Estimation of P- and E-Values for Profile-HMMs

The three dimensional structures of proteins are usually associated with their properties and functions. Certain patterns in the DNA sequence of proteins are known to be closely correlated with their 3D structure and therefore their function. This enables biologists to identify a protein's function by comparing its DNA sequence with the sequences of other, known proteins. A considerable amount of proteins and protein families has already been made available in online databases like PFAM [1], Swiss-Prot [2] and Astral [3].

Hidden Markov Models (HMMs) are complex stochastic models that can be used to extract information about similarity or evolution of protein sequences from these databases. Especially in the form of Profile HMMs, they allow the classification and identification of hitherto unknown protein sequences. Common programs for training and using HMMs are HMMER [4] and SAM [5]. Both of these programs work highly automated and do not easily allow the user to include his or her expert knowledge into the HMMs they generate.

Basis for our research is the HMModeller, an UCSF Chimera [6] plugin for creating Profile HMMs. Starting in 2005, it has been developed by the University of Applied Science Salzburg (represented by Dr. Stefan Wegenkittl) in collaboration with the University of Salzburg, Department of Molecular Biology (represented by Dr. Peter Lackner). HMModeller has since been extended with the help of former MP grants (Samuel Shepard, Felix Auer, Werner Umshaus, Martin Schnöll). Contrary to HMMER and SAM, HMModeller allows biologists with little or no technical expertise to generate HMMs while taking into account their expert biological knowledge.

This is a follow up to the work done by Samuel Shepard [7]. He described a system consisting of both simulation and curve-fitting to estimate the significance of scores obtained in a database search. The results so far are obtained by several helper applications and need to be integrated seamlessly into HMModeler. This will allow exploiting the concept into further detail because the new framework constructed by the project group the author is currently working in allows obtaining information on the families or models in a more efficient way and calibrating the models before doing a database search.

The primary task of the research project is to integrate the aforementioned algorithms for estimating P- and E-Values into HMModeller. The secondary task is to test the algorithms with data supplied by the University of Salzburg when the integration with HMModeller has proven successful. The details and the results of this research project will be documented in form of a MP report and a MA thesis.

References

[1] http://pfam.sanger.ac.uk/

[2] http://expasy.org/sprot/

[3] http://astral.berkeley.edu/

[4] http://hmmer.janelia.org/

[5] http://compbio.soe.ucsc.edu/sam.html

[6] http://www.cgl.ucsf.edu/chimera/

[7] Research Report for the Austrian Marshall Plan Foundation: Efficient estimation of p-values for HMModeller, Samuel S. Shepard, January 12, 2011