Régression binomiale négative avec termes d'interaction: un modèle pour étudier le lien entre protéines et boucles ADN

Raphaël Mourad*[1]

[1]Laboratoire de biologie moléculaire eucaryote du CNRS (LBME) – CNRS : UMR5099, Université Paul Sabatier (UPS) - Toulouse III – Bât IBCG 118 Route de Narbonne 31062 TOULOUSE CEDEX 4, France

Résumé

Introduction

La conformation en 3 dimensions (3D) du génome joue un rôle central dans un certain nombre de processus cellulaires clefs tels que la régulation de son expression, la réplication de l'ADN ou encore la réparation de l'ADN. Récemment, l'avènement de nouvelles méthodes de séquencage à haut débit des contacts chromatiniens (high-throughput chromosome conformation capture, Hi-C) ont permis de caractériser physiquement la présence de milliers de boucles chromatiniennes chez l'homme, notamment entre frontières de domaines 3D, et entre promoteurs et enhanceurs. Ces analyses ont aussi montré que ces boucles impliquaient le plus souvent des motifs de fixation de la protéine insulatrice CTCF qui permet le recrutement d'un cofacteur essentiel, la cohésine. La méthode actuelle de référence pour l'identification des déterminants et facteurs moléculaires des boucles ADN consiste en l'analyse différentielle des interactions chromatiniennes entre cellules sauvages et cellules mutées pour le gène suspecté. Un frein majeur à l'identification systématique des déterminants moléculaires est le coût prohibitif, ainsi que la complexité, des expériences Hi-C à réaliser pour chaque facteur à tester. C'est une des raisons majeures pour laquelle peu de facteurs ont été découverts à ce jour.

Résultats

Une solution au problème consiste à intégrer données de ChIP-seq (fixation de protéines sur l'ADN) aux données d'interactions Hi-C. La régression binomiale négative avec termes d'interactions offrent de nombreux avantages à cette fin. Cette régression permet notamment (1) de lier quantitativement la présence de protéines à l'augmentation du nombres de contacts entre deux loci, (2) de prendre en compte la colocalisation entre protéines grâce à l'indépendance conditionnelle, (3) d'identifier les facteurs majeurs par lasso ou ridge regression, ou encore (4) de comparer différents modèles de boucles ADN à l'aide de critères de type AIC (Akaike information criterion) ou BIC (Bayesian information criterion). Les résultats préliminaires sont prometteurs puisqu'ils ont permis d'identifier des boucles jusqu'alors inconnues telles que les boucles impliquant la protéine DREF avec elle-même (bêta=147, p< 10-20), mais aussi d'autres boucles entre protéines différentes telles que DREF et BEAF-32 (bêta=133, p< 10-20), ou DREF et GAF (bêta=112, p< 10-20). L'analyse graphique des

*Intervenant

paramètres bêtas estimés a aussi pu révéler l'importance de DREF dans la plupart des boucles, et suggèrent un rôle de médiateur ubiquitaire dans la majorité des boucles.

Discussion

Pour l'instant, nous n'avons inclus dans le modèle que des termes d'interaction de second ordre qui permettent de détecter la présence de boucles entre deux protéines différentes (DREF et BEAF-32) ou entre une protéine et elle-même (DREF et lui-même). Mais nous pouvons aller plus loin en incluant par exemple des termes d'interaction d'ordre 4 où l'on teste l'influence d'une autre protéine comme un cofacteur qui permet de stabiliser les boucles entre les deux protéines se fixant à des motifs sur l'ADN. C'est par exemple le cas pour deux sites distants fixant des protéines CTCF et dont la boucle est stabilisée par la cohésine. Dans ce contexte, une approche par sélection de modèles par critère AIC/BIC et exploration d'un espace de recherche permettrait d'identifier de nouveaux modèles complexes de formation de boucles.