A novel structure-based encoding for machine-learning applied to the prediction of SH3 domain specificity

Ferraro E, Via A, Ausiello G, Helmer-Citterich M

Department of Biology, University of Tor Vergata, Roma

Motivation

Protein recognition modules (PRMs) play a key role in the frame of protein-protein interactions. PRMs are protein domains that focus their binding targets on short protein sequences of about ten residues. SH3 domains are well-studied PRMs that bind proline-rich short sequences characterized by the PxxP consensus. The binding information is typically encoded in the conformation of the domain surface and in the short sequence of the peptide. We extracted this information as significant pair of residues involved in the interaction and defined a numerical encoding scheme in order to build a predictive model for the SH3 domain specificity.

Methods

In the first step of the methodology, pairs of contact residues between an SH3 domain and a ligandpeptideare identified. The contact information is extracted from SH3-peptide complexes of known structure and can also be derived for complexes whose structure is unknown, but can be built with homology modeling techniques From pep-spot and phage-display experiments (Landgraf C. et al., (2004) PLOS Biol. 2, 94-103, Brannetti B & Helmer-Citterich M., Nucleic Acids Res. 2003 Jul 1;31(13):3709-11), we obtained a dataset of interacting and non-interacting domain-peptide pairs.

Binding information is then numerically encoded and used to train a neural network. The encoding procedure is based on the frequency of contact residues characterizing binders and non-binders. The neural model prediction is given in terms of the binding propensity of a domain-peptide pair. We focused our analysis on a group of sixteen yeast SH3 domains and a library of 780 peptides from the yeast proteome, resulting in 8797 domain-peptide pairs. Of such pairs, 649 are interacting and 8148 are non-interacting. The method is applicable to other organisms and even to other families of PRMs, whenever at least one complex of known structure and some experimental data on domainpeptide interactions are available. To verify the generalization capability of the model we scanned the peptide interaction library of the human SH3 domain Abl.

Results

The model was tested with five-fold cross validation on the entire yeast dataset. The results are very encouraging: the global accuracy of the model, measured by the area under the ROC curve (AUC), exceeds 80%.

Contact email: enrico@cbm.bio.uniroma2.it

.