Chapitre 6

Illustrations

First cicada :
life is
cruel, cruel, cruel.
Issa.

Come see
real flowers
of this painful world.
Basho.

The Penguin Book of Zen Poetry.
Lucien Stryk, Takashi Ikemoto (Eds. et trads.). Penguin, 1981.

Ce chapitre est consacré à l'illustration sur des cas concrets des divers algorithmes qui ont été décrits dans le chapitre 5.

Notre but n'est pas de présenter des résultats nouveaux, mais plutôt de tenter de porter un regard critique, au travers d'exemples variés, sur la méthode d'analyse de chaînes macromoléculaires par la recherche de mots communs similaires que nous avons adoptée, sur les définitions de ressemblance que nous avons établies et, enfin, sur le choix que nous avons fait d'explorer de manière combinatoire l'espace de ces chaînes.

Ces exemples concernent les deux grandes familles de macromolécules qui nous intéressent dans le cadre de ce travail : les acides nucléiques (ADN/ARN) et les protéines. D'une manière générale, nous connaissions *a priori* ce qu'il fallait 'trouver'. Certains exemples sont 'faciles' dans la mesure où ce qu'il y avait (de connu) à repérer était relativement aisé à localiser. Nous les présentons car ils permettent d'illustrer, soit certains points de notre approche, soit certaines difficultés liées à la nature des objets que nous devons analyser ou aux opérations situées en amont ou en aval de la comparaison proprement dite (établissement d'un ensemble de données, présentation et interprétation des résultats). D'autres exemples ont été choisis parce qu'ils sont considérés comme 'difficiles', ce qui nous a permis de tester nos méthodes dans des conditions plus extrêmes et, par conséquent, de mieux en mesurer leurs limites actuelles. Ces limites sont parfois techniques mais elles peuvent également être biologiques. Nous discutons les deux aspects.

La présentation adoptée suit l'ordre des exemples et non celui des méthodes. Une première partie est ainsi consacrée à l'ADN et à l'ARN et examine divers cas rencontrés dans la littérature ou suggérés par des biologistes proches. Des exemples concernant les protéines sont traités dans une seconde partie. Chacun de ces ensembles d'exemples est soumis à un ou plusieurs des algorithmes que nous avons développés et, lorsque cela est possible, à ceux élaborés par d'autres. Chaque analyse est précédée d'une courte présentation de la famille de macromolécules examinée et du 'problème' biologique. Les résultats obtenus, ainsi que le champ d'application et les limitations des divers algorithmes employés, sont ensuite examinés. Cette analyse nous fournit parfois une occasion d'affiner la description sommaire donnée au début de ce travail des mécanismes biologiques mis en jeu et a pour objectif de mieux en faire sentir la complexité ainsi que la nature des problèmes algorithmiques que cette complexité soulève.

Par deux fois nous avons ressenti le besoin d'étendre de manière importante les algorithmes décrits dans le chapitre précédent. Nous avons reporté la présentation de ces extensions à ce chapitre-ci car nous voulions ainsi illustrer le fait que, même lorsque nous décidons de travailler avec des algorithmes combinatoires exacts et exhaustifs dans un certain sens, c'est la biologie qui doit préciser ce sens et influencer les définitions de ressemblance. Nous souhaitions également montrer la nécessité d'aller plus loin dans ce qui a été fait. L'une de ces extensions concerne exclusivement les chaînes nucléiques et est donc présentée en introduction aux exemples traitant de l'ADN et de l'ARN. L'autre extension pourrait s'appliquer aux deux types de chaînes macromoléculaires, mais n'a été utilisée pour l'instant que dans le cadre d'une étude sur des protéines. Elle est donc introduite au début de la section consacrée à ces molécules.

Observons enfin que l'approche est ici très naïve, ce n'est pas celle d'un biologiste. En particulier, les analyses que nous faisons servent les propos indiqués ci-dessus et uniquement ceux-ci. Du point de vue de la biologie, elles ne sont en aucun cas fines ou complètes.

6.1 Une observation à propos de la notation : un retour au vocabulaire de la biologie

Avant de présenter les illustrations de nos méthodes, il convient d'établir quelques points concernant le vocabulaire que nous allons employer dans ce chapitre. Jusqu'à présent, nous avons utilisé le terme de chaîne pour dénoter un certain codage des chaînes nucléotidiques ou polypeptidiques. La raison de ce choix était que ce codage pouvait représenter la structure, primaire aussi bien que tertiaire, des macromolécules. Nous ne pouvions ainsi nous servir du terme de séquence plus couramment employé en biologie puisque celui-ci est exclusivement associé à la structure primaire. Cependant dans les illustrations qui vont suivre, c'est toujours avec la structure primaire que nous allons travailler, aussi nous avons décidé de revenir à l'expression de séquence qui semble plus naturelle dans un contexte biologique que celle d'une chaîne. Pour la même raison, nous emploierons plus volontiers le terme de motif pour désigner un mot, surtout lorsque celui-ci sera associé à un signal biologique. Lorsque ce motif fera référence à la structure secondaire ou tertiaire plutôt qu'à la séquence, nous préciserons qu'il s'agit d'un motif structural. Enfin, lorsqu'aucune ambiguité n'est possible, nous utiliserons le terme de séquence pour désigner un motif considéré comme tel. Par exemple, nous parlerons de la séquence impliquée dans la régulation d'un gène même si cette séquence représente en fait un motif dans une séquence plus longue.

6.2 ADN et ARN

6.2.1 Introduction et nouvelle extension algorithmique

6.2.1.1 Le problème général

Lorsque nous travaillons avec des séquences d'ADN ou d'ARN, nous recherchons très fréquemment à identifier des signaux, par exemple de promotion de la transcription ou de la traduction. En général, nous essayons de localiser des mots qui apparaissent plus conservés dans un ensemble de séquences dont nous pouvons supposer qu'elles possèdent la même activité biologique. Repérer des signaux, les caractériser, n'est bien sûr pas la seule problématique associée à ce type de séquences, mais c'est une des plus importantes et c'est essentiellement celle qui nous concerne ici.

Le plus souvent, pour un lot de séquences à comparer, nous devons localiser un signal, ou un ensemble relativement restreint de signaux situés les uns à la suite des autres le long des séquences. Une approche naturelle du problème consiste alors, à partir d'une certaine définition de ressemblance entre mots, à rechercher tous les plus longs mots qui sont similaires suivant cette définition. Dans notre cas, cette démarche se ramène, ainsi que nous l'avons vu, à une recherche des plus longs modèles présents dans au moins un certain nombre de ces séquences.

Dans cette optique, les séquences d'ADN présentent un problème supplémentaire lié à leur structure en double hélice composée de deux brins anti-parallèles, et au fait que l'on ne connaît pas nécessairement lequel de ces deux brins il faut 'lire'. Les séquences présentes dans les banques ne correspondent pas toujours au brin 'actif', c'est-à-dire, porteur de l'information recherchée. De plus, dans certains mécanismes, les deux brins peuvent se trouver impliqués simultanément. Nous en verrons un exemple avec les promoteurs de la transcription.

Cela signifie que, dans nos algorithmes traitant de séquences d'ADN, nous devons pouvoir 'lire' les séquences que nous comparons dans les deux sens (c'est-à-dire, lire la séquence donnée au départ, et lire son complémentaire inversé qui correspond à la séquence du brin opposé de l'hélice). En termes de modèles et de leurs occurrences, ceci signifie qu'un mot va désormais être une occurrence d'un modèle soit parce que, comme avant, il est lui-même une occurrence dans le sens d'une des définitions de ressemblance que nous avons établies, soit parce que son complémentaire inversé l'est. Par exemple, si nos modèles sont des mots (notés m), et que nous avons $m = AAGTCT$, alors les mots $AAGTCT$ et $AGACTT$ sont tous deux des occurrences de m, le premier de manière exacte et le second en tant que complémentaire inversé. Nous esquissons donc dans la section suivante l'extension à apporter à nos algorithmes afin qu'ils puissent effectuer cette 'lecture dans les deux sens'.

6.2.1.2 Esquisse de la nouvelle extension algorithmique

Cette extension s'applique à ceux parmi nos algorithmes qui sont habituellement employés pour comparer des séquences nucléiques. Elle concerne donc H/LeMoivre (section 5.4.2.1), Combi (section 5.4.2.2) et LeCombi (section 5.4.2.3). Nous notons l'un quelconque de ces algorithmes par X.

Une façon simple de procéder serait bien sûr de faire tourner l'algorithme X deux fois, une fois sur les chaînes données au départ, une seconde fois sur ces mêmes chaînes inversées et complémentées, puis de vérifier le quorum sur l'union des deux ensembles en ne comptant qu'une fois la présence d'un modèle en même temps sur une chaîne et sur son complémentaire inversé.

Cette approche n'est cependant pas appropriée lorsque nous voulons obtenir les plus longs modèles et il est donc plus intéressant d'étendre l'algorithme de telle sorte que la double lecture

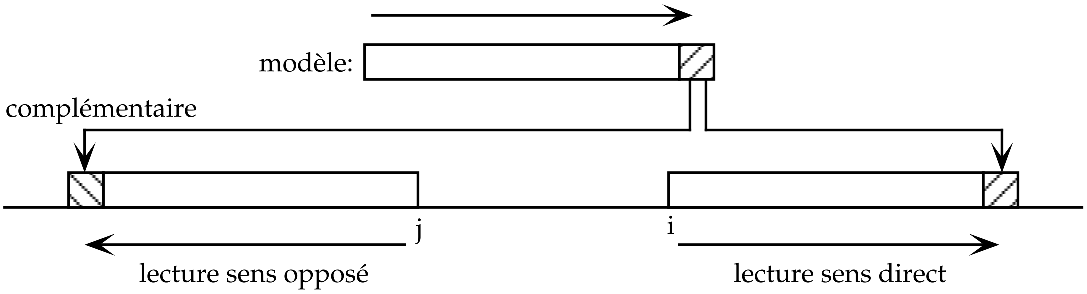

Figure 6.1: Illustration de la lecture dans les deux sens réalisée par l'algorithme ADN-X.

se fasse de manière simultanée lors de l'établissement de chaque modèle. Cette façon de faire présente un avantage supplémentaire : elle nous permet d'envisager une modification beaucoup plus importante de ces algorithmes permettant de localiser des structures symétriques miroir avec les mêmes définitions de ressemblance utilisées dans ce travail. Ces symétries peuvent représenter, soit des palindromes biologiques, soit faire partie de régions composées d'une suite contiguë de répétitions d'un même motif court, direct ou inversé (ces régions sont appelées des micro-satellites [Duret, 1995]).

L'extension des algorithmes de manière directe consiste alors à attacher une variable booléenne D ('Drapeau') à chaque occurrence d'un modèle M, et à procéder à la lecture du prochain symbole sur une séquence, soit dans le sens direct, soit dans le sens inverse (dans ce cas, ce sont les bases complémentaires de celles 'vues' qui sont effectivement 'lues'), suivant la valeur de D. Une illustration de cette double lecture est donnée dans la figure 6.1. La vérification du quorum se réalise ensuite de manière similaire aux versions originales. Nous notons ces extensions ADN-X. Leur complexité théorique est la même que celle des versions originales. Dans le pire des cas, elle est multipliée par une constante égale à 2.

6.2.2 Les promoteurs de *Bacillus subtilis*

6.2.2.1 Introduction

Nous avons vu dans le chapitre 2 que l'expression des gènes correspond à un transfert de l'information génétique de l'ADN aux ARN, puis des ARN aux protéines. Les molécules d'ARN sont synthétisées en utilisant la séquence nucléotidique de l'un des brins de l'ADN comme matrice dans une réaction de polymérisation réalisée par des enzymes appelées ARN polymérases. Le processus par lequel les molécules d'ARN sont initiées, allongées et terminées est appelé la transcription.

La première étape de cette transcription consiste en la fixation de l'ARN polymérase sur la molécule d'ADN. Cette fixation est réalisée au niveau de sites particuliers appelés promoteurs constitués de séquences spécifiques qui correspondent à des parties de la macromolécule d'ADN qui interagit avec l'enzyme. Un site en particulier connu pour être présent chez tous les organismes se situe à une distance de 5 à 10 bases en amont de la première base transcrite en ARN. Il est appelé la boîte de Pribnow pour les procaryotes ou la boîte de Hogness pour les eucaryotes (on parle aussi plus simplement de la séquence -10). Toutes les séquences identifiées à ce jour correspondant à cette boîte représentent des variations plus ou moins floues du motif

TATAAT ou *TATAAAT* respectivement. Cette boîte est supposée être une séquence permettant d'orienter l'ARN polymérase. C'est également la région où la double hélice de l'ADN s'ouvre pour former le complexe enzyme-ADN ouvert et autoriser ainsi le processus de transcription.

Une seconde séquence importante, présente dans la plupart des promoteurs de procaryotes, est appelée la séquence -35 parce qu'elle est localisée à environ 35 bases en amont de la première base copiée en ARN. La fonction précise de ce site n'est pas claire mais il est probable qu'une sous-unité de l'enzyme se fixe en premier lieu au niveau de la séquence -35 permettant ensuite à la région appropriée de la polymérase d'entrer en contact avec la boîte de Pribnow. Une fois fixée sur cette boîte, l'ARN polymérase se dissocie probablement du site de reconnaissance précédent.

Lorsque le promoteur est dépourvu de la séquence -35, un effecteur activateur (voir la section 2.1.1.3) est nécessaire à l'initiation de la transcription. Bien que le mécanisme d'action de cet effecteur ne soit pas encore tout à fait compris, il est probable que celui-ci se fixe en premier sur une autre séquence spécifique du promoteur puisque c'est cette interaction qui doit faciliter celle de la polymérase avec la boîte de Pribnow. La protéine CRP constitue un exemple d'une telle protéine effectrice et nous parlerons de son site de fixation un peu plus loin dans le chapitre. La protéine CRP est également impliquée dans la régulation de nombreux gènes intervenant dans le métabolisme des sucres, en l'absence ou présence de la séquence -35, en tant qu'activateur ou en tant que répresseur.

Outre les séquences situées à -10 et -35 du début de la transcription et celles où viennent parfois se fixer des effecteurs, il est possible que d'autres caractéristiques de l'ADN proches de la région devant être transcrite affectent également cette transcription. Galas *et al.* discutent ainsi [Galas *et al.*, 1985] l'existence d'une séquence qui interviendrait dans la promotion du processus de transcription chez *Escherichia coli* et qui se localiserait à -44, c'est-à-dire environ 9 bases en amont de la séquence -35.

Dans l'exemple discuté ici, nous avons d'abord voulu tester si nos algorithmes étaient capables de repérer la présence des séquences 'classiques' (celles localisées à -10 et -35) des promoteurs de *Bacillus subtilis* qui sont, en général, bien conservées. Nous avons ensuite voulu essayer de déterminer si d'autres séquences ne se trouvaient pas également anormalement conservées en amont ou en aval de ces sites et pourraient ainsi être impliquées dans la promotion de la transcription.

L'ensemble avec lequel nous avons travaillé est composé d'un lot de 142 séquences de longueur moyenne 120 nucléotides correspondant à 100 bases en amont de la première base transcrite et de 20 bases en aval. Ces séquences ont été répertoriées par Helmann [Helmann, 1995] et sont disponibles par ftp (URL http://www.bio.cornell.edu/microbio/helmann/helmann.htlm). Helmann présente en fait 246 séquences dont 94 sont hypothétiques ('putative promoters'). Nous avons choisi de travailler uniquement avec les 142 dont la fonction de promotion a pu être validée expérimentalement. En réalité, nous avons réduit cet ensemble aux 131 séquences du lot ayant une longueur supérieure à 100 bases (celles éliminées ne comportent en moyenne qu'une cinquantaine de bases et ne présentent que les sites à -10 et -35). Enfin, nous n'avons conservé sur chaque séquence que les 100 bases en amont de la première base transcrite.

Sur ces séquences, Helmann a réalisé un travail de comptage des divers nucléotides présents à chaque position des séquences après alignement de celles-ci sur la première base transcrite. Cela lui a permis de mettre en évidence :

- les sites -10 et -35 (très exactement -12 et -35 et dont les consensus sont respectivement *TAtAAT* et *TTGaca* — les bases en majuscules sont présentes dans plus de 70% des promoteurs);

- la présence d'un dinucléotide, TG, assez conservé à -15;
- l'existence de plusieurs régions riches en A en amont du site -35 (situées autour des positions -43, -54 et -65) ainsi que des régions riches en T localisées entre celles riches en A.

Sa méthode d'analyse ne lui donne toutefois pas les moyens de vérifier si une structure plus complexe existe à cet endroit.

Nous résumons dans la prochaine section les premiers résultats obtenus.

6.2.2.2 Résultats

Notre première approche lorsque nous avons voulu tester nos algorithmes sur ce jeu d'essai a consisté tout naturellement à rechercher les modèles présents dans au moins un certain nombre de séquences. À l'époque nous développions Combi (section 5.4.2.2); c'est donc avec des modèles définis sur une couverture combinatoire pondérée que nous avons initialement travaillé. Nous avions en fait établi un certain nombre de couvertures pondérées dont un exemple est la couverture suivante :

$$CCP = \left\{ \begin{array}{l} (\{X\}, \infty)\ \forall X \in \Sigma = \{A, C, G, T\}, \\ (\{\Sigma\}, 0), \\ (S, 1)\ pour\ tous\ les\ (10)\ autres\ ensembles \end{array} \right\}.$$

De plus, nous avions travaillé avec la variante de Combi décrite dans la section 5.4.2.2.6.1, ce qui veut dire que la pondération sur les ensembles était limitée par la pondération sur le joker lors de la première étape de l'algorithme, ceci afin d'éviter d'obtenir des modèles trop dégénérés. Cette pondération initiale sur le joker était égale à 4, et donc le nombre total d'ensembles 'spéciaux' permis dans les modèles (les non singletons) était au plus de 4.

Nous avions alors recherché les modèles les plus longs présents dans au moins 70% des séquences d'un sous-ensemble des 142 composé de 10, 50, 100 et 140 séquences [Sagot and Viari,1996]. Ainsi que nous l'avons indiqué dans l'article et dans la section 5.4.2.2.7, le nombre de modèles trouvés lorsque $k = k_{max}$ varie considérablement suivant le nombre de séquences faisant partie du lot test. Nous avons argumenté que cette instabilité des résultats obtenus par rapport aux paramètres choisis, en particulier l'explosion du nombre de modèles dans certaines circonstances, n'est pas nécessairement une mauvaise chose et peut apporter une information importante. Cela est en particulier le cas lorsque, comme ici, nous bénéficions du fait que les séquences sont cadrées, c'est-à-dire, qu'elles peuvent être alignées de manière 'naturelle' sur la première base à transcrire.

Il est en effet possible, dans ce cas, d'établir un graphique indiquant la position où les modèles sont trouvés, ce qui nous permet, surtout lorsque le nombre de ces modèles est grand, d'avoir une idée des régions qui peuvent éventuellement être associées à un signal. Le graphique que avons obtenu pour un lot de 140 séquences, avec la couverture indiquée ci-dessus et q égal à 70%, est celui donné dans la figure 6.2. Cette courbe présente quatre pics nets, deux faisant référence au même signal situé à -10 (boîte de Pribnow) et -14 (dinucléotide conservé TG), un second pic à -33 et enfin un troisième autour de -46. Sachant d'après Helmann que la région autour de la position -45 est particulièrement riche en A et en T (plus que le reste de la séquence qui présente déjà un biais en composition en faveur de ces deux bases), la question que nous nous sommes posée alors était : est-ce que ce pic est dû au fort biais en composition des séquences dans cette région ou bien y a-t-il une structure, c'est-à-dire un vrai motif, à cet endroit?

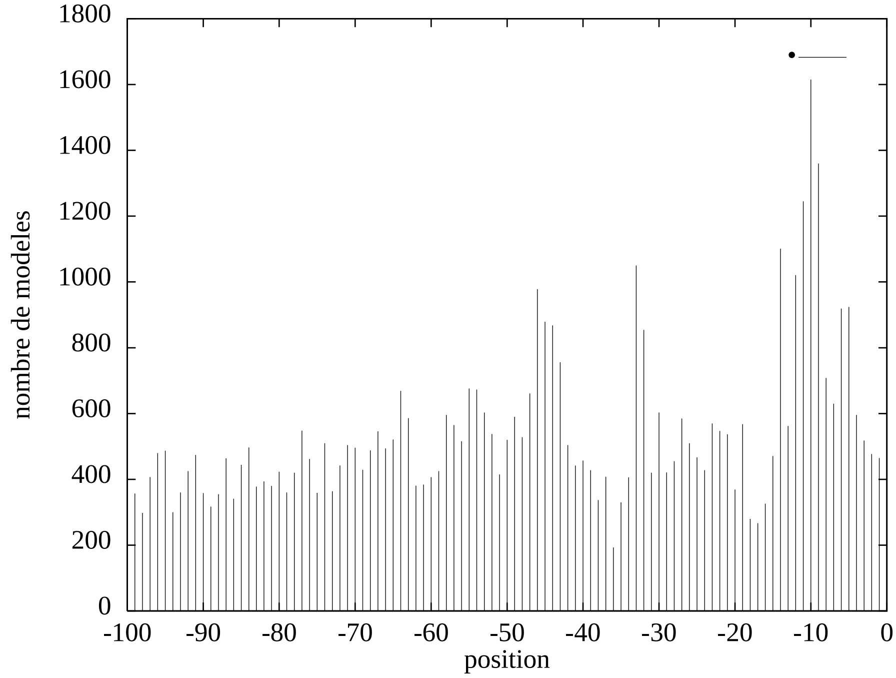

Figure 6.2: Graphe du nombre des modèles (plus longs trouvés avec la couverture CCP donnée dans le texte, $k = k_{max}$, $q = 70\%$ et pour 140 séquences). La position 0 en abscisse indique l'endroit où débute la transcription.

Pour essayer d'y répondre, nous avons par la suite fait une recherche plus exhaustive des modèles présents dans ces séquences et pour cela nous avons adopté la stratégie suivante :

- utiliser non seulement Combi mais aussi HMoivre (section 5.4.2.1.3) et LeCombi (section 5.4.2.2) car une de ces approches pouvait être plus appropriée que les autres pour repérer les signaux présents dans ces régions;
- découper les séquences en trois fragments de longueur 15, centrés sur les positions -10, -33 et -46 respectivement, constituant ainsi trois nouveaux lots de séquences;
- rechercher dans chacun des trois lots tous les modèles de longueur supérieure à une certaine valeur (entre 4 et 6 suivant la méthode);
- enfin, essayer de déterminer dans quelle mesure les modèles trouvés sont dus au hasard, cela en effectuant la même recherche sur des séquences aléatoires de même longueur et de même composition en mono et en dinucléotides que les fragments originaux (cent de ces séquences aléatoires sont produites pour chaque fragment de l'ensemble de départ).

Cette dernière évaluation a été réalisée en comparant des tables de contingence qui permettent de déterminer si une première variable (ici la présence d'un modèle dans un ensemble de séquences) dépend d'une seconde variable (ici le fait que ces séquences sont de 'vraies' séquences où un certain nombre de signaux se trouvent présents). Si cela est le cas, nous pouvons supposer que la présence du modèle dans les séquences n'est pas dû au hasard mais peut être liée au signal recherché [Clarke and Cooke, 1992]. Pour chaque modèle M, nous disposons alors de deux tables de 2×2 cellules chaque. La première, notée $\mathcal{O}$ (observation), indique le nombre de séquences où le modèle M est présent au moins une fois (resp. absent) dans l'ensemble des 'vraies' séquences, et le nombre de fois où il est présent (resp. absent) dans l'ensemble des 'fausses' séquences, c'est-à-dire, des séquences aléatoires. La seconde table, notée $\mathcal{A}$ (attendu), correspond à l'hypothèse nulle, c'est-à-dire à l'hypothèse d'indépendance des deux variables indiquées ci-dessus. Les cellules de cette table valent $\frac{n_i.n_j}{N}$ où n_i indique le nombre total de séquences où M est présent (absent), n_j le nombre de séquences 'vraies' ('fausses') et N le nombre total de séquences (voir figure 6.3). Ces deux tables sont ensuite comparées en effectuant un test du χ^2 (à 1 degré de liberté) :

$$\chi^2 = \sum_{\substack{toutes\ les \\ cellules\ i,j}} \frac{(\mathcal{O}_{ij} - \mathcal{A}_{ij})^2}{\mathcal{A}_{ij}}.$$

Les résultats obtenus sont approximativement les mêmes que l'on travaille avec HMoivre ou avec Combi. Dans les deux cas, une interprétation vraiment claire n'est possible que lorsque relativement peu de souplesse est autorisée. Le nombre d'erreurs permis dans HMoivre (substitutions uniquement car il n'y a, en principe, pas de trous à l'intérieur des sites de promotion) est donc égal à 0 ou 1 et les couvertures combinatoires adoptées dans le cas de Combi correspondent à la variante de la section 5.4.2.2.6.2 (pondération par cardinalité) et sont très faiblement dégénérées (au maximum 2 sous-ensembles autres que les singletons sont permis dans un modèle). Ces résultats sont les suivants :

- les modèles trouvés dans chaque cas sont de longueur relativement petite (6 ou 7 bases) même pour des quorums assez bas (5-20%);
- les modèles trouvés dans les régions situées autour des pics à -10 et -35 sont significatifs par rapport à l'hypothèse nulle (une indication de quelques-uns de leurs noms, nombre

Table $\mathcal{O}$:

	Présents	Absents	
Vraies	P_V	A_V	n_V
Fausses	P_F	A_F	n_F
	n_P	n_A	

$$n_V + n_F = n_P + n_A = N$$

Table $\mathcal{A}$:

	Présents	Absents	
Vraies	$\frac{n_V.n_P}{N}$	$\frac{n_V.n_A}{N}$	n_V
Fausses	$\frac{n_F.n_P}{N}$	$\frac{n_F.n_A}{N}$	n_F
	n_P	n_A	

Figure 6.3: Tables de contingence pour évaluer les modèles de promoteurs obtenus avec nos algorithmes.

de 'vraies' et de 'fausses' séquences où ils sont présents et probabilité par rapport à l'hypothèse nulle (i.e. probabilité d'obtenir une valeur de χ^2 supérieure à celle observée) pour HMoivre avec 0 erreur, q = 5% et un mélange en dinucléotides est donnée dans le tableau 6.1);

- les modèles trouvés dans la région du pic à -45 sont présents dans un nombre très faible de séquences (environ 5%) et ne sont pas significatifs lorsqu'évalués par rapport à des séquences mélangées en préservant la composition en dinucléotides.

Ces résultats semblent donc indiquer que, du moins en ce qui concerne les petits motifs, la région autour de -45 ne présente aucune structure qui ne pourrait être expliquée par le simple biais de la composition des séquences à cet endroit.

Les observations réalisées avec LeCombi sont, par contre, très différentes, surtout lorsqu'un nombre relativement grand de jokers est autorisé. La couverture utilisée est maintenant la suivante :

$$CCPR = \left\{ \begin{array}{l} (\{X\}, \infty) \, \forall X \in \Sigma = \{A, C, G, T\}, \\ (\{A, C, G, T\}, J), \\ (S, 0) \textit{ pour tous les autres ensembles} \end{array} \right\}$$

avec J entre 4 et 20. Les modèles trouvés, même lorsqu'aucune erreur n'est autorisée, sont plus longs (11 pour 4 jokers et $q = 20\%$, 27 pour 20 jokers et $q = 20\%$) et une nette structure apparaît alors avant le site -35. La démarche que nous avons adoptée pour essayer de la caractériser a consisté à :

- établir deux lots de 131 séquences, le premier (lot A) est l'original (longueur des séquences égale à 100 bases) et présente les 3 pics, le second (lot B) représente les mêmes séquences mais chacune ne comporte plus que 30 bases entre -35 et -65 (ce second lot ne présente donc que le pic à -45 et correspond à la région riche en A et T);

Modèle (site -10)	Prés. séq. vraies	Prés. séq. fausses	Probabilité χ^2
$TATAAT$	23	3,6	10^{-15}
$CTATAAT$	9	0,8	10^{-9}
$ATAAT$	50	21	10^{-9}
$TATAATA$	17	3,3	10^{-9}
$TATAATG$	10	1,6	10^{-7}
Modèle (site -35)	Prés. séq. vraies	Prés. séq. fausses	Probabilité χ^2
$TTGACA$	21	1,6	10^{-23}
$TTGAC$	39	10,2	10^{-15}
$GACA$	31	6,9	10^{-14}
$GTTGAC$	15	2,9	10^{-10}
$ATTGAC$	13	1,8	10^{-10}

Tableau 6.1: Quelques modèles trouvés relatifs aux sites de promoteurs -10 et -35. L'algorithme utilisé ici est HMoivre avec 0 erreur et un quorum de 5%. Les modèles devaient avoir une longueur supérieure ou égale à 4. Le nombre d'occurrences fausses est une moyenne sur cent mélanges.

- rechercher ensuite les plus longs modèles présents dans au moins 20% des séquences du lot A ou B avec 0 erreur et 4, 8 ou 20 jokers;
- effectuer un classement des modèles trouvés basé sur l'écart observé entre le nombre d'occurrences sur l'ensemble des 'vraies' séquences et celui sur l'ensemble des séquences aléatoires qui préservent la composition en dinucléotides, cet écart étant mesuré par un χ^2;
- enfin, dans certains cas, 'lire' les séquences dans les deux sens : dans le sens direct et dans le sens complémentaire inversé.

La justification pour ce dernier point dans le cas des promoteurs est la suivante. Nous avons vu auparavant que le site -10 est la séquence d'orientation de la polymérase, et il est probable que l'orientation du site -35 soit également importante. Il est très possible cependant que cela ne soit plus le cas pour un signal qui serait présent autour de la position à -45, et que la protéine qui viendrait éventuellement s'y fixer pourrait tout aussi bien se mettre en contact avec le motif correspondant sur l'autre brin de la double hélice d'ADN. C'est ce qui semble être le cas pour certains promoteurs de *coli* ainsi que le suggère Galas [Galas *et al.*, 1985]. Il est donc nécessaire de 'lire' la séquence dans les deux sens, sachant que cela correspond à une lecture simultanée des deux brins. Notons que, dans le cas de cet exemple, cette double lecture ne change que peu les résultats pour les petits modèles observés dans la région -45 qui demeurent non significatifs. Ceux repérés par LeCombi par contre sont plus sensiblement affectés, surtout pour les motifs moins longs obtenus en autorisant 4 jokers : certains représentent des occurrences du modèle sur le brin opposé de l'ADN.

Nous avons choisi de présenter sous forme de graphiques les résultats de quelques modèles sélectionnés. Considérons un de ces modèles : $TTGN_{20}TATA$. Nous pouvons nous poser plusieurs questions à son propos :

- lorsqu'un symbole (un singleton) apparaît dans ce modèle, par exemple le premier T, pouvons-nous affirmer que ce symbole est réellement important, c'est-à-dire, qu'il ne peut être substitué par un autre? Pour essayer de le déterminer, l'idée est ici de remplacer un à un, de manière indépendante, chacun des symboles non joker du modèle par les trois autres symboles possibles et de compter le nombre d'occurrences du nouveau modèle ainsi constitué. Cette opération produit les graphiques notés 'a'.

- lorsqu'un modèle présente un joker à une certaine position, en quoi cette position est-elle vraiment non spécifique? Cette fois l'idée est de calculer la matrice des fréquences observées de chaque symbole en chaque position des occurrences du modèle lorsqu'aucune substitution n'est autorisée. D'une manière plus générale, nous pouvons également calculer cette matrice pour les occurrences identifiées lorsque $e > 0$ substitutions sont permises. Dans les graphiques notés 'b' donnés ci-après, nous avons choisi $e = 1$.

- enfin, lorsque plusieurs jokers consécutifs apparaissent dans un modèle (par exemple N_{20} dans $TTGN_{20}TATA$), est-ce que ce nombre de jokers contigus est précisément fixé, ou bien peut-il varier dans un certain intervalle? En effet, une suite de jokers contigus peut être considérée comme un 'espace' (en anglais 'spacer') entre deux motifs plus stricts, et nous pouvons légitimement nous interroger sur l'optimalité de la longueur de cet 'espace'. L'idée afin de nous assurer de cette optimalité est de rechercher le modèle en faisant varier la longueur de l''espace'. La quantitée mesurée pour chaque valeur de cette longueur est alors le χ^2 entre le nombre d'occurrences observées, d'une part sur les 'vraies' séquences, d'autre part sur les séquences aléatoires. Ceci définit les graphiques notés 'c' donnés ci-après.

En résumé, nous avons établi pour chaque modèle sélectionné trois types de graphiques :

Graphique a : le nombre d'occurrences observées (lot A ou B) lorsque une (une seulement) des bases (pas le joker) est mutée en une autre base à tour de rôle;

Graphique b : fréquences des bases observées en chaque position dans les occurrences du modèle (lot A ou B) avec cette fois 1 substitution autorisée;

Graphiques c : χ^2 entre le nombre d'occurrences observées (lot A ou B), d'une part sur les séquences 'vraies' et d'autre part sur les séquences aléatoires, lorsque la longueur l des suites contiguës d'au moins 2 jokers (une suite seulement à chaque fois) est variée avec un pas de 1 entre deux valeurs égales à $min\{l - 10, 1\}$ et $l + 10$.

Le tableau 6.2 résume les noms, le nombre d'occurrences et le χ^2 des modèles de longueur maximale parmi les plus intéressants identifiés par LeCombi. Les graphiques (de types 'a', 'b' ou 'c') associés à chacun d'eux sont donnés dans les pages suivantes.

Par exemple, la figure 6.4 représente les graphiques relatifs au modèle $TTGN_{20}TATA$, qui est un des modèles les plus significatifs trouvés dans le lot A avec $q = 20\%$ et en autorisant 20 jokers mais pas d'erreurs (5798 secondes sur une Sparc Station 20). Ce modèle correspond aux trois premières bases du site -35 suivies d'une suite de jokers (notés N) puis des quatre premières bases du site -10. Il est exactement présent dans 28 des 131 séquences, et nous pouvons observer que les bases associées aux jokers sont plus fréquemment des A ou des T (graphique b) sauf aux positions 5 et 23 du modèle où les bases C et G respectivement sont légèrement

Modèle	Occ. sur vraies	Occ. sur fausses	χ^2
$TTGN_{20}TATA$	28	2.69	238.14
$TNTN_3AAAAN_{16}A$	33	8.46	71.18
$TTTTN_2ANAAN_5T$	16	1.24	175,69
$ANAN_3TNTN_3AAAA$	10	0.92	89,62
$AAANAN_7AAAA$	10	1.48	49,05
$TNTTTN_2AANA$	30	6.83	78,60
$TNTTN_2AAANA$	30	6.83	78,60
$TNTTN_3AAAA$	34	8.65	74,29

Tableau 6.2: Quelques modèles trouvés avec LeCombi pour $k = k_{max}$, $q = 20\%$ avec 20, 8 et 4 jokers autorisés.

plus souvent rencontrées. Ce G correspond en fait à celui du dinucléotide conservé TG à -14 (voir la figure 6.2). D'autre part, il est intéressant de noter que le nombre de jokers centraux (N_{20} dans le modèle) est rigoureusement conservé à 20. Cet écart de 20 bases représente en réalité un écart de 17 entre la fin du premier site et le début du second si l'on considère que le site -35 possède 6 bases, les trois dernières n'apparaissant pas dans ce modèle. Cet écart est ainsi pratiquement constant pour les promoteurs dont le site à -35 débute par TTG et celui à -10 par $TATA$. Helmann montre que cela demeure vrai en général et que cette distance de 17 bases entre les deux sites correspond à une valeur optimale chez *Bacillus subtilis* de la même manière que chez *Escherichia coli*.

Les figures suivantes (figures 6.5 à 6.11) correspondent toutes à des modèles repérés dans la région autour du pic observé à -46 (le lot utilisé est le A pour la figure 6.5 et le B dans tous les autres cas, la lecture est directe pour les quatre premières figures et dans les deux sens pour les trois dernières), pour un nombre de jokers égal à 20 (figure 6.5), 8 (figures 6.6, 6.7 et 6.8) et 4 (figures 6.9, 6.10 et 6.11) et toujours avec $q = 20\%$ et $e = 0$ (les temps d'exécution ont été de 1038 secondes pour 8 jokers et de 129 secondes pour 4). Ces figures permettent de dégager la présence d'une nette structure dans la région à -45, du moins en ce qui concerne une partie des promoteurs de *subtilis*. Cette structure peut être caractérisée par :

- un motif central M_C composé d'une suite de 4 à 5 T, suivie par 2 à 3 N puis par une suite de 4 à 5 A, c'est-à-dire $M_C = T_{4-5}N_{2-3}A_{4-5}$;
- deux motifs latéraux, chacun d'eux englobant le motif central mais apparemment ne pouvant apparaître tous deux simultanément :
  - $A_{4-5}N_3M_C$;
  - $M_CN_{3-6}T_{2-3}$;
- un motif composé d'une suite de A, suivie d'une suite de jokers puis de la seconde moitié du motif central : $A_{4-5}N_{6-8}A_{4-5}$.

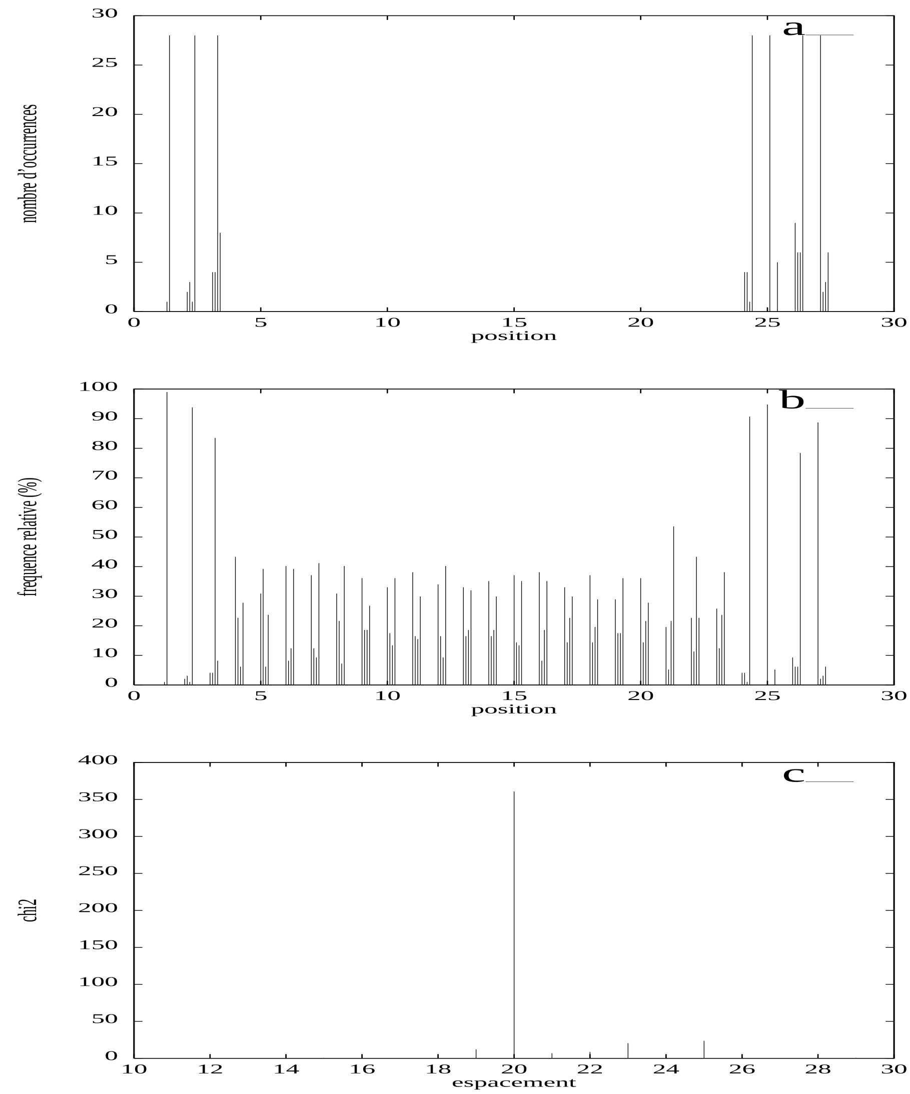

Figure 6.4: Graphiques correspondant au modèle $TTGN_{20}TATA$ (χ^2 = 238,14) identifié dans le lot A par LeCombi (lecture directe uniquement) avec la couverture CCPR donnée dans le texte, J (nombre de jokers) = 20, q = 20% et pas d'erreurs. Chaque groupe de barres des graphiques a et b représente les nucléotides A, C, G et T dans cet ordre.

Notons également que les 2 ou 3 N situés entre les T et les A représentent le plus souvent un T suivi par un A, avec éventuellement une position au centre pouvant être indifféremment un T ou un A. Enfin, le pic à -45 correspond généralement au motif central; plus précisément, c'est le premier symbole de la suite A_{4-5} qui se trouve en général à cet endroit.

Cette étude demeure bien sûr tout à fait préliminaire et une analyse plus précise serait nécessaire pour confirmer ces premiers résultats. Il pourrait être éventuellement intéressant de classer les gènes sur la base des différents motifs observés autour du pic à -45.

6.2.2.3 Discussion

Ce premier exemple nous conduit à faire un certain nombre d'observations concernant nos algorithmes.

La première est que nous avons été amenés, au cours de cette analyse, à tester plusieurs de ces algorithmes : certains (HMoivre, Combi) se sont révélés plus intéressants pour repérer les petits motifs plus conservés correspondant aux sites -10 et -35, alors que c'est LeCombi (avec des jokers) qui a permis de dégager une structure dans la région à -45. Le site à cet endroit, s'il se confirme qu'il existe, semble être plus long et, surtout, il présente des positions apparemment non-spécifiques. Ces positions forment souvent des suites contiguës qui peuvent ne pas être de taille constante (bien que celles des modèles que nous repérons sont en général quasi-optimales — voir les graphiques c des figures 6.4 à 6.11 précédentes). Il serait alors intéressant de disposer d'un algorithme qui permette de traiter de telles suites, c'est-à-dire, qui permette de traiter le problème des jokers de longueur variable (en anglais 'variable length don't care symbols'). Nous pouvons à l'heure actuelle le faire de manière indirecte en autorisant des trous, mais le concept n'est pas tout à fait le même et il est essentiel de maintenir les deux idées séparémment. D'autre part, une telle approche serait importante pour l'identification des motifs à -10 et -35 puisque la promotion de la transcription est associée en réalité non à chacun de ces deux motifs isolément, mais à la structure générale : séquence -35, séquence -10 et espacement optimal (17 bases) entre les deux.

La seconde observation est que nous avons été obligés, dans tous les cas, de faire tourner plusieurs fois chacun des algorithmes en faisant varier les divers paramètres (quorum, nombre d'erreurs, couvertures et pondérations, nombre de jokers). Or le choix de ceux-ci demeure pour l'instant délicat. Est-ce que l'intuition et une certaine expérience des données sont les seuls guides dont nous puissions disposer? Est-ce que la seule stratégie possible consiste à lancer le maximum de tests, essayant de couvrir au mieux l'arsenal à notre disposition, pour ne conserver à la fin que les résultats qui paraissent les plus intéressants?

Quant à ces résultats eux-mêmes, ils se trouvent encore, d'une certaine façon, à l'état brut. Ils possèdent une richesse d'information qui n'est pas exploitée, un exemple plus précis de cela sera donné plus tard. Ils représentent également une quantité de données qui est parfois presque aussi considérable que celle dont nous disposions au départ, simplement mieux organisée, ou organisée différemment. Il faut donc souvent faire appel à d'autres techniques pour essayer de trier ce qui est 'vu' et de l'interpréter. Ces techniques (par exemple ici statistiques) sont en général automatisables, mais les liaisons entre les diverses étapes de l'analyse doivent encore souvent être faites à la main. Est-il possible de les automatiser, du moins en partie? En particulier, est-il possible de fournir une représentation des modèles repérés autre qu'une simple liste comme nous le faisons pour l'instant? Par exemple, nous avons fourni plusieurs modèles correspondant aux sites -10 et -35 et obtenus avec HMoivre qui visiblement sont des variants les uns par rapport aux autres. Pourrait-on les résumer en préservant toute l'information qu'ils contiennent? Pourrait-on faire de même avec les modèles fournis par Combi ou LeCombi?

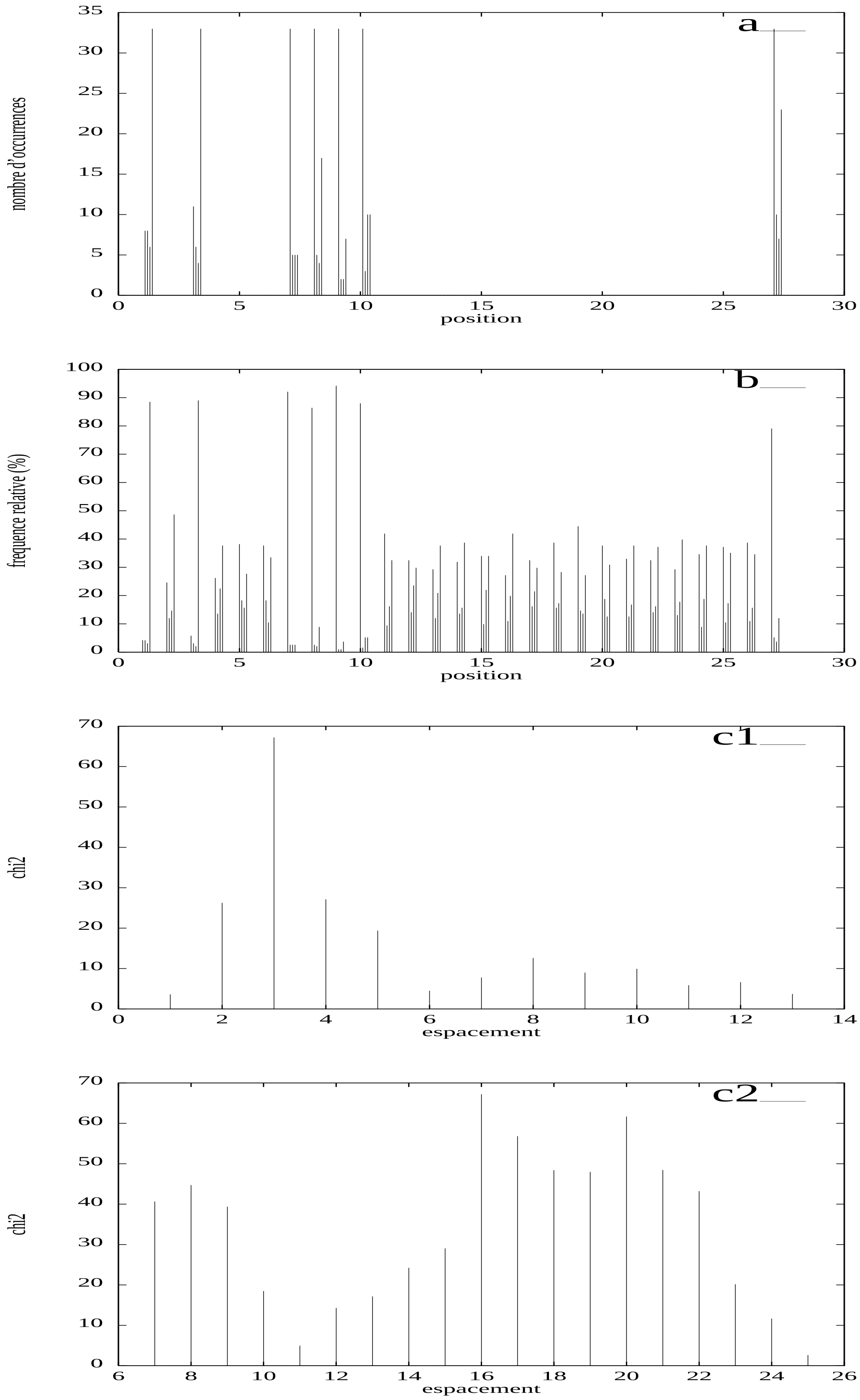

Figure 6.5: Graphiques correspondant au modèle $TNTN_3AAAAN_{16}A$ ($\chi^2 = 71{,}18$) identifié dans le lot A par LeCombi (lecture directe uniquement) avec la couverture CCPR donnée dans le texte, J (nombre de jokers) $= 20$, $q = 20\%$ et pas d'erreurs. Chaque groupe de barres des graphiques a et b représente les nucléotides A, C, G et T dans cet ordre.

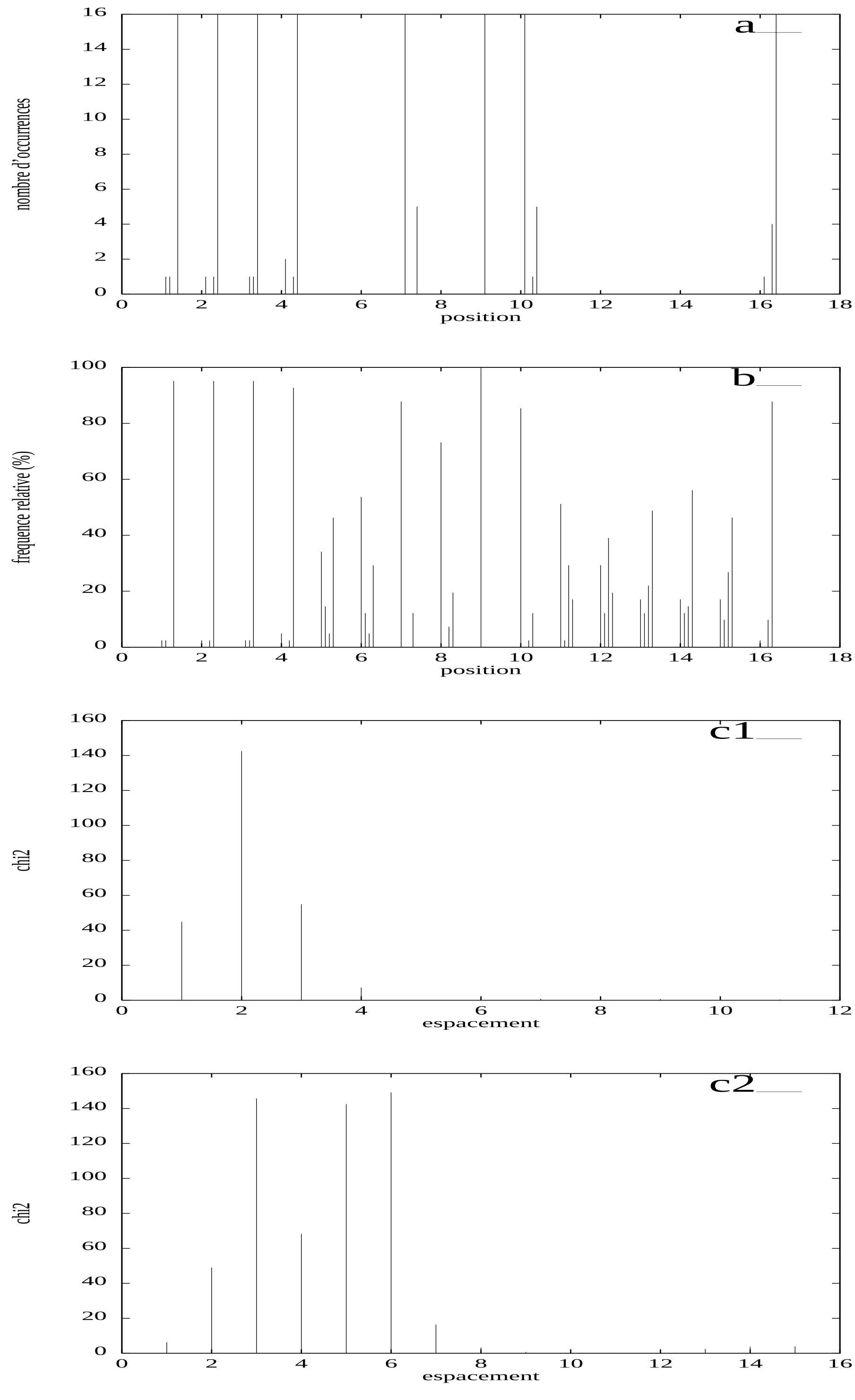

Figure 6.6: Graphiques correspondant au modèle $TTTTN_2ANAAN_5T$ ($\chi^2 = 175{,}69$) identifié dans le lot B par LeCombi (lecture directe uniquement) avec la couverture CCPR donnée dans le texte, J (nombre de jokers) = 8, $q = 20\%$ et pas d'erreurs. Chaque groupe de barres des graphiques a et b représente les nucléotides A, C, G et T dans cet ordre.

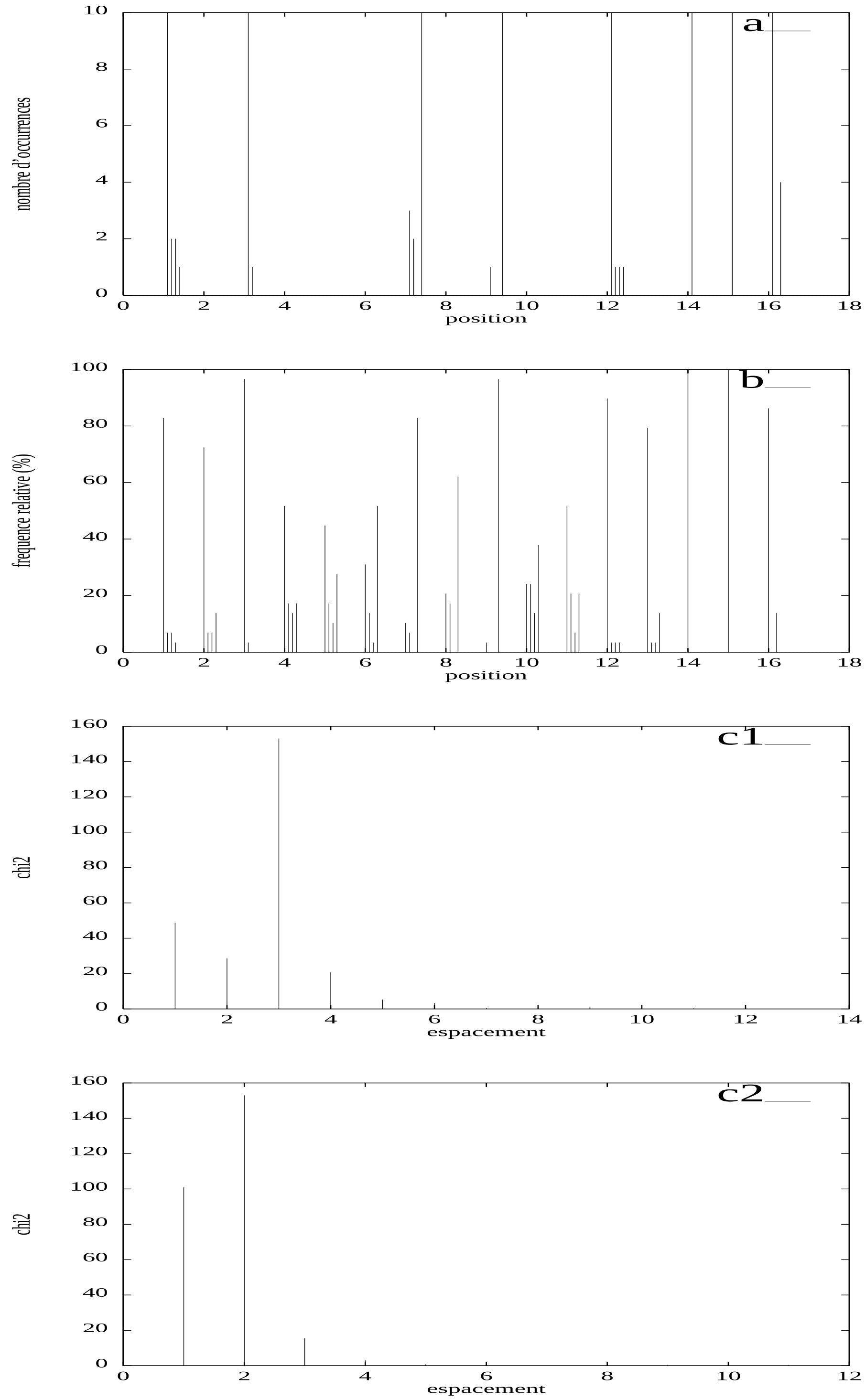

Figure 6.7: Graphiques correspondant au modèle $ANAN_3TNTN_3AAAA$ (χ^2 = 89,62) identifié dans le lot B par LeCombi (lecture directe uniquement) avec la couverture CCPR donnée dans le texte, J (nombre de jokers) = 8, q = 20% et pas d'erreurs. Chaque groupe de barres des graphiques a et b représente les nucléotides A, C, G et T dans cet ordre.

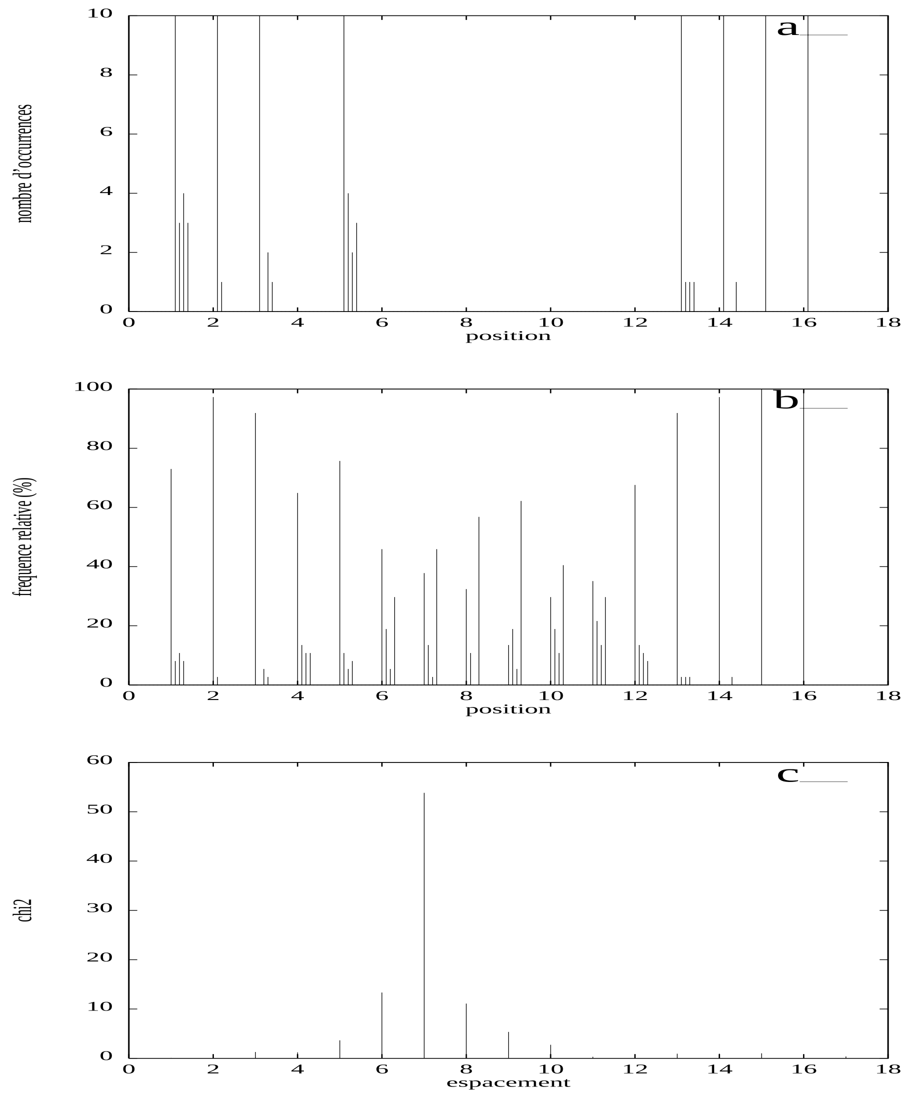

Figure 6.8: Graphiques correspondant au modèle $AAANAN_7AAAA$ (χ^2 = 49,05) identifié dans le lot B par LeCombi (lecture directe uniquement) avec la couverture CCPR donnée dans le texte, J (nombre de jokers) = 8, q = 20% et pas d'erreurs. Chaque groupe de barres des graphiques a et b représente les nucléotides A, C, G et T dans cet ordre.

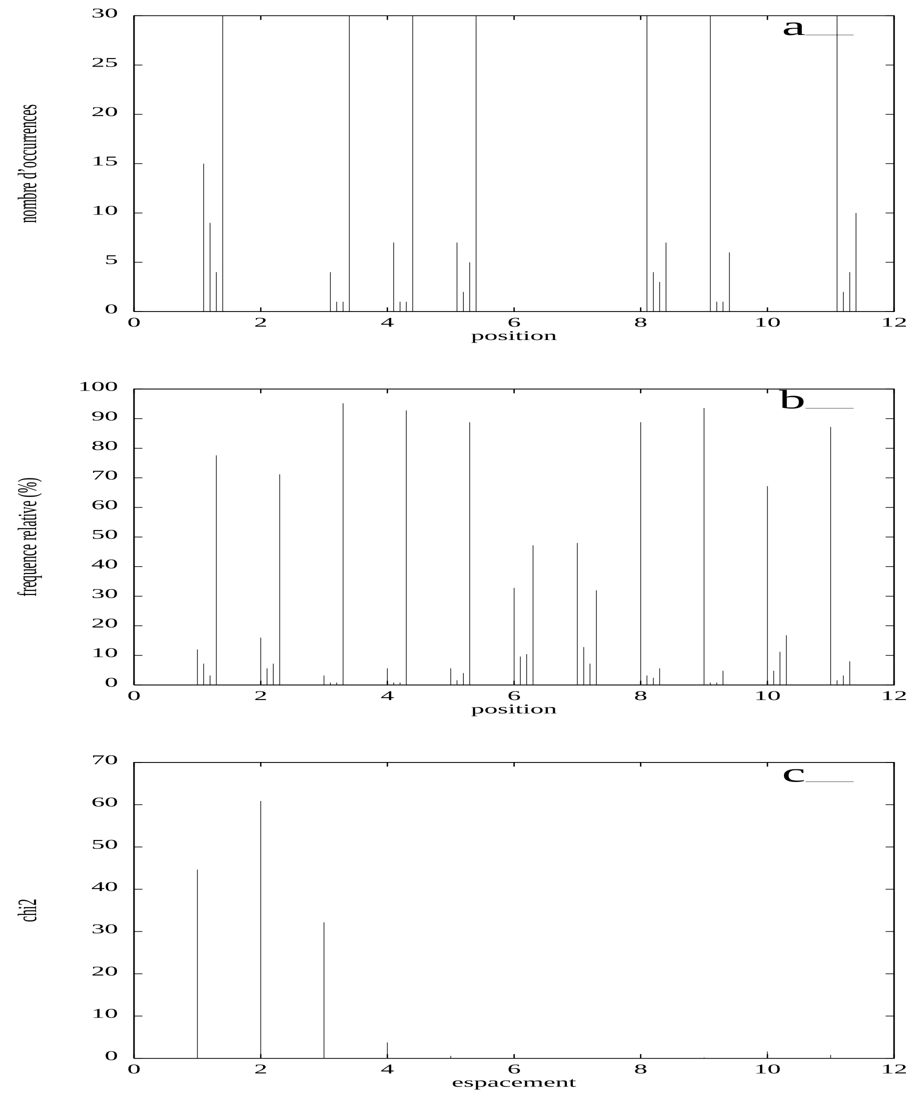

Figure 6.9: Graphiques correspondant au modèle $TNTTTN_2AANA$ ($\chi^2 = 78{,}60$) identifié dans le lot B par ADN-LeCombi (lecture dans les deux sens) avec la couverture CCPR donnée dans le texte, J (nombre de jokers) $= 4$, $q = 20\%$ et pas d'erreurs. Chaque groupe de barres des graphiques a et b représente les nucléotides A, C, G et T dans cet ordre.

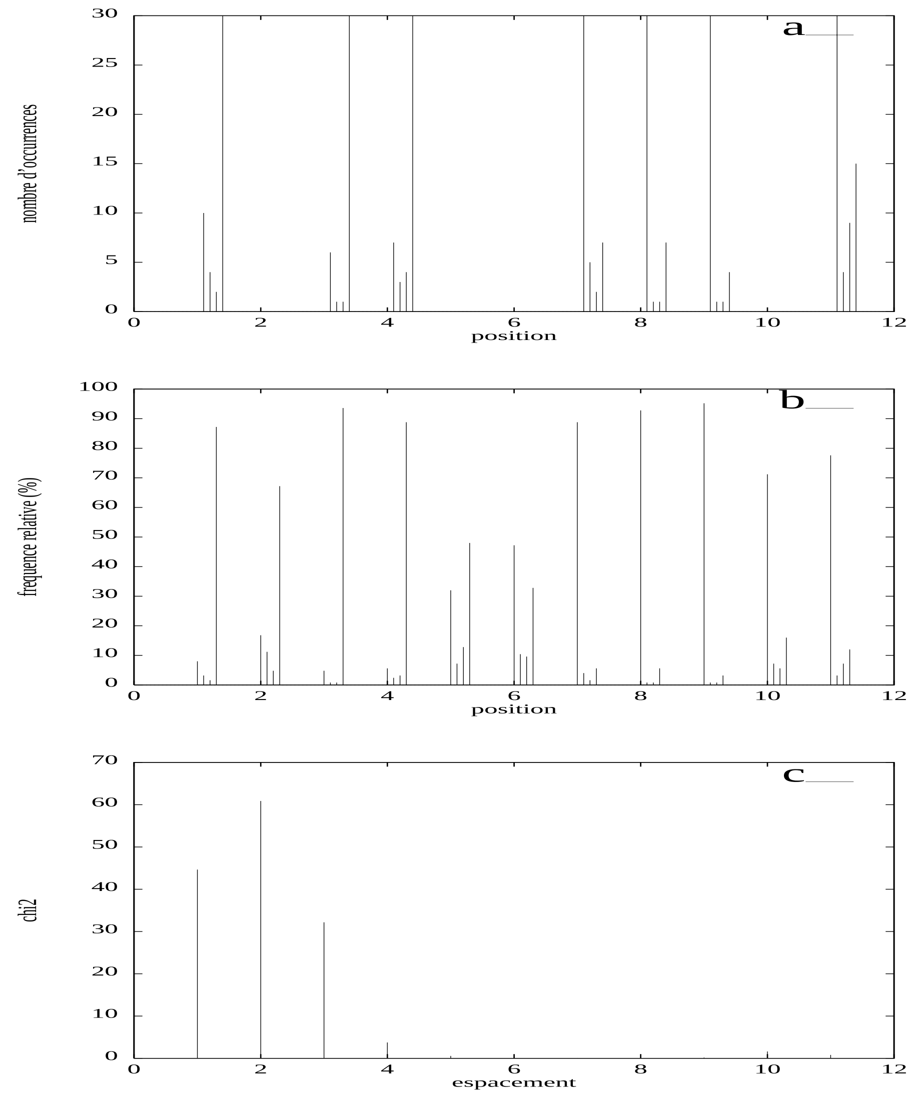

Figure 6.10: Graphiques correspondant au modèle $TNTTN_2AAANA$ ($\chi^2 = 78{,}60$) identifié dans le lot B par ADN-LeCombi (lecture dans les deux sens) avec la couverture CCPR donnée dans le texte, J (nombre de jokers) $= 4$, $q = 20\%$ et pas d'erreurs. Chaque groupe de barres des graphiques a et b représente les nucléotides A, C, G et T dans cet ordre.

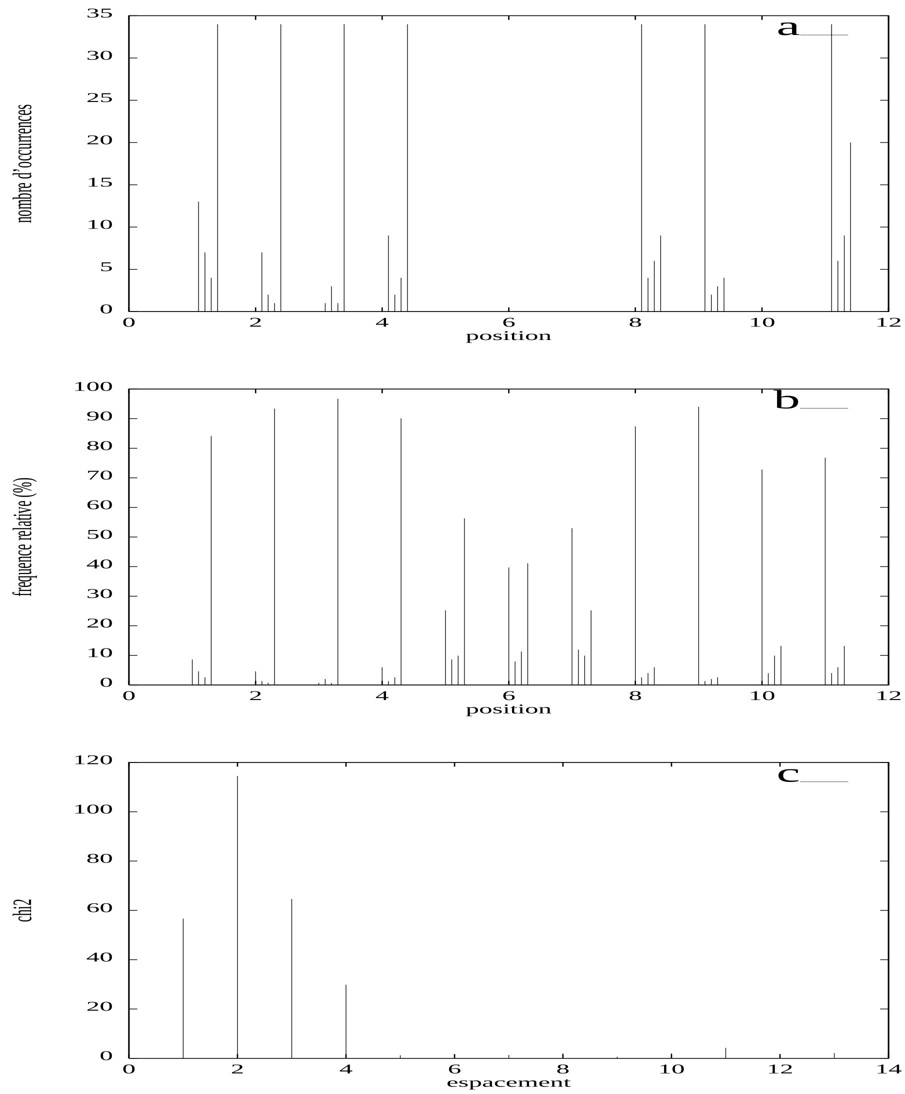

Figure 6.11: Graphiques correspondant au modèle $TNTTN_3NAAAA$ ($\chi^2 = 74{,}29$) identifié dans le lot B par ADN-LeCombi (lecture dans les deux sens) avec la couverture CCPR donnée dans le texte, J (nombre de jokers) = 4, $q = 20\%$ et pas d'erreurs. Chaque groupe de barres des graphiques a et b représente les nucléotides A, C, G et T dans cet ordre.

Enfin, le bon grain se trouve souvent mélangé à l'ivraie dans ce qui est obtenu. Cela est particulièrement le cas lorsque nous devons autoriser beaucoup de souplesse dans la recherche de mots communs similaires. Ainsi le motif à -35, qui est bien plus 'faible' que celui à -10, n'a pu être repéré avec HMoivre et 0 erreur que parce que nous avons fait usage *a posteriori* de statistiques. Celles-ci permettent de le 'voir' parce qu'il comporte deux bases — une Cytosine et une Guanine — peu fréquentes dans toute la région des promoteurs. Le fait que les modèles correspondants aient moins d'occurrences est ainsi 'compensé' par le caractère 'anormal' de leur présence et permet de dire avec une certaine assurance qu'ils sont probablement intéressants. Si nous avions voulu utiliser HMoivre tout seul, nous aurions du autoriser au moins 2 substitutions et, surtout, fixer la longueur des modèles à une valeur relativement courte (autour de 6 ou 7 nucléotides) afin de localiser cette séquence -35 avec des quorums plus élevés. Dans ce cas cependant, et sur cet exemple, les signaux commencent alors à se mélanger et le bruit augmente. D'une façon générale, Combi et LeCombi semblent être plus discriminants dans la mesure où ils permettent plus de souplesse tout en fixant de manière plus forte certaines positions. Est-ce que cela signifie que la définition de ressemblance sur laquelle ils sont établis correspond mieux aux processus de reconnaissance biologique? C'est une question à laquelle nous ne savons répondre.

6.2.3 Les promoteurs dépendents d'une protéine d'activation catabolique de *Escherichia coli*

6.2.3.1 Introduction

Le nombre de protéines produites en une unité de temps par un organisme n'est pas constant. Il varie, pour un même gène, au cours du temps, et d'un gène à l'autre, de telle sorte que les besoins de la cellule soient satisfaits tout en évitant de réaliser des synthèses inutiles. Les mécanismes de régulation de l'expression des gènes peuvent être du type 'marche-arrêt' ou plus élaborés. Malgré leur diversité, il est possible de regrouper tous ces mécanismes en deux catégories : ceux dont la régulation est positive (activation) et ceux dont la régulation est négative (inhibition) (voir la section 2.1.1.3). Dans un système de régulation négatif, un inhibiteur présent dans la cellule bloque la transcription ou la traduction. Un antagoniste de l'inhibition, généralement appelé inducteur, est alors nécessaire pour initier un de ces deux processus. Dans un système de régulation positif, une molécule effectrice (qui peut être une protéine, une petite molécule ou un complexe moléculaire) active un promoteur. Les deux systèmes ne sont pas exclusifs et l'expression d'un gène peut être à la fois positivement et négativement régulée.

La bactérie *Escherichia coli* possède ainsi une protéine d'activation catabolique (CAP, en abbréviation de l'anglais 'Catabolic Activator Protein'), appelée également la protéine réceptrice de la protéine cyclique AMP (CRP, de l'anglais 'Cyclic-AMP Receptor Protein'), qui est impliquée dans la régulation de la transcription de quelques-uns de ses gènes. Ces derniers (isolés ou regroupés en une unité fonctionnelle appelée opéron) ne peuvent en effet être transcrits que si le complexe cAMP-CRP est fixé à l'ADN dans la région du promoteur (dans le cas des opérons, les gènes composant l'unité sont alors co-régulés et co-transcrits). En fait, ce complexe peut également agir comme un répresseur de la transcription de certains gènes, cela est le cas en particulier du gène qui réalise la propre synthèse des CRP [Botsford and Harman, 1992] [Combrugghe *et al.*, 1984], cependant cette répression est probablement due au fait que le site promoteur situé à -35 du gène inhibé chevauche celui où doit se fixer le complexe afin d'activer la transcription d'un autre gène situé à proximité du premier. Par ailleurs, deux sites ont parfois été identifiés pour un même gène, un site primaire et un site secondaire qui fixe la CRP moins fortement et qui peut participer à la localisation du site primaire par le complexe [Combrugghe *et al.*, 1984]. D'une façon plus générale, les promoteurs

CAP-dépendants peuvent être groupés en trois ou quatre classes [Combrugghe *et al.*, 1984] [Ebright, 1993] [Gaston *et al.*, 1990] suivant la localisation des sites fixateurs de la CRP, la fonction de celles-ci et leur mode d'action (le contact entre molécules pouvant intervenir uniquement entre la CRP et l'ADN, ou avoir lieu également entre la CRP et l'ARN polymérase, ou encore impliquer d'autres protéines qui viendraient se fixer à la suite les unes des autres sur l'ADN entre la CRP et la polymérase).

Ce qu'il est important de noter dans le cadre de notre analyse, c'est que tous ces sites présentent deux motifs principaux qui sont séparés par, en général, 6 bases. Le premier motif, $TGTGA$, est assez fortement conservé tandis que le second motif, moins conservé, est soit une version complémentaire inversée du second ($TCACA$), soit un autre type de séquence [Combrugghe *et al.*, 1984]. La possibilité d'une symétrie miroir entre les deux séquences composant le site s'explique par la nature dimérique de la protéine CRP qui interagit avec l'ADN sur 22 à 25 bases [Berg and von Hippel, 1988] [Combrugghe *et al.*, 1984]. Cette nature dimérique ne signifie cependant pas que les deux séquences nucléotidiques établissant le contact avec la CRP sont d'égales importances pour la fixation du complexe. La moindre conservation du second motif, résultant d'une asymétrie du site, peut ainsi jouer un rôle non négligeable dans cette fixation, en particulier lorsque qu'il y a également contact entre CRP et ARN polymérase [Berg and von Hippel, 1988] [Combrugghe *et al.*, 1984]. Dans ce cas, la fixation de la CRP à l'ADN du côté où viendra se placer la polymérase doit être suffisamment flexible pour que le contact entre les deux protéines puisse ensuite s'établir [Berg and von Hippel, 1988].

Cette relative non-conservation est caractéristique des séquences impliquées dans la promotion de la transcription de gènes pour lesquels la quantité de produit final obtenue (les protéines) doit être finement régulée. Dans le cas des gènes de *coli* dont les promoteurs sont CAP-dépendants, c'est tout le site de fixation de la CRP, plus les sites situés à -10 et -35 qui sont relativement peu conservés. Si cela n'était pas le cas, c'est-à-dire si les signaux présents dans ces sites étaient très 'forts', le promoteur serait trop efficace et donc difficile à réguler [Botsford and Harman, 1992] [Combrugghe *et al.*, 1984]. Cette situation ne facilite bien sûr pas l'identification de ces sites par une analyse comparative des séquences, elle rend également délicate l'interprétation des résultats obtenus ainsi que nous le verrons plus loin. À cela s'ajoute parfois la difficulté de savoir sur quel brin de l'ADN est localisé le site. Le motif le plus fort n'est pas toujours celui qui se trouve le plus éloigné de la première base transcrite et ne peut servir d'indicateur du sens de lecture. Par ailleurs, dans certains cas, on ignore quel gène est effectivement régulé par la CRP [Combrugghe *et al.*, 1984]. Ces deux aspects, faible conservation qui n'est pas liée à la présence d'erreurs mais est biologiquement nécessaire, et ignorance du sens dans lequel les séquences dont on dispose doivent être 'lues', posent ainsi le problème de l'établissement d'un ensemble de données 'propres' sur lequel tester les algorithmes. Ce problème est très important et est souvent négligé. Il n'est cependant pas toujours facile à résoudre, le cas traité ici en est un exemple.

Comme pour les promoteurs de la transcription dans notre illustration précédente, l'exposition des résultats obtenus va suivre le déroulement chronologique. Nous allons ainsi commencer par travailler sur un ensemble de données établi par Stormo [Schneider *et al.*, 1986] et repris par Lawrence [Lawrence and Reilly, 1990] dans le cadre de son algorithme d'optimisation stochastique (sections 3.4.3.3 et 4.2.2.2). Au départ, ces données ne faisaient que constituer un jeu d'essai particulièrement intéressant pour nos algorithmes car les motifs communs qui correspondent aux sites de fixation de la CRP, et que nous devions être capables de repérer, sont, pour les raisons que nous avons mentionnées, très dégénérés. Ils nous permettaient ainsi de tester au mieux l'efficacité et la robustesse de nos méthodes. Ce n'est qu'ensuite que nous avons été amenés, grâce à A. Danchin [Danchin, 1995], à remettre en question ce jeu d'essai, à en établir un autre (composé en fait des mêmes séquences mais 'lues' parfois dans un autre sens), puis, en fin de compte, à élaborer l'extension ADN-X décrite dans la section 6.2.1.2. Observons que nous ne sommes concernés ici que par l'identification du site de fixation de la CRP, pas par la localisation des séquences à -10 et -35 des promoteurs correspondants.

Le premier jeu d'essai avec lequel nous travaillons, celui de Stormo, est composé de 17 séquences de 110 bases chacunes sélectionnées dans GenBank (correspondant aux gènes *ara*, *bgl*R mut, *col*E1, *cat*, *crp*, *cya*, *deo*P2, *gal*, *ilv*B, *lac*, *mal*E, *mal*K, *mal*T, *omp*A, *p*BR-P4, *tna*A, *uxu*AB — nous avons éliminé une $18^{ième}$ séquence correspondant au gène *tdc* qui ne provenait pas de GenBank). Ces séquences avaient été découpées par Stormo *et al.* de façon à ce que le motif soit situé à des positions quelconques sur chacune des séquences. Le but des auteurs était de montrer que leur méthode est capable de localiser la position de motifs correspondant à des sites même lorsque les séquences ne sont pas alignées.

Le second jeu d'essai est composé des mêmes 17 séquences, certaines cependant 'lues' en ordre inversé complémentaire et correspondant ainsi à des séquences présentes sur l'autre brin de l'ADN. Ce jeu a été sélectionné par Berg [Berg and von Hippel, 1988], d'après un choix effectué antérieurement par Combrugghe [Combrugghe *et al.*, 1984] auquel 4 séquences ont été ajoutées.

Bien que cela ne soit pas clairement indiqué, nous pouvons supposer que Berg a réalisé le choix du brin sur lequel lire une séquence d'après la séquence qui présentait une plus forte conservation du premier motif *TGTGA* du site de fixation de la CRP. Quant à Stormo, nous pensons que la sélection qu'il a effectuée (celle de GenBank) correspond plutôt à celle pour laquelle le site est correctement orienté par rapport à la direction de transcription du gène qu'il régule lorsque celui-ci est connu (et unique!). Seule la sélection des sites relatifs aux gènes *mal*K et *p*BR-P4 est douteuse ainsi que nous l'avons vérifié sur la base de données Colibri [Médigue *et al.*, 1993].

Dans tous les cas, la position de ces sites sur chacune des séquences a pu par ailleurs être déterminée expérimentalement. Nous comparons ainsi nos résultats à ces données expérimentales et, lorsque les deux diffèrent, aux résultats obtenus par Lawrence avec sa méthode (celle correspondant à l'article de 1990 [Lawrence and Reilly, 1990]).

6.2.3.2 Résultats

Nous commençons par donner une figure qui permet de mieux comprendre en quoi le problème traité ici est difficile. La figure 6.3 montre un alignement des sites que nous devons être capables d'identifier. Nous ne présentons qu'un site par gène lorsque celui-ci en possède deux. Nous avons, pour le propos de cette figure, choisi celui repéré par l'algorithme de Lawrence (site 2), le choix du site 1 ou un choix arbitraire produirait un résultat analogue. En-dessous de l'alignement, nous indiquons la matrice du nombre des bases observées pour chaque colonne de l'alignement. Nous pouvons voir qu'aucune position parmi les 22 que le site comporte n'est strictement conservée, que sur seulement 3 des 22 positions on observe uniquement 2 bases, et qu'à 11 des 22 positions, toutes les bases sont présentes. Par ailleurs, la même base est présente à plus de 50% dans seulement 13 colonnes de la matrice. Enfin, si nous considérons deux de ces motifs et les comparons entre eux, par exemple le onzième motif (associé au gène *mal*K) et le treizième (associé au gène *omp*A), nous pouvons observer qu'il faut 15 substitutions pour passer de l'un à l'autre, 12 si l'on ne considère que les positions centrales (c'est-à-dire, le motif probablement palindromique et les bases intermédiaires). Notons que la matrice donnée ici est approximativement celle qu'obtient Lawrence [Lawrence and Reilly, 1990] à la fin de son algorithme d'optimisation stochastique, mais nous verrons un peu plus loin qu'il a en fait besoin

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	T	T	T	T	T	T	G	A	T	C	G	T	T	T	T	C	A	C	A	A	A	A
	T	T	A	T	T	T	G	C	A	C	G	G	C	G	T	C	A	C	A	C	T	T
	A	A	C	T	G	T	G	A	G	C	A	T	G	G	T	C	A	T	A	T	T	T
	G	T	A	T	G	C	A	A	A	G	G	A	C	G	T	C	A	C	A	T	T	A
	A	G	G	T	G	T	T	A	A	A	T	T	G	A	T	C	A	C	G	T	T	T
	T	T	A	T	T	T	G	A	A	C	C	A	G	A	T	C	G	C	A	T	T	A
	T	A	A	T	T	T	A	T	T	C	C	A	T	G	T	C	A	C	A	C	T	T
	A	A	A	C	G	T	G	A	T	C	A	A	C	C	C	C	T	C	A	A	T	T
	T	A	A	T	G	T	G	A	G	T	T	A	G	C	T	C	A	C	T	C	A	T
	T	T	C	T	G	T	A	A	C	A	G	A	G	A	T	C	A	C	A	C	A	A
	T	T	C	T	G	T	G	A	A	C	T	A	A	A	C	C	G	A	G	G	T	C
	A	A	T	T	G	T	G	A	C	A	C	A	G	T	G	C	A	A	A	T	T	C
	A	T	G	C	C	T	G	A	C	G	G	A	G	T	T	C	A	C	A	C	T	T
	G	A	T	T	G	T	G	A	T	T	C	G	A	T	T	C	A	C	A	T	T	T
	T	G	T	T	G	T	G	A	T	G	T	G	G	T	T	A	A	C	C	C	A	A
	C	G	G	T	G	T	G	A	A	A	T	A	C	C	G	C	A	C	A	G	A	T
	A	C	C	T	G	T	G	A	C	G	G	A	A	G	A	T	C	A	C	T	T	C
A	6	6	6	0	0	0	3	15	6	4	2	11	3	4	1	1	13	3	12	2	5	4
C	1	1	4	2	1	1	0	1	4	7	4	0	4	3	2	15	1	13	2	6	0	3
G	2	3	3	0	12	0	13	0	2	6	6	3	8	5	3	0	2	0	2	2	0	0
T	8	7	4	15	4	18	1	1	5	2	5	3	2	5	12	1	1	1	1	7	11	9

Tableau 6.3: Sites de fixation des CRP (un seul site sélectionné par séquence) et matrice des bases observées (les lignes corespondent aux symboles A, C, G, et T respectivement) : une indication de la difficulté que pose la localisation de ces sites.

de donner un 'coup de pouce' initial à sa méthode afin qu'elle puisse fournir un tel résultat.

Avec la version originale de H/LeMoivre dont nous disposions à l'époque où nous nous sommes intéressés à ce problème, nous n'étions pas capables de le 'résoudre' car le nombre d'erreurs (même s'il s'agissait uniquement de substitutions) que nous pouvions permettre était limité à 2 ou 3. Comme nous l'avons vu, cela n'est pas suffisant pour repérer ce site. Les versions ultérieures de l'algorithme, en particulier avec la première et la seconde des modifications présentées dans la section 5.4.2.1.5, nous ont cependant permis de travailler avec jusqu'à 6 ou 7 substitutions en un temps qui est long mais demeure acceptable (1 journée pour cet exemple de 17 séquences de longueur moyenne 110 sur une Silicon Graphics Indigo WorkStation (R4000)).

Le premier résultat sur ce jeu d'essai a donc été obtenu en cherchant avec HMoivre les plus longs modèles présents dans au moins 95% des séquences et avec 6 substitutions autorisées au plus. Avec ces paramètres, nous sommes capables d'identifier 14 modèles de longueur 16 présents dans 16 des 17 séquences (la $17^{ième}$ séquence où aucun modèle n'est localisé est associée au gène *cat*). Parmi ces modèles, 2 correspondent au site exactement (ce sont les modèles $\underline{TGTGA}ACTATT\underline{TCACA}$ et $\underline{TGAGA}TCTAGG\underline{TCACA}$ — nous soulignons l'emplacement du motif probablement palindromique). Les 12 autres modèles correspondent au site mais les occurrences sont incorrectes sur une ou plusieurs séquences (jusqu'à un maximum de 6 dans un cas). Par 'incorrecte', nous entendons qu'elles ne représentent pas le site expérimental. Il s'agit parfois du site considéré comme étant le second le plus fort par le modèle statistique de Lawrence. Cela est en particulier le cas pour le site de fixation de la CRP du gène *gal*, qui est presque toujours trouvé 18 positions en amont du site expérimental. D'une manière générale, aucun modèle ne rate le site complètement. Le plus intéressant cependant est que 11 des 14 modèles satisfont en tant qu"étiquette' au motif (hypothétiquement) palindromique $\underline{TGTGA}NNNNNN\underline{TCACA}$ et que les 3 autres le satisfont en grande partie. Enfin aucun autre modèle n'est rencontré. La solution apportée par HMoivre paraît donc être suffisamment sélective malgré la grande souplesse autorisée dans la définition de ressemblance.

Cette 'solution' est cependant, du point de vue biologique, très contestable pour les raisons indiquées dans l'introduction. Le jeu établi par Stormo (celui utilisé par Lawrence) obéit à un critère de sélection des données, en particulier quant au brin sur lequel il faut lire la séquence, qui n'est pas le seul possible. À la suite d'une suggestion de A. Danchin, nous avons établi un second jeu d'essai dont le critère, correspondant à celui de Combrugghe et de Berg, est le suivant : la séquence où se trouve le site doit être 'lue' sur le brin où le premier motif $TGTGA$ apparaît le mieux conservé. L'ensemble obtenu est composé des mêmes séquences que celles du jeu de Stormo, simplement dans 6 cas (séquences associées aux gènes *ara*, *crp*, *deo*P, *gal*, *mal*E et *omp*A) la séquence de Combrugghe est le complémentaire inversé de celle du jeu de Stormo.

Par ailleurs, si le premier jeu d'essai que nous avons réalisé nous a fourni une bonne occasion de tester l'algorithme HMoivre modifié, il nous a aussi montré qu'il serait sans doute plus efficace de travailler, non avec HMoivre, mais avec LeCombi, puisque plusieurs des positions à l'intérieur du site paraissent ne pas intervenir dans la reconnaissance du signal par la protéine CRP et sont donc non-spécifiques.

Pour compléter notre analyse préliminaire, nous avons alors :

- testé une version implémentée par Viari [Viari, 1995] de l'algorithme de Lawrence (version de l'article de 1990) sur le second jeu d'essai (celui de Berg et Combrugghe);

- utilisé LeCombi avec la couverture

$$CCPR = \left\{ \begin{array}{l} (\{X\}, \infty)\ \forall X \in \Sigma = \{A, C, G, T\}, \\ (\{A, C, G, T\}, 8), \\ (S, 0)\ pour\ tous\ les\ autres\ ensembles \end{array} \right\}$$

  un quorum de 95% et en autorisant 2 erreurs;

- utilisé ADN-LeCombi (lecture dans les deux sens) dans les mêmes conditions que ci-dessus.

Avec l'algorithme de Lawrence nous avons recherché un motif (c'est-à-dire en fait une matrice) de longueur 22 puisque dans ce cas la taille du site doit être connue à l'avance [Lawrence and Reilly, 1990]. Dans nos algorithmes, nous avons demandé à chaque fois les modèles les plus longs.

Commençons par les résultats obtenus avec l'algorithme de Lawrence. Lorsque nous recherchons avec ce dernier les sites de fixation de la CRP, soit sur les séquences de Stormo, soit sur celles de Berg, nous nous rendons compte qu'en fait l'algorithme ne peut pas vraiment partir de positions initiales quelconques afin de construire la matrice spécifique du site, du moins dans sa version de 1990. Nous rappelons (voir la section 4.2.2.2) que la différence fondamentale entre cette version et celle de 1993 (échantillonnage de Gibbs) porte sur la question du choix du site, que nous noterons 'variable', lorsque les positions de ce site sur les autres séquences sont maintenues fixes (fin de l'étape d'échantillonnage). Dans la méthode d'échantillonnage de Gibbs, la position x sur la séquence 'variable' est choisie avec une certaine probabilité $\frac{L_x}{\sum_x L_x}$ qui dépend à la fois de la quantité d'information L_x que le mot à cette position ajoute à la matrice à une étape donnée, et de celle ajoutée (ou retranchée) par les autres mots de la séquence. Dans la méthode EM de 1990, les positions du site ne sont choisies qu'à la fin. La matrice représente alors les fréquences observées à chaque position du site hypothétique, pondérées par la probabilité que ce site démarre en position x pour tout x (voir la section 4.2.2.2). Si la matrice de départ est trop fausse, la méthode peut rapidement mener à des maximums locaux et produire des résultats complètement faux, comme nous en avons fait l'expérience. Par contre, il suffit souvent de donner un 'coup de pouce' à cette matrice initiale pour que le résultat devienne bien meilleur. Le coup de pouce que nous lui avons donné a consisté tout simplement à forcer la présence d'un T de manière fortement majoritaire (85%) en quatrième position du site, c'est-à-dire, en première position du motif hypothétiquement palindromique. C'est une stratégie un peu *ad hoc*, et les résultats obtenus que nous présentons maintenant sont donc à prendre avec une certaine prudence. Ces résultats sont :

- par rapport au jeu de Stormo, le site est correctement localisé sur 11 des 17 séquences et manqué sur 6 séquences, correspondant aux gènes *bgl*R, *cat*, *gal* (placé 18 positions en amont comme dans HMoivre), *ilv*B, *mal*T et *omp*A;
- par rapport au jeu de Berg, le site est correctement localisé sur 13 des 17 séquences, il est manqué sur *bgl*R, *cat*, *mal*K et *ilv*B.

Le temps d'exécution de l'algorithme est très court : quelques minutes sur une Silicon Graphics Indigo WorkStation (R4000). Il serait intéressant de voir ce que la version de 1993 (Échantillonnage de Gibbs) est capable de repérer, en particulier s'il est alors possible de partir de n'importe quelles positions initiales comme l'affirment les auteurs.

Quand nous utilisons cette fois LeCombi sur le jeu d'essai de Stormo, nous obtenons un seul plus long modèle (couvrant 17 bases) qui repère le site correctement sur 15 des 17 séquences et le place ailleurs sur la séquence associée au gène *gal*. L'occurrence trouvée dans ce cas se situe à nouveau 18 bases en amont du site expérimental. Rappelons que c'est celle la plus couramment observée avec HMoivre et qu'elle correspond au choix initial de Lawrence dans son article [Lawrence and Reilly, 1990]. L'algorithme de Lawrence ne place le site de *gal* au bon endroit que lorsqu'on y introduit des informations supplémentaires, en particulier concernant la nature hypothétiquement symétrique du site [Lawrence and Reilly, 1990]. Le modèle que nous trouvons, $\underline{GTGA}NNNNNN\underline{TCNCA}NT$, ne reflète plus cette symétrie de manière parfaite. Quant au site qui est 'manqué', il correspond une fois de plus à celui de fixation de la CRP pour la régulation du gène *cat* (il y a en fait deux sites relatifs à *cat*, aucun des deux n'est correctement localisé). Si nous faisons l'hypothèse que ce qu'il fallait identifier dans cet exemple était les sites de la CRP indiqués dans [Lawrence and Reilly, 1990], alors le modèle identifié avec LeCombi paraît légèrement moins 'bon' que les deux meilleurs modèles repérés par HMoivre, cependant c'est le seul que nous trouvons et nous pouvons donc argumenter qu'il est plus sélectif. Par ailleurs, ce modèle est plus long d'une unité que les modèles construits par HMoivre. Il est important de noter également que ces résultats ont été produits en 4 fois moins de temps que pour HMoivre (environ 6 heures sur une Silicon Graphics Indigo WorkStation (R4000)).

Avec le jeu de Berg, nous obtenons, avec des temps sensiblement identiques, 7 modèles, également de longueur 17, qui sont tous associés au site : le premier modèle localise correctement le site sur 16 des 17 séquences, les six suivants placent le site ailleurs sur 1 ou 2 séquences. Seul le modèle $AN\underline{TGTGA}NNTNNN\underline{NCNC}$, dont toutes les occurrences correspondent correctement au site, n'identifie pas ce site sur *cat*. La caractéristique la plus frappante des sept modèles trouvés est qu'ils comportent tous le premier motif $TGTGA$, mais jamais le second complètement, et qu'un T est toujours présent en troisième position après le motif $TGTGA$, à un endroit où il ne semble pas y avoir de symbole majoritaire dans le jeu de Stormo et Lawrence (voir la matrice de la figure 6.3).

Enfin, lorsque nous utilisons ADN-LeCombi sur le jeu de Stormo ou sur celui de Berg (puisque nous lisons les séquences dans les deux sens, le résultat obtenu est le même, quel que soit l'ensemble sélectionné), nous obtenons, en à peu près le double du temps, deux modèles de longueur 18, identiques en termes d'ensembles d'occurrences. Ils repèrent le site correctement sur 15 des 17 séquences, le manquent complètement sur *cat* et le placent ailleurs sur *gal*. Si nous nous intéressons au 'sens' de lecture maintenant, certains sites sont des occurrences dans les deux sens, ce sont les sites associés à *crp*, *col*E, *lac*, *mal*E et *tna*A. Les autres peuvent être groupés en deux ensembles en termes de leur sens de lecture. Le premier ensemble est constitué des occurrences sur les séquences de *bgl*R, *cya*, *mal*T, *pbr*P4 et *uxu*A. Le second est formé des occurrences sur les séquences de *ara*, *deo*P2, *ilv*B, *mal*K et *omp*A. Observons qu'à l'exception de *mal*K et de *ilv*B, les éléments de ce second ensemble correspondent à des sites qui devraient, selon le critère établi par Combrugghe et Berg, être 'lus' sur le brin opposé de l'ADN par rapport au jeu de Stormo. Selon eux, cela s'appliquerait également aux sites de *mal*E et *crp*. Cependant par rapport aux modèles trouvés ici, le sens de lecture de ces sites nous apparaît plutôt indifférent. Il est par ailleurs intéressant de noter qu'aucun des modèles trouvés ne comporte les deux motifs $TGTGA$ et $TCACA$ exactement. En effet, les deux modèles sont $TA\underline{NGAGA}NNNNNN\underline{TCNCA}$ et $\underline{TGNGA}NNNNNN\underline{TCACN}TA$.

6.2.3.3 Discussion

Les résultats obtenus avec DNA-LeCombi semblent donner une indication quant au brin sur lequel se situerait effectivement le site de fixation de la CRP, mais il faut être prudent quant à leur interprétation. Il est en effet délicat de conclure hâtivement en ignorant l'influence du choix des paramètres sur ces résultats. Ainsi, lorsqu'aucune substitution n'est autorisée (mais toujours 8 jokers) et qu'en conséquence, la longueur des modèles trouvés diminue même pour des quorums bas (par exemple pour $q = 70\%$, cette longueur est de 12), l'orientation des sites suit celle du jeu de Berg simplement parce que le premier motif (le seul 'vu' par l'algorithme dans ce cas) y est toujours très fort. Par contre, lorsque plus de substitutions sont autorisées, et que la longueur maximale des modèles repérés s'allonge, l'orientation peut changer.

Il est important d'observer également que la valeur de certains paramètres a dû être fixée de manière empirique, en particulier celle du quorum. Dans le cas précis de cet exemple, il aurait été plus logique d'établir ce dernier à 100%. C'est ce que nous avions fait initialement. Les résultats obtenus alors sont cependant moins bons car les deux sites associés au gène *cat* sont plus dégénérés et donc difficilement identifiables en même temps que les autres. Exiger leur présence en tant qu'occurrences des modèles construits suffit à diminuer sensiblement la qualité du résultat. Celui-ci dépend ainsi beaucoup d'une certaine homogénéité des données, ou exige que le quorum varie suivant les besoins. Le cas du gène *cat* est intéressant pour une autre raison. Il a été suggéré [Berg and von Hippel, 1988] [Schneider *et al.*, 1986] que deux facteurs peuvent contribuer à cette plus grande dégénérescence apparente des sites de *cat*. Le premier est que ceux-ci sont asymétriques (seul le motif $TGTGA$ est conservé). Le second facteur est que l'espacement entre les deux motifs $TGTGA$ et $TCACA$ serait en réalité de 9 bases au lieu de 6 pour le site primaire et de 7 pour le site secondaire [Berg and von Hippel, 1988]. Nous avons donc voulu relancer nos algorithmes à la recherche de ces sites en permettant cette fois un certain nombre de substitutions <u>et</u> de trous. La difficulté est que, plus nous autorisons de trous entre un modèle et ses occurrences, plus ces occurrences sont dégénérées. Nous en avons ainsi obtenu plusieurs, localisées aux mêmes positions sur chaque séquence mais de longueurs différentes, et nous n'avions pas les moyens de déterminer quelle occurrence était la bonne. Ce problème des trous est en fait plus général que ce qui est observé sur ce simple jeu d'essai. L'autorisation de trous dans un alignement ou un bloc est un problème qui n'a pas, à notre sens, été traité de manière vraiment satisfaisante jusqu'à présent, nous reviendrons sur ce point dans le dernier chapitre. Dans cet exemple en tout cas, comme dans celui des promoteurs dans la section précédente, il est clair qu'il serait plus intéressant de travailler avec le concept d'une 'suite de jokers de longueur variable' localisée entre des motifs sans trou.

6.2.4 Les éléments de régulation par le fer

6.2.4.1 Introduction

Une des fonctions exercées par les protéines consiste à transporter des petites molécules ou des ions à l'intérieur de l'organisme, ainsi par exemple les protéines de la famille des globines avec lesquelles nous travaillerons plus loin interviennent dans le transport de l'oxygène. Le fer est un autre élément essentiel à la vie qui, chez les organismes supérieurs, est transporté par la transferrine dans le plasma du sang et doit former un complexe avec la ferritine afin de pouvoir être stocké dans le foie.

Ces deux protéines, transferrine et ferritine, subissent une régulation de leur expression qui dépend de la quantité de fer présent à un moment donné à l'intérieur des cellules. Cette régulation est traductionnelle, elle a donc lieu au moment de la traduction de l'ARN messager en protéines et non au moment de la transcription.

Nous avons fait observer dans le chapitre 2 que tout le matériel qui est transcrit n'est pas ensuite traduit en protéine, la partie non traduite concerne en particulier la région initiale 5' et la région terminale 3', ainsi que les introns (dans le cas des eucaryotes). Les régions 5' et 3' sont celles qui interviennent dans le processus de régulation de la biosynthèse des protéines au niveau de la traduction. Celui de la ferritine et de la transferrine est ainsi controlé par deux éléments situés dans ces régions (5' pour la ferritine et 3' pour la transferrine). Ces éléments ont été appelés des 'éléments de régulation par le fer'. Ils sont également dénotés IRE de l'anglais 'Iron Responsive Elements'. L'IRE responsable de la régulation de la ferritine est constitué d'une séquence comportant environ 28 nucléotides qui se replie en forme d'épingle à cheveux avec une excroissance vers le milieu de la partie amont de la tige (voir la figure 6.12). La boucle de l'épingle est absolument conservée en termes de séquence, de même que le nucléotide, une Cytosine, formant l'excroissance et qui se situe toujours 5 positions avant le premier nucléotide de la boucle. Le reste de la séquence n'est réellement conservé qu'en tant que structure, du moins entre familles différentes de gènes. La régulation de la transferrine est, quant à elle, réalisée par 5 séquences qui ressemblent à l'IRE de la ferritine.

Ces éléments de régulation sont donc en général recherchés par des programmes permettant de décrire des structures secondaires d'ARN à travers une grammaire qui inclut à la fois des motifs structuraux (pour la tige) et lexicaux (pour la boucle et l'excroissance) [Billoud *et al.*, 1996] [Dandekar *et al.*, 1991] [Dandekar and Hentze, 1995] [Gautheret *et al.*, 1990]. Un exemple d'une telle description dans le cas des IRE est donnée dans la figure 6.12 [Dandekar *et al.*, 1991]. C'est celle employée par le programme de Billoud appelé Palingol.

Notre but avec cet exemple a été au départ de voir dans quelle mesure il n'y avait pas, y compris dans des gènes de familles différentes, une conservation en séquence au niveau de la tige de l'IRE ainsi qu'au niveau des régions situées en amont ou en aval de l'IRE.

Nous avons pour cela travaillé avec deux jeux d'essais parmi les 4 utilisés par Billoud [Billoud *et al.*, 1996]. Les séquences composant ces deux jeux ont été extraites de la banque EMBL version 43 par la base de donnée HOVERGEN [Duret *et al.*, 1994]. Ces deux jeux sont composés de :

- 48 séquences contenant le mot 'ferritin' parmi leurs mots-clés associés à la séquence dans la banque, dont 22 seulement contiennent un vrai IRE (lot A);
- 33 séquences correspondant à des régions 5' et qui comprennent toutes des IRE (certains associés à des ferritines) (lot B).

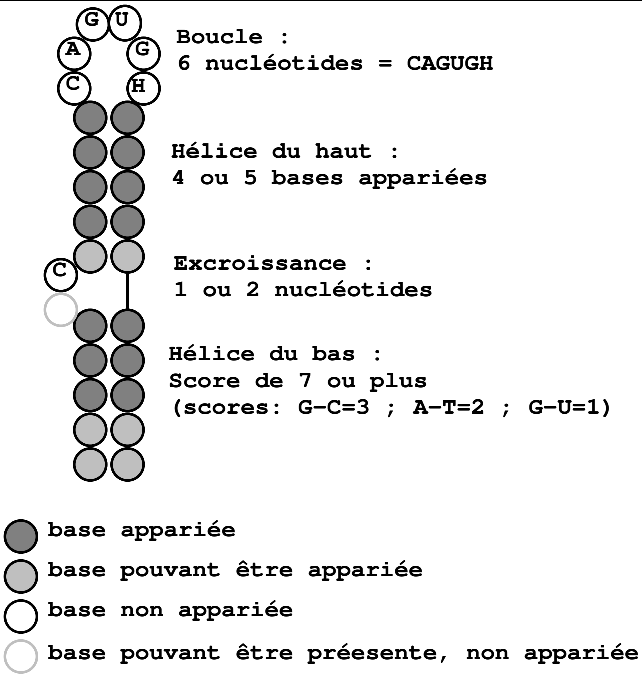

Figure 6.12: Description d'un IRE par Dandekar.

Observons que le programme de Billoud [Billoud *et al.*, 1996] réussit à identifier 19 des 22 séquences du premier jeu qui contiennent des vrais IRE et localise les 33 séquences du second jeu parmi un ensemble plus grand de régions 5' comprennant 17642 séquences en tout. Les trois 'bonnes' séquences du premier jeu qui sont manquées le sont, soit parce qu'elles contiennent une excroissance ou une substitution dans le pied de la tige qui ne sont pas permises dans la description utilisée par Palingol, soit parce que cette partie inférieure de la tige n'est pas suffisamment longue (séquences tronquées).

6.2.4.2 Résultats

Nous avons recherché les modèles les plus longs présents dans les séquences de chacun des deux lots avec HMoivre, Combi (dans sa version originale et la variante qui travaille avec une pondération par cardinalité d'ensemble - section 5.4.2.2.6.2) et LeCombi. Nous avons également fait varier le quorum entre 20% et 70% avec un pas de 10%. Les résultats les plus intéressants ont été obtenus avec Combi (ils sont sensiblement les mêmes avec l'algorithme original ou la variante) et ce sont ceux que nous présentons ici (19 secondes sur une Sparc Station 20).

Nous avons utilisé plusieurs couvertures; une de celles qui se sont montrées les plus 'informatives' sur cet exemple est la suivante :

$$CCP = \left\{ \begin{array}{l} (\{X\}, \infty)\ \forall X \in \Sigma = \{A, C, G, T\}, \\ (\{A, C, G, T\}, 0), \\ (S, 1)\ \textit{pour tous les}\ (10)\ \textit{autres ensembles} \end{array} \right\}.$$

De plus, nous n'avons permis que 4 jokers dans la première étape de l'algorithme (variante de la section 5.4.2.2.6.1), ce qui signifie que le nombre total d'ensembles 'spéciaux' (les non singletons) autorisés dans les modèles était également limité à 4.

Avec un quorum de 40%, les 21 occurrences distinctes des sept modèles de longueur k_{max} égale à 27 identifiés sur le lot A correspondent toutes à l'IRE. Certains de ces modèles peuvent en fait être regroupés en un seul car ils ont exactement les mêmes occurrences et diffèrent uniquement par leurs noms. Les trois modèles obtenus après cette opération sont alors les suivants :

Modèle A1 $C[CT]TG\underline{C}TTCAA\underline{CAGTG}[CT]TTG[AG]ACGGAA[CG]$
(observons que le second ensemble $[CT]$ apparaît comme $[CGT]$ ou $[ACT]$ dans le modèle puisque l'ensemble $[CT]$ n'y est autorisé qu'une fois);

Modèle A2 $C[CT]TG\underline{C}TTCA[AG]\underline{CAGTG}[CT]TTG[AG]ACGGAAC$
(à nouveau ici le second ensemble $[CT]$ apparaît comme $[CGT]$ ou $[ACT]$ dans le modèle et le second ensemble $[AG]$ comme $[AGT]$ ou $[ACG]$);

Modèle A3 $C[CT]TG\underline{C}[GT]TCAA\underline{CAGTG}[CT]TTG[AG]ACGGAAC$
(le second $[CT]$ apparaissant comme $[CGT]$ ou $[ACT]$).

Nous avons souligné les 5 nucléotides de la boucle et le C de l'excroissance qui sont absolument conservés dans tous les IRE.

Ces modèles sont très similaires et peuvent être résumés dans le modèle suivant :

$$C[CT]TG\underline{C}[GT]TCA[AG]\underline{CAGTG}[CT]TTG[AG]ACGGAA[CG].$$

Parmi les ensembles d'occurrences de ces trois modèles (chaque ensemble en comporte 19), 18 occurrences sont communes et représentent 16 IRE identifiés par Palingol ('noms' mnémoniques dans GenBank et positions IRE : HSFERG1 (207-239), HSFERP1 (118-150), HSFERP2 (118-150), HSFERRITH (29-61), HSFHC122 (30-62), MMFERHC (229-261), MMFERHG (967-999), MMFERRH (967-999), OCFERL5 (31-61), RCFERH (24-54), RNFERA1 (205-237), RNFERL1 (400-430), RNFERUTR (20-50), SSFE (12-44), XLFERHSU (7-37) et XLFERRIT (166-198)) ainsi que 2 des 3 IRE ratés par Palingol ('noms' mnémoniques et positions IRE : MMFERLSUB (1198-1213) et MMFERLA (6-21)). Chaque modèle identifie un IRE de plus. Le modèle 1 identifie l'IRE HSAFLP1 (2-28), le modèle 2 l'IRE HSFERHX (949-981) et enfin le modèle 3 l'IRE GGFERH (1306-1336).

Sur le lot B et avec un quorum également de 40%, Combi avec la même couverture CCP que pour le lot A repère initialement 6 modèles de longueur 28 qui représentent deux ensembles d'occurrences distincts et identifient 15 des 33 IRE présents dans le jeu. Ces modèles sont les suivants (après regroupement) (7 secondes sur une Sparc Station 20) :

Modèle B1 $TC[CT]TG\underline{C}TTCA[AG]\underline{CAGTG}[CT]TTG[AG]ACGGAAC$;

Modèle B2 $TC[CT]TG\underline{C}[GT]TCAA\underline{CAGTG}[CT]TTG[AG]ACGGAAC$.

Ils correspondent donc aux modèles 2 et 3 trouvés avec le lot A auxquels un singleton de plus à été ajouté en première position. Sauf pour une des séquences d'origine inconnue (de 'nom' S64727 (482-552)), les IRE correspondants sont tous associés à des ferritines.

La stratégie que nous avons adoptée à partir de ce point sur ce même lot B est, une fois de plus, empirique, mais elle nous a permis d'obtenir des résultats supplémentaires qui peuvent être intéressants. Elle a consisté à retirer du lot B les 15 séquences où le modèle B1 ou B2 est présent, et à relancer Combi à la recherche du plus long modèle présent dans au moins 40% des 18 séquences restantes. Cette seconde recherche a localisé deux modèles de longueur 37 représentant un seul ensemble de 9 occurrences (351 secondes sur une Sparc Station 20). Le modèle après regroupement s'écrit :

Groupe	Modèle	Prés. seq.
1	$TC[CT]TG\underline{C}TTCA[AG]\underline{CAGTG}[CT]TTG[AG]ACGGAAC$	14
	$TC[CT]TG\underline{C}[GT]TCAA\underline{CAGTG}[CT]TTG[AG]ACGGAAC$	14
2	$GAAGGCGTGGCTCCCTC\underline{C}CGGGC\underline{CAGTG}AGCC[CT][GT]G[CG][CG]$	9
3	$C\underline{C}CAGG\underline{CAGTG}CC[CT]TG[GT]G[CT][AG]A$	3
4	$GT\underline{C}[GT][CG]TG[CT]\underline{CAGTG}TGTG[GT]$	2
5	$TCC[AT]\underline{CAGTG}C[AT]GGG[AC][AT]$	2

Tableau 6.4: Modèles trouvés avec Combi sur le lot B de séquences (voir texte pour la couverture CCP et la stratégie suivie — seuls les groupes de modèles ayant au moins deux occurrences sont indiqués).

$$GAAGGCGTGGCTCCCTC\underline{C}CGGGC\underline{CAGTG}AGCC[CT][GT]G[CG][CG]$$

et couvre une partie de l'IRE (environ 3 à 4 bases du côté 3' sont absentes) plus 9 à 12 bases en amont de l'IRE. Il serait intéressant de savoir si cette conservation apparente des séquences avant l'IRE a un lien avec l'activité exercée par ce dernier. Observons aussi que ce modèle est très différent des modèles B1 et B2 caractérisant les ferritines. Les 'noms' mnémoniques des séquences concernées et leurs positions dans le contig d'où elles ont été extraites sont les suivants : HSEP3A1 (1-62), HSEP3A2 (1-62), HSEP3B (1-62), HSEP3C (1-62), HSEP3D (1-62), HSEP3E (1-62), HSEP3F (1-62), HUMPEIR (1-68) et S69200 (1-68). Il s'agit ici essentiellement de gènes codant pour des récepteurs de prostaglandines. Nous avons ensuite renouvelé l'opération, c'est-à-dire que nous avons retiré ces 9 séquences du lot B et relancé Combi dans les mêmes conditions qu'auparavant sur les 9 séquences restantes. Cependant, comme nous ne pouvons plus construire aucun modèle de longueur supérieure à 12, nous avons décidé de baisser le quorum à 30%. Dans ce cas, deux plus longs modèles sont trouvés (en 12 secondes sur une Sparc Station 20), représentant un seul ensemble de 3 occurrences longues de 21 bases et couvrant une partie de l'IRE. Le modèle regroupé s'écrit $C\underline{C}CAGG\underline{CAGTG}CC[CT]TG[GT]G[CT][AG]A$ et est présent dans les séquences HSNMYC (782-853), MMNMYC (1-68) et RATAE3A (3380-3453). Les occurrences du modèle sont identiques sur les deux premières séquences, la troisième sur RATAE3A diffère dans les positions où un ensemble de taille 2 apparaît dans le modèle. Encore une fois, ce modèle ne ressemble à aucun des 2 précédents et il serait intéressant de savoir si le regroupement qu'il opère est fortuit ou non. Parmi les 6 séquences restantes enfin, les IRE ne peuvent plus être regroupés que par des modèles de longueur 18 (2 IRE — modèle $GT\underline{C}[GT][CG]TG[CT]\underline{CAGTG}TGTG[GT]$) et 16 (2 IRE — modèle $TCC[AT]\underline{CAGTG}C[AT]GGG[AC][AT]$). Les IRE des deux dernières séquences sont complètement différents de tous les autres. Tous les modèles trouvés sur le lot B en suivant la stratégie décrite ici se trouvent résumés dans le tableau 6.4.

6.2.4.3 Discussion

Nous ignorons si les résultats indiqués dans la section précédente et la 'classification' primaire qu'ils fournissent des IRE en région 5', ont une valeur biologique quelconque. De notre point de vue cependant, cet exemple est important car seule l'utilisation d'une couverture combinatoire pondérée permet de repérer des caractéristiques susceptibles d'être intéressantes, ce qui n'est pas le cas des autres techniques dont nous disposons (distance de Hamming entre modèles et mots, couverture pondérée restreinte avec uniquement le joker en plus de l'alphabet des nucléotides). Cependant, nous pouvons également voir à quel point l'information contenue dans les modèles est étroitement liée à la définition de ressemblance adoptée, et au choix des valeurs à attribuer aux divers paramètres. Considérons en effet les trois plus longs modèles identifiés sur le lot A. Ils se ressemblent beaucoup et, ainsi que nous l'avons vu, peuvent être résumés en un seul :

$$C[CT]TG\underline{C}[GT]TCA[AG]\underline{CAGTG}[CT]TTG[AG]ACGGAA[CG].$$

En faisant cela cependant, nous gommons le fait que, par exemple, aucun IRE correspondant à $CCTGCTTCA\underline{G}CAGTGCTTGAACGGAA\underline{G}$ n'est observé parce que ce mot réalise un mélange des modèles A1 et A2 (le premier G souligné provenant du modèle A2, le second du modèle A1). Mais le modèle A1 lui-même a perdu l'information qu'un seul IRE comporte un G en $12^{ième}$ position après le G de la boucle $CAGTG$, et que cet IRE avait une Thymine aux deux endroits où le modèle est composé de l'ensemble $[CT]$ et une Guanine là où il présente un $[AG]$. Est-ce que cette information devait être préservée? La question de la pertinence de cette information ne se rapporte pas seulement aux objets qui sont localisés mais également à leur caractérisation. Cette dernière est essentielle, en particulier si nous voulons pouvoir nous en servir ultérieurement pour identifier d'autres objets du même type. Enfin, une question plus générale concerne les relations de dépendance entre les positions d'une séquence. Pour l'instant nous n'identifions ces relations que par hasard avec nos modèles. Y aurait-il un moyen de formaliser ces relations directement dans les définitions de ressemblance et dans l'étape de recherche des modèles?

Une autre observation à propos de la caractérisation des objets que nous recherchons, dans le cas précis de cet exemple, porte sur une comparaison de celle utilisée par des programmes comme Palingol avec celle que nous obtenons (la liste des modèles trouvés). Rappelons que Palingol n'est pas un programme permettant de déduire une caractérisation à partir d'un ensemble d'exemples (ce que nos algorithmes font d'une certaine façon), mais est un programme qui part d'une description d'un objet et qui recherche les instances vérifiant cette description dans un ensemble de vrais et faux exemples. Il existe cependant une autre différence importante entre notre approche et celle de Palingol qui réside dans le fait que la description que ce dernier utilise est plus riche. Elle fait ainsi intervenir des concepts plus complexes que la simple succession des symboles et peut y compris décrire des objets non encore observés. Ces concepts peuvent également avoir été préalablement extraits d'exemples comme le sont nos modèles. Même dans ce second cas cependant, l'extraction de la description utilisée par Palingol a porté sur des critères purement structuraux et sont différents de ceux qui nous permettent de construire les modèles. Pourrait-on incorporer de tels critères dans nos définitions de ressemblance? Par ailleurs, est-ce que les séquences en tant que telles, et la technique utilisée dans cette illustration en particulier, peuvent apporter quelque chose d'autre?

Enfin, nous nous sommes limités ici à rechercher les modèles les plus longs qui se trouvent présents dans au moins un certain pourcentage des séquences de nos jeux d'essai. Afin de déterminer s'il existe d'autres régions conservées, en particulier en aval de l'IRE, semblables à celle identifiée en amont de quelques-unes des séquences du lot B, il faudrait examiner les modèles ayant une longueur inférieure à la longueur maximale. Une autre façon de procéder serait d'appliquer récursivement l'algorithme non sur des ensembles de plus en plus réduits de séquences comme nous l'avons fait pour ce même lot B, mais sur le lot de départ après avoir masqué les occurrences des modèles trouvés à l'étape antérieure. Cette stratégie générale est celle utilisée par plusieurs heuristiques qui construisent un alignement multiple à partir d'une recherche initiale de blocs de similarité, en particulier [Landraud *et al.*, 1989] [Viari, 1995]. Elle est importante mais présente plusieurs difficultés. La plus importante est sans doute de déterminer sur quel critère appuyer le choix des blocs successivement sélectionnés. Est-ce sur leur longueur ou sur une valeur statistique de leur contenu, et dans ce second cas, comment calculer cette valeur? Ce problème devient d'autant plus aigu que la ressemblance entre séquences, ou entre mots dans les séquences, devient faible. C'est ce que nous allons voir dès la section suivante avec les illustrations portant sur des protéines.

6.3 Protéines

6.3.1 Introduction

L'analyse de séquences protéiques ressemble à celle des séquences d'ADN ou d'ARN et est en même temps très différente sous certains aspects. Elle lui ressemble car le problème central reste la recherche des mots communs dans un ensemble de séquences. Ces mots peuvent être associés aux sites actifs de la protéine et donc à son activité, ou correspondre à des éléments structuraux conservés (et qui, en général, seront alors plus dégénérés au niveau de la séquence). Souvent un site actif est préservé dans sa forme mais une forme conservée peut ne pas être caractéristique d'un site. L'ensemble des formes conservées qui possèdent essentiellement les mêmes positions relatives parmi tous les membres alignés d'une famille est appelé le noyau structural de la protéine (en anglais, le 'core') [Gerstein and Altman, 1995]. Ce noyau représente ainsi une espèce d'armature plus ou moins rigide de la macromolécule.

La première différence entre les deux types d'analyse (ADN/ARN et protéines) est que les définitions de ressemblance s'appliquant aux protéines ne sont pas les mêmes que celles utilisées pour comparer des séquences de nucléotides. En outre, il y a généralement, pour un même ensemble de données, plus de mots 'intéressants' à repérer dans le cas des protéines (celles-ci peuvent comporter de nombreux sites ou motifs structuraux) et ces mots sont en général encore moins faciles à localiser et à distinguer du bruit. Il y a plusieurs raisons à cela. La première est que l'alphabet des protéines est plus grand et que les relations de similarité entre les symboles de cet alphabet sont plus complexes et souvent ambiguës. La seconde est qu'un mot 'intéressant' est en général composé de peu de positions réellement importantes pour la forme ou l'activité de la protéine et qu'en revanche, ce mot possède plusieurs positions non-spécifiques ou hautement variables. Plus souvent encore que dans le cas des séquences nucléiques, nous allons donc devoir faire appel à des techniques de filtrage des premiers résultats obtenus par nos algorithmes.

En fait, dans la plupart des cas, nos algorithmes vont maintenant représenter une première étape d'exploration de l'espace des exemples que nous traitons. Elle sera suivie ensuite par une seconde étape d'évaluation des éléments ramassés lors de la première phase. Observons que cette première étape <u>demeure combinatoire</u>. Ainsi que nous l'avons indiqué dans les chapitres 4 et 5, certaines méthodes, notamment MACAW (section 4.2.2.1.2.2) et l'algorithme de Neuwald (section 5.4.2.2.6.4) mélangent les deux phases : évaluation et recherche des mots communs. Nous pensons qu'il est important de maintenir ces deux étapes séparées. En particulier, la recherche initiale des modèles doit être aussi large et complète que possible, et les propriétés que ces modèles vérifient doivent demeurer claires et précises. Le plus souvent, cela signifie que nous allons conserver plus d'information que nécessaire. La recherche initiale n'est donc pas toujours très sélective, l'essentiel est qu'elle soit sensible, c'est-à-dire, qu'elle 'rate' le moins possible de choses. Dans la seconde phase, il ne nous reste plus alors qu'à déterminer lesquels parmi les modèles obtenus dans la première étape ont une chance d'être réellement pertinents.

Cette seconde étape peut être ajoutée à n'importe lequel de nos algorithmes, à savoir LeCombi, LePoivre ou Klast (dans la suite, nous noterons X cet algorithme). Elle constitue une extension algorithmique qui demeure à ce jour préliminaire aussi bien au niveau de l'implémentation que des résultats, mais que nous utilisons déjà sous une forme simple dans trois des quatre illustrations présentées dans la suite de ce chapitre. Nous esquissons les idées principales de cette extension dans la section suivante.

6.3.2 Esquisse de la nouvelle extension algorithmique

Cette seconde phase que nous introduisons dans nos algorithmes consiste donc à évaluer la pertinence des modèles trouvés dans la première phase. L'idée sur laquelle nous nous appuyons pour réaliser cela est que, si un modèle localisé par X est effectivement en relation avec un caractère conservé en termes de structure ou de fonction parmi un ensemble de protéines, alors un mot correspondant à ce caractère doit se trouver présent de manière approchée dans toutes ou presque toutes les protéines appartenant à la même famille, ayant la même structure locale, ou exerçant la même activité. Le principe de la méthode se résume donc à repérer avec l'algorithme X les mots communs à un premier ensemble de protéines (i.e. les occurrences d'un même modèle), à associer à chacun de ces ensembles de mots une matrice (qui sera précisée plus loin), puis à calculer le score obtenu par cette matrice lorsqu'elle est successivement alignée avec toutes les positions possibles d'un second ensemble de séquences fonctionnellement proches de celles du premier ensemble. Ce second ensemble peut être réduit à un seul élément. Il peut également être identique au premier. Nous appelons le premier ensemble celui des séquences exemples et le second celui des séquences tests.

Plusieurs méthodes existent pour construire une matrice à partir d'un bloc de mots — on appelle une telle matrice un profil — puis pour rechercher la meilleure occurrence de ce profil sur une ou plusieurs autres séquences. Les profils sont en général établis à partir de mots alignés sans trou, mais leurs occurrences peuvent éventuellement en comporter. La technique de Gribskov [Gribskov *et al.*, 1987] [Gribskov *et al.*, 1988] [Gribskov *et al.*, 1990] autorise des trous dans la phase de recherche (en utilisant la programmation dynamique). Elle peut même partir d'un bloc initial avec trous. Néanmoins, ce qui nous intéresse ici est moins de localiser les occurrences d'un profil que de lui assigner une valeur statistique permettant d'en mesurer la pertinence. Or aucune théorie n'est capable à l'heure actuelle de prendre en compte les trous de manière satisfaisante.

La technique que nous utilisons alors est celle de Waterman, décrite dans [Goldstein and Waterman, 1994] et pour laquelle chaque entrée (i, j) du profil $\mathcal{P}$ obtenu à partir d'un bloc d'occurrences d'un modèle est définie pour tout acide aminé i et position j par :

$$\mathcal{P}(i, j) = \sum_{a=1}^{20} f(a, j) \times \mathcal{M}(i, a) \tag{6.1}$$

avec $f(a, j)$ égal à la fréquence observée de l'acide aminé a à la position j du bloc et $\mathcal{M}$ une matrice de similarité (par exemple, $\mathcal{M}$ peut être la matrice PAM250 de la figure 3.3, ou une des matrices BLOSUM de Henikoff — en général BLOSUM62).

La localisation du meilleur emplacement du profil $\mathcal{P}$ — notons cette position x — sur une séquence s de longueur n, est alors celle ayant le meilleur score S_0 donné par :

$$S_0 = \sum_{j=1}^{k} \mathcal{P}(s_{x+j-1}, j) \tag{6.2}$$

où k désigne la largeur du profil (qui est aussi, dans notre cas, la longueur du bloc et donc celle du modèle). Enfin, la théorie développée par Goldstein *et al.* [Goldstein and Waterman, 1994] permet d'évaluer la signification statistique de ce score en estimant la probabilité qu'un score supérieur ou égal à S_0 soit observé sur une séquence aléatoire de même longueur et même composition en acides aminés (p-value). La valeur de cette p-value — notée plus simplement p — est égale à :

$$p = 1 - \exp^{-\exp^{-a.(w-c)}} \tag{6.3}$$

avec :

$$a = \sqrt{2.\log(n-k+1)}$$
$$c = a - \frac{\log(\log(n-k+1) + \log(4\pi))}{2a}$$
$$w = \frac{(S_0 - \bar{S})}{\sigma_S}$$
$$\bar{S} = \sum_{j=1}^{k} \sum_{i=1}^{20} f(i) \times \mathcal{P}(i,j)$$
$$\sigma^2_{S_0} = \sum_{j=1}^{k} \sum_{i=1}^{20} f(i) \times \mathcal{P}^2(i,j)$$

et $f(i)$ égal à la fréquence de l'acide aminé i, soit dans la séquence s, soit dans l'ensemble des séquences tests, soit encore dans toute une banque (nous avons opté pour ce dernier choix).

Ceci étant établi, voyons maintenant comment fonctionne notre algorithme en deux étapes. La première, ainsi que nous l'avons dit, consiste à appliquer l'algorithme X (LeCombi ou Klast) au premier ensemble de séquences appelées exemples. Cette application produit un certain nombre de modèles, de longueur fixe ou maximale, présents dans au moins q séquences. Dans la seconde étape, chacun de ces modèles est ensuite considéré à tour de rôle. Pour chacun d'eux, un profil est construit à partir de ses occurrences suivant la formule 6.1 ci-dessus, puis le meilleur emplacement de ce profil est recherché sur chacune des séquences d'un ensemble test. La moyenne des scores obtenus par l'ensemble des meilleurs placements va représenter le score du modèle. Enfin, les modèles sont classés par ordre de p-values croissantes et ceux considérés pertinents sont tous ceux pour lesquels cette p-value est inférieure à une valeur seuil (généralement 10^{-3} ou 10^{-5}).

Lorsqu'un modèle est présent plus d'une fois sur une des séquences, seule est conservée l'occurrence de score maximal. Dans le cas où le quorum de la première étape est strictement inférieur à 100%, un modèle peut ne pas avoir une occurrence sur toutes les séquences. Lors de la seconde étape de l'algorithme, une pseudo-occurrence de ce modèle est alors recherchée sur chacune des séquences s où il n'est pas présent. La position de cette pseudo-occurrence correspond à celle du meilleur emplacement du profil sur s.

Enfin, si un modèle M possède au moins une occurrence qui chevauche l'occurrence d'un autre modèle M', alors seul le modèle ayant la plus petite p-value est conservé.

Nous montrons maintenant l'application de ce nouvel algorithme, que nous avons choisi d'appeler CoSamp-X (de 'Combinatorial Sampling' à partir de l'algorithme X), dans le cadre de trois illustrations représentant des cas considérés difficiles en analyse de séquences : celui des cytochromes P450, celui des globines et enfin celui des protéines contenant un motif structural HTH ('Helix-Turn-Helix'). Dans les trois cas, nous comparons le comportement de CoSamp à celui de CLUSTAL W (section 4.1.3.3.3) et de la méthode d'échantillonnage de Gibbs de Lawrence (section 4.2.2.2) intégrée dans MACAW (section 4.2.2.1.2.2). La première approche, celle de CLUSTAL, est globale et est donc très différente de la notre ou de celle de MACAW mais la comparaison nous a néanmoins semblé intéressante.

6.3.3 Les P450

6.3.3.1 Introduction

Les cytochromes P450 (appelés de manière abrégée les P450) sont des protéines qui catalysent l'oxydation d'une grande variété de substrats hydrophobes et qui sont ainsi impliquées, entre autres, dans le métabolisme des acides gras, des stéroides et des vitamines et dans l'élimination de substances exogènes toxiques. Ces protéines possèdent également une action négative puisqu'elles peuvent exercer un rôle dans l'activation pro-carcinogène [Hasemann *et al.*, 1995] [Stryer, 1981].

Comme pour beaucoup d'autres protéines intervenant dans les processus d'oxydation, les P450 contiennent un groupe hème, qui est un groupement prosthétique composé d'une partie organique et d'un atome de fer qui se lie à l'un des atomes de la chaîne polypeptidique. Dans le cas des P450, contrairement à la plupart des autres protéines qui possèdent un hème, cet atome est le soufre d'une cystéine (C) (les P450 sont ainsi appelés des protéines hème-thiolate). L'acide aminé auquel se lie le groupe hème s'appelle le ligand proximal du groupe.

Les P450 sont regroupés en familles et en sous-familles. Une famille est constituée des P450 présentant une similarité de séquence supérieure à 40%. On définit une sous-famille lorsque cette identité atteint 55% [Nelson *et al.*, 1993]. À l'heure actuelle, on compte 59 familles et 105 sous-familles. Les P450 avec lesquels nous allons travaillé ici sont des éléments isolés n'appartennant à aucune famille précise (ils ont moins de 20% de résidus identiques lorsque alignés optimalement par paires). En termes de séquences, ces protéines se trouvent donc dans ce que Doolittle a désigné la 'twilight zone' [Doolittle, 1981] [Doolittle, 1986], c'est-à-dire qu'elles forment un groupe de protéines probablement homologues et exerçant des activités similaires qui toutefois possèdent moins de 25% de résidus identiques.

Jusque vers la fin 1995, seules trois P450, tous bactériens, avaient pu être cristallisés. Ce sont les P450$_{Cam}$, P450$_{Terp}$ et P450$_{BM_3}$. Ce nombre est actuellement de quatre, le quatrième étant le P450$_{Eryf}$ également bactérien. Des analyses 'manuelle' [Hasemann *et al.*, 1995] et automatique [Jean *et al.*, 1996] des structures des trois premiers P450 cristallisés ont révélé un certain nombre d'éléments structuraux, c'est-à-dire de motifs, relativement bien conservés, alors que les mots lexicaux correspondant à ces motifs sont très dégénérés. Ces éléments structuraux conservés peuvent être groupés de manière parfois différente selon la méthode utilisée pour les repérer. Nous suivons ici la classification donnée dans [Jean *et al.*, 1996] et obtenue avec l'algorithme KMRC/S adapté aux structures (voir section 4.2.3.5) ainsi que celle de [Hasemann *et al.*, 1995] là où il y a divergence. Treize motifs sont ainsi identifiés, dont deux sont en fait composés de deux blocs distincts. La figure 6.13 montre un alignement des trois premiers P450 cristallisés et l'emplacement des blocs structuraux correspondant aux motifs structuraux trouvés sur les trois premiers par KMRC/S. La figure 6.14 montre ces motifs structuraux superposés et la figure 6.15 montre leur arrangement dans l'espace pour le P450$_{Eryf}$. Enfin, le tableau 6.5 fournit la liste des motifs structuraux et les positions qu'ils recouvrent pour le P450$_{Cam}$, le P450$_{Terp}$, le P450$_{BM_3}$ et le P450$_{Eryf}$. Là où il y a désaccord entre P. Jean et Hasemann à propos de ces positions (hélices F et G), nous indiquons les deux. Observons que ces blocs structuraux correspondent souvent à des assemblages de structures secondaires, et l'alignement des séquences montre bien que celles-ci sont très peu conservées entre les P450$_{Cam}$, P450$_{Terp}$ et P450$_{BM_3}$. En effet, seules 26 positions, parmi les 255 que couvre l'ensemble des blocs, présentent le même acide aminé dans les trois P450 (parmi ces 26, 8 se situent dans les blocs 12A et 12B de la région de l'hème, y compris la cystéine où se fixe l'atome de fer). Ces éléments structuraux eux-mêmes ne possèdent pas tous le même degré de conservation. Les hélices G et surtout F sont plus variables. Jean *et al.* ont ainsi décidé de ne pas considérer la région

Bloc	Él^t. de structure	Position sur P450$_{BM3}$	Position sur P450$_{Cam}$	Position sur P450$_{Terp}$	Position sur P450$_{Eryf}$	Long. du bloc
1	β_{1-2} + B	48-62	61-75	50-64	38-52	15
2A	C	94-111	106-123	108-125	96-113	18
2B	D	112-129	124-141	126-143	114-131	18
3	E	144-156	154-166	157-169	144-156	13
4	F	171-185 (J) 174-184 (H)	171-185 (J) 174-184 (H)	174-188 (J) 180-190 (H)	(indéterminé)	15 (J) 11 (H)
5	G	205-225 (J) 201-223 (H)	193-213 (J) 193-215 (H)	212-232 (J) 208-230 (H)	185-205	21 (J) 23 (H)
6	H	228-239	214-225	233-244	206-217	12
7	β_{5-2} + I + J	248-292	232-276	251-295	225-269	45
8	K + β_{6-1}	312-328	279-295	298-314	272-288	17
9	β_{1-4} + β_{2_1}	330-343	296-309	316-329	290-303	14
10	β_{2-2} + β_{1-3} + K'	345-367	310-332	330-352	304-326	23
11	Meander	370-377	334-341	354-361	328-335	8
12A	Cys-pocket	390-398	347-355	367-375	341-349	9
12B	L	399-416	356-373	376-393	350-366	18
13	β_{4-2} + β_{3-2}	441-449	399-407	418-426	395-403	9

Tableau 6.5: Blocs structuraux trouvés par P. Jean [Jean *et al.*, 1996] sur les P450. Là où il y a désaccord avec Hasemann, les deux positions sont indiquées (notées J et H respectivement). Une lettre majuscule pour les éléments structuraux dénote une hélice, β un feuillet.

correspondant à l'hélice F dans leur travail de modélisation de la structure du P450$_{Eryf}$ (non cristallisé à l'époque) à partir de l'alignement multiple structural et en séquence des P450$_{Cam}$, P450$_{Terp}$ et P450$_{BM_3}$ [Jean *et al.*, 1996].

Certains de ces motifs structuraux sont étroitement associés à l'activité des P450, ainsi par exemple, la région de la cystéine (également appelé 'Cys-pocket' parce qu'elle forme une poche ou cavité) dans laquelle se trouve placé l'hème, les hélices F, G et H qui se trouvent proches de la surface de la macromolécule et dont on pense qu'elles contiennent le site de reconnaissance du substrat, la région appelée 'meander' proche de la 'Cys-pocket' qui exerce probablement un rôle dans les réactions d'oxydation et de réduction des électrons venant s'attacher à l'hème (en anglais 'redox-partner protein interaction'), enfin la longue hélice centrale I qui possède sans doute une fonction d'activation de l'oxygène moléculaire pour l'hème.

Notons finalement que tous ces motifs structuraux ont également été retrouvés de manière pratiquement identique (la longueur des blocs est parfois une ou deux unités plus petite) sur le P450$_{Eryf}$, plus récemment cristallisé. L'alignement des quatre P450 cette fois (figure 6.16) montre que la conservation en termes de séquence au niveau de ces blocs est encore plus faible : seules 15 positions présentent le même acide aminé, dont 7 dans les blocs 12A et 12B ('Cys-pocket' et hélice L). Dans ce cas, un bloc structural de plus est trouvé par P. Jean, entre les blocs 1 et 2A précédemment identifiés sur les trois premiers P450 cristallisés.

L'intérêt que présentait pour nous ces protéines était justement de déterminer dans quelle mesure nos algorithmes sont capables de repérer une partie au moins de ces motifs structuraux, en ne considérant que les séquences, c'est-à-dire en ignorant les informations structurales. Cette étude était une bonne occasion également pour examiner et comparer le comportement des deux principales familles d'algorithmes destinés à l'analyse des protéines que

```
BM3    tikempqpktfgelknlpllntdkpvqalmkiadelgeifkfeapgr---
Cam    nlaplpphvpehlvfdfdmynpsnlsagvqeawavlqesnvpdlvwtrcn
Terp   mdaratipehiartvilpqgyaddeviypafkwlrdeqplamahiegyd-

              BLOC 1
BM3    -VTRYLSSQRLIKEACdesrfdknlsqalkfvrdfagdglftswthe---
Cam    gGHWIATRGQLIREAYedyrhfssecpfipreageaydfiptsmdp----
Terp   -PMWIATKHADVMQIGkqpglfsnaegseilydqnneafmrsisggcphv

                 BLOC 2A  |      BLOC 2B
BM3    ---------KNWKKAHNILLPSFSQQAMKGYHAMMVDIAVQLVQKwerln
Cam    ---------PEQRQFRALANQVVGMPVVDKLENRIQELACSLIESlrpqg
Terp   idsltsmdpPTHTAYRGLTLNWFQPASIRKLEENIRRIAQASVQRlldfd

                    BLOC 3                     BLOC 4
BM3    adehievpeDMTRLTLDTIGLCgfnyrfnsfyrdqpHPFITSMVRALDEA
Cam    qcnfted--YAEPFPIRIFMLLaglp----------EEDIPHLKYLTDQM
Terp   gecdfmtd-CALYYPLHVVMTAlgvp----------EDDEPLMLKLTQDF

                                     BLOC 5
BM3    Mnklqranpddpaydenkrq----FQEDIKVMNDLVDKIIADRKAsgEQS
Cam    Trpdgsmt----------------FAEAKEALYDYLIPIIEQRRQ--KPG
Terp   FgvhepdeqavaaprqsadeaarrFHETIATFYDYFNGFTVDRRS--CPK

        BLOC 6              BLOC 7 (hélices I+J)
BM3    DDLLTHMLNgkdpetgePLDDENIRYQIITFLIAGHETTSGLLSFALYFL
Cam    TDAISIVANgqvngr--PITSDEAKRMCGLLLVGGLDTVVNFLSFSMEFL
Terp   DDVMSLLANskldgn--YIDDKYINAYYVAIATAGHDTTSSSSGGAIIGL

                                   BLOC 8 (hélice K)
BM3    VKNPHVLQKAAEeaarvlvdpvpsykqvkqlKYVGMVLNEALRLWPTApA
Cam    AKSPEHRQELIErp------------------ERIPAACEELLRRFSLV-A
Terp   SRNPEQLALAKSdp------------------ALIPRLVDEAVRWTAPVkS

          BLOC 9     BLOC 10 (hélice K') BLOC 11 (meander)
BM3    FSLYAKEDTVLGGeYPLEKGDELMVLIPQLHRDKTIWgdDVEEFRPErfe
Cam    DGRILTSDYEFHG-VQLKKGDQILLPQMLSGLDERENa-CPMHVDFSrqk
Terp   FMRTALADTEVRG-QNIKRGDRIMLSYPSANRDEEVFs-NPDEFDITrfp

              BLOCs 12A et 12B (Cys-pocket + hélice L)
BM3    npsaipqhaFKPFGNGQRACIGQQFALHEATLVLGMmlkhfdfedhtnye
Cam    vs-------HTTFGHGSHLCLGQHLARREIIVTLKEwltripdfsiapga
Terp   nr-------HLGFGWGAHMCLGQHLAKLEMKIFFEEllpklksvelsgpp

                   BLOC 13
BM3    ldiketltlk-PEGFVVKAKskkiplgg
Cam    qiqhksgivsgVQALPLVWDpattkav
Terp   rlvatnfvgg-PKNVPIRFTka
```

Figure 6.13: Arrangement des motifs structuraux trouvés par P. Jean [Jean *et al.*, 1996] dans une comparaison avec KMRC/S (section 4.2.3.5) des structures codées en chaînes de coordonnées internes des trois premiers P450 cristallisés.

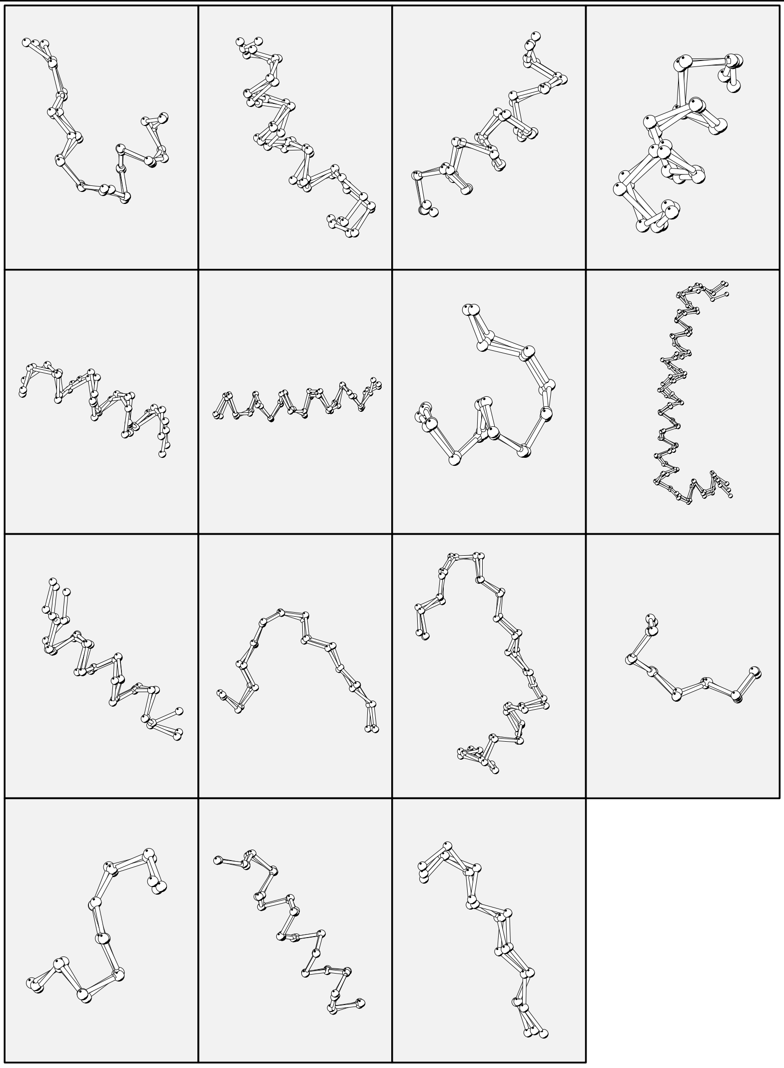

Figure 6.14: Motifs structuraux identifiés sur les P450$_{Cam}$, P450$_{Terp}$ et P450$_{BM_3}$ par KMRC/S [Jean *et al.*, 1996]. Ces motifs sont montrés superposés et correspondent (de la gauche vers la droite et du haut vers le bas) aux blocs 1, 2A, 2B, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12A, 12B, 13.

Figure 6.15: Structure du $P450_{Eryf}$.

```
BM3  tikempqpktfgelknlpllntdkpvqalmkiadelgeifkfeapgr---
Cam  nlaplpphvpehlvfdfdmynpsnlsagvqeawavlqesnvpdlvwtrcn
Terp mdaratipehiartvilpqgyaddeviypafkwlrdeqplamahiegyd-
Eryf mttvpdlesdsfhvdwyrtyaelretapvtpvrflgq------------

BM3  -VTRYLSSQRLIKEACdesrfdknlsqalkfvrdfagd------------
Cam  gGHWIATRGQLIREAYedyrhfssecpfipreageaydf-----------
Terp -PMWIATKHADVMQIGkqpglfsnaegseilydqnneafmrsisggcphv
Eryf -DAWLVTGYDEAKAALsdlrlssdpkkkypgvevefpaylgfpedvrnyf

     NOUVEAU BLOC
BM3  --GLFTSWTHekNWKKAHNILLPSFSQQAMKGYHAMMVDIAVQLVQKwer
Cam  --IPTSMDPP--EQRQFRALANQVVGMPVVDKLENRIQELACSLIESlrp
Terp idSLTSMDPP--THTAYRGLTLNWFQPASIRKLEENIRRIAQASVQRlld
Eryf atNMGTSDPP--THTRLRKLVSQEFTVRRVEAMRPRVEQITAELLDEvgd

BM3  lnadehievpeDMTRLTLDTIGLCgfnyrfnsfyrdqphpfitSMVRALD
Cam  qgqcnfted--YAEPFPIRIFMLLaglpeedip----------HLKYLTD
Terp fdgecdfmtd-CALYYPLHVVMTAlgvpeddep----------LMLKLTQ
Eryf sgvvdivdr--FAHPLPIKVICELlgvdekyrg----------EFGRWSS

BM3  EAMnklqranpddpaydenkrq----FQEDIKVMNDLVDKIIADRKAsge
Cam  QMTrpdgsmt----------------FAEAKEALYDYLIPIIEQRRQkp-
Terp DFFgvhepdeqavaaprqsadeaarrFHETIATFYDYFNGFTVDRRScp-
Eryf EILvmdperaeq--------------RGQAAREVVNFILDLVERRRTep-

BM3  qSDDLLTHMLngkdpetgePLDDENIRYQIITFLIAGHETTSGLLSFALY
Cam  -GTDAISIVAngqvngr--PITSDEAKRMCGLLLVGGLDTVVNFLSFSME
Terp -KDDVMSLLAnskldgn--YIDDKYINAYYVAIATAGHDTTSSSSGGAII
Eryf -GDDLLSALIrvqddddg-RLSADELTSIALVLLLAGFEASVSLIGIGTY

BM3  FLVKNPHVLQKAAEeaarvlvdpvpsykqvkqlKYVGMVLNEALRLWPTA
Cam  FLAKSPEHRQELIErp-----------------ERIPAACEELLRRFSLV
Terp GLSRNPEQLALAKSdp-----------------ALIPRLVDEAVRWTAPV
Eryf LLLTHPDQLALVRRdp-----------------SALPNAVEEILRYIAPP

BM3  pafsLYAKEDTVLGGeYPLEKGDELMVLIPQLHRDKTIWgddVEEFRPer
Cam  adg-RILTSDYEFHG-VQLKKGDQILLPQMLSGLDERENac-PMHVDFsr
Terp ksfmRTALADTEVRG-QNIKRGDRIMLSYPSANRDEEVFsn-PDEFDItr
Eryf etttRFAAEEVEIGG-VAIPQYSTVLVANGAANRDPKQFpd-PHRFDVtr

BM3  fenpsaipqhaFKPFGNGQraciGQQFALHEATLVLGMMLkhfdfedhtn
Cam  qkvs-------HTTFGHGShlclGQHLARREIIVTLKEWLtripdfsiap
Terp fpnr-------HLGFGWGAhmclGQHLAKLEMKIFFEELLpklksvelsg
Eryf dtrg-------HLSFGQGIhfcmGRPLAKLEGEVALRALFgrfpalslgi

BM3  yeldiketltlk---PEGFVVKakskkiplggipspsteqsakkvr
Cam  gaqiqhksgivsg--VQALPLVwdpattkav
Terp pprlvatnfvgg---PKNVPIRftka
Eryf daddvvwrrslllrgIDHLPVRldg
```

Figure 6.16: Arrangement des motifs structuraux trouvés par P. Jean [Jean *et al.*, 1996] dans une comparaison avec KMRC/S (section 4.2.3.5) des structures codées en chaînes de coordonnées internes des quatre premiers P450 cristallisés.

nous avons développés : la famille LePoivre-LeCombi travaillant avec une notion de distance et dont l'unité de comparaison est l'ensemble des symboles de l'alphabet des acides aminés, et Klast travaillant avec une mesure de similarité et dont l'unité de comparaison est l'ensemble des sous-mots de longueur w.

Nous avons pour ce faire utilisé deux jeux d'essais comme ensembles de séquences exemples. Le premier est composé des $\mathrm{P450}_{Cam}$, $\mathrm{P450}_{Terp}$ et $\mathrm{P450}_{BM_3}$ (séquences extraites des entrées PBD [Bernstein *et al.*, 1977] 3cpp, 1ept et 2hpd). Le second est composé de ces trois mêmes séquences plus celle du dernier P450 cristallisé, le $\mathrm{P450}_{Eryf}$ (séquence extraite de l'entrée PDB 1oxa). L'ensemble de séquences tests est constitué, soit du $\mathrm{P450}_{Eryf}$ tout seul, soit d'un P450 humain de la famille III considérée comme étant la plus proche des P450 bactériens en termes de séquence (de nom CP34_HUMAN dans SWISS PROT [Bairoch and Boeckmann, 1991]), soit d'un jeu composé de tous les 9 P450 humains de la famille III (de noms CP31_RAT, CP32_RAT, CP33_HUMAN, CP34_HUMAN, CP35_HUMAN, CP36_RABIT, CP37_HUMAN, CP38_MACFA, CP3C_CANFA dans SWISS PROT), soit enfin de l'ensemble des quatre P450 cristallisés.

6.3.3.2 Résultats

Nous présentons maintenant quelques-uns des résultats préliminaires que nous avons obtenus sur ces séquences, tout d'abord ceux obtenus avec CoSamp, puis ceux produits par CLUSTAL W et MACAW + Gibbs.

Les résultats peuvent être rangés en trois groupes comportant chacun un ou deux essais. Dans tous les cas, la figure correspondant à chacun des essais montre les modèles trouvés par p-value croissante. Ces figures indiquent :

- dans une première ligne, le score et la p-value associée au modèle (i.e. celle du profil construit à partir de son ensemble d'occurrences) ainsi que l'élément structural concerné (numéro du bloc dans l'article de P. Jean et dans le tableau 6.5 et nom de la structure secondaire);
- le bloc (en séquence) correspondant au modèle avec, dans chaque cas, le nom des P450, les positions et les occurrences elles-mêmes.

Lorsqu'une de ces occurrences est mal placée sur une des séquences, nous signalons à côté de celle-ci le décalage observé par rapport à la bonne position, ou marquons simplement 'faux' si ce décalage est trop grand (plus de 10 positions). La position considérée comme correcte est celle établie par Jean sauf dans deux cas, clairement indiqués, où les positions du bloc lexical que nous trouvons correspondent à celles des blocs structuraux de Hasemann.

Par ailleurs, les algorithmes, paramètres, ensembles d'exemples et tests avec lesquels nous avons travaillé ont été :

Groupe 1

- Séquences exemples : $P450_{BM3}$, $P450_{Cam}$, $P450_{Terp}$
- Séquence test : $P450_{Eryf}$
- Algorithme utilisé en première phase : Klast dans sa version 'seuil sur la moyenne des scores des sous-mots' (section 5.4.3.5.2)
- Paramètres :
  - quorum : 100%

- matrice : PAM250 normalisée (figure 3.7)
- seuil t : 62
- longueur des sous-mots w : 3
- longueur des modèles : 10

- Temps d'exécution de Klast : 379 secondes (sur une Silicon Graphics Indigo WorkStation (R4000))
- Nombre de modèles initiaux : 26668 correspondant à uniquement 196 ensembles distincts d'occurrences

Groupe 2

- Séquences exemples : $P450_{BM3}$, $P450_{Cam}$, $P450_{Terp}$
- Séquences tests :

  essai 1 $P450_{Eryf}$

  essai 2 neuf P450 III humains
- Algorithme utilisé en première phase : LeCombi
- Paramètres :
  - quorum : 100%
  - nombre de jokers : 8
  - nombre de substitutions autorisées : 1
  - longueur des modèles : 15
- Couverture pondérée restreinte :

$$CCPR = \left\{ \begin{array}{l} (\{X\}, \infty)\ \forall X \in C, \\ (\Sigma, 8), \\ (S, 0)\ pour\ tous\ les\ autres\ ensembles \end{array} \right\}$$

  avec :

$$C = \left\{ \begin{array}{ll} \{I, L, M, V\} & (hydrophobe), \\ \{A, G\} & (très\ petit), \\ \{A, S, T\} & (petit), \\ \{C\} & (cystéine), \\ \{F, Y, W\} & (aromatique), \\ \{K, H, R\} & (basique), \\ \{E, D\} & (acide), \\ \{N, Q\} & (glutamine\ et\ glutamate), \\ \{G, P\} & (apparaît\ dans\ les\ structures\ contraintes) \end{array} \right\}$$

- Temps d'exécution de LeCombi : 3843 secondes (sur une Sparc Station 20)
- Nombre de modèles initiaux : 188906 correspondant en fait à 15452 ensembles distincts d'occurrences

Groupe 3

- Séquences exemples : $P450_{BM3}$, $P450_{Cam}$, $P450_{Terp}$, $P450_{Eryf}$

- Séquences tests :

  essai 1 les mêmes séquences exemples

  essai 2 le P450 humain CP34_HUMAN
- Algorithme utilisé en première phase : LeCombi
- Paramètres :
  - quorum : 75%
  - nombre de jokers : 6
  - nombre de substitutions autorisées : 1
  - longueur des modèles : 14
- Couverture pondérée restreinte : la même que pour le Groupe 2
- Temps d'exécution de LeCombi : 3275 secondes (sur une Sparc Station 20)
- Nombre de modèles initiaux : 128622 modèles correspondant à 4281 ensembles d'occurrences distincts

Notons que les p-values de la figure 6.20 sont très basses relativement à celles observées avec les autres essais. Nous suggérons comme explication à ce comportement le fait que les séquences tests sont ici les mêmes que les séquences exemples. Cela pose d'ailleurs la question de savoir comment traiter un tel choix dans la seconde étape, de telle sorte que ce phénomène ne se produise pas. Observons que tous les modèles repérés avec cet essai sont présentés, sauf trois ayant des p-values légèrement plus élevées et qui sont complètement faux. Aucun autre modèle n'est identifié.

Les résultats obtenus avec CLUSTAL W et MACAW + Gibbs sont quant à eux donnés dans les figures 6.22 et 6.23. Dans les deux cas, le jeu d'essai utilisé est composé des $P450_{Cam}$, $P450_{Terp}$ et $P450_{BM3}$. Comme CLUSTAL est un algorithme d'alignement global et non un programme de recherche de blocs, ceux indiqués sur la figure (correspondant à des blocs structuraux qu'il aligne correctement) ont été ajoutés par nous *a posteriori*. CLUSTAL W aussi bien que MACAW + Gibbs ont été utilisés avec leurs paramètres par défaut, sauf dans le cas de MACAW + Gibbs où les longueurs minimale et maximale des blocs recherchés doivent être fixées par l'utilisateur. Nous les avons établies à 4 et 18 respectivement.

6.3.3.3 Discussion

Considérons d'abord les résultats fournis par CoSamp.

La première observation que nous pouvons faire est que ces résultats sont d'inégales qualités selon l'algorithme X, les paramètres utilisés en première phase et les ensembles employés en exemple et en test. Le 'meilleur' est celui de la figure 6.18 pour deux raisons. La première est que cet essai localise partiellement 8 des 13 blocs structuraux, y compris la région du 'meander'. La seconde est qu'il ne présente aucun modèle totalement faux parmi ceux dont la p-value est inférieure à 10^{-6}. De plus, pour les trois derniers modèles qui possèdent une p-value inférieure à 10^{-3}, la valeur de celle-ci chute brutalement. Dans les autres essais, cette valeur, soit décroît de manière continue, soit chute bien plus tôt alors que des 'bons' modèles existent encore.

Il nous est difficile pour l'instant d'expliquer pourquoi le comportement de CoSamp varie ainsi avec les données d'entrée et avec la méthode employée en première phase. L'algorithme lui-même, et surtout sa seconde partie, a certainement besoin d'être plus profondément réfléchi. Une première amélioration qui pourrait lui être apportée concerne les modèles éliminés parce qu'une de leurs occurrences chevauche celle d'un autre modèle ayant une plus petite p-value. Cette élimination ne devrait pas avoir lieu lorsque l'ensemble des occurrences d'un des modèles

```
score 29.0 p-value -10.38  bloc 12A 'Cys-pocket'
BM3    392    PFGNGQRACI
Cam    349    TFGHGSHLCL
Terp   369    GFGWGAHMCL
-
score 18.3 p-value -7.91   bloc 2A Hélice C
BM3    220    EQSDDLLTHM  faux
Cam    100    PTSMDPPEQR
Terp   102    LTSMDPPTHT
-
score 20.3 p-value -7.31   totalement faux
BM3    290    AAEEAARVLV
Cam    284    ACEELLRRFS
Terp   128    KLEENIRRIA
-
score 19.0 p-value -6.48   bloc 5 Hélice G (selon Hasemann)
BM3    198    YDENKRQFQE
Cam    193    FAEAKEALYD  décalé de +3
Terp   205    ADEAARRFHE
-
score 15.3 p-value -4.83   bloc 7 Hélices I et J
BM3    262    LIAGHETTSG
Cam    34     LSAGVQEAWA  faux
Terp   265    ATAGHDTTSS
-
score 13.0 p-value -3.92   totalement faux
BM3    180    ALDEAMNKLQ
Cam    326    GLDERENACP
Terp   34     LRDEQPLAMA
-
score 15.0 p-value -3.89   bloc 13 Hélice L
BM3    402    GQQFALHEAT
Cam    359    GQHLARREII
Terp   379    GQHLAKLEMK
-
score 13.3 p-value -3.17   bloc 4 un peu avant Hélice F
BM3    246    GEPLDDENIR  faux
Cam    168    GLPEEDIPHL
Terp   171    GVPEDDEPLM
-
score 15.7 p-value -3.15   bloc 10 Feuillets β2-2 et β1-3
BM3    348    EKGDELMVLI
Cam    313    KKGDQILLPQ
Terp   333    KRGDRIMLSY
-
score 12.3 p-value -3.14   bloc 8 Feuillet β6-1
BM3    318    LNEALRLWPT
Cam    235    SDEAKRMCGL  faux
Terp   304    VDEAVRWTAP
-
score 12.3 p-value -2.96   bloc 7 Hélice J
BM3    278    YFLVKNPHVL
Cam    262    EFLAKSPEHR
Terp   281    IGLSRNPEQL
-
```

Figure 6.17: Résultat première série P450 (Klast, trois séquences, quorum 100%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
score 42.7 p-value -14.30  bloc 12B 'Cys-pocket'
BM3    393    FGNGQRACIGQQFAL
Cam    350    FGHGSHLCLGQHLAR
Terp   370    FGWGAHMCLGQHLAK
-
score 31.7 p-value -11.25  bloc 8 Hélice K
BM3    282    KNPHVLQKAAEEAAR  faux
Cam    276    ERPERIPAACEELLR
Terp   295    SDPALIPRLVDEAVR
-
score 28.7 p-value -10.15  bloc 6 Hélice H
BM3    225    ASGEQSDDLLTHMLN
Cam    211    RRQKPGTDAISIVAN
Terp   230    RRSCPKDDVMSLLAN
-
score 20.7 p-value -8.05   bloc 2A Hélice C
BM3    112    MKGYHAMMVDIAVQL  faux
Cam    104    DPPEQRQFRALANQV
Terp   106    DPPTHTAYRGLTLNW
-
score 22.0 p-value -7.99   blocs 10-11 Hélice K' et 'meander'
BM3    369    DDVEEFRPERFENPS  décalé de +7
Cam    328    DERENACPMHVDFSR
Terp   348    DEEVFSNPDEFDITR
-
score 19.3 p-value -7.10   bloc 3 Hélice E
BM3    342    GGEYPLEKGDELMVL  faux
Cam    155    AEPFPIRIFMLLAGL
Terp   158    ALYYPLHVVMTALGV
-
score 21.3 p-value -6.82   bloc 7 Hélices I et J
BM3    260    TFLIAGHETTSGLLS
Cam    244    LLLVGGLDTVVNFLS
Terp   263    AIATAGHDTTSSSSG
-
score 13.3 p-value -3.80   totalement faux
BM3    143    EDMTRLTLDTIGLCG
Cam    132    QELACSLIESLRPQG
Terp   206    DEAARRFHETIATFY
-
score 12.3 p-value -3.72   totalement faux
BM3    174    ITSMVRALDEAMNKL
Cam    399    VQALPLVWDPATTKA
Terp   395    LPKLKSVELSGPPRL
-
score 13.0 p-value -3.51   totalement faux
BM3    431    LDIKETLTLKPEGFV
Cam    194    AEAKEALYDYLIPII
Terp   1      MDARATIPEHIARTV
-
```

Figure 6.18: Résultat seconde série — premier essai P450 (LeCombi, trois séquences, quorum de 100%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
score 38.9 p-value -12.76  bloc 12A 'Cys-pocket'
BM3    392    PFGNGQRACIGQQFA
Cam    349    TFGHGSHLCLGQHLA
Terp   369    GFGWGAHMCLGQHLA
-
score 31.4 p-value -11.64  bloc 7 Hélices I et J
BM3    263    IAGHETTSGLLSFAL
Cam    247    VGGLDTVVNFLSFSM
Terp   266    TAGHDTTSSSSGGAI
-
score 17.1 p-value -5.23   Hélice A de Hasemann?
BM3    303    PSYKQVKQLKYVGMV  faux
Cam    40     EAWAVLQESNVPDLV
Terp   29     PAFKWLRDEQPLAMA  décalé de +4
-
score 15.0 p-value -4.88   bloc 9 Feuillets β1-4 et β2-1
BM3    284    PHVLQKAAEEAARVL  faux
Cam    298    GRILTSDYEFHGVQL
Terp   318    MRTALADTEVRGQNI
-
score 18.1 p-value -4.63   bloc 10 Feuillets β2-2 et β1-3 et Hélice K'
BM3    349    KGDELMVLIPQLHRD
Cam    314    KGDQILLPQMLSGLD
Terp   334    RGDRIMLSYPSANRD
-
score 14.8 p-value -4.17   bloc 4 un peu avant Hélice F
BM3    18     PLLNTDKPVQALMKI  faux
Cam    168    GLPEEDIPHLKYLTD
Terp   171    GVPEDDEPLMLKLTQ
-
```

Figure 6.19: Résultat seconde série — deuxième essai P450 (LeCombi, trois séquences, quorum de 100%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
score 53.0 p-value -19.17  bloc 12A 'Cys-pocket'
BM3    393    FGNGQRACIGQQFA
Cam    350    FGHGSHLCLGQHLA
Terp   370    FGWGAHMCLGQHLA
Eryf   344    FGQGIHFCMGRPLA
-
score 34.0 p-value -14.03  bloc avant 2A Hélice C
                           nouveau bloc P. Jean
BM3    1      TIKEMPQPKTFGEL  faux
Cam    101    TSMDPPEQRQFRAL
Terp   103    TSMDPPTHTAYRGL
Eryf   91     GTSDPPTHTRLRKL
-
score 36.8 p-value -13.71  bloc 8 Hélice K
BM3    283    NPHVLQKAAEEAAR  faux
Cam    277    RPERIPAACEELLR
Terp   296    DPALIPRLVDEAVR
Eryf   270    DPSALPNAVEEILR
-
score 31.3 p-value -12.67  blocs 5 et 6 Hélices G et H
BM3    320    EALRLWPTAPAFSL  faux
Cam    209    EQRRQKPGTDAISI
Terp   228    VDRRSCPKDDVMSL
Eryf   201    ERRRTEPGDDLLSA
-
score 28.6 p-value -12.10  bloc 7 Feuillet β5-2 et Hélice I
BM3    263    IAGHETTSGLLSFA
Cam    247    VGGLDTVVNFLSFS
Terp   266    TAGHDTTSSSSGGA
Eryf   285    IAPPETTTRFAAEE faux
-
score 29.8 p-value -11.51  bloc 4 avant Hélice F
BM3    245    TGEPLDDENIRYQI faux
Cam    167    AGLPEEDIPHLKYL
Terp   170    LGVPEDDEPLMLKL
Eryf   377    LGIDADDVVWRRSL faux
-
score 26.8 p-value -11.30  entre blocs 4 et 5 entre Hélices F et G
BM3    201    NKRQFQEDIKVMND
Cam    189    GSMTFAEAKEALYD
Terp   208    AARRFHETIATFYD
Eryf   41     LVTGYDEAKAALSD faux
-
score 29.6 p-value -11.24  bloc 10 Feuillets β2-2 et β1-3
BM3    350    GDELMVLIPQLHRD
Cam    315    GDQILLPQMLSGLD
Terp   335    GDRIMLSYPSANRD
Eryf   309    YSTVLVANGAANRD
-
```

Figure 6.20: Résultat troisième série — premier essai P450 (LeCombi, quatre séquences, quorum de 75%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
score 28.5 p-value -12.29  bloc 7 Hélices I et J
BM3    264    AGHETTSGLLSFAL
Cam    248    GGLDTVVNFLSFSM
Terp   267    AGHDTTSSSSGGAI
Eryf   227    SADELTSIALVLLL  faux
-
score 35.8 p-value -11.92  bloc 12A 'Cys-pocket'
BM3    393    FGNGQRACIGQQFA
Cam    350    FGHGSHLCLGQHLA
Terp   370    FGWGAHMCLGQHLA
Eryf   344    FGQGIHFCMGRPLA
-
score 21.8 p-value -7.96   bloc 8 Hélice K et Feuillet β6-1
BM3    333    LYAKEDTVLGGEYP
Cam    299    RILTSDYEFHGVQL
Terp   319    RTALADTEVRGQNI
Eryf   293    RFAAEEVEIGGVAI
-
score 20.5 p-value -7.87   totalement faux
BM3    284    PHVLQKAAEEAARV
Cam    86     PFIPREAGEAYDFI
Terp   123    PASIRKLEENIRRI
Eryf   119    PRVEQITAELLDEV
-
score 21.3 p-value -7.25   juste avant bloc 8 Hélice K
BM3    310    QLKYVGMVLNEALR
Cam    130    RIQELACSLIESLR  faux
Terp   296    DPALIPRLVDEAVR
Eryf   270    DPSALPNAVEEILR
-
score 16.0 p-value -5.53   totalement faux
BM3    175    TSMVRALDEAMNKL
Cam    400    QALPLVWDPATTKA
Terp   168    TALGVPEDDEPLML
Eryf   213    SALIRVQDDDDGRL
-
score 16.3 p-value -5.22   bloc 10 Feuillets β2-2 et β1-3 et Hélice K'
BM3    354    MVLIPQLHRDKTIW
Cam    222    IVANGQVNGRPITS  faux
Terp   339    MLSYPSANRDEEVF
Eryf   313    LVANGAANRDPKQF
-
score 16.0 p-value -4.40   bloc 12B Hélice L
BM3    206    QEDIKVMNDLVDKI  faux
Cam    365    REIIVTLKEWLTRI
Terp   385    LEMKIFFEELLPKL
Eryf   190    REVVNFILDLVERR  faux
-
score 12.0 p-value -3.61   totalement faux
BM3    10     TFGELKNLPLLNTD
Cam    7      SNANLAPLPPHVPE
Terp   137    AQASVQRLLDFDGE
Eryf   19     TYAELRETAPVTPV
-
```

Figure 6.21: Résultat troisième série — second essai P450 (LeCombi, quatre séquences, quorum de 75%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
BM3         TIKEMPQPKTFGELKNLPLLNTDKPVQALMKIADELGEIFKFEAPGRVTRYLS-SQRLIK
Cam         ----------NANLAPLPPHVPEHLVFDFDMYNPSN--LSAGVQEAWAVLQESNVPDLVW
Terp        ----MDARATIPEHIARTVILPQGYADDEVIYPAFK--WLRDEQPLAMAHIEGYDPMWIA
                     .                                                .

BM3         EACDESRFDKNLSQALKFVR--DFAGDGLFTSWTHEKNWKKAHNILLPSFSQQAMKGYHA
Cam         TRCNGGHWIATRGQLIREAY-EDYRHFSSECPFIPREAGEAYDFIP-----TSMDPPEQR
Terp        TKHADVMQIGKQPGLFSNAEGSEILYDQNNEAFMRSISGGCPHVIDS---LTSMDPPTHT
                                  .                     *             .

BM3         MMVDIAVQLVQKWERLNADEHIEVPEDMTRLTLD-TIGLCGFNYRFNSFYRDQPHPFITS
Cam         QFRALANQVVGMPVVDKLENRIQELACSLIESLR-PQGQCNFTEDYAEPFPIRIFMLLAG
Terp        AYRGLTLNWFQPASIRKLEENIRRIAQASVQRLLDFDGECDFMTDCALYYPLHVVMTALG
                .. .          .  *          *    * * *       .  .

                                                           BLOC 5
BM3         MVRALDEAMNKLQRANP--DDPAYD----------ENKRQFQEDIKVMNDLVDKIIADRK
Cam         LPEEDIPHLKYLTDQMT---RPD-------------GSMTFAEAKEALYDYLIPIIEQRR
Terp        VPEDDEPLMLKLTQDFFGVHEPDEQAVAAPRQSADEAARRFHETIATFYDYFNGFTVDRR
            .       .  *         *                  * *      *        *.

                     BLOC 6                       BLOC 7
BM3         ASGEQSDDLLTHMLNGKDPETGEPLDDENIRYQIITFLIAGHETTSGLLSFALYFLVKNP
Cam         QKP--GTDAISIVANGQVN--GRPITSDEAKRMCGLLLVGGLDTVVNFLSFSMEFLAKSP
Terp        SCP--KDDVMSLLANSKLD--GNYIDDKYINAYYVAIATAGHDTTSSSSGGAIIGLSRNP
                   * .. . * .    *  .               * .*      ..   * . *

                                            BLOC 8           BLOC 9
BM3         HVLQKAAEEAARVLVDPVPSYKQVKQLKYVGMVLNEALRLWPTAPAFSLYAKEDTVLGGE
Cam         EHRQELIERPE-----------------RIPAACEELLRRFS-IVADGRILTSDYEFHGV
Terp        EQLALAKSDPA-----------------LIPRLVDEAVRWTAPVKSFMRTALADTEVRGQ
                                         .     * .*      .           *    *

                      BLOC 10                               BLOC 12A
Terp        YPLEKGDELMVLIPQLHRDKTIWGDDVEEFRPERFENPSAIPQHAFKPFGNGQRACIGQQ
Cam         Q-LKKGDQILLPQMLSGLDEREN-ACPMHVDFSRQK-------VSHTTFGHGSHLCLGQH
Terp        N-IKRGDRIMLSYPSANRDEEVF-SNPDEFDITRFP-------NRHLGFGWGAHMCLGQH
              . .** ...       *              *               ** * . *.**.

                 BLOC 12B                          BLOC 13
BM3         FALHEATLVLGMMLKHFDFEDHT-NYELDIKETLTLKPEGFVVKAKSKKIPLGG
Cam         LARREIIVTLKEWLTRIPDFSIAPGAQIQHKSGIVSGVQALPLVWDPATTKAV-
Terp        LAKLEMKIFFEELLPKLKSVELS-GPPRLVATNFVGGPKNVPIRFTKA------
             *  *  .     * .      . .                 .
```

Figure 6.22: Alignement multiple global obtenu avec CLUSTAL W sur les trois premiers P450 cristallisés. L'indication des blocs structuraux correctement alignés a été réalisée par nous, *a posteriori*. Une étoile en-dessous d'une colonne indique la présence de résidus identiques à cet endroit de l'alignement, un point la présence de résidus équivalents.

```
BM3    tikempqpktfgelknlpllntdkpvqalmkiadelgeifkfeapgrvtrylssqrlike
Cam    nanlaplpphvpehlvfdfdmynpsnlsagvqeawavlqesnvpdlvwtrcng
Terp   mdaratipehiartvilpqgyaddeviypafkwlrdeqplamahiegydpmwiatkhadv

BM3    acdesrfdknlsqalkfvrdfagdglftswtheknwkkahnillpsfsqqamkgyhammv
Cam    ghwiatrgqlireayedyrhfssecpfipreageaydfiptsmdppeqrqfralanqvvg
Terp   mqigkqpglfsnaegseilydqnneafmrsisggcphvidsltsmdppthtayrgltlnw

                                         BLOC FAUX
BM3    diavqlvqkwerlnadehievpedmtrltLDTIGLCGFNYRFNSFYRdqphpfitsmvra
Cam    mpvvdklenriqelacslies--------LRPQGQCNFTEDYAEPFPirifmllaglpee
Terp   fqpasirkleenirriaqasvqrl-----lDFDGECDFMTDCALYYPlhvvmtalgvped

BM3    ldeamnklqranpddpaydenkrqfqedikvmndlvdkiiadrkasgeqsddllthmlng
Cam    diphlkyltdqmtrpdgsmtfaeakealydylipiieqrrqkpgtdaisivangqvngrp
Terp   deplmlkltqdffgvhepdeqavaaprqsadeaarrfhetiatfydyfngftvdrrscpk

BM3    kdpetgeplddeniryqiitfliaghettsgllsfalyflvknphvlqkaaeeaarvlvd
Cam    itsdeakrmcglllvggldtvvnflsfsmeflakspehrqelierperipaaceellrrf
Terp   ddvmsllanskldgnyiddkyinayyvaiataghdttssssggaiiglsrnpeqlalaks

BM3    pvpsykqvkqlkyvgmvlnealrlwptapafslyakedtvlggeyplekgdelmvlipql
Cam    slvadgriltsdyefhgvqlkkgdqillpqmlsglderenacpmhvdfsrqk--------
Terp   dpaliprlvdeavrwtapvksfmrtaladtevrgqnikrgdrimlsypsanrdeevfsnp

                                       BLOCS 12A ET 12B
BM3    hrdktiwgddveefrperfenpsaipqHAFKPFGNGQRACIGQQFALHEATlvlgmmlkh
Cam    ---------------------------VSHTTFGHGSHLCLGQHLARREIIvtlkewltr
Terp   defditrfp------------------NRHLGFGWGAHMCLGQHLAKLEMKiffeellpk

BM3    fdfedhtnyeldiketltlkpegfvvkakskkiplgg
Cam    ipdfsiapgaqiqhksgivsgvqalplvwdpattkav
Terp   lksvelsgpprlvatnfvggpknvpirftka------
```

Figure 6.23: Blocs trouvés par la méthode d'échantillonnage de Gibbs sur les trois premiers P450 cristallisés.

représente un simple décalage de l'ensemble des occurrences de l'autre : il est clair qu'alors un des deux modèles est un prolongement de l'autre et tous les deux devraient être conservés dans la liste finale. Ce cas n'est pas rare, et il risque d'être d'autant plus fréquent que certains des blocs structuraux sont longs par rapport à la longueur des modèles recherchés. Cela est le cas en particulier ici où la longueur des motifs structuraux varie entre 8 et 23.

Cette variation dans la longueur des blocs structuraux est sans doute un des facteurs qui font que certains blocs ne sont pas repérés en tant que blocs lexicaux. Considérons le cas des essais des groupes 2 et 3. La longueur des modèles recherchés par LeCombi est de 15 et 14 respectivement. Cela nous donne assez peu de chances de localiser les blocs 1, 2B, 3, 4, 6, 9, 11 et 13 sauf s'ils sont collés à un autre bloc comme cela est le cas de la 'Cys-pocket' (bloc 12A) qui est juxtaposée au bloc 12B ce qui lui permet, outre sa plus forte conservation, d'être aisément identifiée. Ces blocs courts correspondent effectivement à ceux qui ne sont jamais, ou presque jamais, repérés bien que dans certains cas (e.g. blocs 1, 2B, 4), cela puisse être plutôt dû au fait qu'ils sont très variables en séquence (ils le sont également en structure).

L'argument que nous venons d'avancer expliquant pourquoi certains blocs ne sont pas trouvés paraît moins valable lorsque l'on considère l'essai 1 réalisé avec Klast pour lequel la longueur des modèles recherchés est égale à 10. En effet, malgré cette plus petite longueur, Klast ne parvient pas à localiser certains blocs structuraux, comme les blocs 1, 3, 6 et 9 qui sont maintenant 'à sa portée'. Cet 'échec' est cependant lié à une faiblesse qui est intrinsèque à Klast et n'invalide pas le raisonnement sur la longueur des modèles. Le résultat obtenu avec Klast est en effet moins bon dans la mesure où des modèles totalement faux se trouvent mélangés, parfois avec une p-value petite, aux modèles corrects ou partiellement corrects. Ce phénomène est également observé pour l'essai de la figure 6.21, mais la raison, dans ce cas, réside plutôt dans le choix de l'ensemble test.

Ce choix semble en effet être important pour la qualité des résultats. Ce n'est pas le seul, et la difficulté principale dans l'utilisation de CoSamp (que ce soit avec LeCombi ou Klast) demeure dans la sélection de la méthode à employer en première phase, le choix des paramètres à fixer pour cette méthode (en particulier longueur des modèles et quorum) et de l'ensemble des séquences exemples et tests. Notons par ailleurs que l'évaluation statistique des modèles peut être plus difficile à établir lorsque ceux-ci sont plus courts. Il est ainsi sans doute un peu prématuré pour l'instant de conclure que le modèle de ressemblance de LeCombi est plus approprié que celui de Klast au problème que nous voulions résoudre ici.

Nous voulons également attirer l'attention sur un phénomène observé deux fois dans cet exemple des P450 (et qui le sera encore plus dans le suivant) et qui concerne la question des hélices. À deux reprises en effet, une occurrence se trouve décalée de +3 par rapport au bloc structural que le modèle identifie. Un tel décalage de 3 ou 4 positions correspondant à un tour d'hélice par rapport à sa position correcte n'est pas observé uniquement lorsque nous essayons de trouver des blocs dans les séquences. Les algorithmes qui recherchent des motifs 3D répétés de manière automatique produisent souvent de tels décalages. Ceux-ci doivent alors être 'corrigés' *a posteriori* par une inspection visuelle des structures superposées.

Voyons maintenant les résultats obtenus par les deux autres méthodes. Celle de l'échantillonnage de Gibbs intégré dans MACAW n'est capable de repérer correctement que le bloc 12A correspondant à la 'Cys-pocket' ainsi que le début de l'hélice L. Tous les autres blocs construits sont faux, et d'ailleurs donnés comme étant statistiquement non 'significatifs'. Un de ces blocs est montré dans la figure 6.23. Il est intéressant de noter que, lorsque nous recherchons avec Klast ou LeCombi seulement (c'est-à-dire sans la seconde étape de CoSamp) les plus longs modèles présents à 100% dans l'un ou l'autre des jeux d'essais que nous avons utilisés ici, ce sont justement ces blocs que nous retrouvons toujours et exclusivement.

Le résultat produit par CLUSTAL W est quant à lui un peu surprenant. Comme nous l'avons mentionné, CLUSTAL ne cherche pas des blocs, mais réalise un alignement global des séquences. Cet alignement établit toutefois une mise en correspondance correcte des blocs 5, 6, 7, 8, 9, 10, 12A, 12B et 13 des $P450_{Cam}$, $P450_{Terp}$ et $P450_{BM3}$ ainsi que nous avons pu le vérifier manuellement. L'alignement obtenu avant le bloc 5 est, en termes de structure, incorrect. Cette partie N terminale de la protéine correspond en effet à celle de moindre conservation en séquence aussi bien qu'en structure. Les blocs lexicaux que nous y repérons nous-mêmes sont toujours faux sur une des séquences (en général sur le $P450_{BM3}$) ou, dans un cas, décalé d'un tour d'hélice (bloc 5, hélice 6 sur la figure 6.17). La question que nous pouvons nous poser par rapport à ce résultat de CLUSTAL W concerne une observation que nous avions faite dans le chapitre 4 à propos des méthodes globales par rapport aux locales. Nous avions indiqué alors (section 4.2.1) que certains auteurs (notamment Posfai et Barton) considèrent important de comparer les séquences globalement même lorsque ce sont les similarités locales qui sont recherchées. Cette expérience semble être un argument en faveur de cette opinion. L'illustration de la section suivante le sera également, mais nous verrons avec les motifs 'Helix-Turn-Helix' que l'approche défendue par Posfai et Barton peut ne pas être appropriée dans tous les cas de figure.

6.3.4 Les globines

6.3.4.1 Introduction

Les hémoglobines et les myoglobines (abrégées en globines) sont les principaux transporteurs d'oxygène à l'intérieur des cellules des vertébrés, la première protéine circulant dans le sang et la seconde étant localisée dans les muscles. La capacité des deux protéines à fixer l'oxygène dépend de la présence d'un groupe hème dont le ligand proximal est cette fois une histidine (H) et non plus une cystéine comme dans le cas des P450.

Comme pour les P450, une étude des globines qui ont pu être cristallisées a montré que celles-ci sont très conservées en termes de structure alors que les séquences ne présentent qu'une très faible ressemblance entre elles. En effet, deux acides aminés seulement sont connus pour être absolument conservés dans toutes les séquences (dont l'histidine liée à l'hème), et le pourcentage de résidus identiques entre paires de séquences alignées n'excède en moyenne pas 16% [Bashford *et al.*, 1987] [Lesk and Chothia, 1980]. En 1987, Bashford *et al.* ont réalisé une étude de la variation des acides aminés situés aux positions correspondant aux éléments de structure secondaire conservés dans un jeu de 7 globines (celles qui avaient pu être cristallisées à l'époque). Ces éléments correspondent à 8 hélices notées de A à H. À partir de cette étude, ils ont pu construire deux 'super-motifs', composés chacun d'une suite de six mots lexicaux. Chacun de ces mots est constitué soit des acides aminés réellement observés à ces positions, soit des acides aminés qui, bien que non observés, auraient pu se placer à une de ces positions sans que la structure sans trouve modifiée parce qu'ils sont compatibles en termes de taille (volume), caractère hydrophobe et polarité. Ils ont ensuite criblé la banque de séquences disponibles (composée alors de 3512 séquences dont 226 globines) en recherchant chacun de ces deux 'super-motifs' avec une méthode similaire à celle des profils de Gribskov (sans trou) [Gribskov *et al.*, 1988] [Gribskov *et al.*, 1990] [Gribskov *et al.*, 1987] ou de Waterman [Goldstein and Waterman, 1994] que nous utilisons dans CoSamp-X. Dans leur cas cependant, le profil est constitué d'une suite de profils. Ils ont pu montrer ainsi que les 'super-motifs' permettaient de repérer les 226 globines connues avec un très faible nombre de faux-positifs. Ces faux-positifs sont pour la plupart des polyprotéines qui, à cause de leur longueur et de leur nature, finissent par ressembler à presque tout. La construction de ces 'super-motifs' dépendait cependant de la réalisation d'un alignement préalable des séquences, qui lui-même était obtenu à partir d'un alignement manuel des structures. Ultérieurement, de tels alignements ont pu être réalisés de manière quasi-automatique [Gerstein and Altman, 1995] ou automatique [Escalier *et al.*, 1996], soit sur les même sept globines de Bashford [Escalier *et al.*, 1996], soit sur ces globines plus une huitième cristallisée après 1987 [Gerstein and Altman, 1995]. Les deux algorithmes permettent ainsi de localiser les fragments de structure tertiaire des globines les plus conservés et qui forment le noyau structural de ces protéines. Les résultats obtenus par Gerstein et Escalier sont essentiellement les mêmes, sauf en ce qui concerne l'hélice appelée F qui fixe l'hème et que Gerstein estime non conservée. Il considère en effet que les résidus composant cette hélice n'occupent pas toujours strictement les mêmes positions relatives par rapport aux résidus des autres éléments structuraux du noyau lorsque les huit globines de son jeu sont mises en correspondance. Par contre, Gerstein aussi bien qu'Escalier considèrent que l'hélice D ne fait pas partie du noyau.

Encore une fois, notre but avec cet exemple a été de voir si nos algorithmes étaient capables de localiser les mots lexicaux correspondant aux motifs structuraux du noyau à partir des séquences uniquement et sans alignement préalable de celles-ci. La méthode employée a été la même que pour les P450. Cette fois cependant, le jeu d'essai comprennait plus de séquences puisque nous avons choisi comme ensemble d'exemples six des sept globines de Bashford (séquences extraites des entrées PDB [Bernstein *et al.*, 1977] 2lhb, 2hgb et 1ecb (hémoglobines), 1mdb (myoglobine) et 2hhb (hémoglobines, chaînes A et B)) et utilisé comme séquence test soit la huitième globine cristallisée après 1987 (séquence extraite de l'entrée PDB 1mba (hémoglobine)), soit les mêmes six séquences prises comme exemples. Une septième séquence de Bashford (séquence extraite de l'entrée PDB 2lh4 (leghémoglobine)) a été laissée de côté dans ces premiers essais car elle pose problème même dans un alignement structural [Escalier *et al.*, 1996].

Un alignement des six séquences exemples est montré dans la figure 6.24. Celle-ci indique l'emplacement des six fragments de structure secondaire considérés comme faisant partie du noyau par Gerstein et/ou Escalier. Le même alignement est montré dans la figure 6.25 montrant cette fois l'emplacement des structures secondaires (correspondant aux 8 hélices) telles que celles-ci sont présentées dans l'article de Bashford. La figure 6.26 montre l'arrangement de ces motifs dans l'espace pour la structure de la chaîne α de l'hémoglobine humaine. Enfin, le tableau 6.6 fournit la liste des motifs structuraux trouvés par Escalier et les positions qu'ils recouvrent pour les six globines exemples. Il est important d'observer ici que, bien que chaque motif du noyau forme une suite contiguë de résidus, ce n'est pas ainsi qu'Escalier repère ces motifs. Son algorithme, appelé 'mescan', permet en fait d'identifier de vraies formes tridimensionnelles répétés parmi un ensemble de macromolécules. Ses blocs structuraux peuvent ainsi comporter des résidus non contigus sur la chaîne. Les numéros des blocs dans le tableau 6.6 ne correspondent donc pas aux blocs identifiés par mescan mais aux éléments de structures secondaires contenus dans les blocs trouvés par mescan.

Observons enfin que les blocs du noyau des globines sont plus purement structuraux que dans le cas des P450. L'hélice F semble ainsi être la seule liée à l'activité de la protéine.

6.3.4.2 Résultats

La présentation des résultats obtenus avec les globines suit celle adoptée pour les P450. Comme auparavant, nous donnons d'abord ceux produits par CoSamp, puis les résultats de CLUSTAL W et MACAW + Gibbs.

Les résultats de CoSamp représentent trois essais qui peuvent être rangés en deux groupes. Cette fois nous n'avons fait usage que de LeCombi en première phase de CoSamp. Les paramètres

```
1mbd     VLS---------EGEWQLVLHVWAKVEADVAG--HGQDILIRLFKSHPETLEKFDRF
2lhb     PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYET--SGVDILVKFFTSTPAAQEFFPKF
2hbg     GLS---------AAQRQVIAATWKDIAGADNGAGVGKKCLIKFLSAHPQMAAVFGFS
1ecd     LS----------ADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQF
2hhbA    VLS---------PADKTNVKAAWGKVGAHAGE--YGAEALERMFLSFPTTKTYFPHF
2hhbB    VHLT--------PEEKSAVTALWGKVNVDE----VGGEALGRLLVVYPWTQRFFESF

1mbd     KHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH---HEAELKPLAQSHATKHKIP
2lhb     KGLTTADELKKSADVRWHAERIINAVDDAVASMDDTEKMSMKLRNLSGKHAKSFQVD
2hbg     GASDPG--------VAALGAKVLAQIGVAVSHLGDEGKMVAQMKAVGVRHKGYGNKH
1ecd     AGKDLESIKGTAP-FETHANRIVGFFSKIIGELPN---IEADVNTFVASHKPRGVTH
2hhbA    DLSHGSAQ------VKGHGKKVADALTNAVAHVDD---MPNALSALSDLHAHKLRVD
2hhbB    GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN---LKGTFATLSELHCDKLHVD

1mbd     IK--YLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
2lhb     PE--YFKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY
2hbg     IKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS
1ecd     D---QLNNFRAGFVSYMKAHTDFAGA---EAAWGATLDTFFGMIFSKM
2hhbA    PV--NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
2hhbB    PE--NFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

Figure 6.24: Éléments de la structure secondaire des globines faisant partie du noyau. Les boîtes en gris foncé, moyen ou clair ou en blanc indiquent les blocs 3D trouvés par Escalier [Escalier *et al.*, 1996] (4 motifs repérés, d'abord le gris foncé, puis le moyen, le clair et enfin le blanc). Les traits noirs au-dessus des séquences indiquent les blocs de Gerstein.

Bloc	Hélice	Position 1mbd	Position 2lhb	Position 2hbg	Position 1ecd	Position 2hhbA	Position 2hhbB	Long. du bloc
1	A	4-17	13-26	4-17	3-16	4-17	5-18	14
2	B-C	24-43	33-52	26-45	19-38	24-43	23-42	20
3	E	58-72	70-84	55-69	55-69	55-69	60-74	15
4	F	79-91	94-106	79-98	76-88	76-88	81-93	13
5	G	100-107	115-122	102-109	96-103	97-104	102-109	8
6	H	125-140	131-146	127-142	118-133	122-137	127-142	16

Tableau 6.6: Blocs structuraux trouvés par V. Escalier [Escalier *et al.*, 1996] sur les globines.

```
                         hélice A             hélice B     hé-
1mbd   vl---------SEGEWQLVLHVWAKVEa--DVAGHGQDILIRLFKShpet
2lhb   pivdtgsvaplSAAEKTKIRSAWAPVYs--TYETSGVDILVKFFTStpaa
2hbg   gl---------SAAQRQVIAATWKDIAgadNGAGVGKKCLIKFLSAhpqm
1ecd   l----------SADQISTVQASFDKVKg--    DPVGILYAVFKAdpsi
2hhbA  vl---------SPADKTNVKAAWGKVGa--HAGEYGAEALERMFLSfptt
2hhbB  vhl--------TPEEKSAVTALWGKV ---NVDEVGGEALGRLLVVypwt

       lices C/D               hélice E
1mbd   lekFdrFkHlkteaemkaSEDLKKHGVTVLTALGAILKkkghheae----
2lhb   qefFpkFkGlttadelkkSADVRWHAERIINAVDDAVAsmddtekmsmk-
2hbg   aavFg-FsGas-------DPGVAALGAKVLAQIGVAVShlgdegkmvaq-
1ecd   makFtqFaGkdlesikg-TAPFETHANRIVGFFSKIIGelpniead----
2hhbA  ktyFphF-Dlshg-----SAQVKGHGKKVADALTNAVAhvddmpna----
2hhbB  qrfFesFgDlstpdavmgNPKVKAHGKKVLGAFSDGLAhldnlkgt----

         hélice F               hélice G
1mbd   -LKPLAQSHATkhki--PIKYLEFISEAIIHVLHSRhpgdfgadaQGAMN
2lhb   -LRNLSGKHAKsfqv--DPEYFKVLAAVIADTVAAG---------DAGFE
2hbg   -MKAVGVRHKGygnkhiKAQYFEPLGASLLSAMEHRiggkmnaaaKDAWA
1ecd   -VNTFVASHKPrgv---THDQLNNFRAGFVSYMKAHtdfaga---EAAWG
2hhbA  -LSALSDLHAHklrv--DPVNFKLLSHCLLVTLAAHlpaeftpavHASLD
2hhbB  -FATLSELHCDklhv--DPENFRLLGNVLVCVLAHHfgkeftppvQAAYQ

         hélice H
1mbd   KALELFRKDIAAKYkelgyqg
2lhb   KLMSMICILLRSAY
2hbg   AAYADISGALISGLqs
1ecd   ATLDTFFGMIFSKM
2hhbA  KFLASVSTVLTSKYr
2hhbB  KVVAGVANALAHKYh
```

Figure 6.25: Éléments de la structure secondaire des globines : huit hélices localisées manuellement par Bashford [Bashford *et al.*, 1987].

Figure 6.26: Structure de la chaîne α de l'hémoglobine humaine.

utilisés ainsi que l'indication des ensembles tests employés est donnée ci-dessous :

Groupe 1

- Séquences exemples : 2hhbA, 2hhbB, 2lhb, 1mdb, 2hgb, 1ecb
- Séquences tests :

  essai 1 1mba

  essai 2 les mêmes séquences prises comme exemples
- Algorithme utilisé en première phase : LeCombi
- Paramètres :
  - quorum : 65%
  - nombre de jokers : 8
  - nombre de substitutions autorisées : 0
  - longueur des modèles : 12
- Couverture pondérée restreinte :

$$CCPR = \left\{ \begin{array}{l} (\{X\}, \infty)\ \forall X \in C, \\ (\Sigma, 8), \\ (S, 0)\ pour\ tous\ les\ autres\ ensembles \end{array} \right\}$$

  avec :

$$C = \left\{ \begin{array}{ll} \{I, L, M, V\} & (hydrophobe), \\ \{A, G\} & (très\ petit), \\ \{A, S, T\} & (petit), \\ \{C\} & (cystéine), \\ \{F, Y, W\} & (aromatique), \\ \{K, H, R\} & (basique), \\ \{E, D\} & (acide), \\ \{N, Q\} & (glutamine\ et\ glutamate), \\ \{G, P\} & (apparaît\ dans\ les\ structures\ contraintes) \end{array} \right\}$$

- Temps d'exécution de LeCombi : 10 secondes (sur une Sparc Station 20)
- Nombre de modèles initiaux : 850 représentant en fait 666 ensembles d'occurrences distincts

Groupe 2

- Séquences exemples : 2hhbA, 2hhbB, 2lhb, 1mdb, 2hgb, 1ecb
- Séquences tests : les mêmes séquences exemples
- Algorithme utilisé en première phase : LeCombi
- Paramètres :
  - quorum : 100%
  - nombre de jokers : 8
  - nombre de substitutions autorisées : 1

```
score 17.2 p-value -7.11   bloc 1 Hélice A
1mbd   3      SEGEWQLVLHVW
2lhb   12     SAAEKTKIRSAW
2hbg   3      SAAQRQVIAATW
1ecb   110    AHTDFAGAEAAW  faux
2hhbA  3      SPADKTNVKAAW
2hhbB  4      TPEEKSAVTALW
-
score 18.2 p-value -6.89   bloc 3 Hélice E
1mbd   57     ASEDLKKHGVTV
2lhb   67     KSADVRWHAERI
2hbg   51     SDPGVAALGAKV
1ecp   51     GTAPFETHANRI
2hhbA  51     GSAQVKGHGKKV
2hhbB  56     GNPKVKAHGKKV
-
score 13.7 p-value -5.44   bloc 4 Hélice F
1mbd   82     HEAELKPLAQSH
2lhb   95     MSMKLRNLSGKH
2hbg   79     MVAQMKAVGVRH
1ecb   76     IEADVNTFVASH
2hhbA  76     MPNALSALSDLH
2hhbB  81     LKGTFATLSELH
-
score 12.2 p-value -3.87   bloc 5 Hélice G
1mbd   104    LEFISEAIIHVL
2lhb   117    FKVLAAVIADTV  décalé +3
2hbg   103    FEPLGASLLSAM
1ecb   38     FTQFAGKDLESI  faux
2hhbA  98     FKLLSHCLLVTL
2hhbB  103    FRLLGNVLVCVL
-
```

Figure 6.27: Résultat première série — premier essai globines (LeCombi, six séquences, quorum de 65%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

- longueur des modèles : 14

- Couverture pondérée restreinte : la même que pour le Groupe 2
- Temps d'exécution de LeCombi : 783 secondes (sur une Sparc Station 20)
- Nombre de modèles initiaux : 2471 représentant 1889 ensembles d'occurrences distincts

Les figures 6.27 et 6.28 correspondent au groupe 1, la figure 6.29 correspond au groupe 2.

Les figures 6.30 et 6.31 quant à elles présentent les résultats obtenus avec CLUSTAL W et la méthode d'échantillonnage de Gibbs intégrée dans MACAW. Dans le cas de ce dernier, la longueur des blocs recherchés a varié entre 4 et 12. Le jeu d'essai pour les deux algorithmes est le même (six séquences) que celui utilisé dans CoSamp.

6.3.4.3 Discussion

Les commentaires généraux que nous pouvons faire concernant les résultats observés avec CoSamp sur les globines demeurent essentiellement les mêmes que ceux relatifs aux P450.

```
score 24.3 p-value -10.94  bloc 1 Hélice A
1mbd   3      SEGEWQLVLHVW
2lhb   12     SAAEKTKIRSAW
2hbg   3      SAAQRQVIAATW
1ecb   110    AHTDFAGAEAAW  faux
2hhbA  3      SPADKTNVKAAW
2hhbB  4      TPEEKSAVTALW
-
score 22.1 p-value -10.34  bloc 2 et au-delà
                           Hélices C et D
1mbd   35     SHPETLEKFDRF
2lhb   45     STPAAQEFFPKF
2hbg   37     AHPQMAAVFGFS
1ecb   30     ADPSIMAKFTQF
2hhbA  35     SFPTTKTYFPHF
2hhbB  34     VYPWTQRFFESF
-
score 21.9 p-value -9.65   bloc 3 Hélice E
1mbd   56     KASEDLKKHGVT
2lhb   66     KKSADVRWHAER
2hbg   78     KMVAQMKAVGVR  faux
1ecb   50     KGTAPFETHANR
2hhbA  50     HGSAQVKGHGKK
2hhbB  55     MGNPKVKAHGKK
-
score 19.9 p-value -8.81   bloc 4 Hélice F
1mbd   90     AQSHATKHKIPI
2lhb   103    SGKHAKSFQVDP
2hbg   94     GNKHIKAQYFEP  décalé de +6
1ecb   9      VQASFDKVKGDP  faux
2hhbA  84     SDLHAHKLRVDP
2hhbB  89     SELHCDKLHVDP
-
score 20.7 p-value -7.60   bloc 3 Hélice E
1mbd   103    YLEFISEAIIHV  faux
2lhb   78     IINAVDDAVASM
2hbg   62     VLAQIGVAVSHL
1ecb   62     IVGFFSKIIGEL
2hhbA  62     VADALTNAVAHV
2hhbB  67     VLGAFSDGLAHL
-
score 14.8 p-value -6.27   bloc 6 Hélice H
1mbd   15     AKVEADVAGHGQ  faux
2lhb   122    AVIADTVAAGDA  décalé de -11
2hbg   135    ADISGALISGLQ  décalé de +7
1ecb   78     ADVNTFVASHKP  faux
2hhbA  123    ASLDKFLASVST
2hhbB  128    AAYQKVVAGVAN
-
```

Figure 6.28: Résultat première série — deuxième essai globines (LeCombi, six séquences, quorum de 65%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
score 28.8 p-value -11.68  bloc 1 Hélice A
1mbd   2      LSEGEWQLVLHVWA
2lhb   11     LSAAEKTKIRSAWA
2hbg   2      LSAAQRQVIAATWK
1ecb   1      LSADQISTVQASFD
2hhbA  2      LSPADKTNVKAAWG
2hhbB  3      LTPEEKSAVTALWG
-
score 25.9 p-value -10.47  bloc 3 Hélice E
1mbd   58     SEDLKKHGVTVLTA
2lhb   68     SADVRWHAERIINA
2hbg   87     GVRHKGYGNKHIKA  faux
1ecb   52     TAPFETHANRIVGF
2hhbA  52     SAQVKGHGKKVADA
2hhbB  57     NPKVKAHGKKVLGA
-
score 22.5 p-value -10.45  bloc 2 Hélices C et D
1mbd   35     SHPETLEKFDRFKH
2lhb   45     STPAAQEFFPKFKG
2hbg   37     AHPQMAAVFGFSGA
1ecb   30     ADPSIMAKFTQFAG
2hhbA  35     SFPTTKTYFPHFDL
2hhbB  34     VYPWTQRFFESFGD
-
score 20.0 p-value -8.65   bloc 6 Hélice H
1mbd   125    ADAQGAMNKALELF
2lhb   125    ADTVAAGDAGFEKL  décalé de -3
2hbg   124    AAAKDAWAAAYADI
1ecb   115    AGAEAAWGATLDTF
2hhbA  115    AEFTPAVHASLDKF  décalé de -4
2hhbB  120    KEFTPPVQAAYQKV  décalé de -4
-
score 18.3 p-value -7.81   bloc 4 Hélice F
1mbd   81     HHEAELKPLAQSHA
2lhb   94     KMSMKLRNLSGKHA
2hbg   72     HLGDEGKMVAQMKA  décalé de -6
1ecb   75     NIEADVNTFVASHK
2hhbA  75     DMPNALSALSDLHA
2hhbB  80     NLKGTFATLSELHC
-
```

Figure 6.29: Résultat deuxième série — globines (LeCombi, six séquences, quorum de 100%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
                             BLOC 1                    BLOC 2
1mbd          ---------VLSEGEWQLVLHVWAKVE-A-DVAGHGQDILIRLFKSHPETLEKFDRFKHL
2lhb          PIVDTGSVAPLSAAEKTKIRSAWAPVY-SXXYETSGVDILVKFFTSTPAAQEFFPKFKGL
2hbg          ---------GLSAAQRQVIAATWKDIAGADNGAGVGKKCLIKFLSAHPQMAAVFGFSG--
1ecd          ----------LSADQISTVQASFDKVK-----GDP-VGILYAVFKADPSIMAKFTQFAG-
2hhbA         ---------VLSPADKTNVKAAWGKVG-A-HAGEYGAEALERMFLSFPTTKTYFPHF---
2hhbB         --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESFGDL
                        *.      .      .             *       *     *

                             BLOC 3                  BLOC 4
1mbd          KTEAEMKASEDLKKHGVTVLTALGAILKKKG---HHEAELKPLAQSHA--TKHKIPIKYL déc-1
2lhb          TTADELKKSADVRWHAERIINAVDDAVASMDDTEKMSMKLRNLSGKHA--KSFQVDPEYF
2hbg          ----ASD--PGVAALGAKVLAQIGVAVSHLGDEGKMVAQMKAVGVRHKGYGNKHIKAQYF
1ecd          KDLESIKGTAPFETHANRIVGFFSKIIGELP---NIEADVNTFVASHK---PRGVTHDQL
2hhbA         ---DLSHGSAQVKGHGKKVADALTNAVAHVD---DMPNALSALSDLHA--HKLRVDPVNF
2hhbB         STPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLKGTFATLSELHC--DKLHVDPENF
                              .       .                   *       .

               BLOC 5                       BLOC 6
1mbd          EFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
2lhb          KVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
2hbg          EPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
1ecd          NNFRAGFVSYMKAHT--DFAG-AEAAWGATLDTFFGMIFSKM-------
2hhbA         KLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
2hhbB         RLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
                        .                          .
```

Figure 6.30: Alignement multiple global obtenu avec CLUSTAL W sur les six globines de notre jeu d'essai. L'indication des blocs structuraux correctement alignés a été réalisée par nous, *a posteriori*. Une étoile en-dessous d'une colonne indique la présence de résidus identiques à cet endroit de l'alignement, un point la présence de résidus équivalents.

```
                    BLOC 1                       BLOC 2
1mbd   v---------LSEGEWQLvlhvwakveadvaghgq--DILIRLFKSHPetlekfdrfkhl
2lhb   pivdtgsvapLSAAEKTKirsawapvysxxyetsgv-DILVKFFTSTPaaqeffpkfkgl
2hbg   g---------LSAAQRQViaatwkdiagadngagvgkKCLIKFLSAHPqmaavfgfsgas
1ecd   ----------LSADQISTvqasfdkvkgdpv------GILYAVFKADPsimakftqfagk
2hhba  v---------LSPADKTNvkaawgkvgahageyga--EALERMFLSFPttktyfphfdls
2hhbb  vh--------LTPEEKSAvtalwgkvnvdevggealgrllvvypwtqrffesfgdlstpd

                                                           BLOC 5
1mbd   kteaemkasedlkkhgvtvltalgailkkkghheaelkplaqshatkhki-PIKYLEFis
2lhb   ttadelkksadvrwhaeriinavddavasmddtekmsmklrnlsgkhaksfQVDPEYFkv
2hbg   dpgvaalgakvlaqigvavshlgdegkmvaqmkavgvrhkgygnk------HIKAQYFep
1ecd   dlesikgtapfethanrivgffskiigelpnieadvntfvashkprgvt--HDQLNNFra
2hhba   hgsaqvkghgkkvadaltnavahvddmpnalsalsdlhahkl-------RVDPVNFkl
2hhbb  avmgnpkvkahgkkvlgafsdglahldnlkgtfatlselhcdkl------HVDPENFrl

1mbd   eaiihvlhsrhpgdfgadaqgamnkalelfrkdiaakykelgyqg
2lhb   laaviadtvaagdagfeklmsmicillrsay--------------
2hbg   lgasllsamehriggkmnaaakdawaaayadisgalisglqs---
1ecd   gfvsymkahtdfagaeaawgatldtffgmifskm-----------
2hhba  lshcllvtlaahlpaeftpavhasldkflasvstvltskyr----
2hhbb  lgnvlvcvlahhfgkeftppvqaayqkvvagvanalahkyh----
```

Figure 6.31: Blocs trouvés par la méthode d'échantillonnage de Gibbs sur les six globines de notre jeu d'essai.

Une observation supplémentaire importante concerne la 'meilleure qualité' des modèles trouvés par CoSamp dans ce cas. Nous proposons quelques explications à cela, mais il faut noter que celles-ci sont pour l'instant purement intuitives. La première est que les blocs structuraux à identifier en tant que blocs lexicaux sont de longueurs plus petites et surtout plus homogènes. Une seconde explication est que nous disposons dans ce cas d'un ensemble plus grand de séquences. Ceci nous permet de travailler avec un quorum relativement bas (65%, c'est-à-dire $q \geq 4$). L'algorithme de la première phase peut alors plus facilement localiser les blocs lexicaux correspondant aux blocs structuraux sur les q séquences de l'ensemble qui sont les plus proches les unes des autres. Il lui suffit ensuite de retrouver les occurrences sur les séquences manquantes, mais arrivé à ce stade, il dispose d'un alignement de q mots qui contient plus d'information que le simple nom du modèle. Cette identification plus aisée des blocs sur les séquences d'un ensemble qui sont les plus proches suggère qu'il pourrait être intéressant d'introduire dans le cadre d'une recherche de mots communs l'idée d'une comparaison ayant pour base un arbre de classification. De telles méthodes existent, surtout dans le cas global, mais ainsi que nous l'avons vu dans le chapitre 4, l'arbre utilisé est toujours, soit connu dès le départ, soit construit avant l'alignement. Signalons que dans le cas de CLUSTAL W, l'arbre construit ne réussit à regrouper que 2 des 6 séquences du jeu. La raison en est probablement que ces séquences sont trop peu conservées pour que des comparaisons deux à deux (les seules pouvant servir de base à l'établissement de l'arbre avant l'alignement multiple) puissent être vraiment utiles. Construire l'arbre et l'alignement en même temps est bien sûr un problème considérablement plus difficile, mais cet exemple est peut-être une illustration de son importance.

À nouveau ici, l'alignement global réalisé par CLUSTAL W positionne en partie correctement les six blocs considérés comme appartenant au noyau structural des globines par Escalier, ainsi que l'hélice D que CoSamp identifie également. Par contre aussi bien CLUSTAL

W que CoSamp placent incorrectement un résidu de cette hélice : d'après Bashford, le bloc contient en effet un trou en avant-dernière position sur la séquence 2hgb, trou qu'aucun des deux algorithmes n'est capable de placer correctement (LeCombi autorise sur cet exemple une substitution mais pas de trou). D'une façon générale, l'alignement de CLUSTAL W présente plusieurs différences importantes avec l'alignement structural global réalisé par Bashford [Bashford *et al.*, 1987].

6.3.5 Les HTH ('Helix-Turn-Helix')

6.3.5.1 Introduction

L'exemple traité dans cette section est le dernier que nous présentons dans la série des protéines qui possèdent une forte conservation au niveau de leurs structures, mais très faible au niveau de leurs séquences.

Les protéines avec lesquelles nous travaillons ici appartiennent en fait à des familles souvent très diverses dont le seul point commun est leur capacité à se fixer à l'ADN (la CRP dont nous avons parlé dans la section 6.2.3 en est un exemple parmi d'autres). Pour que cette fixation puisse avoir lieu, il faut que la protéine reconnaisse un site particulier sur la macromolécule d'ADN. Rappelons que celle-ci est formée de deux brins anti-parallèles composés de bases complémentaires qui s'apparient à l'intérieur de l'hélice. La face externe de cette double hélice ne porte donc pratiquement aucune trace de l'identité des nucléotides situés à l'intérieur puisqu'elle constituée essentiellement du squelette phosphate-sucre. Pour que la protéine puisse établir un contact avec les bases nucléotidiques, il faut donc que la structure locale de l'ADN soit déformée (coude, ouverture) ou que la protéine adopte une conformation lui permettant de pénétrer à l'intérieur du grand sillon de l'ADN. Dans le cas présent, cette forme lui est conférée par un motif structural composé d'une hélice suivie par un coude puis par une autre hélice (voir la figure 6.32) et appelé un motif HTH (de l'anglais 'Helix Turn Helix').

Comme ces protéines sont d'origines diverses, elles ne présentent souvent aucune autre ressemblance structurale autre que ce motif. Bien sûr, elles ne possèdent que peu, ou pas du tout, de ressemblance en termes de séquence, sauf éventuellement au niveau du coude du motif HTH. En effet, ce dernier est en général formé d'acides aminés hydrophobes. Observons cependant que cette relative similarité ne concerne que 3 ou 4 résidus. C'est toutefois ce coude qui, la plupart du temps, permet d'identifier le motif à partir des séquences uniquement. La spécificité de la reconnaissance du site sur l'ADN par la protéine est quant à elle liée principalement à des différences situées dans les chaînes latérales du début de la seconde hélice. Cette région est appelée l'hélice de reconnaissance [Creighton, 1993]. L'autre hélice établit un contact non-spécifique avec l'ADN.

Le jeu d'essai que nous avons adopté afin de tester la capacité de nos algorithmes à repérer de faibles similarités est celui employé par Lawrence, dans son article sur la méthode d'échantillonnage de Gibbs [Lawrence *et al.*, 1993]. Ce jeu est composé de 30 séquences (dont les numéros d'accès dans la banque NFBR/PIR sont : JEBY1, DNECFS, B26499, BVECDA, A32837, C29010, A25944, A25399, QCBP2L, C55249, RPECL, RPECDU, JWEC, RGKBAP, RPECW, RPECTN, Z1BPC2, JU0083, S11945, RPECEG, RGECH, RPECCT, RPECG, RGECF, QRECC, RGECA, RGKBCP, RPECDO, RCBPL, RGBP22). La longueur du motif structural est d'environ 18 acides aminés et son emplacement sur chacune des séquences, déterminé par l'algorithme de Lawrence, est en accord avec les données expérimentales ou avec un examen direct de la structure de la protéine lorsque celle-ci a pu être cristallisée. La figure 6.33 montre un alignement des séquences à l'endroit du motif ainsi que les positions de début et de fin de ce dernier.

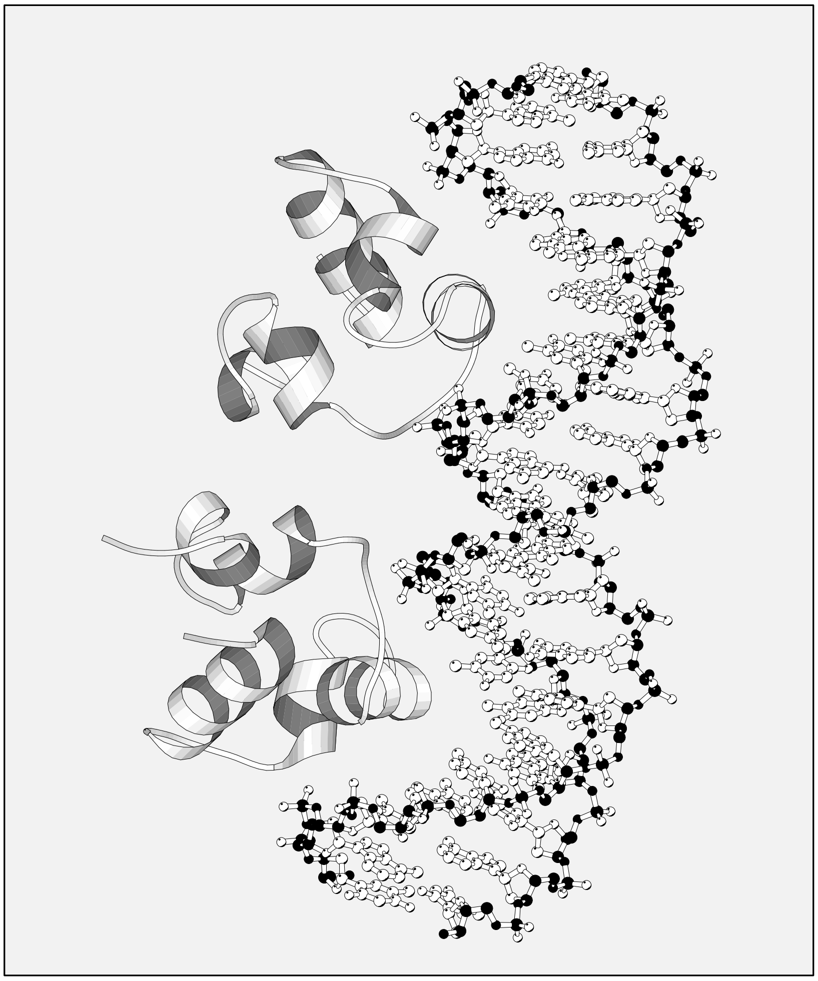

Figure 6.32: Structure hélice-coude-hélice (protéine Cro) avec son ADN.

	Noms dans PIR	Pos.	Motif	Noms 'usuels'
1	JEBY1	100	KEEVAKKCGITPLQVRVW	MAT a1
2	DNECFS	75	QTRAALMMGINRGTLRKK	Fis
3	B26499	450	QIRAADLLGLNRNTLRKK	Ntrc (Brady.)
4	BVECDA	23	QRSLAKALKISHVSVSQW	DicA
5	A32837	23	VSITAENLGLTQPAVSNA	Nahr
6	C29010	6	VSRLALDAGVSVHIVRDY	MerD
7	A25944	224	QKETGDILGISQMHVSRL	Sigma-37
8	A25399	327	RIEIAHALCLTERQIKIW	Antennapedia
9	QCBP2L	26	TEKTAEAVGVDKSQISRW	Lambda cII
10	C55249	95	QREIAKELGISRSYVSRI	SpoIIIC
11	RPEC	6	LYDVAEYAGVSYQTVSRV	LacI
12	RPECDU	4	IKDVAKRANVSTTTVSHV	PurR
13	JWE	161	RQKVAIIYDVGVSTLYKR	Pin
14	RGKBAP	496	QAKAARLLGMTPRQVAYR	NifA
15	RPECW	68	QRELKNELGAGIATITRG	TrpR
16	RPECTN	27	TRKLAQKLGVEQPTLYWH	TetR
17	Z1BPC2	26	QRKVADALGINESQISRW	P22 cI
18	JU0083	206	QKDVADMMGISQSYISRL	SpoIIG
19	S11945	28	RAEIAQRLGFRSPNAAEE	LexA
20	RPECEG	4	LKDIAIEAGVSLATVSRV	EbgR
21	RGECH	253	LQELADRYGVSAERVRQL	HtpR
22	RPECCT	12	MKDVALKAKVSTATVSRA	CytR
23	RPECG	4	IKDVARLAGVSVATVSRV	GalR
24	RGECF	197	RGDIGNYLGLTVETISRL	FnR
25	QRECC	170	RQEIGQIVGCSRETVGRI	Crp (CAP)
26	RGECA	197	IASVAQHVCLSPSRLSHL	AraC
27	RGKBCP	445	KQEAARLLGWGRNTLTRK	Ntrc (K.a.)
28	RPECDO	24	LKDAAALLGVSEMTIRRD	DeoR
29	RCBPL	16	QTKTAKDLGVYQSAINKA	Lambda Cro
30	RGBP22	13	QRAVAKALGISDAAVSQW	P22 Cro

Figure 6.33: Alignement des trente séquences du jeu à l'endroit du motif HTH.

Notons que Lawrence a choisi de construire un jeu d'essai duquel ont été éliminées toutes les séquences considérées trop proches les unes des autres.

Voyons maintenant les résultats obtenus avec CoSamp, ainsi qu'avec l'algorithme de Lawrence (MACAW + Gibbs) et avec CLUSTAL W.

6.3.5.2 Résultats

Comme l'ensemble des séquences du jeu d'essai est bien plus grand que dans le cas des deux illustrations précédentes, et comme l'identification du motif HTH en tant que motif lexical est connu pour être difficile, nous avons décidé initialement de fixer le quorum à une valeur très basse : 17%. Un modèle n'a donc besoin d'être présent que dans 5 parmi les 30 séquences de l'ensemble pour être valide. Nous avons également choisi de travailler avec LeCombi exclusivement. Enfin, dans un second temps, nous avons augmenté le quorum à 60%. Les résultats produits par CoSamp peuvent ainsi être rangés en deux groupes suivant la valeur attribuée à ce quorum, le premier contient deux essais et le second groupe un seul essai. Les paramètres correspondant à chacun des groupes sont les suivants :

Groupe 1

- Séquences exemples : les 30 séquences sélectionnées par Lawrence
- Séquences tests : les mêmes séquences
- Algorithme utilisé en première phase : LeCombi
- Paramètres :

  essai 1

  - quorum : 17%
  - nombre de jokers : 8
  - nombre de substitutions autorisées : 0
  - longueur des modèles : 16

  essai 2

  - quorum : 17%
  - nombre de jokers : 5
  - nombre de substitutions autorisées : 1
  - longueur des modèles : 16
- Couverture pondérée restreinte :

$$CCPR = \left\{ \begin{array}{l} (\{X\}, \infty) \ \forall X \in C, \\ (\Sigma, 8), \\ (S, 0) \ pour\ tous\ les\ autres\ ensembles \end{array} \right\}$$

avec :

$$C = \left\{ \begin{array}{ll} \{I, L, M, V\} & (hydrophobe), \\ \{A, G\} & (très\ petit), \\ \{A, S, T\} & (petit), \\ \{C\} & (cystéine), \\ \{F, Y, W\} & (aromatique), \\ \{K, H, R\} & (basique), \\ \{E, D\} & (acide), \\ \{N, Q\} & (glutamine\ et\ glutamate), \\ \{G, P\} & (apparaît\ dans\ les\ structures\ contraintes) \end{array} \right\}$$

- Temps d'exécution de LeCombi :

  essai 1 232 secondes (sur une Sparc Station 20)

  essai 2 11866 secondes (*idem*)

- Nombre de modèles initiaux :

  essai 1 875 représentant 77 ensembles d'occurrences distincts

  essai 2 6642 représentant 139 ensembles d'occurrences distincts

Groupe 2

- Séquences exemples : les mêmes que pour le groupe 1
- Séquences tests : les mêmes séquences
- Algorithme utilisé en première phase : LeCombi
- Paramètres :
  - quorum : 60%
  - nombre de jokers : 8
  - nombre de substitutions autorisées : 1
  - longueur des modèles : longueur maximale (égale dans ce cas à 15)
- Couverture pondérée restreinte : la même que pour le groupe 1
- Temps d'exécution de LeCombi : 10923 secondes (sur une Sparc Station 20)
- Nombre de modèles initiaux : 36

Cette fois, des résultats intéressants apparaissent dès la première phase, avant même l'étape d'évaluation.

En effet, dans le premier essai, c'est-à-dire parmi les 77 modèles associés à des ensembles d'occurrences distincts obtenus en autorisant 8 jokers, 4 seulement ne sont pas associés au motif HTH. Ces quatre modèles correspondent en réalité à un autre site présent dans un sous-ensemble des protéines du jeu : il s'agit de cinq répresseurs de *coli* (*cyt*, *ebg*, *gal*, *lac* et *pur*). 15 des 72 modèles restants possèdent une seule occurrence fausse du motif HTH. Par ailleurs, 4 des 30 séquences du jeu d'essai ne présentent jamais aucune occurrence, il s'agit de : JEBY1 ('mating hormone' - *yeast*), A25399 ('homeotic protein Antennopedia' - *fruit fly*), QCBP2L ('regulatory protein cII' -*phage lambda*) et JWEC ('DNA-invertase' - *coli*). Les résultats obtenus avec le second essai, autorisant cette fois 5 jokers, et une substitution au plus, sont sensiblement les mêmes. Enfin, avec l'essai du second groupe, lorsque le quorum est fixé à 60%, le nombre de jokers est égal à 8 et au plus une substitution est autorisée, les modèles obtenus sont tous associés au motif HTH et possèdent chacun au maximum 3 occurrences 'fausses'. Seules 3 séquences sont totalement 'manquées' (pas d'occurrences ou occurrences systématiquement fausses) : JEWC, RPECW ('*trp* repressor' - *coli*) et S11945 ('*lex*A repressor' - *coli*).

À partir des résultats de l'essai 1 du premier groupe (8 jokers, 0 erreur, $q = 17\%$), nous avons alors construit une table $\mathcal{T}$ de dimension 30×30 avec $\mathcal{T}(i, j)$ indiquant le nombre de fois où un modèle est présent à la fois sur la séquence i et la séquence j pour tout i et j. Nous pouvons y distinguer (voir la figure 6.34) trois groupes différents de séquences qui se trouvent mis ensemble. Ces groupes sont les suivants :

Ensemble 1

- *cyt* 'repressor'
- *ebg* 'repressor'
- *gal* 'repressor'
- *lac* 'repressor'
- *pur* 'repressor'
- 'regulatory protein' *fnr*

Ensemble 2

- 'transcription initiation factor' sigmaK
- 'transcription initiation factor' sigmaB
- 'transcription initiation factor' sigmaE 'precursor'

Ensemble 3

- *dic*A 'protein'
- *deo* 'operon repressor'
- 'nitrogen assimilation regulatory protein'
- 'DNA binding protein' *fis*
- 'regulatory protein' *cro*

L'ensemble 1 regroupe ainsi tous les répresseurs du jeu d'essai sauf *trp*, *Tet*R et *deo* de *coli*, tandis que l'ensemble 2 réunit les facteurs intervenant dans l'initiation de la transcription. Enfin, l'ensemble 3 réunit des protéines apparemment complètement différentes et il serait intéressant de savoir si elles possèdent en fait un point commun autre que le motif HTH. Toutes les autres séquences du jeu semblent former des groupes isolés.

Lorsque l'étape d'évaluation est maintenant utilisée à la suite, le résultat devient encore plus intéressant et, d'une certaine façon, surprenant. Quel que soit l'essai (Groupe 1 ou 2), CoSamp obtient dans tous les cas un seul profil dont la p-value est inférieure à 10^{-5}. Ces profils sont composés à la fois d'occurrences originalement trouvées par LeCombi et de pseudo-occurrences ajoutées à l'ensemble lors de la seconde étape. Ils sont indiqués dans les figures 6.35, 6.36 et 6.37. Le meilleur résultat est obtenu avec le premier essai du Groupe 1 ($q = 17\%$) où seule est fausse l'occurrence trouvée sur la séquence de numéro d'accès C29010 dans PIR ('mercuric

		23	20	12	11	22	24	10	18	7	21	25	26	17	4	14	15	6	19	1	9	3	28	27	2	5	29	16	8	13
RPECG	**23**	65	61	46	37	33	19	22	22	9	4	6	4	2	3	1	2	2	1	1	1	4	3	2	0	0	0	0	0	0
RPECEG	**20**	61	61	43	35	31	18	21	21	8	4	6	4	2	2	1	2	2	1	1	0	3	3	1	0	0	0	0	0	0
RPECDU	**12**	46	43	46	25	33	8	10	9	4	1	2	4	1	3	0	1	1	1	1	1	4	2	1	0	0	0	0	0	0
RPECL	**11**	37	35	25	37	15	19	7	8	4	4	6	1	0	1	0	0	2	0	0	1	1	1	1	0	0	0	0	0	0
RPECCT	**22**	33	31	33	15	33	2	5	5	3	1	0	3	1	1	0	1	1	1	1	1	2	1	1	0	0	0	0	0	0
RGECF	**24**	19	18	8	19	2	20	8	9	4	2	7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
C55249	**10**	22	21	10	7	5	8	30	29	11	2	2	0	6	4	2	2	0	0	0	0	2	1	1	0	0	0	0	0	0
JU0083	**18**	22	21	9	8	5	9	29	30	12	2	2	0	6	3	2	2	0	0	0	0	2	1	1	0	0	0	0	0	0
A25944	**7**	9	8	4	4	3	4	11	12	14	0	1	0	1	2	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
RGECH	**21**	4	4	1	4	1	2	2	2	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
QRECC	**25**	6	6	2	6	0	7	2	2	1	0	8	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0
RGECA	**26**	4	4	4	1	3	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Z1BPC2	**17**	2	2	1	0	1	0	6	6	1	0	0	0	6	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0
BVECDA	**4**	3	2	3	1	1	0	4	3	2	0	0	0	1	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RGKBAP	**14**	1	1	0	0	0	0	2	2	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RPECW	**15**	2	2	1	0	1	0	2	2	0	0	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
C29010	**6**	2	2	1	2	1	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
S11945	**19**	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
JEBY1	**1**	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
QCBP2L	**9**	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
B26499	**3**	4	3	4	1	2	0	2	2	0	0	1	0	1	0	0	0	0	0	0	0	9	2	4	4	2	4	1	0	0
RPECDO	**28**	3	3	2	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	2	5	1	2	1	2	0	0	0
RGKBCP	**27**	2	1	1	1	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	4	1	5	2	0	2	0	0	0
DNECFS	**2**	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4	2	2	4	2	4	1	0	0
A32837	**5**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	2	2	2	1	0	0
RCBPL	**29**	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4	2	2	4	2	4	1	0	0
RPECTN	**16**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0
A25399	**8**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JWEC	**13**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6.34: Tableau indiquant le nombre de fois où un modèle est présent à la fois sur la séquence i et la séquence j pour tout i et j (8 jokers, 0 erreur, $q = 17\%$). Pour avoir les noms usuels des protéines, se reporter à la figure 6.33.

resistance operon regulatory *mer*D protein' - *Serratia marcescens plasmid*). Observons que les séquences sur lesquelles le motif n'a pas été identifié dans la première étape apparaissent ici avec des occurrences correctes. Dans le groupe 1, le profil couvre sept résidus de la première hélice, trois résidus du coude et les six premiers résidus de l'hélice de reconnaissance tandis que dans le groupe 2, les occurrences possèdent un résidu de moins.

Enfin, la figure 6.38 présente le résultat obtenu avec l'échantillonnage de Gibbs (observons que la recherche a été faite en fixant la longueur de la matrice à 18 comme dans l'article original de Lawrence). Celui de CLUSTAL W n'est pas montré ici — il est discuté dans la section suivante.

6.3.5.3 Discussion

Notre première observation est que l'utilisation de la seconde étape dans cette illustration est, d'une certaine façon, contraire au principe de CoSamp. En effet, celui-ci est constitué d'une phase d'exploration combinatoire que l'on souhaite aussi large que possible suivie par une phase d'évaluation et de classement des modèles trouvés dans la première étape. Si le quorum n'est pas égal à 100%, les ensembles d'occurrences des modèles doivent être complétés par des pseudo-occurrences mais le 'noyau' principal du modèle final reste composé des occurrences originales. Cependant, lorsque le quorum est fixé à une valeur très basse (17% dans les deux premiers exemples ici), les modèles finaux sont constitués essentiellement de telles pseudo-occurrences ce qui semble contradictoire avec la 'philosophie' originale de l'algorithme.

Par ailleurs, cet exemple suggère que d'autres approches que l'évaluation sont possibles pour la seconde étape de l'algorithme. Nous pourrions ainsi envisager de regrouper deux modèles lorsqu'ils possèdent au moins un certain nombre d'occurrences qui se chevauchent du même

```
score 17.8 p-value -6.91
JEBY1  101    EEVAKKCGITPLQVRV
DNECFS 75     TRAALMMGINRGTLRK
B26499 451    IRAADLLGLNRNTLRK
BVECDA 24     RSLAKALKISHVSVSQ
A32837 24     SITAENLGLTQPAVSN
C29010 79     DETAAQLAVLRQFVER  faux
A25944 225    KETGDILGISQMHVSR
A25399 328    IEIAHALCLTERQIKI
QCBP2L 27     EKTAEAVGVDKSQISR
C55249 96     REIAKELGISRSYVSR
RPECL  7      YDVAEYAGVSYQTVSR
RPECDU 5      KDVAKRANVSTTTVSH
JWEC   162    QKVAIIYDVGVSTLYK
RGKBAP 497    AKAARLLGMTPRQVAY
RPECW  69     RELKNELGAGIATITR
RPECTN 28     RKLAQKLGVEQPTLYW
Z1BPC2 27     RKVADALGINESQISR
JU0083 207    KDVADMMGISQSYISR
S11945 29     AEIAQRLGFRSPNAAE
RPECEG 5      KDIAIEAGVSLATVSR
RGECH  254    QELADRYGVSAERVRQ
RPECCT 13     KDVALKAKVSTATVSR
RPECG  5      KDVARLAGVSVATVSR
RGECF  198    GDIGNYLGLTVETISR
QRECC  171    QEIGQIVGCSRETVGR
RGECA  198    ASVAQHVCLSPSRLSH
RGKBCP 446    QEAARLLGWGRNTLTR
RPECDO 25     KDAAALLGVSEMTIRR
RCBPL  17     TKTAKDLGVYQSAINK
RGBP22 14     RAVAKALGISDAAVSQ
-
```

Figure 6.35: Résultat première série — premier essai motif HTH (LeCombi, trente séquences, quorum de 17%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
score 17.7 p-value -6.91
JEBY1  101    EEVAKKCGITPLQVRV
DNECFS 75     TRAALMMGINRGTLRK
B26499 451    IRAADLLGLNRNTLRK
BVECDA 24     RSLAKALKISHVSVSQ
A32837 24     SITAENLGLTQPAVSN
C29010 79     DETAAQLAVLRQFVER  faux
A25944 225    KETGDILGISQMHVSR
A25399 100    AQTNGQLGVPQQQQQQ  faux
QCBP2L 27     EKTAEAVGVDKSQISR
C55249 96     REIAKELGISRSYVSR
RPECL  7      YDVAEYAGVSYQTVSR
RPECDU 5      KDVAKRANVSTTTVSH
JWEC   162    QKVAIIYDVGVSTLYK
RGKBAP 497    AKAARLLGMTPRQVAY
RPECW  69     RELKNELGAGIATITR
RPECTN 28     RKLAQKLGVEQPTLYW
Z1BPC2 27     RKVADALGINESQISR
JU0083 207    KDVADMMGISQSYISR
S11945 29     AEIAQRLGFRSPNAAE
RPECEG 5      KDIAIEAGVSLATVSR
RGECH  254    QELADRYGVSAERVRQ
RPECCT 13     KDVALKAKVSTATVSR
RPECG  5      KDVARLAGVSVATVSR
RGECF  198    GDIGNYLGLTVETISR
QRECC  171    QEIGQIVGCSRETVGR
RGECA  198    ASVAQHVCLSPSRLSH
RGKBCP 446    QEAARLLGWGRNTLTR
RPECDO 25     KDAAALLGVSEMTIRR
RCBPL  17     TKTAKDLGVYQSAINK
RGBP22 14     RAVAKALGISDAAVSQ
-
```

Figure 6.36: Résultat première série — deuxième essai motif HTH (LeCombi, trente séquences, quorum de 17%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
score 14.8 p-value -5.84
JEBY1  101    EEVAKKCGITPLQVR
DNECFS 75     TRAALMMGINRGTLR
B26499 187    RRNGPFVAVNMAAIP  faux
BVECDA 24     RSLAKALKISHVSVS
A32837 24     SITAENLGLTQPAVS
C29010 53     VRAAFEAGIGLGALA  faux
A25944 225    KETGDILGISQMHVS
A25399 328    IEIAHALCLTERQIK
QCBP2L 27     EKTAEAVGVDKSQIS
C55249 96     REIAKELGISRSYVS
RPECL  7      YDVAEYAGVSYQTVS
RPECDU 5      KDVAKRANVSTTTVS
JWEC   162    QKVAIIYDVGVSTLY
RGKBAP 497    AKAARLLGMTPRQVA
RPECW  69     RELKNELGAGIATIT
RPECTN 28     RKLAQKLGVEQPTLY
Z1BPC2 27     RKVADALGINESQIS
JU0083 207    KDVADMMGISQSYIS
S11945 29     AEIAQRLGFRSPNAA
RPECEG 5      KDIAIEAGVSLATVS
RGECH  254    QELADRYGVSAERVR
RPECCT 13     KDVALKAKVSTATVS
RPECG  194    RLQGYYDALAESGIA  faux
RGECF  198    GDIGNYLGLTVETIS
QRECC  171    QEIGQIVGCSRETVG
RGECA  198    ASVAQHVCLSPSRLS
RGKBCP 187    RAKAPFIALNMAAIP  faux
RPECDO 25     KDAAALLGVSEMTIR
RCBPL  17     TKTAKDLGVYQSAIN
RGBP22 14     RAVAKALGISDAAVS
-
```

Figure 6.37: Résultat deuxième série — motif HTH (LeCombi, trente séquences, quorum de 60%). Un 'faux' à côté de l'occurrence d'un modèle indique que celle-ci est mal placée sur la séquence correspondante.

```
JEBY1  EKEEVAKKCGITPLQVRVwfinkrmrsk------------------
DNECFS NQTRAALMMGINRGTLRKklkkygmn--------------------
B26499 NQIRAADLLGLNRNTLRKkirdldiqvyrsgg--------------
BVECDA TQRSLAKALKISHVSVSQwergdseptgknlfalskvlqcsptwil
A32837 RVSITAENLGLTQPAVSNalkrlrtslqdplfvrthqgmeptpyaa
C29010 TVSRLALDAGVSVHIVRDyllrgllrpvacttggyglfddaalqrl
A25944 SQKETGDILGISQMHVSRlqrkavkklrealiedpsmelm------
A25399 RRIEIAHALCLTERQIKIwfqnrrmkwkkenktkgepgsggegdei
QCBP2L GTEKTAEAVGVDKSQISRwkrdwipkfsmllavlewgvvdddmarl
C55249 TQREIAKELGISRSYVSRiekralmkmfhefyraekekrkkakgk-
RPEC   TLYDVAEYAGVSYQTVSRvvnqashvsaktrekveaamaelnyipn
RPECDU TIKDVAKRANVSTTTVSHvinktrfvaeetrnavwaaikelhysps
JWE    PRQKVAIIYDVGVSTLYKrfpagdk---------------------
RGKBAP VQAKAARLLGMTPRQVAYriqimditmprl----------------
RPECW  SQRELKNELGAGIATITRgsnslkaapvelrqwleevllksd----
RPECTN TTRKLAQKLGVEQPTLYWhvknkralldalaiemldrhhthfcple
Z1BPC2 GQRKVADALGINESQISRwkgdfipkmgmllavlewgvedeelael
JU0083 TQKDVADMMGISQSYISRlekriikrlrkefnkmv-----------
S11945 TRAEIAQRLGFRSPNAAEehlkalarkgvieivsgasrgirllqee
RPECEG TLKDIAIEAGVSLATVSRvlnddptlnvkeetkhrileiaekleyk
RGECH  TLQELADRYGVSAERVRQleknamkklraaiea-------------
RPECCT TMKDVALKAKVSTATVSRalmnpdkvsqatrnrvekaarevgylpq
RPECG  TIKDVARLAGVSVATVSRvinnspkaseasrlavhsameslsyhpn
RGECF  TRGDIGNYLGLTVETISRllgrfqksgmlavkgkyitienndalaq
QRECC  TRQEIGQIVGCSRETVGRilkmledqnlisahgktivvygtr----
RGECA  DIASVAQHVCLSPSRLSHlfrqqlgisvlswredqrisqaklllst
RGKBCP HKQEAARLLGWGRNTLTRklkelgme--------------------
RPECDO HLKDAAALLGVSEMTIRRdlnnhsapvvllggyivleprsashyll
RCBPL  GQTKTAKDLGVYQSAINKaihagrkifltinadgsvyaeevkpfps
RGBP22 TQRAVAKALGISDAAVSQwkevipekdayrleivtagalkyqenay
```

Figure 6.38: Blocs trouvés par la méthode d'échantillonnage de Gibbs sur les trente séquences de notre jeu d'essai.

nombre de positions. Un tel regroupement est réalisé par Neuwald [Neuwald and Green, 1994] pendant la recherche des motifs lexicaux. Il est intéressant de noter que Neuwald n'utilise dans son illustration qu'un sous-ensemble de 15 parmi les 30 séquences du jeu de Lawrence. Il justifie cette réduction par le désir d'éliminer de l'ensemble de Lawrence les séquences considérées comme étant encore trop proches.

Le résultat produit par l'échantillonnage de Gibbs est légèrement meilleur que celui fourni par CoSamp puisque dans ce cas le motif est correctement placé sur toutes les séquences.

Avec CLUSTAL W par contre, ce que nous obtenons est cette fois complètement faux dans la mesure où l'alignement global réalisé par l'algorithme ne réussit à mettre le site correctement en correspondance que sur 6 parmi les 30 séquences. Sur ces six séquences-là, le motif est situé sur la partie C terminale des protéines ce qui explique probablement pourquoi CLUSTAL les aligne correctement. Cet exemple illustre donc un point faible de l'approche défendue par Posfai, qui consiste à établir d'abord un alignement global afin de rechercher les blocs conservés localement. Un tel alignement permet la localisation plus précise des parties conservées d'un ensemble de protéines dans la mesure où celles-ci sont relativement fixées les unes par rapport aux autres et uniformément distribuées sur les séquences. Cette approche n'est donc viable que si les séquences sont globalement similaires. Observons que cela n'est pas le cas ici, et n'est pas le cas non plus de la partie N terminale des P450 qui est probablement spécifique à chaque sous-famille et dans laquelle CLUSTAL ne parvient pas à identifier quoi que ce soit.

6.3.6 Nitrogénases et Hydrogénases

6.3.6.1 Introduction

Notre dernière illustration sera plus succinte. Elle est également plus fragile du point de vue biologique car nous ne pouvons valider ce que nous observons avec nos algorithmes avec aucun résultat antérieur. Nous ne sommes pas capables non plus, faute d'avoir des connaissances plus approfondies des protéines impliquées, de développer notre analyse. Cette illustration nous a semblé néanmoins importante parce qu'elle soulève le problème des trous.

En effet, un des partis le plus fréquemment pris est celui de considèrer qu'un bloc conservé associé à un signal ne peut présenter de trou. Or le résultat présenté ici concerne justement un motif indiqué dans PROSITE [Bairoch, 1992] comme étant spécifique d'un des sites actifs d'une famille de protéines, que LePoivre repère dans une famille différente de la première avec une délétion lorsque les modèles communs à ces deux familles sont recherchés. Plus que l'importance de ce motif en lui-même, nous voulons donc nous servir de cette illustration pour mettre en avant la nécessité de permettre la présence de trous dans la recherche de motifs. La raison pour laquelle les trous ne sont généralement pas autorisés dans les blocs est essentiellement technique : autoriser des trous augmente le temps d'exécution des algorithmes. Avec les algorithmes que nous avons développés ici, ce temps demeure cependant raisonnable dans la plupart des cas, ainsi que nous l'avons montré dans les sections de performance du chapitre 5.

Les deux ensembles de protéines dont il va être question dans cette dernière illustration font partie des protéines de la famille des nitrogénases et des hydrogénases.

Les nitrogénases sont des protéines enzymatiques responsables entre autres de la fixation de l'azote. Elles catalysent les réactions de réductions :

$$N_2 + 6H^+ \longrightarrow 2NH_3$$

ou :

$$N_2 + 8H^+ \longrightarrow 2NH_3 + H_2.$$

Cette seconde formule permet de voir que la réduction s'accompagne d'une libération d'hydrogène qui peut être couplée à une oxydation par les hydrogénases. Les nitrogénases sont des protéines dites 'fer-soufre', c'est-à-dire qu'elles contiennent toutes différents assemblages (en anglais 'cluster') constitués d'atomes de fer et de soufre, nécessaires à leur fonction. De plus, elles ont besoin d'un autre ion métallique (qui vient se fixer sur l'assemblage fer-soufre). Il s'agit, soit du molybdène (sous-familles des nitrogénases de noms mnénomiques nifK et nifD), soit, plus rarement, du vanadium (sous-familles des nitrogénases de noms mnénomiques vnfD, vnfK et vnfG). Certaines nitrogénases paraissent cependant ne fixer que le fer; elles ont pour noms anfD, anfK et anfG. Dans le cas des hydrogénases, c'est le nickel qui est nécessaire à la réaction d'oxydation. Il semble enfin que dans les deux cas, d'autres protéines interviennent dans les réactions de réduction et d'oxydation dont le rôle serait d'apporter ces métaux aux nitrogénases et aux hydrogénases [Fay, 1992] [Howard and Rees, 1994] [Robson and Postgate, 1980]. Ce sont ces protéines qui vont nous concerner ici.

Les premières, celles qui fournissent le molybdène aux nitrogénases, leur sont évolutivement liées. Elles ont pour nom mnénomique nifE et nifN. Les protéines de noms hypA, hypB, hypC, hypD, hypE et hypF, apparemment responsables quant à elles de l'apport du nickel aux hydrogénases, sont également impliquées dans la régulation de leur expression. Les deux sous-familles de ces protéines qui vont particulièrement nous intéresser sont les nifE et les hypD. Elles forment un jeu de 5 + 6 séquences dont les noms et numéros d'accès dans SWISS PROT [Bairoch and Boeckmann, 1991] sont respectivement : NIFE_AZOVI, NIFE_BRAJA, NIFE_CLOPA, NIFE_KLEPN, NIFE_RHOCA, HYPD_ALCEU, HYPD_AZOVI, HYPD_BRAJA, HYPD_ECOLI, HYPD_RHILV et HYPD_RHOCA.

6.3.6.2 Résultats

Les deux familles (nifE et hypD) ont été initialement regroupées un peu par hasard. Nous travaillions alors sur les nifE seules, et cherchions des occurrences 'avec trou' d'un des motifs fonctionnels (telles que celles observées sur les hypD). Dans les résultats présentés ici, nous avions recherché les modèles les plus longs présents dans au moins 9 des 11 séquences du jeu d'essai (fixant ainsi le quorum à 80%). Enfin, nous avions utilisé LePoivre avec la couverture suivante :

$$C = \left\{ \begin{array}{ll} \{I,L,M,V\} & (hydrophobe), \\ \{A,G\} & (très\ petit), \\ \{A,S,T\} & (petit), \\ \{C\} & (cystéine), \\ \{F,Y,W\} & (aromatique), \\ \{K,H,R\} & (basique), \\ \{E,D\} & (acide), \\ \{N,Q\} & (glutamine\ et\ glutamate), \\ \{G,P\} & (apparaît\ dans\ les\ structures\ contraintes) \end{array} \right\}.$$

Nous avions alors obtenu 11 modèles de longueur 9 représentant en fait le même ensemble d'occurrences. Cet unique bloc est donné dans la figure 6.39 avec les noms des séquences où il est présent ainsi que les positions. Observons que ce bloc est absent des séquences HYPD_ECOLI et HYPD_RHOCA. Il est important de noter que ce modèle est inclus dans un des deux motifs indiqués dans PROSITE comme étant associés à un site actif de la famille des nitrogénases (le motif de PROSITE est le suivant : [LIVMFYH][LIVMFST]H[AG][AGSP][LIVMQA][AG]C). La question que nous nous sommes alors posée était : est-ce que cette association est fortuite

HYPD_ALCEU	57	RMIHGP-GC
HYPD_AZOVI	57	RLIHGP-GC
HYPD_BRAJA	24	RMIHGP-GC
HYPD_RHILV	57	RMIHGP-GC
NIFE_AZOVI	54	HIVHGPIAC
NIFE_BRAJA	62	HLVHGPIAC
NIFE_CLOPA	71	HLVHGPIGC
NIFE_KLEPN	54	HLVHGPIGC
NIFE_RHOCA	63	HLVHAPAAC
		* * *

Figure 6.39: Motif trouvé avec LePoivre dans le jeu d'essai de 11 séquences, avec 1 erreur au plus autorisée et q égal à 80%.

ou constitue-t-elle une information concernant les protéines de la sous-famille des nifE et des hypD, et éventuellement, à travers eux, celle des nitrogénases et des hydrogénases?

Pour essayer de le déterminer, nous avons commencé par résumer ces modèles sous la forme d'un unique motif $[KHR][ILMV][ILMV]H[AG]P[ILMV][AG]C$ que nous avons ensuite recherché dans toute la banque SWISS PROT (version 31) en autorisant une erreur (substitution ou trou) avec une version simplifiée d'Agrep [Wu and Manber, 1992a] [Wu and Manber, 1992b] que nous avions implémentée. Nous avons alors découvert que celui-ci est présent également dans 12 nifD, 1 nitrogénase fixant le fer (anfD), 2 fixant le vanadium (vnfD) d'espèces différentes, ainsi que dans une protéine hypothétique de *coli* complètement différente (YJGB_ECOLI). La liste de ces séquences est donnée dans la figure 6.40.

Notons que si nous rajoutons la contrainte que les 3 positions strictement conservées dans le bloc initial (celles marquées d'une étoile dans la figure 6.39) ne doivent pas tolérer d'erreur, alors la recherche dans la banque fournit le même résultat à l'exception de la protéine hypothétique de *coli*, dont nous pouvons supposer qu'elle constituait un faux-positif.

Nous avons ensuite poursuivi notre étude de ce même jeu de six séquences initial en recherchant la présence éventuelle de mots communs avec LeCombi. Nous avons donc cette fois permis des jokers en plus d'erreurs (une substitution au plus dans ce cas). La couverture pondérée restreinte avec laquelle nous avons travaillé est :

$$CCPR = \left\{ \begin{array}{l} (\{X\}, \infty)\ \forall X \in C, \\ (\Sigma, 5), \\ (S, 0)\ pour\ tous\ les\ autres\ ensembles \end{array} \right\}$$

avec :

$$C = \left\{ \begin{array}{ll} \{I, L, M, V\} & (hydrophobe), \\ \{A, G\} & (très\ petit), \\ \{A, S, T\} & (petit), \\ \{C\} & (cystéine), \\ \{F, Y, W\} & (aromatique), \\ \{K, H, R\} & (basique), \\ \{E, D\} & (acide), \\ \{N, Q\} & (glutamine\ et\ glutamate), \\ \{G, P\} & (apparaît\ dans\ les\ structures\ contraintes) \end{array} \right\}.$$

Nous avons, par ailleurs, fixé le quorum à 100%. Huit modèles de longueur maximale (13) ont alors été identifiés, représentant 4 ensembles d'occurrences distincts, dont 2 sont en fait

```
+ sp|P16266|ANFD_AZOVI NITROGENASE IRON-IRON PROTEIN ALPHA CHAIN
+ sp|P31903|HYPD_ALCEU HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P31882|HYPD_AZOVI HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P31904|HYPD_BRAJA HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P40598|HYPD_RHILV HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P00464|NIFD_ANASP NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P25313|NIFD_AZOBR NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P07328|NIFD_AZOVI NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P06121|NIFD_BRAJA NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P06120|NIFD_BRASP NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P00467|NIFD_CLOPA NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P20620|NIFD_METTL NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|Q00239|NIFD_PLEBO NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P06769|NIFD_RHICP NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P00465|NIFD_RHILT NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P08717|NIFD_RHOCA NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P06662|NIFD_THIFE NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P08293|NIFE_AZOVI NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P26506|NIFE_BRAJA NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P10996|NIFE_CLOPA NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P08737|NIFE_KLEPN NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P19055|NIFE_RHOCA NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P15332|VNFD_AZOCH NITROGENASE VANADIUM-IRON PROTEIN ALPHA CH
+ sp|P16855|VNFD_AZOVI NITROGENASE VANADIUM-IRON PROTEIN ALPHA CH
- sp|P27250|YJGB_ECOLI HYPOTHETICAL 36.5 KD PROTEIN IN GNTV-LEUX
```

Figure 6.40: Séquences de la banque où le motif de la figure 6.39 est trouvé en permettant une erreur. Ce motif n'est plus retrouvé sur la séquence marquée d'un '-' lorsque l'on ne tolère pas d'erreur à certaines positions du motif (celles indiquées par une étoile dans la figure 6.39).

```
HYPD_ALCEU   220   YGKPVVIAGFEPL
HYPD_AZOVI   219   YGTPLVIAGFEPL
HYPD_BRAJA   187   FGKPVVIAGFEPL
HYPD_ECOLI   218   FHRPLVVAGFEPL
HYPD_RHILV   220   FRKPVVVAGFEPL
HYPD_RHOCA   218   YGKPLVVAGFEPT
NIFE_AZOVI   265   YGTPWFEGSFYGI
NIFE_BRAJA   263   YGIPFFEGSFYGI
NIFE_CLOPA   270   FGIPYIKVSFYGV
NIFE_KLEPN   265   YGTPWFEGSFYGI
NIFE_RHOCA   265   YGIPYFEGSFYGI
                      *     *
```

Figure 6.41: Motif trouvé avec LeCombi dans le jeu d'essai de 11 séquences, avec 5 jokers et 1 substitution au plus autorisée et q égal à 100%.

identiques à un décalage de positions près. Parmi les 3 ensembles restants, 2 mettent en correspondance des fragments des deux familles de protéines situés à des extrémités opposées des macromolécules et nous avons donc décidé de les ignorer pour l'instant. Par contre, le troisième ensemble nous a paru intéressant (il est donné dans la figure 6.41). On peut le résumer sous la forme du motif $[FY]GXPXXX[AG]XFX[GP][ILVM]$ où X dénote n'importe quel acide aminé. Observons que le G en troisième position représente l'intersection des deux ensembles de la couverture, $\{A, G\}$ et $\{G, P\}$, qui sont rencontrés dans les modèles à cet endroit-là). Nous avons recherché ce motif dans la banque avec une substitution au plus (pas de trous cette fois puisque ceux-ci n'étaient pas permis dans LeCombi). Le résultat de cette recherche est donné dans la figure 6.42.

En fixant comme précédemment les positions où un symbole unique apparaît dans les occurrences de l'ensemble de départ (celles marqués d'une étoile dans la figure 6.41), c'est-à-dire, en interdisant des substitutions en ces positions, le résultat se réduit alors aux séquences indiquées par un '+' dans la figure 6.42.

Nous pouvons observer que ce second motif semble être moins sélectif des deux familles de protéines. Par contre, il est présent dans toutes les séquences de notre jeu initial, y compris HYPD_ECOLI et HYPD_RHOCA d'où le premier motif était, lui, absent.

Nous avons alors voulu déterminer dans quelle mesure ces deux motifs considérés ensemble permettent de sélectionner les protéines qui nous intéressent (les nifE, nifN et hypD ainsi que les nitrogénases liées évolutivement aux nifE et nifN) et celles-là seulement.

Pour cela, il nous fallait un algorithme qui nous permette de rechercher ces motifs de manière plus souple que ne le fait Agrep, c'est-à-dire, un algorithme qui nous fournisse un score et non plus seulement une indication de la présence ou non de ces motifs parmi les séquences de la banque. Nous avons alors commencé par établir, pour chacun des motifs, un profil comme celui que nous utilisons dans CoSamp d'après la technique de Waterman [Goldstein and Waterman, 1994]. Puis nous avons calculé le score maximal de ces profils sur chacune des séquences. Deux différences fondamentales ont cependant été introduites à ce stade. La première est que les deux profils correspondant aux deux motifs sont recherchés simultanément et dans l'ordre : le premier motif doit se localiser avant le second puisque c'est ainsi qu'ils apparaissent dans les séquences du jeu d'essai. D'autre part, comme nous voulons cette fois permettre des trous (à cause du premier motif), nous utilisons un algorithme de programmation

```
+ sp|Q00831|AMPD_CITFR AMPD PROTEIN.
+ sp|P13016|AMPD_ECOLI AMPD PROTEIN.
+ sp|P30013|AMPD_SALTY AMPD PROTEIN (FRAGMENT).
+ sp|P20073|ANX7_HUMAN ANNEXIN VII (SYNEXIN).
+ sp|Q07076|ANX7_MOUSE ANNEXIN VII (SYNEXIN).
- sp|Q03025|APRE_PSEAE ALKALINE PROTEASE SECRETION PROTEIN APRE.
+ sp|Q04656|AT7A_HUMAN COPPER-TRANSPORTING ATPASE 1 (EC 3.6.1.36)
+ sp|P35670|AT7B_HUMAN COPPER-TRANSPORTING ATPASE 2 (EC 3.6.1.36)
- sp|P29705|ATPZ_PROMO ATP SYNTHASE PROTEIN I, SODIUM ION SPECIFI
- sp|P08125|CA1A_CHICK COLLAGEN 1(X) CHAIN PRECURSOR.
- sp|Q04539|CXEP_ALCEU RIBULOSE-PHOSPHATE 3-EPIMERASE, PLASMID (E
+ sp|P28639|DHN1_PEA DEHYDRIN DHN1.
+ sp|P07918|DPOL_HSV21 DNA POLYMERASE (EC 2.7.7.7).
+ sp|Q03763|DSG1_BOVIN DESMOGLEIN 1 PRECURSOR (DESMOSOMAL GLYCOPR
+ sp|Q02413|DSG1_HUMAN DESMOGLEIN 1 PRECURSOR (DESMOSOMAL GLYCOPR
+ sp|P17486|EGL5_CAEEL HOMEOBOX PROTEIN EGL-5 (HOMEOBOX PROTEIN C
+ sp|P06131|FDHA_METFO FORMATE DEHYDROGENASE ALPHA CHAIN (EC 1.2.
+ sp|P14324|FPPS_HUMAN FARNESYL PYROPHOSPHATE SYNTHETASE (FPP SYN
+ sp|P40223|GCSR_MOUSE GRANULOCYTE COLONY STIMULATING FACTOR RECE
+ sp|P30870|GLNE_ECOLI GLUTAMATE-AMMONIA-LIGASE ADENYLYLTRANSFERA
- sp|P38631|GLS1_YEAST 1,3-BETA-GLUCAN SYNTHASE COMPONENT GLS1 (E
- sp|Q03460|GLSN_MEDSA GLUTAMATE SYNTHASE (NADH) PRECURSOR (EC 1.
- sp|P24944|GLTT_BACCA PROTON/SODIUM-GLUTAMATE SYMPORT PROTEIN (G
- sp|P05373|HEM2_YEAST DELTA-AMINOLEVULINIC ACID DEHYDRATASE (EC
+ sp|P31903|HYPD_ALCEU HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P31882|HYPD_AZOVI HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P31904|HYPD_BRAJA HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P24192|HYPD_ECOLI HYDROGENASE ISOENZYMES FORMATION PROTEIN H
+ sp|P40598|HYPD_RHILV HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P26411|HYPD_RHOCA HYDROGENASE EXPRESSION/FORMATION PROTEIN H
+ sp|P13129|LUCI_LUCCR LUCIFERIN 4-MONOOXYGENASE (EC 1.13.12.7) (
+ sp|Q01158|LUCI_LUCLA LUCIFERIN 4-MONOOXYGENASE (EC 1.13.12.7) (
+ sp|P08659|LUCI_PHOPY LUCIFERIN 4-MONOOXYGENASE (EC 1.13.12.7) (
+ sp|P00467|NIFD_CLOPA NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P20620|NIFD_METTL NITROGENASE MOLYBDENUM-IRON PROTEIN ALPHA
+ sp|P08293|NIFE_AZOVI NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P26506|NIFE_BRAJA NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P10996|NIFE_CLOPA NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P08737|NIFE_KLEPN NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
+ sp|P19055|NIFE_RHOCA NITROGENASE IRON-MOLYBDENUM COFACTOR BIOSY
```

Figure 6.42: Séquences de la banque où le motif de la figure 6.41 est trouvé en permettant une substitution. Ce motif n'est plus retrouvé sur les séquences marquées d'un '-' lorsque l'on ne tolère pas de substitution à certaines positions du motif (celles indiquées par une étoile dans la figure 6.41) — Continue sur la page suivante.

```
- sp|P34755|PHYB_ASPAW 3-PHYTASE B PRECURSOR (EC 3.1.3.8) (MYO-IN
- sp|P34754|PHYB_ASPNG 3-PHYTASE B PRECURSOR (EC 3.1.3.8) (MYO-IN
+ sp|P22221|PODK_FLATR PYRUVATE,ORTHOPHOSPHATE DIKINASE PRECURSOR
- sp|Q05057|POLG_PYFV1 GENOME POLYPROTEIN (CONTAINS: 22.5 KD PROT
+ sp|P06186|PRGR_RABIT PROGESTERONE RECEPTOR (PR).
+ sp|P05168|PSBE_TOBAC CYTOCHROME B559 ALPHA CHAIN.
- sp|P37189|PTKC_ECOLI PTS SYSTEM, GALACTICOL-SPECIFIC IIC COMPON
+ sp|P38972|PUR4_YEAST PHOSPHORIBOSYLFORMYLGLYCINAMIDINE SYNTHASE
+ sp|P21750|SALA_DROME PROTEIN SPALT-ACCESSORY.
+ sp|P21748|SALA_DROOR PROTEIN SPALT-ACCESSORY.
+ sp|P21749|SALA_DROSI PROTEIN SPALT-ACCESSORY.
+ sp|P07819|SCRB_BACSU SUCROSE-6-PHOSPHATE HYDROLASE (EC 3.2.1.26
+ sp|P33388|SEFC_SALEN SEFC PROTEIN PRECURSOR.
- sp|P04802|SYDC_YEAST ASPARTYL-TRNA SYNTHETASE, CYTOPLASMIC (EC
- sp|Q03577|SYD_CAEEL PROBABLE ASPARTYL-TRNA SYNTHETASE (EC 6.1.1
- sp|P21889|SYD_ECOLI ASPARTYL-TRNA SYNTHETASE (EC 6.1.1.12) (ASP
- sp|P14868|SYD_HUMAN ASPARTYL-TRNA SYNTHETASE (EC 6.1.1.12) (ASP
- sp|P15178|SYD_RAT ASPARTYL-TRNA SYNTHETASE (EC 6.1.1.12) (ASPAR
- sp|P36419|SYD_THETH ASPARTYL-TRNA SYNTHETASE (EC 6.1.1.12) (ASP
- sp|P30053|SYH_STREQ HISTIDYL-TRNA SYNTHETASE (EC 6.1.1.21) (HIS
- sp|P14825|SYK2_ECOLI LYSYL-TRNA SYNTHETASE, HEAT INDUCIBLE (EC
- sp|P15180|SYKC_YEAST LYSYL-TRNA SYNTHETASE, CYTOPLASMIC (EC 6.1
- sp|P32048|SYKM_YEAST LYSYL-TRNA SYNTHETASE, MITOCHONDRIAL PRECU
- sp|P37477|SYK_BACSU LYSYL-TRNA SYNTHETASE (EC 6.1.1.6) (LYSINE-
- sp|P41258|SYK_CAMJE LYSYL-TRNA SYNTHETASE (EC 6.1.1.6) (LYSINE-
- sp|P37879|SYK_CRILO LYSYL-TRNA SYNTHETASE (EC 6.1.1.6) (LYSINE-
- sp|P41255|SYK_THETH LYSYL-TRNA SYNTHETASE (EC 6.1.1.6) (LYSINE-
+ sp|P28176|TYSY_MYCTU PUTATIVE THYMIDYLATE SYNTHASE (EC 2.1.1.45
- sp|P09842|UGST_HORVU GLYCOGEN (STARCH) SYNTHASE PRECURSOR (EC 2
- sp|P04713|UGST_MAIZE GLYCOGEN (STARCH) SYNTHASE PRECURSOR (EC 2
- sp|P19395|UGST_ORYSA GLYCOGEN (STARCH) SYNTHASE PRECURSOR (EC 2
- sp|P27736|UGST_WHEAT GLYCOGEN (STARCH) SYNTHASE PRECURSOR (EC 2
+ sp|P27870|VAV_MOUSE VAV PROTO-ONCOGENE.
+ sp|P03224|VGP8_EBV PROBABLE MEMBRANE ANTIGEN GP85.
- sp|P03234|VP40_EBV CAPSID PROTEIN P40 (VIRION STRUCTURAL PROTEI
- sp|P37340|YDHE_ECOLI HYPOTHETICAL PROTEIN IN RIBC 5'REGION (FRA
- sp|P40468|YIM9_YEAST HYPOTHETICAL 269.9 KD PROTEIN IN RPL13-STH
- sp|P40893|YJ67_YEAST HYPOTHETICAL 22.0 KD PROTEIN IN CRT1 5'REG
- sp|P34428|YL37_CAEEL HYPOTHETICAL 45.5 KD PROTEIN F44B9.7 IN CH
```

Figure 6.42: Séquences de la banque où le motif de la figure 6.41 est trouvé en permettant une substitution. Ce motif n'est plus retrouvé sur les séquences marquées d'un '-' lorsque l'on ne tolère pas de substitution à certaines positions du motif (celles indiquées par une étoile dans la figure 6.41).

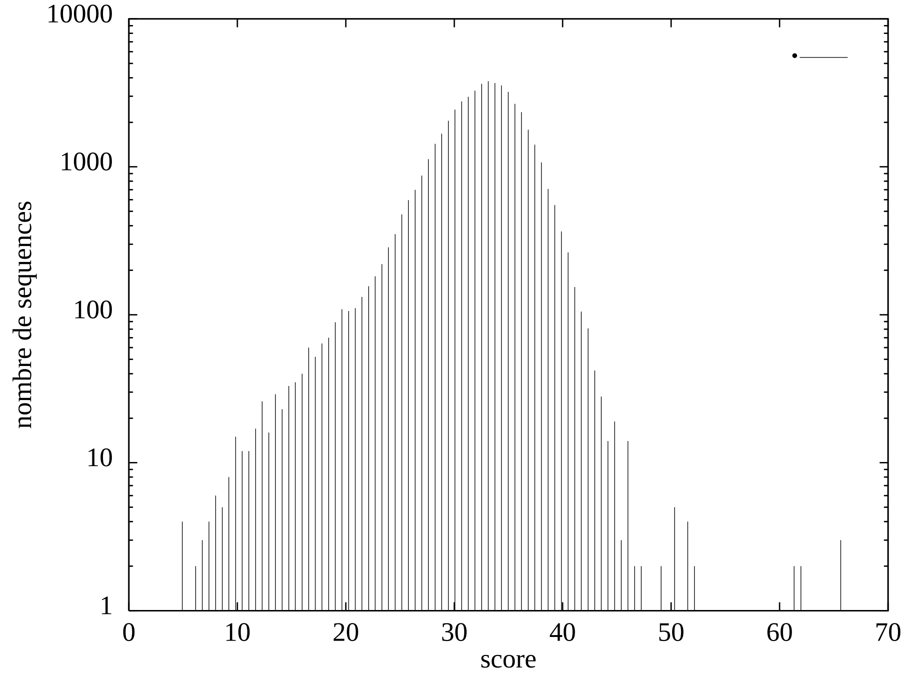

Figure 6.43: Distribution des scores obtenus en recherchant dans la banque les deux motifs des figures 6.39 et 6.41 simultanément avec l'algorithme de programmation dynamique de Viari [Viari, 1995].

dynamique pour obtenir ce meilleur placement des deux motifs [Viari, 1995]. Observons que le poids attribué à un trou dans un motif est égal à la plus forte valeur observée dans le profil. La figure 6.43 donne la distribution des scores qui sont obtenus sur SWISS PROT (version 33).

La queue de cette distribution, c'est-à-dire, les séquences ayant un score supérieur à 47, sont indiquées dans la figure 6.44. La séquence de plus fort score est une protéine du même type que les hypD, mais présente dans un autre organisme (les hypD sont trouvés dans les bactéries comme *Escherichia coli* et les hupD dans des organismes de plantes). Parmi les séquences restantes, apparaissent toutes les hypD, toutes les nifE, 13 parmi les 17 nifD, 6 nitrogénases (1 nifK, 2 nifN, 2 vndF et 1 anfD) et une seule protéine hypothétique et de fonction inconnue.

6.3.6.3 Discussion

Une étude beaucoup plus approfondie serait nécessaire pour savoir si ces premiers résultats présentent un réel intérêt biologique. Ainsi que nous l'avons dit, la principale difficulté rencontrée dans cette analyse est que nous ignorons à près tout de ces protéines, qui semblent par ailleurs exercer une activité très complexe dont nous n'avons fait ici que mentionner l'aspect le plus connu. S'il s'avèrait que ce que nous observons est indicatif d'une vraie homologie de fonction entre les hypD et les nifE et nifD (ainsi que les nitrogénases en général), cela renforcerait bien sûr notre conviction de l'importance des trous. Il serait alors intéressant de voir de façon plus générale dans quelle mesure les trous ne sont pas beaucoup plus courants dans les blocs

```
sp|P42033|HUPD_AZOCH HYDROGENASE score : 66.25
[1]      57       70.11     RLIHGP-GC
[2]      219      62.38     YGTPLVIAGFEPL
sp|P31882|HYPD_AZOVI HYDROGENASE score :  66.25
[1]      57       70.11     RLIHGP-GC
[2]      219      62.38     YGTPLVIAGFEPL
sp|P08737|NIFE_KLEPN NITROGENASE score : 66.05
[1]      54       75.56     HLVHGPIGC
[2]      265      56.54     YGTPWFEGSFYGI
sp|P31903|HYPD_ALCEU HYDROGENASE score : 65.52
[1]      57       68.11     RMIHGP-GC
[2]      220      62.92     YGKPVVIAGFEPL
sp|P31904|HYPD_BRAJA HYDROGENASE score : 64.82
[1]      24       68.11     RMIHGP-GC
[2]      187      61.54     FGKPVVIAGFEPL
sp|P26506|NIFE_BRAJA NITROGENASE score : 64.19
[1]      62       72.00     HLVHGPIAC
[2]      263      56.38     YGIPFFEGSFYGI
sp|P40598|HYPD_RHILV HYDROGENASE score :  62.21
[1]      57       68.11     RMIHGP-GC
[2]      220      56.31     FRKPVVVAGFEPL
sp|P08293|NIFE_AZOVI NITROGENASE score : 62.16
[1]      54       67.78     HIVHGPIAC
[2]      265      56.54     YGTPWFEGSFYGI
sp|P10996|NIFE_CLOPA NITROGENASE score :  61.89
[1]      71       75.56     HLVHGPIGC
[2]      270      48.23     FGIPYIKVSFYGV
sp|P26411|HYPD_RHOCA HYDROGENASE score : 61.39
[1]      62       62.56     EFIHGP-GC
[2]      218      60.23     YGKPLVVAGFEPT
sp|P19055|NIFE_RHOCA NITROGENASE score : 59.25
[1]      63       61.89     HLVHAPAAC
[2]      265      56.62     YGIPYFEGSFYGI
sp|P24192|HYPD_ECOLI HYDROGENASE score : 58.59
[1]      62       62.56     EFIHGP-GC
[2]      218      54.62     FHRPLVVAGFEPL
sp|P20620|NIFD_METTL NITROGENASE score :  57.12
[1]      64       73.56     HMVHGPIGC
[2]      275      40.69     FDIPYTRATFFGL
sp|P00467|NIFD_CLOPA NITROGENASE score : 53.60
[1]      70       65.67     HITHGPIGC
[2]      276      41.54     YGIPWIKCNFIGV
sp|P06662|NIFD_THIFE NITROGENASE score : 52.38
[1]      89       62.00     HISHGPVGC
[2]      299      42.77     YGIPYIEFNFFGP
sp|P08738|NIFN_KLEPN NITROGENASE score:  52.24
[1]      36       56.78     PLVHGAQGC
[2]      391      47.69     FALPLVRAGF-PL
sp|Q00239|NIFD_PLEBO NITROGENASE score :  52.04
[1]      88       62.00     HISHGPVGC
[2]      303      42.08     YGIPWLEYNFFGP
sp|P07328|NIFD_AZOVI NITROGENASE score : 51.62
[1]      79       62.00     HISHGPVGC
[2]      289      41.23     YGIPWMEYNFFGP
sp|P15332|VNFD_AZOCH NITROGENASE score : 51.61
[1]      65       65.22     QMIHGPLGC
[2]      425      38.00     NG-PYM--GFEGF
sp|P16855|VNFD_AZOVI NITROGENASE score : 51.61
[1]      66       65.22     QMIHGPLGC
[2]      426      38.00     NG-PYM--GFEGF
sp|P25313|NIFD_AZOBR NITROGENASE score : 50.92
[1]      80       62.00     HISHGPVGC
[2]      291      39.85     FGIPWMEYNFFGP
sp|P06769|NIFD_RHICP NITROGENASE score : 50.77
[1]      85       62.00     HISHGPVGC
[2]      298      39.54     FGIPWCEYNFFGP
sp|P06120|NIFD_BRASP NITROGENASE score :  50.77
[1]      85       62.00     HISHGPVGC
[2]      298      39.54     FGIPWCEYNFFGP
sp|P06121|NIFD_BRAJA NITROGENASE score : 50.77
[1]      85       62.00     HISHGPVGC
[2]      298      39.54     FGIPWCEYNFFGP
sp|P00464|NIFD_ANASP NITROGENASE score : 50.69
[1]      82       62.00     HISHGPVGC
[2]      297      39.38     YGMPWMEFNFFGP
sp|P16266|ANFD_AZOVI NITROGENASE score : 50.04
[1]      67       62.00     HISHGPVGC
[2]      272      38.08     YGIPrLDIDGF-GF
sp|Q07933|ANFD_RHOCA NITROGENASE score : 49.65
[1]      67       61.22     HISHGPNGC
[2]      272      38.08     YGIPrLDIDGF-GF
sp|P00466|NIFD_KLEPN NITROGENASE score : 49.26
[1]      81       62.44     HISHGPAGC
[2]      300      36.08     FG-PTKIA--ESL
sp|Q02452|NIFD_FRAAL NITROGENASE scrore : 48.41
[1]      82       56.67     TISHGPVGC
[2]      289      40.15     FGTPWTEFNFFGP
sp|P06122|NIFK_BRASP NITROGENASE score : 47.48
[1]      85       52.11     PFVHGSQGC
[2]      452      42.85     TGTPLIRIGF-PI
sp|P30268|YKAB_BACFI HYPOTHETICAL score : 47.36
[1]      74       59.33     HLTHGPIGV
[2]      127      35.38     TKRPYVV-SFTGC
sp|P08717|NIFD_RHOCA NITROGENASE score : 47.23
[1]      95       62.00     HISHGPVGC
[2]      344      32.46     Y-RPLV-DGI--L
sp|P19077|NIFN_RHOCA NITROGENASE score : 47.10
[1]      36       50.89     PLFHGAQGC
[2]      392      43.31     LHLPLVRAGF-PI
```

Figure 6.44: Séquences de la banque où les deux motifs des figures 6.39 et 6.41 sont trouvés simultanément par programmation dynamique [Viari, 1995] avec un score supérieur ou égal à 47. Les séquences sont classées par ordre décroissant des scores.

de similarité que l'on peut associer à l'activité d'une même famille de protéines, ou de familles évolutivement distantes.