PROTEINS: Structure, Function, and Genetics 52:118–122 (2003)

Protein–Protein Docking Predictions for the CAPRI Experiment

Jeffrey J. Gray,* Stewart E. Moughon, Tanja Kortemme, Ora Schueler-Furman, Kira M.S. Misura, Alexandre V. Morozov, and David Baker

Howard Hughes Medical Institute and Department of Biochemistry, University of Washington, Seattle, Washington

ABSTRACT **We predicted structures for all seven targets in the CAPRI experiment using a new method in development at the time of the challenge. The technique includes a low-resolution rigid body Monte Carlo search followed by high-resolution refinement with side-chain conformational changes and rigid body minimization. Decoys ($\sim 10^6$ per target) were discriminated using a scoring function including van der Waals and solvation interactions, hydrogen bonding, residue–residue pair statistics, and rotamer probabilities. Decoys were ranked, clustered, manually inspected, and selected. The top ranked model for target 6 predicted the experimental structure to 1.5 Å RMSD and included 48 of 65 correct residue–residue contacts. Target 7 was predicted at 5.3 Å RMSD with 22 of 37 correct residue–residue contacts using a homology model from a known complex structure. Using a preliminary version of the protocol in round 1, target 1 was predicted within 8.8 Å although few contacts were correct. For targets 2 and 3, the interface locations and a small fraction of the contacts were correctly identified. Proteins 2003;52:118–122.**

Key words: protein binding; protein interactions; biomolecular free energy function; high-resolution refinement; flexible side chains

INTRODUCTION

Protein–protein interactions are vital to almost all cellular processes. Indeed, protein interactions underlie signaling, regulation, immunogenic recognition, and other important biochemical functions. Thus, the ability to model the docking of two proteins is fundamental to the understanding of the operation of biochemical systems. The Critical Assessment of Predicted Interactions (CAPRI) challenge is an excellent opportunity to evaluate the capability of docking algorithms. It also spurs the development of new algorithms, as is the case for our work. We had just begun to study the protein docking problem when the organizational meeting in Charleston was held in 2000.[1] As the targets for CAPRI were released, we developed new code to address concerns specific to the targets, and we enriched our general method. As a result, our CAPRI predictions represent "snapshots" of the ability of our developing algorithm.

The next section briefly describes our computational approach and then we detail the assumptions, manipulations, and results for each of the seven targets.

MATERIALS AND METHODS

A manuscript describing our complete algorithm and its performance on a large benchmark set is forthcoming; therefore, only an overview of our method is presented here. Figure 1 shows the general algorithm used at the time of the CAPRI experiment. Creation of a decoy begins with a random orientation of each partner and a translation of the ligand along the line of protein centers to create glancing contact between the partners. In the first stage of the algorithm, we use a rigid body Monte Carlo search, translating and rotating the ligand around the surface of the stationary receptor. The low-resolution, residue-scale interaction potentials[2] include residue environment and residue–residue interaction terms derived from a database of interfaces, a contact score to reward contacting residues, a bump score to penalize overlapping residues, and constraint scores if any knowledge is known about a particular target. All scores at this stage are formulated for a reduced representation of the amino acids based on side-chain centroid positions. After the low-resolution search, explicit side chains are added to the protein backbones using a backbone-dependent rotamer packing algorithm.[3,4] For CAPRI round 2, an explicit minimization step optimizes the rigid body displacement after the side-chain packing step. In the refinement steps (packing and minimization) and for decoy discrimination, the full-atom scoring function[3] includes van der Waals interactions with the repulsive part of the potential partially replaced with a linear term to avoid singularities; solvation using a pairwise Gaussian solvent-exclusion model;[5] hydrogen-bonding energies using a function derived from structural statistics;[6] residue–residue pair interactions derived statistically from a database of protein structures; a rotamer probability term; and a surface area and atomic solvation term (for decoy discrimination only due to the expense of calculation).[7] Although the weights of most of the terms in the scoring function are of the same order of magnitude, the dominant contributions to discrimination seem to be the

*Correspondence to: Jeffrey J. Gray, Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 N. Charles, Baltimore, MD 21218.

Received 30 October 2002; Accepted 12 November 2002

packing (van der Waals) interactions, followed by the hydrogen bonding.

The search procedure was repeated to create ~10^6 decoys for each target. The 200 best scoring decoys were then clustered on the basis of pairwise root-mean-square distance (RMSD) using a hierarchical clustering algorithm.[8] Structures within a 2.5 or 5 Å clustering threshold were designated as a set, and the lowest scoring decoy within the set represented that position. In general, the largest clusters were selected and submitted to CAPRI evaluators.

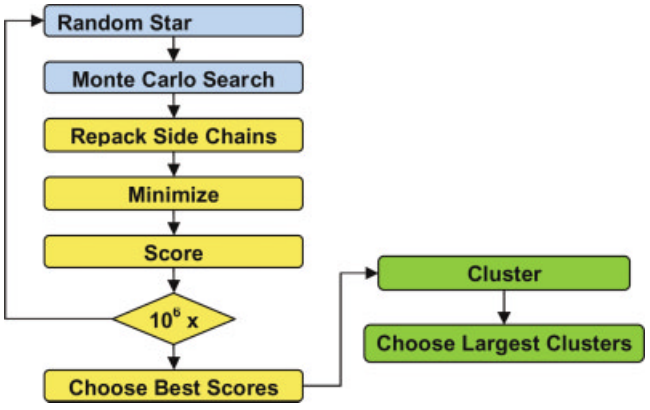

Fig. 1. Docking protocol. Blue backgrounds indicate low-resolution steps that use residue-scale potential functions, and yellow backgrounds indicate high-resolution steps that use atom-scale potential functions. The minimize step was only used on round 2 targets.

We tailored our standard strategy in several ways for the CAPRI experiment. For antibodies, we incorporated knowledge about the usual antibody-binding interface into our scoring function. To determine which antibody residues may contact the antigen, we aligned a set of 55 known antibody-antigen complex structures and created a profile of the likelihood of each antibody residue position to contact the antigen (Fig. 2). We then used this information to guide the low-resolution Monte Carlo search by adding a bonus score for each favored residue at the interface, a penalty for residues in positions never observed to be at the interface, and a neutral score for residue positions that rarely appeared at the interfaces. In practice, this score encouraged the interface to involve the complementarity determining regions (CDRs) of the antibody. Finally, before clustering we rejected decoys with interfaces that contained multiple antibody residues that never occurred in interfaces in our calibration set of complexes.

To choose some models for the antibody targets of round 1, an additional filter was used before clustering to select decoys with interfaces composed of 15 or fewer antibody residues (most antibody-antigen complexes have interfaces with <20 antibody residues). This filter eliminated many incorrect decoys that achieved a top score by creating an unreasonably large number of nonspecific contacts, and we hoped it would preserve biologically correct inter-

Round 1

```
T-02 L   diqmtqspsslsaslgqrvsltcrasq--eiSgylswlqqkpdgtikrliyaastldsgvpkrf
T-03 L   qivltqspaimsaspgekvtltcsasstitss-flywyqqkpgsspklwiYstsnLaSgVparf
profile  ·~·····························+~++++++·~···~·~·~~~·~~···++·~+~~++~~~··

T-02 L   sgtrsgsdysltisslesedfadyyclqyA--sYpFtfgggtkleikra
T-03 L   sgsgsgtsysltissleaedgasyfchqwe--tfprtfgggtkleiar-
profile  ·····················~~·········++~++++·············

T-02 H   evqlqesgpevvrp-gvsvkisckgsgyTfTD-yAmhwvkqshakslewigvISpYNgntnynq
T-03 H   evhlqesgpglvkp-sqslsltcyvtGYSiTSgYYwtwirqfpgnklewmgyis-Ydgstnynp
profile  ~+·····························~+++~++++·~·············+~++++++++~~~

T-02 H   kfkgkatmtvdkssstaymelarltsedsaiyycg--llL----Fpywgqgttvtvssa-
T-03 H   slknrisitrdtsknqfflklnsvtaedtasyycaafYyDYDFfFdywgqgttltvsaak
profile  ~·~~··~····~~·~·····················+++++++·~~~·~···········
```

Round 2

```
T-04 VHH -vqlvesgggsvpaggslrlscaasgntlctydmswYrrapgkgrDFvsgiDn-DgTTTyVdsv
T-05 VHH qvqlvasgggsvqaggslrlscaastfss--YpmgwYrqapgkECELsaRiFs-dgsaNYADsv
T-06 VHH QVqlvesgggsvqaggslslscaasTY---tstvgwfrqapgkeregvaaiYRRTgYtysadsv
profile  ~~~~~·····················~~~~++++++++~~~·····~~~~++++++~~·····

T-04 VHH agrftisqgnakntaylqmnslkpddtamyyckpSLRYGLpgCp------Iipwgqgtqvtvs
T-05 VHH kgrftisrdnaantaylqmdslkpedtavyycaagPGSgKlvvagrtCYgpNYwgqgtqvtv-
T-06 VHH KgrftlsqdnnkntvylqmnslapedtgiyycatgNsVRlAswe-----gyFywgqgtqvtv-
profile  ·····~~~~~~~~~~·····················~~~++++++++++++++++++~~~·····
```

Fig. 2. Antibody sequences with the expected contact profiles. In the profiles, + indicates a position expected to be at the interface, − indicates a position expected not to be at the interface, and ~ represents a position without an imposed preference. In the sequences, bold capital letters represent interface positions in the experimental structure, with blue lettering for residues in favored positions and red lettering for residues in unexpected positions. Yellow shading denotes interface residues in the predicted structure (target 2, model 6; target 3, model 7; targets 4, 5, and 6, model 1). Interface residues in the experimental and predicted structures are defined by intermolecular atom–atom distances of <4 Å.

faces. In round 2, large, nonspecific interfaces were avoided by adjusting the shape of the repulsive part of the van der Waals potential[9] and increasing the weight of this term in the full-atom scoring.

In round 1, some submitted decoys were chosen with an alternate selection and scoring procedure[10] that repacked side chains using a larger rotamer set (varying all χ angles) and discriminated strictly on hydrogen-bonding, solvation, and van der Waals interactions *across* the interface. In round 2, the larger rotamer set was incorporated into the standard docking tool.

To select the final models for submission, the largest clusters and best scoring decoys were examined manually. Features considered included specific contacts (i.e., close contacts, hydrogen bonds, or hydrophobic packing), chemical environment (exposed hydrophobic groups or buried polar groups), overall fit (size and shape of interface or the presence of voids at the interface), and general arrangement (the number of CDR loops interacting with the antigen). In this manner, some high-ranking decoys were removed from the list, and some intermediate-ranked decoys were placed among the final submissions. In all cases, however, model 1 was designated for the lowest scoring member of the largest cluster of structures.

Calculations were performed on clusters of ~50-processor Linux workstations with clock speeds near 1 GHz. Complete processing required between 1 and 10 days of cluster time for each target.

TARGETS AND PREDICTIONS

Target 1, HPr + HPr Kinase

Biochemical information played a significant role in our treatment of target 1. HPr kinase catalyzes the phosphorylation of Ser46 in HPr,[11] and Ser157 on the kinase is part of the P-loop (Walker A motif), which has been shown to bind the phosphate in ATP.[12] Therefore, we constrained our Monte Carlo search to configurations with a distance of 14 Å between Ser46 of the HPr and Ser157 of the kinase. Next, in our manual selection of models for this target, we preferentially selected structures that would allow each kinase protomer of the oligomeric assembly to bind and phosphorylate an HPr molecule. For example, several top ranked structures placed the HPr in the unlikely cavity at the axis of the kinase trimer; only one such structure was submitted as a model. Removal of the kinase residues 236–240 in a partially resolved loop helped the algorithm create structures allowing one-to-one interactions of HPr with each kinase protomer. Finally, an evolutionary trace calculation[13,14] detected conserved surface residues that might be likely to occur at the interface. After the automated procedure created a list of top clusters, the evolutionary information was considered to hand-select submissions that had contacts on a conserved patch of the HP-kinase. However, many of these conserved residues appeared on the terminal helix of the kinase, distant from the phosphorylation site. It was difficult to find decoys that satisfied the phosphorylation distance constraint and simultaneously included these conserved contacts, and indeed, in retrospect, the experimental structure revealed that the terminal kinase helix actually moves on docking.

Nevertheless, the predicted model 8 is a close solution. This model was from the 6th largest cluster and ranked 12th overall by our alternate method of repacking side chains with extra rotamers and scoring hydrogen bonds, solvation, and van der Waals interactions across the interface. Under manual inspection, model 8 seemed to contain fewer close contacts than other top ranked decoys, and it included several of the conserved interface residues on the hydrophobic patch of the kinase. In fact, the interface patches of both partners are correctly predicted, and the prediction superimposes on the experimental structure with a 55° rotation and 2.6 Å translation, for an RMSD of 8.8 Å. Unfortunately, the model submitted did not include the side-chain rotamers from the alternate program; thus, the number of correct atomic contacts in the submitted model 8 is small.

Target 2, Rotavirus Capsid Protein + Antibody

On the basis of the superstructure of the rotavirus and its capsid proteins,[15] we restricted our search to the outer cap of the rotavirus capsid protein, which is likely to be accessible to antibodies in vivo. In addition, we required our decoys to contain contacts with at least one of the epitope residues 172, 305, 306, or 315 (chains A, B, or C of the antigen). Our model 6, chosen with the alternate method with extra rotamer packing and intermolecular scoring, correctly predicted the general interface patches and 16 of 52 interface contacts. This model was from the third largest cluster and was individually the 16th rank decoy from the set of top decoys by intermolecular score with supplemental rotamers, and it represented the third largest cluster in a set of top decoys scored both by the standard program and the alternate program. The antibody light chain of model 6 makes very few contacts with the antigen, but the model met the other, general manual selection criteria.

Target 3, Hemagglutinin + Antibody

On the basis of mutagenesis information, we required our decoys to contain contacts with at least one of the epitope residues of hemagglutinin identified by Wiley and Skehel.[16] In addition, because of the large size of hemagglutinin, we examined the antigen in parts, separately searching the monomer or the head region of the trimer. For this target in particular, our algorithm tended to create large, nonspecific interfaces by burying the antibody inside a concave region of the hemagglutinin. To overcome this, we capped the score obtainable for contacts during our low-resolution search, and we chose some models by first filtering out interfaces containing >15 antibody residues. Our models 3, 7, and 10 correctly predicted a significant fraction of the binding patches on both the antibody and antigen, and models 7 and 10 predicted 8 and 6 of the 63 specific contacts, respectively. Each of these models was selected using the alternative method with additional rotamers and intermolecular scoring. Model 7 represented the second largest cluster and ranked 11th overall by the alternate scoring method, and it was in the largest cluster using a filter combining the best ranks from both standard and alternate program scores. Model 3 was the top ranked

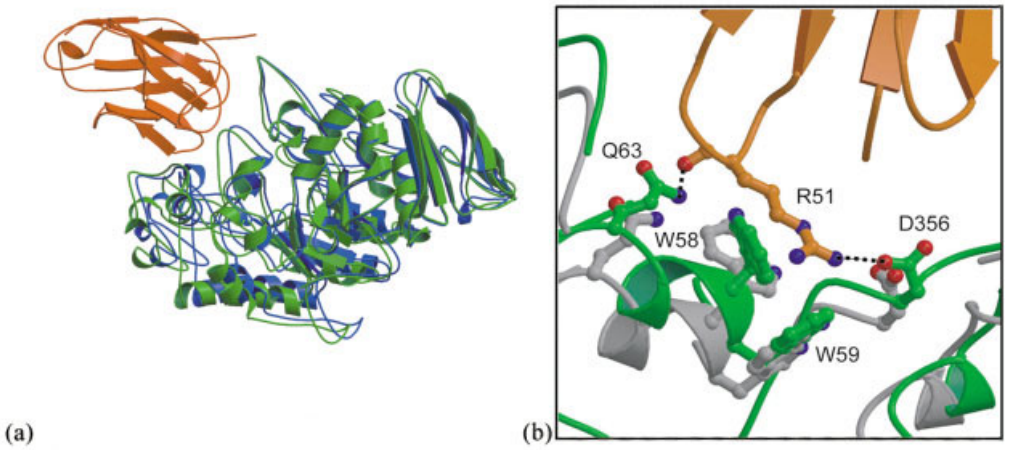

Fig. 3. Target 6 prediction. **a:** Native (blue) and model #1 (green) α-amylase superimposed with the given camelid antibody (orange). **b:** Detail of the side-chain packing (experimental structure in gray; hydrogen bonds indicated by dotted lines.)

decoy by the intermolecular score. Model 10 passed the filter for small antibody interfaces and would not have been ranked without that filter; models 7 and 3 earned their rankings without any filter for interface size.

Targets 4, 5, and 6, α-Amylase + Camelid Antibodies (Round 2)

Before the CAPRI challenge, there were only four nonredundant published structures of camelid antibodies with their antigens. Although perhaps this should have encouraged us to make no assumptions about the behavior of these unusual biomolecules, we instead aligned their sequences to our profile created from a database of mostly typical, two-chain antibodies to classify the residue positions for antibody scoring and filtering. We classified the unique additional residues in the CDR3 region as preferentially interfacial residues, and we manually smoothed the profiles relative to those used in round 1 to allow for additional uncertainty. It is surprising that the experimental structures for targets 4 and 5 were dramatically different from previously known antibody structures, making multiple contacts with framework (non-CDR) residues (Fig. 2). By filtering out decoys that had two or more interface residues in positions that never contacted an antigen in our calibration set, we eliminated any correct decoys for targets 4 and 5. Therefore, we submitted no good predictions for these targets.

For target 6, model 1 is an excellent prediction (Fig. 3). This model was the top ranked decoy and also represented the largest cluster of top scoring decoys. In addition, this model also ranked first in the alternate program intermolecular score. Model 1 contains 48 of 65 specific contacts, 25 of 29 correct interface residues on the antibody, and 35 of 37 correct interface residues on the antigen. The model can be superimposed on the experimental structure through an α-amylase translation of 1.4 Å and rotation of 3°, for an RMSD of 1.5 Å between the prediction and experimental structure. Figure 3(b) presents a detail of the side chains at the interface. The position of Arg51 (all residue numbering follows the deposited PDB record 1KXQ) on the camelid antibody was given, but the side chains on the α-amylase were packed by our algorithm. The rotamers are correctly chosen for Gln63 and incorrectly chosen for Asp356, but both residues make good hydrogen bonds with Arg51. The tryptophan residues at positions 58 and 59 create a hydrophobic pocket around the Arg51, with the rotamer for Trp59 matching that of the experimental structure.

Target 7, Streptococcal Pyrogenic Exotoxin A + T Cell Receptor β-Chain

Because streptococcal pyrogenic exotoxin A is a superantigen that is known to disrupt the normal function of the T-cell receptor, we examined other superantigen structures for clues about the interaction.[17] Model 1 was created using the structural alignment program Mammoth[18] to align the given protein components to the experimental structure of staphylococcal enterotoxin B and T cell receptor β-chain[19] (PDB code 1SBB, chains C and D). The resulting model has excellent shape complementarity and several good hydrogen bonds across the interface. It correctly contains 22 of 37 contacts and the ligand is translated 3.6 Å and rotated 11° from the native. We attempted to refine this interface using our protocol, but none of the refined interfaces were as correct as the simple structural alignment.

DISCUSSION

The best predictions for each target are summarized in Table I. Overall, the methods are satisfactory, especially considering that the protocol was still in development at the time of the CAPRI challenge. In round 1, correct binding patches were identified for each target, although residue–residue contacts were not predicted well. Although the protocol created a sufficient diversity of decoys to include some structures close to native, the best predictions were not ranked highly. The alternate program[10] with additional side-chain rotamers and intermolecular scoring was crucial in identifying our best models. In round 2, the main algorithm was significantly improved

TABLE I. Summary of Best Predictions[†]

		Contacts	Interface1	Interface2	Distance (Å)	Angle	RMSD (Å)
Round 1:	T1-model 8	2/52	13/26	16/25	2.6	55°	8.8
	T2-model 6	16/52	21/27	23/27	9.7	52°	16.6
	T3-model 7	8/63	15/33	27/34	6.4	158°	30.3
Round 2:	T4			Bad CDR assumption			
	T5			Bad CDR assumption			
	T6-model 1	48/65	25/29	35/37	1.34	3°	1.5
	T7-model 1	22/37	19/21	15/17	3.6	11°	5.3

[†]Contacts is the number of correct residue–residue contacts across the interface compared to that of the experimental structure; Interface 1 and 2 represent the correct interface residues of each docking partner compared to that of the experimental structure; Distance and angle represent the translation distance and rigid body rotation angle needed to superimpose the experimental and predicted ligand; and RMSD is the root-mean square distance between the C_α atoms of the predicted and experimental ligand, in the fixed coordinate space of the superimposed receptors.

with the addition of the minimization step, the incorporation of the expanded rotamer set, and adjustments to the scoring functions. The improved program was proven by its success with target 6. The contribution of human interventions in round 2, however, is mixed. For target 7, background research directly led to a close answer via homology modeling, but in the cases of targets 4 and 5, an incorrect assumption prevented any chance of obtaining a relevant prediction.

Several areas of modeling still need to be addressed. Clearly, target 1 shows the need for algorithms that can identify and model flexible protein backbones. The ability to compensate for side-chain or backbone conformational change on binding in antibodies has not been tested here, because bound forms of the antibodies were provided. Finally, nature presents a wide range of interfaces of varying character,[20] and more tests of the algorithm are clearly required to certify a general docking tool. Still, in general the present results are encouraging. Round 1 predictions were satisfactory, and round 2 predictions, with the exception of targets where incorrect assumptions were made, show positive accomplishments. We look forward to further testing of docking algorithms in future CAPRI challenges.

ACKNOWLEDGMENTS

We gratefully acknowledge the efforts of the organizers and evaluators of the CAPRI challenge, and we thank the experimental contributors for allowing the theoretical community to prove once again the value of experimentalists. J.J.G. was supported by a National Institutes of Health K01 Mentored Quantitative Research Fellowship in Genomics, T.K. received long-term funding from the European Molecular Biology Organization and the Human Frontier Science Program Organization (grant LT00358/2000-M), K.M.S.M. was supported by the Helen Hay Whitney Foundation, and O.S.-F. was supported by a Damon-Runyon Fellowship (DRG-1704-02) from the Damon Runyon Cancer Research Foundation. Additional funding was provided by the National Institutes of Health. The supercomputer system administration by Keith Laidig of Formix, Inc. was crucial to the success of this endeavor.

REFERENCES

1. Vajda S, Vakser IA, Sternberg MJ, Janin J. Modeling of protein interactions in genomes. Proteins 2002;47:444–446.
2. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
3. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97: 10383–10388.
4. Dunbrack RL Jr, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
5. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
6. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein complexes. J Mol Biol 2003; 326:1239–1259.
7. Koehl P, Delarue M. Polar and nonpolar atomic environments in the protein core: implications for folding and binding. Proteins 1994;20:264–278.
8. Ihaka R, Gentleman RR. A language for data analysis and graphics. J Comp Graph Stat 1996;5:299–314.
9. Misura K, Wedemeyer B, Baker D. Unpublished results (work in progress). 2002.
10. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci USA 2002;99: 14116–14121. (www.pnas.org/cgi/doi/10.1073/pnas.202485799).
11. Fieulaine S, Morera S, Poncet S, Monedero V, Gueguen-Chaignon V, Galinier A, Janin J, Deutscher J, Nessler S. X-ray structure of HPr kinase: a bacterial protein kinase with a P-loop nucleotide-binding domain. EMBO J 2001;20:3917–3927.
12. Walker JE, Saraste M, Runswick MJ, Gay NJ. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J 1982;1:945–951.
13. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
14. Shi J, Innis CA, Chen L, Blundell TL. Evolutionary Trace Server (TraceSuite II). www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html, accessed September 2001.
15. Mathieu M, Petitpas I, Navaza J, Lepault J, Kohli E, Pothier P, Prasad BVV, Cohen J, Rey FA. Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. EMBO J 2001;20:1485–1497.
16. Wiley DC, Skehel JJ. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu Rev Biochem 1987;56:365–394.
17. Li H, Llera A, Malchiodi EL, Mariuzza RA. The structural basis of T cell activation by superantigens. Annu Rev Immunol 1999;17:435–466.
18. Ortiz AR, Strass CEM, Olmea O. MAMMOTH (Matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
19. Li H, Llera A, Tsuchiya D, Leder L, Ysern X, Schlievert PM, Karjalainen K, Mariuzza RA. Three-dimensional structure of the complex between a T cell receptor beta chain and the superantigen staphylococcal enterotoxin B. Immunity 1998;9:807–816.
20. Larsen TA, Olson AJ, Goodsell DS. Morphology of protein-protein interfaces. Structure 1998;6:421–427.