Université de Montréal

Recherche d'une empreinte phylogénétique reliée à la fonctionnalité des récepteurs couplés aux protéines G

par

Guy Larochelle

Département de Biochimie, Université de Montréal

Faculté de Médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maîtrise ès sciences (M.sc.)
en Bio-informatique

Mai, 2007

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Recherche d'une empreinte phylogénétique reliée à la fonctionnalité des récepteurs couplés aux protéines G

présenté par :
Guy Larochelle

a été évalué par un jury composé des personnes suivantes :

Hervé Philippe, directeur de recherche
Michel Bouvier, co-directeur
Sylvie Hamel, membre du jury
Nikolaus Heveker, membre du jury

Résumé

Les récepteurs avec sept domaines transmembranaires qui portent le nom de récepteurs couplés aux protéines G (RCPG) peuvent former des hétérodimères. Cette propriété a peut avoir une implication fonctionnelle profonde, telle que la modification de leur spécificité face aux ligands. Puisqu'il est difficile et qu'il coûte cher de déterminer à grande échelle, de façon biochimique, si deux récepteurs peuvent hétérodimériser, il serait utile de développer une méthode de prédiction *in silico*. Pour que deux récepteurs forment un hétérodimère, ceux-ci doivent être à proximité l'un de l'autre. En d'autres termes, leurs gènes doivent être transcrits dans la même cellule à une période de temps similaire; les deux ARNm correspondants doivent être exportés du noyau et doivent être traduits dans la même partie de la cellule dans une période de temps semblable.

Nous faisons l'hypothèse qu'il existe des signaux au niveau de la séquence d'ADN (au niveau du promoteur ou de la séquence codante) et qu'ils sont spécifiques aux gènes des RCPG capables de former des hétérodimères. Pour trouver de telles signatures, nous proposons d'employer une méthode qui est inspirée de l'empreinte phylogénétique. Pour chaque RCPG paralogue nous allons analyser les régions codantes et promotrices chez 6 espèces de vertébrés différentes. Nous allons rechercher des motifs de taille variant entre 6 et 15 nucléotides qui sont conservés à travers toutes ces espèces. Selon notre hypothèse, des paires de paralogues qui partagent un excès de motifs devraient être exprimées et traduites dans un cadre spatio-temporel semblable. Cette approche a été appliquée à 42 récepteurs, dont certains d'entre eux sont connus pour former des hétérodimères.

Les résultats que nous avons obtenus, pour les régions promotrices, ne sont pas du tout concluants. Par contre, ceux pour la région codante peuvent nous laisser entrevoir une nouvelle piste de recherche. En effet, il semble y avoir de l'information pertinente qui serait susceptible d'être impliquée dans l'hétérodimérisation des RCPG.

Mots-clés : RCPG, empreinte phylogénétique, motifs, promoteur, régulation, génomique comparative, hétérodimères, traduction

Abstract

The receptors with seven transmembranes domains called G-protein coupled receptor (GPCR) can form heterodimers. This property has profound functional implication, such as the modification of ligand specificity. Since it is difficult and expensive to biochemically determine at a large scale whether two receptors can heterodimerize, an *in silico* prediction method would be helpful. To heterodimerize, two receptors must be in physical proximity. In other words, their genes must be transcribed in the same cell at a similar time period; the two corresponding mRNAs should be exported from the nucleus at a similar period and should be translated in the same part of the endoplasmic reticulum at a similar time period.

We make the hypothesis that there are signals in the DNA sequence (either at the level of the promoter or at the level of the mRNA). Those are specific to GPCR genes able to form heterodimers. To find such signatures, we propose to use a method which is inspired by phylogenetic footprinting. For each GPCR paralog we will analyze the coding and non-coding mRNA sequences from 6 different vertebrate species. Motifs of size from 6 to 15 nt which are conserved across all the species will be searched. According to our hypotheses, pairs of paralogs sharing an excess of motifs should be expressed and translated in a similar spatio-temporal framework. This approach was applied to 42 receptors, some of which being known to form heterodimers.

The results which we obtained for the promoter regions are not conclusive. On the other hand, those for the coding region, can let to us foresee a new avenue of research. Indeed, it seems to have relevant information which would be likely to be implied in the heterodimerisation of the RCPG.

Keywords : GPCR, phylogenetic footprinting, motifs, promoter, regulation, comparative genomic, heterodimers, translation

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ADN	Acide désoxyribonucléique
AMIN	Serotonine/dopamine/adrénergique/trace amines
ARN	Acide ribonucléique
ARNm	ARN messager
ARNr	ARN ribosomal
ARNt	ARN de transfert
ATP	Adénosine triphosphate
BRET	Transfert d'énergie par résonance de bioluminescence
Ca^{++}	Ion calcique
CHEM	Chemokine
CP1	Facteur de transcription pour la globine γ
CPSF	Facteur de spécificité de clivage et de polyadénylation
DR	Répétition directe
EM	Expectation maximization
EPAC	Protéine interchangée directement lors de l'activation par l' AMPc
ER	Répétition renversée
FRET	Transfert d'énergie de fluorescence par résonance
GABA	Acide-γ-amino-butyrique
GDP	Guanosine diphosphate
GFP	Protéine fluorescente verte
GMPc	Guanosine monophosphate cyclique
GRK	Kinase des RCPG
GTP	Guanosine triphosphate
IP_3	Inositol 3-phosphate
IR	Répétition inversée
MECA	Mélanocortine/endogline/adénosine/cannabinoid
MEME	Multiple EM for motif elicitation
NJ	Neighbor-Joining
OOPS	Une occurrence par séquence

P_i	Phosphate i
PIP_2	Phosphatidyl inositol bi-phosphate
PKA	Protéine kinase dépendante de l'AMP_c
PKC	Protéine kinase dépendante du Ca^{++}
POLII	ARN polymérase II
POLIII	ARN polymérase III
Poly(A)	Polyadénylation
RCPG	Récepteurs couplés aux protéines G
A1	Adénosine A1
A2A	Adénosine A2A
AT1	Angiotensine 1
AT2	Angiotensine 2
B1	Adrénocepteur Beta-1
B2	Adrénocepteur Beta-2
CB1	Cannabinoid 1
CB2	Cannabinoid 2
CCR1	Chemokine 1
CCR2	Chemokine 2
CCR3	Chemokine 3
CCR4	Chemokine 4
CCR5	Chemokine 5
CCR6	Chemokine 6
CCR7	Chemokine 7
D1	Dopamine 1
D2	Dopamine 2
D3	Dopamine 3
D4	Dopamine 4
Delta	Opioid delta
H1	Histamine 1
H2	Histamine 2
H3	Histamine 3

H4	Histamine 4
Kappa	Opioid kappa
M1	Acétylcholine (muscarinic) 1
M2	Acétylcholine (muscarinic) 2
M3	Acétylcholine (muscarinic) 3
M4	Acétylcholine (muscarinic) 4
M5	Acétylcholine (muscarinic) 5
Mela1	Mélatonine 1a
Mela2	Mélatonine 1b
Meta1	Métabotropic glutamate 1
Meta5	Métabotropic glutamate 5
Mu	Opioid Mu
Soma1	Somatostatin 1
Soma2	Somatostatin 2
Soma3	Somatostatin 3
Soma4	Somatostatin 4
Soma5	Somatostatin 5
RDR	Répétition directe-inversée
RE	Réticulum endoplasmique
SP1	Facteur de transcription à doigt de zinc
SOG	Somatostatine/opioid/galanine
TBP	Boîte TATA qui lie les protéines
TCM	Mélange de deux composants
TFIIB	Facteur de transcription B et II pour des gènes de classe II
TFIID	Facteur de transcription D et II pour des gènes de classe II
TFIIE	Facteur de transcription E et II pour des gènes de classe II
TFIIF	Facteur de transcription F et II pour des gènes de classe II
TFIIH	Facteur de transcription H et II pour des gènes de classe II
TSS	Site de départ de la transcription
UTR	Région non-traduite
ZOOPS	Zero ou une occurrence de motifs par groupe de séquences

Remerciements

J'aimerais remercier mon directeur Hervé Philippe et mon co-directeur Michel Bouvier pour l'aide précieuse qu'ils m'ont apportée tout au cours de ma maîtrise. Leurs connaissances, leur patience ainsi que leurs conseils ont beaucoup été appréciés. J'aimerais remercier aussi tous les membres du laboratoire pour les nombreuses discussions et conseils qui ont eu lieu lors des réunions. De plus, je voudrais remercier plus particulièrement Fabrice Baro et Olivier Jeffroy qui m'ont beaucoup aidé à résoudre certains problèmes informatiques. Finalement, je remercie toute ma famille : Claude, Diane et Marc pour leur support et leur patience.

Guy Larochelle

Chapitre 1

Introduction

De nos jours, avec la technologie qui avance de façon fulgurante, il est devenu beaucoup plus facile d'accéder à toutes sortes d'information et ce dans plusieurs domaines différents. Le domaine de la biologie moléculaire n'y échappe pas. Avec le nombre sans cesse grandissant de génomes séquencés, une idée a commencé à émerger, il y a quelques années, dans le domaine de la recherche scientifique. Pourquoi ne pas comparer les différents génomes entre eux et voir quelle information peut être déduite de ces analyses? Ce type d'approche se base sur l'idée que l'information qui est conservée à travers l'évolution (entre les espèces) doit nécessairement être importante.

Dans cette ligne de pensées, nous avons utilisé ce type d'approche pour mieux comprendre les mécanismes qui permettent de réguler l'expression des gènes. Nous savons pour l'instant que la régulation s'effectue, la plupart de temps, à l'aide d'éléments régulateurs présents en amont des gènes dans la séquence d'ADN. Mais puisqu'un humain possède environ 25 à 30 mille gènes et que chacun peut être régulé de plusieurs manières différentes et produire plusieurs protéines, il est important de cibler les recherches que nous voulons faire. En ce qui nous concerne, nos recherches se concentreront sur une famille de protéines appelée récepteurs couplés aux protéines G (RCPG) et qui suscite beaucoup d'intérêt de la part de la communauté scientifique puisque les protéines de cette famille sont impliquées dans des voies métaboliques vitales pour le bon fonctionnement des cellules chez l'être humain et plusieurs autres espèces. Pour preuve de leur importance, ces protéines représentent la cible d'environ 40 à 50 % des médicaments produits (Fredriksson et al., 2005). Notre compréhension du mécanisme de régulation de ces protéines est loin d'être complète. L'un des problèmes majeurs est que certains de ces récepteurs peuvent former des homodimères et/ou des hétérodimères ce qui peut modifier leur

spécificité face aux médicaments existants. De plus, certains autres semblent posséder une certaine sélectivité face au mécanisme d'hétérodimérisation. Par conséquent, il serait opportun d'avoir une méthode pouvant prédire si deux récepteurs sont susceptibles de dimériser.

Pour l'instant, il existe certaines méthodes biochimiques qui nous permettent d'accomplir cette tâche, mais elles sont longues et coûteuses. Il existe aussi beaucoup trop de possibilités à tester : environ 1 000 000 (1000 gènes X 1000 gènes). Notre but, ici, est de développer une méthode *in silico* qui permettra d'évaluer le niveau de susceptibilité de dimérisation entre deux récepteurs. Il est important de mentionner que notre projet consiste plus à une approche exploratrice qu'à un projet de type plus conventionnel. Pour y parvenir, nous avons essayé de développer une nouvelle manière d'aborder le problème. Au lieu de seulement étudier la régulation de la transcription comme dans Nikitenko et al. (2003), par exemple, nous avons voulu analyser la régulation de la traduction.

Pour mieux comprendre notre approche, il est impératif de définir quelques notions de biologie et de bio-informatique. Dans ce mémoire, nous aborderons les principaux concepts entourant le Dogme central de la biologie ainsi que ceux des RCPG (Chapitre 2). Ensuite, il y aura une description de ce qu'est la phylogénie et une comparaison des différentes méthodes et logiciels qui existent pour rechercher des motifs régulateurs dans des séquences d'ADN (Chapitre 3). Le chapitre suivant (Chapitre 4) sera consacré à une description plus approfondie de la méthode que nous avons employée. Par la suite, nous énumèrerons les différents résultats que nous avons obtenus (Chapitre 5) et pour terminer nous discuterons des résultats (Chapitre 6).

Chapitre 2

Notions biologiques

Afin de mieux comprendre le travail qui a été effectué durant notre recherche, il est important d'introduire certains concepts biologiques. Alors ce chapitre se veut une brève introduction sur le dogme central de la biologie et les récepteurs transmembranaires.

Dogme central de la biologie

Voici un schéma qui permet de résumer le dogme central de la biologie :

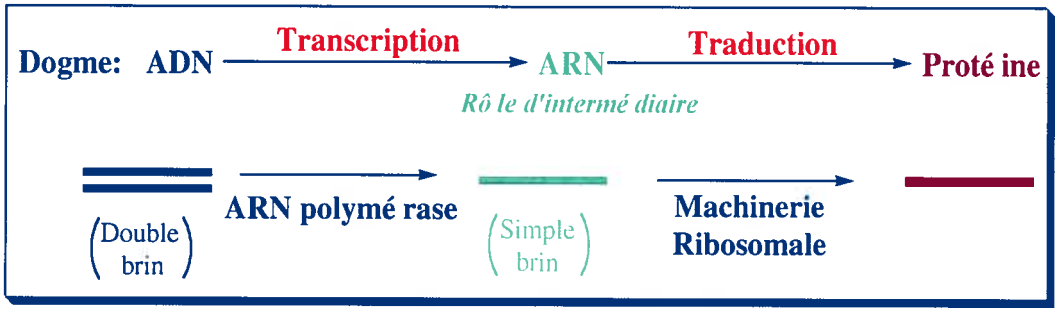

FIG. 2.1 : Dogme central [Notes de cours BCM1501 (Universtié de Montréal)]

On commence avec la transcription de l'ADN, qui est double brin, grâce à l'ARN polymérase et qui produira l'ARNm (simple brin). Ensuite, l'ARNm est traduit par la machinerie ribosomale pour produire la protéine correspondante.

2.1 Transcription

La transcription est la biosynthèse de l'ARNm à partir de l'ADN, qui sert de matrice, et s'effectue grâce à la complémentarité des bases. On décrit généralement une séquence d'ADN comme une suite de bases qui commence en 5' et se termine en 3' pour le brin codant. Tandis que pour l'autre brin, celui qui sera transcrit, c'est le contraire (3' vers 5'). Pour réguler cette biosynthèse, il existe plusieurs mécanismes complexes. La régulation se fait aussi bien au niveau de la compacité de l'ADN que

celui de la transcription. Par conséquent, une brève description de l'organisation de l'ADN est nécessaire.

2.1.1 Chromatine

De nos jours, il est bien connu que le noyau d'une cellule humaine est composé d'une multitude de protéines et de l'ADN, qui est fractionné en 23 paires de chromosomes et qui est condensé sous forme de chromatine (Johnson et al., 1998). Cette compacité est nécessaire puisque chacun des chromosomes contient entre 48 et 240 millions de paires de bases, ce qui équivaut à une longueur de 1,6 à 8,2 cm pour chaque brin. Donc, si l'on veut que l'ADN soit présent dans chaque cellule, il faut qu'elle puisse entrer dans cette dernière. Cette condensation est organisée de manière séquentielle et ordonnée (figure 2.2).

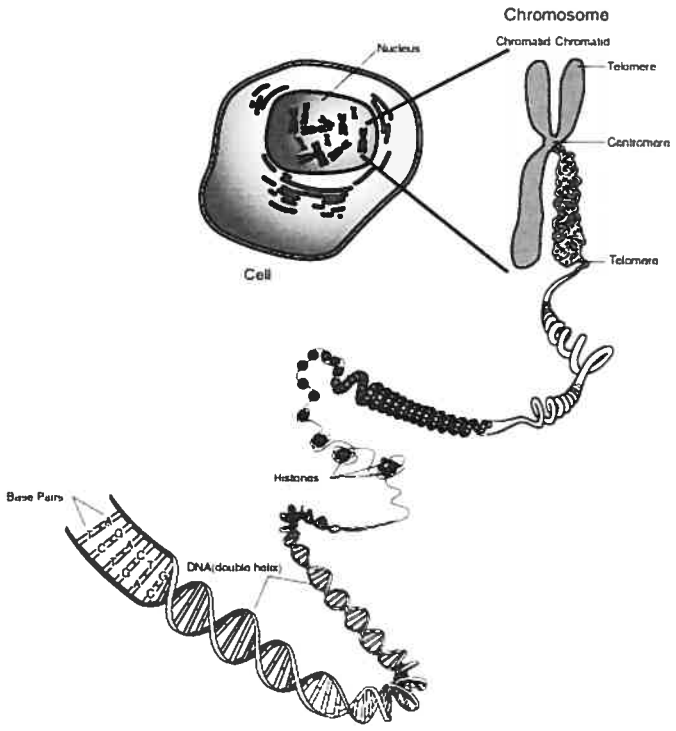

FIG. 2.2 : Condensation de l'ADN [http://fr.wikipedia.org/wiki/Chromosome]

Il existe deux types de chromatine : l'euchromatine et l'hétérochromatine. La différence entre ces deux types est que l'euchromatine est moins compacte que l'autre. Seuls les gènes qui sont situés au plus bas niveau de compacité (euchromatine) peuvent être transcrits.

La chromatine est composée d'ADN et de protéines. La protéine la plus répandue dans la chromatine est l'histone. Elle se divise en 5 classes : histones H1, H2A, H2B, H3 et H4. Chacune d'entre-elles est chargée positivement ce qui leur permet d'établir des liaisons ioniques avec les groupements phosphates de l'ADN qui sont chargés négativement. Par conséquent, l'ADN s'enroule autour des histones et cet enroulement est considéré comme le premier niveau de compacité de l'ADN que l'on nomme nucléosome. Il se compose d'un octamère d'histones (groupe de 8 histones) avec une séquence d'environ 200 paires de bases d'ADN. Le second niveau d'organisation est un filament de plusieurs nucléosomes qui forment un zigzag d'une longueur de 300 Å. Finalement, le dernier niveau consiste dans l'enroulement des filaments sur eux-même pour former des boucles radiales.

Une telle structure peut, par conséquent, être considérée comme un mécanisme de régulation de la transcription des gènes. Pour qu'un gène soit transcrit, il doit y avoir une altération ou dénaturation de la chromatine dans la région où se situe le ou les gènes que l'on veut transcrire. Cette modification de la chromatine permettrait à la machinerie transcriptionnelle (protéines régulatrices et ARN polymérase) de pouvoir atteindre l'ADN et d'effectuer leur travail.

Un exemple d'une telle régulation est l'acétylation des histones. L'acétylation a pour but de réduire l'affinité qui existe entre les histones et l'ADN en modifiant la charge positive des histones. Ainsi le niveau de compacité diminuera et le nucléosome ne pourra plus empêcher les facteurs de transcription d'accéder aux régions régulatrices de l'ADN.

Il existe aussi une méthode qui permet de réguler les gènes en inhibant leur transcription. Il s'agit de la méthylation de l'ADN. Ce mécanisme pourrait avoir deux effets : soit réduire l'affinité entre les facteurs de transcription activateurs et la séquence d'ADN, soit lier des répresseurs, qui ont une grande affinité à l'ADN méthylé.

La compacité de l'ADN, l'acétylation des histones et la méthylation de l'ADN ne sont que quelques-unes des manières de contrôler le processus de transcription des gènes (Voet et Voet, 2002).

2.1.2 Structure génique

Pour que la transcription ait lieu de façon adéquate, chez l'humain, il doit y avoir plusieurs protéines, activatrices et répressives, qui régulent cette étape.

Pour commencer, il existe trois types d'ARN polymérase : PolI pour les ARNr, POLII pour les ARNm et POLIII pour les ARNt. Elles permettent de dénaturer (séparation des deux brins) l'ADN et de synthétiser la séquence d'ARN associée. L'ARN polymérase II est un complexe qui comportent entre 12 et 14 protéines différentes et vont après se fixer sur le TSS (Transcription starting site) du gène. Ce site se situe environ à 30 paires de bases de la TATA-box (flèche rouge sur la figure 2.3). L'une des sous-unités de la polymérase permet de reconnaître le début de la séquence codante. Cette position indique à l'ARN polymérase où elle doit se lier et dans quel sens se diriger. C'est l'ARN polymérase qui dénature la double hélice et commence à synthétiser l'ARN complémentaire.

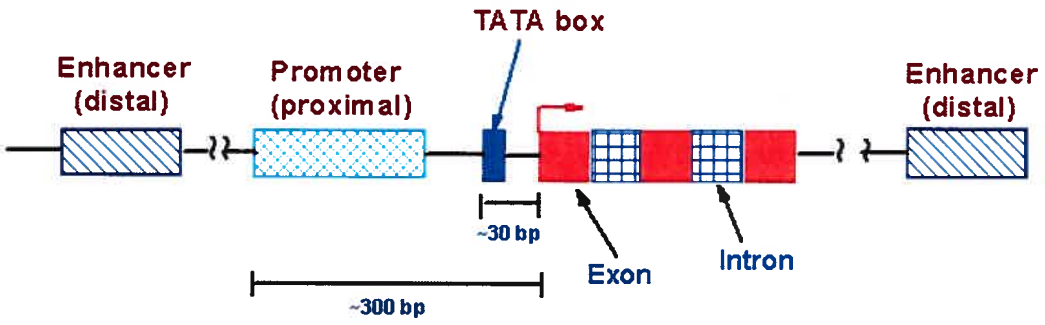

FIG. 2.3 : Structure d'un gène
[http://bioinfo.unice.fr/enseignements/www2005/documentation/outils/recherche_promoteurs/rappels_biologie.html]

La section promotrice est composée de deux éléments : le promoteur principal (core promoter) et le promoteur proximal [http://web.indstate.edu/thcme/mwking/gene-

regulation.html]. Le promoteur principal contient la région TATA-box qui est une séquence des 7 nucléotides suivants : TATAAAA. C'est à cet endroit que se lie un complexe d'environ 50 protéines (machinerie transcriptionnelle basale, TFIID, TBP, TFIIB, etc). On retrouve ce promoteur dans la plupart des gènes. En ce qui concerne le promoteur proximal, on y retrouve deux types de séquences : CCAAT et GGGCGG. Elles sont situées à environ 100-200 bases en amont de la TATA-box et permettent la liaison de certaines protéines : facteur de transcription CP1 (globine γ) et facteur de transcription à doigt de zinc (SP1).

Il existe aussi deux autres classes d'éléments qui peuvent influencer le processus de transcription chez les eucaryotes : les activateurs et les répresseurs. Comme leur nom l'indique, les activateurs augmentent l'activité de la transcription et les répresseurs l'inhibent de façon partielle ou complète. Même si, parfois, on peut les retrouver à plus d'un million de bases du TSS, ils peuvent tout de même exercer leur influence.

Plusieurs gènes partagent les même sites de liaisons aux facteurs de transcription et les mêmes régions promotrices. Mais c'est la combinaison de tous les éléments (ARN polymérase, activateurs, répresseurs, etc.), qui interviennent dans la régulation, qui permet d'avoir une spécificité d'expression. Leur nombre et leur positionnement varient selon les gènes et permettent de réguler l'expression d'un gène en fonction du temps et du lieu (tissu). Par conséquent, un même gène peut être exprimé de plusieurs manières différentes dans des types cellulaires distincts.

2.1.3 Mécanisme général

La transcription commence en un point précis de l'ADN (TSS) et un seul des deux brins est transcrit. C'est l'ARN polymérase II qui permet la transcription des gènes codants pour des protéines (ARNm, figure 2.4). Au tout début, le complexe formé de l'ARN polymérase II et de certains facteurs de transcription (TFIIB, TFIID, TFIIE, TFIIF et TFIIH) doit se fixer à la région promotrice (Voet et Voet, 2002). Ensuite, le facteur TFIIH permet la dénaturation des deux brins.

Une fois cela accompli, il est maintenant possible pour les ribonucléotides libres de venir s'apparier au brin de la matrice (ADN) et de débuter la transcription.

Au fur et à mesure que la transcription progresse, le complexe avance d'une paire de base à la fois. Après l'ajout du quatrième ribonucléotide, le complexe prétranscriptionnel subit un changement de conformation qui entraîne une stabilisation du complexe ouvert (Langelier et al., 2002). A ce moment, il n'est plus nécessaire d'avoir le facteur de transcription TFIIH.

Maintenant, l'ARN polymérase II entre dans un cycle d'initiations avortées. Cela lui permet de s'arrêter à n'importe quel moment, de relâcher l'ARNm, de reculer sur la matrice d'ADN pour repositionner son centre catalytique au site d'initiation et d'entreprendre un nouveau cycle. Lorsque le complexe est en position adéquate, environ au $11^{ième}$ ribonucléotide, le promoteur se referme rapidement et le complexe entre dans le mode d'élongation (Langelier et al., 2002). Lors de cette transition tous les facteurs de transcription, sauf TFIIF, sont relâchés progressivement et la polymérase peut alors s'associer aux facteurs d'élongation. Ces facteurs préviennent les pauses de la polymérase durant la synthèse de l'ARNm. Cette étape se continue jusqu'au moment où le complexe rencontre un site qui lui indiquera que la transcription doit se terminer. Le processus exact de terminaison est mal défini dû au fait que les transcrits primaires ont des séquences 3' hétérogènes.

Une fois la transcription achevée, on se retrouve avec une séquence de transcrit primaire qui n'est pas, nécessairement, fonctionnelle. Pour devenir active, elle doit subir des modifications post-trancriptionnelles. Ces modifications se produisent dans le noyau avant que l'ARNm soit transporté à l'extérieur de celui-ci et traduit. L'ensemble de ces mécanismes de modification de transcrits primaires se nomme la maturation et n'est présent que chez les eucaryotes.

2.1.4 Maturation

Il existe quatre types de modifications post-transcriptionnelles : l'ajout d'une « coiffe » au messager, l'ajout d'une queue polyadénylée, l'épissage et l'édition. Tous les ARNm des cellules eucaryotes possèdent une structure, au tout début, que l'on nomme la « coiffe » et qui a été ajoutée de façon enzymatique. Cela permet de définir le site d'initiation de la traduction. Ensuite, les messagers possèdent presque tous, aussi, des queues poly(A) de 20 à 50 nucléotides. Ces queues sont ajoutées de façon enzymatique grâce à 2 réactions :

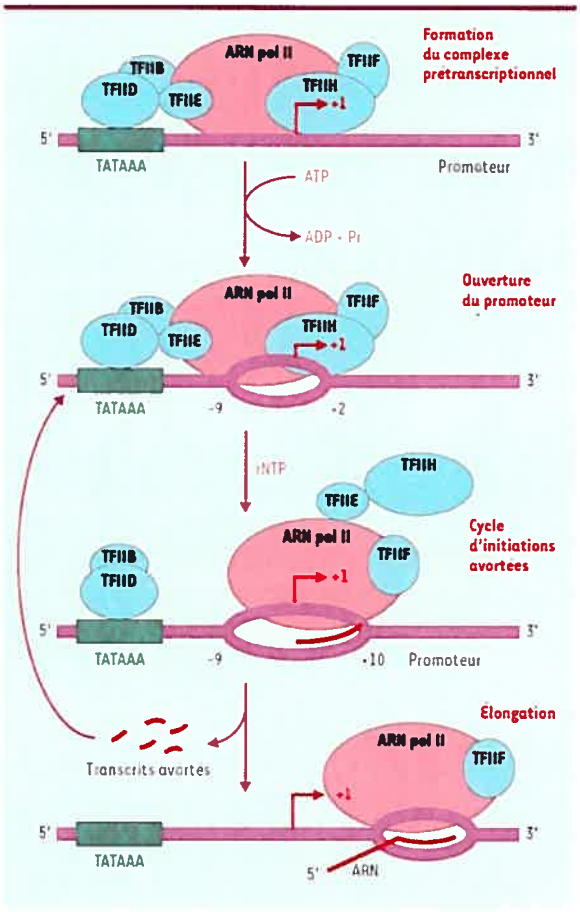

FIG. 2.4 : Étapes de la transcription (Langelier et al., 2002)

1- Un transcrit est clivé (15 à 25 nucléotides) après une séquence très bien conservée (AAUAAA).
2- Ajout de la queue qui est formée à partir d'ATP grâce à la poly(A) polymérase. Elle est activée suite à la reconnaissance du site AAUAAA par un facteur de spécificité de clivage et de polyadénylation (CPSF) (Voet et Voet, 2002).

La troisième modification consiste à réorganiser l'information contenue dans le transcrit. Ce dernier est divisé en plusieurs zones que l'on nomme exons ou introns selon le cas (figure 2.5).

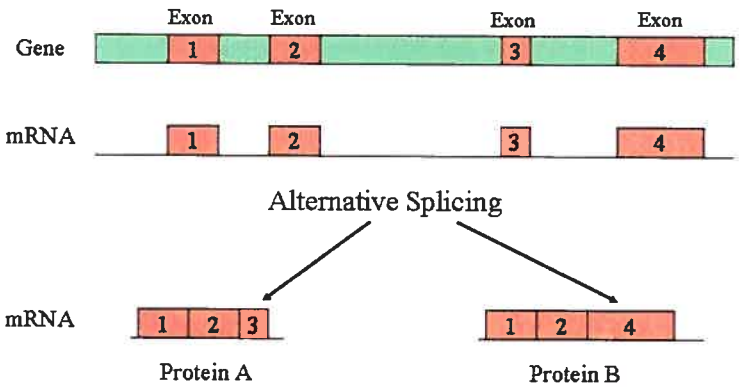

FIG. 2.5 : Épissage alternatif
[http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/alternative_splicing.html]

Les exons (rouge) représentent les régions qui vont coder pour la protéine en question, tandis que les introns (vert) n'ont, généralement, pas de signification pour la dite protéine. Alors l'épissage consiste à enlever les introns du transcrit et à coller les exons les uns à la suite des autres. Il existe, aussi, ce qu'on appelle l'épissage alternatif et qui est représenté sur la figure 2.5. Ce mécanisme permet, à un même gène, de pouvoir produire plusieurs protéines différentes. La protéine est déterminée en fonction des exons qui sont excisés et conservés.

Finalement, il y a l'édition qui consiste en la modification de seulement un ou deux nucléotides. Ce mécanisme permet de corriger ou réparer l'information encodée par le génome, et même parfois, il permet de diversifier l'information contenue dans le messager ce qui offre un plus grand potentiel de complexité à l'organisme (Bass, 2002).

Une fois que toutes les modifications ont été faites, l'ARNm peut être transporté à l'extérieur du noyau et être traduit dans le cytoplasme de la cellule.

2.2 Traduction

Une fois que l'ARNm est arrivé dans le cytoplasme, il y a toute une machinerie qui prend la relève pour traduire ce dernier en séquence protéique. Il est important de mentionner qu'il faut passer d'un alphabet à 4 lettres (ARNm) à un alphabet à 20 lettres (protéine). Il y a plusieurs éléments importants qui participent à la traduction : l'ARNm, le ribosome, les ARNt et les aminoacyl-ARNt synthétases en particulier.

2.2.1 Code génétique

La correspondance entre les lettres n'est pas de un pour un, mais trois acides nucléiques correspondent à un acide aminé. Par conséquent, si l'on fait le calcul des différentes possibilités pour chaque triplet (codon), alors on obtient : 4 X 4 X 4 = 64 possibilités. Il y a 64 possibilités et seulement 20 acides aminés, donc le code génétique est dit dégénéré. Il suffit de regarder dans le tableau qui suit (Tableau I) pour comprendre que plusieurs codons peuvent coder pour le même acide aminé.

First position (5' end)	Second position				Third position (3' end)
	U	**C**	**A**	**G**	
U	UUU Phe UUC Phe	UCU Ser UCC Ser	UAU Tyr UAC Tyr	UGU Cys UGC Cys	U C
	UUA Leu UUG Leu	UCA Ser UCG Ser	UAA Stop UAG Stop	UGA Stop UGG Trp	A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met[b]	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

Tableau I : Code génétique (Voet et Voet, 2002)

a) Les acides aminés sont colorés selon la caractéristique de leur chaîne latérale : orange → non polaire, bleu → basique, rouge → acide, violet → polaire non chargé

b) AUG fait partie des signaux d'initiation d'une protéine et code pour une méthionine

Comme il a été mentionné auparavant, les codons sont composés de trois nucléotides, ce qui amène la possibilité de lire l'ARNm de trois manières différentes (cadre de lecture). De plus, si on ajoute à ce calcul qu'au moment de la transcription, il y a deux sens de lecture possible, on obtient 6 cadres de lecture possible pour un fragment d'ADN. Par contre, dans la réalité, il y en a généralement seulement un qui va coder pour la bonne protéine et celui-ci sera déterminé par les séquences régulatrices en amont lors de la transcription et de la traduction. En général, pour déterminer où la protéine commence, il suffit de trouver le premier codon ATG et elle se termine avec un des trois codons stop.

2.2.2 ARN de transfert (ARNt)

Afin de décrypter l'information contenue dans l'ARNm, il est essentiel d'avoir un outil capable de lire ce « langage » puisqu'il n'y a aucune affinité entre l'ARNm et les acides aminés. Cet outil est l'ARN de transfert (figure 2.6). Il existe un ARNt pour chaque codon possible, donc 64 possibilités. Chacun d'entre eux contient une séquence de trois nucléotides que l'on nomme anticodon. Cette séquence est complémentaire à un codon de l'ARNm et indique quel est l'acide aminé porté par l'ARNt.

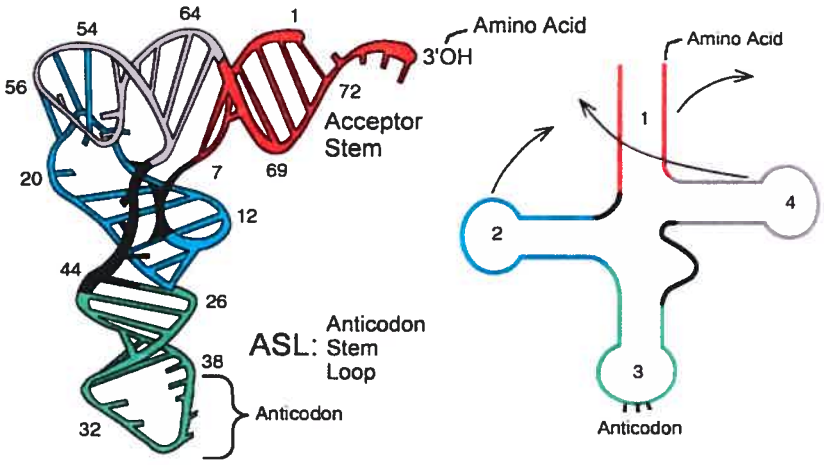

FIG. 2.6 : ARN de transfert [http://biology.kenyon.edu/courses/biol63/ribo/ribo_elong.html]

(Figure gauche) structure tertiaire d'un ARNt, (Figure droite) structure secondaire d'un ARNt

C'est à l'extrémité 3'OH qu'est situé l'acide aminé (figure 2.6). Cet acide aminé ne s'attache pas tout seul à cette extrémité. Ce sont des enzymes (les aminoacyl-ARNt synthétases) qui associent le bon acide aminé au bon ARNt et activent l'extrémité COOH des acides aminés en vue de leur attachement à la chaîne polypeptidique (Voet et Voet, 2002).

2.2.3 Ribosome

Les ribosomes des eucaryotes sont des structures de grande taille. Un ribosome est composé d'ARN et de protéines et il est divisé en deux sous-unités : petite sous-unité (40S) et grande sous-unité (60S) (figure 2.7). Chacune des sous-unités a une fonction spécifique au cours de la traduction et elles s'assemblent d'elles-mêmes au moment opportun. La petite sous-unité permet au ribosome de se lier à l'ARNm et à l'aminoacyl-ARNt, qui s'occupe de la reconnaissance des codons, tandis que la grande sous-unité catalyse la formation de la liaison peptidique.

2.2.4 Synthèse de la chaîne polypeptidique

La synthèse des protéines se fait, généralement, en trois étapes : initiation, élongation et terminaison. Elle s'exécute de l'extrémité N-terminale vers l'extrémité C-terminale et les ribosomes lisent l'ARNm dans le sens 5' → 3'.

Lors de l'initiation de la synthèse, plusieurs facteurs d'initiation ainsi que l'ARNt de départ, celui qui porte une méthionine, vont aller se positionner sur la petite sous-unité ribosomale. Ensuite, le complexe se fixe à l'extrémité 5' de l'ARNm et se déplace jusqu'au premier codon AUG. L'extrémité de départ est reconnue grâce à la coiffe qui a été ajoutée lors de la maturation de l'ARNm. Une fois arrivée au codon de départ, les facteurs d'initiation se détachent et la grande sous-unité du ribosome vient se fixer. Maintenant c'est la seconde étape qui commence : l'élongation de la chaîne polypeptidique.

Durant cette étape, les acides aminés sont ajoutés les uns à la suite des autres, du côté C-terminal. Pour effectuer ce travail, le ribosome possède deux sites de

liaison pour les ARNt : le site P (lie le peptidyl-ARNt) et le site A (lie le nouveau aminoacyl-ARNt) (figure 2.7).

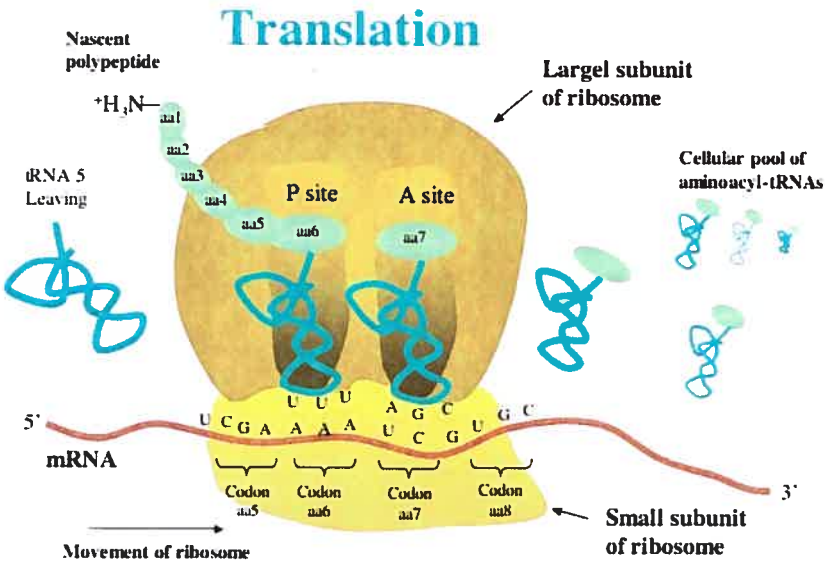

FIG. 2.7 : Traduction
[http://www.langara.bc.ca/biology/mario/Biol2315notes/biol2315chap12.html]

L'extrémité C-terminal se sépare de l'ARNt du site P et réagit avec l'extrémité NH2 de l'ARNt qui se situe dans le site A. La formation de cette liaison amide est catalysée par la grosse sous-unité. Ensuite, il y a une translocation du ribosome sur l'ARNm du 5' vers le 3' qui a pour effet de transférer le peptidyl-ARNt du site A vers le site P. Le mécanisme se poursuit jusqu'à la rencontre d'un codon stop.

Finalement, arrive l'étape de terminaison, lorsque le ribosome rencontre un codon stop, ce qui permet à un facteur de libération de lier l'ARNm. Par la suite, la peptidyl transférase ajoute une molécule d'eau pour former l'extrémité COOH du peptide et la protéine est libérée (Voet et Voet, 2002). La dernière étape de la traduction consiste en la dissociation du ribosome.

Il est important de mentionner que plus d'un ribosome peut travailler sur un même ARNm, ce qui permet de synthétiser plusieurs protéines à la fois. Ce mécanisme permet d'amplifier la production des protéines par rapport à la quantité d'ARNm présente dans le cytoplasme.

Pour devenir mature et fonctionnelle, les protéines doivent subir des modifications post-traductionnelles. Ces modifications (établissement de ponts disulfures, liaison hydrogène, etc.) leur permettront d'adopter leur conformation native et de la conserver. D'autre part, il peut y avoir des ajouts de certaines molécules, non protéiques : des lipides, des glucides ou des groupements phosphate. Parfois, il peut, aussi, y avoir un clivage d'un peptide situé à l'extrémité N-terminale. Ce type de clivage a pour but, par exemple, d'enlever un peptide signal (entre 13 et 36 résidus) qui aurait dirigé la protéine non mature dans un compartiment particulier de la cellule (Voet et Voet, 2002).

Toutes les protéines doivent être traduites pour ensuite pouvoir effectuer leur tache. Mais une fois qu'elles sont traduites, elles peuvent agir de différentes manières dans les cellules. Dans la prochaine section, nous expliquerons le fonctionnement des récepteurs couplés aux protéines G, qui constituent une des familles les plus étudiées présentement par la communauté scientifique.

2.3 Récepteurs Couplés aux Protéines G

L'une des familles de protéines parmi les plus connues et les plus importantes chez les animaux est celle des récepteurs couplés aux protéines G (RCPG). Nous allons décrire leur structure, leur mode d'activation et les voies métaboliques activées, ainsi que le phénomène de dimérisation des récepteurs.

2.3.1 Structure

La principale caractéristique qui distingue ces protéines des autres est qu'elles possèdent toutes sept domaines transmembranaires: une partie N-terminal extracellulaire, une partie C-terminal intracellulaire et trois boucles de part et d'autre de la membrane cytoplasmique (Fig. 2.8). Ces protéines ont une taille qui se situe entre 200 et 1500 acides aminés. Leur structure tridimensionnelle est schématiquement montrée à la figure 2.8. Elle change de conformation lorsque le récepteur est activé et ce changement est généralement dû à la fixation d'un ligand sur le site actif du récepteur. Outre ces ressemblances entre les divers récepteurs, ils possèdent tous des caractéristiques qui leur sont propres.

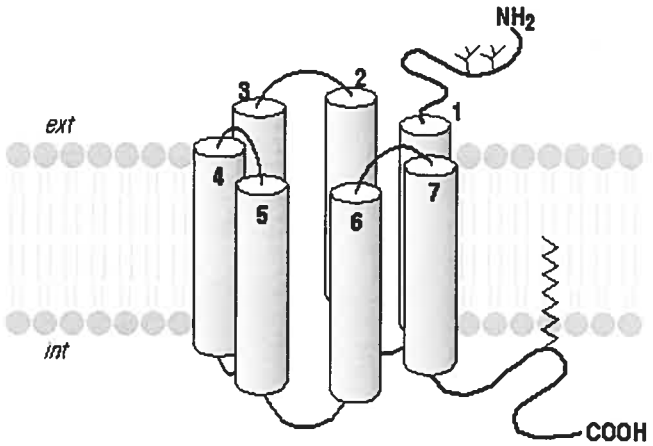

FIG. 2.8 Exemple de RCPG. [http://www.chez.com/rcpg/partie4.html]

Il y a environ 1000 gènes qui sont reconnus pour coder un de ces récepteurs chez l'homme et ceux-ci sont impliqués dans la régulation de plusieurs voies de signalisation importantes pour le développement, la prolifération, la différentiation et la survie des cellules. Au vu de leur grande importance, les RCPG constituent l'une des familles de protéines les plus étudiées de nos jours. Par conséquent, pour mieux comprendre leur fonctionnement, une classification a été produite en tenant compte de l'homologie des diverses séquences.

Il a été remarqué que certains récepteurs avaient des acides aminés en commun, aux même positions le long des domaines transmembranaires ainsi que dans les boucles qui les relient. De plus, en règle générale, leurs régions transmembranaires étaient similaires à environ 25 %. La classification (Kolakowski, 1994) qui fût réalisée était composée de 6 classes différentes, dont quatre principales : A, B, C, F/S. La classe A étant la plus grande avec plusieurs centaines de récepteurs comme ceux pour la lumière, pour l'adrénaline et les récepteurs olfactifs. On peut facilement les reconnaître grâce aux motifs DRY, parfois ERY, que l'on retrouve dans la boucle entre le domaine 3 et 4 (Fig. 2.9). Ensuite, la classe B comporte environ une trentaine de récepteurs comme des peptides hormonaux gastro-intestinaux, de la calcitonine, la parathyroïde et l'hormone corticotropine. Ceux-ci sont caractérisés par leur large domaine N-terminal qui comporte un grand nombre de cystéines. La classe C, qui constitue l'une des plus petites classes, contient

les récepteurs métabotropique du glutamate, du GABA , les récepteurs du goût et ceux du calcium, qui eux aussi ont un large domaine N-terminal et un domaine C-terminal très structurés. La classe F/S ne contient pratiquement que des récepteurs impliqués dans l'embryogenèse, tel que les FRIZZLED et les Smoothened (Pierce et al., 2002).

Chacune des classes peut être divisée en sous-familles. Pour la classe A et B, il y a environ 16 sous-familles chacune et la classe C comporte 8 sous-familles. Ces sous-familles sont déterminées, tout comme les classes, en fonction des séquences et de leurs fonctions respectives dans le corps humain pour chacun des récepteurs. On retrouve, par exemple, les amines et les peptides dans la classe A, les calcitonines et les sécrétines dans la classe B et les glutamates et les GABA-B dans la classe C.

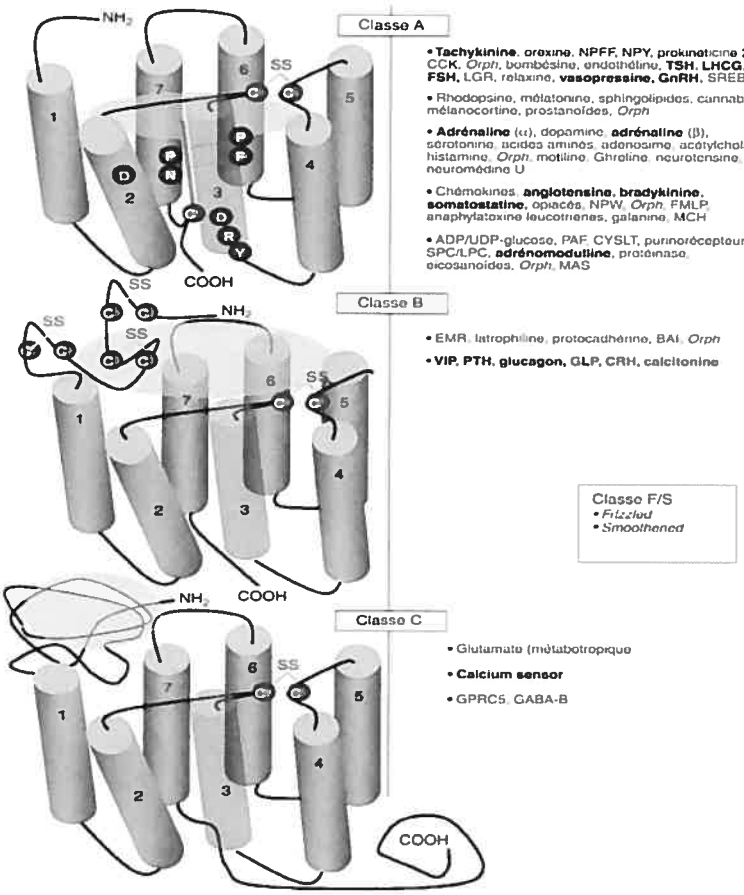

FIG. 2.9 Familles des récepteurs couplés aux protéines G (RCPG)(Assié et al., 2004).

2.3.2 Activation

La plupart des récepteurs couplés aux protéines G (RCPG) sont activés de la même manière et stimulent différentes cascades métaboliques. Pour qu'il y ait une activation du récepteur, il faut généralement (RCPG de la famille A) qu'un ligand vienne se fixer au site actif de ce dernier, qui est situé dans la membrane cytoplasmique. Il existe une très grande variété de stimuli qui peuvent activer un RCPG, par exemple : des ions (Ca^{++}), des stimuli sensoriels (lumière et odeurs), des petites molécules endogènes (lipides), des composés exogènes et des protéines (hormones glycoprotéiques)(Fig. 2.10).

Les protéines G existent sous forme d'hétérotrimère constitué de trois sous-unités: α, β et γ. C'est la sous-unité α qui se lie aux nucléotides. La sous-unité β est en forme d'hélice à 7 lames et la sous-unité γ a deux hélices alpha qui s'enroulent autour de la sous-unité β. Les sous-unités α et γ sont habituellement accrochées à la membrane par des ancrages. L'activation du RCPG permet de catalyser l'échange de la guanosine diphosphate (GDP) pour la guanosine triphosphate (GTP). Le récepteur interagit avec la protéine et permet l'ouverture du site de liaison du nucléotide, ce qui permet au GDP de partir et au GTP, en solution, de s'y attacher. En même temps, la sous-unité α se dissocie du dimère βγ. Cette dissociation permet de transmettre au reste de la cellule que le récepteur s'est lié à son ligand. Ainsi, la sous-unité α et le complexe βγ peuvent aller interagir avec d'autres effecteurs (Fig. 2.10).

La fin du signal d'activation se produit lorsque le GTP est hydrolysé en GDP et que le complexe GDP-α s'associe avec une autre sous-unité βγ. Cette propriété de pouvoir s'auto-désactiver est primordiale pour empêcher l'activation continuelle de la protéine G. C'est la sous-unité α qui possède la propriété de pouvoir hydrolyser le GTP en GDP et P_i. Mais cette réaction est tout de même assez lente (quelques secondes à quelques minutes), ce qui donne le temps au Gα d'activer d'autres composantes de la voie de la transduction. Par contre, il existe un facteur d'échange appelé Epac qui permet d'accélérer la réaction.

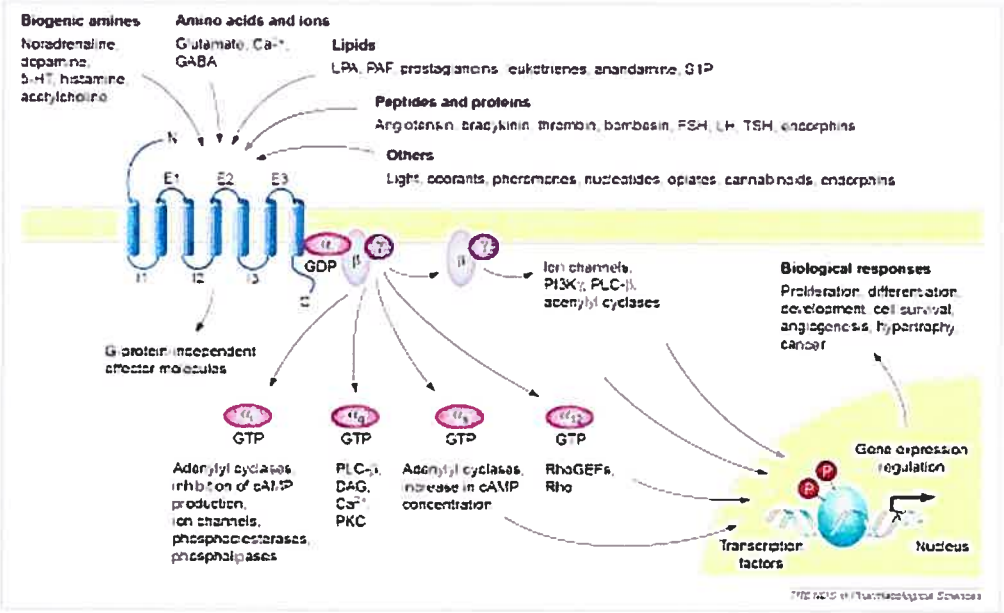

FIG. 2.10 Activation d'un RCPG (Marinissen et al., 2001).

Lorsqu'une protéine G est activée, il peut y avoir une amplification du signal. Cette amplification varie d'un récepteur à un autre et il en existe plusieurs catégories. Cela dépend du type de sous-unité qui a été activée. La sous-unité α est composée de deux domaines et c'est le domaine supérieur qui permet à la sous-unité α d'avoir une interaction avec la sous-unité β et le récepteur. Les sous-unités β et γ sont imbriqués l'une dans l'autre et elles ne peuvent pas être séparées.

Il existe plusieurs gènes qui codent pour les différentes sous-unités. En tout, il y en a 16 pour les sous-unités α, 5 pour les sous-unités β et 12 pour les sous-unités γ (Assié et al., 2004). De plus, il peut y avoir plusieurs combinaisons possibles entre les différentes sortes de sous-unités. Ce mécanisme de combinaison s'effectue en fonction de la compatibilité entre les sous-unités. Chaque type de sous-unité α s'associe de façon exclusive avec un certain type de sous-unité βγ. Par contre, une protéine Gα peut s'associer à plusieurs Gβγ différents. Mais l'activation ou l'inhibition des effecteurs est spécifique selon la sous-unité α impliquée.

Comme il a été mentionné dans la section précédente, les protéines G peuvent cibler plusieurs types d'effecteurs, en voici quelques-uns :

- Adénylates cyclases : c'est une enzyme membranaire qui est activée par les sous-unités de type $G\alpha_s$ et inhibée par celles de type $G\alpha_i$. L'adénylate cyclase catalyse la réaction de formation de l'AMPc, qui est le second messager produit à partir d'ATP. Cet AMPc est l'activateur d'une protéine kinase AMPc-dépendante qui est capable de phosphoryler, et ainsi de moduler l'activité de nombreux substrats protéiques (Rawn, 1990). Epac peut aussi activer la protéine kinase.
- Phospholipase C : c'est une enzyme catalysant la réaction d'hydrolyse du PIP_2 (phosphatidyl inositol bi-phosphate) en IP_3 (inasitol-3-phosphate) et 1,2-diacylglycérol. Elle est activée par les sous-unités de la classe $G\alpha_q$. Une fois PIP_2 hydrolysé, l'IP_3 va se fixer à un canal calcique de la membrane du réticulum endoplasmique et il permet aux ions de calcium contenu dans les vésicules du RE de revenir à l'intérieur du cytosol (Rawn, 1990). Le diacylglycérol quant à lui, active la protéine kinase C, qui peut phosphoryler des protéines afin d'en moduler l'activité tout comme la protéine kinase A (Freeman, 1997).
- cGMP phosphodiestérase : c'est l'enzyme qui permet d'hydrolyser la cGMP à partir du GTP, ce qui permet de transformer le signal lumineux en signal électrique. Elle est activée par les transducines G_{t1} et G_{t2} (Rawn, 1990).
- Canaux ioniques : certains canaux à conductance potassique ou calcique voient leur activité modulée par certaines sous-unités de la classe $G\alpha_i$.

2.3.3 Dimérisation

Une des particularités des RCPG est qu'ils peuvent homodimériser et/ou hétérodimériser. Cela signifie qu'ils peuvent former des oligomères avec d'autres récepteurs identiques (homodimère) ou avec des récepteurs différents (hétérodimères) (Bouvier, 2001). La plupart de ces dimères sont formés lors de la synthèse des RCPG dans le réticulum endoplasmique. Même aujourd'hui, le mécanisme qui permet à certains récepteurs de former des dimères n'est pas très bien compris. Il a été démontré que certains récepteurs peuvent exister sous forme de

monomère (exemple : rhodopsine), mais l'homodimère est plus stable (Chabre et al., 2005). Certains RCPG ne sont fonctionnels que lorsqu'ils forment des dimères ($GABA_B$-R1 et $GABA_B$-R2). Il y en a d'autres qui ne maturent qu'à l'état de dimère (récepteur de l'adrénomédulline) et certains autres possèdent des propriétés de liaison et d'activation différentes lorsqu'une telle association se produit (opioid delta-kappa) (Assié et al., 2004). Par conséquent, plusieurs chercheurs tentent de comprendre les divers mécanismes qui peuvent régir ce phénomène, puisque certaines paires de récepteurs peuvent former des hétérodimères tandis que d'autres ne peuvent pas.

De nos jours, nous ne connaissons qu'une trentaine d'hétérodimères et qu'une dizaine de couples de récepteurs qui ne peuvent pas hétérodimériser sur environ un million de possibilités (Prinster et al., 2005). La principale raison de ce manque de connaissance est due au fait qu'il est difficile et coûteux de déterminer à grande échelle de manière biochimique si deux récepteurs peuvent ou non former une telle association.

Il existe plusieurs technologies qui peuvent nous permettre de détecter des dimères *in vivo*, mais certaines d'entre-elles peuvent comporter certaines limites. Par exemple, l'utilisation de la co-immunoprécipitation et le western-blot ont tous deux besoin que les récepteurs soient parfaitement solubilisés (Bouvier, 2001). Si ce n'est pas le cas, il peut y avoir formation d'agrégats qui seront interprétés comme étant des dimères, ce qui n'est pas nécessairement le cas. Pour remédier à cette situation, plusieurs chercheurs ont commencé à utiliser deux technologies, le FRET (fluorescence resonance energy transfert) et le BRET (bioluminescence resonance energy transfert), qui se basent sur le transfert d'énergie de résonance de la lumière et qui consiste en un transfert non-radioactif d'énergie d'excitation entre les dipôles électromagnétiques d'une molécule donneuse et d'une molécule réceptrice (Bouvier, 2001).

En ce qui concerne le FRET, il s'agit d'un transfert d'énergie entre deux molécules fluorescentes. Il y a une molécule qui donne (énergie) et une qui accepte. On attache à la molécule donneuse un fluorochrome dont la longueur d'onde d'émission correspond à la longueur d'onde d'excitation du fluorochrome placé sur la seconde molécule. Les récepteurs ont, bien entendu, été choisis en fonction de leur

susceptibilité d'interaction. Par conséquent, on placera les molécules dans un environnement qui peut favoriser leur interaction et si la distance et l'orientation entre les deux fluorochromes le permettent, il y aura un transfert d'énergie qui pourra être détecté à l'aide d'un appareil (Fig. 2.11 a) (Trugnan et al., 2004).

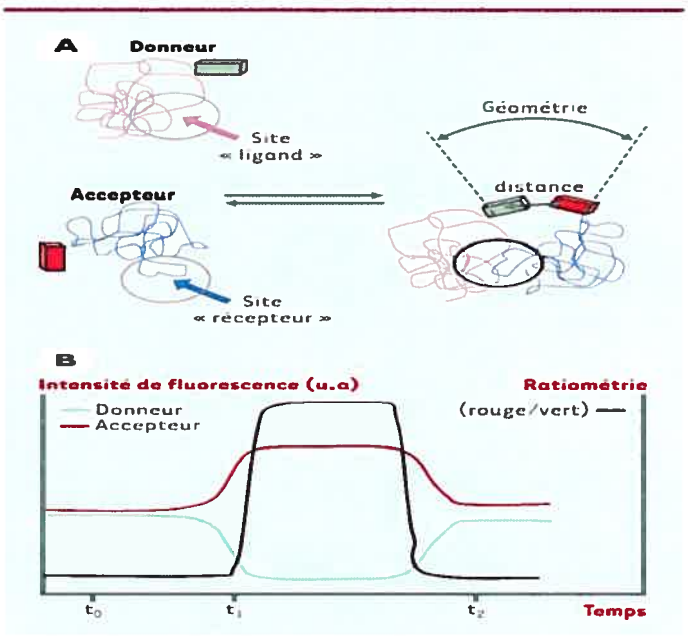

FIG. 2.11. Principe du FRET (Trugnan et al., 2004).

« A : Voir texte ci-dessus. B : Entre t0 et t1, les deux molécules fluorescentes n'interagissent pas et émettent chacune leur fluorescence propre, verte pour l'une et rouge pour l'autre. Au temps t1, sous l'effet d'une stimulation, les deux molécules vont interagir et le phénomène de transfert d'énergie conduit à la diminution de l'intensité de fluorescence du donneur (tracé vert), parallèlement à l'augmentation de l'intensité de fluorescence de l'accepteur (tracé rouge). La mesure du rapport d'intensité des deux fluorochromes (ratiométrie, tracé noir) est également représentée. Au temps t2, l'interaction entre les deux molécules est interrompue (présence d'un compétiteur, par exemple) et le transfert d'énergie disparaît » (Trugnan et al., 2004).

Le BRET se base sur le même principe biophysique de transfert d'énergie que le FRET. Par contre, l'énergie de la molécule donneur provient d'une molécule bioluminescente, la luciférase. Lorsque la luciférase interagit avec son substrat, elle émet des photons. Ceux-ci sont alors transférés sur une molécule fluorescente, GFP (Green Fluorescent Protein), qui peut émettre une fluorescence si la condition essentielle est atteinte : que les molécules soient à proximité l'une de l'autre.

Pour essayer de mieux comprendre le fonctionnement de la dimérisation des RCPG, nous avons décidé d'utiliser la bio-informatique pour parvenir à notre but.

Avant de décrire en détails le protocole que nous avons suivi, il est important d'introduire quelques concepts (phylogénie, recherche de motifs, etc.) de bio-informatique qui seront décrits à la section suivante.

Chapitre 3

Notions bio-informatiques

Suite à l'introduction des concepts biologiques, il est aussi important d'en introduire quelques-uns en bio-informatique. Dans ce chapitre, il y aura une description sur le concept de l'évolution des espèces, de la phylogénie ainsi qu'un aperçu de certaines méthodes et certains outils bio-informatiques qui sont utilisés pour la recherche de motifs dans des séquences d'ADN (acides désoxyribonucléiques). Nous traiterons aussi les avantages et inconvénients de chacun de ces logiciels.

3.1 Évolution

3.1.1 Définition

« Processus qui a transformé la vie sur terre depuis les tout débuts jusqu'à la diversité apparemment sans limites d'aujourd'hui, constitue le fil conducteur de l'apparition et du développement des caractéristiques de la vie. » (Campbell, 1995)

3.1.2 Historique

Le premier point tournant de cette discipline fut sûrement lorsque Charles Darwin publia *De l'origine des espèces par voie de sélection naturelle ou la lutte pour l'existence dans la nature* en 1859. Dans ce livre, il a démontré, en premier lieu, que les espèces que nous connaissons aujourd'hui sont le résultat d'une transformation progressive d'une génération à l'autre des ancêtres de chacune des espèces (descendance). Ensuite, il présenta, dans son livre, la théorie qui explique, selon lui, l'évolution des êtres vivants. Cette théorie se nomme la sélection naturelle. Elle comporte, selon Darwin, « deux faits indéniables et une conclusion incontournable » :

- « Dans une population donnée, les individus diffèrent les uns des autres. » (Campbell, 1995)
- « Toute population peut produire une descendance beaucoup trop nombreuse par rapport à ce que l'environnement offre en matière de nourriture, d'espace et d'autres ressources. Cette surnatalité entraîne inévitablement une lutte pour la survie entre les différents membres de la population. » (Campbell, 1995)
- « Les individus possédant les caractéristiques les mieux adaptées au milieu de vie laissent généralement une progéniture beaucoup plus nombreuse que les autres. Cette reproduction sélective augmente la représentation de certaines variations héréditaires chez la génération suivante. C'est cette survivance différentielle que Darwin appela « sélection naturelle » et considéra comme la cause de l'évolution. » (Campbell, 1995)

Suite à cette publication beaucoup d'autres scientifiques se sont penchés sur la question et ont suivi peu à peu les traces de Darwin avec sa théorie de l'évolution, même si elle comporte certaines faiblesses (par exemple : personne n'a pu observer l'ancêtre commun des Mammifères et des Oiseaux).

3.1.3 Représentation

L'hypothèse centrale de la biologie évolutive est que toutes les formes de vie sur Terre partagent un ancêtre commun. Cette hypothèse est fortement corroborée par de nombreux faits : toutes les espèces utilisent les mêmes molécules organiques, les mêmes polymères, possèdent le même code génétique, etc. Tout ceci n'est pas arrivé par hasard, mais bien parce qu'il y a eu un processus évolutif. Ce processus évolutif a produit un patron de relations entre les diverses espèces. Au fil du temps, une lignée évolue, se sépare et hérite des modifications de leurs ancêtres. Ainsi le « chemin » que prend l'évolution se diversifie. Cela produit un patron d'embranchements qui représente les relations d'évolution et que l'on nomme arbre phylogénétique. En étudiant les caractéristiques héréditaires et d'autres évidences historiques, on peut reconstruire les relations qu'il y a entre les espèces. On peut

représenter sur un seul arbre toutes les relations de descendance entre les diverses espèces. Cela nous a permis de constater que l'on pouvait classifier toutes les espèces dans trois grands domaines : Archées, Bactéries et Eucaryotes (Woese et al., 1990). Il est important de noter que cet arbre peut comporter certaines erreurs et qu'il n'est pas complètement résolu (déterminé).

3.2 Phylogénie

Dans cette section, nous aborderons les concepts de base dans le domaine de la phylogénie et les différentes méthodes de reconstruction d'arbres phylogénétiques.

3.2.1 Définition

Étude des relations de parenté des organismes vivants.

3.2.2 Notions de base

La phylogénie se base sur une hypothèse générale et qui s'énonce comme suit : il y a une origine unique de la vie et tous les êtres vivants ont plus ou moins des relations de parentés étroites. Il est important de définir certaines notions qui permettront de mieux comprendre la phylogénie en général. Le premier élément à définir est le terme groupe monophylétique, qui est un mot très utilisé dans ce domaine et qui représente une unité (genre ou espèce) qui regroupe tous les organismes qui descendent d'un ancêtre commun exclusif et donc qui possèdent, généralement, des caractéristiques communes.

Dans le but de faciliter l'étude des relations de parenté entre les espèces, les scientifiques ont inventé une représentation graphique de ces relations : l'arbre phylogénétique. Les feuilles de l'arbre représentent les taxons observés (actuels ou fossiles), tandis que les nœuds internes, eux, représentent des ancêtres hypothétiques (Fig. 3.1). Les liens (branches) entre les nœuds ou bien entre une feuille et un nœud peuvent être associés à plusieurs mesures différentes, cela dépend de ce que l'on veut représenter. Ils peuvent signifier une distance génétique, une durée ou bien une vitesse d'évolution. Une règle importante à respecter, lorsque l'on construit un arbre

phylogénétique, est qu'il est interdit d'y avoir des cycles, donc un seul lien arrive dans un nœud, mais plus d'un lien peut partir d'un même nœud. Par contre, il y a une exception à cette règle. Au cours de l'évolution, il peut y avoir des échanges de matériels génétiques entre différentes espèces et on nomme ce phénomène, transfert horizontal de gène. Par conséquent, si on essaie de représenter la phylogénie d'un gène en particulier sur l'arbre des espèces, il est possible qu'on se retrouve avec un ou des cycles dans l'arbre.

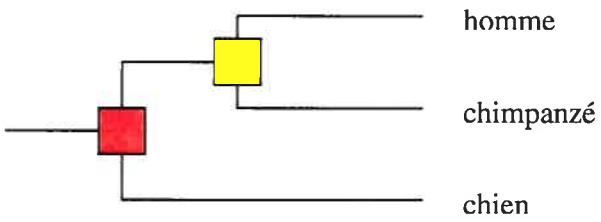

FIG. 3.1 : Exemple d'arbre phylogénétique
Ici, on peut voir que l'homme est plus proche parent du chimpanzé, car il partage un ancêtre commun (carré jaune) exclusif. On peut aussi constater que ces deux derniers ont une relation de parenté avec le chien dû à un ancêtre commun plus ancien (carré rouge). Ici le chien représente le groupe extérieur sur lequel on s'appuie pour créer l'arbre.

Il existe deux types d'arbre : avec ou sans racine. L'arbre qui a été montré à la figure 3.1 est un arbre avec racine qui correspond aux ancêtres communs de tous les taxons représentés. Mais on peut très bien représenter le même arbre sans racine (Fig 3.2).

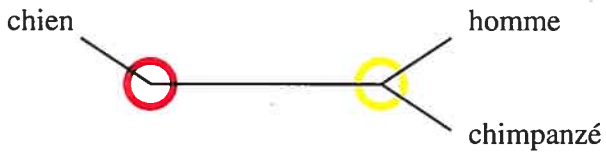

FIG. 3.2 : Arbre sans racine

Par conséquent, il est possible dans un arbre non-raciné de placer la racine où l'on désire. Ainsi, le même arbre peut prendre différentes « formes », mais les connexions sont tout à fait équivalentes.

3.2.3 Homologie et orthologie

La construction des arbres en phylogénie est basée sur les caractères homologues qui existent entre les séquences des diverses espèces. Il existe deux types d'homologie : primaire et secondaire. L'homologie primaire se détermine à partir de structures ou caractères qui se ressemblent et par conséquent on fait l'hypothèse qu'ils sont hérités d'un ancêtre commun. L'homologie secondaire est déterminée avec l'arbre phylogénétique qui maximise la cohérence des données (caractères) tout en minimisant le nombre de changements nécessaires pour passer du premier arbre au second. De plus, il faut qu'il y ait suffisamment de similarité entre les séquences pour que ce ne soit pas dû au hasard mais bien d'un héritage dû à un ancêtre commun. Par conséquent, certains arbres, par exemple, sont créés à partir des caractères morphologiques des espèces et d'autres à partir de leurs séquences protéiques pour certains gènes donnés. Une fois que la phylogénie a été inférée, il est possible de comparer les divers gènes ou espèces présents dans l'arbre et pour cela un certain vocabulaire a été défini.

Par exemple, une paire de gènes est dite orthologue lorsque les deux gènes proviennent d'un évènement de spéciation (Sp1 et Sp2), tandis que deux gènes provenant d'un évènement de duplication sont paralogues (Dp1 et Dp2) (Fig. 3.3). Par conséquent, C2 et C3 sont paralogues entre eux et les deux sont orthologues par rapport à B2.

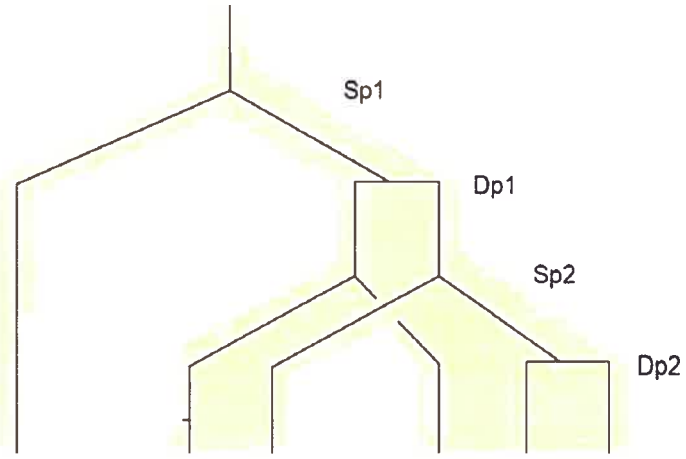

FIG. 3.3 Orthologue et Paralogue (Fitch, 2000)

Suite à une duplication, les deux gènes résultants ont la même fonction. Mais pour que la copie soit conservée, elle doit procurer un avantage sélectif. Il existe quatre résultats finaux possibles : 1) disparition de la copie (ce qui arrive le plus fréquemment), 2) conservation de la même fonction que la copie originale ce qui permet d'avoir un effet sur le dosage des protéines (production plus ou moins élevée d'ARNm), 3) il peut se produire une séparation des fonctions entre les deux copies (sous-fonctionnalisation) et 4) la nouvelle copie peut acquérir une nouvelle fonction. Par contre, en pratique, on fait souvent l'approximation suivante : deux gènes orthologues ont sensiblement la même fonction, mais dans deux espèces différentes, et deux gènes paralogues n'ont pas la même fonction, mais appartiennent à la même espèce.

3.2.4 Méthodes de construction

Pour pouvoir construire un arbre phylogénétique, on doit suivre une certaine procédure pour s'assurer que le résultat sera valide. Avant la construction proprement dite, il faut effectuer les trois étapes suivantes :

- Récupérer les séquences homologues dans une base de données (ex : NCBI).
- Aligner les séquences avec un programme fiable pour déterminer l'homologie primaire de chaque positon.
- Vérifier manuellement si l'alignement semble correct et éliminer les positions dont l'homologie semble incertaine.

Une fois terminée, il faut choisir une méthode d'inférence pour l'arbre que l'on veut produire. Il en existe trois :

1- Maximum de parcimonie : Les arbres les plus parcimonieux sont ceux qui expliquent la distribution des données entre les espèces avec le minimum de changements nécessaires pour passer d'une espèce à l'autre. L'arbre qui aura la plus petite longueur (nombre minimum de changements) sera le plus parcimonieux. En fin de compte, cette méthode se base sur le principe du

Rasoir d'Occam : «l'explication la plus simple possible est toujours la meilleure».

2- Méthodes de distances : La donnée de départ est une matrice qui contient les distances évolutives entre les différents taxons. Ces distances peuvent être calculées de plusieurs manières différentes : en calculant les similitudes et différences entre chaque paire de séquences (distance observée) ou en calculant le nombre de substitutions qui se sont produites depuis leur ancêtre commun en utilisant un modèle probabiliste d'évolution des séquences (distance évolutive). Une fois la matrice déterminée, on peut construire l'arbre. Il existe plusieurs manières de faire, mais seulement une sera décrite, NJ (Neighbor Joining) (Saitou et al, 1987). Le NJ consiste à réunir les taxons qui déforment le moins la matrice de distances, en minimisant la longueur des branches de l'arbre. A chaque étape, les deux taxons sont regroupés pour former un « nouveau » taxon par rapport aux autres restant. La nouvelle distance est calculée en fonction de la moyenne des taxons qui sont regroupés. On recommence le traitement jusqu'au moment où tous les taxons ont été regroupés.

3- Méthodes probabilistes : Il existe deux méthodes d'inférence d'arbres phylogénétiques que se base sur une approche probabiliste. Il y a le maximum de vraisemblance et les méthodes bayésiennes (http://www.biani.unige.ch/msg/teaching/evolution.htm). Ces deux méthodes ont la propriété d'utiliser des modèles d'évolution moléculaire explicites et sont basées sur des calculs statistiques complexes, ce qui allonge considérablement leur temps d'exécution comparativement aux autres méthodes (parcimonie, méthodes de distances).

Le maximum de vraisemblance évalue la topologie de différents arbres et choisit le meilleur en se basant sur un modèle d'évolution spécifique, c'est-à-dire celui qui fournit la plus grande probabilité d'observer les données. Ce modèle essaie de refléter le processus évolutif que l'on croit responsable

de la conversion d'une séquence en une autre. Les longueurs de branches sont un des principaux paramètres à estimer, outre la topologie (Holder et al., 2003).

L'objectif de l'inférence Bayésienne est d'obtenir la distribution des phylogénies en fonction de leurs probabilités postérieures et non pas de chercher le « meilleur » arbre. On peut obtenir ce résultat en combinant la vraisemblance et la distribution des probabilités « a priori » des paramètres d'évolution. L'utilisation des méthodes de chaînes de Markov Monte Carlo s'avère nécessaire et efficace pour calculer la distribution des probabilités postérieures.

3.2.5 Théorie neutraliste de l'évolution

La Théorie Neutraliste de l'évolution (Kimura, 1977) se base sur l'énoncé suivant : la plupart des variations que l'on remarque au niveau moléculaire sont sélectivement neutres, c'est-à-dire qu'il n'y a pas d'avantage ou de désavantage associé à un allèle particulier. De plus, ils les neutralistes affirment que la dérive génétique est la principale cause de changements dans la fréquence des allèles au cours du temps et que la sélection naturelle n'est que très peu impliquée dans ce phénomène. Ils ne disent pas que toutes les mutations sont neutres, au contraire, ils croient plutôt que la plupart des mutations sont délétères pour l'organisme et qu'elles ne restent pas assez longtemps dans la population pour contribuer de façon quantitative à la variation de cette dernière. Les seules mutations qui n'ont pas d'effets néfastes sur les organismes sont les mutations silencieuses et par conséquent elles ont une chance d'être conservées assez longtemps pour qu'on puisse les observer. Cela peut se produire, lorsqu'il y a une mutation du troisième nucléotide d'un codon et que l'acide aminé résultant reste inchangé.

Il existe aussi la sélection positive qui favorise les organismes qui la présentent et du même coup augmente la probabilité qu'elle se fixe dans cette population. Par contre, pour la théorie neutraliste, la sélection positive est considérée comme un événement très rare et s'appuie plutôt sur la dérive génétique et la sélection négative comme mentionné ci-dessus (Lopez et al., 2002).

3.2.6 Empreinte phylogénétique

Avec le nombre grandissant de génomes séquencés et annotés, il devient de plus en plus intéressant de les comparer entre eux (génomique comparative). Une approche qui a fait ses preuves ces dernières années, dans ce domaine, est l'empreinte phylogénétique qui fut introduit pour la première fois par Tagle et al. (1988). Elle consiste à rechercher les positions qui ont été conservés au cours de l'évolution. Elle se base sur la théorie neutraliste de l'évolution (Kimura, 1977) qui dit que la plupart des mutations ponctuelles n'ont pas d'effets biologiques sur l'organisme. Par conséquent, l'empreinte phylogénétique fera plutôt ressortir les positions qui, quand mutées, ont un effet défavorable sur l'organisme.

On utilise couramment cette méthode pour rechercher des régions régulatrices ou des sites potentiels de fixation de facteurs de transcription (Lenhard et al., 2003; Sauer et al., 2006). L'idée principale est de comparer les séquences de gènes de diverses espèces pour déterminer les régions conservées dans l'ensemble des séquences. Mais il ne faut pas comparer n'importe quelle séquence, il faut que les séquences soient orthologues. La règle générale sur laquelle cette approche s'appuie est que les éléments de régulation dans les régions non-codantes subissent une pression sélective purificatrice très forte au cours de l'évolution comparativement aux régions non-fonctionnelles (Sauer et al., 2006). Donc, on s'attend à ce que les régions non-fonctionnelles accumulent plus de substitutions au cours de l'évolution que les régions régulatrices et que l'alignement des séquences orthologues fassent ressortir les régions conservées qui seraient sensées avoir un rôle fonctionnel similaire chez les différentes espèces. Évidemment, ceci découle du fait que les gènes orthologues, comme il a déjà été mentionné auparavant, possèdent en général des fonctions similaires dans chacune des espèces. Donc, on s'attend aussi qu'ils soient régulés plus ou moins de la même façon.

Lorsque l'on utilise l'empreinte phylogénétique comme approche pour trouver des régions régulatrices, il est important de suivre les trois points suivants :

1- Trouver les séquences orthologues

2- Choisir un algorithme d'alignement
3- Déterminer un critère de conservation

En ce qui concerne les deux premiers points, de nos jours, ceux-ci sont plus faciles à résoudre parce qu'il y a beaucoup de données génomiques et que les outils d'alignement de séquences se sont beaucoup améliorés (Bulyk, 2003; Tompa et al, 2005). Pour le dernier point, il peut être plus difficile à déterminer, car il dépend de ce qu'on recherche. Pour résoudre ce problème, on peut tester plusieurs seuils de conservation et ensuite comparer nos résultats pour ainsi déterminer lequel des seuils correspond le plus à la réalité biologique.

3.2.7 Arbre phylogénétique des RCPG

La figure 3.4 présente l'arbre phylogénétique des récepteurs que nous avons inféré à partir des séquences que nous avons récupérées (voir Matériels et Méthodes). Il a été réalisé avec la méthode de maximum de vraisemblance. On peut remarquer que les récepteurs sont regroupés en 4 sous-familles : MECA (mélanocortine / endogline / adénosine / cannabinoide), CHEM (chemokine-like), AMIN (sérotonine / dopamine / adrenergique / trace amines) et SOG (somatostatine / opioid / galanine). Ces sous-familles de la classe A des RCPG ont déjà été identifiées auparavant (Fredriksson et al., 2005). De plus, il ne contient que deux espèces dû au manque d'espace (l'arbre avec les 6 espèces a été réalisé et on obtient des résultats similaires). On peut aussi constater que tous les récepteurs de l'humain ont comme groupe frère leur orthologue chez la souris et vice versa. Les récepteurs métabotropiques glutamates ont été utilisés comme groupe extérieur puisqu'ils viennent de la classe C contrairement aux autres.

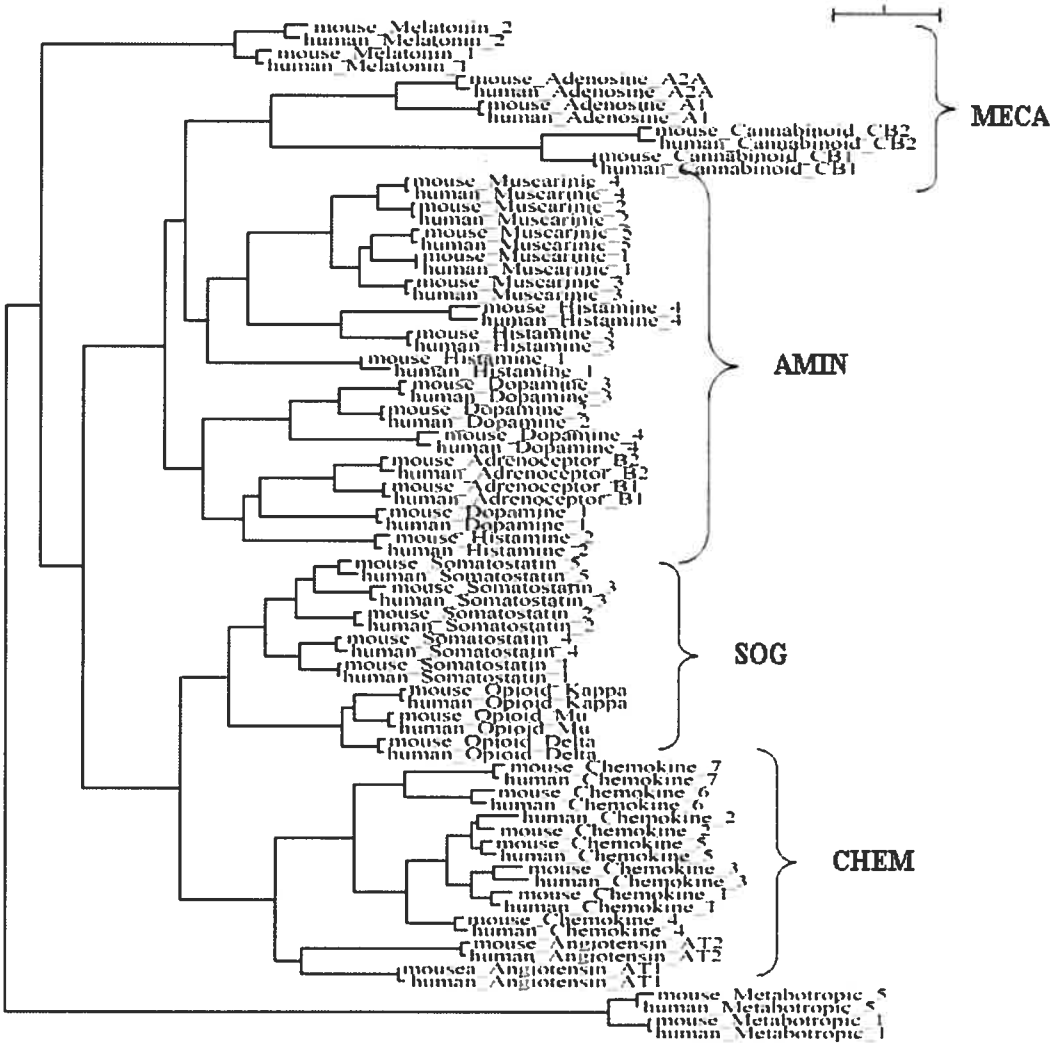

FIG 3.4 : Arbre phylogénétique des RCPG pour la région codante (900-2400 nt). Le logiciel MUSCLE a été utilisé pour réaliser l'alignement des séquences et ensuite le logiciel Phylip a été utilisé pour faire l'arbre avec la méthode du maximum de vraisemblance et le modèle de distribution Gamma.

3.3 Recherche de motifs conservés

L'un des premiers problèmes qu'ont tenté de résoudre les bio-informaticiens est celui de déterminer si deux gènes peuvent être régulés de la même manière. Plusieurs approches ont été développées. L'une d'elles se base sur une analyse plus mathématique et voit le problème comme une comparaison de mots (motifs) qui seraient conservés dans chacune des séquences et/ou dans leurs régions régulatrices en amont ou en aval de leur séquence codante.

La recherche d'une solution pour cette approche fût abordée par plusieurs scientifiques différents (Stojanovic et al., 1999; Jensen et al., 2004a) et cela a donc produit plusieurs solutions. Donc, dans cette section, il sera question des méthodes et outils qui existent pour la recherche de motifs conservés, ainsi que leurs avantages et inconvénients.

3.3.1 Méthodes

Il existe plusieurs stratégies pour effectuer une recherche de motifs régulateurs dans une séquence d'ADN. Ici, nous discuterons des trois approches théoriques les plus utilisées.

Séquences Consensus

Cette méthode consiste à utiliser une représentation (motif général) des motifs régulateurs connus pour en trouver de nouveaux. Un motif est représenté sous la forme d'une chaîne de caractères et chaque caractère provient d'un alphabet (Σ) spécifique (Tableau II). Le but, ici, est de pouvoir trouver toutes les occurrences de ce motif (M) dans une série de chaînes de caractères (séquence : S_i). Une particularité de cette approche est qu'un caractère dans le motif peut correspondre à plusieurs caractères différents dans les séquences. Donc, le problème se définit comme suit :

Σ : Alphabet

M : $m_1 m_2 \ldots m_p$: Motif de taille p

$S_i : s_1 s_2 \ldots s_t$: Séquence i de taille t

On peut voir les différentes séquences comme une seule longue séquence et ainsi, on recherche les occurrences du motif M dans cette séquence. Plusieurs algorithmes peuvent être utilisés pour rechercher un mot dans un texte (exemple : Boyer-Moore (Boyer et al., 1977), Morris-Pratt (Morris et al., 1970), Horspool (1980), etc.), mais nous n'irons pas plus loin dans la description de ceux-ci.

On détermine le motif consensus à l'aide d'une liste de motifs reconnue comme étant des facteurs de transcriptions et on recherche ce dernier dans la liste de séquences que l'on veut analyser. Chaque mot qui correspond au motif consensus est considéré comme un motif régulateur potentiel.

IUPAC nomenclatures for DNA consensus

A	Adenine	C	Cytosine
G	Guanine	T	Thymine
R	Purines (A, G)	Y	Pyrimidines (C, T)
W	Weak hydrogen bond (A, T)	S	Strong hydrogen bond (C, G)
M	Amino group (A, C)	K	Keto group (G, T)
B	Not A (C, G, T)	D	Not C (A, G, T)
H	Not G (A, C, T)	V	Not T (A, C, G)
N	Any (A, C, G, T)		

Tableau II : Nomenclature (Jensen et al., 2004a)

Le principal point négatif de cette approche est qu'elle peut amener plusieurs faux-positifs et peut manquer certains motifs qui ont une variation qui n'est pas représentée dans le motif consensus. Par exemple, on peut utiliser le mot consensus BDAM. Nous pouvons avoir des faux-positifs dans le cas où nous aurions dû mettre un D, en $2^{ième}$ position du mot consensus, car seulement un de nos motifs régulateurs avec une thymine (T) à cette position. Alors nous risquons de trouver beaucoup de motif avec une thymine comme $2^{ième}$ nucléotide, si nous utilisons notre mot consensus pour la recherche même si ce n'est pas très représentatif de l'ensemble des facteurs de transcription que nous possédons. Nous pouvons aussi manquer des occurrences. Il suffit qu'un motif régulateur ne soit pas représenté par notre ensemble de départ, c'est-à-dire qu'il existe des motifs régulateurs qui ont une cystosine (C) en $2^{ième}$ position par exemple, mais avec notre motif consensus nous ne

permettons pas d'avoir une cystosine à cet endroit. Dans ce cas nous allons manquer son occurrence.

Au cours des années 1980, la plupart des recherches qui utilisaient ce genre d'approche cherchaient seulement à identifier une courte séquence (motif) qui était sur-représentée par rapport au reste des autres séquences (Jensen et al., 2004a). Plusieurs améliorations ont été apportées à cette approche. Par exemple, on attribue une valeur selon la longueur du motif et le nombre de fois qu'il apparaît dans les séquences. On ne conserve que ceux qui ont un pointage élevé et ceux-ci sont examinés de plus près pour savoir s'ils sont vraiment des motifs régulateurs. Il y eut aussi comme amélioration (Sinha et al., 2000), autour de l'an 2000, une modification qui rendait l'approche plus permissive face aux variations qu'il peut y avoir entre différents motifs. Ainsi, il est permis d'avoir un « mismatch » entre le motif consensus et les motifs recherchés.

Il y a eu beaucoup d'autres modifications qui ont été apportées (Liu et al., 2001; Keich et al., 2002) et qui ne seront pas discutées ici. Il reste tout de même le fait que cette approche comporte certaines lacunes comme le fait qu'il faille connaître à l'avance ce que l'on veut rechercher. De plus, sa permissivité face aux variations des motifs n'est pas la plus adéquate. Avec certaines approches, il est possible d'avoir qu'un seul « mismatch » dans le motif consensus ce qui peut nous faire manquer beaucoup d'occurrences de motifs. Par conséquent, d'autres approches, plus permissives, ont été développées et l'utilisation de matrices de positions spécifiques en est un exemple.

Matrices de poids selon les positions (Position-Specific weight matrix)

Afin de mieux représenter les variations de nucléotides qui peuvent avoir lieu dans les motifs d'ADN recherchés, une approche basée sur l'utilisation de matrices a été élaborée (Lawrence et al., 1990). La matrice permet de représenter la distribution des nucléotides pour chacune des bases à une position particulière en fonction des sites alignés pour le motif recherché mais sans prendre en compte la dépendance phylogénétique existante entre les séquences (Tableau III). Il est possible, une fois cette matrice déterminée, de la transformer en matrice de fréquence ou de poids.

Position	A	C	G	T
1	3	3	2	0
2	1	4	2	1
3	2	1	2	3
4	8	0	0	0
5	2	2	2	2
6	3	2	1	2

Tableau III : Exemple de matrice

Pour obtenir la matrice de fréquence, il suffit, pour chaque case de la matrice, de diviser ce chiffre par la somme de tous les nombres sur la même ligne. Par exemple, pour la case (1,A) = 3 /(3 + 3 + 2 + 0) = 0,38. Ensuite, si l'on veut obtenir la matrice de poids, il faut utiliser la formule suivante : log [fréquence$_{ij}$ / θ_{oj}] où θ_{oj} est la proportion de la base j pour les positions qui n'appartiennent pas aux motifs.

Que l'on choisisse l'une ou l'autre de ces manières de faire (Liu et al., 1995; Hertz et al., 1999), elles sont toutes plus précises que l'utilisation de séquences consensus parce qu'elles permettent une meilleure représentation, basée sur les fréquences des nucléotides des motifs recherchés et qu'elles donnent une meilleure approximation de la similarité entre les motifs puisqu'un pointage leur ait associé en fonction des valeurs de départ de la matrice. Cette approche comporte tout de même des faiblesses. On doit déterminer un seuil pour le pointage minimal qui sera admis en tant que motif potentiel. De plus, on doit aussi déterminer la matrice de départ qui sera utilisée pour la recherche. Pour solutionner ce problème, plusieurs méthodes ont été mises au point (Lawrence et al., 1990; Liu et al., 1995), mais elles ne seront pas énumérées ici. Par conséquent, il serait intéressant d'avoir des outils qui pourraient à la fois déterminer la matrice de départ et trouver les motifs associés, ce qui a amené le développement d'approches comme le « motif matrix update ».

Matrices (Motif Matrix Updating)

L'idée principale qui se cache derrière cette approche (Jensen et al., 2004a) est l'utilisation d'une matrice de probabilité pour la représentation d'un motif et qui, au

tout début, est initialisée de façon aléatoire. Ensuite on utilise notre jeu de données pour raffiner cette matrice et ainsi trouver le ou les motifs conservés.

Lors de la recherche, l'algorithme commence par rechercher les motifs dans les deux premières séquences. On assume qu'il n'y a qu'un motif par séquence. Ensuite, chacune des occurrences trouvées sera classée selon un pointage. Ce pointage peut être calculé de plusieurs manières distinctes, l'une d'entre elles consiste à calculer l'entropie relative entre les deux motifs :

$$\text{ENTROPIE} = \sum_{i=1}^{m} \sum_{j=1}^{T} f_{ij} \log [f_{ij}/\theta_{oj}]$$

où f_{ij} est la fréquence de la base j dans la position i, le log $[f_{ij}/\theta_{oj}]$ est la matrice de poids, m la longueur du mot et T la longueur du texte ou séquence (Jensen et al., 2004a). La matrice de poids est déterminée à partir de la matrice des fréquences qui, à son tour, est déterminée à partir de la matrice d'alignement (exemple : Tableau III).

Une fois qu'une occurrence a été trouvée et que celle-ci a un pointage raisonnable, alors ce motif est recherché dans les autres séquences. A la fin, lorsque toutes les séquences auront été traitées, les motifs ayant le meilleur pointage seront reportés.

Une seconde approche (Hertz et al., 1999) pour évaluer la pertinence des motifs consiste à calculer la « p-value » de chacun. Cette valeur (p-value) est définie comme la probabilité d'observer un motif similaire de même taille dans un alignement de séquences aléatoires. Seuls les motifs ayant une « p-value » très petite sont considérés comme des occurrences du motif recherché.

Plusieurs outils bio-informatiques (Liu et al., 2001; Bailey et al., 2006) utilisent ce genre d'approche. Il y a eu certaines modifications dans quelques-uns d'entre eux, mais le principe de base reste le même. Ce type de méthode est beaucoup plus flexible que les autres qui ont été présentées auparavant. Par contre, elle ne garantit pas la convergence de l'algorithme ou qu'elle va trouver le motif optimal.

Il existe deux algorithmes principaux qui sont utilisés pour améliorer la matrice jusqu'au moment de sa convergence : l'algorithme EM (Expectation Maximization) et l'algorithme d'échantillonnage (Gibbs sampling).

L'algorithme EM (Expectation Maximization) est un algorithme d'apprentissage non-supervisé qui garantit la convergence de ce dernier vers un maximum local (Bailey et al., 1995b). L'algorithme prend en entrée un groupe de séquences non-alignées et une longueur de motif (W) et retourne un modèle probabiliste du motif partagé.

Une particularité de EM est qu'on peut utiliser plusieurs modèles différents de départ comme : 1) trouver les motifs qui n'apparaissent qu'une seule fois par séquence (modèle OOPS); 2) trouver les motifs qui apparaissent 0 ou 1 fois par séquence (modèle ZOOPS); 3) trouver les motifs qui peuvent apparaître entre 0 et plusieurs fois (modèle TCM). Une fois le modèle choisi, l'algorithme alterne entre deux étapes : l'étape E (expectation) qui calcule la valeur attendue de l'information manquante (probabilité (z) que l'occurrence d'un motif commence à la position j dans une séquence X_i) et l'étape M (maximization) qui estime le maximum de vraisemblance des paramètres (p) en utilisant le résultat obtenu à l'étape précédente (étape E). Ensuite, les résultats obtenus à l'étape M sont utilisés pour la prochaine étape E et ainsi de suite jusqu'à convergence de la matrice (fig. 3.5).

```
EM (jeu de données, W){
      Choisir un point de départ(p)
      Faire{
            Ré-estimer z à partir de p
            Ré-estimer p à partir de z
      } Tant que (changement dans p < ε)
       Retourner
}
```

FIG. 3.5 : Pseudo-code de l'algorithme EM (Adapté de Bailey et al., 1995b)

La stratégie Gibbs sampling, elle, commence par initialiser une matrice de probabilités pour les motifs en utilisant un alignement, produit au hasard, des séquences de départ. Ensuite, elle essaie d'améliorer cette matrice de façon itérative

et stochastique (Lawrence et al., 1993). Elle prend en entrée N séquences de longueur L et recherche un motif de longueur W. Ensuite, elle effectue les étapes suivantes :

1) Choisir une position de départ dans chaque séquence, au hasard : a_1 dans la séquence 1, a_2 dans la séquence 2, …, a_N dans la séquence N
2) Choisir une séquence au hasard dans le groupe de séquences de départ (exemple : seq 1)
3) Faire une matrice de poids de longueur W avec tous les sites de toutes les séquences, sauf celle choisie à l'étape 2.
4) Assigner une probabilité à chaque position de la séquence 1 en utilisant la matrice de poids construite à l'étape 3 : $p = \{ p_1, p_2, p_3, \ldots, p_{L-W+1} \}$
5) Choisir une position de départ dans la séquence 1 basée sur la distribution de la probabilité (p) et affecter a_1 à cette nouvelle position.
6) Choisir une séquence au hasard dans le groupe de séquences de départ (exemple : seq 2)
7) Faire une matrice de poids de longueur W avec tous les sites de toutes les séquences, sauf celle choisie à l'étape 6.
8) Assigner une probabilité à chaque position de la séquence 2 en utilisant la matrice de poids construite à l'étape 7 : $p = \{ p_1, p_2, p_3, \ldots, p_{L-W+1} \}$
9) Choisir une position de départ dans la séquence 2 basée sur la distribution de la probabilité (p).
10) Recommencer jusqu'au moment où la matrice converge (Lawrence et al., 1993).

En résumé, l'algorithme « Gibbs sampling » retourne une série de motifs, qui sont représentés sous forme de matrices de poids, et qui sont sur-représentés dans le jeu de données de départ comparativement aux autres motifs (Tompa et al., 2005).

3.3.2 Outils

Il existe une panoplie de programmes, dont plusieurs sont résumés dans Tompa et al. (2005), qui peuvent effectuer une recherche de motifs dans une séquence d'ADN et/ou protéique, mais certains se démarquent des autres. Alors, nous décrirons cinq logiciels, qui sont parmi les plus utilisés dans leur domaine. Ils peuvent être classés en fonction de l'approche utilisée pour leur recherche de motifs. AlignACE (Hughes et al., 2000) et BioProspector (Liu et al., 2001) sont basés sur un algorithme d'échantillonnage (Gibbs sampling), MEME (Bailey et al., 2006) utilise l'algorithme EM (Expectation Maximization) pour trouver le maximum de vraisemblance pour un modèle statistique donné, CONSENSUS (Hertz et al., 1999) utilise une stratégie de recherche séquentielle afin d'optimiser l'information trouvée (Che et al., 2005) et Bipad (Bi et al., 2004) se base sur la maximisation de l'information.

AlignACE

Ce programme, qui est implémenté en C++, permet de trouver des motifs communs dans un ensemble de séquences d'ADN. Après avoir utilisé l'algorithme Gibbs sampling pour trouver les motifs conservés, il évalue la pertinence de chacun des motifs à l'aide d'un pointage appelé MAP (maximum *a priori* log likelihood). On peut faire une approximation de cette valeur en utilisant la formule suivante : N log R, où N est le nombre de sites alignés et R est le degré de sur-représentation du motif dans la séquence donnée (Hughes et al., 2000). Par exemple, si on s'attend à retrouver un motif pour chaque 100 bases, selon la fréquence des différents nucléotides à l'arrière-plan, et que 30 motifs sont retrouvés pour une séquence de 3000 bases, alors R = 10. Une formule plus détaillée peut être retrouvée dans (Liu et al., 1995).

Cet algorithme fonctionne autant pour les séquences protéiques que pour une séquence d'ADN. Par contre, il ne garantit pas la convergence sur le même motif à chaque fois. Donc, on doit le faire « tourner » plusieurs fois et comparer les résultats. Par conséquent, si nous utilisons AlignACE cela peut nous prendre beaucoup plus de temps pour obtenir un résultat satisfaisant qu'avec un autre outil.

BioProspector

BioProspector est un autre logiciel qui implémente aussi une stratégie « Gibbs sampling ». Il permet d'examiner les régions en amont des gènes pour y trouver des motifs régulateurs. Il peut prendre en entrée une panoplie de données :

- Fichier de départ qui contient toutes les séquences d'ADN
- Fichier qui contient des séquences ou probabilités qui caractérisent la distribution des nucléotides en arrière plan (background).
- Longueur des motifs recherchés et gap entre les deux motifs
- Nombre de copies du motif par séquence
- Le type de recherche effectuée (sur un ou deux brins)
- La recherche de palindromes

Il utilise des modèles de Markov pour lequel les paramètres sont soit définis par l'utilisateur soit estimés à partir des séquences fournies au départ. Par contre, plus nous donnons d'informations et plus notre recherche s'avèrera précise et comportera peu de faux-positifs.

A chaque itération de BioProspector, on appelle un processus (threshold sampler) qui adopte la stratégie « Gibbs sampling ».

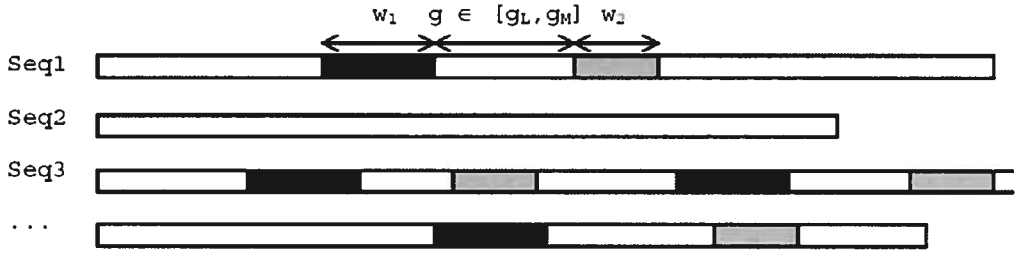

FIG. 3.6 : Modèle utilisé par BioProspector (Liu et al, 2001).

Le motif est composé de 2 parties de différentes longueurs : w_1 et w_2. Elles sont séparées par un nombre d'espaces minimal (g_L) ou maximal (g_M).

Tous les motifs trouvés sont, ensuite, évalués à l'aide d'un pointage estimé par une méthode Monte Carlo et seuls ceux qui ont un bon pointage sont conservés. A la fin le logiciel retourne l'information suivante :

- Le pointage des motifs, leur valeur significative et le nombre de segments alignés.
- Une expression régulière du motif consensus et dégénéré et une matrice de probabilité du motif.
- Le nombre de segments qui a contribué au motif ainsi que la position de départ et la séquence de chacun de ces segments.

Une des particularités de BioProspector est que le programme a été conçu, tout d'abord, pour rechercher des motifs à 2-blocs. Un motif à 2-blocs est composé de deux motifs simples (standard) qui sont séparés par un espacement d'une longueur variable (Fig. 3.6). Il a été démontré que ce type de motif peut réguler certains gènes (Helmann et al., 2002; Jensen et al., 2005). Il peut tout de même rechercher des motifs simples, il suffit que la valeur pour la longueur du $2^{ième}$ motif et pour le gap soit égale à zéro.

MEME

MEME implémente la méthode de Bailey et al. (1993) pour trouver un ou plusieurs motifs qui caractérisent un groupe de séquences. L'élément principal de MEME est l'utilisation de l'algorithme EM. Tous les motifs trouvés par le logiciel, et qui correspondent aux critères préétablis, seront classés en fonction de leur pointage. De plus, il est permis de définir une plage pour la longueur des motifs recherchés, ce qui est un de ses principaux avantages.

Pour prendre en compte la subtilité des différents modèles, l'ajout d'une variable a été nécessaire, λ. Cette dernière représente la probabilité *a priori* que n'importe quelle position, dans une séquence donnée, puisse être le point de départ d'une occurrence d'un motif. Donc, pour le modèle OOPS, elle ne sera pas nécessaire et, par conséquent, la boucle interne de l'algorithme ne se fera qu'une

seule fois. En ce qui concerne les deux autres modèles (ZOOPS et TCM), leur valeur respective est : $\lambda_{min} = 1/(m\sqrt{n})$, $\lambda_{max} = 1/m$ et $\lambda_{min} = 1/(m\sqrt{n})$, $\lambda_{max} = 1/(W+1)$. W représente la longueur du motif, m est le nombre de positions de départ possibles (m= L –W +1), L équivaut à la longueur de la séquence analysée et la variable n représente le nombre de séquences de départ.

```
procedure MEME ( X: dataset of sequences )
  for pass = 1 to pass_max do
    for W = W_min to W_max by ×√2 do
      for λ(0) = λ_min to λ_max by ×2 do
        Choose good θ(0) given W and λ(0).
        Run EM to convergence from chosen
        value of φ(0) = (θ(0), λ(0), W).
        Remove outer columns of motif
        and/or apply palindrome constraints
        to maximize G(φ).
      end
    end
    Report model which maximizes G(φ).
    Update prior probabilities U_i,j to
    approximate multiple-motif model.
  end
end
```

FIG. 3.7 : L'algorithme de MEME (Bailey et Elkan, 1995a)

Il est important de mentionner que l'implémentation de la boucle interne ainsi que de l'algorithme pour raccourcir les motifs et appliquer la contrainte de palindrome, ne sont pas représentés ici dans le but de faciliter la compréhension. On peut retrouver la version détaillée dans Bailey et Elkan (1995a).

En résumé, MEME raffine son approche en choisissant un sous-ensemble de solutions et en appliquant une itération de l'algorithme EM sur chaque solution. Ensuite, il choisit une solution dans l'ensemble précédent comme son meilleur candidat et applique l'algorithme EM pour converger à partir de ce point. Lorsque MEME cherche un point de départ, il ne considère pas toutes les possibilités qui sont dans l'intervalle de longueur des motifs recherchés. Il utilise une heuristique qui se base sur l'algorithme EM et il ne prend que certaines valeurs dans l'intervalle.

A la fin, il renvoie un fichier qui contient une série de modèles de probabilités des séquences qui chacun correspond à un motif trouvé qui a été estimé à l'aide de l'algorithme EM.

Un inconvénient d'utiliser l'algorithme EM est que le maximum qu'il trouve est possiblement seulement local. Une solution pour essayer d'éviter ce problème est d'utiliser plusieurs fois le logiciel et de comparer les résultats.

CONSENSUS

Comme son nom l'indique, il utilise une approche par consensus de séquences. L'algorithme utilisé est de type glouton et commence par trouver une paire de séquences qui partagent un motif. Il est déterminé en fonction de l'information qu'il contient et celle-ci doit être la plus grande possible. Ensuite, CONSENSUS prend une troisième séquence et recherche ce même motif. Il continue ainsi jusqu'au moment où il a passé toutes les séquences. Par la suite, il masque le motif qui a été trouvé et recommence au début. Les motifs sont représentés sous forme de matrices de poids. Par conséquent, il recherche la matrice qui contient le maximum d'information comparativement à l'arrière-plan (background).

BiPad

Bipad (Bi et al., 2004) est un algorithme stochastique implémenté en C++ et qui permet de rechercher des motifs à 2-blocs. Il définit ces motifs sensiblement de la même manière que BioProspector (Fig. 3.8a).

Comme avec la plupart des logiciels, il est possible de fixer certains paramètres comme : 1) la longueur de l'espacement entre les deux blocs (gap), 2) la longueur des motifs recherchés, 3) la composition des blocs (homogène ou hétérogène) 4) l'orientation des motifs selon le brin analysé RDR « reverse-direct repeats », DR « direct repeats », IR « inverted repeats », ER « everted repeats » (Fig. 3.8b).

Pour pouvoir déterminer l'occurrence des motifs, BiPad produit un modèle mathématique basé sur la maximisation de l'information ou sur la minimisation de

l'entropie de Shannon. Il existe deux modèles de base que l'on peut utiliser : OOPS et ZOOPS. Chaque moitié du motif est représentée par une matrice de poids.

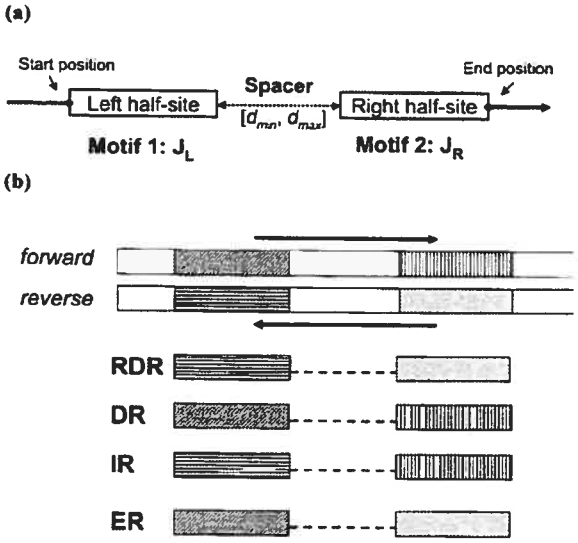

FIG. 3.8 : Fonctionnement de BiPad (Bi et al., 2004).

BiPad détermine les positions initiales de façon aléatoire. Pour chacune de ces positions, il prend une séquence à la fois et énumère toutes les possibilités de position de départ selon les longueurs de gap qui ont été spécifiées. Ensuite, il calcule l'entropie de chaque séquence. L'entropie minimale de chacune des séquences est conservée et les matrices de fréquences sont mises à jour. Ce calcul est exécuté tant et aussi longtemps que la différence entre deux passages ne dépasse pas un certain seuil. Après un tour complet, l'alignement qui contient le maximum d'information est conservé et la procédure ci-dessus recommence un nombre de fois déterminée par l'utilisateur. Finalement, la liste des motifs finaux est bâtie à l'aide de l'ensemble des meilleurs alignements trouvés dans tous les cycles qui ont été réalisés. On peut retrouver les détails de l'algorithme dans Bi et al. (2004).
La sortie est composée de deux matrices de poids, une pour chaque bloc du motif, la position des motifs et les séquences des motifs.

Maintenant que tous les concepts de biochimie et de bio-informatique ont été introduits, nous allons décrire, en détail, dans la section suivante la démarche que nous avons utilisé tout au cours de ce projet.

Chapitre 4

Matériels et méthodes

4.1 Problèmes et Objectif

Comme nous l'avons mentionné dans l'introduction, les RCPG ont une grande importance biologique (système métabolique, système nerveux central, inflammation, etc.) et ils sont nombreux (environ 1000). De plus, ils ont la propriété de pouvoir hétérodimériser ce qui change leur spécificité face aux divers ligands existants. Par conséquent, dans le but de mieux comprendre ce phénomène et parce qu'il est difficile de détecter ces interactions *in vivo*, nous avons voulu étudier la question à l'aide d'une approche *in silico*.

Pour commencer nous avons émis deux hypothèses :

(1) Si deux récepteurs hétérodimèrisent, ils ont plus de chance d'avoir un patron d'expression et surtout de traduction spatio-temporelle similaire.
(2) Il existe des séquences d'ADN qui sont des signatures de patrons d'expression/traduction spatio-temporelle.

Ensuite, il a fallu déterminer quel type de signature nous voulions analyser. Puisqu'il n'existait pas beaucoup de recherches dans ce domaine, nous avons décidé de commencer le plus simplement possible : rechercher des motifs conservés. Par la suite, nous avons aussi déterminé, où il fallait faire ces recherches : dans l'ARNm, en amont et en aval de la région codante. (On se différencie donc de l'approche standard qui recherche les motifs conservés dans les régions promotrices situées en général en 5' du gène).

Finalement, cela nous amène à déterminer l'objectif principal de cette recherche :

- Prédire des patrons d'expression/traduction spatio-temporelles similaires à l'aide de l'empreinte phylogénétique (qui pourraient nous informer sur les possibilités l'hétérodimérisation).

4.2 Jeu de données

Pour réaliser notre étude, nous avons décidé de choisir un ensemble de six espèces différentes. Évidemment, le choix de chacune de ces espèces s'est fait en fonction des génomes séquencés qui étaient disponibles au moment où nous avons débuté cette recherche. En premier lieu, puisque nous voulions étudier les récepteurs chez l'humain, il était important d'avoir ce dernier dans notre jeu de données. Ensuite, nous avions besoin de quelques génomes de mammifères (rat, souris, chien, chimpanzé) pour pouvoir déterminer si les motifs étaient au moins conservés chez ce groupe d'espèces. Finalement, une dernière espèce a été choisie à cause de sa distance évolutive sensiblement plus élevé, le poulet (fig. 4.1)

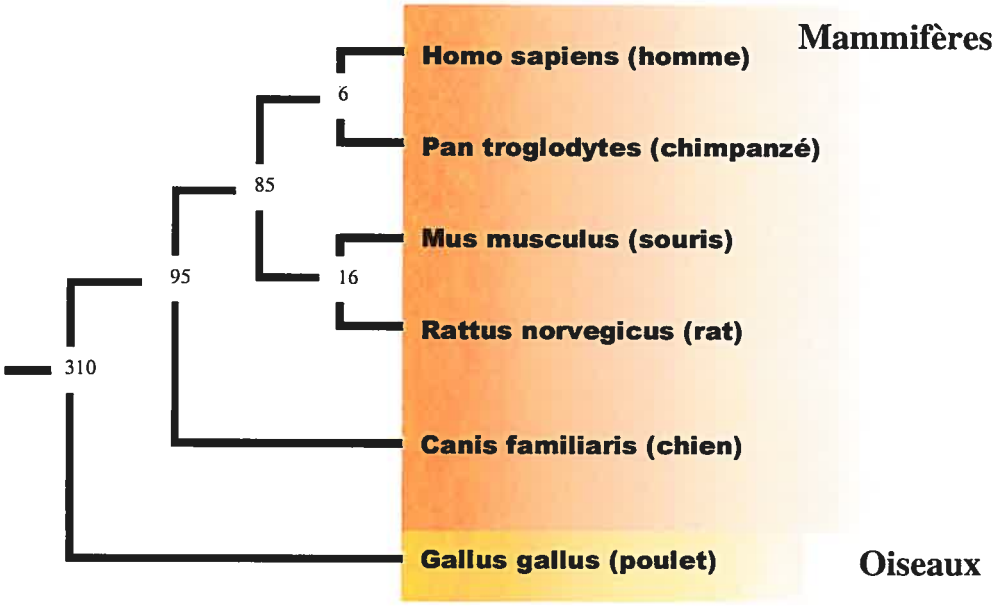

FIG. 4.1 : Arbre des espèces : Les nombres représentent la date de divergence, en terme de millions d'années d'évolution, entre les ancêtres des espèces (Springer et al., 2003).

Une fois les espèces déterminées, il a fallu choisir un groupe de récepteurs couplés aux protéines G. Pour ce faire, nous avons décidé de nous attarder, en particulier, sur la famille A qui contient le plus grand nombre de récepteurs et qui est la plus étudiée. Le choix de ces récepteurs s'est fait de manière aléatoire. Ensuite, nous voulions des récepteurs « témoins », donc nous avons choisi quatre récepteurs de la famille C, dont deux ont été sélectionnés à cause de leur propriété exclusive d'hétérodimériser seulement l'un avec l'autre ($GABA_B$-R1 et $GABA_B$-R2). Voici la liste des récepteurs qui ont été choisis :

- Adenosine (A1, A2A)
- Adrenoceptor (B1, B2)
- Angiotensin (AT1, AT2)
- Cannabinoid (CB1, CB2)
- Chemokine (CCR1 - CCR7)
- Dopamine (D1 - D4)
- Histamine (H1 - H4)
- Muscarinic (M1 - M5)
- Melatonin (MT1, MT2)
- Metabotropic Glutamate (GluM1, GluM5)
- Opioid (Delta, Kappa, Mu)
- Somatostatin (S1 - S5)
- GABA-B (R1, R2)

Il y a en tout 42 récepteurs, dont 38 font partie de la famille A. Pour chacune des espèces, nous avons récupéré une séquence d'une longueur de 1000 et 3000 nucléotides en amont et en aval du gène codant pour chacun des récepteurs de la liste ci-dessus. De plus, nous avons aussi récupéré la séquence codante de chaque récepteur. Les séquences proviennent des bases de données Ensembl (http://www.ensembl.org/index.html) et ainsi que celle du UCSC Genome Browser (http://genome.ucsc.edu/).

Comme la recherche d'orthologues est dans ce cas une tâche facile, nous nous sommes contentés d'utiliser le logiciel Blast en partant des séquences chez l'humain.

4.3 Recherche de motifs

Pour commencer, nous avons décidé de rechercher des motifs simples qui ne sont pas nécessairement connus pour ne pas éliminer dès le départ des éléments qui pourraient

être importants et qui ne sont pas connus. Par la suite, nous avons aussi recherché le nouveau type de motif que nous avons définis plutôt, les motifs à 2-blocs.

Pour effectuer la recherche des motifs conservés à travers les diverses espèces, il a fallu choisir parmi plusieurs outils disponibles, dont certains ont été décrits dans l'introduction. Nous en avons choisi deux : MEME et BioProspector. On a sélectionné ces deux outils, car ils ont été, tous les deux, utilisés à plusieurs reprises dans d'autres recherches et par conséquent, même s'ils ne sont pas parfaits, ils ont fait leurs preuves. De plus, MEME permet la recherche de motifs de différente longueur en même temps et BioProspector permet de rechercher des motifs à 2-blocs.

Ensuite, nous avons déterminé la valeur de quelques paramètres pour nous permettre de faire une recherche plus concise et appropriée. Le premier à avoir été déterminé est la longueur des motifs que nous voulions rechercher. En se basant sur la littérature scientifique (Van Hellemont et al., 2005 ;Venkatesh et al., 2004 ; Zhang et al., 2003), dans laquelle on démontre qu'il existe plusieurs facteurs de transcription qui reconnaissent une séquence d'ADN dont la longueur varie entre 4 et 30 nucléotides, nous avons convenu d'utiliser ces bases comme point de départ. Par contre, très peu sont de petite longueur (4 ou 5) et très peu sont de grande longueur (25 et +), par analogie aux facteurs de transcriptions, alors nous avons décidé que des motifs de 6 à 15 nucléotides de long seraient adéquats. Le second paramètre que nous avons fixé est celui du nombre de motifs recherchés dans une séquence donnée. Pour éviter de bloquer la recherche de motifs parce que le nombre de motifs recherchés est atteint, nous avons tout simplement recherché tous les motifs possibles qui ne se chevauchent pas. Ainsi, dans le cas de MEME, le logiciel effectuait une recherche tant et aussi longtemps qu'il pouvait trouver des motifs. Cela nous a donné environ 200 motifs par séquence codante et environ 100 motifs pour chacune des régions non-codantes (en amont et en aval). Par la suite, ces nombres ont été utilisés comme paramètre pour BioProspector.

En ce qui concerne les motifs à 2-blocs, nous avions deux paramètres principaux à déterminer : la longueur de chacun des blocs qui forment le motif à 2-blocs et la longueur du gap entre ces deux blocs. Pour limiter le nombre de

comparaison de motifs que nous devions faire, nous avons décidé de conserver la même longueur pour les deux blocs du motif à 2-blocs. Ces longueur auront les mêmes valeurs que pour les motifs standards (bloc de longueur de 6 à 15 nt). Par contre, il fut un peu plus difficile de déterminer les longueurs de gap possibles entre les deux blocs. Nous n'avons rien trouvé de précis, dans la littérature, sur ce problème, sauf un article (Jensen et al., 2004b), ce qui nous a amené à définir nos valeurs de gaps à 5, 10 et 15.

4.4 Comparaison des motifs

Une fois que tous les motifs conservés au cours de l'évolution ont été déterminés, il faut pouvoir comparer leur niveau de similarité. Pour faire cette comparaison, nous avons décidé d'utiliser un logiciel d'alignement de séquences appelé NEEDLE et qui provient du package EMBOSS, car il est très simple à utiliser et relativement rapide en terme d'exécution. On fait tout simplement un alignement de chacun des motifs trouvés pour un récepteur en particulier avec les motifs d'un autre récepteur et on détermine le pourcentage d'identité (fig 4.2).

```
#=======================================
# Aligned_sequences: 2
# 1: m1
# 2: m2
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 14
# Identity:      7/14 (50.0%)
# Similarity:    7/14 (50.0%)
# Gaps:          6/14 (42.9%)
# Score: 21.0
#=======================================
m1             1 GTTCTT-GGTGT--    11
                 || ||-||
```

FIG. 4.2 : Exemple de comparaison de motifs avec NEEDLE.

Il est important de mentionner que pour la comparaison des motifs avec NEEDLE, les paramètres qui ont été utilisés sont ceux par défaut (pénalité d'ouverture de gap égale à 10, extension de gap égale à 0,5). Ensuite, les motifs ont été classés en fonction de leur pourcentage de conservation (50, 60, 70, 80 et 90%).

Concernant les motifs à 2-blocs, ceux-ci ont été comparés par bloc (bloc 1 du motif 1 avec bloc 1 du motif 2 et bloc 2 du motif 1 avec bloc 2 du motif 2) et la moyenne des deux comparaisons a ensuite servi à les classer.

4.5 Représentation et analyse des motifs conservés

Afin de faciliter l'analyse, nous avons décidé de représenter les motifs partagés entre les différents récepteurs avec un graphe de connexion. Pour réaliser ce graphe, nous avons fait une matrice d'adjacence dont les entrées représentent le nombre de motifs partagés entre chaque paire de récepteurs avec un niveau de conservation déterminé. Ensuite, nous transformons cette matrice en graphe de connexion : chaque point (rectangle) représente un récepteur, l'arête entre les deux points indique que ces deux récepteurs partagent des motifs communs et le poids sur l'arête indique le nombre de motifs partagés. De plus, les points (récepteurs) sont colorés en fonction de leur répartition dans l'arbre phylogénétique des RCPG (fig. 3.4) qui a été présenté dans l'introduction. Les graphes ont été réalisés en C++ avec le package LEDA [http://www.algorithmic-solutions.com/enleda.htm].

Pour évaluer la validité des résultats obtenus, nous avons réalisé quelques tests. Pour commencer, nous avons fait deux séries de tableaux. Dans la première série (figure 5.1) nous avons représenté la distribution de la taille des motifs en fonction de la région d'où ils proviennent et de leur pourcentage de conservation. Pour la seconde série (figure 5.2), nous avons calculé la composition en nucléotides des séquences de départ et des motifs trouvés dans ces dernières. Les résultats sont regroupés en fonction de leur provenance pour les séquences et en fonction de leur longueur et de leur provenance pour les motifs conservés. Ensuite, nous avons voulu évaluer nos résultats avec une approche statistique. Il nous fallait une manière de pouvoir déterminer si les interactions obtenues entre les récepteurs n'étaient pas du simplement au hasard. Par conséquent, nous avons décidé de faire une simulation

Monte Carlo sur nos données. Cette simulation consiste à générer de façon aléatoire des valeurs pour les variables qui sont incertaines. Ainsi nous pouvons déterminer le niveau de fiabilité de nos résultats. Notre but était de pouvoir évaluer la quantité moyenne de liens (qui représente un partage de motifs communs) que pouvait avoir un récepteur avec les autres. Alors pour chacune des régions étudiées (région 5', région codante et région 3') et pour chacun des pourcentages de conservation, nous avons généré 100 matrices de façon aléatoire dont chacune contient le même nombre de liens que nos résultats expérimentaux. Par exemple, pour la région codante et avec un pourcentage de conservation de 70, nous avons obtenu 57 arêtes entre les récepteurs. Donc, nous avons généré 100 matrices avec 57 liens chacune et évaluer la moyenne de liens par récepteur. Il est important de mentionner que nous n'avons pas respecté le fait que certains nœuds ont plus d'arêtes que d'autres lors de la simulation. Par conséquent, notre distribution des arêtes est uniforme.

Nous avons aussi effectué un test de chi^2 sur les motifs trouvés avec MEME pour les trois régions : la région 5' avec des motifs conservés à 50 %, la région codante avec des motifs conservés à 80% et la région 3' avec des motifs conservés à 50 %. Seuls les résultats de MEME ont été utilisés pour ce test puisque les résultats obtenus avec BioProspector ne nous permettaient pas de retrouver le signal phylogénétique. De plus, nous savons que présentement il y a 27 hétérodimères connus possibles avec les 42 récepteurs que nous avons choisis. Ce test nous a permis de déterminer si le nombre de paires de récepteurs qui sont connus pour hétérodimériser que nous avons retrouvé dans nos graphes est plus élevé que ce à quoi nous pouvions nous attendre à trouver simplement par hasard.

Par la suite, afin de déterminer si les séquences de départ utilisées étaient bien annotées, nous avons comparé nos motifs trouvés avec des facteurs de transcription connus. A l'aide du site Consite (http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite; Sandelin et al., 2004), nous avons pu comparer les motifs conservés à 80 % et trouvés par MEME pour la région 5' et 3'. Le seuil de comparaison (cutoff) utilisé était de 80 et les motifs considérés comme conservés avec des pointages (score) au-delà de 7.0.

Finalement, nous voulions aussi analyser la corrélation entre les niveaux d'expression des récepteurs dans les divers tissus du corps humain. Pour réaliser cette analyse, nous avons utilisé les données de SymAtlas (http://symatlas.gnf.org/SymAtlas/). Pour chacun des récepteurs, nous avons évalué, à l'aide de la fonction fournie sur le site, l'expression de ceux-ci et reportés ceux qui avaient un niveau de corrélation entre 0.5 et 0.9 (par tranche de 0.1).

Chapitre 5

Résultats

5.1 Distribution de la taille des motifs

Une fois que la recherche de motifs conservés entre les espèces fût terminée, nous avons voulu classer les motifs qui ont été trouvés. La première idée qui nous vient à l'esprit est un classement selon la région où ils ont été retrouvés et selon leur longueur (figure 5.1). Ces résultats s'appliquent seulement pour les motifs trouvés par MEME, car pour BioProspector il faut spécifier le nombre de motifs que l'on recherche ainsi que leur longueur exacte. Les résultats obtenus sont conformes avec ce à quoi nous pouvions nous attendre, c'est-à-dire que la région codante comporte beaucoup de motifs de longueur 15 (de grandes zones sont conservées) et peu de motifs de petites longueurs, tandis que les régions en amont (upstream) et en aval (downstream) ont moins de motifs de longueur 15 et la distribution des longueurs de motifs est beaucoup plus uniforme.

5.2 Composition en nucléotides

Ensuite, nous voulions vérifier s'il n'y avait pas de biais de composition, en terme de nucléotide (surplus d'une ou deux bases), dans nos séquences de départ ainsi que dans les motifs trouvés (figure 5.2). On peut remarquer qu'en général, pour les motifs de petites longueurs, il y a souvent une saturation d'un nucléotide en particulier et ces excès sont plutôt rares en 5' ou 3'. Par contre, au fur et à mesure que l'on augmente la longueur des motifs il semble y avoir une stabilisation du niveau de composition des nucléotides pour les motifs des trois régions concernées. Il est important de noter que dû à la saturation des petits motifs, il peut y avoir une augmentation du nombre de motifs conservés entre les récepteurs augmentant du

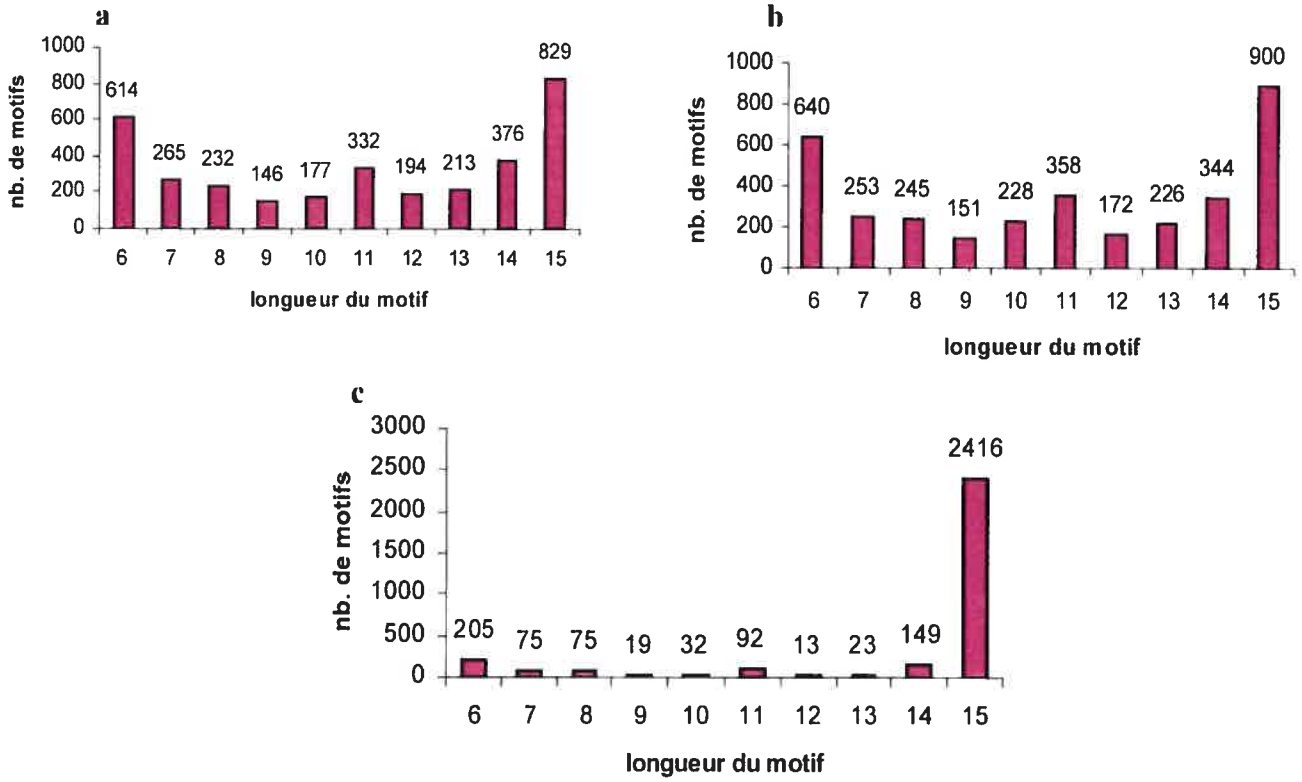

FIG. 5.1 : Distribution du nombre de motifs en fonction de leur taille.**a** | région en 5' de la région codante. **b** | région en 3' de la région codante. **c** | région codante.

même coup le nombre de faux-positifs. Ceci est facile à comprendre puisque si on modifie un seul nucléotide dans un motif de 4 lettres, on peut augmenter ou diminuer de 25 % le niveau de similarité avec un autre motif de même longueur. On peut aussi remarquer la sur-représentation des C et G en 5' (figure 5.2 a) qui est très surprenante puisque cette région est riche en A et T. Ce phénomère pourrait être expliqué par un biais méthodologique qui n'a pas été détecté. En ce qui concerne les séquences de départ (figure 5.2 d), leur composition semble assez bien distribuée, donc les résultats obtenus avec la recherche de motifs ne provient pas nécessairement des séquences initiales.

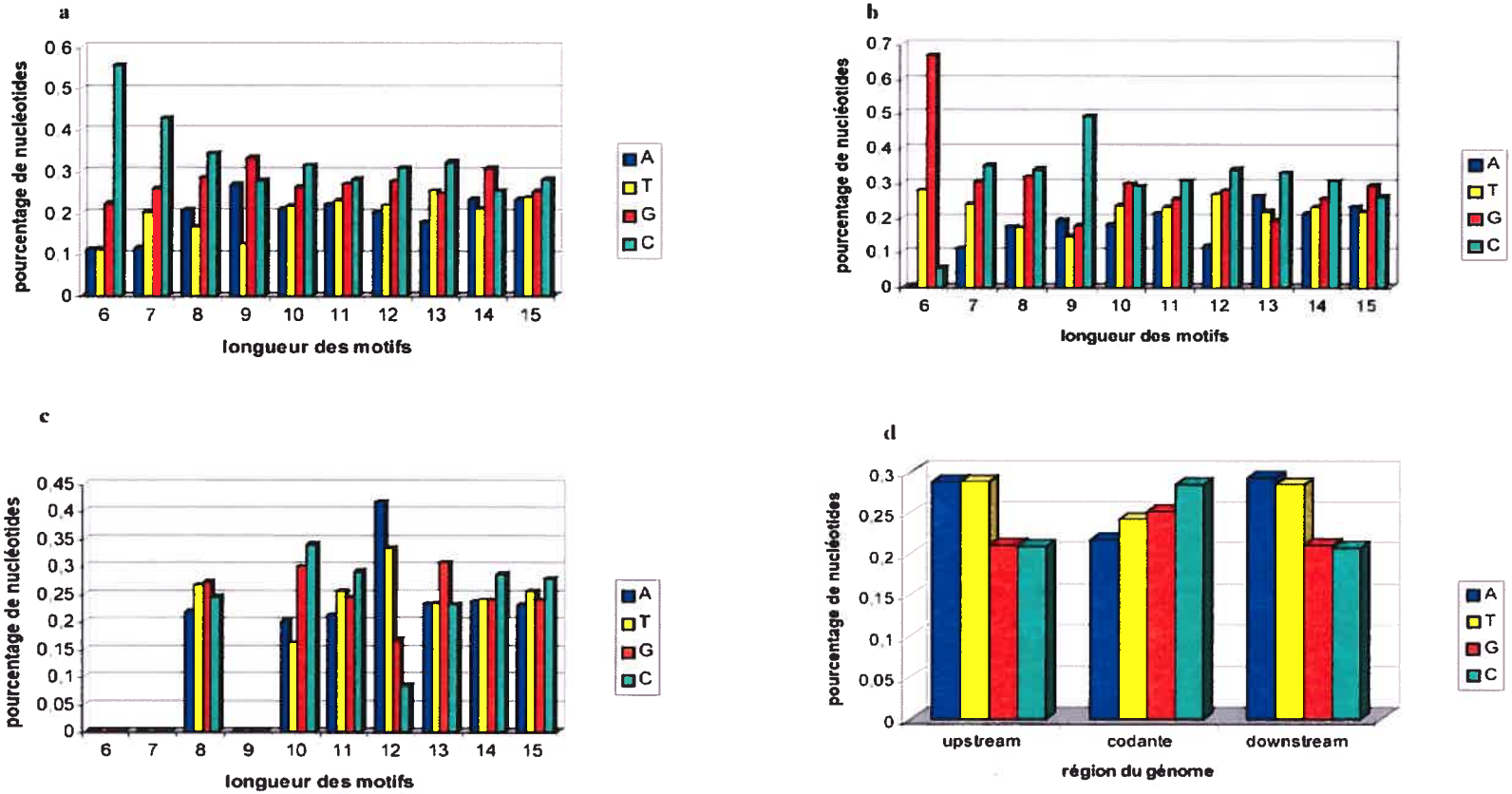

FIG. 5.2 : Composition en nucléotides. **a** | motifs de la région en 5'. **b** | motifs de la région en 3'. **c** | motifs de la région codante. **d** | séquences de tous les récepteurs.

5.3 Réseaux d'interactions

Avant de commencer l'analyse des résultats obtenus pour cette section, la figure 5.3 résume la méthode qui a été suivie pour effectuer la recherche, la comparaison et la classification des motifs trouvés.

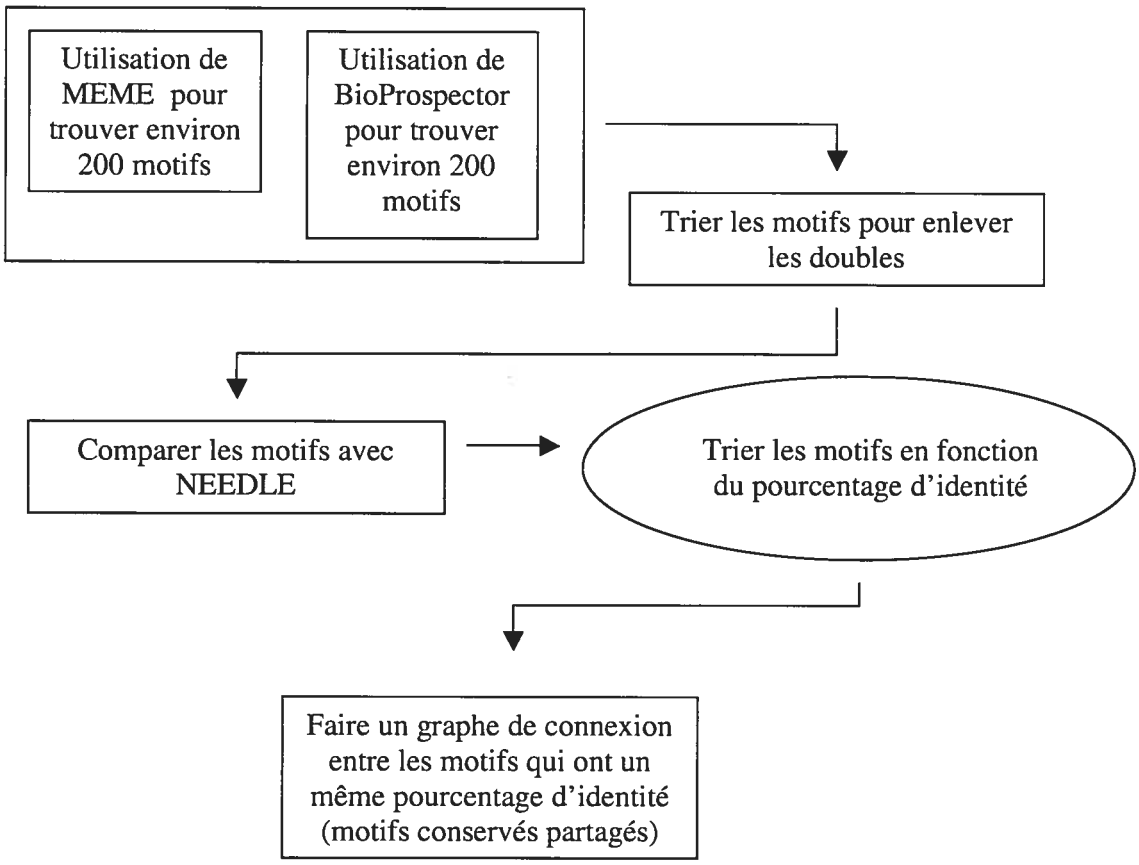

FIG. 5.3 : Organigramme de la méthode pour réaliser les réseaux de motifs conservés partagés.

Plusieurs réseaux ont été créés, mais seulement certains d'entre eux seront présentés dans cette section. Pour commencer, chaque graphe comporte une valeur de « cutoff » qui indique le nombre minimal de motifs partagés (chiffre sur l'arête) entre deux récepteurs. Cette valeur a été déterminée afin de conserver environ les 50 meilleurs résultats du graphe. Ensuite, les récepteurs sont colorés selon la sous-

classe à laquelle ils appartiennent. Ces sous-classes sont les mêmes que dans l'arbre phylogénétique des RCPG (figure 3.4). Finalement, chaque réseau a aussi un graphique, dans le coin supérieur droit, qui montre la distribution du nombre de motifs partagés (nombre sur l'arête) entre toutes les paires de récepteurs et combien de paires de récepteurs ont ce nombre de motifs partagés.

Les résultats vont être découpés en trois sections : 1) comparaison des résultats de MEME et BioProspector avec les régions en amont et aval de la région codante d'une longueur de 1KB et la région codante, 2) comparaison des résultats de MEME pour des régions (5' et 3') de longueurs 1KB et 3KB et 3) comparaison des résultats de BioProspector pour des motifs à 2-blocs avec ceux obtenus à la première section.

Pour la première section, le premier graphe est celui de MEME pour la région codante (figure 5.4). On peut constater que les récepteurs se sont regroupés plus ou moins selon leur sous-classe (figure 5.4 a). Cette méthode retrouve le signal phylogénétique qui a permis leur classification. Ce n'est pas surprenant, car cela revient en grande partie à mesurer la similarité entre deux séquences. Et comme en général deux séquences qui sont groupe-frère sont plus similaires entre elles, elles doivent être connectées par beaucoup de liens. Il est cependant intéressant de noter que ce n'est pas systématique. En particulier, la famille MECA n'est pas du tout retrouvée. Cela suggère que soit la saturation est importante et que du même coup elle modifie le niveau de similarité entre les motifs, soit il existe d'autres signaux dans les séquences codantes qui n'ont pas été nécessairement détectés par l'analyse phylogénétique. Ensuite, à l'aide du graphique (figure 5.4 b) on peut voir aussi que la moyenne des chiffres sur les arêtes (nombre de motifs partagés) pour une paire de récepteur est proche de 7 et que seulement quelques arêtes se retrouvent avec ce nombre ou plus (par exemple : CCR2-CCR5, M1-M5 et GluM1-GluM5). Donc, il serait intéressant d'observer comment ces paires de récepteurs se comportent dans les autres réseaux.

Nous avons aussi voulu déterminer le niveau de confiance que l'on pouvait avoir pour les réseaux d'interactions en fonction des éléments corrélés qui sont

reconnus pour hétérodimériser et que nous avons trouvés avec MEME (1KB) (Tableau IV). Pour ce faire nous avons effectué un test de chi^2

	Hétérodimère	Non-Hétérodimère
Nombre Observé (O)	9	30
Nombre Attendu au hasard(A)	2	37

Tableau IV : Test de chi^2 sur les éléments corrélés de la région codante trouvés à l'aide du logiciel MEME.

Ici, la valeur de $\chi^2 = 25.82$ excède la valeur maximale du tableau pour les valeurs de chi^2 (pour p=0.001, χ^2=10.83). Par conséquent, nous pouvons être confiants à 99,9% que notre résultat n'est pas dû au simple hasard. Le partage de motifs dans la région codante est donc un très bon prédicateur de la dimérisation des récepteurs et ceci est un peu surprenant. C'est-à-dire, qu'à première vue il semble que la dimérisation soit simplement une information contenue dans la région codante.

Par la suite, si l'on compare le graphe de la figure 5.4 a avec celui de Bioprospector (figure 5.5 a) pour la même région et le même pourcentage de conservation, on peut remarquer certaines similitudes (arêtes oranges) entre les deux. Par contre, si l'on essaie de corréler les arêtes similaires en fonction de leur nombre de motifs conservés à l'aide d'un graphe, on constate qu'il n'y en a pas (figure 5.6). Mais dans le cas de BioProspector, les récepteurs partagent plus de motifs (figure 5.5 b) et ils sont moins ordonnés, c'est-à-dire que lorsque que l'on compare les paires de récepteurs trouvés avec BioProspector et les groupes frères de l'arbre phylogénétique (figure 3.4), on retrouve seulement 6 paires de récepteurs conservés entre les deux figures tandis qu'avec MEME on en retrouve 9. Donc le signal phylogénétique est plus faible pour la figure 5.5 a que pour la figure 5.4 a.

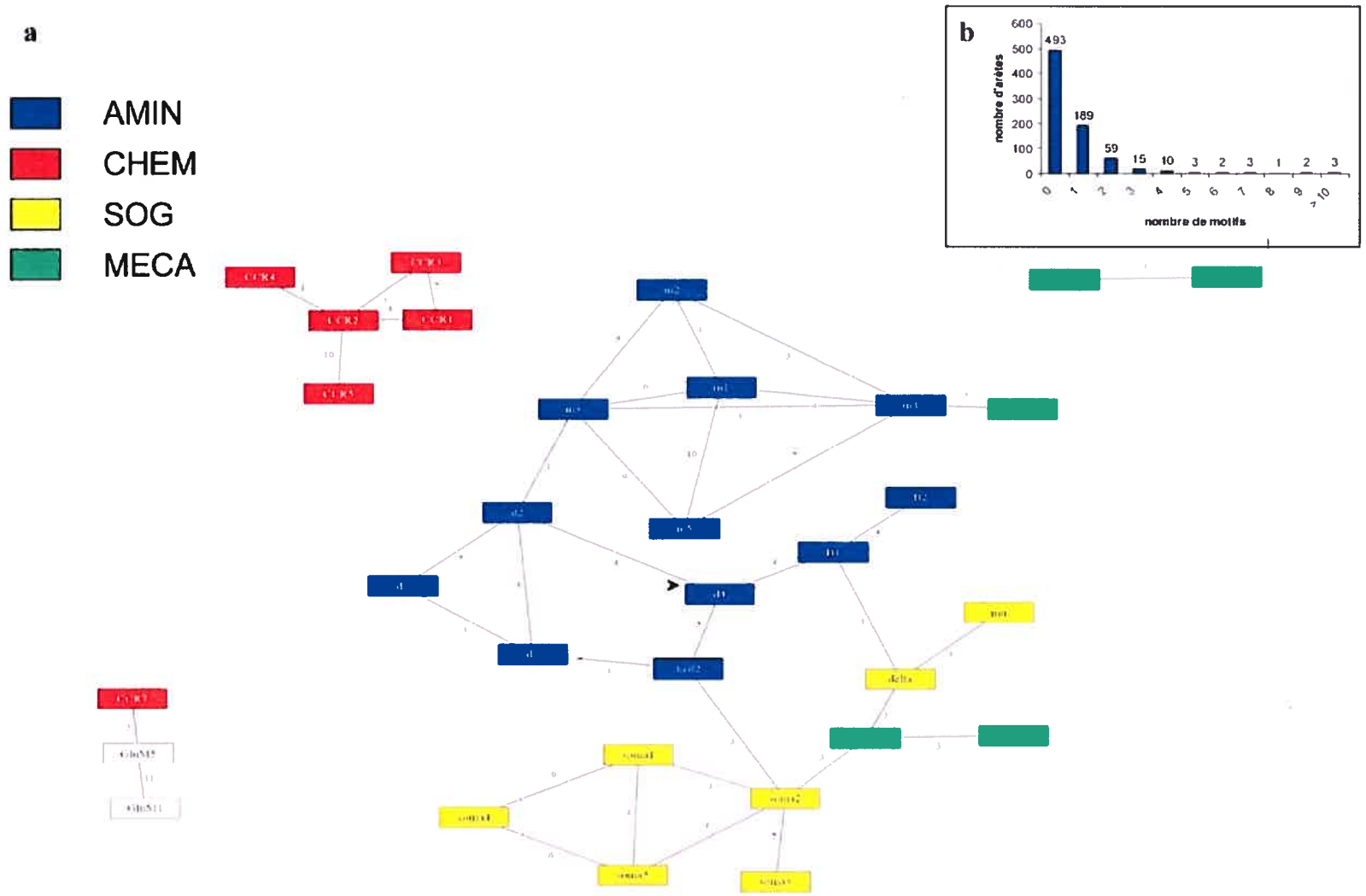

FIG. 5.4: **a** | Distribution des récepteurs qui partagent des motifs conservés à 80% pour la région codante et le cutoff est de 3 (MEME-1KB). **b** | distribution du poids des arêtes.

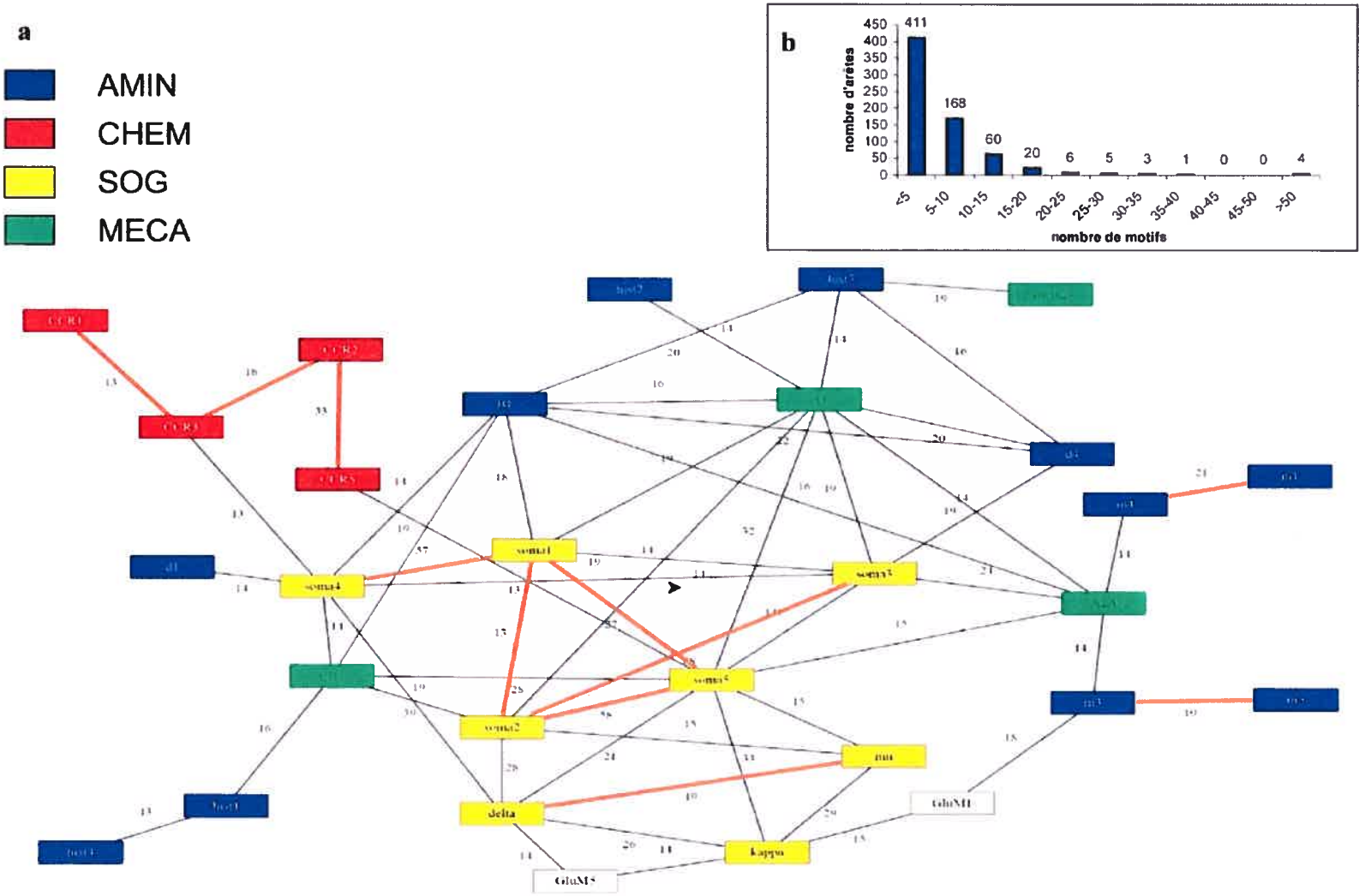

FIG. 5.5: **a** | Distribution des récepteurs qui partagent des motifs conservés à 80% pour la région codante et le cutoff est de 13 (BioP-1KB). **b** | distribution du poids des arêtes.

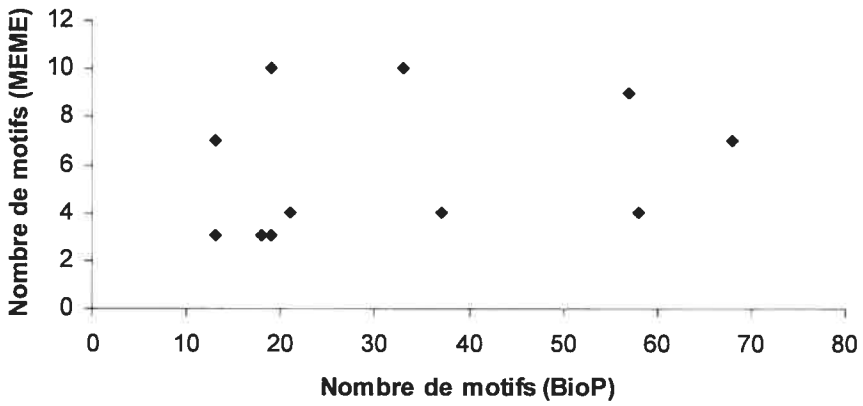

FIG. 5.6 Nombre de motifs trouvés par les deux outils (MEME et BioP) pour les arêtes similaires (oranges) de la figure 5.5

Les résultats obtenus pour les régions 5' et 3', avec MEME, sont représentés à la figure 5.7 et à la figure 5.8. Les similitudes entre les deux sont encore une fois montrées à l'aide d'arêtes orange sur la figure 5.8. On peut remarquer dans ces deux graphes que certains récepteurs semblent agir comme des « hubs », c'est-à-dire que ces récepteurs ont beaucoup plus de liens que les autres (par exemple : Kappa et CCR2 sur la figure 5.7) et qu'il y a très peu d'arêtes communes.

C'est surtout le récepteur Kappa qui est intrigant puisqu'il a un comportement similaire (agit comme un hub) dans les deux régions (5' et 3') à la fois. Soit il s'agit d'un biais méthodologique ou bien il y a une raison biologique derrière tout cela.

Pour pouvoir déterminer la cause, il faut évaluer le niveau de signification du résultat obtenu à l'aide d'une méthode Monte Carlo. Cela nous permettra de déterminer les nombres d'arêtes attendus au hasard pour chacun des récepteurs.

Pour la région en 5', nous avons généré 100 matrices (42 X 42) comportant chacune le même nombre de liens que la figure 5.7, tous distribués au hasard (figure 5.9). On peut remarquer que seulement 3 récepteurs sur 4200 ont 8 arêtes. Donc, il est très peu probable qu'une telle situation (Kappa qui a 24 arêtes) se produise par hasard, surtout si un ou des récepteurs ont beaucoup plus d'arêtes que le nombre maximal (8) atteint lors de la simulation.

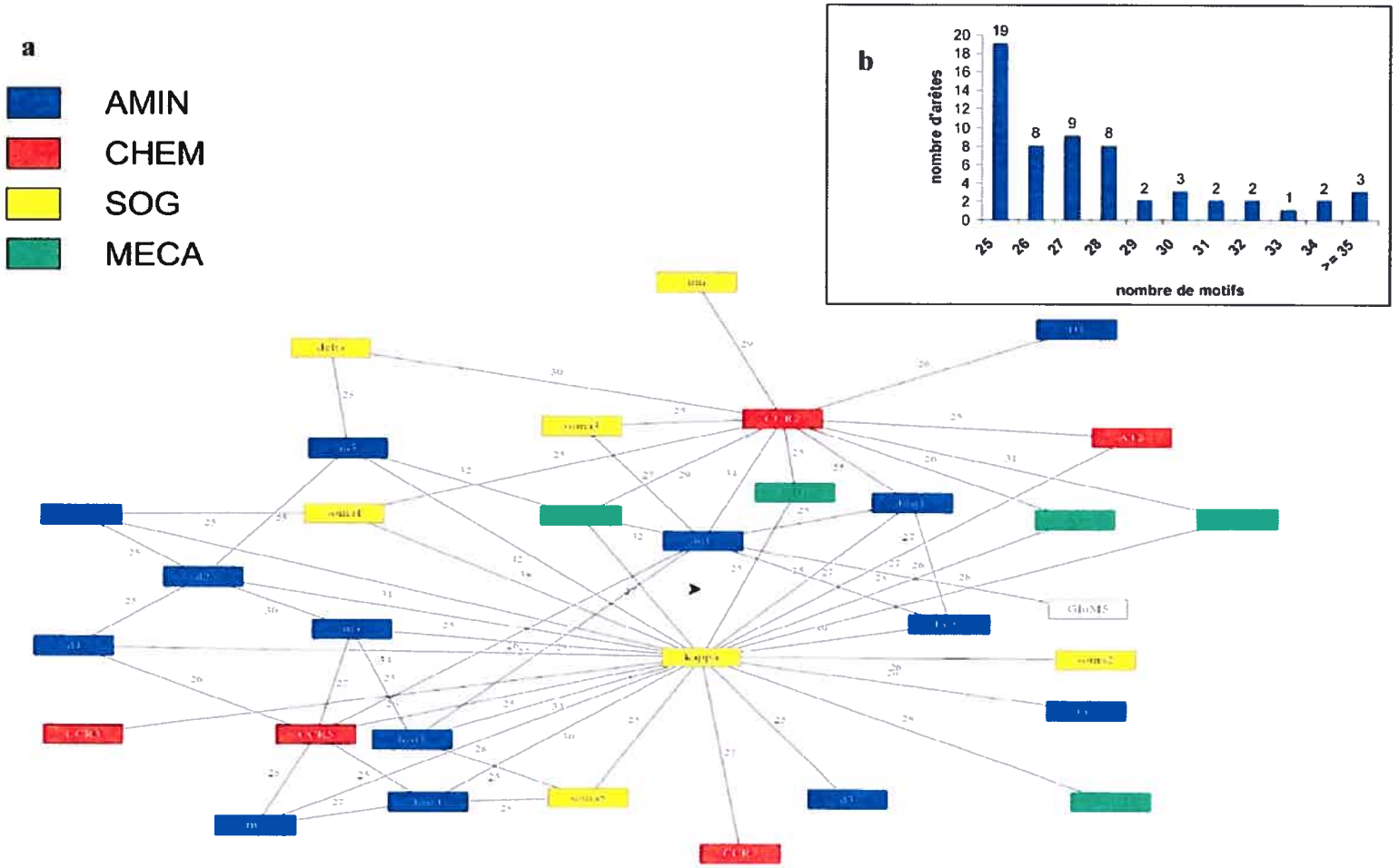

FIG. 5.7: **a** | Distribution des récepteurs qui partagent des motifs conservés à 50% pour la région 5' (upstream) et le cutoff est de 25 (MEME-1KB). **b** | distribution du poids des arêtes.

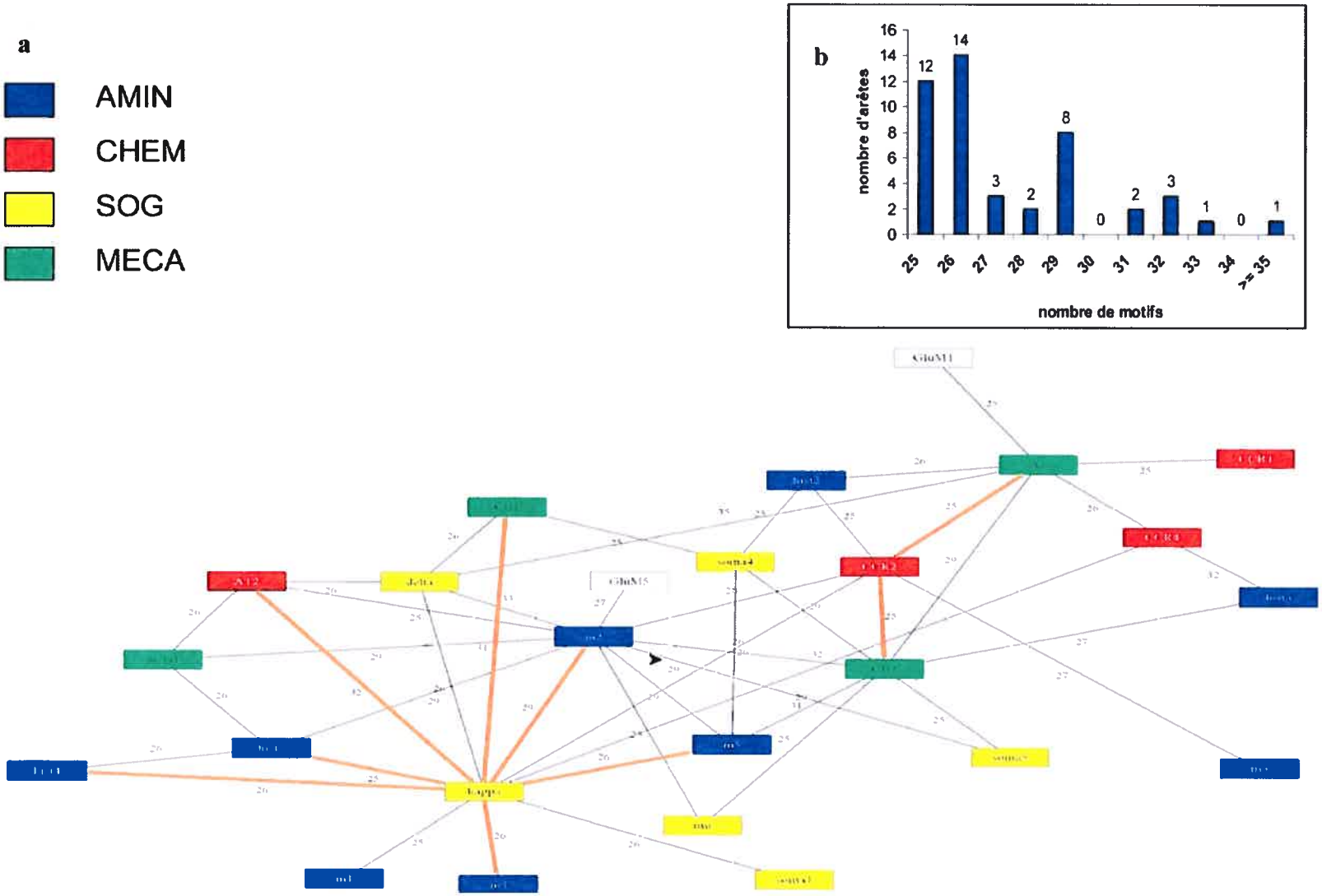

FIG. 5.8: **a** | Distribution des récepteurs qui partagent des motifs conservés à 50% pour la région 3' (downstream) et le cutoff est de 25 (MEME-1KB). **b** | distribution du poids des arêtes.

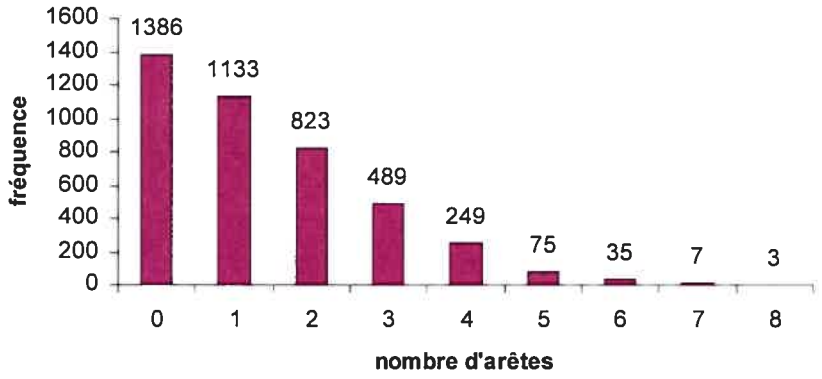

FIG. 5.9 : Simulation Monte Carlo pour la région en 5' de la région codante (MEME-1KB).

Il y a deux récepteurs dont la p-value < 0.01 et il s'agit du récepteur Kappa et CCR2. Ils ont respectivement 24 et 12 liens chacun. Ce sont les deux maximums observés pour la figure 5.7. Par conséquent, on peut affirmer que ces deux résultats sont significatifs. Alors si ce n'est pas dû au hasard, il se peut toujours que la cause soit un biais méthodologique non encore identifié (comme un biais de composition de Kappa), mais il semble tout de même y avoir une explication biologique. Une explication plausible serait que Kappa est présent dans pratiquement tous les tissus et que par conséquent il est normal qu'il partage des motifs similaires avec tous les autres récepteurs. Par contre, on pourrait dire que Delta et Mu sont aussi des opioids et qu'ils ne partagent pas autant de motifs. Alors il serait intéressant de creuser d'avantage la question.

Ensuite, nous avons fait le même test pour la région en 3' et avons obtenu des résultats similaires (figure 5.10).

Une fois les simulations terminées, nous avons aussi fait un test de chi^2 pour les régions 5' et 3'. Alors pour la région 5' (Tableau V), considérant que nous avons un degré de liberté de 1 et que notre de valeur de $\chi^2 = 0.518$, nous pouvons affirmer que nous avons une valeur non-significative (p-value = 0.95), c'est-à-dire que les résultats que nous retrouvons sont simplement dû au hasard.

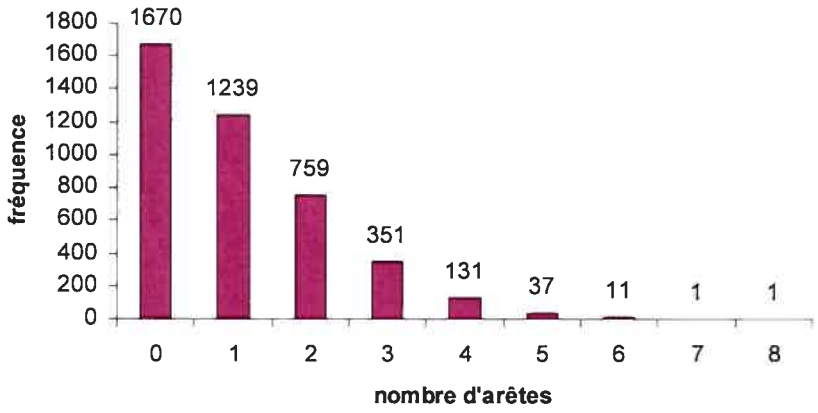

FIG. 5.10 : Simulation Monte Carlo pour la région en 3' de la région codante (MEME-1KB).

	Hétérodimère	Non-Hétérodimère
Nombre Observé (O)	3	56
Nombre Attendu au hasard (A)	2	57

Tableau V : Test de chi^2 sur les éléments corrélés de la région 5' trouvés à l'aide du logiciel MEME.

	Hétérodimère	Non-Hétérodimère
Nombre Observé (O)	3	43
Nombre Attendu au hasard (A)	2	44

Tableau VI : Test de chi^2 sur les éléments corrélés de la région 3' trouvés à l'aide du logiciel MEME.

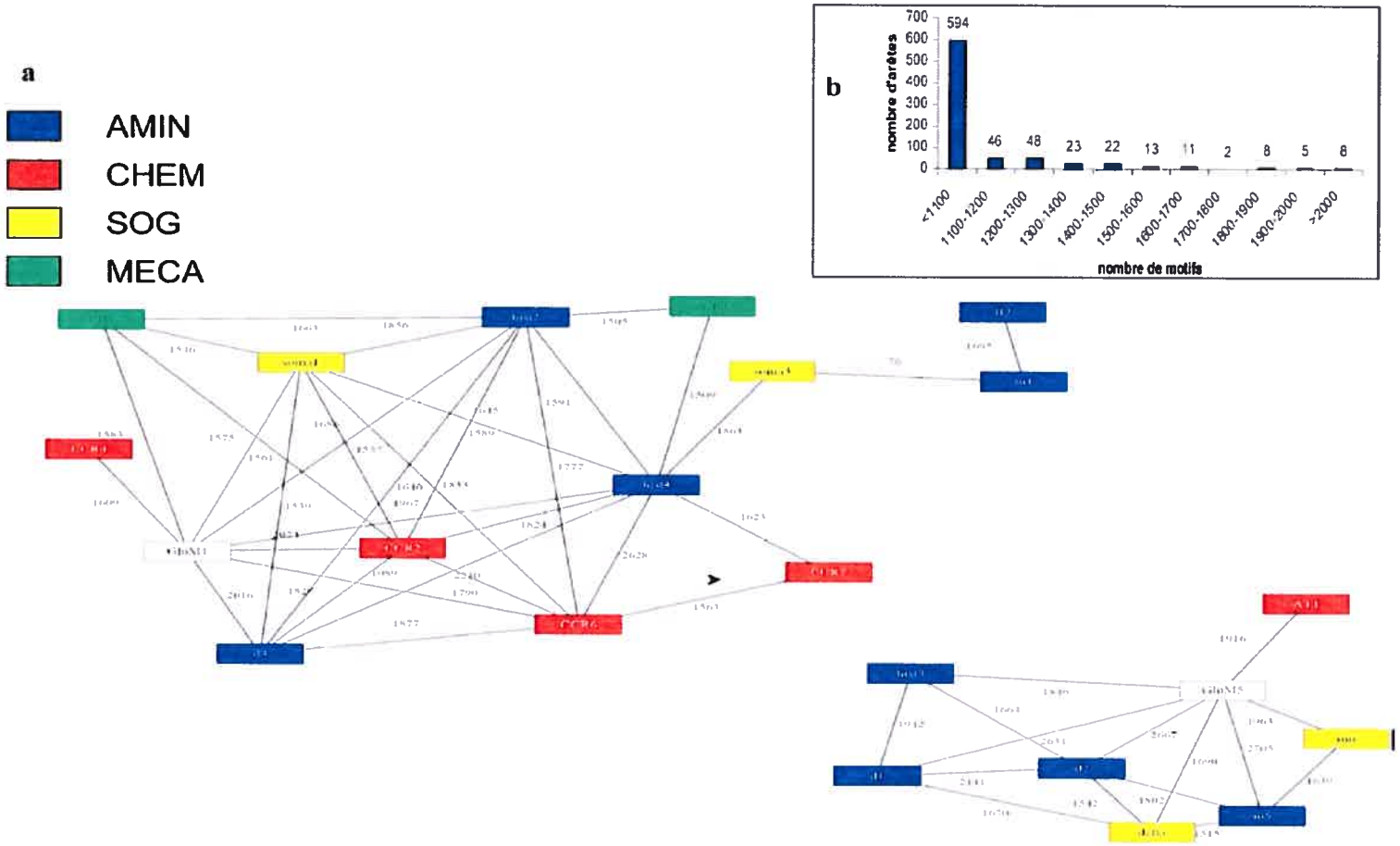

FIG. 5.11: **a** | Distribution des récepteurs qui partagent des motifs conservés à 50% pour la région 5' (upstream) et le cutoff est de 1500 (BioP-1KB). **b** | distribution du poids des arêtes.

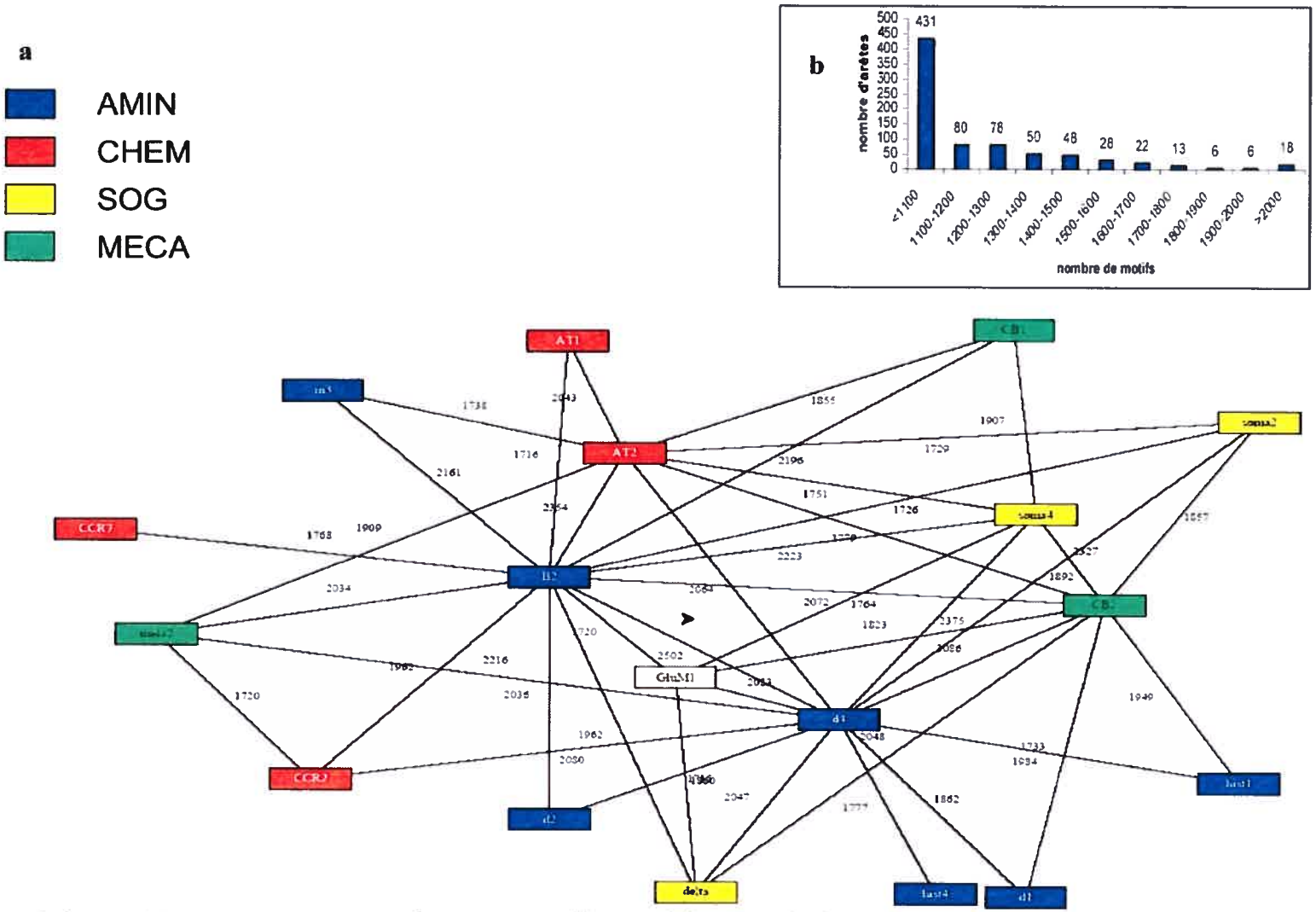

FIG. 5.12: **a** | Distribution des récepteurs qui partagent des motifs conservés à 50% pour la région 3' (downstream) et le cutoff est de 1700 (BioP-1KB). **b** | distribution du poids des arêtes.

On peut affirmer la même chose avec les résultats obtenus pour la région en 3' de la région codante (Tableau VI).

Les résultats de BioProspector, pour ces mêmes régions, sont à la figure 5.11 et 5.12. Lorsque l'on compare ces réseaux à ceux obtenus avec MEME, on remarque que certains récepteurs agissent (comme un « hub ») de la même façon, mais qu'ils ne sont pas les mêmes et qu'ils semblent non significatifs pour la plupart (figures 5.13 et 5.14) puisque leur nombre d'arêtes est similaire aux valeurs obtenues par hasard, sauf pour le récepteur B2 de la figure 5.12 qui se démarque avec 12 arêtes.

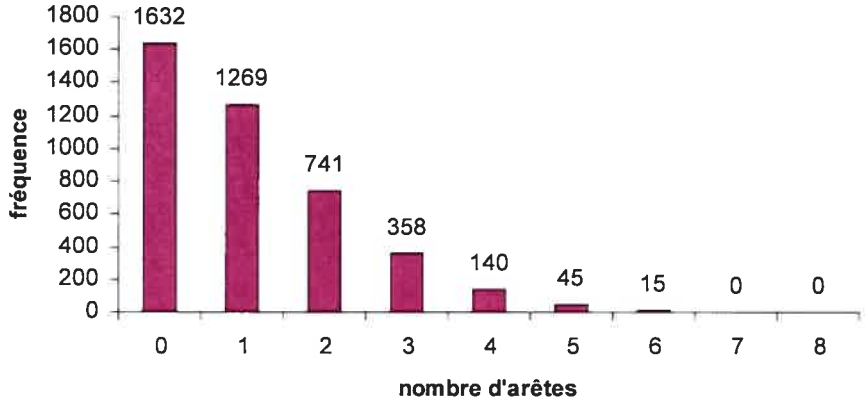

FIG. 5.13 : Simulation Monte Carlo pour la région en 5' de la région codante (BioP-1KB).

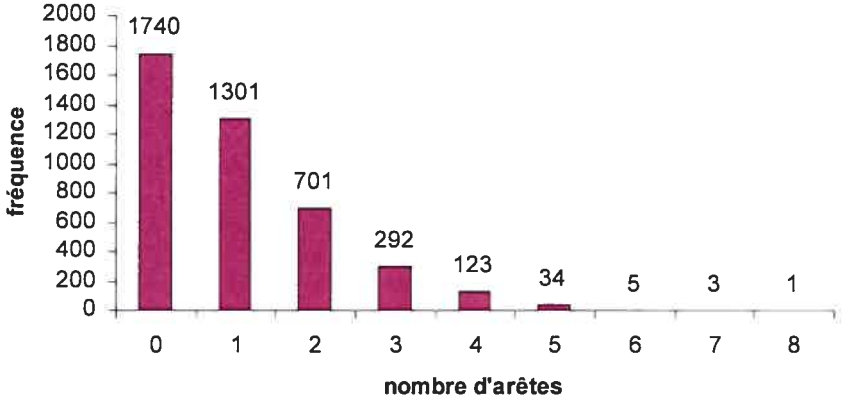

FIG. 5.14 : Simulation Monte Carlo pour la région en 3' de la région codante (BioP-1KB).

De plus, on peut constater qu'il n'y a aucune similarité entre les deux réseaux (5' et 3') générés à l'aide des motifs trouvés avec BioProspector. Par contre, les récepteurs partagent beaucoup plus de motifs conservés que ceux de MEME.

Maintenant pour la section 2, on compare les réseaux dont les motifs ont été trouvés avec MEME, mais avec des longueurs différentes pour les régions en 5' et 3'. La figure 5.15 représente la simulation Monte Carlo pour les séquences de longueur de 3 KB et les figures 5.16 et 5.17 représentent les motifs partagés entre les récepteurs pour les mêmes séquences. Encore une fois, les lignes oranges de la figure 5.17 représentent ce que ces deux graphes ont en commun. On voit ici, aussi, plusieurs récepteurs qui forment des « hubs », mais seulement deux sont communs aux deux graphes (M4 et M5). Lorsque l'on compare ces récepteurs avec la simulation Monte Carlo, il y en a que deux qui semble significatif : A2A (figure 5.16) et M5 (figure 5.17).

Si l'on compare ces deux réseaux aux figures 5.11 et 5.12, qui sont les réseaux obtenus avec MEME pour des séquences de longueur 1KB, il n'y a que très peu de ressemblance. Pour la région 5', il n'y que deux arêtes que l'on retrouve dans les deux réseaux (1KB et 3KB) : CCR5-M4 et D2-M4. Il en est de même pour la région 3', M2-M5 et M5-Soma4. On peut trouver cela problématique puisque le 1KB est inclus dans le 3 KB.

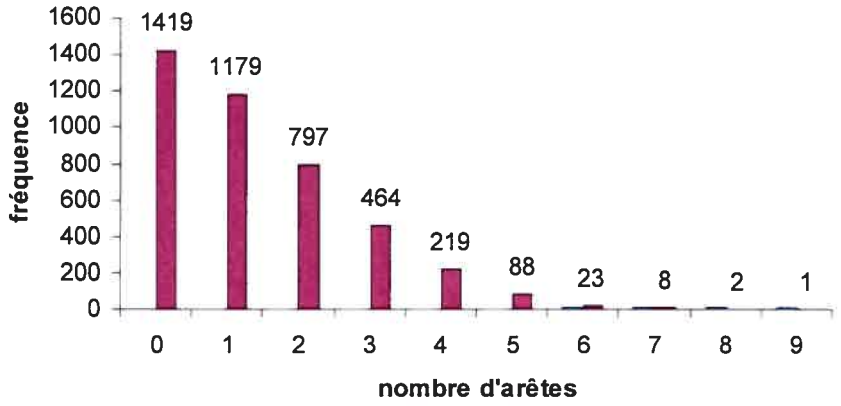

FIG. 5.15 : Simulation Monte Carlo pour la région en 5' et 3' de la région codante (MEME-3KB).

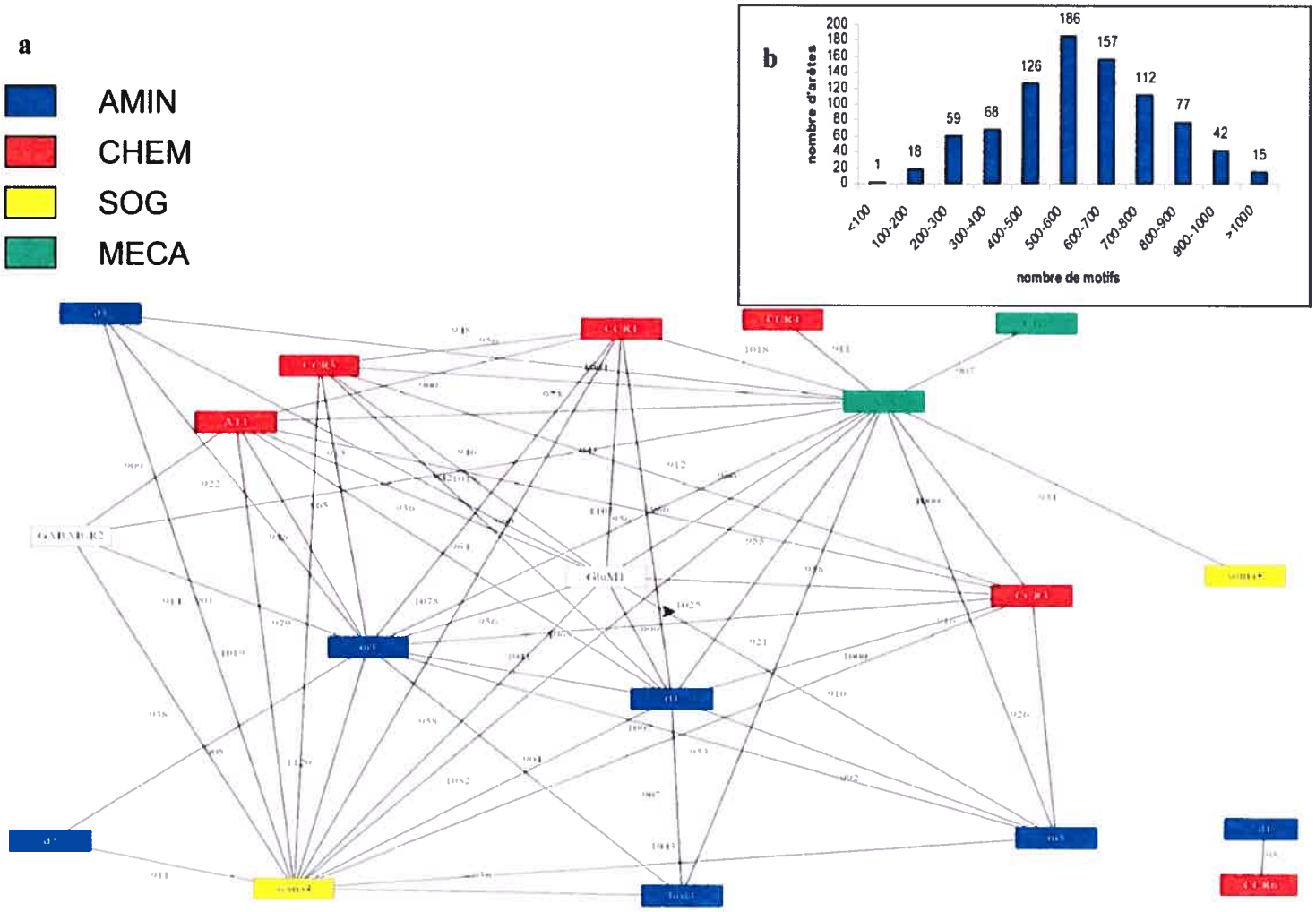

FIG. 5.16: **a** | Distribution des récepteurs qui partagent des motifs conservés à 50% pour la région 5' (upstream) et le cutoff est de 900 (MEME-3KB). **b** | distribution du poids des arêtes.

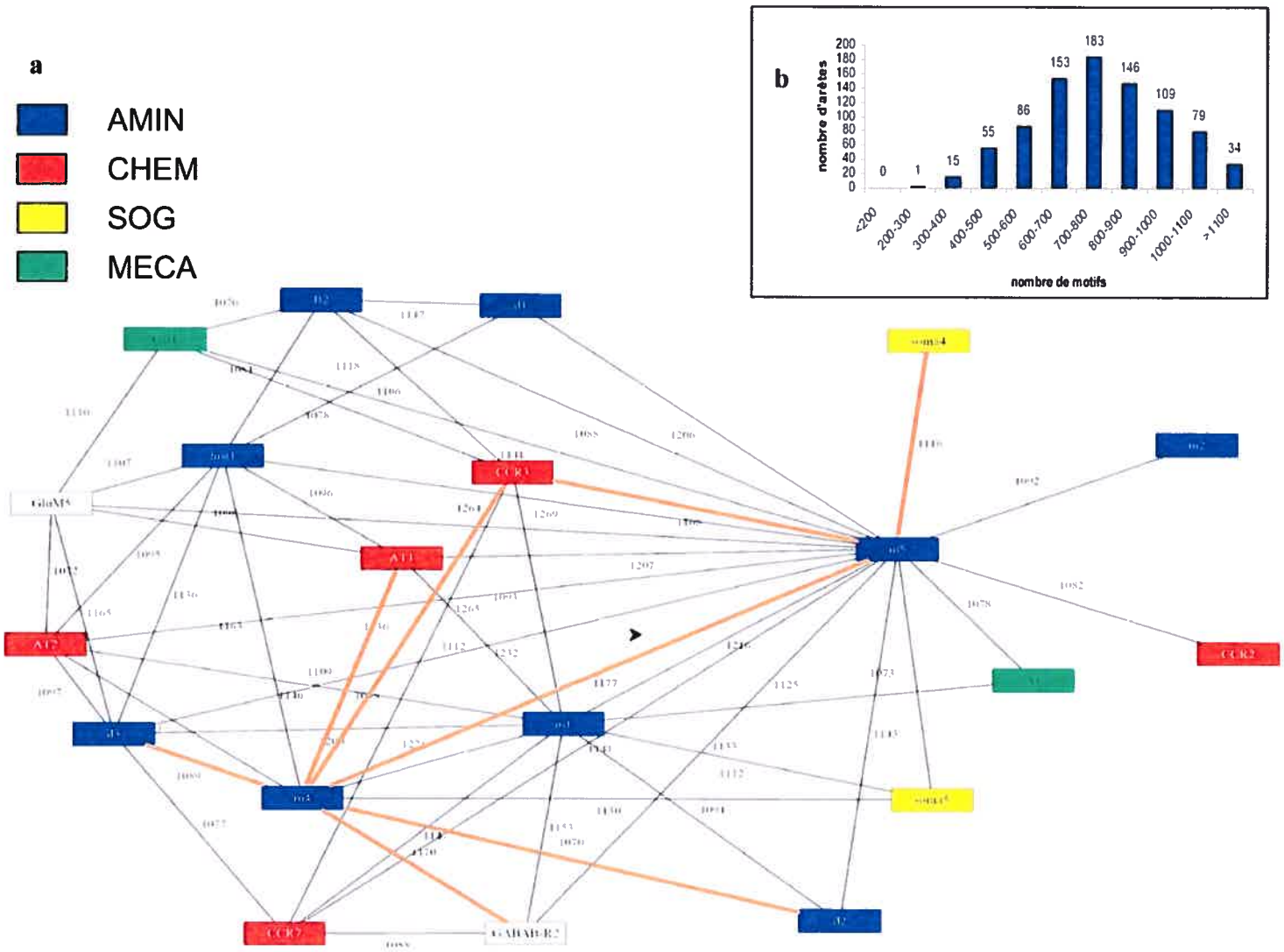

FIG. 5.17: **a** | Distribution des récepteurs qui partagent des motifs conservés à 50% pour la région 3' (downstream) et le cutoff est de 1070 (MEME-3KB). **b** | distribution du poids des arêtes.

Pour la section 3, nous avons décidé de comparer les motifs standards avec les motifs à 2-blocs, en utilisant les résultats de BioProspector (1 KB) comme motifs standards. Nous avons effectué cette comparaison dans le but de déterminer si les deux méthodes permettaient de retirer la même information ou non. Pour chacune des régions, des motifs à 2-blocs avec des longueurs de gap de 5, 10 et 15 ont été recherchés. Suite à la comparaison des motifs pour une même région, les réseaux qui ont été construits avec les motifs qui contenaient un gap de 10 semblaient les plus appropriés à une comparaison avec les motifs standards de BioProspector. Ce choix se base principalement sur le fait que la plupart des motifs découverts avec un gap de 5 ou 15 sont retrouvés dans ceux qui ont un gap de 10. De plus, si nous les comparons aux motifs standards de BioProspector, ils s'en rapprochent beaucoup plus, en terme du nombre d'arêtes similaires, que les autres réseaux à gaps de 5 et 15.

La figure 5.18 représente les motifs trouvés pour la région codante avec un gap de 10 et les lignes oranges représentent les arêtes qui sont communes, pour la même région, avec la figure 5.5 (BioP-1KB). Il en est de même pour la figure 5.19 (région en 5' de la région codante) et la figure 5.20 (région en 3' de la région codante). À la figure 5.19b, il y a un résultat très intéressant ; on peut voir qu'il y a 34 arêtes qui ont plus de 1000 motifs conservés et que ce résultat est en dehors d'une distribution gaussienne. En effet, la distribution gaussienne des autres figures suggèrent très fortement qu'il y a peu ou pas de signal contrairement à cette dernière.

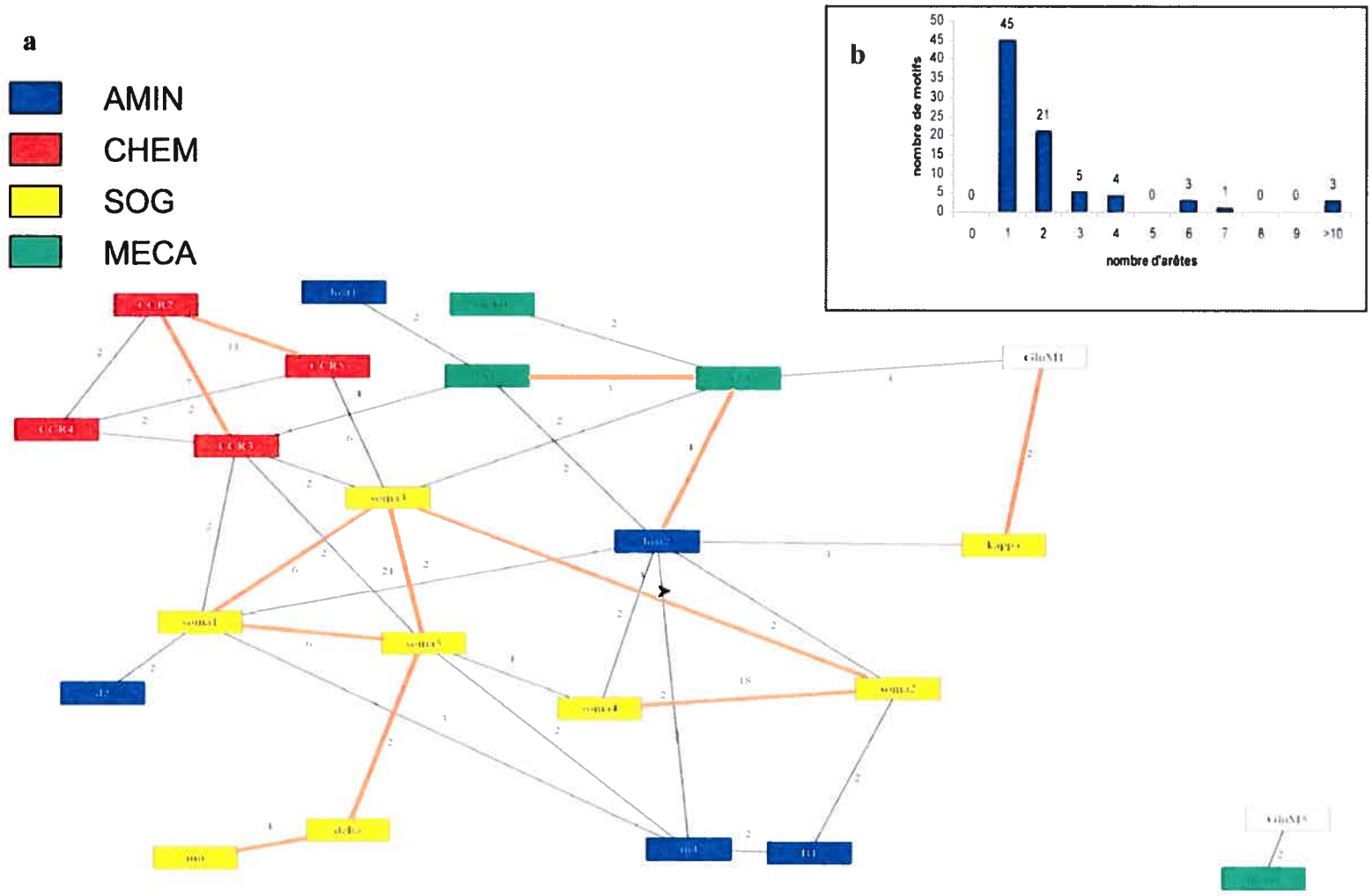

FIG. 5.18: **a** | Distribution des récepteurs qui partagent des motifs conservés à 80% pour la région codante et le cutoff est de 2 (BioP-2-blocs). **b** | distribution du poids des arêtes.

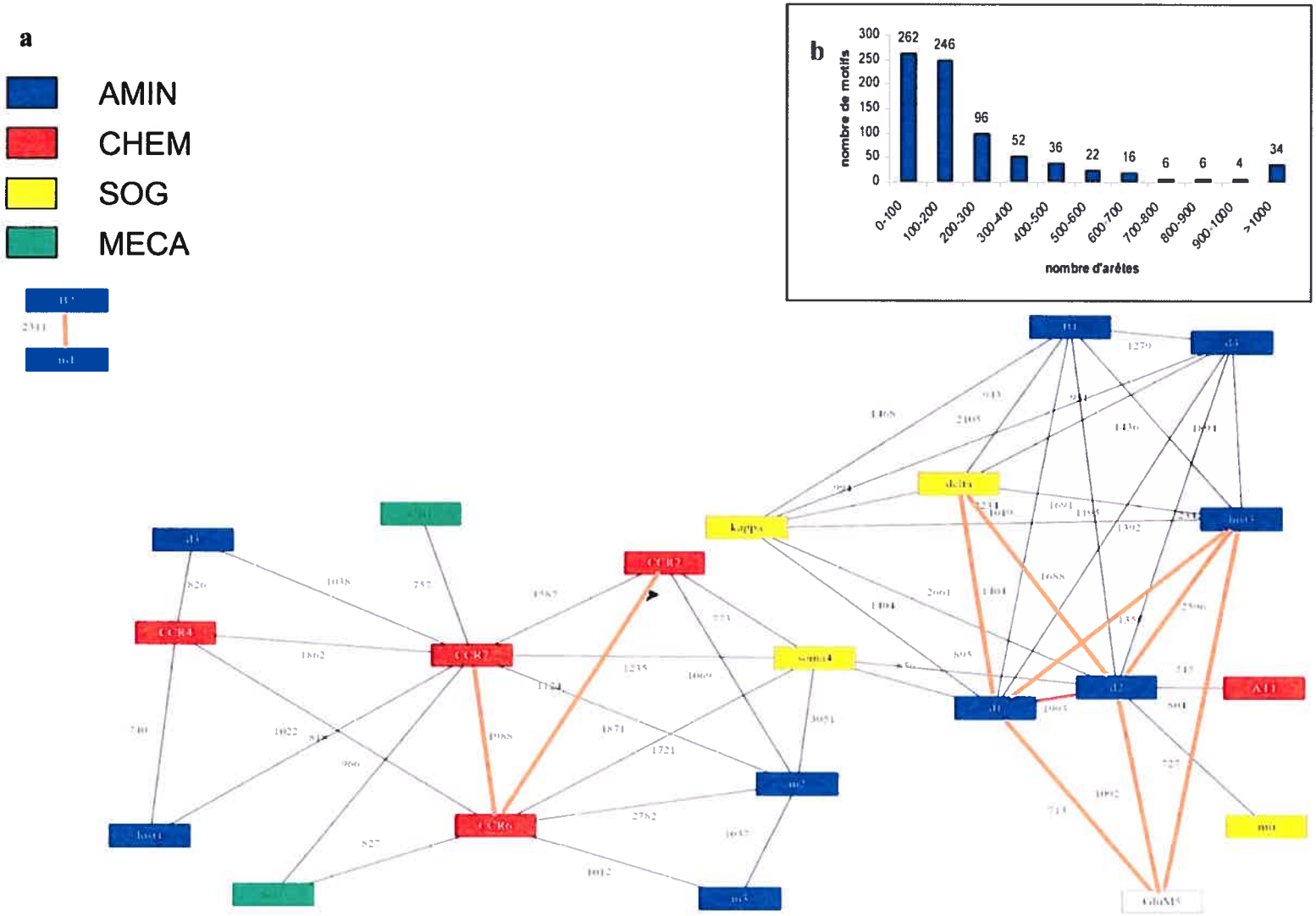

FIG. 5.19: **a** | Distribution des récepteurs qui partagent des motifs conservés à 50% pour la région 5' et le cutoff est de 700 (BioP-2-blocs). **b** | distribution du poids des arêtes.

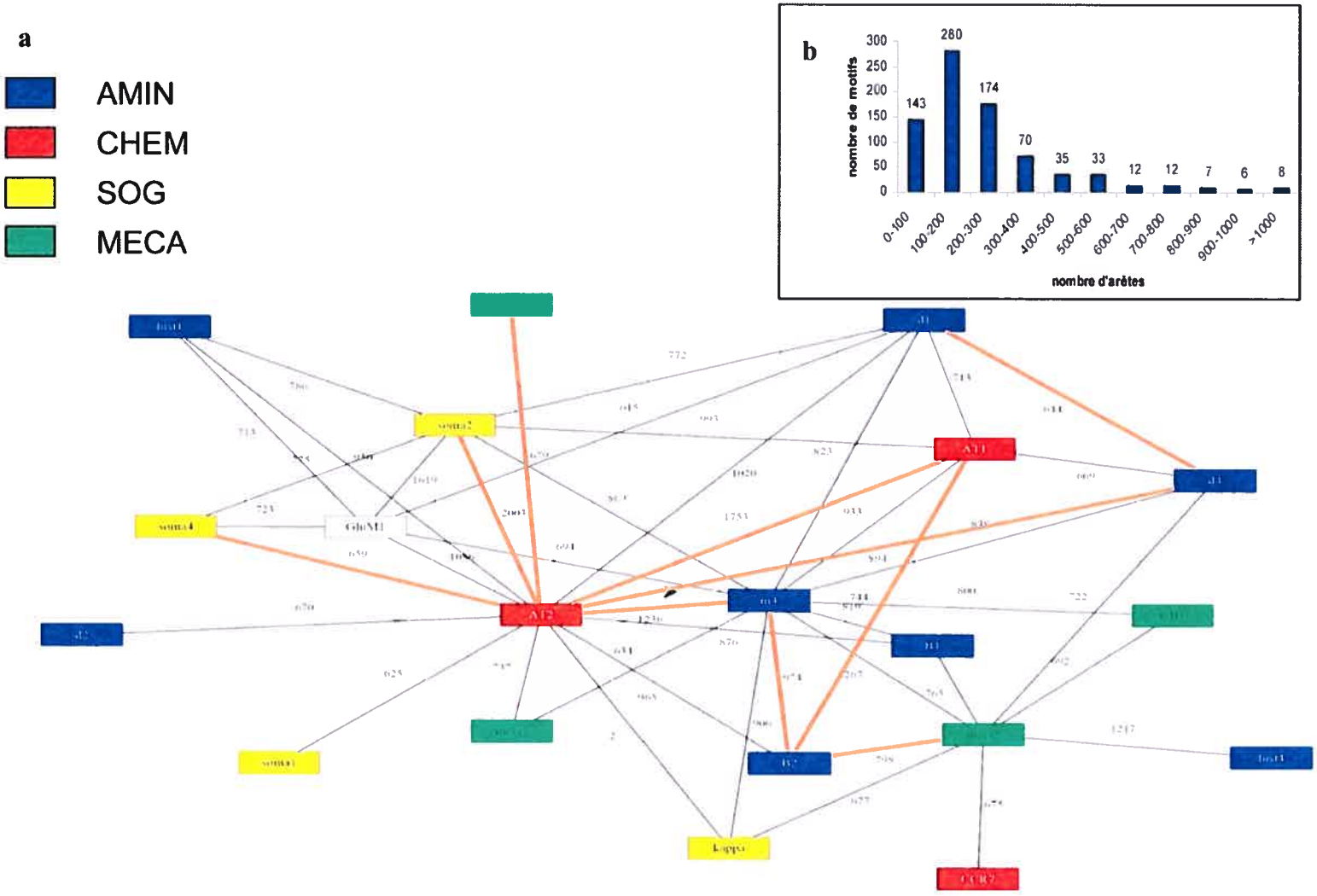

FIG. 5.20: **a** | Distribution des récepteurs qui partagent des motifs conservés à 50% pour la région 3' et le cutoff est de 600 (BioP-2-blocs).
b | distribution du poids des arêtes.

5.4 Résumé des récepteurs corrélés

Au tout début de ce mémoire, nous avons indiqué que deux récepteurs peuvent hétérodimériser et qu'il serait intéressant de pouvoir prédire une telle interaction. Par conséquent, on retrouve ici trois tableaux, un pour chacune des régions étudiées, qui indiquent le nombre d'arêtes qui existait dans un graphe en particulier et dont les deux récepteurs associés à cette arête sont reconnus pour être capable de former un hétérodimère.

En 5' de la région codante (upstream)

	Pourcentage de simillarité (%)	MEME (1KB)	MEME (3KB)	BioProspector (1KB)	BioProspector (2-blocs)
Nombre de récepteurs corrélés (arêtes)/nombre d'arêtes total /nombre attendu par hasard	50	4/59/2	0/57/2	1/47/2	3/50/2
	60	1/41/1	1/42/1	2/42/1	3/56/2
	70	0/23/1	1/35/1	2/54/2	2/46/2
	80	0/23/1	3/49/2	1/48/2	1/37/1
	90	0/2/0	1/40/1	0/46/2	1/11/0

Tableau VII : Pour la région 5'; nombre d'arêtes connectant deux récepteurs peuvant hétérodimériser. Le classement se fait en fonction du pourcentage de conservation entre les motifs ainsi qu'avec la méthode qui a servi pour détecter les motifs.

La région codante (gène)

	Pourcentage de simillarité (%)	MEME (1KB)	BioProspector (1KB)	BioProspector (2-blocs)
Nombre de récepteurs corrélés (arêtes)/nombre d'arêtes total /nombre attendu par hasard	50	1/61/3	2/44/2	3/48/2
	60	1/41/2	5/57/3	3/48/2
	70	7/55/3*	4/48/2	4/52/3
	80	9/39/2*	6/56/3*	2/36/2
	90	4/29/1*	4/50/3	2/23/1

Tableau VIII : Pour la région codante; cf tableau VII

En 3' de la région codante (downstream)

	Pourcentage de simillarité (%)	MEME (1KB)	MEME (3KB)	BioProspector (1KB)	BioProspector (2-blocs)
Nombre de récepteurs corrélés (arêtes)/nombre d'arêtes total /nombre attendu par hasard	50	4/46/2	0/54/2	2/43/2	1/45/2
	60	2/33/1	0/40/2	2/56/2	3/52/2
	70	1/33/1	0/43/2	2/51/2	4/48/2
	80	0/31/1	0/40/2	2/56/2	2/28/1
	90	0/6/0	0/39/2	1/41/2	1/6/0

Tableau IX : Pour la région 3';cf tableau VII

* résultats qui ont un test de chi^2 significatif.

Suite à ces résultats, on peut remarquer qu'on retrouve beaucoup plus de récepteurs corrélés avec la région codante qu'avec les deux autres régions. Par conséquent, il y aurait peut-être une piste à explorer dans cette direction.

5.6 Comparaison avec des facteurs de transcription

Afin de déterminer si les séquences d'ADN que nous avons utilisées étaient bien annotées, nous avons voulu comparer certains motifs trouvés avec des facteurs de transcription déjà connus. Nous avons donc voulu effectuer un contrôle négatif qui consiste à rechercher des facteurs de transcription dans la liste de motifs que nous avons trouvés en 5' et en 3' à l'aide de MEME (1KB) et qui sont conservés à 80 %. Nous avons comparé nos motifs avec les données du site ConSite (http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite). Évidemment, on ne s'attend pas à trouver des facteurs de transcription dans les séquences (1KB) en 5' et 3' parce que la plupart des facteurs de transcription chez les mammifères se retrouvent à beaucoup plus de 1KB de distance du site d'initiation de la traduction (Carey et Smale., 2000).

Suite à cette analyse, nous avons remarqué que plusieurs motifs trouvés avaient une similarité élevée (80%) avec plusieurs facteurs de transcription connus :

SP1	
	GGGGGTGGGA
	GGGCCTGGGT
	TCCCAGCCTT
	TCCCCCCCTC
MZF_1-4	TCCCCA
	TCCCCG
	TCCCCC
GATA-3	AGATAG
AP2alpha	GCCCCTGGG

Tableau X : Exemple de facteurs de transcription reconnus pour la région 5'.

Il est important de mentionner que la plupart des facteurs de transcription trouvés sont similaires à ceux que l'on retrouve chez l'humain, mais que certains d'entre eux

ont plus de ressemblance avec ceux de la souris ou du poulet. Une vue d'ensemble des récepteurs qui partagent des motifs de transcription (chez l'humain) reconnus est représentée à la figure 5.21 (upstream) et 5.22 (downstream).

Pour déterminer quel serait un nombre élevé de motifs conservés qui sont reconnus comme des facteurs de transcription nous avons fait un test qui consiste à utiliser les motifs que nous avons trouvés à l'aide de MEME et pour chacun d'eux mélanger au hasard l'ordre des lettres tout en conservant le même pourcentage de chacun des nucléotides pour chaque motif. Nous avons fait ce test à 5 reprises et on peut voir les résultats à la figure 5.23. Le fait que très peu de motifs conservés sont reconnus comme des facteurs de transcription (18 % en 5' et 30 % en 3') comparativement à ce que nous pouvions nous attendre à trouver au hasard (figure 5.23) nous suggère que ceux-ci ne sont pas très significatifs et ce résultat concorde avec nos attentes. Nous n'avons pas fait beaucoup de réplicats pour des raisons techniques, mais les résultats semblent néanmoins significatifs.

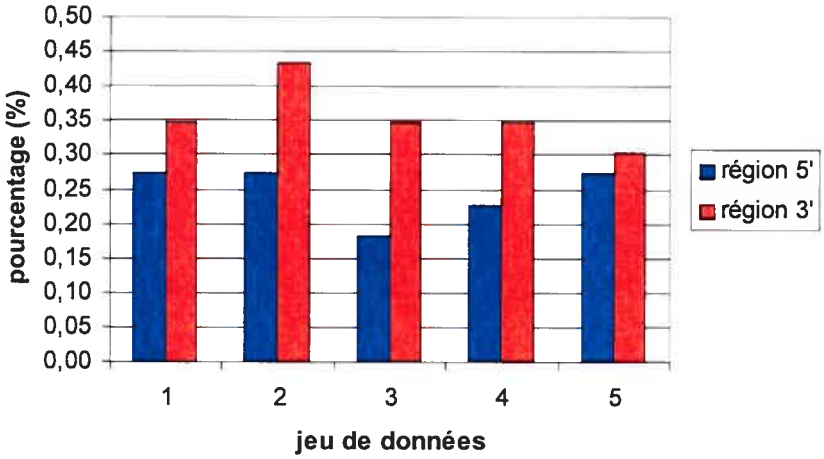

FIG. 5.23 : Randomisation des motifs trouvés pour les régions 5' et 3' (conservés à 80%) et comparaison (% de motifs reconnu comme facteurs de transcription) avec les facteurs de transcription du site web ConSite.

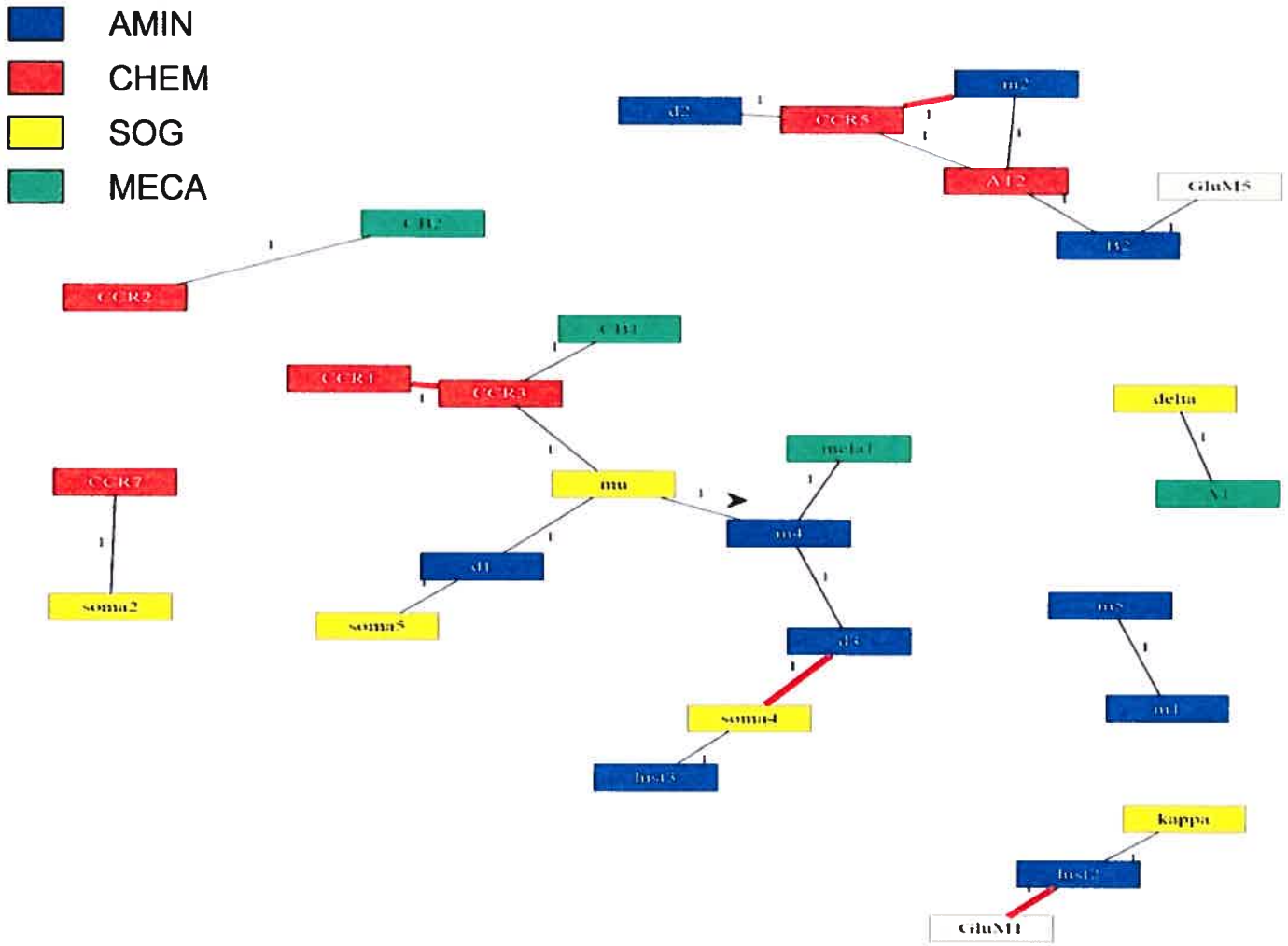

FIG. 5.21: Distribution des récepteurs qui partagent des motifs de transcription reconnus pour la région 5'. Les arêtes rouges représentent les motifs qui ne partagent pas de motifs de transcription, mais qui partagent tout de même des motifs conservés à 80 % trouvés avec MEME (1KB).

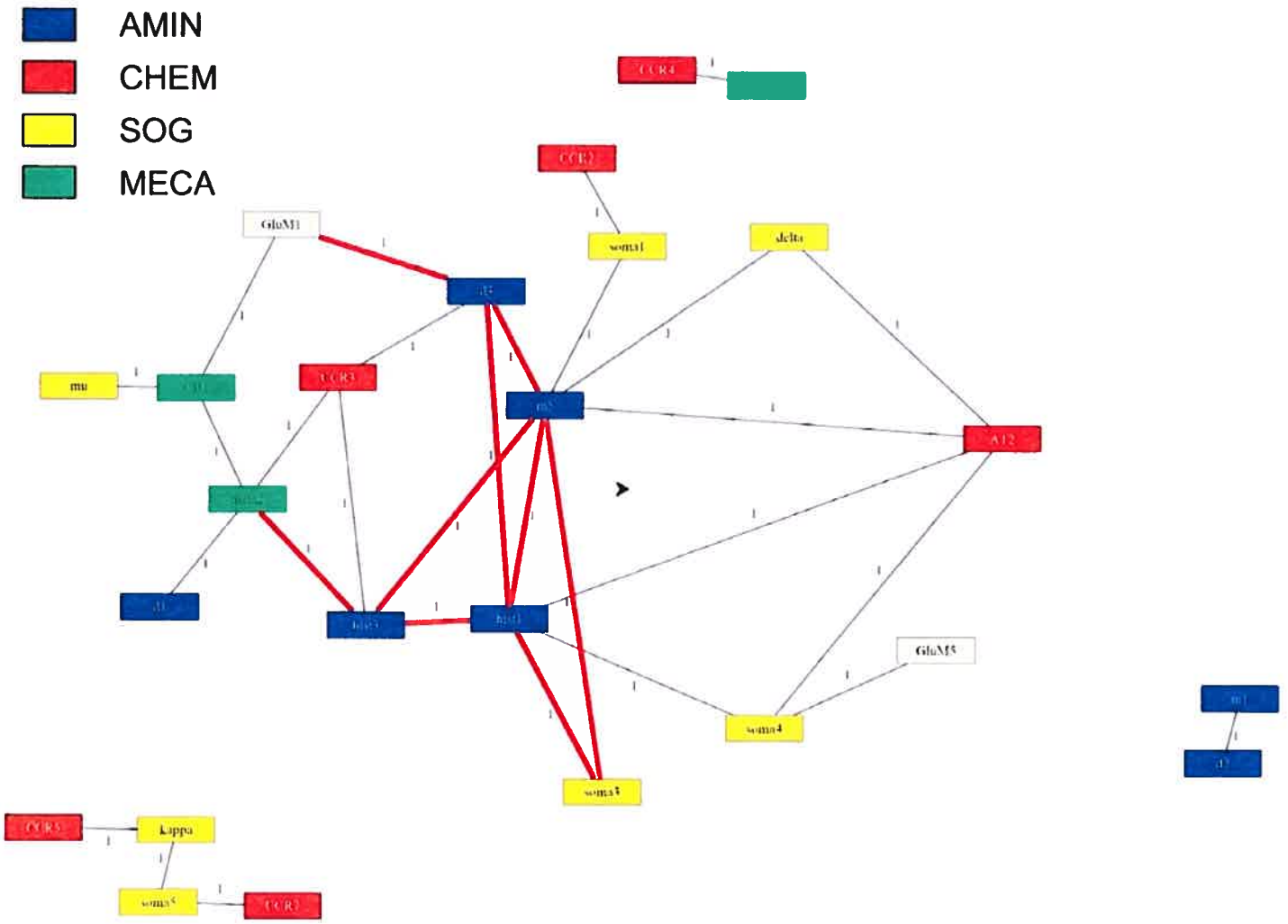

FIG. 5.22: Distribution des récepteurs qui partagent des motifs de transcription reconnus pour la région 3'. Les arêtes rouges représentent les motifs qui ne partagent pas de motifs de transcription, mais qui partagent tout de même des motifs conservés à 80 % trouvés avec MEME (1KB).

Par conséquent nous pouvons éliminer le fait que les résultats que nous avons obtenus contiennent des facteurs de transcription et cela nous indique aussi que les séquences semblent avoir été bien annotées. Le fait que nos données ne contiennent pas de facteurs de transcription nous permet de mieux comprendre nos résultats car il y a moins de bruit de fond.

5.7 Expression tissulaire des récepteurs

L'un des objectifs de ce mémoire est de déterminer des profils d'expression/traduction pour les récepteurs. Pour réaliser un tel objectif, il serait particulièrement intéressant d'avoir accès à des données de co-traduction pour tous les récepteurs. Mais de telles données sont plutôt rares et difficiles à trouver. C'est pourquoi nous nous sommes tournés vers l'expression tissulaire. En effet, si deux récepteurs ne peuvent pas être exprimés dans un même tissu, il n'y a aucune chance qu'ils puissent hétérodimériser.

Pour chaque récepteur, nous avons évalué leur niveau d'expression tissulaire avec la base de données SymAtlas et comparer ceux-ci entre eux. Les données utilisées sont celles du : Human GeneAtlas GNF1H, gcRMA et comporte environ 80 tissus différents.

Le résultat obtenu est présenté à la figure 5.24 et la légende représente le niveau de corrélation de l'expression tissulaire entre les deux récepteurs associés, 0,9 étant considéré comme une très bonne corrélation. Ensuite, nous avons fait un graphe de connexion avec les résultats corrélé à 0,9 pour le comparer avec les autres.

Finalement, nous avons aussi fait une simulation Monte Carlo (figure 5.25) pour déterminer si les « hubs » que l'on retrouve dans la figure 5.26 sont significatifs.

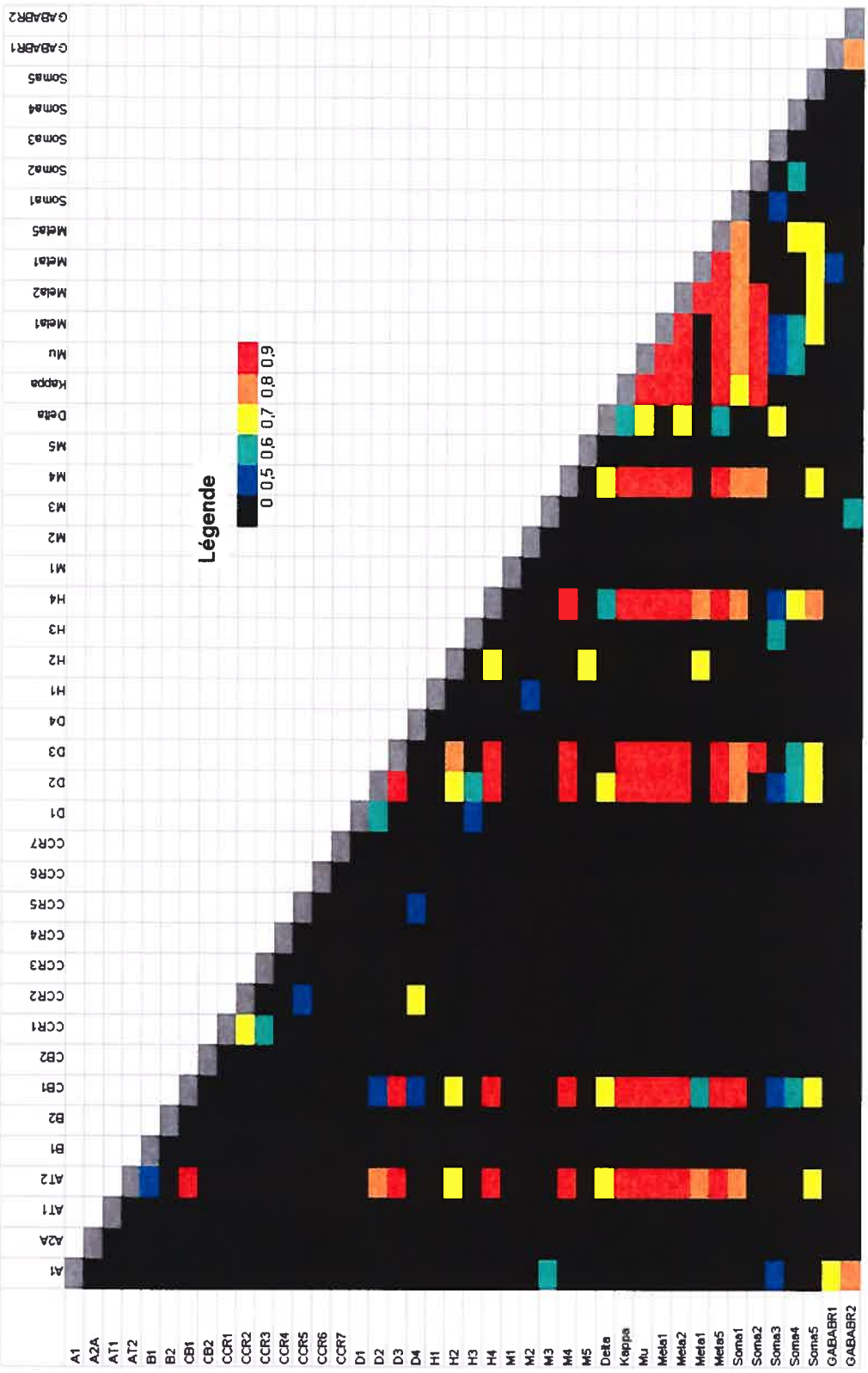

FIG. 5.24 : Niveau de corrélation de l'expression tissulaire des divers récepteurs.

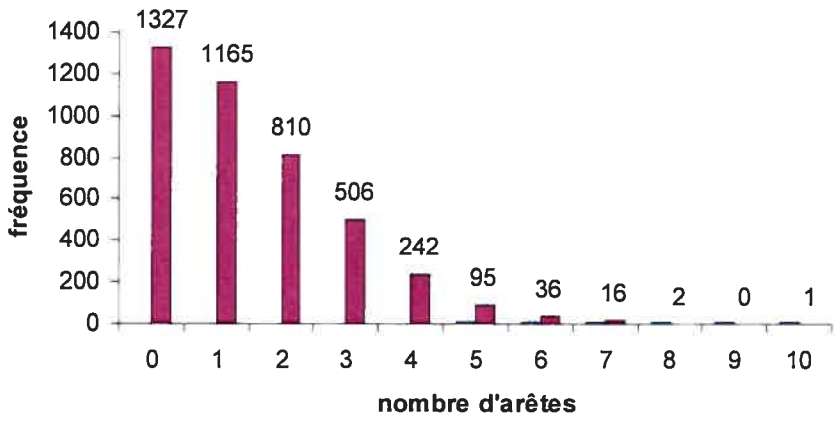

FIG. 5.25 : Simulation Monte Carlo pour les récepteurs corrélé à 0,9 (figure 5.24).

Lorsque l'on compare le graphe de la figure 5.26 avec la simulation Monte Carlo (figure 5.25), on constate que certains des « hubs » peuvent être dus au hasard. Par contre, ce qui est surprenant c'est qu'il y en ait autant. Celui qui se démarque le plus des autres est le récepteur Mela2 et lorsque l'on compare ce « hub » aux autres obtenus dans les autres graphes, c'est la première fois que ce récepteur a un tel comportement. Aucun des « hubs » qui sont significatifs dans les autres graphes ne semblent l'être dans la figure 5.26.

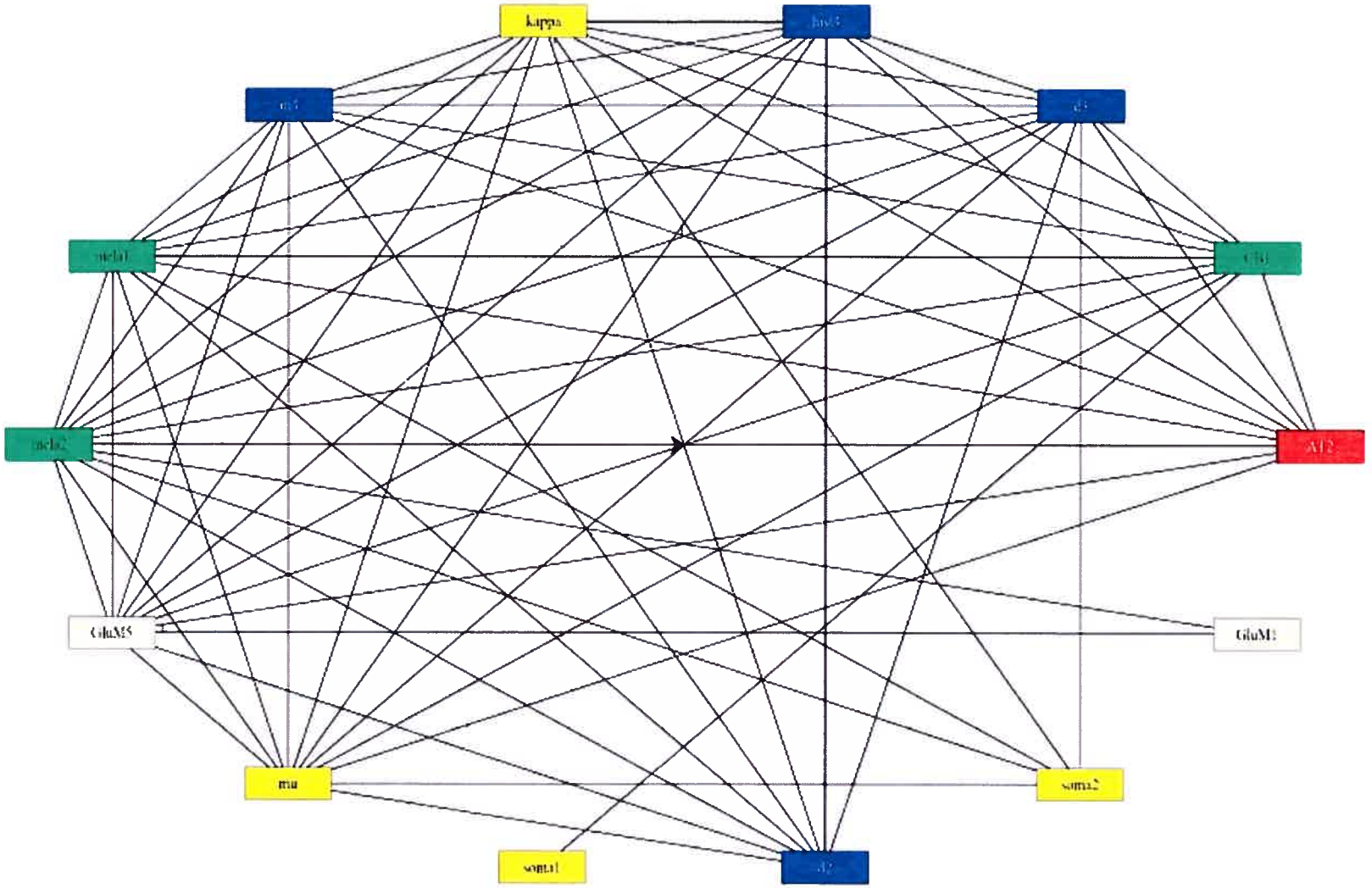

FIG. 5.26 : Distribution des récepteurs corrélés à 0.9 pour l'expression tissulaire (figure 5.24)

Chapitre 6

Discussion

Notre but principal étant de prédire des profils d'expression/traduction pour les récepteurs choisis, la première idée qui nous est venu à l'esprit était de chercher des motifs communs dans la région codante et surtout les régions en amont (5') et en aval (3') de cette dernière. Par contre, la simple recherche de motifs ne s'avère pas suffisante pour déterminer les profils des récepteurs. C'est pourquoi nous avons choisi de comparer des séquences orthologues entre elles, dans le but de réduire le bruit de fond lors de la recherche des motifs conservés.

6.1 Motifs et séquences

Au tout début de l'analyse, il faut s'assurer que les données recueillies ne contiennent pas de biais quelconque. C'est pourquoi nous avons effectué une analyse de la distribution de la taille des motifs en fonction de leur région (figure 5.1). Cette étape avait pour but principal de déterminer de manière globale la qualité de la recherche de motifs effectuée par le logiciel MEME. Contrairement aux autres logiciels, MEME permet de rechercher des motifs de plusieurs longueurs différentes en même temps. Cela entraîne certaines contraintes comme le fait que les motifs ne puissent pas se chevaucher et que la recherche débute toujours par la recherche des plus grands motifs possibles spécifiés. Dans notre cas, tous les motifs de longueur 15 étaient recherchés en premier. Donc, nous nous attendions à ce que la région codante, qui est mieux conservée que les deux autres, ait beaucoup plus de motifs conservés de longueur 15 et c'est ce que nous avons obtenu.

Suite à ce test, il nous semblait important de vérifier la composition en nucléotides des séquences de départ utilisées ainsi que de tous les motifs trouvés (figure 5.2). En effectuant cette deuxième vérification, nous pouvons vérifier s'il y a une saturation d'un nucléotide quelconque. Une saturation d'un ou plusieurs

nucléotides pourraient entraîner une augmentation non-négligeable de faux-positifs dans nos résultats. Par exemple, plusieurs séquences qui contiendraient un niveau de guanosine plus élevé que les autres entraînerait la détection de plusieurs motifs conservés qui sont considérés de faible complexité et par conséquent pas nécessairement important pour la régulation de la traduction. Par contre, l'inverse peut être vrai aussi, c'est-à-dire qu'un excès de GC peut être très significatif biologiquement parlant, car il y a trois liaisons hydrogènes et c'est donc plus stable. C'est dans ce but que cette vérification a été faite. On peut constater que pour l'ensemble des données, il ne semble pas y avoir de saturation, sauf peut-être pour les petits motifs et que par conséquent cela ne devrait pas affecter les autres résultats, c'est-à-dire que nous avons éliminé les liens dans les réseaux d'interactions qui étaient dû aux petits motifs (longueur 4 et 5), car ceux-ci comportaient les plus hauts niveaux de saturation.

6.2 Réseaux d'interactions

Pour représenter les récepteurs qui partagent des motifs similaires, nous avons décidé d'utiliser une représentation graphique, les réseaux d'interactions. Afin de pouvoir valider notre approche (figure 5.3), nous avons utilisé deux mécanismes de contrôle pour nos résultats. Le premier consistait en l'analyse de la région codante pour tous les récepteurs. Notre but, ici, était de retrouver le signal phylogénétique qu'il y a dans les séquences avec une recherche de motifs dans la région codante. Le résultat obtenu (figure 5.4) étant satisfaisant, nous avons décidé de poursuivre dans cette direction. Le second mécanisme de contrôle était l'utilisation des récepteurs GABA-B-R1 et GABA-B-R2 dans notre analyse. Ce sont deux récepteurs étant reconnus pour hétérodimériser ensemble et seulement ensemble, ils constituent un bon groupe contrôle. De plus, pour mieux représenter les arêtes qui sont significatives, nous avons déterminé certains seuils (nombre d'interactions sur un graphe) pour ne conserver qu'environ les cinquante meilleures arêtes. Ce qui rendait l'interprétation des graphes plus facile à faire et plus claire en même temps.

La première comparaison de réseaux (figures 5.4 et 5.5) avait comme objectif de déterminer quel logiciel retrouverait le mieux le signal phylogénétique des

séquences de départ. Comme nous pouvons le constater MEME semble avoir des meilleurs résultats que BioProspector. La principale cause de cette différence se cache derrière leur méthode de recherche de motifs. Comme nous l'avons mentionné plus tôt, MEME ne permet pas aux motifs trouvés de se chevaucher et cherche les motifs les plus longs en premier, ce qui permet une analyse plus étendue sur les séquences utilisées. BioProspector, lui, recherche des motifs d'une seule longueur à la fois. Par conséquent, lorsqu'il recommence une recherche pour une longueur différente il ne prend pas en ligne de compte les anciens motifs trouvés. Il peut prendre pratiquement le même motif, il suffit seulement de lui ajouter un nucléotide. Donc, les motifs trouvés avec BioProspector peuvent être redondants dans un certain sens même s'ils ne sont pas exactement identiques et la recherche de motifs peut parfois cibler qu'une partie de la séquence. Ainsi, MEME semble mieux adapté pour la situation présente.

Ensuite, nous avons continué la comparaison des réseaux pour les régions 5' et 3' pour chacun des logiciels. Un fait intéressant s'est produit avec les résultats obtenus avec MEME (figure 5.7 et 5.8). Les récepteurs CCR2 et Kappa, pour la région 5', possèdent beaucoup plus d'arêtes, 12 et 24 respectivement, que les autres récepteurs. Ils partagent des motifs conservés avec toutes les sous-classes (AMIN, CHEM, SOG et MECA) de la classe 1 des RCPG. De plus, on retrouve sensiblement le même phénomène avec la région 3' pour le récepteur Kappa. Comme il a été mentionné à la section 5.5, cela ne devrait pas être simplement dû au hasard. Les simulations Monte Carlo (figure 5.9 et 5.10) sont là pour le prouver. Si l'on veut vérifier l'hypothèse de la cause biologique, il faut tout d'abord éliminer celle d'un biais méthodologique. Une possibilité pour la provenance d'un biais méthodologique est la manière dont est effectuée la recherche des motifs conservés. Pour éliminer cette possibilité, nous avons appliqué le même procédé, mais avec les motifs trouvés par BioProspector (figure 5.11 et 5.12). Nous pouvons constater que le même phénomène se produit, mais de façon un peu moins évidente. Par contre, les deux réseaux n'ont aucune arête commune et les récepteurs qui agissent comme des « hubs » ne sont pas les mêmes. Cette preuve n'élimine pas la possibilité que le biais soit dû à la méthode de recherche des motifs, mais elle nous permet de continuer de

croire qu'il y a peut-être une raison biologique derrière tout cela. Donc, un autre test s'impose pour appuyer l'hypothèse de la cause biologique. Nous avons opté pour le calcul du niveau d'expression tissulaire des récepteurs (figure 5.24). Le graphe (figure 5.26) montre que rien n'est significatif mais Kappa semble se détacher des autres. La plupart des récepteurs qui sont liés à Kappa sont fortement corrélés à celui-ci en terme de niveau d'expression tissulaire en plus d'avoir le même comportement dans les figures 5.11 et 5.12. Par conséquent, cela semble indiquer que ce résultat n'est pas dû au hasard et que l'hypothèse de la cause biologique est renforcée.

Dans le même ordre d'idées, nous avons décidé d'analyser les régions 5' et 3', mais avec des séquences plus longues. Au lieu de prendre seulement 1 KB de chaque côté des gènes qui codent pour les RCPG, nous avons pris 3 KB. Nous ne retrouvons pas les mêmes résultats (figure 5.16 et 5.17) que lorsque nous avons analysé avec des séquences de 1KB, mais les résultats se ressemblent. Les réseaux ont certaines arêtes en commun et certains récepteurs partagent plus de liens que d'autres. Pour la région 5', il n'y a pas d'arêtes qui sont communes et pour la région 3', il y en a que 8. De plus, les arêtes communes semblent, pour quelque-unes, n'avoir aucun lien en particulier lorsqu'on les compare avec l'arbre phylogénétique (figure 3.4) (exemple : M5 - Soma4). Au tout début, nous pensions que le fait d'augmenter la longueur des séquences pour les régions régulatrices allait augmenter du même fait la quantité/qualité du signal que nous pouvions détecter, mais il semble que ce soit le contraire qui s'est produit.

Par la suite, nous avons tenté d'utiliser une recherche de motifs un peu différente, c'est-à-dire de rechercher des motifs à 2-blocs pour les trois régions. Une fois ces motifs détectés, nous avons comparé ces réseaux avec ceux de BioProspector-1KB (figure 5.5, 5.11 et 5.12). Mais avant de faire cette comparaison, nous avons comparé les réseaux obtenus entre eux pour les différentes longueurs de gap (5, 10 et 15). Il en est ressorti quelques ressemblances, mais en général c'était assez différent et il n'avait aucun récepteur qui se démarquait des autres. C'est pourquoi nous avons décidé de comparer seulement les réseaux avec des longueurs de gap 10 (figure 5.18, 5.19 et 5.20) avec ceux de BioProspector standards. Une fois

encore, il y a quelques ressemblances, des arêtes communes et des récepteurs agissant comme des « hubs », mais rien de surprenant. Ce qui nous laisse croire que nous retrouvons sensiblement le même signal même s'il y a des gaps obligatoires entre les motifs.

6.3 Support statistique

Les tests de chi^2 réalisés sur les données nous ont permis d'entrevoir une nouvelle piste de recherche. En effet, lorsque l'on analyse les résultats on remarque que la plupart des éléments corrélés qui sont reconnus pour hétérodimériser (Tableau VIII) se retrouvent dans la région codante et que ce nombre est significatif (Tableau IV). Alors il nous est venu à l'idée d'observer la position des récepteurs dans l'arbre phylogénétique des RCPG (figure 3.4). Il semble à première vue que les récepteurs qui peuvent hétérodimériser sont souvent des groupes frères. Par conséquent, il serait intéressant d'approfondir cette piste de recherche, mais dû à un manque de temps nous n'avons pas pu procéder.

Le dernier test que nous avons réalisé est celui de comparer les motifs obtenus avec des facteurs de transcription déjà connus (figure 5.21 et 5.22). Nous l'avons effectué seulement sur une petite partie de nos résultats, mais ici le but principal était seulement de voir si des facteurs de transcription se retrouvaient dans les séquences de 1KB que nous avons prises pour les différentes espèces. Il s'est avéré que très peu de motifs conservés peuvent s'apparenter à des facteurs de transcription et ce résultat n'est pas surprenant.

6.4 Perspective

Au commencement de cette recherche, nous ne savions pas ce qui serait le plus significatif à rechercher dans les séquences d'ADN qui allait nous permettre d'établir les profils d'expression/traduction des RCPG. Au fur et à mesure que le projet avançait, il nous semblait de plus en plus évident que la simple recherche de motifs conservés à travers les séquences ne nous permettrait pas de trouver un signal prédisant la dimérisation des RCPG. Une nouvelle idée fut apportée : pourquoi ne

pas analyser les structures secondaires des ARNm des RCPG ? Cette idée nous semblait meilleure que la simple recherche de motifs, car elle intervient à un niveau supérieur aux séquences d'ADN et est en étroite corrélation avec les fonctions des ARNm. Au lieu de se concentrer sur les motifs de régulation, nous nous attardons plutôt sur la structure secondaire des ARNm ainsi que leur niveau dans les cellules (Pesole et al., 2001; Mignone et al., 2002). De plus, certains outils de recherche et comparaison de structure secondaire commençaient à apparaître dans la communauté scientifique (Yang et al., 2004; Pavesi et al., 2004; Siepel et al., 2005).

Depuis plusieurs années, des chercheurs utilisent la méthode du minimum d'énergie libre (Mathews et al., 2006) pour prédire les structures secondaires. Il n'est pas très recommandé d'utiliser ce genre d'approche, car ces structures ne sont pas très fiables et, par conséquent cela entraînerait des résultats peu concluants. Il faut plutôt s'appuyer sur la prédiction de structures secondaires à l'aide de substitutions compensées (Siepel et al., 2005) et ensuite rechercher les motifs de structures secondaires qui sont conservés entre les récepteurs.

Par exemple, la figure 6.1 représente un motif de structure secondaire conservé entre le récepteur GABA-B-R1 et le récepteur GABA-B-R2 pour la région 5' UTR.

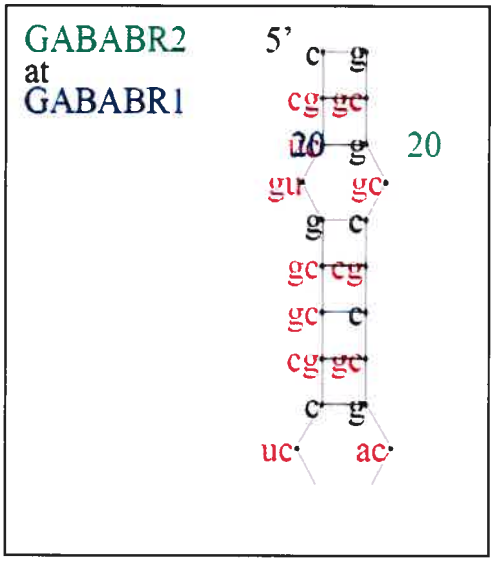

FIG. 6.1 : Motif de structure secondaire conservé entre les récepteurs GABA-B-R1 et GABA-B-R2. Les nucléotides noirs sont ceux conservés dans les 2 structures et ceux en rouges indiquent une substitution du premier par le deuxième.

Par conséquent, il serait possible de faire cette démarche pour tous les récepteurs et ensuite créer les réseaux d'interactions en fonction des motifs conservés comme nous l'avons fait. Ensuite, une analyse et comparaison avec les résultats que nous avons obtenus s'avèreraient une bonne continuation. Nous avons commencé de façon très brève cette approche et à première vue elle semble prometteuse.

Chapitre 7

Conclusion

Nous avons tenté de prédire des profils d'expression/traduction pour un groupe de RCPG, plus particulièrement ceux de la classe 1. Pour ce faire nous avons tenté d'établir des listes de motifs partagés entre les séquences de plusieurs espèces différentes. Nous avons analysé la région codante, la région 5' et la région 3' des récepteurs. Ceux-ci ont été classés en fonction de la région d'où ils provenaient et du pourcentage de conservation qu'ils partageaient avec les autres récepteurs. Ceci nous a amené à faire des réseaux de connexions pour pouvoir mieux distinguer les récepteurs qui partageaient des motifs et les comparer entre eux.

Le principal résultat est que la région codante semble contenir de l'information qui serait susceptible d'être impliquée dans la dimérisation des RCPG, tandis que les régions 5' et 3' sont beaucoup moins ordonnées. Il ne semble pas à première vue que les motifs régulateurs trouvés selon leur seule structure primaire et qui se trouveraient dans les régions étudiées soient la principale cause pour que certains récepteurs dimérisent ou non.

Les réseaux d'interactions, l'analyse de facteurs de transcription connus ainsi que de l'expression tissulaire des récepteurs n'ont pas apporté assez d'information nous permettant d'établir avec certitude des profils d'expression/traduction. Il est tout de même possible d'établir une liste d'hétérodimères possibles en prenant tous les réseaux dont les motifs sont conservés à 80 % pour la région codante :

- CCR1-CCR3
- CCR2-CCR3
- CCR2-CCR5*
- Delta-Mu*
- M1-M4
- M3-M5
- Soma1-Soma4
- Soma1-Soma5*
- Soma2-Soma3*
- Soma2-Soma5

Dans cette liste, il y en a déjà 4 qui sont reconnus pour hétérodimériser (*).

Nous n'avons pas trouvé de résultat qui nous permet de mieux comprendre le mécanisme de dimérisation entre les RCPG, mais nous avons tout de même découvert deux nouvelles pistes de recherche qui seraient intéressantes de suivre : la comparaison de motifs, au niveau des structures secondaires des RCPG pourrait être prometteuse et l'analyse, plus en profondeur, de la phylogénie des RCPG en comparant les différents groupes frères, leur vitesse d'évolution, leur taux de substitutions, etc. pourrait apporter une meilleure compréhension du phénomène.

Bibliographie

Assié, G., Rosenberg, D., Clauser, E. (2004). "Biochimie des hormones et leurs mécanismes d'action : récepteurs membranaires. " EMC-Endocrinologie 1 : 169-199.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol 2: 28-36.

Bailey, T. L. and C. Elkan (1995a). "The value of prior knowledge in discovering motifs with MEME." Proc Int Conf Intell Syst Mol Biol 3: 21-9.

Bailey, T. L. and C. Elkan (1995b). "Unsupervised learning of multiple motifs in biopolymers using expectation maximization." Machine Learning Journal 21: 51-83.

Bailey, T. L., N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res 34(Web Server issue): W369-73.

Barnes, P. J. (2006). "Receptor heterodimerization: a new level of cross-talk." J Clin Invest 116(5): 1210-2.

Bass, B. L. (2002). "RNA editing by adenosine deaminases that act on RNA." Annu Rev Biochem 71: 817-46.

Bi, C. and P. K. Rogan (2004). "Bipartite pattern discovery by entropy minimization-based multiple local alignment." Nucleic Acids Res 32(17): 4979-91.

Blanchette, M. and M. Tompa (2002). "Discovery of regulatory elements by a computational method for phylogenetic footprinting." Genome Res 12(5): 739-48.

Bouvier, M. (2001). "Oligomerization of G-protein-coupled transmitter receptors." Nat Rev Neurosci 2(4): 274-86.

Boyer R.S., Moore J.S., (1977). "A fast string searching algorithm." Communications of the ACM. 20:762-772.

Breitwieser, G. E. (2004). "G protein-coupled receptor oligomerization: implications for G protein activation and cell signaling." Circ Res 94(1): 17-27.

Bulyk, M. L. (2003). "Computational prediction of transcription-factor binding site locations." Genome Biol 5(1): 201.

Campbell, N. A. (1995). "Biologie", Éditions du renouveau pédagogique.

Carey, M., Smale, S.T., (2000). "Transcriptional regulation in Eucaryotes".

Chabre, M. and M. le Maire (2005). "Monomeric G-protein-coupled receptor as a functional unit." Biochemistry 44(27): 9395-403.

Che, D., S. Jensen, et al. (2005). "BEST: binding-site estimation suite of tools." Bioinformatics 21(12): 2909-11.

Fitch, W. M. (2000). "Homology a personal view on some of the problems." Trends Genet 16(5): 227-31.

Frazer, K. A., L. Elnitski, et al. (2003). "Cross-species sequence comparisons: a review of methods and available resources." Genome Res 13(1): 1-12.

Fredriksson, R., M. C. Lagerstrom, et al. (2003). "The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints." Mol Pharmacol 63(6): 1256-72.

Fredriksson, R. and H. B. Schioth (2005). "The repertoire of G-protein-coupled receptors in fully sequenced genomes." Mol Pharmacol 67(5): 1414-25.

Helmann,J.D. and Moran,C.P.,Jr (2002) RNA polymerase and sigma factors. In Sonenshein,A.L., Hoch,J.A., Losick,R. (eds) Bacillus subtilis and Its Closest Relatives, Chapter 21. ASM Press, Washington, D.C.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics 15(7-8): 563-77.

Holder, M. and P. O. Lewis (2003). "Phylogeny estimation: traditional and Bayesian approaches." Nat Rev Genet 4(4): 275-84.

Horn, F., J. Weare, et al. (1998). "GPCRDB: an information system for G protein-coupled receptors." Nucleic Acids Res 26(1): 275-9.

Horspool, R.N., (1980). "Practical fast searching in strings." Software - Practice & Experience, 10(6):501-506.

Hughes, J. D., P. W. Estep, et al. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae." J Mol Biol 296(5): 1205-14.

Hughes, T. A. (2006). "Regulation of gene expression by alternative untranslated regions." Trends Genet 22(3): 119-22.

Jensen-Seaman, M. I., T. S. Furey, et al. (2004a). "Comparative recombination rates in the rat, mouse, and human genomes." Genome Res 14(4): 528-38.

Jensen, S.T., X. S. Liu, et al (2004b). "Computational Discovery of Gene Regulatory Binding Motifs : A Bayesian Perspective." Statistical Science 19(1) : 188-204.

Jensen, S. T., L. Shen, et al. (2005). "Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes." Bioinformatics 21(20): 3832-9.

Johnson, W. et Jameson, J.L. (1998) Transcriptional Control of Gene Expression. In Principles of Molecular Medicine. Jameson, J.L. (ed.) Totowa: Humana Press Inc., pp. 25-41.

Joost, P. and A. Methner (2002). "Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands." Genome Biol 3(11): RESEARCH0063.

Keich, U. and P. A. Pevzner (2002). "Finding motifs in the twilight zone." Bioinformatics 18(10): 1374-81.

Kimura, M. (1977). "Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution." Nature 267(5608): 275-6.

Kolakowski, LF Jr (1994). "GCRDb: a G-protein-coupled receptor database". Receptors Channels 2:1-7.

Langelier, M-F., Trinh, V., et al. (2002). "Gros plan sur l'ARN polymérase II" Med Sci (Paris) 18(2): 210-16.

Lawrence, C. E. and A. A. Reilly (1990). "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences." Proteins 7(1): 41-51.

Lawrence, C. E., S. F. Altschul, et al. (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." Science 262(5131): 208-14.

Lenhard, B., A. Sandelin, et al. (2003). "Identification of conserved regulatory elements by comparative genome analysis." J Biol 2(2): 13.

Liu, J. S., Neuwald, A. F. & Lawrence, C. E. (1995). Bayesian models for multiple

local sequence alignment and Gibbs sampling strategies. J. Am. Stat. Assoc. 90, 1156-1170

Liu, X., D. L. Brutlag, et al. (2001). "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput: 127-38.

Liu, Y., X. S. Liu, et al. (2004). "Eukaryotic regulatory element conservation analysis and identification using comparative genomics." Genome Res 14(3): 451-8.

Lopez, P., Casane, D. et Philippe, H. (2002). "Phylogénie et évolution moléculaires." Med Sci (Paris) 18: 1146-54.

Marinissen, M. J. and J. S. Gutkind (2001). "G-protein-coupled receptors and signaling networks: emerging paradigms." Trends Pharmacol Sci 22(7): 368-76.

Mathews, D. H. and D. H. Turner (2006). "Prediction of RNA secondary structure by free energy minimization." Curr Opin Struct Biol 16(3): 270-8.

Mignone, F., C. Gissi, et al. (2002). "Untranslated regions of mRNAs." Genome Biol 3(3): REVIEWS0004.

Morris, (Jr) J.H., Pratt, V.R., (1970). "A linear pattern-matching algorithm." Technical Report 40, University of California, Berkeley.

Pesole, G., S. Liuni, et al. (2002). "UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002." Nucleic Acids Res 30(1): 335-40.

Pesole, G., F. Mignone, et al. (2001). "Structural and functional features of eukaryotic mRNA untranslated regions." Gene 276(1-2): 73-81.

Pierce, K. L., R. T. Premont, et al. (2002). "Seven-transmembrane receptors." Nat Rev Mol Cell Biol 3(9): 639-50.

Prinster, S. C., C. Hague, et al. (2005). "Heterodimerization of g protein-coupled receptors: specificity and functional significance." Pharmacol Rev 57(3): 289-98.

Rawn, J. D. (1990). "Traité de Biochimie", DeBoeck Université.

Saitou, N., and M. Nei. (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol. Biol. Evol. 4:406–425.

Sandelin, A., W. W. Wasserman, et al. (2004). "ConSite: web-based prediction of regulatory elements using cross-species comparison." Nucleic Acids Res 32(Web Server issue): W249-52.

Sauer, T., E. Shelest, et al. (2006). "Evaluating phylogenetic footprinting for human-rodent comparisons." Bioinformatics 22(4): 430-7.

Siepel, A., G. Bejerano, et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res 15(8): 1034-50.

Sinha, S. and M. Tompa (2000). "A statistical method for finding transcription factor binding sites." Proc Int Conf Intell Syst Mol Biol 8: 344-54.

Springer, M. S., W. J. Murphy, et al. (2003). "Placental mammal diversification and the Cretaceous-Tertiary boundary." Proc Natl Acad Sci U S A 100(3): 1056-61.

Stojanovic, N., L. Florea, et al. (1999). "Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions." Nucleic Acids Res 27(19): 3899-910.

Tagle, D. A., B. F. Koop, et al. (1988). "Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints." J Mol Biol 203(2): 439-55.

Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol 23(1): 137-44.

Trugnan, G., P. Fontanges, et al. (2004). "[FRAP, FLIP, FRET, BRET, FLIM, PRIM.new techniques for a colourful life]." Med Sci (Paris) 20(11): 1027-34.

Van Hellemont, R., P. Monsieurs, et al. (2005). "A novel approach to identifying regulatory motifs in distantly related genomes." Genome Biol 6(13): R113.

Venkatesh, B. and W. H. Yap (2005). "Comparative genomics using fugu: a tool for the identification of conserved vertebrate cis-regulatory elements." Bioessays 27(1): 100-7.

Voet, D., Voet, J. (2002). "Biochimie. " Seconde Édition, DeBoeck Université.

Wasserman, W. W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements." Nat Rev Genet 5(4): 276-87.

Woese, C. R., O. Kandler, et al. (1990). "Towards a natural system of organisms:

proposal for the domains Archaea, Bacteria, and Eucarya." Proc_Natl Acad Sci U S A 87(12): 4576-9.

Yang, Q. and M. Blanchette (2004). "StructMiner: a tool for alignment and detection of conserved secondary structure." Genome Inform 15(2): 102-11.

Zhang, Z. and M. Gerstein (2003). "Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements." J Biol 2(2): 11.

Internet :

http://fr.wikipedia.org/wiki/Chromosome

http://bioinfo.unice.fr/enseignements/www2005/documentation/outils/recherche_promoteurs/rappels_biologie.html

http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/alternative_splicing.html

http://biology.kenyon.edu/courses/biol63/ribo/ribo_elong.html

http://www.langara.bc.ca/biology/mario/Biol2315notes/biol2315chap12.html

http://www.decisioneering.com/monte-carlo-simulation.html

http://evolution.berkeley.edu/

Notes de cours BCM1501 (Université de Montréal)

http://en.wikipedia.org/wiki/G_protein-coupled_receptor

http://www.chez.com/rcpg/partie4.html

http://www.algorithmic-solutions.com/enleda.htm

http://web.indstate.edu/thcme/mwking/gene-regulation.html

http://www.ensembl.org/index.html

http://genome.ucsc.edu/

http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite

http://symatlas.gnf.org/SymAtlas/

http://www.biani.unige.ch/msg/teaching/evolution.htm