Consecutive KEGG pathway models for the interpretation of high-throughput genomics data

Alexey V. Antonov[1*], Sabine Dietmann[1], Hans W. Mewes[1,2]

[1] Helmholtz Center Munich, Institute for Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

[2]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany

[*]Corresponding author.

Abstract: A common strategy to deal with the interpretation of gene lists is to look for overrepresentation of Gene Ontology (GO) terms or pathways. In related computational approaches the cell is formalized as genes that are grouped into functional categories. As output, a list of interesting biological processes is provided, which seems to be mostly covered by the supplied gene list. However, it is more natural to model the cell as a network that reflects relations between genes. For many biological processes such information is available, but it is not used to the full extent in interpretational analyses. In this paper, we propose to interpret gene lists in network terms to provide the most probable scenario of gene interactions based on the available information about the topology of metabolic pathways. The proposed approach is an effort to exploit the biological information available in public resources to a greater extent in comparison to the existing techniques. Applying our approach to experimental data, we demonstrate that the currently widely employed strategy produces an incomplete interpretation, whilst our procedure provides deeper insights into possible molecular mechanisms behind the experimental data.

1. Introduction

In the post-genomic era the targets of many experimental studies are complex cell disorders. A standard experimental strategy is to compare the genetic signatures of the cells in normal and anomalous states. As a result, a set of genes, whose measured activity differs between considered cell states, is delivered. In the next step, an interpretation of the identified genes is required. A common bioinformatics strategy is to infer biological processes that are most relevant to the analyzed gene list. The inference is based on a prior knowledge about individual gene properties, such a molecular functions or biological processes.

In the standard bioinformatics framework, the cell is modeled as a set of genes that splits into known functional categories (Khatri et al., 2007; Antonov and Mewes, 2006; Khatri and Draghici, 2005; Subramanian et al., 2005; Berriz et al., 2003; Khatri et al., 2002). We will refer to this approach as "categorical". It is obvious, that this approach has a number of shortcomings. First, the categorical approach discards from consideration a lot of valuable information that is available in public databases. The relations between genes inside each category, like the pathway topology, as well as the relations between genes from different categories are not considered within the standard categorical framework. Second, the output of the categorical approach is a list of categories that are overrepresented among the analyzed genes. This is evidently helpful information; however, in most cases, this is not exactly what experimentalists are looking for. A basic premise in the application of high-throughput methodologies for studying molecular mechanisms of complex cell disorders is cooperative gene behavior. The change in the state of one (or

several) gene(s) leads to cooperative changes in the state of several dependent genes, and so on. Ideally, as an interpretational model of the gene list, the experimentalist would prefer to obtain a network model that proposes the most probable scenario of gene relations, which cover most of the genes from the supplied experimental list, i.e. gene A interacts with gene B, gene B interacts with gene C, gene C interacts with gene D, and genes A, B, C are from the same metabolic pathway, while genes C and D are regulatory genes. Thus, one seeks not only the information that the corresponding metabolic and regulatory pathways are enriched within a gene list, but also the way, in which the genes are interacting between and within the pathways.

Efforts have been made to overcome the first limitation of the categorical approach in order to take into account the pathway topology. Rahnenfuhrer et al., 2004 used, in addition to pathway categories, the distance between genes within the metabolic pathway. In this case, the impact of a pair of genes was weighted with respect to the distance between genes within a metabolic pathway. Another procedure, proposed recently by Draghici et al., 2007, exploited the hierarchical structure of signaling pathways and weighted the impact of genes with respect to their position in a pathway hierarchy. Genes at the top of signaling cascade received higher impacts in comparison to downstream genes. However, in both cases the second limitation has not been overcome, i.e. both approaches did not provide significant relationships between genes from different pathways. The output was still a list of categories that were enriched.

We propose a fundamentally different technique for the analysis of gene list referred to as the network-based approach (vs. categorical). The cell is modeled as set of genes that are connected into a global network. The input gene list is translated into a network model according to the global network, which reflects the most probable scenario of how genes affect the state of each other. As output, along with a list of enriched categories, our procedure provides a model of gene interactions that present a description of how different and apparently independent biological processes are interconnected. The statistical significance of the inferred network model is computed by a random simulation procedure. We demonstrate on several experimental data sets that our approach provides deeper insight into biological mechanisms that unites the supplied gene lists in comparison to currently available methods.

2. Network based approach

In general, an enrichment analysis is based on the available information about individual gene properties. In most cases, the experimental knowledge is formalized in a categorical format, as provided by several functional classification schemes (Mewes et al., 2004; Apweiler et al., 2001; Ashburner et al., 2000). Genes are subsequently assigned to the pre-defined classes. A straightforward way to use this information is to select those categories that have a statistically significant intersection with the analyzed gene list.

In some cases, like for metabolic processes, the experimental knowledge is stored in more complex forms to represent, for example, associations between genes and metabolites. This information can be easily converted into a pairwise distance between genes, and can be used to infer the optimal network model from a gene list. The distance between genes can be counted as the minimal number of consecutive steps required to get from one gene to another by working through existing paths on the global metabolic network. The inferred network model has several statistical properties which reflect the closeness of connected genes in the network. Based on the distribution of these properties for random gene lists one can estimate the statistical significance of the inferred model.

We used KEGG reference maps (Ogata et al., 1999) of metabolic pathways to generate pairwise distances between available genes. For each gene the set of associated compounds was defined. Genes and compounds are considered to be associated if they are assigned to the same reaction, e.g. a compound is either a substrate or product of the reaction and the gene is mapped to the enzyme that catalyses the reaction. In the same way, for each compound the set of associated genes is defined. A pair of genes is referred to as neighbors, if they have at least one common associated compound. While connecting neighbors via edges, we generate a global Gene Association Network (GAN) by integrating all available metabolic pathways. The distance between neighbors is set to “1” (one step to get from one gene to

another). The distance between two arbitrary genes is computed as a minimal number of steps required to get from one gene to another through available paths on the GAN.

Given a gene list, our purpose is to infer the network that minimizes the distance between each connected gene pair according to the GAN. To solve this problem, we propose to infer the network by a simple iterative procedure. In the first step, we connect by edges all gene pairs with distance 1. In the second step, isolated genes with distance 2 are connected. Genes are referred to as isolated, if there is no path in the network that connects them. Otherwise genes are referred to as connected. In the third step, isolated genes with distance 3 are connected. From our experience with experimental data a distance larger than 3 indicates that the statistical significance of the edge is low and that the genes can be considered independent. At each step (1, 2, 3) we look for connected sub-networks and identify the one with the maximal size (number of nodes or edges). The sub-network is referred to as connected, if it has only connected genes. The sub-network with maximal size is referred to as a maximal sub-network. We also refer to the size of maximal sub-network as the size of the inferred model. The model size is considered as a statistics, which is used to estimate the statistical significance of the model.

The statistical significance of the inferred model is estimated based on the distribution of the model size derived from random gene lists. The distribution is computed by a random simulation procedure (Westfall and Young, 1993). In the first step the random gene list of the size equal to the size of the input list is generated. The iterative network inference procedure described above is applied to the generated gene list. At each step (1, 2, 3) the size of the maximal sub-network is determined. By repeating the random procedure k times we get the background distribution for model size of random gene list and can estimate the statistical significance of the inferred network model up to the confidence level 1/k.

In addition to the genes from the supplied list, the inferred network model includes intermediate metabolites and genes. Intermediate genes are genes that, according to inferred model, connect genes from the list. If the distance between two genes from the list is 2 than they are not neighbors and connected via intermediate gene. Each pair of gene neighbors has a common metabolite (or several metabolites) used to connect them.

Known metabolic genes represent only 10 to 40 percent of genes from the whole genome (depending on the organism analyzed). For other genes there is no reliable information available about their network associations. Therefore, we propose to combine both approaches. Those genes from the analyzed list that are mapped to KEGG pathways are assessed by the network approach. In addition, standard enrichment analysis of GO categories (Ashburner et al., 2000) is performed. In the last step, both models are united in a final graphical representation. Significantly enriched GO terms that additionally have a statistically significant overlap with genes from the network model are selected and added to represent relationships between metabolic and other biological processes.

3. Results

We present several examples of data analyses by the network approach. We start with a simple illustrative example to demonstrate the advantages of the network approach in comparison to the categorical one. In the next step, we bring together two independent studies that performed experimental analyses to identify over- or underrepresented genes related to different biological problems. In each case, we collect the set of differentially expressed genes originally identified in each study and reanalyze them by the network approach.

3.1 Artificial data example

Let us consider an illustrative example to highlight the advantages of the network approach. Assume that as a result of some experiment one gets a list of nine metabolism-related genes, namely *ME3, MDH1, FH, ASL, ASS1, CTH, CDO1, CBS,* and *SHMT1*. Standard enrichment analysis will report several metabolic pathways as being enriched. Three genes (*CTH, SHTM1, CBS*) are mapped to "*glycine, serine and*

threonine metabolism". Two genes (*ASL, ASS1*) are mapped to "*urea cycle*" and two genes (*ME3, MDH1*) are mapped to "*citrate cycle*". No functional model that unites all 9 genes together would be supplied by any currently available analytical tool or approach.

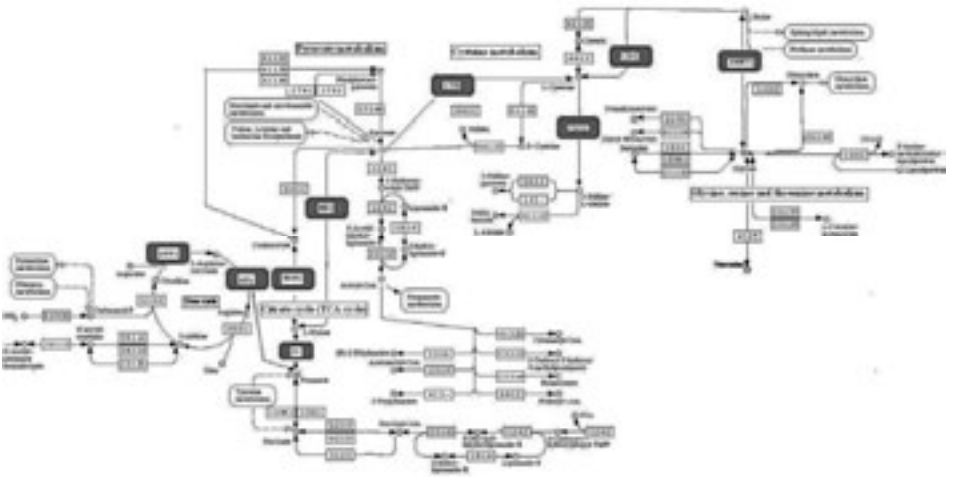

Figure 1: The genes *ME3, MDH1, FH, ASL, ASS1, CTH, CDO1, CBS, SHMT1* are presented as red boxes. Five KEGG pathway wiring diagrams ("*urea cycle*", "*citrate cycle*", "*pyruvate metabolism*", "*cysteine metabolism*", "*glycine, serine and threonine metabolism*") are linked together to demonstrate that all 9 genes are located on a consecutive metabolic path.

However, according to the KEGG pathway wiring diagrams, all 9 genes are consecutively connected via metabolites (Figure 1) and form a non-interrupted path, which runs through five canonical metabolic pathways ("*urea cycle*", "*citrate cycle*", "*pyruvate metabolism*", "*cysteine metabolism*", and "*glycine, serine and threonine metabolism*"). This illustrative example demonstrates that in many cases the knowledge of enriched individual pathways may be insufficient to get a complete understanding of the relation among genes from the supplied list. The consideration of the global gene metabolic network to interpret gene list as a network may be much more informative.

3.2. Analysis of long-lived *C. elegans daf-2* mutants using serial analysis of gene expression

Halaschek-Wiener *et al*., 2005 identified genes that are associated with longevity in a long-lived Caenorhabditis elegans daf-2 (insulin/IGF receptor) mutant using serial analysis of gene expression (SAGE). SAGE libraries were prepared from daf-2 worms at days 1, 6, and 10 of adulthood. The day 6 library represents gene expression in mid-adult life, whereas day 10 marks the latest time before the occurrence of dead animals in the population. To identify gene expression differences and metabolic changes that may lead to the increased life expectancy of *daf-2* adults, the *daf-2* and control worms at the same chronological age at day 6 were analyzed. SAGE libraries were screened for tags that had an abundance of at least 10 in one of the libraries and were differentially expressed by > 2.5-fold between *daf-2* and controls, with a P-value < 0.05. The number of selected genes was about 250 (Halaschek-Wiener *et al.*, 2003, Supplementary Data).

A standard enrichment analysis provides several GO terms that are overrepresented among the analyzed genes. Some terms were related to development and regulatory processes. Among the interesting biological processes, which may have direct links to molecular mechanisms that underlie longevity, one should

mention "*embryonic development ending in birth or egg hatching*", "*lipid transport*", and "*larval development*". Seventeen differentially expressed genes map to KEGG metabolic pathways. However, only the "*glycolysis pathway*" was identified as enriched (P-value ~ 0.05), 4 genes (*F01F1.12, GPD-4, T03F1.3* and *GPD-1*) out of 24 pathway-related genes were among those that were differentially expressed. Other 13 metabolism-related genes were not interpreted, as they represent a statistically insignificant share of genes from pathways they belong to.

In contrast, the application of our network approach reveals that 15 (out of 17) metabolism-related genes are connected into a network model with a distance between each gene pair not exceeding 2 (each pair of genes connected in the network is separated by a maximum of two metabolites). The network model runs through several canonical metabolic pathways, as presented in the overall graphical model in Figure 2. In addition, 6 genes from the inferred network model were also annotated as *"embryonic development ending in birth or egg hatching",* the GO term that was enriched among the 250 differentially expressed genes.

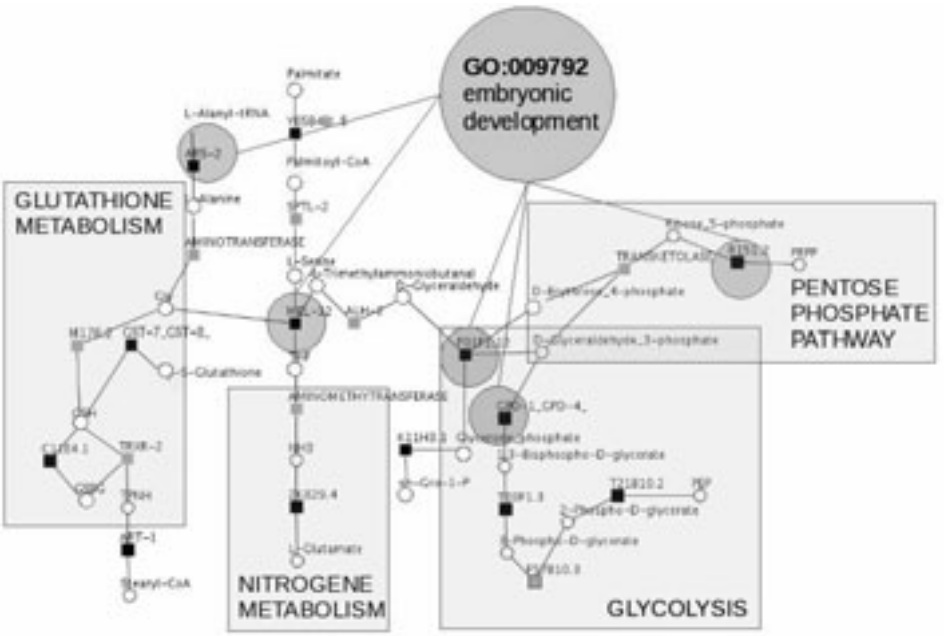

Figure 2: Network model of longevity-associated genes (Halaschek-Wiener et al., 2005) spanning four KEGG pathways. Differentially expressed genes are indicated by black rectangles, intermediate genes in the network model by brown rectangles, and chemical compounds by white circles. Genes that are involved in embryonic development (GO term: GO:009792) are highlighted by circles.

In total, 797 *C. elegans* genes can be mapped to KEGG pathways; and 154 of them are annotated with the GO term GO:009792 ("*embryonic development*"). Our network model covers 15 genes, and 6 are annotated with the term GO:009792. Based on the incidence of genes from the GO category GO:009792 among the KEGG genes (0.2 ~ 154/797) and the same incidence among network model genes (0.4 ~ 6/15), we can propose that the inferred network model has an overrepresentation of genes annotated as *"embryonic development ending in birth or egg hatching"* (P-value ~ 0.03, hypergeometric test). However, we would like to point out, that the correct estimation of the statistical significance in this case requires a non-trivial model which is beyond the scope of this paper.

The interpretational model supplied by the network approach is apparently more instructive in comparison to the categorical approach. The graphical representation of the inferred network allows one to track naturally the relation between metabolic and developmental processes.

To validate the statistical significance of the inferred network model, we computed a background distribution by a random simulation procedure. As described in the previous section, we generated randomly 1000 times the set of 17 genes from the set of C. elegans genes, which mapped to KEGG metabolic pathways. Each time we applied two steps of the proposed network inference procedure to the random set. As a result, all gene pairs from the randomly generated list with a distance equal to 2 (genes in the network model connected via 1 intermediate gene) and 1 (gene that relate to a common metabolite) were connected. Each time we computed the size (number of nodes) of the maximally connected sub-network. We considered these 1000 values as the background distribution for estimating the significance of the inferred network model. In total, 6 times the size of the network model inferred from the randomly generated gene list was greater or equal to 15. Therefore, the P-value of the inferred network model estimated by the random simulation procedure was less than 0.01 (6/1000). Figure 3 presents a plot of the generated background distribution.

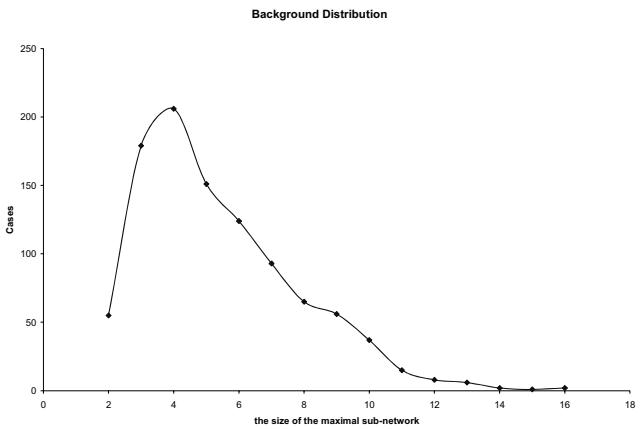

Figure 3: The distribution of the network model size generated by random simulation. The size (x-axis) of the network model is plotted against the number of times (y-axis) the network with this size was inferred during 1000 random simulations.

3.3. Pathway analysis of kidney cancer using proteomics and metabolic profiling

Perroud et al., 2006 performed a proteomic analysis of tumors to determine which pathways and processes are likely to be operative in renal cell carcinoma (RCC). By using 2-dimensional electrophoresis and mass spectrometric analysis, 31 proteins were identified to be differentially expressed in clear cell RCC as compared to adjacent non-malignant tissue. The standard categorical approach applied by the authors identified groups of genes and proteins which are organized into metabolic and signaling pathways relevant to the oncogenesis or progression of ccRCC. Several metabolic pathways closely associated with gluconeogenesis, such as "*pyruvate metabolism*", "*pentanoate metabolism*", "*butanoate metabolism*", as well as "*arginine and proline metabolism*" and the "*urea cycle*", were reported to be enriched among down-regulated genes in ccRCC. Similarly, the glycolysis pathway was identified as being significantly altered in ccRCC. In addition, a statistically significant alteration of the non-metabolic p53 signaling pathway was identified.

The authors of the paper suspect that the identified proteins from different enriched metabolic pathways are dependent. Indeed, 15 out of 16 proteins that were mapped to KEGG pathways form a statistically significant network model (P-value < 0.001). The inferred network contains 19 edges, 6 edges of length 1 and 13 edges of length 2. The models provided by the network approach are evidently more informative in comparison to the categorical one. For example, the authors report that 6 proteins (*HSP1, PKM2, GAPDH, LDHA, ANXA4 and ANXA5*) participate in the p53 signaling pathway. Three of these proteins are involved in the inferred network model. Using visualization capabilities of the network approach we can get an idea of how metabolic and signaling processes are linked in the altered cancer cells. Figure 4 presents a graphical visualization of the inferred model.

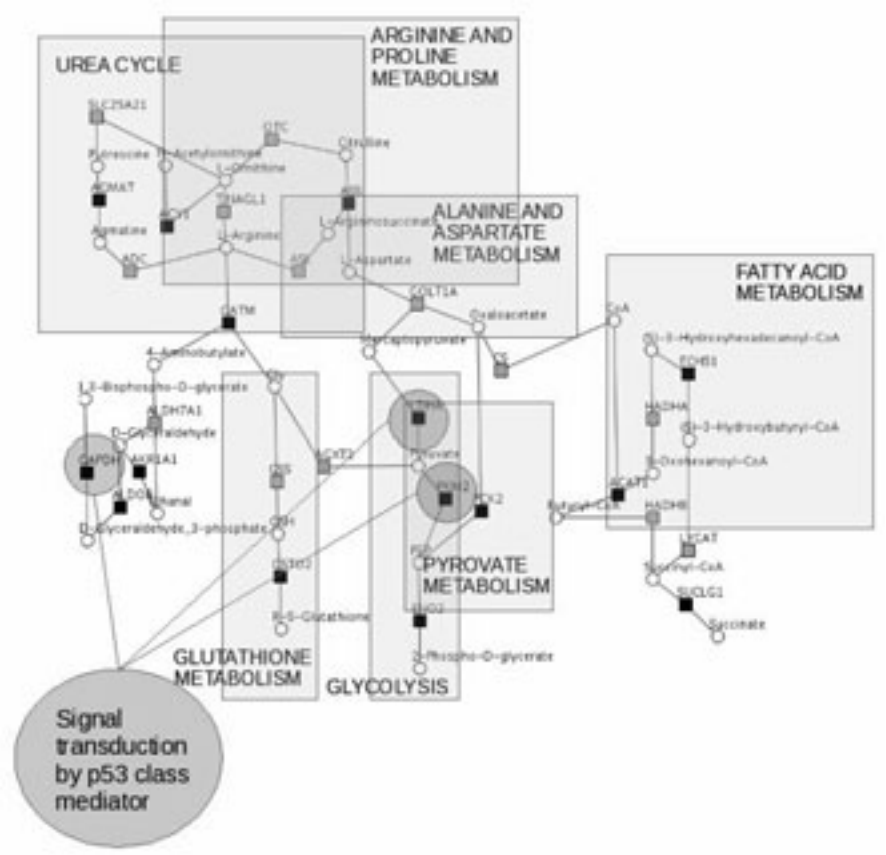

Figure 4: Network models of genes indicating a high risk of kidney cancer (Perroud et al., 2006). Differentially expressed genes are shown by black rectangles, intermediate genes by brown rectangles, and chemical compounds by white circles. Genes that are known to be involved in p53-mediated signaling are highlighted by circles.

The graphical representation of the inferred network is amenable to further analyses. In particular, it may be useful for decisions regarding potential therapy. Brief analysis of the network in Figure 4 identifies that it naturally splits into several sub-networks. Each of these sub-networks connects to the other nodes by single or double paths. For example, the sub-network, which is mostly related to fatty acid metabolism, is connected to the remaining nodes via the path ACAT1 – CoA – CS -- Oxaloacetate. Disruption of this path

may normalize fatty acid metabolism in cancer cells and, thus, reduce the potential of cancer cells to multiply. In the same way, targets for normalizing other metabolic processes may be selected. For example, to affect the urea cycle metabolism in cancer cells, at least two paths must be interrupted, i.e. the path ASS -- Laspartate -- COLT1A - Oxaloacetate and the path Larginine -- GATM.

4. Discussion

The importance of the development of network strategies for the analysis of biological systems was stressed in many studies (Lu et al., 2007; Chuang et al., 2007; Loscalzo et al., 2007; Ergun et al., 2007) . Here, we presented a network strategy for interpretational modeling of results of high-throughput genomics data. We demonstrated that the proposed procedure for translating gene lists into gene network models has a number of advantages in comparison to the widely used categorical approach. First, the coverage of the network model is higher in comparison to the categorical approach. As demonstrated, the network model usually covers a large fraction of genes that are mapped to metabolic pathways. For example, in the first case of C. elegans *daf-2* mutants among 250 selected genes 17 were mapped to metabolic pathways. The standard categorical approach was able to identify the enrichment model for only 4 of them from the glycolysis pathway. However, the network approach infers statistically valid model demonstrating that 15 genes are involved in close metabolic relation. Second, the output network model provides detailed information on pairwise gene relations among the analyzed genes. In the categorical approach this information is limited to the size of individual pathway.

At this stage, we used only metabolic pathway data for interpretational network modeling because this is one of the most reliable resources of genomics data available in network format for most model organisms. To take into account biological process other than metabolic, we combined the network approach with the standard categorical procedure. This allows for generating statistically valid hypotheses of how changes of metabolic processes interact with other non-metabolic biological processes mostly affected in the studied phenomena. However, there are no principle limitations to expand the network approach to comprise gene regulatory networks and protein interaction data of various natures. We consider extending the network approach with this kind of data in the nearest future.

References

Antonov,A.V. and Mewes,H.W. (2006) Complex functionality of gene groups identified from high-throughput data. *J. Mol. Biol.*, 363, 289-296.

Apweiler,R. et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29, 37-40.

Ashburner,M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25, 25-29.

Berriz,G.F. et al. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics.*, 19, 2502-2504.

Chuang,H.Y. et al. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3, 140.

Draghici,S. et al. (2007) A systems biology approach for pathway level analysis. *Genome Res.*, 17, 1537-1545.

Ergun,A. et al. (2007) A network biology approach to prostate cancer. *Mol. Syst. Biol.*, 3, 82.

Halaschek-Wiener,J. et al. (2005) Analysis of long-lived C. elegans daf-2 mutants using serial analysis of gene expression. *Genome Res.*, 15, 603-615.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics.*, 21, 3587-3595.

Khatri,P. et al. (2002) Profiling gene expression using onto-express. *Genomics*, 79, 266-270.

Khatri,P. et al. (2007) Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res.*, 35, W206-W211.

Loscalzo,J., Kohane,I. and Barabasi,A.L. (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol. Syst. Biol.*, 3, 124.

Lu,X. et al. (2007) Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol. Syst. Biol.*, 3, 98.

Mewes,H.W. et al. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, 32, D41-D44.

Ogata,H. et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27, 29-34.

Perroud,B. et al. (2006) Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Mol. Cancer*, 5, 64.

Rahnenfuhrer,J. et al. (2004) Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol.*, 3, Article16.

Subramanian,A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A*, 102, 15545-15550.

Westfall,P.N. and Young,S.S. (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value adjustment. John Wiley&Sons, In.,New York.