BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 4 2009, pages 519–525
doi:10.1093/bioinformatics/btp004

Systems biology

A divide-and-conquer approach to analyze underdetermined biochemical models

Oliver Kotte[1,2] and Matthias Heinemann[1,*]

[1]Institute of Molecular Systems Biology, ETH Zurich, 8093 Zurich and [2]Institute of Process Engineering, ETH Zurich, 8092 Zurich, Switzerland

Received on September 25, 2008; revised on December 12, 2008; accepted on December 30, 2008

Advance Access publication January 6, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: To obtain meaningful predictions from dynamic computational models, their uncertain parameter values need to be estimated from experimental data. Due to the usually large number of parameters compared to the available measurement data, these estimation problems are often underdetermined meaning that the solution is a multidimensional space. In this case, the challenge is yet to obtain a sound system understanding despite non-identifiable parameter values, e.g. through identifying those parameters that most sensitively determine the model's behavior.

Results: Here, we present the so-called divide-and-conquer approach—a strategy to analyze underdetermined biochemical models. The approach draws on steady state omics measurement data and exploits a decomposition of the global estimation problem into independent subproblems. The solutions to these subproblems are joined to the complete space of global optima, which can be easily analyzed. We derive the conditions at which the decomposition occurs, outline strategies to fulfill these conditions and—using an example model—illustrate how the approach uncovers the most important parameters and suggests targeted experiments without knowing the exact parameter values.

Contact: heinemann@imsb.biol.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Mathematical models are capable to reproduce and predict complex cellular responses, and are as such invaluable in advancing our understanding of living cells (Kitano, 2002a). Differential equation models are a common type of mathematical models that are especially suited to investigate the dynamic behavior arising from molecular interactions. Such models often contain many parameters whose values are uncertain but affect the simulated responses (Ingram *et al*., 2006). Therefore, these parameters are usually either directly measured or collectively estimated from experimental data, a process which most commonly involves the maximization of the maximum likelihood, often in the form of the minimization of a least squares distance between the simulation and the data, and the proper pre- and post-estimation diagnostics (Jaqaman and Danuser, 2006).

*To whom correspondence should be addressed.

Because differential equation models of biochemical systems typically contain many uncertain parameters whereas the availability of measurement data is often limited (van Riel, 2007), the parameter estimation problem is often underdetermined and remains a major bottleneck in the development of useful models. However, recent research suggests that the knowledge of all parameter values may not be necessary to obtain good predictions. First, the model structure can tightly constrain the possible responses such that astonishingly accurate predictions are possible even without estimating the parameters (Brown *et al*., 2004). Second, 'sloppiness' seems to be a universal property of systems biology models, meaning that most parameter values are unimportant because the system response is sensitively determined by the combination of only few parameter values (Gutenkunst *et al*., 2007). These observations lead to a question: 'If I do not have enough measurement data to identify my parameter values, can I still obtain a sound system understanding and derive good predictions despite of my problem being underdetermined?'

In this article, we present an approach that is capable to achieve exactly this, given that certain conditions on the underdetermined parameter estimation problem can be fulfilled. This so-called divide-and-conquer approach exploits a division of the estimation problem (not the model itself) into many independent subproblems of smaller dimension. The decomposition yields the *complete* solution space of the underdetermined estimation problem in a structured form, which facilitates a subsequent, systematic analysis of that solution space. The analysis can reveal the possible responses within the solution space and identify which parameters most sensitively determine these, and what effect a variation of these parameter values has on the response.

This article is structured as follows. First, we derive the necessary and sufficient conditions to trigger a decomposition into subproblems. Next, we show how this decomposition can be exploited through the divide-and-conquer approach, and discuss its application to real-world problems in systems biology. Then, to demonstrate the approach, we establish and analyze the complete solution space of an underdetermined model that overarches the metabolic and transcriptional regulation levels.

2 CONDITIONS FOR THE DECOMPOSITION

The divide-and-conquer approach exploits a decomposition of the global parameter estimation problem into smaller subproblems.

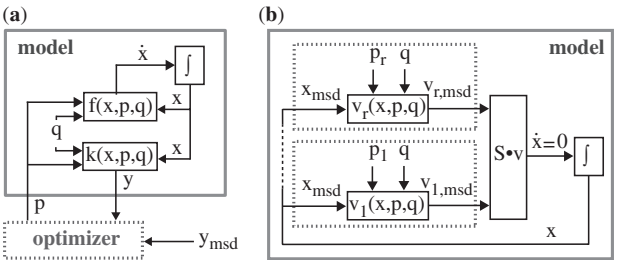

Fig. 1. Comparison of the divide-and-conquer approach with the conventional parameter estimation strategy. (**a**) The conventional strategy. The optimizer (dotted box) uses the output y of a model simulation to optimize the parameters p with respect to the experimental data y_{msd}. (**b**) The divide-and-conquer approach. Given certain conditions, the estimation problem decomposes into multiple independent subproblems (dotted boxes), for which the complete analytical solution spaces can be derived by solving a system of algebraic equations.

This decomposition occurs only when certain conditions are fulfilled. To derive these conditions, we successively specialize the general formulation of the global estimation problem to a formulation composed of independent subproblems. The conditions imposed during this specialization are the necessary and sufficient conditions to trigger the decomposition.

The general parameter estimation problem is stated as finding the set of parameters $\mathbf{p}$ within upper and lower bounds, $\mathbf{p}^U$ and $\mathbf{p}^L$, that minimizes a scalar cost function J. The cost function measures the goodness of the model prediction $\mathbf{y}(\mathbf{p},t)$ with respect to an experimentally measured dataset $\mathbf{y}_{msd}(t)$, and may include a diagonal scaling matrix $\mathbf{W}(t)$ with non-negative elements. The model prediction $\mathbf{y}$, which is calculated from the differential state variables $\mathbf{x}$ with the predictor function $\mathbf{k}$, is constrained by the system dynamics $\mathbf{f}$, which governs the time progression of $\mathbf{x}$. The problem can also include a set of parameters $\mathbf{q}$ that are not estimated. The mathematical formulation of this problem is:

Find $\mathbf{p}$ to minimize the sum of squared errors

$$J=\int_{t_0}^{t_f}(\mathbf{y}_{msd}(t)-\mathbf{y}(\mathbf{p},t))^T\mathbf{W}(t)(\mathbf{y}_{msd}(t)-\mathbf{y}(\mathbf{p},t))\mathrm{d}t \tag{1}$$

subject to the constraints

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}-\mathbf{f}\left(\mathbf{x},\mathbf{p},\mathbf{q},t\right)=0 \tag{2}$$

$$\mathbf{y}-\mathbf{k}\left(\mathbf{x},\mathbf{p},\mathbf{q},t\right)=0 \tag{3}$$

$$\mathbf{x}(t_0)=\mathbf{x}_0 \tag{4}$$

$$\mathbf{p}^L\le\mathbf{p}\le\mathbf{p}^U. \tag{5}$$

Mathematically, this is a nonlinear optimization problem with differential–algebraic constraints, which is commonly solved using a suitable optimizer (Fig. 1a).

Before we specialize Equations (1–5) to a formulation composed of independent subproblems, we take all the summands appearing in $\mathbf{f}$ and list them in a rate vector $\mathbf{v}$. We define a stoichiometric matrix $\mathbf{S}$ such that $\mathbf{f}=\mathbf{S}\cdot\mathbf{v}$, and rewrite Equation (2) to

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}-\mathbf{S}\cdot\mathbf{v}\left(\mathbf{x},\mathbf{p},\mathbf{q},t\right)=0. \tag{6}$$

As most biological measurements are taken at discrete time points, we limit our investigation to cost functions of the form

$$J=\sum_{i=1}^{m}(\mathbf{y}_{msd}(t_i)-\mathbf{y}(\mathbf{p},t_i))^T\mathbf{W}(t_i)(\mathbf{y}_{msd}(t_i)-\mathbf{y}(\mathbf{p},t_i)) \tag{7}$$

where t_i is the i-th of m measurement time points.

Next, we specialize the general formulation given by Equations (3–7) through imposing a condition on the measurement dataset.

Condition 1. At all measurement time points t_i, the measurement dataset must consist of all differential state variables x and all rates v, such that

$$\mathbf{y}_{msd}^T=\left(\mathbf{x}_{msd}^T\,\mathbf{v}_{msd}^T\right), \tag{8}$$

which implies

$$\mathbf{k}^T=\left(\mathbf{x}^T\,\mathbf{v}^T\right). \tag{9}$$

In a later section, we comment on how these conditions can be fulfilled in real-world problems. We continue with including a condition on the model structure.

Condition 2. The model structure must allow for an *exact* fit to *all* measurement data points, such that

$$\begin{aligned}J=\textstyle\sum_{i=1}^{m}&\begin{pmatrix}\mathbf{x}_{msd}(t_i)-\mathbf{x}(\mathbf{p},t_i)\\ \mathbf{v}_{msd}(t_i)-\mathbf{v}(\mathbf{p},t_i)\end{pmatrix}^T\mathbf{W}(t_i)\\ &\begin{pmatrix}\mathbf{x}_{msd}(t_i)-\mathbf{x}(\mathbf{p},t_i)\\ \mathbf{v}_{msd}(t_i)-\mathbf{v}(\mathbf{p},t_i)\end{pmatrix}=0.\end{aligned} \tag{10}$$

As the sum-of-squares J is strictly non-negative, a parameter set leading to $J=0$ must be a global optimum. In practice, this condition requires an underdetermined estimation problem.

With these conditions fulfilled, the global estimation problem reduces to finding a solution $\mathbf{p}$, $\mathbf{p}^L\le\mathbf{p}\le\mathbf{p}^U$, that satisfies

$$v_j\left(\mathbf{x}_{msd}(t_i),\mathbf{p},\mathbf{q},t_i\right)-v_{msd,j}(t_i)=0 \tag{11}$$

for $i=1\ldots m$ and $j=1\ldots r$, where r is the number of components in $\mathbf{v}$. Thus, Equation (11) comprises $m\cdot r$ equations. As $\mathbf{p}$ is the only unknown, this in practice underdetermined equation system can be solved to derive the complete solution space of $\mathbf{p}$. Because these equations are coupled solely through $\mathbf{p}$, this potentially very large-dimensional solution space can be decomposed into many smaller-dimensional subspaces by removing the coupling through a further condition on the model structure.

Condition 3. The parameters to be estimated, $\mathbf{p}$, consist of $1\le s\le r$ disjunct sets

$$\mathbf{p}^T=(\mathbf{p}_1^T\,\mathbf{p}_2^T\ldots\mathbf{p}_s^T) \tag{12}$$

such that each set $(\mathbf{p}_k,\mathbf{q})$ fully parameterizes a subset of the rate equations. If disjunct subsets do not exist ($s=1$), then the complete solution space can be derived through Equation (11) but not decomposed into smaller-dimensional subspaces. If each p_j appears in only one rate equation ($s=r$), then the solution space can be maximally decomposed into pairwise independent subspaces.

The parameter estimation problem is thus reduced to finding $\mathbf{p}_k$ such that

$$v_j\left(\mathbf{x}_{msd}(t_i),\mathbf{p}_k,\mathbf{q},t_i\right)-v_{msd,j}(t_i)=0 \tag{13}$$

for $i=1\ldots m$, $j=1\ldots r$ and $k=1\ldots s$.

As Equation (13) comprises $m\cdot r$ algebraic equations in s decoupled sets, the global problem has been successfully decomposed into independent subproblems of smaller dimension.

Equation (13) states that a parameter set $\mathbf{p}$ is a global optimum if all its subsets $\mathbf{p}_k$ parameterize the rate equations v_j in which they appear such that the measured rates are reproduced exactly for the measured states, as illustrated in Figure 1b. Note that a verbose description of this section can be found in the Supplementary Material.

3 THE DIVIDE-AND-CONQUER APPROACH

The divide-and-conquer approach is the consistent exploitation of the decomposition of the global estimation problem into independent subproblems. This approach encompasses both the derivation of the complete solution space of an underdetermined problem, and the efficient analysis of that space.

The derivation of the complete solution space can be structured into three steps. In the first step, complete sets of state and rate data are obtained to fulfill Condition 1. In the second step, the degree of decomposition is chosen according to Condition 3. In the third step, the complete solution space of the underdetermined problem is derived by fulfilling Condition 2. Here, it is important to understand that the divide-and-conquer approach does not give the solution to the estimation problem as upper and lower bounds on single parameters values, but as an analytic description of the parameter subspaces that together form the complete solution space.

Once the complete solution space is analytically known, it can be quickly and systematically analyzed to obtain a sound understanding of the possible system responses within this parameter space. It is also important to note that the divide-and-conquer approach first neglects the noisy nature of biochemical measurements, then derives insights based on the assumed 'perfect' dataset, and finally assesses the robustness of the derived insights with respect to data noise.

The following three sections discuss the derivation of the complete solution space in detail. After that, we illustrate the systematic analysis of that solution space with an example model, and show how the robustness of the obtained insights with respect to data noise can be assessed.

3.1 Step 1: obtaining complete steady state datasets

Recent developments in high-throughput experimental methods provide us with ever more comprehensive measurement datasets in steady state conditions, e.g. of the cell's proteome and metabolome [as presented e.g. by Ishii *et al*. (2007)]. However, such datasets, whether assembled from literature and/or own measurements, are to date often incomplete. To fulfill Condition 1, which requires complete datasets, we therefore propose to extend incomplete measurement datasets to larger sets of *observed* data. Depending on the problem, these datasets of observables can be complete and therefore applicable for the subsequent parameter estimation [as suggested by Gadkar *et al*. (2005)].

To obtain a complete set of observables, the unmeasured data can be inferred from the measured data with the help of models, simple or sophisticated, that are based on biological knowledge. For instance, a computational model can be used to observe metabolic reaction rates from measured ^{13}C-labeling patterns of amino acids. Similarly, an incomplete set of measured metabolite concentrations can be extended to a complete set of observed metabolite concentrations by using network-embedded thermodynamic (NET) analysis (Kümmel *et al*., 2006), which is capable to infer unmeasured metabolite concentrations within certain limits.

To observe missing rates, the consistency condition

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v_+ - v_- = 0, \tag{14}$$

can be exploited, which states that in steady state, the sum of all compound production rates v_+ must equal the sum of all compound consumption, dilution and degradation rates, v_-. If one of the rates in v_- or v_+ is unknown, then Equation (14) can be applied to determine the missing rate from the known rates. This equation underlies flux balance analysis, which is capable to observe metabolic reaction rates from physiological data using a stoichiometric metabolic network model. On a smaller scale, Equation (14) can be used to, for instance, observe a steady state protein production rate from known protein degradation and dilution rates.

If more than one of the rates in v_- or v_+ is unknown, then some of the missing rates can be observed with simple linear models. For instance, a compound dilution rate can be observed through $v_{\mathrm{dil},x} = \mu \cdot x$ with x and the growth rate μ known, or a compound degradation rate through $v_{\mathrm{degr},x} = k_{\mathrm{degr}} \cdot x$ with x and the degradation rate constant k_{degr} known. Note that if these simple linear models contain parameters of the dynamic computational model (e.g. μ and k_{degr}), then these parameter values are already fixed and may not appear in the estimation problem of the divide-and-conquer approach (i.e. these parameters are included in $\mathbf{q}$, not $\mathbf{p}$).

In the unlikely case that all of the rates in Equation (14) are measured, then these measurements will most likely not add up to 0 due to measurement errors, implying that the model cannot reproduce the steady state *exactly* (Condition 2 is violated). Therefore, these measurement errors must be 'corrected', e.g. by minimally adjusting the data to fulfill Equation (14). Note that at a later stage, it can be assessed if such data 'correction' and the neglected measurement noise sensitively affect the insights obtained with the assumed 'perfect' dataset.

Lastly, datasets such as transcriptome data are often acquired only as relative measures. Fortunately, relative data of a compound concentration x is sufficient if in the model x appears always paired with a multiplicative parameter p. Then, rate equations of x are of the form $r = f(p \cdot x)$, and the parameter estimation of p can correct for an arbitrarily chosen absolute concentration of x. Such situations occur, for instance, with x as an enzyme or mRNA concentration, p as the respective rate constants and r as metabolic reaction rate or translation rate, respectively.

3.2 Step 2: choosing the degree of decomposition

To increase the degree of decomposition, and thereby the transparency of the later solution space to the modeler, the parameters are divided into three disjunct sets, $\mathbf{p_A}$, $\mathbf{p_B}$ and $\mathbf{q}$, such that $\mathbf{p_A}$ contains those parameters that appear in only one equation v_j, $\mathbf{p_B}$ contains those that appear in more than one equation v_j and $\mathbf{q}$ contains the parameters that are not subject to the estimation. In biochemical models, parameters typically have a specific mechanistic meaning and as such tend to appear in only one equation v_j. Therefore, $\mathbf{p_B}$ usually contains only few parameters but is not necessarily empty.

The maximal degree of decomposition ($s = r$) is reached when $\mathbf{p_B} = \emptyset$. If $\mathbf{p_B} \neq \emptyset$, then the degree of decomposition can be increased by excluding a parameter $p \in \mathbf{p_B}$ from the estimation, which moves this parameter from $\mathbf{p_B}$ into $\mathbf{q}$. For instance, if p is known to

be poorly identifiable, which is probable as p appears in multiple underdetermined equations, then assuming a literature value for p may be the better alternative anyway. If this is not justified, then all equations containing p form one subproblem.

3.3 Step 3: determining the complete solution space

To obtain the complete parameter space that reproduces the data exactly, the complete datasets from step 1 are plugged into the kinetic model equations. These equations thus become algebraic functions of the parameters. Due to the decomposition of the parameter space performed in step 2, this set of functions consists of independent subsets. Each of these independent subsets comprises a system of equations with α_j unknowns (the number of parameters to be estimated in the subset) and β_j constraints (the number of equations derived from plugging the data into the subset's kinetic model equations). The estimation problem is thus decomposed into independent subproblems, which correspond to solving each subset's system of algebraic equations. These subproblems can be either overdetermined, exactly determined, or underdetermined:

- If $\beta_j > \alpha_j$, then the j-th subproblem is overdetermined, which prohibits an exact fit to the data and therefore the application of the divide-and-conquer approach. However, this case is unlikely to occur, as complete steady state datasets are usually obtained for only few conditions, whereas rate equations typically contain multiple uncertain parameters. An exception to this rule is first-order kinetics, which usually either approximate higher-dimensional kinetics and can be substituted by those, or do not contain any parameters to be estimated, such as a linear dilution rate equation with known growth rate μ.
- If $\beta_j = \alpha_j$, then the j-th subproblem is exactly determined. If a solution exists, it is unique—the data constrains the α_j-dimensional parameter space to a single point.
- If $\beta_j < \alpha_j$, then the j-th subproblem is underdetermined. If a solution exists, the data reduce the α_j-dimensional parameter space to a solution space of dimension $\alpha_j - \beta_j$.

If for any subproblem j, an exact fit to the data cannot be achieved and the subproblem is not overdetermined, then a discrepancy between the model structure and the available data has been identified and localized. The discrepancy can be removed either by changing the model structure, e.g. to a rate law that reproduces the data, or, if there is reason to doubt the quality of the data, by resorting to another set of measurements.

When all discrepancies between the model structure and the data are removed, then Condition 2 is fulfilled. The solutions to the subproblems are then joined to the global solution Ω of the parent parameter estimation problem. If all subproblems are exactly determined, the solution is a single point. In most cases, however, at least one of the subproblems is underdetermined, and the solution is therefore a multidimensional space. A significant advantage of the divide-and-conquer approach is that it yields the *complete* solution space of the parameters in the form of *analytically* known manifolds on which all global solutions are located. Using an example, we next illustrate how this analytically known solution space can be efficiently and thoroughly analyzed to derive a sound system understanding.

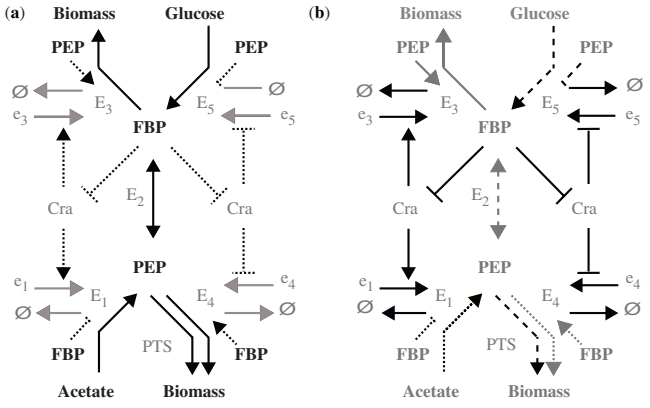

Fig. 2. The example model. (**a**) The topology of the example model. Metabolites and metabolic reactions are black; genes (e_i), proteins (Cra, E_i) and protein production and degradation rates are gray; regulatory interactions are dotted. (**b**) Decomposition of the kinetic equations into six independent subproblems. Subproblem 1 (according to Table 1); regular black lines; 2, dotted black lines; 3, dashed black lines; 4, regular gray lines; 5, dotted gray lines; and 6, dashed gray line.

4 EXAMPLE

We illustrate the application of the divide-and-conquer approach by deriving and analyzing the complete solution space of the small model system depicted in Figure 2a. This model describes a core section of *Escherichia coli*'s central metabolism and covers allosteric and transcriptional regulation. It simulates the reversal of carbon flow through the Embden–Meyerhoff-Pathway, which is required to switch between growth on glycolytic and gluconeogenic substrates, e.g. glucose and acetate.

The model consists of five enzymes (E_i), one transcription factor (Cra), four genes (e_i) and two metabolites (phosphoenolpyruvate, PEP, and fructose bisphosphate, FBP). It contains 16 rates: 4 enzyme production rates, 6 compound dilution and degradation rates, 5 metabolic reaction rates and 1 transcription factor–metabolite binding rate (of Cra to FBP). It further includes a simplified representation of the phosphotransferase system (PTS), which couples the uptake and phosphorylation of extracellular glucose to the conversion of PEP to pyruvate. The model is centered on the transcription factor Cra, whose activity controls the expression of four of the modeled enzymes and is itself controlled by the metabolite FBP. For the model equations, refer to the Supplementary Material.

4.1 Derivation of the complete solution space

To estimate the parameters of the kinetic equations and to analyze the complete solution space with the divide-and-conquer approach, we apply the steps presented in the previous section.

Step 1 is fulfilled with the complete state and rate measurement datasets listed in Supplementary Table S1, which were obtained from literature for balanced growth on either the glucolytic substrate glucose or the gluconeogenic substrate acetate.

In step 2, the degree of decomposition is chosen. Overall, the system comprises 39 parameters. Of these, the growth rate μ and the concentrations of the carbon sources *Glucose* and *Acetate* are directly measured, literature values are assumed for

Table 1. Decomposition of the global estimation problem into six independent subproblems, and the division of the parameters into free and dependent parameters

Sub-problem	α	β	$\alpha-\beta$	Free parameters	Dependent parameters
1	10	8	2	$K_{Cra,\mathrm{FBP}}$ n_{Cra}	$v_{e1,max}$ $v_{e3,max}$ $v_{e4,max}$ $v_{e5,max}$ K_{e1,Cra_A} K_{e3,Cra_A} K_{e4,Cra_A} K_{e5,Cra_A}
2	4	1	3	L_{E_1} $K_{E_1,\mathrm{PEP}}$ $K_{E_1,Acetate}$	k_{cat,E_1}
3	4	1	3	L_{E_5} $K_{E_5,\mathrm{FBP}}$ $K_{E_5,Glucose}$	k_{cat,E_5}
4	4	2	2	L_{E_3} $K_{E_3,\mathrm{FBP}}$	k_{cat,E_3} $K_{E_3,\mathrm{PEP}}$
5	4	2	2	L_{E_4} $K_{E_4,\mathrm{PEP}}$	k_{cat,E_4} $K_{E_4,\mathrm{FBP}}$
6	4	2	2	$K_{E_2,\mathrm{PEP}}$ $K_{E_2,\mathrm{FBP}}$	$v_{E_2,f}$ $v_{E_2,r}$

α: Number of parameters; β: number of constraints; $\alpha-\beta$: degrees of freedom.

ρ and k_{degr}, and n_{E_i} (the number of subunits in the quaternary structure of an enzyme) is set to four for all tetrameric enzymes. Therefore, $\mathbf{q}=[\mu, Glucose, Acetate, \rho, k_{\mathrm{degr}}, n_{E_1}, n_{E_3}, n_{E_4}, n_{E_5}]$. Of the 30 parameters in $\mathbf{p}$, only $K_{Cra,\mathrm{FBP}}$ and n_{Cra}, which describe the binding of FBP to Cra, appear in more than one rate equation. Therefore, all rate equations containing these two parameters are merged to a composite subproblem. Table 1 and Figure 2b summarize the resulting six independent subproblems into which the global estimation problem decomposes.

In step 3, the parameters of the six subproblems are constrained by a system of algebraic equations. Each of these equations reduces a subproblem's degree of freedom by one, and can be rearranged such that one of the subproblem's parameters becomes dependent on the others. This process is described in detail in the Supplementary Material, with Table 1 summarizing the resulting (arbitrary) division into free and dependent parameters. The complete space of global solutions Ω comprises all parameter vectors within admissible bounds $\mathbf{p}^L \le \mathbf{p} \le \mathbf{p}^U$ that solve the obtained system of algebraic equations [Equations (S11–S17) in Supplementary Material], i.e. are located on the solution manifolds. Because of the division into free and dependent parameters, a global solution can be easily generated by choosing a combination of free parameter values and calculating the dependent parameters with Equations (S11–S17) of Supplementary Material. Note that the identification of these parameter dependencies is an active research area by itself (Hengl *et al*., 2007; Liebermeister and Klipp, 2005).

4.2 Analysis of the model behavior

To analyze the model behavior, we exploit the decomposition of the solution space into independent and analytically known manifolds.

First, we verify that the solution space was correctly determined and that the two measured steady states exist and are stable. To do so, we randomly generated 1000 global solutions within admissible parameter bounds ($0.1 \le n_{Cra} \le 4$, $0.1 \le K_i \le 10$, $1 \le L_i \le 10^7$) by assigning random values to the free parameters and calculating the dependent parameters with Equations (S11–S17) of Supplementary Material. We then simulated the model with initial conditions equal to the two measured conditions. As expected, we obtained perfectly level lines for both conditions and all compound concentrations (data not shown). Therefore, as all of the sampled parameter vectors reproduce the steady state measurement data exactly, the solution space has been correctly determined. Furthermore, because the simulations remain in the steady states indefinitely, we can conclude that both steady states exist and are stable (at least for these 1000 samples).

Next, we obtain a general overview of the possible system responses to a sequence of perturbations. This is necessary because although all global solutions reproduce the *stationary* measurement data exactly, different parameter combinations may lead to very distinct *dynamic* responses. As our example model describes the reversal of carbon flow through a metabolic pathway, we are interested in a complete picture of the possible dynamics during such flux reversals. Therefore, we chose to perturb the system by switching the carbon source from acetate to glucose at $t=0$ h, and back to acetate at $t=35$ h.

Figure 3a shows the simulated responses of the FBP concentration to these perturbations (with the previously sampled parameters). While all simulations successfully adapt from glucose to acetate, the dynamic behavior of the adaptation from acetate to glucose varies widely and can be categorized into two response families: responses of Family A converge to the measured steady state on glucose at 6.6 μmol/gDW, and responses of Family B converge to a second steady state on glucose with a parameter-dependent concentration below 4.5 μmol/gDW. Therefore, when adapting from the measured steady state on acetate, the measured steady state on glucose—although it has been verified to exist and be stable—is only attractive for the parameter subset belonging to Family A. In this context, note that the existence of the second steady state on glucose was identified by a sampling strategy and could have remained unnoticed if merely a point solution had been determined.

Next, we exploit the division into two response families to identify those parameters that most sensitively shape the dynamic response. If a parameter sensitively shapes the response, its value should determine the response family. We therefore compared the parameters' distributions in the two response families using the Student's *t*-test. The parameters' *P*-values (Supplementary Table S2) span many orders of magnitude, with $K_{E_2,\mathrm{FBP}}$ and $v_{E_2,f}$ exhibiting extremely low *P*-values. Therefore, the distribution of these two parameters is significantly different between the two response families. Using the derived solution manifolds, this result can be graphically illustrated. In most cases, as in Figure 3b, the parameter values of either response family are evenly distributed across the manifold. However, in the case of $K_{E_2,\mathrm{FBP}}$ and $v_{E_2,f}$ (Fig. 3c and Supplementary Fig. S1), the parameter values of response Family B are clustered in a particular region of the manifold. We therefore suspect that these two parameters dominantly shape the response.

To test if $K_{E_2,\mathrm{FBP}}$ and $v_{E_2,f}$ indeed sensitively determine the system response, we first set the free parameter $K_{E_2,\mathrm{FBP}}$ to its maximal admissible value and randomized all other free parameters as before. By fixing only this free parameter, we were able to constrain the possible responses tightly: all trajectories rapidly converge to the measured steady state on glucose, i.e. belong to Family A (Fig. 3d). We then arbitrarily selected one of these trajectories and kept its parameters constant with the exception of $K_{E_2,\mathrm{FBP}}$, which we decreased stepwise across its entire admissible range. By varying only this free parameter (and the two dependent

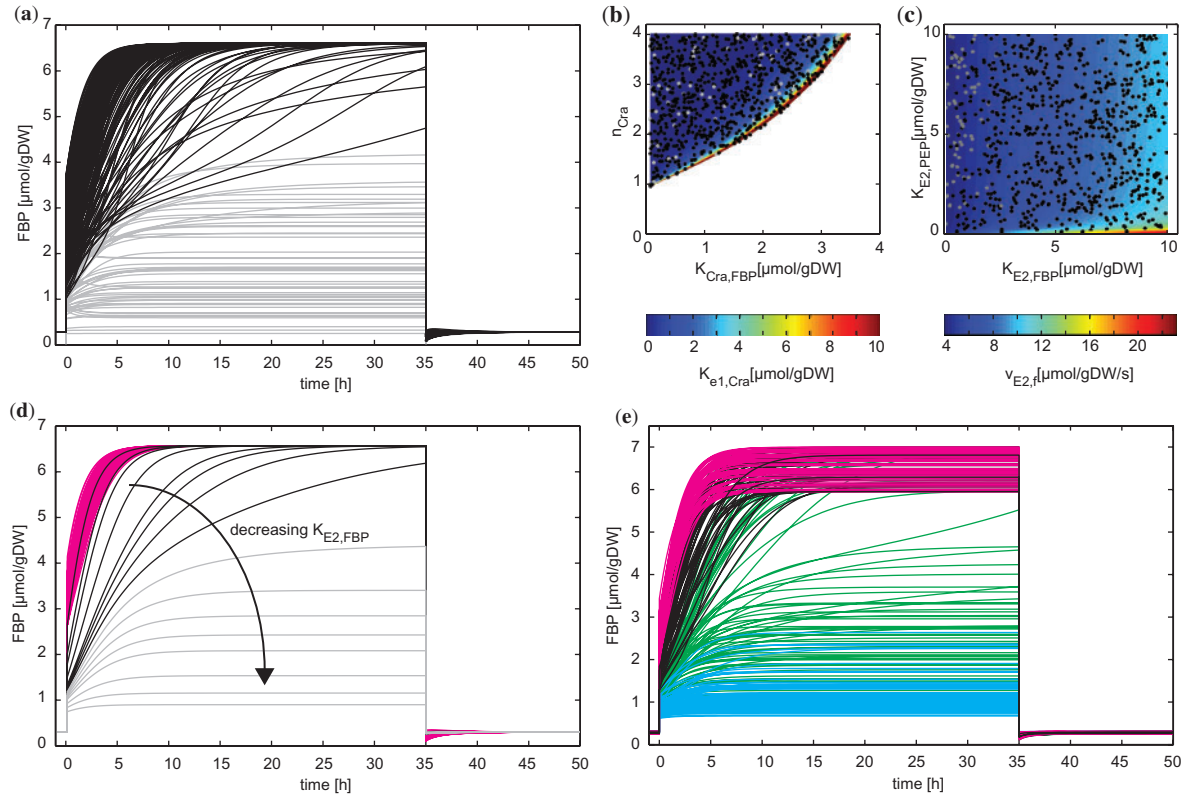

Fig. 3. Analysis of the solution space. (**a**) Simulated responses of the FBP concentration to carbon source shifts from acetate to glucose (at 0 h) and back to acetate (at 35 h). A random sampling of 1000 parameter vectors $\in \Omega$ reveals that the possible system responses vary widely. The responses can be classified into two response families: 63% of the simulations converge (on different trajectories) to the measured steady state on glucose at 6.6 µmol/gDW (black lines, Family A). The remaining 37% converge (on different trajectories) to a second steady state at a parameter-dependent concentration <4.5 µmol/gDW (gray lines, Family B). (**b**) This manifold from subproblem 1 shows a dependent parameter as a function of the two free parameters. Black dots denote the sampled parameters that lead to responses in Family A, whereas gray dots denote those that belong to Family B. The locations of both the black and gray dots are evenly distributed across the entire manifold. (**c**) On this manifold from subproblem 6, the gray dots cluster in a region with low values of both $K_{E_2,\mathrm{FBP}}$ and $v_{E_2,f}$. (**d**) The system response is sensitively determined by the value of a single free parameter, $K_{E_2,\mathrm{FBP}}$. A randomization of all free parameters except $K_{E_2,\mathrm{FBP}} = 10$ only marginally affects the trajectories (1000 purple lines), whereas a stepwise reduction of only this parameter ($K_{E_2,\mathrm{FBP}} = 10; 4; 3; 2; 1.5; 1.4; 1.3; 1.2; 1.1; 1.0; 0.9; 0.8; 0.7; 0.5; 0.3; 0.1$) with all other free parameters constant sensitively shapes the response. (**e**) The value of the free parameter $K_{E_2,\mathrm{FBP}}$ remains the decisive factor in determining the shape of the trajectories even in the presence of 10% measurement noise ($K_{E_2,\mathrm{FBP}} = 10$; 3; 1; 0.1 for the purple, black, green and blue curve sets, respectively).

parameters $v_{E_2,f}$ and $v_{E_2,r}$ with it), i.e. by moving the parameter vector in the direction of the negative $K_{E_2,\mathrm{FBP}}$-axis of the solution manifold shown in Fig. 3c, we were able to move the trajectory across the entire range of the possible responses (Fig. 3d). Thus, in addition to having identified the two most important parameters, we also understand how their variation affects the system response.

Next, we assessed if these two important parameters retain their dominant role in determining the response in the presence of 10% measurement noise. We generated four sets of trajectories with $K_{E_2,\mathrm{FBP}}$ (and thereby $v_{E_2,f}$) at different levels and all other free parameters and the measurement data randomized. Figure 3e shows that despite of these sources of variation, $K_{E_2,\mathrm{FBP}}$ still largely determines the response: only for the green curve set with $K_{E_2,\mathrm{FBP}} = 1$, which is in the transition region between the response families A and B (Fig. 3d), do these sources of variation have a considerable impact on the trajectories. Therefore, the obtained understanding of how the system response is dominantly shaped by $K_{E_2,\mathrm{FBP}}$ (and $v_{E_2,f}$) is reasonably robust with respect to measurement noise. Note that instead of assuming a flat noise magnitude of e.g. 10%, more detailed information about the uncertainties of individual data points can be used, if available.

To conclude, by exploiting the solution manifolds derived with the divide-and-conquer approach, we were capable to obtain a profound system understanding even though the parameter values were not identifiable due to limited and noisy measurement data. In general, the discovery of the most important parameter values suggests targeted experiments to measure these values and may already provide a valuable insight by itself. Before drawing biological conclusions from this particular example system, however, its predictive power should be tested, or alternatively, it should be ensured that the observed effect extends to other model variants and is thus not specific to the chosen model structure, which is merely one among many possible mathematical representations of the available biochemical knowledge. During this process, which

is demonstrated e.g. by Kuepfer *et al*. (2007) and Kremling *et al*. (2008), the divide-and-conquer approach can be repeatedly applied.

5 DISCUSSION

In this article, we presented the divide-and-conquer approach for the analysis of underdetermined biochemical models. This approach exploits a 'trivial point' at which the complete solution space of the global parameter estimation problem can be derived analytically. Using an example system, we have demonstrated how the complete solution space can be derived and subsequently analyzed. This strategy resulted in a sound system understanding and the identification of targeted experiments.

The main difficulty in applying this approach is to move a real-world estimation problem onto that 'trivial point', i.e. to fulfill Conditions 1 and 2. This can be achieved by various means. First, an incomplete measurement dataset can be extended to a complete set of observables by incorporating additional biological knowledge. Second, the measurement noise can be initially neglected and the robustness of the derived insights with respect to measurement noise assessed at a later stage. Third, additional, possibly time-course measurement data that does not belong to a complete steady state dataset can also be included in the divide-and-conquer approach. This can be achieved by providing a global optimizer with the derived equality and inequality constraints on the parameters [Equations (S11–S17) of Supplementary Material]. Then, the optimizer can determine that parameter combination on the solution manifolds which best reproduces the additional measurements *in addition to* exactly reproducing the complete steady state datasets.

To enable analytical solution spaces of large-dimensional parameter estimation problems, the divide-and-conquer approach decomposes the whole solution space via Condition 3 into independent subspaces, for which analytical solutions are feasible. This decomposition occurs automatically when few parameters, whose number in our experience increases only slightly with model size, are fixed at literature values and thereby excluded from the anyway underdetermined parameter estimation problem. The proposed analytical approach is thus well scalable to the often large sizes of realistic models.

Although this approach is not confined to any specific type of model, it is best suited for application areas where models are typically underdetermined, yet omics datasets are available. Due to the many parameters of enzyme kinetics and the availability of metabolomics and fluxomics data, this approach is especially suited for models of metabolism.

The key advantage of the divide-and-conquer approach is that the global solution space of the parameters can be structured in *manageable subspaces*, which are known *completely* and *analytically*. This greatly facilitates the analysis of the possible system responses within the solution space. Of particular interest is the identification of those few (Gutenkunst *et al*., 2007) parameter combinations that most sensitively shape the system response—in fact, this task is one of the major problems raised in systems biology (Kitano, 2002b). Therefore, by focusing directly on the system responses and not on the parameter values, the divide-and-conquer approach is a practical strategy to extract valuable insights from underdetermined biochemical models.

ACKNOWLEDGEMENT

The authors would like to thank Stefan Jol and Jörg Stelling for helpful discussions.

Funding: YeastX project within the Swiss Initiative in Systems Biology (SystemsX.ch).

Conflict of Interest: none declared.

REFERENCES

Brown,K.S. *et al*. (2004) The statistical mechanics of complex signaling networks: Nerve growth factor signaling. *Phys. Biol.*, **1**, 184–195.

Gadkar,K.G. *et al*. (2005) Iterative approach to model identification of biological networks. *BMC Bioinform.*, **6**, 155.

Gutenkunst,R.N. *et al*. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comp. Biol.*, **3**, 1871–1878.

Hengl,S. *et al*. (2007) Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, **23**, 2612–2618.

Ingram,P.J. *et al*. (2006) Network motifs: structure does not determine function. *BMC Genomics.*, **7**, 108.

Ishii,N. *et al*. (2007) Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, **316**, 593–597.

Jaqaman,K. and Danuser,G. (2006) Linking data to models: data regression. *Nat. Rev. Mol. Cell Biol.*, **7**, 813–819.

Kitano,H. (2002a) Computational systems biology. *Nature*, **420**, 206–210.

Kitano,H. (2002b) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

Kremling,A. *et al*. (2008) A feed-forward loop guarantees robust behavior in *Escherichia coli* carbohydrate uptake. *Bioinformatics*, **24**, 704–710.

Kuepfer,L. *et al*. (2007) Ensemble modeling as a novel concept to analyze cell signaling dynamics. *Nat. Biotechnol.*, **25**, 1001–1006.

Kümmel,A. *et al*. (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.*, **2**, 2006.0034 (doi:10.1038/msb4100074).

Liebermeister,W. and Klipp,E. (2005) Biochemical networks with uncertain parameters. *IEE Proc. Syst. Biol.*, **152**, 97–106.

van Riel,N.A.W. (2007) Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief. Bioinform.*, **7**, 364–374.