Copyright © 1997 Ivo Ivanov

SCHOOL OF PHARMACY
MEDICAL UNIVERSITY OF SOFIA
Dunav st. 2, 1000 Sofia, Bulgaria
tel/fax: +359 2 987 9874
http://www.pharmfac.net

Modelling of Peptide – Protein Interactions by BlueGene-P

Atanas Patronov, Mariyana Atanasova, Ivan Dimitrov and Irini Doytchinova*

School of Pharmacy, Medical University of Sofia, 2 Dunav st., 1000 Sofia, Bulgaria,
*Contact: idoytchinova@pharmfac.net

Modelling of protein – protein or peptide – protein interactions always was a very attractive area of Computational Chemistry. During the last years, it became even more relevant because of the many new protein drugs entering the pharmaceutical market. Protein drugs, known also as Biologicals, are considered to be the drugs of the future. Usually, they have better therapeutic profile and fewer side effects than the conventional small molecule drugs. Biologicals comprise therapeutic proteins, monoclonal antibodies and vaccines. The global biologics market is valued at an estimated in 149 bln USD in 2010 and is expected to reach 239 bln USD by 2015 (Figure 1).

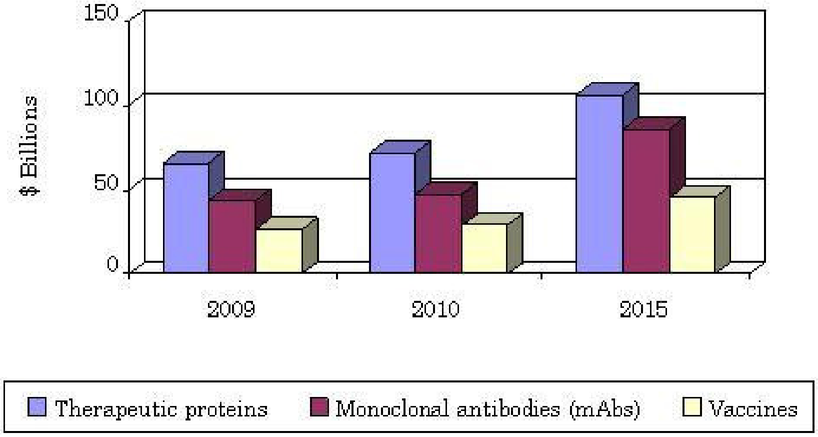


Figure 1. Global biologicals market. Source: BCC Research

Immunogenicity is the ability of a given molecule to provoke an immune response in the organism. It is quite necessary to know the immunogenicity of proteins administered as drugs or vaccines. For the therapeutic proteins and monoclonal antibodies, immunogenicity is an undesired effect shortening their half-life. For the vaccines, however, immunogenicity is the main pharmacological effect ensuring the development of immunological memory.

The smallest part of an antigen recognizable by the immune system is called epitope. Epitopes belong to both foreign and self proteins, and they can be categorized as conformational or linear, depending on their structure and integration with the paratope. T-cell epitopes are presented on the surface of an antigen-presenting cell, where they are bound to MHC molecules in order to induce immune response. MHC class I molecules usually present peptides between 8 and 11 amino acids in length, whereas the peptides binding to MHC class II may have length from 12-25 amino acids. MHC class II proteins bind oligopeptide fragments derived through the proteolysis of pathogen antigens, and present them at the cell surface for recognition by CD4+ T-cells (Figure 2). If sufficient quantities of the epitope are presented, the T cell may trigger an adaptive immune response specific for the pathogen. Class II MHCs are expressed on specialized cell types, including professional Antigen Presenting Cells (APCs), such as B cells, macrophages and dendritic cells while class I MHC are found on every nucleated cell of the body.

The recognition of epitopes by T-cells and the induction of immune response have a key role for the individual’s immune system. Even the slightest deviation from the normal functioning can have a great impact on the organism. In case of autoimmune disease the T-cells recognize the cell’s native peptides as foreign and attack and eventually destroy the organism’s own tissues. One of the key issues in T-cell epitope prediction is the prediction of MHC binding, as it is considered as a prerequisite for T-cell recognition. All T-cell epitopes are good MHC binders, but not all good MHC binders are T-cell epitopes.

MHCs are amongst the most polymorphic protein in higher vertebrates, with more than 8000 class I and class II MHC molecules listed in IMGT/HLA. Determining the peptide binding preferences exhibited by this extensive set of alleles is beyond the present capacity of experimental techniques, necessitating the development of bioinformatics prediction methodologies. The methods for T-cell epitope prediction are sequence-based and structure-based. T-cell epitope prediction typically involves defining the peptide binding specificity of specific class I or class II MHC alleles and then predicting epitopes *in silico*. Using peptide sequence data, experimentally-determined affinity data has been used in the construction of many T-cell epitope prediction algorithms.

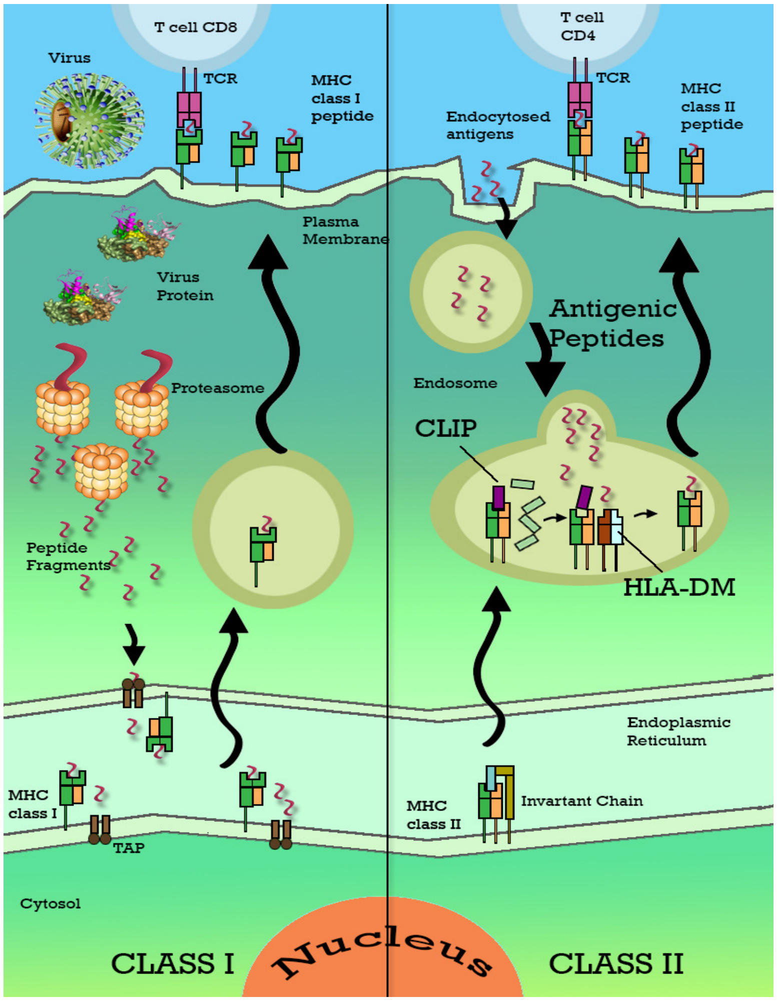


Figure 2. Antigen-processing pathways in the cell. Left: intracellular pathway. Protein is cleaved into oligopeptides in the proteasome, the peptides enter the endoplasmic reticulum via TAP-protein, bind to MHC class I and the complex peptide – MHC-protein is presented on the cell surface. Right: extracellular pathway. Protein is endocytozed, cleaved into oligopeptides in the endosome, bind to MHC class II protein and presented on the cell surface (http://rsob.royalsocietypublishing.org/content/3/1/120139.full).

We applied two structure-based methods to predict peptide binding to MHC class II proteins. As starting information we used the X-ray structure of a peptide – MHC protein complex (Figure 3). The binding core of the peptide from the complex was used to generate a combinatorial peptide library based on the principle "single amino acid substitution". Each amino acid (aa) at each of the 9 positions was substituted by the remaining 19 naturally occurring amino acids and thus, a library of 172 ligands (9 positions x 19 aas + 1 original ligand) was created.

Each peptide in the library was placed on the peptide binding site of the MHC protein and the binding energy of the complex was calculated by molecular dynamics (MD) simulations or by molecular docking. MD is a simulated method where the interaction between two molecules is simulated over the time. It is time-consuming method and requires a high performance computing. Docking is a non-simulated method which scores the energy of the complex as a difference between the final and initial states of the molecules. It is fast and could be run on a single PC. Both methods were used in our study to assess the binding energy and the results were compared in terms of their ability to predict binders and non-binders.

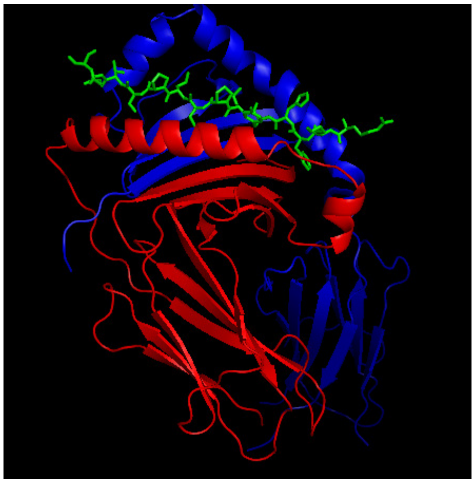


Figure 3. X-ray structure of peptide – HLA-DP2 protein complex (pdb code: 3lqz). Peptide is given in green, the α-chain DPA1*0103 – in red, the β-chain DPB1*0101 – in blue.

GROMACS v.4.0.7 on BlueGene-P was used for the MD simulations. The MD protocol starts with converting the pdb file into gmx (Figure 4). GROMOS force field is used. Next, a box around the peptide – MHC protein complex is created; filled with water molecules; the charge of the complex is neutralized by counterions and the energy of the complex is minimized for a short period of time. The next step is to run a position-restrained MD simulation, i.e. the atom positions of the complex are frozen and only the water molecules move and adjust around the complex. Finally, a MD simulation is run with simulated annealing from 100 to 310°K; the final interaction energies are recorded. Two types of energies are considered: Lenard-Jones short-range (LJ-SR) and Coulomb short-range (Coul-SR).

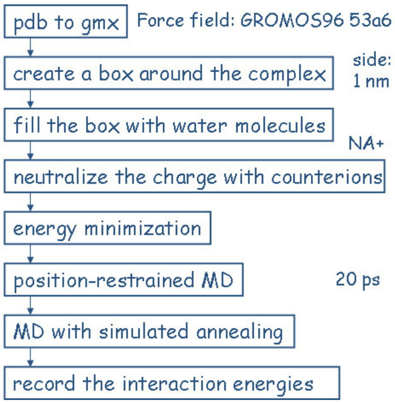

pdb to gmx
Force field: GROMOS96 53a6
create a box around the complex
side:
1 nm
fill the box with water molecules
NA+
neutralize the charge with counterions
energy minimization
position-restrained MD
20 ps
MD with simulated annealing
record the interaction energies

Figure 4. MD protocol.

MD simulations are performed over the whole peptide library; the energies are normalized and compiled into a quantitative matrix (QM). The favourable aas have positive values, the disfavourable ones take negative. Thus created QM is ready to be used for virtual screening of different proteins and for prediction of the most probable binders to HLA-DP2.

During the MD simulations two questions arose: “Which energy to use for prediction?” and “How long does it take for the complex to reach equilibrium?”

To answer the first question we used a test set of 457 known binders to HLA-DP2 protein and applied the derived QM to identify the binders among the top 5% of the best predicted binding peptides. The LJ-SR energies identified 33% of the known binders, the Coul-SR recognized 25% of them and the sum of both energies – 29%. Hence, the LJ-SR energies gave better predictions and they were used further in the study.

To answer the second question, the X-ray complex was simulated for 50 ns and the coordinates were recorded every 10 ns. The overlapped coordinates of the complex is presented in Figure 5. It is evident that the peptide binding core is stable, only the flanking residues at the C-terminal move up and down. Based on these results, we decided to compromise between accuracy and time for calculation and used 1 ns simulation time in all further simulations.

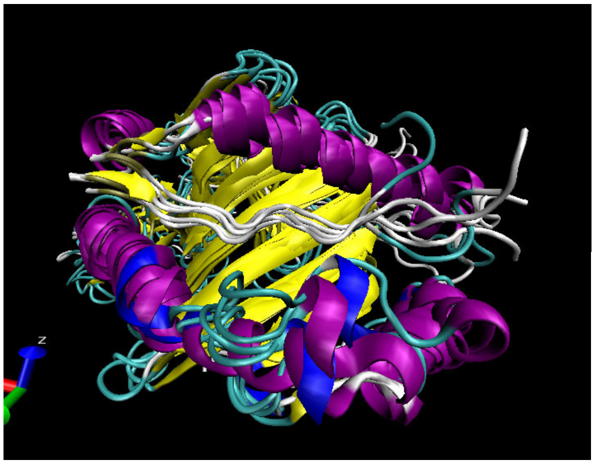
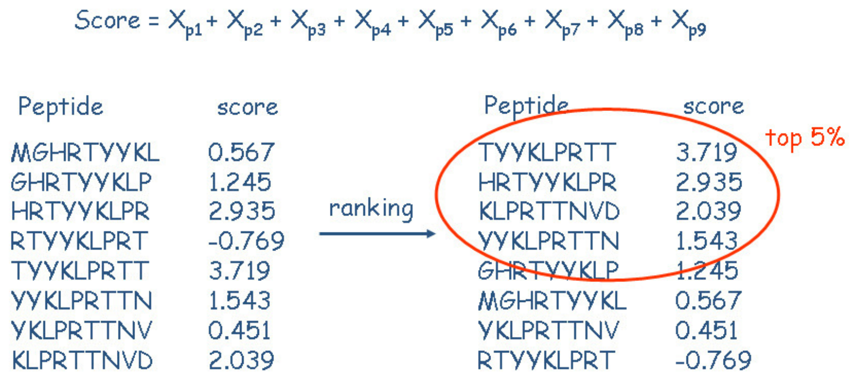


Figure 5. Overlapped coordinated of the complex peptide – HLA-DP2 protein recorded every 10 ns. Peptide is given in grey, MHC helices – in purple, MHC β-sheets – in yellow.

The same protocol was applied in the molecular docking study. We used AutoDock 4.5 for BlueGene-P to dock each peptide from the peptide library into the HLA-DP2 binding site. The energies of the complexes were recorded, normalized and compiled into docking score-based QM (DS-QM).

In order to compare the predictive ability of the two QMs, derived by MD and docking, a test set of 457 known binders to HLA-DP2 originating from 24 proteins was extracted from the ImmuneEpitope database (http://www.immuneepitope.org) and used for external validation. Each origin protein was represented as a set of overlapping nonamers and the score of each nonamer was calculated as a sum of the binding scores of each aa at each position (Figure 6). Two QMs were used to score the peptides – MD-based QM and DS-based QM. Peptides were arranged in a descending mode according to their binding score, the top 5% were selected and compared to the known binders. If any of the predicted binders is among the known ones, it is considered as true predicted binder. The ratio between the true predicted and all known binders gives *sensitivity* of predictions. Results are given in Figure 7.



$$\text{Score} = X_{p1} + X_{p2} + X_{p3} + X_{p4} + X_{p5} + X_{p6} + X_{p7} + X_{p8} + X_{p9}$$

Peptide	score
MGHRTYYKL	0.567
GHRTYYKLP	1.245
HRTYYKLPR	2.935
RTYYKLPRT	-0.769
TYYKLPRTT	3.719
YYKLPRTTN	1.543
YKLPRTTNV	0.451
KLPRTTNVD	2.039

Peptide	score
TYYKLPRTT	3.719
HRTYYKLPR	2.935
KLPRTTNVD	2.039
YYKLPRTTN	1.543
GHRTYYKLP	1.245
MGHRTYYKL	0.567
YKLPRTTNV	0.451
RTYYKLPRT	-0.769

Figure 6. External validation by test set of 457 known binders originating from 24 proteins.

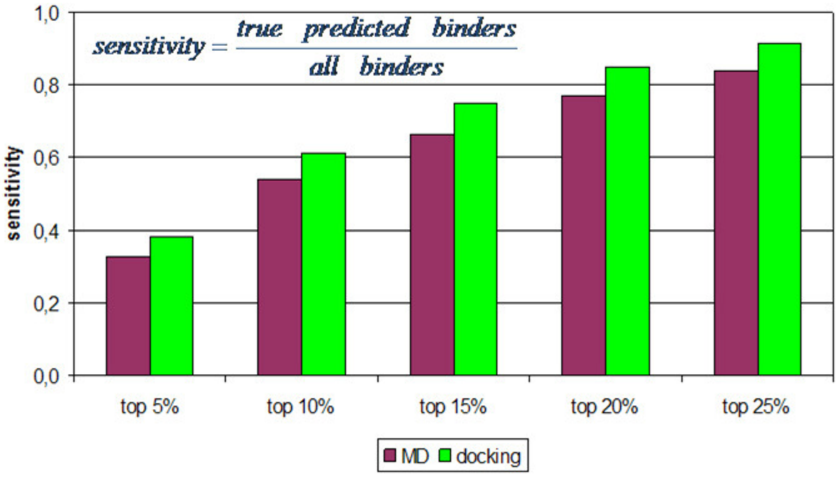


Figure 7. *Sensitivity* of predictions made by MD-based QM (purple) and DS-based QM (green) at several different cutoffs (top 5%, 10%, 15%, 20% and 25%).

Results clearly showed that the DS-based QM gives better predictions than the MD-based QM. DS-QM recognized 38% of the known binders in the top 5% of the best predicted peptides. MD-QM recognized 33% at the same cutoff. Apart from being better predictor, docking takes a less time for calculation (Table 1). While the MD simulation of one peptide – MHC protein complex for 1 ns takes 11 hours on BlueGene-P, a batch of 20 complexes run by AutoDock on BlueGene-P takes 10 – 15 min.

Table 1. Sensitivity of predictions and time for calculation by MD and docking.

Method	Sensitivity at the top 5%	Time for calculation
Molecular dynamics by GROMACS on BlueGene-P	33%	one complex for 1 ns 11 hours
Molecular docking by AutoDock on BlueGene-P	38%	a batch of 20 complexes 10-15 min

Further in the study we concentrated on docking calculations and generated another 22 QMs covering the most frequent human MHC class II proteins. The QMs were implemented into a specially designed site for docking-based prediction of peptides binding to MHC class II. The site was named **EpiDOCK** and is freely available at: http://epidock.ddg-pharmfac.net.

Acknowledgement: Authors thank Peicho Petkov from Sofia University, Bulgaria, for his assistance with BlueGene-P, and Dr. Darren R. Flower from Aston University, UK, for his advising on the study.

Further reading:

1. Doytchinova, I., Petkov, P., Dimitrov, I., Atanasova, M., Flower, D. R. HLA-DP2 binding prediction by molecular dynamics simulations. *Protein Sci.*, 20(11), 1918-1928, 2011.
2. Atanasova, M., Patronov, A., Dimitrov, I., Flower, D. R., Doytchinova, I. EpiDOCK - a molecular docking-based tool for MHC class II binding prediction. *PEDS*, published online May 9, 2013.