IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10400-2

Colloquia: CSFI 2008

A distributed approach for parameter estimation in Systems Biology models

E. Mosca, I. Merelli, R. Alfieri and L. Milanesi
Istituto Tecnologie Biomediche, CNR - Via Fratelli Cervi 93, 20090, Segrate (MI), Italy

(ricevuto il 22 Giugno 2009; pubblicato online l'11 Settembre 2009)

Summary. — Due to the lack of experimental measurements, biological variability and experimental errors, the value of many parameters of the systems biology mathematical models is yet unknown or uncertain. A possible computational solution is the parameter estimation, that is the identification of the parameter values that determine the best model fitting respect to experimental data. We have developed an environment to distribute each run of the parameter estimation algorithm on a different computational resource. The key feature of the implementation is a relational database that allows the user to swap the candidate solutions among the working nodes during the computations. The comparison of the distributed implementation with the parallel one showed that the presented approach enables a faster and better parameter estimation of systems biology models.

PACS `82.39.Rt` – Reactions in complex biological systems.
PACS `82.20.Wt` – Computational modeling; simulation.
PACS `87.10.Ed` – Ordinary differential equations (ODE), partial differential equations (PDE), integrodifferential models.

1. – Introduction

Nowadays a systems biology approach is essential to understand complex biological processes, such as cell cycle regulation [1]. This approach relies on mathematical models used to describe biological systems and to make useful predictions. Due to the lack of experimental measurements, experimental errors and biological variability, the value of many parameters of the models is yet unknown or uncertain [2]. A possible computational solution is the parameter estimation, which can be informally stated as the identification of the parameter values such that the model dynamics fits the experimental data. Considering models based on a non-linear Differential Algebraic Equations (DAEs) system, the parameter estimation problem is formulated as a nonlinear programming (NLP) problem with DAEs constraints, in which the objective function is very often nonconvex. Due to nonconvexity the NLP-DAE problem solution must be searched with a global optimization (GO) method, since it is very likely that a local method would identify a solution of local nature. Stochastic GO approaches can provide good solutions in modest computation time, are quite simple to implement and do not require a transformation of the original problem [3]. However, large instances of the NLP-DAE problem cannot be solved in a reasonable time, even considering stochastic approaches. We have already developed an automated system to face the problem of parameter estimation of systems biology DAE models on high performance computing platforms [4, 5]. Here we present a distributed approach to compute an Evolution Strategy algorithm for parameter estimation of DAE models.

2. – Methods

Let us consider the DAE system $\mathbf{f} = (\mathrm{d}\mathbf{x}/\mathrm{d}t, \mathbf{x}, \mathbf{y}, \mathbf{p}, \mathbf{v}, t)$, where $\mathbf{x}$ is the vector of the differential variables, $\mathbf{y}$ is the n-length vector of the output state variables, $\mathbf{p}$ is the m-length vector of parameters to be estimated and $\mathbf{v}$ is the vector of the known parameters. The parameter estimation is stated as the minimization of the cost function J: $\arg\min_{\mathbf{p}\in S} J = D(A, Y(\mathbf{p})) = \sum_{i=1}^{n}\sum_{j=1}^{T}(A_{i,j}(t) - Y_{i,j}(\mathbf{p}, t))^2$, where $S \in \Re^m$ is the search space, $A_{n\times T}$ is the matrix containing a number T of experimental data for each state variable, $Y_{n\times T}(\mathbf{p})$ is the matrix containing the DAE system solutions at the same time points collected in A and $D(X, Y)$ is the operator which measures the distance between A and $Y(\mathbf{p})$ matrices. The constraints are the DAE system, the initial conditions $\mathbf{x}(t = 0) = \mathbf{x}_0$ and boundaries on parameter values $\mathbf{p}^L < \mathbf{p} < \mathbf{p}^H$.

In order to solve the GO problem we use an Evolution Strategy (ES), a sub-class of nature-inspired optimization methods belonging to the class of Evolutionary Algorithms (EAs) [6]. In order to evolve better and better solutions, EAs apply mutation, recombination, and selection operators to a population of individuals that represent candidate solutions. Within this formalism individuals are evaluated thanks to a fitness function that indicates the goodness of the solutions (in the presented work the fitness is the cost function J). In particular, we chose the Stochastic Ranking Evolution Strategy (SRES) [7] algorithm, which has shown good performance when applied to parameter estimation in biochemical pathways [3].

Although this algorithm is efficient, the time to compute a sufficiently high number of iterations in order to find a satisfying solution increases with the NLP-DAE problem size and complexity. An interesting approach to improve the search of candidate solutions (and hence the expected execution time) relies on running different instances of the algorithm (*i.e.* evolutions) simultaneously, swapping periodically the best results among the processes. Using this method, the convergence to the optimal solution speed up thanks to a wider search on the solutions space. To accomplish this task, we have designed an environment to distribute each run of the evolution algorithm on a different computational resource. This is achieved using a relational database that enables asynchronous communication among processes.

The presented approach was compared to a parallel implementation (in which the population of individuals is splitted in a number of subsets and the fitness of the individuals belonging to the same subset is calculated on the same processor). Experiments with the parallel implementation have been performed on a shared linux cluster of 280 Opteron AMD cores 275 at 2,2 GHz (provided by the Eurotech Group, Amaro Ud, Italy), composed by 60 blades diskless, each one equipped with an Infiniband 4X network card and 8 GB of RAM. Experiments with the distributed implementation have been performed on HPxw6600 equipped with Intel 2.2 GHz processors with 10 GB of RAM.

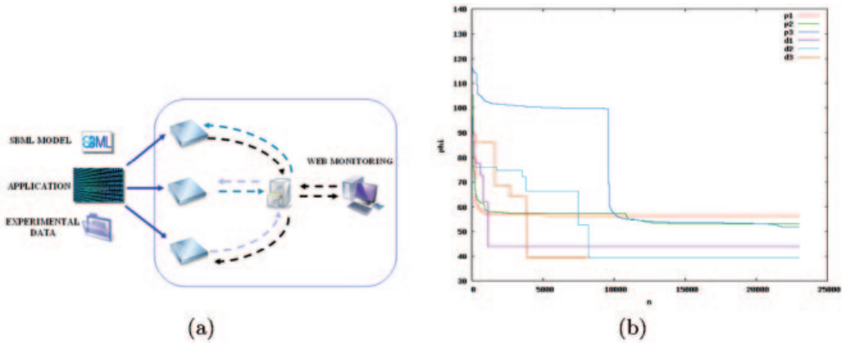

Fig. 1. – a) Schema of the functioning of the distributed approach. b) Best fitness value (phi) during the parameter estimation computed using the parallel (20 processors) and the distributed (8 processes) implementation. The ES setting is the same in all the simulations. Legend: p[1-3], parallel computations, d[1-3], distributed computations.

3. – Implementation

Relying on this conceptual framework, we have developed an environment to coordinate different runs of the evolution algorithm on many computational resources of a distributed platform, fig. 1a. The key feature of the implementation is a relational database that allows to swap the individuals among populations of the different runs of the algorithm during the computations.

In detail, each process is performed on a computational resource which can be completely independent of the others. The only requirement is the availability of a communication network to establish a connection to the central database. Each process starts by contacting the main database to inform the system of its presence, then it random initializes the population and run the optimization. After the accomplishment of each iteration each process stores its solutions along with the corresponding fitness. If the other processes are ready to swap their results, the algorithm downloads from the database the best solutions (sorted by the fitness value) from another randomly selected process and mixes them with its results. This approach improves the variability of the population which is essential to better explore the search space of the NLP-DAE problem.

The swap coordination among the processes is delegated to a script which runs in close association with the database. This script queries the database to check if all processes are ready to exchange their individuals, which means that they have performed a minimum number of iterations from the beginning or after a previous swap. If this is the case, the script flags a specific field in the database that enables the exchange as each process complete the current step.

4. – Results

The presented approach was tested on a published mathematical model of the cell cycle process. The model [8] represents a set of biochemical processes that drive the cell cycle regulatory system dynamics among eukaryotic cells. The model is formalised as a DAE system and is constituted by 14 nonlinear ODEs that contain 90 parameters and 2 algebraic equations. 15 variables represent protein concentrations while 1 is used for cell mass. Moreover, the model contains a rule that is used to implement the cell division: the variable representing the cell mass is reduced by a half whenever the value of another variable, representing the Cyclin B protein concetration, decreases under a predefined threshold.

The parameter estimation was carried on considering "pseudoexperimental" data generated simulating the model: in particular 20 time points for each state variable were considered. Among the 90 parameters, all the kinetic constants (87) were included in the parameter estimation. We repeated the parameter estimation on the parallel and distributed platforms three times. The presented approach outperformed the parallel one even if it was run considering a less number of processes, fig. 1b.

5. – Conclusions

We have developed an automated system to manage NLP-DAE problem that is based on a distributed computing approach. The system is made of two components: the application that handles the NLP-DAE problem and the infrastructure that manages the distribution over the computational resources relying on a relation database. The system accepts as input mathematical models formulated as DAE systems and supports the SBML standard [9]. The results indicate that this approach can successfully lead to the parameter estimation of more complex DAE models, by means of a faster identification of a better candidate NLP-DAE solution.

* * *

This work has been supported by the NET2DRUG, EGEE-III, BBMRI, EDGE European projects, by the MIUR FIRB LITBIO (RBLA0332RH), ITALBIONET (RBPR05ZK2Z), BIOPOPGEN (RBIN064YAT) and CNR-BIOINFORMATICS initiatives.

REFERENCES

[1] Alfieri R., Merelli I., Mosca E. and Milanesi L., *Nucleic Acids Res.*, **36** (2008) 641.

[2] Liebermeister W. and Klipp E., *Systems Biology, IEE Proc.*, **152** (2005) 97.

[3] Moles C. G., Mendes P. and Banga J. R., *Genome Res.*, **13** (2003) 2467.

[4] Mosca E., Merelli I., Alfieri R. and Milanesi L., *Sysbiohealth Symposium 2007* (Locomia Innovazione, Milano) 2007.

[5] Mosca E., Merelli I., Alfieri R. and Milanesi L., *The Sixth International Conference on Bioinformatics of Genome Regulation and Structure, BGRS 2008, June 22-28, Novosibirsk.*

[6] Fogel D. B., *Evolutionary computation: Toward a new philosophy of machine intelligence* (IEEE Press, New York) 2006.

[7] Runarsson T. P. and Yao X., *IEEE Trans. Evolut. Comput.*, **4** (2000) 274.

[8] Csikasz-Nagy A., Battogtokh D., Chen K., Novak B. and Tyson J. J., *Biophys. J.*, **90** (2006) 4361.

[9] Hucka M. *et al.*, *Bioinformatics*, **19** (2003) 524.