Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

Protein Science

Monte Carlo refinement of rigid-body protein docking structures with backbone displacement and side-chain optimization

Stephan Lorenzen and Yang Zhang

Protein Sci. 2007 16: 2716-2725; originally published online Oct 26, 2007;
Access the most recent version at doi:10.1110/ps.072847207

References This article cites 46 articles, 17 of which can be accessed free at:
http://www.proteinscience.org/cgi/content/full/16/12/2716#References

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**

Notes

To subscribe to *Protein Science* go to:
http://www.proteinscience.org/subscriptions/

© 2007 Cold Spring Harbor Laboratory Press

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

Monte Carlo refinement of rigid-body protein docking structures with backbone displacement and side-chain optimization

STEPHAN LORENZEN[1–3] AND YANG ZHANG[1,2]

[1]Center for Bioinformatics, University of Kansas, Kansas 66047, USA
[2]Department of Molecular Biosciences, University of Kansas, Kansas 66047, USA
[3]Macromolecular Modelling Group, Free University, 14195 Berlin, Germany

(RECEIVED March 2, 2007; FINAL REVISION August 29, 2007; ACCEPTED August 30, 2007)

Abstract

Structures of hitherto unknown protein complexes can be predicted by docking the solved protein monomers. Here, we present a method to refine initial docking estimates of protein complex structures by a Monte Carlo approach including rigid-body moves and side-chain optimization. The energy function used is comprised of van der Waals, Coulomb, and atomic contact energy terms. During the simulation, we gradually shift from a novel smoothed van der Waals potential, which prevents trapping in local energy minima, to the standard Lennard-Jones potential. Following the simulation, the conformations are clustered to obtain the final predictions. Using only the first 100 decoys generated by a fast Fourier transform (FFT)-based rigid-body docking method, our refinement procedure is able to generate near-native structures (interface RMSD <2.5 Å) as first model in 14 of 59 cases in a benchmark set. In most cases, clear binding funnels around the native structure can be observed. The results show the potential of Monte Carlo refinement methods and emphasize their applicability for protein–protein docking.

Keywords: protein–protein docking; fast Fourier transformation; scoring; refinement; smoothed potential

Most proteins associate with other proteins to fulfill their function in the living cell. The understanding of the function of a protein thus crucially depends on structural knowledge of its complexes with interaction partners. However, since many protein–protein complexes are hard to crystallize, experimental structural data of the complexes is often missing. With the structures of the respective monomers available in the Protein Data Bank (PDB) (Berman et al. 2000), it is the task of computational protein–protein docking to reassemble the monomers into complexes. Generally, there are three levels of degrees of freedom in complex formation from monomers: First, both chains can be treated as rigid bodies and translated and rotated against each other, which results in six degrees of freedom. Further on, the conformation of side chains can be adjusted to optimize the interface between the two monomers. The highest number of degrees of freedom is introduced by the possibility of backbone rearrangements upon complex formation.

Thus, the "redocking" of monomers obtained from a cocrystallized complex structure (bound–bound docking) is much easier than the prediction of a complex structure from independently crystallized monomers (unbound–unbound docking). While for the bound–bound cases, shape complementarity as the major determinant seems to be sufficient for reliable results (Norel et al. 1994), the more difficult unbound–unbound cases have to be considered as real-world examples. The introduction of the fast Fourier transform (FFT) technique by Katchalski-Katzir et al. (1992) made the computational search of the six-dimensional conformation space possible for the first time. Until today, most rigid-body docking algorithms depend on the FFT algorithm (Gabb et al. 1997; Mandell

Reprint requests to: Yang Zhang, 2030 Becker Drive, Lawrence, KS 66047, USA; e-mail: yzhang@ku.edu; fax: (785) 864-5558.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.072847207.

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

et al. 2001; Heifetz et al. 2002; Chen et al. 2003a; Tovchigrechko and Vakser 2006).

While the initial scanning of possible docking conformations by these methods is quick and efficient, the major problem lies in the scoring of thousands of decoys and in the missing treatment of structural flexibility. A rescoring of docking decoys using more elaborate energy functions and filters (Murphy et al. 2003; Camacho et al. 2006) was shown to improve the selection of models. Structural flexibility can be treated by soft potentials (Vakser 1995, 1996), multicopy representations of side chains (Lorber et al. 2002), or flexible loops (Zacharias 2003; Bastard et al. 2006). Initial results of docking programs can subsequently be refined by conjugate gradient minimization (Tovchigrechko and Vakser 2005) and are then usually clustered to identify near-native structures. The basic idea behind clustering is the notion that in most cases native structures lie in broad energy wells (Camacho et al. 1999), and it can be shown that clustering significantly enhances the performance of docking algorithms (Comeau et al. 2004; Lorenzen and Zhang 2007).

A different approach to the docking problem is the Monte Carlo method. Instead of screening all possible conformations with a Fourier-transformable energy function, random starting decoys are refined by applying random translational and rotational moves and deciding on their acceptance using the Metropolis criterion (Metropolis et al. 1953; Gray et al. 2003; Schueler-Furman et al. 2005). While FFT methods have the advantage of great speed and complete sampling of the conformational space, Monte Carlo methods are able to generate more physical decoy distributions, can involve arbitrary energy functions, and might allow for structural flexibility. However, an exhaustive sampling of conformational space can be very time consuming, if not beyond computational possibility.

Here, we present a hierarchical approach of initially scanning the conformational space by a quick FFT-based ZDOCK run, followed by the refining of the resulting top 100 decoys by ROTAFIT, a Monte Carlo refinement program, combined with subsequent clustering. Of 59 cases in a common benchmark set (Chen et al. 2003b), we are able to obtain 14 cases with a near-native solution (interface RMSD below 2.5 Å) as the highest scoring model, compared with six near-native solutions obtained as first hits by ZDOCK.

Similar two-step approaches have been realized by Fernandez-Recio and colleagues (ICM-DISCO) (Fernandez-Recio et al. 2002, 2003) and Gray and colleagues (RosettaDock) (Gray et al. 2003). In ICM-DISCO, the investigators start with a rigid-body docking searched by a pseudo-Brownian Monte Carlo simulation (Abagyan and Totrov 1994). The second step is the Monte Carlo refinement of ligand side-chain torsion angles. In RossetaDock, the authors start from random ligand-receptor orientations, followed by a low-resolution rigid-body docking. In a second step, RosettaDock optimizes side chains and rigid-body orientations simultaneously, based on the simulated annealing Monte Carlo simulation (Kirkpatrick et al. 1983). In both of the approaches, the rigid-body docking is performed by Monte Carlo searches. In our approach, however, the first step of rigid-body conformation is taken directly from the FFT-based docking (Chen et al. 2003a), which is supposed to cover the whole complex space and may include a larger diversity of initial conformations. In the second step, while ICM-DISCO uses the pseudo-Brownian Monte Carlo approach and RosettaDock the simulated annealing, we exploit the replica-exchange Monte Carlo simulation (Swendsen and Wang 1986), a method that has been demonstrated to be more efficient in biomolecule simulations than other Monte Carlo methods (Gront et al. 2000; Zhang et al. 2002).

Another novelty in our approach is the smoothing of the Lennard-Jones potential and a gradual roughening to finally reach the original potential. Since in unbound–unbound docking, the conformations of side chains, and to some extent, also the main chain, often differ from the bound conformation, the ‘‘hard’’ Lennard-Jones potential needs to be smoothed in some way to allow some overlap between the structures or tolerate inaccuracies. Previous methods to smooth the Lenard-Jones potential for small atom distances in docking refinement include linear extrapolation below some threshold (Gray et al. 2003) or truncation above a maximal positive energy (Fernandez-Recio et al. 2003). Our smoothing approach was inspired by Zacharias's study (Zacharias et al. 1994; Riemann and Zacharias 2005) (see Materials and Methods and Fig. 8, below). In addition to dampening the positive part of the potential, the minimum well is also broadened, which leads to a less restrictive location of the minimum, and thus enables a smooth direction of the structure toward the true minimum in the course of roughening the potential.

Results

The principle of our method is summarized in Figure 1. First, the conformational space is scanned by ZDOCK (Chen et al. 2003a), a rigid-body FFT docking program. Prior to docking, structures were translated and rotated randomly. The top-scoring decoys are then refined by a replica-exchange Monte Carlo approach (Swendsen and Wang 1986). To ensure a smooth refinement process starting from the rigid-body decoys, the simulation starts with a smoothed van der Waals energy function, which is gradually roughened to finally represent the standard 12/6 Lennard-Jones potential (see Materials and Methods). In this way, the repulsive part is small at the beginning of the simulation when incorrect side-chain conformations might prevent perfect geometrical complementarity and

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

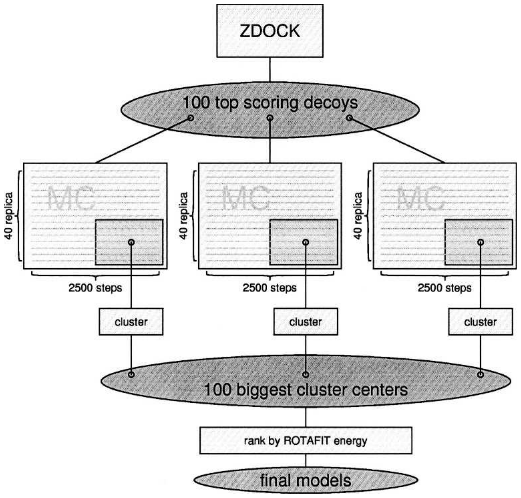

Figure 1. Flowchart of the ROTAFIT refinement procedure: Starting orientations are obtained from rigid-body docking by ZDOCK (Chen et al. 2003a). Each decoy is then refined by a Monte Carlo run with subsequent clustering. The resulting models are ranked by their energy.

a rough scan is desired. During the simulation, it becomes more restrictive gradually to induce a tight fit between receptor and ligand and select correct side-chain conformations. The median energies of the lowest temperature replicas for the 59 cases versus the simulation time are shown in Figure 2. Note the jumps in the energy value at the times when the energy function is roughened (every 150 steps). At around half the simulation length (1250 steps), the energy values start to be stable, and longer simulations did not obviously change the results. Generally, our replica exchange method leads to a complete exploration of the simulated temperature range by all replicas in the course of the simulation. As an illustrative example, the temperature progression of two replicas of case 1ACB is plotted in Figure 3. Initially, both replicas start exploring the complete range of temperatures. Only after around 1600 steps, Replica 16 stays in the colder temperature regions.

Performance of ROTAFIT versus ZDOCK

Table 1 summarizes the results of the procedure on the benchmark set. For all 59 cases, the interface RMSD of the top-1 and top-5 and the best decoys by ZDOCK (columns 2–4) and the subsequent ROTAFIT refinements (columns 7–13) are listed. In the bottom of the table, we also list the number of cases with an interface RMSD below some thresholds as well as the average RMSD for the best 10–50 cases.

In most cases, the ROTAFIT Monte Carlo simulations can significantly improve the interface RMSD. For example, the average interface RMSD of the best decoy in the best 50 targets by ZDOCK is 3.6 Å. During the Monte Carlo refinements, the average RMSD decreases to 2.6 Å. If we consider the best 40/30/20/10 targets, the average RMSD of the best decoys is reduced from 2.4/1.4/0.9/0.7 Å to 1.6/0.9/0.5/0.3 Å, respectively.

The ranking of Monte Carlo decoys by ROTAFIT energy also performs better than the ranking of ZDOCK decoys by the ZDOCK energy. As shown in Table 1, the average interface RMSD of the first model in the best 20 targets by ZDOCK is 5.3 Å; while based on the ROTAFIT ranking, the average RMSD of the first model is reduced to 2.9 Å. If we count the number of targets that have the first model with an interface RMSD <2.5 Å, ZDOCK has six cases and ROTAFIT has 13 cases in this criterion. The structural clustering of the low-temperature replica decoys can further improve the ROTAFIT ranking. As shown in column 10 of Table 1, the average interface RMSD of the first cluster in the best 20 targets is 2.4 Å. For 14 targets, the first cluster center has an interface RMSD of <2.5 Å, and for 19 targets, at least one of the first five cluster centers has an interface RMSD of <2.5 Å (Table 1).

Figure 4 presents three representative examples of interface RMSD versus ZDOCK-score of the ZDOCK decoys (left column) and interface RMSD of refined decoys versus ROTAFIT energy (right column). These data show that the top-ranking decoys in ROTAFIT have a much lower interface RMSD than that of ZDOCK decoys, which explains why the average ranking of ROTAFIT performs better than that of ZDOCK shown in Table 1.

Figure 5 summarizes the comparison of ROTAFIT and ZDOCK results for the first (left) and top five (right) decoys. In two cases, the interface RMSD of the top-1 decoy by ROTAFIT is increasing significantly (1DFJ: 2.5 versus

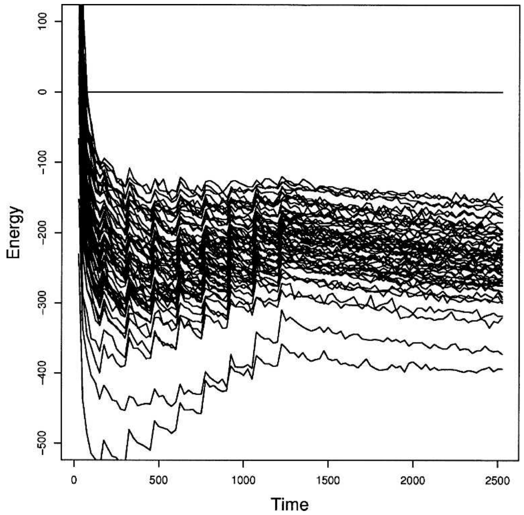

Figure 2. Median energies of the lowest temperature replica versus simulation time (in the unit of moving attempts) for all 59 benchmark protein complexes. Jumps in the energy function in every 150 moving attempts indicate a roughening of the energy function (see Materials and Methods).

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

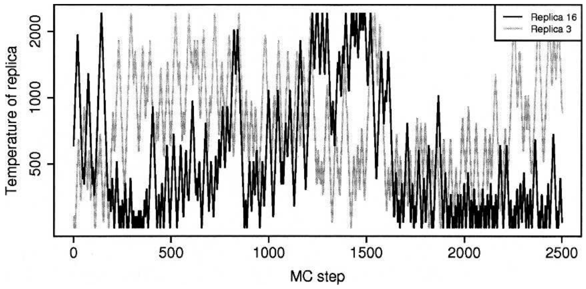

Figure 3. Temperature progression for two replicas of decoy 4 of 1ACB (RMSD 1.5 Å at the end of the simulation). Replica 16 (black) initially explores the complete temperature range and, due to the favorable conformation and energy, stays at low-temperature ranges after around step 1600. Replica 3 (gray) keeps traveling through the complete temperature range in the whole course of the simulation.

14.9 and 1WQ1: 2.73 versus 5.57), whereas in eight cases, ROTAFIT obviously outperforms ZDOCK (1ACB, 1ATN, 1BRC, 1KXQ, 1MEL, 1NCA, 1UDI, 2SIC). The solid lines show the average interface RMSD of the best-performing ROTAFIT cases versus the average interface RMSD of the best-performing ZDOCK cases. Numbers indicate the average of the best 10, 20, 30, 40, and 50 cases. All points are below the diagonal line, which again shows a clear improvement of interface RMSD by ROTAFIT compared with the initial ZDOCK input.

Comparison with other methods

Several groups reported methods to refine initial docking decoys. For example, starting from conformations generated by ZDOCK, Li et al. (2003) use short CHARMM energy minimizations to refine and rescore the initial decoys (called RDOCK). The final scoring function is comprised of electrostatic interaction energies between ligand and receptor as well as an ACE term, with filtering for unfavorable van der Waals energies. Using a benchmark set of 49 protein complexes, the authors report 21 cases with solutions within the first 100 ZDOCK decoys and were able to rank 11 of them as top-ranking decoy and two others within the top five decoys (Li et al. 2003). However, the ZDOCK version used by Li et al. (2003) was 2.1, whereas we are using version 2.3. To have a fair comparison, we downloaded and ran RDOCK based on the same set of 100 decoys (not counting the ones discarded because of unfavorable van der Waals interactions). We obtained 12 hits in the rank-1 decoys and 19 hits in the best of top five decoys, which are comparable to 14 and 19 hits by ROTAFIT. The average RMSD of the rank-1 decoys by ROTFIT is slightly lower, i.e., 0.6/2.4/4.7/6.8/8.8 Å for the top 10/20/30/40/50 cases versus 1.0/3.4/5.9/7.8/9.6 Å by RDOCK. The detailed RDOCK results are listed in columns 5 and 6 in Table I.

Gray et al. (2003) tested RosettaDock on a benchmark of 54 proteins, a subset of the benchmark used in this study. The investigators report seven cases with the first model of <5 Å ligand C_α RMSD after superimposing the receptor structures and a further nine cases with this criterion within the first five models. Using ROTAFIT, we found 12 cases with a complete ligand RMSD of <5 Å as top-1 hit and 16 hits within the first five decoys (see columns 12 and 13 of Table 1).

Performance of ROTAFIT in CAPRI round 11

As a blind test for our algorithm, we took part in round 11 of the community-wide protein–protein docking experiment, CAPRI (Janin et al. 2003; Janin 2005; Mendez et al. 2005). The target was a complex of Huntingtin-interacting protein 2 (Hip2), a ubiquitin-conjugating enzyme with Ubc9, a SUMO transferase (Desterro et al. 1997; Johnson and Blobel 1997; Schwarz et al. 1998). SUMO, a ubiquitin-like protein of 101 residues (Muller et al. 2001), is transferred from cystein 93 of Ubc9 to lysin 14 of Hip2 (Pichler et al. 2005). In a crystal structure of Ubc9 with another substrate, RanGAP (PDB code 1KPS), the acceptor Lysine residue lies in a pocket with contacts to residues Asp127, Pro128, Ala129, and Tyr87 of Ubc9 (Bernier-Villamor et al. 2002). Residues 129–135 of Ubc9 have also been shown to interact with Ubc9 substrates by chemical-shift analysis (Lin et al. 2002).

We first performed a rigid-body docking with ZDOCK and filtered for decoys with a distance of <10 Å between Lys14 of Hip2 and Cys93 of Ubc9. After filtering the 10,000 top-ranking models generated with default and high-density rotational sampling, we obtained 31 and 101 docking structures, respectively, which are used as the starting conformations in the subsequent ROTAFIT refinements. During the simulations, we also add a distance restraint proportional to the square of the distance between N_ε of Lys14 and S_γ of Cys93 to our potential. Here, one issue is the determination of the weight factor of the restraint, which should be strong enough to move the monomer structures, but should not dominate the other inherent physics-based ROTAFIT potentials. To do this, we tested three weight factors of 0.1, 0.5, and 5. Figure 6A shows the Lys–Cys distance distribution of the simulated ROTAFIT decoys with various weights compared with the original selected ZDOCK decoys. A restraint weight factor of 0.5 seems to work best, which is sufficient to guide the monomer movement, but does not modify significantly the energy distribution of the final models (Fig. 6B). So 0.5 has been exploited in our simulations. Finally, the ROTAFIT decoys were clustered with a cutoff of 3.5Å ligand RMSD, which are ranked based on the cluster size. Figure 7 shows our first submitted model to CAPRI,

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

Table 1. *Performance of ZDOCK, RDOCK, and ROTAFIT on 59 benchmark protein complexes*

	ZDOCK			RDOCK		ROTAFIT decoys			ROTAFIT cluster			
	I-1[a]	I-5[b]	I-B[c]	I-1[a]	I-5[b]	I-1[a]	I-5[b]	I-B[c]	I-1[a]	I-5[b]	L-1[d]	L-5[e]
1A0O	12.4	9.1	6.0	12.9	12.9	12.6	11.9	4.7	12.5	12.3	26.8	26.8
1ACB	4.8	1.9	0.9	1.3	1.2	1.4	1.0	0.7	1.5	0.9	6.7	3.5
1AHW	16.0	9.0	1.5	17.4	1.0	20.4	14.7	0.9	27.1	15.2	57.7	43.9
1ATN	12.1	12.1	0.5	0.5	0.5	0.5	0.5	0.3	0.4	0.4	1.4	1.4
1AVW	17.4	1.7	1.6	17.5	1.8	17.8	7.5	0.5	17.6	1.7	49.9	6.8
1AVZ	11.8	11.6	6.9	10.3	9.8	11.8	10.0	6.2	12.3	11.7	30.6	30.6
1BQL	7.9	1.1	1.0	1.4	1.0	19.1	0.7	0.5	14.4	0.7	47.9	1.6
1BRC	5.7	2.4	1.5	4.6	2.6	2.4	2.3	0.8	2.4	2.3	10.3	9.6
1BRS	8.0	8.0	5.7	14.5	5.7	14.9	9.7	4.0	11.4	11.4	35.5	35.5
1BTH	9.4	7.7	3.4	7.3	6.6	8.9	6.6	3.2	6.6	4.2	17.5	10.8
1BVK	18.3	15.9	9.5	14.4	11.6	17.8	17.8	8.1	17.8	15.2	63.8	63.7
1CGI	8.7	2.4	2.1	12.4	2.1	8.8	7.7	1.7	9.8	2.6	23.1	6.4
1CHO	9.3	1.6	1.0	1.5	1.3	6.9	1.5	0.6	6.9	1.1	15.1	3.8
1CSE	11.6	5.8	4.1	17.3	4.1	11.1	5.7	3.2	11.2	5.8	38.6	13.3
1DFJ	2.5	2.0	1.4	2.3	2.0	15.2	14.9	1.0	14.9	14.9	24.3	24.2
1DQJ	14.2	13.5	9.0	14.6	11.5	14.5	14.1	5.9	14.6	14.1	31.6	30.1
1EFU	14.0	13.9	11.3	28.1	12.5	29.1	12.5	10.9	29.2	13.4	64.6	50.7
1EO8	9.7	9.7	9.6	17.6	9.6	17.2	15.0	8.0	12.6	12.6	33.0	33.0
1FBI	13.9	10.3	4.0	13.2	12.5	15.3	11.9	2.5	21.1	11.8	58.5	24.7
1FIN	21.3	13.8	10.4	12.3	12.3	21.6	9.5	8.7	21.6	15.8	56.3	55.8
1FQ1	17.2	10.0	9.3	20.7	16.4	17.0	13.7	8.1	17.2	14.0	30.7	27.4
1FSS	16.4	6.3	1.4	10.4	8.4	5.7	5.4	0.8	5.8	5.4	14.7	13.6
1GLA	19.3	17.8	7.1	19.5	18.7	20.8	11.9	4.5	21.1	20.6	51.5	51.5
1GOT	11.0	11.0	5.6	16.4	9.3	13.3	10.1	4.0	13.2	10.8	37.9	30.8
1IAI	13.4	13.3	7.3	11.3	11.3	11.8	7.3	5.9	12.1	11.9	31.9	24.4
1IGC	16.1	15.9	8.3	16.6	12.7	13.3	13.1	3.5	13.2	13.0	37.9	37.7
1JHL	14.6	8.4	5.1	17.5	17.5	9.4	9.1	2.3	9.2	8.8	18.6	18.3
1KKL	20.2	16.3	11.5	20.1	15.8	17.5	15.7	9.6	18.3	16.7	49.4	48.4
1KXQ	10.2	4.1	1.1	9.3	4.2	1.2	0.7	0.4	1.1	1.1	1.9	1.9
1KXT	19.2	17.2	8.6	12.5	12.5	15.7	11.9	7.3	15.4	15.4	42.4	42.4
1KXV	16.9	16.5	5.2	18.3	16.8	14.4	14.3	4.1	18.6	5.0	45.2	12.4
1L0Y	17.5	12.0	10.2	11.9	11.9	27.3	13.5	9.9	14.6	14.6	67.6	55.7
1MAH	12.4	10.5	1.0	1.3	1.2	5.8	1.1	0.6	5.7	0.6	14.7	1.0
1MEL	12.2	9.7	0.9	14.0	1.2	1.2	1.2	0.7	1.1	1.1	2.1	2.1
1MLC	16.2	9.4	4.6	12.0	12.0	9.6	6.9	3.1	9.9	7.0	23	21.1
1NCA	14.5	1.3	1.1	22.8	15.7	18.8	18.8	0.3	0.5	0.5	1.0	1.0
1NMB	17.7	14.0	13.9	22.7	21.4	24.4	17.9	11.1	24.4	20.2	72.0	43.9
1PPE	0.6	0.6	0.6	1.0	0.8	1.2	0.6	0.3	1.1	0.5	3.3	1.3
1QFU	22.6	13.2	10.6	22.5	13.4	17.5	13.6	9.7	18.1	17.0	53.8	52.2
1SPB	0.6	0.5	0.5	0.7	0.5	0.4	0.4	0.4	0.5	0.4	0.6	0.4
1STF	1.1	0.9	0.7	1.2	0.9	0.4	0.4	0.3	0.4	0.4	1.1	0.9
1TAB	8.5	7.5	0.9	18.4	11.6	7.9	7.9	0.3	8.0	7.9	20.2	20.2
1TGS	8.4	8.2	1.5	9.0	1.9	8.1	8.1	1.4	8.2	8.1	16.2	15.9
1UDI	15.2	1.2	0.8	7.5	0.8	16.4	0.5	0.4	0.5	0.5	0.8	0.8
1UGH	6.4	6.0	1.5	2.2	1.6	2.0	2.0	0.7	8.2	2.2	23.7	5.0
1WEJ	15.8	10.4	8.2	10.5	10.5	10.6	10.6	4.3	11.6	10.8	21.2	19.5
1WQ1	2.7	2.5	1.2	6.4	4.9	5.4	5.4	0.9	5.6	5.6	10.3	10.3
2BTF	0.8	0.7	0.5	1.0	0.7	0.7	0.7	0.3	0.8	0.4	1.9	1.1
2JEL	9.5	9.5	9.0	12.2	10.6	9.9	9.7	8.8	9.7	9.7	16.6	16.6
2KAI	13.5	7.2	4.7	10.2	9.0	4.8	4.8	4.0	4.8	4.8	14.4	14.4
2MTA	20.3	17.7	6.9	16.8	16.8	17.5	16.4	6.8	17.5	17.4	55.6	55.1
2PCC	20.3	14.9	7.0	13.5	7.0	18.5	15.3	4.0	22.3	10.8	47.4	29.3
2PTC	12.6	8.5	2.6	12.6	2.6	12.9	5.3	1.5	13.3	8.0	40.0	18.4
2SIC	10.3	10.3	1.4	8.7	1.4	1.0	0.9	0.6	1.0	1.0	4.8	4.8
2SNI	8.4	5.0	4.4	8.8	5.8	9.2	8.1	3.2	8.4	8.1	18.9	17.8
2TEC	1.0	0.6	0.5	0.9	0.9	0.5	0.3	0.2	0.3	0.3	0.8	0.6
2VIR	18.7	14.9	11.2	21.1	20.5	18.4	15.8	11.2	18.3	18.3	53.6	49.1

(*continued*)

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

Table 1. *Continued*

	ZDOCK			RDOCK		ROTAFIT decoys			ROTAFIT cluster			
	I-1[a]	I-5[b]	I-B[c]	I-1[a]	I-5[b]	I-1[a]	I-5[b]	I-B[c]	I-1[a]	I-5[b]	L-1[d]	L-5[e]
3HHR	12.3	12.0	11.5	13.5	13.4	17.5	12.3	10.6	21.2	17.2	58.4	52.3
4HTC	0.9	0.9	0.7	16.1	16.1	0.6	0.6	0.6	0.7	0.6	0.8	0.7
<2.0[f]	6	13	25	10	18	12	16	27	13	17	9	12
<2.5	6	16	26	12	19	13	17	28	14	19	10	13
<3.0	8	16	27	12	21	13	17	29	14	20	10	13
<4.0	8	16	29	12	21	13	17	38	14	20	11	15
<5.0	9	18	33	13	24	14	18	42	15	23	12	16
<8.0	12	24	43	16	28	19	28	47	20	29	13	19
Best 10[g]	2.1	0.9	0.7	1.0	0.8	0.8	0.6	0.3	0.6	0.5	1.2	0.9
Best 20	5.3	2.2	0.9	3.4	1.1	2.9	1.5	0.5	2.4	1.0	6.2	2.7
Best 30	7.3	4.1	1.4	5.9	2.3	5.3	3.4	0.9	4.7	2.7	11.0	6.6
Best 40	8.9	5.6	2.4	7.7	4.2	7.5	5.2	1.6	6.8	4.7	16.6	10.9
Best 50	10.5	7.1	3.6	9.6	5.9	9.5	6.8	2.6	8.8	6.5	22.9	16.2

[a] (I-1) Interface RMSD (Å) of the highest-scoring decoys.
[b] (I-5) Best interface RMSD (Å) in five top-scoring decoys.
[c] (I-B) Interface RMSD (Å) of best decoy.
[d] (L-1) Ligand RMSD (Å) of the highest-scoring decoys.
[e] (L-5) Best ligand RMSD (Å) of five top-scoring decoys.
[f] (<2.0) Number of cases with an interface RMSD of <2.0 Å.
[g] (Best 10) Mean RMSD (Å) of the 10 best-performing cases in the benchmark set.

which has an interface RMSD of 2.5 Å to the native structure.

Discussion

We present an algorithm to refine docking decoys generated by ZDOCK, a state-of-the-art FFT rigid-body docking program. The goal of the work was to combine the advantages of different methods in protein–protein docking: While the fast Fourier transform technique allows a rapid scanning of the complete six-dimensional translational and rotational space, it relies on simple energy functions that can be expressed as sums of products between grid values to allow a fast Fourier transformation. While the final ranking of decoys might not be optimal, in most cases the FFT algorithm provides some reasonable starting points for further refinements. Another drawback of FFT methods is the inherent rigidity of receptor and ligand molecules.

On the other hand, Monte Carlo simulations allow the implementation of any form of composite energy terms as well as the conformational flexibility. Here, we chose the Monte Carlo approach to obtain not only information about the energies of docking decoys, but also about their distribution. It is well known that native structures mostly lie in broad energy wells rather than a narrow minimum (Camacho et al. 1999), which can be detected by clustering the resulting Monte Carlo decoys (Lorenzen and Zhang 2007).

Another novelty in our approach is the gradual roughening of the energy landscape, starting from a smoothed Lennard-Jones potential. In this way, structures can be smoothly guided into tight-fit positions with rising impact of side-chain positions (since overlap and clashes are punished stronger) during the simulation process. Since the average negative part of the van der Waals energy is kept constant by our scaling method, the shape of the potential smoothly changes from having considerable attractive forces even for r >> r_0 and only a few repulsive

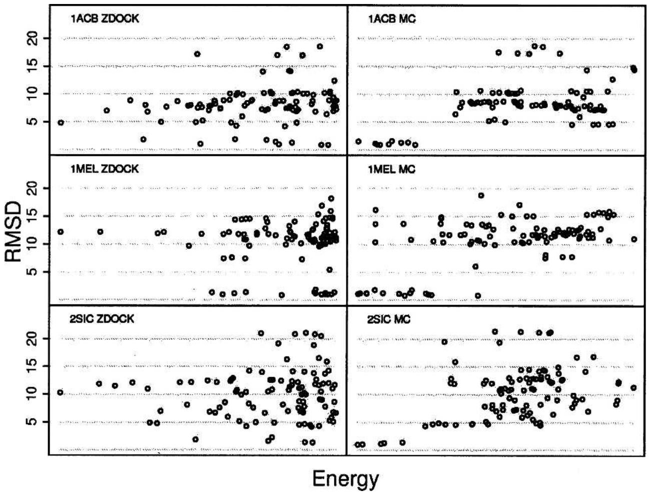

Figure 4. Interface RMSD versus energy for the first 100 ZDOCK decoys before (*left*, negative ZDOCK score interpreted as energy) and after (*right*) the ROTAFIT Monte Carlo refinement. The figure shows three representative cases (1ACB, 1MEL, and 2SIC). The energy values have been rescaled to fit the united scale of different targets.

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

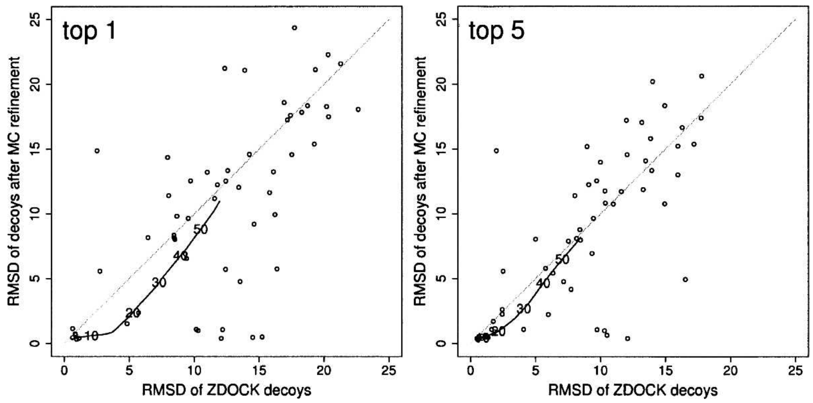

Figure 5. Interface RMSD of refined decoys by ROTAFIT versus RMSD of initial ZDOCK decoys for the 59 benchmark targets. (*Left*) First decoy; (*right*) the best in top five decoys. The Monte Carlo decoys show more near-native cases. The solid lines indicate average interface RMSDs of the best 10, 20, 30, 40, and 50 cases by both ZDOCK and ROTAFIT. The lines *below* the diagonal line indicate that the average interface RMSD has been improved by the ROTAFIT simulations.

forces for clashes caused by incorrect side-chain conformations to being more restrictive toward the distance of minimal energy, r_0.

Another advantage of a smoothed energy function is the possibility of a broader distribution of decoys in the beginning phase of the simulation and subsequent driving toward local minima. In this way, a bigger part of the conformational space can be scanned, while cluster sizes of the final conformations give information about the width of the energy well.

As shown in Figure 5, the RMSD of the highest-scoring decoys could be improved by our method, and in eight cases, near-native decoys were identified that were missed by ZDOCK. Refinement results with gradual roughening were significantly better than with a static energy function (data not shown), which shows the potential of our approach. An increase in the number of starting decoys is promising to further improve the results, since 100 ZDOCK decoys often do not include structures close enough to the native structure to be refined. The improved version of the ROTAFIT algorithm is being developed and will be made publicly available in future work.

Materials and Methods

Docking algorithm

ROTAFIT starts from the FFT-based rigid-body docking structures by ZDOCK (Chen et al. 2003a). The conformational space in ROTAFIT is searched by the replica-exchange Monte Carlo simulation algorithm (Swendsen and Wang 1986), where 40 replicas are used and each takes 2500 Monte Carlo moving attempts. Our testing data show that longer simulation with

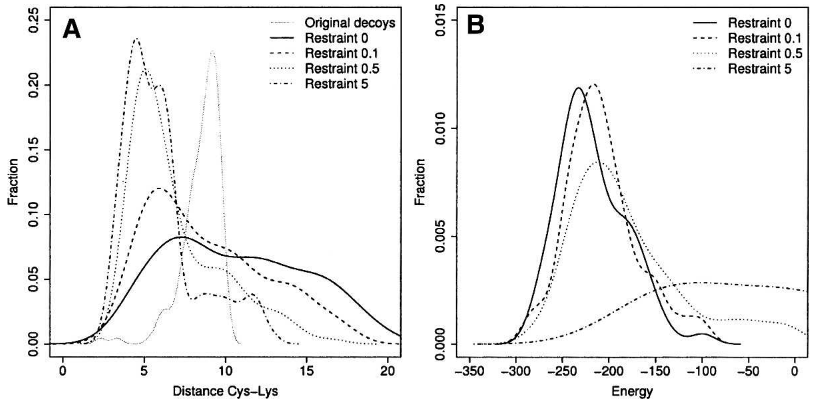

Figure 6. (*A*) Distribution of distances between Cys93 of Ubc9 and Lys14 of Hip2 in the initial decoy set and after Monte Carlo refinement with different restraint weights. A restraint ≥0.5 leads to a significantly closer distance between the two residues. (*B*) Energy distribution of refined decoys using different restraint weights for the Cys–Lys distance. A restraint below 0.5 does not significantly impair the final energies of the models.

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

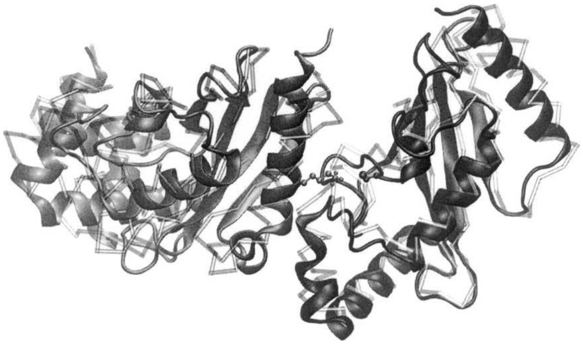

Figure 7. The first ROTAFIT model for T27 submitted to round 11 of the CAPRI experiment, which has an interface RMSD of 2.5 Å. The predicted model is shown in cartoon and the target in backbone. *Left* is the structure of Ubc9 and *right* for Hip2. The catalytic residue Cys93 and the acceptor site Lys14 are shown as sticks in the model.

more replicas does not improve the results (see Fig. 2). In each moving attempt, a translation along a random vector and a rotation around a random axis are performed. The translation distance and the rotation angle are scaled so that the displacement of interface atoms is distributed according to a Gaussian function with a standard deviation of 0.5 Å. Following the backbone displacement, conformations of two randomly selected side chains on the ligand-receptor interface are randomly changed based on Dunbrack's rotamer library (Dunbrack Jr. and Cohen 1997). The acceptance of the composite movements is decided by the Metropolis criterion (Metropolis et al. 1953) using an energy function comprised of van der Waals, Coulomb, and atomic contact energy. Replica exchange is performed after every moving attempt, which leads to an average of 1000 exchanges between neighboring replicas. After the Monte Carlo simulations, the structures in the five lowest temperature replicas obtained in the last 600 steps of the refinement simulations are clustered. For each cluster, the structure at the cluster center is selected. The structures from 100 different clusters are ranked by their energy. The overall procedure of ROTAFIT can be seen in Figure 1.

Energy function

Receptor and ligand are modeled with implicit hydrogen atoms (only polar hydrogen atoms treated explicitly). The energy function of ROTAFIT consists of three terms: a Lennard-Jones term $E_{ij} = \varepsilon_{ij}\left(\left(\frac{r_{0_{ij}}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{0_{ij}}}{r_{ij}}\right)^{6}\right)$ with parameters of ε_{ij} and $r_{0_{ij}}$ taken from CHARMM (Brooks et al. 1983), a Coulomb term $E_{Cou} = \frac{1}{4\pi\varepsilon\varepsilon_0}\frac{q_i q_j}{r_{ij}}$, and an atomic contact energy term (Zhang et al. 1997). For the Lennard-Jones and Coulomb terms, r_{ij} is the distance between two atoms i and j, q_i and q_j are their charges, and $r_{0_{ij}}$ is the optimal distance between two atoms of type i and j. For the Coulomb term, a distant dependent dielectric $\varepsilon = 4r_{ij}$ was used. Partial charges for amino acids used in the Coulomb term were also taken from CHARMM. The atomic contact energy is a statistical potential with 19 distinct atom types derived from PDB structures. For the Coulomb and Lennard-Jones interaction energies, a distance cutoff at 1 Å was used to avoid singularities, which sets the potential at shorter distances equal to that at 1 Å.

Smoothing of the van der Waals energy function

The smoothing approach for the van der Waals function was inspired by the potential scaling approach of Zacharias (Zacharias et al. 1994; Riemann and Zacharias 2005). Basically, the authors used a modified Lennard-Jones potential of the form $V(r_{ij}) = \varepsilon_{ij}(1-\lambda)\left(\left(\frac{r_{0_{ij}}^2+\lambda\delta}{r_{ij}^2+\lambda\delta}\right)^{6} - 2\left(\frac{r_{0_{ij}}^2+\lambda\delta}{r_{ij}^2+\lambda\delta}\right)^{3}\right)$ with ε_{ij} being the minimum Lennard-Jones energy for the interaction between atoms i and j, $r_{0_{ij}}$ the optimal distance between the two atoms, and δ, a shifting parameter. The value of λ can vary between 0 (full Lennard-Jones potential) and 1 (interaction energy vanished) during the simulations.

Our potential was designed to include only one scalable parameter λ instead of the two parameters, λ and δ, and keep the average attractive forces constant while increasing the repulsive forces. Assuming an equidistribution of particles in three-dimensional space, the number of pairwise interactions between particles with distance r would be proportional to r^3. The integral over the attractive (negative) part of our modified potential would thus be:

$$\int_{\sqrt{2^{-\frac{1}{3}}\left(r_{0_{ij}}^2+\lambda\right)-\lambda}}^{\infty} r^3\left(\left(\frac{r_{0_{ij}}^2+\lambda}{r_{ij}^2+\lambda}\right)^{6} - 2\left(\frac{r_{0_{ij}}^2+\lambda}{r_{ij}^2+\lambda}\right)^{3}\right)dr = \tag{1}$$

$$\frac{3\sqrt[3]{4}}{10}\lambda\left(r_{0_{ij}}^2+\lambda\right) - \frac{3\sqrt[3]{2}}{10}\left(r_{0_{ij}}^2+\lambda\right)^2$$

A scalable potential function with constant average attractive forces can be obtained by scaling:

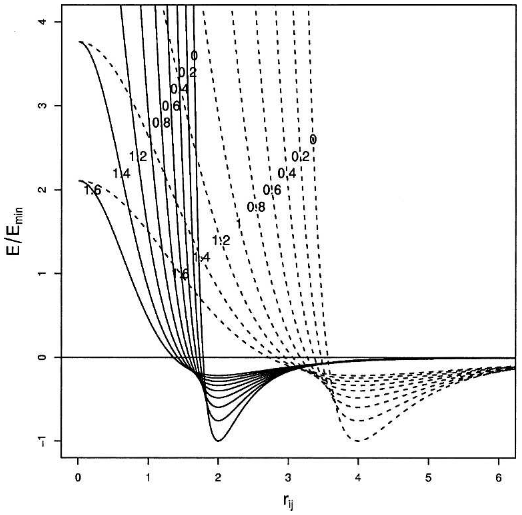

Figure 8. Smoothed Lennard-Jones potential (see Materials and Methods) with $r_{0_{ij}} = 2$ (solid lines) and 4 (dashed lines). λ' ranges from 0 to 1.6, as indicated in the figure. The spatial sum of attractive forces only depends on r, not on λ'. A λ' value of 0 is equivalent to the original Lennard-Jones potential.

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

$$V(r_{ij}) = \varepsilon_{ij} \frac{5r_{0_{ij}}^4}{5\left(r_{0_{ij}}^2+\lambda\right)^2 - 2\sqrt[3]{2}\lambda\left(r_{0_{ij}}^2+\lambda\right)} \times \left(\left(\frac{r_{0_{ij}}^2+\lambda}{r_{ij}^2+\lambda}\right)^6 - 2\left(\frac{r_{0_{ij}}^2+\lambda}{r_{ij}^2+\lambda}\right)^3\right) \qquad (2)$$

To obtain comparable degrees of smoothness for different values of r_{ij}, we used smoothing parameters λ' which depends on $r_{0_{ij}}$ with $\lambda = \lambda' \, r_{0_{ij}}^2$. Figure 8 shows potential functions of different smoothing values.

The current setting starts with a smoothing value $\lambda' = 0.4$, which is decreased by 0.05 in every 150 Monte Carlo steps. This way, the initial setting allows a rough orientation of receptor and ligand toward each other, while later in the simulation, restrictive settings direct orientations and side chains toward a tight fit between receptor and ligand. Stronger smoothing in the beginning of the simulation decreased the specificity of the interaction and worsened the results (data not shown). The roughening of the energy landscape was designed in such a way that the original Lennard-Jones potential was reached and briefly equilibrated after half of the simulation time.

Acknowledgments

Part of the calculations in this work have been performed on the KU ITTC clusters, where help from Drs. V. Frost and A. Hock is gratefully acknowledged. The project is supported by KU Start-up Fund 06194.

References

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

Bastard, K., Prevost, C., and Zacharias, M. 2006. Accounting for loop flexibility during protein-protein docking. *Proteins* **62:** 956–969.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Bernier-Villamor, V., Sampson, D.A., Matunis, M.J., and Lima, C.D. 2002. Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. *Cell* **108:** 345–356.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4:** 187–217.

Camacho, C.J., Weng, Z., Vajda, S., and DeLisi, C. 1999. Free energy landscapes of encounter complexes in protein-protein association. *Biophys. J.* **76:** 1166–1178.

Camacho, C.J., Ma, H., and Champ, P.C. 2006. Scoring a diverse set of high-quality docked conformations: A metascore based on electrostatic and desolvation interactions. *Proteins* **63:** 868–877.

Chen, R., Li, L., and Weng, Z. 2003a. ZDOCK: An initial-stage protein-docking algorithm. *Proteins* **52:** 80–87.

Chen, R., Mintseris, J., Janin, J., and Weng, Z. 2003b. A protein—protein docking benchmark. *Proteins* **52:** 88–91.

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

Desterro, J.M., Thomson, J., and Hay, R.T. 1997. Ubch9 conjugates SUMO but not ubiquitin. *FEBS Lett.* **417:** 297–300.

Dunbrack Jr., R.L. and Cohen, F.E. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6:** 1661–1681.

Fernandez-Recio, J., Totrov, M., and Abagyan, R. 2002. Soft protein-protein docking in internal coordinates. *Protein Sci.* **11:** 280–291.

Fernandez-Recio, J., Totrov, M., and Abagyan, R. 2003. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins* **52:** 113–117.

Gabb, H.A., Jackson, R.M., and Sternberg, M.J. 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272:** 106–120.

Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., and Baker, D. 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331:** 281–299.

Gront, D., Kolinski, A., and Skolnick, J. 2000. Comparison of three Monte Carlo conformational search strategies for a protein-like homopolymer model: Folding thermodynamics and identification of low-energy structures. *J. Chem. Phys.* **113:** 5065–5071.

Heifetz, A., Katchalski-Katzir, E., and Eisenstein, M. 2002. Electrostatics in protein-protein docking. *Protein Sci.* **11:** 571–587.

Janin, J. 2005. Assessing predictions of protein–protein interaction: The CAPRI experiment. *Protein Sci.* **14:** 278–283.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I., and Wodak, S.J. 2003. CAPRI: A critical assessment of predicted interactions. *Proteins* **52:** 2–9.

Johnson, E.S. and Blobel, G. 1997. Ubc9p is the conjugating enzyme for the ubiquitin-like protein Smt3p. *J. Biol. Chem.* **272:** 26799–26802.

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.* **89:** 2195–2199.

Kirkpatrick, S., Gelatt Jr., C.D., and Vecchi, M.P. 1983. Optimization by simulated annealing. *Science* **220:** 671–680.

Li, L., Chen, R., and Weng, Z. 2003. RDOCK: Refinement of rigid-body protein docking predictions. *Proteins* **53:** 693–707.

Lin, D., Tatham, M.H., Yu, B., Kim, S., Hay, R.T., and Chen, Y. 2002. Identification of a substrate recognition site on Ubc9. *J. Biol. Chem.* **277:** 21740–21748.

Lorber, D.M., Udo, M.K., and Shoichet, B.K. 2002. Protein-protein docking with multiple residue conformations and residue substitutions. *Protein Sci.* **11:** 1393–1408.

Lorenzen, S. and Zhang, Y. 2007. Identification of near-native structures by clustering protein docking conformations. *Proteins* **68:** 187–194.

Mandell, J.G., Roberts, V.A., Pique, M.E., Kotlovyi, V., Mitchell, J.C., Nelson, E., Tsigelny, I., and Ten Eyck, L.F. 2001. Protein docking using continuum electrostatics and geometric fit. *Protein Eng.* **14:** 105–113.

Mendez, R., Leplae, R., Lensink, M.F., and Wodak, S.J. 2005. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins* **60:** 150–169.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21:** 1087–1092.

Muller, S., Hoege, C., Pyrowolakis, G., and Jentsch, S. 2001. SUMO, ubiquitin's mysterious cousin. *Nat. Rev. Mol. Cell Biol.* **2:** 202–210.

Murphy, J., Gatchell, D.W., Prasad, J.C., and Vajda, S. 2003. Combination of scoring functions improves discrimination in protein-protein docking. *Proteins* **53:** 840–854.

Norel, R., Lin, S.L., Wolfson, H.J., and Nussinov, R. 1994. Shape complementarity at protein–protein interfaces. *Biopolymers* **34:** 933–940.

Pichler, A., Knipscheer, P., Oberhofer, E., van Dijk, W.J., Korner, R., Olsen, J.V., Jentsch, S., Melchior, F., and Sixma, T.K. 2005. SUMO modification of the ubiquitin-conjugating enzyme E2-25K. *Nat. Struct. Mol. Biol.* **12:** 264–269.

Riemann, R.N. and Zacharias, M. 2005. Refinement of protein cores and protein-peptide interfaces using a potential scaling approach. *Protein Eng. Des. Sel.* **18:** 465–476.

Schueler-Furman, O., Wang, C., and Baker, D. 2005. Progress in protein—protein docking: Atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins* **60:** 187–194.

Schwarz, S.E., Matuschewski, K., Liakopoulos, D., Scheffner, M., and Jentsch, S. 1998. The ubiquitin-like proteins SMT3 and SUMO-1 are conjugated by the UBC9 E2 enzyme. *Proc. Natl. Acad. Sci.* **95:** 560–564.

Downloaded from www.proteinscience.org on December 28, 2007 - Published by Cold Spring Harbor Laboratory Press

Swendsen, R.H. and Wang, J.S. 1986. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.* **57:** 2607–2609.

Tovchigrechko, A. and Vakser, I.A. 2005. Development and testing of an automated approach to protein docking. *Proteins* **60:** 296–301.

Tovchigrechko, A. and Vakser, I.A. 2006. GRAMM-X public Web server for protein-protein docking. *Nucleic Acids Res.* **34:** W310–W314. doi: 10.1093/nar/gki206.

Vakser, I.A. 1995. Protein docking for low-resolution structures. *Protein Eng.* **8:** 371–377.

Vakser, I.A. 1996. Low-resolution docking: Prediction of complexes for underdetermined structures. *Biopolymers* **39:** 455–464.

Zacharias, M. 2003. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* **12:** 1271–1282.

Zacharias, M., Straatsma, T.P., and McCammon, J.A. 1994. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.* **100:** 9025–9031.

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

Zhang, Y., Kihara, D., and Skolnick, J. 2002. Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* **48:** 192–201.