Comparing Various Evolutionary Algorithms on the Parameter Optimization of the Valine and Leucine Biosynthesis in *Corynebacterium glutamicum*

Andreas Dräger*‡ Jochen Supper* Hannes Planatscher* Jørgen B. Magnus† Marco Oldiges† Andreas Zell*

***Abstract*— Parameter estimation for biochemical model systems has become an important problem in systems biology. Here we focus on the metabolic subnetwork of the valine and leucine biosynthesis in *C. glutamicum*. Due to the lack of indisputable information regarding reversibility of the reactions in the pathway we derived two alternative ordinary differential equation models based on the formalisms of the generalized mass-action rate law. We introduced two alternative modeling approaches for feedback inhibition and evaluated the applicability of six optimization procedures (multi start Hill Climber, binary and real valued Genetic Algorithm, standard and covariance matrix adaption Evolution Strategy as well as Simulated Annealing) to the problem of parameter fitting. The model considering irreversible reactions performed worse and was therefore rejected from further analysis. We benchmarked the impact of different mutation and crossover operators as well as the influence of the population size on the remaining system and the two best optimization procedures namely binary Genetic Algorithm and the Evolution Strategy. The GA performed best on average and found the best total result based on the relative squared error.**

I. Introduction

Modeling the dynamic behavior of complex biological reaction systems has become a challenging task in systems biology. The parameters in those models describe constant enzyme properties. These can be measured *in vitro* experimentally. However, this is expensive, time consuming and also often impractical. As the parameters define the thermodynamical dynamic behavior of the system, a common method is the inference of suitable parameter values so that the resulting curves approximate the measurements. This method is based on the assumption that biological systems are optimized for the given environmental circumstances. Thus, the parameters represent the optimal set of enzyme features. Many studies have constructed a set of differential equations or postulated further network properties [1]–[5]. Due to the high nonlinearity of most common model equations, Evolutionary Algorithms (EA) have been successfully applied to similar problems [2], [3]. However, less attention has been drawn to the optimization process of the model parameters. During the last decades many derivatives of EAs have been proposed. Each of them has certain advantages and is therefore more or less appropriate for a specific problem.

Here we consider a metabolic network of the valine (Val) and leucine (Leu) biosynthesis in *C. glutamicum* based on *in vivo* data obtained in a glucose stimulus-response experiment with the use of a rapid sampling device and advanced mass spectrometry. Measurements of 13 metabolites were taken at sub-second intervals for a time period of 25 s, showing a high degree of accuracy [4]. Two mathematical models were developed using the formalism of the mass-action kinetics including different generic formulas for feedback inhibition for cases in which the exact mechanism of the inhibition is unknown. The Val/Leu biosynthesis pathway has already been modeled using linlog kinetics [4], which is an approximation of the metabolic process [6]. By introducing a model description based on mass-action kinetics the formalism is similar to the traditional biochemical modeling.

To explore the applicability of various optimization algorithms to this problem it is necessary to study the influence of their different settings. Four Evolutionary Algorithms (binary and real valued Genetic Algorithm as well as standard and covariance matrix adaption Evolution Strategy) and two non-evolutionary optimization algorithms (multi start Hill Climber and Simulated Annealing) were applied to this inference problem. These algorithms were tested systematically on the data set using default settings. The two best performing algorithms and the best performing model were selected and investigated subsequently in more detail. The influence of different mutation and crossover operators, probabilities p_m and p_c as well as the impact of the population size was examined and compared. The qualities and drawbacks of both model descriptions and optimization procedure were exposed.

II. Methods

A. The System under Consideration

1) Biochemical Model: C. glutamicum is an important industrial producer species of amino acids. Due to this fact it is desirable to gain a better understanding of the chemical processes during the formation of its metabolic products. Fig. 1 shows the biochemical reactions of the Val and Leu biosynthesis according to the MetaCyc database [8] and Magnus *et al.* [4]. Our consideration of the Val and Leu biosynthesis starts with pyruvate (Pyr), which is subsequently consumed to form 2-ketoisovalerate (KIV) in two reactions. There are two different ways to form Val and one to convert KIV to 2-isopropylmalate (2IPM). The latter is the starting substance for the Leu production in four following reaction steps. Both Val and Leu can be used for biomass production or can be pumped out of the cell if not needed. Here we only consider the transport out of the cell, which is the industrially

*Center for Bioinformatics Tübingen (ZBIT), 72076 Tübingen, Germany
†Forschungszentrum Jülich, Institute of Biotechnology, Germany
‡Corresponding author, andreas.draeger@uni-tuebingen.de

1-4244-1340-0/07/$25.00

interesting reaction. In four feedback loops Val and Leu downregulate their own production rate. The transport of Leu and Val accross the cell wall is actually performed by one enzyme, for which both substrates compete. For modeling purposes two distinct reactions are necessary in which the competition is included as inhibition.

Since the reaction $2\,\mathrm{IPM} \rightleftharpoons 3\,\mathrm{IPM}$ is fast, it is assumed to be in equilibrium. This and the two following reactions $3\,\mathrm{IPM} + \mathrm{NAD}^+ \longrightarrow 2\,\mathrm{I_3OS} + \mathrm{NADH_2}$ and (2S)-2-isopropyl-3-oxosuccinate ($2\,\mathrm{I_3OS}$) $\longrightarrow$ 2-ketoisocaproate (KIC) $+\mathrm{CO_2}$ that depend only on the concentration of 2 IPM were lumped together introducing the symbol IPM for both derivates as suggested by Magnus *et al.* [4]. The KEGG database [9] mentions two additional reaction steps not included in METACYC [8]: Pyr reacts to 2-hydroxyethyl-thio-dipyrophosphate first before forming (S)-2-acetolactate (AcLac) which then turns over in 3-hydroxy-3-methyl-2-oxobutanoate before it further reacts to (R)-2,3-dihydroxy-3-metylbutanoate (DHIV).

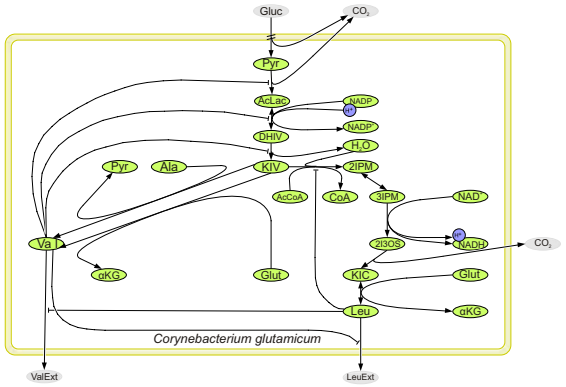


Fig. 1. Process diagram of the Val and Leu synthesis in *C. glutamicum* Metabolites outside the cell are not directly included in the model system. Both amino acids can be formed from Pyr, the end product of the glycolysis.

2) Glucose Stimulus-Response Experiment: After a 10 min starvation period a glucose impulse was added to the culture medium increasing the glucose concentration from 0 to 3.5 g/l. This caused a glucose shock leading to a spontaneous increase of metabolic products linked to this central nutrient. 47 samples of the fermenter broth were taken over 25 s starting 4 s before the glucose pulse. Immediate quenching and cooling with methanol to -50 °C prevented the metabolites from further reactions. Mass spectrometry (HPLC MS/MS) was used to quantify the metabolite concentrations in the probes [4].

No measurements have been taken for intermediates of the glycolysis before Pyr. It was technically not possible to obtain measurements of $\mathrm{NADH_2}$ and $\mathrm{NADPH_2}$ with a high degree of exactness. Due to this fact we follow the suggestion of Magnus *et al.* [4] taking into account that both couples $\mathrm{NAD^+}$ and $\mathrm{NADH_2}$ as well as $\mathrm{NADP^+}$ and $\mathrm{NADPH_2}$ follow a conservation relation so that the total amount of

TABLE I

THE REACTION SYSTEM IN MORE DETAIL

The reactions in KEGG were lumped together to result in this reaction scheme which is in accordance with METACYC besides the question of irreversibility. Two of the resulting reactions contain reversible steps.

No.	Reaction	Enzyme	Inhib.
R_1	$2\,\mathrm{Pyr} \longrightarrow \mathrm{AcLac} + \mathrm{CO_2}$	AHAS	Val
R_2	$\mathrm{AcLac} + \mathrm{NADPH_2} \rightleftharpoons \mathrm{DHIV} + \mathrm{NADP^+}$	AHAIR	Val
R_3	$\mathrm{DHIV} \longrightarrow \mathrm{KIV} + \mathrm{H_2O}$	DHAD	Val
R_4	$\mathrm{KIV} + \mathrm{Glut} \longrightarrow \mathrm{Val} + \alpha\mathrm{KG}$	$\mathrm{BCAAT_{ValB}}$	
R_5	$\mathrm{KIV} + \mathrm{Ala} \longrightarrow \mathrm{Val} + \mathrm{Pyr}$	$\mathrm{BCAAT_{ValC}}$	
R_6	$\mathrm{Val} \longrightarrow \mathrm{Val_{ext}}$	$\mathrm{Trans_{Val}}$	Leu
R_7	$\mathrm{KIV} + \mathrm{AcCoA} \longrightarrow \mathrm{IPM} + \mathrm{CoA}$	IPMS	Leu
R_8	$\mathrm{IPM} + \mathrm{NAD^+} \longrightarrow \mathrm{KIC} + \mathrm{NADH_2} + \mathrm{CO_2}$	IPMDH	
R_9	$\mathrm{KIC} + \mathrm{Glut} \rightleftharpoons \mathrm{Leu} + \alpha\mathrm{KG}$	$\mathrm{BCAAT_{LeuB}}$	
R_{10}	$\mathrm{Leu} \longrightarrow \mathrm{Leu_{ext}}$	$\mathrm{Trans_{Leu}}$	Val

both coupled metabolites remains constant during the 25 s of interest. Thus, we can assume that the concentration of $\mathrm{NADH_2}$ equals $0.8\,\mathrm{mM} - [\mathrm{NAD^+}]$ and $[\mathrm{NADPH_2}] = 0.04\,\mathrm{mM} - [\mathrm{NADP^+}]$. It has also not been possible to measure the concentration of AcetylCoA and CoA in sufficient quality. We assume there is a constant pool of these central metabolites involved in many reactions and this pool does not vary over the short time period of 25 s. Depending on what model we use this yields either an additional parameter to be estimated or these constants can be lumped together with another constant. The steady-state concentrations [4] of the seven metabolites to be simulated serve as initial values for the models.

B. Mathematical Modeling

We benchmarked a reversible generalized mass-action kinetics (GMAK) model as well as its irreversible alternative with a different inhibition function on this metabolic system.

The topology of the reaction system (Tab. I) can be represented by a stoechiometric matrix $\mathbf{N}$. The change of the metabolite concentrations over time can be calculated by combining $\mathbf{N}$ with the vector of reaction velocities $\mathbf{v}$ linearly

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{S} = \mathbf{N}\mathbf{v}(\mathbf{S}(t), t, \mathbf{p})$$

with $\mathbf{S}$ being the vector of reacting species and the parameter vector $\mathbf{p}$. The equations (1) through (7) show the linear combination of the reaction velocities to the resulting metabolite concentrations for this reaction system.

$$\frac{\mathrm{d}[\mathrm{AcLac}]}{\mathrm{d}t} = v_1 - v_2 \quad (1)$$

$$\frac{\mathrm{d}[\mathrm{DHIV}]}{\mathrm{d}t} = v_2 - v_3 \quad (2)$$

$$\frac{\mathrm{d}[\mathrm{KIV}]}{\mathrm{d}t} = v_3 - v_4 - v_5 - v_7 \quad (3)$$

$$\frac{\mathrm{d}[\mathrm{Val}]}{\mathrm{d}t} = v_4 + v_5 - v_6 \quad (4)$$

$$\frac{\mathrm{d}[\mathrm{IPM}]}{\mathrm{d}t} = v_7 - v_8 \quad (5)$$

$$\frac{\mathrm{d[KIC]}}{\mathrm{d}t} = v_8 - v_9 \quad (6)$$

$$\frac{\mathrm{d[Leu]}}{\mathrm{d}t} = v_9 - v_{10} \quad (7)$$

1) Generalized Mass-Action Kinetics (GMAK): To model enzyme catalyzed reactions with more than one substrate one has to consider the exact reaction mechanism in the kinetic model which is often unknown. Due to the fact that the Michaelis-Menten kinetics is a special case of the GMAK it may be desired to neglect the enzyme catalysis. The mechanism can be described by a mass-action kinetics instead. However, if any kind of inhibition is involved in the reaction, this cannot easily be included in the kinetic equation. Here we apply an inhibition function that fits in the generalized mass-action rate law as proposed by Schauer and Heinrich (1983) [10]

$$v_j(\mathbf{S},\mathbf{p}) = F_j(\mathbf{S},\mathbf{p}) \left(k_{+j} \prod_i S_i^{n_{ij}^-} - k_{-j} \prod_i S_i^{n_{ij}^+} \right) \quad (8)$$

The function $F_j(\mathbf{S},\mathbf{p})$ was defined as any positive function of the substrate concentrations $\mathbf{S}$ and the parameter vector $\mathbf{p}$ to introduce saturation or inhibition effects to the common mass-action kinetics written in brackets [10]. For convenience of notation the matrices $\mathbf{N}^{\pm}$ were introduced, whose elements $n_{ij}^{\pm}$ express the absolute values of the positive or negative stochiometric coefficients, respectively.

Feedback inhibition loops can be included using one of the following approaches

$$F_j(\mathbf{S},\mathbf{p}) = \frac{1}{1 + K_j^{\mathrm{I}} \cdot [\mathrm{I}]} \quad (9)$$

$$F_j(\mathbf{S},\mathbf{p}) = \exp(-K_j^{\mathrm{I}} \cdot [\mathrm{I}]) \quad (10)$$

with $K_j^{\mathrm{I}} \geqslant 0$. The inhibition term (9) was derived from the competing reactions of the enzyme with its substrate or the inhibitor, respectively. Using the equilibrium constant for the inhibition reaction and the conservation law of the enzyme and the enzyme-inhibitor complex concentrations yields the first equation. Applying Eq. (8) combined with Eq. (9) to reaction system R_1 through R_{10} leads to an ODE system with 24 parameters $k_{\pm i}$, K_j^{I}:

$$v_1 = \frac{k_{+1}[\mathrm{Pyr}]^2 - k_{-1}[\mathrm{AcLac}]}{1 + K_1^{\mathrm{I}}[\mathrm{Val}]} \quad (11)$$

$$v_2 = \frac{k_{+2}[\mathrm{AcLac}][\mathrm{NADPH_2}]}{1 + K_2^{\mathrm{I}}[\mathrm{Val}]} - \frac{k_{-2}[\mathrm{DHIV}][\mathrm{NADP^+}]}{1 + K_2^{\mathrm{I}}[\mathrm{Val}]} \quad (12)$$

$$v_3 = \frac{k_{+3}[\mathrm{DHIV}] - k_{-3}[\mathrm{KIV}]}{1 + K_3^{\mathrm{I}}[\mathrm{Val}]} \quad (13)$$

$$v_4 = k_{+4}[\mathrm{KIV}][\mathrm{Glut}] - k_{-4}[\mathrm{Val}][\alpha\mathrm{KG}] \quad (14)$$

$$v_5 = k_{+5}[\mathrm{KIV}][\mathrm{Ala}] - k_{-5}[\mathrm{Val}][\mathrm{Pyr}] \quad (15)$$

$$v_6 = \frac{k_{+6}[\mathrm{Val}]}{1 + K_4^{\mathrm{I}}[\mathrm{Leu}]} \quad (16)$$

$$v_7 = \frac{k_{+7}[\mathrm{KIV}][\mathrm{AcCoA}] - k_{-7}[\mathrm{IPM}][\mathrm{CoA}]}{1 + K_5^{\mathrm{I}}[\mathrm{Leu}]} \quad (17)$$

$$v_8 = k_{+8}[\mathrm{IPM}][\mathrm{NAD^+}] - k_{-8}[\mathrm{KIC}][\mathrm{NADH_2}] \quad (18)$$

$$v_9 = k_{+9}[\mathrm{KIC}][\mathrm{Glut}] - k_{-9}[\mathrm{Leu}][\alpha\mathrm{KG}] \quad (19)$$

$$v_{10} = \frac{k_{+10}[\mathrm{Leu}]}{1 + K_6^{\mathrm{I}}[\mathrm{Val}]} \quad (20)$$

2) GMAK, irreversible with exp *Inhibition:* Eq. (10) was derived more intuitively driven by the assumption that the exponent function constitutes an important growth and shrinkage function in biology. By setting all product concentrations apart from R_2 and R_9 to zero and applying Eq. (10) to Eq. (8) we obtain the irreversible version of this equation system with 18 parameters $k_{\pm i}$, K_j^{I}:

$$v_1 = k_{+1}[\mathrm{Pyr}]^2 \exp(-K_1^{\mathrm{I}}[\mathrm{Val}]) \quad (21)$$

$$v_2 = \exp(-K_2^{\mathrm{I}}[\mathrm{Val}]) \cdot (k_{+2}[\mathrm{AcLac}][\mathrm{NADPH_2}] - k_{-2}[\mathrm{DHIV}][\mathrm{NADP^+}]) \quad (22)$$

$$v_3 = k_{+3}[\mathrm{DHIV}] \exp(-K_3^{\mathrm{I}}[\mathrm{Val}]) \quad (23)$$

$$v_4 = k_{+4}[\mathrm{Glut}][\mathrm{KIV}] \quad (24)$$

$$v_5 = k_{+5}[\mathrm{Ala}][\mathrm{KIV}] \quad (25)$$

$$v_6 = k_{+6}[\mathrm{Val}] \exp(-K_4^{\mathrm{I}}[\mathrm{Leu}]) \quad (26)$$

$$v_7 = k_{+7}[\mathrm{KIV}] \exp(-K_5^{\mathrm{I}}[\mathrm{Leu}]) \quad (27)$$

$$v_8 = k_{+8}[\mathrm{NAD^+}][\mathrm{IPM}] \quad (28)$$

$$v_9 = k_{+9}[\mathrm{Glut}][\mathrm{KIC}] - k_{-9}[\alpha\mathrm{KG}][\mathrm{Leu}] \quad (29)$$

$$v_{10} = k_{+10}[\mathrm{Leu}] \exp(-K_6^{\mathrm{I}}[\mathrm{Val}]) \quad (30)$$

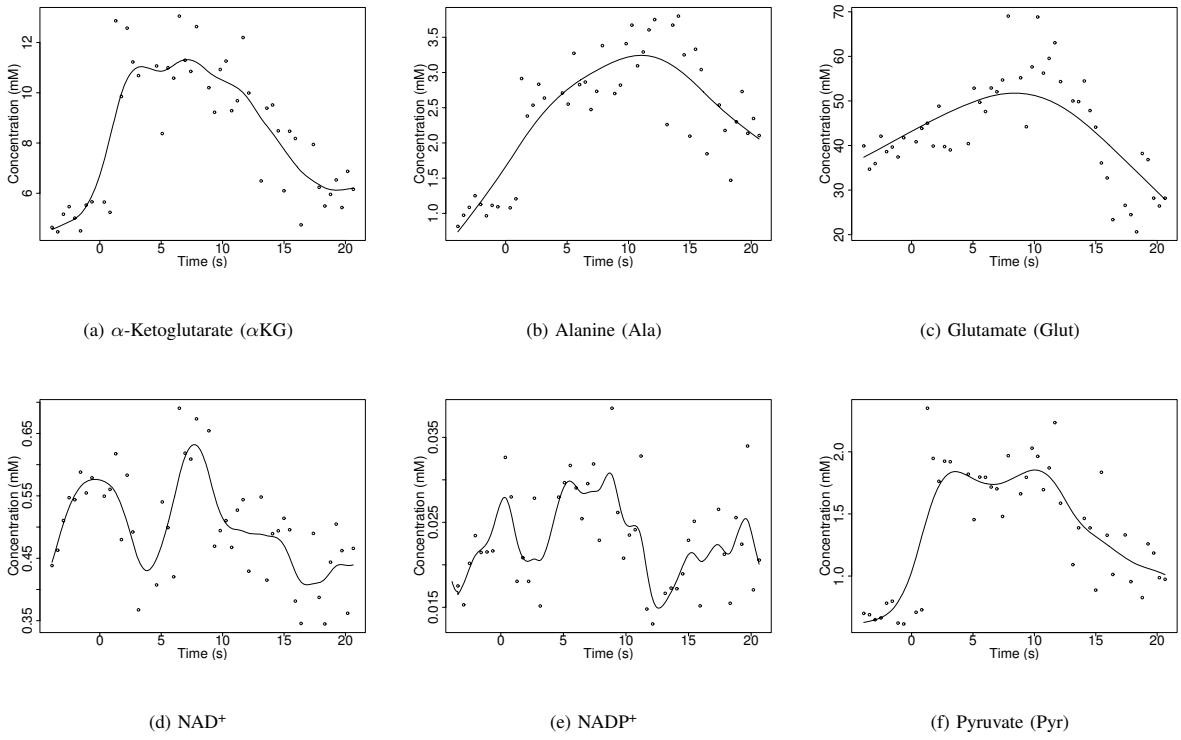
3) Representing external Metabolites with Splines: As suggested by Magnus *et al.* [4], metabolites, whose concentrations cannot be explained by the model itself, are approximated using splines. These metabolites are considered external, i. e., they are an input to the model but they are involved in several other reactions outside this system (Fig. 1).

We used cubic approximation splines to smooth the measurements. The degree of smoothness λ and the weight vector ω define the shape of every spline. To weight all measurements equally, ω_i was set to 1 for all i. Due to the different ranges of the concentrations of the six metabolites it is not possible to find one appropriate parameter λ that leads to equally smooth curves. Hence, we transformed all concentrations into the range $[0, 1]$, set $\lambda = 1$, computed the spline coefficients and retransformed the result back into the original range (Fig. 2).

C. Fitness Function and Search Space Restrictions

Due to the differences in the concentrations of certain metabolites in the system under study the choice of the right fitness function is a crucial step. The Euclidian distance between the model values and the measurements is not applicable here because metabolites in higher concentration would tend to dominate the fitness whereas those in lower concentration would not play any role during the curve fitting procedure. The relative squared error or relative standard error (RSE)

$$f_{\mathrm{RSE}}(\hat{\mathbf{x}},\mathbf{X}) = \sum_{i=1}^{\dim(\hat{\mathbf{x}})} \sum_{t=1}^{T} \left(\frac{\hat{x}_i(\tau_t) - x_{ti}}{x_{ti}} \right)^2 \quad (31)$$



(a) α-Ketoglutarate (αKG) (b) Alanine (Ala) (c) Glutamate (Glut)

(d) NAD^+ (e) $NADP^+$ (f) Pyruvate (Pyr)

Fig. 2. Representing external metabolites using splines

Metabolites whose dynamic change of their concentration cannot be explained in terms of the modeled metabolic subnetwork, were considered external and approximated using splines. The six external metabolites describe the system's input. Splines smooth the fluctuating measurements but do not depend on any biologically relevant model since their coefficients were computed individually for each chemical species.

to be minimized overcomes these limitations. The first sum runs over all dimensions of $\hat{\mathbf{x}}$, i.e., all compounds of the system and T is the number of measurements taken. Here, $\hat{\mathbf{x}}(\tau_t)$ is a vector describing the model output at each sample time τ_t and $\mathbf{X} = (x_{ti})$ is the given data matrix. The fitness (31) has already been used in several publications for similar problems [7].

The fourth order Runge-Kutta method was used to solve the ordinary differential equation systems to obtain $\hat{\mathbf{x}}(\tau_t)$ for every t.

In biology the parameter space is limited to values greater than or equal to zero and cannot exceed the diffusion rate. Therefore, it is necessary to define ranges for every parameter to restrict the search space for the optimizer. Here, all parameter values were limited to the range $[0, 2000]$ covering 98.748 % of all known kinetic parameters in the BRENDA database [11]. In more detail, 99.958 % of all K^{I} values, 99.957 % of all K^{M} values and 96.328 % of all k^{cat} values are lower than 2000. All known parameters in BRENDA are greater than or equal to zero. For the alternative $\exp(-K^{\mathrm{I}}[\mathrm{I}])$ inhibition formalism for the GMAK model we set the search space for $K^{\mathrm{I}} \in [0, 8]$. The focus here was to restrict the search space.

Previous Monte Carlo searches showed that initialization plays an important role due to the high nonlinearity of all considered models. Parameter values choosen completely by chance often lead to instable systems. Hence, all parameters were initialized with low values, assuming that large parameter values are rather infrequent in nature. This assumtion is also supported by the entries of the BRENDA database, showing that 64.807 % of the known parameters are lower than or equal to 2. A Gaussian distribution with $\mu = \sigma = 1$ guarantees low initial values and ensures stable initial populations. Each parameter was set to the boundary values if it broke any of the restrictions of the search space.

D. Standard Settings for the Optimization Algorithms

Following the framework of Spieth *et al.* [7], [12], we applied six optimization algorithms, implemented in the JAVAEVA framework[1] [13] to the inference problem.

- (multi start) Hill Climber (HC): we varied the number of multi starts from 1, 10, 25, 50, 100 to 250. All used mutation with a fixed step size $\sigma = 0.2$ and a mutation probability $p_m = 1.0$.

[1] http://www-ra.informatik.uni-tuebingen.de/software/JavaEvA

- Binary Genetic Algorithm (binGA): we utilized one-point mutation, $p_m = 0.1$, and one-point crossover, $p_c = 0.7$.
- Real valued Genetic Algorithm (realGA): was employed with global mutation, $p_m = 0.1$ and UNDX crossover, $p_c = 0.8$. Both GAs used tournament selection with a group size of 8 in a population of 250 individals.
- Standard Evolution Strategy (stdES): ($\mu = 5$, $\lambda = 25$)-ES used global mutation, $p_m = 0.8$ and discrete one-point crossover, $p_c = 0.2$.
- Evolution Strategy with covariance matrix adaption (cmaES): (5, 25)-ES, the mutation rate was set to $p_m = 1.0$ without crossover. Both ESs applied best selection to choose the next generation.
- Simulated Annealing (SA): We employed the SA with $\alpha = 0.1$ and an initial temperature $T = 5$ using a linear annealing schedule and a population size of 250 individuals.

For all algorithms with population sizes lower than 250 individuals, a pre-population with 250 parameter vectors was generated and the best were selected to generate the initial population. This step is crucial to obtain comparable results for algorithms with different population sizes [7]. Each experiment was repeated 20 times to obtain statistically significant results with 100,000 fitness evaluations.

E. Optimized Settings for the Evolutionary Algorithms

To study the influence of different mutation and crossover operators on the ES and the binGA on the rev. GMAK model system, a grid search was performed with 100,000 fitness evaluations and 20 multi starts.

For the ES the values for p_m and p_c were set to 0.8 and 0.2, respectively. To study the impact of mutation alone, p_m was set to one and p_c to zero. To investigate the influence of crossover without mutation the mutation probability was set to one. In the grid search, the operators correlated mutation (main vector adaption), covariance matrix adaption, local and global mutation as well as the $1/5$ success rule for mutation paired with one- and n-point as well as UNDX crossover were systematically benchmarked.

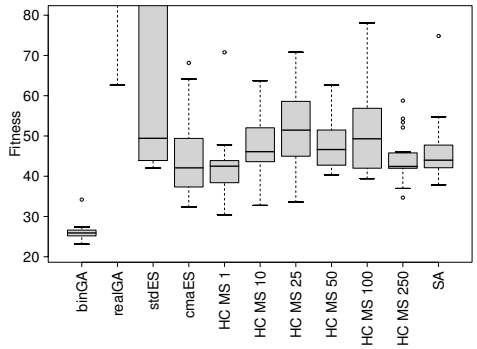
We also performed a grid search on the following mutation operators, using $p_m = 0.1$ and $p_c = 0.7$ evaluating adaptive and one-point mutation paired with bit-simulated, one- and n-point ($n = 3$) as well as uniform crossover. Likewise, to study the influence of crossover and mutation operators alone for the binGA the mutation and crossover probabilities were either set to one or zero depending on what influence was investigated. This setting is reasonable, since neither adaptive mutation, which modifies individual mutation probabilities similar to ES step-size adaption, nor one-point mutation, which flips one randomly choosen bit, inverts all bits of the bit string to the opposite value. The mutation probability is rather the chance of an opperator to be invoked. We selected adaptive muation and bit-simulated crossover for the reversible model and evaluated all pairs of p_m and p_c each varied from 0.0 through 1.0 in 0.1 steps excluding $(0, 0)^T$ in 20 repeats. Subsequently we tested the impact of the population size $\in \{50, 100, 250, 500, 1000\}$ with $(p_m, p_c)^T = (0.2, 1)^T$ each with 20 multi starts.

F. Hard- and Software Configuration

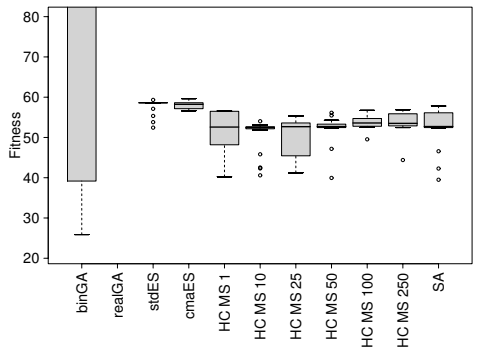
All experiments were run on a cluster with 16 AMD dual Opteron CPUs with 2.4 GHz, 1 MB level 2 cache and 2 GB RAM per node under the Sun Grid Engine and JVM 1.5.0 with Scientific Linux 4 as operating system. An experiment with 20 runs took a computation time of approximately 1.5 h.

III. Results

Fig. 3 shows the capabilities of the aforementioned optimization algorithms on the two optimization problems. All optimization attempts on the irreversible alternative performed significantly worse than on the reversible model. The irreversible model was therefore discarded from further investigations (Tab. II).



(a) GMAK, rev.



(b) GMAK with exp inhibition, irrev.

Fig. 3. Comparison of the different optimization algorithms
The binGA finds the best parameters for both models in total, on the reversible model even on average. The realGA performs worst whereas the other algorithms yield moderate results in similar ranges.

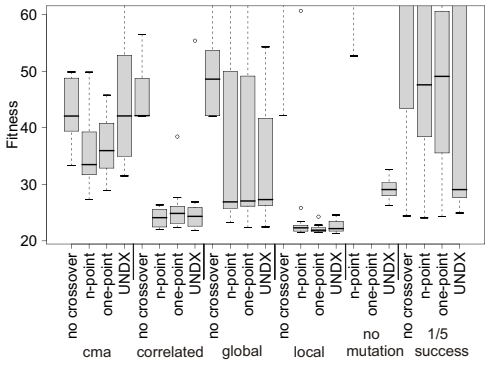
Two algorithms were selected for further analysis: the binGA, that clearly outperformes the other procedures on the remaining system, and the ES, whose median yields the second best result (42.081) when invoked with covariance matrix adaption. However, the median of the HC with 250 multi starts (42.437) is only slighly worse than that of cmaES followed by the HC with only one start (42.497) and SA (43.988). As the differences of the performance ratios between the algorithms mentioned before are very small, but the ES provides a large number of alternative settings, we investigated if the rather bad performance of the stdES can be improved by choosing different mutation and crossover operators.

The realGA, which performes worst, SA and the HC, that yield only moderate results, were therefore discarded from being further analysed.

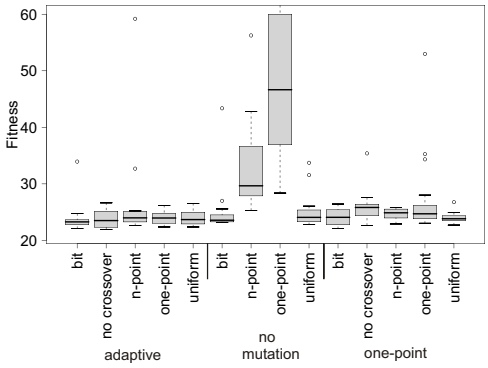
The comparison of different mutation and crossover operators on the selected algorithms ES and binGA (Fig. 4) showed that the rev. GMAK model achieved significantly better fitness values in total using ES or binGA with optimized settings of the algorithms.

Fig. 4 also highlights that the binGA finds reasonable results with lower variance no matter what combination of mutation and crossover operators are applied. The only two exceptions are the combinations of no mutation with one- or n-point crossover. In case of the ES, which performs best with local mutation and n-point crossover in total—best result of 21.481—and with one-point crossover according to the median (21.884), more than half of all operator combinations give median fitness values above 30.

For that reason we also examined the influence of p_m

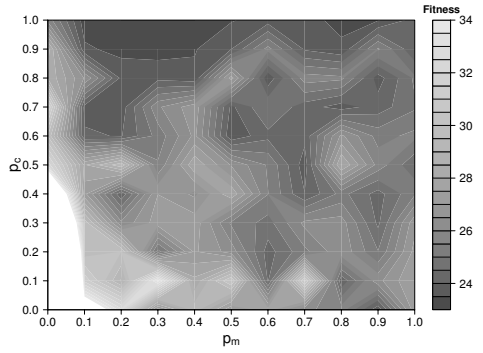


(a) Evolution Strategy

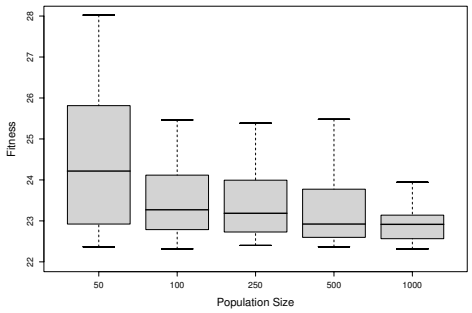


(b) binary Genetic Algorithm

Fig. 4. Dependency of the fitness on mutation and crossover operators
Both box plots depict the dependency of the fitness on different combinations of mutation and crossover operators available for ES (4(a)) or binGA (4(b)), respectively. Both plots were limited to a fitness of 60. The binGA (4(b)) found on average better results than the ES.

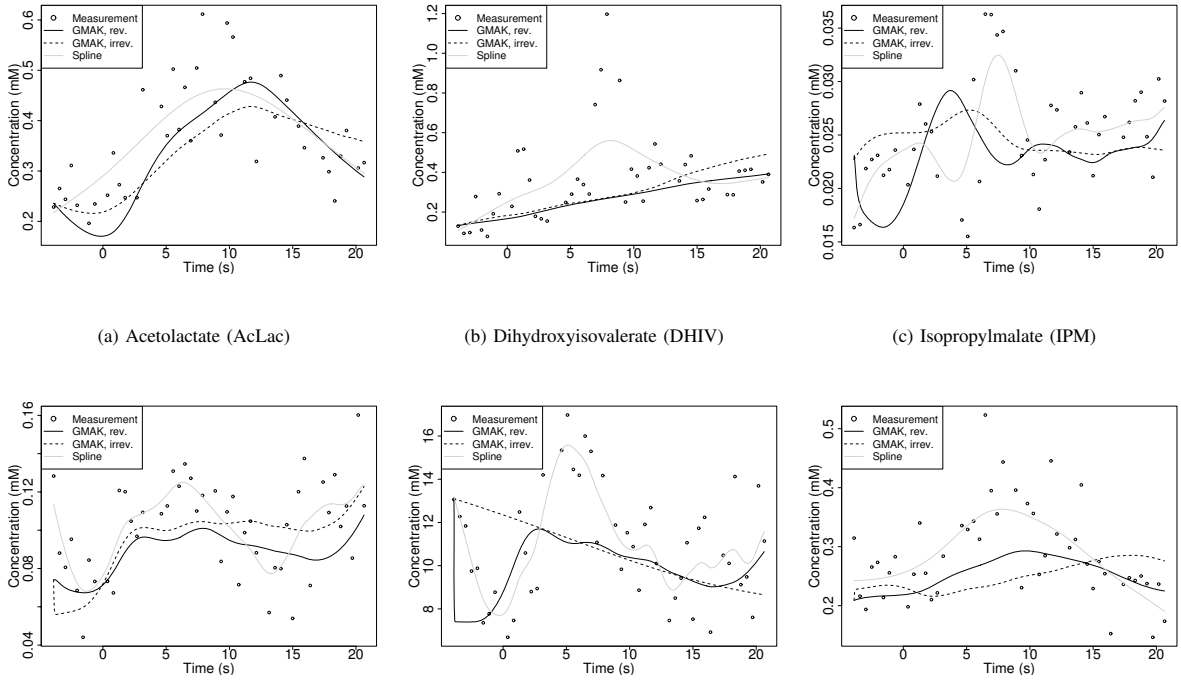


(a) Impact of p_m and p_c rev. GMAK averaged over 20 repeats



(b) Influence of the population size rev. GMAK

Fig. 5. The fitness with respect to p_m and p_c and the population size
A grid search using binGA of the influence of p_c and p_m for the rev. GMAK model 5(a) showed the best average results for $p_m = 0.2$ and $p_c = 1$ using adaptive mutation and bit-simulated crossover. This setting was employed to study the impact of the population size 5(b).



(a) Acetolactate (AcLac)

(b) Dihydroxyisovalerate (DHIV)

(c) Isopropylmalate (IPM)

(d) 2-Ketoisocaproate (KIC)

(e) 2-Ketoisovalerate (KIV)

(f) L-Leucine (Leu)

Fig. 6. The best fit of two different models

For better visual representation, splines were placed in the figure (light gray curves). These were constructed using the same settings as described in Section II-B.3 for external metabolites and give an impression of what one would expect for a good model fit.

and p_c on the rev. GMAK model for the binGA with adaptive mutation and bit-simulated crossover (Fig. 5(a)). The best total fitness 21.241 was found for $p_m = p_c = 0.3$. An increasing population size improves the optimization performance to a median of 22.916 (Fig. 5(b)).

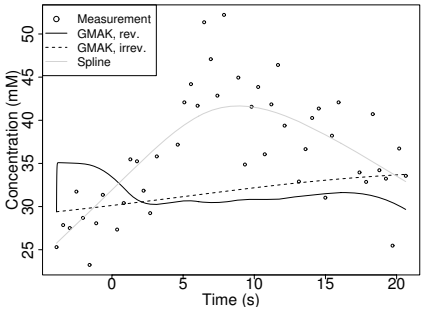
TABLE II

BEST FITNESS VALUES AND ALGORITHMS FOR EVERY MODEL SYSTEM

For each model the minimal fitness and the corresponding standard algorithm are listed. The algorithm that reached the best average fitness and the corresponding average fitness are written in the last two columns together with the standard deviation.

Model	Minimum	Algorithm	Mean	Std. dev.	Algorithm
rev.	23.097	binGA	26.106	2.205	binGA
irrev.	25.891	binGA	50.093	4.669	HC MS 25

The resulting model systems for the parameter values yielding the lowest fitness values are plotted in Fig. 6. For a better visualization we added splines to the figures to indicate a plausible fit (RSE of the splines: 19.670). The irreversible system is unable to follow the dynamic behavior of the measurements for KIV, DHIV, Leu and Val. Instead it results in straight lines, best fitness found: 25.981 (Tab. II). The rev. GMAK model (fitness 21.241) fits the data best.



(g) L-Valine (Val)

Fig. 6. The best fit of the different models

IV. CONCLUSIONS

We performed a systematical benchmark of six optimization procedures with their specific settings on the problem

of network inference with two model alternatives for the Val and Leu biosynthesis in *C. glutamicum*. We highlighted the advantages and drawbacks of every model system and every optimization algorithm. None of the models was able to reproduce the time series of Val and DHIV with high accuracy. Certainly, biological data always show a more or less wide range of measurement noise. Thus, it cannot be expected that any deterministic model will explain every data point exactly. Instead a model should provide a basis for a predictive description of the dynamic network behavior.

Spline curves were fitted individually to each time series, so they are uncoupled and do not underly any biological model system. However, they indicate how we would expect an ideal model to fit the data. Compared to splines, which show an RSE of 19.670, the best model, the reversible generalized mass-action model with inhibition (9), yielded a fitness of 21.241 only slightly higher than that of the splines.

The advantage of the derived inhibition formalism (9) for the rev. GMAK is that the model requires only 24 model parameters, a small number compared to other approaches that we will investigate in subsequent publications, but we neglect the fact that the reactions are catalyzed by enzymes.

We also recognized that the second model based on the assumption of irreversible reactions—except R_2 and R_9—does not lead to a good model fit. The best fitness of this model was found with an error of 25.981, a significantly worse fitness than the best we found. This model includes only 18 parameters, but is unable to reproduce the dynamics of the system. Instead it mostly results in straight lines (Fig. 6), which are a local optimum as well but biologically implausible. If we trust the information stored in KEGG, this behavior suggests other reactions not included in our model system which interact with the metabolites on the pathway.

We identified the optimization procedures that provided the best results for the studied optimzation problems:

1) The binary Genetic Algorithm with adaptive mutation $p_m = 0.2$ and bit-simulated crossover $p_c = 1$ and a population size of 1,000 individuals.
2) The Evolution Strategy with local mutation $p_m = 0.8$ and one-point crossover $p_c = 0.2$ and a population size of $(\mu = 5, \lambda = 25)$.

The choice of the right model system and optimization procedure with appropriate settings is a crucial step for model identification and parameter fitting, which is also confirmed by Spieth *et al.* [7], [12], who investigated the applicability of EAs on artificial biological networks. In their study, the realGA with UNDX crossover performed better than the binGA, which is not the case for this metabolic model based on another formalism and *in vivo* data. On artificial networks the cmaES achieved the best results followed by Differential Evolution, which was not considered in this study. Here, the optimized ES performed finally better than the optimized binGA according to the median—difference of 1.032, and found similar good single results with a difference of only 0.24. The binGA finds regions of good local optima with almost all settings and also the best total result after fine tuning. Due to the manifold settings the ES is harder manageable even though it performs as good with optimized adjustments. These results may also highlight the differences between *in silico* and *in vivo* data.

For our future work this study provides a valuable basis to further extend this model system and to apply optimization algorithms to similar problems with appropriate settings.

ACKNOWLEGEMENTS

A. D. and J. S. are funded by the NGFN-II EP project no. 0313323. We are grateful to Ralf Takors, Klaus Beyreuther and Oliver Kohlbacher.

REFERENCES

[1] R. Guthke, W. Schmidt-Heck, G. Pless, R. Gebhardt, M. Pfaff, J. C. Gerlach, and K. Zeilinger, "Dynamic Model of Amino Acid and Carbohydrate Metabolism in Primary Human Liver Cells," in *VII International Symposium on Biological and Medical Data Analysis*, 2006.

[2] E. Klipp, B. Nordlander, R. Kruger, P. Gennemark, and S. Hohmann, "Integrative model of the response of yeast to osmotic shock," *Nature Biotechnology*, vol. 23, no. 8, pp. 975–982, August 2005. [Online]. Available: http://dx.doi.org/10.1038/nbt1114

[3] C. Chassagnole, N. Noisommit-Rizzi, J. W. Schmid, K. Mauch, and M. Reuss, "Dynamic modeling of the central carbon metabolism of *Escherichia coli*," *Wiley Periodicals, Inc.*, pp. 54–73, January 2002 2002.

[4] J. B. Magnus, D. Hollwedel, M. Oldiges, and R. Takors, "Monitoring and Modeling of the Reaction Dynamics in the Valine/Leucine Synthesis Pathway in *Corynebacterium glutamicum*," *Biotechnology Progress*, vol. 22, no. 4, pp. 1071–1083, 2006. [Online]. Available: http://dx.doi.org/10.1021/bp060072f

[5] R. Guthke, K. Zeilinger, S. Sickinger, W. Schmidt-Heck, H. Buentemeyer, K. Iding, J. Lehmann, M. Pfaff, G. Pless, and J. C. Gerlach, "Dynamics of amino acid metabolism of primary human liver cells in 3D bioreactors," *Bioprocess Biosystem Engineering*, vol. 28, no. 5, pp. 331–340, April 2006.

[6] D. Visser and J. J. Heijnen, "Dynamic simulation and metabolic redesign of a branched pathway using linlog kinetics." *Metab Eng*, vol. 5, no. 3, pp. 164–176, Jul 2003.

[7] C. Spieth, R. Worzischek, F. Streichert, J. Supper, N. Speer, and A. Zell, "Comparing Evolutionary Algorithms on the Problem of Network Inference," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, 2006.

[8] R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. Rhee, C. Tissier, P. Zhang, and P. D. Karp, "MetaCyc: a multiorganism database of metabolic pathways and enzymes." *Nucleic Acids Research*, vol. 34, no. Database issue, pp. D511–D516, January 2006. [Online]. Available: http://dx.doi.org/10.1093/nar/gkj128

[9] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG," *Nucl. Acids Res.*, vol. 34, no. suppl_1, pp. D354–357, 2006. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/34/suppl_1/D354

[10] R. Heinrich and S. Schuster, *The Regulation of Cellular Systems*. 115 Fifth Avenue New York, NY 10003: Chapman and Hall, 1996.

[11] J. Barthelmes, C. Ebeling, A. Chang, I. Schomburg, and D. Schomburg, "BRENDA, AMENDA and FRENDA: the enzyme information system in 2007," *Nucl. Acids Res.*, vol. 35, no. suppl_1, pp. D511–514, 2007. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/35/suppl_1/D511

[12] C. Spieth, N. Hassis, F. Streichert, J. Supper, K. Beyreuther, and A. Zell, "Comparing Mathematical Models on the Problem of Network Inference," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, 2006.

[13] F. Streichert and H. Ulmer, "JavaEvA - A Java Framework for Evolutionary Algorithms," Center for Bioinformatics Tübingen, University of Tübingen, Technical Report WSI-2005-06, 2005. [Online]. Available: http://w210.ub.uni-tuebingen.de/dbt/volltexte/2005/1702/