Defining Modeling Strategies for Systems Biology

Authors: FutureSysBio Workshop Participants:

Jörg Stelling, ETH, Switzerland (joerg.stelling@bsse.ethz.ch)
Pedro Mendes, Manchester University, UK (pedro.mendes@manchester.ac.uk)
Frank Tobin, Tobin Consulting LLC, USA (frank@tobins.org)
Edda Klipp, Humboldt University, Germany (Edda.Klipp@biologie.hu-berlin.de)
Riccardo Zecchina, Politecnico di Torino, Italy (riccardo.zecchina@polito.it)
Matthias Heinemann, ETH, Switzerland (heinemann@imsb.biol.ethz.ch)
Natasa Przulj, Imperial, UK (natasha@imperial.ac.uk)
Judith Wodke, Humboldt University, Germany (judith.wodke@crg.es)
Szymon Stoma, Humboldt University, Germany (szymon.stoma@gmail.com)
Hans-Michael Kaltenbach, ETH, Switzerland, (hans-michael.kaltenbach@bsse.ethz.ch)
Joachim Almquist, FCC Sweden (joachim.almquist@fcc.chalmers.se)
Andreas Raue, University of Freiburg, Germany (andreas.raue@fdm.uni-freiburg.de)
Jonas Hagmar, FCC, Sweden (jonas.hagmar@fcc.chalmers.se)
Marcus Krantz, GU, Sweden (marcus.krantz@cmb.gu.se)
Andrea Pagnani, Politecnico di Torino, Italy (andrea.pagnani@gmail.com)
Sven Nelander, GU, Sweden (sven.nelander@gu.se)
Marija Cvijovic, Chalmers, Sweden (marija.cvijovic@chalmers.se)
Mats Jirstrand, FCC, Sweden (mats.jirstrand@fcc.chalmers.se)

This document represents an outcome of the FutureSysBio Workshop: Defining Modeling Strategies, held in Göteborg, Sweden in January 2011. The purpose of this report is to inform and guide funding organisations, such as the European Commission services and the national funding bodies and foundations, of potential future directions of systems biology and hence possible funding priorities.

Based on this report a journal paper is in preparation and will be submitted to the international peer reviewed journal Nature Biotechnology.

Göteborg, June 2011

Introduction

Systems biology aims at creating mathematical models, i.e. computational reconstructions of biological systems and processes that will result in a new level of understanding - the elucidation of the basic and presumably conserved “design” and “engineering” principles of bio-molecular systems. Thus systems biology will move biology from a phenomenological to a predictive science. Mathematical modelling of biological networks and processes has already greatly improved our understanding of many cellular processes. However, given the massive amount of qualitative and quantitative data currently produced and number of burning questions in health care and biotechnology needed to be solved, is still in its early phases. The field requires novel approaches for abstraction, for modeling bioprocesses that follow different biochemical and biophysical rules, and for combining different modules into larger models that still allow realistic simulation with the computational power available today.

To address these questions, a focused Workshop - “Defining modeling strategies for systems biology”, was held in Göteborg, Sweden (January 20-21, 20011). The Workshop was organized with support of the FutureSysBio project (http://sysbio.se/FutureSysBio/), funded by the European Commission Coordination and Support Action that aims at shaping and predicting the future development of systems biology.

This workshop brought together experts from mathematics, theoretical physics, systems science, engineering and computer sciences, and industry. During the workshop three currently most prominent problems in systems biology were identify and discussed: (i) How to bridge different scales of modelling abstraction, (ii) How to bridge the gap between topological and mechanistic modelling, and (iii) How to bridge the web and dry lab gap.

Bridging Scales

Model integration

Mathematical models are often developed for a specific purpose and may be targeted at a small part of the overall cellular biochemical network. In the future, these models will become more useful if they can be integrated into larger, more unified models. For example, the Unicellsys project (1) aims at an integrated model of stress, nutrient signaling on cell cycle and growth of single yeast cell and its impact on population. The Virtual Liver project (2) endeavors the integration of processes from signaling to whole organ level. Yet another example is the integration of individual cell type models of prostaglandin synthesis and their combination into a physiological model to explain second messenger (3). There is also a major effort underway to develop accurate myocyte level ion channel models and link them with accurate 3D models of the heart to be able to model cardiac electrophysiological effects (4).

The complexity of biological systems is so extensive that separate groups of scientists are required to curate separate parts and pathways and to build the corresponding models. It is

useful to model the details in separate models of appropriate granularities and sizes and later merge them; this process is likely to be carried out repeatedly as new knowledge is incorporated in the separate parts. By necessity, the merging of models requires combining the data on which they are based, as well as recalibration and validation efforts for the joined, resulting models.

Benefits and purposes of combining models

The development of models, even if entirely new, benefits from the parallel exploration of alternative hypotheses. For example, a signaling pathway may not be well known and it may be useful to explore models of two alternative sub-networks (5). At some point these alternative versions of a model will need to be resolved and combined – which requires model integration. Since, model development is, in general, a time-consuming process one approach to building new models is to start from existing models. The re-use of existing models is facilitated by model databases such as JWS online (6), BioModels (7), the CellML model repository (8), BioMet Toolbox (9) and other such libraries. The ability to develop models in this way will enable better collaborative processes and robust models.

Generally, one could combine models with the same level of detail – the horizontal merging problem, or combine models with different levels of details – the vertical merging problem. Often both are needed. In horizontal merging, one of the aims is to create models that have larger scope than the components while keeping the same level of detail. For example, combining models of two signaling pathways provides a new model with more coverage of the network. In vertical merging, another aim could be to represent different levels of organization that are related to the same phenomenon. A model of viral infection may require combining a model of the molecular details of the infection of a cell with the details of the viral spread by blood circulation. Yet another aspect is the combination of models that are based on different mathematical frameworks. One might combine an ODE model with a Boolean model, or represent part of a larger model with differential equations and another part with a stochastic simulation algorithm, or combine a single cell model with a full 3D partial differential equation for the spatial behavior of many cells.

Indeed, this type of model building through composition may be necessary to explain more complicated biology. It may be necessary to include a portion of a model into a second, larger one that shares some parts of the biological details (the overlap problem), or to merge two disjoint sets of biology. Horizontal combination of two models at the same level of detail allows covering a number of processes that would be ignored when the two parts are modeled independently. Such separation means that the interaction between the components of the two subsets is lost, but it is needed explicitly when they are combined. Without that interaction, new biological phenomena may not be able to be modeled.

Challenges

There are four complementary major challenges that must be addressed:

(1) objective, consistent and unique descriptions of the biology that determines the appropriate level of model description,

(2) what level of description to adopt in the combined model, given that the two original models ones may be at different levels,

(3) the potential issue of the combination of the mathematical representations, and

(4) the ability to solve the resulting mathematical problem, especially if there are different approaches used.

Each challenge area varies significantly in details from modeling problem to modeling problem, but there are many common themes to be resolved. Where common tools and approaches can be used, this will only stimulate further research in this area and building of larger models. Some of the more salient aspects of these challenges are described below.

Coherency of the combination

Ensuring the coherence of the combined model with respect to the original models is a major challenge. When merging two (or more) models, the combination must make sense. To ensure coherency of the output model several aspects of model merging have to be considered (10). One issue is the matching of biologically identical compartments, species and reactions shared between the two models. This can be solved by the use of proper annotation (11). Even though overlapping species and reactions in the models to be merged can be matched, there may be still conflicts and ambiguity. For instance, species may have different initial concentrations or environmental conditions (e.g. pH). There can also be contradictions in parameter values or the structural form of rate expressions for some of the overlapping reactions. All such conflicts must be resolved either manually or with automated assistance, if such is available. For example, semanticSBML (11) and its predecessor SBMLmerge (12) are two tools that can assist in the merging of models encoded in Systems Biology Markup Language - SBML (13).

Once this definitional aspect of the component models is handled, it will be necessary to evaluate the properties of the output model. Are the key behaviors of the original models still intact? Does it matter if they are not? Are physical and chemical conservation or constitutive laws still obeyed in the merged model? Will the merging change model properties such as robustness, sensitivity, distribution of control in a pathway etc, in a way that was not intended? What about identifiability? Is this a property that can be lost, or perhaps gained, when models are merged? The merged model should of course be able to describe the data or behaviors described by the original models. But there may be further criteria added from the merging purposes. Almost certainly, this requires a re-calibration of model parameters. The merged model should also be able to describe some data or behavior that the original models could not – that is, it should serve the purpose for why it

was created in the first place. In some cases the answers to the questions above will lead to the conclusion that it is not meaningful to merge models.

Another aspect of the coherency issue is the resolution of the different purposes for which the component models were built. There are both biological and mathematical considerations that have to be resolved. It is possible that there may not be a straightforward resolution of the two components and merging is not possible. This area of coherency has several aspects:

- Accuracy (resolution, etc) between the components: e.g. one model has finite compartments and another uses the continuous space.
- The granularity of the biological description: e.g. EGF monomers can either be modeled explicitly, taking into account the various possible paths from monomer to polymer, or simply by two states (a lumped model).
- The model purposes and assumptions: e.g. one model uses an equilibrium assumption and the other has individual forward and backward reactions.
- A mathematical, chemical or biological approximation in one of the component models may not be valid in the merged model. While one model might use a quasi-steady-state assumption, the other model might explicitly use the transient dynamics on the corresponding time-scale.
- Parameter inconsistencies: different values of the same parameter, rate law, initial conditions in the component models). Even if a shared reaction is modeled by the same rate law, parameter values might differ, which also raises the problem of data fusion as one needs to determine whether two parameter or measurement values are identical, given that they were determined with different accuracy.
- Inconsistencies in structural form of the rate expressions: if a certain reaction is contained in two models, it might be modeled with different rate laws, e.g. the rate law might be simplified to Michaelis-Menten kinetics in one model while written explicitly with mass-action kinetics in the other.

The role of standards in model integration

Annotation of model elements is required for matching parts of the two models (14). Currently, both models and data are often available without proper annotation, which makes automation of their integration infeasible. In this particular area, systems biology strongly benefits from applying methods of computer science (15, 16). Various community standards are already available and partly integrated in common tools (17). There is a need for unifying syntax as well as semantics and the latter is crucial for the model integration problem. For example, SBML is the de-facto standard for exchanging dynamic models at the biochemical level, although there are many models represented in CellML, in other emerging or forgotten standards and some stored in none standardized format. With two models using different names for the same species, merging becomes very difficult. However, additionally annotating all species in a model using a unique database identifier (such as UniProt ID, CheBI) easily resolves this issue. Using Resource Description Framework (RDF) relations (18), modified versions of the same biochemical species (such as a protein

and its phosphorylated version) are also possible to annotate. SBML, in particular, provides advanced annotation capabilities that allow one or more unique database identifiers to each species, parameter, and reaction as well as expressing relationships among them. For example, several common (and even more uncommon) names for glucose-6-phosphate exist in metabolic models, including G6P, g-6-p, Gluc6P, and the like.

While there is some work in this area of standards – e.g. SBML Level 3 is being developed (19) - more research needs to be performed to automate where automation is safe and to identify inconsistencies between the component models being merged.

Data access

Most likely, a merged model needs to be calibrated again, preferably using the original data. This is critical, because the original data sets when brought together into the newly merged model may need to be interpreted differently due to the interaction between the component models. Moreover, each individual model is designed for a particular purpose using particular data and different parameter optimization strategies. Therefore, both models and data must be available and accessible. With data stored in relevant databases - e.g. *ArrayExpress* (20)*, caBig* (https://cabig.nci.nih.gov/)*, ImmunoblotDB* (https://www.immunoblot.de/)*,* etc. - the data can be standardized, especially when coming from different biological protocols, e.g., micro-array, 2D-gels, mass spectrometry, etc. A model should be annotated with the unique access numbers of the data used for its calibration. Clearly, with the composed model serving a new purpose, the original data might be re-interpreted, thereby giving new biological insight. It is not just standardization of data that is critical, but also its availability. If the merged model is to be recalibrated and revalidated, the legacy data therefore needs to be available (which currently is often not the case).

Computational methods

As models are merged, so are the underlying mathematical constructs that represent the components of the models. Considering that the components may approach their mathematical descriptions of biology differently, the resulting model may have to handle a different mathematical situation. Such mixtures may be straightforward and not difficult or they may present a variety of challenges to resolve. Such resolutions may be done at the level of the formulation of the rate equations or the basic biological descriptions as they are translated into the mathematical frameworks. Others may have a difficult numerical resolution, sometimes subtle, but often not, and brute force techniques may be woefully inadequate.

One common aspect of this condition is situation of hybrid models – two or more different mathematical frameworks that require very different numerical approaches towards equations integration. Such hybrids may need to not only resolve issues with scale, but the numerical robustness of the resulting combined integration may not always be as good as it was in the component problems. For example, some of these common situations are:

- Combining ordinary differential equations (ODEs) with stochastic kinetics (often represented as master equations, but sometimes as stochastic differential equations) (21)
- Developing and combining discrete (e.g., graph theoretic methods, Boolean networks) and continuous systems (e.g., ODEs)
- Merging of ODEs with partial differential equations (PDEs) is formally fairly straightforward, but may well have serious multiscale numerical problems as a result (4)

All models merged in such a way may potentially create a multiscale integration (22) problem. This occurs when different time or length scales are combined in the same problem and the numerical solution is often very inefficient unless appropriate numerical methods are used to improve the accuracy, speed, memory requirements, etc. For example, modeling osteoporosis may require a large dynamic range of time scales: the parathyroid hormone receptor binding dynamics (milliseconds), receptor internalization (hour), menstrual cycles (month), bone remodeling (1.5 months), aging (decades) and disease progression (decades). It is this mixture of fine grained and coarse grained phenomena that makes many problems numerically challenging. Multiscale integration is a ubiquitous problem in many areas of mathematical modeling and scientific computation, often computationally intensive, yet still not fully understood and an area of active research in the numerical analysis community (23). As better techniques become available, it will help spur model integration.

The activity of combining models will increasingly become a major part of systems biology. Indeed, it already is part of the collaboration between groups, especially when considering very large and complex biological systems, which, by necessity, require collaboration between multiple experimental and theoretical groups. While conceptually simple to understand, model integration poses a number of problems for which there are not yet very good solutions.

Bridging the Gap between Topological and Mechanistic Modeling

Problem Description

As previously mentioned, systems biology aims at generating understanding about how complex functions emerge from the interaction of biomolecules. Such understanding would allow us to predict the effects of drugs or to rationally design microbes for biotechnological applications. To do this, models that have predictive power are ultimately needed and thus these models will have the requirement to capture the respectively important biochemical (mechanistic) detail.

Unfortunately, we are still far from the mentioned vision. Instead the current state in mechanistic modeling is that models are still rather small, i.e. most often they only grasp a few molecular interactions or they can only *describe* known biological behavior, but have no predictive power. In fact, we are still far away of having predictive models for real applications, such as for medicine or biotechnology. Developing such models represents one of the biggest challenges for the field of systems biology. Our current modeling techniques are limited due to the following reasons:

(i) The problems are provably computationally intractable (NP-hard), so they must be solved approximately (i.e., heuristically) so that they would scale with the size of the data.

(ii) Models require a certain critical level of base knowledge (e.g. about the biological components and their interactions – knowledge, which we typically do not have for larger and more complex systems); it is unclear if methods can be proposed that can guide us towards finding the correct level of abstraction.

(iii) General lack of methods that can guide us in finding the correct level of abstraction.

Next to the problems on the modeling side, we also have the problem – particularly if we step to larger and more complex systems – that there is an inverse relation between the system size and the biological insight/knowledge that we have about the systems: While we have mostly a good understanding about the biological components, we have often only vague notions how the networks are wired. Also, some parts of biological networks might be well biochemically characterized, while for other parts we might know the connectivity of the components but not the molecular details. Although the current *omics* techniques might alleviate some of these problems, it is important to note that it is extremely unlikely that we will ever be able to observe all states in larger biological systems, and we thus will have to develop predictive models despite this uncertainty.

Thus, in order to deliver what is ultimately expected from systems biology – predictive models of (relevant) biological systems - we need novel modeling/computational approaches that can develop such models despite provable computational intractability and existing uncertainties about the molecular details of a given biological system. These models do not necessarily need to be correct in all mechanistic detail but they should have at least predictive power for certain functions.

Ways Ahead

Data Collection

To reduce and characterize uncertainty in dynamic models there is a need for more time series data. Comparing to the situation for classical applications of system identification there is still very little time series data available in most dynamic model efforts in systems biology. Since good time series data is crucial for building high quality well validated dynamic models today's rapid developments of novel and improved measurement techniques provide hope for the future.

Recently, a new, pragmatic approach was proposed for development of large-scale kinetic models on the basis of quantitative omics data acquired at steady-state conditions (24). Powerful *omics* techniques now allow to generate such data sets (25). Unlike classical parameter estimation, the so-called divide-and-conquer approach decomposes large models into small independent subproblems, for which – given measurements on state variables and rates – even analytical solutions can be derived. The solutions to these subproblems are joined to the complete space of global optima, which can be easily analyzed. The practical applicability of this approach was demonstrated by the recent development of a large-scale kinetic model for *E. coli* central metabolism and its enzymatic, transcriptional, posttranslation regulation was developed (26).

New Mathematical and Computational Methods

Various graph-theoretic models have been proposed for protein-protein interactions (PPI) networks (27). The problem is that network comparison is computationally intractable and hence easily computable heuristics are used to assess the fit of the model to the data. However, different heuristics might point to different models as being the best fitting and it is unclear how to resolve such discrepancies, even though integration of different heuristics and the use of machine learning classifiers on them has been proposed (28). The importance of identifying well-fitting models is that they not only help with algorithmic development, but can also provide understanding that can be biologically exploited. Even though current network models are still quite limited and need further refinement, they have already been successfully used, e.g., to de-noise PPI network data (29). Further development and refinement of such models would enable more sophisticated biomedical applications.

Another approach to comparing networks is that of network alignment (30, 31, 32), that has a promise to be as useful as sequence in discovering biological information, but from a new type of biological data, network topology, rather than sequence data. In addition to allowing comparison of network structure (i.e., topology) and discovery of the structurally conserved network parts, network alignment would also enable knowledge transfer from unannotated to annotated parts of aligned networks. Designing fast and reliable network alignment algorithms remains a foremost challenge, even though MI-GRAAL alignment algorithm has demonstrated that species as distant as yeast and human contain a surprising amount of network topology: 78% of yeast proteins participate in a connected sub-network that is topologically identical to that of the human PPI network (32). This suggests broad

similarities in PPI network topology across life on earth that may be due to exposure to the same selective pressures through evolution. Understanding and incorporating them into graph-theoretic models is a way towards mechanistic modeling of networked biological systems (33). Furthermore, new functional information can be extracted purely from network topology, but this requires the design and use of advanced mathematical concepts, rather than simple network statistics, such as node degrees or node adjacency information (27).

Incorporating uncertainty into mathematical models

When building mathematical models of biochemical networks, we can face at least three interlinked levels of uncertainty (34): (i) uncertainty in the network structure, e.g. which components and reactions to include and at which level of detail, (ii) uncertainty with respect to the choice of kinetics, e.g. mass action, Michaelis-Menten, reversible or irreversible, and (iii) uncertainty of the parameter values. Obviously, kinetic laws depend to some extent on the chosen granularity (e.g., two lumped mass-action reactions may conveniently be described by one Michaelis-Menten-type reaction) and the nature of the parameters is determined by the choice of the rate law.

The vast majority of today's mechanistic models does not incorporate or try to quantify the uncertainty inherent in the observed systems. Simulation approaches based on Monte Carlo methods and sensitivity analysis of ODEs can be used to understand implications on system behavior of uncertainty in parameters, initial conditions, and to some extent also network structure. However, it is not until proper descriptions of uncertainty are combined with experimental data we really can conclude what we have learned about the system and what remains uncertain in an integrated modeling and experimental study.

A formal way of specifying uncertainty is through the use of probability distributions. The most common use of concepts from statistics and probability theory in systems biology is to describe variability that stems from the stochastic nature of single reactions, cell to cell variation, noise and variability introduced in the process of taking single or repeated measurements, etc. However, probability distributions can also be used to characterize uncertainty in a broader sense, in particular in combination with measurement data. Assigning probability distributions to entities such as parameters, state variables, and system and measurement noise variables in dynamical models of biochemical reaction networks provides understanding of both the quantity itself and how precise or uncertain it is known. This approach becomes even more useful if one not only uses these *prior* probability distributions and studies their time evolution but also updates them using time series experimental data to obtain so called *posterior* distributions. This more elaborate way of representing the system under study naturally comes with the cost of more complex computations and requires more advanced mathematics but the gain of obtaining not only single numerical values for quantities but also some measure of quality or quantified uncertainty is a tremendous advantage. To fully exploit the described methodology there is a need to abandon systems of ordinary differential equations, ODEs, in favor of systems of stochastic differential equations, SDEs.

These are more complex mathematical objects than ODEs since they represent not only a single solution trajectory but a family of solutions, so called realizations, what statistical properties can be captured in terms of time varying distributions for the state variables. From a modeling perspective the mechanistic part of the SDEs is similar to the ODE description with the exception of additional stochastic variables representing uncertainty in parameters, reaction rates, or network structure. The computational effort to fully solve, i.e., compute the time varying prior/posterior probability distributions for the state variables, the SDEs for a realistic modeling problem are not to be underestimated and are in most cases not even tractable (requires the solution of a high dimensional partial differential equation known as the Kolmogorow forward equation). However, often quantities such as the peaks or mean of a distribution together with a measure of its spread is what is required to draw conclusions about how well the model characterizes a system under study. Hence, much of today's research in this area target methods for efficient and accurate approximate computations of the distributions or their derived features. This approach is what is taken in the field of nonlinear filtering with longstanding successful applications in various engineering fields and tools such as extended and unscented Kalman filters, particle filters, etc.

Model reduction/model expansion

An important aspect of modeling is to adapt the complexity of a model to the available knowledge and data to be used for estimation and model validation. Drafting a model from mechanistic hypotheses often leads to models with quite large number of parameters to be determined. The situation becomes even more problematic when certain gaps in the knowledge of how the system is wired are to be filled with partly competing or overlapping mechanistic hypotheses. To reduce the complexity of a model to better fit the available data situation various methods of model reduction can be applied. Here knowledge about orders of magnitude of different parameters as well as qualitative behavior of the system under study is important. Proper sensitivity analysis to rank the importance of certain parts of the model on the overall behavior is also an important tool to gain information about what could be removed from the model or simplified without sacrificing the model's descriptive power. Building dynamic models by aggregating subsystems developed separately is a natural method for obtaining comprehensive models. However, if such models are to be used for parameter estimation and model validation using time series data it is important to understand if the system is persistently excited by the perturbations or experimental situation at hand, i.e., if the generated data contain enough information to infer values of the parameters to be determined. A simple example is to consider a reaction rate given by a Michaelis-Menten expression. If the experimental conditions are such that this particular rate only operates in its linear range (and never becomes saturated) there will be no information in the data that let us distinguish between V_{max} and K_m and we will only be able to determine their ratio. The toolbox of model reduction techniques includes linearization, separation of time-scales leading to applications of singular perturbation theory, quasi-steady state assumptions, etc., which all need to be used on a more regular basis in modeling efforts seriously addressing comprehensive dynamic models.

Bringing Constraints into Models

Kinetic parameters in biochemical network models are thermodynamically dependent. K_m-values and maximal rates of reversible reactions are linked by the equilibrium constants of

those reactions. Along a series of reactions, the drop or gain of free energy provides a restriction for possible choices of kinetic parameters. If we would estimate the parameters independently from experimental data, we run into the danger of violating thermodynamic constraints. This flaw can be prevented by, e.g., parameter balancing (35) or tackled by network embedded thermodynamic analysis (36).

The accumulating information on parameter values as collected in databases such as BRENDA (37) or SABIO-RK (38) can be used to obtain typical distributions of parameter values specific for reactions, organisms or experimental conditions. Such distributions can be considered as prior probability densities (short: priors) for estimating parameter values in a Bayesian approach. New experimental data can be quantified as the likelihood function and a combination of both types of information can be used to calculate a posterior density distribution (39). This approach takes into account the uncertainty of parameter values as discussed above and makes at the same time use of information obtained in previous and unrelated experimental work.

Bridging the wet/dry lab gap

Discrepancy between data produced and data needed in models

As discussed in the previous sections, direct inference of network functionality from network topology is a non-trivial problem (40). In fact, even a completely defined topological network provides no more than a static view of the analysed system. The topology defines the possibilities within the network, but does not include the information on cause-effect that is absolutely required to understand the network's function. Tools are being developed for inference of missing parts of network topologies, but this is primarily an experimental issue and large efforts have been made to define reaction topology using large-scale methods. To determine the information flow through the network, the topology must be complemented with the cause-effect information, in what we can call a causal topology. This requires experimental data not only on which reactions may occur but also on how they are regulated. Also this is primarily an experimental challenge, and this data is typically not possible to generate with high-throughput methods. Hence, it likely needs to be addressed by dedicated experimentation; which would be time consuming but feasible.

A conceptually greater challenge is the discrepancy between the states used in mathematical models and the states explored experimentally. The clearest examples come from the global PPI studies in which experiments give information on whether a single pair of proteins interacts. The experiments may reveal a large number of such interaction partners, but give no information about which combinations may/must occur. In contrast, mathematical models typically contain highly defined specific states, in which reactions are defined for each possible combination of interaction partners (in fact, not defining the rule makes that combination impossible, which is another, implicit statement). Even dedicated experiments explore few such combinations, and the discrepancy between the combinatorial state space in the models and the relatively small state space explored experimentally leads to uncertainty and implicit assumptions when data is mapped on models. For example, say that we have experimental evidence that a protein has three interaction partners. A model would typically either include or exclude, explicitly, the single proteins, the tertiary complex and/or the intermediate states. However, the experimental evidence proves neither the absence nor the presence of any dimmers, trimmers or the tetramer. Hence, any definition of model structure would be based on guessing. While arbitrary model reductions may be required, it becomes an issue when these are implicit and cannot be distinguished from the real knowledge base of the model. Since it is inconceivable that the empirical data will ever cover the entire possible state space of the network, it will be necessary to adapt the modeling strategy to fit the available data, but also to include new types of data or scientific approaches such as molecular modeling. Hence, this is primarily a challenge for the modeling community.

The different scientific cultures lead to different approaches to - and expectations on - experimental design and model predictions. Importantly, models are only fleshed out hypotheses and cannot be used to verify the hypotheses they are based on - only check them for internal consistency. Failure to understand how – and why – models are built leads to overinterpretation; but also to collection of data that are less than optimal for modeling creation and validation. Data suitable for modeling might on its own not be the

most efficient way to reach biological conclusions but the iterative process between *in silico* hypothesis generation and experimental evaluation leads to even better understanding of the analysed system. Besides, qualitative data is of limited use for systems biology and quantitative biological measurements, as traditionally done are often based on measurements of relative changes. Even if highly precise, fold induction measurements are merely better than qualitative data for the purpose of model fitting and relatively little effort would be required to relate such measurements to entities per cell. In addition, it is imperative to stress the importance of time resolved data; not only to be certain to hit the peak change, but – more importantly – to be able to decipher causality which should be reflected in the temporal order. Overall, a stronger understanding of the modelling process would allow experimentalists to increase the usefulness and impact of the data they produce – with little or no additional cost.

Finally, even if the described data exists it can be difficult to use due to limited or inaccurate documentation/annotation. The implicit data obscured by faulty documentation are therefore not communicated from experimentalist to modeller. Standards for annotation of experimental data including detailed specification of experimental conditions are needed. Already a large effort has been made to define standards for specific subjects and these are collected in the Minimum Information for Biological and Biomedical Investigations (MIBBI, www.mibbi.org) (41). The definition of standards and their comprehensive usage is primarily a communication issue that can be bridged by cross-disciplinary training, as accomplished in bioinformatics or biophysics for example.

Finding common languages for different scientific communities

In order to bring systems biology to a higher level we see two possible strategies. We could start training undergraduate students in systems biology, trying to provide them with detailed knowledge in biology, chemistry, physics and computer science (42, 43). This bears the problem of educating scientists not specialized in any of the mentioned fields. The alternative would be to qualify highly skilled specialists with excellent understanding of their complementary disciplines and strong experience of interdisciplinary work; which would be crucial to reach a systems level of understanding. To do so, closer collaboration between different research areas would be necessary on all levels, i.e., as well for the general lab organization as for individual scientists. Different labs interested in answering the same biological questions need to establish common project planning, including the definition of all aspects to be addressed, and allowing for building multi-disciplinary networks. Within these networks elementary cross-disciplinary training at the beginning of a project will enable scientists to gain a better understanding of the biological problem and help defining a common language. Paired working on the same topic/project(s) with complementary tools will be very helpful to ensure smooth communication and trust between the involved researchers. Especially graduate students will benefit from learning by doing in interdisciplinary research projects with dual supervision (mentors from complementary disciplines).

References

1. Alberghina, L., and Cirulli, C. (2010) Proteomics and systems biology to tackle biological complexity: Yeast as a case study, *Proteomics* 24, 4337-4341.

2. Abbott, A. (2010) Germans Cook Up Liver Project, *Nature* 468.

3. Ten, J.H., Hazelton, W.D., Sparks, R., and Ulrich, C. (2005) A Michaelis-Menten-style model for the autocatalytic enzyme prostaglandin H synthase *Bull Math Biol*. 67(4):683-700

4. Bassingthwaighte, J., Hunter, P., and Noble, D. (2009) The Cardiac Physiome: perspectives for the future, *Experimental Physiology* 94, 597-605.

5. Flöttmann, M., Schaber, J., Hoops, S., Klipp, E., and Mendes, P. (2008) ModelMage: a tool for automatic model generation, selection and management, *Genome Inform.* 20:52-63.

6. Olivier, B., and Snoep, J. (2004) Web-based kinetic modelling using JWS online, *Bioinformatics* 20, 2144.

7. Novre, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J., and Hucka, M. (2006) BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems, *Nucleic Acids Research* 34, D689-D691.

8. Lloyd, C., Lawson, J., Hunter, P., and Nielsen, P. (2008) The CellML Model Repository, *Bioinformatics* 24, 2122-2123.

9. Cvijovic, M., Olivares-Hernández, R., Agren, R., Dahr N., Vongsangnak, W., Nookaew, I., Patil, K.R., and Nielsen, J. (2010) BioMet Toolbox: genome-wide analysis metabolism, *Nucleic Acid Research* 38: W144-9.

10. Liebermeister, W. (2008) Validity and combination of biochemical models, *Proceedings of 3rd International ESCEC Workshop on Experimental Standard Conditions on Enzyme Characterizations. Frankfurt: Beilstein Institute*.

11. Krause, F., Uhlendorf, J., Lubitz, T., Schulz, M., Klipp, E., and Liebermeister, W. (2010) Annotation and merging of SBML models with semanticSBML, *Bioinfromatics* 26, 421-422.

12. Schulz, M., Uhlendorf, J., Klipp, E., and Liebermeister, W. (2006) SBMLmerge, a system for combining biochemical network models, *Genome Informatics* 17, 62-71.

13. Hucka, M., Finney, A., Sauro, H., H. Bolouri, Doyle, J. C., H. Kitano, A. P. Arkin, Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmey, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novre, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Scha, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., J. Wang, and S.B.M.L.Forum. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models, *Bioinformatics* 19, 524-531.

14. Stein, L. (2001) Genome annotation: from sequence to biology, *Nature Reviews Genetics* 2, 493-501.

15. Bornholdt, S. (2005) Less is More in Modeling Large Genetic Networks, *Science* 310, 449-451.

16. Klipp, E., Herwig, R., Kowald, A., and Wierling, C. (2005) *Systems Biology in Practice: Concepts, Implementation and Application*, John Wiley & Sons.

17. Novre, N. L., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., P. Nielsen, Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D., and Wanner, B. L. (2005) Minimum information requested in the annotation of biochemical model (MIRIAM), *Nature Biotechnology* 23, 1509-1515.

18. (W3C), W. W. W. C. (2004) Resource Description Framework (RDF): Concepts and Abstract Syntax, *http://www.w3.org/TR/rdf-concepts/*.

19. Smith, L., and Hucka, M. (2010) SBML Level 3 Hierarchical Model Composition, *COMBINE 2010, Informatics Forum, University of Edinburgh, 09 October 2010*.

20. Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N., Brazma, A. (*2010*) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments *Nucleic Acids Research* 39: D1002-4.

21. Resat, H., Petzold, L., and Pettigrew, M. (2009) Kinetic Modeling of Biological Systems *Methods in Molecular Biology* 541*, Part 3*, 1-25.

22. Southerna, J., J. Pitt-Francisb, Whiteley, J., Stokeley, D., H.Kobashid, Nobes, R., Kadookad, Y., and Gavaghan, D. (2008) Multi-scale computational modelling in biology and physiology *Progress in Biophysics and Molecular Biology* 96, 60-89.

23. Barth, T., Chan, T., and Haimes, R. (2001) *Multiscale and Multiresolution Methods: Theory and Applications*, Springer.

24. Kotte, O., and Heinemann. (2009) A divide–and–conquer approach to analyze underdetermined biochemical models *Bioinformatics* 25: 519-525.

25. Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., Hirasawa, T., Naba, M., Hirai, K., Hoque, A., Ho, P.Y., Kakazu, Y., Sugawara, K., Igarashi, S., Harada, S., Masuda, T., Sugiyama, N., Togashi, T., Hasegawa, M., Takai, Y., Yugi, K., Arakawa, K., Iwata, N., Toya, Y., Nakayama, Y., Nishioka, T., Shimizu, K., Mori, H., and Tomita, M. (2007) Multiple high - throughput analyses monitor the response of E. coli to perturbations *Science* 316(5824):593-7.

26. Kotte. O., Zaugg, J.B., and Heinemann, M. (2010) Bacterial adaptation through distributed sensing of metabolic fluxes *Mol Syst Biol.* 6, 355.

27. Przulj, N. (2011) Protein-protein interactions: making sense of networks via graph - theoretic modeling, *Bioessays* 33(2).

28. Memisevic, V., Milenkovic, T., and Przulj, N. (2010) An integrative approach to modeling biological networks *Journal of Integrative Bioinformatics*, 7(3):120.

29. Kuchaiev, O., Rasajski, M., Higham, D.J., and Przulj, N. (2009) Geometric De-noising of Protein-Protein Interaction Networks *PLoS Computational Biology* 5:8.

30. Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W., and N. Przulj. (2010) Topological network alignment uncovers biological function and phylogeny *Journal of the Royal Society Interface* 7:1341-1354.

31. Milenkovic, T., Wong, W.L., Hayes, W., and Przulj, N. (2010) Optimal network alignment with graphlet degree vectors *Cancer Informatics* 9:121-137.

32. Kuchaiev, O., and Przulj, N. (2011) Integrative network alignment reveals large regions of global network similarity in yeast and human *Bioinformatics* 27(10):1390-6.

33. Przulj, N., Kuchaiev, O., Stevanovic, A., and Hayes, W. (2010) Geometric Evolutionary Dynamics of Protein Interaction Networks *Proceedings of the 2010 Pacific Symposium on Biocomputing (PSB), Big Island, Hawaii, January 4-8*

34. Schaber, J., Liebermeister, W., and Klipp, E. (2009) Nested uncertainties in biochemical models *IET Syst Biol* 3(1):1-9.

35. Lubitz, T., Schulz, M., Klipp, E., and Liebermeister, W. (2010) Parameter balancing in kinetic models of cell metabolism *J Phys Chem B.* 11(49): 16298 303

36. Kümmel, A., Panke, S., and Heinemann, M. (2006) Putative regulatory sites unraveled network-embedded thermodynamic analysis of metabolome data *Mol Syst Biol* 2:2006 0034.

37. Chang, A., Scheer, M., Grote, A., Schomburg, I., Schomburg, D. (2008) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009 *Nucleic Acids Research* 37:D588-592.

38. Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J. and Rojas, I. (2006) SABIO-RK: Integration and Curation of Reaction Kinetics Data Lecture Notes in Bioinformatics, 4075: 94-103.

39. Liebermeister, W., and Klipp, E. (2006) Bringing metabolic networks to life: Integration of kinetic, metabolic, and proteomic data *Theor Biol Med Mode* 15;3:42.

40. Cotterell, J*.,* and Sharp, J. (2010) An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients *Molecular Systems Biology* 6:425.

41. Taylor, C.F., Field, D., Sansone, S.A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.A., Bogue, M., Booth, T., Brazma, A., Brinkman, R.R., Michael Clark, A., Deutsch, E.W., Fiehn, O., Fostel, J., Ghazal, P., Gibson F., Gray, T., Grimes, G., Hancock, J.M., Hardy, N.W., Hermjakob, H., Julian, R.K Jr., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Le Novère, N., Leebens-Mack, J., Lewis, S.E., Lord, P., Mallon, A.M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J.M., Robertson, D.G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R.H., Schober, D., Smith, B., Snape, J., Stoeckert, C.J Jr., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J., and Wiemann, S. (2009) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project *Nat Biotechnol* 26, 889-9.

42. Botstein, D. *Integrated Science* - "Any budding researcher needs a foundation in several fields to be able to work on the most important problems confronting scientists today…".

43. Ledley, R. (1960) Report on The use of computers in biology and medicine *National Academy of Science - National Research Council.*