Master project in bioinformatics:
Benchmarking the next generation of homology inference tools

at Stockholm Bioinformatics Centre, Science for Life Laboratory

Background: Much biological work today rests on our ability to infer or rule out homology, or common descent, from a pair of replicators, such as genes or their corresponding proteins. This is because common descent, particularly orthologous common descent, where no duplicate genes have been retained, is considered to imply conservation of functional role, which scientists wish to infer bioinformatically in the absence of experimental validation. Recent developments have introduced several methods employing sequence profile-based methods for searching sequence databases for even remote homologs to a query sequence. These include PSI-BLAST, PHMMER and CS-BLAST. Because of their great utility, independent evaluation of their performance, particularly regarding the parameters of precision and recall, is required. In a previous project, we compared different settings for the sequence similarity search tool BLAST with respect to achieved precision and recall, on a benchmark dataset built by classifying pairs of proteins as homologous or non-homologous based on their domain architectures. In this project, this dataset will be updated and improved, then used to evaluate precision and recall – and possibly time and memory requirements as well – for the "next generation" of profile-based tools for sequence homology searches. We hope to be able to publish the outcome of this study in a peer-reviewed scientific journal.

Approach: The benchmark dataset is built from protein pairs taken from the UniProt resource. Pairs are considered homologous if their domain sequences, in order, are homologous, and non-homologous if none of their domains are homologous. To verify that the outcome of this benchmark approach is not strongly dependent on the choice of domain definitions used, three different versions of the benchmark dataset is constructed for the same set of UniProt proteins, based on Pfam, CATH/Gene3D, and SCOP/SUPERFAMILY assignments, respectively. As a secondary benefit, this allows testing of the degree of agreement and disagreement between the databases regarding which protein pairs will be considered homologous versus non-homologous under the approach used here. For each domain schema, agreements are considered on the highest level reliably considered homologous, i.e. Pfam clans, CATH topologies and so forth. For all homologous pairs, and an equal number of non-homologous pairs, each tested method – including CS-BLAST, PHMMER, PSI-BLAST, as well as standard NCBI BLAST as a comparison, possibly also with the recently introduced option for masking repeats in the database as well – is applied and an e-value is generated. By ranking all pairs by e-value, ROC plots and AUC performance measures can be generated for each method on each dataset.

Requirements: The applicant should be familiar with basic statistics (specifically significance testing of enrichment and multiple testing correction), basic bioinformatics scripting (we prefer Perl but accept other languages) and sufficient biology/bioinformatics to understand the question we want to answer. We work in a Linux/UNIX environment. The applicant should be able to work independently (but with supervision available on request) and must be prepared to verify, check and double-check their work as we must be completely sure the results are genuine if we are to be able to publish them. Please apply be sending your CV and the email address of a reference person to:

Erik Sonnhammer, Ph.D.
Professor of Bioinformatics
Director of Stockholm Bioinformatics Centre
Science for Life Laboratory
Box 1031, SE-17121 Solna, Sweden
Tel: +46-(0)8-52481184
Email: Erik.Sonnhammer@sbc.su.se
http://sonnhammer.sbc.su.se/

References: Forslund K and Erik L.L. Sonnhammer. (2009) Benchmarking homology detection procedures with low complexity filters. Bioinformatics, 25:2500-2505