Clustering Protein Sequences: Structure Prediction by Transitive Homology

Eva Bolten[1,2] **Alexander Schliep**[2] **Sebastian Schneckener**[3]
Eva.Bolten@uni-koeln.de schliep@zpr.uni-koeln.de schneckener@yahoo.com
D. Schomburg[1] **Rainer Schrader**[2]
D.Schomburg@uni-koeln.de schrader@zpr.uni-koeln.de

[1] Institut für Biochemie, Universität zu Köln, Zülpicher Straße 47, D-50674 Köln, Germany
[2] ZAIK/ZPR, Universität zu Köln, Weyertal 80, D-50931 Köln, Germany
[3] Science Factory, Neusser Str. 368, D-50733 Cologne, Germany

Keywords: structure prediction, protein sequences, clustering

1 Introduction

It is widely believed that for two proteins A and B a sequence identity above some threshold implies structural similarity [7]. It not fully understood whether in the case that sequence similarity between A and B is below this threshold the existence of a third protein with a level of sequence similarity with A and with B which is high enough suffices for inferring structural similarity of A and B. The extend to which transitivity holds in the case of several intermediate sequences has only been partly investigated [1, 2, 4, 5, 6, 7, 8].

2 Method and Results

We examined the protein sequences in the SwissProt database. Their similarity was determined using the Smith-Waterman algorithm. This data was transformed into a directed graph where protein sequences constitute vertices. A directed edge (cf. Figure 1b) was drawn from vertex A to vertex B if the sequences A and B showed similarity above a fixed threshold. One concern in clustering protein-sequences are multi-domain proteins which form unwanted "bridges" between distinct clusters in protein space (cf. Figure 1a). By use of a length dependent scaling of the alignment scores we have a criterion to avoid such edges.

To deal with the resulting large graphs (more than 70,000 vertices and 20,000,000 edges) we have developed a very efficient library. Methods include both a novel graph-based clustering algorithm — based on the concept of strongly connected components — capable of handling multi-domain proteins and cluster comparison algorithms. The parameters of above algorithms used were fine-tuned with extensive computations by using SCOP as a test set.

As a performance measure we use sensitivity and specificity [3, 2] with favorable results. We also address the question whether our method provides a substantial improvement compared to pair-wise sequence comparisons. Out of pairs of truly homologous sequences from the same SCOP super-family with a sequence similarity below a reasonable significance threshold we were able to correctly identify 28% while producing negligible false positive errors.

We will present our algorithmic advances, statistics of the clusterings obtained and also case studies for particular protein families and general methodology relevant for testing our hypothesis.

3 Acknowledgments

We would like to thank Oliver Leven, Sebastian Meller, Frank Nübel, Barthel Steckemetz and Olav Zimmermann for helpful discussion and their support.

References

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

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

[3] Brenner, S.E., Chothia, C. and Hubbard, T.J., Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc. Natl. Acad. Sci. USA*, 95(11):6073–6078, 1998.

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

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

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

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

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

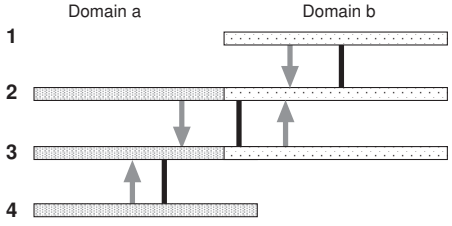

Figure 1a: The problem arising from multi-domain proteins is illustrated. In the undirected case the solid black edges provide a path from protein #1 to protein #4. Directed edges are displayed in grey.

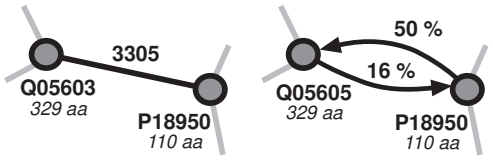

Figure 1b: A single edge (left) weighted with the "raw" Smith-Waterman score is replaced by two *directed* edges in opposite directions (right) weighted by the length of their respective tail vertices. This causes the the distinct similarity scores on the edges.