Modèle de PLNE pour la recherche de cliques de poids maximal

Noël Malod-Dognin, Rumen Andonov, Nicola Yanev, Jean-François Gibrat

Modèle de PLNE pour la recherche de cliques de poids maximal*

N. Malod-Dognin[1**], R. Andonov[1], N. Yanev[2], et J-F. Gibrat[3]

[1] IRISA, Campus de Beaulieu, 35042 Rennes, France
[2] University of Sofia, Bulgaria
[3] MIG - INRA, Domaine de Vilvert, 78350 Jouy en Josas Cedex, France
nmaloddg@irisa.fr, randonov@irisa.fr, choby@math.bas.bg, jean-francois.gibrat@jouy.inra.fr

Mots-clés : Modèle de PLNE, clique de poids maximal, alignement de structures de protéines.

1 Introduction

Nous présentons un nouveau modèle de PLNE[4] pour la recherche de cliques, appliqué aux alignements de structures tridimensionnelles (3D) dans le logiciel VAST[1][5].

En biologie structurelle, il est admis que deux protéines ayant des structures 3D similaires ont de fortes chances de partager la même fonction et de dériver d'un ancêtre commun [2]. De nombreuses méthodes ont été créées pour mettre en évidence cette similarité (par exemple [1,3,4]). La plupart se basent sur un alignement (mariage stable, "one to one matching") entre les éléments de deux protéines (acides aminés, structures secondaires, carbones alpha). Parmi ces méthodes, nous nous intéressons à la méthode VAST (Vector Alignment Search Tool), qui aligne les éléments de structures secondaires (SSEs) des protéines. Cet alignement est modélisé sous la forme d'une recherche de clique de poids maximal sur les arêtes dans un graphe particulier appelé graphe d'alignement (problème NP-Difficile[5]). Pour trouver cette clique, VAST génère et évalue toutes les cliques possibles en utilisant l'algorithme de Bron et Kerbosh[6]. Cependant, la vitesse de cet algorithme restreint l'utilisation de VAST à de petites bases de données de structures 3D.

Pour remédier à cet inconvénient, nous proposons un modèle de PLNE qui résout efficacement la recherche de cliques lorsque le graphe d'alignement est représenté sous forme d'une grille.

2 Reformulation du problème

L'alignement des SSEs de deux protéines, $P1$ et $P2$, peut-être reformulé afin que le graphe d'alignement ait la forme d'une grille, ou chaque ligne représente une SSE de $P1$, et chaque colonne une SSE de $P2$. Les protéines $P1$ et $P2$ sont représentées par deux ensembles N_1 et N_2 finis et ordonnés de SSEs. Soit $G = (N, E)$ le graphe d'alignement de $P1$ et $P2$. Sur la ligne $i \in N_1$ et la colonne $k \in N_2$, le noeud $n_{i,k}$ existe (noté $n_{i,k} \in N$), ssi les SSEs i et k sont alignables[6]. À chaque noeud $n_{ik} \in N$, on associe un poids $S_{ik} \in R$. Une arête e_{ikjl} entre les noeuds n_{ik} et n_{jl} existe ($e_{ikjl} \in E$) ssi l'ordre topologique est conservé ($i < j$ et $k < l$) et si le couple de SSEs (i,j) est alignable avec le couple de SSEs (k,l). À chaque arête $e_{ikjl} \in E$, on associe un poids $C_{ikjl} \in R$.

Pour VAST, le meilleur alignement des SSEs correspond dans le graphe G à la clique de poids maximal sur les arêtes.

3 Modèle de recherche de clique

Soient x_{ik} les variables de décisions associées au noeuds $n_{ik} \in N$, et y_{ikjl} celles associées aux arêtes $e_{ikjl} \in E$. À chaque noeud n_{ik} nous associons l'ensemble $d^{+}_{col}(ik)$ des colonnes contenant des successeurs de n_{ik}, et l'ensemble $d^{-}_{col}(ik)$ des colonnes contenant des prédécesseurs de n_{ik}.

* Recherche effectuée dans le cadre de l'ANR "Calcul Intensif projet PROTEUS", ANR-06-CIS6-008.
** Auteur correspondant, subventionné par la Région Bretagne.

[4] Programmation Linéaire en Nombre Entier.
[5] Disponible via internet. http ://migale.jouy.inra.fr/outils/mig/vast/
[6] Les conditions d'alignement de deux SSEs et de deux couples de SSEs sont explicitées dans [1].

En exploitant la représentation en grille de G, et en se basant sur le modèle présenté dans [7] pour la recherche d'un chemin augmenté, nous proposons le modèle de PLNE suivant :

$$\max \sum_{i,k \mid n_{ik} \in N} S_{ik}\ x_{ik} + \sum_{i,k,j,l \mid e_{ikjl} \in E} C_{ikjl}\ y_{ikjl}. \quad (1)$$

Avec les contraintes :

$$\sum_{k \mid n_{ik} \in N} x_{ik} \leq 1, \quad \forall i \in N_1. \quad (2)$$

$$\sum_{i \mid n_{ik} \in N} x_{ik} \leq 1, \quad \forall k \in N_2. \quad (3)$$

$$x_{ik} \geq \sum_{j \mid e_{ikjl} \in E} y_{ikjl}, \quad \forall n_{ik} \in N,\ \forall l \in d^{+}_{col}(ik). \quad (4)$$

$$x_{jl} \geq \sum_{i \mid e_{ikjl} \in E} y_{ikjl}, \quad \forall n_{jl} \in N,\ \forall k \in d^{-}_{col}(jl). \quad (5)$$

$$\sum_{i \mid n_{ik} \in N} x_{ik} + \sum_{j \mid n_{jl} \in N} x_{jl} - \sum_{i,j \mid e_{ikjl} \in E} y_{ikjl} \leq 1, \quad \forall k \in N_2,\ \forall l \in N_2,\ k < l. \quad (6)$$

$$x_{ik} \in \{0,1\}, \quad \forall n_{ik} \in N. \quad (7)$$

$$y_{ikjl} \in \{0,1\}, \quad \forall e_{ikjl} \in E. \quad (8)$$

Les contraintes (2) et (3) forcent le mariage stable en activant au plus un noeud par ligne et au plus un noeud par colonne. Les contraintes (4) et (5) n'autorisent l'activation d'une arête qu'entre deux noeuds actifs. Enfin, les contraintes (6) forcent l'activation d'une arête entre deux colonnes ayant chacune un noeud actif. Les contraintes d'intégralité des y_{ikjl} (8) peuvent-être relâchées, car les autres contraintes forcent les y_{ikjl} à valoir soit 0 soit 1.

En fonction des valeurs des poids, ce modèle permet de résoudre quatre problèmes de cliques différents : La clique de poids maximal sur les noeuds et les arêtes (lorsque les S_{ik},$C_{ikjl} \in R$). La clique de poids maximal sur les noeuds (lorsque les C_{ikjl} sont nuls). La clique de poids maximal sur les arêtes (lorsque les S_{ik} sont nuls, comme dans VAST). La clique de cardinalité maximale (lorsque les S_{ik} valent 1, et que les C_{ikjl} sont nuls).

4 Conclusion

Ce modèle a été validé dans VAST à l'aide de CPLEX-10. Pour résoudre efficacement ce modèle, nous avons implémenté un algorithme de "branch and bound" dédié, qui permet d'améliorer jusqu'à un facteur 10 la vitesse d'alignement de grosses protéines par rapport à VAST original.

Références

1. J-F. Gibrat, T. Madej and S.H. Bryant. "Surprising similarities in structure comparison". *Curr. Opin. Struct. Biol*, vol 6, p 377-385, 1996.
2. H.J. Greenberg, W.E. Hart and G. Lancia. "Opportunities for Combinatorial Optimization in Computational Biology". *Informs Journal on Computing*, vol 16, n 3, p 211-231, 2004.
3. M.L. Sierk and G.J. Kleywegt. "Déjà Vu All Over Again : Finding and Analyzing Protein Structure Similarities". *Structure*, vol 12, p 2103-2111, 2004.
4. G. Mayr, F.S. Domingues and P. Lackner. "Comparative Analysis of Protein Structure Alignments". *BMC Structural Biology*, vol 7 :50, 2007.
5. R.M. Karp. "Reducibility among combinatorial problem". *Complexity of Computer computations*, p 85-103, 1972.
6. C. Bron and J. Kerbosh. "Finding All Cliques of an Undirected Graph [H]". *Communications of the ACM*, vol 16, n 9, p 575-579, 1973.
7. R. Andonov, N. Yanev and N. Malod-Dognin. "Towards Structural Classification of Proteins based on Contact Map Overlap". *Publication interne IRISA*, n 1872, novembre 2007.