BIOINFORMATICS

Vol. 1 no. 1 2006
Pages 1–1

Using an alignment of fragment strings for comparing protein structures

Iddo Friedberg[1§], Tim Harder[1§], Rachel Kolodny[2,3], Einat Sitbon[4], Zhanwen Li[1] and Adam Godzik[1]*

[1]Program in Bioinformatics and Systems Biology, Burnham Institute for Medical Research, La Jolla, CA USA; [2] Department of Biochemistry and Molecular Biophysics Columbia University, New York, NY USA and [3]Howard Hughes Medical Institute, USA; [4] Department of Molecular Genetics, The Weizmann Institute, Rehovot Israel; [*]Corresponding author
§Joint first authors

ABSTRACT
Motivation: Most methods that are used to compare protein structures use 3D structural information to do so. At the same time, it has been shown that a 1D string representation of local protein structure retains a degree of structural information. This type of representation can be a powerful tool for protein structure comparison and classification, given the arsenal of sequence comparison tools developed by computational biology. However, in order to do so, there is a need to first understand how much information is contained in various possible 1D representations of protein structure.
Results: Here we describe the use of a particular structure fragment library, denoted here as KL-strings, for the 1D representation of protein structure. Using KL-strings, we develop an infrastructure for comparing protein structures with a 1D representation. This study focuses on the added value gained from such a description. We show the new local structure language adds resolution to the traditional three state (helix, strand and coil) secondary structure description, and provides a high degree of accuracy in recognizing structural similarities when used with a pairwise alignment benchmark. The results of this study have immediate applications towards fast structure recognition, and for fold prediction and classification.
Supplementary material: http://iddo-friedberg.org/ECCB06-supplement

1 INTRODUCTION

The computational representation of a protein's 3D structure is a challenging problem because of varying and often conflicting considerations. On the one hand it seems that as far as information is concerned, more is better, hence the drive to atomic level description, temperature factors and crystallization / NMR information. On the other hand we often ask what is the minimal information we need to achieve a specific task. For practical applications, the representation used for a protein structure is goal driven: it is clear that the representation needed for quick over-the-web wire frame backbone display is not the same required for a detailed analysis of a protein-ligand interaction that may include detailed simulations of chemical processes. With the recent explosion of solved protein structures, there is a growing need for a simpler representation of protein structure. This representation should accommodate the high throughput computational functions required by the growing size of protein structure databases, but without undue sacrifice of accuracy. Large structure database scanning is very expensive (Friedberg, et al., 2004; Holm and Sander, 1994) and fast pre-filtering for negatives can reduce search time considerably. Sequence based comparison methods are generally faster than structure based methods. However structure based methods are much more sensitive, as many different and unrelated sequences may adopt the same fold (Rost, 1997). Incorporating more information into a sequence based representation of protein structure will help increase database search sensitivity while maintaining adequate search time.

One popular simplification is encoding the 3D structure using a 1D alphabet, in which each letter represents a backbone fragment. Many studies have been conducted regarding the use of fragment representation of protein structure. The study of protein fragments has become a favorite tool for investigating sequence-structure relationships (Han and Baker, 1995; Rooman, et al., 1990; Unger, et al., 1989) and lead to new approaches for structure prediction (Haspel, et al., 2003) the study of protein folding (Haspel, et al., 2002; Panchenko, et al., 1997; Panchenko, et al., 1996) and protein sequence alignments (Ye, et al., 2003).

At the same time, few studies have directly examined the information contained in a local structure representation using fragments. In this study, we aim to answer two questions. First, can a 1D fragment based representation of protein structure be used for alignment based similarity scoring in a manner analogous to that used in amino-acid sequence based alignments, and if so, how much information is gained by such a representation? The importance of this question lies in understanding our ability to create a fast filtering tool to be used in high throughput applications on structural databases. Second, given a fragment based representation of a protein structure, what can we learn from the pattern of substitutions between fragments? To illustrate the importance of the second question, we should remember that many studies have been performed on amino acid substitutions aiming to capture the relationship between biophysical traits of amino acids and their substitution patterns among homologous sequences (Tomii and Kanehisa, 1996). Here we ask the same question regarding a different building block representation of proteins: that of short backbone fragments.

KL fragments are a series of structure fragment libraries that can be used to represent protein structure similarities with varying degrees of accuracy depending on fragment library size and the fragment size. A KL fragment library can contain between 20 and 200 fragments, of 4-7 residues in length. In previous studies, KL fragments were shown to accurately model protein structures, as well as provide a good decoy library for protein structure modeling (Kolodny, et al., 2002; Kolodny and Levitt, 2003). KL-strings were chosen because, unlike many other fragment libraries, they are generated entirely from structure, disregarding the amino acid sequence, or standard classifications of local structure. Thus, there is no overt sequence or local structure information embedded in the KL fragment generation process.

To answer the first question of this study, we examined the efficacy of KL fragment encoding of protein structure for detecting similarities in a data set using simple dynamic programming-based sequence alignment. We call this encoding KL-strings. First, we created a KL-string representation of 2749 proteins in a pairwise alignment benchmark first described in (Ye and Godzik, 2003). This benchmark consists of some 15,000 pairwise alignments, and contains positive and negative cases of similarities. We then performed a pairwise alignment of the benchmark's protein pairs. We show the performance of KL-string representation and compare it to that of a structure based methods on the one hand, and amino acid sequence based methods on the other. We also describe the construction and traits of two substitution matrices used in this study for scoring KL-string alignments. To answer the second question of this study, we analyzed the substitution matrices themselves. Specifically, we performed an eigenvalue analysis. We discovered that the principal component explaining fragment substitution is secondary structure conservation.

We conclude that a fragment based representation of protein structures can serve well as a first estimate for detection of structurally similar proteins.

2 METHODS

2.1 KL fragments and KL-strings

KL fragments and their performance in approximating protein structures were described extensively in previous studies (Kolodny and Levitt, 2003). The KL fragment libraries were generated from protein structures by clustering fragments of known length in Eucledian space using a variation on *k*-means clustering called *k*-means simulated annealing. Given a library of fragments, a protein chain can be represented with a known approximation by concatenating fragments from the library. Fragments are chosen from a library of representative fragments and those are fit to the structure in a greedy build-up method. This provides a one-dimensional representation of native protein three-dimensional structure whose quality, using RMSD between the model and the template, is known. Each fragment in the library is represented by a character so the protein Cα backbone can be represented as a character string. This representation, which we term a KL-string, was used throughout this study. We used a library of 20 fragments, with an amino acid length of 5: KL-20-5. This library was shown to model proteins with an average global RMSD of 1.85Å. This choice was deliberate: the KL-20-5 library was shown to provide good models, and it would be interesting to compare the utility of a representation using a 20 letter structural alphabet to that of the 20 letter amino acid alphabet. The full set is available with the supplementary material online.

2.2 Data sets

Unless stated otherwise, we used the following data sets in this study:

ASTRAL-40: a subset of the SCOP database, clustered at 40% sequence identity. SCOP is a database of protein structures which provide a hierarchical description of structural relationships, on four hierarchical levels: class (same secondary structure composition), fold (similar overall structure) superfamily (some measure of functional relatedness and possible homology) and family (a clear homologous relationship). The clustering at 40% serves to remove bias.

FSB: the FATCAT-SCOP benchmark, as described in (Ye, et al., 2003). This benchmark was constructed on the basis of the SCOP version 1.61 40% protein set. The benchmark has 6233 pairs of similar proteins and 8769 pairs of dissimilar proteins. The collection of similar protein pairs includes 830 pairs of family-level similarities, 3146 pairs on the superfamily level, and 2257 on the fold level. The collection of dissimilar proteins includes pairs of proteins from different folds. The FSB set is suitable for testing the sensitivity and specificity of pairwise alignment algorithms which aim to discover distant structural homologs.

PDB-SELECT25: PDB entries clustered at no more than 25% sequence similarity. This data set is both non redundant, and contains a minimal number of homologs.

2.3 Substitution matrices

Substitution matrices lie at the core of any sequence alignment algorithm, providing the score associated with substituting one letter for another in an alignment. When comparing amino acid sequences, each letter represents an amino acid. Therefore, the value in a matrix entry is a function of the probability of an amino acid substitution, normalized by the background probability. Each entry value is given as:

$$(1) \quad M_{ij} = \log \frac{P_{ij}}{P_i P_j}$$

Where $P_{i,j}$ is the substitution rate of amino acid i by amino acid j , and P_i and P_j are the frequencies of those amino acids in the data set. Substitution data is usually gathered from carefully curated alignments of proteins with known relationships. For example, the BLOCKS database of gapless multiple sequence alignments was used to generate the BLOSUM series of substitution matrices (Henikoff and Henikoff, 1992).

For the KL-strings, we generated substitution matrices using a similar rationale. We used existing sets of aligned amino acid sequences as a basis. The sequences we used were all of solved protein structures, and we took the alignments from either the HOMSTRAD (Mizuguchi, et al., 1998) multiple structure alignment database, or from the BLOCKS (Henikoff, et al., 1995) multiple sequence alignment database. Thus we have multiple amino acid sequence alignments of proteins whose structures are known. We then used the translation of these proteins to KL-strings and looked at the alignments of the KL fragments, superimposed on the amino-acid sequences. In other words, we had the KL-string representation of the protein sequence using the underlying amino-acid based alignment as a guide. In this manner, we generated another set of multiple alignments: this time, of strings representing KL fragments, rather than amino acids. Each entry in the matrix contains the value as in equation (1), but with i and j representing KL fragments, rather than amino acids.

Two matrices were generated in this fashion: matrix M_H was generated using alignments from the HOMSTRAD database. HOMSTRAD is a database of structure-based alignments for homologous protein families. Protein structures clustered into homologous families (i.e., common ancestry), and the sequences of representative members of each family are aligned on the basis of their 3D structures. In this study, we used data from the top 10 most populated HOMSTRAD families, clustered at no more than 80% identity. 190 sequences were used with 500-2,000 replacements counted per matrix entry. Matrix M_B was generated using the BLOCKS database. BLOCKS are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. We used those BLOCKS which were rich in proteins with solved structures, and removed entries corresponding to proteins with unknown structures, this came to a total of 9254 sequences in 459 alignments, with no less than 5,000 replacements per matrix entry.

2.4 Substitution matrix eigenvalue decomposition

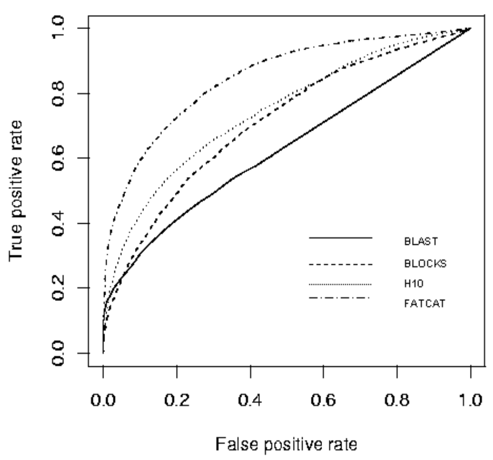

Figure 1: ROC curves describing the performance of the alignment of KL fragments when compared to structure based FATCAT alignments and sequence. Legend: BLAST: results from amino acid sequence pairwise alignments with BLAST; BLOCKS: KL-string alignments using the BLOCKS-based substitution matrix; H10: KL-string alignments using the HOMSTRAD-10 based matrix; FATCAT: structure alignments using the FATCAT structural alignment.

To understand the rules of mutations in the KL-string representation, we performed an eigenvalue decomposition of the substitution matrices. Analogous approaches were used to study sequence based mutation matrices (Kinjo and Nishikawa, 2004). Eigenvalue decomposition of a square matrix yields a series of linearly independent components (eigenvectors) and their weights (eigenvalues). A 20x20 substitution matrix can therefore be decomposed as:

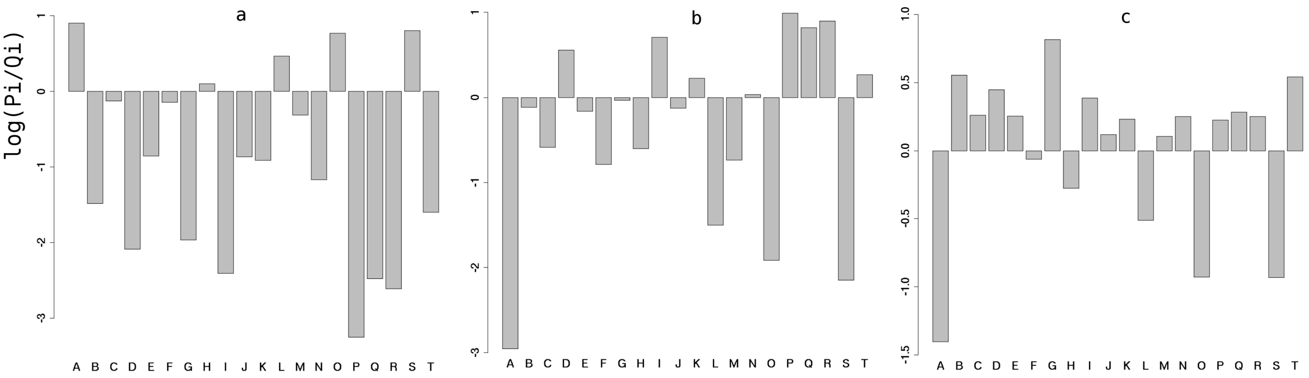

Figure 2: Distribution of KL fragments in secondary structure elements. (a) helices, (b) strands, (c) coils. X-axis letters denote fragments. The Y-axis is the log(Pi/Qi) where Pi is the observed distribution of fragment *i* in a given secondary structure element, and Qi is the frequency of that fragment in the data-set. Note that there is hardly any overlap in the over- or under – representation of fragments in SSE

(2) $$M_{i,j} = \sum_{\alpha=1}^{20} \lambda_\alpha V_{i,\alpha} V_{j,\alpha}$$

where λ_α is the α^{th} eigenvalue and $V_{i\alpha}$ is the i^{th} component of the α^{th} eigenvector. The contribution of the eigenvectors to the matrix is weighted by their associated eigenvalues, which are typically sorted by size. Therefore, it is the first eigenvector that contributes the most to the matrix. In practice the eigenvalue analysis provides us, as shall be seen, with an explanation to the relationships of the fragments on the scoring matrix.

2.5 Pairwise alignments

Pairwise alignments of KL-strings were performed using the Smith Waterman local alignment algorithm (Smith and Waterman, 1981). For each pair of proteins in the FSB set, we took its KL-string representation, and aligned them, scoring the alignment with the substitution matrices which were generates as described in **2.3** above. We normalized the substitution matrices to 1/5th bit each, and provided -10 and -2 penalty values for gap creation and extension penalties which seemed to provide the best results in a series of trials (results not shown). Those were: (1) pairwise amino acid sequence alignment using the BLAST with the BLOSUM62 substitution matrix. (2) Protein structure alignment using the FATCAT structure alignment program. FATCAT compares protein structures using a flexible alignment algorithm, and has been shown to be competitive with other structure alignment programs. In addition to the KL-string alignments, we performed pairwise alignments on the same test set using sequence and structure alignment methods, comparing the performance of these three approaches.

2.6 Significance of KL-string pairwise alignments

We used the negative pairs set from the FSB benchmark to generate a distribution of scores from unrelated proteins, and the pairwise alignments of related proteins to generate a distribution of scores for alignments of related proteins. Thus true negative and true positive score distributions were generated. The quality of an alignment method is judged by its ability to separate the two distributions. The distribution separation was evaluated using ROC curves, as described in the Results section.

Matrix	Helix	Strand	Coil (L+T+B)	Loop	Turn	Bend
M_H	0.89	-0.97	-0.57	-0.82	-0.2	-0.45
M_B	0.88	-0.92	-0.32	-0.67	-0.64	-0.21

Table 1: Correlation of first eigenvector in the substitution matrices, and secondary structure elements. Both the M_H and the M_B matrix display a high correlation between the first eigenvector and helices or strands. For M_H, there is also a mild correlation of the first eigenvector with loops. $P< 0.002$

3 RESULTS

3.1 Performance of KL-string alignment

Note: due to space constraints in this manuscript the substitution matrices are available with the rest of the supplementary material at: http://iddo-friedberg.org/ECCB06-supplement/

We compared the performance of KL fragments to structure based and sequence based methods pairwise alignment methods, using the FSB pairs set. Figure 1 shows the ROC curve resulting from these alignments.

As can be seen, KL-string alignments perform better than the sequence alignment method BLAST, but not as well as the structure alignment method FATCAT. The area under the curve found for each ROC curve was: 0.64 for

BLAST, 0.75 for the KL sequence alignment using the M_B matrix, 0.77 for KL sequence alignment using the M_H matrix and 0.84 for FATCAT.

3.2 Fragments are preferentially distributed in secondary structure elements

Kolodny & Levitt (2004) have reported the generation of protein decoys using the 20-5 fragment library with secondary structure constraints. In their study, they have estimated the probability of each fragment's probability of association with a secondary structure element, in order to generate viable decoys. Here we performed a frequency analysis of KL fragments within SSEs in the PDB-SELECT25 data set. Secondary structure elements were found using DSSP (Kabsch and Sander, 1983), and were expressed in a 3 letter alphabet (H, E, and C) where helices are H, strands are E, and all the rest default to coils, C. The relative frequency of each fragment in an SSE was calculated as the frequency of the fragment in a secondary structure element divided by its frequency in the data-set.

The log of the relative frequencies of fragments in secondary structure elements is shown in Figure 2. The log of the relative frequency was used so that fragments distribution bias can be easily viewed a positive and negative values for higher than expected and lower than expected frequencies respectively. As can be seen in Figure 2, there is a clear preference for four different fragments (*A,L,O,S*) in helices, fragments *D, I, Q, P, K* in strands, and the eleven other fragments in coils. It is important to note that there are no fragments which show a positive preference both in helices and in strands and that only two types of fragments exhibit a positive preference in coils and in strands (D and I). Even then, the positive log frequencies for D and I are the lowest positive values in coils. Finally, no fragment shows a positive preference in coils and in helices. We therefore conclude that KL fragments correspond closely to secondary structure elements. This is illustrated in Figure 3, where an $(\alpha/\beta)_8$ barrel protein is shown, color coded according to its component fragments. The red hues are of fragments which appear in a high frequency in helices, the blue hues are for fragments with a high frequency in strands, and greens are for coils. Note that the colors correspond very well to the secondary structure elements, with the exception of one helix. It should be stressed that the fact that a fragment has a preference to be in a given SSE, does not preclude that it will appear in a different SSE, only that the frequency of this occurring is low. If a fragment overlapped more than a single SSE, it was counted twice for this purpose.

3.3 Eigenvalue analysis of substitution matrices

The analysis of the substitution matrices generated for the KL-string alignments reveals a few interesting findings. As expected, the values along the diagonal are all positive and higher than non self-substitution values, showing that self conservation of fragments is a strong evolutionary determinant.

The eigenvalue decomposition of the matrices revealed three non-trivial eigenvalues per matrix. Since we have seen that certain fragments are over and under- expressed within known secondary structure elements, we decided to examine the correlation of the first eigenvector with the relative frequency of the fragments in secondary structure elements. The results are shown in Figure 4 and Table 1. It is clear that the first eigenvector in both M_B and M_H matrices is related to secondary structure element composition. The relationship is very strong in alpha helices and in beta strands. It is weaker when we look at coils (r= -0.57). Decomposing coils into bends, turns and loops it is interesting to note that for M_H a good correlation (r=-0.82) also exists between the first eigenvector and frequency of fragments in loops (Figure 4c). Table 1 summarizes our findings. Both the M_H and the M_B matrix display a high correlation between the first eigenvector and helices or strands.

3.4 Relative entropy analysis

The relative entropy H of a log-odds substitution matrix is defined as:

$$H = \sum_{1 \le i \le j \le n} P_{ij} \log \frac{P_{ij}}{P_i P_j} \qquad (3)$$

Where P_{ij} is the frequency of substitutions between any two elements i,j, and P_i is the frequency of element i in the data set. An element may be either an amino acid, or, in this case, a KL fragment. The relative entropy describes the amount of information available per aligned element pair.

High relative entropy means that for any given element, the substitution alternatives are clear. This usually happens with substitution matrices generated from alignments of very similar sequences: e.g. BLOSUM80, which is generated from BLOCKS of 80% sequence ID. Conversely, in a matrix with low relative entropy, each element has a larger number of other elements it can be favorably substituted with. This usually happens with substitution matrices generated from alignments with low sequence similarity. The relative entropy of a log odds substitution matrix can be thought of as the ability of the matrix to help distinguish true from chance alignments.

The relative entropy of M_H was found to be 0.68. When examining the relative entropy of a sub matrix constructed from fragments *A, L, O, S* which are preferentially dis-

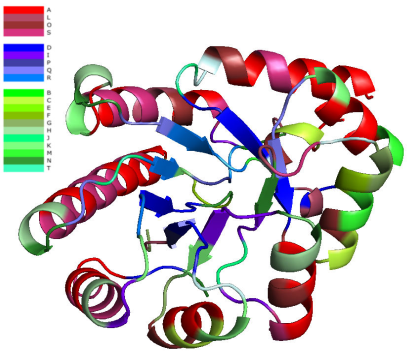

Figure 3: Eukaryotic ornithine decarboxylase (SCOP entry d7odca2), an $(\alpha/\beta)_8$ domain, colored by its modeling KL-fragments. Red hues designate fragments that have a high frequency in helices, blue hues for those fragments that have a high frequency in strands, and greens are for fragments with high frequency in coils. Note the coil transition at the edges of strands and helices. Letters in legend designate fragments.

tributed in helices, we find that the relative entropy is only 0.07. When examining the relative entropy of the sub matrix for strands we find that the relative entropy again is very low: 0.08. When examining the relative entropy of the coils sub matrix, the relative entropy was found to be 0.18 .

4 CONCLUSIONS

In this study we asked the following two questions: First, can a fragment based representation of protein structure be used for alignment based scoring in a manner analogous to alignments of amino-acid sequences, and, second, how much information is gained by such a representation? The results show that it is feasible, and that KL-strings are more sensitive than a sequence alignment method. This is important, as this means that with few modifications, the arsenal of 1D tools available for computational biologists could be used on structure data as well. There is an expected information loss when comparing this method with full structure information alignments. However, as our study is rather preliminary, we expect that the alignment capability will improve by generating substitution matrices more carefully, examining different fragment libraries, and employing sequence profile based tools instead of sequence alone.

Second, what can we learn from the pattern of substitutions between fragments? The eigenvalue analysis has shown that the fragment substitution tendency can be explained primarily by their secondary structure preference. This is in agreement with studies such as (Jeong, et al., 2006; Przytycka, et al., 1999) in which secondary structure has been shown to be a strong attribute for protein structure classification. The subsequent relative entropy analysis has shown that within strands and helices, there are few cases of preferential substitution, but not so in coils. For M_H, there is also a mild correlation of the first eigenvector with loops. We hypothesize that the lower correlation of the M_B first eigenvector with fragment frequency in loops is probably due to the mixture of conserved and non-conserved loops in the BLOCKS database. In contrast, HOMSTRAD alignments include more non conserved regions, and therefore the conserved loops signal moiety may be weaker, and would not affect the general loop signal. This hypothesis is yet to be tested.

The fact that secondary structure information is strongly maintained and fragments associated with known local structure elements can be identified. This finding can be exploited to improve the alignments, as another source of information for creating a 3D-1D database search tools.

The work described here is very preliminary, and can be extended in several directions. One is increasing the sensitivity and specificity of fragment based alignments. How dependent this process is on the size of the library,

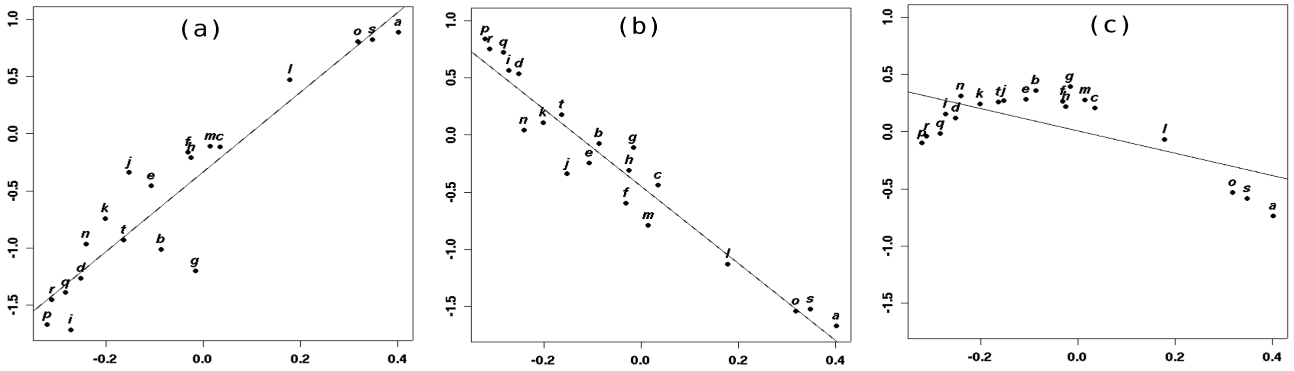

Figure 4: Correlation of the first eigenvector of the HOMSTRAD-10 matrix (X-axis) with the frequency of KL fragments (Y-axis, as in figure 2) in (a) helices (r=0.89) (b) strands (r= -0.97) (c) loops (r= -0.82) Note the parabolic shape in the latter. P <0.002 (by Fisher's Z-transform). Letters next to points denote fragments

and on the fragment length? Previous works (Friedberg and Godzik, 2005; Harrison, et al., 2002) have shown that protein structure space can be partitioned in a non-discrete non hierarchical fashion, unlike the canonical SCOP, CATH and DALI databases. This was done based on alignment of long (Harrison, et al., 2002) or short (Friedberg and Godzik, 2005) fragments associated with specific structures. It would be interesting to re-examine the findings of those studies using the KL library. Better yet, having a good alignment and distance measure on hand, we could cluster the entire protein fold space using the method described in this paper, with KL alignment scores serving as a clustering distance measure: how would the resulting cluster compare with the canonical partitioning of protein fold space? Another direction is increasing the sensitivity of the KL-string database search: using sequence profiles, for example. Seeing protein structure and especially evolution through the perspective of short structural fragments has untapped potential and is very much worth exploring.

ACKNOWLEDGEMENTS

We are grateful to Barry Grant and Łukasz Jaroszewski for their assistance in analyzing the substitution matrices. Thanks to Mickey Kosloff for critical reading and helpful suggestions. We would like to thank the members of the Godzik Lab for stimulating discussions and valuable input. The work of ES is supported by the Israel Science Foundation and by the Leo and Julia Forchheimer Center for Molecular Genetics. This study was funded by NIH grant P01-GM62308.

REFERENCES

Friedberg, I. and Godzik, A. (2005) Connecting the protein structure universe by using sparse recurring fragments, Structure (Camb), 13, 1213-1224.

Friedberg, I., Jaroszewski, L., Ye, Y. and Godzik, A. (2004) The interplay of fold recognition and experimental structure determination in structural genomics, Curr Opin Struct Biol, 14, 307-312.

Han, K.F. and Baker, D. (1995) Recurring local sequence motifs in proteins, J Mol Biol, 251, 176-187.

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.

Haspel, N., Tsai, C.J., Wolfson, H.J. and Nussinov, R. (2002) From the Building Blocks Folding Model to Protein Structure Prediction. In Tsigelny, I. (ed), Protein Structure Prediction: BIoinformatic Approach. International University Line.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, Proc Natl Acad Sci U S A, 89, 10915-10919.

Henikoff, S., Henikoff, J.G., Alford, W.J. and Pietrokovski, S. (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences, Gene, 163, GC17-26.

Holm, L. and Sander, C. (1994) Searching protein structure databases has come of age, Proteins, 19, 165-173.

Jeong, J., Berman, P. and Przytycka, T. (2006) Fold classification based on secondary structure --- how much is gained by including loop topology?, BMC Struct Biol, 6, 3.

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

Kinjo, A.R. and Nishikawa, K. (2004) Eigenvalue analysis of amino acid substitution matrices reveals a sharp transition of the mode of sequence conservation in proteins
10.1093/bioinformatics/bth297, Bioinformatics, 20, 2504-2508.

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.

Kolodny, R. and Levitt, M. (2003) Protein decoy assembly using short fragments under geometric constraints, Biopolymers, 68, 278-285.

Mizuguchi, K., Deane, C.M., Blundell, T.L. and Overington, J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families, Protein Sci, 7, 2469-2471.

Panchenko, A.R., Luthey-Schulten, Z., Cole, R. and Wolynes, P.G. (1997) The foldon universe: a survey of structural similarity and self-recognition of independently folding units, J Mol Biol, 272, 95-105.

Panchenko, A.R., Luthey-Schulten, Z. and Wolynes, P.G. (1996) Foldons, protein structural modules, and exons, Proc Natl Acad Sci U S A, 93, 2008-2013.

Przytycka, T., Aurora, R. and Rose, G.D. (1999) A protein taxonomy based on secondary structure, Nat Struct Biol, 6, 672-682.

Rooman, M.J., Rodriguez, J. and Wodak, S.J. (1990) Automatic definition of recurrent local structure motifs in proteins, J Mol Biol, 213, 327-336.

Rost, B. (1997) Protein structures sustain evolutionary drift, Fold Des, 2, S19-24.

Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences, J Mol Biol, 147, 195-197.

Tomii, K. and Kanehisa, M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, Protein Eng, 9, 27-36.

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.

Ye, Y. and Godzik, A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists, Bioinformatics, 19 Suppl 2, II246-II255.

Ye, Y., Jaroszewski, L., Li, W. and Godzik, A. (2003) A segment alignment approach to protein comparison, Bioinformatics, 19, 742-749.