Application de l'analyse des mots et de l'analyse des graphes à l'interprétation des données d'expression génomique

Jacques van Helden[1,2], David Gilbert[2,3], Lorenz Wernisch[2] et Shoshana Wodak[1,2]

[1] UCMB. Université Libre de Bruxelles CP160/16. 50 av F.D. Roosevelt. B-1050 Bruxelles. Belgique. e-mail: jvanheld@ucmb.ulb.ac.be

[2] European Bioinformatics Institute. Genome Campus - Hinxton Cambridge CB10 1SD - UK.

[3] Department of Computing, City University, Northampton Square, London EC1V 0HB, UK.

Résumé

Nous abordons ici deux types de questions qui peuvent se poser pour l'interprétation de groupes de gènes co-regulés, tels ceux obtenus au moyen des puces à ADN et de techniques apparentées.

° Mécanisme de co-regulation : quels sont les sites cis-régulateurs susceptibles d'être responsables de la co-régulation transcriptionnelle ?

° Interprétation fonctionnelle: quelle est l'utilité, pour la cellule, d'exprimer de façon coordonnée le groupe de gènes considéré ?

Nous présentons des outils informatiques d'analyse de séquences et d'analyse de graphes permettant de répondre à ces questions, et montrons quelques résultats d'applications aux données de la levure *Saccharomyces cerevisiae*. Ces outils sont disponibles aux sur internet aux adresses http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools et http://www.ebi.ac.uk/research/pfmp

1. Introduction

Le développement des puces à ADN et de techniques apparentées permet de mesurer la réponse transcriptionnelle du génome complet d'un organisme à une perturbation contrôlée de l'environnement (milieu de culture, action d'un agent pharmacologique), ou du génotype (mutation, surexpression de gènes choisis). On peut de la sorte isoler des groupes de gènes co-régulés, caractérisés par leur réponse commune à un même signal. Nous abordons deux types de questions qui se posent pour l'interprétation de ces groupes de gènes :

° **Mécanisme de régulation**: quel est le mécanisme qui permet au groupe de gènes de répondre simultanément au stimulus ? L'approche consiste à tenter de découvrir des sites de régulation potentiels dans les régions en amont des gènes considérés.

° **Interprétation fonctionnelle**: quelle est l'utilité, pur la cellule, de réguler de façon coordonnée le groupe de gènes ? L'approche consiste à tenter de découvrir des voies métaboliques, connues ou inconnues, qui pourraient être catalysées par les gènes considérés.

Nous avons développé des outils permettant d'aborder ces questions, et montrons ici leur application pour l'interprétation de familles de gènes co-régulés chez la levure *Saccharomyces cerevisiae*.

2. Analyse des séquences régulatrices

2.1 *Caractéristiques des sites cis-régulateurs*

La régulation du niveau de transcription des gènes est assurée par une classe de protéines, les facteurs transcriptionnels, qui ont la capacité de reconnaître des segments d'ADN spécifiques, de s'y fixer, et d'interagir avec l'ARN polymérase pour augmenter (activation) ou diminuer (répression) le niveau d'expression des gènes avoisinants. La spécificité du site est déterminée par la structure du domaine protéique de liaison à l'ADN. La plupart des sites connus peuvent se regrouper en deux classes. La première classe consiste en une courte séquence de nucléotides adjacents fortement conservés (typiquement sur une longueur de 6 bases), entourés de quelques nucléotides partiellement conservés. Ce type de site est commun pour les facteurs de type Zn finger, homeodomain, leucine zipper, bHLH. Une autre classe de sites consiste en deux oligonucléotides très courts et bien conservés (typiquement 3 paires de bases), séparés par une région de taille fixe mais de contenu variable. Ces sites sont typiques des facteurs à domaines Zn cluster (qu'on trouve chez les fungi) et HTH (chez les prokaryotes). Certains de ces sites montrent une symétrie interne (répétitions en tandem, palindrome réverse), liée au fait que le facteur se lie à l'ADN sous la forme d'un homodimère.

Chez la levure, les sites cis-régulateurs se retrouvent exclusivement en amont des gènes qu'ils contrôlent, dans un intervalle de 800 paires de bases à partir du codon d'initiation. Leur efficacité ne dépend généralement ni de leur position précise, ni de leur orientation. Il est fréquent de trouver de multiples sites de fixations pour un même facteur transcriptionnel dans la région en amont d'un gène.

2.2 *Extraction de sites cis-régulateurs par analyse des fréquences de mots et paires de mots*

En nous basant sur ces propriétés des sites cis-régulatuers de levure, nous avons développé deux programmes spécialisés dans la découverte de sites potentiels au sein d'un groupe de séquences en amont.

Le premier programme, **oligo-analysis**, détecte dans un groupe de séquences tous les mots (typiquement les hexanucleotides) pouvant être considérés comme sur-représentés. Nous avons montré (van Helden *et al*., 1998) que cette approche, en dépit de sa simplicité, permet d'extraire de façon efficace un bon nombre de sites de régulation, tout en donnant un nombre très restreint de faux positifs.

Ce programme échoue toutefois dans la détection de certains sites cis-régulateurs, en particulier ceux où se fixent des facteurs à domaine Zn cluster. Ceci n'a rien d'étonnant, puisque ces sites comportent deux régions très courtes et très conservées, séparées par une région variable. Nous avons donc développé une stratégie complémentaire, qui consiste à compter la fréquence de toutes les paires de mots plus courts (typiquement trinucléotides) avec différentes valeurs d'espacement (entre 0 et 16). Ce programme, appelé **dyad-analysis** (van Helden *et al*., 2000) s'avère très efficace, non seulement pour la détection de sites de fixation de facteurs à Zn

cluster et à HTH, mais également pour les sites reconnus par d'autres classes de facteurs transcriptionnels.

2.3 *Critères statistiques pour l'analyse de fréquences de mots (et de paires de mots)*

L'efficacité des programmes ci-dessus dépend crucialement du choix des paramètres et d'une statistique appropriée.

Sélection de séquences non-redondantes. Avant même d'entamer l'analyse des séquences en amont, il est essentiel de s'assurer qu'on n'a pas inclus de séquences redondantes dans ces données. En effet, l'applicabilité des tests statistiques repose sur l'indépendance mutuelle des séquences. Deux sources de redondance peuvent se présenter :

- duplication récente d'un gène avec sa région en amont. De telles duplications sont particulièrement fréquentes dans les régions télomériques de la levure.
- région intergénique situé simultanément en amont de deux gènes voisins transcrits de façon divergente. Si les deux gènes font partie de l'ensemble initial, ladite région se retrouvera deux fois dans l'ensemble de séquences à analyser.

Il convient donc de "purger" les séquences de départ, en écartant celles qui présenteraient, sur l'un ou l'autre des deux brins, une trop forte similitude avec une autre séquence de l'ensemble.

Comment compter les mots ? Sur un seul ou deux brins ? Faut-il ou non considérer les occurrences successives d'un même mot qui chevauchent mutuellement ? Le choix du mode de comptage dépendra des caractéristiques attendues pour les sites régulateurs, en fonction de l'organisme considéré. Chez la levure, les meilleurs résultats sont obtenus par un comptage sur deux brins et sans chevauchement.

Comment évaluer la fréquence attendue pour chaque mot ? L'approche la plus simple, qui consiste à considérer tous les mots comme équiprobables, donne de piètres résultats, du fait de la haute fréquence en A et T des séquences de levure. La prise en compte de fréquences spécifiques pour chaque nucléotide donne déjà de meilleurs résultats, mais ne corrige pas certains effets d'agrégation préférentielle (par exemple les fréquentes chaînes poly-A/T dans le génome de la levure), et la réponse comporte beaucoup de faux positifs. L'approche la plus efficace consiste à constituer des tables de fréquences attendues, en se basant sur la fréquence observée pour chaque mot dans l'ensemble des régions intergéniques de levure.

Comment comparer la fréquence attendue (F_{att}) et la fréquence observée (F_{obs}) ? Plusieurs statistiques sont *a priori* envisageables (rapport F_{obs}/F_{att}, log likelihood, distribution de Poisson, Z-values, distribution binomiale). Pour l'analyse de petites familles de séquences, l'approche la plus appropriée consiste à utiliser la distribution binomiale.

Comment choisir le seuil de significativité ? L'analyse d'une seule famille nécessite de comparer les fréquences observée et attendue pour quelques milliers de mots. Le seuil de significativité doit être adapté au nombre de mots analysés, sous peine de surévaluer le niveau de sur-représentation.

Comment choisir la longueur des mots à analyser ? Les mots trops petits présentent un biais marqué par rapport aux distributions théoriques. A l'opposé, l'analyse de mots trop longs empêche la détection d'aucun mot significatif. En pratique, nous avons constaté que l'analyse des hexanucléotides donne d'excellents résultats dans la majorité des cas. Même en se restreignant à l'analyse d'hexanucléotides, le programme est à même d'isoler des sites plus longs, qui apparaissent sous la forme d'une série d'hexanucléotides mutuellement chevauchants.

Les concepts ci-dessus s'étendent facilement aux dyades de mots espacés (van Helden et al., 2000).

2.4 *Détection de sites cis-régulateurs à partir de familles issues de puces à ADN*

Après avoir présenté les méthodes et illustré leur résultats sur quelques régulons bien caractérisés chez la levure, nous discuterons brièvement de leur application à des familles issues des expériences de mesure d'expression génomique. Ceci nous permettra de discuter très brièvement d'un autre critère essentiel : la méthode utilisée pour le groupement des gènes en familles fonctionnelles.

3. Analyse des voies métaboliques

3.1 *Principes de base de régulation métabolique*

Les organismes vivants ont la capacité de modifier rapidement leur concentration interne en petites molécules (métabolites) grâce à la catalyse enzymatique des réactions. Le contrôle des flux de métabolites constitue un élément essentiel de la viabilité cellulaire, permettant de maintenir les métabolites essentiels à un niveau stationnaire de concentration (homéostasie). Une large panoplie de mécanismes moléculaires interviennent dans la régulation du métabolisme, et les enzymes et transporteurs sont contrôlés à de multiples niveaux: taux de transcription, stabilité de l'ARN, taux de traduction, activité protéique, localisation intracellulaire, dégradation de la protéine. Plusieurs mécanismes sont souvent associés pour le contrôle d'une même voie métabolique.

3.2 *Représentation des voies métaboliques sous forme de graphe*

L'ensemble des réactions biochimiques peut être représenté comme un graphe, dont les noeuds sont les métabolites et les réactions tandis que les arcs représentent les relations substrat-réaction et réaction-produit. Ce graphe comporte quelque 10.000 noeuds et 14.000 arcs. La complexité d'un tel graphe est énorme et un nombre virtuellement infini de chemins pourraient y être tracés.

Il faut toutefois réaliser que le nombre de voies effectivement suivies pour assurer le métabolisme ne représente qu'un sous-ensemble très restreint des voies potentielles. Par exemple, la base de données EcoCyc, qui recense de façon exhaustive les voies métaboliques connues chez *Escherichia coli*, comporte seulement 159 voies distinctes. Même en admettant

que certaines voies restent à découvrir, on ne s'attend pas à ce que ce nombre augmente considérablement pour cet organisme particulier, qui a constitué le modèle classique d'étude du métabolisme depuis plusieurs décennies.

Un bon nombre de voies restent cependant à découvrir chez d'autres organismes, dont le métabolisme présente des variations par rapport à celui d'*E.coli*. Il n'est pas rare en effet d'observer l'utilisation de voies alternatives pour une transformation donnée. Par exemple, la bactérie *Escherichia coli* synthétise la méthionine à partir de l'aspartate en 7 étapes. Chez la levure *Saccharomyces cerevisiae*, cette biosynthèse est effectuée en 6 étapes, dont 4 sont communes avec *E.coli*. Un exemple plus extrême est la biosynthèse de la lysine, réalisée selon des voies totalement différentes par *E.coli* et *S.cerevisiae* respectivement. En outre, des pans entiers du métabolisme demeurent largement inexplorés, par exemple les mécanismes de dégradation des substances toxiques par certaines bactéries, ou de résistance à des conditions extrêmes de l'environnement.

3.3 *Application de l'analyse de graphes à l'interprétation fonctionnelle des données d'expression génomique*

On observe fréquemment une régulation coordonnée de la transcription de l'ensemble des enzymes et transporteurs impliqués dans une voie métabolique commune. Sur base de ce paradigme, on peut s'attendre à ce que certains des groupes de gènes co-régulés obtenus par des mesures d'expression génomique correspondent à des voies métaboliques connues ou inconnues. La question est donc de découvrir, à partir d'un groupe de gènes, les voies métaboliques qu'ils pourraient catalyser. Partant d'un groupe de gènes co-régulés, il est aisé d'identifier ceux qui codent pour des enzymes, et les réactions que ces enzymes sont susceptibles de catalyser.

L'approche la plus simple consiste alors à détecter, parmi les voies métaboliques connues, celles qui regroupent une proportion significative de ces réactions. On peut ainsi associer la réponse transcriptionnelle donnée à une fonction métabolique particulière. Cette approche est toutefois limitée aux voies connues.

Une approche beaucoup plus flexible consisterait à appliquer l'analyse de graphe pour détecter toutes les voies métaboliques théoriques qui pourraient regrouper les réactions sélectionnées. Les réactions catalysées par le groupe de gènes co-régulés constituent un sous-ensemble de noeuds du graphe contenant l'ensemble des réactions possibles. La méthode consiste à extraire, à partir du graphe complet des réactions, un sous-graphe qui assure le maximum de connections entre ces noeuds de départ. Les connections peuvent être directes ou indirectes, en permettant d'intercaler un nombre limité de pas intermédiaires. Cette intercalation peut se justifier pour différentes raisons. D'une part, les techniques de mesure de l'expression à large échelle ne garantissent pas une réponse précise pour tous les gènes, et certains peuvent échapper à la détection. D'autre part, certaines enzymes pourraient figurer parmi les gènes de fonction indéterminée faisant partie du groupe de gènes co-régulés. Enfin, il n'est pas rare d'observer une voie métabolique dont certaines étapes ne sont pas régulées au niveau

transcriptionnel. Il est donc essentiel de pouvoir déduire une voie métabolique sur base d'un sous-ensemble de ses enzymes.

Après avoir extrait le sous-graphe de réactions et de métabolites, on peut le comparer aux voies métaboliques connues, entreposées dans les base de données. On s'attend à retrouver, pour une partie des voies déduites, les voies de biosynthèse et de dégradation classiques. Dans ce cas, l'approche la plus simple aurait suffi à établir la correspondance. Dans d'autres cas par contre, on se trouvera en présence de voies métaboliques partiellement différentes de celles des organismes modèles, ou même totalement nouvelles, qui pourront être soumises à une caractérisation expérimentale plus spécifique.

Cette méthode permettrait de caractériser rapidement le métabolisme des organismes dont la biochimie a été peu étudiée, et trouvera un champs d'application particulièrement intéressant pour découvrir les mécanismes de résistance à des conditions extrêmes.

Conclusions

Dans le contexte d'une approche génomique, l'analyse des séquences codantes est insuffisante pour assigner de façon systématique une fonction à chaque gène. Cette fonction dépend non seulement de la structure de la protéine, mais également du contexte dans lequel cette protéine exercera son activité. La prédiction de la fonction des gènes requiert l'intégration de différents niveaux d'informations.

La possibilité de mesurer la réponse transcriptionnelle à l'échelle d'un génome offre de nouvelles perspectives pour découvrir la fonction des gènes en se basant non pas sur leurs caractéristiques individuelles mais sur leur association dans des groupes fonctionnels. La combinaison de l'analyse des séquences régulatrices et de l'analyse des voies métaboliques devrait apporter deux types d'indices indépendants et complémentaires pour l'interprétation de ces groupes de gènes. Ces méthodes peuvent également s'appliquer à des groupes fonctionnels obtenus par d'autres approches, telles l'analyse des profils phylogénétiques, ou celle des fusions de gènes.

Références

VAN HELDEN, J., ANDRE, B., AND COLLADO-VIDES, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281:827-842.

VAN HELDEN, J., GILBERT, D. R., WERNISCH, L. & WODAK, S. (1999). Logical tools for querying and assisting annotation of a metabolic and regulatory pathway database. ISMB poster.

VAN HELDEN, J., ANDRE, B. AND COLLADO-VIDES, J. (2000). A web site for the computational analysis of yeast regulatory sequences. Yeast 16(2), 177-187.

VAN HELDEN, J., OLMO, M. AND PEREZ-ORTIN, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28(4), 1000-1010.

VAN HELDEN, J., RIOS, A. F. AND COLLADO-VIDES, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. *accepté*.

VAN HELDEN, J., GILBERT, D., WERNISCH, L., AND WODAK, S. Graph-based analysis of biochemical networks. *soumis.*