

Connecting the Protein Structure Universe by Using Sparse Recurring Fragments

Iddo Friedberg and Adam Godzik

Running title: Connecting the protein structural universe

Keywords: Protein structural fragments, protein structure analysis, Gene Ontology, protein function analysis

Address correspondence to:

Adam Godzik or Iddo Friedberg
{adam, idoerg}@burnham.org
Program in Bioinformatics and Systems Biology
The Burnham Institute
10901 N. Torrey Pines Road, La Jolla, CA 92037
Phone: (858) 646 3168 Fax: (858) 713 9930

Abbreviations used in this manuscript: FBFS: Fragment Based Fold Similarity;
GBFS Gene Ontology Based Fold Similarity;

Summary

The quest to order & classify protein structures has led to various classification schemes, focusing mostly on hierarchical relations between structural domains. At the coarsest classification level such schemes typically identify hundreds of types of fundamental units called folds. As a result, we picture protein structure space as a collection of isolated fold islands. It is obvious however, that many protein folds share structural and functional commonalities. Locating those commonalities is important for our understanding of protein structure, function and evolution. Here we present an alternative view of the protein fold space, based on an inter-fold similarity measure that is related to the frequency of fragments shared between folds. In this view, protein structures form a complicated, cross connected network with very interesting topology. We show that inter-fold similarity based on sequence/structure fragments correlates well with similarities of functions between protein populations in different folds.

Introduction

How should protein structures be classified? Structure classification is important for understanding structural, functional and evolutionary relations between proteins. With the advent of high throughput approaches in structural biology and dawn of structural genomics we are deluged with protein structures, making this problem all the more important. The most popular and currently the only practical solution is the hierarchical, tree-like classification based on a concept of a unique three dimensional arrangement of a polypeptide chain called a fold, which can be further subdivided into superfamilies and families (Murzin et al., 1995; Orengo et al., 1997). Typically, known protein structures are classified into somewhere between 800-1200 elementary folds. Actually, most proteins have to be first divided into domains, which are elemental structural units of typically 50-150 amino acids in length that retain their overall structure, and also their function when found in different proteins (Rossmann and Argos, 1981). Most classifications work on the domain level, with larger proteins described as collections of specific domains.

When classifying the protein structure universe using hierarchical classification schemes one perceives a picture of a very discrete world, consisting of isolated fold islands which have an internal classification (superfamilies and families), but have sharp boundaries between them. While there usually is a coarser classification level, it typically focuses on a dominant secondary structure type (all-alpha, all-beta, alpha/beta -mainly antiparallel beta sheets- alpha+beta) and has a purely organizational role with no implied functional or evolutionary meaning. The jump from a handful of classes to several hundreds of folds is universal for most classification schemes and can be seen in SCOP

(Murzin et al., 1995) or CATH (Orengo et al., 1997). However, even simple analyses of few protein structures show many folds connected by similarities on an intermediate level: helical hairpins, alpha/beta/alpha units, etc. It was recently shown by Harrison *et al.* (Harrison et al., 2002) that there are a few folds which share a substantial local similarity with many other folds. Harrison *et al.* coined the term “gregarious folds”, to describe domains with folds which have significant structural similarities with domains from several other folds.

Recently, there have been a few studies offering a global view of protein structure space using other approaches. One is a principal component analysis of a distance matrix generated by an all vs. all comparison of SCOP representatives (Hou et al., 2003). Hou *et al.*'s view easily reconstructs the class-level view of protein structure space, and determines that the distance between folds is class, domain size and topology dependent. In a complementary work, Choi *et al.* extracted a representative set of local feature patterns using submatrices constructed from the distance matrices from representatives of protein folds. Each structure is then represented by the frequency distribution of these matrices, thus mapping protein structures into a common Euclidian space without secondary structure assignment or any structural alignment (Choi et al., 2004). However, in both these works the interfold distance was not interpreted in any way and was not related to function.

The study of protein fragments has become a favorite tool for investigating sequence-structure relationships (Han and Baker, 1996; Kolodny et al., 2002; Unger et al., 1989) and lead to new approaches for structure prediction (Haspel et al., 2003) and protein sequence alignments (Ye et al., 2003). Fragments, in contrast to domains, are not

expected to exist independently and to have unique evolutionary relationships. Similarities between fragments of different proteins are typically viewed as resulting from a limited number of structural choices in three dimensional space and fragment libraries are treated as sort of structural alphabets. As in words in human speech, no relation of any type is assumed between proteins sharing some number of letters in this structural alphabet.

Can fragments be applied to develop a better structure classification or at least to complement existing classifications? In this study, we use a specially constructed fragment data set to identify relations between different protein folds. In an approach we call **fragment origin agnostic** (Fragnostic) we selected fragments that simultaneously fulfill several criteria, including statistically significant sequence similarity and a presence in proteins with different folds. The rationale being that proteins which have overall different folds, but which share even a partial sequence and structural similarity, may be related in some --not necessarily evolutionary-- sense. We aim to complement the current hierarchical classification of proteins by adding additional connections between folds based on shared fragments between proteins in different folds. This we do using a novel fold-fold similarity measure based on a normalized count of fragments shared between folds. Consequently we amalgamate protein structure definitions at the most fundamental classification level, presenting a different and revealing view of protein structure space.

Another, not insignificant issue in protein classification is function. While by definition, functions of all proteins in a family are the same or highly similar, things get complicated at the fold level. There are many proteins within the same fold with different

functions, but there are also many proteins with different folds but the same (or similar) function. Some argue that it is not the overall fold, but rather selected local structural features that define protein function. While on a general level this statement is easily accepted, to the best of our knowledge this observation has not been used yet in enhancing our understanding of fold and function space. Using structural similarity (real or predicted) to infer functional similarity is seldom done beyond the family (i.e. obvious homology) level.

To hopefully rectify the situation, in this study we investigate a link between fragment based similarities and functional similarity. In addition to the aforementioned fragment based fold similarity measures, we propose a measure for fold-fold function similarity based on Gene Ontology. We then show that these two measures are correlated, and that inter-fold similarity can be associated with functional similarity. Finally, we discuss the structural and evolutionary implications of these findings.

The full results of the study, including an interactive viewer of the protein fold space as seen with the fragment based similarity measure, are available on a web server at <http://ffas.burnham.org/Fragnostic> .

A note on terminology: we use the term fragment data set in this study, rather than fragment library used commonly in literature. This is to distinguish the specific set of fragments used in this work from other works in the field, which have arrived at a smaller number of fragments, by clustering a data set into a smaller representative library (e.g. (Han and Baker, 1996; Kolodny et al., 2002; Tsai et al., 2000)).

Results

Towards a Different View of Protein Structure Space

This study aims at utilizing structure fragments in order to provide a view of protein structure space complementing the view offered by the hierarchical structure classification. To this end we have developed a specific set of fragments (Figure 1) and a fragment-based fold similarity score (FBFS) between folds. Since the Protein Data Bank is heavily biased in sequence and structure representation (Noguchi and Akiyama, 2003), we used the PDB-SELECT 25 dataset for our study, which is a set of all PDB protein sequences clustered at 25% identity. The 1668 proteins in this data set were used to produce sliding-window fragments of lengths 5, 10, 15 and 20 amino acids. Fragments were selected from this group by several criteria (see the Methods section) and then aligned to all the proteins in the PDB-SELECT 25 set. The FBFS similarity measure between folds was derived by counting the number of identical fragments aligned to proteins from these specific folds, normalized by the number of fragments shared by proteins within a fold. Thus FBFS can vary from 1 (strong similarity) to 0 (no similarity). Generation of the fragment data-set is described in figure 1, and explained in detail in the Methods section.

The fragment based similarity measure between folds was used to represent protein fold space as a weighted graph, whose nodes represent folds (or more precisely sets of PDB-SELECT25 proteins in a given fold as defined in SCOP), and the edges and their weights represent fold similarity as measured by the FBFS score. Figure 2 shows three sub-graphs for fragment length 10, with an FBFS threshold of 0.2 (figure 2a) and 0.15 (figure 2b). We present the protein fold space graph at two different thresholds in order to

demonstrate how the connectivity depends on the FBFS-based score. Both Figure 2a and 2b show the same area of the fold space, with more folds connected at the lower threshold (2b) than at the higher one (2a). Custom thresholds may be applied, and the structure of the sample fold space can be explored using the Fragnostic web site at <http://ffas.burnham.org/Fragnostic>.

We illustrate several general observations in Figures 3-5. First, the fold space graph does not have full connectivity. Figure 3a shows the distribution of connections per fold as a function of fragment length. Generally, we see that there are few nodes with many connections, and many nodes with few connections. When fitting a regression curve to the data points with one or more connections, the best fit was obtained using a power-law ($y=ax^n$) function. Figure 3b shows the dependency of this distribution upon the FBFS threshold for the network formed by a fragment length of 10, a similar picture can be obtained for other fragment lengths. As discussed earlier, the higher the FBFS threshold, the fewer connections there are between the nodes of the graph.

Within the global fold space graph, there are smaller groups of folds forming highly connected subgraphs. One such subgraph is composed of g.41, g.50, g.37, g.39 (see Table 1 for more details on these folds), which share a common structural feature. Figure 4a shows the ribbon diagram of the four folds from this example, with shared fragments' color highlighted. All four proteins have a dimetal binding loop, performing its function via variations on the CXXC and/or CXXH motif. The connecting fragments overlap wholly or almost completely with each other as shown on Figure 4b. These functional loops are the common denominator of otherwise different structures. The actual function

of each protein is may be different, although the structural implement (a loop plus something else, chelating a dimetal) is the same.

Other highly connected sub-graphs, may form due to sharing of several different structural features, rather than a single motif. For instance, users of the SCOP database know that the Rossmann fold proteins are classified into several different SCOP folds. Clearly, the SCOP group decided after closer examination that the differences between these proteins warrant a classification into separate folds, but the similarity is acknowledged in the SCOP annotation to many of these folds, which are often called Rossmann-like. This type of classification problem can be alleviated by the graph based view of the fold space, as shown in figure 5. Here, several Rossmann-like folds are connected; showing that inter-fold similarity can be quantified using the FBFS similarity measure. Here we see c.2 which is the NAD(P) binding Rossmann fold domain, c.4 is a nucleotide binding domain and c.3 is an FAD/NAD(P) binding domain. The three folds are known to be related, and the graph in fig 4 illustrates that. Additionally, we have found that c.32, the GTPase domain of Tubulin, and d.108 the Acyl-CoA N-acetyltransferases fold share a significant similarity with c.2. It is interesting to note that c.111 (Molybdenum cofactor biosynthesis protein MoeB fold) shares fragments with both c.2 and c.3. All three are three layered alpha-beta-alpha folds, which differ mainly in the topology and number of the beta strands in the sheet. It is still not clear if situations like this should be interpreted as suggestions of a common evolutionary origin for seemingly disparate folds, examples of forces of function driven convergent evolution at work or illustration of physical limitations of the fold space.

Functional Association in the Fragnostic Graph

FBFS delineates relationships between different folds that define an interesting non-tree like organization of protein structural space, but the question remains whether this relationship has any biological relevance. We examine whether the inter-fold similarity (as defined by FBFS) correlates with the similarity of functions between proteins populating the folds. Proteins within the same superfamily --by definition-- have the same function, and functional similarity often extends to cover an entire fold. However, proteins which have different folds rarely have similar functions and when they do it is usually considered a random event (non-homologous replacement). Therefore, a correlation between FBFS and functional similarity would go a long way towards suggesting that these coarser-than-the-fold-level structural similarities are not spurious and may be connected to distant homologies or parallel evolution of functional features.

In order to test this hypothesis, we need to first define a quantitative functional similarity measure based on Gene Ontology (Harris et al., 2004). Gene Ontology annotates proteins using a hierarchy of terms which progresses from general descriptions to more specific ones. In the case of functional annotation, this means going from a general description such as “enzyme activity”, through a more specific one such as “phosphatase”, towards a specific description such as “protein tyrosine phosphatase”. Such framework allows developing a quantitative measure of functional similarity between folds we call here GO Based Fold Similarity (GBFS). Briefly, this functional similarity measurement is built upon Lord *et al*'s semantic similarity based measure (Lord et al., 2003a; Lord et al., 2003b). Lord *et al*'s similarity measure defines a pairwise similarity between two terms from the GO graph, based on the probability of the minimal

subsuming term to those two terms in the GO hierarchy. Here we extend this measurement in order to measure the similarity between two folds. Each fold is assigned a set of GO terms from the “molecular function” ontology. This is done based on GOA-PDB, a GO-based PDB annotation provided by EBI (Camon et al., 2004). Given two sets of annotations, one from each fold, the function-based inter-fold similarity is the mean semantic similarity between the two folds. See Methods: fragment based fold similarity, for details.

Figure 6a shows the correlation between the fragment based fold similarity (FBFS) and the GO-based fold similarity score (GBFS). As can be seen, there is a significant correlation between the two similarity measurements for fragments of length 10 and 15. To verify the advantage of a fragment-based approach in identifying functional commonalities as compared to other more widely used tools, such as distant homology prediction, we looked for a correlation between mean inter-fold FFAS03 scores and GO-based fold similarity scores. As shown in figure 6b, in this case there is no significant correlation.

Traits of the Fragment Data Set

We now turn to examine particular characteristics of the fragment sets developed in this work. In this section we show that our sets provide only sparse coverage of protein structures and that there is a good correlation between structural compactness of a fragment and the number of different folds it fits into. Both are very interesting because they were never built in as assumptions and actually came as complete surprises. We suggest that taken together, these traits indicate that some fragments in these sets are

precursors for “structural building blocks”, which may be shared even by proteins with different folds.

Structure Coverage

How much of protein structure space does our fragment data set cover? We can define coverage as composed of two parameters, namely *length fraction*, and *fold diversity*. The information we would like to obtain is best phrased by the following questions:

- 1) Length fraction: what fraction of the protein length is covered by fragments from our set? This question is somewhat complicated by the fact that segments of a protein can be covered by one or more partially overlapping fragments. To avoid this problems we calculate the coverage as: L_f/L , where L_f is the total length of the protein covered by fragment disregarding the overlap, and L is the protein chain's length.
- 2) Fold diversity: to how many different folds can a given fragment be aligned? This is a very interesting point, since, as we will show, certain fragments show very high fold diversity, meaning that they can be aligned with fragments originating in many other folds: aligned both in profile and in structure. Taken together, surveying length fraction coverage and fold diversity will give us a picture of how well our fragment data set samples structure space. We are interested in fragments which have a fold diversity of two or greater, because we are looking for commonalities between different folds.

Most of the previous applications of fragment libraries were in modeling and structure prediction, where by design the aim was to cover the entire length of proteins. In this context, length fraction coverage vs. quality of structure predictions has been addressed previously (Hubbard, 1999; Rychlewski and Godzik, 1997). Here we maintained rigorous criteria in generating the fragment data set (see the Methods section for details), so the length fraction is a simple result of the thresholds in choosing the fragments. Figure 7a shows the distribution of length coverage by data-set fragments for different fragment lengths. As can be seen, the length coverage is particularly low using fragments of length 5, with a median coverage of 25%. Fragments of lengths 10-20 exhibit higher length coverage, although the median coverage is still only 34%.

The question of fold diversity is illustrated in figure 7b. As can be seen, fold diversity decreases with fragment length increase. This is expected, as longer fragments will be more fold specific. The contrast between extremes as dependent on fragment length is quite striking though: 90% of the 20-length fragments cover only ten folds, whereas 90% of the 5-length fragments cover some 40 folds.

A Possible Structural Role for High Fold Diversity Fragments

As shown earlier, many fragments in the fragment data set have a high fold diversity. Such fragments clearly contain a strong local structure signal that is present in proteins with very different global structure. It is an interesting phenomenon, possibly pointing to a fundamental role of such fragments in protein structure and folding. The building block model of protein folding has been used to explain theoretical and experimental observations (Pedersen and Moulton, 1995; Tsai et al., 2000). Here we wanted to test if the

high diversity fragments fulfill the requirements for the building blocks in this model. We evaluate their compactness using the radius of gyration (R_g) and ask the question whether compact fragments characterized by low R_g value will have the same or larger fold diversity as those with a high R_g . To answer this question, each fragment was scored based on the size of its R_g relative to the median of the R_g of all the fragments from the protein of origin. This is determined using a sliding window along the chain from which that fragment was taken, the sliding window having the same length as the investigated fragment. The results are shown in figure 8. For fragments of length 10, 15 and 20, it is shown a small radius of gyration is associated with high fold diversity, suggesting that indeed some of those fragments may play a role as protein building blocks appearing in unrelated folds. Despite the preference for compactness, there is no concurrent preference for fragment burial. For fragment length 5, these conclusions do not hold, since they are probably too short and non-specific for this type of analysis. See Methods and the legend of figure 8 for details of the sliding window R_g analysis.

Secondary structure content of fragments

Another topic of interest is the distribution of secondary structure elements (SSEs) in the fragment data sets of different length. Is there a bias in SSE content between the sets, and if so can this bias explain the differences between the data sets? To investigate this issue we first examined the secondary structure (SSE) distribution in PDB-SELECT25. We found that 37% of the residues are in alpha helices, 22% in beta strands and 41% in turns, loops & coils. When compared to PDB-SELECT25, in fragments of length five, strands are over-represented by 12% and helices are under-represented by 12% (see Table

2). In fragments of length 15 and 20, helices are over-represented by 13%, and turns & loops are under-represented by the same proportion. The least bias occurs in fragments of length 10, in which helices and strands are slightly (+5% each) over-represented, at the expense of loops & turns. These differences are probably due to the fact that during the generation of fragment data sets, the best structural fits are found within ordered secondary structure elements. Since beta strands are shorter than alpha helices, the fragments of length five are more biased towards beta-strands, and the longer fragments are more biased towards alpha-helices. For fragment length 5, an over-representation of strands may indeed explain some of the behavior of this fragment set, especially with respect to the non-specificity the set seems to have. However for the longer fragment lengths, the differences in SSE distribution appear to be minor when compared to the distribution in PDB SELECT25, and the bias, if any, is towards the more ordered elements (strands and helices) at the expense of loops and turns.

Discussion

In this work we have applied a specially developed fragment data sets towards a new view of the protein structure universe. This view breaks out of the traditional view of this space as composed of isolated fold-islands, and instead shows it as having a graph-like structure with connections of varying strengths between different folds, Furthermore the strength of these connections can be quantified and is shown to be correlated with functional similarity.

The fragment lengths of 5, 10, 15 and 20 amino acids were chosen in order to provide a “good spread” of coverage of folds. Length five fragments are on the lower border of any kind of fold specificity, as is suggested by figure 6b. Additionally, the 5-length fragments are the least representative of secondary structure content of proteins, and FBFS based on length five fragments correlates poorly with GBFS. Figure 6b shows that at the other extreme -- fragments of length 20 -- the fragments are quite fold specific, where 90% of the fragments cover only ten folds. Between those extremes, the 10 and 15-length libraries seem to produce the most biologically interesting results. We conclude that using our fragment data set generation system, fragments of length five are probably too non-specific and non-representative to convey a meaningful picture of a fragment-based view of structure space, whereas fragment of length 20 are too specific to provide an adequate representative picture of fold space.

Several examples of the data-sets’ fragments suggest that they may correspond to local functional sub-domains, such as a dimetal-ion binding loop, present in domains with completely different folds & functions. This is an example of an identical structural element (tight loop) with a similar basic molecular function (ion binding), but used in different biochemical functions in overall different protein folds. We expect that further analysis would identify more such examples. Other interesting characteristics of our fragment data set include a very broad fold distribution of some fragments, and the tendency for those fragments to be compact despite lack of preference for being buried or localized in the protein core.

We have also shown that there is a statistically significant correlation between fragment based fold similarity (FBFS) and the Gene Ontology based fold similarity (GBFS) for fragments of lengths ten and longer. This suggests that fragment sharing is correlated with functional similarity in a way that extends beyond what can be detected using current sensitive profile-profile alignment methods and what would fit into traditional view of protein evolution. At least two different explanations come to mind: the first is that such fragments correspond to extremely distant homologies between functional fragments that are being shuffled and exchanged between unrelated proteins, and the second being that such fragments represent examples of a convergent, function driven evolution on the sub-domain level. Interestingly, these findings complement those of other studies (Shakhnovich et al., 2003a; Shakhnovich et al., 2003b), which have shown that structural clusters have a unique “functional fingerprint”, and even if the structural cluster is increased by relaxing the similarity threshold, the expanded cluster still maintains a similar functional fingerprint.

Despite the significant correlation between GO based fold similarity and fragment-based fold similarity, many of the fragments contained in the data-set cannot be readily explained for their role in protein structure or function. They may, however, reveal a phenomenon of common substructures necessary for maintenance of protein structure or function, regardless of the actual fold of origin. A good analogy to this phenomenon would be the prevalence of common architectural motifs such as pillars or arches in many types of buildings, although the buildings they appear in are unrelated in their overall appearance and purpose. This hypothesis is also supported by the observation that the

Fragnostic graphs at low (< 0.3) FBFS exhibit few folds which have many connections, and many folds with few connections. These results merit further investigation: can we identify in those elements in the “hub folds” which are shared by many other folds structural building blocks? And how come a few folds have become “hub folds”? Our results of sequence-structure mapping of fold space synthesize an apparent discrepancy between the studies of Harrison *et al* (2002) and of Kihara & Skolnick (Kihara and Skolnick, 2003), both which deal with inter-fold similarities. On the one hand, Harrison *et al* describe fold space as mostly discrete, with a few “gregarious” folds serving as a hub for several other folds. On the other hand, Kihara & Skolnick have discovered that by relaxing structural similarity conditions, structure space can be viewed as continuous. Here we offer a technique for studying the continuity *versus* discreteness of fold space by using a fragment data set. We propose that future work in this area can be carried out with at least two aspects: one would be a computational level, optimizing the fragment data sets and similarity measure. Another would be a biological level, the study of the hub folds and tracing the biological implication of their position in protein structure space.

Methods

Construction and Structure of the Fragment Data set

The PDB files from the PDB-SELECT25 list were used in the construction of the fragment data set. Each chain in PDB-SELECT25 was subject to a PSSM construction using FFAS03 (Rychlewski et al., 2000) (also Jaroszewski *et al*, in preparation). FFAS03 produces specific PSSM using PSI-BLAST identified homologs from a subset of NCBI's

non-redundant protein sequence database (nr), and certain corrective measures enhancing sensitivity. (For details regarding FFAS03, see section *FFAS03*, below). Sliding windows of lengths 5-20 were passed over the profiles and the resulting sub-profiles were compared to each other using FFAS03 profile alignment scoring method. Each fragment pair was assigned a score based on its distance from the mean of the score distribution in its length category. Only fragments a p-value of < 0.001 were included in the fragment data set termed Screen 1.

The following data were obtained for each fragment: secondary structure, solvent accessibility, atomic coordinates of C α amino-acid sequence, SCOP family classification and radius of gyration. For each fragment pair, the FFAS03 alignment score and the RMSD of the structural alignment were noted. The filtering process was then taken further, and only those fragment pairs with an RMSD $< 1 \text{ \AA}$ were examined. Fragments were structurally aligned using the Superpos algorithm for calculating the optimal matrix of rotation (Sippl, 1991), C-alpha atoms were used for the fragment superimposition. Fragments with a C α -RMSD $< 1.0 \text{ \AA}$ along their alignment length were deemed structurally well-aligned. The data set, used for the construction of the Fragnostic graphs and their analysis is the one derived from both profile and structure similarities, and is called Screen-2.

Secondary structure analysis

Three SSE types were used: helix, strand and coil. Secondary structure along the fragments was determined according to DSSP (Kabsch and Sander, 1983). As DSSP provides more than the three SSE types used in this study, we mapped the DSSP SSEs

onto a three letter SSE alphabet as follows: DSSP codes G, H, I are “helix”; E and B are “strand”, T, S and no code are “other”.

SCOP family classification of fragments

Each fragment was tagged with the four-position code of its origin in the SCOP database of structural protein classification (Murzin et al., 1995). Due to the vagaries of PDB residue numbering, the ASTRAL compendium (Brenner et al., 2000) was used to assign the correct fold of origin for each fragment, overcoming problems presented by multi-domain chains.

Radius of Gyration

Radius of gyration for a fragment was determined according to:

$$Rg = \sqrt{\frac{\sum (m_i \cdot R_i^2)}{\sum m_i}}$$

Where R_i is the distance between the backbone atom i and the fragment’s center of mass. m_i is the atomic mass of backbone atom i . Only backbone atoms were considered in calculating a fragment’s center-of-mass and radius of gyration. The center-of-mass was calculated as the weighted vector sum of all backbone atoms in a fragment.

In order to calculate a normalized R_g score for a fragment, a sliding window was used to record the radii of gyration for all fragments in the protein domain of the fragment to be scored. Sliding window size was 5, 10, 15 or 20, as appropriate. The fragment's score is in standard deviations from the median, as the distribution of the radii of gyration was found to be skewed with a long right tail. The calculation was performed as follows:

$$\bar{Rg} = \text{median}([Rg(1,1+k), Rg(2,2+k), \dots, Rg(L-k+1,L)])$$

$$Z(Rg) = \frac{Rg - \bar{Rg}}{\sigma}$$

Where k is the fragment length, $Rg(i, i+k)$ is the radius of gyration of a fragment starting at residue i , L is the domain length, σ is the standard deviation from the median and $Z(Rg)$ is the Z-score of Rg .

Fragment Based Fold Similarity

We have developed a fragment-based fold similarity score (FBFS) between folds which measures the similarity between folds based on the number of fragment pairwise alignments between proteins populating those folds. FBFS is a pairwise similarity measure based on a normalized count of fragments shared between the proteins from two different folds.

Given n folds, indexed $(1, \dots, n)$

Each fold will have a set of fragments shared with other folds: (X_1, X_2, \dots, X_n)

X_i being the set of all fragment pairs which are shared in fold i . $|X_i|$ is the number of those pairs.

$X_{i,j}$ is the set of all fragment pairs shared between fold i and fold j and $|X_{i,j}|$ is a number of such pairs.

FBFS is defined as follows:

if $|X_{i,j}| > \text{thresh}$:

$$FBFS(i, j) = \begin{cases} i \neq j : \max \left[\frac{|X_{i,j}|}{|X_i|}, \frac{|X_{i,j}|}{|X_j|} \right] \\ i = j : 1 \end{cases} \quad (1)$$

The number of shared fragments $|X_{i,j}|$ is normalized by the number of fragments in the fold with the smaller fragment population. Thus in cases when

$$|X_j| \ll |X_i|$$

so that

$$|X_{i,j}| / |X_i| \ll 1 \text{ but } |X_{i,j}| / |X_j| \leq 1,$$

then the larger measure is taken. In other words, if fold j shares a large fraction of its fragments with fold i , but fold i shares a small fraction of its fragments with fold j , then normalizing by $|X_j|$ will preserve that information. The FBFS can be thresholded, as many folds share fragments spuriously. A higher FBFS threshold means that only folds sharing a high normalized count of fragments would be displayed.

However, this may create a bias towards folds which have few representative proteins, or a single representative one, so that even a single fragment shared between

fold might pass a high FBFS threshold, if it was the only fragment shared between that fold and another one. To avoid that, FBFS can be corrected by inserting a threshold requirement of an absolute number of shared fragments.

The measure of fold self-similarity by using FBFS is not self-evident: $FBFS(i,i) = 1$ is not necessarily true. The reason is that intra fold similarity is only being checked between different chains in the fold. Thus, for example, in the extreme case that a fold i is populated by a single chain, $FBFS(i,i) = 0$. Therefore, we state explicitly that when $i=j$, $FBFS=1$.

GO-Based Fold Similarity (GBFS)

As Per Lord *et al* (2003) the similarity between any two terms in a given GO ontology $gosim(c1, c2)$ is defined as:

$$P_{ms}(c_1, c_2) = \min_{c_i \in S(c_1, c_2)} \{P(c_i)\}$$

$$gosim(c_1, c_2) = -\log_2(P_{ms}(c_1, c_2))$$

Where:

c_1, c_2 : two terms from the same ontology

$P(c_i)$: the frequency of term c_i in the PDB-SELECT25 database

$S(c_1, c_2)$: the set of all subsuming GO terms for terms c_1, c_2 .

$P_{ms}(c_1, c_2)$: the frequency of the minimal subsuming term for terms c_1, c_2 in the ontology.

Given two sets of proteins, set A has n protein and set B has m proteins from folds A and B respectively:

$$A=[a_1, \dots, a_n]$$

$$B=[b_1, \dots, b_m]$$

Each protein is associated with one or more functional terms from GO. Thus we have a set of GO terms for the proteins in folds A and B.

$$A' = [g_1, g_2, \dots, g_n]$$

$$B' = [h_1, h_2, \dots, h_m]$$

The GO based fold similarity score (*gbfs*) was then calculated as follows:

$$gbfs(A, B) = \frac{\frac{2}{n+m} \sum_i \sum_j gosim(g_i, h_j)}{\frac{1}{n} \sum_{i>j} gosim(g_i, g_j) + \frac{1}{m} \sum_{i>j} gosim(h_i, h_j)}$$

Determining graph connectivity using FBFS as a similarity measure

For the data depicted in figure 3, the graph nodes and edges were counted, and set in a histogram of node frequency vs. number of edges. For each fragment length, a logarithmic regression, and exponential regression, and a power regression were fit, and the correlation coefficient for each fit determined. Power regression ranked highest (data not shown). Significance of the correlation coefficients was determined using the paired values t-test. The fitting was carried out using Grace 5.1.14 (<http://plasma-gate.weizmann.ac.il/Grace>)

FFAS03

The FFAS program for sequence profile based comparison has been a subject of several publications (Jaroszewski et al., 2000; Rychlewski et al., 2000). Several modifications implemented in its newest version (FFAS03) are described in an upcoming publication (Jaroszewski *et. al.* in preparation). Briefly, a multiple alignment recovered from PSI-BLAST run on the NR sequence database clustered at 85% sequence identity (NR85) is reanalyzed to calculate a variant of a Position Specific Scoring Matrix (PSSM) or a profile. A special two-dimensional weighting scheme is implemented to assure equal contribution from unevenly matched groups of sequences. Also, in contrast to the PSI-BLAST weighting scheme, global similarities between sequences in the multiple alignments are used in weighting contributions on single residue level. The resulting PSSM is then rescaled to ensure a uniform mean and distribution of scores in each column and across the entire database. Finally, a special scoring system between columns in the two different PSSMs is introduced, which changes the scalar product of two columns for profiles with large sequence statistics, to a product weighted by a similarity matrix for profiles based on small number of sequences.

The calculation of multiple sequence alignment

Each representative sequence from PDB-SELECT25 is used as an input for five PSI-BLAST iterations, or until convergence. An e-value inclusion threshold of < 0.001 was used. PSI-BLAST was run over NR85, a subset of NCBI's protein sequence non-redundant database, with sequences clustered so that no two have a positional sequence

identity of more than 85%. To obtain NR85, the sequences in NR were clustered using CD-HIT (Li et al., 2001).

Fragment Alignment and Scoring

Given two profiles, F1 and F2, obtained from sequences S1 and S2 with lengths L1 and L2 respectively, a matrix C of size L1xL2 was created. Each cell in the matrix, C_{m,n} has the following value:

$$C_{m,n} = \sum_{a=1}^{20} \sum_{b=1}^{20} f_{m,b} \cdot B_{a,b} \cdot f_{n,a}$$

Where *a,b* are any of the 20 amino acids, and *m, n* are positions along profiles F1 and F2, respectively.

f_{m,b} is the fractional score for amino acid *b* in position *m* in profile F1.

f_{n,a} is the fractional score for amino acid *a* in position *n* in profile F2.

B_{a,b} is the BLOSUM62 score for amino acids *a,b*.

Scoring the Fragment alignments

Given a fragment size *r*, the scores for a sliding window fragment are calculated as follows:

$$S_{m,n} = \sum_{i=n-r/2}^{n+r/2} C_{m,i}$$

Figure and Table Legends

Table 1:

The folds participating in the sub-graph containing predominantly dimetal-chelating loops depicted in figure 3.

Table2:

Percentage of secondary structure elements (SSEs) in the Fragnostic data sets, and in PDB-SELECT25. H: helix; E: strand; T: other (mainly loops or turns). The values in the boxes are the percentage of each element for fragments of length 5, 10, 15 and 20.

Figure 1: construction of the fragment data set

(A) Each sequence in PDB_SELECT25 (I) is used to generate a PSI-BLAST based profile, using FFAS03 (II). A profile is a $20 \times L$ matrix, each cell representing the normalized amino-acid frequencies for each position in the query sequence, L being the query sequence's length. In this cartoon, the cell values are shown as a grayscale heat map. Each profile is subject to a sliding window, of lengths 5, 10, 15 and 20 (III & IV).

(B) The profile fragments from all PDB_SELECT25 are compared, all vs. all, and the high-scoring matches (see text, fig 1b for details), are saved separately according to fragment length. Screen-1 are high scoring sequence-profile based alignments, FFAS03 based p -value < 0.001 . The Screen-2 data set contains Screen-1 fragments which have been structurally aligned, and found to have an RMSD $< 1 \text{ \AA}$. The final data set contains only those pairwise alignments which have a $p < 0.001$ FFAS03 score and an RMSD $< 1 \text{ \AA}$.

(C): Distribution of FFAS03 scores in the Screen-1 data set

Sliding window of lengths 5, 10, 15 and 20 were taken along whole chain profiles generated of the PDB_SELECT25 chains by FFAS03. All fragments were scored in pairwise alignments. The X-axis shows the FFAS03 score, and the Y-axis is the percent of total. Samples were taken from the extreme of the long tail.

Figure 2: Illustration of graph connectivity dependent on FBFS threshold

(a) Part of a Fragnostic graph for fragment length 10, and FBFS threshold of 0.2. Circles are the SCOP fold populations, color coded according to SCOP class. Red: all alpha; Blue: all beta, Orange: alpha/beta; Green: alpha + beta; purple: small. (b) Same for a lower FBFS threshold, 0.15. Note more connected folds. The area of fold space surveyed is the same one.

Figure 3: Graph connectivity, fragment length and FBFS threshold

(a) Histogram of node connectivity for different fragment lengths. When excluding the nodes with zero connections, the graph was found to fit the general form of $y=a^x$ for fragment length 5 with a correlation coefficient of -0.8; for fragment length 10 $cc=-0.92$; for fragment length 15 $cc=-0.93$, for fragment length 20 $cc=-0.75$. The p values were determined using a paired values t-test, $p < 0.00001$ for all correlations. (b) Histogram of node connectivity for different FBFS thresholds, fragment length 10. The x-axis is the number of edges per node; the y-axis is the percent-of-total nodes having that number of edges.

Figure 4: Shared contigs between different folds containing cysteine-rich metal binding loops

- (a) Four structures of cysteine-rich metal binding domains are shown, each taken from a different fold. Contigs making up the loop, and the metals binding them are shown in color, the rest of each chain is shown in white. Color codes, PDB codes and SCOP folds are: Yellow, 1yuj-A (Omichinski et al., 1997), g.37; green, 1ryt (deMare et al., 1996), g.41; red 1g47-A (Velyvis et al., 2001); g.39; blue, 1vfy-A (Misra and Hurley, 1999), g.50. See table 1 for details.
- (b) Shows the backbone of the structural alignment of the underlined positions of the sequences in (a). Mean and maximal alignment C α -RMSD between any two fragments is $< 1\text{\AA}$.
- (c) The region in the fold connectivity graph from which the four representatives were taken. The circles represent the graph vertices, and are colored according to fold, the lines are edges, the shorter the edge, the larger the fragment based fold similarity between the two folds the edge connects. See also figures 3, 5 and text for details.

Figure 5: Graph-based representation of inter-fold similarities between Rossmann-like folds

c.2, c.3, c.4 and c.111 are all three layered alpha-beta-alpha folds varying mainly in the number and topology of beta strands in the middle sheet. This graph representation enables to quantify the similarity between those folds. Other folds also share fragments. See text for details.

Figure 6: Correlation of FBFS with GBFS

- (a) The linear correlation of fragment-based fold similarity with inter-fold functional similarity. For brevity, each data point is the mean values of all FBFS values in a 0.2 sized bin of GBFS. A significant correlation for all points is seen for fragment lengths 5 (black circles), 10 (red squares) and 15 (green diamonds). The correlation coefficients were found to be $r_5=0.17$, $r_{10}=0.45$ and $r_{15}=0.46$ respectively. We tested the null hypothesis that $r_n = 0$ using Fisher's Z transformation for correlation coefficients, n being a fragment length. The null hypothesis was rejected for all cases with a p-value of $p < 0.0001$ for each fragment length. Correlation was not found to be significant for length 20 (triangles).

(b) Comparing the correlation of FBFS for fragment lengths of 15, and mean FFAS03 scores between folds. Data points represent the same as in (a). The correlation coefficient, $r=0.12$ for FFAS03 vs. GBFS, with a p value $=0.1$, which shows no significant correlation (black circles). For FBFS vs. GBFS $r_{15}=0.46$ with $p < 0.0001$ the correlation is significant (red squares). FFAS03 scores along the x-axis are normalized on a 0--1 scale by the highest score obtained.

Figure 7: Cumulative distribution of sequence coverage and fold diversity using different fragment lengths

(a) Cumulative distribution of sequence coverage. X-axis: bins of sequences by percent coverage; Y-axis: cumulative percent of total sequences per fragment length. Note the particularly low coverage provided by fragment length 5. See text for details.

(b): Cumulative distribution of fold diversity. X-axis: fold diversity: number of folds; Y-axis: cumulative percent of total fragments for each fragment length. Note the high fold diversity exhibited by fragment length 5.

Figure 8: Enrichment of the low R_G fragment population with fragments with high Fold Diversity

The Radius of gyration (R_g) for each fragment was noted. Then, each fragment's R_g was expressed in standard deviations from the median R_g of all fragments for the chain of origin. The median R_g was determined from the R_g of a population of sliding windows along the chain of origin. See Methods for details. Each data point is a single fragment. X-axis: R_g score, in standard deviations from the median. Y-axis: fold diversity: number of different folds in which a given fragment finds a pairwise profile/structure alignment to at least another fragment. (a) for length 5 fragments; (b) for length 10 fragments; (c) for length 15 fragments; (d) for length 20 fragments. No fragment enrichment was apparent in fragment length 5 (e-value > 10). However, enrichment was significant for fragments length 10 (e-value $< 1 \times 10^{-50}$), 15 (e-value $< 1.81 \times 10^{-6}$) and 20 (e-value $< 9.96 \times 10^{-6}$) χ^2 goodness-of-fit test was used to determine the significance of the enrichment. See text for details.

References

- Brenner, S. E., Koehl, P., and Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* 28, 254-256.**
- Camon, E., Barrell, D., Lee, V., Dimmer, E., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol* 4, 5-6.**
- Choi, I. G., Kwon, J., and Kim, S. H. (2004). Local feature frequency profile: a method to measure structural similarity in proteins. *Proc Natl Acad Sci U S A* 101, 3797-3802.**

deMare, F., Kurtz, D. M., Jr., and Nordlund, P. (1996). The structure of *Desulfovibrio vulgaris* rubrerythrin reveals a unique combination of rubredoxin-like FeS₄ and ferritin-like diiron domains. *Nat Struct Biol* 3, 539-546.

Han, K. F., and Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci U S A* 93, 5814-5818.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., *et al.* (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32 *Database issue*, D258-261.

Harrison, A., Pearl, F., Mott, R., Thornton, J., and Orengo, C. (2002). Quantifying the similarities within fold space. *J Mol Biol* 323, 909-926.

Haspel, N., Tsai, C. J., Wolfson, H., and Nussinov, R. (2003). Reducing the computational complexity of protein folding via fragment folding and assembly. *Protein Sci* 12, 1177-1187.

Hou, J., Sims, G. E., Zhang, C., and Kim, S. H. (2003). A global representation of the protein fold space. *Proc Natl Acad Sci U S A* 100, 2386-2390.

Hubbard, T. J. (1999). RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. *Proteins Suppl* 3, 15-21.

Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000). Improving the quality of twilight-zone alignments. *Protein Sci* 9, 1487-1496.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-2637.

Kihara, D., and Skolnick, J. (2003). The PDB is a covering set of small protein structures. *J Mol Biol* 334, 793-802.

Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* 323, 297-307.

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17, 282-283.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003a). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19, 1275-1283.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003b). Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, 601-612.

Misra, S., and Hurley, J. H. (1999). Crystal structure of a phosphatidylinositol 3-phosphate-specific membrane-targeting motif, the FYVE domain of Vps27p. *Cell* 97, 657-666.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536-540.

Noguchi, T., and Akiyama, Y. (2003). PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res* 31, 492-493.

- Omichinski, J. G., Pedone, P. V., Felsenfeld, G., Gronenborn, A. M., and Clore, G. M. (1997). The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. *Nat Struct Biol* *4*, 122-132.
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* *5*, 1093-1108.
- Pedersen, J. T., and Moulton, J. (1995). Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms. *Proteins* *23*, 454-460.
- Rossmann, M. G., and Argos, P. (1981). Protein folding. *Annu Rev Biochem* *50*, 497-532.
- Rychlewski, L., and Godzik, A. (1997). Secondary structure prediction using segment similarity. *Protein Eng* *10*, 1143-1153.
- Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* *9*, 232-241.
- Shakhnovich, B. E., Dokholyan, N. V., DeLisi, C., and Shakhnovich, E. I. (2003a). Functional fingerprints of folds: evidence for correlated structure-function evolution. *J Mol Biol* *326*, 1-9.
- Shakhnovich, B. E., Harvey, J. M., Comeau, S., Lorenz, D., DeLisi, C., and Shakhnovich, E. (2003b). ELISA: structure-function inferences based on statistically significant and evolutionarily inspired observations. *BMC Bioinformatics* *4*, 34.
- Sipl, M. J. (1991). Superposition of three-dimensional objects: A fast and numerically stable algorithm for the calculation of the matrix optimal rotation. *Computers and Chemistry* *15*, 73-78.
- Tsai, C. J., Maizel, J. V., Jr., and Nussinov, R. (2000). Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc Natl Acad Sci U S A* *97*, 12038-12043.
- Unger, R., Harel, D., Wherland, S., and Sussman, J. L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* *5*, 355-373.
- Velyvis, A., Yang, Y., Wu, C., and Qin, J. (2001). Solution structure of the focal adhesion adaptor PINCH LIM1 domain and characterization of its interaction with the integrin-linked kinase ankyrin repeat domain. *J Biol Chem* *276*, 4932-4939.
- Ye, Y., Jaroszewski, L., Li, W., and Godzik, A. (2003). A segment alignment approach to protein comparison. *Bioinformatics* *19*, 742-749.

Table 1

Details of the predominantly dimetal-chelating loop proteins forming a high connectivity subgraph.

Repr. Protein	SCCS	Fold	SCOP Description
rubrerythrin	g.41	Rubredoxin-like	metal(zinc or iron)-bound fold; sequence contains two CX(n)C motif
PI ₃ P Binding Domain	g.50	FYVE/PHD zinc finger	dimetal(zinc)-bound alpha+beta fold

DNA binding domain of the GAGA factor	g.37	C2H2 and C2HC zinc fingers	alpha+beta metal(zinc)-bound fold: beta-hairpin + alpha-helix
PINCH domain	g.39	Glucocorticoid receptor-like (DNA-binding domain)	alpha+beta metal(zinc)-bound fold
--	b.88	Mss4-like	Complex fold made of several coiled beta-sheets
--	g.49	Cysteine rich domain	Dimetal (zinc) bound alpha+beta fold
--	g.52	Inhibitor of apoptosis repeat	Caspase inhibitor

Table2

Percentage of secondary structure elements in the Fragnostoc data sets, and in PDB-SELECT25

SSE/ Fragment length	H	E	T
5	22	37	41
10	41	27	32
15	51	21	28
20	51	20	29
PDB- SELECT25	37	22	41

Figures

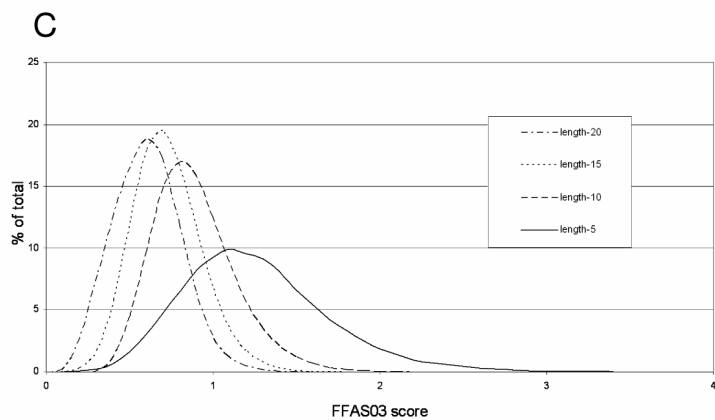
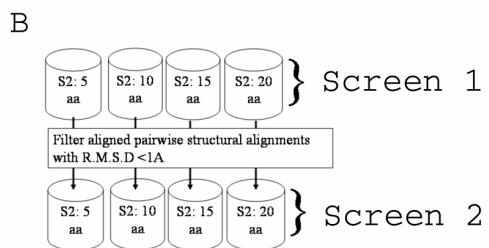
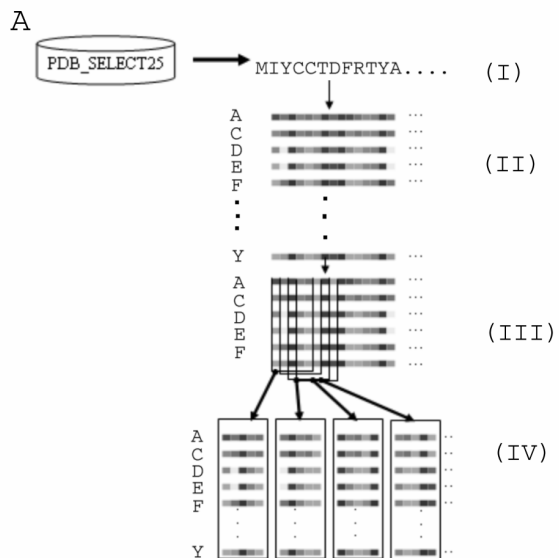


Figure 2

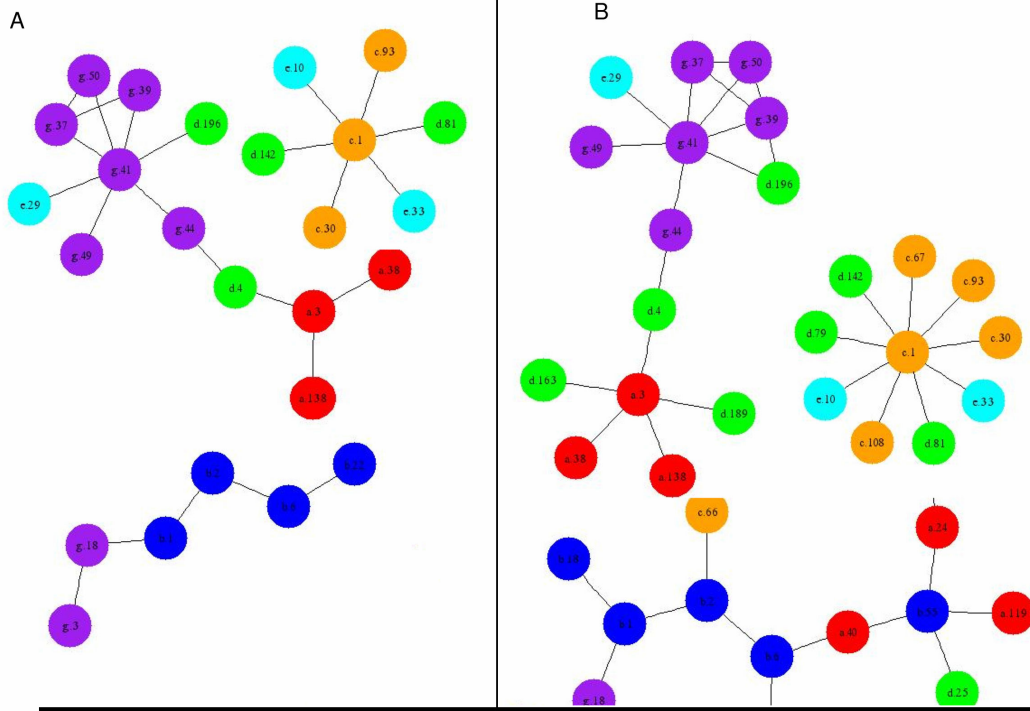


Figure 3

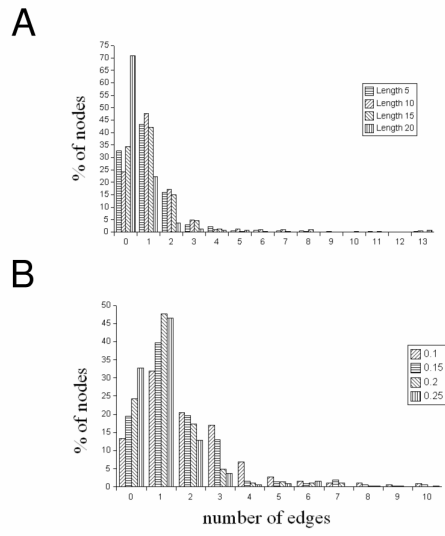
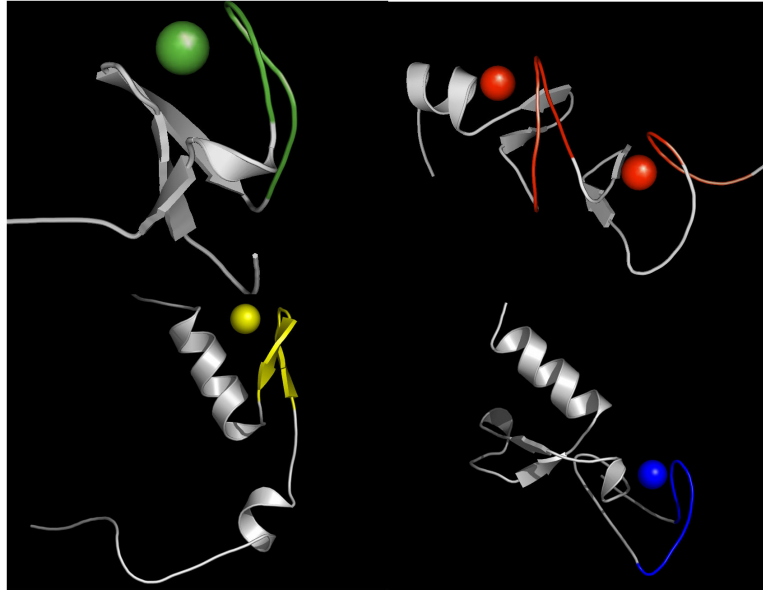
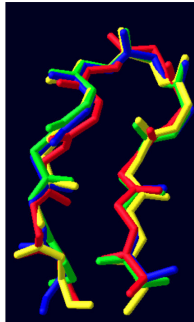


Figure 4

A



B



C

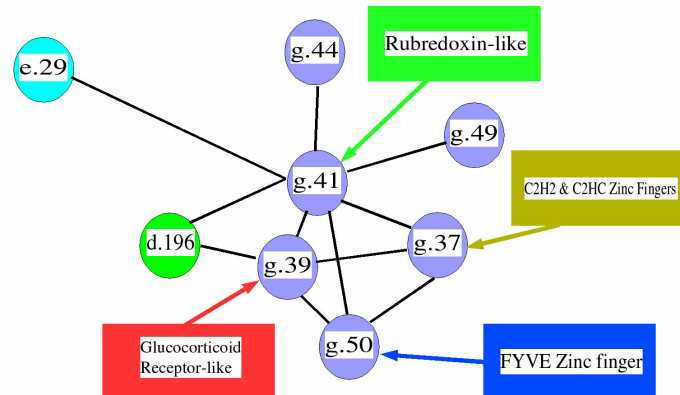


Figure 5

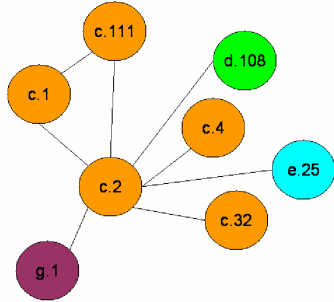


Figure 6

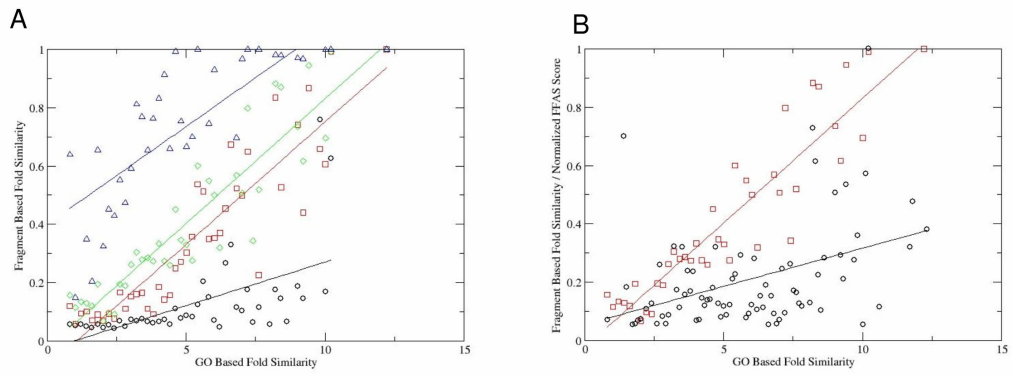


Figure 7

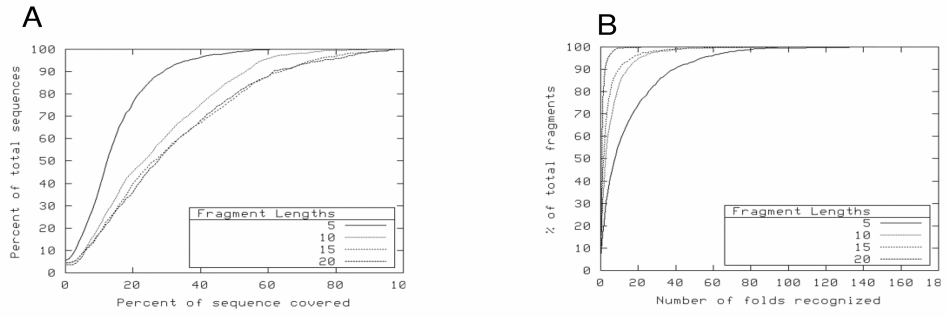


Figure 8

