Analyse des groupes de gènes co-exprimés (AGGC) : un outil automatique pour l'interprétation des expériences de biopuces

Ricardo Martinez[1], Nicolas Pasquier[1], Claude Pasquier[2], Martine Collard[1], Lucero Lopez[3]

[1] Laboratoire I3S,
Université de Nice Sophia Antipolis,
2000, route des lucioles,
06903 Sophia-Antipolis, France.

[2] Laboratoire Biologie Virtuelle,
Université de Nice Sophia Antipolis
Centre de Biochimie, Parc Valrose,
06108 Nice cedex 2, France.

[3] Projet Odyssee
INRIA Sophia Antipolis,
2004 route des Lucioles
06905 Sophia Antipolis, France.

RÉSUMÉ. La technologie des biopuces permet de mesurer les niveaux d'expression de milliers de gènes dans différentes conditions biologiques générant ainsi des masses de données à analyser. De nos jours, l'interprétation de ces volumineux jeux de donnés à la lumière des différentes sources d'informations est l'un des principaux défis dans la bio-informatique. Nous avons développé une nouvelle méthode appelée AGGC (Analyse des Groupes de Gènes Co-exprimés) qui permet de constituer de manière automatique des groupes de gènes à la fois fonctionnellement riches, i.e. qui partagent les mêmes annotations fonctionnelles, et co-exprimés. AGGC intègre l'information issue des biopuces, i.e. les profils d'expression des gènes, avec les annotations fonctionnelles des gènes obtenues à partir des sources d'informations génomiques comme Gene Ontology. *Les expérimentations menées avec cette méthode ont permis de mettre en évidence les principaux groupes de gènes fonctionnellement riches et co-exprimés dans des expériences de biopuces*[1].

MOTS-CLÉS : biopuces, ontologie, co-expression, gène et annotation.

1 Introduction

L'analyse de données de biopuces en utilisant les diverses sources d'informations génomiques, continuellement alimentées par des volumes croissants de données, représente un challenge important. Ces sources d'informations sont sémantiques (taxonomies, thésaurus et ontologies), littéraires et bibliographiques (articles, librairies en ligne, etc.), et constituées de bases de données d'expériences et de nomenclatures. L'un des défis majeurs actuels dans ce domaine est l'intégration automatique des connaissances biologiques issues des sources d'informations mentionnées ci-dessus avec les données d'expression de gènes [ATT 01]. Un premier bilan des méthodes développées pour répondre à ce défi a été fait par Chuaqui [CHU 02].

Nous ciblons ici l'enrichissement de deux axes de recherche récemment développés, *séquentiel* et *a priori*, qui exploitent de multiples sources d'annotations telles que *Gene Ontology* (GO)[2]. Ces annotations sont des informations fonctionnelles, relationnelles et syntaxiques sur les gènes.

Dans l'axe séquentiel, partant des clusters de gènes co-exprimés (groupes de gènes qui ont un profil d'expression similaire), des sous-ensembles de gènes co-annotés (partageant la même annotation) sont détectés. Ensuite, la significativité statistique de ces sous-ensembles de gènes co-annotés est testée.. Parmi les méthodes dans cet axe citons Onto express [DRA 03], EASE [HOS 03] et THEA [PAS 04].

Dans l'axe a priori, partant des groupes fonctionnellement riches (GFR), i.e. des groupes de gènes co-annotés, l'information contenue dans les profils d'expression est intégrée. La significativité statistique des

[1] Informations supplémentaires et programme exécutable AGGC *:* http://www.i3s.unice.fr/~rmartine/AGGC

[2] Ontologie d'annotation de gènes *Gene Ontology project* : http://www.geneontology.org/

GFR est ensuite testée en utilisant un test basé sur un score enrichi [MOO 03], un test issu d'un z-score [KIM 05] ou un test basé sur une pc-value (distribution hypergéométrique) [BRE 04].
Notre approche, appelée AGGC (Analyse des Groupes de Gènes Co-exprimés), est inspirée de l'axe a priori : les GFR sont d'abord formés à partir de la GO est une fonction qui synthétise l'information contenue dans les données d'expression est appliquée afin d'obtenir une liste ordonnée de gènes [BRE 04]. Dans cette liste, les gènes sont triés par variabilité d'expression décroissante. La significativité statistique des GFR obtenus est alors testée à l'aide d'une preuve d'hypothèse de manière similaire à *Onto express* [DRA 03]. Finalement, nous obtenons des GFR co-exprimés et statistiquement significatifs. La méthode AGGC est une extension de la méthode IGA permettant d'obtenir tous les sous-ensembles possibles de GFR de gènes co-exprimés, sans se limiter au GFR constitué des gènes les plus exprimés.
Cet article est organisé de la manière suivante : dans la section 2 nous décrivons les données de validation ainsi que les outils utilisés; l'algorithme AGGC est décrit dans la section 3; les résultats obtenus sont présentés dans la section 4; la section 5 conclu l'article.

2 Données et Méthodes

2.1 Jeux de données et prétraitement

Afin d'évaluer notre approche, l'algorithme AGGC a été appliqué à des jeux de donnés dérivés de celui de DeRisi [DER 97] qui est l'un des plus étudiés dans ce domaine. Ce jeu mesure la variation d'expression des gènes durant le processus cellulaire de « diauxic shift » pour la levure *Saccharomyces Cerevisiae*. Ce processus correspond à la transition de la phase de fermentation du sucre en éthanol (croissance anaérobique) vers la phase de respiration aérobique de la levure .
Ces données indiquent les niveaux d'expression des 6199 ORF's (Opening Reading Frame) de la levure, qui est un organisme entièrement séquencé, pour 7 points temporels durant le processus. Les données ont été prétraitées en prenant le $\log_2$ des ratios (pour considérer les inductions et les répressions cellulaires de façon numériquement égale) et en appliquant l'algorithme d'imputation des K plus proches voisins afin de traiter les valeurs manquantes (1.9% du total).

2.2 Groupes de gènes fonctionnellement riches (GFR)

Nous avons généré une base de données (SGOD) contenant toutes les annotations GO pour chacun des gènes de la levure à partir de GO et SGD. Pour chaque gène sont stockées toutes les annotations du gène et de ses parents. L'ensemble des GFR a été construit à partir de requêtes exécutées sur le SGOD : chaque GFR correspond a un couple constitué d'une annotation GO (*go-term*) et de la liste des gènes annotés par celle-ci.

2.3 Mesure des profils d'expression des gènes

Afin d'incorporer les profils d'expression des gènes, nous nous sommes servi d'une mesure de variabilité d'expression, le *F-score*, qui est plus robuste que d'autres mesures telles que l'*anova*, le *fold change* ou les statistiques *t-student* [RIV 05]. Cette mesure nous permet d'établir une liste des gènes, *g-rank*, ordonnés par variabilités d'expression décroissantes. Nous avons utilisé le programme SAM [TUS 01] pour calculer le *F-score* associé à chaque gène.

3 Analyse des groupes de gènes co-exprimés (AGGC)

AGGC est basé sur l'idée que tout changement affin (co-expression) d'un sous-ensemble de gènes appartenant à une GFR est physiologiquement important. Nous disons que deux gènes sont co-exprimés s'ils sont proches par rapport à la métrique de variabilité d'expression (*F-score*). L'algorithme AGGC permet de déterminer pour chaque GFR la *pc-value* qui estime sa cohérence (à partir de *g-rank*) et donc de détecter les groupes statistiquement significatifs.

3.1 Algorithme AGGC

AGGC commence par déterminer la liste *g-rank* à partir des niveaux d'expression et les GFR à partir de la SGOD. Pour chaque GFR constitué de n gènes, l'algorithme détermine les $n(n+1)/2$ sous-ensembles de gènes dont nous voulons tester la co-expression. Pour chacun de ces sous-ensembles nous calculons sa *pc–value* à partir du test suivant décrit ci-dessous.

H_0 : probabilité que les x gènes d'un de ces sous-ensembles aient été associés par hasard.

Cette probabilité correspond à la distribution hyper-géométrique suivante :

$$p(X = x \mid N, \mathrm{R}_{g(x)}, n) = \frac{\binom{R_{g(x)}}{x}\binom{N - R_{g(x)}}{n - x}}{\binom{N}{n}} \quad où \quad p(X = 0 \mid N, \mathrm{R}_{g(x)}, n) = 0$$

N : nombre total de gènes dans le jeu de données.

n : nombre de gènes dans le GFR.

x : position (n° d'ordre) du gène dans le GFR.

$\mathrm{r}_{g(x)}$: rang du gène de position x dans *g-rank.*

$R_{g(x)}$: nombre de rangs qui séparent le gène x de son prédécesseur (dans le GFR) dans le *g-rank.*

$R_{g(x)}$ est obtenu par : $R_{g(x)} = \mathrm{r}_{g(x)} - \mathrm{r}_{g(x-1)} + 1$ où $R_{g(0)} = \mathrm{r}_{g(0)} = 1$.

La *pc–value* correspondant à cette preuve d'hypothèse est [DRA 03] :

$$pc - value(x) = 1 - \sum_{k=1}^{x} p(X = k \mid N, R_{\mathrm{g(k)}}, n)$$

Afin d'accepter ou rejeter l'hypothèse H_0 nous utiliserons comme seuil de significativité : *p–value = Min* $\{N^{-1}, |\Omega|^{-1}\}$ où $|\Omega|$ est la cardinalité de l'ensemble de tous les annotations fonctionnelles. Ainsi pour chaque GFR, si *pc–value(x) < p–value* alors on rejete H_0, i.e. le GFR est statistiquement significatif.

4 Résultats

Afin d'évaluer notre méthode, nous avons comparé les résultats obtenus par DeRisi, par IGA et AGGC. Les résultats obtenus avec AGGC pour les gènes sur-exprimés sont présentés dans le tableau 1. Les groupes identifiés par AGGC et DeRisi sont en gras, les groupes identifiés seulement par AGGC sont en italique, et le seul groupe identifié par AGGC et IGA est souligné. AGGC a permis de retrouver sept des neuf groupes de gènes obtenus manuellement par DeRisi. Les deux groupes annotés « glycogen metabolism » et « glycogen synthase » n'ont pas été identifiés par AGGC car ils s'expriment uniquement dans la phase initiale du processus et que nous n'avons pas intégré les informations sur les voies métaboliques. Toutefois AGGC a identifié huit groupes statistiquement significatifs et cohérents vis a vis du processus étudié.

Groupe GO fonctionnellement riche	*n* gènes	*x* gènes sur-exprimés	*pc-value*
proton-transporting ATP synthase complex	2	2	4.38E-06
invasive growth (sensu Saccharomyces)	5	3	6.13E-06
signal transduction during filamentous growth	2	2	8.77E-06
respiratory chain complex II	4	4	3.75E-05
succinate dehydrogenase activity	4	4	3.75E-05
mitochondrial electron transport	4	4	3.75E-05
aerobic respiration	36	10	3.30E-05
tricarboxylic acid cycle	14	5	5.09E-05
tricarboxylic acid cycle	14	5	6.54E-05
gluconeogenesis	12	2	9.64E-05
response to oxidative stress	10	3	1.55E-06
filamentous growth	8	4	9.06E-05
vacuolar protein catabolism	4	2	2.63E-05
respiratory chain complex IV	8	2	4.05E-04
cytochrome-c oxidase activity	8	2	4.05E-04

Tableau 1 : GFR sur-exprimés obtenus par AGGC avec une *p-value* de $7X10^{-4}$.

Des résultats similaires, accessibles sur la page du projet, ont été obtenus pour les GFR sous-exprimés.

5 Conclusion

L'algorithme AGGC présenté dans cet article permet d'identifier automatiquement les groupes de gènes co-exprimés significatifs et fonctionnellement riches sans avoir de connaissance a priori des résultats. Il est extensible aux annotations biologiques de toutes natures et aux diverses mesures de variabilité proposées dans le domaine.

AGGC analyse tous les sous-ensembles possibles de chaque GFR, accroissant ainsi la sensibilité de la détection des groupes de gènes co-exprimés, même en présence de données très bruitées. A l'extrême il peut produire des résultats statistiques significatifs sans avoir besoin de répliquer les expériences. Il est également robuste contre les mauvaises assignations lors de la création des groupes fonctionnels à partir des sources publiques (annotations erronées) ou bien de processus automatiques (erreurs de nommage, fautes d'orthographe, etc.).

Les résultats expérimentaux ont montré la validité de l'approche et ont permis d'identifier des informations pertinentes sur les processus biologiques étudiés. Afin d'identifier les groupes de gènes s'exprimant seulement dans certaines phases du processus, nous prévoyons ultérieurement d'intégrer les informations concernant les voies métaboliques.

6 Bibliographie

[ATT 01] ATTWOOD T., MILLER C.J, *Which craft is best in bioinformatics?* Compute. Chem., 25, 2001, p. 329-339.

[BRE 04] BREITLING R., AMTMANN A., HERZYK P., *IGA : A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments*, BMC Bioinformatics, 5:34, 2004.

[CHU 02] CHUAQUI R., *Post-analysis follow-up and validation of microarray experiments*. Nature Genetics, 32, 2002, p. 509 – 514.

[DER 97] DERISI J., IYER L., BROWN V., *Exploring the metabolic and genetic control of gene expression on a genomic scale*. Science, n° 278, 1997, p. 680-686.

[DRA 03] DRAGHICI S., KHATRI P., et al. *Global functional profiling of gene expression,* Genomics, 81, 2003, p. 1-7.

[HOS 03] HOSACK D., DENNIS G., et al., *Identifying biological themes within lists of genes with EASE,* Genome Biology, 4, R70, 2003.

[KIM 05] KIM S., VOLSKY D., *PAGE : Parametric Analysis of Gene Set Enrichment*, BMC Bioinformatics, 6:144, 2005.

[MOO 03] MOOTHA V., LINDGREN C., ERIKSSON K., SUBRAMANIAN A., et al., *PGC-l'alpha-reponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes,* Nat Genet., 34(3), 2003, p. 267-273.

[PAS 04] PASQUIER C., GIRARDOT F., JEVARDAT K., CHRISTEN R., *THEA : Ontology-driven analysis of microarray data*. Bioinformatics, vol.20, issue 16, 2004.

[RIV 05] RIVA A., CARPENTIER A., TORRESANI B., HENAUT A., *Comments on selected fundamental aspects of microarray analysis*, Computational Biology and Chem. 29, 2005, p. 319-336.

[ROB 02] ROBINSON M., et al., *FunSpec : a Web based cluster interpreter for yeast*. BMC Bioinformatics, 3, 35, 2002.

[TUS 01] TUSHER V., TIBSHIRANI R., CHU G., *Significance analysis of microarrays applied to the ionizing radiation response*, Proc. Nat. Acad. Sci. USA, 98 (9), 2001, p. 5116-21.