doi:10.1016/j.jmb.2005.05.066 J. Mol. Biol. (2005) 351, 431–442

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Limit of Accuracy of Protein Modeling: Influence of Crystal Packing on Protein Structure

Eran Eyal*, Sergey Gerzon, Vladimir Potapov, Marvin Edelman and Vladimir Sobolev*

Department of Plant Sciences Weizmann Institute of Science 76100, Rehovot, Israel

**Corresponding authors*

The size of the protein database (PDB) makes it now feasible to arrive at statistical conclusions regarding structural effects of crystal packing. These effects are relevant for setting upper practical limits of accuracy on protein modeling. Proteins whose crystals have more than one molecule in the asymmetric unit or whose structures were determined at least twice by X-ray crystallography were paired and their differences analyzed. We demonstrate a clear influence of crystal environment on protein structure, including backbone conformations, hinge-like motions and side-chain conformations. The positions of surface water molecules tend to be variable in different crystal environments while those of ligands are not. Structures determined by independent groups vary more than structures determined by the same authors. The use of different refinement methods is a major source for this effect. Our pair-wise analysis derives a practical limit to the accuracy of protein modeling. For different crystal forms, the limit of accuracy (C^{α}, root-mean-square deviation (RMSD)) is ~0.8 Å for the entire protein, which includes ~0.3 Å due to crystal packing. For organized secondary elements, the upper limit of C^{α} RMSD is 0.5–0.6 Å while for loops or protein surface it reaches 1.0 Å. Twenty percent of exposed side-chains exhibit different χ_{1+2} conformations with approximately half of the effect also resulting from crystal packing. A web based tool for analysis and graphic presentation of surface areas of crystal contacts is available (http://ligin.weizmann.ac.il/cryco).

Keywords: PDB; crystal contacts; protein flexibility; structure refinment

Introduction

The densely packed environment of globular proteins in crystal structures is likely to affect protein structure. Crystal contacts bury a significant portion of the solvent accessible surface of a protein[1–4] and this might induce structural changes. What are the characteristics and the extent of such changes? A direct way to approach this question would be to compare X-ray structures and those modeled by NMR (see, for example Smith *et al*.[5]). Unfortunately, there are relatively few examples of proteins resolved by both methods, nor is it yet clear how best to compare a multitude of NMR models with a single X-ray model. An indirect way is to analyze differences between structures of the same protein in different crystal environments (i.e. when the arrangement of molecules in the crystal lattice is different). This has been done for several small datasets[6–8] or for several specific proteins.[4,9–19] In some of these, different crystal packing resulted in rigid body motion of large structural units[9,17] or loop conformational changes.[18] In most others, crystal packing had a role in local structural differences. Sometimes, the structural difference was much larger than that originating from point mutation.[11]

The few database studies carried out regarding the influence of crystal packing on protein structure deal mainly with side-chain conformation. Bower *et al*.[20] examined the side-chain angle differences between lysozyme crystallized in different space groups in order to derive an upper limit for

Abbreviations used: PDB, Protein Data Bank; SSE-complete, same structural environment; DSE-complete, different structural environment; SSE-da, same structural environment, different authors; DSE-da, different structural environment, different authors.

E-mail addresses of the corresponding authors: vladimir.sobolev@weizmann.ac.il; eran.eyal@weizmann.ac.il

prediction accuracy of side-chains in modeling programs. They found that about 25% of the lysozyme residues differ in χ_1 by $>40°$ in different crystal environments, and about 40% differ in χ_1 or χ_2. Jacobson *et al.*[21] found that side-chains with close intermolecular contacts tend to have different conformations more often if their crystal environment is different. This might result in an underestimation of predictive accuracy, especially for surface residues. Moreover, it is not clear if regions of protein structures that participate in intermolecular crystal contacts are less or more mobile than other surface regions, since it was variously claimed that the average *B*-factor is larger in the region of contacts[11] as well as the opposite.[2]

The growth of the Protein Data Bank (PDB) database makes it now feasible to arrive at statistical conclusions regarding structural effects of crystal packing. Here, we analyze proteins whose crystals have more than one molecule in the asymmetric unit or whose structures were determined at least twice by X-ray crystallography. The comparison of different structures of the same protein, in identical or different structural environments, is the main tool available for examining the amount of structural variability associated with crystal packing. However, a major problem is that in almost all cases there is some degree of dependence between structures resolved more than once. The potential limits of accuracy of structure prediction as evaluated by crystal structure comparisons is discussed.

Results

A visual example of the crystal packing phenomenon we are addressing is illustrated in Figure 1 using basic fibroblast growth factor as a case in point. At least two different crystal forms of this protein exist.[22,23] In these two crystal forms, different regions of the protein surface are shown to be involved in crystal contact (Figure 1(a)). In fact, in the case of pancreatic ribonuclease and its six crystal forms, it was found that almost any surface residue can be involved in crystal contact in one or other of the forms.[10] Thus, two structures of the same protein, resolved in different crystal forms, are likely to demonstrate greater diversity between them than two structures of the protein resolved in the same crystal form. Figure 1(b) demonstrates this for basic fibroblast growth factor.

In order to statistically investigate the effect of crystal environment on protein structure several datasets (Table 1) composed of pairs of PDB chains having the same sequence were constructed as detailed in Methods. Sets were differentiated based not only on structural environment but also on the source of data. Complete datasets of paired structures with the same, or different, structural environment (i.e. SSE-complete, DSE1-complete, DSE2-complete) were used to maximize statistical significance. In a SSE-complete pair, the two protein

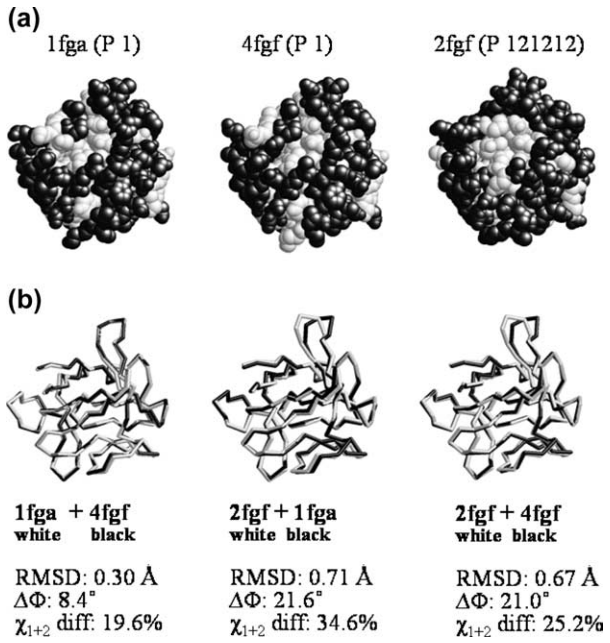

Figure 1. Different crystal structures of basic fibroblast growth factor. Two of the structures (PDB files 1fga and 4fgf) were resolved in the same crystal form (triclinic space group *P* 1).[22] The third structure (2fgf) was obtained from a crystal of a different space group (orthorhombic space group *P* 21 21 21).[23] (a) Space-filled structure models. Atoms of residues forming crystal contacts are shown in black. The majority of surface residues are in contact; however, a different set of residues is involved in crystal contacts for each structure. Note that the contacts of structures derived from different crystal forms differ to a greater degree than the contacts from within the same form. (b) Backbone superimposition of structure pairs. Note that the loop (residues 98–104) at the top of the images is positioned differently in different crystal forms. Parameters of global comparison are shown: C^{α} RMSD values, $\Delta\Phi$ (mean angle difference) and χ_{1+2} diff. (percentage of side-chains differing in χ_1 or χ_2 by more than 60°). Structural superpositions were made using MultiProt.[46] Pictures were created using RasMol.[47]

structures have the same crystal form and the protein molecules have the same environment. Therefore, differences in the two structures can be considered mainly as inaccuracy in the structure

Table 1. Datasets of structure pairs used in this study

Description	Number of pairs
Same structural environment (SSE-complete)	404
Different structural environment (DSE1-complete)	107
Different structural environment (DSE2-complete)	148
Same structural environment, different authors (SSE-da)	45
Different structural environment, different authors (DSE2-da)	43

See http://ligin.weizmann.ac.il/~eyale/cryco/datasets

determination. In DSE-complete datasets, the environment of the two molecules is different. In DSE1-complete the two structures are taken from the same crystal while in DSE2-complete they are from different ones. By comparing the results for these two datasets we seek to obtain a measure of the influence of crystal packing *versus* crystallization conditions such as pH, temperature, ligand occupancy, etc. The datasets of protein pairs resolved by different authors (i.e. SSE-da, DSE2-da) were used in an attempt to maximize data independence.

B-factor analysis

B-factor analysis was carried out in two ways to assess the mobility of regions involved in crystal contact. The first analysis probed SSE-complete (the largest data set) for normalized *B*-factor (B' values), comparing surface exposed atoms of residues having crystal contacts with those that do not. The results are shown in Figure 2. The difference between the average B' values for residues with (0.48), or without (0.89), crystal contact is highly significant ($p<0.002$). In the second approach, we exploited the DSE2-complete dataset to compare B' values of identical atoms between the two members of each pair. Here, only atoms in crystal contact in one structure, but not the other, were used. The mean value of $B'_{\text{contact}}-B'_{\text{no contact}}$ for the more than 30,000 atoms in the dataset was -0.28, which differs significantly from zero ($p<10^{-5}$). Both methods indicate that, as anticipated, exposed atoms participating in crystal contacts are less mobile (lower *B*-factor) than atoms without contacts.

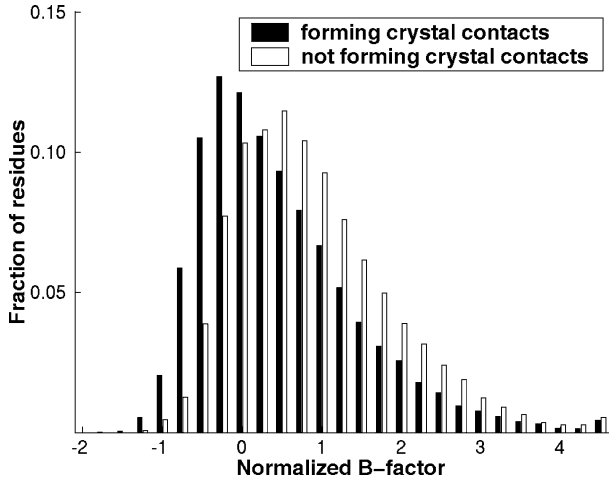

Figure 2. The *B*-factor at the protein surface. Residues from one member of each pair in the SSE-complete dataset having solvent accessibility >40% were separated into two subsets: those forming crystal contacts (20,775 residues) and those that do not (12,231 residues). The histogram shows the distribution of the normalized *B*-factor (B') averaged for each residue. B' is presented on the *x*-axis in bins of 0.25. The last bin represents all cases with B' larger than 4.50.

Backbone structural variability

The influence of the crystal environment on the overall protein structure was roughly estimated by superposition of pair members (minimization of the RMSD value of the C^{α} atoms). Figure 3 is a histogram of the average C^{α} RMSD values obtained for the complete datasets. It is clear that structures of DSE2-complete exhibit a higher tendency for backbone change (C^{α} RMSD = 0.83 Å) than those of SSE-complete (C^{α} RMSD = 0.30 Å), while those of DSE1-complete exhibit an intermediate level (C^{α} RMSD = 0.57 Å). The differences in C^{α} RMSD are significant to $p<10^{-5}$. There were three outliers (0.5%) in the DSE2 dataset with a C^{α} RMSD value larger than 10 Å; these were not considered in the statistics.

The average C^{α} RMSD value for paired structures resolved in the same crystal environment was larger when structures were resolved by different authors (SSE-da, 0.50 Å *versus* SSE-complete, 0.30 Å; $p<10^{-4}$). "Different authors" is defined as no author in common in the AUTHOR field of the two PDB files. While the average C^{α} RMSD difference between DSE2-da and SSE-da is also clearly significant ($p<0.01$), the authorship phenomenon is neutral for pair members resolved in different structural environments (Table 2).

Intuitively, when the environment is different in a pair of files, exposed regions of the protein that are not in crystal contact should be structurally more similar than regions in contact. Indeed, we found that for DSE1-complete and DSE2-complete, residues in contact with other molecules in the crystal exhibit a slightly greater spatial deviation than exposed regions without such contacts (C^{α} RMSD difference of ~0.2 Å, $p<10^{-3}$). However, no difference is seen for SSE-complete pairs.

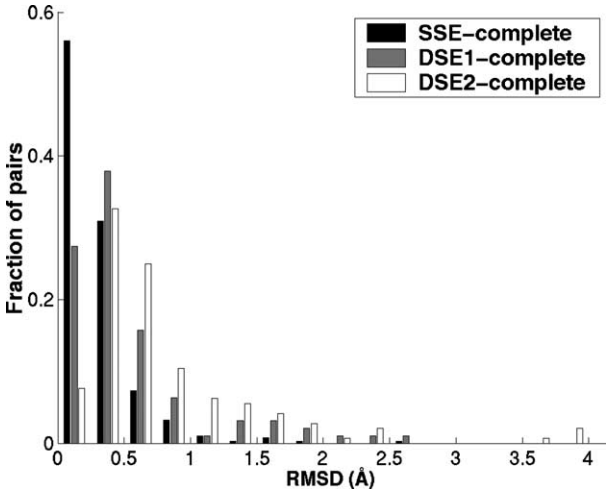

Figure 3. Distribution of C^{α} RMSD values of protein pairs from the SSE-complete, DSE1-complete and DSE2-complete datasets. C^{α} RMSD is presented on the *x*-axis in bins of 0.25 Å. The last bin represents all cases with C^{α} RMSD values more than 3.75 Å.

Table 2. Average C^{α} RMSD (Å) of secondary structure elements

Dataset	All residues[a]	Alpha helix	Beta strand	Coil region
SSE-complete	0.30±0.01 (0.23)	0.22±0.01	0.18±0.01	0.39±0.02
DSE1-complete	0.57±0.05 (0.41)	0.38±0.05	0.25±0.05	0.70±0.07
DSE2-complete	0.84±0.06 (0.59)	0.69±0.06	0.49±0.05	1.04±0.08
SSE-different authors	0.50±0.07 (0.32)	0.29±0.04	0.20±0.03	0.67±0.11
DSE2-different authors	0.82±0.09 (0.60)	0.61±0.09	0.49±0.09	1.05±0.11

Values are given±standard error of the mean.
[a] Median values are given in parentheses.

The C^{α} RMSD value does not reveal the entire picture, since local displacements could affect the results in a non-proportional fashion. Therefore, an analysis of Φ and Ψ backbone angles in each pair was carried out. The average difference of Φ and Ψ angles is a softer index and more forgiving for local changes. The mean difference of Φ angles between pair members was found to be significantly larger ($p<10^{-5}$) for DSE2-complete (7°) and DSE1-complete (6.5°) pairs as compared to the SSE-complete ones (<4°). The results for Ψ angles were very similar (excluding proline). For both the DSE1-complete and the DSE2-complete datasets, contacting residues exhibit larger backbone angle differences than non-contacting ones (for example, for Φ angles the average differences are 9.7° and 8.5°, respectively; $p<10^{-5}$). Therefore, the results for backbone angle analysis and C^{α} RMSD are in agreement.

We determined the structural variability within domains (as defined by SCOP) *versus* that between domains by superimposing separately each domain and summing the squared deviations (equation (1)). This index, called *iRMSD*, reflects the intra-domain variability while the difference between the total *RMSD* and *iRMSD* (i.e. *hRMSD*) reflects the inter-domain variability (hinge-like motion). Figure 4 shows that the mean fraction of the total deviation derived from hinge-like motion is almost twice as great for DSE1-complete pairs than for SSE-complete ones, and almost three times greater for DSE2-complete pairs. This indicates that the environment influences the relative orientation of the domains. Such changes in orientation are a significant source of the variation between multi-domain protein structures resolved in different crystal forms.

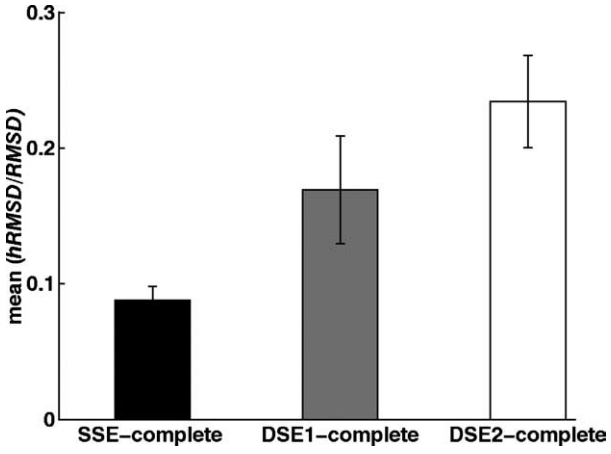

Figure 4. Relative domain (i.e. hinge-like) motions. The mean fraction of *hRMSD* out of the total *RMSD* is shown for the SSE-complete, DSE1-complete and DSE2-complete datasets.

The effect of secondary structure was examined. Assignments of secondary structures were taken from the PDB files. A difference in structural variability between DSE and SSE datasets was observed for all secondary elements (Table 2). As expected, it was largest for coiled regions (loops and turns) and smallest for the structured beta strands, with alpha helices being somewhat more variable.

Side-chain conformational changes

Accurate conformations of amino acid side-chains are important for docking, molecular design and understanding protein stability. Variations in amino acid side-chains can be estimated by measuring differences in χ_1 and χ_2 side-chain dihedral angles between members of a protein pair.[8,24,25] Side-chain conformations are especially sensitive to refinement methods and to the degree of independence in determining the structures.[7,21] In general, side- chain flexibility is much smaller for the SSE datasets than for the DSE ones. For example, the probability for χ_{1+2} (χ_1 or χ_2) conformational change is 0.05 for SSE-complete compared to 0.09 and 0.11 for DSE1-complete and DSE2-complete, respectively. Surprisingly, some increase in flexibility is observed for buried side-chains. For example, for χ_{1+2}, in SSE-complete and DSE2-complete the probabilities are 0.02 *versus* 0.04, respectively, although, as expected, the more significant effects are with exposed side-chains.

The highly exposed side-chains (accessibility >0.4) were divided into populations having or not having crystal contacts. The probability for conformational differences in χ_1 or χ_{1+2} between these populations is given in Table 3. In SSE pairs there is no significant difference between the two populations. Likewise, no differences were observed for DSE2-da or with χ_1 for the DSE2-complete dataset. In the other DSE sets, the difference between the contacting and non-contacting populations was small although statistically significant. On average, 70% of highly exposed residues from the complete datasets form crystal

Table 3. Summary of the fraction of exposed side-chains having different conformations

	Complete									Different authors					
	SSE-complete			DSE1-complete			DSE2-complete			SSE-da			DSE2-da		
	χ_1	χ_{1+2}	n[a]	χ_1	χ_{1+2}	n[a]	χ_1	χ_{1+2}	n[a]	χ_1	χ_{1+2}	n[a]	χ_1	χ_{1+2}	n[a]
Contact[b]	0.06	0.11	14675	0.15	0.23	1925	0.16	0.25	4404	0.12	0.20	1295	0.19	0.29	1178
No contact[c]	0.07	0.11	8050	0.11	0.18	842	0.14	0.22	1419	0.16	0.23	495	0.20	0.30	309

Different conformations are defined as dihedral angle changes of $>60°$.
[a] Sample size.
[b] Residues with crystal contacts.
[c] Residues without crystal contacts.

Table 4. Fraction of exposed side-chains from different amino acid types undergoing conformational change in complete datasets

	SSE-complete						DSE1-complete						DSE2-complete					
	Contact[a]			No contact[b]			Contact[a]			No contact[b]			Contact[a]			No contact[b]		
	χ_1	χ_{1+2}	n[c]	χ_1	χ_{1+2}	n[c]	χ_1	χ_{1+2}	n[c]	χ_1	χ_{1+2}	n[c]	χ_1	χ_{1+2}	n[c]	χ_1	χ_{1+2}	n[c]
Arg	0.06	0.10	1377	0.08	0.12	570	0.18	0.29	161	0.19	0.28	47	0.16	0.28	340	0.14	0.21	80
Asn[d]	0.05	0.07	1131	0.04	0.07	678	0.12	0.15	163	0.04	0.07	68	0.11	0.15	385	0.10	0.15	125
Asp	0.03	0.05	1592	0.03	0.06	1140	0.12	0.17	239	0.08	0.10	116	0.10	0.16	511	0.08	0.14	205
Cys	0.04	0.04	68	0.04	0.04	26	0.00	0.00	4	0.00	0.00	4	0.11	0.11	19	0.00	0.00	5
Gln	0.08	0.14	1169	0.07	0.12	587	0.15	0.31	137	0.09	0.24	55	0.16	0.32	346	0.16	0.26	112
Glu	0.09	0.16	2114	0.11	0.17	1325	0.21	0.31	277	0.11	0.18	136	0.24	0.37	601	0.19	0.33	205
His[d]	0.03	0.04	385	0.05	0.07	185	0.10	0.14	51	0.06	0.11	18	0.13	0.15	105	0.06	0.06	32
Ile	0.06	0.14	304	0.02	0.13	99	0.18	0.36	39	0.00	0.15	13	0.17	0.30	100	0.00	0.27	15
Leu	0.03	0.10	554	0.02	0.09	172	0.02	0.22	50	0.13	0.35	23	0.07	0.20	135	0.08	0.26	38
Lys	0.07	0.17	2140	0.07	0.15	1284	0.16	0.31	340	0.09	0.20	140	0.19	0.38	639	0.16	0.32	224
Met	0.08	0.13	186	0.05	0.07	42	0.18	0.45	11	0.14	0.43	7	0.15	0.29	41	0.00	0.00	5
Phe	0.02	0.02	264	0.01	0.02	85	0.10	0.10	31	0.00	0.00	11	0.06	0.09	67	0.00	0.00	10
Ser	0.12	0.12	1255	0.10	0.10	751	0.18	0.18	163	0.24	0.24	80	0.22	0.22	473	0.23	0.23	162
Thr	0.05	0.05	1145	0.07	0.07	757	0.13	0.13	167	0.13	0.13	86	0.11	0.11	365	0.13	0.13	142
Trp	0.00	0.04	94	0.05	0.05	20	0.00	0.00	9	0.00	0.00	2	0.06	0.06	34	0.00	0.00	7
Tyr	0.02	0.02	365	0.01	0.02	108	0.08	0.11	36	0.00	0.00	9	0.07	0.07	110	0.00	0.00	14
Val	0.06	0.06	532	0.06	0.06	221	0.15	0.15	47	0.19	0.19	31	0.16	0.16	133	0.08	0.08	38

Defined as dihedral angle changes of $>60°$.
[a] Residues with crystal contacts.
[b] Residues without crystal contacts.
[c] Sample size.
[d] The fraction of χ_{1+2} change is underestimated for these two residues (see Methods).

contacts. This represents ~20% of total number of protein residues, similar to that reported by others.[11] Table 4 summarizes the results for individual amino acid types for the complete datasets. The side-chains of Lys, Gln and Glu, large polar amino acids which are usually the most flexible,[25] appear to be more affected by crystal contacts.

Pair member dependency

A major issue in this study is the degree of dependence between pair members in our datasets. As implied by others,[7] and shown in Tables 2 and 3, differences between structures resolved in different laboratories are more evident than those resolved in the same laboratory. Identifying the source of these differences is important. We found that in 211 of the 640 pairs in the SSE and DSE2 datasets, the structure of one member of a pair was determined at least in part based on the structure of the other; or, both were determined based on a common third structure. We found very few pairs that were clearly resolved independently. In the majority of cases, the PDB documentation was insufficient for us to classify.

If we assume that the 211 detectibly dependent cases and the very few independent ones are representative of our datasets as a whole, then the question remains why the statistics for different authors (da datasets) differ from those of all authors (complete datasets). One plausible hypothesis is that different authors use different programs for structure refinement more often than the same authors. The section of the PDB file dealing with refinement is regularly populated and, therefore, can be used to establish whether pair members were refined using substantially the same or different programs. Analysis of our datasets validates our hypothesis: less than 30% of pairs resolved by different authors were refined using the same program for both pair members, compared to more than 70% for all authors. Table 5 further reveals that pair members refined by the same program are structurally more similar than those refined by different programs. The differences for the three parameters shown are significant (p values <0.001). This establishes that refinement is an important (but not exclusive) component of the "different author" effect. We note that while different refinement approaches for pair members may result in structures with different conformations, each is likely to correspond to a different local minimum and each could exist in crystallized form.[26]

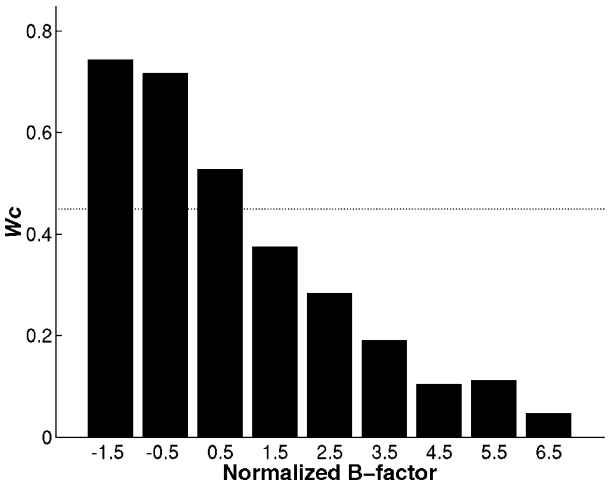

Figure 5. Water correspondence index (W_c) between pair members as a function of the normalized B-factor (B') in the DSE2-complete dataset. Dotted line, overall W_c.

Position of associated ligands

Binding site flexibility and positional variability of associated ligands between pair members in the complete datasets were examined following superposition of the binding site residues. Ligands were divided into two groups: metal ions and other ligands (non-ions). For non-ions, binding site RMSD was significantly higher in the DSE datasets then in the SSE one; however, the ligand RMSD values were not significantly different (Table 6). This is in agreement with the frequent functionality of proteins in crystals. The results obtained with metal ions were inconsistent. For DSE-da and SSE-da pairs, the data were insufficient for statistical comparison, and likewise for ligands having or not having crystal contacts. The average binding-site C^{α} RMSD value between pair members is small

Table 5. Characteristics of SSE-complete subsets

Dataset	Subset	Number of pairs	C^{α} RMSD[a] (Å)	$\Delta\phi$[b] (°)	χ_{1+2}[c]
SSE-complete	Full dataset[d]	404[e]	0.30[f]	4.5	0.10
	Same refinement program	306	0.27	4.1	0.09
	Different refinement program	92	0.39	5.9	0.15
SSE-da	Full dataset[d]	45[e]	0.50[f]	6.6	0.20

[a] Average C^{α} RMSD between pairs.
[b] Average of the ϕ difference in each pair.
[c] Average fraction of residues having different χ_1 or χ_2 conformation.
[d] Data for SSE-complete and SSE-da are listed for comparison.
[e] Taken from Table 1.
[f] Taken from Table 2.

Table 6. RMSD of ligands and their binding sites

Dataset	Ligand type	Average ligand RMSD (Å)	Average binding site RMSD (Å)
SSE-complete	Metal ion	0.26±0.04	0.12±0.01
	Non-ion	0.37±0.02	0.11±0.01
DSE1-complete	Metal ion	0.17±0.03	0.10±0.01
	Non-ion	0.43±0.05	0.24±0.05
DSE2-complete	Metal ion	0.23±0.04	0.22±0.03
	Non-ion	0.43±0.05	0.23±0.02

Pair member C^{α} atoms of binding site residues were superimposed and the RMSD value of the ligand between pair members was then measured.

Table 7. Water correspondence index (W_c)

Dataset		Contact[a]	No contact[b]
Complete	SSE	0.69±0.01	0.72±0.01
	DSE1	0.42±0.01	0.55±0.01
	DSE2	0.34±0.01	0.55±0.01
Different authors	SSE	0.56±0.01	0.59±0.01
	DSE2	0.29±0.01	0.44±0.02

Values are given±standard error of the mean.
[a] Water molecules that contact two protein molecules in the crystal.
[b] Water molecules that have contact only with a single protein molecule in the crystal.

(~0.2 Å; Table 6). While this indicates rigidity of the bound state it does not negate the known flexibility of the binding site during the binding process.

Differences in the water shell

The variability of the water shell around the protein was measured by a correspondence index, W_c, defined in equation (4). W_c indicates the fraction of resolved water molecules that appear at about the same place in the two members of a pair. The entire water shell is significantly more conserved for pairs having the same structural environment (SSE pairs, W_c=0.6–0.7) as compared to those resolved in different environments (DSE pairs, W_c=0.4), irrespective of authorship. However, W_c is strongly dependent on B' values (Figure 5). Thus, the B' value in the PDB files is a good indication of whether water molecule keeps its position in different environments.

The difference in W_c between water molecules involved in crystal contact (i.e. bridging two protein molecules in the crystal), and those not involved, is illustrated in Table 7. The data show that the former are less positionally conserved than the latter in the DSE pairs. There is a much smaller difference between the two populations in the SSE pairs. Interestingly, crystal contacts do not stabilize water molecules that share the same crystal environment, as might have been be expected. It should also be noted that water molecules not involved in crystal contact in DSE pairs still are less positionally conserved than their equivalents in SSE pairs.

A tool for analyzing and visualizing crystal contacts

Our program for constructing the crystal environment and analyzing the atomic contacts is available†. The site builds coordinate files, in a PDB format, for the unit cell as well as for the complete crystal environment of one molecule. Detailed analyses of atomic contacts are available based on contact surface areas using the CSU program.[27] As well, interactive visualization options and coordinate output files are provided. Our site complements and supplements existing tools, such as the WHAT IF web server that lists the crystal contacts,[28] and the xpack VRML-based program.[29]

† http://ligin.weizmann.ac.il/cryco

Discussion

Crystal packing effects

The results of our database study show that, on average, crystal environment influences protein structure. A variety of parameters not previously examined at a statistical level were used. These include: local backbone and hinge-like motions, side-chain flexibility, B' values, positional conservation of associated water and ligands. The variability between proteins resolved in different crystal forms was clearly higher than between those having the same form. However, despite the different packing environments, structure pairs were very similar by and large. Only rarely was an overall C^{α} RMSD of more than 2 Å observed between structures determined in different crystal environments.

Our B' value analyses (Figure 2) support the hypothesis that atoms, or residues, involved in crystal contacts are less mobile than others on the surface. A previous study on basic pancreatic trypsin inhibitor[11] did not report a propensity for residues in crystal contacts to have smaller *B*-factors, but this is not a counter-example. In that study, the authors compared *B*-factors at crystal contacts to those of the rest of the molecule that included a large number of residues buried in the protein interior.

Differences in B' values should be accompanied, in principal, by structural differences in contacting and non-contacting regions between pair members. Concerning backbone, the exposed residues of SSE datasets have the same level of structural variability irrespective of their contact state. In contrast, when crystal environments are different (DSE datasets), backbone differences in regions of crystal contact are larger than for non-contacting surface regions.

From the data in Table 3, we can deduce that the percentage of highly exposed residues forming crystal contacts is high (~70%). However, only a

minor part (~30%) of the solvent accessible surface is buried upon crystal formation.[2,10] This indicates that a large number of contacts have only moderate to low contact area. Might their inclusion in our definition of contact be biasing our results? An analysis of variability between pair members as a function of contact area showed no correlation. Moreover, we found variability to be more or less independent of crystal contact (Table 3). However, Jacobson *et al.*[21] reported a much greater structural variability of solvent exposed side-chains in crystal contact. We repeated our side-chain analysis with the set of proteins used by these authors and obtained the same result as they did (structural variability of 31% in contacting regions *versus* 20% in non-contacting ones for χ_{1+2}). Thus, the source of difference is in the datasets used by Jacobson *et al.*[21] and by us: 12 *versus* 148 proteins, none *versus* stringent homologous protein filtering, and differences in the percent of surface residues in crystal contact. It should be emphasized, however, that both studies found increased flexibility of side-chains in DSE datasets compared to SSE ones.

The reportage of water molecules in the PDB is known to be problematic. Therefore, our water correspondence index (W_c) most probably overestimates their mobility. Nonetheless, we note that the differences shown for the DSE pairs in Table 7 are highly significant ($p \ll 10^{-5}$) and we find no reason to suspect that the W_c indices of crystal contacting and non-contacting water molecules are differentially affected. Even water molecules not in crystal contact in the DSE datasets of Table 7 changed position with higher probability than those in SSE ones. This might imply that the entire hydrogen bond network around the protein is different in different crystal environments. In this context, it should be mentioned that most of the crystal contacts are formed by polar side-chains.[30] All considered, we speculate that positions of water molecules are significantly affected by crystal environment.

Our study may offer a hint for resolving a "chicken or egg"-type paradox for structural differences observed between identical proteins in different crystals. Is it small differences in protein structure, resulting from different solvent conditions, that causes a different arrangement of the crystal or, rather, incipient nucleation events that mold the protein to the growing crystal, leading to slightly different conformations in different crystal forms? According to the first hypothesis, the level of structural differences between members of a protein pair should be similar for contacting and non-contacting regions. However, our analysis for backbone conformations suggests that this is not the case (contacting regions are more structurally divergent). Therefore, we argue that the first hypothesis can be rejected as the sole interpretation. A combination of both hypotheses is, of course, also possible.

Accuracy of protein modeling

The results from this study may be applied to estimating a practical limit to the accuracy of protein modeling. This limit corresponds to the experimental inaccuracy in the determination of the crystal structure and the real difference between two structures caused by different crystallization conditions and/or crystal packing. The average C^α RMSD value for the SSE-complete dataset (0.30 Å; Table 2) gives a measure of differences due mainly to experimental inaccuracy. This average is comparable to values previously obtained for a few individual samples.[5,31] Similarly, the average C^α RMSD value for pair members of our SSE-da dataset (0.50 Å; Table 2) compares favorably with a comparable sampling of 13 PDB pairs (0.51 Å).[7]

The average C^α RMSD value for DSE1-complete (0.57 Å; Table 2) is significantly larger than for SSE-complete (0.30 Å). The difference (0.27 Å) represents the crystal packing effect, since DSE1 pair members are from the same crystal and therefore free from the effect of different crystallization conditions. The DSE1 C^α RMSD value can be compared with the results of Kleywegh[32] who reported a value of 0.46 Å for core C^α atoms of 476 pairs derived from single crystals. We consider the two results to be similar, since the core value included only superimposed C^α atoms within 3.5 Å one from another while the slightly higher DSE1 average includes all C^α atoms. Earlier, Chothia & Lesk[6] reported values ranging from 0.25 Å to 0.40 Å for a set of five pairs derived from single crystals. This range is lower than the average C^α RMSD value for our much larger DSE1 dataset, but is in line with the median (Table 2) and modal values (Figure 3). Finally, average and median C^α RMSD values that we obtained for DSE2-da (0.82 Å and 0.60 Å, respectively; Table 2) also compare favorably with three earlier reported pairs whose members were resolved independently (0.51 Å, 0.55 Å and 0.88 Å).[7]

We note that both in our study and that of Kleywegh[32] there are several cases with C^α RMSD >1.5 Å (e.g. 3sdp, 8fab, 1shf;[32] 1amc, 1eer, 1ba2 (from DSE1-complete)). In fact, analysis of Figure 3 reveals that ~25% of pair members in DSE datasets differ by 1.0 Å or more. Such large differences probably result from two intrinsically different conformations of a protein which both fit well in the crystal cell. Such crystallized conformers may represent a snapshot of some of the oligomerized conformers that exist in solution[33] under different conditions of pH, temperature, salt concentration, etc. However, current modeling approaches do not account for crystallization conditions or crystal packing effects, therefore accuracy of prediction is often effectively restricted to C^α RMSD values >1 Å.

Limits for prediction of side-chain conformations on the protein surface can also be estimated from our results. Given that exposed side-chains are highly sensitive to the refinement procedure, it is probably more accurate to derive such limits from the statistics of the complete datasets where the two members of each pair are usually resolved in the

same laboratory and, presumably, by the same refinement program. In both DSE-complete datasets, about 85% of exposed side-chains from pair members are found in similar χ_1 conformation and about 80% of them are found in similar χ_{1+2} conformation (Table 3). The remaining percentages represent uncertainty in rotamer designation, inherent flexibility and crystal packing. A comparison of the total fraction of side-chains undergoing conformational change in Table 3 suggests that about half of the remaining percentages are due to crystal packing. Current side-chain modeling methods[34] have a prediction accuracy of ~70% for χ_1 and ~60% for χ_{1+2} for exposed residues, leaving room for improvement. Likewise, a comparison of the conserved ligand positions between pair members having different crystal forms (~0.4 Å RMSD; Table 6, Non-ion) with the currently accepted accuracy of docking procedures (1.5–2.0 Å RMSD)[35] shows that there is room for improving docking methods as well.

Exposed side-chains not in crystal contact (i.e. not having a common contact surface area with atoms of other symmetry related molecules in the crystal) were, surprisingly, strongly affected by the crystal environment (e.g. for χ_{1+2}, 18% and 22% of DSE-complete side-chains changed conformation compared to 11% of SSE-complete ones (Table 3)). In fact, these side-chains are affected almost as much as those in crystal contact (e.g. 18% and 22% compared to 23% and 25%). This is even more apparent for the different author datasets. Our results imply that in addition to short-range interactions, long-range effects are involved. Indeed, when crystal contact was taken into account, our side-chain placement program[34] that utilizes short-range interaction (atomic contacts), only slightly improved predictions while inclusion of long-range terms (electrostatics and sophisticated solvation)[21] permitted increased improvement.

Methods

Datasets

Datasets for this study were created by first extracting all pairs of PDB[36,37] chains identical in sequence (February 2004 version) and whose structures were determined by X-ray crystallography to a resolution equal or better than 2.5 Å. Pairs were then separated into those whose partners share, or do not share, the same structural environment (determined from the space group, unit cell dimensions and atomic contacts between the molecules[27]).

We derived two lists of pairs having different structural environments. The first (DSE1) is composed of proteins with two molecules found in the asymmetric unit of a single crystal structure. To derive this list, we collected all multi-chain files that were not reported in the PDB as biological multimers. From this list we eliminated all cases of interface contact area >500 Å^2 to further reduce the possibility of including biological complexes. Pairs that exhibited a similar pattern of contacts in the crystal (>75% of contacts are the same) were also eliminated to ensure that the two structures were actually located in different crystal environments.

The second list (DSE2) is composed of paired structures whose members come from different crystals with different crystal forms. The crystal form was characterized by the space group, unit cell dimensions and atomic contacts. We considered only structures having a single chain in the asymmetric unit. This eliminated the hetero-biological complexes, but not homo-oligomers that have crystal symmetries. While we lost some potential examples by this filtering, it still allowed sufficient pairs for statistical analysis and automation.

In addition, a control list of pairs having the same structural environment (SSE) was constructed. This list is composed of pairs of structures obtained from different crystals having identical crystal forms.

Pair members in all lists were allowed to have the same or different ligands (determined from the HET field of the PDB files). Accepting different ligands provided a threefold increase in data and the results for such pairs were similar to those having identical ligands.

The PISCES site[38] was used to remove pairs that had a high *R*-factor (>0.30) in at least one pair member, or a sequence identity >25% to another pair. Although it is not clear what the identity threshold should be, it is apparent that some threshold is needed to eliminate overrepresentation of proteins with many mutant structures.

Two additional lists were derived in which the structures in each pair were resolved by "different authors" (i.e. no common author in the AUTHOR field for the two PDB files). Datasets in which pair members have different authorship are termed SSE-da and DSE2-da. When the author criterion was not applied, the datasets are termed SSE-complete, DSE1-complete and DSE2-complete. The sizes of the final datasets, and the URL where they can be found, are listed in Table 1.

The automatic procedure we applied in attempts to extract pairs resolved independently is as follows. The entire PDB was divided into groups of files. In each group there is one file, which, based on PDB documentation, was not reported as being determined based on another structure. All additional members of the group were resolved based on an existing member of the same group. The information for this procedure was taken from the "starting model" field of the REMARK200 line of the PDB file. Pair members in the datasets shown in Table 1 are dependent if they are in the same group. If pair members are in different groups they constitute a potential independent pair. For these pairs, we performed a manual check (e.g. scanned HEADERs for common publication, authorship, date of submission, special remarks) to assess the presumption.

The information in the REMARK3 line of the PDB file was used to decide if pair members were refined using basically the same or different programs. If several refinement programs were used for a single structure, we considered pair members to be refined by the same program as long as there was at least one common program applied to both. If two structures were refined by different program versions, they also were considered as being refined by the same program.

Structural environment

The structural environment was built in the following steps: (1) symmetry related molecules were created using the PDBSET program from the CCP4 suite;[39] (2) if necessary, all molecules were translocated to the same unit cell; (3) the 26 adjacent cells in the crystal lattice were

then constructed by translation; (4) any atom farther than 10 Å from the closest atom of the chosen central molecule was removed.

Solvent accessibility and contact analysis

Solvent accessible surface was calculated analytically using the Voronoi procedure.[40] Relative solvent accessibility is the solvent accessible surface divided by the theoretical maximum obtained from an extended GGXGG peptide. Unless otherwise indicated, residues with an accessibility value >0.2 are defined as exposed.

CSU software[27] was used to analyze contacts between symmetry related molecules in the crystal. CSU defines an atom of one molecule as being in contact with an atom of another if the distance between the two is less that the sum of their van der Waals radii plus two radii of a solvent molecule, and if there is no third atom between them.

Superposition and backbone structural changes

The SVD method[41] was used for superposition of protein structures by applying an existing implementation.[42] The positions of the C^{α} atoms were used for calculating overall protein or domain RMSD. For superimposed binding sites, only the positions of the C^{α} atoms of residues in contact with the ligand in at least one of the structures were included. Differences in backbone angles between the two pair members was established by measuring the mean difference of Φ and Ψ angles in a chosen set of residues.

Detection of hinge-like motion

iRMSD is defined as a function of the C^{α} RMSD values obtained by separate superimpositions of each of the domains found in the protein:

$$iRMSD = \sqrt{\frac{\sum_{j=1}^{k}(RMSD_j^2 \times N_j)}{\sum_{j=1}^{k} N_j}} \quad (1)$$

where k is the number of structural domains in the protein, $RMSD_j$ is the C^{α} RMSD of domain j, and N_j is the number of amino acid residues in domain j.

iRMSD is a quantitative indication for structural changes occurring within domains of two compared proteins, such as small loop motions. It can be easily proven that:

$$iRMSD \leq RMSD \quad (2)$$

where *RMSD* is the total C^{α} RMSD obtained by superimposing all the residues of the protein.

The difference between *RMSD* and *iRMSD*:

$$hRMSD = RMSD - iRMSD \quad (3)$$

reflects changes originating from the relative domain motion. Domain assignments were taken from the SCOP database.[43]

Side-chain flexibility

Side-chain flexibility was measured as the fraction of cases where χ_1 or χ_2 differ by more than 60° between the two structures of a pair. This value is well accepted under a variety of conditions.[7,8,20,21,24,25,44] In order to eliminate the strong influence of backbone changes on side-chains, residues that differed by more that 0.4 Å in C^{α}–C^{α} distance to any of their neighbors between the two files were not considered.

Asp, Tyr and Phe are symmetrical in their terminal dihedral angle. In addition, for His and Asn the position of the C, N or O atoms in some cases might be wrongly assigned, causing incorrect calculation of the associated dihedral angles.[8] For all of these cases, there are two possible values for the differences between the dihedral angles. The smaller of the two was always taken as the actual difference.

Ligand position

Binding sites occupied by identical ligands in each of the two structures of a pair were used for evaluating difference in ligand positions. The sites were superimposed using only the C^{α} atoms of the binding site residues. The binding site and ligand RMSD values were collected and used for analysis. No restriction was placed on ligand size. For symmetrical ligands such as PO_4, only the central atom was used for ligand RMSD calculation.

Water position

A "water correspondence" index was defined to indicate the spatial conservation of two sets of water molecules. The index indicates the fraction of well-defined water molecules (whose coordinates explicitly appear in the PDB file) that appear close in space in two compared water sets following superimposition of the protein pair (by RMSD minimization of their C^{α} atoms). The index has the following terms:

$$W_c = \tfrac{1}{2}\left(\frac{C_{ij}}{N_i} + \frac{C_{ji}}{N_j}\right) \quad (4)$$

where N_i is the total number of water molecules in set i (oxygen atoms) and C_{ij} is the number of the water molecules in subset i which are coupled within 1 Å in set j. We analyzed only water molecules present within a shell of radius 3.5 Å around the protein (including those whose coordinates do not explicitly appear in the PDB file but do exist in the crystal). Only water molecules in contact with residues whose C^{α} atoms in the superimposed pair members are less than 0.7 Å apart were included in this analysis.

B-factor

Normalization of the temperature factor (B factor)[45] in each file was done using the equation:

$$B' = \frac{B - \langle B \rangle}{\sigma(B)} \quad (5)$$

where $\langle B \rangle$ and $\sigma(B)$ are the mean and the standard deviation, respectively, of the B-factor values reported in the PDB files. When the analysis was performed at the residue level we considered the mean B' value of the atoms of each residue.

Acknowledgements

We thank Dr Zippora Shakked for valuable discussions.

References

1. Janin, J. & Rodier, F. (1995). Protein–protein interaction at crystal contacts. *Proteins: Struct. Funct. Genet.* **23**, 580–587.
2. Carugo, O. & Argos, P. (1997). Protein–protein crystal-packing contacts. *Protein Sci.* **6**, 2261–2263.
3. Valdar, W. S. J. & Thornton, J. M. (2001). Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* **313**, 399–416.
4. Kuglstatter, A., Oubridge, C. & Nagai, K. (2002). Induced structural changes of 7SL RNA during the assembly of human signal recognition particle. *Nature Struct. Biol.* **9**, 740–744.
5. Smith, L. J., Redfield, C., Smith, R. A. G., Dobson, C. M., Clore, G. M., Gronenborn, A. M. *et al*. (1994). Comparison of four independency determined structures of human recombinant interleukin-4. *Nature Struct. Biol.* **1**, 301–310.
6. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
7. Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**, 1811–1826.
8. Betts, M. J. & Sternberg, M. J. E. (1999). An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng.* **12**, 271–283.
9. Heinz, D. W., Priestle, J. P., Rahuel, J., Wilson, K. S. & Grutter, M. G. (1991). Refined crystal structures of subtilisin novo in complex with wild-type and two mutant eglins. *J. Mol. Biol.* **217**, 353–371.
10. Crosio, M. P., Janin, J. & Jullien, M. (1992). Crystal packing in six crystal forms of Pancreatic Ribonuclease. *J. Mol. Biol.* **228**, 243–251.
11. Kossiakoff, A. A., Randal, M., Guenot, J. & Eigenbrot, C. (1992). Variability of conformations at crystal contacts in BPTI represent true low-energy structures: correspondence among lattice packing and molecular dynamics structures. *Proteins: Struct. Funct. Genet.* **14**, 65–74.
12. Kishan, K. V. R., Zeelen, J. P., Noble, M. E. M., Borchert, T. V. & Wierenga, R. K. (1994). Comparison of the structures and the crystal contacts of trypanosomal triosephosphate isomerase in four different crystal forms. *Protein Sci.* **3**, 779–787.
13. Raghunathan, S., Chandross, R. J., Kretsinger, R. H., Allison, T. J., Penington, C. J. & Rule, G. S. (1994). Crystal structure of human class mu glutathione transferase GSTM2-2. Effects of lattice packing on conformational heterogeneity. *J. Mol. Biol.* **238**, 815–832.
14. Zhang, X. J., Wozniak, J. A. & Matthews, B. W. (1995). Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozymes. *J. Mol. Biol.* **250**, 527–552.
15. Cody, V., Galitsky, N., Luft, J. R., Pangborn, W., Rosowsky, A. & Blakley, R. L. (1997). Comparison of two independent crystal structures of human dihydrofolate reductase ternary complexes reduced with nicotinamide adenine dinucleotide phosphate and the very tight-binding inhibitor PT523. *Biochemistry*, **36**, 13897–13903.
16. Oki, H., Matsuura, Y., Komatsu, H. & Chernov, A. A. (1999). Refined structure of orthorhombic lysozyme crystallized at high temperature: correlation between morphology intermolecular contacts. *Acta Crystallog. sect. D*, **55**, 114–121.
17. Bertrand, J. A., Fanchon, E., Martin, L., Chantalat, L., Auger, G., Blanot, D. *et al*. (2000). Open structures of MurD: Domain movements and structural similarities with folylpolyglutamate synthetase. *J. Mol. Biol.* **301**, 1257–1266.
18. Taylor, P., Dornan, J., Carrello, A., Minchin, R. F., Ratajczak, T. & Walkinshaw, M. D. (2001). Two structures of cyclophilin 40: folding and fidelity in the TPR domains. *Structure*, **9**, 431–438.
19. Diener, J. L. (2003). Complex conformations and crystal contacts. *Nature Struct. Biol.* **10**, 494–494.
20. Bower, M. J., Cohen, F. E. & Dunbrack, R. L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* **267**, 1268–1282.
21. Jacobson, M. P., Friesner, R. A., Xiang, Z. X. & Honig, B. (2002). On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **320**, 597–608.
22. Eriksson, A. E., Cousens, L. S. & Matthews, B. W. (1993). Refinement of the structure of human basic fibroblast growth factor at 1.6 Å resolution and analysis of presumed heparin binding sites by selenate substitution. *Protein Sci.* **2**, 1274–1284.
23. Zhang, J. D., Cousens, L. S., Barr, P. J. & Sprang, S. R. (1991). Three-dimensional structure of human basic fibroblast growth factor, a structural homolog of interleukin 1B. *Proc. Natl Acad. Sci. USA*, **88**, 3446–3450.
24. Najmanovich, R., Kuttner, J., Sobolev, V. & Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins: Struct. Funct. Genet.* **39**, 261–268.
25. Eyal, E., Najmanovich, R., Edelman, M. & Sobolev, V. (2003). Protein side-chain rearrangement in regions of point mutations. *Proteins: Struct. Funct. Genet.* **50**, 272–282.
26. DePristo, M. A., de Bakker, P. I. W. & Blundell, T. (2004). Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure*, **12**, 831–838.
27. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E. & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
28. Rodriguez, R., Chinea, G., Lopez, N., Pons, T. & Vriend, G. (1998). Homology modeling, model and software evaluation: three related resources. *Bioinformatics*, **14**, 523–528.
29. Fu, T. Y. & Chen, Y. W. (1996). Visualization of macromolecular crystal packing using Virtual Reality Modelling Language (VRML). *J. Appl. Crystallog.* **29**, 594–597.
30. Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E. & Bell, J. A. (1997). Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins: Struct. Funct. Genet.* **28**, 494–514.
31. Ohlendorf, D. H. (1994). Accuracy of refined protein structures. II. Comparison of four independently refined models of human interleukin 1B. *Acta Crystallog. sect. D*, **50**, 808–812.
32. Kleywegt, G. J. (1996). Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallog. sect. D*, **52**, 842–857.
33. Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M. & Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature*, **433**, 128–132.
34. Eyal, E., Najmanovich, R., McConkey, B. J., Edelman,

M. & Sobolev, V. (2004). Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J. Comp. Chem.* **25**, 712–724.
35. Kellenberger, E., Rodrigo, J., Muller, P. & Rognan, D. (2004). Comparative evaluartion of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct. Funct. Genet.* **57**, 225–242.
36. Bernstein, F. K. & Williams, G. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
37. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
38. Wang, G. L. & Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
39. Collaborative Computational Project Number 4. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallog. sect. D*, **50**, 760–763.
40. McConkey, B. J., Sobolev, V. & Edelman, M. (2002). Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics*, **18**, 1365–1373.
41. Arun, K. S., Huang, T. S. & Blostein, S. D. (1987). Least-Squares fitting of two 3-D point sets. *IEEE Transact. Pattern Anal. Machine Intellig.* **9**, 699–700.
42. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1997). *Numerical Recipes in C*. Cambridge University Press, New York.
43. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structrural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
44. Zhao, S. R., Goodsell, D. S. & Olson, A. J. (2001). Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation. *Proteins: Struct. Funct. Genet.* **43**, 271–279.
45. Parthasarathy, S. & Murthy, M. R. N. (1997). Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci.* **6**, 2561–2567.
46. Shatsky, M., Nussinov, R. & Wolfson, H. J. (2002). MultiProt—a multiple protein structural alignment algorithm. *Lecture Notes Comput. Sci.* **2452**, 235–250.
47. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374–376.

Edited by M. Levitt

(Received 11 January 2005; received in revised form 26 May 2005; accepted 30 May 2005)

Available online 27 June 2005