Diplomarbeit Andreas Bender, Johann Wolfgang Goethe-University Frankfurt
Research Group Prof. Dr. Gisbert Schneider, April 2002 – September 2002

Abstract

The malaria causing protozoan *Plasmodium falciparum (P. falciparum)* contains mitochondrial genes encoded in its nuclear genome. With the recent sequence completion of its genome, it is desirable to have software tools at hand for prediction of subcellular locations for all proteins. Established tools for the prediction of mitochondrial transit peptides like MitoProtII and TargetP were shown to perform poorly when applied to *P. falciparum* sequences. Therefore, methods specifically designed for this organism had to be developed. Nuclear-encoded mitochondrial protein precursors of *P. falciparum* were analyzed by statistical methods, principal component analysis, self-organizing maps and supervised neural networks and compared to those of other eukaryotes. Two types of descriptions were used, namely relative amino acid frequencies and 19 physicochemical properties. A general distinct amino acid usage pattern has been found in *P. falciparum,* compared to that of other organisms. Glycine, Alanine, Proline and Arginine are underrepresented, whereas Isoleucine, Tyrosine, Asparagine and Lysine are overrepresented, compared to the Swiss-Prot database, Version 36. These patterns were, with variations, also observed in all targeting sequences considered. Using Principal Component Analysis and Self-Organizing Maps, cytosolic N-terminal sequences showed considerable differences to mitochondrial, extracellular and apicoplastical targeting sequences, where the latter were difficult to distinguish from each other. A neural network system (PlasMit) for prediction of mitochondrial transit peptides in *P. falciparum* was developed based on the relative amino acid frequency in the first 24 N-terminal amino acids, yielding a Matthews correlation coefficient of 0.74 (86% correct prediction) in a 20-fold cross-validation study. This system predicted 2449 (24%) mitochondrial genes, based on 10276 predicted open reading frames in the *P. falciparum* genome. A network with the same topology has been trained to give a lower numer of false positive sequences in the training set. This second, more stringent network achieved a Mathews correlation coefficient of 0.51 (84% correct prediction) in a 10-fold cross-validation study. It predicted 903 (8.8%) mitochondrial genes, based on 10276 predicted open reading frames in the *P. falciparum* genome.