Apprentissage de fonctions de tri pour la prédiction d'interactions protéine-ARN

Adrien Guilhot-Gaudeffroy[(1),(2),(3),♯], Jérôme Azé[(1),(3),(4)], Julie Bernauer[(2),(3)], Christine Froidevaux[(1),(3)]

[(1)]LRI, CNRS UMR 8623, Université Paris-Sud, 91405 Orsay
[(2)]LIX, CNRS UMR 7161, École Polytechnique, 91120 Palaiseau
[(3)]Projet AMIB, INRIA Saclay-Île de France, 91120 Palaiseau
[(4)]LIRMM, CNRS UMR 5506, UM2, CC 477, 34095 Montpellier
[♯]adrienguilhot@lri.fr

Résumé. Les fonctions biologiques dans la cellule mettent en jeu des interactions 3D entre protéines et ARN. Les avancées des techniques exérimentales restent insuffisantes pour de nombreuse applications. Il faut alors pouvoir prédire *in silico* les interactions protéine-ARN. Dans ce contexte, nos travaux sont focalisés sur la construction de fonctions de score permettant d'ordonner les solutions générées par le programme d'amarrage protéine-ARN RosettaDock. La méthodologie d'évaluation utilisée par RosettaDock impose de trouver une fonction de score s'exprimant comme une combinaison linéaire de mesures physico-chimiques. Avec une approche d'apprentissage supervisé par algorithme génétique, nous avons appris différentes fonctions de score en imposant des contraintes sur la nature des poids recherchés. Les résultats obtenus montrent l'importance de la signification des poids à apprendre et de l'espace de recherche associé.

1 Introduction

La plupart des mécanismes cellulaires mettent en jeu des complexes protéine-ARN. La compréhension de leurs fonctions dans un but thérapeutique ne peut se faire que par une connaissance fine des mécanismes moléculaires. Même si plus d'un millier de structures 3D de complexes protéine-ARN sont disponibles dans la *Protein Data Bank* [1], base de données de référence des structures 3D, la résolution expérimentale reste longue et coûteuse, parfois même impossible. Les travaux présentés dans cet article sont focalisés sur l'amélioration d'une des approches de référence dans le domaine de la prédiction de l'amarrage (docking) de structures 3D *in silico* : RosettaDock (Gray et al. (2003)). L'objectif de ces approches est de modéliser la protéine et l'ARN et d'en prédire les assemblages 3D les plus probables. De nombreuses méthodes, dont RosettaDock, fonctionnent en deux phases imbricables l'une dans l'autre : (1) génération d'un large ensemble de candidats [2] et (2) évaluation de ces candidats pour ne retenir que les plus plausibles. La "qualité" des candidats est évaluée avec une fonction de score

1. `http://www.wwpdb.org/`
2. assemblage de la protéine et de l'ARN