W460–W464 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm363

PrDOS: prediction of disordered protein regions from amino acid sequence

Takashi Ishida[1],* and Kengo Kinoshita[1,2]

[1]Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan and [2]Structure and Function of Biomolecules, SORST JST, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

Received January 31, 2007; Revised April 11, 2007; Accepted April 25, 2007

ABSTRACT

PrDOS is a server that predicts the disordered regions of a protein from its amino acid sequence (http://prdos.hgc.jp). The server accepts a single protein amino acid sequence, in either plain text or FASTA format. The prediction system is composed of two predictors: a predictor based on local amino acid sequence information and one based on template proteins. The server combines the results of the two predictors and returns a two-state prediction (order/disorder) and a disorder probability for each residue. The prediction results are sent by e-mail, and the server also provides a web-interface to check the results.

INTRODUCTION

Recent progress in structural genomics has revealed that many proteins have regions with very flexible and unstable structures, even in their native states. Such proteins or regions are referred to as being natively disordered or unstructured (1). Disordered protein regions often lead to difficulties in purification and crystallization, and become a bottleneck in high throughput structural determination (2). Therefore, it would be quite useful to identify the disordered regions of target proteins from their amino acid sequences.

The prediction of disordered regions is also important for the functional annotation of proteins. In the sense of the classical 'lock-and-key' theory (3), it is hard to imagine that natively disordered regions have some biological meaning. However, disordered regions are reportedly involved in many biological processes, such as regulation, signaling and cell cycle control (4,5). The primary role of natively disordered regions seems to be the molecular recognition of proteins or DNA. Upon binding with ligands, disorder-to-order transitions are frequently observed, where the flexibility of the disordered regions may be necessary to facilitate interactions with multiple partners with high-specificity and low-affinity (6). In addition, recent research has indicated that phosphorylation sites are frequently found in disordered regions, and thus the prediction of phosphorylation sites is expected to be improved by the accurate identification of disordered regions (7).

There are some particular amino acid sequence characteristics in protein disordered regions, such as a higher frequency of hydrophilic and charged residues, or low sequence complexity (4). Thus, the disordered regions are predictable based on these characteristics, and various prediction methods have been reported (8–10).

We have also developed a system to predict disordered regions from the amino acid sequence. Our system is composed of two predictors, that is, a predictor based on the local amino acid sequence, and one based on template proteins (or homologous proteins for which structural information is available). The first part is implemented using a support vector machine (SVM) algorithm (11) for the position-specific score matrix (or profile) of the input sequence. More precisely, a sliding window is used to map individual residues into a feature space. A similar idea has already been used in secondary structure prediction, as in PSIPRED (12). The second part assumes the conservation of intrinsic disorder in protein families (13,14), and is simply implemented using PSI-BLAST (15) and our own measure of disorder, as described later. The final prediction is done as the combination of the results of the two predictors.

The performance of disorder prediction methods has been evaluated since 2002 by the structural biology community at the CASP benchmark, that is, critical assessment of techniques for protein structure prediction (16). In 2006, the seventh round of the CASP benchmark was held, and the assessors also evaluated our method. As a result, our methods achieved high performance [estimated accuracy (Q2) (>90%) with the sensitivity of 0.56], especially for short disordered regions. The details are available at the CASP7 meeting web page at http://predictioncenter.org/casp7/meeting/presentations/Presentations_assessors/CASP7_DR_Bordoli.pdf (our group number is 443, team name is fais). PrDOS is the web interface of this prediction system.

*To whom correspondence should be addressed. Tel: +81-3-5449-5131; Fax: +81-3-5449-5133; Email: t-ishida@hgc.jp

Inputting data and accessing results

The server requires protein amino acid sequences in either plain text or FASTA (17) format as the input. The user can submit a multiple FASTA formatted input to predict disordered regions of multiple proteins. The number of sequences in the multiple FASTA formatted input is limited to 100, due to the limitation of the computational resources. The server accepts the 20 single letter codes for standard amino acids and the code 'X' generally used for non-standard amino acids. The server automatically replaces other letters such as 'U' for a selenocystein by 'X'. The user can choose to receive the prediction result by either e-mail or web-interface, if the user submits a single protein amino acid sequence. The user can also select the prediction false positive rate, which is the rate of residues incorrectly predicted as disordered residues. The allowed rate of false positives strongly depends on the purpose of the prediction. Therefore, the user has to decide on a false positive rate threshold of the classifier, according to the application of the user, but the user can also change this parameter at the result web page. The user can check the true positive rate of each false positive rate from the receiver operating characteristic (ROC) curve on the web page. This ROC curve was derived by calculating the true positive rate at each false positive rate by varying its order/disorder threshold, using the results of the 5-fold cross-validation test for the training set. The default value of this parameter is set to 5%.

Although the calculation time is sensitive to the length of the query protein and the server conditions, a typical prediction will take from 5 to 10 min. The user can check the estimated calculation time on the submission confirmation page. The e-mail results also include the URL of the result web page. The result web page contains the result of the two-state prediction with the given false positive rate, and the disorder profile plot (Figure 1). The user can also download the raw prediction results in the CSV format or the CASP format from the same page.

Figure 1 shows a typical result page as an output. The query protein is HIV-1 negative factor protein, which is known to have disordered regions at the N-terminus in the monomer, and this region is critically important for binding with an SH3 domain (18).

Prediction flow

Step 1: *Making the sequence profile*

The information content of a single amino acid sequence is vastly enriched by using information about homologous proteins. For this purpose, multiple alignments with the homolog are more useful than a single amino acid sequence. In our system, a position-specific score matrix (PSSM or a profile) is used as a more convenient representation of similar information, as compared to a multiple alignment of the homologues. Therefore, in the first step, the target amino acid sequence is converted into a PSSM, using two rounds of PSI-BLAST searches against NCBI non-redundant (nr) amino acid sequence databases (19) with default parameters. Then, the following two predictions are performed using the PSSM.

Step 2: *Prediction based on local amino acid sequence information*

In the first predictor, the prediction is done using SVM, which is a supervised machine learning technique. The SVM was trained using a non-redundant protein chain set from the Protein Data Bank (PDB) (20), using the PISCES server (21). The training set was selected by the following criteria: determined by X-ray crystallography, resolution ≤2.0 Å, *R*-factor ≤0.25, sequence identities to each other ≤20% and sequence length >50. Disordered regions for these proteins were identified as the missing residues denoted at the REMARK 465 lines in the PDB. The residues with crystal or biological contacts with other chains were excluded, because such contacts may stabilize disordered residues into an ordered state. As a result, 1954 chains with 5110 disordered residues (4.8%) and 109 921 ordered residues (95.2%) were used as the training set. The protein sequences information was then converted into the input vector. The input vector consisted of PSSM information and spacers in a 27-residue window centered at the residue (Figure 2). A spacer represents whether the site is beyond N- or C-terminus or not. If the site of a residue was beyond the N- or C-terminus, then the spacer was set to 1; otherwise it was set to 0. Each element of PSSM was converted into the range from −1.0 to 1.0 by dividing by 10. Finally, the dimension of the input vector was 567 [$=(20+1)\times 27$].

For the query sequence, the same encoding is carried out, and using the trained SVM, the disorder propensity of each residue is predicted. It should be noted that SVM is a binary classifier, and thus it returns only order or disorder as prediction results. We use the distances from decision planes in feature spaces called the decision value, as a prediction value.

Step 3: *Template-based prediction*

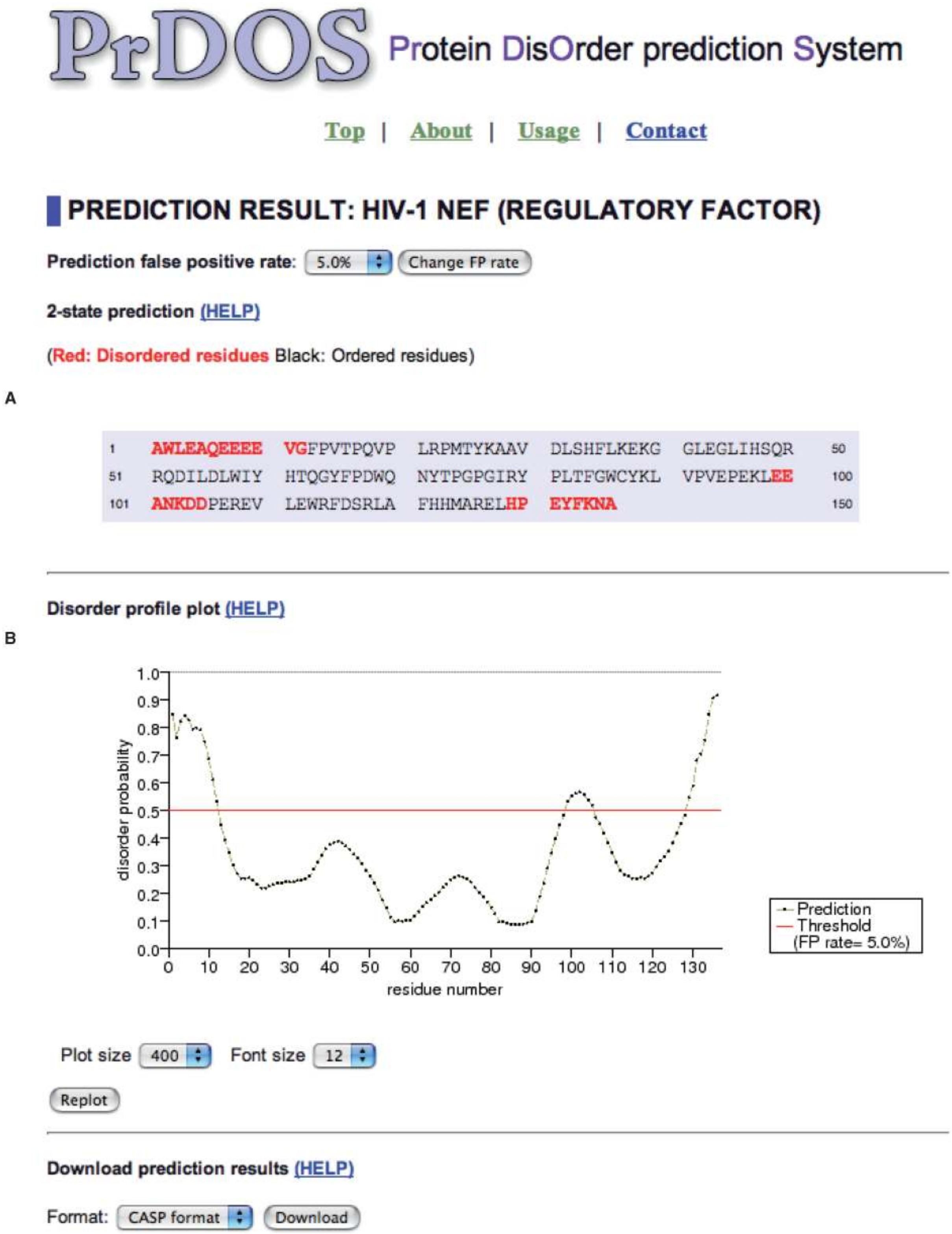
In the second predictor, the prediction is done using the alignments of homologues with structures that have been determined. The sequence homologues are searched against the PDB, using a PSI-BLAST search with the PSSM obtained in the first step. The alignments of the hit sequences with e-values <1.0e-3 are used for the prediction. If there are no significant hits, then this prediction is skipped. The disorder tendency of the *i*th residue, *Pi*, is defined by the following equation:

$$P_i = \frac{\sum_{j=1}^{n} \alpha_j I_j}{n}$$

where n is the number of alignments, Ij is the sequence identity of the *j*th hit and αj is set to 1 if the aligned residue in the *j*th hit is disordered; otherwise, it is 0. In other words, Pi evaluates the weighted ratio of disordered residues among the homologous proteins.

Step 4: *Combining prediction results*

To combine the results of the two independent predictions, the weighted average between the results of the two predictions is calculated. The weight for template-based prediction equals about 0.11, and the weight for prediction based on local amino acid sequence information equals 1.0. These weights are obtained by optimizing



PrDOS Protein DisOrder prediction System

Top | About | Usage | Contact

PREDICTION RESULT: HIV-1 NEF (REGULATORY FACTOR)

Prediction false positive rate: 5.0% Change FP rate

2-state prediction (HELP)

(Red: Disordered residues Black: Ordered residues)

A

```
1     AWLEAQEEEE  VGFPVTPQVP  LRPMTYKAAV  DLSHFLKEKG  GLEGLIHSQR    50
51    RQDILDLWIY  HTQGYFPDWQ  NYTPGPGIRY  PLTFGWCYKL  VPVEPEKLEE   100
101   ANKDDPEREV  LEWRFDSRLA  FHHMARELHP  EYFKNA                   150
```

Disorder profile plot (HELP)

B

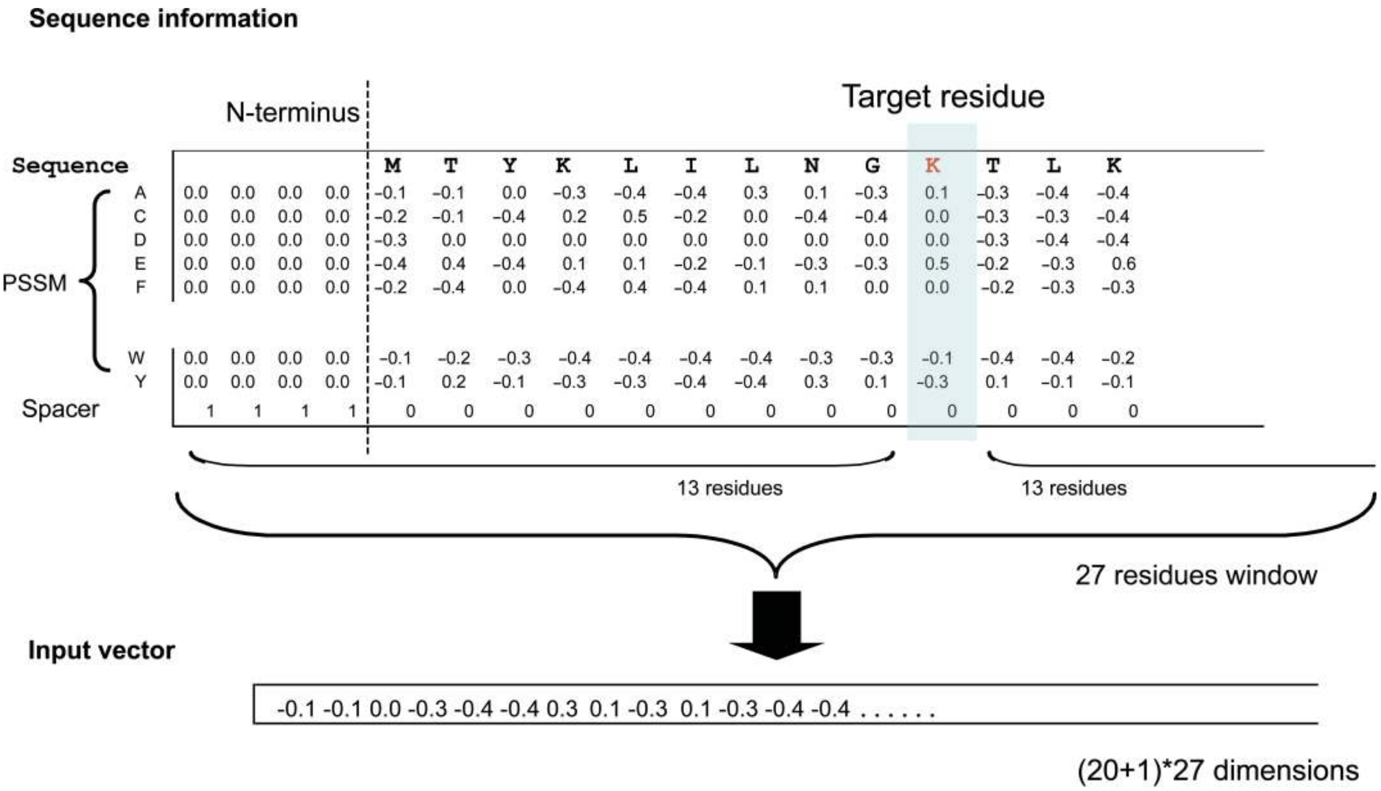
Plot size 400 Font size 12

Replot

Download prediction results (HELP)

Format: CASP format Download

Figure 1. An example of the prediction result page for HIV-1 NEF (PDB code: 2NEF). (**A**) The prediction result of the two-state prediction (disorder/order) is shown in this part. The red residues are predicted to be disordered at the given prediction false positive rate. (**B**) The plot of disorder probability of each residue along the sequence is shown in this part. Residues beyond the red threshold line in this plot are predicted to be disordered. The user can change the size of the plot through the web-interface.



Sequence					M	T	Y	K	L	I	L	N	G	K	T	L	K
A	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	-0.3	-0.4	-0.4	0.3	0.1	-0.3	0.1	-0.3	-0.4	-0.4
C	0.0	0.0	0.0	0.0	-0.2	-0.1	-0.4	0.2	0.5	-0.2	0.0	-0.4	-0.4	0.0	-0.3	-0.3	-0.4
D	0.0	0.0	0.0	0.0	-0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.3	-0.4	-0.4
E	0.0	0.0	0.0	0.0	-0.4	0.4	-0.4	0.1	0.1	-0.2	-0.1	-0.3	-0.3	0.5	-0.2	-0.3	0.6
F	0.0	0.0	0.0	0.0	-0.2	-0.4	0.0	-0.4	0.4	-0.4	0.1	0.1	0.0	0.0	-0.2	-0.3	-0.3
W	0.0	0.0	0.0	0.0	-0.1	-0.2	-0.3	-0.4	-0.4	-0.4	-0.4	-0.3	-0.3	-0.1	-0.4	-0.4	-0.2
Y	0.0	0.0	0.0	0.0	-0.1	0.2	-0.1	-0.3	-0.3	-0.4	-0.4	0.3	0.1	-0.3	0.1	-0.1	-0.1
Spacer	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

-0.1 -0.1 0.0 -0.3 -0.4 -0.4 0.3 0.1 -0.3 0.1 -0.3 -0.4 -0.4 . . . . . .

Figure 2. Diagram of sequence encoding scheme. The sequence information in a 27-residue window is converted into an input vector by aligning the elements in a certain order. For each site, the value of each element of PSSM for 20 amino acid types and the spacer information are appended to the input vector, thus total dimension of the input vector is 567 [=(20 + 1) × 27].

the ROC score (22) of the result of the 5-fold cross-validation test. Next, a low-pass filter by moving-average is applied along the sequence to smooth the prediction results. This smoothing process is performed to avoid unrealistic predictions, such as the case that an isolated ordered residue exists in a long disordered region. Finally, the prediction values are scaled from 0.0 to 1.0, so the values can correspond to the disorder probability used in the CASP.

ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports and Technology of Japan. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, The University of Tokyo. Funding to pay the Open Access publication charges for this article was provided by Japan Science and Technology Agency, Institute for Bioinformatics Research and Development (JST-BIRD).

Conflict of interest statement. None declared.

REFERENCES

1. Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 523–533.
2. Oldfield,C.J., Ulrich,E.L., Cheng,Y., Dunker,A.K. and Markley,J.L. (2005) Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins*, **59**, 444–453.
3. Fischer,E. (1894) Einfluss der configuration auf die wirkung der enzyme. *Ber. Dt. Chem. Ges.*, **27**, 2985–2993.
4. 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.
5. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
6. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
7. Lakoucheva,L.M., Radivojac,P., Brown,C.J., O'Connor,T.R., Sikes,J.G., Obradovic,Z. and Dunker,A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.
8. Garner,E., Cannon,P., Romero,P., Obradovic,Z. and Dunker,A.K. (1998) Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Inform. Ser. Workshop Genome Inform.*, **9**, 201–213.
9. 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*, **11**, 1453–1459.
10. Jones,D.T. and Ward,J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **53**, 573–578.
11. Vapnik,V. (1998) *Statistical Learning Theory*. John Wiley & Sons, New York.
12. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **17**, 195–202.
13. 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.

14. Chen,J.W., Romero,P., Uversky,V.N. and Dunker,A.K. (2006) Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J. Proteome Res.*, **5**, 879–887.
15. 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.
16. Melamud,E. and Moult,J. (2003) Evaluation of disorder predictions in CASP5. *Proteins*, **53**, 561–565.
17. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci.*, **85**, 2444–2448.
18. Lee,C.H., Saksela,K., Mirza,U.A., Chait,B.T. and Kuriyan,J. (1996) Crystal structure of the conserved core of HIV-1 Nef complexed with a Src family SH3 domain. *Cell*, **14**, 931–942.
19. McEntyre,J. and Ostell,J. (2005) *The NCBI Handbook*. National Library of Medicine (US), NCBI, Bethesda, MD.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. Wang,G. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
22. Zweig,M.H. and Campbell,G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, **39**, 561–577.