

Challenges in reengineering of hierarchical biological networks for drug discovery and development

Dr. Andreas Schuppert
Bayer Technology Services GmbH
PT AS CS

Summary

Understanding the regulation of biological systems plays a key role for further progress of drug discovery and development, molecular diagnostics as well as in biotechnology. The significant impacts of adverse drug reactions as well as the attrition rates in the clinical development phases in the drug development workflow on the costs of new drugs and, on the long term, on our healthcare system show that the economic impact of further progress in understanding the underlying, complex biological interaction networks leading to unexpected biological reactions on new pharmaceutical therapies can not be overestimated.

Biological interaction networks are not only coupled within a single, “horizontal” layer of systems parameter (like gene expression or protein - protein interaction). Obviously significant impacts on the regulatory system stems from “vertical “ interaction between networks of different layers of systems parameters. The complexity of the mutual coupling of regulation networks in both directions, “horizontally” and “vertically”, leads to challenging scientific tasks for the development of analysis and modelling tools for biological regulation networks.

Network interactions and the resulting impact on the R&D process in diagnostics and drug discovery will be sketched on examples.

0) Introduction

Understanding the regulation of biological systems plays a key role for further progress of drug discovery and development, molecular diagnostics as well as in biotechnology. The current lacks in our understanding of regulatory interactions of the large variety of entities in biological systems is one reason of the current economic issues in drug discovery and development as well as for drug safety aspects. These issues and their economic impacts on the health care systems are widely discussed and will therefore not be stressed in detail.

Today we see an increasing requirement for support from in silico model based predictions of the responses of biological systems on drug therapies, e.g.

- Prediction of individual risks of adverse drug reaction of the patients
- Prediction of individual “therapeutic windows”
- Prediction of responder / non responder groups in patient population
- Prediction of toxicity properties of new chemical entities

Comparable in silico predictions are today state of the art in lots of engineering oriented industries, such as airplane or car industries [1] and it is obvious that the ability to support these application areas with in silico methods would have dramatic effects on the efficiency of the health care systems.

The generic requirement, however, is the quantitative prediction of the input – output properties of large, complex biological systems which show a complexity which is orders of magnitude higher than the complexity of the largest engineering systems. Moreover, the poor impact of the biological reductionist approach to the study of isolated subsystems, like the genome, on the abovementioned application areas shows that in most cases the interaction of the components of the system plays a

crucial role which can not be compensated by a very detailed understanding of the isolated subsystems. Therefore appropriate systems decomposition and reverse integration methods as developed successfully for physical and engineering systems are the only chance to tackle successfully the biological complexity challenge.

Biological systems show, like most of the engineering systems, a hierarchical, top down structure in terms of “vertical integration”.

Vertically integrated systems are characterised by a hierarchy of decompositions of a system into subsystems, where on each hierarchy level the decomposition procedure leads to classes of subsystems with very similar properties inside each class. Each of the subsystems themselves may be decomposed in further hierarchical decomposition steps into even smaller structures.

On each hierarchical level the systems properties can be described by parameters and model equations which are adapted to the respective system level.

For example, in solid state physics the properties of the entire (non-crystalline) solid, like a polymer, can be described in terms of mechanical properties like elasticity modules. These parameters are overall properties of the solid state system on the macroscopic level and can be measured experimentally on the same level or can be calculated from properties of the subsystems. One subsystem may be the polymer crystallite phase (consisting of lots of individual crystallites) imbedded in an amorphous phase. All the crystallites show similar mechanical properties themselves, but they are very different to the properties of the amorphous phase. The properties of each of the subsystems, crystallites and amorphous phase, are typically very different to the properties of the overall solid state, which can be calculated from the subsystem’s properties by use of homogenisation theory for multiscale systems.

It is obvious that each of the subsystems, crystalline phase and amorphous phase, can be decomposed into smaller subsystems, such as molecular polymer chains with an ordered or disordered structure. On the molecular level the subsystems can be described in terms of new parameters, like stiffness of chemical bonds, which can be aggregated again to the mechanical parameters of the crystalline and amorphous phase each.

According to the decomposition approach described above the quantitative prediction of the properties of macroscopic systems can be performed on different levels:

If the model equations of each subsystem on any isolated hierarchical level can be determined experimentally or from the scratch by theory (e.g. quantum chemistry methods), then application of homogenisation theory leads to the overall properties of the entire system. Therefore the decomposition level used for modelling is simply a matter of practise and closure of the description of the system.

The basic requirement for successful application of the decomposition approach to systems modelling is the ability to describe the model equations of the subsystems on each isolated decomposition level without a strong interaction to the decomposition levels in higher levels of the systems hierarchy. E.g. the mechanical properties of crystallites do not depend strongly on the mechanical stress of the macroscopic solid polymer.

In a similar way a vertical decomposition of biological systems can be performed [2] such that similar modelling approaches might be applied. There are, however, two significant differences with respect to modelling features:

- Homogenisation theory as described above requires more or less homogeneous interaction features inside each hierarchical level: the embedding features of all polymer crystals in the amorphous polymer matrix are very similar. This is, however, not true in biological systems: the interaction structures inside each hierarchical systems level are far from being homogeneous. There are, e.g. few genes whose expression rates correlate strongly to the expression rates many other genes. In

contrast, the expression rates of most genes are strongly correlated to only a few “neighbour” genes. Such an inhomogeneous organisation structure is typical for biological systems and it is obviously far away from the structure which are tackled successfully by homogenisation methods.

- The second difference is the interaction of the properties of the subsystem on a given hierarchical decomposition level with systems properties of decomposition levels at a higher hierarchical rank. As discussed above the weak “top – down” interaction is a necessary prerequisite for a successful modelling of the system on a given hierarchical level and a following “back – integration “ to the overall properties of the entire systems via homogenisation. This prerequisite is satisfied in lots of physical and engineering systems, but it is not always satisfied, however, in biological systems. E.g. each cell (the subsystem) in an organ contains a full set of genes. There are, however, only approximately 10% of the genes activated in the cells. This is per se no contradiction to the “weak interaction” requirement of hierarchical modelling. The set of activated genes, however, depends strongly on environmental conditions of the entire system and the respective control signals are generated top down by the upper hierarchical levels. Therefore the properties of the subsystems “cell” depend strongly on the properties of the “upper” system organ which itself depends on the properties of the entire system “man”. Therefore it is not trivial to prove that modelling of the gene expression on the cell level and their impact on man without explicit integration of the possible top – down control mechanisms will lead to reliable results at it does in physical systems.

Both of these specific properties of interaction networks in biological systems lead to new challenges in analysis and modelling with significant impact on drug discovery and development, which will be sketched on two examples from biomarker development.

1) Analysis of gene expression rates for cancer / non cancer tissue characterisation

Analysis of gene expression rates is a common approach to develop biomarkers allowing the diagnostics and prediction biological responses on external and disease stress, such as tumour diagnostics, tumour progression, drug efficacy as well as toxicity issues.

The analysis and interpretation of gene expression rates with respect to the desired phenotypes suffers from two challenges:

- the severe lack in stability of the gene based biomarkers which are due to the extremely small sample sets compared to the overwhelming complexity of the possible gene expression patterns leading to poor p-values if the complexity is properly taken into account (Bonferroni correction)
- the “non-uniqueness” of the markers which can be developed from the available large sets of gene expression rates. Various available “generic” machine learning tools allow to predict the same phenotype from an identical, large set of genes. Each tool generates an own, predictive biomarker with similar performance compared to the other tools, but all the different biomarkers are typically based on poorly overlapping subsets of genes, which are automatically selected by the machine learning tools as an “optimal” subsets. Even worse, often the same tool is able to generate two biomarkers A and B with similar performance, but based on disjoint gene sets, when forced to neglect genes used for A when generating B. Obviously such a “biomarker continuum” may lead to severe pitfalls for IP strategies for gene diagnostics as well as lacks of reliability and confidence in applications.

Several experimental causes may contribute to the biomarker-continuum mentioned above:

- noisy measurements of gene expression rates by the available chip technology
- heterogeneous tissue samples
- differences in sample preparation methods

These experimental issues may explain the differences in the markers developed from expression rates measured in different studies with different measurement techniques, they can not explain, however, why data from the same data set may itself lead to a biomarker-continuum. Therefore the intrinsic coregulation network structure of gene expression may contribute to the application issues described above.

Recently intensive studies on large biological regulation networks for genomics, transcriptomics, proteomics and metabolomics have been done as described in the recent review from Barabasi and Oltvai [4]. It has been shown that biological regulation networks show a common structure, so called scale free networks, not depending on the tissue or organism [5,6,7]. Moreover it has been shown that the location of a gene in the network is related to the evolutionary history, the genetic stability as well as the physiological damage caused by knocking out the respective gene. Additionally, so called network motifs could be identified with biological functions [8].

To discuss these aspects a data set from colon tissues containing 2000 genes and cancer / control discrimination as clinical outcome has been analysed. This data set is one of the first examples in the literature where a biomarker – continuum has been discussed by Alon et al. [3, 9] and is used as a benchmark for pattern recognition.

Since the analysis is based on 2000 genes used to discriminate a set of only 62 tissues, a typical $n \ll p$ (n : number of data sets, p : number of parameters) has to be solved. Such data structures are typical for gene expression studies. Therefore the findings and methods developed to tackle the inherent issues of $n \ll p$ problems on this data set should be helpful to overcome similar issues in a large set of applications.

In [3], an efficient clustering method has been presented allowing to identify coexpression clusters as well as cluster based biomarkers allowing to discriminate between cancer and control tissues. The cluster based markers showed highly significant p values for cancer – control discrimination. However, it has been shown that omitting the 1500 leading genes involved in the markers with highest significance do not reduce significantly the performance of the markers derived from the reduced data set. Therefore this paper provides one of the first descriptions of a biomarker continuum.

Since most of the genes showed log-normal distribution in the expression rates, all gene expression rates have been transformed to their normalised logarithm.

Based on the logarithmic data the linear pair correlation between the genes have been calculated and the correlation interaction structure has been analysed. The connectivity of the genes in the correlation network shows a log – log distribution as is expected in scale free networks which are common in intracellular biological interaction networks [4].

Additionally the univariate p –values of cancer – control discrimination $p(g)$ for all genes using a standard Wilcoxon test [10] have been calculated. The number of genes $N(p)$ with given p shows again a log – log distribution.

Let $N(g,c)$ describe the number of genes with pair correlation coefficient $r > c$ with respect to a given gene g . $N(g,c)$ describes the connectivity of the genes in the correlation network. For high values of c ($c > .9$) the density $\rho(N)$ of $N(g,c)$ shows a power law leading to a linear log – log plot as depicted in fig. 1a, compared to exponential distributions leading to a linear single – log plot. The single log plot in fig 1b. shows clear nonlinearities in contrast to the log - log plot in fig 1a., therefore a power law distribution for the connectivity can be assumed as reality.

According to [4] this finding is a typical feature of scale free networks which are quite common in intracellular regulated networks in contrast to random networks which show an exponential distribution. This means that there exist a small set of genes, the hubs, which are highly correlated to an atypically high number of genes. In contrast to the “embedding” features of most of the genes in the coregulation network.

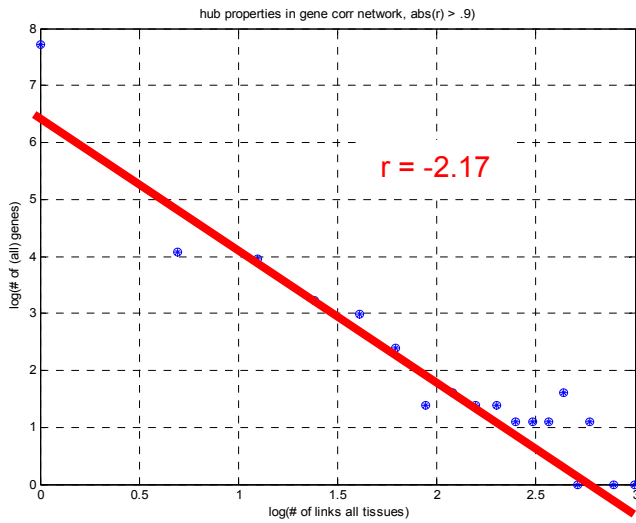


Fig. 1a: log – log plot of the density of $\rho(N)$ over the connectivity $N(g,9)$

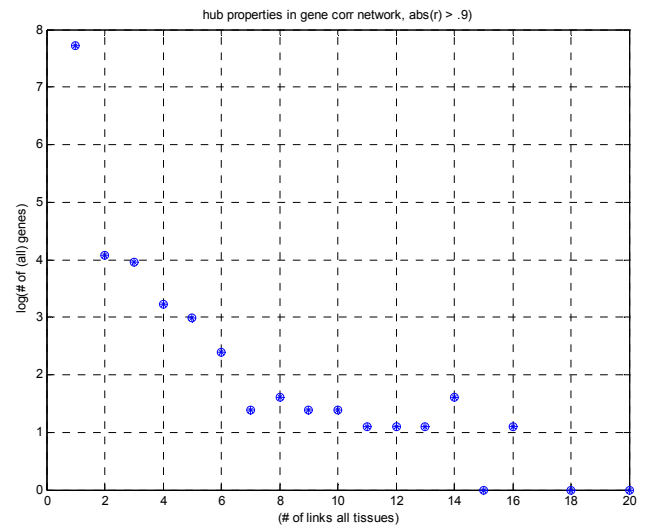


Fig 1b: single log plot

The univariate p values of each gene with respect to the distribution of the respective expression rates in cancer and control tissues has been calculated by a standard Wilcoxon test. Surprisingly the distribution of the number of genes $N(p)$ with given p shows a power law :

$$N(p) \sim p^{-\alpha}$$

leading to a linear log – log relation as depicted in fig 2.

In fig. 3 the $\log_{10}(p)$ – values of the all genes are plotted against the number genes $N(g,9)$. Obviously the set of strongly cancer – related genes (low p – values) are disjointed from the hub – genes (high $N(g,9)$). The only cluster of cancer related genes with medium p values showing higher $N(g,9)$ – values (green box) consists of genes which are related to the activity of protein expression in the proteasome. They are, however, not tightly integrated in the correlation network.

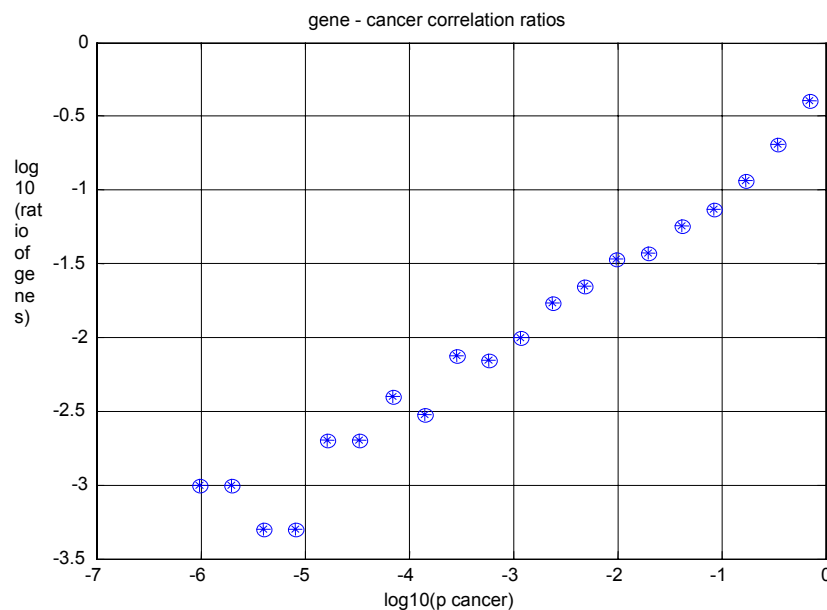


Fig.2: log – log plot of $N(p)$ over the p-value for cancer – control tissue discrimination

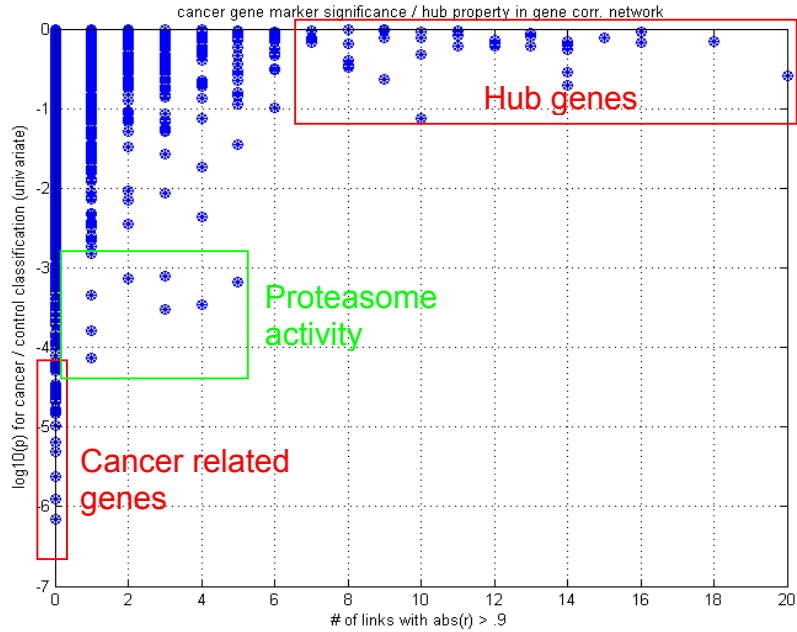


Fig. 3: distribution of log(p) over the connectivity $N(g, .9)$ for all genes

The disjointed distribution of cancer related genes and hub genes is in accordance to [4,5] showing that hub genes are closely related to functions which are critical for the survival of the cell. Therefore diseases which don't tend to kill the cells should not affect the hub genes.

The results described above show the inhomogeneous distribution of the gene – gene interactions on the same decomposition level “gene expression (transcriptome)” of a cell. Obviously, a “homogenization” approach assuming that each subsystem “gene” has similar “embedding” properties (as used successfully in material science) will not cover the features in living cells properly. Therefore such an approach will not lead to reliable predictions of the overall systems properties (in this example “cancer / non cancer cell”).

In contrast, however, as discussed above predictions on the basis of isolated genes (or sets of them) show notorious lacks on reliability leading to severe bottlenecks for clinical applications on the bedside. Obviously the interaction structures between the genes contribute significantly to the overall properties of the entire systems.

Things are even worse. As depicted in Fig. 3, the hub genes controlling large sets of genes are far from being significant to the system properties, a feature, which is not an isolated result of the studied data set. Therefore the most relevant parts of the genetic network are obviously located in the “intermediate” area of the gene expression network, far away from the hubs. In contrast to lots of the results on hub properties in scale free networks, today it is not clear how to describe the features of such intermediate, but obviously important, areas in scale free networks properly.

2) Prediction of adverse drug reactions from genotype and physiological data

The second example shows the interaction between physiological data from a higher level of biological system decomposition with genotype data on the genome level, a much lower level of biological system decomposition. The example stems from an proprietary study and therefore not be described in detail. It is, however, well suited to show the top – down impact of vertical system decomposition levels in biology.

A clinical study was performed to identify a biomarker allowing the prediction of an adverse drug reaction (ADR) under therapy. The patients were selected such that the percentage of ADR patients enclosed in the study has been significantly higher than in reality. From each patient, the genotype of genes has been determined which are supposed to be most relevant. The studied adverse drug reaction (ADR) is measured in terms of the concentration of an enzyme, named here CE. If CE is significantly higher after drug therapy than before, then the patient has a high risk of to have an ADR implicating a change in therapy.

The control group, patients which show no ADR, have approximately the same CE level after therapy than before. The “cases”, however, show significantly increased CE levels after therapy not depending on the CE levels before. This indicates that an additional mechanism for CE production has been induced by the drug therapy in the cases, but not in the control group.

All attempts to predict the ADR directly from the genotype data alone did not yield reliable results. The sensitivities and selectivities reached values of about 70%. This means, that 30% of all patients have been wrongly classified, a value which is not acceptable for broad applications. This finding is mentioned here because in many publications with respect to the prediction of clinical phenotypes from genotype data values of about 70% true classifications are reported indicating that significant contributions to clinical phenotypes are not encoded in the genome, but have to be identified in subsystems of other hierarchical decomposition levels. These findings are not sensitive with respect to the classification method, the variation between the classification methods are typically small compared to the gap required to reach reliable results.

Therefore one approach is to combine patient data from various levels of system hierarchy to improve the prediction quality. The idea behind this approach is that the additional data, e.g. from physiology or body constitution, may be not causal for the phenotype to be predicted, but they may be parameters which are correlated to the hidden (and mostly unknown) systems parameters which are causal and which have an control impact on the subsystems on the low level hierarchy. If, for example, a patient suffers from a liver damage like Hepatitis with an impact on the metabolisation of the drug, then this damage on the organ level can not be found in the genotype, but it impacts the patient's response on drug additionally to the genotype impact. It may additionally shift the expression levels of the genes such that the entire regulation structures on the lower hierarchical levels may be shifted.

Such a combinatorial approach has been successfully reported on gene expression analysis in oncology by [11]. They showed a significant improvement in prediction quality with respect to the predictions based on genetic level data alone, but they did not use any structured approach.

The pitfall in the pure addition of additional data structures is the increase of the numbers of parameters. The classification in clinical studies always suffers significantly from the large number p of measured parameters in each patient compared to the number n of patients available in the study. Mathematically these are typical $n \ll p$ problems leading to losses in reliability of the identified patterns. Therefore a mere addition of parameters may be not necessarily a benefit for the classification problem.

To overcome this pitfall we used a technique which has been developed in the framework of process modelling which allows basically to improve the $n \ll p$ pitfall significantly.

It has been shown that the explicit integration of hierarchical system structures into the data analysis and modelling problem of hierarchically structured systems improves the reliability significantly. The mathematical idea behind this so called “structured hybrid modelling (SHM)” approach is that hierarchical networks with a fixed interaction structure can be represented by a directed network structure. Each node represents an unknown, nonlinear function depending only on these variable which are represented by the directed edges into the node. Since the function space, which can be spanned by functional networks with fixed interaction structures, can be characterised in terms of invariant functional structures depending only on the interaction network structure, the SHM modelling approach shows superior features for modelling of complex systems [12].

If the vertical interaction structure of the hierarchical system can be properly represented by such a network, then (under some restrictions of the structure of the system graph) the network model can be properly identified and the number of data required for the identification can be reduced dramatically compared to unstructured model types. Therefore the hierarchical network based SHM models allow an interpolation in the modelling technologies between unstructured models (like neural networks) and fully mechanistic models. The superior features of SHM models for modelling of hierarchically structured systems have been shown in various applications in modelling of complex processes [13].

Although the computational effort to identify SHM models is higher than for unstructured models, a formal decomposition into physiological parameters split according to their physiological origin on the one side and genotype parameters on the other side allowed us to identify a model with superior performance. The sensitivity and selectivity could be improved from 70% to more than 90% now allowing practical application.

3) Conclusion

As discussed on the latter examples, network reengineering plays a crucial role for reliable modelling of biological systems. Since reliable quantitative and predictive models for the response of biological systems on various external stress factors is expected to improve the efficiency and costs of drug research and development significantly, providing appropriate tools for the described tasks are not only intellectually challenging, but will have a great economic impact, too.

Biological systems show a twofold interaction structure: vertically and horizontally. In contrast to physical or engineering systems, network reengineering of biological systems suffers from a significant inhomogeneity inside horizontal layers of system decomposition and additionally strong vertical top – down control interactions. Therefore the established methods for network reengineering suffer in practise from significant reliability bottlenecks.

Although no generic solution to both of the network reengineering challenges is available today, the current progress on identification of generic features of scale free networks as well as integration of data structures from heterogeneous hierarchical system levels via structured hybrid modelling show that the challenge of biological network reengineering with respect to drug discovery and development may be successfully tackled by future efforts.

References

- [1] P.W. Swaan, S. Ekins, *Reengineering the pharmaceutical industry by crash-testing molecules*, Drug Discovery Today, Vol. 17, pp. 1191-1200, Sept. 2005
- [2] J.W. Lengeler, *Metabolic networks: a signal oriented approach to cellular models*, Biol. Chemistry, Vol. 381, 911-920, 2000
- [3] U. Alon, N. Barkai, D.A. Nottermann, K. Gish, S. Ybarra, D. Mack, A.J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proc. Natl. Acad. Sci, USA, Vol. 96, pp. 6745-6750, June 1999
- [4] A. L. Barabási, Z. N. Oltvai, *Network biology: understanding the cell's functional organization*, Nature Reviews Genetics, Vol. 5, pp. 101-113, February 2004
- [5] A. Wagner, D.A. Fell, *The small world inside large metabolic networks*, Proc. R. Soc. Lond. B 268, pp 1803-1810, 2001
- [6] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi, *The large-scale organization of metabolic networks*, Nature, Vol. 407, pp. 651-654, 2000

- [7] S. Maslov, K. Sneppen, *Specificity and stability in topology of protein networks*, Science, Vol. 296, pp 910-913, 2002
- [8] R. Milo, S.S: Shen-Orr, S. Itzkovitz, N. Kashtan, U. Alon, Network motifs: simple building blocks of complex networks, Science, Vol 298, pp. 824-827, 2002
- [9] <http://www.weizmann.ac.il/mcb/UriAlon/>
- [10] Matlab, Statistics package, Version 14.1
- [11] J. Pittman et al., Integrated modeling of clinical and gene expression informatin for personalized prediction of disease outcomes, , Proc. Natl. Acad. Sci, USA, Vol. 101, no. 22, pp. 8431-8436, June 1, 2004
- [12] Schuppert, A., *Extrapolability of structured hybrid models: a key to optimization of complex processes*, in: Proceedings of EquaDiff'99, B.Fiedler, K.Gröger, J.Sprekels Eds., (2000), pp. 1135-1151
- [13] Mogk, G.; Mrziglod, Th.; Schuppert, A.: *Application of Hybrid Models in Chemical Industry*, Proceedings of the ESCAPE 12, Proceedings of ESCAPE 12, J. Grievink ed., Elsevier 2002