Template-based Recognition of Natively Disordered Regions in Proteins

Alessandro Vullo, Cliona P.Roche, and Gianluca Pollastri*

University College Dublin, Technical Report UCD-CSI-2012-01, April 2012

Abstract

Disordered proteins are increasingly recognised as a fundamental component of the cellular machinery. Parallel to this, the prediction of protein disorder by computational means has emerged as an aid to the investigation of protein functions. Although predictors of disorder have met with considerable success, it is increasingly clear that further improvements are most likely to come from additional sources of information, to complement patterns extracted from the primary sequence of a protein. In this article, a system for the prediction of protein disorder that relies both on sequence information and on structural information from homologous proteins of known structure (templates) is described. Structural information is introduced directly (as a further input to the predictor) and indirectly through highly reliable template-based predictions of structural features of the protein. The predictive system, based on Support Vector Machines, is tested by rigorous 5-fold cross validation on a large, non-redundant set of proteins extracted from the Protein Data Bank. In these tests the introduction of structural information, which is carefully weighed based on sequence identity between homologues and query, results in large improvements in prediction accuracy. The method, when re-trained on a 2004 version of the PDB, clearly outperforms the algorithms that ranked top at the 2006 CASP competition.

1 Introduction

Over the last years, intrinsically disordered proteins (IDPs) have been repeatedly shown to be an important piece of knowledge in functional proteomics. Recent progresses in the field very frequently involve the recognition of an IDP as being a fundamental actor of some molecular process, such as recognition, post-translational modification and assembly. At a higher level, IDPs make signalling, regulation and control pathways possible.

*School of Computer Science and Informatics and Complex and Adaptive Systems Laboratory, University College Dublin, Ireland

The relatively recent discovery and analyses of many IDPs have prompted the development of methods for identifying the location of disordered segments in protein sequences [12]. The problem is generally formulated as a binary classification task, with "disordered" and "ordered" being the two classes, and several systems have achieved relatively high rates of correct classification at it, see for instance [25, 41, 10, 44, 6]. This seems to imply that there are fairly strong sequence patterns (e.g. sequence complexity, net charge and hydropathy, sequence patterns), which are remarkably different in ordered and disordered regions.

In spite of this success, some factors indicate that prediction of intrinsic disorder from sequence is close to a point where only little improvements are possible [32]. As emerged in the CASP7 assessment of disorder [3], some key problems still persist, such as the reliable identification of disordered regions outside sequence termini, and overprediction i.e. the fact that more residues are predicted to be disordered than is really the case. Moreover, no clear improvement has been observed between CASP6 and CASP7. Regardless of the sophistication of the statistical or learning algorithm employed, improvements in disorder prediction have primarily come from a careful combination of informative features and what may be termed the "specialise, combine, filter" paradigm. That is, respectively: training different models on different aspects of disorder (e.g. short vs. long); combining their outputs by some voting or weighting scheme; and finally filtering the resulting outputs to eliminate unlikely predictions, either by a second-stage machine learning model or based on some hard rule [41, 25, 38, 13, 15, 17].

As pointed out in [32], novel types of data and computational techniques are required to improve protein disorder prediction, for it to assist more effectively our understanding of functional mechanisms. Given that none of the algorithms tested on the problem seems to have a clear edge over the others [12], the exploration of alternative sources of data, other than the sequence, is a natural candidate for current research efforts.

All methods for predicting disorder use sequence information in some form, and can be distinguished according to the granularity and level of complexity employed: average physicochemical features [37, 35], sequence composition [7, 10], evolutionary information from homologous proteins [21, 6] and in combination with reduced amino acid alphabets [43, 36, 34]. Prediction from a multiple alignment of protein sequences rather than a single sequence has long been recognised as a way to improve prediction accuracy for virtually all protein structural features: secondary structure [33, 8, 29, 28], solvent accessibility and coordination number [24, 26, 27], β-sheet pairing [2], contact maps [11, 38, 5, 39], etc. The role of evolutionary information for the case of disorder prediction is less clear, even though regions of conserved disorder can be found in protein families and domains [4]. Experimental validation confirms the benefit of inputting profiles instead of just the sequence [42].

Structural knowlege about a protein has also been considered for locating disordered regions, mainly in the form of predicted structural features such as secondary structure and solvent accessibility [42, 6]. So far, the contribution of

predicted (rather than exact) structural features has not proven to be especially beneficial. As pointed out in [42], this may depend on the fact that the presence of certain structural features is implicitly encoded into the profile of homologous sequences that is input to all the most successful predictors. Thus, explicitly providing predictions of these features yields no further gains. Nonetheless, it is clear that structural information about a protein, if accurate enough, should contribute to accurate prediction of disorder: for instance, by definition a disordered region cannot be found within a regular local structure such as a helix or a strand, hence accurate knowledge of locations of helices and strands is bound to help disorder prediction.

As conjectured in [6], other forms of structural knowledge can also be taken into account, if one assumes that disorder is somewhat conserved across members of the same family [4, 17, 3].

In this article, it is argued that significant improvements in protein disorder prediction may come from: (1) the use of highly reliable predictions of structural features such a secondary structure and solvent accessibility; (2) careful selection of information from evolutionarily related proteins of known structure (and, by extension, disorder), when available. In [27], secondary structure and solvent accessibility were showed to be predicted with high reliability by providing the predictors with both evolutionary information from homologous sequences and weighted structural profiles derived from sets of homologues of known structure (templates) from the Protein Data Bank (PDB) [9]. For sequence similarity exceeding 30%, secondary structure prediction quality is approximately 90%, close to its theoretical maximum, and 2-class solvent accessibility roughly 85%. This approach was extended to the prediction of protein contact maps [39], protein domains [40], and to the extraction of inoformation from remote homologues [23]. The solution presented here draws on this work, extending it and applying it to disorder prediction. The disorder predictor is implemented as a simple, single Support Vector Machine (hence no voting scheme) with no specialisation on long or short disorder, and no second stage prediction or post-processing of predictions. The first step is encoding secondary structure and solvent accessibility information into the input of the predictor. This information is predicted by Porter and PaleAle [28, 27, 23], that use templates from the PDB alongside the sequence, hence indirectly contain template information. The second step is inputting templates, in the form of a weighed profile, directly to the Support Vector Machine.

Different combinations of sequence and structural inputsare compared by 5-fold cross validation experiments on a large subset of non-redundant chains extracted from a recent version of the PDB. The results of these experiments indicate that both structural feature predictions and disorder templates contribute to large and significant improvements compared with the use of sequence information alone. The best results are obtained when all sources of information are combined: the sequence, secondary structure and solvent accessibility predictions, and the templates. The predictive system is also trained on a publicy available dataset of disordered chains [6], compiled before 2005, and compared on the targets from the CASP7 assessment [3] against the top-ranking predic-

tors at the same CASP. Surprisingly enough, in spite of its lack of specialisation, voting, or filtering (which leaves room for further improvements), this simple SVM outperforms all the best methods at CASP7. Although the predictor performs especially well in the case when templates are available, it also achieves state-of-the-art performances when these are not. A first implementation of the system described, dubbed *Punch*, is freely accessible on the Web at the address `http://distill.ucd.ie/punch`.

2 Materials and Methods

2.1 Datasets

Given a dataset of proteins with unfolded sequence fragments, training a learning algorithm on it in supervised fashion requires the adoption of an unambigous definition of disorder. This is difficult in the absence of an experimental characterization of disorder. In order to solve this problem, an operational defintion of disorder has to be devised, based of some form of resistence to self organisation and/or structural determination. The most popular of these definitions prescribes the a residue be considered as disordered if there are no coordinates for it in the PDB file (REMARK 465 or no ATOM record for a residue listed in the SEQRES record). Here a slightly different definition of disorder is considered, which takes into account information from alternative data sources other than the PDB, although derived from it. Specifically, disodered residues are identified by aligning the sequence of each chain in the PDB SEQRES records with the corresponding DSSP sequence in the PDBFinderII database [16]. The main difference with respect to the more conventional definition is that non standard or modified residues (like TRN) having unique coordinates are classified as ordered, in spite of being assigned a HETATM field in the PDB. The use of this defintion prevents from: assigning as disordered many regions of one or two amino acids which could affect stability but are not likely to be interesting for functional studies of disorder; learning the concept that disorder is consistently associated with the presence of non conventional residues.

Two datasets are considered in this paper. One is a large set of disordered protein chains derived from a recent version of the PDB. This set is used to make a robust comparison of performance between different combinations of input features, using 5-fold cross validation. In particular, this article investigates whether template information significantly improves over sequence-based inputs and the extent to which sequence similarity between the query and the templates contributes to predictive accuracy. In another set of experiments, the same simple binary classifier is trained on the publicy available DISpro dataset [6], which includes disordered proteins in the PDB as of May 2004. By training on this set a fair comparison is obtained with the DISpro predictor and with the other methods participating in the last CASP7 assessment of disorder prediction, which took place in 2006. The DISpro dataset is filtered in order to contain chains where DSSP could produce an output and the DSSP sequence

and PDBFinder sequence match, ending up with a total of 690 disordered chains compared to the 723 sequences of the original set.

The proteins used for cross-validation are obtained from the 47812 entries present in the PDB on December 17 2007. DNA chains, proteins sequences shorter than 30 amino acids, and those for which no information in the PDBFinderII database (as available on January 16 2008) could be found are filtered out. Structures not solved by X-ray crystallography or with a resolution $\geq$ 2.0 Å, and without any disordered region at least three residues in length are also removed. The PDB subset thus obtained is then redundancy reduced at 25% pairwise sequence similarity, resulting in a final set of 2649 protein chains with 652417 residues of which 42033 (6.4%) are disordered and the remaining 93.6% are ordered. The dataset resulting from the filtering procedure is called here PDBD2649.

To perform 5-fold cross validation, this set is split into five subsets with an equal number of chains. As a result of the large dimension of the subsets, the number of residues in each fold is also approximately the same and the distribution of disordered/ordered residues closely mimics that of the whole set, thus increasing the robustness of the experiment. PDBD2649, the lists of the chains used in each fold and their labelling into disordered/ordered regions, are available for download at the address `http://distill.ucd.ie/punch`.

2.2 Sequence inputs

Evolutionary information in the form of frequency profiles compiled from multiple alignments of homologous sequences is relied on in the tests. The alignments for the datasets described above are extracted from the NR database as available on March 3 2004, containing over 1.4 million sequences. The database is first redundancy reduced at a 98% threshold, leading to a final 1.05 million sequences. The alignments are generated by three runs of PSI-BLAST with parameters $b = 3000$ (maximum number of hits), $e = 10^{-3}$ (expectation of a random hit) and $h = 10^{-10}$ (expectation of a random hit for sequences used to generate the PSSM).

2.3 Structural inputs

Secondary structure and solvent accessibility are predicted respectively by Porter and PaleAle [28, 27], which both exploit homology to proteins of known structure, where available, in the form of simple structural frequency profiles extracted from sets of PDB templates. Relying on structural predictions of this kind can be seen as an indirect way of using information from PDB templates.

2.4 Template generation

For each of the proteins in PDBD2649, DISpro and the targets of CASP5 and CASP7, structural templates are searched for in the PDB. An obvious problem arising is that all proteins of these datasets are expected to be in PDB (barring

name changes), hence every protein will have a perfect template. To avoid this, every protein that appears in the above sets is expluded from the search. All entries shorter than 10 residues are also excluded, leading to a final 98962 chains for PDBD2649, 100131 for DISpro, 101454 for CASP5 and 101417 for CASP7.

To generate the actual templates for a protein, two rounds of PSI-BLAST are run against the version of the redundancy-reduced NR database described above, with parameters $b = 3000$, $e = 10^{-3}$ and $h = 10^{-10}$. A third round of PSI-BLAST is then run against the filtered PDB using the PSSM generated in the first two rounds. In this third round a high expectation parameter ($e = 10$) is deliberately used to include hits that are beyond the usual Comparative Modelling scope ($e < 0.01$, at the CASP6 competition [18]). From the set of templates for a query all those with sequence similarity exceeding 95% over the whole query are also removed, to exclude PDB resubmissions of the same structure at different resolution, other chains in N-mers and close homologues.

2.5 Predictive architecture

The tests in this article aim at verifying the contribution of predicted structural features and templates, rather than producing a state-of-the-art system, thus a minimal disorder predictor is implemented with no specialisation and voting scheme, no second stage prediction and post-processing. The predictor is implemented by an SVM with a linear kernel, meaning that the search space of the algorithm is directly in the feature space. As a consequence, fewer hyperparameters have to be optimised. This results in vast savings of computing time, an aspect that is particularly important given the large dataset considered here. With the linear kernel one can also learn about interesting aspects of the data, such as non-linear separation, thus giving the opportunity to evaluate potential noise introduced by some feature.

In order to take into account the local context of each amino acid, the input is built by considering sequence and/or structural features of the residues over a fixed-width window centered in the amino acid itself. The class predicted by the SVM is the sign of the decision value, which corresponds to the distance from the hyperplane separating predicted ordered and disordered residues.

2.5.1 Encoding sequence and template information

Input i_j associated with the j-th residue contains primary sequence and evolutionary information ($i_j^{(E)}$), predicted structural information (secondary structure and solvent accessibility) derived from PDB templates and direct structural information derived from disordered PDB templates ($i_j^{(T)}$):

$$i_j = (i_j^{(E)}, i_j^{(T)}) \tag{1}$$

where, assuming that e inputs are devoted to sequence and evolutionary information, and t to structural information:

$$i_j^{(E)} = (i_{j,1}^{(E)} \dots i_{j,e}^{(E)}), \tag{2}$$

and:

$$i_j^{(T)} = (i_{j,1}^{(T)} \dots i_{j,t}^{(T)}) \tag{3}$$

Hence i_j contains a total of $e + t$ components.

Evolutionary information for the j-th residue is encoded as the frequency profiles of the amino acids in a window of width w, and centered in residue j. The size of the sequence-based part of the input is set to $e = 25 * w$, meaning that the 20 amino acids are considered alongside 5 non standard symbols: B (aspartic acid or asparagine), U (selenocysteine), X (unknown), Z (glutamic acid or glutamine) and $\cdot$ (gap). A profile in the window is given by the frequency of each of the 24 non-gap symbols, plus the overall frequency of gaps in the corresponding column of the alignment. I.e., if n_{jk} is the total number of occurrences of symbol j in column k, and g_k the number of gaps in the same column, the jth profile component of column k is:

$$\frac{n_{jk}}{\sum_{v=1}^{24} n_{vk}} \tag{4}$$

for $j = 1...24$, while the 25th input is:

$$\frac{g_k}{g_k + \sum_{v=1}^{24} n_{vk}} \tag{5}$$

This input coding scheme is richer than simple 20-letter schemes and a similar approach has proven effective in [28] and [27].

For representing structural information, $t = 10 * w$. The first $3 * w$ inputs contain predicted secondary structure probabilities in three classes (H, E and C) for each residue in the window. Other $4 * w$ inputs are devoted to represent predicted solvent accessibility in four classes for the same residues. The remaining $3 * w$ structural inputs are used to represent disorder information from the templates, using 3 numbers for each residue in the window. For a residue in position j of the sequence, one number (d_j) is the average 2-class disorder composition in the aligned position of the PDB templates, while the last two (c_j and s_j) encode the average quality of the template column. Suppose $d_{p,j} \in \{0, 1\}$ represents disorder for the j-th residue in the p-th template. Then, if P is the total number of templates for a protein:

$$d_j = \frac{\sum_{p=1}^{P} w_p d_{p,j}}{\sum_{p=1}^{P} w_p}, \tag{6}$$

where w_p is the weight attributed to the p-th template. If the identity between template p and the query is id_p and the quality of a template (measured as X-ray resolution + R-factor/20, as in [14] - the lower the better) is q_p, then it is $w_p = id_p^3/q_p$. Taking the cube of the identity between template and query drastically reduces the contribution of low-similarity templates when good templates are available. For instance a 90% identity template is weighed two orders of magnitude more than a 20% one. Weighted average coverage and similarity of a column of the template profile are also encoded, as follows:

$$c_j = \frac{\sum_{p=1}^{P} w_p c_p}{\sum_{p=1}^{P} w_p}, \tag{7}$$

where c_p is the coverage of the sequence by template p (i.e. the fraction of non-gaps in the alignment), and:

$$s_j = \frac{\sum_{p=1}^{P} w_p id_p}{\sum_{p=1}^{P} w_p}. \tag{8}$$

It is worth noting how both disorder information from templates and the two indices of template quality above are residue-based. For this reason, the case in which only templates covering fragments of a protein exist does not pose a problem for the method; the residues not covered by templates will simply have the section of the input with template information blank, and predictions will be based only on the sequence and predicted structural features. Note how this differs from the approach described in [17], where prediction in this case is skipped. Here there is no need to employ different predictors for different types of input.

Finally, to encode both sequence and structural inputs of a window position beyond the N- and C-terminus, its corresponding value is simply set to -1 in the SVM feature vector. Note how alternative solutions may be devised, e.g. setting to 0 the feature value and placing an additional input to mark the position type (terminus or not), as in [17]. The scheme described requires fewer inputs and enforces the similarity between positions near the ends of the sequence, which typically show sequence patterns different from central positions. At the same time, it is possible that the approach used here could bias the classifier to almost always predict disorder for the extreme positions, so that alternative approaches (not tested here) are worth trying.

2.5.2 Training

SVM-light [20] is used to implement the disorder classifier. Training a support vector machine with a linear kernel mainly requires to adjust the parameter C. This controls the trade-off between expected generalisation performance and the

need to reduce the error on the training set. During preliminary experiments, samples of the training set were selected, and line search [30] in the space of C was performed to find its optimal value. The search frequently stalled when training with specific subsets of C, especially using high values. To perform cross validation on PDBD2649 and training on the DISpro dataset, the default C value computed by SVM-light was adopted, always set in the range $(0, 1)$.

Additional parameters to consider are the cost factor and the width of the window used in the input vector. In order to alleviate the problem of the unbalanced data set, the cost factor is set to ratio of negative (ordered) to positive (disordered) examples as found in the training set. This is a reasonable strategy that prevents to learn classifiers predicting the negative class almost exclusively. Similarly to the case of parameter C, preliminary experiments were run with diffent values of w (the width of the input window) and it was found that the performance of the classifier is fairly stable for $w \in [11, 19]$ for various combinations of sequence and structural inputs, with $w = 13$ resulting the best value. The results reported in the next section are all obtained by setting $w = 13$.

2.6 Performance measures

To estimate performances and compare the effect of various inputs and predictors, global accuracy (Q_2), the Matthews' correlation coefficient (MCC), area under the ROC curve (AUC), specificity (or precision, P) and sensitivity (or recall, R) [1] are used. Let t_p, f_p, t_n, f_n denote respectively the number of true positives, false positives, true negatives and false negatives. Q_2 is defined as $(t_p + t_n)/(t_p + f_p + t_n + f_n)$ and measures the fraction of correctly predicted residues irrespective of their class. Specificity, $P = t_p/(t_p + f_p)$, estimates the fraction of correctly predicted disordered residues, whereas sensitivity, $R = t_p/(t_p + f_n)$, refers to the fraction of actual disordered residues correctly identified. The Matthews' correlation coefficient [22] is given by:

$$MCC = \frac{t_p t_n - f_p f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}, \tag{9}$$

where MCC = +1 means perfect predictions, -1 totally incorrect predictions and 0 a prediction which is undistinguishable from random, on average. As the predictor normally outputs a probability or a decision value, the measures above can also be computed for some false positive hit rate (FPR), by setting the decision threshold to a value such that a specified percentage of ordered examples on a validation set are predicted above it. Setting the threshold at 5% is expected to limit the problem of disorder overprediction on unseen examples.

The above measures cannot account for the large disproportion between positive and negative examples observed in this binary classification task. The recommended CASP S-score [19] provides a measure of performance which rewards correctly predicted disordered residues rather than ordered ones, and is defined as:

				5% FPR		
	AUC	MCC	Q_2	P	R	S_w
Prof	0.89	0.49	92.8	45.8	61.3	0.55
Prof+(SS,SA)	0.92	0.54	93.2	48.2	67.5	0.62
Prof+Templates	0.92	0.55	93.4	49.0	69.9	0.64
Prof+(SS,SA)+Templates	0.93	0.57	93.6	50.1	72.8	0.67

Table 1: Comparison of cross validation results using different combinations of evolutionary information (Prof), predicted structural features (SS=secondary structure, SA=solvent accessibility) and disorder from weighted templates (Templates). The + sign between input types indicates that both inputs are used.

$$S_w = \frac{w_1 t_p - w_0 f_p + w_0 t_n - w_1 * f_n}{w_1 n_1 + w_0 * n_0}, \tag{10}$$

where n_0 (resp. n_1) is the number of ordered (resp. disordered) examples, w_0 and w_1 weight predictions on the corresponding classes. The CASP assessment usually set $w_0 = 6$ and $w_1 = 94$ [19, 3]. As for MCC, $S_w \in [-1, 1]$, with $S_w = 0$ indicating generalisation no better than random.

Finally, the area under the Receiver Operator Characteristics curve (AUC-ROC or AUC) [31] represents a global measure of predictor behaviour in terms of the trade-off of having a low fraction of false positives (i.e. by setting a high decision threshold) and a high fraction of disordered residues correctly recovered. These are obviously two conflicting goals.

3 Results and Discussion

3.1 Cross validation on the PDBD2649 dataset

Firstly the impact of sequence and structural features are tested by 5-fold cross validation on the PDBD2649 dataset. Table 1 reports a summary of the results using the performance measures described in the previous section and obtained using different combinations of evolutionary inputs, structural features predicted by Porter (secondary structure) and PaleAle (solvent accessibility), and weighted information from disordered templates. The row labelled "Prof" refers to the results of the ab initio predictor, i.e. not using direct (weigthed disorder profile) or indirect (secondary structure and solvent accessibility) information from homologous proteins, while the remaining rows refer to predictors employing different combinations of structural information. The values for accuracy, specificity, sensitivity and the S_w score are shown for FPR=5%.

All performance measures reported in Table 1 clearly indicate that structural information, in the form of predicted structural features or encoded into

templates or both, improves classification performances with respect to the use of sequence information alone. Noticeably, secondary structure and solvent accessibility predictions (second row in Table 1) yield large gains over sequence information alone, by all quality measures. According to the hypothesis in [42], this indicates that the structural features employed here encode information about disorder beyond that contained in profiles extracted from multiple sequence alignments. The third row shows the result obtained when using as inputs sequence information and weighted templates. The improvement over sequence alone is comparable to that achieved through secondary structure and solvent accessibility. When sequence information, secondary structure and solvent accessibility predictions and templates are combined (fourth row of Table 1) a further improvement is observed over the other cases. This suggests that the information carried by secondary structure and solvent accessibility predictions and that contained in the templates are at least partially independent.

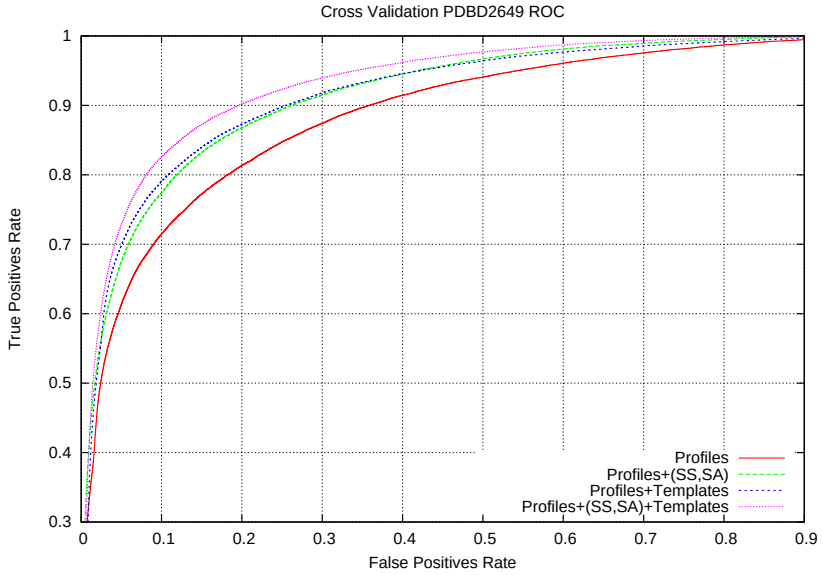

Figure 1: Comparison of Receiver Operator Characteristic curves obtained by cross validation on PDBD2649 and using different combinations of evolutionary information (Prof), predicted structural features (SS=secondary structure, SA=solvent accessibility) and disorder from weighted templates (Templates). The + sign between input types indicates that both inputs are used.

In Figure 1, the same combinations of features are compared using the ROC curve. Even in this case the predictor trained with all features is consistently

	AUC	MCC	Q_2 (5% FPR)
Punch	0.95	0.56	94.0%
DISpro	0.94	0.51	93.2%
DISOPRED2	0.90	0.52	93.1%
VL3	0.80	0.38	92.1%

Table 2: Comparison of methods on CASP5 targets. Values for: ROC-area under curve (AUC), Matthews correlation coefficient (MCC), accuracy (Q_2) at 5% false positive rate (FPR). Results for others methods reported by [6].

better than the other versions. The classifier using the sequence and templates is slightly better than the one using the sequence and structural features for smaller FPRs, and vice versa. This suggests that the template-based predictor overestimates disorder. This is probably due to the combination of: high confidence of the predictor when templates are available; balanced training (a higher cost is attached to the disorder class, which is 15 times smaller than its complement), which may lead to overprediction when the model is highly confident.

To further explore the relationship between performance and structural information from templates, in Figure 2 the histogram of the AUC is plotted as a function of sequence identity between the query and its best template, for different combinations of features. The histogram shows that the improvements observed when using all inputs are roughly idependent on the amount of detectable sequence identity between queries and templates. The figure reveals a trade-off between employing predicted secondary structure and solvent accessibility, and weighted disorder from templates. For identity values below 50%, the improvements of the classifier that uses all features are mainly (fully, below 20%) due to structural feature predictions from Porter and PaleAle. The low error rates of these predictors compensate for less informative inputs from low-quality or missing templates. For similarity values above 50%, the templates tend to weigh more, thus providing richer inputs which can be exploited by the predictor.

3.2 Training on the DISpro dataset

The predictor described here is also trained and tested on the DISpro dataset. The predictor using all input features (thereby termed Punch) is compared to the top-performing methods at CASP7. Note how this is fair, since this took place after the version of DISpro adopted here (dating back to 2004) was public.

First, Punch is compared with the DISpro method (eponymous of the database) and best methods from CASP5 (Table 2). In this case the comparison is fair with DISpro (which is trained on our same data set) but potentially slightly unfair on the other predictors which were necessarily trained on older data sets, as CASP5 took place in 2002. Nevertheless, using the measures reported in [6],

	AUC	(Q_2, MCC, P, R, S_w)$^{5\%}$	(Q_2, MCC, P, R, S_w)
Punch	0.86	(92.4, 0.41, 39.6, 51.0, 0.46)	(86.8, 0.36, 26.3, 66.3, 0.54)
ISTZORAN	0.86	(92.0, 0.37, 37.2, 45.8, 0.41)	(83.0, 0.33, 22.2, 72.1, 0.56)
CBRC-DR	0.84	(92.3, 0.41, 39.3, 50.6, 0.46)	(93.5, 0.41, 45.3, 44.4, 0.41)
Fais	0.85	(92.3, 0.40, 39.1, 50.1, 0.45)	(90.3, 0.38, 32.4, 56.7, 0.49)
DISOPRED	0.83	(91.9, 0.35, 35.6, 43.2, 0.38)	(92.1, 0.35, 36.7, 42.8, 0.38)

Table 3: Comparison of Punch with other methods on 96 CASP7 targets. Values for: area under curve ROC curve(AUC), Matthews correlation coefficient (MCC), accuracy, specificity, sensitivity and S-score at 5% FPR (Q_2, P, R, S_w)$^{5\%}$, (Q_2, P, R, S_w) same as before but using predictor decision threshold.

the approach described here outperforms DISpro, marginally on the AUC and Q_2 and more consistently on Matthews' correlation coefficient. This is likely to indicate a significant decrease of the number of false negatives (i.e. disordered residues incorrectly predicted as ordered) with respect to the other methods. Punch outperforms the other CASP5 methods by healthy margins and by all quality measures.

Punch is then compared with the top methods at CASP7 (ISTZORAN, CBRC-DR, Fais and DISOPRED). Figure 3 plots the comparison in terms of ROC curve, while Table 3 shows the summary of a detailed comparison based on AUC, and various performance indices evaluated using both the 5% false positives rate and the decision thresholds of each specific predictor. The ROC plots of Figure 3 show that Punch compared favourably with all top-ranked CASP methods across a wide range of FPR values. The result is confirmed by Table 3, where Punch shows the highest value of the AUC, together with ISTZORAN. The other performance indices indicate similar trends, especially at the low false positives rate. The results obtained considering the predictor decision threshold are also competitive (S_w). This results from a relatively high number of false positives as shown by the precision (i.e. incorrectly predicted ordered residues), which is compensated by a high fraction of actual disordered residues correctly recovered (recall), thus confirming the findings emerged from the CASP5 comparison. Figure 4 compares Punch with Fais and DISpro, two methods that seemed to benefit from knowledge of the structure of related proteins. Similarly to [3], the AUC values obtained on two target sub-sets are reported, the 59 targets with homologous PSI-BLAST templates (3D-homologous), and the 37 targets having no related protein with known structure (3D-nonhomologous). It is clear from Figure 4 that Punch benefits from structural information. Nevertheless, it outperforms the other predictors by roughly equal, clear margins on both the 3D-homologous and 3D-nonhomologous set.

Overall, in spite of its lack of specialisation and post-processing, Punch outperforms the best methods at CASP7. This is especially remarkable given its simplicity, and confirms that the structural features and templates Punch adopts are highly informative for disorder prediction.

4 Funding

Health Research Board of Ireland (RP/2005/219), Science Foundation Ireland (05/RFP/CMS0029).

5 Acknowledgments

We would like to thank Alberto Jesus Martin-Martin for helpful discussion and advice during the development of the experiments. And so much for thanking Jesus.

References

[1] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[2] P. Baldi, G. Pollastri, C. A. F. Andersen, and S. Brunak. Matching protein β-sheet partners by feedforward and recurrent neural networks. In *Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00)*, volume 8, pages 25–36, La Jolla, CA, 2000. AAAI Press.

[3] L. Bordoli, F. Kiefer, and T. Schwede. Assessment of disorder predictions in casp7. *Proteins*, 69(S8), 2007.

[4] J. W. Chen, P. Romero, V. N. Uversky, and A. K. Dunker. Conservation of intrinsic disorder in protein domain and families: I. a database of conserved predicted disordered regions. *Journal of Proteome Research*, 5:879–887, 2006.

[5] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, 1(113), 2007.

[6] J. Cheng, M. Sweredoski, and P. Baldi. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery*, 11(3):213–222, 2005.

[7] K. Coeytaux and A. Poupon. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, 21:1891–1900, 2005.

[8] J. A. Cuff and G. J. Barton. Application of multiple sequence alignments profiles to improve protein secondary structure prediction. *Proteins: Structure, Function and Genetics*, 40(3):502–511, 2000.

[9] N. Deshpande, K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, and Z. K. Feng et al. The rcsb protein data bank: a redesigned query system and relational database based on the mmcif schema. *Nucleic Acids Res*, 33:D233–D237, 2005.

[10] Z. Dosztanyi, V. Csizmok, P. Tompa, and I Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol*, 347:827–839, 2005.

[11] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering*, 14:835–839, 2001.

[12] F. Ferron, S. Longhi, B. Canard, D., and Karlin. A practical overview of protein disorder prediction methods. *Proteins*, 65:1–14, 2006.

[13] S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi. Poodle-l: a two-level svm prediction system for reliably predicting long disordered regions. *Bioinformatics*, 23(16):2046–2053, 2007.

[14] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci*, 3:522–24, 1994.

[15] N. Holladay, L. N. Kinch, and N. V. Grishin. Optimization of linear disorder predictors yields tight association between crystallographic disorder and hydrophobicity. *Protein Science*, 16(21):2140–2152, 2007.

[16] R.W.W. Hooft, C. Sander, M. Scharf, and G. Vriend. The pdbfinder database: a summary of pdb, dssp and hssp information with added value. *Bioinformatics*, 12(6):525–529, 1996.

[17] T. Ishida and K. Kinoshita. Prdos: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*, 35(Suppl 2):W460–W464, 2007.

[18] J. Moult J, K. Fidelis, A. Tramontano, B. Rost, and T. Hubbard. Critical asssessment of methods of protein structure prediction (casp)-round vi. *Proteins*, 61(Suppl 6):3–7, 2005.

[19] Y. Jin and R. L. Dunbrack. Assessment of disorder predictions in casp6. *Proteins*, 61:167–175, 2005.

[20] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT Press, 1999.

[21] D. T. Jones and J. J. Ward. Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, 53:573–578, 2003.

[22] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.

[23] C. Mooney and Pollastri. Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins*, 77(1):181–90, 2009.

[24] M. Mucchielli-Giorgi, S. Hazout, and P. Tuffery. Predacc: prediction of solvent accessibility. *Bioinformatics*, 15(2):176–177, 1999.

[25] K. Peng, P. Radivojac, S. Vucetic, K. Dunker, and Z. Obradovic. Lenght-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7(208), 2006.

[26] G. Pollastri, P. Fariselli, R. Casadio, and P. Baldi. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 47:142–235, 2002.

[27] G. Pollastri, A. J. M. Martin, C. Mooney, and A. Vullo. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, 8(201), 2007.

[28] G. Pollastri and A. McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–20, 2005.

[29] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47:228–235, 2002.

[30] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical recipes in C: the art of scientific computing*. Cambridge Univeristy Press, third edition, 2008.

[31] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42:203–231, 2001.

[32] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, and V. N. Uversky. Intrinsic disorder and functional proteomics. *Biophysical Journal*, 92:1439–1456, 2007.

[33] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, 232:584–599, 1993.

[34] K. Shimizu, S. Hirose, and T. Noguchi. Poodle-s: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics*, 23(17):2337–2338, 2007.

[35] D. Stoffer and L. G. Volkert. A neural network for predicting protein disorder using amino acid hydropathy values. In *IEEE symposium on computational intelligence in bioinformatics and computational biology*, pages 482–490, San Diego, CA, 2005.

[36] C. Su, C. Chen, and Y. Ou. Protein disorder prediction by condensed pssm considering propensity for order or disorder. *BMC Bioinformatics*, 7(319), 2006.

[37] V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are natively unfolded proteins unstructured under physiologic conditions? *Proteins*, 41:415–427, 2000.

[38] A. Vullo, O. Bortolami, G. Pollastri, and S. Tosatto. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.*, 34:W164–168, 2006. Web Server Issue.

[39] I. Walsh, D. Baù, A.J.M. Martin, C. Mooney, A. Vullo, and G. Pollastri. Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. *BMC structural biology*, 9(1):5, 2009.

[40] I. Walsh, A.J.M. Martin, C. Mooney, E. Rubagotti, A. Vullo, and G. Pollastri. Ab initio and homology based prediction of protein domains by recursive neural networks. *BMC bioinformatics*, 10(1):195, 2009.

[41] J. J. Ward, L. J. McGuffin, K. Bryson, , B. F. Buxton, and D. T. Jones. The disopred server for the prediction of protein disorder. *Bioinformatics*, 20:2138–2139, 2004.

[42] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol*, 337:635–645, 2004.

[43] E. A. Weathers, M. E. Paulaitisa, T. B. Woolf, and J. H. Hoh. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Letters*, 576(3):348–352, 2004.

[44] Z. R. Yang, R. Thomson, P. MCNeil, and R. M. Esnouf. Ronn: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, 21:3369–3376, 2005.

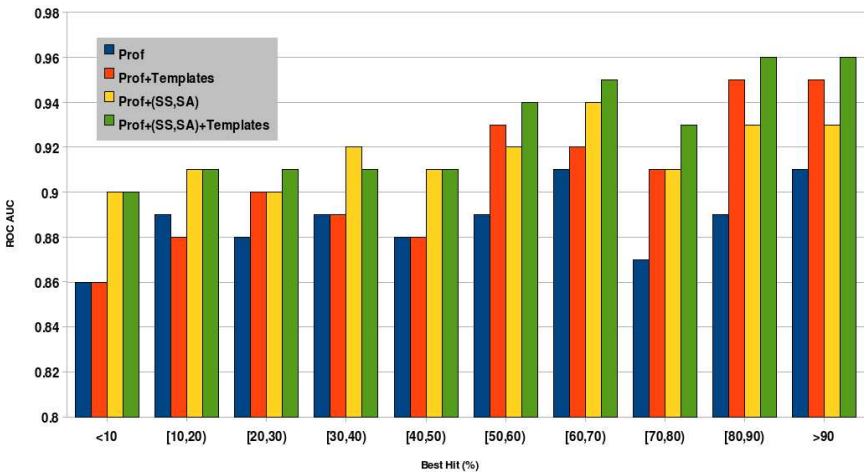

Figure 2: Distribution of the area under the ROC curve as a function of sequence similarity to the best hit in PSI-BLAST templates found using the procedure explained in the "Template generation" subsection. Values compared are those obtained by using different combinations of evolutionary information (Prof), predicted structural features (SS=secondary structure, SA=solvent accessibility) and disorder from weighted templates (Templates). The + sign between input types indicates that both inputs are used.

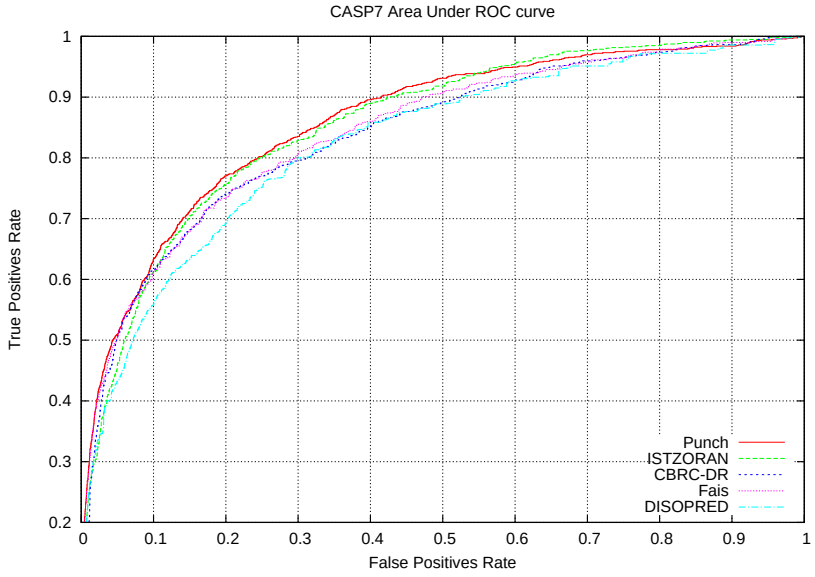

Figure 3: Comparison of ROC curves of different methods on 96 CASP7 targets. Punch is trained on the DISpro dataset [6] and with inputs given by evolutionary information, structural features and weighted disorder templates (Prof+(SS,SA)+Templates).

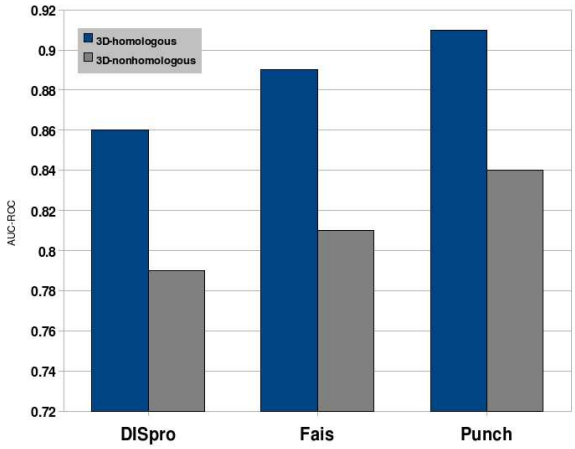

Figure 4: Histogram of area under ROC curve of three methods on two CASP7 target sub-sets: 59 targets with homologous PSI-BLAST templates (3D-homologous) and 37 targets having no related protein with known structure (3D-nonhomologous).