Identification of ORFs from Organelle Genomes: A Data Mining Approach

Sivakumar Kannan*, Genevieve Boucher and Gertraud Burger.

Canadian Institute for Advanced Research, Program in Evolutionary Biology,
Département de Biochimie, Université de Montréal, Montréal, Québec H3C 3J7, Canada.
*e-mail: siva@bch.umontreal.ca

Genomes of mitochondria and chloroplasts from diverse organisms carry on average 5 to 20 ORFs without assigned functions. In order to understand the biological role of these ORFs, we have developed a comprehensive analysis procedure using data mining methods. To do this, a non-redundant dataset of known mitochondrial proteins from diverse species will be compiled from GOBASE (Shimko, et al. 2001), Pfam (Bateman, et al. 2002) and SWISS-PROT (Bairoch and Apweiler, 2000). In order to represent or describe the knowledge embedded in the sequences, various bioinformatics analyses like physico-chemical properties, motifs, domains, structural features, gene linkage etc., will be performed. A table of the results with known proteins in the rows and their features or attributes in the columns will be made with each different function being a class. Then, a data mining algorithm such as C4.5 (Quinlan, 1993) will be used to look for patterns in the data and to learn the rules. These rules will be evaluated and validated. The yet unannotated ORFs (OGMP) will be represented in the same way like known proteins and their putative functions will be inferred by classifying them in to anyone of the functional classes using the learnt rules.

The advantage of data mining algorithms (C4.5 in this case) over other machine learning algorithms like neural networks is that C4.5 classifiers are expressed as decision trees or sets of if-then rules. Such forms are human readable and hence human expertise could be used to improve the efficiency of the rules. The presented work also serves as a pilot study for automating complete cDNA or genome annotation with a minimum of human intervention. Also, this would enable us to experiment different ways of representing a sequence using various attributes for machine learning algorithms.

References:

Shimko N, Liu L, Lang BF and Burger G (2001) GOBASE: the Organelle Genome Database. *Nucleic Acids Res.*, **29**(1):128-132

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S.R, Griffiths-Jones S, Howe K.L, Marshall M, and Sonnhammer E.L.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**(1):276-280.

Bairoch A., Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000 *Nucleic Acids Res.*, **28**:45-48

Quinlan J. R, (1993) "C4.5: Programs for machine learning," Morgan Kaufmann Publishers.

The Organelle Genome Megasequencing Program (OGMP)
URL: http://megasun.bch.umontreal.ca/ogmpproj.html