BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 4 2012, pages 503–509
doi:10.1093/bioinformatics/btr682

Structural bioinformatics
Advance Access publication December 20, 2011

ESpritz: accurate and fast prediction of protein disorder

Ian Walsh, Alberto J. M. Martin, Tomàs Di Domenico and Silvio C. E. Tosatto*
Department of Biology, University of Padua, Viale G. Colombo 3, I-35131 Padova, Italy

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Intrinsically disordered regions are key for the function of numerous proteins, and the scant available experimental annotations suggest the existence of different disorder flavors. While efficient predictions are required to annotate entire genomes, most existing methods require sequence profiles for disorder prediction, making them cumbersome for high-throughput applications.

Results: In this work, we present an ensemble of protein disorder predictors called ESpritz. These are based on bidirectional recursive neural networks and trained on three different flavors of disorder, including a novel NMR flexibility predictor. ESpritz can produce fast and accurate sequence-only predictions, annotating entire genomes in the order of hours on a single processor core. Alternatively, a slower but slightly more accurate ESpritz variant using sequence profiles can be used for applications requiring maximum performance. Two levels of prediction confidence allow either to maximize reasonable disorder detection or to limit expected false positives to 5%. ESpritz performs consistently well on the recent CASP9 data, reaching a S_w measure of 54.82 and area under the receiver operator curve of 0.856. The fast predictor is four orders of magnitude faster and remains better than most publicly available CASP9 methods, making it ideal for genomic scale predictions.

Conclusions: ESpritz predicts three flavors of disorder at two distinct false positive rates, either with a fast or slower and slightly more accurate approach. Given its state-of-the-art performance, it can be especially useful for high-throughput applications.

Availability: Both a web server for high-throughput analysis and a Linux executable version of ESpritz are available from: http://protein.bio.unipd.it/espritz/

Contact: silvio.tosatto@unipd.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 4, 2011; revised on November 24, 2011; accepted on December 3, 2011

1 INTRODUCTION

Protein function has been traditionally thought to be determined by tertiary structure. More recently, an alternative view is emerging with respect to non-folding regions, which suggests a reassessment of the structure-to-function paradigm (Dunker and Obradovic, 2001; Dunker *et al.*, 2008; Schlessinger *et al.*, 2011; Wright and Dyson, 1999). Flexible segments lacking a unique native structure within a protein are known as disordered regions (Tompa, 2002). Disorder has been shown to be widespread within known natural proteins, especially in eukaryotic organisms (Dunker *et al.*, 2000, 2008; Schlessinger *et al.*, 2011). It also plays a key role in human disease (Uversky *et al.*, 2008) where it is thought that 79% of all cancer-associated proteins are at least in part unstructured/disordered (Dunker *et al.*, 2008). Proteins with disordered segments are frequently associated with molecular recognition (Tompa and Fuxreiter, 2008; Tompa *et al.*, 2009). They have also been observed to be common among hub proteins, i.e. those with a large number of interaction partners (Dosztanyi *et al.*, 2006). In addition, protein disorder is also important for protein expression, purification and crystallization since difficulties often arise when long disordered regions are present, as happens frequently at the N and C termini.

Protein disorder is experimentally determined with an assortment of indirect biochemical methods collected in the DisProt database (Sickmeier *et al.*, 2007) currently containing ∼640 proteins. Alternatively, missing residues in X-ray crystallographic structures from ∼70 000 structures deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2007) can be used. The analysis of ∼6000 nuclear magnetic resonance (NMR) ensembles from the PDB is also possible (Martin *et al.*, 2010), but to the best of our knowledge, has never been used to train a prediction method. It is assumed that different flavors of protein disorder exist (Dunker *et al.*, 2008). The most common distinction is between long (DisProt) and short (X-ray) segments (Schlessinger *et al.*, 2011). Alternatively, there has also been an attempt to distinguish flavors based on enrichment for certain amino acid types (Vucetic *et al.*, 2003). The characteristically skewed amino acid distribution of disordered segments, lacking in hydrophobic and enriched in polar and charged residues (Uversky *et al.*, 2000), can be easily exploited for sequence-based predictions. Available prediction methods can be broadly divided into three classes. Biophysical methods (Dosztanyi *et al.*, 2005; Galzitskaya *et al.*, 2006; Lobanov and Galzitskaya, 2011; Obradovic *et al.*, 2005; Prilusky *et al.*, 2005; Uversky, 2002) exploit the sequence distribution to derive pseudo-energy propensities to adopt a disordered state. Of these, IUPred (Dosztanyi *et al.*, 2005) is probably the most widely used due to its availability and efficiency, as it does not require multiple sequence alignments. Machine learning techniques have been widely used for the prediction of protein disorder (Cheng *et al.*, 2005; Hirose *et al.*, 2007; Ishida and Kinoshita, 2007; Linding *et al.*, 2003; McGuffin, 2008; Vullo *et al.*, 2006; Ward *et al.*, 2004; Yang *et al.*, 2005). In most cases, PSI-BLAST sequence profiles (Altschul *et al.*, 1997) are combined with additional features, e.g. predicted secondary structure in the widely used Disopred (Ward *et al.*, 2004). On average, these methods are slower but somewhat more accurate than biophysical predictors. The last, and most recent, category of disorder predictors use a consensus of various biophysical and machine learning methods (Mizianty *et al.*, 2010; Schlessinger *et al.*, 2009; Walsh *et al.*, 2011;

*To whom correspondence should be addressed.

Xue *et al.*, 2010). Here, a further improvement in accuracy is obtained at the cost of running several predictors in parallel and averaging their output.

Among the applications of disorder prediction, we can distinguish at least two different scenarios. The first is represented by the Critical Assessment of techniques for protein Structure Prediction (CASP) experiment, where the methods are used to predict disorder on a relatively small number of proteins with maximum accuracy (Noivirt-Brik *et al.*, 2009). Here, clearly consensus predictors aiming for maximum accuracy should excel. However, a more practical scenario is represented by high-throughput analysis of protein disorder, e.g. on entire genomes (Schlessinger *et al.*, 2011). In this case, the focus is shifted toward fast predictors producing a minimum number of false positives (Sirota *et al.*, 2010). Over the years, most prediction methods have addressed the first problem, with comparatively little attention to the practicalities of large-scale predictions. This has led to a relative paucity of accurate fast predictors, as adding more prediction layers has produced slightly more accurate but increasingly cumbersome methods (e.g. Mizianty *et al.*, 2010; Schlessinger *et al.*, 2009; Walsh *et al.*, 2011; Xue *et al.*, 2010).

Here we present the ESpritz methods for determining disorder based solely on sequence, aimed for high-throughput applications. The predictor is tested on large datasets of different disorder types, including a novel NMR mobility definition. The sole input for the ESpritz method is the amino acid. It does not require sliding windows to capture local contextual information or any complex sources of information and is shown to be both state-of-the-art and efficient.

2 METHODS

ESpritz uses bidirectional recurrent neural networks (BRNN) (Baldi *et al.*, 1999) to predict disorder from sequence information. A BRNN can be likened to an ensemble of three neural networks, learning the N-terminal sequence context, the sequence and the C-terminal sequence context, respectively. Where regular neural networks use a sliding window of predetermined size, BRNNs learn this context information through the recursive dynamics of the network, reducing the number of parameters and extracting information implicitly from the surrounding local context. Another important feature is a top layer filter which takes as input ‘semi-global’ information from the bottom layer in analogy to Pollastri and McLysaght (2005). Parameter learning proceeds by gradient descent and the back-propagation algorithm and the output contains two units producing the probability of order and disorder. The total number of parameters depends on the number of neuronal units in the various network layers. ESpritz never exceeds 5886 which is acceptable and very unlikely to overfit considering the amount of training examples which is between 30 and 100 times the number of parameters (Supplementary Table S1). BRNNs with a similar number of parameters have already been applied to various prediction problems, e.g. secondary structure (Pollastri and McLysaght, 2005). In the following, we introduce the four basic variants and their consensus (Table 1) before describing the datasets used and evaluation measure.

Table 1. Definition of Spritz variants and acronyms used

Acronym	Sequence	Profile	Consensus
ESpritz	Both	Yes	Four-way
ESpritzP	Both	Yes	Two-way
ESpritzS	Both		Two-way
ASpritzP	Attributes	Yes	
ASpritzS	Attributes		
SSpritzP	Identities	Yes	
SSpritzS	Identities		

Definitions for the Spritz variants. Sequence relates to the input information with attributes (five Atchley scales) or identities (20 residue types). Two-way consensus is calculated for the two sequence coding schemes with and without profile. Four-way consensus is calculated among all four basic variants.

2.1 Sequence-only predictions

In analogy to our work on repeat proteins (Marsella *et al.*, 2009), ASpritz uses the five Atchley sequence metrics (Atchley *et al.*, 2005) as numerical sequence attributes for BRNN input. Each scale, listed in Table 2 of Atchley *et al.* (2005), was obtained by clustering almost 500 different amino acid scales from the AAindex database (Kawashima *et al.*, 1999). The scales were shown to reflect polarity, secondary structure, molecular volume, codon diversity and electrostatic charge (Atchley *et al.*, 2005) and may allow for a richer amino acid representation. As the five scales have different, asymmetric, ranges they require normalization in order to be useful as neural network inputs. As in our previous work, normalization is performed so that the squares of the scales sum to 1 (Marsella *et al.*, 2009):

$$\sum_{t=1}^{20} \left[A_t(X)\right]^2 = 1 \quad (t=1,2,\ldots.5) \tag{1}$$

where $X = [A, C, D, E, \ldots, W, Y]$ is the one letter code corresponding to each of the 20 amino acids, and $A_t(X)$ is the sequence metric for amino acid X. ASpritz has five inputs i to the neural network for each sequence position k, each representing one normalized Atchley scale. If position k in the sequence contains amino acid X then the five inputs to this system are as follows:

$$i_k^t = A_t(X) \quad (t=1,2,\ldots.5) \tag{2}$$

Alternatively, SSpritz considers the 20 amino acids in ‘one-hot’ encoding. It consists of 20 inputs i where each unit for sequence position k is allocated for 1 of the 20 amino acids:

$$i_k^{1-20} = R_k(X) \tag{3}$$

where $X$is the residue at position$k$,$R_k(X) \quad \in R^{20}$ is an alphabetically ordered vector of positions R_k^j corresponding to the 20 amino acids (i.e. $[A, C, D, E, \ldots, W, Y]$). R_k^j= 1 if the position amino acid is in the sequence at position k and R_k^j= 0 otherwise. ASpritz and SSpritz are combined into a consensus score ESpritzS. As previously shown for CSpritz (Walsh *et al.*, 2011), simply averaging the two scores proved most effective (data not shown).

2.2 Multiple sequence alignment-based methods

Evolutionary information in the form of multiple sequence alignments is commonly used to improve predictor performance. Here, the two sequence encodings are extended to accommodate sequence profiles. Let a sequence profile $p_k(X)$ give the probability of finding amino acid X in the multiple sequence alignment at position k along the sequence (gaps not considered). For the Atchley scales, the profile-based predictor (ASpritzP) contains six inputs i for sequence position k, one for each scale plus gaps:

$$i_k^t = \sum_{X \in C_k(X)} A_t(X) p_k(X) \quad (t=1,\ldots.5) \quad i_k^6 = \frac{g}{n+l} \tag{4}$$

where $C_k(X)$ is the set of amino acids for position k, g is the number of gaps, n is the number of non-gaps and l is the total number of sequences involved in the multiple sequence alignment. Alternatively, when considering the 20 amino acids, the sequence profile $p_k(X)$ is multiplied against the input vector from SSpritz:

$$i_k^{1-20} = \sum_{X \in C_k(X)} R_k(X) p_k(X) \tag{5}$$

The fraction of gaps in the sequence profile is included as an additional input (i_k^{21}) and the system is called SSpritzP. Averaging the ASpritzP and SSpritzP output into a consensus prediction will be termed ESpritzP, while ESpritz is the ensemble combination of all four single predictors by averaging their probabilities (Table 1). In order to assess the effect of using the BRNN architecture, we also train a standard feed-forward neural network (NN) using 'one-hot' encoding [Equation (3)] and a fixed window size of 23 residues (6484 parameters) found to be the best combination on the X-ray training set.

2.3 Datasets

To train and measure the performance of the predictors, we created several datasets from structures deposited at the PDB (Berman *et al.*, 2007) and from experimental data as deposited in the Disprot database (Sickmeier *et al.*, 2007). Training and testing data are strictly separated, with appropriate separation and redundancy reduction (maximum ∼25% for X-ray and NMR, 40% for DisProt) to present truly unseen data during testing. Unless stated otherwise, all alignments were calculated using PSI-Blast (Altschul *et al.*, 1997) using options -b 3000 -e 0.001 -h 1e-10. The alignment sequence database for PSI-Blast was non-redundant (NR) at a 90% sequence identity level. Low complexity sequences, transmembrane helices and coiled-coil regions were filtered from the sequence database using Pfilt (Jones and Swindells, 2002). All new datasets used for training and testing are available for download from URL: http://protein.bio.unipd.it/espritz/.

2.3.1 X-ray disorder The X-ray training set was constructed from crystallographic structures deposited in the PDB until May 1, 2008, restricted to X-ray protein chains of length between 25 and 2000 amino acids, with resolution at most 2.5 Å and R-factor up to 25%. Disordered residues are defined as those with missing backbone C-alpha atoms. All proteins were classified into those containing at least three consecutive disordered amino acids and those with no disordered regions. Both subsets were sorted by decreasing quality and reduced by sequence identity using UniqueProt (Mika and Rost, 2003) to an HSSP value of 0 (∼25% over 100 aligned residues) giving priority to proteins with better quality (-m option). The resulting lists were merged and redundancy reduced in a similar manner leaving proteins with disordered regions as a priority. The training set contains 3244 proteins with 660 120 residues of which 5.68% are disordered. The test set was created using the same procedure for proteins released by the PDB between May 1, 2008 and September 13, 2010. The test set contains 569 proteins with 94 520 residues of which 7.34% are disordered.

2.3.2 DisProt disorder The training set is based on DisProt version 3.7 (January 28, 2008) in order to ensure that we have sufficient testing data, i.e. proteins annotated between 2008 and 2010. It contains 484 proteins with 219 424 residues where we consider 25.71% disordered. Here we define a residue as disordered if the DisProt curators consider the residue to be disordered at least once while all other residues (including unannotated) are considered structured. Since many residues are unannotated in DisProt, and this could be a potential source of bias in testing, we extend the coverage of the test set by annotating DisProt version 5.7 with PDB structures in analogy to the work of Sirota *et al.* (2010). Briefly, all sequences from DisProt 5.7 with >40% sequence identity to the training set are removed. The remainder was matched to PDB entries through the UniProt accession code from DisProt and linked through the SIFTS database (Velankar *et al.*, 2005). Entries were annotated for disorder considering DisProt definitions where available. Unannotated residues are deemed structured if they exist in both the SEQRES and ATOM sections, and as disordered for regions of length at least five residues missing from the latter. The DisProt and PDB sequences were then aligned to take into account possible variations, and PDB disorder annotations transferred to the DisProt sequence with at least 95% sequence identity. The new test set contains 52 proteins where 49.72% of the residues are unannotated, 41.04% are disordered and 9.24% are ordered for a total of 18 096 residues.

2.3.3 NMR mobility NMR mobility datasets are calculated using the Mobi server (Martin *et al.*, 2010). Mobi is based on a simple algorithm to find regions with different conformations among all models in an NMR ensemble. Briefly put, residues with a variation of atomic coordinates and torsion angles between models above a fixed threshold are marked as mobile. The threshold was optimized to replicate the NMR disorder definition used in CASP8 (Noivirt-Brik *et al.*, 2009). The extraction and redundancy reduction is identical to the X-ray datasets (see above) except that PDB NMR structures are considered (no quality filter). The training set consists of 2187 proteins with 173 154 residues of which 16.90% are considered disordered. The testing set contains 671 proteins with 59 384 residues and 18.70% disorder.

2.3.4 Other datasets The MxD dataset (Mizianty *et al.*, 2010), sharing <25% sequence identity with CASP8, was downloaded from the website at URL: http://biomine-ws.ece.ualberta.ca/MxD.txt. Note that the 5-fold cross-validation used here might differ from the one used in the paper. The *Homo sapiens* protein sequences were downloaded from URL: ftp://ftp.ncbi.nih.gov/genomes/. The total number of proteins as of September 2010 for the human genome was 39 151. The time comparison was calculated for 1% of the human genome (i.e. 391 proteins).

2.4 Comparison with available methods

ESpritz is compared with several other methods which were either downloaded (Disopred, MULTICOM, DisEMBL, IUpred) or used as web server (PONDR-FIT). The original Spritz method (Vullo *et al.*, 2006) and our recently published improvement CSpritz (Walsh *et al.*, 2011) are also shown for comparison. In all cases, the methods were used with default parameters. Multiple sequence alignments for Disopred and ESpritz were calculated on the 90% reduction of the May 2008 non-NR database and preprocessed with the pfilt program. MULTICOM alignments were calculated using an internal database. The CASP9 data was downloaded from the official website (URL: http://predictioncenter.org/casp9/). Note that, in contrast to our previous paper (Walsh *et al.*, 2011), 252 residues marked as 'x' in CASP9 were not considered in the analysis and disordered segments of 1 or 2 residues are considered. ESpritz is available both as a web server and as a pre-compiled executable for Linux machines from URL: http://protein.bio.unipd.it/espritz/.

2.5 Measuring performance

The assessment of our predictions use similar measures as used in CASP8 and previous CASPs (Noivirt-Brik *et al.*, 2009). There are two types of measures. Binary measures are calculated once the probability decision threshold is found. All our disorder probability thresholds were found on the corresponding training sets. We define the binary measures sensitivity (Sens = TP/N_{dis}), specificity (Spec=TN/N_{ord}), selectivity (Sel=TP/TP+FP), F- measure [$F = 2\times$ Sen $\times$ Sel/(Sen + Sel)], Matthews correlation coefficient (MCC), accuracy [Acc=(Sens+Spec)/2] and the score (S_w=Sens+Spec-1) (Lobanov *et al.*, 2010). *TP*, *TN*, *FN* and *FP* are the number of true positives, true negatives, false negatives and false positives, respectively (positive is disorder, negative is order). N_{dis}, N_{ord} are the number of disorder and ordered residues, respectively. We also use area under the receiver operator curve (AUC), calculated between false positive rate (FPR = 1 - specificity; *x* axis) and true positive rate (TPR = sensitivity; *y* axis), as a measure of the quality of the probabilities. As in CASP8 (Noivirt-Brik *et al.*, 2009), the statistical significance of the evaluation scores was determined by bootstrapping (Supplementary Material): 80% of the targets were randomly selected 1000 times, and the standard error of the scores was calculated (i.e. 1.96 × standard_error gives 95% confidence around mean for normal distributions).

3 RESULTS

3.1 Training and consensus building

The different Spritz variants (Table 1) have been trained on pre-CASP8 data from all three flavors of protein disorder (X-ray, DisProt, NMR). A comparison between training and test set performance can be found in Supplementary Table S1, which also show how the training examples is 30–100 times the number of parameters for each predictor. The consistent performance on independent training and test sets is an indication of the good generalization capability of ESpritz. Table 2 and Supplementary Table S2 show the results for the X-ray disorder definition in comparison with various other methods and Table 3 and Supplementary Table S3 for analogous data on DisProt. From the data, it is apparent that the Spritz variants present high specificity and clearly outperform the methods used for comparison. The difference is more pronounced for the DisProt dataset, probably because it contains longer disordered segments on which fewer methods have been trained. This appears to confirm the hypothesis that long (i.e. DisProt) and short (i.e. X-ray) disorder are different flavors (Schlessinger *et al.*, 2011). As expected, the profile-based predictors perform slightly better than the sequence-only ones, although the latter are still competitive. The effect of the BRNN architecture becomes apparent in comparison to the standard neural network (NN), which performs significantly worse. It is also interesting to note that the differences between sequence encoding schemes appear minimal. Each Spritz variant remains, nevertheless, competitive against other state-of-the-art methods such as IUPred and DisoPred. These results are also verified using 5-fold cross-validation on an independent set provided by the MxD database (Mizianty *et al.*, 2010) (Supplementary Table S4).

Table 2. Benchmark results on X-ray disorder test set

Predictor	Sens	Spec	S_w	AUC
CSpritz	79.63	85.05	***64.68***	*0.8993*
SSpritzP	76.48	87.02	***63.50***	*0.8893*
ESpritz	77.23	85.63	***62.85***	***0.8912***
ESpritzP	77.49	85.29	***62.77***	*0.8876*
MULTICOM	81.99	80.37	*62.37*	*0.8879*
ASpritzP	76.06	84.81	*60.86*	*0.8766*
ESpritzS	73.67	86.23	*59.89*	*0.8748*
SSpritz	73.98	85.39	*59.37*	0.8699
ASpritz	73.03	86.23	*59.25*	0.8721
NN (w=23)	69.39	87.74	57.12	0.8645
PONDR-FIT	69.20	86.73	55.92	0.8609
Disopred	56.48	93.87	50.33	0.8391
IUPred (short)	54.00	94.95	48.92	0.8475
Spritz (old)	41.63	93.09	34.69	0.7884
DisEMBL465	31.91	97.67	29.61	0.8320

Performance measured on X-ray disorder for 569 structures. Methods performing at least as well as or not statistically different from ESpritz are highlighted in bold. Methods performing at least as well as or not statistically different from ESpritzS, our best fast predictor, are in italics and underlined.

Once the performance has been established, the question becomes whether the same disorder information is detected to a different degree or slightly distinct signals are picked up by the Spritz variants. This information could then be used to create a consensus predictor. Figure 1 shows how the different methods represent somewhat different predictions and an implicit confidence estimate. Whenever the four variants agree, as they do for ~80% of all residues, the accuracy is close to 100% for order and ~40% for disorder (see Supplementary Table S5 for Pearson's correlation coefficients). The relative rarity of intermediate cases should allow a simple averaging of the probabilities (ESpritz) to outperform each individual method, as shown in Table 2. In order to maintain the efficiency of the sequence-only variants, the two partial combinations between SSpritz/ASpritz (ESpritzS) and SSpritzP/ASpritzP (ESpritzP) are also shown. In the following, for

Table 3. Benchmark results on enhanced DisProt disorder test set

Predictor	Sens	Spec	S_w	AUC
ESpritz	77.51	80.37	***70.58***	***0.892***
ESpritzS	73.78	93.66	***67.44***	***0.901***
ESpritzP	75.47	91.69	*67.15*	0.888
PONDR-FIT	68.89	93.18	*62.08*	0.885
IUPred (long)	61.57	96.83	*58.4*	0.878
Disopred	64.19	93.9	*58.54*	0.824
CSpritz	79.07	78.02	57.09	0.877
MULTICOM	77.35	78.89	56.23	0.853
NN (w=23)	69.05	82.66	51.71	0.815
IUPred (short)	49.17	97.61	46.77	0.855
Spritz (old)	81.74	59.86	41.6	0.770
DisEMBL465	32.51	98.03	30.53	0.792
DisEMBL	46.74	82.89	29.63	0.692

Performance on enhanced Disprot disorder for 52 structures. Methods performing at least as well as or not statistically different from ESpritz are highlighted in bold. Methods performing at least as well as or not statistically different from ESpritzS, our best fast predictor, are in italics and underlined.

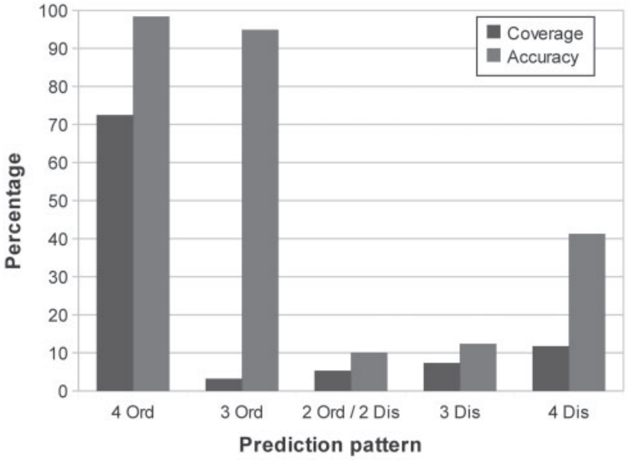

Fig. 1. Agreement between the four Spritz variants. The relative frequency (coverage) of each state distribution for the four predictors is plotted together with the accuracy for that case.

Table 4. Benchmark results on NMR disorder test set

Predictor	Sens	Spec	S_w	AUC
ESpritzP	72.83	79.19	***52.01***	***0.8366***
ESpritz	72.53	79.33	***51.85***	***0.8401***
ESpritzS	66.94	80.77	*47.71*	*0.8179*
CSpritz	71.93	74.74	46.67	0.7964
MULTICOM	75.14	69.55	44.69	0.7976
PONDR-FIT	63.74	75.55	39.29	0.7533
Disopred	48.69	89.19	37.88	0.7556
IUPred (short)	45.07	90.73	35.79	0.7505
NN (w=23)	51.73	80.27	31.99	0.7301
Spritz (old)	29.29	96.49	26.28	0.7481
DisEMBL HL	25.38	82.08	7.46	0.6329

Performance on NMR disorder for 671 structures. Methods performing at least as well as or not statistically different from ESpritz are highlighted in bold. Methods performing at least as well as or not statistically different from ESpritzS, our best fast predictor, are in italics and underlined. IUpred long performs worse than IUPred short and is not shown. DisEMBL HL is the hot loops predictor which performs better than DisEMBL465 on this dataset.

Table 5. Benchmark results on CASP9 targets ranked by S_w

Predictor	Sens	Spec	S_w	AUC
ESpritz	67.41	87.52	***54.82***	***0.8558***
PRDOS2(291)	60.78	90.03	*50.65*	*0.8544*
CSpritz	63.66	86.37	*49.91*	*0.8316*
Multicom-refine(119)	64.98	85.02	*49.89*	0.8217
Biomine(351)	59.63	89.01	*48.48*	0.8213
ESpritzS	59.75	88.83	*48.43*	*0.8308*
GSMETADISORDERMD(374)	65.72	81.93	*47.57*	0.8184
MASON (193)	53.70	92.76	*46.25*	0.7438

Performance on 117 CASP9 targets (19 NMRs and 98 X-rays). The top five performing CASP9 groups are shown with their official group name and number in brackets. Methods performing at least as well as or not statistically different from ESpritz are highlighted in bold. Methods performing at least as well as or not statistically different from ESpritzS, our best fast predictor, are in italics and underlined. Note that group 351 was missing 10 proteins.

simplicity we will only show results for the ESpritz variants. Full data is available in the Supplementary Material.

3.2 Novel NMR mobility flavor

A unique feature of ESpritz is the explicit prediction of NMR mobility through a dedicated predictor. To the best of our knowledge, no other method has been developed to predict disorder defined on mobile residues in NMR structural ensembles, although this has been benchmarked since CASP8 in 2008 (Noivirt-Brik *et al.*, 2009). One of the problems was the unique automatic definition of NMR mobility, which we have recently addressed (Martin *et al.*, 2010). As can be seen from the results shown in Table 4 (and Supplementary Table S6), ESpritz has a strong performance and even the sequence-based predictors outperform existing methods. NMR mobility appears to harbor a distinct signal that is somewhere between short (X-ray) and long (DisProt) disorder. ESpritz is particularly useful to detect this novel flavor of disorder, although the specificity values remain below those of other variants. The latter may be speculatively attributed to the variability of NMR structures, which combine greater structural flexibility than crystal structures with a wider range of experimental conditions. In general, the NMR flavor appears to predict more disorder than the X-ray and DisProt ones, with segments of length somewhere between the other two. Supplementary Figure S1 shows as example ESpritz predictions for the human p53 protein using the three different flavors.

3.3 Comparison on CASP9 data

In order to fully compare our method to the state-of-the-art, we use data from the recent CASP9 experiment. Table 5 and Supplementary Table S7 show the results for all targets, while Supplementary Table S8 shows only the NMR targets. ESpritz is significantly more accurate than all methods using both the S_w and AUC criteria. This strong performance can be partially explained by the use of a dedicated NMR prediction mode. Perhaps not unexpectedly, the ESpritz variants excel on NMR targets thanks to the novel NMR prediction mode, where they outperform the best CASP9 methods by at least 15% on S_w and 7% on AUC (Supplementary Table S8). ESpritz also outperforms our recent consensus-based CSpritz method (Walsh *et al.*, 2011), which combines three different predictors including a preliminary version of SSpritz but lacks an explicit NMR mode.

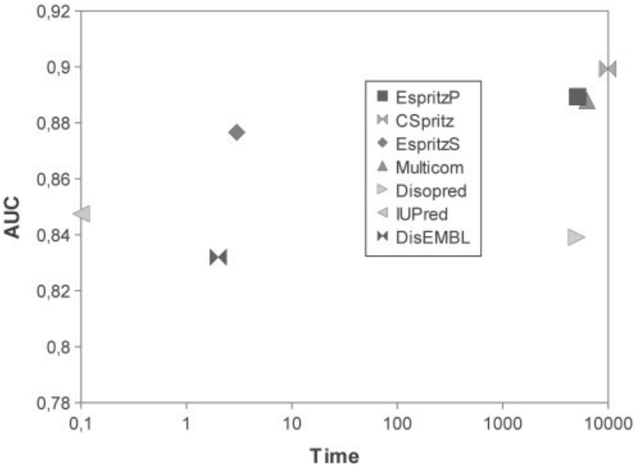

Fig. 2. Time versus performance plot for different predictors. The time in minutes for pedicting 1% of the human genome on a single Intel Xeon processor core is plotted against the AUC for each locally installed method. Note that the time axis uses a logarithmic scale.

3.4 Large-scale predictions

The large-scale analysis across entire genomes is an important application of disorder predictors (Schlessinger *et al.*, 2011; Sirota *et al.*, 2010; Ward *et al.*, 2004), both to further our understanding of disorder as a biological phenomenon and to help establishing protein function. The efficiency in terms of CPU time versus AUC of different methods on a randomly selected 1% of the human genome is shown in Figure 2. As can be seen, the field is divided between fast methods with somewhat lower accuracy and much slower methods using multiple sequence alignment information from PSI-BLAST. The latter improve AUC by up to four percentage points at the cost of four orders of magnitude of computation time. ESpritzS

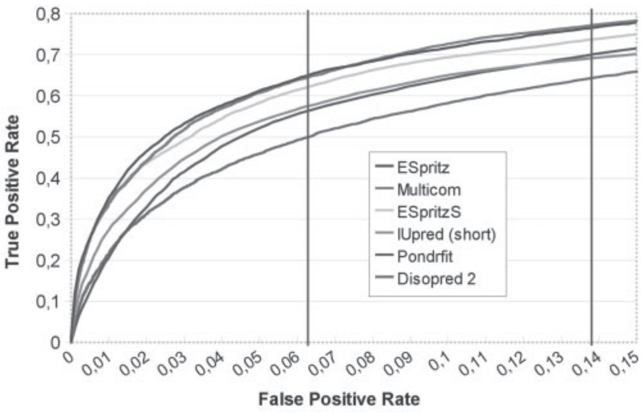

Fig. 3. Receiver-operating characteristic curve for X-ray test set data. The plot shows the FPR in the region from 0% to 15% false positives for various methods. The two vertical lines represent the ESpritz decision thresholds corresponding to a predicted 5% FPR (left) and the optimal S_w threshold (right).

combines the best of both worlds, by maximizing performance for a fast method that does not require multiple sequence alignments.

When analyzing large numbers of sequences, it can be especially useful to be able to limit the number of expected false positives to avoid drawing false conclusions on the prevalence of disorder (Sirota *et al.*, 2010). Figure 3 shows a typical receiver-operating characteristic curve plot on the X-ray dataset. As can be seen, ESpritzS is particularly good at low FPRs up to around 5% FPR, after which the relative FPR increases. The optimal binary S_w decision threshold can be found around 13.5% FPR. An alternative 5% expected FPR threshold was derived on the training dataset for all ESpritz variants. On the testing data this yields ∼63% sensitivity and 45% selectivity at ∼6.5% FPR. The more stringent decision threshold provides a simple way to limit the number of false positives at the expense of somewhat lower sensitivity. At this low FPR threshold, the ESpritz variants perform better than the other tested methods using similar decision thresholds. For the full analysis on all datasets, including F-measure and MCC values, please refer to the Supplementary Material. The 5% expected FPR threshold should prove useful for high-throughput applications requiring low FPRs. It is, therefore, expected that ESpritzS can provide a valid alternative in applications requiring the high-throughput analysis of thousands of sequences or entire genomes.

4 CONCLUSIONS

We have presented a new ensemble of disorder predictors, called ESpritz, having state-of-the-art performance on three different flavors of disorder. Compared with our previous methods Spritz (Vullo *et al.*, 2006) and CSpritz (Walsh *et al.*, 2011), ESpritz combines a more sophisticated BRNN architecture with enhanced definitions of disorder flavors. The BRNN improves performance slightly on X-ray data but substantially on the other two disorder datasets. The comparatively larger improvement on DisProt data may be related to our usage of an enhanced re-annotation of ordered segments in DisProt (Sirota *et al.*, 2010), providing a clearer distinction between the two states. Unsurprisingly, where ESpritz really excels is on NMR mobility. This is a novel definition which, to the best of our knowledge, was never incorporated before in a disorder predictor. Our comparison with existing methods, and the strong performance on CASP9 data, suggest that NMR flexibility is encoded by a somewhat different but related signal to the other two flavors. The NMR flavor appears to capture a larger fraction of amino acids at the borderline between the ordered and disordered states, perhaps at the expense of more false positive predictions. Nevertheless, the differences between disorder datasets support the hypothesis of different flavors being encoded by somewhat different sequence features as suggested by Schlessinger *et al.* (2011). The second major improvement in ESpritz is the creation of a sequence-only predictor which is four orders of magnitude faster than multiple sequence alignment-based methods at the expense of a slight reduction in accuracy. This allows the user to choose between highly accurate predictions for single proteins or high-throughput predictions at genomic scale. The third, and final, improvement in ESpritz is the definition of an alternative, more stringent, disorder threshold limiting the expected FPR to 5%. This allows the user to choose between detection of more disorder or highly selective predictions depending on the data being analyzed. The very high specificity of ESpritz also ensures a low rate of false positives on high-throughput problems, making it even more valuable for this task. This scenario is typically overlooked when developing disorder prediction methods, but accounts for a large part of the biological problems to be addressed. We believe that ESpritz offers an accurate and efficient way to address many biologically relevant problems encountered with disordered proteins.

ACKNOWLEDGEMENTS

The authors are grateful to an anonymous referee for many useful suggestions and to members of the BioComputing UP lab for insightful discussions.

Funding: University of Padova grants (CPDA098382, CPDR097328); FIRB Futuro in Ricerca grant (RBFR08ZSXY) to SC.E.T.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Atchley,W.R. *et al.* (2005) Solving the protein sequence metric problem. *Proc. Natl Acad. Sci. USA*, **102**, 6395–6400.

Baldi,P. *et al.* (1999) Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937–946.

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.

Cheng,J. *et al.* (2005) Accurate prediction of protein disordered regions by mining protein structure data. *Data Min. Knowl. Discov.*, **11**, 213–222.

Dosztanyi,Z. *et al.* (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Dosztanyi,Z. *et al.* (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.*, **5**, 2985–2995.

Dunker,A.K. and Obradovic,Z. (2001) The protein trinity–linking function and disorder. *Nat. Biotechnol.*, **19**, 805–806.

Dunker,A.K. *et al.* (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 161–171.

Dunker,A.K. *et al.* (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.

Galzitskaya,O.V. *et al.* (2006) Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput. Biol.*, **2**, e177.

Hirose,S. *et al.* (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, **23**, 2046–2053.

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

Jones,D.T. and Swindells,M.B. (2002) Getting the most from PSI-BLAST. *Trends Biochem. Sci.*, **27**, 161–164.

Kawashima,S. *et al.* (1999) AAindex: Amino Acid Index Database. *Nucleic Acids Res.*, **27**, 368–369.

Linding,R. *et al.* (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.

Lobanov,M.Y. and Galzitskaya,O.V. (2011) The Ising model for prediction of disordered residues from protein sequence alone. *Phys. Biol.*, **8**, 035004.

Lobanov,M.Y. *et al.* (2010) Library of disordered patterns in 3D protein structures. *PLoS Comput. Biol.*, **6**, e1000958.

Marsella,L. *et al.* (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*, **25**, i289–i295.

Martin,A.J. *et al.* (2010) MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. *Bioinformatics*, **26**, 2916–2917.

McGuffin,L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, **24**, 1798–1804.

Mika,S. and Rost,B. (2003) UniqueProt: Creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Mizianty,M.J. *et al.* (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.

Noivirt-Brik,O. *et al.* (2009) Assessment of disorder predictions in CASP8. *Proteins*, **77** (Suppl. 9), 210–216.

Obradovic,Z. *et al.* (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61** (Suppl. 7), 176–182.

Pollastri,G. and McLysaght,A. (2005) Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21**, 1719–1720.

Prilusky,J. *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.

Schlessinger,A. *et al.* (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **4**, e4433.

Schlessinger,A. *et al.* (2011) Protein disorder–a breakthrough invention of evolution? *Curr. Opin. Struct. Biol.*, **21**, 412–418.

Sickmeier,M. *et al.* (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, **35**, D786–D793.

Sirota,F.L. *et al.* (2010) Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC Genomics*, **11** (Suppl. 1), S15.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.

Tompa,P. and Fuxreiter,M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.*, **33**, 2–8.

Tompa,P. *et al.* (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, **31**, 328–335.

Uversky,V.N. (2002) What does it mean to be natively unfolded? *Eur. J. Biochem.*, **269**, 2–12.

Uversky,V.N. *et al.* (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.

Uversky,V.N. *et al.* (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, **37**, 215–246.

Velankar,S. *et al.* (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.

Vucetic,S. *et al.* (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.

Vullo,A. *et al.* (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.*, **34**, W164–W168.

Walsh,I. *et al.* (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res.*, **39**, W190–W196.

Ward, J.J. *et al.* (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.

Xue,B. *et al.* (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*, **1804**, 996–1010.

Yang,Z.R. *et al.* (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.