DoBo

Introduction

This project was developed to predict domain boundaries from sequence. It's name, DoBo, means exactly that, domain boundaries. One could ask why such a tool is needed since sophisticated sequence searching using Hidden Markov models and PFAM could do more than find domain boundaries but also identify the domain architecture. The difficulty here is that not all domains are known, characterized by sequence and placed in the PFAM database. This can be seen by taking some of the newer targets from CASP and running them through PFAM. Some of the domains will hit while others do not.

This tool is different than other protein domain prediction packages in that it does not try to predict the domain architecture. Some would say that DoBo does not go far enough by only predicting domain boundaries and that it should go further and predict the domain architecture. The problem with predicting domain architecture is that there are non-continuous domain and long linker regions that make this problem very difficult. Hence, for a "real world" protein or application with several domains, possibly interspersed, the predicted domain architecture could be misleading for some tasks.

By predicting domain boundaries, we allow the user to apply his or her knowledge and use the predictions as they best see fit. Some possible applications are in protein structure prediction or determination. By identifying some very likely domain boundaries, it is possible to properly "cut" a protein sequence into more manageable pieces which correspond to underlying functional units (i.e., the domains).

Construction

DoBo works by classifying data found in a multiple sequence alignment (MSA). As shown in the paper, when genes combine or split they create a signal that shows up in the MSA. These sites are identified as possible domain boundaries and then classified using a SVM that was trained to classify candidate sites as in-boundary (i.e., with in 20 residues of a true boundary) or no-boundary.

Configuration

In the programs directory, configure sspro by running the configure script (setting first the path to the nr database and the installation path). Do not try to use a new version of blast, nr database or sspro unless you absolutely know what you are doing.

You also need to install a version of libstdc++ version 5 for sspro or secondary structure predictions will not be generated. The needed shared object is in the lib

folder but the LD_LIBRARY_PATH environmental variable needs to be set to find it (See run-dobo-stage2.sh to see how).

Update the paths in the scripts directory. Paths need updated in generate-msa.sh script.

In run-stage1.sh and run-state2.sh, set the path for the scripts, models and program directories.

Usage

Two scripts run-dobo-stage2.sh and run-dobo-stage1.sh can be used to run dobo locally.

Contents

In addition to the scripts needed to run DoBo, there is a tarball name dobo-development.tar that contains data and scripts related to the training and publication of dobo and some legacy code in dobo.tar.