Complementarity of structure ensembles in protein-protein binding

Raik Grünberg[+], Johan Leckner[+*] and Michael Nilges[†]

23rd May 2005

Unité de Bioinformatique Structurale, Institut Pasteur, Paris, France.

[+]These authors contributed equally.

Abstract

Protein-protein association is often accompanied by changes in receptor and ligand structure. This interplay between protein flexibility and protein-protein recognition is currently the largest obstacle both to our understanding and to the reliable prediction of protein complexes. We performed two sets of molecular dynamics simulations for the unbound receptor and ligand structures of 17 protein complexes and applied shape-driven rigid body docking to all combinations of representative snapshots. The cross docking of structure ensembles increased the chances to find near native solutions. The free ensembles appeared to contain multiple complementary conformations. These were in general not related to the bound structure. We suggest that protein-protein binding follows a three-step mechanism of diffusion, free conformer selection and refolding. This model combines previously conflicting ideas and is in better agreement with the current data on interaction forces, time scales, and kinetics.

Abreviations

MD – molecular dynamics
PCR-MD – principle component restrained molecular dynamics
RMS(D) – root mean square deviation
FNAC – fraction of native atom contacts

1 Introduction

Specific recognition between proteins is a prerequisite for most biological processes. Our current understanding of this fundamental interaction is caught in a contradiction: On the one hand experimental rates of association suggest that, in many cases, almost every collision between two partner proteins leads to the formation of a complex (Northrup and Erickson 1992). On the other hand, even if we know the atomic structure of both proteins, we often fail to predict the structure of the complex because the two free partners simply do not fit sufficiently well. Over the last two decades the computational solution of this protein-protein docking problem has been an area of intense research (reviewed by Halperin et al. (2002)). Advances in docking methods often went hand in hand with new insights into the binding mechanism, and the fact that we often fail to predict the structure of a protein complex with confidence perhaps mirrors our incomplete understanding of the binding process.

Structures of protein complexes reveal intricate shape complementarity between the binding partners, which seemingly confirms Emil Fischer's (1894) key-lock model of biomolecular interaction. However, the free (unbound) receptor and ligand structures are often much less complementary and show significant deviations from their bound conformation (Betts and Sternberg 1999; Lo Conte et al. 1999). Consequently, early rigid-body docking algorithms could re-dock known complexes but were unable to predict them from the free components (Kuntz et al. 1982; Goodford 1985). The key-lock model may hold for the final protein complex but it cannot explain the process of recognition between the free molecules.

Daniel Koshland's (1958) induced fit model acknowledges a certain plasticity of proteins and postulates a mutual adaptation of the two structures. It offers a valid description of recognition if we assume that this process is driven by forces that do not require good shape complementarity to start with (Bosshard 2001). However, protein-protein recognition seems to be controlled, to a large extent, by short range electrostatics (Frisch et al. 2001), desolvation entropy (Camacho et al. 2000), and van der Waals interactions (Gray et al. 2003), which all depend to various degrees on shape complementarity. Induced fit may be appropriate for describing the transformation of receptor and ligand after recognition has occurred, but it cannot explain the process of recognition itself.

A third model, conformational selection, is inspired by the MWC mechanism of allosteric regulation (Monod et al. 1965) and is more compatible with short range interaction forces. Experimental protein structures are only the average of many conformational states (Frauenfelder et al. 1991). The model postulates "recognition" conformers that are hidden in the two structure ensembles and select each other upon binding. Early on, experiments corroborated the MWC model (Kirschner et al. 1966). Later experiments on antibodies showed that, in several cases, binding of an antigen was influenced by an equilibrium of different antibody conformations (e.g. (Lancet and Pecht 1976; Foote and Mil-

[*]Current address: Department of Chemistry and Bioscience/Molecular biotechnology Chalmers University of Technology, Sweden

[†]Correspondence: nilges@pasteur.fr

stein 1994). Experimental evidence was also provided for the inverse case - the selection of antigen conformers by antibodies (Leder et al. 1995; Berger et al. 1999). Kumar et al. (2000) then suggested conformational selection as a mechanism for protein-protein interaction in general. They explicitly postulated that bound conformations of receptor and ligand are part of their free structure ensembles and that recognition occurs between the two bound conformers. Thus, recognition and (apparent) structural adaptation could be explained simultaneously. Evidence for a preexisting equilibrium between free and bound conformations is hard to come by. Recent experimental structures are interpreted in this direction (Goh et al. 2004). However, at closer examination they confirm the existence of distinct conformations in free and bound structure ensembles but only very few suggest overlaps between the two. Since it usually leaves no traces in free crystallographic or NMR structures, the bound conformation, if it is present, must be a rare state.

The elegance of the preexisting equilibrium hypothesis stems from its combination of the modern ensemble view of protein structure with a simple key-lock mechanism for recognition. However, the model is challenged by the usually very fast pace of protein-protein recognition, which does not leave room for many unsuccessful collisions (Northrup and Erickson 1992). Recognition conformations must be frequent enough to occur simultaniously for both receptor and ligand within the short time window during which they are properly aligned in the course of a single random collision. Northrup and Erickson describe a protein encounter as a series of micro-collisions at different orientations. Estimates for the length of a (possibly correctly) aligned micro-collision range from 400 ps as lower bound to 10 ns as upper bound (Northrup and Erickson 1992; Janin 1997). The preexisting equilibrium hypothesis thus implies a certain minimum frequency of bound conformations. According to our rough estimate (see appendix), bound conformations must represent 4% of both free ensembles in order to achieve a 50% recognition success within a 400 ps time window. Even a fairly unrealistic recognition time of 10 ns still requires a frequency close to 1%.

A valid model of protein-protein association needs to explain not only the obvious difference between free and bound protein structures, but must also be compatible with kinetic data. So far, the two problems are usually addressed in isolation. The detailed theoretical studies on the kinetic mechanism of binding have focused on the diffusion of proteins that are rigidly locked into their bound conformation (Northrup and Erickson 1992; Janin 1997; Camacho et al. 1999; Selzer and Schreiber 2001; Zhou 2001). These models can reproduce the kinetics of diffusion-controlled protein-protein associations with some success (Gabdoulline and Wade 2002) but regard structural transitions only as a passive induced fit after recognition has occurred.

Likewise, protein-protein docking algorithms rely on rigid body, rigid segment (Schneidman-Duhovny et al. 2003) or rigid backbone simplifications. Several recent programs consider alternative conformations of some or all exposed amino acid side chains (Fernandez-Recio et al. 2003; Gray et al. 2003) (and others). This strategy often improves predictions, especially in cases where a few side chain rotations account for most of the difference between free and bound structures. It has also spurred interest in the role of side chain flexibility for the process of protein binding (Kimura et al. 2001). However, the distinction between backbone and side chain dynamics is dictated by technical constraints and lacks a physical basis. Side chain and backbone torsions are correlated (Schrauber et al. 1993). Upon binding, side chain and backbone atoms are equally involved in conformational changes (Betts and Sternberg 1999; Lo Conte et al. 1999). Furthermore, also backbone conformations display significant variations across independently determined structures (Chothia and Lesk 1986) and deformations on this scale can already affect docking results (Ehrlich et al. 2004). From this point of view, such a thing as side chain flexibility does, strictly speaking, not exist. Therefore, recent protein-protein docking algorithms still fail if there are substantial differences between free and bound structures. The effective treatment of overall protein flexibility is now the largest obstacle both to our understanding and to the reliable prediction of protein-protein association.

In this study we examine the interplay of complete protein flexibility and protein-protein recognition. We combined two molecular dynamics based sampling strategies with systematic rigid body docking. We derived ensembles from the independently solved (unbound) structures of 17 receptor and 16 ligand proteins and applied shape-driven rigid body docking to all combinations of representative snapshots. We compared the success of this extended but still manageable search with the simple docking of the experimental structures. We show that already very sparse structure ensembles contained several combinations of receptor and ligand conformers that generated more and better near-native solutions. Remarkably, the docking performance of a given combination of receptor and ligand structure was largely uncorrelated with their similarity to the bound conformation. Based on these results we extend and combine the up to now conflicting models of protein-protein binding. We suggest a 3-step mechanism of diffusion, free conformer selection and refolding as working model for flexible recognition.

2 Results and Discussion

2.1 Structural data

We selected a set of 17 protein-protein complexes for which the structures of both the free components and the complex are available (table 1). This set is based on docking benchmarks from Graham Smith (http://www.bmm.icnet.uk/docking/systems.html) and Chen et al. (2003). From these benchmarks we excluded complexes with large non-protein ligands to facilitate the mostly automated modeling procedure. Only the free structures and molecular dynamics ensembles derived from them were used for the rigid body docking. The structure

Table 1. Protein-protein complexes used in this study.

ID[a]	RECEPTOR / LIGAND	PDB codes with chain identifier			Size (residues)		Contacts[b] (residue)	Complex type[c]
		Receptor	Ligand	Complex	Receptor	Ligand		
c01	Trypsin / Amyloid β-protein precursor inhibitor domain	1BRA	1AAP(A)	1BRC(E:I)	223	56	82	EI
c02	α-chymotrypsinogen / Pancreatic secretory trypsin inhibitor	2CGA(A)	1HPT	1CGI(E:I)	245	56	72	EI
c03	Kallikrein A / Pancreatic trypsin inhibitor	2PKA(AB)	5PTI	2KAI(AB:I)	232	58	49	EI
c04	Subtilisin BPN / Subtilisin inhibitor	1SUP	3SSI	2SIC(E:I)	275	108	62	EI
c05	Extracellular domain of tissue factor / Antibody Fab 5G9	1FGN(LH)	1BOY	1AHW(AB:C)	248	211	55	AA
c06	Humanized anti-lysozyme Fv / Lysozyme	1BVL(AB)	3LTZ	1BVK(AB:C)	224	129	26	AA
c08	Anti-lysozyme antibody Hyhel-63 / Lysozyme	1DQQ(AB)	3LTZ	1DQJ(AB:C)	424	129	24	AA
c11	Barnase / Barstar	1A19(A)	1A2P(A)	1BSG(A:E)	108	89	44	EI
c13	Ribonuclease inhibitor / Ribonuclease A	2BNH	7RSA	1DFJ(E:I)	456	124	58	EI
c14	Acetylcholinesterase / Fasciculin-II	1VXR	1FSC(A)	1FSS(A:B)	532	61	53	EI
c15	HIVB-1 NEF / FYN tyrosin kinase SH3 domain	1AVV	1SHF(A)	1AVZ(B:C)	99	59	28	O
c16	Uracil-DNA glycosylase / Inhibitor	1AKZ	1UGI(A)	1UGH(E:I)	223	83	59	EI
c17	RAS activating domain / RAS	1WER	5P21	1WQ1(R:G)	324	166	73	O
c19	Glycosyltransferase / Tendamistat	1PIF	2AIT(mdl1)	1BVM(P:T)	495	74	59	EI
c20	CDK2 cyclin-dependant kinase 2 / Cyclin A	1HCL	1VIN	1FIN(A:B)	294	252	99	O
c21	CDK2 cyclin-dependant kinase 2 / KAP	1B39(A)	1FPZ(A)	1FQ1(A:B)	290	176	45	O
c22	Transductin Gt-α / Heteromeric G-protein	1TAG	1TBG(AE)	1GOT(A:BG)	314	408	80	O

[a] Complex identifier (ID) used throughout the paper (retained from www.bmm.icnet.uk/docking/systems.html).
[b] Number of intermolecular residue contacts calculated with a 4.5 Å cutoff.
[c] Complex types: EI – enzyme / inhibitor; AA – antibody / antigen; O – other

of receptor and ligand solved as a complex served as reference.

2.2 Measuring the quality of docking solutions

We analyzed and compared 2,106,368 solutions from 4114 rigid body docking calculations between 693 conformations of 33 different proteins (c05 and c06 share a ligand). To this end we needed a single metric for the quality of a given solution, i.e. to which extent it resembles the native arrangement of receptor and ligand in the complex. Rmsd-based measures are inappropriate for our purposes because they depend on the size and shape of the binding interface and, furthermore, would also be influenced by the conformational variations in our receptor and ligand ensembles. Criteria based on residue-residue contacts (Mendez et al. 2003) suffer from ambiguity introduced by bulky side chains in the interface. We therefore used a measure based on atom contacts. We define a fraction of native atom contacts (fnac) as the number of pairs of non-hydrogen receptor and ligand atoms that are within a 10 Å distance both in the native and the predicted orientation, divided by the total number of such pairs in the native complex. This value is less ambiguous and correlates better with rmsd-based criteria, shown in figure 1.

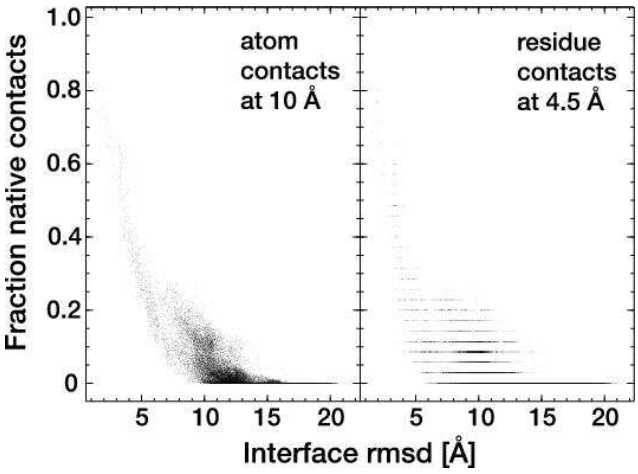

Figure 1: Correlation between rmsd- and contact-based docking quality criteria.
All 61.952 solutions from the cross-docking of 11 Barnase with 11 Barstar conformations derived from the unrestrained MD simulation were compared to the native complex (c11) by an rmsd- and two contact-based criteria. (A) The fraction of native atom contacts (fnac) is less ambiguous and correlates better with the heavy atom interface rmsd than (B) the traditional fraction of native residue contacts.

2.3 Conformational sampling

Rather than by a static structure, proteins are best described by an ensemble of individual conformations (Frauenfelder et al. 1991). In this study we try to incorporate the additional dimensions of receptor and ligand variability into the picture of the protein-protein recognition process. This recognition starts from the unbound components and we therefore concentrate on the conformational ensembles of the free receptor and the free ligand.

Molecular dynamics (MD) simulations offer a way to generate such ensembles, (Frauenfelder and Leeson 1998). We performed two sets of MD simulations for each of the 33 structures of free receptor and ligand. In the first set, 10 independent trajectories of 50 ps length each were calculated with the structure embedded in a 9 Å layer of explicit water.

Large-scale correlated motions usually escape the sampling of MD simulations (Balsera et al. 1996). A second ensemble was calculated with identical protocol except of a weak restraint alleviating this problem. Large-scale correlated motions typically occur along small gradients in the energy landscape. They are hence slow but, on the other hand, can be boosted by small interventions. As described previously (Abseher and Nilges 2000), the restraint acts on the ensemble of 10 concurrent trajectories as a whole and increases the variability along the major principal compo-

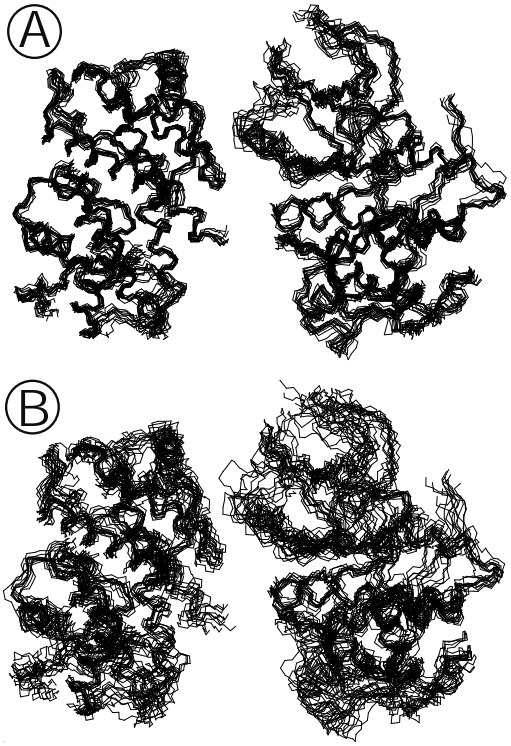

Figure 2: Receptor and ligand ensembles used for the docking of c20.
(A) The 10 receptor (right) and 10 ligand (left) snapshots selected from the unrestrained simulations. (B) The 10 snapshots from the principle component restrained simulations (PCR-MD) cover a wider range of conformations. The receptor and ligand snapshots have been oriented as in the native complex, but are separated horizontally. Side chains have been omitted for clarity.

Table 2. Average rmsd of structure ensembles.

ID	PDB code	MD		PCR-MD		interface RMSD to bound[c]	
		pairwise[a]	to free[b]	pairwise[a]	to free[b]	MD	PCR-MD
c01	1BRA	1.3±0.09	1.2±0.16	2.9±0.52	2.1±0.56	1.6±0.12	2.2±0.48
	1AAP	1.5±0.21	1.4±0.28	2.1±0.35	1.7±0.35	1.6±0.26	1.6±0.17
c02	2CGA	1.3±0.08	1.2±0.16	2.6±0.41	1.9±0.39	3.0±0.08	3.4±0.30
	1HPT	1.6±0.19	1.5±0.25	2.3±0.37	1.8±0.36	3.0±0.19	3.1±0.19
c03	2PKA	1.4±0.09	1.5±0.18	2.8±0.65	2.1±0.58	2.2±0.25	2.5±0.50
	5PTI	1.5±0.22	1.4±0.29	1.9±0.25	1.6±0.33	1.6±0.26	1.5±0.21
c04	1SUP	1.3±0.10	1.2±0.18	2.8±0.48	2.1±0.48	1.5±0.17	2.4±0.45
	3SSI	1.4±0.11	1.4±0.24	2.1±0.28	1.7±0.28	1.8±0.24	1.9±0.30
c05	1FGN	1.7±0.14	1.7±0.30	2.9±0.47	2.4±0.53	1.7±0.23	1.9±0.30
	1BOY	1.7±0.12	1.6±0.28	2.6±0.40	2.1±0.48	1.7±0.18	2.0±0.32
c06	1BVL	1.5±0.09	1.4±0.17	2.7±0.46	2.1±0.38	1.8±0.19	2.2±0.37
	3LZT	1.2±0.10	1.1±0.21	2.5±0.47	1.9±0.53	2.4±0.22	2.7±0.28
c08	1DQQ	1.5±0.13	1.6±0.29	2.6±0.40	2.1±0.47	1.5±0.15	1.6±0.18
	3LZT	1.2±0.10	1.1±0.21	2.3±0.41	1.8±0.38	1.9±0.14	2.3±0.28
c11	1A2P	1.3±0.10	1.2±0.19	2.2±0.38	1.8±0.55	1.7±0.21	2.3±0.62
	1A19	1.5±0.11	1.4±0.15	2.4±0.40	1.9±0.34	1.5±0.15	1.8±0.30
c13	2BNH	1.5±0.10	1.6±0.24	2.8±0.41	2.2±0.46	2.7±0.36	2.9±0.61
	7RSA	1.5±0.13	1.4±0.26	2.3±0.38	1.9±0.46	1.9±0.22	2.4±0.43
c14	1VXR	1.4±0.07	1.4±0.22	3.1±0.58	2.3±0.56	2.1±0.24	2.5±0.50
	1FSC	1.5±0.17	1.4±0.23	2.2±0.36	1.8±0.43	2.0±0.29	2.3±0.38
c15	1AVV	1.6±0.17	1.5±0.23	2.6±0.42	2.0±0.35	1.5±0.18	1.8±0.24
	1SHF	1.5±0.16	1.4±0.22	1.8±0.19	1.6±0.23	2.0±0.21	2.1±0.19
c16	1AKZ	1.3±0.07	1.2±0.17	2.0±0.28	1.6±0.30	1.7±0.18	1.8±0.27
	1UGI	1.6±0.18	1.4±0.25	2.4±0.40	1.8±0.41	1.8±0.14	2.1±0.31
c17	1WER	1.5±0.10	1.5±0.22	2.3±0.36	1.8±0.37	1.7±0.08	2.0±0.30
	5P21	1.4±0.08	1.3±0.21	2.4±0.35	1.9±0.41	2.4±0.28	2.8±0.47
c19	1PIF	1.4±0.06	1.3±0.22	3.1±0.65	2.2±0.61	2.0±0.29	2.6±0.51
	2AIT	1.6±0.15	1.6±0.20	2.4±0.33	2.1±0.34	2.0±0.16	2.1±0.25
c20	1HCL	1.6±0.12	1.5±0.22	2.7±0.40	2.0±0.38	7.9±0.19	8.0±0.38
	1VIN	1.4±0.07	1.4±0.20	2.4±0.33	1.8±0.36	1.7±0.14	1.9±0.28
c21	1B39	1.6±0.11	1.4±0.18	2.3±0.39	1.9±0.40	5.8±0.16	5.8±0.35
	1FPZ	1.6±0.10	1.6±0.23	2.4±0.35	2.0±0.35	2.4±0.21	2.6±0.31
c22	1TBG	1.6±0.14	1.5±0.22	3.4±0.62	2.6±0.58	1.6±0.19	2.2±0.44
	1TAG	1.5±0.09	1.4±0.26	2.4±0.33	1.9±0.41	6.4±0.10	6.4±0.20

[a] Average pairwise heavy atom rmsd in (with standard deviation) between the 10 simulation snapshots.
[b] Average heavy atom rmsd in (with standard deviation) of the 10 simulation snapshots to the free structure.
[c] Average heavy atom rmsd (with standard deviation) of interface residues between the 10 simulation snapshots and the bound structure.

nents of motion. The computational cost of this principal component restrained simulation (PCR-MD) is similar to the classic approach above but the ensemble is considerably more diverse.

We performed c-means fuzzy clustering for each of the two structure ensembles and selected 2 x 10 representative conformations for combinatorical rigid body docking. A representative example of these discretized structure ensembles from the unrestrained (MD) and the restrained (PCR-MD) simulation is shown in figure 2. The snapshots capture considerable variation. Table 2 lists the average (rms) deviation between the members of each docking ensemble and their distance to the free and the bound structure. In tables S1 and S2 of supplemental materials this information is broken up into deviations of backbone and side chain atoms.

2.4 Ensemble (cross-) docking

We tried to mimic the recognition between two flexible molecules by a combinatorical docking of all snapshots from the receptor ensemble against all snapshots from the ligand ensemble. Each of the docking ensembles was supplemented with the free (experimental) structure. Using the docking program HEX (Ritchie and Kemp 2000), we performed 121 rigid body dockings for each complex and MD strategy. HEX represents receptor and ligand by a soft 3D surface skin model and calculates the volume of water that is expelled from the protein surfaces as they come together. In addition there is a penalty for steric overlap. Both terms are combined in a pseudo energy that depends solely on the atomic and water probe radii and is interpreted as an approximation of the desolvation and van der Waals component of the free energy of association. We did not employ any additional (e.g. electrostatic) potentials and dealt therefore only with the contribution of short range, geometry dependent, effects to the interaction free energy. HEX performs a systematic search over all 6 rigid body degrees of freedom and ranks in the order of 109 trial orientations by this interaction energy.

From each of the 121 HEX dockings we analyzed the 512 top ranking solutions provided by default. Since we did not apply any clustering and there was no random element in the search, the amount and quality of near-native orientations within the set of top-ranking solutions effectively depended on: a) how well the two protein conformations matched each other geometrically near the native orientation, b) how tolerant this steric match was to deviations from the optimum orientation, and c) how many non-native

alternative orientations with comparable geometric match existed and competed with the correct arrangement.

2.5 Complementarity across ensembles

Discrimination by shape complementarity alone is usually sufficient to predict the native arrangement of the bound receptor and ligand. In figure 3A the docking of the bound structures from c19 (Glycosyltransferase / Tendamistat) is shown as a representative example. The free structures, on the other hand, are generally much less complementary. For example, the majority of top-ranking solutions from the docking of free Glycosyltransferase and Tendamistat (figure 3B) reproduce no, or only few, native contacts. However, figure 3C shows the fnac (quality) of top-ranking solutions from the docking of the same free receptor structure against one of the alternative inhibitor conformations from the PCR-MD simulation. Clearly, this combination of structures had a better geometric fit in near native orientations. In figure 4A we show the amount and quality of near native solutions for all cross-dockings between the simulation-derived ensembles of the two proteins. Several conformer combinations performed better than the docking of the two experimental structures, both in terms of quantity (indicated by the size of the circle) and quality (indicated by the color). The gain was yet even more pronounced for the cross-docking of the ensemble that had been calculated with the PCR-MD technique (figure 4B).

As a second example we present similar results for the complex between CDK2 and Cyclin A (c20). This complex is one of the difficult docking test cases as the receptor undergoes large structural changes moving from the free to the bound state (C_α displacements of up to 20 Å). All 512 solutions from the docking of the two experimental structures stayed below a fnac of 10%. Nevertheless, as shown in figure 4B and C there were many combinations of MD or PCR-MD snapshots that yielded better solutions with fnac values up to 30%.

The results of all 17 test complexes are provided in supplemental figure S3 and summarized in table 3 and figure 5. We selected 2 dockings each from the cross-docking of MD and of PCR-MD ensembles: The one that generated the single highest fnac within the 512 top-ranking orientations and the one with the best compromise between quantity and quality of near-native solutions. We quantify this "compromise" docking performance with the sum of squared fnac values above 10%, i.e. a simple score strongly biased towards high fnac ranges.

The cross-docking of ensemble snapshots always found more and, in all but one case, also better near native solutions than the docking of the free conformations alone. There were usually several combinations of simulation snapshots, or snapshot and free structure, with better complementarity near the native orientation. Moreover, we can assume that even better fits remained hidden due to the fact that our docking ensembles were artificially sparse. The insufficient shape complementarity between many of the free receptor and ligand pairs could be an artifact of the rigid body or rigid backbone simplification.

2.6 Specificity of docking success

For every complex, we generated 10 random orientations that were distinct both from each other and the native (no contact overlap). We re-analyzed all docking solutions using these random orientations as reference. This allowed us to quantify the probability that the score of the free docking and the best score from the ensemble docking did not occur at random (table 3). All of the best performing conformer pairs reproduced the native complex better than the docking of the free experimental structures. In 9 out of 17 cases, the profound enrichment of high quality solutions from the docking of selected conformer pairs is also specific to the native orientation. In the remaining cases, the improvement is substantial but not significantly higher than what would be expected for a random orientation. We have indications that more specific results can be achieved for some of the 8 latter complexes if the HEX energy function is extended with an electrostatic term.

It should be noted that the consideration of 512 solutions each from 121 docking runs combined with the soft and simplistic energy function provoke a high level of "noise", i.e. similarities to a random orientation. The evaluation of fewer solutions with more detailed energy functions would most likely improve the discrimination. However, the technical (and challenging) problem of scoring docking solutions is not subject of this article.

2.7 Recognition conformations

Our simulations cover a time window that, at least, resembles but probably exceeds the estimated duration of a micro-collision. Already the use of multiple trajectories is expected to increase sampling by a factor of 2 (Caves et al. 1998). The fast equilibration, the method of solvation and, especially, the introduction of principle component restraints further enhance diversity (Abseher and Nilges 2000). We did not find a global transition from free to bound interface conformation in any of our 2 x 33 ensembles (data not shown). There was nevertheless notable variation in the structure ensembles and some conformers were necessarily closer to the bound than others (compare table 2). Binding could be promoted by such shifts towards the bound state (Kumar et al. 2000). In figure 6A we relate the distance from the bound state of a given pair of conformers and its performance in docking. There is no obvious correlation between similarity to the interface of the bound structure and docking performance. This picture remained the same when we expressed the distance between structures as Contact Area Difference (Abagyan and Totrov 1997) (data not shown) and is therefore not an artifact of the rmsd measure.

In figure 6B and C we focus only on those pairs of conformations that yielded the best docking result (score) for each complex. As apparent from table 3, the experimental structure was over-represented among these pairs, albeit only on the side of the larger binding partner. This bias was unique to the native orientation and absent from the conformer pairs with the highest similarity to a random reference (data not shown). Compared to the average ensemble

Table 3. Docking results for the free receptor and ligand, the conformer combination giving the highest fnac value and the best scoring combination.

ID	FREE					HIGHEST FNAC						BEST SCORE						average
		interf. rms[a]					conformer		interf. rms[a]				conformer		interf. rms[a]			
	fnac[b]	rec	lig	score[c,e]		fnac[b]	rec	lig	rec	lig	score[c]	fnac[b]	rec	lig	rec	lig	score[c,e,]	score[d,e]
c01	0.82	1.4	1.5	3.9 *(88+)*	MD	0.88	6	8	1.6	1.5	3.5	0.68	Free	6	1.4	2.2	10.2 *(85+)*	2.9 *(74+)*
					PCR-MD	0.86	Free	2	1.4	1.4	5.3	0.56	3	4	2.5	1.6	10.9 *(78+)*	3.8 *(84+)*
c02	0.35	2.9	2.8	2.9 *(97+)*	MD	0.66	9	2	3.1	2.8	3.3	0.61	6	3	3.1	3.1	10.4 *(96+)*	2.6 *(98+)*
					PCR-MD	0.63	Free	8	2.9	2.8	2.7	0.48	1	6	3.7	3.0	7.0 *(86+)*	2.3 *(97+)*
c03	0.75	1.3	1.4	4.2 *(83+)*	MD	0.75	Free	Free	1.3	1.4	4.2	0.53	Free	3	1.3	1.6	8.2 *(81+)*	2.5 *(80+)*
					PCR-MD	0.75	Free	Free	1.3	1.4	4.2	0.54	Free	4	1.3	1.7	8.5 *(68+)*	2.2 *(79+)*
c04	0.22	0.9	1.3	0.5 *(58+)*	MD	0.77	Free	1	0.9	1.5	0.9	0.70	6	4	1.3	1.5	4.3 *(71+)*	1.1 *(78+)*
					PCR-MD	0.74	10	5	1.3	1.5	0.9	0.50	9	8	2.4	1.9	3.9 *(57+)*	0.6 *(54+)*
c05	0.09	1.1	1.2	—	MD	0.20	6	2	1.8	1.6	0.2	0.20	6	2	1.8	1.6	0.2 *(96-)*	0.0 *(81-)*
					PCR-MD	0.23	3	1	1.7	2.3	0.1	0.18	7	5	2.3	2.5	0.3 *(96-)*	0.0 *(92-)*
c06	0.16	1.4	2.3	0.1 *(3-)*	MD	0.73	8	6	2.0	2.4	0.8	0.35	2	1	2.1	2.3	0.8 *(35-)*	0.1 *(26-)*
					PCR-MD	0.69	2	10	1.7	2.5	1.0	0.40	3	1	2.1	2.3	1.1 *(90-)*	0.1 *(38-)*
c08	0.09	1.2	1.7	—	MD	0.39	6	4	1.5	1.9	0.2	0.39	6	4	1.5	1.9	0.2 *(2-)*	0.0 *(1-)*
					PCR-MD	0.26	Free	6	1.2	2.0	0.2	0.26	Free	6	1.2	2.0	0.2 *(3-)*	0.0 *(1-)*
c11	0.75	1.0	1.0	4.8 *(81+)*	MD	0.81	Free	10	1.0	1.5	14.9	0.81	Free	10	1.0	1.5	14.9 *(88+)*	2.1 *(71+)*
					PCR-MD	0.82	Free	3	1.0	1.9	7.6	0.81	10	9	2.6	1.1	17.2 *(90+)*	2.7 *(79+)*
c13	0.76	1.8	1.2	0.9 *(64+)*	MD	0.83	3	4	2.0	1.8	7.0	0.78	10	3	2.6	1.8	26.1 *(99+)*	1.4 *(78+)*
					PCR-MD	0.83	Free	6	1.8	1.5	3.2	0.67	5	4	2.6	2.3	8.0 *(69+)*	0.7 *(58+)*
c14	0.15	1.4	1.7	0.1 *(2-)*	MD	0.84	8	8	2.0	1.8	2.2	0.84	8	8	2.0	1.8	2.2 *(19-)*	0.3 *(27-)*
					PCR-MD	0.69	6	6	2.3	2.2	2.4	0.69	6	6	2.3	2.2	2.4 *(15-)*	0.4 *(25-)*
c15	0.27	1.1	1.7	1.5 *(45+)*	MD	0.54	4	7	1.4	2.1	0.7	0.31	4	5	1.4	1.8	2.2 *(43-)*	0.9 *(5-)*
					PCR-MD	0.70	2	7	1.7	2.1	1.2	0.54	7	Free	2.2	1.7	2.6 *(30-)*	0.8 *(10-)*
c16	0.73	1.2	1.7	1.8 *(80+)*	MD	0.82	Free	7	1.2	1.6	3.0	0.59	5	10	1.5	1.8	7.5 *(83+)*	1.1 *(64+)*
					PCR-MD	0.85	4	1	1.8	1.5	12.9	0.85	4	1	1.8	1.5	12.9 *(91+)*	1.4 *(70+)*
c17	0.70	1.5	1.7	6.9 *(100+)*	MD	0.80	3	Free	1.8	1.7	2.8	0.74	10	Free	1.6	1.7	8.3 *(99+)*	1.5 *(97+)*
					PCR-MD	0.77	3	Free	1.5	1.7	3.1	0.65	Free	2	1.5	2.0	11.0 *(100+)*	1.3 *(99+)*
c19	0.39	1.2	1.7	0.6 *(49+)*	MD	0.77	Free	6	1.2	1.8	3.4	0.62	Free	1	1.2	1.8	6.8 *(72+)*	0.7 *(57+)*
					PCR-MD	0.81	Free	1	1.2	1.8	22.4	0.81	Free	1	1.2	1.8	22.4 *(92+)*	1.4 *(74+)*
c20	0.09	7.8	1.4	—	MD	0.29	8	7	7.8	1.6	0.3	0.27	8	9	7.8	1.6	0.5 *(38-)*	0.1 *(37-)*
					PCR-MD	0.30	5	9	8.1	2.1	0.5	0.30	5	9	8.1	2.1	0.5 *(46-)*	0.1 *(19-)*
c21	0.24	5.9	1.9	0.5 *(75+)*	MD	0.36	8	7	5.7	2.3	0.9	0.32	8	8	5.7	2.6	1.2 *(5-)*	0.3 *(45-)*
					PCR-MD	0.42	Free	2	5.9	2.6	0.6	0.28	1	7	5.7	2.6	1.1 *(16-)*	0.3 *(41-)*
c22	0.15	1.0	6.2	0.0 *(19-)*	MD	0.22	5	5	1.9	6.6	0.2	0.22	5	5	1.9	6.6	0.2 *(1-)*	0.0 *(16-)*
					PCR-MD	0.26	6	4	2.4	6.4	0.3	0.18	6	10	2.4	6.7	0.6 *(1-)*	0.1 *(23-)*

[a] heavy atom interface rmsd to the bound structure.
[b] fraction of native atom contacts (highest of the docking run).
[c] docking performance score, defined as the sum of squared fnac-values above 0.1, quantifies the amount and quality of near-native solutions captured by the docking run.
[d] average docking score of all 121 dockings.
[e] probability (in percent) that the value is an outlier above (+) or below (-) random expectation. Probabilities for scores around or below random expectations are afflicted with higher error due to the asymmetric shape of the lognormal random distribution. Probabilities also depend on the number of scores considered. The best of 121 combinatorical dockings must accordingly perform much better than the single free docking in order to achieve the same significance (specificity).

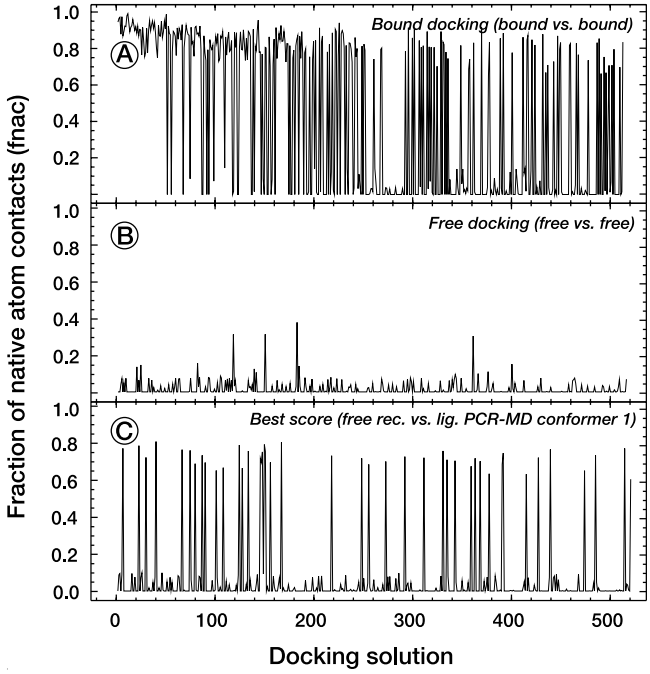

Figure 3: Selected docking results for c19.
Each panel shows the result of a single shape-driven rigid body docking experiment. The similarity to the true complex is measured for the 512 predictions that rank highest in surface complementarity. Data are shown for (A) bound docking, (B) free docking and (C) the highest scoring of the ensemble dockings (see table 3).

member, experimental conformations (open symbols in figure 6B and C) are also closer to the bound structure since the ensembles were moving away from the free without systematically moving towards the bound conformation. The short simulation time sometimes aggravates the effect as it may cause uneven sampling of the conformational space around the starting structure. The preference of experimental receptor structures might be an artifact of the docking protocol being optimized for free and bound crystallographic rather than simulation derived structures - not only in general but actually using the very same test complexes. After excluding experimental structures from the conformations of best complementarity no obvious trend remains, neither to the free experimental nor to the bound state (histogram in figure 6B and C).

Indeed, the systematic dependency on a single, e.g. bound, recognition conformation would impede fast binding. Protein structures move on a flat energy landscape that probably requires ms or even s for adequate sampling (Brooks III et al. 1988). The time window for recognition is short by comparison (Northrup and Erickson 1992; Janin 1997; Camacho et al. 2000). Nevertheless, we often observe deviations between the experimental free and bound structures that can only be bridged by large scale correlated motions, which, in turn, are unlikely to occur spontaneously within this short recognition time.

Our extensive data show that short range forces can drive recognition even where this is not evident from the free structures. Due to the simplistic energy function used we can only speculate that the conformations of highest complementarity are related to actual recognition conformers. Our results nevertheless suggest that different such conformers coexist and can be sampled within the short window of opportunity. The cross-docking of simulation-derived structure ensembles indicates that shape-driven recognition does not, or at least not generally, depend on systematic transitions from free to bound structures. This allows us to refine and combine the current models of the protein-protein binding process.

2.8 A working model of flexible recognition

Gabdoulline and Wade (2002) recently criticized the mutual inconsistency of current models for protein-protein association. Disputed are the nature of the rate-limiting step (diffusion or induced fit), the shape of the association energy landscape (broad funnel or tight channel), and the mechanism of conformational changes (preexisting equilibrium or induced fit). Most of these inconsistencies can be resolved if we describe binding as a 3-step process of diffusion, free conformer selection, and refolding or "induced fit", as shown in figure 7.

Association starts with the diffusional encounter of the two free structure ensembles (Rf and Lf) which, at rate k1, leads to a micro-collision with approximately correct orientation of receptor and ligand (Rf:Lf). The lifetime of this aligned encounter complex allows for gradual desolvation and it could, potentially, be prolonged by random complementarities between sub-populations of the two structure ensembles. Apart from such an unspecific "pre-selection", the structure of the two proteins is still characterized by their free conformation ensembles. This is the point where

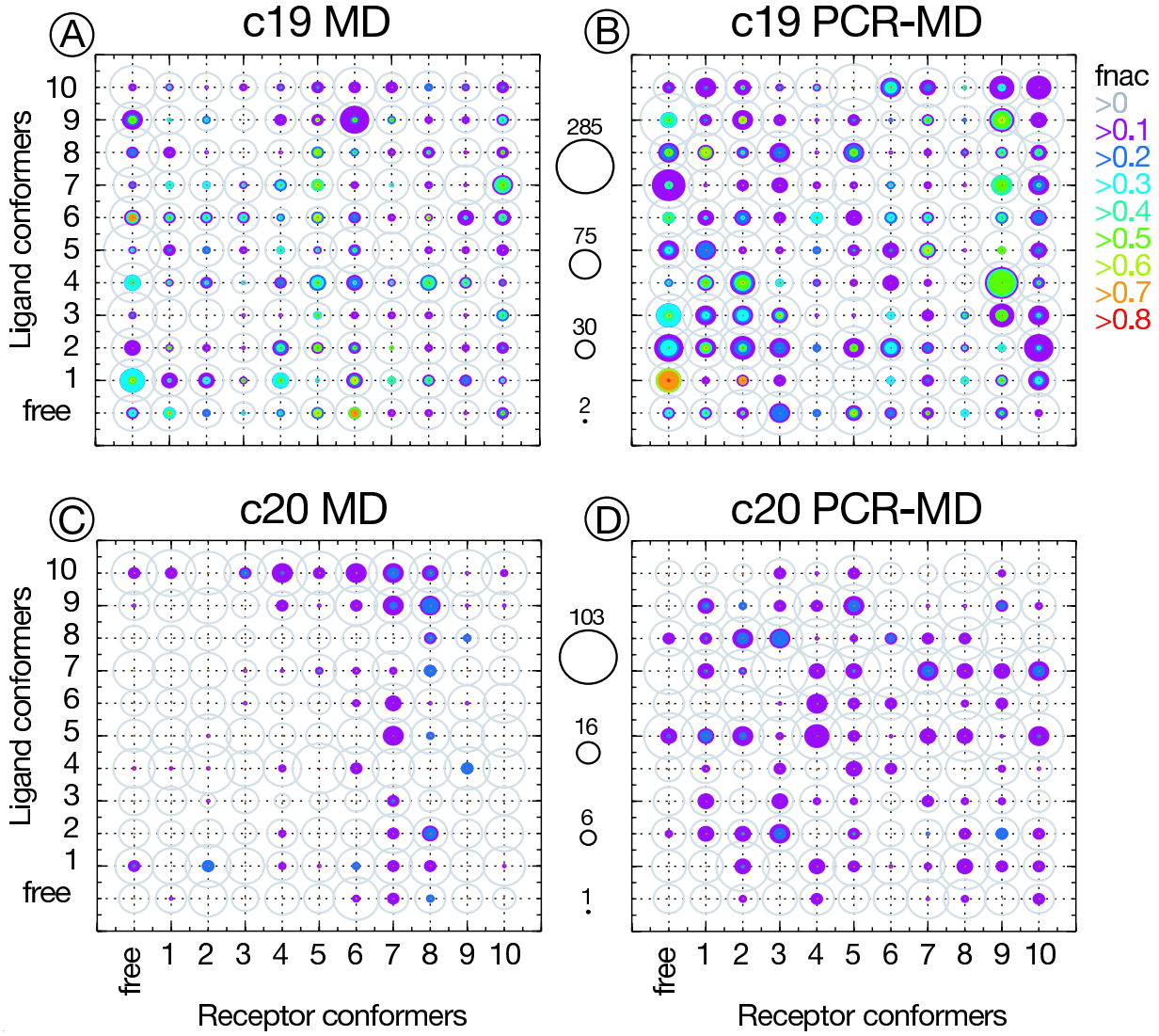

Figure 4: Quantity and quality of near-native solutions in 4 selected ensemble dockings.
The cross-docking of 11 receptor and 11 ligand conformations generates 121 sets of 512 docking solutions. The amount and quality of near native solutions among each set is shown for the ensemble dockings of c19 (A and B) and c20 (C and D). The area of each contour is proportional to the number of solutions (see the separate size legends). The color of a contour indicate solutions above a certain fnac-value (see the color legend). Several conformer-combinations perform better than the traditional docking of the free structures.

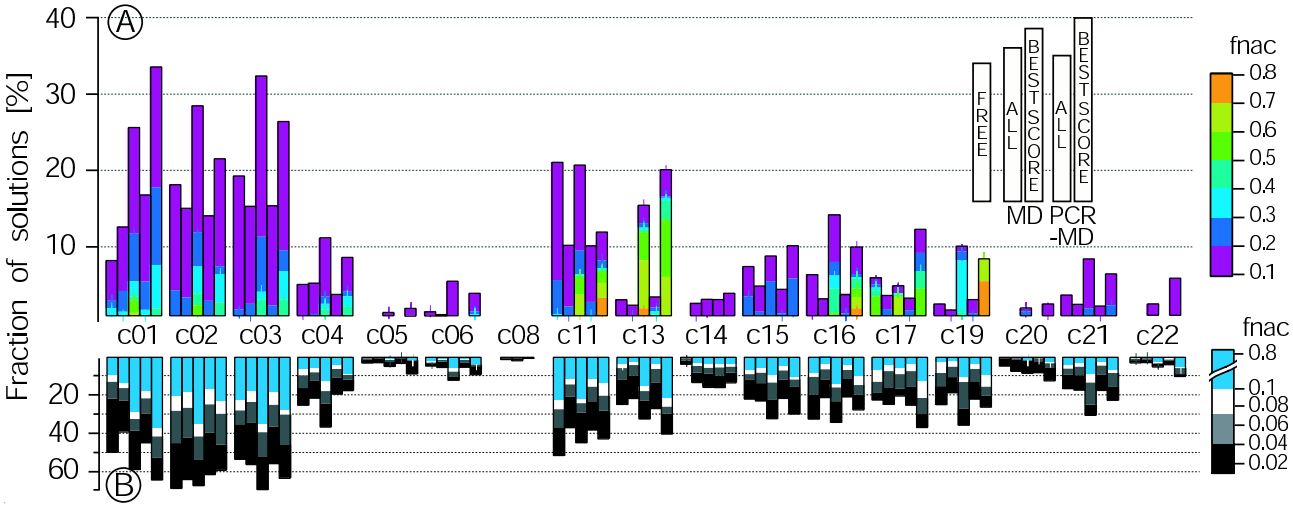

Figure 5: Quantity and quality of near-native solutions in all test cases.
The amount of solutions above a certain quality (fnac) level (see color legend) is given for selected docking runs of all 17 test complexes. Data for each complex are presented using groups of five bars. The first bar describes the free docking (512 orientations), the second and fourth bar show the data for all cross dockings (11 x 11 x 512 orientations). Bars three and five show the data for the best performing conformer-combination (512 orientations) from the MD and PCR-MD ensemble, respectively (see table 3). The upper plot (A) depicts solutions with fnac above 10%, while the lower plot (B) uses a 1% fnac threshold.

short range forces and internal dynamics become important for recognition. Specifically matching conformations will select each other from the free conformation ensembles of the two proteins and form a recognition complex (Rf*Lf*). The recognition complex will quickly be stabilized by progressive desolvation as well as short range electrostatic and van der Waals interactions. At this stage the receptor and ligand structure cannot any longer be considered independent. They are now moving in concert through a potential that has changed from the free to the bound energy landscape. The equilibration into this new landscape requires the transition from the (free) recognition conformations to the more dominant states of the bound structure ensemble (RbLb). This is potentially a time consuming step, depending on the distance between free and bound structure (or the probability of the recognition conformations in the context of the bound energy landscape) and may be considered a folding process.

In figure 7 we attempt to give a schematic view on the free energy profile and the forces that are involved, and compensate each other, at the proposed stages of protein-protein association. This reaction scheme extends earlier 3- and 4-state models (Camacho et al. 2000; Frisch et al. 2001; Schreiber 2002) and combines them with the idea of conformer selection (Monod et al. 1965; Kumar et al. 2000; Gabdoulline and Wade 2001). Existing 4-state models (Camacho et al. 2000; Schreiber 2002) distinguish between the formation of an unspecific (randomly aligned) encounter complex on one side and its correct orientation on the other. For the sake of clarity, we combine these two steps into one. The search for this correctly aligned encounter complex (Rf:Lf) was considered the rate limiting barrier in the previous models. We introduce an additional step of free conformer selection that separates the diffusive search for a correct orientation from the conformational search for the bound state. Both diffusive and conformational search are well studied in isolation - the former by simulations and experiments on diffusion-controlled associations (Gabdoulline and Wade 2002) and the latter by decades of research on protein folding (Dill and Chan 1997). Conformer selection has been observed in experiments (e.g. (Lancet and Pecht 1976; Foote and Milstein 1994; Leder et al. 1995; Berger et al. 1999) and our results suggest the specific recognition via a subset of free conformations. Moreover, the mechanism does not rely on the ad-hoc assumption of preexisting bound conformations and is compatible with the time scale and typical rates of protein-protein association.

The scheme contains the previous models as border cases among several possible kinetic regimes: If the free energy cost of selecting matching conformers is much lower than the cost of finding the correct orientation ($k1 << k2$), the model reverts to the previous 3- or 4-state descriptions (with- or without induced fit, respectively) of a diffusion-controlled reaction. If, on the other hand, we assume that recognition requires bound conformers, the refolding barrier (III in figure 7) would be absent ($k2 << k3$) and we would revert to the preexisting equilibrium model. The proposed 3-step model is the general description of an interaction that can be diffusion controlled, recognition controlled, refolding controlled, or be influenced by a mixture of the three rates.

2.9 Implications of the model

Diffusion-controlled associations have been studied experimentally and relative rates for a given system under different conditions can in many cases be reproduced by Brow-

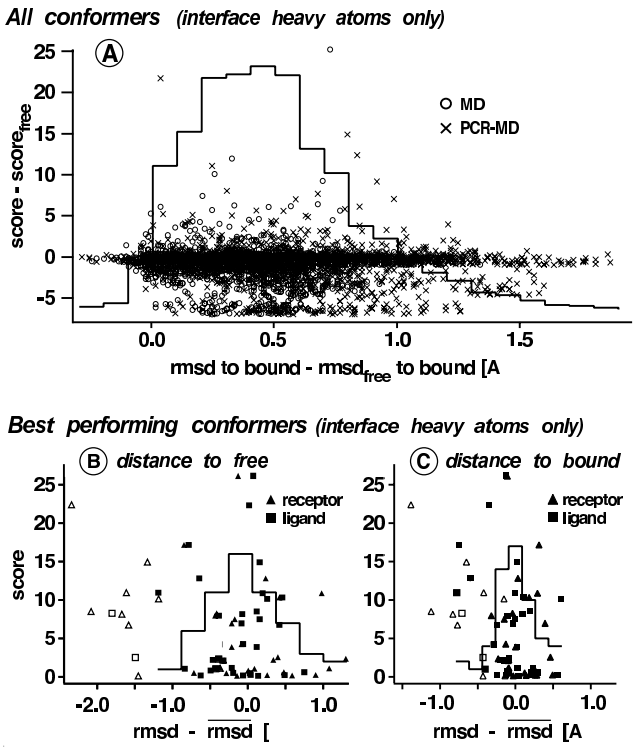

Figure 6: Docking performance and structural changes in the interface.
In panel (A) the combined distance of the receptor and ligand interface regions from their respective bound conformation is expressed as (rmsdrec + rmsdlig)/2 and is plotted against the pair's docking performance. Both values are given relative to the docking of the free conformation pair. Data is shown for each combination of receptor and ligand conformers (11 x 11 x 34). A solid line describes the distribution of rmsd values (distances to the bound structure). Panels (B) and (C) show only the best performing pairs of each ensemble docking. The rmsd of the receptor (triangle) and ligand (square) interface to the free (B) and to the bound structure (C) is given relative to the respective average value of the 10 simulation-derived conformers. High performing conformations seem to be shifted both towards the bound and the free structure. This trend is largely caused by free (experimental) structures (open symbols) that are over-represented on the receptor side of high-performing conformer pairs. Free structures are excluded from the distribution of rmsd shifts (solid lines).

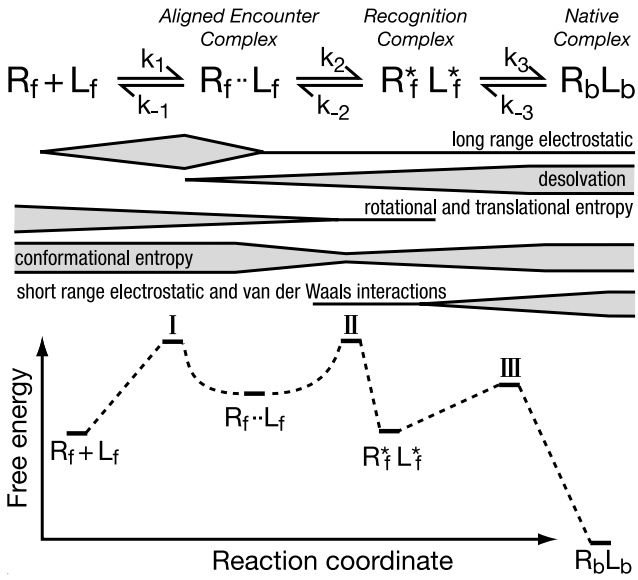

Figure 7: A working model for flexible protein recognition. Protein-protein association may be governed by diffusion, selection of matching conformers, and refolding. Rf and Lf are the free structure ensembles of receptor and ligand, respectively. R*f and L*f are sub-sets of the free receptor and ligand ensembles (recognition conformers). The middle and lower section of the figure suggest, schematically, the forces involved at the different stages and the resulting free energy profile. The widths and barrier heights are not meant to reflect real proportions.

nian Dynamics simulations (Gabdoulline and Wade 2002). An issue with simulations is that association rates are usually overestimated, even if binding is assumed only for orientations very close to the native. Gabdoulline and Wade (2001) showed that this overestimation was different for 5 different protein complexes and concluded that association can be influenced by non-diffusive effects. For the binding of fasciculin-II to acetylcholinesterase in particular, they suggested a mechanism of "conformal gating" by two distinct conformations of a loop. Our working model of diffusion, selection and refolding offers a similar, more general explanation. The recognition barrier (barrier II in figure 7) differs from the free energy of the encounter complex ensemble Rf:Lf by a loss of conformational entropy because it can only be crossed by a sub-set of free conformations. A mixed control by diffusion and recognition should lower observed association rates by a systematic factor (related to the frequency of recognition conformers), such as described by Gabdoullin and Wade. Predominant control by recognition and/or refolding, on the other hand, would uncouple the observed rate from conditions like ionic strength, charge, and viscosity - which they demonstrated for another of the tested complexes.

The 3-step model also helps to refine our description of the transition state ensemble(s) in protein-protein association. Both theoretical (Janin 1997; Camacho et al. 2000) and experimental studies (Frisch et al. 2001) conclude that the transition state closely resembles the structure of the final complex. Less clear is whether or not desolvation is necessary for recognition. According to Camacho et al. (1999, Camacho et al. (2000) partial desolvation is important for the correct positioning and initial stabilization of the encounter complex. However, Frisch et al. (2001) measured activation entropies close to zero for the association of barnase and barstar. They hence assumed that the activated complex remains mostly solvated. This discrepancy may testify to a "special" nature of the barnase - barstar interface (featuring many charged residues and structural waters). It may on the other hand also result from underlying conformer recognition. Following our 3-step model,

recognition occurs at the cost of conformational entropy. A low activation entropy does not rule out desolvation effects but could rather reflect a balance between conformational entropy loss and solvent entropy gain. Our comparison of free and bound MD simulations shows that bound structure ensembles are not generally less diverse than free ones (unpublished results). One can hence speculate that the refolding phase of binding is accompanied by the regain of conformational entropy. A mixed control by diffusion and recognition implies a structurally constrained transition state ensemble that is close to the bound orientation on the one hand, but resembles the two free conformations on the other hand.

2.10 Implications for predictive docking

The idea to use pre-generated conformer libraries of receptor or ligand have already a while ago been implemented for the docking of small molecules against proteins (reviewed by Brooijmans and Kuntz (2003). For predictive protein-protein docking similar strategies are now being tested in many labs. Multiple MD simulations enhanced by principal component restraints is a promising technique for sampling relevant structure ensembles. The application of conformers from such simulations of the free ensemble could be a viable strategy to account for protein flexibility in protein-protein docking. Ensemble cross-docking appears to rather increase the challenge to identify a correct orientation within the large number of false ones - an issue that we do not address in this paper. However, compared to classic single rigid body docking, ensemble docking generates many more solutions with (by comparison) excellent shape complementarity. Therefore it is possible to refine and evaluate candidate orientations with more accurate energy functions that are less forgiving to steric clashes and other artifacts which otherwise have to be tolerated.

2.11 Conclusion

We here examined the impact of overall protein flexibility on protein-protein recognition. The cross-docking of ensemble snapshots derived from MD simulations of the two partner proteins increases the chances to find near native solutions. There appear to exist multiple complementary conformations within the free structure ensembles. Our results suggest that recognition does not depend on the bound structure and such a dependence would also be inconsistent with the time scale of typical protein-protein associations. We propose that protein-protein binding follows a 3-step mechanism of diffusion, free conformer selection and refolding. This working model is an extension and combination of earlier ideas and models. In particular, we mix and generalize the previously conflicting mechanisms of diffusion-controlled binding with passive induced fit on the one hand, and the recognition via preexisting conformations on the other. The combined mechanism appears consistent with current data from simulations and experiments on protein-protein association. However, most of these studies have so far focused on diffusion-controlled interactions without large changes in protein structure. It is now time to move on to systems where association could be dominated by the selection of matching conformers and where recognition is either depending on or followed by large-scale structural re-arrangements.

3 Experimental procedures

3.1 Conformational sampling

Simulations were performed with a modified version of X-PLOR (Brünger 1992; Abseher and Nilges 2000) using the CHARMM19 force field (Brooks et al. 1983) and an electrostatic cutoff of 12 Å with force shifting (Steinbach et al. 1991).

The coordinates of the 51 molecules (table 1) were retrieved from the Protein data bank (Berman et al. 2002). An automated procedure removed duplicate peptide chains and all hetero atoms (but not waters), converted non-standard amino acids to their closest standard residue and identified disulfide bonds. Missing atoms, including polar hydrogens were added and briefly minimized. The protein was surrounded by a 9 Å layer of TIP3 water molecules and the solvent briefly equilibrated. 10 copies were starting point for parallel simulations of 50 ps length summing up to 500 ps total simulation time per system. SHAKE constraints (van Gunsteren and Berendsen 1977) were put on all bonds to hydrogens and on all TIP3 waters. Each copy was heated from 100 K to 300 K in 50 K steps of 1 ps each, followed by additional 5 ps of equilibration with continued re-assignment of velocities every 1 ps. The temperature was kept constant by explicit coupling to a heat bath via Langevin dynamics and a friction coefficient of 20 ps^{-1} for water oxygens and between 0.5 and 5.5 ps^{-1} for protein atoms dependent on their solvent accessible area. A time step of 2 fs was used. The simulation scripts are available upon request.

A second set of simulations was performed with identical setup but adding an additional force onto the potential acting along the principle components of motion, basically as described by Abseher and Nilges (2000). In difference to the published method we re-defined the principal components iteratively during the calculation. Details will be published elsewhere.

100 snapshots spaced 5 ps apart were taken from each 10 trajectories. The snapshots were fitted to their average structure and divided into 10 groups by c-means fuzzy clustering (Gordon and Somorjai 1992) over the coordinates of backbone carbonyl carbon and every second side chain carbon. The clustering method is similar to the simple k-means but gives each item a continuous membership to each cluster instead of a binary membership to one. From each cluster the structure nearest to the center was selected for docking.

3.2 Docking

All protein-protein docking calculations were performed with HEX version 4.2 (Ritchie and Kemp 2000). Orien-

tations were discriminated by shape complementarity only. For all protein independent parameters the default values provided by HEX were used with the exception of the distance range step, which was set to 0.5 Å and the receptor and ligand samples which were set to 720 (Ritchie and Kemp 2000). The initial molecular separation and the distance range to be sampled were calculated from the maximal and minimal distance from the center of mass to any surface atom (any atom with an exposure >95% as determined by WhatIf (Vriend 1990)). In the 7 cases where the receptor had a radius larger than 35 Å HEX "macro docking" was performed with default parameters, i.e. the program docked the ligand to several overlapping fragments of the receptor (Ritchie 2003). The 512 highest scoring solutions were retained from each docking, thus the combination of 11 x 11 conformations always produced 61952 orientations. The docking of a single conformer pair took in the order of 15 min on a dual 2.4 GHz Xeon computer but lasted about 8 h for the "macro docking" cases.

3.3 Randomized reference complexes

For each ligand we generated 100 transformation matrices with randomized euler angles and a random translation onto a sphere around the receptor's center of mass. These randomized orientations were each subjected to 100 steps of rigid body minimization using a soft van der Waals potential and a NOE restraint pulling the two centers of mass together (X-Plor script available upon request). We removed all orientations having any atom contact in common with the native complex (fnac > 0) and performed a hierarchical clustering by the pairwise overlap of atom contacts. The clustering will be described in detail elsewhere. For the present purpose, we applied a clustering threshold of 0.0001 and obtained a set of cluster centers without mutual contact overlap. We selected 10 at random and re-calculated the "fnac" of all HEX solutions with respect to each of the 10 random complexes. From these values we estimated the probability of the score (for reproducing the native complex) being a random observation (details are in supplement S4). The necessary random distribution cannot be deduced from 10 values. However, score values were by definition positive and usually small. A lognormal distribution was hence the least biased assumption.

3.4 Analysis of docking results

All atoms not present in both free and bound receptor or ligand structure were removed before performing the analysis. The interaction interface was defined as any residue with any atom within 4.5 Å from the other molecule. Calculations (docking and analysis) were distributed to between 30 and 90 processors of a Linux cluster. The total computation time for this study amounts to about 8 years on a single 2.4 GHz processor.

3.5 Figures

Figure 1 was prepared with MOLMOL (Koradi et al. 1996). Figures 2 through 4 were created using Biggles (biggles.sourceforge.net) and figure 5 and 6 using IgorPro (www.wavemetrics.com).

4 Appendix

4.1 Minimum frequency of recognition conformations

According to the preexisting equilibrium model, protein recognition relies on the simultaneous occurrence of bound conformations both in receptor and ligand ensemble. The recognition probability R of a correctly aligned microcollision should depend on the average frequencies $\langle fr \rangle$ of recognition conformations in the free ensembles. The probability of recognition failure can be estimated as:

$$1 - R \quad = \quad (1 - \langle fr \rangle^2)^N$$

N is the number of distinct conformations sampled in the course of the correct alignment. The frequency of recognition conformations which is needed for a certain recognition rate is then

$$\langle fr \rangle = \sqrt{1 - \exp(\ln(1-R)/N)}$$

N depends on the lifetime τ of the alignment and on our definition of distinct conformations. The short recognition time will only allow for fairly limited sampling in the flat energy landscape of protein structures. For the sake of simplicity, we assume that N depends linearily on the recognition time τ and that the "recognizability" of a given protein structure changes every 1 ps ($N = \tau/ps$). We thus arrive at the estimates given in the introduction.

4.2 Specifity estimate for docking scores

Given is the score s for the success of a docking experiment to reproduce the native complex and the scores r_1 .. r_{10} to reproduce 10 non-native random complexes. We assume random scores to follow a lognormal distribution. The lognormal density function $f(x)$ can be estimated from the mean α and the standard deviation β of the 10 log-transformed random scores.

$$\alpha = \frac{1}{n} \sum_{i=1}^{n} \ln r_i \tag{1}$$

$$\beta = \sqrt{1/(n-1)} \sum_{i=1}^{n} (\ln r_i - \alpha)^2 \tag{2}$$

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} \exp(-\frac{1}{2}[\frac{\ln x - \alpha}{\beta}]^2) \tag{3}$$

Given $f(x|\alpha,\beta)$, we determined the confidence level κ of the smallest interval still containing s.

$$\kappa = \int_{r_{max}^2/s}^{s} f(x|\alpha,\beta)dx \quad (4)$$

$$r_{max} = \exp(\alpha - \beta^2) \quad (5)$$

$$\kappa = \frac{1}{2}\mathrm{erf}\left(\frac{\ln s - \alpha}{\sqrt{2\beta^2}}\right) - \frac{1}{2}\mathrm{erf}\left(\frac{\ln(r_{max}^2/s) - \alpha}{\sqrt{2\beta^2}}\right) \quad (6)$$

κ is the probability that s is not a random observation such as r_1 .. r_{10}.

Acknowledgements

J.L. and R.G. were supported by fellowships of the Knut and Alice Wallenberg foundation and the Boehringer Ingelheim Fonds, respectively. M.N. acknowledges support from the E.U. (QLG2 CT00-01313). We thank D. Ritchie for helpful discussions and his prompt assistance with any HEX-related problems. We also appreciate useful comments from J. Janin. We thank Tru Hyn for keeping our computer software and hardware running. We are grateful to Michael Habeck and Wolfgang Rieping in particular for help with statistics, structure clustering and parallelization of calculations.

References

Abagyan, R. and M. Totrov (1997). Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J Mol Biol 268*(3), 678–85.

Abseher, R. and M. Nilges (2000). Efficient sampling in collective coordinate space. *Proteins 39*(1), 82–8.

Balsera, M., W. Wriggers, Y. Oono, and K. Schulten (1996). Principal component analysis and long time protein dynamics. *J Phys Chem 100*, 2567–72.

Berger, C., S. Weber-Bornhauser, J. Eggenberger, J. Hanes, A. Pluckthun, and H. Bosshard (1999). Antigen recognition by conformational selection. *FEBS Lett 450*(1-2), 149–53.

Berman, H., T. Battistuz, T. Bhat, W. Bluhm, P. Bourne, K. Burkhardt, Z. Feng, G. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. Westbrook, and C. Zardecki (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr 58*(Pt 6 No 1), 899–907.

Betts, M. and M. Sternberg (1999). An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng 12*(4), 271–83.

Bosshard, H. (2001). Molecular recognition by induced fit: how fit is the concept? *News Physiol Sci 16*, 171–3.

Brooijmans, N. and I. Kuntz (2003). Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct 32*, 335–73.

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

Brooks III, C., M. Karplus, and B. Pettitt (1988). *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics*. New York: Wiley.

Brünger, A. (1992). *X-PLOR. A System for X-ray crystallography and NMR*. New Haven: Yale University Press.

Camacho, C., S. Kimura, C. DeLisi, and S. Vajda (2000). Kinetics of desolvation-mediated protein-protein binding. *Biophys J 78*(3), 1094–105.

Camacho, C., Z. Weng, S. Vajda, and C. DeLisi (1999). Free energy landscapes of encounter complexes in protein-protein association. *Biophys J 76*(3), 1166–78.

Caves, L., J. Evanseck, and M. Karplus (1998). Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci 7*(3), 649–66.

Chen, R., J. Mintseris, J. Janin, and Z. Weng (2003). A protein-protein docking benchmark. *Proteins 52*(1), 88–91.

Chothia, C. and A. Lesk (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J 5*(4), 823–6.

Dill, K. and H. Chan (1997). From Levinthal to pathways to funnels. *Nat Struct Biol 4*(1), 10–9.

Ehrlich, L. P., M. Nilges, and R. Wade (2004). The impact of protein flexibility on protein-protein docking. *Proteins in press*.

Fernandez-Recio, J., M. Totrov, and R. Abagyan (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins 52*(1), 113–7.

Fischer, E. (1894). Einfluss der configuration auf die wirkung der enzyme. *Ber Dt Chem Ges 27*, 2985–93.

Foote, J. and C. Milstein (1994). Conformational isomerism and the diversity of antibodies. *Proc Natl Acad Sci U S A 91*(22), 10370–4.

Frauenfelder, H. and D. Leeson (1998). The energy landscape in non-biological and biological molecules. *Nat Struct Biol 5*(9), 757–9.

Frauenfelder, H., S. Sligar, and P. Wolynes (1991). The energy landscapes and motions of proteins. *Science 254*(5038), 1598–603.

Frisch, C., A. Fersht, and G. Schreiber (2001). Experimental assignment of the structure of the transition state for the association of barnase and barstar. *J Mol Biol 308*(1), 69–77.

Gabdoulline, R. and R. Wade (2001). Protein-protein association: investigation of factors influencing association rates by brownian dynamics simulations. *J Mol Biol 306*(5), 1139–55.

Gabdoulline, R. and R. Wade (2002). Biomolecular diffusional association. *Curr Opin Struct Biol 12*(2), 204–13.

Goh, C., D. Milburn, and M. Gerstein (2004). Conformational changes associated with protein-protein interactions. *Curr Opin Struct Biol 14*(1), 104–9.

Goodford, P. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem 28*(7), 849–57.

Gordon, H. and R. Somorjai (1992). Fuzzy cluster analysis of molecular dynamics trajectories. *Proteins 14*(2), 249–64.

Gray, J., S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. Rohl, and D. Baker (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol 331*(1), 281–99.

Halperin, I., B. Ma, H. Wolfson, and R. Nussinov (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins 47*(4), 409–43.

Janin, J. (1997). The kinetics of protein-protein recognition. *Proteins 28*(2), 153–61.

Kimura, S., R. Brower, S. Vajda, and C. Camacho (2001). Dynamical view of the positions of key side chains in protein-protein recognition. *Biophys J 80*(2), 635–42.

Kirschner, K., M. Eigen, R. Bittman, and B. Voigt (1966). The binding of nicotinamide adenine dinucleotide to yeast glyceraldehyde-3-phosphate dehydrogenase: Temperature-jump relaxation studies on the mechanism of an allosteric enzyme. *Proc Natl Acad Sci USA 56*, 1661–67.

Koradi, R., M. Billeter, and K. Wuthrich (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph 14*(1), 51–5, 29–32.

Koshland, D. (1958). Application of a theory of enzyme specifity to protein synthesis. *Proc Natl Acad Sci USA 44*, 98–104.

Kumar, S., B. Ma, C. Tsai, N. Sinha, and R. Nussinov (2000). Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci 9*(1), 10–9.

Kuntz, I., J. Blaney, S. Oatley, R. Langridge, and T. Ferrin (1982). A geometric approach to macromolecule-ligand interactions. *J Mol Biol 161*(2), 269–88.

Lancet, D. and I. Pecht (1976). Kinetic evidence for hapten-induced conformational transition in immunoglobin MOPC 460. *Proc Natl Acad Sci U S A 73*(10), 3549–53.

Leder, L., C. Berger, S. Bornhauser, H. Wendt, F. Ackermann, I. Jelesarov, and H. Bosshard (1995). Spectroscopic, calorimetric, and kinetic demonstration of conformational adaptation in peptide-antibody recognition. *Biochemistry 34*(50), 16509–18.

Lo Conte, L., C. Chothia, and J. Janin (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol 285*(5), 2177–98.

Mendez, R., R. Leplae, L. De Maria, and S. Wodak (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins 52*(1), 51–67.

Monod, J., J. Wyman, and J. Changeux (1965). On the nature of allosteric transitions: a plausible model. *J Mol Biol 12*, 88–118.

Northrup, S. and H. Erickson (1992). Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc Natl Acad Sci U S A 89*(8), 3338–42.

Ritchie, D. (2003). Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins 52*(1), 98–106.

Ritchie, D. and G. Kemp (2000). Protein docking using spherical polar Fourier correlations. *Proteins 39*(2), 178–94.

Schneidman-Duhovny, D., Y. Inbar, V. Polak, M. Shatsky, I. Halperin, H. Benyamini, A. Barzilai, O. Dror, N. Haspel, R. Nussinov, and H. Wolfson (2003). Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins 52*(1), 107–12.

Schrauber, H., F. Eisenhaber, and P. Argos (1993). Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol 230*(2), 592–612.

Schreiber, G. (2002). Kinetic studies of protein-protein interactions. *Curr Opin Struct Biol 12*(1), 41–7.

Selzer, T. and G. Schreiber (2001). New insights into the mechanism of protein-protein association. *Proteins 45*(3), 190–8.

Steinbach, P., R. Loncharich, and B. Brooks (1991). The effects of environment and hydration on protein dynamics: A simulation study of myoglobin. *Chem Phys 158*, 383–94.

van Gunsteren, W. and H. Berendsen (1977). Algorithms for macromolecular dymaics and constraint dynamics. *Mol Phys 34*, 1311–27.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J Mol Graph 8*(1), 52–6, 29.

Zhou, H. (2001). Disparate ionic-strength dependencies of on and off rates in protein-protein association. *Biopolymers 59*(6), 427–33.