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]Burnham Institute for Medical Research, La Jolla, CA USA; [2]Columbia University, New-York, NY USA; [3]Howard Hughes Medical Institute USA; [4]The Weizmann Institute of Science, Rehovot Israel

Summary

Here we describe the use of a structure fragment library for the 1D representation of protein structure. 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 pair wise alignment benchmark. The results of this study have immediate applications towards fast structure recognition, and for fold prediction and classification.

Background

The computational representation of a protein's 3D structure is a challenging problem because of varying and often conflicting considerations. A representation 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 [1, 2] 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 [3]. 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. 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? Many studies have been performed on amino-acid substitutions to study the connection between substitution patterns and biophysical traits. To the best of our knowledge, no such study has ben performed with the incrementally larger building blocks, protein fragments.

Results

The Kolodny-Levitt (KL) fragment libraries are a series of backbone fragment libraries used to represent protein structure [4]. In this study we used a library of 20 fragments, each of the length of five amino acids(KL-20-5). Using these fragments, we created a string representation of 2749 proteins in circa 15,000 alignments that contain well-identified positive and negative cases of structure similarities. This is the FSB (FATCAT-SCOP benchmark), as described in [5]. Our goal was to compare the sensitivity of KL-string alignments to that of amino-acid sequence alignments on the one hand, and protein structure based alignments on the other using the FSB set. However, to perform an alignment using KL-strings, we also require at least one substitution matrix. We generated the substitution matrix M_H using alignments from the HOMSTRAD database, and matrix M_B from the BLOCKS database, which contain multiple alignments from structure and sequence considerations, respectively.

Having generated the matrices, we then compared the alignment performance of three different protein representations: amino acid sequence, KL-strings, and structure. The KL-string alignments perform better than amino-acid sequence alignments. This answers our first question positively: a 1D fragment based representation can be used for detecting similarities. To answer the second question, we performed an eigenvalue decomposition of the substitution matrices, which yielded three non-trivial eigenvalues. Since we have observed that certain fragments are over and under represented within secondary structure elements, we examined the correlation of the first eigenvector of M_H and M_B with fragment frequency in secondary structure elements. We show a good correlation between fragment frequency in secondary structure elements, and the first eigenvector of both substitution matrices. The relative entropy of both matrices was found to be high (0.68 bits and 0.71 bits respectively), while that of the sub-matrices containing the fragments associated with alpha helices or beta sheets was found to be low (0.07 and 0.08 bits respectively). This means that although the matrices have the utility to distinguish meaningful from chance alignments overall, the fragments associated with secondary structure elements are relatively interchangeable. The exception is for those fragments that are associated with non-alpha, non-beta elements, for which the relative entropy is higher: 0.18 bits.

Conclusions

We present protein structure representation method that is good for coarse-grained screening for high data volume computations. The fragment based string representation is good for fast alignments using standard sequence based techniques. The fragments themselves are closely associated with secondary structure elements, a strong structure determinant. This study is described in full in [6].

References

1. Holm, L. and C. Sander, *Searching protein structure databases has come of age.* Proteins, 1994. **19**(3): p. 165-73.
2. Friedberg, I., L. Jaroszewski, Y. Ye, and A. Godzik, *The interplay of fold recognition and experimental structure determination in structural genomics.* Curr Opin Struct Biol, 2004. **14**(3): p. 307-12.
3. Rost, B., *Protein structures sustain evolutionary drift.* Fold Des, 1997. **2**(3): p. S19-24.
4. Kolodny, R., P. Koehl, L. Guibas, and M. Levitt, *Small libraries of protein fragments model native protein structures accurately.* J Mol Biol, 2002. **323**(2): p. 297-307.
5. Ye, Y. and A. Godzik, *Database searching by flexible protein structure alignment.* Protein Sci, 2004. **13**(7): p. 1841-50.
6. Friedberg, I., T. Harder, R. Kolodny, E. Sitbon, Z. Li, and A. Godzik, *Using and alignment of fragment strings for comparing protein structures.* Bioinformatics, 2007. **In Press.**