Master project in bioinformatics:
Domain orthology

at Stockholm Bioinformatics Centre, Science for Life Laboratory

Background: Proteins usually consist of some combination of recurring subsequences, corresponding to independently folding domains. Through mechanisms of domain shuffling, the domain composition of proteins can change, leading to situations where parts of a protein has a different evolutionary history than the whole. One interesting property of the evolutionary history of a pair of proteins is whether they are orthologs or paralogs, i.e. whether they are descended from a single protein in the last common ancestral species or not. The phenomenon of domain shuffling could conceivably lead to situations where the orthology status of a pair of proteins, and of their constituent domain pairs, are not the same. This factor has been noted as a potentially large problem for orthology reconstruction. This project aims to determine how common this is.

Goals: Protein domains can easily be assigned to protein sequences, and in most cases is already available in databases. In this case, the alignments of the domain sequences to the sequence profile models used to detect them are crucial, so initial analysis should use sequences represented in the Pfam database. If time permits, results could be validated using Gene3D or SUPERFAMILY assignments instead, but this is secondary. A set of complete genomes should be selected where sequences can be mapped to the UniProt database for which Pfam is built. The InParanoid algorithm should be employed to determine orthology and paralogy status for each pair of proteins defined by each comparison of two genomes. This set of definitions makes up the gene-level orthology set. Next, a distance measure should be developed between domain sequences as aligned to the Pfam Hidden Markov Models, to be used as input to a slightly modified InParanoid run, in order to determine orthology and paralogy relationships between all domain sequences in each comparison of two species, making up the domain-level orthology set. The main goal of the project, then, is to compare the gene-level and domain-level orthology assignments, for orthologous and non-orthologous pairs of proteins that share the same domain architectures, such that corresponding domains in either protein can trivially be identified. How often do corresponding domain pairs not display the same orthology status as their host proteins? Further, is disagreement in this regard more common for some class of domains, such as promiscuous domains, or domains by the termini of the proteins? In what fraction of proteins should we expect overall orthologous relationships to be shared by the constituent domains? Alternately, are there cases where orthology can be inferred from all constituent domains, but not from the sequences as a whole? These cases, if they exist, could form a basis for further improving the sensitivity and accuracy of the InParanoid algorithm.

Requirements: This project involves handling large quantities of structured data files, and applying several complex bioinformatics programs to them. As such, the applicant should be familiar with a scripting language such as Perl or Python, of using such a language to automate multiple operations, as well as have some formality in sequence file formats. Willingness and ability to get into the underlying theory and practice of systems such as Pfam and InParanoid is also necessary. The applicant must be capable of working semi-independently, although tutoring will be available when requested. Experience in Linux/UNIX is important. Please apply be sending your CV and the email address of a reference person to:

Erik Sonnhammer, Ph.D.

Professor of Bioinformatics
SBC, 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/