A Hybrid Approach to Piecewise Modelling of Biochemical Systems

Zujian Wu, Shengxiang Yang, and David Gilbert

School of Information Systems, Computing and Mathematics
Brunel University, Uxbridge, Middlesex UB8 3PH, UK
{zujian.wu, shengxiang.yang, david.gilbert}@brunel.ac.uk

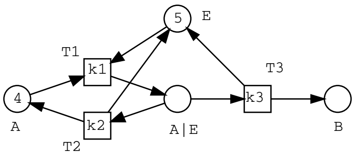
Abstract. Modelling biochemical systems has received considerable attention over the last decade from scientists and engineers across a number of fields, including biochemistry, computer science, and mathematics. Due to the complexity of biochemical systems, it is natural to construct models of the biochemical systems incrementally in a piecewise manner. This paper proposes a hybrid approach which applies an evolutionary algorithm to select and compose pre-defined building blocks from a library of atomic models, mutating their products, thus generating complex systems in terms of topology, and employs a global optimization algorithm to fit the kinetic rates. Experiments using two signalling pathways show that given target behaviours it is feasible to explore the model space by this hybrid approach, generating a set of synthetic models with alternative structures and similar behaviours to the desired ones.

1 Introduction

Models of biochemical systems can be used in systems biology to predict and explain behaviour, or as templates for designing novel biological systems in synthetic biology. It is still an open question regarding how to build and verify models of biochemical systems, involving intelligent methods and tractable computational tools. Traditionally the structures of models are inferred from various experimental observations, and the kinetic rates are estimated computationally by considering kinetic laws [3, 9].

Much previous research has focused on how to fit the kinetic rates of an existing biochemical model so that its behaviour coincides with the observations of a given physical system [7, 14, 13]. However, another research line is to identify alternative topologies and optimize the topologies [8, 20]. Moreover, a model of a biochemical system can be engineered by modifying and piecewise constructing its network topology, using biological building blocks. As the kinetic rates (parameters) associated with biochemical reactions (forming the structure) are crucial for biochemical systems exhibiting observed behaviours, it is necessary to model the systems in terms of both the topology and kinetic rates by a hybrid method. The challenging aim of our research is the development of a robust method for the automated construction of models from descriptions of the

Table 1: An enzymatic reaction and its components

Enzymatic Reaction	Petri net	Components
$A + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} A\|E \xrightarrow{k_3} B + E$ $[A] = 4$ $[E] = 5$ $[A\|E] = [B] = 0$	 5 E T1 k1 T3 4 k3 A k2 A|E B T2	$A + E \xrightarrow{k_1} A\|E$ $A\|E \xrightarrow{k_2} A + E$ $A\|E \xrightarrow{k_3} B + E$

observed or desired behaviours of the biochemical systems, by the manipulation of both the topology and kinetic rates.

Some recent research applying evolutionary methodologies to model biological systems can be found in [18, 19, 2]. Evolutionary computation and functional Petri nets have been applied to infer metabolic pathways by Kitagawa and Iba [10]; however their approach relies on starting with an existing network model which is then modified, whereas our approach is to incrementally piecewise construct a network from a single node. Previously [21] we have developed a method to piecewise construct the topology of networks using simulated annealing (SA); in the research reported here we use a hybrid approach which employs evolution strategy (ES) to derive the topology and SA for the kinetic rates.

2 Components and Composition Rules

2.1 Pre-defined Components

Components are defined according to the semantics and syntax of the biological building blocks in [21]. There are two patterns for generating the reusable components in a library: (1) binding pattern $P_1 + P_2 \xrightarrow{k_i} P_3$; (2) unbinding pattern $P_3 \xrightarrow{k_j} P_1 + P_2$. The parameters k_i and k_j are the kinetic rates of binding and unbinding reactions, and usually $k_i \gg k_j$. The two patterns illustrate how a complex can be synthesized from substrates or broken down into substrates. A basic enzymatic reaction can be represented by one instantiation of the binding pattern and two instantiations of the unbinding pattern, as shown in Table 1. The concentrations of substrates in the enzymatic reaction are indicated with labels within square brackets, such as '[A]' and '[A|E]', where the symbol '|' means that the biochemical complex 'A|E' is made from two substrates A and E.

2.2 Composition Rules

Given an existing biochemical network model *BioN* and a library of biological components *CompLib*, the operation of piecewise composition can be the addition of one component C_a from *CompLib* to *BioN*, or the subtraction of one component C_s from *BioN*. We have further developed the original composition rules proposed in [21] to permit component subtraction and greater flexibility in composition. The rules developed are performed by comparing and replacing parts of the labels of the added component. In this paper, L_i $(i = 1, 2, 3)$ is the label of places P_i from the added component C_a, and L_B is the label of a place P_B of a component C_B in *BioN*. The details of composition rules are as follows.

1. Given a binding component $P_1 + P_2 \xrightarrow{k1} P_3$ or an unbinding component $P_3 \xrightarrow{k2} P_1 + P_2$, where L_1, L_2 and L_3 are labels of P_1, P_2 and P_3, respectively.
   (a) If $L_B = L_1$ or $L_B = L_2$ or $L_B = L_1|L_2$, the component C_a is added to the existing network *BioN* by adding its reaction equations directly;
   (b) If $L_B \neq L_1$ or $L_B \neq L_2$, all L_1 (L_2) in C_a are replaced by L_B in C_a and the modified reaction equations are added to *BioN*;
   (c) If $L_B \neq L_3$ and P_B is a complex, L_3 in C_a is replaced by L_B, L_1 is replaced by L_{B1}, and L_2 is replaced by L_{B2} where $\{L_{B1}, L_{B2} | L_{B1} \cap L_{B2} = 0$ and $L_{B1} \cup L_{B2} = L_B\}$. The corresponding modified reaction equations of C_a are added to *BioN*;
   (d) If $L_B \neq L_3$ and P_B is not a complex, the reaction equations of C_a are added into *BioN*, and a new component C'_a is created by binding P_3 with P_B to produce $P_B|P_3$ $(P_3 + P_B \xrightarrow{k1} P_B|P_3)$, and reaction equations of C'_a are added to *BioN*.
2. A component C_s is selected randomly from *BioN* for subtraction.
   (a) If C_s is the only component in *BioN*, no subtraction is applied to *BioN*;
   (b) If C_s is not the only component in *BioN*, the transition and its incident arcs in C_s are removed directly. The *BioN* is checked for connectivity. Non-connected parts of *BioN* are linked by creating a binding component with species selected randomly from the non-connected parts.

3 Hybrid Piecewise Modelling

Current research has focused on generating topologies [16] and fitting kinetic rates [17], or both [5]. In this paper, we aim to solve a topology optimization problem by iteratively piecewise assembling components represented by quantitative Petri nets from a user pre-defined library, combined with optimizing the kinetic rates.

A hybrid evolutionary and heuristic approach has been developed using a two layer framework: firstly, this hybrid approach evolves the topology of the model representing the target system by performing ES at the outer layer, and then SA is applied at the inner layer to optimize the kinetic rates of the evolved model. The piecewise modelling stops after a pre-defined number of generations and returns a set of the best synthetic models, offering alternative topologies with similar behaviours to the target system. The pseudo-code of the hybrid framework between ES and SA is shown in Algorithm 1. The details of evolving the topology by the ES layer and optimizing kinetic rates by the SA layer are described in Section 3.1 and Section 3.2, respectively.

Algorithm 1 A hybrid piecewise modelling framework

```
Require: CompLib, Composition Rules
Ensure: $BioN_{best}$
Initiate the population;
while Not reached maximum generation (ES layer) do
    for Each individual in the population do
        Mutate the topology of individual by Addition or Subtraction;
        Check the mutated topology of the individual;
        Evaluate the mutated individual;
        if The kinetic rates are required to be optimized then
            while Not reach minimum temperature (SA layer) do
                Optimize the kinetic rates of individual by Gaussian distribution;
                Evaluate the mutated kinetic rates;
            end while
        end if
    end for
    Crossover the individuals;
    Select offspring for next generation;
end while
Return $BioN_{best}$
```

3.1 Evolution Strategy Based Topology Optimization

The (μ+λ)-ES is utilized to iteratively piecewise assemble the components for the model construction, where μ and λ are the number of parents and children respectively. The (μ+λ)-ES starts from an initial population of individuals and each individual is a single component selected randomly from the library. The individuals are mutated by genetic operators adapted from evolutionary algorithms: *Addition* ($\oplus$), *Subtraction* ($\ominus$) and *Crossover* ($\otimes$). The individuals with the best fitness are selected to generate offspring for the next generation.

The three genetic operators are concepts taken from genetic algorithms, and the implementation of these operators in this paper is inspired by nature. The addition operator is used to integrate a component to an existing topology of model. The subtraction operator is used to remove the transition with incident arcs in a component selected randomly from the model for a removal. The crossover operator is used to apply a ‘cut and splice’ method to reproduce offspring from two models under construction. The set of composition rules has been introduced in Section 2.2 for the components composition carried out by the three genetic operators.

3.2 Simulated Annealing Based Kinetic Rates Fitting

SA is a heuristic optimization algorithm for searching a global optimum solution in a very large solutions space, avoiding local optimum solutions. In our previous work [21] we have applied the SA to piecewise construct and explore the topologies of the biological systems. In this paper, the SA layer is integrated within the ES layer to estimate the kinetic rates of the synthetic models. The topologies of these models are fixed while in the SA layer, having been passed down from the ES based outer layer after mutating their structures.

The rates of reactions in a given model are coded as follows: a vector $K(M) = (k_1^t, k_2^t, ..., k_l^t)$ is used to record the rate values in a model, where l is the number of reactions, t is the current cooling temperature, and k_i^t is a constant rate of the ith chemical reaction r_i $(i = 1, 2, ..., l)$. The vector $K(M)$ is mutated by the *Gaussian* distribution $N(\mu, \sigma)$ by N iteration times at each cooling temperature. The mutated $K(M)$ of the model is evaluated after each iteration, by comparing the behaviour of the synthetic model and the target system.

Due to the probabilistic behaviour of the random procedure of SA [1], a mutated vector $K(M)$ could be generated which causes a bad estimated fitness of the model. This is because there is a chance that the model with a fixed topology and optimized kinetic rates returned from the SA layer to the ES layer could be worse than the one passed into the SA layer.

3.3 Model Evaluation

A synthetic model is evaluated by comparing its behaviours with target behaviours of a biochemical system. The behaviours are represented by time series data of the concentrations of species, e.g. enzymes, other proteins, and complexes. The behaviours of the species in the target system can be obtained from a reference model or by observations of a biochemical system from the wet-lab.

Given a set of reference data for the target system M_T, there are N generated time series $X_T = (X_1, X_2, ..., X_N)$ which represent the behaviours of N species, $N \geq 1$. There are P data points in each time series $X_i = (x_i^1, x_i^2, ..., x_i^P)^T$, $i = 1, ..., N$. There are M time series $X_G = (\hat{X}_1, \hat{X}_2, ..., \hat{X}_M)$ describing the behaviours of M species in a constructed model M_G, with P data points for each time series $\hat{X}_j = (\hat{x}_j^1, \hat{x}_j^2, ..., \hat{x}_j^P)^T$, $j = 1, ..., M$. The intersection between M_T and M_G of species is defined by $X_C = X_T \cap X_G = (X_1, X_2, ..., X_n)$, $1 \leq n \leq N$. The difference between the behaviours of M_T and M_G is calculated by averaging the difference of behaviours of each species in X_C by a paired comparison of the P data points. As shown in Eq. (1), the difference of behaviours for one species $X_k \in X_C$ is measured by the Euclidean distance function, where η is the total number of compared substrates in X_C:

$$d_{M_T, M_G}(X_k) = \frac{1}{\eta} \sum_{k=1}^{\eta} \sqrt{\sum_{t=1}^{P} (x_k^t - \hat{x}_k^t)^2}. \qquad (1)$$

While evaluating the generated model, the species for behaviour comparison can be specified by the user and are stored in X'_C $(|X'_C| = n')$. In this scenario, there could be some synthetic substrates in M_G which do not exist in M_T. Therefore, if a substrate is specified for comparison in M_G but does not exist in M_T, then M_G should be punished. If a species for comparison exists both in M_T and M_G, a reward can be given to M_G. A *Reward and Penalty* function $\Phi(X_k)$

is used to improve the objective function as a complement of the Euclidean distance function: $\Phi(X_k) = -\varepsilon_1$ if $X_k \in X_G \wedge X_k \notin X_T$, where ε_1 is a non-negative real value for the reward; $\Phi(X_k) = \varepsilon_2$ if $X_k \in X_G \wedge X_k \in X_T$, where ε_2 is a non-negative real value for the punishment. The reward and penalty can be defined by the user at the initial stage. The return result of $\Phi(X_k)$ will partly contribute to the fitness evaluation of a generated model M_G by an objective function $f(M_G)$ in Eq. (2):

$$f(M_G) = d_{M_T,M_G}(X_k) + \frac{1}{\eta}\sum_{k=1}^{\eta}\Phi(X_k) \tag{2}$$

where $\eta = n$ if the compared substrates are from the intersection X_C, and $\eta = n'$ if the compared substrates are from the specific X'_C. In this paper, modelling is a minimization problem, therefore the smaller the fitness value, the better the generated model.

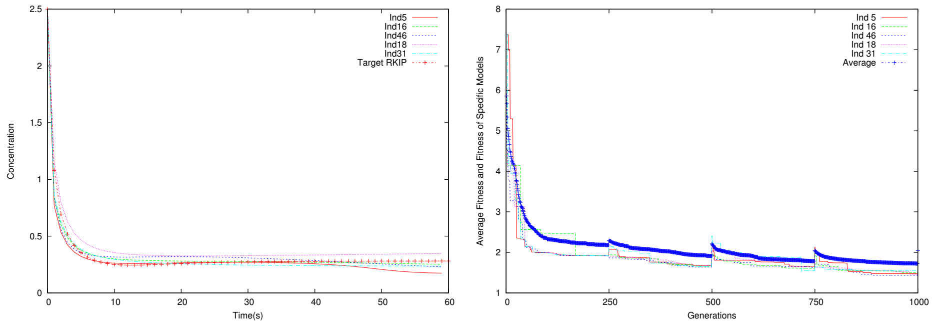
4 Experimental Study

In this section, we present simulation results for the implementation of the hybrid modelling approach on two signalling pathways: (1) the RKIP pathway, which is a mathematical model taken from Cho et al [6] for representing the fragment of the mitogen-activated protein kinase (MAPK) signal transduction pathway concerned with the inhibition of the extracellular signal regulated kinase (ERK) by the Raf1 kinase inhibitor protein (RKIP); (2) the Levchenko pathway [11] for quantitatively analyzing the signal propagation regulated by the formation of scaffold kinase complexes in the core MAPK cascade. ERK is one of the MAP Kinases (mitogen activated protein kinases), and can also be referred to as 'MAPK'; it is a player in both the Cho et al model of the RKIP fragment of the MAPK cascade as well as in the Levchenko model of the MAPK cascade.

4.1 Generation of Similar Behaviours

The main aim of our approach is to construct models with similar behaviours to the target biochemical systems. The RKIP pathway transfers the mitogenic signals from the cell membrane to the nucleus. The hypothesis is that RKIP can inhibit activation of Raf1 by binding to it, disrupting the interaction between Raf1 and MEK, thus playing a part in regulating the activity of the ERK.

Figure 1a shows that the behaviours of substrates in the generated models are similar to the target behaviour in terms of Euclidean distance. Because the behaviours of RKIP in the 50 synthetic models are similar both to each other and also to the target behaviour, we only illustrate the behaviours of RKIP from the five best generated models (obtained in a single run). The construction of the models can be driven to approach to that of the target pathway by increasing the fitness in terms of reducing the Euclidean distance between behaviours employed to evaluate the models. As shown in Fig. 1b, the fitness of each model converges



(a) Behaviours of RKIP

(b) Fitness of synthetic models

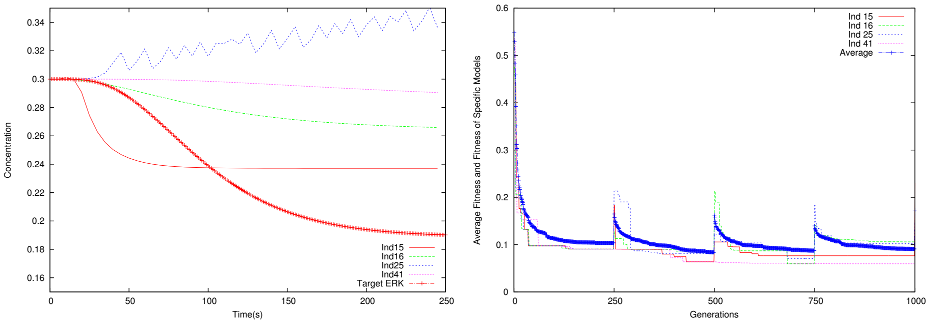
Fig. 1: (a) Behaviours of the RKIP from five best synthetic models and target RKIP pathway; (b) Average fitness of all 50 synthetic models of RKIP pathway, and fitness of the five best synthetic models.

to a minimum value with the increased number of generations in the simulation. In our current implementation, the hybrid modelling process is set to call the SA layer to optimize the kinetic rates of each model at every 250 generations; different settings are under study in ongoing research. Due to the probabilistic mechanism of accepting a worse solution by the SA, there is a jump of fitness convergence for most models. These fitness values converge again after move back to the ES layer, following a traditional evolutionary process, see Fig. 1b.

Our results for the Levchenko pathway given in Figure 2a show that models can be generated with unexpected behaviours regarding ERK which are similar in terms of Euclidean distance to the target in the MAPK cascade [11]. Although the target system does not exhibit oscillations, there is an oscillating substrate behaviour from one of the synthetic models, as supported by Kholodenko's model [12]. This suggests that feedbacks could exist in solution space, and are indeed incorporated in many MAPK models, e.g. [4], although missing in our target Levchenko model. Again, the fitness of constructed models converges with the increased number of generations in the simulation as shown in Fig. 2b.

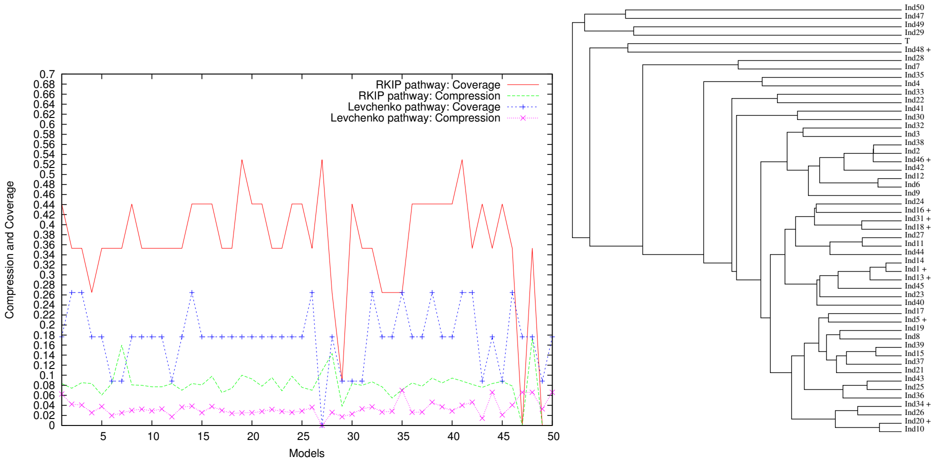
4.2 Exploration of Alternative Topologies

One of main aims of modelling biochemical systems is to explore alternative topologies, for understanding the relationships among the compounds in wet-lab. Our approach can search the model space and suggest a set of alternative topologies with similar behaviours to the target. The results can be analysed in terms of the structural difference between models, using *Compression* and *Coverage* measures. *Compression* (adapted from [21]) is a metric which computes the distance between two networks in terms of the proportion of matched (common) arcs (between the generated and target model) with respect to the maximum number of arcs in the generated or target model. *Coverage* computes inclusion in



(a) Behaviours of ERK (b) Fitness of synthetic models

Fig. 2: (a) Behaviours of ERK from four best and interesting synthetic models and target Levchenko pathway; (b) Average fitness of all 50 synthetic models of Levchenko pathway, and fitness of four best and interesting synthetic models in terms of ERK behaviours.



(a) RKIP and Levchenko pathways (b) Clusters

Fig. 3: (a) Compression and coverage of RKIP and Levchenko pathways; (b) A clustering of 50 synthetic models and target RKIP pathway (T).

terms of the ratio of arcs in the target model which are matched in the generated model. Both measures vary from 0 (worst) to 1 (best). If either compression or coverage is low for a particular model, then its topology is very different to the target, even if their behaviours are similar.

Figure 3a illustrates the compression and coverage of two signalling pathways. Most coverage of synthetic models of RKIP and Levchenko pathways is in the ranges of [0, 0.53] and [0, 0.27], respectively. Compression for both the RKIP and the Levchenko generated models is very poor, ranging over [0, 0.18], indicating that the generated models are very different to the target ones in terms of topologies. Figure 3b is a dendrogram of hierarchical pairwise clustering based on similarity and complete linkage over compression among 50 generated models and the target RKIP pathway, and illustrates the generation of a wide range of alternative topologies by our hybrid approach; the closest 10 models in terms of fitness are shown with a '+'. Although none of the generated topologies are close to the target one, the nearest being individual 48 which is 10th closest regarding fitness, there are 9 other models which are closer in terms of behaviour despite being poorly related to the target structurally, and are also fairly widely scattered over structural space. Thus our approach is able to search model space for networks which have similar behaviours to the target, even though they may differ quite significantly in terms of structure.

5 Conclusions and Future Work

Our study addresses the evolution of quantitative Petri nets and could thus be applied to stochastic and hybrid Petri nets as well as the continuous Petri nets, which can benefit mathematical modelling. We have applied the proposed approach to two signalling pathways. The experiments show that it is feasible to iteratively piecewise model biochemical systems using our hybrid approach and explore the solution space of alternative models with different topologies but similar behaviours to the target ones.

One important issue to be investigated in our future research is to study the switching policy between ES and SA layers, in order to obtain models with good quality in terms of both topology and kinetic rates. Furthermore, implementation of the genetic operators can result in different model sizes, and thus one of our future aims is to exploit the potential tradeoff of the combinatorial application of the genetic operators. More biological constraints will be considered for defining the components and the composition rules, thus improving the biological relevance of the synthetic models. Finally, the generated models can be used as design templates to guide the construction of synthetic biological systems which may have quite different topologies from existing natural systems.

Acknowledgements ZW is partially supported by a grant from the China Scholarship Council and a grant from Brunel University. We would like to thank Crina Grosan for her helpful comments on the text.

References

1. Anily, S., Federgruen, A.: Simulated annealing methods with general acceptance probabilities. J. Appl. Prob., 24(3), 657–667 (1987)
2. Balsa-Canto, E., Banga, J.R., Egea, J.A., Fernandez-Villaverde, A., de Hijas-Liste, G.M.: Global optimization in systems biology: stochastic methods and their applications. In Goryanin, I.I., Goryachev, A.B. (eds.) Advances in Systems Biology, Adv. Exp. Med. Biol. 736, 409-424 (2012)

3. Breitling, R., Gilbert, D., Heiner, M., Orton, R.: A structured approach for the engineering of biochemical network models, illustrated for signalling pathways. Brief Bioinform., 9(5), 404–422 (2008)
4. Brightman, F.A., Fell, D.A.: Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells. FEBS letters, 482(3), 169–174. Elsevier (2000)
5. Cao, H., Romero-Campero, F., Heeb, S., Camara, M., Krasnogor, N.: Evolving cell models for systems and synthetic biology. Syst. Synth. Biol., 4(1), 55–84 (2010)
6. Cho, K.H., Shin, S.Y., Kim, H.W., Wolkenhauer, O., Mcferran, B., Kolch, W.: Mathematical modeling of the influence of RKIP on the ERK signaling pathway. In: Priami, C. (eds.) CSMB 2003, LNCS 2602, 127–141 (2003)
7. Feng, X.J., Hooshangi, S., Chen, D., Li, G., Weiss, R., Rabitz, H.: Optimizing genetic circuits by global sensitivity analysis. Biophys., 87(4), 2195–2202 (2004)
8. Francois, P., Hakim, V.: Design of genetic networks with specified functions by evolution in silico. PNAS, 101(2), 580–585 (2004)
9. Gilbert, D., Breitling, R., Heiner, M., Donaldson, R.: An introduction to BioModel Engineering, illustrated for signal transduction pathways. In: WMC 2008, LNCS 5391, 13–28 (2009)
10. Kitagawa, J., Iba, H.: Identifying metabolic pathways and gene regulation networks with evolutionary algorithms. In Fogel, G.B., Corne, D.W. (eds.) Evolutionary Computation in Bioinformatics, 255–278 (2003)
11. Levchenko, A., Bruck, J., Sternberg, P.W.: Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. In: Proc. of the National Academy of Sciences of the United States of America, 97(11), 5818–5823 (2000)
12. Kholodenko, B.N.: Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. Eur. J. Biochem, 267: 1583–1588 (2000)
13. Manca, V., Marchetti, L.: Log-Gain stoichiometric stepwise regression for MP systems. J. Found. Comput. Sci., 22(1), 97–106 (2011)
14. Maria, G.: A review of algorithms and trends in kinetic model identification for chemical and biochemical systems. Chem. Biochem. Eng. Q., 18(3), 195–222 (2004)
15. Murata, T.: Petri Nets: properties, analysis and applications. In: Proc. of the IEEE, 77(4), 541–580 (1989)
16. Rodrigo, G., Carrera, J., Jaramillo, A.: Genetdes: automatic design of transcriptional networks. Bioinformatics, 23(14), 1857–1858 (2007)
17. Schulz, M., Bakker, B.M., Klipp, E.: TIde: a software for the systematic scanning of drug targets in kinetic network models. BMC Bioinformatics, 10(1), 344–353 (2009)
18. Sendin, J.O.H., Exler, O., Banga, J.R.: Multi-objective mixed integer strategy for the optimisation of biological networks. Systems Biology, IET, 4(3), 236-248 (2010)
19. Sun, J., Garibaldi, J.M., Hodgman, C.: Parameter estimation using metaheuristics in systems biology: a comprehensive review. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 9(1), 185–202 (2012)
20. Vyshemirsky, V., Girolami, M.: Bayesian ranking of biochemical system models. BMC Bioinformatics, 24(6), 833–839 (2008)
21. Wu, Z., Gao, Q., Gilbert, D.: Target driven biochemical network reconstruction based on petri nets and simulated annealing. In: Quaglia, P. (eds.) CMSB 2010, 33–42. ACM (2010)