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

Ricardo Martinez*, Nicolas Pasquier*,
Claude Pasquier**, Martine Collard*, Lucero Lopez***

*Projet Execo
Laboratoire I3S, UNSA
2000, route des lucioles,
06903 Sophia-Antipolis cedex, France
{rmartine, pasquier, mcollard}@i3s.unice.fr

**Laboratoire Biologie Virtuelle, UNSA
Centre de Biochimie, Parc Valrose,
06108 Nice cedex 2, France
claude.pasquier@unice.fr

***Projet Odyssee
INRIA Sophia Antipolis,
2004 route des Lucioles
06905 Sophia Antipolis, France
lucero.lopez@gmail.com

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 données à la lumière des différentes sources d'information 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'information 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].

[1] Programme et informations annexes sur AGGC : http://www.i3s.unice.fr/~rmartine/AGGC.

1 Introduction

L'analyse de données de biopuces en utilisant les diverses sources d'information génomiques représente un challenge important. Ces sources d'information, continuellement alimentées par des volumes croissants de données, sont :

- Des taxonomies, thésaurus et ontologies fournissant des sources d'information sémantiques sur les gènes : *Gene Ontology* (GO)[2], *Unified Médical Language System* (UMLS), Taxonomy etc.
- Des bases de données bibliographiques tels que des articles, corpus de documents, des abstraits etc.): *Pubmed-Medline, Biosis,* etc.
- Des bases de données d'expériences contenant des résultats de précédentes analyses : *Arrayexpress*, *Gene Expression Omnibus* (GEO), etc.
- Des bases de données moléculaires contenant des collections de données structurales et/ou fonctionnelles au sujet des chaines de nucléotides (ADN, ARN, gènes etc.) ou bien des protéines : GenBank, Embl, Unigene, etc.
- Des sources d'information relatives aux gènes ou protéines en concernant un domaine spécifique d'étude tels que : KEGG pour les systèmes biologiques, GeneCards qui contiennent des relations entre des gènes et des maladies, etc.
- L'information minimal en concernant l'expérience de biopuces tel que : les identifiants des gènes, les conditions biologiques et leur caractéristiques tels que le type de tissue, le gendre, âge, etc., les caractéristiques du prétraitement des données etc. Plus de détailles en Brazma et al. (2001).

Plusieurs approches statistiques et d'analyse de données permettant d'identifier des groupes de gènes co-exprimés en comparant les niveaux d'expression des gènes, i.e. sans prendre en compte la connaissance a priori, ont été proposées : DeRisi et al. (1997), Cho et al. (1998), Eisen et al. (1998), Tamayo et al. (1999), etc. Une caractéristique commune à ces approches purement numériques est qu'elles identifient des groupes (appelés *clusters*) de gènes d'intérêt mais laissent à l'expert la tache de découvrir et interpréter les similarités biologiques cachées dans ces groupes. Si ces méthodes sont bénéfiques, car elles guident le processus d'analyse des groupes fonctionnellement riches, les résultats obtenus sont parfois incomplets car ces approches n'identifient pas des groupes de gènes qui s'expriment dans quelques conditions biologiques seulement et non dans tous les conditions biologiques (Liu et al., 2004).

L'un des défis majeurs actuels dans ce domaine est l'intégration automatique des connaissances biologiques issues des diverses sources d'information avec les données d'expression de gènes (Attwood et Miller, 2001). Un premier bilan des méthodes développées pour répondre à ce défi se trouve dans (Chuaqui, 2002).

Nous ciblons ici l'enrichissement de deux axes de recherche récemment développés, *standard* ou *basé sur l'expression* et *a priori* ou *basé sur la connaissance*, qui exploitent de multiples sources d'annotations telles que *Gene Ontology*. Ces annotations sont des informations fonctionnelles, relationnelles et syntaxiques sur les gènes.

Dans l'axe *standard* (*basé sur l'expression)*, 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

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

de ces sous-ensembles de gènes co-annotés est testée. Parmi les méthodes dans cet axe citons *Quality tool* (Gibbons et al., 2002), *Onto express* (Draghici et al., 2003), *EASE* (Hosack et al., 2003), *THEA* (Pasquier et al., 2004) et *Graph modeling* (Lee et al., 2004).

Dans l'axe *a priori* (*basé sur la connaissance)*, 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 GFR est ensuite testée en utilisant un test basé sur un score enrichi (Mootha et al., 2003), un test issu d'un *z-score* (Kim et al., 2005) ou un test basé sur une *p-value* calculé a partir de la distribution hypergéométrique (Breitling et al., 2004).

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. 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*. Finalement, nous obtenons des GFR co-exprimés et statistiquement significatifs.

La méthode AGGC est une extension de la méthode IGA (Breitling et al. 2004). En effet AGGC permet d'obtenir tous les sous-ensembles possibles de GFR de gènes co-exprimés, contrairement a la méthode IGA qui se limite à trouver des GFR des gènes les plus ou les moins exprimés. IGA élimine tous les GFR co-exprimés qui ne se trouvent pas dans les premiers (ou les derniers) rangs de la liste de gènes ordonnée sur les niveaux d'expression par ordre décroissant ou croissant. Par contre AGGC permet de sélectionner également des GFR classés au milieu de cette liste ordonnée. Pour cela AGGC prend en compte le niveau relatif (rang relatif) de classement dans cette liste. La recherche de tous les possibles sous-ensembles de gènes co-exprimés et co-annotés dans une expérience biologique quelconque, sans se limiter à quelques groupes seulement, augmentent les chances de compréhension du phénomène biologique par l'expert.

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 et al. (1997) 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.

La technique utilisée est celle des biopuces par double coloration par les fluorochromes à spectres d'émission distincts Cy3 et Cy5. 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 à 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, la *statistique t modifiée*, qui est plus fiable que d'autres mesures telles que le simple *fold change* (Riva et al., 2005). Cette mesure nous permet d'établir une liste des gènes, *g-rank*, ordonnés par variabilités d'expression décroissantes. Nous avons utilisé le programme SAM (Tusher et al., 2001) pour calculer la *statistique t modifiée* associé à chaque gène.

La *statistique t modifiée* utilisé par SAM est une statistique t avec un terme correcteur dans le dénominateur. Une fois que SAM a calculé la *statistique t modifiée* pour chaque gène, les gènes *potentiels* sont choisis comme étant ceux qui possèdent un score supérieur à un seuil fixé à l'avance. Afin de contrôler les faux positifs, SAM utilise les permutations de mesures d'expression de gènes pour estimer le taux de faux positifs (TFP). Le score correspondant au seuil de sélection de gènes est alors ajusté itérativement d'après les valeurs des TFP jusqu'à ce qu'un ensemble de gènes significatifs soit identifié. Plus de détail sur la méthode SAM se trouve dans Tusher et al., 2001.

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

AGGC est basé sur l'idée que tout changement affine (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 (*statistique t modifiée*). L'algorithme AGGC permet de déterminer pour chaque GFR la *p-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 évalue les 2^n sous-ensembles de gènes dont nous voulons tester la co-expression. Pour chacun de ces sous-ensembles nous calculons son probabilité de changement ou *p–value* à partir du test suivant

décrit ci-dessous. La probabilité de changement d'un sous-ensemble d'une GFR équivaut au couramment nommé *p-value* dans le langage des preuves d'hypothèses.

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.

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

avec :

- *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,
- $r_{g(x)}$: rang absolu du gène de position *x* dans la liste *g-rank*,
- $R_{g(x)}$: nombre de rangs dans *g-rank* qui séparent le gène de position *x* de son prédécesseur dans le GFR.

L'intervalle entre rang, $R_{g(x)}$, est calculé a partir des rangs absolus $r_{g(x)}$ selon la formule :

$$R_{g(x)} = r_{g(x)} - r_{g(x-1)} + 1 \quad où \quad R_{g(0)} = r_{g(0)} = 1$$

Le *p–value(x)* correspondant à cette preuve d'hypothèse est (Draghici et al., 2003) :

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

Afin d'accepter ou rejeter l'hypothèse H_0 nous proposons le seuil de significativité : $\alpha = 1/|\Omega|$, où $|\Omega|$ est le nombre de GFR obtenus à partir de toutes les annotations fonctionnelles sur les *N* gènes de l'expérience biologique.
Ainsi pour chaque sous-ensemble d'une GFR de gènes, nous testons l'inégalité :

$$p - value\ (x) < \alpha,$$

pour rejeter H_0, i.e. l'hypothèse que le GFR est statistiquement significatif.
Une alternative pour fixer ce seuil serait de faire déterminer une valeur significative par l'expert du processus biologique étudié. Ceux-ci en évitant l'élection d'un seuil correspondant a un choix par hasard des groupes significatifs.

Le pseudo-code de l'algorithme AGGC est présenté dans la FIG. 1. Cet algorithme a été implanté en langage Perl. Il reçoit en entrée les listes d'annotations de chaque gène (générés par une requête sur la base de données SGOD contenant toutes les annotations GO) et la liste ordonné *g-rank* des *N* gènes. Il renvoie la liste des groupes de gènes co-exprimés significatifs.

Entrée : Liste *annotations*(*G*) des annotations pour chaque gène *G*.
Liste ordonnée *g-rank* des *N* gènes.
Sortie : Ensemble *résultat* des GFR de gènes co-exprimés.

- **début**
- **déterminer** $\alpha = 1/|\Omega|$
- **pour chaque** annotation *A* de GO **faire**
- **pour chaque** gene *G* **faire**
- **si** $A \in$ *annotations*(*G*) **alors**
- $GFR_A \leftarrow GFR_A \cup G$
- **finsi**
- **finpour**
- **finpour**
- **pour chaque** GFR_A **faire**
- **pour chaque** sous-ensemble *S* de GFR_A **faire**
- **calculer** *p-value*(*S*)
- **si** *p-value*(*S*) < α **alors**
- *résultat*(GFR_A) ← *résultat*(GFR_A) ∪ *S*
- **finsi**
- **finpour**
- **supprimer** de *résultat*(GFR_A) les *S* non maximaux vis-à-vis de l'inclusion
- **finpour**
- *résultat* ← $\cup_{i=A}$ *résultat*(GFR_i)
- **fin**

FIG.1 – *Algorithme AGGC*

L'algorithme commence par déterminer la valeur de $\alpha = 1/|\Omega|$ (étape 2) et générer les GFR à partir des annotations GO (étapes 3 à 9). Il considère ensuite successivement chaque GFR (étapes 10 à 18). Pour chaque GFR, il détermine tous ses sous-ensembles non vides et calcule la *p-value* de chacun (étapes 11 à 16). Si la *p-value* calculée est inférieure à α, le sous-ensemble est inséré dans le résultat du GFR (étapes 13 à 15). Les sous-ensembles insérés dans le résultat du GFR qui ne sont pas maximaux vis-à-vis de l'inclusion sont ensuite supprimés (étape 17). Supposons par exemple que pour $GFR_A = \{g_1, g_2, g_3\}$ nous ayons *résultat*(GFR_A) = $\{\{g_1\}, \{g_2\}, \{g_3\}, \{g_1, g_2\}, \{g_2, g_3\}, \{g_1, g_2, g_3\}\}$. Alors, tous les

ensembles de $\{g_1, g_2, g_3\}$ sont supprimés de *résultat*(GFR_A). Finalement, le résultat global est constitué de tous les groupes de gènes co-exprimés et significatifs (étape 19).

3.2 Exemple

L'exemple de l'analyse d'un groupe de gènes co-annotés par l'algorithme AGGC est présenté dans TAB. 1. Les données utilisées sont celles de l'expérience d'analyse du processus de la « diauxic shift » pour la levure *Saccharomyces Cerevisiae* réalisée par DeRisi et al. en 1997 (voir section 2.1). La liste ordonnée *g-rank* a été calculée à partir de la *statistique t modifiée* obtenu par le programme SAM (voir section 2.3). Les données sur le GFR étudié, annoté *vacuolar protein catabolism*, ont été obtenues par une requête sur la base de données SGOD (voir section 2.2). Ce GFR contient 4 gènes (n = 4) dont les rangs dans la liste globale g-rank varient de 6 à 424.

Dans le TAB. 1 sont données les valeurs des paramètres nécessaires pour déterminer les sous-ensembles significatifs de gènes au sein du GFR. En gras est indiqué le seul sous-ensemble de gènes de ce GFR que AGGC a trouvé significativement co-exprimé.

Liste *g-rank*	*x*	Gène ID (SGD)	Annotation GO	$r_{g(x)}$	$R_{g(x)}$
1				1	
2				2	
…				…	
6	**1**	**S000000490**	**vacuolar protein catabolism**	**6**	**1**
7		-	-	7	
8	**2**	**S000001586**	**vacuolar protein catabolism**	**8**	**3**
…				…	
69	3	S000000786	vacuolar protein catabolism	69	62
…				…	
424	4	S000006075	vacuolar protein catabolism	424	356
…				…	
N				*N*	

TAB. 1 – *Analyse du GFR des gènes annotés* vacuolar protein catabolism *par AGGC.*

AGGC a testé l'hypothèse H_0 pour les (4*5)/2=10 sous-ensembles possibles de ce GFR en déterminant la *p-value* de chacun. Par exemple, la *p-value* du sous-ensemble {S000000490, S000001586} de rang 6 et 8 dans *g-rank* est de 2.63E-05 (cf. TAB. 2). Cette *p-value* est inférieure à α fixé à 6.88E-04, i.e. $\alpha = 1/1453 = 6.88\text{E-}04$, où le nombre d'annotations fonctionnelles pour les 6199 gènes est de 1453 (cf. section 3.1). Alors, AGGC rejette l'hypothèse H_0 et le groupe de gènes {S000000490, S000001586} est statistiquement significatif et co-exprimé et co-annoté. Intuitivement nous apercevons que le sous-ensemble avec les gènes de rang 6 et 8 est très proche et pourtant co-exprimé. Par contre les gènes de rang 69 et 424 sont assez éloignés de leurs plus proches voisins, c'est-à-dire les groupes auxquels ils participent ne sont pas co-exprimés de manière significative.

4 Résultats

Afin d'évaluer notre méthode, nous avons comparé les résultats obtenus par DeRisi et al. (1997), par IGA (Breitling et al. 2004) et AGGC. Les résultats obtenus avec AGGC pour les gènes sur-exprimés et sous-exprimés sont présentés dans le TAB. 2 et TAB. 3 respectivement. 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.

Groupe GO fonctionnellement riche	*n* gènes	*x* gènes sur-exprimés	*p-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

TAB. 2 – GFR sur-exprimés obtenus par AGGC avec un α = 6.88E-04.

Dans le cas de gènes sur-exprimés (TAB. 2), 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. Toutefois AGGC a identifié huit autres groupes statistiquement significatifs et cohérents vis-à-vis du processus étudié. Un seul de ces huit autres groupes avait été identifié par IGA et aucun lors de l'étude de DeRisi et al.

Pour le cas de gènes sous-exprimés, AGGC a retrouvé sept des huit groupes de gènes obtenus manuellement par DeRisi et al. Comme pour les gènes sur-exprimés, un groupe, annoté « ribosome biogenesis », n'a pas été identifié par AGGC car s'exprimant seulement durant la phase finale du processus. AGGC a également identifié sept autres groupes statistiquement significatifs et cohérents vis-à-vis du processus étudié qui n'ont pas été identifiés lors de l'analyse de DeRisi et al. ni par IGA.

Groupe GO Fonctionnellement Riche	*n* gènes	x gènes sous-exprimés	*p-value*
chromatin modification	6	5	2.35E-06
mitochondrial inner memb. prot. inser. complex	3	2	3.60E-06
regulation of nitrogen utilization	4	2	7.20E-06
acid phosphatase activity	4	2	7.20E-06
histone acetylation	4	4	7.95E-06
nucleolus	52	10	3.41E-05
rRNA modification	10	3	2.75E-05
transcription initiation from RNA poly. II prom.	14	3	1.00E-05
mitochondrial matrix	15	3	1.25E-05
processing of 20S pre-rRNA	11	2	1.97E-04
ribosomal large subunit biogenesis	9	4	3.17E-04
small nucleolar ribonucleoprotein complex	20	3	2.52E-04
cytosolic large ribosomal subunit	69	13	2.87E-04
ribosomal large subunit assembly and maint.	21	2	2.52E-04

TAB. 3 – GFR sous-exprimés obtenus par AGGC avec une $\alpha = 6.88E\text{-}04$.

Les trois groupes identifiés par DeRisi et al. que AGGC n'a pas identifié, à savoir les groupes sur-exprimés « glycogen metabolism » et « glycogen synthase », et le groupe sous-exprimé « ribosome biogenesis » partagent deux propriétés importantes. Premièrement, ils contiennent des gènes appartenant à une structure hétérogène, i.e. des gènes appartenant à plusieurs groupes fonctionnels. Deuxièmement ces GFR s'expriment uniquement durant une phase spécifique du processus étudié et non pas tout au long de celui-ci. La détection de ces groupes ne sera donc possible qu'en intégrant les informations sur les voies métaboliques (KEGG, EMP, CFG, etc.).

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.

AGGC analyse tous les sous-ensembles possibles de chaque GFR, accroissant ainsi la sensibilité de la détection des groupes de gènes co-exprimés. 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 annotations fonctionnelles fournies par AGGC constituent un outil efficace et rapide permettant de réduire la complexité du problème de l'analyse de données de biopuces en intégrant des informations de diverses natures sur les gènes. Il peut être utilisé pour valider et comparer les résultats d'expériences de biopuces avec ceux stockés dans les bases de

données et les bases documentaires publiques en identifiant les groupes de gènes d'intérêt pour l'expérience en question.

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

Références

Attwood T. et Miller C.J. (2001). *Which craft is best in bioinformatics?* Compute. Chem., 25:329-339.

Breitling R., Amtmann A. et Herzyk P. (2004). *IGA : A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments*. BMC Bioinformatics, 5:34.

Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., et al. (2001) *Minimum Information about a microarray experiment MIAME - toward standards for microarray data.* Nature Genetics, 29: (365--371).

Cho R., Campbell M., Winzeler E., et al. (1998). *A genome-wide transcriptional analysis of the mitotic cell cycle*. Mol. Cell., 2:65-73.

Chuaqui R. (2002). *Post-analysis follow-up and validation of microarray experiments.* Nature Genetics, 32:509-514.

DeRisi J., Iyer L. et Brown V. (1997). *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 278:680-686.

Draghici S., Khatri P., et al. (2003). *Global functional profiling of gene expression*. Genomics, 81:1-7.

Eisen M., Spellman P., Brown P., Botsein D., et al. (1998). *Cluster analysis and display of genome wide expression patterns*. Proc. Nat. Acad. Sci., 95 (25):14863-8.

Gibbons D., Roth F., et al. (2002). *Judging the quality of gene expression-Based Clustering Methods Using Gene Annotation*. Genome Research, 12:1574-1581.

Hosack D., Dennis G., et al. (2003). *Identifying biological themes within lists of genes with EASE*. Genome Biology, 4:R70.

Kim S., Volsky D. et al. (2005). *PAGE : Parametric Analysis of Gene Set Enrichment*. BMC Bioinformatics, 6:144.

Lee S., Hur J., et Kim S. (2004). *A graph theoretic modeling on GO space for biological interpretation of gene clusters*. Bioinformatics, 3: 381-386.

Liu J., Yang J., et Wang, W. (2004). *Biclustering in gene expression data by tendency. In* Proceedings of Computational Systems Bioinformatics Conference (CSB), 182-193.

Masys D., et al. (2001*). Use of keyword hierarchies to interpret gene expressions patterns*. Bioinformatics, 17:319-326.

Mootha V., Lindgren C., Eriksson K., Subramanian A. et al. (2003). *PGC-l'alpha-reponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes*. Nat Genet. 34(3): 267-273.

Pasquier C., Girardot F., Jevardat K. et Christen R. (2004). THEA : *Ontology-driven analysis of microarray data*. Bioinformatics, vol.20, issue 16.

Quackenbush J. (2002). *Microarray data normalization and transformation*. Nat. Genet., 32 (suppl.):496-501.

Riva A., Carpentier A., Torresani B. et Henaut A. (2005) *Comments on selected fundamental aspects of microarray analysis*. Computational Biology and Chemistry 29:319-336.

Robinson M., et al. (2002). *FunSpec : a Web based cluster interpreter for yeast.* BMC Bioinformatics, 3:35.

Storey J. et Tibshirani R. (2003). *Statistical significance for genomewide studies*. Proc. Natl. Acad. Sci., 100 (16): 9440-5.

Tusher V., Tibshirani R., Chu G., et al. (2001). *Significance analysis of microarrays applied to the ionizing radiation response*. Proc. Nat. Acad. Sci. USA, 98 (9):5116-21.

Tamayo P., Slonim D., et al. (1999). *Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation.* Proc. Natl. Acad. Sci., 96:2907-2912.

Summary

Microarray technology produces vast amounts of data by measuring simultaneously the expression levels of thousands of genes under hundreds of biological conditions. Nowadays, one of the principal challenges in bioinformatics is the interpretation of this large amount of data using different sources of information.

We have developed a new data analysis method named CGGA (Coexpressed Gene Group Analysis) which finds automatically groups of genes, that are functionally enriched, i.e. have the same functional annotations, and that are coexpressed.

CGGA automatically integrates the information of microarrays, i.e. gene expression profiles, with the functional annotations of the genes obtained by the genome-wide information sources as Gene Ontology.

By applying CGGA to several microarray experiments, we have discovered the principal functionally enriched and coexpressed gene groups, and we have shown that this approach enhances and accelerates the interpretation of DNA microarray experiments.