Chapitre 7

Critiques et perspectives

We play at Paste -
Till qualified, for Pearl -
Then, drop the Paste -
And deem ourself a fool -

The Shapes - though - were similar -
And our new Hands - Learned Gem-*Tactics -*
Practicing Sands -

Poème #320. Emily Dickinson.
The Complete Poems of Emily Dickinson.
Thomas H. Johnson (Ed.). Faber and Faber, 1975.

Ce travail a soulevé plus de problèmes qu'il n'a réussi à en résoudre. Certains ont été mentionnés auparavant, nous souhaitons dans ce dernier chapitre les reprendre de manière plus systématique. L'examen critique de notre approche, tant au niveau informatique qu'à celui de la biologie, nous conduira parfois à remettre en cause un certain nombre de concepts, comme celui d'une comparaison multiple. Nous proposerons également des développements possibles des définitions de ressemblance et des algorithmes introduits précédemment.

Ce chapitre est divisé en deux parties. La première considère les aspects plutôt algorithmiques et les suggestions qui y sont faites pour l'avenir sont donc le plus souvent des extensions immédiates de ce travail. La seconde partie traite les aspects plus étroitement liés à la biologie et les perspectives correspondantes exigent une réflexion plus approfondie de la notion de ressemblance. Bien sûr, la frontière entre ces deux parties peut être moins nette que nous ne l'affirmons ici.

7.1 Les critiques et perspectives portant sur l'aspect algorithmique

Les critiques présentées dans cette section font exclusivement référence à l'approche lexicale que nous avons adoptée. Notons toutefois que les problèmes rencontrés, et qui ont été principalement indiqués dans le chapitre consacré aux illustrations, s'appliquent pour la plupart à d'autres types d'approches également — par exemple, celles purement structurales. Les développements que ces critiques suggèrent sont le plus souvent de même nature.

Les difficultés algorithmiques que nous avons rencontrées sont essentiellement liées à la complexité (due en particulier à la méthode de construction par récurrence des modèles) et à la forme parfois insatisfaisante des résultats produits.

En ce qui concerne la complexité algorithmique, celle obtenue par nos algorithmes est intrinsèquement attachée au problème que nous essayons de résoudre. Cela est principalement le cas en ce qui concerne le nombre de modèles (et de leurs occurrences) que ces algorithmes doivent manipuler pour parvenir aux résultats demandés. Ainsi, les facteurs les plus importants dans ce nombre et, par conséquent, ceux ayant la plus forte influence sur le temps d'exécution, interviennent sous la forme d'un terme exponentiel. Il s'agit du nombre e d'erreurs autorisées (e.g. k^e où k est la longueur des modèles) et/ou de cette longueur k lorsque l'alphabet sur lequel les modèles sont définis est dégénéré (e.g. g^k où g mesure le degré de non-transitivité d'une couverture). Or le poids de ces facteurs ne peut être diminué lorsque nous recherchons toutes les solutions possibles du problème initial (c'est-à-dire, tous les modèles). Par contre, ces facteurs n'interviennent en pratique que lorsque la longueur des modèles est encore relativement petite (autour de 4-6). Ils cessent d'avoir une influence aussi importante dès que la longueur des modèles qui demeurent valides devient grande, car ceux-ci sont alors peu nombreux. Modifier le mode de construction des modèles, par exemple en engendrant directement tous ceux suffisamment longs, n'apporte pas une solution au problème. Il y a essentiellement deux raisons à cela. La première est qu'il est difficile de déterminer à partir de quelle longueur il n'y aura plus qu'un nombre réduit de modèles encore valides. La seconde est qu'en général cette longueur est trop grande pour qu'une génération automatique de tous les modèles possibles puisse se faire en un temps raisonnable. L'algorithme de Lawrence (section 4.2.2.2), qui produit les résultats les plus comparables aux notres, résoud le problème, c'est-à-dire, il évite d'avoir à passer par le 'pic' des petits modèles (il s'agit pour lui de matrices), essentiellement parce qu'il n'est pas exhaustif. Il n'explore ainsi qu'une partie de l'espace de toutes les solutions possibles compatibles avec la définition de ressemblance qu'il établi. Dans la plupart des cas, la partie explorée est la 'bonne', c'est-à-dire, la matrice obtenue à la fin est bien celle qui maximize la quantité observée, mais ceci n'est pas garanti absolument. Cependant nous avons vu dans la section 4.2.3.3 qu'il existe des techniques permettant de trouver tous les mots identiques communs à un ensemble de chaînes qui, par l'utilisation qu'ils font d'une structure de données spéciale (arbre des suffixes ou DAWG), sont plus efficaces que les notres en l'absence d'erreurs ou de dégénérescence de l'alphabet. La complexité obtenue est ainsi en $O(n.N)$ alors que la notre est en $O(n.N.\log k)$ ou en $O(n.N.k)$ où n est la longueur moyenne des chaînes et N leur nombre. Cela nous amène à poser un certain nombre de questions qui sont :

- est-il possible d'attribuer des noms aux mots répétés de longueur k ou k_{max} de N chaînes en $O(n.N)$ temps (ou en moins de $O(n.N.\log k)$ opérations) sans avoir besoin de construire une structure élaborée comme un arbre des suffixes (ni utiliser une table indexée ou de hachage), c'est-à-dire donc, en procédant 'à la volée'?

- pouvons-nous envisager la construction d'un arbre/automate des facteurs/suffixes qui tienne compte d'une relation de similarité non-transitive entre mots d'une ou de plusieurs chaînes? Est-ce que l'utilisation d'une telle structure nous permettrait d'éliminer le terme en k ou $\log k$ comme pour les relations d'identité? Est-ce que toute construction utile sera nécessairement trop coûteuse en mémoire?

- d'une manière plus simple, est-ce qu'un arbre/automate des facteurs/suffixes exacts pourrait être utilisé uniquement pour faciliter le parcours dans l'arbre des modèles, et non pour stocker ces derniers, cela quelle que soit la définition de ressemblance adoptée?

Ces questions demeurent ouvertes pour l'instant.

Il paraît assez clair que, lorsqu'une contrainte de quorum q strictement supérieure à 1 est appliquée, des économies plus grandes que celles que nous avons déjà réalisées devraient être possibles. Il suffit pour cela d'observer que lorsqu'un certain modèle m (un modèle comme un mot) ou M (un modèle comme un produit cartésien d'ensembles d'une couverture) ne vérifie plus cette contrainte, alors ce n'est pas seulement tous les modèles m'/M' ayant m/M comme préfixe dont nous pouvons affirmer qu'ils ne vérifient pas non plus cette contrainte, mais tous les modèles m'/M' ayant m/M comme sous-modèle! Or cette observation n'est jamais prise en considération dans aucun de nos algorithmes, simplement parce que l'arbre des modèles est parcouru en profondeur et qu'il n'est jamais effectivement construit. Le faire exigerait bien sûr plus de mémoire, la question est alors, y a-t-il un équilibre à trouver entre mémoire occupée et temps de calcul?

En ce qui concerne les résultats obtenus, les problèmes rencontrés ont été de natures diverses. Le premier est que ces résultats représentent souvent une quantité d'information encore assez considérable qu'il faut essayer de condenser d'une certaine façon. La 'condensation' des modèles peut se faire de deux manières principales :

- les modèles peuvent être regroupés en 'méta-modèles verticaux' : c'est ce que nous faisons déjà d'une première manière simple lorsque nous réunissons en un seul deux modèles ayant exactement les mêmes ensembles d'occurrences, ou éliminons les modèles strictement inclus dans un autre en termes de ces ensembles;
- les modèles peuvent être réunis en 'méta-modèles horizontaux', c'est-à-dire en trains de modèles dont les occurrences se succèdent dans le même ordre sur chacune des chaînes de l'ensemble que nous comparons.

Cette seconde manière impose l'établissement d'un ordre (partiel) de succession des modèles le long des chaînes. Elle est particulièrement intéressante lorsque ce qui est recherché représente une succession de répétitions contiguës directes ou miroir, un palindrome flou, ou encore un signal constitué par une suite de signaux séparés entre eux d'une certaine distance. Nous avons rencontré un exemple d'un tel signal avec les promoteurs de *subtilis*, il en existe de nombreux autres. Pour l'instant, nous n'avons pas traité cet aspect sauf dans des cas très particuliers (comme pour les sites -10 et -35 de ces promoteurs lorsque ceux-ci sont à une distance constante l'un de l'autre — voir la section 6.2.2). Il y a au moins trois façons différentes de le faire. La première s'applique aux répétitions et aux palindromes et consiste à travailler avec nos algorithmes tels quels (version 'normale' ou ADN-X — voir section 6.2.1.2) en y ajoutant une contrainte sur la position relative des occurrences dans la chaîne. Quant aux signaux, nous pouvons, soit les identifier séparément pour les mettre ensemble ensuite, soit essayer de les caractériser de manière simultanée (avec un seul modèle). Cette seconde façon est bien sûr la plus délicate. Une approche possible du problème consisterait à introduire des jokers comme nous l'avons fait, en autorisant cependant cette fois une mise en correspondance de suites de jokers de longueurs variables ('variable length don't care symbols') à l'intérieur d'un bloc de similarité. Enfin, une approche encore plus générale exigerait l'introduction de 'méta-définitions' de ressemblance : des définitions portant cette fois sur les objets des définitions que nous avons établies dans le cadre de ce travail, c'est-à-dire, sur les modèles eux-mêmes. Nous pourrions ainsi caractériser un complexe de signaux plutôt qu'un seul signal. Ces 'méta-définitions' fourniraient alors l'équivalent des descriptions utilisées par Myers dans [Mehldau and Myers, 1993] et principalement dans [Knight and Myers, 1995] et qu'il appelle des 'super-patterns'. Dans le cas de Myers, de telles descriptions sont connues à l'avance et il s'agit d'en rechercher les instances dans des séquences biologiques. Dans notre cas, il faudrait extraire les descriptions des séquences.

Obtenir des descriptions même dans les cas simples (i.e. lorsqu'il n'y a qu'un seul motif) n'est cependant pas une chose facile à faire. Ces descriptions sont pourtant importantes. En effet, la plupart du temps l'identification de signaux ou de motifs structuraux dans les macromolécules biologiques à l'aide de modèles ne représente qu'une première étape dans une étude complète de ces objets. Cette étape préliminaire doit ensuite être suivie par une utilisation des modèles trouvés afin de localiser des signaux ou motifs du même genre ailleurs. Nous avons indiqué dans le chapitre 3 (section 3.4.4) que la représentation des motifs lexicaux associés à des signaux est traditionnellement donnée, soit au moyen d'expressions régulières, soit en utilisant des matrices ou des profils. Nos modèles sont un mélange des deux, ils sont une matrice lorsqu'on ne considère que les ensembles d'occurrences, et sont des expressions régulières limitées lorsqu'on les considère comme des produits cartésiens des ensembles d'une couverture. Aucune des deux formes, matrice ou expression régulière limitée, n'est parfaite. Les matrices perdent en effet toute information concernant la dépendance entre monomères situés à des positions différentes du motif, tandis que les expressions régulières ne sont souvent pas suffisamment flexibles. En particulier, elles ne peuvent permettre de repérer que ce qui a été effectivement 'vu' dans les exemples ayant servi à sa construction. Une matrice ou un profil peuvent être ajustés (à travers des régularisateurs par exemple) pour tenir compte de ce qui n'est pas réellement observé mais que l'on suppose seulement être possible. Par ailleurs, les algorithmes, y compris les notres, biaisent en faveur des motifs les plus forts. Or nous avons vu avec les CRP (section 6.2.3) qu'un signal peut être 'délibérément' faible. Il est donc essentiel de pouvoir en tenir compte, ce que nos modèles ne sont pas capables de faire pour l'instant. Ils représentent le centre d'une boule et peuvent indiquer seulement (lorsqu'ils sont 'corrects') que les signaux recherchés doivent se trouver quelque part à l'intérieur de la boule. Dans la plupart des cas cependant, la boule est creuse : le signal ne peut être 'trop bon', c'est-à-dire, en termes pratiques, trop ressemblant au modèle.

Reste enfin le problème du tri des modèles obtenus, quels qu'ils soient, afin de séparer 'l'ivraie du bon grain'. La pertinence d'un modèle est bien sûr de nature purement biologique et, dans ce sens, nous ne pouvons rien faire au niveau des algorithmes. Cependant, il est possible d'en obtenir une première évaluation statistique. Nous en avons proposée une dans le chapitre précédent. Nous avons alors également expliqué notre sentiment concernant le moment où cette évaluation devait être réalisée, à savoir *a posteriori* et non en même temps que la recherche des modèles comme cela est le cas pour d'autres approches. Ce sentiment n'a pas changé, cependant nous pensons que l'algorithme que nous avons élaboré (CoSamp — section 6.3.2) n'est pas encore suffisamment mûr conceptuellement. Outre le fait que l'implémentation peut être rendue plus efficace, nous ne sommes pas complètement satisfaits de la stratégie adoptée. Cette insatisfaction est indépendante du problème du choix de certaines données ou paramètres (comme celui des ensembles exemples et tests), dans lesquels sans doute seule l'intuition biologique peut intervenir. Il serait intéressant également de voir dans quelle mesure l'étape d'évaluation ne pourrait être retardée. Notre intuition est qu'il est possible de pousser l'exploration combinatoire plus loin et d'obtenir des résultats plus fins avant d'avoir à les évaluer. Il y a ainsi des facteurs que nous n'essayons pas d'optimiser (par exemple, le quorum par rapport à la longueur des modèles) et qui nous permettraient peut-être d'éliminer des 'mauvais' modèles sur la base de critères non nécessairement statistiques. Observons en outre que la méthode telle qu'actuellement conçue ne permet en fait d'évaluer que des modèles très simples, en particulier sans trous, et qu'elle ne résoud pas le problème de la 'condensation' des modèles à quelque niveau que ce soit.

7.2 Les critiques et perspectives portant sur l'aspect biologique

Les critiques et perspectives de notre travail portant sur les questions de biologie se réfèrent aux définitions de ressemblance que nous avons établies et qui, malgré leur diversité, restent parfois insuffisantes. Nous allons en donner ici plusieurs exemples.

Le premier concerne la notion de comparaison multiple. Considérons le cas de la valeur à attribuer à un bloc $\mathcal{B}$. Supposons que nous décidons d'affecter à $\mathcal{B}$ un score SP, c'est-à-dire, sa 'valeur' est égale à la somme des scores SP de chacune des colonnes du bloc, où le score SP d'une colonne est lui-même la somme des scores de toutes les paires de monomères que la colonne comporte. Supposons également que nous disposons de deux blocs, en fait, de deux modèles M et M' ayant même scores SP mais tels que M est de longueur k et est présent dans $\frac{N}{2}$ chaînes distinctes tandis que M' est de longueur $2k$ et est présent dans $\frac{N}{4}$ chaînes distinctes. Lequel des deux est le plus pertinent? Une évaluation statistique, si tant est qu'elle soit possible, ne nous permettrait de répondre à cette question que dans le cadre d'un modèle dont la validité est toujours sujette à caution. Par ailleurs, si nous changeons la définition du score multiple, par exemple si nous adoptons des scores 'Étoile Centrale' (section 3.4.1.5) à la place des scores SP, il est possible que les deux blocs ne présentent plus le même score et la réponse à la question de la pertinence peut changer. Lorsque nous comparons un ensemble de chaînes relativement homogène dans lequel ce que nous recherchons doit se trouver présent partout ou presque, nous pouvons, soit considérer que le modèle le plus long est celui qui nous intéresse le plus (c'est que ce nous avons souvent fait dans les illustrations — par exemple pour les CRP et les IRE), soit envisager d'optimiser à la fois sur la longueur des modèles et sur le quorum ainsi que nous l'avons suggéré pour CoSamp dans la section précédente. Il existe toutefois au moins une situation pour laquelle il nous est pour l'instant algorithmiquement impossible de décider lequel des modèles M et M' donnés en exemple ci-dessus est le 'bon'. Cette situation est celle d'un criblage multiple de banque.

Rappelons en effet que le problème du criblage est à la fois plus simple et plus délicat à traiter que celui d'une analyse multiple portant sur un ensemble plus ou moins homogène de chaînes. Il est plus simple car il ne s'agit plus ici de localiser précisément les régions qui se trouvent impliquées dans une fonction biologique, encore moins d'essayer de comprendre de manière plus fine le mécanisme de reconnaissance qui permet à une molécule d'établir les contacts aux travers lesquels elle exerce son activité. Le but du criblage est 'uniquement' de savoir, de façon même grossière, si une molécule nouvellement séquencée ressemble au moins partiellement à quelque chose de déjà connu. Par contre, le problème qu'un tel criblage pose est plus compliqué pour au moins deux raisons. La première est de nature algorithmique. En effet, les algorithmes introduits dans ce travail ne supportent, à l'heure actuelle, qu'un nombre relativement restreint de chaînes ((jusqu'à 1000 environ). Ce nombre est très inférieur à ce que contient une banque complète (par exemple, 52 000 chaînes pour une banque de protéines). La seconde raison, probablement plus importante, est qu'une banque est un ensemble <u>très</u> hétérogène de chaînes. Par hétérogène, nous entendons ici que, d'une part, la banque contient de nombreuses chaînes qui n'ont rien à voir entre elles (et surtout n'ont rien à voir avec la chaîne requête) et, d'autre part, qu'elle est souvent très redondante, c'est-à-dire, qu'elle contient de multiples exemplaires ou variantes d'une même protéine. Enfin, les banques sont généralement biaisées, dans la mesure où les familles de protéines qui y sont stockées ne constituent pas nécessairement un échantillonnage représentatif de celles qui sont exprimées chez les différents organismes. Si la chaîne requête dont la fonction est inconnue est vraiment 'nouvelle', c'est-à-dire, si elle n'a pas de 'parent' très proche dans la banque, alors l'établissement de son identité, et donc de sa fonction, en présence d'autant de bruit est difficile. L'utilisation de comparaisons multiples, plutôt que deux à deux, permet d'amplifier les faibles ressemblances entre objets mais cet avantage s'effondre d'une certaine façon lorsque ces comparaisons portent sur un aussi grand nombre d'objets trop hétérogènes. En effet, les similarités qui dénotent de 'vraies' homologies se retrouvent alors très souvent noyées parmi les 'fausses', et il devient très difficile de les départager. Ce problème n'est bien sûr pas seulement lié au fait qu'il y a plus d'objets à comparer, ou à la nature de ces objets, mais est également associé au système de score choisi. En particulier, il ne se pose pas dans le cas des structures tridimensionnelles parce que l'extension d'une distance entre deux objets se fait de manière naturelle et intuitive au cas de plus de deux objets. Cela n'est plus vrai pour les séquences. Le prototype de cribleur multiple que nous avons développé (et qui constitue une extension naturelle des algorithmes introduits dans ce travail) 'marche' ainsi relativement bien en tant qu'algorithme, mais produit des résultats biologiques insatisfaisants car il n'est pas capable de distinguer, parmi tous les modèles qu'il trouve, lesquels apportent une information quant à l'activité de la chaîne requête, et lesquels ne sont dus qu'au hasard. Deux voies paraissent possibles pour essayer de contourner cet obstacle. La première consiste à mettre en place un système d'évaluation des modèles obtenus qui serait plus adapté au type de problème rencontré lors d'un criblage. La seconde exige une remise en cause des définitions de ressemblance pour le cas spécifique du criblage, en particulier la notion de distance ou de similarité entre plus de deux objets à la fois. Aucun des systèmes proposés à l'heure actuelle (scores SP, distance de Levenshtein ou autre) ne semble réellement adéquat.

L'inadéquation de certaines des mesures de ressemblance adoptées dans ce travail ne s'applique pas qu'aux cribleurs. Considérons par exemple la question des trous. Le problème n'est pas seulement de savoir quelle valeur numérique affecter à la présence d'un trou dans un alignement global ou à l'intérieur d'un bloc, et en particulier comment compter les trous qui se succèdent; le problème est plutôt de savoir comment 'interpréter' ces trous. Dans les études phylogénétiques, les trous sont en général interdits parce que, fondamentalement, on ne sait pas quelle signification accorder à de tels événements mutationels. Lorsque ceux-ci sont autorisés, ils sont traités de la même manière que les substitutions, et les algorithmes cherchent simplement à minimiser le nombre de ces événements. Il a été montré cependant que cette minimisation de la distance entre deux séquences en termes de mutations peut produire des alignements qui sont souvent inconsistants avec leur similarité structurale [Creighton, 1993], pour ne pas parler de leur histoire évolutive. Ce problème devient d'autant plus aigu que nous cherchons à aligner non plus deux, mais plus de deux séquences simultanément puisque dans ce cas nous ne disposons même plus d'une notion de distance aussi intuitivement claire que pour le deux à deux. Pourtant, il est important de pouvoir tenir compte des trous, y compris dans les études phylogénétiques. Nous ne pouvons cependant pas le faire de la façon habituelle. Ainsi, Barriel [Barriel, 1994] par exemple suggère de considérer de manière différente la présence d'une suite de plusieurs trous dans une des chaînes d'un ensemble que l'on aligne de manière multiple suivant que d'autres types de mutations ont été ou non observées aux positions placées en correspondance avec eux sur les autres chaînes de l'ensemble étudié. Son idée est que tout système de poids doit pouvoir tenir compte d'événements emboîtés, c'est-à-dire, de mutations subséquentes à l'insertion ou délétion de plusieurs nucléotides à un endroit donné sur une des séquences.

Ce traitement différent à accorder à un trou selon que celui-ci est isolé, fait partie d'une suite contiguë d'autres trous ou se trouve dans des zones très 'mutables' de la chaîne, nous rappelle également que nous ne travaillons en fait qu'avec une partie des pièces du puzzle biologique. En ne considérant que les informations 'élémentaires' (les mutations ponctuelles), nous négligeons le fait que les objets biologiques se transforment avec le temps par des opérations portant sur des 'morceaux' plus grands de ces objets : duplication, inversion ou translocation d'un segment entier d'ADN etc. La difficulté dans ce cas n'est pas seulement d'élaborer des algorithmes qui puissent traiter ces transformations, mais est surtout de formaliser leur nature exacte, les règles auxquelles elles obéissent et l'importance à accorder à chacune.

Les liens évolutifs entre macromolécules en général sont un autre aspect qui n'a pu être exploité jusqu'à présent que de manière heuristique par les programmes d'alignement ou de comparaison de chaînes, cela même lorsque nous pouvons faire l'hypothèse que l'arbre phylogénétique est connu *a priori*. Lorsque l'arbre est inconnu et est construit avant l'alignement par des méthodes de regroupement ('clustering'), c'est à la fois cette première étape et la seconde, réalisant l'alignement ou la comparaison à partir de l'arbre, qui sont alors effectuées de manière heuristique. Par ailleurs, dans les deux cas, un tel arbre phylogénétique n'a été utilisé pour l'instant que dans le cadre d'un alignement global, jamais dans celui de la recherche de blocs de similarité. Or certaines des illustrations présentées dans le chapitre précédent (par exemple celle concernant les motifs 'Helix Turn Helix' de la section 6.3.5) semblent indiquer que la connaissance d'un tel arbre pourrait apporter une aide considérable à la recherche de blocs. Ces illustrations montrent également que, lorsque l'arbre phylogénétique n'est pas connu à l'avance (ce qui est le plus souvent le cas), il serait intéressant de pouvoir le construire <u>en même temps</u> que l'alignement et non préalablement à celui-ci. La raison fondamentale est que nous pouvons alors bénéficier des informations plus riches qu'apporte une comparaison multiple, les constructions d'arbre effectuées préalablement à la recherche de similarités ne pouvant en effet se faire que sur la base de comparaisons deux à deux.

Un autre problème important soulevé au cours de ce travail est, à nouveau, lié à la question de l'évaluation des modèles mais est considéré, cette fois, sous un autre angle. L'évaluation dont nous avons parlé dans la section précédente est de nature statistique. Dans le cas de CoSamp, une telle mesure a été établie directement à partir des objets obtenus (les modèles et leurs occurrences) par rapport à ce qui serait observé sur des objets aléatoires (elle aurait pu également s'appuyer sur une autre hypothèse nulle). À aucun moment cependant, nous ne profitons du fait qu'un objet peut être 'intéressant' non à cause de ce qu'il est en termes absolus, mais parce qu'il possède des caractéristiques qui le distinguent des autres objets. Watanabe [Watanabe, 1985] propose ainsi que la reconnaissance des frontières ('boundary recognition') séparant un objet de son entourage est un facteur important dans l'identification de celui-ci : *"Texture may be determined only by a statistical relation between neighboring points"*. Watanabe considère des critères statistiques, ou probabilistes, mais il est sans doute possible d'imaginer des critères plus 'déterministes'. Ceux-ci indiqueraient ce que les modèles qui nous intéressent ne peuvent <u>pas</u> être et établiraient ainsi les limites de l'espace où ces modèles peuvent se situer. Ces critères pourraient alors guider la recherche dès le départ <u>sans courir le risque de perdre de l'information</u>. Un exemple d'un tel critère (mais que nous ne savons 'quantifier' pour l'instant) pourrait être celui d'un taux de mutation (ou d'un type de conservation) différent dans les régions impliquées dans une activité biologique [Johnson and Overington, 1993]. Un tel phénomène est sans doute ce qui est observé dans les sites de fixation de la CRP (section 6.2.3) dont nous avons vu que la relative conservation possède un rôle essentiel dans le mécanisme de régulation de la transcription. Si nous sommes capables de limiter d'une certaine façon l'espace à explorer autrement que par la contrainte de quorum (voir l'algorithme Combi — section 5.4.2.2), nous ne savons pas vraiment comment reconnaître un objet par ce qui l'oppose au reste des objets de l'univers ou l'en différencie. Nos modèles sont ainsi toujours construits de manière 'positive' : leurs occurrences sont tous les mots qui vérifient une certaine propriété par rapport aux modèles et par rapport à eux seulement.

L'idée développée dans le paragraphe précédent est bien sûr liée à celle du contexte dans lequel se situe un objet. Il existe une autre notion de contexte très différente qui concerne cette fois la manière avec laquelle l'environnement peut modifier l'identité même des objets ou leurs caractéristiques essentielles. Il a été montré, par exemple, que la propension que possède un acide aminé à être substitué par un autre acide aminé quelconque varie selon l'élément de structure secondaire où se situe le résidu. Gribskov [Gribskov, 1992] va même plus loin que cela puisqu'il suggère que les scores de substitution d'un monomère par un autre devraient tenir compte des contraintes intervenant au niveau de la structure tertiaire de la macromolécule, et donc de nature non plus locale mais globale. Nous sommes impuissants à capter ce genre d'information d'origine géométrique avec nos modèles. Même les matrices de substitutions établies sur la base d'alignements structuraux (section 3.2.2.2) n'en tiennent pas totalement compte. Observons toutefois que notre incapacité n'est pas liée à la contrainte d'ordre que nous nous sommes imposée. L'espace occupé par un monomère à l'intérieur d'une macromolécule ne dépend après tout que de ce monomère et de la molécule. Il ne dépend donc pas d'un système de référence relatif mais est une quantité que nous pourrions mesurer dans l'absolu si nous connaissions la structure cristallographique de la molécule. De la même façon, nous pourrions associer de manière absolue à chaque monomère la liste de ses 'voisins immédiats' (les monomères avec lesquels il est en contact). Considérer le contexte n'est ainsi pas la même chose que se libérer de l'ordre de succession des monomères le long de la chaîne polymérique. Le fait que nous ne le fassions pas à l'heure actuelle est une limitation des définitions que nous employons pour les modèles et pour la similarité entre modèles et mots, ce n'est pas une limitation de notre approche elle-même.