Journal of Mathematics in Industry (2011) 1:6
DOI 10.1186/2190-5983-1-6

Journal of Mathematics in Industry
a SpringerOpen Journal

RESEARCH Open Access

Efficient reengineering of meso-scale topologies for functional networks in biomedical applications

Andreas A Schuppert

Received: 17 December 2010 / Accepted: 23 June 2011 / Published online: 23 June 2011

Abstract Despite the deluge of bioinformatics data, the extraction of information with respect to complex diseases remains an open challenge. The development of efficient tools allowing the re-engineering of functional biological networks will therefore be crucial for the future of the pharmaceutical and biotech industry. In this paper we present a method for efficient re-engineering of meso-scale network topologies for biomedical systems from stationary data. We show that the meso-scale topology is related to functional structures of the input-output data of the entire system, which can be unravelled from high throughput screening experiments, without information with respect to intermediate variables. Analysis of the functional structure of the data provides a complementary approach to established network reengineering methods based on combinatorial optimization. A combination of both approaches will help to overcome the drawbacks of the established network reengineering algorithms.

1 Introduction

The health care systems of western ageing societies suffer from a continuously increasing frequency of complex diseases, such as cancer, metabolic syndrome, auto immune diseases or diseases of the central nervous system. In contrast to infectious diseases, all these diseases are characterized by a dysfunction of the biological regulation systems of patients. They cannot be reduced to single root causes and we still lack a sound mechanistic understanding of even the un-diseased function of the relevant regulatory systems. Consequently little progress has been seen in drug research

AA Schuppert (✉)
Aachen Institute for Advanced Studies in Computational Engineering Sciences, RWTH University of Aachen, Schinkelstrasse 2, 52062 Aachen, Germany
e-mail: schuppert@aices.rwth-aachen.de

AA Schuppert
Process Technology, Bayer Technology Services GmbH, Bldg. 9115, 51368 Leverkusen, Germany

and there are no ‘silver bullets’ for cancer or Parkinson’s disease as compared to antibiotic therapy of microbial infections. In all complex systemic diseases the medical need is still very high. Despite the deluge of genome or proteome data accessible today, the extraction of biologically relevant information is an open challenge. On the background of the estimated cost of $ 1,000 for sequencing of an individual human genome, this challenge was named the ‘one-million-dollar-interpretation’.

Despite steadily increasing investments into drug research and development operations and the introduction of novel technology platforms like high throughput and high-content screening, up to date the output of novel, effective drugs for complex diseases is not only low but shows a continuous downturn. As a direct consequence of this lack of R&D efficiency, the average investment for research and development per drug newly approved by the regulatory agencies already exceeds $ 1,000 mil. From project initiation to marketing authorization, a normal Pharma R&D project takes more than 10 years. Even worse, up to 83% of the drug candidates which are successful in pre-clinical tests fail in the clinical development phase where the drug candidate is tested in human volunteers and patients, and a still significant proportion fails in the most expensive late pivotal trials.

High attrition rates in clinical development are an important contributor to the overall costs of novel drugs. Our inability to predict these failures is, at least partially, caused by the lack of tools which allow the prediction of the efficacy in patients based on lab and pre-clinical animal data.

This situation is mainly caused by the lack of understanding of the mutual interactions of the biological entities which are involved in disease development as well as in drug action. Neither the combinatorial effects of abiotic stress, genotype variations and drug action nor the induced long term stress response of the cells on drug action can be predicted. In consequence, this lack of predictive models leads to unexpected adverse drug reactions or insufficient efficacy of the drugs which are observed in the late clinical trials at high costs. Thus, efficient network re-engineering methods leading to reliable predictive models would have a tremendous economic impact.

Over the last years it has been shown that biological entities such as proteins or genes show a strong interaction in order to guarantee the survival of the cells. The respective interaction networks show a small world topology [1] leading to strongly cooperative effects which are not fully understood.

Moreover, the biological processes controlling drug efficacy or development of diseases are based on networks of heterogeneous, yet interacting, biological functionalities. Modelling and prediction of the efficacy of drugs will therefore require the re-engineering of the respective functional networks, which are far less understood than the protein-protein interaction networks. The established methods for unravelling of biological networks are based on combinatorial optimization and statistical algorithms [2]. Despite significant progress in network reengineering of small and medium-sized networks [3], for large-scale networks the one-step methods suffer from the exponential increase of complexity with the number of involved functionalities. So far, the established methods for complex processes are far from being satisfactory or from being ready for use in a standardized workflow of any industrial R&D processes.

For these reasons the development of efficient tools allowing the systematic re-engineering of functional biological networks from the massive deluge of data which is available today will be crucial for the future of the pharmaceutical and biotech industry.

In order to overcome the complexity gap of a direct network re-engineering approach, our approach aims to establish an efficient meso-scale network re-engineering procedure. In order to reduce size and complexity of detailed network models, meso-scale modelling aims to lump sub-processes and sub-networks to 'effective' functional nodes without loosing the accuracy of the overall model. Meso-scale models provide an interpolation between detailed and black box models representing the dominating functionalities by 'effective' input-output models, connected by their interactions [4, 5]. Based on the meso-scale structure, the network may be decomposed into small, separate sub-networks which can be directly re-engineered with significantly lower complexity. So far, meso-scale network re-engineering may provide a step towards efficient multi-scale network re-engineering workflows. In this paper we will describe novel mathematical approaches which allow the efficient re-engineering of network topologies for biomedical applications from high-throughput data. In contrast to one-step combinatorial methods using minimization of residuum functionals, the novel meso-scale network reengineering approach is based on the functional or algebraic structures of input-output functions of the entire system. These structures can be identified from modern high throughput experimentation facilities. They are numerically less demanding than the combinatorial optimization approaches and show improved stability with respect to small errors in the data due to the focus on the meso-scale network topology only.

We will first describe a direct approach for reengineering of the structure of hierarchical functional networks. Hierarchical functional networks allow the establishment of models linking data and functionalities from heterogeneous levels of a system structure. It has been shown that combining data from the genome and the physiology level in a systematic approach can result in significantly improved predictions of 'macroscopic' biological phenotypes [6, 7].

We will then develop a method for the reengineering of meso-scale structures for non-hierarchical networks, which allow to model cooperative interaction on a homogeneous level of the system structure, for example, phosphorylation of signalling proteins in response to external stimuli and inhibition.

2 Re-engineering of hierarchical functional networks with feed-forward structure

The identification of quantitative models f linking biological stress factors and molecular markers, such as mutations on the genome, with macroscopic biomedical phenotypes plays a crucial role for a broad range of applications in biomedicine. This requires the identification of a quantitative model describing the readout y as a function f depending on multivariate input variables $\underline{x} : y = f(\underline{x})$, $y \in \Re$, $\underline{x} \in \Re^n$, $n \gg 1$. The readout y shall quantify the observed reaction of a biological system in response to biotic or abiotic stress factors as well as molecular markers of the system, which are quantified by the input variables represented by the components of $\underline{x}$. For these applications, it is not necessary to map the detailed biological mechanisms in the model. It is sufficient to develop the so called biomarker models representing only the overall input-output relation of the system. Examples arising in drug research or biotechnology are:

- High throughput experiments in drug discovery, where the input of the system consists of the set of structural descriptors of the chemical compounds whereas the output is given by the respective biological activity of the compounds.
- Genome-wide association studies, where the set of mutations forms the input vector $\underline{x}$ and the output is given by the classification of the biological status, for example, the disease or drug action which are associated to the respective genotype.
- Combinatorial stress experiments, where various combinations of stimuli and/or inhibitors are applied to cellular systems forming the input vector $\underline{x}$. The respective output is given by the cellular response which can be quantified by means of phosphorylation of signalling proteins [3], gene or protein expression.

The straightforward approach for biomarker identification uses machine learning algorithms such as support vector machines, neural networks or logic models [8]. These so-called black-box approaches provide algorithms which allow the construction of quantitative input-output relations from data for all sufficiently smooth functions without any mechanistic understanding of the underlying mechanisms. The drawback of black-box approaches, however, is that the data demand increases (in the worst case) exponentially with the dimension of the input variables $\underline{x}$ (curse of dimensionality). In biomedical applications, where the number of input variables (for example, genes, mutations or proteins) can easily exceed 10^4, this approach can result in unaffordable data demands. It is therefore a fundamental challenge for mathematics to develop modelling approaches which allow a systematic combination of a priori mechanistic knowledge and black box algorithms in order to provide tools with a controlled ratio between the demands on a priori knowledge and data.

3 First step: modelling of hierarchical functional networks

Suppose the system under consideration is controlled by n input variables $\underline{x} \in \Omega \subseteq \Re^n$ and produces one output variable $y = y(\underline{x}) =: \Re^n \to \Re$. The input-output relation of the system can be modelled using black-box approaches where no a priori knowledge with respect to the system is required. However, black box modelling suffers from a data demand increasing exponentially with the number of input variables, which has therefore been called the 'curse of dimensionality' [9]. Although it has been shown [10] that restrictions on the input-output functions can reduce the data demand significantly, the tremendous dimensionality of biological data sets lead to unsatisfactory results yet. Improved modelling approaches, compared to the pure black box modelling, are urgently required here.

In functional network models the system is decomposed into interacting subsystems which are characterized by their input-output behaviour described by the set of functions $\underline{u}(\underline{x})$, where the function representing a node l depends only on a subset of components of $\underline{x}$: $u^l = u^l(\underline{x}^l)$, $\underline{x}^l \in \Re^{m_l} \subset \Re^n$, $m_l < n$. Each input-output function u^l can be represented by a given mechanistic model or, alternatively, by a black-box model. The mutual interaction of the sub-systems is represented by a directed graph

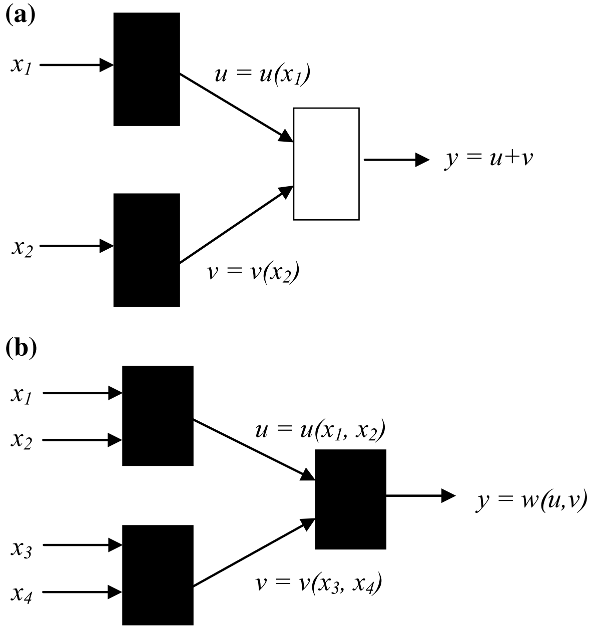

Fig. 1 Structures of functional networks. (**a**) Functional network consisting of two black-box nodes, represented by the functions $u(x_1)$ and $v(x_2)$, and a mechanistic model, exemplified by the function $y(x_1, x_2) = u(x_1) + v(x_2)$. u and v are the input-output functions of the respective nodes. The outputs u and v are input variables of the downstream nodes as well, indicating that a functional network represents a concatenation of functions. (**b**) Functional network consisting of three black-box nodes, represented by the functions $u(x_1, x_2)$, $v(x_3, x_4)$ and $w(u, v)$, depending on two input variables each.

S, the nodes of which represent the sub-systems and the edges the respective input and output variables. In neural networks the input-output functions of the nodes are fixed up to a small set of parameters and the structure S is used for the adaption to the data. In contrast, in functional networks the structure S is fixed and the input-output functions of the nodes are fit such that the overall model represents the data (Figure 1a, b).

Functional networks show highest benefits if the systems can be decomposed into sub-systems which are controlled by a few input and output variables, whereas the mechanisms inside the functionalities show a significantly higher interaction between the components. A functional network thereby provides a meso-scale model for the systems with significantly reduced complexity.

Such functional networks can always be established, if the system to be modelled consists of well-defined subsystems and the connections between the subsystems are known. Various industrial applications have been realized successfully [11–14], and software implementations are available as well.

The analysis of the properties of functional networks goes back to Hilbert's 13th problem, which was solved by Vitushkin [15]. He found that the so-called Vitushkin-Entropy of a functional network allows the decision whether all functions depending on n variables can be represented or only a constrained set of functions. However, he did not discuss the consequences for modelling and network reengineering.

4 Second step: direct reengineering of functional networks with tree structure

If S is of tree structure, it has been shown before [4, 5] that for all such functional networks there are low-dimensional manifolds $M \subset \Re^n$ such that it is sufficient to measure data in a U_ε-environment of M to in order to identify the model properly. Such manifolds M are called data bases. The same authors have proven that the minimal dimension of data bases is equal to the maximum number of input edges of any black-box node in the network. Moreover, almost all differentiable, monotonic submanifolds $M \subset \Omega$ with $\dim(M) =$ maximum number of input variables in a black box node have (at least locally) the properties of a data base. Additionally, direct as well as indirect identification procedures have been analysed and implemented in software [13].

This result is based on the structure of S which guarantees that, despite all nodes in S may be black box models, the overall functional network model cannot represent any smooth function $y = y(\underline{x})$ depending on n input variables. Now we show that this intrinsic property of hierarchical functional networks is a specific property of the topology of S and allows, if large enough data sets are available, a direct reconstruction of the topology of S from data.

In all functional models, where S has a tree structure, there will be a unique path P_i connecting each input variable x_i to the output node. As the paths from inputs i and j to the output node may join in a node k, P_i and P_j are not necessarily disjoined. Suppose all node functions are strictly monotonic in all variables with bounded second derivatives. Then the partial derivatives of the output function $y = y(\underline{x})$ with respect to x_i are the product of the partial derivatives of all i-o-functions u_k along the path P_i starting at the input node of x_i and ending with the output node of the entire model:

$$y_{x_i} = \left(\prod_{k=2:length(P_i)} \partial_{u_{k-1}} u_k \right) \partial_{x_i} u_l =: \partial_i P_i \, \partial_{x_i} u_l,$$

where u_l is the input-node of x_i. The term $\partial_i P_i$ represents the product of the partial derivatives of the functional nodes along the path P_i with respect to x_i. Let P_{ij} be the common part of the paths P_i and P_j, then it holds $\partial_i P_{ij} = \partial_j P_{ij}$.

Let the input variables x_i and x_j be input variables to the same input node l whose input-output relation is represented by the function $u_l = u_l(\ldots, x_i, \ldots, x_j, \ldots) =: u_l(\underline{x}^l)$ and x_k be an input variable to any other node. Then application of the chain rule for derivations with respect to x_i, x_j and x_k leads to the following set of partial differential equations (PDEs) for the output function $y = y(\underline{x})$:

$$\begin{aligned} y_{x_i} &= \partial_i P_i \, \partial_{x_i} u^l, \\ y_{x_j} &= \partial_j P_j \, \partial_{x_j} u^l. \end{aligned} \tag{1}$$

Since the variables i and j are inputs of the same node u_l, P_i and P_j are identical. The respective products of the partial derivatives along both pathways are the same for i and j, leading to the relation:

$$\frac{y_{x_1}}{y_{x_j}} = \frac{\partial_i P_i u_{x_i}}{\partial_j P_j u_{x_j}} = \frac{u^l_{x_i}}{u^l_{x_j}} (\underline{x}^l). \tag{2}$$

All partial derivatives of (2) with respect to any variable x_k which is not part of $\underline{x}^l$ will vanish everywhere:

$$\partial_{x_k}\left(\frac{y_{x_i}}{y_{x_j}}\right) = 0, \quad \forall x_k \notin \underline{x}^l. \tag{3}$$

Therefore, all functions $y = y(\underline{x})$ which can be represented by the functional network have to satisfy the set of PDEs:

$$y_{x_j}\, \partial_{x_k} y_{x_i} - y_{x_1}\, \partial_{x_k} y_{x_j} = 0 \tag{3a}$$

for all triplets $i, j, k \in [1, \ldots, n]$ where x_i and x_j are inputs to the same node, whereas x_k is the input to another node.

Generalizing this argument, we show that S is associated with an even larger set of structural PDEs that $y(\underline{x})$ has to satisfy. Now let the root and rank be defined as follows:

Definition 1 Node k shall be the **root** T_{ij} of the input variables x_i and x_j, if the pathways from x_i to the output of the entire system z and from x_j to y join for the first time in node k. As in tree structures the pathways from each input variable to the output are unique, all pairs of input variables will have a unique root.

The **rank** $Rg(k)$ of a node k shall be given by the length of the path from k to the output z of the entire system. In tree structures each node will have a unique rank.

Then, in tree structures with n input variables x_i and one output variable y the following theorem holds:

Theorem 1 (Structure-Constraint Theorem) For each triplet of input variables $\{x_i, x_j, x_k\}$, $i, j, k = 1, \ldots, n$, the conditions:

(i) $y_{x_i}\, \partial_{x_k} y_{x_j} - y_{x_j}\, \partial_{x_k} y_{x_i} = 0$

and:

(ii) $\{Rg(T_{ij}) > Rg(T_{ik})\} \wedge \{Rg(T_{ij}) > Rg(T_{jk})\}$

are equivalent.

Remark Eq. (3a) is a special case of the structure-constraint theorem, where $Rg(T_{ij})$ is maximal.

Proof For all triplets i, j, k satisfying (ii) the pathways P_i, P_j and P_k must be at least partially disjoined. As (ii) is satisfied, each of the pathways can be decomposed into three components with specific overlaps:

$$\begin{aligned} P_i &= P_i^0 \circ P_i^1 \circ P_i^2, \\ P_j &= P_j^0 \circ P_j^1 \circ P_j^2, \\ P_k &= P_k^0 \circ P_k^1 \circ P_k^2, \end{aligned} \tag{4a}$$

$$P_i^1 = P_j^1, \quad P_i^2 = P_j^2 = P_k^2 \tag{4b}$$

with

$$\partial_j P_i^0 = \partial_k P_i^0 = \partial_i P_j^0 = \partial_k P_j^0 = \partial_i P_k^0 = \partial_j P_k^0 = 0,$$
$$\partial_k P_i^1 = \partial_k P_j^1 = \partial_i P_k^1 = \partial_j P_k^1 = 0$$

and, because of the partial coincidence of the pathways: $P_i^1 = P_j^1$, $P_i^2 = P_j^2 = P_k^2$, it holds:

$$\partial_i P_i^1 = \partial_j P_j^1,$$
$$\partial_i P_i^2 = \partial_j P_j^2.$$

Equation (2) leads to

$$\frac{y_{x_1}}{y_{x_j}} = \frac{\partial_i P_i u_{x_i}}{\partial_j P_j u_{x_j}} = \frac{\partial_i P_i^0 \times \partial_i P_i^1 \times \partial_i P_i^2 \times u_{x_i}^{l_i}}{\partial_j P_j^0 \times \partial_j P_j^1 \times \partial_j P_j^2 \times u_{x_j}^{l_j}} = \frac{\partial_i P_i^0 \times u_{x_i}^{l_i}}{\partial_j P_j^0 \times u_{x_j}^{l_j}}.$$

Because of (4b) the last term does not depend on x_k, and it holds:

$$\partial_k \frac{y_{x_i}}{y_{x_j}} = \partial_k \frac{\partial_i P_i^0 \times u_{x_i}^{l_i}}{\partial_j P_j^0 \times u_{x_j}^{l_j}} = 0 \Rightarrow y_{x_i}\, \partial_{x_k} y_{x_j} - y_{x_j}\, \partial_{x_k} y_{x_i} = 0$$

On the other side, if (i) holds, then we can find a decomposition of the respective pathways P_i, P_j and P_k according to eq. (4a) and (4b), resulting in (ii). □

Based on the Structure-Constraint Theorem, the structure S of the functional network can be unravelled from the data as follows:

Algorithm 1
Direct hierarchical functional network reconstruction:

i. Test for any triplet of input variables i, j, k whether condition (i) of the structure-constraint theorem is globally satisfied leading to a full set of satisfied rank-root conditions for the structure S.

ii. Pick all double combinations i, j where for no $k = 1, \ldots, n$ the condition (ii):

$$\{Rg(T_{ik}) > Rg(T_{ij})\} \wedge \{Rg(T_{jk}) > Rg(T_{ij})\}$$

holds. Then i and j are inputs to the same input node. Use this combinatorial information to distribute all input variables onto their respective input nodes.

iii. Join the outputs of each input node l to one 'child' variable x_l'. The roots for a 'child' variable x_l' are equal to those roots of the respective 'parent' variables which are not yet identified as input nodes. The respective ranks for the roots of the 'child' variables are the ranks of the respective roots of the parent variables minus 1. So we arrive at a new, smaller structure S' which consists of all nodes which have not been identified in step (ii) as input nodes. Therefore, S' is identical to the respective part of S, the input variables of S' are the 'child' variables of the input nodes. The respective roots and ranks can be determined from the roots and ranks from S.

iv. Distribute the 'child' variables as input variables of S' on their input nodes in S'. This can be performed as described in step (ii) leading to novel 'grand-child' variables. To do so, go to step (ii).

v. In each tree-structure there exists m, $m < \infty$, such that m loops of steps ii-iv described above will lead to a structure $S^{m\prime}$ where all new input variables have the same root node. Then this common root is the output node of the entire system structure S and the algorithm stops.

Notes

a. If for all triplets of input variables $\{x_i, x_j, x_k\}$ the rank-root relations are known, then the adjoint tree structure of S can be directly reengineered from this set of relations. Therefore, if very large sets of data are given (for example, from high-throughput experimentation) such that a reliable test on truth of the conditions (i, ii) for all triplets can be performed, then the structure of the underlying functional network can be directly reconstructed. This direct approach is much more effective than the approach of identifying quantitatively the model for all possible model structures S, then selecting the structure of the model with the lowest residues.

b. The results described above can be transferred to models with discrete, for example, binary outputs. Then it allows the direct identification of the structure of the functional mechanisms behind the measured data in various scientific applications, if, for example, in the identification of pharmacological mechanisms from high-throughput screening data [16].

The direct network identification algorithm provides a very efficient approach to hierarchical network reengineering. It is superior to one-step reengineering approaches which need the minimization of an error functional of residues, which leads to a highly nonlinear, combinatorial optimization problem. As the algorithm can be generalized to discrete variables, it may be an efficient method for the analysis of next generation sequencing data when large data sets will be available. However, its drawbacks are the existing limitation to tree structures as well as the required estimates for condition (i) which is an ill-posed problem. Further research will be necessary for the development of stable routines which can be applied by non-experts in a standardized workflow.

5 Re-engineering of meso-scale structures for non-hierarchical networks

Intracellular signalling networks provide a mechanism for regulating cellular cross-talk and gene transcription. Protein phosphorylation plays the dominant role in activation of cellular signalling. Development of an efficient modelling and simulation of the response of signalling protein phosphorylation on multiple, complex combinations of stimuli and inhibitors is crucial for improved research for targeted drugs and may play an important role in systematic development of direct reprogramming of cells in future. Moreover, insight into the structure of mutual protein-protein interactions can provide direct information into multifactorial stimulation-response relations which are crucial for experimental design in drug research and therapies. The reconstruction of a stimulation-inhibition network between signalling proteins will lead to a significantly improved benefit compared to direct response modelling of individual proteins.

The established network reconstruction algorithms for reconstruction of signalling networks using phosphorylation data in response to external stimuli typically solve a combinatorial, mixed-integer optimization problem in order to minimize the error of a network-based signalling model with given experimental data. Nodes represent target proteins and edges (connections between nodes) represent the cascade direction of stimulated protein phosphorylation. However, if the number n of network nodes increases, then the number of potential networks to be analyzed will increase at least exponentially with n. Thus, any algorithm using an exhaustive search analyzing all possible networks with n nodes will become impractical even at modest n. Since most mechanisms which are relevant for applications involve multiple pathways and their crosstalk, there is a need for algorithms which avoid the pitfalls of detailed network reengineering in only one step.

In order to avoid computationally exhaustive one-step searches, network reconstruction has been tackled by others using a variety of methods, such as heuristic combinatorial optimization algorithms [17], efficient linear programming algorithms using sparsity constraints [3] or Boolean network modelling [18]. The interaction models describing the transfer of stimulation and inhibition across the network can be binary, logarithmic or kinetic (as in Michaelis-Menten models). These approaches are motivated by the kinetics of protein activation and lead to good fits for protein phosphorylation in terms of stimulation and inhibition [19]. However, this approach requires the explicit integration of all ‘hidden’ proteins unaccounted for in the network, but which are likely involved in the entire signalling mechanism of the network model, even if their phosphorylation status is not experimentally available. Moreover, depending on cellular status, the structure of the network may change, such that only subsets of proteins are expressed. Therefore, a fine-grained model may provide very detailed insight, however it requires networks with very high complexity. Moreover, as proteins may be taken into account whose phosphorylation levels have not been measured, the direct network reengineering algorithms may become ill-posed hampering the stability and numerical efficiency of the network reconstruction. Additionally, incorrect signal transfer models along edges can result in unstable network models as well.

We here present an algorithm which allows direct extraction of topological mesoscale features of a functional network using combinatorial stimulation-inhibition data without dynamic information. The concept is based on the functional network reengineering concept (as described above), but the focus is on the development of additional modules in order to overcome the drawbacks of the hierarchical network reconstruction algorithm in the special case of signalling network reengineering from stimulation-inhibition data.

In this case, a functional network refers to a group of inter-dependent protein kinases and their associated level of activation by phosphorylation status. The motivation is threefold: First to establish a hierarchical, reengineering approach for networks on a coarse-grained level, second to describe the system response upon multiple stimulations and inhibitions of the network proteins and third, to describe co-regulation by a network of minimal complexity. The resulting network allows for network-resolution adjustment based on available data and provides a starting point for further, more detailed, network reengineering with less computational complexity. Moreover, since we restrict on patterns that focus strictly on induced sub-networks where experimental data are available and which are expressed in the cells, we avoid complications from uncertainties and assumptions as seen in detailed, fine-grained model quantification.

Since both drug research and cellular engineering rely on the identification of targets as potential points of pathway inhibition, we represent the coarse-grained (or large-scale) network by inhibition-induced network topology. Two topological features will be analyzed in detail:

- Effect(s) of inhibition on the topology of a co-regulatory network
- Localization of the apparent inhibition of a signalling protein upstream or downstream of crosslinks between stimulation pathways connecting respective receptors with signalling proteins

Our approach is based on the observation that the topology of interaction networks connecting stimuli and targets induces characteristic patterns with relation to a function which quantifies the combinatorial stimulation-activation. In contrast to direct reconstruction we use the model as a ‘model-based filter’ in order to reconstruct the coarse-grained network topology. The advantage of this intermediate filtering step is that we can establish a tight balance between the ‘coarseness’ of the network reconstruction and the statistics of error minimization. Moreover we can reduce redundancies thus enabling us to avoid a large net error in network reconstruction.

Since the model is based on direct pattern identification within a network, it provides an efficient method to identify functional interaction or crosstalk between stimulation pathways (also referred to as ‘structures’) which can be used to define constraints for more detailed network reconstruction. Since this functional structure of stimulation and inhibition is derived from data, it provides direct insight into the network structure expressed by cells given their respective microenvironment and/or disease status.

As a full reengineering of functional networks without tree structure is not available, the meso-scale reengineering approach will be focused on the extraction of various specific topological features of the signalling network which induce detectable structural signals in the stimulation response functions, without claim for completeness. The overall network has to be established by combination of the identified topological substructures.

6 First step: identifying downstream pathway inhibition

The phosphorylation shift (from baseline phosphorylation state), z, of a protein in reaction to stimuli $x_1 \dots x_n$ and inhibitor y is expressed by the stimulation-inhibition response function (SRF):

$$z = F(x_1, \dots, x_n, y), \tag{5}$$

where $x_i \equiv$ *stimulus* i (applied $= 1$, none $= 0$) and $y \equiv$ *inhibitor* (active $= 1$, none $= 0$). Thus, if the stimulus i is applied on the cell, $x_i = 1$, and if the inhibitor is active, $y = 1$, otherwise the respective values are set to zero. The activation function, F, represents the dependence of the stimulation response function (SRF), z, on the stimuli and inhibitors. The baseline value of F is normalized to zero if no stimuli and no inhibitors are applied.

Since all variables are assumed to be binary, F is fully equivalent to a multilinear-form:

$$F = \sum_{k=1,\dots,n} \sum_{i_1,\dots,i_k} (\gamma_{i_1 \dots i_k} + \eta_{i_1 \dots i_k} y) x_{i_1} \dots x_{i_k}, \tag{6}$$

where $\eta \equiv$ effect of inhibition on the extent of phosphorylation, $\gamma_I \equiv$ effect of stimulation in the wildtype protein (without inhibition) on the extent on the phosphorylation (with respect to the baseline). If the experiments cover all combinations of stimuli and inhibitors from $k = 1, \dots, m$, then all coefficients of F up to interactions of order m can be estimated from the data. In each experiment coefficients will be 'active' (and contribute to the phosphorylation, F) only if all stimuli $x_{i_1} \dots x_{i_k}$ are applied. Hence, if the experiments cover linear and binary combinations of stimuli only, with and without inhibitors, the terms γ_i, γ_{ij}, η_i, η_{ij} can be identified from the experiments using standard linear estimators as the universal representation of the activation function F.

We will now show that the algebraic structure of the coefficients contains structural information regarding the interaction of functional pathways between stimuli and inhibitors.

Definition 2 Let $I =: i_1, \dots, i_k$ be a set of indices for parameter γ_I. Then, the sum of all $\gamma_{I'}$ with index-combinations $I' \subseteq I$ forms the spray Γ_I of I:

$$\Gamma_I = \sum_{I' \subseteq I} \gamma_{I'}. \tag{7}$$

Theorem 2 (Downstream sub-network topology identification) Assuming that the input-output relations of all functional nodes in the network are strictly monotonic, the following holds: Suppose $S = [i_1, \dots, i_k]$ is a set of stimuli, such that for all subsets of S the sprays $H + \Gamma$ for the response with inhibition and Γ without inhibition satisfy

$$H_{i_1 \cdots i_m} + \Gamma_{i_1 \cdots i_m} = h(\Gamma_{i_1 \cdots i_m}) \tag{8}$$

with a strictly monotonic function h, (where $h(0) = 0$ since no stimulus will always lead to the baseline level of stimulation), then the data are consistent with a functional network where the pathways from the receptors (for all stimuli in S) to the protein merge together before any effect from the inhibitor on these pathways becomes relevant (Figure 2a, b).

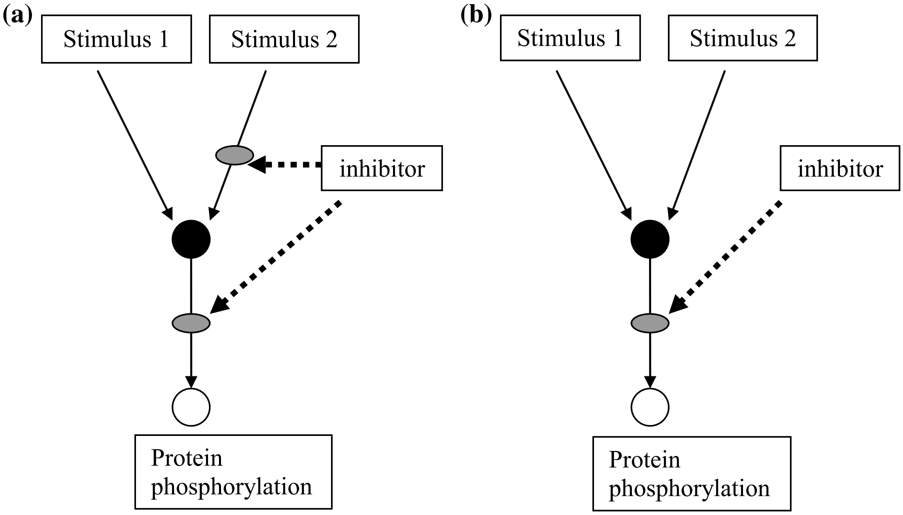

Fig. 2 Scheme of upstream/downstream inhibition. (**a**) Inhibitor acting on a signalling pathway upstream and downstream of a 'merger' node (black circle). If the 'merger' node has nonlinear characteristics, the inhibition can affect the entire response of the stimulation. (**b**) Inhibitor acting downstream of the 'merger' node only. Then the strength of the 'joint' stimulation signal is affected without change of the respective rate of the phosphorylation induced by the combined stimulation.

Proof All combinations of stimuli from S generate a combinatorial subset containing all coefficients γ, such that the respective sprays represent the outcomes of all (possible) stimulation experiments with respect to the stimuli in S. However, even for moderate numbers of stimuli not all possible combinations will be realized even in experimental high-throughput settings. Let us now denote G_S to be the set of all available γ representing combinations of stimuli from S. Sorting the elements of G_S with respect to size leads to a monotonic sequence G'_S representing the ordered stimulation intensities from S without inhibition. If we additionally apply an inhibitor (that is, $y = 1$), then the ordered sequence G'_S may be rearranged due to (possibly multiple) interactions between the inhibitor and the functional nodes which are involved in signalling of the stimulation towards the phosphorylated protein (Figure 3a). If all pathways from the stimuli to the protein merge upstream of a 'merger' node (Figure 2b), then the activation of the 'merger' node will be a monotonic function $F(G_S)$ representing the overall activation by all stimuli. Therefore any inhibition downstream of the 'merger' node will affect only the overall activation, such that the ordered sequence G'_S will not be affected by the inhibition. Then there will be a monotonic function h such that the quantitative effect of the inhibition on the stimulation of the protein can be expressed as before (Figure 3b). If, comparing the same stimuli with and without inhibition, the data show that the sequence G'_S is not permuted by the application of the inhibitor, then the phosphorylation data with inhibition can be explained by only one inhibition mode acting on the entire set of pathways from all stimuli in S towards the protein.

However, the inverse is not guaranteed: For any finite set of stimuli, small inhibitions may occur which do not change the sequence of order of stimulation, even

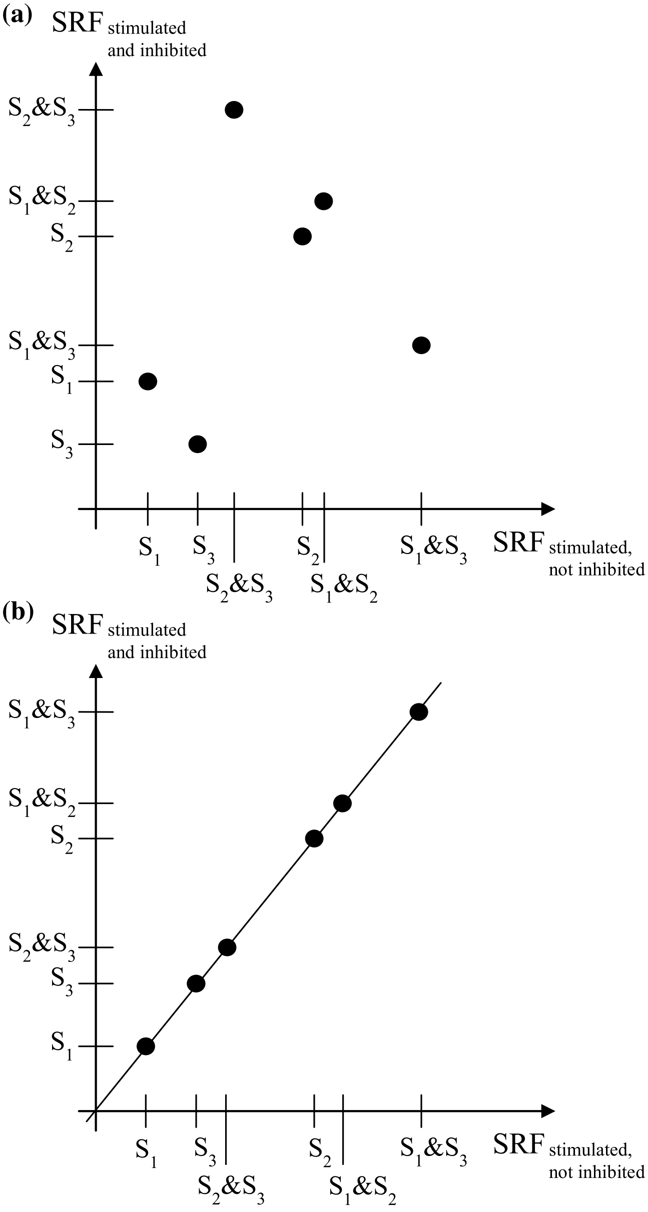

Fig. 3 Effects of inhibition on induced phosphorylation. (**a**) Phosphorylation of a protein, stimulated by stimuli 1-3 and combinations without inhibition (x-axis) and with inhibition (y-axis). The sequence of the phosphorylations for all combinations of stimuli is different with and without inhibition indicating that the inhibition acts upstream of the joint node of the stimulation pathways. (**b**) Phosphorylation of a protein, stimulated by stimuli 1-3 and combinations without inhibition (x-axis) and with inhibition (y-axis). The sequence of the phosphorylations for all combinations of stimuli is the same with and without inhibition. This indicates that the inhibition acts downstream of the joint node of the stimulation pathways. The slope of the resulting line quantifies the impact of the inhibitor on the response of the protein with respect to stimulation.

if the inhibition acts upstream of the merger node. However, if the sequence G'_S is rearranged by the impact of the inhibitor, then the inhibitor must affect pathways upstream of the merger node. □

As discussed, if the sequence G'_S is not rearranged by inhibition, other models may be possible as well. However, all other models will be more complex. According to Occam's razor we assume that the model with minimum complexity should be used as the first working hypothesis.

Corollary 1 If the $\gamma + \eta$ and the γ terms of the combinations of S are linearly correlated, then the sprays can be substituted by the coefficients directly. Moreover, in case of linear correlation, the function h is given by

$$h = a + b\gamma = b\gamma.$$

The zero-shift a vanishes as we set the shifts for all proteins to zero if no stimulation is applied. Then the parameter b quantifies the damping (if $b < 1$) or increasing (if $b > 1$) of the protein phosphorylation sensitivity with respect to the stimulations from the stimuli in S. This way we get quantitative, direct information on the impact of the inhibitor upon the cellular response to external stimuli.

Corollary 2 If the stimulated phosphorylations z of a protein satisfy Theorem 2 for all stimulations (S contains all stimuli, as depicted in Figure 3b, in contrast to Figure 3a), then the respective inhibition acts downstream of any 'merger' nodes where all stimulations join together. The inhibition may even act directly on the protein and not on intermediate functionalities along the signalling pathways between the receptors of the stimuli and the protein.

The advantage of this approach is the identification of structural features within the stimulation-inhibition network through combinations of either stimulation and/or inhibition. In contrast to the established network reengineering algorithms from dynamic data, which require time course measurements of phosphorylation, this approach utilizes static data. Even if dynamic data are unavailable, it allows functional network reengineering. If dynamic data are available, allowing a time resolution, then theorem 2 will hold as well. However, the time resolution may provide additional information with respect to the signalling network which cannot be unravelled by the stationary method described above.

Additionally, such an approach can be used to establish future models for clinically relevant applications where dynamics are slow such as lagging cellular responses to stimuli and inhibitors. Moreover, we can reconstruct these structures without any assumptions about the intermediate signalling network expressed in the cell or assumptions about signalling transfer mechanisms. Since we use only pathways or sub-networks that are directly extracted from data without any combinatorial optimization, it provides an efficient method to generate constraints needed to improve the stability and reliability of detailed network reconstruction approaches.

7 Second step: identification of pathway and subnetwork-specific inhibition

In order to establish a consistent method for topological sub-network identification we suppose that the functional network response F can be described by the combination of sub-networks with the respective response functions $f_1 \dots f_m$:

$$F = f_1 \circ \cdots \circ f_m \tag{9}$$

such that the effective parameters of the model shall be a linear combination of the respective parameters of the sub-models:

$$\underline{\gamma} = \sum_{k=1,\dots,m} a_k \underline{\gamma}^k, \tag{10}$$

$$\underline{\eta} = \sum_{k=1,\dots,m} a_k \underline{\eta}^k. \tag{11}$$

We can conclude that the parameters γ_i^k, η_i^k are mutually inclusive within the structure of the sub-network k. The parameters $a_1 \dots a_m$ describing the overall activation of the various sub-networks can then be extracted by using linear models to fit the data. However, due to the low dimensionality of the ‘effective’ model parameter space spanned by the parameters $\underline{\gamma}$, the direct feature extraction may become highly unstable, except where the parameter set $\{a_k\}$ is sparse or the parameters $\underline{\gamma}^k$, $\underline{\eta}^k$ representing the sub-models show highly restricted variability.

If none of the exceptions hold, the sub-network which is affected by the inhibitors can still be identified if the action of inhibitors is decomposed into a non-sparse component acting downstream of crosslinks of all stimulation pathways and sparse upstream contributions.

The effective shift of the parameters induced by the inhibition can then be quantified by:

$$F' = \sum_{k=1,\dots,n} \sum_{i_1,\dots,i_k} (\gamma_{i_1 \dots i_k} + \eta_{i_1 \dots i_k} y) x_{i_1} \dots x_{i_k} = \lambda \left(F + \sum_{k=1,\dots,n} \sum_{i_1,\dots,i_l} \eta_{i_1 \dots i_l} x_{i_1 \dots i_l} \right). \tag{12}$$

If only a few sub-networks are affected by the inhibition, then the set $\{\eta_{i_1 \cdots i_l}\}$ will be sparse, thus reducing the complexity of the parameter identification problem as well.

8 Third step: network clustering in multi-protein situations

If the functional pathways from the stimuli and inhibitors towards the proteins coincide for protein groups $G_1 \dots G_m$, then the responses of the proteins inside each group will coincide for all stimulations up to an individual sensitivity of each protein. The respective SRF’s are expected to show a high degree of mutual correlation.

In order to identify the respective protein groups, the correlation between SRF’s of proteins k and l on the stimuli from group $\Gamma : h_k(\Gamma), h_l(\Gamma)$ can be used:

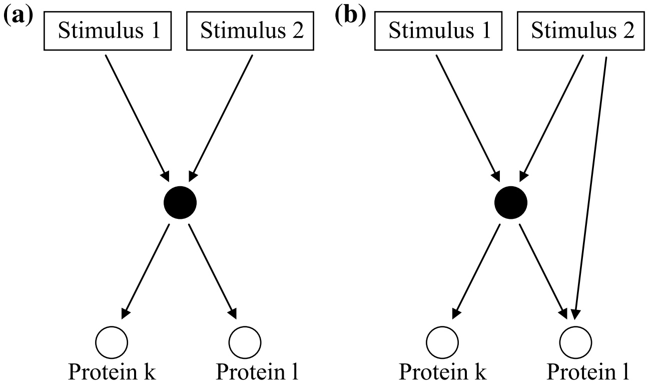

Fig. 4 Network structures and correlated stimulation. (**a**) The stimulations of proteins k and l are correlated. (**b**) The stimulations of proteins k and l are not correlated.

Corollary 3 If the stimulation response functions $h_k(\Gamma)$, $h_l(\Gamma)$ are correlated, then the functional pathways from the stimuli of the respective stimulation group S to the proteins k and l join in one control node (Figure 4a, b).

The stimulation of protein clusters with high mutual correlation of phosphorylation under a given set of stimuli is expected to depend on a joint functional node as shown in Figure 4a. The number of functional nodes (control nodes) controlling the entire protein network under a given set S of stimuli is equal to the number of correlation clusters.

9 Fourth step: identifying dominant stimulation pathways

In order to identify dominant stimulation pathways for each signalling protein, we assume that pathways with no affection on a protein do not significantly contribute to the respective activations (phosphorylation response). If a stimulus S_i (or set of stimuli) can be identified where, for all types of inhibition, the amount of the induced (or decreased) phosphorylation is significantly less than for the rest of the stimuli (both for the single-stimuli experiments as well as for the two-stimuli experiments where the stimulus S_i is involved), then we claim that this S_i does not affect the protein.

To ensure robust statistical results in case of only a few dominating stimuli leading to a poor combinatorial statistics, we developed the following workflow:

1. We quantify the difference between stimulations by the logarithms of the ratio of protein response to a single stimulus S_i with respect to the mean of all other single stimulations for each inhibition j, named $Q_{i,j}$. In order to estimate the significance of differential response between the stimulus i under all inhibitions, we calculate the p-value for the difference between the set $\{Q_{i,.}\}$ for all inhibitors j and the respective set $\{Q_{i',.}\}$ for all other stimuli i' and inhibitors j, using a two sided Wilcoxon ranksum test:
2. $p_i = \mathit{ranksum}(\{Q_{i,.}\}, \{Q_{i',.}\})$.

3. The same procedure is performed for the two-stimuli experiments
4. The p_i-values are calculated for the joint sets of single- and double stimulation experiments.

For each signalling protein and each type of stimulation a set of three p-values are calculated to indicate the difference between the impact of the stimulus S_i and the other stimuli. They allow the identification of either a dominant stimulus for a protein or, in contrast, that a stimulus does not affect the protein. As in our experimental setting only three independent dominating stimuli (for any protein) could be found, only six double-stimuli combinations could be analyzed leading to poor statistics. In order to achieve more stable results we therefore used all three tests as described above as an input to a weighted mean between the logarithms of the three p-values. This step can be omitted if more combinations are available.

Using p and Q-values, stimuli can be classified as either inactive or dominant. Inactive stimuli may display a mean p-value less than 0.05 and a negative Q-value whereas dominant stimuli display p-values less than 0.05 and positive Q-values (the other stimuli were set as inactive).

10 Conclusion

In this paper we describe direct network reengineering approaches which avoid the drawbacks of the established one-step network reengineering methods for networks with medium and high complexity. We show that the functional structure of the input-output functions of the entire system provides information with respect to the meso-scale topology of the functional network which is responsible for the behaviour of the system. Direct reengineering of the network topology using the functional structures of the input-output functions provides a mean to establish meso-scale network models which may be used as start point of further detailed models. So they can help to overcome the complexity issues of the existing approaches.

Although the algorithm for reengineering of the full network topology, presented in the first section of this paper, requires hierarchical tree structures, we have shown on the example of reengineering of signalling networks from protein kinase phosphorylation that the basic concept can be applied in a broad range of applications if problem specific adjustments are developed.

Protein kinase phosphorylation and dephosphorylation is critical in maintaining intracellular homeostasis and largely determines the cellular phenotype. The cellular environment can trigger intracellular phosphorylation signalling cascades to make needed adjustments, often through gene transcription. Cells showing deregulated phosphorylation due to protein misfolding, genetic mutations or other factors display aberrant phenotypes. Network re-engineering may serve as a useful tool in the elucidation and discovery of phosphorylation regulatory mechanisms as well as in the validation of observed cellular phenotypes.

We have presented adjustments of the rigorous network reengineering approach which allow the unravelling of important topological features of the functional network describing the interference of multiple stimuli and inhibitors on the global and sub-network scales. The approach is based solely on algebraic structures between the coefficients of multi-linear models as well as clustering, which can be effectively estimated from data. It therefore avoids some drawbacks of combinatorial network reconstruction approaches: the strong increase of computational efforts with network size as well as uncertainties with respect to the signalling transfer modes and 'hidden' signalling proteins. The meso-scale network topology may provide a starting point for further, more detailed network reengineering algorithms helping to improve their computational performance.

Combinatorial stimulation-inhibition analysis using a direct functional reengineering approach can aid in rapidly unravelling stable information about the coarse-grained structure of large-scale drug action from high throughput screening experiments. Future work in this area has the potential to provide more insight into the complex interactions between drug action and mutations in the signalling protein network in order to improve the R&D processes in pharmaceutical and biotech industry.

Competing interests

The author declares that he has no competing interests.

Acknowledgements I thank Leonidas Alexopoulos for providing the data used to validate the methods. I thank Nilou Arden and Angela Schuppert for helpful comments and assistance in preparing the manuscript.

References

1. Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5**, 101–113 (2004)
2. Camacho, D., Licona, P.V., Mendes, P., Laubenbacher, R.: Comparison of reverse-engineering methods using an in silico network. Ann. N.Y. Acad. Sci. **1115**, 73–89 (2007)
3. Mitsos, A., Melas, I.N., Siminelakis, P., Chairakaki, A.D., Saez-Rodriguez, J., Alexopoulos, L.G.: Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. PLoS Comput. Biol. **5**, e1000591 (2009). doi:10.1371/journal.pcbi.1000591
4. Schuppert, A.: Extrapolability of structured hybrid models: A key to the optimization of complex processes. In: Fiedler B., Groeger K., Sprekels J. (eds.) Proceedings of the International Conference on Differential Equations, 1-7 August 1999, Berlin, Germany, pp. 1135–1151. World Scientific Publishing, Singapore (1999)
5. Fiedler, B., Schuppert, A.: Local identification of hybrid models with tree structure. IMA J. Appl. Math. **73**, 449–476 (2008)
6. Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., De Moor, B.: Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics **22**, e184–e190 (2006)
7. Schuppert, A., Burghaus, R., v.Törne, C., Schwers, S., Stropp, U. Kallabis, H.: Method for identifying predictive biomarkers from patient data. Patent WO/2007/07/9875 (2006)
8. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Heidelberg (2009)
9. Rojas, R.: Theorie der neuronalen Netze. Springer, Heidelberg (1996)
10. Barron, A.R.: Approximation and estimation bounds for artificial neural networks. Mach. Learn. **14**, 115–133 (1994)

11. Schuppert, A.: New approaches to data-oriented reaction modelling. In: Jaeger W. (ed.) Proceedings of the 3rd Workshop on Modelling of Chemical Reaction Systems, July 28 1996, Heidelberg, Germany, pp. 1135–1151. Springer, Heidelberg (1999)
12. Mogk, G., Mrziglod, T., Schuppert, A.: Application of hybrid models in the chemical industry. Comput.-Aided Chem. Eng. **10**, 931–936 (2002)
13. Schopfer, G., Kahrs, O., Marquardt, W., Warncke, M., Mrziglod, T., Schuppert, A.: Semi-empirical process modelling with integration of commercial modelling tools. In: Puigjaner L., Espuña A. (eds.) Proceedings of ESCAPE-15, 29 May – 1 June 2005, Barcelona, Spain, pp. 595–600. Elsevier, Amsterdam (2005)
14. Kahrs, O., Marquardt, W.: The validity domain of hybrid models and its application in process optimization. Chem. Eng. Proc. **62**, 6222–6233 (2007)
15. Vitushkin, A.G.: On Hilbert's thirteenth problem. Dokl. Akad. Nauk SSSR **95**, 701–704 (1954)
16. Schuppert, A.: Method for the identification of pharmacophores. Patent EP 1451750 (2001)
17. Battle, A., Jonikas, M.C., Walter, P., Weissman, J.S., Koller, D.: Automated identification of pathways from quantitative genetic interaction data. Mol. Syst. Biol. **6**, 379 (2010). doi: 10.1038/msb.2010.27
18. Saez-Rodriguez, J., Alexopoulos, L.G., Epperlein, J., Samaga, R., Lauffenburger, D.A., Klamt, S., Sorger, P.K.: Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. Mol. Syst. Biol. **5**, 331 (2009). doi:10.1038/msb.2009.87
19. Alexopoulos, L.G., Saez-Rodriguez, J., Cosgrove, B.D., Lauffenburger, D.A., Sorger, P.K.: Networks inferred from biochemical data reveal profound differences in TLR and inflammatory signalling between normal and transformed hepatocytes. Mol. Cell. Proteomics (2010). doi: 10.1074/mcp.M110.000406