Model selection and adaptation for biochemical pathways

R. Brause

J.W.G.-University, 60054 Frankfurt, Germany
RBrause@cs.uni-frankfurt.de

Abstract

In bioinformatics, biochemical signal pathways can be modeled by many differential equations. It is still an open problem how to fit the huge amount of parameters of the equations to the available data. Here, the approach of systematically obtaining the most appropriate model and learning its parameters is extremely interesting.

One of the most often used approaches for model selection is to choose the least complex model which "fits the needs". For noisy measurements, the model which has the smallest mean squared error of the observed data results in a model which fits too accurately to the data – it is overfitting. Such a model will perform good on the training data, but worse on unknown data.

This paper propose as model selection criterion the least complex description of the observed data by the model, the minimum description length. For the small, but important example of inflammation modeling the performance of the approach is evaluated.

Keywords

biochemical pathways, differential equations, septic shock, parameter estimation, overfitting, minimum description length.

1. Introduction

In living organisms many metabolisms and immune reactions depend on specific, location-dependent interactions. Since the interactions occur in a timed transport of matter and molecules, this can be termed as a network of biochemical pathways of molecules. In Bioinformatics, these pathways or signal interactions are modeled by many differential equations. For complicated systems, differential equations systems (DES) with up to 7,000 equations and 20,000 associated parameters exist and model reality. The motivation for life science industry to use such systems is evident: A prediction of reactions and influences by simulated models helps avoiding time-consuming, expensive animal and laboratory experiments, decrease the high costs for developing new drugs and therefore may save millions of Euros. For small signal transduction networks, this has already been done by estimating the parameters by data-driven modeling of expression profiles of DNA microarrays, see e.g. [2],[3],[4]. Interestingly, no problems were reported fitting the models to the data.

Although the basic idea is quite seducing, the practical problems associated with the simulation approach are difficult to solve: How do we know that our selected model is valid and how can all parameters be set to the correct values? And if all parameters are different for each individual, how can they be adapted to the real values based only on a small set of measured data per organism?

In this paper we will try to answer some of theses questions for the example of the small but important problem of inflammation and septic shock.

2. The differential equation neural network of inflammation and septic shock

The symptoms of septic shock contain low blood pressure, high ventilation and high heart rates and may occur after an infection or a trauma (damage of tissue). The septic shock research has no convincing results yet; there is still a high mortality of about 50% on the intensive care units (ICU) and nobody knows why. It is only possible to predict the outcome for a patient in advance just for 3 days, see [5] . In 1999, about 250,000 death were associated with sepsis in the USA.

The septic shock state is produced by a confusing myriad of immune pathways and molecules. For studying the basic problems we restrict ourselves first to a simplified but still

functional version of the model which uses only three variables and 12 constant parameters [6]. Let *P* be the pathogen influence, *M* the immunological response, e.g. the macrophages involved and *D* the obtained cell damage. Then, using some basic assumptions [7], we might combine them into a coupled system of three first order differential equations:

$$P'(t) = a_1(1-P)P + a_2MP \tag{1}$$

$$M'(t) = a_3M + a_4M(1-M)P + a_5M(1-M)D \tag{2}$$

$$D'(t) = a_6D + a_7\, h(\,(M-a_9)/a_8) \tag{3}$$

The plot of the time course for the three outputs (three variables) for the set of parameters shown in Tab. 1 is shown in Fig. 1. For this, the differential equations were numerically integrated using the Runge-Kutta method.

$a_1 = 0.054$	$a_3 = -1.0$	$a_6 = -0.01$
$a_2 = -0.2155$	$a_4 = 5.0$	$a_7 = 0.00384$
$a_9 = 0.2018$	$a_5 = 1.0$	$a_8 = 0.1644$

Tab. 1 The constant parameter values

It can be concluded that an infection (P) causes cell damage (D) and a delayed activity of the macrophages (M). The infection is defeated by the macrophages which decrease to a sufficient level afterwards. In this case (parameter regime), the infection remains chronically and the cell damage reaches a stable level.

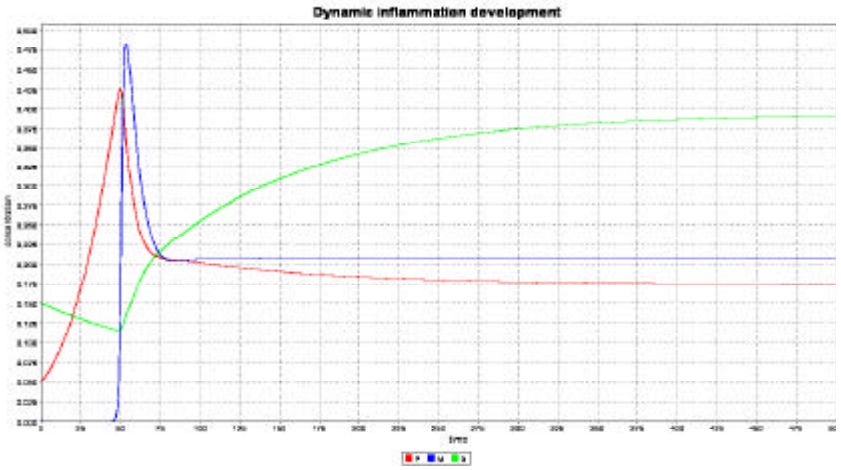

Fig. 1 The time dynamics of the equations (4),(5) and (6).

Now, how can the parameters, which correspond to the weights of the second layer, be learned? It is well known that the non-linear transfer function of deterministic chaotic systems can be efficiently be learned in order to predict a chaotic time series, see for instance [8]. Therefore, all dynamics which evolve by recurrent influence may be modeled by recurrent neural nets containing delayed signals, implemented e.g. in the discrete case by delay elements like tapped delay lines. In this case, the learning can be done by simple error reducing algorithms.

In the next section let us regard the adaptation of the parameters more closely.

3. Learning the Parameters

Generally, the biochemical pathways are very complex. It is not clear, which influences are important and which are not important. For the analytical description by equations this means that the number of terms ("model selection") and the values of its parameters ("model adaptation") are not given *a priori*, but have to be estimated ("learned") by the real observed data. How can this be done?

First, we are troubled by the fact that we do not have the full data set of Fig. 1 but only the small set of observed data given in table 2.
This situation is different from the previous one of learning the unknown parameters: the time scales of the observed training data and of the iteration cycles are different. For instance, the dynamics of inflammation might be in the reach of hours, whereas the observed data is taken once each day.

Time step	P	M	D
0	0.050000	0.001000	0.150000
100	0.201215	0.206079	0.254347
200	0.183751	0.206844	0.342027
300	0.177270	0.206750	0.374282
400	0.174876	0.206680	0.386141
500	0.173995	0.206649	0.390500

Tab. 2 The observed sparse data

In Fig. 2 this situation is shown. Here, the variable y(t) changes after each time tick, but it is only measured at time points t_i.

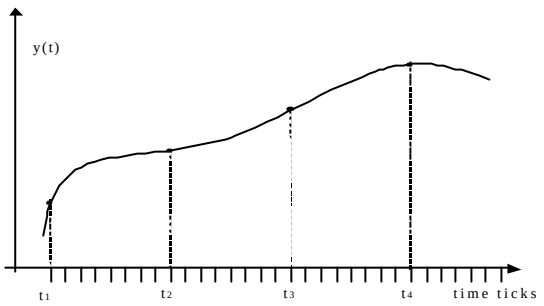

Fig. 2 The different time intervals for the differential equation and the observations

The different time scales will change heavily the approximated coefficients and difference equations, see [7] . Therefore, if we ignore the time steps between the observations and assume that system iterates once for one observation we will not be able to predict the best fitting parameters a_i for the difference equations that have several time steps between the observations.
Now, how can we proceed to approximate the unknown parameters from sparse observations? Obviously, the direct approach of a gradient descend for reducing the mean squared error between a simulated time sequence and an observed one used for instance in chaotic time sequence parameter estimation [8] is not possible here because we have no knowledge of the intermediate samples.
Instead, let us consider a variant of the classical evolutionary approach as it was introduced by Rechenberg 1973 [7], see [10].

(i) Generate a new set of random weights numerically by incrementing the old value with a random number, e.g. a Gaussian deviation.
(ii) test it: does it decrease the objective function?
(iii) If no, inverse the sign of the increment and test it: does it decrease the objective function?
(iv) If no, take the old weight values and choose another weight set.
(v) Continue until the error has sufficiently decreased or the number of iterations has exceeded a predefined maximum.

In order to avoid getting stuck in a suboptimum, the whole process might be repeated several times using different random starts.

The advantage of this approach is its independency of the complexity of the objective function. The disadvantage is its high computational burden: we have to recompute the objective function each time we change only one parameter, and we can not adapt the step width in advance. Nevertheless, for a given DES and given observed data this approach shows good performance, see [7].

For a given model, this is fine. If the model is not given we are in trouble: How should we select the model and adapt the parameters at the same time? The initial idea of first adapting the parameters and then selecting the model by pruning all terms that has very small parameter values might work. Consider for instance our model of eqs. (1),(2),(3). We might add the following ideas:

(1) If the pathogen influence (microbes) is present at the location where cell damage occurs, the pathogen influence will be increased: $P' \sim PD$

(2) Macrophages will die due to toxic influence of microbes, proportional to the co-occurrence probability and the microbe concentration : $M' \sim -P^2M$

These two possible extensions of the model will be translated into the modified differential equations

$$P'(t) = a_1(1-P)P + a_2MP + a_{10}PD \quad (4)$$

$$M'(t) = a_3M + a_4M(1-M)P + a_5M(1-M)D + a_{11}P^2M \quad (5)$$

$$D'(t) = a_6D + a_7\, h((M-a_9)/a_8) \quad (6)$$

On the other hand, we might have a more simple model in reality than we expect. For instance, we might have a model without influence of variable D to variable M, i.e. $a_5 = 0$, or a changed model with both $a_{10}, a_{11} \neq 0$ and $a_5 = 0$.

With these ideas, we have now four different possible models. How can we decide which model is implemented by reality? How can we choose the best model?

4. Model selection

The choice of the model is important for all diagnosis and therapies of the septic process. First, we have to discuss several possibilities for selecting the appropriate model and then we will select one strategy of our choice.

4.1 Model selection by parameter pruning

As the first, naïve approach let us consider the case where we have the pure differential equations (1),(2),(3) or (4),(5),(6) encountering no noise and we have recorded observation samples. How do we know which model is the right one for the observations? In this case, we might expect that the additional terms produce an error in modeling the observations. In the other way, reducing the error in the parameter adaptation process might result in setting the unnecessary parameters to zero, if they exist: the describing model is automatically tailored to the observed data.

Strategy 1: Adapt the parameters of the most complex model. All unnecessary parameters will automatically become zero and can be pruned.

Let us make an experiment to review this approach. For a time series produced by equations (1),(2),(3) we start the adaptation process, once for the small model of equations (1),(2),(3) and once for the augmented model of or equations (4),(5),(6). Now, what parameter values will we encounter for the new parameters a_{10} and a_{11}? We expect them to become zero. For good starting points and short approximation runs, this is true. Even for k = 1000 cycles the deviations are not huge: in Table 1 the mean squared error is shown for a certain number of cycles, once with the additional terms clamped to zero (i.e. without the terms) and once with the terms.

k	MSE	a_{10}	a_{11}
1,000	2.4633 **E-7** 7.6321 **E-7**	0.0 0,006796	0.0 0,645184
10,000	2.2348 **E-9** 3.9750 **E-7**	0.0 0,004453	0.0 0,549848

Table 1 The modeling error and the development of additional parameter terms

We observe that in the long run the approximation with additional terms does not improve while the correct model does. How can this be explained? The additional interactions that are caused by the additional parameters a_{10} and a_{11} in the augmented model will produce small disturbances that will deviate the approximation process: the approximation will slow down in relation to the non-augmented model which fits well to the observed data. This is shown in Fig. 3.

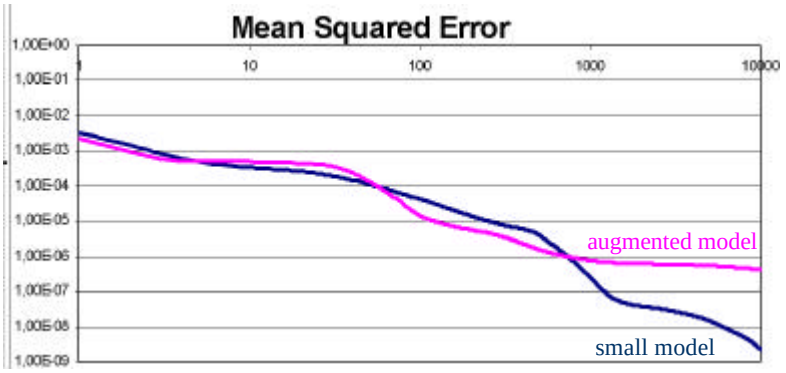

Fig. 3 The error development with and without additional terms

This might inspire us to the second strategy:

Strategy 2: Adapt the parameters of all models. Select the model, which converges best.

Do we have discovered a good selection criterion for the model? The answer is no: the additional interactions slow the convergence down, but the inverse is also true for too simple models which can not approximate the observed samples well, but initially converge faster than the true model.

Additionally, in nearly almost all natural systems we encounter noise that is not considered in this approach. So, we have to relay on other approaches.

4.2 Model selection by minimum description length

In the previous section we have seen that the convergence of the parameters cannot automatically replace the model selection process. Instead, we have to evaluate the performance of each model, i.e. each form of DES separately related to the observed time course samples. What kind of performance measure should we choose? We know that the deviation of the samples to the predicted values, the mean squared error, is not a good approach: by the additional parameters the more complex models will tend to overfit on adapting to the observed values perfectly whereas the best model will produce sample differences within the variance of the samples. This leads to our

Strategy 3: Adapt the parameters of all models to fit the observed data. Select the model which gives the shortest description of the observed data, on average and asymptotically.

So, we are looking for a model which neither fits too good nor too bad and needs only a small amount of information to describe the observations. How do we evaluate this?

Let us formalize our problem: For each of the *k* subjects and each variable, we observe values at different time steps t_1, t_2, …,t_n. For example as shown in Fig. 4, we might measure the dynamics at four times. All the four samples of one subject might be grouped together in one set. The set of observations for one subject is called a sample $\mathbf{x} = (x_1,\ldots,x_n)$ of all possible observations $\{\mathbf{x}\}$. Each model m which has been fit to the sample also produces by simulation a sample $\mathbf{f} = (f_1,\ldots,f_n)$ for the designated *n* time steps.

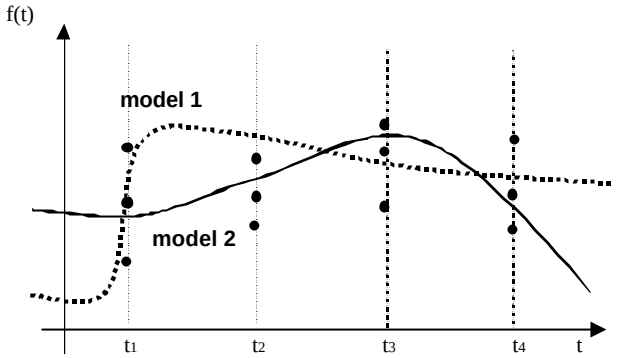

Fig. 4 *Selecting the best fitting model*

The deviation of the i-th observed sample $\mathbf{x}(i)$ from its adapted model $\mathbf{f}(i)$ of the same type is for all time steps $t = 1\ldots N$ its empirical variance [11]

$$\sigma_i^2 = \frac{1}{N-1}\sum_{t=1}^{N}(x_t(i) - f_t(i))^2 = \frac{1}{N-1}(\mathbf{x}(i) - \mathbf{f}(i))^2 \qquad (7)$$

Observing S subjects, the variance for all subjects from their approximated models is

$$\sigma^2 = \sum_{i=1}^{S}\sigma_i^2 \qquad (8)$$

Since each sample of each subject contain measurements at the same time period, we might draw these samples together in one chart as it is done in Fig. 4.

Assuming that the deviations at each time step are differently distributed, i.e. time dependent, we might compute the variance for one time step *t* for the set of all observed subjects to all models *i* by

$$\sigma_t^2 = \frac{1}{S-1}\sum_{i=1}^{S}(x_t(i) - f_t(i))^2 \text{ and } \sigma^2 = \sum_{t=1}^{N}\sigma_t^2 \qquad (9)$$

Now, we might analyze the system by two different approaches:

a) Either we have only one model for all subjects. Then, all samples are random deviations of the true model sample **f**. We select as best model the one which best "fits the data ". Biologically, the approach of only one fixed model is improbable.

b) Or we assume a different model $\mathbf{f}(i)$ for each subject *i*. This means that either the parameters of m are different or even the basic model type m might be different. Then, each observed sample $\mathbf{x}(i)$ deviate slightly from the best fitting model sample $\mathbf{f}(i)$. As best model m* we might select the one which, after the individual parameter adaptation, "fits best" for all subjects.

Now, in order to evaluate the fitting of the model we have to compute the description length L of the data, given the model. This is the average information of the data. It can be shown that the description length L, is bounded below by the entropy of the probability distribution of the data set X

$$L \geq H(X)$$

Thus, the minimal description length of the data is obtained for the model which provides the smallest entropy of the observed data. For normally distributed data according to [13],

we know that for the compound random variable X we have

$$H(X) = \ln \sqrt{(2\pi e)^n \det C_{XX}}$$

For uncorrelated samples **x** we have

$$\det C_{XX} = \prod_t \sigma_t^2$$

and therefore

$$H(X,m) = \tfrac{1}{2} \ln (2\pi e)^n + \tfrac{1}{2} \ln \prod_t \sigma_t^2(m) = A + \tfrac{1}{2} \sum_{t=1}^{n} \ln \sigma_t^2(m) \qquad (10)$$

Therefore, the best model m^* is the one which minimizes H(X,m), i.e. which has the smallest variances for all time steps.

The information can be measured for two cases: the situation for training and the situation for testing. For training, we have the mean squared error between the observations of the training sample and those of the model sample, averaged over all variables *v* and all models *k* of the same type

$$MSE_{train} = \frac{1}{S \cdot m \cdot N} \sum_{k=1}^{S} \sum_{v=1}^{m} \sum_{t=1}^{N} \left(x_t(k,v) - f_t(k,v)\right)^2 \qquad (11)$$

For the test case, we compare the model sample with all other possible observations, i.e. the rest of the training set, also averaged over all variables and all models of the same type

$$MSE_{test} = \frac{1}{S \cdot m \cdot N \cdot (S-1)} \sum_{k=1}^{S} \sum_{v=1}^{m} \sum_{t=1}^{N} \sum_{j \neq k} \left(x_t(j) - f_t(k)\right)^2 \qquad (12)$$

For the two cases, the information of eq.(10) is averaged over all models *k* of same type m and variables *v* and becomes

$$H(m) = \langle H(k,v) \rangle_{k,v} = \frac{1}{S} \sum_{k=1}^{S} \frac{1}{m} \sum_{v=1}^{m} \left(\tfrac{1}{2} N \cdot \ln (2\pi e) + \tfrac{1}{2} \sum_{t=1}^{n} \ln \sigma_t^2(k,v) \right)$$

with $\sigma_t^2(k,v) = \left(x_t(k,v) - f_t(k,v) \right)^2$ *for training* (13)

or $\sigma_t^2(k,v) = \frac{1}{S-2} \sum_{j \neq k}^{S} \left(x_t(j,v) - f_t(k,v) \right)^2$ *for testing* (14)

Therefore, we get for the averaged information *H* of a model of type m

$$H(m) = \frac{1}{S \cdot m \cdot N} \sum_{k=1}^{S} \sum_{v=1}^{m} \sum_{t=1}^{N} \ln(\sigma_t^2(k,v)) + \frac{N}{2} \ln(2\pi e) \qquad (15)$$

5. Evaluating the data simulations

For the simulation, we generated four data files for two different model types and $M = 10$ subjects. The *M* individually generated time courses start with the same initial values, but differentiate in the following aspects:

a) All subjects have the same standard model parameters and all observations are the same.
b) All subjects have the same standard model parameters, but the observations have random deviations (N(0,0.02) distribution) of the true values.
c) All subjects have individual, different model parameters (N(0,0.001) distribution);

the observations are the true values.

d) All subjects have individual, different model parameters; but the observations have random deviations of the true values.

These four model assumptions are used to generate four observation files. Each set of observations contain n = 5 sampled values of all three variables for each of the M = 10 subjects.

The four data files of observations are analyzed by four different model types:

m_1) The smaller model with $a_5 = 0$.

m_2) The "standard" model with $a_5 \neq 0$ and $a_{10}, a_{11} = 0$.

m_3) The augmented model with $a_5, a_{10}, a_{11} \neq 0$.

m_4) The changed (dropped and added terms) model with $a_{10}, a_{11} \neq 0$ and $a_5 = 0$.

Each of the four observation files is used to adapt the parameters of each of the M = 10 subjects of the same model type to the observations by the evolutionary method described in section 3. So, we get 16 result sets for 10 subjects each.

For each adaptation try we use 100 cycles of adapting all parameters in order to minimize the mean squared error R between the model prediction and the observations. For each subject, 10 tries are performed and the one with the smallest R is recorded in order to avoid getting stuck in a suboptimum. After adaptation, the performance of the models was evaluated by computing the minimum description length, i.e. entropy H for the model adaptations. The evaluated values for the entropy H for the four model types m_1, m_2, m_3, m_4 adapting to the data of the four observed situations a),b),c) and d) are presented in table 1. The results of the standard model is shown in bold face.

Table 1 *The evaluated observations for N = 6 samples*

Simulation		Model	MSE train	MSE test	H train	H test
a)	1	1	1.4686 E-4	1.4686 E-4	3.3993	3.2162
	2	**2**	**4.6640 E-4**	**4.6640 E-4**	**3.5141**	**3.3310**
	3	3	2.4891 E-3	2.4891 E-3	4.5359	4.3528
	4	4	4.8081 E-2	4.8081 E-2	5.4169	5.2338
b)	5	1	2.0826 E-3	2.2171 E-3	4.3146	4.2963
	6	**2**	**8.2153 E-4**	**1.0491 E-3**	**4.2346**	**4.2163**
	7	3	4.2131 E-2	4.2006 E-2	4.7925	4.7742
	8	4	1.0672 E-2	1.0916 E-2	4.9448	4.9264
c)	9	1	2.7123 E-4	5.3854 E-3	3.5523	3.3692
	10	**2**	**2.5080 E-3**	**7.5664 E-3**	**3.7953**	**3.6122**
	11	3	3.5737 E-3	8.4245 E-3	4.2872	4.1041
	12	4	3.0380 E-2	3.3430 E-2	5.1236	4.9405
d)	13	1	1.3788 E-2	2.3057 E-2	5.4111	5.3744
	14	**2**	**1.3771 E-2**	**2.3096 E-2**	**5.4219**	**5.3853**
	15	3	4.9014 E-2	5.9496 E-2	5.3382	5.3016
	16	4	6.9419 E-2	7.7748 E-2	5.5937	5.5571

What can we conclude by these results?

Keeping in mind that m_2 is the standard model type that was used to produce all data, we see that this model type has the smallest entropy of all other models in the context of cases a) and b) – it turns out as the best model to select. Therefore, our model selection criterion is valid in our example.

For cases c) and d) of different parameter regimes and random deviations the smaller model fits slightly better to the data. Why? The reason behind is that the random deviations and the systematic deviations are in the same range; for only a small number of observations for one individual (N = 6) the difference becomes hard to detect: the proposed method reaches its limits.

Here we encounter a fundamental problem of data modeling: how do we know that for a given observation variance the number of observed data points are sufficient to select a model properly? What difference of complexity should be taken as argument for a model to be more valid than another one? Theses questions are still open for research.

6. Discussion

Data driven modeling is an important attempt to rationalize the efforts of creating models guided not by assumptions but by reality. The paper shows some of the problems involved in this kind of modeling and proposes the minimum description length of the observed data as selection criterion.

For the small but important problem of inflammation and septic shock differential equations we consider four different models types: a standard model, the model with one term dropped, the model with two additional terms and a changed model. These four models are confronted with synthetic data, generated by random versions of the standard model. Here, all four possible model types converge more or less fast to fit the data; no terms can be pruned due to small parameter values; an automatic tailoring of the model to the data is not possible.

Thus, the model selection can neither be based on the convergence speed nor on the "complexity" of the formulas (is a multiplication more complex than an addition?) but have to be based on another criterion. In this paper we chose the minimum description length MDL of the data using the model as performance criterion. Assuming normally distributed deviations we computed the entropy as lower limit of the MDL by using the observed variance between the adapted models and the observed data. The simulation results validated our approach: The analyzing model describes the data with the lowest MDL if data generation model type and analyzing model type coincide.

Nevertheless, for a small amount of observed data, many different parameter regimes of the same model type and many random deviations the difference between the models cannot be detected by the MDL criterion any more. Here, more problem-specific information is needed.

7. References

[1] R. Brause, E.Hanisch, J.Paetz, B. Arlt: The MEDAN project: results and practical meanings, 3. Int. Symposium "*Sepsis, SIRS, Immune Response - Concepts, Diagnostics and Therapy*", A. Nierhaus, J. Schulte am Esch (Eds.), PABST Science Publishers, Lengerich, Germany, (2003), pp.122-129

[2] D'haeseleer, P.; Liang, S.; Somogyi, R. (2000) *Genetic network inference: from co-expression clustering to reverse engineering*. Bioinformatics 16, 707-26.

[3] Steffen, M., Petti, A., D'haeseleer, P., Aach, J., and Church, G.M. (2002) *Automated Modeling of Signal Transduction Networks*. BMC Bioinformatics 3:34-44.

[4] Yeung, M.K.; Tegner J.; Collins, J.J. (2002) *Reverse engineering gene networks using singular value decomposition and robust regression*. Proc Natl Acad Sci USA. 99, 6163-8.

[5] J. Paetz, R. Brause: "A Frequent Patterns Tree Approach for Rule Generation with Categorical Septic Shock Patient Data"; in: J. Crespo, V. Maojo, F. Martin, *Medical Data Analysis*, Springer Verlag Berlin Heidelberg (2001), pp.207-212

[6] C. Chow: Development of a Dynamical Systems Model of Acute Inflammation, E. Neugebauer: 2nd workshop on complex systems: *Analysis in Shock and Trauma Research*, University of Cologne, (2003)

[7] R. Brause: *Adaptive modeling of biochemical pathways*. IEEE 15th Int. Conf on Tools with Art. Intell. ICTAI-2003, IEEE Press 2003, pp.62-68, (2003)

[8] A. Lapedes, R. Farber: "How Neural Nets Work"; Report LA-UR-88-418, Los Alamos Nat. Lab. 1988; and in Y.C. Lee (Ed.): *Evolution, Learning and Cognition*; World Scientific, Singapore, New Jersey, London (1988)

[9] I. Rechenberg: *Evolutionsstrategie*; problemata frommann-holzboog, (1973)

[10] H.-P. Schwefel: *Evolution and Optimum Seeking* , J.Wiley, (1995)

[11] I. N. Bronstein et al., *Handbook of Mathematics*, Van Nostrand Reinhold; 3rd edition (1991)

[12] M. Hansen, B. Yu: *Model selection and the principle of Minimum Description Length*, Journal of the American Statistical Association, Vol.96 (454),pp. 746-774, 2001

[13] T.Cover, J.Thomas: *Elements of Information Theory*, John Wiley, New York 1991