Identification rapide de familles protéiques par dominance

Mathilde Le Boudic-Jamin, Noël Malod-Dognin, Alexandre Cornu, Jacques Nicolas, Rumen Andonov

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Identification rapide de familles protéiques par dominance

Mathilde Le Boudic-Jamin — Noël Malod-Dognin — Alexandre Cornu — Jacques Nicolas — Rumen Andonov

N° 7696

Mars 2011

_ Computational Biology and Bioinformatics _

ISSN 0249-6399 ISRN INRIA/RR--7696--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SOPHIA ANTIPOLIS - MÉDITERRANÉE

Identification rapide de familles protéiques par dominance

Mathilde Le Boudic-Jamin *, Noël Malod-Dognin †, Alexandre Cornu*, Jacques Nicolas*, Rumen Andonov*

Theme : Computational Biology and Bioinformatics
Computational Sciences for Biology, Medicine and the Environment
Équipes-Projets Symbiose et ABS

Rapport de recherche n° 7696 — Mars 2011 — 5 pages

Résumé : La comparaison de stuctures protéiques est une opération fréquente et importante dans le domaine de la bioinformatique. Elle apporte des informations aidant à la détermination des fonctions d'une protéine. Néanmoins, le problème sous-jacent est NP-complet. Différentes approches d'analyse existent: certaines basées sur la superposition de coordonnées (e.g. VAST) et d'autres sur les distances internes conservées dans les structures. L'objectif est donc d'identifier et de classer rapidement des structures similaires. Nous avons classé des structures de la base de données CATH avec un programme nommé A_purva qui utilise l'approche CMO (Conctact Map Overlap). Nous montrons que ce dernier a permis de prédire correctement la classification de 92% des structures soumises et que l'introduction de la notion de dominance a réduit considérablement les temps de classement des protéines.

Mots-clés : Comparaison de protéines, alignement, séparation et évaluation, dominance

* Symbiose, INRIA Rennes - Bretagne Atlantique, France.
† A.B.S., INRIA Sophia Antipolis - Méditerranée, France.

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Fast Protein Family Identification Using Dominance

Abstract: Structural comparison of proteins is a frequent and important operation in bioinformatics, giving precious information for determining the possible functions of proteins. Unfortunately, the corresponding optimization problems are often NP-Hard. Different analysis approaches exist: Most are based on the superimposition of residue coordinates (like VAST) or on the comparison of internal distances. The objective is to quickly identify and classify similar structures. We used the comparison tool A_purva, which is based on Contact Map Overlap (CMO), to classify protein structure coming from the CATH database. The obtained results show that A_purva was able to correctly classify 92% of the structures, and that introducing the notion of dominance drastically reduces the computational time needed for classifying the protein structures.

Key-words: Protein comparison, alignment, branch and bounds, dominance

1 Introduction

La comparaison de stuctures protéiques est une opération fréquente et importante dans le domaine de la bioinformatique. Elle apporte des informations aidant à la détermination des fonctions d'une protéine. Néanmoins, le problème sous-jacent est NP-complet [1]. Différentes approches d'analyse existent : certaines basées sur la superposition de coordonnées (e.g. VAST [2]) et d'autres sur les distances internes conservées dans les structures. L'objectif est donc d'identifier et de classer rapidement des structures similaires. Nous avons classé des structures de la base de données CATH [3] avec un programme nommé A_purva [4, 5] qui utilise l'approche CMO [6] (Conctact Map Overlap). Nous montrons que ce dernier a permis de prédire correctement la classification de 92% des structures soumises et que l'introduction de la notion de dominance a réduit considérablement les temps de classement des protéines.

2 Cartes de contacts, approche CMO & A_purva

Une structure protéique peut être representée par une carte de contact : un graphe G=(V,E) où l'ensemble des sommets V correspond aux résidus de la protéine et E est l'ensemble des arêtes ou contacts ; un contact est créé entre deux sommets lorsque les résidus correspondants sont spatialement proches (distance euclidienne inférieure à 7.5 Å) (voir figure (1a,1b)).

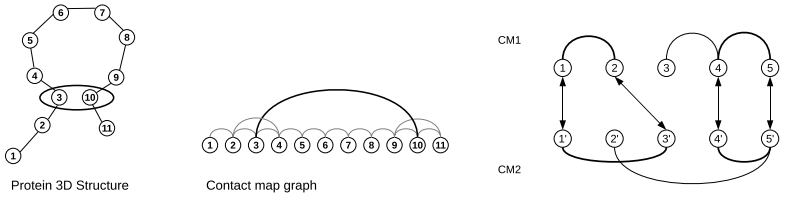

FIGURE 1 – a. Contact entre deux résidus, b. Carte de contacts, c. Alignement de deux cartes de contacts (Quatre résidus sont alignés et deux contacts sont conservés (NCC=2)).

CMO compare deux structures protéiques en cherchant l'alignement des résidus qui maximise le nombre de contacts communs (NCC) comme représenté sur la figure (1c). A_purva est un solveur exact pour CMO, utilisant un algorithme de type séparation et évaluation (Branch & Bound). Pour un temps de calcul donné (T), A_purva aligne deux cartes de contacts (CM1,CM2) et renvoie un alignement, des bornes supérieure (UB) et inférieure (LB) de NCC, et un score de similarité SIM(CM1,CM2) = 2 * LB/($|E_1|$+$|E_2|$).

3 Optimisation des résultats en sortie d'A_purva

Lorsque deux structures sont similaires, leur score SIM est élevé et le temps de calcul d'A_purva est petit, en revanche des structures dissimilaires ont un score SIM faible et le temps de comparaison varie énormément. Lorsqu'une instance (comparaison structure-structure) est résolue de manière optimale,

LB=NCC=UB. Par contre, lorsque le calcul est arrêté par le paramètre temps, LB≤NCC≤UB.

Soit Q une requête qu'on cherche à classer et P={P_1,P_2,..., P_n} un jeu de structures protéiques de classification connue. L'identification de la famille de Q passe par la recherche de son plus proche voisin ((nearest neighbour)) dans P tel que (nearest neighbour) = arg max SIM(Q,P_i) avec $P_i \in$ P. Afin de réduire les temps de calcul, nous effectuons un premier calcul borné par un temps très court et nous introduisons un nouveau score approché : le score potentiel (SPOT) calculé à partir de la borne UB fournie par A_purva tel que SPOT= 2 * UB/(|E_1|+|E_2|). On peut alors filtrer les structures rapidement sur la base de ces deux scores.

Dominance entre des instances : Soit P_1 et $P_2 \in$ P. Si SIM(Q,P_2) > SPOT(Q,P_1) alors l'instance (Q,P_2) *domine* l'instance (Q,P_1). Il est inutile de résoudre exactement l'instance (Q,P_1) car cela démontre que P_1 est moins similaire que P_2 à Q. Le calcul de l'instance (Q,P_1) ne sera pas poursuivi.

L'introduction de la dominance a permis de réduire de dix fois l'analyse des 50 000 instances utilisées lors du concours SHREC'10 [7] tout en garantissant l'exactitude des résultats.

4 Conclusion

Dans cette étude nous avons introduit la notion de dominance entre instances qui permet d'améliorer significativement les temps de calculs et garantit l'obtention dans tous les cas du (nearest neighbour). L'une des perpesctives de cette approche est l'utilisation efficace d'A_purva sur de grands jeux de données protéiques.

Références

[1] RH. Lathrop. The protein threading problem with sequence amino acid interactions preferences in NP-complete. *Protein Eng.*, 7(9) :1059-68, 1994.

[2] JF. Gibrat,T. Madej, SJ. Bryant. Surprising similarities in structure comparison. *Curr Opin Struct Biol*, 6(3):377-385, 1996.

[3] CA. Orengo, AD. Michie, S. Jones, DT. Jones, MB. Swindells and JM Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093-108, 1997.

[4] R. Andonov, N. Malod-Dognin, and N. Yanev. Maximum contact map overlap revisited. *J. Comput. Biol.*, 18(1):1–15, 2011.

[5] R. Andonov, N. Yanev, and N. Malod-Dognin. An efficient lagrangian relaxation for the contact map overlap problem. In *WABI '08 : Proc. of the 8th Int. Workshop on Algorithms in Bioinformatics*,162–173. Springer-Verlag, 2008.

[6] A. Godzik and J. Skolnick. Flexible algorithm for direct multiple alignment of protein structures and sequences. *CABIOS*, 10:587–596, 1994.

[7] L. Mavridis, V. Venkatraman, D.W. Ritchie, N. Morikawa, R. Andonov, A. Cornu, N. Malod-Dognin, J. Nicolas, M. Temerinac-Ott, M. Reisert, H. Burkhardt, and A.Axenopoulos. Shrec-10 track : Protein models. In I. Pratikakis,

M. Spagnuolo, T. Theoharis, and R. Veltkamp, editors, *3DOR : Eurographics Workshop on 3D Object Retrieval*, 117–124, Norrköping, Sweden, 2010. The Eurographics Association.

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399