DRUG DISCOVERY TODAY BIOSILICO

Bioinformatic methods for integrating whole-genome expression results into cellular networks

Duccio Cavalieri and Carlotta De Filippo

Extracting a comprehensive overview from the huge amount of information arising from whole-genome analyses is a significant challenge. This review critically surveys the state of the art methods that are used to connect information from functional genomic studies to biological function. Cluster analysis methods for inferring the correlation between genes are discussed, as are the methods for integrating gene expression information with existing information on biological pathways and the methods that combine cluster analysis with biological information to reconstruct novel biological networks.

Regulation of gene expression and protein activity is central to the function of molecular and cellular systems. High-throughput methods (e.g. cDNA microarrays, ChIP-chips, yeast two-hybrid interaction studies) provide various views of the participating genes, RNAs and proteins and their interactions, and it is becoming increasingly clear that turning the data deluge arising from these global analyses into knowledge represents an unprecedented challenge. This review discusses three main bioinformatic approaches (Figure 1) that aim to answer three major questions:

1. How can we properly visualize gene expression information in the context of existing knowledge of pathways or networks?
2. What are the advantages and the limitations of the application of existing probabilistic and graphical models to 'omics' data, and how can we improve these methods?
3. How can we infer regulatory networks from heterogeneous sources of data (including gene expression data, promoter sequence information, comparative genomics, functional annotations, and genetic and protein interaction data)?

Duccio Cavalieri*
Carlotta De Filippo
Department of Pharmacology, University of Florence, Viale Pieraccini 6, 50139 Florence Italy
*e-mail: duccio.cavalieri@unifi.it

Methods that use genomic information to infer correlations between genes and clusters of genes

One of the challenges in interpreting microarray data is to group genes on the basis of similar regulation and function, or similar cellular state and biological phenotype. This is a multivariate problem of extremely high dimensionality that has proven attractive to biostatisticians, so that in the past few years statistical analysis has been among the most active fields of microarray research. Identification of unknown classes of co-regulated genes using whole-genome expression profiles (unsupervised learning, cluster analysis) or classification into known classes of functionally or structurally related genes (supervised learning, discriminant analysis) are two common techniques used in gene expression experiments.

Clustering algorithms based on probability models are methods in which the data are assumed to be generated by a finite mixture of underlying probability distributions. One important feature of mixture modelling is that posterior probabilities (i.e. the probability to belong to the class, given the observed

1359-6446/05/$ – see front matter PII: S1359-6446(05)03433-1

gene expression data) of class membership are obtained. Following the first papers on hierarchical clustering [1,2], research in cluster analysis has used graphical Gaussian modeling [3], clustering using reliable, informative data subsets [4], partitional clustering and motif discovery [5], clustering based on singular value decomposition [6,7], simulated annealing [8], graph theory [9], self-organizing maps [10] and scale-free networks [11]. Some clustering methods, developed to identify subsets of genes with expression patterns that vary in a similar fashion across conditions, can be partially or fully supervised by using known properties of the genes or samples to assist in finding meaningful associations [12,13]. Tested on standard datasets, these methods all identify similar core clusters of genes that are most strongly affected by coordinate regulation. All these methods are based on the assumption that functionally related genes have similar gene expression levels, an assumption that might mean that important genes, in particular regulators, are overlooked. In fact, transcription factors can change in function without necessarily changing in expression, but rather in state, through binding to an activator or by proteolytic activation. Another limit of clustering is that these methods do not directly address the functional relationships among genes.

Methods that integrate functional genomics result in metabolic or cellular pathways

Genes never act alone in a biological system, but participate in a cascade of networks. Pathways represent the biologist's way of describing the nature of biological interactions and control networks. Biochemical networks link enzymes to the flow of substrates and products of the different reactions; regulatory networks describe how the expression of the genes encoding the enzymes of the pathways are regulated; protein–protein interaction networks annotate interacting proteins. One of the main differences between a network and a pathway is that a pathway is a linked

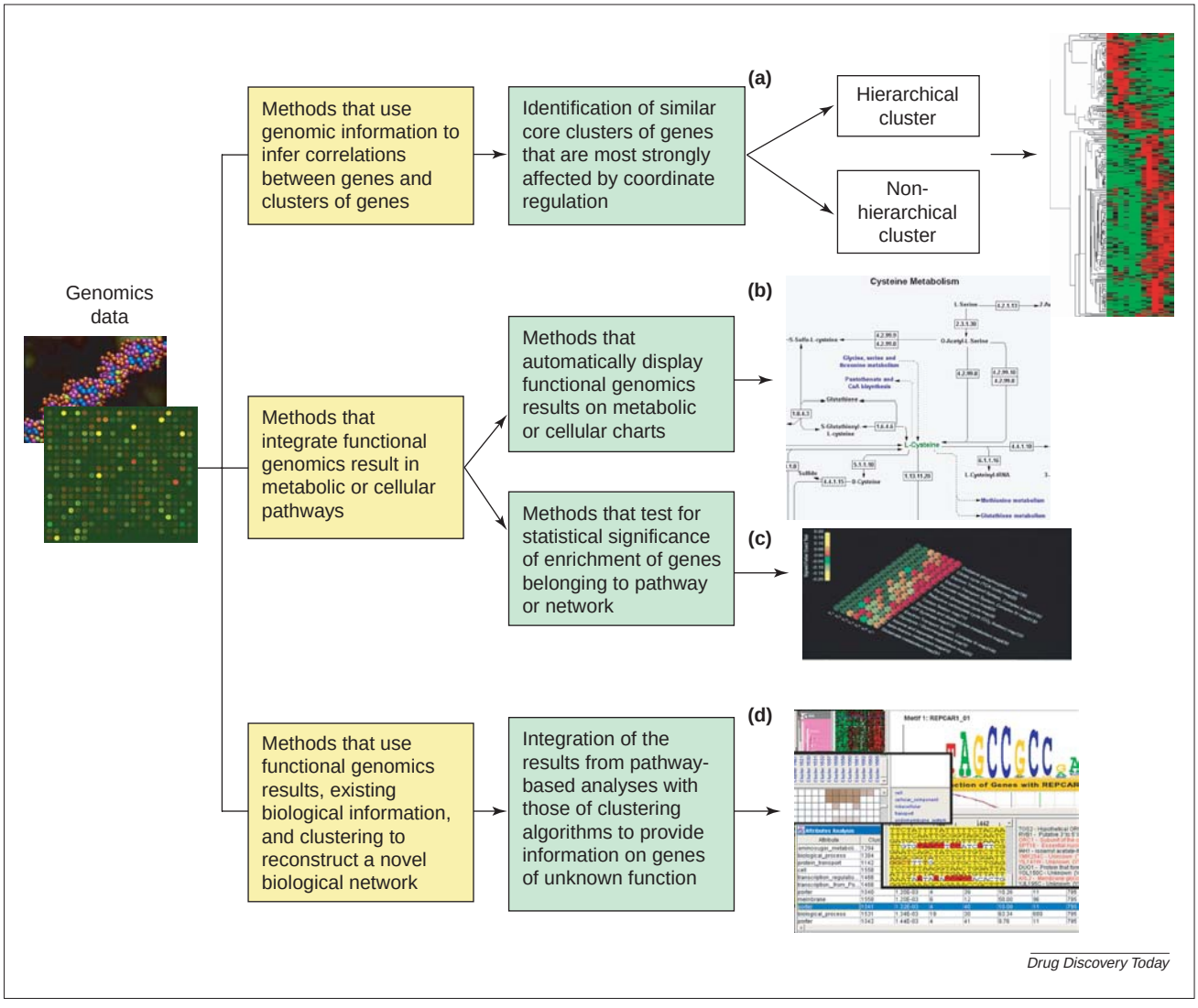

FIGURE 1

Flowchart illustrating the three main approaches for microarray data analysis in the context of cellular and biochemical networks. Visualization of output generated by **(a)** Hierarchical analysis, **(b)** GeneMapp software, **(c)** Pathway Processor software and **(d)** GeneXPress software.

set of biochemical reactions, and includes the idea of directionality, where lines become arrows, and contain information on the energy and metabolite flow, and where the product of a reaction is either the substrate or the enzyme that catalyzes a subsequent reaction. Several efforts have therefore addressed the analysis of microarray data in the context of state-of-the-art knowledge of biological networks or pathways.

Pathway-based microarray analysis methods look for patterns of expression variation in predefined classes of genes, such as those involved in metabolism, the cytoskeleton, cell-division control, apoptosis, membrane transport, sexual reproduction, signalling, and so forth, with the aim of integrating information obtained on a genomic scale with the biological information accumulated through years of research of molecular genetics, biochemistry and cell physiology. The integration of genomic and physiological information is now increasingly important with the emergence of 'systems biology', which attempts to simultaneously study the expression patterns and activity of all genes, together with proteomic and metabolomic data. The analysis of different kinds of genomic data from a network perspective promises to foster a new level of understanding of the system as a whole.

Efforts to establish proper gene ontology (GO) [14] are becoming increasingly important with the progress of the various genome sequencing projects, and are relevant to the interpretation of genomics data in their biological context. The GO database provides a useful tool to annotate and analyze the functions of a large number of genes.

The availability of properly annotated pathway databases is one of the requirements for analyzing microarray data in the context of biological pathways. The development of pathway databases is continuing apace, with several resources available publicly, such as the Kyoto encyclopedia for genes and genomes, KEGG [15], MetaCyC EcoCyCand AraCyC [16–19], MIPS yeast pathways, or commercially, such as BioSilico [20], and GeneGo (see Box 1 for URLs).

From the need for a unique annotation of all the biochemical reactions stems another effort, Reactome, a joint project of Cold Spring Harbor Laboratory, EBI and GO Consortium [21]. Reactome is a curated database of biological processes, covering biological pathways ranging from the basic processes of metabolism to high-level processes such as hormonal signalling. Although it is targeted at human pathways, it also includes many individual biochemical reactions from rat and mouse. The information in this database is provided by bench biologists who are experts in a specific domain of biology, edited and entered into a relational database which is cross-referenced with PubMed, GO, and the sequence databases LocusLink, Ensembl and SwissProt. The information is then reviewed by other biological researchers for consistency and accuracy and made public.

In parallel to the development of pathway databases, the development of a dedicated universal standard data exchange format for pathway information is essential for sharing, evaluating and developing pathway information resources and pathway-based models. The first effort in this direction has been the Systems Biology Markup Language (SBML) [22]. SBML is a format for representing models of biochemical reaction networks such as metabolic networks, cell-signalling pathways and regulatory networks, which has been evolving since mid-2000 through the efforts of an international group of software developers and users, the Systems Biology Workbench. SBML is mainly a support for integrating different modelling and simulation tools. It is particularly suited for building programs that model how two genes construct an oscillatory circuit or a feedback inhibition loop. This language is very detailed and specifically describes the biological variables, including biochemical reaction, substrate of the reaction, products, cell localization, K_m, etc.

Another effort to realize a pathway exchange and representation format is the BioPAX project. Established in 2002, BioPAX is a workgroup with the goal of developing a common exchange format for biological pathways data. The initial release of BioPAX level 1 is now available as a Web Ontology Language (OWL) file. The OWL Web Ontology Language is designed for use by applications that need to process the content of information and facilitates greater machine interpretability of Web content than that supported by XML, RDF and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. The main difference between BioPAX and SBML is that BioPAX is essentially a format for exchange of information between different databases, whereas SBML is a support for modelling and simulation. BioPAX has been developed to respond to the need to interpret genomics data, and aims to represent the biochemical reactions in the big picture of cellular metabolism, describing biochemical pathways and protein–protein interaction data. SBML, by contrast, has not been designed specifically to represent microarray or genomics data, but is more suited for the modelling of a series of defined dynamic biological events.

Current efforts to develop pathway-based approaches for the analysis of microarray data can be divided into two main classes: first, methods that automatically display functional genomics results on metabolic or cellular charts; and second, methods that test for statistical significance of enrichment of genes belonging to the same class, pathway or networks.

Methods that automatically display functional genomics results on metabolic or cellular charts

A pathway map is a diagram showing biological relationships between genes or gene products based on organizing principles, such as metabolic pathways, signal transduction cascades or subcellular locations. Many authors have

BOX 1

Summary of the names of the various applications discussed in the article and the relative URLs

Name	URL
Acuity	http://www.axon.com/
AraCyC	http://www.*Arabidopsis*.org/tools/aracyc/
ArrayDB	http://genome.nhgri.nih.gov/arraydb/
BIND	http://bind.ca/
BioPAX	http://www.biopax.org/
BioSilico	http://www.biosilico.com/
CARRIE	http://zlab.bu.edu/CARRIE-web
Cytoscape	http://www.cytoscape.org
DAVID	http://www.david.niaid.nih.gov
EcoCyCand	http://ecocyc.PangeaSystems. com/ecocyc/
Gene Ontology	http://www.geneontology.org/
GeneGo	http://www.genego.com
GeneMerge	http://www.oeb.harvard.edu/hartl/laboratory/publications/GeneMerge/GeneMerge.html
GeneSpring	http://www.silicongenetics.com
GeneXpress	http://genexpress.stanford.edu/
GenMapp	http://www.genmapp.org/
Gostat	http://gostat.wehi.edu.au/
KEGG	http://www.genome.ad.jp/kegg/
MetaCyC	http://www.genome.jp/kegg/
MIPS Yeast Pathways	http://mips.gsf.de/proj/yeast/CYGD/db/pathway_index.html
OpenDX	http://www.opendx.org/
OpenDX	http://www.opendx.org
Pathway Miner	http://www.biorag.org/pathway.html
Pathway Processor	http://cgr.harvard.edu/cavalieri/pp.html
Proteome Bioknowledge	http://www.incyte.com/
PSI-MI	http://psidev.sourceforge.net/
Reactome	http://www.reactome.org/
Rosetta Resolver	http://www.rosettabio.com/
SBML plug-in	http://sbml.org/index.psp/
SGD	http://www.yeastgenome.org/
SBML	http://sbml.org/index.psp/
TRANSFAC	http://transfac.gbf.de/

manually mapped transcriptional changes to metabolic charts; this time-consuming process has prompted the development of automated methods. Efforts have been made to display expression data with pathway information in databases, such as ArrayDB [23], SGD [24] and KEGG [15]. Some commercial microarray analysis or bioinformatics packages, such as Rosetta Resolver™ (Rosetta Inpharmatics LLC), GeneSpring™ (Silicon Genetics, Agilent Technologies), Acuity™ (Axon instruments), or GeneGO™ (GeneGo) and the Proteome Bioknowledge Library™ (Incyte) have developed features that enable the display of gene expression data in the context of metabolic, genetic or interaction maps.

GeneMapp [24] is one of the most interesting freely available computer applications. This program is designed to visualize global gene expression or other kinds of genomic data in the context of hundreds of existing pathway MAPPs (see below) and thousands of Gene Ontology Terms, and to facilitate the exchange of pathway-related data among investigators.

A MAPP is a special file format, assigning to each gene an identification (ID) taken from GenBank, or a user-defined ID system. The MAPP format is independent of the gene expression data and of the organizing principle, and allows the visualization of well known pathways from curated databases or the building of custom MAPPs according to criteria defined by the user, without necessarily specifying the number and type of interactions between the elements. The gene ID is the link (hyperlink) between the gene object on the MAPP, the gene expression data and the available information for that gene contained in the GenMAPP database, a gene library that includes annotation and biological and functional information on the gene of interest available on public databases. Information for genes not in the database may be entered by the user. One of the most interesting features is MAPPBuilder, a program that enables the user to group genes by any organizing principle and provides graphical objects that can be placed and manipulated on a 'drafting board'. These include general objects, such as lines and arrows, as well as biological items, such as receptor- and ligand-binding symbols and subcellular components, and the 'gene', which represents a gene or gene product, thus allowing the creation of detailed custom MAPPs. The visualization tools of MAPPBuilder are indeed a strong point of GeneMApp, in respect to the other publicly available visualization tools, as they clearly define the nature of connections of the elements in the pathway, discriminating

the use of a biochemical intermediate from an interaction between different proteins, or a feedback loop. To visualize whole-genome expression on the MAPPs, the gene objects can be color-coded and labelled on the basis of microarray data [25]. One of the limits of the color-coding visualization is that it provides only qualitative but not quantitative information, without the ability to display simultaneously genomics data from different sources, such as proteomics or metabolomics, and relying on the use of the hyperlink to obtain more information on a particular gene or reaction.

GenMAPP provides a unique additional feature in respect to the existing pathway resources, as it allows the user to modify pathways for their own use or to design new pathways. Thus, it is a powerful tool for freely exchanging pathway-related data among investigators. One of the limits of this approach is the heterogeneity of the pathways stored in the database. Ideally, the MAPPS should have a well-established common annotation and should be properly validated. The need for a controlled pathway definition suggests that the potentials of GeneMapp will be greatly enhanced by the interaction with other sources, such as BioPAX or Reactome, addressed to properly annotate and curate the different pathways.

Cytoscape is another useful tool for the integration of genomic data onto gene networks. This open-source tool allows the visualization, drawing and editing of molecular interaction networks [26]. Molecular interaction networks are constructed from raw interaction files containing lists of protein–protein and/or protein–DNA interaction pairs. This is particularly useful for yeast studies, which benefit from large sources of pairwise interactions, available through the BIND and TRANSFAC databases. The program allows the loading and saving of previously constructed two-dimensional interaction networks in GML format (Graph Markup Language), and importation of gene functional annotations from the Gene Ontology (GO) database, and has an SBML plug-in and PSI-MI plug-in, thus making the tool compatible with upcoming community standards for describing and modelling molecular interaction. The program allows superposing gene expression ratios and p-values on the interaction network. A variety of layout algorithms and analysis tools enable filtering of the network to select subsets of nodes and/or interactions. One limitation is that the graphical visualization is not necessarily intuitive and could be greatly improved, and thus currently it is useful only to investigate interaction networks. This tool has a great potential and could be adapted to design other kinds of networks, such as regulatory networks, and integrate them with expression or proteome data.

A new and interesting tool to graphically display genomic results is the Database for Annotation, Visualization, and Integrated Discovery (DAVID) [27]. This is an integrated database that associates data from several public databases to lists of genes and displays graphic summaries of functional information assigning genes to KEGG metabolic processes (KEGG charts) and Gene Ontology functional categories (GO charts) at different term specificity levels, and a tool called 'DomainCharts' that groups genes according to conserved protein family domains. The limit of this tool compared with GeneMapp is that it restricts the analysis to a pre-defined set of pathways, and that the graphical representation of the data on GO charts is not the most appropriate to describe a pathway.

One of the common limitations of all the visualization programs presented so far is that they do not automatically indicate the statistical significance of the change of a pathway, making it hard to select the most interesting of the different pathway maps.

Methods that test for statistical significance of enrichment of genes belonging to the same class, pathway, or network

The development of statistical methods to assess the significance of alteration in expression in diverse cellular pathways is of increasing interest. Several microarray papers have reported activation or repression of a given pathway, but too often the researcher finds what is expected, or what is already known, and the assumption has not been properly supported from a statistics standpoint.

To assess the significance of the genes of a pathway to be coordinately changed in expression in a given experiment, several factors have to be taken into account: first, the number of open reading frames (ORFs) for which expression has altered in each pathway; second, the total number of ORFs contained in the pathway; third, the proportion of the ORFs in the genome contained in a given pathway; fourth, the correlation of the pathways, to select the most appropriate statistical test.

Pathway Processor [28] was the first program that implements both a statistical method to determine which pathways are most affected by transcriptional changes and a visualization tool to map expression data from multiple whole-genome expression experiments on metabolic pathways.

The method automatically associates an ORF with a given biochemical step according to the information contained in 92 pathway files from KEGG database. KEGG has been chosen for the concise and clear way in which the genes are interconnected and for the great effort in keeping the information up to date. Thus, this approach allows the proper testing of the ability of the statistical method to score a limited number of well curated and biochemically annotated pathways.

The 'Pathway Analyzer' program of 'Pathway Processor' uses the Fisher Exact Test to generate a *P*-value which indicates the probability that the pathway would contain as many or more affected genes than actually observed, the null hypothesis being that the relative changes in gene expressions in the pathway are a random subset of those observed in the experiment as a whole. The value

of the statistical test is multiplied by +1 or –1, to indicate whether the majority of the genes in a particular pathway are upregulated or downregulated, and is therefore called the 'Signed Fisher Exact Test', whose value is a positive real number (between 0 and 1) corresponding to the *P*-value of the Fisher Exact Test for the pathway. Sorting for the Signed Fisher Exact Test can be used to compare different experiments; the comparison can be represented graphically using programs as common as Excel™ or TreeView, or more sophisticated ones such as OpenDX.

Common practice suggests that the validity of the statistical methods has to be tested on a well known biological problem. The yeast *Saccharomyces cerevisiae* represents the best model for testing tools connecting expression data with biochemical pathways. The potential of the Fisher exact test and of Pathway Processor as a whole was tested by examining the differential expression during the diauxic shift, a well known physiological switch from fermentation to respiration of the yeast cell upon exhaustion of the available carbon sources, described in one of the proof-of-concept papers for microarray technology [29]. The analysis highlighted affected pathways that had previously been detected only through cumbersome analysis of the results with repeated references to KEGG, MIPS and SGD.

The *P*-values can also be used to select pathways to examine more closely using Expression Mapper, the second program of Pathway Processor, displaying differences in expression on KEGG metabolic charts. One of the major limits of this program is that it has been initially designed only for the analysis of expression data from two well established model organisms, yeast and *Bacillus subtilis*, according to the KEGG biochemical pathways. This limitation is derived from the choice to use only well annotated pathways, with minimal overlapping, for the statistical tests. A new updated version (to be made public in autumn 2005) will soon enable researchers to analyze metabolic and cellular pathways, other than those reported in KEGG, in several organisms including *Mus musculus* and *Homo sapiens*.

Several methods for assessing the significance of expression changes in diverse cellular pathways have been recently developed. GeneMerge [30] uses gene lists from KEGG, GO, MIPS or other sources, and rank scores for functional or categorical over-representation within the study set of genes and is obtained using the hypergeometric distribution. The program is extremely useful, as it extends the analysis to any favorite list of genes, the major limitations being the absence of any form of pathway visualization. Similar to GeneMerge, Gostat [31] also takes input from all genes on a microarray and automatically obtains the GO annotations from a database. The program automatically obtains the GO annotations from a database and generates statistics (Wilcoxon test and Kolmogorov-Smirnov statistics) of which annotations are overrepresented in the analyzed list of genes.

Pathway Miner [32] is another freely available tool that uses the Fisher exact test to rank genes that are defined as part of the same pathway, on the basis of their role in metabolic, cellular and regulatory pathways, or as groups pre-defined by the user. The genes are then mapped onto pathways with a graphical output similar to that of Cytoscape, and gene product association networks are extracted for genes that co-occur in pathways.

One of the major limitations of these methods is the application of the Fisher exact test and/or hypergeometric distribution to the analysis of highly interconnected pathways with a high level of redundancy. The use of statistics in the current approaches considers all the annotations as independent categories, which is clearly not the case. As an example, several gene lists may include entirely sublists of genes, and some genes could be part of more than one gene list. When using programs that test for enrichment in GO annotations, it is important to keep in mind that the GO annotations do not define a pathway, but are rather a classification based on existing knowledge, where a gene can be attributed to many different classes, and the ontology defines different hierarchical levels, that do not necessarily describe causality, directionality, or the type of interaction between the components of a given class. As a result, classes defined by the GO terms can be highly redundant. The fact that one gene can be contained in several categories particularly affects the application of hypergeometric statistics and makes it difficult to properly correct for multiple hypotheses. Therefore, the statistical values obtained should be considered merely as indicative, and lead to loss of important information. Our experience on the Diauxic shift dataset and KEGG indicates that the application of a standard Bonferroni correction is often too restrictive, and clearly discards a great deal of important biological information (D Cavalieri, unpublished).

This consideration highlights the need to tailor the tool to the source of the pathway information, and in particular to estimate the connection between different pathways. This connection can be the result of partial or total overlap, or from sharing of common reaction precursors or the potential flow of intermediate products into more than one pathway. An estimate of the connectivity between pathways could be used to apply multiple hypothesis correction to the statistics. Integration of expression information pathway-based analysis with Flux balance analysis can be converted to *in silico* regulatory models that can be further combined with genome-scale metabolic models to build integrated models of cellular function including both metabolism and its regulation [33].

On the other hand, it will be interesting to see the results of the application of Bayesian networks for representing statistical dependencies to the discovery of the interactions between genes in pathways and discover novel pathways, according to the same framework used to describe interactions between genes [34].

Methods that use functional genomics results, existing biological information, and clustering to reconstruct a novel biological network

One of the limits of pathway-based methods is that they do not provide information on genes of unknown function or the transcriptional regulatory networks. The integration of the results from pathway-based analyses with those of clustering algorithms could indicate which genes of unknown function cluster together with genes assigned to a given pathway, suggesting that their functions are metabolically related, and affording a new approach for attribution of functions to unknown genes.

GeneXpress is a recently developed tool that enables the user to combine cluster analysis with the analysis of biological attributes. Following cluster analysis of microarray data, GeneXPress tests what biological processes are represented in a given cluster, what *cis*-regulatory motifs are shared by genes within a cluster, and how significant these associations are, while allowing the possibility of performing a global analysis. This is a substantial improvement compared with other tools because statistical analysis of the clusters can associate the cluster to one or more biological processes, provide a p-value for the association and integrate the pathway results with the structure of the transcriptional regulatory networks, testing for enrichment in known transcription factor binding motifs. This tool has been recently applied to the collection of available expression data related to cancer progression, demonstrating an extremely promising mine of biological and clinical information [35].

Another recently developed program (CARRIE) uses promoter sequence and expression data to infer a regulatory network shown on an interactive graph [36]. The network data can be analyzed in the context of the cellular response visualizing the genes on KEGG metabolic pathways.

Conclusions

The advantage of analyzing a network of genes is that altered genes which co-vary even to a small extent can be considered significant, rather than focusing on the probability of the change of every single element. The hypothesis that genes in the same pathway are more likely to be coordinatedly regulated than a randomly selected gene set has been recently demonstrated using coherence indicators estimated in 96 pathways in tumour and normal samples [37].

Ideally, methods that analyze expression data according to a pathway-based logic should give: (1) an indication of the statistical significance of the conclusions; (2) provide a user-friendly interface for visualizing the results; (3) be able to encompass the largest number of well-established pathways; (4) define novel pathways; (5) assign unknown genes to a pathway; and (6) reconstruct the hierarchical structure of a group of pathways.

The efforts of the consortiums developing common exchange formats for biological pathways, annotations of all the biochemical reactions and databases of validated pathways are instrumental to the integration of biochemical function and gene expression. The improvement of pathway-based methods will require improved statistics, the connection of biological pathways with information on transcription factors, metabolite flow, networks of protein–protein interaction, and better visualization and graphical tools. Progress in this field will provide a crucial contribution to the application of microarrays to clinical studies and to systems biology.

Acknowledgements

We thank Philippe Rocca-Serra from EBI, for the useful discussions on Reactome; Rob Stierum and Marion Van Erk from TNO, and Chris Evelo from Maastricht University for the critical evaluation of GeneMapp; Andrea Splendiani for his precious insights on BioPAX; and Annibale Biggeri from Florence University for stimulating debate on the statistical approaches to data analysis. We acknowledge the EU networks of excellence NUGO and DC-Thera for promoting the discussion of fundamental issues on the analysis of genomics data.

References

1 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
2 Horimoto, K. and Toh, H. (2001) Statistical estimation of cluster boundaries in gene expressions profile data. *Bioinformatics* 17, 1143–1151
3 Toh, H. and Horimoto, K. (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* 18, 287–297
4 Kim, S. *et al.* (2000) Multivariate measurement of gene expression relationships. *Genomics* 67, 201–209
5 Tavazoie, S.J.D. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285
6 Alter, O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106
7 Wall, M.E. (2001) SVDMAN: singular value decomposition analysis of microarray data. *Bioinformatics* 17, 566–568
8 Lukashin, A.V. and Fuchs, R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* 17, 405–414
9 del Rio, G. *et al.* (2001) Mining DNA microarray data using a novel approach based on graph theory. *FEBS Lett.* 509, 230–234
10 Torkkola, K. *et al.* (2001) Self-organizing maps in mining gene expression data. *Inf. Sci.* 139, 79–96
11 Featherstone, D.E. and Broadie, K. (2002) Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioessays* 24, 267–274
12 Dougherty, E.R. *et al.* (2002) Inference from clustering with application to gene-expression microarrays. *J. Comput. Biol.* 9, 105–126
13 Hastie, T. *et al.* (2001) Supervised harvesting of expression trees.Genome Biol. 2, RESEARCH0003
14 Harris, M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261
15 Kanehisa, M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280
16 Karp, P.D. *et al.* (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* 28, 56–59
17 Krieger, C.J. *et al.* (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 32, D438–D442
18 Karp, P.D. *et al.* (2004) An evidence ontology for use in pathway/genome databases. *Pac. Symp. Biocomput.* 2004, 190–201

19 Mueller, L.A. *et al.* (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol.* 132, 453–460
20 Hou, B.K. *et al.* (2004) BioSilico: an integrated metabolic database system. *Bioinformatics* 20, 3270–3272
21 Robertson, M. (2004) Reactome: clear view of a starry sky. *Drug Discov. Today* 9, 684–685
22 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531
23 Ermolaeva, O. *et al.* (1998) Data management and analysis for gene expression arrays. *Nat. Genet.* 20, 19–23
24 Dwight, S.S. *et al.* (2004) Saccharomyces genome database: underlying principles and organisation. *Brief. Bioinform.* 5, 9–22
24 Dahlquist, K.D. *et al.* (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* 31, 19–20
25 Doniger, S.W. *et al.* (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* 4, R7
26 Shannon, P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504
27 Dennis, G., Jr *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4, P3
28 Grosu, P. *et al.* (2002) Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res.* 12, 1121–1126
29 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
30 Castillo-Davis, C.I. and Hartl, D.L. (2003) GeneMerge-post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* 19, 891–892
31 Beissbarth, T. and Speed, T.P. (2004) Gostat: find statistically overrepresented Gene Ontologie within a group of genes. *Bioinformatics* 20, 1464–1465
32 Pandey, R. *et al.* (2004) Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics* 20, 2156–2158
33 Herrgard, M.J. *et al.* (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol.* 15, 70–77
34 Friedman, N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620
35 Segal, E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* 36, 1090–1098
36 Haverty, P.M. *et al.* (2004) CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Res.* 32, W213–W216
37 Yang, H.H. *et al.* (2004) A computational approach to measuring coherence of gene expression in pathways. *Genomics* 84, 211–217