SAMPLING GEOMETRIES OF PROTEIN-PROTEIN COMPLEXES

AYSAM GUERLER
guerler@chemie.fu-berlin.de

STEPHAN LORENZEN
lorenzen@chemie.fu-berlin.de

FLORIAN KRULL
fkrull@chemie.fu-berlin.de

ERNST-WALTER KNAPP
knapp@chemie.fu-berlin.de

Frie Universität Berlin, Department of Chemistry and Biochemistry, Fabeckstr. 36a, 14195, Berlin-Dahlem, Germany

Protein-protein docking is a major task in structural biology. In general, the geometries of protein pairs are sampled by generating docked conformations, analyzing them with scoring functions and selecting appropriate geometries for further refinement. Here, we present an algorithm in real space to sample geometries of protein pairs. Therefore, we initially determine uniformly distributed points on the surfaces of the two protein structures to be docked and additionally define a set of uniformly distributed rotations. Then, the sampling method generates structures of protein pairs as follows: (i) We rotate one protein of the protein pair according to a selected rotation and (ii) translate it along a line connecting two surface points belonging to different proteins such that these surface points coincide. The resulting protein pair geometries are then analyzed and selected using a scoring function that considers residues and atom pairs. We applied this approach to a set of 22 enzyme-inhibitor complexes and demonstrate that a discretisation of the rigid-body search in real space provides an efficient and robust sampling scheme. Our method generates decoy sets with a considerable fraction of near-native geometries for all considered enzyme-inhibitor complexes.

Keywords: protein-protein docking; rigid-body geometry search; interface analysis

1. Introduction

Proteins are important regulators of biochemical processes in biological cells. They are for instance used to catalyze chemical reactions, to transport substrates through membranes and to stabilize cellular structures. Interactions with other molecules can affect a protein's macromolecular structure and functionality. For proteins, whose function is to form specific complexes with other proteins, the shape of the contact surface and the residue pair interactions at the contact surface are especially relevant [1]. This protein-protein interaction obeys the key-lock principle and is driven by free energy contributions, resulting in high binding affinities. Binding can influence the function of proteins in diverse ways from total inhibition to enhancement or induction.

Although genome-wide proteomics studies indicate that many proteins interact with each other, the number of complexes in the Protein Data Bank (PDB) increases very slowly. Possibly, this is related to the instability of transient protein-protein interactions, which make a crystallographic analysis difficult. Therefore, theoretical approaches for the identification and prediction of protein-protein interactions can be of great importance. Many efforts have been made to find a computational solution to this problem. Unlike the prediction of the binding modes for small molecules (i.e. FlexX [2], ICM [3] and Fado [4]), most protein-protein docking approaches consider the structures of the individual proteins in the complex to be rigid. Initially, a wide variety of docked conformations are generated and simultaneously evaluated by scoring functions. In general, these methods perform well when applied on individual protein conformations that are directly taken from the corresponding co-crystallized structures. However, predicting protein complex geometries using protein structures obtained from separate crystallizations essays remains difficult, often leading to many false positives. The binding process often involves conformational changes. Although these are generally subtle, they make it more difficult to find the proper complex geometry. Therefore, a further refinement of the proposed complex geometries by other methods, e.g. Monte Carlo approaches, is often necessary.

Currently, most established methods for rigid-body analysis of protein-protein interactions are based on the convolution technique in Fourier space as initially utilized by Katchalski-Katzir *et al.* in 1992 [5]. These approaches include ZDOCK [6], MolFit [7], 3D-Dock [8], DOT [9], GRAMM [10] and others. These methods use a scoring function defined on a discrete grid for each of the two proteins. Instead of evaluating the scoring function in real space, which is computationally expensive, the values of the scoring function are obtained by multiplication the corresponding Fourier transformed grids. This is done by assigning the atomic interaction parameters for each protein on separate grids, which are subsequently transformed by the fast Fourier transform (FFT) algorithm. In the Fourier space the Fourier coefficients are multiplied and the results are transformed back to real space. This is done for a large set of protein orientations [5]. Besides the FFT-based approaches, a variety of other procedures have also been applied on the protein-protein docking problem. Nussinov *et al.* proposed an algorithm based on geometric matching of knobs on the interacting surfaces [11]. Others, such as Baker [12] and Abagyan [13] have developed highly accurate methods using Monte Carlo simulations. The protein complex geometries are clustered [14] and their stability is analyzed by perturbation studies using different scoring functions [15].

The development of proper scoring functions is a non-trivial problem in protein-protein docking. A large variety of scoring functions attempt to capture the biophysically relevant properties for protein complex formation, such as e.g. interactions based on physical principles, on residue pair distributions or on geometric fit [16-20].

In this work, we describe a real space rigid-body protein-protein docking approach. Instead of assigning atom specific interaction parameters to each grid point, as necessary for FFT methods, we can take into consideration all interactions of atom pairs within a certain cutoff distance from the protein surfaces. In order to reduce the computational costs in real space, an efficient sampling strategy of the search space is used, which in turn allows to consider additional parameters in the scoring function. Two proteins are translated and rotated by a discrete set of transformations. To obtain the corresponding parameters for the transformations, the protein surfaces are uniformly covered by surface points. In addition, a set Q of uniformly distributed quaternions is generated from which the rotations are obtained. The translational vector is defined by the line connecting the pair of surface points selected from each of the two proteins. The residues interacting in the resulting geometry are evaluated by a statistical scoring function, which comprises geometrical and physicochemical components by considering residue pairs and atom pairs. The parameters of the scoring function were determined by Heuser *et al.* for enzyme-inhibitor complexes [20, 21].

2. Methods

2.1. *Preparing surface and grid representation*

From now on, we call the smaller of both proteins ligand (L) and the larger receptor (R). We embed both proteins by a grid with grid constant of 1.0 Å. Points of the receptor grid G_R, which are in the van der Waals (vdW) sphere of a receptor atom (radius of 1.8 Å for all atoms) are inside the receptor and marked as receptor points. If the receptor grid points are outside of the vdW volume of the corresponding protein they contain a neighbor list of protein atoms, which are within a distance cutoff of r_{cut}(neighbor) = 7 Å. This neighbor list provides an efficient way to find atomic interaction partners between the two proteins in the complex structure.

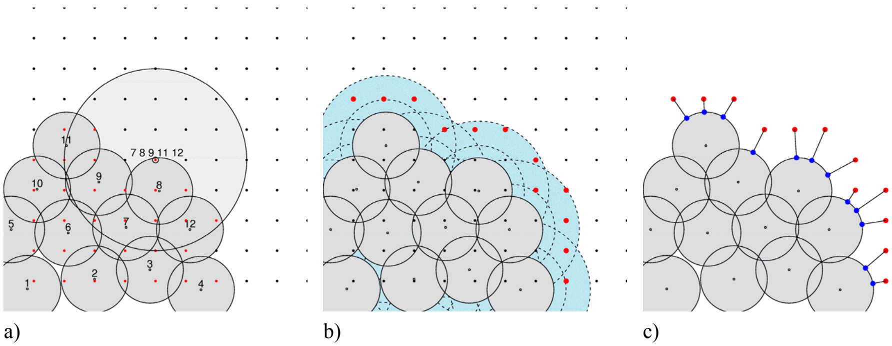

Fig. 1. Generation of neighbor list and surface points. Small spheres denote the protein atoms. a) Atom neighbor list of a reference grid point (center of large sphere) contains the numbers of atoms within the cut-off distance (largest sphere). b) Initial surface points (thicker red points of the grid) are all grid points, which are within a specified minimal and maximal distance (medium size blue spheres denoted by dashed lines) to the nearest protein atoms. c) The initial surface points are translated towards the center of the nearest protein atom until the vdW surface of the atom is reached (blue points on the surface of the gray spheres).

For both proteins (ligand and receptor) the grids are also used to determine surface points and surface normal vectors (see Fig. 1 for more details). In a first approximation the protein surface points are those grid points whose distances to the nearest protein atoms are between 4.0 and 6.0 Å. These points are then projected on the vdW surface of the nearest atom sphere. For each such surface point, we calculate a surface normal vector connecting the assigned atom center with the surface point. Then, we compute for all atoms of a residue the average of the surface normal vectors. Now we reduce the number of surface points. To obtain an even distribution of surface points we randomly select a single surface point and delete all other surface points within a distance of r_{cut}(surface) = 7 Å. Next, we select the nearest remaining surface point and repeat the procedure until all surface points have been selected or deleted. We denote the resulting sets of surface points S_R and S_L and of corresponding normal vectors V_R and V_L for the receptor and ligand, respectively. For the rotations a set Q of 8000 uniformly distributed quaternions is calculated with the approach described by Kuffner [22].

2.2. *Sampling strategy*

During the generation of the protein-protein geometries (called decoys), the receptor stays fixed, while the ligand is moved, i.e. translated and rotated. A decoy is defined by the triplet [q(k), $\mathbf{s}_R$(i), $\mathbf{s}_L$(j)], of quaternion q(k) ∈ Q and surface points $\mathbf{s}_R$(i) and $\mathbf{s}_L$(j) of receptor and ligand, respectively. For each pair [$\mathbf{s}_R$(i), $\mathbf{s}_L$(j)] of surface points we compute the angle $\alpha_{i,j}$ between the corresponding normal vectors $\mathbf{v}_R$(i) and $\mathbf{v}_L$(j). If this angle is smaller than a threshold value of $\alpha_{threshold}$ that is typically only slightly below 180°, we discard this decoy. If not, the ligand is translated by the difference vector $\Delta\mathbf{v}_{RL}(i, j) = \mathbf{s}_R(i) - \mathbf{s}_L(j)$. For the resulting protein complex structure we count the number of receptor points $n_{overlap}$, which are inside the vdW volume of the ligand. If $n_{overlap}$ exceeds 10% of the total number of ligand atoms the corresponding decoy is discarded as well, else the decoy is accepted and its score is computed. For each translated atom, we obtain the interacting atoms using the neighbor list and summing up the weighted contacts. The scoring function g used in the present study is defined by

$$g_{feature} = \sum_{m,n \in features} c_{feature}(m,n)\, W_{feature}(m,n) \tag{1}$$

where $c_{feature}$ is the number of interactions occurring for an atom pair type with the features m and n and $W_{feature}$(m,n) is corresponding element of the weighting matrix. The total score g^{total} of a generated decoy is defined by

$$g^{total} = g_{atom} * g_{residue} \tag{2}$$

where atom-based and residue-based weighting matrices W_{atom} and $W_{residue}$ are employed.

3. Results

3.1. *Docking performance*

We applied the described sampling approach on a set of 22 enzyme-inhibitor complexes (see Fig. 2 for a list of the corresponding PDB codes) from the ZDOCK 1.0 benchmark set [23]. We generated a set of uniformly distributed surface points for each individual protein structure using r_{cut}(surface) = 7 Å. Thus, we obtained on average 60 surface points for the receptors and 25 for the ligands (Fig. 2) yielding about 1500 pairs of surface points per protein complex on average. Hence, we consider 1500 translations and 8000 rotations and check that the normal vectors of the selected surface point pair possess an angle larger than $\alpha_{\text{threshold}}$. This yields decoys in the range of 10^7 per protein pair. For each of these decoys we verified that at most 10% of the ligand atoms overlap with receptor points. For the remaining decoys the scoring function g^{total}, eq. (2), was evaluated, keeping for each rotation the decoy with the highest score only. This results in about 8000 decoys per protein-protein complex. Figure 3 shows the number of generated near-native receptor-ligand geometries with an interface root mean square displacement (iRMSD) relative to the native complex structure below 5.0 Å. On average, about 50 near-native decoys out of the 8000 were generated per protein structure pair. For 1UDI, only 15 near-native decoys were generated, while the maximum number of 650 near-native decoys was obtained for 1BRC (Fig. 3). With a higher density of surface points using r_{cut}(surface) = 3 Å the results remained qualitatively similar.

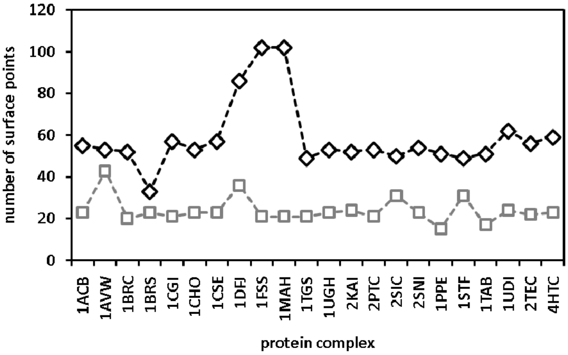

Fig. 2. Number of surface points of the considered 22 protein complexes consisting of receptor and ligand protein pairs (receptors: diamonds; ligands: squares).

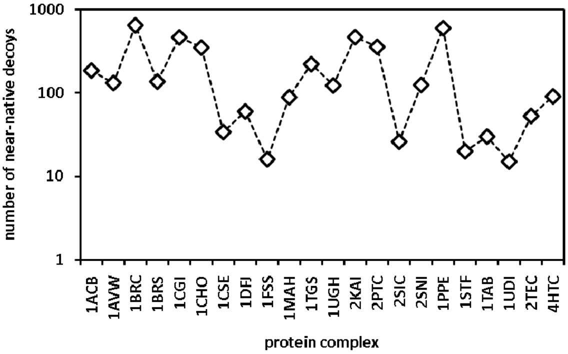

Fig. 3. Results of the protein docking approach. For each protein complex the highest ranked decoy per rotation was kept (about 8000 decoys per complex in total). The diamonds illustrate the number of decoys with an interface RMSD (iRMSD) below 5.0 Å.

3.2. *Sampling of a serine-protease-inhibitor complex*

In the following, we briefly illustrate the sampling results obtained for the first enzyme-inhibitor complex of the ZDOCK 1.0 benchmark set, which is a serine-protease-inhibitor complex (1ACB) [24]. We applied the algorithm on the separately crystallized protein structures 5CHA and 1CSE. The surface of the serine-protease was covered with 55 the inhibitor with 23 surface points. With the uniform set of 8000 rotations, more than 10^7 decoys were generated. Less than 5% (387047 in total) of these decoys fulfilled the geometrical criteria probing the ratio of receptor points with ligand atoms and the angle between the normal vectors of assigned surface points. We calculated the iRMSD of these decoys relative to the native reference complex, which was generated by aligning the separately crystallized protein structures on the co-crystallized true native complex structure. The iRMSD of this reference structure with the true native complex structure is 0.7 Å. About 10% of them have an iRMSD below 10 Å to the reference complex. The decoys were scored and the highest ranked decoy per rotation was kept (see details in 2.2) resulting in 8000 decoys. Figure 4 shows the scores with respect to the iRMSD for the 2000 highest ranked decoys. The complete set of 8000 decoys comprises 186 cases with an iRMSD below 5.0 Å, whereby the decoy with the lowest iRMSD of 4.8 Å is ranked at position 33. Considering all 8000 decoys, in eleven cases an iRMSD below 2.5 Å was detected. Hereby, the highest rank is 1743 with an iRMSD of 2.1 Å.

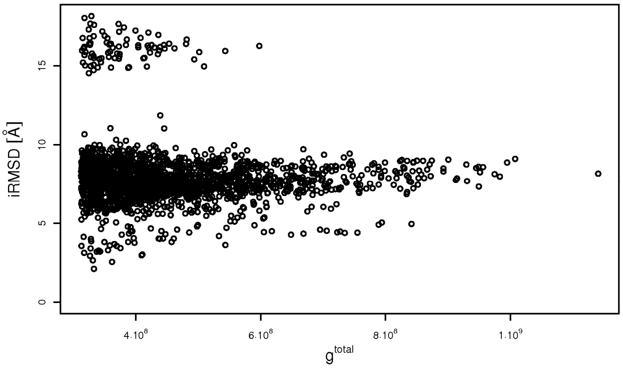

Fig. 4. Diagram correlating for the protein complex 1ACB [24] the iRMSD of the 2000 highest ranked decoys with the corresponding scores given by eq. (2).

Figure 5a shows the surface of the serine-protease and the center of masses of the inhibitor coordinates (dots) in the 8000 generated decoys. In Fig. 5b the serine-protease is shown together with the inhibitor in the native reference structure. The conserved residues of the serine-protease detected with BLAST [25] and CLUSTALW [26] were highlighted in dark red (Fig. 5b). It is evident that the residues in the interface between serine-protease and the inhibitor are highly conserved. Furthermore, we find that the binding cavity allows a better geometric match between the two protein structures than any other region detected on the serine-protease surface. Probably, the physicochemical specificity and the geometrical fit contribute to the large number of hits in the generated decoy set.

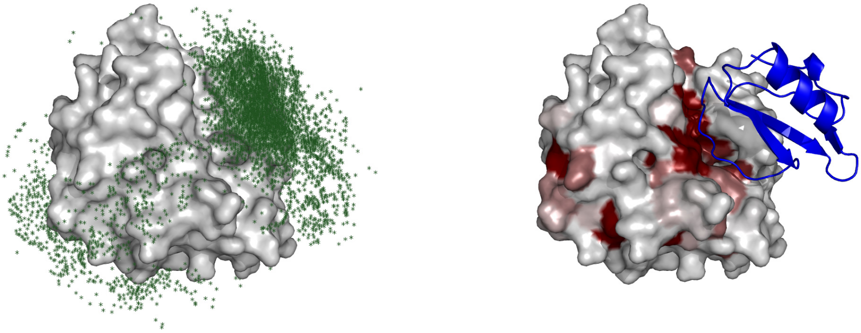

Fig. 5. Illustration of the docking results for the protein complex 1ACB. a) Surface of the receptor with the center of masses of the 8000 highest ranked decoys (green dots). b) Surface of the receptor and cartoon of the ligand structure (dark blue). The conserved residues of the receptor are highlighted in dark red.

To refine and rerank the decoys obtained with the initial sampling procedure, we performed a Monte Carlo stability analysis [27] using the program ROTAFIT [28]. Briefly, this procedure uses the 2000 top ranked decoys to perform 500 steps of a replica exchange Monte Carlo simulation using 10 replica. After the simulation, the pair-wise iRMSDs of the last 250 time steps of the five lowest temperature replica of each decoy are calculated. We then plot the number of structure pairs below a given iRMSD threshold versus the iRMSD threshold. The structural stability score of the decoys is calculated as the integral under this curve. Near-native decoys show a considerably higher structural stability score then false hits (Fig. 6).

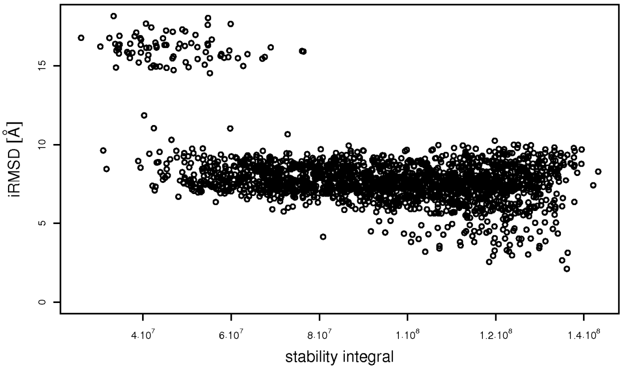

Fig. 6. Structural stability scores of the first 2000 decoys versus iRMSD.

4. Discussion

Initial-stage approaches in protein-protein docking are commonly based on the Fourier transform technique (FFT approach). This method is well established and capable to search an extensive variety of receptor-ligand geometries. However, the FFT approach carries inherent limitations. It can account only for interactions referring to pairs of coinciding grid points from the two proteins where the contribution to the scoring function is given as product of the parameters of the corresponding two grid points. The real space sampling technique of decoys allows using more general expressions for the scoring function.

The present application for a set of 22 enzyme-inhibitor complexes demonstrated that efficient sampling and scoring of receptor-ligand geometries in real space is computationally feasible. The method provides decoy sets with near-native geometries for all of the considered 22 enzyme-inhibitor complexes. The analysis of 1ACB, a serine-protease-inhibitor complex, emphasizes that the method is capable to generate a large fraction of near-native binding modes (see Fig. 3). In 186 out of 8000 cases, the protein-protein decoys exhibit an iRMSD of less than 5.0 Å. The highest number of 650 near-native binding modes has been generated for the protein complex 1BRC. The subsequent rescoring by structural stability analysis greatly improves the rank of near-native decoys. Interestingly, the stability integral proved to be a better way of identifying near-native decoys than various energy functions (data not shown).

In future studies, we plan to utilize our method for the evaluation of a variety of other all-atom, respectively heavy-atom, or residue-based scoring functions, which can be described as summation of weighted amino acid or atom pair interactions (see 2.2). We will also try to implement new scoring schemes. Thereby, the preliminary analysis of potential interface residues can be of particular interest. This can significantly improve the performances, since the described real space approach is capable to acquire preliminary residue selections to reduce the search space or to increase the surface resolution at particular protein surface sites. In addition, clustering the generated decoys can be used to improve detection of near-native complex structures. Finally, we aim to incorporate further rigid-body optimization procedures and perturbation studies to evaluate the stability of docked conformers and approaches to model the intramolecular flexibility of the two interacting protein structures.

Acknowledgments

We like to thank Connie Wang for useful contributions. This study was funded by the International Research Training Group (IRTG) on “Genomics and Systems Biology of Molecular Networks” (GRK1360, Deutsche Forschungsgemeinschaft (DFG)).

References

[1] Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. J., Spatial chemical conservation of hot spot interactions in protein-protein complexes, *BMC Biology*, 5, 2007.

[2] Rarey, M., Kramer, B., Lengauer, T., and Klebe, G., A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol.*, 261:470–89, 1996.

[3] Totrov, M. M., Abagyan, R. A., and Kuznetsov, D., ICM—a new method for protein modeling and design. Applications to docking and structure prediction from the distorted native conformation, *J. Comp. Chem.*, 15:488–506, 1994.

[4] Guerler, A., Moll, S., Weber, M., Meyer, H., and Cordes, F., Selection and flexible optimization of binding modes from conformation ensembles, *Biosystems*, 92:42-8, 2008.

[5] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A., Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques, *PNAS*, 89:2195–9, 1992.

[6] Chen, R. L. L. and Weng, Z., ZDOCK: an initial-stage protein-docking algorithm, *Proteins*, 52:80-7, 2003.

[7] Heifetz, A., Katchalski-Katzir, E., and Eisenstein, M., Electrostatics in protein–protein docking, *Protein Science*, 11:571-87, 2002.

[8] Gabb, H. A., Jackson, R. M., and Sternberg, M. J. E., Modelling protein docking using shape complementarity, electrostatics and biochemical information *J. Mol. Biol.*, 272, 1997.

[9] Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., and Eyck, L. F. T., Protein docking using continuum electrostatics and geometric fit, *Protein Eng.*, 14:105-13, 2001.

[10] Tovchigrechko, A. and Vakser, I. A., Development and testing of an automated approach to protein docking, *Proteins*, 60, 2005.

[11] Duhovny, D., Nussinov, R., and Wolfson, H. J., Efficient unbound docking of rigid molecules, *Lecture Notes in Computer Science*, 2452:185-200, 2002.

[12] Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D., Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations, *J. Mol. Biol.*, 331:281-99, 2003.

[13] Fernández-Recio, J., Totrov, M., and Abagyan, R., ICM-DISCO docking by global energy optimization with fully flexible side-chains, *Proteins*, 52:113-7, 2003

[14] Lorenzen, S. and Zhang, Y., Identification of near-native structure by clustering protein docking conformations, *Proteins*, 68:187-94, 2007.

[15] Kozakov, D., Schueler-Furman, O., and Vajda, S., Discrimination of near-native structure in protein-protein docking by testing the stability of local minima, *Proteins*, 72(3):993-1004, 2008.

[16] Kortemme, T. and Baker, D., Computational design of protein-protein interactions, *Curr. Opin. in Struct. Biol.*, 8:91-7, 2004.

[17] Lei, H. and Duan, Y., Incorporating intermolecular distance into protein-protein docking, *Protein Eng.*, 17:837-45, 2004.

[18] Keskin, O., Ma, B., Nussinov, R., Hot regions in protein-protein interactions: The organization and contribution of structurally conserved hot spot residues, *J. Mol. Biol.*, 345:1281-94, 2005.

[19] Shulman-Peleg, A., Shatsky, M., Nussinov, R., Wolfson, H. J., Spatial chemical conservation of hot spot interactions in protein-protein complexes, *BMC Biology*, 5, 2007.

[20] Heuser, P., Schomburg, D., Combination of scoring schemes for protein docking, *BMC Bioinformatics*, 8, 2007

[21] Heuser, P., Schomburg, D., Optimised amino acid specific weighting factors for unbound protein docking, *BMC Bioinformatics*, 7, 2006

[22] Kuffner, J. J., Effective sampling and distance metrics for 3D rigid body path planning, *In Proc. IEEE Int. Conf. on Robotics and Automation*, 2004.

[23] Chen, R., Mintseris, J., Janin, J., Weng, Z., A protein-protein docking benchmark, *Proteins*, 52:88-91, 2003.

[24] Frigerio, F., Coda, A., Pugliese, L., Lionetti, C., Menegatti, E., Amiconi, G., Schnebli, H. P., Ascenzi, P., Bolognesi, M., Crystal and molecular structure of the bovine a-chymotrypsin-eglin c complex at 2.0 A resolution, *J. Mol. Biol.*, 225:107-23, 1992.

[25] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25:3389-402, 1997.

[26] Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., Higgins, D. G., ClustalW and ClustalX version 2.0, *Bioinformatics*, 21:2947-8, 2007.

[27] Lorenzen, S., Detecting near-native docking decoys by Monte Carlo stability analysis, *Genome Informatics*, 18:206-14, 2007.

[28] Lorenzen, S., Zhang, P. F., Monte Carlo refinement of rigid-body protein docking structures with backbone displacement and side-chain optimization, *Protein Science*, 16:2716-25, 2007.