# Particularities of Evolutionary Parameter Estimation in Multi-stage Compartmental Models of Thymocyte Dynamics

Daniela Zaharie
Dept. Computer Science,
West University of Timişoara
300223 Timişoara, Romania
dzaharie@info.uvt.ro

Lavinia Moatar-Moleriu
Dept. Computer Science,
West University of Timişoara
300223 Timişoara, Romania
lmoatar@info.uvt.ro

Viorel Negru
Dept. Computer Science,
West University of Timişoara
300223 Timişoara, Romania
vnegru@info.uvt.ro

## ABSTRACT

The aim of this paper is twofold. Firstly, it presents an extension of a multi-stage compartmental model in order to make it more appropriate in modelling various perturbations of thymocyte dynamics.Secondly, it proposes an evolutionary approach, based on the JADE algorithm, for simultaneously estimating the number of division stages, the rates associated to cellular processes (e.g. proliferation, death, migration) and the parameters corresponding to the proposed perturbation functions. Several quality of fit measures are investigated and their relationship with the variability of experimental data is exploited in order to select the optimization criterion.

## Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence—*Problem Solving, Control Methods, and Search*; J.3 [**Computer Applications**]: Life and Medical Sciences

## General Terms

Algorithms

## Keywords

Compartmental models; thymocyte dynamics; parameter estimation; differential evolution

## 1. INTRODUCTION

Simulations of computational models inferred from experimental data can provide valuable information on various biological systems. In immunology, constructing models and simulating the processes involved in the intrathymic cell development are essential in understanding mechanisms of immune reactions. The intrathymic development of thymocytes (T cells) begins with an influx of bone marrow precursor cells and continues with several complex processes involving cell proliferation, differentiation and death [5]. During these processes the cells express some membrane markers (e.g. CD4 and CD8) which are specific to mature T cells. Depending on the absence/presence of these markers there are four main populations of cells in the thymus: DN (double-negative cells - they lack both markers), DP (double-positive cells - they have both markers), single positive M4 cells (which express only CD4) and single positive M8 cells (which express only CD8). The pathway of thymocyte development starts with DN cells which differentiate into DP cells, which further lead to single positive cells (either of M4 or M8 type). In order to shed some light on the thymocyte dynamics, two main classes of mathematical models have been proposed. The first class relies on the usage of ordinary differential equations (ODEs) to describe quantitative changes in each T cell population, the most representative models being the compartmental ones proposed in [5, 11]. The second class contains discrete models which take into account the spatial distribution of T cells in the thymic micro-environment and use rules to control the differentiation routes of cells. The main approaches belonging to this class are the reactive animation method [4] and the cellular automata model [8]. Choosing a model depends both on the type of information we are looking for and on the particularities of the available experimental data which can be used to estimate the model parameters.

The overall aim of our research was to obtain information on the impact on the thymocyte dynamics induced by the administration of a glucocorticoid (e.g. dexamethasone). The experimental data contain estimations of the number of T cells in each of the four populations (DN, DP, M4, M8) collected at several time moments starting with the dexamethasone administration. The experiments were conducted on mice and the numerical estimates of each population size were obtained by flow cytometry.

This paper presents an adaptation of the compartmental model introduced in [11] to the particularities of experimental data and proposes an evolutionary approach to explore the parameter space of the model. Due to their ability to deal with nonlinear multimodal objective functions, evolutionary algorithms (EAs) have been successfully used in estimating parameters of various biological models [2, 3, 9, 10]. However, at our best knowledge, none of the previous works on computational models of thymocyte dynamics used EAs for parameter estimation. As long as initial

*GECCO'13,* July 6–10, 2013, Amsterdam, The Netherlands.
Copyright 2013 ACM 978-1-4503-1963-8/13/07 ...$15.00.

parameter values are provided by prior studies, using local search methods seems to be the best choice. On the other hand, when such an information is lacking the global search ability of EAs could be beneficial. Moreover, exploring extensively the parameter space could provide suggestions for further experimental design or ideas for new hypotheses on the involved mechanisms. These motivated us to investigate the particularities of evolutionary parameter estimation in compartmental thymus models.

The rest of the paper is organized as follows. Section 2 presents a compartmental model involving multiple division stages adapted starting from the model proposed in [11]. Section 3 contains a short review of related work on evolutionary design of computational models for biological systems and the description of the selected evolutionary algorithm. Details on the components of the evolutionary approach (e.g. optimization criteria, search strategy) are given on Section 4, while their influence on the simulation results is presented in Section 5. The last section concludes the paper.

## 2. COMPARTMENTAL MODELS FOR THYMOCYTE DYNAMICS

The development of T cells involving several populations and migration between them can be easily modeled using communicating compartments. The first mathematical model of the selection and differentiation processes in the thymus has been proposed by Mehr at al. [5]. In its simplest form it consists of four compartments, one for each thymocyte population. The dynamics of the population in each compartment is described by a differential equation involving terms modeling proliferation, competition, death and transfer/migration between compartments. A particularity of this approach is the fact that it models cell proliferation by a logistic-like growth term characterized by parameters related to the carrying capacity. On the other hand, the model proposed by Thomas-Vaslin et al. in [11] uses several division stages to model the proliferation process.

### 2.1 The model proposed in [11]

In this multi-stage compartmental model, the dynamics of each population is described by several differential equations, each division stage having associated an equation. Thus, the number of equations is related to the number of division stages, while this number is further correlated with the values of the proliferation, death and migration rates. Eqs. (1)-(4) correspond to the processes taking place into the thymus (the original model presented in [11] includes also equations corresponding to a spleen compartment).

$$\begin{aligned}
\dot{N}_0(t) &= \sigma_N - (r_N + d_N)N_0(t) \qquad (1)\\
\dot{N}_i(t) &= 2\gamma(t)r_N N_{i-1}(t) - (r_N + d_N + \mu_N(i))N_i(t),\\
&\qquad i = \overline{1, n_N}
\end{aligned}$$

$$\dot{P}_0(t) = \sum_{i=1}^{n_N} \mu_N(i)N_i(t) + 2\gamma(t)r_N N_{n_N}(t) - (r_P + d_P)P_0(t)$$

$$\begin{aligned}
\dot{P}_i(t) &= 2\gamma(t)r_P P_{i-1}(t) - (r_P + d_P + \mu_P(i))P_i(t), \quad (2)\\
&\qquad i = \overline{1, n_P - 1}
\end{aligned}$$

$$\dot{P}_{n_P}(t) = \sum_{i=1}^{n_P-1} \mu_P(i)P_i(t) + 2\gamma(t)r_P P_{n_P-1}(t) - \mu_{LP}P_{n_P}(t)$$

$$\begin{aligned}
\dot{M}_{40}(t) &= \alpha_4\mu_{LP}P_{n_P}(t) - (r_4 + d_4)M_{40}(t) \qquad (3)\\
\dot{M}_{4i}(t) &= 2\gamma(t)r_4 M_{4,i-1}(t) - (r_4 + d_4 + e_4(i))M_{4i}(t),\\
&\qquad i = \overline{1, n_4 - 1}\\
\dot{M}_{4n_4}(t) &= 2\gamma(t)r_4 M_{4,n_4-1}(t) - (d_4 + e_4(n_4))M_{4n_4}(t)
\end{aligned}$$

$$\begin{aligned}
\dot{M}_{80}(t) &= \alpha_8\mu_{LP}P_{n_P}(t) - (r_8 + d_8)M_{80}(t) \qquad (4)\\
\dot{M}_{8i}(t) &= 2\gamma(t)r_8 M_{8,i-1}(t) - (r_8 + d_8 + e(i))M_{8i}(t),\\
&\qquad i = \overline{1, n_8 - 1}\\
\dot{M}_{8n_8}(t) &= 2\gamma(t)r_8 M_{8,n_8-1}(t) - (d_8 + e_8(n_8))M_{8n_8}(t)
\end{aligned}$$

In these equations the proliferation rates are denoted by $r_N$, $r_P$, $r_4$ and $r_8$ (for DN, DP, M4 and M8 populations, respectively), while the rates corresponding to natural death are denoted by $d_N$, $d_P$, $d_4$ and $d_8$. The process of migration between populations and that of exporting cells outside the thymus are controlled by some rates which depend on the stage number $i$, i.e. $\mu_N(i) = (\alpha_N \cdot i)^n$, $\mu_P(i) = (\alpha_P \cdot i)^n$, $e_4(i) = (\alpha_{e4} \cdot i)^n$ and $e_8(i) = (\alpha_{e8} \cdot i)^n$. If the computed migration/export rate is higher than 1 then it is set to 1. When solving this system of equations, it is supposed that at the initial time moment all cells corresponding to a population are in stage 0 (i.e. $N_0$, $P_0$, $M_{40}$ and $M_{80}$).

In [11] the function $\gamma$ is used to model the influence on the proliferation process of a treatment applied continuously for a given time interval $T$ (e.g. 7 days). More specifically, $\gamma(t) = 0$ for $t \leq T$ (meaning that the proliferation is completely inhibited) and $\gamma(t) = 1$ for $t > T$. Such an approach cannot be applied in the case of a single dose treatment. In such a case it is not easy to set the value for $T$, as it is not a priori known when the impact of the treatment vanishes. Moreover the impact of the treatment can vary between populations.

### 2.2 Extending the model applicability

In order to model the transient influence induced on the thymocyte populations by the administration of the substance in an unique dose we propose to use a continuous perturbation function, $\gamma(t)$ (Eq. (5)), which can model both the depletion and the rebound of each thymocyte population. The process of population depletion, initiated immediately after the substance administration, is modeled by an exponential function depending on a decay rate denoted by $\delta_0$. The recovery process (starting at a moment $\tau_0$) is modeled by a logistic function depending on two parameters: $\delta_1$ (corresponding to the recovery rate) and $\tau_1$ (corresponding to the time moment when the proliferation rate attains half of its original value). In order to ensure the continuity of $\gamma(t)$ one of the four parameters $\delta_0$, $\tau_0$, $\delta_1$ and $\tau_1$ has to be chosen based on the other three parameters such that $\exp(-\delta_0\tau_0) = 1/(1 + \exp(-\delta_1(\tau_0 - \tau_1)))$. In our simulations we considered $\delta_0$, $\tau_0$ and $\delta_1$ as free parameters, while $\tau_1$ was computed using the above mentioned constraint.

$$\gamma(t) = \begin{cases} \exp(-\delta_0 t) & \text{if } t < \tau_0 \\ 1/(1 + \exp(-\delta_1(t - \tau_1))) & \text{if } t \geq \tau_0 \end{cases} \qquad (5)$$

Since the treatment can have different impact on the T cell populations, we consider different perturbing functions, i.e. $\gamma_N(t)$ (for DN), $\gamma_P(t)$ (for DP), $\gamma_4(t)$ (for M4) and $\gamma_8(t)$ (for M8). The modified model will thus have $\gamma_N$, $\gamma_P$, $\gamma_4$ and $\gamma_8$ instead of $\gamma$ in Eqs. (1)-(4), respectively. Therefore there will be 12 new free parameters to be estimated in order to fit the model to the data. It should be also mentioned that, motivated by the variability of the experimental data corresponding to M4 and M8 populations, we considered different proliferation/death/migration parameters for these two populations (unlike the model fitted in [11] where same parameters were used for both populations).

# 3. EVOLUTIONARY DESIGN OF BIOLOGICAL MODELS

Inferring models from data is usually formulated as an optimization problem aiming to identify the model structure and the parameters which minimize the distance between simulated and experimental data. As model simulation typically means numerically solving differential equations, the objective function cannot be computed symbolically and it usually is highly nonlinear and multimodal. This limits the effectiveness of local search methods and explains the increasing interest in using population-based stochastic optimization methods.

## 3.1 Related work

In [9] is analyzed the ability of several evolutionary algorithms to estimate the kinetic parameters of gene regulatory networks, modeled through S-systems and H-systems. Six EAs (binary and real encoded genetic algorithms - GA, standard evolution strategy - ES, covariance matrix adaptation evolution strategy - CMA-ES, differential evolution - DE and particle swarm optimization - PSO) were involved in the comparative study. Best behavior is reported for CMA-ES, closely followed by DE.

The same set of EAs, together with a Simulated Annealing (SA), a Hill Climbing (HC) and a Tribes algorithm (a settings free variant of PSO) have been investigated in [3] in the context of parameter estimation in several mathematical models of metabolic networks. In all experiments, the DE algorithm has been ranked either the best or the second best (after PSO). Based on these results the authors conclude that PSO and DE represent a good choice when optimizing parameters in biological models.

In [10] the parameters of an ODE model describing the dynamics of endocytosis are estimated using a local-derivative based method (A717) and three meta-heuristic algorithms (differential ant-stigmergy algorithm - DASA, PSO and DE). All meta-heuristics significantly outperformed the local search method, while DE proved to be the best of them with respect to the quality of fit. Despite the large number of reported results on applying EAs in designing models for biological systems, there are no reported results on using EAs in designing models for thymocyte dynamics.

## 3.2 Selected EA: JADE [12]

The effectiveness of DE for parameter estimation in biological systems, as reported in several comparative studies, together with its simplicity and flexibility represent strong points which recommend its usage. The sensitivity of the standard DE [7] to its control parameters, has been significantly reduced by recent adaptive variants such as JADE [12], which is currently one of the most effective DE variant [6]. The JADE overall structure, for minimization of $f : S \subset \mathbf{R}^d \to \mathbf{R}$, is described in Algorithm 1 and its main features are: (i) the elements used in the recombination rule described in Eq.(6) are chosen such that a new candidate is created in a neighborhood of a good population element but away from a worse one ($x_{rbest}$ is selected from the $p\%$ elites of the current population and $x_{r2}$ is one of the inferior elements which were discarded in a previous selection step); (ii) the scale factor ($F$) and the crossover probability ($CR$) are generated for each element of the population using a probability distribution (Gaussian and Cauchy, respectively) whose mean is recomputed at each generation using information from successful elements (sets $\mathcal{F}$ and $\mathcal{CR}$ in Algorithm 1).

**Algorithm 1** JADE overall structure

```
Population initialization $X(0) \leftarrow \{x_1(0), \dots, x_m(0)\}$
Compute $\{f(x_1(0)), \dots, f(x_m(0))\}$
Control parameters initialization ($k = \overline{1,m}$):
$F_k(0) \leftarrow \mathcal{N}(a_F(0), \sigma)$, $CR_k(0) \leftarrow \mathcal{C}(a_{CR}(0), \sigma)$
Archive initialization: $\mathcal{A} \leftarrow \emptyset$; $\mathcal{F} \leftarrow \emptyset$; $\mathcal{CR} \leftarrow \emptyset$
$g \leftarrow 0$
while ⟨the stopping condition is false⟩ do
  for $k = \overline{1,m}$ do
    Construct $z_k$ using Eq.(6); Compute $f(z_k)$
    if $f(z_k) < f(x_k(g))$ then
      $x_k(g+1) \leftarrow z_k$; $\mathcal{A} \leftarrow \mathcal{A} \cup x_k(g)$
      $\mathcal{F} \leftarrow \mathcal{F} \cup F_k(g)$; $\mathcal{CR} \leftarrow \mathcal{CR} \cup CR_k(g)$
    else
      $x_k(g+1) \leftarrow x_k(g)$
    end if
  end for
  $g \leftarrow g + 1$
  Compute $a_F(g)$ and $a_{CR}(g)$ using $\mathcal{F}$ and $\mathcal{CR}$
  Control parameters adjustment ($k = \overline{1,m}$):
  $F_k(g) \leftarrow \mathcal{N}(a_F(g), \sigma)$, $CR_k(g) \leftarrow \mathcal{C}(a_{CR}(g), \sigma)$
  Archive pruning by random selection; $\mathcal{F} \leftarrow \emptyset$; $\mathcal{CR} \leftarrow \emptyset$
end while
```

$$z_k^l = \begin{cases} x_k^l + F_k \cdot (x_{rbest}^l - x_k^l) + & F_k \cdot (x_{r1}^l - x_{r2}^l) \\ & \text{if } rand() \leq CR_k \\ x_k^l & \text{otherwise} \end{cases} \tag{6}$$

The only parameters to be specified are: the population size ($m$), the fraction of the top-ranked elements used in the recombination rule ($p \in [0.05, 0.2]$) and a parameter for adjusting the mean of the distribution probabilities used to generate values for $F_k$ and $CR_k$ (a value of 0.5 can be used).

# 4. INFERRING THE MODEL FROM DATA

This section presents specific elements in selecting the optimization criterion and the parameter space search strategy.

## 4.1 Data variability and optimization criteria

The experimental dataset consists of estimates of the number of cells in each of the four thymocyte populations collected at eleven time moments starting with the moment when the substance was injected and ending at day fourteen after this moment ($t_1 = 0$, $t_2 = 0.25$, $t_3 = 0.5$, $t_4 = 1$, $t_5 = 2$, $t_6 = 3$, $t_7 = 4$, $t_8 = 5$, $t_9 = 8$, $t_{10} = 9$ and $t_{11} = 14$).For each

**Table 1: Variability of experimental data corresponding to each thymocyte population.**

| Measure | DN data | DP data | M4 data | M8 data | Avg. |
|---|---|---|---|---|---|
| $Var_0$ | 0.164 | 1.346 | 0.339 | 0.017 | 0.390 |
| $Var_1$ | 0.031 | 0.486 | 0.115 | 0.105 | 0.184 |
| $Var_2$ | 0.007 | 0.008 | 0.015 | 0.006 | 0.009 |

time moment, two to four values are available, the number of data instances corresponding to each thymocyte population being 35. Same number of data instances were provided for each thymocyte population, thus the total number of data is 140. The dataset is characterized by several types of variability: (i) variability over time in the data corresponding to each population, mainly induced by the influence of the administrated substance on the thymocyte dynamics; (ii) variability between the values collected for the same time point, caused by the differences between the mice used in the experiments; (iii) variability over the four thymocyte populations.

The last two variability types can influence the results of a data fitting procedure if a Mean Squared Error (MSE) is used as optimization criterion. Therefore, we analyzed several MSE variants and their corresponding variability measures. Let us denote by $y_{ij}$ the value of sample $j$ corresponding to time $t_i$ ($i = \overline{1,k}$, $j = \overline{1,n_i}$, $n = n_1 + n_2 + \ldots + n_k$) and by $y(t_i, x)$ the estimation, based on the parameter set $x$, of the number of cells at time $t_i$. The general form of MSE is described in Eq. (7).

$$MSE_w(x) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij} (y_{ij} - y(t_i, x))^2 \tag{7}$$

For $w_{ij} = 1$ one obtains the classical mean squared error ($MSE_0$). In several works [1, 3, 9] presenting computational approaches in parameter estimation of biological models is used the relative MSE characterized by $w_{ij} = 1/y_{ij}^2$ (denoted here as $MSE_1$). Taking into account the variability between the data corresponding to the four thymocyte populations we could also consider to weigh differently the $MSE$ terms corresponding to different populations. The variant we consider uses $w_{ij}^{(p)} = 1/(y_{max}^{(p)})^2$ where $y_{max}^{(p)}$ denotes the largest experimental data corresponding to population $p$. In this case there are four weights, one corresponding to each population and the optimization criterion is denoted by $MSE_2$.

The objective function used in parameter estimation is the average of $MSE$ values corresponding to the four populations. In order to decide which $MSE$ variant is less sensitive with respect to the variability over the populations we conducted a preliminary analysis by replacing $y(t_i, x)$ in Eq. (7) with the average of experimental data for each time moment ($\overline{y}_i$). The variability measures corresponding to the three $MSE$ variants described above are denoted by $Var_0$, $Var_1$, $Var_2$ and their values are presented in Table 1, suggesting that $MSE_2$ is the least sensitive.

## 4.2 Searching the parameter space

There are three types of parameters in the multi-stage compartmental model described in Eqs. (1-4): (i) parameters influencing the structure of the model (i.e. the number of division stages for each population: $n_N$, $n_P$, $n_4$, $n_8$ which control the number of differential equations); (ii) parameters corresponding to the rates of proliferation, death and differentiation mechanisms (when no interrelation constraints are imposed on them there are 17 parameters); (iii) parameters involved in the functions used to perturb the proliferation mechanism of each population (12 free parameters if the function is as described in Eq. (5)).

Usually, the structural parameters are estimated first and the other parameters (e.g. rates) are estimated in a second step using fixed values for the structural parameters. In an evolutionary approach this could be done by using a nested algorithm consisting of an outer EA aiming to evolve the structure and an inner EA aiming to estimate the parameters corresponding to each structure. Such an approach, based on two EAs with specific operators, has been proposed in [2] for cell models design. The particularity of the model investigated in this paper is that for any number of division stages the number of non-structural parameters is the same. This property is ensured by the fact that the only difference between the equations describing the dynamics corresponding to different division stages is related to the migration rate value, $\mu(i)$ which in the model proposed in [11] is computed as: $\mu(i) = (\alpha \cdot i)^n$, thus for all values of $i$ there are the same two migration parameters per population ($n$ and $\alpha$). Moreover, same value of $n$ is used for all populations leading to five different parameters related to division stages ($n$, $\alpha_N$, $\alpha_P$,$\alpha_{e4}$ and $\alpha_{e8}$). Therefore the number of stages and all the other parameters can be estimated simultaneously, by interpreting the optimization problem as a mixed integer-continuous one. Previous studies [7] proved that DE can be effectively applied for mixed integer-continuous optimization problems by conducting the search in a continuous space and by converting the values of discrete parameters to integers only for evaluation.

Before deciding on the search variant we conducted a preliminary comparative analysis between a nested variant (based on JADE both at the outer and the inner level) and a simple simultaneous search. The same computational budget (50000 MSE evaluations) has been divided as follows. In the nested case the outer JADE used a population of $m_o = 10$ elements and $g_o = 5$ generations while the inner JADE (called to compute the fitness of each set of number of stages) evolved for $g_i = 100$ generations a population of $m_i = 10$ elements (sets of non-structural parameters).

In the simultaneous evolution approach a population of 20 elements (encoding all parameters) was evolved for 2500 generations. The average quality of the parameters generated by the simultaneous evolution was significantly better than that of parameters estimated by the nested variant ($MSE_2 = 0.0187 \pm 0.001$ vs. $MSE_2 = 0.0265 \pm 0.002$). A possible explanation is that, as Figures 1 and 2 illustrate, in the simultaneous evolution good sets of numbers of stages are evolved continuously for more than $g_i = 100$ generations (the number allowed in the nested case). Moreover, in the nested variant at most $m_o \cdot g_o = 50$ distinct configurations (out of $s^4 = 1296$ possible configurations corresponding to the case when the number of stages belongs to $\{2, 3, \ldots, 7\}$) are explored. On the other hand, the experiments suggest that the number of configurations explored in the simultaneous evolution is higher, as there are runs where more than 50 configurations were selected as best ones at least for one generation. Therefore the results presented in the following section were obtained using the strategy based on

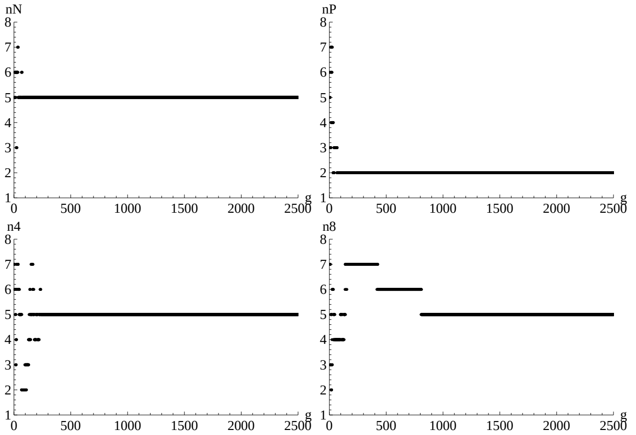

**Figure 1: Evolution of the number of stages in the simultaneous estimation variant. Quality of fit:** $MSE_0 = 0.481$, $MSE_2 = 0.015$.

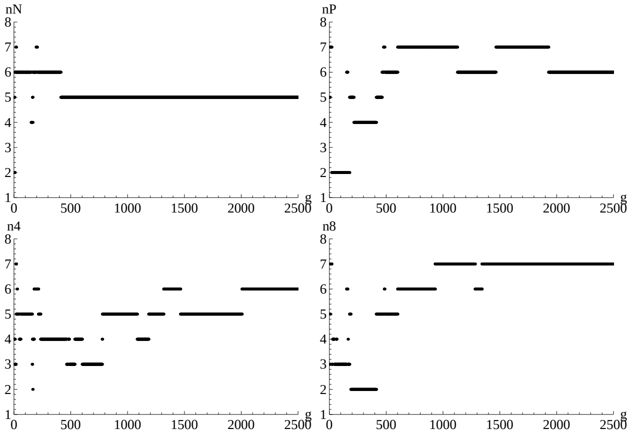

**Figure 2: Evolution of the number of stages in the simultaneous estimation variant. Quality of fit:** $MSE_0 = 0.488$, $MSE_2 = 0.019$.

simultaneously searching for the number of stages and for the non-structural parameters.

## 5. SIMULATION RESULTS AND DISCUSSION

All simulation results reported in this section have been obtained by applying JADE on a population of 20 elements for 2500 generations (i.e. the computational budget was set to 50000 function evaluations). The only parameters of JADE to be set by the user were $a_F = a_{CR} = 0.5$ (the initial value of the mean used to generate the scale factor and the crossover probability) and the percent of elements used by the rand/current-to-pbest strategy ($p = 10\%$).

Each objective function evaluation requires solving a system of up to 30 differential equations. To numerically solve such systems, the ODE solver from Mathematica 7.0 was used. The statistical estimates were obtained by at least 10 independent runs of the parameter estimation procedure.

### 5.1 Influence of the quality of fit measure

Since the optimization criterion used in the evolutionary parameter search is the main guiding element, we compared the results obtained when using each of the three quality of fit measures ($MSE_0$, $MSE_1$, $MSE_2$) presented in subsection 4.1. In each case we computed the values of all three criteria but only one of them was used as optimization criterion. From the results in Table 2 it follows that:

- when using $MSE_0$ or $MSE_2$ as optimization criteria, their values are quite close; also the Pearson correlation coefficient between the $MSE_0$ and $MSE_2$ values collected during 12 runs is over 0.85 in both cases; on the other hand, the correlation coefficient between $MSE_0$ and $MSE_1$ is 0.32 while that between $MSE_2$ and $MSE_1$ is 0.24;
- when using $MSE_1$ as optimization criteria, the estimated parameters lead to values of $MSE_0$ and $MSE_2$ significantly larger than the values obtained when they were used to guide the search.

As a consequence of this preliminary analysis the following simulation results are based on using $MSE_0$ and $MSE_2$ as optimization criteria (but not simultaneously, the problem remaining of single-objective type).

### 5.2 Influence of the migration rate rule

The estimation of the number of stages is related to the dependence between the migration rate and the stage number. In [11] the migration rate is computed using $\mu(i) = (\alpha \cdot i)^n$ and the number of stages, $n_{stages}$, is established using the assumption that it should maximize the probability $Prob(i)$ that a cell completes exactly $i$ division stages. Supposing that $r$ denotes the proliferation rate and $d$ the death rate this probability is described in Eq. (8).

$$Prob(i) = \left(1 - \frac{r}{r + d + \mu(i+1)}\right) \prod_{j=1}^{i} \frac{r}{r + d + \mu(j)} \qquad (8)$$

Table 3 presents the quality of fit obtained in three variants: (i) the number of stages is established at each evolutionary step based on the probabilistic approach described above (this means that only the non-structural parameters are evolved); (ii) the number of stages is evolved simultaneously with the other parameters; (iii) a similar approach as the previous one but based on the assumption that the migration rate depends linearly on the stage number (Eq. 9). In this last case instead of $\alpha$ and $n$, values for $\alpha_{min}$ and $\alpha_{max}$ are to be evolved (two parameters for each T cells population).

$$\mu(i) = \alpha_{min} + (i-1)\frac{\alpha_{max} - \alpha_{min}}{n_{stages}} \qquad (9)$$

The results in Table 3 were obtained by using $MSE_2$ as optimization criterion and based on them one can state:

- the two migration rate rules (the linear one and that based on the power law) lead to parameters of similar quality (the Mann-Whitney statistical test returned a p-value larger than 0.8 for both quality measures);
- evolving the number of stages simultaneously with the other parameters leads to a slightly better fitting than by estimating them using the probability given in Eq. (8) (the Mann-Whitney statistical test returned a p-value smaller than 0.05 in all tested cases).

**Table 2: Quality of fit measures and optimization criteria**

| Optimization<br>criterion | $MSE_0$<br>avg±stdev (min) | $MSE_1$<br>avg±stdev (min) | $MSE_2$<br>avg±stdev (min) |
|---|---|---|---|
| $MSE_0(w_{ij} = 1)$ | $0.511 \pm 0.020$ $(0.487)$ | $1.390 \pm 0.413(1.202)$ | $0.023 \pm 0.001(0.020)$ |
| $MSE_1(w_{ij} = 1/y_{ij}^2)$ | $0.971 \pm 0.272(0.829)$ | $0.203 \pm 0.008(0.187)$ | $0.036 \pm 0.005(0.039)$ |
| $MSE_2(w_{ij}^{(p)} = 1/(y_{max}^{(p)})^2)$ | $0.512 \pm 0.038(0.477)$ | $1.187 \pm 0.317(1.250)$ | $0.018 \pm 0.001(0.017)$ |

**Table 3: Influence of the migration rate rules and the number of stages estimation on the quality of fit**

| Quality | $\mu(i) = (\alpha \cdot i)^n$ | | $\mu(i) = \alpha_{min} + (i-1) * \dfrac{\alpha_{max} - \alpha_{min}}{n_{stages}}$ |
|---|---|---|---|
| measure | Prob. estim. of $n_{stages}$<br>avg±stdev | Evol. estim. of $n_{stages}$<br>avg±stdev | Evolutionary estimation of $n_{stages}$<br>avg±stdev |
| $MSE_0$ | 0.6162±0.0777 | 0.5514±0.0593 | 0.577±0.125 |
| $MSE_2$ | 0.0200±0.0017 | 0.0187±0.0014 | 0.0189±0.0022 |

## 5.3 Exploring the outputs of the evolutionary search

Since the stochastic character of EAs requires repeated runs, such a parameter estimation procedure provides a significant set of candidate solutions which can be explored in order to collect information about the particularities of the model. The results presented in this section are based on 40 independent runs of the evolutionary procedure aiming to simultaneously estimate both the number of stages and all the other parameters (a total of 33 parameters). Half of the runs were based on using $MSE_0$ as optimization criterion and the other half used $MSE_2$. The overall quality of fit is illustrated in the last two columns of Table 3. A close inspection of the estimated thymocyte dynamics corresponding to these results revealed that small $MSE$ values do not necessarily ensure a biologically plausible behavior. Therefore a filtering of the results involving constraints concerning the long term behavior of the population would allow to select relevant outputs. As the thymocyte dynamics is expected to stabilize after 2-3 weeks since the glucocorticoid administration, constraints on the values of the derivatives of $N(t)$, $P(t)$, $M_4(t)$ or $M_8(t)$ seems natural. Using $MSE_0 \leq 0.55$, $MSE_2 \leq 0.02$ and $\dot{M}_8(20) < 0.1$ as filtering criteria one obtained 8 sets of parameters corresponding to biologically plausible simulation models. Figure 3 illustrates two examples corresponding to the smallest values of $MSE_0$ and $MSE_2$. In the same time they illustrate two potential behaviours: (i) a first one when all populations reach a stationary state as soon as proliferation became fully active; (ii) a second one characterized by an involution stage following the rebound process.

The estimated values of the parameters and the statistics collected by independent runs of the EA can be used both to validate the model from a biological point of view and to refine the model or the estimation procedure. Table 4 presents averages, standard deviations and value ranges for all free parameters of the model. They were computed starting from the 8 best fitted models selected from results of 40 independent runs of the evolutionary procedure. A first analysis allows to check if the estimated values are in accordance with existing knowledge and to obtain preliminary and rough information concerning the model sensitivity. For instance, the estimated values of the death rates $d_N$ and $d_P$ are in accordance with the fact that the DP cells death rate is significant, while the rate corresponding to DN is almost negligible. Also the small value of $\mu_{LP}$ combined with the small values of $\alpha_4$ and $\alpha_8$ are in accordance with experimental observations concerning the dramatic loss of DP cells by negative selection [5, 11].

On the other hand, there are parameters with a large range of estimated values, suggesting that the model is less sensitive to their values. Such a parameter is the exponent $n$ involved in the rule used to compute the migration rate. The small sensitivity of the model to this parameter has been also remarked in [11] (even if the model and the data are different). The fact that the model is almost insensitive to the migration rate rule is supported also by the comparative results presented in Table 3. Looking at the estimated values of the parameters involved in the perturbation functions $\gamma(t)$ one can remark that $\tau_{40}$ and $\tau_{80}$ (time moments marking when the depletion of the proliferation process stops) are significantly smaller than $\tau_{N0}$ and $\tau_{P0}$ suggesting that the influence of the administrated substance on cells in populations $M4$ and $M8$ is shorter in time, while the most affected cells are those of DP. The different impact of the treatment on the four thymocyte populations is also illustrated in Figure 4 where are plotted the perturbation functions corresponding to all 8 selected sets of estimated parameters. The variability in the set of estimated values for the same parameter is mainly caused by the correlations between different parameters which make their true values practically unidentifiable.

This means that models with significantly different values of the parameters can behave similarly, making difficult the parameter identification in the absence of a priori biological constraints. A first step in dealing with this problem is to limit the search range of some parameters. In the context of thymocyte dynamics modelling, the parameters which are not involved in the perturbation terms and which are supposed to control the normal thymocyte dynamics can be estimated using experimental data on normal thymus. Therefore we analyzed the effectiveness of the following two-steps approach:

*Step 1.* The parameters of the non-perturbed model (i.e. $\gamma(t) = 1$ for any $t$) are estimated using experimental data corresponding to the normal thymus behaviour and using large search ranges (as those mentioned in Table 4); the values estimated in 20 independent runs are used to define new ranges for parameters.

*Step 2.* All parameters of the perturbed model are estimated

**Table 4: Estimated parameters. Statistics based on 8 runs selected from 40 runs such that $MSE_0 \leq 0.55$, $MSE_2 \leq 0.02$, $\dot{M}_8(20) < 0.1$. Search ranges: $s_n \in [0.0001, 0.05]$, $r_P \in [0, 5]$, $n \in [5, 50]$, $n_N, n_P, n_4, n_8 \in [1.51, 7.49]$, $\tau_{N0}, \tau_{P0}, \tau_{40}, \tau_{80} \in [0, 7]$, for all $\delta$ values the search range is $[0, 20]$ and $[0, 1]$ for all other parameters.**

| Param. | Avg±StDev | Estimated range | Param. | Avg±StDev | Estimated range | Param | Avg±StDev | Estimated range |
|---|---|---|---|---|---|---|---|---|
| $r_N$ | 0.55±0.1 | [0.38,0.79] | $d_N$ | 0.07±0.04 | [$10^{-6}$,0.13] | $\alpha_N$ | 0.21±0.09 | [0.07,0.32] |
| $r_P$ | 0.79±0.19 | [0.47,1.10] | $d_P$ | 0.92±0.05 | [0.84,0.99] | $\alpha_P$ | 0.63±0.22 | [0.29,0.97] |
| $r_4$ | 0.80±0.23 | [0.27,0.99] | $d_4$ | 0.42±0.08 | [0.28,0.51] | $\alpha_4$ | 0.20±0.06 | [0.11,0.27] |
| $r_8$ | 0.87±0.14 | [0.59,0.99] | $d_8$ | 0.36±0.15 | [0.12,0.62] | $\alpha_8$ | 0.16±0.05 | [0.07,0.23] |
| $s_N$ | 0.04±0.004 | [0.03.0.05] | $n$ | 76±36 | [36,144] | $\alpha_{e4}$ | 0.44±0.17 | [0.12,0.73] |
| $n_N$ | 4.87±0.59 | [4,6] | $\mu_{LP}$ | 0.13±0.05 | [0.07,0.25] | $\alpha_{e8}$ | 0.25±0.07 | [0.14,0.38] |
| $n_P$ | 3.87±2.20 | [2,7] | $n_4$ | 4.87±1.05 | [3,6] | $n_8$ | 5.62±1.21 | [3,7] |
| $\tau_{N0}$ | 1.27±0.39 | [0.68,1.82] | $\delta_{N0}$ | 2.75±1.91 | [1.24,7.05] | $\delta_{N1}$ | 12.43±4.88 | [6.42,19.99] |
| $\tau_{P0}$ | 1.37±0.39 | [0.85,1.98] | $\delta_{P0}$ | 14.28±3.20 | [9.46,19.88] | $\delta_{P1}$ | 14.31±3.80 | [8.34,19.08] |
| $\tau_{40}$ | 0.30±0.12 | [0.16,0.51] | $\delta_{40}$ | 17.10±2.24 | [13.77,20] | $\delta_{41}$ | 11.38±5.25 | [4.68,19.32] |
| $\tau_{80}$ | 0.35±0.11 | [0.19,0.56] | $\delta_{80}$ | 14.17±4.45 | [5.81,19.99] | $\delta_{81}$ | 12.72±3.20 | [6.38.16.59] |

using, for the parameters of the non-perturbed model, the restricted ranges obtained at the previous step.
This approach limited significantly the number of overfitting cases (models with small MSE but not biologically plausible long term behavior): in all 20 runs the fitted model is characterized by a biologically plausible long term behavior of all populations and the variability between runs is also smaller ($MSE_2 = 0.0183 \pm 0.0003$ vs. $MSE_2 = 0.0187 \pm 0.001$).

## 6. CONCLUSIONS

The usage of parameterized continuous functions to perturb the proliferation term of the model introduced in [11] allos its usage to model the perturbation induced by single dose administration of a glucocorticoid. The evolutionary procedure for simultaneously estimating the structural and non-structural parameters leads to models which fit well the experimental data as long as in the optimization criterion choice the data variability is taken into account. The good quality of fit of the evolved model, especially when the search range is controlled, makes it reliable for predictions. However not the same can be said about the estimated parameters if their intercorrelation is not carefully taken into account during the estimation procedure. Therefore one of the problems to be addressed in a further research is to conduct a multiple correlation analysis and to investigate the parameters identifiability.

## 7. ACKNOWLEDGMENTS

This work was supported by grant PN-II-ID-PCE-2011-3-0571 and by the strategic grant POSDRU/CPP107/DMI1.5/S/78421, Project ID 78421(2010), co-financed by the European Social Fund-Investing in People, within the Sectorial Operational Programme Human Resources Development 2007-2013. The authors thank dr. F. Mic for providing the experimental data.

## 8. REFERENCES

[1] C. Baker, G. Bocharov, J. Ford, P. Lumb, S. Norton, C. Paul, T. Junt, P. Krebs, and B. Ludewig. Computational approaches to parameter estimation and model selection in immunology. *J. of Comput. and Appl. Math.*, 184:50–76, 2005.

[2] H. Cao, F. Romero-Campero, S. Heeb, M. Camara, and N. Krasnogor. Evolving cell models for systems and synthetic biology. *Syst. Synth. Biol.*, 4:55–84, 2010.

[3] A. Drager, M. Kronfeld, M. Ziller, J. Supper, H. Planatscher, J. Magnus, M. Oldiges, O. Kohlbacher, and A. Zell. Modelling metabolic networks in c. glutanicum: a comparison of rate laws in combination with various parameter optimization strategies. *BMC Systems Biology*, 3(5), 2009.

[4] S. Efroni, D. Harel, and I. Cohen. Towards rigorous comprehension of biological complexity: Modeling, execution and visualization of thymic t-cell maturation. *Genome Res.*, 13:2485–2497, 2003.

[5] R. Mehr, A. Globerson, and A. Perelson. Modelling positive and negative selection and differentiation processes in the thymus. *J. Theor. Biol.*, 175:103–126, 1995.

[6] P. Posik and V. Klema. Jade, an adaptive differential evolution algorithm, benchmarked on the bbob noiseless testbed. In *GECCO'12 Proc.*, pages 197–204, 2012.

[7] K. Price, R. Storn, and J. Lampinen. *Differential Evolution. A Practical Approach to Global Optimization.* Springer, Berlin, Heidelberg, 2005.

[8] H. Silva, W. Savino, R. Feijoo, and A. Vasconcelos. A cellular automata-based mathematical model for thymocyte development. *PLoS ONE*, 4(12:e8233), 2009.

[9] C. Spieth, R. Worzischek, and F. Streichert. Comparing evolutionary algorithms on the problem of network inference. In *GECCO'06 Proc.*, pages 305–306, 2006.

[10] K. Tashkova, P. Korosek, J. Silc, L. Todorovski, and S. Dzeroski. Parameter estimation with bio-inspired meta-heuristic optimization: modeling the dynamics of endocytosis. *BMC Systems Biology*, 5(159), 2011.

[11] V. Thomas-Vaslin, H. K. Altes, R. de Boer, and D. Klatzmann. Comprehensive assessment and mathematical modeling of t cell population dynamics and homeostatis. *J Immunol.*, 180(4):2240–2250, 2008.

[12] J. Zhang and A. Sanderson. Jade: adaptive differential evolution with optional external archive. *IEEE Trans. on Evol. Comput.*, 13(5):945–95, 2009.

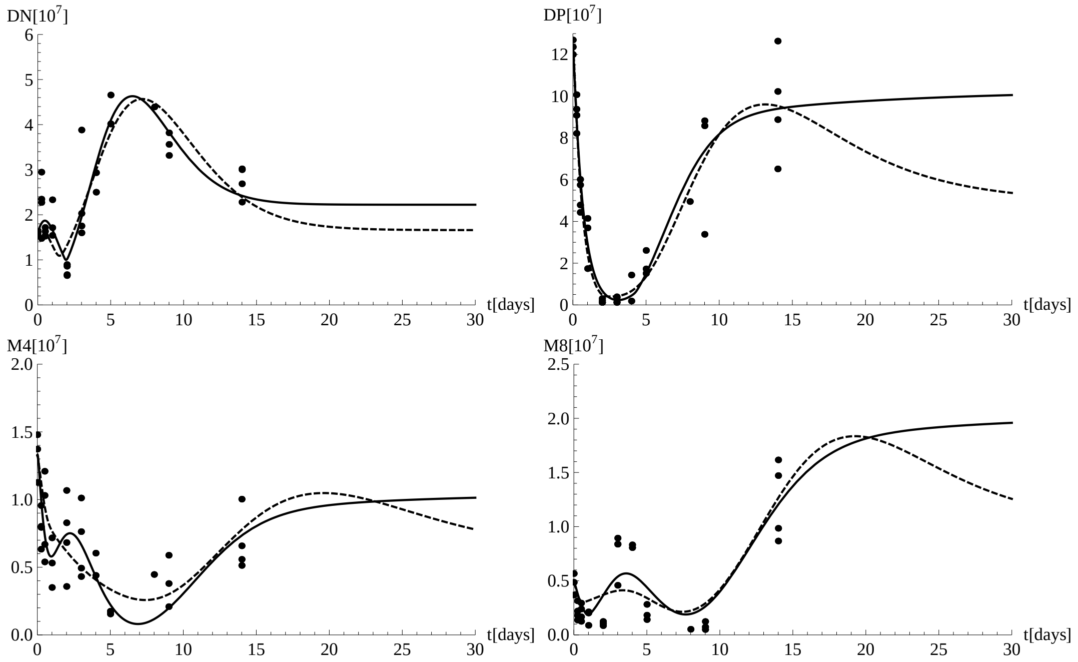

**Figure 3: Estimated T cell populations dynamics (examples with smallest MSE values in 40 independent runs). Points: experimental data (number of cells/$10^7$). Continuous line: $MSE_0 = 0.481$, $MSE_2 = 0.015$, number of division stages: $n_N = 5$, $n_P = 2$, $n_4 = 5$, $n_8 = 5$. Dashed line: $MSE_0 = 0.488$, $MSE_2 = 0.019$, number of division stages: $n_N = 5$, $n_P = 6$, $n_4 = 6$, $n_8 = 7$.**

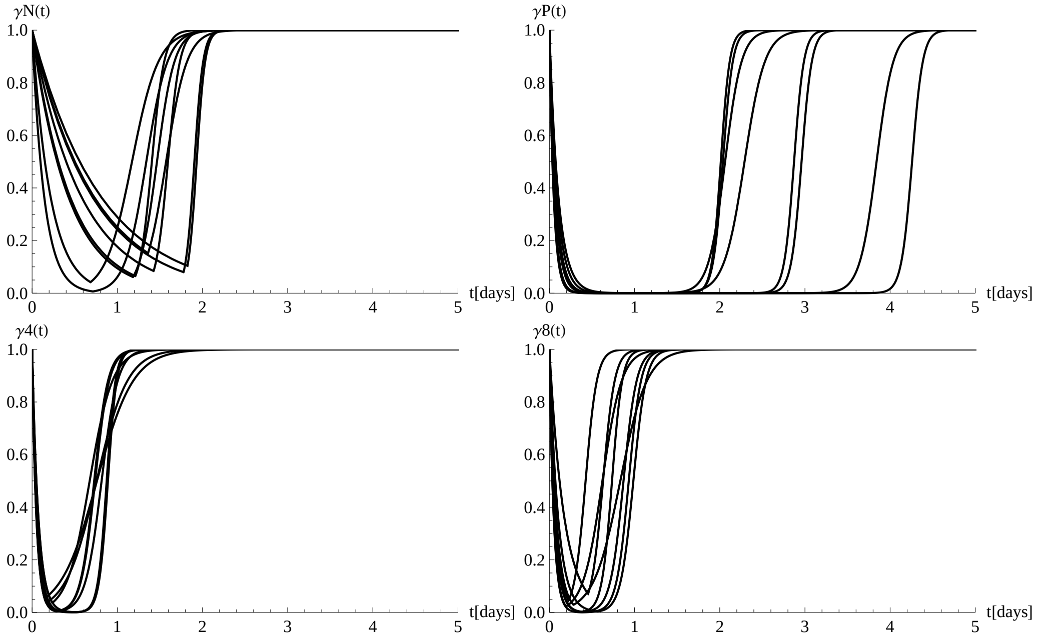

**Figure 4: Perturbation functions corresponding to selected models ($MSE_0 \leq 0.55$, $MSE_2 \leq 0.02$, $\dot{M}_8(20) \leq 0.1$).**