Pattern recognition of gene expression data on biochemical networks with simple wavelet transforms

Gunnar Schramm[1,2], Marcus Oswald[3], Hanna Seitz[3], Sebastian Sager[4], Marc Zapatka[2], Gerhard Reinelt[3], Roland Eils[1,2] and Rainer König[1,2]
[1] Department of Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular Biotechnology, University of Heidelberg; [2] Theoretical Bioinformatics, German Cancer Research Center (DKFZ); [3] Institute of Computer Science, University of Heidelberg; [4] Interdisciplinary Center for Scientific Computing, University of Heidelberg
69120 Heidelberg, Germany
g.schramm@dkfz.de, Marcus.Oswald@Informatik.Uni-Heidelberg.de, Hanna.Seitz@Informatik.Uni-Heidelberg.de, Sebastian.Sager@iwr.uni-heidelberg.de, m.zapatka@dkfz.de, Gerhard.Reinelt@Informatik.Uni-Heidelberg.de, r.eils@dkfz.de, r.koenig@dkfz.de

Abstract

Biological networks show a rather complex, scale-free topology consisting of few highly connected (hubs) and many low connected (peripheric and concatenating) nodes. Furthermore, they contain regions of rather high connectivity, as in e.g. metabolic pathways. To analyse data for an entire network consisting of several thousands of nodes and vertices is not manageable. This inspired us to divide the network into functionally coherent sub-graphs and analysing the data that correspond to each of these sub-graphs individually. We separated the network in a two-fold way: 1. clustering approach: sub-graphs were defined by higher connected regions using a clustering procedure on the network; and 2. connected edge approach: paths of concatenated edges connecting striking combinations of the data were selected and taken as sub-graphs for further analysis. As experimental data we used gene expression data of the bacterium *Escherichia coli* which was exposed to two distinctive environments: oxygen rich and oxygen deprived. We mapped the data onto the corresponding biochemical network and extracted disciminating features using Haar wavelet transforms for both strategies. In comparison to standard methods, our approaches yielded a much more consistent image of the changed regulation in the cells. In general, our concept may be transferred to network analyses on any interaction data, when data for two comparable states of the associated nodes are made available.

1 Introduction

Modern high throughput methods in biotechnology allow to profile cellular mechanisms not only for a single, focused, but for rather large numbers of aspects and compounds. This is especially given by the DNA microarray technology. It allow us to explore the expression levels for a major subset or all genes of an organism under a variety of conditions such as alternative treatments, mutants, developmental stages and time points. For example, the technique enables us to classify tumour samples [Van 'T Veer, *et al.*, 2002], to define small sets of potential marker genes to distinguish leukemias [Stephanopoulos, *et al.*, 2002], and to discover regulatory mechanisms [Gasch, *et al.*, 2000, Spellman, *et al.*, 1998]. In general, such studies use supervised methods to support diagnosis [Stephanopoulos *et al,* 2002, Van 'T Veer *et al*., , 2002] and yield gene lists that are crucial for the classifications [Thuerigen, *et al.*, 2006]. However, lists of single genes are rather tedious to analyse for yielding a general functional meaning. Therefore, methods were developed that map these gene lists onto functionally relevant reaction and signaling cascades [Manoli, *et al.*, 2006]. Furthermore, regulatory networks could be defined with co-expressed genes. E.g., without prior information, the structure and function of the network that regulates the SOS pathway in *E. coli* could be elucidated with transcription profiles [Gardner, *et al.*, 2003]. Another way to easier extract the functionality of expression patterns is to first map the data onto networks that consist of nodes bearing the expression data and vertices that link nodes with common functionality and directed or undirected interdependence. E.g. physically interacting proteins may be related functionally by being succeeding nodes in a signaling cascade. Such knowledge of protein-protein interaction from high-throughput techniques [Uetz, *et al.*, 2000] was applied to analyse gene expression data and revealed novel regulatory circuits [Ideker, *et al.*, 2002]. On a statistical basis, this data is useful for inferring changed signal transduction of e.g. diseased situations. Besides this, over the last four decades, biochemical investigations have discovered an increasingly consistent image of the cellular metabolism (see e.g. [Berg, *et al.*, 2002]). These biochemical reactions can be functionally linked together by setting a reaction r_i as the precursor of a reaction r_j if and only if one of the reaction-products of r_i is needed as a substrate for r_j. This yields a biochemical or metabolic network. Microarray expression data for each gene can then be mapped on its corresponding enzyme and the reaction the enzyme it is catalysing. Such interaction knowledge from the biochemical network has been used to support the clustering procedure for gene expression profiles of yeast [Hanisch, *et al.*, 2002, Zien, *et al.*, 2000]. Pattern analyses on such networks advantage from

the fact that such interactions are well defined and established. This is especially true for less complex organisms such as yeast or *Escherichia coli* [Karp, *et al.*, 2002]. Simple clustering of gene expression data on these metabolic networks can yield sub-graphs that are either commonly stimulated or repressed as we showed previously for tryptophan treated cells [König and Eils, 2004]. With this method we were able to find the biosynthesis pathway of tryptophan as an expression pattern in the network having a common response to the environmental changes. Note that such a clustering method discovers patterns of co-expressed genes. We now developed methods enabling the discovery of more complex patterns. As a case study, we investigated the response of the hetero-fermentative bacterium *E. coli* in response to oxygen deprivation. The regulatory machinery can react on this environmental change in different ways. One basic response changes the catabolism of glucose, switching off or down-regulating the respiratory sub-graphs such as the glyoxylate cycle and switching on the fermentation and production of acid end products (see e.g. [Neidhardt, 1996]). This is supported by several signalling concepts, e.g. by inducing inhibitors for glyoxylate cycle (TCA) genes, down-regulating glyoxylate cycle genes or activating and up-regulating genes for the fermentation processes.
Within the first approach (clustering approach), we performed a clustering of the network without gene expression data to define regions with high connectivity. Gene expression data was mapped onto these regions. We wanted to explore all possible relevant expression level combinations in these regions. For this, we used the adjacency matrix representations for these regions, mapped the data onto them and calculated simple Haar wavelet transforms applying a standard procedure for two-dimensional images (see e.g. [Theodoridis and Koutroumbas, 1998]). The extracted features were ranked due to their ability to distinguish between the two environmental conditions (oxygen rich versus oxygen deprived). In so doing, we were able to reveal interesting switches that are posted at process bifurcations, in rather good agreement to the expected anaerobic response. However, with this approach we were only able to extract interesting data patterns in highly connected sub-graphs. Therefore, we applied a second method to reveal crucial data patterns also of linearly ordered reactions. Features were generated by applying the one dimensional Haar-wavelet transform onto each pair of nodes. With this method we were able to detect expected up-regulated pathways of formate fermentation and C6 nutrients metabolism. Furthermore, our method revealed a down-regulation of the iron processing parts of the metabolic network as well as the up-regulation of the histidine biosynthesis pathway which constitutes a response for enriched acidic products during anaerobic growth.

2 Methods

The description of the clustering method will be briefly described here. For a description in detail, see [König, *et al.*, 2006, Schramm, *et al, in prep.*]. Metabolic reactions were extracted from the EcoCyc database (Version 9, [Keseler, *et al.*, 2005]). A graph was established by defining neighbours of metabolites. Two metabolites were neighbours if and only if an enzymatic reaction existed that needed one of the metabolites as input (needed substrate) and produced the other as output (product). Note, that in this representation, enzymes are edges and metabolites the nodes. This network was clustered to group enzymes into parts of the network with their major connections (the clustering algorithm is described in [König *et al.*, 2006]). The clustering algorithm produced a symmetrical sub-matrix of the cluster matrix for each cluster, whose rows and columns were the metabolites. The matrix contained a "1" entry at position (i, j) if an enzyme existed that combined metabolites of row i and column j. Otherwise a "0" entry was set.

2.1 Mapping gene expression data onto the cluster-matrices

For our case study, we collected raw intensity values of gene expression data from the work of Covert et al. [Covert, *et al.*, 2004]. We normalised them with an established variance normalisation method [Huber, *et al.*, 2002] and selected the data for 43 hybridisations of the following samples: strain K-12 MG 1655, wild-type, Δ*arcA*, Δ*appY*, Δ*fnr*, Δ*oxyR*, Δ*soxS* single mutants and the Δ*arcA*Δ*fnr* double mutant. The mutated genes are key transcriptional regulators of the oxygen response [Covert *et al*., 2004]. They effect a major portion of all genes in *E. coli* and therefore supported a variance stimulation of the respiratory and fermentative control of the investigated strain. All gene expression experiments were done in triplicate under aerobic and anaerobic conditions, respectively, except for anaerobic wild-type which was repeated four times. The gene expression data of each data-set was mapped onto the corresponding reactions of the transcribed proteins. Mean values were taken if a reaction was catalysed by a complex of proteins. The expression data of all samples was mapped onto each cluster-matrix, yielding 43 different patterns for each cluster.

2.2 Pattern discovery: defining the features with the Haar wavelet transform

We wanted to calculate a value for every possible expression pattern of neighbouring genes and groups of genes within a cluster that may show essential differences between samples of different conditions. Therefore, we performed a Haar-wavelet transform for each cluster-matrix. The wavelet transformed expression values served as features for the classifier (classification method, see next section). This allowed the identification of regions with a varying pattern between aerobic and anaerobic conditions. The wavelet-transformation is described in the following. Each cluster-matrix was divided into 2x2 pixeled disjoint sub-sections (e.g. a cluster matrix of size 8 x 8 was divided into 16 sub-sections). Clusters with non-fitting sizes (e.g. 3x3, 5x5, ...) were extended with rows and columns of zeros to yield matrices that could be divided into 2x2 pixeled sub-sections. For each sub-section, all combinations of row-wise and column-wise means and differences, respectively, were calculated. This yielded 4 combined values for each 2x2 pixeled sub-section: 1st: mean of the mean of the upper and mean of the lower row, 2nd: difference of the mean of the upper and the mean of the lower row, 3rd: mean of the difference of the upper and the difference of the lower row, and, 4th: difference of the difference of the upper and the difference of the lower row. All four combined values for each 2x2 pixeled sub-

section were stored and applied as features for the classifier. This was done for all sub-sections of the matrix. All 1st combined values (mean of means) were taken for a new matrix and were again grouped into 2x2 fractions that were combined in the same manner, yielding again 4 new features for every fraction. This procedure was repeated until no further grouping was possible. Such a "Haar" wavelet transform can be regarded as a low pass filter when calculating the mean, and a high pass filter when calculating the difference between neighbouring value pairs. The transform applied a filter in horizontal and subsequently in vertical direction. The procedure consisted of repeatedly applying high and low pass filters on the image. Therefore, either high frequency or low frequency portions of the signal were calculated and stored, until the maximal possible compositions were obtained. This procedure was carried out for all clusters of every sample and the results of the transforms were stored as the corresponding features for every sample.

2.3 Extracting essential features and their sub-graphs with the classifier

The SAM method [Tusher*, et al.*, 2001] as a modified t-test was performed to rank the features according to their p-values. Higher ranking features (low p-values) were selected focusing the classifier on the most relevant patterns (9,996 out of 70,912). For classification, we applied the Support Vector Machine implementation as provided by the R MCRestimate package [Ruschhaupt*, et al.*, 2004]. To receive a suitable feature extraction result, a 10-fold cross validation was performed and repeated 10 times with different splittings of the data, respectively. A linear kernel was applied for the feature extraction as described elsewhere [Ruschhaupt *et al*., 2004]. Parameter optimisation was performed for the regularisation term that defined the costs for false classifications (9 steps, range: 2^n, n = -4, -2, ..., 8, 10). This optimisation was realised by an internal three-fold cross validation during every iteration. To determine the most relevant features, a recursive feature elimination [Ruschhaupt *et al*., 2004] was applied during the parameter optimisation procedure. This yielded a set of discriminating features for every run. These features were ranked due to their selection frequency of all 100 runs. Note, that high-ranking features yielded the corresponding sub-graphs (cluster of the cluster matrix) of the reaction network that contained well discriminating patterns of the expression data. We defined a cut-off criterion for selecting only substantial features by comparing the selection frequency of each feature with random selections. We assumed a binomial distribution, neglecting the cases that the same feature may have been chosen twice in one run. The overall number of drawings was the sum of all selections (8,191 selections). The probability to draw the respective feature was the reciprocal value of the number of all features (1/9,996). The number of drawings for the respective feature was its selection frequency. As we calculated this for every feature, the resulting p-values where corrected for multiple testing by multiplying them with the number of all features (Bonferroni correction [Bonferroni, 1935a]).

2.4 Generating and assembling the features for the second approach

The Haar-wavelet was used to extract discriminating reaction pairs. We added and subtracted the values of neighbouring pairs of nodes yielding low pass and high pass filtered features for each pair, respectively. All generated features were ranked via a multiple t-test between aerobic and anaerobic conditions: to correct for potential influences coming from individual mutants, t-tests were performed for every constellation of samples excluding the sample of one particular mutant, respectively. The wild type sample was never excluded. From this outcome the worst (highest) p-value for each feature was selected. All p-values were corrected for multiple testing (Bonferroni [Bonferroni, 1935b]). Features were then ranked according to their p-value. Sub-graphs were put up by connecting found significant features (reaction-pairs) having a p-value ≤ 0.01. This resulted in 5 sub-graphs. To facilitate the interpretation of the found sub-graphs, nodes with equal expression behaviour (up-, down-regulation) were grouped together, and functionally described (see Results) if the group size was ≥ 5 focusing only on larger patterns. In total 10 such clusters were found. Reaction-pairs having one up- and one down-regulated node were regarded as switches. They were extracted if their p-value was ≤ 0.01 and are also functionally characterised (see Results).

3 Results

We will give a brief description of the functional meanings for the found sub-graphs focussing on integrative aspects (for details see [König *et al*., 2006, Schramm *et al*., *in prep*.]). From the clustering approach, we yielded 973 (not necessarily disjoint) clusters of sizes between 2 and 46 reactions and 160,264 features. After deleting features that consisted only of zeros, 70,912 features remained. A modified t-test was performed [Tusher et al., 2001] to reduce the remaining features and focus the classifier on the most relevant patterns. As a threshold, a false discovery rate of 2e-05 was chosen to further analyse the 9,996 most significant features. With these features, the SVM was trained and tested by a ten-time's ten-fold cross-validation. A recursive feature elimination [Ruschhaupt *et al.,* 2004] was applied for each run, yielding 100 lists of the most discriminating features. These features were ranked according to their selection frequency. 8,191 out of 9,996 features were selected at least once. To help us focusing on the most relevant features, only features with a significant selection frequency were used (p-value ≤ 0.05, in comparison to a random selection, Bonferroni corrected for multiple testing [Bonferroni, 1935a]). This yielded 181 features. Network clusters that contained these features were extracted and are further referred to as extracted clusters. They were listed in accordance to their selection frequency. Extracted clusters that contained less than six nodes were not considered to focus on larger patterns. Reactions were regarded as up-regulated (green in figures) if the corresponding genes were significantly up-regulated under anaerobic conditions (p-value ≤ 0.05 of a t-test), down-regulated if significantly down-regulated (red in figures), and not significantly differentially regulated otherwise (grey in figures, red/green frames indicate a non-significant tendency). Note that not differentially expressed nodes were discarded.

With the second approach (connected edge approach) we yielded 660 significantly discriminating reaction-pairs

(applied p-value cut-off: 0.01). Features that occurred twice due to the generation method were considered only once. All significant reaction-pairs were mapped onto the complete metabolic graph to extract sub-graphs consisting of connected pairs. In total, 5 such sub-graphs were identified consisting of 165 reactions.
Neighbouring, connected nodes, that showed identical regulation (up or down) were grouped together. Only groups of a minimum of five reactions were selected for functional interpretation focussing on major regulation patterns. We will refer to this groups as clusters in the following. Furthermore all significant switches were extracted. As switches were deemed pairs of reactions. We yielded 20 significant switches which were defined as reaction-pairs that showed opposing regulatory behaviour.

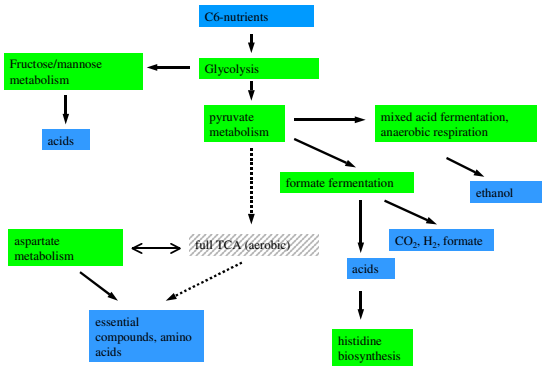

Figure 1. Carbohydrate metabolism and stress response to acids, up-regulated sub-graphs are green (filled), compounds are blue (dark). The TCA cycle (grey, fasciated) has got limited function for oxygen deprived conditions (see text).

3.1 Functional characterisation

Basically, the clustering approach found six clusters with crucial patterns in the following metabolic pathways: 1. formate fermentation, 2. Aspartate metabolism, 3. Lysine biosynthesis and C6 nutrients uptake, 4. Glycolysis and glucose storing, and 5. Glycolysis and NAD switch, 6. branched chain amino acid transporters. The connected edge approach found ten clusters: 1. formate fermentation and anaerobic electron transport chain, 2. Pyruvate metabolism, anaerobic synthesis of deoxyribonucleotides and electron transport, 3. C6 nutrients metabolism, i.e. glycolysis and fructose/mannose metabolism, 4. C6 nutrients metabolism: Glycolysis and Entner-Doudoroff pathway, 5. aerobic iron processing and transport, 6. aerobic iron processing: FE-S biogenesis, 7. histidine biosynthesis, aspartate metabolism and NAD switch, 8. processing of guanine nucleosides, 9. processing of uracil nucleosides, and 10. C1-processing changes and glutathione synthesis. Finally, we analysed the significant switches. They could be functionally combined yielding seven groups: 1. formate fermentation, 2. mixed acid fermentation, 3. C1 processing changes, 4. C6 nutrients, 5. branched chain amino acids transporters, 6. and 7. These groups consisted of two pairs each with ambiguous functionality and are not discussed here (for details, see *[*Schramm *et al, in prep*.].
The results are sketched in Figure 1. Under aerobic conditions, glycolysis and the TCA cycle are the major producers of energy. The TCA cycle depends heavily on oxygen and is therefore of limited use for anaerobic conditions. To keep the energy production going, glycolysis is up-regulated and similarily the fructose and mannose metabolism (C6 nutrients up-take pathways). The pyruvate metabolism switches the compound flow from the TCA cycle to acids and ethanol production. The formate fermentation degrades the acid formate or expells it into the outside of the cell. Additionally, the TCA normally supplies essential compounds for e.g. producing amino acids.For oxygen deprived conditions, this is taken over by the up-regulated aspartate metabolism. The cell responses to the higher concentration of acids by up-regulating histidine biosynthesis to enable an induced buffering by histidine. Under oxygen rich conditions, oxygen is causing oxidative stress. During anaerobic conditions, oxygen is reduced and therefore the oxidative stress response is down-regulated, i.e. the production of Fe-S and glutathione reduction (Figure 2). NAD switch: even though NAD may be more constitutively produced, up-regulation of quinolate synthetases which are the starting point of the NAD biosynthesis makes sense, as it could be shown that quinolate synthetases become inactive when exposed to oxygen [Ollagnier-De Choudens*, et al.*, 2005] and NAD may be primarily produced via the tryptophan biosynthesis pathway under aerobic conditions. Oxygen limitation limits energy production and therefore reduces producing energy intensive nucleosides. Reduced energy supply may also explain a switching of the branched chain amino acid transporters from ATP-dependend ABC-transporters to sodium-gradient dependend transporters. The C1 processing changes are according to a down-regulated glycin cleavage system and may be due to the fact that the reaction involved reduces NAD^+ to NADH, an oxygen costly reaction [Madigan*, et al.*, 2003]. The production of the one-carbon units, for which the glycine cleavage system is used [Stauffer, 1987], was taken over by serine hydroxymethyltransferase.

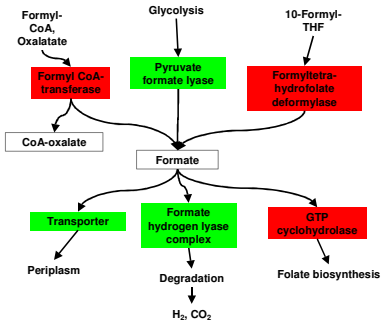

Figure 2. Formate fermentation. Green boxes indicate significant up-regulation (p-value ≤ 0.05) under anaerobic conditions. Red boxes indicate significant down-regulation. Glucose is catabolised into pyruvate. Under anaerobic conditions, pyruvate is degraded to formate which is either expelled, or further degraded into H_2 and CO_2. The reactions for these processes were up-regulated whereas the biosynthesis and degradation of costly compounds were down-regulated (folate and 10-formyl-THF, respectively).

3.2 Comparison to a standard method

To compare the findings of the method described here to a standard method for analysing gene expression data, a t-test was run on the gene expression levels for the corresponding reactions (without any network information). Extracted features were ranked due to the calculated p-value. The first 40 highest ranking reactions were: 1. formate hydrogenlyase complex, 2. FocA formate FNT transporter, 3. pyruvate formate-lyase, 4. aminomethyltransferase, 5. gcv system, 6. 3-methyl-2-oxobutanoate hydroxymethyltransferase, 7. glycine dehydrogenase (decarboxylating), 8. PFL-deactivase, 9. acetaldehyde dehydrogenase, 10. pyruvate kinase, 11. fumarate reductase, 12. enolase, 13. N-acetylmuramyl-L-alanine amidase, 14. formate dehydrogenase, 15. glutamate dehydrogenase (NADP+), 16. mannonate dehydratase, 17.+18. pyruvate formate-lyase activating enzyme (2x), 19. triose phosphate isomerase, 20. glutamyl-tRNA reductase, 21. histidine-phosphate aminotransferase, 22. 2-keto-4-hydroxyglutarate aldolase, 23. 2-keto-3-deoxy-6-phosphogluconate aldolase, 24. oxaloacetate decarboxylase, 25. putative NAD+ kinase, 26. 6-phosphofructokinase-1, 27. mannose-6-phosphate isomerase, 28. Outer Membrane Ferrichrome Transport System, 29. NADH oxidoreductase, 30. isocitrate dehydrogenase kinase, 31. isocitrate dehydrogenase phosphatase, 32. RhtB homoserine Rht Transporter, 33. histidinol-phosphatase, 34. imidazoleglycerol-phosphate dehydratase, 35. Outer Membrane Ferric Enterobactin Transport System, 36. phosphoenolpyruvate carboxylase, 37. tetrahydrodipicolinate succinylase, 38. imidazole glycerol phosphate synthase, 39. 3-hydroxy acid dehydrogenase, and 40. branched chain amino acids ABC transporter. At the top are three reactions involved in fermentation of formate that were also found with our method. Six reactions (10, 13, 15, 25, 29, 32) were not extracted by our method. Five of these reactions were not found due to the network creation method. Unspecific metabolites were deleted resulting in the deletion of reactions that catalyse such unspecific metabolites, such that pyruvate kinase, glutamate dehydrogenase (NADP+), NAD kinase, NADH oxidoreductase and RhtB homoserine Rht transporter were not included into the metabolic network. Putative reactions with not defined metabolites like N-acetyl-anhydromuramyl-L-alanine-amidase, the sixth not found reaction, were also not included into the metabolic network and could therefore not be found. With this caculated list from the standard method, we could not get any reactions for the iron processing response. Furthermore, the interesting histidine pathway was entirely found by our connected edge method. In contrast, with the standard method we found four out of ten reactions which are scattered in the list (21, 33, 34, 38) making it rather difficult to infer a combined regulation of the defined histidine pathway.

4 Conclusions

The methods described here facilitate the extraction of interesting and complex sub-graphs within a metabolic network by applying image-processing methods onto gene expression data. It suits well for less complex organisms like *E. coli*, for which the metabolic network is well established and reaction levels can be better estimated from the gene expression levels. In our case study several interesting and essential sub-graphs with differential expression patterns for *E. coli* when exposed to oxygen deprived conditions were identified like the fermentation of formate, processing of C6 nutrients, biosynthesis of histidine and iron metabolism. Thus a huge variety of anaerobic responses were discovered ranging from fermentation to energy and iron metabolism to acidic buffering. This covered not only direct regulations but also patterns originating from more complex environmental influences following the adaptation to oxygen deprivation like the response to excreted acids and thus the change in pH. Nevertheless, essential sub-graphs are not detected isolated but might interfere with related or connected pathways depending on the metabolites. The cluster containing histidine biosynthesis consisted also of parts of the aspartate and glutamine metabolism. This is due to the unspecific hub-like character of some metabolites connecting a huge variety of pathways. However, to give the found sub-graphs and reaction chains functional meaning was rather tedious and time consuming. We had to scan the appropriate literature and extract the specific information in a very detailed and long lasting procedure. We see an automated processing of this as a major task for the future.

Acknowledgments

We thank EcoCyc, Covert and his co-workers, and the ASAP team for making their data online available. The work was funded by the German National Genome Research Network (NGFN 01 GR 0450) and the Deutsche Forschungsgemeinschaft (Optimization-based control of chemical processes BO 864/10).

References

J.M. Berg, Tymoczko J.L., Stryer L.: *Biochemistry.* Fifth Edition edn. New York: W. H. Freeman; 2002.

C. E. Bonferroni: *Il Calcolo Delle Assicurazioni Su Gruppi Di Test.* In *Studi in Onore Del Professore Salvatore Ortu Carboni.* Rome, Italy; 1935a: 13-60

C. E. Bonferroni: *Il Calcolo Delle Assicurazioni Su Gruppi Di Teste.* In *Studi in Onore Del Professore Salvatore Ortu Carboni.* Rome; 1935b: 13-60

M. W. Covert, Knight E. M., Reed J. L., Herrgard M. J., Palsson B. O. *Integrating High-Throughput and Computational Data Elucidates Bacterial Networks***.** Nature, 429:92-96, 2004.

T. S. Gardner, Di Bernardo D., Lorenz D., Collins J. J. *Inferring Genetic Networks and Identifying Compound Mode of Action Via Expression Profiling***.** Science, 301:102-105, 2003.

A. P. Gasch, Spellman P. T., Kao C. M., Carmel-Harel O., Eisen M. B., Storz G., Botstein D., Brown P. O. *Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes***.** Mol Biol Cell, 11:4241-4257, 2000.

D. Hanisch, Zien A., Zimmer R., Lengauer T. *Co-Clustering of Biological Networks and Gene Expression Data***.** Bioinformatics, 18 Suppl 1:S145-154, 2002.

W. Huber, Von Heydebreck A., Sultmann H., Poustka A., Vingron M. *Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression***.** Bioinformatics, 18 Suppl 1:S96-104, 2002.

T. Ideker, Ozier O., Schwikowski B., Siegel A. F. *Discovering Regulatory and Signalling Circuits in Molecular Interaction Networks***.** Bioinformatics, 18 Suppl 1:S233-240, 2002.

P. D. Karp, Riley M., Paley S. M., Pellegrini-Toole A. *The Metacyc Database***.** Nucleic Acids Res, 30:59-61, 2002.

I. M. Keseler, Collado-Vides J., Gama-Castro S., Ingraham J., Paley S., Paulsen I. T., Peralta-Gil M., Karp P. D. *Ecocyc: A Comprehensive Database Resource for Escherichia Coli***.** Nucleic Acids Res, 33:D334-337, 2005.

R. König, Eils R. *Gene Expression Analysis on Biochemical Networks Using the Potts Spin Model***.** Bioinformatics, 20:1500-1505, 2004.

R. König, Schramm G., Oswald M., Seitz H., Sager S., Zapatka M., Reinelt G., Eils R. *Discovering Functional Gene Expression Patterns in the Metabolic Network of Escherichia Coli with Wavelets Transforms***.** BMC Bioinformatics, 7:119, 2006.

T. M. Madigan, Martinko J. M., Parker J.: *Biology of Microorganisms.* 10th edn: Prentice Hall; 2003.

T. Manoli, Gretz N., Grone H. J., Kenzelmann M., Eils R., Brors B. *Group Testing for Pathway Analysis Improves Comparability of Different Microarray Data Sets***.** Bioinformatics, 2006.

F.C. Neidhardt: *Escherichia Coli and Salmonella: Cellular and Molecular Biology.* Washington D.C.: American Society for Microbiology; 1996.

S. Ollagnier-De Choudens, Loiseau L., Sanakis Y., Barras F., Fontecave M. *Quinolinate Synthetase, an Iron-Sulfur Enzyme in Nad Biosynthesis***.** FEBS Lett, 579:3737-3743, 2005.

M. Ruschhaupt, Huber W., Poustka A., Mansmann U. *A Compendium to Ensure Computational Reproducibility in High-Dimensional Classification Tasks***.** Stat Appl Genetics Mol Biol, 3:37, 2004.

G. Schramm, Eils R., König R. *E. Coli's Crucial Switches, Pathways and Clusters of Gene Expression During Oxygen Deprivation***.** in preparation,

P. T. Spellman, Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D., Futcher B. *Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization***.** Mol Biol Cell, 9:3273-3297, 1998.

G. V. Stauffer: *Biosynthesis of Serine and Glycine.* In *Escherichia Coli and Salmonella Typhimurium Cellular and Molecular Biology. Volume* 1. Edited by F. C. Neidhardt: American Society for Microbiology; 1987: 412-418

G. Stephanopoulos, Hwang D., Schmitt W. A., Misra J. *Mapping Physiological States from Microarray Expression Measurements***.** Bioinformatics, 18:1054-1063, 2002.

S. Theodoridis, Koutroumbas K.: *Pattern Recognition.* London: Academic Press; 1998.

O. Thuerigen, Schneeweiss A., Toedt G., Warnat P., Hahn M., Kramer H., Brors B., Rudlowski C., Benner A., Schuetz F., Tews B., Eils R., Sinn H. P., Sohn C., Lichter P. *Gene Expression Signature Predicting Pathologic Complete Response with Gemcitabine, Epirubicin, and Docetaxel in Primary Breast Cancer***.** J Clin Oncol, 24:1839-1845, 2006.

V. G. Tusher, Tibshirani R., Chu G. *Significance Analysis of Microarrays Applied to the Ionizing Radiation Response***.** Proc Natl Acad Sci U S A, 98:5116-5121, 2001.

P. Uetz, Giot L., Cagney G., Mansfield T. A., Judson R. S., Knight J. R., Lockshon D., Narayan V., Srinivasan M., Pochart P., Qureshi-Emili A., Li Y., Godwin B., Conover D., Kalbfleisch T., Vijayadamodar G., Yang M., Johnston M., Fields S., Rothberg J. M. *A Comprehensive Analysis of Protein-Protein Interactions in Saccharomyces Cerevisiae***.** Nature, 403:623-627, 2000.

L. J. Van 'T Veer, Dai H., Van De Vijver M. J., He Y. D., Hart A. A., Mao M., Peterse H. L., Van Der Kooy K., Marton M. J., Witteveen A. T., Schreiber G. J., Kerkhoven R. M., Roberts C., Linsley P. S., Bernards R., Friend S. H. *Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer***.** Nature, 415:530-536, 2002.

A. Zien, Kuffner R., Zimmer R., Lengauer T. *Analysis of Gene Expression Data with Pathway Scores***.** Proc Int Conf Intell Syst Mol Biol, 8:407-417, 2000.