COMBINED OPTIMIZATION TECHNIQUE
FOR BIOLOGICAL DATA FITTING

Gursky V.V., Kozlov K.N.*1, Samsonov A.M.2

1 Dept. of computational biology, State Polytechnical University, St.Petersburg, 195251 Russia;
2 A.F. Ioffe Physico-technical Institute of the Russian Academy of Sciences, St.Petersburg, 194021, Russia
* Corresponding author: e-mail: kozlov@spbcas.ru

Keywords: optimization, regulatory gene networks, optimal control, mathematical model, computer analysis

Summary

Motivation: Modern molecular biology has massive amounts of quantitative data already at its disposal. The crucially important problem for getting closer insights into mechanisms of development is to reduce the complexity of finding the parameters of mathematical models by fitting to experimental data.

Results: The new Combined Optimization Technique (COT) showed a high accuracy in reconstruction of phenomenological parameters of equations and saved about 30% of the most time consuming operations in computation that allow to propose the COT as quite attractive instrument for processing big amounts of experimental data of various nature.

Availability: available on request from the authors

Introduction

Modern molecular biology has massive amounts of quantitative data already at its disposal, and robust and reliable algorithms’ development to treat them becomes a foreground job. Mathematical modeling is essential for systematic treatment of experimental results and for getting insights into the structure of underlying natural objects.

We perform the gene expression data fitting in the context of one biological system namely the segment determination gene network of a fruit fly Drosophila embryo. The experiments were performed to acquire data on segmentation gene expression at cellular resolution, see (Reinitz, Sharp, 1995). The dynamical model of gene expression is described by a system of highly nonlinear reaction-diffusion (NRD) equations. We present new results of experimental data fitting for finding phenomenological parameters capable of pattern formation and propose a new Combined Optimization Technique (COT) for processing large amounts of experimental data. The developed algorithm combines advantages of random search method and steepest descent approach. Main idea is as follows: firstly a rough approximation of parameters is to be found by the random search, afterwards it is subjected to refinement by the Optimal Steepest Descent Algorithm (OSDA) developed recently (Kozlov, Samsonov, 2003) and applicable to problems of various physical nature.

Main goal is to evaluate the efficiency and convergence speed of the method, and for this reason we start with an estimation of the parameters of small (2 genes) network for simplicity and expanded the number of genes involved after successful numerical experiments. The parameters we have found in 2-gene network allow to keep permanent patterns of gene expression when time tends to infinity. In larger networks the optimization results lead to different asymptotic behavior of solution (Gursky et al., 2004).
Methods and Algorithms

The dynamics of the model is described by the system of coupled differential-difference NRD equations formulated in (Reinitz, Sharp, 1995).

\[
\frac{\partial v_n^a}{\partial t} = R_a g \left( \sum_{b=0}^{G-1} T^{ab} v_n^b + m^a v_n^d + h^a \right) + D^a \left( v_{n-1}^a - 2 v_n^a + v_{n+1}^a \right) - A^a v_n^a . \tag{1}
\]

The equation is written for each gene product \( a \) and each nucleus \( n \), and \( G \) is the number of genes. A matrix element \( T^{ab} \), one for each pair of proteins, and coefficients \( m^a, h^a, R^a, D^a, \lambda^a \) for each protein are unknown parameters which should be determined by means of minimization of a functional equal to the sum of squared differences between the concentrations of the gene products (say, proteins), observed experimentally and calculated independently, e.g., by means of gene network approach.

Constraints in the form of inequalities are used to be imposed to the parameters \( R^a, D^a, \lambda^a \) for each protein \( a \), that does not allow to include them directly into an extended Lagrangian. Therefore to apply the Lagrange technique for optimization the constraints are to be transformed into equations, e.g., for \( R^a \) as follows:

\[
R_{\text{low}} \leq R^a \leq R_{\text{up}} \quad \Rightarrow \quad R^a = \alpha_r + \beta_r \tanh(\gamma^a r^a) .
\]

Constants \( \alpha \) and \( \beta \) are defined for parameter \( R^a \) by the following formulae

\[
\alpha_r = (R_{\text{up}} + R_{\text{low}}) / 2 \quad \text{and} \quad \beta_r = (R_{\text{up}} - R_{\text{low}}) / 2 .
\]

Constraints for other parameters (\( D^a \) and \( \lambda^a \)) are transformed similarly, however the transformation is not unique, and other representations involving bounded functions can be used.

Combined Optimization Technique for data fitting consists in application of the Simulated Annealing (SA) method using a weak quality criterion to obtain the rough approximation of the parameter set, which is refined afterwards by the OSDA, see (Kozlov, Samsonov, 2003).

The transformation coefficient values \( \gamma_r, \gamma_d \) and \( \gamma_\lambda \) are renewed after each of \( M \) steps using an empirical rule. If the functional value has been changed greater than the value of the corresponding parameter during the last \( M \) steps, then the corresponding value of \( \gamma \) should be increased in order to make the transformation function (\( \tanh \)) steeper and vice versa.

To make a close comparison of SA and COT we used the quality criterion for COT similar to that which was proposed in (Reinitz, Sharp, 1995) for SA. Namely, when the functional value decreased less than a predefined value \( \theta \) during the last \( S \) steps the set of parameters obtained at the very last step \( N \) is identical to the solution of the problem, and the optimal parameters are the components of this vector.

Implementation and Results

The numerical results are given in the Table below. Because of stochastic nature of the SA method the results provided are averaged over all performed experiments. Parameters were recovered with 6.1 % accuracy. The average number of functional evaluations used by COT equals 1,949,198. To obtain similar results with smaller accuracy of 6.9 % using the SA only it required 2,708,485 functional evaluations, that is ~28 % more.

61
To study the COT convergence in lab conditions we produced artificial gene expression data for the network of two genes in eight nuclei by integrating the model equations, using the set of parameters that represents already known solution. We took the model output for 9 time moments to calculate the functional value.

We performed optimization for 100 random initial approximations of parameter set \( q = \{ q_i \} \) and introduced the following criterion \( \kappa \) to measure the precision of numerical simulation \( q^{opt} \):

\[
\kappa = \max_i \left| \frac{q_i^{true} - q_i^{opt}}{q_i^{true}} \right| \times 100\% ,
\]

where \( q^{true} \) is the known solution.

The weak quality criterion used by the SA part of COT was: \( M = 100, S = 5, \theta = 10^3 \) and the final one for COT was: \( M = 14, S = 500, \theta = 10^9 \). To estimate the efficiency of COT we performed numerical simulations with SA using the following quality criterion: \( M = 100, S = 5, \theta = 10^5 \).

The average precision of the parameter set obtained by COT is less than 10\% if only those rough approximations produced by SA were taken into account, for which \( \kappa < 30 \% \). This coincides with the fact that the gradient method converges to the local minimum in general.

**Discussion**

The widely used Simulated Annealing method converges to the global minimum at the cost of very intensive computations. The number of functional evaluations determines the time necessary to obtain the solution of the data fitting problem because of a huge number of species and, therefore, the differential equations that are to be integrated. In real gene networks this number can exceed three hundreds, and to include more genes in the network under consideration is crucially important for getting closer insights into mechanisms of development.

![Fig.](image)

The time evolution of patterns from gastrulation time to infinity is shown for the first and the second proteins in the network on panels A and B respectively. The concentration of both proteins in all nuclei does not change in time.
The proposed new Combined Optimization Technique showed a high level of accuracy in reconstruction of phenomenological parameters of equation and saved about 30% of the most time consuming operations in computation, which may be equal to several days for a large scale problem simulation on a high performance computer. These features allow to propose COT as quite attractive instrument for processing big amounts of experimental data of various nature.

The parameters we have found in 2-gene network allow to keep permanent patterns of gene expression when time tends to infinity. The time evolution of patterns from gastrulation time to infinity is shown in Figure for the first and the second proteins in the network on panels $A$ and $B$ respectively. In larger networks the optimization results lead to different asymptotic behavior of solution (Gursky et al., 2004).

Acknowledgements

Our sincere thanks to Dr J. Reinitz and J. Jaeger for many valuable discussions. The support of the study by the NIH Grants RR07801, TW01147, the CRDF GAP Awards RBO685, RBO1286 is gratefully acknowledged.

References