Crystal contacts as nature's docking solutions

Evgeny Krissinel

European Bioinformatics Institute, Genome Campus, Hinxton,
Cambridge CB10 1SD, United Kingdom
Tel: +44 (1223) 494628, Fax: +44 (1223) 494468, E-mail: keb@ebi.ac.uk

August 14, 2009

An expanded version of article:
E. Krissinel, Crystal contacts as nature's docking solutions,
Journal of Computational Chemistry (2009), in press;
published on-line 6 May 2009; DOI 10.1002/jcc.21303

Abstract

The assumption that crystal contacts reflect natural macromolecular interactions makes a basis for many studies in structural biology. However, the crystal state may correspond to a global minimum of free energy where biologically relevant interactions are sacrificed in favour to unspecific contacts. A large-scale docking experiment was performed in order to assess the extent of misrepresentation of natural (in-solvent) protein dimers by crystal packing. As found, the failure rate of docking may be quantitatively interpreted if both calculation errors and misrepresentation effects are taken into account. The failure rate analysis is based on the assumption that crystal structures reflect thermodynamic equilibrium between different dimeric configurations. The analysis gives an estimate of misrepresentation probability, which suggests that weakly bound complexes with $K_D \geq 100$ μM (some 20% of all dimers in the PDB) have higher than 50% chances to be misrepresented by crystals. The developed theoretical framework is applicable in other studies, where experimental results may be viewed as snapshots of systems in thermodynamic equilibrium.

Keywords: macromolecular crystals, protein-protein interactions, protein-protein docking, crystal misrepresentation effects, failure rate analysis, thermodynamic equilibrium.

1 Introduction

Many important processes in biology are associated with the ability of proteins to interact with each other and form complexes [1]. Protein-protein interactions are thought to be specific [2], which means that a given protein is likely to interact only with particular protein types and in particular regions of protein surface. This feature is important for research and applications. It is commonly assumed that data on potentially interacting proteins and structural details of protein binding may bring about a better understanding of biochemical processes and give a clue for drug discovery and design [3].

Most of our today's knowledge on structural aspects of protein-protein interactions (PPIs) comes from protein crystallography [4]. Because the crystalline state represents an energetically optimal arrangement of molecular units, one could expect that favourable protein interactions are preserved by crystal packing. In simple words, this means that crystals are likely to exhibit natural protein contacts, or interfaces, which are formed in protein's native, "working" environment. This assumption is exploited in most, if not all, studies where structural aspects of PPIs are inferred from crystals.

Two problems arise when inferring on PPIs from crystallographic data. Firstly, distinguishing between significant crystal interfaces (i.e. those supposedly representing the natural interactions) and artifacts of crystal packing is not always a simple task [5]. To a certain degree, the problem may be helped by crystallographic considerations. For example, a hetero-chain asymmetric unit and non-crystallographic symmetry rotations may indicate a complex, while a pure translation almost always (except for naturally infinite polymers, such as muscle proteins) identifies an artifactual, unspecific interface. Also, it is widely assumed that if a given interface is found in a few different crystal forms then it is likely to be the "real" one [6]. Such recipies lack quantitative description and obviously are not applicable in many cases, e.g. when only a single crystal form is available.

A more rigorous approach to the identification of significant interfaces in crystal packing is based on the analysis of interface properties [5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. Ideally, such type of analysis should be performed in energy terms. However, accurate energy estimates for protein-protein binding represent a challenging theoretical problem, therefore, most methods use various descriptors, such as interface planarity, shape, surface complementarity, propensity, residue composition, area etc. in attemt to find a combination of properties that would reliably identify significant interfaces. This line was researched in many studies, including those cited above, with different degrees of success.

In Ref. [22], we addressed a closely related problem, the identification of macromolecular assemblies in crystal packing, using empirical estimates for the dissociation free energy of macromolecular complexes. As was found, despite a relatively simple nature of the estimates, our procedure, PISA (Protein Interfaces, Surfaces and Assemblies) reproduces about 90% of complex structures verified by independent (non-crystallographic) experimental studies. This success rate is higher than initially expected, which can be hardly attributed to the quality of free energy estimates used. More probably, the success is due to the ready geometry of PPIs provided by crystal packing. PISA does not dock macromolecular units, but rather assumes that crystal-given dockings (the interfaces) are the optimal ones. Then, even approximate free energy estimates appear to be sufficient for the successful identification of complexes.

The question of whether crystal interfaces correspond to the natural interactions (or whether they, indeed, are the optimal dockings) is the second problem in crystallography-based analysis of PPIs. The relationship between natural complexes and their representation in crystals is almost always assumed to be a straightforward one. However, crystals exemplify thermodynamic systems in global minimum of free energy, taking into account both natural and unspecific interactions. Therefore, if energy of a natural interaction does not compete with the combined effect of unspecific crystal contacts, then such interaction may be sacrificed in the course of crystallization. If this happens, an apparently significant crystal interface does not represent the natural PPI. In such cases, we will say that the PPI (or a complex) is misrepresented by crystal packing.

One can view the change of complex configuration in crystal environment as interaction-induced shift in energy landscape. These effects have been thoroughly discussed in literature in application to conformational changes in proteins upon binding [23, 24, 25, 26]. Recently, these theoretical concepts have received experimental verification [27, 28]. As pointed out in Refs. [23, 24, 25, 26], most proteins exist in dynamic equilibrium between several conformations, which may be classed into four energy landscape patterns. Analysis of these patterns suggests that a conformation, different from the lowest-energy one, may be selected for structure-specific (lock-and-key) binding, subject to energy and kinetic barriers between the conformations. These results are directly tranferable to protein complexes in crystal packing, where "conformational change" refers to a wide spectra of complex configurations.

Direct assessment of misrepresentation effects in crystals is difficult because of a rather limited number of protein complexes with 3D structure experimentally verified by both crystallographic and non-crystallographic (NMR [29], EM [30], small-angle scattering [31, 32]) studies. Thus, in Ref. [22], we were able to use only 430 PDB (Protein Databank [33]) entries, reviewed in other studies [20, 34], where structure of macromolecular complexes was thoroughly investigated using complementary experimental techniques. If highly accurate free energy calculations were available, then significant crystal contacts could be verified by computational docking [35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 48, 49, 50, 51, 52, 53, 54, 55, 56]. However, the accuracy of existing docking procedures is not well understood. As a rule, various parameters of docking programs are calibrated on a limited set of selected targets [57, 58], normally chosen as significant crystal interfaces. This type of procedure does not guarantee universality of calibrated parameters, and the overall success of macromolecular docking is known to be rather limited [59].

Docking failures are most often attributed to the inevitable errors in free energy calculations or imperfectness of other scores used. However, docking procedures are tested on targets of primarily crystallographic origin [57, 58]. Therefore, there is a hypothetical possibility of docking failure (that is, non-arrival at a significant crystal interface) due to misrepresentation of PPIs in crystals. In this study, we attempt to identify contributions from both calculation errors and crystal misrepresentation effects to the failure rate of docking by analysing docking results on a large set of structures. As will be shown, these contributions may be identified because of the differences in their dependences on free energy of complex dissociation. Only dimeric complexes are considered, because of their far greater population in the PDB, as compared with complexes of higher mutiplicity, and also because of a simpler theoretical analysis they require. An estimate of misrepresentation effects in crystals will be given, which suggests that a considerable part of weak dimers in the PDB may not correspond to natural complexes, and that the probability of seeing transient interactions in crystalline state is rather slim.

2 Method

In this study, we aim to conclude on the reproduceability of protein dimers, identified in crystal packing ("crystal dimers"), with a computational docking procedure. An ideal, error-free docking is supposed to arrive at a natural dimer configuration. If crystal dimer corresponds to the natural one, then no conformational modelling is required and it should be reproduceable by the simplest rigid-bidy bound docking procedure [35, 36, 37, 38]. However, in reality not all dimers may be reproduced due to computational errors and possible difference between crystal and natural complexes.

As mentioned in the Introduction, the number of protein complexes with independent (non-crystallographic) verification of their 3D structures is limited. Therefore, we will use protein dimers identified as significant crystal contacts by PISA software [22]. PISA employs certain physical-chemical models of PPIs for the identification of chemically stable complexes in crystal packing. In order to maintain consistency between the models used for the identification of crystal dimers and docking and minimize computational artifcats due to the difference in underlying principles, we develop a docking method based on PPI models that are close, as much as possible, to those used in PISA. Below we sketch the method.

An optimal docking position (orientation and translation) of proteins A and B corresponds to the maximum of Gibbs free energy ΔG_0 dissipated by the solvent upon formation of dimeric complex AB:

$$\Delta G_0 = -\Delta G_{int} - T\Delta S \tag{1}$$

where ΔG_{int} is binding energy and ΔS is the entropy cost of dimerization. In PISA, ΔS is estimated as [22]

$$\begin{aligned}\Delta S &= C + \frac{3}{2}R\log\left(\frac{m(A)m(B)}{m(AB)}\right) + \frac{1}{2}R\log\left(\frac{\prod_k J_k(A)\prod_k J_k(B)}{\prod_k J_k(AB)}\right) + \\ &\quad R\log\left(\frac{\gamma(AB)}{\gamma(A)\gamma(B)}\right) + 2F\Delta\sigma \end{aligned}\tag{2}$$

where $m(X)$, $J_k(X)$ and $\gamma(X)$ stand for the mass, kth principal moment of inertia and symmetry number of molecule X, respectively, $\Delta\sigma$ is buried surface area (BSA), and C and F are constants.

It may be shown that orientation dependence of ΔS is rather weak. Indeed, first two terms in Eq. (2) do not depend on the orientation, and so does the 3rd term in case of globular proteins. In the worst hypothetical case of elongated molecules, approximated with cylinders, the 3rd term shows variations of about

0.5 kcal/mol with the geometry of a dimer (from side-to-side to end-to-end orientations of the cylinders) at room temperatures. The fourth term of Eq. (2) equals to zero in the case of asymmetric dimers and reaches some 0.41 kcal/mol in the case of symmetric complexes. The last term, corresponding to the entropy of surface side-chains, was found to be quite small [22], contributing about 1 kcal/mol per 10^4 Å^2 of BSA. Thus, the total error may amount to $1-2$ kcal/mol, which is below the expected precision of PISA models (±5 kcal/mol or worse [22]). Therefore, we neglect orientation dependence of ΔS in our docking procedure.

Binding energy ΔG_{int} in PISA is estimated as [22]:

$$\Delta G_{int} = \Delta G_{solv} + N_{hb}E_{hb} + N_{sb}E_{sb} + N_{db}E_{db} \tag{3}$$

where ΔG_{solv} stands for the solvation energy gain upon complex formation, N_{hb}, N_{sb}, N_{ds} are numbers of formed hydrogen bonds, salt bridges and disulphide bonds, respectively, and E_{hb}, E_{sb}, E_{ds} stand for their free energy effects. The following approximation for ΔG_{solv} is used in PISA [22]:

$$\Delta G_{solv} = \sum_k \omega_k \left(\Delta\sigma_k^A + \Delta\sigma_k^B\right) \tag{4}$$

where ω_k is atomic sovation parameter (ASP) of kth atom type and $\Delta\sigma_k^X$ is the sum BSA of atoms of kth type belonging to molecule X.

All terms of Eq. (3) essentially depend on docking position. Assuming position-independent ΔS, one can formulate the docking problem as finding a relative position of molecules A and B that minimizes ΔG_{int} at zero (subject to tolerance) overlap of the molecules. It may be shown that all terms of Eq. (3) may be regarded as properties of molecular surface. Then, minimization of ΔG_{int} may be conveniently solved by a shape correlation technique, described in Ref. [35]. Below we sketch our approach, which is based on the original work [35] and its modifications reported in Refs. [42, 47].

Firstly, geometrical and interaction properties of docking molecules are represented as discret functions on a 3D grid, constructed as a sufficiently large molecule-embedding cube, divided into N cells in each dimension. First function describes the protein core (see Fig. 1):

$$\rho_{jkl}^c(A) = \begin{cases} 1 & \text{protein } A \text{ inside} \\ 0 & \text{open space} \end{cases} \tag{5}$$

Non-zero correlation of core functions:

$$\Theta_{\alpha\beta\gamma}^c(A,B) = \sum_{jkl} \rho_{jkl}^c(A)\rho_{j+\alpha,k+\beta,l+\gamma}^c(B) \tag{6}$$

indicates an overlap of protein structures at given mutual orientation and displacement $(\alpha\beta\gamma)$. Second function describes the solvent-excluded volume and what we will call a "solvation layer" (Fig. 1):

$$\rho_{jkl}^s(A) = \begin{cases} i = \sqrt{-1} & \text{solvent-excluded volume } W(A) \text{ of protein } A \\ f_{jkl}(A) & \text{solvation layer } F(A) \text{ of protein } A \\ 0 & \text{open space} \end{cases} \tag{7}$$

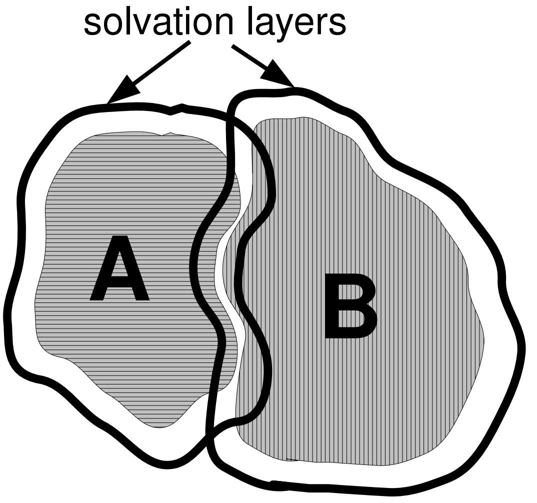

Figure 1: Schematic of docked molecules A and B. Hashed areas represent the molecule inside, the "solvation layers" are boundaries of the molecule solvent-excluded volumes. In our method, the "solvation layers" are assigned hydrophobic properties, and the total hydrophobic effect is measured by parts of "solvation layers" buried in the solvent-excluded volume.

"Solvation layer" represents the surface of closest approach of solvent molecules to protein surface. Each cell of the solvation layer is assigned a hydrophobic effect

$$f_{jkl} = \sum_n \omega_n \Delta\sigma_n \tag{8}$$

where n enumerates atoms that contribute the van der Waals surface patches $\Delta\sigma_n$ into the cell, and ω_n stand for the atomic solvation parameter of nth atom. Note that "solvation layer", as well as the molecule inside, belongs to the solvent-excluded volume. Imaginary part of the correlation of solvation layers ρ^s gives the solvation energy gain:

$$\begin{aligned} Im\left(\Theta^s_{\alpha\beta\gamma}(A,B)\right) &= Im\left(\sum_{jkl}\rho^s_{jkl}(A)\rho^s_{xyz}(B)\right) \\ &= \sum_{\substack{jkl\in F(A)\\ xyz\in W(B)}} f_{jkl}(A) + \sum_{\substack{jkl\in W(A)\\ xyz\in F(B)}} f_{xyz}(B) \\ &= \Delta G_{solv} \end{aligned} \tag{9}$$

where $x = j + \alpha$, $y = k + \beta$ and $z = l + \gamma$.

The remaining terms in Eq. (3) represent the effects hydrogen bonds, salt bridges and disulphides. They should be calculated with respect to geometrical features of the corresponding bonds. This increases the dimensionality of the problem, which makes it computationally unfeasible. Therefore, we simplify the situation by assuming a hydrogen bond and salt bridge wherever the distance between suitable electron donor and acceptor is less than $r_{hb} = 4$ Å, and a disulphide bond if sulphurs in contacting CYS residues are separated by less than $r_{ds} = 3$ Å. These distance cut-offs were chosen empirically such as to approximate most closely the more accurate PISA calculations. r_{hb} and r_{ds} do not coincide with the equivalent PISA parameters because they also accomodate the effect of the finite resolution of 3D grids. The hydrogen bond function is discretized on a grid with resolution r_{hb} as follows:

$$\rho^{hb}_{jkl}(A) = \begin{cases} \sqrt{n} & \text{if cell } (j,k,l) \text{ contains } n \text{ electron donors} \\ & \text{on the surface of protein } A \\ i\sqrt{n} & \text{if cell } (j,k,l) \text{ contains } n \text{ electron acceptors} \\ & \text{on the surface of protein } A \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

and imaginary part of the corresponding correlation:

$$N^{hb}_{\alpha'\beta'\gamma'} = Im\left(\Theta^{hb}_{\alpha'\beta'\gamma'}(A,B)\right) = Im\left(\sum_{ijk} \rho^{hb}_{ijk}(A)\rho^{hb}_{i+\alpha',j+\beta',k+\gamma'}(B)\right) \tag{11}$$

approximates the number of potential hydrogen bonds at translation vector $(\alpha'\beta'\gamma')$. The salt bridge function ρ^{sb}_{ijk} is constructed similarly to Eq. (10) but in respect only to polar residues LYS, ARG, HIS (donor-nitrogens) and GLU, ASP (acceptor-oxygens). The disulphide function is discretized on a grid with resolution r_{ds} as

$$\rho^{ds}_{jkl}(A) = \begin{cases} 1 & \text{if cell } (j,k,l) \text{ contains a sulphur in CYS residue} \\ & \text{on the surface of protein } A \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

and the corresponding correlation $\Theta^{ds}_{\alpha'\beta'\gamma'}$ gives an approximate number of potential disulphide bonds $N^{ds}_{\alpha'\beta'\gamma'}$.

Using the above-described 3D correlation functions, one can represent the binding energy ΔG_{int}, Eq. (3), in the following form:

$$\Delta G^{\alpha\beta\gamma}_{int} = Im\left(\Theta^{s}_{\alpha\beta\gamma}\right) + Im\left(\Theta^{hb}_{\alpha\beta\gamma}\right) E_{hb} + Im\left(\Theta^{sb}_{\alpha\beta\gamma}\right) E_{sb} + \Theta^{ds}_{\alpha\beta\gamma} E_{ds} \tag{13}$$

where translation $(\alpha\beta\gamma)$ for the last three terms is calculated from translation $(\alpha'\beta'\gamma')$, used in Eqs. (10,12), according to the respective grid resolutions. The minimum $\Delta G^{\alpha_0\beta_0\gamma_0}_{int}$, achievable at zero structure overlap $\Theta^{c}_{\alpha_0\beta_0\gamma_0} = 0$, corresponds to the optimal translation of molecules A and B, $(\alpha_0\beta_0\gamma_0)$, at given relative orientation Ω. The minimum $\Delta G^{\alpha_*\beta_*\gamma_*}_{int}$ chosen over all possible orientations $\{\Omega_j\}$ of A and B identifies the optimal docking position as a combination of orientation Ω_* and translation vector $(\alpha_*\beta_*\gamma_*)$.

In our practical implementation, we have found it necessary to allow a small overlap of core functions ρ^c in order to compensate the imperfectness of discretized functions due to finite resolution of 3D grids. We do this by addition of a penalty term to the binding energy:

$$\Delta G^{\alpha\beta\gamma}_{S} = \Delta G^{\alpha\beta\gamma}_{int} + \xi\left(\frac{\Theta^{c}_{\alpha\beta\gamma}(A,B)}{\max^{2/3}\left(\Theta^{c}_{000}(A,A), \Theta^{c}_{000}(B,B)\right)}\right)^2 \tag{14}$$

and subsequent minimzation of $\Delta G_S^{\alpha\beta\gamma}$ instead of $\Delta G_{int}^{\alpha\beta\gamma}$. The penalty in Eq. (14) is interpreted as a parabolic potential built on a weighted measure of overlapped surface cells. At $\xi = 5\cdot 10^3$ kcal/mol, this penalty allows only occassional overlap of a few cells but leads to better docking positions.

We have chosen to sample the orientation space with resolution of 2°, which was empirically found to be a good compromise between computation time and accuracy, shifted generously to the latter. All correlations are calculated using FFTW (Fastest Fourier Transfrom in the West) software [60]. FFT is most efficient on dimensions $N = 2^n$, out of which the calculations were found practical with discretizing protein molecules on 3D grids with $N = 256$ [35]. This keeps the grid resolution below 1 Å for most protein structures. Due to the necessity to calculate several FFT correlations [42, 47], our method is not expected to be faster than some other docking algorithms. As found, a parallel implementation of the method on a 60-node cluster of 2.8 GHz AMD CPUs yields a docking solution in 20-30 minutes. Here, we sacrifice performance for a description of PPIs that is consistent with PISA software [22], used for the selection of dimeric structures in the PDB, as described in the next Section.

3 The Dataset

The dataset was initially composed of stable protein dimers ($\Delta G_0 \geq 0$), calculated by PISA software [22] in the absence of any ligands (unless covalently linked). Then clusters of similar dimers were identified, and only one central structure from the cluster was left in the set. The structure similarity criteria used for clustering were identical to those employed in PISA, where structures A and B are considered similar if their structural alignment yields the following values of quality score Q and sequence identity SI [22]:

$$Q = \frac{N_{align}^2}{(1 + (RMSD/3)^2)N_A N_B} \geq 0.65 \qquad SI = \frac{N_{ident}}{N_{align}} \geq 0.9 \qquad (15)$$

In these expressions, N_X stands for the number of residues in structure X, $RMSD$ is r.m.s.d. between aligned C_α's at best structure superposition, N_{align} is the number of aligned residue pairs, of which N_{ident} pairs are formed by identical residues. SSM (Secondary Structure Matching) software [61] was used to perform the alignments.

Conditions (15) correspond to a rather high structure similarity. However, we use these criteria because even moderate structure changes may significantly influence the interface properties and have a drastic effect on complexation. The final structure set used in our study includes 4065 dimeric complexes, covering the range of $\Delta G_0 = 0 \ldots 211$ kcal/mol. 3431 (84%) structures in the dataset are homodimers.

Many PDB entries represent only parts of natural proteins. Quite typically, only selected protein domains are crystallized, either those of interest or those that are crystallizable. Therefore, most probably, not all structures in the selected dataset are "truly" dimeric. This, however, is not significant for the purpose of our study. Indeed, PISA treats all PDB structures as if they were complete proteins, and derives oligomeric states that are likely to correspond to *given* macromolecules, whether they represent the natural polypeptides or not. Therefore it is possible to treat them as true dimeric structures in our docking experiment as well, disregarding the fact that they may be the artifacts of sample preparation.

4 Results and Discussion

The developed docking procedure has been applied to all dimeric complexes in the selected dataset. Before the docking, orientation of one subunit in each complex was randomized in order to eliminate the possibility of docking by a trivial translation. For each protein pair, we analyse only one docking solution with maximum free energy of dissociation ΔG_0 (1), in difference of many other studies, where success is traditionally measured by the occurence of correct solution among 10 or so top-ranked alternatives. Then, docking solutions were compared to the original complexes by calculating the r.m.s.d. of the corresponding C_α atoms at best superposition of the original and docked dimers. Docking solutions with r.m.s.d. $\leq$ 10 Å were counted as acceptable, others were classed as failures. The 10 Å threshold has been chosen after visual inspection of a considerable number of dockings. Figure 2 shows the r.m.s.d. distribution of all dockings. As seen from the Figure, the chosen threshold corresponds roughly to the minimum between a pronounced low-r.m.s.d. peak ("successful" dockings) and a long hill in the high-r.m.s.d. end ("failed" dockings). The Figure suggests that the exact value of the threshold r.m.s.d. should not make a significant effect on final conclusions, as less than 5% of all dockings fall into r.m.s.d. region of $5-10$ Å, normally suggested for the discriminating threshold.

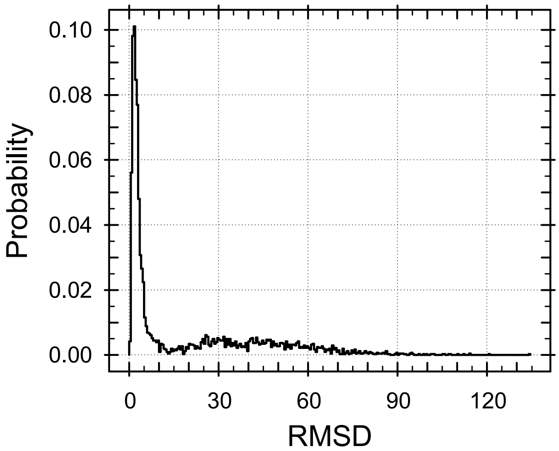

Figure 2: Distribution of docking solutions over r.m.s.d. from the corresponding crystal dimers, built on 0.25 Å bins.

In Figure 2, 38% of dockings belong to the high-r.m.s.d. hill, meaning precisely that for 38% of structures, the maximum free energy dimer was found to differ substantially from the most significant crystal contact. This figure looks confusingly high, taking into account that it was obtained for the simplest rigid-body bound docking, with no conformational effects involved. A seemingly plausible explanation of docking failures is that optimal dockings are missed be-

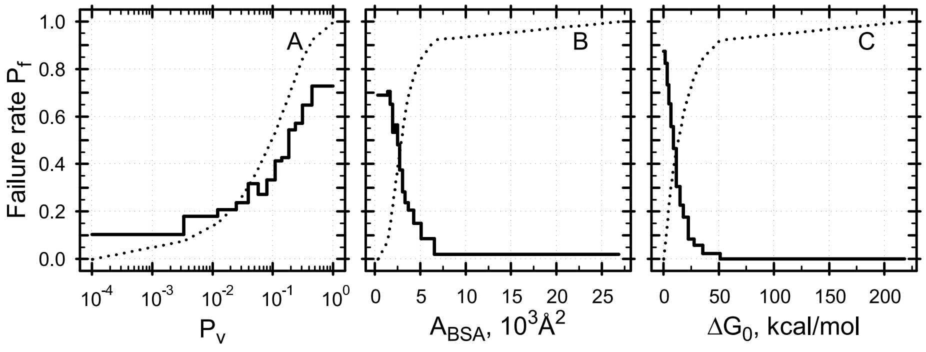

Figure 3: Relative fraction of dockings (solide lines), failed to arrive at the corresponding crystal contact, as a function of (A) hydrophobic P-value P_v (B) buried surface area A_{BSA} and (C) free energy of dissociation ΔG_0. All values presented are for the corresponding crystal dimers, P_f is calculated by averaging within equipopulated ΔG_0 bins and every step of the solid curve indicates the corresponding bin. Dotted lines indicate the cumulative number of docked pairs, divided by the total number of dockings (4065). See details in the text.

cause of limited accuracy of free energy calculations and finite resolutions of 3D grids and angular search procedure. However, it is also possible that the overall low success rate reflects mainly the composition of the dataset, if, for example, it overrepresents classes of crystal dimers not reproduceable by docking.

In order to illustrate that, let us examine how the failure rate of docking P_f depends on a few different parameters. Figure 3 suggests that docking success increases consistently with decreasing hydrophobic P-value (Fig. 3A), decreasing buried surface area A_{BSA} (Fig. 3B) and increasing free energy of dissociation ΔG_0 (Fig. 3C). Consider first data in Fig. 3A. The hydrophobic P-value P_v of an interface is defined as probability to find a same-area patch on protein surface that would be more hydrophobic than the interface. Therefore, low P_v indicate specific hydrophobic spots, which are likely to be preferential in protein-protein interactions and for this conserved by crystal packing. In the Figure, the failure rate reaches maximum at $P_v \approx 0.5$. This corresponds to the situation when the chances to find patches on protein surface that are more or less hydrophobic than the dimer interface, are equal, and, therefore, hydrophobic properties of the interface are not “surprising”. At $P_v \geq 0.5$ protein binding is not specific, which means that there is no strong preference to any particular dimer configuration among few permitted by structural features. As seen from Fig. 3A, the failure rate of docking P_f is maximal in these conditions.

Generally speaking, structural promiscuity of protein contacts does not imply a weak binding. Hypothetically, two proteins may form a few different complexes with close values of ΔG_0 [23, 24, 25, 26, 27, 28]. Such complexes would then exist in a dynamic equilibrium [23, 24, 25, 26]. In this case, the docking objectives are ill-defined because of ambiguity of target selection. Due to the finite accuracy of practical calculations, docking program may pick any of the similar-energy dimers. One can imagine, however, that the same may happen in the course of crystallization if, subject to the crystallization regime or precipitation agents used, the procedure arrives at structurally different but

PDB	Crystal dimer			Docked dimer		
entry	A_{BSA}	ΔG_0	P_v	A_{BSA}	ΔG_0	P_v
1ea9	6580	19.7	0.446	3360	33.1	0.068
1xr4	6930	26.5	0.092	5080	36.5	0.072
2j6h	6960	16.1	0.473	4260	26.4	0.259
2cst	7150	35.8	0.338	6130	49.2	0.111
1sgk	8350	32.7	0.206	4390	32.8	0.325

Table 1: Summary of failed dockings from the highest-area bin in Figure 3B. Middle column shows the buried surface area A_{BSA}, in Å^2, dissociation free energy ΔG_0, in kcal/mol and hydrophobic P-value P_v for crystal dimers identified by PISA software [22]. Right column shows the same data calculated for docked complexes. The crystal and docked dimers are shown in Figure 4.

energetically close packings. As seen from Fig. 3A, about 50% of dimers in the dataset have $P_v \geq 0.1$. This indicates a moderate interaction specificity and, therefore, reproduceability of these dimers in docking may be impaired in presence of alternative configurations.

Buried surface area A_{BSA} (Fig. 3B) is a traditional measure of interface significance. As may be found from the Figure, docking fails to reproduce crystal contacts with A_{BSA} larger than 6500 Å^2 in only $\approx$ 2% of instances. Further analysis shows that no failures are found at $A_{BSA} > 8400$ Å^2. Docking failures with $A_{BSA} \geq 6500$ Å^2 are summarized in Table 1, and Figure 4 shows the corresponding dimeric complexes. As seen from Table 1, in all cases, docking arrives at non-crystal dimers because they show a higher ΔG_0 than the corresponding crystal interfaces. The ΔG_0 difference between crystal and docked dimers is rather high but within the 3σ confidence limits for the anticipated accuracy of PISA models ($\sigma \approx 5$ kcal/mol). At the same time, BSA of docked dimers is less than that of the corresponding crystal interfaces. This results in lower P-values, indicating an apparently higher specificity of interactions in docked complexes. The only exception here is PDB entry 1SGK [66], where a higher value of ΔG_0 is due to the formation of a higher number of hydrogen bonds, rather than a higher hydrophobic specificity. Visual inspection of docked dimers in Fig. 4 suggests that docked 1EA9 [62] and 1SGK [66] are asymmetric and therefore unlikely to be the real dimers. Docking of 1XR4 [63] is an artifact due to the treatment of flexible "arms" as rigid structures. However, docked 2J6H [64] and 2CST [65] represent well-packed symmetric complexes, which could be the locally-stable alternative dimers.

At BSA below 1700 Å^2, the failure rate of docking reaches 70% (cf. Fig. 3B). Remarkably, P_f shows a consistent growth with decreasing A_{BSA}. About 50% of crystal dimers in the selected dataset have $A_{BSA} \leq 3000$ Å^2, of those less than 50% are reproduced by docking. In order to interpret these results, note that the underlying reason for taking A_{BSA} as a measure of interface significance is that it correlates with the binding properties: smaller BSA implies weaker binding. Then, the smaller A_{BSA}, the smaller should be the absolute difference in ΔG_0 between alternative docking solutions. Hence, one possible explanation for data in Fig. 3B is that the accuracy of energy calculations becomes increasingly insufficient for discrimination between the alternatives at decreasing BSA. Figure

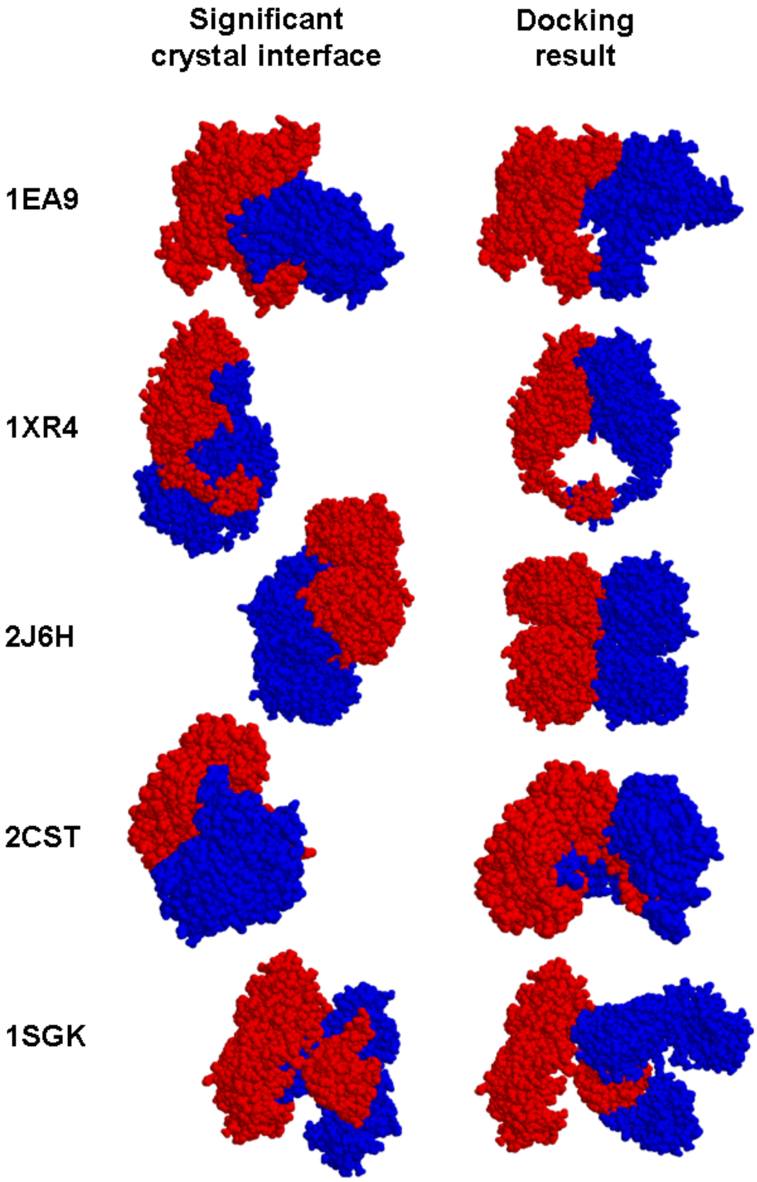

Figure 4: Comparison of docking failures from the highest-area bin in Figure 3B (right column of structures), with the corresponding crystal dimers identified by PISA software [22]. The summary of the corresponding docked and crystal interfaces is given in Table 1.

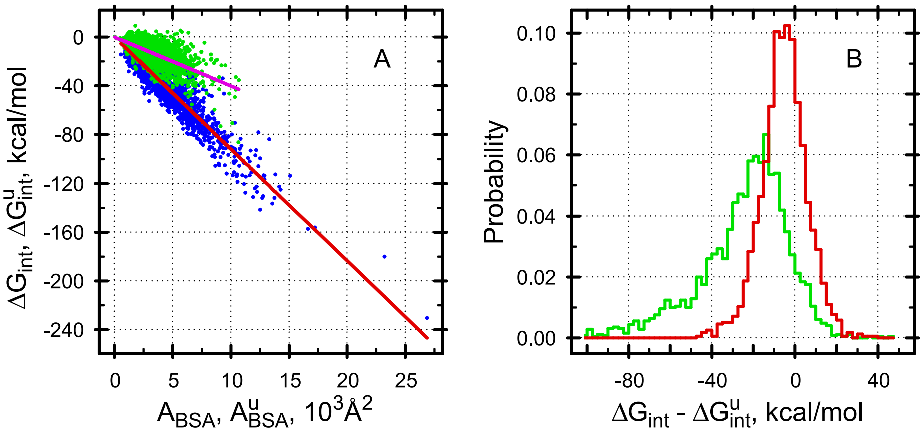

Figure 5: (A) The relationship between buried surface area A_{BSA} and binding energy. Each protein pair is represented by one blue and one green dots. Blue dots $(A_{BSA}, \Delta G_{int})$ correspond to "significant" crystal interfaces. Green dots $(A^u_{BSA}, \Delta G^u_{int})$ correspond to sum energy and sum BSA of all other, unspecific, crystal contacts. Solid lines represent the corresponding linear fits. The data were calculated using the dataset of 4065 PDB entries described in Section 3. (B) Distributions of successful (green line) and failed (red line) dockings over the difference in binding energy between significant and all unspecific interfaces. The distributions are calculated from data in (A).

5A illustrates the situation. In the Figure, blue dots represent A_{BSA} and ΔG_{int} of significant crystal interfaces, and green dots show the sum energy ΔG^u_{int} and sum BSA A^u_{BSA} of all unspecific (inter-dimer) contacts for all PDB entries in the dataset. As may be seen from the Figure, significant interfaces provide, on average, twice more binding energy per Å^2 of BSA than the unspecific crystal contacts. Also, significant interfaces may have considerably larger BSA than the combined area of unspecific contacts. Analysis of Figs. 3B and 5A suggests that crystal dimers are reproduced by docking in areas where clusters of blue and green dots are clearly separated. Where blue and green dots are mixed, the failure rate increases in approximate proportion to the degree of mixing. Figure 5B provides further insight into the situation. It may be seen from the Figure that most docking failures happen when ΔG_{int} is close to ΔG^u_{int} (the red-line distribution of $\Delta G_{int} - \Delta G^u_{int}$ of failed dockings is centered almost at 0), while for the most of successful dockings ΔG_{int} clearly prevails (the green-line distribution of successful dockings is shifted into the area of $\Delta G_{int} \leq \Delta G^u_{int}$).

One can, again, suggest an alternative explanation of $P_f(A_{BSA})$ dependence in Fig. 3B, arguing that close values of ΔG_{int} and ΔG^u_{int} may enable substantial structural changes during crystallization, particularly on the right-hand slope of the red curve in Fig. 3B, where the unspecific interactions prevail. If that happens, then the maximum energy dimer in solution may differ from the one represented by the most significant interface in crystal. Note that an accurate docking procedure is expected to reproduce complexes in solution because it takes no unspecific inter-complex interactions into account. Therefore, the difference between crystal and natural dimers would be seen as a docking failure.

It has been concluded in a number of studies [20, 68, 69, 70] that BSA larger than 600-850 Å^2 indicates a biologically relevant interface. A lower figure of 400 Å^2 was found in Ref. [9] and then used in the Protein Quaternary Structure (PQS) server [5]. The minimal BSA of potentially stable crystal dimers in our dataset is found to be 390 Å^2 (PDB entry 1SDX [67]), which agrees with the literature data. However, it follows from Figs. 3B,5A and above considerations that unspecific interactions may prevail at $A_{BSA} \leq 3000$ Å^2, causing substantial changes to the original complexes, and, therefore, dimeric structures with low A_{BSA} may be misrepresented by crystals.

Figure 3C shows the dependence of failure rate P_f on the free Gibbs energy of dissociation ΔG_0 (1). This dependence is similar to the one in Fig. 3B and may be interpreted in the same terms as above. It appears, however, that this dependence is more suitable for quantitative interpretation thanks to the fact that free Gibbs energy is an ultimate state function for thermodynamic systems. Two observations in Fig. 3C are important for such analysis. Firstly, there is a non-zero chance to reproduce crystal dimer with any ΔG_0, and maximal $P_f \approx 0.88 < 1$ is attained at $\Delta G_0 \approx 0$. The near-zero values of free energy indicate a very low reactivity of molecules, which makes the selection of a preferable complex configuration extremely difficult. In this situation, the fact that crystal dimers are reproduced in about 12% of all dockings should be interpreted in pure probabilistic terms. This implies that an average protein pair may form about $N =$ 8-10 different dimers, identified as principal local minima ΔG_0^i of the free Gibbs energy, and docking procedure arrives "randomly" at one of them when calculation errors are larger than the differences between the minima. The term "principal local minima" here refers to the essentially different docking solutions, as measured by the r.m.s.d. threshold. The figure of 8-10 principal docking solutions appears to be reasonably close to the most probable number of contacts per chain in crystal packings, as illustrated by the distribution shown in Figure 6, where the distribution peak and center are found at 7 crystal contacts

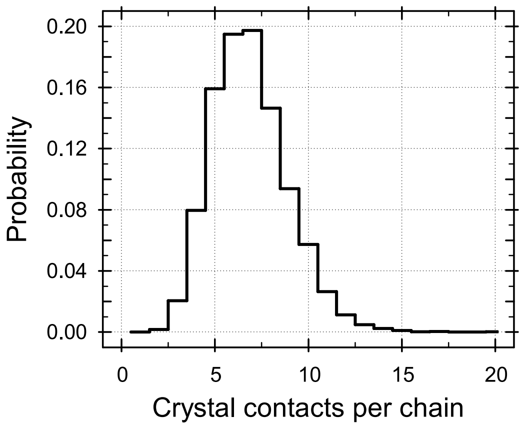

Figure 6: The distribution of the number of interfaces per protein chain in X-ray entries of the PDB. The peak and mass center of the distribution are found at 7 contacts per chain.

per chain. Crystal contacts represent geometrically optimized docking solutions and are expected to be binding, therefore, they may correspond to principal local minima ΔG_0^i. At $\Delta G_0 \approx 0$, the free energy of all other crystal dimers $\Delta G_0^i \approx 0$ as well, which means that docking has to make a pick from $N \approx 7$ energetically close configurations. This seems to be a plausible explanation of a limited failure rate P_f in the zero energy end of Fig. 3C.

The second interesting feature of $P_f(\Delta G_0)$ dependence in Fig. 3C is that it shows a nearly perfect, to the quality of docking data, exponential fall (see also Figure 7). This type of behaviour suggests an idea about its possible origin. Imagine that an average protein pair makes N different stable dimers D_i [23, 24, 25, 26, 27, 28] with free Gibbs energies of dissociation ("energy states") $\Delta G_0^i \geq \Delta G_0^{i+1} \geq 0$, where index $i \in [0..N)$ enumerates the dimers. In thermodynamically equilibrated solutions, the occurence probability of ith dimer is

$$P_D^i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}, \qquad x_i = \frac{\Delta G_0^i}{RT}, \qquad x_i \geq x_{i+1} \geq 0 \tag{16}$$

Each dimer D_i represents a docking solution. An ideal docking procedure arrives at the most probable (highest-energy) dimer, i.e. D_0. Imagine next that dimer D_i may crystallize as a significant crystal contact subject to the occurence probability P_D^i. Then the ideal docking solution D_0 will differ from crystal dimer with probability

$$P_f = 1 - P_D^0 = \frac{\sum_{j>0} \exp(x_j - x_0)}{1 + \sum_{j>0} \exp(x_j - x_0)} \tag{17}$$

which is the failure rate of docking. In the limit of $\Delta G_0 = 0$, Eq. (17) yields $P_f(0) = (N-1)/N < 1$. At high ΔG_0, when $\exp(x_j - x_0) \ll 1$, Eq. (17) reduces to a single exponent $P_f \approx \exp(x_1 - x_0)$. Assuming that the free energy spectra $\{\Delta G_0^i\}$ scales uniformly with ΔG_0, so that $\Delta x = x_0 - x_1 \approx \alpha x_0$, obtain

$$P_f(x_0) \approx \exp(-\alpha x_0) \tag{18}$$

Fitting Eq. (18) to docking results (solid line in Fig. 7) yields $\alpha \approx 0.053$ (dashed line). Formally, approximation (18) is valid at $\exp(-\Delta x) \ll 1$, or $P_f \ll 1$. However, Fig. 7 suggests that it agrees with docking results at considerably higher $P_f \leq 0.85$. This allows us to postulate that energy states $\{x_i\}$ are equidistant:

$$x_i = x_0 - i \cdot x_0/N \tag{19}$$

in which case P_f becomes exponential almost everywhere. Indeed, denote $z = \exp(-x_0/N)$, then

$$P_D^0 = \frac{z^{-N}}{\sum_{i=1}^{N} z^{-i}} = \frac{1}{1 + \sum_{i=1}^{N-1} z^{N-i}} \approx \frac{1}{1 + z/(1-z)} \tag{20}$$

$$P_f = 1 - P_D^0 \approx z = \exp\left(-\frac{\Delta G_0}{N \cdot RT}\right) \tag{21}$$

where approximation is valid for large N. We will refer Eqs. (20,21) as "PDIC model" (Perfect Docking, Imperfect Crystals). Docking data in Fig. 7 are best reproduced by PDIC with $N = 19$ energy states, which corresponds to $\alpha \approx 0.053$

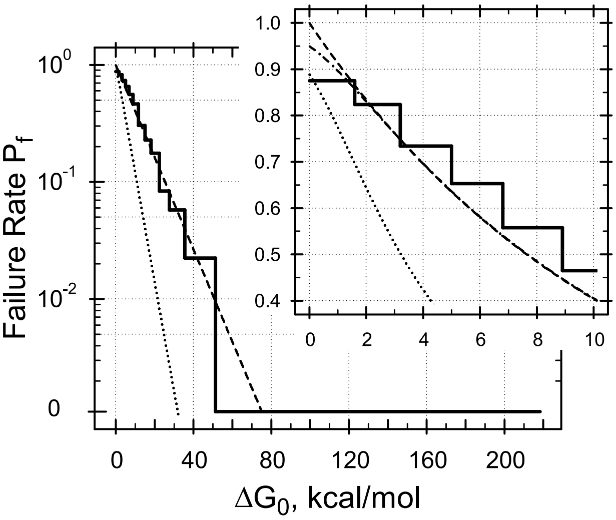

Figure 7: Failure rate of docking P_f fitted with PDIC (Perfect Docking, Imperfect Crystals) model (Eqs. (20,21)). Solid line shows the same data as in Fig. 3C. Dashed line shows the large-N approximation in Eq. (20) and center line corresponds to the exact PDIC model, both for $N = 19$ average number of principal docking solutions per pair. Dotted line shows the exact PDIC model for $N = 9$, which agrees best with the maximum $P_f \approx 0.88$ reached at $\Delta G_0 \approx 0$. See discussion in text.

quoted above. In the Figure, dashed and center lines show P_f calculated with and without the large-N approximation, respectively. As seen from the Figure, the assumption of thermodynamically equilibrated system of N dimeric configurations $\{D_i\}$ with equidistant energy states $\{\Delta G_0^i\}$ allows one to reproduce the exponential fall of $P_f(\Delta G_0)$ everywhere except very low $\Delta G_0 \leq 2$ kcal/mol.

At $\Delta G_0 \approx 0$, PDIC gives a higher failure rate of docking (≈ 0.95) than what is observed in the docking experiment. In order to reproduce the "experimental" value of $P_f(0) \approx 0.88$, an average of $N = 9$ principal docking solutions per protein pair should be assumed in PDIC. This, however, would lead to a significantly lower failure rate at higher ΔG_0, as shown by dotted line in Fig. 7. This disagreement between PDIC and docking results suggests one to include the effect of calculation errors into consideration.

In practical docking, energy states $\{x_i\}$ are calculated with errors $\{\xi_i\}$. Therefore, docking procedure arrives at dimer D_c with free energy $x_c = \max_i (x_i + \xi_i)$, which does not coincide with the highest-energy docking solution D_0 if $x_c > x_0 + \xi_0$. If, e.g., D_c corresponds to D_i, then P_f is calculated as in Eq. (17): $P_f = 1 - P_D^i$. However, in our analysis we can consider only a probability to associate D_c with D_i, treating this as a hypothesis. Besides, the value of x_0 is not given by docking and should be treated as a hypothesis as well. Each such hypothesis corresponds to the exponential solution (17), and then the failure rate is calculated as sum effect of all possible associations and

x_0-hypotheses:

$$P_f(x_c) = \sum_{i=0}^{N-1} \int_0^\infty \left(1 - P_D^i(x_0)\right) \phi_i(x_c, x_0) \frac{N-i}{N} \,\mathrm{d}x_0 \tag{22}$$

where $(N-i)/N \,\mathrm{d}x_0$ stands for $\mathrm{d}x_i$, and $\phi_i(x_c, x_0)$ is the probability density to associate docking solution D_c with dimer D_i. D_c is associated with D_i if $x_i + \xi_i = x_c$ and energies of all other docking solutions $x_j + \xi_j < x_c$, $j \neq i$. Let $\omega(\xi)$ be the free energy error function. Then

$$\phi_i(x_c, x_0) = \omega(x_c - x_i) \prod_{j \neq i} \int_{-\infty}^{x_c - x_j} \omega(\xi) \,\mathrm{d}\xi, \qquad x_i = x_0 - i \cdot x_0/N \tag{23}$$

Assuming normal error $\varepsilon = \Delta G_0^\varepsilon / RT$ for free energy calculations, $\omega(\xi)$ may be estimated as

$$\omega(\xi) = \frac{\sqrt{2}\,\exp\left(-\frac{\xi^2}{2\varepsilon^2}\right)}{\sqrt{\pi}\,\varepsilon\,\mathrm{erfc}\left(-\frac{x_c}{\sqrt{2}\varepsilon}\right)} \tag{24}$$

where denominator is chosen from the condition that free energy x_i of any principal docking solution D_i is non-negative:

$$\int_0^\infty \omega(x_c - x_i) \,\mathrm{d}x_i = 1 \tag{25}$$

Finally, substituting Eq. (24) into Eq. (23), obtain

$$\phi_i(x_c, x_0) = \frac{\sqrt{2}\,\exp\left(-\frac{(x_c - x_i)^2}{2\varepsilon^2}\right)}{\sqrt{\pi}\,\varepsilon\,\mathrm{erfc}^N\left(-\frac{x_c}{\sqrt{2}\varepsilon}\right)} \prod_{j \neq i} \left(1 + \mathrm{erf}\left(\frac{x_c - x_j}{\sqrt{2}\varepsilon}\right)\right) \tag{26}$$

which may be further used in Eq. (22) to calculate the failure rate of docking. We will call Eqs. (16,22,26) as "IDIC model" (Imperfect Docking, Imperfect Crystals).

It is useful for further analysis to understand the effect of calculation errors on the identification of docking solutions. Figure 8 shows the probability Φ_i to associate docking solution D_c with dimer D_i, calculated as follows:

$$\Phi_i(x_c) = \int_0^\infty \phi_i(x_c, x_0) \frac{N-i}{N} \,\mathrm{d}x_0 \tag{27}$$

The calculations verify that $\sum_i \Phi_i = 1$. As seen from Fig. 8, $\Phi_i > \Phi_{i+1}$, so that D_c is most likely associated with D_0. The Figure also shows that Φ_i hardly depends on x_c if x_c is less than calculation error ε. Indeed, at $x_c \ll \varepsilon$, energy states $\{x_i\}$ at most probable $x_0 \approx x_c$ are found well within each other's error margines and cannot be discriminated. Here, the difference between Φ_i is due to the contribution from higher $x_0 > x_c + \varepsilon$ in the integral (27), which barely depends on x_c if $x_c \ll \varepsilon$. On contrary, if $x_c \gg \varepsilon$ then the separation of energy states $\{x_i\}$ is larger than the calculation error ε and Φ_0-curve becomes dominant. In the limit of $x_c/\varepsilon \to \infty$ or $\varepsilon \to 0$, $\{x_i\}$ are clearly discriminated, which, effectively, means reduction to PDIC, Eqs. (20,21). Indeed,

$$\lim_{\varepsilon \to 0} \phi_i(x_c, x_0) = \delta_{i,0} \delta(x_c - x_0) \tag{28}$$

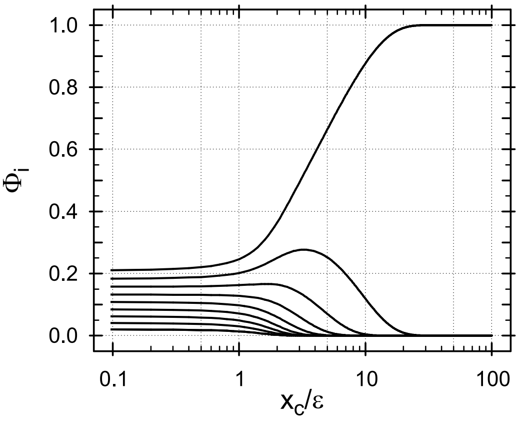

Figure 8: Probabilities Φ_i (Eq. (27)) to associate a docking solution D_c with ith dimeric form D_i, as a function of dissociation free energy $x_c = \Delta G_0^c/RT$ of D_c, in the units of the calculation error ε. $N = 9$ dimeric forms are assumed, ith line from top corresponds to Φ_i. See discussion in text.

and then Eq. (22) reduces to Eq. (17), leading further to PDIC (20,21).

It is interesting to see whether docking results may be explained only by calculation errors. In the absence of crystal misrepresentation effects, the probability to find dimer D_i as a significant crystal interface is $P_D^i = \delta_{i,0}$. Substituting this into Eq. (22), obtain

$$P_f(x_c) = \sum_{i=1}^{N-1} \int_0^\infty \phi_i(x_c, x_0) \frac{N-i}{N} \,\mathrm{d}x_0 = \sum_{i=1}^{N-1} \Phi_i(x_c) = 1 - \Phi_0(x_c) \qquad (29)$$

which we will address to as "IDPC model" (Imperfect Docking, Perfect Crystals). IDPC is an antipode to PDIC, both being special cases of IDIC.

Dashed line in Figure 9 shows best IDPC fit to docking results. As seen from the Figure, at $N = 17$ and $\Delta G_0^\varepsilon = 1.25$ kcal/mol, IDPC fit is nearly as goods as that of PDIC (shown by center line). The root-mean square deviation:

$$rmsd = \sqrt{\frac{1}{N_{bins}} \sum_{i=1}^{N_{bins}} \left(P_{f,i} - P_{f,i}^c\right)^2} \qquad (30)$$

is only marginally better in IDPC (cf. Table 2). In Eq. (30), $P_{f,i}$ is failure rate of docking in ith ΔG_0 bin (cf. Fig. 3) and $P_{f,i}^c$ is the corresponding model approximation calculated in the mass center of the bin.

At $\Delta G_0 \approx 0$, IDPC shows a much closer, than PDIC, match with docking results. However, in difference of PDIC, the success rate of docking at low ΔG_0 in IDPC cannot be interpreted as a mere chance to pick the "correct" dimer from N energetically close alternatives. Indeed, association probabilities Φ_i in IDPC are not equal at $\Delta G_0 \to 0$ (cf. Fig. 8). Since $\Phi_0(x_c \approx 0) > 1/N$, docking solution D_c has higher, than random, chances to be associated with the "correct" dimer D_0. According to Eq. (29), this results in lower $P_f(\Delta G_0 \approx 0)$

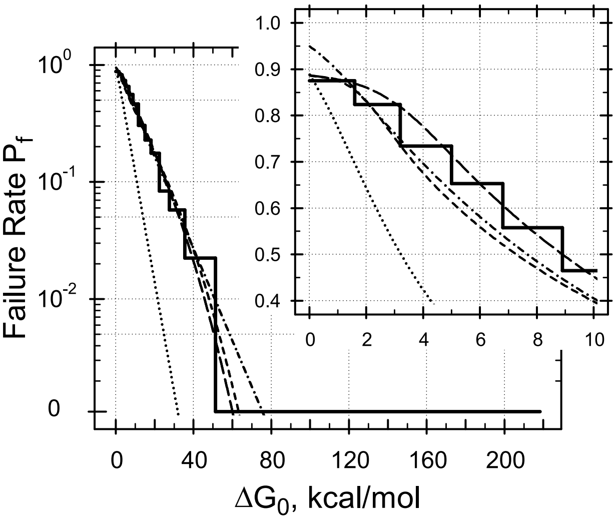

Figure 9: Comparison of fits to the failure rate of docking P_f (solid line), calculated in the framework of 3 different models. The solid, center and dotted lines are the same as in Fig. 7. The center, dashed and long-dashed lines correspond to PDIC (Eqs. (20,21), pure misrepresentation effects), IDPC (Eqs. (26,29), pure docking calculation errors) and IDIC (Eqs. (22,26), both misrepresentation effects and calculation errors) models, respectively. The fit parameters are summarized in Table 2. See discussion in text.

on comparison with PDIC at similar values of N, which is indeed seen in Fig. 9.

Long dashed line in Fig. 9 shows the best fit of docking results in the framework of IDIC, which takes both calculation errors and misrepresentation effects into account. As seen from the Figure, IDIC fit is visibly better than those given by PDIC and IDPC, which is also confirmed by a considerably lower *rmsd* (30) (cf. Table 2). Interestingly enough, IDIC and PDIC give very close values of $P_f(0)$ at the same number of energy states $N = 9$. In both models, this is a direct consequence of indiscrimination between alternative dimers at low x_c. Indeed, note that association functions $\phi_i(x_c, x_0 \approx 0)$ (26) fade at $x_c \geq \varepsilon$, until when the association probabilities $\Phi_i(x_c)$ (27) stay almost constant (cf. Fig. 8).

Model	N	ΔG_0^ε	$rmsd$
PDIC	19	N/A	0.049
IDPC	17	1.25	0.046
IDIC	9	2.3	0.019

Table 2: Summary of best fits to the failure rate of docking, presented in Figure 9. N stands for the average number of principal docking solutions, ΔG_0^ε is the normal error of free energy calculations, in kcal/mol, and *rmsd* measures the difference between the observed and model failure rates, see Eq. (30).

This allows one to represent the IDIC master equation (22) as

$$P_f(x_c \leq \varepsilon) \approx \sum_{i=0}^{N-1} \left(1 - P_D^i(0)\right) \Phi_i(0) = \frac{N-1}{N} \tag{31}$$

where equalities $P_D^i(0) = 1/N$ and $\sum_i \Phi_i = 1$ are used.

As follows from the above analysis, the free-energy dependence of failure rate of docking may be explained by either calculation errors (IDPC) or crystal misrepresentation effects (PDIC), or a combination of both factors (IDIC). It should be admitted that accuracy of our P_f calculations (solid line in Fig. 9) is not quite sufficient for unambiguous discrimination between alternative interpretations. A considerable improvement in P_f calculations may be achieved only by a substantial increase in the number of docked structures. This, however, is not possible due to the limited size of the PDB. In this situation, one can choose the most plausible alternative, which seems to be IDIC for the following reasons. Firstly, the best-fit calculation error ΔG_0^ε in IDIC amounts to 2.3 kcal/mol, which is higher than that obtained in IDPC (1.25 kcal/mol) and PDIC (effectively 0). Higher values of free energy calculation error are more acceptable here because it is unlikely to have ΔG_0^ε much lower than what was estimated previously for PISA models (±5 kcal/mol [22]). Secondly, $N = 9$, obtained in IDIC (cf. Table 2), is closer to the average number of contacts per chain in the PDB (see distribution in Fig. 6), than $N = 19$ and $N = 17$ obtained for PDIC and IDPC, respectively. Thirdly, IDIC gives a better-quality fit to docking results as compared with PDIC and IDPC, with more than twice lower r.m.s.d. (cf. Table 2). Finally, the presence of both calculation errors and crystal misrepresentation effects is logically justified.

Assuming that IDIC provides a more realistic interpretation of docking results than PDIC and IDPC, one can conclude that an average protein pair has $N = 9$ principal docking solutions. Then, if docking were exact, the failure rate would be given by PDIC with $N = 9$, shown by dotted line in Fig. 9. This line represents the probability that crystal and natural dimers are different, i.e. the pure misrepresentation effect. The Figure suggests that the effect is limited to weakly bound complexes. For example, some 12% of crystal dimers with $\Delta G_0 \approx 10$ kcal/mol seem to misrepresent their natural forms, while for crystal dimers with $\Delta G_0 \approx 20$ kcal/mol these expectations are as low as 1%. It is worth noting that many crystal dimers in the PDB appear to be weakly bound. The fraction of misrepresented dimers in a dataset of M protein pairs may be estimated as

$$F_c = \frac{1}{M} \sum_{i=1}^{M} P_f\left(\Delta G_0^{(i)}\right) \tag{32}$$

where $\Delta G_0^{(i)}$ is the dissociation free energy of ith dimer, and P_f is calculated as in PDIC (Eqs. (20,21)). For the dataset used in present study, $F_c = 0.19$, which means that 19% of non-redundant dimers in the PDB may be misrepresented by crystal packing.

Weak protein-protein complexes, which may readily dissociate or associate depending on precise physiological condition or environment, play an important role in many biological processes, such as signal transduction [71], electron transport [72, 73, 74], transcriptional regulation [75, 76], growth factors [77, 78, 79, 80], molecular switches [81, 82, 83], cell-cell recognition [84] and many others [85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99]. The dissociation constant K_D of weak complexes may reach a few hundred μM [90], which corresponds to ΔG_0 of only a few kcal/mol. Experimental identification of structural features of such complexes is difficult because of their transient nature (see, e.g., [92, 93, 96, 99]). In a number of studies, it was found that weak PPIs manifest themselves in highly condensed, pre-crystal, solutions and crystalline state, which implies that protein crystallography may be used for studying weak associations (see examples in [74, 79, 80, 90, 94, 95, 96, 98]). However, the overall probability of seeing a weak biological interaction as a crystal interface remains unclear. Our results provide an estimate of such probability, which suggests that capturing transient PPIs in crystals may be less likely than anticipated. Therefore, weak complexes, obtained from crystallographic data, should be always verified by complementing studies.

If accurate, 1 kcal/mol and less, calculations of dissociation free energy ΔG_0 were possible, a first and easy step to verify the structure of a weak complex would be to attempt to reproduce the corresponding crystal contact by bound (conformation-less) docking. If a crystal dimer is close to the natural, in-solvent one, no conformation modelling is required for reproducing it by docking. Then, a negative docking result should be taken as an indication of substantial structural changes induced by crystallization. In this connection, the importance of accurate docking programs cannot be underestimated.

Over last 8 years, quality of different docking algorithms is tested in CAPRI (Critical Assessment of PRedicted Interactions) competition [58]. In this "blind" docking experiment, final targets are unknown to contestants and the corresponding structures are offered at generosity of their authors prior the deposition into the PDB. The experiment aims essentially on unbound docking, with unbound molecules from different crystals given as starting points. The success rate of such experiments has been found well below 10% on average, which was mainly attributed to the difficulties of unbound docking [59]. Figure 10 shows CAPRI targets, reviewed in [59], superimposed on the results of present study. As seen from the Figure, most targets represent very weak complexes, half of which are not stable in PISA estimates ($\Delta G_0 < 0$). Only two targets: No. 9 and 14, fall in the region of dissociation free energy where probability of misrepresentation by crystals is low ($\approx 1\%$). These dimers were successfully docked by program used in this study. All other targets may be misrepresented by crystals with probability 50% and higher, according to the above analysis. Our docking program, being applied to these targets, failed to reproduce the corresponding crystal contacts.

Even assuming considerable ($\approx$ 10 kcal/mol) errors in PISA energy estimates, many CAPRI targets in Fig. 10 show a rather marginal stability, which means weak binding and a possible co-existence of alternative dimeric forms in solution. As has been shown, in this situation one can expect about-90% failure rate of docking even if target structures correspond exactly to natural complexes. On the other hand, it does not seem completely unrealistic that CAPRI targets were identified mainly from crystallographic considerations, while the possibility of crystallization-induced structural changes of protein complexes was not always taken into account. As found above, misrepresentation of weak complexes by crystals may also result in about-90% failure rate even if docking programs are perfect. Whatever is the reason, the approximate correspondence between failure rates in CAPRI and our results suggests that, at least for some tar-

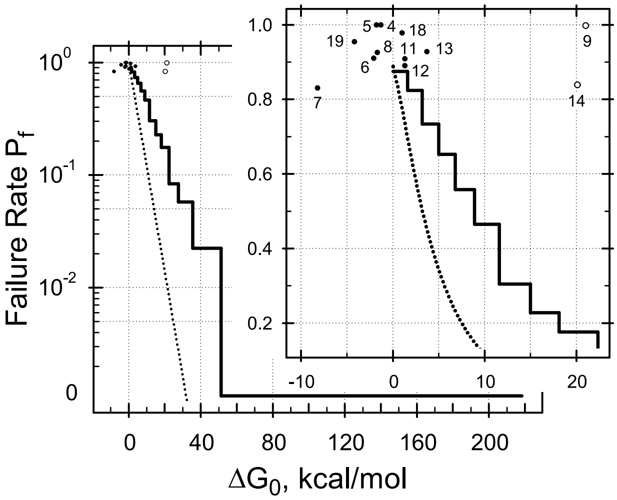

Figure 10: Failure rate of docking in CAPRI experiment [58], superimposed on the results of present study. Solid and dotted lines are the same as in Fig. 9. Numbered dots and circles correspond to CAPRI targets reviewed in Ref. [59]. Only dimeric targets were used, and only circle-marked targets were successfully docked by program used in this study. See discussion in text.

gets and/or some programs, successful dockings may emerge merely by chance, therefore falling into about-10% success region in Fig. 10.

5 Conclusion

Broadly speaking, both crystals and docking programs give us models of protein complexes, and it appears that they both have limits to the accuracy of models they provide. In present study, we attempted to estimate these limits by comparing protein dimers identified in crystal packing with the results of computational docking. In our analysis, we assumed the existence of alternative dimeric structures with equidistant energy spectra for each protein pair, and hypothesized a reflection of their thermodynamic equilibrium in crystal packings. These assumptions were necessary for deriving a theoretical model for the failure rate of docking and the likelihood of misrepresentation of protein dimers in crystals. Therefore, quantitative aspects of our results may have a limited value, however, the model is useful for general understanding and qualitative analysis.

The underlying reasons for misrepresentation effects in crystals and docking errors are quite similar. In crystals, misrepresentation may happen if energy gap between alternative complex configurations is too narrow on comparison with the binding power of crystal contacts. Docking is likely to fail when energy gap between principal docking solutions compares with free energy calculation error. We have shown that crystals and docking agree very well on strongly bound complexes, where alternative configurations are well separated on energy scale. However, in case of weak complexes, crystal and docked dimers may differ in up to $\approx 90\%$ of instances, and free energy trend of this disagreement is best explained if imperfectness of both crystals and docking calculations is assumed.

It is widely accepted that crystals give a much better idea about complex structure than does the computational docking, and our results confirm this in general. As appears, docking errors and misrepresentation effects have very similar rate at $\Delta G_0 \approx 0$, however, the latter fade with increasing ΔG_0 much faster than the former.

As found, weak complexes may be significantly misrepresented by crystal packing. Transient complexes with $K_D > 100$ μM ($\Delta G_0 \leq 5$ kcal/mol) are estimated to have only $10-15\%$ chances to retain their structure in crystalline state. Reliable ($1-2\%$ errors) representation of complexes in crystals is expected at $\Delta G_0 > 15-20$ kcal/mol. The misrepresentation effects disappear only at $\Delta G_0 > 35-40$ kcal/mol, when, in good agreement with physical considerations, binding forces become nearly as strong as covalent linking. In computational docking, the free energy benchmarks are higher. For the docking program used, no errors were recorded at $\Delta G_0 \geq 50$ kcal/mol, relatively reliable results are obtainable at $\Delta G_0 \geq 40$ kcal/mol, and the program is expected to produce more errors than correct answers if $\Delta G_0 \leq 10$ kcal/mol.

Different datasets of macromolecular complexes are used in the literature to calibrate or test computational procedures related to the prediction of macromolecular interactions and complexes, docking, active site recognition and similar. Our results emphasize that independent, non-crystallographic, evidence for weak 3D interactions should be secured prior including them into the dataset. While examination of such datasets is beyond the scope of this paper, we presented an example of CAPRI competition, where weak complexes appear to dominate and, therefore, the contest results could reflect the composition of CAPRI dataset rather than the quality of docking programs.

Finally, our theoretical framework may be applicable in other studies, where experimental results may be viewed as snapshots of thermodynamically equilibrated systems. An obvious field of appication includes comparative analysis of protein folds obtained from protein crystallography, NMR studies and computational modelling (CASP competition [100]). A major advantage of our approach to such sort of analysis is that it estimates the quality of the dataset and indicates the principally achievable rate of success. Therefore, we believe that the method presented is more rigorous and conceptually correct than simple estimates of success used traditionally. The method requires a sizable dataset in order to achieve a reasonable accuracy in the failure rate calculations (4065 protein pairs were used in present study), but the outcome is worth the computational cost.

Acknowledgement

E.K. thanks Dr. Kim Henrick for reading the manuscript and helpful discussion. The work was supported by the research grant No. 721/B19544 from the Biotechnology and Biological Sciences Research Council (BBSRC) UK.

References

[1] Berg, J.M., Tymoczko, J.L. and Stryer, L. Biochemistry; W.H. Freeman and Co., New York, 2002.

[2] Jones, S. and Thornton, J.M. Proc. Natl. Acad. Sci. USA, 1996, **93**, 13–20.

[3] Krogsgaard-Larsen, P., Tommy Liljefors, T. and Madsen, U. Textbook of Drug Design and Discovery; Taylor & Francis, 2002.

[4] Blundell, T.L. and Johnson, L.N. Protein Crystallography; Academic Press Inc. London, 1976.

[5] Henrick, K. and Thornton, J. Trends in Biochem. Sci., 1998, **23**, 358–361.

[6] Xu, Q., Canutescu, A.A.A., Wang, G., Shapovalov, M., Obradovic, Z. and Dunbrack, R.L.L. J. Mol. Biol., 2008, **381**, 487–507.

[7] Argos, P. Protein Eng., 1988, **2**, 101–113.

[8] Janin J. and Chothia, C. J. Biol. Chem., 1990, **265**, 16027-16030.

[9] Jones, S. and Thornton, J.M. Prog. Biophys. Molec. Biol., 1995, **63**, 31–65.

[10] Miller, S. Protein Eng., 1989, **3**, 77–83.

[11] Padlan, E.A. Proteins: Struct.Funct.Genet., 1990, **7**, 112–124.

[12] Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A. and Thornton, J.M. 2006, Proteins, **62**, 479–488.

[13] Gutteridge, A. and Thornton, J.M. Trends Biochem Sci., 2005, **30**, 622–629.

[14] Gutteridge, A., Bartlett, G.J. and Thornton, J.M. J.Mol.Biol., 2003, **330**, 719–734.

[15] Tsai, C.J., Lin, S.L., Wolfson, H. and Nussinov, R. Critical Reviews in Biochem. Mol. Biol., 1996, **31**, 127–152.

[16] Lo Conte, L., Chotia, C. and Janin, J. J.Mol.Biol., 1999, **285**, 2177–2198.

[17] Chakrabarti, P. and Janin, J. Proteins, 2002, **47**, 334–343.

[18] Keskin, O., Tsai, C.J., Wolfson, H. and Nussinov, R. Protein Sci., 2004, **13**, 1043–1055.

[19] Ogmen, U., Keskin, O., Aytuna A.S., Nussinov, R. and Gursoy, A. Nucl. Acids Res., 2005, **33**, W331–W336.

[20] Ponstingl, H., Henrick, K., and Thornton, J. Proteins, 2000, **41**, 47–57.

[21] Ponstingl, H., Kabir, T. and Thornton, J. J. Appl. Cryst., 2003, **36**, 1116–1122.

[22] Krissinel, E. and Henrick, K. J. Mol. Biol., 2007, **372**, 774–797.

[23] Tsai, C.J., Kumar, S., Ma, B. and Nussinov, R. Protein Sci., 1999, **8**, 1181–1190.

[24] Ma, B., Kumar, S., Tsai, C.J. and Nussinov, R. Protein Eng., 1999, **12**, 713–720.

[25] Tsai, C.J., Ma, B. and Nussinov, R. Proc. Natl. Acad. Sci. USA, 1999, **96**, 9970–9972.

[26] Kumar, S., Ma, B., Tsai, C.J., Sinha, N. and Nussinov, R. Protein Sci., 2000, **9**, 10–19.

[27] Boehr, D.D. and Wright, P.E. Science, 2008, **320**, 1429–1430.

[28] Lange, O.F., Lakomek, N.A., Fares, C., Schröder, G.F., Walter, K.F.A., Becker, S. Meiler, J., Grubmüller, H., Griesinger, C. and de Groot, B.L. Science, 2008, **320**, 1471–1475.

[29] Cavanagh, J., Fairbrother, W.J., Palmer III, A.G. and Skelton N.J. Protein NMR Spectroscopy; Academic Press, 1996.

[30] Frank, J. Three-Dimensional Electron Microscopy of Macromolecular Assemblies; New York: Oxford University Press, 2006.

[31] Feigin L.A. and Svergun D.I. Structure Analysis by Small Angle X-ray and Neutron Scattering; New-York: Plenum press, 1987.

[32] Svergun, D.I. and Koch, M.H.J. Cur.Opin.Struct.Biol., 2002, **12**, 654-660.

[33] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. Nucleic Acids Res., 2000, **28**, 235–242.

[34] Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. Genome Biol., 2000, **1**, 1–37.

[35] Katchalski-Katzir, E., Shariv, I. Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I. Proc. Natl. Acad. Sci. USA, 1992, **89**, 2195–2199.

[36] Ehrlich, L. P. and Wade, R. C. Rev. Comp. Chem., 2001, **17**, 61–97.

[37] Lengauer, T. and Rarey, M. Curr. Opin. Struct. Biol., 1996, **6**, 402–6.

[38] Chen, R., Li, L. and Weng, Z. Proteins, 2003, **52**, 80–87.

[39] Li, L., Chen, R. and Weng, Z. Proteins, 2003, **53**, 693–707.

[40] Sternberg, M. J., Gabb, H. A., Jackson, R. M., Moont, G. Methods Mol. Biol., 2000, **143**, 399–415.

[41] Smith, G. R. and Sternberg, M. J. Curr. Opin. Struct. Biol., 2002, **12**, 28–35.

[42] Gabb, H. A., Jackson, R. M. and Sternberg, M.J. J. Mol. Biol., 1997, **272**, 106–120.

[43] Mandell, J.G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I. and Ten Eyck L. F. Protein Eng., 2001, **14**, 105–113.

[44] Ritchie, D. W. and Kemp, G. J. Proteins, 2000, **39**, 178–194.

[45] Palma, P. N., Krippahl, L., Wampler, J. E. and Moura, J. J. Proteins, 2000, **39**, 372–384.

[46] Taylor, J. S. and Burnett, R. M. Proteins, 2000, **41**, 173–191.

[47] Chen, R. and Weng, Z. Proteins, 2002, **47**, 281–294.

[48] Camacho, C. J. and Vajda, S. Proc. Natl. Acad. Sci. USA, 2001, **98**, 10636–10641.

[49] Cherfils, J. and Janin, J. Curr. Opin. Struct. Biol., 1993, **3**, 265–269.

[50] Halperin, I., Ma, B., Wolfson, H. and Nussinov, R. Proteins, 2002, **47**, 409–443.

[51] Garzón, J. I., Kovacs, J. A., Abagyan, R. and Chacón, P. Bioinformatics, 2007, **23**, 427–433.

[52] Abagyan, R. and Totrov, M. Curr. Opin. Chem. Biol., 2001, **5**, 375–382.

[53] Shoichet, B. K., Kuntz, I. D. and Bodian, D. L. J. Comp. Chem., 2004, **13**, 380–397.

[54] Meng, E. C., Shoichet, B. K. and Kuntz, I. D. J. Comp. Chem., 2004, **13**, 505–524.

[55] Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. J. Comp. Chem., 1998, **19**, 1639–1662.

[56] Ruvinsky, A.M. and Kozintsev, A. V. J. Comp. Chem., 2005, **26**, 1089–1095.

[57] Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. and Weng, Z. Proteins, 2005, **60**, 214–216.

[58] Janin J., Henrick K., Moult J., Eyck L. T., Sternberg M. J., Vajda S., Vakser I., Wodak S. J, Proteins, 2003, **52**, 2–9.

[59] Vajda, S. Proteins, 2005, **60**, 176–180.

[60] Frigo, M. and Johnson, S. G. Proceedings of the IEEE, 93, **2005**, 216–231.

[61] Krissinel, E. and Henrick, K. Acta Cryst. D, 2004, **60**, 2256–2268.

[62] Lee, H.S., Kim, M.S., Cho, H.S., Kim, J.I., Kim, T.J., Choi, J.H., Park, C., Lee, H.S., Oh, B.H. and Park, K.H. J. Biol. Chem., 2002, **277**, 21891–21897.

[63] Osipiuk, J., Quartey, P., Moy, S., Collart, F. and Joachimiak, A. Protein Databank entry 1XR4, 2004, , unpublished.

[64] Mouilleron S., Badet-Denisot M. A. and Golinelli-Pimpaneau B. J. Biol. Chem., 2006, **281**, 4404–4412.

[65] Malashkevich, V. N., Strokopytov, B.V., Borisov, V.V., Dauter, Z., Wilson, K.S. and Torchinsky, Y.M. J. Mol. Biol., 1995, **247**, 111–124.

[66] Bell, C.E. and Eisenberg, D. Biochemistry, 1997, **36**, 481–488.

[67] Jabeen, T., Sharma, S., Singh, N., Bhushan, A. and Singh, T.P. Acta Cryst. D, 2005, **61**, 1107–1115.

[68] Janin, J., Miller, S., and Chothia, C. J. Mol. Biol., 1988, **204**, 155–164.

[69] Janin, J. Nat. Struct. Biol., 1997, **4**, 973–974.

[70] Janin, J. and Rodier, F. Proteins, 1995, **23**, 580–587.

[71] Gomperts, B. D., Kramer, I. M., Tatham, P. E. R. Signal transduction; Academic Press, 2002. .

[72] Brown, K., Nurizzo, D., Besson, S., Shepard, W., Moura, J. et al. J. Mol. Biol., 1999, **289**, 1017–1028.

[73] Doyle, M. L., Gill, S. J. and Cusanovich, M. A. Biochemistry, 1986, **25**, 2509–2516.

[74] Ren, Z., Meyer, T. and McRee, D. E. J. Mol. Biol., 1993, **234**, 433–445.

[75] Huang, D. B., Huxford, T., Chen, Y. Q. and Ghosh, G. Structure, 1997, **5**, 1427–1436.

[76] Sengchanthalangsy, L. L., Datta, S., Huang, D. B., Anderson, E., Braswell, E. H. and Ghosh, G. J. Mol. Biol., 1999, **289**, 1029–1040.

[77] Lu, H. S., Chang, W. C., Mendiaz, E. A., Mann, M. B., Langley, K. E. and Hsu, Y. R. Biochem. J., 1995, **305**, 563–568.

[78] Hsu, Y. R., Wu, G. M., Mendiaz, E. A., Syed, R., Wypych, J., Toso, R. et al. J. Biol. Chem., 1997, **272**, 6406–6415.

[79] Bianchet, M. A., Ahmed, H., Vasta, G. R. and Amzel, L. M. Proteins, 2000, **40**, 378–388.

[80] Blundell, T. L., Burke, D. F., Chirgadze D., Dhanaraj, V., Hyvnen, M., Innis, C. A., Parisini, E., Pellegrini, L. Sayed, M. and B. Lynn Sibanda, B. L. Biol. Chem., 2000, **381**, 955–959.

[81] Darling, P. J., Holt, J. M. and Ackers, G. K. Biochemistry, 2000, **39**, 11500–11507.

[82] Pan X. and Heitman J. Mol Cell Biol., 2002, **22**, 3981–3993.

[83] Jianpeng M. and Karplus M. PNAS, 1997, **994**, 11905–11910.

[84] Alattia, J. R., Ames, J. B., Porumb, T., Tong, K. I., Heng, Y. M., Ottensmeyer, P. et al. FEBS Letters, 1997, **417**, 405–408.

[85] Waas, W. F. and Dalby, K. N. J. Biol. Chem., 2002, **277**, 12532–12540.

[86] Cho K.-I., Lee K., Lee K. H., Kim D., and Lee D. Proteins, 2006, **65**, 593–606.

[87] Johannes, R. Nature Rev. Cancer, 2007, **7**, 202–211.

[88] Bonet, J., Caltabiano, G., Khan, K. A., Johnston, M. A., Corb, C., Gomez, A., Rovira, X., Teyra, J. and Vill-Freixa, J. Proteins, 2006, **63**, 65–77.

[89] Ansari, S. and Helms, V. Proteins, 2005, **61**, 344–355.

[90] Nooren, M. A. and Thornton, J. M. J. Mol. Biol., 2003, **325**, 991–1018.

[91] Schnarr, A. A. and Khosla, C. ACS Chem. Biol., 2006, **1**, 679–680.

[92] Vaynberg, J. and Qin, J. TRENDS in Biotechnology, 2006, **24**, 22–27.

[93] Fuentes, M., Mateo, C., Pessela, B. C. C., Guisan, J. M. and Fernandez-Lafuente, R. Proteomics, 2006, **5**, 4062–4069.

[94] Boelens, W., Scherly, D., Beijer, R. P., Jansen, E. J., Dathan, N. A., Mattaj, I. W. and van Venrooij, W. J. Nucleic Acids Res., 1991, **19**, 455–460.

[95] Ceres, P. and Zlotnick, A. Biochemistry, 2002, **41**, 11525–11531.

[96] Buts, L., Dao Thi, M. H., Loris, R., Wyns, L., Etzler, M. and Hamelryck, T. J. Mol. Biol., 2001, **309**, 193–201.

[97] Nyfeler, B., Michnick, S. W. and Hauri, H.-P. PNAS, 2005, **102**, 6350–6355.

[98] Hamelryck, T. W., Moore, J. G., Chrispeels, M. J., Loris, R. and Wyns, L. J. Mol. Biol., 2000, **299**, 875–883.

[99] Fuentes, M., Mateo, C., Pessela B. C., Batalla P., Fernandez-Lafuente, R. and Guisn, J. M. J. Chromatogr. B, 2007, **849**, 243–250.

[100] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T. and Tramontano, A. Proteins, 2007, **69**, 3–9.