Clustering protein sequences with a novel metric transformed from sequence similarity scores and sequence alignments with neural networks

Qicheng Ma, N. R. Nirmala, Gung-wei Chirn, Richard Cai [1]

Keywords: comparative genomics, neural networks, protein sequence clustering

1 Introduction.

Clustering of protein sequences from different organisms has been used to identify orthologous and paralogous protein sequences, to find protein sequences unique to an organism, and to derive the phylogenetic profile for a cluster of protein sequences. These are some of the essential components of a comparative genomics study of protein sequences across several genomes.

Algorithms used to cluster protein sequences can be either domain-based or family-based. All the clustering methods start with an all-against-all pairwise protein sequence similarity searches. The domain-based clustering methods organize the protein sequence universe into domain clusters where domains are the structural units of proteins, e.g., COG [1]. Family-based clustering methods group protein sequences into families, which contain a group of evolutionarily related proteins that share similar domain architecture, e.g., PROTONET [2].

We propose a novel family-based clustering method to address two problems: how to detect whether two aligned sequences have similar domain structures; and how to quantify transitive homologies through intermediate sequences to detect remote homologies at the superfamily level. These two problems are simultaneously solved by a new metric for clustering of protein sequences.

2 Method

From the all against all pairwise sequence similarity searches, we extract four sets of features to represent the homology between a pair of sequences. The first two sets of input features detect the homology of two aligned sequences, the last two sets of input features test whether two aligned sequences have similar domain structures. We use neural networks to map these input features to a new metric, a probability value which scales from 0 to 1. This metric is interpreted as the likelihood that two sequences are of the same homologous superfamily.

The first input feature is the log scale of the pairwise E-value. To model the correlation between two consecutive positions in the alignment, we use the 2-gram encoding [3] of the aligned regions as the second input. Intuitively, if two aligned sequences have similar domain structures, the alignment will divide the two aligned sequences in similar proportions. Thus the third input measures how similarly the two aligned sequences are cut by their aligned regions. Furthermore, the more similar domain structure two aligned sequences have, the more similar neighbour sets two aligned sequences have. The last input feature is to measure the overlap of the two neighbor sets of two aligned sequences.

[1] Life Science Informatics, Functional Genomics Area, Novartis Institute of Biomedical Research Inc, 100 Technology Squere, Cambridge, MA 02139, USA. E-mail: {qicheng.ma,nanguneri.nirmala,gung-wei.chirn, richard.cai}@pharma.novartis.com

After we represent the sequence homology between a pair of sequences by a set of input features, we can train the neural network. Each homologous pair of sequences in the training dataset is labelled as 1 if it belongs to the same Interpro [4] superfamily or the same domain if they are single domain proteins, and 0 otherwise. The neural network we use is fully connected feed-forward back propagation neural network and has one hidden layer with sigmoid activation functions. The network is trained with the scaled conjugate gradient algorithm implemented in MATLAB, and has a specificity of 94.18% and a sensitivity of 91.81%.

To take advantage of the transitive homology between sequence A and C through the third intermediate sequence B, we calculate the product of the metric score for A & B, $P(A,B)$ and the metric score for B & C, $P(B,C)$. If the metric score between A and C is smaller than $P(A,B)P(B,C)$, it is replaced by $P(A,B)P(B,C)$. Then the hierarchical average linkage clustering method is applied to clustering of the protein sequences in the new metric space using the geometric mean of the metric value as the merging rule.

3 Results.

The benchmark data set consists of all Swissprot sequences which satisfied the following criteria. One criterion is that the Interpro annotation for the sequence is consistent, e.g., the same superfamily or domain assignment in at least two member databases which include PROSITE, Pfam, SMART, TIGRFAM, and PRINTS. In addition, the alignment of the sequence with respect to either a hidden Markov model or a profile is at least 30 amino acids long. 41480 Swissprot sequences satisfied these criteria.

We evaluated the performance measure at different threshold values. At the metric score threshold of 0.5, 2073 clusters are generated with specificity 98.7%, sensitivity 99.1%, goodness is 72.6%. There were 59 orphan clusters which can not be mapped to the corresponding Interpro family or domain. Details of the clustering results with regard to specific families together with an analysis of false positives and false negatives will shed light on the strength and weakness of our clustering algorithm, and will be presented in the poster.

4 Conclusion.

This poster describes a novel clustering method of protein sequences into families based on the new metric derived from the prediction by neural networks and further utilizing the metric to model the transitive sequence homologue to detect the remote homologue. Good performance with respect to the Interpro protein sequence database has been achieved on the benchmarking dataset.

References

[1] Tatusov R. L. , Koonin E. V. and Lipman D. J. 1997. A genomic perspective on protein families. *Science* 278(5338), 631-637.

[2] Sasson O., Linial N. and Linial M. 2002. The metric space of proteins-comparative study of clustering algorithms. *Bioinformatics, Suppl 1.* S14-21.

[3] Wang T.J., Ma Q., Shasha D. and Wu C. 2001. New techniques for extracting features from protein sequences. *IBM Systems Journal, Special Issue on Deep Computing for the Life Sciences.* 40(2), 426-441.

[4] Zdobnov E.M. and Apweiler R. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 17(9), 847-848.