Improving the discrimination of near-native complexes for protein rigid docking by implementing interfacial water into protein interfaces

Su, Tran To Chinh

2014

Su, T. T. C. (2014). Improving the discrimination of near-native complexes for protein rigid docking by implementing interfacial water into protein interfaces. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/62097>

<https://doi.org/10.32657/10356/62097>

Improving the Discrimination of Near-Native Complexes for Protein Rigid Docking by Implementing Interfacial Water into Protein Interfaces

Su Tran To Chinh

**A thesis submitted to the Nanyang Technological University
In partial fulfillment of the requirement for the degree of
Doctor of Philosophy**

**School of Computer Engineering
2014**

Acknowledgements

Had I been alone on this route, I would have lost or stuck in some corner along it. I am blessed having people around who are there caring, encouraging, and helping me keep my feet on the ground.

Firstly I am grateful to my supervisor, Prof. Kwoh Chee Keong for his patient guidance and fatherly advice and encouragement. I thank Nguyen Thuy Diem, Ouyang Xuchang, Stephanus Daniel Handoko, who always offer help and share useful discussion, and Mr. Tan Sing Yau for his assistance in resolving every lab issues.

I appreciate the friendship with Divya Jain and Shamima Banu who accompany me in any walk, share the talks and the laughter, and with Fanny Zhang who was a wonderful roommate.

My thanks deeply go to my family for always thinking of me, especially to Sammy for such a caring brother. I thank Raymond, my little nephew for just being there and smiling. And finally I thank both my Mom and god Mom for everything.

Table of contents

Acknowledgements	i
Table of contents	i
Abstract	v
List of Figures	vii
List of Abbreviations	xiii
Abbreviations of amino acids	xiv
Chapter 1	1
Introduction	1
1.1. Background	1
1.1.1. Protein-Protein Docking: an overview	1
1.1.2. Summarized current issues of protein-protein docking approach in improving the results	4
1.2. Motivation and Objectives	6
1.3. Overview of this Thesis	8
Chapter 2.....	11
Various re-ranking techniques to improve the initial docking results	11
2.1. Re-ranking algorithm in protein docking: ZRANK	12
2.1.1. Overview	12
2.1.2. Comparison with the initial docking ZDOCK.....	14
2.2. Re-ranking using Generalized Born model: F ² Dock.....	17
2.2.1. Overview	17
2.2.2. Comparison with the initial docking ZDOCK.....	19
2.3. Concluding remark	20
Chapter 3.....	21
Docking analysis with implicit and explicit presence of water molecules	21

3.1. Applications of protein docking in vaccine discovery for 2009 H1N1 influenza A virus (<i>implicit water</i>).....	22
3.1.1. A possible mutation that enables the H1N1 virus to escape antibody recognition ..22	
3.1.1.1. Background.....	22
3.1.1.2. Overview of the docking algorithm ZDOCK.....	22
3.1.1.3. Docking experiments and results	25
3.1.1.4. Summary and concluding remark.....	28
3.1.2. Implications for T-cell Receptor (TCR) cross-recognition and vaccine development for 2009-H1N1 and 2004-H5N1 influenza virus HLA-B*4405-restricted HA epitope candidates.....	28
3.1.2.1. Background.....	28
3.1.2.2. Overview of the docking algorithm PIPER.....	29
3.1.2.3. Docking experiments and results	31
3.1.2.4. Summary and concluding remark.....	34
3.1.3. Concluding remark for this section.....	34
3.2. Protein docking analysis of HIV-1 Gag and Protease interactions (<i>explicit water</i>).....	35
3.2.1. Background	35
3.2.2. Overview of the docking algorithm HADDOCK.....	36
3.2.3. Gag-Protease interactions	39
3.2.4. Summary and concluding remark	39
3.3. Conclusion for this chapter	40

Chapter 441

Modeling studies of various potentials on different physicochemical properties of protein complexes41

4.1. An overview of potential terms used in scoring functions of the initial docking algorithms.....	42
4.1.1. Shape complementarity (SC).....	42
4.1.2. Electrostatics (ELEC).....	43
4.1.3. Atomic Contact Energy (ACE).....	44
4.1.3.1. Conventional ACE	45
4.1.3.2. Interface Atomic Contact Energy (IFACE).....	46

4.1.4. Desolvation Free energy (DE)	47
4.1.5. Pairwise interactions with Decoys as Reference States (DARS)	49
4.1.6. Comparison between atom-level (ACE) and residue-level (DARS) potentials.....	50
4.2. Comparisons of various scoring functions.....	52
4.2.1. Data	52
4.2.2. A comparison of docking performances among different scoring functions.....	54
4.3. Rigid docking results refined by Molecular Dynamics simulation	57
4.3.1. Molecular Dynamics (MD) simulation gives better discrimination of near-native structures from false positive ones.....	57
4.3.2. Molecular Dynamics simulation allows more flexibility	58
4.4. Conclusion for this chapter	62
 Chapter 5.....	65
Improving the discrimination of near-native complexes for protein rigid docking by implementing interfacial water into protein interfaces	65
5.1. Introduction	66
5.2. Process of the implementation	68
5.2.1. Data collection and preprocessing	68
5.2.2. Implementation of interfacial water in the proposed energy-based scoring function.....	69
5.2.2.1. Combinations of pairwise shape complementarity, electrostatics, and desolvation	70
5.2.2.2. Water-mediated contacts and the water effect on the protein interfaces	71
5.2.2.3. Free energy change of the protein interfaces $\Delta G_{\text{interface}}$	75
5.2.3. Estimation of weights w_1, w_2, w_3, w_4	76
5.3. Results and Discussion	77
5.3.1. Optimal weights w_1, w_2, w_3, w_4 of the proposed scoring function	77
5.3.1.1. Reflection of weights on different physicochemical properties	77
5.3.1.2. Interdependencies among the components	78
5.3.2. The proposed water-implemented method improves results of protein rigid docking.....	79
5.3.2.1. Comparisons of the IFACEwat and ZDOCK3.0.2 and ZRANK.....	79
5.3.2.2. Comparisons of the IFACEwat and F ² Dock	91
5.4. Conclusion for this chapter	97

Chapter 6	98
Evaluation of the proposed interfacial water-implemented scoring function.....	98
6.1. The IFACEwat improved the first Hit rank for either with- or without water-containing structures	99
6.2. Is the IFACEwat over-fitting to the data?	102
6.3. Free energy change ($\Delta G_{\text{interface}}$) plays the most important role in the IFACEwat scoring function	103
6.4. The IFACEwat maintains sufficiently the computational efficiency yet generate improved ranking of the first hit.....	106
Chapter 7	110
Conclusion and future work	110
7.1. Summary and Conclusion.....	110
7.2. Future Work	112
Author's publications.....	114
Appendix	115
A1. A possible mutation that enables the H1N1 virus to escape antibody recognition	115
A2. Implications for TCR cross-recognition and vaccine development for 2009-H1N1 and 2004-H5N1 influenza virus HLA-B*4405-restricted HA epitope candidates	122
A3. Scores of other false positive and incorrect docked conformations	139
A4. Torsion angles (ω, ϕ, ψ) of the 10-mer peptide backbone from various docked complexes	140
A5. Statistics of interfacial water of complexes extracted from benchmark 4.0	141
A6. Re-rankings results that are equivalent with or better than the F2Dock (full table). Results of F2Dock are done by Chowdhury et al.	146
A7. Details of software used in this thesis.....	151
References	153

Abstract

Protein-protein docking is an *in silico* method to predict the formation of protein complexes. Due to the limited computational resources, the protein docking approach has been originated and developed under the assumption of rigid docking, in which one of the two protein partners remains rigid during the protein associations. The rigid docking has successfully predicted structures of various protein complexes, but often failed if the proteins acquire conformational changes or are driven by influences of other factors (e.g. solvent) while interacting.

Formulation of the initial rigid docking contains two main stages: (1) *searching* for all possible surface matches on a rotational and translational sampling space and (2) *ranking* those possible solutions to distinguish the correct predictions by locating them on the top high ranks. To obtain better results of the protein rigid docking, one can improve it by optimizing the solutions in the search space or by developing more effective ranking methodology to discriminate the correct predictions from the incorrect or false positive ones. However, while development and improvement are reported in the *searching* stage, it seems to-date that most initial docking algorithms find it difficult or even fail to locate successfully the correct predictions apart from the others in the *ranking* stage, especially for Antigen/Antibody complexes.

To tackle this issue, a new energy-based scoring function is proposed in this research, namely IFACEwat, to re-rank the results of an initial rigid docking algorithm and therefore further improve the discrimination of the near-native structures from the other false positives. Unlike other re-ranking techniques, in which the solvent effect is ignored or implicitly presented, the IFACEwat implements interfacial water into the protein interfaces and explicitly reflects the water influences on the protein interactions.

The IFACEwat was implemented based on the interface Atomic Contact Energy (IFACE) of the initial rigid docking algorithm ZDOCK3.0.2 and the derived energies of water-involved interactions at the protein interfaces. The IFACEwat therefore not only takes advantages of shape complementarity from the initial rigid docking algorithm for protein recognitions but also accounts for the water-mediated contacts during the protein associations.

Evaluated for various types of protein complexes, the IFACEwat both increases the numbers of the near-native structures and improves their ranks (i.e. most of them ranked at the top-1) as compared to the original rigid docking. In fact, it achieves a success rate of 83.8% for

Antigen/Antibody complexes, which is 10% better than ZDOCK3.0.2. As compared to another re-ranking technique ZRANK, the IFACEwat obtains success rates of 92.3% (8% better) and 90% (5% better) respectively for medium and difficult cases. When comparing with the latest published re-ranking method in F²Dock, it performs equivalently well or better for several Antigen/Antibody complexes.

In conclusion, with the inclusion of interfacial water, the IFACEwat improves significantly the results of the initial rigid docking, especially for Antigen/Antibody complexes. The improvement is achieved by explicitly accounting the contribution of water during the protein interactions, which was ignored or not fully presented by the initial rigid docking and other re-ranking techniques. In addition, the IFACEwat maintains the sufficient computational efficiency of the initial docking algorithm, yet improves the ranks as well as the number of the near-native structures found.

List of Figures

Figure 1-1	Illustration of protein docking of TCR-HLAp complex [PDB:3DXA]	2
Figure 1-2	Overview of protein-protein docking process	4
Figure 1-3	Limitation of ranking techniques of conventional initial rigid docking algorithms	6
Figure 1-4	Development of the re-ranking technique to improve the docking results	8
Figure 2-1	Re-scoring and re-ranking flowchart of ZRANK	13
Figure 2-2	Illustration of comparisons between ZRANK and ZDOCK on the system level.	16
Figure 2-3	Detailed comparisons of ZRANK and ZDOCK3.0.2 using the benchmark 4. Success rate is defined as percentage of complexes that obtain at least 1 near-native structure in the top 100 predictions.	16
Figure 2-4	Re-scoring and re-ranking flowchart of F ² Dock-Generalized Born	17
Figure 2-5	Schematic presentation of atomic interactions in implicit solvent following the Generalized Born model, adapted from (Case, III <i>et al.</i> 2005).	18
Figure 2-6	Comparisons of F ² Dock and ZDOCK3.0.2 using the benchmark 4. Success rate is defined as percentage of complexes that obtain at least 1 near-native structure in the top 100 predictions.	19
Figure 3-1	Docking experiment of the epitope-containing HA to the 2D1-Antibody	26
Figure 3-2	Influence of His ¹³⁰ side-chain on antibody binding of residue Lys ¹⁵⁷ in 2D1-epitope-containing HA complex (A) and 2D1-1918 HA protein complex (B). <i>Top panel:</i> differences in binding regions of epitope-containing HA and the 1918 HA protein is shown above. Matching residues are marked by asterisks. Mismatches in binding regions (131-133) are colored in red. The B-cell epitope is underlined. <i>Middle panel:</i> atom-based residue coloring: C (white), O (red), and N (blue). Backbones are presented by sticks. Side-chains are shown as lines. Interactions are presented as yellow-dotted lines. A. Internal contacts of Ser ¹³² and His ¹³⁰ reduce the distance of His ¹³⁰ to Lys ¹⁵⁷ to 1.8 Å, thereby blocking binding of the antibody. B. Lys ¹⁵⁷ interacts with Asn ⁵⁴ of the antibody. The distance His ¹³⁰ is of 4.7 Å. <i>Bottom panel:</i> Different binding orientations of the 2D1 antibody to the epitope-containing HA protein (A) and	

to the 1918 influenza HA (B). The bound complexes of antibody and HA monomer are shown as surface presentation. HA1 and HA2 domains are colored in green and cyan. The 2D1 antibody chains are shown in yellow (heavy chain) and orange (light chain). The predicted B-cell epitope is colored in red.	27
Figure 3-3 Docking experiment of the T-cell epitope to the HLA-B*4405 and the DM1-TCR	32
Figure 3-4 Examples of conformational TCR adaption to differently recognize both embedded and solvent-exposed epitopes.	34
Figure 3-5 Illustration of Accessible Surface Area calculation, adapted from (WHAT-IF 2013)	38
Figure 3-6 HADDOCK – the solvated docking protocol (Dijk and Bonvin 2006).	38
Figure 3-7 Different structural orientations of Protease (ligand) docked to Gag (receptor) at the same cleavage site. Protease generated from HADDOCK (green) interacts with the buried cleavage site (red) of Gag (white surface) while the docked orientation of the Protease resulted by ZDOCK (blue) cannot access to this buried site.	39
Figure 4-1 Schematic presentation of surface recognition by Katchalski-Katzir <i>et al.</i> While the left panels <i>a</i> and <i>c</i> show presentations of no interaction and penetration respectively, the panels <i>b</i> and <i>d</i> show different surface complementarity of the two proteins. The plot on the right depicts the surface complementarity according to the SC algorithm.	43
Figure 4-2 Docking performance using DARS (green) and ACE (red) for Enzyme/Inhibitor complexes according to percentage of obtained number of near-native structure found (y-axis) on 2000 docked conformations (x-axis).	51
Figure 4-3 Docking performance using DARS (green) and ACE (red) for Antigen/Antibody complexes according to percentage of obtained number of near-native structure found (y-axis) on 2000 docked conformations (x-axis).	52
Figure 4-4 HLA-B*4405/EBV-peptide complex [PDB:3DX8]. HLA chains A, B, and the EBV peptide are presented in green, cyan, and red, respectively in the front view (left) and the side view (right).	53
Figure 4-5 Docking performances (<i>without refinement</i>) of different scoring functions in terms of percentage of number of near-native structures found in 2000 docked conformations	54
Figure 4-6 Docking performance of various docking scoring functions SC/ELEC, PSC,	

	PSC+ELEC+DE _{ACE} , PSC+ELEC+DE _{IFACE} , and DARS in terms of corrects and false positives for 2000 generated complex structures.	57
Figure 4-7	Structural presentations of the 10 lowest energy structures. The complexes 1 and 3 show correct bindings of their peptides (red and blue, respectively) to the binding groove, while the peptides of the other 8 complexes (in other colors) are located outside of the groove. The complexes are presented in cartoons in which the HLA structure is in gray.	59
Figure 4-8	The Asn ⁷⁷ -Phe ¹⁰ contact that occurs only in the MD-refined HLA/peptide complex. The 5ns, 10ns, and 20ns conformations are shown in red, green, and blue lines, respectively. The PSC and PSC+ELEC+DE _{IFACE} conformations are perfectly superimposed and presented in purple stick. The dotted lines represent the polar contact between the Asn ⁷⁷ (HLA) and Phe ¹⁰ (peptide) and the distances are shown. There is no such the contact in both the PSC and PSC+ELEC+DE _{IFACE} complexes. (A) and (B) have the same color scheme.	61
Figure 4-9	A closer view of the Asn ⁷⁷ -Phe ¹⁰ contact in the HLA binding groove. The residues Asn ⁷⁷ and Phe ¹⁰ of the PSC, PSC+ELEC+DE _{IFACE} , and MD (5ns, 10ns, 20ns) complexes are shown in orange, purple, red, green, and blue line presentations. The original HLA-B*4405/peptide bound complex [PDB:3DX8] is presented in yellow stick and ribbon. The Asn ⁷⁷ of the PSC and PSC+ELEC+DE _{IFACE} conformations are totally fitted with the original one.	61
Figure 4-10	Combination of PSC+ELEC+DE _{IFACE} and MD shows better performance than the combined PSC and MD.	62
Figure 5-1	Process of generating 2000 predictions to calculate the Score _{IFACE}	70
Figure 5-2	Flow chart of the process of estimating the contributions of each potential term in the proposed scoring function to improve rankings for protein docking results using interfacial water.....	73
Figure 5-3	Angle and distance constraints to define a water-mediated Hydrogen bond according to the solvated rotamer library. <i>d</i> is the distance between the water (W) oxygen and polar (acceptor A/donor D) atoms. H and AB are Hydrogen and Acceptor-Base atoms respectively.	75

Figure 5-4	Process of estimating the weights w_1 , w_2 , w_3 , w_4 . In this process, the free energy change term ($\Delta G_{interface}$) is expected in significant contributions as observed in section 5.2.3.3. The input sets of the 4 potential terms are derived from 318,000 predictions (i.e. 159 cases x 2000 predictions).....	76
Figure 5-5	Performances of ZDOCK3.0.2, ZRANK, and IFACEwat method in terms of success rate of different complex types	86
Figure 5-6	Performances of ZDOCK3.0.2, ZRANK, and IFACEwat method in terms of numbers of hits found (against total numbers of hits shown in brackets) for each type of complexes.	87
Figure 5-7	Best RMS of conformations in the top 10, 100, and 1000 ranks found by the IFACEwat against ZDOCK (left) and ZRANK (right). Each dot represents a protein complex case according to difficulty level (easy, medium , and difficult).....	89
Figure 5-8	Best RMS of conformations in the top 10, 100, and 1000 ranks found by the IFACEwat against ZDOCK (left) and ZRANK (right). Each dot represents a protein complex case according complex types: Antigen/Antibody (A , AB), Enzyme/Inhibitor (E), Others (O).....	90
Figure 5-9	Improvement of the IFACEwat and F2Dock against the initial docking ZDOCK. The success rate is defined the same as in Figure 5-5. It is observed that the IFACEwat performs better for Antigen/Antibody (A/AB) complexes while the F2Dock is better for Enzyme/Inhibitor (E).....	96
Figure 6-1	Presence of the water-mediated Hydrogen bonds at the interfaces of protein complexes, the first hit of which is improved in ranking. The complex interfaces (in grey shaded surfaces) contain interfacial water molecules (red spheres) that mediate the hydrogen bonds (black dotted lines) bridging the interactions between the receptor (green) and the ligand (cyan).....	100
Figure 6-2	Advantages of the IFACEwat re-ranking over the initial docking ZDOCK	106
Figure 6-3	Receiver Operating Characteristic (ROC) curves to illustrate improvement of the IFACEwat over ZDOCK3.0.2 in terms of ranking the first hit as top-1	109

List of Tables

Table 1-1	Percentages of correct predictions versus false positives for some docking algorithms	5
Table 4-1	Atomic Contact Energy e_{ij} value (in energy unit of $RT/(q_i/2))^*$ of 18 atom types, adapted from Zhang <i>et al.</i>	48
Table 4-2	Percentage of hits obtained in the top 2000 docked Enzyme/Inhibitor complexes using DARS (PIPER) and ACE (ZDOCK2.3).	51
Table 4-3	Percentage of hits obtained in the top 2000 docked Antigen/Antibody complexes using DARS (PIPER) and ACE (ZDOCK2.3)	51
Table 4-4	The first 10 docked conformations used as illustrations of the best, correct, and false positive results of the docking methods.....	56
Table 4-5	Ten optimal structures obtained after 5ns MD refinement of the PSC results	59
Table 4-6	Overall summaries about limitations and performances of various scoring functions. Performances were evaluated for 2000 generated conformation.	64
Table 5-1	van der Waals parameters (based on AMBER force field) used to calculate the Lennard-Jones repulsive energy in this project	74
Table 5-2	Interdependencies among components for Antigen/Antibody (<i>bottom-left triangle</i>) and Enzyme/Inhibitor (<i>up-right triangle</i>) complexes	79
Table 5-3	Overall results of ZDOCK3.0.2, ZRANK, and the IFACEwat in terms of numbers of near-native structures (<i>hits</i>) found and the rank of the first hit. Cases which are in bold indicate that IFACEwat method performs equivalently with or better than both of the others.	80
Table 5-4:	Different implementation between the IFACEwat and F ² Dock in determining the near-native structures (hit).	91
Table 5-5:	Some re-rankings results that are equivalent with or better than the F ² Dock (full table can be found in the Appendix A6). The results of F ² Dock are performed by Chowdhury <i>et al.</i>	92
Table 6-1	The IFACEwat still improves both the number of hits and the rank of the first hit for	

	either protein complex cases of with- or without interfacial water.....	101
Table 6-2	The IFACEwat still maintain the ability of locating the first hit at top-1 for most of unseen data.....	103
Table 6-3	Decrease of the number of near-native structures for all types of protein complexes Antigen/unbound-Antibody (A), Antigen/bound-Antibody (AB), Enzyme/Inhibitor (E) and others (O): ↓ (less than 5% hits drop), ↓↓ (less than 30% hits drop), ↓↓↓ (noticeably drop less than 50% hits), ↓↓↓↓ (significantly drop less than 70% hits), ↓↓↓↓↓ (dramatically drop more than 95% hits), - (unchanged). Experiment #0 shows the results of number of hits obtained from the full set of weights for each complex type and is used as the reference.	104
Table 6-4	Real running time T_R of the IFACEwat in re-ranking some of the protein complexes and the estimated formula of the MD time for the structural refinement stage of the IFACEwat results as compared to ZDOCK.....	107

List of Abbreviations

A/AB	Antigen-unbound Antibody/Antigen-bound Antibody complex
ACE/ACE	Atomic Contact Energy/Potential
AMBER	Assisted Model Building with Energy Refinement
CAPRI	Critical Assessment of Predicted Interaction
E	Enzyme/Inhibitor complex
DARS	Decoys As the Reference States
DE	Desolvation Energy
ELEC	Electrostatics/Electrostatic interaction
FFT	Fast Fourier Transform
GB	Generalized Born
GSC	Grid-based Shape Complementarity
HA	Hemagglutinin
HLA	Human Leukocyte Antigen
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
O	Other complexes
PDB	Protein Data Bank
PSC	Pairwise Shape Complementarity
RMS/RMSd	Root Mean Square deviation
SC	Shape Complementarity
TCR	T-cell receptor

Abbreviations of amino acids

Ala	Alanine (A)
Arg	Arginine (R)
Asn	Asparagine (N)
Asp	Aspartic Acid (D)
Cys	Cysteine (C)
Gln	Glutamine (Q)
Glu	Glutamic Acid (E)
Gly	Glycine (G)
His	Histidine (H)
Ile	Isoleucine (I)
Leu	Leucine (L)
Lys	Lysine (K)
Met	Methionine (M)
Phe	Phenylalanine (F)
Pro	Proline (P)
Ser	Serine (S)
Thr	Threonine (T)
Trp	Tryptophan (W)
Tyr	Tyrosine (Y)
Val	Valine (V)

Chapter 1

Introduction

1.1. Background

1.1.1. Protein-Protein Docking: an overview

Proteins involve in all cellular processes and play significant roles in multiple biological functions (Rodwell and Kennelly 2003, Ritchie 2008), most of which are only performed when proteins form into complexes (Jones and Thornton 1996). For an example, eukaryotic translation process is only initiated by the assembly of two subunits of a the ribosome protein. Various associations of proteins driven by diverse protein interactions subsequently form a huge number of protein complexes. However, while knowledge of a lot of protein interactions and molecular pathways have been discovered experimentally, a very limited number of structures of known protein complexes could be elucidated. In addition, due to the transient or obligatory associations, not every protein complex could be experimentally crystalized. Hence, predicting the formation of protein complexes has become a complement in studying protein interactions so far.

There are two types of protein complex: homo– and hetero complexes (Jones and Thornton 1996). Homo complex consists of same protein components such as homodimers, trimers, etc. Hetero complexes are formed by different proteins, e.g. enzyme/inhibitor or antigen/antibody complexes. Experiments have been developed to investigate the protein interactions such as high throughput Yeast Two Hybrid (Y2H) (Walhout and Vidal 2001), Tandem Affinity Purification Mass Spectrometry (TAP-MS) (Kaiser, Meierhofer *et al.* 2008), Protein Microarray (MacBeath and

Schreiber 2000), and Phage Display (Smith 1985). Although these *in vivo* large-scale experiments can account for high-throughput data and detect protein physical interactions (Shoemaker and Panchenko 2007), accuracy is the main limitation; for an instance, among these methods, the Y2H obtains the least reliability—around 50% (Sprinzak, Sattath *et al.* 2003). Therefore, *in silico* methods are complementarily developed to predict protein-protein interactions and formations of protein complexes (Shoemaker and Panchenko 2007) before one can conduct the protein binding experiments to aim for both time and cost efficiency. One of the most popular computational methods is *protein docking*.

Protein docking predicts structures of protein complexes. Docking algorithm simulates the protein-protein association by reconstituting the complex from two component molecules (Cherfils and Janin 1993, Ritchie 2008) and attempts to find the best matching between them: a receptor and a ligand (Figure 1-1).

Figure 1-1 Illustration of protein docking of TCR-HLA complex [PDB:3DXA]

Protein-protein docking simulates molecular recognition between two protein molecules (Halperin, Ma *et al.* 2002). There are two docking types: rigid-body docking and flexible docking, based on substantial conformational changes occurring during the complex formation. If parameters—such as bond angles, bond length, and torsion angles of the components—keep updated and modified during process of complex generation, it is known as flexible docking. However, due to the sizes of the protein molecules, it is very challenging and possibly impractical to perform entire conformational search due to the large number of degrees of freedom. Hence, most of current protein-protein docking algorithms utilize rigid-body method (Halperin, Ma *et al.* 2002). In this thesis, rigid-body docking is therefore assumed unless otherwise specified.

Rigid docking is applied to compromise between accuracy and limitations of current computational expenses in the docking field. The two-stage docking process of *searching* and *ranking* is

introduced in this approach (Figure 1-2). *Searching* stage generates a number of all possible docked conformations (or called “predictions”) of protein complexes from 6-degree-of-freedom searching space while *ranking* helps to distinguish the correct predictions from other incorrect and false positives (Halperin, Ma *et al.* 2002, Smith and Sternberg 2002).

Initially, achieving an effective search strategy in protein docking requires great efforts. Since there are various ways to bring two molecules together (due to three translational and three rotational degrees of freedom), a huge number of possibilities that grow exponentially with the sizes of the components could be generated (Cherfils and Janin 1993). Therefore, most of the current protein-protein docking algorithms treat the molecules as rigid bodies and apply correlation-based scoring functions in the ranking stage to rank the predictions. The top ranked candidates are assumed as the near-native structures, which then undergo a refinement process using molecular dynamics simulation to reach the optimal conformations.

Molecular dynamics (MD) simulation is conventionally applied for conformational refinement as it is able to offer complementary flexibility for the rigidly docked protein complexes during a time-course simulation. The MD optimizes the structures dynamically and searches for their global minimum energy values using various constraints and parameterized models (for examples, the impact of solvent is considered in periodic boundary condition, force constants are applied to restrain hydrogen-involved bonds, etc.). However, this step requires intensive computational resources and is thus very time-consuming, making it the last step and often separated or excluded from some docking packages. Moreover it still requires a compromise between speed and accuracy due to its energy evaluation process (Halperin, Ma *et al.* 2002), which adds a significant portion of computational time and complexity into the whole docking protocol. In fact, it is impractical for the large-scale system, e.g. macromolecules, as it may take days to months for all the calculations to reach the convergences.

Therefore, an efficient docking algorithm is required to optimize the searching speed but still cover relevant conformational space while maintaining effective discrimination of correct solutions from others within reasonable computational time. In fact, most of rigid-body docking programs only perform the *initial docking* and do not include the refinement step in their deliverable packages, but instead more efforts are put in improving the searching optimization or the ranking efficiency to obtain the predictions that are close to the native complex as much as possible. In this thesis,

the MD simulation is thus little mentioned, but still a part of MD results will be shown in the literature review to demonstrate the necessity of this step to obtain an optimal conformation of the protein complexes in terms of the structural flexibility.

Figure 1-2 Overview of protein-protein docking process

1.1.2. Summarized current issues of protein-protein docking approach in improving the results

Protein-protein docking has been widely applied to study protein interactions. However, due to limited computational resources against the sizes of the macromolecules, the docking technique is bound to rigid or semi-rigid, in which only some extent of flexibility is allowed (Halperin, Ma *et al.* 2002, Schnedman-Duhovny, Nussinov *et al.* 2006). This leads to a need of structural refinement to further optimize the conformations of the protein complexes using energy-based evaluations such as the molecular dynamics (MD) simulation (Halperin, Ma *et al.* 2002, Smith and Sternberg 2002) (an example of optimizing the docked protein complexes using MD simulation and how it improves the flexibility of the complex structures is demonstrated in Chapter 4). However, it is a major challenge as this refinement step is also computationally intensive.

While predicting formations of protein complexes, docking algorithms also need to discriminate the correct predictions from the incorrect and false positive ones using the ranking technique by putting the correct structures on the top ranks as high as possible from the others. Nonetheless, summarized results of some current rigid docking algorithms (see Chapter 4 for more details) indicate that while they obtain acceptable numbers of correct predictions, their limitation is obvious due to the high percentages of false positives (which are defined as numbers of wrong but highly ranked predictions in this report). For examples, percentages of correct predictions against false positives of some docking programs are shown in Table 1-1 below. Therefore, given the high computational cost required for the refinement step, improvement of ranking method is important to reduce as many as possible numbers of false positives that do not need to be refined before heading to the refinement steps.

Table 1-1 Percentages of correct predictions versus false positives for some docking algorithms

Docking algorithm	Correct prediction (%)	False positive (%)
FTDOCK	36.4	66
ZDOCK2.3	60.65	38
PIPER	52.85	29

$$\text{Correct prediction} = \frac{\# \text{ correctly bound structures of protein complex}}{\# \text{ total number of predictions (} N=2000 \text{)}}$$

$$\text{False positive} = \frac{\# \text{ wrong predictions in the top 100 ranks}}{\# \text{ top 100 predictions}}$$

Another effort to improve the docking results is to include the solvent effect to simulate the concept of protein associations: proteins interact in the solvent-surrounding environment (more detail of this example is shown in section 3.2). This involves large degrees of freedoms and again render the docking algorithms unaffordable for including water explicitly; so instead, they take into account the water effect via different potential terms, e.g. electrostatics, desolvation or solvation energy. The results are improved, as shown in an example of comparisons between *without*- and *with* the solvent effect of different docking programs, i.e. ZDOCK2.1 (46.65% correct) and ZDOCK2.3 (60.65% correct) respectively, yet numbers of false positives are still unacceptable high (35% and 38% respectively).

In general, if the refinement is a crucial step that every docking algorithm needs to undergo, it is necessary to improve the ranking method that can filter out the correct predictions from others.

Figure 1-3 Limitation of ranking techniques of conventional initial rigid docking algorithms

1.2. Motivation and Objectives

Since the first automated docking program (Wodak and Janin 1978) was developed, protein docking has been significantly improved (Pierce, Phillips *et al.* 2007). Different strategies such as Fast Fourier Transform correlation, geometric hashing, Monte Carlo technique, and molecular dynamics simulation are employed. However, due to the sizes of the macromolecules, it becomes more complicated and mostly computationally infeasible if all properties of protein complexes are taken into account. Hence, some assumptions are applied; for examples, the proteins need to be rigidly handled for six degrees of freedom from their translations and rotations, only certain flexibility of protein side chains or very small conformational changes upon binding is considered.

Most of current rigid docking programs employ these assumptions and achieve structural prediction results from acceptable to good for certain protein complexes in CAPRI (Janin 2005), a community-wide critical assessment platform for predicting protein interactions. Nonetheless, according to CAPRI, most of the docking algorithms encountered wrong predictions for difficult cases in which large conformational changes were involved (Janin 2005). The typical cases are

Antigen/Antibody complexes which contains water in the complex interfaces. Due to computational efficiency, most of the current docking programs either ignore or only implicitly take into account the impact of the interfacial water. Therefore, it will be challenging if the effect of water could be implemented explicitly into the scoring function to be able to predict correctly the protein interactions mediated by interfacial water, and still maintain computational efficiency.

Furthermore, limitations of ranking techniques of the initial docking (Figure 1-3) render current docking algorithms impossible to obtain the best performances, results of which are mostly affected by large number of the false positives, which are wrong but highly ranked predictions. Although re-ranking techniques are developed to improve results of those current docking algorithms, it is noticed that there is still a need to improve further the ranking step in order to achieve more correct predictions in the top highly ranks (more details in section 2.1).

Combining the two limitations above, the need of developing a re-ranking step using explicit interfacial water is motivated to further discriminate the correct predictions from the other incorrect and false positive ones, particularly in improving the results of predicting complex formations for Antigen/Antibody complexes.

To explore possible solutions for this challenge, interfacial water is implemented in a scoring function and applied in the ranking stage of a rigid docking algorithm (Figure 1-4) in this work. The water effect is expected to improve the ranking process, and therefore improving the docking results. In this report, interfacial water effect is interpreted by three energy terms in a proposed energy-based scoring function, i.e. Lennard-Jones repulsive (to represent the protein-water dispersion), water-mediated hydrogen bond and the free energy change of the protein interfaces. It is also expected that the complex flexibility is induced in the presence of the added water.

As a result, this implementation aims to (1) obtain better rankings for the top complex docked conformation, and thus achieve better discrimination of the native structures from the false positive ones, as well as (2) complement with the initial protein docking algorithm, hence does not affect the computational efficiency of the docking program.

Note:

- More correct structures are refined while number of false positives is reduced

Figure 1-4 Development of the re-ranking technique to improve the docking results

1.3. Overview of this Thesis

This thesis consists of 7 chapters, which are summarized as below:

Chapter 2 – Various re-ranking techniques to improve the initial docking results, provides an overall picture of the re-ranking approach which aims to improve the ranks of the predictions of protein complexes generated from the initial docking algorithms. In this chapter, two different techniques, i.e. ZRANK and F²Dock are discussed. Although each technique approaches from different angles of employing potential energy terms, their scoring functions reflect the solvent effect and show significant improvement of the complex predictions as compared to the initial

docking algorithm.

Chapter 3 – Docking analysis with implicit and explicit presence of water, highlights some docking applications in various biological issues. In these analyses, both implicit and explicit effect of the solvent (water) are presented in several docking algorithms, and different results are demonstrated to put an emphasis on the important role of water molecules in docking proteins, especially in detecting protein-protein interactions.

Chapter 4 – Modeling studies of various potentials on different physicochemical properties of protein complexes, provides an overview of popular different potential terms used in various scoring functions of the current rigid docking algorithms. Comparisons among them are performed and it is shown that the atom-based potential performs well for Antigen/Antibody complexes (especially the interface Atomic Contact Energy or IFACE), whereas the residue-based potential (DARS) is better for Enzyme/Inhibitor complexes. Different performances of various scoring functions employed in some rigid docking algorithms are also evaluated, and it is observed that combination of shape complementarity, electrostatics, and desolvation using the IFACE performs the best. The combination is therefore selected as a candidate for the further implementations. In addition, an experiment of molecular dynamics (MD) simulation is conducted to highlight that MD-required structural refinement is essential to obtain the optimal conformations of protein complexes.

Chapter 5 – Improving the discrimination of near-native complexes for protein-rigid docking by implementing interfacial water in protein interfaces, describes the proposed re-ranking techniques and the in-depth discussion on the weight estimations. Contributions of the four potential terms (i.e. combined score of potentials in the rigid docking, protein-water dispersion energy, water-mediated Hydrogen bond energy, and free energy change of the complex interfaces) in the energy-based scoring function are interpreted and estimated. As expected, different physicochemical characteristics of protein complexes are partially reflected in this implementation via the weight values. It is observed that the proposed method improves results of the initial rigid docking in increasing both numbers of near-native structures found and the ranks of the correct predictions.

Chapter 6 – Evaluation of the proposed interfacial water-implemented scoring function, discusses further on its ability of discriminating the correct complex predictions of the initial

docking results by the proposed water-implemented scoring function via the hydrogen bond network and the interface free energy change that facilitate the protein interactions at the interfaces. Further, it is also observed that the scoring function is able to maintain the sufficient computation time during the re-ranking, and shows significant improvement of computational efficiency for the MD simulation when the predicted conformations undergo the structural refinement.

Chapter 7 – Conclusion and future work, summarizes and concludes the results and also suggests some further implementations that help improve more the success rate as well as the robustness for the proposed method.

Chapter 2

Various re-ranking techniques to improve the initial docking results

This chapter provides an overall picture of the re-ranking approach, which aims to improve the ranks of the predictions of protein complexes generated from the initial docking algorithms. Two different techniques are described in terms of the scoring functions employed to rank the complexes. They are ZRANK and F²Dock. Although each of the techniques approaches from different angles of employing potential energy terms, they both interpret the solvent effect in their energy-based scoring functions and show significant improvement as compared to the initial docking algorithm ZDOCK. However, observed in the implementations of both the techniques, the water is still presented *implicitly*. It is therefore inspiring for the idea of using *explicit* water in the re-ranking step in this thesis to improve further the results of the initial docking algorithm.

2.1. Re-ranking algorithm in protein docking: ZRANK

2.1.1. Overview

ZDOCK is one of the current protein-protein docking algorithms, which are based on shape complementarity. Although being rigid and shape complementarity-dependent makes it susceptible to some particular characteristics of protein bindings (more details are shown in Chapter 4), ZDOCK is one of the few protein docking algorithms developed in the earlier decade and has achieved successful predictions for most of protein complexes evaluated by the CAPRI (Shen 2013).

Established from a correlation scoring function (Chen, Li *et al.* 2003) including shape complementarity (SC), electrostatics (ELEC), and desolvation (DE) terms:

$$f_{ZD} = w_1 SC + w_2 ELEC + w_3 DE \quad (1)$$

in which

$$SC = \text{Re} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{SC}) \cdot \text{DFT}(L_{SC})) \right] - \text{Im} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{SC}) \cdot \text{DFT}(L_{SC})) \right]$$

$$ELEC_{\alpha,\beta,\gamma} = \sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N \phi_{l,m,n} \cdot q_{l+\alpha,m+\beta,n+\gamma} \quad \text{and}$$

$$DE = \frac{1}{2} \times \text{Im} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{DE}) \cdot \text{DFT}(L_{DE})) \right]$$

are correlation functions representing both the receptor (R) and ligand (L) on a N^3 grid and approximated by using Fast Fourier Transform (FFT) techniques of inverse (IFT) and discrete (DFT) functions. While the *SC* and *DE* are combinations of IFT and DFT, the *ELEC* is presented by a Coulomb model, in which the static charges of the ligand $q_{l,m,n}$ move through the electric field of the receptor $\phi_{l,m,n}$ (Gabb, Jackson *et al.* 1997). More details of the implementation will be described in Chapter 3 (section 3.1.1.2).

ZDOCK team further improves the ranking step with an optimized energy function (Pierce and Weng 2007, 2008) by developing a re-ranking technique to better discriminate the correct structures from incorrect predictions. The re-ranking algorithm is called ZRANK (Figure 2-1).

Figure 2-1 Re-scoring and re-ranking flowchart of ZRANK

ZRANK aims for a more accurate and quick re-ranking of the rigid docking predictions from ZDOCK, (i.e. ZDOCK2.3 at the time) it therefore employs an energy-derived scoring function of linear weighted energy terms of van der Waals interactions (*vdW*), electrostatics (*elec*), and desolvation (*desol*).

$$f_{ZR} = w_4 vdW + w_5 elec + w_6 desol \quad (2)$$

However, while each term of the function f_{ZD} (Equation 1) consists of grid-based discrete functions derived from both the receptor and the ligand, the scoring function f_{ZR} of ZRANK (Equation 2) includes linear sum of van der Waals energies (with attractive and repulsive interactions), electrostatics and desolvation of the whole protein complex. The only commonality between both the programs is that they both apply CHARMM19 potentials and pairwise Atomic Contact Energy (ACE) for the desolvation terms (Mintseris, Pierce *et al.* 2007, Pierce and Weng 2007). More details about the ACE and how it is implemented in the desolvation term are described in Chapter 4 (section 4.1.3 and 4.1.4).

The electrostatics in ZRANK is broken down into short and long range attractive and repulsive energies, which are weighted separately. Moreover, a noticeable difference between ZDOCK2.3 and ZRANK is that ZDOCK2.3 does not account for the presence of hydrogen atoms whereas ZRANK requires polar hydrogens prior to the rescoring.

The scoring function in ZRANK is interpreted into details in Equation 3 below:

$$f_{ZR} = 1.0 \text{ vdW}_{\text{attr}} + 0.009 \text{ vdW}_{\text{rep}} + 0.31 \text{ elec}_{\text{short_attr}} + 0.34 \text{ elec}_{\text{short_rep}} + 0.44 \text{ elec}_{\text{long_attr}} + 0.5 \text{ elec}_{\text{long_rep}} + 1.02 \text{ desol} \quad (3)$$

in which the van der Waals energy is calculated using the Lennard-Jones 6-12 potentials for all atoms i, j having an inter-atom distance $r_{ij} < 8\text{\AA}$.

$$E_{\text{vdW}}(i, j) = \varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

with the well-depth ε_{ij} and vdW radius σ_{ij} are from CHARMM19. A linear repulsive is applied when $r_{ij} < 0.6\sigma_{ij}$.

Similar to ZDOCK2.3, a Coulomb model is used to calculate the electrostatics in ZRANK; yet unlike ZDOCK2.3, instead of a grid-based presentation in which a point charge of each ligand atom interacts with the electric field of the receptor, pairwise charge interactions between atoms are accounted with a $1/r$ distance dependent dielectric, and the electrostatics calculation is separated into short range (mostly for polar-hydrogen bonding) for atoms i, j in distance $r_{ij} < 5\text{\AA}$, and long range electrostatics for atoms at the greater distances (Pierce and Weng 2007).

$$E_{\text{elec}}(i, j) = 332 \frac{q_i q_j}{r_{ij}^2}$$

In general, the scoring function f_{ZR} employs different potential terms to re-score and re-rank the complex predictions generated by the initial docking ZDOCK. However, unlike ZDOCK, the scoring function f_{ZR} is a linear sum of the potential terms and does not require the grid-based presentations of both the receptor and ligand before the scoring; therefore does not significantly affect the overall complexity of the initial docking algorithm. This is an important improvement of the ZRANK in terms of computational efficiency because it can compute quickly and effectively the scores of each of the predicted protein complexes while there is no Fast Fourier Transform (FFT) technique required.

2.1.2. Comparison with the initial docking ZDOCK

To compare with the initial docking ZDOCK, protein-protein docking benchmark 2.0 (Mintseris, Wiehe *et al.* 2005) and 4.0 (Hwang, Vreven *et al.* 2010) are discussed in this part. ZDOCK

successfully generates hits for 62/76 cases for the Benchmark 2.0 (with ZDOCK 2.3) and for 139/159 cases for the Benchmark 4.0 (with ZDOCK 3.0.2). The benchmark dataset includes protein complexes retrieved from the Protein Data Bank and classified as different categories – easy, medium, and difficult— based on CAPRI criteria of conformational change level at the protein interfaces. More details of the protein structures and their group classifications are in Chapter 5 - section 5.2.1.

ZRANK achieved remarkable improvements for most of the cases in the protein complex benchmark 2.0 with a significant increase of number of cases that obtain hits in the top 20 ranks over the two precedent docking ZDOCK versions 2.1 and 2.3, i.e. from 7 (11.3%) cases to 24 (38.7%) cases and from 14 (22.6%) to 22 (35.5%) cases, respectively (Pierce and Weng 2007). Further, ZRANK outperforms ZDOCK version 3 in both the benchmark 2.0 (Pierce and Weng 2008) and the latest benchmark 4.0 (Figure 2-2 and Figure 2-4) by improving the rankings of the top 2000 conformations (which are from dockings of 15° rotational sampling that yields 3600 predictions), especially number of the “top20-rank hit” cases increases from 71 to 86 cases (out of 119 easy and medium cases), in which the near-native structure was found and ranked in the top 20 from results of ZDOCK3.0.2 (Figure 2-2) below.

Inclusion of ZRANK as a re-ranking step shows noteworthy improvement for results of the ZDOCK alone (Pierce and Weng 2008). Especially when combining the atomic contacts within protein chains (i.e. the conventional Atomic Contact Energy or ACE) (Zhang, Vasmatzis *et al.* 1997) with the Interface Atomic Contact Energy (IFACE – detail is in Chapter 4) (Mintseris, Pierce *et al.* 2007) to represent the water exclusion via the desolvation term in the scoring function, performance of the protein-protein docking is significantly improved (Pierce and Weng 2008, Vreven, Hwang *et al.* 2011). It therefore suggests that implementing potential terms of ZDOCK into a new scoring function to re-ranking the docked conformations be practical and efficient.

Figure 2-2 Illustration of comparisons between ZRANK and ZDOCK on the system level.

Figure 2-3 Detailed comparisons of ZRANK and ZDOCK3.0.2 using the benchmark 4. Success rate is defined as percentage of complexes that obtain at least 1 near-native structure in the top 100 predictions.

2.2. Re-ranking using Generalized Born model: F²Dock

2.2.1. Overview

Taking into account the solvent effect in the desolvation terms of both ZDOCK and ZRANK in ranking proves to improve the results of the initial rigid docking. Another re-ranking technique using an implicit solvent model (Generalized Born or GB) implemented in a rigid docking program F²Dock also shows a better performance by locating more correct predictions in higher ranks (Chowdhury, Rasheed *et al.* 2013) than the conventional initial rigid docking.

Figure 2-4 Re-scoring and re-ranking flowchart of F²Dock-Generalized Born

F²Dock applies a Generalized Born (GB) model to estimate the changes in the solvent energy (ΔE_{sol}) at the protein surfaces of protein complexes, and uses this solvation energy value to re-rank the docking predictions (Figure 2-4).

$$\Delta E_{sol} = E_{sol_complexAB} - (E_{solA} + E_{solB}) \quad (4)$$

where each $E_{sol} = E_{cavity} + E_{vdw_solute-solvent} + E_{pol}$ contains the energy to form the cavities in the solvent (E_{cavity}), the van der Waals interactions between the protein atoms and the solvent ($E_{vdw_solute-solvent}$), and a polarization energy E_{pol} which represents the electrostatic potential energy change due to the solvent (Chowdhury, Rasheed *et al.* 2013). Among these terms, the GB model is applied to approximate the E_{pol} between the atoms for the macromolecular systems (Chowdhury

and Bajaj 2010).

$$E_{\text{pol}} = -\frac{1}{2} \left(1 - \frac{1}{\epsilon}\right) \sum_{i,j} \frac{q_i q_j}{\left[r_{ij}^2 + R_i R_j \exp\left(-\frac{r_{ij}^2}{4R_i R_j}\right) \right]^{\frac{1}{2}}} \quad (5)$$

in which q_i is the atomic charge of atom i , r_{ij} is the distance between atoms i and j , R_i is the effective Born radius of atom i to show how deep a charge is buried inside the molecule (Chowdhury and Bajaj 2010).

The GB model is frequently employed to approach a more reasonable and computational efficiency compared to the standard numerical Poisson-Boltzmann equation (Tsui and Case 2001). In the GB model, each atom is represented by a sphere with radius a_i and charge q_i . The interior of the atom is assumed to carry dielectric constant $\epsilon_{\text{in}} = 1$. The molecule is surrounded by solvent with higher dielectric constant ϵ_{out} (~ 78.5 for water at 300K) (Figure 2-5) (Ponder and Case 2003, Case, III *et al.* 2005).

Figure 2-5 Schematic presentation of atomic interactions in implicit solvent following the Generalized Born model, adapted from (Case, III *et al.* 2005).

The GB-based polarization energy E_{pol} in Equation 5 is thus interpreted as:

$$E_{\text{pol}} = -\frac{1}{2} \left(1 - \frac{1}{\epsilon_{\text{out}}}\right) \sum_{i,j} \frac{q_i q_j}{f_{\text{GB}}} \quad (6)$$

Where f_{GB} is a smooth function to show the burial level of the atom i inside the molecule.

$$f_{\text{GB}} = \left[r_{ij}^2 + R_i R_j \exp\left(-\frac{r_{ij}^2}{4R_i R_j}\right) \right]^{\frac{1}{2}}$$

For an isolated ion, $R_i = a_i$ (van der Waals radius), so $f_{\text{GB}} = a_i$, then:

$$E_{\text{pol}} \approx -\frac{1}{2} \left(1 - \frac{1}{\epsilon_{\text{out}}}\right) \frac{q^2}{a_i} \quad (7)$$

If the two spheres are separated in a sufficient large distance compared to their radii, the polarization energy E_{pol} in Equation 5 will be a sum of individual Born terms and pairwise Coulombic potentials:

$$E_{\text{pol}} = -\left(1 - \frac{1}{\epsilon_{\text{out}}}\right) \sum_i \frac{q_i^2}{2a_i} - \frac{1}{2} \left(1 - \frac{1}{\epsilon_{\text{out}}}\right) \sum_{ij} \frac{q_i q_j}{r_{ij}} \quad (8)$$

In brief, the Generalized Born model attempts to achieve a relatively simple formulation to simulate the molecular system with an effort to capture as much property as possible to the Poisson-Boltzman law (Ponder and Case 2003), yet still maintains efficient simplicity.

2.2.2. Comparison with the initial docking ZDOCK

Figure 2-6 Comparisons of F²Dock and ZDOCK3.0.2 using the benchmark 4. Success rate is defined as percentage of complexes that obtain at least 1 near-native structure in the top 100 predictions.

The GB-based re-ranking of F²Dock shows better results for two out of 3 types of protein complexes than those of the initial rigid docking, i.e. ZDOCK3.0.2 (Figure 2-6) in their comparisons (Chowdhury and Bajaj 2010). The comparison was performed on the protein complexes in the benchmark 4.0 with 15° rotational sampling for each of protein complex types,

i.e. Antigen/Antibody (A/AB), Enzyme/Inhibitor (E), and others (O).

The estimation of the ΔE_{sol} using the GB model for the polarization energy E_{pol} however neglects the conformational changes upon the complex formation (Chowdhury and Bajaj 2010). Therefore some biases may be expected due to the rigidity of the protein backbones and side-chains.

2.3. Concluding remark

Two different re-ranking techniques described above provide several good observations as well as limitations that could be useful for the proposed implementation in this project.

- The re-ranking approach using an energy-based scoring function is practically efficient in improving the results of the initial docking algorithm, particularly for the protein-protein docking ZDOCK.
- The linear sum of various energy terms provides more efficient computations for the ranking step against the conventional FFT-based discrete functions.

However,

- The water molecules is implicitly presented; therefore not fully interpreting the solvent effect, especially the interfacial water molecules and the water-mediated Hydrogen bonds in the protein interfaces, which are expected to induce the conformational changes during the protein associations, particularly in the Antigen/Antibody complexes.

Chapter 3

Docking analysis with implicit and explicit presence of water molecules

This chapter showcases some protein docking experiments in vaccine and drug discoveries for various biological issues. In these applications, both implicit and explicit presentations of water molecules are described according to different docking algorithms. In addition, protein dockings followed by the structural refinement using minimization or molecular dynamics simulation for different types of protein complexes is also demonstrated. It is observed from the results of these experiments that although docking with implicit solvent shows acceptable predictions of protein complexes that involve surface-bindings, it is unable to predict the interactions occurring in the deep binding pockets. The docking algorithm with explicit water presentation is, on the other hand, able to detect these bindings. It therefore demonstrates the important role of water molecules in docking proteins, especially in detecting protein-protein interactions.

3.1. Applications of protein docking in vaccine discovery for 2009 H1N1 influenza A virus (*implicit water*)

3.1.1. A possible mutation that enables the H1N1 virus to escape antibody recognition¹

3.1.1.1. Background

Studies that have found cross-neutralizing antibodies between the 1918 and 2009 pandemic flu elicit a basis of pre-existing immunity against the 2009 H1N1 virus in old population. The cross-reactivity occurs due to conserved antigenic epitopes shared between the two pandemic viruses. However, evolutionary mutation can enable the virus to elude human immunity system, making these antibodies probably no longer effective. In this study, it suggested that a possible mutation in B-cell epitope could reveal a chance for the virus to escape the 1918 antibody recognition (more details about this analysis are shown in Appendix A1).

The mutated epitope was mapped to a Hemagglutinin (HA) protein using homology mapping. Subsequently, the epitope-containing HA protein was docked to a structure of a known antibody extracted from the Fab 2D1-1918 influenza virus HA complex [PDB:3LZF]. The 1918 influenza virus HA was chosen because the antigenic drift of the H1N1 influenza A virus involved in the 2009 pandemic started in the year 1918 (Garten, Davis *et al.* 2009, Wei, Boyington *et al.* 2010). The binding ability of the mutated epitope and antibodies was demonstrated using ClusPro, a multistage protocol-docking program with ZDOCK implemented at the time to perform the rigid-body docking.

3.1.1.2. Overview of the docking algorithm ZDOCK

ZDOCK is a rigid docking method to optimize functions of shape complementarity, electrostatics, and desolvation as combination of correlations using FFT algorithm (Chen and Weng 2002). In this section, only how these potential terms were implemented in ZDOCK is shown. More details of these potentials are reviewed in Chapter 4.

¹ Chinh TT Su, Stephanus D Handoko, Christian Schonbach, Xiaoli Li, Chee-Keong Kwoh, "A possible mutation that enables H1N1 Influenza A to escape antibody recognition", IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Hongkong, December 18-21, 2010

a. Shape Complementarity (SC)

As similar to previous studies of Katchalski-Katzir *et al.* (Katchalski-Katzir, Shariv *et al.* 1992) and another rigid docking algorithm FTDOCK (Gabb, Jackson *et al.* 1997), the searching space applied in ZDOCK covers exhaustively the entire rotational and translational space of the ligand with respect to the receptor. Two discrete functions R_{SC} and L_{SC} are used to present the geometric properties on a N^3 grid for the receptor and ligand, respectively, and are presented as:

$$R_{SC} = \begin{cases} 1 & \text{surface of the receptor} \\ \rho_i & \text{core} \\ 0 & \text{open space} \end{cases} \quad \text{and} \quad L_{SC} = \begin{cases} 1 & \text{surface of the ligand} \\ \rho_i & \text{core} \\ 0 & \text{open space} \end{cases}$$

with $\rho=9$ and $i=\sqrt{-1}$. Grid spacing is 1.2Å, and number of grid points $N=100$.

The surface and core atoms are determined by computing their solvent accessible area (ASA) with the radius of the water probe set to 1.4Å (Chen and Weng 2002). If the atom obtains $ASA > 1\text{\AA}^2$, it is considered “surface”; otherwise it is a core atom. The ρ_i value is assigned for the grid point which is within $\sqrt{1.5}$ van der Waals (vdW) radii of a core atom in both the receptor and ligand, or within $\sqrt{0.8}$ vdW radii of a surface atom in the receptor.

The shape complementarity (SC) is presented as a correlation function of the R_{SC} and L_{SC} and is estimated using several discrete and inverse FFT:

$$SC = \text{Re} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{SC}) \cdot \text{DFT}(L_{SC})) \right] - \text{Im} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{SC}) \cdot \text{DFT}(L_{SC})) \right]$$

The algorithm obtains a complexity of $N^3 \log_2(N^3)$.

Rotating the ligand around its each three Cartesian angles by an interval of 15° performs the rotational sampling ($360 \times 360 \times 360 / 15^3$), which yields 3600 predictions after the redundant angle degeneration.

b. Electrostatic interaction (ELEC)

Not only does shape complementarity play an important role in interactions, but electrostatics contributes to molecular binding as well. A simple Coulombic model represents the electrostatic complementarity, which is calculated using Fourier correlation.

Partial charges (from CHARMM19) of the ligand interacting with the electric field of the receptor are measured instead of measuring individual charge-charge interactions (Gabb, Jackson *et al.*

1997, Jackson, Gabb *et al.* 1998). The receptor is placed in a grid and atomic charges are assigned to its atoms. The electric field at each grid node (excluding those at the protein core) is calculated as:

$$\phi_i = \sum_j \frac{q_j}{\varepsilon(r_{ij})r_{ij}}$$

where ϕ_i is the field strength at node i , q_j is the partial charge of atom j , r_{ij} is the distance between i and j ($r_{ij} \geq 2\text{\AA}$), and $\varepsilon(r_{ij})$ is a distance-dependent dielectric function as below:

$$\varepsilon(r_{ij}) = \begin{cases} 4 & \text{if } r_{ij} \leq 6\text{\AA} \\ 38r_{ij} - 224 & 6\text{\AA} < r_{ij} < 8\text{\AA} \\ 80 & r_{ij} \geq 8\text{\AA} \end{cases}$$

The ligand is also projected onto a grid, and the charges are assigned as $q_{l,m,n}$. Then, two discrete functions of receptor and ligand for electrostatic interactions are as below:

$$R_{l,m,n} = \begin{cases} \phi_{l,m,n} & \text{entire grid excluding core} \\ 0 & \text{core of the receptor} \end{cases} \quad \text{and} \quad L_{l,m,n} = q_{l,m,n}$$

The correlation function of the electrostatic interaction (ELEC) represents that the static charges of the ligand ($q_{l,m,n}$) move through the electric field of the receptor (Gabb, Jackson *et al.* 1997). α, β, γ are grid coordinates from which the ligand shifts with respect to the receptor.

$$\text{ELEC}_{\alpha,\beta,\gamma} = \sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N \phi_{l,m,n} \cdot q_{l+\alpha,m+\beta,n+\gamma}$$

c. Desolvation Free energy (DE)

The Atomic Contact Energy (ACE) developed by Zhang *et al.* (Zhang, Vasmatzis *et al.* 1997) is implemented in this method. ACE (detail in section 4.1.3.1) is the free energy of replacing a protein-atom/water contact by a protein-atom/protein-atom contact. The ACE is obtained for all pairs of 18 atom types. The total desolvation free energy (DE) is calculated by summing all the non-pairwise ACE scores (meaning score between one atom of a specific type and another atom of any type) of all atom pairs between the receptor and the ligand that are within $6\text{\AA}$ (Chen and Weng 2002). Using the FFT algorithm, the *DE* correlation is:

$$\text{DE} = \frac{1}{2} \times \text{Im} \left[\frac{1}{N^3} \text{IFT}(\text{IFT}(R_{\text{DE}}) \cdot \text{DFT}(L_{\text{DE}})) \right]$$

The R_{DE} and L_{DE} are set similarly to the shape complementarity. They are both initialized to 0. After looping over every grid point, the sum of ACE scores of all receptor atoms that are within 6Å will be assigned to the *real* parts of the R_{DE} , and “1” is assigned to the *imaginary* part of the R_{DE} for the nearest grid points. The L_{DE} is done in that same fashion.

d. Final target scoring function

As combining SC, ELEC, and DE, the final target scoring function is:

$$S_{SC+DE+ELEC} = w_1 SC + w_2 ELEC + w_3 DE$$

in which $w_1 = 0.01$, $w_2 = 0.06$ and $w_3 = 1$ are scaling factors.

3.1.1.3. Docking experiments and results

Firstly, various HA protein sequences of the pandemic 2009 H1N1 strains were retrieved from (NCBI Influenza sequence database), aligned using MAFFT v.6 (Kato 2008), and used to construct the Neighbor Joining tree (*bootstrap* value = 1,000). Results suggested that HA subtypes of the pandemic influenza strains differed from those of circulating seasonal strains. The 2009 H1N1 pandemic strains also differed from WHO recommended strains that had been used for vaccine development and group-2 HA subtype 14 (detailed phylogenetic tree in Appendix A1). Therefore, mutations derived from the 2009 H1N1 pandemic strains might not have been exposed and would lead unexpected consequences if emerged.

We hence studied computationally the mutations and predicted the possible consequences. Various potential epitopes were predicted using IEDB Analysis Resource Tools for B-cell epitope prediction for linear epitopes by applying different epitope characteristics and prediction methods—Beta-Turn (Chou and Fasman 1978), Surface Accessibility (Emini, Hughes *et al.* 1985), Flexibility (Karplus and Schulz 1985), Antigenicity (Kolaskar and Tongaonkara 1990), Hydrophilicity (Parker, Guo *et al.* 1986), and Bepipred Linear Epitope Prediction (Larsen, Lund *et al.* 2006). The B-cell epitope candidate with score higher than a threshold ($\theta = 0.75$) and resulting in most of the methods was chosen for further investigation. Homology modeling was then done using I-TASSER (Zhang 2007, 2008) to obtain 3D structure for the selected B-cell epitope.

Table 3-1 The potential B-cell epitopes

Epitope	Methods	Sequence 1	Sequence 2
PNHDSNKG	Bepipred	TSSW PNHDSNKG VTAACPHAGAKSF	
	Chou	TSSWPNH	PNHDSNK
	Karplus	PKTSSWP	NHDSNKG
	Parker	PNHDSNK	NHDSNKG

Using Epitope prediction tools in Immune Epitope Database (IEDB), B-cell epitope PNHDSNKG was found (Table 3-1). This epitope sequence obtained highest prediction score and predicted by multiple tools indicating that it is likely to be recognized by the immune system.

The binding ability of the mutated epitope and antibodies was demonstrated using ClusPro, a multistage protocol-docking program. ClusPro (with ZDOCK implemented at the time) performs a rigid-body docking, filters and clusters the docked conformations, and refines the complex results using CHARMM minimization (Comeau, Gatchell *et al.* 2004, Comeau, Gatchell *et al.* 2004, Kozakov, Brenke *et al.* 2006, Comeau, Kozakov *et al.* 2007). The dockings were performed with an antibody-binding mode. The lowest docked energy model was selected for detection with an assumption that the minimum binding energy-obtained model is assumed to be the best docked conformation and nearest to the native binding state.

Figure 3-1 Docking experiment of the epitope-containing HA to the 2D1-Antibody

The docked complex had different antibody-binding orientation from the Fab 2D1-1918 influenza virus HA complex (Figure 3-2). In the docked complex, the 2D1 antibody bound to HA1 domain, but it did not interact with the predicted B-cell epitope as shown in Figure 3-2A.

120 130 140
(A) –EKFE IFPKTSSWPPNHDSNKGVTAACSYAGA–
 *
(B) –ERFE IFPKTSSWPNHETT^{TT}KGVTAAACPYAGA–

Figure 3-2 Influence of His¹³⁰ side-chain on antibody binding of residue Lys¹⁵⁷ in 2D1-epitope-containing HA complex (A) and 2D1-1918 HA protein complex (B). *Top panel:* differences in binding regions of epitope-containing HA and the 1918 HA protein is shown above. Matching residues are marked by asterisks. Mismatches in binding regions (131-133) are colored in red. The B-cell epitope is underlined. *Middle panel:* atom-based residue coloring: C (white), O (red), and N (blue). Backbones are presented by sticks. Side-chains are shown as lines. Interactions are presented as yellow-dotted lines. A. Internal contacts of Ser¹³² and His¹³⁰ reduce the distance of His¹³⁰ to Lys¹⁵⁷ to 1.8 Å, thereby blocking binding of the antibody. B. Lys¹⁵⁷ interacts with Asn⁵⁴ of the antibody. The distance His¹³⁰ is of 4.7 Å. *Bottom panel:* Different binding orientations of the 2D1 antibody to the epitope-containing HA protein (A) and to the 1918 influenza HA (B). The bound complexes of antibody and HA monomer are shown as surface presentation. HA1 and HA2 domains are colored in green and cyan. The 2D1 antibody chains are shown in yellow (heavy chain) and orange (light chain). The predicted B-cell epitope is colored in red.

3.1.1.4. Summary and concluding remark

Variability analysis of H1N1 2009 influenza A HA proteins highlighted the phylogenetic distinctness to other seasonal influenza and the WHO recommended vaccine strains. It is shown that HA subtype H1 is not considered as a potential inhibitory drug target as confirmed by Russell *et al.* (Russell, Kerry *et al.* 2008). A potential B-cell epitope was predicted using HA protein sequences of the current virus strains and it was assumed to have undergone an evolutionary mutation. In this study, it was shown that this B-cell epitope failed to be recognized by the antibodies. In fact, intermediate neighbor contacts of the Lys¹⁵⁷ residue make it unavailable for binding to the antibody. Hence, it could suggest that if the virus actually obtained this mutation, the antibodies could no longer recognize it. This finding could be useful for vaccine designs in later years.

3.1.2. Implications for T-cell Receptor (TCR) cross-recognition and vaccine development for 2009-H1N1 and 2004-H5N1 influenza virus HLA-B*4405-restricted HA epitope candidates²

3.1.2.1. Background

Studies on pandemic influenza vaccines have shown that T cell recognition to conserved epitopes and cross-reactive T cell responses are important when new strains emerge, especially in the absence of antibody cross-reactivity. In this work, using Human Leukocyte Antigen type B*4405 (HLA-B*4405) and DM1-TCR structure model, high confidence conserved 2009-H1N1 T cell epitope candidates were systematically generated and investigated on their potential cross-reactivity against H5N1 avian flu virus. The results are novel with regard to HA epitopes and useful for developing possible vaccination strategies against the rapidly changing influenza viruses. Yet, the challenge of identifying epitope candidates that result in heterologous T cell immunity under natural influenza infection conditions can only be overcome if more structural data on the TCR repertoire become available (more detail about this analysis is shown in Appendix A2).

²Chinh T.T Su, Christian Schonbach, Chee Keong Kwoh, "Molecular docking analysis of 2009-H1N1 and 2004-H5N1 influenza virus HLA-B*4405-restricted HA epitope candidates: Implications for TCR cross recognition and vaccine development", BMC Bioinformatics, 14(Suppl 2):521 (2013)

The pipeline included docking the epitopes to HLA-B*4405 [PDB:3DX8] and followed by docking of the HLA/epitope complex as a ligand to DM1-TCR [PDB:3DXA]. The binding ability of the predicted epitopes was further analyzed using ClusPro v.2 (Kozakov, Brenke *et al.* 2006, Kozakov, Hall *et al.* 2010), a web-based automated docking program performing a multistage protocol: rigid PIPER docking.

3.1.2.2. Overview of the docking algorithm PIPER

PIPER is a rigid docking algorithm, in which approximation of the interactions is represented in an energy expression as the sum of a few correlation functions (Kozakov, Brenke *et al.* 2006). Applying this assumption, the energy-like scoring function defined on a grid is expressed as the sum of P correlation functions for all possible translations α, β, γ of the ligand respected to the receptor:

$$E(\alpha, \beta, \gamma) = \sum_P \sum_{i,j,k} R_P(i, j, k) L_P(i + \alpha, j + \beta, k + \gamma)$$

where $R_P(i, j, k)$ and $L_P(i, j, k)$ are components of the receptor and ligand correlation functions, respectively. This expression is calculated using P discrete Fourier transforms (DFT) and one inverse Fourier transform (IFT) as below:

$$E(\alpha, \beta, \gamma) = \text{IFT} \left\{ \sum_P \text{DFT}^*(R_P) \text{DFT}(L_P) \right\} (\alpha, \beta, \gamma)$$

and the complexity is $N^3 \log(N^3)$, in which N is the dimension of the grid.

The energy scoring function above is expressed in terms of shape complementarity, electrostatic interactions, and pairwise potentials.

$$E = E_{\text{shape}} + w_2 E_{\text{elec}} + w_3 E_{\text{pair}}$$

The shape complementarity term (E_{shape}) contains two components: attraction and repulsion interactions, the latter of which helps to eliminate atomic overlaps (Kozakov, Brenke *et al.* 2006).

$$E_{\text{shape}} = E_{\text{attraction}} + w_1 E_{\text{repulsion}}$$

with w_1 , w_2 , and w_3 are weight coefficients in different contributions to the scoring function.

The repulsion interactions are cut off at the van der Waals radius ($r_{\text{vdW}} + 2\text{\AA}$) to allow differences

between bound and unbound structures, while the attraction part has the same cutoff radius ($6\text{\AA}$) for all atom types. The two discrete functions for the receptor and the ligand on a grid are:

$$R_{l,m,n} = -c_{l,m,n} + w_1 r_{l,m,n} \quad \text{and} \quad L_{l,m,n} = \begin{cases} 1 & \text{if } (l, m, n) \ni (a_j \in J) \\ 0 & \text{otherwise} \end{cases}$$

Where $c_{l,m,n}$ is the number of atoms that are at distance $r_{vdW} + 2\text{\AA} < r < 6\text{\AA}$

$r_{l,m,n}$ is the number of atoms that are at distance $r < r_{vdW} + 2\text{\AA}$

$(l, m, n) \ni (a_j \in J)$ means the grid point (l, m, n) overlaps with atom j of the type J

A simplified Generalized Born model with constant Born radii was used to calculate the electrostatic interactions between the two molecules surrounded by solvent. The approximation of the electrostatic interaction is expressed as a correlation function of the electric field of the receptor and the charges of the ligand atoms.

$$E_{\text{elec}} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \frac{q_i q_j}{\left[r_{ij}^2 + D^2 \exp\left(\frac{-r_{ij}^2}{4D^2}\right) \right]^{\frac{1}{2}}}$$

where N_R and N_L indicate the numbers of atoms in the receptor and the ligand, respectively. And

$$R_{l,m,n} = \sum_{i=1}^{N_R} \frac{q_i}{\left[\hat{r}_{i,(l,m,n)}^2 + D^2 \exp\left(\frac{-\hat{r}_{i,(l,m,n)}^2}{4D^2}\right) \right]^{\frac{1}{2}}} \quad \text{and} \quad L_{l,m,n} = \begin{cases} q_j & \text{if } (l, m, n) \ni (a_j \in J) \\ 0 & \text{otherwise} \end{cases}$$

$\hat{r}_{i,(l,m,n)} = \max(r_{i,(l,m,n)}, 2D)$ when $r_{i,(l,m,n)}$ is the distance between atom a_i and the grid point (l, m, n) .

The cutoff distance $D=6\text{\AA}$.

The pairwise potentials is:

$$E_{\text{pair}} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \varepsilon_{ij}$$

The E_{pair} does not have the form of a correlation function, but it can be written as a sum of correlation functions, which is based on eigenvalue-eigenvector decomposition of the pairwise interaction matrix of the element ε_{IJ} (when interacting atoms belong to atom type I and J).

The element matrix ε_{IJ} is expressed as:

$$\varepsilon_{IJ} = \sum_{p=1}^K \lambda_p u_{pI} u_{pJ}$$

This element matrix is symmetric and has K (number of different atom types) real eigenvalues. λ_p is the p^{th} eigenvalue of the interaction matrix and u_{pI} is the I^{th} component of the p^{th} eigenvector. Therefore, for any pairwise potential, it can be calculated using K real and $K/2$ complex FFT.

The energy term with the p^{th} eigenvalue of the pairwise potential on the grid is presented as:

$$R_{l,m,n} = \sum_{i=1}^{N_R} u_{pI} \delta_i \quad \text{and} \quad L_{l,m,n} = \begin{cases} u_{pJ} & \text{if } (l, m, n) \ni (a_j \in J) \\ 0 & \text{otherwise} \end{cases}$$

$\delta_i=1$ if the grid point (l, m, n) is at the distance within $D=6\text{\AA}$ from atom i of the receptor.

3.1.2.3. Docking experiments and results

Potential epitopes were predicted using T cell epitope prediction tool (NetCTL v1.2) from IEDB (Larsen, Lundergard *et al.* 2005, Larsen, Lundegaard *et al.* 2007). NetCTL, a neural network architecture-based tool, was used to predict T cell epitope candidates for HA proteins of current H1N1 influenza A strains according to HLA-B44 supertype. Weight parameters on C-terminal cleavage (=0.15) and transporter associated with antigen processing (TAP) efficiency (=0.05) were as default. There are 3 threshold scores (0.75, 0.9, 1.0) that give both high sensitivity (0.8, 0.74, 0.7) and specificity (0.97, 0.98, 0.985) accordingly. While the first 2 scores (0.75 and 0.9) obtained similar results of T cell epitopes, the score of 1.0 resulted only 7 epitopes (positions 240, 188, 128, 482, 400, 131, and 50) shown in Table 3-2. Our docking results of HLA-B*4405 and the T cell epitope candidates showed that epitope candidates at positions 259, 251, and 229, which were located in the top 10 HLA-binding energy rankings in Table 3-2, would be missed if the 1.0 score was used. Finally, as comparing to the score of 0.9, we selected 0.75 as the threshold for its better sensitivity (0.8) to predict the potential T cell epitopes. Therefore candidates with prediction scores greater than 0.75 were chosen for further investigation. I-TASSER (Roy, Kucukural *et al.* 2010) was used to obtain homology models of the selected T cell epitopes.

The pipeline included docking the epitopes to HLA-B*4405 [PDB:3DX8] and followed by docking of the HLA/epitope complex as a ligand to DM1-TCR [PDB:3DXA]. The binding ability of the predicted epitopes was further analyzed using ClusPro v.2 (Kozakov, Brenke *et al.* 2006,

Kozakov, Hall *et al.* 2010), a web-based automated docking program performing a multistage protocol: rigid PIPER docking, filtering and clustering of docked conformations, and stabilizing using Monte Carlo simulations (Kozakov, Hall *et al.* 2010). During the docking, HLA β -domain was masked but remained surface accessible since it is not involved in interaction with the TCR (Reinherz, Tan *et al.* 1999). The results were clustered according to their binding energies. The binding energy score (hereby called “binding energy”) is generated from an energy function of PIPER. This function is a sum of potential terms of shape complementarity, electrostatics, desolvation contributions, and Decoys as reference states (DARS) (Kozakov, Brenke *et al.* 2006).

Figure 3-3 Docking experiment of the T-cell epitope to the HLA-B*4405 and the DM1-TCR

According to Kozakov *et al.*, the core idea of the knowledge-based potentials DARS counts on observed numbers of intermolecular interactions (Kozakov, Brenke *et al.* 2006), thus docking results were filtered by selecting docked complexes that belonged to the most populated clusters of interacting complexes but with lowest binding energy scores for the final results. It was assumed that the model with minimum binding energy was the optimal conformation. Then, molecular dynamics (MD) simulations using AMBER 10 force field ff99SB was performed (Case,

CheathamIII *et al.* 2005) to improve the bound conformation of rigid docking (Figure 3-3).

Table 3-2. Total binding energies of the 19 epitope candidates (in descending HLA-B*4405 binding energy ranks).

Position	Sequence	Docked to HLA-B*4405		Docked to DM1-TCR	
		Energy	Rank	Energy	Rank
240	QEGRMNYW	-1,006	1	-541.7	17
188	KEVLVLWGI	-921.8	2	-573.9	13
128	FERFEIFPK	-869.9	3	-594.6	12
482	FEFYHKCDN	-846.8	4	-743.9	2
259	FEATGNLVV	-815.2	5	-608.9	10
400	IEKMNTQFT	-775.8	6	-688.5	5
131	FEIFPKTSS	-752.9	7	-560.5	14
251	VEPGDKITF	-743.3	8	-599.1	11
229	PEIAIRPKV	-687.7	9	-548.9	15
50	LEDKHNGKL	-661.4	10	-552.1	16
274	MERNAGSGI	-644	11	-	-
514	REEIDGVKL	-643.8	12	-666.3	7
81	PECESLSTA	-633.3	13	-	-
446	LENERTLDY	-603.1	14	-621.1	9
412	KEFNHLEKR	-599.5	15	-737.2	3
493	MESVKNQTY	-543	16	-645	8
475	KEIGNGCFE	-519.6	17	-715.4	4
463	YEKVRSQLK	-475.3	18	-751.3	1
421	IENLNKKVD	-464.6	19	-666.7	6

“-“ means TCR could not recognize the corresponding epitope

As expected, all the predicted epitopes bound to the helical antigen-presenting groove of the HLA-B*4405 (Reinherz, Tan *et al.* 1999). Eleven out of 19 epitopes, residues of which occupy the centre of the binding groove are solvent-exposed, bound to HLA-B*4405 in similar orientations with Glu at position 2 and residues at the C-terminus binding to the B and F pockets of the HLA-B*4405, respectively (DiBrino, Parker *et al.* 1995, Reinherz, Tan *et al.* 1999, Macdonald, Purcell *et al.* 2003). These exposed epitopes likely to bind to the TCR (Reinherz, Tan *et al.* 1999, Wahl, McCoy *et al.* 2009). In contrast, most residues of other six epitopes interacted with the HLA-B*4405 peptide binding pockets. In this DM1-TCR/HLA-B*4405/epitope model, the deeply embedded residues were not exposed to the solvent. Therefore TCR recognition of HLA-B*4405/epitope complexes is expected to be conformationally restrained (Figure 3-4).

Figure 3-4 Examples of conformational TCR adaption to differently recognize both embedded and solvent-exposed epitopes.

3.1.2.4. Summary and concluding remark

The DM1-TCR was docked to the HLA-B*4405/epitope complexes to assess their binding conformation. In the models with minimum binding energies, DM1-TCR V_{α} and V_{β} domains interacted with HLA-B*4405 α_1 and α_2 domain residues and solvent exposed residues of the epitopes. In the case of epitope candidates that were embedded in the binding groove of HLA-B*4405, the interactions appeared to be mediated by a change of the TCR binding position to HLA-B*4405 α_1 - and α_2 -domain residues, resulting in contacts with polar and charged residues of the HLA-B*4405. As a result, it is expected in a future pandemic that apart from a few mutated epitopes, heterologous immunity (Selin and Welsh 2004) mediated by pre-existing cross-reactive T cell responses to seasonal influenza virus will limit its severity and extent.

3.1.3. Concluding remark for this section

The protein-protein docking programs applied in the two applications of vaccine discovery for the 2009 H1N1 influenza A virus above are rigid docking, and some of the results were refined using MD simulation to obtain the optimal conformations of the protein complexes. Although the results are acceptable and reasonable, the protein complexes were selected from the top predictions of the corresponding docking algorithm. There might be other correct binding candidates (details in Appendix A2), which are not on the top predictions and therefore were not found, especially in the ZDOCK results. It would be a more complete analysis if the results were further improved in locating more correct predictions against the incorrect ones.

3.2. Protein docking analysis of HIV-1 Gag and Protease interactions (*explicit water*)

3.2.1. Background

Every year the AIDS epidemic infects 33.4 million and takes 2 million lives worldwide. Despite significant research and developments in drug targets and novel therapeutic candidates, the disease remains incurable to date. The rapid development of resistance against current drugs by targeted viral enzymes play a major contributing factor to the ineffectiveness of current therapies. Although circumvented by the use of cocktail drugs, recent studies have implicated substrates of these targeted enzymes. Increasing evidences have shown that mutations in both the cleavage and non-cleavage sites of the polyprotein Gag, a substrate of Protease, compensates in the resistance against Protease Inhibitors. As a result, drugs targeting the maturation of the virus require the in-depth analysis of both Protease and Gag in order to be designed effectively.

Gag and Protease are encoded in the gag-pol genes of the HIV genome and are involved in the viral maturation process. The protease protein is inhibited by Protease Inhibitors that bind to the enzyme site where the substrate Gag cleavage occurs, thus preventing the release of viral proteins essential for the viral maturation cycle. Despite the common usage of Highly Active Antiretroviral Therapy (HAART), attempts to control the infection typically end in failure. Studies in the last decade have shown that the viral substrate Gag plays a significant role in HIV-1 Protease Inhibitor resistance and have supported evidence that suggest Gag undergoes compensatory mutations to overcome those inhibitors without significant mutations in Protease (Gatanaga, Suzuki *et al.* 2002, Maguire, Guinea *et al.* 2002, Tamiya, Mardy *et al.* 2004). In fact, pharmaceuticals companies now face a unique co-evolutionary model in Gag-Protease, which may explain why current attempts to target protease alone had largely been unsuccessful. As such, there is an increased need to understand the Protease-Gag interactions at the cleavage site of Gag. In this application, the interactions of Gag and Protease are performed using HADDOCK with explicit presence of water molecules during the docking process.

3.2.2. Overview of the docking algorithm HADDOCK

Water plays an important role in protein interactions and associations. It is therefore expected that excluding the water effect in protein docking algorithm will worsen its accuracy and thus its performance. On the other hand, adding water into the docking protocol increases significantly the overall complexity. In fact, it was impractical to dock solvated macromolecules while taking into account all the degrees of freedom and the water-induced flexibility due to the limited computational resources of the last decades. As a result, protein-protein docking algorithm has gone through a process of ignoring totally the water to implicitly accounting it by simulating water as a continuum solvent or adding desolvation effect (Gabb, Jackson *et al.* 1997, Jackson, Gabb *et al.* 1998, Chen and Weng 2002, Chen, Li *et al.* 2003, Chen and Weng 2003, Gray, Moughon *et al.* 2003, Chuang, Kozakov *et al.* 2008, Kozakov, Schueler-Furman *et al.* 2008, Garzon, Lopez-Blanco *et al.* 2009, Li, Guo *et al.* 2011). Therefore, current docking algorithms are able to achieve moderate to accurate predictions for most of protein complexes in reasonable timing. Nonetheless, it remains challenging when the targets acquire large conformational changes during the complex formation, leading to incorrect predictions or failure in discriminating the correct structures from the false positive ones. To tackle the issue, HADDOCK introduces water explicitly into the protein-protein docking protocol to improve flexibility during the protein interactions.

HADDOCK is the first protein-protein docking that explicitly accounts the presence of water in protein-protein complexes (Dijk and Bonvin 2006, Kastiris, Visscher *et al.* 2013). The process mimics the physical concept of protein complex formation in which a water layer is firstly presented in-between the two molecules, and when the molecules are assembled, most of the interfacial waters will be expelled.

HADDOCK implements Ambiguous Interaction Restraints (AIRs) to drive the docking (Dominguez, Boelens *et al.* 2003, Dijk and Bonvin 2006). AIRs are derived from experimental information of residues involved in intermolecular interactions. Each AIR is represented by an ambiguous intermolecular distance d_{iAB} (with $d \leq 2\text{\AA}$) between any atom m of an active residue i of protein A (m_{iA}) and any atom n of residue k of protein B (n_{kB}). The effective distance d_{iAB}^{eff} for each restraint is calculated as:

$$d_{iAB}^{\text{eff}} = \left(\sum_{m_{iA}=1}^{N_{\text{atoms}}} \sum_{k=1}^{N_{\text{resB}}} \sum_{n_{kB}=1}^{N_{\text{atoms}}} \frac{1}{d_{m_{iA}k_{kB}}^6} \right)^{-\frac{1}{6}}$$

in which the active residues are considered as active site residues or residues involved in the protein interactions (as a little differently defined as compared to the original version of HADDOCK when it was firstly dependent on chemical shift perturbation data from NMR experiments (Dominguez, Boelens *et al.* 2003).

Both the proteins are solvated before being docked rigidly. An optimization of water positions using molecular dynamics simulation is performed. The results are then refined in simulated annealing stages to offer flexibility for both the side chains and protein backbone (Figure 3-6). During the docking process, HADDOCK evaluates docked complexes using an energy-based scoring function combined with the ambiguous interaction restraints described above. The energy potentials include non-bonded energies (van der Waals *vdW* and electrostatics *elec*), buried surface area *bsa*, and desolvation *desolv* energy. The overall function *E* is achieved by:

$$E = 0.01 \text{ vdW} + 1.0 \text{ elec} + 0.1 \text{ AIR} + (-0.01) \text{ bsa} + 1.0 \text{ desolv}$$

in which the buried surface area *bsa* is calculated as the difference between the sum of solvent accessible surface area (ASA) for each molecule and the ASA of the whole complex using a 1.4Å radius solvent probe (Figure 3-5).

$$\text{bsa} = (\text{ASA}_{\text{molecule_receptor}} + \text{ASA}_{\text{molecule_ligand}}) - \text{ASA}_{\text{molecule_complex}}$$

In addition to docking solvated molecules, influences of water on the interactions were also detected. Correlation of water-mediated contacts for all 20x20 amino acid combinations was analyzed (Dijk and Bonvin 2006). It is shown that there exist preferences of water-mediated contact among the amino acid pairs. For instances, Gly-Gly is less affected by water while His-Glu has very high fraction of the water-mediated contacts. This analysis is very helpful for the docking of solvated proteins to determine the preferred docked conformation in favor of water impact on the protein assembly.

Figure 3-5 Illustration of Accessible Surface Area calculation, adapted from (WHAT-IF 2013)

Figure 3-6 HADDOCK – the solvated docking protocol (Dijk and Bonvin 2006).

3.2.3. Gag-Protease interactions

The 198-residue Protease was docked as a ligand into cleavage sites of a 500-residue Gag protein using the solvated docking HADDOCK—online web server (Vries, Dijk *et al.* 2010). Dockings of Gag and Protease focus on the symmetric triad Asp-Thr-Gly (DTG) active site of the Protease. Results of the top 200 docked Gag-Protease complexes are clustered using a cutoff distance of 7.5 Å. The best structure of the top cluster—assumed with the most reliable HADDOCK score (using the overall function E above)—is chosen as the best candidate.

For the comparison purpose between the dockings with implicit and with explicit water, the similar dockings of Gag and Protease were performed using ZDOCK3.0.2 at the same time and in the same fashion.

Structural comparisons of the docked Gag-Protease complexes reveal that the buried surface area and interaction restraints in HADDOCK could help the Protease (ligand) access deeply into the buried binding pockets of the Gag protein (receptor) whereas this observation cannot be achieved from the shape complementarity-based docking, e.g. in ZDOCK (Figure 3-7).

Figure 3-7 Different structural orientations of Protease (ligand) docked to Gag (receptor) at the same cleavage site. Protease generated from HADDOCK (green) interacts with the buried cleavage site (red) of Gag (white surface) while the docked orientation of the Protease resulted by ZDOCK (blue) cannot access to this buried site.

3.2.4. Summary and concluding remark

According to the 5th evaluation report of the latest CAPRI (EMBL-EBI 2013), a community-wide protein docking assessment, referred by Dr. Yang Shen (Shen 2013), HADDOCK team is one of the top docking teams (in fact, rank number 1 and 12 among 63 teams) who obtain either accurate or acceptable predictions for most of the CAPRI targets, especially target T47 that

required predictions of water-mediated interactions between proteins. As a result, the presence of water improves docking results makes it reinforced the fact that implementing interfacial water and water-mediated interactions to pre-evaluate docked complex structures could be a crucial step to improve the number of near-native structures for the further refinement.

3.3. Conclusion for this chapter

Water plays an important role in determining the correct binding orientations of the ligand respect to the receptor in protein-protein docking. With the inclusion of water into the docking protocol combined with optimization of water positions and side chain refinement, it suggests that docking proteins with the presence of water be practical and feasible for macromolecules, especially when the prior knowledge of interactions is known, as demonstrated by the results from HADDOCK.

However, including directly water into the docking protocol makes it difficult for conventional rigid docking programs, e.g. ZDOCK, PIPER, FTDock etc., which have been developed for decades without considering the water or simply accounting it implicitly. If doing so, the algorithms would need to be reconstructed almost totally from the scratch. Therefore in order to achieve better docking results, these algorithms have instead been improved in either *searching* or *ranking* stages. Between the two approaches, the ranking (or re-ranking) demonstrates significant improvements in obtaining more correct predictions as well as distinguishing them from other incorrect ones. In fact, for this aspect, there is still more room to develop and improve for either achieving better docking results or reducing numbers of false positives in the top ranks which will be selected for further intensive refinement.

Chapter 4

Modeling studies of various potentials on different physicochemical properties of protein complexes

This chapter serves as a brief review of various potential terms used in different scoring functions of some current rigid docking algorithms. The potentials are interpreted in terms of mathematic equations that are employed to estimate the contributions of the physicochemical characteristics of protein complexes as well as of the solvent effect on the protein interactions.

Comparisons among the potentials are performed and it is shown that the atom-based energy (ACE) performs better for Antigen/Antibody complexes whereas the residue-based potential (DARS) is better for Enzyme/Inhibitor complexes. Performances of various combinations of the potentials are also evaluated, and it is indicated that the combination of shape complementarity, electrostatics, and desolvation using the interface ACE performs the best. Therefore, the combination of potentials is chosen as a candidate for the further implementation in this research. In addition, an extra experiment of molecular dynamics (MD) simulation is conducted to highlight that MD-required structural refinement is essential to obtain the optimal conformations of protein complexes.

4.1. An overview of potential terms used in scoring functions of the initial docking algorithms

4.1.1. Shape complementarity (SC)

Association of receptor and ligand is involved in inter- and intra-molecular interactions, solvation effects, and conformational changes (Katchalski-Katzir, Shariv *et al.* 1992). One of the simple approaches to resolve such complexity is based on geometry to identify molecular surface complementarity between the receptor and the ligand (as shown in Figure 4-1). The first geometry-based algorithm was developed by Katchalski-Katzir's group to predict structure of protein complexes. This established a foundation for most of current protein docking methods (Gabb, Jackson *et al.* 1997, Vakser 1997, Mandell, Roberts *et al.* 2001, Kozakov, Schueler-Furman *et al.* 2008, Pierce and Weng 2008, Ritchie, Kozakov *et al.* 2008). The algorithm relies on the function correlation and Fast Fourier transform (FFT) techniques used in the field of pattern recognition. According to this algorithm, shape complementarity (SC) estimates correlated surface matching between receptor (R) and ligand (L), and is calculated using a correlation function representing the receptor and the ligand on a N^3 grid with l, m, n coordinates.

$$SC_{\alpha,\beta,\gamma} = \sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N R_{l,m,n} \cdot L_{l+\alpha,m+\beta,n+\gamma} \quad (11)$$

where α, β, γ are numbers of grid points by which the ligand is shifted with respect to the receptor in each dimension. The SC value receives 0 when there is no contact between the two proteins (Figure 4-1a) or positive value when the two molecules are in interactions (panels b and d). However, deep penetration or overlapping (Figure 4-1c) is avoided when the ligand shifts into the receptor by setting some penalty values that lead to a negative contribution to the overall correlation SC. Consequently, a better surface match (Figure 4-1d) obtains a higher SC value. All the SC values are demonstrated in a 3D-plot, in which the highest positive peak represents the best surface match between the molecules.

Figure 4-1 Schematic presentation of surface recognition by Katchalski-Katzir *et al.* While the left panels *a* and *c* show presentations of no interaction and penetration respectively, the panels *b* and *d* show different surface complementarity of the two proteins. The plot on the right depicts the surface complementarity according to the SC algorithm.

Although shape complementarity is sufficient to reconstitute separated components of a known bound complex, it is not sufficient to dock unbound components (Smith and Sternberg 2002) whose interactions are driven by other characteristics than surface complementarity alone. Hence, the current docking algorithms further implement extra physicochemical criteria, e.g. electrostatic interactions, pairwise amino acid contacts, solvation energy, nonpolar buried surface area, etc., into the scoring functions to find the desired correlation values, or in other words, to find the best matches between the two molecules. For examples, besides applying shape complementarity, FTDock (Gabb, Jackson *et al.* 1997) uses electrostatics as a filter to refine the rankings, ZDock (Chen and Weng 2002) combines shape complementarity with electrostatics and desolvation in the final scoring function, and PIPER (Kozakov, Brenke *et al.* 2006) utilizes decoys as reference states, and so on.

4.1.2. Electrostatics (ELEC)

Not only does shape complementarity play an important role in interactions, but electrostatics contributes to molecular binding as well. A simple Coulombic model representing the electrostatic complementarity is often used. In this model, point charges of the ligand interacting with the electric field of the receptor are measured instead of measuring individual charge-charge interactions (Gabb, Jackson *et al.* 1997, Jackson, Gabb *et al.* 1998). The receptor is placed in a

grid and atomic charges are assigned to its atoms. The electric field at each grid node is calculated as:

$$\phi_i = \sum_j \frac{q_j}{\varepsilon(r_{ij})r_{ij}}$$

where ϕ_i is the field strength at node i , q_j is the charge of atom j , r_{ij} is the distance between i and j , and $\varepsilon(r_{ij})$ is a distance-dependent dielectric function. In the grid-based docking approach, approximation of the electrostatic interactions is expressed as a correlation function of the electric field of the receptor and the charges of the ligand atoms. Among three docking algorithms mentioned above, there exist two differences including (1) how the *ELEC* potential term is implemented and (2) different charge models are used.

The *ELEC* and *SC* terms are combined in the scoring functions of both ZDOCK and PIPER, but in FTDOCK, aside from shape complementarity (*SC*), electrostatics is only used as an independent binary filter (Gabb, Jackson *et al.* 1997) to pinpoint the false positives, which obtained high shape correlation scores but unfavorable electrostatics scores. In addition, although both FTDOCK and ZDOCK apply the Coulombic model to estimate the electrostatics, the atomic charges q_i of the atom i is replaced by partial charges from the CHARMM19 potential in ZDOCK (Chen and Weng 2002). On the other hand, another simplified Generalized Born model with constant Born radii is used to calculate the electrostatic interactions between the two molecules in PIPER (Kozakov, Brenke *et al.* 2006).

$$E_{\text{elec}} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \frac{q_i q_j}{\left[r_{ij}^2 + D^2 \exp\left(\frac{-r_{ij}^2}{4D^2}\right) \right]^{\frac{1}{2}}} \quad (12)$$

where N_R and N_L indicate the numbers of atoms in the receptor and the ligand, respectively, and D is a cut-off distance.

4.1.3. Atomic Contact Energy (ACE)

Water plays an important role in protein interactions and associations. Accounting for water contribution in the protein interactions leads to more accurate and practical docking solutions for protein complexes (as shown in Chapter 3 in the solvated docking HADDOCK). However, this

also increases significantly the complexity of the docking problem, especially for various conventional docking algorithms which have been developed since the era of limited computational power. To accommodate this issue, the Atomic Contact Energy (ACE) term or potential was developed to estimate the energy required to replace the protein-water contacts by the protein-protein contacts. In other words, the ACE term is used to present the contributions of water in the protein associations.

4.1.3.1. Conventional ACE

The conventional atom-level ACE developed by Zhang *et al.* (Zhang, Vasmatzis *et al.* 1997) is adapted and re-defined from the residue-level potential (Miyazawa and Jernigan 1985). The ACE is established from 18 different atom types of proteins, which are clustered based on heavy atoms of the common 20 amino acids. The 18 clusters of atom types were referred to the protein ($i = 1, \dots, 18$) and an extra atom type was referred to the water ($i = 0$). Each atom type was assigned a coordination value q_i representing the average number of possible nearest neighbor contacts (Equation 13).

$$\frac{1}{2} q_i n_i = \sum_{i,j=0}^{18} n_{ij} \quad (13)$$

Where n_i and n_{ij} are number of atoms of type i and numbers of contacts between 2 atoms of different atom types i, j , respectively.

Let E_{ij} be the absolute contact between atoms of type i, j ($i, j = 1, \dots, 18$), the relative Atomic Contact Energy e_{ij} was defined as:

$$e_{ij} = E_{ij} + E_{00} - E_{i0} - E_{j0} \quad (14)$$

in which E_{00} and E_{ij0} are absolute contacts of water-water atoms and of atom type i or j and water, respectively. The Equation 14 therefore represents the energy change when a protein-protein contact is formed from protein-water atom pairs.

The ACE e_{ij} can be approximated in terms of the observed contact numbers:

$$\exp(-e_{ij}) = \frac{\bar{n}_{ij} \bar{n}_{00}}{\bar{n}_{i0} \bar{n}_{j0}} \quad (15)$$

where $\bar{n}_{ij}$ is estimated as the observed contacts of protein atoms of type i, j (Zhang, Vasmatzis *et al.* 1997). The full table of ACE values of 18 types of protein atoms from Zhang *et al.* is summarized in Table 4-1.

4.1.3.2. Interface Atomic Contact Energy (IFACE)

The conventional ACE has been further improved (Mintseris, Pierce *et al.* 2007) to include atoms that have contacts involved in the protein interfaces. Following the procedure of defining the conventional ACE, the IFACE applies the coordination number q_i (as in Equation 13), but with an extra modification.

$$q_i n_i = \sum_{i,j=1}^{18} n_{ij} + n_{i0} \quad (16)$$

In which the n_{i0} represents contacts of protein atom with water at the interfaces. Therefore, n_{i0} is derived from the Equation 16 as:

$$n_{i0} = q_i n_i - \sum_{i,j=1}^{18} n_{ij} \quad (17)$$

The IFACE is a knowledge-based statistical potential. The IFACE is estimated using the reference states that represent all possible contacts between surface atoms of the interacting proteins (Mintseris, Pierce *et al.* 2007), and is defined as:

$$e_{\text{IFACE};ij} = \ln \frac{n_{ij}}{c_{\text{surface-fraction};ij}} \quad (18)$$

with

$$c_{\text{surface-fraction};ij} = \frac{\sum n_{i0;p1} n_{j0;p2}}{\sum_1^{18} n_{i0;p1} \sum_1^{18} n_{i0;p2}} \cdot n_{rr;c}$$

where the summations are applied for all the complexes c and the interacting partners p_1 and p_2 . The value n_{rr} represents interactions of any atoms of any types. In the equation 18, the potential term “*surface-fraction*” is extrapolated from the possible representing interface contacts of the reference states—which are mathematically defined by (Mintseris, Pierce *et al.* 2007) for atom-level potentials.

The IFACE is designed and implemented in the latest ZDOCK3.0.2 docking algorithm, which is

hence employed by this research.

4.1.4. Desolvation Free energy (DE)

Diversity of experimental structures of protein complexes reveals that shape complementarity and electrostatics are still not sufficient to distinguish the correct predictions from other false positives. Hence, a third potential term is introduced into the scoring function of the current docking algorithms.

The DE potential is derived from the Atomic Contact Energy (ACE) (Zhang, Vasmatzis *et al.* 1997) described in part 4.1.3 above. ACE is the free energy of replacing a protein-atom/water contact by a protein-atom/protein-atom contact; hence the total contact energy of all atoms with the water represents the energy required to desolvate the protein complex. The desolvation DE is defined:

$$DE = \sum_{i=1}^{18} \sum_{j=1}^{18} e_{ij} n_{ij} \quad (19)$$

in which the ACE is obtained for all pairs of 18 atom types. The total desolvation free energy (DE) is calculated by summing all the non-pairwise ACE scores between one atom of a specific type and another atom of any type of all atom pairs between the receptor and the ligand that are within 6Å (Chen and Weng 2002).

Table 4-1 Atomic Contact Energy e_{ij} value (in energy unit of $RT/(q_r/2)$)^{*} of 18 atom types, adapted from Zhang *et al.*

	N	C ^α	C	O	GC ^α	C ^β	KN ^ζ	KC ^δ	DO ^δ	RN ^η	NN ^δ	RN ^ε	SO ^γ	HN ^ε	YC ^ζ	FC ^ζ	LC ^δ	CS ^γ
N	-0.724	-0.903	-0.722	-0.322	-0.331	-0.603	1.147	0.937	0.752	0.502	0.405	0.495	-0.116	-0.092	-0.412	-0.499	-1.005	-2.060
C ^α		-0.842	-0.850	-0.332	-0.365	-0.531	1.139	0.918	0.852	0.452	0.445	0.436	-0.044	-0.165	-0.539	-0.590	-1.123	-2.028
C			-0.704	-0.371	-0.308	-0.499	1.140	0.846	0.788	0.490	0.408	0.418	-0.045	-0.089	-0.415	-0.478	-0.963	-2.033
O				-0.016	0.205	-0.059	1.414	1.075	1.152	0.831	0.758	0.649	0.383	0.241	-0.073	-0.166	-0.650	-1.650
GC ^α					0.182	-0.009	1.502	1.385	1.192	0.912	0.872	0.943	0.434	0.392	-0.062	-0.021	-0.342	-1.212
C ^β						-0.469	1.310	1.010	0.859	0.671	0.591	0.565	0.122	0.000	-0.438	-0.533	-1.034	-1.800
KN ^ζ							3.018	2.911	1.157	2.635	1.848	2.699	1.587	1.557	1.004	1.340	1.252	0.700
KC ^δ								2.811	0.989	2.439	1.622	2.525	1.361	1.277	0.685	0.938	0.814	0.411
DO ^δ									-1.978	0.695	1.386	0.641	1.002	0.642	0.618	0.894	0.792	-0.029
RN ^η										-1.589	1.395	1.515	1.022	0.948	0.457	0.707	0.586	0.021
NN ^δ											1.300	1.283	0.931	0.829	0.484	0.633	0.389	-0.046
RN ^ε												1.309	0.921	0.777	0.265	0.484	0.274	-0.253
SO ^γ													0.559	0.319	0.301	0.212	-0.155	-0.949
HN ^ε														-0.301	-0.159	-0.132	-0.338	-1.324
YC ^ζ															-0.314	-0.478	-0.964	-1.410
FC ^ζ																-0.687	-1.240	-1.784
LC ^δ																	-1.873	-2.402
CS ^γ																		-3.742

^{*}1 $RT/(q_r/2) = 1/21$ kcal/mol, in which q_r is the average of the coordination number in Equation 13

4.1.5. Pairwise interactions with Decoys as Reference States (DARS)

One feature that makes PIPER different from previous docking approaches is the development of potentials with Decoys as the Reference State (DARS) (Chuang, Kozakov *et al.* 2008). DARS is a derived structure-based intermolecular potentials. With the assumption that the atom-type related properties determining the distribution of pairwise interactions should be removed for the reference state, a large decoy set of docked conformations based only on good shape complementarity as the scoring function was generated, and the frequency of interactions in the decoys was observed (Chuang, Kozakov *et al.* 2008).

Unlike the implementation of the shape complementarity, the pairwise potential E_{pair}

$$E_{pair} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \epsilon_{ij} \quad (20)$$

does not have the form of a correlation function, but it can be written as a sum of correlation functions, which is based on eigenvalue-eigenvector decomposition of the pairwise interaction matrix of the element ϵ_{IJ} (when interacting atoms belong to atom type I and J). According to Boltzmann approach, a statistical potential between atoms a_i (type I) and a_j (type J) within a certain cutoff distance is defined as

$$\epsilon_{IJ} = -RT \ln(P_{IJ})$$

where R and T are gas constant and temperature, respectively. P_{IJ} is the probability of two interacting atoms. This probability can be approximated by the frequency:

$$P_{IJ} = \frac{\text{observed number of interacting atom pairs}}{\text{expected number of interacting atom pairs in reference state}} = \frac{v_{IJ}^{obs}}{v_{IJ}^{ref}}$$

Assume the state of the protein is fully determined by the interactions among its interaction sites,

E_{pair} is satisfied when:

$$\begin{cases} \epsilon_{ij} = \epsilon_{IJ} & \text{if } r_{vdw} + 2\text{\AA} < r_{ij} < 6\text{\AA} \\ \epsilon_{ij} = 0 & \text{if } r_{ij} > 6\text{\AA} \end{cases}$$

4.1.6. Comparison between atom-level (ACE) and residue-level (DARS) potentials

The structure-based potentials can be categorized into two groups: atom-level potentials (using Atomic Contact Energy or ACE) and residue-level potentials (using DARS) (Kozakov, Brenke *et al.* 2006, Chuang, Kozakov *et al.* 2008). In this report, comparisons of 2000 predictions of both the Enzyme/Inhibitor and Antibody/Antigen complexes are performed using the two potentials DARS (in PIPER) and ACE (in ZDOCK2.3). Performance is evaluated using “percentage of hits”, which is the near-native structures ($\text{RMS} \leq 10\text{\AA}$) as similar to Chuang’s study (Chuang, Kozakov *et al.* 2008). The overall results shows that docking using ACE potential outperforms the other, except for one case of antibody/antigen complex (*1FBI*).

$$\text{Percentage of hits} = \frac{\# \text{ number of near-native structures (RMS} \leq 10\text{\AA})}{\# \text{ total number of predictions (N=2000)}}$$

The ACE-employing docking maintains high percentages of hits ($\geq 70\%$) for most of the Enzyme/Inhibitor complexes as shown in Table 4-2 and Figure 4-2. In the two cases of *1AVW* and *1BRS*, both ACE and DARS result in lower performance (i.e. 31.75% and 37.5% for ACE, and 0% and 3.25% for DARS, respectively). Binding interfaces of the two complexes reveal that interactions between the enzymes and the inhibitors in these complexes are driven by most of hydrophilic and planar residues. This observation explains why the shape complementarity based scoring functions (e.g. in both ZDOCK and PIPER) do not work well, especially for DARS which favors pairing of hydrophobic patches in the interface (Chuang, Kozakov *et al.* 2008).

In most cases of the Antigen/Antibody complexes, DARS as well as ACE did not give good results (Table 4-3 and Figure 4-3). These results were also observed by (Chuang, Kozakov *et al.* 2008). Docking Antigen/Antibody complex is much more challenging than docking Enzyme/Inhibitor complexes. Especially in DARS, the training set cannot represent specific interactions in Antigen/Antibody complexes as it is biased toward complexes of homodimers containing good complementarity (Chuang, Kozakov *et al.* 2008).

Table 4-2 Percentage of hits obtained in the top 2000 docked Enzyme/Inhibitor complexes using DARS (PIPER) and ACE (ZDOCK2.3).

Complex*	Percentage (%)		Complex*	Percentage (%)	
	DARS	ACE		DARS	ACE
1ACB	70.2	78.6	1CHO	93.8	100
1AVW	0	31.75	1CSE	34.2	88.95
1BRC	1.6	85.05	1DFJ	86.95	92.75
1BRS	3.25	37.5	1FSS	71.5	86.9
1CGI	27.55	87.5	1MAH	0	77.5

*Experiments are performed using complexes as similar to Chuang's study (Chuang, Kozakov et al. 2008).

Table 4-3 Percentage of hits obtained in the top 2000 docked Antigen/Antibody complexes using DARS (PIPER) and ACE (ZDOCK2.3)

Complex*	Percentage (%)		Complex*	Percentage (%)	
	DARS	ACE		DARS	ACE
1AHW	28.75	31.7	1FBI	63.75	54.35
1BQL	43.05	52.65	1IAI	0	1.7
1BVK	0	40.75	1JHL	0	50.05
1DQJ	7.05	56.05	1MEL	5.8	12.05
1EO8	0	5.25	1MLC	27.45	45.1

*Experiments are performed using complexes as similar to Chuang's study (Chuang, Kozakov et al. 2008).

Figure 4-2 Docking performance using DARS (green) and ACE (red) for Enzyme/Inhibitor complexes according to percentage of obtained number of near-native structure found (y-axis) on 2000 docked conformations (x-axis).

Figure 4-3 Docking performance using DARS (green) and ACE (red) for Antigen/Antibody complexes according to percentage of obtained number of near-native structure found (y-axis) on 2000 docked conformations (x-axis).

4.2. Comparisons of various scoring functions

A comparison of performances (*without the refinement*) among different scoring functions, i.e.

Potentials used in the scoring function	Docking algorithm used
SC and ELEC as a filter	FTDOCK
SC+ELEC+DE using ACE/IFACE	Several versions of ZDOCK
SC+ELEC+DARS	PIPER

demonstrates that the combination of shape complementarity, electrostatics, and desolvation using IFACE (e.g. as in ZDOCK3.0.2) outperforms the others. In addition, results of protein complex structures obtained from the ZDOCK3.0.2 and followed by MD refinement are also showcased to highlight the importance of the MD refinement to obtain better optimal structures, especially in terms of structural flexibility and conformational changes.

4.2.1. Data

Human Leukocyte Antigen (HLA) is a major histocompatibility complex (MHC) in human. The HLA

is responsible for presenting the antigen (as a peptide) to the T-cell receptor (TCR), triggering the immune responses (Rudolph, Stanfield *et al.* 2006). Therefore, binding of the peptide to the HLA plays a critical point in the antigen presentation. The stronger the binding is, the more likely the peptide is recognized by the TCR. The interactions between the HLA and its peptide occur at the helical binding groove, which is made of two domains α_1 , α_2 and eight stranded β -sheet arising from both the domains of the HLA (Figure 4-4). The peptide binds to this region using its two termini as anchors to the binding pockets of the HLA (DiBrino, Parker *et al.* 1995, Reinherz, Tan *et al.* 1999, Macdonald, Purcell *et al.* 2003); consequently making the middle part of the peptide expose to the solvent and recognized by the TCR.

Figure 4-4 HLA-B*4405/EBV-peptide complex [PDB:3DX8]. HLA chains A, B, and the EBV peptide are presented in green, cyan, and red, respectively in the front view (left) and the side view (right).

In this part, interactions of the Human Leukocyte Antigen (HLA-B*4405) and its 10-mer Epstein-Barr Virus (EBV) epitope [PDB:3DX8] are simulated. It was found that the T-cell mediated immunity was more enhanced when this peptide bound to the HLA-B*4405 compared to other HLAs in the HLA-B44 family (Archbold, Macdonald *et al.* 2009). The bound HLA-B*4405/EBV-peptide complex is separated into two components, i.e. HLA-B*4405 (hereafter called HLA) and the EBV peptide (or called peptide), and their bindings are simulated using various docking algorithms together with the MD simulation for structural refinement. Correct bound conformations and false positives of these complexes are evaluated and followed by cross confirmation using the original crystalized complex structure 3DX8.

The HLA consists of chain A (276 residues) and chain B (99 residues). Chain A is a HLA class I

histocompatibility complex and chain B is a β -2-microglobulin. The HLA chain A contains two domains α_1 and α_2 that constitute a helical antigen-presenting groove (or called the binding groove) to which the 10-residue peptide will bind.

4.2.2. A comparison of docking performances among different scoring functions

Dockings of the HLA (receptor) and the peptide (ligand) are performed using some current rigid docking algorithms that apply various correlation-based scoring functions for their ranking step. In addition to the core shape complementarity (SC), these algorithms employ other physicochemical properties that are implemented as different potentials into the scoring functions using Fast Fourier Transform technique. The potentials are electrostatics (ELEC), desolvation (DE), pairwise shape complementarity (PSC), atomic contact potential (ACE), or decoys as reference states (DARS). Performance of these scoring functions SC/ELEC (FTDOCK), PSC (ZDOCK2.1), PSC+ELEC+DE using the conventional ACE (ZDOCK2.3), PSC+ELEC+DE using IFACE (ZDOCK3.0.2), and DARS (PIPER) to simulate structural formations of the HLA/peptide complexes is demonstrated in Figure 4-5 below.

Figure 4-5 Docking performances (*without refinement*) of different scoring functions in terms of percentage of number of near-native structures found in 2000 docked conformations.

In the results of the HLA/peptide complex prediction, the docking performances follow an increasing trend when number of formed conformations increases (shown in Figure 4-5). For examples, SC/ELEC, PSC, PSC+ELEC+DE_{ACE}, PSC+ELEC+DE_{IFACE}, and DARS obtains small numbers of near-native structures (or hits with root-mean-square deviation $RMS \leq 5\text{\AA}$) in the first 10 conformations, i.e. 2, 8, 7, 10, and 10 structures, respectively. These values quickly increase when number of conformations reaches 1000 conformations (377, 511, 629, 726, and 539) and 2000 conformations (728, 933, 1213, 1181, and 1054).

In addition, it is observed that the combination of PSC+DE+ELEC performed better than the PSC alone as shown in Figure 4-5. This result is in agreement with that of Chen's study (Chen, Li *et al.* 2003). Further, as compared to DARS, which also applies pairwise shape complementarity, the extra contributions of desolvation (DE) potential helps increase number of correct bindings for this type of HLA/peptide complexes (Figure 4-6).

In fact, it is demonstrated that the PSC+ELEC+DE_{IFACE} shows the best performance as compared to the other potential combinations. In order to evaluate the outperformance of the PSC+ELEC+DE_{IFACE} over the others, their performances are detailed using the best, correct, and false positive conformations. In this result, the terms "best", "correct", and "false positive" conformations are *locally* defined according to the 2000 retrieved structures. The "correct" structures are the conformations in which the peptide binds to the helical binding groove as illustrated in Figure 4-4. The "best" structure is the "correct" conformations that obtain the highest rank (top-1). The "false positive" is defined as the conformation in which the peptide lies out of the binding groove, but is highly ranked. In fact, the number of false positives is assumed as the total number of the incorrect-bound conformations found in the first 100 ranks. Examples are shown in Table 4-4.

The values in Table 4-4 are docking scores determined according to the docking algorithms. Only scores of the "best" and "correct" structures among the top 10 docked conformations are given as illustrations in the table. It depicts that for the PSC, the first conformation is the "best" structure (i.e. the peptide is in the binding groove and ranked 1), the "correct" structures are conformations 1 and 3, and there are 8 false positives (i.e. conformations 2 and 4-10). Similarly, the combination of PSC+ELEC+DE_{ACE} obtains 4 correct and 6 false positives. DARS results in 6 correct and 4 false positives respectively, while the PSC+ELEC+DE_{IFACE} generates all correct, and the

SC/ELEC does not obtain any correct bindings in the first 10 generated structures.

Table 4-4 The first 10 docked conformations used as illustrations of the best, correct, and false positive results of the docking methods

Rank	SC/ELEC	PSC	PSC+ ELEC+DE _{ACE}	PSC+ ELEC+DE _{IFACE}	DARS
1	-	16.64	41.34	1672.71	996.0
2	-	-	30.29	1448.03	908.9
3	-	13.9	28.76	1366.46	845.6
4	-	-	-	1339.40	-
5	-	-	-	1316.02	836.3
6	-	-	26.44	1314.59	794.6
7	-	-	-	1289.02	783.7
8	-	-	-	1261.95	-
9	-	-	-	1241.97	-
10	-	-	-	1239.80	-

Only scores of the “best” and “correct” structures among the top 10 ranked conformations are shown. The docked scores of the other conformations of the first 10 structures could be found in Appendix A3.

According to these definitions, percentages of correct and false positive conformations in the 2000 structures are calculated and plotted in Figure 4-6 to evaluate the docking performance for all the scoring functions.

Illustrated in Figure 4-5 and Figure 4-6, the combination of PSC+ELEC+DE_{IFACE} performs the best: 59.05% correctly—for total 2000 predictions of HLA/peptide complexes—and with fewest number of false positives (2%). While Chuang *et al.* proved that the DARS potential were superior over the atomic contact energy (ACE) (Chuang, Kozakov *et al.* 2008), in our results DARS obtains 52.85% and 29% of correct and false positives respectively. Further, it is also observed that DARS even achieves a lower success rate than the PSC+ELEC+DE_{ACE} although it generates fewer false positives.

Following these brief comparisons, the combination of pairwise shape complementarity (PSC), desolvation energy (DE), and electrostatics (ELEC) is considered as a potential scoring function to determine good surface matches between the receptor and the ligand. However, it is noticed in these results and also supported by Chen *et al.* (Chen, Li *et al.* 2003, Chen and Weng 2003) that the PSC is prone to identify large concave binding pockets, consequently resulting the false positives.

Figure 4-6 Docking performance of various docking scoring functions SC/ELEC, PSC, PSC+ELEC+DE_{ACE}, PSC+ELEC+DE_{IFACE}, and DARS in terms of corrects and false positives for 2000 generated complex structures.

4.3. Rigid docking results refined by Molecular Dynamics simulation

4.3.1. Molecular Dynamics (MD) simulation gives better discrimination of near-native structures from false positive ones

The MD refinement results in a solid discrimination of the correct conformations from the false positive ones. In this result part, 10 HLA/peptide complex structures generated from the scoring function PSC undergo the MD refinement using AMBER10. The process of 5 nano-second (ns) MD equilibration stage creates 500 conformations for each of the 10 complexes. Then, from these 500 conformations, the lowest energy conformation (assumed as the optimal structure) is located and extracted.

Structural presentation of the optimal structures of the 10 complexes (Figure 4-7) indicates that the peptides in complexes 1 and 3 are clearly located at the helical binding groove compared to the rest 8 complexes. The two structures also reach equilibrium with significant lower energies (around -117,000 kcal/mol) than the others (i.e. around -114,000 kcal/mol and higher) (Table 4-5).

Confirmed by using the known native crystal structure [PDB:3DX8], the two complexes 1 and 3 are the near-native bound structures. In fact, the two correct structures are only 1.48 Å and 1.42 Å deviated from the native structure, respectively, whereas the others are more than 4.8 Å in difference in the same time scale of 5 ns. This suggests that the near-native structures be remarkably discriminated from the other false positives using MD energy evaluation.

However, there is an orphan case. The complex 4 obtains its lowest energy $\sim -118,500$ kcal/mol, which is even lower than the energies of the two correct bound complexes (Table 4-5). Yet, the RMSD value of this complex is 4.56 Å, which reflects an incorrect binding. It is a false positive case, and this is confirmed using the native structure of the HLA/peptide bound complex.

4.3.2. Molecular Dynamics simulation allows more flexibility

Flexibility plays a focal point to induce conformational changes upon protein associations during docking process. However, only certain flexibility is offered in rigid docking for computational simplicity and efficiency, so some assumptions are needed. For example, backbones are kept rigid and flexibility is restricted to side chains only (Wang, Bradley *et al.* 2007).

The impact of flexibility driven by MD refinement is demonstrated in a comparison of backbone conformational changes of the best structures obtained above, i.e. rank-1 conformations of the PSC and the PSC+ELEC+DE_{IFACE} before and after the MD simulation. To ensure the complex structure would not be trapped in any local minima, the MD equilibration stage is performed in longer time (20ns); and the lowest energy structures at milestones 5ns, 10ns, and 20ns are extracted to invest the changes in backbone conformations of both the peptide (ligand) and the HLA (receptor).

In the HLA-B*4405/peptide complex, two peptide residues Glu² and Phe¹⁰ represent the two anchors that bind to the important binding pockets B and F on the HLA, respectively. These bindings cause the solvent-exposure of the middle part of the peptide to facilitate T-cell receptor contacts (Archbold, Macdonald *et al.* 2009); hence, trigger the immune responses. Therefore, it is expected that conformational changes in these residues could also affect the binding of the peptide to the HLA.

Figure 4-7 Structural presentations of the 10 lowest energy structures. The complexes 1 and 3 show correct bindings of their peptides (red and blue, respectively) to the binding groove, while the peptides of the other 8 complexes (in other colors) are located outside of the groove. The complexes are presented in cartoons in which the HLA structure is in gray.

Table 4-5 Ten optimal structures obtained after 5ns MD refinement of the PSC results

HLA/peptide complex	Energy (kcal/mol)	RMS (Å)	Quality of the binding*
1	-117,504.5341	1.48	Correct
2	-113,367.0612	5.52	False positive
3	-116,977.9455	1.42	Correct
4	-118,503.8579	4.56	False positive
5	-114,167.8167	6.65	False positive
6	-112,002.8978	5.21	False positive
7	-113,712.2248	4.95	False positive
8	-113,270.7119	4.82	False positive
9	-113,226.9971	5.25	False positive
10	-113,355.7885	5.57	False positive

*The quality of the binding (correct and false positive) is defined in the previous result part. The RMS is the deviation of each structures from the native bound HLA/peptide complexes retrieved from Protein Data Bank [PDB:3DX8].

The peptide interacts with the helical binding groove of the HLA using residues Glu¹, Glu², Asn³, Arg⁹, and Phe¹⁰. These interactions occur in all the three complexes of both the scoring functions and the MD simulation. However, there is an extra polar contact between the residues Phe¹⁰ (peptide) and Asn⁷⁷ (HLA) that could only be found in the MD-derived complex Figure 4-8). The extra Asn⁷⁷-Phe¹⁰ contact does not occur in the rigidly docked complex from both the scoring functions, the conformations of which are totally fitted (Asn⁷⁷) (Figure 4-8A and Figure 4-9). It was observed that the HLA backbone is kept rigid during the docking, so there are no conformational

changes expected in the HLA backbone among these docked complexes. On the other hand, during the MD simulation of 5ns, 10ns, and 20ns, the Asn⁷⁷ conformations keeps changing compared to the rigid docked complexes (Figure 4-9). This fluctuation therefore slightly changes the distances, yet maintains the interactions (Figure 4-8B).

In addition to detecting the flexibility of the receptor (e.g. the residue Asn⁷⁷), flexibility of the peptide is also observed. Three torsion angle (ω, ϕ, ψ) values of the 10-mer peptides (Appendix A4) indicate that both the scoring function keep the peptide backbone neatly rigid when it is docked to the HLA. Since the dockings are performed using two protein components derived from an original bound HLA/peptide complex, the backbone torsion angles of the peptides in these two docked complexes are more than 99% similar to that of the original complex.

On the other hand, more movement is observed in the peptide backbone of the MD complex compared to the original bound complex (Appendix A4). Most changes occur at two dihedral angles ϕ, ψ around the bonds C α -N and C α -C respectively, especially at residues Leu⁴, Leu⁵, Asp⁶, and Phe⁷ which are all the peptide solvent-exposed residues. These flexible contributions might also intervene and facilitate the peptide recognition from the T-cell receptor. Generally, these preliminary results show that MD simulation could give flexibility not only to the protein side chains, but also to the backbones, compared to the other rigid dockings.

Further, taking advantage from previous results in which the PSC+ELEC+DE_{IFACE} outperforms the other scoring functions, a combination of the PSC+ELEC+DE_{IFACE} and MD refinement is applied. Initially, 2000 docked HLA/peptide complexes are generated using the PSC+ELEC+DE_{IFACE}. As it showed the best performance among various scoring functions shown above, the top-ranked complexes are most likely correct bindings. However, in order to obtain reasonable computational cost for this part, only 10 top complexes are selected to perform the MD refinement. These 10 best HLA/peptide complex structures undergo an explicit solvent MD process of short minimization, heating, and equilibration. To avoid local minimum trap, the MD is performed more intensively in 20ns.

Figure 4-8 The Asn⁷⁷-Phe¹⁰ contact that occurs only in the MD-refined HLA/peptide complex. The 5ns, 10ns, and 20ns conformations are shown in **red**, **green**, and **blue** lines, respectively. The PSC and PSC+ELEC+DE_{IFACE} conformations are perfectly superimposed and presented in **purple** stick. The dotted lines represent the polar contact between the Asn⁷⁷ (HLA) and Phe¹⁰ (peptide) and the distances are shown. There is no such the contact in both the PSC and PSC+ELEC+DE_{IFACE} complexes. (A) and (B) have the same color scheme.

Figure 4-9 A closer view of the Asn⁷⁷-Phe¹⁰ contact in the HLA binding groove. The residues Asn⁷⁷ and Phe¹⁰ of the PSC, PSC+ELEC+DE_{IFACE}, and MD (5ns, 10ns, 20ns) complexes are shown in **orange**, **purple**, **red**, **green**, and **blue** line presentations. The original HLA-B*4405/peptide bound complex [PDB:3DX8] is presented in **yellow** stick and ribbon. The Asn⁷⁷ of the PSC and PSC+ELEC+DE_{IFACE} conformations are totally fitted with the original one.

There are 2000 conformations generated during the equilibrium state. All of them obtain an average energy of $\sim -117,500$ kcal/mol, and the lowest energy is $-118,363$ kcal/mol at time point of 4.63ns. Compared to the previous combination of the PSC and MD (with the energy of the optimal structure at $-117,672$ kcal/mol), this optimal structure obtains much lower energy in the same simulation time of 20ns (Figure 4-10).

How the two resulted structures (from the PSC+ELEC+DE_{IFACE}/MD and the PSC/MD) are near to the native structure is estimated by calculating ligand RMS (LRMS). The LRMS is the deviation of ligand positions in the two complexes as compared to the ligand position in the original bound complex. It shows that the combined PSC+ELEC+DE_{IFACE}/MD complex is $0.21\text{\AA}$ ($1.934\text{\AA}$ before MD) deviated from the original crystal structure while the deviation of the other is $0.1\text{\AA}$ ($0.636\text{\AA}$ before MD). This suggests that the combination of the PSC+ELEC+DE_{IFACE} and MD bring a better prediction for protein structure association in both energy and structure deviation evaluations.

Figure 4-10 Combination of PSC+ELEC+DE_{IFACE} and MD shows better performance than the combined PSC and MD.

4.4. Conclusion for this chapter

Different potentials of shape complementarity (SC), electrostatics (ELEC), desolvation (DE), decoys as reference states (DARS) are developed for a better ranking solution to discriminate the correct predicted structures from the false positive ones. Comparisons among the potentials, especially in atom-level (i.e. using atomic contact energy or ACE) and in residue-level (i.e. using

DARS) reveals the close dependency of these potentials on physiochemical characteristics of the binding interfaces. In fact, while the ACE is prone to identifying large concave binding pockets, DARS is on the other hand more favored in hydrophobic interfaces. Further, both these potentials are implemented along with shape complementarity, making them more selective for Enzyme/Inhibitor complexes rather than Antibody/Antigen complexes. This results in a remaining challenge in protein docking area, especially in predicting structures of the Antibody/Antigen complexes, which require induced conformational changes during the protein associations.

These potentials are implemented in various scoring functions of different docking algorithms, which firstly perform the molecular surface matching on grid presentations and then scoring the matches in reasonable calculation time. These algorithms therefore provide substantial accuracy but computational efficiency. However the exhaustive searching properties result in the performances that make the approach still far from an ideal model for the predictions of complex formation due to its flexibility limitations (Table 4-6) and its difficulty in discriminating near-native structures from the false positives in some protein complexes. Thus the further MD refinement step is still required to obtain the optimal docked conformations.

Table 4-6 Overall summaries about limitations and performances of various scoring functions. Performances were evaluated for 2000 generated conformation.

Scoring function	Complexity	Limitations	Performance (%)	
			Correct	False pos
SC & uses ELEC as a filter criterion	$N^3 \ln(N^3)$	An exhaustive search is performed; consequently producing 'noisy' outputs in which number of incorrect conformations is larger than that of correct ones. Association at atomic level is not considered, so the two docked components need to be close to their native conformations in order to achieve the best results.	36.4	66
PSC+ELEC+DE _{IFACE}	$N^3 \log_2(N^3)$	DE does not performs well for <i>Antibody/Antigen</i> complexes because these complexes mostly interact predominantly via charge-charge and charge/polar interactions and contain interfacial water molecules which are not explicitly accounted . Combination of pairwise shape complementarity (PSC), electrostatics, and DE cannot dock bound complex with weak binding affinities because PSC is prone to identifying large concave binding pockets.	59.05	2
PSC+ELEC+DARS	$N^3 \log(N^3)$	Training dataset is <i>Enzyme/Inhibitor</i> biased. Hence, the results can be best achieved for oligomeric and <i>Enzyme/Inhibitor</i> complexes. DARS is used effectively if the interfaces are in good complementarity and hydrophobic	52.7	29

Chapter 5

Improving the discrimination of near-native complexes for protein rigid docking by implementing interfacial water into protein interfaces

This chapter describes the proposed re-ranking technique and how it is implemented in this project to improve the ability of discriminating the correct predictions of protein formation from other incorrect ones. The effect of solvent is **explicitly** reflected by using “*actual*” water molecules that are added into the protein interfaces during the re-ranking. The combinations of $PSC + ELEC + DE_{IFACE}$ together with the water contributions via the energies of protein-water dispersion and water-mediated Hydrogen bonds as well as the free energy change of the complex interfaces are implemented and weighted in an energy-based scoring function. As expected, different physicochemical characteristics of protein complexes are partially reflected in this implementation via the weight values. It is indicated that the proposed method improves results of the initial rigid docking in both numbers of near-native structures found and the ranks of the correct hits.

5.1. Introduction

Importance of protein-protein interactions has been long established by either experiments or computational simulations (Kastritis and Bonvin 2013). Understanding well the protein interactions is crucial for designing drugs and finding drug targets. While knowledge of protein interactions and their molecular pathways have been discovered experimentally, limited information about structures of known protein complexes could be elucidated. In addition, due to the transient or obligatory associations, not every protein complex could be experimentally crystalized. Tackling this issue of predicting the formation of protein complexes has actually given plenty room for various protein docking algorithms to evolve and has therefore placed a solid foundation in studying protein interactions.

When proteins interact, however, there are uncountable numbers of issues that cannot be foreseen, especially by *in silico* methods due to the flexibility of the molecules against limited computational resources. Hence, protein-protein docking algorithms have branched into rigid docking with limited flexibility but further refinement and optimization required for computational efficiency (Gabb, Jackson *et al.* 1997, Jackson, Gabb *et al.* 1998, Chen, Li *et al.* 2003, Chen and Weng 2003, Wang, Bradley *et al.* 2007, Chuang, Kozakov *et al.* 2008, Kozakov, Schueler-Furman *et al.* 2008, Ritchie, Kozakov *et al.* 2008, Li, Guo *et al.* 2011, Solernou and Fernandez-Recio 2011) or flexible docking with intensive computations (Taylor and Burnett 2000, Fernandez-Recio, Totrov *et al.* 2001, Noy and Goldblum 2010). For efficient computations, rigid docking followed by refinement and optimization is therefore more concerned and developed.

To further improve results of rigid docking, not only are optimal searching solutions required (Gray, Moughon *et al.* 2003, Solernou and Fernandez-Recio 2010), effective ranking protocols are also essential to successfully discriminate the correct predictions from the wrong or false positive ones (Pierce and Weng 2007, 2008, Bourquard, Bernauer *et al.* 2011, Vreven, Hwang *et al.* 2011). For examples, ZRANK, a re-ranking method improves significantly the results of the rigid docking program ZDOCK (see Chapter 2 section 2.1 for the comparison). One of the factors that have been implemented and shown to improve dramatically the quality of protein docking algorithm is the effect of solvent by using solvation energy or desolvation term. However, the inclusion of the solvent or water is only implicitly accounted or in other words, the water molecules

are not actually presented in the protein-protein interactions.

Water is crucial for protein activities and plays as an important solvent for most of interactions in the cell. Water molecules are found in protein–protein and protein–DNA interfaces. Despite quick exchanges between interfacial water and the bulk solvent, some water-mediated interactions contribute directly to the stability of protein formation and specificity of protein recognitions (Janin 1999). However, water effect on the behavior of proteins is only partially understood. The issue therefore leads to various studies to investigate the relation between dynamics of the hydration water and the dynamics of protein (Franzese, Bianco *et al.* 2010, Ahmed, Spyraakis *et al.* 2011).

Water contributes to conformational stability of proteins due to its ability of simultaneously accept and donor hydrogen bonds; consequently either facilitates a hydrogen bond network to bridge polypeptide atoms or prevents interactions of protein atoms with the bulk solvent (Xu, Tsai *et al.* 1997, Petukhov, Cregut *et al.* 1999). Further, water molecules at protein interfaces are also well studied in protein design and packing (Joachimiak, Kortemme *et al.* 2006, Jiang, Althoff *et al.* 2008). In protein interfaces, water is known for different roles such as bridging the two proteins, simply interacting with one of the two proteins, or even just being trapped in the interfaces without any effect on the protein interactions (Ahmed, Spyraakis *et al.* 2011).

However, for some complexes such as Antigen/Antibody or Enzyme/Inhibitor complexes, water plays an important role in their recognitions and interactions as well (Bhat, Bentley *et al.* 1994, Huang, James *et al.* 1995). Frequently, Antigen/Antibody interfaces are prone to less hydrophobic while the Enzyme/Inhibitor interfaces more hydrophobic; however, studies have shown that both types of complexes exist either wet or dry interfaces (Janin 1999) and that wet interfaces with water-filled cavities maintain the close pack of the structures and facilitate the protein interactions via water-mediated hydrogen bonds or indirect interactions (Ikura, Urakubo *et al.* 2004).

So far water is only explicitly implemented in protein design for side chain packing (e.g. Rosetta packages (Commons, Dantas, Kuhlman *et al.* 2003), Solvated Rotamer (Jiang, Kuhlman *et al.* 2005)) or in protein-ligand docking (Lemmon and Meiler 2013). And the docking protocol HADDOCK v2.0 by van Dijk *et al.* (Dijk and Bonvin 2006), which introduced water into modeling of biomolecular complexes elicits that the presence of explicit water in protein docking is feasible and promising.

As described in Chapter 3, docking approach of HADDOCK is different from the conventional FFT-based docking algorithms and while there is more room to improve and develop from the current rigid docking programs, e.g. ZDOCK, this project aims to improve the discrimination of near-native structures from the false positives resulted from the rigid docking, i.e. ZDOCK3.0.2, by implementing interfacial water into the protein interfaces. With advantages of water in the interfaces and its important role in the protein-protein interactions, especially for Antigen/Antibody complexes, the implementation will present the presence of water explicitly during the evaluation of protein interactions using a proposed energy-based scoring function for re-ranking. Therefore, some extent of flexibility upon protein associations will be expected as the water-mediated interactions are exclusively accounted.

5.2. Process of the implementation

5.2.1. Data collection and preprocessing

A dataset extracted from the latest protein-protein benchmark 4.0 (Hwang, Vreven *et al.* 2010) is used in this work. The dataset includes non-redundant structures of protein-protein complexes of Enzyme/Inhibitor (E), Antibody/Antigen (A/AB), and other types (O). These structures are classified into 3 groups (easy, medium, and difficult) according to the CAPRI-defined difficulty level that the current protein-protein docking algorithms can handle. In addition, because only the interfaces of one-receptor versus one-ligand complexes are interested and investigated in this project, only complexes that contain monomer ligands from the benchmark 4.0 are selected and included in the dataset, which yields a total of 159 cases. The dataset contains 113 easy cases (for which simple rigid-body docking algorithms can successfully predict the protein complex structures), 26 medium, and 20 difficult cases (predictions of which are significantly deviated from the native structures, i.e. interfacial RMS > 2.5Å (Hwang, Vreven *et al.* 2010)).

To further quantify water level at the protein interfaces, statistics of interfacial water is calculated for each of the crystalized protein complexes to highlight the effect of water on the protein interactions. It is observed that a percentage of 67% (130/159 complexes) of the dataset includes water in the structures (Appendix A5). Among the difficult cases, 16 out of 20 cases contain at least 6 water molecules buried in the protein interfaces. Hence, it is expected that taking into

account the interfacial water molecules and water-mediated interactions contributes significantly to the prediction of protein-protein associations.

5.2.2. Implementation of interfacial water in the proposed energy-based scoring function

The proposed scoring function is implemented based on a linear combination model (Vogt and Cottrell 1999, Wu, Bi *et al.* 2009). According to this model, the relevance ρ of an individual x to a system q can be evaluated using a set of weights $W=(w_1, w_2, \dots, w_k)$ given to each of the k factors and their individual relevance ρ_i :

$$\rho(W, x, q) = \sum_{i=1}^k w_i \rho_i(x, q)$$

with $\sum_{i=1}^k w_i \neq 1$ to indicate that the factors are independent.

Hence in this report, to demonstrate the influence of protein interaction (x) onto the overall protein complex structure (q), the proposed scoring function is described as a linear combination of potential terms (k), each of which is given a weight (w_i) and its corresponding estimated value (ρ_i). As shown in Equation 21, the potential terms k consist of the complex-binding scores and free energies of the protein complex interfaces (including protein-protein and protein-water interactions), and are derived from the protein-protein docking algorithm and the interfacial interactions (applying the solvated rotamers from (Jiang, Kuhlman *et al.* 2005). The individual weights w_i are estimated using Monte Carlo simulation method.

The proposed scoring function is used to re-rank all the protein complexes and is estimated as:

$$f = w_1 \text{Score}_{\text{IFACE}} + w_2 E_{\text{LJ_repulsive}} + w_3 E_{\text{water-mediated-Hbond}} + w_4 \Delta G_{\text{interface}} \quad (21)$$

where the $\text{Score}_{\text{IFACE}}$ is derived from the combination of shape complementarity, electrostatics, and desolvation ($\text{SC}+\text{ELEC}+\text{DE}_{\text{IFACE}}$) employed in the initial rigid docking algorithm of ZDOCK3.0.2, $E_{\text{LJ_repulsive}}$ and $E_{\text{water-mediated-Hbond}}$ are energies of the Lennard-Jones repulsive and Hydrogen bonds involving water to represent the protein-water interactions, and $\Delta G_{\text{interface}}$ is the free energy change of the interface representing the interactions between the protein partners. The four parameters w_1, w_2, w_3, w_4 are separate weights of the according potentials in the

proposed scoring function f (as shown in Figure 5-2).

5.2.2.1. Combinations of pairwise shape complementarity, electrostatics, and desolvation

ZDOCK is the rigid docking algorithm to optimize functions of shape complementarity, electrostatics, and desolvation as linear combination of correlations using FFT algorithm (Chen and Weng 2002). ZDOCK3.0.2 is the latest version of this protein docking program in which the Interface Atomic Contact Energies (IFACE) is implemented in the desolvation term (Mintseris, Pierce *et al.* 2007, Pierce, Hourai *et al.* 2011). The implementation of the pairwise shape complementarity (SC), electrostatics (ELEC) in the program ZDOCK3.0.2 is similar to the description in Chapter 3 – section 3.1.1.2, and of the desolvation (DE) term is described in section 4.1.4.

The $Score_{IFACE}$ is approximated as a sum of the three potentials with different contributions:

$$Score_{IFACE} = 0.01 SC + 0.06 ELEC + DE_{IFACE} \quad (22)$$

which is obtained using the FFT algorithm for discrete functions on a 3D-grid of both the receptor and the ligand (Mintseris, Pierce *et al.* 2007). The searching space is performed on 15° rotational sampling, which yields 3,600 predictions and the top 2,000 predictions are selected for the further implementation.

Input : Atomic coordinates of Receptor R and Ligand L , *structure*

List of complex interfaces, *comp_interface*

List of ligand rotation, *lig_rot*, and translation, *lig_trans*

Output: Top 2000 predictions of the protein complexes

1. # Assign IFACE value for each atom in *structure R* or *L* using Table 4-1
 2. **for** each amino acid *aa* in *structure do*
 3. **if** *aa* $\in$ [*comp_interface*] **then**
 4. Assign IFACE according to Table 4-1
 5. **else**
 6. Ignore
 7. **end if**
 8. **end for**
 9. # Calculating the $Score_{IFACE}$
 10. **for** each (*lig_rot*, *lig_trans*) **do**
 11. Calculate $Score_{IFACE}$
 12. **end for**
 13. # Ranking
 14. Rank the $Score_{IFACE}$ in descending order
 15. Take predictions in the top 2000 ranks
-

Figure 5-1 Process of generating 2000 predictions to calculate the $Score_{IFACE}$

In this research, the searching is set exhaustively to cover the entire binding regions of the ligand. During this process, the searching space is limited to the rotational and translational space of the ligand interfaces, which involve only the interacting residues with respect to the receptor. The purpose of doing this is to minimize the chances of getting false positives which are the wrong predictions but obtain high ranks.

Observations of the FFT-based docking algorithms indicate that the docking score is obtained based on correlations of two grid-based discrete functions of the ligand and the receptor. Therefore, in order to sum linearly all the terms shown in the proposed function f (Equation 21) and also for the simplifying purpose, the ZRANK platform is chosen instead of the ZDOCK and the terms will be weighted separately. Hence, obtaining the $Score_{IFACE}$ is firstly performed for all the protein complexes of the dataset using ZDOCK3.0.2 (Figure 5-1).

5.2.2.2. Water-mediated contacts and the water effect on the protein interfaces

a. Preprocessing the predicted protein complexes: extracting protein interfaces and adding water molecules into the protein interfaces

The water effect and contacts are presented by the energies of the protein-water dispersion and the water-mediated Hydrogen bonds at the interfaces. Therefore, the interfaces of the protein complexes are extracted and followed by a process of adding water molecules. There is no water molecule found in the interfaces of the predicted complexes because no water is so far involved in the dockings of the protein complexes (Figure 5-3).

Extracting the complex interfaces are performed using INTERVOR (Loriot and Cazals 2010), a Voronoi-based algorithm modeling and computing macromolecular interfaces. Repeated interface extractions for all the crystalized complexes in the dataset are firstly performed and then mapped back to the original structures to confirm the reliability of the INTERVOR package. The results show that 159/159 interfaces of the protein complexes were successfully extracted and superimposed with the original structures. Only after being cross-confirmed of its fidelity, INTERVOR is used to extract the complex interfaces of all the 2000 predicted complexes in the same protocol.

Similarly, DOWSER is applied to explicitly add water molecules into the protein complex interfaces. DOWSER is a modeling package that investigates free energy of hydrophilic cavities and surfaces

of proteins and adds water molecules into the cavities in which water exists (Zhang and Hermans 1996). A simulation of adding water into interfacial cavities of protein complexes is performed initially for the dataset (159 crystal structures of protein complexes) to test the reliability and accuracy of the application in recovering internal water of the protein interfaces. It is demonstrated that 91.3% of internal water molecules are successfully recovered (where the predicted water positions are within a distance $\leq 1.5\text{\AA}$ from the water positions in the crystal structures). Full results of this water-recovering test are shown in Appendix A5. Therefore, it suggests that DOWSER is a reliable tool to recover the water in the protein crystalized structures and thus is used to predict the water positions for the interfaces of protein complexes in this project.

After water is added into each of the extracted interfaces of the protein complexes, structural minimization is conducted using AMBER10 (force field *ff99SB*) to stabilize the water positions with respect to the whole structure of the protein interface. Initially, 2500 steps of steepest descent are performed and followed by another 500 steps of conjugate gradient minimization. During the minimization, no restraint is applied and random initial velocities are used to maintain the robustness. The minimized water-containing structures of the protein interfaces are then subject to the next process of energy calculations.

b. Protein-water dispersion: $E_{LJ_repulsive}$

In the presence of the interfacial water, there exist repulsive forces between the water molecules (oxygen atom) and any other atoms in a distance of $3.6\text{\AA}$ (which is the maximum distance of the equilibrium inter-nuclear separation between oxygen and any atom). Contribution of the force to the energy changes of the protein interfaces is presented by the Lennard-Jones repulsive energy ($E_{LJ_repulsive}$), which involves interfacial water interactions. In the proposed scoring function, this energy terms is defined as the linear van der Waals repulsive potential:

$$E_{LJ_repulsive} = 4\varepsilon \left(\frac{\sigma}{r} \right)^{12}$$

where ε is the well-depth, r (with $r \leq 3.6\text{\AA}$) is the inter-nuclear distance, and σ is the sum of van der Waals radii of water ($1.4\text{\AA}$) and the interacting atom (Table 5-1).

Figure 5-2 Flow chart of the process of estimating the contributions of each potential term in the proposed scoring function to improve rankings for protein docking results using interfacial water

Table 5-1 van der Waals parameters (based on AMBER force field) used to calculate the Lennard-Jones repulsive energy in this project

Atom	σ (Å)	ϵ (kcal/mol)
C	1.7	0.173
N	1.55	0.179
O	1.52	0.2
S	1.8	0.2
H	1.2	0.063

c. Water-mediated Hydrogen contacts: $E_{\text{water-mediated-Hbond}}$

A solvated rotamer library developed by Jiang *et al.* offers this project a modeling approach, which is extended from the side-chain packing methods of protein folding, to evaluate water impact on the protein interfaces and to describe water-mediated Hydrogen bonds (Jiang, Kuhlman *et al.* 2005). Studies have shown that water attaching to protein side chains and mediating the Hydrogen bond network at protein interfaces contribute significantly to the changes of free energies in protein interactions (Schreiber and Fersht 1995, Burnett, Kellogg *et al.* 2000), hence stabilizing structures of protein complexes. Therefore, accounting the effect of interfacial water and energies of the water-mediated Hydrogen bonds into the proposed scoring function is expected to improve the ranks of the predicted protein complexes.

According to this library, a hydrogen bond that is considered as water-mediated needs to satisfy a number of geometric constraints of angles and distances around the water molecules (shown in Figure 5-3). The water-mediated Hydrogen bond energy is therefore distance-dependent and follows the energy function below:

$$E_{\text{water-mediated-Hbond}} = \begin{cases} E_{\text{Hb}} & \text{when constraints in Figure 5 – 3 are satisfied} \\ 0 & \text{otherwise} \end{cases}$$

where E_{Hb} is the Hydrogen bond energy evaluated by Jiang *et al.* from the solvated rotamer library.

Figure 5-3 Angle and distance constraints to define a water-mediated Hydrogen bond according to the solvated rotamer library. d is the distance between the water (W) oxygen and polar (acceptor A/donor D) atoms. H and AB are Hydrogen and Acceptor-Base atoms respectively.

5.2.2.3. Free energy change of the protein interfaces $\Delta G_{\text{interface}}$

The last energy term in the proposed scoring function (Equation 21) represents interactions of protein partners at the interfaces. In order to take into account the flexibility of the interfaces and the trade-off of the initial rigid docking, interactions that involve side-chains are included. Derived from the solvated rotamers by Jiang *et al.* (Jiang, Kuhlman *et al.* 2005), the free energy change of the interface ($\Delta G_{\text{interface}}$) includes: van der Waals interactions of protein atoms (E_{vdW}), hydrogen bond energies (E_{Hbond}) involving in side-chain_backbone (*sc-bb*), side-chain_side-chain (*sc-sc*), solvation energy (E_{sol}), which is used to complement the desolvation (DE from the $\text{Score}_{\text{IFACE}}$), and potential of side-chain and backbone torsions (E_{torsion}). The $\Delta G_{\text{interface}}$ is calculated using Equation 23 below:

$$\begin{aligned} \Delta G_{\text{interface}} = & E_{\text{vdW_attraction}} + 0.005 E_{\text{vdW_repulsive}} + \\ & E_{\text{Hbond_sc-bb}} + E_{\text{Hbond_sc-sc}} + \\ & 0.37 E_{\text{sol}} + 0.29 E_{\text{torsion}} \end{aligned} \quad (23)$$

in which the contributions of each term are referenced from the solvated rotamer library (Jiang, Kuhlman *et al.* 2005) and ZRANK (Pierce and Weng 2007). In this report, these term contributions are pre-estimated using 113 easy cases of the dataset. Observations of this pre-estimation results show that the $\Delta G_{\text{interface}}$ values noticeably discriminate the top-10 ranked structures from the other ranks. For examples, in the *1DQJ* case (easy, Antigen/Antibody complex), most of the structures in the top-10 ranks obtain $\Delta G_{\text{interface}} \leq \theta$ (with $\theta \approx -100$ kcal/mol) while most of the other lower ranks obtain the $\Delta G_{\text{interface}}$ greater than this value. The discrimination is also observed in some other

cases of protein complexes even though with different θ values. As a result, those parameters of weights in the Equation 23 are used for further calculations in this project.

The calculation of this $\Delta G_{interface}$ term is performed using PyRosetta (Chaudhury, Lyskov *et al.* 2010), the Rosetta package on the Python platform.

5.2.3. Estimation of weights w_1, w_2, w_3, w_4

The proposed scoring function to re-rank the protein complexes

$$f = w_1 \text{Score}_{\text{IFACE}} + w_2 E_{\text{LJ_repulsive}} + w_3 E_{\text{water-mediated-Hbond}} + w_4 \Delta G_{\text{interface}}$$

contains four separate weights w_1, w_2, w_3 , and w_4 representing different contributions of the four potential terms. Optimal values of these weights are obtained by performing Monte Carlo experiments for the whole dataset of 159 protein complexes which are clustered into four types of complexes, i.e. Antigen/Antibody (A), Antigen/bound-Antibody (AB), Enzyme/Inhibitor (E), and Others (O). The weights are estimated during the recurring experiments until the maximal numbers of hits converges. In this project, hits are defined as the near-native structures, RMS of which are within 2.5Å from the crystalized structures ($\text{RMS} \leq 2.5\text{\AA}$).

Input : 4 types of protein complexes (A,AB,E,O), *type*

sets of values ($\text{Score}_{\text{IFACE}}, E_{\text{LJ_repulsive}}, E_{\text{water-mediated-Hbond}}, \Delta G_{\text{interface}}$)

numbers of hits in the ground truth, *numGT*

Output: weights w_1, w_2, w_3, w_4

```

1.  for each type of protein complexes do
2.    numHitMax  $\leftarrow$  0
3.    iteration  $\leftarrow$  0
4.    while (numHitMax < numGT or numHitMax converges) do
5.      set random weights for ( $w_1, w_2, w_3, w_4$ )
6.       $f \leftarrow w_1 \text{Score}_{\text{IFACE}} + w_2 E_{\text{LJ\_repulsive}} + w_3 E_{\text{water-mediated-Hbond}} + w_4 \Delta G_{\text{interface}}$ 
7.      rank f in descending order
8.      numHit  $\leftarrow$  number of structures ( $\text{RMS} \leq 2.5\text{\AA}$ ) against the ground truth
9.      if numHit > numHitMax then
10.        numHitMax  $\leftarrow$  numHit
11.      end if
12.      iteration  $\leftarrow$  iteration + 1
13.    end while
14.    return ( $w_1, w_2, w_3, w_4$ )
15.  end for
```

Figure 5-4 Process of estimating the weights w_1, w_2, w_3, w_4 . In this process, the free energy change term ($\Delta G_{interface}$) is expected in significant contributions as observed in section 5.2.3.3. The input sets of the 4 potential terms are derived from 318,000 predictions (i.e. 159 cases x 2000 predictions).

5.3. Results and Discussion

5.3.1. Optimal weights w_1, w_2, w_3, w_4 of the proposed scoring function

$$f = w_1 \text{Score}_{\text{IFACE}} + w_2 E_{\text{LJ_repulsive}} + w_3 E_{\text{water-mediated-Hbond}} + w_4 \Delta G_{\text{interface}} \quad (21)$$

5.3.1.1. Reflection of weights on different physicochemical properties

As observed in Chapter 4 about the dependency of various potentials on different physicochemical properties of the protein complexes (i.e. interactions between the enzymes and the inhibitors are driven mostly by hydrophilic and planar residues, or methods based on the Atomic Contact Energy (ACE in the desolvation term) does not likely work well in predicting interactions of antigen and antibody complexes), it is expected that contributions of the four potential terms in Equation 21 would be diverse. Therefore, the weights of the proposed function f are estimated differently according to different types of protein complexes, i.e. Antigen/Antibody (A), Antigen/bound-Antibody (AB), Enzyme/Inhibitor (E), and Others (O).

After a number of iterations (averagely 40,000 steps) to estimate the weights using Monte Carlo method, the sets of weights (w_1, w_2, w_3, w_4) obtained for each of types of protein complexes are [0.96, 0.002, 0.6, -6.2] for type A, [0.96, 0.95, 0.6, -6.2] for type AB, [1.6, 0.12, 0.15, -1.6] for type E, and [2.0, 0.12, 0.15, -1.6] for type O.

Among the four weights contributing to the interactions of the protein complexes, w_4 is negative to maintain the positive f scores. The free energy change $\Delta G_{\text{interface}}$ represents the interactions of the protein partners at the interfaces. It is considered as a complementary potential terms to either (1) the $\text{Score}_{\text{IFACE}}$, which favors the shape complementarity of the two molecules, or (2) the complex flexibility due to the trade-off of the rigid docking. The more negative free energy change $\Delta G_{\text{interface}}$ favors the more stable interactions ($\Delta G < 0$) and also reflects the difficulty level that drives the two molecules to interact with each other, e.g. the value $w_4 = -6.2$ for type A/AB as compared to $w_4 = -1.6$ for type E/O.

In fact, different sets of weight values implicitly reflect different characteristics between the *Antigen/Antibody* and *Enzyme/Inhibitor* complexes. In addition to the noticeable difference of w_4 between the two types, the weight of the first term w_1 also indicates the susceptibility to shape

complementarity of type E/O ($w_1 = 1.6$ or 2.0) over the other type A/AB ($w_1 = 0.96$). The Enzyme and Inhibitor interactions are characterized as geometrical matching or shape complement whereas the Antigen and Antibody interactions are known as “induced fit” and require conformational changes during the associations. These properties however cannot be totally interpreted using rigid docking although some “soft” docking techniques are applied (i.e. adjusted surface layers for receptor and ligand when their grids are prepared, some slight overlaps are allowed, added electrostatics terms in the scoring function, etc. (Jackson, Gabb *et al.* 1998, Fernandez-Recio, Totrov *et al.* 2001, Chen, Li *et al.* 2003)).

Contribution of interfacial water (w_2, w_3) shows clearer differences between the types A and AB than between the other two types E and O, of which the shape complementarity is more predominant. Especially for the Lennard-Jones repulsive energy (w_2), the weight of 0.002 is much insignificant in type A as compared to 0.95 of that of type AB. Structural observations reveal that steric constraints of Antibody conformations between the unbound (A) and bound (AB) might lead to the difference of the weights. Interactions of the Antigen with the bound-conformation (AB) may have formed more spacious interfaces where there is room for the presence of water. In fact for the AB structures, interactions of the Antibody and its substrate may re-orient the overall conformational packing and therefore expose more hydrophilic regions on its surface; consequently forming hydrophilic cavities when interacting with the Antigen. Increase of the hydrophilicity in those cavities may facilitate occupation of one or more water molecules, substantially increasing effect of water onto the interfacial free energy of these AB complexes.

5.3.1.2. Interdependencies among the components

The scoring function (Equation 21) targets the ranks of the correct predictions –assumed as the near-native structures. Individual contribution of each component therefore involves in the overall rank improvement. In this part, Spearman rank correlation (R_{ij}) is used (Table 5-2) to evaluate the interdependency among these components: shape complementarity ($\text{Score}_{\text{IFACE}}$), protein-water dispersion ($E_{\text{LJ-repulsive}}$), water-mediated interaction ($E_{\text{wat-mediatedHbond}}$), and protein-protein interaction at the interface ($\Delta G_{\text{interface}}$).

As expected, each component is independent, yet somewhat correlates with others as complements; for examples $\text{Score}_{\text{IFACE}}$ and $\Delta G_{\text{interface}}$. This correlation is more pronounced in the

case of Enzyme/Inhibitor ($R=0.22$) than in the case of Antigen/Antibody complexes ($R=0.12$) due to the more shape complementarity prone in the Enzyme/Inhibitor complexes.

Table 5-2 Interdependencies among components for Antigen/Antibody (*bottom-left triangle*) and Enzyme/Inhibitor (*up-right triangle*) complexes

R_{ij}	$\text{Score}_{\text{IFACE}}$	$E_{\text{LJ-repulsive}}$	$E_{\text{wat-mediatedHbond}}$	$\Delta G_{\text{interface}}$
$\text{Score}_{\text{IFACE}}$	-	0.06	-0.03	0.22
$E_{\text{LJ-repulsive}}$	0.02	-	-0.35	0.31
$E_{\text{Wat-mediated Hbond}}$	-0.006	-0.3	-	-0.12
$\Delta G_{\text{interface}}$	0.12	0.33	-0.12	-

Noticeable differences of the correlation between $\text{Score}_{\text{IFACE}}$ and $E_{\text{wat-mediated Hbond}}$ again reflect different physicochemical properties between Enzyme/Inhibitor and Antigen/Antibody complexes: $R=-0.03$ vs $R=-0.006$, respectively. In the Enzyme/Inhibitor complex, shape complementarity is more predominant. In fact, the more “perfect” matching –e.g. as in “lock and key”– two protein partners result, the less chance water is present at the complex interfaces; consequently forming a reciprocal correlation between the two components. This correlation is however less significant (i.e. $R=-0.006$) in the Antigen/Antibody complexes perhaps also due to the less involvement of the shape complementarity.

In general, results of Spearman rank correlation highlight the interdependency among the potential terms of the scoring function (Equation 21): they are independent yet correlated to some extent to each others, subsequently eliciting their own significant contribution to the whole improvement of the overall ranks.

5.3.2. The proposed water-implemented method improves results of protein rigid docking

With the sets of weight values obtained above, ranking performances of each method, i.e. ZDOCK3.0.2 (using IFACE), ZRANK, and the water-implemented approach (briefly called IFACEwat) are demonstrated in further analysis.

5.3.2.1. Comparisons of the IFACEwat and ZDOCK3.0.2 and ZRANK

The proposed scoring function containing linearly combined energy terms is used to calculate the f score for each complex in the dataset (159 cases of various types) in order to perform a re-

ranking process for all the docked conformations generated by the docking program ZDOCK3.0.2.

Table 5-3 Overall results of ZDOCK3.0.2, ZRANK, and the IFACEwat in terms of numbers of near-native structures (*hits*) found and the rank of the first hit. Cases which are in **bold** indicate that IFACEwat method performs equivalently with or better than both of the others.

Complex	Cat*	ZDOCK		ZRANK		IFACEwat		
		#hits	rank of 1st hit	#hits	rank of 1st hit	#hits	rank of 1st hit	% (#hits/total hits)
Easy (113)								
1AHW	A	3	3	2	1	0	-	0
1BVK	A	5	1	0	-	1	2	10
1DQJ	A	6	1	1	2	7	1	41.18
1E6J	A	1	1	1	1	1	2	33.33
1JPS	A	2	1	2	1	1	1	25
1MLC	A	11	1	7	1	6	1	30
1VFB	A	2	1	1	4	2	4	25
1WEJ	A	6	1	2	1	6	1	31.58
2FD6	A	1	1	1	1	1	1	25
2I25	A	14	1	5	1	13	1	68.42
2VIS	A	2	1	1	2	1	3	33.33
1FSK	AB	0	-	0	-	0	-	-
1IQD	AB	0	-	1	1	1	1	100
1K4C	AB	1	1	3	1	3	1	42.86
1KXQ	AB	0	-	1	1	1	1	16.67
1NCA	AB	0	-	0	-	0	-	-
1NSN	AB	0	-	0	-	0	-	0
2JEL	AB	0	-	0	-	0	-	0
1AVX	E	4	1	4	1	4	1	36.36
1AY7	E	0	-	0	-	0	-	-
1BVN	E	10	1	9	1	9	1	28.13
1CGI	E	0	-	0	-	0	-	0
1CLV	E	140	1	119	1	141	1	33.98
1D6R	E	0	-	0	-	0	-	0
1DFJ	E	3	1	2	1	3	1	27.27
1E6E	E	1	1	3	1	1	1	11.11
1EAW	E	3	3	0	-	2	2	5.56
1EWY	E	0	-	0	-	0	-	0
1F34	E	0	-	0	-	0	-	-
1FLE	E	1	6	0	-	1	4	8.33
1GL1	E	0	-	0	-	0	-	0

Complex	Cat*	ZDOCK		ZRANK		IFACEwat		
		#hits	rank of 1st hit	#hits	rank of 1st hit	#hits	rank of 1st hit	% (#hits/total hits)
1GXD	E	3	1	3	1	3	1	50
1HIA	E	0	-	0	-	0	-	0
1JTG	E	3	1	4	1	3	1	37.50
1MAH	E	0	-	1	1	0	-	0
1N8O	E	0	-	0	-	0	-	-
1OC0	E	0	-	1	29	0	-	0
1OPH	E	4	1	2	2	3	1	37.50
1OYV	E	0	-	0	-	0	-	-
1PPE	E	2	25	0	-	2	30	4
1R0R	E	1	38	4	17	2	39	2.70
1TMQ	E	3	1	3	1	3	1	30
1UDI	E	0	-	0	-	0	-	-
1YVB	E	0	-	1	1	1	2	33.33
2ABZ	E	1	2	1	15	1	2	5.56
2B42	E	3	1	3	1	3	1	60
2J0T	E	0	-	0	-	0	-	-
2MTA	E	0	-	0	-	0	-	0
2O8V	E	0	-	0	-	0	-	-
2OUL	E	0	-	0	-	0	-	-
2PCC	E	0	-	1	1	0	-	0
2SIC	E	0	-	0	-	0	-	-
2SNI	E	0	-	0	-	0	-	0
2UUY	E	0	-	1	9	0	-	0
3SGQ	E	0	-	0	-	0	-	0
4CPA	E	3	51	4	2	3	4	5
7CEI	E	1	5	4	1	2	5	22.22
1AK4	O	0	-	2	1	1	3	33.33
1AZS	O	1	1	1	1	1	1	100
1B6C	O	0	-	0	-	0	-	0
1BUH	O	0	-	0	-	0	-	0
1E96	O	0	-	0	-	0	-	-
1EFN	O	0	-	0	-	0	-	0
1F51	O	0	-	0	-	0	-	0
1FC2	O	16	1	8	1	20	1	47.62
1FCC	O	1	27	0	-	0	-	0

Complex	Cat*	ZDOCK		ZRANK		IFACEwat		
		#hits	rank of 1st hit	#hits	rank of 1st hit	#hits	rank of 1st hit	% (#hits/total hits)
1FFW	O	8	2	3	1	7	2	35
1FQJ	O	2	1	2	1	4	1	50
1GCQ	O	9	1	4	1	7	1	30.43
1GHQ	O	0	-	0	-	0	-	-
1GLA	O	0	-	0	-	0	-	0
1GPW	O	1	1	1	1	1	1	33.33
1H9D	O	0	-	1	1	1	1	50
1HCF	O	0	-	1	1	1	1	50
1HE1	O	0	-	1	1	0	-	0
1J2J	O	0	-	0	-	0	-	0
1JWH	O	0	-	0	-	0	-	-
1KAC	O	4	1	5	1	4	1	40
1KLU	O	0	-	0	-	0	-	-
1KTZ	O	0	-	2	4	0	-	0
1KXP	O	2	1	1	1	2	1	66.67
1ML0	O	0	-	0	-	0	-	0
1OFU	O	1	1	1	1	1	1	33.33
1PVH	O	1	1	0	-	1	1	50
1QA9	O	1	2	1	1	1	1	12.50
1RLB	O	0	-	0	-	0	-	0
1RV6	O	0	-	1	1	1	3	33.33
1S1Q	O	0	-	0	-	0	-	-
1SBB	O	0	-	0	-	0	-	-
1T6B	O	3	1	3	1	5	1	55.56
1US7	O	2	1	2	1	2	1	66.67
1WDW	O	0	-	0	-	0	-	-
1XD3	O	0	-	0	-	0	-	0
1XU1	O	4	85	10	11	3	41	3.45
1Z0K	O	0	-	1	1	1	1	33.33
1Z5Y	O	2	1	2	1	3	1	37.50
1ZHH	O	1	1	1	1	1	1	50
1ZHI	O	3	1	3	2	3	1	37.50
2A5T	O	2	1	1	1	2	1	50
2A9K	O	0	-	2	1	1	2	33.33
2AJF	O	0	-	1	2	0	-	0

Complex	Cat*	ZDOCK		ZRANK		IFACEwat		
		#hits	rank of 1st hit	#hits	rank of 1st hit	#hits	rank of 1st hit	% (#hits/total hits)
2NZ8	O	3	1	1	1	2	1	40
2OZA	O	1	1	1	1	2	1	100
2Z0E	O	6	1	3	1	8	1	72.73
3CPH	O	4	1	5	1	3	1	37.50
<i>Difficult (20)</i>								
1E4K	A	1	1	1	1	1	1	50
1F6M	E	0	-	0	-	0	-	0
1FQ1	E	5	1	1	2	5	1	55.56
1PXV	E	4	1	6	1	4	1	57.14
1ZLI	E	0	-	0	-	0	-	0
2O3B	E	1	7	0	-	1	5	12.50
1ATN	O	3	1	2	1	3	1	75
1BKD	O	6	1	3	1	8	1	88.89
1FAK	O	4	1	1	1	4	1	80
1H1V	O	2	1	1	4	2	1	50
1IBR	O	6	1	2	1	6	1	66.67
1IRA	O	3	1	1	1	3	1	42.86
1JK9	O	2	1	2	1	3	1	50
1JZD	O	4	1	2	1	4	1	44.44
1R8S	O	7	1	2	1	6	1	60
1Y64	O	1	1	1	1	1	1	100
2C0L	O	6	1	3	1	6	1	66.67
2I9B	O	3	1	3	1	3	1	50
2IDO	O	10	1	7	1	9	1	64.29
2OT3	O	7	1	5	1	7	1	70

*A = Antigen/Antibody complex, AB = Antigen/bound-Antibody complex, E = Enzyme/Inhibitor complex, O = others
 - means no hit was found

Results of initial dockings by ZDOCK3.0.2 show that each of 139 cases (93 easy, 26 medium, 20 difficult or 18 Antigen/Antibody, 40 Enzyme/Inhibitor, 81 Others) out of 159 cases of protein complexes obtains at least one near-native structure called a “hit” (RMS≤2.5Å as compared to the according crystalized structure) among 2000 predictions.

The *success rate* reflects how well the method can rank the correct predictions in the top ranks, and therefore it shows how effective the method can discriminate the correct structures from the

false positives. It is important to achieve the higher ranks for the correct predictions because those top ranked predictions are then selected for further refinement using more intensive computing, e.g. MD simulation.

The *success rate* is defined as:

$$\text{Success rate} = \frac{\# \text{ cases that obtain at least 1 hit in the top 100 predictions}}{\# \text{ total cases of the particular complex type}}$$

The example in which ZDOCK obtains a success rate of 72.2 % (~13/18) for the complex type A/AB in Figure 5-5 means that in the top 100 predictions generated by ZDOCK3.0.2 there is at least one hit found among 13 out of 18 cases of type A/AB.

Comparisons of success rate between ZDOCK3.0.2 and the proposed method in this project indicate that the IFACEwat does improve the ranking of the initial rigid docking of ZDOCK3.0.2 for most of the complexes classified either by physicochemical properties (A/AB, E, and O) or by CAPRI difficulty level (easy, medium, and difficult). Except for the difficult cases in which both the methods achieve a tight percentage of 90% of cases obtaining at least one hit in the top 100 predictions, the IFACEwat shows better ranking outcomes, and therefore can discriminate better the correct from the false positive predictions.

In addition, it is demonstrated in Figure 5-5 (left panel) that the IFACEwat performs better than ZDOCK3.0.2 in cases of Antigen/Antibody complexes. While both the methods are equivalent or even the IFACEwat is slightly better than ZDOCK3.0.2 in ranking results of Enzyme/Inhibitor complexes (i.e. 60% vs 57%), the former obtains higher success rate in type A/AB (83.3% as compared to 72.2% of the latter). This improvement is the result of the inclusion of interfacial water effect and taking into account the flexibility of the interfaces, which are the two important characteristics of the Antigen/Antibody interactions—i.e. most of Antigen/Antibody complexes contain wet-interfaces and conformational changes are induced during their complex formation (Janin 1999, 2005).

For examples, two Antigen/Antibody cases, i.e. *1IQD* (easy) and *1BGX* (medium) show a clear improvement of the IFACEwat against the ZDOCK3.0.2 (Table 5-3). For both these two cases, while there is only one hit that could be generated from the initial docking, the IFACEwat method detect it and ranked it in the top 1, but ZDOCK3.0.2 could find none. Furthermore, in those difficult cases, while both the methods could find the same numbers of hits, the IFACEwat ranked the first

hit in better ranks in some cases (e.g. *2O2B* has 1 hit found in rank 7 by ZDOCK3.0.2, but rank 5 by the IFACEwat). Noticeably, the IFACEwat found 2 hits more than the other for the difficult case of *1BKD* (Table 5-3). Compared to ZRANK, the IFACEwat performs equivalently well for both the Antigen/Antibody and Enzyme/Inhibitor complexes, yet slightly better for medium and difficult cases.

*Only 139/159 cases could generate hits according to the ZDOCK

	Complex type (total cases*)					
	A/AB (18)	E (40)	O (81)	Easy (93)	Medium (26)	Difficult (20)
ZDOCK	13	23	60	55	23	18
ZRANK	15	23	67	66	22	17
IFACEwat	15	24	66	63	24	18

Figure 5-5 Performances of ZDOCK3.0.2, ZRANK, and IFACEwat method in terms of success rate (*top*) and of number of cases (*bottom*) for different complex types.

The IFACEwat could detect more numbers of near-native structures than the other two methods for most of protein complexes in the dataset (Figure 5-6). Especially for the type AB (complexes of Antigen and the bound-Antibody), while ZDOCK3.0.2 was able to locate only one hit of 1 complex (i.e. *1K4C* shown in Table 5-3 above), the IFACEwat (and also ZRANK) found 5 hits for 3 out of 7 cases and ranked the first hit as rank 1 for all these 3 cases.

Figure 5-6 Performances of ZDOCK3.0.2, ZRANK, and IFACEwat method in terms of numbers of hits found (against total numbers of hits shown in brackets) for each type of complexes.

Comparisons between results of ZDOCK3.0.2 and ZRANK, it is observed that ZDOCK3.0.2 can obtain more hits than the other, but ZRANK performs better in locating the hits in top ranks, thus filtering the wrong predictions more effectively than ZDOCK3.0.2, especially for Antigen/Antibody complexes (type A/AB). Further, observations of result comparisons among the three methods reveals that ZRANK obtains less number of hits than ZDOCK3.0.2, while performing equivalently well with the IFACEwat proposed in this project, indicating that ZRANK makes a significant improvement in re-ranking approach. Established on the same platform which is a linear combination of various potentials, the basic difference between the scoring functions of ZRANK and the IFACEwat is the way to account for the electrostatics and desolvation terms. In ZRANK, the two terms are weighted separately whereas in the scoring function of the IFACEwat they are included in the $Score_{IFACE}$ term and complemented by the other energies (derived from the added interfacial water). Therefore, implementing the electrostatics separately could be a future step to improve further the re-ranking ability for the proposed scoring function.

To demonstrate further the performance comparisons among the three methods, best RMS conformations are used (Figure 5-7 and 5-8). In the top 10, 100, and 1000 predictions, the conformations with the lowest RMS to the corresponding crystalized complex structure are noted. It is indicated that, for either types of complex classification (easy, medium, difficult or Antigen/Antibody, Enzyme/Inhibitor, Others), the IFACEwat outperforms ZDOCK3.0.2 in locating

the best RMS conformations (left panels in both the Figures). For the top 10 ranks, the two methods are equivalent although the IFACEwat already shows better performance, particularly for type AB (Figure 5-7-*left*). When the number of predictions increases to 1000, the trend is clearer for all the types, especially for medium and difficult cases.

As comparing with ZRANK, the trend is also more pronounced when the number of predictions increases, especially for the medium and difficult cases (Figure 5-7-*right*). However, both the methods perform equivalently well in locating the best RMS conformations for Antigen/Antibody cases, yet sometimes ZRANK is found with slightly better performance.

Figure 5-7 Best RMS of conformations in the top 10, 100, and 1000 ranks found by the IFACEwat against ZDOCK (left) and ZRANK (right). Each dot represents a protein complex case according to difficulty level (easy, medium, and difficult)

Figure 5-8 Best RMS of conformations in the top 10, 100, and 1000 ranks found by the IFACEwat against ZDOCK (left) and ZRANK (right). Each dot represents a protein complex case according complex types: Antigen/Antibody (A, AB), Enzyme/Inhibitor (E), Others (O).

5.3.2.2. Comparisons of the IFACEwat and F²Dock

While the IFACEwat is being developed, there is another re-ranking method published in March 2013 and included in the F²Dock package. This method applies the Generalized Born model to estimate the changes in solvation energy of the protein interfaces, and also considers separately the characteristics of protein complexes. Therefore, a comparison between the two methods is performed.

There is different implementation in F²Dock in ways of defining hits (or the near-native structures) and re-ranking the hits (Chowdhury, Rasheed *et al.* 2013) as compared to the IFACEwat. For examples, while F²Dock defines the hits based on RMS of the interface-involved residues (with interface RMS $\leq 5\text{\AA}$), hits in the IFACEwat are defined with the RMS of the whole ligand orientation against ligand position in the crystalized structures (and with RMS $\leq 2.5\text{\AA}$). In addition, in the IFACEwat, the success rate is estimated based on numbers of hits, which are considered only among the numbers of near-native structures in the ground truth (Table 5-4). The purpose of this strict constraint is to guarantee that the first hit needs to acquire the rank as high as possible.

For examples, for the case of *1BGX* (Antigen/Antibody and medium classified), there is only 1 hit generated according to the ground truth (so *number_of_hits* = 1). The constraint is set to the case that only the conformation ranked at the top-1 (and with RMS $\leq 2.5\text{\AA}$) by the IFACEwat is considered as a hit. And this hit is the first and the only hit according to the method. Other conformations out of this window length of ranking (e.g. here *window_length* = 1) will not be considered as hits found by the IFACEwat although their conformations obtain RMS $\leq 2.5\text{\AA}$. On the other hand, F²Dock considers any conformations whose interface-RMS $\leq 5\text{\AA}$ as hits found. As a result, as shown in the Table 5-5, some first hits of F²Dock obtain low ranks, e.g. for the case *1E6J* (Antigen/Antibody and easy), and the first hit is ranked at rank 126 by F²Dock, but ranked at rank 2 by the IFACEwat. For the case *1BGX* above F²Dock does not obtain any hit. Therefore, to maintain a rational assessment, results between each of the methods and ZDOCK3.0.2 separately are used as the intermediate comparison since all the 3 methods are applied for the same dataset of protein complexes (the benchmark 4.0) with 15° rotational sampling.

Table 5-4: Different implementation between the IFACEwat and F²Dock in determining the near-native structures (hit).

IFACEwat	Input: number of hits in ground truth, <i>numGT</i> list of conformations with rank and <i>RMS</i> , <i>C(r,RMS)</i> Output: hit found 1. Sort conformations according to rank in ascending order 2. for each conformation $c_i(r_i, RMS_i) \in C$ do 3. if $r_i \in [1, numGT]$ and $RMS_i \leq 2.5$ then 4. c_i is a hit 5. else: 6. c_i is not a hit 7. end if 8. end for
F ² Dock	Input: list of conformations with rank and <i>interface RMS</i> , <i>C(r,iRMS)</i> Output: hit found 1. Sort conformations according to rank in ascending order 2. for each conformation $c_j(r_j, iRMS_j) \in C$ do 3. if $iRMS_j \leq 5$ then 4. c_j is a hit 5. else: 6. c_j is not a hit 7. end if 8. end for

*The differences are in blue. In the IFACEwat algorithm, the number of hits in the ground truth (*numGT*) was defined as the number of complex conformations generated by the initial rigid docking ZDOCK3.0.2 that obtained the ligand orientation $RMS \leq 2.5\text{\AA}$ as compared to the original ligand orientation in the crystal structure. In both algorithms, the ranks were obtained according to their potential scores (i.e. the IFACEwat uses *f* score; F²Dock uses solvation energy change ΔE_{sol}).

Table 5-5: Some re-rankings results that are equivalent with or better than the F²Dock (full table can be found in the Appendix A6). The results of F²Dock are performed by Chowdhury *et al.*

Comp	Cat	Rank of the first Hit				RMS of the first Hit			
		ZDOCK [‡]	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
Easy									
1BVK	A	1	2	184	205	0.94	1.61	3.6	4.9
1DQJ	A	1	1	374	74	0.23	1.56	4.0	4.9
1E6J	A	1	2	3	126	0.42	0.42	4.1	5
1JPS	A	1	1	1266	186	0.3	0.3	2.1	2.7
1MLC	A	1	1	57	11	0.48	0.48	2	3.8
1WEJ	A	1	1	9	5	0.5	0.5	1.5	3.2
2FD6	A	1	1	3	62	0.55	0.55	5.0	4.4
2I25	A	1	1	2	122	0.6	0.6	3.0	3.9

[‡] Comparisons with ZDOCK were separately performed by the IFACEwat and by Chowdhury *et al.* with the same docking parameters—i.e. 15° rotational sampling—but with different constraints to define the hits. Therefore, ZDOCK results of both the experiments are shown in the Table 5-5 and level of improvement of each method—IFACEwat or F²Dock—over the ZDOCK will be used for comparison purposes.

Comp	Cat	Rank of the first Hit				RMS of the first Hit			
		ZDOCK [‡]	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
2VIS	A	1	3	-	-	0.68	0.68	-	-
1IQD	AB	-	1	18	4	-	0.35	4.3	3.0
1K4C	AB	1	1	583	105	0.49	0.49	2.9	4.4
1KXQ	AB	-	1	2	1	-	0.09	1.2	1.6
1NSN	AB	-	-	1267	-	-	-	1.6	-
1AVX	E	1	1	25	1	0.31	0.31	3.5	4.5
1BVN	E	1	1	3	1	0.21	0.21	1.2	3.2
1CLV	E	1	1	3	1	0.35	0.35	2.3	2.5
1DFJ	E	1	1	1	9	0.47	0.47	4.1	4.4
1E6E	E	1	1	5	20	0.11	0.11	3.2	4.7
1GXD	E	1	1	1173	-	0.8	0.8	4.9	-
1JTG	E	1	1	1	7	0.19	0.19	2.6	4.6
1OC0	E	-	-	1590	-	-	-	4.8	-
1OPH	E	1	1	1694	-	0.49	1.81	3.9	-
1TMQ	E	1	1	16	1	0.33	0.33	3.6	4.8
1YVB	E	-	2	1	-	-	1.86	2.4	-
2ABZ	E	2	2	-	5	2.25	2.25	-	2.8
2B42	E	1	1	3	1	0.24	0.24	4.2	3.9
1AK4	O	-	3	1090	964	-	0.36	3.5	4.3
1AZS	O	1	1	42	-	1.99	1.99	2.9	-
1EFN	O	-	-	-	-	-	-	-	-
1F51	O	-	-	589	-	-	-	4.6	-
1FC2	O	1	1	-	1190	0.44	0.44	-	5.0
1FCC	O	27	-	-	-	8.33	-	-	-
1FFW	O	2	2	73	325	2.56	0.35	4.5	4.7
1FQJ	O	1	1	-	-	0.31	0.31	-	-
1GCQ	O	1	1	1105	-	0.42	0.42	1.4	-
1GHQ	O	-	-	-	-	-	-	-	-
1GLA	O	-	-	1708	-	-	-	3.9	-
1GPW	O	1	1	3	1	0.48	0.48	3.6	3.7
1H9D	O	-	1	1006	-	-	0.63	4.5	-
1HCF	O	-	1	175	1225	-	0.29	4.0	4.7
1JWH	O	-	-	7	-	-	-	3.6	-
1KAC	O	1	1	592	8	0.32	0.32	4.5	4.4
1KLU	O	-	-	1957	-	-	-	3.4	-

Comp	Cat	Rank of the first Hit				RMS of the first Hit			
		ZDOCK [‡]	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
1KXP	O	1	1	1	7	0.57	0.57	1.6	3.5
1OFU	O	1	1	84	-	0.45	0.45	4.5	-
1PVH	O	1	1	748	-	2.08	2.08	4.5	-
1QA9	O	2	1	-	-	0.13	0.13	-	-
1S1Q	O	-	-	756	-	-	-	1.9	-
1SBB	O	-	-	-	-	-	-	-	-
1T6B	O	1	1	58	525	0.15	0.15	3.6	4.0
1US7	O	1	1	74	-	0.48	0.48	1.1	-
1XU1	O	85	41	912	-	2.01	2.01	5.0	-
1Z0K	O	-	1	8	307	-	0.22	3.3	3.3
1Z5Y	O	1	1	20	-	0.31	0.31	3.4	-
1ZHH	O	1	1	-	-	0.43	0.43	-	-
1ZHI	O	1	1	65	202	0.23	0.23	4.4	4.0
2A5T	O	1	1	-	268	0.73	0.73	-	3.6
2A9K	O	-	2	-	558	-	0.41	-	3.4
2AJF	O	-	-	475	-	-	-	3.6	-
2AYO	O	32	10	37	1108	2.06	2.31	3.3	2.0
2B4J	O	1	1	-	-	0.54	0.54	-	-
2BTF	O	6	1	53	95	0.45	0.45	4.7	4.5
2HLE	O	2	1	31	4	2.16	0.44	4.1	3.8
2HQS	O	1	1	-	27	0.55	0.55	-	4.1
2OOB	O	-	-	-	-	-	-	-	-
2OOR	O	-	1	766	16	-	0.21	4.4	4.0
2VDB	O	57	18	5	-	2.06	2.36	1.2	-
3BP8	O	11	3	-	474	2.69	2.46	-	5.0
<i>Medium</i>									
1BGX	A	-	1	-	-	-	0.17	-	-
1JIW	E	5	5	-	-	5.21	2.15	-	-
1KKL	E	-	-	-	-	-	-	-	-
1M10	E	1	1	-	-	0.78	0.78	-	-
1NW9	E	1	1	-	321	0.76	0.76	-	4.9
1GRN	O	1	1	1299	401	0.52	0.52	4.3	4.8
1HE8	O	1	1	-	-	0.31	0.31	-	-
1I2M	O	1	1	267	545	0.56	0.56	2.2	2.6
1IB1	O	1	1	-	-	0.51	0.51	-	-

Comp	Cat	Rank of the first Hit				RMS of the first Hit			
		ZDOCK [‡]	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
1K5D	O	1	1	-	521	0.35	0.35	-	4.3
1LFD	O	1	1	85	990	0.24	0.24	4.6	4.6
1MQ8	O	1	1	1455	-	0.27	0.27	3.2	-
1R6Q	O	1	1	-	180	0.28	0.28	-	3.7
1SYX	O	1	1	211	2	0.56	0.56	4.8	4.7
1WQ1	O	1	1	81	-	0.29	0.29	4.0	-
1XQS	O	1	1	19	61	0.44	0.44	3.8	4.2
1ZM4	O	1	2	6	-	0.33	0.33	4.1	-
2CFH	O	1	1	1	119	0.59	0.59	3.8	2.6
2H7V	O	1	1	1112	-	0.44	0.44	4.6	-
2HRK	O	1	1	3	-	0.39	0.39	3.7	-
2J7P	O	1	1	-	-	0.53	0.53	-	-
2NZ8	O	1	1	64	-	0.35	0.35	4.5	-
2OZA	O	1	1	-	-	0.53	0.53	-	-
2Z0E	O	1	1	-	169	0.61	0.61	-	3.9
3CPH	O	1	1	-	250	0.44	0.44	-	4.3
<i>Difficult</i>									
1E4K	A	1	1	-	-	0.27	0.27	-	-
1F6M	E	-	-	-	-	-	-	-	-
1FQ1	E	1	1	-	-	0.68	0.68	-	-
1PXV	E	1	1	-	-	0.12	0.12	-	-
1ZLI	E	-	-	-	-	-	-	-	-
2O3B	E	7	5	-	-	2.47	2.47	-	-
1ATN	O	1	1	-	1307	0.26	0.26	-	2.7
1BKD	O	1	1	-	-	0.45	0.45	-	-
1FAK	O	1	1	-	-	0.17	0.17	-	-
1H1V	O	1	1	-	-	0.59	0.59	-	-
1IBR	O	1	1	-	-	0.32	0.32	-	-
1IRA	O	1	1	-	-	0.65	0.65	-	-
1JK9	O	1	1	510	422	0.15	0.15	4.2	2.5
1JZD	O	1	1	44	144	0.53	0.53	4.6	4
1R8S	O	1	1	-	-	0.47	0.47	-	-
1Y64	O	1	1	-	-	1.08	1.08	-	-
2C0L	O	1	1	-	-	0.46	0.46	-	-
2I9B	O	1	1	-	-	0.65	0.65	-	-

Comp	Cat	Rank of the first Hit				RMS of the first Hit			
		ZDOCK [‡]	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
2IDO	O	1	1	130	156	0.38	0.38	3.6	4.5
2OT3	O	1	1	121	-	0.6	0.6	4.6	-

Further, to obtain a fair comparison, the relatively significant improvement of the success rate (as defined above) achieved by each of the methods is evaluated against the initial docking program ZDOCK (Figure 5-9).

Figure 5-9 Improvement of the IFACEwat and F2Dock against the initial docking ZDOCK. The success rate is defined the same as in Figure 5-5. It is observed that the IFACEwat performs better for Antigen/Antibody (A/AB) complexes while the F2Dock is better for Enzyme/Inhibitor (E).

It is observed in Figure 5-9 that different implementations of the role of water in the scoring function might result in different performances between the two techniques. In particular, the IFACEwat performs better for some Antigen/Antibody cases (e.g. improves 11.1% over ZDOCK versus 4% by F²Dock) and also for some medium and difficult cases. Both the methods account for the water effect, but while the F²Dock approximates the changes of solvation energy only (Chowdhury, Rasheed *et al.* 2013), the IFACEwat includes the free energy change of the complex interfaces by applying solvated rotamers, subsequently accounting for the side chain and backbone flexibility of the proteins; especially for some of difficult cases that required induced conformational fit during the protein associations.

The F²Dock on the other hand performs better than the IFACEwat in some Enzyme/Inhibitor cases (13.5% versus 2.5% in Figure 5-9). This might be the result of considering the energy to form the cavities that are frequently observed in the Enzyme and Inhibitor interactions.

5.4. Conclusion for this chapter

This chapter proposes a new re-ranking technique using a new energy-based scoring function, namely IFACEwat – a combined Interface Atomic Contact Energy (IFACE) and water effect. The IFACEwat aims to further improve the discrimination of the near-native structures of the initial rigid docking algorithm ZDOCK3.0.2. Unlike other re-ranking techniques, the IFACEwat explicitly implements interfacial water into the protein interfaces to account for the water-mediated contacts during the protein interactions.

The overall results show that the IFACEwat increases both the numbers of the near-native structures and improves their ranks as compared to the initial rigid docking ZDOCK3.0.2. In fact, the IFACEwat achieves a success rate of 83.8% for Antigen/Antibody complexes, which is 10% better than ZDOCK3.0.2. As compared to another re-ranking technique ZRANK, the IFACEwat obtains success rates of 92.3% (8% better) and 90% (5% better) respectively for medium and difficult cases. When comparing with the latest published re-ranking method F²Dock, the IFACEwat performs equivalently well or even better for several Antigen/Antibody complexes.

Chapter 6

Evaluation of the proposed interfacial water-implemented scoring function

This chapter discussed further on the ability of discriminating the correct complex predictions of the initial docking results by the proposed water-implemented scoring function in this project. The IFACEwat has improved the results not only in increasing number of near-native structures (or hits) found but also in locating the first hit on the top ranks. The improvement is even found for protein complexes that do not contain water in their original structures. For those structures containing water molecules, the implementation is able to reflect the water-mediated hydrogen networks that facilitate the bindings at the protein interfaces. Further, the IFACEwat is also able to maintain the sufficient computation time during the re-ranking, and shows significant improvement of computational efficiency for the MD simulation when the predicted conformations undergo the structural refinement.

6.1. The IFACEwat improved the first Hit rank for either with- or without water-containing structures

The solvent effect is explicitly presented in the complex interfaces during the re-ranking using the scoring function IFACEwat. Complement with the interface Atomic Contact Energy (IFACE) of the initial docking ZDOCK, the water contribution via the energies of protein-water dispersion and water-mediated Hydrogen bonds has fulfilled the overall picture of water-facilitating interactions between the protein partners, particularly in the Antigen/Antibody complexes.

Especially, the IFACEwat is strictly constrained in defining the *hit* to locate the near-native structures of protein complexes in as high ranks as possible (as describe above in section 5.3.2.2). Together with the inclusion of the interfacial water, the IFACEwat has improved not only the number of hits found but also the ranks of the first hit, which represents the best complex predicted and the nearest to the corresponding original protein complex (as shown in Table 5-3). Among the protein cases that are improved in ranking the first hit as compared to the initial docking ZDOCK, structural observations of their complex interfaces reveal the presence of the hydrogen bond network, which is mediated by the interfacial water (Figure 6-1). For the Antigen/Antibody case of *1KXQ*, while the initial docking cannot locate any hit, the IFACEwat found the best hit (with RMS = 0.09Å) and ranks it as the top-1. The interface of this protein complex contains 6 water molecules, two of which contribute to the hydrogen bond network that bridges the protein interactions between the interfacial residues. Similarly, in the cases of *1FLE* (Enzyme/Inhibitor) and *2BTF* (Other type), it is observed that the interfacial water molecules also form the mediated hydrogen bonds that connect the protein chains.

Further, results of the re-ranking using the IFACEwat show that the improvement applies for some of the protein complexes that do not contain interfacial water in the original crystal structures (Table 6-1). In fact, the re-ranking effectiveness of the proposed scoring function is still maintained for such cases when taking into account the free energy change of the protein complex interfaces (i.e. $\Delta G_{\text{interface}}$).

Figure 6-1 Presence of the water-mediated Hydrogen bonds at the interfaces of protein complexes, the first hit of which is improved in ranking. The complex interfaces (in grey shaded surfaces) contain interfacial water molecules (red spheres) that mediate the hydrogen bonds (black dotted lines) bridging the interactions between the receptor (green) and the ligand (cyan).

Table 6-1 The IFACEwat still improves both the number of hits and the rank of the first hit for either protein complex cases of with- or without interfacial water.

Crystal structure contains WAT	Complex	Cat	↑ number of hits		↑ rank of the first hit		$\Delta G_{\text{interface}}(1^{\text{st}} \text{ hit})$
			ZDOCK	IFACEwat	ZDOCK	IFACEwat	
NO	1FC2	O	16	20	-	-	-
	4CPA	E	-	-	51	4	-26.14
	1QA9	O	-	-	2	1	-76.04
	2AYO	O	-	-	32	10	-47.54
	2HLE	O	-	-	2	1	-178.17
	2VDB	O	-	-	57	18	-58.42
	3BP8	O	-	-	11	3	-32.08
YES	1DQJ	A	6	7	-	-	-
	1BGX	A	0	1	-	1	-348.73
	1K4C	AB	1	3	-	-	-
	1KXQ	AB	0	1	-	1	-191.97
	1CLV	E	140	141	-	-	-
	2O3B	E	-	-	7	5	-47.95
	1YVB	E	0	1	-	2	-55.84
	1FLE	E	-	-	6	4	-47.85
	1AK4	O	0	1	-	3	-44.51
	1FQJ	O	2	4	-	-	-
	1H9D	O	0	1	-	1	-131.85
	1HCF	O	0	1	-	1	-175.53
	1RV6	O	0	1	-	3	-110.72
	1T6B	O	3	5	-	-	-
	1Z0K	O	0	1	-	1	-133.32
	1Z5Y	O	2	3	-	-	-
	2A9K	O	0	1	-	2	-161.63
	2B4J	O	6	8	-	-	-
	1MQ8	O	4	5	-	-	-
	1R6Q	O	3	4	-	-	-
	2CFH	O	9	10	-	-	-
	2OZA	O	1	2	-	-	-
	2ZOE	O	6	8	-	-	-
	2BTF	O	-	-	6	1	-142.31
	2OOR	O	0	3	-	1	-141.03

“-” represents “not applicable” for the corresponding cases. For examples, in the first case [PDB:1FC2], it is found that IFACEwat improves the number of hits (i.e. 4 hits more than ZDOCK), but does not improve the rank of the first hit (e.g. the first hit found by both IFACEwat and ZDOCK is ranked at top-1). Also noted in this table that $\Delta G_{\text{interface}}$ value is only shown for the first hit of the case, in which the IFACEwat improves the rank of the first hit.

In addition to highlighting the contributions of water in the interfaces, the scoring function of the IFACEwat (Equation 21) includes the free energy change as the last potential term to account for the interactions occurring at the complex interfaces. Observed in some complex structures that do not contain water but still achieve the ranking improvement (Table 6-1), it is indicated that the contribution of the free energy change term help improve significantly the rank of the first hit.

Among the components that shape the free energy change term (in Equation 23), the energies of sidechain- and backbone-involved hydrogen bonds contribute noticeable portions (with the weights equaling to 1) in the overall value of the $\Delta G_{\text{interface}}$ and help determine the influence of side chain flexibility in the interfaces. Except for the Enzyme/Inhibitor complex (4CPA) in Table 6-1, the interactions occurring at the interfaces of the other non-water protein complexes are induced by significant conformational changes (Wang, Smolyar *et al.* 1999, Hu, Li *et al.* 2005, Chrencik, Brooun *et al.* 2006, Lejon, Cramer *et al.* 2008, Nam, Jung *et al.* 2008). Therefore, it is expected that taking into account these energy contributions would help distinguish the good-binding complexes from the others. For examples, in the two cases of 1QA9 ($\Delta G_{\text{interface}} = -76.04$ kcal/mol) and 2HLE ($\Delta G_{\text{interface}} = -178.17$ kcal/mol), which obtain the first hit ranking at top-1, the breakdown of the hydrogen bond-involving energies of -2.805 kcal/mol and -5.712 kcal/mol, respectively indicate the assistance of the hydrogen bonds in the protein interactions. However, those energies are ignored by the initial docking ZDOCK.

Similarly, for some water-containing complex cases, the first hit of which is ranked in the top-1 by the IFACEwat, the free energies of the complex interfaces also obtain significantly low values (as the more negative the more favorable). Particularly, in 2 cases of Antigen/Antibody complexes, i.e. 1BGX ($\Delta G_{\text{interface}} = -348.73$ kcal/mol). and 1KXQ ($\Delta G_{\text{interface}} = -191.97$ kcal/mol), the IFACEwat locates the first hit at rank 1 while ZDOCK could not find any hit. Therefore, it suggests that including the consideration of the flexibility of protein side chains during the protein associations improve the discrimination of good binding complexes from the others, thereby improving the rank of the first hit or the best near-native structure of the protein complexes.

6.2. Is the IFACEwat over-fitting to the data?

The sets of weights of the IFACEwat scoring function are optimized over the whole dataset using Monte Carlo experiment. Therefore, 13 complexes—in which the ligand is not a monomer and

therefore are "unseen" to the algorithm—are used to test if the IFACEwat is over-fitting. The test data contains 8 easy, 2 medium, and 3 difficult cases.

Among these test cases, the IFACEwat can rank the first hit at top-1 rank for most of the complexes; except for *1I9R* (type AB–easy) and *2HMI* (type AB–difficult), in which the IFACEwat cannot locate any hits (Table 6-2). Therefore, it is suggested that the estimated weights of the IFACEwat are optimal and unlikely over-fits to the given data.

Table 6-2 The IFACEwat still maintains the ability of locating the first hit at top-1 for most of unseen data.

Complex	Cat	Number of hits found	Rank of the 1 st Hit	Complex	Cat	Number of hits found	Rank of the 1 st Hit
<i>Easy</i>				<i>Medium</i>			
1BJ1	AB	1	1	1IJK	E	1	1
1I9R	AB	0	-	1GP2	O	1	1
1QFW	AB	1	1	<i>Difficult</i>			
1EZU	E	4	1	2HMI	AB	0	-
2PCC	E	1	2	1DE4	O	1	1
1A2K	O	8	1	1EER	O	5	1
1AKJ	O	2	1				
1I4D	O	4	1				

“–” means no hit was obtained in corresponding cases.

6.3. Free energy change ($\Delta G_{\text{interface}}$) plays the most important role in the IFACEwat scoring function

The IFACEwat scoring function includes various potential terms of the complex-binding score and the protein interactions at the interfaces (i.e. protein-protein and protein-water interactions). The contributions of these terms are evaluated given a set of weights (w_1, w_2, w_3, w_4), each of which reflects an independent influence of each potential onto the overall f score (Equation 21).

As discussed in the section 5.2.2 and 6.1 above, the free energy change ($\Delta G_{\text{interface}}$) contributes the most among the four potential terms, especially in Antigen/Antibody complexes (i.e. $w_4 = 6.2$). The free energy change ($\Delta G_{\text{interface}}$) is used to evaluate the interactions occurring at the complex interfaces, therefore determining the more favorable bound complex conformations (the more

negative values).

To evaluate the contribution of each potential term, “weight knock-out” experiments are performed. In those experiments, each potential contribution i is alternately disabled, i.e. $w_i=0$ (Table 6-3), while the others are used to calculate the total f score and estimate the number of the found near-native structures (or hits). Result observations show that the contribution of the free energy change ($\Delta G_{\text{interface}}$) plays the most important role among the potential terms. When $w_4 = 0$, the protein-protein interactions that occur at the interfaces are mostly ignored and the side chain flexibility is limited (in experiment #4); consequently leading to the noticeable drop of the total number of hits found (Table 6-3), especially for the Antigen/Antibody complexes (i.e. type A and AB obtain only 17.1% and 37.5% number of hits, respectively as compared to the experiment #0). Differences of Antibody conformations between the two types (unbound in A and bound in AB) might explain the different decrease of the number of hits. For the Antigen/bound-Antibody structures (type AB), interactions of the Antibody and its possible substrate may have re-oriented the conformation of the Antibody itself, which substantially leads to conformational changes before interacting with the Antigen. The significant drop of the number of hits in type A implies that the free energy change $\Delta G_{\text{interface}}$ provides significant influence to determine the number of near-native structures for the Antigen/unbound-Antibody complexes.

Table 6-3 Decrease of the number of near-native structures for all types of protein complexes Antigen/unbound-Antibody (A), Antigen/bound-Antibody (AB), Enzyme/Inhibitor (E) and others (O): ↓ (less than 5% hits drop), ↓↓ (less than 30% hits drop), ↓↓↓ (noticeably drop less than 50% hits), ↓↓↓↓ (significantly drop less than 70% hits), ↓↓↓↓↓ (dramatically drop more than 95% hits), - (unchanged). Experiment #0 shows the results of number of hits obtained from the full set of weights for each complex type and is used as the reference.

#	$w_i = 0$	Number of hits found (in %)* and level of hit decrease			
		A	AB	E	O
0	(w_1, w_2, w_3, w_4)	100	100	100	100
1	$(0, w_2, w_3, w_4)$	↓↓↓↓ (41.5)	↓↓ (87.5)	↓↓ (73.8)	↓↓ (74.4)
2	$(w_1, 0, w_3, w_4)$	-	↓↓ (75.0)	↓ (97.1)	↓ (98.2)
3	$(w_1, w_2, 0, w_4)$	-	-	↓ (97.0)	↓ (98.0)
4	$(w_1, w_2, w_3, 0)$	↓↓↓↓ (17.1)	↓↓↓ (37.5)	↓↓ (78.0)	↓↓ (90.0)
5	$(w_1, 0, 0, w_4)$	-	↓↓ (87.5)	↓ (97.0)	↓ (97.7)
6	$(0, w_2, w_3, 0)$	↓↓↓↓↓ (4.87)	↓↓↓ (50.0)	↓↓↓↓ (30.0)	↓↓↓↓ (18.7)
7	$(0, 0, 0, w_4)$	↓↓↓↓ (43.9)	↓↓ (75.0)	↓↓ (71.9)	↓↓ (75.8)
8	$(w_1, 0, 0, 0)$	↓↓↓↓ (17.0)	↓↓↓↓↓ (0.0)	↓↓ (74.2)	↓↓ (89.0)

* % of hits as compared to the reference (i.e. results of the experiment #0)

Further, results of both the experiments (#4 and #6) with $w_4=0$ show the significant decrease of the number of hits of the types A and AB. Results of the Enzyme/Inhibitor complexes also indicate the similar decreasing trend, especially when both the first potential term (e.g. shape complementarity) and the free energy change ($\Delta G_{interface}$) are disabled in experiment #6, the results of which significantly become worse. In the experiment #6, number of hits obtained for each complex type are 4.87% (A), ~50% (AB), ~30% (E), and 18.7% (O) as compared the results of experiment #0. Similarly, results of the experiments #8 reinforce the important role of the free energy change ($\Delta G_{interface}$) in the IFACEwat scoring function, in particularly for the Antigen/Antibody complexes. Decrease of the number of hits therefore implies increases of the false positives when the unfavorable complex conformations are not be detected.

It is observed in Table 6-3 that restriction of the combinations of shape complementarity, electrostatics and desolvation also contributes importantly to the decrease of the number of hits for the type A complexes. When $w_1 = 0$ in the experiment #1, except for type A (41.5%), most of the complexes obtain only around >70% (nearly 30% drops) of the number of near-native structures as compared to the experiment #0 results. Comparisons of the two experiments #1 (with $w_1 = 0$) and #4 (with $w_4 = 0$) again reveal the different influences of both shape complementarity and free energy change on protein complexes of different physicochemical properties. Although the shape complementarity surprisingly affects much on the complexes type A (i.e. more than 50% drop) as shown in experiment #1, ignoring the contribution of the free energy change ($\Delta G_{interface}$) results in decreasing more hits (i.e. more than 82% drop shown in experiment #4). Similar decreasing trend is observed in type AB in experiments #1 and #4 as well as in the experiments #7 and #8. On the other hand, the shape complementarity shows more predominant role in determining the number of hits in the Enzyme/Inhibitor (type E) complexes.

It is also observed in the Antigen/Antibody complexes that the effect of water molecules is clearer on the complexes of type AB than on the complexes of type A. As discussed in the section 5.3.1, interactions of the Antigen with the bound-conformation of the Antibody may result in more spacious interfaces where there can be occupied by the water molecules, and therefore increasing the effect of water onto the protein interactions at the complex interfaces of type AB. Hence, when $w_2 = 0$ (experiment #2), especially if there exists water molecules at the complex interfaces, this may cause the decreases of number of hits in type AB, but not in type A.

6.4. The IFACEwat maintains sufficiently the computational efficiency yet generate improved ranking of the first hit

The IFACEwat scoring function is developed as a re-ranking technique using explicit interfacial water to further discriminate the correct predictions from the other incorrect and false positive ones of an initial docking algorithm, particularly in improving the results of predicting complex formations for Antibody/Antigen complexes.

Figure 6-2 Advantages of the IFACEwat re-ranking over the initial docking ZDOCK

The nature of the rigid initial docking is to keep the receptor rigid and allow limited flexibility of the ligand in order to achieve computational efficiency due to the large sizes of both the molecules. However, at the same time, it limits the ability of finding the correct bindings of the protein complexes as well as increases the probability of generating the false positive ones. This leads to a need of structural refinement to optimize the conformations of the protein complexes and further eliminate the decoys. Nonetheless, such the process, e.g. using energy-based evaluations by MD simulation, for the refinement purposes requires intensive computations, especially for the Antibody/Antigen complexes. Therefore, for this work the IFACEwat is expected as the prerequisite to reduce as many as possible numbers of false positives that do not need to be refined before heading to the refinement steps (Figure 6-2).

Table 6-4 Real running time T_R of the IFACEwat in re-ranking some of the protein complexes and the estimated formula of the MD time for the structural refinement stage of the IFACEwat results as compared to ZDOCK

Comp	Cat	IFACEwat re-ranking time T_R (second/~minute)	Expected MD refinement time to obtain at least 1 near-native structure using *			
			ZDOCK		IFACEwat	
			Rank of 1 st Hit	MD time	Rank of 1st Hit	MD time
4CPA	E	489.9 / 8.2	51	$51 * T_{Di}$	4	$4 * T_{Di}$
1QA9	O	542.9 / 9.1	2	$2 * T_{Di}$	1	T_{Di}
2AYO	O	1109.8 / 18.5	32	$32 * T_{Di}$	10	$10 * T_{Di}$
2HLE	O	819.0 / 13.6	2	$2 * T_{Di}$	1	T_{Di}
2VDB	O	864.0 / 14.4	57	$57 * T_{Di}$	18	$18 * T_{Di}$
3BP8	O	500.8 / 8.3	11	$11 * T_{Di}$	3	$3 * T_{Di}$
1BGX	A	1215.4 / 20.2	-	$2000 * T_{Di}$	1	T_{Di}
1KXQ	AB	1289.5 / 21.5	-	$2000 * T_{Di}$	1	T_{Di}
2O3B	E	632.4 / 10.5	7	$7 * T_{Di}$	5	$5 * T_{Di}$
1YVB	E	855.5 / 14.2	-	$2000 * T_{Di}$	2	$2 * T_{Di}$
1FLE	E	564.0 / 9.4	6	$6 * T_{Di}$	4	$4 * T_{Di}$
1AK4	O	290.4 / 4.8	-	$2000 * T_{Di}$	3	$3 * T_{Di}$
1H9D	O	954.9 / 15.9	-	$2000 * T_{Di}$	1	T_{Di}
1HCF	O	738.7 / 12.3	-	$2000 * T_{Di}$	1	T_{Di}
1RV6	O	848.2 / 14.1	-	$2000 * T_{Di}$	3	$3 * T_{Di}$
1Z0K	O	859.6 / 14.3	-	$2000 * T_{Di}$	1	T_{Di}
2A9K	O	678.9 / 11.3	-	$2000 * T_{Di}$	2	$2 * T_{Di}$
2BTF	O	1130.1 / 18.8	6	$6 * T_{Di}$	1	T_{Di}
2OOR	O	1297.8 / 21.6	-	$2000 * T_{Di}$	1	T_{Di}

*assume that each conformation of each complex takes equivalent time during the MD simulation, i.e. T_{Di}

Results of re-ranking time (T_R) by the IFACEwat (Table 6-4) indicate that the IFACEwat is able to maintain sufficient timing (less than approximately 22 minutes) to re-rank 2000 conformations of each protein complex. The two Antigen/Antibody complexes (i.e. *1BGX* and *1KXQ*) are among the complexes, which take the longest re-ranking time and of which the near-native could not be located by the initial docking ZDOCK.

Let T_{Di} be the MD simulation time for each conformation that would undergo the structural refinement and assume that each of them would take an equivalent time. In the cases of *1BGX* and *1KXQ*, there would be $2000 * T_{Di}$ computational time needed for the MD simulation to find the near-native complex structure. However, after re-ranked by the IFACEwat, there would need only

$1 * T_{Di}$ simulation time to obtain at least one near-native structure of protein complex because this conformation is re-ranked among the top (in fact rank-1). Similarly for other complexes, which obtain the higher rank for the first hit, the MD simulation timing is significantly improved if the structures undergo the refinement stage.

Besides that, in terms of computational complexity, the IFACEwat maintains the computational efficiency of the initial docking algorithm ZDOCK3.0.2. Considering the individual computational contribution of each component, except for the $\text{Score}_{\text{IFACE}}$ that is estimated based on all atoms of the whole protein complexes, the other potential terms are only involved partially in the complex interfaces.

$$f = w_1 \underbrace{\text{Score}_{\text{IFACE}}}_{O(N^3 \log_2 N^3)} + w_2 \underbrace{E_{\text{LJ_repulsive}} + w_3 E_{\text{water-mediated-Hbond}}}_{O(cN), c \lll 1} + w_4 \underbrace{\Delta G_{\text{interface}}}_{O(kN^2), k \lll 1}$$

The “add-ons” complexity of the scoring function f —by the last 3 components—is derived from the number of water molecules c at the interfaces and of pair-wise interactions k of interfacial residues as shown in the schematic above. Therefore, they contribute insignificantly into the overall complexity, resulting in equivalent complexity as a whole when compared to the complexity of the ZDOCK3.0.2 –i.e. $O(N^3 \log_2 N^3)$.

Figure 6-3 Receiver Operating Characteristic (ROC) curves to illustrate improvement of the IFACEwat over ZDOCK3.0.2 in terms of ranking the first hit as top-1

To benchmark more accurately the ability of discriminating the high rank on the top, ROC curves quantifying the results between ZDOCK and the IFACEwat is performed (Figure 6-3). It is observed that both the methods can obtain locate correctly the hits—with ratio of true positive/false positive $\sim 50/0.3$ —yet ZDOCK cannot maintain this trend for the whole dataset: the ratio decreases and results in its AUC (area under curve) around 0.81. On the other hand, the IFACEwat still maintains the better performance (AUC=0.96).

It is noted in this analysis that the ROC is evaluated based on the ranking improvement of the top-1 hit, in which the IFACEwat is constrained to prioritize the first hit location. Therefore, it suggests—based on the ROC curve—that the IFACEwat significantly improves the ability of discriminating the near native structure from the other incorrect or false positive ones.

Chapter 7

Conclusion and future work

7.1. Summary and Conclusion

Limitations of the ranking step of several current initial docking algorithms render them impossible to obtain the best performances, results of which are mostly affected by large number of the false positives, which are wrong but highly ranked predictions. Therefore, it is essentially required for developing an effective ranking protocol to successfully discriminate the correct predictions from the wrong or false positive ones.

Predicting the protein associations of Antigen/Antibody complexes is the most challenging among the protein complex types. As the interfaces of this complex type are either dry or wet (Janin 1999) and the wet interfaces with water-filled cavities maintain the close pack of the structures and facilitate the protein interactions via water-mediated hydrogen bonds or indirect interactions (Ikura, Urakubo *et al.* 2004). It leads therefore to a more challenging issue in discriminating the correct binding predictions from the incorrect ones, especially using the initial rigid docking, because the interactions involve several conformational changes or the water-mediated interactions during the complex associations.

In this research, an energy-based scoring function is proposed for the re-ranking step to improve the initial rigid docking results in locating as more as possible the near-native structures and bringing them as high ranks as possible, particularly for the Antigen/Antibody complexes.

In the implementation, interfacial water is added into protein interfaces and its effect is accounted explicitly in the proposed energy-based scoring function.

$$f = w_1 \text{Score}_{\text{IFACE}} + w_2 E_{\text{LJ_repulsive}} + w_3 E_{\text{water-mediated-Hbond}} + w_4 \Delta G_{\text{interface}}$$

In the presence of water at the interfaces, some extent of flexibility upon protein associations is evaluated accordingly via the approximations of the free energy change of the complex interfaces, and the water effect is presented via the water-mediated Hydrogen bond and protein-water dispersion energies.

Observations of the physicochemical-dependent potential energy terms reveal a possibility of different contributions of each potential term in the function f according to different types of protein complexes, i.e. Antigen/Antibody or Enzyme/Inhibitor. Therefore, the four weights w_1 , w_2 , w_3 , and w_4 are evaluated separately for each complex type. For the Antigen/Antibody type, the set of weights is [0.96, x , 0.6, -6.2] with $x=0.002$ or $x=0.95$ for the unbound and bound Antibody, respectively. For the Enzyme/Inhibitor type, the set of weights is [1.6, 0.12, 0.15, -1.6].

The changes in the Antibody conformations may affect the overall orientations of the interacting partners, consequently affects the cavity formation and may change the hydrophilicity of these cavities. Therefore it results in a significant difference of the weights w_2 between the unbound and bound Antibodies. In addition, the differences of the weight values also reflect the shape complementarity predominance of the complexes in which this characteristic is more favored, i.e. in Enzyme/Inhibitor type.

Using those sets of evaluated weights to calculate the f scores for each predictions generated by the initial rigid docking and thereby re-ranking them, it is demonstrated that the re-ranking improves the docking results and the correct predictions are highly ranked. Most of the near-native structures found for each protein complexes are ranked at the top-1. In fact, the proposed re-ranking technique (or named IFACEwat) achieves a success rate of 83.8% for Antigen/Antibody complexes, which is 10% better than ZDOCK3.0.2. As compared to another re-ranking technique ZRANK, the IFACEwat obtains success rates of 92.3% (8% better) and 90% (5% better) respectively for medium and difficult cases. These cases are actually characterized as conformational change required during the protein associations. For the number of near-native structures (#hits) found, the IFACEwat obtains a slightly more number of hits than other two

methods. When comparing with the current re-ranking method in F²Dock, which employs a Generalized Born (GB) model in its energy-based function, it is shown that the IFACEwat performs equivalently with the GB-rerank F²Dock. In fact, it performs better in Antigen/Antibody type, but on the other hand, the GB-rerank F²Dock achieves better results for the Enzyme/Inhibitor type.

In general, although the IFACEwat achieves equivalently or better than some other current methods used in the comparisons, it is still required to improve further its performance to obtain better success rate (which is currently 83.3% for Antigen/Antibody complex type, 60% for Enzyme/Inhibitor type, or even slightly more than 90% for medium and difficult cases).

Furthermore, the scoring function IFACEwat is able to reflect the water-mediated hydrogen networks that facilitate the bindings at the protein interfaces. In terms of computational efficiency, the IFACEwat obtains sufficient computation time during the re-ranking and maintains equivalently the computational complexity of the initial docking algorithm; subsequently highlighting the potential of computational efficiency for the MD simulation if the predicted conformations need to undergo the structural refinement.

7.2. Future Work

While separately evaluating the weight values of each potential term in the proposed scoring function f , it is observed that the contribution of the fourth term, i.e. the free energy change of the complex interfaces ($\Delta G_{interface}$), significantly increases the number of hits for each protein complexes (for example, the number of hits increases from 12 to 46 for the Antigen/Antibody complexes). This indicates a high possibility of increasing more the number of hits if the contribution of interface free energy change is further exploited.

Therefore, a future investigation of the change of the free energy change (e.g. $\Delta\Delta G_{complex}$) during a time-course simulation can be conducted after the complex prediction. The term $\Delta\Delta G_{complex}$ can be used as a ranking filter to further re-rank the results. As it is expected that the quality of interactions could be reflected by the difference of the free energy change between different states, as e.g. the more negative the more favorable, the correct predictions will be again discriminated from the other incorrect ones. Further, another improvement can be conducted on the first potential term of the proposed function. Currently, the $Score_{IFACE}$ term is derived from the

scoring function ($SC+ELEC+DE_{IFACE}$) employed by the initial docking ZDOCK3.0.2 to represent mostly the shape complementarity characteristics of the protein complexes. It is also observed that the weight values are separately evaluated, therefore the terms are independent. Hence, it is possible that the first term ($Score_{IFACE}$) could be replaced by any other scores derived from any other scoring functions of any initial docking algorithms in order to achieve the robustness for the method.

Thus, with these further implementations, the proposed method is expected to achieve better results and more improvement in the discrimination of the near-native structures from the false positive ones.

Author's publications

1. **Chinh Tran-To Su**, Thuy-Diem Nguyen, Jie Zheng, Chee Keong Kwoh, "IFACEwat: the interfacial water-implemented re-ranking algorithm to improve the discrimination of near-native structures for protein rigid docking", *BMC Bioinformatics*, 15(Suppl 16):S9 (2014).
2. **Chinh Tran-To Su**, Xuchang Ouyang, Jie Zheng, Chee Keong Kwoh, "Structural analysis of the novel influenza A (H7N9) viral Neuraminidase interactions with current approved neuraminidase inhibitors Oseltamivir, Zanamivir, and Peramivir in the presence of mutation R289K", *BMC Bioinformatics*, 14(Suppl 16):S7 (2013).
3. **Chinh T.T Su**, Christian Schonbach, Chee Keong Kwoh, "Molecular docking analysis of 2009-H1N1 and 2004-H5N1 influenza virus HLA-B*4405-restricted HA epitope candidates: Implications for TCR cross recognition and vaccine development", *BMC Bioinformatics*, 14(Suppl 2):521 (2013).
4. Xuchang Ouyang, Shuo Zhuo, **Chinh Tran To Su**, Zemei Ge, Runtao Li, Chee Keong Kwoh, "CovalentDock: Automated covalent docking with parameterized covalent linkage energy estimation and molecular geometry", *J. Comput Chem* (2012).
5. Stephanus Daniel Handoko, Xuchang Ouyang, **Chinh Tran To Su**, Chee Keong Kwoh, Yew Soon Ong, "QuickVina: Accelerating AutodockVina using gradient-based heuristics for global optimization", *IEEE/ACM Trans. Comput. Biology. Bioinform.* 9(5):1266-1272 (2012).
6. Qian Liu, Steven C.H. Hoi, **Chinh T.T Su**, Zhenhua Li, Chee Keong Kwoh, Limsoon Wong, Jinyan Li, "Structural analysis of the hot spots in the binding between H1N1 HA and the 2D1 antibody: do the mutations of the H1N1 from 1918 to 2009 affect much on this binding?", *Bioinformatics*, 27(18):2529-2536 (2011).
7. **Chinh T.T Su**, Stephanus D. Handoko, Christian Schonbach, Xiaoli Li, Chee Keong Kwoh. "A Possible Mutation that Enables H1N1 Influenza A Virus to Escape Antibody Recognition", *IEEE International Conference on Bioinformatics & Biomedicine, Hongkong* (2010).
8. Stephanus D. Handoko, **Chinh T.T. Su**, Chee Keong Kwoh, and Yew Soon Ong, "Structural Analysis of (TCR-)HLA/peptide Complexes: An Initial Study," in *Proceedings of 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pp.63-66, Dec 2010.

Appendix

A1. A possible mutation that enables the H1N1 virus to escape antibody recognition

A1.1. Abstract

The H1N1 influenza A 2009 pandemic caused a global concern as it has killed more than 18,000 people worldwide so far. Studies that have found cross-neutralizing antibodies between the 1918 and 2009 pandemic flu elicit a basis of pre-existing immunity against the 2009 H1N1 virus in old population. The cross-reactivity occurs due to conserved antigenic epitopes shared between the two pandemic viruses. However, evolutionary mutation can enable the virus to elude human immunity system, making these antibodies probably no longer effective. In our study, we found that a possible mutation in B-cell epitope (the sequence PNHDSNKG) could be the chance for the virus to escape the 1918 antibody recognition. Hence, this finding can be helpful for further vaccine designs against the H1N1 2009 influenza A virus.

A1.2. Introduction

The outbreak of H1N1 influenza A 2009 was caused by a re-assorted strain from swine, avian, and seasonal human flu viruses (Neumann, Noda *et al.* 2009, Smith, Vijaykrishna *et al.* 2009). Various studies have been conducted, especially on Hemagglutinin (HA) protein which is essential for virus to attach and infect cells (Neumann, Noda *et al.* 2009), making it a key target for developing neutralizing antibodies and vaccines to prevent the infection. The studies also revealed that the 2009 HA contains conserved epitopes that existed in the 1918 Spanish flu HA protein (Ekiert, Bhabha *et al.* 2009, Wei, Boyington *et al.* 2010, Xu, Ekiert *et al.* 2010). The conservation has explained cross-reactivity of antibodies between the two pandemic viruses and prior immune memory against the current pandemic strain of the older population (Greenbaum, Kotturi *et al.* 2009, Wei, Boyington *et al.* 2010). However, no research could make sure that evolutionary mutation would not render the antibodies ineffective and enable the virus to escape human immunity system. In this paper we examined and analyzed variability in HA sequences of current and past strains of pandemic influenza virus. A potential mutation in B-cell epitope was

found to show possible escape mechanism of the H1N1 influenza A virus from the current antibody recognition.

A1.3. Method

A1.3.1. Phylogenetic Analysis of Pandemic (H1N1) 2009 Strains

The HA protein sequences of the pandemic (H1N1) 2009 strains were retrieved from (NCBI Influenza sequence database)(NCBI Influenza sequence database)(NCBI Influenza sequence database)(NCBI Influenza sequence database)(NCBI Influenza sequence database)(NCBI Influenza sequence database). MAFFT version 6 (Katoch 2008) was used to do multiple sequence alignment and construct the Neighbour Joining tree (bootstrap value = 1,000). The tree was visualized using Archaeopteryx tree viewer (Han and Zmasek 2009). The potential epitopes were predicted using the IEDB Analysis Resource Tools, and the B-cell candidate was found using linear epitope prediction tools such as Chou & Fasman Beta-Turn Prediction (Chou and Fasman 1978), Emini Surface Accessibility Prediction (Emini, Hughes *et al.* 1985), Karplus & Schulz Flexibility Prediction (Karplus and Schulz 1985), Kolaskar & Tongaonkar Antigenicity (Kolaskar and Tongaonkara 1990), Parker Hydrophilicity Prediction (Parker, Guo *et al.* 1986), and Bepipred Linear Epitope Prediction (Larsen, Lund *et al.* 2006). The candidate with score higher than the default threshold and predicted by most of the methods was chosen for further investigation. Homology modeling was then done with I-TASSER (Zhang 2007, 2008) on the selected B-cell epitope.

A1.3.2. Molecular docking for the predicted epitopes

The binding ability of the mutated epitope and antibodies was demonstrated using ClusPro (Comeau, Gatchell *et al.* 2004, Comeau, Gatchell *et al.* 2004), a docking server of Boston University. ClusPro v.2 is a multistage docking program. It performs rigid docking using ZDOCK programs, filters and clusters docked conformations, and refines the results using CHARMM minimization (Comeau, Kozakov *et al.* 2007). The dockings were performed using all default parameters and antibody-binding mode was applied. The lowest docked energy model was selected for detection with an assumption that the minimum binding energy-obtained model is assumed to be the best docked conformation and nearest to the native binding state.

A1.4. Results and Discussion

A1.4.1. Phylogenetic relationship of current pandemic influenza

Typically, HA subtypes of the pandemic influenza strains differ from that of circulating seasonal strains. To demonstrate that, multiple sequence alignment and phylogenetic analysis of HA H1 clade subtype 1 to strains of subtype 2 and 3 were done and it confirmed that the clade of the current pandemic virus strains is distinct from strains of seasonal influenza strains. The current pandemic strains also differ from WHO recommended strains that had been used for vaccine development and group 2 HA subtype 14 (Fig. A1-1). Group 2 but not Group 1 HA subtypes are recognized by tert-butyl hydroquinone (TBHQ), a small molecule that inhibits virus entry (Russell, Kerry *et al.* 2008). Considering that the binding sites of monoclonal antibody CR6261 is close to the TBHQ binding site, it is unlikely that passive immunization would protect against the current pandemic strains. Molecular docking further confirmed that the CR6261 did not recognize the presence of the epitope in the current pandemic HA proteins.

Figure A1-1 Neighbor-joining tree of H1, H2, H3, and H14 subtype HA sequences with 1,000 bootstrap replicates. The current strains (black) are distinct to seasonal influenza strains used in WHO vaccines (blue) and the past pandemic influenza (purple). HA of strains derived from swine (orange) and birds (green) are relatively close to current pandemic strain subtype H14 HA (red) which is known to bind to TBHQ.

A1.4.2. Investigation of interactions between known neutralizing antibodies and epitope-containing HA protein

Using Epitope prediction tools in Immune Epitope Database (IEDB), B-cell epitope PNHDSNKG was found (Table A1-1). This epitope sequence obtained highest prediction score and predicted by multiple tools indicating that it is likely to be recognized by the immune system.

Table A1-1 The predicted potential B-cell epitopes

Epitope	Methods	Sequence 1	Sequence 2
PNHDSNKG	Bepipred	TSSWPNHDSNKGVTAAACPHAGAKSF	
	Chou	TSSWPNH	PNHDSNK
	Karplus	PKTSSWP	NHDSNKG
	Parker	PNHDSNK	NHDSNKG

A1.4.2.1. Docking with the 2D1 antibody from the complex of Fab 2D1-1918 influenza virus HA

The binding ability of the chosen epitope was predicted using a docking server ClusPro. Only clusters of docked models with minimum binding energies were considered. The predicted epitope sequence was first modelled by I-TASSER (Zhang 2007, 2008) to obtain the 3D-structures. It was then mapped back to a HA protein using homology mapping (Beaver, Bourne *et al.* 2007). The epitope-containing HA protein was docked to a structure of a known antibody extracted from the Fab 2D1-1918 influenza virus HA complex [PDB:3LZF] (Xu, Ekiert *et al.* 2010). The 1918 influenza virus HA was chosen because the antigenic drift of the H1N1 influenza A virus involved in the 2009 pandemic started in the year 1918 (Garten, Davis *et al.* 2009, Wei, Boyington *et al.* 2010).

Figure A1-2 Different binding orientations of the 2D1 antibody to the epitope-containing HA protein (A) and to the 1918 influenza HA (B). The bound complexes of antibody and HA monomer are shown as surface presentation. HA1 and HA2 domains are colored in green and cyan. The 2D1 antibody chains are shown in yellow (heavy chain) and orange (light chain). The predicted B-cell epitope is colored in red.

The docked complex had different antibody-binding orientation from the Fab 2D1-1918 influenza virus HA complex (Figure A1-2). In the docked complex, the 2D1 antibody bound to HA1 domain, but it did not interact with the predicted B-cell epitope as shown in Figure A1-2A. The binding regions of the two HA protein sequences contain different residues between the positions 130-133

(as shown in Figure A1-3). In the 2D1antibody-1918 HA protein complex, the antibody interacts with residue Lys¹⁵⁷ of the HA protein. The residues Thr¹³² and Thr¹³³ interact with the upstream Glu¹³¹ residue, pulling the His¹³⁰ forward and creating a distance of 4.7Å between the His¹³⁰ and the Lys¹⁵⁷, therefore making the Lys¹⁵⁷ free to bind to the antibody.

In case of the epitope-containing HA protein, it was noticed that the Lys¹⁵⁷ made internal polar contacts of 1.8Å distance with the His¹³⁰ (Figure A1-3A). Epitope binding indirectly changes the side-chain orientation of His¹³⁰ via Ser¹³², thereby inhibits the interaction of Lys¹⁵⁷ with the antibody in our docking experiment.

Figure A1-3 Influence of His130 side-chain on antibody binding of residue Lys157 in 2D1-epitope-containing HA complex (A) and 2D1-1918 HA protein complex (B). Differences in binding regions of epitope-containing HA and the 1918 HA protein is shown above. Matching residues are marked by asterisks. Mismatches in binding regions (131-133) are colored in red. The B-cell epitope is underlined. In the below panel, atom-based residue coloring: C (white), O (red), and N (blue). Backbones are presented by sticks. Side-chains are shown as lines. Interactions are presented as yellow-dotted lines. A. Internal contacts of Ser132 and His130 reduce the distance of His130 to Lys157 to 1.8Å, thereby blocking binding of the antibody. B. Lys157 interacts with Asn54 of the antibody. The distance His130 is of 4.7Å.

A1.4.2.2. Docking with neutralizing antibody CR6261

Another docking of epitope-containing HA protein was performed with the neutralizing antibody (CR6261) of the 1918 H1N1 pandemic influenza. CR6261 differs from other antibodies because it does not interfere with the binding of virus to the host cell. The antibody inhibits the conformational rearrangement of the HA protein that is necessary for membrane fusion (Ekiert,

Bhabha *et al.* 2009). The heavy chain of the antibody made contacts with the helical turns of the HA2 domains of the epitope-containing HA (Figure 1-5) as it did in case of the 1918 H1N1 viral HA protein, confirmed by Ekiert's group (Ekiert, Bhabha *et al.* 2009).

Figure A1-4 Binding of CR6261 antibody heavy chain to helical turn of HA2 domain of the epitope-containing HA protein. The interaction is shown in ribbon (A) and surface presentation (B). HA1: green, HA2: cyan, heavy chain: yellow, light chain: orange, and B-cell epitope: red.

Generally, docking results of neutralizing antibodies to the epitope-containing HA protein showed that the HA protein was inactivated by the CR6261 antibody; however, the predicted B-cell epitope was not involved in this binding. This suggests a possibility that future H1N1 strains with these mutations might escape an adaptive immune response. If the currently effective epitopes accumulate this mutation, existing vaccines are likely to become ineffective.

A1.5. Concluding remark

Variability analysis of H1N1 2009 influenza A HA proteins highlighted the phylogenetic distinctness to other seasonal influenza and the WHO recommended vaccine strains. It is shown that HA subtype H1 is not considered as a potential inhibitory drug target as confirmed by Russell *et al.* (Russell, Kerry *et al.* 2008). A potential B-cell epitope was predicted using HA protein sequences of the current virus strains and it was assumed to have undergone an evolutionary mutation. In this study, it was shown that this B-cell epitope failed to be recognized by the antibodies. In fact, intermediate neighbour contacts of the Lys¹⁵⁷ residue make it unavailable for binding to the antibody. The docking results of the CR6261 antibody and the epitope-containing HA protein also showed that the epitope was not involved in inactivating the function of the HA protein. Hence, it could suggest that if the virus actually obtained this mutation, the antibodies could no longer recognize it. This finding could be useful for vaccine designs in later years.

A2. Implications for TCR cross-recognition and vaccine development for 2009-H1N1 and 2004-H5N1 influenza virus HLA-B*4405-restricted HA epitope candidates

A2.1. Abstract

Studies on pandemic influenza vaccines have shown that T cell recognition to conserved epitopes and cross-reactive T cell responses are important when new strains emerge, especially in the absence of antibody cross-reactivity. In this work, using HLA-B*4405 and DM1-TCR structure model, high confidence conserved 2009-H1N1 T cell epitope candidates were systematically generated and investigated on their potential cross-reactivity against H5N1 avian flu virus. The results are novel with regard to HA epitopes and useful for developing possible vaccination strategies against the rapidly changing influenza viruses. Yet, the challenge of identifying epitope candidates that result in heterologous T cell immunity under natural influenza infection conditions can only be overcome if more structural data on the TCR repertoire become available.

A2.2. Introduction

In 2009, the outbreak of a new swine-origin strain of influenza A H1N1 caused widespread human infection (Smith, Vijaykrishna *et al.* 2009). One of the most important surface proteins, hemagglutinin (HA) permits the virus to bind to cell membrane and infect the cells. Since mutations enable the virus to escape from either T cell or antibody recognition, current flu vaccines were not effective against the emerging virus. Sequence analyses showed that the HA sequence of the pandemic 2009-H1N1 underwent an antigenic shift (Gatherer 2009) that altered its antigenicity in context of the seasonal flu vaccine.

The antigenicity of HA 2009-H1N1 remained highly conserved to pandemic 1918-H1N1 and partially conserved to seasonal flu strains of the 1930s. Therefore, the majority of infected individuals who were vaccinated with the WHO recommended seasonal flu vaccine did not produce neutralizing antibodies against the new influenza strain. However, elderly and individuals born before 1950 were less affected than expected. The lower infection rate of these age groups has been interpreted as the results of cross-reactive T cells (Gras, Kedzierski *et al.* 2010) and antibody (Xu, Ekiert *et al.* 2010) responses to the pandemic 1918-H1N1 and partially cross-

reactive T cell response to seasonal flu strains of the 1930s. A study by Boon *et al.* (Boon, Mutsert *et al.* 2004) on CD8⁺ T cell recognition of heterosubtypic H1N1 variants indicated that repeated infection with heterologous viruses may increase cross-reactive Cytotoxic T Lymphocytes (CTL) and thus confer protection against newly emerging strains in absence of a cross-reactive antibody response. Further support for this concept comes from a study of subjects who were vaccinated against seasonal influenza and showed *in vitro* cross-reactive T cell responses against HA of the pandemic 2009-H1N1 (Subbramanian, Basha *et al.* 2010).

In growing recognition of the role of T cell responses to H1N1, several groups conducted large-scale Human leukocyte antigen (HLA) binding motif scanning analyses of pandemic and seasonal strains to predict and identify conserved peptides that elicit cross-reactive HLA class I and/or class II restricted T cell responses (Assarsson, Bui *et al.* 2008, Groot, Ardito *et al.* 2009, Gupta, Srivastava *et al.* 2010). While the affinity-based approach allows a broad coverage of HLA supertypes and epitope bindings (Assarsson, Bui *et al.* 2008, Groot, Ardito *et al.* 2009), structural approach gives better insight view onto T cell recognition of the HLA-restricted T cell epitopes (Chang, Smolyar *et al.* 1997, Wang and Reinherz 2001, Archbold, Macdonald *et al.* 2009). This study focuses on the immunogenicity that depends on the quality of T-cell receptor (TCR) interaction with the HLA/peptide complexes rather than HLA-binding peptide affinity only. Therefore, a combined affinity-based epitope prediction with molecular docking was performed to generate conserved high confidence HA T cell epitope candidates of current and past pandemic strains, and consequently to analyze the potential TCR cross-recognition of 2009-H1N1 and 2004-H5N1.

According to Archbold *et al.* complex of DM1-TCR and HLA-B*4405/peptide showed significant enhancement in T cell-mediated responses among micro-polymorphisms in the HLA-B*44 family, and as such they are key factors in controlling persistent viral infections (Archbold, Macdonald *et al.* 2009). Thus, to perform the experiments HLA-B*4405 and DM1-TCR were used as models. Results of structural models of TCR cross-recognition between 2009-H1N1 and 2004-H5N1 revealed steric and topological effects of TCR contact residue mutations on TCR binding affinity. While these results are novel with respect to HLA-B*4405-restricted H1N1 HA epitopes and DM1-TCR, yet with limited available structural data upon the TCR repertoire, more investigations and experimental analyses are still recommended for further broad perspective of their utility in vaccine development against the emerging virus strains.

A2.3. Methods

A2.3. 1. Obtaining HA candidate epitopes from pandemic (H1N1) 2009 sequences

Potential epitopes were predicted using T cell epitope prediction tool (NetCTL v1.2) from IEDB (Larsen, Lundergard *et al.* 2005, Larsen, Lundegaard *et al.* 2007). NetCTL, a neural network architecture-based tool, was used to predict T cell epitope candidates for HA proteins of current H1N1 influenza A strains according to HLA-B44 supertype. Weight parameters on C-terminal cleavage (0.15) and transporter associated with antigen processing (TAP) efficiency (0.05) were as default. There are 3 threshold scores (0.75, 0.9, 1.0) that give both high sensitivity (0.8, 0.74, 0.7) and specificity (0.97, 0.98, 0.985) accordingly. While the first 2 scores (0.75 and 0.9) obtained similar results of T cell epitopes, the score of 1.0 resulted only 7 epitopes (positions 240, 188, 128, 482, 400, 131, and 50) shown in Table A2-1. The docking results of HLA-B*4405 and the T cell epitope candidates showed that epitope candidates at positions 259, 251, and 229, which were located in the top 10 HLA-binding energy rankings in Table A2-1, would be missed if the 1.0 score was used. Finally, as comparing to the score of 0.9, the value 0.75 was selected as the threshold for its better sensitivity (0.8) to predict the potential T cell epitopes. Therefore candidates with prediction scores greater than 0.75 were chosen for further investigation. I-TASSER (Roy, Kucukural *et al.* 2010) was used to obtain homology models of the selected T cell epitopes.

A2.3. 2. Molecular docking of predicted HLA-B*4405-binding epitopes to T cell receptor DM1-TCR

The pipeline included docking the epitopes to HLA-B*4405 [PDB:3DX8] and followed by docking of the HLA/epitope complex as a ligand to DM1-TCR [PDB:3DXA]. The binding ability of the predicted epitopes was further analyzed using ClusPro v.2 (Kozakov, Brenke *et al.* 2006, Kozakov, Hall *et al.* 2010), a web-based automated docking program performing a multistage protocol: rigid PIPER docking, filtering and clustering of docked conformations, and stabilizing using Monte Carlo simulations (Kozakov, Hall *et al.* 2010). During the docking, HLA β -domain was masked but remained surface accessible since it is not involved in interaction with the TCR (Reinherz, Tan *et al.* 1999). The results were clustered according to their binding energies. The binding energy score (hereby called "binding energy") is generated from an energy function of

PIPER. This function is a sum of potential terms of shape complementarity, electrostatics, desolvation contributions, and Decoys as reference states (DARS) (Kozakov, Brenke *et al.* 2006). According to Kozakov *et al.*, the core idea of the knowledge-based potentials DARS counts on observed numbers of intermolecular interactions (Kozakov, Brenke *et al.* 2006), thus docking results were filtered by selecting docked complexes that belonged to the most populated clusters of interacting complexes but with lowest binding energy scores for the final results. It was assumed that the model with minimum binding energy was the optimal conformation.

Then, molecular dynamics (MD) simulations using AMBER 10 force field *ff99SB* was performed (Case, CheathamIII *et al.* 2005) to improve the bound conformation of rigid docking. The TCR-HLA/epitope docked complexes underwent a 3-stage MD simulation (minimization, heating, and equilibration) using explicit solvent model under periodic boundary condition. In the minimization stage, a weak positional restraint of a force constant of 500kcal/molÅ² was put on the whole complex for the first 1,000 steps to hold them fixed while initially minimizing positions of solvent and sodium ions. Yet in the next 2,500 steps of minimization, this restraint was removed. In this minimization stage, constant volume was set. In the heating stage of 20ps, the restraint was again put on the complex, but with only 10kcal/molÅ² to avoid wild fluctuations in the structure. The system was heat up from 0K to 300K and applied the Langevin temperature equilibration scheme to control the temperature. Then, a short equilibration stage (1ns) without the restraints was performed in constant pressure of 1 atm and at 300K. SHAKE was used in both heating and equilibration stages to constrain bonds that involves hydrogen. As a result, complexes obtained from the MD simulation above were considered final bound conformations of the docked TCR-HLA/epitope complexes in our study.

A2.4. Results and Discussion

A2.4. 1. Conserved HA T cell epitope candidates generated for HLA-B*4405 and DM1-TCR

After various HA sequences among the pandemic 2009-H1N1, 1918-H1N1, and WHO recommended vaccine strains (Supplement A2.6.1) were examined and aligned, potential 9-mer T cell epitopes were predicted by using NetCTL v1.2 from Immune Epitope Database (IEDB). Nineteen candidates, which scored greater than 0.75, were more likely well presented by HLA-B44 family (Supplement A2.6.2). These 19 epitopes were then docked to HLA-B*4405, and

subsequently the HLA-B*4405/epitope complex as ligand to DM1-TCR. The 19 predicted epitope candidates differ in their physico-chemical properties. Eight candidates starting at positions 128, 188, 251, 131, 229, 400, 240, and 259 contain mostly hydrophobic and polar residues, whereas other 11 candidates (positions 50, 412, 446, 463, 493, 514, 421, 475, 482, 81, and 274) are composed of predominantly hydrophilic residues (Supplement A2.6.3) which result in higher HLA-B*4405 binding energies (Table A2-1). As expected, all the 19 epitopes bound to the helical antigen-presenting groove of the HLA-B*4405 (Reinherz, Tan *et al.* 1999). Eleven epitopes (50, 128, 251, 412, 463, 493, 514, 131, 229, 400, and 259) bound in similar orientations to HLA-B*4405 with Glu at position 2 and residues at the C-terminus binding to the B and F pockets of the HLA-B*4405, respectively (DiBrino, Parker *et al.* 1995, Reinherz, Tan *et al.* 1999, Macdonald, Purcell *et al.* 2003). Residues which occupy the center of the binding groove are solvent-exposed and likely to bind to TCR (Reinherz, Tan *et al.* 1999, Wahl, McCoy *et al.* 2009). In contrast, most residues of epitopes at positions 188, 446, 421, 475, 482, and 240 interacted with the HLA-B*4405 peptide binding pockets. In this model, the deeply embedded residues were not exposed to the solvent. Therefore TCR recognition of HLA-B*4405/epitope complexes is expected to be conformationally restrained.

Table A2-1 Total binding energies of the 19 epitope candidates (in descending HLA-B*4405 binding energy ranks).

Position	Sequence	Docked to HLA-B*4405		Docked to DM1-TCR	
		Energy	Rank	Energy	Rank
240	QEGRMNYW	-1,006	1	-541.7	17
188	KEVLVLWGI	-921.8	2	-573.9	13
128	FERFEIFPK	-869.9	3	-594.6	12
482	FEFYHKCDN	-846.8	4	-743.9	2
259	FEATGNLVV	-815.2	5	-608.9	10
400	IEKMNTQFT	-775.8	6	-688.5	5
131	FEIFPKTSS	-752.9	7	-560.5	14
251	VEPGDKITF	-743.3	8	-599.1	11
229	PEIAIRPKV	-687.7	9	-548.9	15
50	LEDKHNGKL	-661.4	10	-552.1	16
274	MERNAGSGI	-644	11	-	-
514	REEIDGVKL	-643.8	12	-666.3	7
81	PECESLSTA	-633.3	13	-	-
446	LENERTLDY	-603.1	14	-621.1	9
412	KEFNHLEKR	-599.5	15	-737.2	3
493	MESVKNGTY	-543	16	-645	8

475	KEIGNGCFE	-519.6	17	-715.4	4
463	YEKVRSQLK	-475.3	18	-751.3	1
421	IENLNKKVD	-464.6	19	-666.7	6

DM1-TCR was docked to the HLA-B*4405/epitope complexes to assess their binding conformation. In the models with minimum binding energies, DM1-TCR V_α and V_β domains interacted with HLA-B*4405 α_1 and α_2 domain residues and solvent exposed residues of the epitopes. Epitope candidates 240 and 482 were embedded in the binding groove of HLA-B*4405, and interacted with DM1-TCR via Tyr⁷ and Tyr⁸ (epitope 240) and Lys⁶ (epitope 482) as shown in Table A2-2. These interactions appeared to be mediated by a change of the TCR binding position to HLA-B*4405 α_1 - and α_2 -domain residues, resulting in contacts with polar and charged residues Asn⁷⁰, Glu⁷⁶, and Gln¹⁵⁵ of the HLA-B*4405. Hydrogen bond lengths (in Å) and dihedrals around these H-bonds made by DM1-TCR and HLA-B*4405 domains in cases of exposed and embedded epitope types were calculated. In case of epitope 259, residues at positions 4, 5, and 6 were exposed while the TCR V_α domain interacted with Asn⁷⁰ (2.1Å) and Glu⁷⁶ (2.2Å). In contrast, for the embedded-type epitope 240, binding of TCR V_β domain to Asn⁷⁰ and Glu⁷⁶ at 2.5Å H-bond lengths was observed. Epitope candidates 400 (exposed-type) and 482 (embedded-type) also showed similar noticeable H-bond length alterations. Distances of the TCR to the residues Gln¹⁵⁵ (HLA- α_2) decreased when the contacts with the HLA domains shifted from the TCR V_β to TCR V_α . Meanwhile, the dihedrals of residues that made the H-bonds also behaved a decreasing trend when the TCR changed its binding positions to the HLA-B*4405 from V_β to V_α domain (e.g. from -133.7° to -54.1° in cases of epitope 240 and 259 respectively) (Supplement A2.6.4). It could suggest that the changes in H-bond lengths and dihedral angles indicate an attempt of the TCR to adjust its binding access to the HLA-B*4405/epitope complexes when the epitopes change from exposed to embedded type.

Table A2-2 2009-H1N1 HA epitope interacting residues with DM1-TCR and HLA-B*4405 complexes.

Position	Epitope sequence	HLA/epitope (solvent-exposed residues)	TCR-HLA/epitope (interacting residues)
50	LEDKHNGKL	Lys ⁴	Lys ⁴
128	FERFEIFPK	Arg ³ , Glu ⁵	Arg ³
188	KEVLVLWGI	-	Lys ¹ , Leu ⁴ , Leu ⁶ , Gly ⁸
251	VEPGDKITF	Asp ⁵ , Lys ⁶	Asp ⁵ , Lys ⁶

412	KEFNHLEKR	Asn ⁴ , His ⁵ , Glu ⁷	Glu ² , Asn ⁴ , His ⁵ , Glu ⁷
446	LENERTLDY	-	Leu ¹ , Glu ⁴ , Asp ⁸
463	YEKVRSQLK	Arg ⁵	Arg ⁵ , Lys ⁹
493	MESVKNGTY	Met ¹	Met ¹ , Lys ⁵
514	REEIDGVKL	Lys ⁸	Lys ⁸
131	FEIFPKTSS	Phe ¹	Phe ¹ , Lys ⁶
229	PEIAIRPKV	Ala ⁴ , Ile ⁵	Ile ⁵
400	IEKMNTQFT	Lys ³ , Met ⁴	Glu ² , Lys ³ , Met ⁴ , Asn ⁵ , Thr ⁶
421	IENLNKKVD	-	Lys ⁷ , Asp ⁹
475	KEIGNGCFE	-	Lys ¹ , Glu ²
482	FEFYHKCDN	-	Lys ⁶
240	QEGRMNYYW	-	Asn ⁶ , Tyr ⁷ , Tyr ⁸ , Trp ⁹
259	FEATGNLVV	Thr ⁴ , Gly ⁵ , Asn ⁶	Glu ² , Thr ⁴ , Gly ⁵

Epitope candidates at positions 81 and 274 in 1918 and WHO vaccine sequences (Supplement A2.6.2) yielded rather low HLA-B*4405 binding scores (less than threshold 0.75) and could not be docked to DM1-TCR. Therefore it is unlikely that the two candidate epitopes were antigenic to the 1918 and WHO strains. Mutations at positions 81 and/or 274 in HA sequences of the 2009 viral strains increased the HLA-B*4405 binding scores above the threshold, but did not facilitate TCR recognition in our model.

A2.4. 2. Recognition of computationally inferred optimal epitope KEKMNTEFW by HLA-B4405 and DM1-TCR

In this work, a designed the mosaic epitope candidate KEKMNTEFW was computationally designed from five epitope sequences, which were in the top 5 ranks of DM1-TCR binding energy (Table A2-1). These epitopes are at positions 463, 482, 412, 475, and 400, whose corresponding residue positions were favorably bound by the DM1-TCR model. In fact, the epitope IEKMNTQFT at position 400 was selected as the starting point because most of its residues made direct contacts with DM1-TCR, i.e. -EKMNT--- (Table 1-3). Also, Archbold *et al.* (Archbold, Macdonald *et al.* 2009) suggested that the second positioned residue Glu (E²) be required for preferential binding to HLA-B44. According to the top 5 DM1-TCR docking results of this work, 2 out of 5 epitopes (412 and 475) contain residue Lys (K) at the first position, and Lys^{epitope475} directly interacted with the DM1-TCR. Therefore, Lys (K¹) was used for the first position of the mosaic epitope. Similarly, Glu (E) was chosen for the 7th position since it was a TCR-interacting residue of

the epitope 412. Finally, bulky side-chains of F⁸ (Phe) and W⁹ (Trp) served as anchors for HLA-B*4405 binding. Thus, using Deep-View (Guex and Peitsch 1997), the residues I¹, Q⁷, and T⁹ were substituted by K, E, and W respectively.

Docking of the mosaic epitope KEKMNTEFW to HLA-B*4405 and DM1-TCR showed that it bound favorably to both HLA-B*4405 (binding energy -849.5 kcal/mol; rank 4) and DM1-TCR (-684.4 kcal/mol; rank 6) with Asn⁵ and Thr⁶ exposed to the solvent and directly interacting with DM1-TCR. Although the HA peptide appears to be a good candidate for inclusion in multivalent peptide vaccine against the H1N1 influenza A, its efficacy as a protective epitope on population level depends on the TCR repertoire which could be only tested experimentally.

A2.4. 3. Cross-recognition of 2009-H1N1 and 2004-H5N1 HA T cell epitopes

A study by Kreijtz *et al.* (Kreijtz, Mutsert *et al.* 2008) showed that T cell responses to seasonal H1N1 and H3N2 strains were largely cross-reactive to avian H5N1. According to WHO Global Influenza Program (*Availability of new H5N1 prototype strain for influenza pandemic vaccine development* 2006), H5 HA viral strain A/Vietnam/1194/2004 of the avian flu outbreak in Vietnam was one of the H5N1 prototype vaccine strains in 2005 and recommended candidate of pre-pandemic H5N1 vaccine. Therefore, HA protein of this strain was used as a model to test if the 2009-H1N1 T cell epitope candidates above would be cross-reactive for 2004-H5N1.

As in the same pipeline, NetCTL v1.2 was used to extract the 2004-H5N1 HLA-B*4405 restricted T cell epitopes candidates from IEDB. Of all candidates with predicted scores greater than 0.75, seven were highly similar (greater than 78% similarity) to the 2009-H1N1 T cell epitope candidates (Table A2-3) and likely to be recognized by DM1-TCR.

Docking of the seven HLA-B*4405/2004-H5N1 epitope complexes to DM1-TCR revealed that the DM1-TCR predominantly interacted with the 2004-H5N1 epitopes at conserved positions 50, 446, 493, and 475. DM1-TCR binding energies of the 2004-H5N1 epitope candidates were lower and thus more favorable than these of the corresponding 2009-H1N1 epitope candidates (Table A2-3). Most mutations that occurred between these two influenza outbreaks did not appear to affect DM1-TCR recognition of HLA-B*4405 presented epitope candidates.

Table A2-3 Seven 2009-H1N1 T cell epitope candidates that could be cross-reactivity with 2004-H5N1 T cell responses.

position	Strain	Sequence	NetCTL score	HLA-B*4405 docked	DM1-TCR
				Energy	Energy
50	2009-H1N1	LE <u>D</u> K HNGKL	1.0508	-661.4	-552.1
	2004-H5N1	LE <u>K</u> T HNGKL	1.1478	-715.0	-682.2
412	2009-H1N1	<u>K</u> EFN <u>H</u> LEKR	0.9764	-599.5	-737.2
	2004-H5N1	<u>R</u> EFN <u>N</u> LEKR	0.9949	-638.9	-640.8
446	2009-H1N1	<u>L</u> ENERTLD <u>Y</u>	1.1929	-603.1	-621.1
	2004-H5N1	<u>M</u> ENERTLD <u>F</u>	1.5813	-563.5	-667.2
493	2009-H1N1	M ESV <u>K</u> NGTY	1.73	-543.0	-645.0
	2004-H5N1	M ESV <u>R</u> NGTY	1.6645	-595.0	-667.5
421	2009-H1N1	IENLNK <u>K</u> V D	0.8563	-464.6	-666.0
	2004-H5N1	IENLNK <u>K</u> M E	0.7968	-466.1	-642.0
475	2009-H1N1	K E <u>I</u> GNGCFE	0.9892	-519.6	-715.4
	2004-H5N1	K E <u>L</u> GNGCFE	0.8258	-548.9	-733.8
482	2009-H1N1	FEFYH K CDN	1.1128	-846.8	-743.9
	2004-H5N1	FEFYH K CDN	1.1129	-846.8	-743.9

*Different residues are underlined. The residues that interacted with DM1-TCR are highlighted in bold. Two significant changes in TCR binding energy are marked in bold.

However, a significant decrease in TCR binding energy of the epitope candidate 2004-H5N1 complex at position 50 was observed (Table A2-3). The mutation K4T changed the structural conformation of the epitope's exposed region. The side chain of the mutated Thr⁴ pointed toward the HLA- α_2 domain and is embedded in the helical binding groove of the HLA-B*4405. It might therefore induce the interactions of Lys⁸ and Leu⁹ at its C-terminal with the DM1-TCR (Figure A2-1A). This could consequently suggest a more favorable binding of DM1-TCR to the HLA-B*4405/epitope50 in 2004-H5N1 than in 2009-H1N1 viral strain.

In addition, it was noticed that the binding affinity of DM1-TCR to the 2004-H5N1 epitope candidates considerably decreased when the TCR interactions occurred directly at mutated residues of the epitope candidates at position 412 (H5N) and 421 (V8M and D9E). Substitution of a positively charged His with a smaller-sized neutral Asn reduced the contact surface of the epitope candidate's exposed region (Figure A2-1B), resulting in a conformationally constrained contact of DM1-TCR with 2004-H5N1 candidate epitope 412 and a higher binding energy (-640.8 kcal/mol) compared to the 2009-H1N1 candidate epitope 412. In a future pandemic it is expected that apart from a few mutated epitopes, heterologous immunity (Selin and Welsh 2004) mediated by pre-existing cross-reactive T cell responses to seasonal influenza virus will limit its severity and extent.

Figure A2-1 Conformational changes of DM1-TCR/HLA-B*4405/epitope complexes caused by epitope mutations. (A) Mutation K4T in the exposed region of 2004-H5N1 epitope 50 induced interaction of the C-terminal Lys8 and Leu9 with DM1-TCR. (B) Reduced exposed DM1-TCR (grey) contact area caused by mutation H5N of 2004-H5N1 epitope 412. Epitopes are shown in green (2004-H5N1) and magenta (2009-H1N1). Residue side chains are presented as lines.

A2.5. Conclusions

The HLA-B*4405 and DM1-TCR docking models showed differential recognition of the 2009-H1N1 HA T cell epitope candidates, reflecting the topological constraints of the epitopes. This information was used to derive the synthetic H1N1 epitope KEKMNTEFW with optimal recognition of both HLA-B*4405 and DM1-TCR models and to identify likely cross-reactive 2004-H5N1 epitopes. While the results are novel with regard to HLA-B*4405-restricted H1N1 HA epitopes, their utility in vaccination strategies against influenza viruses is limited by the fact that the T cell responses to viruses depend on the TCR repertoire, and in particular on the nature of TCR alpha chain and their conformation as shown in a study by Zhong et al. (Zhong, Dixit *et al.* 2007). To simulate a T cell response to H1N1 epitopes on population level as it is desirable for vaccine design, a large number of crystal structure data on TCR V α and V β chains and their heterodimers would be necessary to computationally assess epitope candidates for their potential to induce a broad T cell response.

A2.6. Supplements

A2.6.1. MAFFT multiple sequence alignment of pandemic strains (2009 – the top 7 sequences)

HA proteins to HA (1918 – South Carolina) and WHO vaccine HA (the last 3 sequences).

<i>A/California/04/2009</i> (GQ117044)/1-566	1	M	K	A	I	L	V	V	L	L	Y	T	F	A	T	A	N	A	D	T	L	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43
<i>A/Mexico/4108/2009</i> (GQ223112)/1-566	1	M	K	A	I	L	V	V	L	L	Y	T	F	A	T	A	N	A	D	T	L	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43
<i>A/Paris/2573/2009</i> (238505669)/1-566	1	M	K	A	I	L	V	V	L	L	Y	T	F	A	T	A	N	A	D	T	L	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43
<i>A/England/195/2009</i> (GQ166661)/1-566	1	M	K	A	I	L	V	V	L	L	Y	T	F	A	T	A	N	A	D	T	L	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43
<i>A/Israel/276/2009</i> (CY041968)/1-566	1	M	K	A	I	L	V	V	L	L	Y	T	F	A	T	A	N	A	D	T	L	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43
<i>A/Singapore/TLL01/2009</i> (GQ392017)/1-566	1	M	E	A	I	L	V	V	L	L	Y	T	F	A	T	A	N	A	D	T	L	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43
<i>A/Canada-NS/RV1535/2009</i> (FJ998207)/1-566	1	M	K	A	I	L	V	V	L	L	Y	T	F	A	T	A	N	A	D	T	L	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43
<i>A/South_Carolina/1/18</i> (AF117241)/1-566	1	M	E	A	R	L	L	V	L	L	C	A	F	A	A	T	N	A	D	T	I	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43
<i>A/New_Caledonia/20/1999</i> (AY289929)/1-565	1	M	K	A	K	L	L	V	L	L	C	T	F	T	A	T	A	D	T	I	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43	
<i>A/Brisbane/59/2007</i> (CY030232)/1-565	1	M	K	V	K	L	L	V	L	L	C	T	F	T	A	T	A	D	T	I	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43	
<i>A/Solomon_Islands/3/2006</i> (EU124177)/1-565	1	M	K	V	K	L	L	V	L	L	C	T	F	T	A	T	A	D	T	I	C	I	G	Y	H	A	N	N	S	T	D	T	V	D	T	V	L	E	K	N	V	T	V	43	
<i>A/California/04/2009</i> (GQ117044)/1-566	44	T	H	S	V	N	L	L	E	D	K	H	N	G	K	L	C	K	L	R	G	V	A	P	L	H	L	G	K	C	N	I	A	G	W	I	L	G	N	P	E	C	E	S	86
<i>A/Mexico/4108/2009</i> (GQ223112)/1-566	44	T	H	S	V	N	L	L	E	D	K	H	N	G	K	L	C	K	L	R	G	V	A	P	L	H	L	G	K	C	N	I	A	G	W	I	L	G	N	P	E	C	E	S	86
<i>A/Paris/2573/2009</i> (238505669)/1-566	44	T	H	S	V	N	L	L	E	D	K	H	N	G	K	L	C	K	L	R	G	V	A	P	L	H	L	G	K	C	N	I	A	G	W	I	L	G	N	P	E	C	E	S	86
<i>A/England/195/2009</i> (GQ166661)/1-566	44	T	H	S	V	N	L	L	E	D	K	H	N	G	K	L	C	K	L	R	G	V	A	P	L	H	L	G	K	C	N	I	A	G	W	I	L	G	N	P	E	C	E	S	86
<i>A/Israel/276/2009</i> (CY041968)/1-566	44	T	H	S	V	N	L	L	E	D	K	H	N	G	K	L	C	K	L	R	G	V	A	P	L	H	L	G	K	C	N	I	A	G	W	I	L	G	N	P	E	C	E	S	86
<i>A/Singapore/TLL01/2009</i> (GQ392017)/1-566	44	T	H	S	V	N	L	L	E	D	K	H	N	G	K	L	C	K	L	R	G	V	A	P	L	H	L	G	K	C	N	I	A	G	W	I	L	G	N	P	E	C	E	S	86
<i>A/Canada-NS/RV1535/2009</i> (FJ998207)/1-566	44	T	H	S	V	N	L	L	E	D	K	H	N	G	K	L	C	K	L	R	G	V	A	P	L	H	L	G	K	C	N	I	A	G	W	I	L	G	N	P	E	C	E	S	86
<i>A/South_Carolina/1/18</i> (AF117241)/1-566	44	T	H	S	V	N	L	L	E	D	S	H	N	G	K	L	C	K	L	K	G	I	A	P	L	L	G	K	C	N	I	A	G	W	L	L	G	N	P	E	C	D	L	86	
<i>A/New_Caledonia/20/1999</i> (AY289929)/1-565	44	T	H	S	V	N	L	L	E	D	S	H	N	G	K	L	C	L	L	K	G	I	A	P	L	L	G	N	C	S	V	A	G	W	I	L	G	N	P	E	C	E	L	86	
<i>A/Brisbane/59/2007</i> (CY030232)/1-565	44	T	H	S	V	N	L	L	E	D	S	H	N	G	K	L	C	L	L	K	G	I	A	P	L	L	G	N	C	S	V	A	G	W	I	L	G	N	P	E	C	E	L	86	
<i>A/Solomon_Islands/3/2006</i> (EU124177)/1-565	44	T	H	S	V	N	L	L	E	D	S	H	N	G	K	L	C	L	L	K	G	I	A	P	L	L	G	N	C	S	V	A	G	W	I	L	G	N	P	E	C	E	L	86	
<i>A/California/04/2009</i> (GQ117044)/1-566	87	L	S	T	A	S	S	W	S	Y	I	V	E	T	P	S	S	D	N	G	T	C	P	G	D	F	I	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/Mexico/4108/2009</i> (GQ223112)/1-566	87	L	S	T	A	S	S	W	S	Y	I	V	E	T	P	S	S	D	N	G	T	C	P	G	D	F	I	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/Paris/2573/2009</i> (238505669)/1-566	87	L	S	T	A	S	S	W	S	Y	I	V	E	T	P	S	S	D	N	G	T	C	P	G	D	F	I	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/England/195/2009</i> (GQ166661)/1-566	87	L	S	T	A	S	S	W	S	Y	I	V	E	T	P	S	S	D	N	G	T	C	P	G	D	F	I	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/Israel/276/2009</i> (CY041968)/1-566	87	L	S	T	A	S	S	W	S	Y	I	V	E	T	P	S	S	D	N	G	T	C	P	G	D	F	I	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/Singapore/TLL01/2009</i> (GQ392017)/1-566	87	L	S	T	A	S	S	W	S	Y	I	V	E	T	P	S	S	D	N	G	T	C	P	G	D	F	I	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/Canada-NS/RV1535/2009</i> (FJ998207)/1-566	87	L	S	T	A	S	S	W	S	Y	I	V	E	T	P	S	S	D	N	G	T	C	P	G	D	F	I	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/South_Carolina/1/18</i> (AF117241)/1-566	87	L	L	T	A	S	S	W	S	Y	I	V	E	T	S	N	S	E	N	G	T	C	P	G	D	F	I	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/New_Caledonia/20/1999</i> (AY289929)/1-565	87	L	I	S	K	E	S	W	S	Y	I	V	E	T	P	N	P	E	N	G	T	C	P	G	Y	F	A	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/Brisbane/59/2007</i> (CY030232)/1-565	87	L	I	S	K	E	S	W	S	Y	I	V	E	K	P	N	P	E	N	G	T	C	P	G	H	F	A	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/Solomon_Islands/3/2006</i> (EU124177)/1-565	87	L	I	S	R	E	S	W	S	Y	I	V	E	K	P	N	P	E	N	G	T	C	P	G	H	F	A	D	Y	E	E	L	R	E	Q	L	S	S	V	S	S	F	E	29	
<i>A/California/04/2009</i> (GQ117044)/1-566	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	D	S	N	K	G	V	T	A	A	C	P	H	A	G	A	K	S	F	Y	K	N	L	I	W	L	V	K	K	G	172
<i>A/Mexico/4108/2009</i> (GQ223112)/1-566	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	D	S	N	K	G	V	T	A	A	C	P	H	A	G	A	K	S	F	Y	K	N	L	I	W	L	V	K	K	G	172
<i>A/Paris/2573/2009</i> (238505669)/1-566	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	D	S	N	K	G	V	T	A	A	C	P	H	A	G	A	K	S	F	Y	K	N	L	I	W	L	V	K	K	G	172
<i>A/England/195/2009</i> (GQ166661)/1-566	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	D	S	N	K	G	V	T	A	A	C	P	H	A	G	A	K	S	F	Y	K	N	L	I	W	L	V	K	K	G	172
<i>A/Israel/276/2009</i> (CY041968)/1-566	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	D	S	N	K	G	V	T	A	A	C	P	H	A	G	A	K	S	F	Y	K	N	L	I	W	L	V	K	K	G	172
<i>A/Singapore/TLL01/2009</i> (GQ392017)/1-566	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	D	S	N	K	G	V	T	A	A	C	P	H	A	G	A	K	S	F	Y	K	N	L	I	W	L	V	K	K	G	172
<i>A/Canada-NS/RV1535/2009</i> (FJ998207)/1-566	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	D	S	N	K	G	V	T	A	A	C	P	H	A	G	A	K	S	F	Y	K	N	L	I	W	L	V	K	K	G	172
<i>A/South_Carolina/1/18</i> (AF117241)/1-566	130	K	F	E	I	F	P	K	T	S	S	W	P	N	H	E	T	T	K	G	V	T	A	A	C	S	Y	A	G	A	S	S	F	Y	R	N	L	L	W	L	T	K	K	G	172
<i>A/New_Caledonia/20/1999</i> (AY289929)/1-565	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	-	T	V	T	G	V	S	A	S	C	S	H	N	G	K	S	S	F	Y	R	N	L	L	W	L	T	G	K	N	171
<i>A/Brisbane/59/2007</i> (CY030232)/1-565	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	-	T	V	T	G	V	S	A	S	C	S	H	N	G	E	S	S	F	Y	R	N	L	L	W	L	T	G	K	N	171
<i>A/Solomon_Islands/3/2006</i> (EU124177)/1-565	130	R	F	E	I	F	P	K	T	S	S	W	P	N	H	-	T	T	T	G	V	S	A	S	C	S	H	N	G	E	S	S	F	Y	K	N	L	L	W	L	T	G	K	N	171
<i>A/California/04/2009</i> (GQ117044)/1-566	173	N	S	Y	P	K	L	S	K	S	Y	I	N	D	K	G	E	V	L	V	L	W	G	I	H	H	P	S	T	S	A</														

<i>A/California/04/2009</i> (GQ117044Y1-566	216	V	F	V	G	S	S	R	Y	S	K	K	F	K	P	E	I	A	I	R	P	K	V	R	D	E	G	R	M	N	Y	Y	W	L	V	E	P	G	D	K	I	T	258		
<i>A/Mexico/4108/2009</i> (GQ223112Y1-566	216	V	F	V	G	S	S	R	Y	S	K	K	F	K	P	E	I	A	I	R	P	K	V	R	D	E	G	R	M	N	Y	Y	W	L	V	E	P	G	D	K	I	T	258		
<i>A/Paris/2573/2009</i> (238505669Y1-566	216	V	F	V	G	S	S	R	Y	S	K	K	F	K	P	E	I	A	I	R	P	K	V	R	D	E	G	R	M	N	Y	Y	W	L	V	E	P	G	D	K	I	T	258		
<i>A/England/195/2009</i> (GQ166661Y1-566	216	V	F	V	G	S	S	R	Y	S	K	K	F	K	P	E	I	A	I	R	P	K	V	R	D	E	G	R	M	N	Y	Y	W	L	V	E	P	G	D	K	I	T	258		
<i>A/Israel/276/2009</i> (CY041968Y1-566	216	V	F	V	G	S	S	R	Y	S	K	K	F	K	P	E	I	A	I	R	P	K	V	R	D	E	G	R	M	N	Y	Y	W	L	V	E	P	G	D	K	I	T	258		
<i>A/Singapore/TLL01/2009</i> (GQ392017Y1-566	216	V	F	V	G	S	S	R	Y	S	K	K	F	K	P	E	I	A	I	R	P	K	V	R	D	E	G	R	M	N	Y	Y	W	L	V	E	P	G	D	K	I	T	258		
<i>A/Canada-NS/RV1535/2009</i> (FJ998207Y1-566	216	V	F	V	G	S	S	R	Y	S	K	K	F	K	P	E	I	A	I	R	P	K	V	R	D	E	G	R	M	N	Y	Y	W	L	V	E	P	G	D	K	I	T	258		
<i>A/South_Carolina/1/18</i> (AF117241Y1-566	216	V	S	V	G	S	S	K	Y	N	R	R	F	T	P	E	I	A	A	R	P	K	V	R	D	E	A	G	R	M	N	Y	Y	W	L	L	E	P	G	D	T	I	T	258	
<i>A/New_Caledonia/20/1999</i> (AY289929Y1-565	215	V	S	V	V	S	S	H	Y	S	R	R	F	T	P	E	I	A	K	R	P	K	V	R	D	E	G	R	I	N	Y	Y	W	L	L	E	P	G	D	T	I	T	257		
<i>A/Brisbane/59/2007</i> (CY030232Y1-565	215	V	S	V	V	S	S	H	Y	S	R	K	F	T	P	E	I	A	K	R	P	K	V	R	D	E	G	R	I	N	Y	Y	W	L	L	E	P	G	D	T	I	T	257		
<i>A/Solomon_Islands/3/2006</i> (EU124177Y1-565	215	V	S	V	V	S	S	H	Y	S	R	K	F	T	P	E	I	A	K	R	P	K	V	R	D	E	G	R	I	N	Y	Y	W	L	L	E	P	G	D	T	I	T	257		
<i>A/California/04/2009</i> (GQ117044Y1-566	259	F	E	A	T	G	N	L	V	V	P	R	Y	A	F	A	M	E	R	N	A	G	S	G	I	I	S	D	T	P	V	H	D	C	N	T	T	C	Q	T	P	K	G	301	
<i>A/Mexico/4108/2009</i> (GQ223112Y1-566	259	F	E	A	T	G	N	L	V	V	P	R	Y	A	F	A	M	E	R	N	A	G	S	G	I	I	S	D	T	P	V	H	D	C	N	T	T	C	Q	T	P	K	G	301	
<i>A/Paris/2573/2009</i> (238505669Y1-566	259	F	E	A	T	G	N	L	V	V	P	R	Y	A	F	A	M	E	R	N	A	G	S	G	I	I	S	D	T	P	V	H	D	C	N	T	T	C	Q	T	P	K	G	301	
<i>A/England/195/2009</i> (GQ166661Y1-566	259	F	E	A	T	G	N	L	V	V	P	R	Y	A	F	A	M	E	R	N	A	G	S	G	I	I	S	D	T	P	V	H	D	C	N	T	T	C	Q	T	P	K	G	301	
<i>A/Israel/276/2009</i> (CY041968Y1-566	259	F	E	A	T	G	N	L	V	V	P	R	Y	A	F	A	M	E	R	N	A	G	S	G	I	I	S	D	T	P	V	H	E	C	N	T	T	C	Q	T	P	K	G	301	
<i>A/Singapore/TLL01/2009</i> (GQ392017Y1-566	259	F	E	A	T	G	N	L	V	V	P	R	Y	A	F	A	M	E	R	N	A	G	S	G	I	I	S	D	T	P	V	H	D	C	N	T	T	C	Q	T	P	K	G	301	
<i>A/Canada-NS/RV1535/2009</i> (FJ998207Y1-566	259	F	E	A	T	G	N	L	V	V	P	R	Y	A	F	A	M	E	R	N	A	G	S	G	I	I	S	D	T	P	V	H	D	C	N	T	T	C	Q	T	P	K	G	301	
<i>A/South_Carolina/1/18</i> (AF117241Y1-566	259	F	E	A	T	G	N	L	V	V	P	R	Y	A	F	A	M	E	R	N	A	G	S	G	I	I	S	D	T	P	V	H	D	C	N	T	T	C	Q	T	P	K	G	301	
<i>A/New_Caledonia/20/1999</i> (AY289929Y1-565	258	F	E	A	N	G	N	L	I	A	P	W	Y	A	F	A	L	S	R	G	F	G	S	G	I	I	T	S	N	A	P	M	D	E	C	D	A	K	C	Q	T	P	G	300	
<i>A/Brisbane/59/2007</i> (CY030232Y1-565	258	F	E	A	N	G	N	L	I	A	P	W	Y	A	F	A	L	S	R	G	F	G	S	G	I	I	T	S	N	A	P	M	D	E	C	D	A	K	C	Q	T	P	G	300	
<i>A/Solomon_Islands/3/2006</i> (EU124177Y1-565	258	F	E	A	N	G	N	L	I	A	P	W	Y	A	F	A	L	S	R	G	F	G	S	G	I	I	T	S	N	A	P	M	D	E	C	D	A	K	C	Q	T	P	G	300	
<i>A/California/04/2009</i> (GQ117044Y1-566	302	A	I	N	T	S	L	P	F	Q	N	I	H	P	I	T	I	G	K	C	P	K	Y	V	K	S	T	K	L	R	L	A	T	G	L	R	N	I	P	S	I	Q	S	R	344
<i>A/Mexico/4108/2009</i> (GQ223112Y1-566	302	A	I	N	T	S	L	P	F	Q	N	I	H	P	I	T	I	G	K	C	P	K	Y	V	K	S	T	K	L	R	L	A	T	G	L	R	N	V	P	S	I	Q	S	R	344
<i>A/Paris/2573/2009</i> (238505669Y1-566	302	A	I	N	T	S	L	P	F	Q	N	I	H	P	I	T	I	G	K	C	P	K	Y	V	K	S	T	K	L	R	L	A	T	G	L	R	N	V	P	S	I	Q	S	R	344
<i>A/England/195/2009</i> (GQ166661Y1-566	302	A	I	N	T	S	L	P	F	Q	N	I	H	P	I	T	I	G	K	C	P	K	Y	V	K	S	T	K	L	R	L	A	T	G	L	R	N	V	P	S	I	Q	S	R	344
<i>A/Israel/276/2009</i> (CY041968Y1-566	302	A	I	N	T	S	L	P	F	Q	N	I	H	P	I	T	I	G	K	C	P	K	Y	V	K	S	T	K	L	R	L	A	T	G	L	R	N	V	P	S	I	Q	S	R	344
<i>A/Singapore/TLL01/2009</i> (GQ392017Y1-566	302	A	I	N	T	S	L	P	F	Q	N	I	H	P	I	T	I	G	K	C	P	K	Y	V	K	S	T	K	L	R	L	A	T	G	L	R	N	V	P	S	I	Q	S	R	344
<i>A/Canada-NS/RV1535/2009</i> (FJ998207Y1-566	302	A	I	N	T	S	L	P	F	Q	N	I	H	P	I	T	I	G	K	C	P	K	Y	V	K	S	T	K	L	R	L	A	T	G	L	R	N	V	P	S	I	Q	S	R	344
<i>A/South_Carolina/1/18</i> (AF117241Y1-566	302	A	I	N	S	S	L	P	F	Q	N	I	H	P	V	T	I	G	E	C	P	K	Y	V	R	S	T	K	L	R	A	T	G	L	R	N	I	P	S	I	Q	S	R	344	
<i>A/New_Caledonia/20/1999</i> (AY289929Y1-565	301	A	I	N	S	S	L	P	F	Q	N	V	H	P	V	T	I	G	E	C	P	K	Y	V	R	S	A	K	L	R	M	V	T	G	L	R	N	I	P	S	I	Q	S	R	343
<i>A/Brisbane/59/2007</i> (CY030232Y1-565	301	A	I	N	S	S	L	P	F	Q	N	V	H	P	V	T	I	G	E	C	P	K	Y	V	R	S	A	K	L	R	M	V	T	G	L	R	N	I	P	S	I	Q	S	R	343
<i>A/Solomon_Islands/3/2006</i> (EU124177Y1-565	301	A	I	N	S	S	L	P	F	Q	N	V	H	P	V	T	I	G	E	C	P	K	Y	V	R	S	A	K	L	R	M	V	T	G	L	R	N	I	P	S	I	Q	S	R	343
<i>A/California/04/2009</i> (GQ117044Y1-566	345	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	L	K	S	T	Q	N	387
<i>A/Mexico/4108/2009</i> (GQ223112Y1-566	345	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	L	K	S	T	Q	N	387
<i>A/Paris/2573/2009</i> (238505669Y1-566	345	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	L	K	S	T	Q	N	387
<i>A/England/195/2009</i> (GQ166661Y1-566	345	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	L	K	S	T	Q	N	387
<i>A/Israel/276/2009</i> (CY041968Y1-566	345	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	L	K	S	T	Q	N	387
<i>A/Singapore/TLL01/2009</i> (GQ392017Y1-566	345	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	L	K	S	T	Q	N	387
<i>A/Canada-NS/RV1535/2009</i> (FJ998207Y1-566	345	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	L	K	S	T	Q	N	387
<i>A/South_Carolina/1/18</i> (AF117241Y1-566	345	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	I	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	Q	K	S	T	Q	N	387
<i>A/New_Caledonia/20/1999</i> (AY289929Y1-565	344	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	Q	K	S	T	Q	N	386
<i>A/Brisbane/59/2007</i> (CY030232Y1-565	344	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	Q	K	S	T	Q	N	386
<i>A/Solomon_Islands/3/2006</i> (EU124177Y1-565	344	G	L	F	G	A	I	A	G	F	I	E	G	G	W	T	G	M	V	D	G	W	Y	G	Y	H	H	Q	N	E	Q	G	S	G	Y	A	A	D	Q	K	S	T	Q	N	386
<i>A/California/04/2009</i> (GQ117044Y1-566	388	A	I	D	E	I	T	N	K	V	N	S	V	I	E	K	M	N	T	Q	F	T	A	V	G	K	E	F	N	H	L	E	K	R	I	E	N	L	N	K	K	V	D	430	
<i>A/Mexico/4108/2009</i> (GQ223112Y1-566	388	A	I	D	E	I	T	N	K	V	N	S	V																																

A2.6.2. Predicted HA T cell epitopes of H1N1 2009 strains compared to 1918 and WHO-1999-2006-2007 strains using NetCTL v1.2.

Position	Strains	Sequence	Conservancy (%)	MHC-binding affinity	Rescaled MHC-binding affinity	C-terminal cleavage affinity	TAP efficiency	Overall Score
50	2009	LEDKHN GKL	63.6	0.1253	0.8663	0.9978	0.695	1.0508
	1918	LEDSHN GKL		0.1227	0.8482	0.9983	0.695	1.0327
	WHO	LENSHN GKL		0.1407	0.9727	0.9982	0.777	1.1612
128	2009	FERFEIF PK	100	0.1336	0.9238	0.7674	0.337	1.0558
	1918	FEKFEIF PK		0.1386	0.9582	0.8127	0.286	1.0944
	WHO	FERFEIF PK		0.1336	0.9238	0.9326	0.337	1.0806
188	2009	KEVLVL WGI	63.6	0.152	1.0509	0.8717	0.554	1.2094
	1918	KEVLVL WGV		0.1473	1.0184	0.9885	0.334	1.1833

Position	Strains	Sequence	Conservancy (%)	MHC-binding affinity	Rescale MHC-binding affinity	C-terminal cleavage affinity	TAP efficiency	Overall Score
	WHO	KEVLVL WGV		0.1473	1.0184	0.9949	0.197	1.1774
251	2009	<u>VEPGDKI</u> TF	63.6	0.1047	0.724	0.9929	2.494	0.9976
	1918	<u>LEPGDTI</u> TF		0.1103	0.7628	0.9988	2.436	1.0344
	WHO	<u>LEPGDTI</u> IF		0.1058	0.7313	0.9971	2.436	1.0026
412	2009	KEFNHL EKR	63.6	0.1166	0.8062	0.6221	1.538	0.9764
	1918	KEFNHL ERR		0.1208	0.8354	0.6364	1.538	1.0078
	WHO	KEFNKL ERR		0.0938	0.6488	0.7765	1.538	0.8422
446	2009	LENERTL DY	100	0.1312	0.9073	0.988	2.747	1.1929
	1918	LENERTL DF		0.161	1.1134	0.573	2.357	1.3172
	WHO	LENERTL DF		0.161	1.1134	0.4891	2.357	1.3046
463	2009	YEKVR QLK	63.6	0.0896	0.6193	0.9868	0.355	0.7851
	1918	YEKVK QLK		0.0969	0.6698	0.9897	0.355	0.836
	WHO	YEKVK QLK		0.0969	0.6698	0.9897	0.355	0.836
493	2009	MESVKN GT	90.9	0.2081	1.4391	0.998	2.825	1.73
	1918	MESVRN GT		0.1986	1.3735	0.9976	2.825	1.6644
	WHO	MESVKN GT		0.2081	1.4391	0.998	2.825	1.73
514	2009	REEIDGV KL	72.7	0.1046	0.723	0.997	1.06	0.9256
	1918	REEIDGV KL		0.1046	0.723	0.9914	1.06	0.9247
	WHO	REKIDGV KL		0.0992	0.6857	0.9922	1.202	0.8946
131	2009	FEIFPKT SS	72.7	0.162	1.12	0.0182	-2.275	1.009
	1918	FEIFPKT SS		0.162	1.12	0.0128	-2.306	1.0066
	WHO	FEIFPKE SS		0.1425	0.9856	0.0101	-2.275	0.8733
229	2009	PEIAIRP KV	63.6	0.1132	0.7828	0.9532	-0.131	0.9192
	1918	PEIAARP KV		0.1452	1.0043	0.715	-0.222	1.1005
	WHO	PEIAKRP KV		0.1113	0.7693	0.8004	-0.222	0.8782
400	2009	IEKMNT QFT	100	0.1504	1.0402	0.0188	-0.773	1.0043
	1918	IEKMNT QFT		0.1504	1.0402	0.0187	-0.773	1.0043
	WHO	IEKMNT QFT		0.1504	1.0402	0.0167	-0.773	1.004
421	2009	<u>IENLNKK</u> VD	72.7	0.1372	0.9484	0.0159	-1.89	0.8563
	1918	IENLNKK VD		0.1372	0.9484	0.0159	-1.89	0.8563
	WHO	<u>MENLNK</u> KVD		0.1679	1.1613	0.0183	-1.866	1.0707
475	2009	KEIGNG CFE	100	0.1528	1.0563	0.0098	-1.371	0.9892
	1918	KEIGNG CFE		0.1528	1.0563	0.0098	-1.371	0.9892
	WHO	KEIGNG CFE		0.1528	1.0563	0.0098	-1.371	0.9892
482	2009	FEFYHK CDN	63.6	0.1711	1.1834	0.0097	-1.441	1.1128
	1918	FEFYHK CDD		0.207	1.4315	0.0102	-1.941	1.336
	WHO-2006-2007	FEFYHK CND		0.2043	1.4127	0.0103	-1.941	1.3172

Position	Strains	Sequence	Conservancy (%)	MHC-binding affinity	Rescaled MHC-binding affinity	C-terminal cleavage affinity	TAP efficiency	Overall Score
	WHO-1999	FEFYHK CNN		0.1693	1.1706	0.0099	-1.441	1.1
240	2009	QEGRMN YYW	63.6	0.1926	1.3318	0.8493	0.46	1.4822
	1918	QAGRMN YYW		0.0204	0.1412	0.829	0.648	0.2979
	WHO 1999, 2007	QEGRIN YYW		0.1784	1.2333	0.9916	0.46	1.405
	WHO 2006	<u>REGRIN</u> YYW		0.1757	1.2148	0.9887	0.691	1.3976
259	2009	FEATGN LVV	63.6	0.1014	0.7011	0.8355	0.111	0.832
	1918	FEATGN LIA		0.0874	0.6041	0.0246	-0.739	0.5709
	WHO	FEANGN LIA		0.0713	0.4927	0.2799	-0.74	0.4977
81	2009	<u>PECESL</u> STA	63.6	0.1057	0.7307	0.7034	-1.001	0.7862
	1918	<u>NPECDL</u> LLT		0.014	0.0965	0.0507	-1.147	0.0468
	WHO	PECELLI SR		0.0695	0.4807	0.4991	1.019	0.6065
274	2009	<u>MERNAG</u> SGI	63.6	0.1138	0.7869	0.5974	0.649	0.9089
	1918	<u>LNRGSG</u> SGI		0.0166	0.1151	0.0507	0.589	0.1521
	WHO	<u>LSRGFG</u> SGI		0.014	0.0971	0.0864	0.782	0.1492

A2.6.3. Physicochemical properties of NetCTL-predicted HLA-B44 restricted T cell epitope candidates (A) and non-epitopes (B).

A		B	
50_2009/1-9	LEDKHNQKL	81_WHO-1999-2007/1-9	PECELLISK
50_WHO-1999-2006/1-9	LEDSHNQKL	81_WHO-2006/1-9	PECELLISR
50_1918/1-9	LEDSHNQKL	81_1918/1-9	PECDLLLTA
50_WHO-2007/1-9	LENSHNQKL	259_WHO-1999-2006-2007/1-9	FEANGNLIA
493_2009/1-9	MESVKNGTY	259_1918/1-9	FEANGNLIA
493_WHO-1999-2006-2007/1-9	MESVKNGTY	240_WHO-1999-2006-2007/1-9	QEGRIYYW
493_1918/1-9	MESVRNGTY	240_1918/1-9	QAGRMYYW
251_2009/1-9	VEPGDKITF	274_WHO-1999-2006-2007/1-9	LSRGFGSGI
251_WHO-1999-2006-2007/1-9	LEPGDTITF	274_1918/1-9	LNRGSGSGI
251_1918/1-9	LEPGDTITF		
446_2009/1-9	LENERTLDY		
446_WHO-1999-2006-2007/1-9	LENERTLDF		
446_1918/1-9	LENERTLDF		
463_2009/1-9	YEKVRSQLK		
463_1918/1-9	YEKVRSQLK		
463_WHO-1999-2006-2007/1-9	YEKVRSQLK		
400_2009/1-9	IEKMNTQFT		
400_WHO-1999-2006-2007/1-9	IEKMNTQFT		
400_1918/1-9	IEKMNTQFT		
421_2009/1-9	IEKLNKKVD		
421_WHO-1999-2006-2007/1-9	IEKLNKKVD		
128_2009/1-9	FERFEIFPK		
128_WHO-1999-2006-2007/1-9	FERFEIFPK		
128_1918/1-9	FERFEIFPK		
131_2009/1-9	FEIFPKTSS		
131_WHO-1999-2006-2007/1-9	FEIFPKTSS		
482_2009/1-9	FEFYHKKCN		
482_WHO-1999-2006-2007/1-9	FEFYHKKCN		
259_2009/1-9	FEATGNLVV		
514_2009/1-9	REEIDGVKL		
514_WHO-1999-2006-2007/1-9	REEIDGVKL		
274_2009/1-9	MERNAGSGI		
412_2009/1-9	KEFNHLEKR		
412_WHO-1999-2006-2007/1-9	KEFNHLEKR		
412_1918/1-9	KEFNHLEKR		
475_2009/1-9	KEIGNGCFE		
475_WHO-1999-2006-2007/1-9	KEIGNGCFE		
475_1918/1-9	KEIGNGCFE		
188_2009/1-9	KEVLVLWGVI		
188_WHO-1999-2006-2007/1-9	KEVLVLWGVI		
229_2009/1-9	PEIAIRPKV		
229_WHO-1999-2006-2007/1-9	PEIAIRPKV		
229_1918/1-9	PEIAIRPKV		
81_2009/1-9	PECESLSTA		
240_2009/1-9	QEGRMYYW		

Aliphatic/hydrophobic	ILVAM
Aromatic	FWY
Positive	KRH
Negative	DE
Hydrophilic	STNQ
conformationally special	PG
Cysteine	C

A2.6.4. Residues of domain $\alpha 1$, $\alpha 2$ of HLA-B*4405 interacting with DM1-TCR

exposed/embedded	HLA- α_1	HLA- α_2	TCR V $_{\alpha}$	TCR V $_{\beta}$	H-bond Length (Å)	Dihedral ($^{\circ}$)
259	Asn ⁷⁰		Glu ⁷⁰		2.1	32.3
	Glu ⁷⁶		Arg ⁸⁰		2.2	-54.1
240	Asn ⁷⁰			Pro ²⁵	2.5	-119.7
	Glu ⁷⁶			Gln ⁶	2.5	-133.7
400		Gln ¹⁵⁵	Gln ⁵⁴		2.2	-16.3
482		Gln ¹⁵⁵		Ser ²⁷	2.5	84.1

A3. Scores of other false positive and incorrect docked conformations

Dockings to predict the structures of the HLA/peptide complexes were performed using different docking programs FTDock, ZDOCK version 2.1, 2.3, 3.0, and PIPER. Each program obtained the best conformation, which was 1st ranked, some correct bindings, as well as false positives (the conformations that were defined as incorrect conformations, but stayed in the first 100 ranks). Percentages of false positives were 4.75%, 4.7%, 4.65%, 3.09%, and 4.05% respectively for each docking program above. The scores of the false positive conformations were tabulated as below.

Rank	FTDOCK (SC/ELEC)	ZDOCK2.1 (PSC)	ZDOCK2.3 (PSC+DE+ELEC using ACE)	ZDOCK3.0.2 (Improved ACE)	PIPER (DARS)
1	66/20.75	-	-	-	-
2	62/12.8	14.04	-	-	-
3	60/37.57	-	-	-	-
4	57/18.95	13.4	26.8	-	844.8
5	56/24.7	13.34	26.5	-	-
6	56/15.24	13.18	-	-	-
7	55/35.93	13.12	26.19	-	-
8	55/18.05	13.10	25.88	-	779.4
9	55/3.34	13.04	25.71	-	779.4
10	55/3.04	13.02	25.30	-	776.6

*Scores of the correct conformations were given in Table 2-3.

This result could not clearly support the accuracy of these performances. However, it showed that ZDOCK3.0.2 was one of the two programs that obtained fewer numbers of false positives compared to the others.

A4. Torsion angles (ω, ϕ, φ) of the 10-mer peptide backbone from various docked complexes

Residue	ω				ϕ				φ			
Glu ¹	-	-	-	-	-	-	-	-	139.23	139.24	139.24	98.6
Glu ²	-178.4	-178.38	-178.38	169.03	-79.93	-79.95	-79.95	-90.77	164.62	164.61	164.61	159.87
Asn ³	172.78	172.76	172.76	-171.59	-90.68	-90.67	-90.67	-99.43	126.57	126.57	126.57	142.4
Leu ⁴	169.72	169.69	169.69	165.76	-95.63	-95.59	-95.59	-140.52	86.88	86.89	86.89	121.35
Leu ⁵	-172.62	-172.60	-172.60	170.99	-84.09	-84.11	-84.11	-119.43	125.14	125.15	125.15	152.75
Asp ⁶	-169.34	-169.32	-169.32	172.04	-84.89	-84.89	-84.89	-69.29	125.93	125.92	125.92	153.41
Phe ⁷	178.13	178.13	178.13	156.7	57.18	57.21	57.21	58.3	26.82	26.75	26.75	6.71
Val ⁸	-179.72	-179.72	-179.72	174.62	-85.72	-85.68	-85.68	-96.7	118.05	118.04	118.04	106.25
Arg ⁹	178.33	178.37	178.37	-176.14	-148.02	-148.01	-148.01	-141.19	156.61	156.59	156.59	160.79
Phe ¹⁰	-176.21	-176.27	-176.27	171.89	-61.53	-61.51	-61.51	-64.16	-	-	-	-

The torsion values of the **original**, **ZDOCK2.1**, **ZDOCK3.0.2**, and the **20ns-MD** complexes are presented in **black**, **red**, **green**, and **blue**, respectively

A5. Statistics of interfacial water of complexes extracted from benchmark 4.0

Complex	Category*	#water	#interfacial water	#internal water	#internal water recovered ($\leq 1.5\text{\AA}$)
<i>Easy (113)</i>					
1AHW	A	0	0	0	0
1BVK	A	0	0	0	0
1DQJ	A	467	39	17	17
1E6J	A	0	0	0	0
1JPS	A	579	48	18	18
1MLC	A	117	10	11	11
1VFB	A	48	39	12	12
1WEJ	A	637	41	18	18
2FD6	A	314	17	15	15
2I25	A	135	26	10	9
2VIS	A	0	0	0	0
1FSK	AB	0	0	0	0
1IQD	AB	477	40	22	22
1K4C	AB	469	33	18	18
1KXQ	AB	700	64	58	58
1NCA	AB	72	2	22	22
1NSN	AB	0	0	0	0
2JEL	AB	67	11	11	11
1AVX	E	121	20	19	19
1AY7	E	190	28	6	6
1BVN	E	161	19	41	41
1CGI	E	52	6	11	11
1CLV	E	273	37	66	66
1D6R	E	116	9	21	21
1DFJ	E	46	7	8	6
1E6E	E	140	15	12	12
1EAW	E	43	10	6	6
1EWY	E	55	7	1	1
1F34	E	102	20	14	14
1FLE	E	157	15	14	14
1GL1	E	107	11	22	22

Complex	Category*	#water	#interfacial water	#internal water	#internal water recovered ($\leq 1.5\text{\AA}$)
1GXD	E	0	0	0	0
1HIA	E	163	20	20	20
1JTG	E	232	43	28	28
1MAH	E	0	0	0	0
1N8O	E	217	19	22	22
1OC0	E	137	11	17	17
1OPH	E	377	28	29	29
1OYV	E	52	5	4	4
1PPE	E	140	24	23	23
1R0R	E	405	40	13	13
1TMQ	E	331	34	51	51
1UDI	E	47	12	10	10
1YVB	E	14	0	3	3
2ABZ	E	0	0	0	0
2B42	E	65	16	18	18
2J0T	E	21	3	1	1
2MTA	E	104	10	34	34
2O8V	E	0	0	0	0
2OUL	E	171	23	14	14
2PCC	E	249	30	27	27
2SIC	E	258	34	16	16
2SNI	E	168	17	10	10
2UUY	E	510	45	25	25
3SGQ	E	152	19	5	5
4CPA	E	0	0	0	0
7CEI	E	162	26	1	1
1AK4	O	174	5	33	0
1AZS	O	69	4	15	15
1B6C	O	22	1	5	5
1BUH	O	127	9	9	8
1E96	O	77	8	7	6
1EFN	O	45	6	7	6
1F51	O	0	0	0	0
1FC2	O	0	0	0	0
1FCC	O	0	0	0	0
1FFW	O	87	17	2	2

Complex	Category*	#water	#interfacial water	#internal water	#internal water recovered ($\leq 1.5\text{\AA}$)
1FQJ	O	209	33	12	12
1GCQ	O	128	24	1	1
1GHQ	O	542	33	17	0
1GLA	O	0	0	0	0
1GPW	O	169	35	15	15
1H9D	O	26	11	7	7
1HCF	O	49	13	3	3
1HE1	O	361	61	15	15
1J2J	O	278	34	8	8
1JWH	O	135	1	6	6
1KAC	O	70	9	6	6
1KLU	O	380	29	27	27
1KTZ	O	163	28	4	4
1KXP	O	388	67	25	25
1ML0	O	43	2	1	0
1OFU	O	283	18	18	18
1PVH	O	104	14	7	7
1QA9	O	0	0	0	0
1RLB	O	0	0	0	0
1RV6	O	139	17	10	10
1S1Q	O	153	32	3	3
1SBB	O	87	10	4	4
1T6B	O	139	9	23	23
1US7	O	176	23	7	7
1WDW	O	46	1	13	13
1XD3	O	495	52	19	19
1XU1	O	114	9	13	13
1Z0K	O	222	37	9	8
1Z5Y	O	252	30	6	6
1ZHH	O	333	52	17	17
1ZHI	O	66	11	7	7
2A5T	O	342	53	16	16
2A9K	O	210	30	18	18
2AJF	O	40	5	7	6
2AYO	O	0	0	0	0
2B4J	O	171	15	18	18

Complex	Category*	#water	#interfacial water	#internal water	#internal water recovered ($\leq 1.5\text{\AA}$)
2BTF	O	0	0	0	0
2FJU	O	486	18	51	51
2G77	O	416	59	33	0
2HLE	O	79	13	10	10
2HQS	O	713	64	22	21
2OOB	O	134	17	0	0
2OOR	O	91	13	22	22
2VDB	O	0	0	0	0
3BP8	O	0	0	0	0
3D5S	O	69	1	11	11
<i>Medium (26)</i>					
1BGX	A	181	10	0	0
1ACB	E	142	18	25	25
1JIW	E	592	54	44	44
1KKL	E	149	10	9	9
1M10	E	0	0	0	0
1NW9	E	0	0	0	0
1GRN	O	72	12	13	13
1HE8	O	6	1	4	4
1I2M	O	242	49	34	34
1IB1	O	72	10	9	9
1K5D	O	98	23	19	19
1LFD	O	141	16	5	5
1MQ8	O	2	1	2	2
1R6Q	O	86	17	5	5
1SYX	O	85	5	10	10
1WQ1	O	35	14	8	8
1XQS	O	36	15	1	1
1ZM4	O	0	0	0	0
2CFH	O	28	3	1	1
2H7V	O	41	2	0	0
2HRK	O	274	36	9	9
2J7P	O	277	52	28	27
2NZ8	O	242	61	15	15
2OZA	O	60	13	10	10
2Z0E	O	184	34	20	20

Complex	Category*	#water	#interfacial water	#internal water	#internal water recovered ($\leq 1.5\text{\AA}$)
3CPH	O	76	11	7	7
<i>Difficult (20)</i>					
1E4K	A	0	0	0	0
1F6M	E	60	4	4	4
1FQ1	E	0	0	0	0
1PXV	E	228	34	11	11
1ZLI	E	505	66	29	0
2O3B	E	227	24	21	21
1ATN	O	0	0	0	0
1BKD	O	26	4	8	8
1FAK	O	317	58	40	0
1H1V	O	346	38	3	3
1IBR	O	198	53	17	16
1IRA	O	86	22	9	9
1JK9	O	6	2	0	0
1JZD	O	124	20	8	8
1R8S	O	350	57	10	10
1Y64	O	0	0	0	0
2C0L	O	99	14	3	3
2I9B	O	25	2	0	0
2IDO	O	212	39	3	2
2OT3	O	463	44	32	31
Recovered =					91.3%

*A = Antigen/Antibody complex, AB = Antigen/bound-Antibody complex, E = Enzyme/Inhibitor complex, O = others

A6. Re-rankings results that are equivalent with or better than the F2Dock (full table). Results of F2Dock are done by Chowdhury et al.

Complex	Cat*	Rank of the first Hit				RMS of the first Hit			
		ZDOCK	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
Easy (113)									
1AHW	A	3	-	354	8	2.49	-	4.5	4.4
1BVK	A	1	2	184	205	0.94	1.61	3.6	4.9
1DQJ	A	1	1	374	74	0.23	1.56	4.0	4.9
1E6J	A	1	2	3	126	0.42	0.42	4.1	5
1JPS	A	1	1	1266	186	0.3	0.3	2.1	2.7
1MLC	A	1	1	57	11	0.48	0.48	2	3.8
1VFB	A	1	4	250	191	0.4	0.4	3.1	4.8
1WEJ	A	1	1	9	5	0.5	0.5	1.5	3.2
2FD6	A	1	1	3	62	0.55	0.55	5.0	4.4
2I25	A	1	1	2	122	0.6	0.6	3.0	3.9
2VIS	A	1	3	-	-	0.68	0.68	-	-
1FSK	AB	-	-	1	1	-	-	2.9	3.2
1IQD	AB	-	1	18	4	-	0.35	4.3	3.0
1K4C	AB	1	1	583	105	0.49	0.49	2.9	4.4
1KXQ	AB	-	1	2	1	-	0.09	1.2	1.6
1NCA	AB	-	-	11	168	-	-	1.7	3.7
1NSN	AB	-	-	1267	-	-	-	1.6	-
2JEL	AB	-	-	4	1	-	-	3.5	3.3
1AVX	E	1	1	25	1	0.31	0.31	3.5	4.5
1AY7	E	-	-	577	2	-	-	2.5	4.0
1BVN	E	1	1	3	1	0.21	0.21	1.2	3.2
1CGI	E	-	-	10	76	-	-	4.0	3.4
1CLV	E	1	1	3	1	0.35	0.35	2.3	2.5
1D6R	E	-	-	-	59	-	-	-	4.6
1DFJ	E	1	1	1	9	0.47	0.47	4.1	4.4
1E6E	E	1	1	5	20	0.11	0.11	3.2	4.7
1EAW	E	3	2	68	1	2.04	2.06	3.4	1.0
1EWY	E	-	-	53	14	-	-	4.2	3.2
1F34	E	-	-	62	2	-	-	3.4	4.3
1FLE	E	6	4	31	3	2.34	2.34	5.0	3.7

Complex	Cat*	Rank of the first Hit				RMS of the first Hit			
		ZDOCK	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
1GL1	E	-	-	73	326	-	-	2.6	3.8
1GXD	E	1	1	1173	-	0.8	0.8	4.9	-
1HIA	E	-	-	-	18	-	-	-	3.4
1JTG	E	1	1	1	7	0.19	0.19	2.6	4.6
1MAH	E	-	-	1	1	-	-	3.1	2.7
1N8O	E	-	-	7	11	-	-	3.4	4.8
1OC0	E	-	-	1590	-	-	-	4.8	-
1OPH	E	1	1	1694	-	0.49	1.81	3.9	-
1OYV	E	-	-	15	7	-	-	4.9	3.6
1PPE	E	25	30	1	1	2.21	2.21	2.9	2.3
1R0R	E	38	39	138	39	2.32	2.32	2.2	4.3
1TMQ	E	1	1	16	1	0.33	0.33	3.6	4.8
1UDI	E	-	-	24	1	-	-	3.5	3.1
1YVB	E	-	2	1	-	-	1.86	2.4	-
2ABZ	E	2	2	-	5	2.25	2.25	-	2.8
2B42	E	1	1	3	1	0.24	0.24	4.2	3.9
2J0T	E	-	-	-	19	-	-	-	2.8
2MTA	E	-	-	76	90	-	-	4.4	4.3
2O8V	E	-	-	29	654	-	-	5.0	3.7
2OUL	E	-	-	1	1	-	-	1.7	4.9
2PCC	E	-	-	496	10	-	-	2.6	4.3
2SIC	E	-	-	5	1	-	-	1.1	1.1
2SNI	E	-	-	177	1	-	-	3.8	4.7
2UUY	E	-	-	693	7	-	-	4.4	4.1
3SGQ	E	-	-	428	110	-	-	4.0	2.6
4CPA	E	51	4	1	1	2.0	2.04	4.4	4.8
7CEI	E	5	5	1	1	1.29	1.29	4.4	4.1
1AK4	O	-	3	1090	964	-	0.36	3.5	4.3
1AZS	O	1	1	42	-	1.99	1.99	2.9	-
1B6C	O	-	-	1	3	-	-	2.9	4
1BUH	O	-	-	30	431	-	-	3.6	4.8
1E96	O	-	-	1171	278	-	-	3.8	5.0
1EFN	O	-	-	-	-	-	-	-	-
1F51	O	-	-	589	-	-	-	4.6	-
1FC2	O	1	1	-	1190	0.44	0.44	-	5.0

Complex	Cat*	Rank of the first Hit				RMS of the first Hit			
		ZDOCK	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
1FCC	O	27	-	-	-	8.33	-	-	-
1FFW	O	2	2	73	325	2.56	0.35	4.5	4.7
1FQJ	O	1	1	-	-	0.31	0.31	-	-
1GCQ	O	1	1	1105	-	0.42	0.42	1.4	-
1GHQ	O	-	-	-	-	-	-	-	-
1GLA	O	-	-	1708	-	-	-	3.9	-
1GPW	O	1	1	3	1	0.48	0.48	3.6	3.7
1H9D	O	-	1	1006	-	-	0.63	4.5	-
1HCF	O	-	1	175	1225	-	0.29	4.0	4.7
1HE1	O	-	-	1141	574	-	-	4.7	4.9
1J2J	O	-	-	-	182	-	-	-	4.6
1JWH	O	-	-	7	-	-	-	3.6	-
1KAC	O	1	1	592	8	0.32	0.32	4.5	4.4
1KLU	O	-	-	1957	-	-	-	3.4	-
1KTZ	O	-	-	535	98	-	-	2.8	3.9
1KXP	O	1	1	1	7	0.57	0.57	1.6	3.5
1ML0	O	-	-	4	2	-	-	3.1	4.3
1OFU	O	1	1	84	-	0.45	0.45	4.5	-
1PVH	O	1	1	748	-	2.08	2.08	4.5	-
1QA9	O	2	1	-	-	0.13	0.13	-	-
1RLB	O	-	-	3	555	-	-	4.6	5.0
1RV6	O	-	3	2	2	-	0.16	1.3	4.0
1S1Q	O	-	-	756	-	-	-	1.9	-
1SBB	O	-	-	-	-	-	-	-	-
1T6B	O	1	1	58	525	0.15	0.15	3.6	4.0
1US7	O	1	1	74	-	0.48	0.48	1.1	-
1WDW	O	-	-	2	1	-	-	1.2	2.5
1XD3	O	-	-	8	1	-	-	4.0	4.2
1XU1	O	85	41	912	-	2.01	2.01	5.0	-
1Z0K	O	-	1	8	307	-	0.22	3.3	3.3
1Z5Y	O	1	1	20	-	0.31	0.31	3.4	-
1ZHH	O	1	1	-	-	0.43	0.43	-	-
1ZHI	O	1	1	65	202	0.23	0.23	4.4	4.0
2A5T	O	1	1	-	268	0.73	0.73	-	3.6
2A9K	O	-	2	-	558	-	0.41	-	3.4

Complex	Cat*	Rank of the first Hit				RMS of the first Hit			
		ZDOCK	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
2AJF	O	-	-	475	-	-	-	3.6	-
2AYO	O	32	10	37	1108	2.06	2.31	3.3	2.0
2B4J	O	1	1	-	-	0.54	0.54	-	-
2BTF	O	6	1	53	95	0.45	0.45	4.7	4.5
2FJU	O	-	-	261	228	-	-	3.2	4.2
2G77	O	-	-	15	8	-	-	1.5	3.7
2HLE	O	2	1	31	4	2.16	0.44	4.1	3.8
2HQS	O	1	1	-	27	0.55	0.55	-	4.1
2OOB	O	-	-	-	-	-	-	-	-
2OOR	O	-	1	766	16	-	0.21	4.4	4.0
2VDB	O	57	18	5	-	2.06	2.36	1.2	-
3BP8	O	11	3	-	474	2.69	2.46	-	5.0
3D5S	O	-	-	71	1	-	-	3.1	3.2
<i>Medium (26)</i>									
1BGX	A	-	1	-	-	-	0.17	-	-
1ACB	E	-	-	126	22	-	-	4.4	3.2
1JIW	E	5	5	-	-	5.21	2.15	-	-
1KKL	E	-	-	-	-	-	-	-	-
1M10	E	1	1	-	-	0.78	0.78	-	-
1NW9	E	1	1	-	321	0.76	0.76	-	4.9
1GRN	O	1	1	1299	401	0.52	0.52	4.3	4.8
1HE8	O	1	1	-	-	0.31	0.31	-	-
1I2M	O	1	1	267	545	0.56	0.56	2.2	2.6
1IB1	O	1	1	-	-	0.51	0.51	-	-
1K5D	O	1	1	-	521	0.35	0.35	-	4.3
1LFD	O	1	1	85	990	0.24	0.24	4.6	4.6
1MQ8	O	1	1	1455	-	0.27	0.27	3.2	-
1R6Q	O	1	1	-	180	0.28	0.28	-	3.7
1SYX	O	1	1	211	2	0.56	0.56	4.8	4.7
1WQ1	O	1	1	81	-	0.29	0.29	4.0	-
1XQS	O	1	1	19	61	0.44	0.44	3.8	4.2
1ZM4	O	1	2	6	-	0.33	0.33	4.1	-
2CFH	O	1	1	1	119	0.59	0.59	3.8	2.6
2H7V	O	1	1	1112	-	0.44	0.44	4.6	-
2HRK	O	1	1	3	-	0.39	0.39	3.7	-

Complex	Cat*	Rank of the first Hit				RMS of the first Hit			
		ZDOCK	IFACEwat	ZDOCK	F ² Dock	ZDOCK	IFACEwat	ZDOCK	F ² Dock
2J7P	O	1	1	-	-	0.53	0.53	-	-
2NZ8	O	1	1	64	-	0.35	0.35	4.5	-
2OZA	O	1	1	-	-	0.53	0.53	-	-
2Z0E	O	1	1	-	169	0.61	0.61	-	3.9
3CPH	O	1	1	-	250	0.44	0.44	-	4.3
<i>Difficult (20)</i>									
1E4K	A	1	1	-	-	0.27	0.27	-	-
1F6M	E	-	-	-	-	-	-	-	-
1FQ1	E	1	1	-	-	0.68	0.68	-	-
1PXV	E	1	1	-	-	0.12	0.12	-	-
1ZLI	E	-	-	-	-	-	-	-	-
2O3B	E	7	5	-	-	2.47	2.47	-	-
1ATN	O	1	1	-	1307	0.26	0.26	-	2.7
1BKD	O	1	1	-	-	0.45	0.45	-	-
1FAK	O	1	1	-	-	0.17	0.17	-	-
1H1V	O	1	1	-	-	0.59	0.59	-	-
1IBR	O	1	1	-	-	0.32	0.32	-	-
1IRA	O	1	1	-	-	0.65	0.65	-	-
1JK9	O	1	1	510	422	0.15	0.15	4.2	2.5
1JZD	O	1	1	44	144	0.53	0.53	4.6	4
1R8S	O	1	1	-	-	0.47	0.47	-	-
1Y64	O	1	1	-	-	1.08	1.08	-	-
2C0L	O	1	1	-	-	0.46	0.46	-	-
2I9B	O	1	1	-	-	0.65	0.65	-	-
2IDO	O	1	1	130	156	0.38	0.38	3.6	4.5
2OT3	O	1	1	121	-	0.6	0.6	4.6	-

A7. Details of software used in this thesis

ZDOCK3.0.2

ZDOCK version 3.0.2 is the latest version of the docking program ZDOCK—one of the most currently popular *ab initio* rigid docking algorithms. The ZDOCK 3.0.2 program can be retrieved from <http://zdock.umassmed.edu/software/download/> after registering for academic use.

Further, a brief instruction of how to do docking with ZDOCK is presented in http://zlab.umassmed.edu/zdock/how_to_run_zdock.shtml.

The process to obtain 2000 complex predictions by ZDOCK is presented as the Pseudocode 1 as below:

Pseudocode 1. Process of generating 2000 docked complexes using ZDOCK 3.0.2

Input : Atomic coordinates of Receptor *R* and Ligand *L*, *structure*

List of complex interfaces, *comp_interface*

List of ligand rotation, *lig_rot*, and translation, *lig_trans*

Output: Top 2000 predictions of the protein complexes

1. # Assign IFACE value for each atom in *structure R* or *L* using Table 4-1
 2. **for** each amino acid *aa* in *structure* **do**
 3. **if** *aa* $\in$ [*comp_interface*] **then**
 4. Assign IFACE according to Table 4-1
 5. **else**
 6. Ignore
 7. **end if**
 8. **end for**
 9. # Calculating the Score_{IFACE}
 10. **for** each (*lig_rot*, *lig_trans*) **do**
 11. Calculate Score_{IFACE} using Equation 22
 12. **end for**
 13. # Ranking
 14. Rank the Score_{IFACE} in descending order
 15. Take predictions in the top 2000 ranks
-

ZRANK

ZRANK is the re-ranking technique developed by the ZDOCK team to improve further results of the docking program ZDOCK version 2. The ZRANK program can also be retrieved from <http://zdock.umassmed.edu/software/download/> after registering for academic use.

One additional important step before performing re-ranking by ZRANK is to add the polar

Hydrogen using PDB2PQR program, which can be retrieved from <https://sites.google.com/a/poissonboltzmann.org/software/pdb2pqr/d> for local experiment or for webserver interface.

The user guide and install instruction can be found here <http://www.ics.uci.edu/~dock/pdb2pqr/userguide.html>.

INTERVOR

INTERVOR is a package to calculate and present the molecular interfaces based on the Voronoi diagram. The INTERVOR program can be obtained from <http://biskit.pasteur.fr/install/applications/intervor>.

DOWSER

DOWSER is the program to add water in the protein interfaces. Although the DOWSER project started a decade ago, it maintains very high accuracy in recovering water positions at the protein interfaces. The program and manual can be downloaded from <http://www.ks.uiuc.edu/Research/vmd/plugins/dowser/>.

AMBER 10

AMBER is a molecular dynamics simulation package (<http://ambermd.org>), which is used for the minimization stage in this thesis. However, unlike other programs above, the AMBER package is not free or open source. Hence, for the purpose of reproducing or comparing results of this thesis (if any), it is suggested that one can apply other simulation packages, which are free such as GROMACS (<http://www.gromacs.org>) or NAMD (<http://www.ks.uiuc.edu/Research/namd/>), to perform minimization of the predicted solvated complexes resulted from previous stages mentioned in the Figure 5-2.

References

- Ahmed Mostafa H., Francesca Spyraakis, Pietro Cozzini, Parijat K. Tripathi, Andrea Mozzarelli, J. Neel Scarsdale, Martin A. Safo and Glen E. Kellogg (2011). *Bound water at Protein-Protein Interfaces: partners, roles, and hydrophobic bubbles as conserved motif*. PLoS ONE **6**(9): e24712.
- Archbold Julia K., Whitney A. Macdonald, Stephanie Gras, Lauren K. Ely, John J. Miles, Melissa J. Bell, Rebekah M. Brennan, Travis Beddoe, Matthew C.J. Wilce, Craig S. Clements, Anthony W. Purcell, James McCluskey, Scott R. Burrows and Jamie Rossjohn (2009). *Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition*. J. Exp. Med **206**(1): 209-219.
- Assarsson Erika, Huynh-Hoa Bui, John Sidney, Qing Zhang, Jean Glenn, Carla Oseroff, Innocent N. Mbawuikie, Jeff Alexander, Mark J. Newman, Howard Grey and Alessandro Sette (2008). *Immunomic analysis of the repertoire of T-cell specificities for Influenza A Virus in Humans*. Journal of Virology **82**(24): 12241-12251.
- Availability of new H5N1 prototype strain for influenza pandemic vaccine development*. 2012, <http://www.who.int/influenza/vaccines/virus/2strains2006/en/>.
- Beaver John E., Philip E. Bourne and Julia V. Ponomarenko (2007). *EpitopeViewer: a Java application for the visualization and analysis of immune epitopes in the Immune Epitope Database and Analysis Resource (IEDB)*. Immunome Research **3**(3).
- Bhat T. Narayana, Graham A. Bentley, Ginette Boulot, Mark I. Greene, Diana Tello, William Dall'Acqua, Helene Souchon, Frederick P. Schwarz, Roy A. Mariuzza and Roberto J. Poljak (1994). *Bound water molecules and conformational stabilization help mediate an antigen-antibody associations*. Proc. Natl. Acad. Sci. USA **91**: 1089-1093.
- Boon Adrianus C. M., Gerrie de Mutsert, Debbie van Baarle, Derek J. Smith, Alan S. Lapedes, Ron A.M. Fouchier, Kees Sintnicolaas, Albert D.M.E Osterhaus and Guus F. Rimmelzwaan (2004). *Recognition of homo- and heterosubtypic variants of influenza A viruses by human CD8+ T lymphocytes*. The Journal of Immunology **172**: 2453-2460.
- Bourquard Thomas, Julie Bernauer, Jerome Aze and Anne Poupon (2011). *A collaborative filtering approach for Protein-Protein Docking scoring functions*. PLoS ONE **6**(4): e18541.

Burnett JC., Glen E. Kellogg and DJ. Abraham (2000). *Computational methodology for estimating changes in free energies of biomolecular association upon mutation: the importance of bound water in dimer-tetramer assembly for beta 37 mutant hemoglobins*. Biochemistry **39**: 1622-1633.

Case D. A., T. E. CheathamIII, T. Darden, H. Gohlke, R. Luo, Jr. K. M Merz, A. Onufriev, S. Simmerling, B. Wang and R. Woods (2005). *The Amber biomolecular simulation programs*. J. Computat. Chem **26**: 1668-1688.

Case David A., Thomas E. Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz Jr., Alexey Onufriev, Carlos Simmerling, Bing Wang and Robert J. Woods (2005). *The Amber Biomolecular Simulation Programs*. J Comput Chem **26**: 1668-1688.

Chang Hsiu-Ching, Alex Smolyar, Rebecca Spoerl, Torsten Witte, Yasuko Yao, Earl C. Goyarts, Stanley G. Nathenson and Ellis L. Reinherz (1997). *Topology of T cell Receptor-peptide/Class I MHC interaction defined by charge reversal complementation and functional analysis*. J. Mol. Biol **271**: 278-293.

Chaudhury Sidhartha, Sergey Lyskov and Jeffrey J. Gray (2010). *PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta*. Bioinformatics **26**(5): 689-691.

Chen Rong, Li Li and Zhiping Weng (2003). *ZDOCK: An Initial-Stage Protein-Docking Algorithm*. Proteins: Structure, Function, and Genetics **52**: 80-87.

Chen Rong and Zhiping Weng (2002). *Docking Unbound Proteins Using Shape Complementarity, Desolvation, and Electrostatics*. Proteins: Structure, Function, and Genetics **47**: 281-294.

Chen Rong and Zhiping Weng (2003). *A Novel Shape Complementarity Scoring Function for Protein-Protein Docking*. Proteins: Structure, Function, and Genetics **51**: 397-408.

Cherfils Jacqueline and Joel Janin (1993). *Protein docking algorithms: simulating molecular recognition*. Current Opinion in Structural Biology **3**: 265-269.

Chou P. Y. and G. D. Fasman (1978). *Prediction of the secondary structure of proteins from their amino acid sequence*. Adv. Enzymol. Relat. Areas. Mol. Biol. **47**: 45-48.

Chowdhury Rezaul Alam and Chandrajit Bajaj (2010). Algorithms for faster Molecular Energetics, Forces, and Interfaces. ICES Report, The Insititute for Computational Engineering and Sciences, The University of Texas at Austin: 10-32.

Chowdhury Rezaul, Muhibur Rasheed, Donald Keidel, Maysam Moussalem, Arthur Olson, Michel Sanner and Chandrajit Bajaj (2013). *Protein-Protein Docking with F2Dock 2.0 and GB-rerank*. PLoS ONE **8**(3): e51307.

Chrencik Jill E., Alexei Brooun, Michelle L. Kraus, Michael I. Recht, Anand R. Kolatkar, Gye Won Han, Jan Marcus Seifert, Hans Widmer, Manfred Auer and Peter Kuhn (2006). *Structural and biophysical characterization of the EphB4.EphrinB2 protein-protein interaction and receptor specificity*. The journal of Biological Chemistry **281**: 28185-28192.

Chuang Gwo-Yu, Dima Kozakov, Ryan Brenke, Stephen R. Comeau and Sandor Vajda (2008). *DARS (Decoys As the Reference State) Potentials for Protein-Protein Docking*. Biophysical Journal **95**: 4217-4227.

Comeau S. R. , D. W. Gatchell, S. Vajda and C. J. Camacho (2004). *ClusPro: a fully automated algorithm for protein-protein docking*. Nucleic Acids Research **32**(Web Server Issue).

Comeau S. R. , D. W. Gatchell, S. Vajda and C. J. Camacho (2004). *ClusPro: an automated docking and discrimination method for the prediction of protein complexes*. Bioinformatics **20**(1): 45-50.

Comeau S. R., D. Kozakov, R. Brenke, Y. Shen, D. Beglov and S. Vajda (2007). *ClusPro: Performance in CAPRI rounds 6-11 and the new server*. Proteins **69**: 781-785.

Rosetta - The premier software suite for macromolecular modeling. 2013, <https://http://www.rosettacommons.org/>.

Dantas Gautam, Brian Kuhlman, David Callender, Michelle Wong and David Baker (2003). *A large scale test of computational protein design: Folding and Stability of Nine Completely Redesigned Globular Proteins*. J. Mol. Biol **332**: 449-460.

DiBrino Marianne, Kenneth C. Parker, David H. Margulies, Joseph Shiloach, Richard V. Turner, William E. Biddison and John E. Coligan (1995). *Identification of the peptide binding motif for HLA-B44, one of the most common HLA-B alleles in the Caucasian Population*. Biochemistry **34**: 10130-10138.

Dijk Aalt DJ. van and Alexandre MJJ. Bonvin (2006). *Solvated docking: introducing water into the modelling of biomolecular complexes*. Bioinformatics **22**(19): 2340-2347.

Dominguez Cyril, Rolf Boelens and Alexandre M.J.J. Bonvin (2003). *HADDOCK: a protein-protein docking approach on biochemical or biophysical data*. J. Am. Chem. Soc. **125**: 1731-1737.

Ekiert Damian C., Gira Bhabha, Marc-Andre Elsliger, Robert H. E. Friesen, Mandy Jongeneelen, Mark Throsby, Jaap Goudsmit and Ian A. Wilson (2009). *Antibody Recognition of a highly conserved influenza virus epitope*. Science **324**: 246-251.

CAPRI: Critical Assessment of PRediction of Interactions. 2013, <http://www.ebi.ac.uk/msd-srv/capri/>.

Emini Emilo A., Joseph V. Hughes, Debra S. Perlow and Joshua Boger (1985). *Induction of Hepatitis A Virus-Neutralizing Antibody by a Virus-specific Synthetic Peptide*. Journal of Virology **55**(3): 836-839.

Fernandez-Recio Juan, Maxim Totrov and Ruben Abagyan (2001). *Soft protein-protein docking in internal coordinate*. Protein Science **11**: 280-291.

Franzese Giancarlo, Valentino Bianco and Svilen Iskov (2010) "Water at interface with proteins."

Gabb Henry A., Richard M. Jackson and Michael J.E. Sternberg (1997). *Modelling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information*. J. Mol. Biol **272**: 106-120.

Garten Rebecca J., C. Todd Davis, Colin A. Russell, Bo Shu, Stephen Lindstrom, Amanda Balish, Wendy M. Sessions, Xiyan Xu, Eugene Skepner, Varough Deyde, Margaret Okomo-Adhiambo, Larisa Gubareva, John Barnes, Catherine B. Smith, Shannon L. Emery, Michael J. Hillman, Pierre Rivaller, James Smagala, Miranda de Graaf, David F. Burke, Ron A. M. Fouchier, Claudia Pappas, Celia M. Alpuche-Aranda, Hugo López-Gatell, Hiram Olivera, Irma López, Christopher A. Myers, Dennis Faix, Patrick J. Blair, Cindy Yu, Kimberly M. Keene, P. David Dotson Jr., David Boxrud, Anthony R. Sambol, Syed H. Abid, Kirsten St. George, Tammy Bannerman, Amanda L. Moore, David J. Stringer, Patricia Blevins, Gail J. Demmler-Harrison, Michele Ginsberg, Paula Kriner, Steve Waterman, Sandra Smole, Hugo F. Guevara, Edward A. Belongia, Patricia A. Clark, Sara T. Beatrice, Ruben Donis, Jacqueline Katz, Lyn Finelli, Carolyn B. Bridges, Michael Shaw, Daniel B. Jernigan, Timothy M. Uyeki, Derek J. Smith, Alexander I. Klimov and Nancy J. Cox (2009). *Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans*. Science **325**: 197-201.

Garzon Jose Ignacio, Jose Ramon Lopez-Blanco, Carles Pons, Julio Kovacs, Ruben Abagyan, Juan Fernandez-Recio and Pablo Chacon (2009). *FRODOCK: a new approach for fast rotational protein-protein docking*. *Bioinformatics* **25**(19): 2544-2551.

Gatanaga Hiroyuki, Yasuhiro Suzuki, Hsinyi Tsang, Kazuhisa Yoshimura, Mark K. Kavlick, Kunio Nagashima, Robert J. Gorelick, Sek Mardy, Chun Tang, Michael F. Summers and Hiroaki Mitsuya (2002). *Amino Acid Substitutions in Gag Protein at Non-cleavage sites Are Indispensable for the Development of a high Multitude of HIV-1 Resistance against Protease Inhibitors*. *The journal of Biological Chemistry* **277**(8): 5952-5961.

Gatherer D. (2009). *The 2009 H1N1 influenza outbreak in its historical context*. *J. Clin Virol* **45**(3): 174-178.

Gras S., L. Kedzierski, SA. Valkenburg, K. Laurie, YC. Liu, JT. Denholm, MJ. Richards, GF. Rimmelzwaan, A. Kelso, PC. Doherty, SJ. Turner, J. Rossjohn and K. Kedzierska (2010). *Cross-reactive CD8+ T-cell immunity between the pandemic H1N1-2009 and H1N1-1918 influenza A viruses*. *Proc. Natl. Acad. Sci. USA* **107**(28): 12599-12604.

Gray Jeffrey J., Stewart Moughon, Chu Wang, Ora Schueler-Furman, Brian Kuhlman, Carol A. Rohl and David Baker (2003). *Protein-Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations*. *J. Mol. Biol* **331**: 281-299.

Greenbaum Jason A., Maya F. Kotturi, Yohan Kim, Carla Oseroff, Kerrie Vaughan, Nima Salimi, Randi Vita, Julia Ponomarenko, Richard H. Scheuermann, Alessandro Sette and Bjoern Peters (2009). *Pre-existing immunity against swine-origin H1N1 influenza viruses in the general human population*. *PNAS Early Edition*.

Groot Anne S. De, Matt Ardito, Elizabeth M. McClaine, Leonard Moise and William D. Martin (2009). *Immunoinformatic comparison of T-cell epitopes contained in novel swine-origin influenza A (H1N1) virus with epitopes in 2008-2009 conventional influenza vaccine*. *Vaccine* **27**: 5740-5747.

Guex N. and M.C. Peitsch (1997). *Swiss-Model and the Swiss-Pdb Viewer: An environment for comparative protein modeling*. *Electrophoresis* **18**: 2714-2723.

Gupta Shishir K., Mugdha Srivastava, Bashir A. Akhoon, Suchi Smita, Ulf Schmitz, Olaf Wolkenhauer, Julio Vera and Shailendra K. Gupta (2010). *Identification of immunogenic consensus T-cell epitopes in globally distributed influenza A H1N1 neuraminidase*. *Infection, Genetics, and Evolution* **11**(2): 308-319.

Halperin Inbal, Buyong Ma, Haim Wolfson and Ruth Nussinov (2002). *Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions*. Proteins: Structures, Functions, and Genetics **47**: 409-443.

Han M. V. and C. M. Zmasek (2009). *PhyloXML: XML for evolutionary biology and comparative genomics*. BMC Bioinformatics **10**: 356.

Hu Min, Pingwei Li, Ling Song, Philip D. Jeffrey, Tatiana A. Chenova, Keith D. Wilkinson, Robert E. Cohen and Yigong Shi (2005). *Structure and mechanisms of the proteasome-associated deubiquitinating enzyme USP14*. The EMBO journal **24**(21): 3747-3756.

Huang Kui, Michael N.G. James, Wuyuan Lu, Michael Laskowski Jr. and Stephen Anderson (1995). *Water molecules participate in proteinase-inhibitor interactions: Crystal structures of Leu18, Ala18, and Gly18 variants of turkey ovomucoid inhibitor third domain complexed with Streptomyces griseus proteinase B*. Protein Science **4**(10): 1985-1997.

Hwang Howook, Thom Vreven, Joel Janin and Zhiping Weng (2010). *Protein-protein Docking Benchmark Version 4.0*. Proteins **78**(15): 3111-3114.

Ikura Teikichi, Yoshiaki Urakubo and Nobutoshi Ito (2004). *Water-mediated interaction at a protein-protein interface*. Chemical Physics **307**: 111-119.

Jackson Richard M., Henry A. Gabb and Michael J.E. Sternberg (1998). *Rapid Refinement of Protein Interfaces Incorporating Solvation: Application to the Docking Problem*. J. Mol. Biol **276**: 265-285.

Janin Joel (1999). *Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition*. Structure: R277-R279.

Janin Joel (2005). *Assessing predictions of protein-protein interaction: the CAPRI experiment*. Protein Science **14**: 278-283.

Jiang Lin, Eric A. Althoff, Fernando R. Clemente, Lindsey Doyle, Daniela Rothlisberger, Alexandre Zanghellini, Jasmine L. Gallaher, Jamie L. Betker, Fujie Tanaka, Carlos F. Barbas III, Donald Hilvert, Kendall N. Houk, Barry L. Stoddard and David Baker (2008). *De novo computational design of retro-aldol enzymes*. Science **319**: 1387-1391.

Jiang Lin, Brian Kuhlman, Tanja Kortemme and David Baker (2005). *A "Solvated Rotamer" Approach to Modeling Water-Mediated Hydrogen Bonds at Protein-Protein Interfaces*. Proteins: Structure, Function, and Bioinformatics **58**: 893-904.

Joachimiak Andrzej, Tanja Kortemme, BL Stoddard and David Baker (2006). *Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface*. J. Mol. Biol **361**: 195-208.

Jones Susan and Janet M. Thornton (1996). *Principles of protein-protein interactions*. Proc. Natl. Acad. Sci. USA.

Kaiser P., D. Meierhofer, X. Wang and L. Huang (2008). *Tandem affinity purification combined with mass spectrometry to identify components of protein complexes*. Methods Mol Biol **439**: 309-326.

Karplus P. A. and G. E. Schulz (1985). *Prediction of chain flexibility in proteins*. Naturwissenschaften **72**(4): 212-213.

Kastritis Panagiotis L. and Alexandre M.J.J. Bonvin (2013). *On the binding affinity of macromolecular interactions: daring to ask why proteins interact*. J R Soc Interface **10**.

Kastritis Panagiotis L., Koen M. Visscher, Aalt D.J. van Dijk and Alexandre M.J.J. Bonvin (2013). *Solvated protein-protein docking using Hyte-Doolittle-based water preferences*. Proteins **81**: 510-518.

Katchalski-Katzir Ephraim, Isaac Shariv, Miriam Eisenstein, Asher A. Friesem, Claude Aflalo and Ilya A. Vakser (1992). *Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques*. Proc. Natl. Acad. Sci. USA **89**: 2195-2199.

Katoh Toh (2008). *Recent development in the MAFFT multiple sequence alignment program*. Briefings in Bioinformatics **9**: 286-298.

Kolaskar A.S. and Prasad C. Tongaonkara (1990). *A semi-empirical method for prediction of antigenic determinants on protein antigens*. FEBS Letters **276**(1-2): 172-174.

Kozakov D., R. Brenke, S. R. Comeau and S. Vajda (2006). *PIPER: An FFT-based protein docking program with pairwise potentials*. Proteins: Structure, Function, and Bioinformatics **65**: 392-406.

Kozakov Dima, Ryan Brenke, Stephen R. Comeau and Sandor Vajda (2006). *PIPER: An FFT-Based Protein Docking Program with Pairwise Potentials*. Proteins: Structure, Function, and Bioinformatics **65**: 392-496.

Kozakov Dima, David R. Hall, Dmitri Beglov, Ryan Branke, Stephen R. Comeau, Yang Shen, Keyong Li, Jiefu Zheng, Pirooz Vakili, Ioannis Ch. Paschalidis and Sandor Vajda (2010). *Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19*. Proteins **78**: 3124-3130.

Kozakov Dima, Ora Schueler-Furman and Sandor Vajda (2008). *Discrimination of near-native structures in protein-protein docking by testing the stability of local minima*. Proteins **72**: 993-1004.

Kreijtz J.H.C.M., G. de Mutsert, C.A. van Baalen, R.A.M. Fouchier, A.D.M.E. Osterhaus and G.F. Rimmelzwaan (2008). *Cross-Recognition of Avian H5N1 Influenza Virus by Human Cytotoxic T-Lymphocyte Populations Directed to Human Influenza A Virus*. Journal of Virology **82**(11): 5161-5166.

Larsen Jens Erik Pontoppidan, Ole Lund and Morten Nielsen (2006). *Improved method for predicting linear B-cell epitopes*. Immunome Research **2**(2).

Larsen M., C. Lundergard, K. Lamberth, S. Buus, S. Brunak, O. Lund and M. Nielsen (2005). *An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions*. European Journal of Immunology **35**(8): 2295-2303.

Larsen Mette, Claus Lundegaard, Kasper Lamberth, Soren Buus, Ole Lund and Morten Nielsen (2007). *Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction*. BMC Bioinformatics **8**(1): 424.

Lejon Sara, Jacob Flyvholm Cramer and Peter Nordberg (2008). *Structural basis for the binding of naproxen to human serum albumin in the presence of fatty acids and the GA module*. Acta Cryst. Sect F Struct Biol Cryst Commun **64**(Pt 2): 64-69.

Lemmon Gordon and Jens Meiler (2013). *Towards Ligand Docking including Explicit Interface Water Molecules*. PLoS ONE **8**(6): e667536.

Li Lin, Dachuan Guo, Yangyu Huang, Shiyong Liu and Yi Xiao (2011). *ASPDock: protein-protein docking algorithm using atomic solvation parameters model*. BMC Bioinformatics **12**: 36.

Loriot Sebastien and Frederic Cazals (2010). *Modeling macro-molecular interfaces with Intervor*. Bioinformatics **26**(7): 964-965.

MacBeath G. and S.L. Schreiber (2000). *Printing proteins as microarrays for high-throughput function determination*. Science **289**: 1760-1763.

Macdonald Whitney A., Anthony W. Purcell, Nicole A. Mifsud, Lauren K. Ely, David S. Williams, Linus Chang, Jeffrey J. Gorman, Craig S. Clements, Lars Kjer-Nielsen, David M. Koelle, Scott R. Burrows, Brian D. Tait, Rhonda Holdsworth, Andrew G. Brooks, George O. Lovrecz, Louis Lu, Jamie Rossjohn and James McCluskey (2003). *A Naturally Selected Dimorphism within the HLA-B44 Supertype Alters Class I Structure, Peptide Repertoire, and T Cell Recognition*. The Journal of Experimental Medicine **198**(5): 679-691.

Maguire Michael F., Rosario Guinea, Philip Griffin, Sarah Macmanus, Robert C. Elston, Josie Wolfram, Naomi Richards, Mary H. Hanlon, David JT Porter, Terri Wrin, Neil Parkin, Margaret Tisdale, Eric Furfine, Chris Petropoulos, B. Wendy Snowden and Jorg-Peter Kleim (2002). *Changes in Human Immunodeficiency Virus Type 1 Gag at Positions L449 and P453 Are Linked to I50V Protease Mutants In Vivo and Cause Reduction of Sensitivity to Amprenavir and Improved Viral Fitness In Vitro*. Journal of Virology **76**(15): 7398-7406.

Mandell Jeffrey G., Victoria A. Roberts, Michael E. Pique, Vladimir Kotlovyyi, Julie C. Mitchell, Erik Nelson, Igor Tsigelny and Lynn F. Ten Eyck (2001). *Protein docking using continuum electrostatics and geometric fit*. Protein Engineering **14**(2): 105-113.

Mintseris J., B. Pierce, K. Wiehe, R. Anderson, R. Chen and Z. Weng (2007). *Integrating Statistical Pair Potentials into Protein Complex Prediction*. Proteins: Structure, Function, and Bioinformatics **69**(3): 511-520.

Mintseris Julian, Kevin Wiehe, Brian Pierce, Robert Anderson, Rong Chen, Joel Janin and Zhiping Weng (2005). *Protein-protein Docking Benchmark 2.0: An Update*. Proteins: Structure, Function, and Bioinformatics **60**: 214-216.

Miyazawa Sanzo and Robert L. Jernigan (1985). *Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation*. Macromolecules **18**: 534-552.

Nam Tae-Wook, Ha Il Jung, Young Jun An, Young-Ha Park, Sang Hee Lee, Yeong-Jae Seok and Sun-Shin Cha (2008). *Analyses of Mlc-IIBGlc interaction and a plausible molecular mechanism of Mlc inactivation by membrane sequestration*. Proc. Natl. Acad. Sci. USA **105**(10): 3751-3756.

NCBI Influenza sequence database.
<http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi>.

Neumann Gabriele, Takeshi Noda and Yoshihiro Kawaoka (2009). *Emergence and pandemic potential of swine-origin H1N1 influenza virus*. Nature **459**(18).

Noy Efrat and Amiram Goldblum (2010). *Flexible Protein-Protein Docking based on Best-First Search Algorithm*. Journal of Computational Chemistry **31**(9): 1929-1943.

Parker J.M., D. Guo and R.S. Hodges (1986). *New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites*. Biochemistry **25**(19): 5425-5432.

Petukhov Michael, David Cregut, Claudio M. Soares and Luis Serrano (1999). *Local water bridges and protein conformational stability*. Protein Science **8**: 1982-1989.

Pierce Brian G., Yuichiro Hourai and Zhiping Weng (2011). *Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library*. PLOS ONE **6**(9): e24657.

Pierce Brian, AT. Phillips and Zhiping Weng (2007). *Structure Prediction of Protein Complexes*. Computational Methods for Protein Structure Prediction and Modeling, New York: Springer. Volume 2: Structure Prediction: 109-134.

Pierce Brian and Zhiping Weng (2007). *ZRANK: Reranking Protein Docking Predictions with an optimized energy function*. Proteins: Structure, Function, and Bioinformatics **67**: 1078-1086.

Pierce Brian and Zhiping Weng (2008). *A combination of rescoring and refinement significantly improves protein docking performance*. Proteins: Structure, Function, and Bioinformatics **72**: 270-279.

Pierce Brian and Zhiping Weng (2008). *A combination of rescoring and refinement significantly improves protein docking performance*. Proteins: Structure, Function, and Bioinformatics **72**: 270-279.

Ponder Jay W. and David A. Case (2003). *Force fields for Protein Simulation*. Advances in Protein Chemistry **66**.

Reinherz Ellis L., Kemin Tan, Lei Tang, Petra Kern, Jin-huan Liu, Yi Xiong, Rebecca E. Hussey, Alex Smolyar, Brian Hare, Rongguang Zhang, Andrzej Joachimiak, Hsiu-Ching Chang, Gerhard Wagner and Jia-huai Wang (1999). *The Crystal Structure of a T Cell Receptor in Complex with Peptide and MHC Class II*. Science **286**: 1913-1921.

Ritchie David W. (2008). *Recent Progress and Future Directions in Protein-Protein Docking*. Current Protein and Peptide Science **9**: 1-15.

Ritchie David W., Dima Kozakov and Sandor Vajda (2008). *Accelerating and Focusing Protein-Protein Docking Correlations Using Multi-Dimensional Rotational FFT Generating Functions*. Bioinformatics **24**: 1865-1873.

Rodwell Victor W. and Peter J. Kennelly (2003). *Proteins: Determination of Primary Structure*. Harper's Illustrated Biochemistry, The McGraw-Hill Companies.

Roy Ambrish, Alper Kucukural and Yang Zhang (2010). *I-TASSER: a unified platform for automated protein structure and function prediction*. Nature Protocols **5**(4).

Rudolph Markus G., Robyn L. Stanfield and Ian A. Wilson (2006). *How TCRs bind MHCs, Peptides, and Coreceptors*. Annu. Rev. Immunol. **24**: 419-466.

Russell Rupert J., Philip S. Kerry, David J. Stevens, David A. Steinhauer, Stephen R. Martin, Steven J. Gamblin and John J. Skehel (2008). *Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion*. PNAS **105**(46): 17736-17741.

Schnedman-Duhovny Dina, Ruth Nussinov and Haim J. Wolfson (2006). *Prediction Molecular Interaction in silico: Protein-Protein and Protein-Drug Docking*. Frontier in Medicinal Chemistry **3**: 585-613.

Schreiber G. and AR. Fersht (1995). *Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles*. J. Mol. Biol **248**: 478-486.

Selin Liisa K. and Raymond M. Welsh (2004). *Plasticity of T Cell Memory Responses to Viruses*. Immunity **20**: 5-16.

CAPRI report. 2013, <http://web.mit.edu/sheny/capri.html>.

Shoemaker BA. and AR. Panchenko (2007). *Deciphering protein-protein interactions. Part I. Experimental techniques and databases*. PLOS Computational Biology **3**(3): e42.

Shoemaker Benjamin A. and Anna R. Panchenko (2007). *Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners*. PLOS Computational Biology **3**(4): e43.

Smith Gavin J. D., Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J. Lycett, Michael Worobey, Oliver G. Pybus, Siu Kit Ma, Chung Lam Cheung, Jayna Raghvani, Samir Bhatt, J. S. Malik Peiris, Yi Guan and Andrew Rambaut (2009). *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic*. Nature **459**(25).

Smith GP (1985). *Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface*. Science **228**: 1315-1317.

Smith Graham R. and Michael JE Sternberg (2002). *Prediction of protein-protein interactions by docking methods*. Current Opinion in Structural Biology **12**: 28-35.

Solernou Albert and Juan Fernandez-Recio (2010). *Protein docking by Rotation-based Uniform Sampling (RotBUS) with fast computing of intermolecular contact distance and residue desolvation*. BMC Bioinformatics **11**.

Solernou Albert and Juan Fernandez-Recio (2011). *pyDockCG: New Coarse-Grained Potential for Protein-Protein Docking*. The journal of Physical Chemistry B **115**: 6032-6039.

Sprinzak E., S. Sattath and H. Margalit (2003). *How reliable are experimental protein-protein interaction data?* J Mol Biol **327**(5): 919-923.

Subbramanian RA., S. Basha, MT. Shata, RC. Brady and DI. Bernstein (2010). *Pandemic and seasonal H1N1 influenza hemagglutinin-specific T-cell responses elicited by seasonal influenza vaccination*. Vaccine **28**(52): 8258-8267.

Tamiya Sadahiro, Sek Mardy, Mark F. Kavlick, Kazuhisa Yoshimura and Hiroaki Mitsuya (2004). *Amino acid insertions near Gag cleavage sites restore the otherwise compromised replication of human immunodeficiency virus type 1 variants resistant to Protease Inhibitors*. J Virol **78**(21): 12030-12040.

Taylor Jeffrey S. and Roger M. Burnett (2000). *DARWIN: A Program for Docking Flexible Molecules*. Proteins: Structure, Function, and Genetics **41**: 173-191.

Tsui V. and D. A. Case (2001). *Theory and applications of the generalized Born solvation model in macromolecular simulations*. Biopolymers **56**(4): 275-291.

Vakser Ilya A. (1997). *Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex*. Proteins(Suppl1): 226-230.

Vogt Christopher C. and Garrison W. Cottrell (1999). *Fusion via a linear combination of scores*. Information Retrieval **1**: 151-173.

Vreven Thom, Howook Hwang and Zhiping Weng (2011). *Integrating atom-based and residue-based scoring functions for protein-protein docking*. Protein Science **20**: 1576-1586.

Vries Sjoerd J de, Marc van Dijk and Alexander MJJ Bonvin (2010). *The HADDOCK web server for data-driven biomolecular docking*. Nature Protocols **5**(5): 883-897.

Wahl Angela, William McCoy, Fredda Schafer, Wilfried Bardet, Rico Buchli, Daved H. Fremont and William H. Hildebrand (2009). *T-Cell Tolerance for Variability in an HLA Class I-Presented Influenza A Virus Epitope*. Journal of Virology **83**(18): 9206-9214.

Walhout Albertha JM and Marc Vidal (2001). *High Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping*. Methods **24**: 297-306.

Wang Chu, Philip Bradley and David Baker (2007). *Protein-protein docking with backbone flexibility*. J Mol Biol **373**(2): 503-519.

Wang Jia-huai and Ellis L. Reinherz (2001). *Structural basis of T cell recognition of peptides bound to MHC molecules*. Molecular Immunology **38**: 1039-1049.

Wang Jia-huai, Alex Smolyar, Kermin Tan, Jin-huan Liu, Mikyung Kim, Zhen-yu J Sun, Gerhard Wagner and Ellis L. Reinherz (1999). *Structure of a heterophilic adhesion complex between the human CD2 and CD58 (LFA-3_ counterreceptors*. Cell **97**(6): 791-803.

Wei Chin-Jen, Jeffrey C. Boyington, Kaifan Dai, Katherine V. Houser, Melissa B. Pearce, Wing-Pui Kong, Zhi-yong Yang, Terrence M. Tumpey and Gary J. Nabel (2010). *Cross-neutralization of 1918 and 2009 influenza viruses: Role of Glycans in viral evolution and Vaccine design*. Sci Transl Med **2**(24ra21).

Accessibility. 2013, http://swift.cmbi.ru.nl/teach/theory/Theory_3.html.

Wodak Shoshana J. and Joel Janin (1978). *Computer Analysis of Protein-Protein Interaction*. J. Mol. Biol **124**: 323-342.

Wu Shengli, Yaxin Bi, Xiaoqin Zeng and Lixin Han (2009). *Assigning appropriate weights for the linear combination data fusion method in information retrieval*. Information Processing and Management **45**(4): 413-426.

Xu Dong, Chung-Jung Tsai and Ruth Nussinov (1997). *Hydrogen bonds and salt bridges across protein-protein interfaces*. Protein Engineering **10**(9): 999-1012.

Xu R., D. C. Ekiert, J. C. Krause, R. Hai, J. E. Crowe Jr. and I. A. Wilson (2010). *Structural Basis of Preexisting Immunity to the 2009 H1N1 Pandemic Influenza Virus*. Science **328**(357).

Zhang C., G. Vasmatzis, J.L. Cornette and C. DeLisi (1997). *Determination of atomic desolvation energies from the structures of crystallized proteins*. J Mol Biol **267**: 707-726.

Zhang Li and Jan Hermans (1996). *Hydrophilicity of Cavities in Proteins*. Proteins: Structure, Function, and Bioinformatics **24**: 433-438.

Zhang Yang (2007). *Template-based modeling and free modeling by I-TASSER in CASP7*. Proteins **69**: 108-117.

Zhang Yang (2008). *I-TASSER server for protein 3D structure prediction*. BMC Bioinformatics **9**(40).

Zhong Weimin, Surjit B. Dixit, Robert J. Mallis, Haribabu Arthanari, Alexey A. Lugovskoy, David L. Beveridge, Gerhard Wagner and Ellis L. Reinherz (2007). *CTL Recognition of a Protective Immunodominant Influenza A Virus Nucleoprotein Epitope Utilizes a Highly Restricted V β but Diverse V α Repertoire: Functional and Structural Implications*. J. Mol. Biol **372**: 535-548.