Side chain substitution benchmark for peptide/MHC interaction

BERNHARD KNAPP, ULRICH OMASITS, AND WOLFGANG SCHREINER

Unit for Medical Statistics and Informatics, Section for Biomedical Computer Simulation and Bioinformatics, Medical University of Vienna, A-1090 Vienna, Austria

(RECEIVED December 17, 2007; FINAL REVISION February 12, 2008; ACCEPTED February 15, 2008)

Abstract

The prediction of T-cell epitopes is an essential part in virtual immunology. Apart from sequence-based techniques, which achieve good results but fail to give insight into the binding behavior of a certain peptide binding to a major histocompatibility complex, structure-based approaches are another important technique. An essential step is the correct placement of the side chains for a given peptide in cases where no experimental data for the structure are available. To our knowledge, no benchmark for side chain substitution in the area of HLA has been reported in the literature. Here, we present a comparison of five different tools (SCWRL, SCATD, SPDBV, SCit, IRECS) applicable for side chain substitution. Each tool is tested on 29 different HLA-A2 structures with experimentally known side chain positions. Parts of the benchmark are correctness, reliability, runtime, and usability. For validation, the root mean square deviations between X-ray structures and predicted structures are used. All tools show different strengths and weaknesses.

Keywords: MHC; benchmark; side chain substitution; side chain placement; RMSD; rotamers

Peptides (p) binding to major histocompatibility complex (MHC) molecules play an essential role in immunology. The MHC presents a peptide on the cell surface to the T-cell receptor (TCR) of cells of the immune system. While MHC class I molecules are recognized by CD8+ cytotoxic T-cells, MHC class II molecules are recognized by CD4+ “helper” T-cells. The binding between MHC and peptide is a prerequisite for invoking and eliciting any adaptive immune response. In the case of MHC class I, the binding of the peptide in the MHC groove is stabilized at amino and carboxyl terminus by a cluster of tyrosine residues forming hydrogen bonds and ionic interactions. The second major stabilization factors are the anchor residues of the peptide interacting with the binding pockets of the MHC. Correct

modeling of peptide residues is essential for any kind of structural binding prediction.

To model the binding interaction between peptide and MHC, there are two major techniques (Tong et al. 2007) in virtual immunology: (1) Sequence-based approaches are usually calibrated on experimental data and obtain, if enough calibration data are available, good results in quite a short computational time. The drawback of sequence-based approaches is that they fail to give insight into the binding behavior of the pMHC. They can give an answer if a certain peptide is binding but not why it is binding. (2) Structure-based approaches give insight into the binding but require three-dimensional data of the target pMHC molecule. To our knowledge, there are currently ~70 HLA-A2 structures available in the Protein Data Bank (PDB) (Bergman et al. 2000). To create other pMHC complexes, computational methods for side chain substitutions are needed. Most tools for that purpose have been developed in the background of homology modeling of whole biocomplexes. For this study, we evaluate their performance in the area of pMHC. None of these tools

Reprint requests to: Bernhard Knapp, Unit for Medical Statistics and Informatics, Section for Biomedical Computer Simulation and Bioinformatics, Medical University of Vienna, General Hospital, Room BT88-88.03.712, Spitalgasse 23, A-1090 Wien, Austria; e-mail: bernhard.knapp@meduniwien.ac.at; fax: 43-1-40400-6677.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.073402508>.

has been created exactly for this task. It is not the aim of this study to generally judge the tools but to provide a benchmark of which tool is most applicable in this special area of pMHC interactions where even a small difference in the side chain positioning can discriminate between binding and nonbinding peptides.

Results

For each of the five tools, we evaluate different aspects: (1) the correctness, i.e., the root mean square deviation (RMSD) to the “correct” X-ray structure; (2) the reliability, which means if a tool is able to deal with different pMHC complexes; (3) the runtime; and (4) the usability, i.e., how easy it is to use this tool in an automated approach for high data-throughput.

Correctness

Due to the setup of this benchmark (see Materials and Methods), for every predicted complex a corresponding

experimentally determined structure is available. The RMSD is calculated between all original complexes and the corresponding five predicted complexes. In the original X-ray structure of complex 28 (PDB-id 2git), the C ϵ atom and the ϵ amino group of LYS5 are missing in the peptide. Therefore, these two atoms are excluded from RMSD calculation, but due to the relatively low number of test complexes available (see Materials and Methods), we still decided to include complex 28 into the results.

Box plots for each tool are shown in Figure 1A, whereas the average RMSD values for each amino acid type are illustrated in Figure 1B. Figure 1C shows the performance of the five tools on each single complex, and in Figure 1D, the different positions in the peptide are compared. All paired t-tests between the RMSDs of the tools are significant except the one between IRECS and SPDBV. It can be seen that SCWRL (Canutescu et al. 2003) outperforms the other tools on almost each complex, which makes it the favorite regarding correctness. SCATD (Xu 2005) obtains results not much worse as

Figure 1. RMSD between predicted and X-ray structure. (A) Box plots of the RMSD values grouped for each tool. (B) Average RMSD values for each amino acid type. (C) RMSD values for each complex. (D) RMSD values for each position. Positions 2 and 9 are anchor residues, while position 6 is an auxiliary anchor.

SCWRL. SPDBV (Guex and Peitsch 1997) and IRECS (Hartmann et al. 2007) are nearly ranked equal, which is quite surprising considering that SPDBV is 8 years older. A possible reason is that IRECS rearranges all side chains (using `-m` option) or does not create all side chains (using `-s` option and lowercase letters). The highest RMSD is produced by SCit (Gautier et al. 2004); this can partly be explained by the flexible receptor side chains. Although there is an option of leaving single residues unaffected, unfortunately all side chains are repositioned by the tool. Especially in IRECS and SCit, it can be seen that these tools have not been designed for pMHC interaction, they may have their strengths elsewhere.

As a second measure of correctness, we use the dihedral angles χ_1 and χ_2 . We screen the different thresholds with a step of 1° . This is done due to the fact that setting the threshold fixed to, for example 40° , does not really reveal the performance of a tool because one tool could be better at 30° threshold and the second one at 40° . The overall results of the dihedral angles (Fig. 2) are similar to the ones achieved by the RMSD. It is interesting that IRECS seems to perform a little bit better in the dihedral angles than in the RMSD. In the χ_1 angle, SCATD is a little bit better than SCWRL, but in χ_2 SCWRL seems to be on top.

Reliability

For all 29 complexes, all tools except SCit could position all side chains without reporting any difficulties. For the

complexes with PDB identification 1jht and 1p79 SCit encountered problems for the α -chain of MHC and the β_2 -microglobulin chain. In the complex 2bsv, missing backbone atoms were detected. All three problems could be solved by manually removing the criticized part of the complex (which was not interacting with the peptide anyway). Only for complex 1hhi did SCit crash without giving a reason.

For further testing the reliability of the tools, we substitute the side chains of all peptides with the sequence WWWWWWWW, which is of course spatially not fitting to the different binding grooves. In this case the side chains of the MHCs are considered flexible to allow adjustment of the receptor. The target of these runs is to test the reaction of the tools to a situation with many spatial clashes of the rotamers. SCit, IRECS, SPDBV, as well as SCATD could generate a structure, on average, faster than a minute. Only SCWRL took an average of ~ 14 min to finish, probably caused by a too big cluster size (for details, refer to the SCWRL publication by Canutescu et al. [2003]). On the complex with PDB identification 1b0g, the runtime was ~ 30 min. We suppose that there may also be complexes where the runtime is even longer. This behavior has pros and cons: On the one hand, one needs to implement a supervising program to discard a badly converging run after a certain timeout, but on the other hand, this timeout may be used as an additional way of discriminating binding peptides from nonbinding ones.

Figure 2. Dihedral angles: Chi1 and Chi2 prediction accuracy in percentage over all amino acids under the assumption that a certain threshold between 0° and 180° is chosen. A residue is classified as correct if its angle is below the current threshold. Residues that do not have a Chi1 and/or Chi2 angle are excluded.

Runtime

In order to characterize the runtime performance, we adopted the average runtime over the 29 complexes to build the peptide. Due to its web-based design, SCit takes the longest time for side chain positioning, but these estimates are biased because the Internet connection influences calculation time. The other tools are available offline and widely differ in runtime. IRECS (Hartmann et al. 2007) benefits from its fast heuristic implementation, but SCATD is definitely the fastest algorithm (Fig. 3). SPDBV suffers from its graphical user interface, which has to be loaded repeatedly. Measurements are done on a two-core, 2.80-GHz standard desktop machine using 2 GB of RAM and an openSuse operating system.

Usability

SCWRL, SCATD, and IRECS offer a console-based user interface, which makes automated usage easy. The parameterization is complaisant and intuitive. SCit is web-based and offers, according to our intuitive judgment, the easiest to use graphical-user interface. Batch jobs have to be carried out by self-developed scripts using the web interface. The side chain substitution feature of SPDBV is integrated into a graphical user interface. Alternatively, the integrated scripting language of SPDBV can be used to automate batch jobs.

Discussion

From our point of view, no tool can come out as the overall winner for the area of peptide modeling for MHC binding. Every tool has advantages and disadvantages,

and users need to decide which property is most important for them. The lowest RMSD was clearly achieved by SCWRL. In that case, our results agree well with a previous study (Wallner and Elofsson 2005) that showed a superior performance of SCWRL in side-chain prediction compared with other homology modeling tools. On the other hand, SCWRL suffers from extreme runtimes in cases where a peptide clearly does not fit in the binding groove of the MHC. This could be explained by too many active residues building too big clusters to be solved in reasonable time (for details refer to the original publication by Canutescu et al. [2003]). This can be a problem for the automated screening of large sets of peptides but can also be seen as an advantage to discriminate fitting from nonfitting peptides in docking studies. A good balance between very fast runtime and still good RMSD is provided by SCATD. Apart from that, SCit is clearly the easiest tool to handle and applicable also for scientists with non-information technical backgrounds. Very surprising was the overall good performance of the aged tool SPDBV.

Materials and Methods

The aim of our study was to evaluate the performance of several tools for side-chain prediction within the grooves of peptide-MHC complexes. For that purpose, the HLA-A2 allele is chosen. The workflow of the benchmark is illustrated in Figure 4, and the single steps are described in detail in the following sections.

Part of this benchmark is only the substitution of the side chains. Mainly, this was done for two reasons: (1) It would be hard to calculate a correct backbone without the knowledge of the appropriate way to position side chains, and (2) pure side-chain prediction is very fast (see Fig. 3) compared with the full threading of peptides (Bordner and Abagyan 2006; Bui et al. 2006). One would argue that not changing the backbone would probably lead to wrong results, but it has been shown that different peptides binding to the same MHC molecule share similar backbone structures (Madden et al. 1992). We superimposed the β -sheet floor of the peptide-binding groove (as done by Rudolph et al. [2006]) to align the peptides. The calculated average RMSD between the average peptide structure and the single peptides is 0.93 Å. These superimposed peptides can be seen in Figure 5. Full threading procedures achieved an average backbone RMSD of 1.09 Å on the HLA-A*0201 allele (Bordner and Abagyan 2006). This should not devalue full threading procedures but show that working with the original backbones was valid in our case. Only this approach would enable structural high-throughput screening. In further studies, we will show the application of different scoring functions in discriminating binding peptides from nonbinding ones via this fast side-chain-only modeling approach. These results will be published elsewhere.

Data acquisition and preparation

The first step was to select as many complexes as possible for the test set. A search in the PDB (Bergman et al. 2000) was

Figure 3. Average runtime over all complexes.

Figure 4. Workflow of the benchmark.

performed, and the resulting complexes were downloaded and visually inspected to check if they really contained what their label indicates: a single HLA-A2 molecule presenting a nonameric peptide. T-cell-receptor-peptide-MHC (TCRpMHC) complexes were excluded because at the current state we were only interested in the interaction between peptide and MHC. Simply removing the TCR was not sufficient because the side chains of the peptide in the direction of the TCR would then remain in an unnatural state. Moreover, double molecules and co-receptors like CD8 were removed. Only peptide and MHC remain in the PDB file. The effect of removing domains apart from the binding groove before doing calculations is controversial in the literature. Some investigators are of the opinion it affects the results (Wan et al. 2004) and some are not (Rognan et al. 1992; Gregoire et al. 1996; Toh et al. 2000; Zoete and Michielin 2007). We decided to leave them where they are because in contrast to molecular dynamics simulations, larger complexes will not dramatically increase the runtime. All complexes were checked by a script if they really contain the amino acid sequence of HLA-A2. Some complexes contained a single or more mutations. If mutations were detected near (~ 10 Å) the binding groove, these complexes were removed from the test set. Only complexes presenting distinct peptides (different amino acid sequences) were used to avoid any bias in the benchmark. Finally, those complexes were selected that carry nonamer peptides. The majority of peptides binding to the HLA-A2 molecules were nonamers. But also peptides with a length of 8, 10, or 11 amino acids may bind and were reported in the PDB. Due to the structural properties of the MHCI binding groove, peptides longer than 9 amino acids usually have to accommodate their structure by kinking their backbone or binding outside the groove with the carboxyl terminus. This, however, increased the degrees of freedom for the side chains and made more states in phase space accessible. Therefore, only nonamers were considered in this study. The remaining complexes were visually confirmed again, and we ended up with 29 pMHC complexes obtained from the PDB identification: 1b0g, 1eey, 1eez, 1hhg, 1hhi, 1ilf, 1i7r, 1i7t, 1i7u, 1im3, 1jht, 1p7q, 1qew, 1qr1, 1s8d, 1s9w, 1s9x, 1s9y, 1t1w, 1t1x, 1t1y, 1t1z, 1t20, 1t21, 1tvb, 1tvh, 2bsv, 2git, and 2gtw.

Removing the side chains

Although most tools predict the positions of the side chains without cheating by inspecting the original position, all side chains of a peptide were removed initially to ensure reliability. The whole MHC and the backbone of the peptide remained as they were. The backbone of the peptide can be considered rigid for two reasons: (1) The structures of backbones and the MHC molecules are already in the correct position because in the original structure exactly the same side chains were included, and also (2) different peptides binding to the same MHC molecule would share a similar backbone structure (Madden et al. 1992). The results of this step were pMHC complexes having backbones and side chains for the MHC, whereas only backbones were for the peptide.

Remodeling of the side chains

Five different tools (SCWRL [Canutescu et al. 2003], SCATD [Xu 2005], SPDBV [also known as DeepView] [Guex and Peitsch 1997], SCit [Gautier et al. 2004], and IRECS [Hartmann et al. 2007]) were used to predict the structure of the previously removed side chains of the peptides. The tools were parameterized to consider the whole MHC as rigid since its backbone as well as the side chains of the MHC were already in the correct position because they had been recorded presenting exactly the same peptide which was re-substituted. The results of this step were five sets of predictions containing 29 complexes, respectively.

Evaluation of the predictions

A standard technique to compare two similar complexes is the RMSD being defined for three-dimensional structures as:

Figure 5. Peptides/MHC complex: The β -sheet floors of the binding grooves were superimposed to align the 29 peptides. The α_1 helix of the MHC is hidden to allow direct insight into the binding groove. White indicates MHC α -chain and β_2 -microglobulin.

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^{i=n} (X_{ix} - Y_{ix})^2 + (X_{iy} - Y_{iy})^2 + (X_{iz} - Y_{iz})^2}$$

where n is the number of atoms, i is the current atom, X is the first complex, and Y is the second one.

In this study, the `g_rms` function of GROMACS (Berendsen et al. 1995) was used to calculate the RMSD between two complexes.

Additionally, we used the dihedral angles (χ_1 , χ_2) to show the correctness of the predictions. The `g_chi` function of GROMACS was used to calculate the angles. To decide if a predicted angle was correct or not, usually an arbitrary threshold was used. From our point of view, it is more convenient to systematically screen all possible thresholds instead of using a single one; hence this may hide strengths and weaknesses of tools. A residue is classified as correct if the difference between the known angle from the X-ray structure and the predicted one is smaller than the current threshold.

Acknowledgment

This work is supported in part by the Austrian Grid Project of the Austrian Ministry of Education, Science and Culture (contract no. GZ 4003/2-VI/4c/2004).

References

- Berendsen, H.J.C., van der Spoel, D., and van Drunen, R. 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**: 43–56.
- Bergman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**: 235–242.
- Bordner, A.J. and Abagyan, R. 2006. Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins* **63**: 512–526.
- Bui, H.H., Schiewe, A.J., von, G.H., and Haworth, I.S. 2006. Structural prediction of peptides binding to MHC class I molecules. *Proteins* **63**: 43–52.
- Canutescu, A.A., Shelenkov, A.A., and Dunbrack Jr., R.L. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**: 2001–2014.
- Gautier, R., Camproux, A.C., and Tuffery, P. 2004. SCit: web tools for protein side-chain conformation analysis. *Nucleic Acids Res.* **32**: W508–W511. doi: 10.1093/nar/gkh388.
- Gregoire, C., Lin, S.Y., Mazza, G., Rebai, N., Luescher, I.F., and Malissen, B. 1996. Covalent assembly of a soluble T cell receptor-peptide-major histocompatibility class I complex. *Proc. Natl. Acad. Sci.* **93**: 7184–7189.
- Guex, N. and Peitsch, M.C. 1997. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**: 2714–2723.
- Hartmann, C., Antes, I., and Lengauer, T. 2007. IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci.* **16**: 1294–1307.
- Madden, D.R., Gorga, J.C., Strominger, J.L., and Wiley, D.C. 1992. The three-dimensional structure of HLA-B27 at 2.1 Å resolution suggests a general mechanism for tight peptide binding to MHC. *Cell* **70**: 1035–1048.
- Rognan, D., Zimmermann, N., Jung, G., and Folkers, G. 1992. Molecular dynamics study of a complex between the human histocompatibility antigen HLA-A2 and the IMP58-66 nonapeptide from influenza virus matrix protein. *Eur. J. Biochem.* **208**: 101–113.
- Rudolph, M.G., Stanfield, R.L., and Wilson, I.A. 2006. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**: 419–466.
- Toh, H., Kamikawaji, N., Tana, T., Muta, S., Sasazuki, T., and Kuhara, S. 2000. Magnitude of structural changes of the T-cell receptor binding regions determine the strength of T-cell antagonism: Molecular dynamics simulations of HLA-DR4 (DRB1*0405) complexed with analogue peptide. *Protein Eng.* **13**: 423–429.
- Tong, J.C., Tan, T.W., and Ranganathan, S. 2007. Methods and protocols for prediction of immunogenic epitopes. *Brief. Bioinform.* **8**: 96–108.
- Wallner, B. and Elofsson, A. 2005. All are not equal: A benchmark of different homology modeling programs. *Protein Sci.* **14**: 1315–1327.
- Wan, S., Coveney, P., and Flower, D.R. 2004. Large-scale molecular dynamics simulations of HLA-A*0201 complexed with a tumor-specific antigenic peptide: Can the $\alpha 3$ and $\beta 2m$ domains be neglected? *J. Comput. Chem.* **25**: 1803–1813.
- Xu, J. 2005. Rapid side-chain prediction via tree decomposition. *RECOMB* **3500**: 423–439.
- Zoete, V. and Michielin, O. 2007. Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein-protein association using MM-GBSA: Application to the TCR-p-MHC complex. *Proteins* **67**: 1026–1047.