

**ETUDE STRUCTURALE DES CASSURES
D'HELICES ET SON APPLICATION A LA
MODELISATION DES RECEPTEURS
COUPLES AUX PROTEINES G (RCPG)**

THESE DE DOCTORAT

Spécialité : Bioinformatique

ECOLE DOCTORALE D'ANGERS

Présentée et soutenue publiquement

Le : 19/12/2007

à : Angers

par

Julie Devillé

Devant le jury ci-dessous :

Dr J.F. Gibrat (rapporteur), DR, INRA, Jouy-en-Josas

Dr I. Milazzo (rapporteur), MCU, UMR CNRS 6014, Rouen

Pr C. Delamarche (examinateur), PU, UMR CNRS 6026, Rennes

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
Rue Haute de Reculee
49045 ANGERS CEDEX 01

TABLE DES MATIERES

INTRODUCTION GENERALE.....	7
1 LES RECEPTEURS COUPLES AUX PROTEINES G.....	11
1.1 La famille des récepteurs couplés aux protéines G.....	12
1.2 Classification des RCPG.....	13
1.2.1 Diversité des récepteurs couplés aux protéines G :.....	13
1.2.2 Classification historique par ligands de GPCRdb.....	13
1.2.3 Le système de classification GRAFS.....	15
1.2.3.1 Les cinq familles de la classification GRAFS.....	17
1.2.3.2 La famille de la rhodopsine.....	18
1.2.4 Classification par Analyse par Composante Principale (ACP).....	21
1.3 Nos récepteurs d'intérêt.....	21
1.3.1 Les récepteurs de l'angiotensine : AT1 et AT2.....	23
1.3.1.1 Fonctionnement.....	23
1.3.1.2 Les récepteurs AT1 et AT2.....	23
1.3.2 Les autres récepteurs du groupe 1.....	25
1.3.2.1 Les autres récepteurs de peptides vasoactifs.....	25
1.3.2.2 Les récepteurs des chimiokines.....	26
1.3.2.3 Les récepteurs opioïdes/somatostatines.....	26
1.3.2.4 Les récepteurs MCH.....	26
1.3.2.5 Les récepteurs purinergiques.....	27
1.4 Structure des RCPG de classe A.....	27
1.4.1 La structure de la rhodopsine et les RCPG.....	27
1.4.2 Analyse des segments transmembranaires de la famille de la rhodopsine.....	28
1.4.3 Les déformations des hélices de la rhodopsine.....	30
1.4.3.1 Hélice transmembranaire 1.....	31
1.4.3.2 Hélice transmembranaire 2.....	32

1.4.3.3	Hélice transmembranaire 3.....	32
1.4.3.4	Hélice transmembranaire 4.....	34
1.4.3.5	Hélice transmembranaire 5.....	34
1.4.3.6	Hélice transmembranaire 6.....	34
1.4.3.7	Hélice transmembranaire 7.....	35
1.4.4	Conclusion.....	35
2	OUTILS MIS EN PLACE.....	36
2.1	Outils d'analyse de la PDB	37
2.1.1	Outils d'analyse structurale et stéréochimique.....	37
2.1.1.1	Calcul des angles dièdres	37
2.1.1.2	Calcul des vecteurs normaux parallèles à l'axe d'un segment d'hélice	39
2.1.1.3	Calcul des angles d'inclinaison et de giration entre 2 hélices successives	39
2.1.2	Outils d'analyse des séquences	41
2.1.2.1	Formule pour le calcul de la propension des acides aminés.....	41
2.1.2.2	Formule Z-score pour analyser les propensions.....	41
2.1.2.3	Prédiction des conformations	43
2.1.3	La base de données des motifs de hélices cassées : HXH.....	43
2.1.4	Recherche de motifs : SPASM.....	45
2.2	Création de la base de données relationnelle des RCPG.....	46
2.2.1	Modèle conceptuel de données (MCD).....	47
2.3	Conclusion.....	49
3	RESULTATS	50
3.1	L'analyse du motif hélice-X-hélice au sein des protéines solubles.....	51
3.1.1	Article I	51
3.1.2	Conclusion.....	74
3.2	L'évolution de l'hélice transmembranaire 2 chez les RCPG	76
3.2.1	Article II	76
3.2.2	Conclusion.....	95
4	CONCLUSIONS ET PERSPECTIVES.....	96
	LISTE DES ABREVIATIONS	100

A Marie et Charlotte.

REMERCIEMENTS

Au terme de ce travail, je tiens à remercier Mme le Dr. Marie Chabbert qui m'a accueilli dans l'équipe de Bioinformatique. Elle m'a donné le goût de la recherche et m'a toujours encouragée. Ses conseils, son encadrement et les nombreuses discussions que nous avons eues, scientifiques ou non, m'ont toujours été bénéfiques. Que vous trouviez dans ce manuscrit le témoignage de ma profonde et sincère reconnaissance.

Je remercie M le Dr. Daniel Henrion pour m'avoir accueilli dans son unité de recherche.

Je remercie M. le Dr Gibrat et Mme le Dr Milazzo qui ont accepté d'être rapporteurs de mon travail. Je remercie M. le Pr. Delamarche et M. le Dr. Legros qui ont bien voulu participer au jury.

Je remercie Mme Chartier, administrateur délégué régional INSERM, pour m'avoir permis de réussir mon transfert de laboratoire.

Je remercie les étudiants de l'équipe de bioinformatique avec qui j'ai travaillé pour leur disponibilité et leur gentillesse, ainsi que l'ensemble de l'UMR CNRS 6214/ INSERM 771 pour les séances de bavardages toujours positives, que ce soit humainement ou scientifiquement. Un merci particulier à Mme le Dr Fromy pour son amitié et les longues discussions que nous avons pu partager.

Je remercie David Perret, Fabien Barbier et François Rousseau, mes anciens collègues de l'Unité INSERM 564 avec qui je garde toujours de bons contacts. J'espère ne jamais rompre nos liens.

Je remercie Francky, toujours enjoué à accepter mes invitations à dîner, Farfouille le roi de l'« abusation », Mattmatt, le plus gentil des amis, Gazman et son décalage intemporel, la « meuf » parce qu'elle le vaut « bieng », Bourguignon et son amour des escaliers et toutes les personnes qui se retrouvent à la cafétéria de Montclair à midi et qui me permette de décompresser.

Je remercie Antoine Fouillet pour être le meilleur des meilleurs amis, mon colocataire adoré. Tu as toujours été là quand j'en avais besoin et je sais que tu le seras encore et toujours. Ton amitié m'est très chère et je te souhaite tout le bonheur du monde.

Je remercie Arnaud Favre pour mettre du soleil dans mon cœur et me redonner de la joie quand plus rien ne va.

Je voudrais remercier mes parents qui m'ont soutenue tout au long de mon parcours. Ils sont deux rocs au milieu de la tempête. Ils ont toujours été là pour ramasser les morceaux et les réassembler et je ne serai pas arrivée jusque là sans leur précieuse aide. Ce manuscrit est le résultat de votre indéfectible soutien.

INTRODUCTION GENERALE

Les récepteurs couplés aux protéines G (RCPG) constituent une famille protéique majeure des membranes biologiques. C'est la plus grande famille de récepteurs membranaires dans le génome humain avec plus de 1000 protéines répertoriées [1, 2]. Ces récepteurs sont impliqués dans la transduction du signal d'une très grande variété de stimuli endogènes ou exogènes, y compris des photons, des ions, des molécules organiques odorantes, des amines, des lipides, des nucléotides, des peptides et des protéines [3]. Ils sont impliqués dans de nombreuses fonctions physiologiques naturelles aussi diverses et variées que la phototransduction, l'olfaction, l'immunité, la digestion, le système cardio-vasculaire. Néanmoins, on a également mis en évidence leur rôle dans de nombreux processus pathologiques telles des maladies inflammatoires, des cancers, des troubles neurodégénératifs et cardio-vasculaires ou encore des infections virales. Certaines de ces pathologies peuvent être liées à une mutation ou à une surexpression du récepteur, à une dérégulation de l'activation de ces récepteurs ou à leur utilisation comme récepteur ou co-récepteur de virus. Leur importance physiopathologique explique que ces récepteurs soient la cible d'un grand nombre de médicaments. Il a été estimé que 40% des médicaments modernes ciblent ces récepteurs et plusieurs ligands de ces RCPG sont compris dans le haut des ventes des produits pharmaceutiques [4]. Des exemples notables sont le Zyprexa® de chez Eli Lilly® utilisé pour traiter la schizophrénie, le Clarinex® de chez Schering-Plough® qui est un anti-histaminique, le Zantac® de chez GlaxoSmithKline® utilisé dans la prévention des ulcères ou encore le Zelnorm® de chez Novartis® utilisé dans le traitement de la constipation [5].

De nombreux peptides vasoactifs, comme l'angiotensine II, interagissent par l'intermédiaire de récepteurs spécifiques, qui appartiennent à la famille des récepteurs couplés aux protéines G [6]. Ces peptides jouent un rôle physiologique important en régulant la tonicité, la réactivité et la structure vasculaire. En conditions pathologiques, les altérations dans la régulation des peptides vasoactifs ont comme conséquence le dysfonctionnement endothélial, le remodelage vasculaire et l'inflammation vasculaire, qui sont des processus fondamentaux importants dans la maladie cardiovasculaire. Parmi les nombreux peptides vasoactifs impliqués dans la biologie (patho)vasculaire, l'angiotensine II, l'endotheline et les peptides natriurétiques semblent être particulièrement importants en raison de leurs nombreuses actions pléiotropiques et parce qu'ils ont été identifiés en tant que cibles thérapeutiques potentielles dans les maladies cardiovasculaires.

L'intérêt du laboratoire pour les peptides vasoactifs et en particulier l'angiotensine II, nous a amené à étudier les récepteurs couplés aux protéines G de la famille de la rhodopsine (classe A) auxquels appartiennent les récepteurs AT1 et AT2 [7], en vue de leur modélisation.

Le récepteur de l'angiotensine II de type 1 (AT1) induit une vasoconstriction via l'activation de la phospholipase C et l'inhibition de l'adénylcyclase alors que le récepteur de l'angiotensine II de type 2 (AT2) induit une vasodilatation via l'ouverture de canaux potassiques ou l'activation de la guanylate cyclase [8].

Bien que les RCPG représentent une cible pharmaceutique importante, les seules structures à haute résolution actuellement disponibles sont celles de la rhodopsine bovine [9] et du récepteur adrénérgique $\beta 2$ humain, qui vient d'être résolue [10, 11]. La raison de ce paradoxe tient essentiellement au fait que ces protéines sont naturellement très faiblement exprimées dans leur tissu d'origine et qu'il existe des barrières expérimentales majeures à chaque étape de la procédure qui, de l'expression à la solubilisation et à la purification, rendent particulièrement difficile la réalisation d'études structurales.

Jusqu'en Octobre 2007, la rhodopsine bovine était le seul récepteur de la famille des RCPG dont la structure cristallographique était résolue. Cette structure est employée couramment comme modèle en modélisation par homologie des récepteurs couplés aux protéines G de classe A. La validité de cette structure modèle est ainsi une question cruciale pour la conception de médicaments. La structure de la rhodopsine est constituée de sept hélices transmembranaires (TMH1 à 7), structure représentative de la famille des RCPG. Comme généralement observé au sein des protéines, la plupart de ces hélices ne sont pas droites, mais cassées ou incurvées. Plusieurs cassures sont liées aux prolines fortement conservées situées dans les hélices transmembranaires 5-7. En ce qui concerne l'hélice transmembranaire 2 (TMH2) de la rhodopsine, la cassure se fait au niveau d'un motif GG correspondant à un renflement π . Ce motif GG est situé aux positions 2.56-2.57, selon la nomenclature de Ballesteros [12] (c'est-à-dire 6 et 7 résidus en aval du résidu ASP fortement conservé de TMH2) et n'est pas conservé parmi les RCPG. Une proline est fréquemment observée dans cette hélice aux positions 2.58, 2.59 et 2.60. Cependant, sa présence n'est pas obligatoire et environ 20% des classes des RCPG n'ont pas de proline à ces positions. Nos récepteurs d'intérêt AT1 et AT2 possèdent quant à eux une proline à la position 2.58. La question se pose alors de savoir comment modéliser cette cassure d'hélice avec le remplacement d'un motif GG par une proline pouvant se situer à différentes positions.

Le but de mon travail de thèse a été de comprendre dans un premier temps quelles peuvent être les structures possibles pour les motifs de cassures d'hélice. Pour cela, nous avons créé une base de données d'hélices cassées qui nous a permis d'effectuer une étude exhaustive de ces motifs. Cette connaissance des hélices cassées nous a permis dans un deuxième temps

de résoudre la modélisation de TMH2 en fonction de la position de la proline en 2.58, 2.59 ou 2.60.

Nous présenterons tout d'abord les récepteurs couplés aux protéines G dans la première partie de cette thèse. Puis nous présenterons les outils développés dans la deuxième partie. Enfin, nous présenterons les résultats de notre travail et nous envisagerons les perspectives de recherche pour le laboratoire.

1 LES RECEPTEURS COUPLES AUX PROTEINES G

Nous présenterons les récepteurs couplés aux protéines G dans la première partie de cette introduction, tout d'abord de manière générale, puis en détaillant nos récepteurs d'intérêt. Ensuite, nous présenterons les données connues sur la structure des récepteurs couplés aux protéines G de classe A.

1.1 La famille des récepteurs couplés aux protéines G

Les RCPG sont certainement parmi les plus anciens transducteurs de signaux. En effet, ils sont présents dans les plantes, les levures, les champignons, ainsi que les protozoaires et les métazoaires diploblastiques. Ils interviennent dans tous les grands systèmes de communication intercellulaire. Les ligands des RCPG sont d'une très grande diversité chimique. Ils incluent les photons (rhodopsine des bâtonnets et opsines rouge, verte et bleue des cônes), des ions (Ca^{2+}), des stimuli sensoriels (molécules olfactives, gustatives et phéromones), des petites molécules endogènes (acides aminés, amines, nucléotides, lipides et peptides endogènes), des composés exogènes (cannabinoïdes, peptides d'amphibiens : ranatensine ou bombésine), des composés impliqués dans les réactions du système immunitaire (les chimiokines, les anaphylatoxines C3a et C5a du complément, les peptides N-formylés chimiotactiques) et des protéines (hormones glycoprotéiques, protéases) [13].

Le séquençage du génome humain a permis d'identifier la famille des RCPG comme une des plus grandes classes de protéines du génome des mammifères. Le nombre total de RCPG dans le génome humain est estimé à environ 1000 dont plus de 400 sont des récepteurs olfactifs [14]. Chez l'homme, on a constaté que plus de 1% du génome code pour plus de 1000 protéines ayant une structure à 7 domaines transmembranaires. Cette famille correspond à 1% des gènes de la *Drosophila* et plus de 5% de tous les gènes de *Caenorhabditis elegans* (dont le génome est maintenant entièrement séquencé).

Les RCPG sont des protéines à sept domaines transmembranaires (1 à 7) possédant une structure en hélice α , reliés par trois boucles externes (e1, e2, e3) et trois boucles internes (i1, i2, i3). Ces sept segments transmembranaires constituent le "corps central" de ces récepteurs et le changement de conformation de cette région, qui fait suite à la liaison du ligand, est probablement responsable de l'activation des RCPG [15]. L'activation d'un RCPG par son ligand entraîne la transduction d'un message à l'intérieur de la cellule par un mécanisme faisant intervenir les protéines G hétérotrimériques ($G\alpha$, $G\beta\gamma$) liant, suivant leur état d'activation, les nucléotides guanyliques GDP ou GTP [16]. La liaison du ligand à son récepteur entraîne un remaniement de la structure du récepteur. Le changement de

conformation du complexe agoniste-récepteur permet la liaison de la protéine G hétérotrimérique-GDP au récepteur couplé aux protéines G. Cette liaison entraîne l'activation de l'échange du GDP (état inactif) en GTP (état actif) au niveau du site nucléotidique de la sous-unité $G\alpha$. La liaison du GTP au site actif de la sous-unité $G\alpha$ induit la dissociation de $G\alpha$ et de $G\beta\gamma$. Les différentes sous-unités sont alors capables de moduler des effecteurs intracellulaires aussi divers que des enzymes, des canaux ou encore des échangeurs ioniques. L'hydrolyse du GTP en GDP au niveau du site actif de la sous-unité $G\alpha$ par une activité GTPasique intrinsèque, stimulée par des protéines RGS (Regulator of G-protein Signaling) permet le découplage de la sous-unité $G\alpha$ des effecteurs. L'hétérotrimère $G\alpha$ - $G\beta\gamma$ se reforme à nouveau et permet le retour à l'état inactif initial (figure 1) [17].

1.2 Classification des RCPG

1.2.1 Diversité des récepteurs couplés aux protéines G :

Les RCPG représentent la plus grande famille des récepteurs transmembranaires impliqués dans la transmission du signal. Cette famille de protéines est la plus vaste mais aussi la plus ubiquitaire et la plus diversifiée trouvée dans la nature, toutes classes de protéines confondues, ce qui conforte l'idée de leur importance physiologique. Une comparaison directe entre les RCPG humains et murins révèle un degré d'homologie très élevé : la préservation de cette famille de récepteurs au cours de l'évolution est aussi un argument allant dans le sens de leur importance pour la cellule [18]. Du fait du grand nombre de récepteurs couplés aux protéines G, une classification de ces récepteurs est nécessaire pour leur étude.

1.2.2 Classification historique par ligands de GPCRdb

Les RCPG ont commencé à être regroupés systématiquement dans des bases de données dans la fin des années 1980 et le début des années 1990. Les deux principales bases sont Uniprot (<http://www.expasy.uniprot.org/>) et la GPCRdb (<http://www.gpcr.org/>). La base Uniprot [19] regroupe la base de données Swiss-Prot (environ 200 000 entrées) et TrEMBL (environ 2 100 000 entrées). Swiss-Prot est une base de données de séquences protéiques de tous les organismes qui vise à assurer un niveau élevé d'annotations (telles que la description

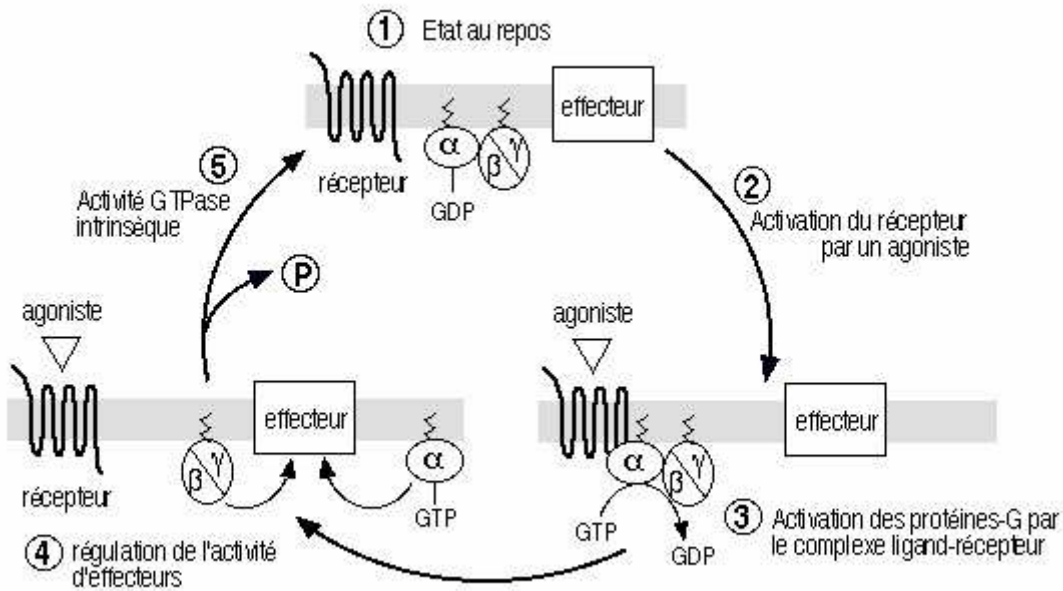


figure 1 : Mécanisme d'activation des RCPG

Un RCPG au repos (1) est activé par la liaison d'un agoniste spécifique (2). Le changement de conformation du complexe agoniste-récepteur, induit par cette liaison, permet l'activation de l'échange du GDP par du GTP et donc l'activation de la protéine-G hétérotrimérique (sous-unités $G\alpha$ et $G\beta/\gamma$) intracellulaires (3) qui vont aller réguler l'activité de divers effecteurs (4) membranaires ou cytosoliques. Le déclenchement de l'activité phosphatase, intrinsèque à la sous-unité $G\alpha$ entraîne la réassociation des sous-unités $G\alpha$ et $G\beta/\gamma$ (5) et le retour à l'état initial (1).

de la fonction d'une protéine, ses domaines structuraux, ses modifications post-transcriptionnelles, ses variants, etc), un niveau minimum de redondance et un haut niveau d'intégration avec d'autres bases de données. TrEMBL est une base de données complémentaire à Swiss-Prot comprenant des entrées supplémentaires (pas encore annotées) résultant de la traduction automatique de l'EMBL.

La GPCRdb (G-Protein-Coupled Receptors Database) [20], mise en place en 1994, est spécifique des récepteurs à 7 hélices transmembranaires, et en particulier les RCPG. Elle regroupe de nombreuses informations, telles que les séquences, mais aussi des modèles 3D, une classification basée sur des similarités de séquences et aussi des liens vers d'autres bases de données.

De nombreux systèmes de classification des RCPG ont été proposés. Ils divisent les récepteurs en fonction de leur ligand et/ou de leur séquence. Dans un des systèmes les plus utilisés, les RCPG ont été classés en 6 classes notées de A à F, sur la base des identités de séquences. Ce système est nommé système de Kolakowski [21]. Ces classes sont: (A) les récepteurs ressemblant à la rhodopsine, qui ont un petit domaine N-terminal, (B) les récepteurs ressemblant à celui de la Sécrétine, qui ont un grand domaine N-terminal structuré de façon complexe, (C) les récepteurs ressemblant à celui du Glutamate, qui ont un grand domaine N-terminal sous forme de dimère, (D) des récepteurs de phéromones fongiques (non humains), (E) des récepteurs de AMPc (non humains), et enfin (F) des récepteurs dénommés en anglais Frizzled et Smoothened, qui possèdent de très longs domaines N-terminaux.

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

Chaque famille est partagée en sous-familles ou sous-classes, d'après la nature chimique des ligands et des spécificités pour les agonistes et antagonistes connus. Le [tableau 1](#) donne la classification de la classe A de la GPCRdb en familles, basée sur les ligands connus.

1.2.3 Le système de classification GRAFS

Deux travaux proposent des classifications phylogénétiques des RCPG. Le premier est celui de Joost et Methner en 2002 [22] qui présente une classification de 277 RCPG humains non olfactifs. La méthode utilisée est une méthode de distance phylogénétique entre

Class A Rhodopsin like
Amine
Peptide
Hormone protein
(Rhod)opsin
Olfactory
Prostanoid
Nucleotide-like
endogènes
Platelet activating factor
Gonadotropin-releasing hormone
Thyrotropin-releasing hormone & Secretagogue
Melatonin
Viral
Lysosphingolipid & LPA (EDG)
Leukotriene B4 receptor
Class A Orphan/other

Tableau 1 : Classe A des récepteurs ressemblant à la Rhodopsine

Les différentes sous-famille de la famille de la Rhodopsine.

séquences appelée « neighbour-joining » (joindre les voisins) [23]. Le jeu de données utilisé comportait 196 récepteurs non orphelins, et 81 récepteurs orphelins. La seconde classification a été établie par Fredriksson et collaborateurs en 2003 [24]. Plus complète dans le jeu de données car plus récente, de méthode semblable, elle présente une révision de la classification des RCPG répartissant 342 récepteurs humains non olfactifs selon 5 familles (figure 2). Elle est baptisée GRAFS, acronyme des noms des 5 familles (Glutamate, Rhodopsin, Adhesion, Frizzled/taste2, Secretin).

1.2.3.1 Les cinq familles de la classification GRAFS

La première famille de la classification GRAFS de Fredriksson comprend les récepteurs du Glutamate. Elle regroupe 8 récepteurs métabotropiques au Glutamate, 2 récepteurs du GABA et leurs variants d'épissage, le récepteur du calcium (CASR), et 5 des récepteurs du goût (TAS1). CASR et TAS1 sont regroupés avec GABA, formant la branche basale de la famille, suggérant que les récepteurs du glutamate sont apparus plus tardivement lors de l'évolution.

La seconde famille du système de classification GRAFS est la famille de la rhodopsine, comprenant le plus grand nombre de membres. Cette famille se divise en 4 groupes principaux : α , β , χ et δ . Cette famille comprend plusieurs caractéristiques tels que le motif NSxxNPxxY dans l'hélice transmembranaire 7, le motif D(E)-R-Y(F) à la frontière entre l'hélice transmembranaire 3 et la boucle intracellulaire 2. Seul un petit nombre de récepteurs ne comprennent pas ces motifs, mais ils ont d'autres "empreintes digitales" qui les lient phylogénétiquement à la famille de la rhodopsine.

La troisième famille du système de classification GRAFS comprend les récepteurs d'adhésion. Cette famille a été dans un premier temps comparée aux récepteurs de la sécrétine [25]. Ces derniers forment désormais leur propre famille.

Dans la classification de Fredriksson, les récepteurs Frizzled/TAS2 forment une famille divisée en 2 clusters : un pour les récepteurs Frizzled et l'autre pour les récepteurs TAS2.

La dernière famille de la classification GRAFS est la famille de la sécrétine. Elle comprend les récepteurs de la sécrétine, de la calcitonine, du glucagon, du CRH (corticotropin-releasing hormone), de la parathyroïde, etc.

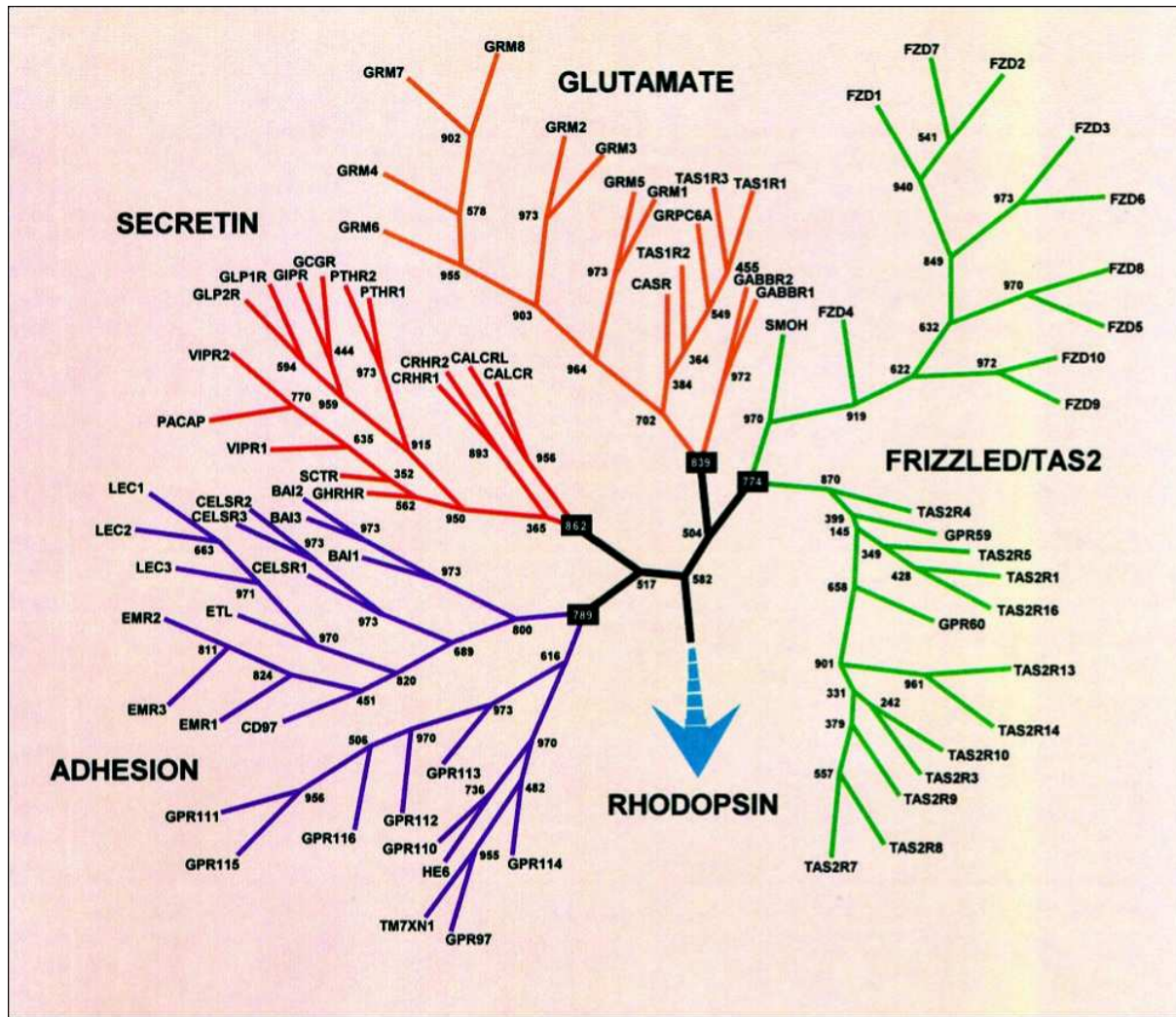


Figure 2 : Arbre phylogénétique des GPCR (TMI-TMVI) dans le génome humain [24]

Les RCPG sont classés selon le système GRAFS qui regroupent les récepteurs en 5 familles : La famille des récepteurs du glutamate, la famille des récepteurs de la sécrétine, la famille des récepteurs d'adhésion, la famille des récepteurs Frizzled/Tas2 et enfin, la famille de la rhodopsine.

1.2.3.2 La famille de la rhodopsine

Le groupe α se divise en 5 branches principales : les récepteurs des prostaglandines, amines, opsines, mélatonines et les récepteurs MECA (figure 3). La branche des récepteurs des prostaglandines comprend les 8 récepteurs des prostaglandines et 7 récepteurs orphelins. Les récepteurs des prostaglandines ont entre 19 et 41% d'identité de séquence et partagent les mêmes motifs au niveau de TMH7 (IXDPW) et de TMH1 (LXXTDXXG). La branche des récepteurs des amines comprend les récepteurs de la sérotonine (HTR), les récepteurs à la dopamine (DRD), les récepteurs muscariniques (CHRM), les récepteurs des histamines (HRH), les récepteurs adrénergiques (ADR), les récepteurs TAR ou TAAR (trace amine receptor ou trace amine associated receptor), et quelques récepteurs orphelins. Tous les ligands connus de ce groupe de récepteurs ont une structure similaire aux amines avec un seul noyau aromatique. Le degré de conservation des séquences varie selon les types de récepteurs. La branche des récepteurs des opsines comprend le récepteur visuel des bâtonnets (la rhodopsine : RHO), les récepteurs des 3 types de cônes (OPN1SW, OPN1LW, OPN1MW), la péropsine (RRH), l'encéphalopsine (OPN3), la mélanopsine (OPN4) et les récepteurs de la rétine couplés aux protéines G (RGR). Les opsines sont les seuls RCPG connus pour répondre à la lumière et aucun de ces récepteurs ne fixe de ligand physique. La branche des récepteurs des mélatonines comprend les récepteurs des mélatonines (MTNR1A, MTNR1B) et un récepteur orphelin (GPR50). La branche des récepteurs MECA se compose des récepteurs de la mélanocortine (MCR), des RCPG de la différenciation endothéliale (EDGR), des récepteurs cannabinoïdes (CNR) et des récepteurs se liant à l'adénosine (ADORA). Trois récepteurs orphelins appartiennent aussi à ce groupe (GPR - 3, -6, et -12). Il est intéressant de noter que les récepteurs de ce groupe lient des ligands structurellement différents : des hormones stimulant les mélanocytes (peptide, MCR), des acides lysophosphatidiques (lipides, EDGR), l'anandamide (arachidonylethanolamide, CNR) et l'adénosine.

Le groupe β de la famille de la rhodopsine ne contient pas de branches principales et comprend 36 récepteurs. Tous les ligands connus de ces récepteurs sont des peptides. Ce groupe inclut les récepteurs à l'hypocrétine (HCRTR), les récepteurs des neuropeptides FF (NPFF), les récepteurs de la tachykinine (TACR), les récepteurs de la cholécystokinine (CCK), les récepteurs du neuropeptide Y (NPYR), les récepteurs des endothélines (EDNR et ETBRLP1/2), le récepteur *gastrin-releasing peptide* (GRPR), le récepteur de la neuromédine B (NMBR), le récepteur de la bombésine (BRS3), les récepteurs des neurotensines (NTSR), le récepteur de l'hormone libérant la thyrotropine (TRHR), le récepteur de la ghréline (GHSR), les récepteurs à l'arginine-vasopressine (AVPR), les récepteurs de la *gonadotropin-releasing hormone* (GHRHR), le récepteur de l'oxytocine (OXTR) et un récepteur orphelin.

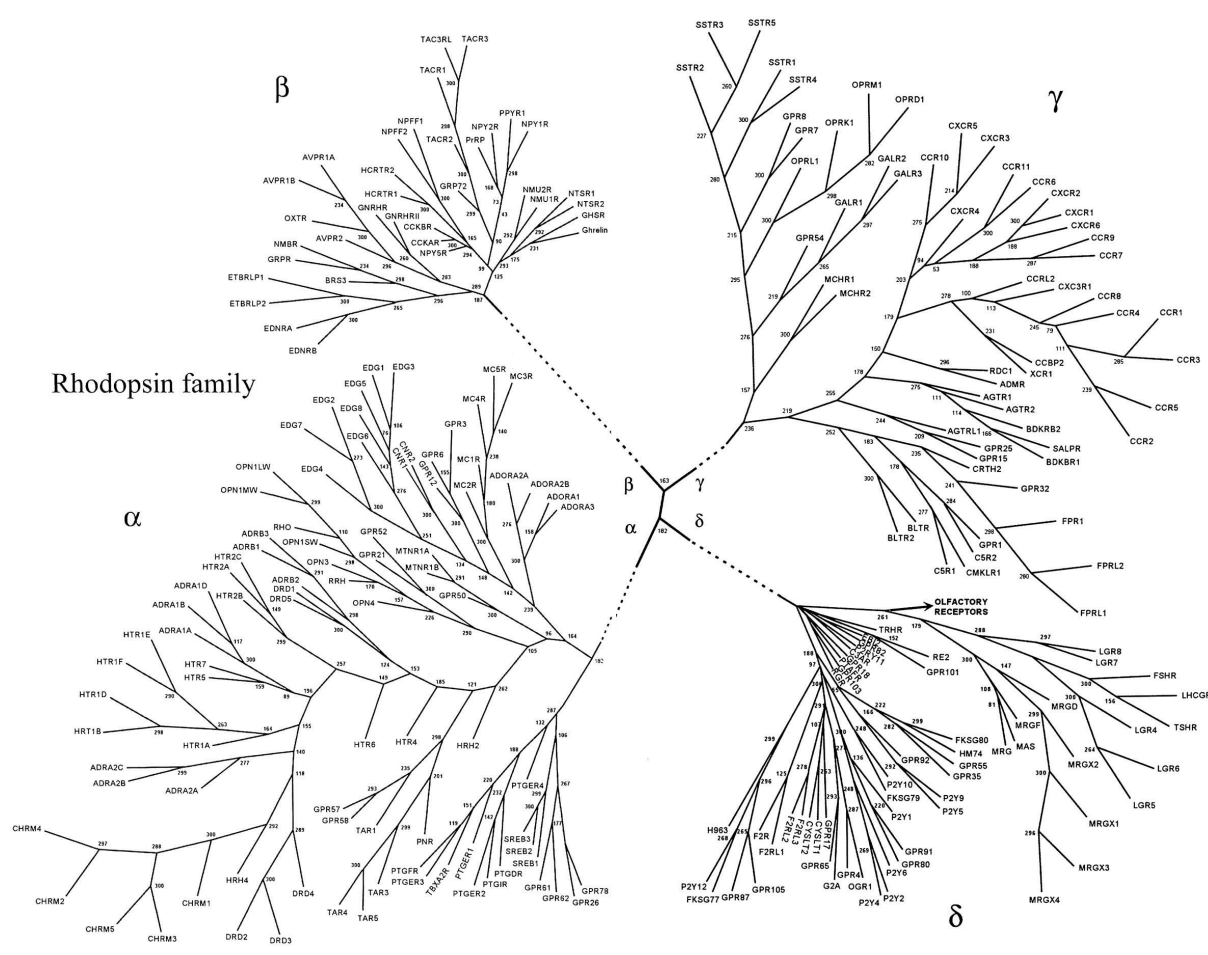


Figure 3 : Classification de la famille de la Rhodopsine selon Fredriksson [2]

La famille de la rhodopsine se décompose en 5 branches : 1) La branche α comprenant les récepteurs des prostaglandines, des amines, des opsines et les récepteurs MECA. 2) La branche β comprend les récepteurs des peptides. 3) La branche γ comprend les récepteurs SOG, les récepteurs MCH, et les récepteurs des chimiokines. 4) La branche δ comprend les récepteurs des glycoprotéines, de la famille MAS, purinergiques et les récepteurs olfactifs.

Le groupe χ de la famille de la rhodopsine se divise en 3 branches principales : les récepteurs SOG, les récepteurs MCH et les récepteurs de chimiokines. La branche des récepteurs SOG contient les récepteurs de la somatostatine (SSTR), les récepteurs des opioïdes (OPR), les récepteurs de la galanine (GALR) et le récepteur GPR54. GPR7 et GPR8 se lient au neuropeptide W. Tous les ligands connus de cette branche sont des peptides mais ils ne partagent pas de similitudes structurales. La branche des récepteurs MCH a comme ligand l'hormone de mélanocortine qui est un neuropeptide cyclique de 19 acides aminés. La branche des récepteurs de chimiokines contient les récepteurs de chimiokines classiques (CCR, CXCR), les récepteurs de l'angiotensine (AGTR ou ATR), les récepteurs de la bradykinine (BDKRB), le récepteur de l'APJ (APJR) et un grand nombre de récepteurs orphelins. La plupart des ligands sont des peptides (chimiokine, angiotensine, apéline, bradykinine).

Le groupe δ de la famille de la rhodopsine se divise en 4 branches principales : les récepteurs de la famille MAS, des glycoprotéines, des purines, et les récepteurs olfactifs. La branche des récepteurs de la famille MAS comprend le récepteur de l'oncogène MAS1 (MAS) et les récepteurs apparentés à la famille MAS (MRG et MRGX). La branche des glycoprotéines contient les récepteurs des hormones gonadotropes (FSHR, TSHR et LHCR) et les *leucine-rich-repeat-containing GPCR* (LGR). La branche des récepteurs purinergiques contient les récepteurs des peptides formylés (FPR), les récepteurs des nucléotides (P2Y), les récepteurs activés par la thrombine (F2R), les récepteurs de lipides comme les leucotriènes cysteinyles (CYSLT) et un grand nombre de récepteurs orphelins. La branche des récepteurs olfactifs est estimée à 460 membres.

1.2.4 Classification par Analyse par Composante Principale (ACP)

Le laboratoire a développé son propre système de classification par ACP [26]. Cette classification fait apparaître 3 groupes distincts au sein de la famille de la rhodopsine dont un est caractérisé par la présence d'une proline à la position 2.58 (figure 4). Ce groupe réunit les récepteurs de chimiokines, 14 des 18 récepteurs SOG, les récepteurs MCH appartenant au groupe χ selon la classification de Fredriksson, les récepteurs purinergiques appartenant au groupe δ selon la classification de Fredriksson et 7 récepteurs orphelins.

1.3 Nos récepteurs d'intérêt

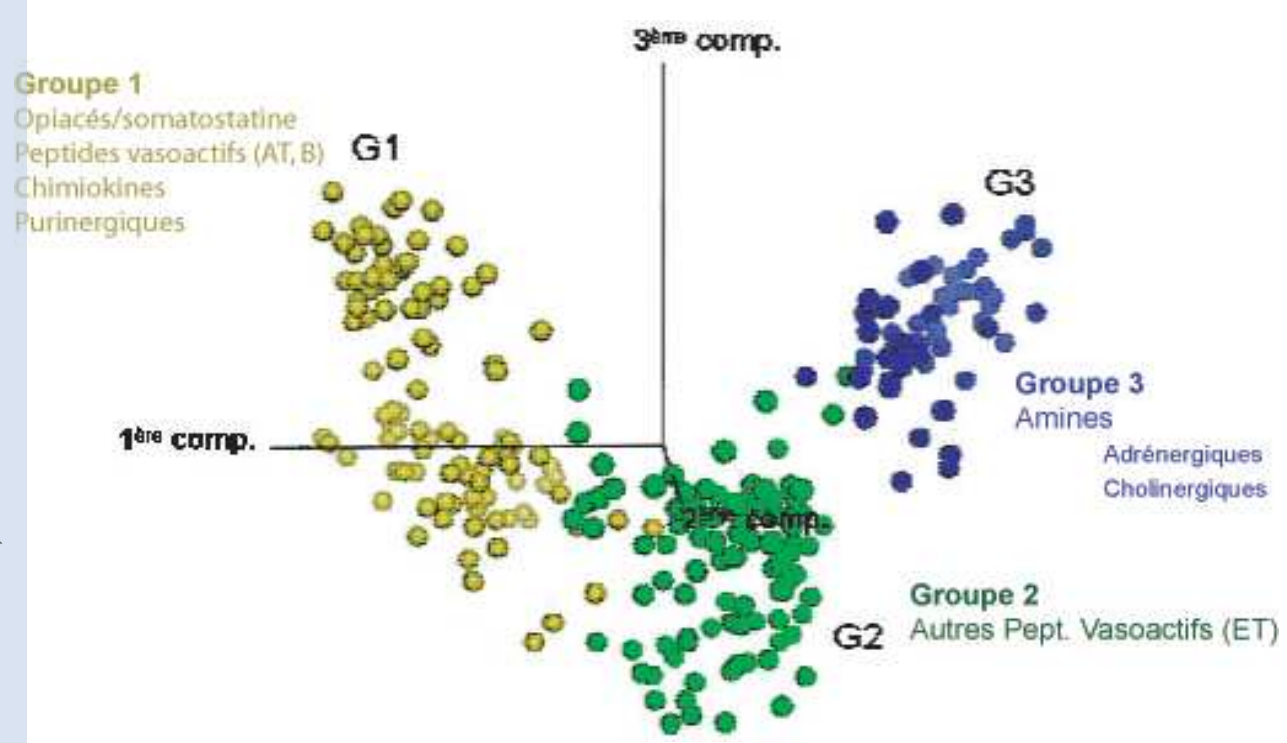


Figure 4 : Classification de la famille de la Rhodopsine par ACP

Le laboratoire a développé un système de classification propre en se basant sur l'Analyse par Composante Principale (ACP). Cette classification regroupe les récepteurs de classe A en 3 sous groupes nommés groupe 1, 2 et 3.

Notre laboratoire étudie les voies de transduction impliquées dans la réponse microvasculaire à la pression et au débit. Ces voies font intervenir les récepteurs de l'angiotensine II : AT1 et AT2. Une étude structurale serait intéressante pour une meilleure compréhension du mécanisme d'action de ces récepteurs, mais aussi pour envisager une étude de « drug-design » sur ceux-ci. Ces récepteurs appartiennent au groupe 1 de la classification par ACP. Les récepteurs de ce groupe 1 possèdent une proline conservée au sein de TMH2 à la position 2.58. Cette proline remplace le motif GG de la rhodopsine bovine impliqué dans le renflement π de TMH2.

1.3.1 Les récepteurs de l'angiotensine : AT1 et AT2

Les récepteurs de l'angiotensine ont comme ligand l'angiotensine, peptide vasoactif. L'angiotensine a un rôle important dans le système rénine-angiotensine-aldostérone (RAA). Le système rénine-angiotensine-aldostérone est une cascade de régulation endocrinienne et enzymatique. C'est un système hormonal que l'on trouve dans le rein et qui sert à préserver l'homéostasie hydrosodée (l'équilibre entre les ions Na^+ et l'eau).

1.3.1.1 Fonctionnement

L'angiotensinogène, protéine inactive produite par le foie, circule dans le sang. C'est le précurseur des peptides actifs, et le seul substrat de la rénine. D'autres protéases (cathepsine G, kallikréine tissulaire) peuvent dégrader l'angiotensinogène et permettre la libération de peptides actifs (figure 5).

En cas de baisse de la pression dans l'artère rénale (il existe d'autres stimuli : baisse de la natrémie au niveau du tube contourné distal et stimulation des cellules juxta-glomérulaires par le système bêta-adrénergique), la rénine (une enzyme parfois considérée comme une hormone) est sécrétée par le rein. L'angiotensinogène, sécrété par le foie, est clivé par la rénine pour donner un décapeptide appelé *Angiotensine I*, inactif. L'angiotensine I sera ensuite transformée en angiotensine II principalement au niveau du poumon par l'enzyme de conversion de l'angiotensine (ACE) qui est une carboxypeptidase. L'angiotensine agit en se fixant sur ses récepteurs transmembranaires (dont il existe deux types AT1 et AT2 qui ont des rôles parfois antagonistes).

1.3.1.2 Les récepteurs AT1 et AT2

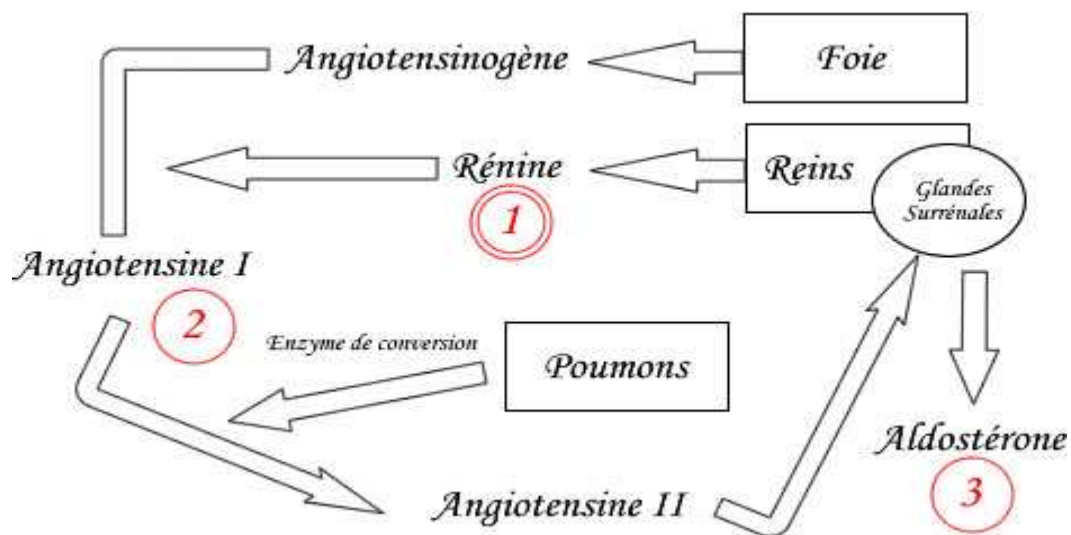


Figure 5 : Système Rénine-Angiotensine-Aldostérone

Le système rénine-angiotensine-aldostérone (RAA) est une cascade de régulation endocrinienne et enzymatique. 1) En cas de baisse de pression dans l'artère rénale, la rénine va cliver l'angiotensinogène en angiotensine I. 2) L'angiotensine I sera transformée en angiotensine II par l'ACE (enzyme de conversion). L'angiotensine II agit sur les récepteur AT1 et AT2. 3) Cette fixation va notamment stimuler la sécrétion d'aldostérone qui provoquera l'augmentation de la volémie par réabsorption du sodium (Na).

Les récepteurs AT1 et AT2 ont une identité de séquence d'environ 30%, mais ont une affinité semblable pour l'angiotensine II, qui est leur ligand principal.

Le récepteur AT1 est le récepteur de l'angiotensine le plus étudié. Il est couplé à la phospholipase C augmentant ainsi le niveau cytosolique de Ca^{2+} . Il inactive également l'adénylate cyclase et active diverses tyrosines kinases. Les effets médiés par le récepteur AT1 incluent la vasoconstriction, la synthèse et la sécrétion d'aldostérone, l'augmentation de la sécrétion de vasopressine, l'hypertrophie cardiaque, l'augmentation de l'activité noradrénergique périphérique, la prolifération des cellules musculaires lisses vasculaires, la diminution du flux sanguin rénal, l'inhibition de la rénine rénale, le relargage de sodium dans les tubules rénaux, la modulation de l'activité du système nerveux central sympathique, la contractilité cardiaque, le contrôle osmotique central et la formation de la matrice extracellulaire.

Les récepteurs AT2 sont plus présents chez le fœtus et le nouveau-né. Les effets médiés par le récepteur AT2 incluent l'inhibition de la croissance cellulaire, le développement des tissus foetaux, la modulation de la matrice extracellulaire, la régénération neuronale, l'apoptose, la différenciation cellulaire et peut-être la vasodilatation.

1.3.2 Les autres récepteurs du groupe 1

1.3.2.1 Les autres récepteurs de peptides vasoactifs

Les récepteurs des peptides vasoactifs du groupe I comprennent les récepteurs de l'angiotensine II, les récepteurs de la bradykinine et le récepteur de l'APJ. Les récepteurs de l'angiotensine II ont été décrits ci-dessus. Les récepteurs de la bradykinine fixe la bradykinine, une kinine fabriquée sous l'action d'une enzyme libérée dans la sueur. Elle a un effet vasodilatateur. La bradykinine est un puissant vasodilatateur endothélium-dépendant, qui provoque la contraction des muscles lisses non vasculaires, augmente la perméabilité vasculaire et est également impliquée dans le mécanisme de la douleur. Chez les mammifères, on connaît deux types de récepteurs de la bradykinine. Le récepteur B_1 est uniquement exprimé suite à une lésion tissulaire et on suppose qu'il joue un rôle dans la douleur chronique. Le récepteur B_2 est constitutivement actif et participe à l'action vasodilatatrice de la bradykinine. Le récepteur de l'APJ fixe l'apéline. L'apéline et son récepteur sont présents dans le SNC et le système cardiovasculaire. L'apéline et la vasopressine sont régulés en sens opposé aussi bien dans le cerveau que dans la circulation sanguine. Cette régulation croisée

semble avoir pour finalité de maintenir l'équilibre hydrique de l'organisme, en optimisant la sécrétion de vasopressine et en évitant ainsi une perte d'eau supplémentaire au niveau rénal. L'apéline a des effets cardiovasculaires propres. Elle augmente la force contractile du myocarde par un effet inotrope positif et a un effet vasodilatateur.

1.3.2.2 Les récepteurs des chimiokines

Les récepteurs des chimiokines sont des récepteurs chimiotactiques. Le rôle majeur des chimiokines est de guider la migration des cellules. Certaines chimiokines contrôlent les cellules du système immunitaire au cours du processus de surveillance immunitaire, comme le fait de diriger les lymphocytes vers les ganglions lymphatiques. Leur nom est dérivé de leur capacité à induire le chimiotactisme, ce sont des cytokines chimiotactiques (**chemotactic cytokines**). Les chimiokines sont des protéines de petite taille (environ 8-10 kDa) caractérisées par la présence de 4 résidus cystéine situés à des localisations conservées.

1.3.2.3 Les récepteurs opioïdes/somatostatines

La branche des récepteurs opioïdes/somatostatines présent dans le groupe 1 de la classification du laboratoire contient les récepteurs de la somatostatine (SSTR), les récepteurs des opioïdes (OPR) et les récepteurs du neuropeptide W. Ces récepteurs sont les plus anciens au niveau phylogénique. Toutes les autres branches présentes dans le groupe 1 de la classification du laboratoire sont apparues plus récemment chez les Vertébrés. Les récepteurs de la somatostatine se lient à la somatostatine. Les récepteurs des opioïdes fixent entre autres les opioïdes endogènes tels que les endorphines, les endomorphines, les dynorphines et les enképhalines. Ces récepteurs agissent sur la neurotransmission GABAergique. Les récepteurs du neuropeptide W agissent au niveau de la régulation de l'homéostasie.

1.3.2.4 Les récepteurs MCH

La branche des récepteurs MCH comporte deux récepteurs : MCHR1 et MCHR2. Ces récepteurs ont comme ligand l'hormone de mélan-concentration, neuropeptide cyclique de 19 acides aminés synthétisé dans l'hypothalamus. Les rôles physiologiques de cette hormone sont nombreux chez l'homme : la régulation centrale du comportement alimentaire, la

régulation de l'axe hypothalamo-hypophyso-surrénalien (axe HPA) durant le stress, la régulation des processus sensitifs, etc....

1.3.2.5 Les récepteurs purinergiques

Ces récepteurs participent à plusieurs fonctions telles que la réactivité vasculaire, l'apoptose ou la sécrétion de cytokines. Les membres de cette famille sont les récepteurs P2Y, les récepteurs des peptides formylés (FPR), les récepteurs des nucléotides (P2Y), les récepteurs activés par la thrombine (F2R) et les récepteurs de lipides comme les leukotriènes cysteinyles (CYSLT). Les récepteurs P2Y sont activés par des nucléotides, tels que ATP, ADP, UTP, UDP et UDP-glucose. Les ligands des récepteurs FPR sont des peptides formylés, facteurs chimiotactiques des neutrophiles permettant leur activation. Les récepteurs activés par la thrombine, F2R, induisent l'hydrolyse des phosphoinositides. Ils jouent un rôle dans l'activation plaquettaire et le développement vasculaire. Les récepteurs de lipides comme les leukotriènes cysteinyles (CYSLT) interviendraient dans la relaxation de l'endothélium vasculaire pulmonaire.

En conclusion, la majorité des récepteurs de ce groupe 1 sont impliqués dans la régulation vasculaire (récepteurs des peptides vasoactifs, récepteurs purinergiques), dans la coagulation du sang (récepteurs purinergiques, activation plaquettaire) et dans le système immunitaire (récepteurs des chimiokines et récepteurs chimiotactiques). Leur apparition est liée à celle des vertébrés avec son système cardiovasculaire et son système immunitaire spécifiques.

1.4 Structure des RCPG de classe A

1.4.1 La structure de la rhodopsine et les RCPG

Les RCPG possèdent un domaine central commun constitué de 7 hélices α transmembranaires et hydrophobes (composés de 25 à 35 résidus) que nous noterons TMH1 à

TMH7 [27]. Ces domaines transmembranaires sont connectés par trois boucles extracellulaires et trois boucles intracellulaires de taille variable. La région N-terminale est localisée au niveau extracellulaire tandis que la région C-terminale est située dans la région intracellulaire. La séquence protéique comporte plusieurs sites de N-glycosylation situés dans l'extrémité N-terminale et dans la 2^{ème} boucle extra cellulaire (E2), et deux cystéines dans les boucles extracellulaires E1 et E2 impliqués dans un pont disulfure. Les hélices transmembranaires possèdent une forte identité de séquence tandis que les boucles extracellulaires et intracellulaires présentent une grande variabilité de séquence, conduisant à la spécificité de chaque récepteur (figure 6).

La première structure cristallographique d'un RCPG fut résolue en 2000 avec l'obtention de la structure de la rhodopsine (2.3 Å)[9, 28, 29].

La rhodopsine agit au niveau de la transduction du signal lumineux. À l'état normal, l'apoprotéine opsine (environ 40 kD) et le rétinol 11-*cis* (un dérivé de la vitamine A) se lient pour former la rhodopsine. Le rétinol 11-*cis* agit comme agoniste inverse et l'activité intrinsèque de l'apoprotéine se trouve donc réduite. En présence de lumière, le rétinol 11-*cis* se photo-isomérisse en rétinol tout-*trans*, entraînant un changement conformationnel de la protéine. L'absorption du photon par la rhodopsine active une protéine G, la transducine (Td) entraînant le processus de phototransduction avec signal électrophysiologique passant à travers la membrane des cellules photoréceptrices.

La structure de la rhodopsine bovine a permis de confirmer la prédiction de l'arrangement ainsi que de l'orientation des sept hélices transmembranaires des RCPG.

1.4.2 Analyse des segments transmembranaires de la famille de la rhodopsine

L'analyse statistique des résidus formant les hélices transmembranaires de la famille de la rhodopsine montre de nombreux motifs conservés [30]. Au moins un des acides aminés est très fortement conservé dans chaque hélice transmembranaire : N au sein de TMH1 (100%), D au sein de TMH2 (94%), R au sein de TMH3 (96%), W au sein de TMH4 (96%), P au sein de TMH5 (77%), P au sein de TMH6 (100%), et P au sein de TMH7 (96 %). Cet acide aminé est facilement identifiable sur un alignement multiple de séquences et sert de point de référence pour définir un code numérique applicable à toute la famille de la rhodopsine dans la nomenclature de Ballesteros [12]. Cela facilite la comparaison entre les résidus des segments transmembranaires des différents récepteurs. Chaque résidu est identifié par deux chiffres : le premier (de 1 à 7) correspond à l'hélice correspondante; le second indique la position relative au

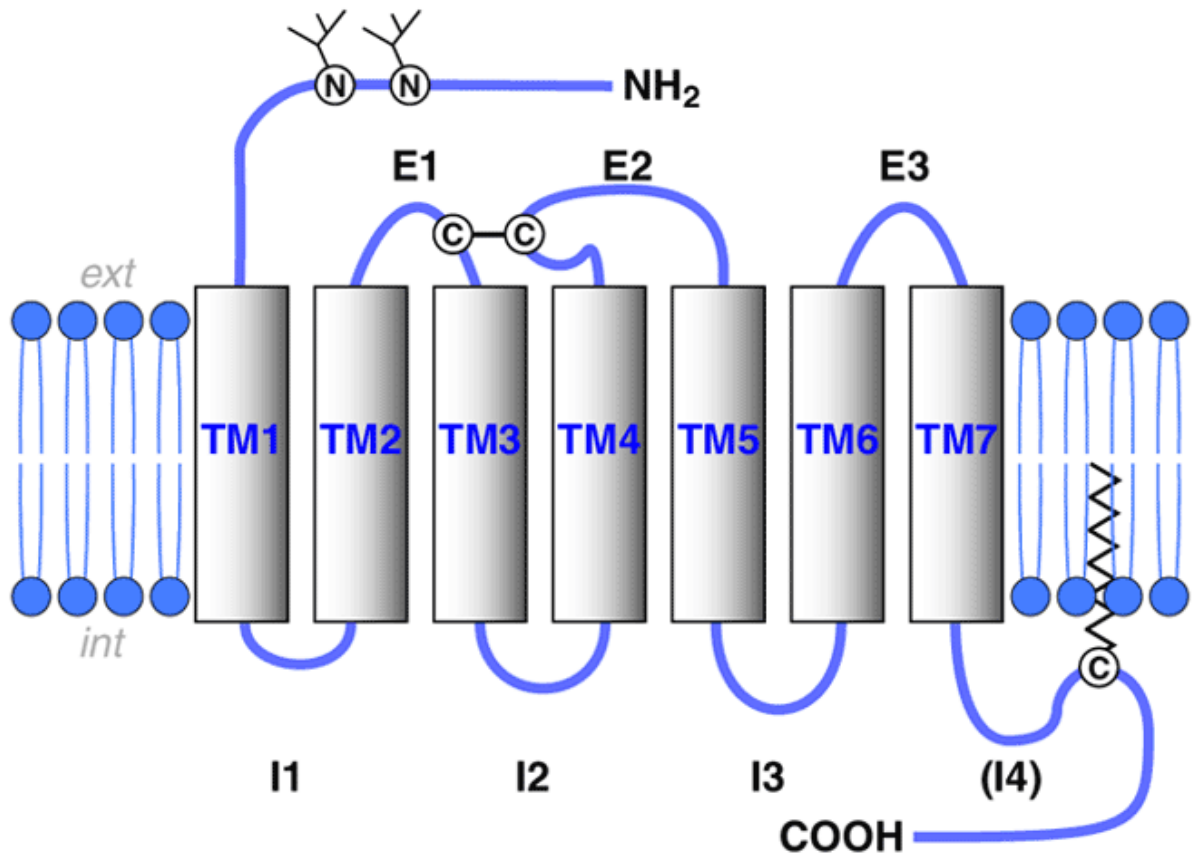


Figure 6 : Représentation schématique en deux dimensions des récepteurs couplés aux Protéines G. [31]

L'extrémité amino-terminale (NH₂) est extracellulaire et est représentée avec deux sites de Nglycosylation. Les sept hélices α transmembranaires (1 à 7) sont reliées par 3 boucles extracellulaires (E1 à E3) et 3 boucles intracellulaires (I1 à I3). Un pont disulfure relie la boucle E1 à la boucle E2. L'extrémité carboxy-terminale (COOH) est intracellulaire. Cette dernière peut présenter un ancrage lipidique (palmitoylation d'une cystéine) dans la membrane.

résidu conservé servant de point de référence arbitrairement assigné à 50. Par exemple, N7.49 correspond à l'asparagine se situant dans la 7^{ème} hélice transmembranaire (TMH7) localisée 1 résidu avant la proline de référence P7.50. La Figure 7 montre la position, dans la structure cristalline de la rhodopsine, de ces acides aminés de référence (en rouge), et d'autres acides aminés conservés avec plus de 80% d'identité (en vert). Un fait remarquable est la présence d'une proline comme résidu de référence au sein des 3 hélices transmembranaires : TMH5, TMH6, TMH7 indiquant un rôle majeur des prolines dans la structure/fonction de ces récepteurs. La proline a une très faible probabilité d'être incluse dans une hélice α [32], où elle induit une cassure [33]. La cassure est produite afin d'éviter un encombrement stérique entre l'anneau de pyrrolidine de la proline (position i) et l'oxygène du groupe carbonyle du résidu précédent la cassure (Position $i-4$) [34], ce qui produit une inclinaison de 27° en moyenne (figure 8) [35, 36]. Le nombre important de motifs de séquences conservés au niveau des hélices transmembranaires suggère une structure commune aux RCPG de classe A. Cette homologie de structure entre la rhodopsine et les autres RCPG ne peut pas être étendue au domaine extracellulaire, du fait de la très faible conservation de séquence.

1.4.3 Les déformations des hélices de la rhodopsine

Les hélices transmembranaires de la rhodopsine ne sont pas des hélices α régulières [37, 38]. Au contraire, ces hélices montrent des cassures qui facilitent le repliement au sein de la membrane. Nous présenterons ici une analyse détaillée de ces déformations et de leurs implications pour modéliser d'autres RCPG [39].

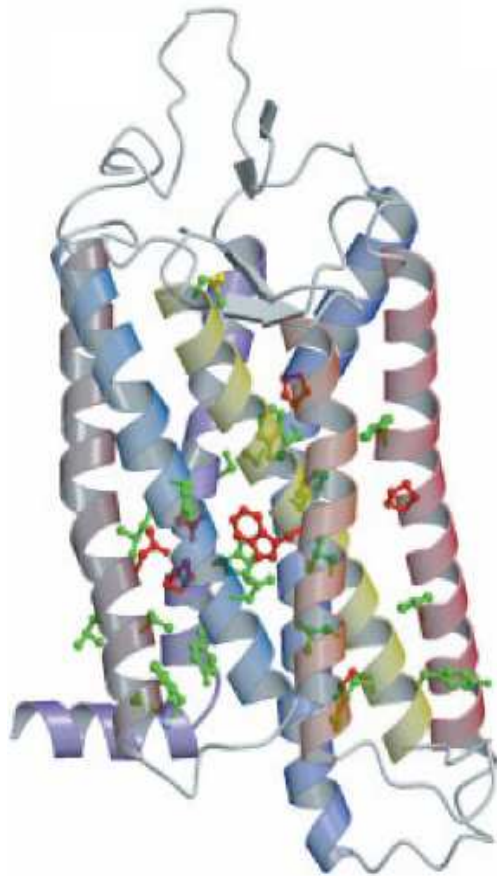


Figure 7 : Structure cristalline de la Rhodopsine Bovine

Les acides aminés de référence pour la famille de la Rhodopsine sont représentés en rouge : N1.50 (100%), D2.50 (94%), R3.50 (96%), W4.50 (96%), P5.50 (77%), P6.50 (100%), et P7.50 (96%). Les acides aminés conservés dans plus de 80% des séquences des récepteurs de la famille de la Rhodopsine sont représentés en vert.



Figure 8 : Cassure d'hélice induite par la proline

1.4.3.1 Hélice transmembranaire 1

L'hélice transmembranaire 1 (TMH1) de la rhodopsine présente une cassure au niveau du milieu de l'hélice (figure 9). Cette cassure est induite par une proline située à la position 1.48. Un grand nombre de RCPG n'ont pas de proline en 1.48 (P1.48 est seulement présent dans 9% des séquences de RCPG [30]). Dans la rhodopsine, l'asparagine fortement conservée

de l'hélice 1 (N1.50) forme des liaisons hydrogène avec les oxygènes du groupe carbonyle des résidus en position 1.46 et 7.46, liant TMH1 à TMH7.

1.4.3.2 Hélice transmembranaire 2

L'hélice transmembranaire 2 (TMH2) de la rhodopsine montre une cassure vers l'extrémité C-terminale de l'hélice. Cette cassure est induite par 2 résidus glycine successifs situés en positions 2.56 et 2.57. Cette cassure adopte une conformation en renflement π . Ce motif GG n'est pas conservé au sein des RCPG. Il peut être remplacé par une proline en 2.58 (41%), en 2.59 (36%), en 2.60 (3%) mais 20% des RCPG ne possèdent ni proline, ni glycine à ce niveau-là.

1.4.3.3 Hélice transmembranaire 3

L'hélice transmembranaire 3 (TMH3) de la rhodopsine présente une faible torsion vers l'extrémité C-terminale. Cette hélice possède une serine en 3.42 dont la chaîne latérale est impliquée dans une liaison hydrogène avec la chaîne latérale de l'asparagine 2.45. Cette liaison hydrogène entraîne une faible torsion de l'hélice α . Des serines et des thréonines sont observées fréquemment dans la séquence de TMH3 des RCPG. Par exemple, Thr3.37 est conservé dans 85% des séquences de la famille des neurotransmetteurs et une analyse bioinformatique montre que ce résidu est probablement impliqué dans une torsion comme celle observée chez la rhodopsine [40, 41].

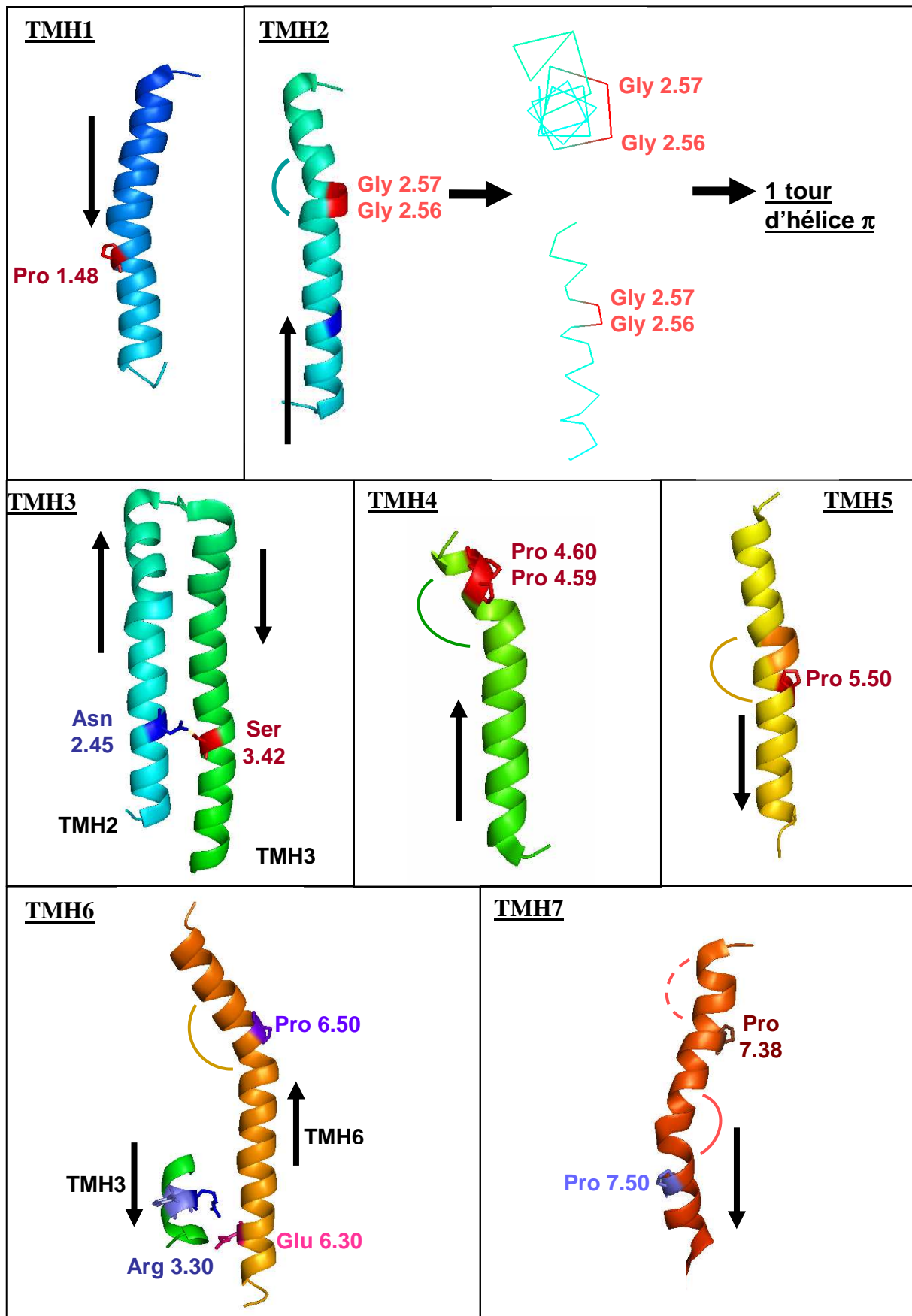


Figure 9 : Les résidus impliqués dans les cassures des 7 hélices transmembranaires de la Rhodopsine

1.4.3.4 Hélice transmembranaire 4

L'hélice transmembranaire 4 (TMH4) de la rhodopsine présente une cassure de 35° à l'extrémité C-terminale. Ceci est dû à la présence de 2 prolines consécutives aux positions 4.59 et 4.60. Ce motif Pro-Pro n'est pas conservé et se retrouve uniquement dans quelques opsines des Vertébrés et quelques sous-familles de récepteurs d'amines. En revanche, les autres RCPG possèdent un résidu Pro unique, ou un motif Pro-X-Pro à ces positions.

1.4.3.5 Hélice transmembranaire 5

L'hélice transmembranaire 5 (TMH5) de la rhodopsine présente une faible torsion de 13° seulement situé au milieu de l'hélice. Cette cassure est induite par la proline fortement conservée à la position 5.50. L'angle de torsion est nettement inférieur à l'angle moyen d'un repli induit par une proline. L'analyse structurale indique un motif structural particulier dans la rhodopsine où la courbure induite par la proline de TMH5 est diminuée grâce à l'ouverture de l'hélice (>3.6 résidus/tour, angle de torsion <100°) au tour 5.45-5.48. Cela permet de diminuer le « encombrement » stérique entre l'anneau de pyrrolidine de P5.50 et l'oxygène du groupe carbonyle à la position 5.46.

1.4.3.6 Hélice transmembranaire 6

L'hélice transmembranaire 6 (TMH6) de la rhodopsine présente une cassure vers l'extrémité C-terminale de l'hélice. Cette cassure est induite par la proline fortement conservée située à la position 6.50. Cette cassure est stabilisée dans la rhodopsine par l'interaction ionique entre Glu6.30 et Arg3.30 du motif (D/E)RY de TMH3 [42]. La rupture de cette interaction ionique, facilitée par la protonation de (D/E)3.49 [43], induit de grands changements conformationnels de TMH3 et TMH6 [44], considérés comme une étape essentielle de l'activation des RCPG. Le résidu acide à la position 6.30 est seulement conservé dans 32% des séquences (D, 7% ; E, 25%). Cependant, de nombreux RCPG possèdent un résidu basique à cette position 6.30 (34% ; K=18% ; R=16%) empêchant l'interaction directe avec R3.30. Ces récepteurs possèdent probablement un réseau totalement différent d'interactions inter-hélicales sur le côté intracellulaire qui reste à identifier.

1.4.3.7 Hélice transmembranaire 7

L'hélice transmembranaire 7 (TMH7) de la rhodopsine présente deux cassures distinctes situées à l'extrémité N-terminale et C-terminale de l'hélice. La cassure située à l'extrémité N-terminale est induite par une proline à la position 7.38. Cette proline n'est pas conservée au sein des RCPG. La cassure située à l'extrémité C-terminale est induite par une proline fortement conservée à la position 7.50. Les résidus précédant la proline ont une conformation très déformée, déplaçant la cassure induite par la proline 5 résidus en amont au niveau de la serine 7.45 avec un angle très marqué de 45°. Cette proline appartient au motif NPXXY présent dans la rhodopsine qui est fortement conservé au sein des RCPG.

1.4.4 Conclusion

La famille de la rhodopsine est caractérisée par un certain nombre de résidus fortement conservés au sein des hélices transmembranaires, dont certains sont impliqués dans des hélices cassées. Ces cassures sont induites pour la plupart par des prolines mais d'autres résidus peuvent aussi en induire. Il serait intéressant de savoir s'il y a une corrélation entre la structure des cassures d'hélices et les séquences correspondantes. Mon travail consistera dans un premier temps à effectuer une analyse exhaustive des hélices cassées au sein d'une base de données de structures non redondantes dans le but de comparer les cassures induites par les prolines et celles observées en absence de proline.

Une question additionnelle concerne les récepteurs de peptides vasoactifs. Ces récepteurs possèdent une proline en position 2.58 au sein de l'hélice transmembranaire 2 (TMH2), cruciale pour l'activation du récepteur. Mon travail consistera à savoir si cette proline, non conservée au sein de tous les RCPG, induit une réorganisation complète du renflement de TMH2 observé dans la rhodopsine bovine.

Pour répondre à ces questions, il a été nécessaire de mettre en place un certain nombre d'outils.

2 OUTILS MIS EN PLACE

L'analyse exhaustive des hélices cassées a nécessité la mise en place d'outils informatiques. D'une part, nous avons créé des scripts permettant d'effectuer des analyses structurales et séquentielles automatisées. Il a aussi fallu concevoir une base de données d'hélices cassées nous permettant de nous assurer de la non redondance des données tout en conservant un panel exhaustif de structures d'hélices cassées. Des recherches de motifs structuraux similaires à un motif précis d'hélice cassée ont été ensuite effectuées à l'aide du programme SPASM (<http://xray.bmc.uu.se/usf/spasm.html>). Enfin, nous avons conçu une base de données relationnelle des RCPG, non redondante, et permettant une sélection des récepteurs par nos propres critères.

2.1 Outils d'analyse de la PDB

L'analyse des hélices cassées a nécessité la création de scripts permettant d'effectuer des analyses structurales, stéréochimiques, et séquentielles automatisées. Tous ces scripts ont été écrits dans le langage de programmation Perl (version 5.8.0). Une base de données de structures d'hélices cassées a été mise en place.

2.1.1 Outils d'analyse structurale et stéréochimique

Les analyses structurales effectuées ont été le calcul des angles dièdres, le calcul de l'axe d'un segment d'hélice et le calcul des angles d'inclinaison et de giration entre deux hélices.

2.1.1.1 Calcul des angles dièdres

La chaîne principale contient trois liaisons covalentes par acide aminé. La liaison peptidique étant une liaison plane, il reste deux liaisons simples autour desquelles la rotation est possible. On peut donc déterminer la conformation du squelette d'un acide aminé à partir de deux angles dièdres, φ et ψ (figure x). L'angle dièdre φ du résidu est défini par les quatre atomes successifs du squelette : $C_{i-1}-N_i-C\alpha_i-C_i$, le premier carbonyle étant celui du résidu précédent. L'angle dièdre ψ est défini par les quatre atomes successifs du squelette : $N_i-C\alpha_i-C_i-N_{i+1}$, le second amide étant celui du résidu suivant (figure 10).

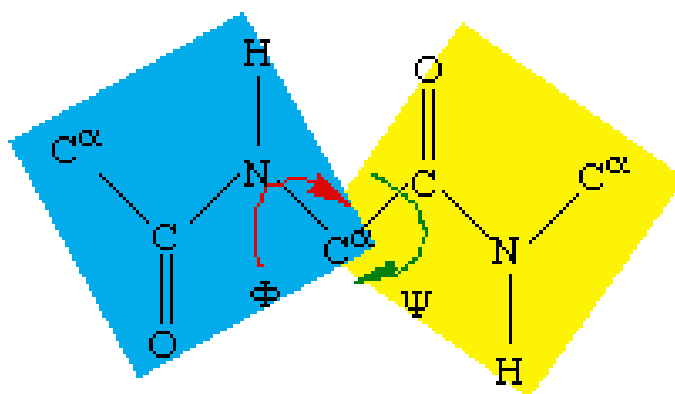


Figure 10 : Angles dièdres ϕ et ψ dans une chaîne protéique

Angle ϕ = Rotation N-C α

Angle ψ = Rotation C α -C'

Toutes les valeurs des angles φ et ψ ne sont pas possibles car certaines conduisent à des contacts trop proches entre atomes qui sont énergétiquement très défavorables. Une étude systématique des combinaisons admissibles d'angles φ et ψ a été réalisée par le biophysicien indien Gopalamudram Narayana Ramachandran en 1963 [45]. Il a imaginé une représentation sous forme graphique de l'espace (φ, ψ) qui porte aujourd'hui le nom de diagramme de Ramachandran (figure 11). Ce diagramme montre trois principales zones énergétiquement favorables. Lorsqu'on analyse une structure de protéine, on observe que la majeure partie des acides aminés ont des combinaisons d'angles (φ, ψ) qui s'inscrivent à l'intérieur de ces zones. Les deux principales régions correspondent aux structures secondaires régulières qui sont principalement observées dans les protéines : la région des hélices α et celle des feuillets β . La troisième région, plus petite, correspond à une conformation en hélice gauche ($\varphi > 0$).

2.1.1.2 Calcul des vecteurs normaux parallèles à l'axe d'un segment d'hélice

Le vecteur normal parallèle à l'axe d'un segment d'hélice a été calculé par la méthode de Kahn [46]. Ce vecteur est obtenu à partir des coordonnées des $C\alpha$ de 4 acides aminés consécutifs de l'hélice : $C\alpha_1$, $C\alpha_2$, $C\alpha_3$ et $C\alpha_4$. La bissectrice de l'angle créé par les 3 premiers $C\alpha$ ($C\alpha_1$, $C\alpha_2$ et $C\alpha_3$) est calculée, permettant d'obtenir un vecteur \vec{V}_1 perpendiculaire à l'axe de l'hélice. Cette opération est répétée en se déplaçant d'un résidu avec $C\alpha_2$, $C\alpha_3$ et $C\alpha_4$, permettant d'obtenir un deuxième vecteur \vec{V}_2 perpendiculaire à l'axe de l'hélice. Le vecteur directeur de l'axe de l'hélice \vec{H} est le produit vectoriel de ces deux vecteurs \vec{V}_1 et \vec{V}_2 (figure 12).

2.1.1.3 Calcul des angles d'inclinaison et de giration entre 2 hélices successives

Une cassure entre deux hélices H1 et H2 peut être définie par deux angles : l'angle d'inclinaison et l'angle de giration. L'angle d'inclinaison représente l'angle entre l'axe de l'hélice N-terminale (H1), et l'hélice C-terminale (H2). L'angle d'inclinaison entre H1 et H2 est défini par l'angle θ_b . Le cosinus de θ_b est donné par le produit scalaire des vecteurs normaux parallèles à l'axe de l'hélice H1 et à l'axe de l'hélice H2. L'angle de giration est l'angle qui définit l'orientation de l'axe de l'hélice H2 dans l'espace tridimensionnel par rapport à la section plane de l'hélice H1 (le plan perpendiculaire à l'axe de cette hélice H1).

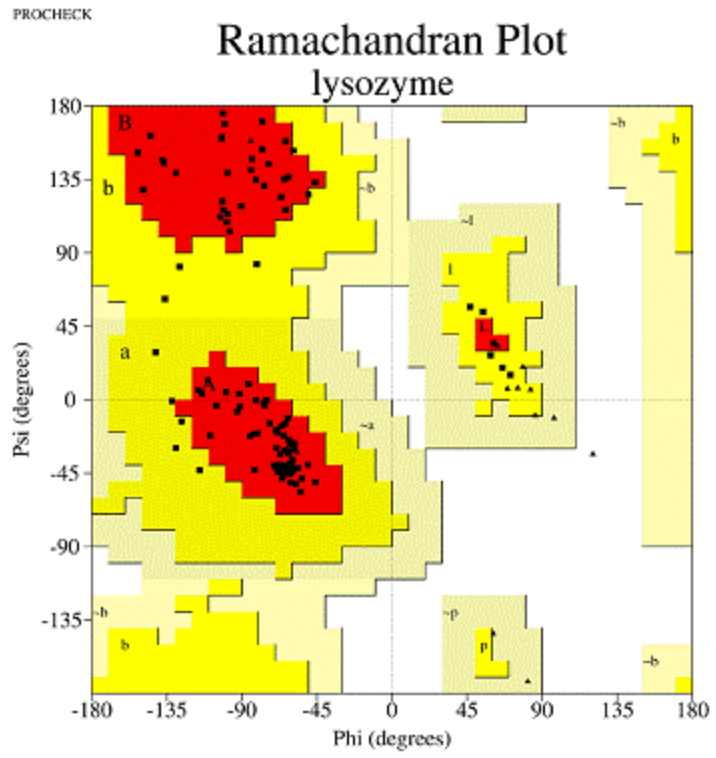


Figure 11 : Diagramme de Ramachandran

Représentation des valeurs ϕ et ψ des angles diédriques des conformations les plus stables

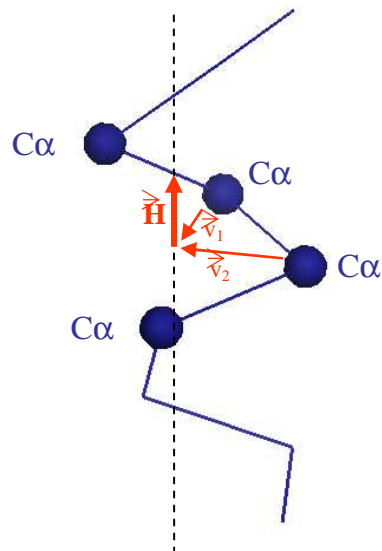


Figure 12 : Détermination du vecteur directeur \vec{H} de l'axe local de l'hélice selon la Méthode de Kahn

L'angle de giration entre H1 et H2 est défini par l'angle θ_w . Le vecteur de référence du plan perpendiculaire à l'axe de H1 joint le centre de H1 au carbone C α du résidu C4 de H1 (figure 13). Il a été choisi car il permet une description intuitive de la géométrie de la cassure. Pour calculer cet angle, nous avons utilisé les coordonnées polaires et leur relation avec les coordonnées cartésiennes.

2.1.2 Outils d'analyse des séquences

2.1.2.1 Formule pour le calcul de la propension des acides aminés

La propension P_{ak} de l'acide aminé a de se produire à la position k d'un motif K a été déterminée par le rapport du nombre d'acide aminé observé à cette position k (n_{ak}) au nombre d'acide aminé prévu (n_{aexp}) à partir de la distribution des acides aminés dans le set de données entier :

$$P_{ak} = \frac{n_{ak}}{n_{aexp}}$$

$$P_{ak} = \frac{\binom{n_{ak}}{n_k}}{\binom{n_a}{n}}$$

avec P_{ak} , la propension de l'acide aminé a à la position k , n_k est le nombre de motifs k , n_a est le nombre d'acide aminé a dans le set de données entier et n , le nombre total d'acides aminés dans le set de données entier.

2.1.2.2 Formule Z-score pour analyser les propensions

Pour éviter les biais dus à des échantillons de petite taille, nous avons calculé un Z-score pour chaque acide aminé a à la position k . Le Z-score est défini comme suit :

$$Z_{ak} = \frac{n_{ak} - n_{aexp}}{\sigma_a}$$

où σ_a , l'écart type du nombre observé d'un acide aminé a , correspond à la racine carrée de la valeur attendue. Cette formule, initialement proposée par Engel et DeGrado [47], a été validée expérimentalement (figure 14). Le pourcentage de chaque acide aminé a a été calculé pour l'ensemble de la PDB25. Puis, 100 tirages aléatoires d'un total de 50, 100, 200 et 500 acides aminés ont été effectués dans la PDB25. Pour chaque tirage, le nombre observé de chaque

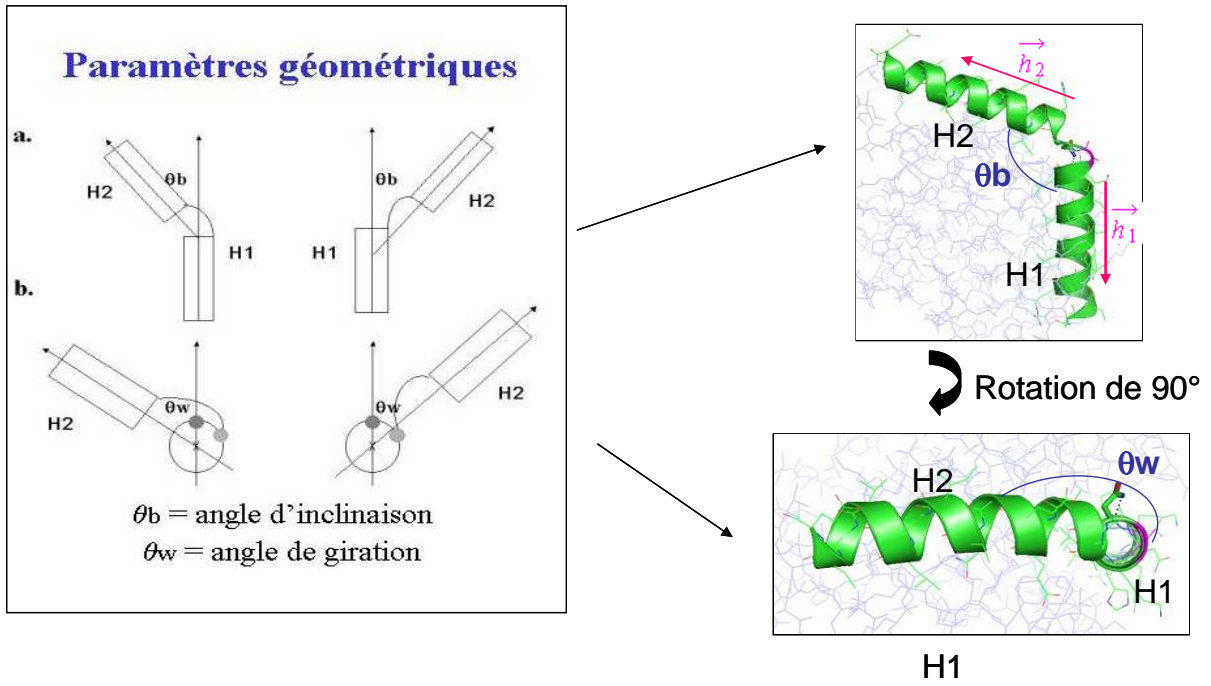


Figure 13 : Angles de giration et d'inclinaison de la cassure

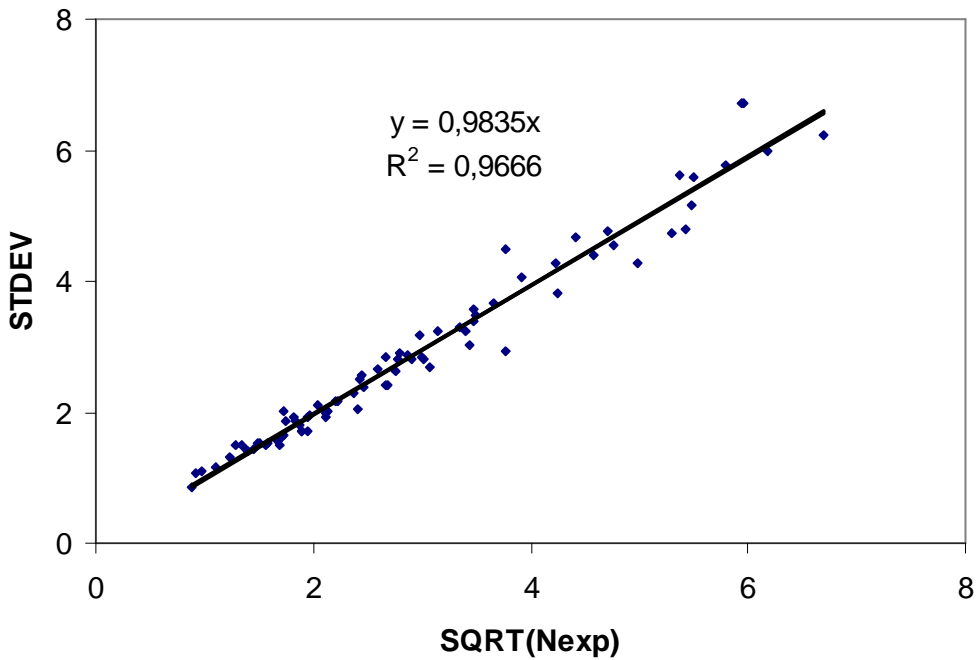


Figure 14 : Validation de la formule de Engel et Degrado

acide aminé a , $N_{\text{obs}}(a)$, a été calculé et comparé à la valeur attendue, $N_{\text{exp}}(a)$, à partir du pourcentage de a dans la base de données. Il en résulte une distribution gaussienne. La déviation standard a été calculée. Nous avons reporté sur le graphique de la figure 14 la déviation standard en fonction de la racine carrée du nombre attendu de chaque acide aminé. Cette courbe valide la formule $\sigma_a = (N_{\text{exp}}(a))^{1/2}$.

2.1.2.3 Prédiction des conformations

Pour pouvoir prédire une conformation C , des matrices de scores spécifiques des positions ont été développées. Le score $S(C)$ d'une séquence $a_1 \dots a_i \dots a_n$ pour une conformation C est donné par :

$$S(C) = \sum w_C(a_i)$$

où a_i est l'acide aminé à la position i dans la séquence considérée et $w_C(a_i)$, son poids relatif à la même position i dans la conformation C .

Une séquence a été assignée à la conformation donnant le meilleur score. L'exactitude de la prédiction a été définie par la matrice Qscore [48] qui donne le rapport Qx^{obs} du nombre de conformations correctement prévues x au nombre de conformations observées x :

$$Qx^{\text{obs}} = M_{xx} / (\sum_y M_{xy})$$

et le rapport Qx^{pred} du nombre de conformations correctement prévues x au nombre de conformations prévues x :

$$Qx^{\text{pred}} = M_{xx} / (\sum_y M_{yx})$$

où M_{xy} est le nombre de séquences observées dans la conformation x et prédites dans la conformation y .

2.1.3 La base de données des motifs de hélices cassées : HXH

Il existe différents types de cassure d'hélice : des hélices tordues mais considérées tout de même comme hélices continues, des hélices cassées au niveau d'un seul résidu et enfin des hélices cassées au niveau de plusieurs résidus. Pour nous assurer de la cohérence de nos résultats, nous avons décidé de créer une base de données de motifs d'hélices cassées par un seul résidu dans un premier temps, cette cassure d'hélice étant nette et permettant de travailler sur des données propres et sûres. Pour uniformiser les définitions des structures secondaires et plus particulièrement des hélices α , nous avons utilisé le programme DSSP (Dictionary of Secondary Structure of Proteins, [49]). DSSP est basé sur le réseau de liaisons hydrogènes.

L'angle entre l'oxygène du groupe carbonyle (C=O) en i et le groupement NH en $i+4$ doit être supérieur à 120° , la distance entre l'hydrogène de l'amide et l'oxygène du groupe carbonyle doit être inférieur à 2.5 \AA et l'énergie de liaison hydrogène E doit être inférieure à -0.5 kcal/mol .

L'analyse statistique des cassures d'hélices nécessite un grand nombre de motifs structuraux d'hélices cassées. Le choix de la Protein Data Bank (PDB, [50]) a été immédiatement écarté. Cette base de donnée est extrêmement redondante avec 43 633 structures répertoriées et certaines de ces structures sont très incertaines. Nous nous sommes donc tournées vers des sous-ensembles de la PDB : les PDBselect ([51],[52]). La PDB_90select comprend toutes les structures ayant moins de 90% d'identité de séquences. Malgré ce filtre, cette base de données reste encore redondante avec 8152 structures répertoriées (18% de la PDB totale). La PDB_25select comprend les structures de protéines ayant moins de 25% d'identité de séquences. Mais avec 2810 structures répertoriées (6% de la PDB), cette base de données ne permet pas d'avoir suffisamment de motifs de hélices cassées pour effectuer une analyse statistique. Il a donc fallu créer notre propre base de données nous permettant de filtrer les motifs tout en gardant un nombre suffisant de ces motifs pour effectuer l'analyse.

La PDB_90select a été choisie pour initier de notre base de données (8152 structures). Le premier filtre a été de ne garder que des structures protéiques obtenues par cristallographie à rayon X (7363 structures). Seules les protéines solubles ont été retenues du fait du trop faible nombre de protéines membranaires pour des études statistiques (6775 structures restantes). Les structures ayant une résolution trop faible ont été exclues (limite : résolution $<2.5 \text{ \AA}$ et Rvalue <0.250) (5189 structures restantes). Seules les structures contenant deux hélices α de 5 résidus minimum séparées par un seul résidu ont été retenues (1164 structures restantes). Afin de s'assurer de la bonne structure des hélices α , un filtre Phi/Psi a été mis en place pour les résidus des hélices entourant la cassure et ce jusqu'au résidu en +5 pour l'hélice suivant la cassure et -5 pour l'hélice précédant la cassure (1123 structures restantes). Enfin, ces structures sont alignées au niveau de la cassure avec une extension de 5 résidus de chaque côté. Tous les segments ayant plus de 50% d'identité de séquences ont été regroupés et celui du groupe ayant la plus haute résolution a été sélectionné, les autres étant éliminés de la base de données. Ce dernier filtre nous a permis de nous assurer de la non redondance des données. Notre base de données : HXH (hélice – acide aminé quelconque – hélice) comprend finalement 837 structures (figure 15).

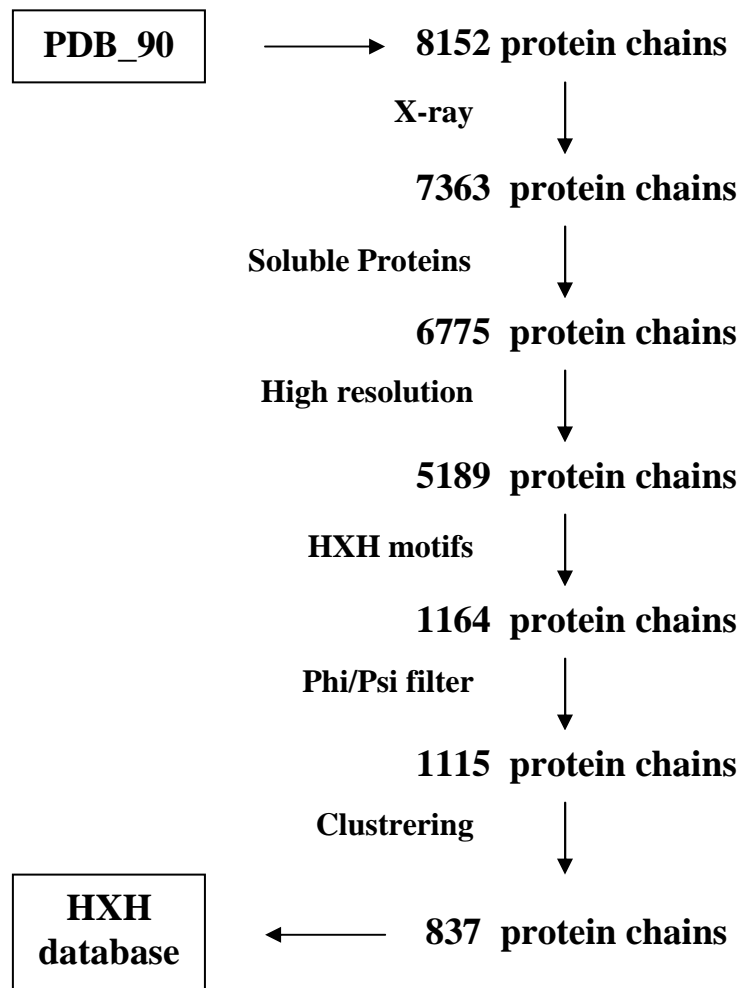


Figure 15 : Filtres choisis pour la création de la base de données HXH

2.1.4 Recherche de motifs : SPASM

La recherche des motifs d'hélices cassées similaires à la cassure de l'hélice transmembranaire 2 de la rhodopsine a été faite grâce au programme SPASM (<http://xray.bmc.uu.se/usf/spasm.html>) signifiant « SPatial Arrangements of Side chains and Main chains » [53]. Ce programme est utilisé pour trouver des motifs structuraux tridimensionnels. Nous avons écrit des script en Perl afin d'automatiser l'analyse des données.

SPASM est un programme qui fonctionne uniquement avec des résidus d'acides aminés (reconnus par le fait qu'ils contiennent au moins 3 atomes du squelette carbonés). Ils sont représentés par les coordonnées de leur atome C α et, pour les résidus autres que la glycine, par un pseudo-atome localisé au centre de gravité de la chaîne latérale. Les avantages de cette représentation sont multiples. Premièrement, la base de données des structures est nettement plus petite qu'avec une représentation complète des structures. Deuxièmement, les opérations sur la base de données se font plus beaucoup plus vite. Troisièmement, cette approche évite de nombreux problèmes associés à des ambiguïtés dans l'identification expérimentale de certains atomes.

Des contraintes peuvent être rajoutées lors de la recherche d'un motif. La première contrainte impose la conservation de la direction de la séquence. La seconde contrainte exige que les résidus voisins dans le motif soient conservés dans les structures de la base de données. Enfin la dernière contrainte porte sur les interruptions de séquence. Le ou les interruptions de séquence présentes dans le modèle peuvent rester de la même longueur lors de la recherche des motifs dans la base de données.

La superposition des structures retrouvées par SPASM avec la structure de référence, l'hélice TMH2 de la rhodopsine bovine a été réalisé grâce au programme MODELLER (<http://www.salilab.org/modeller/>) [54].

2.2 Création de la base de données relationnelle des RCPG

La majorité des travaux effectués au laboratoire nécessite l'utilisation de groupes de récepteurs couplés aux protéines G. Les classifications générales établies dans les bases de données publiques sont trop globales et ne répondent pas aux exigences du laboratoire sur ce point. Il était donc nécessaire de créer un outil, sous forme d'une interface de base de

données, permettant d'établir des classifications sur mesure auquel seraient ajoutées des options de gestion, d'exportation et de traitement des données.

Cela permettra notamment de classer les protéines selon la présence ou non de proline en positions 2.58, 2.59 ou 2.60 au niveau de TMH2.

Cette base de données a été réalisée avec l'aide précieuse d'un étudiant du DESS EGOIST de Rouen, Matthieu Moreau.

L'ensemble des données des RCPG provient de la base de données GPCRdb [20] La nomenclature a été uniformisée avec la base de données Uniprot. Le langage utilisé est le langage SQL (Structure Query Language). C'est un langage standard et normalisé permettant la définition (LDD), la manipulation (LMD) et le contrôle (LCD) des données d'une base de données relationnelle. Il est intégré dans de nombreux SGBD (Système de Gestion de Base de Données) tels que MySQL, PostgreSQL ou encore Oracle. Le SGBD utilisé lors de ce travail est la version 4.1.9 de MySQL. La version 3.23, par défaut sur Red Hat 9, a été remplacée par la 4.1.9 afin d'harmoniser l'utilisation de l'interface et de la base de donnée sur le serveur et le développement local réalisé à l'aide de la plateforme de développement Web EasyPHP 1.8 comprenant Apache (serveur http), MySQL, PHP ainsi que le gestionnaire de base de données phpMyAdmin. Elle permet d'exécuter des pages PHP sur un serveur Web installé localement et facilite donc la gestion d'un projet. L'interface a donc été développée sous PHP 4.3.10, la version 5.0 n'ayant pas été choisie à cause de nombreuses différences syntaxiques avec les versions précédentes.

2.2.1 Modèle conceptuel de données (MCD)

En concertation avec tous les membres de l'équipe de bioinformatique et en fonction des différents besoins propres au laboratoire, un modèle conceptuel de données (MCD) (figure 16) a été réalisé à l'aide du logiciel libre dbdesigner4. Il comprend notamment les entités suivantes :

- l'entité *gpcr* qui contient l'ensemble des informations relatives à la protéine comme son identifiant UniProt, sa séquence, la date d'insertion dans la base... Elle possède de nombreuses associations vers d'autres entités (*classe*, *bibliographie*...). Chaque occurrence, correspondant à un RCPG, possède un identifiant unique empêchant la redondance.
- l'entité *classification* permettant d'insérer des classifications originales et propres au laboratoire. Pour permettre une complète autonomie de cette entité, elle utilise

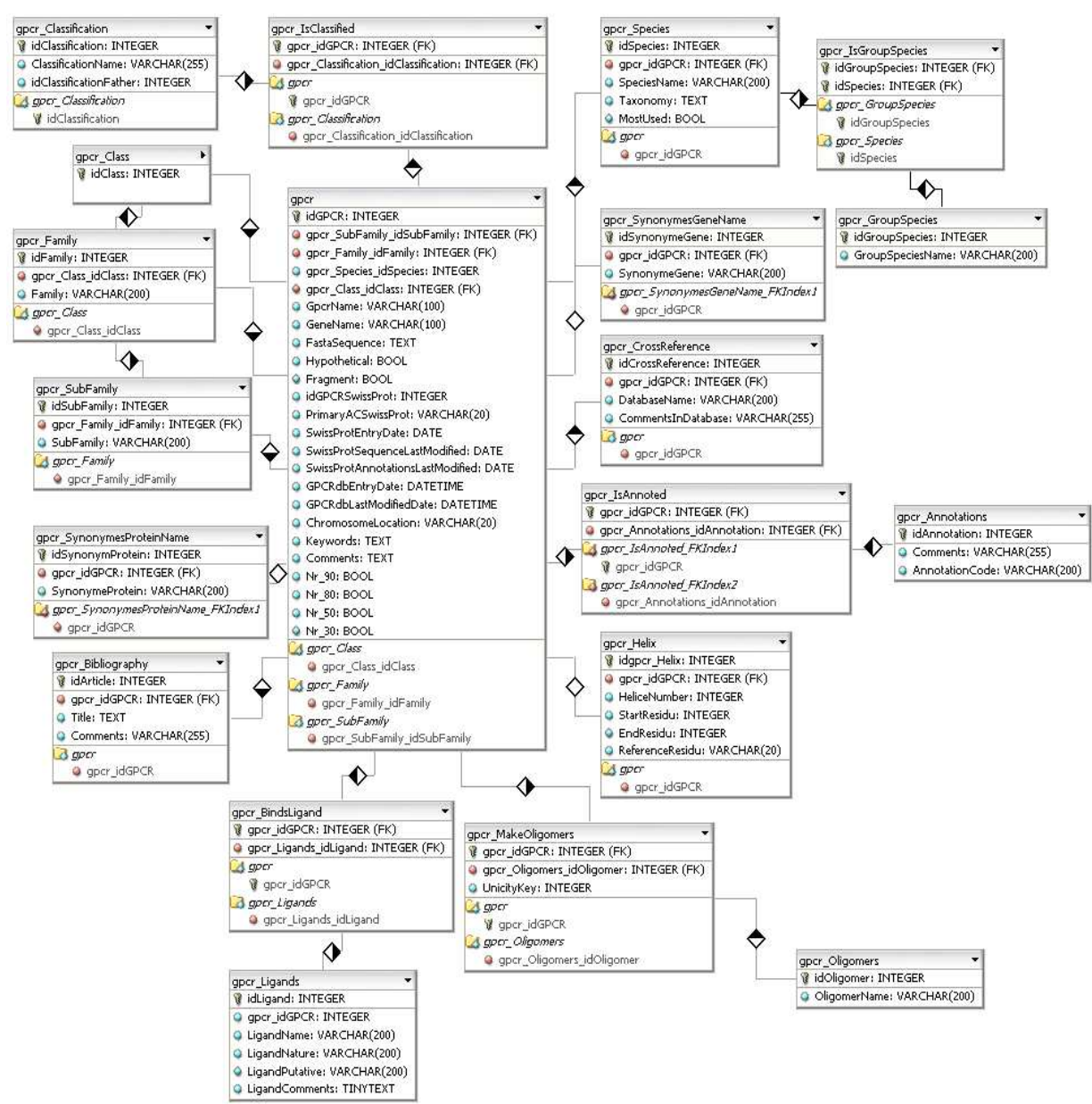


Figure 16 : Modèle conceptuel de données

- une relation réflexive. Ceci permet entre autres d'introduire la classification des RCPG en fonction de la position de la proline au sein de TMH2 : positions 2.58, 2.59, 2.60 ou l'absence de proline.
- les entités *espèce* et *groupeEspèce* qui contiennent, d'une part l'ensemble des espèces dont proviennent les RCPG et, d'autre part, des groupes d'espèces définis par l'utilisateur (par exemple, l'ensemble des virus ou des vertébrés...).
- les entités *classe*, *famille* et *sous – famille* intégrant les classifications déjà établies officiellement (comme la classe A, la famille des chimiokines...). Ces entités auraient pu être intégrées à l'entité *classification*, mais il a été décidé de les faire devenir indépendantes dans un souci de lisibilité et de rapidité d'accès.
- L'entité *NR* permettant de regrouper les protéines avec une identité de séquences inférieure à un seuil donné.

Les autres entités permettent de stocker les données annexes contenues dans les fichiers UniProt ou des paramètres susceptibles d'être utiles au laboratoire dans le futur.

L'équipe de bioinformatique dispose, avec la mise en place de cette plateforme, d'un outil central pour la majorité de ses travaux. Elle intervient aux différents niveaux de la recherche : de l'élaboration des jeux de données, correspondant aux sous-ensembles de RCPG dans notre cas, grâce à la base de données, sa fonction de recherche et la possibilité d'établir des classifications propres, au stockage des résultats et leur mise en commun. Elle a permis d'assurer un gain de temps et une cohérence plus importante entre les différents membres.

2.3 Conclusion

Tous ces outils développés ont permis d'étudier de manière exhaustive les hélices cassées. Cette étude a donné lieu à une première publication acceptée dans *PROTEINS* : «Comprehensive analysis of the helix-X-helix motif in soluble proteins ». Puis, en s'appuyant sur ces résultats, une analyse de la cassure de l'hélice transmembranaire 2 a été réalisée pour étudier les possibles différences structurales existant entre les récepteurs possédant la proline en positions 2.58, 2.59 ou 2.60. Cette analyse a donné lieu à une seconde publication en cours de soumission dans *PROTEINS* : « Structural conservation in the GPCR family: Evolution of transmembrane helix 2 ».

3 RESULTATS

3.1 L'analyse du motif hélice-X-hélice au sein des protéines solubles

L'intérêt du laboratoire pour les peptides vasoactifs et en particulier l'angiotensine II, nous a amené à étudier les récepteurs couplés aux protéines G de la famille de la rhodopsine (classe A) auxquels appartiennent les récepteurs AT1 et AT2 [7], en vue de leur modélisation. La rhodopsine bovine est employée couramment pour la modélisation par homologie des récepteurs couplés aux protéines G de classe A car elle fut jusqu'en Octobre 2007 la seule structure cristalline d'un RCPG. De nombreuses hélices coudées ou cassées sont présentes dans la rhodopsine. Pour modéliser nos récepteurs d'intérêts, une meilleure compréhension de ces cassures est nécessaire. Une étude exhaustive des hélices cassées trouvées au sein des structures de la PDB résolues par rayon X a été initiée afin d'analyser plus particulièrement l'influence de certains résidus, en particulier les prolines et les glycines, au niveau des cassures. Cette étude s'est portée sur les protéines solubles car les protéines membranaires sont trop peu représentées au sein de la PDB pour permettre d'obtenir des résultats significatifs. Le motif étudié se définit par 2 hélices séparées par un résidu de jonction X, aussi noté HXH. En effet, de nombreuses études se sont portées sur les propriétés de motifs définis par 2 hélices séparées par 2 ou 3 résidus, en revanche, peu d'études existent encore sur les hélices séparées par un seul résidu. Pourtant, une analyse de la PDB montre que le motif structural HXH est presque aussi fréquent que les motifs définis par 2 hélices reliées par 2 ou 3 résidus. Comme expliqué dans le chapitre 3.1, il a fallu créer notre propre base de données pour avoir suffisamment de motif structuraux non redondants afin de permettre une analyse qualitative, mais surtout quantitative de ces motifs. L'analyse s'est portée sur notre base de données comprenant 837 motifs (HXH). Nous avons analysé les propriétés stéréochimiques et les relations séquence-structure de ce motif grâce au programme présenté en 3.2. Plus particulièrement, nous avons étudié la géométrie de la chaîne principale (les angles dièdres ϕ et ψ), la géométrie des deux hélices impliquées (définie par les deux angles d'inclinaison et de giration), la présence de liaisons hydrogène entre les atomes de la chaîne principale, entre eux et avec les atomes des chaînes latérales voisines, la préférence des acides aminés pour différentes positions du motif HXH, le profil d'hydrophobicité et l'accessibilité au solvant. Cette étude a donné lieu à une publication acceptée dans PROTEINS et présentée ci-après.

3.1.1 Article I : "Comprehensive analysis of the helix-Xhelix motif in soluble proteins"

COMPREHENSIVE ANALYSIS OF THE HELIX-X-HELIX MOTIF IN SOLUBLE PROTEINS

Julie DEVILLE, Julien Rey and Marie CHABBERT*

UMR CNRS 6214 - INSERM U771, Université d'Angers, Faculté de Médecine, 3 rue Haute de Reculée, 49045 ANGERS, FRANCE.

SHORT TITLE: The helix-X-helix motif

KEY WORDS: 3D data mining, protein structure, helix kink, proline, glycine

* to whom correspondance should be addressed ; Dr Marie Chabbert, UMR CNRS 6214 – INSERM U771, Faculté de Médecine d'Angers, 3 rue Haute de Reculée, 49045 ANGERS, France; Tel : 33 2 41 73 58 73; Email : marie.chabbert@univ-angers.fr.

ABSTRACT

α -Helices are the most common secondary structures found in globular proteins. In this report, we analyze the stereochemical and sequence properties of helix-X-helix motifs in which two α -helices are linked by a single residue, in search of characteristic structures and sequence signals. The analysis is carried out on a database of 837 non-redundant helix-X-helix motifs. The kinks are characterized by the bend angle between the axes of the N-terminal and C-terminal helices and the wobble angle corresponding to the rotation of C-terminal helix axis on the plane perpendicular to the N-terminal one. The phi-psi dihedral angles of the linker residue are clustered in six distinct areas of the Ramachandran plot: two areas are located in the additional allowed alpha region (α_1 and α_2), two areas are in the additional allowed beta region (β_1 and β_2) and two areas have a positive phi values (α_L and β_M). Each phi/psi region corresponds to characteristic bend and wobble angles and amino acid distributions. Bend angles can vary from 0 to 160°. Most wobble angles correspond to a counter-clockwise rotation of the C-terminal helix. Proline residues are rigorously excluded from the linker position X but have a high propensity at position X+1 of the β_1 and β_2 motifs (12 and 7, respectively) and at position X+3 of the α_1 motifs (9). Glycine linkers are located either in the α_L region (20%) or in the β_M region (80%). This latter conformation is characterized by a marked bend angle ($124 \pm 18^\circ$) and a clockwise wobble. Among other amino acids, Asn is remarkable for its high propensity (>3) at the linker position of the α_2 , β_1 and β_2 motifs. Stabilization of HXH motifs by H-bonds between polar side chains of the linker and polar groups of the backbone is determined. A method based on position-specific scoring matrices is developed for conformational prediction. The accuracy of the predictions reaches 80% when the method is

applied to Proline-induced kinks or to kinks with bend angles in the 50°-100° range.

INTRODUCTION

Understanding the relationship between amino acid sequence and protein structure is crucial to develop methods aimed at improving structure predictions and designing *de novo* proteins. Secondary structure (SS) is a crucial level in the hierarchical classification of protein structure. α -Helices and β -sheets, the major secondary structure elements, allow a simple description of protein structures and are used for protein classification (e.g. the SCOP or the CATH databases[55, 56]). These regular backbone structures are linked by loops or turns.

Prediction of the secondary structure of a protein from its sequence is a key step for structure prediction. The different propensities of each amino acid for helical, strand or random coil conformations have been acknowledged long ago[57, 58] and form the bases of numerous secondary structure prediction programs[59-62]. All the SS prediction programs rely on pattern recognition techniques combined to statistical analysis of known protein structures. The accuracy rate for a three state prediction (helix, strand and coil) is about 60% for the initial algorithms based on a single sequence analysis[57, 60, 63] and rises to about 75% for current algorithms based on multiple sequence alignment[62, 64-68]. One limitation of these programs is that they rely on the analysis of the sequence properties of a window surrounding each individual amino acid, implying a poor precision. For example, Jnet predictions are based on a 17 residue long window[65]. Although α -helices are generally better predicted than β -strands or coils, it is difficult to identify start and

stop signals correctly, in spite of strong capping sequence signals[69]. Another caveat due to the window size is that short irregular elements joining similar secondary structures can be missed. Two helices joined by a few residue linker may be predicted to form a single long helix.

The capability to develop bio-informatics tools able to predict the position and the structure of such kinks between two α -helices should be very valuable for both structure predictions and protein design. Development of these tools requires a comprehensive analysis of these structures. The properties of two or three residue long linkers joining α -helices have been widely investigated[47, 70, 71], but one residue linkers did not receive much attention yet. However, an analysis of the Protein Data Bank[47] indicated that a structural motif consisting of two consecutive helices separated by one residue linker (the helix-X-helix or HXH motif) is almost as frequent as motifs with two helices joined by a two or three residue linker.

In this study, we thus focus on one residue linkers joining α -helices in soluble proteins. We analyze the stereochemical and sequence properties of helix-X-helix motifs in a non-redundant database that we have developed. We show that the relative orientation between the two helices is determined by a few possible dihedral conformations of the protein backbone at the linker position and that each dihedral conformation exhibits distinct amino acid distributions at and around the linker. We analyse H-bonds within the HXH motif to determine the interactions stabilizing each conformation. Finally, we develop a method based on position-specific scoring matrices for conformational prediction.

MATERIALS AND METHODS

Terminology

The helices are named as helix 1 (H1) or helix 2 (H2) depending upon their relative position in the protein primary structure. Following the common nomenclature, the first i positions of an α -helix are called N1, N2, N3, ..., Ni, the first preceding position Ncap and the second preceding position N', whereas the last j positions are Cj, ..., C3, C2 and C1, followed by Ccap and C'. The linker position is called X and corresponds both to the Ccap position of helix 1 (Ccap(H1)) and to the Ncap position of helix 2 (Ncap(H2)). The i^{th} position upstream is called X- i and the j^{th} position downstream is X+ j . Position X-1 corresponds to C1(H1) and to N'(H2), whereas position X+1 corresponds to N1(H2) and to C'(H1).

Data sets

PDB_25 and PDB_90 refer to non homologous sets of protein structures selected from the 25% and the 90% threshold lists, compiled by Hobohm and Sander[52] and accessible at <http://bioinfo.tg.fh-giessen.de/pdbselect> (March 2006 release). PDB_25S and PDB_90S refer to subsets of PDB_25 and PDB_90, respectively, containing only soluble proteins whose structure was determined by crystallography with a resolution ≤ 2.5 Å and an R-factor ≤ 0.25 . Helix definition was determined by DSSP[49].

We developed a database of helix-X-helix (HXH) motifs from PDB_90, according to the procedure summarized in Fig. 1. First, soluble proteins whose structure was determined by X-ray crystallography with a resolution of 2.5 Å and a R-factor < 0.25 or better were selected from PDB_90. Proteins containing two helices separated by a single residue, with each helix at least five residue long, were selected. Identification numbers were based on HXH motifs. When the same protein contained several HXH motifs, it was identified several times. For each helix-X-helix motif, the phi/psi dihedral angles of the C-terminal five residues of helix 1 and of the N-terminal five residues of helix 2 were calculated. The structures for which either one of these angles did not match α -helical values ($-140^\circ < \phi < -30^\circ$ and $-100^\circ < \psi < 20^\circ$) were removed from the data set. The sequences of the helix-X-helix motifs were extracted and aligned, using the linker residue as an anchor. The sequence identities of the minimal HXH motifs comprising the C-terminal five residues of helix 1, the linker residue and the N-terminal five residues of helix 2 were calculated. The motifs were clustered on the basis of their sequence identity (50% threshold), using a home-developed clustering algorithm. For each cluster, the structure with the best resolution was selected.

Kink analysis

As we were interested in local properties of the kinks, the axes \mathbf{H}_1 and \mathbf{H}_2 of helices 1 and 2 were determined by Kahn's method[72] for the last helical turn before the linker residue and the first helical turn after the linker, respectively. This method uses the coordinates of four consecutive C α atoms to locally determine helix axis. The C α atoms of residues C4, C3, C2 and C1 of H1, and of residues N1, N2, N3 and N4 of H2 were used to determine the axis of helix 1 and helix 2, respectively. We verified that shifting the selected atoms by one residue (from C5 to C2 of H1 and from N2 to N5 of H2) did not alter the results.

The motif was characterized by the bend angle θ_b and the wobble angle θ_w . Definition of these angles allows a description of the relative orientation of the two helices in spherical coordinates. The bend angle θ_b defines the angle between the axes of the two helices (Fig. 2a). It is measured as:

$$\cos \theta b = \mathbf{H}_1 \cdot \mathbf{H}_2$$

A bend angle of 0° indicates that there is no change in the direction of the helix axes, whereas a bend angle of 180° indicates a total reversal in the direction of helix 2 (anti-parallel helices).

The wobble angle θw (Fig. 2b) corresponds to the rotation of the projection of \mathbf{H}_2 in a plane perpendicular to \mathbf{H}_1 . The reference vector in this plane joins the helix centre to the C α coordinate of the C4 residue of H1. This vector corresponds to the bisector of the H1 C5-C4-C3 angle. It was chosen as it allows an intuitive description of the kink geometry. Left-handed kinks correspond to a positive θw whereas right-handed kinks correspond to a negative θw . For small bend angles, determination of the wobble angle could be difficult. In these cases, the wobble angle was verified by shifting the reference atoms by one or two residues and visual inspection. In particular, each kink with a negative wobble angle was carefully controlled by visual inspection and detailed structural analysis.

Search of reverse false positives

The search for residues that are identified as helical by DSSP but have phi and psi values similar to those of the HXH linker (reverse false positives) was carried out on PDB_90S. We considered helices that were at least eleven residue long. We measured the backbone dihedral angles of each residue included from position N6 to position C6 along with the dihedral angles of the neighbour residues. Residues surrounded by at least one residue (from position -5 to -1 and from position +1 to +5) out of the α -helical range were removed from the data set, as previously done with the HXH motifs, to consider only helix distortions due to a single residue.

The search of α_2 -like structures was carried out on PDB_25S. We first searched for residues located within α -helices and with the sum of the dihedral angles equal to $-90^\circ \pm 10^\circ$. Among the 2200 residues found, we selected those with the additional criteria of $\psi > -20^\circ$. This led to a set of 59 structures used for visual inspection and further characterisation of the dihedral angles of the residues surrounding the α_2 -like residue.

Proline-containing helices

The analysis was carried out on PDB_25S. We considered helices with a minimal length of 11 residues that contained a proline located from position N5 to position C5. This led to a set of 85 structures, used for determination of the sequence and of the dihedral angles of the residues surrounding the proline.

Amino acid propensity

The propensity P_{ak} of aminoacid a to occur at position k of a motif K was determined by the ratio of the relative number of aminoacids a observed at this position k (n_{ak}) to the relative number of amino acids a expected (n_{aexp}) from the amino acid distribution in the entire data set:

$$P_{ak} = n_{ak}/n_{aexp}$$

$$P_{ak} = (n_{ak}/N_K) / (n_a/N)$$

where n_{ak} is the number of amino acids a at position k , N_K is the number of motifs K , n_a is the total number of amino acids a in the full data set and N the total number of amino acids in the full data set. The amino acid distribution of PDB_25 was used as reference after verifying the absence of any significant change in the amino acid distributions of PDB_25, PDB_90 and of the database of the proteins containing the HXH motifs.

The number N_K of the motifs analysed varied from 37 to 218 and the expected number of amino acids could be very low. To avoid biases due to small sample sizes, for each amino acid a at position k we calculated a Z-score defined as:

$$Z_{ak} = (n_{ak} - n_{aexp}) / \sigma_a$$

where σ_a is the standard deviation of n_{ak} . The standard deviation of the observed number of an amino acid a corresponds to the square root of the expected value. This formula, initially proposed by Engel and DeGrado[47], was experimentally validated by random trials in PDB_25. We considered amino acids with $|Z| > 2.0$ and 2.6 (95% and 99% confidence level, respectively).

Hydrophobicity profiles

The hydrophobicity index H at each position surrounding the reference residue X corresponds to the average of the hydrophobicity index at this position for all the motifs in the subset considered. The Eisenberg's consensus scale[73] was used after verifying that different scales[74, 75] led to similar results. Heterogeneity between subsets was tested using χ^2 homogeneity tests. For this analysis, amino acids were grouped in three classes: hydrophobic (L, M, I, V, F, Y, W), polar (H, K, R, E, Q, D, N, S, T) and other (P, G, A, C).

Solvent accessibility

Solvent accessibility was calculated with the NACCESS program[76] (<http://wolf.bms.umist.ac.uk/naccess>). The program calculates the atomic accessible surface when a 1.4 Å probe is rolled around the van der Waal's surface of the protein, according to the Lee & Richard method[77]. Residue accessibility corresponds to the sum of the atomic accessibilities.

H-bond analysis

Analysis of hydrogen bonds was carried out with the HBPLUS v3.0 program accessible at <http://www.biochem.ucl.ac.uk/bsm/hbplus/home.html>[78]. Default parameters were used throughout the analysis. We analysed main chain – main chain H-bonds involving the carbonyl oxygens of residues X-4 to X and the amide nitrogens of residues X to X+4, in search of disruption of helical pattern and alternative H-bonds. When several H-bonds were possible, the H-bond conserving the helical pattern was privileged. The only exception was for H-bonds typical of the Schellman motif for which bifurcate H-bonds were taken into account.

We analysed side chain – main chain hydrogen bonds for polar side chain Asn, Asp, Gln, Glu, Ser, Thr and His located at the linker position X. We considered hydrogen bonds involving the γ -hydroxyl oxygen of Ser/Thr residues as donor or acceptor, the δ - and ϵ -nitrogens of Asn, Gln and His as donor, the δ - and ϵ -oxygens of Asp/Asn and Glu/Gln as acceptor. We searched for putative H-bond partners among polar atoms of the backbone from position X-4 to X+4. Hydrogen bonds linking donor side chain atoms at position X to acceptor backbone oxygen at position X-*i* are denoted by O(X-*i*). Similarly, hydrogen bonds linking donor backbone nitrogen at position X+*i* to acceptor side chain atoms at position X are denoted by N(X+*i*).

Prediction of the linker conformation

Position-specific scoring matrices, based on the relative amino acid weights, were developed to predict linker conformations. The score $S(C)$ of a sequence $a_1 \dots a_i \dots a_n$ for the conformation C was given by:

$$S(C) = \sum_i w_C(a_i)$$

where a_i is the amino acid at position i in the sequence considered and $w_C(a_i)$ its relative weight at the same position i in the conformation C .

A sequence was assigned to the conformation giving the highest score. The accuracy of the prediction was defined by the Q-score matrix which gives both the ratio Q_x^{obs} of the number of correctly predicted conformations x to the number of observed conformations x :

$$Q_x^{\text{obs}} = M_{xx} / (\sum_y M_{xy})$$

and the ratio Q_x^{pred} of the number of correctly predicted conformations x to the number of predicted conformations x :

$$Q_x^{\text{pred}} = M_{xx} / (\sum_y M_{yx})$$

where M_{xy} is the number of sequences observed in conformation x and predicted in conformation y .

In order to build and test the scoring matrices on different data sets and estimate standard

deviations of the Q-scores, we used a tenfold, cross validation procedure, i.e., 9/10th of the data was used to build the matrices and the remaining 1/10th of the data was used to test them.

Molecular graphics

PYMOL (DeLano Scientific LLC, San Francisco, USA) was used for figure preparation and molecular graphics analysis. Representative structures were chosen after visual inspection. They correspond to: (α_1) O-acetylserine sulfhydrylase from *Thermotoga maritima* (PDB code: 1O58, chain D)[79]; (α_2) Gamma-glutamyl phosphate reductase from *Saccharomyces Cerevisiae* (PDB code: 1VLU, chain A) (to be published); (β_1) 28KDA Glutathione S-transferase from *Schistosoma haematobium* (PDB code: 1OE8, chain A)[80]; (β_2) Human glutathione synthetase (PDB code: 2HGS, chain A)[81]; (α_L) S-Adenosylmethionine-dependent methyltransferase from *Thermotoga maritima* (PDB code: 1M6Y, chain A)[82]; (β_M) Periplasmic iron-binding protein from *Neisseria gonorrhoeae* (PDB code, 1XC1, chain A)[83]. The structure with the typical translation observed for α_2 motifs with negative psi corresponds to chorismate mutase from *Saccharomyces cerevisiae* (PDB code: 5CSM, chain A).

RESULTS

1. The helix-X-helix database

Using the DSSP definition of secondary structure elements, we analysed the length distribution of short segments (up to three residues) joining two α -helices in a subset of PDB_25 containing 1200 protein chains of soluble proteins solved by X-ray crystallography at a resolution of 2.5 Å or better. The subset contained 8044 helices with a length of at least five residues. This length was chosen as it corresponds to the minimal length required to reliably determine α -helices with DSSP[84].

A linker length of 0 residue corresponds to two contiguous helices without any residue between them. It is seldom observed (3 examples). On the other hand, the numbers of observations for one, two or three residue linkers are very similar (160, 144 and 192 observations, respectively). This result is in agreement with a previous study[47] using a different helix definition. The distribution of the bend angles θ_b between the two helices as a function of the linker length is shown in Fig. 3. Kinks with one residue linkers overwhelm the kink distribution when the bend angle is $< 40^\circ$. Their weight plateau in the 30-40% range for θ_b comprised between 40 and 120° then progressively decreases up to a bend angle of 160°. Interestingly,

the number of observations with θ_b between 20° and 40° (31) is similar to that with θ_b between 100 and 120° (28).

These data indicate that HXH motifs may lead to large amplitude kinks and that they are frequently found in protein structures. They are present in more than 10% of protein structures and 4% of α -helices are involved in such a motif. A preliminary analysis of the distribution of the dihedral angles at the linker position X indicated the presence of six distinct conformations for this residue. The number of HXH motifs found in PDB_25 (160) was not sufficient for a detailed analysis of the sequence properties of these conformations. On the other hand, the number of helix-X-helix motifs found in PDB_90 (>1000) was sufficient for quantitative analysis but data could be severely biased. To overcome these limitations, we developed a helix-X-helix database from PDB_90, according to the procedure described in Materials and Methods (Fig. 1). This ensured a sequence identity $\leq 45\%$ (5 out of eleven residues) for the minimal eleven residue long motif. This identity rate was the best balance between the size of the sample and the redundancy. A filter based on the backbone dihedral angles of the five residues surrounding the linker residue X was added to remove false positives, i.e. structures with at least one residue in a non-helical conformation from positions X-5 to X-1 and X+1 to X+5 after the DSSP-based selection process. Non-helical dihedral angles were found for 2% of positions X+1, 1% of positions X-1 and about 1% of all the other positions. This led to the removal of 4% of the motifs, without altering the overall results. The resulting database is composed of 837 HXH motifs with associated protein structures. 47% of the sequence pairs of the minimal HXH motifs have no sequence identity, whereas 94% of them have a sequence identity $\leq 27\%$ (3 out of eleven residues), enlightening the sequence diversity of the database.

2. Backbone geometry

Using the HXH database, we analysed the geometry of the backbone at the linker position X. The phi/psi dihedral angles of this residue cluster in six distinct areas of the Ramachandran plot, indicating six possible conformations of the protein backbone (Fig. 4). None of these conformations is located in the most favoured regions of the Ramachandran plot. Two of these conformations, α_1 and α_2 , correspond to distortions of the α conformation and are in the additional allowed region surrounding the most favoured α -helix region. Together, they represent about 50% of the motifs (212 and 188 occurrences out of 837, for the α_1 and α_2 conformations, respectively). The α_1 conformation is distorted as compared to a canonical α helix ($\phi = -58^\circ$, $\psi = -47^\circ$) with a large shift of the phi angle to $-118^\circ \pm 15^\circ$, whereas the psi angle corresponds to canonical values ($-58^\circ \pm 14^\circ$). Interestingly, this conformation corresponds

to the α conformation initially proposed by Ramachandran[85]. The α_2 conformation is located at the upper limit of the allowed α region. It is characterized by a strong correlation between the phi and psi angles whose sum is almost constant ($-90^\circ \pm 10^\circ$). The average value of psi is slightly positive ($8^\circ \pm 19^\circ$) while phi is shifted to $-97^\circ \pm 20^\circ$. This conformation encompasses the 3_{10} helix area ($\phi = -74^\circ$, $\psi = -4^\circ$) and corresponds to the γ conformation reported by Efimov[70]. It makes a transition from the α to the β regions of the Ramachandran plot.

Two areas of the β region of the Ramachandran plot are also possible for the linker. They are not in the core, but in the additional allowed β region. The β_1 conformation represents 26% of the motifs (218 occurrences). It is characterized by a phi angle of $-137^\circ \pm 15^\circ$ and a psi angle of $83^\circ \pm 24^\circ$. This region was first described by Karplus[86] and is largely associated with residues preceding proline[87]. The β_2 conformation is less frequent and represents only 12% of the kinks (104 occurrences). Its dihedral angles, $\phi = -73^\circ \pm 11^\circ$ and $\psi = 122^\circ \pm 26^\circ$ are close to the values observed for residues in polyProline type II helix conformation (PPII)[87, 88]. PPII helix is a left handed helical structure formed when sequential residues adopt backbone dihedral angles centered around -75° and 145° [89, 90]. This area is also frequently associated with pre-Pro residues[86, 87].

Thirteen percent of the kinks have a positive phi value. They correspond to two distinct conformations. The β_M conformation ($\phi = 96^\circ \pm 15^\circ$, $\psi = 157^\circ \pm 23^\circ$), observed in 9% of the kinks (78 occurrences), corresponds to a mirror region of the β conformation[87, 88]. This conformation has been recently described by Lovell *et al.*[88]. It is associated with glycine residues, because the lack of a side chain for Gly produces mirror symmetry in steric constraints and in the phi/psi distribution[88]. Finally, the α_L conformation, corresponding to left α -helices with $\phi = 73^\circ \pm 24^\circ$ and $\psi = 25^\circ \pm 25^\circ$, is observed for 4 % of the kinks (37 cases).

It is worth to note that the relative weights of the six conformations are very similar in the PDB_25S (α_1 : 0.22, α_2 : 0.21, β_1 : 0.29, β_2 : 0.14, α_L : 0.04, β_M : 0.09) and in the HXH sets (α_1 : 0.25, α_2 : 0.22, β_1 : 0.26, β_2 : 0.13, α_L : 0.04, β_M : 0.09), strongly suggesting that the procedure for building the database does not lead to significant bias of the data.

3. Distribution of bend and wobble angles

The distribution of bend and wobble angles describing the relative orientation of the two α -helices was analysed for each conformation of the linker residue (Fig. 4b). Examples of each conformation are shown in Fig. 5a-f. The α_1 conformation (Fig. 5a) allows only moderate deviations from linearity with an average bend

angle of $30 \pm 15^\circ$. The α_2 conformation (Fig. 5b) allows a much larger range of bend angles from 10° to 100° with an average value of $52^\circ \pm 23^\circ$. Most α_1 and α_2 motifs have a positive wobble angle, corresponding to a left-handed helix motion. Negative wobble angles, corresponding to a right-handed helix motion, are observed only for a minor part of these motifs (3% and 6% for α_1 and α_2 motifs, respectively), usually with small bend angles ($\theta_b < 30^\circ$). The average wobble angle of these conformations is $77^\circ \pm 48^\circ$ and $79^\circ \pm 50^\circ$ for the α_1 and α_2 conformations, respectively.

In addition to the α_2 conformation, bend angles in the 40 - 100° range can be reached through the β_2 and the α_L conformations (Fig. 5d and 5e, respectively). However, the average wobble angles of these conformations are very different. The wobble angle is near zero for the α_L conformation with an average value of $22^\circ \pm 31^\circ$ (Fig. 5e). For the β_2 conformation (Fig. 5d), large amplitude counter-clockwise rotation of H2 leads to θ_w values that can be larger than 180° (average value of $162^\circ \pm 35^\circ$) and thus to positive and negative wobble values.

Bend angles ranging from 90° to 160° can be reached through either the β_1 or the β_M conformations (average values of $113^\circ \pm 16^\circ$ and $124^\circ \pm 18^\circ$, respectively). However, these two conformations correspond to very different wobble angles. The β_1 conformation leads to a left-handed motion of helix 2 with an average value of $88^\circ \pm 36^\circ$ (Fig. 5c), whereas the β_M conformation leads to a right-handed motion of helix 2 with an average value of $-55^\circ \pm 35^\circ$. This conformation leads to a reversal of helix direction. In the example of β_M conformation displayed in Fig. 5f, the bend angle between the two helices reaches 134° .

A striking result of this analysis is the anisotropy of the wobble motion. Most kink conformations correspond to a counter-clockwise rotation of helix 2. This is the case for the α_1 and α_2 conformations, except for a few cases, and for the β_1 and β_2 conformations. The negative values measured for the β_2 conformations do correspond to counter-clockwise rotation larger than 180° . The α_L conformation displays small amplitude wobble rotation. Only the β_M conformation allows a marked clockwise rotation of helix 2 with a reversed wobble motion as compared to the other conformations.

4. Main chain – main chain hydrogen bonds

HBPLUS was used for a detailed analysis of the H-bonds involving the backbone NH and CO groups of the HXH motif from position X-4 to position X+4. We checked the conservation of the conventional NH(*i*) to CO(*i*-4) H-bonds within the motif and, when these H-bonds were not conserved, alternative interactions. We thus analysed the H-bonds involving the carbonyl groups of helix 1, the

amine groups of helix 2 and either group for residue X for the six conformations (Table I). In most cases, the helical H-bond pattern involving the linker X is conserved and this residue is involved both in NH to CO H-bond with residue X-4 and in CO to NH H-bond with residue X+4. The only exception is the β_2 conformation. For this conformation, only 50% of the NH(X) to CO(X-4) H-bonds are conserved, but the overwhelming majority (92%) of NH(X+4) to CO(X) H-bonds are conserved. On the other hand, the helical *i* to *i*-4 H-bond pattern disappears between residues X+1 to X+3 and residues X-3 to X-1. The lack of these bonds is observed even for conformations with small bend angles as α_1 and α_2 .

Alternative interactions are possible for some conformations. The α_1 conformation is characterized by NH(*i*) to CO(*i*-5) H-bonds observed mainly between residues X+2 and X-3 in 77% of the α_1 motifs. A few examples of *i* to *i*-5 H-bonds are also observed between positions X+3 and X-2 when residue X+3 is not a proline. In most cases, however, the amide group of residue X+3 is not involved in H-bonds with neighbour carboxyl groups. Examples of NH(*i*) to CO(*i*-3) H-bonds are commonly found in the α_2 motif, especially from position X+1 to position X-2. 70% of the α_2 motifs involved in this interaction. NH(*i*) to CO(*i*-3) H-bonds are also observed from position X to position X-3 in 42% of the β_2 motifs. H-bonds from NH(X+3) to CO(X) and from NH(X) to CO(X-3) can also be observed for the β_1 conformation (19 and 12%, respectively). For β_M motifs, involvement of residues X-3 to X-1 and X+1 to X+3 in main chain – main chain H-bonds is very marginal.

The occurrence of Schellman motifs[91] was checked for the α_L conformation. When a α -helix terminates by a Ccap residue in the α_L conformation, a characteristic H-bond pattern links the NH of the C' and Ccap residues to the CO of the C4 and C3 residues, respectively. The occurrence of these two NH(C') to CO(C4) and NH(Ccap) to CO(C3) H-bonds forms the so-called Schellman motif [91, 92]. In the α_L motifs, possibilities of X+1 to X-4 and/or X to X-3 H-bonds, typical of the Schellman motif, are observed in about 60% of the structures. Complete Schellman motif with both H-bonds is however observed only in 25% of the HXH motifs in the α_L conformation, either in presence or not of glycine at position X. An example of a Schellman motif with an Arg linker is displayed in Fig. 5e.

5. Comparison with normal helices

Kinks with bend angles smaller than 30 - 40° are commonly included within DSSP definition of α -helices. Larger bend angles are infrequent but examples of extreme distortion have been reported[93]. We thus searched for reverse false positives, i.e. residues included in DSSP defined

contiguous helices and having dihedral angles similar to those of the linker residue of the HXH motif. The PDB_90S subset was used for this analysis in order to obtain the maximal limit of these reserved false positives. The search was carried out on helices with a length of at least 11 residues, which corresponds to the minimal length of the HXH motifs. The central residue of each 11 residue long window sliding from position N1 to position C1 was considered, after ensuring that the other residues of the window were helical. This led to 99920 residues whose Ramachandran plot is shown in Fig. 6. Comparison with the Ramachandran plot of the linker residues in HXH motifs (Fig. 4a) indicates the absence of significant overlap or of reverse false positives for most HXH conformations, except for the α_2 and β_M motifs. However, they correspond to two very different cases.

The tail of the α_2 conformation ($\psi < 0^\circ$, equivalent to 35% of the α_2 HXH motifs) overlaps the α -helix area, with average values of $-77^\circ \pm 15^\circ$ and $-14^\circ \pm 11^\circ$ for the ϕ and ψ dihedral angles, respectively. The average bend angle of these motifs (α_{2-}) is $27^\circ \pm 15^\circ$, consistent with the bend angles frequently found in normal bent helices. The search for α_{2-} -like structures in contiguous α -helices led to a set of 59 structures, with $\phi = -76^\circ \pm 9^\circ$ and $\psi = -14^\circ \pm 7^\circ$, very similar to those observed in α_{2-} motifs. These structures were visually inspected and compared to the α_{2-} motifs. The α_{2-} motifs have a very typical translation of the helix axis that can reach up to 1 Å (Fig. 7). This shift was not observed in any of the α_{2-} -like motifs that appeared either as “normally” bent helices or as straight helices (22 cases out of 59). Analysis of the dihedral angles of the residues surrounding the α_{2-} -like residue indicates that this aspect is due to a compensative distortion of the following residue whose ϕ is significantly different from standard helix ($77^\circ \pm 16^\circ$). In straight helices, this distortion is still larger with ϕ reaching $80^\circ \pm 7^\circ$. On the other hand, in α_{2-} motifs, the twist of the protein backbone is due to distortion of both the linker residue and the preceding residue ($\phi = -73^\circ \pm 18^\circ$; $\psi = -24^\circ \pm 10^\circ$). This combination of dihedral angles leads to the translation of the helix axis, preventing formation of the helical H-bonds between the residues surrounding the linker but favouring H-bonds between the amide group at position X+1 and the carbonyl group at position X-2. 82% of the α_{2-} motifs are involved in this alternative interaction.

When ψ is positive, the overall aspect of the α_2 motifs is overwhelmed by the bend resulting from the severe distortion of the linker residue. These α_{2+} motifs are characterized by an average bend angle of $65^\circ \pm 15^\circ$, resulting in a markedly bend motif (Fig. 5b). It is noteworthy that there is a continuum between these conformations as the bend angle varies linearly with the ψ/ϕ dihedral angles.

Concerning the β_M conformation, 36 residues with dihedral angles in this area were found in contiguous helices. We analyzed the corresponding structures. In 20% of these structures, H-bonds between the NH group at position X+3 and the CO group at position X-1 were detected with HBLUS. In most cases, however, these motifs could not be distinguished from the β_M motifs of our HXH database. In particular, the dihedral angles of the surrounding residues were very similar, with the preceding residue markedly different from standard helix ($\phi = -98^\circ \pm 20^\circ$, $\psi = -18^\circ \pm 16^\circ$). Clearly, DSSP may be mistaken by the very severe folding back of the protein chain and underestimates the weight of the β_M HXH motifs. Analysis of the DSSP data did not reveal any clear pattern for rationalizing the different assignments.

The α_1 conformation is clearly distinct from the conformations accessible to central positions of α -helices, with no significant overlap (Fig. 4a and 6), albeit the resulting bend angle is small ($30^\circ \pm 15^\circ$). Visual inspection of the α_1 motifs reveals an “opening” of the helix at this position (Fig. 5a), in agreement with the pattern of alternative H-bonds involving residues X+2 and X-3 (Table I). Finally, only a few reverse false positives are observed for the β_1 conformations, whereas the β_2 and α_L conformations are excluded from internal positions of DSSP-defined α -helices.

6. Amino acid propensities

The propensities p of the different amino acids at and around the linker position X were analysed from position X-4 to X+4 for the six geometrically defined conformations (Table II). To take into account the low expected number of some amino acids, data in Table II are highlighted by their Z-scores. Five out of six conformations have a characteristic glycine or proline residue with a very high propensity ($p > 6$, corresponding to $Z > 13$). The glycine residue is located at the linker position X of the α_L and β_M conformations, whereas the proline is located either at position X+1 of the β_1 and β_2 conformations or at position X+3 of the α_1 conformation. In addition to these hallmark residues, each conformation has its own amino acid distribution.

More than forty percent of the HXH motifs in the α_1 conformation possess a proline located at position X+3. Fig. 5a displays such a HXH motif with proline located three residues downstream the linker X. In addition, 9% of the α_1 motifs possess a proline at position X+2. This conformation is the only one for which the linker residue has a marked preference for hydrophobic residues, especially Val and Phe. Positions X-4, X+1 and X+4 also have a marked preference for hydrophobic residues. Aromatic residues are overrepresented at position X-4 where they are observed in 24% of the α_1 motifs. Position X-1 is polar with a high propensity

for Asn. Other positions do not display clear tendencies.

The α_2 conformation does not possess a specific overwhelming Gly or Pro residue. Proline can be observed at position X+2 with a propensity of 2.5, corresponding to only 12% of the observations. The α_2 linker has a high propensity for Asn and His ($p > 2.8$). These amino acids correspond to amino acids most favoured at the Ccap position of α -helices at the exception of glycine[94, 95]. An example of α_2 motif with an Asn residue at position X is shown in Fig. 5b. Positions X-2, X-1 and X+2 have a marked tendency for polar residues, whereas position X+1 displays a high propensity for Ala, Lys and Arg and position X+4 has a marked preference for hydrophobic residues.

Both the β_1 and β_2 conformations have a high propensity for proline at position X+1. This position corresponds to the N1 position of helix 2 and to the C' position of helix 1. The high propensity of proline for both the N1 and the C' positions has been widely reported[69, 94-97]. In these conformations, the dihedral angles of the X linker correspond to values accessible to pre-Pro residues[87, 88]. Fifty-six percent of the β_1 motifs have a proline at position X+1. An example of a β_1 motif with a proline is displayed in Fig. 5c. The linker residue has a high propensity for Asp, Asn, His and Tyr ($Z > 2.6$) and Ser ($Z > 2.0$), consistent with propensities for Ccap positions in the β -strand conformation[97]. These five residues are found at position X in two thirds of the β_1 motifs. Positions X-2, X-1 and X+2 have a marked tendency for polar residues, whereas positions X-4, X-3, X+3, X+4 are preferentially hydrophobic. When only β_1 motifs without proline at position X+1 are considered, this position has a preference for bulky charged or polar residues, specially Lys ($p = 2.9$).

Prolines are present in 32% of the β_2 motifs at the position X+1. Glu has also a high propensity for this position and is observed in 18% of the β_2 motifs. The β_2 linker has high propensities for Ser, Asp and Asn ($Z > 2.6$), typical of Ncap positions[69, 94, 96, 98]. These three residues are found at position X in 68% of the β_2 motifs, while aliphatic residues are uncommon and aromatic residues rigorously excluded at this position (no case out of 188). Thr that is usually favoured at the Ncap position of α -helices is not favoured at the linker position ($p = 0.5$). This behaviour of Thr is also observed for the other conformations.

In spite of similarities in the amino acid propensities at the linker position, the β_1 and β_2 conformations display a phase shift in the hydrophobicity of the residues surrounding the kink, observable at positions X-4, X-2 and X+4. In β_1 motifs, positions X-4 and X+4 are preferentially hydrophobic, whereas position X-2 is preferentially polar. In β_2 motifs, position X-4 displays a preference for polar residues or alanine and position X+4 has a high propensity for Arg and Gln which

are present in 40% of these motifs, whereas position X-2 has a preference for hydrophobic residues.

Both the α_L and β_M conformations have a very high propensity for glycine at the linker position. This residue is observed in 54% of α_L motifs and in all the β_M motifs except one. This single exception is a Gln residue (position 369) in the isocitrate lyase from *Mycobacterium tuberculosis*[99] (PDB access number: 1F8M, chain A) with phi/psi angles equal to $41^\circ, -122^\circ$. This high propensity for glycine is also observed in the β_M reverse false positives (34 cases out of 36). The β_M region is highly specific for glycine residues, while the α_L region has a high propensity for glycine but can accommodate other residues, in agreement with literature data[87, 88, 100]. The number of α_L motifs (37) is too low to have high Z-scores and levels of significance. However, qualitative comparison of amino acid propensities for α_L and β_M motifs indicates a marked difference in the hydrophobicity pattern of these motifs. In α_L motifs, positions X-1 and X+4 have a preference for polar residues and position X+2 for hydrophobic residues. On the other hand, in β_M motifs, position X-1 is preferentially hydrophobic with a high propensities for Leu and Phe ($p = 2.3$ and 2.9 , respectively), but Lys, whose part of the side chain is aliphatic, can also be observed. Hydrophobic residues and Ala are favoured at position X+4 ($p = 2.7, 2.1$ and 3.2 for Ala, Leu and Phe, respectively), whereas position X+2 has a marked polar character with high propensity ($p > 3$) for negatively charged residues.

7. Hydrophobicity profiles

To further compare the different conformations, we analyzed the average hydrophobicity profiles from position X-4 to position X+4, using Eisenberg's consensus scale (Fig. 8). The use of different scales did not lead to significant differences in data (not shown). For this analysis, the α_1 , β_1 , β_2 and α_L conformations were shared in two groups depending upon the presence or not of proline at position X+1 (β_1 and β_2) or X+3 (α_1) or of glycine at position X (α_L). This led to the α_{1+} and α_{1-} subsets (121 and 97 motifs, respectively), to the β_{1+} and β_{1-} subsets (121 and 97 motifs, respectively), to the β_{2+} and β_{2-} subsets (33 and 71 motifs, respectively) and to the α_{L+} and α_{L-} subsets (20 and 17 motifs, respectively). In addition, we shared the α_2 motifs in two equal subsets by the median of θ_b (55°). This parameter appeared as the most pertinent because of its linear relationship between the phi/psi dihedral angles. This led to two α_{2inf} and α_{2sup} subsets (94 motifs each one) with average θ_b of $33 \pm 15^\circ$ and $72 \pm 10^\circ$, respectively, allowing to test the homogeneity of this conformation.

As previously noted, the α_1 motif is mainly hydrophobic or neutral (Fig. 8a). The presence or not of proline at position X+3 has, however, a marked effect at positions X-1 and X+1 (99%

significance level for χ^2 homogeneity test). These positions have polar and hydrophobic characters, respectively, in the presence of proline, whereas no clear tendency is observed in its absence. The presence of proline at position X+1 does not alter significantly the hydrophobicity profile of the β_1 and β_2 motifs (Fig. 8c-d). The only exception is position X-4 of β_1 motif that is markedly hydrophobic in the presence of Pro. For the other positions except position X+1 corresponding to proline, differences observed in the average hydrophobicity index H do not correlate with significant changes in the distribution of hydrophobic or polar residues.

Major differences in the hydrophobicity of the α_{2inf} and α_{2sup} subsets are observed, especially for positions X-2 and X+2 to X+4 (99% confidence level). The hydrophobicity profile of the α_{2inf} subset is similar to that observed for the α_{1-} subset with only position X-1 exhibiting a polar character. By contrast, the hydrophobicity profile of α_{2sup} indicates that the positions surrounding the X linker from position X-2 to X+2 are polar. This is corroborated by the detailed analysis of the amino acid distribution from position X-4 to X+4 (not shown). In particular, at position X, the propensities of His and Asn raise from 0.47 and 1.21, respectively, for the α_{2low} motifs, to 5.2 and 5.6, respectively, for the α_{2sup} motifs.

The polar character at and around the linker X of the α_{2sup} subset is also observed for the β_1 motifs (positions X-2 to X+2) and for the β_2 motifs (positions X to X+2), either in the presence or not of proline (Fig. 8 and Table II). The average hydrophobicity index at position X of the β_{1+} motifs has a moderate polar character due to the increased propensity of Tyr and Phe in that case (3.0 and 1.9, respectively, vs 1.4 and 1.3 in the absence of proline). Similarly, the average hydrophobicity index of position X-1 for the β_{2+} motif is related to the high propensity of Ala ($p = 3.4$ in the presence of proline) and underestimates the polar character of this position. The length of the polar stretch decreases, however, from the α_{2sup} to the β_1 and β_2 conformations. In either case, position X-3 is hydrophobic. This residue is followed by a polar position in the α_{2sup} and β_1 motifs and by a hydrophobic position for β_2 motifs. On the other hand, position X+4 has a marked hydrophobic character in α_{2sup} motifs and polar character in β_2 motifs. This corroborates the differences in the phases of the helices observed by analysing the amino acid propensities (Table II).

Comparison of hydrophobicity profiles of the α_{L+} and α_{L-} motifs does not lead to significant differences (except for the Gly position), because of the small number of observations. On the other hand, marked differences are observed between the α_{L+} and β_M motifs for positions X-1, X+2 and X+4 (99% confidence level). This reversed hydrophobicity corroborates the observations drawn for amino acid propensities (Fig. 8e-f and Table II).

8. Solvent accessibility

The average accessible solvent area (ASA) of the residues located from position X-4 to position X+4 of the HXH motifs was measured for the different subsets (Fig. 9). The most buried motifs are the α_{1-} ones. In that case, the average solvent accessibility is lower than 50% for any position of the motif, in agreement with the hydrophobicity profile indicating low polarity (Fig. 8a). The α_{2inf} motifs are slightly more solvent exposed with the average ASA raising to 60% for positions X-1 and X+2. Helix 1 is clearly more solvent exposed in the α_{1+} motifs than in α_{1-} motifs, with an average ASA of 75% at position X-1, in agreement with the polar character of this position. This increased solvent accessibility is not observed for helix 2 (ASA \leq 40%).

The α_{2sup} , β_1 and β_2 conformations are very solvent exposed with accessibilities \geq 80% for individual positions. The presence of proline at position X+1 of the β_1 and β_2 motifs does not alter the solvent accessibility (Fig. 9c-d). Both helices have highly solvent exposed positions. Position X-1 is the most solvent exposed position of helix 1 for the α_{2sup} , β_1 and β_2 motifs, whereas a difference is observed for helix 2 whose the most solvent exposed position is either X+2 for the α_{2sup} and β_1 motifs or X+1 for the β_2 motifs (Fig. 9b-d). These solvent accessibilities are consistent with the hydrophobicity profiles and amino acid propensities indicating that polar residues (or Ala) are favoured at these positions. It is worth to note that, in spite of the polar character of the linker residue X in any of these three conformations, this residue has a limited solvent accessibility (<50%). Indeed, the solvent exposed residues delimitate the polar stretch observed for these conformations but, within this stretch, positions X and X+1 of the α_{2sup} motifs and position X of the β_1 and β_2 motifs have low solvent accessibilities.

Concerning the α_L and β_M conformations, the low accessible surface area of glycine is related to its absence of side chain. This caveat leads to very different ASA for position X of the α_{L+} and α_{L-} subsets (20% and 73%, respectively) that are not related to changes in the solvent exposed location of this residue on the protein surface. Either the α_L or the β_M conformation leads to a break in the helical pattern of H_1 , with exposure of the linker residue (see Fig. 5e and 5f as examples). Major differences between the two motifs come from the second helix. In the α_L motifs, H_2 initiates from the buried side, whereas in β_M motifs, H_2 initiates from the exposed side, leading to very different solvent exposures at positions X+2 and X+4 (Fig. 9e-f), in agreement with the hydrophobicity profiles (Fig. 8e-f).

9. Side chain – main chain hydrogen bonds

Side chains of polar and negatively charged residues (Asn, Asp, Gln, Glu, Ser, Thr and His) can

be involved in the formation of closed-loop conformations through side chain – main chain hydrogen bonds[101]. They have the capability to form C10 to C17 membered ring conformations, through H-bonding of the side chain oxygen or nitrogen to the backbone polar groups of residues located up to four positions upstream or downstream. We thus screened the HXH motifs in search of H-bonds between polar or negatively charged side chains of the linker X and polar groups of neighbour backbone (Table III). As previously observed[101], side chains acting as donor can form a H-bond only with the carbonyl oxygen of upstream residues whereas side chains acting as acceptor can interact only with amide nitrogen of downstream residues.

Gln and Glu residues located at position X are seldom involved in side chain – main chain H-bonds (3%), although these residues have usually the capability to be involved in such H-bonds[101]. For Asn, Asp, Ser, Thr and His, the H-bonding pattern depends upon the conformation of the linker (Table III). The percent of polar side chains involved in H-bonds raises from less than 15% for α_L and α_1 motifs up to 80% for β_2 conformation motifs.

For α_1 motifs, only Thr or Ser at position X can be involved in H-bond interactions. These H-bonds involve only carbonyl groups, mainly at position X-4 (9 cases out of 10). The H-bond linking the Ser/Thr side chain to the X-4 carbonyl is typically observed for Ser or Thr residues located within α -helices[101]. Such an interaction involving a Thr side chain and the carbonyl group at position X-4 is shown in Fig. 5a. When they are present at position X of α_1 motifs, Asp, Asn and His are not involved in H-bonds.

Ser and Thr are not favoured at position X of the α_2 motifs (Table II). However, when present, most of them (80%) are involved in H-bonds with either the X-4 or X-3 carbonyl groups. The interaction between the Ser/Thr side chains and the carbonyl groups at position X-3 is typical of helix C-terminus[101]. In addition, Asn and His, which have high propensities for position X of α_2 motifs, can interact with the carbonyl group of residue X-4. These latter interactions are also typical of the C-terminal end of α -helices[101]. Two thirds of the Asn residues at position X of α_2 motifs are involved in such H-bond interactions. An example of this interaction is given in Fig. 5b. Asp, which cannot form this H-bond, is seldom present in the α_2 conformation ($p = 0.65$). No example of interaction of Asn or Asp side chains with the amide group at position X+3, typical of N-capping, is observed.

In β_1 motifs, polar side chains present a different H-bonding pattern and can be involved in interaction with either the carbonyl groups of residue X-4 (Asn, Ser, Thr and His) or with the amide groups of residues X+2 or X+3 (Asn, Asp,

Gln, Ser and His). When present, most Ser side chains (77%) are involved in H-bonds with the amide group at position X+3. An example of this interaction is given in Fig. 5c. Thr is seldom present at position X of the β_1 conformation but two out of the three cases are involved in H-bonds with the carbonyl groups at X-4. Either in β_{1+} or β_{1-} motifs, about 80% of Asn and Asp side chains are involved in H-bonds. Interestingly, the presence of proline favours H-bonding of Asn with carbonyl groups (10 cases out of 16) whereas its absence favours H-bonding with amide groups (11 cases out of 15) (Table IV). Asp is involved in H-bonds with amide nitrogens, either at position X+2 or X+3. About 30% of His are involved in H-bond with carbonyl groups at X-4. Five cases out of six for this interaction are observed for β_{1+} motifs (Table IV).

Either in the presence or not of proline, polar residues Asn, Asp, Ser, Thr and His are present in about 70% of the β_2 motifs and more than 90% of their side chains are involved in H-bonds (Tables III and IV). These H-bonds involve almost exclusively the NH groups of residue X+2 or X+3. Only 2 out of 75 examples of H-bonds involve carbonyl groups. Most H-bonds involving the amide group at position X+2 are observed for β_{2+} motifs (7 cases out of 11), in spite of the limited number of these motifs (Table IV). This effect of proline is especially marked for Asp which can interact with amide groups either at position X+2 or X+3 (4 and 7 cases, respectively) in β_{2+} motifs whereas only H-bonds with amide groups at position X+3 are observed in β_2 motifs (18 cases).

DISCUSSION

1. The helix-X-helix motif

One of the difficulties in analysing secondary structures of proteins relies on the definition of secondary structure elements, in particular when the attention is focused on the limits of these elements. Different algorithms, based on H-bond pattern, C α geometry, backbone dihedral angles or a combination of different criteria have been developed[49, 102-106]. The analysis of the HXH motif is dependent upon the definition of the secondary structure elements. For example, in their analysis of α - α linking motifs, Engel and DeGrado[47] used a broad phi/psi based definition of α -helix with psi ranging up to 45° and thus could not observe the α_1 and α_2 conformations that we have determined.

In this article, we relied on the DSSP definition of secondary structure elements[49]. DSSP is based on the detection of H-bonds and was developed in order to define secondary structure elements in which not all possible H-bonds are formed, for example in bended or curved helices[49]. In addition, we added a phi/psi filter to insure that the five residues surrounding the linker were in a helical conformation and to remove false positive

motifs in which one of the helices was not correctly defined. This filter suppressed also potential problems with the definition of the helix termini. As a matter of fact, in 3% of the HXH motifs initially found with DSSP, the C-terminus of helix 1 or the N-terminus of helix 2 were out of the α region. These structures corresponded to helices linked by two residues and were erroneously assigned as HXH motifs by DSSP.

Analysis of the dihedral angles of residues located in the middle of α -helices (Fig. 6) indicates that the β_M conformation is the only one for which a significant number of reverse false positives can be found. Although the reasons are not clear, DSSP does not cope well with the extreme distortion of the protein chain observed in these motifs. This is not the case for the other conformations. Only a few cases are observed for the β_1 conformation and none for the α_L and β_2 conformations. The dihedral angles of the α_1 and α_{2+} motifs, albeit included in the additional allowed α -helical region, can only marginally be accessed by residues located in the middle of contiguous helices (Fig. 6). Finally, detailed analysis of the α_2 motifs indicate that, albeit the dihedral angles of the linker overlap the α -helix area, they are clearly distinct from kinks included in contiguous α -helices (Fig. 7) and correctly assigned as HXH motifs by DSSP.

The precise determination of the H-bond pattern of the HXH motif was carried out with HBPLUS (Table I). This motif is characterized by the lack of NH(*i*) to CO(*i*-4) hydrogen bonds between the three residues downstream and upstream the linker residue. The linker residue X is usually involved in H-bond interactions both with residue X-4 and residue X+4 but there is complete disruption of the helical pattern between residues N1 - N3 of helix 2 and residues C3 - C1 of helix 1.

The limited number of motifs found in PDB_25 led us to develop an alternative strategy to build a HXH database (Fig. 1). Although this may introduce some bias in the quantitative results, several lines of evidence strongly suggest that the procedure should not significantly affect the general conclusions of this study: (1) sequence diversity is high, with 94% of the sequence pairs having a sequence identity $\leq 27\%$ (to be compared to 96% for the PDB_25 set); (2) the relative weights of the six conformations are very similar in both data sets; (3) sequence properties can be rationalized by energetic considerations, in relation with the specific structural properties of these motifs.

A striking property of the HXH motifs is their solvent exposure (Fig. 9). Disruption of the helical H-bonding pattern makes several polar groups of the protein backbone free (Table I). The high solvent exposure of most motifs is probably related to this property and required for energetic reasons. The α_1 and α_{2inf} motifs, for which alternative H-bonds are the most frequent, are the most buried ones (Fig. 9). Other motifs are solvent exposed. However, the linker residue, in spite of its usual

polar character, is more buried than its neighbours in most conformations (Fig. 6 and 7). This may be related to its H-bonding properties. First, the amide and carbonyl groups of this residue are involved in main chain – main chain H-bonds with residues X-4 and X+4 (Table I). This makes easier the burying of the backbone at this position. Second, polar side chains at position X are involved in H-bonds with neighbour polar groups of the protein backbone and may contribute to stabilize the kink.

The propensities of Asn and Asp at position X are very consistent with the H-bonding patterns. Both Asn and Asp have high propensities in β_1 and β_2 motifs where their side chain can interact with downstream amide groups, indicating that these N-capping interactions are stabilizing. Similarly, the high propensity of Asn in α_2 motifs can be related to C-capping interactions with upstream carbonyl groups. His has also a high propensity at position X in motifs (α_2 , β_1 and β_2) where it can form H-bonds. Possibility of H-bonds between Ser/Thr and carbonyl groups in α_1 and α_2 motifs is not related with an increased propensity of these residues ($p \leq 1$). However, these H-bonds are concurrent of the main chain – main chain H-bonds involving the X-4 and X-3 oxygens (Table I) and thus may not be stabilizing. On the other hand, favourable propensity for Ser is observed when it can be involved in interactions with amide groups ($p=1.6$ and 3.3 for the β_1 and β_2 conformations, respectively), indicating that these N-capping interactions stabilize these motifs.

2. Proline-induced kinks

Proline is rigorously excluded from the linker position X and the preceding residues in HXH motifs but is found from position X+1 to X+3 in almost forty percent of these motifs, enlightening the role of this residue as helix breaker. Proline has an average propensity of 1.4 at position X+2 and is observed in 6% of the motifs mainly for the α_1 and α_2 conformations. Most proline induced kinks, however, correspond either to a α_1 motif with proline at position X+3 (12% of the 837 motifs) or to a β_1 or β_2 motif, with proline located at position X+1 (20% of the 837 motifs).

Because of the bulkiness of the pyrrolidine ring, the backbone of the residue preceding a proline can adopt only a limited range of dihedral conformations. This includes a narrow range in the α region and two subsets of the β conformation corresponding to the β_1 and β_2 regions defined in this study[87, 88, 100]. When proline residues are found in the middle of α -helices, steric constraints lead to a kink in the helix to avoid a clash between the C δ atom of Pro at position *i* and the carbonyl oxygen at position *i*-4. These kinks have bend angles in the 20-30° range, with an average value of 26° [107], similar to the average bend angle of the α_1 motifs. Analysis of the dihedral angles of the residues located around the proline reveals that,

when proline is included in a contiguous helix, the distortion of the backbone is shared by the residues located two positions ($\phi = -80^\circ \pm 12^\circ$, $\psi = -30^\circ \pm 11^\circ$) and three positions upstream the proline ($\phi = -77^\circ \pm 13^\circ$, $\psi = -35^\circ \pm 11^\circ$). On the other hand, the α_1 motifs correspond to a large distortion of the residue located three positions upstream the proline ($\phi = -122^\circ \pm 9^\circ$, $\psi = -56^\circ \pm 7^\circ$) while the residue located two positions upstream is not affected ($\phi = -64^\circ \pm 6^\circ$, $\psi = -52^\circ \pm 7^\circ$). However, it is worth to note that these differences in the dihedral angles do not induce any significant difference in the amino acid distribution or hydrophobicity pattern of the residues surrounding the proline.

When proline is present in β_1 or β_2 motifs, the linker conformation corresponds to one of the two conformations accessible to pre-Pro residues in the β area [87, 88, 100] and proline is located at the next position. The linker is mainly in the β_1 conformation (78% of the motifs with proline at X+1) which allows a dramatic change in the helix orientation with bend angles larger than 90° . The β_1 conformation corresponds to the well described Pro C-capping motif [108]. This conformation allows a stabilizing electrostatic interaction of residues X and X+1 with the helix dipole. The high propensity of His and aromatic amino acids at position X for this motif in the presence of Pro (4.4, 1.9 and 3.0 for His, Phe and Tyr, respectively) is consistent with stabilization of the Pro C-capping motif by interaction of these rings with the carbonyl group located 4 residues upstream [108]. The analysis of the H-bonds between His at the linker position and the carbonyl group at position X-4 (Table IV) provides an additional evidence of such interaction.

Analysis of the amino acid distributions and of the hydrophobicity patterns (Table II and Fig. 8) strongly suggests that position-specific scoring matrices could be used to predict the backbone conformation of the pre-Pro residues in proline-containing sequences (see Materials and Methods). The matrix for pre-Pro residues in the α conformation (α matrix) was built from the sequences of the α_1 motifs and of Pro-containing helices in PDB_25 (89 and 85 sequences, respectively). The matrix for pre-Pro residues in the β conformation (β matrix) was built from the sequences of the β_1 and β_2 motifs (121 and 33 sequences, respectively). The limited number of β_2 motifs did not allow considering them separately. In either case, the window ranged from 5 residues upstream to 3 residues downstream the proline. Predictions were based on a tenfold, cross-correlation procedure. The Q-score matrix is shown in Table V and corresponds to an average accuracy of 0.81 ± 0.06 . Indeed, up to 85% of sequences with the pre-Pro residue in the β region were predicted successfully, clearly indicating the usefulness of these position-specific scoring matrices for prediction purpose.

3. Glycine-induced kinks

Thirteen percent of the HXH motifs have a glycine at the linker position. This corresponds to an average propensity of 1.7. However, Gly is seldom observed in α_1 , α_2 , β_1 and β_2 motifs and has a propensity $p < 1$ at position X of these conformations (Table II). In most cases, the dihedral angles of glycine linkers are either in the β_M or in the α_L conformation, characterized by positive phi values. In addition to its high propensity for the Ccap position of α -helices, glycine is known to have a high propensity for the C' position [94, 96, 98]. However, a favourable propensity for Gly at position X+1, corresponding to the C' position of helix 1, is not observed in any of the conformations (the value of 1.4 in α_L motif is not significant, due to the limited number of observations).

Glycine is characterized by the absence of side chain that allows its backbone dihedral angles to experience a much broader range than for other residues [87, 88, 100]. Its phi dihedral angle can be positive and glycine can access either the α_L or the β_M regions of the Ramachandran plot, corresponding to mirror regions of the α -helix or of the β -strand, respectively. In particular, the β_M region is very specific of Gly residues [87].

Termination of α -helices by a glycine residue in the α_L conformation at the Ccap position is commonly observed in proteins [91, 92, 97]. This conformation allows the formation of the Schellman motif [91, 92], which involves two main chain – main chain hydrogen bonds joining NH(C') to CO(C4) and NH(Ccap) to CO(C3). The propensity of Gly in α_L motifs, 7.1, correlates well with the propensity of 8 observed for Gly at the Ccap position when this one is in the α_L conformation [97]. However, α_L motifs represent only 20% of the HXH motifs with a Gly at the linker position. This is in agreement with the limited number of α_L -terminated helices followed by a second helix initiating at the next position (2%) [97].

In most cases (80%), when a second helix initiates after a glycine, the dihedral angles of the glycine linker are in the β_M conformation. This dihedral conformation has been recently described as a glycine specific conformation [88]. To the best of our knowledge, this specific motif, in which two α -helices are linked by a Gly residue in the β_M conformation, has not been described yet. This conformation allows large bend angles ($124 \pm 18^\circ$) and a clockwise wobble rotation. It is the only conformation allowing such wobble rotation, reversed as compared to the other conformations. The weight of the β_M motifs is underestimated by DSSP (Fig. 6). Nevertheless, they represent 9% of all the HXH motifs in our database, highlighting their structural importance.

Because of the very limited number of Gly-containing α_{L+} (20 cases), position-specific scoring matrices could not be developed for quantitative

prediction of the linker conformation. However, comparison of the amino acid propensities (Table II) and hydrophobicity profiles (Fig. 8) of α_L and β_M motifs clearly shows the reversed hydrophobicity of the helix following the glycine linker, especially at positions X+2 and X+4, and suggests rules of the thumb to differentiate these motifs. In particular, in β_M motifs, the residue located at position G+4 is hydrophobic or Ala and stabilizes the motif by interaction with hydrophobic residues in helix 1 (residues G-5 and/or G-4). Alanine, with its small size, favours the folding back of the protein backbone (Fig. 5f) and is usually observed for high bend angles, with an average θ_b of 134° . This position is polar and solvent exposed in α_L motifs. It is also interesting to note the reversed polarity of the position preceding the glycine. Although partly exposed on the protein surface in both cases, position G-1 has a preference for hydrophobic residues in β_M motifs and for polar residues in α_{L+} motifs. In particular, Ser is found at position G-1 in 30% of the α_{L+} motifs (propensity of 4.6).

4. Non-Gly, non-Pro motifs

In spite of the high number of proline or glycine induced kinks, about half the HXH motifs involve neither proline or glycine residues. Among these motifs, the most frequent one is α_2 (40%) and the less frequent one is α_L (5%). The α_1 , β_1 and β_2 motifs have a weight in the 15-20% range. The α_2 motif is the only one without a characteristic proline or glycine residue. It has however high propensities for Asn and His at position X (Table II). These propensities are dependent upon the bend angle and these two residues represent 10% and 36% of linker residues when θ_b is lower and higher than 55° , respectively. These residues may be involved in H-bond interactions with the carbonyl group at position X-4 (Table III). These H-bonds are typical of C-capping stabilization[97]. The increased propensity of His and Asn in α_{2sup} motifs indicates that the C-capping H-bonds involving these residues stabilize the α_2 motifs with high bend angles.

Either in the β_1 or β_2 conformation, the absence of proline at position X+1 is correlated with an increased propensity of Asn, Asp and Ser residues, with the only exception of Asp for which propensity is not significantly altered by the presence of Pro (Table IV). These three residues represent 38% and 55% of the β_{1+} and β_{2+} linkers, respectively, and 63% and 75% of β_{1-} and β_{2-} linkers. The presence of proline does not alter the percent of these residues involved in H-bonds (80 and 95% for the β_1 and β_2 motifs, respectively). However, it alters the H-bond patterns (Table IV). In particular, the preference for H-bonds involving the amide group at position X+3 is more marked in β_{1-} and β_{2-} motifs (80 and 92%, respectively) than in β_{1+} and β_{2+} motifs (56 and 71%, respectively) (Table IV). These H-bonds are typical of helix N-

terminus[69, 101, 109-111] and appear to stabilize β_1 and β_2 motifs, especially in the absence of proline.

Asn has unique properties as helix breaker. Its propensity at the linker position is 5.6, 4.9 and 5.1 for the α_{2sup} , non-Pro β_1 and non-Pro β_2 motifs, respectively. These high propensities can be related to the H-bond pattern of its side chain. It may be involved in C-capping stabilizing interactions in the α_2 conformation or in N-capping stabilizing interactions in the β_1 or β_2 conformations. As the Asp side chain can only be involved in H-bond interaction with upstream amides, it is seldom observed in the α_2 motifs but displays high propensity in the β_1 and β_2 motifs. On the other hand, although Ser can be involved in C-capping interactions in α_2 motifs, it has a low propensity (0.75) for this motif and is mostly observed in β_1 and β_2 motifs, in which it is involved in N-capping interactions.

Comparison of the hydrophobicity patterns of the α_{2sup} , β_1 - and β_2 - motifs (Fig. 8c) enlightens the general properties of HXH motifs when a large bend angle is observed between two α -helices in the absence of proline or glycine breakers. In the three cases, the linker has a marked polar character and is located within a polar stretch. Position X-3 is buried and helix 1 ends at its solvent exposed side. In β_1 and β_2 motifs, helix 2 initiates from the solvent exposed side. In α_{2sup} motifs, the solvent exposed side of helix 2 starts at position X+2. However, position X+1 displays a preference for either polar residues or alanine. This leads to a characteristic 5 residue long polar stretch (which may include Ala at position X+1), which appears as a hallmark of the α_{2sup} motifs.

The α_{2sup} and β_2 motifs correspond to kinks with bend angles in the same 50 - 100° range, but with different wobble values (Fig. 4b). A tool able to discriminate between these two conformations should be very useful both for protein modelling and protein design. We thus tested the capability of the position-specific scoring matrices to differentiate them. Accuracy of the prediction reached 85% (Table V), indicating that this method may be used to estimate the wobble motion of kinks with bend angle in the 50 - 100° range.

CONCLUSIONS

The present work describes the first systematic description of a structural motif characterized by two helices linked by a single residue. This motif is commonly found in soluble proteins and about 10% of the proteins possess such a helix-X-helix structure. Most importantly, only a few backbone conformations are allowed at the linker position, leading to a classification of these kinks in six classes with characteristic amino acid distributions. The α_1 conformation is mainly limited to small bend angles ($\theta_b < 30^\circ$), and displays a high propensity for proline at position N3 of helix 2.

Larger amplitude kinks are usually located either at Gly, Ser, Asp or Asn residues or at positions preceding proline residues. Bend angles larger than 90° can be obtained through the β_M or the β_L conformations and are usually related either to the presence of a glycine at the linker position or to the presence of a proline at the following position. It is noteworthy that, when a glycine residue is located between two α -helices, the unconventional β_M conformation is more frequent than the α_L conformation.

The analysis of the HXH motifs developed here should provide useful information both for molecular modelling and for *de novo* design of protein structures. Among possible applications in the modelling field, it should contribute (1) to determine correctly the location of the putative kink between two helices when secondary structure predictions lead to an unrealistic long helix and (2) to determine the relative orientation of the two helices when the kink position is determined. In the protein design field, it will be possible to test the compatibility of a sequence with specific conformations, especially for Pro-induced kinks. The position-specific scoring matrices that we developed should be particularly useful for this purpose.

ACKNOWLEDGMENTS: We thank NEC Computers Services SARL (Angers, FRANCE) for the kind availability of a multiprocessor server. We thank D. Thybert for the clustering algorithm. J. D. was supported by fellowships from INSERM-Région des Pays-de-la-Loire and from the Association pour la Recherche sur le Cancer (ARC). J. R. is supported by a fellowship from CNRS.

REFERENCES

1. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247(4):536-540.
2. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. *Structure* 1997;5(8):1093-1108.
3. Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry* 1974;13(2):222-245.
4. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* 1974;13(2):211-222.
5. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* 1978;47:45-148.
6. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 1978;120(1):97-120.
7. Albrecht M, Tosatto SC, Lengauer T, Valle G. Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng* 2003;16(7):459-462.
8. Ouali M, King RD. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci* 2000;9(6):1162-1176.
9. Kabsch W, Sander C. How good are predictions of protein secondary structure? *FEBS Lett* 1983;155(2):179-182.
10. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 1999;34(4):508-519.
11. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000;40(3):502-511.
12. Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 1997;27(3):329-335.
13. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A* 1993;90(16):7558-7562.
14. Salamov AA, Solovyev VV. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J Mol Biol* 1995;247(1):11-15.
15. Wilson CL, Boardman PE, Doig AJ, Hubbard SJ. Improved prediction for N-termini of alpha-helices using empirical information. *Proteins* 2004;57(2):322-330.
16. Brazhnikov EV, Efimov AV. [Structure of alpha-spiral hairpins with short connections in globular proteins]. *Mol Biol (Mosk)* 2001;35(1):100-108.
17. Engel DE, DeGrado WF. Alpha-alpha linking motifs and interhelical orientations. *Proteins* 2005;61(2):325-337.
18. Lahr SJ, Engel DE, Stayrook SE, Maglio O, North B, Geremia S, Lombardi A, DeGrado WF. Analysis and design of turns in alpha-helical hairpins. *J Mol Biol* 2005;346(5):1441-1454.
19. Hobohm U, Sander C. Enlarged representative set of protein structures. *Protein Sci* 1994;3(3):522-524.
20. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22(12):2577-2637.
21. Kahn PC. Defining the axis of a helix. *Computers Chem* 1989;13:185-189.
22. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol* 1984;179(1):125-142.
23. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;157(1):105-132.
24. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci U S A* 1981;78(6):3824-3828.
25. Hubbard SJ, Beynon RJ, Thornton JM. Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng* 1998;11(5):349-359.
26. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 1971;55(3):379-400.
27. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol* 1994;238(5):777-793.
28. Heine A, Canaves JM, von Delft F, Brinen LS, Dai X, Deacon AM, Elsliger MA, Eshaghi S, Floyd R, Godzik A, Grittini C, Grzechnik SK, Guda C, Jaroszewski L, Karlak C, Klock HE, Koesema E, Kovarik JS, Kreuzsch A, Kuhn P, Lesley SA, McMullan D, McPhillips TM, Miller MA, Miller MD, Morse A, Moy K, Ouyang J, Page R, Robb A, Rodrigues K, Schwarzenbacher R, Selby TL, Spraggon G, Stevens RC, van den Bedem H, Velasquez J, Vincent J, Wang X, West B, Wolf G, Hodgson KO, Wooley J, Wilson IA. Crystal structure of O-acetylserine sulfhydrylase (TM0665) from *Thermotoga maritima* at 1.8 Å resolution. *Proteins* 2004;56(2):387-391.
29. Johnson KA, Angelucci F, Bellelli A, Herve M, Fontaine J, Tsernoglou D, Capron A, Trottein F, Brunori M. Crystal structure of the 28 kDa glutathione S-transferase from *Schistosoma haematobium*. *Biochemistry* 2003;42(34):10084-10094.
30. Polekhina G, Board PG, Gali RR, Rossjohn J, Parker MW. Molecular basis of glutathione synthetase deficiency and a rare gene permutation event. *Embo J* 1999;18(12):3204-3213.

31. Miller DJ, Ouellette N, Evdokimova E, Savchenko A, Edwards A, Anderson WF. Crystal complexes of a predicted S-adenosylmethionine-dependent methyltransferase reveal a typical AdoMet binding domain and a substrate recognition domain. *Protein Sci* 2003;12(7):1432-1442.
32. Zhong W, Alexeev D, Harvey I, Guo M, Hunter DJ, Zhu H, Campopiano DJ, Sadler PJ. Assembly of an oxo-zirconium(IV) cluster in a protein cleft. *Angew Chem Int Ed Engl* 2004;43(44):5914-5918.
33. Wilson CL, Hubbard SJ, Doig AJ. A critical assessment of the secondary structure alpha-helices and their termini in proteins. *Protein Eng* 2002;15(7):545-554.
34. Ramachandran GN, Venkatachalam CM, Krimm S. Stereochemical criteria for polypeptide and protein chain conformations. 3. Helical and hydrogen-bonded polypeptide chains. *Biophys J* 1966;6(6):849-872.
35. Karplus PA. Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci* 1996;5(7):1406-1420.
36. Ho BK, Brasseur R. The Ramachandran plots of glycine and pre-proline. *BMC Struct Biol* 2005;5:14.
37. Lovell SC, Davis IW, Arendall WB, 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by C α geometry: phi,psi and C β deviation. *Proteins* 2003;50(3):437-450.
38. Adzhubei AA, Sternberg MJ. Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol* 1993;229(2):472-493.
39. Cubellis MV, Caillez F, Blundell TL, Lovell SC. Properties of polyproline II, a secondary structure element implicated in protein-protein interactions. *Proteins* 2005;58(4):880-892.
40. Schellman C. The α L-conformation at the ends of helices. in *Protein Folding*. Amsterdam: Jaenicke, R.; 1980. 53-61 p.
41. Aurora R, Srinivasan R, Rose GD. Rules for alpha-helix termination by glycine. *Science* 1994;264(5162):1126-1130.
42. Cubellis MV, Cailliez F, Lovell SC. Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinformatics* 2005;6 Suppl 4:S8.
43. Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of alpha helices. *Science* 1988;240(4859):1648-1652.
44. Kumar S, Bansal M. Dissecting alpha-helices: position-specific analysis of alpha-helices in globular proteins. *Proteins* 1998;31(4):460-476.
45. Engel DE, DeGrado WF. Amino acid propensities are position-dependent throughout the length of alpha-helices. *J Mol Biol* 2004;337(5):1195-1205.
46. Gunasekaran K, Nagarajaram HA, Ramakrishnan C, Balaram P. Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals. *J Mol Biol* 1998;275(5):917-932.
47. Kumar S, Bansal M. Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys J* 1998;75(4):1935-1944.
48. Sharma V, Sharma S, Hoener zu Bentrup K, McKinney JD, Russell DG, Jacobs WR, Jr., Sacchettini JC. Structure of isocitrate lyase, a persistence factor of *Mycobacterium tuberculosis*. *Nat Struct Biol* 2000;7(8):663-668.
49. Anderson RJ, Weng Z, Campbell RK, Jiang X. Main-chain conformational tendencies of amino acids. *Proteins* 2005;60(4):679-689.
50. Eswar N, Ramakrishnan C. Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. *Protein Eng* 2000;13(4):227-238.
51. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins* 1995;23(4):566-579.
52. Labesse G, Colloc'h N, Pothier J, Mornon JP. P-SEA: a new efficient assignment of secondary structure from C α trace of proteins. *Comput Appl Biosci* 1997;13(3):291-295.
53. Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. *J Mol Biol* 1977;114(2):181-239.
54. Martin J, Letellier G, Marin A, Taly JF, de Brevern AG, Gibrat JF. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct Biol* 2005;5:17.
55. Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins* 1988;3(2):71-84.
56. Barlow DJ, Thornton JM. Helix geometry in proteins. *J Mol Biol* 1988;201(3):601-619.
57. Prieto J, Serrano L. C-capping and helix stability: the Pro C-capping motif. *J Mol Biol* 1997;274(2):276-288.
58. Aurora R, Rose GD. Helix capping. *Protein Sci* 1998;7(1):21-38.
59. Doig AJ, MacArthur MW, Stapley BJ, Thornton JM. Structures of N-termini of helices in proteins. *Protein Sci* 1997;6(1):147-155.
60. Eswar N, Ramakrishnan C. Secondary structures without backbone: an analysis of backbone mimicry by polar side chains in protein structures. *Protein Eng* 1999;12(6):447-455.

LEGENDS TO FIGURES

Fig. 1: Flow chart used to create the non-redundant helix-X-helix database. PDB_90 corresponds to the March 2006 release of the PDB selection with a 90% threshold[52].

Fig. 2: Definitions of the bend angle θ_b and of the wobble angle θ_w . (a) and (b) correspond to two perpendicular views of a schematic HXH motif, parallel and perpendicular to the axis of H1, respectively. The C α atoms of the linker X and of residue C4 of H1 are shown as light grey and dark grey spheres, respectively.

Fig. 3: Distribution of the bend angles between two helices separated by a linker of one residue (black bars), two residues (white bars) and three residues (grey bars). N represent the number of observations in a subset of PDB_25 containing 1200 protein chains.

Fig. 4: (a) Ramachandran plot of the linker residue X and (b) distribution of the bend and wobble angles between the two helices, for the 837 helix-X-helix motifs of the database. The color code indicates the HXH conformation (α_1 : dark blue; α_2 : sky blue; β_1 : violet; β_2 : pink; α_L : green; β_M : red)

Fig. 5: Typical HXH motifs for each linker conformation. (a) α_1 motif: residues 136-153 of 1O58 chain D, X = Thr-145, $\theta_b = 33^\circ$, $\theta_w = 128^\circ$; (b) α_2 motif: residues 23-57 of 1VLU chain A, X = Asn-40, $\theta_b = 67^\circ$, $\theta_w = 99^\circ$; (c) β_1 motif: residues 183-206 of 1OE8 chain A, X = Ser-196, $\theta_b = 107^\circ$, $\theta_w = 109^\circ$; (d) β_2 motif: residues 69-91 of 2HGS chain A, X = Asn-82, $\theta_b = 90^\circ$, $\theta_w = 137^\circ$; (e) α_L motif: residues 205-228 of 1M6CY chain A, X = Arg-217, $\theta_b = 53^\circ$, $\theta_w = 1^\circ$; (f) β_M motif: residues 140-168 of 1XC1 chain A, X = Gly-154, $\theta_b = 134^\circ$, $\theta_w = -82^\circ$. The helix-X-motif is shown as a green ribbon. The axis of helix 1 is vertical. The side chain of the linker residues and prolines, when present, are shown as sticks. In (e), the C α of Gly at position X is shown as a sphere and the side chain of Ala at position X+4 is shown as a stick. Dashed lines represent H-bonds either between polar linker side chains and the protein backbone (a-d) or typical of the Schellman motif (e). Polar atoms involved in H-bonds are shown as spheres.

Fig. 6: Ramachandran plot of the residues located from position N6 to position C6 of α -helices. The analysis was carried out on the PDB_90S subset.

Fig. 7: Typical translation of the helix axis in α_2 motifs. The HXH motif (shown as a ribbon) corresponds to residues 140-170 of 5CSM chain A, X = Phe-160, $\theta_b = 10^\circ$, $\theta_w = 61^\circ$. The phi and psi angles of the linker are -72° and -10° , respectively.

Fig. 8: Average hydrophobicity index H from position X-4 to position X+4 for the α_1 (a), α_2 (b), β_1 (c), β_2 (d), α_L (e) and β_M (f) motifs. Open bars represent the α_{1+} motifs with proline at position X+3, the β_{1+} and β_{2+} motifs with proline at position X+1, the α_{L+} motifs with glycine at position X and the α_{2inf} motifs with $\theta_b < 55^\circ$. Closed bars represent the α_{1-} motifs without proline at position X+3, the β_{1-} and β_{2-} motifs without proline at position X+1, the α_{L-} motifs without glycine at position X and the α_{2sup} motifs with $\theta_b > 55^\circ$. The 1F8M structure with X=Gln was removed from the β_M data set.

Fig. 9: Average accessible solvent area (ASA) from position X-4 to position X+4 for the α_1 (a), α_2 (b), β_1 (c), β_2 (d), α_L (e) and β_M (f) motifs. Open bars represent the α_{1+} motifs with proline at position X+3, the β_{1+} and β_{2+} motifs with proline at position X+1, the α_{L+} motifs with glycine at position X and the α_{2inf} motifs with $\theta_b < 55^\circ$. Closed bars represent the α_{1-} motifs without proline at position X+3, the β_{1-} and β_{2-} motifs without proline at position X+1, the α_{L-} motifs without glycine at position X and the α_{2sup} motifs with $\theta_b > 55^\circ$.

TABLE I

Main chain – main chain hydrogen bonds within the HXH motif ^a

Conformation	α-Helical bonds NH(<i>i</i>) to CO(<i>i</i> -4)		N _{bonds}	Alternative H-bonds NH(<i>i</i>) to CO(<i>j</i>)		N _{bonds}
α ₁ (212)	X	X-4	194	X+1	X-4	10
	X+1	X-3	4	X+2	X-3	164
	X+2	X-2	0	X+3	X-2	12
	X+3	X-1	0	X+3	X	13
	X+4	X	177			
α ₂ (188)	X	X-4	135	X-1	X-4	9
	X+1	X-3	1	X	X-3	42
	X+2	X-2	2	X+1	X-2	130
	X+3	X-1	0	X+2	X-1	25
	X+4	X	157	X+3	X	26
β ₁ (218)	X	X-4	183	X-1	X-4	18
	X+1	X-3	0	X	X-3	27
	X+2	X-2	0	X+1	X-2	1
	X+3	X-1	0	X+3	X	43
	X+4	X	168			
β ₂ (104)	X	X-4	50	X-1	X-4	25
	X+1	X-3	0	X	X-3	44
	X+2	X-2	0	X+3	X	6
	X+3	X-1	0			
	X+4	X	96			
α _L (37)	X	X-4	24 ^b	X+1	X-4	14 ^b
	X+1	X-3	0	X	X-3	18 ^c
	X+2	X-2	0	Schellman motif ^d		9
	X+3	X-1	0	Partial motif ^e		23
	X+4	X	34			
β _M (78)	X	X-4	74	X-1	X-4	2
	X+1	X-3	0	X	X-3	1
	X+2	X-2	0	X+2	X-1	1
	X+3	X-1	0	X+3	X	8
	X+4	X	65			

^a The number N_{bonds} of H-bonds was determined with HBPLUS as described in Materials and Methods for the six conformations of the linker X. The numbers between brackets represent the number of motifs for each conformation.

^b For 10 of these H-bonds, interactions from NH(X) or NH(X+1) to CO(X-4) are equally probable.

^c For 10 of these H-bonds, interactions from NH(X) to CO(X-3) or CO(X-4) are equally probable.

^d Presence of both NH(X+1) to CO(X-4) and NH(X) to CO(X-3) H-bonds.

^e Presence of either NH(X+1) to CO(X-4) and NH(X) to CO(X-3) H-bonds.

TABLE II
Amino acid propensities^a

α_1	X-4	X-3	X-2	X-1	X	X+1	X+2	X+3	X+4
P	0.0	0.2	0.1	0.0	0.0	0.2	2.0	9.2	0.0
G	0.4	0.1	0.2	0.5	0.5	1.1	0.9	0.3	0.4
A	0.7	1.1	0.9	0.8	0.6	0.5	1.0	0.8	1.6
L	1.5	1.6	1.6	0.9	1.2	1.8	1.5	0.5	1.7
M	1.1	0.7	1.6	0.4	0.7	1.6	1.6	0.9	1.8
I	1.8	0.8	1.4	0.5	1.6	1.7	1.1	0.2	1.3
V	1.1	1.1	0.7	0.5	1.8	1.4	0.9	0.8	1.4
F	2.9	1.7	0.6	1.0	2.3	1.9	0.7	1.2	2.4
Y	2.6	2.1	1.2	0.7	1.4	2.4	0.7	1.1	0.7
W	2.6	2.0	0.7	0.3	0.3	1.7	3.0	1.0	0.7
H	1.0	1.7	1.0	1.7	1.9	0.4	1.0	1.3	0.4
K	0.4	0.7	1.8	1.1	0.5	0.8	0.9	0.9	0.5
R	0.4	0.8	1.5	0.9	0.9	0.5	1.1	0.5	0.7
E	0.8	1.3	1.2	1.6	1.0	1.1	0.4	0.3	1.2
Q	0.5	0.8	0.8	1.8	1.5	0.6	1.3	1.3	0.9
D	0.6	1.8	0.8	0.8	0.3	0.5	0.5	0.2	0.4
N	0.9	0.9	1.7	3.9	1.0	0.1	0.5	0.3	0.9
S	0.6	0.5	0.7	0.9	0.9	0.4	1.0	0.3	0.6
T	1.1	0.8	0.8	1.5	1.0	1.1	0.6	0.9	1.0
C	1.4	0.3	0.6	0.3	0.8	0.6	1.4	0.0	0.6

α_2	X-4	X-3	X-2	X-1	X	X+1	X+2	X+3	X+4
P	0.3	0.3	0.0	0.0	0.0	0.5	2.6	1.3	0.0
G	0.6	0.1	0.4	0.5	0.3	0.9	0.8	0.1	0.1
A	1.5	1.3	1.1	1.6	1.0	2.4	1.1	1.0	0.7
L	1.6	2.6	1.3	1.0	1.5	1.3	0.5	1.3	2.2
M	1.5	1.8	1.8	0.3	1.8	1.0	1.0	0.5	1.0
I	1.1	1.6	0.9	0.6	0.8	0.9	0.4	1.2	3.9
V	1.0	0.9	0.9	0.5	0.4	0.4	0.3	1.3	1.3
F	1.2	1.5	0.4	0.7	0.9	0.7	0.1	1.3	2.0
Y	0.7	0.7	0.6	0.0	1.6	0.9	0.1	1.0	0.7
W	1.9	0.4	0.4	0.4	0.0	1.5	0.0	1.5	1.9
H	0.9	0.2	0.5	0.9	2.8	0.2	0.7	1.4	1.4
K	0.7	0.8	1.4	1.6	1.5	1.6	1.6	1.2	0.5
R	1.3	2.4	2.2	2.3	1.0	2.1	0.6	0.7	1.0
E	0.7	0.6	2.2	2.1	0.6	0.7	2.2	1.8	0.4
Q	1.1	1.1	1.9	1.9	0.9	1.4	1.3	0.7	0.1
D	0.7	0.2	0.5	1.3	0.7	0.2	2.5	1.2	0.6
N	0.7	0.4	0.8	1.6	3.4	0.4	1.0	0.5	0.4
S	0.8	0.6	0.6	0.7	0.7	0.8	1.3	0.7	0.5
T	1.0	0.6	0.8	0.4	0.5	0.5	0.5	0.7	0.6
C	0.6	0.6	0.6	0.0	1.6	0.6	0.0	0.0	0.6

β_1	X-4	X-3	X-2	X-1	X	X+1	X+2	X+3	X+4
P	0.2	0.0	0.0	0.0	0.0	12.1	0.7	0.2	0.0
G	0.3	0.2	0.2	0.5	0.0	0.3	0.5	0.5	0.7
A	1.3	1.2	1.4	0.8	0.6	0.6	1.4	2.1	1.8
L	1.9	2.3	0.8	0.6	0.3	0.3	0.3	1.7	1.5
M	1.1	2.2	0.2	1.3	0.4	0.0	0.9	0.4	0.9
I	1.1	1.1	0.8	0.5	0.3	0.4	0.4	1.4	0.6
V	1.3	0.5	0.7	0.4	0.1	0.4	0.3	1.4	0.6
F	2.1	3.5	0.7	0.3	1.6	0.1	0.5	1.2	3.5
Y	1.2	2.2	1.4	0.4	2.3	0.4	0.4	1.7	2.2
W	2.2	1.0	0.3	0.3	0.3	0.0	1.0	1.0	2.2
H	0.2	0.4	0.2	2.4	3.7	0.8	1.0	0.4	0.2
K	0.8	0.8	2.7	2.7	0.4	1.3	0.8	0.8	0.8
R	1.1	0.7	2.0	1.8	0.5	0.8	1.3	0.8	1.7
E	1.4	0.4	1.5	1.7	0.4	0.6	3.9	1.1	0.6
Q	1.0	1.2	1.0	1.7	1.1	0.7	2.0	1.2	1.4
D	0.8	0.4	0.6	0.8	3.5	0.2	2.4	0.6	0.2
N	0.4	0.5	1.0	1.0	4.3	0.0	0.5	0.3	0.3
S	0.5	0.8	1.2	0.9	1.6	0.5	0.8	0.3	0.4
T	0.5	0.3	0.8	1.6	0.2	0.5	0.5	0.9	0.8
C	0.3	2.5	0.8	0.5	1.9	0.0	0.0	0.0	0.3

β_2	X-4	X-3	X-2	X-1	X	X+1	X+2	X+3	X+4
P	0.2	0.0	0.0	0.0	0.0	6.9	0.6	0.0	0.0
G	0.8	0.1	0.4	0.8	0.3	0.3	0.5	0.5	0.4
A	1.9	0.5	1.3	1.9	0.2	0.7	1.8	1.2	1.2
L	0.6	2.9	2.9	0.4	0.4	0.3	0.6	1.7	1.1
M	0.5	2.7	1.8	1.4	0.9	0.5	0.0	1.4	0.5
I	1.0	2.0	1.0	0.5	0.5	0.2	1.0	1.4	0.7
V	0.4	1.7	0.8	0.5	0.0	0.0	0.8	1.9	0.7
F	0.2	1.7	1.0	0.7	0.0	0.2	0.7	4.6	0.0
Y	0.3	1.3	1.6	0.3	0.0	0.3	1.1	1.9	0.5
W	1.3	2.0	2.7	0.0	0.0	0.0	1.3	1.3	0.7
H	0.4	0.9	0.0	1.7	2.1	1.7	0.0	1.7	0.9
K	1.1	0.3	1.0	1.3	0.3	1.8	1.8	0.0	1.8
R	1.8	0.4	1.0	1.4	0.6	0.4	0.8	0.6	4.4
E	1.8	0.4	1.2	0.9	0.6	2.8	1.8	0.7	0.1
Q	0.8	1.0	0.5	1.3	0.5	0.8	1.0	0.0	5.4
D	1.5	0.2	0.3	0.2	5.0	1.5	2.4	0.0	0.5
N	2.0	0.7	0.2	2.6	4.2	0.2	0.4	0.7	0.4
S	0.5	0.9	1.1	1.8	3.4	0.8	0.8	0.3	0.6
T	1.1	0.7	0.0	1.5	0.5	0.3	0.7	0.7	0.2
C	1.1	0.0	1.7	0.0	0.6	0.0	0.0	0.6	0.6

α_L	X-4	X-3	X-2	X-1	X	X+1	X+2	X+3	X+4
P	0.6	0.0	0.0	0.0	0.0	1.2	1.2	0.6	0.0
G	1.1	0.0	0.0	0.7	7.2	1.4	0.4	0.7	0.4
A	0.3	0.7	1.3	1.0	0.0	1.3	0.7	1.0	0.7
L	1.5	1.2	0.9	0.9	0.3	0.9	2.5	0.6	1.9
M	2.6	2.6	0.0	0.0	1.3	2.6	2.6	0.0	1.3
I	1.9	2.4	1.0	1.0	0.0	1.4	1.0	0.5	0.0
V	0.8	0.8	1.1	0.8	0.4	1.5	0.8	0.0	0.8
F	1.4	0.0	0.7	0.0	0.0	2.0	1.4	0.0	0.7
Y	0.8	1.5	0.8	0.8	0.0	2.3	2.3	0.8	0.0
W	3.8	1.9	0.0	0.0	0.0	1.9	1.9	3.8	0.0
H	0.0	2.4	0.0	0.0	2.4	0.0	2.4	2.4	1.2
K	0.0	1.8	2.7	0.9	1.8	0.4	0.4	0.4	2.2
R	1.1	1.1	1.1	2.3	1.1	0.0	0.6	0.6	3.4
E	0.4	1.3	2.1	0.8	0.4	1.7	0.8	2.9	0.4
Q	0.7	1.4	0.7	0.7	1.4	0.0	0.0	3.6	0.7
D	2.4	0.9	0.9	0.9	0.5	0.0	1.4	0.5	0.5
N	1.2	1.2	1.8	1.2	0.6	0.0	0.6	0.6	1.8
S	0.4	0.9	0.4	2.6	0.0	0.4	1.3	1.3	0.9
T	0.9	0.0	0.9	2.3	0.0	0.5	0.9	0.5	1.4
C	0.0	0.0	1.6	0.0	0.0	0.0	1.6	1.6	0.0

β_M	X-4	X-3	X-2	X-1	X	X+1	X+2	X+3	X+4
P	0.6	0.0	0.0	0.0	0.0	2.5	0.6	0.0	0.0
G	0.3	0.0	0.3	0.0	13.1	0.2	0.5	0.5	0.5
A	2.4	1.3	1.3	1.3	0.0	0.3	1.3	2.2	2.7
L	1.2	1.2	0.9	2.4	0.0	0.9	0.3	0.9	2.1
M	0.0	0.6	1.8	0.0	0.0	1.2	0.6	1.2	1.2
I	1.6	0.7	0.5	1.4	0.0	0.5	0.2	0.9	1.4
V	1.4	0.4	0.7	1.8	0.0	1.4	0.7	0.5	1.8
F	1.9	1.3	0.6	2.9	0.0	1.0	0.3	1.6	3.2
Y	1.1	1.8	0.7	2.1	0.0	0.7	0.4	1.4	0.4
W	0.9	0.0	0.0	0.9	0.0	0.0	0.0	1.8	2.7
H	0.6	0.6	2.8	1.1	0.0	0.0	0.0	0.6	0.0
K	0.9	2.1	2.3	1.3	0.0	2.6	1.7	2.6	0.2
R	0.5	2.1	0.8	0.3	0.0	0.5	0.5	1.6	0.3
E	0.8	1.8	1.4	0.6	0.0	2.6	3.0	0.6	0.4
Q	0.7	1.7	2.7	0.7	0.3	2.1	0.7	0.3	0.3
D	0.4	1.8	1.6	0.2	0.0	0.4	3.4	0.7	0.0
N	0.3	0.0	0.3	0.3	0.0	0.3	1.2	1.8	0.0
S	0.6	0.8	0.4	0.2	0.0	0.8	1.0	0.0	0.6
T	1.3	0.4	1.1	0.7	0.0	0.7	0.7	0.4	0.7
C	0.8	0.0	0.0	1.5	0.0	0.0	0.8	0.8	0.8

^a The amino acid propensities were calculated from position X-4 to position X+4 for the six conformations of the linker X as described in Materials and Methods. Hallmark residues ($Z > 13$) are highlighted in yellow, favourable residues with $Z > 2.0$ and 2.6 are highlighted in light green and dark green, respectively, and disfavoured residues with $Z < -2.0$ and -2.6 are highlighted in pink and red, respectively.

TABLE III

Side chain – main chain H-bonds within the HXH motif^a

Residue	Donor/Acceptor	Conformation	N	N _{bonds}	Acceptor/donor
Asn	Oδ1	α ₁	9	0	
		Nδ2	28	17	17 O(X-4)
	Nδ2	β ₁	41	31	14 O(X-4), 4 N(X+2), 13 N(X+3)
		β ₂	19	19	1 O(X-3), 3 N(X+2), 15 N(X+3)
		α _L	1	0	
Asp	Oδ	α ₁	4	0	
		α ₂	7	0	
		β ₁	44	36	11 N(X+2), 25 N(X+3)
		β ₂	30	29	6 N(X+2), 23 N(X+3)
		α _L	2	2	2 N(X+3)
Gln	Oε1	α ₁	12	1	1 O(X-1)
		Nε2	6	1	1 N(X+2)
	Nε2	β ₁	9	1	1 N(X+2)
		β ₂	2	0	
		α _L	2	1	1 O(X-3)
Glu	Oε	α ₁	13	0	
		α ₂	7	0	
		β ₁	6	0	
		β ₂	4	1	1 N(X+3)
		α _L	1	0	
Ser	Oγ	α ₁	12	4	4 O(X-4)
		α ₂	8	7	3 O(X-4), 4 O(X-3)
		β ₁	22	19	1 O(X-4), 1 N(X+2), 17 N(X+3)
		β ₂	22	20	1 O(X-3), 19 N(X+3),
		α _L	0	0	
Thr	Oγ1	α ₁	13	6	5 O(X-4), 1 O(X-3)
		α ₂	6	5	2 O(X-4), 3 O(X-3)
		β ₁	3	2	2 O(X-4)
		β ₂	3	3	3 N(X+3)
		α _L	0	0	
His	Nδ1	α ₁	9	0	
		Nε2	12	5	5 O(X-4)
	Nε2	β ₁	18	6	5 O(X-4), 1 N(X+2)
		β ₂	5	3	2 N(X+2), 1 N(X+3)
		α _L	2	0	

^a The number N_{bonds} of H-bonds was determined with HBPLUS as described in Materials and Methods for the indicated atoms of polar side chains at position X of HXH motifs and neighbour polar groups of the protein backbone. O(X-*i*) indicates a H-bond with the carbonyl group at position X-*i*. N(X+*j*) indicates a H-bond with the amide group at position X+*j*. N represents the number of amino acids *a* at position X in each conformation of the linker.

TABLE IV

Effect of proline on side chain - main chain H-bonds^a

Residue	Conformation	Pro	N	<i>p</i>	N _{bonds}	Acceptor/donor
Asn	β_1	-	21	4.9	15	4 O(X-4), 1 N(X+2), 10 N(X+3)
		+	20	3.8	16	10 O(X-4), 1 N(X+2), 5 N(X+3)
	β_2	-	16	5.18	16	1 O(X-3), 2 N(X+2), 13 N(X+3)
		+	3	2.1	3	1 N(X+2), 2 N(X+3)
Asp	β_1	-	27	4.9	22	4 N(X+2), 18 N(X+3)
		+	17	2.5	14	4 N(X+2), 10 N(X+3)
	β_2	-	19	4.7	18	18 N(X+3)
		+	11	5.8	11	4 N(X+2), 7 N(X+3)
Ser	β_1	-	13	2.1	13	1 O(X-4), 12 N(X+3)
		+	9	1.2	6	1 N(X+2), 5 N(X+3)
	β_2	-	18	4.0	17	1 O(X-3), 16 N(X+3)
		+	4	1.9	3	3 N(X+3)
Thr	β_1	-	0	0.0	0	
		+	3	0.4	2	2 O(X-4)
	β_2	-	1	0.2	1	1 N(X+3)
		+	2	1.0	2	2 N(X+3)
His	β_1	-	6	2.7	1	1 O(X-4)
		+	12	4.4	5	4 O(X-4), 1 N(X+2)
	β_2	-	2	1.2	0	
		+	3	4.0	3	2 N(X+2), 1 N(X+3)

^aThe number N_{bonds} of H-bonds was determined with HBPLUS as described in Materials and Methods for the indicated polar side chains at position X and neighbour polar groups of the protein backbone in β_1 and β_2 motifs, as a function of the presence (+) or not (-) of proline at position X+1. Propensities *p* with *Z* > 2.0 are in bold font. N represents the number of amino acids *a* at position X in the considered subset. The total number of β_{1+} and β_{1-} motifs are 121 and 97, respectively. The numbers of β_{2+} and β_{2-} motifs are 33 and 71, respectively.

TABLE V
Conformational predictions ^a

	Conformations	Q^{obs}	Q^{pred}
Pro motifs ^b	α pre-Pro	0.78 ± 0.07	0.86 ± 0.06
	β pre-Pro	0.85 ± 0.07	0.77 ± 0.06
NonP, nonG motifs ^c	$\alpha_{2\text{sup}}$	0.87 ± 0.13	0.89 ± 0.07
	β_2	0.86 ± 0.10	0.85 ± 0.13

^a The Q-scores were determined by a 10-fold cross correlation procedure as described in Materials and Methods.

^b The position-specific scoring matrices included the five positions preceding the proline residue and the three positions following it. In the α pre-Pro conformation, proline could be in α_1 motifs or in contiguous helices. In the β pre-Pro conformation, proline could be either in β_1 or β_2 motifs.

^c The position-specific scoring matrices included the four positions surrounding the linker residue.

3.1.2 Conclusion

Ce travail décrit de manière systématique le motif structural caractérisé par deux hélices séparées par un résidu de jonction. Ce motif est communément trouvé dans les protéines solubles et environ 10% des protéines possèdent un motif HXH. De plus, le résidu de jonction n'admet qu'un nombre de conformations limitées conduisant à la classification de ces cassures en 6 motifs bien distincts.

Deux de ces zones correspondent à la déformation de la conformation α et sont situées dans la région avoisinante de la conformation α . Elles sont notées $\alpha 1$ et $\alpha 2$. La région $\alpha 1$ est décalée vers la gauche dans le diagramme de Ramachandran par rapport à la zone α canonique. Cette conformation est surtout caractérisée par la présence d'une proline à la position X+3 dans plus de 40% des motifs. Cette conformation correspond à une ouverture de l'hélice entraînant une apparence de renflement plus ou moins marqué en fonction de la valeur de l'angle phi. La région $\alpha 2$ est située à la limite supérieure de la région α . Cette région marque la transition entre la région α qu'elle chevauche légèrement et la région β du diagramme de ramachandran. La conformation $\alpha 2$ ne présente pas de surexpression des résidus proline ou glycine. Par contre, le résidu de jonction a une forte propension pour l'asparagine et l'histidine ($p > 2.8$).

Deux régions se situant sur les bords de la région β du diagramme de Ramachandran sont observées pour le résidu de jonction. La région $\beta 1$ est localisée à gauche de la région β (angles phi plus négatifs). La région $\beta 2$ est localisée à droite de la région β du diagramme de Ramachandran (angles phi moins négatifs). Ces deux régions présentent une forte propension de proline en position X+1 (32% des cas).

13% des cassures possèdent un angle phi positif. Ces cassures correspondent à 2 régions distinctes : αL et βM . La région βM correspond à la région miroir de la conformation β . Cette région est associée au résidu glycine. C'est la seule conformation qui présente un angle de giration inverse avec une rotation de la 2^{ème} hélice vers la droite. Enfin la région αL correspond à la région des hélices α gauches. Environ la moitié des structures αL et la quasi-totalité des structures βM ont un résidu Gly comme linker.

L'analyse des séquences montre une préférence de certains acides aminés à diverses positions en fonction du motif de cassure. Plus particulièrement, la proline est exclue de la position X et de la position X-1 et est fortement représentée en position X+1 ($\beta 1$ et $\beta 2$) et en position X+3 ($\alpha 1$). La présence d'une glycine au niveau du résidu de jonction caractérise les

motifs HXH se trouvant dans la partie droite du diagramme de Ramachandran. En dépit de grand nombre de cassures induites par une proline ou une glycine, environ la moitié des motifs HXH n'impliquent ni proline, ni glycine. Un bon candidat pour la position X est alors l'asparagine. Ce résidu peut en effet former des interactions stabilisantes dans les 3 cas aussi bien avec l'hélice N-terminale qu'avec l'hélice C-terminale.

Enfin, nous avons mis en place une méthode pour prédire (1) la conformation du résidu se trouvant avant une proline dans la séquence et (2) du résidu en position X lorsque la cassure présente a un fort angle d'inclinaison. Ces prédictions sont effectuées à l'aide de matrices de scores spécifiques des positions construites à partir des séquences de notre base de données HXH et de la PDB_25.

L'analyse des motifs HXH développée ici devrait fournir des informations utiles aussi bien pour la modélisation moléculaire que pour la conception de structures des protéines. Par exemple, il sera possible de tester la compatibilité d'une séquence avec des configurations spécifiques, en particulier pour les déformations induites par une proline. Cette étude va nous permettre d'initier l'analyse détaillée de la cassure de l'hélice transmembranaire 2 au sein des RCPG de classe A.

3.2 L'évolution de l'hélice transmembranaire 2 chez les RCPG

Nos récepteurs d'intérêts, les récepteurs de l'angiotensine II, possèdent une proline en position 2.58 selon la nomenclature de Ballesteros au niveau de l'hélice transmembranaire 2. La rhodopsine bovine, quant à elle, présente un renflement π dû à deux résidus glycine successifs aux positions 2.56 et 2.57. Alors que nous terminions cette étude, la structure d'un 2^{ème} RCPG, le récepteur β 2 adrénergique, a été résolue. La structure de l'hélice transmembranaire 2 du récepteur β 2 adrénergique est similaire à celle de la rhodopsine avec la présence d'un renflement π , lié à la présence d'une proline en position 2.59. Ce motif GG observée au sein de la rhodopsine n'est effectivement pas conservé au sein des RCPG. Il peut être remplacé soit par une proline en position 2.58 (comme pour nos récepteurs d'intérêts), 2.59 (comme pour le récepteur β 2 adrénergique), 2.60 mais la présence d'une proline n'est pas indispensable (20% des récepteurs humains). Nous avons voulu savoir si la position de la proline en 2.58 entraînait des modifications de la conformation locale et/ou globale de l'hélice transmembranaire 2. Une analyse de la présence ou non de cette proline aux différentes positions a été effectuée dans le génome humain, puis dans les génomes de 4 espèces qui sont apparues progressivement durant l'évolution : un nématode (*Caenorhabditis elegans*), un insecte (*Drosophila melanogaster*), un protochordé (*Ciona intestinalis*) et un poisson osseux correspondant aux plus anciens vertébrés (*Danio rerio*). Enfin nous avons effectué une analyse exhaustive des motifs 3D similaires au renflement π observé au niveau de TMH2 de la rhodopsine bovine à l'aide du programme SPASM. Cette étude a donné lieu à une publication en cours de soumission dans PROTEINS et présentée ci-après.

3.2.1 Article II : "Structural conservation in the GPCR family: Evolution of transmembrane helix 2"

STRUCTURAL CONSERVATION IN THE G PROTEIN COUPLED RECEPTOR FAMILY: EVOLUTION OF TRANSMEMBRANE HELIX 2

Julie Devillé, Julien Rey, Matthieu Moreau, David Thybert and Marie Chabbert*

CNRS UMR 6214 – INSERM U771, Faculté de Médecine d'Angers, F-49045 ANGERS, FRANCE

* To whom correspondance should be addressed : UMR CNRS 6214 – INSERM U771, 49045 Angers, FRANCE;

Tel : 33 2 41 73 58 73 ; Email : marie.chabbert@univ-angers.fr

Key words: G protein coupled receptors, helix kink, helix bulge, proline

ABSTRACT

Bovine rhodopsin is the only G protein coupled receptor whose crystallographic structure has been resolved. The validity of this structure as a template for homology modelling of class A GPCRs depends upon the conservation of the structural features of the transmembrane helices. In this paper, we focus on the transmembrane helix 2 (TMH2) of GPCRs. In rhodopsin, this helix is bent by a π bulge located at a GG motif (positions 2.56-2.57) which is not conserved among GPCRs. On the other hand, in the human GPCRs, 80% of class A receptors possess a proline in TMH2 located either at positions 2.58, 2.59 or 2.60.

The analysis of class A GPCRs from four additional fully sequenced genomes indicates that proline at position 2.60 is marginal throughout evolution, whereas the relative number of receptors with no TMH2 proline or a 2.59 proline is highly variable and does not show any evolutionary trend. On the other hand, the weight of GPCRs with a proline at position 2.58 is low (10-20%) up to vertebrates where it markedly increases with the apparition of chemotactic and purinergic receptors. In invertebrate species, most of these receptors are orthologs of the somatostatin or opioid receptors. In insects, some opsins also have a proline at position 2.58. Detailed sequence analysis strongly suggests that the presence of a proline at position 2.58 may be related to the deletion of one residue from an initial bulge structure.

To address this point, we performed an exhaustive search of kinks and bulges consistent with the rhodopsin TMH2 structure in α -helices of the Protein Data Bank and we analyzed the positions of the prolines in the hits. Three-dimensional data mining indicates that the presence of a proline at position 2.58 is not compatible with a bulge structure but is consistent with a typical proline-induced kink. The implications for GPCR modelling are discussed.

INTRODUCTION

G protein coupled receptors (GPCR) represent the main family of transmembrane receptors in the human genome [1, 2]. These receptors are involved in signal transduction of a large number of different endogenous or exogenous stimuli, including photons,

ions, organic odorants, amines, lipids, nucleotides, peptides and proteins [3] and have a major pharmaceutical importance as the target of more than 40% of presently available drugs. GPCRs are classified in several classes on the basis of sequence similarity. One of the most frequent classification uses classes A – F designed to cover both vertebrate and invertebrate GPCRs [21]. Rhodopsin is the prototype of the largest GPCR subfamily that includes about 90% of human GPCRs and corresponds to class A. In the human genome, class A includes around 800 members. About 60% of them correspond to olfactory receptors and 40% to non olfactory receptors.

Bovine rhodopsin is the only GPCR whose crystallographic structure has been resolved[9]. This structure is widely used as a template for homology modeling of class A GPCRs. The validity of this structure as a template is thus a crucial issue for drug design. The rhodopsin structure is composed of seven transmembrane helices (TMH). As commonly observed in proteins, most of these helices are not straight, but kinked or curved. Several kinks are related to highly conserved prolines, located in helices 5 to 7. However, a major concern is related to TMH2. In rhodopsin, this helix is kinked at a GG motif corresponding to a π bulge. This GG motif is located at positions 2.56-2.57, according to Ballesteros' nomenclature[112] (i.e. 6 and 7 residues downstream the highly conserved Asp residue of TMH2) and is not conserved among GPCRs. On the other hand, a proline residue is frequently observed in this helix, at positions 2.58, 2.59 or 2.60. However, its presence is not strictly required and about 20% of class GPCRs do not have a proline at these positions.

As prolines are known to induce helix kinks[113], we wondered how the bent of TMH2 observed in the crystal structure of rhodopsin could be mimicked in different GPCRs depending upon the presence and the position of the proline residue. To answer this question, we initiated a systematic analysis of TMH2, first in class A GPCRs from five fully sequenced genomes then in all known sequences of opsin orthologs. Our data strongly support the assumption that change in proline positioning is related to the deletion/insertion of one aminoacid at the level of the helix elbow. Three dimensional data mining of the Protein Data Bank reveals that the rhodopsin kink

can be mimicked either by a “typical” proline kink for GPCRs possessing at proline at position P2.58, or by a bulge when proline is located at position 2.59 or 2.60. Implications for GPCR modelling are discussed.

MATERIALS AND METHODS

GPCR sequence analysis. The sequences of class A, non olfactory GPCRs from *H. Sapiens*, *C. elegans*, *D. melanogaster*, and *D. rerio* were obtained from the UniProt database (<http://www.expasy.org>). Sequences from *H. sapiens* were clustered with the nrdb program[114] (90% sequence identity) to remove doublets, splice or polymorphism variants. Selection of one sequence in each cluster was carried out manually in order to select the SWISSPROT sequence wherever it is possible. The sequences of this non redundant set were aligned with ClustalX[115] and manually refined using Genedoc (<http://www.psc.edu/biomed/genedoc>). This led to a non redundant set of 282 sequences from *H. sapiens*.

Most of the 434 sequences from *C. elegans* assigned as class A GPCRs in UnitProt do not have the conserved residues typical of class A GPCRs from higher organisms and comparison with human class A GPCRs is meaningless. We thus built a Hidden Markov Model (HMM) of class A GPCRs from the aligned sequences of the non redundant set of human class A GPCRs using HMMER[116]. This profile was used to select sequences consistent with the HMM profile of human class A GPCR using an E value of 10 as cutoff. Only 142 sequences from *C. elegans* were selected by this filter. The same profile was used to filter sequences of class A GPCRs from *D. rerio* and *D. melanogaster*. In these cases, however, only a few highly suspicious sequences were removed (6 cases out of 239 and 8 cases out of 147 for *D. rerio* and *D. melanogaster*, respectively).

The number of GPCR sequences from *C. intestinalis* present in the UniProt being too small (18 hits), class A GPCRs were searched for in the translation of the *C. intestinalis* genome available at the ENSEMBL database (www.ensembl.org). Search was carried out with HMMER, using the HMM profile of human class A GPCRs (E value of 10), leading to 79 hits.

These sequences were clustered with nrdb and manually selected to have non redundant sets with sequence identity <90%. For each genome, the sequences were aligned against the profile of human GPCRs using Clustal X and the resulting alignment manually refined with Genedoc. At this step, a few additional sequences appearing as suspicious because they lacked the anchor residues typical of class A GPCRs were removed. Usually, their E-value obtained with HMMER was in the 0-10 range. Finally, the non redundant sets of class A GPCRs were composed of 108 sequences for *C. elegans*, 54 sequences for *C. intestinalis*, 74 sequences for *D. melanogaster* and 179 sequences for *D. rerio*.

Opsin sequences from any species (195 hits) were obtained from the Swiss-Prot database at www.expasy.org. As previously, these sequences were clustered with the nrdb program (90% sequence

identity), leading to a non redundant set of 82 sequences that were aligned with clustalX.

The amino acid numbering scheme was based on Ballesteros' system[112]. The position of the most conserved residue in each helix n was assigned the number $n.50$ and was used as relative position reference. The corresponding positions for bovine rhodopsin were N55 (1.50), D83 (2.50), R135 (3.50), W161 (4.50), P215 (5.50), P267 (6.50) and P303 (7.50).

Neighbor-Joining trees. Neighbor-joining trees were calculated with the MEGA3.1 program[117]. Distances between opsins were calculated from the multiple alignment of full length sequences, using the Dayhoff matrix model. Five hundred replicates were obtained and a consensus tree was calculated.

Classification of class A GPCRs: Human receptors with known ligands were classified in twelve groups according to the classification developed by Fredriksson et al[24], using the same nomenclature as these authors^[24, 118] for clarity purpose. Orphan receptors and receptors that were not included in Fredriksson's study were classified according to the sequence identity of the transmembrane domain with closest homologs. Unclassified receptors (UC) refer to receptors without close known paralogs (threshold of 30% for the transmembrane domain). Relaxin-3 and urotensin II receptors were classified with somatostatin/opioid/galanin receptors on the basis of their high sequence identity with the somatostatin receptors (40%)

GPCRs from each non-human genome were aligned with human GPCRs on a genome-by-genome basis. Classification was based on the sequence identity of the transmembrane domain with the closest human orthologs. A threshold value of 30% was used throughout this analysis. This value was selected as it allowed an unambiguous assignment to one of the human groups. Receptors whose sequence identity with human receptors was below this threshold were not classified, several groups being equally probable.

3D data mining. Three dimensional data mining of structures similar to the transmembrane helix 2 of bovine rhodopsin was carried out with the SPASM ("SPatial Arrangements of Side chains and Main chains") program[53] using the SPASM server accessible at <http://portray.bmc.uu.se/cgi-bin/spasm/scripts/spasm.pl>.

The search database contained 4946 entries from the protein Data Bank with sequence identities < 25% (Jan 2007 release). Search was limited to X-ray resolved structures with a resolution better than 2.5 Å. The search motifs contained only the backbone atoms from residue 79 (2.46) to 99 (2.66). Data parsing was carried out with home developed scripts written in Perl. The analysis was limited to the structures with a rmsd ≤ 1 Å. When the search motif was missing the elbow residues 88-90 between the N and C terminal helices (positions 2.55 – 2.57), the structures and the sequences of the missing part were recovered directly from the Protein Data Bank using home developed scripts. The superposition of the fragments including the elbow residues was carried out with MODELER[54].

Three-dimensional data mining of proline induced kinks in α -helices was carried out locally on a subset of the Protein Data Bank based on the 25% threshold list compiled by Hobohm and Sander[52] (March 2006 release). This list is accessible at <http://bioinfo.tg.fh-giessen.de/pdbselect>. The subset contained only soluble proteins whose structure was determined by crystallography with a resolution ≤ 2.5 Å and an R-factor ≤ 0.25 . Helix definition was determined by DSSP[49].

Molecular Modeling. PYMOL (DeLano Scientific LLC, San Francisco, USA) was used for molecular graphics analysis and figure preparation. Molecular modelling was carried out with MODELLER[54] (version 8v1). The “typical” proline kink used for data mining with SPASM was modelled from the structure of bovine rhodopsin transmembrane helix 2 (PDB code: 1L9H). Briefly, the N-term and C-term parts of rhodopsin TMH2 (up to residue F88 (2.55) and from residue T93 (2.60), respectively) were used as a template to model a proline kink in which these two helices were separated by a XXP motif.

RESULTS

1. Analysis of proline positioning in TMH2 of human class A GPCRs

We analysed the presence or not of proline in TMH2 and its position in a non redundant set of 282 class A, non olfactory GPCRs. For clarity purpose, the receptors were grouped according to the classification of class A GPCRs developed by Frederickson et al. [24, 118] (Table I). Proline could be located at positions 2.48 (41%), 2.59 (36%) or, in a few cases, 2.60 (2%). Absence of proline at these positions was observed in about 20% of the sequences. Four cases of doublets (P2.58P2.59 or P2.59P2.60) were also observed. In these cases, the receptors were classified with the other members of their groups.

Most groups are characterized by consensus behaviour of TMH2 proline (Table I). The receptors for chemokines and vasoactives peptides as angiotensin II or bradykinin along with chemotactic receptors, which are closely related[24], display a proline at position 2.58. A single exception is observed for the Chemokine receptor-like 2 (CML2) which has no proline in transmembrane 2. The PUR sub-family, which includes purinergic receptors, closely related receptors with a wide variety of ligands as Succinate receptor 1 (SUCR1) and Sphingosylphosphorylcholine receptor (SPR1) and orphan receptors (45 members) display also a proline at position 2.58 which is strictly conserved. Proline at position 2.59 is mainly observed for the PEP and the AMIN sub-families. The PEP sub-family is constituted by receptors for a wide variety of peptides, including neuropeptides. Most PEP receptors have a proline at position 2.59 (29 cases) but members of this sub-family can also have a proline at position 2.58 (2 cases) or 2.60 (5 cases) or no proline (3 cases). The AMIN sub-family is constituted by receptors for biogenic amines. Most AMIN receptors have a proline at position 2.59 (40 cases), except the acetylcholine receptors and the trace amine TAAR8, which do not have a proline. Sub-families

characterized by the absence of proline include glycoprotein hormone receptors and leucine-rich repeat receptors (LGR, 8 members), the receptors related to the proto oncogene MAS (MRG, 9 members) and the receptors for phospholipids (EDG), melanocortin, cannabinoids and adenosine (MECA, 18 members).

A few sub-families, however, do not have a consensus behavior. The SOG sub-family includes the somatostatin, the opioid and galanin receptors[24]. The somatostatin and opioid receptors have a proline at position 2.58, whereas the galanin (GAL1-3) and the closely related metastin receptor (KISSR) receptors have a proline at position 2.59. Figure 2a shows the multiple alignment of the TMH2 sequences of these receptors, with aligned proline, which strongly suggests that an insertion/deletion of the residue preceding proline occurred during evolution. Similarly, in the prostaglandin sub-family, the presence of a proline at position 2.59 is not obligate. Alignment of the TMH2 sequences of this sub-family (Figure 2b) indicates that substitution of proline at this position is possible. It is interesting to note that the sequences without proline have a glycine residue located at position 2.58. Alignment of the closely related trace amine receptors provides another example in which the proline usually observed at position 2.59 is mutated to a leucine residue in TAAR8 (not shown). Finally, in the opsin sub-family, members have different behaviors and proline may be present at positions 2.59 or 2.60 or not at all.

2. Evolution of the TMH2 proline

Although GPCRs are present in almost any eukaryotic organism, including insects and plants, the five A-F classes are not represented in all the organisms[21]. An analysis of GPCRs in 13 species with fully sequenced genomes[118] has shown that the class A initiates with nematodes[118]. To analyse the evolution of TMH2, we thus focused on the genomes of four species that appeared progressively during evolution: a nematode (*Caenorhabditis elegans*), an insect (the fruit fly *Drosophila melanogaster*), a protochordate from the tunicate lineage (*Ciona intestinalis*) and a bony fish corresponding to most ancient vertebrates (*Danio rerio*).

The sequences of class A GPCRs from these species were clusterized at 90% sequence identity, to insure non redundant data sets, and aligned against the profile of human GPCRs. We searched for the presence of proline from position 2.58 to 2.60 in the different genomes and compared the resulting distribution to that observed in *H. sapiens* (Fig. 2). A striking result is the increase in the relative number of receptors with proline at position 2.58 in most recent species. Only 4% of class A receptors from *C. elegans* have a proline at position 2.58. The weight of these receptors is still marginal in *D. melanogaster* (11%) but reaches 25 % in *D. rerio*. The relative number of receptors with proline at position 2.59 or no proline is highly variable among species and does not display any evolutionary trend. This ratio varies from 1 in *C. elegans*, to about 1.5 in *C. intestinalis* and *H. sapiens* and to 2.5 in *D. melanogaster* and *D. rerio*. This suggests that proline at position P2.59 is not obligate and

can be substituted by other residues. In any species, proline at position 2.60 is very marginal (<5%).

Assignment of the sequences was undertaken for further analysis of the evolution of TMH2 sequences as a function of sub-families (Table II). 12% of GPCR sequences from *H. sapiens* could not be classified according to the sub-families defined by Fredriksson[24]. In the other species, the ratio of unclassified sequences was only 5% in the vertebrate *D. rerio* but increased to 20-25% for *D. melanogaster* and *C. intestinalis*. Evolutionary distance made this assignment difficult for *C. elegans* and only half the sequences could be unambiguously assigned in this species. Nevertheless, general tendencies can be drawn from these data (Table II).

Four out of the eight receptors of *C. elegans* with a proline at position 2.58 could be clearly assigned as members of the somatostatin/opioid/galanin sub-family. They were most closely related to somatostatin receptors with sequence identity in the 30-35% range, whereas receptors of this sub-family with no proline or a proline at position 2.59 were closer to galanin receptors. This was also observed in *D. melanogaster*. Two of the three receptors of the SOG sub-family with proline at P2.58 had a high sequence identity with somatostatin receptors (47%) whereas the other receptors were closer to galanin receptors (30-42%). In *C. intestinalis*, SOG receptors with proline at position 2.58 could be assigned either as somatostatin receptors (3 cases with 35% sequence identity with human orthologs) or opioid receptor (1 case with 38% identity with the human kappa opioid receptor).

The weight of the P2.58 receptors increased in *D. rerio* with the apparition of the CHEM, PUR and MCH sub-families. The *D. rerio* orthologs of the receptors for chemokines or vasoactive peptides had the typical proline at position 2.58 observed in *H. sapiens*. The only exception was a bradykinin receptor 2 ortholog with a proline at position 2.59. Purinergic receptors also showed the consensual position of proline at position 2.58 observed for the human orthologs. Receptors of these families are involved in vascular regulation (vasoactive peptide receptors and purinergic receptors), in blood clotting (purinergic receptors, platelet activation) and in the immune system (chemokine and chemotactic receptors). Their apparition is related to that of the vertebrate blood system with its specific immune system.

Concerning other receptor sub-families, it is noteworthy that the positioning of TMH2 proline observed in *H. sapiens* is usually observed throughout evolution. For example, biogenic amine receptors have usually a proline at position 2.59, or in a few cases, no proline in TMH2. Similarly, most PEP receptors have a proline at 2.59. The LGR receptors have no proline throughout evolution. The MECA receptors which are present in *C. elegans* and *D. melanogaster* are closely related to adenosine receptors and, as their human orthologs, have a proline at position 2.59. The other members of this sub-family (cannabinoid, EDG, melancortin receptors), which do not have a proline in

TMH2, appeared most recently with chordates. In addition to the CHEM, PUR and MCH sub-families, melatonin, and prostaglandin receptors appeared with vertebrates, as previously reported[118]. Mas-related receptors are the most recent ones and appeared with mammals[118].

Interestingly, opsins constitute the receptor sub-family in which most variability at the level of TMH2 is observed, since it includes receptors with proline at either position 2.58, 2.59 or 2.60 or no proline. In particular, opsin receptors with proline at position P2.58 are observed in *D. melanogaster*.

3. Analysis of the TMH2 sequence of the opsin sub-family

Analysis of the positioning of TMH2 proline in different genomes reveals a wide variety in the opsin sub-family. However, this analysis relies on translated sequences which might correspond to pseudo-genes and not being expressed. To further characterize the opsin sub-family, we analyzed all the opsin sequences available in the curated SWISS-PROT database of protein sequences and corresponding to expressed proteins. From an initial set of 195 sequences of opsins, we obtained a representative ensemble of 82 sequences with an identity < 90% for the full length sequence. The neighbour-joining tree of these sequences is displayed on Fig. 3. When present, markers indicate the position of the TMH2 proline. The tree enlightens the very high similarity between vertebrate rhodopsins (Id > 70%). The identity between placental mammals being larger than 90%, bovine rhodopsin was chosen as representative of this cluster. It is closely related to the rhodopsin of the marsupial opossum (SMICR: 89%) and of chicken (CHICK: 85%). In spite of these very high identities, the GG motif of bovine rhodopsin is not strictly conserved. In the ten closely related sequences shown in Fig2b (id >79%), the GG doublet observed for placental mammals is observed only in six sequences. It is not observed in the sequences of opossum (SMICR), xenopus (XENLA), lamprey (PETMA) and chameleon (ANOCA). In two cases, a proline is present at position 2.59, in salamander (AMBTI) and spotted catshark (SCYCA), whereas it is not observed in the very closely related sequences of xenopus (AMBTI vs XENLA: 88%) and of blackmouth catshark (SCYCA vs GALML: 88%).

A similar behaviour is observed in the opsin subset shown in Fig. 6c. In these sequences, the GG doublet is not present and the anchor residue of TMH2 is a Gly instead of the wide spread Asp. However, the Asn present in the opsin family at position 2.45 is conserved along with the GYF motif in the extracellular loop 1 (ECL1), allowing comparison with rhodopsins. In these eight sequences whose identity is > 60%, the sequences of the human blue-sensitive opsin (OPSB) and of xenopus violet-sensitive opsin (OPSV) display a proline at position 2.60 which is not observed in closely related proteins (OPSB_HUMAN vs OPSB_MOUSE: 84%; OPSV_XENLA vs OPSUV_MELLUD:78%).

Analysis of the sequences of arthropod opsins (id >30% with an average of 41%) further enlightens the variability occurring in the opsin family (Fig. 3d). Proline can be observed as a singlet at position 2.58, 2.69 or 2.60 or as a doublet at positions 2.59 and 2.60. Presence of a

NxF/Y pattern allows an unambiguously alignment of the TMH2 sequence downstream the proline residue. Clearly, proline at position 2.58 results from the deletion of one residue between residue 2.55 and 2.59 as compared to sequences with proline at position 2.59. When the gap is taken into account, the NxF/Y pattern located at position 2.64-2.66 is conserved within arthropod opsins. On the other hand, proline at position 2.59 or 2.60 or at both positions can be observed in closely related sequences (Id > 50%).

4. Three-dimensional data mining of π bulges

The analysis of opsin sequences strongly suggests that evolution has yielded two main structural motifs for TMH2. Proline at position 2.59 or 2.60 could be accommodated in π bulge structures whereas proline at position 2.58 would result from deletion of one amino acid in the elbow, leading to a "typical" proline kink. To address this point, we performed an exhaustive search of structures in the Protein Data Bank consistent with the rhodopsin TMH2 structure and analyzed the positions of the prolines found in the hits. This search was carried out with the SPASM program[53] which can be used to find similar arrangements of helices or strands.

Six high resolution structures of bovine rhodopsin are presently available: 1U19[119] (2.2 Å), 1L9H[120] (2.6 Å), 1HZX[29] (2.6 Å), 1GZM[37] (2.6 Å, trigonal crystal form), 2HPY[121] (2.8 Å, lumirhodopsin) and 1F88[9] (2.8 Å). Four additional structures are also available with a lower resolution: 2J4Y[122] (3.4 Å, stabilizing mutant), 2I35[123] (3.8 Å, rhombohedral crystal form), 2I36[123] (4.1 Å, trigonal crystal form) and 2I37[123] (4.1 Å, photoactivated intermediate). The dihedral angles of the TMH2 residues from these structures were analyzed. In the six structures with high resolution (Fig. 4), the phi angle of Gly90 (2.57) is strongly shifted to negative values ($\phi = -108^\circ \pm 10^\circ$). In some of these structures, specially 1L9H and 1F88, the phi angle of Phe88 (2.55) is also markedly shifted, resulting in an average value of $84^\circ \pm 12^\circ$. In none of the six structures, Gly89 is significantly altered ($\phi = -61 \pm 13^\circ$). In either case, psi is in the $-40^\circ - -50^\circ$ range. The two low resolution structures 2I36 and 2I37 are characterized by a marked shift of phi at position 89 (2.56) ($\phi = -165^\circ$ and -100° , respectively). Whether this corresponds to local structural changes or is an artefact due to the resolution (>4 Å) cannot be presently determined. However, it is worth to note that, in these structures, the orientations of the helices are similar to those observed in high resolution structures[123]. In any case, the rmsd between the different TMH2 structures were ≤ 0.6 Å. However, comparison of the dihedral angles from different structures indicate that the same orientation of the N and C terminal parts of TMH2 can result from different combinations of dihedral angles and suggest that the bulge region might have conformational variability.

To overcome possible biases due to conformational variability, two search strategies were used. In the first one, the search motif corresponded to the backbone coordinates of residues 79 to 99 (2.46 to 2.66) from the high resolution structures of bovine

rhodopsin. The number of hits with $\text{rmsd} < 1$ Å ranged from 62 for 1L9H to 76 for 1GZM. Sequence analysis indicated that the presence of proline is favourable at positions equivalent to the rhodopsin positions 2.59 and 2.60 (Fig. 5). For any structure investigated, 46 ± 1 % of the hits possessed a proline located at position 2.59 or 2.60. However, the ratio of these hits was highly variable and ranged from 0.26 for 1L9H to 0.60 for 1U19. The average propensity of proline was 2.8 ± 0.8 at position 2.59 and 7.5 ± 0.7 at position 2.60. The propensity of proline was either null or very low (≤ 0.3) at the other positions of the motif. In particular, no proline was observed at position 2.58.

In the second strategy, the search motif corresponded to the backbone coordinates of the helical residues preceding and following the rhodopsin TMH2 π bulge. The N-terminal helix included residues 79-87 (H1), whereas the C-terminal helix included residues 91-99 (H2). The program searched for two helices having the same relative orientation as H1 and H2 and separated by 3 residues, without any constraint on the structure of the linker between the two helices. The number of hits with $\text{rmsd} < 1$ Å was higher and ranged from 83 for 1L9H to 101 for 1GZM. The percent of hits with a proline at positions 2.59 or 2.60 was similar to previously observed (44 ± 1 %). However, the relative weights were altered and ranged from 0.40 for 1L9H to 0.72 for 1U19. This led to an average propensity for proline of 3.5 ± 0.6 and 6.2 ± 0.4 at positions 2.59 and 2.60, respectively (Fig. 5a). As previously observed, no hit possessed a proline at position 2.58. These data clearly indicates that, in bulge motifs, the presence of proline is favourable at positions equivalent to positions 2.59 and 2.60 of rhodopsin, whereas it is very unfavourable at position 2.58. This corroborates our assumption that the π bulge structure observed in bovine rhodopsin is not present in GPCRs with proline at position 2.58.

Figure 6 displays the best matching structures of the rhodopsin bulge that do not possess a proline (Fig. 6a) or with a proline located at positions equivalent to position 2.59 (Fig. 6b) or 2.60 (Fig. 6c). This figure enlightens the wider structural variability of the hits with a proline at position 2.59 or without proline as compared to hits with a proline at position 2.60. To better quantify this observation, the dihedral angles of the hits were analyzed, as a function of the proline positioning. Very similar results were obtained whatever the rhodopsin structure and the search strategy used. These parameters affected the relative weights of the three data sets, but did not affect the average behaviour of each data set. A typical example is shown in Fig. 7. In the absence of proline, distortion is observed from position 2.55 to 2.57, but standard deviations are very high. In the hits with a proline at position 2.59, the distortion is better defined with a strong decrease in the dihedral angle phi at position 2.56 ($\phi = -114^\circ \pm 12^\circ$). When proline is located at position 2.60, a very strong decrease in the phi value of position 2.57 is observed ($\phi = -121^\circ \pm 7^\circ$) and is accompanied with an increase in the psi value of the preceding residue ($-20^\circ \pm 8^\circ$). This behaviour is due to a van der Waals contact between the pyrrolidine ring of the proline and the carbonyl group located three residues

upstream, leading to a very strictly defined geometry (not shown). On the other hand, for the hits with proline at position 2.59, the ring of the proline is located between the carbonyl groups of residues located 3 and 4 positions upstream, leading to increased conformational flexibility (not shown).

5. Three-dimensional data mining of proline kinks

In our initial attempts to find proline kinks compatible with rhodopsin structure, we searched for two helices having the same relative orientation as H1 and H2 and separated by a gap missing one residue as compared to rhodopsin. This approach failed, because distortion of the helix backbone cannot be located at the same position in proline bulges and kinks. As a matter of fact, only a very narrow range of dihedral angles are accessible to the proline preceding residue in the helical area[87] and the helical distortion to accommodate the proline ring is located upstream this residue. This implies an obligate local reorganisation of the protein backbone that has to be correctly modelled before 3D data mining. We thus modelled a “typical” proline kink with MODELLER (see Materials and Methods). The rhodopsin TMH2 was used as template with no restraint on the spatial structure of the 3 residues located from position 2.55 to position 2.57 and joining the 2 helices. This led to a proline kink optimizing the MODELLER molecular probability density function. This one combines spatial restraints based on the C α atoms of helices H1 and H2 and stereochemical restraints due to the presence of proline on the kink residues. The resulting structure was used as a template for 3D data mining with SPASM. As for the bulge, searches were carried out either with or without the coordinates of the kink residues. Both strategies led to similar data, with 109 and 158 hits for the search including all the residues or excluding the kink residues, respectively. In either case, proline was observed with high propensity at position 2.58 (7.5 and 5.3, respectively) (Fig. 5b). A few examples of hits with a proline at position 2.59 were also observed, but the resulting propensity was ≤ 0.6 , indicating that the presence of proline is not favourable at this position. Proline was rigorously excluded from the other positions investigated.

The best matching structures without proline or with a proline at positions 2.58 or 2.59 are shown in Fig. 6d-f. They are superimposed on the structure of the rhodopsin bulge, clearly indicating that the bend angle between the N and C terminal parts of TMH2 can accommodate the deletion of one residue in the bulge elbow. Dihedral angles were measured for the different hits and compared to the dihedral angles observed in “typical” proline-induced kinks observed in contiguous helices (Fig. 8). In the absence of proline, distortion is mainly located at position 2.56 with decrease in ϕ ($-75^\circ \pm 16^\circ$) and increase in ψ ($-26^\circ \pm 14^\circ$). However, standard deviations are very high and make further analysis difficult. When a proline is present at position equivalent to 2.58, a strong distortion is located at position 2.56 (2 residues upstream the proline) with $\phi = -84^\circ \pm 10^\circ$ and $\psi = -22^\circ \pm 13^\circ$. The dihedral angles of the preceding residue are also significantly different from

standard helical values ($\phi = -71^\circ \pm 9^\circ$; $\psi = -29^\circ \pm 13^\circ$). For the three hits with a proline at position 2.59, a distortion is also observed at position 2.56 ($\phi = -77^\circ \pm 11^\circ$, $\psi = -22^\circ \pm 11^\circ$), corresponding to three residues upstream the proline. When a proline is located in contiguous helices, it induces a distortion of the dihedral angles of the residues located two or/and three positions upstream the proline, with a significant decrease in ϕ (Fig. 8d), as previously observed (Devillé et al., in press).

DISCUSSION

Analysis of protein families indicates that, during evolution, structural conservation is much stronger than sequence conservation[124, 125]. Indeed, structure is usually strongly conserved up to sequence identity of 25-30% and structural divergence progressively increases below this limit. However, similar overall fold of homologous proteins can occur even when sequence conservation is as low as 10%, e.g. cytokines of the IL-6 family. The “evolutionary triumphant” GPCR family is an example of a very wide diversity in sequences. Within class A GPCRs, the average sequence identity is about 20%, but may be as low as 10-12% between some sub-families. During evolution, divergence from a common ancestor has led to the sequence diversity necessary to specifically interact with very varied ligands while maintaining a similar seven helix bundle fold. In addition, about 4% of the residues are highly conserved. In particular, each helix has a specific highly conserved residue that is used to anchor helix sequence, even when the overall sequence identities are very low. These residues are involved in electrostatic interactions which are impaired upon activation[121]. Mutation of these anchor residues may lead to constitutive activation or inactivity, indicating a key role in the activation process conserved throughout evolution.

Determination of the reliability of the rhodopsin structure as a template for class A GPCRs is crucial for the design of drugs targeted against this family of receptors. It was suggested that the overall conservation of the seven helix bundle fold structure might accommodate local diversity in the structure of some receptors [39]. In particular, the presence of a proline at different positions in TMH2 might result in a divergence in the structure of some GPCRs compared to rhodopsin as this proline might bend the transmembrane helix 2 differently, depending upon its position [39]. Several studies were focused on the proline at position 2.58 found in the chemokine sub-family. In this sub-family, the proline is part of a TXP motif that is crucial for receptor activation[126]. Molecular dynamics simulations of a model α -helix containing a TXP motif suggested that this motif might induce a change in the kink orientation of TMH2, yielding a reorganisation of the TMH bundle as compared to rhodopsin[126].

Distortion of α -helical structures by proline has been observed several decades ago[113]. When a proline is present in a α -helix, it induces a distortion of the upstream protein backbone to avoid steric clash with the pyrrolidine ring, resulting in a bend of about 30° of the

helix axis[113]. However, the dihedral angles of the residue preceding the proline must be located in a narrow area of the α region[87, 88, 100] plot, implying that distortion arises upstream this residue. Evolutionary models have shown that once this proline-induced kink is stabilized by favourable interaction of its three-dimensional environment, proline can be substituted by other amino acids without alteration of the overall bend of the helix[127].

Deformation of α -helices leading to bulges did not receive the same attention, probably because of the difficulty to detect these structures by secondary structure analysis programs as DSSP or STRIDE[128]. Identification of π -bulges or π -helices is based on locally developed algorithms aimed at searching hydrogen bonds between the NH group of residue i and the CO group of residue $i-5$ [128-130]. These studies have shown that the π -helix pattern is related to structural parameters corresponding to a shift of the phi dihedral angle to the $-120^\circ \pm 30^\circ$ range with psi in the $-60^\circ \pm 20^\circ$ range. These dihedral angles lead to a helix "opening" with a shift from the α to the π helical hydrogen bonding pattern. This deformation may be located to one residue, leading to a π -turn[129]. Alternatively, the α -helix can resume after a succession of π -turns or a coil-like structure, leading to π -helix[128] or π -bulge[130]. For clarity purpose, we will refer to all of these structures as π -structures.

π -structures can be detected by DSSP as a structural motif in which the two α -helices are linked by a single residue (HXH motif) (Devillé et al., in press). Indeed, in our analysis of the helix-X-helix motif, the so-called α_1 conformation of the linker residue does correspond to a π -turn with the associated unwinding of the helix and i to $i-5$ H-bonding pattern (Devillé et al., in press). In the six rhodopsin structures with a resolution lower than 3.0 Å, the TMH2 bulge could be assigned as a HXH motif (linker residue located at Gly90), albeit in several of these structures severe distortion was observed both for residues 88 and 90 (Fig. 4).

Analysis of π structures indicates the high propensity of proline at the beginning of the C-terminal helix[128-130]. Differences in numbering make difficult direct comparison of proline positions. However, when the reference residue X corresponds to the position with the most marked shift in the phi dihedral angle, all the studies point to a high propensity of proline for position X+3[128-130], in agreement with our data on helix-X-helix motifs (Devillé et al., in press). Our search of π bulges similar to the rhodopsin TMH2 bulge corroborates the high propensity of proline downstream the bulge. In our analysis, proline has a high propensity for two positions, equivalent to the rhodopsin positions 2.59 or 2.60 (Fig. 5a). Detailed analysis of the dihedral angles of these structures indicate that, in either case, the proline is located three residues downstream the most distorted residue (Fig. 7b,c). However, the dihedral angles of the bulge are different in both cases. Motifs with proline at position 2.60 have a very strong distortion located at position 2.57 with low conformational variability ($\phi = -$

$121^\circ \pm 7^\circ$). Graphical analysis of the corresponding structures indicates that these motifs correspond to a contact between the proline ring and the carbonyl group at position 2.57. On the other hand, for motifs with proline at position 2.59, the distortion is of smaller amplitude but involves more residues (positions 2.55 to 2.57 with a maximum for position 2.56). These motifs have an increased conformational variability as compared to the P2.60 motifs, with phi reaching $-114^\circ \pm 12^\circ$ at position 2.56. (Fig. 7b). Larger conformational variability is also observed in the absence of proline with phi reaching $-96^\circ \pm 20^\circ$ at position 2.56 (fig. 7a).

Interestingly, position 2.59 is not the most favourable position for proline in π bulges structures but has been selected by GPCRs during evolution. It is consistent with the observation that the relative weights of GPCRs with a proline at position 2.59 or without proline in TMH2 do not show any evolutionary tendency. The structures with proline at position equivalent to 2.60 display a very low conformational variability whereas a larger variability is observed for proline at position 2.59 or without proline (Fig. 7). Conformational changes at the helix kinks, without gross reorganisation of the helix bundle, have been pointed out in the structure of a photoreaction intermediate of rhodopsin[121] (PDB code: 2HPY). Our analysis suggests that the TMH2 bulge might participate to such conformational changes in latter stages of the activation process.

While prolines at position 2.59 or 2.60 (Table I) are consistent with a bulge structure similar to that observed in TMH2 rhodopsin, the presence of a proline at position 2.58 rules out the possibility of a bulge structure. Our 3D data mining does not support the assumption of a reorganisation of TMH2 depending upon the position of the proline, but clearly indicates that, when a proline is located at position 2.58, the bulge found in the rhodopsin structure has to be substituted by a proline kink (Fig. 6). These two structures induce similar bend angle in a regular α -helical structure, and results in similar tertiary structure of the N and C-terminal parts of the helix, when the deletion of one residue between the bulge and the kink is taken into account.

When proline is present at position 2.58, this residue plays a key role in the activation mechanism of the receptor. Indeed, mutation of this residue in the chemokine receptor CCR5[126] or in the angiotensin receptor AT1[131] markedly impairs receptor activation. Binding of angiotensin II to the P2.58A receptor is not altered [131], whereas the affinity of the P2.58A CCR5 receptor depends upon the considered chemokine with no effect on CCL5 but a marked decrease for CCL3 and CCL4[126]. This proline is part of a T(S)XP motif which is conserved in 80% of the GPCRs with a proline at position 2.58. The importance of this threonine in the activation mechanism of CCR5[126] suggests that receptors with a proline kink in TMH2 may have evolved a specific mechanism to trigger activation.

In non-vertebrate genomes, only a few G protein coupled receptors have a proline at position 2.58. Most of these receptors are related to opioid or somatostatin receptors or cannot be classified (Table I). An exception

is the occurrence of proline at position 2.58 in insect opsins (Fig. 3c). The weight of these receptors markedly increases in vertebrates with the apparition of chemokine receptors (including closely related vasoactive peptide receptors) and purinergic receptors. These receptors are involved in the vascular and immune system characteristic of the vertebrates.

Finally, the importance of the TMH2 proline positioning for GPCR modelling has to be outlined. Two different structures can mimic the rhodopsin TMH2 elbow: Kinks are consistent with a proline at position 2.58, whereas bulges are consistent with a proline at position 2.59 or 2.60. The structure of the rhodopsin bulge can be directly used as a template to model a proline induced bulge, when the proline is located at position 2.59. In the absence of proline, both structures are possible. Insertions or deletions can be positioned at the helix elbow. Careful analysis of orthologous and paralogous sequences may be required to correctly model TMH2, especially for receptors with no proline in this helix.

ACKNOWLEDGMENTS: We thank NEC Computers Services SARL (Angers, FRANCE) for the kind availability of a multiprocessor server. J. D. was supported by fellowships from INSERM-Région des Pays-de-la-Loire and from the Association pour la Recherche sur le Cancer (ARC). J. R. is supported by a fellowship from CNRS.

ABBREVIATIONS: GPCR: G protein coupled receptor; TMH: transmembrane helix; rmsd: root mean square deviation.

LEGENDS TO FIGURES

Figure 1: Positioning of TMH2 proline in human GPCR sub-families: (a) the somatosatin/opioid/galanin sub-family; (b) the prostaglandin subfamily. The star indicates the position of the TMH2 anchor (position 2.50). Positions that are at least 80% conserved or type-conserved are enlightened in light grey.

Figure 2: Relative weight of GPCR sequences with proline at position 2.58, 2.59 and 2.60 or no proline with the genome of *C. elegans* (white bars), *D. melanogaster* (dotted bars), *C. intestinalis* (grey bars), *D. rerio* (hatched bars) and *H. sapiens* (black bars).

Figure 3: Evolution of the rhodopsin transmembrane helix 2 (a) Unrooted Neighbour Joining tree of 82 non redundant sequences of rhodopsin (Id < 90% for the full length sequence). The tree represents the consensus tree obtained after 500 replicates. The label located near the sequence name indicates the presence of proline at position 2.58 (circle), 2.59 (up triangle), 2.60 (down triangle) or both 2.59 and 2.60 (diamond). Absence of label indicates absence of proline; (b) Alignment of the TMH2 sequences of vertebrate rhodopsins; (c) Alignment of the TMH2 sequences of vertebrate blue, violet and ultraviolet opsins; (d) Alignment of the TMH2 sequences of insect opsins. The star indicates the position of the TMH2 anchor (position 2.50). Positions that are at least

80% conserved or type-conserved are enlightened in light grey. When present, the glycines of the GG motif and the prolines are enlightened in dark grey and in black, respectively.

Figure 4: Average phi and psi dihedral angles of transmembrane helix 2 in the rhodopsin structures with resolution <3.0 Å (PDB access numbers: 1U19, 1L9H, 1HZX, 1GZM, 2HPY and 1F88[9]).

Figure 5: (a) Propensity of proline along the bulge positions. The numbering is done by similarity with rhodopsin. Positions 2.59 and 2.60 correspond to superposition with Thr-92 and Thr-93, respectively. The open and closed bars correspond to 3D data mining carried out with or without the residues 87-90, respectively; (b) Propensity of proline along the kink positions. Position 2.58 corresponds to eight residues downstream the anchor D2.50 in the proline kink modeled from the rhodopsin TMH2. The open and closed bars correspond to 3D data mining carried out with or without the residues 87-89, respectively.

Figure 6: Superposition of the C α trace of rhodopsin TMH2 (black) with the best matching structures obtained with SPASM for bulge search (a-c) and kink search (d-f). In the bulge search, the search motif was based on residues 2.46-2.54 and 2.58-2.66 of the rhodopsin structure 1U19. The best seven fits are shown with a colour code based on the rmsd. For the matching structures without proline (a), the rmsd increases from dark blue to cyan structures. When a proline is present at position 2.59 (b), the rmsd increases from green to yellow structures. When a proline is present at position 2.60 (c), the rmsd increases from orange to red structures. In the kink search, the search motif was based on residues 2.46-2.53 and 2.57-2.64 of the "typical" proline kink (grey) modelled as described in Materials and Methods. For the matching structures without proline (d), the rmsd of the best seven structures increases from dark blue to cyan structures. When a proline is present at position 2.58 (e), the rmsd of the best seven structures increases from green to orange structures. The three matching structures with a proline at position 2.59 are shown in (f) with an rmsd increasing from light red to dark red structures. The C α of the rhodopsin residues 89 and 90 are shown as spheres.

Figure 7: Typical average phi and psi dihedral angles of the hits without proline (a) or with a proline located at position equivalent to 2.59 (b) or 2.60 (c) when structures similar to the rhodopsin bulges were searched for with SPASM. The search motif was based on residues 2.46-2.54 and 2.58-2.66 of the rhodopsin structure 1U19.

Figure 8: Average phi and psi dihedral angles of the hits without proline (a) or with a proline located at positions equivalent to 2.58 (b) or 2.59 (c) when structures similar to a proline-induced kink were searched for with SPASM. The search motif was based on residues 2.46-2.53 and 2.57-2.64 of the "typical" proline kink modelled as described in Materials and Methods. In (d), the dihedral angles of the residues surrounding a proline located in contiguous helices are given for comparison.

REFERENCES

- Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
- Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).
- Gether, U. Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocr Rev* **21**, 90-113 (2000).
- Kolakowski, L.F., Jr. GCRDb: a G-protein-coupled receptor database. *Receptors Channels* **2**, 1-7 (1994).
- Palczewski, K. et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289**, 739-45 (2000).
- Sealfon, S.C. et al. Related contribution of specific helix 2 and 7 residues to conformational activation of the serotonin 5-HT_{2A} receptor. *J Biol Chem* **270**, 16683-8 (1995).
- MacArthur, M.W. & Thornton, J.M. Influence of proline residues on protein conformation. *J Mol Biol* **218**, 397-412 (1991).
- Holm, L. & Sander, C. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14**, 423-9 (1998).
- Thompson, J.D., Higgins, D.G. & Gibson, T.J. 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**, 4673-80 (1994).
- Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755-63 (1998).
- Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150-63 (2004).
- Fredriksson, R., Lagerstrom, M.C., Lundin, L.G. & Schiöth, H.B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* **63**, 1256-72 (2003).
- Fredriksson, R. & Schiöth, H.B. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* **67**, 1414-25 (2005).
- Kleywegt, G.J. Recognition of spatial motifs in protein structures. *J Mol Biol* **285**, 1887-97 (1999).
- Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815 (1993).
- Hobohm, U. & Sander, C. Enlarged representative set of protein structures. *Protein Sci* **3**, 522-4 (1994).
- Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637 (1983).
- Okada, T. et al. The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *J Mol Biol* **342**, 571-83 (2004).
- Okada, T. et al. Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. *Proc Natl Acad Sci U S A* **99**, 5982-7 (2002).
- Teller, D.C., Okada, T., Behnke, C.A., Palczewski, K. & Stenkamp, R.E. Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs). *Biochemistry* **40**, 7761-72 (2001).
- Li, J., Edwards, P.C., Burghammer, M., Villa, C. & Schertler, G.F. Structure of bovine rhodopsin in a trigonal crystal form. *J Mol Biol* **343**, 1409-38 (2004).
- Nakamichi, H. & Okada, T. Local peptide movement in the photoreaction intermediate of rhodopsin. *Proc Natl Acad Sci U S A* **103**, 12729-34 (2006).
- Standfuss, J. et al. Crystal structure of a thermally stable rhodopsin mutant. *J Mol Biol* **372**, 1179-88 (2007).
- Salom, D. et al. Crystal structure of a photoactivated deprotonated intermediate of rhodopsin. *Proc Natl Acad Sci U S A* **103**, 16123-8 (2006).
- Ho, B.K. & Brasseur, R. The Ramachandran plots of glycine and pre-proline. *BMC Struct Biol* **5**, 14 (2005).
- Aloy, P., Oliva, B., Querol, E., Aviles, F.X. & Russell, R.B. Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics. *Protein Sci* **11**, 1101-16 (2002).
- Devos, D. & Valencia, A. Practical limits of function prediction. *Proteins* **41**, 98-107 (2000).
- Ballesteros, J.A., Shi, L. & Javitch, J.A. Structural mimicry in G protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structure-function analysis of rhodopsin-like receptors. *Mol Pharmacol* **60**, 1-19 (2001).
- Govaerts, C. et al. The TXP motif in the second transmembrane helix of CCR5. A structural determinant of chemokine-induced activation. *J Biol Chem* **276**, 13217-25 (2001).
- Anderson, R.J., Weng, Z., Campbell, R.K. & Jiang, X. Main-chain conformational tendencies of amino acids. *Proteins* **60**, 679-89 (2005).
- Lovell, S.C. et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**, 437-50 (2003).
- Yohannan, S., Faham, S., Yang, D., Whitelegge, J.P. & Bowie, J.U. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc Natl Acad Sci U S A* **101**, 959-63 (2004).
- Fodje, M.N. & Al-Karadaghi, S. Occurrence, conformational features and amino acid propensities for the pi-helix. *Protein Eng* **15**, 353-8 (2002).
- Rajashankar, K.R. & Ramakumar, S. Pi-turns in proteins and peptides: Classification, conformation, occurrence, hydration and sequence. *Protein Sci* **5**, 932-46 (1996).
- Cartailler, J.P. & Luecke, H. Structural and functional characterization of pi bulges and other short intrahelical deformations. *Structure* **12**, 133-44 (2004).
- Reis, R.I. et al. Participation of transmembrane proline 82 in angiotensin II AT1 receptor signal transduction. *Regul Pept* **140**, 32-6 (2007).

TABLE I

Position of TMH2 proline in a non redundant set of class A GPCRs from the human genome¹

Proline	Family	N	Receptors
2.58	CHEM ²	43 (44)	CCR5 CCR2 CCR3 CCR1 CCR4 CCR8 CX3C1 CCRL2 CCBP2 XCR1 CCR9 CCR7 CCR6 CCRL1 CXCR4 CXCR2 CXCR1 CXCR5 CCR10 CXCR3 CXCR6 RDC1 ADMR AGTR1 AGTR2 BKRB1 BKRB2 APJ GPR25 GPR15 C5ARL C5AR C3AR GPR44 FPRL1 FPRL2 FPR1 LT4R1 LT4R2 CML1 GPR32 GPR33 GPR1
	PUR ³	45 (45)	P2RY1 P2RY2 P2RY4 P2RY5 P2RY6 P2RY8 P2RY9 P2Y10 P2Y11* P2Y12 P2Y13 P2Y14 PTAFR SUCR1 OXER1 OXGR1 G109A PSYR SPR1 CLTR1 CLTR2 PAR1 PAR2 PAR3 PAR4* EBI2 FFAR1 FFAR2 FFAR3 GPR4 GPR17 GPR18 GPR20 GPR31 GPR34 GPR35 GPR55 GPR81 GPR87 GPR92 GP132 GP141 GP174 GP171 Q5KU21
	SOG ⁴	14 (18)	OPRM OPRD OPRK OPRX SSR1 SSR2 SSR3 SSR4 SSR5 NPBW1 NPBW2 RL3R1 RL3R2 UR2R
	MCHR ⁵	2 (2)	MCHR1 MCHR2
	PEP ⁶	2 (39)	MTLR GHSR
	UC ⁷	6 (34)	GPBAR Q14968 GP120 Q5KU14 GPR82 GP146
P2.59	MTN ⁸	3 (3)	MTR1A MTR1B MTR1L
	SOG ⁴	4 (15)	GALR1 GALR2 GALR3 KISSR
	PEP ⁶	29 (39)	NMUR1 NMUR2 NTR1 NTR2 GPR39 EDNRA EDNRB ETBR2 GPR37 PKR1 PKR2 NPY1R NPY2R NPY4R NPY5R BRS3 GRPR NMBR CCKAR GASR QRFPR OX1R OX2R NPFF1 NPFF2 PRLHR GNRR2 GNRHR GPR83
	OPN ⁹	2 (8)	OPN4 OPSX
	PTGR ¹⁰	5 (8)	PE2R2 PE2R3 PE2R4 PD2R PI2R
	AMIN ¹¹	36 (42)	5HT1B 5HT1D 5HT1E 5HT1F 5HT1A 5HT7R 5HT4R 5HT2A 5HT2C 5HT2B 5HT5A 5HT6R* HRH1 HRH2 HRH3 HRH4 DRD1 DRD2 DRD3 DRD4 DRD5 ADA1A ADA1B ADA1D ADA2A ADA2B ADA2C ADRB1 ADRB2 ADRB3* TAAR1 TAAR2 TAAR3 TAAR5 TAAR6 TAAR9
	MECA ¹²	4 (22)	AA2AR AA2BR AA1R AA3R
UC ⁷	19 (34)	GPR19 GPR22 GPR26 GPR27 GPR45 GPR61 GPR62 GPR63 GPR75 GPR78 GPR84 GPR85 GPR88 GP101 GP135 GP151 GP161 GP173 GP176	
P2.60	PEP ⁶	5 (39)	V1AR V1BR V2R OXYR TRFR
	OPN ⁹	1 (8)	OPSB
	UC ⁷	2 (34)	GPR21 GPR52
No P	CHEM ²	1 (47)	CML2
	PEP ⁶	3 (39)	NK1R NK2R NK3R
	OPN ⁹	5 (8)	OPN3 OPN5 RGR OPSR OPSD
	LGR ¹³	8 (8)	LGR4 LGR5 LGR6 RXFP1 RXFP2 TSHR LSHR FSHR
	PTGR ¹⁰	3 (8)	TA2R PF2R PE2R1
	MRG ¹⁴	9 (9)	MAS MAS1L MRGRF MRGX1 MRGX2 MRGX3 MRGX4 MRGRD MRGRE
	MECA ¹²	18 (22)	ACTHR MSHR MC3R MC4R MC5R CNR1 CNR2 EDG1 EDG2 EDG3 EDG4 EDG5 EDG6 EDG7 EDG8 GPR3 GPR6 GPR12
	AMIN ¹¹	6 (46)	TAAR8 ACM1 ACM2 ACM3 ACM4 ACM5
	UC ⁷	7 (34)	GP119 GP139 GP142 GP148 GP150 GP152 GP160

¹ The classification was based on Fredriksson study. N refers to the number of receptors of a group with the proline at a given position. The number between brackets indicates the total number of receptors in this group. The stars indicate receptors with a PP doublet.

² CHEM: chemokine receptors, vasoactive peptide receptors and chemotactic receptors

³ PUR: purinergic receptors, proteinase activated receptors and acid receptors

⁴ SOG: Somatostatin, opioid and galanin receptors

⁵ MCHR: melanocyte concentrating hormone receptors

⁶ PEP: peptide receptors

⁷ UC: unclassified

⁸ MTN: melatonin receptors

⁹ OPN: opsins

¹⁰ PTGR: prostaglandin receptors

¹¹ AMIN: biogenic amine receptors

¹² MECA: receptors for phospholipids (EDG), melanocortin, cannabinoids and adenosine

¹³ LGR: glycoprotein hormone receptors and leucine-rich repeat receptors.

¹⁴ MRG: Mas-related receptors

TABLE II

Position of TMH2 proline in five genomes

Classification ¹	<i>C. elegans</i>	<i>D. melanogaster</i>	<i>C. intestinalis</i>	<i>D. rerio</i>	<i>H. sapiens</i>
AMIN	No P: 2 P2.59: 16	No P: 4 P2.58: 1 P2.59: 16	No P: 3 P2.58: 1 P2.59: 7	No P: 2 P2.58: 1 P2.59: 52	No P: 6 P2.59: 36
MECA	P2.59: 1	P2.59: 1	No P: 2 P2.59: 3	No P: 18 P2.59: 4	No P: 18 P2.59: 4
PEP	No P: 8 P2.59:16 P2.60: 1	No P: 4 P2.59: 19	No P: 1 P2.58: 2 P2.59: 6 P2.60: 1	No P: 2 P2.59: 14 P2.60: 2	No P: 3 P2.58: 2 P2.59: 29 P2.60: 5
SOG	No P: 2 P2.58: 4 P2.59: 2	No P: 1 P2.58: 3 P2.59: 2	No P: 2 P2.58: 4	P2.58: 9 P2.59: 2	P2.58: 11 P2.59: 4
LGR	No P: 1	No P: 4	No P: 2	No P: 2	No P: 8
OPS		P2.58: 4 P2.59: 3	No P: 2	No P: 13 P2.59: 5 P2.60: 2	No P: 5 P2.59: 2 P2.60: 1
CHEM				P2.58: 15 P2.59: 1	No P: 1 P2.58: 46
PUR				No P: 1 P2.58: 15	P2.58: 45
MCH				P2.58: 2	P2.58: 2
MLT				P2.59: 3	P2.59: 3
PTG				P2.59: 4	No P: 3 P2.59: 5
MAS					No P: 9
UC	No P: 28 P2.58: 4 P2.59: 23	No P: 5 P2.59: 7	No P: 10 P2.58: 3 P2.59: 5	P2.58: 1 P2.59: 6 P2.60:1	No P: 7 P2.58: 6 P2.59: 19 P2.60: 2

¹ The classification was based on Fredriksson study. Assignment was carried out as described in Materials and Methods.

Figure 1

a

*

```

OPRM_HUMAN : ATNIYIFNLALADALATS-TLPPFQSVNYLMGTW
OPRD_HUMAN : ATNIYIFNLALADALATS-TLPPFQSAKYLMETW
OPRK_HUMAN : ATNIYIFNLALADALVTT-TMPFQSTVYLMNSW
OPRX_HUMAN : ATNIYIFNLALADTLVLL-TLPPFQGTDILLGFW
SSR3_HUMAN : VTNVYILNLALADELFML-GLPFLAAQNALSYW
SSR5_HUMAN : VTNIYILNLAVADVLYML-GLPFLATQNAASFV
SSR2_HUMAN : ITNIYILNLAIADELFML-GLPFLAMQVALVHW
SSR4_HUMAN : ATNIYLLNLAVADELFML-SVPFVASSAALRHW
SSR1_HUMAN : ATNIYILNLAIADELLML-SVPFLVTSTLLRHW
NPBW2_HUMAN : VTNVFILNLAVADGLFTL-VLPVNIAEHLQYW
NPBW1_HUMAN : VTNLFILNLAIADELFTL-VLPINIADFLLRQW
KISSR_HUMAN : VTNFYIANLAATDVTFLCCVPFTALLYPLPGW
GALR2_HUMAN : TTNLFILNLGVADLCFILCCVPFQATIYTLDGW
GALR3_HUMAN : TTDLFILNLAVADLCFILCCVPFQATIYTLDAW
GALR1_HUMAN : TTNLFILNLSIADLAYLLFCIPFQATVYALPTW
    
```

b

*

```

PE2R1_HUMAN : TFLLFVASLLATDLAGHVIPGALVLRLYTAGR---
PE2R3_HUMAN : SFLLCIGWLALTDLVGQLLTPEVVIVVYLSKQRWE
PE2R4_HUMAN : TFYTLVCGLAVTDLLGTLVSPVTIATYMKGQ---
PE2R2_HUMAN : LFHVLVTELVFTDLLGTCLISPVVLAAYARNQTLV
PI2R_HUMAN : AFAVLVTGLAATDLLGTSFLLSPAVFVAYARNSSLL
PD2R_HUMAN : VFYMLVCGLTVTDLLGKCLLSPVVLAAAYAQNRSR
TA2R_HUMAN : SFLTFLCGLVLTDFLGLLVTGTIVVSQHAALFEWH
PF2R_HUMAN : SFLLLASGLVITDFFGHLLINGAIAVVFVYASDKEWI
    
```

Figure 2

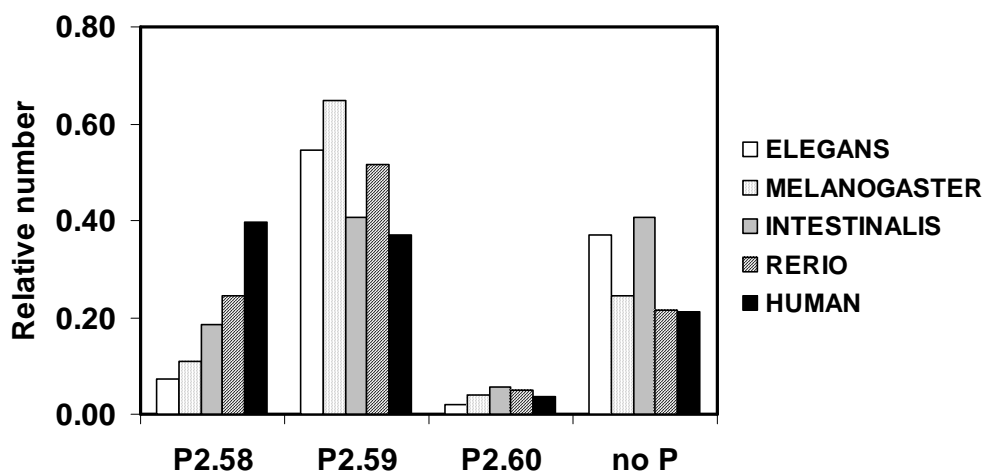


Figure 4

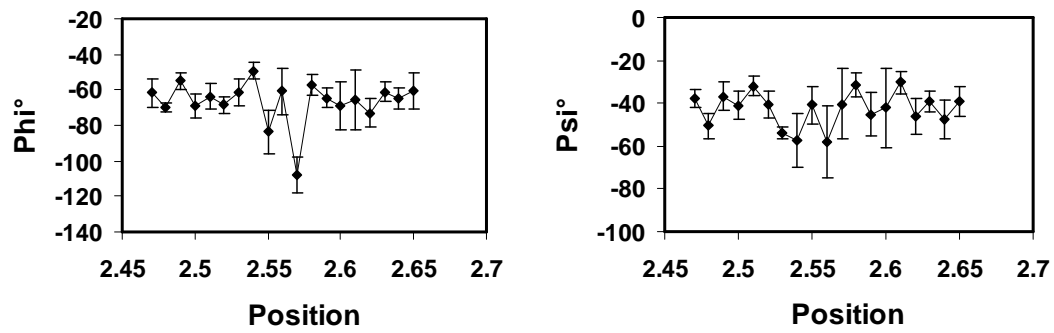
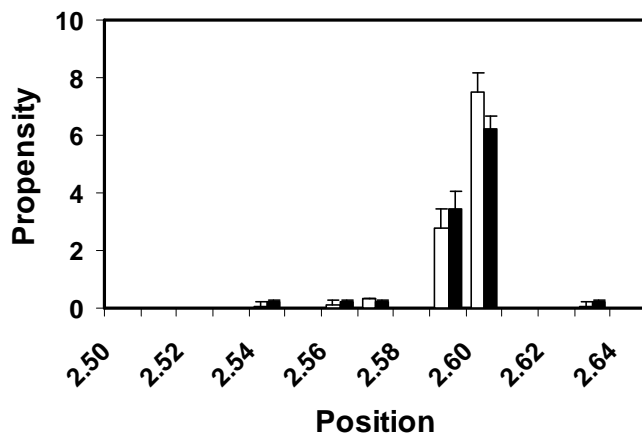


Figure 5

a



b

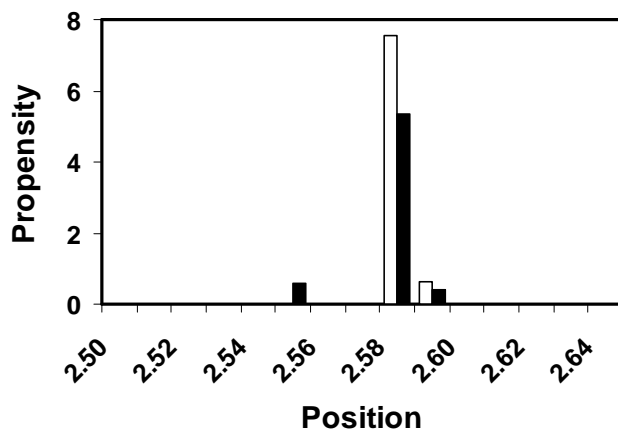


Figure 6

a



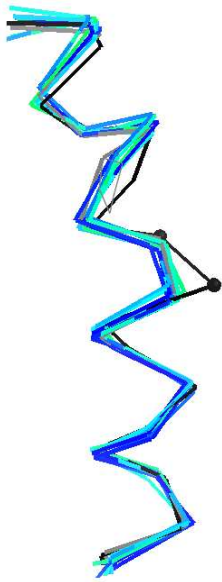
b



c



d



e

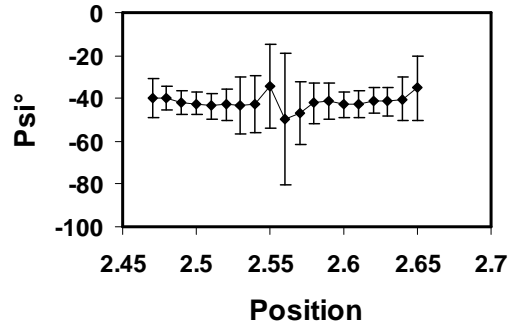
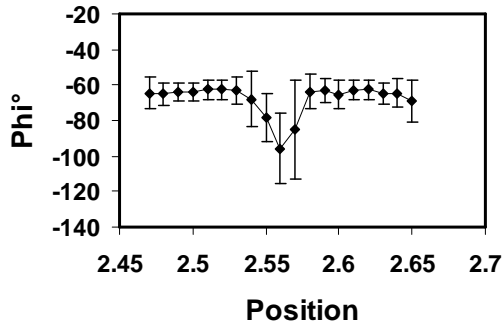


f

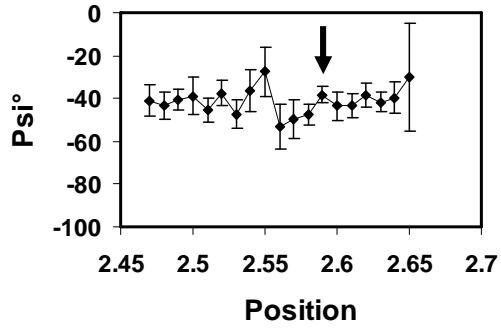
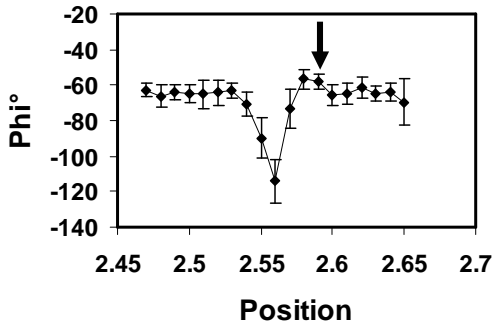


Figure 7

a



b



c

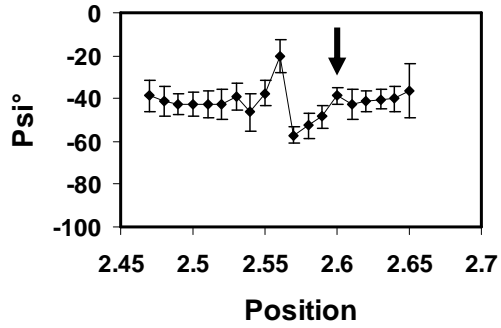
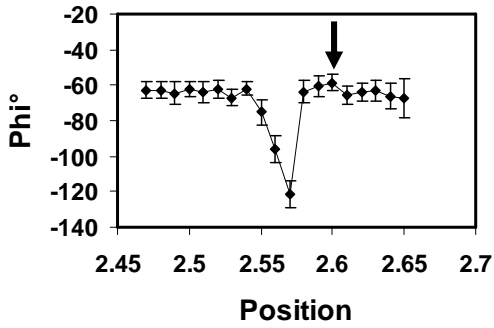
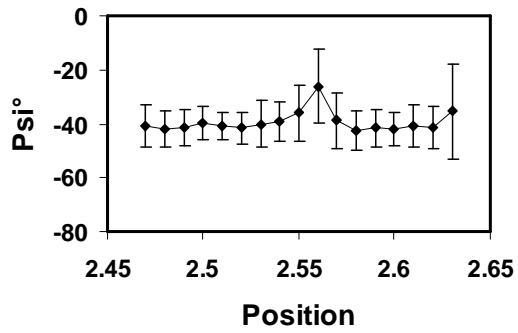
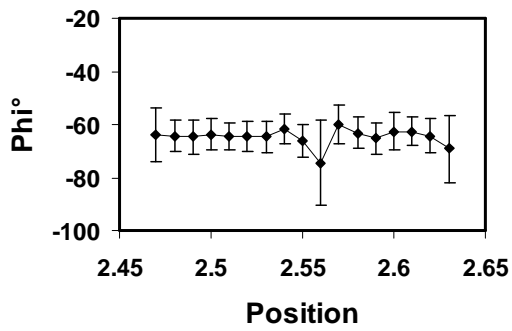
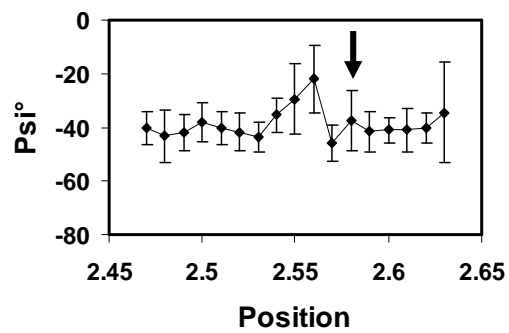
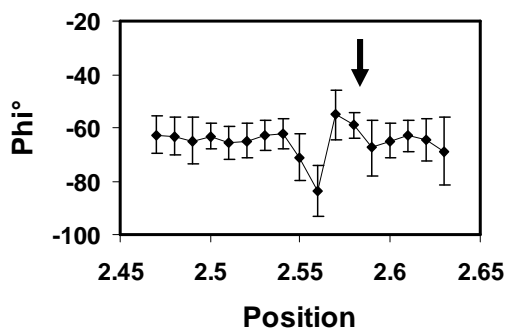


Figure 8

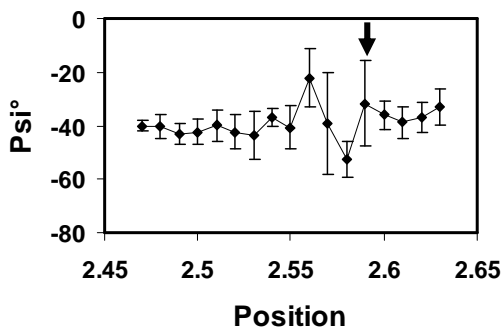
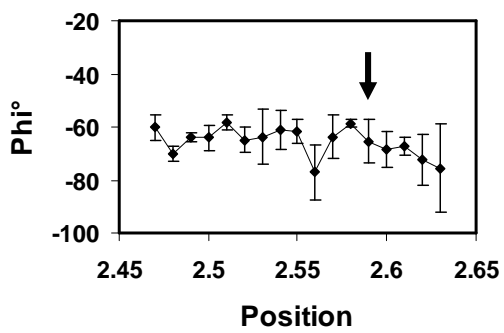
a



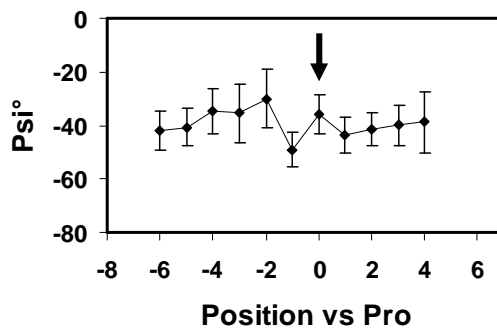
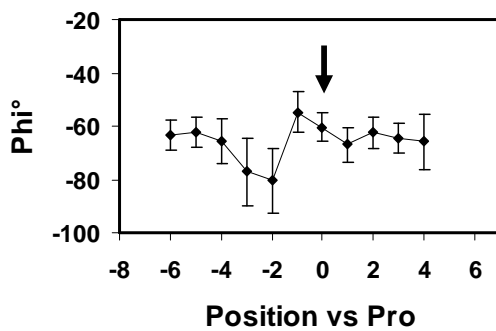
b



c



d



3.2.2 Conclusion

L'étude de l'évolution de l'hélice transmembranaire 2 au sein de la famille des RCPG de classe A montre que le poids des récepteurs ayant une proline en position en 2.58 a augmenté au cours de l'évolution. En effet, elle est présente dans 4% des RCPG de classe A chez *C. elegans*, 11% chez *D. melanogaster*, pour atteindre 25% chez *D. rerio* et 40% chez *H. sapiens*. De plus, l'apparition des familles de récepteurs avec une proline en position 2.58 est corrélée avec l'apparition des systèmes vasculaire et immunitaire des vertébrés. Les récepteurs avec une proline en position 2.59 ou sans proline ne suivent pas une ligne d'évolution et ont des proportions variables en fonctions des espèces. Enfin, quelle que soit l'espèce, la position 2.60 est très marginale (<5%). Il est intéressant de noter que la sous-famille des récepteurs des opsines est celle qui présente le plus de variabilité au niveau de la position de la proline avec des prolines en positions 2.58, en 2.59, en 2.60, en doublet 2.59-2.60, mais aussi le remplacement de la proline par le motif GG, une seule glycine, ou enfin l'absence de glycine et de proline.

Nos analyses structurales montrent donc que le modèle de la rhodopsine bovine peut être utilisé lorsque nous avons des prolines en 2.59 ou 2.60 correspondant à des renflements π . Ceci est validé par la structure cristalline nouvellement obtenue du récepteur $\beta 2$ adrénergique qui présente effectivement un renflement π alors qu'il possède une proline en position 2.59. Bien que la structure de la cassure soit différente en ce qui concerne les récepteurs avec une proline en 2.58, la structure tertiaire globale de l'hélice TMH2 est similaire à celle de la rhodopsine. La rhodopsine peut donc aussi être utilisée comme modèle structural pour modéliser les hélices TMH2 des récepteurs avec une proline en position 2.58 à **condition** de prendre en compte la délétion d'un résidu au niveau du renflement. Cette délétion permet d'obtenir une cassure proline en coude similaire à celles observées dans les hélices continues au lieu du renflement observé dans la rhodopsine et le récepteur $\beta 2$ adrénergique.

4 CONCLUSIONS ET PERSPECTIVES

Les travaux effectués au cours de cette thèse concernent l'étude de la structure des récepteurs couplés aux protéines G (RCPG), et plus particulièrement, les récepteurs de peptides vasoactifs. Les informations structurales sur les RCPG sont basées sur la structure cristalline de la rhodopsine bovine. Cette structure cristalline était jusqu'à très récemment la seule structure résolue pour les RCPG. Cette structure corrobore la structure prédite en 7 hélices α transmembranaires mais elle révèle que ces hélices sont tordues ou cassées à certaines positions spécifiques. Mon travail avait donc comme objectifs de déterminer : (1) si ces motifs présents au sein de la rhodopsine étaient conservés dans les différentes familles des RCPG et (2) si les différentes sous-familles pouvaient adopter différentes conformations. Cette question est fondamentale pour les récepteurs de peptides vasoactifs comme AT1 et AT2. En effet, ces récepteurs possèdent une proline en position 2.58 au sein de l'hélice transmembranaire 2 (TMH2), qui est cruciale pour l'activation de ces récepteurs. Il est essentiel de savoir si cette proline, non conservée au sein de tous les RCPG, induit une réorganisation complète de la structure en sept hélices.

La première étape de mon travail a été d'effectuer une analyse exhaustive des hélices cassées au sein de la PDB dans le but de comparer les cassures induites par les prolines et celles observées en absence de proline. Les prolines induisent un petit nombre de motifs de cassures possibles mais leur présence n'est pas obligatoire et les mêmes conformations sont observées en leur absence. Lorsqu'elles sont présentes, les prolines peuvent induire trois motifs de cassure de type hélice-X-hélice, correspondant aux trois conformations distinctes du résidu de jonction : $\alpha 1$, $\beta 1$ et $\beta 2$. Plus particulièrement, la conformation $\alpha 1$ est caractérisée par une grande distorsion de l'angle phi du résidu se situant trois résidus avant la proline. Celui-ci est décalé d'environ 50° par rapport à la zone α canonique. Cette distorsion entraîne une ouverture de l'hélice se traduisant par un motif en renflement plus ou moins marqué.

La seconde étape a été d'analyser la cassure de l'hélice TMH2. Dans la rhodopsine bovine, l'hélice TMH2 est cassée par un motif en renflement π localisé aux positions 2.56-2.57 avec les résidus GG. Ce motif n'est pas conservé au sein des RCPG. 80% des RCPG humains possèdent une proline pouvant être localisée à 3 positions distinctes : 2.58, 2.59 et 2.60. Une classification par ACP effectuée au laboratoire a regroupé les RCPG du génome humain en 3 groupes. Le groupe 1 est caractérisé par la présence de la proline en position 2.58 et comprend principalement des récepteurs des systèmes vasculaire et/ou immunitaire spécifiques des vertébrés. Ce groupe de récepteurs a donc pris de l'ampleur tardivement au cours de l'évolution. Ce groupe comprend notamment nos récepteurs d'intérêt : AT1 et AT2.

Un criblage de la PDB a été effectué pour rechercher les motifs structuraux de cassure compatibles avec le changement d'orientation observé dans l'hélice 2 de la rhodopsine. Nous avons montré que 2 structures locales différentes étaient possibles. La première structure locale correspond à un renflement π . Cette structure est compatible avec la présence d'une proline aux positions 2.59 ou 2.60 mais pas en position 2.58. Dans ces conformations en renflement, on observe une conformation α_1 pour le résidu situé trois positions en amont de la proline. La deuxième structure locale correspond à une cassure proline en coude, comme on peut en observer dans des hélices continues. Cette structure est possible lorsque la proline est en position 2.58. Dans ce cas, il existe une délétion d'un résidu dans le coude de la cassure par rapport à la structure en renflement.

Cette approche bioinformatique permet le développement de modèles structuraux pour les récepteurs du groupe I, caractérisés par une proline en position 2.58. Un récepteur de ce groupe d'intérêt majeur est le récepteur de chimiokine, CCR5. Ce récepteur, notamment connu pour être le co-récepteur du virus du SIDA, est donc le récepteur de chimiokine le plus étudié [132]. Les ligands de CCR5 sont des chimiokines de type CC qui sont responsables du chimiotactisme des monocytes, des lymphocytes et des éosinophiles. Les ligands principaux de CCR5 sont RANTES, MIP-1 α , et MIP-1 β [133]. La modélisation moléculaire a permis de développer un modèle d'interaction entre CCR5 et son ligand RANTES. Nous sommes en train de valider ce modèle par une approche expérimentale en collaboration avec l'équipe du Dr M. Mellado du département d'Immunologie et d'Oncologie du Centre National des Biotechnologies de Madrid en Espagne.

L'analyse détaillée des propriétés du groupe I de notre classification des RCPG sera développée. Une analyse des corrélations de séquences de ce groupe sera effectuée. Cette analyse des résidus corrélés permettra de déterminer leur organisation. Nous pourrons notamment déterminer l'organisation tridimensionnelle des résidus corrélés avec la proline en position 2.58. Cette étude permettra de déterminer les interconnexions entre les résidus conservés au sein des RCPG de classe A et les résidus spécifiques du groupe I et sera poursuivie à différents niveaux hiérarchiques. Cela permettra notamment de déterminer le réseau de résidus impliqués dans la transduction du signal et l'activation du récepteur, ce qui nous fournira des indices quant au mécanisme d'action de ces récepteurs.

La structure tridimensionnelle de l'hélice TMH2 sera validée par RMN sur des peptides basés sur la séquence des récepteurs du groupe I avec la proline en position 2.58. Cette validation par RMN se fera dans des systèmes modèles mimant l'environnement de la membrane plasmique (solvants organiques et micelles). Ceci se fera en collaboration avec

l'équipe du Dr I. Milazzo du laboratoire de Chimie Organique et Biologie Structurale (UMR CNRS 6014) de l'Université de Rouen.

Toutes ces données nous permettront finalement de raffiner nos modèles structuraux et d'obtenir une meilleure compréhension du mécanisme d'activation et de transduction du signal. Nous pourrions proposer un modèle d'activation des récepteurs du groupe I qui sera validé expérimentalement au sein du laboratoire en collaboration avec l'équipe du Dr Daniel Henrion. Enfin, dans le cas du récepteur AT1, nous pourrions aussi envisager une étude de « drug-design ». Les molécules trouvées pourront être testées *in vitro* et *in vivo* par l'équipe du Dr. Daniel Henrion. Cette collaboration apportera une meilleure connaissance des voies de transduction impliquées dans la réponse microvasculaire à la pression et au débit.

LISTE DES ABREVIATIONS

Å : Angstrom
ACE : Angiotensin Conversion Enzyme
ACP : Analyse par Composante Principale
AMPc : Adénosine Monophosphate Cyclique
DSSP : Dictionary of Secondary Structure of Proteins
GABA : gamma-aminobutyric acid
GDP : Guanosine Diphosphate
GPCRdb : G-Protein-Coupled receptors Database
GTP : Guanosine Triphosphate
kD : kiloDalton
LCD : Langage de Contrôle de Données
LDD : langage de définition de Données
LMD : Langage de Manipulation de Données
MCD : modèle Conceptuel de Données
PDB : protein Data Bank
RAA : Système Rénine Angiotensine Aldostérone
RCPG : récepteurs Couplés aux Protéines G
RGR : Retinal G protein-coupled Receptor
RGS : Regulator of G-protein Signaling
RMN : Résonance Magnétique Nucléaire
SGBD : Système de Gestion de Base de Données
SIDA : Syndrome d'immuno-Déficience Acquise
SNC : Système Nerveux Central
SPASM : Spatial Arrangements of Side chains and Main chains
SQL : Structure Query language
TMH : Hélice Transmembranaire

ABREVIATIONS DES RECEPTEURS

ADOR : Adenosine Receptor
ADR : Adrenergic Receptor
AGTR/ATR : Angiotensin Receptors
APJR : APJ (apelin) Receptor
AT1 : Récepteur de l'Angiotensine 1
AT2 : Récepteur de l'Angiotensine 2
AVPR : Arginine Vasopressin Receptor
BDKR : Bradikinin Receptor
BRS3 : bombesin Receptor Subtype 3
CASR : Calcium-Sensing Receptor
CCK : Cholecystokinin receptor
CCR : CC chemokine Receptor
CHRM : Cholinergic Receptor, Muscarinic

CNR : Cannabinoid Receptor
 CRH : Corticotropin-Releasing Hormone
 CXCR : CXC chemokine Receptor
 CYSLT : Cysteinyl Leukotriene Receptor
 DRD : Dopamine D Receptor
 EDGR : Endothelial Differentiation G protein-coupled Receptor
 EDNR : Endothelin Receptor
 ETBRLP : endothelin Type B Receptor Like-protein
 F2R : Coagulation Factor II (thrombin) Receptor
 FPR : Formyl Peptide Receptor
 FSH : Follicle-Stimulating Hormone
 FSHR : Follicle-Stimulating Hormone Receptor
 GALR : Galanin Receptor
 GHSR : Growth Hormone Secretagogue Receptor
 GNRHR : Gonadotropin-Releasing Hormone receptor
 GRPR : Gastrin-Releasing Peptide Receptor
 HCRTR : Hypocretin Receptor
 HRH : Hypothalamic Releasing Hormone receptor
 HTR : Thyroid Hormone receptor (recepteur de la serotonine)
 LGR : leucine-rich-repeat-containing GPCR
 LHCGR : Luteinizing Hormone/Choriogonadotropin Receptor
 MCH : Melanin Concentrating Hormone Receptor
 MCR : Melanocortin Receptor
 MECA : Melanocortin-Endoglin-Cannabinoid-Adenosin
 MRG : Mas Related Gene receptor
 MTNR : Melatonin Receptor
 NMBR : Neuromedin B Receptor
 NPFF : neuropeptide FF Receptor
 NPYR : Neuropeptide Y receptor
 NTSR : neurotensin Receptor
 OPN1LW : Opsin 1 (cone pigments) Long-Wave-sensitive receptor
 OPN1MW : Opsin 1 (cone pigments) Medium-Wave-sensitive receptor
 OPN1SW : Opsin 1 (cone pigments) Short-Wave-sensitive receptor
 OPN3 : Encephalopsin (panopsin) receptor
 OPN4 : Melanopsin receptor
 OPR : opioid Receptor
 OXTR : Oxytocin Receptor
 RHO : rhodopsine
 RRH : Visual pigment-like receptor peropsin
 SOG : Somatostatin/Opioid/Galanin Receptors
 SSTR : Somatostatin Receptor
 TACR : Tachykinin Receptor
 TAR/TAAR : Trace Amine Receptor/Trace Amine Associated Receptor
 TAS : Taste receptor
 TRHR : Thyrotropin-Releasing Hormone Receptor
 TSH : Thyroid-Stimulating Hormone
 TSHR : Thyroid-Stimulating Hormone Receptor

REFERENCES

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
2. Venter, J.C., et al., *The sequence of the human genome*. Science, 2001. **291**(5507): p. 1304-51.
3. Gether, U., *Uncovering molecular mechanisms involved in activation of G protein-coupled receptors*. Endocr Rev, 2000. **21**(1): p. 90-113.
4. Flower, D.R., *Modelling G-protein-coupled receptors for drug design*. Biochim Biophys Acta, 1999. **1422**(3): p. 207-34.
5. Filmore, D., *it's a GPCR world*, in *Modern Drug Discovery*. 2004, American Chemical Society: 1. p. 24-28.
6. Kusserow, H. and T. Unger, *Vasoactive peptides, their receptors and drug development*. Basic Clin Pharmacol Toxicol, 2004. **94**(1): p. 5-12.
7. Chiu, A.T., et al., *Identification of angiotensin II receptor subtypes*. Biochem Biophys Res Commun, 1989. **165**(1): p. 196-203.
8. Timmermans, P.B., et al., *Angiotensin II receptor subtypes*. Am J Hypertens, 1992. **5**(6 Pt 1): p. 406-10.
9. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor*. Science, 2000. **289**(5480): p. 739-45.
10. Cherezov, V., et al., *High-Resolution Crystal Structure of an Engineered Human β_2 -Adrenergic G Protein Coupled Receptor*. Science, 2007.
11. Rasmussen, S.G., et al., *Crystal structure of the human β_2 adrenergic G-protein-coupled receptor*. Nature, 2007.
12. Ballesteros, *Integrated methods for the construction of three dimensional models and computational probing of structure-function relations in G-protein coupled receptors*. Methods Neurosci., 1995. **25**: p. 366-428.
13. Liebmann, C., *G protein-coupled receptors and their signaling pathways: classical therapeutic targets susceptible to novel therapeutic concepts*. Curr Pharm Des, 2004. **10**(16): p. 1937-58.
14. Okuno, Y., et al., *GLIDA: GPCR-ligand database for chemical genomic drug discovery*. Nucleic Acids Res, 2006. **34**(Database issue): p. D673-7.
15. Bockaert, J. and J.P. Pin, *Molecular tinkering of G protein-coupled receptors: an evolutionary success*. Embo J, 1999. **18**(7): p. 1723-9.
16. Bockaert, J. and J.P. Pin, *[Use of a G-protein-coupled receptor to communicate. An evolutionary success]*. C R Acad Sci III, 1998. **321**(7): p. 529-51.
17. Yeagle, P.L. and A.D. Albert, *A conformational trigger for activation of a G protein by a G protein-coupled receptor*. Biochemistry, 2003. **42**(6): p. 1365-8.
18. Vassilatis, D.K., et al., *The G protein-coupled receptor repertoires of human and mouse*. Proc Natl Acad Sci U S A, 2003. **100**(8): p. 4903-8.
19. *The Universal Protein Resource (UniProt)*. Nucleic Acids Res, 2007. **35**(Database issue): p. D193-7.
20. Horn, F., et al., *GPCRDB: an information system for G protein-coupled receptors*. Nucleic Acids Res, 1998. **26**(1): p. 275-9.
21. Kolakowski, L.F., Jr., *GCRDb: a G-protein-coupled receptor database*. Receptors Channels, 1994. **2**(1): p. 1-7.
22. Joost, P. and A. Methner, *Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands*. Genome Biol, 2002. **3**(11): p. RESEARCH0063.

23. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees*. Mol Biol Evol, 1987. **4**(4): p. 406-25.
24. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints*. Mol Pharmacol, 2003. **63**(6): p. 1256-72.
25. Ishihara, T., et al., *Molecular cloning and expression of a cDNA encoding the secretin receptor*. Embo J, 1991. **10**(7): p. 1635-41.
26. Chabbert, D.T.a.M., *Etude des récepteurs couplés aux protéines G par Analyse en Composantes Principales*. 2006.
27. Baldwin, J.M., *The probable arrangement of the helices in G protein-coupled receptors*. Embo J, 1993. **12**(4): p. 1693-703.
28. Okada, T., A. Terakita, and Y. Shichida, [*Structure-function relationship in G protein-coupled receptors deduced from crystal structure of rhodopsin*]. Tanpakushitsu Kakusan Koso, 2002. **47**(8 Suppl): p. 1123-30.
29. Teller, D.C., et al., *Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs)*. Biochemistry, 2001. **40**(26): p. 7761-72.
30. Mirzadegan, T., et al., *Sequence analyses of G-protein-coupled receptors: similarities to rhodopsin*. Biochemistry, 2003. **42**(10): p. 2759-67.
31. Spiegel, A.M., *Mutations in G proteins and G protein-coupled receptors in endocrine disease*. J Clin Endocrinol Metab, 1996. **81**(7): p. 2434-42.
32. O'Neil, K.T. and W.F. DeGrado, *A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids*. Science, 1990. **250**(4981): p. 646-51.
33. Monne, M., M. Hermansson, and G. von Heijne, *A turn propensity scale for transmembrane helices*. J Mol Biol, 1999. **288**(1): p. 141-5.
34. Chakrabarti, P. and S. Chakrabarti, *C--H.O hydrogen bond involving proline residues in alpha-helices*. J Mol Biol, 1998. **284**(4): p. 867-73.
35. Deupi, X., et al., *Ser and Thr residues modulate the conformation of pro-kinked transmembrane alpha-helices*. Biophys J, 2004. **86**(1 Pt 1): p. 105-15.
36. Cordes, F.S., J.N. Bright, and M.S. Sansom, *Proline-induced distortions of transmembrane helices*. J Mol Biol, 2002. **323**(5): p. 951-60.
37. Li, J., et al., *Structure of bovine rhodopsin in a trigonal crystal form*. J Mol Biol, 2004. **343**(5): p. 1409-38.
38. Filipek, S., et al., *The crystallographic model of rhodopsin and its use in studies of other G protein-coupled receptors*. Annu Rev Biophys Biomol Struct, 2003. **32**: p. 375-97.
39. Ballesteros, J.A., L. Shi, and J.A. Javitch, *Structural mimicry in G protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structure-function analysis of rhodopsin-like receptors*. Mol Pharmacol, 2001. **60**(1): p. 1-19.
40. Ballesteros, J.A., et al., *Serine and threonine residues bend alpha-helices in the chi(1) = g(-) conformation*. Biophys J, 2000. **79**(5): p. 2754-60.
41. Lopez-Rodriguez, M.L., et al., *Design, synthesis and pharmacological evaluation of 5-hydroxytryptamine(1a) receptor ligands to explore the three-dimensional structure of the receptor*. Mol Pharmacol, 2002. **62**(1): p. 15-21.
42. Ballesteros, J.A., et al., *Activation of the beta 2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6*. J Biol Chem, 2001. **276**(31): p. 29171-7.
43. Scheer, A., et al., *Constitutively active mutants of the alpha 1B-adrenergic receptor: role of highly conserved polar amino acids in receptor activation*. Embo J, 1996. **15**(14): p. 3566-78.

44. Farrens, D.L., et al., *Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin*. *Science*, 1996. **274**(5288): p. 768-70.
45. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, *Stereochemistry of polypeptide chain configurations*. *J Mol Biol*, 1963. **7**: p. 95-9.
46. Kahn, *Defining the helix axis*. *Computers & Chemistry*, 1989. **13**(3): p. 185-189.
47. Engel, D.E. and W.F. DeGrado, *Alpha-alpha linking motifs and interhelical orientations*. *Proteins*, 2005. **61**(2): p. 325-37.
48. Moore, R.E., M.K. Young, and T.D. Lee, *Qscore: an algorithm for evaluating SEQUEST database search results*. *J Am Soc Mass Spectrom*, 2002. **13**(4): p. 378-86.
49. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. *Biopolymers*, 1983. **22**(12): p. 2577-637.
50. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures*. *J Mol Biol*, 1977. **112**(3): p. 535-42.
51. Hobohm, U., et al., *Selection of representative protein data sets*. *Protein Sci*, 1992. **1**(3): p. 409-17.
52. Hobohm, U. and C. Sander, *Enlarged representative set of protein structures*. *Protein Sci*, 1994. **3**(3): p. 522-4.
53. Kleywegt, G.J., *Recognition of spatial motifs in protein structures*. *J Mol Biol*, 1999. **285**(4): p. 1887-97.
54. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints*. *J Mol Biol*, 1993. **234**(3): p. 779-815.
55. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures*. *J Mol Biol*, 1995. **247**(4): p. 536-40.
56. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures*. *Structure*, 1997. **5**(8): p. 1093-108.
57. Chou, P.Y. and G.D. Fasman, *Prediction of protein conformation*. *Biochemistry*, 1974. **13**(2): p. 222-45.
58. Chou, P.Y. and G.D. Fasman, *Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins*. *Biochemistry*, 1974. **13**(2): p. 211-22.
59. Chou, P.Y. and G.D. Fasman, *Prediction of the secondary structure of proteins from their amino acid sequence*. *Adv Enzymol Relat Areas Mol Biol*, 1978. **47**: p. 45-148.
60. Garnier, J., D.J. Osguthorpe, and B. Robson, *Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins*. *J Mol Biol*, 1978. **120**(1): p. 97-120.
61. Albrecht, M., et al., *Simple consensus procedures are effective and sufficient in secondary structure prediction*. *Protein Eng*, 2003. **16**(7): p. 459-62.
62. Ouali, M. and R.D. King, *Cascaded multiple classifiers for secondary structure prediction*. *Protein Sci*, 2000. **9**(6): p. 1162-76.
63. Kabsch, W. and C. Sander, *How good are predictions of protein secondary structure?* *FEBS Lett*, 1983. **155**(2): p. 179-82.
64. Cuff, J.A. and G.J. Barton, *Evaluation and improvement of multiple sequence methods for protein secondary structure prediction*. *Proteins*, 1999. **34**(4): p. 508-19.
65. Cuff, J.A. and G.J. Barton, *Application of multiple sequence alignment profiles to improve protein secondary structure prediction*. *Proteins*, 2000. **40**(3): p. 502-11.
66. Frishman, D. and P. Argos, *Seventy-five percent accuracy in protein secondary structure prediction*. *Proteins*, 1997. **27**(3): p. 329-35.
67. Rost, B. and C. Sander, *Improved prediction of protein secondary structure by use of sequence profiles and neural networks*. *Proc Natl Acad Sci U S A*, 1993. **90**(16): p. 7558-62.

68. Salamov, A.A. and V.V. Solovyev, *Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments*. J Mol Biol, 1995. **247**(1): p. 11-5.
69. Wilson, C.L., et al., *Improved prediction for N-termini of alpha-helices using empirical information*. Proteins, 2004. **57**(2): p. 322-30.
70. Brazhnikov, E.V. and A.V. Efimov, [*Structure of alpha-spiral hairpins with short connections in globular proteins*]. Mol Biol (Mosk), 2001. **35**(1): p. 100-8.
71. Lahr, S.J., et al., *Analysis and design of turns in alpha-helical hairpins*. J Mol Biol, 2005. **346**(5): p. 1441-54.
72. Kahn, P.C., *Defining the axis of a helix*. Computers Chem, 1989. **13**: p. 185-189.
73. Eisenberg, D., et al., *Analysis of membrane and surface protein sequences with the hydrophobic moment plot*. J Mol Biol, 1984. **179**(1): p. 125-42.
74. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein*. J Mol Biol, 1982. **157**(1): p. 105-32.
75. Hopp, T.P. and K.R. Woods, *Prediction of protein antigenic determinants from amino acid sequences*. Proc Natl Acad Sci U S A, 1981. **78**(6): p. 3824-8.
76. Hubbard, S.J., R.J. Beynon, and J.M. Thornton, *Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures*. Protein Eng, 1998. **11**(5): p. 349-59.
77. Lee, B. and F.M. Richards, *The interpretation of protein structures: estimation of static accessibility*. J Mol Biol, 1971. **55**(3): p. 379-400.
78. McDonald, I.K. and J.M. Thornton, *Satisfying hydrogen bonding potential in proteins*. J Mol Biol, 1994. **238**(5): p. 777-93.
79. Heine, A., et al., *Crystal structure of O-acetylserine sulphydrylase (TM0665) from Thermotoga maritima at 1.8 A resolution*. Proteins, 2004. **56**(2): p. 387-91.
80. Johnson, K.A., et al., *Crystal structure of the 28 kDa glutathione S-transferase from Schistosoma haematobium*. Biochemistry, 2003. **42**(34): p. 10084-94.
81. Polekhina, G., et al., *Molecular basis of glutathione synthetase deficiency and a rare gene permutation event*. Embo J, 1999. **18**(12): p. 3204-13.
82. Miller, D.J., et al., *Crystal complexes of a predicted S-adenosylmethionine-dependent methyltransferase reveal a typical AdoMet binding domain and a substrate recognition domain*. Protein Sci, 2003. **12**(7): p. 1432-42.
83. Zhong, W., et al., *Assembly of an oxo-zirconium(IV) cluster in a protein cleft*. Angew Chem Int Ed Engl, 2004. **43**(44): p. 5914-8.
84. Wilson, C.L., S.J. Hubbard, and A.J. Doig, *A critical assessment of the secondary structure alpha-helices and their termini in proteins*. Protein Eng, 2002. **15**(7): p. 545-54.
85. Ramachandran, G.N., C.M. Venkatachalam, and S. Krimm, *Stereochemical criteria for polypeptide and protein chain conformations. 3. Helical and hydrogen-bonded polypeptide chains*. Biophys J, 1966. **6**(6): p. 849-72.
86. Karplus, P.A., *Experimentally observed conformation-dependent geometry and hidden strain in proteins*. Protein Sci, 1996. **5**(7): p. 1406-20.
87. Ho, B.K. and R. Brasseur, *The Ramachandran plots of glycine and pre-proline*. BMC Struct Biol, 2005. **5**: p. 14.
88. Lovell, S.C., et al., *Structure validation by C α geometry: phi,psi and C β deviation*. Proteins, 2003. **50**(3): p. 437-50.
89. Adzhubei, A.A. and M.J. Sternberg, *Left-handed polyproline II helices commonly occur in globular proteins*. J Mol Biol, 1993. **229**(2): p. 472-93.
90. Cubellis, M.V., et al., *Properties of polyproline II, a secondary structure element implicated in protein-protein interactions*. Proteins, 2005. **58**(4): p. 880-92.

91. Schellman, C., *The α L-conformation at the ends of helices*. in *Protein Folding*. Elsevier/North Holland Biochemical Press ed. 1980, Amsterdam: Jaenicke, R. 53-61.
92. Aurora, R., R. Srinivasan, and G.D. Rose, *Rules for alpha-helix termination by glycine*. *Science*, 1994. **264**(5162): p. 1126-30.
93. Cubellis, M.V., F. Cailliez, and S.C. Lovell, *Secondary structure assignment that accurately reflects physical and evolutionary characteristics*. *BMC Bioinformatics*, 2005. **6 Suppl 4**: p. S8.
94. Richardson, J.S. and D.C. Richardson, *Amino acid preferences for specific locations at the ends of alpha helices*. *Science*, 1988. **240**(4859): p. 1648-52.
95. Kumar, S. and M. Bansal, *Dissecting alpha-helices: position-specific analysis of alpha-helices in globular proteins*. *Proteins*, 1998. **31**(4): p. 460-76.
96. Engel, D.E. and W.F. DeGrado, *Amino acid propensities are position-dependent throughout the length of alpha-helices*. *J Mol Biol*, 2004. **337**(5): p. 1195-205.
97. Gunasekaran, K., et al., *Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals*. *J Mol Biol*, 1998. **275**(5): p. 917-32.
98. Kumar, S. and M. Bansal, *Geometrical and sequence characteristics of alpha-helices in globular proteins*. *Biophys J*, 1998. **75**(4): p. 1935-44.
99. Sharma, V., et al., *Structure of isocitrate lyase, a persistence factor of Mycobacterium tuberculosis*. *Nat Struct Biol*, 2000. **7**(8): p. 663-8.
100. Anderson, R.J., et al., *Main-chain conformational tendencies of amino acids*. *Proteins*, 2005. **60**(4): p. 679-89.
101. Eswar, N. and C. Ramakrishnan, *Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures*. *Protein Eng*, 2000. **13**(4): p. 227-38.
102. Frishman, D. and P. Argos, *Knowledge-based protein secondary structure assignment*. *Proteins*, 1995. **23**(4): p. 566-79.
103. Labesse, G., et al., *P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins*. *Comput Appl Biosci*, 1997. **13**(3): p. 291-5.
104. Levitt, M. and J. Greer, *Automatic identification of secondary structure in globular proteins*. *J Mol Biol*, 1977. **114**(2): p. 181-239.
105. Martin, J., et al., *Protein secondary structure assignment revisited: a detailed analysis of different assignment methods*. *BMC Struct Biol*, 2005. **5**: p. 17.
106. Richards, F.M. and C.E. Kundrot, *Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure*. *Proteins*, 1988. **3**(2): p. 71-84.
107. Barlow, D.J. and J.M. Thornton, *Helix geometry in proteins*. *J Mol Biol*, 1988. **201**(3): p. 601-19.
108. Prieto, J. and L. Serrano, *C-capping and helix stability: the Pro C-capping motif*. *J Mol Biol*, 1997. **274**(2): p. 276-88.
109. Aurora, R. and G.D. Rose, *Helix capping*. *Protein Sci*, 1998. **7**(1): p. 21-38.
110. Doig, A.J., et al., *Structures of N-termini of helices in proteins*. *Protein Sci*, 1997. **6**(1): p. 147-55.
111. Eswar, N. and C. Ramakrishnan, *Secondary structures without backbone: an analysis of backbone mimicry by polar side chains in protein structures*. *Protein Eng*, 1999. **12**(6): p. 447-55.
112. Sealfon, S.C., et al., *Related contribution of specific helix 2 and 7 residues to conformational activation of the serotonin 5-HT_{2A} receptor*. *J Biol Chem*, 1995. **270**(28): p. 16683-8.
113. MacArthur, M.W. and J.M. Thornton, *Influence of proline residues on protein conformation*. *J Mol Biol*, 1991. **218**(2): p. 397-412.
114. Holm, L. and C. Sander, *Removing near-neighbour redundancy from large protein sequence collections*. *Bioinformatics*, 1998. **14**(5): p. 423-9.

- tel-00346950, version 1 - 12 Dec 2008
115. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
 116. Eddy, S.R., *Profile hidden Markov models*. Bioinformatics, 1998. **14**(9): p. 755-63.
 117. Kumar, S., K. Tamura, and M. Nei, *MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment*. Brief Bioinform, 2004. **5**(2): p. 150-63.
 118. Fredriksson, R. and H.B. Schioth, *The repertoire of G-protein-coupled receptors in fully sequenced genomes*. Mol Pharmacol, 2005. **67**(5): p. 1414-25.
 119. Okada, T., et al., *The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure*. J Mol Biol, 2004. **342**(2): p. 571-83.
 120. Okada, T., et al., *Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography*. Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5982-7.
 121. Nakamichi, H. and T. Okada, *Local peptide movement in the photoreaction intermediate of rhodopsin*. Proc Natl Acad Sci U S A, 2006. **103**(34): p. 12729-34.
 122. Standfuss, J., et al., *Crystal structure of a thermally stable rhodopsin mutant*. J Mol Biol, 2007. **372**(5): p. 1179-88.
 123. Salom, D., et al., *Crystal structure of a photoactivated deprotonated intermediate of rhodopsin*. Proc Natl Acad Sci U S A, 2006. **103**(44): p. 16123-8.
 124. Aloy, P., et al., *Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics*. Protein Sci, 2002. **11**(5): p. 1101-16.
 125. Devos, D. and A. Valencia, *Practical limits of function prediction*. Proteins, 2000. **41**(1): p. 98-107.
 126. Govaerts, C., et al., *The TXP motif in the second transmembrane helix of CCR5. A structural determinant of chemokine-induced activation*. J Biol Chem, 2001. **276**(16): p. 13217-25.
 127. Yohannan, S., et al., *The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors*. Proc Natl Acad Sci U S A, 2004. **101**(4): p. 959-63.
 128. Fodje, M.N. and S. Al-Karadaghi, *Occurrence, conformational features and amino acid propensities for the pi-helix*. Protein Eng, 2002. **15**(5): p. 353-8.
 129. Rajashankar, K.R. and S. Ramakumar, *Pi-turns in proteins and peptides: Classification, conformation, occurrence, hydration and sequence*. Protein Sci, 1996. **5**(5): p. 932-46.
 130. Cartailier, J.P. and H. Luecke, *Structural and functional characterization of pi bulges and other short intrahelical deformations*. Structure, 2004. **12**(1): p. 133-44.
 131. Reis, R.I., et al., *Participation of transmembrane proline 82 in angiotensin II AT1 receptor signal transduction*. Regul Pept, 2007. **140**(1-2): p. 32-6.
 132. Berger, E.A., P.M. Murphy, and J.M. Farber, *Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease*. Annu Rev Immunol, 1999. **17**: p. 657-700.
 133. Blanpain, C., et al., *CCR5 binds multiple CC-chemokines: MCP-3 acts as a natural antagonist*. Blood, 1999. **94**(6): p. 1899-905.

RESUME :

Nos récepteurs d'intérêt, les récepteurs à l'angiotensine AT1 et AT2, appartiennent à la classe A de la grande famille des récepteurs couplés aux protéines G (RCPG). Jusqu'à très récemment, la rhodopsine bovine était le seul RCPG dont la structure cristallographique était résolue. La structure de la rhodopsine est employée couramment comme modèle en modélisation par homologie des RCPG de classe A. La structure de la rhodopsine est constituée de sept hélices transmembranaires. La plupart de ces hélices ne sont pas droites, mais cassées ou incurvées. Pour comprendre quels sont les motifs structuraux possibles pour les cassures d'hélices, nous avons réalisé une étude exhaustive des motifs d'hélices cassées au niveau d'un seul résidu de jonction (motif HXH) grâce à une base de données de structures d'hélices cassées développée localement. Les résultats montrent que le résidu de jonction n'admet qu'un nombre limité de conformations conduisant à la classification de ces cassures en six motifs bien distincts. Un de ces motifs correspond à une cassure au niveau d'un renflement. Ce motif se retrouve dans l'hélice transmembranaire 2 (TMH2) de la rhodopsine où une cassure se fait au niveau d'un motif GG correspondant à un renflement π . Ce motif GG, situé aux positions 2.56-2.57, n'est pas conservé parmi les RCPG mais une proline est fréquemment observée aux positions 2.58, 2.59 ou 2.60. Nos récepteurs d'intérêt AT1 et AT2 possèdent une proline à la position 2.58. L'étude de l'évolution de l'hélice transmembranaire 2 au sein de la famille des RCPG suggère fortement que la position en 2.58 correspond à une délétion d'un résidu au niveau de la cassure. Ceci est confirmé par des analyses structurales de la Protein Data bank. Ces résultats indiquent que la structure de la rhodopsine peut être utilisée directement pour modéliser l'hélice 2 lorsque la proline est en position 2.59 ou 2.60 (renflement π). Lorsque la proline est en position 2.58, la rhodopsine peut aussi être utilisée comme modèle structural à condition de prendre en compte la délétion d'un résidu au niveau du renflement, pour obtenir une cassure proline en coude classique.

Mots-clés : bioinformatique, criblage 3D, modélisation moléculaire, RCPG, hélices cassées, proline

ABSTRACT :

Our receptors of interest, the angiotensin receptors AT1 and AT2, belong to the family of class A G protein coupled receptors (GPCR). Up to very recently, bovine rhodopsin has been the only GPCR whose crystal structure was resolved. This structure is widely used as template for homology modelling of class A GPCR. The structure of rhodopsin consists of a bundle of seven transmembrane helices, common to all GPCRs. Most of these helices are not straight, but kinked or bent. For a better understanding of the structural motifs related to kinked helices, we analyzed the properties of helix-X-helix motifs in which two α -helices are linked by a single residue in a home-built structural database of HXH motifs. The linker residue can be classified in six distinct motifs. One of these motifs corresponds to π bulge. This GG motif, located at positions 2.56-2.57, is not conserved among GPCRs, but a proline is frequently observed in this helix, either at positions 2.58, 2.59 or 2.60. The receptors AT1 and AT2 possess a proline at position 2.58. The analysis of the evolution of the TMH2 proline positioning within the GPCR family strongly suggests that proline at position 2.58 corresponds to the deletion of one residue in the bulge elbow. This assumption is corroborated by 3D data mining of the Protein data bank. These results indicate that the rhodopsin structure can be directly used as a template to model TMH2 when proline is located at position 2.59 or 2.60. When proline is located at position 2.58, the rhodopsin structure can also be used, but the deletion of one residue in the elbow is taken into account. In that case, the helix bulge of rhodopsin can be mimicked by a typical "proline" kink.

Mots-clés : bioinformatics, 3D data mining, molecular modelling, GPCR, helix kink, bulge, proline