Using tetrahedral grid-based protein models in docking

2007 J. Phys.: Condens. Matter 19 285209

(http://iopscience.iop.org/0953-8984/19/28/285209)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 285209 (8pp) doi:10.1088/0953-8984/19/28/285209

Using tetrahedral grid-based protein models in docking

G Wieczorek[1] **and P Zielenkiewicz**[1,2]

[1] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5a, 02-106 Warszawa, Poland
[2] Plant Molecular Biology Department, Warsaw University, Pawinskiego 5a, 02-106 Warszawa, Poland

Received 9 October 2006, in final form 29 December 2006
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285209

Abstract

Protein–protein recognition, leading to the formation of specific functional complexes, involves complementary surfaces of interacting subunits. Current docking protocols employ complex scoring functions, neglecting proper shape matching by the use of cubic grids. In the present paper a docking algorithm based on the tetrahedral grid model of proteins is described, allowing a more precise description of shape complementarity. The software was tested on the docking benchmark, giving excellent results for rigid-body docking. Extension of the present methodology to flexible docking is in progress.

1. Introduction

Interactions between macromolecules have been the subject of extensive research in recent years [1–9]. There are several programs available for trying to solve the problem of finding the structure of protein complexes. The first algorithms addressing this challenging target treated docked macromolecules as rigid bodies [10–17]. Recently more and more procedures that try to take into account the flexible nature of biological macromolecules have appeared [18], but rigid body docking is still the most popular amongst *in silico* docking approaches. A very popular methodology of resolving the docking problem consists of three stages: the first is digitization of the molecules of interest, the second is the search for shape complementarity of the obtained models, and finally there is the clustering of results [19–21]. The digitization part is done by projecting the molecules onto a grid. The grid used—its topology and spacing—seems to strongly influence the results of further steps. The most popular type of grid in docking applications is the cubic grid. Tetrahedral grids, on the other hand, have been applied in many fields requiring accurate description of 3D geometries, such as hydrodynamics, aerodynamics, computer modelling and visualization (for example [22–25]). Cubic grids are not used in such applications at all because of their poor performance in 3D object modelling. Moreover, the employment of cubic grids in molecular surface representation has been discouraged [26]. Also, the chemical nature of carbons with four covalent bonds (four-fold tetrahedral coordination) leads to their having approximate tetrahedral geometry, which influences the packing and surface landscape of the molecule. Since carbons make

up the majority of atoms in proteins this makes the tetrahedral grid a particularly appropriate choice [27]. The reason for using a cubic grid in docking instead of any other is the low computational cost of creating such a grid, and the method used for docking of protein models, which in most cases is calculation of a correlation function that assesses the degree of molecular surface overlap by means of Fourier transformation. The aim of this work was to create a program for macromolecule docking based on shape complementarity estimations that would be based on the tetrahedral grid model of proteins and allow the density of search space sampling to be precisely adjusted according to the hardware and amount of time available to the user. Also, since a pure shape complementarity approach of rigid body representation is not sufficient for finding correct solutions to docking problems in which significant conformational changes occur upon the binding process [18, 28, 29], the program described was conceived as a framework for including the information about the flexibility of molecules into docking calculations. This part of the work is still in progress.

2. Methods

2.1. Shape complementarity estimation algorithm

Structure descriptions containing atomic coordinates of molecules **A** and **B**, for example in PDB format, are the starting point of the procedure. Before any further steps, atomic radii have to be assigned. The procedures exported in the Gromacs library [30] have been adapted for this purpose. The atomic radius is estimated as half of the distance between two atoms of the same kind, for which the Lennard-Jones potential equals zero:

$$C_{ii}^{(12)}/r_{ii}^{12} - C_{ii}^{(6)}/r_{ii}^{6} = 0$$

where $C_{ii}^{(12)}$ and $C_{ii}^{(6)}$ are Lennard-Jones parameters, r_{ii} is the distance between a pair of atoms of the same type. Molecules in certain orientations are then projected onto tetrahedral grids consisting of N points a_i, $i \in \langle 1, 2, 3, \ldots, N\rangle$ for molecule **A**, and M points b_j, $j \in \langle 1, 2, 3, \ldots, M\rangle$ for molecule **B**, respectively. If the grid point is inside the molecule, it is assigned the value 1, otherwise it takes the value 0. In the tetrahedral grid every point (besides those located near the grid boundary) has 12 equally distant neighbouring points. Together they form a cuboctahedron with a centre a_{i0} and vertices a_{i1} to a_{i12}. For further analysis, for molecule **A**, points of the grid are selected which possess the value 1 and at least one of their neighbouring points has the value 0. For molecule **B**, on the contrary, grid points of value 0 are selected, of which at least one of their neighbouring grid points has the value 1. Most of the selected points are just above the surface of molecule **A** and just under the surface of molecule **B**. The rest of the thus defined points reside in cavities not accessible to the solvent, and are excluded from the list of grid points for shape complementarity analysis. Besides the information about the state of the grid points a_i or b_i (0 or 1) and states of all 12 neighbouring points, there is a vector $\boldsymbol{V}\boldsymbol{a}_i$ associated with every a_i ($\boldsymbol{V}\boldsymbol{b}_j$ and b_j, respectively) joining the geometrical centre of the molecule with this a_i (or b_j). Thus, every grid point (with all mentioned data attached to them) on the molecular surface carries information about the shape of a small patch of that surface and its position in space. The surface grid point b_j is then superposed onto a_i and the score is calculated, estimating the shape complementarity of these patches of molecular surface (see figure 1 and table 1). Vector $\boldsymbol{P}\boldsymbol{a}_i b_j = \boldsymbol{V}\boldsymbol{a}_i - \boldsymbol{V}\boldsymbol{b}_j$ relates the centres of molecules **A** and **B**, given the fact that molecule **B** undergoes a shift in order to superpose b_j on a_i, and specifies the point in another three-dimensional tetrahedral grid, in which the calculated score is accumulated. Superposing and score calculation is repeated for every pair of previously selected surface grid points. The cumulative scores are sorted

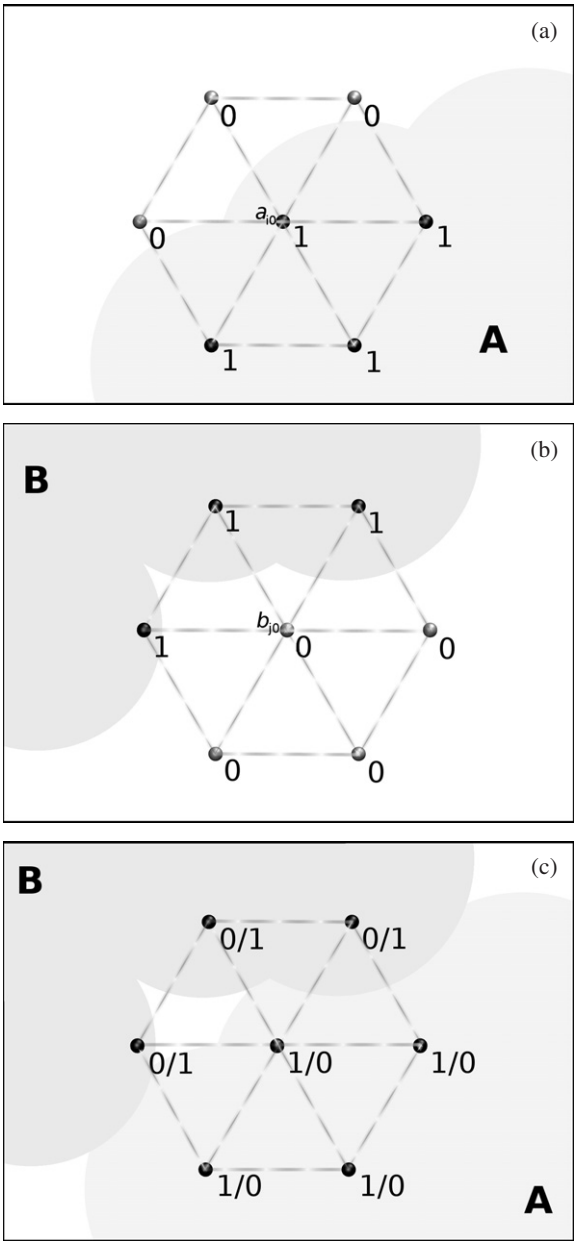

Figure 1. Schematic (2D) description of the scoring algorithm. (a) The way the grid on molecule **A** is built. Points of the tetrahedral grid inside the molecule, and having at least one neighbour outside the molecule, are selected for further processing. (b) On molecule **B**, on the other hand, points just above the surface of the molecule and having at least one neighbour inside the molecule are selected. (c) Superposing two cuboctahedra (hexagons in this 2D schematic), one from the grid of molecule **A** and one from molecule **B**, simulates a shift of molecule **B** towards molecule **A**. This allows for estimation of shape complementarity for small patches of molecular surfaces around points a_{i0} and b_{j0}.

and orientations of molecules of the best shape match are found. Then, the molecule **B** is rotated, grids for the molecule and for the scores are rebuilt, and the score calculation procedure

Table 1. An example of 'rewards' and 'penalties' in score calculation.

a_i	0	0	1	1
b_j	0	1	0	1
Score[a]	0	1	1	−10

[a] When a_i and b_j are both 0, the molecule surfaces are separated. There is an overlap if they are both 1. Otherwise there is a surface match.

is repeated. Even though the number of computations needed for complementarity searches described here is limited only to grid points at the surface of molecules **A** and **B**, the scoring procedure is highly computationally expensive. One of the ways to reduce the number of computations is to use external information such as data regarding putative binding sites. In the program described it is possible to exclude from computations parts of the molecule that are not suspected of being a part of the interface between docked molecules, which may lead to a reduction in the time needed for calculation by more than an order of magnitude. An example of a molecular surface and corresponding tetrahedral grid obtained in the way just described is shown in figure 2.

2.2. Rotation of molecules

The algorithm calculates shape complementarity for certain orientations of docked proteins. The program therefore has to perform many docking attempts and rotate molecules each time. The implemented methods of molecule rotation allow for extensive control over the density of rotational space sampling. In the representation axis-angle, it is possible to manipulate both the number of axes and the angles the molecule has to be rotated by. Special care was taken to exploit the symmetry properties of the tetrahedral grid to reduce the number of rotations needed. A cuboctahedron has 24 symmetry rotations, one of which is identity operation. Having one grid model of a molecule, it is sufficient to apply fast bit-wise symmetry rotation to obtain 23 other models of the protein in different orientations. To take advantage of symmetry operations though, it was necessary to precisely reduce the rotational space and select $1/24$ of it in such a way that the selected part after applying symmetry transformations covered the whole rotational space. Figure 3 shows an example of rotational space sampling using the algorithm described. In the case of monomultimer searching, the program accepts an option which tells what order of multimer (dimeric, trimeric or higher) the user is interested in. Having this information, it is possible to significantly reduce the rotational space that has to be taken into account.

2.3. Clustering

In the simplest cases, the results of the described calculation, consisting of information about the relative position of proteins and the score they achieved in such a position, are sufficient for discriminating the correct solution of the docking problem (see results below). However, most of the docking tasks require further data analysis, one of the most important parts of which is geometric clustering of docked proteins. The assumption underlying this part of the procedure is that events occurring in clusters are probably not random [31]. The clustering methods most extensively used here were simple linkage and Jarvis–Patrick algorithms, which are available in the g_cluster program from the Gromacs distribution. The total score for every cluster as a sum of scores of individual cluster members was calculated, which was the main assessment criterion of the solution. As a final result, the complex with the highest score in the cluster was proposed.

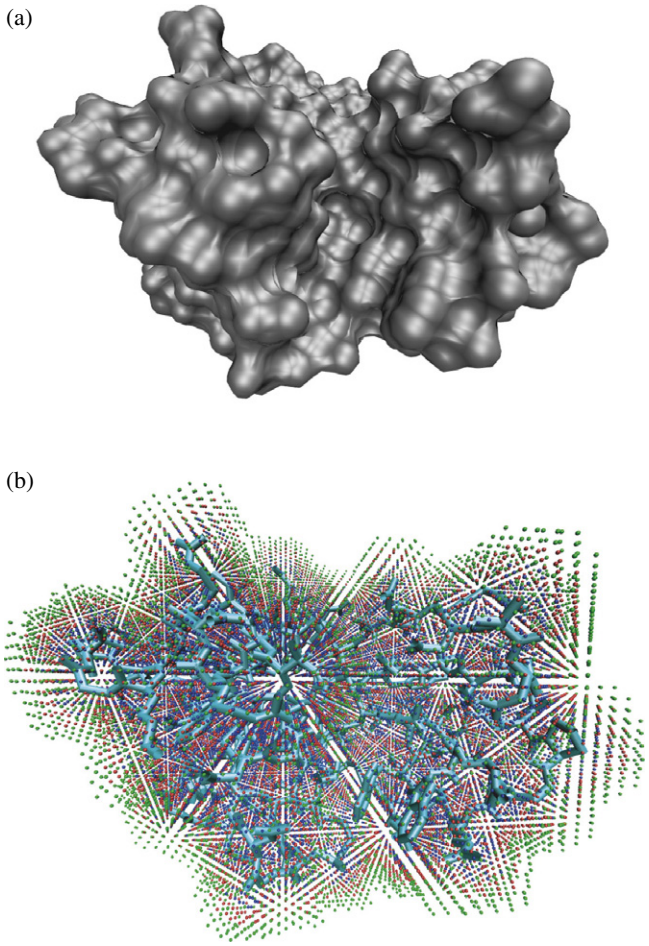

Figure 2. An example of the van der Waals surface of a protein (hen egg lysozyme) (a), and the bond representation of the same molecule in the same orientation submerged in a tetrahedral grid of 1 Å spacing (b). The molecule was treated as molecule **B**. Red grid points are just above the van der Waals surface of the molecule, the remaining points are the neighbours, where blue ones are grid points inside the molecule and green are outside the molecule.

3. Calculations and results

Some of the results obtained during tests of the described algorithm will be presented here. As the docking targets, proteins from a second version of the benchmark for docking programs were used [32]. The benchmark consists of 84 pairs of proteins that have been shown to create complexes. The protein pairs have been divided into three groups: rigid body (63 pairs), medium difficulty (13) and difficult (8). Every pair has been crystallized as a complex and as unbound monomers, and the structures of such proteins were resolved by means of x-ray crystallography.

In the program one has to set parameters such as atomic parameters, grid spacing, penalty for overlap of molecules, density of rotational space sampling, number of best solutions taken into account coming from every orientation, the clustering method used and its parameters, etc. For the purpose of the test, the proteins were converted to an all-atom optimized potential

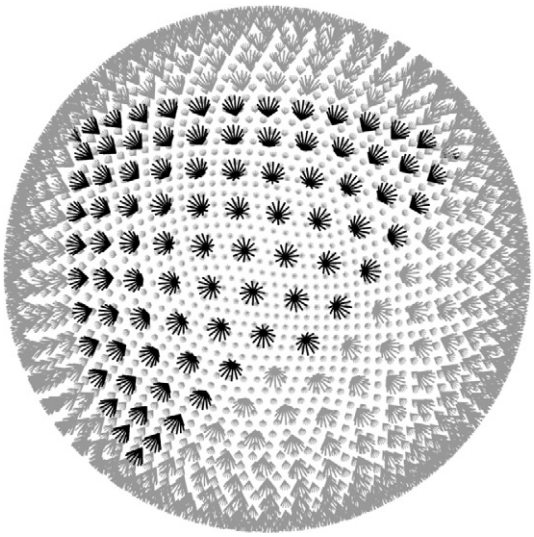

Figure 3. An example of coverage of the rotational space, perspective view. Rotations are represented as axes and angles, where axes are vectors (not shown) from the centre of the sphere to points from which small vectors spread. Small vectors represent several angles of rotation (30° in this case, which is a value good for such a visualization) [34]. The black part is the base set of rotations, the rest is an effect of 23 symmetry operations on the base set.

for liquid simulations (OPLS) forcefield [33] Gromacs run files. During docking molecule **B** was rotated by 5.9°. The measure of correctness of results was root mean square deviation (RMSD) of the displacement of atoms of molecule **B**, compared to the reference structure. For simplicity, results with RMSD values below 4.0 Å were named correct solutions of the docking problem. The overlap penalty (table 1) was set to −10 *a priori*, just to be significantly more influential than the complementarity reward, which was set to 1. For each orientation, the first 200 results were stored. The tests started with proteins in the bound state. Twenty pairs of proteins have been docked with a grid spacing set to 1.28 Å. Such a spacing was found at the beginning of the development of the docking algorithm to give reasonable results for simple cases. For 17 pairs, the results with the highest score were correct solutions, for one pair the first correct solution was in the second place according to the score, and for two pairs in the third place. After clustering of the first 10 000 results using simple linkage and Jarvis–Patrick algorithms, both with 3 Å cutoff, the biggest clusters contained correct solutions in all cases. The results are shown in table 2.

Next, docking of unbound structures was attempted. With a grid spacing of 1.28 Å and an overlap penalty of −10, the program did not usually rank the correct solution high enough to make it distinguishable from false positives. In docking unbound structures, the sidechains (and the backbone in difficult cases) do not have the same structure as when the proteins are docked. So at the orientation close to the bound state there are many overlaps between two molecules. Here is the place for taking flexibility into account, which is the main focus of the development. It is possible to improve the rigid body docking even in those difficult cases. When tweaking the parameters, mainly decreasing grid spacing and imposing less rigorous overlap penalties, the program was able in some cases to make the correct solution more significant when compared with false positives.

Table 2. Results of shape complementarity calculation.

Name	ΔASA[a] (Å^2)	Score[b]	Rank	Cluster score	Cluster rank	cs/fs[c]
1ACB	1554	1289	1	323 624	1	1.87
1AHW	1899	1133	1	170 709	1	6.95
1AK4	1029	944	3	115 945	1	2.06
1AKJ	1995	1066	1	134 863	1	2.85
1AY7	1237	1002	1	242 500	1	4.31
1B6C	1752	1209	1	266 832	1	3.83
1BJ1	1731	1600	2	448 419	1	1.42
1BUH	1324	938	3	132 602	1	2.26
1BVN	2222	1182	1	353 242	1	8.38
1CGI	2053	1687	1	582 256	1	12.50
1DQJ	1765	1492	1	268 541	1	10.11
1E6J	1245	1104	1	112 309	1	1.96
1EAW	1866	1265	1	283 109	1	2.59
1F34	3038	1768	1	774 699	1	6.64
1FAK	3363	1906	1	110 680	1	4.55
1FSK	1623	1244	1	109 314	1	1.31
1GCQ	1208	1223	1	316 975	1	3.75
1HE1	2113	1621	1	491 648	1	11.87
1HIA	1737	1255	1	648 164	1	6.87
1I2M	2779	1890	1	192 432	1	5.01

[a] Change in accessible surface area upon complex formation calculated using NACCESS [32].
[b] Highest score of the proper solution.
[c] Ratio of the cluster score of the proper solution (cs) to the highest false cluster score (fs).

4. Discussion

Although the chosen test cases are considered simple, and the performance is far from what one would expect from a 'working' docking program, it is necessary to remember that the only criterion used here was a geometrical one. All modern docking procedures take electrostatic interactions and desolvation terms into account. However, in most of the existing docking programs based on grid models of proteins, a cubic grid is used, which has been shown not to be precise. The importance of proper shape matching seems to be neglected in favour of building complicated scoring functions, in most cases combining different quantities in a completely nonphysical way, or building metaservers, which sometimes work for other types of problems, such as protein folding, but do not enable us to understand the mechanisms behind them any better. Thus, the aim of this project was to take one step back and take advantage of shape matching to its edge. When this is done, taking long range interactions into account might be a way of development, as well as working on the parameters. For the authors of this program the most tempting choice is to continue the exploration of shape complementarity taking into account full flexibility information in the docking procedure, for example coming from molecular dynamics, without a preliminary rigid body docking step.

References

[1] Janin J and Wodak S J 2002 The structural basis of macromolecular recognition *Adv. Protein Chem.* **61** 9–73
[2] Keanthous C 2000 *Protein–Protein Recognition* (Oxford: Oxford University Press)
[3] Jones S and Thornton J M 1996 Principles of protein–protein interactions *Proc. Natl Acad. Sci. USA* **93** 13–20
[4] Marcotte E M, Pellegrini M, Thompson M J, Yeates T O and Eisenberg D 1999 A combined algorithm for genome-wide prediction of protein function *Nature* **402** 83–6

[5] Xenarios I, Fernandez E, Salwinski L, Duan X J, Thompson M J, Marcotte E M and Eisenberg D 2001 DIP: The database of interacting proteins: 2001 update *Nucl. Acids Res.* **29** 239–41
[6] Enright A J, Iliopoulos I, Kyrpides N C and Ouzounis C A 1999 Protein interaction maps for complete genomes based on gene fusion events *Nature* **402** 86–90
[7] Bader G D, Donaldson I, Wolting C, Ouellette B F, Pawson T and Hogue C W 2001 Bind—the biomolecular interaction network database *Nucl. Acids Res.* **29** 242–5
[8] Park J, Lappe M and Teichmann S A 2001 Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the pdb and yeast *J. Mol. Biol.* **307** 929–38
[9] Mendez R, Leplae R, De Maria L and Wodak S J 2003 Assessment of blind predictions of protein–protein interactions: current status of docking methods *Proteins* **52** 51–67
[10] Wodak S J and Janin J 1978 Computer analysis of protein–protein interaction *J. Mol. Biol.* **124** 323–42
[11] Janin J and Wodak S J 1985 Reaction pathway for the quaternary structure change in hemoglobin *Biopolymers* **24** 509–26
[12] Jiang F and Kim S H 1991 Soft docking: matching of molecular surface cubes *J. Mol. Biol.* **219** 79–102
[13] Cherfils J, Duquerroy S and Janin J 1991 Protein–protein recognition analyzed by docking simulation *Proteins* **11** 271–80
[14] Shomhet B K and Kuntz I D 1991 Protein docking and complementarity *J. Mol. Biol.* **221** 79–102
[15] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem A A, Aflalo C and Vakser I A 1992 Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques *Proc. Natl Acad. Sci. USA* **89** 2195–9
[16] Norel R, Fischer D, Wolfson H and Nussinov R 1994 Molecular surface recognition by computer vision-based technique *Protein Eng. Design Selection* **7** 39–46
[17] Norel R, Petrey D, Wolfson H and Nussinov R 1999 Examination of shape complementarity in docking of unbound proteins *Proteins* **36** 307–17
[18] Bonvin A M J J 2006 Flexible protein–protein docking *Curr. Opinion Struct. Biol.* **16** 194–200
[19] Eyck T L F, Mandell J, Roberts V A and Pique M E 1995 Surveying molecular interactions with dot *Supercomputing '95: Proc. 1995 ACM/IEEE Conf. on Supercomputing (CDROM)* (New York: ACM Press) p 22
[20] Chen R, Li L and Weng Z 2003 Zdock: an initial-stage protein-docking algorithm *Proteins* **52** 80–7
[21] Camacho C J and Gatchell D W 2003 Successful discrimination of protein interactions *Proteins* **52** 92–7
[22] Frink N and Pirzadeh S Z 1999 Tetrahedral finite-volume solutions to the Navier–Stokes equations on complex configurations *Int. J. Numer. Methods Fluids* **31** 175–87
[23] Wilmoth R G, LeBeau G J and Carlson A B 1996 Dsmc grid methodologies for computing low-density, hypersonic flows about reusable launch vehicles *AIAA* **96** 1812
[24] Gerstner T and Pajarola R 2000 Topology preserving and controlled topology simplifying multiresolution isosurface extraction *Proc. Visualization 2000* ed T Ertl, B Hamann and A Varshney, pp 259–66
[25] Sharov D and Nakahashi K 1996 Hybrid prismatic/tetrahedral grid generation for complex geometries *AIAA* **96** 2000
[26] Chan S L and Purisima E O 1998 Molecular surface generation using marching tetrahedra *J. Comput. Chem.* **19** 1268–77
[27] Bagci Z, Kloczkowski A, Jernigan R L and Bahar I 2003 The origin and extent of coarse-grained regularities in protein internal packing *Proteins* **53** 56–67
[28] Chern-Sing G, Milburn D and Gerstein M 2004 Conformational changes associated with protein–protein interactions *Curr. Opin. Struct. Biol.* **14** 104–9
[29] Parak F G 2003 Proteins in action: the physics of structural fluctuations and conformational changes *Curr. Opin. Struct. Biol.* **13** 552–7
[30] Berendsen H J C, van der Spoel D and van Drunen R 1995 Gromacs: a message-passing parallel molecular dynamics implementation *Comput. Phys. Commun.* **91** 43–56
[31] Kozakov D, Clodfelter K H, Vajda S and Camacho C J 2005 Optimal clustering for detecting near-native conformations in protein docking *Biophys. J.* **89** 867–75
[32] Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J and Weng Z 2005 Protein–protein docking benchmark 2.0: an update *Proteins* **60** 214–6
[33] Kaminski G A, Friesner R A, Tirado-Rives J and Jorgensen W L 2001 Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides *J. Phys. Chem.* B **105** 6474
[34] Kuffner J 2004 Effective sampling and distance metrics for 3d rigid body path planning *ICRA 2004: Proc. 2004 IEEE Int. Conf. on Robotics and Automation* (Piscataway, NJ: IEEE)