Definition of a neural strategy for the prediction of protein interaction specificity

Enrico Ferraro, Gabriele Ausiello, Simona Panni, Gianni Cesareni and Manuela Helmer-Citterich

Centre for Molecular Bioinformatics, University of Rome Tor Vergata
enrico@cbm.bio.uniroma2.it
http://cbm.bio.uniroma2.it/

Keywords. PepSPOT, Neural networks, Support vector machines,

We are working at the development of a neural network strategy for the prediction of peptide recognition specificity by SH3 domains. As a training set we use the results of a large number of SH3-peptide binding experiments obtained by the SPOT synthesis technique (PepSPOT). As input for the neural network, we consider the sequence of both the domain and the hypothetical ligand peptide, in order to infer for each domain peptide combination the likelihood that they form a complex in a binding reaction. The method will be applied to predict the affinity of any peptide for domains of unknown specificity.
We analyzed data from PepSPOT experiments [1] for nine SH3 domains each tested against several hundred peptides: we decided to construct a proper dataset where each data point includes the domain and peptide sequence, and a figure in arbitrary BLU units that correlates with binding affinity. In order to translate this information in a format that can be easily captured from a neural network, we focused on three main problems: i) the information coding; ii) the dimension of the input space; iii) the correct identification of the two classes (binding and not binding). We decided to use the orthogonal representation [2] of the sequences and, in order to reduce the huge dimensionality, of the domains residues we only considered those positions that make contact with the ligand peptide. The contact positions are identified from the analysis of the SH3-peptide complexes of known structure and extended to other SH3 domains of known sequence by multiple alignment [3]. For the peptide sequences we restricted our representation to the most significant positions, excluding the two consensus prolines from the input. Finally we identified the binding class considering all the peptides that show spot intensity higher than 10000 BLU units. The resulting dataset was strongly unbalanced and this implies the pursuit of different methodological strategies: usual feed-forward neural networks requires the balancing of the training set, while kernel methods (support vector machine) perform classification even on unbalanced sets but with the correct choice of a non-linear kernel [4].
We will verify the performance of the neural strategy with respect to regular expressions, position weight matrices, position specific scoring matrices (PSSMs) and the SPOT procedure [3].

References

[1]Landgraf C., Panni S., Montecchi-Palazzi L., Castagnoli L., Schneider-Mergener J., Volkmer-Engert R., Cesareni G. Protein Interaction Networks by Proteome Peptide Scanning. *PLoS Biol.* 2004 Jan;2(1):E14.

[2]Baldi P., Brunak S. *Bioinformatics. The machine learning approach*. 1998 MIT Press Cambridge, Massachusetts

[3]Brannetti B., Via A., Cestra G., Cesareni G. and Helmer-Citteirch M. SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family, *J Mol Biol.* 2000 Apr 28;298(2):313-28.

[4]Vapnik V. *Statistical Learning Theory* Springer, N.Y., 1998.