Strongly Connected Components can Predict Protein Structure

Eva Bolten [a] Alexander Schliep [b] Sebastian Schneckener [c]
Dietmar Schomburg [a] Rainer Schrader [b]

[a] *Institut für Biochemie, Universität zu Köln*
[b] *ZAIK/ZPR, Universität zu Köln, Weyertal 80, D-50937 Köln, Germany*
[c] *Science Factory, Köln, Germany*

Key words: Structure prediction, Proteins, Clustering, Graphs

1 Introduction

Finding the three-dimensional structure of proteins is one of the fundamental problems in molecular biology today. The improvements in throughput of classical methods for determining the structure — e.g., using x-ray diffraction analysis or NMR — could not keep up with the ever-increasing speed with which proteins are sequenced. This resulted in a desire for methods for structure prediction solely from sequence data, either ab initio, modeling the molecular folding process, or homology based, using protein sequences with known structures as a template. The main idea is based on the fact that sequence similarity allows to detect homology, i.e., the existence of a common evolutionary predecessor, and thus to infer similar structure and even function virtue of this shared history [10,15]. Note, that the same structure or function does not imply a common ancestor, likewise a common ancestor does not imply a common function, but probably a shared fold.

The relation of sequence similarity as obtained by pair-wise alignments and structural or functional properties has been the goal of a number of publications [3,10,9]. The established and widely accepted rule-of-thumb seems to be that 30% identity over aligned regions [4] is sufficient. More recent studies [13,11] qualified this rule. We will call a sequence similarity above this threshold significant. There are many examples of homologue proteins with

Email address: schliep@zpr.uni-koeln.de (Alexander Schliep).

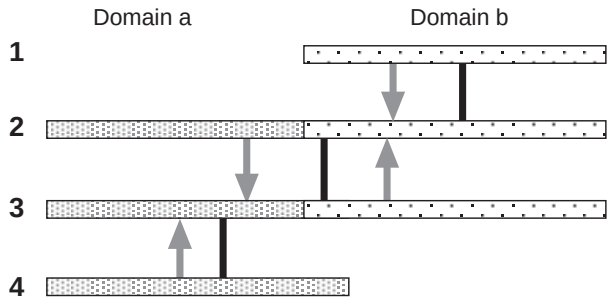

Fig. 1. The problem arising from multi-domain proteins is illustrated. In the un-directed case the solid black edges provide a path from protein #1 to protein #4. Directed edges are displayed in grey.

a sequence similarity below any reasonable significance threshold. Detecting those *distant homologues*, sheading light into the so-called *twilight zone* of low similarity has been approached in a number of different ways [1,8,10,5,6,12,2]: *transitivity* of homology was a common concept.

That is, if proteins A and B as well as B and C have the same ancestor than also A and C have the same mutual ancestor. Even if A and C have a sequence similarity below the threshold for a pair-wise comparison then the existence of a third sequence B with a large enough similarity to both A and C can be used to infer the homologous relation between A and C. The question remains if transitivity extends to arbitrary numbers of intermediate sequences.

2 Algorithm

To answer this question we designed a method for partitioning the sequence space into maximal transitive sets, where clusters correspond to vertex sets of a threshold graph. We identified protein sequences with nodes and each node was labelled with a sequence identifier and weighted with the length of the corresponding sequence. At first a complete undirected graph G was considered where each edge was weighted with the raw Smith-Waterman [14] alignment score, denoted by $raw(P, Q)$. Note, that an arbitrary similarity measure can be used as input for the clustering.

One concern in clustering protein-sequences are multi-domain proteins which form unwanted "bridges" between distinct clusters in protein space. As Fig. 1 shows, not using a directed relation between proteins, i.e., protein A is a domain of protein B instead of A and B are similar at a certain level, will result in false positives during clustering [8]. A computationally inexpensive method for reliable prediction of domains would be highly desirable to accurately establish such a relation between protein sequences. Unfortunately, no such method exists to our knowledge.

A very simple heuristic for approximating the effect of a domain-prediction method for use in our clustering algorithm was obtained in the following way. Noting that there has to be a difference in length between sequences in such a multi-domain situation we decided to *direct* the edges in the graph. Each undirected edge was replaces by two *directed* edges where the weight of the edge from P to Q, (P, Q), was computed as

$$w(P, Q) = \frac{\text{raw}(P, Q) * 100}{\text{raw}(P, P)}$$

resulting in a similarity score between zero and one-hundred percent scaled by the raw score of an alignment of P with itself. If P and Q are two sequences of distinct length, then the weights of the edges (P, Q) and (Q, P) will differ. The resulting graph is denoted by G_d.

Clustering

The next step in the procedure is to proceed to a threshold graph. That is all edges from G_d with a similarity score of less or equal than some fixed threshold τ were removed resulting in the graph $G_d(\tau)$. All similarity values below this threshold are assumed to be produced by chance and not to be an indicator of true structural homology. Again motivated by the problems with multi-domain proteins and after observing the size and composition of the resulting clusters obtained with a single-link cluster algorithm, we decided to use strongly connected component (SCC), a standard concept in graph theory.

Note, that in Fig. 1 only proteins number two and three are in a SCC and thus using a SCC as a cluster discards a substantial amount of information. Nevertheless, we choose to evaluate the performance of our algorithm on the basis of the SCCs alone to establish the validity of our approach. An extension of the method is in preparation.

The Structural Classification of Proteins (SCOP) data base [7] provides very high quality hand-crafted partitions of protein sequences at different levels. For this paper the relevant level are *family*, i.e, sequences with more than 30% sequence identity and possible functional identity, *super-family*, i.e., sequences likely to have a common ancestor, low sequence identity, but structural and functional similarity, and *fold*, i.e., sequences having structural similarity. We evaluated out clustering procedure on SCOP and a number of additional data sets.

References

[1] R. A. Abagyan and S. Batalov. Do aligned sequences share the same fold? *J Mol Biol*, 273(1):355–68, Oct 17 1997.

[2] L. Arvestad, L. Ivansson, J. Lagergren, and A. Elofsson. in preparation, 1999.

[3] S. E. Brenner, C. Chothia, and T. J. Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A*, 95(11):6073–8, May 26 1998.

[4] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–6, April 1986.

[5] M. Gerstein. Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics*, 14(8):707–14, 1998.

[6] A. Krause and M. Vingron. A set-theoretic approach to database searching and clustering. *Bioinformatics*, 14(5):430–8, June 1998.

[7] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, Apr 7 1995.

[8] J. Park, S. A. Teichmann, T. Hubbard, and C. Chothia. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol*, 273(1):349–54, Oct 17 1997.

[9] W. R. Pearson. Comparison of methods for searching protein sequence databases. *Protein Sci*, 4(6):1145–60, June 1995.

[10] W. R. Pearson. Identifying distantly related protein sequences. *Comput Appl Biosci*, 13(4):325–32, August 1997.

[11] B. Rost. Twilight zone of protein sequence alignments [in process citation]. *Protein Eng*, 12(2):85–94, February 1999.

[12] A. A. Salamov, M. Suwa, C. A. Orengo, and M. B. Swindells. Combining sensitive database searches with multiple intermediates to detect distant homologues [in process citation]. *Protein Eng*, 12(2):95–100, February 1999.

[13] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68, 1991.

[14] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, Mar 25 1981.

[15] G. Yona, N. Linial, N. Tishby, and M. Linial. A map of the protein space–an automatic hierarchical classification of all protein sequences. *Ismb*, 6:212–21, 1998.