Détection de domaines dans des séquences génomiques: un problème de couverture optimale

P. Veber, S. Tempel, R. Andonov, D. Lavenier, et J. Nicolas

IRISA, Campus de Beaulieu 35042 Rennes Cedex
Jacques.Nicolas@irisa.fr

1 Introduction: un problème de segmentation

Le problème de la segmentation d'une séquence d'ADN en unités biologiquement pertinentes s'est posé dès l'origine de la production du code génétique de chromosomes par séquençage. Il s'agit d'extraire de l'enchaînement des acides nucléiques qui compose chaque séquence une vue plus abstraite en délimitant dans celle-ci un enchaînement de sous-séquences caractéristiques. Des travaux très divers ont permis ainsi de segmenter les séquences suivant leur utilisation de l'alphabet (isochores, ilôts CpG), de détecter les zones correspondant à des origines ou des terminaisons de réplication ou encore d'effectuer dela séparation entre régions codantes et non codantes [LBGHG02].

Toutes les méthodes reposent sur une analyse compositionnelle de la séquence, basée sur différentes statistiques de mots. Les deux principales techniques sont l'estimation puis la segmentation par modèles de Markov cachés (HMM), et la segmentation binaire récursive de séquence, en partant de la séquence complète puis en déterminant récursivement le meilleur point de coupure des fragments [ARLR02], [LBGHG02], [OCHBG04].

Cependant, il est possible de considérer le problème de la segmentation à un niveau plus fin, où les segments sont caractérisés par l'ensemble des mots qu'ils représentent, c'est à dire un langage. Dans tous les cas, le problème peut être posé dans un cadre commun et Gionis et Mannila ont proposé récemment une formulation claire de ce cadre [GM03]. L'analyse de l'architecture en domaines des séquences peut être posé comme un problème d'optimisation, le "(k,h)-segmentation problem", et qui consiste à trouver la meilleure segmentation d'une séquence de longueur n en k fragments, chaque fragment appartenant à un ensemble de h sources avec $h < k$. Ce problème est démontré NP-complet sous des hypothèses assez larges. Nous proposons un cadre légèrement différent fondé sur l'hypothèse que les sources sont des séquences qui ont été copiées puis ont divergé par mutation dans les génomes pour former des familles identifiables. Nous supposons donc l'observation d'un ensemble de séquences S et d'un ensemble de sources D (les domaines) et recherchons un codage optimal de S comme une concaténation d'éléments de D. Du fait que la segmentation prenne en compte l'ordre sur la séquence et plusieurs séquences en parallèle, on peut espérer produire ainsi un partitionnement plus fin que par une analyse basée sur la seule composition d'une séquence.

2 Couverture d'un ensemble de séquences

En pratique, les domaines sur un ensemble de séquences biologiques S d'une même famille peuvent être repérés par des "patterns" caractéristiques que l'on retrouve sur plusieurs séquences de la famille avec des variations mineures. On suppose ici que ces patterns ont été déterminés par apprentissage automatique et forment un ensemble D.

L'ensemble des occurrences des domaines de D dans une séquence S_l de S forme un graphe orienté G_l, appelé *graphe de couverture* si l'on décide de relier deux occurrences lorsqu'elles peuvent se suivre dans une segmentation valide. À ce graphe, on ajoute un sommet initial et un sommet terminal virtuels s_l et t_l, et les arcs $s_l \to o$ (resp. les arcs $o \to t_l$) pour tout occurence o pouvant débuter (resp. terminer) une segmentation. G_l est un graphe acyclique et tout chemin de s_l à t_l représente une segmentation valide. De plus les arcs de G_l sont pondérés : soient un arc $i \to j$ de G_l et notons (d_i, f_i) et (d_j, f_j) les positions de début et de fin des occurrences i et j respectivement dans S_l. Le poids de $i \to j$ est $c^l_{ij} = |d_j - f_i|$.

Si l'on cherche alors à redécrire le plus complètement possible les séquences de S à l'aide d'occurrences de domaines, tout en utilisant un minimum de domaines différents, on peut chercher

un chemin de s_l à t_l dans chaque graphe G_l en minimisant la somme des distances obtenues et d'une pénalité fonction des domaines utilisés. À cette fin, on pose d_k le coût (positif ou nul) associé à l'utilisation du domaine $k \in D$.

Spécifions à présent ce problème d'optimisation sous la forme d'un programme linéaire :

Variables

On introduit les variables binaires

- x_{ij}^l où l est une des séquences de S, $i \to j$ un arc de G_l. On pose $x_{ij}^l = 1$ ssi le chemin de s_l à t_l passe par l'arc $i \to j$
- y_k où k est un domaine de D. On pose $y_k = 1$ ssi le domaine k est utilisé pour segmenter l'une des séquences de S.

Modèle

Min

$$\sum_{l,i,j} c_{ij}^l x_{ij}^l + \sum_k d_k y_k$$

Subject to

$Ax = b$ (contraintes de flot)

$y_k \geq x_{ij}^l$ avec i ou j occurrence de k dans S_l

3 Premiers résultats sur des séquences de transposons

Les transposons sont des éléments qui se déplacent ou sont copiés d'une position chromosomique à une autre [FJW02]. Reconstituer la dynamique au cours de l'évolution de ces éléments abondants voire majoritaires dans les génomes eucaryotes [KL01], est un problème qui peut être posé dans le cadre de la modélisation que nous venons de définir.

Nous nous sommes-nous attachés à l'étude de familles particulière de transposons appelées AtREP chez *Arabidopsis thaliana* et caractérisées par une grande variabilité de leurs séquences.

Il existe ainsi 48 copies de la famille AtREP21 dans le génome d'*Arabidopsis thaliana*. Ces éléments mesurent de 312 à 1012 pb. L'algorithme d'identification des domaines a découvert 121 domaines qui correspondent à des fragments répétés avec erreurs. Le coût associé à un domaine est la taille moyenne de ses occurrences. Nous avons choisi d'appliquer une simple approche gloutonne qui procède séquentiellement sur l'ensemble des séquences. Pour chaque séquence, l'ensemble des combinaisons possibles de fragments est exploré et le minimum calculé. L'ensemble des domaines utilisé pour une séquence est conservé pour la suivante. Afin de diminuer l'impact du premier choix de séquence, qui détermine en général une bonne partie des domaines, on réitère le processus pour chaque choix possible de première séquence et on conserve le minimum de l'ensemble des résultats. On obtient ainsi 76 domaines avec un score global de 5811. Pour vérifier la précision de notre approche, nous avons testé toutes les permutations possibles des deux premières séquences. Le score minimum obtenu est de 5599 avec 73 domaines, ce qui représente un faible écart par rapport au premier score que nous avons trouvé.

Il existe des familles plus importantes dont la segmentation ne peut être résolue parl'approche gloutonne. Ainsi, ATtREP3 contient 141 copies et l'algorithme de détection des domaines produit 695 domaines : de tels jeux de données constituent un bon test de différentes méthodes d'optimisation et sont mis à disposition de la communauté sur http://genoweb.univ-rennes1.fr/Serveur-GPO/outils/repeatsAnalysis/DOMAIN/AtREP.

Références

[ARLR02] Rajeev K. Azad, J. Subba Rao, Wentian Li, and Ramakrishna Ramaswamy. Simplifying the mosaic description of dna sequences. *Physical Review E*, 2002.

[FJW02] C. Feschotte, N. Jiang, and S.R. Wessler. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.*, 3:329–341, 2002.

[GM03] A. Gionis and H. Mannila. Finding recurrent sources in sequences. In *RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 123–130, New York, NY, USA, 2003. ACM Press.

[KL01] M.G. Kidwell and D.R Lisch. Perspective: transposable elements and host genome evolution. *Trends Ecol. Evol.*, 15:95–99, 2001.

[LBGHG02] W. Li, P. Bernaola-Galvan, F. Haghighi, and I. Grosse. Applications of recursive segmentation to the analysis of dna sequences. *Computers & Chemistry*, 2002.

[OCHBG04] José L. Oliver, Pedro Carpena, Michael Hackenberg, and Pedro Bernaola-Galván. Isofinder: computational prediction of isochores in genome sequences. *Nucleic Acids Research*, 32(Web-Server-Issue):287–292, 2004.