High-Throughput Function Assignment for Novel Gene Products Using Annotation Clustering

Alexander Renner **Hilmar Lapp** **András Aszódi**
alexander.renner@pharma.novartis.com

Laboratory of Computational Biology and Chemistry Inflammatory Diseases,
Novartis Forschungsinstitut Vienna, Brunnerstr. 59 - 1235 Austria

Keywords: function assignment, clustering algorithm

1 Introduction

We have designed and implemented a software package for the automatic high-throughput function prediction for genes. This system attempts to assign a biological function to protein sequences by carrying out searches in sequence databanks and by locating functionally relevant motifs in the query sequences. The results produced by the various prediction methods consist of the annotations of matching sequences and/or motifs, which are free-format texts written by humans and therefore may describe the same concept with synonymous words. It was considered desirable to present the results in such a way that the annotations describing the same biological function are grouped together so that the user does not need to read through all of them. To this end we devised an algorithm that enables the hierarchical clustering of free-format documents based on the similarity of their contents. This poster presents an enhanced version of our previously published method [1].

2 Methods

Two documents were considered similar if they contained terms (or keywords) which were related to the same biomedical concept. Concepts were represented by term clusters that grouped related terms together. Relatedness of the terms was measured by how often they occurred together in a set of training documents (the so-called "co-occurrence criterion"). We compiled a term list from the standard SwissProt keyword list and from the functional hierarchy of biomolecules in the InCyte database, containing about 2500 terms in total. For each term, Medline records (max. 1000 per term) were retrieved using a simple keyword query and the union of these records constituted the training document set for building the term clusters. These were obtained using a specialized clustering algorithm that allowed a term to belong to several clusters (expressing ambiguity) and assigned a probability value to each cluster member ("fuzziness").

With the help of the term clusters describing biomedical concepts, two documents could be compared by checking which terms occurred in them, and to which term clusters these terms belonged. If both documents contained terms belonging to the same term cluster (a "term cluster match"), then the documents probably described the same phenomenon and could be clustered together. An appropriate document distance measure was constructed based on term cluster matches, and documents were clustered according to their distances by using Ward's hierarchical clustering algorithm.

3 Results

The term clustering procedure, which had to be carried out only once, produced 890 clusters that indeed described meaningful biomedical concepts. Document clustering was tried on a hand-made test set of Medline records known to contain related and unrelated terms, and on annotations of similarity matches generated by different prediction methods. Biologically meaningful groupings were generated, indicating that the system is capable of detecting concordant and/or conflicting annotations and thus will speed up the interpretation of the function prediction results.

References

[1] Renner, A. and Aszódi, A., High-throughput function assignment for novel gene products using document clustering, *Pacific Symposium on Biocomputing*, 5:54–65, 2000.