Biological networks to the analysis of microarray data*

FANG Zhuo[1], LUO Qingming[1]**, ZHANG Guoqing[1] and LI Yixue[2]**

(1. Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan 430074, China; 2. Shanghai Center for Bioinformation and Technology, Shanghai 200235, China)

Received April 10, 2006; revised July 3, 2006

Abstract Microarray technology, which permits rapid and large-scale screening for patterns of gene expressions, usually generates a large amount of data. How to mine the biological meanings under these data is one of the main challenges in bioinformatics. Compared to the pure mathematical techniques, those methods incorporated with some prior biological knowledge generally bring better interpretations. Recently, a new analysis, in which the knowledge of biological networks such as metabolic network and protein interaction network is introduced, is widely applied to microarray data analysis. The microarray data analysis based on biological networks contains two main research aspects: identification of active components in biological networks and assessment of gene sets significance. In this paper, we briefly review the progress of these two categories of analyses, especially some representative methods.

Keywords: biological networks, microarray, data analysis, subnetwork, gene set.

The genomic expression, which has a profound impact on identifying regulation relationship, gene function prediction, investigation of pathogenic mechanisms, drug discovery and so on, is one of the research focuses in the post-genomic era. High-throughput methodologies such as oligonucleotide and cDNA microarrays, which can monitor global expression changes of thousands of genes, are thus of great significance to the research of modern life science, and have attracted extensive studies[1]. Due to the valuable yet complicated information involved in the expression data, how to manage, integrate and interpret these data correctly is becoming the main challenge.

Originally, pure mathematical approaches are introduced to help the microarray data analysis. In general, it includes two aspects: (1) identification of differentially expressed genes according to sample classes; (2) classifying samples or genes according to similar expression patterns. Several statistic methods are adopted to determine differentially expressed genes, including *t*-test, non-parametric test, Bayesian model and so on[2]. These approaches produce a list of significant genes, which is however difficult to interpret without any combined biological theme. As we know, cellular processes are often carried out through interactions among many genes, thus the analysis of single gene may miss some important information. Accordingly, clustering or classification processes[3—6] are introduced to describe the global change of the expressions of many genes. Similar to previous differential expressed gene identification, this global change cannot reflect the cellular response directly either. Moreover, these two processes both depend very much on the particular algorithm designs. Different designs always lead to different results with little overlap. So far, it is difficult to understand the biological meaning of microarray data by pure mathematical approaches only.

Many subsequent studies illustrate that incorporating prior biological knowledge into microarray data analysis can effectively avoid the above problems of pure mathematical methods[7—9]. Biological knowledge includes all aspects covering sequence alignments, protein structures and biological functions, which can be either generic or species-specific. Recently, one kind of prior knowledge, biological network, which characterizes the annotation relationships among genes as network structure, is becoming very popular. The microarray data analysis based on biological network integrates expression profiles with network-wide annotations, such as metabolic networks and protein interaction networks. According to the well-accepted assumption that the co-expression among genes attributes to common biological func-

* Supported by the National Program on Key Basic Research Projects (No. 2004CB518606), the Fundamental Research Program of Shanghai Municipal Commission of Science and Technology (No. 04DZ14003), and the National Key Technologies R&D Program of China (No. 2005BA711A04)

** To whom correspondence should be addressed. E-mails: qluo@mail.hust.edu.cn, yxli@sibs.ac.cn

tions, the microarray data analysis based on biological network may identify the genes which perform a certain function through their expression profiles. Furthermore, it can be used to evaluate the significance of a certain biological function by checking the expression state of all related genes.

1 Biological networks

As we mentioned in the previous section, biological network represents the annotation relationships among genes or gene products, such as proteins. Generally, there are three main sources of biological networks: metabolic network, molecular interaction networks and Gene Ontology. The knowledge of these biological networks can be obtained from public databases. Metabolic networks as well as chemical reactions can be found in KEGG database. The Kyoto Encyclopaedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/)[10] provides a reference knowledge base for linking genomes to biological systems and wiring diagrams of interaction networks and reaction networks, which can be used for modeling and simulation as well as for browsing and retrieval. The Biomolecular Interaction Network Database (BIND) (http://bind.ca)[11] and the Database of Interacting Proteins (DIP) (http://dip.doembi.ucla.edu)[12] provide the molecular interactions information. The BIND database stores interactions and reactions arising from biopolymers (protein, RNA and DNA), as well as small molecules, lipids and carbohydrates. The DIP database aims to supply binary protein-protein interactions. Gene Ontology (GO) (http://www.geneontology.org)[13] is a cross-species, controlled vocabulary describing three domains of molecular biology: molecular function, cellular component and biological process. It is specifically intended for annotating gene products and independent of any biological species. Besides the databases of these three biological networks, GenMapp (http://www.genemapp.org)[14] allows visualization of gene expression data on alliance for cellular signaling, BioCarta, EcoCyc, MetaCyc, KEGG and PathDB pathways, associated with analysis tools such as MAPPFinder[15]. The TRANSFAC (http://www.gene-regulation.de/)[16] database provides transcription factors and their DNA-binding sites and profiles. Pfam (http://www.sanger.ac.uk/Software/Pfam/)[17] is a database of protein families represented by multiple sequence alignments and hidden Markov models.

2 Assessment of gene sets significance

A preliminary stage of analyzing microarray data on biological networks is the analysis focusing on gene sets level. Gene sets denote the genes with correlations in biological networks, for example, participating in the same metabolic pathway, sharing the same biological function, the common chromosomal location or regulation. The main goal of gene sets analysis is to determine whether a large number of genes from the gene sets are significantly regulated. These analyses, by assessing the global differentially expressed levels of the genes, can catch the fact that cellular processes often affect sets of genes acting in concert, thus avoiding the shortage of single-gene analyses. Moreover, the analyses at the gene set level can detect consistent but subtle changes in gene expression, which is also worth adverting to.

Curtis et al.[18] have reviewed some methods of gene set analysis, including hypergeometric probability (binomial distribution)[19], fisher exact test[20] and χ^2 test[21], Z scores[15] and odds ratio[22], gene set enrichment analysis[23] and so on, which are not discussed in this paper any more. In the following sections, we will summarize some recent progress made on some methods in gene sets significance assessment and their additional information is listed in Table 1.

2.1 Pathway scores

Pathway is an intricate network consisting of the chemical reactions and interacting molecules that perform specified biological functions. It is the key to understanding how an organism reacts to perturbations from its environment or internal changes. Zien et al. have presented an approach to evaluate pathways by scoring the gene expressions from conspicuousness, synchrony and combined effect[24]. In the conspicuousness score, a normal distribution is used to model the expression change. The conspicuousness score of a gene is valued by the expression levels under all conditions, while the conspicuousness score of a pathway is the average over the score of the genes included in the pathway. Pearson's correlation coefficient is introduced as expression similarity in synchrony score. The synchrony score of a gene is calculated as the average correlation coefficient to other genes in the pathway it belongs to, while the synchrony score of a pathway is the average of the gene scores. The combined scoring function is a modified

form of synchrony score by replacing the standard deviation of particular genes with the holistic standard deviation, which can scale the covariance between genes in a union.

Kurhekar et al. have made some improvements on Zien et al. s' pathway scoring[25]. Different from the conspicuousness score of Zien et al., the activity score of Kurhekar et al. counts in the active gene numbers in a pathway. The score will be higher if there are more genes over-expressed or under-expressed in the pathway. The coregulation score measures the slopes among genes in a pathway, avoiding the problem that pairwise correlations cannot capture the simultaneous co-expression of all genes in a pathway. To incorporate the structure of a pathway into the analysis, a cascade score is introduced to measure the interaction levels among active genes in a pathway. Each path in a pathway is scored by the active gene numbers occurred in it, while the highest score is assigned as the cascade score of this pathway.

The analysis of pathway scores completely considers pathways in three aspects. Nevertheless, previous approaches do not synchronize these three aspects well. The authors always computed the three scores separately, where different pathways might gain different significant scores according to the three aspects. For example, pathways A and B are the most significant pathways by conspicuousness score; pathways C and D are the most significant pathways by synchrony score; pathways E and F are the most significant pathways by combined score. Hence the interpretation of results is challengeable. Because of such inconsistency, a uniform criterion is definitely needed to involve these three aspects.

2.2 Function module analysis

It is well known that in most cases only a subset of the whole gene set may contribute to its expression signature, and different sets may have similar signatures in the same experiment. To extract the core parts of each set, function module analysis is introduced. Segal et al. firstly adopted this method to identify conditional active expression modules in different types of cancer[26]. There, the significance of a function module (a set of genes) is determined by the fraction of active genes in it. The significance of a module for specified cancer type is determined by the fraction of experiments in which the module is active.

The function module analysis combines function modules (gene sets) with cancer types (experiment sets) through the concepts of gene enrichment and gene set enrichment. As the reference pointed out, the function module analysis method can characterize the modules shared across multiple tumor types, which may be related to general tumorigenic processes, and modules specific to particular tumors. Also, each cancer type can be described as a particular combination of modules.

2.3 Gene set enrichment analysis

Another important aspect of gene set analysis is the gene set enrichment analysis (GSEA), which was proposed by Mootha et al. in 2003[23]. GSEA determines whether prior defined gene sets are enriched at the top of a list of genes ordered by the expression difference (signal to noise ratio, SNR) between two classes. A normalized Kolmogorov-Smirnov statistic is used to define the enrichment score. For a gene set S containing G members and a gene list $R_1, \cdots, R_N$ ordered by the differential expression levels, the score is $X_i = -\sqrt{\frac{G}{N-G}}$ if R_i does not belong to S, and $X_i = \sqrt{\frac{N-G}{G}}$ if R_i belongs to S. The enrichment score of a pathway is the running sum of X_i for N genes with the maximum absolute value.

In subsequent studies, some statistical problems about GSEA procedure are concerned by Damian et al.[27]. The most interesting one is that the enrichment score will be influenced by the size of a gene set. Another limitation is that the GSEA process cannot treat the top genes in the gene list with the same significance as the bottom genes of the list, where the top genes and bottom genes represent up-regulated and down-regulated respectively and should be equally significant for analysis. To avoid these problems, Sunramanian et al. introduced an enhanced GSEA, in which weighted score is applied to each step instead of equal score[28]. Here $X_i = \frac{r_i}{N_R}$ is the score for genes included in set S, while $X_i = \frac{1}{N-G}$ is the score for genes excluded in set S, where r_i is the rank for gene i and N_R is the rank sum of all genes in S. The enrichment score is still the maximum running sum deviating from zero. The improved version of GSEA can identify gene sets that have enriched subsets both at the top and bottom of the gene lists, which will be ignored by the over penalization of bottom genes in the

original GSEA algorithm.

Kim et al. developed a parametric gene set enrichment analysis (PAGE) by normal distribution[29]. Their theoretical base is the Central Limit Theorem: the distribution of the average of randomly sampled n observations tends to follow normal distribution as the sampling size n becomes larger, no matter whether the parent distribution is normal or not. As a result, the statistical significance of gene sets can be assessed by normal distribution if the gene numbers in the sets are sufficiently large (10 at least). Compared with GSEA, PAGE can obtain more significant gene sets whose p-value is lower than that of GSEA. It might be caused by that GSEA uses permutation of original data set (1000 times) to get background distribution of each enrichment score and evaluates the significance of each enrichment score from the permutated data set, thus the best p-value in GSEA cannot be smaller than 0.001 (which is 1 over 1000).

Table 1. Summary of some methods for gene sets significance assessment

Methods	Annotations	Databases	Data sets
Statistical tests[7, 15, 20, 21]	Gene Ontology Metabolic network GenMapp Biocarta Pfam MIPS SMART	GO[13] KEGG[10] GenMapp[14] Biocarta Pfam[17] MIPS[30] SMART[31]	Arabidopsis Yeast Mouse Human GEPAS
Pathway scores[24]	Glycolysis and gluconeogenesis pathway in *Saccharomyces* *Cerevisiae* described in [32] Metabolic pathways	KEGG[10]	6178 *Saccharomyces cerevisiae* ORFs during 18 time points[32—34] 6178 *Saccharomyces cerevisiae* ORFs during 4 time courses[32—34]
Function module analysis[26]	Gene Ontology Metabolic pathway GenMapp Tissue-specific expressed gene set[35] P-clusters[36]	GO[13] KEGG[10] GenMapp[14]	14145 genes in 1975 arrays spanning 17 cancer categories[26]
Gene set enrichment analysis[23, 28, 29]	Gene Ontology GenMapp Biocarta Signaling pathway Signaling gateway Transduction information Protein reference Sigma-Aldrich pathways Human cancer genome anatomy information Gene arrays Regulatory-Motifs in the promoter regions[37] Neighborhoods around cancer associated genes Novartis normal tissue compendium[37] Novartis carcinoma compendium[38] Global cancer map[39]	GO[13] GenMapp[14] Biocarta SPAD[40] AfCS-Nature Signaling Gateway[41] STKE[42] Human protein reference database[43] Sigma-Aldrich[44] SupperArray[45] CAGECancer[46]	22000 genes in skeletal muscle biopsy samples from 43 males[23] 12625 probes on HGU95Av2 chip for 50 NCI60 cell lines, 17 normal and 33 p53 mutations[47] 12625 probes on HGU95Av2 chip for 24 acute lymphoid leukemia and 24 acute myeloid leukemia[48] 12625 probes on HGU95Av2 chip for 62 lung adenocarcinomas[49] 7129 probes on HU6800 chip for 86 lung adenocarcinoma[50]

3 Identification of active components in biological networks

Recently, another direction of microarray data analysis based on biological networks, namely identifying active components in biological networks, becomes very popular. In these analyses, gene expression information is transformed and mapped to biological networks. Based on the assumption that genes with shared function will be activated together and thus show correlated expression profiles, the connected regions (subnetworks) inside the network, which show significant changes over particular conditions, can be selected by applying certain optimization algo-

rithms. In this sense, the main effort in this topic focuses on the optimization algorithms as the following sections illustrate. Identification of active components in biological networks is supposed to help understanding of the underlying mechanisms governing the observed changes in gene expression. Some methods in identification of active components in biological networks and their additional information are listed in Table 2.

Table 2. Summary of some methods for identification of active components in biological networks

Methods	Annotations	Databases	Data sets
Simulated annealing[51]	Protein-protein interactions Protein-DNA interactions	BIND[11] TRANSFAC[16]	997 mRNAs responding to 20 systematic perturbations of the yeast galactose-utilization pathway[51]
Expectation maximization[53]	Protein-protein interactions	DIP[12]	3589 *Saccharomyces cerevisiae* genes with 173 micrarrays[65] 3589 *Saccharomyces cerevisiae* genes with 77 micrarrays[32]
Kernel function analysis[56]	Pathways	KEGG[10]	6178 *Saccharomyces cerevisiae* ORFs during 18 time points[32—34]
Graph-iterative group analysis[63]	Gene Ontology Metabolic network	GO[13] Swiss-Prot[66]	6178 *Saccharomyces cerevisiae* ORFs during 18 time points[32—34]
Wavelet transform[64]	Metabolic network	EcoCyc[67]	4345 *E coli*. ORFs of 43 samples[68]

3.1 Simulated annealing

Algorithm based on simulated annealing is introduced by Ideker et al.[51]. In their algorithm, the genes are firstly scored by their differential expression levels. The score of a subnetwork is defined as adjusted average of scores of genes inside. After Monte Carlo random approach, the score of each pathway is normalized. High score indicates active biological subnetwork. Because the problem of finding maximal subnetwork is NP-hard, the simulated annealing approach is used to look for the high score subnetworks. The initialization of simulated annealing randomly sets each gene node in the network as active/inactive state, where the active genes form an initialized subnetwork. In each step, the score of the current subnetwork is calculated by randomly toggling the state of a gene in the network. After sufficient iterations, the subnetwork with the highest score is exported as one of the results.

Patil et al.[52] have adopted this approach to metabolic network. The complete metabolic network is represented as a bipartite graph. In this graph, metabolites and enzymes are represented as nodes, while interactions between them are represented as edges. Another unipartite graph is constituted by enzymes, and any two enzymes sharing a common substrate in the corresponding reactions are connected to each other. Scoring and sorting genes with their expression data generate reporter metabolites originated from metabolic graph. Simulated annealing similar with Ideker's process identifies the significantly correlated subnetworks.

The simulated annealing method has several drawbacks. Firstly, this method cannot guarantee to find the optimal subnetwork. Theoretically, if the number of iterations is large enough, the final solution will be the global optimum. Thus the number of iterations is always set very large (for example, 100000 by Ideker et al.) to assure the quality of subnetworks, which will lead to great computational demand. Secondly, the simulated annealing requires relatively complex parameter estimation. It is difficult to get appropriate parameters. Currently, the parameters are mainly set empirically.

3.2 Expectation maximization

Segal et al. have proposed another approach to detect gene groups whose expression profiles are correlated and protein products interacted[53]. The partitions are modeled with relational Markov networks, which contain two components: one for the expression data and the other for the protein interaction data. Gene expression profiles are modeled using Naïve Bayes models. In this model, genes are clustered into disjoint classes. The conditional probability for the attribution of one experiment to a certain gene belonging to a certain cluster is assumed to follow Gaussian distribution. The probabilistic model for protein interaction data is based on the assumption that interacting proteins are likely to be in the same pathway. Thus binary Markov random fields are used to model the protein interaction framework. Finally, a unified model integrating gene expression model and protein interaction model can then be naturally defined as their product.

Some parameters, such as the ones in the model

probabilistic distributions, need to be estimated in the unified model. The parameters are estimated by Expectation Maximization (EM) algorithm. The EM procedure iterates between Expectation (E) and Maximization (M) steps. In the E step, the unified probability is calculated with the current parameters. In the M step, parameters are estimated so as to maximize the probability obtained in the E step. When the EM algorithm converges, each gene is assigned to the groups with the maximum conditional probability.

Despite the success of this method in searching for functional gene groups, some limitations still remain. The main one is that the method is based on probabilistic models, thus relies heavily on the assumption that the data set fits a particular distribution. This may not be true in many practical cases. For example, Yeung et al.[54] studied three gene expression data sets with several data transformations and found that the data sets all fit the Gaussian model poorly. Another example is that, the model constrains each gene to be in exactly one group, which cannot capture the biological fact that many gene products participate in more than one biological process[55].

3.3 Kernel function analysis

Kernel function analysis method is introduced by Vert et al.[56] There, the gene networks and expression data are transformed into two kernel functions, and consequently active pathways are extracted by performing a regularized form of canonical correlation analysis. The correlation between two different elements, nodes in pathway graph and expression profiles, are used to assess the relationship between gene expression and pathway function information. The main goal of this algorithm can be simplified as finding vectors, which denote the variations among gene expression profiles as large as possible and the expression features among adjacent genes in a pathway as continuous as possible. These vectors are characterized as linear combinations of expression profiles. By encoding the expression data and pathways into two kernel functions, the problem is solved by canonical component analysis.

The kernel function analysis has been generalized to many kinds of data, including aminoacid sequences[57], phylogenetic profiles[58] and promoter regions[59]. All these data are represented as kernel functions and take correlation analysis. Furthermore, the form of correlations is not limited to two variables[60]. An attempt of multiple kernel functions has been adapted for operon detection in bacterial genomes[61].

The kernel function analysis is efficient but really complicated. With the increase of factors, the dimension and complexity of analysis will rise greatly. Besides, the integration of multiple kernel functions is also a problem. Currently the solution is to form a convex combination of kernels by setting nonnegative weight to each kernel. Lanckriet et al. gave an optimized algorithm to estimate the kernel weights by semidefinite programming, yet the process is seriously time and space consuming[62].

3.4 Graph-iterative group analysis

The graph-iterative group analysis (GiGA) by Breitling et al. has also provided a statistic way to identify active subgraphs in the biological knowledge graph[63]. Genes that share a Gene Ontology annotation or participate in a metabolic pathway are connected to build the graph. At the same time, a rank list of genes sorted by differential expression is provided and each node is ranked according to the gene allocated to it. Then local minima are identified in the graph as the nodes with a lower rank than all of their direct neighbors. The local minima nodes are considered to be significant centers of the subsequent subnetworks. From each local minima, a subnetwork is extended by including the neighboring node with the next highest rank (m) and, if present, all adjacent nodes of ranks equal to or smaller than m. To assess the significance of each extension, a p-value is calculated as the probability of observing n genes whose ranks are equal to or better than m from all N genes. The extension process ends when no more nodes can be included. The subnetwork with the highest score is output as ideal relevant regions.

In GiGA, the calculation can only deal with one condition at one time, which means that it cannot capture the complete variances of gene expression through multiple conditions. The method will not perform very well in many experiments, especially in the case of time series dataset. Another disadvantage of GiGA is that the genes are ranked by the log value of expression levels, from positive to negative. However, in fact, a biological pathway contains both induced expressed genes and repressed expressed genes.

In a pathway, there are both up-regulated and down-regulated relationships among genes. Thus the operation that ranks genes from positive to negative cannot detect the pathways including both remarkable induced genes and repressed genes, whereas these genes and their regulation relationships are usually extremely important.

3.5 Wavelet transform

Early this year, Konig et al. published a novel method to discover central components of metabolic networks[64], which identified distinct expression patterns from *E. coli* under both the aerobic and anaerobic conditions using wavelet transform. In this method, the metabolic network is represented as a graph, with the enzymes as edges and metabolites as nodes. After applying a clustering method, the metabolic graph is grouped into several sub-graphs. Expression data of all samples are mapped and features for every sample are gained by performing Haar-wavelet transformations on the gene expression patterns of sub-graphs. The most significant features are extracted by modified *t*-test and SVM process. Finally, the sub-graphs containing the most significant features are considered as relevant ones that can represent the adaptation of the cells to changing environmental conditions.

This wavelet transform method combines some machine learning approaches and elucidates relevant sub-graphs by testing all possible patterns within the metabolic network. However, the sub-graphs are extracted in advance by graph clustering method, which is based on the topology of the metabolic network only. This kind of initialization limits the selection of sub-graphs as there are many other important factors need to be considered to extract sub-graphs besides topological connections.

4 Validation problem

Different methods may lead to different results. How to assess the significance of the results is a serious task. For the first category of analysis, gene sets significance assessment, the most widely used validation method is to calculate a *p*-value through multiple randomized data, which is used to verify whether the analysis result is more significant than expected. The false discovery rate (FDR) is another guideline for gene sets significance evaluation. FDR can predict the number of false discovered results, for example, numbers of pathways that would be expected by chance to have a notable *p*-value. Usually, there are both false-positive and false-negative rate that should be calculated separately.

For the second category of analysis, active components identification, there are no standard validation criteria. The commonly treatment compares resulted active components with existing evidence, for example, known regulatory circuits, gene clusters in other literatures and so on. Besides, there are also other validation criteria for some special approaches. In expectation maximization method of Segal et al.[53], three principles are used to evaluate the learned model: prediction of removed interactions, functional annotations enrichment of pathways and coverage of protein complexes. It is generally agreed that a good molecular pathway should have three properties: (1) stable when a small portion of interactions are taken away; (2) coherent in functional annotations and (3) as many as possible protein complexes are assigned to the same pathway[53]. These validation criteria are only applied in expectation maximization method and might be adapted to other procedures.

5 Comparison of methods

In order to examine the overlaps and differences in the results of various methods, comparisons among different methods are necessary. Curtis et al. have compared some methods for gene sets assessment[18]. They found that binomial distribution and *z*-scores have similar results while more gene sets are shown to be downregulated by GSEA[18]. Here, we try to compare two latest methods of active components identification: GiGA[63] and wavelet transform[64].

We used the microarray dataset in wavelet transform. The dataset is from Covert et al.[68], which denotes *E. coli* genome expression under aerobic and anaerobic conditions. Knoig et al. have normalized the dataset and selected 43 hybridizations of one wild-type sample and six strains with knockouts of key transcriptional regulators in the oxygen response: $\triangle arcA$, $\triangle appY$, $\triangle fnr$, $\triangle oxyR$, $\triangle soxS$ and double knock $\triangle arcA\ \triangle fnr$[64]. Signal-to-noise ratio (SNR) in GSEA[28] is introduced to produce the ranked list of reactions for GiGA (the genes are represented as reactions in wavelet transform, thus the reactions instead of genes are analyzed in GiGA for comparing purpose). We also used other rank proce-

dures such as standard deviation for GiGA and found that the results from different rank procedures agree closely, especially in the most significant subgraphs.

We compared the subgraphs extracted from GiGA and wavelet transform. Possibly because wavelet transform allows overlapping subgraphs while GiGA dose not, there are many more numbers of subgraphs resulted from wavelet transform than from GiGA. Only one subgraph (formate metabolism) exists in the results of both GiGA and wavelet transform, which is the most significant subgraph in both methods. However, there are many overlaps between the results of GiGA and the 40 first ranking reations listed in wavelet transform method. The results of comparison are summarized in Fig. 1. This may support the concept that wavelet transform is designed to identify complex expression patterns that cannot be founded straightforward[64].

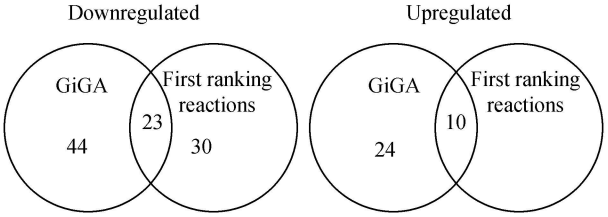

Fig. 1. Comparison between the results of GiGA and the 40 first ranking reations listed in wavelet transform method. The figure shows the number of either downregulated or upregulated reactions in subgraphs by GiGA and first ranking reations list in wavelet transform, as well as their overlaps. GiGA was performed on a gene list, ranked by SNR.

6 Discussions and prospects

With the development of microarray technology, more and more algorithms on genome-wide data are becoming available. How to efficiently and thoroughly mine the knowledge under the expression data is the most interesting challenge up to date. Incorporating biological knowledge seems to be the tendency in expression data analysis, as the biological knowledge can provide guidance to connect expression profiles with biological functions. In this review, we have summarized one form of biological knowledge, biological networks, and two main research aspects for microarray data analysis based on biological networks: identification of active components in biological networks and assessment of gene sets significance.

The microarray data analysis based on biological networks integrates biological annotation knowledge to expression profiles, which show more potential to obtain biological significant results. The structure of biological networks intuitively reflects the annotation relationships among genes. For example, neighbor nodes in the biological networks (nodes with connections) represent genes with annotation correlations, which are convenient to next analysis. Furthermore, the microarray data analysis based on biological networks involves contributions of multiple genes, which capture the biological fact that cellular processes often act as interactions of many genes.

A possible improvement deserving attention is that all kinds of clues should be integrated regularly. For example, in pathway scores, the expression activity level, the expression similarity and structure character of pathways are all considered, yet no criteria with all these three properties are provided. In the subnetwork identification methods, either expression activity or expression similarity is judged, while there are also no approaches incorporating both properties. As an attempting, we have incorporated gene expression synchrony assessment to GSEA, which shows better results than current GSEA method (manuscript in preparation).

Another potential improvement is to add some details of biological annotations. For both categories of analyses, detailed biological information is usually ignored. The active biological subnetworks and functional gene sets are considered as a whole and the gene relationships inside the subnetworks and gene sets are not exhibited. The subnetworks or gene sets analysis is designed to avoid the limitations of single-gene analysis. However, it is obvious that the detailed information will help to analyze and interpret results. These information may include the degree of nodes (the number of edges connected to each node) in biological networks, the structure of subnetworks, the connections among active genes in subnetworks, the co-regulated regions in a subnetwork and so on. Such information may bring in some heuristic idea. For example, in graph theory, the degree of nodes represents the central level of the nodes in the network, thus integrating this character may obtain gene significance in biological network structure in addition to expression profiles.

In addition, to the biological annotations, there is so far no consistent standard for different annotation sources. Situations will probably arise such as some expression data tend to show more significance to Gene Ontology gene sets, while others show more

tendency to KEGG gene sets. The function module refining process by Segal et al.[26] is perhaps the initiatory attempt for this problem. Moreover, in the identification of subnetworks, annotations from multiple sources could be thought to integrate together. That is to say, one might build a network involving all or some of protein interactions, metabolic information, Gene Ontology categories, and regulatory relationships so on. Using this large network, not only the discovery of better subnetworks but also the correlation analysis among different annotations can be imagined potentially.

For further consideration, the two categories of analysis can be integrated to some extent. The principles used in two categories such as expression scores can be adapted to each other. Moreover, the results from the second category of analysis can be validated by the first category of analysis, that is to say, if some active subnetworks are obtained from certain active components identification methods, then these subnetworks can be assessed by some gene sets assessment methods to verify whether the resulted subnetworks are significant or not.

Last but not the least, there are some exterior facts that will affect the analyses. As mentioned in [18], there is no single repository for various gene identifiers, such as Affymetrix probe IDs, gene symbols, accession numbers and so on. Some of them are too obsolete or redundant to map an enzyme or protein to a microarray probe ID. Besides, there are large proportions of genes that have no function annotations yet, thus all the analyses are processed on a small subset of genes. Hopefully, the analysis of microarray data with biological networks will be more powerful when the annotations become more complete.

Acknowledgement The authors would like to thank Dr. Lu Qiang for helpful comments and suggestions.

References

1 Duggan D. J., Michael B., Chen Y. et al. Expression profiling using cDNA microarrays. Nature Genetics, 1999, 21(1 Suppl): 10—14.
2 Efron B. and Tibshirani R. Empirical bayes methods and false discovery rates for microarrays. Genet. Epidemiol., 2002, 23(1): 70—86.
3 Eisen M. B., Spellman P. T., Patrick O. B. et al. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 1998, 95(25): 14863—14868.
4 Getz G., Levine E. and Domany E. Coupled two-way clustering analysis of gene microarray data. Proc. Natl. Acad. Sci. USA, 2000, 97(22): 12079—12084.
5 Cheng Y. and Church G. M. Biclustering of expression data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB). 2000, 93—103.
6 Jiang D., Tang C. and Zhang A. Cluster analysis for gene expression data: a survey. IEEE Transsactions on Knowledge and Data Engineering, 2004, 16(11): 1370—1386.
7 Adryan B. and Schuh R. Gene-Ontology-based clustering of gene expression data. Bioinformatics, 2004, 20(16): 2851—2852.
8 Cheng J., Cline M., Martin J. et al. A knowledge-based clustering algorithm driven by Gene Ontology. Journal of Biopharmaceutical Statistics, 2004, 14(3): 687—700.
9 Fang Z., Yang J., Li Y. X. et al. Knowledge guided analysis of microarray data. Journal of Biomedical Informatics, accepted.
10 Kanehisa M., Goto S., Hattori M. et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res., 2006, 34(Database issue): D354—357.
11 Bader G. D., Betel D. and Hogue C. W. BIND: the biomolecular interaction network database. Nucleic Acids Res., 2003, 31(1): 248—250.
12 Xenarios I., Salwinski L., Duan X. J. et al. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res., 2002, 30(1): 303—305.
13 Ashburner M., Ball C. A., Blake J. A. et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 2000, 25(1): 25—29.
14 Dahlquist K. D., Salomonis N., Vranizan K. et al. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat. Genet., 2002, 31(1): 19—20.
15 Doniger S. W., Salomonis N., Dahlquist K. D. et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biology, 2003, 4: R7.
16 Wingender E., Chen X., Fricke E. et al. The TRANSFAC system on gene expression regulation. Nucleic Acids Res., 2001, 29(1): 281—283.
17 Finn R. D., Mistry J., Schuster-Bockler B. et al. Pfam: clans, web tools and services. Nucleic Acids Res., 2006, 34(Database issue): D247—251.
18 Curtis R. K., Oresic M. and Vidal-Puig A. Pathways to the analysis of microarray data. Trends Biotechnol., 2005, 23(8): 429—435.
19 Tavazoie S., Hughes J. D., Campbell M. J. et al. Systematic determination of genetic network architecture. Nat. Genet., 1999, 22(3): 281—285.
20 Zeeberg B. R., Feng W., Wang G. et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biology, 2003, 4: R28.
21 Khatri P., Bhavsar P., Bawa G. et al. Onto-tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. Nucleic Acids Res., 2004, 32: W449—W456.
22 Choi J. K., Choi J. Y. Kim D. G. et al. Integrative analysis of multiple gene expression profiles applied to liver cancer study. FEBS Lett., 2004, 565(1—3): 93—100.
23 Mootha V. K., Lindgren C. M., Eriksson K. F. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet., 2003, 34(3): 267—273.
24 Zien A., Kuffner R., Zimmer R. et al. Analysis of gene expression data with pathway scores. In: Proc. Int. Conf. Intell. Syst. Mol. Biol., 2000, 8: 407—417.
25 Kurhekar M. P., Adak S., Jhunjhunwala S. et al. Genome-wide pathway analysis and visualization using gene expression data. Pac. Symp. Biocomput., 2002: 462—473.

26 Segal E., Friedman N., Koller D. et al. A module map showing conditional activity of expression modules in cancer. Nat. Genet., 2004, 36(10): 1090—1098.

27 Damian D. and Gorfine M. Statistical concerns about the GSEA procedure. Nat. Genet., 2004, 36(7): 663.

28 Subramanian A. Tamayo P., Mootha V. K. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA, 2005, 102(43): 15545—15550.

29 Kim S. Y. and Volsky D. J. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics, 2005, 6: 144.

30 Mewes H. W., Frishman D., Mayer K. F. X. et al. MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res., 2006, 34: D169—172.

31 Letunic I., Copley R. R., Pils B. et al. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res, 2006, 34(Database issue): D257—260.

32 Spellman P. T., Sherlock G., Zhang M. Q. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell., 1998, 9(12): 3273—3297.

33 DeRisi J. L., Iyer V. R., Brown P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 1997, 278(5338): 680—686.

34 Chu S., DeRisi J., Eisen M. et al. The transcriptional program of sporulation in budding yeast. Science, 1998, 282(5389): 699—705.

35 Su A. I., Cooke M. P., Ching K. A. et al. Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA, 2002, 99(7): 4465—4470.

36 Segal E., Shapira M., Regev A. et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet., 2003, 34(2): 166—167.

37 Su A. I., Wiltshire T., Batalov S. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA, 2004, 101(16): 6062—6067.

38 Su A. I., Welsh J. B., Sapinoso L. M. et al. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res., 2001, 61(20): 7388—7393.

39 Ramaswamy S., Tamayo P., Rifkin R. et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA, 2001, 98(26): 15149—15154.

40 Higashi-ku H. Signaling Pathway Database. http://www.grt.kyushu-u.ac.jp/spad/menu.html [1998-10-16]

41 Li J., Ning Y., Hedley W. et al. The molecule pages database. Nature, 2002, 420(6916): 716—717.

42 Signal transduction knowledge environment. http://stke.sciencemag.org [2006]

43 Human protein reference database. www.hprd.org [2006]

44 Sigma-Aldrich. http://www.sigmaaldrich.com/Area-of-Interest/Biochemicals/Enzyme-Explorer/Key-Resources.html [2006]

45 SupperArray. www.superarray.com [2006]

46 Brentani H., Caballero O. L., Camargo A. A. et al. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc. Natl. Acad. Sci. USA, 2003, 100(23): 13418—13423.

47 Olivier M., Eeles R., Hollstein M. et al. The IARC TP53 database: new online mutation analysis and recommendations to users. Hum. Mutat., 2002, 19(6): 607—614.

48 Armstrong S. A., Staunton J. E., Silverman L. B. et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet., 2002, 30(1): 41—47.

49 Bhattacharjee A., Richards W. G., Staunton J. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA, 2001, 98(24): 13790—13795.

50 Beer D. G., Kardia S. L., Huang C. C. et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med., 2002, 8(8): 816—824.

51 Ideker T., Ozier O., Schwikowski B. et al. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics, 2002, 18(Suppl 1): S233—240.

52 Patil K. R. and Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc. Natl. Acad. Sci. USA, 2005, 102(8): 2685—2689.

53 Segal E., Wang H. and Koller D. Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics, 2003, 19(Suppl 1): i264—271.

54 Yeung K. Y., Fraley C., Murua A. et al. Model-based clustering and data transformations for gene expression data. Bioinformatics, 2001, 17(10): 977—987.

55 Hvidsten T. R., Lagreid A. and Komorowski J. Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. Bioinformatics, 2003, 19(9): 1116—1123.

56 Vert J. P. and Kanehisa M. Extracting active pathways from gene expression data. Bioinformatics, 2003, 19(Suppl 2): II238—II244.

57 Jaakkola T., Diekhans M. and Haussler D. A discriminative framework for detecting remote protein homologies. J. Comput. Biol., 2000, 7(1—2): 95—114.

58 Vert J. P. A tree kernel to analyze phylogenetic profiles. Bioinformatics, 2002, 18: S276—S284.

59 Pavlidis P., Furey T. S., Liberto M. et al. Promoter region-based classification of genes. In: Proceedings of the Pacific Symposium on Biocomputing, 2001, 151—163.

60 Francis R. B. and Jordan M. I. Kernel independent component analysis. J. Machine Learning Res., 2002, 3: 1—48.

61 Yamanishi Y., Vert J. P., Nakaya A. et al. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics, 2003, 19(Suppl 1): 323—330.

62 Lanckriet G. R., De B. T., Cristianini N. et al. A statistical framework for genomic data fusion. Bioinformatics, 2004, 20(16): 2626—2635.

63 Breitling R., Amtmann A. and Herzyk P. Graph-based iterative Group Analysis enhances microarray interpretation. BMC Bioinformatics, 2004, 5: 100.

64 Konig R., Schramm G., Oswald M. et al. Discovering functional gene expression patterns in the metabolic network of Escherichia coli with wavelets transforms. BMC Bioinformatics, 2006, 7: 119.

65 Gasch A. P., Werner-Washburne M. The genomics of yeast responses to environmental stress and starvation. Funct. Integr. Genomics, 2002, 2(4—5): 181—192.

66 Bairoch A., Boeckmann B., Ferro S. et al. Swiss-Prot: juggling between evolution and stability. Brief Bioinform., 2004, 5(1): 39—55.

67 Keseler I. M., Collado-Vides J., Gama-Castro S. et al. EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res., 2005, 33(Database issue): D334—337.

68 Covert M. W., Knight E. M., Reed J. L. et al. Integrating high-throughput and computational data elucidates bacterial networks. Nature, 2004, 429(6987): 92—96.