Local Structural Comparison with Global Structural Descriptors

Degui Zhi[1]*, Maxim Shatsky[1,2], Steven E. Brenner[1,2]

[1]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA.
[2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, CA 94720, USA.

Protein sequence and structure are fundamental objects in computational biology. The sequence comparison problem has been widely addressed, resulting in a spectrum of algorithms ranging from the sensitive ones such as profile-HMM to fast ones such as k-mer indexing, arguably culminated in BLAST, where a practical balance of sensitivity and speed is achieved. Current structural comparison methods achieve results generally satisfactory to biologists. However, fast and accurate data base searches, in spirit to BLAST, are not possible due to the nature of the structural comparison methodology.

Similarity of protein structures is typically measured at the residue level via structural alignment, whose goal is to find a 3D transformation that brings into correspondence the largest number of atoms. The quality of a 3D superposition is typically measured by the number of matched C-alpha atoms and their RMSD. The exact solution for the pairwise structural alignment is computationally expensive [1]. Therefore, heuristic approaches have been developed to find a good solution efficiently (for a review see [3]).

An alternative approach to assess protein structure similarity is based on global topological properties, for example, by means of writhe number [2] and Gauss integrals (GIs) [5], or by means of secondary structure footprints [6]. The advantage of this approach is that each structure is represented by a constant number of features. This concise representation tolerates small structural distortions. More importantly, unlike in the structural alignment approach, global topological features of proteins can be trivially compared in constant time, e.g. by the Euclidean distance of GI vectors [5]. This offers potential for fast database search.

The global descriptor approach may suffer from drawbacks. First, it is unable to detect local similarities, i.e., matching of substructures. For example, it cannot detect the similarity between a single domain protein to one of the domains in a multidomain protein. Second, certain, relatively small, structural changes in a protein structure, e.g. loop movement or loop indels, may cause significant changes in a global descriptor.

We propose a new scheme that unifies the above two approaches for structural comparison. Instead of using one global descriptor for the entire protein backbone we consider descriptors for all possible fragments $[i, j]$. The overall similarity between two structures can be defined as the sum of matching scores of a set of sequential, non-overlapping (or not-so-much overlapping) fragment pairs, normalized by their lengths. We designed a dynamic programming algorithm variant to calculate the optimal matching. The similarity between a pair of segments is measured in the same fashion as in [5].

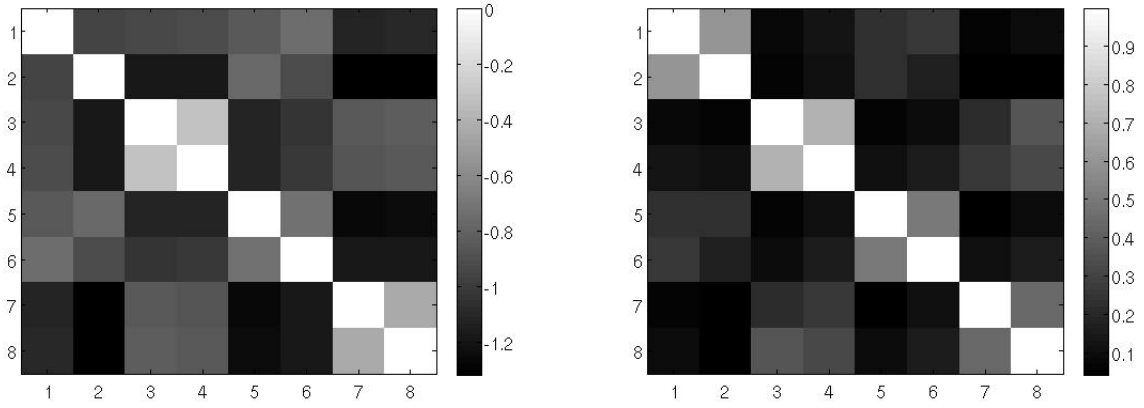
We reduce the running time by exploiting the redundancy in the set of $[i, j]$ descriptors. The running time of the dynamic programming method is $\Theta(n^4)$ if all $\Theta(n^2)$ fragments from each protein are considered. However, we notice that the number of fragments whose descriptors are sufficient

*Corresponding authors email dzhi@compbio.berkeley.edu.

to assess the structural similarity is substantially smaller than the number of all fragments. This observation allows us to significantly reduce the running times.

We tested our algorithm on a set of 8 protein structures, chosen to include both similar and dissimilar ones. As shown in Figure 1, the alignment scores given by our method clearly separate similar ones from dissimilar ones. The global GI measure used in [5], however, fails to identify some very similar structures (e.g., structure 1 and 2).

To summarize, in our approach we overcome the primary drawbacks of the global descriptor methods. Our result showed that the global descriptors for a set of representative fragments capture the essential information needed for structure comparison. The proposed methodology has the potential to be extended into an efficient structure indexing scheme, complementing existing SSE-based structural indexing methods such as the 3-D lookup method [4], which could make structure database queries as efficient as BLAST.



(a) (b)

Figure 1: Test with 8 protein structures from different SCOP categories. Their SCOP ids (1-8) are: a.1.1.1 d1dlwa_; a.1.1.2 d1a6m__; b.1.1.4 d1fcga1; b.1.1.4 d1fcga2; c.23.1.1 d1a04a2; c.23.1.1 d1dz3a_; g.7.1.1 d1chvs_; g.7.1.1 d1coe__. (a) Similarity matrix according to [5](entry $s_{ij} = -d_{ij}^{1/3}$, where d_{ij} is the Euclidian distance of GI vectors for protein i and j). (b) Similarity matrix computed by our method.

References

[1] Christoph Ambuhl, Samarjit Chakraborty, and Bernd Gartner. Computing largest common point sets under approximate congruence. In *Proceedings of the 8th Annual European Symposium on Algorithms*, pages 52–63. Springer-Verlag, 2000.

[2] Gustavo A. Arteca and Orlando Tapia. Characterization of fold diversity among proteins with the same number of amino acid residues. *Journal of Chemical Information and Computer Sciences*, 39(4):642–649, 1999.

[3] I. Eidhammer, I. Jonassen, and WR. Taylor. Structure Comparison and Structure Patterns. *J Comput Biol.*, 7:685–716, 2000.

[4] L. Holm and C. Sander. 3-D lookup: Fast protein structure database searches at 90% reliability. In Christopher J. Rawlings, Dominic A. Clark, Russ B. Altman, Lawrence Hunter, Thomas Lengauer, and Shoshana J. Wodak, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 179–187, Menlo Park, California, 1995. The AAAI press.

[5] Peter Rogen and Boris Fain. Automatic classification of protein structure by using Gauss integrals. *PNAS*, 100(1):119–124, 2003.

[6] Elena Zotenko, Dianne O'Leary, and Teresa Przytycka. Secondary structure spatial conformation footprint: a novel method for fast protein structure comparison and classification. *BMC Structural Biology*, 6(1):12, 2006.