A new neural network approach for the inference of SH3 domains specificity

Ferraro E. (1), Via A. (1), Ausiello G. (1) Helmer-Citterich M. (1)

(1) Department of Biology, University of Tor Vergata, Roma

Motivation

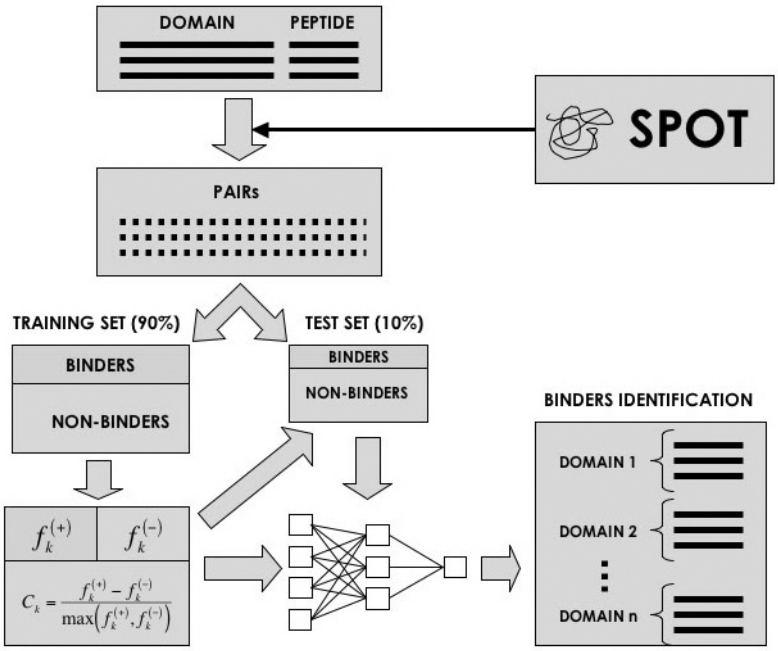
SH3 domains bind polyproline II peptides characterized by the PxxP consensus (P is proline and x in any amino acid). Single domain specificities display a preference for peptides within a range of variability on the common structural theme and different domains may interact with common peptides. We defined a new neural strategy to extract information from interacting partner sequences to improve the identification of SH3 domains specificity.

Methods

The methodology follows the identification of pairs of contact residues between an SH3 domain and a ligand-peptide. Such information, properly encoded, is used to train a neural network whose output expresses the binding propensity of the two partners. Contact information is extracted from SH3-peptide complexes of known structure and can also be applied to complexes whose structure are not known, but can be built with homology modeling techniques. The encoding procedure focuses on the frequency of contact residues characterizing binders and non-binders (see Figure). Such information is deduced 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 restrict our analysis to a group of sixteen yeast SH3 domains with a library of 8797 peptides from the yeast proteome, where 649 are binders and 8148 are non-binders. The method is applicable to other organisms and even to other families of protein domains if at least one complex of known structure and experimental data on domain-peptide interaction are available for network training.

Results

The model was tested by five randomly sampled datasets and an average performance was evaluated with respect to different values of decision threshold. The global accuracy of the model, measured by the area under the ROC curve (AUC) exceeds 90% with respect to a standard logistic regression applied to the same data (88%) and a random network (52%) trained with randomly paired domain-peptide couples.



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