Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2008) 957–961
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Toward the atomistic simulation of T cell epitopes Automated construction of MHC: Peptide structures for free energy calculations

Sarah J. Todman [a], Mark D. Halling-Brown [a], Matthew N. Davies [b,*], Darren R. Flower [b], Melis Kayikci [a], David S. Moss [a]

[a] Department of Crystallography, University of London, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom
[b] Edward Jenner Institute, University of Oxford, Compton, Newbury, Berkshire RG20 7NN, United Kingdom

Received 17 May 2007; received in revised form 25 July 2007; accepted 25 July 2007
Available online 28 July 2007

Abstract

Epitopes mediated by T cells lie at the heart of the adaptive immune response and form the essential nucleus of anti-tumour peptide or epitope-based vaccines. Antigenic T cell epitopes are mediated by major histocompatibility complex (MHC) molecules, which present them to T cell receptors. Calculating the affinity between a given MHC molecule and an antigenic peptide using experimental approaches is both difficult and time consuming, thus various computational methods have been developed for this purpose. A server has been developed to allow a structural approach to the problem by generating specific MHC:peptide complex structures and providing configuration files to run molecular modelling simulations upon them. A system has been produced which allows the automated construction of MHC:peptide structure files and the corresponding configuration files required to execute a molecular dynamics simulation using NAMD. The system has been made available through a web-based front end and stand-alone scripts. Previous attempts at structural prediction of MHC:peptide affinity have been limited due to the paucity of structures and the computational expense in running large scale molecular dynamics simulations. The MHCsim server (http://igrid-ext.cryst.bbk.ac.uk/MHCsim) allows the user to rapidly generate any desired MHC:peptide complex and will facilitate molecular modelling simulation of MHC complexes on an unprecedented scale.

Keywords: T cell epitopes; MHC molecules; Homology modelling; Molecular dynamics; Free energy calculations

1. Introduction

1.1. Major histocompatibility complex molecules

Major histocompatibility complex (MHC) molecules are glycoproteins derived from a highly polymorphic region of chromosome 6. MHC molecules play a vital role in the adaptive immune system by forming complexes with self and non-self peptides and displaying such peptides on the cell surface for inspection by T cell receptors. This enables T cell recognition of cells displaying foreign antigenic peptides derived from infective pathogens. Recognition is achieved following the degradation of antigenic proteins into short peptides and the co-localisation of these peptides with MHC molecules to form stable complexes. These complexes are then transported to the cell surface where they can interact with T cell receptors and thus stimulate an immune response specific to infection. A healthy cell will exhibit only self-peptides and in normal circumstances does not stimulate an immune response. Through interactions with T cell receptors, the repertoire of MHC:peptide complexes shapes the specificity of the T cell response. In humans, MHC molecules are also known as human leukocyte antigens (HLA) molecules.

There are two main types of MHC molecule, which have similar pathways but differ both structurally and functionally. Two different types of MHC molecules: MHC Class I molecules are present on most cells while MHC Class II molecules are only

* Corresponding author.
E-mail addresses: sarah_todman@yahoo.co.uk (S.J. Todman), m.halling-brown@mail.cryst.bbk.ac.uk (M.D. Halling-Brown), m.davies@mail.cryst.bbk.ac.uk (M.N. Davies), darren.flower@jenner.ac.uk (D.R. Flower), kayikci.melis@gmail.com (M. Kayikci), d.moss@mail.cryst.bbk.ac.uk (D.S. Moss).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.07.005

found on so-called "professional" antigen presenting cells (APCs), most importantly: macrophages, B lymphocytes and dendritic cells [1]. Class I and Class II are recognized by distinct sets of T cells: CD8+ and CD4+, respectively. MHC Class I molecules, which predominantly present antigenic material derived from the cytosol, are composed of an α heavy chain, a light chain (β_2-microglobulin or β_2m) and a peptide between 8–11 amino acids in length [2,3]. Class II molecules are composed of non-covalently bound α and β chains that typically bind peptides 12–24 amino acids in length [4,5].

1.2. MHC binding prediction

The ability to predict the binding affinity of antigenic peptides to MHC molecules accurately is a key objective of vaccine design as well as having applications for research into autoimmunity, protein therapeutics and transplantation. The traditional experimental approach to determining antigenic peptides is to scan the whole sequence of the target antigen by synthesising overlapping peptide fragments and assaying each for an immune response. Although the technique is potentially accurate, it requires a great deal of time, labour, and financial resources. An alternative is to reduce the number of peptides required for scanning by predicting high affinity peptides using computational techniques. One possibly method is through using reverse vaccinology to identify bacterial proteins that have secreted or surface-like properties that may contain epitope sequences. Another attempt is to identify specific protein sequences that have the capacity to bind to an MHC molecule and hence the potential to stimulate a T cell response. Broadly, the MHC binding peptide prediction methods can be divided into three main groups: (a) motif-based methods [6], (b) statistical/mathematical expression-based methods [7–11], and (c) structure-based methods [12–14]. These methods derive their parameters from experimentally determined IC_{50} and BL_{50} binding data [15]. Binding motifs describe general position-based patterns of recurrent amino acids favourable for MHC:peptide binding. Statistical/mathematical expression-based methods include quantitative matrix, neural network-based approaches, and support vector machines (SVM). Quantitative matrices provide a linear model with easy to implement capabilities while neural networks and SVMs are more complex, non-linear self-learning systems. Structure based methods represent an entirely different approach. The technique calculates the binding energy of the peptide–MHC complex by using a molecular dynamics simulation to evaluate the atomic interactions within the MHC and the peptide. Potentially, they offer a greater level of precision than the other techniques but are computationally expensive and difficult to initiate.

Currently many vaccines are targeted at the most prevalent HLA allotypes within a particular population. Individuals normally have between six and eight MHC molecules (which vary tremendously between populations, regions and even families) and each molecule has the potential to bind to hundreds or thousands of peptides [16]. Compared to the large number of MHC allotypes, there is only sufficient data to develop motif or machine based learning methods for a very limited number of alleles. In this paper we describe an application that enables the prediction of the binding affinity of antigenic peptides to any known HLA allotype. It focuses on the automated construction of MHC:peptide structures, even when a crystallised structure of the required allotype is absent, as well as automatically constructed configuration files necessary for free energy binding calculations. These calculations will be achieved using molecular dynamics (MD) simulations with the aim of assessing the strength of the interaction between the peptide and the MHC molecule. This method will form part of an antigenic peptide prediction pipeline enabling a high throughput peptide screening process to identify MHC binding peptides.

2. Methodology

2.1. Data extraction

Official sequences for the WHO HLA Nomenclature Committee for Factors of the HLA System were obtained from the IMGT/HLA Database as part of the ImMunoGeneTics project (IMGT). These sequences were stored in a database and considered to be accurate for the various allotypes.

An extensive search of the Protein Data Bank (PDB) [17] was undertaken to determine the available HLA crystal structures. Once acquired, these entries were parsed to extract monomers where appropriate, isolate each chain and surrounding water and confirm that a completed peptide structure was available. Specific information was then extracted from each PDB file and stored within the database. In order to ensure each structure was classified with the correct allotype, a series of pairwise alignments were carried out, using the Smith–Waterman algorithm supplied as part of the BioPerl toolkit [18], between the heavy HLA chain and every single sequence from the IMGT/HLA database.

In order to facilitate the positioning of peptides with the binding groove of MHC class II structures, a visual analysis was undertaken to determine the position of pocket one and nine and the relative positions of the amino acid side chains with which they interact (see Fig. 1).

2.2. Construction of MHC:peptide structures

Software was developed in Perl which allows the creation of relevant PDB files and Protein Structure Files (PSF files) which contain respectively the solvated structure and topology of any HLA allotype with any peptide sequence. The software initially requires the class of the molecule to be identified (i.e. class I or II), the necessary allotype(s) to be chosen and the peptide to be defined (in which case it must be of length nine residues) or that the original peptide from the crystal structure is to be retained. Additionally there are some optional preferences that can be implemented. These include the solvation of the structure by placing it within a water box (thus creating a full protein–water system) and the removal of the membrane proximal domains, which reduces the computational time of MD simulations at the cost of full structural flexibility. Also, if the structure is a class II molecule, there is an option to ensure the α-chain does not

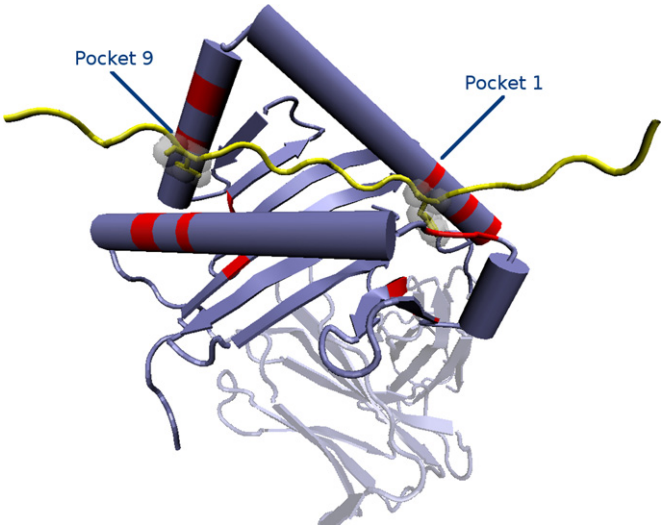

Fig. 1. Pockets 1 and 9. The P1 and P9 peptide residues, shown here in yellow liquorice surrounded by semi transparent VDW spheres, can be located by the highlighting the MHC residues from the pockets.

undergo any mutations. Fig. 2 shows an overview of the MHC:peptide construction process.

An initial scan determines whether a crystal structure exists within the database for the desired allotype. If there is a structure, the most accurate is used, in terms of insertions, deletions and substitutions. If there is no structure available, a series of pairwise alignments are run between the requested allotype sequence and every crystal structure sequence to obtain a closest match. This structure is then mutated to the correct allotype using *psfgen*. The psfgen tool is a plug-in for the VMD program [19] which generates structures and coordinates using the CHARMM force field. The psfgen structure building tool consists of a portable library of structure and file manipulation routines including the construction of missing atomic coordinates. The side chains of the MHC

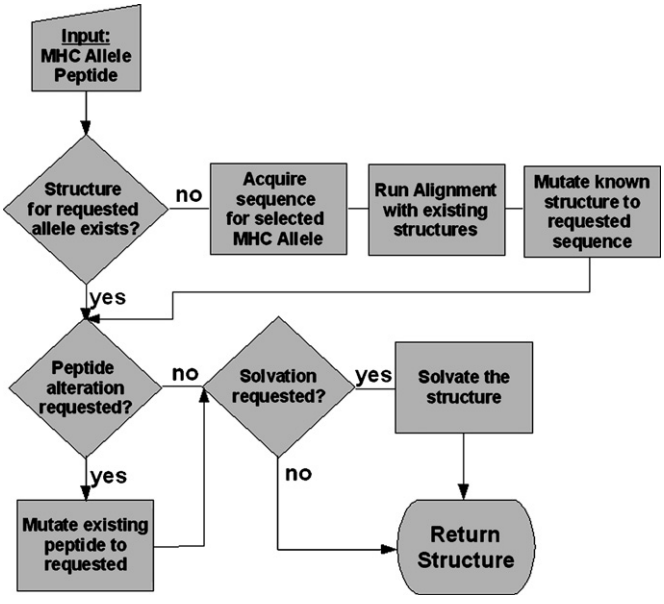

Fig. 2. Dataflow. Dataflow diagram of the MHC:peptide construction process.

molecules and then of the peptide are mutated as necessary. The new side chains are build in an extended conformation which avoids any spatial clashes.

If it is a class II molecule and it is requested that the α-chain should undergo mutations, then the correct sequence is found for this chain so that it can be mutated to the correct allotype. If a peptide has been defined and it is a class I molecule, only crystal structures that have an original peptide of length 9 will have been used and again *psfgen* is used to mutate the peptide to the correct sequence. If it is a class II molecule, the region of the peptide between positions P1 and P9 will be mutated to the desired nine length sequence and the flanking residues will be mutated to alanines. If, however, no peptide sequence is defined then the original peptide, which can be of any length, is retained. There is an option to solvate the resulting MHC:peptide structure in a water box of a requested size around the molecule. The standard is taken to be a minimum distance of 5 Å between the outermost atom and the edge of the water box. Additionally there is the option to remove the membrane proximal domain(s) of the MHC molecule and generate only the two membrane-distal, peptide binding regions.

2.3. Construction of NAMD configuration files

In addition to providing the structure files for any MHC:peptide complex, techniques to produce the configuration files for a NAMD simulation [19] were implemented. This involved the preparation of the various files required for an energy minimisation, MD or ABF simulation. These files are tailored for each structure file that is generated and presented in such a manner that every file that is required by the simulation is present. NAMD was selected as it is publicly available under a non-exclusive, non-commercial user license.

2.4. Web interface for the construction of an MHC:peptide complex

A web-based interface, MHCsim, was constructed in PHP. The interface presents the user with options to select the class of the HLA molecule, which in turn populates the drop down boxes where the user can select the allotype(s). The peptide can then be entered, which is restricted to a length of 9, or a box can be checked which retains the original peptide from the resultant source structure. The final stage is to select optional extras such as solvating the structure or the removal of the membrane proximal domain(s). Once generated, this structure can then be downloaded directly or used to create simulation configuration files for free energy simulations. The structure and configuration files generated by a user are kept on the server for two hours before they are deleted.

3. Results

3.1. Database statistics

A database which was constructed to contain all official HLA sequences and all available HLA crystal structures

currently holds 1839 HLA class I sequences and 875 HLA class II sequences from the IMGT-HLA and 104 crystal structures. There are 104 suitable HLA crystal structures, 81 of which are class I molecules and 23 are class II. There are 15 distinct class I allotypes, two distinct class II α-chain allotypes and six distinct class II β-chain allotypes. The most prevalent class I allotype structure within the PDB is HLA-A*0201, which occurs 36 times. HLA-DRB1*010101 is the most commonly occurring class II β-chain molecule with 15 structure entries and HLA-DRA*0101 is the most common class II α-chain molecule with 22 entries.

3.2. Data validation

In order to access the quality of the structures generated by the servers, it was necessary to compare the RMSD values of them with original MHC crystal structures that were used by the server as a template. The SuperPose software (http://wishart.biology.ualberta.ca/SuperPose/) was used to generate RMSD values. The software superimposed the structures of the MHC Class I heavy chain and calculates the RMSD between the two structures. The software was used on four different datasets, each containing 15 combinations of structures. The first compared structures of HLA*A_0201 generated by MHCsim using the A_1101 templates against genuine A_0201 structures. The second compared random combinations of A_0201 structures, the third random compared combinations of MHC Class I structures and the fourth compared random combinations of MHC Class II structures. It can be seen from Table 1 that average RMSD value of the MHC Class I structures are all >1. Although the average value for the generated structure is higher than the crystal structures, it is not a significant distance and lower than the average value of superimposed MHC Class II structures. This suggests that the server is generating structures that are consistent with predetermined crystal structures.

3.3. The graphical user interface

The MHCsim server is simple to use and has a step by step guide that allows a user to easily and quickly create an MHC:peptide structure. MHCsim is freely available via the URL: http://igrid-ext.cryst.bbk.ac.uk/MHCsim. The server offers the user two interlinked stages. First is to construct PDB and PSF files containing details of the MHC:peptide structures, while the second is to create MD simulation configuration files. All files are directly downloadable and do not contain absolute paths so they can be run from any directory on a personal computer.

Table 1
RMSD values of four datasets compared using Superpose, generated HLA*A_0201 against genuine A_0201 structures, random combinations of A_0201 structures, random combinations of MHC Class I structures and random combinations of MHC Class II structures

	Average	Min	Max
Generated A_0201 structures with known A_0201 structures	0.907	0.73	1.15
Randomly selected known A_0201 structures	0.819	0.48	1.17
Randomly selected known class I structures	0.837	0.38	1.18
Randomly selected known class II structures	1.146	0.53	1.92

4. Discussion

We have created a system for the automated production of MHC:peptide structure files and configuration files for utilising these in a molecular dynamics simulation. This system is available through a web-based interface or as standalone scripts. Motif-based methods are very imperfect predictors of T cell epitopes. Their main advantage is allele coverage, yet even they are only able to predict the binding of peptides to a limited number of MHC Alleles. While there are over 2500 known alleles [20], the various motif methods only cover between 20 and 120 alleles. This leaves many alleles for which binding affinity cannot be predicted. By producing a system to facilitate MD simulations with any combination of MHC allele and peptide, we seek to surmount such problems. Currently there are no published applications that use free energy binding calculations, for example thermodynamic integration (TI), free energy perturbation (FEP) or adaptive biasing force (ABF), to assess the binding affinities of MHC:peptide molecules. Although these energy calculations are computationally expensive [21,22], previous work [13,14] has indicated the validity of structure based affinity calculations for MHC:peptide binding prediction. The bottleneck of free energy calculations has shifted from a purely computational aspect to a human one due to the need for qualified modellers to set these calculations up.

It is anticipated that this system will be integrated into an antigenic peptide prediction pipeline (APPP) that is being developed as part of the ImmunologyGrid project [23]. This pipeline currently utilises computationally inexpensive techniques, such as motif-based biding prediction, to predict antigenic peptides from protein sequences. We expect the system discussed here to be used to automatically build structures and configuration files to assess the predictions emerging from this pipeline. Work is also ongoing to develop techniques to detect a non-binding peptide early on in a free energy simulation in order that it can be aborted, with concomitant savings in CPU time. It is anticipated that these capabilities will be integrated into the existing pipeline. The APPP offers important advances in the automation and potential for high throughput screening techniques. It will soon be developed further to utilise Grid-based supercomputing technologies to reduce simulation wall clock times by increasing processing power [24–26].

A further development that is expected is the integration of automated submission of prepared simulations onto Grid enabled clusters. This will be implemented using a piece of middleware called application hosting environment (AHE) [27]. AHE allow submission, monitoring and retrieval of applications over a grid environment. This allows access to the National Grid Service (NGS), DEISA resources and web service enabled servers. It is expected that this increase of available computing power will allow a considerable increase in the number of simulations that can be submitted through the APPP.

5. Conclusions

Successful T cell epitope prediction is a vital part of the rational design and optimisation of next-generation vaccines. Motif-based techniques lack insufficient subtlety to discriminate properly peptide binders from non-binders. Machine based learning requires large quantities of experimental data that is available for a very limited number of MHC alleles. Only simulation methods offer the opportunity to predict the peptide specificity of any allele with a minimum of extant experimental data.

The APPP that is being developed will allow for a far more flexible and diverse range of MHC:peptide affinity calculations than has previously been possible. The work described here forms part of the ImmunoGrid project [23]. ImmunoGrid is a European Union project that started in 2006 and aims to eventually simulate the whole human immune system using Grid technologies. It hopes to provide useful tools for clinicians and vaccine/immunotherapy developers to easily identify the optimal immunisation protocols necessary for the prevention and treatment of human disease. It is anticipated that the simulation techniques developed here will be utilised within this project to help design and optimise future vaccines.

Acknowledgements

The authors gratefully acknowledge the funding of the European Union ImmunoGrid project (MDHB and DSM), the ESPRC grant EP/D501377/1 (DRF and MND) and the student support of the Biology Department, University of York (SJT).

References

[1] G.J. Kersh, P.M. Allen, Essential flexibility in the T-cell recognition of antigen, Nature 380 (1996) 495–498.
[2] M.A. Saper, P.J. Bjorkman, D.C. Wiley, Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution, J. Mol. Biol. 219 (1991) 277–319.
[3] P. Sliz, O. Michielin, J.C. Cerottini, I. Luescher, P. Romero, M. Karplus, D.C. Wiley, Crystal structures of two closely related but antigenically distinct HLA-A2/melanocyte–melanoma tumor-antigen peptide complexes, J. Immunol. 167 (2001) 3276–3284.
[4] B.J. McFarland, C. Beeson, Binding interactions between peptides and proteins of the class II major histocompatibility complex, Med. Res. Rev. 22 (2002) 168–203.
[5] D.R. Madden, The three-dimensional structure of peptide-MHC complexes, Annu. Rev. Immunol. 13 (1995) 587–622.
[6] H. Rammensee, J. Bachmann, N.P. Emmerich, O.A. Bachor, S. Stevanovic, SYFPEITHI, database for MHC ligands and peptide motifs, Immunogenetics 50 (1999) 213–219.
[7] K. Udaka, H. Mamitsuka, Y. Nakaseko, N. Abe, Prediction of MHC class I binding peptides by a query learning algorithm based on hidden Markov models, J. Biol. Phys. 28 (2002) 183–194.
[8] I.A. Doytchinova, M.J. Blythe, D.R. Flower, Additive method for the prediction of protein-peptide binding affinity. Application to the MHC Class I molecule HLA-A*0201, J. Proteome Res. 1 (2002) 263–272.
[9] P. Donnes, A. Elofsson, Prediction of MHC class I binding peptides, using SVMHC, BMC Bioinformatics 3 (2002) 25.
[10] P.A. Reche, J.P. Glutting, E.L. Reinherz, Prediction of MHC class I binding peptides using profile motifs, Hum. Immunol. 63 (2002) 701–709.
[11] J. Salomon, D.R. Flower, Predicting Class II MHC–peptide binding. A kernel based approach using similarity scores, BMC Bioinformatics 7 (2006) 501.
[12] O. Schueler-Furman, Y. Altuvia, A. Sette, H. Margalit, Structure-based prediction of binding peptides to MHC class I molecules, application to a broad range of MHC alleles, Protein Sci. 9 (9) (2000) 1838–1844.
[13] M.N. Davies, C.E. Sansom, C. Beazley, D.S. Moss, A novel predictive technique for the MHC Class II peptide-binding interaction, Mol. Med. 9 (2003) 220–225.
[14] M.N. Davies, C.K. Hattotuwagama, D.S. Moss, M.G. Drew, D.R. Flower, Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity, BMC Struct. Biol. 6 (2006) 5.
[15] A. Sette, J. Sidney, Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism, Immunogenetics 50 (1999) 201–212.
[16] O. Lund, M. Nielsen, C. Kesmir, J. Christensen, C. Lundegaard, P. Worning, S. Brunak, Web-based tools for vaccine design, HIV Mol. Immunol. (2002) 45–51.
[17] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[18] J.E. Stajich, D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C.J. Mungall, B.I. Osborne, M.R. Pocock, P. Schattner, M. Senger, L.D. Stein, E. Stupka, M.D. Wilkinson, E. Birney, The Bioperl toolkit, Perl modules for the life sciences, Genome Res. 12 (2002) 1611–1880.
[19] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, K. Schulten, Scalable molecular dynamics with NAMD, J. Comput. Chem. 26 (2005) 1781–1802.
[20] J. Robinson, M.J. Waller, P. Parham, N. de Groot, R. Bontrop, L.J. Kennedy, P. Stoehr, S.G.E. Marsh, IMGT/HLA and IMGT/MHC, sequence databases for the study of the major histocompatibility complex, Nucleic Acids Res. 31 (2003) 311–314.
[21] C. Chipot, J. Henin, Exploring the free-energy landscape of a short peptide using an average force, J. Chem. Phys. 123 (2005) 244906.
[22] J. Henin, C. Chipot, Overcoming free energy barriers using unconstrained molecular dynamics simulations, J. Chem. Phys. 121 (2004) 2904–2914.
[23] ImmunologyGrid Project. http://www.immunologygrid.org.
[24] S. Wan, P.V. Coveney, D.R. Flower, Peptide recognition by the T cell receptor, Comparison of binding free energies from thermodynamic integration. Poisson–Boltzmann and linear interaction energy approximations, Philos. Trans. A: Math. Phys. Eng. Sci. 363 (2005) 2037–2053.
[25] S. Wan, P.V. Coveney, D.R. Flower, Molecular basis of peptide recognition by the T-cell receptor. Affinity differences calculated using large scale computing, J. Immunol. 175 (2005) 1715–1723.
[26] S. Wan, P.V. Coveney, D.R. Flower, Large scale molecular dynamics simulations of HLA-A*0201 complexed with a tumour-specific antigenic peptide. Can the α3 and β2m domains be neglected? J. Comput. Chem. 25 (2004) 1803–1810.
[27] J. Cohen, A.S. McGough, J. Darlington, N. Furmento, G. Kong, A. Mayer, RealityGrid an integrated approach to middleware through ICENI, Philos. Trans. A: Math. Phys. Eng. Sci. 363 (2005) 1817–1827.