

# Modelling the Human Immune System by Combining Bioinformatics and Systems Biology Approaches

Nicolas Rapin · Can Kesmir · Sune Frankild ·  
Morten Nielsen · Claus Lundegaard ·  
Søren Brunak · Ole Lund

Received: 10 April 2006 / Revised: 18 May 2006 / Accepted: 2 June 2006 /  
Published online: 27 October 2006  
© Springer Science + Business Media B.V. 2006

**Abstract** Over the past decade a number of bioinformatics tools have been developed that use genomic sequences as input to predict to which parts of a microbe the immune system will react, the so-called epitopes. Many predicted epitopes have later been verified experimentally, demonstrating the usefulness of such predictions. At the same time, simulation models have been developed that describe the dynamics of different immune cell populations and their interactions with microbes. These models have been used to explain experimental findings where timing is of importance, such as the time between administration of a vaccine and infection with the microbe that the vaccine is intended to protect against. In this paper, we outline a framework for integration of these two approaches. As an example, we develop a model in which HIV dynamics are correlated with genomics data. For the first time, the fitness of wild type and mutated virus are assessed by means of a sequence-dependent scoring matrix, derived from a BLOSUM matrix, that links protein sequences to growth rates of the virus in the mathematical model. A combined bioinformatics and systems biology approach can lead to a better understanding of immune system-related diseases where both timing and genomic information are of importance.

**Key words** mathematical modelling · HIV · bioinformatics · simulation · immunology · immune system

## Introduction

Many “virtual organism” simulation projects are currently underway in the US, Japan and Europe: Visible Human, Virtual Soldier, and Virtual Patient projects – with titles that indicate both the area of focus and the level of abstraction. These projects address mainly

---

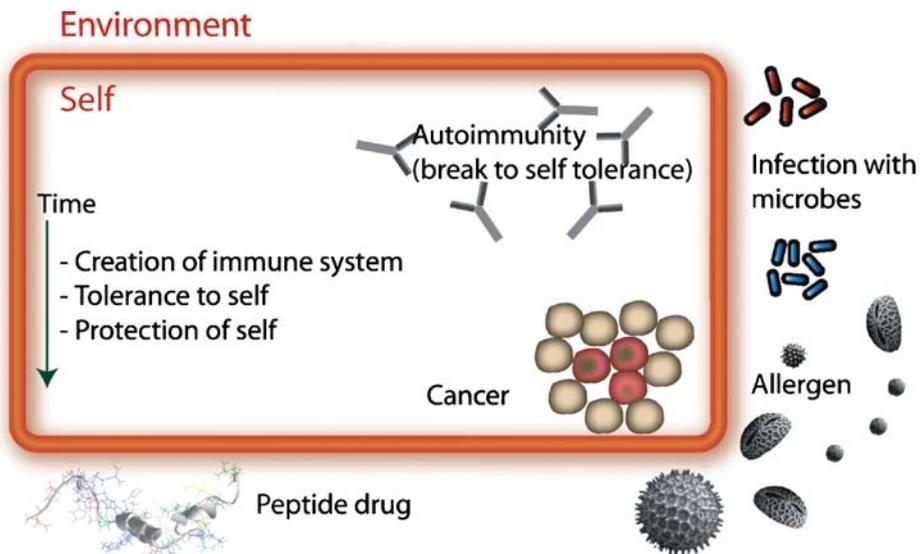
N. Rapin · S. Frankild · M. Nielsen · C. Lundegaard · S. Brunak · O. Lund (✉)  
Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark,  
Building 208, 2800 Lyngby, Denmark  
e-mail: lund@cbs.dtu.dk

C. Kesmir  
Theoretical Biology/Bioinformatics, Utrecht University, Utrecht, The Netherlands

physical and physiological aspects of the human body and responses to stimuli or injury – for example, mechanical simulation of the human heart. The immune system is another example where simulations are performed at high organization levels such as in the European ImmunoGrid project (<http://www.immunogrid.org/>). An immune system is essential to the survival of the organism, and it is a vital area of research in relation to the development of more effective and safer treatments for cancer, autoimmunity, and infectious diseases.

As illustrated in Figure 1, the human immune system has to cope with a lot of challenges both externally and internally. The major assignment of an immune system is to defend the host against infections, a task that is clearly essential to any organism. At the same time, the immune system must create a tolerance to the recognition of “self” to avoid autoimmunity. The primary role of the immune system is obviously to be able to recognize and fight any foreign invader, but it should not overreact, since inadequate responses may lead to autoimmunity [1, 2] or to allergic responses.

Part of the immune system, known as the adaptive immune system, has the ability to adapt to specific microbes and remember them so that it can react more efficiently if it encounters the same threat again. The adaptive immune system is mainly composed of T and B lymphocytes and it can generate humoral and cellular responses. In the humoral immune response, the B lymphocytes are activated to produce a soluble form of their antigen receptors called antibodies. Antibodies neutralize their target (the antigen) by binding to it. The cellular immune response is generated by T lymphocytes, which recognize fragments of proteins (peptides) expressed on the surface of the host cells. The peptides are bound to the Major Histocompatibility Class I (MHC I) and Class II (MHC II) molecules. The cytotoxic T lymphocytes (CTL) will kill infected cells that present non-self



**Figure 1** Schematic view of the main challenges to the human immune system. The environment is everything that is non-self, including microorganisms, viruses, allergens, and drugs. Those agents are primary targets for the immune system upon penetration to the inside of the body. Cancer and autoimmunity are forms of aggression that come from within the organism, when normal internal milieu functions are disrupted.

peptides bound to MHC I. The helper T lymphocytes recognize non-self peptides presented on MHC II, become activated, and generate cytokines that will skew the immune system towards humoral or cellular responses. Both CTLs and B cells need this helper component to become fully activated.

### **Diversity of the Immune System**

While many other traits of the human organism can be linked to particular genes, immune systems have always been viewed as systems, in the sense that their genetic foundation is complex and based on a multitude of proteins in many pathways, which interact with each other to coordinate the defense against infection. The immune system is a “combinatorial” system with a large number of products, typically about  $10^9$  different antibodies and more than  $10^{8-9}$  different T lymphocytes (clones) in a given individual. The diversity of the immune system forms the basis for its ability to fight a very large number of pathogens and the ability to discriminate between self and non-self.

Each T lymphocyte is generated with a unique T cell receptor (TCR) that can recognize peptides presented on MHC molecules. The human T cells are matured in the thymus where all the T cells with a TCR that cannot bind to a self-MHC molecule are killed. This process is called positive selection. It ensures that the host does not waste energy on T cells that have no chance of functioning as they are supposed to in the immune system. The other process that occurs in the thymus is the negative selection where cells that react strongly to self-peptides are killed. This prevents autoimmunity.

### **Modelling the Immune System**

Several methodologies are used to model immune systems. Complex generalized cellular automata are simulations based on the global consequences of local interactions of members of a population and have been proposed as models of the immune system [3, 4]. The IMMSIM program [3, 4] uses a generalized cellular automaton to simulate clonotypic cell types and their interactions with other cells, and with antigens and antibodies. Pappalardo et al. [5] explicitly implement the cellular and humoral immune responses in a set of rules relating to the spatio-temporal interaction of cellular and molecular entities. These models typically consist of an environment or framework, in which the interactions occur, and individuals are defined in terms of their behaviors (procedural rules) and characteristic parameters.

This stands in contrast to mathematical modelling techniques where the characteristics of the population are averaged. Two main kinds of mathematical models have emerged over the years: descriptive and analytical models. The descriptive models aim to fit biological data, and generate mathematical constants within a defined experimental setup in order to achieve prediction through a statistical model calibration. The analytic models, based on systems of ordinary differential equations (ODEs), take into consideration the mechanisms involved in the studied system, and for which accuracy and predicting power are sacrificed in favour of abstract concepts. A famous example of the latter models is the Lotka–Volterra model for predator/prey interactions. In the case of immunology, the simplest models have only one compartment, representing the whole body. In order to ameliorate the over-simplifications embedded in ODE models, it is possible to increase the number of

compartments, which can explain phenomena that cannot be reduced to one-compartment models. For example, the increase in the number of helper T cells ( $CD4^+$  T cells) in the peripheral blood after anti-HIV treatment has been shown to be consistent with a mathematical model where the redistribution between lymph nodes and peripheral blood is a function of the viral burden [6]. Modelers have to estimate the parameters in their simulations if they intend to achieve realistic behavior of their system. The common practice in mathematical biology, apart from guessing, is to review experimental data, and extract numerical values from the lab-bench, e.g., estimates of average turnover rates for cells can be extracted from papers such as [7, 8]. The turnover rates for specific clones may be modeled as a function of the concentration of different antigens [9]. Newer experimental techniques such as DNA microarrays may also be used to generate data for modelling the interaction between pathogens and their host. Recently, DNA microarrays have been used to estimate immunoregulatory gene networks in human herpesvirus type 6-infected T cells [10].

An alternative approach to immune system modelling might be to consider differential equations as agents, and have a system of equations that dynamically updates itself. The example provided later in this paper is an attempt to do so. Finally, immunological bioinformatics is a new discipline emerging from the growing knowledge gathered for decades in experimental immunology and immunogenomics [11]. So far the main focus of immunological bioinformatics research has been to develop methods that can identify immunogenic regions in any pathogen genome.

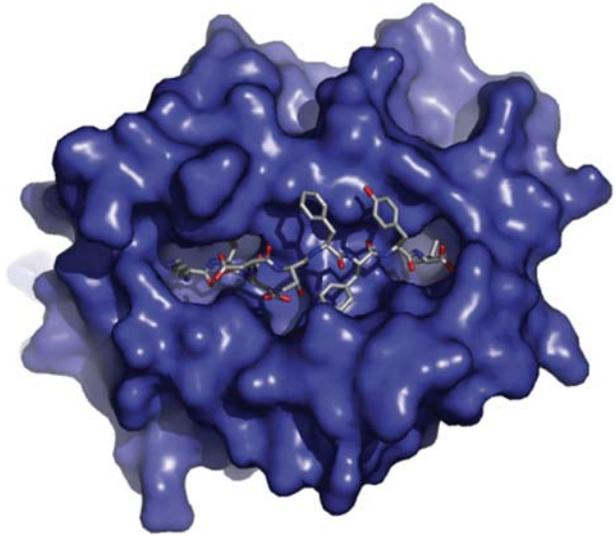
Immune system models have recently been used to answer a number of immunologically relevant questions and to investigate controversial new theories or mechanisms. Examples include:

- Identification of the key factors that contribute to the success of the Mycobacterium tuberculosis [12].
- Evaluation of the effectiveness of antihepatitis C virus treatment.
- Elucidation of the optimal tradeoffs between using energy to fight a disease rather than accepting it [13].
- Determination of the criteria for success of gene therapy against HIV [14].
- Exploration mechanisms for self-tolerance induction [15].
- Explanation of how appropriate concentrations of immune cells are maintained [16].
- Description of molecular dynamics leading to recognition of an antigen [17, 18].

Existing models of the immune system do not cover much detail at the molecular level, such as protein sequences and structures. The lymphocyte receptors are often represented by one-dimensional bit strings, and the model immune system can recognize complementary “epitope” (the parts recognized by the lymphocyte receptors) bit-strings in microbes. However, epitopes of microbes are not bit strings in reality but rather strings of 20 different amino acids with complex contextual binding properties. The lymphocyte receptors undergo a genetic rearrangement so that each clone of lymphocytes has a different receptor and recognizes a different amino acid string. An example of an MHC molecule presenting a CTL epitope can be seen in Figure 2.

Thus, models using bit strings cannot describe infection by any specific virus, with a given genome, but rather models a “generic” virus. Using bit strings makes simulation of the immune system computationally more tractable, but also makes it difficult to accurately predict the outcome of specific diseases and to use the results for rational vaccine design

**Figure 2** Molecular model of an MHC molecule (in blue surface representation) presenting a peptide (in stick representation) with the amino acid sequence KVDDTFYYV to the immune system. The peptide has been used as a vaccine [19]. Figure courtesy of Anne Mølgaard.

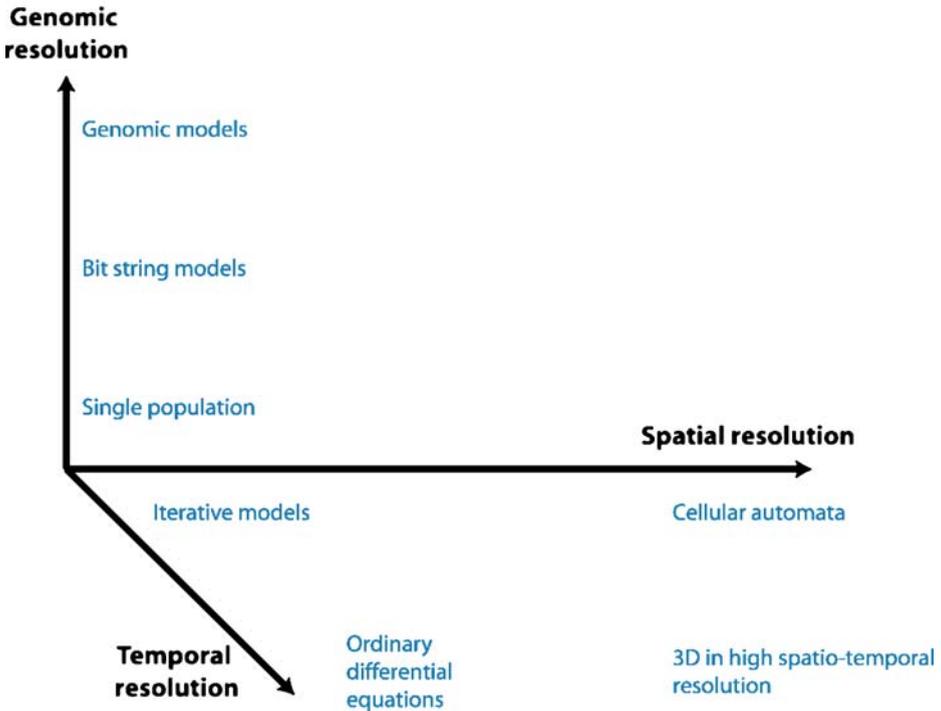


[20]. In addition to focusing on obtaining a good resolution in time and space, we believe that it is important to focus on what we may call “genomic resolution” (see Figure 3). Adding genomic resolution will allow us to understand resistance and susceptibility patterns between pathogens and genetically different individuals, because such models will relate directly to specific pathogens and specific parts of the human genome. Obviously, vaccine design will benefit enormously from integration of genomic resolution into current models. However, adding genomic resolution will be computationally much more demanding. One possibility of overcoming this computational problem is to keep the models as simple as possible. Full-scale computational modelling of the entire immune system is complicated because it relies on integration of many different components. However, many of these components may be much simpler models that describe how immune systems - step by step - deal with pathogens. These simpler models may successfully represent important subcomponents of the immune system, provided that the biological mechanisms controlling the various model steps are sufficiently well understood.

The time is ripe to add genomic resolution to models of the immune system, because we and others have developed methods for predicting which amino acid strings the immune system can react to (HTL/MP [21, 22], CTLs [23], B cells [24, 25]). We are also developing methods based on protein–protein docking to predict the parts of viral, bacterial or tumoral antigens that antibodies bind to. These can be formulated as simple iterative models (before/after infection), as ordinary or partial differential equations, or as cellular automata. In the first round of simulations it may be helpful to concentrate on T-cell epitopes, and especially the CTL epitopes, since these are the best described in terms of experimental data.

We hope that by developing simulation models of the human immune system where genomic resolution is taken into account, one can obtain both a better understanding of immune system-related diseases and a more rational approach to advanced vaccine design.

# Model resolution

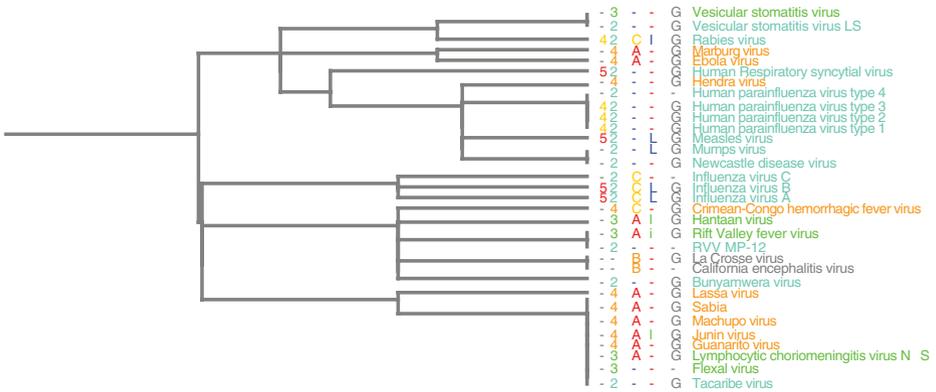


**Figure 3** Schematic view of the potential resolution in different dimensions of an immune system model. Three types of dimensions are taken into account here, namely, temporal, spatial and genomic resolutions. Temporal resolution can vary from discrete to continuous. Genomic resolution represents how close the model is to immunological variability. Finally, spatial resolution explains which level of organization is taken into account and which sub levels are included (cellular compartments, cells, organs or even populations).

## Realistic Models of Pathogens and Hosts

In order to make more realistic simulations we have developed a database with all known microbes that can cause disease in humans (pathogens). This database includes all proteins and all available complete genomes for these pathogens. Figure 4 shows an overview of some of the viruses in the database.

Natural immune systems have presumably been selected for survival of the population (such that no individual pathogen will be able to bring down an entire species), not primarily for the survival of the individual. Most human vaccines today are also designed to fight pathogens on a statistical basis, in the sense that they are not equally effective for all individuals in a population. A main target for immune system modelling should not just be the “prototypical” features of these defense systems; rather, the models should be designed to take key differences of individuals into account. In the future the genome of any human may be obtained fast and at low cost, but for now, for each gene, we only have the few publicly available sequences from the human genome project. These need to be modified in order to be used in simulations. For example, the human genome in the NCBI ([www.ncbi.nlm.nih.gov/genome/guide/human/](http://www.ncbi.nlm.nih.gov/genome/guide/human/)) only contains one copy of each gene, while



**Figure 4** The Dodo database: Pathogenic viruses. *First column:* log<sub>10</sub> of the number of deaths caused by the pathogen per year. *Second column:* DNA Advisory Committee (RAC) classification. DNA Advisory Committee guidelines [41] includes those biological agents known to infect humans, as well as selected animal agents that may pose theoretical risks if inoculated into humans. RAC divides pathogens into four classes. Risk group 1 (RG1): Agents that are not associated with disease in healthy adult humans. Risk group 2 (RG2): Agents that are associated with human disease which is rarely serious and for which preventive or therapeutic interventions are often available. Risk group 3 (RG3): Agents that are associated with serious or lethal human disease for which preventive or therapeutic interventions may be available (high individual risk but low community risk). Risk group 4 (RG4): Agents that are likely to cause serious or lethal human disease for which preventive or therapeutic interventions are not usually available (high individual risk and high community risk). *Third column:* CDC/NIAID bioterror classification. Classification of the pathogens according to the Centers for Disease Control and Prevention (CDC) bioterror categories A–C, where category A pathogens are considered the worst bioterror threats. *Fourth column:* vaccines available. A letter indicates the type of vaccine if one is available (A acellular/adsorbent, C conjugate, I inactivated, L live, P polysaccharide, R recombinant, S staphage lysate, T toxoid). Lower case indicates that the vaccine is released as an investigational new drug (IND). *Fifth column:* G complete genome is sequenced. This database covers all known pathogens (figure shows part of the virus component only; data derived from [www.cbs.dtu.dk/databases/Dodo](http://www.cbs.dtu.dk/databases/Dodo)).

the IPI database ([www.ebi.ac.uk/IPI/IPIhelp.html](http://www.ebi.ac.uk/IPI/IPIhelp.html)) contains many alleles (different versions) for polymorphic genes. A human genome normally contains two copies of most genes, with the genes on the sex chromosomes in males being the most notable exception. To study the outcome of diseases in different individuals we are presently developing a computer program to construct a human genome that takes into account polymorphic genes and the sex of an individual. Special focus will be on the polymorphic genes of the immune system such as the MHC molecules.

The “systems biology” approaches described here can be designed to model immune systems for individual persons, and this is likely to change the situation dramatically, leading to an optimal interaction between an individualized vaccine and the immune system. In the rest of this article, we develop a model of HIV infection to demonstrate the usefulness and feasibility of the ideas presented above.

**Mathematical and Bioinformatic Model of HIV Infection**

To give a concrete example of the combined approach we presented above, we have chosen to model CTL escape mutant dynamics during HIV infection. The model to be presented in this paper was inspired from the well-known model developed by Perelson [26] which traditionally follows the evolution of the viral load, target cells and infected cells over time.

It is an ODE-based mechanistic model that has received wide acceptance in the mathematical biology community. Our model combines this mathematical modelling with bioinformatics. We investigate how the region containing the major HIV gag 77–85 epitope ‘SLYNTVATL’ [27] evolves together with the immune response against it.

**Model Assumptions**

Healthy CD4<sup>+</sup> T cells (*T*, for Target cells) are assumed to be produced at a fixed rate  $\sigma$  and die at rate  $\delta_T$ . The virus (*V*) is able to infect healthy target cells, turning them into infected cells (*I*).  $f$  is the efficiency of infection by viral particles. Infected cells trigger an effector cellular immune response (*E*), which kill infected CD4<sup>+</sup> cells. Each infected cell produces  $b$  virus particles when it dies.

**Processes Involved in the Model**

- The production of infected CD4<sup>+</sup> cells from healthy cells, due to the interaction of the latter with virus particles with efficiency  $f$ , and collision rate  $\beta$ , following the law of mass action.
- The death of infected cells *I* at rate  $\delta_I$ .
- The interaction of infected CD4<sup>+</sup> and effector CD8<sup>+</sup> leading to the killing of infected cells at rate  $k$ . The logic behind this term is extracted from enzyme kinetics assuming that none of the variables are limiting the maximal rate  $k$  [28–30].
- The creation of virus particles from dying infected CD4<sup>+</sup> cells. It is assumed that the bursting of an infected cell releases  $b$  functional viral particles.
- The natural decay of virus, at rate  $c$ .
- The proliferation of the CD8<sup>+</sup> cells, with maximum rate set to  $\alpha$ . The same logic is applied as in the modelling of infected cell killing.
- The natural loss of CD8<sup>+</sup> cells at rate  $\delta_E$ .
- The production of healthy CD4<sup>+</sup> cells at a rate of  $\sigma$ .
- The natural death of target cells *T*, at rate  $\delta_T$ .
- The loss of target cells by infection.

With parameters defined as above, the full system is defined by the following set of ODEs:

**Equations**

$$\begin{cases} \frac{dT}{dt} = f\beta VT - \delta_I I - \frac{kEI}{h+E+I} \\ \frac{dV}{dt} = b\delta_I I - cV \\ \frac{dE}{dt} = \frac{\alpha EI}{k_m + E + I} - \delta_E E \\ \frac{dT}{dt} = \sigma - f\beta VT - \delta_T T \end{cases} \tag{1}$$

We have used the ODE23S solver from Matlab (The MathWorks, Inc.), which is an implementation of an explicit Runge–Kutta pair of Bogacki and Shampine to solve these differential equations.

## Initial Values of Variables

The total concentration of target  $CD4^+$  cells is assumed to be  $10^7 \text{ ml}^{-1}$ . This value is typical for *in vitro* experiments, though activated lymph nodes might show higher concentrations. There is no infected cell in the beginning, no immune response, and the viral load is set to  $10^4 \text{ ml}^{-1}$  particles. One HIV-specific  $CD8^+$  T cell is introduced in the beginning, i.e., we assume that the immune system is always able to generate at least one  $CD8^+$  clone specific for non-self. The parameters used in the model are given in Table I.

## Model results

When the virus is introduced into the system, the viral load and infected cells increase very rapidly causing a dramatic drop in the target cell population. The system oscillates for approximately 1 year and becomes stable. The effector cellular response is triggered at a slower rate and increases until day 10 to a stable steady state. The increase in effector response leads to a decrease in infected cells and viral loads, which results in an increase in the healthy cell population. The disease shows a pseudo-chronic state until day 1,200, i.e., the virus and infected cells are not cleared but remain in a stable steady state with a constant amount of effector cells and a depleted pool of target cells reaching about less than 10% of the initial value (Figure 5a).

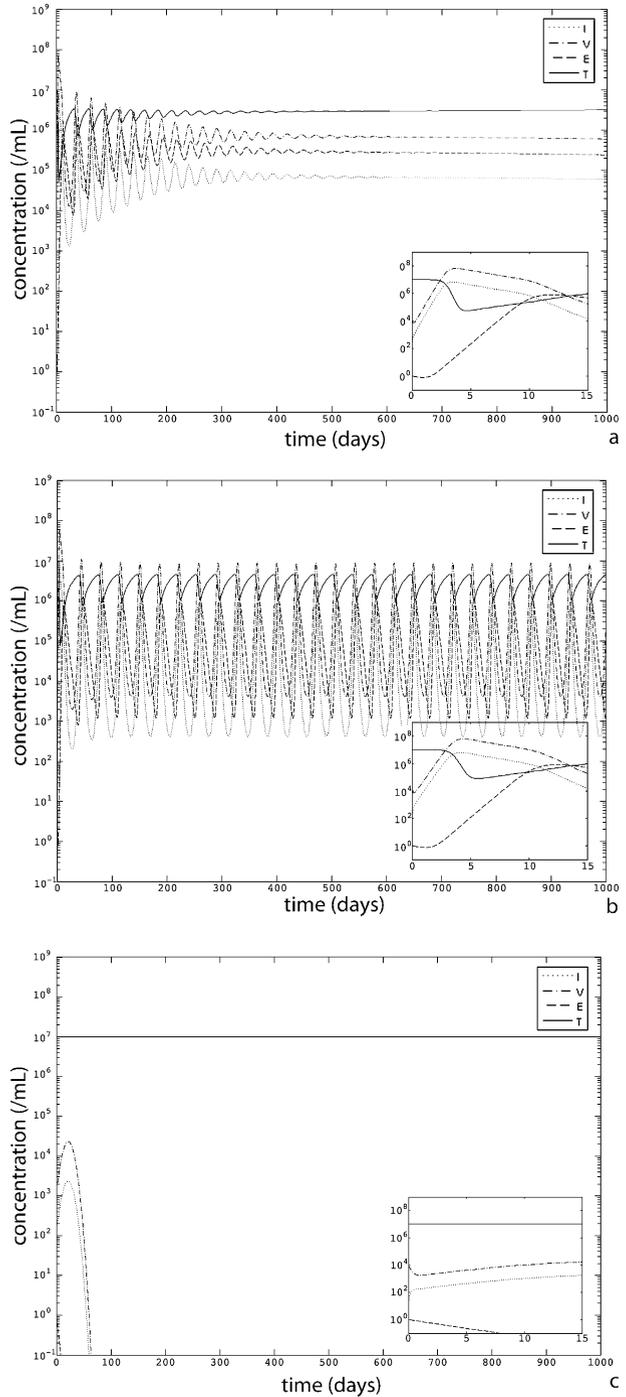
To investigate further the importance of the efficiency of infectivity, we decrease the  $f$  parameter ( $f = 0.8$ , Figure 5b). The system oscillates between two attractors and never stabilizes. While the system now approaches a limit cycle instead of a steady state, on average the steady-state values of the variables remain similar to when  $f = 1$ . The first attractor corresponds to a state where clearance of the virus might be possible, and the second one is a chronic state. Indeed, Figure 5c shows that by decreasing  $f$  by one order of magnitude ( $f = 0.1$ ), the virus and the infected cells pool are cleared, and the system goes back to a “healthy state” with target cells at their initial values in the absence of virus, infected or effector cells. Clearly, viruses with a low value for  $f$  cannot expand, and are cleared rapidly by the immune system. The oscillatory state may be an artifact of the model: such large oscillations are not likely in the HIV infections, although some “viral blips” are

**Table I** Parameter values used in the model

Parameter	Value	Unit	Description	Source
$\beta$	$7 \cdot 10^9$	$\text{ml}^{-1} \text{day}^{-1}$	Rate of cells/virus particles collision	[31]
$k$	1	$\text{cell day}^{-1}$	Lysing rate of infected $CD4^+$ cells by $CD8^+$ cells	Estimated
$\delta_I$	0.4	$\text{day}^{-1}$	Death rate of infected $CD4^+$ cells	[31, 32]
$C$	4	$\text{day}^{-1}$	Decay rate of free virus particles	[31, 32]
$\alpha$	2	$\text{day}^{-1}$	Maximum proliferation rate of $CD8^+$ cells	Estimated
$\delta_E$	0.3	$\text{day}^{-1}$	Death rate of $CD8^+$ cells	[33]
$\sigma$	$2 \cdot 10^5$	$\text{day}^{-1}$	Source of target cells	Estimated
$\delta_T$	0.02	$\text{day}^{-1}$	Death rate of target cells	Estimated <sup>a</sup> [34]
$b$	100	Virion nb	Virus particles produced by one dying $CD4^+$ cells	[31]
$f$	1	N/A	Contamination efficiency for virus / healthy cell interaction	[31]
$k_m$	$10^3$	cell	Michaelis-Menten constant for triggering of effector cells	Estimated <sup>a</sup>
$h$	$10^5$	cell	Michaelis-Menten constant for infection of target cells	Estimated <sup>a</sup>

<sup>a</sup>Personal communication with R. de Boer.

**Figure 5** Results from a simulation using the parameters in Table I. Target cells ( $T$ , solid lines), effector cells ( $E$ , dashed lines), infected cells ( $I$ , dotted line) and virus particles ( $V$ , dash-dot lines) concentration (log scale) are shown as a function of time since infection. (a) The results show that the virus and the infected cells are not cleared from the system and coexist at a pseudo-steady state. (b) Simulation performed with  $f = 0.8$ . The virus is not cleared and the system oscillates between two attractors. (c) The virus is cleared immediately and the system returns to a normal state, with only target cells.



observed during therapy. As such we will consider that Figure 5b is just a representation of a transition between two biologically relevant phenomenon, chronic disease and clearance. When  $f$  is large enough, the dynamics of acute immune response remains unchanged until day 20.

Other parameters can result in a similar behavior,  $b$  for example, which is the number of viral particles released when a cell dies. This is not surprising, as on average,  $b$  and  $f$  affect the system in a similar fashion, as they both affect the infected cells dynamics.

The number of target cells during the acute phase reaches a local minimum equal to  $c/(f\beta b)$ , therefore, the change in those parameters will have an effect on the number of target cells infected during the acute phase.  $\sigma/\delta_T$  gives the steady state of target cells and this turnover rate has an impact on the system. A high turnover rate of the target cells has a beneficial effect on the HIV population. Indeed, a quickly replenished pool of target cells is a good way for the virus to infect more cells. On the other hand, a slow turnover readily limits the progression of HIV to a single acute phase, where target cells become limiting while the immune response is able to clear the infected cells.

### Extension of the Model with Viral Mutations

An infected individual does not express only one copy of the virus, but rather a number of HIV variants compared to the wild type. New mutants arise at a rate proportional to the amount of newly infected cells. When a virus penetrates into a cell, the retrotranscriptase (RT) translates the viral RNA into DNA. This process is prone to error because the RT lacks proofreading activity. The consensus is that there is one base error each time an HIV genome is transcribed. The immune system is challenged by these constantly moving targets, and adapts accordingly. There is a wealth of genomics data on HIV. However, the link between theoretical systems biology and its data-driven counterpart, bioinformatics, has seldom been focused on by the scientific community.

We have extended our model to include the mutation process. Every mutation of the virus leads to the addition of equations (1), so that for each mutant there is an ODE for the viral load, the infected cells and the corresponding immune response. The total amount of uninfected cells is calculated by only one equation, which includes an infection term with all the mutant viruses of the system. We do not take the entire genome/proteome of the virus into account but only focus on the 9-amino-acid-long region (9mer) containing one known dominant CTL epitope (SLYNTVATL). This epitope is found in the individuals expressing a common human MHC molecule, HLA-A\*0201. We assume that viruses containing a 'SLYNTVATL' sequence are the most fit, i.e., the wild type with  $f=1$ . To get the fitness of a 9mer sequence, we compare the similarity of the mutant with the wild type using values extracted from a BLOSUM matrix [35]. BLOSUM is a general amino acid substitution scoring matrix used for alignment of protein sequences. We generate a  $9 \times 20$  matrix (Table II) with each column representing a position in the 9mer, where the rows correspond to substitution scores for that specific amino acid coming directly from a BLOSUM matrix. For a 9mer, the addition of the values in this matrix for the 9 amino acids gives its score. By comparing the score of the wild-type epitope to the score of any mutated sequence, we calculate the fitness of every mutant.

The formula used to calculate the score of a given peptide ( $m = A_1A_2...A_9$ ) is:

$$S_m = \sum_{i=1}^9 B[A_i, i],$$

where  $B$  is the matrix provided in Table II.

In order to get the fitness value  $f_m$  for the mutant, the ratio between the score of the wild type peptide and the mutant peptide is calculated. As the score of a peptide can be negative, we need to transpose the space of the score values into a positive space, since we want  $f \in [0, 1]$ . We do so by adding a value “ $p$ ”, equal to the opposite sum of the minimum value in each column of the sequence specific matrix. In the case of the BLOSUM62 matrix,  $p = 28$ . Thus,  $f_m = \frac{S_m + p}{S_w + p}$ , where  $S_w$  is the score for SLYNTVATL.

As a mutation at the DNA level is not necessarily linked to a change at the protein level, which in turn is not necessarily linked to a change in the epitope presented on the MHC, a pseudo-arbitrary value for the mutation rate was chosen to get some functionally working mutations in the system. The process of mutation is discrete, and its occurrence is governed by the time-steps used to solve the system of equations. In the model, different genomes differ only at their region corresponding to the immunodominant epitope (SLYNTVATL in the wild type) and the fitness of a mutant is evaluated only at this 9mer. We assume that any combination of amino acids resulting from the mutation process leads to a functional virus, albeit with a lower fitness. This is obviously a simplification of reality. While different components of the model will keep the same parameters with regards to a standard immune response, such as the death rate of immune cells or their activation rate, the extended model attributes different infection efficiencies of the various virus variants. As main hypothesis, it

**Table II** Sequence specific scoring matrix used to determine the fitness value of any 9mer compared to the reference peptide SLYNTVATL

	S	L	Y	N	T	V	A	T	L
A	1.00	-1.00	-2.00	-2.00	0.00	0.00	4.00	0.00	-1.00
R	-1.00	-2.00	-2.00	0.00	-1.00	-3.00	-1.00	-1.00	-2.00
N	1.00	-3.00	-2.00	6.00	0.00	-3.00	-2.00	0.00	-3.00
D	0.00	-4.00	-3.00	1.00	-1.00	-3.00	-2.00	-1.00	-4.00
C	-1.00	-1.00	-2.00	-3.00	-1.00	-1.00	0.00	-1.00	-1.00
Q	0.00	-2.00	-1.00	0.00	-1.00	-2.00	-1.00	-1.00	-2.00
E	0.00	-3.00	-2.00	0.00	-1.00	-2.00	-1.00	-1.00	-3.00
G	0.00	-4.00	-3.00	0.00	-2.00	-3.00	0.00	-2.00	-4.00
H	-1.00	-3.00	2.00	1.00	-2.00	-3.00	-2.00	-2.00	-3.00
I	-2.00	2.00	-1.00	-3.00	-1.00	3.00	-1.00	-1.00	2.00
L	-2.00	4.00	-1.00	-3.00	-1.00	1.00	-1.00	-1.00	4.00
K	0.00	-2.00	-2.00	0.00	-1.00	-2.00	-1.00	-1.00	-2.00
M	-1.00	2.00	-1.00	-2.00	-1.00	1.00	-1.00	-1.00	2.00
F	-2.00	0.00	3.00	-3.00	-2.00	-1.00	-2.00	-2.00	0.00
P	-1.00	-3.00	-3.00	-2.00	-1.00	-2.00	-1.00	-1.00	-3.00
S	4.00	-2.00	-2.00	1.00	1.00	-2.00	1.00	1.00	-2.00
T	1.00	-1.00	-2.00	0.00	5.00	0.00	0.00	5.00	-1.00
W	-3.00	-2.00	2.00	-4.00	-2.00	-3.00	-3.00	-2.00	-2.00
Y	-2.00	-1.00	7.00	-2.00	-2.00	-1.00	-2.00	-2.00	-1.00
V	-2.00	1.00	-1.00	-3.00	0.00	4.00	0.00	0.00	1.00

Data extracted from a BLOSUM62 matrix. The score of the major epitope is 43.

is assumed that the efficiency of infection of the major epitope is optimal, i.e., a few virus particles are enough to infect a cell, while peptides with a low score will be less fit, and therefore will get a lower efficiency. To reflect this, the parameter  $f$  in the  $j$ th set of equations is replaced with the fitness of the mutant,  $f_j$ , that is calculated as described above.

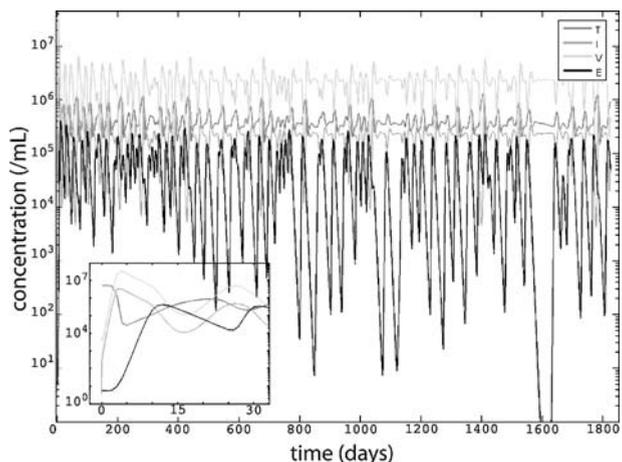
Each simulation is run with a maximum number of different genomes that can coexist. The simulations always start with one genome (SLYNTVATL). As mutations appear, one amino acid is changed in one of the viral genomes present in the system, and the newly formed virus gets a starting value equal to the number of mutants generated by the infected cell (c.f. Table 1). This extended system of equations is given below in Eq. (2).

$$\begin{cases} \frac{dI_j}{dt} = f_j \beta V_j T - \delta_I I_j - \frac{k E_j I_j}{h + E_j + I_j} \\ \frac{dV_j}{dt} = b \delta_I I_j - c V_j \\ \frac{dE_j}{dt} = \frac{\alpha E_j I_j}{k m + E_j + I_j} - \delta_E E_j \\ \frac{dT}{dt} = \sigma - \beta T \sum_j V_j f_j - \delta_T T \end{cases} \quad (2) \quad \forall j \in [1, n]$$

### Model with Mutations

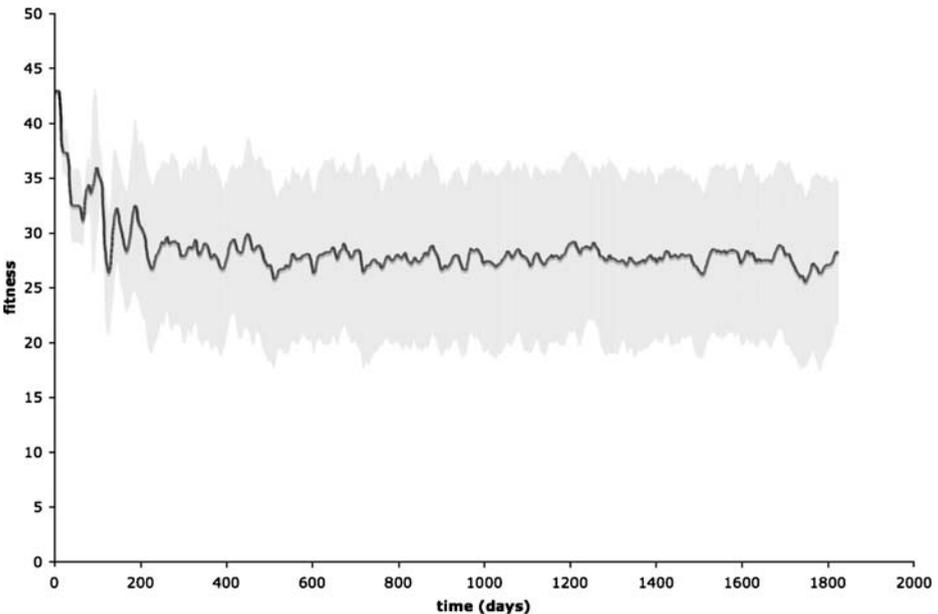
The general results of the extended model are not drastically different from the first one. As the system (2) starts with the same parameters and variable values as before (1), it is logical to find comparable results. The mutation of the virus allows the system to generate mutants that evolve together with the wild type to a steady state-like situation, where the virus species and their specific immune responses coexist (Figure 6). A basic feature of the model is the rapid growth of the wild type virus, followed by an acute response against it, leading to a drop in the viral load. As the virus mutates and new variants arise, the viral load is increased again, followed by a corresponding immune response. The extended model

**Figure 6** Simulation results for the extended model, averaged values of 100 simulations.



allows us to follow which mutants arise and when. The mutants arise at different time points, which is why the viral load shows many peaks that represent the maximal expansion of all the mutants. The disease remains in a chronic state, but it is possible to see a general decrease in the viral population fitness, as shown in Figure 7. Although the signal remains noisy in Figure 6, a pseudo-steady state is observed. At this point, compared to the simple model simulation with the fittest virus ( $f = 1$ ), the values of the variables of the model differ. The mutants are able to infect more target cells; this leads to a higher viral load and a lower value for target cells. Surprisingly, the effector cells counts are much lower, especially when one integrates the total immune response over time, showing that the immune system cannot cope efficiently with all the mutants. Moreover, the number of mutants that remain at significant levels at the end of the simulations never reaches more than 10 to 20 distinct populations, i.e., HIV persists as a quasi-species in the patient as observed clinically. Only the most fit mutants can survive, and the mean fitness score for the viral population drops to 28 (the wild type has a fitness of 43). Thus, viruses with lower fitness values are not able to infect enough cells to persist in the system and are cleared (as in Figure 5c).

The extended model only accounts for one parameter related to a partial sequence of one protein of the HIV virus. Still, by choosing the right parameter to tweak, namely  $f$ , the model is able to account for main characteristics of the HIV virus. A direct extension of the model would be to relate more parameters to real sequences in the same fashion as has been done above for  $f$ .



**Figure 7** General normalized viral fitness (normalized 9mer score values) over time in days, averaged for one hundred simulation runs (solid line, mean value, gray area, standard error). The fitness is maximal in the very beginning and equal to the fitness of the major epitope, but drops quickly because of the appearance of less fit viruses. As the disease progresses, the fitness remains stable at a lower value compared to SLYNTVATL.

## Discussion

Developing a “virtual immune system” is a very challenging task, where several levels of organizations have to be taken into account. The modeller’s task is to reduce the number of variables or rules they include in their system via assumptions that reduce the complexity of the model, making it easier to analyze. So far, the traditional approach to the problem has been to subdivide the immune system into defined components, from the molecular [17] level to the population level [12], and to analyze those reduced problems. With the advance in models at all levels of organization, it is now possible to merge the different components developed and work at the systems level, thus characterizing the interactions that govern the global behavior of the immune system. This is a major task to complete.

The development of computers allowed repetitive computational work to be performed easily and faster. In the modelling field, this improvement led to the development of more complex systems, taking into account more biological data. A good example of this evolution is the one that cellular automata (CA) have undergone since the 1940s. It started with the game of life, with its simple two-state system and a few neighbouring rules and developed into complex agent-based models (ABM), based on the same principles, but including more states and phenomena, compared to the initial system [36]. Despite the potential complexity that can be reached, automata described in the literature do not contain much detail at the molecular level, such as protein sequences and structures. Full-scale computational modelling of the entire immune system is complicated because it relies on integration of many different components. However, many of these components may be much simpler models that describe how immune systems - step by step - deal with pathogenic organisms. These simpler models may successfully represent important sub-components of the immune system, provided that the biological mechanisms controlling the various model steps are sufficiently well understood. The goal is to develop simulation models of the human immune system, and use these to obtain a better understanding of immune system-related diseases. In the past, most models have been built on systems of differential equations, which, by nature tend to have a top-bottom approach, and have the advantage of enabling an accurate study of every parameter used in an analytical way. However, this approach suffers from generalization, since a population of heterogeneous agents is reduced to a single continuous variable [36].

With current immune system models, main trends are captured, but low-impact or rare phenomenon cannot possibly be represented nor studied. We argue that it is possible to create hybrid models including a third dimension, the genomic dimension (Figure 3), where, in the end, real genomic sequences will be used together with mathematical models to achieve a high spatio-temporal resolution. We demonstrate this by developing the very first model of HIV dynamics mixing mathematical modelling and bioinformatics data.

The mathematical model developed in this article accounts for the cell biology of the immune system challenged by an HIV infection and is an extension of a widely accepted mathematical model of HIV dynamics. The immune response is concatenated into the cytotoxic T cell variable, and is of course not a detailed representation of the biology, as one could argue that humoral and cellular response play distinct but overlapping roles during a viral infection. The model is still able to account for biologically relevant phenomena, i.e., chronic infection and clearance, but not the AIDS phase, where the viral load increases by several orders of magnitude and the CD4<sup>+</sup> cell count drops drastically. The AIDS phase is presumably the result of the appearance of HIV mutants with high fitness which are able to escape CTL control. To investigate this idea, we developed the second model presented in

this paper, where the fitness of the mutant strains is taken into account. Indeed, the hyper-variability of HIV is the key factor for chronic disease, meaning that even if the immune system is able to produce a response against one particular HIV and clear it, the constant appearance of new mutants forces the immune system to develop new responses again and again. Several theories have been suggested, namely, the erosion of the immune system [37], or the switch between CCR5-tropic and CXCR4-tropic virus [38] that are known to have preferential affinity for naïve and memory T cells.

The use of a sequence-specific matrix to score a part of a protein in relation to a reaction rate casts a novel insight into immune system modelling. The motivations behind this idea are that, first, epitopes that are easily recognized by the immune system trigger a higher response against them, while more cryptic epitopes tend to stay hidden from the immune system and continue to infect new target cells, albeit with a lower efficiency; second, the automation inherent to such a method can produce more robust models, since a wealth of bioinformatics data is available. While the developed models remain simple, with regard to biological complexity, they prove to mimic results shown in the literature [7]. The interaction of several hundreds of coupled differential equations linked to HIV genomic data allows a scientific study of the genetic dynamics of the system. The major achievement here is the use of bioinformatics data to generate different parameter values for a mathematical model. As stated earlier, we use a BLOSUM62 matrix to generate the fitness parameters. Other BLOSUM matrices exist, based on the percentage of similarities between the sequences in the data set used to produce them. Typically, the BLOSUM matrices are used for alignment of biological sequences, as they measure the similarities between protein sequences with regard to their biological function. While epitopes are usually conserved regions of proteins, that have a specific biological role in the protein, in the case of HIV, the use of BLOSUM matrices may not be so straightforward.

Eventually, the future of HIV modelling lies in the development of more detailed models taking into account many essential biological activities and relating them to protein sequences associated with them.

This is a first attempt into a more extensive integration of bioinformatics and systems biology: we plan to develop in the future a model that would allow complete genomic sequences to be used together with new or existing theoretical models.

## Applications of More Realistic Models of the Immune System

Vaccines are, presumably, the most effective medical technology ever invented. Vaccines are currently only available for approximately one tenth of the microorganisms known to be harmful to humans. New vaccines are lacking for the three main infectious killers in the world: HIV, malaria, and tuberculosis, as well as against diseases such as influenza and pox, which may evolve to be threats either naturally or by intentional development/spread by terrorists. The use of whole live or killed microorganisms as vaccines is in many cases unfeasible because of concerns about safety, efficacy and ease of production. Much focus has therefore been on vaccines composed of parts of a microorganism (subunit vaccines) or genes coding for parts of the microorganism (genetic vaccines). A major challenge when making such vaccines is to select the parts of the microbe to include in the vaccine. The genome-specific models will in particular be used to identify some of the epitope-containing regions of foreign organisms.

Genetic factors contribute to the development of allergy, but environmental factors may also be important. Allergic reactions are caused by a special class of antibodies called

immunoglobulin E (IgE) antibodies. IgE responses are, under normal physiological conditions, protective, especially in response to parasitic worms, which are prevalent in less developed countries. In the industrialized countries, however, due to better hygienic conditions, IgE responses occur almost entirely against allergens. Almost half of the inhabitants of North America and Europe have allergies to one or more common environmental antigens. The reason for the increase in allergy in the Western countries remains largely unknown, but one of the theories is that it is caused by the higher standards of hygienic conditions [39]. We plan to use our model to study the development of an immune system with or without parasitic worms, to find new ways of intervening that may hinder the development of allergy.

Autoimmune diseases occur when the body's defence against non-self is accidentally directed against self-antigens. Genetics are thought to play a role in the development of autoimmune diseases, as a link to the MHC has been reported in many autoimmune diseases such as rheumatoid arthritis or type I diabetes. Environmental factors seem to play a role as well in the susceptibility of these different diseases. In order to limit the effects of these uncontrollable factors, computer models ensure a relatively good and scientific solution to this problem; despite the fact that models cannot be compared directly to their biological counterpart since they rely on assumptions, they do allow a better understanding of biological pathways and underlying mechanisms regulating a given disease [40].

**Acknowledgements** The authors wish to thank Rob de Boer for insightful and inspiring discussions about the mathematical model. This work is supported by a grant for the European Immunogrid project (<http://www.immunogrid.org/>) and the European Union through the Network of Excellence BioSim, contract number LSHB-CT-2004-001537.

## References

1. Holmdahl, R., Bockermann, R., Backlund, J., Yamada, H.: The molecular pathogenesis of collagen-induced arthritis in mice - a model for rheumatoid arthritis. *Ageing Res. Rev.* **1**, 135–147 (2002)
2. Burroughs, N.J., de Boer, R.J., Kesmir, C.: Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics* **56**, 311–320 (2004)
3. Kohler, B., Puzone, R., Seiden, P.E., Celada, F.: A systematic approach to vaccine complexity using an automaton model of the cellular and humoral immune system. I. Viral characteristics and polarized responses. *Vaccine* **19**, 862–876 (2000)
4. Seiden, P.E., Celada, F.: A model for simulating cognate recognition and response in the immune system. *J. Theor. Biol.* **158**, 329–357 (1992)
5. Pappalardo, F., Lollini, P.L., Castiglione, F., Motta, S.: Modeling and simulation of cancer immunoprevention vaccine. *Bioinformatics* **21**, 2891–2897 (2005)
6. Ferguson, N.M., deWolf, F., Ghani, A.C., Fraser, C., Donnelly, C.A., Reiss, P., Lange, J.M., Danner, S. A., Garnett, G.P., Goudsmit, J., Anderson, R.M.: Antigen-driven CD4+ T cell and HIV-1 dynamics: residual viral replication under highly active antiretroviral therapy. *Proc. Natl. Acad. Sci. USA* **96**, 15167–15172 (1999)
7. De Boer, R.J., Homann, D., Perelson, A.S.: Different dynamics of CD4+ and CD8+ T cell responses during and after acute lymphocytic choriomeningitis virus infection. *J. Immunol.* **171**, 3928–3935 (2003)
8. De Boer, R.J., Mohri, H., Ho, D.D., Perelson, A.S.: Estimating average cellular turnover from 5-bromo-2'-deoxyuridine (BrdU) measurements. *Proc. Biol. Sci.* **270**, 849–858 (2003)
9. Fraser, C., Ferguson, N.M., De Wolf, F., Ghani, A.C., Garnett, G.P., Anderson, R.M.: Antigen-driven T-cell turnover. *J. Theor. Biol.* **219**, 177–192 (2002)
10. Takaku, T., Ohyashiki, J.H., Zhang, Y., Ohyashiki, K.: Estimating immunoregulatory gene networks in human herpesvirus type 6-infected T cells. *Biochem. Biophys. Res. Commun.* **336**, 469–477 (2005)
11. Busic, V., Petrovsky, N.: Immunoinformatics - the new kid in town. *Novartis Found Symp.* **254**, 3–13 (2003); discussion 13–22, 98–101, 250–102

12. Kirschner, D., Marino, S.: *Mycobacterium tuberculosis* as viewed through a computer. *Trends Microbiol.* **13**, 206–211 (2005)
13. Romanyukha, A.A., Rudnev, S.G., Sidorov, I.A.: Energy cost of infection burden: An approach to understanding the dynamics of host–pathogen interactions. *J. Theor. Biol.* (2005)
14. Lund, O., Lund, O.S., Gram, G., Nielsen, S.D., Schonning, K., Nielsen, J.O., Hansen, J.E., Mosekilde, E.: Gene therapy of T helper cells in HIV infection: mathematical model of the criteria for clinical effect. *Bull. Math Biol.* **59**, 725–745 (1997)
15. Scherer, A., Noest, A., de Boer, R.J.: Activation-threshold tuning in an affinity model for the T-cell repertoire. *Proc. Biol. Sci.* **271**, 609–616 (2004)
16. Warrender, C., Forrest, S., Segel, L.: Homeostasis of peripheral immune effectors. *Bull. Math Biol.* **66**, 1493–1514 (2004)
17. Jansson, A., Barnes, E., Klenerman, P., Harlen, M., Sorensen, P., Davis, S.J., Nilsson, P.: A theoretical framework for quantitative analysis of the molecular basis of costimulation. *J. Immunol.* **175**, 1575–1585 (2005)
18. McKeithan, T.W.: Kinetic proofreading in T-cell receptor signal transduction. *Proc. Natl. Acad. Sci. USA* **92**, 5042–5046 (1995)
19. Snyder, J.T., Belyakov, I.M., Dzutsev, A., Lemonnier, F., Berzofsky, J.A.: Protection against lethal vaccinia virus challenge in HLA-A2 transgenic mice by immunization with a single CD8+ T-cell peptide epitope of vaccinia and variola viruses. *J. Virol.* **78**, 7052–7060 (2004)
20. Lund, O., Nielsen, M., Lundegaard, C., Kesmir, C., Brunak, S.: *Immunological bioinformatics*. MIT Press, Cambridge, Massachusetts (2005)
21. Nielsen, M., Lundegaard, C., Worning, P., Hvid, C.S., Lamberth, K., Buus, S., Brunak, S., Lund, O.: Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* **20**, 1388–1397 (2004)
22. Sturmiolo, T., Bono, E., Ding, J., Radrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M.P., Sinigaglia, F., Hammer, J.: Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.* **17**, 555–561 (1999)
23. Larsen, M.V., Lundegaard, C., Lamberth, K., Buus, S., Brunak, S., Lund, O., Nielsen, M.: An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* **35**, 2295–2303 (2005)
24. Hopp, T.P., Woods, K.R.: A computer program for predicting protein antigenic determinants. *Mol. Immunol.* **20**, 483–489 (1983)
25. Odorico, M., Pellequer, J.L.: BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J. Mol. Recognit.* **16**, 20–22 (2003)
26. Perelson, A.S.: Modelling viral and immune system dynamics. *Nat. Rev. Immunol.* **2**, 28–36 (2002)
27. Tsomides, T.J., Aldovini, A., Johnson, R.P., Walker, B.D., Young, R.A., Eisen, H.N.: Naturally processed viral peptides recognized by cytotoxic T lymphocytes on cells chronically infected by human immunodeficiency virus type 1. *J. Exp. Med.* **180**, 1283–1293 (1994)
28. Borghans, J.A., de Boer, R.J., Segel, L.A.: Extending the quasi-steady state approximation by changing variables. *Bull. Math Biol.* **58**, 43–63 (1996)
29. Tzafirri, A.R., Edelman, E.R.: The total quasi-steady-state approximation is valid for reversible enzyme kinetics. *J. Theor. Biol.* **226**, 303–313 (2004)
30. Tzafirri, A.R., Edelman, E.R.: On the validity of the quasi-steady state approximation of bimolecular reactions in solution. *J. Theor. Biol.* **233**, 343–350 (2005)
31. Strain, M.C., Richman, D.D., Wong, J.K., Levine, H.: Spatiotemporal dynamics of HIV propagation. *J. Theor. Biol.* **218**, 85–96 (2002)
32. Dixit, N.M., Perelson, A.S.: HIV dynamics with multiple infections of target cells. *Proc. Natl. Acad. Sci. USA* **102**, 8198–8203 (2005)
33. von Boehmer, H., Hafén, K.: The life span of naive alpha/beta T cells in secondary lymphoid organs. *J. Exp. Med.* **177**, 891–896 (1993)
34. Zhang, Z.Q., Notermans, D.W., Sedgewick, G., Cavert, W., Wietgreffe, S., Zupancic, M., Gebhard, K., Henry, K., Boies, L., Chen, Z., Jenkins, M., Mills, R., McDade, H., Goodwin, C., Schuwirth, C.M., Danner, S.A., Haase, A.T.: Kinetics of CD4+ T cell repopulation of lymphoid tissues after treatment of HIV-1 infection. *Proc. Natl. Acad. Sci. USA* **95**, 1154–1159 (1998)
35. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992)
36. Breckling, B.: Individual-based modelling: potentials and limitations. *ScientificWorldJournal* **2**, 1044–1062 (2002)

37. Galvani, A.P.: The role of mutation accumulation in HIV progression. *Proc. Biol. Sci.* **272**, 1851–1858 (2005)
38. Ribeiro, R.M., Hazenberg, M.D., Perelson, A.S., Davenport, M.P.: Naive and memory cell turnover as drivers of CCR5-to-CXCR4 tropism switch in human immunodeficiency virus type 1: implications for therapy. *J. Virol.* **80**, 802–809 (2006)
39. Borchers, A.T., Keen, C.L., Gershwin, M.E.: Hope for the hygiene hypothesis: when the dirt hits the fan. *J. Asthma* **42**, 225–247 (2005)
40. Gonzalez, P.P., Cardenas, M., Camacho, D., Franyuti, A., Rosas, O., Lagunez-Otero, J.: Cellulat: an agent-based intracellular signalling model. *Biosystems* **68**, 171–185 (2003)
41. RAC (Recombinant DNA Advisory Committee), appendix B, [http://www.od.nih.gov/oba/vac/guidelines\\_02/APPENDIX\\_b](http://www.od.nih.gov/oba/vac/guidelines_02/APPENDIX_b)