent de 0 à 1. Elle est appelée fréquence de corrélations (FC pour Frequency Correlation) et sa formule est comme suit :

$$\mathrm{FC}_i = \sum_g \sum_x \frac{f(g) \times \left(f_x(i,g) - f_x(i)\right)^2}{f_x(i)}$$

Concernant la formule de Fodor et Aldrich, la formule est adaptée et devient :

$$\mathrm{OMES2}_i = N \sum_g \sum_x \left(f(g)\right)^2 \times \left(f_x(i,g) - f_x(i)\right)^2$$

S'il existe une corrélation entre les groupes et une position *i*, la formule renvoie une valeur maximale supérieure à 0, à savoir :

$$\mathrm{OMES2}_i = 4N \prod_x f_x(i)^2$$

S'il n'y a aucune corrélation, la formule retourne la valeur 0 car $f_x(i,g) = f_x(i)$.

5.7.3.2 Méthodes basées sur l'information mutuelle

Deux versions de la méthode MI ont été utilisées : la MI classique et l'information mutuelle normalisée (MIr pour Mutual Information removed). La formule correspondant à celle de la MI devient :

$$\mathrm{MI}_i = \sum_g \sum_x f_x(i^g) \times \ln \frac{f_x(i^g)}{f_x(i) \times f(g)}$$

Si la corrélation est parfaite entre les groupes *g1* et $g1^c$ (le complémentaire), et une position *i*, la formule devient :

$$\mathrm{MI}_i = f(gI) \times \ln \frac{1}{f(gI)} + f(gI^c) \times \ln \frac{1}{f(gI^c)}$$

S'il n'y a aucune corrélation entre les groupes et une position *i*, le score est de 0. La MI peut être divisé par la MI maximal (MImax, qui correspond à la formule précédente) pour ne pas dépendre de la taille des groupes. Dans ce cas, le maximum possible est de 1. La version de l'information mutuelle normalisée, MIr, correspond à la MI divisé par l'entropie jointe (JE pour Joint Entropy) :

$$\mathrm{MIr}_i = \frac{\mathrm{MI}_i}{-\sum_g \sum_x f_x(i^g) \times \ln f_x(i^g)}$$

5.7.3.3 Optimisation de l'entropie combinatoire

Historiquement, la méthode de l'Optimisation de l'Entropie Combinatoire (CEO pour Combinatorial Entropy Optimization) fut développée pour résoudre le problème complexe d'identification de résidus spécifiques et, simultanément, celui de la séparation optimale d'un ensemble de séquences protéiques en sous-familles [83]. L'implémentation ne demande qu'un AMS et retourne les sous-familles, séparées par les résidus spécifiques, généralement responsables de la diversité fonctionnelle. Ces résidus sont ensuite visuellement identifiables car ils sont conservés dans chaque sous-famille mais différent entre sous-familles. Dans le cadre de notre étude, seule une partie de l'algorithme de la méthode CEO va nous servir. En effet, la partie concernant la séparation des séquences protéiques en sous-familles est volontairement omise. La méthode est adaptée de telle manière que l'identification des positions clés de l'AMS soit amorcée à partir de notre propre classification, à savoir nos groupes. Les valeurs de CEO sont toujours inférieures ou égales à 0. Si une corrélation entre des groupes et les acides aminés de la position *i* existe, la valeur renvoyée est inférieure à 0. S'il n'y a aucune corrélation, la valeur est de 0.

5.7.3.4 Comparaisons

Les différentes méthodes présentées renvoient des valeurs différentes pour des cas de figures identiques. Les scores ne sont pas exploitables directement. Un score n'a réellement de sens que s'il est mis en relation avec la moyenne de l'échantillon. De plus, des scores venant de différentes distributions doivent être centrés et réduits dans le but d'offrir un moyen de comparaison entre les

différentes méthodes. Pour comparer ces méthodes, les Z-scores sont calculés :

$$\text{Z}-\text{score} = \frac{x - \mu}{\sigma}$$

avec :

x : score brut.

μ : moyenne de l'échantillon.

σ : écart type de l'échantillon.

Quel que soit l'échantillon, les Z-scores ont une moyenne de 0 et un écart-type de 1. Le Z-score est négatif lorsque le score d'origine est en dessous de la moyenne et est positif lorsqu'il est au-dessus. La comparaison est effectuée avec un jeu de données commun (**Figure 12**). On peut voir que les méthodes donnent des résultats très proches malgré des algorithmes différents. Les méthodes basées sur OMES2 et MIr sont écartées car elles ont tendance à trop compacter les Z-scores faibles. Ensuite, les résultats obtenus par les méthodes OMES1 (ou FC, cela revient au même avec les Z-scores), MI et CEO sont très proches. La méthode OMES1 est préférée car il a été démontré que, dans le cadre des AMC, elle donne de meilleurs résultats que la méthode MI [5].

5.8 Méthodes d'analyse des mutations corrélées

Les méthodes d'AMC recherchent les dualités qui peuvent se manifester entre deux acides aminés et fonctionnent à partir d'un AMS (**Figure 13**). A chaque fois qu'un acide aminé donné d'une position d'un AMS est modifié, l'algorithme de la méthode analyse si un changement correspondant affecte un autre acide aminé d'une autre position du même AMS. Les méthodes d'AMC présentent des caractéristiques inhérentes à la définition de la covariation qu'elles proposent. Pour chacune des méthodes, un score de corrélation va être ainsi attribué pour chaque paire de positions de l'AMS. Les méthodes réagissent différemment face à un même jeu de données et ont différentes sensibilités vis à vis de la conservation. Il y a peu, de nombreuses modifications des méthodes d'AMC ont été proposé dans le but de réduire la proportion de faux-positif [85]. Il existe une dizaine de méthodes de calcul des mutations corrélées (**Tableau 3**) fonctionnant sur différents principes statistiques et je me suis intéressé aux principales [86]. Pour toutes les méthodes, seules les séquences ne possédant pas de gap aux positions *i* et *j* sont prises en compte.

Les méthodes d'AMC sont intimement liées à la conservation des résidus. Chaque algorithme

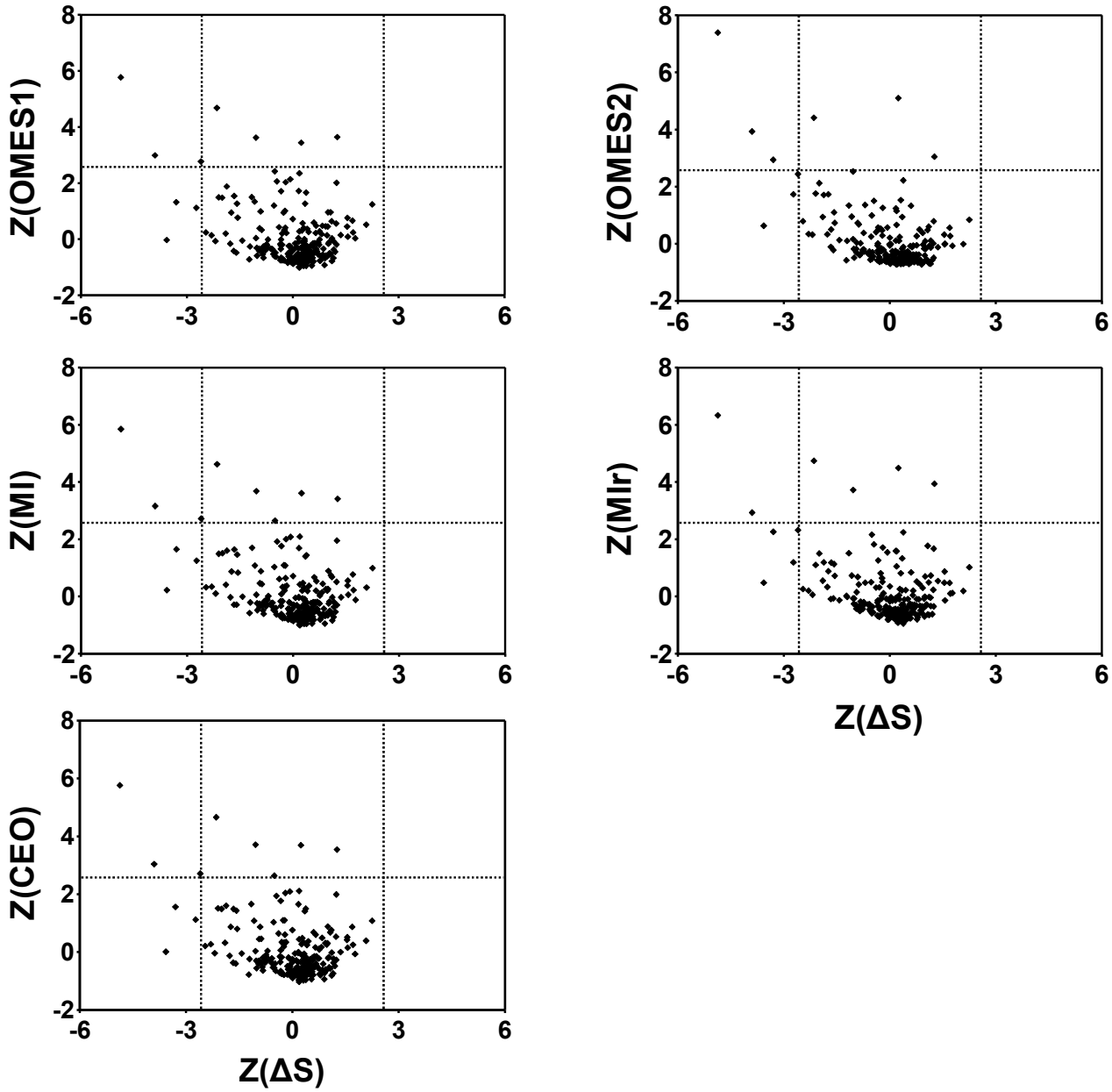

Figure 12 : Comparaison des méthodes de corrélation groupes-séquences

La comparaison est effectuée pour le G1 contre son complémentaire noG1 (voir chapitre 6.1.2). De gauche à droite et de haut en bas, les graphiques correspondant aux méthodes suivants sont affichées : OMES1, OMES2, MI, MIr et CEO. Chaque point matérialise une position des HTM de l'AMS des RCPG de classe A des espèces suivantes : *N. vectensis*, *C. elegans*, *C. intestinalis*, *D. rerio* et *H. sapiens*. Pour chaque graphique, l'axe des abscisses et des ordonnées représentent les Z-scores de la différence d'entropies et de la méthode adaptée, respectivement. Les lignes en pointillés permettent de différencier les positions au-dessus du seuil de confiance à 95%.

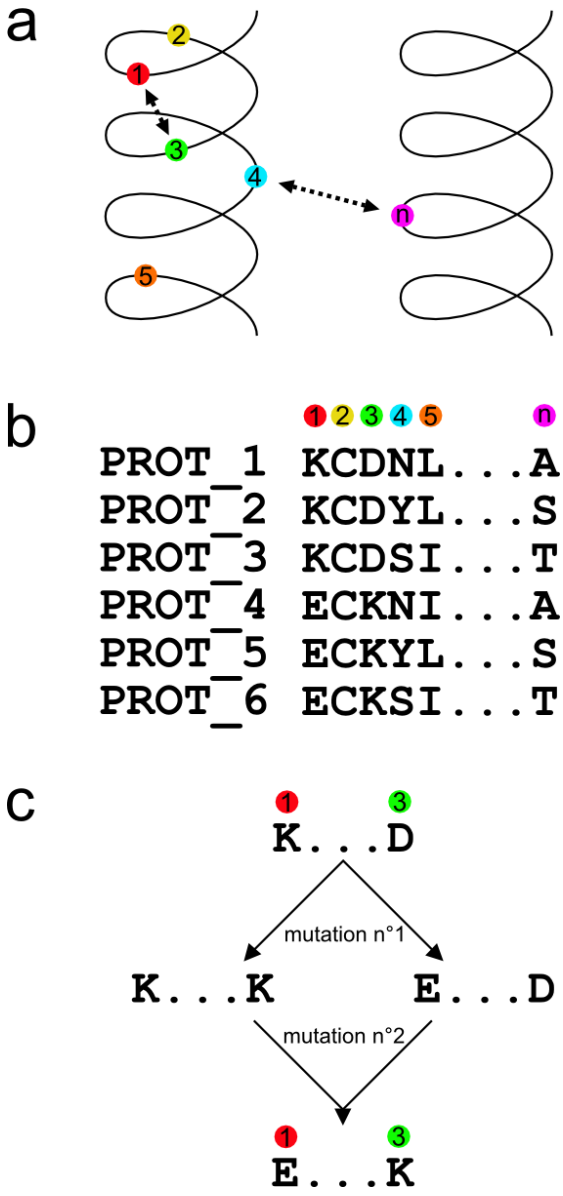

Figure 13 : Analyse des mutations corrélées

(a) Plusieurs positions sont placées sur la structure de la protéine correspondante, par exemple deux hélices α proches dans notre contexte des RCPG. (b) L'AMS qui correspond aux six séquences (PROT_1-6) de ces positions est affiché. Les paires de positions (1,3) et (4,n) montrent une certaine corrélation entre les trois premières et les trois dernières séquences. (c) Lorsque des positions importantes pour la structure subissent des mutations (mutation n°1), les protéines qui en résultent s'en retrouvent instables. Des mutations compensatrices (mutation n°2) permettent de restaurer la stabilité structurale. Les méthodes AMC ne détectent pas les états intermédiaires mais interprètent l'état initial et final en tant que mutations corrélées. Figure inspirée de [84].

Basée sur	Date	Acronyme	Définition	Référence	Utilisé ?
le chi-deux	01/09/2002	OMES1	Observed Minus Expected Squared de Kass et Horovitz	[82]	✔
	01/08/2004	OMES2	Observed Minus Expected Squared de Fodor et Aldrich	[5]	✔
l'information mutuelle	17/01/2000	MI	Mutal Information	[7]	✔
	15/11/2005	MIr	Mutual Information Removed	[90]	✔
	01/02/2008	MIp	Mutual Information Product	[91]	
	01/02/2008	MIa	Mutual Information Additive	[91]	
	12/04/2003	MINT1	Mutual INTerdependency	[92]	✔
	12/04/2003	MINT2	Alternative Mutual INTerdependency	[92]	
l'asymétrie	08/10/1999	SCA	Statistical Coupling Analysis	[93]	✔
	10/07/2004	ELSC	Explicit Likelihood of Subset Covariation	[94]	✔
des matrices de subtitution	18/04/1994	MCBASC1	McLachlan BAsed Substitution Correlation McLachlan	[4,97]	✔
	18/04/1994	MCBASC2	McLachlan BAsed Substitution Correlation Miyata	[4,98]	
	01/03/2005	CMAV1	Correlated Mutation Analysis of Vicatos PRIN1	[99]	
	01/03/2005	CMAV2	Correlated Mutation Analysis of Vicatos PRIN2	[99]	
	01/03/2005	CMAV3	Correlated Mutation Analysis of Vicatos PRIN3	[99]	

Tableau 3 : Les différentes méthodes d'AMC

Le tableau affiche les méthodes implémentées au laboratoire de bio-informatique. Les colonnes indiquent sur quel principe est basé la méthode, la date de publication de l'algorithme, l'acronyme, la définition de l'acronyme et si la méthode a été utilisée dans les résultats, respectivement.

porte une préférence pour un certain niveau de conservation et sélectionne les positions qu'il considère covariantes. Ces algorithmes favorisent un niveau de conservation particulier et attribuent un score élevé pour la paire de positions en question.

Les méthodes d'AMC que nous avons utilisées peuvent être réunies en quatre groupes suivant leur principe de fonctionnement. Du fait de la redondance dans le comportement de certaines méthodes, nous présentons uniquement celles qui ont réellement été utilisées et discutées dans les résultats. Veuillez vous référez aux annexes pour plus de détails sur les algorithmes.

5.8.1 Méthodes basées sur le chi-2

Ces méthodes sont représentées par les méthodes OMES, qui sont symétriques et qui existent sous deux versions comme dit précédemment : OMES1 et OMES2. La différence se situe au niveau de l'algorithme mais le principe général de fonctionnement est identique : elles comparent l'occurrence observée de chaque paire d'acides aminés *xy* présente aux positions *i* et *j* de l'AMS avec leur occurrence attendu. Cette différence est estimée par une formule mathématique proche de celle du test statistique du chi-2 [87]. La version OMES1 est légèrement supérieure à la variante OMES2 dans la prédiction de contacts indépendants entre hélices α transmembranaires [88].

5.8.2 Méthodes basées sur l'information mutuelle

La méthode MI est basée sur la définition donnée par Atchley [7] et est attractive parce qu'elle mesure la dépendance d'une position par rapport à une autre, mais se retrouve limitée par trois facteurs : l'influence de l'entropie, la taille de l'AMS et les contraintes phylogénétiques. Chacun de ces facteurs tend à modérer le signal véritable et contribue à un certain bruit de fond [89,90] :

- l'influence de l'entropie : les positions avec une entropie élevée tendent à afficher de plus hauts niveaux de MI que celles avec une entropie faible. Cela peut s'expliquer en partie par le fait que la méthode gère mal les positions aléatoires, en tout cas très variables, et cet inconvénient est inhérent à la méthode.

- la taille de l'AMS : la valeur de MI entre deux positions exprime le degré de corrélation. La valeur théorique de MI sera de 0 si les acides aminés pour une paire de positions sont indépendants. Cependant, en pratique, la valeur de MI sera de 0 seulement si les fréquences des paires d'acides aminés observées reflètent toutes les possibilités d'appariements pour les fréquences d'acides aminés observées pour chaque position de la paire observée. Ces fréquences sont directement liées au nombre de séquences de l'AMS.

- les contraintes phylogénétiques : aucune paire de positions ne peut être véritablement indépendante parce que deux acides aminés de la même protéine sont généralement apparentés. Lorsque deux positions subissent des substitutions de façon complètement indépendantes, l'héritage partagé contribue à l'apparition d'un bruit de fond important lors de l'estimation de la valeur de MI.

L'influence de l'entropie peut être partiellement diminuée par des méthodes de normalisation comme par exemple MIr. D'autres normalisations sont envisageables [91] pour minimiser les effets des trois facteurs précédents : l'Information Mutuelle produit (MIp pour Mutual Information product) et l'Information Mutuelle additive (MIa pour Mutual Information additive). Une dernière méthode basée sur la MI existe et se base plus spécialement sur la quantification des contraintes phylogénétiques : l'INTerdépendance Mutuelle (MINT pour Mutual INTerdependency). Le programme de la méthode MINT, fourni par Tillier et Lui (2003) [92], propose deux méthodes : celle correspondant à la méthode originelle (MINT1) et une autre qui n'est pas décrite en tant que telle dans l'article et qui constitue une alternative (MINT2). La différence se situe au niveau de la définition du ratio de dépendance [92]. Veuillez vous référez aux annexes pour plus de détails.

5.8.3 Méthodes asymétriques

Les deux méthodes suivantes ont la particularité de ne pas être symétrique contrairement à la plupart des méthodes d'AMC. Ici, le calcul du score se fait toujours d'une positon de référence par rapport à une autre et non en prenant en compte l'ensemble de la paire de positions en tant qu'entité complète. Plusieurs stratégies sont envisageables pour résoudre le problème d'asymétrie (la moyenne, par exemple). Mon implémentation de ces méthodes reste asymétrique parce que l'effet des différentes stratégies n'a pas encore été étudié.

5.8.3.1 Analyse de couplage statistique

La méthode d'analyse de couplage statistique (SCA pour Statistical Coupling Analysis) repose sur les études menées par Lockless et Ranganathan [93]. Des études ont apporté des corrections à l'algorithme pour l'adapter au calcul des mutations corrélées [59,94,95]. Cette méthode est particulière car elle compare la composition en acides aminés d'un sous-alignement à celle d'un alignement total. Le sous-alignement constitue la perturbation et est caractérisé par les séquences de l'AMS contenant l'acide aminé le plus conservé à la position *i*.

5.8.3.2 Probabilité explicite de covariation de sous-ensembles

Le principe de la méthode de probabilité explicite de covariation de sous-ensembles (ELSC pour Explicit Likelihood of Subset Covariation) [94] est similaire à celui de la méthode SCA. La détermination du sous-alignement est définie par une contrainte d'identité d'un acide aminé à la position *i*. Ensuite, l'influence de cette perturbation est observée pour chaque autre position *j*. La différence principale qui réside entre la méthode ELSC et SCA est la manière dont on considère la composition en acides aminés entre le sous-alignement et l'alignement total.

5.8.4 Méthodes basées sur des matrices de substitution

La méthode de corrélation de substitution basée sur McLachlan (MCBASC pour McLachlan BAsed Substitution Correlation) est basée sur la définition mathématique du coefficient de corrélation. Elle se distingue des autres par l'utilisation d'une matrice de similarité donnant un score pour chaque paire d'acides aminés [4,96]. Cette matrice peut être de McLachlan (MCBASC1) [97] ou de Miyata (MCBASC2) [98]. L'algorithme de la méthode MCBASC reste identique. Avec la matrice de McLachlan, la méthode MCBASC exprime de meilleurs résultats comparé aux autres méthodes de calcul des mutations corrélées en ce qui concerne la prédiction de contacts hélice-hélice. Cependant, avec la matrice Miyata, la méthode MCBASC offre de meilleures performances lorsqu'il s'agit de détecter les positions corrélées situées dans le même tour d'hélice. Une autre méthode d'AMC se base sur les caractéristiques physico-chimiques des acides aminés : la méthode d'analyse de mutations corrélées de Vicatos (CMAV pour Correlated Mutation Analysis of Vicatos) mais elle n'est pas utilisée [99,100]. Veuillez vous référez aux annexes pour plus de détails.

5.8.5 Implémentation

Nos implémentations des différentes méthodes d'AMC attribuent un score pour des paires de positions avec gaps. Cet ajustement est nécessaire pour pouvoir comparer les différentes méthodes de calcul des mutations corrélées. Les divergences observées ne seront donc pas dues à une gestion du gap qui différente suivant les méthodes. De plus, les paires de positions dont l'une ou les deux positions sont parfaitement conservées, sont écartées du calcul. Cette stratégie est nécessaire car la méthode MCBASC n'attribue aucun score pour ce type de paire de positions alors que les autres méthodes attribuent un score nul ou proche de 0. De plus, le calcul du score entre une position i et j ne se fera que pour $j > i$. Cela permettra d'accélérer le calcul pour les méthodes symétriques et d'analyser les spécificités des méthodes asymétriques.

Toutes les méthodes ont été implémentées en langage de programmation Perl (Version 5.8.0). Les résultats des différentes méthodes ont été comparés aux programmes déjà existant pour s'assurer de la qualité de l'implémentation :

- code source en Java fourni par Fodor et Aldrich (2004) [5] pour OMES (leur version), MI, SCA, ELSC et MCBASC.

- code source en Perl fourni par Martin *et al.* (2005) [90] pour la MIr.

- code source en Perl fourni par Dunn *et al.* (2008) [91] pour la MIp et la MIa.

- code source en C++ fourni par Tillier et Lui (2003) [92] pour la MINT.

5.9 Base de données relationnelle

L'étude des RCPG de classe A implique le référencement de données de différentes natures. Les séquences en acides aminés des RCPG proviennent de sources différentes et doivent être répertoriées en fonction de leur classification. Chaque séquence appartient à une sous-famille, à une famille et à une classe. Elle appartient également à un ou plusieurs AMS aux caractéristiques spécifiques. La complexité des relations qui existent entre ces différentes données exige une

certaine rigueur. Nos classifications propres, notamment celle des différents groupes suite à la MDS, écartent d'emblée l'utilisation des bases de données publiques existantes. La base de données historique GPCRDB (http://www.gpcr.org/7tm/) [101] ne répondant pas à nos exigences, elle ne sera pas utilisée. Ainsi, la gestion de nos nombreuses données spécifiques et la création de classifications appropriées à notre domaine de recherche ont demandé la mise en place d'une base de données propre au laboratoire. De plus, la difficulté de mettre à jour un grand nombre de séquences a nécessité le développement d'une interface Web couplée à la base de données, pour faciliter la visualisation des requêtes. La réalisation des outils informatiques s'est basée sur les travaux déjà réalisés par l'étudiante en doctorat Julie Devillé et de nombreux étudiants effectuant leurs stages de Master dans notre équipe. Les outils de mise à jour des données seront finalisés, administrés et migrés sur le nouveau serveur par l'étudiant en Master de bioinformatique de Rouen, Jean-michel Bécu.

5.9.1 Modèle conceptuelle de données

L'ensemble de nos données est stocké dans une base dite relationnelle parce que les données sont sous forme de tables et sont mises en relation par des bases théoriques solides, notamment la théorie des ensembles (union, par exemple). L'architecture de la base de données relationnelle a été mise en place à l'aide du langage Structure Query Language (SQL). Né dans les années 80, ce langage normalisé est un langage de définition de données (LDD), un langage de manipulation de données (LMD) et un langage de contrôle de données (LCD). Il est devenu incontournable des systèmes de gestion de bases de données (SGBD) relationnelles, tels que MySQL, PostgreSQL ou encore Oracle. Les SGBD représentent l'ensemble des logiciels nécessaires pour gérer indépendamment le niveau logique (objets) du niveau physique (fichiers) et servent d'interface entre la base et l'utilisateur. La SGBD utilisée lors de ce projet est la version 4.1.9 de MySQL. La représentation schématique de la base de données peut être interprétée comme un modèle conceptuel des données (MCD) amélioré. Il comprend les entités suivantes (**Figure 14**) :

- l'entité *gpcr* contient les informations relatives au récepteur telles que son numéro d'accession, l'espèce à laquelle il appartient ou encore sa séquence en acides aminés. Cette entité est centrale parce qu'elle contient de nombreuses relations avec les autres entités et environ la moitié de ses propriétés sont des clés étrangères. Chaque récepteur est identifié

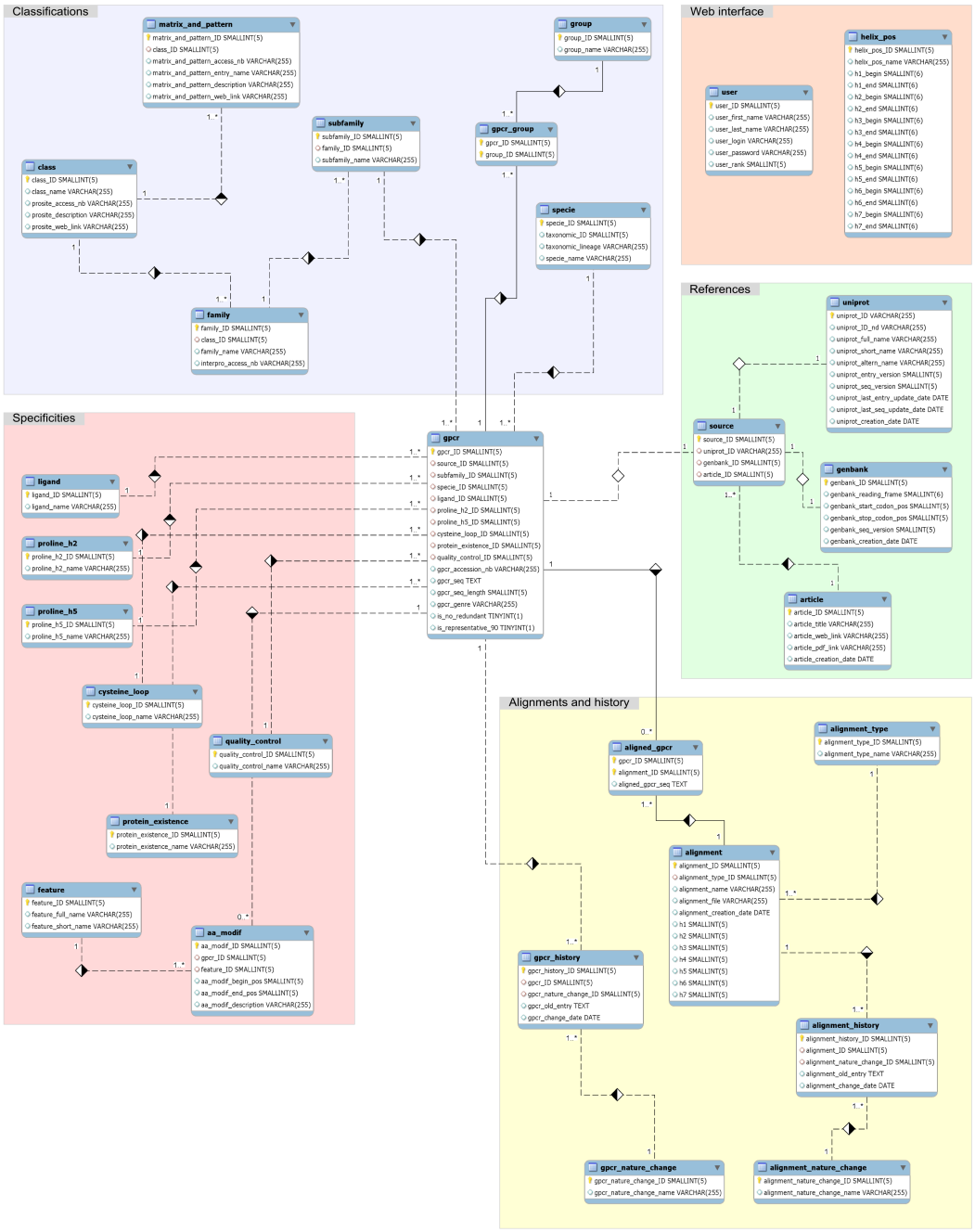

Figure 14 : Modèle conceptuel des données amélioré

Les cadres de couleur regroupent plusieurs entités sous la même thématique. Chaque entité est sous la forme d'un tableau contenant différentes propriétés. Celles avec un symbole de clé, un losange rouge et un losange bleu correspondent aux clés primaires, étrangères et normales, respectivement. Les lignes en continu et en pointillés matérialisent les relations entre entités et les chiffres aux extrémités indiquent les cardinalités. La figure a été réalisée à l'aide de la version 5.1.18 du logiciel MySQL Workbench (successeur de DBDesigner4).

par un ID (identifiant) unique pour empêcher toute redondance.

- le groupe d'entités *Classifications* fait référence aux classifications déjà reconnues dans la littérature, telles que l'espèce, la classe ou la famille, et celles développées spécialement au laboratoire, telles que le groupe MDS ou la sous-famille. L'entité *gpcr_group* est particulière car sa clé primaire est représentée par un couple de deux propriétés. Elle permet, par exemple, de rendre compte de la différence entre les CHEM classiques et non classiques (voir chapitre 6.2.1 pour la correspondance).

- le groupe d'entités *Specificities* permet d'attribuer des caractéristiques de séquences aux séquences en acides aminés. On peut noter la présence de l'entité *proline_h2* qui permet d'attribuer la position de la proline dans l'HTM2 : position 2.58, 2.59, 2.60 ou absence.

- le groupe d'entités A*lignments and history* renseigne sur les informations liées aux AMS et aux historiques de l'activité de la base. En effet, si un récepteur est supprimé de la base, une trace écrite de la suppression sera sauvegardée. L'utilisation de *triggers* (*i.e.*, déclencheurs s'activant lors d'un événement particulier) sous MySQL faciliterait cette automatisation mais cette fonctionnalité n'était pas présente dans notre version de MySQL.

- le groupe d'entités *Web interface* ne possède aucune relation. Elle contient l'entité *user* qui permet de vérifier les informations utilisateurs lors de la connexion à l'interface et l'entité *helix_pos* qui regroupent les positions relatives des différentes HTM (utiles pour la récupération d'AMS).

- le groupe d'entités *References* associe chaque récepteur à sa provenance. Les informations des RCPG peuvent provenir de la base de données protéique Uniprot (la majorité), de la base de données génomique Genbank ou d'articles scientifiques (plutôt rare).

5.9.2 Interface AJAX

L'interface Web a été codée avec la version 5.2.3 du langage Hypertext Preprocessor (PHP) et la technologie Asynchronous Javascript And XML (AJAX) qui permet d'aborder les applications Web

sous une nouvelle approche. AJAX est un terme qui regroupe l'utilisation de technologies Web éprouvées : le JavaScript (JS), la norme Document Object Model (DOM) et l'objet XMLHttpRequest. Ce dernier permet de lancer des requêtes sur notre base de données de façon asynchrone, c'est à dire en arrière plan de la page et de façon complètement transparente pour l'utilisateur. En évitant l'actualisation systématique de la page, l'application gagne en fluidité. La version 3.0.0 de bibliothèque JS Ext a été utilisée pour améliorer la convivialité de l'interface. Suite à une requête, les résultats sont affichés dans un tableau dynamique car l'utilisateur peut choisir les informations qu'il souhaite afficher. La sélection d'un ou de plusieurs récepteurs du tableau permet d'appliquer différentes fonctionnalités (**Figure 15**) :

- *Get Anno* donne un ficher Comma-Separated Values (CSV) avec les informations affichées.

- *Get Seq* renvoie un fichier FAST-All (FASTA) avec les séquences correspondantes.

- *Edit Pro* permet de modifier les informations (annotations, classifications, motifs et séquence). Cette fonctionnalité n'est active que lorsqu'un récepteur est sélectionné.

- *Edit Group* autorise la modification simultanée de certaines informations pour plusieurs récepteurs. Cette fonctionnalité n'est active que lorsque plusieurs récepteurs sont sélectionnés.

- *Delete*, *Print* et *Reload* permettent de supprimer, d'imprimer le tableau entier et de recharger le tableau, respectivement.

Grâce à ces outils, l'équipe de bioinformatique dispose d'une base de données sur mesure et d'une interface Web permettant une récupération facilitée des informations. Elle a permis un gain de temps non négligeable et une cohérence de suivi des données pour passer d'une étude à une autre.

5.10 Package R bio2mds

R [102] (via http://www.r-project.org/) est à la fois un langage de programmation et un environnement pour la manipulation, le calcul et la création de représentations graphiques à partir

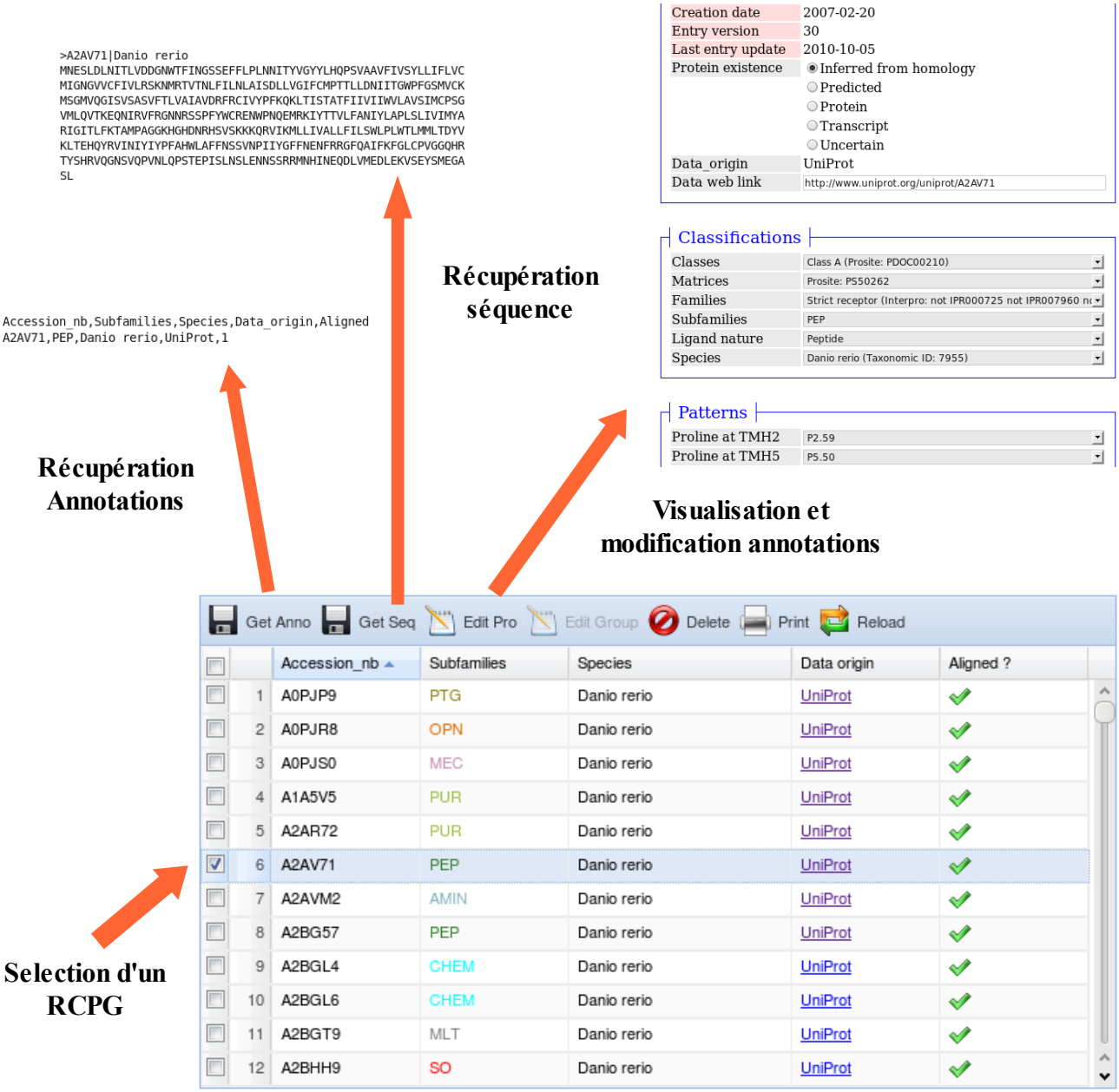

Figure 15 : Interface en AJAX

Le tableau interactif permet de sélectionner un ou des RCPG. Si l'utilisateur ne choisit qu'une seule protéine, plusieurs fonctionnalités deviennent disponibles. *Get Anno* donne un fichier CSV. *Get Seq* renvoie un fichier FASTA. *Edit Pro* permet de visualiser et de changer les attributions du RCPG en question. Une fois le formulaire validé, les changements sont directement répercutés sur la base de données. Également, l'interface AJAX permet de trier les RCPG en fonction des différentes colonnes et d'avoir accès à la fiche UniProt par un lien Web par exemple.

de données. Il peut être considéré comme une implémentation du langage S, développé aux laboratoires Bell dans les années 70 [103]. Le langage S n'est pas exclusif à R et est à la base, par exemple, de l'implémentation commerciale S-PLUS. R fut initialement développé en 1993 par R. Ihaka et R. Gentleman au Département de statistiques de l'Université d'Auckland (Auckland, Nouvelle-Zélande) [104]. R est impliqué dans le projet GNU's Not UNIX (GNU) et s'est davantage spécialisé dans les statistiques. Il s'est rapidement étoffé par le développement et la distribution d'un large ensemble de paquets. Il en existe deux types :

- les standards (environ 25). Ils sont développés par la R Development Core Team et font partie intégrante du code source de R. Ils contiennent les méthodes de base, statistiques et graphiques, nécessaires au bon fonctionnement de R.

- les contribués (plusieurs milliers). Ils sont développés par des personnes tierces, en majorité des chercheurs et sont disponibles notamment sur le site web du Comprehensive R Archive Network (CRAN) via http://cran.r-project.org/.

Ces paquets constituent une des forces principales de R et donnent accès généralement à des méthodes statistiques spécialisées et à des données supplémentaires concernant une thématique précise. R couvre des domaines variés, allant de l'écologie à la finance. Le domaine de la bioinformatique est dominé par le projet Bioconductor [105] (via http://www.bioconductor.org/) mais ne concerne presque qu'exclusivement les données génomiques ; seul le package *bgafun* s'applique aux données protéiques. Les paquets concernant les données protéiques et notre thématique de recherche sont assez peu nombreux. Voici un aperçu des plus notables :

- *seqinr* (« sequences in R ») : créé au Laboratoire de Biométrie et Biologie évolutive (Villeurbanne, France) et publié en 2007 [106], il contient des méthodes d'analyse et de visualisation des séquences protéiques (et génomiques). Il inclut également des utilitaires pour la gestion de données sous le système de recherche d'information ACides NUCléiques (ACNUC). Ce package est référencé sur le CRAN et de la documentation complémentaire est disponible via http://pbil.univ-lyon1.fr/software/seqinr/.

- *bio3d* : créé au Département de Chimie et de Biochimie de l'Université de Californie (San

Diego, États-Unis) et publié en 2006 [107], il contient des méthodes pour traiter, organiser et explorer des structures et des séquences protéiques. Par exemple, *bio3d* est capable de superposer des structures et, en y appliquant une PCA, d'examiner les relations entre les différentes conformations. Ce package n'est pas référencé sur le CRAN et il n'est disponible que via http://mccammon.ucsd.edu/~bgrant/bio3d/.

- *aaMI* : créé en 2005, il contient des méthodes pour le calcul des mutations corrélées. Le CRAN a récemment placé ce package dans ses archives.

Le package R *seqinr* est très complet mais est exempt de fonctions d'analyses exploratoires, *bio3d* se concentre davantage sur l'analyse des structures et *aaMI* est aujourd'hui obsolète. La MDS métrique est implémenté sous R (notamment avec la fonction `cmdscale` dans le package standard *stats*) mais ne gère pas la projection d'éléments supplémentaires. De plus, aucun package n'a encore implémenté la méthode DISTATIS. Le développement d'un package R au sein de notre équipe de bioinformatique permet de lier, à la fois, des données biologiques à des méthodes d'analyses exploratoires inédites sous R et cela constitue un triple intérêt :

- valoriser et mettre à disposition toutes les méthodes statistiques utilisées,
- respecter une certaine norme au niveau du code et de la documentation,
- partager un outil didactique avec les futurs membres de l'équipe.

Le package R *bio2mds* (« biological to mds ») a été développé et est prioritairement dédié à l'analyse des séquences protéiques par MDS métrique avec possibilité de projection d'éléments supplémentaires. Tous les détails qui concernent, à la fois, le package lui-même, les jeux de données et les fonctions sont présents dans la documentation R de *bio2mds*. *Ce* package n'est disponible qu'au laboratoire de l'équipe de bioinformatique. Le code source reste confidentiel car il n'a pas encore été publié. Des exemples simples sont présentés en annexes. La documentation technique du package R n'est pas disponible dans ce présent document.

6. RESULTATS

6.1 L'analyse de l'évolution des RCPG de classe A par MDS métrique

6.1.1 Article : A Novel Multidimensional Scaling Technique Reveals the Main Evolutionary Pathways of Class A G-Protein-Coupled Receptors

La difficulté de résoudre les relations qui existent entre les différentes sous-familles restant inhérent aux arbres phylogénétiques, cela nous a conduit à appliquer des méthodes statistiques originales sur notre jeu de données. De plus, la complexité de la structure hiérarchique des RCPG obtenue amène une approche alternative qui ne se base pas sur l'hypothèse d'une répartition forcement hiérarchique des données. Au cours de l'évolution, l'apparition progressive des différentes sous-familles, en lien avec les étapes évolutives majeures des espèces concernées, nous a poussé à formuler l'hypothèse que les RCPG de classe A pourraient évoluer, non pas de manière bifurquée, mais sur la base d'un système radial. La technique MDS a tout de suite attiré notre attention car elle n'utilise pas d'hypothèse hiérarchique et prend en considération les données brutes dans son ensemble. Dans un premier temps, nous avons appliqué la technique MDS sur les RCPG de classe A humain pour visualiser la manière dont les sous-familles se plaçaient les unes par rapport aux autres. Suite à un clustering par K-moyennes, les RCPG de classe A semblent se regrouper sous la forme de quatre groupes suivant les trois premières composantes principales. Pour suivre l'évolution, nous avons envisagé d'utiliser une méthode de projection permettant de superposer les positions des récepteurs d'une espèce par rapport à une autre. Ainsi, nous avons projeté les RCPG de classe A de différentes espèces plus ou moins éloignées sur l'espace humain, grâce à l'application inédite de la technique de projection d'éléments supplémentaires dans le cadre de la MDS. Dans un second temps, nous avons voulu mettre en évidence les spécificités de séquences principales qui permettent de différencier un groupe de récepteurs d'un autre. Comme nous n'avons pas trouvé de techniques explicites dans la littérature scientifique en bioinformatique, nous avons du adapter différentes techniques, dont des méthodes d'AMC, à notre stratégie de calcul. Certains résidus sont mis en évidence et seraient impliqués dans l'apparition des nouvelles sous-familles, notamment pour les motifs proline dans l'HTM2 et 5. Suite à cela, nous avons émis l'hypothèse d'un schéma évolutif des RCPG de classe A qui placerait la sous-famille PEP en tant que possible origine commune et qui mettrait en évidence l'implication de certains motifs d'HTM. Cette étude a été soumise dans Plos Computational Biology (30/08/10) qui a été refusée par les *reviewers*. Après les modifications demandées et en suivant les conseils de l'éditeur, elle a ensuite été soumise à Plos One (06/12/10) :

A Novel Multidimensional Scaling Technique Reveals the Main Evolutionary Pathways of Class A G-Protein-Coupled Receptors

Julien Pelé[a], Hervé Abdi[b], Matthieu Moreau[a], David Thybert[a,1] and Marie Chabbert[a,2]

[a] CNRS UMR 6214 – INSERM 771, Faculté de Médecine, 3 rue Haute de Reculée, 49045 ANGERS, FRANCE

[b] The University of Texas at Dallas, School of Behavioral and Brain Sciences. 800 West Campbell Road, Richardson, TX 75080-3021, USA

[1] Present address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

[2] To whom correspondence should be addressed. Dr M. Chabbert, CNRS UMR 6214 – INSERM U771, Faculté de Médecine, 3 rue Haute de Reculée, 49045 ANGERS, France. Tel: 33241735873, Fax: 33241735895, Email: marie.chabbert@univ-angers.fr.

Running Title: MDS analysis of GPCR evolution

Keywords: GPCR, Multidimensional scaling, Evolution, Sequence analysis

ABSTRACT

Class A G-protein-coupled receptors (GPCRs) constitute the largest family of transmembrane receptors in the human genome. Understanding the mechanisms which drove the evolution of such a large family would provide invaluable information on the specificity of each GPCR sub-family with applications to drug design. To gain evolutionary information on class A GPCRs, we explored their sequence space by metric multidimensional scaling analysis (MDS). Tri-dimensional mapping of human sequences shows a non-uniform distribution of GPCRs, organized in clusters that lay along four privileged directions. To interpret these directions, we developed a new MDS technique that projects supplementary sequences from different species onto the human space used as a reference. With this technique, we can easily monitor the evolutionary drift of several GPCR sub-families from cnidarians to humans. Results support a model of radiative evolution whose central node is formed by peptide receptors. The privileged directions obtained from the MDS analysis are interpretable in terms of three main evolutionary pathways linked to specific sequence determinants. The first pathway corresponds to the differentiation of the amine receptors. The second pathway was initiated by a deletion in transmembrane helix 2 (TM2) and led to three sub-families by divergent evolution. The third pathway corresponds to parallel evolution of several sub-families in relation with a covarion process involving proline residues in TM2 and TM5. As exemplified with GPCRs, the MDS projection technique helps decipher the evolutionary information hidden within orthologous sequence sets, revealing original details of the history of protein families.

INTRODUCTION

Proteins with a seven transmembrane helix scaffold are widespread in the animal kingdom and are usually assumed to be G-protein-coupled receptors (GPCRs) by similarity with their vertebrate counterparts. Because they transduce signals from a wide variety of chemical or physical stimuli, these receptors are involved in the perception by the cell of its environment and the regulation of most physiological functions [1]. Impaired GPCR signaling characterizes numerous pathologies of the cardiovascular, immune, neurological and metabolic systems. Consequently, GPCRs constitute major therapeutic targets for a wide spectrum of diseases and are subject to intensive investigation aimed at drug discovery.

GPCRs are classified into several classes whose common origin is still debated [2,3]. Within each class, however, receptors are clearly phylogenetically related and share conserved sequence patterns. With about 300 non-olfactory and 400 olfactory members, class A or rhodopsin-like GPCRs represent up to 90% of human GPCRs. Non-olfactory receptors can be further classified into a dozen of sub-families. However, the hierarchy of these sub-families is still unresolved and there is a strong discrepancy between the conclusions of different studies [2,4,5,6]. Understanding the mechanisms that led to the diversification of this family would help decipher the specificity of the sequence-structure-function relationships of each sub-family and would improve drug design targeted to GPCRs.

The phylogeny of a huge family of proteins such as GPCRs is far from obvious. Most current phylogenetic methods implicitely assume that the sequences can be classified according to a binary tree and try to reconstruct this tree. However, evolution may proceed either by bifurcation or by radiation. Radiative evolution, which should be described by polytomic trees, may account for discrepancies between binary trees [7,8]. In addition, evolution works on the sequence level, but proceeds under strong structural and functional constraints. As a consequence, selective pressure on a given amino acid may depend on the identity of amino acids at other sites, resulting in correlated mutations and/or branch specific changes in evolutionary rates [9,10,11]. This so-called covarion process may lead to misinterpretation of parallel/convergent evolution and is responsible of topological biases [12,13]. These difficulties inherent to phylogenetic methods prompted us to consider alternative methods to gain information on the relationships between GPCRs.

One such method is metric multidimensional scaling analysis (MDS) [14,15,16]. MDS is an exploratory multivariate procedure designed to identify patterns in a distance matrix. In this regard, when applied to sequences, MDS can be compared to neighbor joining (NJ) or UPGMA methods.

However, in NJ or UPGMA, sequences are considered by pairwise progression to establish a binary tree, whereas, in MDS, sequences are considered all at once, to determine a sequence space. In that case, sequences are represented, in a low-dimensional Euclidean space, by elements whose respective distances best approximate the original distances.

In this article, we use MDS to explore the sequence space of class A GPCRs. To interpret patterns in link with evolution, we developed a novel MDS technique that projects supplementary elements onto an active space (i.e., a space defined by the set of the data under scrutiny) [15]. Applied for the first time to protein sequences, this projection technique helps reveal the factors underlying the evolution of GPCRs.

RESULTS

1. The sequence space of human GPCRs

In *H. sapiens*, non-olfactory class A G-protein-coupled receptors (thereafter GPCRs) form a non-redundant set of 283 sequences that are referred to as the active sequence set. Most of these sequences (93%) can be classified into the twelve sub-families listed in Table I. From the multiple sequence alignment (MSA) of the active sequences, we computed a matrix of pairwise distances, based on sequence identity. Then, the distance matrix was analyzed by MDS, according to the procedure detailed in the Methods section. Briefly, MDS transforms the distance matrix **D** into a cross-product matrix **S** whose eigendecomposition provides a factor score matrix **F** (Fig. 1). This one gives the coordinates of the active sequences in the active space formed by the eigenvectors (also called principal components) of **S**.

We can map the sequence space of human GPCRs in two or three dimensions (Fig. 2) by using their projection onto the first two or three principal components (i.e. the components which explain the largest proportion of the variance of the original distances). The MDS representation clearly reveals the non-uniform distribution of human GPCRs. These receptors have a radial organization and cluster along a few privileged directions. This organization yields a straightforward classification of the receptors into four groups (named G0 to G3), at an intermediate level between the class and the sub-family levels (Table I).

The most central receptors cluster along the third component to form group G0. This central group includes the PEP, MTN, and OPN sub-families, with these latter two sub-families located on the edges of the group. Group G1 is characterized by positive coordinates on the first component. It

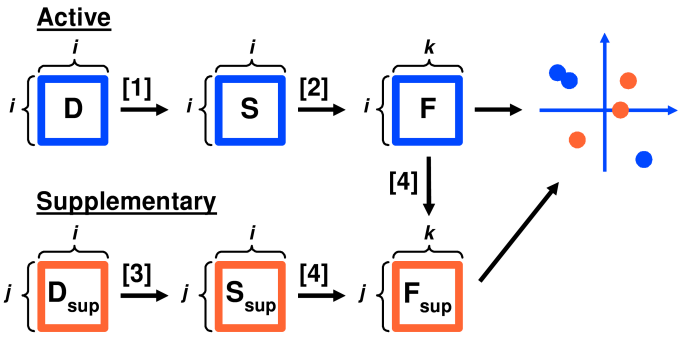

Fig. 1: Schematic representation of the MDS analysis. The analysis of N active and N_{sup} supplementary sequences are represented in blue and orange, respectively. D and D_{sup} represent distance matrices, S and S_{sup} cross-product matrices and F and F_{sup} factor score matrices. The coordinate of the i^{th} active sequence on the k^{th} principal component is directly obtained from the i^{th} element of the k^{th} column of F. The coordinate of the j^{th} supplementary sequence on the k^{th} principal component of the active space is directly obtained from the j^{th} element of the k^{th} column of F_{sup}. The numbers between brackets refer to the equations given in the Methods section.

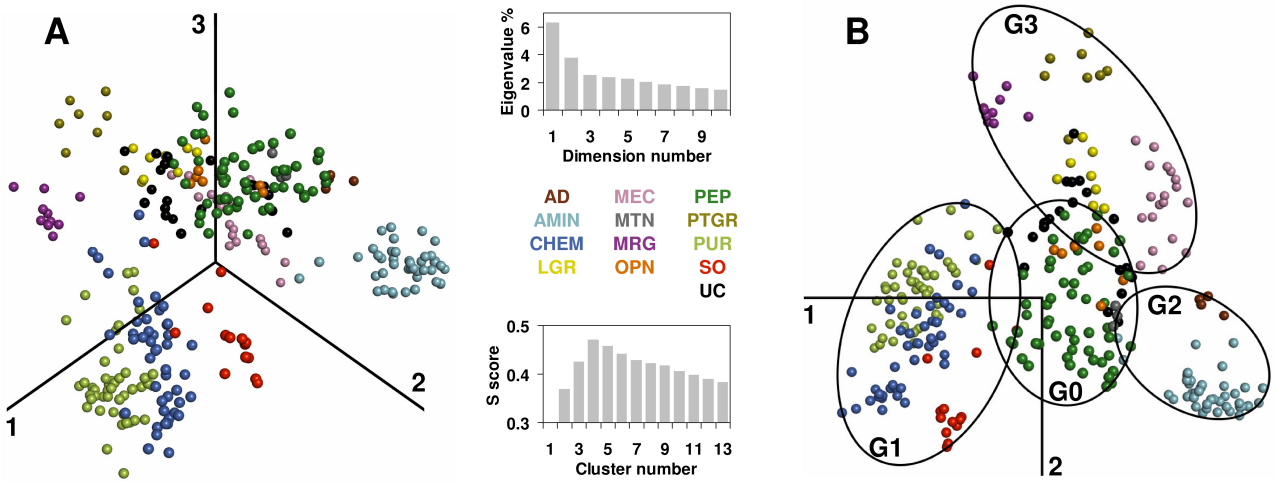

Fig. 2: MDS representation of human non-olfactory class A GPCRs. Data are projected onto the first three components (A) or onto the first two components (B). The inserts display the scree plot of the first ten eigenvalues (expressed as their percent to the total of the eigenvalues) (top) and the Silhouette score S as a function of the number of clusters (bottom). The color code refers to the GPCR sub-families (AD, brown; AMIN: cyan; CHEM: blue; LGR: yellow, MEC: pink; MTN: grey; MRG: violet; OPN: orange; PEP: dark green; PTGR: olive green; PUR: light green; SO: red; UC: black). Spanning ellipses are given for clarity purpose.

is composed of the SO, CHEM, and PUR sub-families which are phylogenetically related [6]. These three sub-families are separated in a 3D space by a combination of the three components. Group G2 is characterized by negative coordinates on the first component. It includes the AMIN and AD receptors. Finally, group G3 is spread along the second component with characteristic negative coordinates on this axis, and includes the LGR, PTGR, MRG and MEC sub-families.

This intuitive clustering based on visual inspection is corroborated by *K*-means analysis. The maximum of the Silhouette score [17] is reached for four clusters (Fig. 2, bottom insert), which correspond to the best description of the data. Receptors are attributed to the same clusters by *K*-means and visual inspection, except a few receptors (about 4%) located at the interface between two groups. For the forthcoming analysis, these receptors are assigned to the group including most members of their sub-family.

The only exception for the assignment of a sub-family to a single cluster is observed for the MECA (melanocortin, S1P, cannabinoid and adenosine) receptors. We and others considered these receptors as forming a single sub-family from phylogenetic data [2,6], but the MDS analysis clearly divides the MECA receptors into two subsets. The adenosine receptors (AD) cluster with the AMIN receptors, as observed in some phylogenetic studies [4,5], whereas the remaining receptors (MEC), whose coordinates on the second component are negative, cluster with group G3. Unclassified receptors (7% of the human set) cluster either with G0 or with G3.

The eigenvalues associated with the first three components explain respectively 6.3%, 3.8% and 2.6% of the total variance (Fig. 2, top insert). To assess the significance of this amount of explained variance, we decided to estimate the amount of variance due to random mutations. To do so, we generated 1000 random MSA with the same characteristics as the human MSA and analyzed the distribution of the eigenvalues obtained by MDS. The resulting distribution of eigenvalues indicates that only the first twenty-one components are significant (95% confidence level). They correspond to 39% of the total variance, to be compared to the 13% explained by the first three components. After the third component, however, the eigenvalues slowly decrease and lower ranking components are not interpretable. On the other hand, the first component clearly discriminates groups G1 and G2 from the remaining receptors, whereas groups G0 and G3 form a continuum, but do not overlap significantly on the second dimension (Fig. 2b). Most details are thus described by the first two components in agreement with the scree plot [18]. However, the third component improves the discrimination performance and clearly separates groups G0 and G3, providing a more detailed view of the GPCR space. Therefore we decided to keep the clustering

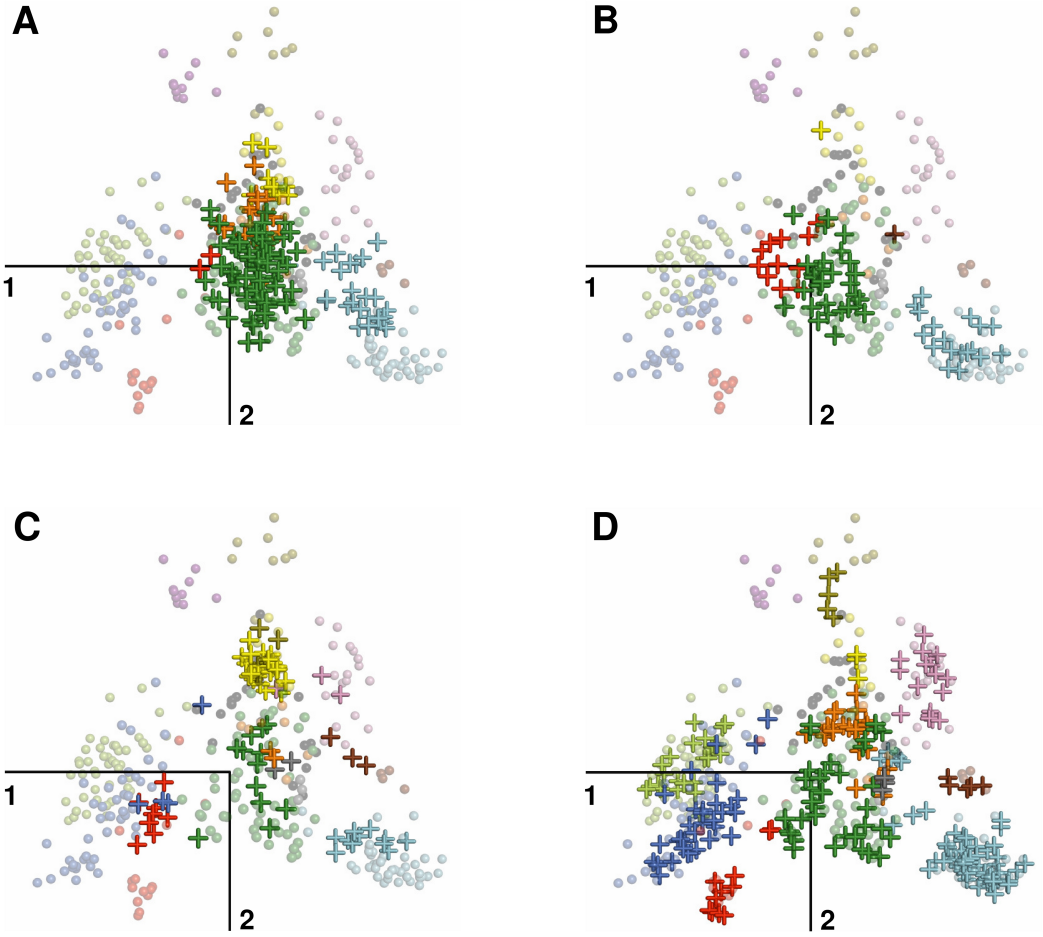

Fig. 3: Projection of supplementary GPCR sequences onto the sequence space of human GPCRs. GPCRs from *N. vectensis* (A), *C. elegans* (B), *C. intestinalis* (C) and *D. rerio* (D) were projected onto the first two components of the human active space. Transparent circles and crosses represent human and supplementary elements, respectively. The color code is defined in Fig. 2.

obtained from the 3D space for subsequent analyses.

2. Evolutionary drift of GPCRs

To understand the organization of the sequence space of human GPCRs, we developed an MDS technique that projects additional sets of sequences (referred to as supplementary sequences) onto the space of the active sequences analyzed by MDS (Fig. 1). As we are interested in the evolution of sub-families present in humans, supplementary sequences correspond to GPCRs from these sub-families in four selected species. These species have fully sequenced genomes and belong to the cnidarian (*N. vectensis*), nematode (*C. elegans*), chordate (*C. intestinalis*) and verbebrate (*D. rerio*) lineages. Five sub-families (PEP, AMIN, LGR, OPN and SO) are present from cnidarians to vertebrates whereas the other sub-families appeared in bilaterians (AD), chordates (MEC, PTGR, CHEM, MTN), vertebrates (PUR) and mammalian (MRG) [6,19]. Supplementary sequences were aligned against the MSA of human GPCRs and the matrix of distances between supplementary and active sequences was calculated from sequence identity. This supplementary distance matrix was transformed as described in the Methods section to obtain the coordinates of the supplementary sequences when they are projected onto the human active space.

The projection of supplementary GPCRs allows the straightforward monitoring of the evolutionary drift undergone by some sub-families while other sub-families remained stable (Fig. 3-4). The central position of the PEP receptors is maintained throughout species while no significant shift is observed for the OPN, LGR and MTN receptors. On the other hand, the drift of the AMIN receptors is obvious when comparing the position of this sub-family in *N. vectensis* and vertebrates. The drift of the SO receptors is still more striking because they move from the left side of G0 in *N. vectensis* and *C. elegans* to an intermediate position in *C. intestinalis* and to their final position in vertebrates (Fig. 3-4).

The first members of the CHEM sub-family appeared with chordates. In *C. intestinalis*, the members of the CHEM sub-family are not clearly separated from the SO receptors, neither by MDS analysis (Fig. 3) nor by phylogenetic analysis [6]. In vertebrates, the ancestral SO group diverged into three sub-families: “modern” SO, CHEM and PUR receptors. The position of these later ones suggests that they evolved from ancestors of CHEM receptors.

The AD receptors are close to G0 in *C. elegans* and progressively move towards the AMIN receptors in vertebrates. Interestingly, compared to the position of the single AD receptor from *C. elegans*, the AD and MEC receptors from *C. intestinalis* are translated along the first and second

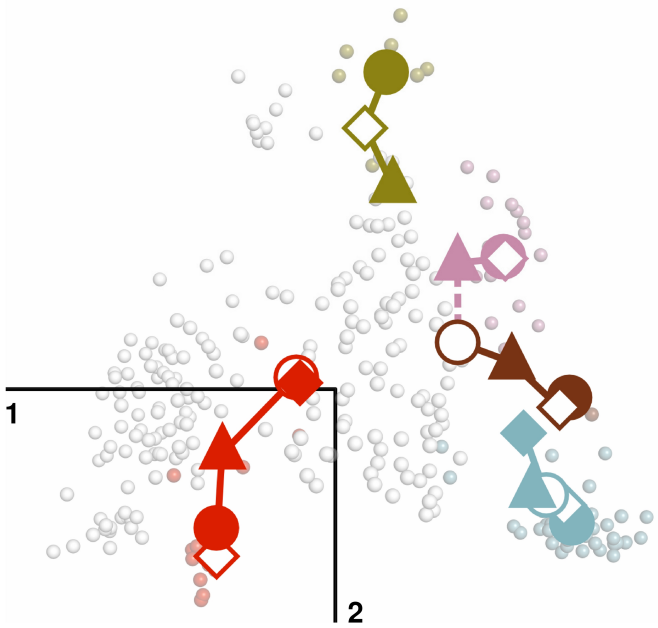

Fig. 4: Evolutionary drift of specific sub-families. The barycenters of the SO (red), AMIN (cyan), AD (brown), MEC (pink) and PTGR (olive green) sub-families are projected onto the first two components of the human **active** space. The symbol code indicates the species (*N. vectensis*: closed diamonds, *C. elegans*: open circles; *C. intestinalis*: closed triangles; *D. rerio*: open diamonds*; H. sapiens*: closed circles). The color lines joining the barycenters are given for clarity purpose. The magenta dashed line indicates the putative phylogenetic relationship between AD and MEC receptors.

components, respectively. Finally, the PTGR receptors shift along the second component from chordates to mammalians (Fig. 4).

It is worth noting that the evolution of orthologous sequences from the oldest ancestor common to an entire protein family can be decomposed into a shared part existing before speciation and a specific part originating after speciation. When sequences from one species are projected onto the sequence space of a reference species, this specific part is expected to be described by coordinates on high dimensions whereas the shared part should correspond to coordinates on the low dimension space of reference (i.e., to the position of the observed projected elements). This assumption is corroborated by the MDS analysis of GPCRs from the non-human sub-families present in *N. vectensis* and *C. elegans* whose projection onto the human space of reference overlaps group G0 (Fig. 5).

3. Sequence determinants of GPCR evolution

To search sequence determinants related to the evolutionary pathways observed by MDS, the entire sequence set was divided into four groups, according to the MDS classification of the human counterparts (Table I). Positions specific of each MDS group (Fig. 6) were searched for by plotting, for each position ℓ of the MSA, the frequency correlation, $FC(\ell)$, as a function of the difference of entropy, $\Delta S(\ell)$ (see Methods). The position numbering is based on Ballesteros' scheme [20]. The most conserved position in each transmembrane helix n (TMn) is numbered n.50 and is used as a relative reference.

A proline residue at position 2.58 in TM2 is the hallmark of G1 receptors. Present in SO receptors from *N. vectensis*, it is conserved in almost any G1 receptor [6]. The P2.58 pattern results from an indel (insertion/deletion) in TM2 [6] which appears as the key event yielding to the emergence of this group. Recently, this indel received experimental validation with the resolution of the crystal structure of CXCR4 [21]. An aliphatic residue is also highly conserved at position 2.57 as a result of the indel. On the other hand, position 3.37 presents interesting characteristics. This position is variable in SO receptors from *N. vectensis* and *C. elegans* whereas it corresponds to Tyr in chordate SO and vertebrate CHEM and PUR receptors and to Phe in vertebrate SO receptors. This suggests that this position might be crucial for the evolution and the diversification of G1 receptors.

Two positions, 3.32 and 7.40, are specific of the AMIN receptors whose weight overwhelms AD receptors in G2. Interestingly, position 3.32 corresponds to an Asp residue in any species,

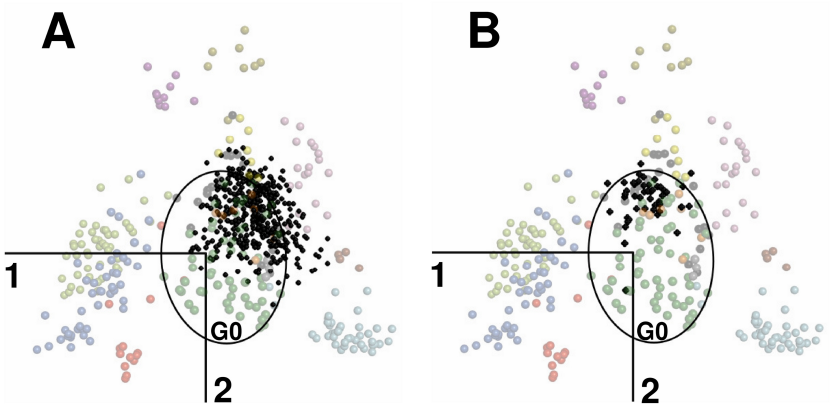

Fig. 5: Projection of GPCRs from non-human sub-families onto the sequence space of human GPCRs. GPCRs from *N. vectensis* (a) or *C. elegans* (b) that cannot be attributed to a sub-family present in humans were projected onto the first two components of the human active space. Projected elements (397 and 47 sequences from *N. vectensis* and *C. elegans*, respectively) are represented by black dots. Transparent circles, whose color code is defined in Fig. 2, represent human elements. The ellipse indicates G0 receptors.

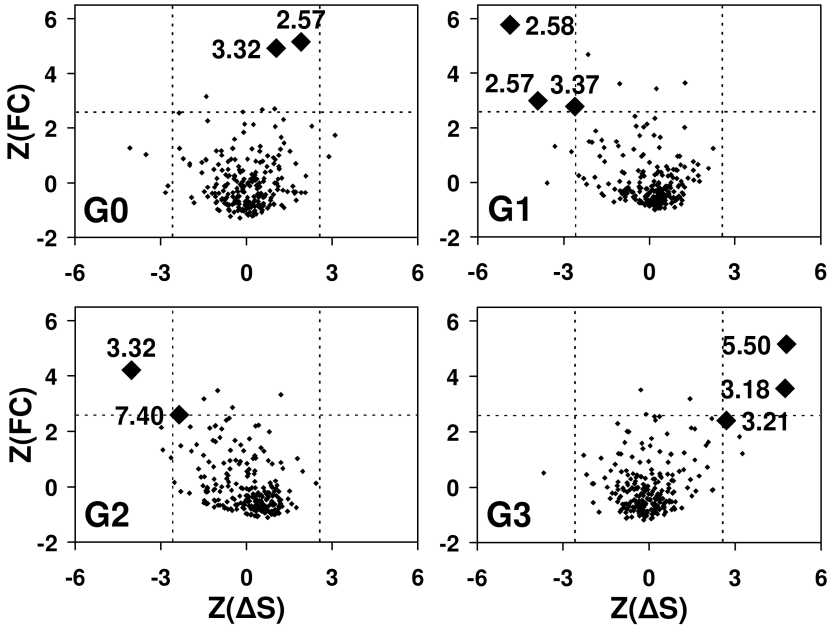

Fig. 6: Sequence analysis of the four receptor groups defined by MDS. For each group Gi (i = 0 to 3) and each position ℓ of the alignment, the Z-score of the correlation function, $FC(\ell)$, is plotted as a function of the Z-score of the entropy difference $\Delta S(\ell)$ between group Gi and its complement Gi^C. The dashed lines correspond to Z-scores of 2.58 (99% confidence level).

whereas position 7.40 is a highly conserved Trp in any species except in *N. vectensis*, suggesting that this position is important in the evolution of AMIN receptors.

Three positions are highly specific of G3 receptors. However, these positions are *variable* in G3, whereas they are highly conserved in the other GPCRs. The hallmark of G3 is the *absence* of the P5.50 proline residue in TM5 which is frequently associated with the absence of the W/FxxG motif at positions 3.18-3.21. In addition, the proline residues in TM2 and TM5 are not independent (p-values < 10^{-10} with the χ^2 test of independence) and the absence of proline in TM2 is also frequent in G3 (Table I). It is interesting to note that the drift of PTGR receptors along the second dimension is correlated with the partial loss of the TM2 proline in most recent species [6].

In contrast with the other groups, G0 does not possess hallmark residues. The positions with highest *FC*, 2.57 and 3.32, are only moderately conserved in G0 (28% Cys and 31% Gln, respectively) whereas they are highly conserved in G1 and G2, respectively. These positions, located within the extracellular side of the TM domain, face the receptor core and are ligand specific [22].

DISCUSSION

Introduced in the field of protein science more than 20 years ago [23], multidimensional scaling analysis was applied to the analysis of protein families [23,24,25,26,27] and of the protein fold space [28,29,30]. Though scarcely used, this method usefully complements phylogenetic techniques and provides important insights into the evolution of protein structural classes. In addition, compared to phylogenetic methods, MDS provides the possibility to project supplementary elements onto a reference space. This projection of supplementary elements has been used previously with principal component analysis [18] and is also routinely used with correspondence analysis [31,32]. The MDS projection technique has been developed for pattern recognition in cognitive sciences [15] and, to our knowledge, has never been applied previously to the field of protein evolution. In this paper, we show that this technique provides invaluable information on the evolution of protein families that is not reachable by classical phylogenetic analysis.

In the MDS representation of the GPCR sequence space, receptors are clustered along a few privileged directions (Fig. 2). Projection of receptors from supplementary species (Fig. 3) helps interpret these directions in terms of evolutionary trends that are corroborated by sequence analysis

Summary of the human GPCR set

Group	Sub-family	Description	Pro in TM2	TM2 Pro position	Pro in TM5	W/FxxG
G0	PEP	Peptide receptors	+++	**2.59**	+++	++
	OPN	Opsins	++	*2.59*	+++	+++
	MTN	Melatonin receptors	+++	**2.59**	+++	++
G1	SO	Somatostatin/opioid receptors	+++	**2.58**	+++	+++
	CHEM	Chemotactic receptors	+++	**2.58**	+++	+++
	PUR	Purinergic receptors	+++	**2.58**	+++	++
G2	AMIN	Amine receptors	+++	**2.59**	+++	++
	AD	Adenosine receptors	+++	**2.59**	+++	–
G3	LGR	Leucine-rich repeat receptors	–	–	–	Δ
	MEC	Melanocortin, S1P and cannabinoid receptors	–	–	–	–
	PTGR	Prostaglandin receptors	++	2.59	–	+
	MRG	MAS-related receptors	–	–	+	–

Table I: Human non-olfactory class A GPCRs were assigned to twelve sub-families according to the detailed classification reported in [6], except for the split of the MECA receptors into the AD and MEC sub-families. 7% of the human receptors could not be classified. Shaded rows indicate sub-families present from cnidarians to mammalians. The symbols indicate the percent of sequences with the pattern considered in human GPCRs (–, +, ++ and +++ correspond to 0%, 0 to 50%, 50 to 80% and ≥80%, respectively). Proline was searched for from position 2.58 to 2.60 in TM2 and at position 5.50 in TM5. The main proline position in TM2 is italic, normal and bold when it is observed in <50%, 50 to 80% and ≥80% of the sequences. Δ indicates that the W/FxxG motif is shifted to positions 3.19-3.22.

(Fig. 6). Several lines of evidence strongly suggest that the PEP sub-family is a central node of GPCR evolution. First, its central position is maintained from cnidarians to vertebrates (Fig. 3). Second, several sub-families (SO, AMIN, AD) are close to central PEP in the most distantly related species, then drift towards their final position in species most closely related to humans (Fig. 4). This is very striking for SO receptors whose vicinity to PEP receptors in non-chordate species corroborates our assumption of a common origin for these two sub-families [6]. Third, groups G1 to G3 are characterized by specific gain or loss of sequence patterns. This is not the case for group G0 (Fig. 6). Fourth, the loss of proline in TM2 and/or TM5 is characteristic of "recent" sub-families, such as the MEC, PTGR or MRG ones. This suggests that the LGR and OPN receptors may have evolved from an ancestor possessing proline residues in both helices whose PEP receptors might be the closest relative. This is consistent with the observation that substitutions from proline are more easily accommodated than substitutions to proline [33,34]. It should be added that there is no evidence of evolutionary linkage between prokaryotic and eukaryotic rhodopsins whose retinal-based photosensory system results from convergent evolution [35].

The sequence space of GPCRs indicates three main evolutionary pathways. The first one is related to the differentiation of the AMIN receptors, present in cnidarians, but whose drift is completed only in vertebrates (Fig. 3). AD receptors are part of this path, either by divergence from AMIN receptors or by convergence from PEP receptors. The second evolutionary pathway is related to an indel process in TM2 [6], leading to the split of P2.59 PEP and P2.58 SO receptors. The present data support the existence of a deletion mechanism that arose very early in GPCR evolution since P2.58 receptors are present in cnidarians. However, the differentiation of SO from PEP receptors was progressive. It involved further mutations (e.g. at position 3.37) and eventually led to "modern" SO, CHEM and PUR receptors by divergence (Fig. 3).

The hallmark of the third evolutionary pathway is the mutation of proline residues in TM2 and/or TM5 (Fig. 7), which is correlated with the mutation of the W/FxxG motif. However, the detailed analysis of these patterns (Table I) does not indicate a unique mechanism. The PTG and MRG sub-families provide an example of reverse order in the mutation of the TM2 and TM5 proline residues. The split of the AD and MEC sub-families, related to the mutation of both proline residues in MEC receptors, is subsequent to the mutation of the W/FxxG motif in AD receptors. This pattern suggests parallel evolution related to a covarion process in which the mutation of one of these sequence motifs releases structural and/or functional constraints and makes easier the subsequent mutation of the other motifs.

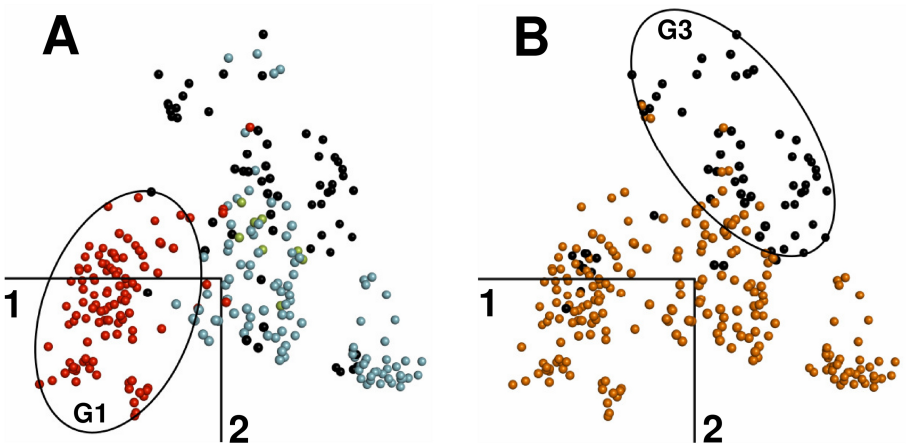

Fig. 7: Proline patterns of TM2 and TM5 from human GPCRs. In (A), receptors with a proline residue at position 2.58, 2.59 or 2.60 in TM2 are red, cyan or light green, respectively. Receptors with no proline in TM2 are black. The ellipse indicates G1 receptors. In (B), receptors with and without a proline residue at position 5.50 in TM5 are orange and black, respectively. The ellipse indicates G3 receptors.

Taken together, our results support a mechanism of radiative evolution from the ancestors of the PEP receptors. This mechanism is consistent with several evolutionary trees obtained by NJ or maximum parsimony (MP) methods for human and non-human species (dog, rat, pufferfish) that display a fan shape with sub-families from G1 on one hand and the AMIN sub-family on the other hand [6,36,37,38]. In particular, this model is supported by the full consensus tree for rat and human GPCRs obtained from both NJ and MP analysis in which the position of the OPN, MRG, PTGR and LGR sub-families is ambiguous [36]. It should be added that a classification of human GPCRs into four groups by MP [2] has enlightened the specificity of PEP receptors as a group. The discrepancy observed for the other groups might be explained by biases due to long branch attraction and/or to parallel evolution.

It is worth noting that two of the main pathways of GPCR diversification are related to proline residues in transmembrane helices (Fig. 7). Proline residues induce helical distortions that are key elements of GPCR structure and mechanism of activation. In particular, structural divergence between receptors may relate to the presence of proline [39] whereas the wobbling motion of TM6, at the level of a highly conserved proline, is a crucial step of rhodopsin activation [40,41]. We have previously proposed that the deletion in TM2 modifies the distortion of this helix from a bulge to a “typical” proline kink [6]. This structural change is now experimentally validated [21]. How it affects the mechanism of receptor activation remains to be determined. However, it is interesting to note that a rotational motion of TM2, reminiscent of TM6, has been observed in the type I angiotensin II receptor which belong to group G1 [42]. Along with the TM2 proline, the TM5 proline appears as a major vector of GPCR evolution. The concomitant loss of several sequence patterns observed in independent sub-families of group G3 is indicative of a covarion process and suggests a release of constraints that might be related to structural and functional changes.

In conclusion, MDS is especially suited for the analysis of large families, such as GPCRs. The projection technique allows a straightforward and spectacular visualization of the drift of several sub-families during evolution and provides a unique opportunity to decipher evolutionary pathways of protein families, in particular in the case of radiative evolution. In addition, MDS emphasizes the usefulness of rare mutational events as indels or mutations of residues with strong structural and/or functional constraints to infer the evolution of protein families.

METHODS

Sequences of class A GPCRs. The non-redundants sets of non-olfactory class A GPCRs from *C. elegans*, *C. intestinalis*, *D. rerio* and *H. sapiens* (109, 90, 236 and 283 sequences, respectively) correspond to the previously determined sets [6], updated with the July 2009 release of Uniprot when necessary. 93% of the human receptors can be assigned to twelve sub-families (Table I), whereas 7% of them remain unclassified (UC). The sub-family nomenclature is adapted from Fredriksson's classification [2]. The ratio of sequences assigned to these twelve sub-families is 57, 87, and 95% for *C. elegans*, *C. intestinalis,* and *D. rerio*, respectively. The sequence set of class A GPCRs from *N. vectensis* was prepared from the July 2009 release of Uniprot, according to the procedure previously described [6]. It is composed of 538 non-redundant (identity < 90%), non-olfactory sequences, 22% of which could be assigned to GPCR sub-families present in humans. The remaining sequences belong to GPCR sub-families specific of cnidarians [43]. The accession numbers of the sequences used for this study are given in the Supplementary Data.

Multiple sequence alignments were carried out with ClustalX (http://www.clustal.org/) and manually refined with GeneDoc (http://www.nrbsc.org/gfx/genedoc/) to insure that the residue anchor of each helix was correctly aligned. The anchor residues corresponding to the most conserved positions are N1.50, D2.50, R3.50, W4.50, P5.50, P6.50 and P7.50 (Ballesteros' numbering [20]). For the less conserved TM5, we used either P5.50 or Y5.58 to insure correct alignment. Sequence analyses were carried out on the MSA positions with less than 2% gaps. These 236 positions include the seven transmembrane helices, the putative eighth intracellular helix and parts of the intracellular and extracellular loops.

Multidimensional scaling analysis. When a set of sequences (referred to as active sequences) are aligned, a distance between each pair of sequences can be calculated from the MSA. The matrix of the pairwise distances can then be analyzed by MDS [15,16]. Formally, if we denote by N the number of sequences, by $\mathbf{D}$ the N by N the matrix of the squared distance between sequences, by $\mathbf{I}$ the N by N identity matrix, and by $\mathbf{1}$ an N by N matrix of ones, the first step is to transform the distance matrix $\mathbf{D}$ into a cross-product matrix denoted $\mathbf{S}$ and computed as:

$$\mathbf{S} = -0.5\,[\mathbf{I} - (1/N)\mathbf{1}] \times \mathbf{D} \times [\mathbf{I} - (1/N)\mathbf{1}]\,. \qquad [1]$$

The eigendecomposition of $\mathbf{S}$ expresses this matrix as the product of the eigenvector matrix $\mathbf{U}$ by the diagonal matrix of the eigenvalues $\mathbf{\Lambda}$ (such as $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{T}$, where ^{T} denotes the transposition operation). The eigenvectors of $\mathbf{S}$, or principal components, form the active space. The factor score

matrix, denoted **F,** is computed as:

$$\mathbf{F} = \mathbf{U}\mathbf{\Lambda}^{1/2} \qquad [2]$$

and gives the coordinates of the active elements in the active space. The eigenvalue of a component represents the variance explained by this component and the ratio of the eigenvalue of a component to the total of the eigenvalues gives the proportion of the data variance explained by this component.

Additional sequences are projected onto the active space as supplementary elements [15], according to the procedure summarized in Fig. 1. First, supplementary sequences are aligned against the active MSA, resulting into a supplementary matrix of distances between the supplementary and active sequences. Then, the supplementary distance matrix is transformed into a supplementary cross-product matrix which is in turn transformed into a factor matrix (Fig. 1). Specifically, if we denote N_{sup} the number of supplementary sequences, $\mathbf{1}_{sup}$ an N_{sup} by N matrix of ones, and $\mathbf{D}_{sup}$ the supplementary squared distance matrix, then the first step is to transform $\mathbf{D}_{sup}$ into a cross product matrix denoted $\mathbf{S}_{sup}$ as:

$$\mathbf{S}_{sup} = -0.5\,[\mathbf{I} - (1/N)\mathbf{1}]\,(\mathbf{D}_{sup}^{T} - (1/N)\mathbf{D}\,\mathbf{1}_{sup}) . \qquad [3]$$

The factor matrix for the supplementary sequences, denoted $\mathbf{F}_{sup}$, is computed as:

$$\mathbf{F}_{sup} = \mathbf{S}_{sup}^{T}\mathbf{F}\mathbf{\Lambda}^{-1} \qquad [4]$$

and gives the coordinates of the supplementary elements in the active space.

The simplest pairwise distance is given by the proportion of sites that differ between the two sequences [44]. It yields a distance very close to an Euclidian distance, because the eigendecomposition of the matrix based on this distance gives a small proportion of negative eigenvalues representing only 3% of the sum of absolute eigenvalues. Distances based on generic or transmembrane specific scoring matrices [45] do not perform as well with negative eigenvalues representing from 4 to 10% of the sum.

Receptor clustering. Following MDS, receptors were mapped in a 3D space and clustered by *K*-means analysis. The *K*-means procedure was reiterated 1000 times with random centroids. The most frequent clustering, in agreement with visual inspection, was selected and used as a reference to assess the reproducibility of the analysis. More than 97% of the receptors were assigned to the same reference clusters in more than 85% of all iterations. The Silhouette score [17] was calculated from 1000 iterations with the number of clusters ranging from 1 to 13 (for the 12 sub-families and UC receptors).

Sequence analysis. When a sequence set is divided into a subset g and its complement g^C, the correlation between a position ℓ of the MSA and the subsets is measured by the frequency correlation $FC(\ell)$, derived from the χ^2 test [46], according to the formula:

$$FC(\ell) = \sum [f(g) \times (f_i(\ell,g) - f_i(\ell))^2 + f(g^C) \times (f_i(\ell,g^C) - f_i(\ell))^2] / f_i(\ell) \qquad [5]$$

where $f(g)$ and $f(g^C)$ are the frequencies of g and g^C, respectively, and $f_i(\ell)$, $f_i(\ell,g)$ and $f_i(\ell,g^C)$ are the frequencies of amino acid i at position ℓ in the entire set, in g and in g^C, respectively. $FC(\ell)$ varies from 0 for totally variable positions to 1 for positions fully correlated with the subsets. In addition, the difference of sequence entropy [47] between g and g^C is given by:

$$\Delta S(\ell) = \sum f_i(\ell,g) \ln f_i(\ell,g) - \sum f_i(\ell,g^C) \ln f_i(\ell,g^C). \qquad [6]$$

Specific conservation or variability in the subset g corresponds to negative and positive values of ΔS, respectively. Sequence determinants of g are searched for by plotting the Z-scores of $FC(\ell)$ as a function of the Z-scores of $\Delta S(\ell)$.

ACKNOWLEDGEMENTS: We thank NEC Computers Services SARL (Angers, France) for the kind provision of a multiprocessor server. We thank C. Raimbault (Angers) for her help in the preparation of the sequence sets. We thank Dr D. Henrion (Angers) for continuous support and stimulating discussions. J. P. is supported by a fellowship from the Conseil Général du Maine-et-Loire. M. M. was supported by a fellowship from CNRS.

REFERENCES

1. Gether U (2000) Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocr Rev 21: 90-113.
2. Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63: 1256-1272.
3. Kolakowski LF, Jr. (1994) GCRDb: a G-protein-coupled receptor database. Receptors Channels 2: 1-7.
4. Surgand JS, Rodrigo J, Kellenberger E, Rognan D (2006) A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. Proteins 62: 509-538.
5. Vassilatis DK, Hohmann JG, Zeng H, Li F, Ranchalis JE, et al. (2003) The G protein-coupled receptor repertoires of human and mouse. Proc Natl Acad Sci U S A 100: 4903-4908.
6. Deville J, Rey J, Chabbert M (2009) An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors. J Mol Evol 68: 475-489.
7. Rokas A, Carroll SB (2006) Bushes in the tree of life. PLoS Biol 4: e352.
8. Rokas A, Kruger D, Carroll SB (2005) Animal evolution and the molecular signature of

radiations compressed in time. Science 310: 1933-1938.
9. Fitch WM (1971) Rate of change of concomitantly variable codons. J Mol Evol 1: 84-96.
10. Studer RA, Robinson-Rechavi M (2010) Large-scale analysis of orthologs and paralogs under covarion-like and constant-but-different models of amino acid evolution. Mol Biol Evol 27: 2618-2627.
11. Tuffley C, Steel M (1998) Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147: 63-91.
12. Susko E, Inagaki Y, Roger AJ (2004) On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Mol Biol Evol 21: 1629-1642.
13. Wang HC, Susko E, Spencer M, Roger AJ (2008) Topological estimation biases with covarion evolution. J Mol Evol 66: 50-60.
14. Togerson WS (1958) Theory and methods of scaling. New York: Wiley.
15. Abdi H (2007) Metric multidimensional scaling. In: Salkind NJ, editor. *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage. pp. 598-605.
16. Takane Y, Jung S, Oshima-Takane Y (2009) Multidimensional scaling. In: Millsap R, Maydeu-Olivares A, editors. Handbook of quantitative methods in psychology. London: Sage Publications. pp. 219-242.
17. Rousseeuw P (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. J Comput Appl Math 20: 53-65.
18. Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdisciplinary reviews: Computational Statistics 2: 433-459.
19. Fredriksson R, Schioth HB (2005) The repertoire of G-protein-coupled receptors in fully sequenced genomes. Mol Pharmacol 67: 1414-1425.
20. Sealfon SC, Chi L, Ebersole BJ, Rodic V, Zhang D, et al. (1995) Related contribution of specific helix 2 and 7 residues to conformational activation of the serotonin 5-HT2A receptor. J Biol Chem 270: 16683-16688.
21. Wu B, Chien EY, Mol CD, Fenalti G, Liu W, et al. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. Science.
22. Ye K, Lameijer EW, Beukers MW, Ijzerman AP (2006) A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors. Proteins 63: 1018-1030.
23. Woolley KJ, Athalye M (1986) A use for principal coordinate analysis in the comparison of protein sequences. Biochem Biophys Res Commun 140: 808-813.
24. Higgins DG (1992) Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. Comput Appl Biosci 8: 15-22.
25. Casari G, Sander C, Valencia A (1995) A method to predict functional residues in proteins. Nat Struct Biol 2: 171-178.
26. Gogos A, Jantz D, Senturker S, Richardson D, Dizdaroglu M, et al. (2000) Assignment of enzyme substrate specificity by principal component analysis of aligned protein sequences: an experimental test using DNA glycosylase homologs. Proteins 40: 98-105.
27. Lu F, Keles S, Wright SJ, Wahba G (2005) Framework for kernel regularization with application to protein clustering. Proc Natl Acad Sci U S A 102: 12332-12337.
28. Hou J, Jun SR, Zhang C, Kim SH (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. Proc Natl Acad Sci U S A 102: 3651-3656.
29. Hou J, Sims GE, Zhang C, Kim SH (2003) A global representation of the protein fold space. Proc Natl Acad Sci U S A 100: 2386-2390.

30. Choi IG, Kim SH (2006) Evolution of protein structural classes and protein sequence families. Proc Natl Acad Sci U S A 103: 14056-14061.
31. Greenacre M (2007) Correspondance analysis in practice. London: Chapman & Hall/CRC
32. Murtagh F (2005) Correspondence Analysis and data Coding with R and Java. London: Chapman & Hall/CRC.
33. Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU (2004) The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc Natl Acad Sci U S A 101: 959-963.
34. Yohannan S, Yang D, Faham S, Boulting G, Whitelegge J, et al. (2004) Proline substitutions are not easily accommodated in a membrane protein. J Mol Biol 341: 1-6.
35. Rompler H, Staubert C, Thor D, Schulz A, Hofreiter M, et al. (2007) G protein-coupled time travel: evolutionary aspects of GPCR research. Mol Interv 7: 17-25.
36. Gloriam DE, Fredriksson R, Schioth HB (2007) The G protein-coupled receptor subset of the rat genome. BMC Genomics 8: 338.
37. Haitina T, Fredriksson R, Foord SM, Schioth HB, Gloriam DE (2009) The G protein-coupled receptor subset of the dog genome is more similar to that in humans than rodents. BMC Genomics 10: 24.
38. Metpally RP, Sowdhamini R (2005) Genome wide survey of G protein-coupled receptors in Tetraodon nigroviridis. BMC Evol Biol 5: 41.
39. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, et al. (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 318: 1258-1265.
40. Park JH, Scheerer P, Hofmann KP, Choe HW, Ernst OP (2008) Crystal structure of the ligand-free G-protein-coupled receptor opsin. Nature 454: 183-187.
41. Scheerer P, Park JH, Hildebrand PW, Kim YJ, Krauss N, et al. (2008) Crystal structure of opsin in its G-protein-interacting conformation. Nature 455: 497-502.
42. Domazet I, Holleran BJ, Martin SS, Lavigne P, Leduc R, et al. (2009) The second transmembrane domain of the human type 1 angiotensin II receptor participates in the formation of the ligand binding pocket and undergoes integral pivoting movement during the process of receptor activation. J Biol Chem 284: 11922-11929.
43. Anctil M, Hayward DC, Miller DJ, Ball EE (2007) Sequence and expression of four coral G protein-coupled receptors distinct from all classifiable members of the rhodopsin family. Gene 392: 14-21.
44. Nei M, Zhang J (2005) Evolutionary Distance: Estimation. Encyclopedia of Life Sciences. Chichester (UK): John Wiley & Sons, Ltd. pp. 1-4.
45. Grishin VN, Grishin NV (2002) Euclidian space and grouping of biological objects. Bioinformatics 18: 1523-1534.
46. Kass I, Horovitz A (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins 48: 611-617.
47. Valdar WS (2002) Scoring residue conservation. Proteins 48: 227-241.

6.1.2 Conclusion

Cet article montre une classification alternative des RCPG de classe A humains par rapport aux arbres phylogénétiques. L'application d'une MDS métrique sur l'ensemble de ces récepteurs permet de les classer en quatre groupes bien distincts :

- le groupe central G0 qui comprend les sous-familles PEP, MTN et OPN ; les deux dernières se situant à la périphérie du groupe,

- le groupe G1 qui est composé des sous-familles SO, CHEM et PUR ; ces sous-familles sont reliées de manière phylogénétique [85],

- le groupe G2 qui réunit les sous-familles AMIN et des récepteurs d'adénosines (AD). En effet, la sous-famille MECA semble se séparer en deux avec les MEC d'un coté et les AD de l'autre,

- le groupe G3 qui rassemble les sous-familles LGR, PTGR, MRG et MEC.

Le nombre de groupe fut confirmé par silhouette score et l'affiliation de tel récepteur à tel groupe fut validé par K-moyennes, aidant notamment à l'association des récepteurs se situant aux l'interfaces à leurs groupes respectifs. Pour suivre l'évolution de ces différents groupes, la projection de RCPG de classe A d'espèces sur l'espace humain a été effectuée grâce à notre nouvelle technique statistique. Les séquences projetées doivent refléter une certaine dynamique évolutive et correspondent aux RCPG d'un cnidaire (*N. vectensis*), d'un nématode (*C. elegans*), d'un chordé (*C. intestinalis*) et d'un vertébré (*D. rerio*). Cinq sous-familles (PEP, AMIN, LGR, OPN et SO) sont communes aux cinq espèces alors que les autres sous-familles sont spécifiques des espèces les plus récentes. L'application de la nouvelle technique de projection permet de mettre en évidence la dérive évolutive de certaines sous-familles alors que d'autres restent stables tout au long de l'évolution :

- pour les sous-familles communes :

    - la sous-famille PEP est centrale et reste très stable. Cette observation est également vraie pour les sous-familles OPN et LGR qui ne présentent pas de changements de localisation vraiment notables.

    - Les sous-familles AMIN et SO affichent quant à elles une importante dérive évolutive. En effet, leurs récepteurs se retrouvent proches de la sous-famille PEP au niveau du cnidaire et retrouvent la localisation finale des humains qu'à partir des vertébrés.

- pour les nouvelles sous-familles :

    - la sous-famille CHEM n'apparaît qu'à partir du chordé et ses récepteurs ne sont pas clairement séparés à ce stade de la sous-famille SO. La séparation s'opère au niveau des vertébrés, simultanément à l'apparition de la sous-famille PUR. On peut émettre l'hypothèse qu'une sous-famille SO ancestrale pourrait être à l'origine des trois sous-familles SO, CHEM et PUR actuelles.

    - la sous-famille AD est proche du groupe G0 pour le nématode et se déplace progressivement, en se rapprochant de la sous-famille AMIN chez les vertébrés. La position singulière de l'unique récepteur des AD chez le nématode est à mettre en relation avec les localisations des sous-familles AD et MEC, cette dernière n'apparaissant qu'à partir du chordé.

    - les sous-familles PTGR et MTN apparaissent au même moment que les sous-familles MEC et CHEM chez le chordé. Cette concordance est remarquable car elle montre l'importance de cette étape évolutive chez les RCPG. La sous-famille PTGR montre une dérive évolutive non négligeable, en s'éloignant progressivement des LGR. La sous-famille MTN apparaît avec les chordés et reste stable.

    - la famille MRG n'apparaît qu'au moment des mammifères (avec la projection des récepteurs de *M. musculus*, par exemple).

Les calculs de la différence d'entropies et de la FC ont permis d'identifier les résidus clés de chaque groupe en lien avec leurs évolutions respectives :

- la proline se situant en position 2.58 représente le marqueur prépondérant, et de loin, du G1. Ce motif résulte d'un indel dans la TM2 [85] et cet évènement majeur concorde avec le rôle clé des motifs proline dans les cassures d'hélices.

- l'acide aspartique en position 3.32 et le tryptophane en position 7.40 semblent jouer un rôle important pour le G2, dominé par la sous-famille AMIN.

- les résultats obtenus pour le G3 sont particuliers car les positions intéressantes sont conservées pour les récepteurs n'appartenant pas au groupe G3. L'observation principale concerne l'absence de proline en position 5.50, qui est fréquemment associée à l'absence de motif W/FxxG aux positions 3.18 et 3.21, positions également révélées par le graphique. Cela met d'autant plus en exergue le fait que les prolines, jouant déjà un rôle clé dans les cassures d'hélices avec la proline dans l'HTM2, sont des marqueurs déterminants pour les RCPG de classe A.

- le G0 possède quelques positions spécifiques mais elles ne sont pas très bien conservées. Toutefois, il est à noter que ces positions se retrouvent impliquées dans la reconnaissance du ligand notamment.

Pour résumé, nous pouvons dire que l'étude des sous-familles des RCPG de classe A est relativement complexe. Au cours de l'évolution, les sous-familles se développent différemment suivant l'embranchement de classification concerné. La MDS a révélé trois chemins évolutifs majeurs :

- le premier concerne la différenciation des AMIN. La sous-famille AD y est impliquée, soit par divergence des AMIN ou convergence des PEP.

- le deuxième est lié à l'indel dans l'HTM2, séparant la sous-famille PEP, avec une proline en 2.59, de la sous-famille SO, avec une proline en 2.58.

- le troisième est la mutation du résidus proline dans TM2 et /ou TM5, amenant à son tour la mutation du motif W/FxxG.

Les techniques de MDS et de projection d'éléments supplémentaires sont bien adaptées lorsqu'il faut analyser une famille protéique affichant de très fortes différenciations et des relations entre sous-familles complexes. Ces méthodes servent également à révéler des informations qui restent inaccessibles par des méthodes classiques d'arbres phylogénétiques. Par exemple, le fait que les prolines jouent un rôle central dans l'évolution des sous-familles n'était que partiellement visible avec les arbres phylogénétiques. Les résultats suggèrent que les RCPG de classe A semblent liés à un mécanisme d'évolution radiale et cela n'est pas visible directement avec les arbres phylogénétiques basés sur une hypothèse d'évolution bifurquée. De plus, la facilité de mise en place des méthodes statistiques utilisées pour cet article et leur disponibilité dans le package R *bio2mds* permettent d'envisager leur utilisation pour d'autres familles protéiques.

6.2 Application des méthodes de mutations corrélées aux RCPG de classe A

Nous voulons étudier les réseaux de corrélation entre résidus pour différents niveaux de classification des RCPG qui ont des caractéristiques de taille et de conservation différentes. Les différentes méthodes d'AMC, qui ont des sensibilités spécifiques, ont été appliquées à nos données, à savoir l'AMS des RCPG de classe A humains et nous avons voulu savoir quelle méthode était la plus robuste par rapport aux différents niveaux de classification des RCPG de classe A : la classe, le groupe et la sous-famille.

6.2.1 Caractéristiques des trois jeux de données

Dans un premier temps, nous avons sélectionné trois jeux de données aux caractéristiques différentes pour pouvoir analyser la sensibilité et la robustesse des méthodes d'AMC :

- tous les RCPG de classe A humains (283 récepteurs), c'est à dire l'ensemble des 12 sous-familles et UC.

- le groupe G1 (107 récepteurs, taille d'environ 40% par rapport au nombre de RCPG de classe A humains), c'est à dire uniquement les sous-familles SO, CHEM et PUR. Pour rappel, le groupe G1 contient nos récepteurs d'intérêt.

- la sous-famille CHEM classique (23 récepteurs, taille d'environ 8% par rapport au nombre de RCPG de classe A humains), c'est à dire contentant seulement les récepteurs aux chimiokines. Les récepteurs de l'angiotensine, de la relaxine, de la bradykinine et de l'apéline ainsi que les récepteurs orphelins sont écartés.

6.2.2 Analyse par entropie

Les différents niveaux de données possèdent des tailles différentes et le nombre de séquences

baisse au fur et à mesure que le niveau de classification décroît. Plus le nombre de séquences de l'AMS augmente, plus la probabilité d'avoir une forte variabilité croit potentiellement. Cela va être visible par le calcul de la valeur d'entropie. De plus, on souhaite tester des spécificités de conservation différentes entre les données pour disposer d'un large panel de valeurs d'entropie et pour permettre d'observer les divergences de comportement entre les différentes méthodes d'AMC. La figure 16 montre un exemple de calcul d'entropie pour trois hélices caractéristiques. Le calcul d'entropie est basé sur le logarithme base 20 pour maintenir les valeurs entre 0 et 1. En général, on peut observer pour une position donnée que l'entropie des RCPG de classe A humains est supérieure à celle du groupe G1, qui est à son tour est supérieure à celle de la sous-famille CHEM classique. Cela est confirmé par le calcul des médianes des entropies : 0,71 pour les RCPG de classe A humains, 0,62 pour le groupe G1 et 0,43 pour la sous-famille CHEM classique. Certaines positions ont cependant des profils atypiques (**Figure 16**). Par exemple :

- la position 2.39 est conservée pour la sous-famille CHEM classique mais pas pour les 2 autres ensembles. Cette position semble donc particulière pour cette sous-famille.

- la position 2.58 est très conservée pour la sous-famille CHEM classique et le groupe G1. Une proline en position 2.58 constitue en effet la propriété de séquence caractéristique des récepteurs du groupe G1, composé des sous-familles SO, CHEM et PUR.

- la position 4.50 est très conservée quel que soit l'ensemble. Cette position correspond à un tryptophane qui est presque toujours présent parmi tous les RCPG de classe A humains.

- la position 6.47 est assez particulière car elle est moins conservée dans la sous-famille CHEM classique que dans les deux autres ensembles, malgré un nombre de récepteurs bien inférieur. Cela est dû à la conservation moindre de la cystéine à cette position dans les récepteurs des chimiokines de la famille CC.

Ces jeux de données possèdent des caractéristiques globales, dues aux nombres de séquences et des caractéristiques locales, au niveau de certaines positions. Toutes ces caractéristiques vont permettre de tester les différentes méthodes d'AMC pour voir comment elles se comportent sous différentes conditions d'entropies.

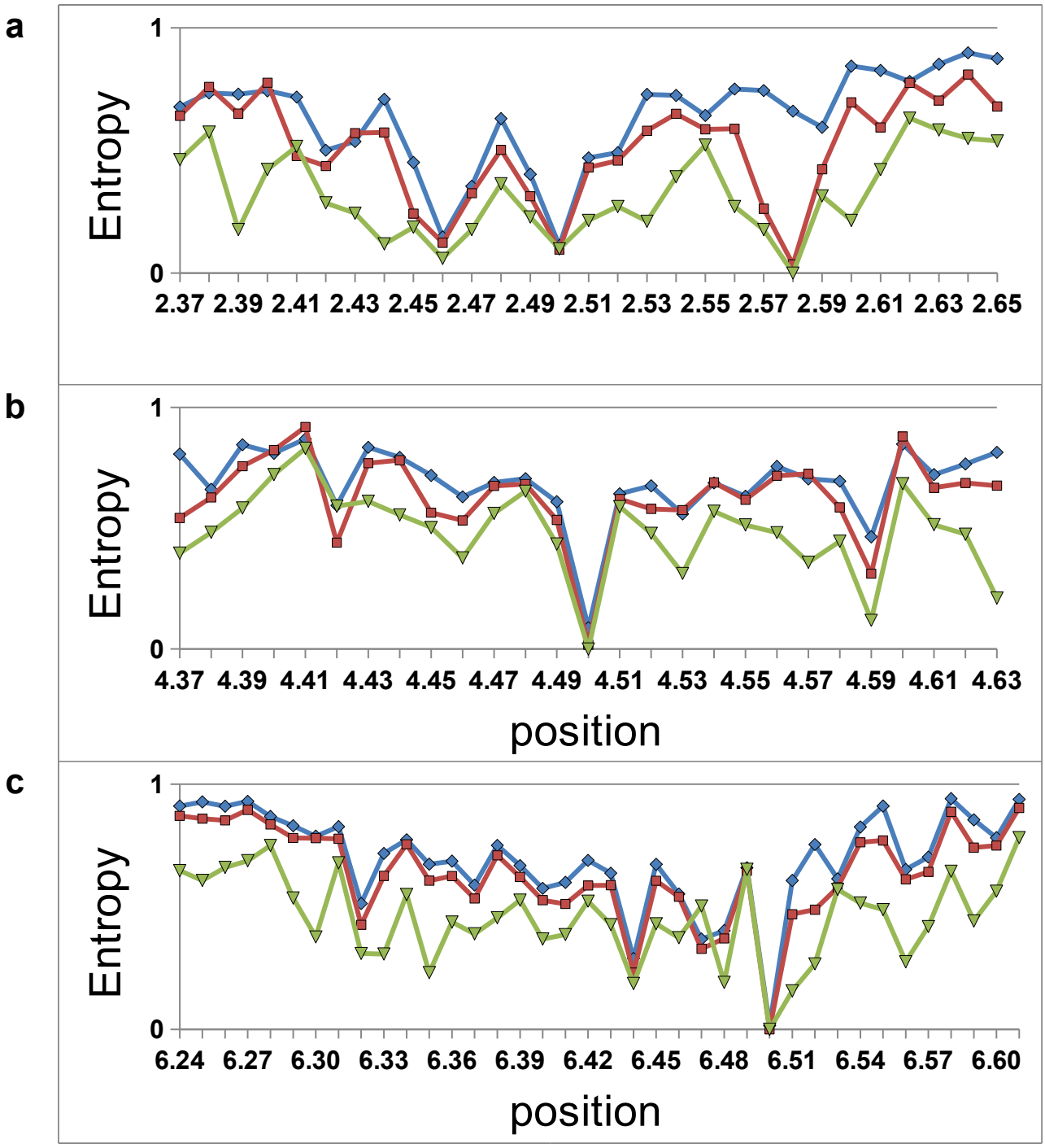

Figure 16 : Mesure d'entropies des HTM2, 4 et 6

Les 3 graphiques affichent les entropies pour l'HTM2 (a), 4 (b) et 6 (c). Les courbes bleue, rouge et verte représentent les entropies des RCPG de classe A humains, du groupe G1 et de la sous-famille CHEM classique, respectivement. Les symboles respectifs indiquent à quelle position de l'hélice la valeur d'entropie fait référence. Le graphique de l'HTM6 affiche environ 30% de positions en plus que celui de l'HTM2 et 4.

6.2.3 Comparaison des méthodes d'AMC

Les jeux de données sont testés avec les différentes méthodes d'AMC implémentées : OMES1, OMES2, MI, MIr, SCA, ELSC, MCBASC1 et MINT1 (**Tableau 3**). Dans un premier temps, nous avons cherché à comparer les valeurs de corrélation entre toutes les paires de positions (*i*,*j*) et les entropies des positions *i* et *j* (**Figure 17**). Sur un graphe à deux dimensions représentant les entropies de *i* et de *j*, nous visualisons la corrélation de la paire (*i*,*j*) par un code couleur. Les paires avec les valeurs de corrélation les plus élevées sont indiquées en bleu foncé et cyan pour les 25 et 275 premières valeurs. Les paires avec les valeurs de corrélation les plus faibles sont indiquées en rouge et en rose pour les 25 et 275 dernières valeurs (**Figure 17**). Pour qu'une méthode soit efficace, les positions les plus corrélées doivent avoir des entropies intermédiaires. En effet, si les positions les plus corrélées ont une entropie élevée (coin supérieur droit du graphe), cela signifie que ces positions sont très variables. Si elles ont une entropie faible (coin inférieur gauche), cela signifie qu'elles sont très conservées. De même, lorsqu'une position est très conservée, une méthode efficace doit indiquer un faible score de corrélation pour les paires de positions l'impliquant.

De manière générale, les valeurs d'entropie pour les RCPG de classe A humains couvrent l'ensemble du spectre possible (de 0 à 1), avec quelques lacunes entre les valeurs 0 et 0,3. Celles du groupe G1 affichent les mêmes lacunes mais les valeurs n'atteignent jamais le maximum 1. Pour la sous-famille CHEM classique, le spectre est plus restreint car les valeurs s'arrêtent à 0,8. De manière détaillée pour chaque méthode d'AMC, on peut observer que :

- la méthode OMES1 n'est pas robuste car les entropies des positions les plus corrélées varient sensiblement suivant le jeu de données. Plus le jeu de données est réduit, plus les positions corrélées ont des entropies élevées (déplacement vers le coin supérieur droit).

- la méthode OMES2 semble la plus robuste pour les trois ensembles de données parce que les entropies des positions les plus corrélées restent à peu près fixes. Les positions les plus corrélées affichent un niveau d'entropie intermédiaire, entre 0,3 et 0,8, alors que les positions les moins corrélées correspondent à des positions très conservées.

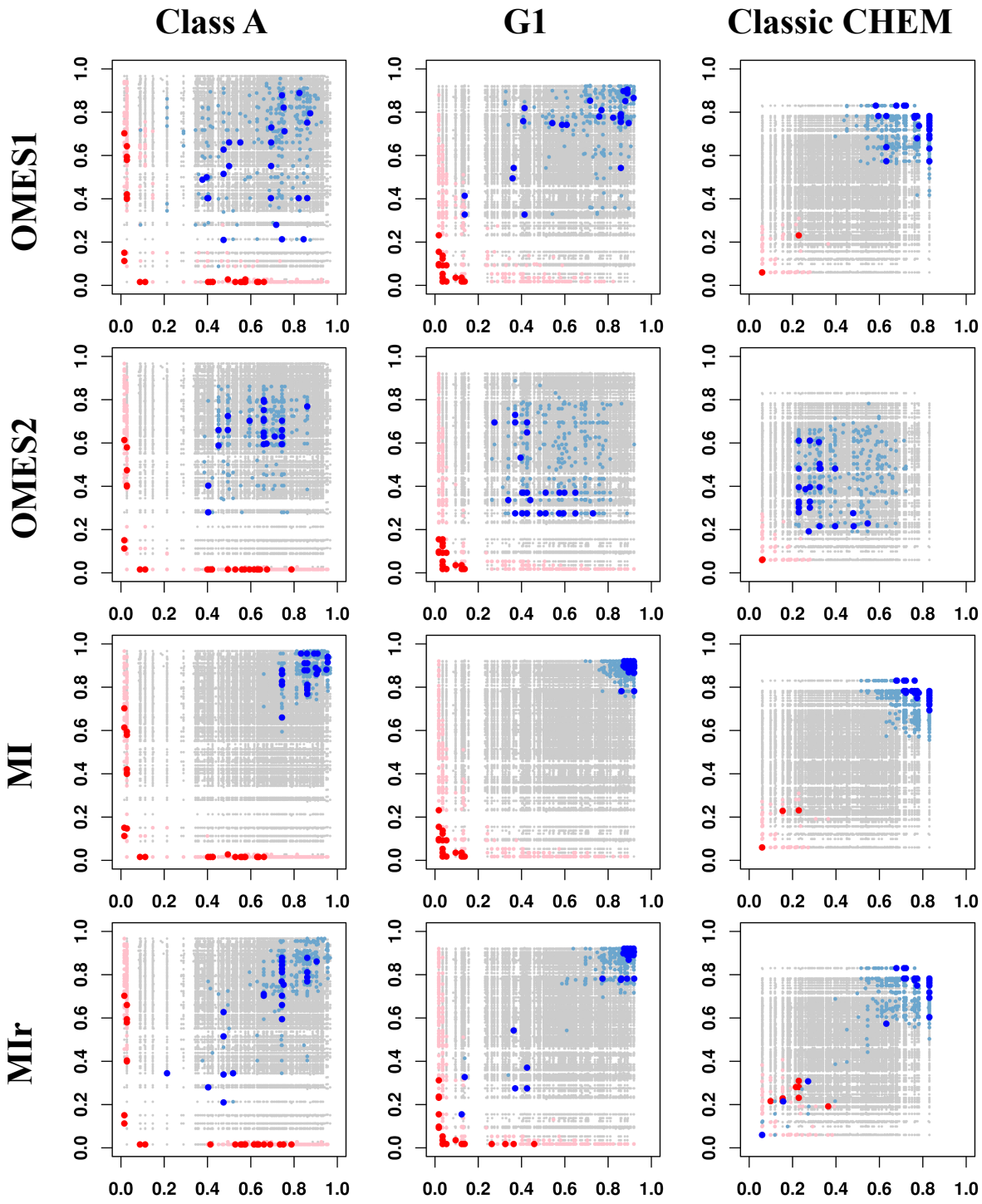

Figure 17 : Comparaison des méthodes d'AMC

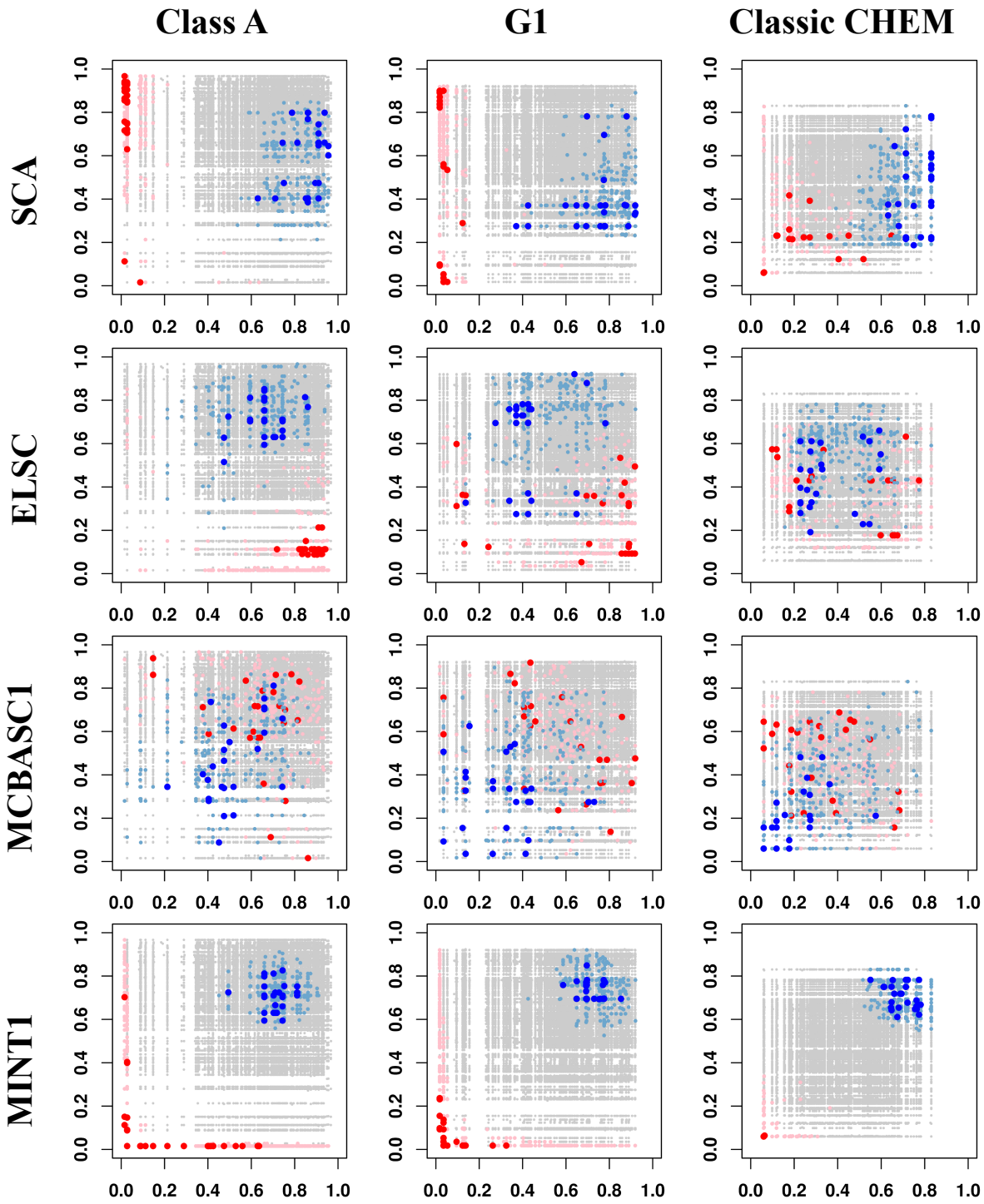

Figure 17 (Suite) : Comparaison des méthodes d'AMC

Les lignes correspondent aux méthodes d'AMC et les colonnes affichent les différents ensembles de données. Pour chaque graphique, les axes des abscisses et des ordonnées correspondent à l'entropie de la position *i* et *j*, respectivement. Les points rouges et roses correspondent aux 25 et 275 paires de positions (*i*,*j*) avec les valeurs de corrélation les plus faibles. Les points bleu foncé et cyan correspondent, à l'inverse, aux 25 et 275 paires avec les valeurs de corrélation les plus élevées.

- la méthode MI n'est pas efficace car les positions les plus corrélées affichent un niveau très élevé d'entropie. Les paires de positions les moins corrélées affichent un comportement similaire à celui de la méthode OMES1 et OMES2.

- la méthode MIr, dérivée de la méthode MI, améliore les résultats car les paires de positions les plus corrélées affichent des entropies plus faibles, en particulier pour le jeu le plus important. Néanmoins, la majeure partie des positions les plus corrélées ont encore une entropie très élevée. C'est le cas en particulier pour le jeu de plus petite taille.

- la méthode SCA n'est pas symétrique et cela s'observe directement sur le graphique. En effet, pour que des positions soient fortement corrélées, la position *i* doit afficher une entropie proche de 1.

- la méthode ELSC n'est pas symétrique et cela se ressent sur le graphique car pour l'ensemble des RCPG de classe A, les positions les moins corrélées impliquent forcement une entropie faible pour la position *j*, exactement l'inverse de la méthode SCA. Cette caractéristique semble s'estomper pour les ensembles avec moins de séquences.

- la méthode MCBASC1 ne semble pas attribuer de valeurs d'entropie spécifiques pour les positions les plus ou les moins corrélées car le calcul se base sur des matrices de substitution.

- la méthode MINT1 montre des résultats intéressants. Pour les CHEM classique, le profil est similaire à celui de la MI. Cependant, plus le nombre de séquences augmente, plus les positions fortement corrélées affichent des entropies intermédiaires.

6.2.4 Alignement multiple de positions corrélées

Dans une troisième étape, nous souhaitons observer la concordance entre les paires de positions corrélées et la composition en acides aminés respective. Pour cela, des alignements multiples de positions corrélées (AMPC) sont générés et correspondent à des AMS mais uniquement pour les paires de mutation corrélées.

L'objectif est de visualiser les paires d'acides aminés corrélés pour mettre en évidence les résidus qui coévoluent. Pour faciliter les interprétations, seules les deux paires d'acides aminés corrélés les plus fréquentes sont mises en valeur et les redondances de positions sont préservées. Les paires de première et deuxième occurrence sont colorées en turquoise et en jaune, respectivement. Les paires contenant un acide aminé impliqué dans l'une des deux paires les plus fréquentes sont également marquées par la couleur correspondante au niveau de cet acide aminé. Les paires d'acides aminés qui restent blanches n'impliquent aucun des acides aminés des deux paires les plus fréquentes. Une paire correspond à deux positions réellement corrélées quand il y a une proportion importante et à un peu près équivalente des paires turquoises et jaunes. Si une paire de positions corrélées est presque entièrement turquoise, cela veut dire que cette paire d'acides aminés est très conservée. Si une paire est presque entièrement blanche, cela veut dire que les acides aminés ne sont pas du tout conservés.

Également, pour faciliter la visualisation, les séquences sont classées en les regroupant par occurrences décroissantes des paires d'acides aminés. Lorsque des séquences affichent des acides aminés identiques pour la première paire de positions corrélées, ces séquences sont ordonnées suivant la deuxième paire et ainsi de suite. Si des séquences sont identiques sur toutes les paires corrélées, elles sont classées suivant leurs noms par ordre alphabétique. Les AMPC sont analysés (**Figure 18**) et nous montrons les quatre méthodes d'AMC (OMES2, MI, SCA et MCBASC1) les plus significatives. On peut observer que :

- les paires de positions les plus corrélées pour la méthode OMES2 présentent une alternance claire de turquoise et de jaune, en particulier pour les jeux G1 et CHEM classique.

- les paires de positions les plus corrélées pour la méthode MI montrent un degré de conservation faible, visualisé par une majorité de paires blanches. Aucune interprétation n'est possible.

- les paires de positions les plus corrélées pour la méthode SCA affichent une alternance de turquoise et jaune. Cependant, comme la méthode n'est pas symétrique, on peut voir que la position *i* est souvent moins conservée que la position *j*. L'interprétation s'en retrouve faussée. De plus, l'alternance est moins marquée que pour OMES2.

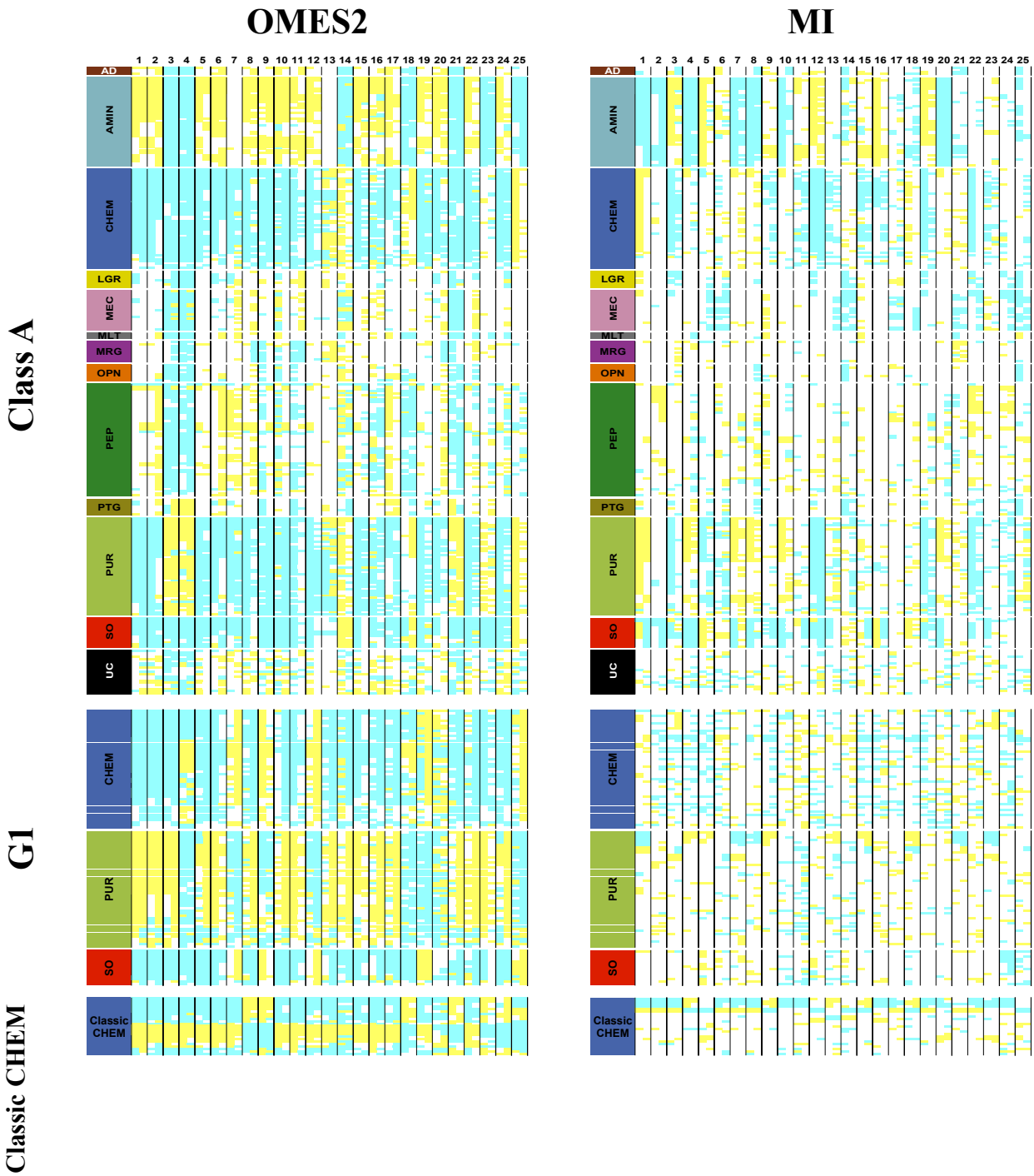

Figure 18 : Alignement multiple de positions corrélées

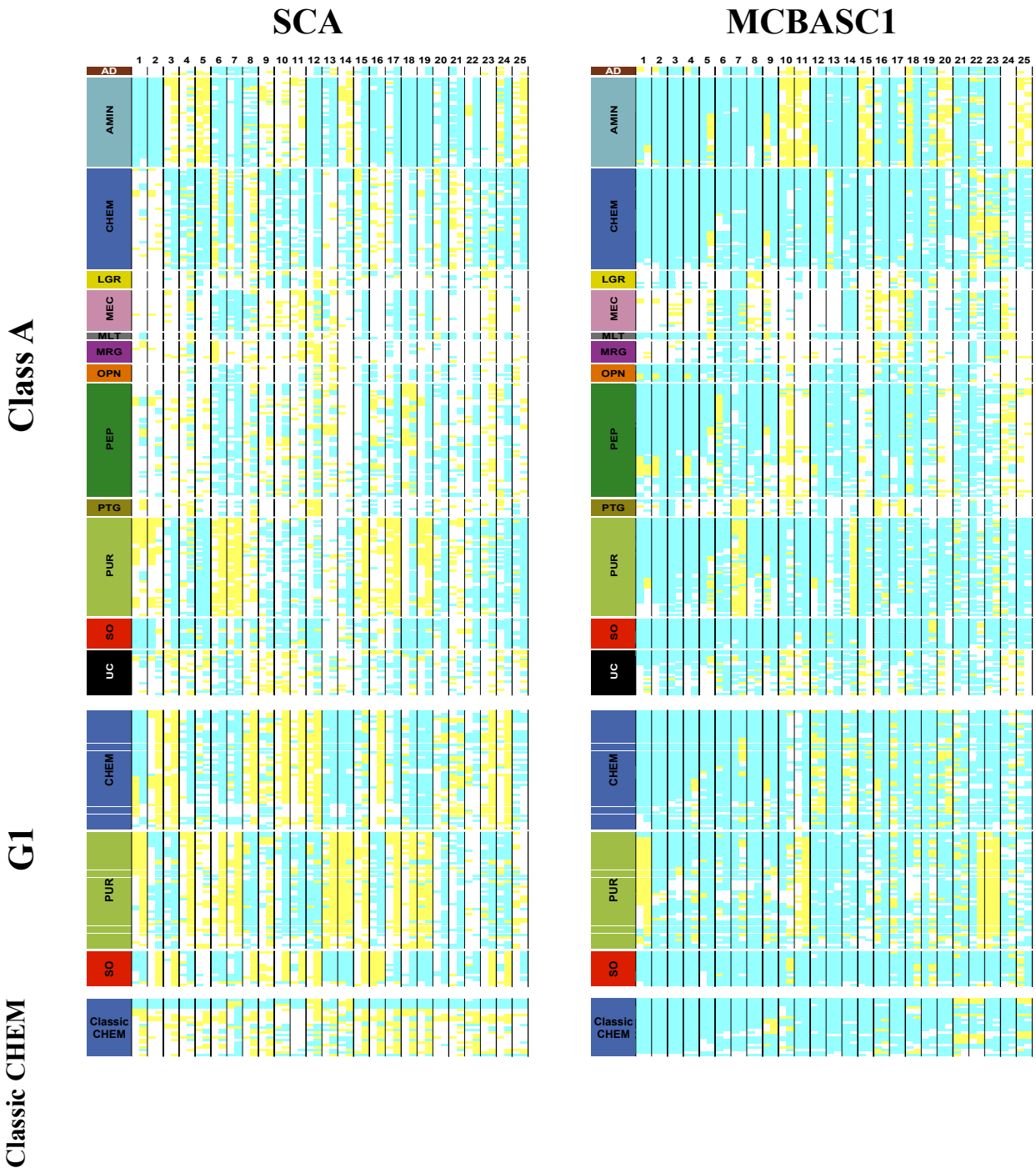

Figure 18 (Suite) : Alignement multiple de positions corrélées

Les lignes correspondent aux récepteurs des différentes sous-familles et les colonnes affichent les 25 paires de positions les plus corrélées. Les résidus en turquoise et en jaune correspondent aux paires de première et deuxième occurrence, respectivement.

- les paires de positions les plus corrélées pour la méthode MCBASC1 sont majoritairement en turquoise, indiquant un degré de conservation trop important.

6.2.5 Application d'OMES2 aux trois jeux de données

Les résultats les plus intéressants se retrouvent dans les AMPC de la méthode OMES2 car on observe que la plupart des paires de positions les plus corrélées sont liées à des paires d'acides aminés spécifiques, mettant en jeu de façon majoritaire les deux paires d'acides aminés les plus fréquentes. Les résultats se comportent différemment suivant l'ensemble testé :

- pour les RCPG de classe A, l'entropie moyenne est de 0,63 et les pourcentages moyens des deux premières paires d'acide aminés sont de 26% et 9%, respectivement. La position 2.58 est impliquée dans 48% des 25 paires de positions les plus corrélées. On peut observer que pour la première paire de positions (2.57,2.58), les sous-familles PUR, SO et CHEM affichent une paire d'acides aminés conservée (paire d'acides aminés LP) différentes des autres sous-familles (paire d'acides aminés VM pour la sous-famille AMIN, par exemple). Pour la deuxième paire de positions (2.58,2.59), on observe que la paire d'acides aminés PF est spécifique des sous-familles PUR, SO et CHEM. Pour l'ensemble des 25 premières paires de positions les plus corrélées, les sous-familles PUR, SO et CHEM affichent globalement un comportement semblable qui conforte le fait que ces trois sous-familles se retrouvent dans le même groupe G1 suite à l'analyse MDS et ont une origine commune [45]. On peut remarquer que la sous-famille AMIN affiche un comportement opposé par rapport aux sous-familles du groupe G1.

- pour le groupe G1, l'entropie moyenne est plus faible (0,44) et les pourcentages moyens des deux premières paires d'acide aminés sont plus élevés (34% et 18%, respectivement). Cette conservation est plus importante que précédemment et peut s'expliquer par le fait que les séquences du groupe G1 sont plus proches entre elles qu'avec les récepteurs des autres groupes. Les positions 7.49 et 6.48 sont impliquées dans 44% et 40% des 25 paires de positions les plus corrélées, respectivement. Cette fois-ci, ce sont les sous-familles SO et CHEM qui se comportent de manière similaire, contrairement à la sous-famille PUR qui

affiche toujours l'une des deux paires d'acides aminés les plus fréquentes opposée. En effet, pour les positions 7.49 et 6.48, la sous-famille PUR affiche un acide aspartique (D) au lieu d'une asparagine (N) et une phénylalanine (F) au lieu d'un tryptophane (W) par rapport aux sous-familles SO et CHEM, respectivement. Ces corrélations mettent clairement en évidence les résidus cruciaux pour la différentiation de la sous-famille PUR.

- pour les CHEM classique, l'entropie moyenne est encore plus faible (0,34) et les pourcentages moyens des deux premières paires d'acide aminés sont encore plus élevés (43% et 27%, respectivement). Cette entropie relativement basse peut s'expliquer par le fait que les séquences des CHEM classique sont très proches car elles constituent une catégorie de la sous-famille CHEM. La position 2.49 est impliquée dans 36% des 25 paires de positions les plus corrélées. Il est intéressant de noter que l'on peut apercevoir deux groupes au sein des CHEM classique de par la répartition des deux paires d'acides aminés les plus fréquents (turquoise et jaune) pour chaque paire de positions. Aucun motif de séquences, mis en évidence dans l'article (voir chapitre 6.1), ne permet de les distinguer car elles possèdent toutes des prolines en positions 2.58 et 5.50. L'AMPC sépare les récepteurs des chimiokines de la famille CC de ceux de la famille CXC. Cette observation souligne l'importance de la coévolution entre récepteurs et ligands.

6.2.6 Visualisation sur la structure de CXCR4

Visualiser les positions corrélées sur la structure des récepteurs peut permettre leur interprétation en termes structuraux et/ou fonctionnels. La figure 19 montre ainsi les positions les plus corrélées obtenues par la méthode OMES2 pour nos trois jeux de données, visualisées sur la structure de CXCR4 qui vient d'être résolue (**Figure 19**).

Dans un premier temps, nous reportons les 25 paires de positions les plus corrélées sur la structure pour les trois jeux de données (« All » sur la Figure 19). On peut observer que chaque niveau de classification des RCPG de classe A met en œuvre des positions corrélées différentes. Il est à remarquer que certaines positions y sont beaucoup plus impliquées : la position 2.58 pour les RCPG de classe A, les positions 6.48 et 7.49 pour le groupe G1 et la position 2.49 pour la sous-famille CHEM classique. Les réseaux de corrélation observés ne sont pas linéaires et concernent majoritairement des positions clés qui représentent de véritables nœuds de corrélation.

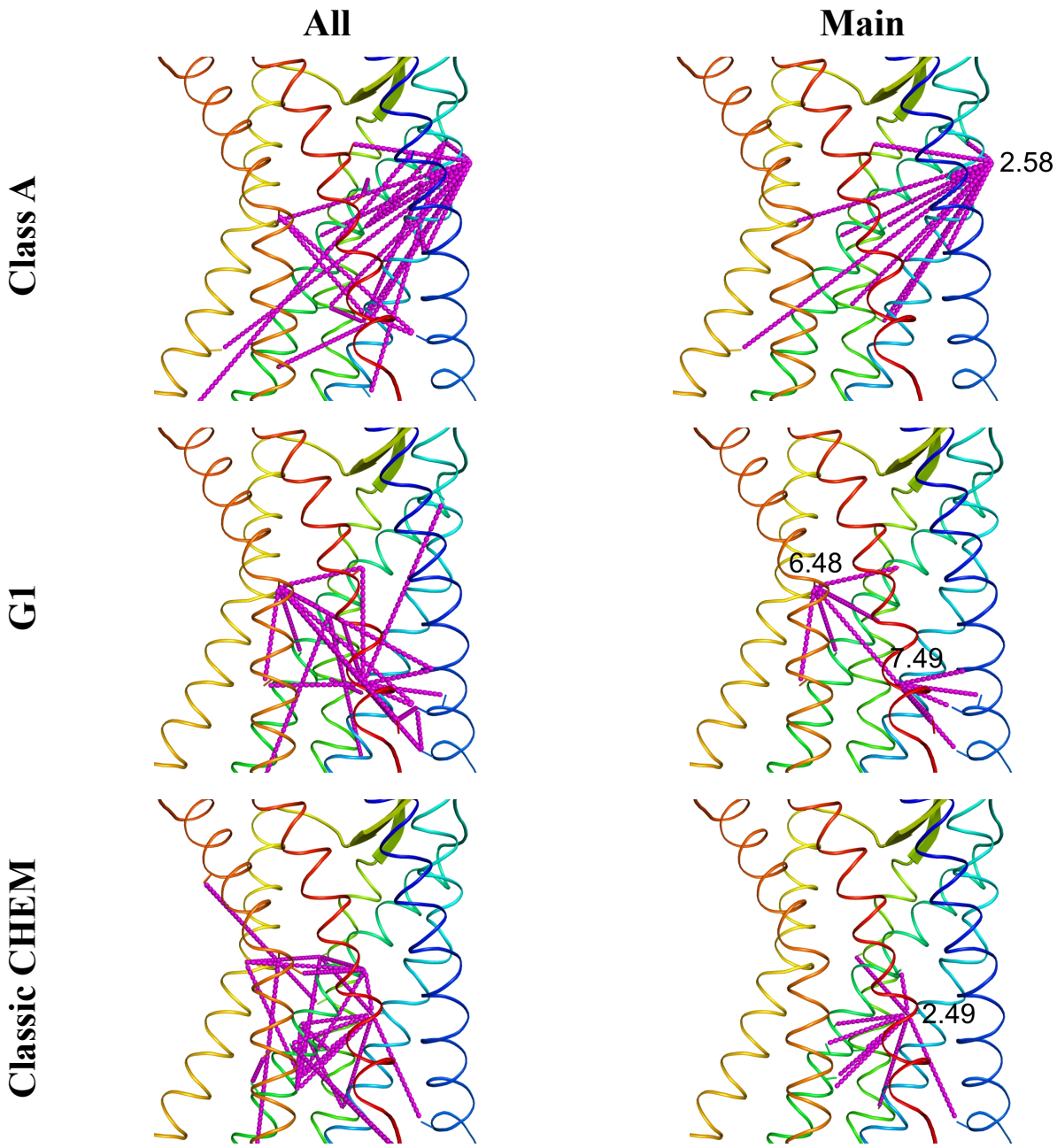

Figure 19 : Visualisation des réseaux de corrélation sur la structure cristalline de CXCR4

Les corrélations ont été calculées à partie de la méthode OMES2 sur les séquences des RCPG de classe A, du groupe G1 et de la sous-famille CHEM classique. Elles sont affichées sur la structure du récepteur CXCR4. Les HTM1, 2, 3, 4, 5, 6 et 7 sont colorées en bleu foncé, bleu clair, vert foncé, vert clair, jaune, orange et rouge, respectivement. Les barres en pointillés, en magenta, matérialisent les corrélations qui font participer les positions les plus fréquemment impliquées dans les 25 corrélations les plus élevées.

Dans un second temps, nous allons visualiser uniquement les corrélations qui mettent en jeu ces positions prépondérantes (« Main » sur la Figure 19). Nous écartons les corrélations qui n'impliquent pas ces positions. Après cette étape, on peut observer des corrélations triangulaires. En effet, lorsqu'une position A est corrélée à une position B qui est elle-même corrélée à une position C, la position A est corrélée à la position C. Les méthodes d'AMC ne font pas de distinction et reportent les corrélations brutes qui peuvent parfois être redondantes. Ainsi, pour chaque corrélation triangulaire, nous éliminons la corrélation qui est observée pour la distance maximale au niveau de la structure. Concernant les résultats, les corrélations qui impliquent les positions les plus redondantes ne sont pas totalement radiales car elles possèdent des directions bien précises, généralement dirigées vers l'espace intracellulaire. Cela est très remarquable pour le groupe G1 qui fait intervenir deux positions prépondérantes qui semblent jouer un rôle de pivot et qui pourraient être impliquées dans la transduction du signal vers l'espace intracellulaire.

6.2.7 Conclusion

Cette étude comparative a permis de mettre en évidence que certaines méthodes d'AMC ne sont pas robustes face à des profils d'entropie différents et que certaines ne sont pas adaptées à nos jeux de données. La méthode OMES2 donne les meilleurs résultats car les positions les plus corrélées affichent un degré de conservation intermédiaire, que l'on peut facilement interpréter grâce aux AMPC. Dans la plupart des cas, les deux paires d'acides aminés les plus fréquentes sont impliquées en montrant de réelles dualités : entre les sous-familles PUR, SO et CHEM et les autres sous-familles pour les RCPG de classe A humain (AMIN, notamment), entre les deux sous-familles SO-CHEM et PUR pour le groupe G1 et entre les familles CCR et CXCR de la sous-famille CHEM classique.

Pour le premier cas, il faut mettre en relation ces observations avec l'hypothèse évolutive de l'article (voir chapitre 6.1). Au cours de l'évolution, la sous-famille SO donne naissance aux sous-familles CHEM et PUR. Ces sous-familles restent tout de même proches car l'analyse MDS les réunit sous le même groupe. La sous-famille AMIN affiche un comportement opposé par rapport aux sous-familles du groupe G1 et cela s'explique par le fait que ces sous-familles se placent aux extrêmes de la première composante de l'analyse MDS pour le génome humain. Cette différenciation importante impliquerait des changements de séquences importants que l'on retrouve dans nos AMPC.

Pour le second cas, la sous-famille SO est proche de la sous-famille CHEM. L'émergence de la sous-famille PUR est caractérisée par des changements marqués au niveau de positions très corrélées et impliquant des résidus très conservés. Par exemple, la corrélation la plus élevée pour ce jeu de données concerne la paire de positions 6.48-7.49 et la sous-famille PUR affiche un acide aspartique (D) au lieu d'une asparagine (N) et une phénylalanine (F) au lieu d'un tryptophane (W) par rapport aux sous-familles SO et CHEM, respectivement.

Pour le troisième cas, la sous-famille CHEM se développe à partir des vertébrés et les AMPC montrent que les CHEM classiques se séparent en deux groupes bien distincts. Il serait intéressant d'étudier l'évolution de cette sous-famille pour expliquer la différenciation entre les récepteurs CCR et CXCR de la sous-famille CHEM. Cette analyse est envisagée prochainement par l'équipe de bioinformatique.

Cette étude méthodologique a permis de sélectionner la méthode qui convient le mieux à nos jeux de données. L'application de la méthode OMES2 permet d'expliquer les changements de séquences qui s'effectuent avec les différents niveaux de classification des RCPG de classe A. La visualisation des réseaux de corrélation sur la structure cristalline de CXCR4 permet de mettre en évidence la prépondérance de certaines positions et leur implication possible, de part leurs positions spatiales, dans le mécanisme d'activation des RCPG.

7. CONCLUSIONS ET PERSPECTIVES

Les travaux effectués au cours de cette thèse de doctorat concernent l'analyse de l'évolution des RCPG, en lien avec nos récepteurs d'intérêt. Mon travail avait comme objectif de déterminer le schéma évolutif des RCPG de classe A en mettant en évidence les relations qui existent entre les différentes sous-familles et en déterminant les résidus importants impliqués dans l'évolution et la diversification de cette famille de récepteurs. Apporter de nouvelles connaissances dans ce domaine aidera grandement à mieux appréhender les relations séquence-structure-fonction des RCPG et éventuellement à mieux comprendre les mécanismes d'activation spécifique de chaque sous-famille. Les liens entre ces travaux et les travaux précédents de l'équipe de bioinformatique, qui concernent les motifs structuraux, vont permettre de mieux comprendre les aspects séquentiels, structuraux et fonctionnels en relation avec l'évolution des RCPG de classe A.

La première étape de mon travail a été d'étudier les différentes sous-familles des RCPG de classe A et de les caractériser par des spécificités séquentielles. Ce travail a conduit à l'élaboration d'une publication. L'analyse des RCPG de classe A humains par MDS a révélé que les sous-familles se réunissent sous forme de quatre groupes bien distincts. L'application inédite de la technique de projection d'éléments supplémentaires dans le cadre de la MDS a permis de projeter des espèces plus ou moins éloignées par rapport à l'homme. Cette approche a montré que les sous-familles évoluent de manière asynchrone et que leurs apparitions sont directement liées aux étapes majeures de l'évolution. L'évolution des sous-familles implique très probablement des modifications au niveau des résidus clés qui jouent un rôle important pour le maintien structural et fonctionnel des récepteurs. Cela est confirmé par l'analyse des AMS de chacun des quatre groupes et met en évidence un petit nombre de résidus spécifiques. Les résidus prolines des HTM2 et 5 y sont grandement impliqués, en plus d'être des résidus critiques pour la conformation des hélices des RCPG de classe A. Une hypothèse de schéma évolutif de type radial des RCPG de classe A a été établie avec la sous-famille PEP comme sous-famille centrale. Des mécanismes de mutagenèse, de délétion dans l'HTM2 et de mutations dans l'HTM2 et/ou l'HTM5 et du motif W/FxxG ont permis de donner naissance à la grande diversité des RCPG de classe A. La technique MDS permet d'analyser de grandes familles protéiques et de résoudre les relations entre sous-familles lorsque la technologie des arbres phylogénétiques ne le permet pas en totalité. L'application de la technique de projection d'éléments supplémentaires donne la possibilité de suivre l'évolution progressive des sous-familles. De plus, la technique MDS permet de conjuguer les relations évolutives entre les sous-familles avec les structures cristallines connues des RCPG de classe A (**Figure 20**) et de mettre en évidence la structure cristalline la plus adaptée pour la modélisation moléculaire par homologie

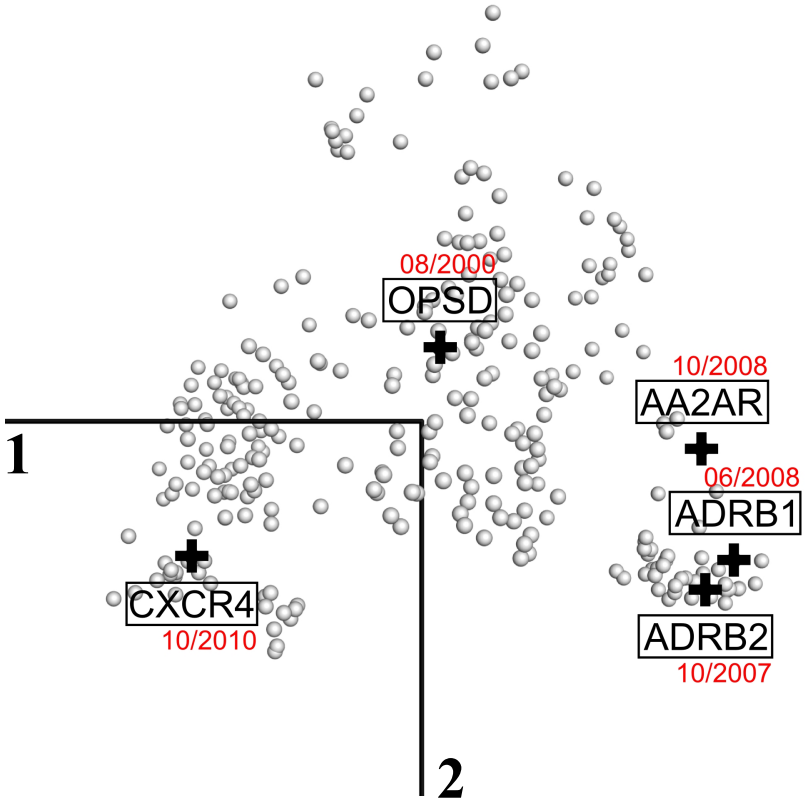

Figure 20 : Localisation des structures connues de RCPG sur l'analyse MDS

Les sphères blanches correspondent aux RCPG de classe A humains. Les croix noires mettent en évidence la localisation des récepteurs dont la structure cristalline est connues : OPSD pour la rhodopsine, ADRB2 pour le récepteur β2 adrénergique, ADRB1 pour le récepteur β1 adrénergique, AA2AR pour le récepteur de l'adénosine A2A et CXCR4 pour le récepteur des chimiokines CXCR4. Les inscriptions en rouge indiquent le mois et l'année de sortie de la structure.

d'un récepteur donné. La première structure qui a été découverte est la rhodopsine et elle appartient à la sous-famille OPN du groupe G0. Puis, les structures des récepteurs β1 et β2 adrénergiques concernant la sous-famille AMIN et celle du récepteur de l'adénosine A2A concernant la famille AD ont été résolues. Elles font partie du groupe G2. Ces quatre structures présentent toutes une conformation de l'HTM2 en renflement π, soit par la présence des résidus GG aux positions 2.56-2.57 pour la rhodopsine, soit par la présence d'une proline en position 2.59 pour les trois autres récepteurs. Récemment, c'est la structure d'un récepteur de la sous-famille CHEM qui a été résolue : le récepteur aux chimiokines CXCR4, qui fait partie du groupe G1. Cette structure confirme notre hypothèse de délétion d'un résidu dans l'HTM2 des récepteurs du groupe G1 [45,108].

Suite à notre étude par MDS, nous avons souhaité étudier davantage le groupe G1, qui contient nos récepteurs d'intérêt. L'étude des mutations corrélées doit permettre d'en étudier les caractéristiques et de mettre en relation l'évolution des séquences de ces récepteurs avec les contraintes structurales et fonctionnelles. Ainsi, la seconde étape de mon travail a été de déterminer les positions impliquées dans la covariation des résidus. Il a fallu mener une étude de méthodologie pour sélectionner la méthode d'AMC qui correspond le mieux aux caractéristiques de nos jeux de données. Nous nous sommes basés sur la notion d'entropie pour pouvoir comparer les méthodes entre elles car elle permet de faciliter la visualisation des paires de résidus corrélés. L'étude a montré que les méthodes d'AMC réagissent différemment et possèdent des comportements bien spécifiques. La méthode OMES2 semble s'accorder le mieux avec les différents niveaux de classification des RCPG de classe A. Dans la plupart des cas, l'analyse des résidus corrélées montre majoritairement deux paires d'acides aminés totalement différentes et ainsi met en évidence les dualités qui peuvent exister au niveau du schéma de mutations. Cette étude de méthodologie fera l'objet d'une publication, qui est en cours de rédaction.

Concernant le package R *bio2mds*, il pourra être amélioré, complété et être publié dans une revue scientifique par l'intermédiaire d'une note d'application. On pourra proposer une interface graphique (GUI) via le package R commander [109] pour les utilisateurs qui sont peu familiers avec l'interface en ligne de commande (CLI). Les méthodes d'AMC qui ont été utilisées lors de mon projet sont généralement fournies par les auteurs mais elles sont implémentées en différents langages et demandent des paramètres différents. Nous pourrons les implémenter sous la forme d'un package R, différent de *bio2mds*, pour les regrouper et offrir une interface utilisateur homogène. Le temps de calcul des mutations corrélées peut parfois être long pour des jeux de données volumineux. Ainsi, les fonctions pourront être codées en C++ et appelées par le package R.

En termes de perspectives, on pourra envisager d'approfondir l'étude des mutations corrélées en lien avec l'évolution des sous-familles du groupe G1. La méthode d'AMC OMES2 pourrait être appliquée à un jeu de données contenant le groupe G1 et la sous-famille PEP pour mettre en évidence les positions corrélées les plus importantes pour la différenciation des récepteurs du groupe G1. On peut également tenter de visualiser les paires de mutations corrélées en fonction de l'évolution des sous-familles de RCPG de classe A en appliquant la méthode d'AMC OMES2 aux récepteurs d'une sous-famille pour l'ensemble des espèces testées dans l'article scientifique (voir chapitre 6.1). Par exemple, l'application à la sous-famille SO, qui affiche la dérive évolutive la plus prononcée, pourrait permettre de relier l'évolution de ces récepteurs à des caractéristiques de séquences particulières. Il serait également intéressant d'appliquer la méthode MDS à une matrice de distances, basées sur les corrélations obtenues après application de la méthode OMES2. Cela permettrait d'analyser l'ensemble des corrélations et pourrait faciliter l'interprétation des réseaux de corrélation.

Les réseaux de corrélation devront être analysés et synthétisés par des méthodes appropriées. Également, les réseaux devront être mis en relation avec les interactions protéine-protéine dans le cadre d'homo-oligomères et les connaissances actuelles sur les mécanismes d'activation des RCPG de classe A. On peut observer que chaque niveau de classification des RCPG de classe A met en œuvre des positions corrélées différentes. Il serait intéressant de visualiser les réseaux de corrélation spécifiques des sous-familles avec les structures cristallines correspondantes actuellement disponibles. Pour les structures manquantes, on pourrait les modéliser avec le programme de modélisation MODELLER [110] et mettre en lien les réseaux de corrélation avec les contraintes structurales propres à chacun des récepteurs. La théorie des graphes devra être appliquée pour faciliter les interprétations [54]. Toutes ces données permettront d'affiner nos modèles structuraux en lien avec nos récepteurs d'intérêt. Les hypothèses pourront être confirmées expérimentalement et contribuer à l'expansion des connaissances sur les RCPG de classe A. Dans une perspective un peu plus lointaine, ces informations qui contribuent à la compréhension des relations séquence-structure-fonction des RCPG pourraient contribuer à la conception des médicaments (*i.e.*, drug design) ciblant ces récepteurs.

Liste des abréviations

ACNUC : ACides NUCléiques
AD : récepteur d'adénosines
ADORA : récepteur de fixation à l'adénosine
ADR : récepteur adrénergique
AGTR : récepteur de l'angiotensine
AGTR : récepteur lié à l'angiotensine
AJAX : Asynchronous Javascript And XML
AMC : Analyse des mutations corrélées
AMIN : récepteur des amines
AMPc : Adénosine MonoPhosphate cyclique
AMPC : Alignement Multiple des Positions Corrélées
AMS : Alignement Multiple de Séquences
APC : Average Product Correction
ASC : Average Sum Correction
AVPR : récepteur de la vasopréssine
BATMAS : BActerial Transmembrane MAtrix of Substitutions
BDKR : récepteur à la bradykinine
BLOSUM : BLOcks of amino acid SUbstitution Matrix
BRS3 : récepteur de l'uterinbombésine
CALCR : récepteur de la calcitonine
CASR : récepteur du calcium
CCK : récepteur de la cholécystokinine
CCR et CXCR : récepteur des chimiokines classiques
CEO : CombinatorialEntropy Optimization
CHEM : récepteur des chimiokines
CHRM : récepteur muscarinique
CLI : Command-line interface
CMAV : Correlated Mutation Analysis of Vicatos
CNR : récepteur de l'anandamide
CNR : récepteur des cannabinoïdes
CRAN : Comprehensive R Archive Network
CSV : Comma-Separated Values
DOM : Document Object Model
DRD : récepteur de la dopamine
EC : boucle ExtraCellulaire
EDGR : récepteur de l'acide lysophosphatidique
EDGR : récepteur de la différenciation endothéliale
EDNR et ETBRLP1/2 : récepteur lié aux endothélines
ELSC : Explicit Likelihood of Subset Covariation
FASTA : FAST-All
FC : Frequency Correlation
FPR : récepteur de peptides formylés
FSHR, TSHR et LHCGR : récepteurs des hormones glycoprotéiques classiques
GABA : acide γ-aminobutyrique
GALR : récepteur de la galanine

GDP : Guanosine DisPhosphate
GHSR : récepteur de l'hormone de croissance sécrétagogue
GHSR : récepteur de la ghréline
GLPR : récepteur du glucagon
GNRHR : récepteur de l'hormone de libération de la gonadotropine
GNU : GNU's Not UNIX
GPR54 : récepteur du RF-amide
GRAFS : Glutamate, Rhodopsin, Adhesion, Frizzled/taste2 et Secretin
GRM : récepteur métabotrope du glutamate
GRPR : récepteur du peptide de libération de la gastrine
GTP : Guanosine TriPhosphate
GUI : Graphical user interface
HCRTR : récepteur de l'hypocrétine
HRH : récepteur des histamines
HTM : Hélice TransMembranaire
HTR : récepteur de la sérotonine
IC : boucle IntraCellulaire
JE : Joint Entropy
JS : JavaScript
LCD : langage de contrôle de données
LDD : langage de définition de données
LGR : récepteur aux glycoprotéines
LGR : récepteur des répétitions riches en leucine
LMD : langage de manipulation de données
MAS : récepteur à l'oncogène MAS1
MCBASC : McLachlan BAsed Substitution Correlation
MCD : modèle conceptuel des données
MCHR : récepteur de l'hormone de concentration de mélanine
MCR : récepteur de l'hormone de stimulation des mélanocytes
MCR : récepteur des mélanocortines
MDS : MultiDimensional Scaling
MECA : récepteur des mélanocortines-endoglines-cannabinoïdes-adénosines
MI : Mutual Information
MIa : Mutual Information additive
MImax : MI maximal
MINT : Mutual INTerdependency
MIp : Mutual Information product
MIr : Mutual Information removed
MRG : récepteur lié à MAS
MSF : Multiple Sequence Format
MTN : récepteur des mélatonines
NJ : Neighbour Joining
NMBR : récepteur de la neuromédine B
NMUR : récepteur de la neuromédine
NPFF : récepteur du neuropeptide FF
NPYR : récepteur du neuropeptide Y
NTSR : récepteur de la neurotensine
OLF : récepteur olfactif

OMES : Observed Minus Expected Squared
OPN : récepteur des opsines
OPN1SW, OPN1LW, OPN1MW : récepteurs dans les trois types de cônes
OPN3 : l'encéphalopsine
OPN4 : la mélanopsine
OPR : récepteur des opioïdes
OXTR : récepteur de l'
P2Y : récepteur de nucléotides
PAM : Point Accepted Mutation
PAM250TM : TransMembrane PAM250
PCA : Principal Component Analysis
PDB : Protein Data Bank
PEP : récepteur des peptides
PHAT : Predicted Hydrophobic And Transmembrane
PHP : Hypertext Prepro
PML : PyMoL
PTGR : récepteur des prostaglandines
PTH : ParaTHormone
PUR : récepteur purinergique
Rd : R documentation
rda : R data
RCPG : Récepteurs Couplés aux Protéines G
RGR : récepteur lié au rétinal
RGS : régulateurs de protéines G
RHO : récepteur visuel dans les bâtonnets
RL3R : récepteur de la relaxine
RMN : Résonance Magnétique Nucléaire
RRH : la péropsine
SCA : Statistical Coupling Analysis
SGBD : système de gestion de bases de données
SLIM : Scorematrix Leading to IntraMembrane domains
SOG : récepteur des somatostatine/opioïde/galanine
SOM : Self Organising Map
SQL : Structure Query Language
SSTR : récepteur de la somatostatine
STATIS : Structuration des Tableaux A Trois Indices de la Statistique
TACR : récepteur de la tachykinine
TAR : récepteur des traces d'amines
TAS : récepteur du goût
TM : TransMembranaire
TRHR : récepteur de l'hormone de libération de la thyrotropine
UC : non-classifié
VIP : Peptide Intestinal Vasoactif

Bibliographie

1. Rokas, A. & Carroll, S. B. (2006), 'Bushes in the tree of life.', PLoS Biol 4(11), e352.

2. Rokas, A.; Krüger, D. & Carroll, S. B. (2005), 'Animal evolution and the molecular signature of radiations compressed in time.', *Science* **310**(5756), 1933--1938.

3. Nei, M., & Kumar S. (2000), 'Molecular Evolution and Phylogenetics.', *Oxford University Press*, Oxford.

4. Göbel, U.; Sander, C.; Schneider, R. & Valencia, A. (1994), 'Correlated mutations and residue contacts in proteins.', *Proteins* **18**(4), 309--317.

5. Fodor, A. A. & Aldrich, R. W. (2004), 'Influence of conservation on calculations of amino acid covariance in multiple sequence alignments.', *Proteins* **56**(2), 211—221.

6. Codoñer, F. M. & Fares, M. A. (2008), 'Why should we care about molecular coevolution?', *Evol Bioinform Online* **4**, 29--38.

7. Atchley, W. R.; Wollenberg, K. R.; Fitch, W. M.; Terhalle, W. & Dress, A. W. (2000), 'Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.', *Mol Biol Evol* **17**(1), 164--178.

8. Horner, D. S.; Pirovano, W. & Pesole, G. (2008), 'Correlated substitution analysis and the prediction of amino acid structural contacts.', *Brief Bioinform* **9**(1), 46--56.

9. Buck, M. J. & Atchley, W. R. (2005), 'Networks of coevolving sites in structural and functional domains of serpin proteins.', *Mol Biol Evol* **22**(7), 1627--1634.

10. Gether, U. (2000), 'Uncovering molecular mechanisms involved in activation of G protein coupled receptors.', *Endocr Rev* 21(1), 90--113.

11. Kolakowski, L. F. (1994), 'GCRDb: a G-protein-coupled receptor database.', *Receptors Channels* **2**(1), 1--7.

12. Davies, M. N.; Gloriam, D. E.; Secker, A.; Freitas, A. A.; Mendao, M.; Timmis, J. & Flower, D. R. (2007), 'Proteomic applications of automated GPCR classification.', *Proteomics* **7**(16), 2800--2814.

13. Bockaert, J. & Pin, J. P. (1999), 'Molecular tinkering of G protein-coupled receptors: an evolutionary success.', EMBO J 18(7), 1723--1729.

14. Flower, D. R. (1999), 'Modelling G-protein-coupled receptors for drug design.', *Biochim Biophys Acta* **1422**(3), 207--234.

15. Drews, J. (2000), 'Drug discovery: a historical perspective.', *Science* **287**(5460), 1960--1964.

16. Gether, U.; Ballesteros, J. A.; Seifert, R.; Sanders-Bush, E.; Weinstein, H. & Kobilka, B. K. (1997), 'Structural instability of a constitutively active G protein-coupled receptor. Agonist-independent activation due to conformational flexibility.', J Biol Chem 272(5), 2587--2590.

17. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Trong, I. L.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M. & Miyano, M. (2000), 'Crystal structure of rhodopsin: A G protein-coupled receptor.', *Science* **289**(5480), 739--745.

18. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K. & Stevens, R. C. (2007), 'High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor.', *Science* **318**(5854), 1258--1265.

19. Topiol, S. & Sabio, M. (2009), 'X-ray structure breakthroughs in the GPCR transmembrane region.', Biochem Pharmacol 78(1), 11--20.

20. Teller, D. C.; Okada, T.; Behnke, C. A.; Palczewski, K. & Stenkamp, R. E. (2001), 'Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs).', *Biochemistry* **40**(26), 7761--7772.

21. Park, J. H.; Scheerer, P.; Hofmann, K. P.; Choe, H.-W. & Ernst, O. P. (2008), 'Crystal structure of the ligand-free G-protein-coupled receptor opsin.', *Nature* **454**(7201), 183--187.

22. Shimamura, T.; Hiraki, K.; Takahashi, N.; Hori, T.; Ago, H.; Masuda, K.; Takio, K.; Ishiguro, M. & Miyano, M. (2008), 'Crystal structure of squid rhodopsin with intracellularly extended cytoplasmic region.', *J Biol Chem* **283**(26), 17753--17756.

23. Murakami, M. & Kouyama, T. (2008), 'Crystal structure of squid rhodopsin.', *Nature* **453**(7193), 363--367.

24. Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G. W.; Tate, C. G. & Schertler, G. F. X. (2008), 'Structure of a beta1-adrenergic G-protein-coupled receptor.', *Nature* **454**(7203), 486--491.

25. Jaakola, V.-P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; Ijzerman, A. P. & Stevens, R. C. (2008), 'The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist.', *Science* **322**(5905), 1211--1217.

26. Wu, B.; Chien, E. Y. T.; Mol, C. D.; Fenalti, G.; Liu, W.; Katritch, V.; Abagyan, R.; Brooun, A.; Wells, P.; Bi, F. C.; Hamel, D. J.; Kuhn, P.; Handel, T. M.; Cherezov, V. & Stevens, R. C. (2010), 'Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists.', *Science* **330**(6007), 1066--1071.

27. Mobarec, J. C. & Filizola, M. (2008), 'Advances in the Development and Application of Computational Methodologies for Structural Modeling of G-Protein Coupled Receptors.', Expert Opin Drug Discov 3(3), 343--355.

28. Baldwin, J. M. (1993), 'The probable arrangement of the helices in G protein-coupled

receptors.', EMBO J 12(4), 1693--1703.

29. Morris, M. B.; Dastmalchi, S. & Church, W. B. (2009), 'Rhodopsin: structure, signal transduction and oligomerisation.', *Int J Biochem Cell Biol* **41**(4), 721--724.

30. Ji, T. H.; Grossmann, M. & Ji, I. (1998), 'G protein-coupled receptors. I. Diversity of receptor-ligand interactions.', J Biol Chem 273(28), 17299--17302.

31. Sealfon, S. C.; Chi, L.; Ebersole, B. J.; Rodic, V.; Zhang, D.; Ballesteros, J. A. & Weinstein, H. (1995), 'Related contribution of specific helix 2 and 7 residues to conformational activation of the serotonin 5-HT2A receptor.', J Biol Chem 270(28), 16683--16688.

32. Ballesteros, J. A. & Weinstein, H. (1995), 'Integrated methods for the construction of three dimensional models and computational probing of structure function relations in G protein-coupled receptors.', Methods in Neurosciences 25, 366--428.

33. Tuteja, N. (2009), 'Signaling through G protein coupled receptors.', *Plant Signal Behav* **4**(10), 942—947.

34. Li, J.; Edwards, P. C.; Burghammer, M.; Villa, C. & Schertler, G. F. X. (2004), 'Structure of bovine rhodopsin in a trigonal crystal form.', J Mol Biol 343(5), 1409--1438.

35. Sansuk, K.; Deupi, X.; Torrecillas, I.; Jongejan, A.; Nijmeijer, S.; Bakker, R.; Pardo, L. & Leurs, R. (2010), 'A structural insight into the reorientation of transmembrane domains 3 and 5 during family A GPCR activation.', *Mol Pharmacol*.

36. Prinster, S. C.; Hague, C. & Hall, R. A. (2005), 'Heterodimerization of g protein-coupled receptors: specificity and functional significance.', *Pharmacol Rev* **57**(3), 289--298.

37. Milligan, G. (2009), 'G protein-coupled receptor hetero-dimerization: contribution to pharmacology and function.', *Br J Pharmacol* **158**(1), 5--14.

38. Milligan, G. (2007), 'G protein-coupled receptor dimerisation: molecular basis and relevance to function.', *Biochim Biophys Acta* **1768**(4), 825--835.

39. Filizola, M. & Weinstein, H. (2005), 'The study of G-protein coupled receptor oligomerization with computational modeling and bioinformatics.', *FEBS J* **272**(12), 2926--2938.

40. Guo, W.; Shi, L.; Filizola, M.; Weinstein, H. & Javitch, J. A. (2005), 'Crosstalk in G protein-coupled receptors: changes at the transmembrane homodimer interface determine activation.', Proc Natl Acad Sci U S A 102(48), 17495--17500.

41. DeWire, S. M.; Ahn, S.; Lefkowitz, R. J. & Shenoy, S. K. (2007), 'Beta-arrestins and cell signaling.', *Annu Rev Physiol* **69**, 483—510.

42. Bhattacharya, S.; Hall, S. E. & Vaidehi, N. (2008), 'Agonist-induced conformational changes in bovine rhodopsin: insight into activation of G-protein-coupled receptors.', *J Mol Biol*

382(2), 539—555.

43. Fredriksson, R.; Lagerström, M. C.; Lundin, L.-G. & Schiöth, H. B. (2003), 'The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints.', Mol Pharmacol 63(6), 1256--1272.

44. McKnight, A. J. & Gordon, S. (1998), 'The EGF-TM7 family: unusual structures at the leukocyte surface.', J Leukoc Biol 63(3), 271--280.

45. Devillé, J.; Rey, J. & Chabbert, M. (2009), 'An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors.', *J Mol Evol* **68**(5), 475--489.

46. Hill, C. A.; Fox, A. N.; Pitts, R. J.; Kent, L. B.; Tan, P. L.; Chrystal, M. A.; Cravchik, A.; Collins, F. H.; Robertson, H. M. & Zwiebel, L. J. (2002), 'G protein-coupled receptors in Anopheles gambiae.', *Science* **298**(5591), 176--178.

47. Josefsson, L. G. & Rask, L. (1997), 'Cloning of a putative G-protein-coupled receptor from Arabidopsis thaliana.', *Eur J Biochem* **249**(2), 415--420.

48. Okada, T. & Palczewski, K. (2001), 'Crystal structure of rhodopsin: implications for vision and beyond.', *Curr Opin Struct Biol* **11**(4), 420--426.

49. Fredriksson, R. & Schiöth, H. B. (2005), 'The repertoire of G-protein-coupled receptors in fully sequenced genomes.', Mol Pharmacol 67(5), 1414--1425.

50. Schiöth, H. B. & Fredriksson, R. (2005), 'The GRAFS classification system of G-protein coupled receptors in comparative perspective.', *Gen Comp Endocrinol* **142**(1-2), 94--101.

51. Gouldson, P. R.; Dean, M. K.; Snell, C. R.; Bywater, R. P.; Gkoutos, G. & Reynolds, C. A. (2001), 'Lipid-facing correlated mutations and dimerization in G-protein coupled receptors.', *Protein Eng* **14**(10), 759--767.

52. Pazos, F.; Helmer-Citterich, M.; Ausiello, G. & Valencia, A. (1997), 'Correlated mutations contain information about protein-protein interaction.', *J Mol Biol* **271**(4), 511--523.

53. Oliveira, L.; Paiva, A. C. M. & Vriend, G. (2002), 'Correlated mutation analyses on very large sequence families.', *Chembiochem* **3**(10), 1010—1017.

54. Kowarsch, A.; Fuchs, A.; Frishman, D. & Pagel, P. (2010), 'Correlated mutations: a hallmark of phenotypic amino acid substitutions.', *PLoS Comput Biol* **6**(9).

55. Ye, K.; Lameijer, E.-W. M.; Beukers, M. W. & Ijzerman, A. P. (2006), 'A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors.', *Proteins* **63**(4), 1018--1030.

56. Fatakia, S. N.; Costanzi, S. & Chow, C. C. (2009), 'Computing highly correlated positions using mutual information and graph theory for G protein-coupled receptors.', *PLoS One*

4(3), e4681.

57. Surgand, J.-S.; Rodrigo, J.; Kellenberger, E. & Rognan, D. (2006), 'A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors.', *Proteins* **62**(2), 509--538.

58. Brown, C. A. & Brown, K. S. (2010), 'Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my!', *PLoS One* **5**(6), e10779.

59. Fodor, A. A. & Aldrich, R. W. (2004), 'On evolutionary conservation of thermodynamic coupling in proteins.', *J Biol Chem* **279**(18), 19046--19050.

60. Choi, K. & Gomez, S. M. (2009), 'Comparison of phylogenetic trees through alignment of embedded evolutionary distances.', BMC Bioinformatics 10, 423.

61. Ye, K.; Lameijer, E.-W. M.; Beukers, M. W. & Ijzerman, A. P. (2006), 'A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors.', Proteins 63(4), 1018--1030.

62. Holm, L. & Sander, C. (1998), 'Removing near-neighbour redundancy from large protein sequence collections.', Bioinformatics 14(5), 423--429.

63. Thompson, J. D.; Higgins, D. G. & Gibson, T. J. (1994), 'CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.', Nucleic Acids Res 22(22), 4673--4680.

64. Eddy, S. R. (1998), 'Profile hidden Markov models.', Bioinformatics 14(9), 755--763.

65. Raghava, G. P. S. & Barton, G. J. (2006), 'Quantification of the variation in percentage identity for protein sequence alignments.', BMC Bioinformatics 7, 415.

66. May, A. C. W. (2004), 'Percent sequence identity; the need to be explicit.', Structure 12(5), 737--738.

67. Grishin, V. N. & Grishin, N. V. (2002), 'Euclidian space and grouping of biological objects.', Bioinformatics 18(11), 1523--1534.

68. Feng, D. F.; Johnson, M. S. & Doolittle, R. F. (1984), 'Aligning amino acid sequences: comparison of commonly used methods.', *J Mol Evol* **21**(2), 112—125.

69. Feng, D. F. & Doolittle, R. F. (1997), 'Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships.', *J Mol Evol* **44**(4), 361--370.

70. Dayhoff, M. O.; Schwartz, R. M. & Orcutt B. C. (1978), 'A model for evolutionary change in proteins.', *Atlas of Protein Sequence and Structure* **5,** 345--352.

71. Henikoff, S. & Henikoff, J. G. (1992), 'Amino acid substitution matrices from protein blocks.', *Proc Natl Acad Sci U S A* **89**(22), 10915--10919.

72. Jones, D. T.; Taylor, W. R. & Thornton, J. M. (1994), 'A mutation data matrix for transmembrane proteins.', *FEBS Lett* **339**(3), 269--275.

73. Ng, P. C.; Henikoff, J. G. & Henikoff, S. (2000), 'PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane.', *Bioinformatics* **16**(9), 760--766.

74. Müller, T.; Rahmann, S. & Rehmsmeier, M. (2001), 'Non-symmetric score matrices and the detection of homologous transmembrane proteins.', *Bioinformatics* **17 Suppl 1**, S182--S189.

75. Sutormin, R. A.; Rakhmaninova, A. B. & Gelfand, M. S. (2003), 'BATMAS30: amino acid substitution matrix for alignment of bacterial transporters.', *Proteins* **51**(1), 85—95.

76. Abdi, H.; O'Toole, A. J.; Valentin, D. & Edelman, B. (2005), 'DISTATIS: The Analysis of Multiple Distance Matrices.', *in* 'Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops - Volume 03', IEEE Computer Society, Washington, DC, USA, pp. 42.

77. Abdi H. & Valentin D. (2007), 'DISTATIS: the analysis of multiple distance matrices.', *in* 'N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics', Thousand Oaks (CA): Sage. pp. 284--290.

78. Abdi, H.; Dunlop, J. P. & Williams, L. J. (2009), 'How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS).', Neuroimage 45(1), 89--95.

79. Abdi H. (2007), 'Metric multidimensional scaling.', in 'N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics', Thousand Oaks (CA): Sage. pp. 598--605.

80. Perner, P., ed., (2007), *Machine Learning and Data Mining in Pattern Recognition, 5th International Conference, MLDM 2007, Leipzig, Germany, July 18-20, 2007, Proceedings*, Vol. 4571, Springer.

81. Rousseeuw, P. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', J. Comput. Appl. Math. 20, 53--65.

82. Kass, I. & Horovitz, A. (2002), 'Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations.', Proteins 48(4), 611--617.

83. Reva, B.; Antipin, Y. & Sander, C. (2007), 'Determinants of protein function revealed by combinatorial entropy optimization.', *Genome Biol* **8**(11), R232.

84. Miller, P. L.; Nadkarni, P.; Singer, M.; Marenco, L.; Hines, M. & Shepherd, G. (2001), 'Integration of multidisciplinary sensory data: a pilot model of the human brain project approach.', *J Am Med Inform Assoc* **8**(1), 34--48.

85. Xu, F.; Du, P.; Shen, H.; Hu, H.; Wu, Q.; Xie, J. & Yu, L. (2009), 'Correlated mutation analysis on the catalytic domains of serine/threonine protein kinases.', *PLoS One* **4**(6), e5913.

86. Halperin, I.; Wolfson, H. & Nussinov, R. (2006), 'Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families.', *Proteins* **63**(4), 832--845.

87. Larson, S. M.; Nardo, A. A. D. & Davidson, A. R. (2000), 'Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions.', *J Mol Biol* **303**(3), 433—446.

88. Fuchs, A.; Martin-Galiano, A. J.; Kalman, M.; Fleishman, S.; Ben-Tal, N. & Frishman, D. (2007), 'Co-evolving residues in membrane proteins.', *Bioinformatics* **23**(24), 3312--3319.

89. Cline, M. S.; Karplus, K.; Lathrop, R. H.; Smith, T. F.; Rogers, R. G. & Haussler, D. (2002), 'Information-theoretic dissection of pairwise contact potentials.', *Proteins* **49**(1), 7--14.

90. Martin, L. C.; Gloor, G. B.; Dunn, S. D. & Wahl, L. M. (2005), 'Using information theory to search for co-evolving residues in proteins.', *Bioinformatics* **21**(22), 4116--4124.

91. Dunn, S. D.; Wahl, L. M. & Gloor, G. B. (2008), 'Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.', *Bioinformatics* **24**(3), 333--340.

92. Tillier, E. R. M. & Lui, T. W. H. (2003), 'Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments.', Bioinformatics 19(6), 750--755.

93. Lockless, S. W. & Ranganathan, R. (1999), 'Evolutionarily conserved pathways of energetic connectivity in protein families.', *Science* **286**(5438), 295--299.

94. Dekker, J. P.; Fodor, A.; Aldrich, R. W. & Yellen, G. (2004), 'A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments.', *Bioinformatics* **20**(10), 1565--1572.

95. Süel, G. M.; Lockless, S. W.; Wall, M. A. & Ranganathan, R. (2003), 'Evolutionarily conserved networks of residues mediate allosteric communication in proteins.', *Nat Struct Biol* **10**(1), 59—69.

96. Olmea, O. & Valencia, A. (1997), 'Improving contact predictions by the combination of correlated mutations and other sources of sequence information.', *Fold Des* **2**(3), S25--S32.

97. McLachlan, A. D. (1971), 'Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551 .', *J Mol Biol* **61**(2), 409--424.

98. Miyata, T.; Miyazawa, S. & Yasunaga, T. (1979), 'Two types of amino acid substitutions in protein evolution.', J Mol Evol 12(3), 219--236.

99. Vicatos, S.; Reddy, B. V. B. & Kaznessis, Y. (2005), 'Prediction of distant residue contacts with the use of evolutionary information.', Proteins 58(4), 935--949.

100. Kawashima, S.; Ogata, H. & Kanehisa, M. (1999), 'AAindex: Amino Acid Index Database.', Nucleic Acids Res 27(1), 368--369.

101. Horn, F.; Weare, J.; Beukers, M. W.; Hörsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F. & Vriend, G. (1998), 'GPCRDB: an information system for G protein-coupled receptors.', *Nucleic Acids Res* **26**(1), 275--279.

102. Team, R. D. C. (2010), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. Available at http://cran.r-project.org/doc/manuals/refman.pdf.

103. Venables, W. N.; Smith, D.M. & Team, R. D. C. (2010), 'An introduction to R.', R Foundation for Statistical Computing, Vienna, Austria. Available at http://cran.r-project.org/doc/manuals/R-intro.pdf.

104. Ihaka, R. (1998), 'R: Past and Future History' in 'Proceedings of the 30th Symposium on the Interface.', S. Weisberg Ed., pp. 392--396. Available at http://cran.r-project.org/doc/html/interface98-paper/paper.html.

105. Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A. J.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. Y. H. & Zhang, J. (2004), 'Bioconductor: open software development for computational biology and bioinformatics.', *Genome Biol* **5**(10), R80.

106. Charif, D. & Lobry, J. (2007), SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis., *in* U. Bastolla; M. Porto; H.E. Roman & M. Vendruscolo, ed., 'Structural approaches to sequence evolution: Molecules, networks, populations', Springer Verlag, New York, pp. 207-232.

107. Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A. & Caves, L. S. D. (2006), 'Bio3d: an R package for the comparative analysis of protein structures.', *Bioinformatics* **22**(21), 2695--2696.

108. Rosenkilde, M. M.; Benned-Jensen, T.; Frimurer, T. M. & Schwartz, T. W. (2010), 'The minor binding pocket: a major player in 7TM receptor activation.', *Trends Pharmacol Sci* **31**(12), 567--574.

109. Fox, J. (2005), 'The R Commander: A Basic-Statistics Graphical User Interface to R.', Journal of Statistical Software 14(9).

110. Sali, A. & Blundell, T. L. (1993), 'Comparative protein modelling by satisfaction of spatial restraints.', *J Mol Biol* **234**(3), 779--815.

ANNEXES

A. Formule du calcul de dissimilarités

Le score de dissimilarités est obtenu auprès du score de similarité qui est obtenu comme suit :

$$V_{ij} = \frac{S_{ij} - S_{ij}^{rand}}{T_{ij} - S_{ij}^{rand}}$$

avec :

$S_{i,j} = \sum_{k \in K_{ij}} s(A_{ik} - A_{jk}) \; / \; l(K_{ij})$ qui est le score par site.

$T_{i,j} = 0.5 \sum_{k \in K_{ij}} s((A_{ik}, A_{ik}) + (A_{jk}, A_{jk})) \; / \; l(K_{ij})$ qui est la limite supérieure moyenne du score par site.

$S_{i,j}^{rand} = \sum_{a=1}^{20} \sum_{b=1}^{20} f^{i}_{j}(a) \; f^{j}_{i}(b) \; s(a,b)$ qui est le score par site attendu pour des séquences aléatoires.

A : alignement de n séquences.
i et j : séquences de chaque paire (i, j).
K_{ij} : position sans gap pour les séquences i et j.
$l(K_{ij})$: nombre d'éléments dans K_{ij}.
K : position dans l'alignement.
$s(a, b)$: score de similarité entre les acides aminés a et b.
$f^{i}_{j}(a)$: proportion de l'acide aminé a de la $i^{ième}$ séquence de A pour tous les sites K_{ij}.

Plus précisément, S_{ij} est la valeur de similarité calculée à partir de la paire de séquences authentiques. T_{ij} est la valeur lorsque chacune des séquences est alignée avec elle-même. S^{rand} est la moyenne des valeurs lorsque les séquences de même longueur et composition sont mélangées puis alignées. Les valeurs de similarité s'échelonnent de 0, pour des séquences aléatoires, à 1, pour des séquences identiques. Pour que la matrice soit applicable aux méthodes statistiques de réduction dimensionnelle, la similarité est convertie en dissimilarité en appliquant la formule suivante : dissimilarité = 1 – similarité.

B. Formules des méthodes de corrélation entre groupes et séquences

B.1 Rappel des notations

x : l'ensemble des acides aminés présents à la position i.

g : l'ensemble des groupes.

N : le nombre total de séquences.

$f(g)$: fréquence du groupe g.

$f_x(i)$: fréquence de l'acide aminé x à la position i.

$f_x(i,g)$: fréquence de l'acide aminé x à la position i pour le groupe g.

$f_x(i^g)$: fréquence de l'acide aminé x présent dans le groupe g à la position i (à ne pas confondre avec la fréquence précédente car celle-ci se calcule par rapport au nombre de séquences totales).

$f_{(x,y)}$: fréquence de la paire d'acides aminés (x,y) aux positions i et j.

B.2 Méthodes basées sur le chi-2

B.2.1 Définitions de N_{obs} et N_{ex}

- N_{obs} et N_{ex} tels qu'ils sont définis dans la littérature pour la méthode d'AMC OMES :

  N_{obs} : le nombre de paires d'acides aminés observées aux positions i et j.

  N_{ex} : le nombre de paires d'acides aminés attendues aux positions i et j.

  Le calcul de N_{ex} est basé sur le calcul de la fréquence de chaque acide aminé à chaque position :

$$N_{ex} = \frac{N_x N_y}{N_{valid}}$$

  N_x : le nombre de fois où l'acide aminé x apparaît en position i.

  N_y : le nombre de fois où l'acide aminé y apparaît en position j.

- N_{obs} et N_{ex} adaptés pour le calcul entre groupes et séquences :

$$N_{obs} = N \times f(g) \times f_x(i,g)$$

$$N_{ex} = N \times f(g) \times f_x(i)$$

B.2.2 Démonstration d'OMES1

La fonction d'AMC pour OMES1 est définie comme suit :

$$\text{OMES1}_{(i,j)} = \sum_{x,v} \frac{\left(N_{obs} - N_{ex}\right)^2}{N_{ex}}$$

En remplaçant N_{obs} et N_{ex} dans la formule d'OMES1, on obtient :

$$\text{OMES1}_i = \sum_g \sum_x \frac{\left(N \times f(g) \times f_x(i,g) - N \times f(g) \times f_x(i)\right)^2}{N \times f(g) \times f_x(i)}$$

$$\text{OMES1}_i = \sum_g \sum_x \frac{N^2 \times (f(g))^2 \times \left(f_x(i,g) - f_x(i)\right)^2}{N \times f(g) \times f_x(i)}$$

$$OMES1_i = N \sum_g \sum_x \frac{f(g) \times \left(f_x(i,g) - f_x(i)\right)^2}{f_x(i)}$$

La formule ci-dessus est celle qui est utilisée pour la mise en évidence des corrélations entre groupes et séquences pour la méthode adaptée à partir d'OMES1 (voir chapitre 5.7.3.1).

B.2.3 Démonstration d'OMES2

La fonction d'AMC pour OMES2 est définie comme suit :

$$\text{OMES2}_{(i,j)} = \sum_{x,v} \frac{\left(N_{obs} - N_{ex}\right)^2}{N}$$

En remplaçant N_{obs} et N_{ex} dans la formule d'OMES2, on obtient :

$$\text{OMES2}_i = \sum_g \sum_x \frac{\left(N \times f(g) \times f_x(i,g) - N \times f(g) \times f_x(i)\right)^2}{N}$$

$$\text{OMES2}_i = \sum_g \sum_x \frac{N^2 \times (f(g))^2 \times \left(f_x(i,g) - f_x(i)\right)^2}{N}$$

$$\text{OMES2}_i = N \sum_g \sum_x (f(g))^2 \times \left(f_x(i,g) - f_x(i)\right)^2$$

La formule ci-dessus est celle qui est utilisée pour la mise en évidence des corrélations entre

groupes et séquences pour la méthode adaptée à partir d'OMES2 (voir chapitre 5.7.3.1).

B.3 Méthodes basées sur l'information mutuelle

B.3.1 Démonstration de la MI classique

La fonction d'AMC pour la MI classique est définie comme suit :

$$\mathrm{MI}_{(i,j)} = \sum_{x,y} f_{(x,y)} \ln \frac{f_{(x,y)}}{f_{(x)} f_{(y)}}$$

En adaptant la formule en sachant que la position *j* correspond à l'appartenance au groupe *g*, on obtient :

$$\mathrm{MI}_i = \sum_g \sum_x f_x(i^g) \times \ln \frac{f_x(i^g)}{f_x(i) \times f(g)}$$

La formule ci-dessus est celle qui est utilisée pour la mise en évidence des corrélations entre groupes et séquences pour la méthode adaptée à partir de MI classique (voir chapitre 5.7.3.2).

B.3.2 Démonstration de la MI normalisée

La fonction d'AMC pour la MI normalisée est définie comme suit :

$$\mathrm{MIr}_{(i,j)} = \frac{\mathrm{MI}_{(i,j)}}{\mathrm{JE}_{(i,j)}}$$

avec la JE (entropie jointe) qui est égale à :

$$\mathrm{JE}_{(i,j)} = -\sum_{xy} f_{(x,y)} \ln f_{(x,y)}$$

En adaptant la formule en sachant que la position *j* correspond à l'appartenance au groupe *g*, on obtient :

$$\mathrm{MIr}_i = \frac{\mathrm{MI}_i}{-\sum_g \sum_x f_x(i^g) \times \ln f_x(i^g)}$$

B.4 Optimisation de l'entropie combinatoire

Cette méthode présente l'avantage de calculer d'emblée la corrélation entre groupes et séquences contrairement aux précédentes méthodes qui sont adaptées historiquement à la corrélation de paires de positions. Les valeurs du score sont toujours inférieures ou égales à zéro et le score est calculé par la différence d'entropie entre une distribution observée et attendue :

$$\Delta S_i = S_i - \tilde{S}_i$$

avec :

$S_i = \sum_g ln Z_{i,g}$ qui est une mesure des différentes distributions d'acides aminés observées.

$\tilde{S}_i = \sum_g ln \tilde{Z}_{i,g}$ qui est la valeur maximale de S_i.

$Z_{i,g} = \frac{N_{g!}}{\prod_{x=1,\ldots,21} N_{x,i,g!}}$ qui est le nombre observé de permutations pour une position *i* du groupe *g*.

$\tilde{Z}_{i,g} = \frac{N_{g!}}{\prod_{x=1,\ldots,21} \tilde{N}_{x,i,g!}}$ qui est le nombre attendu de permutations pour une position *i* du groupe *g*.

$N_{x,i}$: nombre d'acides aminés de type *x* à la position *i*.

N_g : nombre de séquences dans le groupe *g*.

$N_{x,i,g}$: nombre observé d'acides aminés de type *x* à la position *i* pour le groupe *g*.

$\tilde{N}_{x,i,g} = N_g N_{x,i} / N$: nombre attendu d'acides aminés de type *x* à la position *i* pour le groupe *g*.

C. Formules des méthodes d'AMC

C.1 Méthodes basées sur le chi-2

La formule d'OMES1 est définie comme suit :

$$\text{OMES2}_{(i,j)} = \sum_{x,y} \frac{\left(N_{obs} - N_{ex}\right)^2}{N_{valid}}$$

La formule d'OMES2 est légèrement différente :

$$\text{OMES2}_{(i,j)} = \sum_{x,y} \frac{\left(N_{obs} - N_{ex}\right)^2}{N}$$

avec :

x,y : l'ensemble des paires d'acides aminés présentes aux positions i et j.

N_{valid} : le nombre de paires d'acides aminés aux positions i et j.

N_{obs} : le nombre de paires d'acides aminés observées aux positions i et j.

N_{ex} : le nombre de paires d'acides aminés attendues aux positions i et j.

Le calcul de N_{ex} est basé sur le calcul de la fréquence de chaque acide aminé à chaque position :

$$N_{ex} = \frac{N_x N_y}{N_{valid}}$$

N_x : le nombre de fois où l'acide aminé x apparaît en position i.

N_y : le nombre de fois où l'acide aminé y apparaît en position j.

OMES est symétrique car $\text{OMES}_{(i,j)} = \text{OMES}_{(j,i)}$ et le calcul du score ne se fera donc que pour $j > i$. Dans le cas de positions i et j parfaitement conservées, la position i ne dispose que de l'acide aminé x et la position j n'affiche que l'acide aminé y. Le score obtenu sera de 0 car $N_{ex} = N_{valid} = N_{obs}$. Dans le cas où une des positions est parfaitement conservée, le score obtenu sera également de 0 car $N_{ex} = N_{obs}$.

C.2 Méthodes basées sur l'information mutuelle

La formule de la MI classique est définie comme suit :

$$MI_{(i,j)} = \sum_{x,y} f_{(x,y)} \ln \frac{f_{(x,y)}}{f_{(x)} f_{(y)}}$$

avec :

$f_{(x)}$: la fréquence qu'un acide aminé *x* soit présent à la position *i*.

$f_{(y)}$: la fréquence qu'un acide aminé *y* soit présent à la position *j*.

$f_{(x,y)}$: la fréquence qu'une paire d'acides aminés *xy* soit présente aux positions *i* et *j*.

MI est symétrique. Dans le cas de positions *i* et *j* parfaitement conservées, le score est de 0 car $f_{(x)} f_{(y)} = f_{(x,y)}$. Dans le cas où une des positions est parfaitement conservée, le score obtenu sera également de 0 et les positions *i* et *j* seront indépendantes. Si un acide aminé de type *x* n'interagit qu'avec un acide aminé de type *y*, $f_{(x)} = f_{(x,y)}$. La formule de MI devient :

$$MI_{(i,j)} = \sum_{x,y} f_{(x)} \ln \frac{1}{f_{(y)}}$$

D'après cette formule, en sachant que $f_{(y)}$ ne peut être inférieur à $f_{(x)}$, MI favorise les paires de positions avec un nombre maximal de paires d'acides aminés différentes. Si pour une paire de positions donnée, la probabilité d'avoir chacune de toutes les paires d'acides aminés possibles est la même, le score que donnera MI vaudra 1. L'influence de l'entropie peut être partiellement diminuée par la MIr où le dénominateur correspond à la JE :

$$MIr_{(i,j)} = \frac{MI_{(i,j)}}{JE_{(i,j)}}$$

Cette normalisation permet de réduire uniquement l'influence de l'entropie pour la MI. D'autres normalisations sont envisageables pour minimiser les effets des trois facteurs précédents : la MIp et la MIa. Les méthodes MIp et MIa ont pour objectif de séparer le signal véritable (MI_{sf}) du bruit de fond (MI_b) causé par les deux facteurs restants, la taille de l'AMS et les contraintes phylogénétiques. MI_b est exprimée par le terme Average Product Correction (APC) et représente une excellente approximation du bruit de fond partagé par les positions *i* et *j* de la paire. Par conséquent, la

différence entre APC et la MI total pour une paire de positions devrait isoler MI_{sf}. La notation MIp est utilisée pour indiquer la différence :

$$MIp_{(i,j)} = MI_{(i,j)} - APC_{(i,j)}$$

MIp, en plus de réduire l'influence de la taille de l'AMS et les contraintes phylogénétiques, annule également l'influence de l'entropie. L'affinité pour MI_b peut être additive plutôt que multiplicative. Cette hypothèse mène à l'élaboration du terme Average Sum Correction (ASC). La différence entre ASC et la MI totale est annotée par MIa :

$$MIa_{(i,j)} = MI_{(i,j)} - ASC_{(i,j)}$$

Ces corrections améliorent significativement l'identification des positions qui varient de paire et qui sont proximales au niveau de la structure protéique. La correction par ASC fonctionne presque aussi bien que la correction APC dans la détection d'acides aminés proximaux.

Les relations phylogénétiques entre un groupe de séquences proviennent d'un schéma de substitutions qui est similaire à différentes positions dans l'AMS. Bien que les positions dans une protéine peuvent évoluer de manière indépendante par rapport à une autre, elles suivent la même stratégie phylogénétique avec de fortes corrélations entre elles. Ces effets phylogénétiques sont quantifiables par la mesure de la corrélation entre les positions.

Par conséquent, les positions influencées par la phylogénie sont fortement corrélées entre beaucoup d'autres et auront de nombreuses interdépendances, alors que les positions qui tendent à avoir des relations de type fonctionnel affichent des corrélations avec seulement un nombre limité de positions. Ce degré d'interdépendance est mesuré par la méthode MINT et son score est donné par la formule suivante :

$$MINT_{(i,j)} = D_{(i,j)} \times E_{(i,j)}$$

avec :

$D_{(i,j)}$ qui est le ratio de dépendance. Pour déterminer le degré de corrélation entre une position *i* et *j* qui n'est pas du à de la phylogénie, le score de MINT est pondéré par ce ratio.

$E_{(i,j)}$ qui est le facteur d'entropie. Il permet de réduire le ratio de dépendance pour des paires dont les positions affichent des entropies extrêmes.

C.3 Méthodes asymétriques

La formule de la méthode SCA est définie comme suit :

$$\mathrm{SCA}_{(i,j)} = \sqrt{\sum_{y} (\ln P^{y}_{j|\delta\, i} - \ln P^{y}_{j})^2}$$

avec :

y : l'ensemble des acides aminés présents à la position *j*.

P^{y}_{j} : la probabilité binomiale qu'un acide aminé *y* soit présent à la position *j*.

$P^{y}_{j|\delta\, i}$: la probabilité binomiale qu'un acide aminé *y* soit présent à la position *j* en sachant que cet acide aminé appartient au sous-alignement déterminé par la perturbation à la position *i*.

La probabilité binomiale compare la fréquence de la position donnée avec la fréquence moyenne dans toutes les protéines :

$$P^{x} = \frac{N!}{n_x!(N - n_x)!} P_x^{n_x} (1 - P_x)^{N - n_x}$$

avec :

N : valeur arbitraire égale à 100.

n_x : pourcentage numérique de séquences avec l'acide aminé *y* à la position donnée.

P_x : fréquence moyenne de l'acide aminé *y* à toutes les positions, déterminée à partir de 36 498 entrées de Swiss-Prot (protéines eucaryotes).

La formule de la méthode ELSC est définie comme suit :

$$ELSC_{(i,j)} = -\ln \prod_{y}^{20} \frac{C^{n_{y,j}}_{N_{y,j}}}{C^{m_{y,j}}_{N_{y,j}}}$$

avec :

$N_{y,j}$: le nombre d'acides aminés *y* à la position *j* dans l'alignement total.

$n_{y,j}$: le nombre d'acides aminés *y* à la position *j* dans le sous-alignement.

$m_{y,j}$: le nombre d'acides aminés *y* à la position *j* dans le sous-alignement idéalisé, c'est à dire si la composition en acides aminés du sous-alignement était la même que celle de l'alignement total.

C_n^k : le coefficient binomial d'indices *n* et *k*, défini par la formule suivante (avec $0 \leq k \leq n$) :

$$C_n^k = \frac{n!}{k!\ (n - k)!}$$

Des valeurs intermédiaires sont calculées pour obtenir les valeurs de $m_{y,j}$:

- $mf_{y,j}$: $mf_{y,j} = (N_{y,j} / N_{total})$ x n_{total}, en sachant que N_{total} est le nombre de séquences de l'alignement total et n_{total}, le nombre de séquences du sous-alignement (f pour *float*).
- $mi_{y,j}$: $mf_{y,j}$ qui peut être décimal est arrondi à l'entier inférieur (i pour *integer*).
- le reste : $r_{y,j} = (mf_{y,j} - mi_{y,j})$.

Les valeurs de $m_{y,j}$ sont soumises à la contrainte suivante : $\sum_y m_{y,j} = \sum_y n_{y,j}$.

Les valeurs $m_{y,j}$ sont triées par ordre décroissante de leur reste, puis par ordre alphabétique inverse des acides aminés respectifs si une égalité existe. Les valeurs triées de $m_{y,j}$ sont incrémentées de 1, l'une après l'autre, tant que la condition n'a pas été rencontrée. ELSC n'est pas symétrique. Si les positions *i* et *j* sont indépendantes, la distribution d'acides aminés du sous-alignement à la position *j* devrait être similaire à celle de l'alignement total à la position *j*. Le score sera proche de 0. Cependant, si les deux positions covarient, la composition du sous-alignement à la position *j* devrait être influencée par la contrainte de la position *i* et le score sera plus élevé.

C.4 Méthodes basées sur des matrices de substitution

Voici la formule de la méthode MCBASC en sachant que pour chaque paire de positions de l'AMS, nous avons une matrice *N* x *N* dans laquelle chaque élément (*x* et *y* allant de 1 à *N*) représente la similarité de la paire (*x*,*y*) selon :

$$MCBASC_{(i,j)} = \frac{1}{N^2} \frac{\sum_{xy} \left(S_{ixy} - S_i\right)\left(S_{jxy} - S_j\right)}{\sigma_i \sigma_j}$$

avec :

S_{ixy} et S_{jxy} : les scores pour la paire d'acides aminés *xy*, pour les positions *i* et *j*, respectivement.

S_i et S_j : moyennes de tous les scores S_{ixy} et S_{jxy} de la matrice *N* x *N* pour les positions *i* et *j*, respectivement.

σ_i et σ_j : écart-types de tous les scores S_{ixy} et S_{jxy} de la matrice *N* x *N* pour les positions *i* et *j*, respectivement. La formule de l'écart-type est celle de l'échantillon et non de la population.

Les valeurs du score s'échelonnent de -1 à 1, avec un score de -1 ou de 1 indiquant des positions fortement corrélées. Les scores de MCBASC seront donnés en valeur absolue pour pouvoir les comparer à ceux des autres méthodes. Le score est indéfini si les positions *i* et *j* sont parfaitement conservées pour toutes les entrées : $\sigma_i = \sigma_j = 0$. Le score est de 1 lorsque les positions *i* et *j* sont parfaitement identiques : $\sigma_i = \sigma_j$ *et* $S_i = S_j$ et $(S_{ixy} - S_i) = (S_{jxy} - S_j)$. Ce comportement est particulier car les méthodes OMES, MI et SCA donne un score de 0 si les positions sont parfaitement conservées.

La méthode CMAV est capable de rechercher les mutations corrélées dans un AMS à partir des caractéristiques physico-chimiques des acides aminés. Ces caractéristiques sont disponibles sous la forme de descripteurs qui ont été obtenus à partir d'informations contenues dans la AAindex database. L'application d'une analyse en composantes principales sur un certain nombre de données de cette base a permis de distinguer les propriétés, de les séparer et de construire les vecteurs. Les descripteurs sélectionnés sont au nombre de trois : PRIN1 pour les propriétés d'hydrophobie, PRIN2 qui spécifique de la taille et PRIN3 qui est lié au pk_N, c'est à dire relatif aux propriétés ioniques. Le score de mutations corrélées entre deux positions *i* et *j* est défini par une formule proche de celle du coefficient de corrélation de Bravais-Pearson :

$$CMAV_{(i,j)} = \frac{1}{N} \sum_{xy} \frac{(q_i^x - m_i)(q_j^y - m_j)}{\sigma_i \sigma_j}$$

q_i^x et q_j^y : la valeur du descripteur pour l'acide aminé *x* à la position *i* et pour l'acide aminé *y* à la position *j*, respectivement.

m_i et m_j : la moyenne du descripteur pour les position i et j.

L'amplitude, plus que le signe du coefficient de corrélation, est considérée comme étant l'indicateur de la covariation entre les deux positions d'acides aminés. Lorsque une mutation corrélée affecte les deux positions :

- le coefficient de corrélation peut être positif dans le cas où les valeurs de la propriété peuvent simultanément être réduites ou augmentées,

- le coefficient de corrélation peut être négatif dans le cas où la valeur de la propriété de l'une peut baisser alors que celle de l'autre position peut augmenter ; la valeur moyenne de cette propriété reste la même.

Dans les deux cas, la méthode CMAV révèle des interactions possibles entre des acides aminés relativement proches. Ainsi, les scores de CMAV seront donnés en valeur absolue pour pouvoir les comparer à ceux des autres méthodes.

D. Applications du package R bio2mds

D.1 Contenu

Le package se présente sous la forme d'un dossier *bio2mds* contenant quatre dossiers et deux fichiers :

- le dossier *data* contient deux jeux de données au format R data (rda) :
  - `gpcr` : matrices de différences, matrices de dissimilarités, correspondance entre sous-familles et codes couleur pour les RCPG de classe A non-olfactifs de cinq espèces animales.
  - `sub.mat` : différentes matrices de substitution (des PAM, des BLOSUM, PAM250TM et PHAT).

- le dossier *inst* (installation) contient un dossier *ams*, contenant quelques exemples d'AMS provenant de la base de données GPCRDB :

  - GPCR.angiotensin.fa : au format FASTA contenant les RCPG d'angiotensines.
  - GPCR.cannabinoid.fa : au format FASTA contenant les RCPG des cannabinoïdes.
  - GPCR.dopamine.fa : au format FASTA contenant les RCPG des dopamines.
  - GPCR.dopamine.msf : au format Multiple Sequence Format (MSF) contenant les RCPG des dopamines.

- le dossier *man* (manual) contient la documentation pour les fonctions et les jeux de données au format R documentation (Rd), très proche du LaTex.

- le dossier *R* contient 21 fonctions codées en language R :

  - `aa.to.class` : convertit le code des acides aminés à 1 lettre en un code spécifique suivant différentes classifications.
  - `dif` : calcule le degré de différence entre deux séquences protéiques.

- `dis` : calcule le degré de dissimilarité entre deux séquences protéiques.
- `distatis` : réalise une DISTATIS sur une liste de matrices de distances.
- `is.aa` : évalue la présence d'acides aminés dans une séquence.
- `is.gap` : évalue la présence de gaps dans une séquence.
- `kmeans.iter` : réalise des itérations de K-moyennes.
- `mat.dif` : calcule une matrice de differences entre deux AMS.
- `mat.dis` : calcule une matrice de dissimilarité entre deux AMS.
- `mmds` : réalise une MDS métrique avec gestion d'élèments supplémentaires.
- `plot.distatis` : trace un graphique global suite à une analyse DISTATIS.
- `plot.distatis.F` : trace un graphique pour la matrice de compromis F suite à une analyse DISTATIS.
- `plot.distatis.G` : trace un graphique pour la matrice des facteurs G suite à une DISTATIS.
- `plot.mmds` : trace un graphique global suite à une MDS métrique.
- `plot.mmds.point` : trace un graphique pour la matrice des facteurs F (et Fsup) suite à une MDS métrique.
- `plot.scree` : trace un graphique d'éboulis.
- `random.msa` : produit un AMS aléatoire suivant différentes charactéristiques.
- `read.fasta` : lit un AMS au format FASTA.
- `read.msf` : lit un AMS au format MSF.
- `sil.score` : calcule le score de silhouette.
- `write.mmds.pymol` : produit un fichier PDB et PyMoL (PML) suite à une MDS métrique.

- le fichier DESCRIPTION donne des informations générales sur le package (titre, version, date, auteurs, dépendances...).

- le fichier NAMESPACE permet de spécifier, entre autres, quelles variables devront être importées à partir d'autres packages. Par exemple, cela évite les conflits lorsque deux

fonctions arborent le même nom dans le cadre d'une dépendance (*i.e.*, un import).

D.2 Structure

bio2mds affiche une structure ordonnée et possède un fonctionnement spécifique (**Figure 21**). Les fonctions peuvent se diviser en deux types :

- les principales : elles servent à lire des AMS (`read.fasta` et `read.msf`), à calculer des distances entre AMS (`mat.dif` et `mat.dis`), à réaliser des analyses exploratoires (`distatis` et `mmds`) et à tracer des représentations graphiques (`plot.distatis` et `plot.mmds`).

- les secondaires : elles constituent des sous-routines aux fonctions principales (`aa.to.class`, `dif`, `dis`, `is.aa`, `is.gap`, `plot.distatis.F`, `plot.distatis.G`, `plot.mmds.point` et `plot.scree`). Les 4 restantes (`kmeans.iter`, `random.msa`, `sil.score` et `write.mmds.pymol`) n'y sont pas impliquées.

Le jeu de données `gpcr` peut être utilisé par les fonctions d'analyses exploratoires ainsi que pour la réalisation des représentations graphiques. Le jeu de données `sub.mat`, quant à lui, est utilisé par certaines fonctions secondaires. Toutes les fonctions, qu'elles soient principales ou secondaires, peuvent être exécutées seules. De plus, les méthodes exploratoires ne sont pas réservées qu'aux données biologiques et les jeux de données peuvent être exploités par des fonctions extérieures à *bio2mds*. Les éléments de ce package permettent de reproduire la plupart des résultats présentés dans l'article (voir chapitre 6.1).

D.3 Documentation

Tous les détails qui concernent, à la fois, le package lui-même, les jeux de données et les fonctions sont présents dans la documentation R de *bio2mds*. Elle regroupe les informations du fichier DESCRIPTION et celles des fichiers du dossier *man*. En général, cette documentation

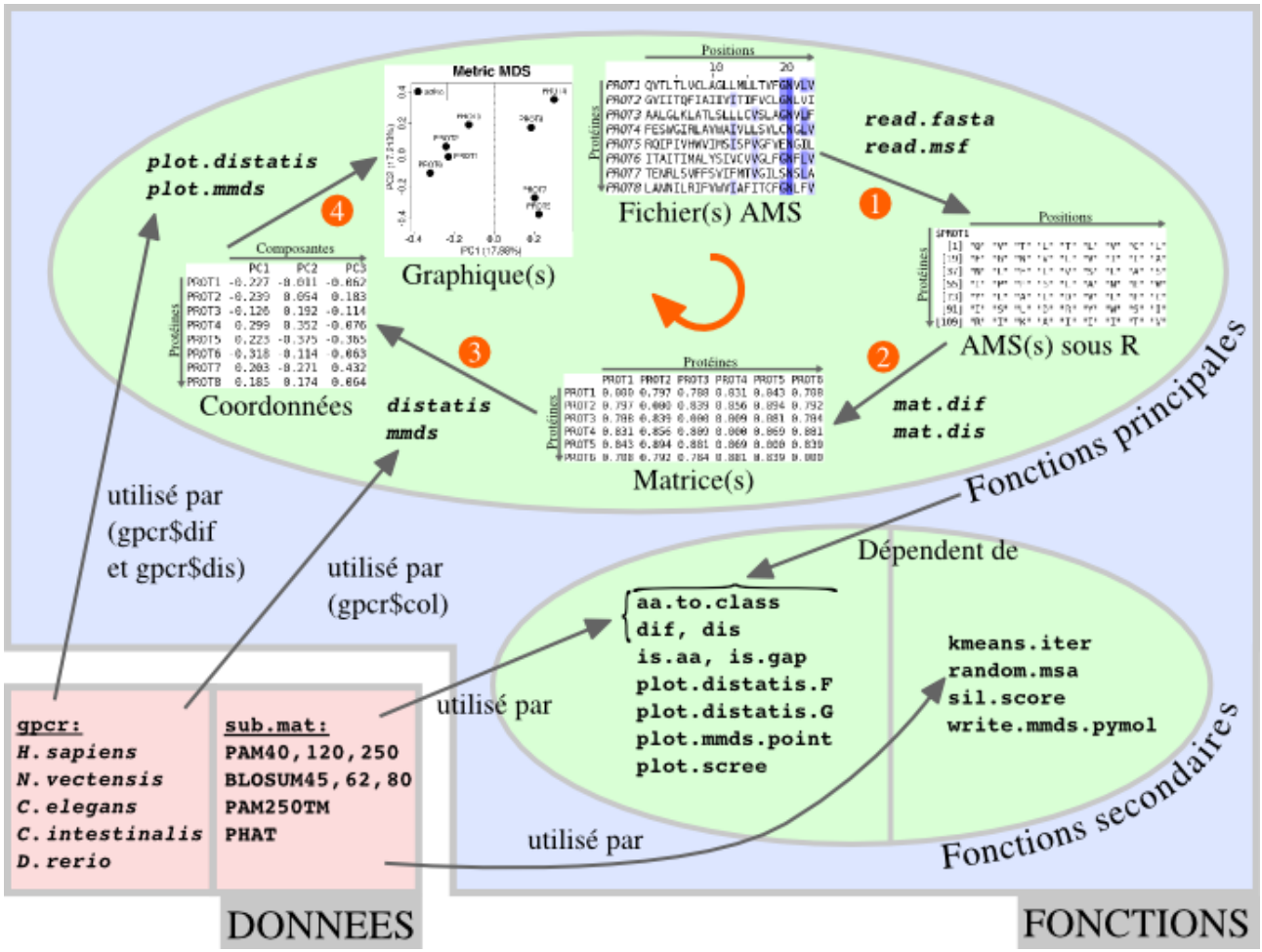

Figure 21 : Workflow du package R bio2mds

Le package est constitué de deux types d'éléments : les jeux de données (en rouge) et les fonctions (en vert). Chacune des données et des fonctions est illustrée en gras. Les ellipses représentent les deux types de fonctions (fonctions principales et secondaires). Les chiffres entourés en orange indiquent le sens de la succession de fonctions qui permettent de passer d'un fichier AMS à une représentation graphique. Le présent exemple concerne une MDS métrique sur quelques séquences de RCPG d'*H. Sapiens*.

contient pour chaque élément du package une série d'informations :

- description : description concise du rôle de la fonction (ou du jeu de données).
- usage : synopsis de la fonction (ou du jeu de données).
- arguments : arguments dans le cadre d'une fonction.
- details : détails techniques, concernant l'algorithme par exemple.
- value : ce que donne la fonction en retour.
- note : des remarques diverses.
- reference : références bibliographiques.
- see also : autres fonctions sous la même thématique.
- examples : exemples fonctionnels.

bio2mds n'est disponible qu'au laboratoire de l'équipe de bioinformatique. Le code source reste confidentiel car il n'a pas encore été publié. La documentation complète n'est disponible qu'au laboratoire. Des exemples illustrés sont présentés dans le chapitre suivant.

D.4 Applications

Après avoir installé et exécuté l'application R, veuillez-vous placer dans la console R pour saisir les commandes des exemples d'applications du package. Pour éviter des erreurs durant l'analyse, il est important de respecter les formats demandés. Les exemples d'AMS du dossier *inst* et le jeu de données `gpcr` permettent d'avoir un aperçu du formatage des données qu'on doit soumettre aux différentes fonctions.

D.4.1 Initialisation

bio2mds est chargé dans la session :

```
>library(bio2mds)
```

Le jeu de données `gpcr` est rendu accessible :

```
>data(gpcr)
```

Pour obtenir le chemin d'accès courant :

```
>getwd()
```

Il est conseillé de modifier le chemin d'accès pour accéder aux fichiers de *bio2mds* :

```
>dir <- system.file("msa", package = "bio2mds")
>setwd(dir)
```

D.4.2 Lecture d'un AMS

L'exemple d'AMS utilisé est au format FASTA. La fonction `read.fasta` lit et enregistre les données dans la variable `aln` :

```
>aln <- read.fasta("GPCR.angiotensin.fa")
```

Pour afficher une séquences de l'AMS, il suffit de placer le signe `$` entre le nom de la variable et le nom de la séquence. Pour afficher certaines positions de la séquence, il suffit de spécifier les indices correspondants. Voici un exemple affichant les 50 premières positions du récepteur agtr1 d'*H. sapiens* :

```
>aln$agtr1_human[1:50]

[1] "D" "D" "C" "P" "K" "A" "G" "R" "H" "N" "Y" "I" "F" "V" "M"
"I" "P" "T" "L" ...
```

D.4.3 Calcul de distances

Les récepteurs d'angiotensines de *M. musculus* et d'*H. sapiens* sont enregistrés dans les variables `mouse` et `human` respectivement :

```
>mouse <- aln[c("a2aut5_mouse", "agtr2_mouse", "q32mf7_mouse",
"agtra_mouse", "agtrb_mouse", "g109a_mouse")]
>human <- aln[c("agtr2_human", "q8tbk4_human", "q6nup5_human",
"agtr1_human", "d3dng8_human", "g109b_human", "g109a_human")]
```

Une matrice de distances, basée sur la différence, est calculée avec la fonction `mat.dif` entre

les récepteurs d'angiotensines de *M. musculus* et d'h. *Sapiens* :

```
>aln.dif <- mat.dif(mouse, human)
```

La matrice est affichée :

```
>aln.dif
              agtr2_human q8tbk4_human q6nup5_human agtr1_human
a2aut5_mouse        0.036        0.645        0.641       0.641
agtr2_mouse         0.036        0.645        0.641       0.641
q32mf7_mouse        0.641        0.043        0.039       0.039
...
```

D.4.4 Comparaison des matrices de distances

Le jeu de données `gpcr` est utilisé pour illustrer la manière dont on peut l'exploiter. Tout d'abord, une liste de matrices `list.mat` est déclarée et contient les matrices de distances calculées à partir des RCPG de classe A non-olfactifs d'*H. Sapiens* issus du jeu de données `gpcr` :

```
>list.mat <- list(
dif = gpcr$dif$sapiens.sapiens,
PAM40 = gpcr$dis$sapiens.sapiens$PAM40,
PAM120 = gpcr$dis$sapiens.sapiens$PAM120 ...
```

Les noms des éléments de `list.mat` sont affichés pour vérifier le contenu de la variable :

```
>names(list.mat)
[1]   "dif"   "PAM40"   "PAM120"   "PAM250"   "BLOSUM45"   "BLOSUM62"
"BLOSUM80" ...
```

`list.mat` contient une matrice de différences et huit matrices de dissimilarités correspondant aux huit matrices de substitution de `sub.mat`. Une DISTATIS est réalisée sur la liste `list.mat` et le résultat est stocké dans la variable `distatis1` :

```
>distatis1 <- distatis(list.mat)
```

La matrice des facteurs G de la variable `distatis1` est affichée :

```
>distatis1$G

             PC1     PC2     PC3
dif       -0.970   0.238 -0.015
PAM40     -0.997   0.019   0.057
PAM120    -0.996 -0.037   0.072
PAM250    -0.991 -0.091   0.090
...
```

La matrice des facteurs G contient trois coordonnées pour chacune des matrices. Ensuite, sa représentation graphique est affichée avec la fonction graphique `plot.distatis.G`, avec couleurs et légendes pour les matrices de distances (**Figure 22**) :

```
>plot.distatis.G(distatis1, col = rainbow(length(distatis1)), lab
= TRUE)
```

Les coordonnées des matrices sur la première composante ont toutes le même signe. La convention veut qu'elles soient positives. La première composante explique ce qui est en commun entre chaque matrice et toutes les autres : plus la coordonnée est élevée, plus la matrice partage de l'information avec les autres matrices. Dans notre cas, la première composante explique 98% de la variance totale et montre que la matrice de différences partage légèrement moins d'informations que les matrices de dissimilarités. La séparation est renforcée entre la matrice de différences et les matrices de dissimilarités sur la deuxième composante même si elle n'explique que 1% de la variance totale. La matrice de différences est donc plus singulière et les matrices de dissimilarités sont proches malgré des matrices de substitution spécifiques. Par la suite, seule la matrice de différences sera utilisée à cause de la difficulté de distinguer les matrices de dissimilarités entre-elles.

D.4.5 MDS métrique en 2 dimensions

Une MDS métrique est réalisée sur la matrice de différences entre les séquences d'*H. sapiens* en tant qu'éléments actifs et le résultat est stocké dans la variable `mmds1` pour les 10 premières composantes principales :

```
>mmds1 <- mmds(active = gpcr$dif$sapiens.sapiens, pc = 10)
```

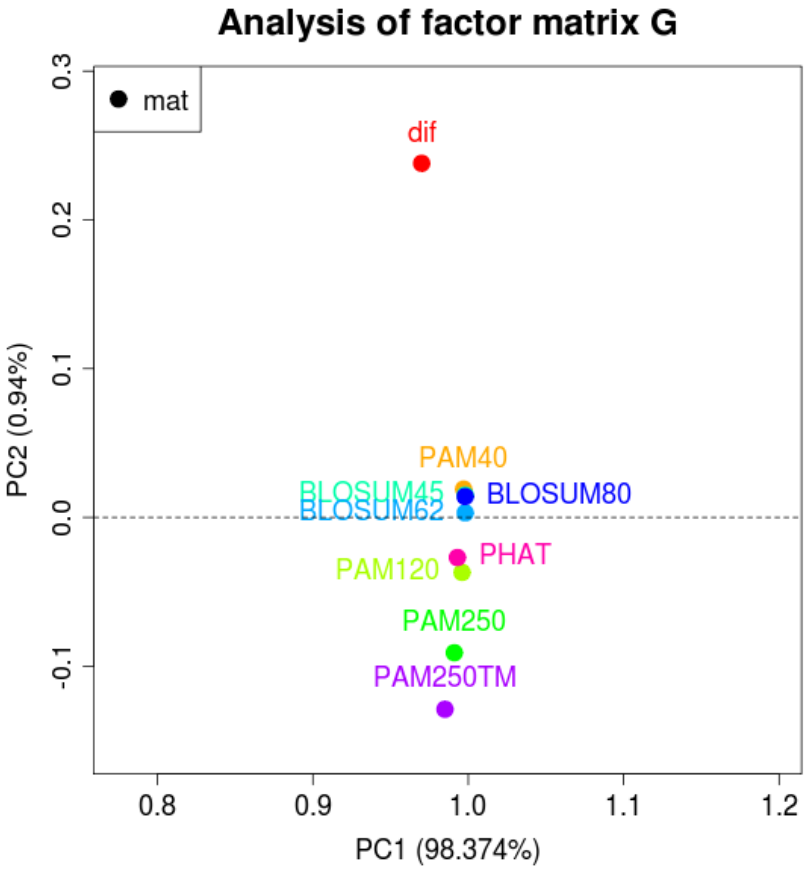

Figure 22 : Comparaison par DISTATIS des matrices de distances

Les points représentent les différentes matrices de distances. La matrice de différences ("dif" sur le graphique) est en rouge. Les matrice de dissimilarités PAM40, PAM120, PAM250, BLOSUM45, BLOSUM62, BLOSUM80, PAM250TM et PHAT sont en jaune, vert clair, vert foncé, turquoise, bleu clair, bleu foncé, violet et magenta, respectivement. L'axe des abscisses et des ordonnées correspondent à la première (PC1) et deuxième composante principale (PC2), respectivement.

Le contenu de la variable mmds1 est affiché et contient trois éléments :

```
>mmds1
$eigen
 [1] 5.633 3.392 2.273 2.136 2.014 1.815 1.678 1.564 1.428 1.317

$eigen.perc
 [1] 6.324 3.809 2.552 2.398 2.261 2.037 1.884 1.756 1.603 1.479

$active.coord
             PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
SA.AD1    -0.194 -0.002 -0.005  0.026 -0.016  0.070 -0.027  0.018
SA.AD2    -0.224 -0.022  0.002 -0.003 -0.068  0.068 -0.015  0.030
SA.AD3    -0.194 -0.008  0.023 -0.001 -0.043  0.021  0.025  0.016
SA.AD4    -0.204  0.002  0.003  0.015 -0.038  0.028 -0.050 -0.009
SA.AMIN1  -0.281 -0.116 -0.078  0.015  0.032 -0.046 -0.020  0.007
...
```

Quatre autres MDS métriques avec les mêmes éléments actifs mais avec projections d'éléments supplémentaires issus des quatre autres génomes (*N. vectensis*, *C. elegans*, *C. intestinalis* et *D. rerio*) sont réalisées et les résultats sont stockés dans les variables `mmds2`, `mmds3`, `mmds4` et `mmds5` respectivement :

```
>mmds2 <- mmds(active = gpcr$dif$sapiens.sapiens,
sup = gpcr$dif$vectensis.sapiens)
>mmds3 <- mmds(active = gpcr$dif$sapiens.sapiens,
sup = gpcr$dif$elegans.sapiens)
...
```

Les représentations graphiques sont affichées dans un seul et même cadre avec la fonction `layout` (**Figure 23**) :

```
>layout(matrix(1:6, 3, 2))
```

Ensuite, un graphique d'éboulis des pourcentages de variance en fonction des composantes

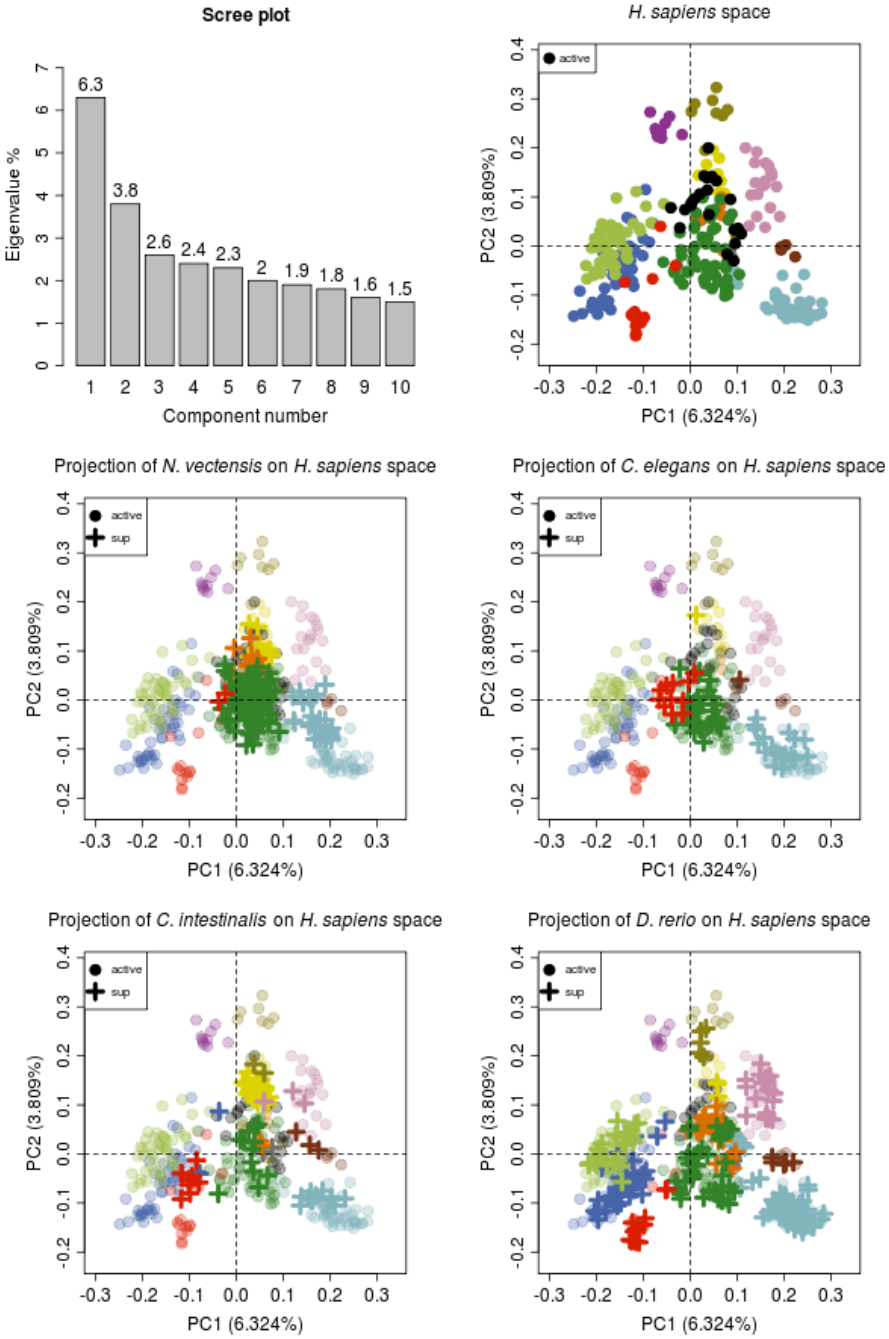

Figure 23 : Graphique d'éboulis et projections par MDS métrique

Pour le graphique d'éboulis, les valeurs sur les barres sont les valeurs précises des pourcentages de valeurs propres. Pour les cinq graphiques suivant, les cercles représentent les RCPG d'*H. sapiens* alors que les croix indiquent la position des RCPG des génomes projetés en tant qu'éléments supplémentaires. Le code couleur correspond aux différentes sous-familles et la concordance est expliquée dans l'article (voir chapitre 6.1).

principales est affiché avec la fonction graphique `plot.scree`, avec légendes :

```
>plot.scree(mmds1$eigen.perc, lab = TRUE, new.plot = FALSE)
```

Pour chaque MDS métrique, un graphique des éléments actifs et des projections successives des éléments supplémentaires est affiché avec la fonction graphique `plot.mmds.point` :

```
>plot.mmds.point(mmds3,
title = expression(paste("Projection of ", italic("C. elegans"),
" on ", italic("H. sapiens"), " space")),
active.col = as.vector(gpcr$col$sapiens$HEX), active.alpha = 0.3,
sup.col = as.vector(gpcr$col$elegans$HEX), active.cex = 3,
new.plot = FALSE)
...
```

Le graphique d'éboulis montre que le point d'inflexion se trouve au niveau de la troisième composante principale. La convention veut qu'il faille conserver uniquement les composantes se situant avant le coude, à savoir les deux premières composantes. Le graphique en haut à droite représente l'espace actif seule. Les quatre suivants concernent les projections d'éléments supplémentaires.

D.4.6 MDS métrique en 3 dimensions

Les codes RGB de chaque séquence pour les cinq génomes sont conservés dans différentes variables, sous forme de listes, à partir du jeu de données `gpcr` :

```
>sapiens.col <- lapply(strsplit(as.vector(gpcr$col$sapiens$RGB),
" "), function(i) {as.integer(i)})
...
```

Le contenu de la variable `sapiens.col` est affiché :

```
>sapiens.col
[[1]]
[1] 120  50  20
[[2]]
```

```
[1] 120  50  20 ...
```

La couleur marron (rouge = 120, vert = 50, bleu = 20) est attribuée aux quatre premiers récepteurs et la couleur cyan (rouge = 130, vert = 180, bleu = 190) est attribuée au cinquième et ainsi de suite pour les récepteurs restants. Ces variables servent à générer des fichiers PDB et PML avec la fonction `write.mmds.pymol` et permettent de visualiser l'espace actif et les projections de la Figure 23 mais en trois dimensions et de manière interactive avec Pymol :

```
>write.mmds.pymol(mmds1, active.col = sapiens.col, file.pdb = "R1.pdb", file.pml = "R1.pml")
>write.mmds.pymol(mmds2, active.col = sapiens.col, active.alpha = 0.3, sup.col = vectensis.col, file.pdb = "R2.pdb", file.pml = "R2.pml")
...
```

Pour chaque fichier PDB est attribué un fichier PML correspondant. Par exemple, les fichiers R4.pdb et R4.pml permettent de visualiser la projection des séquences de *C. intestinalis* sur l'espace actif d'*H. sapiens* (**Figure 24**). *bio2mds* a vocation à être amélioré et complété pour offrir davantage de fonctionnalités sur la même thématique.

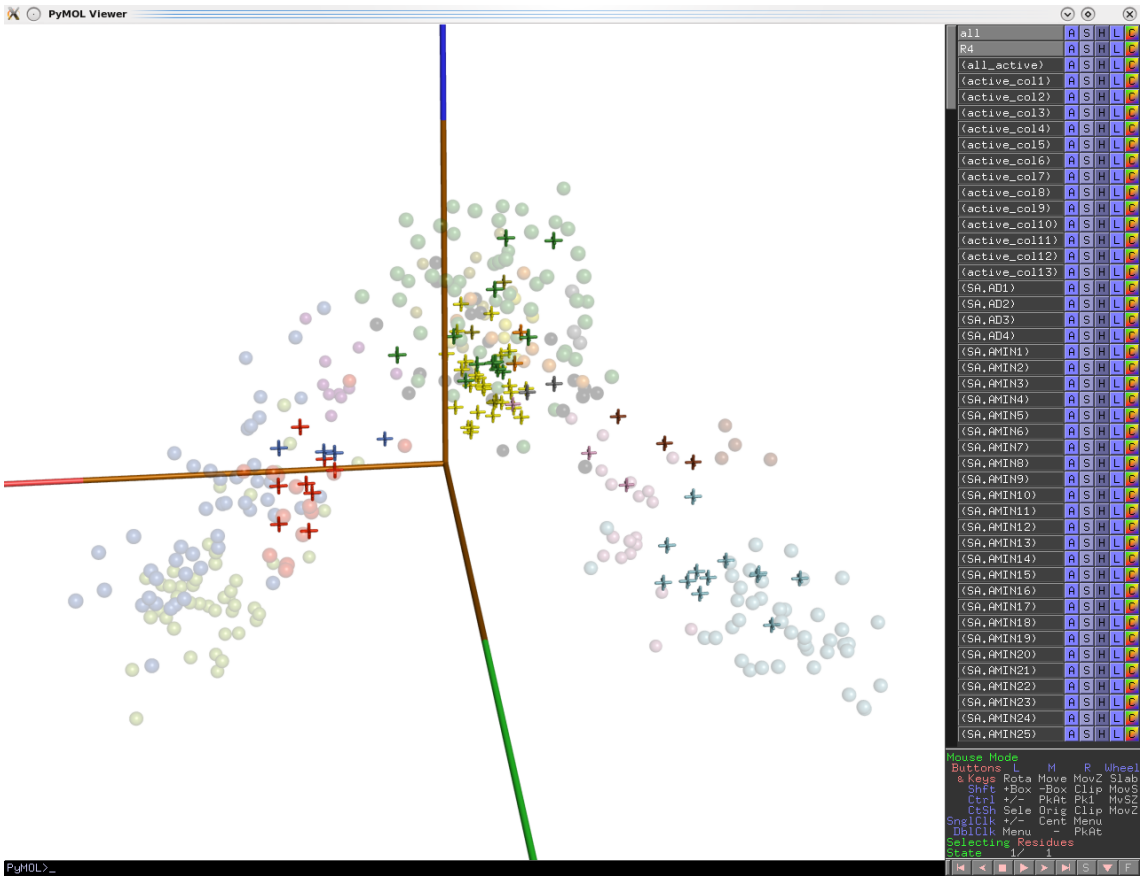

Figure 24 : Visualisation sous Pymol de la projection de séquences

Les axes rouge, bleu et vert représentent les axes de coordonnées x, y et z, respectivement. Les sphères transparentes donnent la position des RCPG d'*H. sapiens* alors que les croix indiquent celles des RCPG de *C. intestinalis*. La colonne de droite montre les différentes possibilités de sélection des objets. Par exemple, active_col1 correspond à la première sous-famille des RCPG d'*H. sapiens* et SA.AD1 correspond au premier RCPG de la sous-famille AD d'*H. sapiens*. Ces sélections permettent de repérer les objets correspondant dans l'espace tri-dimensionnelle.

Résumé

Les récepteurs couplés aux protéines G de classe A (RCPG) constituent la plus grande famille de récepteurs transmembranaires du génome humain et sont impliqués dans la régulation de nombreux mécanismes physiologiques. Comprendre les mécanismes évolutifs qui ont conduit à la diversité de cette famille de récepteurs pourrait permettre une meilleure connaissance des relations séquence-structure-fonction des différentes sous-familles. Pour obtenir des informations sur l'évolution des RCPG, nous avons exploré leur espace de séquences par multidimensional scaling métrique (MDS). Nous avons appliqué une nouvelle technique MDS qui projette des séquences supplémentaires sur un espace de référence et permet ainsi la comparaison des séquences de différentes espèces. Les résultats montrent que les récepteurs se répartissent en quatre groupes et suggèrent que les récepteurs actuels ont évolué à partir d'ancêtres des récepteurs de peptides suivant trois directions évolutives principales. Les prolines des hélices transmembranaires 2 et/ou 5 sont impliquées dans deux de ces directions. Pour comprendre le mécanisme fin ayant abouti à la formation des différentes sous-familles, nous avons analysé les covariations des résidus à différents niveaux hiérarchiques (classe/groupe/sous-famille). Nous avons testé différentes méthodes pour analyser les mutations corrélées afin de sélectionner une méthode robuste pour les différents jeux de séquences. L'application de cette méthode met en évidence des résidus spécifiques qui sont cruciaux pour l'évolution de sous-familles particulières.

Mots-clés : bioinformatique, RCPG, MDS, évolution, proline, mutations corrélées

Abstract

Class A G-protein-coupled receptors (GPCRs) constitute the largest family of transmembrane receptors in the human genome and are involved in the regulation of many physiological functions. Understanding the mechanisms that drove the evolution of this receptor family should allow a better knowledge of sequence-structure-function relationships of the different sub-families. To gain evolutionary information on GPCRs, we explored their sequence space by metric multidimensional scaling (MDS). We applied a new MDS technique which projects supplementary sequences onto a sequence space of reference and allows comparison of sequences from different species. Results show that receptors cluster in four groups and suggest that modern receptors evolved from ancestors of the peptide receptors along three main evolutionary pathways. Proline residues of transmembrane helices 2 and/or 5 are involved in two of these pathways. To further detail the mechanisms that led to the different sub-families, we analyzed covariations of residues at different hierarchical levels (class/group/sub-family). We tested different methods of covariation analysis in order to select a method robust at the different hierarchical levels. This method highlights sequence determinants that are crucial for the evolution of specific sub-families.

Keywords : bioinformatics, GPCR, MDS, evolution, proline, correlated mutations