

Clustering Protein Sequences — Structure Prediction by transitive homology

Eva Bolten^{1,2}, Alexander Schliep², Sebastian Schneckener³,
Dietmar Schomburg¹, Rainer Schrader²

¹ Institut für Biochemie, Universität zu Köln

² ZAIK/ZPR, Universität zu Köln

Weyertal 80, D-50937 Köln, Germany

Tel. +49-221-470-6040 Fax- +49-221-470-5160

schliep@zpr.uni-koeln.de

³ Science Factory, Köln, Germany

It is widely believed that for two proteins A and B a sequence identity above some threshold implies structural similarity. It is 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, that is whether transitivity holds.

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 was drawn from vertex A to vertex B if the sequences A and B showed similarity above a fixed threshold. By use of a length dependent scaling of the alignment scores we have a criterion to avoid clustering errors due to multi-domain proteins.

To deal with the resulting large graphs we have developed a very efficient library. Methods include both a novel graph-based clustering algorithm capable of handling multi-domain proteins and cluster comparison algorithms. The parameters of above algorithms used were fine-tuned by using SCOP as a test set.

We will present our algorithmic advances yielding a 24 percent improvement over pair-wise comparisons, statistics of the clusterings obtained and general methodology relevant for testing our hypothesis.

Keywords: Structure prediction, Proteins, Clustering

Selected References:

L. Arvestad, L. Ivansson, J. Lagergren, and A. Elofsson. In preparation, 1999.

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

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.