Development of a Leishmania-specific phosphorylation sites predictor

Palmeri A[1], Gherardini PF[1], Ausiello G[1], Späth GF[2], Zilberstein D[3], Helmer-Citterich M[1]

Motivation

The high number of Protein Kinases in the genomes of trypanosomatids, together with the dramatical changes that these organisms undergo during their life cycle, underlines the leading role of phosphorylation in the physiology of these parasites. The signal transduction pathways of the trypanosomatids are mainly unknown, but genomic data show that there might be considerable differences in signalling mechanism between host and parasite. Hence the interest in developing a phosphorylation site predictor, specific for these organisms.

Methods

Here we propose a new method, based on Support Vector Machines, to predict phosphorylation sites in trypanosomatids protein sequences. We trained and tested our SVM using the results of several phopshoproteomic experiments performed in Leishmania infantum. We firstly reduced the redundancy at various levels (90%, 60%, 40%) and then splitted the peptides in training (80%) and test sets (20%).The features we included as variables in the SVM were: the aminoacid sequence, the secondary structure and the disorder prediction for the site, and a feature dependent on the composition of a window of +/- 2 residues around the phosphorylation site. We implemented a grid search method for parameters selection. A 10-fold cross-validation was performed to evaluate the performance of each parameters combination.

Results

For each run of cross-validation performed when exploring the parameters, we assessed the variability of our training results by plotting an average ROC curve. The average ROC curves show that the method is stable during the cross validation runs. We tested various combinations of features using sets of peptides with different redundancy. The best results were obtained at the 40% redundancy level with the following combination of features: secondary structure, disorder prediction, residue composition and sequence (in standard orthogonal binary

[1] Centre for Molecular Bioinformatics, Department of Biology, University of Tor Vergata, Roma [2] Laboratoire de Virulence Parasitaire, Department of Parasitology and Mycology, Laboratory of Parasite Virulence, Institut Pasteur, Paris, France [3] Faculty of Biology, Technion-Israel Institute of Technology, Haifa, Israel

encoding). This combination of features achieved an AUC of 0.82 and an MCC of 0.38. The final test was performed using a set of non-redundant peptides (90% identity), totally distinct from the training sets. We used a bootstrap procedure to assess the variability of the results on the test set. The method is able to discriminate between phosphorylated and non phosphorylated sites with an average AUC of 0.737 ± 0.006 and an MCC of 0.227 ± 0.001. We compared the method with the Netphos and NetphosK phosphorylation predictors. The former achieved an average AUC of 0.594 ± 0.006. Since NetphosK gives kinase-specific predictions, it is less appropiate for generic prediction, and obtains an average AUC of 0.497 ± 0.007.

Contact e-mail

antoniopalmeri@tiscali.it