BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 19 2009, pages 2544–2551
doi:10.1093/bioinformatics/btp447

Structural bioinformatics

FRODOCK: a new approach for fast rotational protein–protein docking

José Ignacio Garzon[1], José Ramón Lopéz-Blanco[1], Carles Pons[2,3], Julio Kovacs[4,†], Ruben Abagyan[4], Juan Fernandez-Recio[2] and Pablo Chacon[1,*]

[1]Centro de Investigaciones Biológicas, CSIC, Ramiro de Maeztu, 9. 28040 Madrid, Spain, [2]Barcelona Supercomputing Center, [3]National Institute of Bioinformatics, Computational Bioinformatics, Jordi Girona 29, Barcelona 08034, Spain and [4]Department of Molecular Biology, The Scripps Research Institute La Jolla, CA 92037, USA

Received on March 23, 2009; revised on July 15, 2009; accepted on July 16, 2009

Advance Access publication July 20, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Prediction of protein–protein complexes from the coordinates of their unbound components usually starts by generating many potential predictions from a rigid-body 6D search followed by a second stage that aims to refine such predictions. Here, we present and evaluate a new method to effectively address the complexity and sampling requirements of the initial exhaustive search. In this approach we combine the projection of the interaction terms into 3D grid-based potentials with the efficiency of spherical harmonics approximations to accelerate the search. The binding energy upon complex formation is approximated as a correlation function composed of van der Waals, electrostatics and desolvation potential terms. The interaction-energy minima are identified by a novel, fast and exhaustive rotational docking search combined with a simple translational scanning. Results obtained on standard protein–protein benchmarks demonstrate its general applicability and robustness. The accuracy is comparable to that of existing state-of-the-art initial exhaustive rigid-body docking tools, but achieving superior efficiency. Moreover, a parallel version of the method performs the docking search in just a few minutes, opening new application opportunities in the current 'omics' world.

Availability: http://sbg.cib.csic.es/Software/FRODOCK/

Contact: Pablo@cib.csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The prediction of the structure of a protein–protein complex from the coordinates of unbound components by docking methods is one of the major challenges in current computational structural biology (Bonvin, 2006; Deremble and Lavery, 2005; Gray, 2006). Accurate predictions, properly integrated with experimental data could give new insights into the basic principles of molecular recognition and the mechanism of protein association, which are key to cellular functioning. This can be particularly interesting in the context of structural genomics efforts. Thanks to these initiatives and the advance of modeling, it is increasingly frequent to be able to successfully predict the interaction of a pair of proteins at atomic detail. Moreover, a better understanding of protein–protein interactions will be undoubtedly useful for structure-based drug design and other biotechnological applications.

Using a wide range of different strategies, existing protein–protein docking algorithms are steadily improving in both reliability and accuracy as it can be seen from the reported results of the CAPRI blind docking experiments (Lensink *et al*., 2007; Mendez *et al*., 2005). These docking methods generally have an initial stage during which the components are rigidly combined. During this stage, many predictions are generated, and later assessed in a second refinement stage. Rescoring such large set of predictions with more accurate approximations can eventually filter out false positives produced in initial rigid-body searches. Adding certain side-chain or backbone flexibility and/or constraining the search with experimental information of the binding site are successful strategies that significantly help the filter process.

Despite some promising results, major research effort is still needed in order to improve existing approaches. The biggest challenge is to combine high-accuracy energy calculations with speed and sampling power, while being able to handle induced conformational searches at the protein–protein interfaces. See Bonvin (2006), Camacho and Vajda (2002), Ritchie (2008), Vakser and Kundrotas (2008) for complete reviews of existing docking approaches and their limitations.

Here we focus on the first stage of docking, which consists on rigid-body orientational sampling of a ligand molecule with respect to a fixed receptor molecule while a docking scoring function is maximized. The 6D sampling space of the relative orientations between ligand and receptor is huge, and therefore computationally demanding. To efficiently tackle this search, many of current approaches follow the Fast Fourier Transform (FFT)-based algorithm described by Katchalski-Katzir *et al*. (1992). In that approach, the molecules are represented by 3D grids that carry information of the shape. The ligand and receptor grids are then correlated using FFT to efficiently scan the translational space. After the Fourier-based evaluation has been complemented by an implicit

*To whom correspondence should be addressed.

†Present address: SeaSpace Co., 12120 Kear Place, Poway, CA 92064, USA

orientational search, a large number of docked conformations with favorable surface complementarity can be obtained. This initial shape-based scoring function has been further enhanced by including additive correlation terms to consider electrostatics [FTDOCK (Gabb *et al*., 1997), DOT (Moont *et al*., 1999) and Molfit (Heifetz *et al*., 2002)], solvation [ZDOCK (Chen *et al*., 2003)] or even statistical interaction potentials [PIPER (Kozakov *et al*., 2006)].

Despite significant progress in the FFT-based methods, there is room for improvement in both speed and accuracy of the grid-based scoring function. The efficiency of the 6D FFT-based search depends on several factors. The computational cost increases with size, scaling at $O(N\log N)$, where N is the number of grid cells. The efficiency also decreases with the number of considered interaction potential terms, given that the global energy is computed as a sum of independent FFT correlation functions. Moreover, several rigid-body searches might be performed if the flexibility is explicitly considered or if alternative homology structures are used as docking templates. Thus, the FFT-based search process can take several hours or even more if the sampling is relatively large or several candidates must be docked.

The docking efficiency can be further improved by accelerating the rotational search using spherical harmonics (SH). In the Hex docking correlation algorithm (Ritchie and Kemp, 2000), the rotational docking is accelerated by correlating spherical polar basis functions (SPF) that model the surface shape and charges of docking molecules. Very recently, the same authors (Ritchie *et al*., 2008) presented several improvements for calculating multidimensional multi-property rotational FFT docking SPF correlations. Inspirited by the efficiency achieved by this approach, here we have adapted our original Fast Rotational Method (FRM) (Kovacs and Wriggers, 2002; Kovacs *et al*., 2003), which was previously successfully used to fast fit atomic structures into electron microscopy (EM) density maps (Garzon *et al*., 2007), to protein–protein docking. This approach permitted a superior efficiency and a more exhaustive search by speeding up the three rotational degrees of freedom using SH and a convenient formulation of the 3D rotation group. The application of FRM to protein–protein docking has derived in new mathematical expressions, and hence in a novel docking methodology termed FRODOCK (Fast ROtational DOCKing).

In contrast to other approaches, FRODOCK has the advantage of combining the capability to express the interaction terms into 3D grid-based potentials with the efficiency of a SH-based rotational search. The binding energy upon complex formation is approximated by a sum of three types of potentials: van der Waals, electrostatics and desolvation, each of which can be written as a correlation function. These potentials are conveniently pre-calculated on a 3D grid, using appropriate energy thresholds. The interaction energy minima, and hence the potential docking solutions, are identified by a new fast and exhaustive rotational docking SH-based search combined with a simple translational scanning. A parallel version of FRODOCK can perform the docking search in just a few minutes, and the competitive docking accuracy achieved on standard protein–protein benchmarks demonstrates its applicability and robustness.

2 METHODS

Global energy optimization was performed by 6D (3D rotations + 3D translations) rigid-body exhaustive search of the orientations of a fixed ligand with respect to a mobile receptor. The docking criterion is the minimization of a scoring function based on the interaction energy and composed of several terms. Considering only the rotational part, each energy term can be calculated by a correlation function defined as an integral of the form:

$$E(R)=\int f\cdot\Lambda_R g \tag{1}$$

where f and g correspond to the interaction potential parts of the receptor and ligand, respectively. The operator Λ_R denotes rotation of g by R defined by canonical Euler angles ϕ, θ and ψ. On the unit sphere, the interaction potential can be expressed in terms of SH functions, $Y_{lm}(\beta,\lambda)$, and its corresponding coefficients $\hat{f}_{lm}(r)$ and $\hat{g}_{lm}(r)$:

$$f(r,\beta,\lambda)=\int_{s^2}\hat{f}_{lm}(r)Y_{lm}(\beta,\lambda)\;\; g(r,\beta,\lambda)=\int_{s^2}\hat{g}_{lm}(r)Y_{lm}(\beta,\lambda) \tag{2}$$

where r is the radius of the unit sphere; $l\geq 0$ and $-l\leq m\leq l$ are the SH degree and order, and β and λ are the co-latitude and longitude, respectively. Instead of employing SPF functions Ritchie and Kemp, (2000), here the SH transformation is done discretely in concentric spherical layers (like onion shells) as previously described in (Kovacs and Wriggers, 2002). This radialization process permits a novel volumetric description of an interaction potential defined into a 3D grid in terms of harmonic functions.

The correlation docking function can be expressed in terms of an inverse Fourier transform of the SH functions (Garzon *et al*., 2007; Kovacs and Wriggers, 2002; Kovacs *et al*., 2003):

$$E(R)=FT^{-1}_{m,h,m'}\left[\sum_l d^l_{mh}\, d^l_{hm'} I^l_{mm'}\right] \text{ where } I^l_{mm'}=\int_0^\infty \hat{f}_{lm}(r)\cdot\overline{\hat{g}_{lm'}(r)}\cdot r^2\cdot dr \tag{3}$$

where d^l_{mhn} is the real coefficient that defines the matrix elements of the irreducible representations of the 3D rotation group. This expression can be computed very efficiently by pre-calculating such coefficients and by using as upper limit of integration the maximum shell radius for which a given potential has non-zero values. In addition to a very fast calculation of the rotational docking correlation, Equation (3) permits a deep and exhaustive rotational search. Note that rotational sampling step is limited by twice of the bandwidth (bw) used in the harmonic expansion of the correlated potentials [Equation (2)]. For example, a bandwidth of 32 corresponds to a sampling rotational step of 5.6° which implies the scanning of more than 60 000 distinct rotations.

This fast exhaustive rotational search combined with an implicit translational scan was successfully employed by us to fit atomic structures into low-resolution EM maps (Garzon *et al*., 2007). In that case, we cross-correlated two electron density maps: one that corresponded to the experimental EM map and another that corresponded to a lower-resolution version of the atomic structure. In the case of protein–protein docking, a receptor potential pre-calculated in a 3D grid is correlated with a ligand forcefield property defined at their atomic coordinates. Being L_i such atomic property, the correlation contribution of the ligand, g, can be expressed as a summatory function of the form: $\sum_i^N L_i\cdot\delta_{pi}$, where δ_{pi} is a delta function of the atom i centered at its coordinate position. From this expression the spherical coefficient of the ligand can be reduced to (see Supplementary Appendix for details):

$$\hat{g}_{lm}(r)=\sum_{i=1}^N L_i\int_{s^2}\delta_{pi}(ru)\cdot\overline{Y_{lm'}(u)}\cdot\mathrm{d}\sigma=\sum_{i=1}^N\frac{L_i}{r^2}\cdot\delta_{ri}(r)\cdot\overline{Y_{lm'}(u_i)} \tag{4}$$

Integrating Equation (4) into (3):

$$I^l_{mm'}=\int_0^\infty\hat{f}_{lm}(r)\cdot\sum_{i=1}^N L_i\cdot\delta_{ri}(r)\cdot Y_{lm'}(u_i)\cdot\mathrm{d}r=\sum_{i=1}^N L_i\cdot\hat{f}_{lm}(r_i)\cdot Y_{lm'}(u_i). \tag{5}$$

With this expression we avoid the implicit calculation of the SH coefficients of the ligand from a potential grid map as it is done with the receptor. However, we need to perform a summatory over all the ligand atoms, which can be costly if this number is too high. To overcome this problem and improve the overall efficiency, the integration is done over the atoms grouped

in the spherical layers (or shells) in where the SH transformation is defined. Thus, Equation (5) can be transformed into:

$$I^{l}_{mm'} = \sum_{i=0}^{C} \hat{f}_{lm}(\overline{r_i}) \cdot \sum_{j=0}^{n_{C_i}} L_j^{C_i} \cdot Y_{lm'}(u_j^{C_i}) \tag{6}$$

where C is the number of layers in the spherical representation, $\bar{r}_i$ is the average radii of the atoms in layer i, $L_j^{C_i}$ is the potential value (weight, charge, etc) of atom j in the set of atoms C_i on layer i,and $u_j^{C_i}$ is the spherical coordinates of this atom j. The centre of the spherical radial representation is established on the ligand centre of mass. Centering on the smallest protein reduces the potential distortions that could be produced in the radialized SH transforms of large proteins at points too far off the center. This is the reason why, in the relative translational search, we considered the ligand (smallest protein) fixed with respect to a mobile receptor (biggest protein), as opposed to standard docking FFT algorithms. Taking into account Equations (3) and (6), the final equation:

$$E(R) = FT^{-1}_{m,h,m'} \left[\sum_{l} d^{l}_{mh}\, d^{l}_{hm'} \sum_{i=0}^{C} \hat{f}_{lm}(\overline{r_i}) \cdot \sum_{j=0}^{n_{C_i}} L_j^{C_i} \cdot Y_{lm'}(u_j^{C_i}) \right] \tag{7}$$

gives us a new and efficient way to perform the rotational part of the rigid-body docking search. Notice that for different translational points the computation of Equation (7) only needs to recalculate the receptor SH coefficients, $\hat{f}_{lm}(r)$, the rest of the terms are pre-calculated. Moreover, it is possible to calculate different energy terms with only a single inverse FFT by taking advantage of the linearity of such transforms, hence reducing the overall computational cost.

To complete the 6D exhaustive docking search, the translational search was done implicitly by sampling uniformly the space with a fixed step size grid. We also reduced the translational space by simple masking procedures to prevent exploring points without physical meaning. To this end, we only considered points outside the surface of the receptor within a distance bigger than the minimum radius of the ligand and smaller than the maximal ligand radius. In other words, the sampling was constrained to avoid situations in which the ligand deeply penetrates into the receptor or in which the ligand molecule is not even in contact with the receptor. At this stage of presentation and validation of our methodology, we preferred this extensive translational sampling setup to maintain the general applicability. However, the translational space could be greatly reduced in particular if any geometrical constraint is introduced.

FRODOCK shares the use of SH to accelerate the rotational search with other methods such as HEX (Ritchie and Kemp, 2000). However, in addition to the original merging of SH with grid-based potentials, our approximation [Equation (7)] is different from other methods and quite novel in the protein–protein-docking field.

2.1 Interaction potentials

The binding energy during complex formation was approximated by three types of potentials: van der Waals, electrostatics and desolvation:

$$E = W_W E_W + W_E E_E + W_S E_S \tag{8}$$

where W_W, W_E and W_S weight the different contributions of the protein–protein interaction energy terms.

The van der Waals interactions are described by a Lennard-Jones 6–12 potential with most of the repulsive part truncated by a cut-off to reduce its extreme sensitivity to small conformational changes and hence to introduce some tolerance to conformational flexibility. This soft potential has been successfully used in protein–protein docking (Fernandez-Recio *et al*., 2002). Thus, the receptor soft potential $P_w(p)$ at a grid point p of atom i is given by

$$P_W(p) = \sum_{i}^{N} P_W^{(i)}(p) \tag{9}$$

and

$$P_W^{(i)}(p) = \begin{cases} P_W^{o(i)}(p) & \text{if} \quad P_W^{o(i)}(p) \le 0 \\ \frac{P_W^{o(i)}(p) P_{\max}}{P_W^{o(i)}(p) + P_{\max}} & \text{if} \quad P_W^{o(i)}(p) > 0 \end{cases}, \; P_W^{o(i)} = \frac{A_i}{r_{pi}^{6}} + \frac{B_i}{r_{pi}^{12}} \tag{10}$$

where r_{pi} denotes the distance between the coordinates of atom i to a given grid point p, $P_{\max}$ is a repulsive potential cut-off and A_i and B_i are constants. Using this expression and considering only the heavy atoms, the receptor van der Waals potential map was pre-computed using a generic C atom probe with radius 2.0 Å to model the ligand presence. By computing the receptor SH coefficients $\hat{f}_{lm}(r)$ from this grid, and making use of the ligand atoms mass as ligand scalar property (L_i), the van der Waals docking contribution for a given translational point was evaluated using Equation (7).

The electrostatic contribution was calculated in a similar way. To this end, only the partial charges are needed for the ligand, whereas for the receptor an electrostatic grid potential is approximated using a modified Coulomb's law. Such grid potential is defined by

$$P_E(p) = \sum_{i}^{N} P_E^{(i)}(p) \text{ where } P_E = \frac{q_i}{\varepsilon r_{pi}} \tag{11}$$

where $\varepsilon = 4r_{pi}$ is a distance-dependant dielectric constant and q_i are the receptor partial charges. The soft van der Waals potential used here allows certain overlap between atoms, which can result in unrealistic large electrostatic energy terms. To alleviate this, the electrostatic values were clamped in a range of ± 10 kcal/mol.

The docking desolvation energy is defined by the transfer of surface residues from water to protein–protein interface. Here this was estimated as a sum of per-atomic contributions proportional to the buried solvent accessible surface area, BSA; hence, the grid points of the receptor desolvation energy potential were calculated using

$$P_S(p) = \sum_{i}^{N} BSA_i(p)\sigma_i \tag{12}$$

where σ_i is the atomic solvation parameter for atom type i as previously calculated from linear fitting to experimental octanol/water transfer energies (Abagyan, 1997) and finally optimized for rigid-body docking (Fernandez-Recio *et al.*, 2004). To estimate the receptor buried surface upon binding we modeled the presence of the ligand by locating generic probes of 1.7 Å radius at all the grid points close to the receptor surface (defined by atoms with solvent accessible surface area, SASA > 0). The BSA was computed on the grid as the SASA difference with and without these atom probes. For the L property, we utilized the SASA of the ligand. In the same way, the desolvation contribution of the ligand with respect to the receptor was estimated but now using its accessible surface as reference. Thus, the total interaction desolvation energy is given by the sum of two correlation functions as Equation (12), each of them modeling the receptor–ligand and the ligand–receptor desolvation.

2.2 Implementation details

The method was implemented in three consecutive steps:

(1) *Generation of pre-calculated grid maps*: Three grid potentials were computed from the receptor coordinates (van der Waals, electrostatic and desolvation), whereas only one was needed from ligand coordinates (desolvation). Several *ad-hoc* tools have been developed in order to pre-compute such potential maps. Atomic properties such as van der Waals radius, charges etc. were taken from CHARMM 19 force field. The SASA calculations were performed using analytical methods (Busa *et al*., 2005).

(2) *Performing the docking 6D search*: Once the grid maps were pre-calculated, the docking was performed with a single tool called FRODOCK, which implements the new methodology presented in the Methods section for 6D exhaustive docking search. The rotational

and translational sampling resolutions were fixed to $5.6°(6 \times 10^4$ rotations) and 2 Å, respectively. These values were chosen in order to have a good balance between efficiency and accuracy. Since the translational search can eventually explore 10^5 points, we only considered the best four docking predictions for each translation point in order to avoid a large redundant set of solutions.

(3) *Clustering*: In each docking run, the results were clustered using a explicit comprehensive algorithm (Kozakov *et al*., 2005). Briefly, once the solution set was ranked according to their docking correlation, we formed clusters with all ligand-docking predictions within 5 Å RMSD distance from the first ranked solution (i.e. the lowest energy). The members of this first cluster were removed from the ranking, and we selected the next ranked solution as the centre for next cluster. We iteratively repeated this procedure until a predetermined number of clusters was achieved (to have a manageable number we fixed it to 10 000 clusters). Thus, each of the cluster centres will represent a different potential docking solution. Obtaining 10 000 clusters takes around 3–4 min on a standard pc. The clustering time drops below the minute for 2000 clusters which corresponds to a reasonable number in practical situations.

We further extended the efficiency of the method by parallelizing techniques. The chosen docking setup that splits the exhaustive search in translational and rotational parts is very suitable to run in parallel. Efficient rotational searches of independent translational scanned points can be easily farmed to different processors. Following this approach, we implemented an Message Passing Interface (MPI) version of FRODOCK, which, as shown below, allows for nearly linear performance gain depending on the number of processors used.

2.3 Benchmarks and parameter optimization

To test the method performance, we used the protein–protein benchmark 2.0 (Mintseris *et al*., 2005). This validation test set includes 84 protein–protein interactions with available 3D structures of the complexed and unbound forms, and contains examples of enzyme–inhibitor (E), antigen–antibody (A) and other complexes (O). According to the extension of conformational changes between the unbound and bound forms of the complex components, the test cases are also classified into rigid-body, medium and difficult cases. Since the latter are clearly out of scope of any rigid-body approximation, they have been excluded from our analysis. The remaining 76 protein–protein test cases (listed in Table S1) were used for testing our approximation. In all cases, the unbound forms of the subunits were used for docking. In order to optimize the parameters of the method, we used a random subset of the validation benchmark as training set. This set was formed by 15 test cases in which we found at least an acceptable solution within the top 1000. To balance its composition, five test cases of each class (E, A, O) have been considered.

Some of the most important parameters were the relative weights of the interaction energy terms W_W, W_E and W_S, for which we found optimal values of 1.0, 0.3 and 0.5, respectively. As expected, the optimization results suggested that the most important energetic contributions to the free binding energy in our rigid-body protein–protein docking are, in decreasing order: shape, desolvation and electrostatics. Other parameters, such as the radialization step size of the spherical layers used in the SH expansions (fixed to 1 Å), the bandwidth (32) and the translational step size (2 Å), were chosen to have a good efficiency without compromising the accuracy of the method. Bandwidths above 32 quickly deteriorate the performance and they did not improve the docking results. Note that for this type of initial rigid-body docking, a detailed shape description is probably not required, and certain degree of smoothness is even desirable in order to model small structural changes upon binding. To effectively test the method, the docking was repeated 50 times for each complex, with distinct random initial ligand orientations, thus avoiding pre-alignment situations with reference/original complexes.

An additional validation benchmark was compiled with available rigid-body test cases of the latest CAPRI experiments, which were not already included in the Weng's benchmark. This additional benchmark included targets T11 and T12 of the cohesin–dockerin complex of the cellulosome (PDB ID 1OHZ); T13 of the SAG1-antibody complex (PDB 1YNT); T14 of the protein Ser/Thr phosphatase-1 bound to MYPT1; T18 xylanase-TAXI complex (PDB ID 1T6G); T19 of ovine prion-Fab complex (PDB 1TPX); T25 of Arf1-GTP-ARHGap10 (PDB 2J59); T26 TolB/Pal (PDB 2HQS) and T27 Hip2 bound to a UBC9 (PDB 2O25). For targets T11 and T19, homology models previously built by ICM were used as ligand probes [details of modeling are described in (Fernandez-Recio *et al*., 2005)].

The ligand ($RMSD_L$) and interface ($RMSD_I$) root mean square deviations were computed following CAPRI criteria (Mendez *et al*., 2003). For computing the $RMSD_L$, the receptors were superimposed using all the Cα atoms, with the exception of the T3 case. In this case only the binding domain of the ligand was considered for the $RMSD_L$ calculation, as the other domain, which is not relevant for the interaction, is moved with respect to the bound reference state. A ligand or receptor residue is considered to be at the interface if any of its atoms is within 10 Å of an atom of the receptor or the ligand, respectively. Contacts are defined in the same way but within a shorter distance of 5 Å. None of the interface or contacts residues that fulfill such distance restraints have been excluded by any other criterion.

3 RESULTS

The global success rates shown in Figure 1 provided a first overall view of the performance of our new docking approximation on the unbound 76 targets from Weng's benchmark. We had on average a probability of 90% to find at least an acceptable solution ($RMSD_L \leq 10$ Å) within the 10 000 predictions made for all 50 runs of each docking case, and a probability of 67% for finding a medium quality solution ($RMSD_L \leq 5$ Å). These success rates smoothly diminished to 78 and 53% for finding acceptable and medium solutions, respectively, within top 1000. When only the top 100 were considered, FRODOCK maintains excellent success rates of 51 and 30%, respectively. In a closer view, it can be seen a clear different behavior depending on the complex type: whereas at least one acceptable solution can be found below the first 500 predictions (100% success rate) for enzyme–substrate cases (Fig. 1A; E, solid line), for antibody–antigen (dotted lines) the success rate drops to 78% for top 1000, and it falls even more (63%) for the other type (O, dashed line). These differences are more accentuated when looking at the top 100, in which we found success percentages of 92, 50 and 22% for E, A and O categories, respectively. It is well known that surface complementarity, which is the main docking driving force in this method (as in the majority of rigid docking methods), is a stringent criterion with enzyme–substrate and antibody–antigen docking cases. Nevertheless, it is much less effective with the O type docking cases, which contain the most heterogeneous and difficult test cases of the three categories. Similar observations can be made by looking at $RMSD_I$ instead of $RMSD_L$ (see Supplementary Fig. S1).

Several acceptable solutions have been found in almost all docking cases (see Supplementary Table S1 and S2). There are only five known difficult cases (1BGX, 1I4D, 1SBB, 1HE8, 1IB1) in which practically no acceptable solutions were found with $RMSD_L \leq 10$ Å or $RMSD_I \leq 4$ Å within 10 000 default predictions yielded by FRODOCK. There are also poor accuracy cases such as 1KLU in which some of the predictions are lost because they are ranked beyond the considered 10 000 predictions and/or their RMSD fell out of the limits to consider the solution as acceptable.

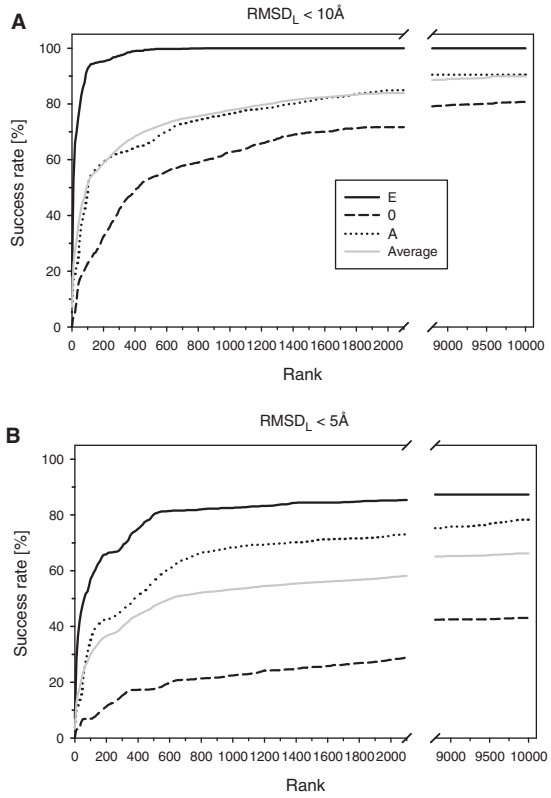

Fig. 1. Success rates obtained for the different types of unbound test cases included in the Weng's benchmark: enzyme–substrate (solid line), antibody–antigen (dotted), other (dashed) and average (grey). (**A**) The success rate is defined as the percentage of cases where at least an acceptable solution with a RMSD$_L \leq 10$ Å is detected with rank smaller than the number given in abscissas. (**B**) Here the success rate is defined to follow the medium solution assessment of CAPRI contest, i.e. have a RMSD$_L \leq 5$ Å.

Variations in the rank, RMSD, and f_{nat} can be also observed, showing the method dependence on the initial positions. Nevertheless, only relatively small variations were typically detected. For example for E cases (Table S1), average standard deviation values of 13, 0.55 and 0.04 are obtained for rank, RMSD$_L$ and f_{nat}, respectively. As expected, this variation becomes larger as predictions are less accurate; for O cases these values are 306, 0.89 and 0.06, respectively. However, in most of the cases, the predictions are essentially the same in all the 50 runs, thus demonstrating the robustness of the method. Only in a few cases, such as 1JPS, 1GCQ or 1MLO, different solution sets can be obtained with an average RMSD$_L$ variation larger than 3 Å. This, otherwise relatively minor, degeneration of solutions is likely a result of grid interpolation errors amplified by the clustering of the solutions.

We tested the comparative efficiency of FRODOCK in the case HyHel-5/lysozyme, previously used as a timing reference in a very recent version of HEX (Ritchie *et al*., 2008), which to our knowledge is the fastest protein–protein exhaustive docking search available. Our approximation is 10% slower than Hex, which takes ~27 min to perform this docking in a standard 2.2 MHz linux workstation. This docking tool slightly outperforms FRODOCK, most likely because it uses a more efficient two-stage protocol using 3D shape FFT scans with bandwidth 20 followed by 1D shape plus electrostatics rescoring with a bandwidth 30. Nevertheless, if FRODOCK employed the two equivalent shape and electrostatic terms used in HEX, the docking time could be reduced to 18 min. In this particular case, FRODOCK found the first acceptable solution at positions 32th and 30th, with and without desolvation, respectively. The different nature of the docking methodology and the diverse parameters employed, including different size of the translational and rotational steps, made difficult a thorough comparison of the docking performance. Nevertheless, it is clear that both methods have comparable performance at least in this representative example. On the same case, the FFT standard ZDOCK (versions 2.3.1 or 3.0.1) docking tool takes more than two hours (using a dense sampling with -d option). In any case, we implemented a parallel version that can take advantage of multiple processors. This version can speed up the docking calculations several folds (see Supplementary Fig. S2). HEX could follow similar strategy but its two-step protocol will be slightly more complex to parallelize than our direct translation space split in multiple processors.

3.1 Validation of docking method on CAPRI targets

For validation purposes, we also tested how FRODOCK would have performed in the CAPRI experiments. For that, we applied our docking protocol to the CAPRI targets 11, 12, 13, 18, 19, 25,26 and 27. In four of nine of the CAPRI test cases, the method predicted at least one acceptable solution within the top 10 (Table 1). Moreover, two of these cases (T19, T25) achieved medium quality by CAPRI standards, and another two (T12, T14) even achieved high accuracy. The goodness of these predictions can be observed in Figure 2B, D, F and G. Acceptable solutions for targets T11 (Fig. 2A) and T13 (Fig. 2C) were also found at 26th and 23rd position, respectively. In the three remaining cases we still found acceptable solutions within the top 500 positions (Fig. 2E, H, I). From these solutions, and using further refining protocols, it is feasible to improve their ranking to top positions. The simplest and most common way to achieve this would be through the screening of the predictions with available experimental information of the complex. For example, in the difficult case of TAXI-Niger Xilanase complex (T18), the successful CAPRI predictions were obtained using an experimental restraint for residues Glu79 and Glu170, which were known to be at the interface. If we use this restraint to filter out the FRODOCK solutions without these two residues at the interface, the first prediction is now ranked 10th with a RMSD$_I$ of 2.6 Å. Following a similar strategy, we obtained top ranked solutions for targets T26 and T27 (see details in Table 1). In summary, these results validate the excellent performance of our initial exhaustive search-docking tool.

4 DISCUSSION

We have developed an initial-stage rigid-body docking program called FRODOCK, which optimizes van der Waals, desolvation, and electrostatics interaction potentials by using a new fast rotational docking algorithm based on SH combined with a systematical translational search.

We have shown that, on a standard benchmark set, our new approach can place an acceptable solution (RMSD$_L \leq 10$ Å) within the top 100 solutions in more than half of the cases (51%), and within the top 20 solutions in almost a third of the cases (30%). These results

Table 1. Results obtained with CAPRI test cases

	Ligand RMSD										Interface RMSD									
	< 10 Å					< 5 Å					< 4 Å					< 2.5 Å				
	N	Rank	$RMSD_L$	f_{nat}	f_{not}	*N*	Rank	$RMSD_L$	f_{nat}	f_{not}	*N*	Rank	$RMSD_I$	f_{nat}	f_{not}	*N*	Pos	$RMSD_I$	f_{nat}	f_{not}
T11	16	26	9.81	0.14	0.59	–	–	–	–	–	6	241	3.01	0.37	0.49	–	–	–	–	–
T12	13	**1**	1.94	0.96	0.33	1	**1**	1.94	0.96	0.33	9	**1**	0.89	0.96	0.33	3	**1**	0.89	0.96	0.33
T13	70	23	7.90	0.40	0.57	5	68	3.91	0.72	0.21	87	23	2.66	0.40	0.57	18	68	0.94	0.72	0.21
T14	8	**1**	1.47	0.54	0.16	1	**1**	1.47	0.54	0.16	5	**1**	0.77	0.54	0.16	2	**1**	0.77	0.54	0.16
T18	6	261	7.36	0.70	1.10	–	–	–	–	–	8	261	2.62	0.70	1.10	1	3049	1.95	0.59	0.63
T18[a]	2	**10**	7.36	0.70	1.10	–	–	–	–	–	3	**10**	2.62	0.70	1.10	1	39	1.95	0.59	0.63
T19	13	**10**	6.86	0.44	0.48	1	401	4.89	0.73	0.19	16	**5**	3.45	0.46	0.79	1	401	1.24	0.73	0.19
T25	31	**3**	3.36	0.69	0.21	4	**3**	3.36	0.69	0.21	35	**3**	1.63	0.69	0.21	9	**3**	1.63	0.69	0.21
T26	11	224	3.92	0.29	0.33	2	224	3.92	0.29	0.33	9	224	2.08	0.29	0.33	3	224	2.08	0.29	0.33
T26[b]	7	32	3.92	0.29	0.33	2	32	3.92	0.29	0.33	7	32	2.08	0.29	0.33	3	32	2.08	0.29	0.33
T27	25	468	3.83	0.59	0.24	1	468	3.83	0.59	0.24	46	335	2.25	0.61	0.63	46	335	2.25	0.61	0.66
T27[c]	4	**2**	8.31	0.41	0.59	–	–	–	–	–	9	**2**	2.33	0.41	0.59	3	**3**	1.51	0.56	0.63

N denotes the number of solutions, and rank the position of the first solution found within the RMSD limit shown in the top of the column. The f_{nat} and f_{not}, the interface ratios were calculated as described in Mendez *et al*. (2003). In all cases the RMSDs were calculated using *C* atoms.

[a]Filter results of T18 considering only the predictions in which residue E70 of TAXI is at <5 Å from residue E179 of Niger Xilanase.

[b]Filter results of T26 considering only the predictions in which residue H246 and T292 of TolB are present in the complex contact interface.

[c]Filter results of T27 considering only the predictions in which residue K14 of Hip2 is at <5 Å from residue C93 of Ubc9.

Test cases in bold had at least a solution ranked in the top ten predictions.

are very competitive, as compared to other exhaustive protein-protein docking approaches. In a comparative blind docking on the same cases of the Weng's benchmark, HEX found 16 acceptable solutions within the top 20 orientations, and 24 cases within the top 100 (Ritchie *et al.*, 2008). FRODOCK results were better, finding 20 and 38 acceptable solutions within the same ranges (see Table S1). HEX also significantly improved the number of acceptable solutions by constraining the search to focus the calculation around the receptor binding site, e.g. up to 28/42 with one constraint. Despite evident benefits of employing constraints during the search, in terms of complexity reduction and enhanced performance, in this work we have chosen to focus on the most general and challenging problem defined by the blind 6D exhaustive docking search.

Apart from procedural differences, ZDOCK and FRODOCK have similar docking accuracy. On the 76-case docking test used here, ZDOCK 2.3 identified 14 cases with a hit ranked in the top 20 orientations, and 24 cases with a hit in the top 100 [see Table 1 of Pierce and Weng (2007)], where hit is defined as a solution having an $RMSD_I \leq 2.5$ Å from the complex reference structure. Here we obtained 13 and 24 cases (see Table S2), respectively, which are comparatively very similar. In addition to shape, electrostatics and desolvation, ZDOCK 3.0 (Mintseris *et al*., 2007) also considers statistical pair potentials, which clearly improved the success rates to achieve 19 cases with a hit within the first 20 predictions [Table II of Pierce and Weng (2008)]. However, the success rates for hits and near-hits ($RMSD_I$ <4 Å) in the top 100 of ZDOCK 3.0 [Figure 1 of Pierce and Weng (2008)] and FRODOCK (see Fig. S1 of Supplementary Data) are both very close to ∼50%. The reason of this ZDOCK over-performance obtaining top 10 hits can be also partially attributed to procedural differences such as the use of search constraints (ZDOCK 2.3 blocks non-CDR regions in the antibody–antigen cases), the definition of interface residues in the evaluation (we strictly follow CAPRI convention), statistical differences in the number of runs (we explore 50 distinct random poses per case) and different sampling sizes (6° and 1 Å for ZDOCK, 5.4° and 2 Å for FRODOCK). However, the considerations of statistical pair potentials have proven to be a successful strategy to improve the number of near-native docked conformations (Kozakov *et al*., 2006; Mintseris *et al*., 2007). Therefore future versions of FRODOCK will pursue its inclusion. Compared to pyDock (Cheng *et al*., 2007), from which FRODOCK inherits part of the potential term definitions, we found again similar performances. Using a combined set of docking poses previously generated by FTDOCK and ZDOCK in a similar benchmark (with four fewer cases than here), pyDock was able to find 23 cases with acceptable solutions within the top 20 orientations, and 35 cases within the top 100. The presented approximation yielded slightly worst results, but the advantage is that they were directly obtained from a fast exhaustive docking search.

The robustness of this novel docking tool was confirmed on another benchmark formed by CAPRI contest test cases. From the exhaustive search, in four cases we obtained predictions within the top 10 and in other two we obtained very close top positions. The remaining three cases can improve their ranking (below 500 positions) to top positions simply by filtering with a few experimental restraints. Taking into account that FRODOCK is an initial exhaustive search tool, these results are quite promising.

Another advantage of this method is its efficiency. The sequential version of FRODOCK is slightly slower than HEX, the fastest exhaustive docking approach that outperformed the classical translational FFT-based methods. Nevertheless, a parallel version of FRODOCK, which can take advantage of multiprocessor architectures and the newer wave of multi-core processors, reduces several folds the docking time. Note that there are much faster alternatives, e.g. Patchdock (Schneidman-Duhovny *et al*., 2005),

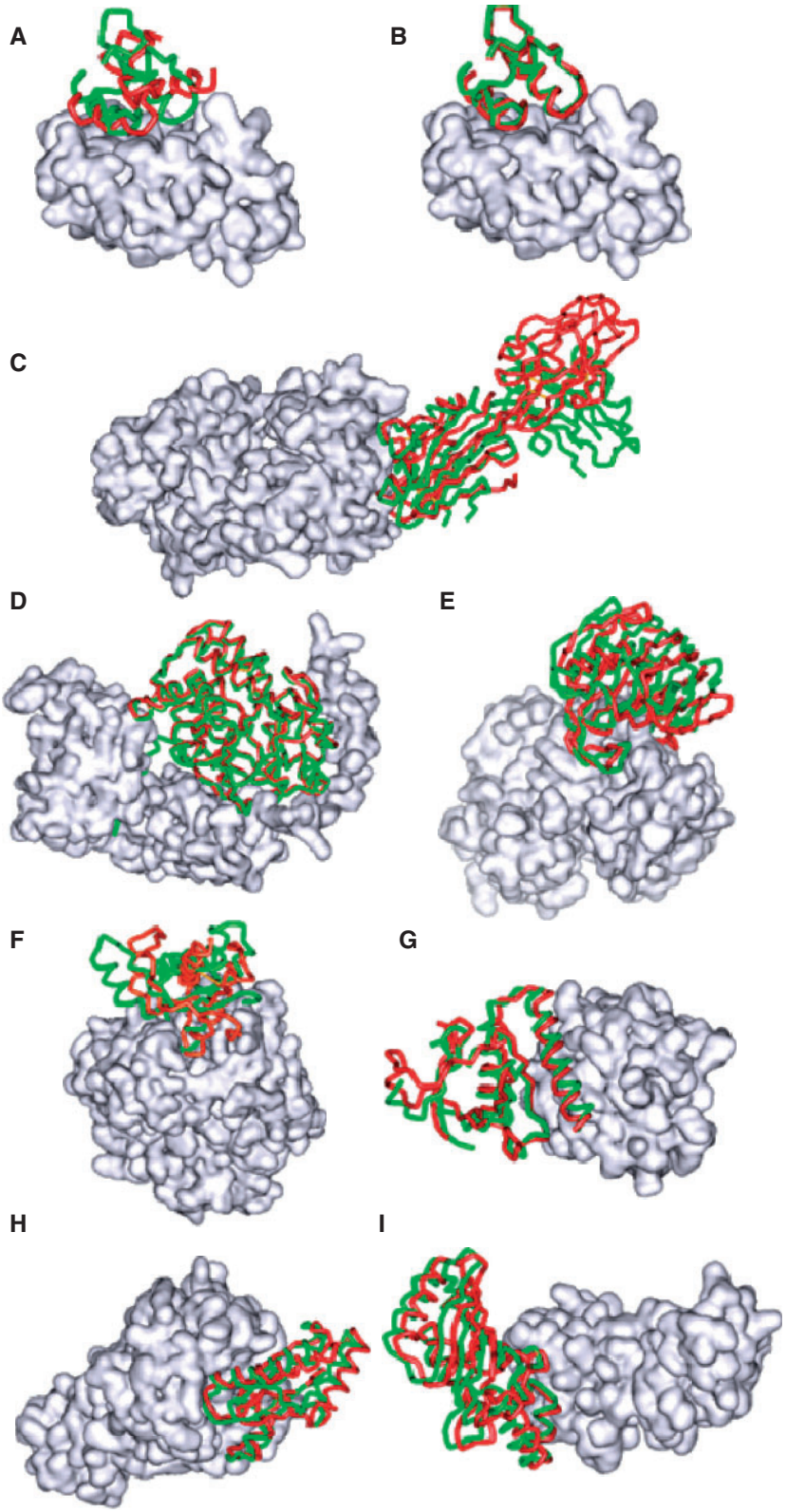

Fig. 2. Docking predictions for rigid-body CAPRI targets T11 (**A**); T12 (**B**); T13 (**C**); T14 (**D**), T18 (**E**), T19 (**F**), T25 (**G**); T26 (**H**) and T27 (**I**). The orientations of predicted ligand (in red) and the corresponding crystal structure (in green) are shown after superposition of their receptors (surface representation in gray). The displayed predictions correspond to the first ranked acceptable solution of Table 1.

but they do not perform an exhaustive search, and therefore there is always the possibility of losing the correct docking pose.

In summary, the competitive docking accuracy and efficiency achieved by our approach can eventually open up new application windows, especially regarding large-scale structural modeling of protein complexes (Aloy and Russell, 2006; Zhu *et al*., 2008). In this context, a tool capable of reducing the protein–protein docking search to a few minutes will be critical to effectively address future high-throughput approaches. Further method improvements will include merging with scoring protocols such ICM (Abagyan and Totrov, 1994) and pyDock (Cheng *et al*., 2007), together with local refinement and rescoring of the atomic coordinates of the FRODOCK predicted complex in order to generate more realistic solutions.

Funding: Spain grants BFU2007-65977 and CAM-BIO-0214-2006 (to P.C.) and BIO2008-02882 (to J.F.R.) and by NIH grant R01-GM071872 (to R.A.).

Conflict of Interest: none declared.

REFERENCES

Abagyan,R. (1997) *Protein Structure Prediction by Global Energy Optimization*. Kluwer Academic Publisher, Dordrecht.

Abagyan,R. and Totrov,M.M. (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, **235**, 983–1002.

Aloy,P. and Russell,R.B. (2006) Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.*, 7, 188–197.

Bonvin,A.M. (2006) Flexible protein-protein docking. *Curr. Opin. Struct. Biol.*, **16**, 194–200.

Busa,J. *et al.* (2005) ARVO: a Fortran package for computing the solvent accessible surface area and the excluded volume of overlapping spheres via analytic equations. *Comp. Phys. Commun.*, **165**, 59–96.

Camacho,C.J. and Vajda,S. (2002) Protein-protein association kinetics and protein docking. *Curr. Opin. Struct. Biol.*, **12**, 36–40.

Chen,R. *et al*. (2003) ZDOCK: an initial-stage protein-docking algorithm. *Proteins*, **52**, 80–87.

Cheng,T.M. *et al*. (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, **68**, 503–515.

Deremble,C. and Lavery,R. (2005) Macromolecular recognition. *Curr. Opin. Struct. Biol.*, **15**, 171–175.

Fernandez-Recio,J. *et al.* (2002) Soft protein-protein docking in internal coordinates. *Protein Sci.*, **11**, 280–291.

Fernandez-Recio,J. *et al.* (2004) Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.*, **335**, 843–865.

Fernandez-Recio,J. *et al*. (2005) Improving CAPRI predictions: optimized desolvation for rigid-body docking. *Proteins*, **60**, 308–313.

Gabb,H.A. *et al*. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, **272**, 106–120.

Garzon,J.I. *et al*. (2007) ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics*, **23**, 427–433.

Gray,J.J. (2006) High-resolution protein-protein docking. *Curr. Opin. Struct. Biol.*, **16**, 183–193.

Heifetz,A. *et al*. (2002) Electrostatics in protein-protein docking. *Protein Sci.*, **11**, 571–587.

Katchalski-Katzir,E. *et al*. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.

Kovacs,J.A. and Wriggers,W. (2002) Fast rotational matching. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 1282–1286.

Kovacs,J.A. *et al*. (2003) Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Crystallogr. D Biol. Crystallogr.*, **59**, 1371–1376.

Kozakov,D. *et al*. (2005) Optimal clustering for detecting near-native conformations in protein docking. *Biophys. J.*, **89**, 867–875.

Kozakov,D. *et al*. (2006) PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, **65**, 392–406.

Lensink,M.F. *et al*. (2007) Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins*, **69**, 704–718.

Mendez,R. *et al*. (2003) Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, **52**, 51–67.

Mendez,R. *et al*. (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, **60**, 150–169.

Mintseris,J. *et al*. (2005) Protein-protein docking benchmark 2.0: an update. *Proteins*, **60**, 214–216.

Mintseris,J. *et al*. (2007) Integrating statistical pair potentials into protein complex prediction. *Proteins*, **69**, 511–520.

Moont,G. *et al*. (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**, 364–373.

Pierce,B. and Weng,Z. (2007) ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins*, **67**, 1078–1086.

Pierce,B. and Weng,Z. (2008) A combination of rescoring and refinement significantly improves protein docking performance. *Proteins*, **72**, 270–279.

Ritchie,D.W. (2008) Recent progress and future directions in protein-protein docking. *Curr. Protein Pept. Sci.*, **9**, 1–15.

Ritchie,D.W. and Kemp,G.J. (2000) Protein docking using spherical polar Fourier correlations. *Proteins*, **39**, 178–194.

Ritchie,D.W. *et al*. (2008) Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics*, **24**, 8.

Schneidman-Duhovny,D. *et al*. (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res*., **33**, W363–W367.

Vakser,I.A. and Kundrotas,P. (2008) Predicting 3D structures of protein-protein complexes. *Curr. Pharm. Biotechnol*., **9**, 57–66.

Zhu,Z. *et al*. (2008) Large-scale structural modeling of protein complexes at low resolution. *J. Bioinform. Comput. Biol.*, **6**, 789–810.