doi:10.1016/j.jmb.2003.10.069 J. Mol. Biol. (2004) 335, 843–865

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Identification of Protein–Protein Interaction Sites from Docking Energy Landscapes

Juan Fernández-Recio[1], Maxim Totrov[2] and Ruben Abagyan[1]*

[1]*Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037 USA*

[2]*Molsoft, LLC, 3366 Torrey Pines Court, La Jolla, CA 92037, USA*

**Corresponding author*

Protein recognition is one of the most challenging and intriguing problems in structural biology. Despite all the available structural, sequence and biophysical information about protein–protein complexes, the physico-chemical patterns, if any, that make a protein surface likely to be involved in protein–protein interactions, remain elusive. Here, we apply protein docking simulations and analysis of the interaction energy landscapes to identify protein–protein interaction sites. The new protocol for global docking based on multi-start global energy optimization of an all-atom model of the ligand, with detailed receptor potentials and atomic solvation parameters optimized in a training set of 24 complexes, explores the conformational space around the whole receptor without restrictions. The ensembles of the rigid-body docking solutions generated by the simulations were subsequently used to project the docking energy landscapes onto the protein surfaces. We found that highly populated low-energy regions consistently corresponded to actual binding sites. The procedure was validated on a test set of 21 known protein–protein complexes not used in the training set. As much as 81% of the predicted high-propensity patch residues were located correctly in the native interfaces. This approach can guide the design of mutations on the surfaces of proteins, provide geometrical details of a possible interaction, and help to annotate protein surfaces in structural proteomics.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: protein–protein interactions; binding site identification; pseudo-Brownian Monte Carlo; docking energy landscapes; hot spots

Introduction

Protein–protein interactions (PPI) play a key role in many biological processes such as signal transduction, gene expression control, enzyme inhibition, antibody–antigen recognition or even the assembly of multi-domain proteins. In the cellular context, these PPIs very often lead to the formation of stable specific protein–protein complexes that are essential to perform their biological functions. Identification of protein–protein interactions can be performed through a combination of experimental and theoretical techniques, and several protein interaction databases are being compiled.[1] However, the exact location and geometry of the interface poses a separate problem, which cannot always be addressed with X-ray crystallography or NMR studies. As a result, the demand for computer modeling is growing, and a number of algorithms that generate models of the macromolecular complex based on the structures of individual subunits (docking methods) have been described.[2–7] However, these docking procedures can only generate the near-native complex structure among a number of alternative docked conformations (false-positives), and some information about the interacting surfaces (biological data, sequence and structure homology) is frequently required to single out the correct solution. The knowledge of the approximate location of the binding interface

Present address: J. Fernández-Recio, Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK.

Abbreviations used: PPI, protein–protein interaction; PDB, Protein Data Bank; ICM, internal coordinate mechanics; RMSD, root-mean-square deviation; ASA, accessible surface area; ASP, atomic solvation parameter; ABS, averaged buried surface; RBS, relative buried surface; IP, interface propensity; NIP, normalized interface propensity; EPO, erythropoietin; EPOR, erythropoietin receptor; MC, Monte Carlo; FNR, ferredoxin-NADP$^+$ reductase; Fd, ferredoxin.

E-mail address of the corresponding author: abagyan@scripps.edu

0022-2836/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.

on the protein surface can be of direct practical use as well as help to improve the accuracy of the docking models. Chemical and physical complementarity between the interacting surfaces is essential, and the existence of "hot spots", e.g. a few residues that confer most of the binding energy, has been reported.[8–12] In this context, the question arises of whether there are specific chemical and physical characteristics on the surfaces of proteins that can be used to identify the PPI sites. Different analyses of known protein–protein interfaces attempted to find specific characteristics that differentiate PPI sites from the rest of the protein surface.[13–19] However, in hetero-complexes the PPI site composition does not differ significantly from the rest of the solvent-accessible surface.[20–22] In addition, the results of continuum electrostatic calculations suggest that protein–protein interfaces are naturally designed to exploit electrostatic and hydrophobic forces in very different ways.[23] Thus, it appears difficult to find chemical or structural patterns on the surface of proteins that unequivocally define PPI sites, and very few computational methods have been developed so far for their prediction. A method based on residue interface propensities and physico-chemical analysis of surface patches, presented a success rate of 54% on a dataset of 59 complexes, which included homo-dimers, hetero-complexes, and antibody–antigen complexes (prediction was considered successful when at least one of the three predicted patches covered >50% of the real interface).[24] More recently, a neural network based on sequence profiles of neighboring residues and solvent-exposure values was trained on a dataset of 615 protein–protein complex structures to predict interacting regions.[25] The method was tested on a set of 35 unbound proteins, where 70% of the predicted interface residues were located correctly in the interfaces. Unfortunately, results obtained from neural networks are difficult to interpret in physical or phenomenological terms. The conformational changes upon protein–protein association are often limited to local movements, which suggests that protein–protein association is driven by rigid-body fit.[21,26–28] Thus, the idea of analyzing the rigid-body docking energy landscape in search of protein recognition areas is highly attractive. The possibility of low-resolution recognition in protein–protein association has been studied recently using geometry-based rigid-body docking simulations.[29]

In the present work, we use rigid-body docking simulations, with a grid-based energy description of the system, to analyze the existence of preferred interaction sites on a protein surface. The soft interaction energy and the pseudo-Brownian conformational sampling procedure used here were tested previously and optimized on a benchmark of known protein–protein complexes for which the uncomplexed structures were determined.[4] The energy function was shown to be sufficiently accurate and tolerant to the induced fit to select the near-native docked solution in most cases, and the search procedure was able to explore thoroughly the rotational and translational degrees of freedom of the ligand in the vicinity of the binding site of the receptor (so called local docking). Here, we address the more challenging task of exploring the conformational space around the whole receptor, and we show that the energy profile for the ensemble of found docked poses can be used to determine accurately interaction sites on protein surfaces.

Results

Global rigid-body docking

Rigid-body docking simulations were applied to a large dataset of protein–protein complexes (Table 1) using the 3-D coordinates of the unbound subunits. The method used here permits a complete sampling of the whole receptor surface, without any restriction, and is based on a previously described multiple-start pseudo-Brownian Monte-Carlo minimization procedure,[4] which was extended to sample entire surfaces for both proteins and to adapt automatically the distribution of starting points to the shape and size of the proteins as described in Materials and Methods. The modified procedure is capable of locating nearly all distinct (farther than 4 Å of interface RMSD apart) local minima of the grid energy function. The number of conformations obtained from the simulations depended on the size and shape of the interacting molecules. For instance, a total of 5958 different conformations were obtained for the complex chymotrypsin/APPI (PDB 1ca0), whereas a total of 12,064 conformations were obtained for the considerably larger complex Fab D44.1/lysozyme (PDB 1mlc).

The RMSD/energy distributions of all docked poses for some of the complexes are shown in Figure 1(a). Despite the rather noisy distribution of the individual energy values, the behavior of the averaged energy for the solutions within a sliding RMSD range suggests an overall dependence of the energy on the proximity to the real structure, which for some complexes points towards the native solution ("funnel" shape).

Inclusion of a detailed solvation term and optimization of the energy estimate on the training set

To accelerate calculations, only surface charge scaling (SChEM)[30] and a crude hydrophobic contact grid potential were used during the docking simulations to account for the solvation effects. To include the desolvation penalty of burying solvent-exposed polar groups, as well as to improve the precision of the hydrophobic energy calculation, the energy value for every docked pose was recalculated after the simulations using an

Table 1. Benchmark of protein–protein complexes used in this work

Receptor	Ligand	Complex	Unbound receptor	Unbound ligand
Parameterization set				
Chymotrypsin	APPI	1ca0	5cha	1aap
Chymotrypsin	BPTI	1cbw	5cha	1bpi
Chymotrypsin	Eglin C	1acb	5cha	1egl
Chymotrypsin	Ovomucoid	1cho	5cha	1omu
Chymotrypsinogen	HPTI	1cgi	1chg	1hpt
Kallikrein A	BPTI	2kai	2pka	1bpi
Subtilisin BPN	CI-2	2sni	2st1	2ci2
Subtilisin BPN	SSI	2sic	2st1	3ssi
Subtilisin Carlsberg	Eglin C	1cse	1sbc	1egl
Thermitase	Eglin C	2tec	1thm	1egl
Trypsin (bovine)	APPI	1taw	5ptp	1aap
Trypsin (bovine)	BPTI	2ptc	5ptp	1bpi
Trypsin (rat)	BPTI	3tgi	1ane	1bpi
Trypsin (rat)	APPI	1brc	1bra	1aap
Acetylcholinesterase	Fasciculin II	1fss	2ace	1fsc
α-amylase	Tendamistat	1bvn	1pif	2ait
Barnase	Barstar	1bgs	1a2p	1a19
Ribonuclease Sa	Barstar	1ay7	1rge	1a19
TEM-1 β-lactamase	BLIP	TEM1[a]	RETM[a]	BLIP[a]
UDG	UGI	1ugh	1akz	2ugi
Cytochrome c peroxidase	Cytochrome *c*	2pcb	1ccp	1hrc
Cytochrome f	Plastocyanin	2pcf	1ctm	1ag6
Fab D44.1	Lysozyme	1mlc	1mlb	1lza
Fv D1.3	Lysozyme	1vfb	1vfa	1lza
Test set				
Ferredoxin-NADP + reductase	Ferredoxin	1ewy	1que	1fxa
Erythropoietin receptor	Erythropoietin	1eer	1ern	1buy
HPr kinase	HPr	1kkl	1jb1	1sph
VP6 (rotavirus)	Fab	VP6FAB[b]	1qhd	VP6FAB[c]
Hemagglutinin	Fab HC63	1ken	2viu	1ken[c]
α-amylase	VHH Amd10	1kxv	1pif	1kxv[c]
α-amylase	VHH Amb7	1kxt	1pif	1kxt[c]
α-amylase	VHH Amd9	1kxq	1pif	1kxq[c]
TCR-β	speA	1l0x	1bec	1b1z
Trypsin	Soybean trypsin inhibitor	1avw	2ptn	1ba7
Ribonuclease inhibitor	Ribonuclease A	1dfj	2bnh	7rsa
Trypsinogen	PSTI	1tgs	2ptn	1hpt
Fab 5G9	Tissue factor	1ahw	1fgn	1boy
Hyhel-63 Fab	Lysozyme	1dqj	1dqq	3lzt
IgG1 E8 Fab	Cytochrome *c*	1wej	1qbl	1hrc
HIV-1 NEF	FYN TK SH3 domain	1avz	1avv	1shf
RAS activating domain	RAS	1wq1	1wer	5p21
Methylamine dehydrogenase	Amicyanin	2mta	2bbk	1aan
Thrombin mutant	Pancreatic trypsin inhibitor	1bth	2hnt	6pti
CDK2	Cyclin	1fin	1hcl	1vin
CDK2	KAP	1fq1	1b39	1fpz

The selection of protein–protein complexes is described elsewhere.[4]

[a] No coordinates deposited in PDB: the structures were kindly provided by the authors.[57–59]

[b] No coordinates deposited in PDB at the present time; the structure of the VP6/Fab complex was made available by the authors (M. C. Vaney & F. Rey, unpublished results) to constitute target 2 in the recent CAPRI competition (http://capri.ebi.ac.uk).

[c] We used for the docking simulations the coordinates of the bound ligand in a random orientation, as provided by the CAPRI organization (http://capri.ebi.ac.uk).

atomic solvent-accessible surface area (ASA)-based solvation model (see Materials and Methods). We had found that the atomic solvation parameters (ASP) derived from vapor/water transfer energies improved the local docking results.[4] Here, we extend and optimize this solvation model for the "global docking" by using ASP derived from octanol/water transfer energies (see Materials and Methods), which better describe the burial of solvent-exposed residues in the interface.[31]

The following energy function was used to re-evaluate the set of conformations generated during the docking simulations:

$$E = \alpha(E_{\mathrm{Hvw}} + E_{\mathrm{Cvw}}) + E_{\mathrm{el}} + \beta E_{\mathrm{hb}} + \gamma E_{\mathrm{polar}} + \delta E_{\mathrm{ar}} + \varepsilon E_{\mathrm{al}} \quad (1)$$

The coefficients (α, β, γ, δ, ε) assigned to each energy term were optimized to obtain the best possible ranking for the near-native solutions in all complexes of the training set listed in Table 1 (the fitting procedure is detailed in Materials and Methods). The same training set was previously

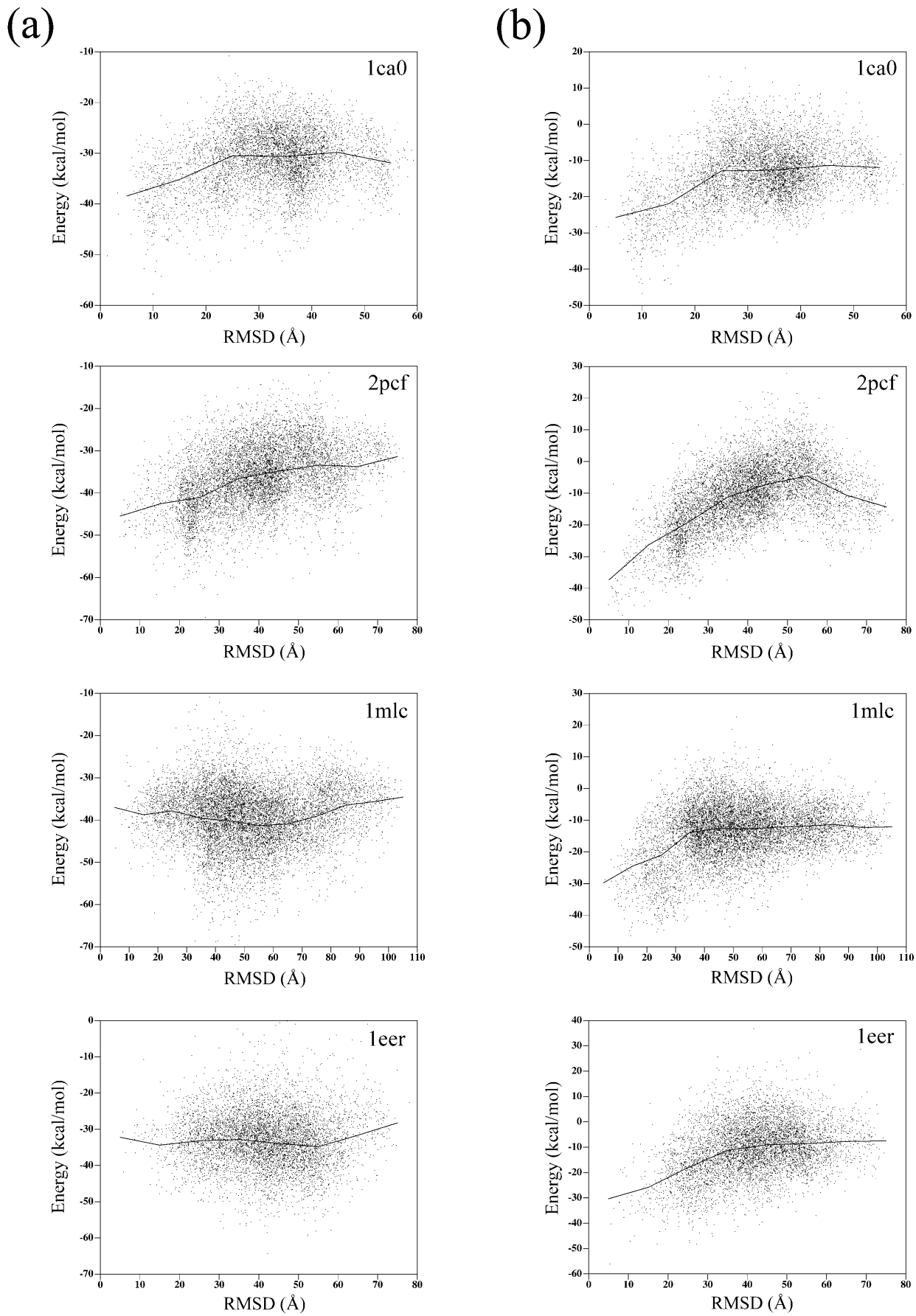

Figure 1. (a) Distribution of the rigid-body docking poses for different complexes after the simulations, as re-evaluated by equation (3) in Materials and Methods. (b) The same docking poses re-evaluated with the optimized energy

used to benchmark the initial docking procedure.[4] The non-solvation energy terms are described elsewhere:[4] E_{Hvw} and E_{Cvw} are the van der Waals potentials pre-calculated for a hydrogen and a heavy-atom probe, respectively, truncated at +1.0 kcal/mol (1 cal = 4.184 J); E_{el} is the electrostatic potential with a distance-dependant dielectric constant $\varepsilon = 4r$;[32] and E_{hb} is the hydrogen-bonding potential. The additional solvation term was split into three components: E_{polar} includes contributions only from polar atoms; E_{ar} and E_{al} include contributions from the aromatic and aliphatic atoms, respectively. Several parallel minimization trajectories of the "amoeba" simplex optimizer with different initial values converged to the following weighting factor values: $\alpha = 0.40$; $\beta = 2.43$; $\gamma = 2.30$; $\delta = 6.26$ and $\varepsilon = 1.27$. The ratio between the van der Waals and electrostatics weighting factors (ratio vdw/el = 0.40) remained similar to the ratio between the values used during simulations (ratio vdw/el = 0.46). The low weighting of the van der Waals term can be explained by (i) its higher intrinsic noise due to the high sensitivity of this short-distance term to small deviations in the structure; and/or (ii) a possible double-counting due to the fact that the solvation term (with parameters derived from octanol/water transfer energies) already takes into account the attractive component of the intermolecular van der Waals (this would be the case assuming well-packed interfaces; however, the van der Waals term is still needed to account for clashes in the rigid-body docking interfaces). The weight of the hydrogen bonding potential with respect to the electrostatics (ratio hb/el = 2.43) approximately doubled (weights used in simulations had the ratio hb/el = 1.17). This is as expected, since the desolvation of hydrogen bond donors/acceptors is now explicitly present and compensates, in part, for the (bigger) energy gain from the hydrogen bond formation. Interestingly, the factor for the aromatic solvation component is considerably larger than the rest of the solvation terms (see more in Discussion).

Figure 1 shows the energy/RMSD distributions of all docking poses for some of the complexes of the training set (1ca0, 2pcf and 1mlc), as re-evaluated with both the original energy function (equation (3) in Materials and Methods) and the optimized one (equation (1) with the parameters described above). The energy profiles become much more funnel-shaped (i.e. conformations closer to the real structure have significantly lower relative energy values) when the solutions are re-evaluated with the optimized energy function and new solvation terms.

In order to demonstrate the general applicability of the optimized parameters, we validated them on a test set (Table 1) formed by complexes whose structures (and those of their unbound subunits) were available to us only after we compiled the training set for benchmarking the docking method. The test set included: the recently solved complexes FNR/ferredoxin and erythropoietin (EPO)/receptor (EPOR), the seven targets of the CAPRI experiment†, and the unbound–unbound test cases of the benchmark compiled by Chen *et al.*,[33] none of which was used in our training set. As an example, the effect of the new energy function on the energy/RMSD profile for the docking of EPO and EPOR is shown in Figure 1. As can be seen, the relative energy values of the conformations in the vicinity of the native complex structure also improved with the new energy function. The energy/RMSD profiles of the rest of complexes in the test set (Table 1) became also more funnel-like (not shown), which indicates that the optimized parameters are applicable to complexes outside the training set.

Distribution of docking poses in the training set and in the test set

The energy/RMSD profile provides only a 1-D representation of the 6-D rigid-body docking energy landscape of the complex and requires prior knowledge of the native structure. Visualization of the spatial distribution of the docked poses around the receptor can result in a more objective and detailed picture. In Figure 2 we plot the positions of the center of mass of the ligand in all of the rigid-body docking poses obtained for a representation of complexes of the training set: a protease–inhibitor (chymotrypsin/APPI; PDB 1ca0); an electron-transfer complex (cytochrome *f*/plastocyanin; PDB 2pcf); and an antibody–antigen (Fab/lysozyme; PDB 1mlc). In the same Figure is shown a representation of complexes of the test set: a protease-inhibitor (ribonuclease inhibitor/ribonuclease A; PDB 1dfj); an electron-transfer complex (FNR/ferredoxin; PDB 1ewy); and an antibody–antigen (IgG1 E8 Fab/cytochrome *c*; PDB 1wej). The dots have been colored from red to blue, according to the energy values of the

function, as defined by equation (1). RMSD is calculated for the ligand interface C^{α} atoms (defined as those with at least one atom at <4 Å from any receptor atom) with respect to the native complex structure. The average energy value for the docking poses within a sliding 5 Å RMSD range is represented as a continuous line. The complexes chymotrypsin/APPI (PDB 1ca0); cytochrome *f*/plastocyanin (PDB 2pcf); and Fab/lysozyme (PDB 1mlc) were included in the initial training set used to derive the optimized parameters. An example of a complex not included in the initial training set, EPO/EPObp (PDB 1eer), is shown.

† http://capri.ebi.ac.uk

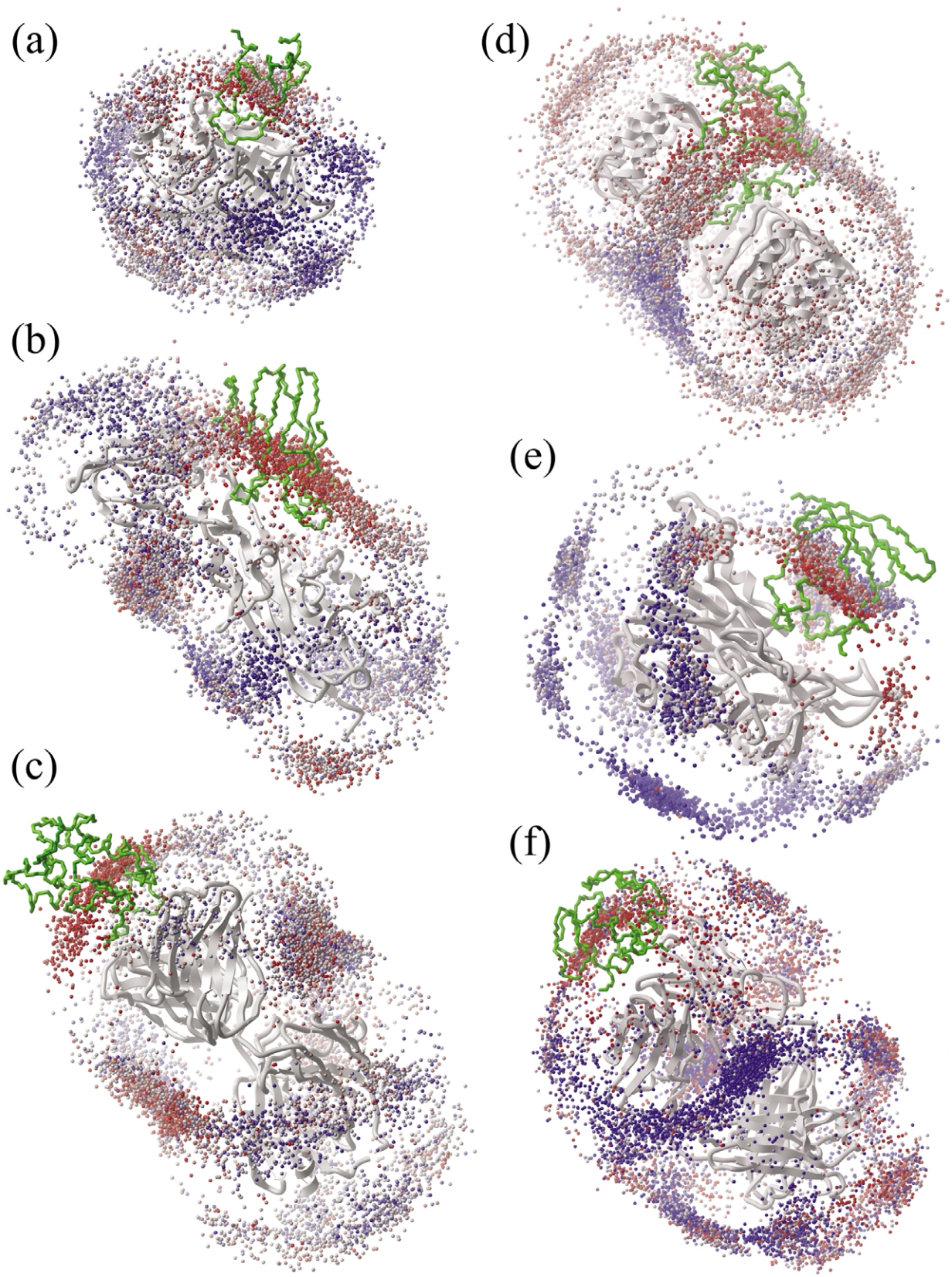

Figure 2. Representation of the docking poses obtained for complexes of the training set: (a) chymotrypsin/APPI; (b) cytochrome *f*/plastocyanin; and (c) Fab/lysozyme; and of the test set: (d) ribonuclease inhibitor/ribonuclease A; (e) FNR/ferredoxin; and (f) IgG1 E8 Fab/cytochrome *c*. The native complex structures (PDB codes: 1ca0; 2pcf; 1mlc; 1dfj; 1ewy; 1wej, respectively) are represented (receptor in gray; ligand in green). The dots represent the center of mass of every ligand pose after the rigid-body simulations. They are colored according to the energy value of each solution from red (lowest energy) to blue (highest energy).

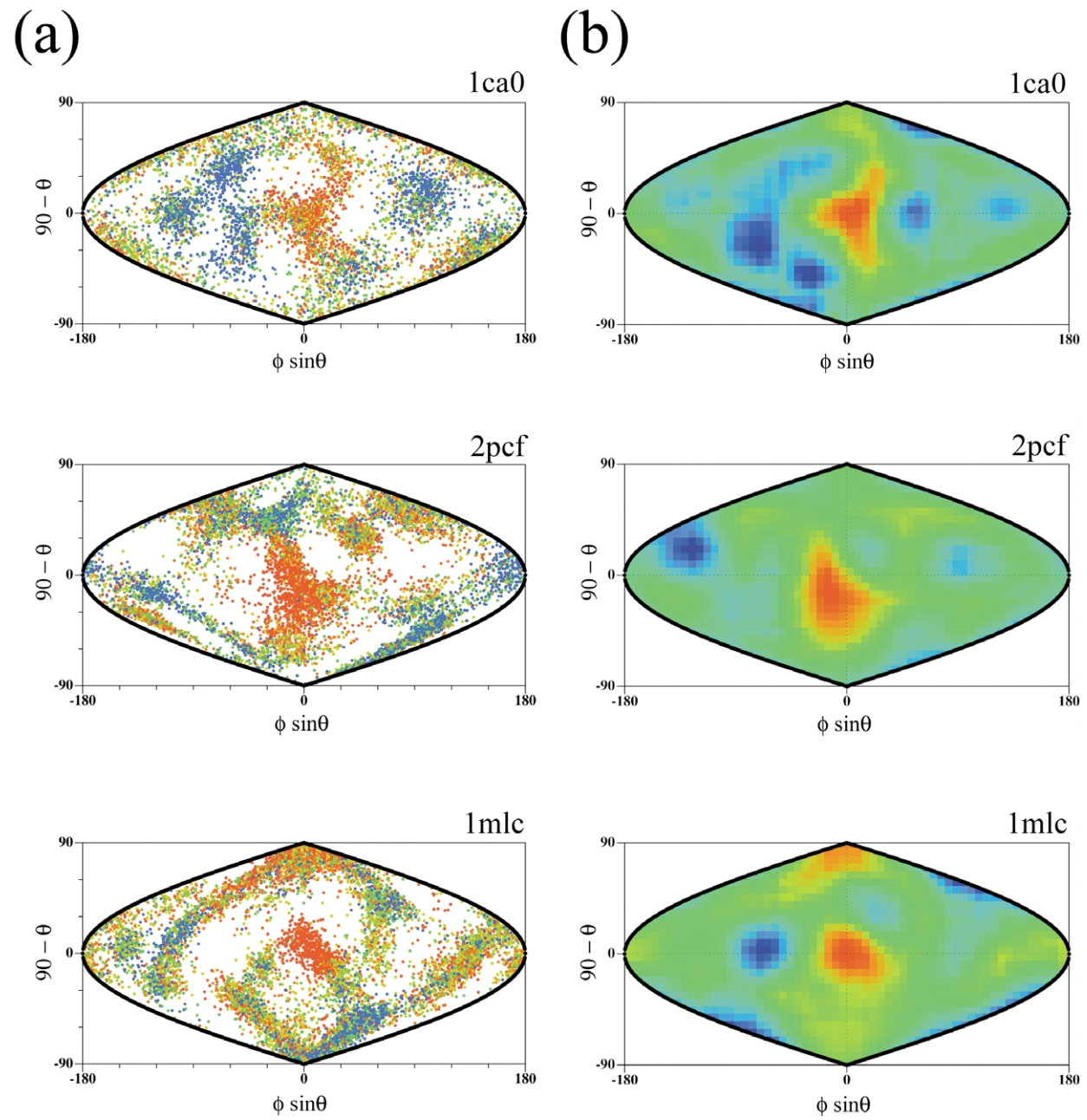

Figure 3. (a) Representation of the rigid-body docking poses for the complexes chymotrypsin/APPI (PDB 1ca0); cytochrome *f*/plastocyanin (PDB 2pcf); and Fab/lysozyme (PDB 1mlc). The position of the center of mass of every ligand docking pose is represented in spherical coordinates as a 2-D sinusoidal equal-area projection, and colored (from red to blue) according to the energy values (low energies in red; high energies in blue). (b) The same docking energy landscapes, averaged as described in Materials and Methods, and colored according to the same criteria.

corresponding pose (from lower to higher energy values, respectively). An accumulation of low-energy solutions around the known position of the ligand in the complex can be observed, both in complexes in the training set and in the test set.

The analysis of these distributions can be further facilitated using a spherical coordinates projection of the positions of the ligand's center of mass (see Materials and Methods). These projected docking landscapes (shown in Figure 3(a) for selected complexes) indicate clearly the accumulation of low-energy solutions around the known binding site. Zones with high density of low-energy solutions are shown in red, while regions with high-energy solutions or no solutions at all are shown in blue. Averaging neighboring regions removes some of the noise and shows more clearly the populated areas of both low and high-energy regions (red and blue zones, respectively), as can be seen in Figure 3(b). In this analysis, the receptor surface is considered spherical (thus overlooking all minor geometrical details). In some cases, this might be an oversimplification, but it provides a simple visualization of the distribution of docked poses on the surface of the receptor, which can be useful to compare docking landscapes of different complexes.

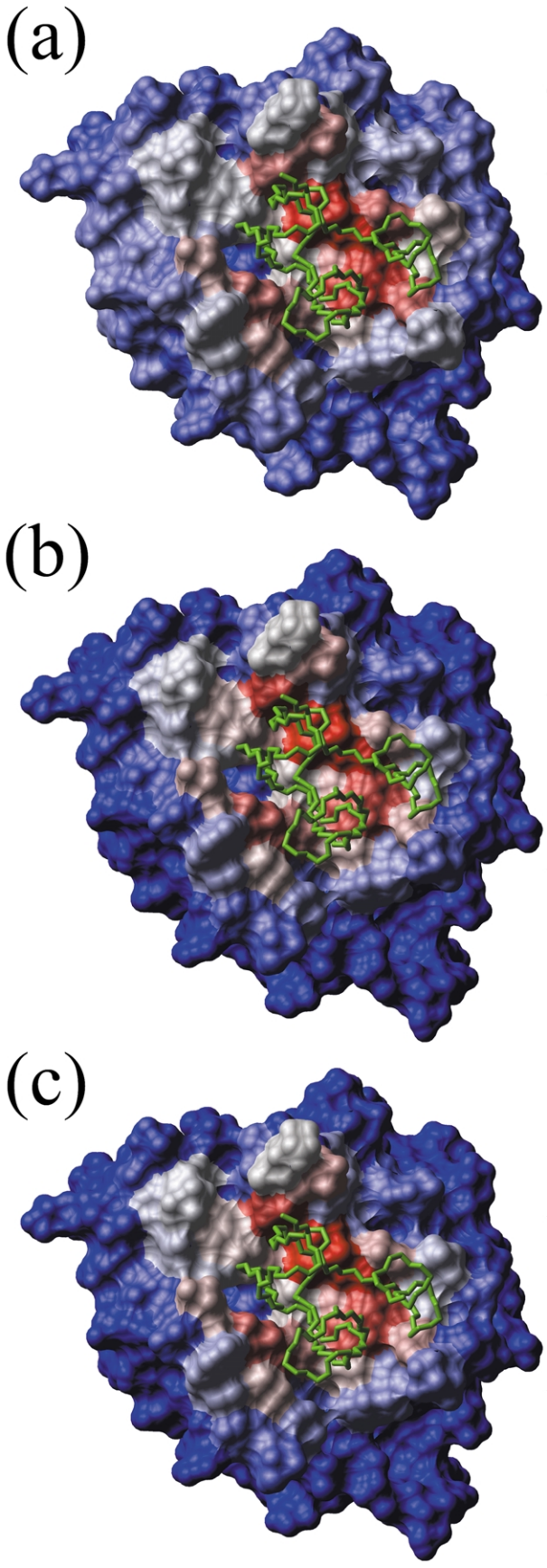

Figure 4. Mapping of the averaged relative buried surface per residue on the receptor molecular surface for the docking of chymotrypsin (PDB 5cha) and APPI (PDB 1aap). The Connolly surface of the receptor is represented, and residues are colored according to their ABS values (residues in red/blue have highest/lowest values, respectively). The position corresponding to the experimental complex structure is marked in green. (a) Plot obtained from all 5958 docking poses, after weighting according to a Boltzmann distribution (equation (4) in Materials and Methods). (b) Plot obtained from the 100 lowest-energy docking poses with the same energy weighting (equation (4) with $N = 100$). (c) Plot obtained from the 100 lowest-energy docking poses with no energy weighting, as defined by equation (5).

Mapping residue interface propensities onto the protein surface

We next proceeded to investigate how the observed trends in docking pose distributions and energy landscapes could be translated into the propensities of the individual receptor surface residues to be buried in the putative complexed structures. For each receptor–surface residue, the relative surface buried upon ligand binding was calculated, and this value was averaged for all docking poses according to a Boltzmann distribution at the elevated simulation temperature, as defined by equation (4). This energy-based weighting system increases the influence of the low-energy solutions in the final average relative buried surface values (for more details, see Materials and Methods). In Figure 4(a) the per-residue segments of the molecular surface of chymotrypsin are colored from red to blue according to the energy-averaged buried surface values calculated using the complete set of docking solutions ($N = 5958$) for the chymotrypsin/APPI complex. This representation reflects the zones of the receptor surface where the lowest-energy solutions accumulate (in red). High-energy solutions have little influence on the final values, so if we compute only the relative buried surface values per residue for the 100 lowest-energy solutions (instead of the total number of 5958), the results are very similar, as can be seen in Figure 4(b). Moreover, if we calculate the same values using the 100 lowest-energy solutions without any energy weighting, as defined by equation (5), the results are practically identical, as can be seen in Figure 4(c). This averaged buried surface (ABS) value is easier to compute and reflects the number of low-energy solutions in which a particular residue is involved in the interface: a value of 0 for a particular residue would mean that it is not found in the interface in any of the 100 lowest-energy solutions, whereas a value of 1 would mean that it is buried completely upon binding in all 100 lowest-energy docking solutions. Using these values and $<$ABS$>$ (estimator for the expected ABS value from a random distribution of the docking interfaces), we can calculate the normalized interface propensity (NIP) per residue (see Materials and Methods). A normalized propensity of 0 will indicate that the residue is found

in the docking interfaces as frequently as expected, while a value of 1 (maximum expected value) would indicate that it is found in absolutely all the docking interfaces. Negative values would indicate that the residue is found in the docking interfaces less frequently than expected. Figure 5 shows the interface propensity maps calculated for complexes from both the training and test sets. The accumulation of low-energy solutions in the vicinity of the known binding sites indicates the existence of preferred binding areas for protein–protein docking. We will now compare these interface propensity maps with the experimental interfaces and explore their potential to predict protein–protein interaction sites.

Interface propensity maps from the docking poses: prediction of binding sites

The observed distribution of the docking interface propensities suggests that they can be used to predict which residues are likely to be involved in the interface. We initially used a loose NIP cutoff of 0.0 to define the theoretical binding residues (i.e. those found in the docking interfaces more frequently than expected). In order to compare this theoretical binding site with the experimental one, we can compute the coverage, i.e. percentage of the experimental binding site that is predicted correctly:

$$\text{COVERAGE} = 100\,\frac{N_r^{EXP} \cap N_r^{PRED}}{N_r^{EXP}} \qquad (2)$$

where N_r^{EXP} and N_r^{PRED} are the number of residues in the experimental and predicted binding site, respectively.

The statistical significance of the predicted coverage values was evaluated by a χ^2 test. Contingency tables were built for each prediction, for which a P significance (probability that the prediction is random) was computed. As an example, in Table 2 is shown the contingency table for the predicted receptor and ligand interfaces of the complex chymotrypsin/APPI (PDB 1ca0) compared to the experimental ones.

In Table 3 are listed the values of the percentage of the experimental interface covered by the predicted one in the complexes of the training and test sets (for both receptor and ligand molecules), as well as their confidence values. As can be seen, the predicted interfaces differ from a random distribution (confidence $>95\%$) in the majority of the molecules both in the training set (69%) and in the test set (81%). If we consider only the significant predictions, the average percentage of the experimental interface covered by the predicted one in the complexes of the training set is 86% (89% for the receptor interfaces; 81% for the ligand interfaces). The results for the test set are very similar: the average coverage is 80% (83% for the receptor interfaces; 78% for the ligand interfaces).

Table 2. χ^2 test for the comparison between predicted and experimental interfaces in the interaction chymotrypsin/APPI

	In experimental interface	Not in experimental interface
Receptor		
In predicted interface	21 (5.3)	21 (36.7)
Not in predicted interface	5 (20.7)	158 (125.6)
$\chi^2 = 66.4$; $P < 10^{-4}$		
Ligand		
In predicted interface	8 (5.8)	14 (16.2)
Not in predicted interface	6 (8.2)	25 (22.8)
$\chi^2 = 1.9$; $P = 0.17$		

The χ^2 test is based on the number of residues inside/outside the predicted and experimental interfaces (expected occurrences for a hypothetically random distribution in parentheses).

High-propensity binding patches in the training set

The predicted binding sites listed in Table 3 are formed by residues found in the docking interfaces more frequently than expected. This yields, in general, theoretical interaction sites larger than the experimental ones. Although we found a significant correlation between the predicted and native interfaces in most of the cases, some concern may arise about the practical application of these large predicted binding sites. Thus, it may be beneficial to reduce their size by selecting residues with stronger interface propensities, that is, by increasing the NIP cutoff value used to define the theoretical binding sites. As can be seen in Figure 6(a), when the *NIP* cutoff value was increased, the accuracy (i.e. percentage of correctly predicted residues) of the theoretical binding sites in the training set improved considerably, although it increased the number of molecules with no detectable high-propensity residues. As can be seen in Table 4, at NIP cutoff of 0.4 the average accuracy of the high-propensity binding patches (found in 65% of the molecules) is about 85%. Higher cutoff values could be imposed to try to find residues even more likely to be in the native interface; e.g. at NIP cutoff of 0.7 the high-propensity binding patches, found in 19% of the molecules, are all correct (results not shown).

High-propensity binding patches in the test set

In order to test the general applicability of the method, we have analyzed the existence of high-propensity binding patches in an independent test set formed by complexes not used in the training set (Table 1; for more details about the test set selection, see Materials and Methods).

As can be seen in Figure 6(b), when the NIP cutoff value was increased, the accuracy (i.e. percentage of correctly predicted residues) of the theoretical binding sites in the test set improved

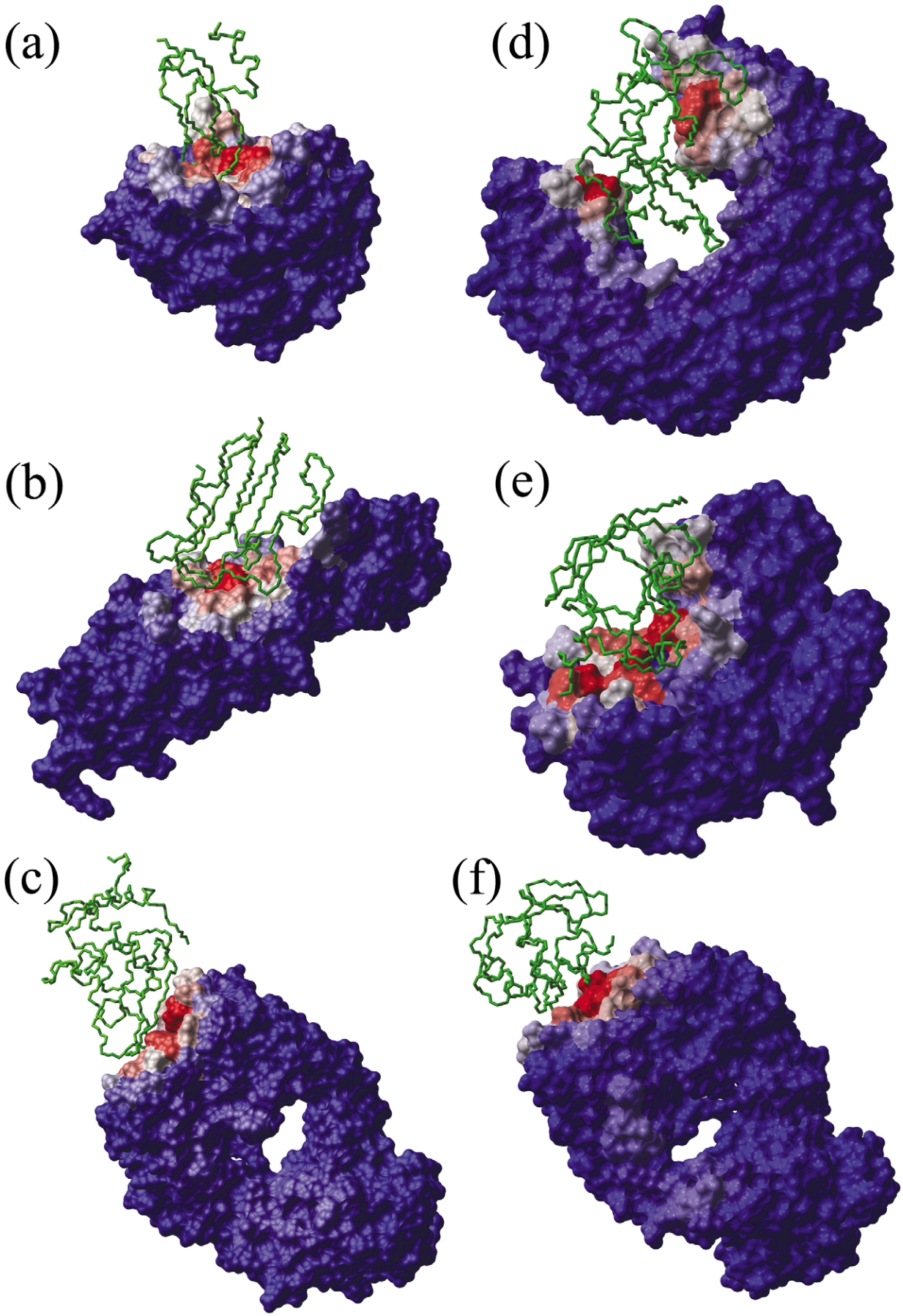

Figure 5 (*legend opposite*)

Table 3. Comparison between the predicted and the experimental interfaces

	Receptor			Ligand		
Complex[a]	Coverage[b]	Size[c]	P[d]	Coverage[b]	Size[c]	P[d]
Parameterization set						
1ca0	81	1.6	$<10^{-4}$	57	1.6	0.17
1cbw	92	1.4	$<10^{-4}$	64	1.9	0.16
1acb	77	1.4	$<10^{-4}$	41	1.9	0.44
1cho	89	1.4	$<10^{-4}$	77	1.6	<0.05
1cgi	92	1.3	$<10^{-4}$	70	1.2	<0.05
2kai	54	1.3	$<10^{-4}$	79	2.2	0.06
2sni	93	2.6	$<10^{-4}$	53	1.4	0.09
2sic	83	2.8	$<10^{-4}$	46	2.5	0.27
1cse	96	2.9	$<10^{-4}$	41	1.8	0.53
2tec	82	2.4	$<10^{-4}$	45	1.5	0.90
1taw	83	1.6	$<10^{-4}$	62	1.8	0.13
2ptc	89	1.7	$<10^{-4}$	36	2.0	0.16
3tgi	29	1.8	0.78	64	2.1	0.44
1brc	80	1.8	$<10^{-4}$	62	1.7	0.09
1fss	100	4.2	$<10^{-4}$	83	1.2	$<10^{-4}$
1bvn	90	2.0	$<10^{-4}$	86	1.4	$<10^{-4}$
1bgs	95	1.3	$<10^{-4}$	94	1.3	$<10^{-4}$
1ay7	100	1.4	$<10^{-4}$	94	1.8	$<10^{-4}$
TEM1[e]	59	2.3	0.07	79	1.2	$<10^{-4}$
1ugh	97	2.2	$<10^{-4}$	75	0.9	$<10^{-4}$
2pcb	92	3.8	$<10^{-4}$	73	2.5	<0.05
2pcf	92	1.6	$<10^{-4}$	73	1.1	$<10^{-4}$
1mlc	100	3.2	$<10^{-4}$	29	3.2	0.11
1vfb	100	1.9	$<10^{-4}$	83	3.1	<0.05
Test set						
1ewy	100	1.9	$<10^{-4}$	68	1.2	$<10^{-4}$
1eer	91	2.6	$<10^{-4}$	75	1.4	$<10^{-4}$
1kkl	11	3.1	0.37	81	1.2	$<10^{-4}$
VP6FAB[f]	91	4.5	$<10^{-4}$	100	4.9	$<10^{-4}$
1ken	97	3.2	$<10^{-4}$	100	3.4	$<10^{-4}$
1kxv	10	2.7	0.19	83	1.9	$<10^{-4}$
1kxt	55	3.9	<0.05	96	1.8	$<10^{-4}$
1kxq	100	1.9	$<10^{-4}$	96	1.7	$<10^{-4}$
1l0x	0	2.8	<0.05	100	6.2	$<10^{-4}$
1avw	100	1.7	$<10^{-4}$	94	2.6	$<10^{-4}$
1dfj	89	1.8	$<10^{-4}$	80	1.2	$<10^{-4}$
1tgs	93	1.5	$<10^{-4}$	82	1.2	$<10^{-4}$
1ahw	89	5.7	$<10^{-4}$	0	1.9	$<10^{-4}$
1dqj	100	5.0	$<10^{-4}$	39	2.7	0.68
1wej	100	4.1	$<10^{-4}$	44	2.4	0.77
1avz	42	2.6	0.49	92	1.7	$<10^{-4}$
1wq1	33	2.9	0.35	53	1.3	<0.05
2mta	93	3.1	$<10^{-4}$	100	1.7	$<10^{-4}$
1bth	61	2.2	$<10^{-4}$	39	1.2	0.84
1fin	68	2.4	$<10^{-4}$	100	1.5	$<10^{-4}$
1fq1	32	3.9	0.84	0	2.1	$<10^{-4}$

Predicted interfaces were defined as those surface residues with NIP > 0.0 (equation (7) in Materials and Methods).

[a] The PDB code of the complex is shown here for clarity, but the unbound subunits were used during docking simulations (see Table 1).

[b] Percentage of residues in the experimental interface predicted correctly.

[c] Relative size of the predicted interface with respect to the experimental one (i.e. total number of residues in the predicted interface divided by the total number of residues in the experimental interface).

[d] P-significance values obtained from a χ^2 test (see Materials and Methods).

[e] Complex TEM-1 β-lactamase/BLIP; see Table 1 for coordinates.

[f] Complex VP6/Fab; see Table 1 for coordinates.

Figure 5. Maps of the residue interface propensities for some receptor molecules. The 100 lowest-energy docking solutions were used to generate the maps. The color-coding is the same as that in Figure 4. Complexes of the training set: (a) chymotrypsin (PDB 5cha)/APPI (PDB 1aap); (b) cytochrome *f* (PDB 1ctm)/plastocyanin (PDB 1ag6); (c) subtilisin (PDB 2st1)/SSI (PDB 3ssi). Complexes of the test set: (d) ribonuclease inhibitor/ribonuclease A (PDB 1dfj); (e) FNR/ferredoxin (PDB 1ewy); and (f) IgG1 E8 Fab/cytochrome *c* (PDB 1wej).

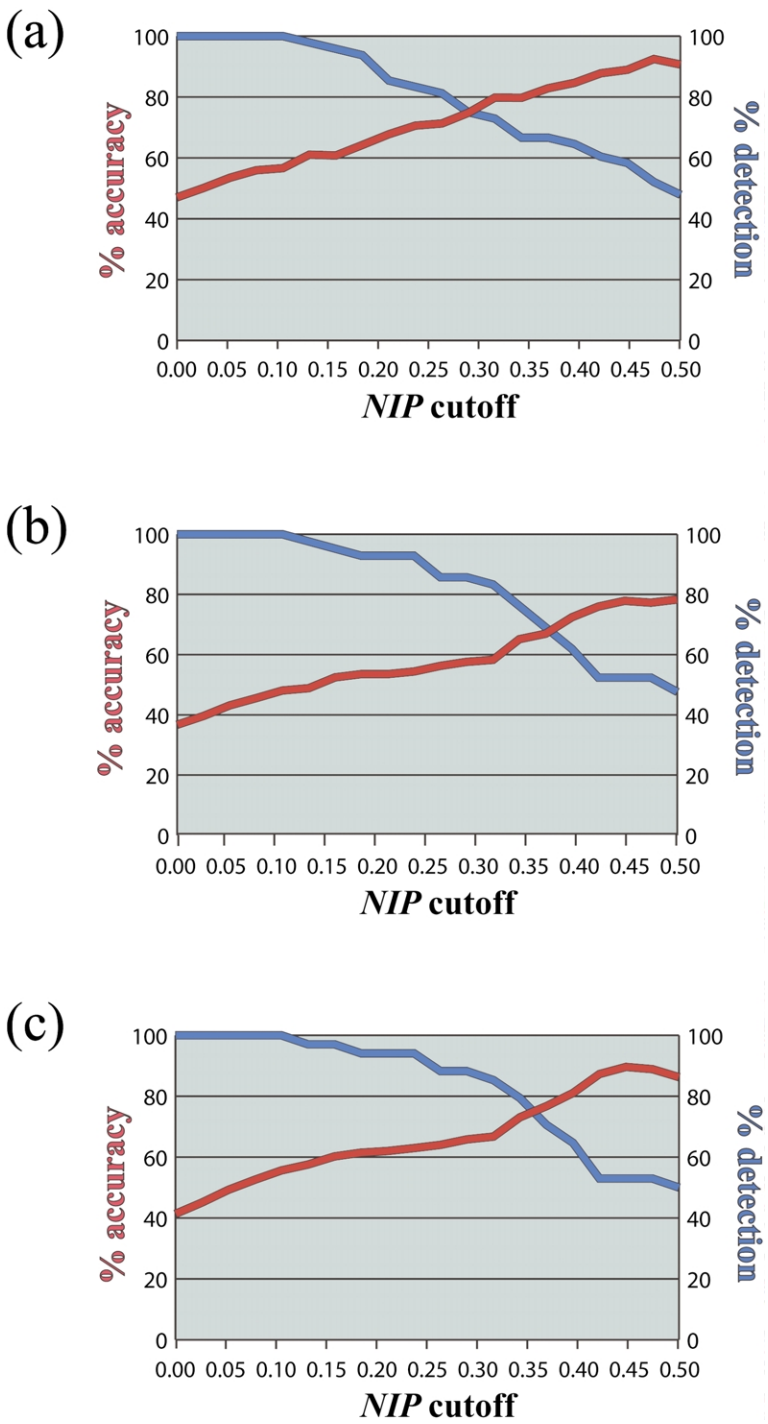

Figure 6. Comparison between the theoretical and the experimental binding sites in (a) the training set; (b) the test set; and (c) the same test set after excluding antigens. The average accuracy (percentage of predicted residues located correctly in the native interface; red line) increases proportionately to the NIP cutoff value that is used to define the theoretical binding sites. The blue line shows the percentage of cases in which a predicted binding patch (formed by at least one residue with NIP > cutoff) is detected.

considerably in a fashion similar to that in the training set (expectedly, it increased the number of molecules with no detectable high-propensity residues). As can be seen in Table 4, at NIP cutoff of 0.4 the average accuracy of the high-propensity binding patches (found in 62% of the molecules) is 73%. If antigen molecules are removed from the analysis (see Discussion), the average accuracy of the high-propensity binding patches (NIP cutoff = 0.4) increases to 81% (Figure 6(c)). The results obtained in the test set are thus comparable to the values obtained in the parameterization set.

Docking non-native ligands

So far, we have generated interface propensity maps starting from the unbound receptor and ligand subunit coordinates of known protein–protein complexes. For test purposes, we have explored the possibility of docking a non-native ligand. For that, we performed docking simulations using the 3-D coordinates of unbound receptors (chymotrypsin, PDB 5cha; cytochrome *f*, PDB 1ctm; UDG, PDB 1akz) and a non-native ligand such as lysozyme (PDB 1lza) instead of their native ligands (APPI, plastocyanin and UGI,

Table 4. High-propensity binding patches

Protein[a]	Number of residues in patch	Patch accuracy[b]
Parameterization set		
1ca0 (receptor)	12	83
1ca0 (ligand)	2	100
1cbw (receptor)	14	93
1acb (receptor)	14	86
1cho (receptor)	11	91
1cho (ligand)	2	100
1cgi (receptor)	15	100
1cgi (ligand)	1	100
2kai (receptor)	7	57
2sni (ligand)	9	44
2sic (ligand)	2	0
1taw (receptor)	9	100
1taw (ligand)	2	100
2ptc (receptor)	11	91
3tgi (receptor)	3	0
1brc (receptor)	3	100
1brc (ligand)	1	100
1fss (ligand)	8	100
1bvn (receptor)	3	100
1bvn (ligand)	5	100
1bgs (receptor)	6	100
1bgs (ligand)	3	100
1ay7 (receptor)	8	100
TEM1 (ligand)[c]	20	95
1ugh (ligand)	10	100
2pcb (receptor)	9	56
2pcb (ligand)	7	29
2pcf (receptor)	10	100
2pcf (ligand)	3	100
Test set		
1ewy (receptor)	11	100
1ewy (ligand)	7	71
1eer (ligand)	6	83
1kkl (ligand)	4	100
VP6FAB (ligand)[d]	4	100
1ken (ligand)	11	100
1kxv (receptor)	11	0
1kxv (ligand)	8	75
1kxt (receptor)	9	0
1kxt (ligand)	9	89
1kxa (receptor)	13	100
1l0x (receptor)	13	0
1avw (receptor)	7	100
1avw (ligand)	7	86
1dfj (receptor)	10	90
1dfj (ligand)	3	100
1tgs (receptor)	6	100
1tgs (ligand)	4	100
1ahw (ligand)	2	0
1dqj (receptor)	1	100
1wej (receptor)	9	89
1avz (receptor)	2	0
1avz (ligand)	1	100
2mta (ligand)	6	100
1bth (ligand)	1	0
1fin (ligand)	9	100

All cases where a high-propensity binding patch was found (as defined by those surface residues with NIP > 0.4; equation (7) in Materials and Methods).

[a] The PDB code of the complex is shown here for clarity, but the prediction results apply to the unbound subunits listed in Table 1 (receptor or ligand as specified).

[b] Percentage of predicted residues located correctly in the experimental interface.

[c] Complex TEM-1 β-lactamase/BLIP; see Table 1 for coordinates.

[d] Complex VP6/Fab; see Table 1 for coordinates.

respectively). The results are quite interesting: the lysozyme docking poses accumulated around the experimental location of the native ligand in some of the receptor molecules. For instance, the predicted receptor interface derived from the chymotrypsin/lysozyme docking simulations (at NIP cutoff $= 0.0$) covered 73% ($P < 10^{-4}$) of the receptor experimental interface (as compared to the value of 81% obtained by using the native APPI ligand; Table 3). Furthermore, a high-propensity binding patch was found on the surface of chymotrypsin (14 residues; 79% accuracy). However, the predicted interface derived from the cytochrome f/lysozyme docking (at NIP cutoff $= 0.0$) did not show a significant correlation with the experimental interface, and the four residue high-propensity binding patch found had a low level of accuracy (25%). Lastly, the predicted interface derived from the UDG/lysozyme docking simulations (at NIP cutoff > 0.0) covered 52% ($P < 0.05$) of the receptor experimental interface (as compared to the value of 97% obtained by using the native UGI ligand; Table 3), but no high-propensity binding patch was found.

Discussion

Global rigid-body docking: exploring the whole protein surface

In this work we have explored the docking energy landscapes of known protein–protein complexes using ensembles of docked structures generated by a rigid-body pseudo-Brownian global energy optimization procedure. Previously[4] we described a two-step docking procedure that allowed sampling of approximately half of the surface of the receptor around the known binding site. We have extended the search to the whole receptor surface. No spatial or biological restrictions were used during simulations, which allowed a complete sampling of the docking landscape around each subunit. A side-chain refinement step was used in our previous work in order to re-evaluate the 400 lowest-energy interfaces in search for the near-native solution. However, the goal here is the analysis of all docking poses for the description of the complete docking landscape. A side-chain refinement of all the thousands of interfaces generated by the rigid-body protocol would have been computationally too expensive, so we have preferred to focus the analysis onto the rigid-body docking landscape.

Optimized energy function for protein–protein association

It was observed repeatedly that force-field energy as well as various solvation and electrostatic terms failed to discriminate consistently the near-native solution from false positives in docking simulations. A number of factors contribute to this

phenomenon: (i) each of the terms used in the energy functions can be estimated only with considerable uncertainty (e.g. the values of partial charges and internal dielectric constant, depths of the minima for van der Waals interactions, and per-atomic solvation parameters vary up to twofold and sometimes more between various estimates/parameterizations); (ii) van der Waals interactions and, to a lesser extent, electrostatics and other terms, are highly sensitive to the inaccuracies of the atomic coordinates, the situation being exacerbated by the lack of relaxation in the structures produced by the rigid-body docking simulations; and (iii) certain phenomena such as aromatic/aromatic interaction might be largely unaccounted for in the existing force-field parameters.

These problems may be alleviated by the appropriate weighting of the terms. Possible benefits of weighting are twofold: (i) it can correct for suboptimal choice of uncertain parameters such as internal dielectric constant, and (ii) it can downweight the terms that carry larger noise. While the latter may seem non-physical, it can be shown easily that the optimal weight of a term, which includes a noise component declines as $1/(1+\sigma_n^2/\sigma_s^2)$ with the increase of the amplitude of noise (σ_n^2 is the mean-square deviation of the noise component and σ_s^2 is that of the signal).

We thus introduced weighting coefficients for van der Waals, electrostatics and hydrogen bonding potentials. Furthermore, we have split the atomic ASA-based solvation term into three components: polar desolvation, aliphatic hydrophobicity and aromatic hydrophobicity. The coefficients for the different energy terms were optimized on a training set using as objective function the ranking of the near-native conformations for a set of 24 protein–protein complexes of known structure, docked from the unbound subunits. After the energy re-evaluation using the optimized parameters, the near-native solutions ranked below 100 in 70% of the cases of the training set (as compared to 33% with the old energy function). In order to exclude the possibility of over-fitting of the parameters on the training set, we explored the effects of the parameter optimization on the ranking of more remote solutions, as well as on the ranking of the near-native solution in a test set formed by complexes not included in the training set. We observed that although the new energy function was optimized to rank as best as possible only a single near-native solution per complex, the energy landscapes in the vicinity of the native structure adopted a more defined funnel-like shape (Figure 1). This suggests that the optimization procedure was not distorted by particular features/inaccuracies of the specific near-native solutions used. On the contrary, the new energy function seems to result in better behavior of the averaged energy for the ensemble of solutions near the native binding sites. Furthermore, we found that the new energy function also improved the docking landscapes of other complexes not included in the initial training set (Figure 1).

It is noteworthy that the optimal weighting factor for the aromatic solvation is considerably larger than those obtained for the rest of the solvation terms. It might indicate that the parameters for water/octanol transfer of aromatic atoms are not optimal to estimate the energy gain of burying aromatic atoms in the (often) non-aliphatic environment of the protein–protein interfaces. As an alternative explanation, the larger weighting value for the solvation of aromatic atoms could be compensating for some specific aromatic interactions underestimated or not considered at all in the rest of the energy function, such as the van der Waals attraction. Indeed, Lomize and co-workers reported recently that the solvation parameter for aromatic carbon should be increased several-fold when no explicit van der Waals interaction is included.[34]

Validation of the binding-site prediction on a test set of 21 complexes

We used 21 complexes not included in the initial parameterization set as an external validation set. Overall, the performance of the method on the test set was comparable to its performance on the training set.

The first example in the test set is a recently solved electron transfer complex formed by the proteins FNR and Fd. As can be seen in Figure 5(e), the predicted binding sites correlated extraordinarily well with the experimental interfaces, and we found highly accurate high-propensity binding patches in both FNR and Fd molecules (100% and 71% accuracy, respectively; see Table 4). These high-propensity binding areas detected in the surface of both molecules, as can be seen in Figure 7(a) and (b), may be explained by the existence of favorable forces that orient the molecules during the association process to allow an efficient electron transfer between the redox centers, as mutational studies suggest.[35] Another interesting example is the interaction between EPO and EPOR. The predicted receptor interaction site (NIP cutoff = 0.0) covered 91% ($P < 10^{-4}$) of the experimental one (see Table 2). Interestingly, we found a six residue high-propensity binding patch in the ligand, located correctly (83% accuracy) in one of the two experimental ligand interaction sites: indeed the one reported to have significantly higher affinity (Figure 7(c)).[36]

The following examples in the test set came from the CAPRI experiment. The first of the CAPRI targets was the HPr kinase/HPr complex, in which the ligand interface was predicted perfectly (Tables 3 and 4). However, no good prediction was found for the receptor interface, probably because of the large conformational movement of the C-terminal helix of the receptor HPr kinase between the unbound and the complexed states. The following

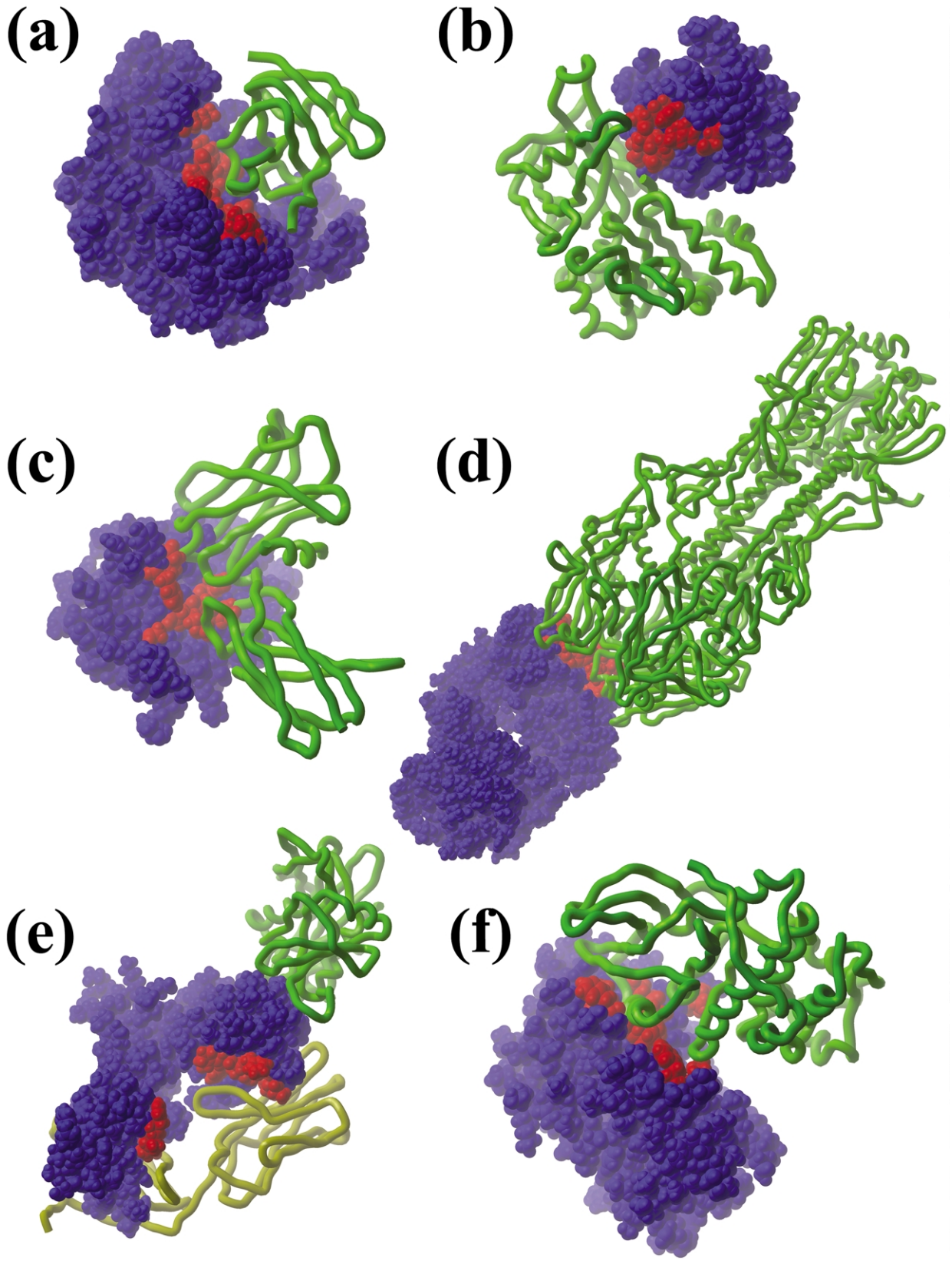

Figure 7. High-propensity binding sites found in different molecules of the test set: (a) FNR (bound ferredoxin is shown in green); (b) ferredoxin (bound FNR in green); (c) EPO (bound EPObp in green); (d) Fab (bound hemagglutinin in green); (e) TCR-β (bound speA in green; in yellow is the expected position of the TCR-α subunit as in PDB 1tcr); and (f) cyclin (bound CDK2 is shown in green). High-propensity binding residues (NIP > 0.4) are shown in red.

five CAPRI targets were antigen–antibody complexes, in which the structure of the bound antibody was provided in a random orientation. Interestingly, the antibody-binding interfaces were predicted excellently (Tables 3 and 4): high-propensity patches were found in four out of a total of five cases, with an average level of accuracy of 91%. For example, Figure 7(d) shows the interface prediction for the hemagglutinin/Fab complex, in which the 11 residue high-propensity patch found in the Fab ligand was located 100% correctly on the experimental interface. However, the predictions were, in general, poorer for the antigen molecules (see the next sub-section). In the last CAPRI target, the TCR-β/speA complex, a high-propensity patch was found in the receptor, although in a location different from the experimental binding site. Interestingly, as can be seen in Figure 7(e), this predicted binding site was actually located in the previously described interface between TCR-β and TCR-α (PDB 1tcr). In the absence of TCR-α during simulations, it seems that low-energy docking poses accumulated in the area that TCR-β uses for the binding to TCR-α, a "protein-phylic" patch that the method thus recognized.

The remaining complexes in the test set were taken from the benchmark independently built by Chen *et al.*[33] After removing those cases already included in our training set, we had a set of 12 complexes divided in four categories (enzyme-inhibitor, antibody–antigen, others, and difficult cases) according to the definition in the original benchmark. It is interesting to compare the results obtained for the different categories. In the three complexes classified as enzyme-inhibitor (PDB codes 1avw, 1dfj, 1tgs), the results were perfect: high-propensity binding patches were found in all of them (both in receptor and ligand molecules) with 96% of the predicted residues located correctly at the interfaces. It is interesting that even in a complex such as the ribonuclease A/inhibitor (PDB 1dfj), whose topology is completely different from that of the standard protease-inhibitor, the predictive results were excellent, as can be seen in Figure 5(d). As for the three antibody–antigen complexes (PDB codes 1ahw, 1dqj, 1wej), the results were similar to those obtained in the rest of the antibody–antigen cases: good prediction of binding sites in the antibody surfaces, but no correct predicted sites in the antigen surfaces (see the next sub-section). In the following three complexes, classified in the original benchmark as others (PDB codes 1avz, 1wq1, 2mta), we found high-propensity binding patches in three of the six molecules involved, but one of them (HIV-1 NEF) yielded incorrect predictions (0% accuracy). This is perhaps a non-adequate case for a prediction test, as the structure of the unbound receptor (PDB 1avv) is lacking three residues that are essential for binding: thus, it is not surprising that the correct binding patch is not found on this incomplete protein structure. In the last three complexes, described as difficult cases (PDB codes: 1bth, 1fin, 1fq1), the results were slightly worse, as expected: high-propensity binding patches were found in only two of the six molecules involved, and one of them (pancreatic trypsin inhibitor) was not located correctly in the interface. All of these difficult cases either have major backbone movement between the unbound and bound conformations, or are lacking an important number of residues in the structure of the unbound subunits. These challenging conditions are pushing the rigid-body docking approach to the limit. Nevertheless, it is remarkable that, in spite of these difficulties, we can find a high-propensity binding patch in cyclin located correctly in the interface with CDK2, as can be seen in Figure 7(f).

Poor prediction of antibody-binding sites on antigen surfaces

Relatively poor performance of the NIP as a predictor of the protein binding patch for the antigens is perhaps not surprising, since the antibody–antigen system represents a rather special case of protein–protein interaction, where only one partner (the antibody) has actively evolved for optimal binding, while the other partner (the antigen) passively provides relatively arbitrary interface. For instance, among the CAPRI targets there were three different α-amylase/VHH complexes (PDB codes 1kxv, 1kxt, and 1kxa), where the VHH antibody bound to different areas of the α-amylase surface in each complex. Interestingly, docking simulations found an identical high-propensity α-amylase binding site in all these three cases. This predicted site correlates very well (89% accuracy) with the experimental interface of the α-amylase/VHH Amd9 complex, indeed the one reported to have the strongest affinity amongst the three complexes.[37] This brings up the interesting question of whether these antibody molecules preferably bind zones of the antigen surface with higher "protein-phylicity". A systematic analysis of high-propensity patches in antigen surfaces, beyond the scope of this work, could help to understand immunogenicity properties, which would be extremely useful for antibody design.

Docking a non-native ligand to predict binding interfaces of a receptor

The surprising results obtained when using a non-natural ligand during the docking simulations indicate that, at least for some complexes, the preferred binding areas derived from the interface propensity maps may not depend on the ligand molecule used, and might suggest the existence of non-specific binding areas in the protein surfaces. This would be expected for proteases, which are functionally required to interact with a variety of polypeptides. It would be highly interesting to analyze if this is a general property of protein–protein interfaces or if, on the contrary, this is a phenomenon associated only with certain protein–protein

binding mechanisms. A complete analysis of the effect of using different probes on the docking landscape is beyond the scope of this work, and will be studied in a forthcoming publication.

Conclusions

The method described here can be applied immediately to suggest "hot spots" in the subunits of the complexes for experimental testing by mutagenesis. The identification of protein–protein interaction sites on a protein surface can be used to restrict the conformational search during the docking simulations. This would allow the selection of a sub-set of rigid-body docking solutions for further refinement when no experimental information about the binding site is available. A further development would be the introduction of genomics information (conservation of residues in the interfaces), which could lead to more accurate predictions. This methodology could be applied to large datasets of individual structures generated by the structural genomics efforts or by homology modeling methods. Prediction of putative protein interfaces can provide additional evidence for candidate protein pairs detected by other techniques, such as the Rosetta Stone method,[38] which could be an important step in building the map of interactions of a given organism at the atomic level.

In conclusion, we have described a new efficient protocol for sampling the complete protein surfaces in protein–protein docking. The ensembles of rigid-body docking solutions for a training set of 24 protein–protein complex structures were used to derive an optimized energy function for protein–protein association. We found that low-energy docking solutions accumulated in the vicinity of the experimental protein–protein interfaces, and devised a graphic way of visualizing the putative binding hot spots by averaging the interfaces of the 100 lowest-energy docking poses. The interface propensity maps generated from the docking poses showed excellent correlation with the experimental protein-binding sites in the training set, and these results were further confirmed on an independent test set of 21 protein–protein complexes. The work presented here indicates the existence of energetically favorable binding areas on the surfaces of proteins, and provides a useful tool to identify protein-binding sites on individual structures.

Materials and Methods

Rigid-body docking simulations: energy function

The rigid-body docking procedure used here is a variation of a previously described method, benchmarked on a set of experimental protein–protein complex structures.[4] Global energy optimization was performed by rigid-body sampling of the orientations of the ligand molecule with respect to the receptor molecule (whose position is fixed). The interaction energy was represented by five types of grid potentials,[39,40] pre-calculated on a grid as described elsewhere:[4,41] the van der Waals potentials for a hydrogen atom probe (E_{Hvw}) and for a heavy-atom probe (E_{Cvw}); an electrostatic potential where the partial atomic charges of the receptor were corrected by the induced surface solvent charge density to account for the solvent screening effect on the intermolecular pairwise electrostatic interactions ($E_{\mathrm{el}}^{\mathrm{solv}}$);[30] the hydrogen-bonding potential (E_{hb}); and a simple hydrophobic potential roughly proportional to the number of hydrophobic atoms of the ligand in contact with the hydrophobic surface of the receptor (E_{hp}). Energy balance between the different terms was optimized previously on a training set of known protein–protein complexes.[4] The solvation energy (E_{solv}) was calculated using an atomic solvent-accessible surface (ASA)-based model[42,43] with per-atomic parameters derived[44] from experimental vapor–water transfer energies for side-chain analogues[45] and additional experimental data for charged solutes.[46] This solvation term was added to the total energy to re-evaluate the docking solutions as described:[4]

$$E = E_{\mathrm{Hvw}} + E_{\mathrm{Cvw}} + 2.16E_{\mathrm{el}}^{\mathrm{solv}} + 2.53E_{\mathrm{hb}} + 4.35E_{\mathrm{hp}} + 0.20E_{\mathrm{solv}} \quad (3)$$

The grid potentials generated from the receptor were calculated in a box covering all the receptor atoms plus a margin of 10 Å (Figure 8). An extended box that depended on the size of the ligand was defined. In order to maintain the ligand molecule in the vicinity of the receptor during simulations, an energy penalty was applied outside this extended box. If a ligand atom is inside the map box, its energy is given by the grid potentials. For all ligand atoms inside the extended box, but outside the grid potential box, the energy is zero. For each ligand atom outside the extended box, a penalty of +1 kcal/mol is added to the energy.

Rigid-body docking simulations: conformational sampling procedure

In order to improve conformational sampling, multiple initial ligand positions were generated around the receptor, by defining a series of points ($i_{\mathrm{Rec}} = 1, 2, \ldots, N_{\mathrm{rec}}^{\mathrm{sf}}$) with an average distance of 15 Å between them, as can be seen in Figure 9(a). These points were generated systematically on the receptor solvent-accessible surface (expanded by 3 Å to overcome minor structural details) by dividing the surface into triangles with average side 15 Å. This distribution of points around the molecule is insensitive to the atomic details of the molecular surface, but reflects the overall shape of the molecule (including sizeable concavities and protuberances), and is more realistic than considering the molecule as a sphere or polyhedron as in our previous work. A similar set of points ($i_{\mathrm{Lig}} = 1, 2, \ldots, N_{\mathrm{lig}}^{\mathrm{sf}}$) was generated around the ligand molecule. Imaginary axes perpendicular to the molecule surface were defined from each receptor and ligand point. For each receptor surface point (i_{Rec}), various ligand orientations "aiming" at this point were generated by superimposing every ligand surface point (i_{Lig}) onto the receptor point and aligning their respective axes, as can be seen in Figure 9(b). Finally, for each of these orientations, six 60° rotations ($i_{\mathrm{Rot}} = 1, \ldots, 6$) were defined around the axis perpendicular

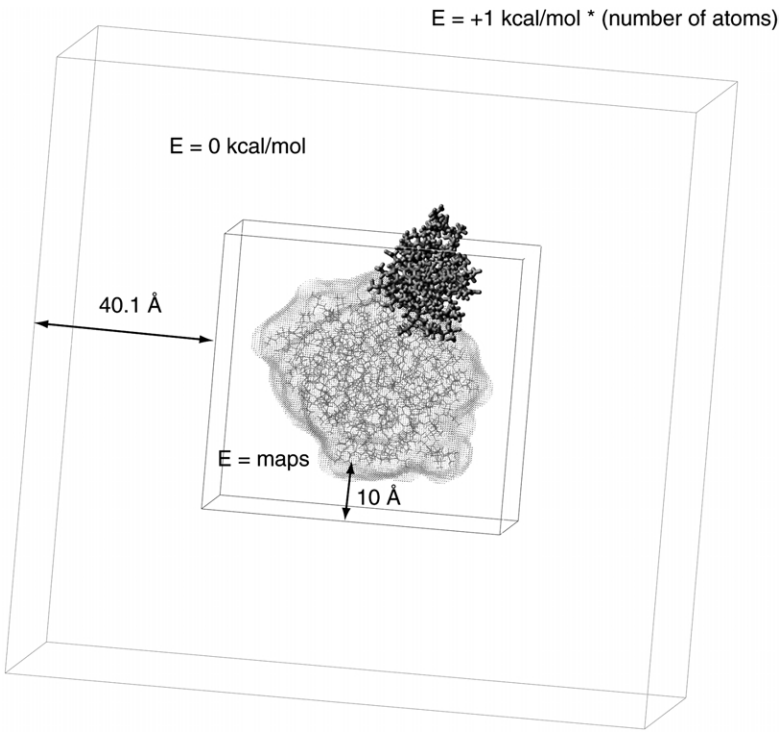

Figure 8. The grid box where the receptor potentials are pre-calculated (small cube) is shown around chymotrypsin, with the van der Waals potential for a heavy-atom probe represented inside. A larger box (defined by extending the grid box by the ligand size: 40.1 Å) where the ligand is allowed to roam freely is shown (large cube). The ligand molecule is shown in an arbitrary position. The energy of the ligand atoms inside the grid box is given by the grid potentials. The energy of the ligand atoms inside the extended box, but outside the grid box, is zero. The penalty energy for each ligand atom outside the extended box is +1.0 kcal/mol.

to both receptor and ligand surfaces. This operation was repeated for all receptor surface points in order to generate the total number of ligand starting conformations ($6N_{lig}^{sf}N_{rec}^{sf}$; which ranged from 1092 in the ribonuclease Sa/barstar complex to 4488 in the Fab F44.1/lysozyme complex).

For each ligand starting position $\{i_{Rec}, i_{Lig}, i_{Rot}\}$, sampling of rotational and translational space of the rigid-body ligand around the receptor molecule was performed by a pseudo-Brownian Monte-Carlo minimization procedure[47] implemented in the MolSoft ICM 2.8 program.[48] The minimization procedure consisted of a series of Monte-Carlo random moves of the position of the ligand, each one followed by up to 200 steps of the local conjugate gradient minimization. The simulation terminated after 20,000 energy evaluations. New conformations generated after each random move and local minimization were selected or dropped according to the Metropolis criteria[49] with a temperature of 5000 K. The final selection of conformations is a variation of a previously described procedure.[4] A simple parallelization protocol was used to accelerate calculations: all simulations originated from a given receptor surface point (i_{Rec}) were run consecutively as a single computational job (two to seven hours on a 667 MHz Alpha processor; four to ten hours on a 700 MHz Pentium III workstation running Linux). Thus, a total number of jobs equal to the number of surface points around the receptor (N_{rec}^{sf}) were run in parallel (ranging from 14 to 44 per complex). For each job, all the conformations accumulated after the different MC runs (starting from $6N_{lig}^{sf}$ ligand orientations/rotations) were merged in a single conformational stack,[50] from which the geometrically similar conformations (RMSD of the ligand rotation and translation positional variables <2 Å) were removed (see the scheme in Figure 10). Later, the 400 lowest-energy solutions were selected and compressed further so that only the lowest-energy conformations with RMSD for the ligand interface C^{α} atoms greater than 4 Å were retained. Finally, for each complex, all conformational stacks obtained from the different jobs were merged into a single file, from which the geometrically similar conformations (RMSD of the ligand rotation and translation positional variables <2 Å) were removed as described above.

Solvation term for protein–protein association

The solvation effect upon complex formation was calculated using an atomic solvent-accessible surface (ASA)-based model (i.e. as a sum of per-atomic contributions proportional to the ASA)[42,43] with parameters derived from linear fitting to octanol/water transfer energies for *N*-acetyl amino acid amide derivatives.[51] These per atomic parameters, listed in Table 5, were calculated for ten classes of atoms in the same fashion as that described for vapor–water transfer parameters.[44] The solvation energy was calculated separately for the polar, aromatic and aliphatic atoms, and separate weighting factors were assigned to these three components. The ASPs derived from octanol/water transfer energies seem to describe better the burial of solvent-exposed residues in the interface.[31] Since this solvation

Table 5. Atomic solvation parameters

σ (cal/(mol Å^2))	Radius (Å)	Atom type
15.1	1.95	C aliphatic
17.7	1.8	C aromatic
−17.0	1.7	N uncharged
−54.8	1.7	N^{+}, N^{ζ} in Lys^{+}
−27.3	1.7	$N^{\eta 1}$, $N^{\eta 2}$ in Arg^{+}
−18.5	1.6	O hydroxyl
−13.6	1.4	O carbonyl
−29.9	1.4	O^{-} in Glu, Asp
11.2	2.0	S in SH
2.2	1.85	S in Met or S–S

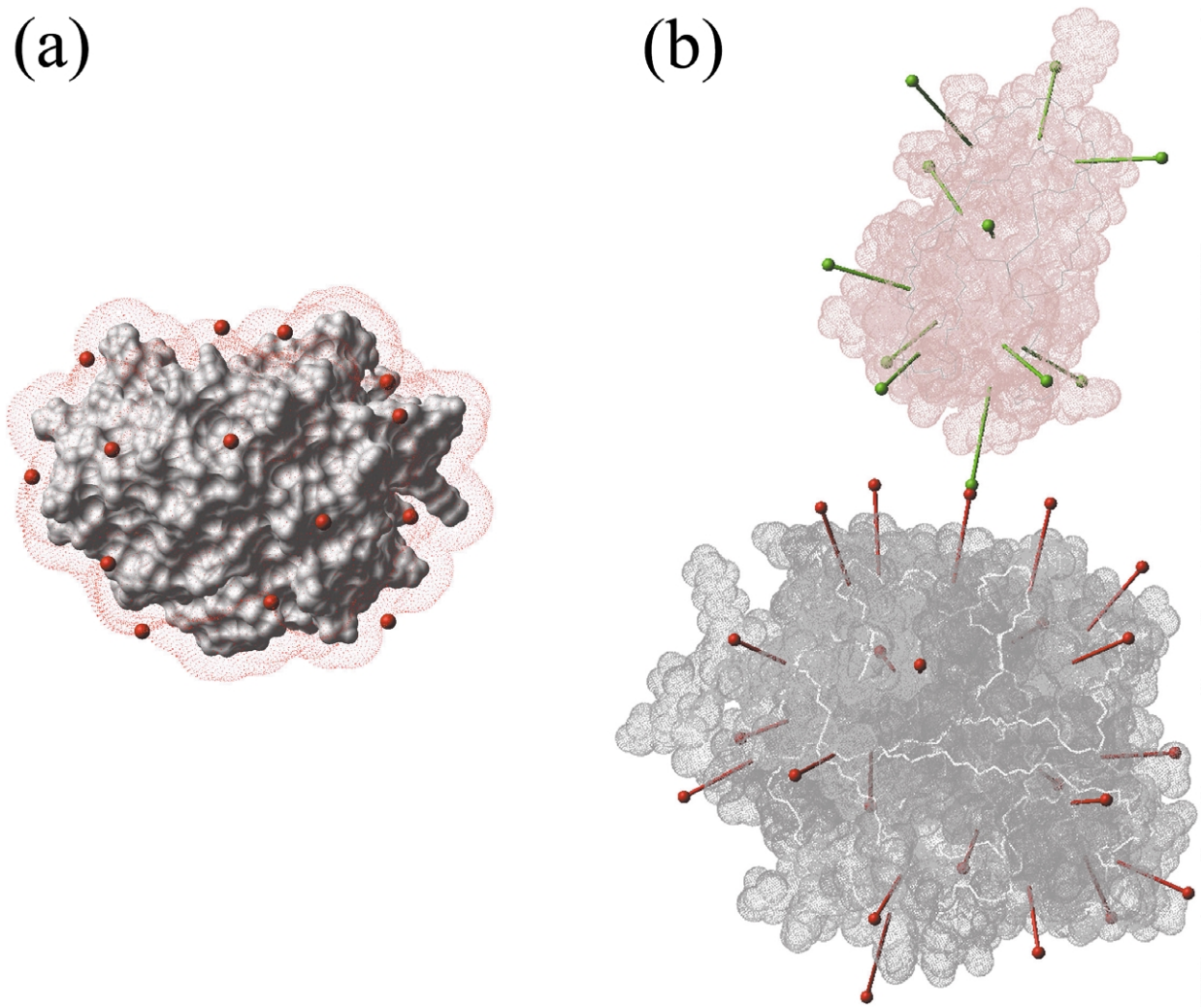

Figure 9. (a) Distribution of the ($i_{\mathrm{Rec}} = 1, 2, \ldots, N^{\mathrm{sf}}_{\mathrm{rec}}$) points around the chymotrypsin receptor (red spheres) that is used to define the starting positions for the docking of the APPI ligand. These points ($N^{\mathrm{sf}}_{\mathrm{rec}} = 23$) were defined on the receptor solvent-accessible surface expanded by 3 Å (represented by the small dots), with a minimum distance of 15 Å between them. (b) A series of points ($i_{\mathrm{Lig}} = 1, 2, \ldots, N^{\mathrm{sf}}_{\mathrm{lig}}$) around the APPI ligand was defined in the same fashion ($N^{\mathrm{sf}}_{\mathrm{lig}} = 12$). Each i_{Rec} point around the chymotrypsin is superimposed onto each i_{Lig} point around the APPI in order to define $N^{\mathrm{sf}}_{\mathrm{rec}} N^{\mathrm{sf}}_{\mathrm{lig}}$ (276) different positions of the APPI ligand. The ligand is also re-oriented to align the axis perpendicular to the solvent-accessible surfaces of ligand and receptor at the match point. Finally, six ligand orientations are generated for every position by rotating the ligand molecule in 60° increments around that axis. A total of $6N^{\mathrm{sf}}_{\mathrm{rec}} N^{\mathrm{sf}}_{\mathrm{lig}}$ (1656) starting positions for the APPI ligand around the chymotrypsin receptor are thus generated (one starting ligand position is shown).

model includes the attractive component of the intermolecular van der Waals interactions, the explicit van der Waals calculations are required here only to avoid large intermolecular clashes in the rigid-body interfaces. Because of that, it was not essential to use the correct geometry of the interacting side-chains during van der Waals calculations, and thus we were more confident in the use of the rigid-body approach.

Optimization of the energy function

In order to optimize the parameters for protein–protein binding, a minimization process was applied, as described below. The conformations obtained from the rigid-body docking simulations were re-evaluated using an energy function consisting of: (i) the van der Waals, hydrogen bonding and electrostatic potentials calculated from the grid maps; and (ii) the solvation energy calculated separately for the polar, aromatic and aliphatic atoms, according to the original per-atomic parameters (Table 5). Because the explicit solvation term was included, neither the SChEM[30] correction for solvent electrostatic screening nor the hydrophobic grid potential (both used during the simulations) were included in the final energy function, defined by equation (1). The weights of the different energy terms (the weight for electrostatics was kept as 1 for a reference) were optimized through the amoeba simplex minimization procedure,[52] using a training set of 24 protein–protein complexes (Table 1). For each iterative step, the new weights defined a new energy function that was used to re-evaluate the docking solutions and to calculate a new ranking value for the near-native solution (i.e. the one

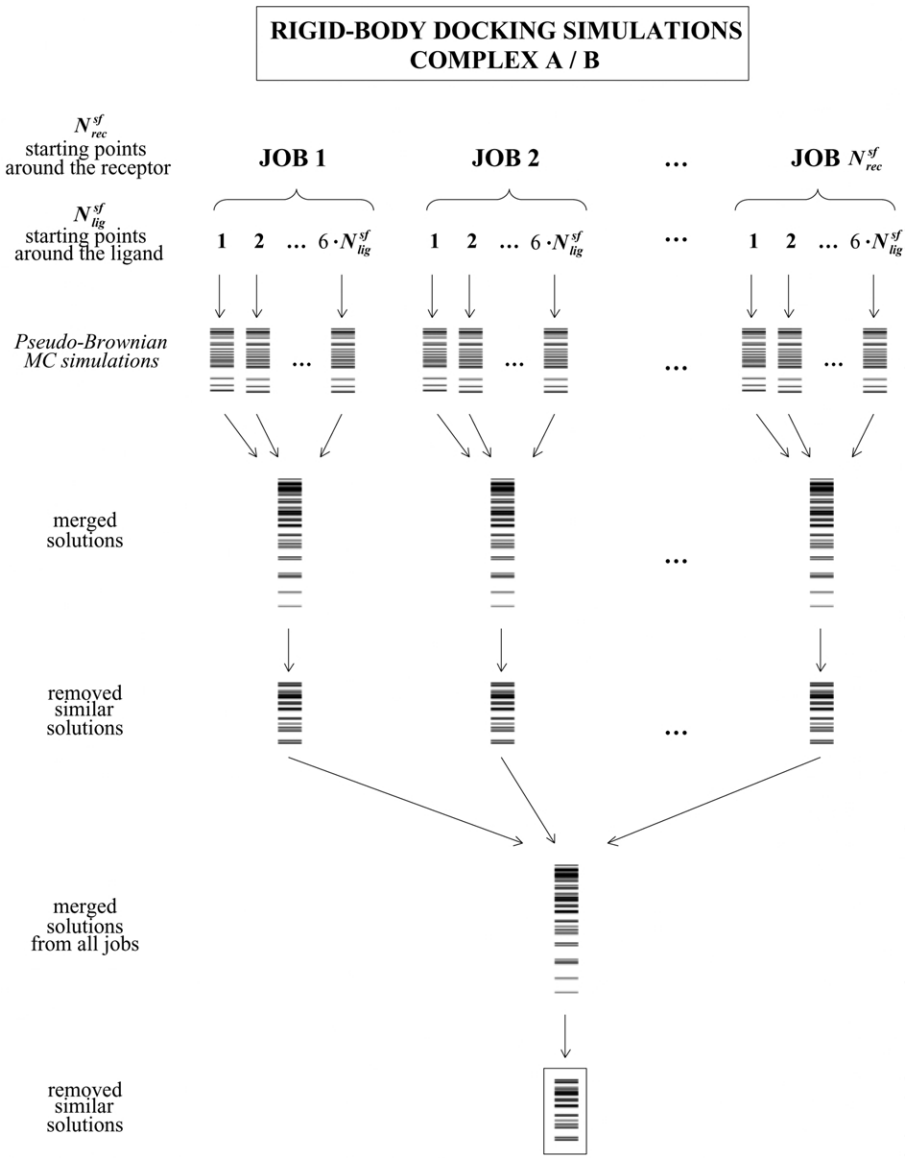

Figure 10. Scheme of the global rigid-body docking procedure. A computational job is launched for each i_{Rec} starting point around the receptor. The resulting conformations are merged and filtered to obtain the final conformational set (for more details, see Materials and Methods).

with the lower RMSD) for each complex. The logarithm of the sum of the ranking values of the 24 near-native solutions (one per complex) was the objective function to optimize (the logarithm function was used here to avoid an excessive influence of the complexes in which the near-native solution was ranked "poorly"). Five independent minimizations were performed, starting from different sets of random weighting values to ensure convergence.

Analysis of the distribution of the docking poses: representation of the energy landscape in spherical coordinates

For each docking solution, the center of mass of the ligand was represented in spherical coordinates with respect to the center of mass of the receptor. A θ angle was calculated with respect to the z axis (which was chosen arbitrarily), and a φ angle was calculated from the projection of the center of mass of the given solution onto the xy plane, with respect to the x axis (defined by the center of mass of the ligand in the known complex structure). As can be seen in Figure 3(a), we can represent the spherical coordinates of the different docking poses in a 2-D sinusoidal projection ($x = \varphi \sin\theta$; $y = 90 - \theta$), in which the position of the ligand in the native complex structure (coordinates: $\{\varphi = 0°, \theta = 90°\}$) is in the center of the plot. This projection is equal-area, the scale is true along all parallels and central meridian, and we can evaluate the density of docking poses at all zones of the plot (because they correspond with the true values on the spherical surface).

We generated a continuous energy map from the previously described point values as follows. The projection of the distribution of docking poses was divided into cells generated by a 2-D grid (60×30). All 2-D unit cells represented portions of the spherical space with equal area. The energy values of the solutions within a particular grid cell were averaged and assigned to this cell. If a

cell had no solutions inside, the highest energy value of all solutions was assigned to that cell. After assigning an energy value to all unit cells, the energy profile was smoothed by averaging the value for each cell over the surrounding 3×3 cell region, as can be seen in Figure 3(b).

Analysis of the distribution of the docking poses: predicted interface propensities

The distribution of docked poses obtained after global docking was projected onto the molecular surface of the receptor as follows. For every solvent-exposed residue of the unbound receptor, the relative surface buried upon binding was calculated for all docking poses and averaged with Boltzmann weighting:

$$\text{Energy-averaged buried surface (EABS}_i\text{)} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\text{ASA}_i^{\text{Unb}} - \text{ASA}_{ik}^{\text{Bnd}}}{\text{ASA}_i^{\text{Unb}}} e^{-(E_k - E_0)/RT}\right) \quad (4)$$

where $\text{ASA}_i^{\text{Unb}}$ is the solvent-accessible surface area for the receptor residue i before ligand binding; $\text{ASA}_{ik}^{\text{Bnd}}$ is the solvent-accessible surface area for the same residue after ligand binding according to the pose k; E_k is the energy value of that pose; E_0 is the lowest energy value; T is the temperature of the simulation (typically 5000 K); R is the gas constant (1.987 cal/mol); and N is the total number of docking poses.

A simplified version of the method was applied as follows. The averaged buried surface was calculated from the 100 lowest-energy solutions, with no energy weighting:

$$\text{Averaged buried surface (ABS}_i\text{)} = \frac{1}{100}\sum_{k=1}^{100}\left(\frac{\text{ASA}_i^{\text{Unb}} - \text{ASA}_{ik}^{\text{Bnd}}}{\text{ASA}_i^{\text{Unb}}}\right) \quad (5)$$

From this value, we calculated the residue interface propensity:

$$\text{Interface propensity (IP}_i\text{)} = \frac{\text{ABS}_i}{\langle \text{ABS} \rangle} = \frac{\text{ABS}_i}{\frac{1}{N_{\text{sar}}}\sum_{i=1}^{N_{\text{sar}}}\text{ABS}_i} \quad (6)$$

where N_{sar} is the number of total solvent-accessible receptor residues. An interface propensity $\text{IP} > 1$ will indicate that the residue is found in the docking interfaces more frequently than expected. Larger IP values would indicate a stronger preference for the interface, but their maximum expected values will be different for each complex. Therefore, in order to compare IP values for residues in different complexes, we have to normalize them between the average value ($\langle \text{IP} \rangle = 1$) and the maximum expected value (IP^{MAX}):

$$\text{Normalized interface propensity (NIP}_i\text{)} = \frac{\text{IP}_i - \langle \text{IP} \rangle}{\text{IP}^{\text{MAX}} - \langle \text{IP} \rangle} = \frac{\frac{\text{ABS}_i}{\langle \text{ABS} \rangle} - 1}{\frac{\text{ABS}^{\text{MAX}}}{\langle \text{ABS} \rangle} - 1}$$

$$= \frac{\text{ABS}_i - \langle \text{ABS} \rangle}{\text{ABS}^{\text{MAX}} - \langle \text{ABS} \rangle} \quad (7)$$

where ABS^{MAX} is the maximum expected ABS value ($\text{ABS}^{\text{MAX}} = 1$). The normalized interface propensity (NIP) value for a specific residue will indicate its preference for the interface.

Definition of experimental interface

In order to compare predicted and experimental interfaces, the latter were defined as the residues in the native complex structure with relative surface buried upon binding RBS > 0.1:

$$\text{Relative buried surface (RBS}_i\text{)} = \frac{\text{ASA}_i^{\text{Unb}} - \text{ASA}_i^{\text{Bnd}}}{\text{ASA}_i^{\text{Unb}}} \quad (8)$$

Statistical significance of the predictions

In order to determine if the global occurrence of the predicted interfaces differs significantly from a random distribution, we evaluated the predicted coverage values by a χ^2 test. For each complex, the distribution of the predicted interface residues was compared to the distribution of the residues in the real interface. A contingency table can be built for the comparison of these two distributions, by computing the number of residues in the predicted interface that are in the real interface, the number of residues in the predicted interface that are not in the real interface, and the number of residues not predicted to be in the interface that are actually in the real interface (as in Table 2). The P significance obtained from applying the χ^2 test (one degree of freedom) to this contingency table indicates the probability that the distribution of the predicted interfaces with respect to the experimental ones differs from a random distribution. This is used here to evaluate if the predicted coverage values are statistically significant, or if, on the contrary, they could have been obtained by chance alone.

Selection of protein–protein complexes in the test set

The method presented here has been validated on a test set formed by recently solved complexes not included in the training of the energy function.

The first test case was the electron transfer complex formed by the proteins FNR and Fd, which play an essential role in photosynthesis.[53] The X-ray structure of the complex between the oxidized FNR and Fd from *Anabaena* PCC7119 has been reported (PDB 1ewy),[54] and suggests a highly probable conformational state of the electron transfer between these two proteins. The structures of the unbound FNR (PDB 1que) and Fd (PDB 1fxa) from *Anabaena* PCC7119 were used to generate the interface propensity maps as described in previous sections. The second test case was the interaction between EPO and EPOR. The structure of EPO is known, both in its unbound form (PDB 1buy) and in complex with EPObp, the extracellular ligand binding domain of EPOR (PDB 1eer).[55] The structure of the unbound dimerized EPObp has been released recently (PDB 1ern). Since each EPObp monomer binds to different areas of EPO, and no contact between EPObp monomers is present in the complex structure, only one of the EPObp monomers was used during simulations.

A further seven complexes were taken from the recent CAPRI competition (Table 1). The coordinates of the unbound subunits were provided by the organizers, but the structures of the protein–protein complexes were not released until all participants had submitted their

models. Therefore, the rigid-body docking simulations used here to obtain the interface propensity maps were run in the complete absence of information about the final solution[56] and therefore constitute a perfect blind validation of the method.

Twelve more complexes were taken from the unbound–unbound test cases in the benchmark compiled by Chen *et al.*[33] The benchmark is formed by 31 unbound–unbound test cases, but we selected only complexes that were significantly different from those we used in our training set. The final selection happened to be formed by 12 complexes, which included three different cases of each one of the four categories in which the benchmark is divided: enzyme-inhibitor, antibody–antigen, others, and difficult cases (Table 1). Thus, this selection constitutes a genuine validating test that effectively covers a broad variety of protein–protein complexes.

Acknowledgements

This work was supported by the Department of Energy (DE-FG03-00ER63042).

References

1. Xenarios, I. & Eisenberg, D. (2001). Protein interaction databases. *Curr. Opin. Biotechnol.* **12**, 334–339.
2. Camacho, C. J. & Vajda, S. (2002). Protein-protein association kinetics and protein docking. *Curr. Opin. Struct. Biol.* **12**, 36–40.
3. Elcock, A. H., Sept, D. & McCammon, J. A. (2001). Computer simulation of protein–protein interactions. *J. Phys. Chem. ser. B*, **105**, 1504–1518.
4. Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2002). Soft protein–protein docking in internal coordinates. *Protein Sci.* **11**, 280–291.
5. Smith, G. R. & Sternberg, M. J. (2002). Prediction of protein–protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12**, 28–35.
6. Sternberg, M. J., Gabb, H. A. & Jackson, R. M. (1998). Predictive docking of protein–protein and protein–DNA complexes. *Curr. Opin. Struct. Biol.* **8**, 250–256.
7. Wodak, S. J. & Janin, J. (1978). Computer analysis of protein–protein interaction. *J. Mol. Biol.* **124**, 323–342.
8. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.
9. Clackson, T., Ultsch, M. H., Wells, J. A. & de Vos, A. M. (1998). Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.* **277**, 1111–1128.
10. Clackson, T. & Wells, J. A. (1995). A hot spot of binding energy in a hormone–receptor interface. *Science*, **267**, 383–386.
11. Hu, Z., Ma, B., Wolfson, H. & Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins: Struct. Funct. Genet.* **39**, 331–342.
12. Novotny, J., Bruccoleri, R. E. & Saul, F. A. (1989). On the attribution of binding energy in antigen–antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry*, **28**, 4735–4749.
13. Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.
14. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins: Struct. Funct. Genet.* **43**, 89–102.
15. Janin, J., Miller, S. & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204**, 155–164.
16. Jones, S. & Thornton, J. M. (1995). Protein–protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63**, 31–65.
17. Jones, S. & Thornton, J. M. (1996). Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
18. Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
19. Young, L., Jernigan, R. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein–protein recognition. *Protein Sci.* **3**, 717–729.
20. Chothia, C. & Janin, J. (1975). Principles of protein–protein recognition. *Nature*, **256**, 705–708.
21. Conte, L. L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
22. Janin, J. & Chothia, C. (1990). The structure of protein–protein recognition sites. *J. Biol. Chem.* **265**, 16027–16030.
23. Sheinerman, F. B. & Honig, B. (2002). On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.* **318**, 161–177.
24. Jones, S. & Thornton, J. M. (1997). Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
25. Zhou, H. X. & Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Struct. Funct. Genet.* **44**, 336–343.
26. Gabdoulline, R. R. & Wade, R. C. (2001). Protein–protein association: investigation of factors influencing association rates by Brownian dynamics simulations. *J. Mol. Biol.* **306**, 1139–1155.
27. Janin, J. (1997). The kinetics of protein–protein recognition. *Proteins: Struct. Funct. Genet.* **28**, 153–161.
28. Norel, R., Petrey, D., Wolfson, H. J. & Nussinov, R. (1999). Examination of shape complementarity in docking of unbound proteins. *Proteins: Struct. Funct. Genet.* **36**, 307–317.
29. Vakser, I. A., Matar, O. G. & Lam, C. F. (1999). A systematic study of low-resolution recognition in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **96**, 8477–8482.
30. Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2002). Screened charge electrostatic model in protein–protein docking simulations. *Pac. Symp. Biocomput.* **7**, 552–565.
31. Vajda, S., Weng, Z., Rosenfeld, R. & DeLisi, C. (1994). Effect of conformational flexibility and solvation on receptor–ligand binding free energies. *Biochemistry*, **33**, 13977–13988.
32. Pickersgill, R. W. (1988). A rapid method of calculating charge–charge interaction energies in proteins. *Protein Eng.* **2**, 247–248.
33. Chen, R., Mintseris, J., Janin, J. & Weng, Z. (2003). A protein–protein docking benchmark. *Proteins: Struct. Funct. Genet.* **52**, 88–91.
34. Lomize, A. L., Riebarkh, M. Y. & Pogozheva, I. D. (2002). Interatomic potentials and solvation parameters from protein engineering data for buried residues. *Protein Sci.* **11**, 1984–2000.
35. Hurley, J. K., Hazzard, J. T., Martinez-Julvez, M.,

Medina, M., Gomez-Moreno, C. & Tollin, G. (1999). Electrostatic forces involved in orienting Anabaena ferredoxin during binding to Anabaena ferredoxin:NADP + reductase: site-specific mutagenesis, transient kinetic measurements, and electrostatic surface potentials. *Protein Sci.* **8**, 1614–1622.
36. Philo, J. S., Aoki, K. H., Arakawa, T., Narhi, L. O. & Wen, J. (1996). Dimerization of the extracellular domain of the erythropoietin (EPO) receptor by EPO: one high-affinity and one low-affinity interaction. *Biochemistry*, **35**, 1681–1691.
37. Lauwereys, M., Arbabi Ghahroudi, M., Desmyter, A., Kinne, J., Holzer, W., De Genst, E. *et al.* (1998). Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. *EMBO J.* **17**, 3512–3520.
38. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
39. Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857.
40. Totrov, M. & Abagyan, R. (1997). Flexible protein–ligand docking by global energy optimization in internal coordinates. *Proteins: Struct. Funct. Genet. Suppl*, **1**, 215–220.
41. Totrov, M. & Abagyan, R. (2001). Protein–ligand docking as an energy optimization problem. In *Drug-Receptor Thermodynamics: Introduction and Applications* (Raffa, R. B., ed.), 1st edit., pp. 603–624, Wiley, New York.
42. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
43. Wesson, L. & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* **1**, 227–235.
44. Abagyan, R. (1997). Protein structure prediction by global energy optimization. Computer Simulations of Biomolecular Systems (van Gunsteren, W. F., *et al.*, eds), vol. 3, pp. 363–394, Kluwer Academic Publisher, Dordrecht.
45. Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. (1981). Affinities of amino acid side chains for solvent water. *Biochemistry*, **20**, 849–855.
46. Pearson, R. G. (1986). Ionization potentials and electron affinities in aqueous solution. *J. Am. Chem. Soc.* **108**, 6109–6114.
47. Abagyan, R., Totrov, M. & Kuznetsov, D. (1994). ICM: a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488–506.
48. MolSoft (2000). *ICM 2.8 Program Manual*, MolSoft LLC, San Diego.
49. Metropolis, N. A., Rosenbluth, A. W., Rosenbluth, N. M., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.
50. Abagyan, R. & Argos, P. (1992). Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.* **225**, 519–532.
51. Fauchere, J. L. & Pliska, V. (1983). Hydrophobic parameters-pi of amino-acid side-chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369–375.
52. Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Comput. J.* **7**, 308–313.
53. Batie, C. J. & Kamin, H. (1984). Electron transfer by ferredoxin:NADP + reductase. Rapid-reaction evidence for participation of a ternary complex. *J. Biol. Chem.* **259**, 11976–11985.
54. Morales, R., Charon, M. H., Kachalova, G., Serre, L., Medina, M., Gomez-Moreno, C. & Frey, M. (2000). A redox-dependent interaction between two electron-transfer partners involved in photosynthesis. *EMBO Rep.* **1**, 271–276.
55. Syed, R. S., Reid, S. W., Li, C., Cheetham, J. C., Aoki, K. H., Liu, B. *et al.* (1998). Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature*, **395**, 511–516.
56. 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.
57. Strynadka, N. C., Adachi, H., Jensen, S. E., Johns, K., Sielecki, A., Betzel, C. *et al.* (1992). Molecular structure of the acyl-enzyme intermediate in beta-lactam hydrolysis at 1.7 Å resolution. *Nature*, **359**, 700–705.
58. Strynadka, N. C., Jensen, S. E., Alzari, P. M. & James, M. N. (1996). A potent new mode of beta-lactamase inhibition revealed by the 1.7 Å X-ray crystallographic structure of the TEM-1-BLIP complex. *Nature Struct. Biol.* **3**, 290–297.
59. Strynadka, N. C., Jensen, S. E., Johns, K., Blanchard, H., Page, M., Matagne, A. *et al.* (1994). Structural and kinetic characterization of a beta-lactamase-inhibitor protein. *Nature*, **368**, 657–660.

Edited by J. Thornton

(Received 16 December 2002; received in revised form 21 October 2003; accepted 29 October 2003)