Computational simulations of the immune system for personalized medicine: State of the art and challenges

F. Pappalardo
P. Zhang
M. Halling-Brown
K. Basford
A. Scalia, et al.
Computational simulations of the immune system for personalized medicine: state of the art and challenges

Francesco Pappalardo\textsuperscript{1}, Ping Zhang\textsuperscript{2}, Mark-Halling Brown\textsuperscript{3}, Kaye Basford\textsuperscript{2}, Antonio Scalia\textsuperscript{1}, Adrian Shepherd\textsuperscript{3}, David Moss\textsuperscript{3}, Santo Motta\textsuperscript{1} and Vladimir Brusic\textsuperscript{2,4}

\textsuperscript{1}Department of Mathematics and Computer Science, University of Catania, Italy;\nEmails: \{francesco, Scalia, motta\}@dmi.unict.it
\textsuperscript{2}School of Land, Crop and Food Sciences, The University of Queensland, Australia;\nEmails: \{p.zhang2, k.e.basford\}@uq.edu.au
\textsuperscript{3}School of Crystallography, Birkbeck College, Malet Street, London;\nEmails: \{m.halling-brown, a.shepherd, d.moss\}@mail.cryst.bbk.ac.uk
\textsuperscript{4}Cancer Vaccine Center, Dana-Farber Cancer Institute;\nemail: vladimir_brusic@dfci.harvard.edu

Abstract

The human immune system is a complex and adaptive learning system which operates at multiple levels: molecules, cells, organs, organisms, and groups of organisms. Each human individual has tens of immune system organs and some 10\textsuperscript{12} immune cells that belong to multiple types and subtypes. These cells produce molecules that act as initiators, regulators, and effectors of the immune function. In addition, each individual has a unique immune system and will respond differently to immune challenges. The complexity of the immune system therefore arises from its hierarchical and combinatorial properties. Immunologic research, both basic and applied, needs to deal with this complexity. We increasingly use mathematical modeling and computational simulation in the study of the immune system and immune responses.

Simulating the immune system is a challenge that involves multidisciplinary efforts. In immunology, simulators are used to get answers to a variety of questions, including understanding the general behavior of the immune system, the course of disease, effects of treatment, analysis of cellular and molecular interactions, and simulation of laboratory experiments. Models and simulations have proven useful in studying the roles of single constituents and simple interactions, planning of experiments, testing theoretical assumptions, and even highly abstract tasks such as
suggesting theory modifications. A variety of mathematical modeling techniques have been
developed for and applied to the immune system simulation during the last fifty years and, more
recently, some have been implemented and made publicly available as software tools.

The ImmunoGrid project integrates molecular and system level models of the immune system and
processes for *in silico* studies of the immune function. The ImmunoGrid simulator is implemented
as a web-based application. It uses Grid technologies, aiming to enable computational simulation of
the immune system at the natural scale, perform a large number of simulated experiments, capture
the diversity of the immune system between individuals, and suggest studies tailored to the
individual genetic make-up. The architecture of the simulator has a modular design enabling easy
implementation of extensions and modifications. This paper reviews the current developments in
the field of computational simulation of the immune system. We use the ImmunoGrid project as a
showcase for integration of immune system models. ImmunoGrid is web accessible and it enables
users to perform simulations for both educational and research purposes.

**Key words:** computational modeling, immune system simulation, ImmunoGrid, grid computing
1 Introduction

Pharmacogenomics studies the effects of genetic variation on patient responses to therapies. It includes genomics profiling of patients and screening their responses to various treatments. Applications of pharmacogenomics include evaluation of safety and efficacy of treatment as well as optimization of therapies and therapeutic regimens. Pharmacogenomics has research (e.g. target validation, lead optimization, and toxicity screening) as well as clinical (selection of optimal dosing, schedule, and duration of therapies) applications [Katz, D.A. 2006; Sioud, M. and Melien, A. 2007]. The common goal is the rational selection and design of diagnostics and therapies that best suit the genetic make-up of both the individual and the disease and thus maximize the benefit to the patient.

Pharmacogenomics is becoming increasingly important in immunology, for the development of new generation vaccines [Rappuoli, R. 2007], immunotherapies [Gu, J. et al. 2005; Yan, L. and Beckman, R.A. 2005], and in transplantation [Dickinson, A.M. et al. 2007]. An ideal vaccine should be cheap to produce, stable to store and distribute, applicable to the global population, and produce long-lasting immunity. Unfortunately, variation in host and pathogens makes such universal vaccines a remote possibility for the majority of pathogens [Brusic, V. and August, J.T. 2004] and cancers [Brusic, V. et al. 2007]. This difficulty is exemplified in the studies of immune responses to vaccines, where a broad spectrum of responses to the same vaccine can be observed across the vaccinated population. The emerging field rational design of vaccines using high-throughput technologies and bioinformatics helps predict the efficacy and adverse effects of vaccines [Poland, G.A. et al. 2007]. Advanced applications, such as optimization of vaccination protocols [Lollini, P.L. et al. 2006] or modeling immune responses in atherogenesis [Pappalardo, F. et al. 2008], require the use of advanced bioinformatic applications, including computational
modeling of the immune system. Computational models are mathematical models implemented in
the form of computer programs that can be used to study complex systems by computational
simulations. Computational simulation technology has provided researchers with new insights into
the variation of immune responses to vaccines within the human population [Kurstak, E. 2003;
Plotkin, S. 2005].

Two principal directions for computational modeling and simulation of the immune system (and its
parts) concentrate on either theoretical or applied aspects of immunology. The theoretical modeling
focuses on understanding the biology of the immune system, while applied modeling assists in the
design of vaccines and immunotherapies [Motta, S. and Brusic, V. 2004; Louzoun, Y. 2007;
Lundegaard, C. et al. 2007]. The models of the immune system need to accurately encode the actual
behavior and functional effects, need to capture both the combinatorial and hierarchical complexity,
and should be technically feasible (Mitcha et al, 2008). A computational model, that is a mapping
from a real-world domain to a computational/mathematical domain, has several essential properties.
A well-designed model must: a) be relevant, b) capture the essential properties of the modeled
entity (e.g. concept, structure, system, process, or relationship), c) have a conceptual framework for
reasoning about the modeled entity, d) be implementable as a computer program, and e) be
extensible so that additional real properties can be added into the model framework.

The usefulness of advanced computational approaches to the study of the immune system has been
demonstrated in many studies [Petrovsky, N. and Brusic, V. 2002; Morel, P.A. et al. 2006;
Castiglione, F. and Piccoli, B. 2007; Cohen, I.R. 2007]. Due to the great complexity of the human
immune system and computational limitations of combinatorial entities, earlier immune system
models were restricted to an idealized small scale, which has several orders of magnitude less
components than the number of molecules and cells within a natural human immune system.
Recent developments, namely the ImmunoGrid (http://www.immunogrid.eu) simulator, enable modeling of the immune system at the natural scale by using the power of Grid technologies.

This paper discusses the current advances in simulation models of the immune system, the extension of capabilities due to Grid technology that enables the performance of very large computational tasks, and the challenges therein.

2 Computational modeling

A significant body of work has been produced in the area of immune system modeling. Most of these modeling systems focus on specific functions or phenomena. From the methodology point of view, we categorize the models into ab initio (theoretical) models that are based on established laws and knowledge of the domain known as first principles, and statistical models that are derived from a limited number of observations or data points where the description of the modeled system is lacking. Theoretical models describe the real immune environment and translate the description of the immune system and its function into a set of mathematical descriptors. Figure 1a shows a simple diagram of theoretical immune system simulation models. The accuracy of the system is achieved by tuning parameters of the system. Statistical models are built using statistics or machine learning approaches, using known data points and any one of a range of underlying modeling structures, to predict the unknown phenomena. They are pattern recognition processes. Figure 1b shows a generalized diagram of the development of data mining models using statistical approaches. The performance of the model on the unknown data depends on the quality and quantity of the data and the training algorithm.
2.1 Theoretical Models

Many theoretical simulation models of the immune systems have been developed in the past. These models start from either one of the two biological existing theories of the immune system, namely the clonal selection theory [Forsdyke, D.R. 1995] and, to the smaller extent, the idotypic network theory [Jerne, N.K. 1974]. They can be classified as continuous models or discrete models according to the underlying mathematical framework. Continuous models use the framework of
both immunological theories [Perelson, A.S. and Weisbuch, G. 1997], while discrete models mostly use clonal selection theory. From the computational implementation point of view, we can categorize the simulation methodologies into mathematical modeling by differential calculus and by agent-based models such as Cellular Automata (CA), which have become the two main streams of theoretical modeling of the immune system.

2.1.1 Differential Equation-Based Models

A mathematical model uses mathematical language, such as formulation to describe a system. An early idea of such modeling of the immune system is from Sercarz and Coons [1962]. The idea was modified by Šterzl [1967] and led to the construction of mathematical models of the immune response [Bell, G.I. 1970; Jílek, M. and Sterzl, J. 1970]. Based on these mathematical models many contemporary models were built using differential equations.

Differential equation based models of the immune system have been recognized as efficient models producing good simulation results. In fact, models using ordinary or partial differential equations have become the great majority of mathematical models of the immune system. They have been very popular, and a wide range of immunological phenomena have successfully been simulated using differential equation based models. Most of these models focus on one specific phenomenon observed in immunology. The behavior of the model is then compared to data obtained by \textit{in vivo} or \textit{in vitro} experiments, and if in agreement, the model can be used for description of immune function as well as for prediction of its behavior.

Differential equation based models are still used widely for simulating many different phenomena. For example, Chaturvedi and Insana [1997] proposed a mathematical model to describe the dynamics of chronic progressive renal disease. With a few parameters, the model presented a
distribution of glomerular responses that is similar to experimental observations. That model included both deterministic and stochastic elements of the disease mechanisms through a set of three differential equations and their initial conditions. Funk et al. [1998], presented a mathematical model with eight nonlinear differential equations to simulate the kinetics of B-cell activation and the virus-neutralizing immunoglobulin response in the spleen of mice after infection with vesicular stomatitis virus (VSV). Rundell and his team [1998] developed a mathematical model focused on simulation of the humoral immune response to haemophilus influenzae Type b (Hib) infection. They represented the dominant dynamics of the humoral immune response to the bacteria, Hib, with an 11th order nonlinear differential delay model.

Analytical techniques allow modelers to define the system behavior and their associated parameters and initial conditions. Bocharov [1998] presented a mathematical model formulated by a nonlinear delay-differential equation, in which the parameters were estimated by assimilating with model data for different levels of infection. The model was used for investigating the effect of variations in virus and CTL response parameters on LCMV infection outcome and suggesting predictions for experimental studies, in particular the phenotype of LCMV-WE infection in C57BL/6 as a function of initial virus doses.

Modifications, such as adding time delays or age-structured partial differential equations [Antia, R. et al. 2003] make these models more faithful representations of biological reality. In Antia’s work, two models describing the CD8 responses by a partial differential equation or a system of ordinary differential equations were reported. Both of these models showed high consistency with the experimental observations.
The parameters of the differential equations-based models are determined from experimental data, but at this stage the dynamics of the system is not carefully validated by experiments. Kleinstein and Singh [2001] analyzed a model of germinal center dynamics and affinity maturation where dividing centroblasts undergo periodic rounds of affinity-based selection as non-dividing centrocytes, as proposed by Oprea and Perelson (OP) [1997]. They developed a discrete/stochastic implementation of the OP models to compare the system dynamics with data collected from individual germinal centers. The discrete/stochastic implementation is based on a fixed-increment time advance framework [Law, A.M. and Kelton, W.D. 1991]. The differential equation implementation was translated into the discrete/stochastic framework by grouping of related terms in the differential equations into independent processes and labeling processes that span multiple equations. That extension improved several aspects of immune system modeling in which the OP model was deficient.

Differential equation models capture the general behaviour of system and the global parameters of the model can be easily adjusted. However, these models are limited to a specific observable function, and they do not capture the complex interplay of various factors influencing observed behaviour. This limitation prevents their application for studies of the immune system as a complex system. A more detailed review of earlier applications of differential equations for modeling the immune system can be found in [Perelson, A.S. and Weisbuch, G. 1997].

2.1.2 Cellular Automata

Cellular Automata (CA) models are fully discretized dynamical systems that are well suited for computer simulations of biological systems [Stauffer, D. and Pandey, R. 1992]. The general framework for capturing dynamic behaviour in the models can be precisely tuned to mimic the behavior of the real system [Manneville, P. et al. 1989].
The initial idea of using a discrete automaton model in immunology was proposed by Kaufman et al. [1985]. In this model, various cellular populations and interactions were represented by Boolean values. Several subsequent discrete models such as [Weisbuch, G. and Atlan, H. 1988; Cohen, I.R. and Atlan, H. 1989; Chowdhury, D. et al. 1990; Sieburg, H.B. 1992] were all developed based on this idea. In [Neumann, A.U. 1989], the Boolean structure was extended to a cellular automaton which adapted the concept of the ‘shape space’ [Wolfram, S. 1986].

Many CA models [Stauffer, D. and Sahimi, M. 1993; Stauffer, D. and Sahimi, M. 1994; Smith, D.J. et al. 1997] of the immune system are based on the work of de Boer et al. [1992] which was derived from Jerne’s immune network theory. De Boer’s network model was composed of an infinite number of B cell and corresponding antibody populations of different shapes that may be ordered in a finite-dimensional shape-space. A lattice mapping was employed in their two-dimensional model, whose rules were derived from a simplified version of the underlying differential equations via a logarithmic transformation of variables. A novel feature of that lattice model was that a cell interacts directly with neighboring cells. In a lattice cellular automaton, the state of a cell at time $t+1$ depends only on the state of its immediate neighbors at time $t$, or any of the earlier time steps $(t-n)$.

The automata that model complex systems or complex behavior are known as agent-based models or multi-agent systems. As in CA, there are rules governing interactive behavior and the agents “operate” in an environment of some sort. Many of the recent computational models of the immune system are agent based systems. The majority of current applications of simulation models are inspired by the early immunological CA proposed by Celada and Seiden [1992] which attempted to build a general immune system simulator. Celada-Seiden model has provided useful insights into the dynamics of the immune system responses [Castiglione, F. et al. 1997].
Since there is no simple mathematical description of Agent Based Models (ABM), there is no real attempt to mathematically formalize simulators based on this approach. The agent-based paradigm can be considered as an extension of the spin concept [Mez’ard, M. et al. 1988] where the widespread use of computers enabled the addition of more realistic descriptions of the internal state-space of the spins (the entities or agents) and the mathematical description of relatively complex interaction rules. There are only a few basic rules in the initial Celada-Seiden cellular automaton, while later developments extended and redefined the set of rules, resulting in improved generalization ability of the immune system cellular automaton. Their simulation program called IMMSIM has been used to investigate a number of immune system phenomena, for example affinity maturation and hypermutation of the humoral immune response [Celada, F. and Seiden, P.E. 1996]. The embedded rules can be extremely complex, as illustrated by the large number and complex expression of the Boolean rules developed in the Kaufman model [Kaufman, M. et al. 1985]. More detailed reviews of CA and ABM can be found in [Perelson, A.S. and Weisbuch, G. 1997] and [Forrest, S. and Beauchemin, C. 2007]. Here we list several well-known examples of agent based models.

**ImmSim and C-ImmSim**

IMMSIM was written in APL2 with the IBMAPL2 runtime environment and thus was not easily portable. A parallelized implementation of IMMSIM called ParImm [Castiglione, F. et al. 1997; Bernaschi, M. et al. 1999] was developed using C language and later extended it to the current version called C-ImmSim. Kleinstein and Seiden modified IMMSIM to C++ version [Kleinstein, S.H. and Seiden, P.E. 2000]
C-IMMSIM can simulate both the innate and adaptive immune responses. The adaptive immune responses act when the innate fails to recognize or eliminate the threat such as pathogen or tumor. Adaptive immune responses are based on the immunity acquired by learning and they utilize immune memory. C-IMMSIM contains agents representing the macrophages, dendritic cells, T-cells and B-cells, antibodies, antigens, and some cytokines. It models mainly adaptive immune responses to the antigens, both cellular and humoral.

C-IMMSIM uses a bit-string polyclonal lattice model. The “bit-string” refers to the representation of molecules and the specificity, “polyclonal” indicates that multiple clones of different specificity of lymphocytes are represented (as opposed to the monoclonal models where only a single population of genetically identical lymphocytes is represented), and “lattice” means that discrete lattice is used to represent the discrete space.

In the two-dimensional version, a single lymph node of a vertebrate animal is mapped onto a two-dimensional $L \times L$ hexagonal (or triangular) lattice (six neighbors), with periodic boundary conditions [Bernaschi, M. and Castiglione, F. 2001]. Physical proximity is modeled through the concept of lattice-site. All interactions among cells and molecules are computed within a lattice-site in a single time step, so that there is no correlation between entities residing on different sites at a fixed time. The diffusion of entities at the end of each time step introduces correlations and it models physical spreading of the interacting molecules in the lymph-node.

**SimTriplex**

Triplex is an immunopreventive HER-2/neu breast cancer vaccine [Lollini, P.L. et al. 2005], which combines the specific target antigen, p185(HER-2/neu) with two non-antigen-specific adjuvants: IL-12 and allogeneic major histocompatibility complex (MHC) class I molecules. Four vaccine
administration schedules (early, late, very late and chronic) have been tested on HER-2/neu transgenic mice, and the chronic schedule showed that it provides complete, long-term protection from mammary carcinoma.

SimTriplex [Pappalardo, F. et al. 2005] is a specialized cellular automaton in modeling mammary carcinoma, Triplex vaccine and the immune system competition based on a modification of Celada-Seiden framework [Celada, F. and Seiden, P. 1992]. It mimics the behavior of immune cells at the cellular level in both vaccinated and in naive mice. The simulation results have shown excellent agreement with in vivo experiments both for the time of formation of solid tumor and for the role of antibody responses in controlling tumor growth [Motta, S. et al. 2005; Pappalardo, F. et al. 2006]. Once SimTriplex was validated, a new component was developed to optimize the Triplex vaccine protocol [Lollini, P.-L. et al. 2006; Pappalardo, F. et al. 2006; Pennisi, M. et al. 2008] and help immunologists find the best timing for administration of the vaccine.

To represent the tumor growth and the vaccine action, the following entities and interactions were modelled in SimTriplex: i) the cancer cells (CC) that encode their tumor associated antigens (TAA); they interact with antibodies (Ab), T-cells (TC) and natural killer cells (NK), ii) the vaccine cells (VC) that include TAAs, IL-12 and allogenic major histocompatibility molecules class I (MHCI); VC interact like CC, with Ab, TC and NK, but the affinity function is modified by the presence of two adjuvants,, iii) NK cells which, in the model, express only CD16a (FcγRIIIA) receptor for ADCC mediated cytotoxicity [Lanier, 2008]; they interact with Ab coated VC and CC. When two entities, which may interact, lie in the same lattice site then they interact following a probabilistic law. All entities which have capacity to mutually interact and are in the same site have a positive interaction.
SIS

SIS-I and SIS-II are two versions of SIS (Synthetic Immune System) that use cellular automata-based immune system models. Currently they are available from the web site http://www.cig.salk.edu. The comparison with C-ImmSim was stated in the paper published by the developers [Mata, J. and Cohn, M. 2007]: C-ImmSim appeals to experimentalists who desire to manipulate a wide variety of parameters, whereas SIS was more theoretically driven and was likely to appeal most to those with non-trivial minimalist inclinations. The authors also pointed out that the rules in ImmSim are hardcoded and only the parameter values were changeable, while SIS allows both the parameter values and the rules to be changed.

SIMISYS

Recently, Kalita et al. [2006] reported a software system called SIMISYS that simulates aspects of the human immune system based on cellular automata. The current version, SIMISYS 0.3, models the innate and adaptive components of the immune system. To our knowledge, SIMISYS is not yet publicly available.

2.1.3 Hybrid models

Cellular automata enable the analysis of selected details of complex immune interactions while differential equation models enable quantitative estimation of global behavior of the immune system [Hu, R. and Ruan, X. 2003]. Realizing the strengths and weaknesses of differential equation-based models and cellular automaton models, researchers have developed hybrid systems to improve modeling performance. SIMMUNE and Cycells are representatives of hybrid models that combine differential equations and cellular automata.
SIMMUNE

SIMMUNE is a two-level immune system simulator combining molecular and cellular level entities [Meier-Schellersheim, M. and Mack, G. 1999]. Molecules such as cytokines are defined as continuous quantities, and their dynamics are modeled using differential equations. The immune system cells are modeled as discrete computational agents. Thus, SIMMUNE is a hybrid of continuous and agent-based modeling (ABM) techniques ([http://www.simmune.org](http://www.simmune.org)). SIMMUNE has no attributes that are unique to immune system cells and it is applicable to the simulation of any living cell system. The software suite allows molecular reactions to be defined directly at the level of interactions between molecular binding sites, using simple graphical representations of molecules and molecular complexes. It allows users to define molecule types as well as to specify the number and properties of their binding and/or enzymatic sites [Meier-Schellersheim, M. 2001; Meier-Schellersheim, M. et al. 2006]. The mechanisms describing cellular behavior in SIMMUNE provide a proof of principle for a new kind of simulation analysis in immunology.

There are similarities between SIMMUNE and SIS [Mata, J. and Cohn, M. 2007]. Both SIMMUNE and SIS allow the initial concentrations of agents or cells to be specified by the users. Both programs are based on rules that are defined by the users, and both also use a three-dimensional (3D) data structure to attempt to model the real world as closely as possible. There is a major difference between these two models: SIMMUNE models molecular interactions and cellular responses to stimuli while SIS, in contrast, encodes cellular responses to stimuli as cellular states.

CyCells

CyCells is a similar modeling framework [Warrender, C.E. 2004; Warrender, C.E. 2005] as SIMMUNE. It is also a hybrid simulator in which molecular concentrations are represented
continuously and cells are represented as discrete entities. CyCells is implemented on a cubic grid. In CyCells, the model is initialized by specifying the simulation geometry, molecular concentrations, and numbers of each cell type (e.g. B cells and macrophages). It has been used to model early infection dynamics of Mycobacterium tuberculosis (Mtb) bacteria, and has reproduced some qualitative results observed in mice in vivo [Warrender, C.E. et al. 2006]. CyCells can be used to model a wide variety of multicellular systems, and the source code is available on https://sourceforge.net/projects/cycells/.

2.2 Statistical Models

Statistical techniques have been broadly used to extract information from the large sets of data generated by experimentation and clinical observation in immunology. Statistical modeling applications in immunology often focus on data mining and pattern recognition [Brusic, V. and Zeleznikow, J. 1999]. Data mining uses machine learning, statistical and visualization techniques to discover the patterns from the large amount of data. Data mining has become popular in bioinformatics during the last two decades.

Adaptive immunity is a hallmark of the diversity of the immune system. One of the main areas in which data mining techniques have used to study the adaptive immune system is identification of MHC-binding peptides and T-cell epitopes which are targets for vaccine development and drug discovery [Muzzi, A. et al. 2007]. All nucleated cells in an organism degrade internal proteins into short peptides and present them for recognition by T-cells of the immune system. In addition, professional antigen presenting cells (macrophages, B lymphocytes and dendritic cells) have the ability to capture extracellular antigens, degrade them into short peptides and to present them on the cell surface for recognition by T cells [Rammensee, H.G. et al. 1993; Cresswell, P. 1994]. Major histocompatibility complex (MHC) is involved in this process because it binds short peptides for
displaying them to T cells. MHC class I bound peptides have the ability to prime cytotoxic T cells that eventually kill infected, mutated or transplanted target cells. MHC class II bound peptides deal mainly with regulation of immune response. Peptides recognized by T cells are termed T-cell epitopes. Different set of MHC genes leads to a different capacity of the immune system to deal with a particular antigen. MHC genes show extensive polymorphism with more 1600 variants of human leukocyte antigen (HLA, human MHC) class I and more than 700 variants of HLA class II molecules have been characterized and named to date (June 2008) [Robinson, J. and Marsh, S.G. 2007].

Moreover in a single human individual there are more than $10^{11}$ different antibody specificities and more than $10^{11}$ T cell receptors. This high variability along with the contribution of genetic processes of recombination, class switching, and somatic hypermutation contribute to an extraordinary large number of different immune system products. The number of different possible combinations of HLA molecules in human population is up to $10^{12}$ given the number of HLA and that each individual human expresses 3 to 6 different HLA class I and perhaps a dosen different HLA class II molecules.

The number of 9-amino acid long linear targets of immune response is greater than $10^{11}$. MHC class I molecules bind 8-11 amino acids long peptides of which majority are 9-mers. MHC class II molecules bind longer peptides most being 10-25 amino acids long. These peptides are targets of immune recognition in the context of presentation by HLA molecules. Due to the combinatorial nature of immune system interactions, the number of experiments required for conducting systematic studies, such as T-cell epitope mapping is prohibitively large. Current approaches focus on a set of selected experiments to generate initial data sets, which are then used to train predictive models. Systematic scanning can be performed in silico and is followed by validation experiments.
New data are then used to refine predictive models making MHC binding and T-cell epitope mapping a showcase of synergy achieved by combining computational modeling and experimentation.

Most epitopes recognized by antibodies or B cells are three-dimensional surface features of an antigen molecule; these features fit precisely into the matching surface of antibodies providing tight binding. Some B-cell epitopes are determined by linear amino acid sequence (the primary structure) of a protein. Predictions of B-cell epitopes is a complex modeling task because it involves matching complex 3D shapes of antibodies with target epitopes on the surface of antigens. The attempts to model and predict B-cell epitopes are still in exploratory stage and practical applications are expected to emerge some time in the future, however a body of work in this field has been developing. The methods for prediction of B-cell epitopes have been reviewed in [Lundegaard, C. et al. 2007]).

A number of computational methods have been developed to facilitate the identification of MHC binding peptides. Matrix-based techniques [Peters, B. et al. 2003; Bui. H.H et al. 2005; Moutaftsi, M. et al. 2006], Artificial Neural Networks [Buus, S. et al. 2003; Nielsen, M. et al. 2003] and Support Vector Machines [Bozic, I. et al. 2005; You, L. et al. 2007] have been reported as the most efficient, predictive and frequently used methods in the developed models for identification of MHC binding peptides and epitopes. More detailed reviews of the data mining predictive algorithms are available [Brusic, V. et al. 2004; Brusic, V. and Flower, D.R. 2004; Peters, B. et al. 2006]. The Center for Biological Sequence Analysis (CBS) at the Technical University of Denmark developed a set of prediction servers including some for immunological features. NetMHC (www.cbs.dtu.dk/servers/NetMHC) [Buus, S. et al. 2003; Nielsen, M. et al. 2003; Nielsen, M. et al. 2004] is one of the servers, which predicts binding of peptides to a number of different HLA alleles.
using artificial neural networks (ANNs) and weight matrices. Similarly, a large number of prediction models are available at the Immune Epitope Database and Analysis Resource (IEDB, www.immuneepitope.org/) [Peters, B. et al. 2005]. Altogether, more than thirty prediction servers have been developed and are accessible via the Internet. The analysis of thirty different prediction servers of MHC class I peptide binding has been reported [Lin, H.H. et al. 2008] with a list of the servers. For some variants of human MHC, notably HLA-A*0201, computational modeling has reached the level where it is as accurate as experimentation. However for a vast majority of MHC class I molecules, predictive accuracy is still lower than experimental identification [Lin, H.H. et al. 2008]. More than 20 prediction servers of peptide binding to MHC class II are also accessible via the Internet. But their prediction accuracy is much lower than the accuracy of MHC class I predictors [Gowthaman, U. and Agrewala, J.N. 2008].

Recently, combined predictors that include several consecutive steps of antigen processing and presentation have been developed as a new generation. Those immune predictors, combine predictors of HLA binding with predictors of peptide binding to transporter associated with antigen processing (TAP) and predictors of proteasomal cleavage (reviewed in [Lundegaard, C. et al. 2007]). The predictor made by the combination of HLA and TAP–binding prediction is useful because in some cases improves the prediction of T-cell epitopes [Peters, B. 2003]. However, a weakness of these predictors is that they eliminate TAP-independent peptides from further analysis, namely those that are processed by lysosomal [Demirel, O. et al. 2007], endosomal Kurotaki, T. et al. 2007] and vacuolar [Demirel, O. et al. 2007] pathways. Methods for prediction of proteasomal cleavage are of a relatively low accuracy [Peters, B. 2003], and are not yet properly validated [Lundegaard, C. et al. 2007]. Currently HLA-binding remains the most useful computational tool for mapping of HLA ligands and T cell epitopes [Lin, H.H. et al. 2008].
The main objective of the ImmunoGrid project is the development of a human immune system simulator using common computational platform to help development of vaccines and immunotherapies. The unique component of this project is the integration of the simulations of immune system processes at molecular, cellular and organ levels (system level) for various applications. The project is a web-based implementation of the Virtual Human Immune System using Grid technologies. It adopts a modular structure that enables easy extensions and modifications. Figure 2 shows the current overall system architecture. Depending on the user’s request, it can be used as an antigen analyzer (T-cell and B-cell), antigen processing simulator, infection simulator, cancer simulator, allergy simulator, vaccine designer, etc.

3 ImmunoGrid -- an integrated immune system simulator
3.1 System Level Modeling

System level modeling produces cellular-level models and software prototypes of the immune system simulator. The models describe processes at cellular and organ levels, including models of immune responses to viruses, models which analyze MHC diversity and its effects on host-pathogen interactions, B cell maturation, and dynamic models that can simulate both cellular and humoral immune responses.

Currently, C-ImmSim and SimTriplex constitute the main part of the system level models. C-ImmSim simulates the immune responses to bacteria and viruses (e.g. HIV-1), and SimTriplex is used to model tumor growth and responses to immunization. The results produced by C-ImmSim and SimTriplex are presented on the web in graphical and text file formats for educational and research users. Figure 3 shows an example web interface of the system level modeling application in ImmuoGrid, which simulates the cancer cells growth when using the early schedule of the triplex vaccine on the mouse. The interface was created for educational users based on the result produced by SimTriplex. Users can select different vaccine schedules from the buttons on the left side of the screen, and the simulation result is shown on the right side of the window simultaneously. A drop down list allows users to have the option of showing the progression of cancer cells, T-cells, B-cells, Dendritic cells, Macrophages and Antibodys when the vaccine is administered.
3.2 Molecular Level Modeling

Molecular level models simulate the immune system’s sequence-based processing and discriminate self and non-self proteins at the molecular level. Currently, these models focus on predictions of T-cell and B-cell epitopes using various modeling algorithms, such as Artificial Neural Networks (ANN), Matrix, Hidden Makov Model (HMM), Support Vector Machine (SVM) and other statistical techniques.

CBS prediction tools make up the core of the molecular level models. Both pre-calculated models and online job submission models are supplied from the web for different users. Figure 4 shows an example interface when it is used for epitope prediction. Users can select the specific proteins and
MHC alleles from the lists and the predicted MHC binders or hotspots will be displayed on the web instantly.

### 3.3 Grid Technologies and Integration

An integral part of the ImmunoGrid project is the use of Grid technologies. The motivation behind implementing a Grid solution for the ImmunoGrid project is twofold: the increases in the complexity and scale of the simulator require more computational power; and the exploration of the parameter space to investigate the varying immune responses within a population requires an extremely large number of simulations to be run. To satisfy both these demands, the provision of large clusters and of connected supercomputers, a considerable number of individual nodes are needed to perform the required simulations.

A main goal of the ImmunoGrid project was that the solution should support all major existing Grid middleware and platforms and that there should be no re-implementation of existing methods. Another was to allow access to a wide variety of modeling resources through a common

![Figure 4: Example ImmunoGrid web interface for epitope prediction](image-url)
mechanism and common interface. These rules led to the adoption and integration of three different Grid solutions: Uniform Interface to Computing Resources (UNICORE, http://www.unicore.eu), Application Hosting Environment (http://www.realitygrid.org/AHE) and web services. These solutions offer the combined coverage of adoptions while there is relative ease of implementation.

The ImmunoGrid project has developed a solution which is made up of a single web-based interface which is coupled with a job broker. This job broker/launcher accesses either UNICORE or AHE via command line tools and web services via standard SOAP protocols. The main benefit of this setup is the ability to launch jobs on different Grid platforms, local resources and web services, all through a single interface. This approach effectively hides the complexity and diversity of accessing these resources from the end user. The only real interaction that the user has with Grid technologies will be the optional requirement of a grid certificate.

Both UNICORE and AHE use X.509 certificates for security and authorisation which allowed us to uniform the security through the web-based interface. This requirement means that any user accessing the system would require a X.509 certificate. NGS (UK, National Grid Service, http://www.grid-support.ac.uk/) and DEISA (Distributed European Infrastructure for Supercomputing Applications, http://www.deisa.org) currently supply facilities and resources to ImmunoGrid. Although a certificate is required to access NGS and DEISA resources, it is anticipated that the majority of project participants would not have access to one. It is for this very reason that access to local group resources is provided. If a user has a Grid certificate, then the interface allows them access to the NGS and DEISA resources that are available. If the user does not have a Grid certificate, then a self issued certificate is automatically generated which allows access to the local group resources. Many of the characteristics of ImmunoGrid are shared by other biological projects: the involvement of multiple international partners (each bringing their own
computational resources to the project); the need to run large numbers of computations, both large and small; and the need to provide an easy-to-use interface for a non-technical user base. The grid requirements of the ImmunoGrid project are as follows:

- To enable the most complex single simulations to be run, requiring access to a large cluster or to a supercomputer.
- To enable large sets of immune system simulations and epitope predictions to be carried out, both to explore the parameter space of the simulator and to investigate the effects of a given clinical scenario on multiple individuals.
- To support smaller-scale simulations, including runs of the ImmunoGrid educational simulators, for which standards workstations are sufficient.

One of the major challenges for projects which demand a large amount of computational power from a Grid is deciding which solution to adopt and the implementation of this solution. The computational demands of systems biology models vary considerably which implies that the type of resources that a project may require will vary considerably also. A large number of Grid solutions are available and these vary with respect to their suitability for rapid development. UNICORE and AHE are described in more detail.

Arguably the most important ingredient in our framework is the role played by “upper middleware” (AHE and DESHL), as it hides much of the complexity of the Grid both from those developing a new Grid infrastructure and from the users of that infrastructure. Without upper middleware, the whole enterprise would be prohibitively complicated and time-consuming for most scientific institutions or consortiums to undertake. AHE and DESHL play a key role in the deployment of software to different computational resources and in the management of jobs. Both AHE and DESHL provide a command-line interface via which a job can be launched, its progress monitored,
and its output (both intermediate and final) retrieved. There are, however, some important differences.

The Application Hosting Environment (AHE) is a lightweight environment designed to run unmodified applications on diverse, distributed Grid resources. An explicit design goal of AHE is to hide the underlying complexity (of the Grid middleware, of the host environment and of how executables are set up) from the end user. Currently this is achieved using GridSAM (http://gridsam.sourceforge.net/2.0.1/index.html) (Grid Job Submission and Monitoring Web Service), but a UNICORE (Uniform Interface to Computing Resources [Almond, J. and Snelling, D. 1999] plugin is at an advanced stage of development and has been used by the Coveney group [Zasada, S.J. et al. 2007]. The UNICORE plugin is not yet available as part of the AHE package. After the initial deployment of AHE, a simple Job Submission Description Language (JSDL) file must be produced for each combination of software and resource that the AHE will have access to. This is the only manual intervention required, and it need only be done once for each software/resource combination. Thereafter, the AHE provides a list of available resources upon which the software has been installed. Jobs can then be run by simply selecting resources from the list (see Resource brokering and job launching); there is no need to access any resources directly. The AHE is easy to install as part of the OMII stack.

DESHL is somewhat less flexible than the AHE, but it does provide essential mechanisms for accessing European supercomputers via UNICORE. In the context of the ImmunoGrid project, such resources are available via DEISA and at CINECA (a member of the ImmunoGrid Consortium). Setting up access to a supercomputer via DESHL is somewhat less transparent than adding a resource using the AHE, as scripts need to be written that manage access to a named user space.
There are two ways that a local resource can be accessed via our system – using the AHE, or as a Web Service. The fundamental difference between these two approaches is that a given resource is made available for any application using the AHE route, whereas the Web Services approach makes a specific application available. The practical differences between these two approaches are relatively minor.

To access a new local resource using the AHE, an Apache Tomcat server needs to be installed on the local machine together with an instance of GridSAM (an Open Source job submission and monitoring Web Service). These are automatically installed and configured (without recourse to special administrative rights) when the OMII stack is installed on the machine. The final step is simply to edit the configuration of the GridSAM instance so that it points to the locally installed software that we wish to use. A local resource can also be accessed using web services. This can be achieved either by ‘wrapping’ the software in a simple web service shell or pointing a web service execution at the local software.

Jobs are launched by a simple job launcher. This executes the appropriate launch command for a given job and resource (this is different for the AHE, DESHL, and Web Services). The launcher also records the details of the job both on the server’s file system and in a local database, and executes the appropriate command line script corresponding to the resource that is selected. The state of the job is stored in the local database along with any information required to uniquely identify that job. This allows the appropriate scripts to be called when the user requests the job state to be refreshed.

The Web interface to our infrastructure provides the end user with simple mechanisms for uploading, launching and monitoring the progress of jobs, as well as for retrieving and displaying
results. The interface comprises PHP Web pages, with AJAX and DHTML used to give them a modern look and feel. The interface is loosely coupled to the resource broker/job launcher and has been developed in a modular way to facilitate it being adapted to handle new applications.

4 Discussion

Various computational models of the immune system, as described in the previous sections, have been made publicly available. Mathematical models and Cellular Automata are mostly used for cellular level simulations, while a range of statistical modeling applications are suitable for the analysis of sequences at molecular level of the immune system.

Grid computing technology brings the possibility of simulating the immune system at the natural scale. ImmunoGrid project integrates the immune system simulation on both molecular and cellular levels using Grid computing. In our opinion, a Grid solution is only as good as the interface provided to the users. For this reason, we have concentrated on providing an easy-to-use web interface which allows users to access every single resource on the Grid, regardless of the middleware used. This is achieved by the separation of the web interface from the job broker and the Grid implementations. Grid computing provides access to practically use unlimited computational resources, enabling the analysis of the immune system at the natural scale and at the same time addressing the combinatorial complexity of the immune system. These technologies also herald a new era in immunology where the immune system of an individual can be profiled in detail and large number of simulated experiments can be performed. This is a necessary step towards tailoring vaccines and immunotherapies to individuals.
5 Acknowledgments

This work was supported under the EC contract FP6-2004-IST-4, No.028069 (ImmunoGrid).
References


