doi:10.1016/j.jmb.2005.01.058 *J. Mol. Biol.* (2005) **347**, 1077–1101

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Relationship between the Flexibility of Proteins and their Conformational States on Forming Protein–Protein Complexes with an Application to Protein–Protein Docking

Graham R. Smith[1], Michael J. E. Sternberg[2] and Paul A. Bates[1]*

[1] *Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute Lincoln's Inn Fields Laboratories 44 Lincoln's Inn Fields, London WC2A 3PX, UK*

[2] *Department of Biological Sciences, Imperial College of Science, Technology and Medicine, South Kensington London SW7 2AZ, UK*

**Corresponding author*

We investigate the extent to which the conformational fluctuations of proteins in solution reflect the conformational changes that they undergo when they form binary protein–protein complexes. To do this, we study a set of 41 proteins that form such complexes and whose three-dimensional structures are known, both bound in the complex and unbound. We carry out molecular dynamics simulations of each protein, starting from the unbound structure, and analyze the resulting conformational fluctuations in trajectories of 5 ns in length, comparing with the structure in the complex.

It is found that fluctuations take some parts of the molecules into regions of conformational space close to the bound state (or give information about it), but at no point in the simulation does each protein as whole sample the complete bound state. Subsequent use of conformations from a clustered MD ensemble in rigid-body docking is nevertheless partially successful when compared to docking the unbound conformations, as long as the unbound conformations are themselves included with the MD conformations and the whole globally rescored. For one key example where subdomain motion is present, a ribonuclease inhibitor, principal components analysis of the MD was applied and was also able to produce conformations for docking that gave enhanced results compared to the unbound.

The most significant finding is that core interface residues show a tendency to be less mobile (by size of fluctuation or entropy) than the rest of the surface even when the other binding partner is absent, and conversely the peripheral interface residues are more mobile. This surprising result, consistent across up to 40 of the 41 proteins, suggests different roles for these regions in protein recognition and binding, and suggests ways that docking algorithms could be improved by treating these regions differently in the docking process.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: molecular dynamics; protein–protein interactions; protein–protein docking; protein–protein interfaces; conformational transitions

Introduction

The formation of protein–protein complexes is perhaps the most common event by which biochemical function is mediated. Structural information is extremely valuable both in giving insight into the biochemical nature of the process for which the components come together, and in facilitating the design of compounds that might influence it, but there is relatively little structural information available about complexes compared to those

Present address: G. R. Smith, Biosystems Informatics Institute, 2nd Floor, Bioscience Centre, International Centre for Life, Times Square, Newcastle upon Tyne, NE1 4EP, UK.

Abbreviations used: MD, molecular dynamics; PDB, Protein Data Bank; ED, essential dynamics; PCA, principal component analysis; RMSD, root mean squared deviation.

E-mail address of the corresponding author: paul.bates@cancer.org.uk

0022-2836/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.

proteins that exist as single chains or form permanent oligomers, mainly due to the greater difficulty in crystallizing them.

Therefore, there has been much activity in developing computational methods of predicting the structure of protein–protein complexes, the protein–protein docking problem.[1–3] Recent studies have focused on the case where the structure of one or both components is known in its unbound state, when the task is made much more difficult by the conformational changes that the unbound proteins undergo when they dock together. These may be large (e.g. cyclin-CDK, ≈4 Å) but even those which are quite small (e.g. a couple of Å RMSD between the bound and unbound forms on all heavy atoms; 1 Å or less on C^{α} atoms) are enough to reduce shape complementarity to a level where it alone can no longer reliably predict the correct form of the complex.

To dock protein components for which some conformational change needs to occur, some methods have been tried that use "soft" scoring functions that accommodate flexibility[4,5] while others explicitly include domain hinging movements[6,7] or side-chain flexibility in the docking.[8–13] The newest approaches have also begun to take flexibility in the protein backbone into account.[14]

Moreover, in relation to the high-throughput generation of protein–protein interactions, the structures of both interacting partners will be known in only a small fraction of cases, but for many more it will be possible to build homology models that may be expected to differ, for a single conformer, from the true structure by 1–3 Å at relatively high levels of homology (40–100% ID).[15] It would therefore be highly desirable to be able to dock such models as part of the validation of the putative protein–protein interactions. Some early work has been done on this.[16] In recent work, it has been shown that, if both components are homologous to components of a known complex, then this information can be used to drastically restrict the conformational search.[17] In an effort to facilitate progress on predictive docking, an international blind trial, CAPRI, has recently been set up†.[18,19]

Conformational changes of the size discussed above are within the size of the fluctuations commonly observed in molecular dynamics (MD) computer simulations, which also rely on atom-based force fields but can include water explicitly, thus increasing the accuracy of the modeling of the crucial hydration effects, and allow flexibility of any kind. Motivated by this, we investigate in detail the fluctuations of the unbound conformations of complex-forming proteins with a particular view to determining the extent to which they resemble the conformational changes on complex formation. There is no reason to expect the simulation to fall exactly into the correct bound state; the conformational change is expected to be, in large measure, induced by the binding partner, but this change takes place on microsecond to millisecond timescales which are as yet inaccessible to classical MD simulation (even though the components may become irrevocably committed to complex formation much more quickly).[2,20,21] Moreover, it is important to point out that MD, and the energy functions that MD calculations employ, capture only imperfectly important aspects of the real energetics of proteins (polarization effects for example). This will certainly affect the details of the results we obtain. Nevertheless we believe that MD can be a valuable tool in addressing the following, important, question: is the conformation of the bound component of the complex within the ensemble of the conformations of the unbound component (the theory of "pre-existing conformational ensembles")[22–24] or are certain conformations never sampled in the absence of the binding partner; in other words, what is the extent of induced fit?

The previous studies that most resemble the current work are from the group of Vajda & Camacho.[25,26] In the first study,[25] certain key side-chain rotamers of six unbound proteins were examined. In three cases side-chains from the interface of the complex visit the rotamers that they occupy in the complex, but only in one case does the side-chain in the simulation spend the majority of its time in the rotamer from the complex. Recently a more extensive analysis has been carried out with longer simulations on 11 proteins.[26] The authors suggest that one or a few key ("anchor") residues frequently sample their bound state and may be critical in the early recognition stage of docking.

Strategy

In this study, a set of 22 protein–protein complexes are used, for each of which the complex is of known structure, and the structures of the components in their unbound states have also been experimentally determined by X-ray crystallography. We have carried out MD of the components, starting each from its unbound state, and compared the structures that are formed along the MD trajectory with the structure of the protein in the complex. We emphasize that each protein is simulated alone in this study; its binding partner is not included, so this work shows to what extent the conformational changes on docking reflect conformational flexibility inherent in the isolated component proteins. Three of the proteins appear in two protein–protein complexes with different partners, so altogether 41 complex-forming proteins have been simulated. The complexes and unbound components are listed in Table 1; most are also present in a published "Protein–Protein Docking Benchmark",[27] with some differences as described in the legend. The set we have simulated includes most of those characterized as

† capri.ebi.ac.uk

Table 1. List of the protein–protein complexes investigated, in order of increasing size of conformational change of interface residues on complex formation

Complex PDB	Receptor	Receptor PDB	Receptor all-atom interface RMSD (Å)	Ligand	Ligand PDB	Ligand all-atom interface RMSD (Å)	Total backbone RMSD in interface (Å)	Class
2PTC	Trypsin	2PTN	0.57	Pancreatic trypsin inhibitor	5PTI[a]	0.98	0.32	Enzyme-inhibitor
1WEJ	IgG1 E8 Fab fragment	1QBL	0.84	Cytochrome *c*	1HRC	1.78	0.32	Antibody-antigen
2SNI	Subtilisin	2ST1[b]	0.69	Novo chymotrypsin inhibitor 2	2CI2	1.41	0.37	Enzyme-inhibitor
2SIC	Subtilisin	2ST1[c]	0.60	BPN subtilisin inhibitor	3SSI	1.28	0.40	Enzyme-inhibitor
2VIR	Igg1-lamda Fab	1GIG[d]	0.73	Influenza virus hemagglutinin	2HMG[e]	3.12	0.41	Antibody-antigen
2PCC	Cytochrome *c* peroxidase	1CCA	1.34	Iso-1-cytochrome *c*	1YCC	1.06	0.44	Others
1BRC	Trypsin	1BRA	1.20	APPI	1AAP	1.40	0.44	Enzyme-inhibitor
1BGS[f]	Barnase	1A2P	1.05	Barstar	1A19	0.99	0.47	Enzyme-inhibitor
1UGH	Human uracil-DNA glycosylase	1AKZ	1.11	Inhibitor	1UGI	1.74	0.53	Enzyme-inhibitor
2KAI	Kallikrein A	2PKA	1.19	Trypsin inhibitor	5PTI[g]	1.40	0.70	Enzyme-inhibitor
1AHW	Antibody Fab 5G9	1FGN	1.11	Tissue factor	1BOY	1.29	0.71	Antibody-antigen
1AVZ	HIV-1 NEF	1AVV	1.13	FYN tyrosin kinase SH3 domain	1SHF	1.69	0.73	Others
1DQJ	Hyhel-63 Fab	1DQQ	1.22	Lysozyme	3LZT	1.69	0.73	Antibody-antigen
1FSS	Snake venom acetylcholinesterase	1VXR[h]	1.42	Fasciculin II	1FSC	1.65	0.75	Enzyme-inhibitor
1WQ1	RAS activating domain	1WER	1.44	RAS	5P21	1.61	0.83	Others
1MLC	IgG1 D44.1 Fab fragment	1MLB	1.01	Lysozyme	3LZT[i]	1.03	0.85	Antibody-antigen
1DFJ	Ribonuclease inhibitor	2BNH	1.72	Ribonuclease A	7RSA	1.13	1.04	Enzyme-inhibitor
1BVK	Antibody Hulys11 Fv	1BVL	1.41	Lysozyme	3LZT	2.19	1.22	Antibody-antigen
1CGI	Chymotrypsinogen	2CGA[j]	2.91	Pancreatic secretory trypsin inhibitor	1HPT	2.88	1.48	Enzyme-inhibitor
1KKL	HPr kinase/phosphatase	1JB1	1.70	Phosphocarrier protein Hpr	1SPH	1.09	2.53	Difficult
1FQ1	CDK2 cyclin-dependant kinase 2	1HCL[k]	6.05	KAP	1FPZ	1.94	3.55	Difficult
1FIN	CDK2 cyclin-dependant kinase 2	1HCL	7.76	Cyclin A	1VIN	1.37	4.66	Difficult

[a] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 5PTI contains one residue more than 6PTI and has higher resolution.
[b] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 2ST1 used instead of 1SUP, which has a distorted unit cell and was found to be unstable in MD.
[c] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 2ST1 used instead of 1SUP, which has a distorted unit cell and was found to be unstable in MD.
[d] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 1GIG is an unbound form of the antibody in the complex 2VIR.
[e] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 2HMG used instead of 2VIU.
[f] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 1BGS used instead of 1BRS.
[g] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 5PTI contains one residue more than 6PTI and has higher resolution.
[h] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 1VXR is longer than 2ACE and has higher resolution.
[i] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 3LZT has higher resolution than 1LZA.
[j] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 2CGA has higher resolution than 1CHG and fewer unresolved residues.
[k] Bound/unbound pairs correspond to the benchmark set of Chen *et al.*,[27] except for: 1HCL used instead of 1B39, which has ATP bound.

"unbound–unbound" in the benchmark set. Six are antibody–antigen and ten are enzyme–inhibitor, of which six are Kazal-type serine protease–inhibitor complexes. Those in the enzyme–inhibitor class that are excluded (nine complexes) are mainly either very similar to another complex or have a disorder-to-order conformational change associated with complex formation (e.g. the transducin–G-protein complex, Protein Data Bank (PDB) code 1GOT). The proteins considered here are all expected to form *binary* protein–protein complexes, although complexes containing more than two components are of course of great importance, and methods capable of dealing with them are being developed.[28]

The approach of simulating the unbound component but comparing with the bound, is represented schematically in Figure 1(a). It will be useful throughout to consider surface residues as non-interface or interface, and to subdivide the interface into core (that is almost completely buried in the complex) and periphery (that retains some solvent accessibility) (Figure 1 (b)). This is similar to

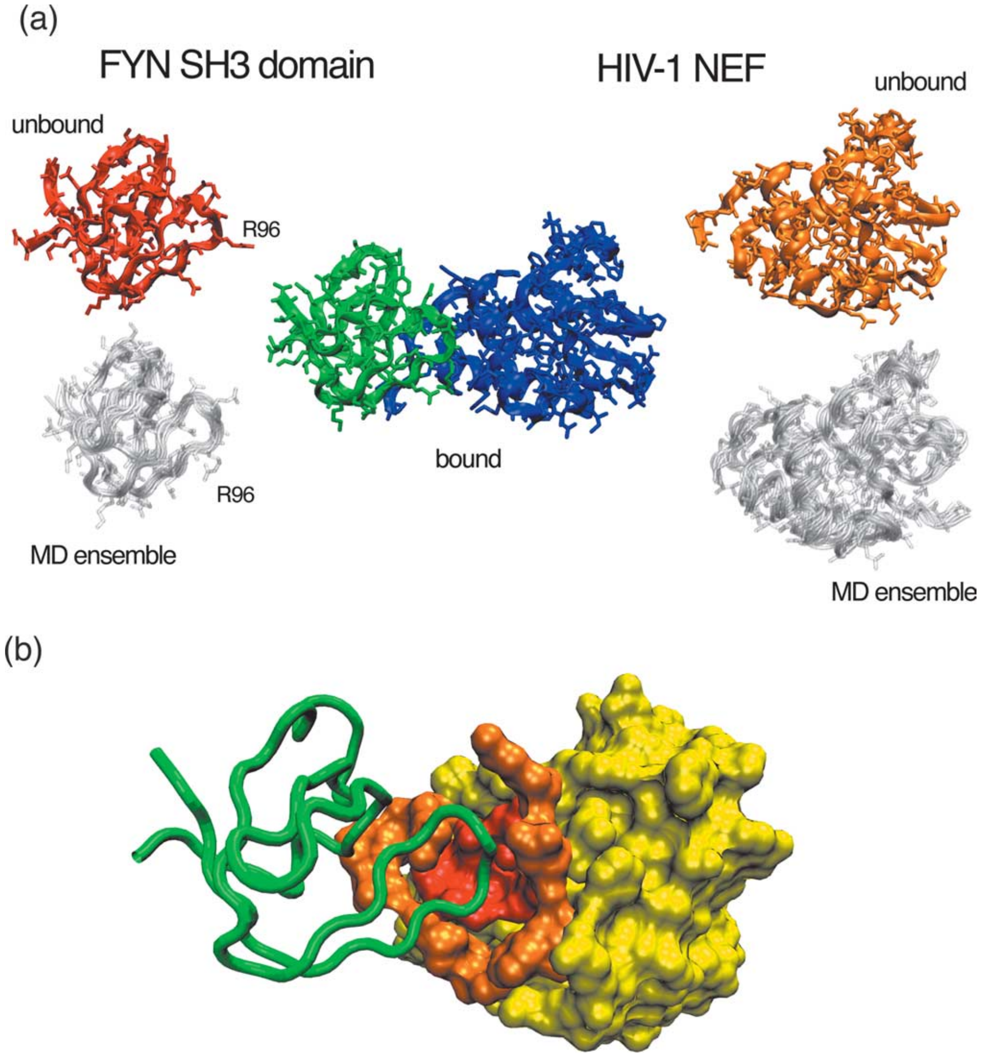


Figure 1. (a) Schematic Figure of protein complex formation. The example shown is FYN SH3 domain (green) in complex with HIV-1 NEF (blue) (PDB code1AVZ). The unbound X-ray structures of both proteins (red and orange, PDB codes 1SHF and 1AVV, respectively) are shown at the top left and top right. The backbone is shown in ribbon representation and side-chains in wire-frame. At the bottom left and right, ten configurations from the 5 ns MD trajectories of each protein (produced by RMS clustering) are shown as backbone traces (transparent white). One conformation from each MD ensemble also has side chains shown in wire-frame representation. At least one bulky side-chain in the interface, Arg96 of the SH3 domain, can be seen to have adopted a conformation in the MD that more closely resembles the bound form of the protein. (b) Figure showing the definitions of various kinds of surface residue for the NEF component. The FYN SH3 domain is shown as a green backbone trace. Surface residues of the NEF (shown in solvent-accessible surface representation) are defined as those that have greater than 5% of their solvent-accessible surface area exposed. The solvent-accessible surface area S is calculated using NACCESS[73] with a 1.4 Å probe, and is expressed as a fraction of S_0, where S_0 is the residue's accessibility in an extended G-X-G peptide. The first division is into residues that lie in the interface (closer than 5 Å to the other component when the protein is in the complex) and those that do not (yellow); then the interface is further divided into periphery (orange), where the residues retain some ($S/S_0 \geq 5\%$) solvent accessibility in the complex, and core (red), where they are completely buried.

the approach of Lo Conte *et al.*,[29] where interface atoms (not residues) were divided into those that make direct van der Waals contacts with the partner, and those that do not; the direct contact atoms being further divided into partially accessible and completely buried (core). Notwithstanding differences in detail, our definitions result in the assignment of 0.37 of the interface residues to the core, while in Lo Conte *et al.*,[29] a very similar fraction, 0.35, of the interface atoms are so assigned. The average number of residues in the core is 8.5 (± standard deviation 5), and the average number in the periphery is 14.0 (±5.3).

Having described the observed fluctuations we then investigate the use of conformations generated in the course of the MD runs, as well as the unbound components, in rigid protein–protein docking with the 3D-Dock suite. This uses surface-complementarity-based scoring in a rigid body search,[30,31] followed by rescoring[32,33] and an optional final side-chain refinement[8] (not used here). Although sometimes, in some regions, the MD simulations do approach closer to the bound conformations, it was necessary to confirm their usefulness directly, since there is not a clear correlation between the size of the conformational change on docking and the difficulty of predicting the complex.[34] This information is used in two different ways: the first is to soften the surface of the complexes; the second is to dock multiple conformations. Both these approaches have precedents, surface softening[4,30] and the use of multiple conformations,[4,5,35–38] but in work on protein–protein docking full-atomic MD has not been employed to generate the conformations (though it has in protein-small molecule docking).[39,40] We have aimed to treat the information as we would in a blind trial, without using our knowledge of the bound state to select the "best" MD conformation.

Results

Part I: description and analysis of the MD trajectories

Stability of MD trajectories and completeness of sampling

The first question to be addressed is whether the MD simulations are sufficiently stable during their course for reliable inferences to be extracted from them. It is usual for an MD trajectory to remain within an RMSD of 2–3 Å of the starting structure; excursions to a higher RMSD are usually regarded as worthy of further investigation, though they do not necessarily indicate technical problems: they could be the result of conformational changes, which are (of course) what we are endeavoring to study here. All the trajectories here keep within 4 Å of the starting structure, barring occasional fluctuations lasting of the order of a few hundred picoseconds, except for 2BNH, 1DQQ, 1MLB, 5PTI, 1JB1 and 2HMG (see Supplementary Table 1). Of these, the ribonuclease inhibitor (2BNH) seems to genuinely undergo a large and functionally important conformational change; the HPr kinase/phosphatase (1JB1) undergoes a large conformational change (perhaps due to being simulated in a different oligomeric state to its *in vivo* state) and experiences large fluctuations in a loop that was not resolved in the crystal structure and was subsequently modeled; 1DQQ and 1MLB are antibodies for which both the variable and constant domains were included in the simulation, and the large RMSD is the result of a hinge-bending between them; and 2HMG is a viral protein for which only one component of a trimer was included in the simulation. A trypsin inhibitor (5PTI) is the only protein for which the large RMSD cannot be readily explained. Another method of determining the stability of a simulation is by the total secondary structure content; all but four simulations retain more than 90% of their initial secondary structure (see Supplementary Table 1).

Since the 5 ns of each simulation is a short timescale for structural changes in proteins, it is important to investigate how extensive the sampling achieved in this time is. There is no method guaranteed to do this, but one way to attempt it is to carry out a principal components analysis (see below and in Methods for further use and explanation of this technique) on the first and second halves of the trajectory, and see how much overlap there is between the two sets of eigenvectors. The first few eigenvectors define the "essential subspace" in which the majority of the motion of the protein takes place, and this subspace often seems to converge quite fast.[41] The overlap of the first five eigenvectors of the first set and the first 50 of the second (a large number, chosen to effectively span the whole of the essential subspace of the second half) is generally high (0.77±0.1), indicating that the essential subspace is quite well defined even on the timescale of a few nanoseconds (Supplementary Table 1).

Does MD sample near both the bound and unbound conformational states?: the global RMSD distribution

The most straightforward way to investigate the similarity of the unbound and bound conformation to the MD is simply to RMS-fit the conformations from the MD trajectory to the bound and unbound X-ray conformations, using an all-residue, all-atom fit. The results are summarized in Figure 2. In parts (a)–(c), we show the histograms of RMSD *versus* the unbound and bound states for three typical examples, 1BOY, where the MD is closer to the bound, 5PTI, where it is roughly the same distance from both, and 1HPT, where it is closer to the unbound. The RMSD between bound and unbound is shown as an arrow on the *x*-axis. Note that neither the bound nor unbound conformation is ever sampled directly by the MD (none of the histograms

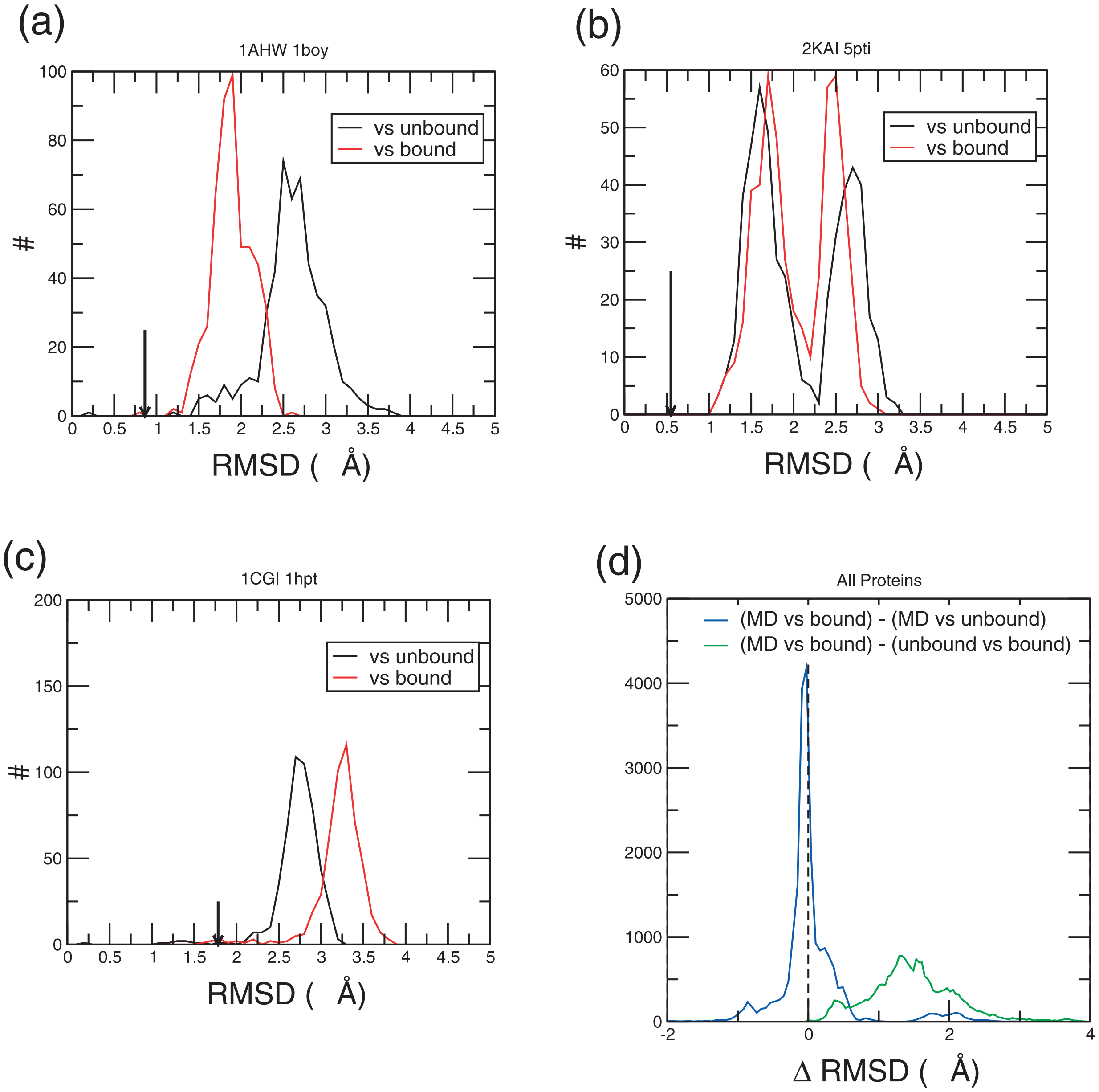


Figure 2. (a)–(c) Histograms of the RMSD between the MD and the unbound X-ray conformation, and between the MD and the bound X-ray conformation, for the three proteins 1BOY, 1HPT and 5PTI. The RMSD of bound *versus* unbound is shown by a downward-pointing arrow. (d) Histogram of the distance of all MD snapshots of all proteins from the respective bound and unbound X-ray structures after all-atom RMS fitting. Blue line: $\text{RMSD}_{(\text{MD } \textit{versus} \text{ bound})} - \text{RMSD}_{(\text{MD } \textit{versus} \text{ unbound})}$; green line: $\text{RMSD}_{(\text{MD } \textit{versus} \text{ bound})} - \text{RMSD}_{(\text{unbound } \textit{versus} \text{ bound})}$.

extends to zero). The bimodal shape of the histogram for 5PTI is indicative of a conformational transition that takes place during the simulation. Shapes similar to this, or even more complicated, are quite common. Corresponding data are shown for all 44 proteins in Supplementary Figure 1. Figure 2(d) shows two histograms for all proteins pooled. The green histogram shows the offset, $\text{RMSD}_{(\text{MD } \textit{versus} \text{ bound})} - \text{RMSD}_{(\text{bound } \textit{versus} \text{ unbound})}$, so it summarizes the difference between the black histogram and the vertical arrow in (a)–(c) for all proteins. As there are no points below zero, a complete MD conformation was never closer to the bound than the unbound to the bound. The blue histogram is $\text{RMSD}_{(\text{MD } \textit{versus} \text{ bound})} - \text{RMSD}_{(\text{MD } \textit{versus} \text{ unbound})}$, and summarizes the difference between the red and black histograms for all proteins. Interestingly, 58% of the blue histogram lies below zero, showing that the MD is slightly closer to the bound conformation than to the unbound. This is significantly different from a 50–50 division if we assume that the conformations generated by the simulations become effectively de-correlated over a timescale of 1 ns.

Although all the MD snapshots are more different from either X-ray structure than the X-ray structures are from each other, it might be that this general tendency to move away from both X-ray conformations disguises the movement at certain instants of certain small regions of the molecule closer towards the bound conformation; these are swamped in a global RMS fit by the larger regions that move away (at that instant). We investigate this hypothesis later. However, we first investigate larger-scale movements, as these are sometimes required for the function of the proteins, and proteins that undergo them are often the most difficult to dock.

Large-scale readjustments: essential dynamics

Information about large-scale conformational fluctuations of proteins can often be derived efficiently from relatively short (∼5 ns) molecular dynamics simulations using essential dynamics (ED), which is a principal component analysis (PCA) of the trajectory.[41–43] The covariance matrix of atomic displacements observed during the simulation is generated (using either all protein atoms or, for a large protein, some subset such as C^α atoms), and then diagonalized. The resulting eigenvectors, the principal components, may then be sorted according to their eigenvalues. It is often found that the molecular motions described by the first few eigenvectors, with large associated eigenvalues, are the most important in producing the fluctuations of the system, and tend to correspond to large-scale motions of the molecule (i.e. motions correlated over a long distance). True induced-fit motion, however, may not correspond to these eigenvectors.

In Table 2, the patterns of atomic displacement in the first 20 principal components (PCs) are compared with the conformational change from unbound to bound. This is done using the correlation and the overlap, defined in the methods section. Correlation and overlap are calculated for the C^α atoms only, so this measure of the conformational change concentrates on the backbone. The correlation is a measure that takes into account only the size of the atomic displacement, whereas the overlap (the modulus of the averaged dot product) considers its direction. The average (standard deviation) of the overlap for the first principal component is 0.13(0.11), and its correlation 0.27(0.18). For the second principal component, average overlap is again 0.13(0.15) and average correlation 0.30(0.22). For the best principal component among the first 20, the average overlap is 0.31(0.12) and the correlation is 0.50(0.17).

Only four proteins, ribonuclease inhibitor (2BNH), Novo chymotrypsin inhibitor 2 (2CI2), an antibody (1BVL) and pancreatic trypsin inhibitor (1HPT) have an overlap >0.5, and only one of these, 2BNH, has an overlap >0.75. The correlation is generally larger than the overlap, but the two measures tend to behave in the same way for a particular protein: again, 2BNH is one of only two that have a correlation >0.75. Taken all together, these results suggest that for the majority of proteins the eigenvectors from an all-residue PCA do not provide a good model for the conformational change on docking.

For a particular protein, any of the first 20 PCs may have the highest overlap, but if the overlap or correlation is high in absolute terms, the associated PC is likely to be one of the first few: all the PCs with an overlap >0.5 are found in the first five. The importance of PCs with low eigenvector indices is consistent with a previous MD study on domain movements in AAA+proteins,[44] though PCs with higher eigenvector indices turned out to be important in a recent similar study of loop movement on ligand binding to a protein kinase.[38] That study, however, may correspond to the protein–protein examples where no PC with a high overlap was found, because it took into account more PCs than the current work and they were derived in a different way (from normal modes rather than MD).

The special status of the ribonuclease inhibitor 2BNH amongst the proteins studied here was confirmed by an analysis with DynDom (see Methods)[45,46] on both the bound–unbound X-ray crystal structures and the structures obtained by projection along the PCs. This analysis was performed for all 22 proteins in the test set (data in Supplementary Table 2), and 2BNH turned out to be unique in having dynamic sub-domains in all of the first five PCA eigenvectors and in the unbound–bound transition, as shown in Figure 3. We show the domains in all five of the eigenvectors, since individual eigenvectors do not reliably converge on a 5 ns timescale,[41] consensus is valuable in establishing any apparent result. There is reasonable agreement between the X-ray sub-domains and those in the PCA eigenvectors, particularly those in eigenvectors 2 and 4 (Figure 3 (b2) and (b4)): at least two sub-domains are found, with a hinge around the eighth of the helix-strand repeats that make up the structure, enabling the inhibitor to wrap around the ribonuclease. There is also some evidence for a further domain in the C-terminal half (present in eigenvectors 2 and 5). In addition, there is good agreement in the unbound–bound transition between the hinge regions identified by DynDom and MolMovDb (a database of protein conformational changes)[47] (Figure 3 (a)). The extent to which the movement along the measured PC eigenvectors actually results in a reduction in the C^α RMSD of the unbound protein relative to the bound is shown in Figure 3(c).

As this protein is unique in our data set in showing substantial sub-domain motion, we have treated it differently by docking conformations derived from the PCA, as described later.

Small-scale readjustments

The above analysis has concentrated on medium to large scale movements of the protein backbone. As stated in the section on the global RMSD distribution, it is also of interest to investigate in detail how smaller regions of the proteins move; in particular short segments of backbone, or the

Table 2. Relationships between unbound-bound conformational change $\Delta\mathbf{r}$ and principal components (PC) $\mathbf{q}_j$ produced by essential dynamics analysis of MD

cplx	Protein	Overlap							Correlation							
		First	Second	Best	Index of best	Index of first:			First	Second	Best	Index of best	Index of first:			Collectivity (ub-b)
						> 0.25	> 0.5	> 0.75					> 0.25	> 0.5	> 0.75	
2PTC	2ptn	0.03	0.09	0.25	9	–	–	–	0.28	0.37	0.46	19	1	–	–	0.57
	5pti	0.29	0.39	0.40	10	1	–	–	0.56	0.88	0.88	2	1	1	2	0.05
1WEJ	1qbl	0.27	0.10	0.27	1	1	–	–	0.38	0.33	0.45	6	1	–	–	0.41
	1hrc	0.14	0.08	0.24	14	–	–	–	0.04	0.00	0.21	7	–	–	–	0.61
2SNI	2st1	0.32	0.02	0.32	1	1	–	–	0.37	0.32	0.51	14	1	14	–	0.45
	2ci2	0.28	0.50	0.50	2	1	2	–	0.45	0.63	0.63	2	1	2	–	0.46
2SIC	2st1	0.16	0.03	0.34	3	3	–	–	0.33	0.43	0.55	14	1	13	–	0.44
	3ssi	0.04	0.01	0.26	5	5	–	–	0.34	0.28	0.48	5	1	–	–	0.38
2VIR	1gig	0.10	0.06	0.24	3	–	–	–	0.36	0.14	0.57	18	1	18	–	0.23
	2hmg	0.39	0.26	0.39	1	1	–	–	0.48	0.33	0.48	1	1	–	–	0.37
2PCC	1cca	0.05	0.27	0.30	12	2	–	–	0.27	0.55	0.68	12	1	2	–	0.30
	1ycc	0.03	0.16	0.18	6	–	–	–	0.09	0.14	0.47	6	3	–	–	0.35
1BRC	1bra	0.14	0.13	0.23	6	–	–	–	0.31	0.29	0.71	6	1	6	–	0.56
	1aap	0.28	0.08	0.37	4	1	–	–	0.25	0.09	0.54	4	1	4	–	0.56
1BGS	1a2p	0.01	0.19	0.42	13	3	–	–	0.31	0.40	0.56	13	1	3	–	0.57
	1a19	0.25	0.27	0.27	2	1	–	–	0.12	0.17	0.43	20	4	–	–	0.65
1UGH	1akz	0.06	0.02	0.31	10	10	–	–	0.16	0.11	0.51	14	5	14	–	0.52
	1ugi	0.23	0.07	0.29	8	8	–	–	0.02	–0.03	0.30	10	10	–	–	0.46
2KAI	2pka	0.14	0.09	0.19	9	–	–	–	0.41	0.25	0.50	11	1	–	–	0.14
	5pti	0.05	0.12	0.24	10	–	–	–	–0.14	–0.07	0.37	11	4	–	–	0.35
1AHW	1fgn	0.10	0.02	0.20	12	–	–	–	–0.06	0.12	0.26	17	17	–	–	0.62
	1boy	0.05	0.06	0.17	6	–	–	–	–0.12	–0.03	0.02	13	–	–	–	0.45
1AVZ	1avv	0.11	0.20	0.25	3	–	–	–	0.17	0.45	0.53	8	2	4	–	0.44
	1shf	0.03	0.00	0.30	4	4	–	–	0.12	0.29	0.37	15	2	–	–	0.65
1DQJ	1dqq	0.30	0.10	0.30	1	1	–	–	0.03	0.04	0.04	2	–	–	–	0.27
	3lzt	0.16	0.03	0.45	3	3	–	–	0.30	0.48	0.58	3	1	3	–	0.37
1FSS	1vxr	0.00	0.02	0.22	5	–	–	–	0.34	0.41	0.46	7	1	–	–	0.42
	1fsc	0.19	0.33	0.35	7	2	–	–	0.37	0.54	0.61	3	1	2	–	0.61
1WQ1	1wer	0.20	0.03	0.35	10	8	–	–	0.44	0.23	0.64	10	1	3	–	0.43
	5p21	0.26	0.02	0.26	1	1	–	–	0.55	0.03	0.56	18	1	1	–	0.34
1MLC	1mlb	0.04	0.12	0.27	9	9	–	–	0.37	0.30	0.51	7	1	7	–	0.42
	3lzt	0.06	0.31	0.31	2	2	–	–	0.38	0.48	0.68	4	1	4	–	0.38
1DFJ	2bnh	0.40	0.77	0.77	2	1	2	2	0.63	0.79	0.79	2	1	1	2	0.51
	7rsa	0.13	0.01	0.33	6	6	–	–	0.41	0.36	0.68	10	1	10	–	0.47
1BVK	1bvl	0.14	0.06	0.54	4	4	4	–	0.22	0.48	0.57	4	2	4	–	0.50
	3lzt	0.15	0.24	0.43	5	5	–	–	0.18	0.69	0.69	2	2	2	–	0.13
1CGI	2cga	0.01	0.12	0.29	3	3	–	–	0.12	0.32	0.40	7	2	–	–	0.11
	1hpt	0.07	0.23	0.51	5	5	5	–	0.52	0.49	0.55	20	1	1	–	0.23
1KKL	1jb1	0.07	0.06	0.18	16	–	–	–	0.34	0.49	0.63	9	1	4	–	0.31

	1sph	0.03	0.02	0.19	7	–	–	–	0.08	0.10	0.25	19	19	–	–	0.69
1FQ1	1hcl	0.02	0.05	0.25	8	–	–	–	0.20	0.04	0.36	13	3	–	–	0.19
	1fpz	0.05	0.01	0.39	3	3	–	–	0.54	0.23	0.60	3	1	1	–	0.40
1FIN	1hcl	0.04	0.06	0.20	8	–	–	–	0.19	0.03	0.31	13	13	–	–	0.15
	1vin	0.02	0.05	0.17	3	–	–	–	0.37	0.23	0.38	12	1	–	–	0.44

The vectors are evaluated for all C$^\alpha$ atoms. The PCs are ordered by decreasing size of the associated eigenvector. In the first half of the Table, the overlaps of the first and second PC are shown, then the highest overlap out of the 20 PCs considered ("best") and the index of that PC. Then, for the three values 0.25, 0.5 and 0.75 of the overlap, we show the index of the first PC having an overlap of such a size. The second half of the Table is similar but for the correlation between $\mathbf{q}_j$ and $\Delta\mathbf{r}$, rather than the overlap.

changes to particular side-chain rotamers. The "global" RMSD analysis of the entire molecules conducted above is liable to miss fluctuations in short segments of the protein chain which take them closer to the bound conformation, because it involves fitting of protein conformers on all (or all C$^\alpha$) atoms, which may well not produce alignment in the segment in question.

Small-scale readjustments: RMSD in moving windows show local movements towards the bound state

We have looked for such local fluctuations of the backbone by fitting the MD trajectory to the bound X-ray structure over a moving window along the polypeptide chain and calculating the RMSD of the same window as a function of time. The width of the window was taken as four residues throughout. The RMSD calculations were performed for all heavy atoms and also for C$^\alpha$ atoms alone.

The results are shown in Figure 4. In Figure 4(a) the simulations are analyzed separately and the fraction of segments shown for which the MD approaches closer to the bound state. In this case we mean "closer to the bound state than the unbound conformation is", i.e. $\mathrm{RMSD}_{(\mathrm{MD}\ \mathit{versus}\ \mathrm{bound})} < \mathrm{RMSD}_{(\mathrm{unbound}\ \mathit{versus}\ \mathrm{bound})}$. Figure 4(a) shows: firstly, the fraction of segments that fluctuate closer to the bound even for a single instant (these events of closest approach generally last only a few tens of picoseconds and will in general occur at different times for different segments); secondly, the fraction of segments that are closer to the bound in the best single trajectory snapshot (i.e. the snapshot that contains the most such segments); and thirdly, the fraction of segments that are closer to the bound on average during the MD. It is found that 52(±17)% of the surface segments fluctuate at some time closer to the bound state (the range being the standard deviation over all 41 proteins). The fraction of surface segments closer to the bound in the best single structure is 17(±10)%, and the fraction closer to the bound on average is 5(±5)%. Figure 4(b) relates to the times of closest approach; the main Figure is a histogram of the time-slices for which the segments make their closest approach to the bound state, and the inset is a histogram of the time-slices at which best single snapshots occur. It is apparent that all these closest approaches tend to be early on in the simulations.

We have also carried out the analysis of Figure 4(a) specifically for the core and periphery of the interface separately (data for individual proteins not shown). The fractions obtained are very similar in the periphery, while in the core 52(±30)% of segments are ever closer to the bound, 30 (±27)% closer in the best single and 3(±6)% closer on average. The difference between closest MD and the unbound is generally small, however, and the glycoslyase inhibitor 1UGI is in fact the only protein in the test set where the RMSD of a reasonably large section of the interface becomes appreciably lower.

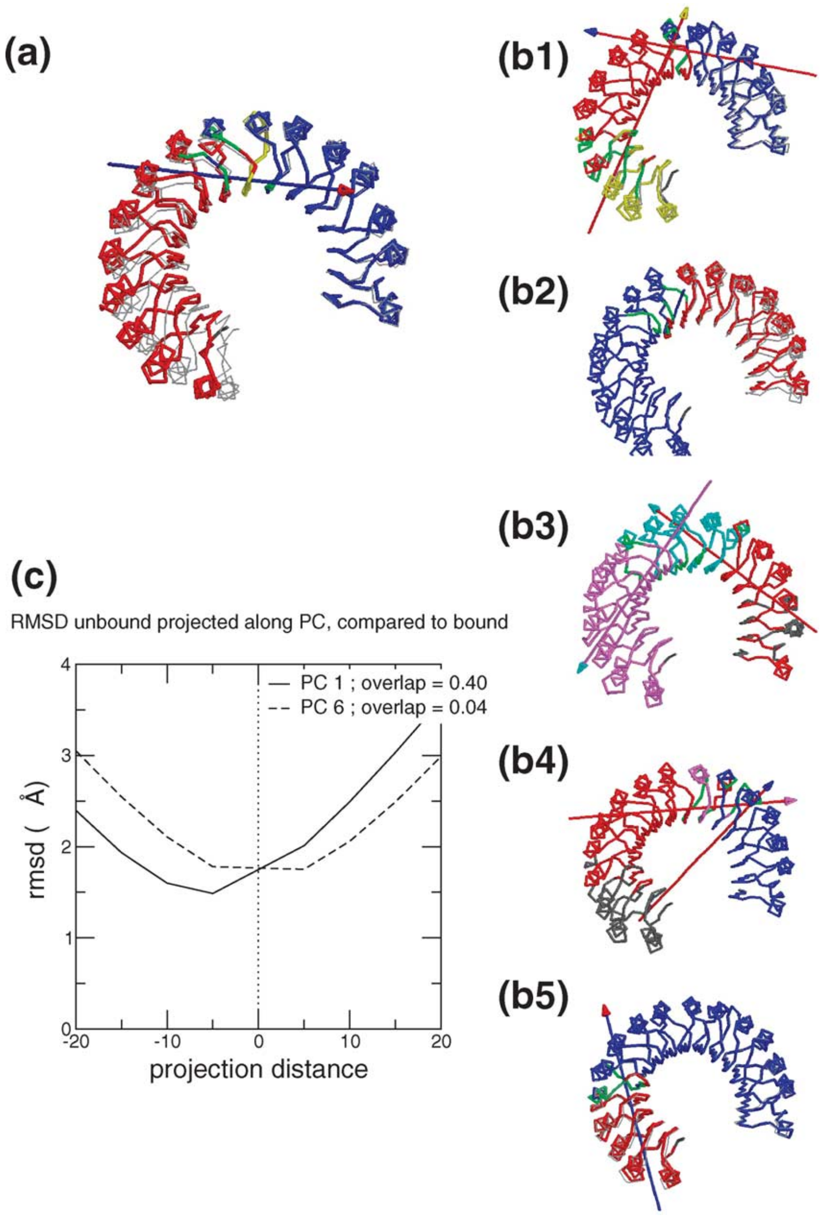


Figure 3. Domain movement and RMSD of ribonuclease inhibitor. (a) Sub-domains identified by DynDom[45,46] from analysis on bound (1DFJ), light grey trace, and unbound (2BNH) conformations: two sub-domains, blue and red, are highlighted by a screw axis that lies approximately midway and within the plane of the molecule. The hinge residues between domains are coloured in green. As a control, the main hinge identified in the database MolMovDb,[47] is indicated to be in a similar region (yellow residues). For (a), and for (b1)–(b5) described below, arrows indicate direction of relative sub-domain motion: each domain pair is coloured the same as the shaft and head of the arrow relating their motion. (b1)–(b5) Sub-domains identified by DynDom from conformations along principal component eigenvectors 1–5: domains blue, red, magenta, cyan and yellow, hinge residues green. (b1) and (b3)–(b5) show the predominant twisting motions between sub-domains within the plane of the molecule while (b2) shows a twisting motion along an axis perpendicular to this plane. There is approximate agreement as to where the domains lie. (c) Effect of projecting the unbound conformation along a PC eigenvector with high (the first, overlap 0.4) or low (the sixth, overlap 0.04, i.e. not significantly different from zero) overlap with the unbound–bound transitions, and comparing the C$^\alpha$ RMSD of the resulting structure with the bound.

The details of this are shown in Figure 4(c) and (d), showing interface segments around residues 15–20 that fluctuate appreciably closer to the bound conformation than the unbound, even on average, during the MD.

Small-scale readjustments: bound side-chain rotamer occupancies higher within the core regions of interface

Here we investigate whether the MD simulation of the unbound structure produces conformations in which the rotamers occupied by the side-chains in the bound structure are visited. It is natural to divide the side-chains into those where the rotamers change between bound and unbound, and those where they do not; and also those where the MD rotamer (starting from the unbound) makes a correct prediction of the bound and those where it does not. This is described in detail in the caption to Figure 5; an example of the cases that need to be considered when the bound and unbound rotamers are different is shown in Figure 5(a). To define a single rotamer for an entire side-chain, we use each unique combination of the χ_1 and χ_2 rotamers, where they are put into bins of width 120° or 180°, as appropriate for the amino acid residue.

The results are shown in Figure 5(b) for all proteins, pooled, normalized, and analyzed in the five groups, surface, interface, surface-but-not-interface, interface core and interface periphery, and three different criteria for true positive/true negative rotamers: that the most frequently visited rotamer by MD must match the bound, that the second most frequently visited must match (so the bound state would be the most frequently visited sub-state), or that any may match. Hatched bars indicate that the rotamer does not change on

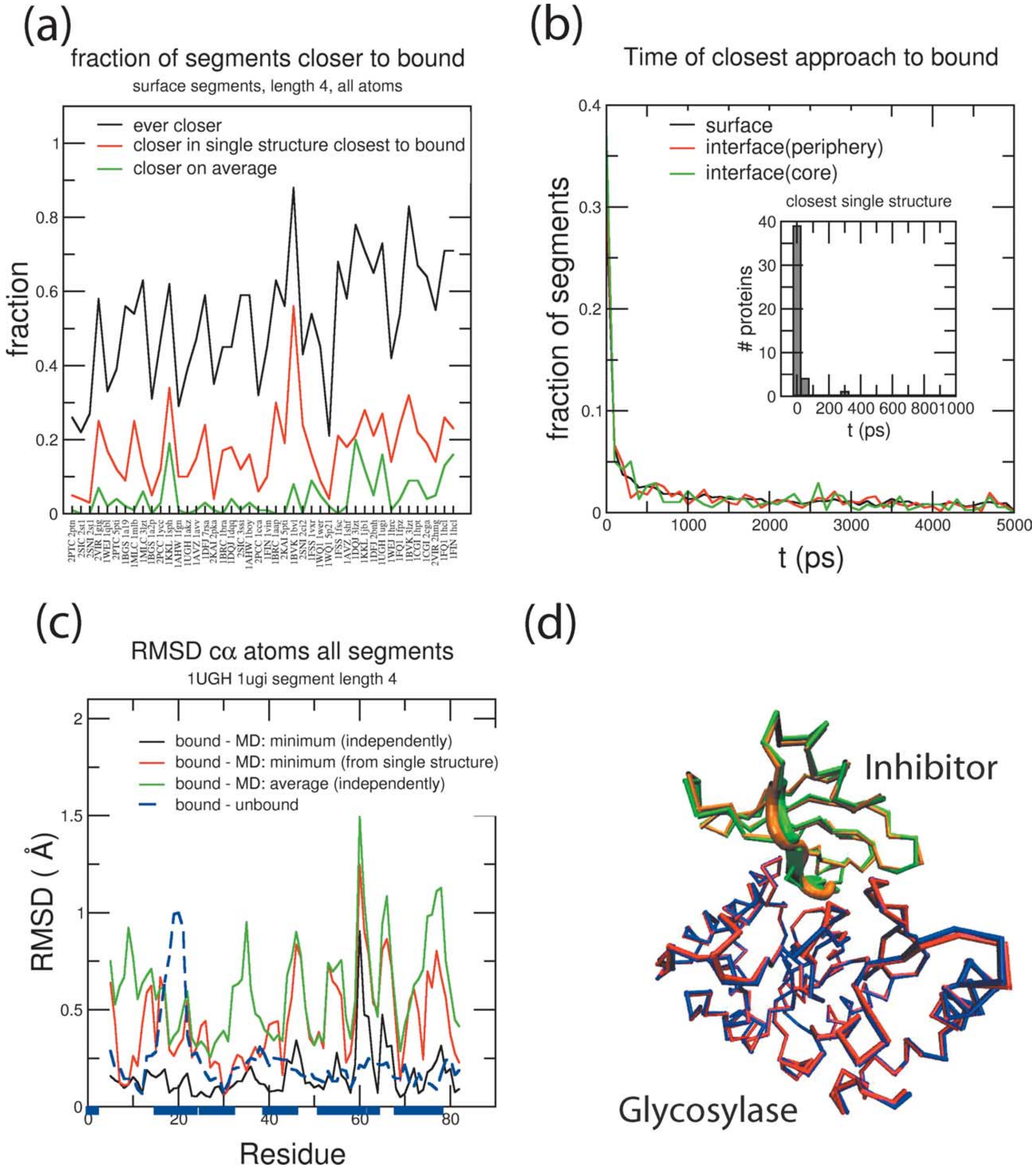


Figure 4. (a) Fraction of segments that are: closer to bound than unbound at any instant (so corresponding not to a single MD snapshot in each protein but to a "mosaic" of segments from different snapshots); in the best single structure; or on average throughout the simulation. (b) Distribution of times corresponding to snapshots of closest approach: (main Figure) treating surface segments independently; (insert) for the single structure containing the most such segments. (c) Bound–unbound RMSD (dotted blue line) for glycosylase inhibitor 1UGI, and comparison with MD fluctuations: showing the closest approach at any instant of the simulation (again corresponding to many MD snapshots); the closest single conformation; and the average RMSD over the whole MD trajectory. (d) Glycosylase (bound, blue; unbound, red) and glycosylase inhibitor (bound, green; unbound, orange). The interface segment around residues 16–21 is shown in protein cartoon representation.

docking (so the unbound side-chain is already in the correct rotamer for docking); blue indicates that the MD has the correct rotamer for docking.

First, we remark that about 1/3 of the residues change rotameric state between the bound and unbound conformations, and this fraction is very nearly independent of whether the residue is in the interface or not, though in the core of the interface it is slightly less, 1/4 (this is essentially consistent with previous work).[48] Using the most-visited rotamer as the criterion, about half of the side-chains are predicted to change rotamer by MD, and also about half are correctly predicted. In general, then, the unbound X-ray conformation provides a better model for the bound rotamers than the MD, but there are many side-chains where the bound rotamer is predicted by MD and not by the unbound. Predictions of the bound rotamer are more likely to be right for side-chains that do not change rotamer than those that do. This, again, is independent of whether the side-chains will lie in the interface or not, but the interface core is predicted better than the rest of the surface, and the interface periphery a little less well. In all regions, the fraction of true positives is lower than the fraction of false positives of type (a), which means that the unbound X-ray has a slightly greater fraction of its side-chains in the correct rotamer for docking than is predicted from the most likely MD rotamer. The region where the MD performs best, however, is the interface core. The fraction correctly predicted falls if the second most visited rotamer in MD is used to compare with bound and unbound. This shows that there are more surface residues for which the bound rotamer is the most probable rotamer in solution, than there are residues for which it is the second most probable. This suggests that if the bound state is present as a

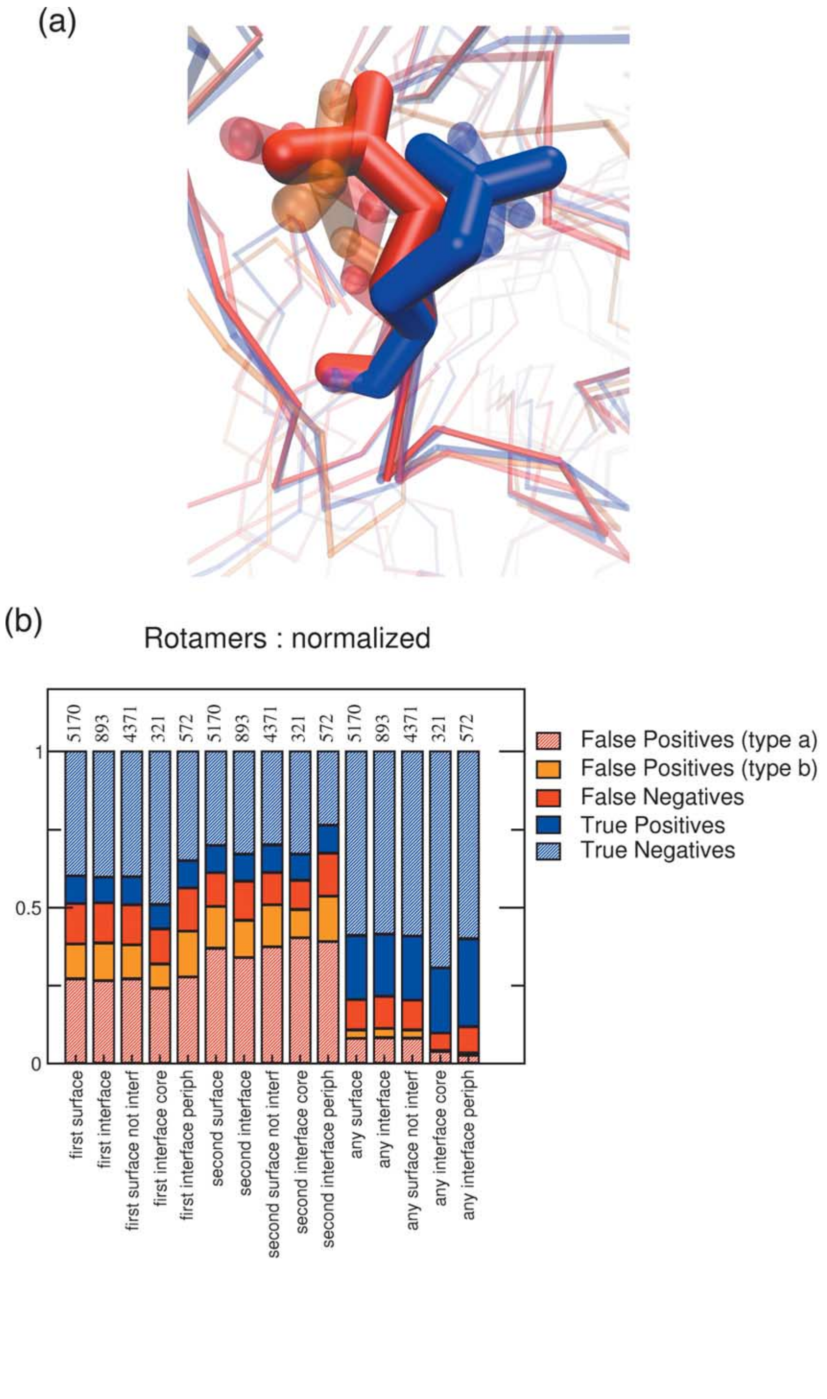


Figure 5. Side-chain rotamer occupancies, MD (rotamer χ_{md}) compared with bound (χ_b) and unbound (χ_{ub}). The rotamer for a particular side-chain is defined as a unique combination of the χ_1 and χ_2 rotamers (χ_3 and higher rotamers, where the side-chain has them, are ignored). False positive (type a) (hatched red): $\chi_b == \chi_{ub}$ and $\chi_b != \chi_{md}$. False positive (type b) (orange): $\chi_b != \chi_{ub}$ and $\chi_b != \chi_{md}$ and $\chi_{ub} != \chi_{md}$. False negative (red): $\chi_b != \chi_{ub}$ and $\chi_{ub} == \chi_{md}$. True positive (blue): $\chi_b != \chi_{ub}$ and $\chi_b == \chi_{md}$. True negative (hatched blue): $\chi_b == \chi_{ub}$ and $\chi_b == \chi_{md}$. Hatched bars indicate that the rotamer does not change on docking (between b and ub); blue (hatched and solid) indicates a correct prediction from the MD. (a) Conformations of a typical side-chain, showing bound X-ray (transparent blue), unbound X-ray (transparent red), and conformations from MD colored as described above, i.e. corresponding to true positives (solid blue), false negatives (solid red), and false positives Type (b) (solid orange). (b) Culmulative results for all proteins, broken down into the five classes all surface residues, all interface, all in surface but not in interface, interface core and interface periphery. The results are normalized by the total number of rotamers in each class (this total number being shown above the bar). "first" means that χ_{md} is the most populated MD rotamer, as in A, "second" means that it is the second most populated, whereas "all" means that any of the rotamers visited in the course of the MD, provided they are occupied for ≥5% of the simulation time, can be used in defining a true positive or true negative.

"conformational sub-state", it does not show a strong tendency to be the second most occupied. Finally, to find if the bound conformer is present as a sub-state at all, we also consider the number of residues whose bound state is amongst any of those rotamers that are visited for a substantial fraction (≥5%) of the simulation time. About 80% of residues satisfy this criterion. Why are not all bound rotamers visited? First, the simulation time may still be too short for side-chains that are tightly packed on the surface of each protein, and so need to move in concert to reorganize, to reach the rotamers observed in the bound state. Second, it may simply be an indication of inaccuracies in the force field. Third, some of the bound state rotamers in regions away from the functional protein–protein interface may themselves be the result of crystal contacts between complexes in different asymmetric units. Finally, all bound rotamer states may only be observed in the presence of the binding partner: the induced fit model.

The complete results analyzed separately for each protein for the most frequently visited rotamer are in Supplementary Table 3.

Analysis of interface core/periphery and other residues

We now analyze several residue-level properties pooled for all 41 of the proteins simulated, breaking them down according to the various regions of the proteins as defined previously. Of particular interest is the interface, and its subdivisions into core and interface, and the rest of the surface. The results are shown in Figure 6.

First the conservation was assessed by running five cycles of PSI-BLAST on each protein and mapping the conservation of each residue. If the conservation is averaged over the regions of the surface, the degrees of conservation are significantly different: the periphery of the interface is less conserved even than the rest of the surface, and the interface core is more so, and residues here may even be as conserved as the buried residues (Figure 6(a)). Because these differences are in different directions, there is no significant difference in the conservation of the interface as a whole and the rest of the surface.

A simple way to quantify local mobility in the same way is to evaluate the root-mean-square fluctuation (RMSF) of each atom, and average over residues. The result is shown in Figure 6(b). There is greater local mobility in the periphery of the interface than in the rest of the surface, and less in the core of the interface, whether all atoms are considered or just the C^{α} atoms of the backbone, the difference being significant as shown by the lack of overlap of the (2× standard error of the mean) error bars. If these averages are evaluated for each protein separately, the mobility of the core is lower than that of the periphery of the interface for 35/41 proteins (though the difference is not usually significant). A specific example of this is shown in Figure 6(c) and (d), where residues in the periphery and core of the interface of protein 1BOY are shown, coloured according to C^{α} RMSF. The exceptions are 1A19, 1AAP, 5PTI, 1CCA, 3SSI and 2CI2. The majority of these are inhibitors. It is also noteworthy that 5/6 of them have less than the average number of residues in the interface core.

Another measure of local mobility comes from the occupancies of the rotamers of a particular side-chain, as investigated in the previous section. It is possible to define an entropy:[49–51]

$$S = -k_{\mathrm{B}} \sum_i p_i \ln p_i$$

and an associated effective number of rotamers for the side chain:[52]

$$n_{\mathrm{eff}} = \exp(S/k_{\mathrm{B}})$$

The results for n_{eff} for the proteins under study are shown, categorized by residue in Figure 6(e) and averaged over all types in Figure 6(f). It is apparent that, most side-chain types (particularly the longer and more mobile acidic, amide, and basic side-chains) are slightly but significantly less mobile in the patches that will form the core of the interface, and slightly more so in the periphery. Once again, the significance of the result disappears if the interface is considered as a whole. This result is in agreement with the RMSF analysis and also agrees with previous work,[53] where it was found that waters in protein–protein interfaces are less mobile; it might be expected that the protein side-chains near to them will also tend to be so. In comparison with the previous work in which the entropy was calculated from the observed statistics of side-chain rotamers in the PDB,[49–51] the numerical values are slightly smaller (one estimate[51] was that the average value of $T\Delta S$ on binding was 1 kcal mol^{-1}, corresponding to $n_{\mathrm{eff}}=5.5$, compared with 4.5 obtained in the current work; it was also estimated that the largest value of $T\Delta S$ was 2.1 kcal mol^{-1} for Gln, corresponding to $n_{\mathrm{eff}}=35$, much larger than $n_{\mathrm{eff}}\approx 7$ for Lys and Met found here). These results suggest that the entropic effects on binding from this source might be slightly lower than previously estimated. The small influences of environment seem not to have been observed before, possibly because a difference between 4 and 4.5 in n_{eff} corresponds to a very small change in $T\Delta S$ of only 0.07 kcal mol^{-1}. Considering n_{eff} for each protein separately (data not shown), the mobility of the core remains lower than that of the periphery (though not significantly so) for 35/41 proteins. The exceptions this time are 1BOY, 1AAP, 1MLB, 1UGI 1QBL and 2PKA, of which only 1AAP was identified before in the RMSF analysis. The consensus result of RMSF and entropy, then, is that 30/41 of the proteins individually have a higher mobility in the periphery of the interface than the core. However, if only one of the two measures is required, then up to 40 of the 41 proteins show this tendency.

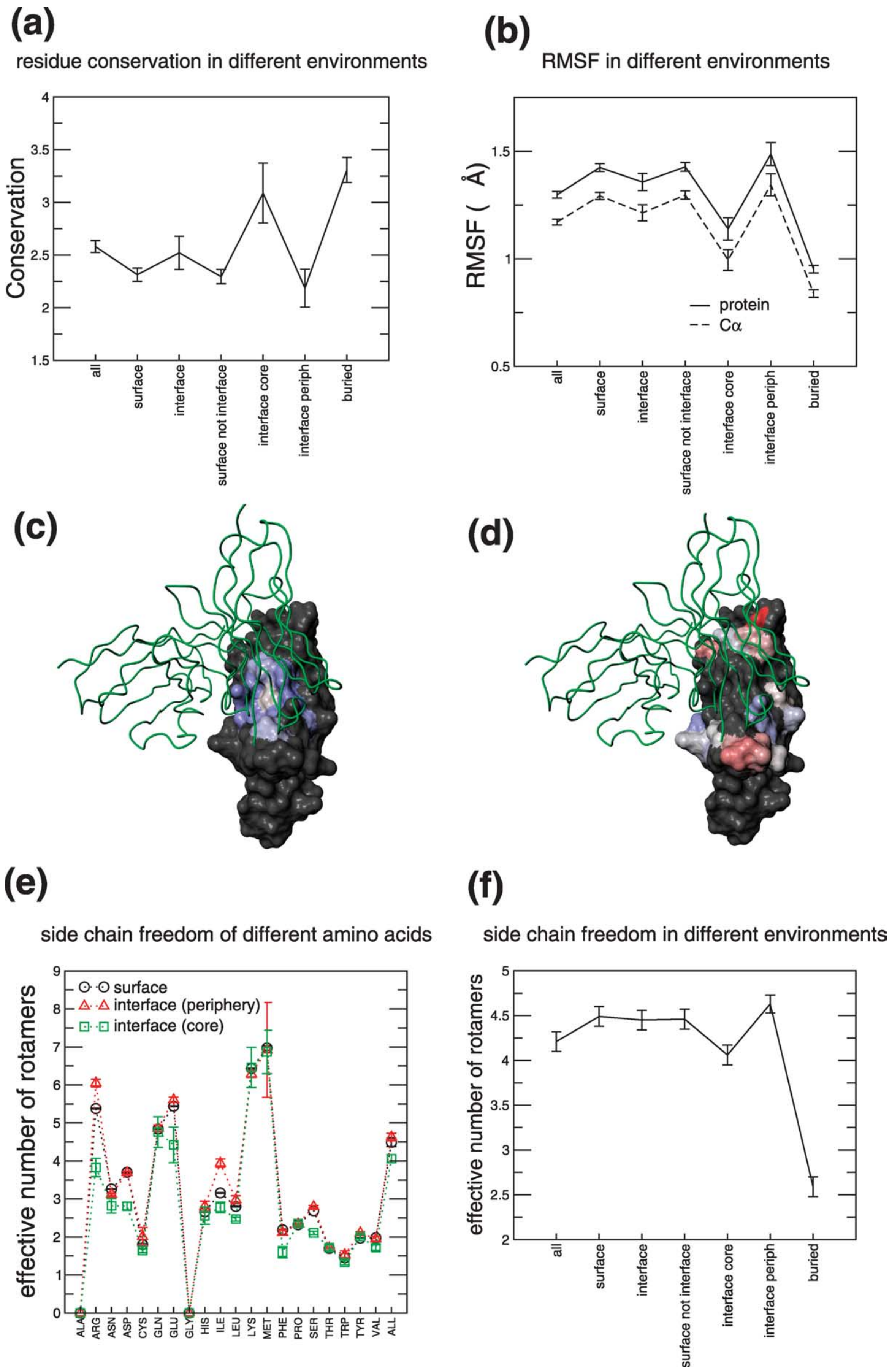


Figure 6 (*legend opposite*)

Cluster analysis of the trajectories

We have used cluster analysis of the trajectory to select conformations that are representative of the MD trajectory as a whole for use in docking (see next section); in most cases only one or two clusters dominate the entire trajectory; so we select the median conformation from the two clusters of longest lifetime (or from a single cluster, if no other has a lifetime longer than 0.5 ns). By "median" we mean the conformation that has lowest total RMSD to all the others in the cluster. Two examples are shown in Supplementary Figure 2(a). The distribution of cluster lifetimes is shown in the first part of Supplementary Figure 2(b), showing that the majority of the clusters are short lived, but there are a few that last for a large proportion of the simulation. The second part of Supplementary Figure 2(b) shows that the two longest-lived clusters together always occupy at least half the trajectory, and frequently more (on average 3.6 ($\pm$0.9) ns, or 72% of the total simulation time); see also Supplementary Table 4. The contingency analysis on the first two side-chain rotamers has also been repeated for the cluster-center conformations (Supplementary Figure 2(c)), with similar results to those for the most-populated rotamer of the entire trajectory.

The MD cluster centers (like the trajectory from which they come) are more distant from the bound than is the unbound in terms of the RMSD of both the interface and the whole molecule (data in Supplementary Table 4). Perhaps the relatively short simulation time, 5 ns, and/or the classical energy functions used in the MD simulations, may be the root cause for this discrepancy. Nevertheless, by sometimes including side-chain or backbone conformations that are closer to the bound, and which may in some cases be of particular importance in establishing surface complementarity, docking may overall be improved, especially in cases where the use of the unbound conformations alone does not lead to a satisfactory result.

Part II: is rigid-body docking improved by information gleaned from the MD trajectories?

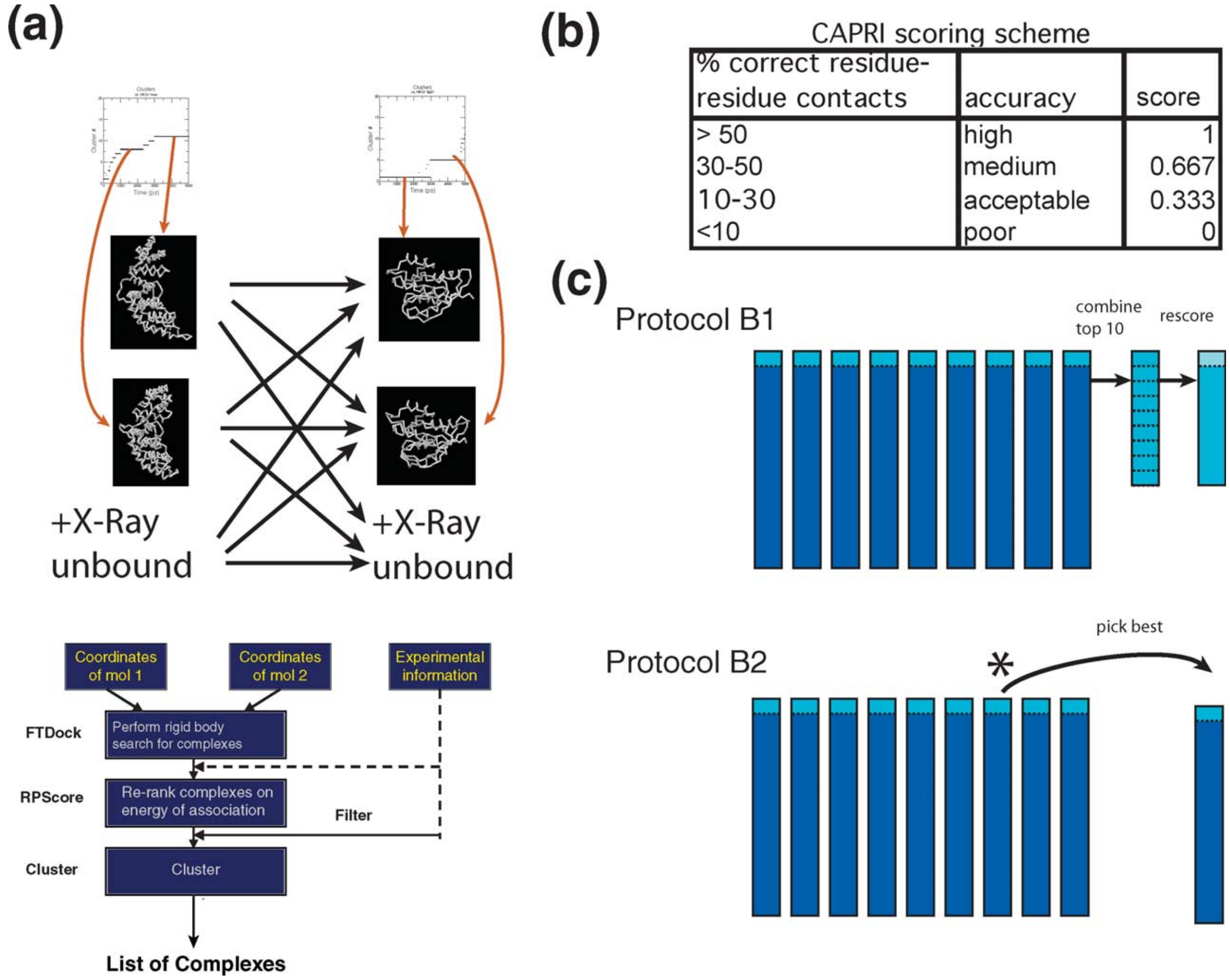
We investigate the effect of use of MD-derived conformations rather than unbound X-ray structures with the protein–protein docking program 3D-Dock.[8,30–33] To summarize, the program exhaustively searches the space of relative translations and rotations of receptor and ligand, keeping several thousand candidate complexes based on high surface complementarity and a negative electrostatic energy. These are then re-assessed using a scoring function called RPScore, derived from the statistics of residue–residue contacts in the PDB, and clustered. To improve the likelihood that the "top ranked" clusters by this method will be close to the native, a biological filter is also applied, which reflects a few residues that are thought to lie in the interface by virtue of mutagenesis studies, a knowledge of the active site, etc. These can easily be found in most cases from an examination of the literature, or by an automatic method based on spatial clusters of evolutionarily conserved or divergent residues on the surface.[53,54] The filters used in this case were those already extracted from the literature, and are listed in Supplementary Table 5.

To assess the results, we primarily use a criterion similar to that used in the first rounds of the CAPRI docking assessment exercise:[55] the nearest-native among the top ten predicted complexes in the ranked list is taken without regard for its rank within the top ten. It is then given a score between 0 and 1 based on the fraction of correct residue–residue contacts that it contains, according to the table in Figure 7(b). A residue–residue contact is taken to occur when any atom in the one residue is closer than a cutoff distance of 5 Å to any atom in the other, and the fraction of correct contacts is calculated relative to the total number of contacts in the native (not the predicted) complex.

Surface softening does not improve docking

The global RMSF in the MD was used to define a local softening of the surface layer used in 3D-Dock on the X-ray unbound components. However, as shown in Table 3 part E, the results are much worse than the default protocol (Table 3 part A). This can be attributed to the loss of detailed complementarity in those parts of the interface that are already a good fit in the unbound components, which is not

Figure 6. Local properties of the surface, pooled for all 41 proteins in the test set. Error bars on graphs are 2×SEM (standard error of the mean) throughout and the lines connecting data points are only guides to the eye. (a) The residue conservation (pooled in this case for the 40 proteins in the test set for which homologues were detected; none was found for the glycosylase inhibitor 1UGI). (b) Root mean square fluctuation of all heavy atoms (continuous line) or C^α atoms (broken line). Data corresponding to residues in different regions of the proteins are analyzed following fitting of trajectory snapshots on all residues of each protein (rather than just the residues in the region). (c, d). RMS fluctuations of C^α atoms of the residues in the core (c) or periphery (d) of the tissue factor, PDB code 1BOY (shown in molecular surface representation). Its binding partner (Antibody 5G9, PDB code 1FGN) is also shown as a backbone cartoon though it was absent in the simulation. The two parts of the interface of 1BOY are colored on a blue–white–red scale, increasingly deeply red or blue according to the extent to which the residues' C^α fluctuation is higher or lower than the average (corresponding to white) for all surface residues of this component. (e) Effective number of rotamers of the side-chain of each amino acid type (considering up to the first three χ angles). Also a weighted average, normalized with respect to composition; black: all side-chains of the surface; red: periphery of interface; green: core of interface. (f) Weighted average for all rotamers in all the environments considered.



CAPRI scoring scheme

% correct residue-residue contacts	accuracy	score
> 50	high	1
30-50	medium	0.667
10-30	acceptable	0.333
<10	poor	0

Figure 7. (a) The cluster center docking protocol. (b) The table shows the score given to a predicted protein–protein complex, depending on the percentage of correct residue–residue contacts in the interface, based on the scheme used for assessing the CAPRI blind trial. (c) The combination and comparison of the results of multiple runs. In the first method (compatible with a blind trial), the top-ranked complexes from the output list of each run are combined and re-ranked. In the second (incompatible with a blind trial), the run containing the complex closest to native is picked.

fully compensated by improved complementarity in other parts of the surface.

Docking of MD conformations can improve docking

A second approach is to dock individual conformations extracted from the 5 ns MD trajectory. 3D-Dock takes about one day to dock a pair of proteins, making it prohibitive to dock more than a few conformations for each complex, so we use the cluster center conformations described above, of which there are only one or two for each protein. For each complex, we then dock these cluster center conformations against each other, and also against the unbound X-ray components, including the two unbounds against each other. Thus there are a maximum of 9 $(=(2+1)\times(2+1))$ dockings for each pair of interacting proteins, though sometimes only six or even four. The protocol is shown in Figure 7 and the results are in Table 3 and Figure 8.

The results of docking the MD cluster centers, are shown in Table 3 part B. Two ways of assessing are shown: first (B1), all the cluster-center dockings are re-ranked by their RPScore (as would be the case in a blind trial), and the top ten overall are used to assess the docking. Secondly (B2), each of the cluster-center runs are re-ranked separately, the top ten of each taken, and then that run chosen which has the highest-scoring complex in its top ten. We call this the "best" run, and clearly it is not compatible with a blind trial; it is used here to assess the quality of the RPScore in identifying good solutions. B1 and B2 are indicated schematically in Figure 7(c).

To compare with this, two control runs have been performed using only the unbound components. The first, protocol A ("Default") (Table 3A) is the result of doing a single 3D-Dock run, while we defer discussion of the second to the next subsection.

The results for protocols A and B are summarized (as histograms of the number of complexes achieving each CAPRI-score) in the first column of Figure 8. The remainder of this Figure compares the different protocols. The most important comparison

Table 3. Results of docking MD cluster center conformations with 3D-Dock, compared with docking unbound conformations, either once or repeatedly, at random starting orientations.

Protocol A: a single 3D-dock run on unbound starting structures, with default parameters

	Default: single 3D-Dock run ub-ub				
Complex	CAPRI score	Correct contacts	Position of best	No. in top 10	$n<5$ Å in top 10,000
2PTC	0	0/68	–	0	4
1WEJ	0	0/43	–	0	0
2SNI	0.333	8/71	1	3	0
2SIC	0.333	8/71	3	1	3
2VIR	0	0/51	–	0	0
2PCC	0.333	4/29	3	2	0
1BRC	1	47/60	1	8	6
1BGS	0.667	20/56	5	4	11
1UGH	0	0/82	–	0	1
2KAI	0.333	13/64	1	3	3
1AHW	1	51/73	1	3	11
1AVZ	0.333	4/33	3	5	1
1DQJ	0	0/72	–	0	0
1FSS	0	0/65	–	0	3
1WQ1	0	0/91	–	0	0
1MLC	0	0/56	–	0	2
1DFJ	0	0/73	–	0	0
1BVK	0	0/53	–	0	0
1CGI	0.667	26/85	1	3	3
1KKL	0.333	3/28	8	1	0
1FQ1	0	0/61	–	0	1
1FIN	0	0/86	–	0	0
	av. score	TP rate	av. posn of best	av. #in top 10	av.
	0.24	0.13	2.7	1.5	2.2

Protocol B: 4–9 runs were done for each complex, cross-docking unbound and MD-cluster center starting structures. In B1, a global reranking was done; in B2 the best of the 4–9 runs was selected (best in the sense that it contains closest-to-native result in the top ten)

	MD cluster centers + ub-ub B1 re-rank all					B2 best single			
Complex	CAPRI score	Correct contacts	Position of best	No. in top 10	$n<5$ Å in top 10,000	CAPRI score	Correct contacts	Position of best	No. in top 10
2PTC	0.333	7/68	7	1	1	0.333	7/68	2	2
1WEJ	0	0/43	–	0	9	0.667	13/43	5	1
2SNI	0.333	12/71	2	3	0	0.333	9/71	2	4
2SIC	0.333	9/71	8	1	1	0.667	27/71	9	2
2VIR	0	0/51	–	0	0	0	0/51	–	0
2PCC	0.667	14/29	4	4	0	0.667	14/29	1	1
1BRC	1	47/60	9	8	29	1	32/60	9	6
1BGS	1	40/56	2	7	27	1	40/56	1	6
1UGH	0	0/82	–	0	0	0.333	11/82	6	1
2KAI	0.333	13/64	5	1	6	0.667	20/64	2	5
1AHW	1	38/73	4	1	31	1	38/73	1	2
1AVZ	0.333	8/33	10	1	0	0.667	10/33	7	6
1DQJ	0	0/72	–	0	2	0.333	8/72	6	1
1FSS	0	0/65	–	0	2	0	0/65	–	0
1WQ1	0	0/91	–	0	1	0	0/91	–	0
1MLC	0	0/56	–	0	3	0.333	8/56	7	1
1DFJ	0.333	19/73	10	1	0	0.667	31/73	9	1
1BVK	0.333	7/53	9	1	2	0.333	8/53	3	1
1CGI	0.333	10/85	9	1	11	0.667	26/85	9	3
1KKL	0.333	3/28	4	2	0	0.333	3/28	5	3
1FQ1	0	0/61	–	0	2	0.333	8/61	2	1
1FIN	0	0/86	–	0	0	0	0/86	–	0
	av. score	TP rate	av. posn of best	av. #in top 10	av.	av. score	TP rate	av. posn of best	av. #in top 10
	0.3	0.17	6.4	1.5	5.8	0.47	0.23	4.8	2.1

for the cluster centers is that between protocols B1 and A, i.e. to rerank the cluster centers + unbound–unbound globally, and compare with the default protocol, as if in a blind trial. By this measure, better results are obtained for five complexes, worse results for one, and the remainder are approximately the same. The probability of this outcome is around 0.1; this is suggestive that an improvement is achieved but is not significantly different from random. In detail, better results are obtained for

Table 3 (*continued*)

Protocol C: the unbound starting structures are docked repeatedly at different random starting orientations, the same number of runs being done as for B. C1: all runs re-ranked; C2: the best of the 4–9 runs selected. Comparison between the different methods is shown in Figure 8

	Multiple randomly spun 3D-Dock runs + ub–ub C1 re-rank all					C2 best single			
Complex	CAPRI score	Correct contacts	Position of best	No. in top 10	$n<5$ Å in top 10,000	CAPRI score	Correct contacts	Position of best	No. in top 10
2PTC	0	0/68	–	0	6	0.333	7/68	3	1
1WEJ	0	0/43	–	0	25	0.333	7/43	4	1
2SNI	0.333	11/71	7	4	0	0.333	11/71	3	3
2SIC	0	0/71	–	0	0	0.333	9/71	3	1
2VIR	0	0/51	–	0	0	0	0/51	–	0
2PCC	0.333	4/29	1	3	0	0.333	4/29	1	4
1BRC	1	46/60	9	10	38	1	48/60	6	5
1BGS	0	0/56	–	0	37	1	32/56	7	4
1UGH	0	0/82	–	0	0	0	0/82	–	0
2KAI	0.333	13/64	2	5	14	0.333	15/64	1	5
1AHW	1	51/73	10	8	80	1	54/73	5	2
1AVZ	0.333	6/33	10	1	0	0.667	11/33	9	4
1DQJ	0	0/72	–	0	1	0.333	12/72	10	1
1FSS	0	0/65	–	0	4	0	0/65	–	0
1WQ1	0	0/91	–	0	0	0	0/91	–	0
1MLC	0	0/56	–	0	1	0	0/56	–	0
1DFJ	0	0/73	–	0	1	0.667	29/73	7	1
1BVK	0	0/53	–	0	0	0.667	16/53	10	3
1CGI	0.667	26/85	5	6	46	1	54/85	2	1
1KKL	0	0/28	–	0	0	0.333	7/28	7	1
1FQ1	0	0/61	–	0	1	0	0/61	–	0
1FIN	0	0/86	–	0	0	0.333	14/86	10	1
	Av. score	TP rate	Av. posn of best	Av. #in top 10	Av.	Av. score	TP rate	Av. posn of best	Av. #in top 10
	0.18	0.12	6.3	1.7	11.5	0.41	0.24	5.5	1.7

Protocol D: docking PCA-derived components for 1DFJ only. D1: global re-ranking D2: best run selected

	Principal components projection D1 re-rank all				D2 best single		
	Complex	CAPRI score	Correct contacts	Position of best	No. in top 10	CAPRI score Correct contacts	Position of best	No. in top 10
1DFJ	0.333	10/73	10	1	0.667	30/73	3	1

Protocol E: Results of docking unbound conformations with 3D-Dock, with the surface thickness at the FTDock stage defined by the amplitude of local fluctuations in MD, compared with default parameters

	Surface softening by RMSF			
Complex	CAPRI score	Correct contacts	Position of best	No. in top 10
2PTC	0	0/68	–	0
1WEJ	0	0/43	–	0
2SNI	0	0/71	–	0
2SIC	0	0/71	–	0
2VIR	0	0/51	–	0
2PCC	0	0/29	–	0
1BRC	0	0/60	–	0
1BGS	0	0/56	–	0
1UGH	0	0/82	–	0
2KAI	0.333	10/64	2	3
1AHW	0.667	36/73	9	1
1AVZ	0.333	4/33	9	1
1DQJ	0	0/72	–	0
1FSS	0	0/65	–	0
1WQ1	0	0/91	–	0
1MLC	0	0/56	–	0
1DFJ	0	0/73	–	0
1BVK	0	0/53	–	0
1CGI	0	0/85	–	0
1KKL	0	0/28	–	0
1FQ1	0	0/61	–	0
1FIN	0	0/86	–	0
	Av. score	TP rate	Av. posn of best	Av. #in top 10
	0.06	0.04	2.2	0.2

In each Table, for each complex, we show the CAPRI score (as defined in Figure 9), the number of correct residue–residue contacts in the interface in the best complex, the position of the highest-scoring complex found in the top ten, and the number of complexes in the top ten that have >10% correct contacts. The bottom row shows averages of all these quantities.

2PTC, 2PCC, 1BGS, 1DFJ and 1BVK, and only 1CGI is worse.

It is worth remarking that if the X-ray unbound to X-ray unbound run is excluded in the cross-docking and re-ranking, then the same five complexes improve but results are worse for three complexes, 1CGI, 2KAI and 1BRC, rather than one. This seems to indicate that, while the MD-cluster-centers are sometimes able to improve docking, they may also in some cases make it worse, as might be expected since they are in general further from the bound structure than the unbound X-ray structure is. However, if the unbound–unbound docking is included and generates a good near-native solution, the scoring functions are often able to pick it out, which rescues the protocol for that complex.

Control runs for docking

Notwithstanding this fact, we believe that part of the reason that a more significant improvement is not found is that the RPScore scoring function, which is presented with between four and nine times as much data in the cluster-center docking runs as the default, is not able to pick out a native-like complex reliably from the very many more possible "false positive" complexes. Some evidence for this comes from comparing the "best" (i.e. closest-to-native) run, B2, with A, when the performance is much better using the cluster centers (13 better, none worse; p-value <0.001). To separate the effects of the amount of data is the function of the second control, randomly spun runs. This is a modified version of the default protocol where multiple runs of the docking are performed with the same conformation of the proteins (the unbound X-ray structure), but starting in different relative spatial orientations. The 3D-Dock algorithm is known to be appreciably sensitive to the starting orientation of the two components, both because the Fourier transform grid is defined relative to the x, y and z axes of the reference coordinate system, and because the angle search step used in each run (9°) is quite coarse and random spinning allows conformations of one run to lie "in between" those generated by another, so these differently oriented dockings act effectively as near-independent runs. The number of such runs is chosen to be the same as the number of cluster-center runs for each complex; four, six, or nine as appropriate (again with one of the runs being the initial unbound–unbound). Thus, the amount of data from which the top ten complexes must be picked is the same for both protocols. The results are in Table 3 protocol C. Again the comparison is done in the same two ways: by re-ranking all complexes from all runs by RPScore (C1), or by picking the best single run for each (C2).

Unlike cluster centers, the re-ranked random spin method (C1) is slightly worse than the default protocol (three worse, the rest the same) and when the best run is picked (C2) it is superior to A in only nine cases, rather than 13 by method B2. Comparing the blind trial methods B1 and C1 directly, results are better with the cluster-center conformations for seven complexes and worse for one (p-value $=0.04$) (in detail, better results are generated for 2PTC, 2SIC, 2PCC, 1BGS, 1DFJ, 1BVK and 1KKL using the cluster centers, while the performance is worse on 1CGI).

Further support to the hypothesis that using cluster-center conformations is helping to find additional good dockings, irrespective of the scoring function used, is given by the comparison of the best of each, protocols B2 and C2. It is found that the cluster-center docking is better for seven complexes and worse for three (p-value $=0.17$). We also remark that a docking with at least 10% correct contacts is generated for several complexes when using MD cluster centers (B2 protocol) where no acceptable docking is produced with the random spins. These are 1UGH, 1MLC and 1FQ1. 1UGH was previously discussed as being the only complex for which a large segment of the polypeptide backbone (in the inhibitor component 1UGI) showed a consistent tendency to approach its bound conformation during the MD; the success could be tentatively ascribed to this. It is also encouraging to have generated an acceptable solution with cluster centers for the difficult case of the CDK2-phosphatase complex 1FQI, where a large conformational change occurs, though this is tempered by the fact that for the very similar CDK2–cyclin complex 1FIN an acceptable complex is generated with protocol C2 but not B2.

Since the skimming off of the top ten conformations is an extreme reduction of the data, we have also evaluated the number of complexes with ligand RMSD <5 Å that are generated in the top 10,000 by each protocol (the top 10,000 being ranked by RPScore over all lists globally). This has been done for protocols A, B1 and C1, and the result is included in the appropriate parts of Table 3. There are more on average in the list from B1 than A, but somewhat surprisingly more still in the list from C1. This may be due to the fact that the superposition of the unbound in positions near to the true binding site will have a smaller RMSD than the corresponding superposition of an MD-generated conformation, and therefore the MD conformations are less tolerant to "near-binding-site" positions.

Docking of MD conformations: essential dynamics may assist in some cases

Yet a third approach, is to take conformations along the first eigenvector(s) identified by essential dynamics simulation and dock these. This is an even more expensive technique than the use of cluster centers, as the distance to project along the eigenvector is not known, so several conformations along it, separated by a few Å in RMSD, must be examined. Therefore, it has been done here for only one of the complexes in the test set, the ribonuclease–ribonuclease inhibitor, 1DFJ. The inhibitor component of this complex (unbound PDB code

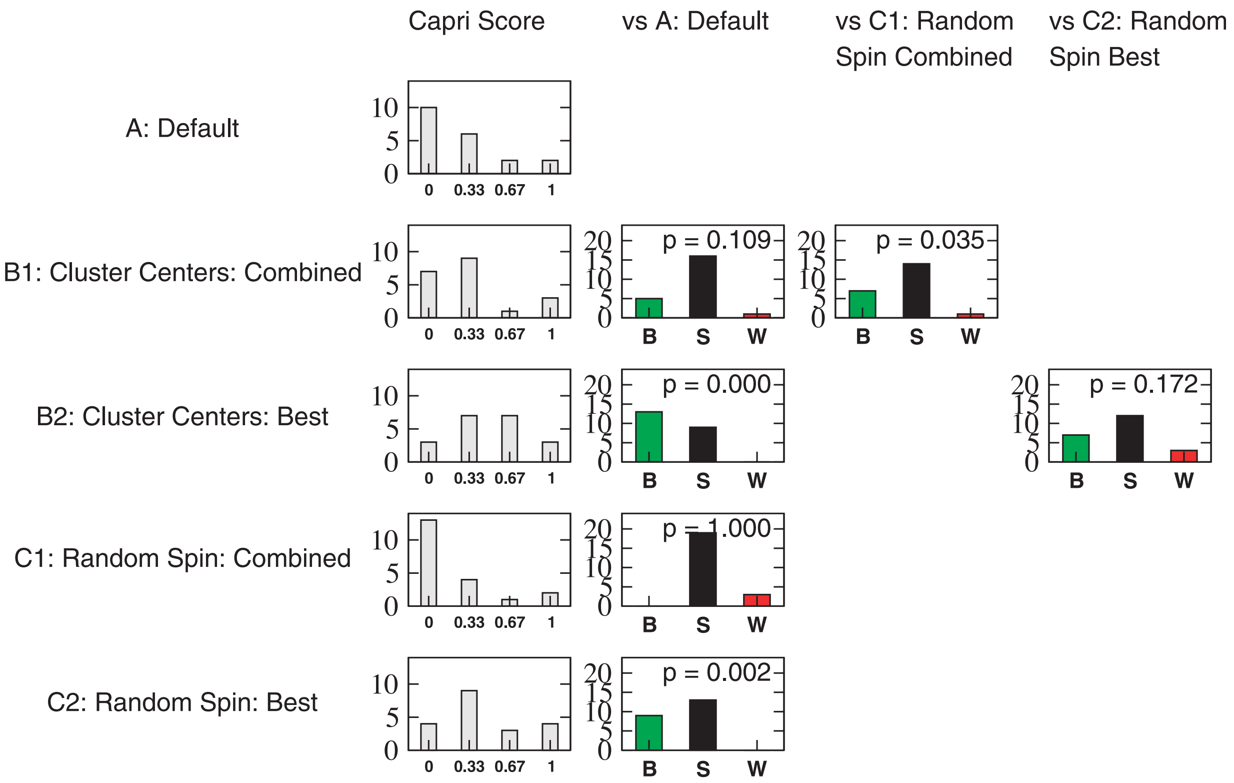


Figure 8. Summary and comparisons of the different docking protocols shown in Table 3. The first column shows a histogram of the CAPRI scores obtained by each method. In the rest of the Figure the protocols are compared. The subfigure at (x, y) shows a histogram of the number of complexes for which a better (B), the same (S) or worse (W) result is obtained by the method in row x, compared to the method in column y as reference. The binomial probability that these results would have if B and W were actually equally likely is also shown.

2BNH) was previously identified as having a domain-movement-like conformational change on binding, and conformations were generated by projecting the C^{α} structure along the eigenvector, rebuilding the side-chains (from the unbound conformer) and energy minimizing. We used only conformations projected along the first principal component eigenvector, and only for the inhibitor. For its ribonuclease partner, the unbound X-ray structure 7RSA was used throughout. The result is shown as protocol D in Table 3. Use of the RPScore function to rank the top complexes pooled over the PCA conformers (D1) succeeds in producing a complex with a CAPRI score of 0.333, better than the default protocol and the random spins (score 0) and equal to the use of cluster centers with RPScore ranking (though in fact the cluster center docking has the higher number of correct contacts). If the best run is selected, (D2) then a complex scoring 0.667 is found; however this is also true for the equivalent cluster centers (B2) and random spin (C2) runs.

It is also interesting that the highest-scoring complexes come from the projection of the complex to -10 (for the global re-ranking D1) and -15 (for the run that actually contains the complex with the greatest number of correct contacts); this is the direction of projection in which the RMSD to the bound decreases, but in both cases the component is actually projected past the minimum and the RMSD has begun to increase again (Figure 3(c)).

Discussion and Conclusion

We have performed a 5 ns simulation for 41 unbound protein complexes to evaluate the extent to which their dynamic structure samples the bound state in the binary complex. A major aim of the study was to develop an approach to introduce flexibility into existing rigid body algorithms. The key conclusions from our study are now described.

Simulation of the test set of 41 proteins: how well does MD sample the bound state?

Our hypothesis was that major regions of most of the proteins would adopt, at least for an instant, the conformation of the bound state. The complete proteins are closer to the bound than to the unbound slightly more than half the time (Figure 2), possibly as a consequence of relieving a bias due to crystallographic contacts. Moreover, when the surface is considered as composed of short segments (Figure 4), half of these segments do indeed sample the bound state, in the sense of fluctuating closer to it than the initial unbound state at some instant, and about 5% are closer even on average. Most surface side-chains, too, sample their bound-state rotamer at some time (Figure 5). However the proteins never, in our simulations, fluctuate closer as a whole to the bound state. This might be a consequence of the use of 5 ns trajectories, which are still very short on a biological or experimental timescale, or of inaccuracies/approximations in the classical Force Field used for MD.

In addition, some of the proteins have high overlap between the fluctuations identified by principal components analysis and their unbound−bound transitions, showing that part of the protein is moving in the right direction even if it does not move all the way.

All the above is in agreement with evidence from other sources that the conformational changes on docking may reflect a conformational sub-state of the unbound component,[47] and that the equilibrium is simply shifted towards this by the presence of the binding partner. The same has also been indicated in the context of protein–ligand docking.[22–24] Of course, in a statistical mechanical description, every conformation of the protein has a finite probability of occurrence in the ensemble, and so any conformation could be regarded as produced by a shift of a pre-existing equilibrium, even though the vast majority of possible conformations are so unlikely that they would not be expected ever to be observed over any reasonable biological timescale. The operational definition, then, must be that the bound state is an observable sub-state, as of course it will be in experimental work. One might then, tentatively categorize the binding behaviour of complex-forming proteins (or regions of them) in three ways: those that undergo very little change (which are less of a challenge, though not trivial, for computational protein–protein docking), those where the docked state is a visited sub-state, and those where it must be induced by a partner.

Docking using MD conformations

Encouragingly, even with classical MD, there is some evidence that this partial capturing of aspects of the bound state has advanced the quality of the docking results, as indicated by our extensive comparison of cross-docking of MD-generated conformations with randomly generated starting unbound conformations. Unfortunately the improvement is on the edge of significance; further developments in protocols, improvements in the scoring function, or a larger data set are required.

Significantly less mobility in the core of the known interface, and more in the periphery

The division of the interfaces into core and periphery regions clearly shows that properties relating to mobility (RMSF and side-chain entropy) are different in the two regions, the core being less mobile and the periphery more so: the average values of the properties in the regions are separated by >2 standard deviations, and up to 40/41 proteins individually showed this behaviour (reduced to 30 if the requirement is that the result for both RMSF and side-chain entropy be consistent). Moreover, in the core, a higher proportion of the side-chains (as identified by MD) than in the rest of the surface are in the correct rotamer for docking. Interestingly, these observations may relate to a previous suggestion, that protein–protein binding sites tend to contain regions which make both high and low contributions to the stability of the protein.[56] In addition, they appear to corroborate the recent findings of Ramajani *et al*.[26] that certain key residues in the core act as "anchors", having a high occupancy of the bound state rotamer in the unbound ensemble. We remark that these differences between different surface regions are likely to be more robust to changes in the MD force field than are absolute values of RMSDs, rotamer occupancies etc. described above.

In future work, to enhance our docking efforts, we intend to apply our findings on core/peripheral mobility to help us to identify potential interface residues. Approaches based on e.g. patches of sequence conservation, or interaction energy[57] are powerful, though a combination of both conservation and mobility could aid in identification. Indeed, use of a genetic algorithm to combine information from a variety of sources, including conservation, has recently proven to be better than any single method in interface prediction.[53] It was also important to consider spatial clusters of the properties of interest, an approach that was also used in the identification of binding sites[54] and homodimer interfaces.[58] We remark that the protein–protein interfaces where antibodies bind seem more difficult to identify than other interfaces:[57] the mobility-based measures used here might help in this problem.

Where did MD improve docking?

When using MD cluster-center conformations globally re-ranked together with the unbound, there was an indication of a small improvement in performance over a single run on the unbound components. Inadequacies of the scoring functions are partly responsible for this: they are swamped by false positives when attempting to cope with the larger volume of data provided by the MD cluster-center runs. This is shown by the fact that, when the random spin runs were used as controls to equalize the volume of data, the MD cluster-centers perform better than the random spins. Moreover, the comparison of results between "blind trial" runs and runs where the best solution is selected also indicate that improvements to the scoring functions would enable selection of better solutions.

Perhaps the application of MD that may prove most useful to docking in cases of large conformational change is to extract information about subdomain motion with essential dynamics, as was done in the case of the ribonuclease inhibitor. The way of using the PCs produced by essential dynamics to generate an ensemble of receptor conformations is similar to recent work by Zacharias;[59] another approach, not investigated here, is to use the PCs as collective coordinates in a minimization algorithm. So far this has only been done in protein–ligand docking.[60] PCs extracted from simplified (not full-atomic MD) models have also given good results, indicating that it may be possible to replace the full MD+ED analysis conducted here with a simpler, and less time-consuming, approach.

In connection with this, we would like to make the observation that PC analysis proved effective for us in a recent round of the CAPRI blind trial of protein–protein docking,[55] in which it was required to predict the structure of the trimeric form of the fusion protein from the virus TBEV given the coordinates of the dimer. There turns out to be a substantial sub-domain movement, straightening the protein, resulting in a backbone RMSD of more than 4 Å.[61] We tackled this problem in a similar way to the ribonuclease, carrying out an MD simulation and projecting along a principal component to obtain structures for docking. One of these captured enough of the conformational change to enable us to submit a complex with $>10\%$ correct contacts, which placed us amongst the best few groups for this target. (We have not included a full analysis of the docking of TBEV fusion protein in this work because the simulation used a different MD protocol and 3D-Dock was also specially modified to take into account the 3-fold symmetry.)

It seems, then, that if one or more of the docking partners undergoes clear domain or sub-domain movement, use of MD configurations can significantly improve docking results.

Future directions in docking

The most important observation made here is that the core of the interface is more rigid and predisposed for docking, while the periphery is more flexible even than the rest of the surface. Our current docking algorithms do not consider this partition. If we were able to dock core regions confidently and then allow small readjustments of these regions, in conjunction with "gelling" of the periphery, this could produce a more tractable approach. This may reflect the physical time course of docking: first a fast core–core recognition, followed by a much longer process of conformational sorting at the periphery.[26] The difference observed in the side-chain entropy in these regions of the interface may have implications for the affinity of binding: since there is less configurational entropy to be lost in the core, its contribution to binding will be greater. However, there is one important caveat: core regions need to be predicted with confidence first.

It may be useful in future work to consider two avenues of enquiry: the further development of the approach we have taken here, utilizing the observed conformational states of the unbound protein in solution; and the obvious need to include both proteins in the MD simulations.

Methods

Molecular dynamics simulations and analysis were carried out with Gromacs v 2.0,[62] 3.0[63] and 3.1 on a farm of Intel Pentium III and IV processors running Linux. The simulations were performed according to the following protocol.

The proteins were taken from the PDB.[64] Non-protein prosthetic groups were removed before the simulation. This affects 1VXR, 1HRC, 5P21, 1YCC and 1CCA. If the protein formed homodimers or homotrimers, a monomer alone was used for simulation, and in certain cases the polypeptide chain was cut into domains and only the domain that forms the interface on docking was included in the simulation. The proteins affected by this were the antibodies 1QBL and 1BVL, where only the variable domain of the L and H chains was used. Hemagglutinin (2HMG) is a functional trimer of dimers, but only a monomer of the larger polypeptide was simulated. It has several sub-domains, a "head" consisting of receptor-binding and esterase sub-domains and a long flexible receptor-binding sub-domain;[65] in our simulations the receptor-binding sub-domain was truncated by removing residues 1–42 and 310–326. HPrK/P (1JB1) was simulated as a monomer. The smaller polypeptide that is present in the hexamer but not our simulations also lies almost entirely in the receptor-binding sub-domain. In general, missing residues at the termini of a chain were not included; those at the center of a chain, in CDK2 (1HCL) and 1JB1 were modeled using spdbv[66] and/or modeler.[67]

Each protein was then solvated by combining with replicas of a pre-equilibrated box of SPC waters, those waters that make bad contacts being removed. Crystal waters present in the initial X-ray structure were retained. Enough Na^+ or Cl^- ions were added to neutralize the protein charge, then further ions were added corresponding to a salt solution of concentration 0.1 M. The ions were added by randomly replacing water molecules. The resulting system was energy minimized by steepest descents, then equilibrated for 40 ps with water and ions free to move but all protein atoms restrained to their crystallographic coordinates with harmonic restraints with spring constant $1000\ kJ\ mol^{-1}\ nm^{-2}$. A further 40 ps equilibration was performed with all atoms free to move, then a 5 ns production run (often run in segments of 1 ns or 2 ns) was carried out, again with no harmonic restraints on the protein. In the first (restrained) equilibration stage the MD integration timestep was 2 fs; in the second equilibration and production stages it was increased to 5 fs by using the heavy hydrogen and dummy atom features of Gromacs.[68] The force field used was GROMOS 96[69] and non-bonded forces were treated using a 1.4 nm cut-off for van der Waals forces and short-range electrostatics, long-range electrostatics being treated by a reaction field with an effective relative dielectric constant of 54.[70] Coordinate frames were saved every 10 ps from the MD trajectory for subsequent analysis. The protocol is similar to one used previously in van Gunsteren's group.[71] The algorithms and force fields currently recommended for use in protein simulations with the Gromacs package are the particle-mesh Ewald method for treatment of long-range electrostatics (though twin-range cut off with reaction field is not deprecated) and the OPLS-AA force field with a 2 fs timestep. These developments, however, were not available in Gromacs when the work reported here was begun, and so we chose to retain our original simulation protocol throughout. However, the OPLS-AA force field may be slightly more accurate, which might affect the details of the statistics of RMSDs, rotamer occupancies, etc.

Essential dynamics calculations were performed by calculating and diagonalizing the matrix of atomic fluctuations observed over the course of the MD trajectory:

$$C_{ij} = \langle (S_i - \langle S_i \rangle)(S_j - \langle S_j \rangle) \rangle$$

where i and j run over the three spatial coordinates S of each atom (3N).[42,43] Only C^α atoms were considered.

The displacement patterns in a normal mode can be compared with the conformational changes on docking using as a measure the overlap I:

$$I_j = \frac{|\sum_{i=1}^{N} \mathbf{q}_{ij} \cdot \Delta \mathbf{r}_i|}{\sum_{i=1}^{N} \mathbf{q}_{ij}^2 \sum_{i=1}^{N} \Delta r_i^2}$$

and the correlation c:

$$c_j = \frac{\mathrm{cov}(|\mathbf{q_j}|, |\Delta \mathbf{r}|)}{\sqrt{\mathrm{var}(|\mathbf{q_j}|)\mathrm{var}(|\Delta \mathbf{r}|)}}$$

where this time i runs over the N atoms in the molecule that were considered in the ED analysis (C^α atoms).[72] $\mathbf{q_{ij}}$ is the displacement vector of the ith atom in the jth normal mode and $\Delta\mathbf{r_i}$ is the conformational change vector of the ith atom, i.e. $\mathbf{r_i}$(bound) − $\mathbf{r_i}$(unbound), where the all coordinate frames (in the calculation of $\Delta\mathbf{r_i}$ and $\mathbf{q_{ij}}$) have previously been aligned by least-squares fitting. In the second equation $|\mathbf{q_j}\ |$ and $|\Delta\mathbf{r}|$ refer to the sets of the moduli of these vectors.

To extract specific large-scale sub-domain motions from a PC eigenvector we use the program DynDom[45,46] operating on two configurations obtained by projection along it; DynDom identifies dynamic sub-domains as regions whose relative movement between the two input configurations can be described by a general rigid-body transformation. This has been done for the first five eigenvectors (ranked by size of associated eigen value) of each protein. To compare with the unbound–bound transition, DynDom was also run on the bound and unbound X-ray crystal structures of the proteins. All DynDom runs were carried out with default parameters.

Side-chain rotamer contingency analysis was carried out by considering the rotamer of an entire side-chain to be each unique combination of the rotamers of the non-fixed torsion angles along it, put into bins of 120 or 180 degree width as appropriate.

Protein–protein docking was done with a modified version of the 3D-Dock suite 2.0.[8,30–33] The initial global shape complementarity search was carried out with a grid spacing of 0.7 Å and an angle step size of 9°. Twenty-seven thousand possible dockings were kept from this stage, and then reranked with a scoring function (RPScore) derived from the statistics of residue–residue contacts in a database of 92 non-redundant complexes. Biological filters were applied defining a few residues likely to be in the interface (detailed in Supplementary Material). Clustering of solutions was carried out using a single-linkage algorithm, the representative structure of each cluster being taken to be that with the best RPScore.

When comparing the results of docking with different protocols, the p-value is evaluated in the following way: We work with the CAPRI scores of the "query method", which can be better, the same or worse than the corresponding scores of the "reference method". The effective number of trials is $N_{\mathrm{e}} = N - N_{\mathrm{s}}$, where N is the number of complexes and N_{s} is the number of complexes where the same CAPRI score is obtained by the two methods. Then if we obtain N_{B} better by the "query method", the significance of this is the probability of getting at least N_{B} better in a model where the single-trial probabilities of better and worse are taken to be the same:

$$p(N_{\mathrm{B}}) = \sum_{i=N_{\mathrm{B}}}^{N_e} P(i) = \sum_{i=N_{\mathrm{B}}}^{N_e} {}^{N_e}C^i \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{N_e - i}$$

Acknowledgements

The authors are grateful to Chris Page of the Biomolecular Modelling Laboratory (BMM) for help with Figure 6 and to Gavin Kelly of the Computational Genome Analysis Laboratory at CR-UK for help with use of the package R. We also acknowledge helpful discussions with all the members of the BMM and with other workers in the docking field.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.01.058

References

1. Smith, G. R. & Sternberg, M. J. (2002). Prediction of protein–protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12**, 28–35.
2. Camacho, C. J. & Vajda, S. (2002). Protein–protein association kinetics and protein docking. *Curr. Opin. Struct. Biol.* **12**, 36–40.
3. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002). Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins: Struct. Funct. Genet.* **47**, 409–443.
4. Palma, P. N., Krippahl, L., Wampler, J. E. & Moura, J. J. G. (2000). BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins: Struct. Funct. Genet.* **39**, 372–384.
5. Heifetz, A. & Eisenstein, M. (2003). Effect of local shape modifications of molecular surfaces on rigid-body protein–protein docking. *Protein Eng.* **16**, 179–185.
6. Sandak, B., Wolfson, H. J. & Nussinov, R. (1998). Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins: Struct. Funct. Genet.* **32**, 159–174.
7. Shatsky, M., Nussinov, R. & Wolfson, H. J. (2004). FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. *J. Comput. Biol.* **11**, 83–106.
8. Jackson, R. M., Gabb, H. A. & Sternberg, M. J. E.

(1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.* **276**, 265–285.

9. Camacho, C. J. & Vajda, S. (2001). Protein docking along smooth association pathways. *Proc. Natl Acad Sci. USA*, **98**, 10636–10641.
10. Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2002). Soft protein–protein docking in internal coordinates. *Protein Sci.* **11**, 280–291.
11. Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins: Struct. Funct. Genet.* **52**, 113–117.
12. Fitzjohn, P. W. & Bates, P. A. (2003). Guided docking: first step to locate potential binding sites. *Proteins: Struct. Funct. Genet.* **52**, 28–32.
13. Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A. & Baker, D. (2003). Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331**, 281–299.
14. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737.
15. Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (2003). Assessment of progress over the CASP experiments. *Proteins: Struct. Funct. Genet.* **53**, 585–595.
16. Tovchigrechko, A., Wells, C. A. & Vakser, I. A. (2002). Docking of protein models. *Protein Sci.* **11**, 1888–1896.
17. Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S. *et al.* (2004). Structure-based assembly of protein complexes in yeast. *Science*, **303**, 2026–2029.
18. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S. *et al.* (2003). CAPRI: a critical assessment of predicted interactions. *Proteins: Struct. Funct. Genet.* **52**, 2–9.
19. Wodak, S. J. & Mendez, R. (2004). Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.* **14**, 1–8.
20. Camacho, C. J., Kimura, S. R., DeLisi, C. & Vajda, S. (2000). Kinetics of desolvation-mediated protein–protein binding. *Biophys. J.* **78**, 1094–1105.
21. Kleanthous, C. (2000). *Protein–Protein Recognition*, Oxford University Press, Oxford.
22. Najmanovich, R., Kutter, J., Sobolev, V. & Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins: Struct. Funct. Genet.* **39**, 261–268.
23. Ma, B., Shatsky, M., Wolfson, H. J. & Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **11**, 184–197.
24. Kern, D. & Zuiderweg, E. R. (2003). The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.* **13**, 748–757.
25. Kimura, S. R., Brower, R. C., Vajda, S. & Camacho, C. J. (2001). Dynamical view of the positions of key side chains in protein–protein recognition. *Biophys. J.* **80**, 635–642.
26. Rajamani, D., Thiel, S., Vajda, S. & Camacho, C. J. (2004). Anchor residues in protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **101**, 11287–11292.
27. Chen, R., Mintseris, J., Janin, J. & Weng, Z. (2003). A protein–protein docking benchmark. *Proteins: Struct. Funct. Genet.* **52**, 88–91.
28. Inbar, Y., Benyamini, H., Nussinov, R. & Wolfson, H. J. (2003). Protein structure prediction *via* combinatorial assembly of sub-structural units. *Bioinformatics*, **19**, i158–i168.
29. Lo Conte, L., Chothia, C. & Janin, J. (1998). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
30. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272**, 106–120.
31. Sternberg, M. J. E., Aloy, P., Gabb, H. A., Jackson, R. M., Moont, G., Querol, E. & Aviles, F. X. (1998). A computational system for modelling flexible protein–protein and protein–DNA docking. In *Proceedings Sixth International Conference on Intelligent Systems for Molecular Biology* (Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D. & Sensen, C., eds), pp. 183–192, AAAI Press, Menlo Park, CA, USA.
32. Moont, G., Gabb, H. A. & Sternberg, M. J. E. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins: Struct. Funct. Genet.* **35**, 364–373.
33. Sternberg, M. J. E. & Moont, G. (2001). Modelling protein–protein and protein–DNA docking. In *Bioinformatics—From Genomes to Drugs* (Lengauer, T., ed.), Vol. 1, pp. 361–404, Wiley-VCH, Weinheim.
34. Heifetz, A., Katchalski-Katzir, E. & Eisenstein, M. (2002). Electrostatics in protein–protein docking. *Protein Sci.* **11**, 571–587.
35. Lorber, D. M. & Shoichet, B. K. (1998). Flexible ligand docking using conformational ensembles. *Protein Sci.* **7**, 938–950.
36. Lorber, D. M., Udo, M. K. & Shoichet, B. K. (2002). Protein–protein docking with multiple residue conformations and residue substitutions. *Protein Sci.* **11**, 1393–1408.
37. Klebe, G., Kramer, O. & Sotriffer, C. (2004). Strategies for the design of inhibitors of aldose reductase, an enzyme showing pronounced induced-fit adaptations. *Cell Mol. Life Sci.* **61**, 783–793.
38. Cavasotto, C. N. & Abagyan, R. A. (2004). Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **337**, 209–225.
39. Lin, J. H., Perryman, A. L., Schames, J. R. & McCammon, J. A. (2003). The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers*, **68**, 47–62.
40. Lin, J. H., Perryman, A. L., Schames, J. R. & McCammon, J. A. (2002). Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J. Am. Chem. Soc.* **124**, 5632–5633.
41. Amadei, A., Ceruso, M. A. & Di Nola, A. (1999). On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Struct. Funct. Genet.* **36**, 419–424.
42. Garcia, A. E. (1992). Large-amplitude nonlinear motions in proteins. *Phys. Rev. Letters*, **68**, 2696–2699.
43. Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. (1993). Essential dynamics of proteins. *Proteins: Struct. Funct. Genet.* **17**, 412–425.
44. Smith, G. R., Contreras-Moreira, B., Zhang, X. & Bates, P. A. (2004). A link between sequence conservation and domain motion within the AAA+ family. *J. Struct. Biol.* **146**, 189–204.
45. Hayward, S. & Berendsen, H. J. (1998). Systematic analysis of domain motions in proteins from

conformational change: new results on citrate synthase and T4 lysozyme. *Proteins: Struct. Funct. Genet.* **30**, 144–154.

46. Hayward, S. & Lee, R. A. (2002). Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J. Mol. Graph. Model.* **21**, 181–183.
47. Echols, N., Milburn, D. & Gerstein, M. (2003). MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucl. Acids Res.* **31**, 478–482.
48. Betts, M. J. & Sternberg, M. J. E. (1999). An analysis of conformational changes on protein–protein docking: implications for predictive docking. *Protein Eng.* **12**, 271–283.
49. Pickett, S. D. & Sternberg, M. J. E. (1993). Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.* **231**, 825–839.
50. Sternberg, M. J. E. & Chickos, J. S. (1994). Protein side-chain conformational entropy derived from fusion data—comparison with other empirical scales. *Protein Eng.* **7**, 149–155.
51. Doig, A. J. & Sternberg, M. J. E. (1995). Side chain conformational entropy in protein folding. *Protein Sci.* **4**, 2247–2251.
52. Shenkin, P. S., Farid, H. & Fetrow, J. S. (1996). Prediction and evaluation of side-chain conformations for protein backbone structures. *Proteins: Struct. Funct. Genet.* **26**, 323–352.
53. Neuvirth, H., Raz, R. & Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **338**, 181–199.
54. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. E. (2001). Automated structure-based prediction of functional sites in proteins—applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408.
55. Mendez, R., Leplae, R., De Maria, L. & Wodak, S. J. (2003). Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins: Struct. Funct. Genet.* **52**, 51–67.
56. Luque, I. & Freire, E. (2000). Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins*, 63–71.
57. Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2004). Identification of protein–protein interaction sites from docking energy landscapes. *J. Mol. Biol.* **335**, 843–865.
58. Valdar, W. S. & Thornton, J. M. (2001). Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Struct. Funct. Genet.* **42**, 108–124.
59. Zacharias, M. (2003). Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* **12**, 1271–1282.
60. Zacharias, M. (2004). Rapid protein-ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: binding of FK506 to FKBP. *Proteins: Struct. Funct. Genet.* **54**, 759–767.
61. Bressanelli, S., Stiasny, K., Allison, S. L., Stura, E. A., Duquerroy, S., Lescar, J. *et al*. (2004). Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. *EMBO J.* **23**, 728–738.
62. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. (1995). GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56.
63. Lindahl, E., Hess, B. & van der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Mod.* **7**, 306–317.
64. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acid Res.* **28**, 235–242.
65. Wiley, D. C. & Skehel, J. J. (1987). The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu. Rev. Biochem.* **56**, 365–394.
66. Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
67. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
68. Feenstra, K. A., Hess, B. & Berendsen, H. J. (1999). Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20**, 786–798.
69. Scott, W. R. P., Huenenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J. *et al*. (1999). The GROMOS biomolecular simulation program package. *J. Phys. Chem. A*, **102**, 3596–3607.
70. Tironi, I. G., Sperb, R., Smith, P. E. & van Gunsteren, W. F. (1995). A generalized reaction-field method for molecular dynamics simulations. *J. Chem. Phys.* **102**, 5451–54459.
71. Voordijk, S., Hansson, T., Hilvert, D. & van Gunsteren, W. F. (2000). Molecular dynamics simulations highlight mobile regions in proteins: a novel suggestion for converting a murine V(H) domain into a more tractable species. *J. Mol. Biol.* **300**, 963–973.
72. Tama, F. & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **14**, 1–6.
73. Hubbard, S. J., Campbell, S. F. & Thornton, J. M. (1991). Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.* **220**, 507–530.

Edited by B. Honig

(Received 13 August 2004; received in revised form 4 January 2005; accepted 21 January 2005)