Master project in bioinformatics:
Protein domain architecture evolution

at Stockholm Bioinformatics Centre, Science for Life Laboratory

Background: Proteins mainly consist of independently folding units called domains. These domains frequently perform distinct functions in the proteins where they occur, and through high-level mutation events such as recombination the domain architectures of proteins can evolve over time. A controversy within protein evolution research is whether two proteins with identical domain architectures must by necessity stem from a single ancestral protein, or whether they have arisen through multiple independent evolutionary events. To accurately answer this question, knowledge of the domain architectures of ancestral proteins is needed. Attempts have been made at reconstructing such architectures through the technique of maximum parsimony along a tree describing the evolution of a family of genes. This effectively amounts to finding the set of ancestral architecture assignments that would require the fewest domain shuffling or gain/loss events in total in order to explain the domain architectures of the present-day sequences. However, parsimony has several weaknesses, notably that it does not explicitly consider time and that it implicitly treats all architecture changes as equally important.

Goals: This project concerns development of a probabilistic model for changes in domain architectures over time, that is, to model the probability of one architecture changing into another depending both on the nature of the changes and on the range of time. Given a basic form for such a function, its parameters could be estimated from a large-scale set of proteins with known evolutionary separation and domain architectures. Once this model exists, it should be applied to gene family trees to determine the set of ancestral domain architectures assignments that have the highest likelihood of resulting in the present-day protein domain architectures. Both parameter estimation and ancestral domain architecture reconstruction are perhaps easiest implemented in a Bayesian Markov Chain Monte Carlo (MCMC) framework. Further expansions of the project include: Comparison between parsimony and maximum likelihood ancestral domain architecture assignments; Assessment of how often multiple instances of the same domain architecture have evolved independently

Requirements: This project will require a student with relatively high bioinformatics skills, or a strong capacity for learning. Primarily, the applicant should be familiar with phylogenetic methods, particularly maximum likelihood and probability. Familiarity with Bayesian phylogenetic tools is very useful. The applicant must be capable of working semi-independently, particularly as he/she will be doing novel method development, although tutoring will be available when requested. Experience in Linux/UNIX and programming/scripting is less crucial, but will come in useful. This avenue of exploration is novel, and should result in publication in a prestigious peer-reviewed journal. 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/