

Problème d'optimisation de recherche de cliques pour caractériser des familles de protéines.

F. Coste et G. Kerbellec

Projet Symbiose, IRISA/INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France
{francois.coste,goulven.kerbellec}@irisa.fr

Mots-clés : Bioinformatique, Cliques, Automates

1 Problématique

La base de données fédératrice UniProt [1] contient plus de 3 millions de séquences de protéines. Des experts ont annoté une partie de ces séquences en les regroupant par familles selon des critères de type fonction ou structure.

Nous avons proposé une nouvelle approche permettant la découverte automatique de signatures de familles de protéines. Basé sur l'inférence grammaticale, notre outil est implémenté dans un programme nommé Protomata [2]. À partir du jeu de séquences fourni, Protomata recherche dans un premier temps des zones caractéristiques de la famille. Puis, à partir d'un automate reconnaissant exactement les séquences données, Protomata produit un nouvel automate par fusions d'états à l'intérieur des zones caractéristiques. Cet automate sert alors à scanner les bases de données à la recherche de nouveaux membres potentiels de la famille d'intérêt.

Nous allons ici, nous focaliser uniquement sur la première partie de l'approche. Pour cela, nous introduisons un modèle nommé PLMA (Partial Local Multiple Alignment). Un PLMA correspond à un ensemble de fragments reliés par des relations de similarités significatives. Entre plusieurs PLMAs il peut exister des incompatibilités [3,2]. Cependant, il est possible de pondérer chaque PLMA de manière à pouvoir départager les PLMAs incompatibles.

Notre problématique est d'extraire un sous-ensemble de PLMAs de score total maximal sous certaines contraintes.

2 Formalisation de la Caractérisation

Soit P le jeu de protéines. On peut décomposer chaque séquence S de P en $\frac{|S|(|S|+1)}{2}$ fragments de positions et de tailles différentes. On construit le graphe $G = (V, E)$ dont les sommets V représentent les fragments F_x de P et dont les arêtes E sont valuées par la similarité entre les paires de fragments. Pour évaluer $w(F_i, F_j)$ nous utilisons l'option de pré-traitement du programme Dialign [4]. Cette pondération est corrélée à la probabilité de correspondance, au sens physico-chimique, induite par l'alignement de F_i et F_j . Le modèle utilisé est alors le graphe G dont seules les arêtes de poids $w > 0$ sont considérées.

On définit un PLMA comme un ensemble de fragments liés par une relation de similarité. Toute composante connexe C dans G est alors un PLMA qui représente une caractérisation locale et partielle de l'ensemble P .

Compte tenu du contexte biologique, un PLMA se doit d'exhiber un consensus fort. C'est pourquoi dans l'approche spécifique nommée Protomata-CL, le choix est le suivant : tout PLMA C est conservé si et seulement si C forme une clique dans G . Nous posons classiquement $W(C)$ le poids du PLMA correspondant à la somme des arêtes de C . Pour rechercher les caractéristiques de P , il s'agit donc, étant donné $G = \{V, E\}$, de trouver l'ensemble des cliques EC maximisant $\sum(W(C_x))$ sous contraintes de compatibilités.

Notre approche heuristique utilise alors un algorithme glouton classique. On dispose tous les PLMAs dans une pile PC que l'on ordonne en fonction de $W(C_x)$. Chaque candidat sera dépilé et ajouté à EC si et seulement s'il est compatible avec les PLMAs déjà présent dans EC . Une approche par recherche exhaustive des cliques par taille décroissante pour choisir PC [2] s'avère coûteuse pour les jeux de données à forte similarité. Dans le nouvel algorithme de Protomata-CL, nous estimons PC par le calcul d'une pile de fragments PF dont l'ordre est basé sur l'arité de chaque fragment à l'intérieur de G . On se sert alors de chaque fragment comme d'une graine pour

choisir les cliques (PLMAs de consensus fort), candidates.

On ne peut pas choisir l'ensemble des PLMAs comme une caractérisation globale d'une famille de protéine, car il existe des contraintes d'incompatibilités entre certains PLMAs. En effet, il existe une première contrainte nommée contrainte d'Unicité qui impose de ne pas aligner deux positions d'une même séquence sur la même position d'une troisième. Ensuite, il existe une contrainte dite d'Inconsistance qui interdit un croisement des positions qui ne respecterait pas l'ordre de chaque séquence. La librairie Gabios [3] est utilisée pour un traitement rapide de ces conditions. Enfin la notion de Préservation de PLMA est utilisée de manière à ne pas altérer les informations obtenues par la fusion des états d'un PLMA dans un automate. On rejette donc toute interférence qui proviendrait de l'association d'une position d'un PLMA déjà choisi avec une position n'appartenant pas à ce PLMA.

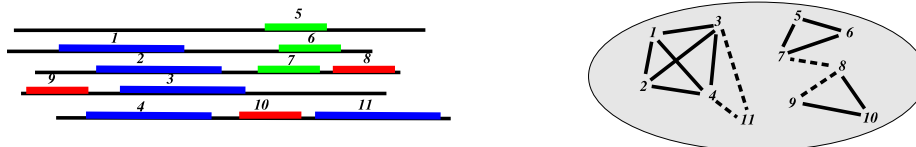


FIG. 1. Jeu P et modèle G . $\{1,2,3,4\}$, $\{5,6,7\}$ et $\{8,9,10\}$ sont des PLMAs de consensus fort. $\{1,2,3,4\}$ et $\{8,9,10\}$ sont incompatibles par contrainte d'Inconsistance.

Algorithme 1 Recherche de PLMAs de consensus fort dans Protomata-CL

ENTRÉES: jeu de séquences de protéines P

SORTIES: ensemble EC des PLMAs choisis pour la caractérisation de P

$EC \leftarrow \emptyset$; $G \leftarrow CreerGraphe(P)$

$PF \leftarrow Sommets(G)$

$PF \leftarrow PF.Trier(w())$

pour tout $F \in PF$ **faire**

$C \leftarrow ChercheClique(F, G)$

si $EstCompatible(C, EC)$ **alors**

$EC \leftarrow C$; $PurgeAretesIncompatibles(C, G)$

fin si

fin pour

 Retourner(EC).

3 Conclusion

La modélisation à partir d'un ensemble de fragments permet de poser le cadre général dans lequel nous évoluons. La nouvelle implémentation heuristique de Protomata-CL a pour but l'obtention de solutions performantes pour des temps d'exécution raisonnables. Ceci a pu être observé sur des données réelles. Cependant, la réalisation d'un algorithme exact pour résoudre ce problème, par programmation linéaire ou autre méthode de recherche opérationnelle, est toujours une question ouverte.

Références

1. Bairoch, A. et al. : The Universal Protein Resource (UniProt). 1362-4962, Nucleic Acids Res (2005)
2. Coste, F. and Kerbellec, G. : Learning Automata on Protein Sequences. JOBIM 2006, Bordeaux (2006)
3. Abdeddaïm S. and Morgenstern B : Speeding up the DIALIGN multiple alignment program by using the 'greedy alignment of biological sequences library' (GABIOS-LIB). Computer Science 2066 (2001)
4. Morgenstern B. : DIALIGN 2 : improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics (1999)