W164–W168 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl166

Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines

Alessandro Vullo, Oscar Bortolami[1], Gianluca Pollastri* and Silvio C. E. Tosatto[1]

School of Computer Science and Informatics, University College Dublin, Ireland and [1]Department of Biology and CRIBI Biotechnology Centre, University of Padova, Italy

Received February 13, 2006; Revised March 1, 2006; Accepted March 20, 2006

ABSTRACT

Intrinsically disordered proteins have long stretches of their polypeptide chain, which do not adopt a single native structure composed of stable secondary and tertiary structure in the absence of binding partners. The prediction of intrinsically disordered regions in proteins from sequence is increasingly becoming of interest, as the presence of many such regions in the complete genome sequences are discovered and important functional roles are associated with them. We have developed a machine learning approach based on two support vector machines (SVM) to discriminate disordered regions from sequence. The SVM are trained and benchmarked on two sets, representing long and short disordered regions. A preliminary version of Spritz was shown to perform consistently well at the recent biannual CASP-6 experiment [Critical Assessment of Techniques for Protein Structure Prediction (CASP), 2004]. The fully developed Spritz method is freely available as a web server at http://distill.ucd.ie/spritz/ and http://protein.cribi.unipd.it/spritz/.

INTRODUCTION

Over the last couple of years there has been growing experimental and theoretical interest into what has come to be known as intrinsic disorder in proteins (1–11). Some proteins appear to have long stretches of their polypeptide chain which do not adopt a single native structure. These disordered fragments instead appear to exist in a state that is unstructured but different from unfolded proteins. Frequently such disordered fragments interact with other proteins and can become structured under certain conditions, such as ligand binding or phosphorylation. Recent studies have determined the functional classes of signal transduction and transcription to be overrepresented among disordered proteins (12,13). Perhaps linked to this functional bias is a higher percentage of disordered regions in Eukaryotes compared to Prokaryotes and Archea, since the former require a more sophisticated control of communication (12,13). From a structural point of view, disordered regions frequently do not appear on the electron density map in X-ray crystallographic studies and are highly flexible according to NMR data. While disorder is not overly represented in the PDB for the above reasons, long 'loopy' regions have also been observed in some structures (13). Predicting the location of such unstructured regions is therefore important in protein structure prediction, as these disordered fragments cannot and should not be predicted. This has prompted the inclusion of a disorder prediction session in the last Critical Assessment of techniques for protein Structure Prediction (CASP) experiments (14), starting with CASP-5 in 2002.

Perhaps due to the inclusion in CASP, there has been recently increased interest in predicting disorder and associated features from sequence (12,15–25). The most prominent feature of many, in particular long, disordered sequences is low sequence complexity. This has been used for a long time to filter sequence database searches with the SEG filter (26). Long disordered regions show a high net charge and have few hydrophobic residues (1,2). A similar trend has also recently been observed for short disordered sequence patterns (21). This sequence bias has been used to develop machine learning based predictors in analogy to secondary structure prediction. A training set of long and/or short disordered regions is used to train a neural net or support vector machine to discriminate disordered from ordered sequence fragments. The results are quite encouraging, as disorder appears to be easier to predict than secondary structure (27). Long disordered regions in particular are consistently predicted by most methods. The performance on short disordered regions is subject to higher

*To whom correspondence should be addressed. Tel: +353 1 716 2926; fax: +353 1 269 7262; Email: gianluca.pollastri@ucd.ie

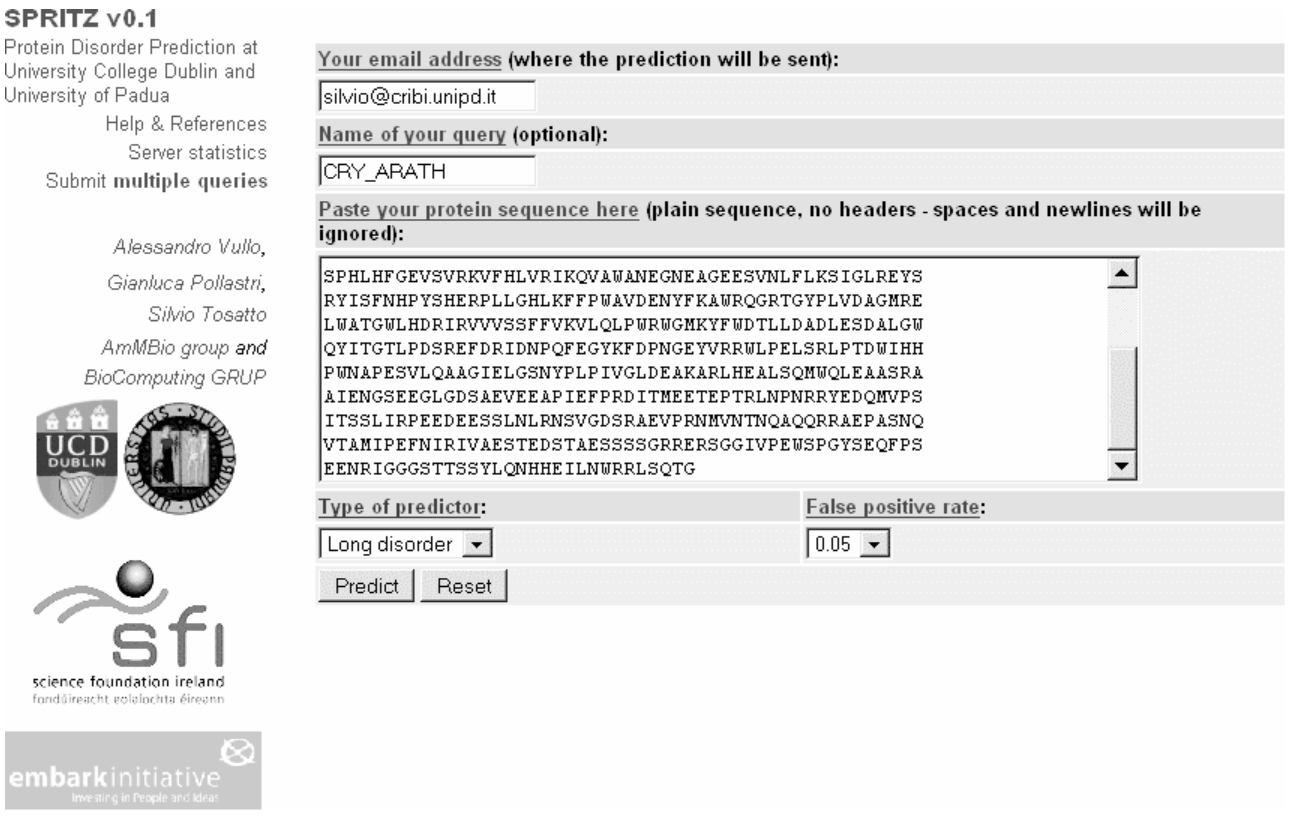

Figure 1. The Spritz web server interface.

fluctuations. Overprediction is still an issue with most methods, i.e. more residues are predicted to be disordered than is really the case. However, at least some of these residues correspond to the more flexible parts of the protein structure, pointing at an intuitive correlation between disorder and flexibility. In analogy to secondary structure prediction (28–30), recently some of the best performing disorder predictors have started to use a combination of predictors. Probably because short and long disordered regions may be the result of different processes, and both may be further classified into different 'flavours' (22), approaches relying on multiple specialized predictors have led to improved performances. In this paper we present our novel server for the prediction of intrinsically disordered region in proteins, based on two separate support vector machines (SVM), one specialized to recognize long disordered regions, one short disordered regions. The server, called Spritz, can be accessed from http://distill.ucd.ie/spritz/ and http://protein.cribi.unipd.it/spritz/.

PROGRAM DESCRIPTION

Overview

The server Spritz is implemented using two specialized binary classifiers, one for short regions of disorder and the other for long disordered fragments. The purpose of doing so is to develop different, disjoint expertize by taking advantage of the different class distributions in the two cases. These classifiers are derived from independent datasets as discussed in the following.

LD dataset

We first assemble a subset of completely disordered sequences, each with over 45 disordered residues, from DISPROT release 1.2 (31) by removing sequences containing errors of annotations and then using the most up to date GI accession numbers. The final set is completed by incorporating an equal sized subset of chains classified as having no disordered fragments and derived from the PDBselect25, March 2002 release. The filtered and balanced set contains 293 sequences corresponding to 34 159 residues, 17 001 (49.7%) of which are classified as belonging to long regions of disorder.

SD dataset

A collection of short disordered sequence fragments is compiled from the November 2004 release of the PDB (32). We filter out proteins that are not solved by X-ray diffraction, are less than 30 AA and have resolution worse than 2.5 Å. In order to obtain non overlapping entries with the LD set, we finally select those chains having no more than 20 disordered amino acids and sharing at most 25% sequence similarity. The final set contains 1017 sequences corresponding to 278 600 residues, 8824 (3.2%) of which are classified as belonging to short regions of disorder. There is strong unbalance towards the ordered class, reflecting the PDB distribution.

Specialized SVM classifiers

The LD and SD datasets are used independently to derive two binary classifiers. These classifiers are both implemented with

probabilistic soft-margin SVM or C-SVMs (33,34) mounted with a Gaussian kernel. Note that differently from (12), we use a non linear kernel which is less biased (i.e. more expressive) than the linear one but also more prone to overfitting. We tackle the overfitting problem by running model selection on independent validation data (35). An additional problem is represented by the strong unbalance of class distribution in the SD set. To mitigate this problem, we train SVMs using asymmetric costs, i.e. a larger penalty for disorder misclassifications. This penalty is equal to the ratio of the number of negative (i.e. ordered) to positive examples in the training set.

Both classifiers use as inputs residue attributes extracted from a window of k residues. For each of the k residues, 20 amino acid frequencies are input into the SVM, computed from multiple alignments obtained from three runs of PSI-Blast (36) against the non-redundant NCBI sequence database redundancy reduced at 90%. For each amino acid in the window, we consider also secondary structure information predicted by the Porter server (37) in the form of three probability values. In our tests secondary structure information leads to gains of roughly 3% in two-state classification accuracy. After initial bootstrapping, all the experiments and the final stage (the server) are implemented using a window of size five (respectively three) for the LD (respectively SD) set predictor.

Table 1. Results of SVM trained on SD and LD dataset

	C	AUC	Q2 (5% FPR)
SD			
5-fold CV	0.44	0.82	93.2
CASP-6	0.44	0.79*	91.5
LD	0.14	0.6	53.9
LD			
5-fold CV	0.59	0.85	72.2
CASP-6	0.35	0.85	91.5
SD	0.21	0.8	92.6
CaspIta			
CASP-6	0.41	0.73	93.2
VSL-1			
CASP-6	0.32	0.88	82.4

*An indicates a probable underestimation of the AUC due to insufficient data in subintervals of [0,1]. The official results for two top scoring CASP-6 groups are shown for comparison.

SERVER DESCRIPTION

The web server has two interfaces, one for single and one for multiple queries (see also Figure 1). The former takes as input a bare protein sequence while the latter accepts input in FASTA format with multiple entries for batch querying. Submissions of up to 32 768 characters are accepted. The user has the option to adjust the classifier output according to the estimated false positive rate (FPR). The fraction of disordered amino acids that are expected to be recovered can be estimated from the FPR and the ROC curve provided in the help page. The user can select either predictor from the interface and may provide an FPR (default 0.05).

The server outputs the following sequences: (i) the prediction of secondary structure in three classes (C = coil; E = strand; H = helix), obtained using Porter (37); (ii) the prediction of protein disorder in two classes (O = ordered; D = disordered). The results in plain ASCII text format are sent by email.

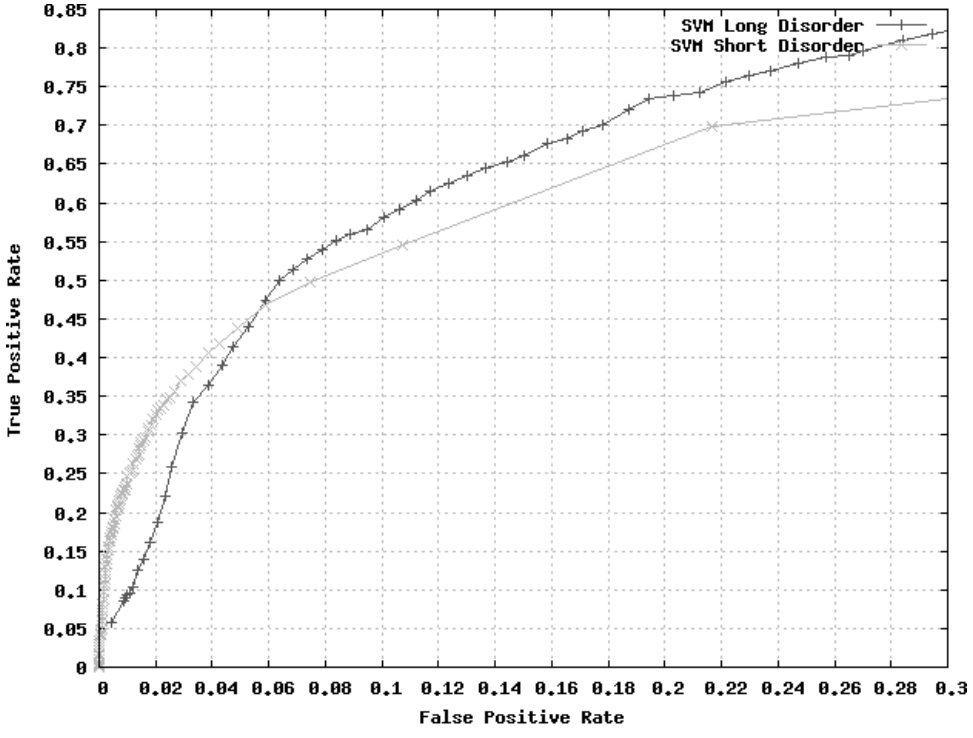

Figure 2. ROC curve of both the long disorder (SVM LD) and the short disorder (SVM SD) experts as computed from the CASP-6 targets.

DISCUSSION

To assess the performance of our method we run a 5-fold cross validation (5-fold CV) procedure with both predictors. CV performance for both short and long disorder experts is given in Table 1 together with an independent assessment of both final predictors (i.e. trained on the whole respective dataset) on the set represented by the DR category targets of the most recent CASP-6 competition (38). We report Matthews correlation coefficient (C), the area under the ROC curve (AUC) and the two state (ordered/disordered) classification accuracy Q2 for a false hit rate set to 5%. To show that these results are competitive with state-of-the-art methods, Table 1 also reports the performance of two top ranking CASP-6 methods. VSL-1 (group id 193) (39) and a preliminary version of Spritz used in CASP-6, participating as group CaspIta (group id 096). We also report the performances of the short disorder expert on the LD set, and of the long disorder expert on the SD set. In both cases the specialized predictor performs significantly better on its own class of disorder, justifying the choice of distinct experts.

Figure 2 shows the ROC curve of both the long disorder (SVM LD) and the short disorder (SVM SD) experts as computed from the CASP-6 targets. This plot is very similar to what we obtain on the 5-folds CV experiments and suggests a threshold switch between the experts, since it shows the long disorder expert consistently outperforming the alternative above 5% FPR. Nonetheless, the choice of the expert is left to the user, to increase the flexibility of the method. To help the user's choice, in the web help page we provide the expected fraction of disordered residues recovered for a given FPR, for both experts.

ACKNOWLEDGEMENTS

A.V. and G.P. are supported by an Embark Fellowship from the Irish Research Council for Science, Engineering and Technology, Science Foundation Ireland grants 04/BR/CS0353 and 05/RFP/CMS0029, grant RP/2005/219 from the Health Research Board of Ireland and UCD President's Award 2004 and Seed Funding scheme 2005 grants. S.C.E.T. is funded by a 'Rientro dei cervelli' grant from the Italian Ministry for Education, University and Research (MIUR). Funding to pay the Open Access publication charges for this article was provided by Science Foundation Ireland's grant 04/BR/CS0353.

Conflict of interest statement. None declared.

REFERENCES

1. Uversky,V.N., Gillespie,J.R. and Fink,A.L. (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.
2. Uversky,V.N. (2002) What does it mean to be natively unfolded? *Eur. J. Biochem.*, **269**, 2–12.
3. Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M. and Obradovic,Z. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.
4. Dunker,A.K. and Obradovic,Z. (2001) The protein trinity—linking function and disorder. *Nat. Biotechnol.*, **19**, 805–806.
5. Bracken,C., Iakoucheva,L.M., Romero,P.R. and Dunker,A.K. (2004) Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Opin. Struct. Biol.*, **14**, 570–576.
6. Fink,A.L. (2005) Natively unfolded proteins. *Curr. Opin. Struct. Biol.*, **15**, 35–41.
7. Dyson,H.J. and Wright,P.E. (2002) Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, **12**, 54–60.
8. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
9. Romero,P., Obradovic,Z. and Dunker,A.K. (2004) Natively disordered proteins: functions and predictions. *Appl. Bioinformatics*, **3**, 105–113.
10. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nature Rev. Mol. Cell Biol.*, **6**, 197–208.
11. Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
12. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
13. Liu,J., Tan,H. and Rost,B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.
14. Moult,J., Fidelis,K., Rost,B., Hubbard,T. and Tramontano,A. (2005) Critical assessment of methods of protein structure prediction (CASP)—round 6. *Proteins*, **61**, 3–7.
15. Radivojac,P., Obradovic,Z., Smith,D.K., Zhu,G., Vucetic,S., Brown,C.J., Lawson,J.D. and Dunker,A.K. (2004) Protein flexibility and intrinsic disorder. *Protein Sci.*, **13**, 71–80.
16. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P., Brown,C.J. and Dunker,A.K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53**, 566–572.
17. Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J. and Dunker,A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.
18. Jones,D.T. and Ward,J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **53**, 573–578.
19. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure (Camb)*, **11**, 1453–1459.
20. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
21. Lise,S. and Jones,D.T. (2005) Sequence patterns associated with disordered regions in proteins. *Proteins*, **58**, 144–150.
22. Vucetic,S., Brown,C.J., Dunker,A.K. and Obradovic,Z. (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.
23. Oldfield,C.J., Cheng,Y., Cortese,M.S., Brown,C.J., Uversky,V.N. and Dunker,A.K. (2005) Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**, 1989–2000.
24. Coeytaux,K. and Poupon,A. (2005) Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, **21**, 1891–1900.
25. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
26. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem.*, **18**, 269–285.
27. Melamud,E. and Moult,J. (2003) Evaluation of disorder predictions in CASP5. *Proteins*, **53**, 561–565.
28. Cuff,J.A. and Barton,G.J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**, 508–519.
29. Albrecht,M., Tosatto,S.C.E., Lengauer,T. and Valle,G. (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng.*, **16**, 459–462.
30. Ward,J.J., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2003) Secondary structure prediction with support vector machines. *Bioinformatics*, **19**, 1650–1655.
31. Vucetic,S., Obradovic,Z., Vacic,V., Radivojac,P., Peng,K., Iakoucheva,L.M., Cortese,M.S., Lawson,J.D., Brown,C.J., Sikes,J.G. *et al.* (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.
32. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z.K. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and

relational database based on the mmCIF schema. *Nucleic Acids Res*., **33**, D233–D237.
33. Shawe-Taylor and Cristianini,N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.
34. Platt,J. (2000) *Probabilistic Outputs of Support Vector Machines and Comparison to Reguralised Likelihood Methods*. MIT press, Cambridge, MA.
35. Keerthi,S.S. and Lin,C.J. (2003) Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput*., **15**, 1667–1689.
36. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
37. Pollastri,G. and McLysaght,A. (2005) Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21**, 1719–1720.
38. Jin,Y. and Dunbrack,R.L.,Jr (2005) Assessment of disorder predictions in CASP6. *Proteins*, **61**, 167–175.
39. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P. and Dunker,A.K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61**, 176–182.