PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Sequence-similar, structure-dissimilar protein pairs in the PDB

Mickey Kosloff[1]* and Rachel Kolodny[1,2]

[1] Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032

[2] Howard Hughes Medical Institute

ABSTRACT

It is often assumed that in the Protein Data Bank (PDB), two proteins with similar sequences will also have similar structures. Accordingly, it has proved useful to develop subsets of the PDB from which "redundant" structures have been removed, based on a sequence-based criterion for similarity. Similarly, when predicting protein structure using homology modeling, if a template structure for modeling a target sequence is selected by sequence alone, this implicitly assumes that all sequence-similar templates are equivalent. Here, we show that this assumption is often not correct and that standard approaches to create subsets of the PDB can lead to the loss of structurally and functionally important information. We have carried out sequence-based structural superpositions and geometry-based structural alignments of a large number of protein pairs to determine the extent to which sequence similarity ensures structural similarity. We find many examples where two proteins that are similar in sequence have structures that differ significantly from one another. The source of the structural differences usually has a functional basis. The number of such proteins pairs that are identified and the magnitude of the dissimilarity depend on the approach that is used to calculate the differences; in particular sequence-based structure superpositioning will identify a larger number of structurally dissimilar pairs than geometry-based structural alignments. When two sequences can be aligned in a statistically meaningful way, sequence-based structural superpositioning provides a meaningful measure of structural differences. This approach and geometry-based structure alignments reveal somewhat different information and one or the other might be preferable in a given application. Our results suggest that in some cases, notably homology modeling, the common use of nonredundant datasets, culled from the PDB based on sequence, may mask important structural and functional information. We have established a data base of sequence-similar, structurally dissimilar protein pairs that will help address this problem (http://luna.bioc.columbia.edu/rachel/seqsimstrdiff.htm).

Proteins 2008; 71:891–902.

Key words: structure comparison; structure alignment; structural differences; nonredundant; structure prediction.

INTRODUCTION

It is often assumed that in the Protein Data Bank (PDB),[1] all the structural representatives of a protein are similar, and more generally that two proteins with similar sequences will also have similar structures. Since the PDB includes many such pairs of structures, it has proved useful to develop subsets of the PDB from which "redundant" structures have been removed, based on a sequence-based criterion for similarity (e.g. Refs. 2–6). These "non-redundant" subsets are often used in statistical and rule-based approaches to protein structure analysis and prediction. The implicit assumptions used in their construction is either that sequence-similar pairs in the PDB have insignificant structural differences or that if significant structural differences between such pairs do exist, the occurrence of this phenomenon is rare enough that it can be safely ignored. Similarly, when predicting protein structure using homology modeling, if a template structure for modeling a target sequence is selected by sequence alone, this implicitly assumes that all sequence-similar templates are equivalent.[7] In particular, this assumption underlies most automated homology modeling servers. Here we investigate the validity of these assumptions.

Some time ago Chothia and Lesk[8] observed that two structures with 50% (100%) sequence identity will align to ~1 Å (0.6 Å) RMSD from each other. Sander and Schneider[9] showed that two structures with more than 35 aligned residues and at least 40% sequence identity will generally structurally align to within 2.5 Å RMSD. Rost[10] used a larger PDB to study the "twilight zone" of low-sequence identities and confirmed that sequence-similar

This Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Mickey Kosloff and Rachel Kolodny contributed equally to this work.

*Correspondence to: Mickey Kosloff, Duke University Medical Center, AERI, 2351 Erwin Rd., Box 3802, Durham, NC 27710. E-mail: mickey.kosloff@duke.edu or Rachel Kolodny, Department of Computer Science, University of Haifa, Haifa 31905, Israel. E-mail: trachel@cs.haifa.ac.il

Received 23 April 2007; Revised 9 July 2007; Accepted 27 July 2007

Published online 14 November 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21770

proteins are expected to be structurally similar. Because these studies measured similarity between protein pairs, they focused on the common substructures and ignored dissimilar parts. Nonetheless, their results suggest that the similar-sequence implies similar-structure paradigm holds.

Of course, there are many well-known examples where proteins undergo significant conformational changes and in such cases the relationship between sequence and structural similarity may no longer be valid (for examples see Refs. 11–13). The molecular motion database of Gerstein and co-workers,[12–15] contains examples of proteins in the PDB with globally similar sequences and dissimilar structures. Most of the entries in this database are from a dataset built as a comprehensive sample of protein flexibility. The motions dataset contains over 3800 SCOP domain pairs sharing a fold, and with a pairwise RMSD that is two standard deviations higher than the average RMSD observed at a given percent identity.[16,17] Recently, Gan *et al.* used structural alignment to compare a representative set of proteins selected from the PROSITE database of protein families and observed over 1700 pairs of structurally-dissimilar proteins in the PDB with sequence identities $\geq$20% and RMSD $\geq$ 2 Å.[18] In these datasets, only a small minority of the structurally-dissimilar pairs have a sequence identity that is above 50% and very few of these have an RMSD $\geq$ 3 Å.

In this study, we further investigate the occurrence of protein pairs with similar-sequences and significant structure dissimilarity, focusing on pairs of proteins with high levels of sequence identity. In contrast to previous studies that used geometry-based structural alignment of protein pairs, our analysis is based on sequence-based structure superpositions, as we show that it better estimates structural differences in sequence-similar proteins. We find numerous protein pairs, of 50–100% sequence identity, that have dissimilar structures, as measured by RMSDs greater than 3 Å or 6 Å. A database of structure-dissimilar pairs is available online at http://luna.bioc.columbia.edu/rachel/seqsimstrdiff.htm. Our results suggest that when creating non-redundant subsets of the PDB or when selecting templates for homology modeling, two proteins or domains in the PDB should be judged as redundant only if both their sequences and structures are similar.

RESULTS

Structure alignment underestimates structural dissimilarity as compared to sequence-based structure superpositioning

It is useful to define the terms *alignment* and *superposition* as used in this work. An alignment of two proteins matches pairs of residues, one from each protein—the alignment refers to the set of these matched residues. *Superposition* refers to the process of superimposing, or overlaying, two protein structures in three dimensions. A *sequence-based superposition* is obtained by optimally superimposing all pairs of residues that are aligned by sequence alone (see Materials and Methods). In contrast, geometry-based superposition methods search for geometric similarities between two proteins while ignoring sequence information. Such programs align and superimpose structurally similar regions and assign gaps to regions that do not superimpose well. The commonly used term *structural alignment* refers to the coupled geometry-based superposition and the alignment it produces.

The RMSD between two superimposed structures is usually measured only over those residues that are considered as aligned, that is, that are not assigned to gaps. The geometry-based alignment of two proteins that are similar in sequence and substantially different in structure will align fewer residues than will be aligned based on sequence. For example, Figure 1 shows two examples of how the RMSD obtained from a geometry-based structure alignment of two sequence-identical, or nearly identical chains, can be lower than that calculated by sequence-based structure superposition. In the first example (panels A–C) a local structurally-divergent region results in an RMSD of 7.13 Å when measured over all residues that are aligned by sequence. In contrast, the RMSD obtained from the structural alignment is much smaller (1.44 Å) since this RMSD is measured only over residues that occupy similar positions in space. In the second example (panels D–F), a hinge motion between domains causes the structure alignment programs to align only one domain and ignore the rest of the protein, resulting in an RMSD of 1.35 Å, which is significantly lower than that measured over all sequence-aligned residues that includes all residues in the full-length protein (10.28 Å).

Figure 2 compares the results of sequence-based structure superpositioning with geometry-based structure alignments obtained using the programs Ska[19] and CE.[20] Each data point shown in the figure corresponds to the best alignment obtained from one of the two programs (see Materials and Methods). The figure is based on a total of 110,068 protein pairs with sequence identities $\geq$70% (see Materials and Methods). As can be seen in Figure 2(A), structural alignment methods consistently align an equal number or fewer residues than sequence alignments. Figure 2(B) compares the RMSDs of the aligned sub-structures obtained from both approaches and further analysis of this data is presented in the Supplementary Material. In almost all cases, geometry-based structure alignments yield a lower RMSD than sequence-based RMSDs.

The situation is reversed when comparing alignments over the same set of residues. Figure 2(C) lists RMSDs calculated from either sequence or geometry-based superpositioning over the set of residues that are matched by

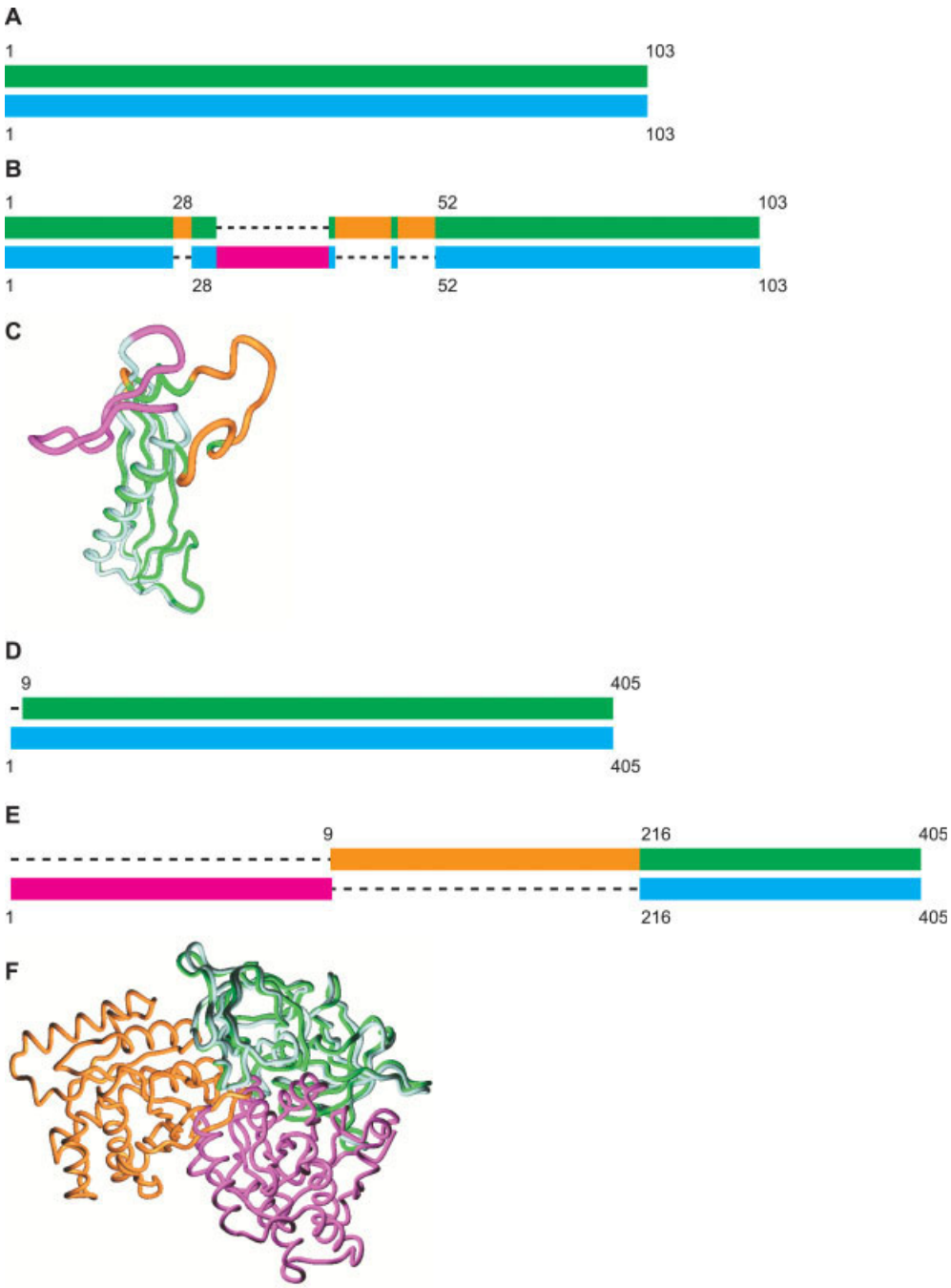

Figure 1

*Two examples of how structure alignment can underestimate structural dissimilarity. (A, B) Schematic representation of the sequence alignment (**A**) versus structural alignment (**B**) of chain A versus chain D from PDB ID 1vr4. The two chains are 100% identical in sequence. The aligned parts are colored green (chain A) and cyan (chain D), while the unaligned parts are colored orange and magenta, respectively. The RMSD of all sequence-aligned residues is 7.1 Å, while that of the structurally-aligned residues is 1.4 Å. (**C**) Structural-alignment based superposition of chains A and D of 1vr4 (colored as in panel B). (D, E) Schematic representation of the sequence alignment (**D**) versus structural alignment (**E**) of two structures of the elongation factor Ef-Tu from Thermus aquaticus—PDB ID 1tui chain A (GDP bound) and 1eft (GTP bound). Inter-domain changes (hinge motion) cause structure alignment programs to align only one domain and ignore the rest of the protein. The aligned parts are colored green (1tui) and cyan (1eft), while the unaligned parts are colored orange and magenta, respectively. The RMSD of all sequence-aligned residues is 10.3 Å, while that of the structurally-aligned residues is 1.3 Å. Note that 1tuiA is not in our dataset because this structure was solved at a resolution of 2.7 Å. Homologs of 1tuiA from E.coli, with ∼70% sequence identity to 1tuiA and to 1eft are in our dataset. These orthologs (e.g. 1dg1G, 1d8tA) are similar in structure to 1tuiA and dissimilar to 1eft, with RMSDs > 10 Å to the latter structure. (**F**) Structural-alignment based superposition of 1tui chain A and 1eft (colored as in panel E).*

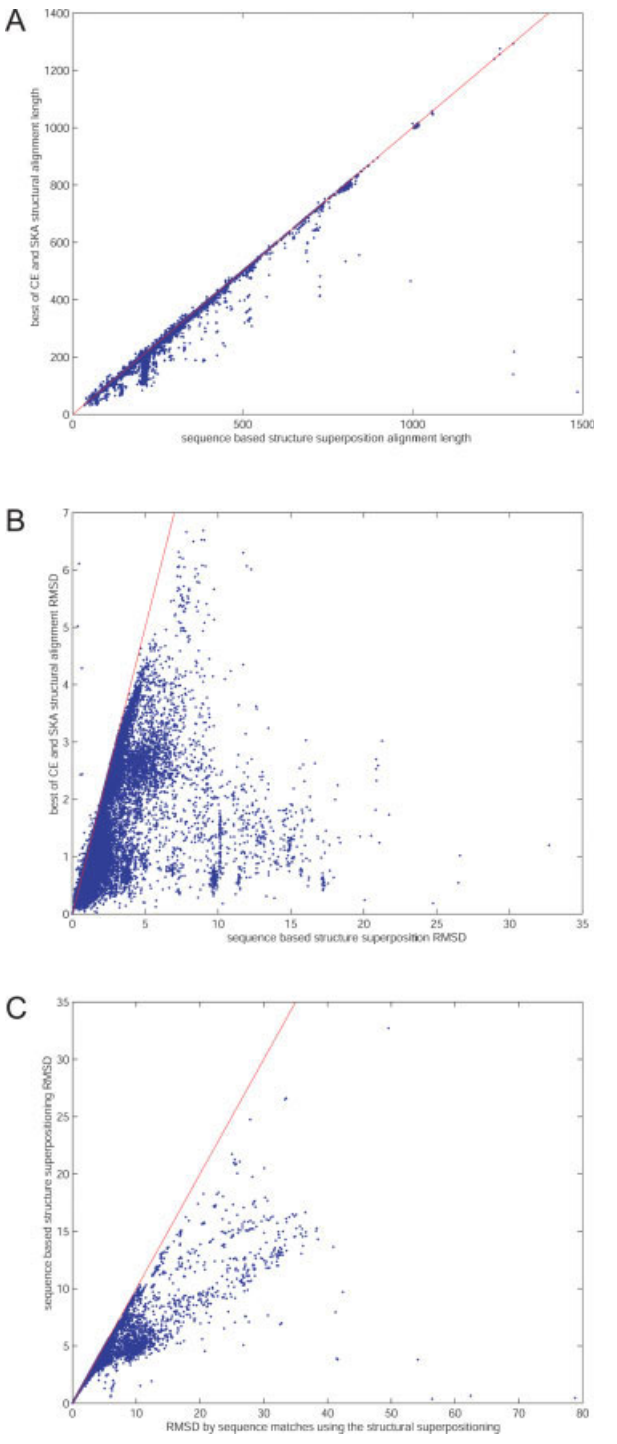

Figure 2

*Comparison of sequence-based structural superpositioning and structural alignments. (**A**) sequence- and structure-alignment lengths. (**B**) RMSDs over these aligned sub-structures. Note the different scales of the x- and y-axis. (**C**) RMSDs of all sequence-aligned residue pairs using two different superpositions: on the x-axis using structural alignment superpositioning, and on the y-axis the sequence-based structural superpositioning. The aligned dataset is of protein pairs with sequence identity ≥70% (see Materials and Methods for details). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

the sequence alignments. That is, the RMSD is calculated for the same set of residues, including residues that are not aligned in the geometry-based alignments. In this case, the RMSD obtained from the geometry-based superposition is always larger than that obtained from the sequence-based superposition. This is expected since the structure alignment makes no attempt to align residues that are identified as equivalent in the sequence-based alignment. Thus, these residues are effectively ignored in the optimization procedure and the need to include them in the RMSD calculation will increase the value that is obtained.

Protein pairs with similar sequences and significant structural differences

Figure 3 shows the distribution of RMSD values obtained from sequence-based structure superpositioning for chain pairs of varying sequence identities. It is evident from the figure that there are many pairs of proteins that have high levels of sequence identity but that are structurally quite dissimilar. Table I lists the total number of pairs in 12 (overlapping) subsets defined by sequence identities ≥50, 70, 99, and 100% and RMSD ≥ 0 Å, 3 Å, 6 Å, showing many protein pairs with similar sequences and substantially different structures. For example, there are over 2600 (11,700) pairs with sequence identity greater or equal to 50% and RMSDs ≥ 6 Å (≥3 Å). Even for 100% sequence identities, there are 158 pairs with RMSD ≥ 6 Å. Note that had we based our analysis on geometry-based structure alignments, much fewer cases would have been detected.

To relate our results to previous work, we used sequence-based structure superpositioning to analyze the "outlier set" in the molecular motions database.[16,17] The majority of the domain pairs in the "outlier set" have sequence identities <50% and many contain NMR entries or structures with resolution worse than 2.5 Å. Only 742 pairs meet our criteria of RMSD ≥6 Å (3 Å), sequence identity ≥50% and resolution better than 2.5 Å. Similarly, the majority of the 1735 structurally dissimilar protein pairs reported by Gan *et al.* using representative probes[18] do not meet our structure resolution, RMSD, and sequence identity criteria. Therefore, the vast majority of the sequence-similar structurally-dissimilar pairs that we report here have not been reported previously.

The complete list of chain pairs in our data set and the eight subsets of structurally-dissimilar chain-pairs (≥6 Å or ≥3 Å RMSD, sequence identity ≥50, 70, 99, and 100%) are available online (http://luna.bioc.columbia.edu/rachel/seqsimstrdiff.htm). Also available online is the sequence-based structural superposition of each pair.

We note that the protein pairs considered in this work cover a significant subset of SCOP families, superfamilies, and folds. Table II lists the number of unique SCOP v.1.69 classes, folds, superfamilies, and families counted

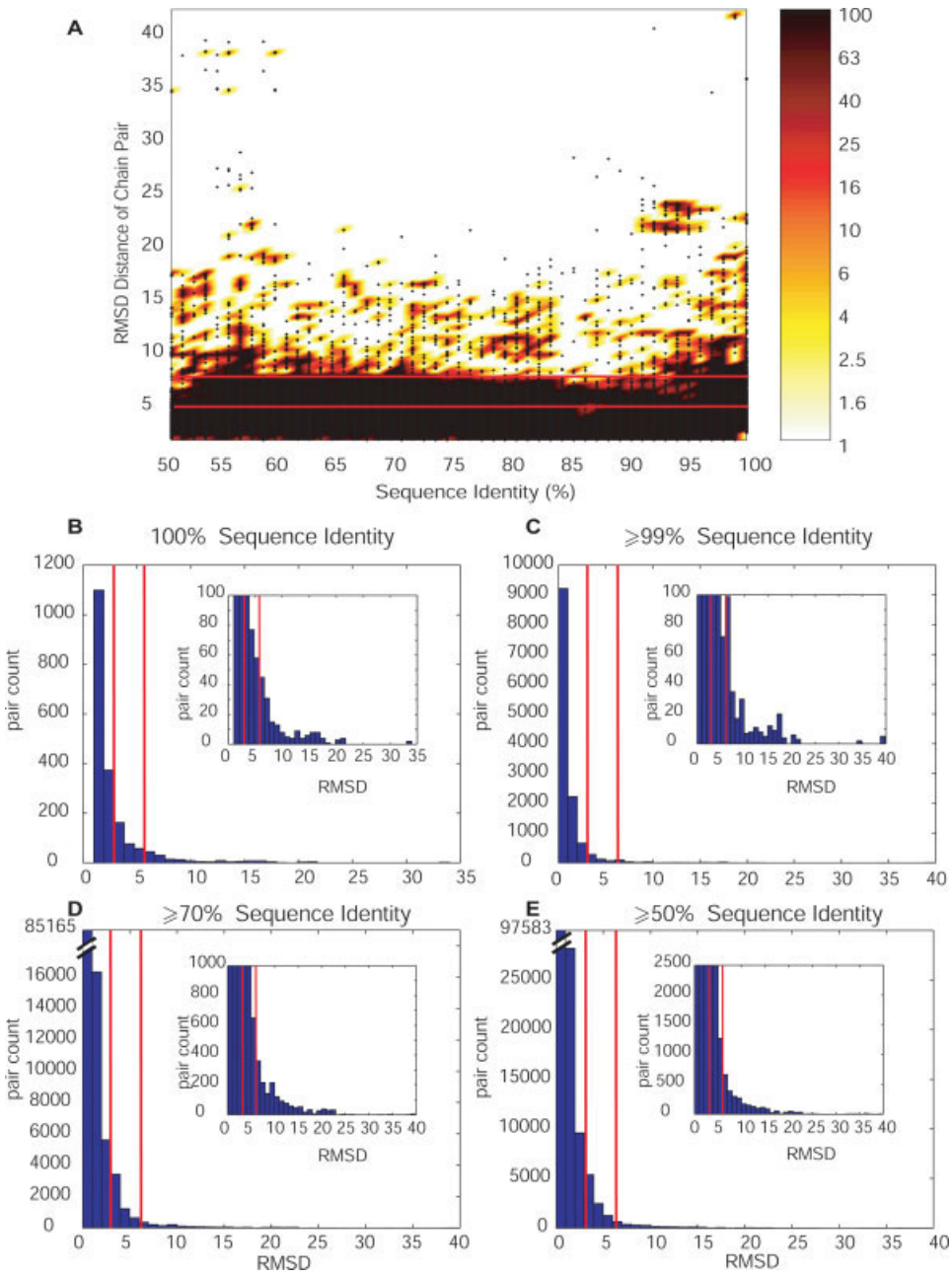

Figure 3

*Abundance of sequence-similar and structurally-dissimilar pairs. (**A**) Sequence-based RMSD of all chain pairs in our data set versus their BLAST sequence identity; the color/gray scale codes the number of pairs in each area of the plot. (**B–E**) show the number of pairs of varying RMSDs and sequence identities ≥100, 99, 70, and 50%; the insets show the same histograms with a magnified y-axis scale. The data shown is the same as in (A), reorganized to quantify the abundance of different pairs with a given RMSD. Lines mark the 6 Å and 3 Å RMSD values. Notice that since we filter pairs with identical sequences and highly similar structures, there are no pairs with 100% sequence identity and less than 1 Å RMSD. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

in each of our subsets: pairs with sequence identities ≥50, 70, 99, and 100%, and RMSD ≥ 3 or 6 Å. The percent of all superfamilies that are found in the subset of pairs with sequence identity ≥50% (= 100%) ranges from 17.2% (7.9%) for RMSD ≥ 3 Å to 8.1% (3.1%) for RMSD ≥ 6 Å.

Inter versus intra-domain structural dissimilarities

Dissimilarities among structures of similar sequences can lie within (intra-) or across (inter-) domains. In the first case, sub-structures within a domain differ [e.g. Figs. 1(A–C)], while in the second there is a hinge

Table I
Sequence-Similar, Structurally-Dissimilar Chain Pairs

Sequence identity (%)	Total pairs[a]		
	≥0 (Å)[b]	≥3 (Å)[c]	≥6 (Å)[c]
100	1,941	444	158
≥99	12,868	757	278
≥70	114,021	6,873	1,575
≥50	147,186	11,749	2,653

[a]Number of pairs after removing redundant structures from the PDB (see Materials and Methods).
[b]The total number of pairs for each of the four subsets.
[c]The total number of structurally-dissimilar pairs, restricted to RMSD ≥ 3 Å or 6 Å.

motion between domains [e.g. Figs. 1(D–F)]. We can distinguish between these cases by comparing the RMSD over the full alignment with that of individually aligned domains. In the case of intra-domain dissimilarity both RMSD values will be the same, and large. For inter-domain dissimilarity the RMSD will be small when measured over individual domains separately [for example, in Fig. 1(D) the RMSD measured only over the superimposed cyan and green domains is small]. Here, we use the domain definitions of SCOP.

Table III lists the number and percentage of domain-pairs that have RMSD ≥ 6 Å or ≥ 3 Å for each of the four chain-pair subsets with chain RMSD ≥ 6 Å and both pairs classified in SCOP v.1.69. In each set, we consider all SCOP v.1.69 domain pairs that overlap more than 35 residues in their sequence alignment, and calculate the RMSD using only the aligned residues in the matched domains. In 60–80% of these structurally-different chain-pairs the RMSD measured over individual SCOP domain-pairs is also greater than 6 Å. Therefore, in the majority of the chain-pairs in this dataset, the structural dissimilarity is due to intra-domain differences.

Table II
The Occurrence of Sequence-Similar, Structurally-Dissimilar Pairs In Different SCOP Classifications

Pairs with sequence identity (%)	Number of SCOP v.1.69			
	Classes (out of 9)	Folds (out of 945)	Super families (out of 1539)	Families (out of 2845)
Containing a structure from a pair with RMSD ≥ 6 Å				
100	8	44	48	54
≥99	8	51	56	63
≥70	9	99	111	129
≥50	9	112	125	150
Containing a structure from a pair with RMSD ≥ 3 Å				
100	8	104	122	143
≥99	8	124	149	179
≥70	9	190	238	306
≥50	9	209	265	351

Table III
Sequence-Similar, Structure-Dissimilar Chain Pairs Containing Structure-Dissimilar SCOP v.1.69 Domains (Intra-Domain Dissimilarity)

Chain pairs with sequence identity (%)	Pairs with chain RMSD ≥ 6 Å[a] (classified by SCOP)	Number of pairs containing at least one SCOP domain-pair with RMSD[b]	
		≥6 (Å)	≥3 (Å)
100	148	106 (72%)	113 (76%)
≥99	259	208 (80%)	219 (85%)
≥70	1,338	987 (74%)	1,030 (77%)
≥50	2,289	1,422 (62%)	1,524 (67%)

[a]RMSD measured over all aligned residues in the chain pairs.
[b]Each chain is separated into SCOP domains and the RMSD is measured independently for each domain. When a chain contains multiple domains, the domain with the maximal *domain* RMSD ≥ 6 Å or 3 Å is counted. In parentheses are the percentages out of the number of chain pairs in each subset.

Factors that lead to structural differences between sequence-identical proteins

Obviously, sequence differences are a major contributor to structural dissimilarity at lower levels of sequence identity. To identify the sources of structural differences between proteins that are essentially identical in sequence, we manually examined the set of 278 pairs in the ≥99% sequence identity and RMSD ≥ 6 Å subset of protein pairs. We clustered these pairs into 66 distinct clusters, based on their SCOP super-family classification and on their biological function, which was derived from the relevant literature. In almost all cases, the biological function dictates a conformational plasticity that results in two or more distinct structures. Figure 4 lists the distribution of causes that account for the structural differences observed for each pair in this subset. The full annotated

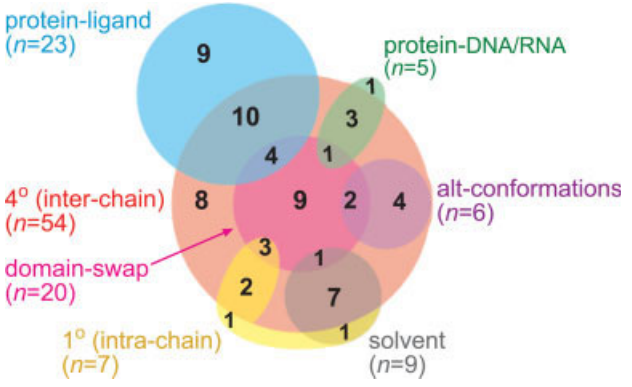

Figure 4
Causes for the marked structural dissimilarity between protein pairs with ≥99% sequence identity and RMSD ≥ 6 Å. The Venn diagram shows the distribution of causes for the structural dissimilarity within pairs. A detailed explanation of each category is given in the text. n refers to the number of occurrences of each cause, out of the 66 separate clusters examined.

subset is available online at (http://luna.bioc.columbia.edu/rachel/pairs_id99-100_rms6.html) and includes the full list of protein pairs and causes.

The causes of structural difference, ordered by frequency, are the following: (1) "Inter-chain (4° structure)"—different quaternary protein–protein interactions (including homomeric interactions). In the majority of cases this involves the presence of a protein chain, which interacts with the relevant chain in only one of the two structures in a pair. A minority of cases involve dissimilar interactions with similar binding partners (usually with an additional cause). "Domain-swap" is a sub-category of "inter-chain" interactions, where only one of the structures in a pair is domain-swapped.[21,22] In rare instances both structures are domain-swapped, but with a different interface. (2) "Protein-ligand"—mostly a ligand-bound protein versus its apo form. Here, ligands are either small molecules, which are nonprotein/nonnucleic acid, or short (<15 residues) peptides. (3) "Solvent"—significant differences in the crystallization conditions (e.g. different pH or salt concentrations). (4) "Alt-conformations"—alternative crystallographic conformations of the same protein. Four of these cases are asymmetric homomers, for which "inter-chain" is an additional cause. One instance corresponds to the same protein crystallized in different space groups, and another corresponds to two alternative fits to the same crystallographic data. (5) "Intra-chain (1° structure)"—the presence/absence of part of a protein chain in one of the structures, a point mutation (combined with an additional cause), or in two instances, oxidized versus reduced intra-chain S—S bonds. (6) "Protein-DNA/RNA"—a DNA-bound protein versus its apo form. One instance involves a restriction enzyme (BamH) bound to specific versus non-specific DNA sequences.

Figure 5 presents selected examples of functional significance that is related to the structural differences between high sequence identity pairs. These include: (a) The bacterial protein TonB (CASP6 target T0240), where the considerable structural difference (20.4 Å RMSD) is a result of intra-protein differences (one structure contains a 14 residue N-terminal stretch, which is absent in the other structure) and a different "inter-chain" quaternary structure, which in this case involves two disparate modes of domain-swapping within each homodimer. Biochemical evidence suggests that in additional to these two conformations, other conformations also exist.[23] This inherent structural plasticity is thought to be central in TonB's function as a transport mediator.[23,24] (b) The apo versus ligand-bound forms of adenylate kinase. The so called "lid" and "NMP" sub-domains change conformation upon ligand binding as part of the catalytic cycle of this enzyme,[25] resulting in a 7.1 Å RMSD. (c) The SH2-SH3 domains of the cABL tyrosine kinase, with or without the C-terminal kinase domain. The presence of the kinase domain in one structure locks the SH2 domain in a specific conformation in relation to the SH3 domain,[26] resulting in an RMSD of 9.5 Å to the second structure, which lacks the kinase domain.[27] These two crystallographic snapshots are representative of a much wider array of possible conformations of cABL.[28,29] (d) Alternative conformations of the monomers in the apo form of the E.coli single-strand DNA-binding (SSB) protein. Each C-terminus of the four chains in this homotetramer, which belongs to the nucleic-acid binding OB-fold superfamily, adopts a different conformation. This conformational plasticity is consistent with the significant conformational changes and refolding events that have been generally associated with the function of nucleic-acid binding by OB-fold proteins.[30] (e) Influenza haemagglutinin, a text book example of functional conformational change,[31] where different pH ("solvent") and differing "inter-chain" interactions result in the largest RMSD difference (39.8 Å) in this high identity subset. (f) The apo form of the transcription factor and proto-oncogene c-Myb versus its DNA bound form. The latter structure also contains an additional transcription factor, C/EBPb, which interacts with c-Myb and the bound DNA.[32]

Examples of sequence-similar, structure-dissimilar templates for homology modeling

The frequent occurrence of sequence-similar, structure-dissimilar proteins in the PDB, which we observe, poses a unique challenge to homology modeling. In particular, we are not aware of an automatic homology modeling server that returns more than one alternative "best" model, if there is more than one sequence-equivalent, but structural dissimilar, template in the PDB. We illustrate how our database can identify such templates with two relatively "easy" examples of homology modeling.

(1) As shown in Figure 5(A), TonB from *E.coli* has been crystallized in two alternative homodimeric forms. If, for example, we want to model a C-terminal part of TonB from *Enterobacter aerogenes* (residues 171–240 of Uniprot entry TONB_ENTAE), searching with this sequence for homologs identifies the *E.coli* structures (1ihrB and 1u07A), both aligning with 75% sequence identity and no gaps to the *E. aerogenes* sequence. Searching our database for either structure identifies the 1ihrB-1u07A pair as having identical sequences and a sequence-based superpositioning RMSD of 20.4 Å. Therefore, these two templates should be treated as non-redundant, and the user modeling the *E. aerogenes* sequence needs to decide between the two alternative templates based on biological or functional criteria. Similarly, an automated prediction server should return both alternative models. Indeed, the assessors in the CASP6 experiment noted that target T0240 (TonB from *E.coli*) was a difficult target for prediction and an "odd case" that

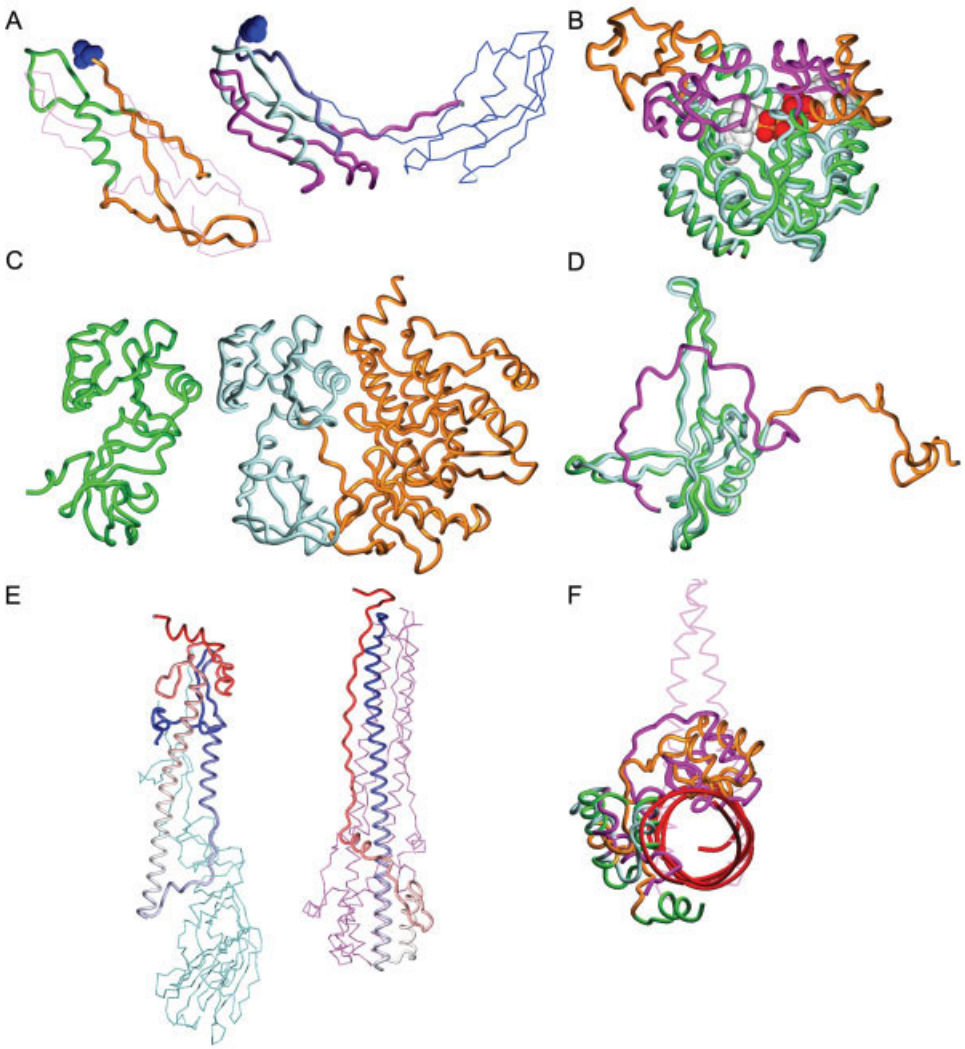

Figure 5

*Examples of pairs with highly-similar sequences and structure dissimilarity that is related to biological function. (**A**) The bacterial protein TonB (1ihrB–1u07A, Inter-chain; Domain-swap; Intra-chain, RMSD of 20.4 Å, 100%): both compared structures are homodimers with a different domain-swapped interface, shown side by side for clarity. The structurally dissimilar regions are colored magenta (1u07A) and orange (1ihrB). 1u07 contains a 14 residue N-terminal stretch, depicted as a purple worm, which is absent in 1ihrB. The N-terminal residue of both compared chains is depicted in CPK model and the second monomer in each structure is depicted in Cα wire representation. (**B**) Adenylate kinase (1akeA–4akeA, Protein-ligand, RMSD of 7.1 Å, 100%): the ligand-bound form is superimposed on the apo form. The so called "lid" and "NMP" domains, which change conformation significantly upon ligand binding, are colored orange (1akeA, the apo form) and magenta (4akeA, the ligand-bound form). The ligand is depicted in CPK model. (**C**) The SH2-SH3 domains of cABL (2abl–1opkA, Intra-chain, RMSD of 9.5 Å, 95%): 1opk contains the cABL kinase domain (orange worm), which is absent in 2abl. This results in different SH2-SH3 domain-domain interaction ("inter-domain" differences). The two structures are displayed side-by-side for clarity. (**D**) The apo structure of the E.coli single-strand DNA-binding (SSB) protein (1qvcA–1qvcB, Inter-chain; Alt-conformations, RMSD of 20.7 Å, 100%): the two compared chains (out of four dissimilar chains in the homo-tetramer) are superimposed, and the variable C-terminus is colored orange (chain A) and magenta (chain B). (**E**) Influenza haemagglutinin (2viuB–1qu1F, Inter-chain; Solvent, RMSD of 39.8 Å, 94%): The two structures are displayed side-by-side for clarity and the two compared chains are colored in a gradient from blue (N-terminal) through white to red (C-terminal). The additional interacting chains are depicted in Cα wire representation. (**F**) The c-Myb transcription factor (1gv2A–1h89C, Inter-chain; Protein-DNA, RMSD of 7.1 Å, 100%): the apo form is superimposed on the DNA bound form, which also includes two chains of the C/EBPβ enhancer protein, depicted in Cα wire representation. The DNA backbone is depicted in red worm and the structurally variable regions are colored magenta (1gv2A) and orange (1h89C). In parenthesis for each example are the two protein chains, designated by their PDB id and chain ID, the causes for the structural differences between the two chains, the sequence-based superpositioning RMSD and the coverage (percentage of the alignment length from the length of the shorter chain). The compared chains are depicted as backbone worms. Unless stated otherwise, the first chain in each pair is colored green and the second chain cyan.*

"fooled" the automatic prediction servers and thus had to be removed from some of the assessments.[33]

(2) In a second example we consider modeling the SH2-SH3 domains of the ABL kinase from *Drosophila melanogaster* (residues 187–346 of Uniprot entry ABL_DROME). Searching for homologs of this query, we find the vertebrate structures (2abl and 1opkA) that align with 74% sequence identity to the *D. melanogaster*

sequence with almost no gaps. Figure 5(C) shows these templates, where the SH2-SH3 domains of the cABL kinase have different conformations, depending on the presence or absence of the kinase domain. Our database shows that the 2abl-1opk chain pair show 100% sequence identity and a sequence-based superpositioning RMSD of 9.5 Å. The graphical representation of the 2abl-1opkA alignment (Fig. S2 of the Supplementary Material) illustrates that the C-terminal kinase domain is present only in 1opk. Here, choosing one of these templates over the other to model the *D. melanogaster* sequence must be based on the biological context of the model.[31] Interestingly, searching for 1opk in our RMSD $\geq$ 3 Å subset identifies two more structure-dissimilar homologs, 1opjB and 1fpuB, that are essentially identical in sequence to the C-terminal domain of 1opk. These are structures of the kinase domain of cABL (i.e., they do not overlap with 2abl, see Fig. S2 in the Supplementary Material), which show an intra-domain dissimilarity to 1opk. The sequence-based superpositioning RMSD of 1opjB and 1fpuB to 1opkA is 5.1 and 3.7 Å, respectively. The reader is referred to Nagar *et al.* for a detailed discussion of the biological significance and cause of the structural differences in the different domains of cABL.[31]

DISCUSSION

In this article we report the existence of a significant number of sequence-similar structurally-dissimilar pairs of proteins in the PDB. Although numerous, these pairs are a minority in our dataset, and by extension in the PDB. The structural dissimilarities range from global rearrangements through inter-domain motion to relatively local structural differences (see also Supplementary Material). The majority of the cases correspond to intra-domain differences. Also, the range of SCOP classifications for the pairs of proteins that we find shows that this phenomenon is found in a wide range of biological families and structural folds.

Many of the pairs of proteins that we identify would not have been found with geometry-based structural alignment programs. As discussed earlier, such programs search for common sub-structures between two proteins, while removing dissimilar parts from the resulting alignment. Thus they will underestimate true geometric differences between structures. In contrast, had we used the results of the geometric superpositions to measure dissimilarities over regions that are well-aligned in sequence, we would have found even more cases of structurally dissimilar pairs. In this regard, it is important to emphasize here that since we are only considering pairs of proteins with high sequence identity ($\geq$50%) and low E-values, the sequence alignments are quite reliable and hence sequence-based structure superpositioning provides a meaningful measure of structural dissimilarities. Furthermore, these high sequence identity alignments typically cover most of the aligned sequences: in the set of $\geq$70% sequence identity, more than 90% of the residues in both proteins are aligned in more than 95% of the pairs.

Interestingly, the vast majority of the sequence-similar structurally-dissimilar pairs reported here were not identified in the studies of Gerstein and co-workers or in the results of Gan *et al.*[17,18] The apparent discrepancy results from a combination of three factors: (1) previous studies used geometry-based structural alignment to calculate RMSDs. (2) The majority of the pairs reported here were not in the PDB five years ago, when other databases were built (data not shown). (3) To reduce the computational cost associated with large-scale structure alignment, previous studies used a representative set of structures, reduced according to SCOP or PROSITE classification. However, our results suggest that many sequence-similar pairs will be overlooked when considering such a reduced set.

Our estimate of the number of sequence-similar structurally-dissimilar protein pairs in the PDB is conservative because: (1) NMR structures were removed from our dataset. (2) The resolution and length thresholds remove many known examples of conformational changes (see examples in Refs. 12,13). (3) We only compare single PDB chains and ignore relative structure changes in a complex of multiple chains within PDB entries.[34] (4) Using global RMSD as a measure of dissimilarity understates relatively local changes in larger proteins.

As is well known, lower sequence identity between pairs contributes to structural differences.[35] This effect is eliminated when focusing on a very high identity/dissimilar structures subset. Our classification of the environmental causes in this subset shows that distinct inter-chain (protein–protein) interactions account for more than half of the dissimilarities and differing protein-ligand interactions for more than one third. This reflects the known fact that binding is often associated with significant conformational changes. As expected, in all of the cases we surveyed that had a known biological function, the conformational plasticity leading to multiple structural states was dictated by that function. It should be noted that the biological function that underlies conformational plasticity is not necessarily the direct cause of the structural differences: for example, a protein that is flexible because of its DNA binding function can adopt two different conformations, even in the absence of bound DNA.

From one perspective, the results of this study are not surprising. The fact that single proteins can exist in more than one conformation is well-known and thus it is expected that some pairs of proteins that are closely related in sequence will have significantly different structures. Disordered proteins are an even more extreme example of structural plasticity.[36–38] However, we believe that the frequency of this phenomenon in the PDB comes as a surprise. It is large enough to suggest that culled databases that do not take structural plasticity

into account may mask important information that can be used, for example, in homology model building. This is particularly relevant to automated structure prediction servers that generally provide a single model as their top answer and usually rely on non-redundant representations of the PDB and to the assessment of structure prediction methods, as in the CASP experiments.[7] The database we have developed as a result of this study (http://luna.bioc.columbia.edu/rachel/seqsimstrdiff.htm) may prove useful in this regard.

Finally, the different results obtained from different alignment protocols raise issues about the meaning of structural alignments. Since geometry-based alignments search for common substructures, they can identify evolutionary related regions of two proteins that do not have a significant sequence similarity. However, when two sequences can be aligned in a statistically meaningful way, the identification of remote evolutionary relationships is not an issue. In this case, sequence-based structural superpositioning provides a meaningful measure of structural differences and of the extent of conformational change that a group of closely related proteins may be expected to undergo. In such cases, geometry-based structure alignments are only useful as a means of identifying common regions between two alternative conformations. Clearly the two approaches to superpositioning reveal different information and it may be useful to use one or both in different applications.

MATERIALS AND METHODS

Data set of protein chains

The dataset used here includes all protein chains from the April 2005 PDB, that are longer than 35 residues, and whose structures were determined to resolution 2.5 Å or better using X-ray crystallography; 38,449 chains from 19,295 proteins satisfy these criteria. Chains with sequence identity of 100% and RMSD lower than 1 Å over their corresponding Cα atoms are defined as redundant. For structures with a resolution of 2.5 Å or better, the Cα RMSD due to the experimental error is well below this threshold.[39–41] In every set of redundant chains, the chain with better resolution was kept. In case of several redundant chains with identical resolution, the longest was kept. The final data set contains 13,193 chains from 9906 protein structures.

Data set of protein chain pairs

The sequences of all chain pairs in the above data set were aligned with BLAST utility bl2seq (version 2.2.10).[42,43] Alignments that had: (1) sequence identity greater or equal to 50%, (2) E-value better than 0.001, and (3) at least 35 matched residues, were selected, resulting in 147,186 pairs. Using more stringent E-value cutoffs up to 10^{-10} and increasing the alignment length cutoffs up to 70 matched residues had a negligible effect on the size of the dataset (data not shown). The sequences were extracted from the PDB coordinates (rather than from the SEQRES fields) and chemically modified residues were translated to standard residues as in ASTRAL.[44] When creating this data set, pairs were filtered by masking low-complexity sub-sequences in the aligned pairs (BLAST filter parameter turned "on"), and then recalculating the correct sequence identity without masking low-complexity sub-sequences (filter parameter turned "off").

Sequence alignment and sequence-based structural superpositioning

As the focus here is on protein pairs that have well-aligned sequences, the set of matching residues can be extracted from their sequence alignment. Each protein pair in the data set was aligned, recording the E-value, BLAST score, sequence identity, sequence similarity, and alignment length. The matching residues were optimally superimposed and the RMSD was calculated using this superposition. Formally, a rotation and translation of one of the chains with respect to the other was calculated, so that it (globally) minimizes the RMSD of the Cα atoms of the sequence-aligned residues.[45] This method is denoted sequence-based structure superpositioning. An implementation of sequence-based structure superpositioning is available within Vistal (http://luna.bioc.columbia.edu/~kolodny/software.html).

The chain pairs were separated into 12 (overlapping) sets based on their level of sequence identity ($\geq$50, 70, 99, and 100%) and structural similarity (RMSD greater than 0, 3, and 6 Å).

Geometry-based structure alignment

Structural superpositions were carried out with Ska,[19] and CE,[20] and the corresponding alignment lengths and RMSDs were recorded. Ska and CE were combined into one alignment method by selecting the alignment with the lowest SAS score (SAS = RMS * 100/(number_matched_residues)).[46] For comparing structure alignment with sequence-based superpositions we focus on cases that are expected to align similar regions using both approaches, restricting the analysis to the 110,068 chain pairs with sequence identities greater or equal to 70%, and with at least 90% of the residues in both chains aligned.

Incorporating SCOP domain assignments into the data set

The SCOP v.1.69[47] domain classification was used to assess if the structural dissimilarities are within domains (intra-domain), or if are they mostly due to inter-domain differences (i.e., rigid body movement of one do-

main relative to another domain in the same chain). For each SCOP-classified sequence-aligned pair, each structure was separated into SCOP domains and RMSDs were calculated independently for each of the sequence-aligned domain pairs, recording the maximum RMSD among the pairs of domains. We also count the different SCOP classifications of the aligned domains to gauge the diversity of the pairs in the sets. As SCOP does not classify all PDB entries, we verified that for all subsets, more than 96% of the chains and more than 85% of the chain-pairs are classified in SCOP v.1.69.

ACKNOWLEDGMENTS

We are grateful to Barry Honig for guidance, support, and seminal contributions to this study, to Michael Levitt, Burkhard Rost, and the members of the Honig group for enlightening discussions regarding this work and to the anonymous reviewers for suggestions that improved the manuscript. Mickey Kosloff gratefully acknowledges the support of the Human Frontier Science Program.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
3. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
4. Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res 2003;31:492–493.
5. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33 (Web Server issue):W94–W98.
6. Mika S, Rost B. UniqueProt: creating representative protein sequence sets. Nucleic Acids Res 2003;31:3789–3791.
7. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round 6. Proteins 2005;61 (Suppl 7):3–7.
8. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
9. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
10. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
11. James LC, Tawfik DS. Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. Trends Biochem Sci 2003; 28:361–368.
12. Goh CS, Milburn D, Gerstein M. Conformational changes associated with protein–protein interactions. Curr Opin Struct Biol 2004;14:104–109.
13. Gerstein M, Echols N. Exploring the range of protein flexibility, from a structural proteomics perspective. Curr Opin Chem Biol 2004;8:14–19.
14. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res 1998;26:4280–4290.
15. Echols N, Milburn D, Gerstein M. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res 2003;31:478–482.
16. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297:233–249.
17. Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu H, Gerstein M. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins 2002;48:682–695.
18. Gan HH, Perlow RA, Roy S, Ko J, Wu M, Huang J, Yan S, Nicoletta A, Vafai J, Sun D, Wang L, Noah JE, Pasquali S, Schlick T. Analysis of protein sequence/structure similarity relationships. Biophys J 2002;83:2781–2791.
19. Petrey D, Honig B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. Methods Enzymol 2003;374:492–509.
20. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
21. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11:1285–1299.
22. Rousseau F, Schymkowitz JW, Itzhaki LS. The unfolding story of three-dimensional domain swapping. Structure (Camb) 2003;11: 243–251.
23. Wiener MC. TonB-dependent outer membrane transport: going for Baroque? Curr Opin Struct Biol 2005;15:394–400.
24. Kodding J, Killig F, Polzer P, Howard SP, Diederichs K, Welte W. Crystal structure of a 92-residue C-terminal fragment of TonB from *Escherichia coli* reveals significant conformational changes compared to structures of smaller TonB fragments. J Biol Chem 2005;280: 3022–3028.
25. Muller CW, Schlauderer GJ, Reinstein J, Schulz GE. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure 1996;4:147–156.
26. Nagar B, Hantschel O, Young MA, Scheffzek K, Veach D, Bornmann W, Clarkson B, Superti-Furga G, Kuriyan J. Structural basis for the autoinhibition of c-Abl tyrosine kinase. Cell 2003;112:859–871.
27. Nam HJ, Haser WG, Roberts TM, Frederick CA. Intramolecular interactions of the regulatory domains of the Bcr-Abl kinase reveal a novel control mechanism. Structure 1996;4:1105–1114.
28. Fushman D, Xu R, Cowburn D. Direct determination of changes of interdomain orientation on ligation: use of the orientational dependence of 15N NMR relaxation in Abl SH(32). Biochemistry 1999;38:10225–10230.
29. Nagar B, Hantschel O, Seeliger M, Davies JM, Weis WI, Superti-Furga G, Kuriyan J. Organization of the SH3-SH2 unit in active and inactive forms of the c-Abl tyrosine kinase. Mol Cell 2006;21: 787–798.
30. Theobald DL, Mitton-Fry RM, Wuttke DS. Nucleic acid recognition by OB-fold proteins. Annu Rev Biophys Biomol Struct 2003;32:115–133.
31. Branden C-I, Tooze J. Introduction to protein structure, Vol. 14. New York, NY: Garland Pub.; 1999. 410 p.
32. Tahirov TH, Sato K, Ichikawa-Iwata E, Sasaki M, Inoue-Bungo T, Shiina M, Kimura K, Takata S, Fujikawa A, Morii H, Kumasaka T, Yamamoto M, Ishii S, Ogata K. Mechanism of c-Myb-C/EBP beta cooperation from separated sites on a promoter. Cell 2002;108:57–70.
33. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;61 (Suppl 7):27–45.
34. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2006;2:e155.
35. Wood TC, Pearson WR. Evolution of protein sequences and structures. J Mol Biol 1999;291:977–995.
36. Fink AL. Natively unfolded proteins. Curr Opin Struct Biol 2005; 15:35–41.
37. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533.

38. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
39. DePristo MA, de Bakker PI, Blundell TL. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure (Camb) 2004;12:831–838.
40. Carugo O. How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. J Appl Crystallogr 2003;36:125–128.
41. Eyal E, Gerzon S, Potapov V, Edelman M, Sobolev V. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J Mol Biol 2005;351:431–442.
42. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
43. Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 1999;174:247–250.
44. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28:254–256.
45. Kabsch W. Solution for Best Rotation to Relate 2 Sets of Vectors. Acta Crystallogr A 1976;32:922–923.
46. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 2005;346:1173–1188.
47. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32 (Database issue):D226–D229.