Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W451–W454***
doi:10.1093/nar/gkm296

SH3-Hunter: discovery of SH3 domain interaction sites in proteins

Enrico Ferraro*, Daniele Peluso, Allegra Via, Gabriele Ausiello and Manuela Helmer-Citterich

Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica, 00133, Rome, Italy

Received February 1, 2007; Revised April 6, 2007; Accepted April 12, 2007

ABSTRACT

SH3-Hunter (http://cbm.bio.uniroma2.it/SH3-Hunter/) is a web server for the recognition of putative SH3 domain interaction sites on protein sequences. Given an input query consisting of one or more protein sequences, the server identifies peptides containing poly-proline binding motifs and associates them to a list of SH3 domains, in order to compose peptide–domain pairs. The server can accept a list of peptides and allows users to upload an input file in a proper format. An accurate selection of SH3 domains is available and users can also submit their own SH3 domain sequence.

SH3-Hunter evaluates which peptide–domain pair represents a possible interaction pair and produces as output a list of significant interaction sites for each query protein. Each proposed interaction site is associated to a propensity score and sensitivity and precision levels for the prediction. The server prediction capability is based on a neural network model integrating high-throughput pep-spot data with structural information extracted from known SH3-peptide complexes.

INTRODUCTION

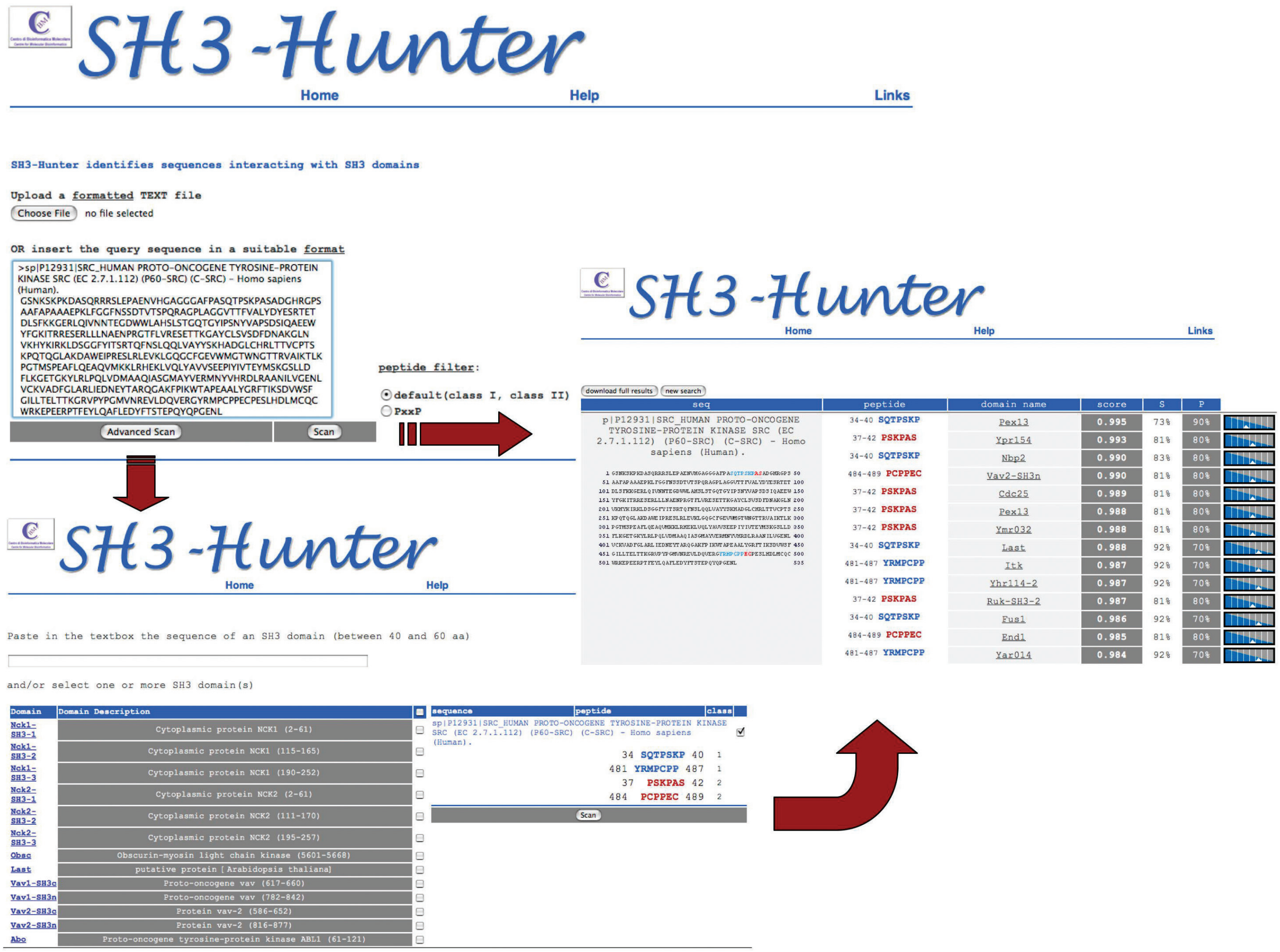
Identifying interacting partners of a given protein is a crucial step towards the discovery of its function. Often proteins communicate by means of protein recognition modules (PRMs), i.e. well-conserved domains characterized by a specific function and interacting with short peptides. The SH3 domain family is one of the most representative PRMs, having a pivotal role in intracellular signal transduction and being widely involved in pathologies such as cancer and AIDS. Several experimental strategies have been proposed to investigate the issue of SH3 domains specificity: from low-throughput analyses focused on specific SH3 domains (1,2) to high-throughput approaches where libraries of peptides are synthesized and their binding ability is confirmed by different *in vitro* experiments (3–5). The high-throughput approaches, however, work within the limits of the current technology for peptide synthesis. The number of short peptides matching the recognition consensus, even in the relatively simple yeast proteome, is in the order of 10^7 (4) while domain or protein family databases contain thousands of SH3 domains. Furthermore, computational methods have been developed (6–8) to help restrict the sequence space of putative SH3 domain binders and to provide experimentalists with powerful tools for the construction of appropriate peptide libraries and for the investigation of domain–peptide interactions.

In such scenario, we present a new web server that permits the inference of SH3 domain interaction specificity on protein sequences. The server is based on a recently published well-performing neural network predictor (8). SH3-Hunter can be used either to predict putative SH3 interactors or to help validating high-throughput experiments, or to support molecular biologists in defining peptide libraries. Furthermore, SH3-Hunter can also be interrogated to investigate the specificity of uncharacterized SH3 domains.

RESULTS

The SH3-Hunter web server analyzes protein sequences to identify putative SH3 domain binders. Users can submit one or more sequences, or even a list of peptides as possible interactors of one or more SH3 domains. To submit large collections of sequences or peptides, users can directly upload an input file. The input sequences can be processed in simple or advanced mode (see Figure 1). In simple mode, a list of inferred interactions is proposed with the whole list of SH3 domains available (see http://cbm.bio.uniroma2.it/SH3-Hunter/help.html). Otherwise, a fine selection of test domains can be prepared with the possibility for the user to submit its own SH3 domain. In both cases, proteins are first scanned by a pattern

*To whom correspondence should be addressed. Tel: +39 06 72594324; Fax: +39 06 72594314; Email: enrico.ferraro@uniroma2.it



SH3-Hunter

SH3-Hunter identifies sequences interacting with SH3 domains

Upload a formatted TEXT file

no file selected

OR insert the query sequence in a suitable format

```
>sp|P12931|SRC_HUMAN PROTO-ONCOGENE TYROSINE-PROTEIN
KINASE SRC (EC 2.7.1.112) (P60-SRC) (C-SRC) - Homo sapiens
(Human).
 GSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPS
 AAFAPAAAEPKLFGGFNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTET
 DLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDSIQAEEW
 YFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLN
 VKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADGLCHRLTTVCPTS
 KPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLK
 PGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLD
 FLKGETGKYLRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENL
 VCKVADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSF
 GILLTELTTKGRVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQC
 WRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL
```

peptide filter:

● default(class I, class II)
○ PxxP

SH3-Hunter

seq	peptide	domain name	score	S	P
p|P12931|SRC_HUMAN PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC (EC 2.7.1.112) (P60-SRC) (C-SRC) - Homo sapiens (Human).	34-40 SQTPSKP	Pex13	0.995	73%	90%
	37-42 PSKPAS	Ypr154	0.993	81%	80%
	34-40 SQTPSKP	Nbp2	0.990	83%	80%
	484-489 PCPPEC	Vav2-SH3n	0.990	81%	80%
	37-42 PSKPAS	Cdc25	0.989	81%	80%
	37-42 PSKPAS	Pex13	0.988	81%	80%
	37-42 PSKPAS	Ymr032	0.988	81%	80%
	34-40 SQTPSKP	Last	0.988	92%	70%
	481-487 YRMPCPP	Itk	0.987	92%	70%
	481-487 YRMPCPP	Yhr114-2	0.987	92%	70%
	37-42 PSKPAS	Ruk-SH3-2	0.987	81%	80%
	34-40 SQTPSKP	Fus1	0.986	92%	70%
	484-489 PCPPEC	End1	0.985	81%	80%
	481-487 YRMPCPP	Yar014	0.984	92%	70%

```
  1 GSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPS  50
 51 AAFAPAAAEPKLFGGFNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTET 100
101 DLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDSIQAEEW 150
151 YFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLN 200
201 VKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADGLCHRLTTVCPTS 250
251 KPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLK 300
301 PGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLD 350
351 FLKGETGKYLRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENL 400
401 VCKVADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSF 450
451 GILLTELTTKGRVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQC 500
501 WRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL                535
```

SH3-Hunter

Paste in the textbox the sequence of an SH3 domain (between 40 and 60 aa)

and/or select one or more SH3 domain(s)

Domain	Domain Description	□
Nck1-SH3-1	Cytoplasmic protein NCK1 (2-61)	□
Nck1-SH3-2	Cytoplasmic protein NCK1 (115-165)	□
Nck1-SH3-3	Cytoplasmic protein NCK1 (190-252)	□
Nck2-SH3-1	Cytoplasmic protein NCK2 (2-61)	□
Nck2-SH3-2	Cytoplasmic protein NCK2 (111-170)	□
Nck2-SH3-3	Cytoplasmic protein NCK2 (195-257)	□
Obsc	Obscurin-myosin light chain kinase (5601-5668)	□
Last	putative protein [Arabidopsis thaliana]	□
Vav1-SH3c	Proto-oncogene vav (617-660)	□
Vav1-SH3n	Proto-oncogene vav (782-842)	□
Vav2-SH3c	Protein vav-2 (586-652)	□
Vav2-SH3n	Protein vav-2 (816-877)	□
Abo	Proto-oncogene tyrosine-protein kinase ABL1 (61-121)	□

sequence	peptide	class	
sp|P12931|SRC_HUMAN PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC (EC 2.7.1.112) (P60-SRC) (C-SRC) - Homo sapiens (Human).			☑
	34 SQTPSKP 40	1	
	481 YRMPCPP 487	1	
	37 PSKPAS 42	2	
	484 PCPPEC 489	2	

Figure 1. The SH3-Hunter web server. The home page in the background presents the input session characterized by the upload file button and, below, by the text area where the user can paste directly the protein sequences. On the right of the text area, the user can select the peptide filter used to identify putative interacting sites and below the two buttons for scan mode and advanced scan mode represents the two available kinds of submission. The first type of submission allows users to analyze the query sequence checking its interaction propensity with the entire list of SH3-domains of the server (see Table H1 in http://cbm.bio.uniroma2.it/SH3-hunter/help.html). The corresponding output page (on the right) represents a list of significant interaction sites on the query protein with different domains. The last three columns in the output page define respectively the significance (score) and the reliability (sensitivity and precision) of the prediction. On the right side of the table, a graphical representation combines sensitivity and precision levels. Above the results table, two buttons allow users respectively to download the results in a text format and to recover the input page for a new search. The advanced scan submission requires an intermediate session (bottom part in the figure) in which users are required to select one or more among the available SH3 domains. An input text line is also provided for the submission of a user SH3 domain. If a list of proteins is provided in input, one or more of them can be selected for prediction. Then, the scan button submits the chosen peptide–domain pairs to the neural network predictor, thus producing the output list.

matching algorithm to detect poly-proline motifs (9,10). The identified motifs are then combined to the complete list of SH3 domain (scan) or to selected domains (advanced scan) to arrange the input information for the neural network predictor (8). The output consists of a list of significant domain–peptide pairs that the predictor recognizes as reliable interacting pairs.

Input

The server requires input in a single protein sequence, a list of proteins or a list of peptides. The submitted input can be pasted on the available textbox area or uploaded as a text file. Four types of formats are allowed for the input sequences: FASTA, bare sequence (sequence without header), interspersed data (as GenBank/GenPept flatfile) and SwissProt flatfile format (as detailed in the server's help). In the quick scan application, this represents the only input information that users have to supply. For advanced scan, after the sequence submission, users are required to submit the sequence of an SH3 domain or to select specific SH3 domains from the available server list and, if a list of proteins or peptides was submitted, specific domain–sequence pairs can be chosen for evaluation. By default, each submitted protein sequence is checked to verify the presence of one or more proline-rich peptides conforming to class I or class II binding motifs ([RKHYFW]xxPxxP and PxxPx[RK], respectively). If consensi are not found, the submitted sequence is considered as non-interacting and a warning message is visualized. However, if the requirement of this filter is considered too stringent, users can relax the filter by choosing the PxxP motif for the peptide selection. If proline-rich peptides are identified, every one of them is combined with an SH3 domain from either the complete server's list or a user-defined sub list. If an SH3 domain is added by the user, its possible interactions with the selected peptides are evaluated. Each resulting peptide–domain pair represents an input for the predictor. Each input is transformed into a set of real numbers (see Methods) that can be classified by the neural network.

Output

Each peptide–domain pair undergoes the predictor evaluation and is reported in output if the score is higher than a given threshold. Therefore, the output consists of a list of peptide–domain pairs, sorted according to the predictor's score, which is a measure of the reliability of the inferred interaction (see Figure 1). For a more correct interpretation of the results, each score is also associated to the sensitivity and precision levels of the neural network prediction. The sensitivity measures the expected true positives rate detected by the neural network with that given score, while the precision measures the reliability of the prediction. The two measures clearly have opposite tendencies and the user can decide whether to collect results with higher sensitivity, involving much more true positives as possible, but with a higher risk of false positives, or select only results with higher precision levels, avoiding false positives but with a higher probability to loose a portion of true positives. A graphical representation of sensitivity/precision levels lies at the right margin of the numerical measures.

Users must be aware of the fact that the absence of any output for their submissions means that no interaction scored above the chosen significance threshold. However, the full list of results can be downloaded as a text file.

METHODS

SH3-Hunter is based on a neural network predictor, which infers the specificity of interaction between a peptide and an SH3 domain (8). The neural model integrates both sequence and structure information of the peptide–domain pair, involving a knowledge-based numerical encoding of the input information. The sequences of each peptide-SH3 pair are processed by selecting only amino acids lying on the interaction surface and involved in an inter-molecular contact. Each peptide–domain pair is represented by a fixed number of contact residue–residue pairs, the former belonging to the peptide, the latter to the domain (8). Contact residues on SH3 domain and peptide can be identified directly on crystallized SH3 domain–peptide complexes or indirectly by homology modeling (8,11), while the numerical encoding of the residue–residue pairs is based on their occurrence in a dataset of interacting and non-interacting peptide–domain pairs (8). Contact information for a list of SH3 domains were previously evaluated and represent a fundamental knowledge for the server prediction (see Table H1 in http://cbm.bio.uniroma2.it/SH3-hunter/help.html). The list will be progressively upgraded in order to extend interaction prediction to a wider number of SH3 domains.

The server application consists of a three-step process aimed at the discovery of SH3 domain–binding sites on protein sequences.

The first step consists of a pattern matching algorithm that scans the submitted proteins in order to check if they contain either the class I [+@]xxPxxP or the class II PxxPx[+] patterns (9,12), where the + identifies positively charged amino acids (His, Arg or Lys), @ corresponds to aromatic amino acids (Phe, Tyr, Trp), x means any amino acid and P is proline. Note that in the class I pattern, the first position is also extended to aromatic residues with respect to the standard motif. Such choice is motivated by pep-spot experimental results (4) on yeast SH3 domains. The result of the first step provides a list of 10-residue long peptides conforming to the SH3 typical binding motifs. The presence of such a filtering procedure is required since the neural network predictor was trained by class I and class II interaction data (4,13). From a methodological point of view, a neural network is able to generalize to some extent its predictive capability (14). Therefore we expect that SH3-Hunter will produce meaningful prediction even for peptides that do not fit precisely with the class I and class II motifs. However, in order to limit the loss of reliability of the server predictions, we allow a different kind of filter based only on the PxxP consensus. Users can select the appropriate filter for their submission. Sequences not conforming to the chosen filter are discarded. It is worth noting that the use of the PxxP filter produces predictions of lower reliability. Besides, the PxxP filter does not avoid the class I and class II distinction: the two types of binding orientations are still considered by selecting class I or class II peptides as showing the PxxP motif respectively at the C terminal or at the N terminal, according to the peptide alignment requirements of the predictor (8).

In the second step, each peptide is combined to the SH3 domains of the server's list, to compose a peptide–domain pair. This corresponds to the simple 'scan' submission. An 'advanced scan' submission is also available, which permits the selection of one ore more SH3 domains. Here the user can submit its own SH3 domain sequence, which can be appended to the selected domains from the server list or analyzed separately (see Figure 1). A previously and accurately evaluated multiple alignment of SH3 domains is used as a profile to align the user domain and infer its contact positions (see earlier discussion and 8). Specifically, the server uses the ClustalW algorithm (15) to provide the alignment and assigns the name Sh3Usr to the user submitted domain. We want to stress that the identification of surface contact positions of the user domain is based only on the domain sequence information and on an automated alignment procedure. For a more reliable prediction, users are encouraged to submit new SH3 domain sequences via email asking for a manual alignment.

Furthermore, if a list of proteins or peptides is submitted, the advanced option allows the selection of one or more list of members. Finally, each peptide–domain pair is transformed in a set of real variables (8) representing the input of the neural network predictor.

The third step applies the neural network described in (8) to the peptide–domain pairs. The neural network is trained by a dataset of experimentally verified interacting and non-interacting peptide–domain pairs (4,13). Input peptide–domain pairs are processed and an output response is given that measures the peptide–domain interaction propensity. Each propensity is then standardized and normalized in order to obtain a score ranging between 0 and 1.

Sensitivity and precision measures

The neural network model is characterized by different levels of sensitivity and precision, corresponding to specific thresholds on its output score. Sensitivity is defined as the rate of true positives recognized by the neural network with respect to the total number of true positives: TP/(TP + FN), where TP and FN represent respectively true positives and false negatives. Similarly, precision is defined as the fraction of true positives recognized by the model with respect to the number of cases that the model classifies as positives: TP/(TP + FP), where FP identifies false positives. TP, FN and FP clearly depend on the value of a decision threshold: if the output of the neural network is higher than or equal to the threshold value, the peptide–domain pair is classified as interacting, otherwise it is classified as non-interacting. We defined a set of thresholds, which can be used to interpret the output of the neural model (i.e. the score assigned to each peptide–domain pair) and the corresponding values of sensitivity and precision (see Table H2 in http://cbm.bio.uniroma2.it/SH3-hunter/help.html).

ACKNOWLEDGEMENTS

This work was supported by Telethon (GGP04273), a PNR 2001-2003 (FIRB art. 8) and a PNR 2003-2007 (FIRB art. 8). Funding to pay the Open Access publication charges for this article was provided by AIRC.

Conflict at interest statement. None declared.

REFERENCES

1. Masumi,A., Aizaki,H., Suzuki,T., DuHadaway,J.B., Prendergast,G.C., Komuro,K. and Fukazawa,H. (2005) Reduction of hepatitis C virus NS5A phosphorylation through its interaction with amphiphysin II. *Biochem. Biophys. Res. Commun.*, **366**, 572–578.
2. Stamenova,S.D., French,M.E., He,Y., Francis,S.A., Kramer,Z.B. and Hicke,L. (2007) Ubiquitins binds to and regulates a subset of SH3 domains. *Mol. Cell*, **25**, 273–284.
3. Kay,B.K., Williamson,M.P. and Sudol,M. (2000) The importance of being proline: The interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J.*, **14**, 231–241.
4. Landgraf,C., Panni,S., Montecchi-Palazzi,L., Castagnoli,L., Schneider-Mergener,J., Volkmer-Engert,R. and Cesareni,G. (2004) Protein interaction networks by proteome peptide scanning. *PLoS Biol.*, **2**, 94–103.
5. You,X., Nguyen,A.W., Jabaiah,A., Sheff,M.A., Thorn,K.S. and Daugherty,P.S. (2006) Intracellular protein interaction mapping with FRET hybrids. *Proc. Natl Acad. Sci. USA*, **103**, 18458–18463.
6. Hou,T., Chen,K., McLaughlin,W.A., Lu,B. and Wang,W. (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. *PLoS Comput. Biol.*, **2**, e1.
7. Lehrach,W.P., Husmeier,D. and Williams,C.K. (2006) A regularized discriminative model for the prediction of protein-protein interactions. *Bioinformatics*, **22**, 532–540.
8. Ferraro,E., Via,A., Ausiello,G. and Helmer-Citterich,M. (2006) A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics*, **22**, 2333–2339.
9. Mayer,B.J. (2001) SH3 domains: Complexity in moderation. *J. Cell Sci.*, **114**, 1253–1263.
10. Musacchio,A. (2002) How SH3 domains recognize proline. *Adv. Protein Chem.*, **61**, 211–268.
11. Brannetti,B., Via,A., Cestra,G., Cesareni,G. and Helmer-citterich,M. (2000) SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. *J. Mol. Biol.*, **298**, 313–328.
12. Cesareni,G., Panni,S., Nardelli,G. and Castagnoli,L. (2001) Can we infer peptide recognition specificity mediated by SH3 domains? *FEBS Letters*, **513**, 38–44.
13. Tong,A.H., Drees,B., Nardelli,G., Bader,G.D., Brannetti,B., Castagnoli,L., Evangelista,M., Ferracuti,S., Nelson,B., Paoluzi,S., Quondam,M., Zucconi,A., Hogue,C.W., Fields,S., Boone,C. and Cesareni,G. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.
14. Bishop,C.M. (1995) *Neural Networks for Pattern Recognition* Oxford University Press.
15. Higgins,D., Thompson,J., Gibson,T., Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.