Poster H-78

Evaluating ORF Function Predictions using Domain-specific Knowledge

Authors:
Sivakumar Kannan (*Robert Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Mon*)
Gertraud Burger (*Robert Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Mon*)

Short Abstract: Evaluating the machine learning based function predictions of hypothetical proteins is a challenging task. Therefore, we have developed evaluation criteria based on domain-specific knowledge. Using these criteria, we were able to rank predictions and thus identifying the most like working hypotheses for experimental validation.

Long Abstract:

Mitochondrial genomes from diverse eukaryotes carry on average 5 to 20 hypothetical proteins (ORFs) whose function has remained elusive. At the time of writing, GenBank stores more than 2,500 mitochondrial ORFs available from over 243 organisms. Using a machine learning based function prediction method, the decision tree algorithm C4.5 (Quinlan 1993), we have assigned function to 2,549 mitochondrial ORFs.

However, evaluation of these function assignments is not trivial. Performance evaluation methods such as ten-fold cross validation or leave-one-out only evaluate the performance of the classifier on the known but not on the unknown data. Sequence similarity based evaluation (multiple sequence alignment, for example) can also not be applied in this case, because these ORFs do not have recognizable homologs or domains. Therefore, we have developed criteria based on domain-specific knowledge to evaluate the function predictions of mitochondrial ORFs.

The ‘solitary rule’ states that a mitochondrial genome contains only a single gene for each function. Consequently, a function prediction for an ORF is unlikely if a gene of the same function already has been found in the mtDNA of question. The second criterion is the ‘solidarity rule’ that corroborates a prediction if the proteome of a closely related species contains a protein with the same predicted function. Finally, the ‘completeness criterion’ considers whether a genome has been completely sequenced or not, in order to determine whether absence of a gene from the dataset means true absence from the genome.

Using the above criteria, we were able to rank, out of 2,549 predictions, 614 as highly trustable, 759 as likely, and 1,152 as wrong predictions. As long as biochemical techniques do not lend themselves to high-throughput validation of function prediction, in silico predictions are of great value for wet-lab biologists, especially if all available evaluation criteria are taken together to establish most likely working hypotheses.