

RESEARCH ARTICLES

An Alternative View of Protein Fold Space

Ilya N. Shindyalov¹ and Philip E. Bourne^{2,3*}

¹*San Diego Supercomputer Center, San Diego, California*

²*Department of Pharmacology, University of California, San Diego, La Jolla, California*

³*The Burnham Institute, La Jolla, California*

ABSTRACT Comparing and subsequently classifying protein structures information has received significant attention concurrent with the increase in the number of experimentally derived 3-dimensional structures. Classification schemes have focused on biological function found within protein domains and on structure classification based on topology. Here an alternative view is presented that groups substructures. Substructures are long (50–150 residue) highly repetitive near-contiguous pieces of polypeptide chain that occur frequently in a set of proteins from the PDB defined as structurally non-redundant over the complete polypeptide chain. The substructure classification is based on a previously reported Combinatorial Extension (CE) algorithm that provides a significantly different set of structure alignments than those previously described, having, for example, only a 40% overlap with FSSP. Qualitatively the algorithm provides longer contiguous aligned segments at the price of a slightly higher root-mean-square deviation (rmsd). Clustering these alignments gives a discreet and highly repetitive set of substructures not detectable by sequence similarity alone. In some cases different substructures represent all or different parts of well known folds indicative of the Russian doll effect—the continuity of protein fold space. In other cases they fall into different structure and functional classifications. It is too early to determine whether these newly classified substructures represent new insights into the evolution of a structural framework important to many proteins. What is apparent from on-going work is that these substructures have the potential to be useful probes in finding remote sequence homology and in structure prediction studies. The characteristics of the complete all-by-all comparison of the polypeptide chains present in the PDB and details of the filtering procedure by pair-wise structure alignment that led to the emergent substructure gallery are discussed. Substructure classification, alignments, and tools to analyze them are available at <http://cl.sdsc.edu/ce.html>. *Proteins* 2000;38:247–260.

© 2000 Wiley-Liss, Inc.

Key words: protein fold space; protein structure comparison; protein structure similarity; protein structure neighbors; combinatorial extension; substructure similarity

INTRODUCTION

The comparative analysis of protein structures is not new. For example, Choitha and Lesk^{1,2} compared the globins in detail as early as 1980 when it was realized, even with just a few structures available in the PDB, that these structures adopted very similar folds even with low sequence identity. Just how low became apparent from later work (circa 1992) when systematic comparisons of all available protein structures were undertaken.^{3,4} These and later studies uncovered 3-D similarity in functionally diverse proteins, for example, glycogen phosphorylase, a protein involved in metabolism, and DNA glucosyltransferase, a protein that protects the DNA of phage T4 against its own nucleases as it degrades the hosts genome.⁵ The 3-D similarity comes from a common chemical mechanism of diphosphate and sugar-based chemistry but substrate specificity, and cellular functions are different. Such 3-D similarity points to a common ancestor not detectable by sequence homology algorithms alone. Promise of finding remote evolutionary relationships, hence the inference of unknown biological function from 3-D structure similarity, and the desire to better understand the scope of experimentally observed protein fold space has led to the development of several different protein structure classification schemes. Each classification scheme is derived from different algorithms and with slightly different goals in mind. Scop⁶ is derived primarily by visual inspection whereas FSSP⁷ is based upon the DALI⁸ algorithm, CATH⁹ has its roots in the SSAP algorithm¹⁰ and MMDB uses the VAST algorithm.¹¹ Many other methodologies and classifications have been reported (see Holm and

Grant sponsor: National Science Foundation; Grant numbers: DBI 9630339 and DBI 9808706.

The substructure gallery specified in this article can be reviewed at <http://cl.sdsc.edu/ce.html/>.

*Correspondence to: Philip E. Bourne, San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537. E-mail: bourne@sdsc.edu.

Received 16 April 1999; Accepted 21 September 1999

Sander, 1994; Godzik, 1996; and Madej et al., 1995^{12–14} for a review). Those highlighted here provide Web-accessible resources that are kept current as new structures become available from the Protein Data Bank (PDB).

Structure comparison methods generally operate at the level of the polypeptide chain, however, it has long been recognized¹⁵ that from a functional perspective proteins should be classified differently. Either because they consist of multiple chains (possibly homodimers) or that they consist of discrete domains that are subject to evolutionary rearrangement and reuse. Multiple domains may appear in a single polypeptide chain or a single domain may involve multiple polypeptide chains, or non-contiguous chain segments from one or more chains. Domains are characterized by a large number of intra-domain contacts and relatively few inter-domain contacts. Given the functional significance of domains a variety of methods have been developed to predict domains.^{16–20} CATH uses a consensus approach based upon three methods²¹ and FSSP provides a domain dictionary.²²

The methodology described here purposely does not take into account domains per se. Rather it simply classifies recurring substructures as contiguous or near-contiguous pieces of a single polypeptide chain. Intuitively, recurring substructures, if they have any biological meaning, would be expected to occur within single domains and this is analyzed subsequently.

The global 3-D alignment of two proteins has been characterized as NP hard.²³ Therefore, each of the methods described above, and others like them, make approximations to make these problems computationally tractable. As Godzik has shown¹³ these assumptions, while providing gross similarities, lead to differences in the details of the structure alignment. So much so that two alignment methods can result in differences at every position in an alignment. Consider the implications of this for alignments based on the Combinatorial Extension (CE) method when compared to other methods.

CE was described in earlier work²⁴ and only a synopsis is given here. CE empirically defines an Aligned Fragment Pair (AFP) as a geometric similarity between two octomeric fragments of carbon-alpha positions. An AFP is a 3-D structure analogue of a dot on a dot matrix plot of sequence similarity, but rather each dot represents a similarity in local geometry based on comparison of octomeric C-alpha fragments. An essential part of the approach is the set of empirical rules that describe how to build the alignment path from AFPs. Applying the same rules to a random set of structures derives the statistical significance of the alignment (z-score).

In this work the CE algorithm is used to compare all non-redundant polypeptide chains in the current PDB to each other and all significant similarities are recorded in a Property Object Model²⁵ (POM) database which includes pairwise alignments of the polypeptide chains. Since the PDB itself is highly redundant so are the structure alignments. These similarities are subsequently filtered and collated as defined subsequently to define a recurrent set of substructures and as shown a useful adjunct to our current knowledge of protein fold space.

The starting point for this work is available as a Web accessible database either directly at <http://cl.sdsc.edu/ce.html> or via the PDB at <http://www.rcsb.org> as one of several 3-D structure neighboring methods.

MATERIALS AND METHODS

All-Against-All Structure Comparison—A First-Order Pairwise Alignment

Using the previously described CE algorithm a comparison of structures was made for the complete set of polypeptide chains from the PDB with at least 30 C-alpha atoms per chain. For NMR structures the first member of an ensemble was chosen. In the course of the search unprocessed chains are randomly selected from the pool of all chains and compared against a list of structure representatives (which is empty in the beginning). If the new chain is similar to one of the structure representatives then it is assigned to that representative and becomes the representing chain. If the new chain is unique it becomes a new representative and is added to the representatives' list. Obviously, in the beginning the first chain examined starts the list of representatives. The following set of criteria were used to define these first-order structure representatives:

- (i) The rmsd (root mean square deviation) between two aligned chains is less than 2 Å;
- (ii) The difference in chain length is less than 10%;
- (iii) The number of aligned positions is at least two-thirds the length of the represented chain;
- (iv) The number of gap positions in the alignment is less than 20% of the number of aligned positions.

A chain that is represented is not used in the subsequent analysis described here, but is available from the database either for use as a probe for finding structure homologues or to compare sequence alignments resulting from structure alignments. If two representatives have m_1 and m_2 represented chains, respectively, then this would imply a total of $(m_1 + 1) \cdot (m_2 + 1)$ structural similarities. Given the computational intractability of this problem, it is necessary to infer similarities between represented chains through the relationships of their representatives. The alignment between two represented chains, R1 and R2 is determined by inference from the overlapping residue positions of the alignments between RE1 and RE2, RE1 and R1, and, RE2 and R2, where RE1 and RE2, are representatives of R1 and R2, respectively. The negative impact of this approach is that the similarity between structures being represented is not as well described as for the representatives themselves. The positive impact of this approach is that the representatives themselves constitute, as a first-order comparison, a non-redundant set of structures for subsequent analysis.

No requirement for sequence similarity was imposed in defining the first-order representative set of structures, however, as expected, many structures being represented by a single representative have substantial sequence similarity. However, the structure similarities are in no way biased by sequence, but represent a pure geometric

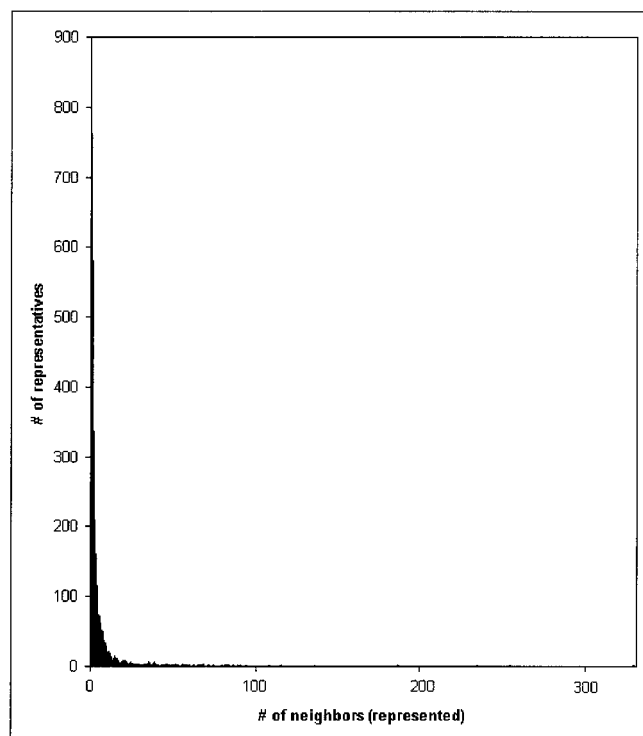


Fig. 1. Distribution of first-order structure representative polypeptide chains by number of represented chains.

comparison. Not performing a sequence alignment first, thereby reducing the number of comparisons, is not a computational burden since structure superposition of highly homologous sequences is fast. It is only the detailed CE alignment using dynamic programming that is slow and this step is only performed between representatives. The search for structure representatives was performed in parallel using 128 processors on a Cray T3E. A total of 24,000 cpu hours were used, with an average time for comparing two polypeptide chains of 39 seconds. A starting data set of 8,557 files from the PDB of August 1998 was used, providing 12,720 polypeptide chains with 30 or more C-alpha atoms. Based on the criteria described above, 2,016 structure representatives were identified, comprising 15.8% of all polypeptide chains. These structure representatives are available as a Property Object Model (POM)²⁵ database for querying via the Web at the URL <http://cl.sdsc.edu/ce.html> or as a flat file containing all alignments for downloading.

The distribution of these 2,016 first-order structure representatives with respect to the polypeptide chains they are representing is given in Figure 1. Eight first-order representatives each have 100 or more structures that they represent. The top first-order representatives are: hemoglobin (2HHB:C)* with 329 polypeptide chains represented; Bacteriophage T4 lysozyme (190L:_) with 254

represented; egg white lysozyme (1HEL:_) with 234 represented; HIV-1 protease (5HVP:A) with 186 represented; and beta trypsin (1TYN:_) with 135 represented. A total of 763 representatives have no represented structures and 402, only one. Contrast this to the set proposed by Hobohm et al.²⁶ (known as PDB_SELECT) which is based on Smith-Waterman pair-wise sequence alignments followed by selection of a representative based on structural quality as determined by resolution and R factor. The Hobohm et al. list of December 1998 contains 1,028 unique entries above a 25% sequence similarity. Comparison of the first-order representatives computed here and the PDB_SELECT set indicate that while the overall fold above a 25% sequence similarity may be similar, details of the structure may show significant variation. Hence the difference in the number of representatives, even given the obvious difference that we were working with a different version of the PDB. For example, cAMP-dependent protein kinase (PKA) is represented by a single entry in PDB_SELECT (1CMK:E), yet there are over 20 representatives with an rmsd > 2 Å in the first order representative set presented here. An especially large number, given the number of kinases that have been studied. Conversely, proteins belonging to the same superfamily may have a sequence similarity of less than 25% and while counted twice by PDB_SELECT are excluded from the first order representatives since their overall folds are very similar. For the substructure characterization presented here these variations need to be captured. As will be shown subsequently, these numbers will be refined but for the first pass it is important to capture a wide range of geometric variation. The first-order pair-wise alignments calculated here again highlight the previously reported structural redundancy in the PDB (Fig. 1). A small number of structures with similar folds have been extensively studied, often through point mutations (e.g., T4 lysozyme) or through structure solution of a native protein with a large number of inhibitors (e.g., HIV-1 protease). Recent estimates^{27,28} suggest that all proteins can be divided into somewhere between 600 to 1,500 superfamilies. Since only a fraction of protein fold space is already present in the PDB, clearly the 2,016 first-order representatives can be further filtered at the superfamily level. However, our efforts were directed not at the overall fold, but at the details of highly similar substructures.

Analyzing Clusters in Fold Space—A Second-Order Pairwise Alignment

Given this initial filtering to define a representative set of polypeptide chains it was possible to compare these first-order representatives further in an effort to better characterize protein fold space.

To reduce the effect of overrepresentation by large protein families a threshold on sequence similarity of 25% was applied to the 15.8% of the polypeptide chains in the PDB that constitute the first-order representative set. Specifically, any two proteins in the first-order representative set were represented by a single entry in the second-order set if the sequence similarity for common positions in the alignment was greater than 25%. A further restriction was applied to gap positions. Two polypeptide chains were

*Protein structures are identified by their four-character PDB identifiers followed by a colon and a single character representing the specific polypeptide chain. In the case of a single unassigned chain it is represented by an underscore (_).



Fig. 2. Structure neighbors for Succinyl-CoA Synthetase chain B (1SCU:B). Blue represents aligned residues except for the first bar, which shows the summary of alignments of the master representative to all neighbors. Green represents gaps: thin line for the neighbor; thick line for

the master representative (in the latter case the single residue before and after the gap is colored in green and the actual gap size is not represented). Gray represents residues not included in the local alignment.



Fig. 3. Structure neighbors for nuclear factor of activated T cells (NFAT) from a protein-DNA complex chain N (1A02:N). Color coding is identical to that shown in Figure 2.

only considered similar if the number of gaps (non-aligned positions) was less than 30% of aligned positions. This criterion can be thought of as analogous to applying a gap penalty when performing a sequence alignment.

Based on these criteria all significant pairwise alignments for the remaining first-order representatives having a z-score above 4.0 and a root-mean-square deviation (rmsd) below 5 Å have been assembled. According to our previous experience, similarities with a z-score above 4.5–5.0 are typical for proteins belonging to a single family, whereas a z-score above 4.0 corresponds to a superfamily or fold-level structural similarity. Similarities with z-scores in the range from 3.5 to 4.0 correspond to the “twilight zone” of structural similarity and may represent some interesting biological feature that needs confirmation by experiment. These are empirical rules, but have nonetheless proven useful comparison indicators. Consider results from recent work that support these empirical rules. A pairwise comparison of the structures of the catalytic subunit of 55 protein kinases (PKA), an enzyme responsible for phosphorylation in the signal transduction pathway was undertaken (http://cl.sdsc.edu/ce/align_db.html). The catalytic subunits of PKA exhibit sequence identity as low as 17%, yet the enzyme retains a common fold in the catalytic domain, albeit with a maximum rmsd of 5.9 Å between two insulin receptor kinases (1IR3:A and 1IRK:_) where there is a major conformational change in the activation loop upon autophosphorylation.²⁹ The z-scores for pairwise alignments between the 55 members of the protein kinase family range from 4.2 to 8.3 with a mean close to 7.0. The z-scores support the idea that these proteins are all from the same family.

Similarly, recent work³⁰ reports the presence of two putative EF-hand motifs in acetylcholinesterases that are indicative of Ca^{2+} binding. An rmsd of 3.8 Å and a z-score of 4.1 is reported between acetylcholinesterase from *Torpedo californica* (2ACE:_) and calmodulin (1CLL:_), a well characterized calcium binding protein, over the

76 residues defining the contiguous region of the two putative EF-hand motifs. While this z-score approaches the twilight zone, there is supporting experimental evidence for a calcium-binding motif. Specifically, a well-defined sequence homology to a related family of neurologins that have been shown experimentally to bind Ca^{2+} has been established. Further, the region of acetylcholinesterase containing the sequence homology includes a region similar to the PROSITE EF-hand signature for calcium binding.

Accepting the empirical criteria defined above, it appears that pairwise second-order alignments tend to form clusters that can be mapped in a pair-wise fashion to a single master representative. Examples illustrating this mapping are provided in Figures 2 and 3 and Tables I and II, respectively. One can clearly see distinctive clustering in the location of pairwise similarities between two typical representatives, succinyl-CoA synthetase (1SCU:B, Figure 2, Table I) and nuclear factor of activated T cells (NFAT) from a protein-DNA complex (1A02:N, Figure 3, Table II) and their structure neighbors. It is important to emphasize that the alignments provided in Figures 2 and 3 are not multiple structure alignments, but rather pairwise alignments between the master representative and its neighbors. In short, these are not consensus alignments but represent common substructures present in all structures within the specified criteria of rmsd and z-score. Similar alignments are seen for other first-order representatives and can be reviewed on the Web site. Figures 2 and 3 indicate that these substructure motifs occur at various points with respect to the N- and C-termini of the polypeptide chain and require analysis with respect to domains defined by other methods. This is discussed subsequently. What is apparent from Tables I and II is that the sequence similarity for these regions is undetectable without a priori knowledge of the structure alignments which occur over extended regions of the polypeptide chain. Given the importance of structure in recognizing distant

TABLE I. Structure Neighbors for Chain 1SCU:B of Size 388[†]

Chain	Size	Z-score	Rmsd Å	Seq. ide. %	Aligned	Gaps	Alignment			
							First Chain (master repr.)		Second Chain	
							Begin	End	Begin	End
1NAW:A	419	4.1	4.8	5.2	77	19	291	373	260	349
1OXP:___	401	5.0	3.2	8.4	107	27	258	374	124	247
2LBP:___	346	5.2	3.3	11.3	106	15	258	374	140	249
1PFK:B	320	4.1	3.4	14.1	71	19	265	352	178	250
1UDR:B	129	5.5	3.4	8.7	115	20	257	386	6	125
1DBP:___	271	5.0	2.8	16.7	102	20	258	372	125	233
1TTP:B	397	4.4	3.8	8.2	85	22	293	386	85	182
1DBQ:A	289	4.4	2.9	12.8	94	28	258	374	9	107
3MIN:A	491	4.6	4.2	13.9	108	26	258	368	77	207
1TAH:B	318	4.2	3.3	12.4	89	24	259	354	9	114
1BAP:___	306	5.0	3.1	12.3	106	23	258	373	135	253
1BMT:A	246	5.3	3.2	6.7	120	23	256	388	95	224
1LBG:A	360	5.0	3.1	12.9	101	24	258	373	181	290
1RNL:___	215	5.3	2.9	10.5	114	23	258	388	7	126
1NBA:A	264	4.4	3.3	4.6	87	18	286	384	139	231
1BRO:A	277	4.6	3.6	9.3	107	32	242	358	1	129
1SRR:C	124	5.2	3.1	9.7	113	21	258	387	5	121
1LCI:___	550	4.7	3.5	5.9	101	21	258	374	235	340
2REQ:B	637	4.7	3.4	5.2	118	19	257	388	510	632
1NBD:___	214	4.9	3.5	6.7	105	23	260	374	65	182
1PEA:___	385	5.2	3.2	9.3	107	17	258	374	142	255
1FCD:A	401	4.4	4.2	7.7	91	13	287	377	123	226
1A2O:A	349	5.2	2.8	9.7	113	31	258	386	5	132
1AMU:A	563	4.7	3.1	4.2	96	26	258	374	223	323

[†]Chain—protein chain identifier in the form PPPP:c, where PPPP—PDB ID, C—chain ID within protein PPPP. Size—chain size as a number of residues. Z-score—z-score of structure alignment. Rmsd—rmsd of structure alignment. Seq. ide.—sequence identities for the region of structure alignment. Aligned—number of positions in the alignment. Gaps—number of gaps (non-aligned positions) in the alignment. Alignment—residues included in the alignment from the master representative and the other chain aligned to the master representative.

sequence relationships it is encouraging that the number of structures is currently growing at about 30% per year. The Web site associated with this work has been automated and will continue to provide a current view of these relationships.

An all-by-all comparison of these first-order representatives with no gap or sequence similarity restrictions and with a z-score > 4 and an rmsd < 5 Å revealed second-order representatives with yet further structure similarity (Fig. 4). Only 239 of the 2,016 first-order representatives have no detectable structural similarity and by the criteria imposed here would be considered unique structure motifs. The distribution given in Figure 4 indicates that many first-order representatives have a significant number of structure neighbors, specifically 796 first-order representatives have from 1 to 10 detectable similarities which form the second-order representative set. In other words, at this level of second-order filtering there are a few representatives with a large number of neighbors, but there is a more even distribution than exists in the PDB as a whole. The second-order representatives with the most neighbors are: methylmalonyl-CoA mutase polypeptide chain C (2REQ:C) with 183 representatives; heat labile enterotoxin type Iib polypeptide chain C (1TII:C) with 181 representatives; methylmalonyl-CoA mutase polypeptide chain A (1REQ:A) with 176 representatives; cryia A (1CIY:_) with 165 representatives; and beta-glucuronidase polypeptide chain A

(1BHGA) with 163 representatives. These groupings are not surprising since they represent well-characterized tertiary structures. Both {1,2}REQ are TIM barrels and 1CIY:_ and 1BHGA each consist of three domains. 1CIY:_ contains a beta sandwich, a mainly beta-aligned prism and a helix bundle, a diverse set of tertiary structures and hence well represented. Likewise, 1BHGA contains a beta sandwich, immunoglobulin-like domain and a TIM Barrel. 1TII contains an uncharacterized, yet common tertiary structure.

In summary, first order representatives represent close structure neighbors that have high structure similarity over a good part of their polypeptide chains. They represent a non-redundant set of protein families. By comparing first order representatives a set of second order representatives emerge which show further similarities between substructures. This progressive comparative analysis of structures unique at one level and redundant at the substructure level has been referred to as the Russian doll effect. With a cataloging of these substructures the question becomes what is the smallest doll that makes sense and what can you do with it?

Common Substructures—A Third-Order Pairwise Alignment

To further characterize this substructure-based view of protein fold space we searched for additional structure

TABLE II. Structure Neighbors for Chain 1A02:N of Size 301[†]

Chain	Size	Z-score	Rmsd Å	Seq. ide. %	Aligned	Gaps	Alignment			
							First Chain (master repr.)		Second Chain	
							Begin	End	Begin	End
1NFK:A	325	4.9	2.7	10.3	107	7	193	301	200	311
1CDI:A	315	4.1	4.0	5.6	90	20	198	300	181	277
1SEB:E	181	4.2	4.2	9.8	92	17	198	299	83	181
1WHP:___	96	4.1	3.9	11.0	82	17	202	299	3	85
4KBP:B	432	4.6	3.6	10.5	95	16	197	301	22	122
1AR2:___	109	4.1	3.7	3.2	94	24	198	300	1	109
1TEN:___	90	4.7	3.4	2.3	88	13	200	299	2	90
1GOG:___	639	5.2	2.8	15.9	94	12	199	300	542	639
1ILN:B	200	4.4	4.3	6.9	102	22	193	301	84	200
3CD4:___	182	4.1	3.8	8.2	85	24	194	301	93	178
1ILN:G	192	4.4	4.4	4.2	94	22	194	300	90	192
1AKP:___	114	4.2	3.8	7.4	94	24	202	300	2	114
1TLK:___	154	4.2	3.9	6.8	88	24	198	301	39	134
1FRT:A	269	4.1	3.7	7.9	88	15	198	296	178	269
1IEA:B	228	4.2	4.1	5.7	89	22	198	299	122	219
1IGY:B	434	5.0	4.3	4.5	112	21	176	298	208	329
1KCW:___	1046	4.1	4.0	7.3	82	20	209	301	793	883

[†]Chain—protein chain identifier in the form PPPP:c, where PPPP—PDB ID, C—chain ID within protein PPPP. Size—chain size as a number of residues. Z-score—z-score of structure alignment. Rmsd—rmsd of structure alignment. Seq. ide.—sequence identities for the region of structure alignment. Aligned—number of positions in the alignment. Gaps—number of gaps (non-aligned positions) in the alignment. Alignment—residues included in the alignment from the master representative and the other chain aligned to the master representative.

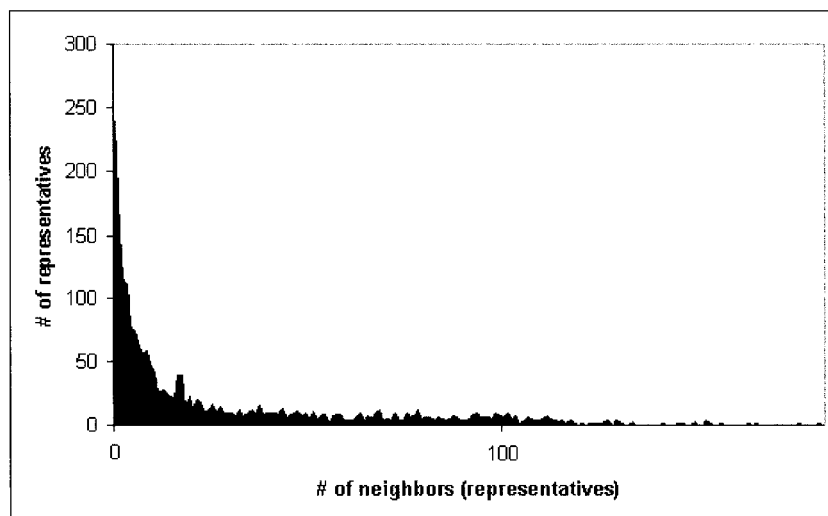


Fig. 4. Distribution of first-order structure representatives by number of neighbors.

similarity in second-order alignments. The resulting characterization of third-order alignments we refer to as the substructure gallery, which was defined as follows. As a starting point, we use second-order representatives expanded with first-order representatives to remove any restriction on sequence similarity. Sequence similarity was not restricted here since the second-order filtering prevented any bias towards sequence homology at the level of the complete polypeptide chain. However, any sequence similarity at the substructure level defined by second-order alignments now becomes valuable informa-

tion. Similarly, the number of gaps was allowed to increase to be up to 50% of the aligned positions.

Given these criteria, common substructures are detected and ordered by selecting first-order representatives with the maximum number of overlapping structure alignments. Given the somewhat continuous nature of protein fold space, yet desiring to characterize distinct regions, it was necessary to introduce an empirical rule to better distinguish substructures from each other. Once a group of first-order representatives is chosen to represent a common substructure then no new substructure can be defined

that has greater than 50% overlap in specific first-order representatives with a substructure already defined. This rule requires some further analysis with respect to how boundaries between distinct substructures are defined.

RESULTS AND DISCUSSION

Properties of Common Substructures

With a z-score > 4.0 and a rmsd threshold < 4.0 Å a total of 75 substructures containing seven or more third-order alignments was found. These are referred to as class I substructures. Increasing the rmsd threshold to < 5.0 Å while maintaining a z-score of > 4.0 , this number goes up to 140, referred to as class II substructures. The top 20 common substructures from class I are shown in Figure 5. What is immediately apparent is that some substructures are clearly identified as classical folds, for example, substructure 1 [1] contains the immunoglobulin fold. Others, for example [2] and [4], contain the Rossman fold, indicative of nucleotide binding, yet have different overall environments. Still others appear as unique substructures, either with a simple or complex tertiary structure. The average length of substructures is 131 (class I) and 134 (class II) amino acids, respectively. The number of first-order representatives that contain each substructure ranges from 7 to 106, with a mean of 21. Consider for example member [10] from the substructure gallery. The substructure exists at various positions within the complete polypeptide chain (Fig. 6) yet has a consistent alignment of secondary structure elements (Fig. 7). Closer examination reveals that this substructure is found consistently in both TIM barrels and Rossman folds, yet is also found in other types of alpha-beta-alpha 3-layered sandwich structures. Similarly, member [14], while obviously all alpha, exists in structures previously classified as bundled and non-bundled. Structures with widely ranging biological function, for example, cytokines, hormones, electron transport and iron storage proteins. In short, substructures span traditional structural and functional boundaries, a fact that is revisited subsequently with respect to current classification schemes. Information like that presented in Figures 6 and 7 is available via the Web site for all members of the substructure gallery.

The extent of known fold space covered by the substructure gallery can be estimated from the total number of first-order representatives aligned to common substructures; 721 for rmsd < 4.0 Å (class I) and 932 for an rmsd < 5.0 Å (class II) out of a total of 2,016 first-order representatives. This represents 36% and 46% coverage of experimentally known fold space for class I and class II, respectively. The remainders are first-order representatives that are not found in common substructures, or are included in substructures with less than seven first-order representatives. At the same time there are first-order representatives containing two or more substructures and in approximately 60% of cases the substructures contain overlapping regions of the polypeptide chain. There were 258 (class I) and 481 (class II) first-order representatives containing more than one substructure. Thus 36% (class I) and 52%

(class II) of first-order representatives include several substructures.

To further characterize this overlap between substructures (and hence first-order representatives) a comparison of substructure overlap based on the number of shared first-order representatives was performed. This distribution is given in Figure 8 for an rmsd between first-order representatives < 4.0 Å for the region of overlap. The actual rmsd for the overlap is given in Figure 9. For substructure overlaps greater than 30% and less than 50% of the substructure length it was found that each common substructure overlaps on average 2.1 other common substructures (Fig. 9). Further, substructures themselves have significant structure similarity. That is, for a z-score above 4.0 each substructure overlaps on average 2.3 other substructures with an average rmsd of 3.4 Å (Fig. 9).

Thus, substructures are not in and of themselves discreet but demonstrate substantial overlap. This in turn implies substantial overlap between structure representatives and hence all proteins of known structure. This notion of the continuity of fold space brings into question how well fold space is described in terms of distinct folds in, for example, the scop or CATH sense. While domain-based classifications are very useful and indeed, as shown subsequently, related, the classification of substructures found in the substructure gallery raises the possibility of an alternative, quantitative and in one sense more natural classification of fold space than folds assigned to whole protein domains. This again raises the question about mapping between folds in the traditional sense and the common substructures presented here. This is discussed with respect to domain assignments as found in the CATH⁹ classification, the scop⁶ classification, and the FSSP⁷ classification.

Comparison of Common Substructures With CATH Domain Assignments

CATH uses a consensus approach based on three methods to assign domains.²¹ The top 20 substructures were compared to the CATH domain classification. In all but one case, [16] biotin carboxylase, substructures were contained within CATH domains. [16] defines a substructure from 2–108 residues, whereas CATH defines three subdomains from residues 1–85, 86–130; 204–446, and 131–203. Interestingly scop defines domains from 1–114, 115–330, and 331–446. Visual inspection indicates either are probable, yet from a functional perspective the authors³² conclude that residues 1–103 form the dinucleotide-binding motif, the substructure detected here. In eleven cases the substructures miss one or more secondary structure elements, five are within a few residues of the specified domain boundaries, and the remainder were not processed by CATH. In summary the methodology described here for defining substructures provides a useful check against other methods for defining domain boundaries. That is, a substructure found to cross a domain boundary might indicate a problem in an alternative automated assignment procedure. Most important, sub-

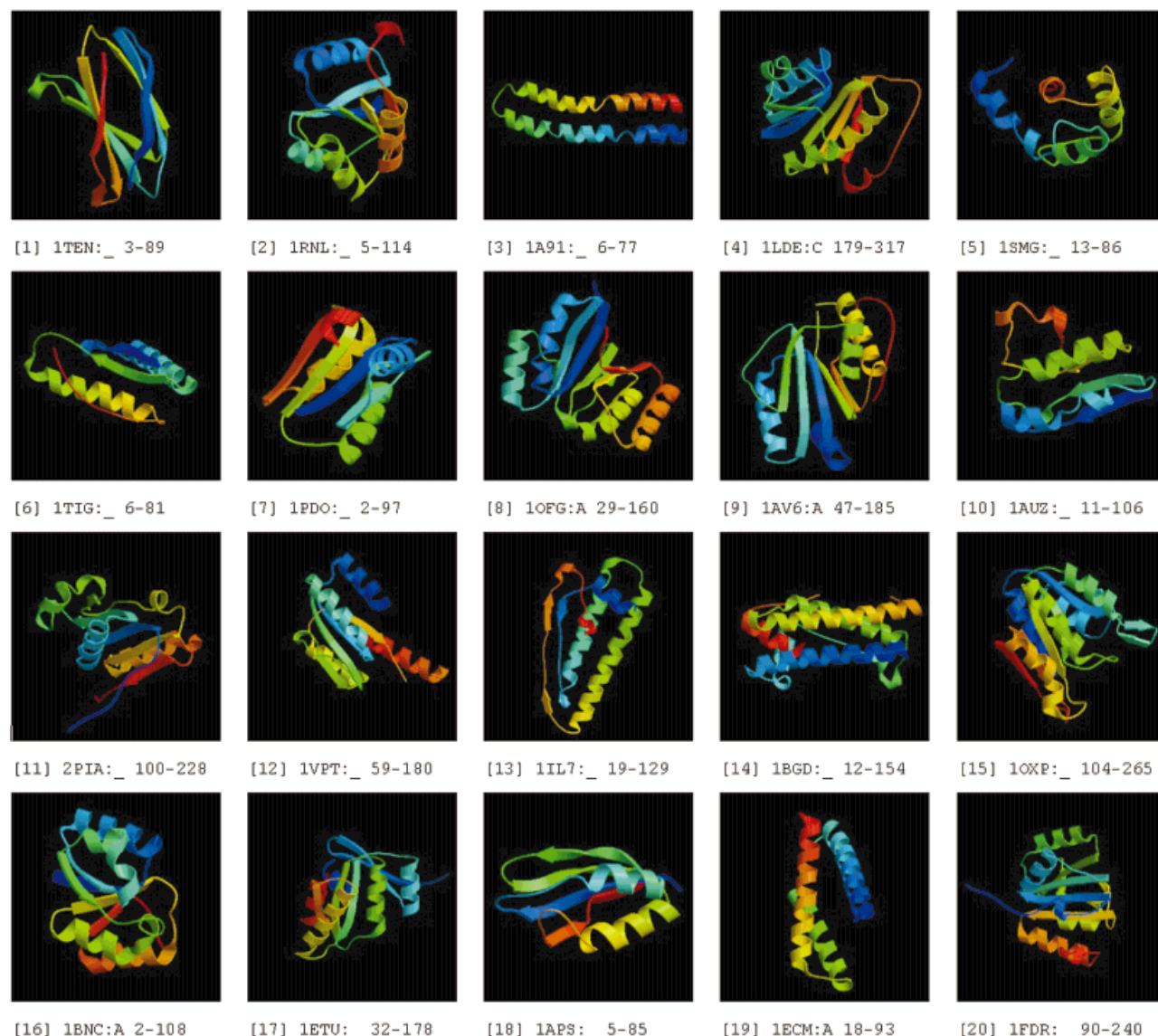


Fig. 5. Top 20 of 75 common substructures based upon a z-score > 4.0 and an rmsd < 4.0 Å.

structures can define a motif present in different domains thereby specifying a finer level of granularity than previously described in a systematic way.

Comparison of Common Substructures With scop Folds

Consider a typical example. Four neighbors (1LDE:C, 2PIA:_, 1VPT:_, and 1BNC:A) containing substructure [9] (1AV6:A), clearly a Rossmann-fold, are assigned to different folds by scop⁶ (Table III). These four neighbors are similar to the master representative 1AV6:A with a rmsd ranging from 0.6 to 3.8 Å. All four folds consist of parallel beta sheets with five or more strands and with alpha helices packed on both sides. In comparing structure topologies (Table III) it can be seen that there is an exact match in one case (2PIA:_). In the three other cases the neighbors from the common substructure have fewer

secondary structure elements than the scop fold: one helix less in 1BNC:A, three secondary structure elements less in 1LDE:C, and six secondary structure elements less in 1VPT:_. Pairwise structure comparison of these four common substructures result in a rmsd ranging from 3.4 to 4.4 Å (Table IV). These neighbors are in turn used in the definition of the other common substructures 1LDE:C - [4], 2PIA:_ - [11], 1VPT:_ - [12], 1BNC:A - [16] and can be found on Figure 6.

Thus common substructures as reported here represent a different clustering than provided by scop. While the overall topologies are the same, the substructures exclude secondary structure elements found in the scop folds based on full domains and assigned to specific functional classes. Again, the hypothesis is that the substructures presented here offer an alternative representation of fold space that transcend previously defined

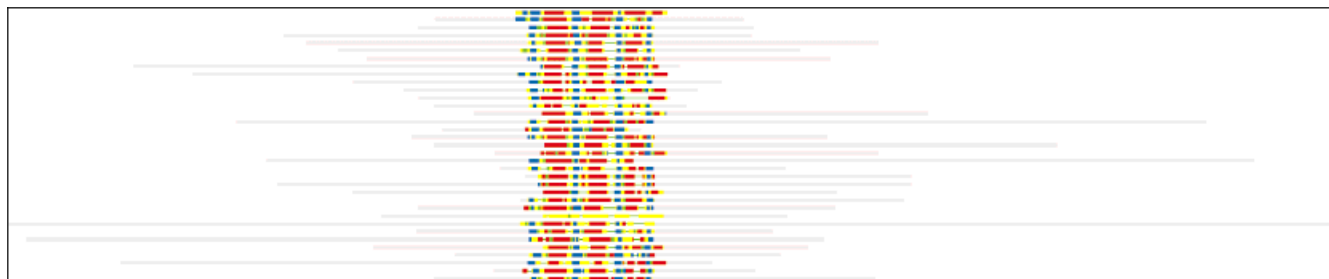


Fig. 6. Alignments of complete polypeptide chains to substructure [10], spoiaa, a phosphorylatable component of the system that regulates transcription factor sigma of bacillus subtilis (1AUZ_., residues 11–106). Color coding is as shown in Figure 2.



Fig. 7. Detailed view of alignments for substructure [10], spoiaa, a phosphorylatable component of the system that regulates transcription factor sigma of bacillus subtilis (1AUZ_., residues 11–106) showing specific region of the alignment. The columns represented are: arbitrary reference index; start and end residue number of subdomain structure;

start and end residue number of neighbor; index for neighbor in the database; overall length of neighbor chain; and resulting sequence alignment. The sequence is color-coded such that red is alpha helix, blue is beta strand, and yellow is unassigned.

structural boundaries defined in the scop sense and may provide structural scaffolds important to different functional classes of proteins.





Comparison With FSSP—All Against All

Scop is biased (in a positive way) by human intuition that brings into play other known features of protein classification. As such scop does not consider the systematic definition, at a finer resolution, the substructure presented here. Certainly they have been alluded to by various authors and characterized by Thornton and colleagues as the Russian doll effect—repetitive substructures contained within and spanning larger recognized folds. Why have these substructures not been seen in other automated structure comparison methods that are based on geometric relationships? A comparison of CE against

FSSP⁴ was made using the FSSP database of May 5, 1998 in an effort to address this question. FSSP is constructed using the Dali algorithm. To avoid the problem of each method using different versions of the PDB only the 11,050 polypeptide chains that were present in both the FSSP and CE databases was used in this analysis. It was necessary to reconcile the different notion of representative versus represented chains since the assignment of chains to these two groups was different for FSSP and CE. We assumed for both FSSP and CE that significant similarity detected for representatives implied similarity with the same level of significance for represented chains. Only similarities detected with a z-score threshold above 2.0 by FSSP and above 3.7 by CE were considered. The z-scores calculated for CE and FSSP are attributed to different statistical models and not comparable, therefore values recom-







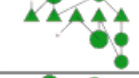







TABLE III. Structure Neighbors from the Definition of Common Substructures [9] 1AV6:A Assigned to Different Folds in Scop^{6†}

Chain		Substructure		Scop Assignment		Fold vs. Substructure
ID	Size	Rmsd ° Å	Begin-End	Fold	Begin-End	
1LDE:C	374	3.8	181–299	NAD(P)-binding Rossmann-fold domains	175–324	
2PIA: _	321	3.5	104–234	Ferredoxin reductase-like, C-terminal NADP-linked domain	104–223	
1VPT: _	348	0.6	47–185	S-adenosil-L-methionine-dependent methyltransferases	1–297	
1BNC:A	449	3.7	1–105	Biotin carboxylase N-terminal domain-like	1–114	

[†]Chain: ID—protein chain identifier in the form PPPP:C, where PPPP—PDB ID, C—chain ID within protein PPPP; Size—chain size as a number of residues. Substructure: Rmsd—rmsd of structure alignment between 1AV6:A and a given chain; Begin-End—substructure boundaries. Scop Assignment: Fold—fold name/description from the scop database; Begin-End—fold boundaries in the chain containing the substructure. Fold vs. Substructure—TOPS³¹ diagrams of folds: the overlap between the scop fold and the common subdomain is shown in green and parts of the structure only included in the fold but not in substructure are shown in red. Circles represent alpha helices and triangles represent beta strands.

TABLE IV. Comparison Between Structure Neighbors from the Definition of Common Substructure [9] 1AV6:A Assigned to Different Folds in Scop^{2†}

Chains	Z-score	Rmsd ° Å	Seq. ide. %	Aligned	Gaps	First Chain	Second Chain
1LDE:C vs 2PIA: _	4.4	3.4	6.2	96	34		
1LDE:C vs 1VPT: _	4.2	4.4	6.3	159	86		
1LDE:C vs 1BNC:A	4.9	4.0	8.9	101	5		
2PIA: _ vs 1VPT: _	4.1	4.0	11.1	108	46		
2PIA: _ vs 1BNC:A	3.9	3.7	7.7	91	36		
1VPT: _ vs 1BNC:A	3.5	4.2	11.1	119	72		

[†]Chains—protein chain identifiers in the form PPPP:C, where PPPP—PDB ID, C—chain ID within protein PPPP. Z-score—z-score of structure alignment. Rmsd—rmsd of structure alignment. Seq. ide.—sequence identities for structure alignment. Aligned—number of positions in the alignment. Gaps—number of gaps (non-aligned positions) in the alignment. First Chain, Second Chain—TOPS¹⁴ diagrams of folds: area of structural similarity between two chains is shown in green. Circles represent alpha helices and triangles represent beta strands.

mended by the authors as being statistically significant were considered appropriate.

A summary of comparisons between FSSP and CE is

given in Table V. From the set of similarities found by one method only 41% are found by the other method and the rest, 59%, are unique. Given these differences it is not

TABLE V. Comparison of Similarities Produced in All Against All Alignment Experiment by FSSP(Dali) and CE[†]

Chain Set (Size)	$n_{ce+fssp+}$	$n_{ce+fssp-}$	$n_{ce-fssp+}$	$n_{ce-fssp-}$
All chains (11,050)	811,633	1,182,016	1,174,450	57,877,626
CE representatives (1,868)	12,273	33,876	16,669	1,427,583
FSSP representatives (1,326)	7,643	9,422	27,422	830,016
CE and FSSP representatives (720)	4,018	4,132	5,393	245,297

[†]The numbers of chain pairs assigned as similar by both CE and FSSP ($n_{ce+fssp+}$), by CE but not FSSP ($n_{ce+fssp-}$), by FSSP but not CE ($n_{ce-fssp+}$) and not assigned by both CE and FSSP ($n_{ce-fssp-}$) are given.

TABLE VI. Top Ten Similarities Detected by FSSP but not Detected by CE[†]

Item	Chain 1	Chain 2	Z-Score	Rmsd (Å)	Alignment Positions		
					Aligned	Gaps	Gaps/Aligned (%)
1	1GOX:___	1AK5:___	22.4	2.5	245	137	56
2	3PTE:___	1PMD:___	13.3	3.7	226	229	101
3	1WAB:___	1ESC:___	13.0	2.8	170	139	82
4	3COX:___	1TRB:___	12.7	2.5	170	472	278
5	1PKM:___	1CTN:___	10.9	3.6	210	327	156
6	1X11:A	1IRS:A	10.7	2.0	93	47	51
7	1TRB:___	1GAL:___	10.4	3.6	197	486	247
8	1DIK:___	1AK5:___	10.3	3.4	199	401	202
9	1QBA:___	1CTN:___	9.3	3.9	241	475	197
10	1PKM:___	1AK5:___	9.7	3.9	181	271	150
Average				3.2	193	298	152

[†]Alignment Positions: Aligned—number of positions that match in both chains; Gaps—number of positions in both chains that do not match.

TABLE VII. Top Ten Similarities Detected by CE but not Detected by FSSP[†]

Item	Chain 1	Chain 2	Z-Score	Rmsd (Å)	Alignment Positions		
					Aligned	Gaps	Gaps/Aligned (%)
1	1NSJ:___	1IGS:___	6.3	2.7	188	28	15
2	1EDG:___	1BHG:A	6.0	3.2	257	130	51
3	2MNR:___	1NSJ:___	5.6	2.9	158	57	36
4	2SPC:A	1DKG:A	5.3	2.2	72	0	0
5	1GOW:A	1BHG:A	5.3	2.7	261	230	88
6	2SPC:A	1VSG:A	5.2	6.1	95	1	1
7	1MTY:B	1CIY:___	5.2	5.5	131	128	98
8	2SAS:___	1CM4:A	5.0	4.1	137	36	26
9	1TMI:___	1BHG:A	5.0	3.9	198	105	53
10	1LIS:___	1CIY:___	5.0	4.7	94	11	12
Average				3.8	159	73	38

[†]Alignment Positions: Aligned—number of positions that match in both chains; Gaps—number of positions in both chains that do not match.

surprising that different views of protein fold space emerge. This then raises the question of why are there so many apparent structure similarities detected by one method but not the other? To address this question we considered the top 10 similarities, based on z-score, that were found by one method and not the other (Tables VI and VII) using only the 720 first-order representative chains found in both databases. For these top ten similarities, those detected by CE have higher rmsd than FSSP, 3.8 Å versus 3.2 Å, respectively, but significantly less gaps 38% versus 152%, respectively. Thus Dali/FSSP does not penalize gaps and hence places a substantial number of gaps into the alignment to create matches and minimize rmsd. With these differences it is not surprising that the long substructures existing in representative structures and detected

with CE have not been characterized in the FSSP database.

Do these differences make one method preferable to the other? Do the results from Dali invalidate the idea of the substructures reported here? We suggest that the answers to these questions is no. Rather, it points to the need for further analysis of different views of protein fold space.

CONCLUSIONS

Using a previously reported Combinatorial Extension (CE) algorithm we compared the structures of all protein polypeptide chains greater than 30 residues in length (12,720) as found in the PDB. This resulted in 2,016 first-order representative structures that were the subject of further analysis. This analysis revealed recurring sub-

structures that were classified and compared to the classification schemes developed by others, specifically CATH, scop, and FSSP. This comparison revealed significant differences that need to be analyzed further. The 75 top substructures are found in from 7 to 106 first-order representatives with lengths ranging from 75 to 281 residues. A detailed analysis, including alignments and description for members of each cluster is available from <http://cl.sdsc.edu/ce.html>.

These substructures provide a systematic representation of fold space at a different granularity that described previously, and raise certain questions:

Is it meaningful to describe protein fold space at this level of granularity? This question can be phrased differently based on efforts by us and others^{33,34} to use structure alignments to strengthen property-based profiles defined by sequence alone. In other words, can these substructures be used to identify sequence motifs for which structures have yet to be determined and do they infer a functional significance? This question is being addressed by current work and will be presented elsewhere. Note here that several substructures [1], [19], [30], [46], and [63] contain immunoglobulin folds which have recently been characterized by structure superposition.³³ While sequence signatures are too weak alone to characterize the whole family, structures of Ig-like domains were subclassified based on structure alignments and manual review. Substructures determined here by automated methods alone cluster to conform to this classification; [1] is found predominantly in C2, C3, and Fn3 subtypes, [46] is found exclusively in C4 subtypes and [30] in C3 and H subtypes. The success of this automated classification and alignment provides strong structure-driven sequence profiles, the use of which are described elsewhere.

Given how different structure comparison algorithms treat gaps, what is the reasonable number of gaps one can put into an alignment and still consider it meaningful? A sufficient number of protein structures now exist to indicate that this is the wrong question to ask. Nature shows an extraordinary diversity of sequence and structure while preserving biological function. Different approaches to the comparison of protein structures would seem warranted in an effort to increase our understanding. One alternative view is presented here.

ACKNOWLEDGMENTS

We thank Tim Bailey, Lynn Ten Eyck, Michael Gribskov, Boojala Reddy, and Victor Solovyev for valuable discussions.

REFERENCES

1. Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 1980;136:225–270.
2. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins *EMBO J* 1986;5:823–826.
3. Orengo CA, Flores TP, Taylor WP, Thornton JM. Identification and classification of protein fold families. *Protein Eng* 1993;6:485–500.
4. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. A database of protein structure families with common folding motifs. *Protein Sci* 1992;1:1691–1698.
5. Holm L, Sander C. Mapping the protein universe. *Science* 1996;273:595–603.
6. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
7. Holm L, Sander C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res* 1998;26:316–319.
8. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993;233:123–138.
9. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. *Structure* 1997;5:1093–1108.
10. Taylor WR, Flores TP, Orengo CA. Multiple protein structure alignment. *Protein Sci* 1994;10:1858–1870.
11. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol* 1996;6:377–385.
12. Holm L, Sander C. Searching protein structure databases has come of age. *Proteins* 1994;19:165–173.
13. Godzik A. The structural alignment between two proteins: is there a unique answer? *Protein Sci* 1996;5:1325–1338.
14. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. *Proteins* 1995;23:356–369.
15. Wetlaufer DB. Nucleation, rapid folding, and globular interchain regions in proteins. *Proc Nat Acad Sci USA* 1973;70:697–701.
16. Islam SA, Luo J, Sternberg MJ. Identification and analysis of domains in proteins. *Protein Eng* 1995;8:513–525.
17. Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci* 1995;4:872–884.
18. Sowdhamini R, Ruffino SD, Blundell TL. A database of globular protein structural domains. *Fold Des* 1996;1:209–220.
19. Taylor WR. Protein structural domain identification. *Protein Eng* 1999;12:203–216.
20. Wernisch L, Hunting M, Wodak SJ. Identification of structural domains in proteins by a graph heuristic. *Proteins* 1999;35:338–352.
21. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM. Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci* 1998;7:233–242.
22. Holm L, Sander C. Dictionary of recurrent domains in protein structures. *Proteins* 1998;33:88–96.
23. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng* 1994;7:1059–1068.
24. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 1998;9:739–747.
25. Shindyalov IN, Bourne PE. Protein data representation and query using optimized data decomposition. *CABIOS* 1997;13:487–496.
26. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. *Protein Sci* 1992;1:409–417.
27. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. *Nature* 1994;372:631–634.
28. Wang ZX. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng* 1998;11:621–626.
29. Hubbard SR. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J* 1997;16:5572–5581.
30. Tsigelny I, Shindyalov IN, Bourne PE, Sudhof T, Taylor P. Do putative EF-hand motifs in cholinesterases and neurologins suggest Ca^{2+} binding sites? In press.
31. Westhead DR, Hatton DC, Thornton JM. An atlas of protein topology cartoons available on the World-Wide Web. *TIBS* 1998;23:35–36.
32. Waldrop GL, Rayment I, Holden HM. Three-dimensional structure of the biotin carboxylase subunit of acetyl-CoA carboxylase. *Biochemistry* 1994;33:10249–10256.
33. Halaby DM, Poupon A, Morron J-P. The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng* 1999;12:563–571.
34. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J Mol Biol* 1997;269:423–439.