



Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data

JIANLIN CHENG

MICHAEL J. SWEREDOSKI

PIERRE BALDI

pfbaldi@ics.uci.edu (<http://www.ics.uci.edu/~pfbaldi>)

*School of Information and Computer Science, Institute for Genomics and Bioinformatics,
University of California Irvine, Irvine, CA 92697, USA*

Abstract. Intrinsically disordered regions in proteins are relatively frequent and important for our understanding of molecular recognition and assembly, and protein structure and function. From an algorithmic standpoint, flagging large disordered regions is also important for *ab initio* protein structure prediction methods. Here we first extract a curated, non-redundant, data set of protein disordered regions from the Protein Data Bank and compute relevant statistics on the length and location of these regions. We then develop an *ab initio* predictor of disordered regions called DISpro which uses evolutionary information in the form of profiles, predicted secondary structure and relative solvent accessibility, and ensembles of 1D-recursive neural networks. DISpro is trained and cross validated using the curated data set. The experimental results show that DISpro achieves an accuracy of 92.8% with a false positive rate of 5%. DISpro is a member of the SCRATCH suite of protein data mining tools available through <http://www.igb.uci.edu/servers/psss.html>.

Keywords: protein structure prediction, disordered regions, recursive neural networks

1. Introduction

Proteins are fundamental organic macromolecules consisting of linear chains of amino acids bonded together by polypeptide bonds and folded into complex three-dimensional structures. The biochemical function of a protein depends on its three-dimensional structure, thus solving protein structures is a fundamental goal of structural biology. Tremendous efforts have been made to determine the three-dimensional structures of proteins in the past several decades by experimental and computational methods. Experimental methods such as X-ray diffraction and NMR (Nuclear Magnetic Resonance) spectroscopy are used to determine the coordinates of all the atoms in a protein and thus its three-dimensional structure. While most regions in a protein assume stable structures, some regions are partially or wholly unstructured and do not fold into a stable state. These regions are labeled as disordered regions by structural biologists.

Intrinsically disordered proteins (IDPs) play important roles in many vital cell functions including molecular recognition, molecular assembly, protein modification, and entropic chain activities (Dunker et al., 2002). One of the evolutionary advantages of proteins with disordered regions may be their ability to have multiple binding partners and potentially partake in multiple reactions and pathways. Since the disordered regions may be determined only when the IDPs are in a bound state, IDPs have prompted scientists to reevaluate the structure-implies-function paradigm (Wright and Dyson, 1999). Disordered regions

have also been associated with low sequence complexity and an early survey of protein sequences based on sequence complexity predicted that a substantial fraction of proteins contain disordered regions (Wootton, 1994). This prediction has been confirmed to some extent in recent years by the growth of IDPs in the Protein Data Bank (PDB) (Berman et al., 2000), which currently contains about 26,000 proteins and 16,300,000 residues. Thus the relatively frequent occurrence of IDPs and their importance for understanding protein structure/function relationships and cellular processes makes it worthwhile to develop predictors of protein disordered regions. Flagging large disordered regions may also be important for *ab initio* protein structure prediction methods. Furthermore, since disordered regions often hamper crystallization, the prediction of disordered regions could provide useful information for structural biologists and help guide experimental designs. In addition, disordered regions can cause the poor expression of a protein in bacteria, thus making it difficult to manufacture the protein for crystallization or other purposes. Hence, disordered region predictions could provide biologists with important information that would allow them to improve the expression of the protein. For example, if the N or C termini regions were disordered, they could be omitted from the gene.

Comparing disorder predictors can be difficult due to the lack of a precise definition of disorder. Several definitions exist in the literature including loop/coil regions where the carbon alpha (C_α) on the protein backbone has a high temperature factor and residues in the PDB where coordinates are missing as noted in a REMARK465 PDB record (Linding et al., 2003). Here, consistent with Ward et al. (2004), we define a disordered residue as any residue for which no coordinates exist in the corresponding PDB file.

Previous attempts at predicting disordered regions have used sequence complexity, support vector machines, and neural networks (Wootton, 1994; Dunker et al., 2002; Linding et al., 2003; Ward et al., 2004). Our method for predicting disordered regions, called DISpro, involves the use of evolutionary information in the form of profiles, predicted secondary structure and relative solvent accessibility, and 1D-recursive neural networks (1D-RNN). These networks are well suited for predicting protein properties and have been previously used in our SCRATCH suite of predictors, including our secondary structure and relative solvent accessibility predictions (Pollastri et al., 2001a, 2001b; Baldi and Pollastri, 2003).

2. Methods

2.1. Data

The proteins used for the training and testing of DISpro were obtained from the PDB in May 2004. At that time, 7.6% (3,587) of the protein chains in the PDB obtained by X-ray crystallography contained at least one region of disorder at least three residues in length. Most of these disordered regions were short segments near the two ends of protein chains (N- and C-termini).

We first filtered out any proteins that were not solved by X-ray diffraction methods, were less than 30 amino acids in length, or had resolution coarser than 2.5 Å. Next, the proteins were broken down into their individual chains. For the creation of our training and testing sets, we selected only protein chains that had sections of disordered regions strictly greater

than three residues in length. The determination of residues as being ordered or disordered is based on the existence of an ATOM field (coordinate) for C_{α} atom of a given residue in the PDB file. If no ATOM records exist for a residue listed in the SEQRES record, the residue is classified as disordered.

We then filtered out homologous protein chains using UniqueProt (Mika and Rost, 2003) with a threshold HSSP value of 10. The HSSP value between two sequences is a measure of their similarity taking into account both sequence identity and sequence length. An HSSP value of 10 corresponds roughly to 30% sequence identity for a global alignment of length 250 amino acids.

Secondary structure and relative solvent accessibility were then predicted for all the remaining chains by SSpro and ACCpro (Pollastri et al., 2001a, b; Baldi and Pollastri, 2003). Using predicted, rather than true secondary structure and solvent accessibility, which are easily-obtainable by the DSSP program (Kabsch and Sander, 1983), introduces additional robustness in the predictor, especially when it is applied to sequences with little or no homology to sequences in the PDB. The filtering procedures resulted in a set of 723 non-redundant disordered chains. To leverage evolutionary information, PSI-BLAST (Altschul et al., 1997) is used to generate profiles by aligning all chains against the Non-Redundant (NR) database, as in (Jones, 1999; Przybylski and Rost, 2002; Pollastri et al., 2001b). As in the case of secondary structure prediction, profiles rather than primary sequences are used in the input, as explained in next section. Finally, these chains were randomly split into ten subsets of approximately equal size for ten-fold cross-validated training and testing. The final dataset is available at: <http://www.ics.uci.edu/~baldig/scratch/>.

The final dataset has 215,612 residues, 6.4% (13,909) of which are classified as disordered. Of the 13,909 disordered residues, 13.8% (1,924) are part of long regions of disorder (≥ 30 AA). Figure 1 shows a histogram of the frequency of disordered region lengths in our dataset.

2.2. *Input and output of neural networks*

The problem of predicting disordered regions can be viewed as a binary classification problem for each residue along a one dimensional (1-D) protein chain. The residue at position i is labeled as ordered or disordered. A variety of machine learning methods can be applied to this problem, such as probabilistic graphical models, kernel methods, and neural networks. DISpro employs 1-D recursive neural networks (1D-RNN)(Baldi and Pollastri, 2003). For each chain, our input is the 1-D array I , where the size of I is equal to the number of residues in the chain and each entry I_i is a vector of dimension 25 encoding the profile as well as secondary structure and relative solvent accessibility at position i . Specifically, twenty of the values are real numbers which correspond to the amino acid frequencies in the corresponding column of the profile. The other five values are binary. Three of the values correspond to the predicted secondary structure class (Helix, Strand, or Coil) of the residue and the other two correspond to the predicted relative solvent accessibility of the residue (i.e., under or over 25% exposed).

The training target for each chain is the 1-D binary array T , whereby each T_i equals 0 or 1 depending on whether residue at position i is ordered or disordered. Neural networks (or

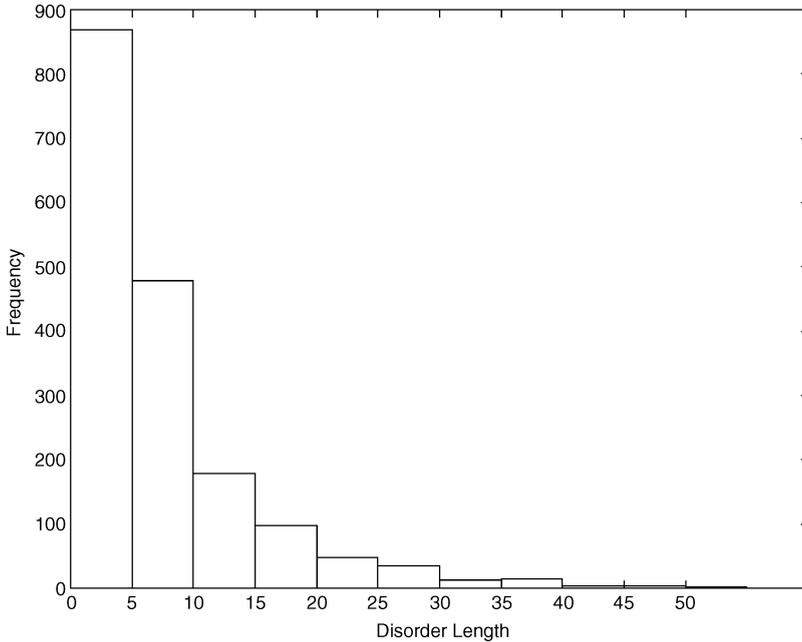


Figure 1. Frequency of lengths of disordered regions.

other machine learning methods) can be trained on the data set to learn a mapping from the input array I onto an output array O , whereby O_i is the predicted probability that residue at position i is disordered. The goal is to make the output O as close as possible to the target T .

2.3. The architecture of 1-D-recursive neural networks (1D-RNNs)

The architecture of the 1D-RNNs used in this study is derived from the theory of probabilistic graphical models, but use a neural network parameterization to speed up belief propagation and learning (Baldi and Pollastri, 2003). 1D-RNNs combine the flexibility of Bayesian networks with the fast, convenient, parameterization of artificial neural networks without the drawbacks of standard feedforward neural networks with fixed input size. Under this architecture, the output O_i depends on the entire input I instead of a local fixed-width window centered at position i . Thus, 1D-RNNs can handle inputs with variable length and allow classification decisions to be made based on contextual long-ranged information outside of the traditional local input window. Since 1D-RNNs use weight sharing in both their forward and backward recursive networks (see below), only a fixed number of weights are required to handle propagation of long-ranged information. This is in contrast to local window approaches, where the number of weights (parameters) typically grows linearly with the size of the window, increasing the danger of overfitting. Nevertheless, it is important to recognize that since the problem of disordered region prediction can be formulated as a

standard classification problem, other machine learning or data mining algorithms such as feed forward neural networks or support vector machines can in principle be applied to this problem effectively, provided great care is given to the problem of overfitting.

The architecture of the 1D-RNN is described in figures 2 and 3 and is associated with a set of input variables I_i , a forward H_i^F and backward H_i^B chain of hidden variables, and a set O_i of output variables. In terms of probabilistic graphical models (Bayesian networks), this architecture has the connectivity pattern of an input-output HMM (Bengio and Frasconi,

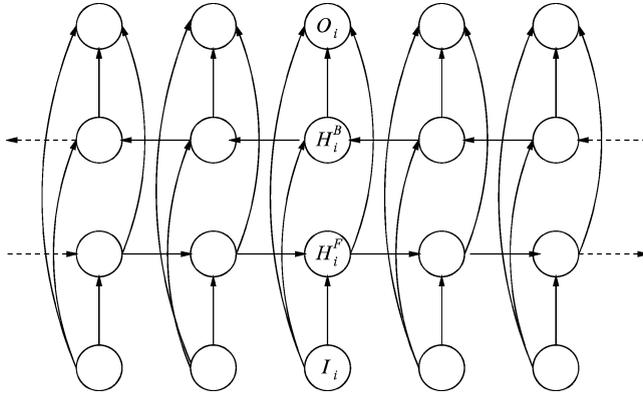


Figure 2. 1D-RNN associated with input variables, output variables, and both forward and backward chains of hidden variables.

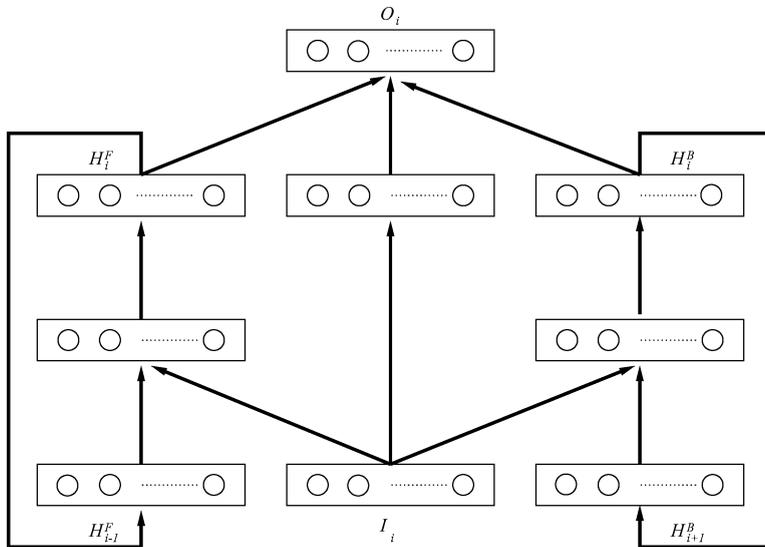


Figure 3. A 1D-RNN architecture with a left (forward) and right (backward) context associated with two recurrent networks (wheels).

1996), augmented with a backward chain of hidden states. The backward chain is of course optional and used here to capture the spatial, rather than temporal, properties of biological sequences.

The relationship between the variables can be modeled using three separate neural networks to compute the output, forward, and backward variables respectively. These neural networks are replicated at each position i ; (i.e., weight sharing). One fairly general form of weight sharing is to assume stationarity for the forward, backward, and output networks, which leads to a 1D-RNN architecture, previously named a bidirectional RNN architecture (BRNN), and is implemented using three neural networks \mathcal{N}_O , \mathcal{N}_F , and \mathcal{N}_B in the form

$$\begin{aligned} O_i &= \mathcal{N}_O(I_i, H_i^F, H_i^B) \\ H_i^F &= \mathcal{N}_F(I_i, H_{i-1}^F) \\ H_i^B &= \mathcal{N}_B(I_i, H_{i+1}^B) \end{aligned} \tag{1}$$

as depicted in figure 3. In this form, the output depends on the local input I_i at position i , the forward (upstream) hidden context $H_i^F \in \mathbb{R}^n$ and the backward (downstream) hidden context $H_i^B \in \mathbb{R}^m$, with usually $m = n$. The boundary conditions for H_i^F and H_i^B can be set to 0, i.e. $H_0^F = H_{N+1}^B = 0$ where N is the length of the sequence being processed. Alternatively these boundaries can also be treated as a learnable parameter. Intuitively, we can think of \mathcal{N}_F and \mathcal{N}_B in terms of two “wheels” that can be rolled along the sequence. For the prediction at position i , we roll the wheels in opposite directions starting from the N- and C-terminus and up to position i . We then combine the wheel outputs at position i together with the input I_i to compute the output prediction O_i using \mathcal{N}_O .

The output O_i for each residue position i is computed by two normalized-exponential units, which is equivalent to one logistic output unit. The error function is the relative entropy between the true distribution and the predicted distribution.

All the weights of the 1D-RNN architecture, including the weights in the recurrent wheels, are trained in supervised fashion using a generalized form of gradient descent on the error function, derived by unfolding the wheels in space. To improve the statistical accuracy, we average over an ensemble of five trained models to make prediction.

3. Results

We evaluate DISpro using ten-fold cross validation on the curated dataset of 723 non-redundant protein chains. The resulting statistics for DISpro are given in Table 1, including a separate report for the special subgroup of long disordered regions (>30 residues), which have been shown to have different sequence patterns than N- and C-termini disordered regions (Li et al., 1999). Performance is assessed using a variety of standard measures including correlation coefficients, area under the ROC curves, Accuracy at 5% FPR (False Positive Rate), Precision [TP/(TP + FP)], and Recall [TP/(TP + FN)]. The accuracy at 5% FPR is defined as [(TP + TN)/(TP + FP + TN + FN)] when the decision threshold is set so that 5% of the negative cases are above the decision threshold. Here, TP, FP, TN, and

Table 1. Results for DISpro on 723 non-homologous protein chains.

Dataset	Corr. Coef.	ROC area	Accuracy		
			(5% FPR)	Precision	Recall
All disorder	0.589	0.878	92.8%	75.4%	38.8%
Long disorder (≥ 30 AA)	0.255	0.789	94.5%	22.1%	25.9%

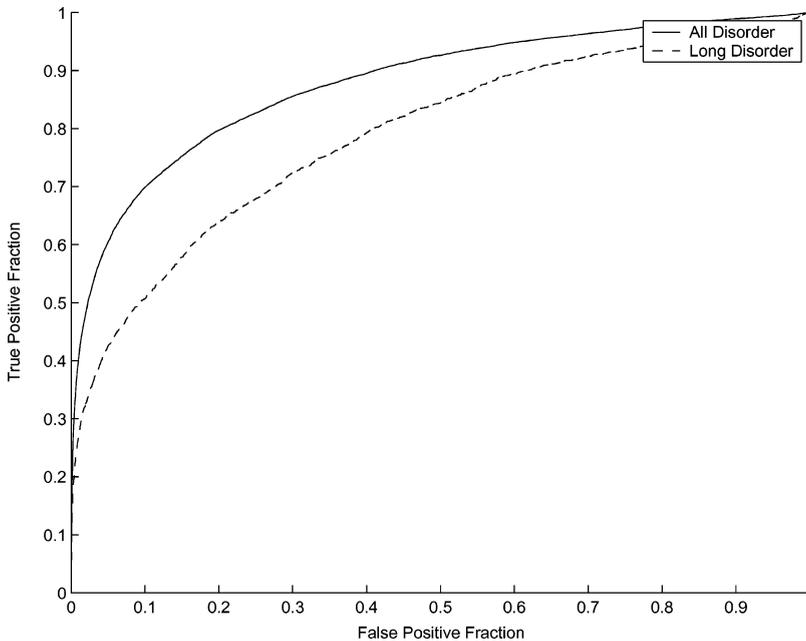


Figure 4. ROC curve for DISpro on set of 723 protein chains.

FN refer to the number of true positives, false positives, true negatives, and false negatives respectively.

The area under the ROC curve of DISpro computed on all regions is .878. An ROC area of .90 is generally considered a very accurate predictor. An area of 1.00 would correspond to a perfect predictor and an area of .50 would correspond to a random predictor. At 5% FRP, the TPR is 92.8% for all disordered regions. DISpro achieves a precision and recall rate of 75.4 and 38.8% respectively, when the decision threshold is set at .5. Figure 4 shows the ROC curves of DISpro corresponding to all disordered regions and to disordered regions 30 residues or more in length. It shows that the long disordered regions are harder to predict than the shorter disordered regions.

We have also compared our results to those of other predictors from CASP5 (Ward et al., 2004) (Critical Assessment of Structure Prediction). The set of proteins from CASP5 should be considered a fair test since each chain had a low HSSP score (< 7) in comparison to our

Table 2. Summary of comparison results for six predictors using the proteins from CASP5. Results for predictors other than DISpro were reported by Ward et al. (2004).

Predictor	Corr. coef.	ROC area	Accuracy (5% FPR)
DISpro	0.51	0.935	93.2%
DISOPRED2	0.52	0.900	93.1%
Dunker VLXT	0.31	0.809	91.4%
Dunker VL2	0.36	0.786	91.8%
Obradovic VL3	0.38	0.801	92.1%
FoldIndex	0.26	0.738	91.0%

training set. Table 2 shows our results in comparison to other predictors. DISpro achieves an ROC area of 0.935, better than all the other predictors. The correlation coefficient of DISpro is 0.51, roughly the same as DISOPRED2. The accuracy of DISpro at a 5% FPR is 93.2% on the CASP5 protein set. Thus, on the CASP5 protein set, DISpro is roughly equal or slightly better than all the other predictors on all three performance measures. DISOPRED2 and DISpro performance appear to be similar and significantly above all other predictors.

4. Conclusion

DISpro is a predictor of protein disordered regions which relies on machine learning methods and leverages evolutionary information as well as predicted secondary structure and relative solvent accessibility. Our results show that DISpro achieves an accuracy of 92.8% with a false positive rate of 5% on large cross-validated tests. Likewise, DISpro achieves an ROC area of 0.88.

There are several directions for possible improvement of DISpro and disordered region predictors in general that are currently under investigation. To train better models, larger training sets of proteins with disordered regions can be created as new proteins are deposited in the PDB. In addition, protein sequences containing no disordered regions, which are currently excluded from our current datasets, may also be included in the training set to decrease the false positive rate.

Our results confirm that short and long disordered region behave differently and therefore it may be worth training two separate predictors. In addition, it is also possible to train a separate predictor to detect whether a given protein chain contains any disordered regions or not using another machine learning technique, such as kernel methods, for classification, as is done for proteins with or without disulphide bridges (Frasconi et al., 2002). Results derived from contact map predictors (Baldi and Pollastri, 2003) may also be used to try to further boost the prediction performance. It is reasonable to hypothesize that disordered regions ought to have poorly defined contacts. We are also in the process of adding to DISpro the ability to directly incorporate disorder information from homologous proteins. Currently, such information is only used indirectly by the 1D-RNNs. Prediction of disordered regions in proteins that have a high degree of homology to proteins in the PDB should

not proceed entirely from scratch but leverage the readily available information about disordered regions in the homologous proteins. Large disordered regions may be flagged and be removed or treated differently in *ab initio* tertiary structure prediction methods. Thus it might be useful to incorporate disordered region predictions into the full pipeline of protein tertiary structure prediction. Finally, beyond disorder prediction, bioinformatics integration of information from different sources may shed further light on the nature and role of disordered regions. In particular, if disordered regions act like reconfigurable switches allowing certain proteins to partake in multiple interactions and pathways, one might be able to cross-relate information from pathway and/or protein-protein interaction databases with protein structure databases.

Acknowledgments

The authors wish to thank anonymous reviewers for helpful comments. Work supported by the Institute for Genomics and Bioinformatics at UCI and a Laurel Wilkening Faculty Innovation award, an NIH Biomedical Informatics Training grant (LM-07443-01), an NSF MRI grant (EIA-0321390), a Sun Microsystems award, a grant from the University of California Systemwide Biotechnology Research and Education Program (UC BREP) to PB.

References

- Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402.
- Baldi, P. and Pollastri, G. 2003. The principled design of large-scale recursive neural network architectures—DAG-RNNs and the protein structure prediction problem. *Journal of Machine Learning Research*, 4:575–602.
- Bengio, Y. and Frasconi, P. 1996. Input-output HMM's for sequence processing. *IEEE Transactions on Neural Networks*, 7(5):1231–1249.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P. 2000. The protein data bank. *Nucleic Acids Research*, 28:235–242.
- Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. and Obradovic, Z. 2002. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582.
- Frasconi, P., Passerini, A., and Vullo, A. 2002. A two-stage svm architecture for predicting the disulfide bonding state of cysteines. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 25–34.
- Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202.
- Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.
- Li, X., Romero, P., Rani, M., Dunker, A., and Obradovic, Z. 1999. Predicting protein disorder for n-, c-, and internal regions. *Genome Inform.*, 42:38–48.
- 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(11):1453–1459.
- Mika, S. and Rost, B. 2003. Uniqueprot: Creating representative protein-sequence sets. *Nucleic Acids Res.*, 31(13):3789–3791.
- Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R. 2001a. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 47:142–153.
- Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. 2001b. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47:228–235.

- Przybylski, D. and Rost, B. 2002. Alignments grow, secondary structure prediction improves. *Proteins*, 46:195–205.
- 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. *Journal of Molecular Biology*, 337(3):635–645.
- Wootton, J. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Computational Chemistry*, 18:269–285.
- Wright, P.E. and Dyson, H.J. 1999. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331.