Empirical study of computational intelligence strategies for biochemical systems modelling

Zujian Wu[1], Crina Grosan[2,3] and David Gilbert[3]

[1]School of Natural and Computing Sciences
University of Aberdeen, UK
[2] Department of Computer Science
Babes-Bolyai University Cluj-Napoca, Romania
[3]Department of Information Systems and Computing
Brunel University London, UK
zujian.wu@abdn.ac.uk, crina.grosan@brunel.ac.uk, david.gilbert@brunel.ac.uk

Abstract. Modelling biochemical networks can be achieved by iteratively analyzing parts of the systems via top-down or bottom-up approaches. It is feasible to piece-wise model the biochemical networks from scratch by employing strategies able to assemble reusable components. In this paper, we investigate a set of strategies that can be employed in a bottom-up piece-wise modelling framework, to obtain synthetic models with similar behaviour to the target systems. A combination of evolution strategies and simulated annealing is employed to optimize the structure of the system and its kinetic rates. Simulation results of different variants of those computational methods on a standard signaling pathway show that it is feasible to obtain a tradeoff between the generation of desired behaviour and similar and alternative topologies.

1 Introduction

In theoretical chemistry and systems and synthetic biology, time-dependent chemical concentration data for large networks of biochemical reactions are important. These data are collected with the purpose to identifying the exact structure of a network of chemical reactions and their corresponding kinetic rates for which the identity of the chemical species present in the network is known but no information is available on the species interactions.

General methods for engineering biochemical networks can be divided into two main approaches: top-down or bottom-up, which allow the modelling of biochemical systems by manipulating parts of the systems. In the top-down (analytical) approach, a whole complex biochemical system is segregated into subunits that can be easily dealt with for further investigation, such as dissecting apoptotic signals [9] and tuning complex signal cascades [14]. In the bottom-up (constructionist) approach, a complex biochemical system is composed from building blocks where the relationships of involved compounds are investigated, such as building synthetic oscillators [15] and transplanting synthetic genomes [2]. The

modelling of biochemical networks involves the optimisation of two main attributes: network topology and kinetic rates.

There exist several approaches dealing with inferring biochemical systems, some of them with limitations and drawbacks [6][12][13]. They mainly include evolutionary algorithms and genetic programming (from the class of evolutionary computation models). Previous research applies a hybrid combining Evolutionary Strategies (ES) and Simulated Annealing (SA) to the optimisation of topology and the kinetic rates of a biochemical system [20]. In this paper, we investigate variants of the ES-SA heuristics for bottom-up systems modelling. Due to the flexibility of these strategies, various combinations of the evolutionary operators, evaluation criteria and design principles can be considered. These variants are presented in detail in Section 4.

2 Biochemical systems

The modelling of biochemical systems has been investigated widely in computational biology, especially in systems biology. In biochemistry, a chemical reaction is a process of converting molecules of reactants into products within a specific time period. The reactants are usually known as substrates. Biochemical systems are composed of interacting molecules (or molecular species), whose dynamic evolution is determined by the occurrence of chemical reactions. A biochemical model is fully characterized by the initial concentration of each molecular species and the description of the reactions with their kinetic rate laws. The representation of the dynamics is given by an ordinary differential equation (ODE) as follows:

$$\frac{dX_i}{dt} = \sum_j \mu_{ij} \cdot \gamma_j \prod_k X_k^{f_{jk}} \tag{1}$$

where X_i represents one species of the model, for instance metabolite concentrations, protein concentrations or levels of gene expression; j represents the biochemical reaction affecting the dynamics of the species; μ_{ij} indicates the stoichiometric coefficient; γ_j indicates rate constants; f_{jk} stands for kinetic orders; and k denotes the number of species.

Mass action kinetics are used in chemistry and chemical engineering to describe the dynamics of chemical reactions. The mass action given in equation MA is used in this work; note that A is the substrate, B the product, E the enzyme and $A|E$ is the intermediary substrate-enzyme complex.

$$A + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} A|E \xrightarrow{k_3} B + E \ . \tag{MA}$$

There are different methodologies employed to describe biochemical systems in computational biology. Petri nets are one of the existing mathematical modelling structures used for the description of biochemical systems as a reaction-system behaviour descriptor, and comprise two types of nodes – places and

transitions – connected via edges. The usage of Petri nets in biological systems comes as a natural solution as biochemical reactions are inherently bipartite, comprising reactions between biochemical entities [11], which are mapped onto transitions and places respectively. A continuous Petri net can be represented by a system of ODEs [8]. We focus on the automatic identification of both network structure and its corresponding kinetic rates from observed time-domain concentrations alone without assuming a given basic structure or any given reaction kinetics.

3 ES-SA metaheuristic for biochemical systems

ES (as well as any of the evolutionary computation methods) are good candidates for evolving biochemical systems. A solution of the ES encodes a Petri net which is a representation of a biochemical system. SA is a powerful optimization method and it is used for optimizing the kinetic rates. The hybrid method ES-SA applied to biochemical systems is described in detail in [19][20]. We reproduce here the main characteristics. In order to understand the main constituents of an ES solution, elements such as *pattern, component, model* and *rules* are required. Any complex biochemical reactions can be described by employing instantiations from the binary patterns. The two general patterns we use describe how two species form a new species, or how one species is decomposed into two species:

- *binding pattern*: two reactants are merged into a complex with a specific kinetic rate
- *unbinding pattern*: a complex is disassociated back to reactants, or converted to a product and an enzyme with a specific kinetic rate.

A *component* for constructing biochemical models is given by $C = \langle P, T, f, v, m_0 \rangle$, which is based on the structure of Petri nets, where:

- P is a disjoint set of three continuous Places
- T is a singleton set containing one continuous Transition
- $f : ((P \times T) \cup (T \times P)) \rightarrow R_0^+$ defines a set of three directed arcs, weighted by non-negative real numbers, such that there is at least one arc of the form $p \rightarrow t$ and at least one of the form $t \rightarrow p$
- $v : T \rightarrow H$ assigns a firing rate function to the transition, whereby the set of all firing rate functions is $H := \bigcup_{t \in T} \left\{ h_t | h_t : R^{|\bullet t|} \rightarrow R \right\}$, and $v(t) = h_t$ is for the transition $t \in T$
- $m_0 : P \rightarrow R_0^+$ gives the initial marking.

A *model* of a biochemical system is a generalized form of a component but with no restrictions on the number of places and transitions. The mathematical interpretation of both component and model is a system of ODEs, illustrating the nonlinear relationship among at least three involved biochemical elements.

The ES part of the ES-SA metaheuristic builds models from single components by using evolutionary mechanisms for composition operators and rules.

A database has been designed and two libraries developed to store the components and models. Components are created at initial stage, according to the predefined patterns. A components library is developed as a table in the database, to preserve the generated components as atomic building blocks. The library maintains detailed information of these atomic components, such as labels of involved species, constants of associated kinetic rates and structures of created components.

The fitness function for a generated model M_G is given by:

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

where

$$d_{M_T,M_G}(X_k) = \frac{1}{\eta}\sum_{k=1}^{\eta}\sqrt{\sum_{t=1}^{P}(x_k^t - \widehat{x}_k^t)} \tag{3}$$

$X_T = (X_1, X_2, ..., X_N)$ represent the behaviour of the N species, P denotes data points in each time series $X_i = (x_{1i}, x_{2i}, ..., x_{Pi})$, $i = 1, ..., N$. There are M time series $X_G = (\widehat{X}_1, \widehat{X}_2, ..., \widehat{X}_M)$ describing the behaviour of M species in a constructed model M_G, and there are P data points for each time series $\widehat{X}_j = (\widehat{x}_{1j}, \widehat{x}_{2j}, ..., \widehat{x}_{Pj}), j = 1, ..., M$. The intersection between M_T and M_G of species is defined by $X_C = X_T \cap X_G = (X_1, X_2, ..., X_n), 1 \leq n \leq N$. $\eta = n$ if the compared species are from the intersection X_C and $\eta = n'$ if the compared substrates are from X'_C, the set which contains the species for behaviour comparison specified by the user. The fitness function has to be minimized, therefore the smaller the evaluated fitness value, the better the generated model.

A set of composition operators are adapted from the evolutionary optimization to fine tune the structures of the models:

- *addition*, represented by $\oplus$: addition rules add a component to a model
- *subtraction*, represented by $\ominus$: subtraction rules remove a component from a model
- *crossover*, represented by $\otimes$: crossover rules combine two models. The crossover rules allow two models be cut and spliced by swapping parts of the models via a "cut and splice" approach.

ES builds solutions, i.e. biochemical systems represented by Petri nets, in a piece-wise manner by applying the operators above to the components library. In this way, ES optimises the topology of the biochemical system. The kinetic rates of the reactions encoded in the Petri net are optimized using simulated annealing. In order to evaluate an ES-SA solution, the fitness function includes both the topology and the kinetic rates. The topology part of the fitness function gives the number of common species and their interactions in the evolved model compared to the target one. Some of the target model species and interactions may be missing from the generated model, as well as extra species not in the

target topology could be generated. For the optimization of the kinetic rates, we employed the BioNessie [10] platform to simulate the model and generate time course data as a set of target behaviour of species in the model. The measurement of behavioural distance is obtained by employing the Euclidean distance function. This part of the fitness involves solving the system of ODEs associated with the reactions. More details on the implementation of the two methods and all the parameters involved can be found in [19][20].

4 ES-SA variants for biochemical systems modelling

Due to a large variety of ways in which evolutionary methods can be designed in terms of performing genetic operators, comparing species behaviour and evaluating generated models during the construction process, we have carried out an empirical investigation of the advantages and disadvantages of some variants for the piecewise modelling, with an emphasis on the effect of genetic operators and evaluation criteria. Five sets of specific modelling variants are compared and general descriptions of these variants are given in what follows.

1. Methods of driving model composition:
   - Fixed: behaviour of a fixed set of species to be compared
   - Dynamic: behaviour of a dynamic set of species to be compared

   Time series data presenting behaviour of species in a target biochemical system is used to drive the modelling process via reducing the behaviour distance between generated and target model. Given a target biochemical system and a generated model which consist of N and M species respectively, there are two sets of time series data describing species behaviour in the target and generated model. It is easy to deduce that species to be compared can be selected via a fixed or a dynamic method.

   In the *fixed method*, the species in a fixed set are specified by users at the initial stage. They are referred to the target biochemical system. Therefore, all the information (names, concentrations and behaviour in time series data format) of these species is provided without uncertainty. Regarding the process of piecewise modelling, a model which is constructed at initial stages or evolved by mutation after many generations could only consist of less species than the target model. Thus some of the species could be absent.

   In the *dynamic method*, the species for comparison are generated and preserved in a dynamic set according to the existence of species in both generated and target models. The number of species is a dynamic variable in a range of $[0, N]$, N denoting the number of species in the target model.
2. Methods of survival selection:
   - SES: standard (1+1)-evolution strategy
   - PES: probabilistic (1+1)-evolution strategy, probabilistically accept a worse model

   A probabilistic evolution strategy (PES) is proposed, which differs from the standard evolution strategy (SES) in the sense that it can accept worse

models by a probability while searching the solution space. This may be helpful in avoiding local optima.
The *SES method* is the standard evolutionary process, selecting model candidates as offsprings for further evolution in following generations. The criteria for survival models is based on fitness value. The *PES method* introduces an acceptance probability into the stages of choosing survival models, which is integrated within the normal model selection stages of SES.

3. Methods of implementing the mutation operator (mutation consists in adding and/or subtracting a component to/from the topology):
   - Fixed: a fixed frequency of switching the addition/removal of a component to/from the model
   - Random: a random way of switching the addition/removal of a component to/from the model

   In the *fixed method*, the two mutation operators can be performed alternatively.
   In the *random method*, addition and subtraction are applied to models at every generation in a random manner.
4. Methods of performing crossover operator:
   - Best: each individual mates with the best individual in the population
   - Random: each individual mates with a randomly selected individual from the population

   The crossover operator mates two individual models under construction by a cut and splice method. New offspring are generated from the combination of parental models in terms of components (reactions and species). Parents and offspring compete and only one of them can be preserved as a model candidate in the population of the next generation. We consider two ways to performing the crossover operator: best and random methods.
   In the *best method*, each model under construction from the population is recombined with the model with best fitness. It is inspired by the elitism based individual selection in genetic algorithm.
   In the *random method*, each model in the population will be crossed over with another model chosen randomly.
5. Methods of evaluating solutions (models):
   - ED: the objective function represents the Euclidean distance function
   - ED+RP: the objective function is a combination of a reward and penalty mechanism and the Euclidean distance function

   The difference between generated and target model is calculated by employing an objective function. In the objective function, there are two methods of evaluating the composed models: Euclidean distance (ED) based method, and Euclidean distance with a reward and penalty mechanism (ED+RP) based method. ED is an ordinary distance between two points on the time series data representing the species behaviour from generated and target model. The inclusion of the reward and penalty in an objective function is intended to prioritize individuals whose components are among the ones existing in the target model. For instance, if a species is generated in a synthetic model and the species is also among the ones existing in the target model, fitness will be improved by giving a reward value.

5 Evaluation metrics

In order to evaluate the synthetic model structures quantitatively, two measures are employed: Compression and Coverage. Both measures vary from 0 (worst) to 1 (best). If either compression or coverage is low for a particular model, it indicates the topology of the generated model is very different from the target biochemical system, even if their behaviours are similar.

Compression (adapted from [3] and [7]) measures the percentage of matched common arcs between synthetic and target model and it is given by:

$$Compression = \frac{|Intersection|}{Max(|Target|, |Generated|)}$$

where $|Intersection|$ represents the number of matched arcs between target and generated topology, $|Target|$ is the number of arcs in the target topology, $|Generated|$ denotes the number of arcs in the generated topology, and $Max(|Target|, |Generated|)$ is the maximum number of arcs in either of the target and generated model.

Coverage calculates the ratio of matched arcs in the target model and it is given by:

$$Coverage = \frac{|Intersection|}{|Target|}$$

where $|Intersection|$ represents the number of matched arcs between target and generated topology, and $|Target|$ is the number of arcs in the target topology.

6 Experiments and comparisons

In order to quantitatively study the modelling variants, we performed statistical analysis of the performance by comparing fitness values, compression and coverage scores. One of the most important and intensively studied signaling pathways is ERK pathway (the Ras/Raf-1/MEK/ERK signaling pathway) which transfers the mitogenic signals from the cell membrane to the nucleus [17]. The ERK pathway is de-regulated in various diseases, ranging from cancer to immunological, inflammatory and degenerative syndromes and thus represents an important drug target. A brief illustration of regulations among proteins and complex based on signaling transduction in the ERK pathway is given as follows. Ras is activated by an external stimulus, via one of many growth factor receptors; it then binds to and activates Raf-1 to become Raf-1*, or activated Raf, which in turn activates MAPK/ERK Kinase (MEK) which in turn activates Extracellular signal Regulated Kinase (ERK). Cell differentiation is controlled by following cascade of protein interactions: Raf-1→ Raf-1* → MEK → ERK. The effect of regulation is dependent upon the activity of ERK. The Raf-1 kinase inhibitor protein (RKIP) inhibits the activation of Raf-1 by binding to it, disrupting the interaction between Raf-1 and MEK, thus playing a part in regulating the activity of the ERK pathway [18]. A number of computational models

have been developed in order to understand the role of RKIP in the pathway and ultimately to develop new therapies [4][5].

Due to the space limitation we present the analysis of a single signaling pathway but other examples could be found in [19].

Figure 1 shows a Petri net of the *RKIP* signaling pathway. Figure 2 displays the behaviour of all the species in the model of *ERK* signaling pathway regulated by *RKIP*, which is generated by simulation on a set of given ODEs and a group of original kinetic rates.

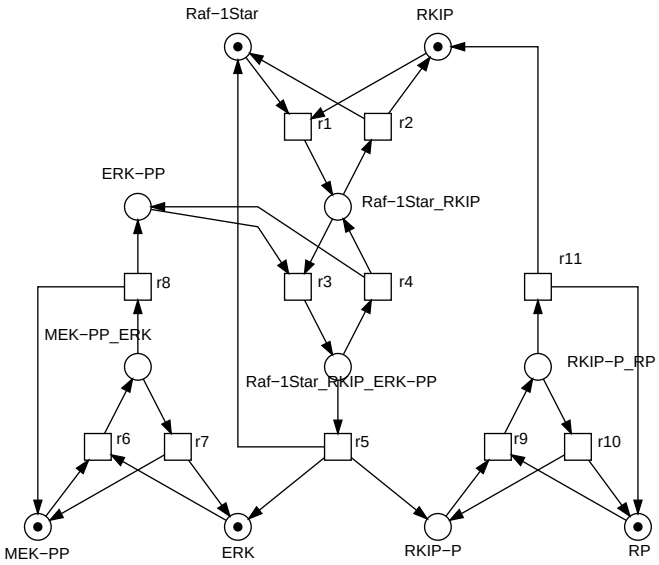


Fig. 1: A Petri net of the *RKIP* signaling pathway. Initial markings are taken from [21]

6.1 Simulation settings

There are five pairs of ES-SA variants compared and investigated. Details of simulation settings are given in Table 1.

The hybrid ES-SA platform calls the subtraction operator at every two generations, Sub@Ge=2; SA is called to optimize kinetic rates at every 25 generations, OptRate@Ge=25; reward ε_1 and penalty ε_2 values are 0.01 and 1000 respectively.The number of generations in one run of ES is 100, GeSi=100; the number of individuals is 50, PopSi=50. Initial SA system temperature is 10, T_{ini}=10; cooling rate of SA system is 0.8, CoRate=0.8; minimum temperature for stopping simulation is 1, T_{min}=1; number of iterations at each temperature is 10, Iter=10. The mean μ and standard deviation σ of Gaussian distribution $N(\mu, \sigma)$ are 0 and 0.00001, μ=0 and σ=0.00001. Other properties of the simulation setting during the modelling process are fixed without modification except

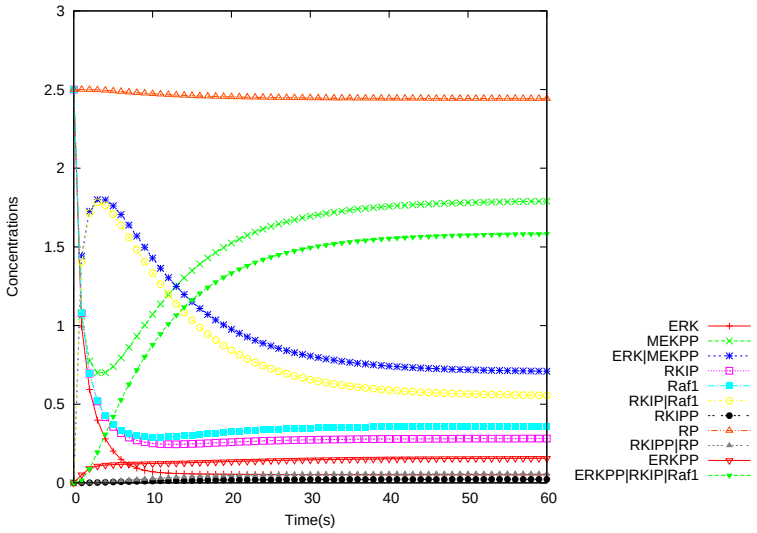


Fig. 2: Behaviour of all species in *RKIP* signaling pathway.

the two compared modelling variants, which allows a fair comparison between two modelling variants in each pair in terms of performance on generation of synthetic models.

We investigate both alternative topologies and similar topologies. By analyzing the number of reactions in target pathway to the ones existing in the model generated by our method, we can quantitatively measure the difference between the alternative topology compared to the target one. A similar topology contains parts identical to the ones in the target network, but it could as well contain parts which are absent in the target network. An alternative topology has a different structure for the network, for instance could contain the same reactants as the target one, but the arcs between them are different.

Table 1: Simulation settings for running modelling variants.

Modelling Variants	Hybrid Modelling	ES	SA	Gaussian $N(\mu, \sigma)$
Data Driven:	♯Runs = 10	GeSi = 100	T_{ini} = 10	μ= 0
Fixed vs Dynamic	Sub@Ge= 2	PopSi = 50	CoRate = 0.8	σ= 0.00001
Survival Selection:	OptRate@Ge = 25		T_{min} = 1	
SES vs PES	ε_1=0.01		Iter = 10	
Mutation:	ε_2=1000			
Fixed vs Random				
Recombination:				
Best vs Random				
Fitness Function:				
ED vs (ED+RP)				

6.2 Statistical analysis

Two statistical measures in the *R* packages [16], 'var.test(X, Y)' and 't.test(X, Y)', are employed to perform the statistical analysis.

Fitness values, compression and coverage scores are used to calculate the P-value in 'var.test(X, Y)' and 't.test(X, Y)' for further statistical analysis. The P-value is compared with a traditional significant level 'p=0.05', and the ratios of variances among generated models are also compared (see Tables 2, 3, 4). Results over 10 independent runs are summarized.

Table 2: Statistical analysis of average fitness sets

NO.	X vs Y	var.test(X, Y)		t.test(X, Y)		
		P-value	$r_{Variances}$	P-value	X	Y
1.1	Dri_{Fixed} vs Dri_{Dyn}	0.0229	0.6309	< 2.2e-16		3.1602
1.2	SES vs PES	0.4574	1.1616	0.837		4.2289
1.3	M_{Fixed} vs M_{Ran}	0.6821	0.9208	0.0262	4.2474	4.035
1.4	$\otimes_{Ran}$ vs $\otimes_{Best}$	1.07e-03	1.9448	0.5737		4.2019
1.5	ED vs (ED+RP)	< 2.2e-16	6.15e-06	< 2.2e-16		348.78

Table 3: Statistical analysis of average compression.

NO.	X vs Y	var.test(X, Y)		t.test(X, Y)		
		P-value	$r_{Variances}$	P-value	X	Y
1.1	Dri_{Fixed} vs Dri_{Dyn}	0.0096	0.4713	< 2.2e-16		0.025
1.2	SES vs PES	0.0461	1.7802	6.78e-16		0.0361
1.3	M_{Fixed} vs M_{Ran}	0.75	1.0958	0.0296	0.0526	0.0567
1.4	$\otimes_{Ran}$ vs $\otimes_{Best}$	1.60e-06	0.2387	< 2.2e-16		0.1033
1.5	ED vs (ED+RP)	1.25e-05	3.6546	0.0004		0.0469

Table 4: Statistical analysis of average coverage.

NO.	X vs Y	var.test(X, Y)		t.test(X, Y)		
		P-value	$r_{Variances}$	P-value	X	Y
1.1	Dri_{Fixed} vs Dri_{Dyn}	6.74e-12	8.4369	< 2.2e-16		0.0731
1.2	SES vs PES	0.4961	1.2161	0.0261		0.2065
1.3	M_{Fixed} vs M_{Ran}	0.062	1.7147	6.63e-05	0.2322	0.2765
1.4	$\otimes_{Ran}$ vs $\otimes_{Best}$	0.3373	1.3178	0.1888		0.2174
1.5	ED vs (ED+RP)	9.39e-05	0.3163	1.05e-14		0.3967

Table 5 shows a comparative example of the reactions obtained in a model generated by ES-SA strategies compared with the ones in the real (target) model.

In the case presented here, four reactions marked with a star in target *RKIP* pathway are generated in the synthetic model. The synthetic model consists of 12 reactions, four of them being identical to the ones in *RKIP* pathway. The ES-SA metaheuristics can obtain alternative topologies exhibiting similar behaviour to the target ones.

Alternative topologies in synthetic models illustrate target biochemical system in a different way, providing templates to biologists in wet-lab for further experimental examination at the properties of the biochemical systems.

Table 5: Comparison of one synthetic model with *RKIP* pathway.

Reactions in *RKIP* pathway	Reactions in One Generated Model
$*Raf1 + RKIP \xrightarrow{k1} RKIP\|Raf1$	$ERK\|RP \xrightarrow{r1} ERKP + RP$
$*RKIP\|Raf1 \xrightarrow{k2} Raf1 + RKIP$	$ERKPP\|MEKPP \xrightarrow{r2} ERKPP + MEKPP$
$RKIP\|Raf1 + ERKPP \xrightarrow{k3} ERKPP\|RKIP\|Raf1$	$ERK\|RP + ERKPP\|RKIPP \xrightarrow{r3} ERK\|ERKPP\|RKIPP\|RP$
$ERKPP\|RKIP\|Raf1 \xrightarrow{k4} RKIP\|Raf1 + ERKPP$	$ERK + RKIP\|Raf1 \xrightarrow{r4} ERK\|RKIP\|Raf1$
$ERKPP\|RKIP\|Raf1 \xrightarrow{k5} Raf1 + ERK + RKIPP$	$*RKIP + Raf1 \xrightarrow{r5} RKIP\|Raf1$
$*ERK + MEKPP \xrightarrow{k6} ERK\|MEKPP$	$*ERK + MEKPP \xrightarrow{r6} ERK\|MEKPP$
$*ERK\|MEKPP \xrightarrow{k7} ERK + MEKPP$	$ERKPP\|MEKPP + MEKPP\|RKIPP \xrightarrow{r7} ERKPP\|MEKPP\|RKIPP$
$ERK\|MEKPP \xrightarrow{k8} MEKPP + ERKPP$	$RKIP + ERK\|RP \xrightarrow{r8} ERK\|RKIP\|RP$
$RKIPP + RP \xrightarrow{k9} RKIPP\|RP$	$*RKIP\|Raf1 \xrightarrow{r9} RKIP + Raf1$
$RKIPP\|RP \xrightarrow{k10} RKIP + RP$	$ERK\|MEKPP \xrightarrow{r10} ERKP + MEKPP$
$RKIPP\|RP \xrightarrow{k11} RKIPP + RP$	$RKIP\|Raf1 + ERKP \xrightarrow{r11} ERKP\|RKIP\|Raf1$
	$*ERK\|MEKPP \xrightarrow{r12} ERK + MEKPP$

6.3 Discussion

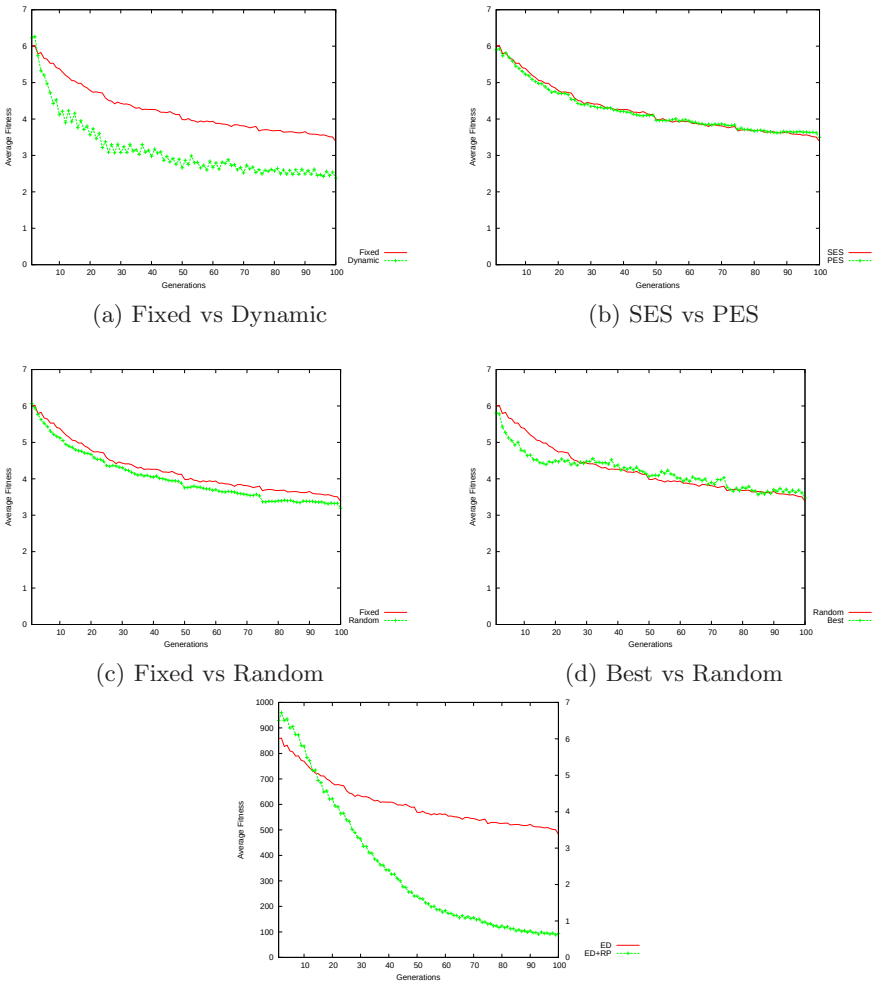
Details of the advantage and disadvantage of applying ES-SA variants to construct models are described below, each pair being considered separately.

Fixed vs Dynamic - Data driven For generating desired behaviour and alternative topologies, dynamic variant, is better than fixed one, but for generating similar topologies, the fixed variant is better than dynamic one.

Figure 3a shows that the dynamic version converges more quickly in terms of fitness function than the fixed one.

SES vs PES - Survival selection For generating desired behaviour, the experiments do not show any difference between SES and PES; for generating similar topologies, SES is better than PES and for generating alternative topologies, SES is better than PES.

Figure 3b shows that SES and PES have a similar performance regarding the convergence of fitness values.



(a) Fixed vs Dynamic

(b) SES vs PES

(c) Fixed vs Random

(d) Best vs Random

(e) ED vs ED+RP

Fig. 3: Fitness convergence of ES-SA variants (the number of generations is shown on the X-axis and the average fitness values on the Y-axis): (a) variants for data driven, Fixed vs Dynamic; (b) variants for survival selection, SES vs PES; (c) variants for applying mutation operator, Fixed vs Random; (d) variants for applying crossover operator, Best vs Random; (e) variants for models estimation, ED vs ED+RP.

Fixed vs Random - Mutation operator For generating desired behaviour and similar topologies random variant is better than fixed one; and for alternative topologies random variant is the same as fixed one.

Figure 3c shows the convergence of the fitness values for the fixed and random variant.

Best vs Random - Crossover operator For generating desired behaviour and similar topologies, a random selection of mate for recombination works the same as the selection of the best individual but selection of best individual for recombination is better than the random selection for generating alternative topologies. Figure 3d shows the convergence of the fitness values. In Table 2 (1.4) and Table 4 (1.4), the two P-values of t.test() are both larger than the significant level 0.05, indicating that the mean fitness and coverage values of the random variant are the same as the ones of the best variant. This suggests that the best and random mechanisms of selecting individual for crossover have the same performance.

ED vs ED+RP For generating similar topologies, ED+RP variant is better than ED but ED is better for generating alternative topologies. In Table 2 (1.5), the P-value is much smaller than 0.05, indicating a significant difference between ED and ED+RP. Figure 1(e) presents the convergence of the fitness value for ED and ED+RP. The average coverage value is larger for the models estimated by ED+RP which suggests that the ED+RP variant can be better than the ED variant in terms of generating similar topologies. However the P-value of var.test() in Table 3 (1.5) is smaller than 0.05 and the ratio of variances is larger than 1.

Note that some of the ES-SA variants are not directly comparable, because the statistical values are not in the same measurement scale. For instance, the ED and ED+RP are not comparable in terms of fitness values, since the mechanism of reward and penalty generates a different fitness scale.

We are aware that sometimes small amendments to the original methods could have an impact upon the final results; this is what we tried to prove in this paper, but with the aim of selecting those forms of operators and evaluation procedures would best fit the biochemical network design. The probabilistic ES makes no difference to the standard one (it is even worse in certain situations) which shows that accepting worse solutions will not bring additional exploration of the search space. The manner in which mutation is performed helps if additional information is known about the problem to be solved. In the case presented in this paper, the imposition of a certain number of steps for adding a component or removing a component is not helpful. This could work better than in the random case if more interaction is provided, i.e. remove a component every certain fixed number of steps only if the size of the network is too big. The step size of the application of an addition or a subtraction is also important, but that requires extra analysis. Elitism plays an important role and in our case it helped in selection the individuals for crossover.

Table 6 shows the overall pair-wise comparison of all the five variants in terms of topologies generation and behaviour.

Table 6: A summary of performance between compared modelling variants.

Modelling Variants	Desired Behaviours	Similar Topologies	Alternative Topologies
Data Driven: Fixed vs Dynamic	Dynamic	Fixed	Dynamic
Survival Selection: SES vs PES	=	SES	SES
Mutation: Fixed vs Random	Random	×	=
Recombination: Best vs Random	=	×	Best
Fitness Function: ED vs (ED+RP)	×	ED+RP	ED

Notes: '×' means not comparable; '=' means the same.

7 Summary and conclusions

The work described in this paper focuses on the empirical analysis of piecewise modelling approaches of signalling pathways, comparing performance of different Evolutionary Strategies – Simulated Annealing variants. Alternative topologies of synthetic models obtained *in silico* can be taken as general guides for biologists to examine and understand biochemical systems by experimental techniques in wet-lab. Moreover, these can be used as templates for researchers in synthetic biology to develop specific functions of biochemical systems. The research presented here aims at guiding biomodel engineers in deciding the computational setup and selecting the right parameters. Our analysis of some of the combinations which could be considered helps in developing models that are useful for further construction with respect to specific characteristics of modelling biochemical systems.

Acknowledgements

The authors acknowledge the scientific support form Monika Heiner (Brandenburg University of Technology Cottbus). Part of this work has been supported by the Romanian National Authority for Scientific Research PN-II-PT-PCCA-2011-3.2-0917.

References

1. E. Aarts, J. Korst, W. Michiels, Simulated Annealing and Boltzmann Machines: a stochastic approach to combinatorial optimization and neural computing, Wiley, pp. 188-202, 1989
2. M. Baker, Synthetic genomes: the next step for the synthetic genome. Nature, 473, 403408 2011

3. A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen, Pattern discovery in biosequences, ICGI98, pp. 257270, 1998.
4. M. Calder, S. Gilmore, J. Hillston, Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA, In: Trans. Computational Systems Biology, pp. 1-23, Springer, 2004.
5. K. H. Cho, S. Y. Shin, H. W. Kim, O. Wolkenhauer, B. Mcferran, W. Kolch, Mathematical modeling of the influence of RKIP on the ERK signaling pathway, In: C. Priami, editor, Computational Methods in Systems Biology (CSMB'03), pp. 127-141, Springer, 2003.
6. G. Fogel, D. Corne, Evolutionary Computation in Bioinformatics, Morgan Kaufmann, 2003, pp. 256-276
7. D. Gilbert, D. Westhead, and J. Viksna, Techniques for comparison, pattern matching and pattern discovery: from sequences to protein topology, In: P. Frasconi and R. Shamir, Eds., Artificial Intelligence and Heuristic Methods in Bioinformatics, pp. 128147, IOS Press, 2003.
8. D. Gilbert, M. Heiner, S. Lehrack, A Unifying Framework for Modelling and Analysing Biochemical Pathways Using Petri Nets. In proceedings CMSB 2007 (Computational Methods in Systems Biology), Editors: M.Calder and S.Gilmore, Springer LNCS/LNBI 4695, pp. 200-216, 2007.
9. K.S. Lau, A.M. Juchheim, K.R. Cavaliere, S.R. Philips, D.A. Lauffenburger, K.M. Haigis, In vivo systems analysis identifies spatial and temporal aspects of the modulation of TNF-alpha-induced apoptosis and proliferation by MAPKs. Sci. Signal., Vol. 4, No. 165, ra16, 2011.
10. X. Liu, J. Jiang, O. Ajayi, X. Gu, and D. Gilbert, BioNessie(G)- A Grid Enabled Biochemical Networks Simulation Environment, Studies in Health Technology and Informatics, Vol. 138, pp. 147157, 2008.
11. T. Murata, Petri nets: Properties, analysis and applications, Proceedings of the IEEE, Vol. 77, No. 4, pp. 541580, 1989.
12. S. Rausanu, C. Grosan, Z. Wu, O. Parvu, D. Gilbert, D., Evolving Biochemical Systems, IEEE Congress on Evolutionary Computation (CEC), Cancun, Mexico, 2013.
13. E. Sakamoto, H. Iba, Inferring a system of differential equations for a gene regulatory network by using genetic programming, In Proceedings of the IEEE Congress on Evolutionary Computation, IEEE Service Center, Piscataway, N.J., 2000
14. E.C. OShaughnessy, S. Palani, J.J. Collins, C.A. Sarkar, Tunable signal processing in synthetic MAP kinase cascades. Cell, Vol. 144, No. 1, pp. 119-131, 2011.
15. Elowitz, M.B., Leibler, S.: A synthetic oscillatory network of transcriptional regulators. Nature, 403, pp. 335-338, 2000.
16. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2009.
17. K. Yeung, P. Janosch, B.McFerran, D.W. Rose, H.Mischak, J.M. Sedivy, W. Kolch. "Mechanism of suppression of the Raf/MEK/Extracellular signal regulated kinase pathway by the Raf kinase inhibitor protein". Molecular and Cellular Biology, Vol. 20, No. 9, pp. 3079-3085, 2000.
18. K. Yeung, T. Seitz, S. Li, P. Janosch, B. McFerran, C. Kaiser, F. Fee, K. D. Katsanakis, D. W. Rose, H. Mischak, J. M. Sedivy, W. Kolch. Suppression of Raf-1 kinase activity and MAP kinase signaling by RKIP, Nature, Vol. 401, pp. 173-177, 1999.
19. Z. Wu, A generic approach to behaviour-driven biochemical model construction, PhD Thesis, Brunel University, 2013.

20. Z. Wu, Q. Gao and D. Gilbert, Target Driven Biochemical Network Reconstruction Based on Petri nets and Simulated Annealing, In proceedings CMSB 2010 (8th International Conference on Computational Methods in Systems Biology), pp 33–42, ACM Digital Library, 2010.
21. Z. Wu, S. Yang, D. Gilbert, A Hybrid Approach to Piece-wise Modelling of Biochemical Systems, 12th International Conference on Parallel Problem Solving From Nature, LNCS 7491, pp. 519-528, 2012.