ImmunoGrid - The Virtual Human Immune System Project

Andrew EMERSON [a,1] and Elda ROSSI [a]
[a] *High Performance Systems Division, CINECA, 40033 Casalecchio di Reno, Italy*

Abstract. ImmunoGrid is a 3 year project funded by the European Union which began in February 2006 and establishes an infrastructure for the simulation of the immune system that integrates processes at molecular, cellular and organ levels. It is designed for applications that support clinical outcomes such as the design of vaccines, immunotherapies and optimization of immunization protocols. The first phase of the project concentrated on improving and extending current models of the immune system. We are now entering the second phase which will design and implement a human immune system simulator. Since the new models are orders of magnitude more complex than the previous ones, grid technologies will be essential in providing the necessary computer infrastructure. The final phase of the project will validate the simulator with pre-clinical trials using mouse models.

Keywords. Immunology, Grid computing,

Introduction

The immune system is a complex and adaptive learning system which has evolved to defend the individual against foreign invaders. It has multiple levels (molecular, cellular, organ and tissue, organism, and organism-to-organism) and is also combinatorial in nature with a large number of products; there are typically more than 10^{15} antibodies and 10^{12} immune system cell clones in a single individual. The function of the immune system depends on both the genetic composition and the previous exposure, i.e. the experience of the organism.

Immune intervention, such as vaccination, is the most effective method for the control of disease and the greatest achievements include eradication of smallpox, near-elimination of polio, and savings of some 170 million person-years. Vaccination has been used in the control of over two dozen diseases by the 50 or so successful vaccines which have been developed to date. These vaccines largely protect against infectious diseases, although recent vaccine developments offer great hope for treatment for a broader range of diseases. Large-scale studies of the immune system, also known as immunomics, is the key factor driving the current wave in vaccine development. These include genomics and proteomics, analysis of the diversity of pathogens or complexity of the human immune system, high-throughput screening or immunoinformatic tools for the management

[1] Corresponding Author: A. Emerson, High Performance Systems Division, CINECA, 40033 Casalecchio di Reno, Italy; E-mail: a.emerson@cineca.it

and analysis of vast quantities of data. Computational models are becoming increasingly important in immunomics:

1. Experimental approaches are expensive and it is impossible to perform systematic experimental studies of immune processes in humans.
2. Because of ethical issues, there are stringent limitations as to what experiments can be performed in humans.
3. Computational approaches can compensate for the limitations of allowable studies.

Computational modeling is ideal for complementing experimental research and clinical observations. The usefulness of computational approaches to the study of immune system has been demonstrated, but computational models that encode the natural-size immune system have not been developed because of the past limitations of computational infrastructures. Grid technologies now enable the complexity of the computational models to be matched with that of the natural human immune system. Such computational models have applications not only in theoretical immunology (better understanding of immune function), but also in clinical medicine in diagnostics and the development of vaccines and therapies for cancer and infectious disease.

This paper is intended as a high level view of the ImmunoGrid project which is currently at the half way stage. We start by giving an overview of the computer models employed in our simulations of the immune system and the computational resources required. We then propose a grid infrastructure and discuss the possible technologies which we will be used in its implementation.

Specific applications of the ImmunoGrid Simulator will appear in subsequent publications.

1. Immune system models

Methods for modeling the behaviour of the immune system generally are designed for different length scales, ranging from interactions at the molecular level (e.g. peptide binding to receptors) to system approaches which operate at cellular and organ dimensions. A novel aspect of the ImmunoGrid project is that it integrates both levels, resulting in the most realistic models yet realised.

The simulator at the heart of the ImmunoGrid project originally derives from the ImmSimm program, a cellular automata model of the human immune response [1]. Variants and parametrisations of this program have been successfully employed in the modeling of, for example, the response to generic bacteria and viruses (C-ImmSimm, [2]) and the Triplex cancer immunoprotection vaccine (SimTriplex, [3]). The stochastic cellular automaton represents many different interacting entities (antigens, antibodies, T cells, B cells, peptides, etc) and hence is capable of modeling many aspects of the immune response. Although very successful, the computational requirements impose limits on the sample sizes and complexity that can be handled in current simulations. For example, very often two dimensional grids are used, typically representing only a few mm^3 of tissue. ImmunoGrid has already started to employ much larger three dimensional models which accurately describe the spatial distribution of the entities. When complete the ImmunoGrid simulator will be able to deal with systems which approach the complexities of real immune systems.

2. Computational requirements for the Simulator

Since the first major objective of ImmunoGrid was the research and validation of improved immune system models, the implementation of the Simulator has only just started. However, from an initial analysis of the new models the following requirements have emerged:

1. **Processing power.** The most important requisite is for high computational processing power. Extending the SimTriplex model (used for the simulation of the cancer immune response), for example, to larger samples based on three dimensional lattices requires powerful computing facilities. Fortunately, parallelisation of the SimTriplex program is possible and efficient and the implementation of this program with the MPI message passing library on the supercomputing clusters available to the ImmunoGrid consortium has has already begun. For the C-ImmSimm program (designed for the response to bacteria and viruses), the other main component of the Simulator, a high-throughput approach is more appropriate. Here the simulation involves a genetic algorithm which runs C-ImmSim for every genotype in the simulation space representing the set of therapies possible until convergence. Thus, a very large number (e.g. many tens of thousands) of simulations are launched, with each run requiring typically about 10min CPU time for the current models. The program is therefore "embarrassingly parallel" and is readily implemented on a parallel computer.
2. **Data federation.** Although less critical, access to data resources, either in the form of primary databases or data produced by the simulations themselves, will need to be made available to all the components of the Simulator. Technologies for performing data federation between geographically separated computing resources, such as SRB [4] or OGSA-DAI[5], are being considered.
3. **Visualisation and user portals.** One goal of the project is to create a tool which will allow clinicians or health researchers to design new vaccination protocols based on the simulations performed while another important objective, strongly promoted by the EU, is to provide facilities for educational purposes. Thus, at a later stage it will be possible for non-specialist audiences such as the general public to have access to ImmunoGrid. For these reasons, user portals and visualization facilities are essential. Basic visualization tools, to show for instance the spatial distribution of the interacting entities, have already been designed, while a "ImmunoGrid user portal" forms part of the grid implementation (see next section).

As mentioned above enhanced versions of the basic simulator tools have already been written and are currently being tested. In the next section we will describe how we are integrating these and further tools to create the Simulator.

3. Grid implementation

It should be mentioned at this stage what we mean by "grid" since the term in computational science can have various meanings and interpretations. In our project, the interpretation is primarily that of a "virtual organisation", an important concept in many grid

projects and implying a single point of access for heterogeneous resources and administrative domains. This is not to suggest that the programs themselves only require modest resources but that for the moment we don't anticipate direct communication between the various grid nodes during execution. This is technically feasible by using, for example, grid or web services, but is in practice fairly complicated and also incurs some costly overheads. In addition, these technologies are still not yet mature and standards are continually evolving. It should be emphasised however that we do expect CPU processing requirements to increase rapidly during the project as models of more complex tissues and organs are simulated, but these will run on separate clusters. These models will themselves be scaled up to the whole organism level, i.e. the mouse, to facilitate validation by the pre-clinical trials (based on transgenic mice, see the next section). Towards the end of the project the Simulator will be extended to the human natural scale, which is about 1,000 times bigger than that of the mouse.

The aim of ImmunoGrid is not to develop new grid middleware so we are evaluating the grid-enabling technologies currently available to the consortium members and these are described below.

3.1. EnginFrame portal

EnginFrame[6] is a commercial product based on "agents" which manage computing resources on behalf of users, interacting with the underlying operating system, batch schedulers, grid middleware, user applications and so on. With this software it is possible to create a web-based portal which provides a single point of access to heterogeneous resources, allowing the user to submit and monitor jobs, browse the user's disk space and obtain information regarding the status of the resources. It is also possible to design specific interfaces for commonly used applications, e.g. the different versions of the simulators we are setting up in this project. But in addition to hiding the complexities of the operating system (usually some dialect of UNIX), portals built with EnginFrame have two features which make the environment very attractive for building grid infrastructures:

1. Single sign-on. When combined with a standard password authentication system (e.g based on LDAP, a common UNIX protocol) a user will need to enter username and password only once, i.e. on entry to the portal, in order to access the resources and run jobs. This may require careful mapping of the user accounts between the hosted computing platforms, but this is generally straightforward, at least at a single site.
2. Transparent access to job schedulers. This is a crucial feature because many schedulers and queuing systems are used in supercomputing centres around the world. A non exhaustive list includes PBS, Torque, LSF and LoadLeveler. Support for some schedulers is already built into the product (e.g. LSF) while for others (e.g. LoadLeveler) a plug-in needs to be written and installed.

A computing portal based on EnginFrame is currently under development at the CINECA computing centre where a trial version is being beta-tested by selected users. The CINECA portal is a good test of the technology because this centre provides four different publicly available supercomputing clusters, with different operating systems (Linux and IBM AIX) and two different batch schedulers (LSF and LoadLeveler). The unified access to the batch schedulers is expected to be popular with users who generally

find them difficult to use. In principle, there should be no problem in incorporating other systems external to CINECA into this portal or one built for ImmunoGrid with a custom interface for submitting C-ImmSimm or SimTriplex jobs.

3.2. Application Hosting Environment

The Application Hosting Environment (AHE) [7] is a sophisticated grid infrastructure currently in use at the site of the UK partner of ImmunoGrid. The AHE has been designed to provide the simplest possible service interface to a client for submitting jobs to highly complex grids. The AHE, is a lightweight web service hosting environment able to operate over multiple administrative domains. The AHE stores all the necessary information about how an application should be run on the various computational resources of a grid and provides a uniform interface to the client for running that application across those resources. The AHE interfaces with GridSAM [8], a job submission and monitoring service. GridSAM provides an abstraction of resources, such as Globus Toolkit and Condor on the resource side. Therefore the AHE interacts in a clear transparent way with complex grid architecture and submission systems, such as queuing systems, without interaction from the client. The AHE provides resource selection, application launching, workflow execution, provenance and data.

3.3. Specific applications

One or both of the above technologies will be used to build the framework for the ImmunoGrid simulator, a prototype of which will be available by the end of 2007. But, although the infrastructure and portal will be expected to cope with the day-to-day running of the ImmunoGrid, we anticipate that there will be simulations which will be beyond the computational resources normally available within the consortium. In these situations applications will be made to some large European Grid-infrastructures like DEISA[9], an e-infrastructure connecting the major supercomputing centres in Europe which provides facilities for particular problems which require substantial computing resources or the NGS, the UK National Grid Service[10]. Thanks to the use of the portal, the applications will run on these large infrastructures in a transparent way, hiding the implementation details on the different computing environments. Thus, some nodes on our grid will in fact be other grid infrastructures.

4. Simulator validation by pre-clinical trials

Validation of the ImmunoGrid simulator will be performed by one of the consortium members who has expertise in performing pre-clinical trials with transgenic mice models, such as BALBneuT which carries the HER-2/neu oncogene. The predictions of the simulator will be tested in vivo by the mice models, with healthy and pathological individuals, and will provide feedback for improving the computer models. Later there will be comparison with experimental data for human systems, where available.

5. Comments

The European Commission has identified the Virtual Physiological Human as one of the challenges of their seventh Framework Programme (FP7). ImmunoGrid will clearly have an important role in this effort and collaborations with other projects are being sought.

The project began in February 2006 and has already extended the immune system models and developed more sophisticated parametrisations. Small-scale simulations on single grid nodes are currently being run. By the end of 2007 a demonstration version of the grid-enabled simulator will be available for general use. With the final version of the simulator it will be possible to perform immunological simulations vastly more complex and detailed than those available currently.

Acknowledgments

The ImmunoGrid consortium consists of the following entities: CINECA (Italy), the University of Queensland (Australia), CNR-Rome (Italy), CNRS-Montpellier (France), Birkbeck College London (UK), Dept. of Experimental Pathology (Bologna, Italy) and the University of Catania (Italy). This project is funded as part of the European Union FP6 programme, contract no. FP6-2004-IST-028069.

References

[1] F. Celada, and P.E. Seiden: A computer model of cellular interaction in the immune system. *Immunology Today* (1992), 13:56–62.

[2] F. Castiglione, M. Bernaschi and S. Succi: Simulating the immune response on a distributed parallel computer. *Int. J. Mod. Phys. C* **8** 527–545.

[3] F. Pappalardo, P.-L Lollini, F. Castiglione and S. Motta: Modeling and simulation of cancer immunoprevention vaccine. *Bioinformatics* Vol. 21 no 12 (2005) 2891–2897.

[4] A. Rajasekar, M. Wan and R. Moore: "MySRB & SRB: Components of a Data Grid," hpdc, p. 301, *11th IEEE International Symposium on High Performance Distributed Computing* (HPDC-11 '02), (2002).

[5] The OGSA-DAI Project, http://www.ogsadai.org.uk/.

[6] For more information see http://www.enginframe.com.

[7] P. V. Coveney, S. K. Sadiq, R. Saksena, M. Thyveetil, S. J. Zasada, M. McKeown and S. Pickles, "A Lightweight Application Hosting Environment for Grid Computing", *Proceedings of the UK e-Science All Hands Meeting 2006* (2006), 217–224.

[8] See for example, http://gridsam.sourceforge.net/2.0.0/index.html.

[9] For more information about DEISA see http://www.deisa.org

[10] See, for example, http://www.grid-support.ac.uk/.