Journal of Bioinformatics and Computational Biology
Vol. 11, No. 2 (2013) 1250023 (12 pages)

DOI: 10.1142/S0219720012500230

Imperial College Press
www.icpress.co.uk

TUNING THE PRECISION OF PREDICTORS TO REDUCE OVERESTIMATION OF PROTEIN DISORDER OVER LARGE DATASETS

ANTONIO DEIANA*,‡ and ANDREA GIANSANTI*,†,§

*Physics Department, Sapienza University of Rome, Rome, Italy

†INFN, Sezione di Roma1, Roma, 00185, Italy

‡Antonio.Deiana@Roma1.infn.it

§Andrea.Giansanti@roma1.infn.it

Received 27 February 2012
Revised 13 August 2012
Accepted 20 September 2012
Published 28 November 2012

This is a study on the precision of four known protein disorder predictors, ranked among the best-performing ones: DISOPRED2, PONDR VSL2B, IUPred and ESpritz. We address here the problem of a systematic overestimation of the number of disordered proteins recognized through the use of these predictors, considered as a standard. Some of these predictors, used with their default setting, have a low precision, implying a tendency to overestimate the occurrence of disordered proteins in genome-wide surveys. Moreover, different predictors often disagree on the evaluation of individual proteins. To cope with this problem and in order to propose a simple procedure that enhances precision based on precision-recall curves, we re-tuned the discriminative thresholds of the predictors by training and cross-validating their performance on a cured dataset. After re-tuning, both the disagreement among predictors and the tendency to overestimate the occurrence of disordered proteins are reduced. This is shown in a dedicated study over the human proteome and a set of cancer-related human proteins, with no *a priori* disorder annotation. Simple quantitative estimates suggest that the occurrence of disorder among cancer-related proteins and other similar large-scale surveys has been overestimated in the past.

Keywords: Intrinsically disordered proteins; disorder predictors; precision recall curves.

1. Introduction

The growing interest in intrinsically disordered proteins (IDPs) is a new and potentially very relevant tendency in the recent protein science.[1,2] IDPs lack a stable three-dimensional structure, globally or in short or long segments of their chain. Different from structured proteins, IDPs can interact with many targets and therefore fulfill important roles in numerous cellular processes such as signal transduction, transcriptional regulation, and translation.[3] Moreover, IDPs are thought to have a key role in several human diseases,[4,5] including cancer,[6] cardiovascular

diseases,[7] neurodegenerative[8] and genetic diseases,[9] and in the formation of amyloidotic fibrils in misfolding diseases.[10] Therefore, considerable enthusiasm has arisen in the study of these proteins, both experimentally[2,11] and through bioinformatics methods, particularly through the action of a few very active groups.[11–15]

To study IDPs, it is important, first of all, to find them in large databases, such as proteomes or interactomes. Great efforts have been devoted to this task, that led to the development of many predictors of protein disorder (reviews in Refs. 12–15) and to their use to scan large databases (reviews in Ref. 2). Disorder predictors aim at identifying unfolded segments in polypeptide chains. Generally, they are trained to dichotomically recognize residues as belonging to structured or unstructured polypeptide segments (named ordered and disordered residues, respectively), based on amino acid composition and several physical–chemical properties (as, for example, hydrophobicity, charge, packing). In this paper we consider four disorder predictors: DISOPRED2,[16] PONDR VSL2B,[17,18] IUPred,[19] and ESpritz.[20] With the exception of ESpritz, they have been widely used to extract IDPs from large sets of proteins. We also include the recent ESpritz since it is trained on different variants of disorder, is fast, and has a good performance.

It has been shown, both by the authors of the predictors and in independent assessments of the CASP experiments,[21–23] that these methods have a high rate of true predictions, i.e. in a large sets of residues, they effectively recognize disordered residues out of ordered ones. However, when the predictors are used to recognize IDPs (based on the presence of long segments of disordered residues), there are remarkable discrepancies. Different predictors reveal different and, in some cases, largely different occurrences of IDPs in the same dataset.[2] Moreover, different predictors quite often do not agree in the exact identification of the boundaries of disordered domains within protein sequences.[12,13]

The aim of this paper is to clearly assess the problem of the inconsistency of the predictions returned by predictors and to propose a re-tuning of their settings to partially address this problem. In particular, we investigate whether the inconsistency can be due to an overestimate of the number of disordered proteins, as indicated by a low selectivity, or precision. First, we re-assessed the performance of the predictors in identifying IDPs on a nonredundant set containing both well-structured proteins selected from the Protein Data Bank (PDB)[24] and disordered proteins from the DisProt database.[25] PONDR VSL2B and ESpritz have low precision, less than 35% of the predictions obtained are correct. This indicates that they can overestimate the number of IDPs in a dataset by classifying many ordered proteins as disordered.

To limit the potential overestimate of disorder, we changed the parameters of the predictors, so to set their precision to be quite close to their sensitivity, on the basis of precision-recall curves. As we discuss below, in this way the number of proteins predicted as disordered becomes quite similar to the number of IDPs actually present in the dataset, and the overestimate of disorder can be controlled.

After this re-tuning, all predictors have sensitivity and precision close to or higher than 0.6 and the rate of false positives is lower than 0.09. The percentage of IDPs in

Homo sapiens (HS) ranges from 26% to 61% before our re-tuning procedure. After re-tuning, we found that the percentage of IDPs is less than 45% and the disagreement between predictors was reduced. In *Homo sapiens* cancer-related proteins (HSCPs), re-tuned predictors found less than 54%.

In conclusion, our study points out that some widely used predictors tend to overestimate the number of disordered proteins, due to their systematic low precision. The re-tuning of the settings to increase precision decreases in general the disagreement among predictors and can limit the overestimation of the number of IDPs in large sets of proteins.

2. Methods

There is no golden standard for the assignment of disorder to a given region of a protein sequence. Protein disorder manifests itself under different experimental signatures and prediction methods depend on the particular flavor of disorder they are trained over.[26,27] Here, a protein is considered as intrinsically disordered if more than 30% of its residues (either predicted or experimentally found, or annotated) are disordered. This criterion finds an interesting validation in a study of the functional regulation of IDPs in eukaryotes.[28] A simple index to express the disagreement between two predictors in estimating the percentage of IDPs in a dataset is given by the ratio:

$$\Delta = (n_d^{\mathrm{A/B}} + n_d^{\mathrm{B/A}})/n_{\mathrm{TOT}}$$

where $n_d^{\mathrm{A/B}}$ is the number of proteins predicted as disordered by predictor A and as ordered by predictor B; conversely $n_d^{\mathrm{B/A}}$ is the number of proteins predicted as disordered by predictor B and as ordered by predictor A, and n_{TOT} is the total number of proteins in the dataset. The Δ value ranges between 0 (perfect agreement) and 1 (total disagreement).

2.1. *Sets of proteins*

We performed our analysis on a nonredundant set of 864 structured proteins from PDB[24] and 132 IDPs from DisProt database.[25] In the following, we call this dataset as ProtSel. The structured proteins we selected contain less than 30% of disordered residues. They were selected from PDBSelect25, version February 2010, a nonredundant set of proteins from the PDB, with less than 25% of sequence identity.[29–31] From PDBSelect25, we filtered out complex proteins (i.e. no "COMPLEX" nor "COMPLEXED" term in the PDB record) and retained only structures with a resolution lower than 2 Å, an R-factor lower than 20%, no X character in their sequences and less than 30% of disordered residues. Operationally, a residue is disordered (e.g. missing, unresolved) if it is present in the SEQRES field but not in the ATOM field of the PDB files. IDPs were extracted from DisProt database, version 1.57.[25] We selected all proteins with more than 30% annotated disordered residues.

We excluded proteins with segments lacking either ordered or disordered annotation. ProtSel is available online (ftp://aglab.phys.uniroma1.it/pub/databases/ProtSel.txt).

For the case studies, we considered two protein sets: the *Homo sapiens* proteome (HS) and the *Homo sapiens* cancer-associated proteins (HSCP). HS contains 20,236 proteins selected from the SwissProt database, version January 2011.[32] HSCP contains 3,176 proteins, selected by searching SwissProt with keywords: tumor, oncogene, anti-oncogene and proto-oncogene. In HS and HSPC no *a priori* signature of disorder is known.

2.2. *Predictors of protein disorder*

Three of the four predictors we consider have been widely used to select out IDPs from large sets of proteins. Just as an indication, let us quote the number of citations found on the ISI Web of Knowledge (http://apps.webofknowledge.com/): 408 for the PONDR family methods; 501 for DISOPRED2; 508 for IUPred. The methods of the PONDR VSL2 family[17,18] (we use the VSL2B version) and DISOPRED2[16] are support vector machines trained to recognize disordered residues from the amino acid composition of the region of the polypeptide chain in which they are embedded. PONDR VSL2B is trained on 1,327 proteins selected both from PDB and DisProt. In the training set, there are both proteins with long disordered segments experimentally identified (>30 disordered amino acids) and proteins with short segments (<30 disordered amino acids). DISOPRED2 is trained on 7,169 structured proteins from the PDB, with less than 95% of sequence similarity. IUPred makes use of a pairwise energy function among residues in a protein and it is based on the empirical observation that known disordered residues have higher total energy than ordered ones. It is trained on 785 structures proteins from PDB, with less than 25% of sequence similarity.[19]

We also consider ESpritz,[20] a recently published predictor that has shown quite promising performances. ESpritz is based on a bi-directional recursive neural network, trained on the flavors of disorder emerging both from crystallographic and NMR structures.

2.3. *Training and testing procedure*

In this paper, we considered a protein as intrinsically disordered if more than 30% of its residues are disordered. To identify disordered residues, we used the scores returned by the predictors. If the score is higher than a fixed discriminative threshold, then a residue is predicted as disordered. We verified that the precision of some predictors in identifying IDPs is low if one uses the default discriminative thresholds, indicated by the authors of the methods. So, we tried to re-evaluate these thresholds to increase the performance of the predictors.

To obtain the new threshold and test the resulting performance of the predictors, we used a five-fold cross-validation procedure.[33] ProtSel was partitioned into five

different subsets. At each cross-validation step, four subsets were combined in a training set, and the remaining subset was used to test predictors.

2.4. *Performance measures*

The performance of disorder predictors as dichotomic classifiers is usually assessed on their ability to identify disordered residues in a test set. In this paper, however, predictors are used to identify IDPs in datasets, not disordered residues. Therefore, we tested the performance of predictors in finding IDPs in our selection of proteins. Let N_d and N_o be number of disordered and ordered proteins effectively present in a set, and n_d and n_o the number of predicted disordered and ordered proteins respectively, returned by a given predictor. Clearly, $n_o + n_d = N_o + N_d$, but in general $n_d \neq N_d$ and $n_o \neq N_o$. In an ideal predictor, n_d and n_o coincide with N_d and N_o, respectively. But in general, this is not the case, and the relative performance of a predictor is evaluated by computing several ratios between correct and incorrect predictions. The first step is to compute the following quantities: TP, number of disordered proteins predicted as disordered (true positives); FN, number of disordered proteins predicted as ordered (false negatives); TN, number of ordered proteins predicted as ordered (true negatives); FP, number of ordered proteins predicted as disordered (false positives). Then the following indexes are evaluated[21–23]:

$$\text{Sensitivity (or recall)} : S_n = \frac{TP}{TP + FN} = \frac{TP}{N_d} \tag{1}$$

is the number of correctly identified disordered proteins normalized to the total number of disordered proteins in the sample

$$\text{Specificity} : S_p = \frac{TN}{TN + FP} = \frac{TN}{N_o} \tag{2}$$

is the ratio between the number correctly identified ordered proteins and the total number of ordered proteins in the sample;

$$\text{Rate of false positives} : f_p = \frac{FP}{TN + FP} = 1 - S_p \tag{3}$$

is the ratio between the number of ordered proteins predicted as disordered and the total number of ordered proteins in the sample;

$$\text{Accuracy} : ACC = \frac{S_n + S_p}{2} \tag{4}$$

that is the average between sensitivity and specificity. It measures the overall performance of the predictor. Then,

$$\text{Precision (or selectivity)} : \text{Pr} = \frac{TP}{TP + FP} = \frac{TP}{n_d} \tag{5}$$

is the ratio between the number of correctly predicted disordered proteins and the total number of proteins predicted as disordered in the sample.

To evaluate if a predictor either overestimates or underestimates the number of IDPs in a dataset, we used the index n_d/N_d. It is easy to verify that

$$n_d/N_d = S_n/Pr. \tag{6}$$

Clearly, a low precision enhances this index and it can indicate an overestimate of the number of IDPs identified in a dataset.

3. Results

3.1. *Performance of disorder predictors*

Initially, we tested predictors with their default thresholds on ProtSel. DISOPRED2 and IUPred had the highest precision. PONDR VSL2B and ESpritz had low precision, not exceeding 0.35 (Table 1). Their ratio n_d/N_d was higher than 2.4, indicating that they predicted as disordered more than twice the real number of IDPs in the dataset. Therefore, VSL2B and ESpritz seriously overestimated the frequency of IDPs in this dataset.

To address this problem, we changed the discriminative thresholds used to identify disordered residues in protein sequences (see Sec. 2.3) so to tune the precision of predictors to be close to the sensitivity ($S_n/Pr \sim 1$). In this way, the number of predicted IDPs is about equal to the number of IDPs present in the dataset [$n_d/N_d \sim 1$, see Eq. (6)] and the overestimation of disorder is kept under control. To select the thresholds, we evaluated precision-recall (PR) curves, in which sensitivity (recall) is plotted against precision (selectivity), for different thresholds. PR curves should be preferred to the generally used receiving operating characteristics (ROC) curves, since, in skew datasets as those considered in the present paper (ProtSel contains 87% structured and 13% disordered proteins), ROC curves are biased toward a low rate of false positives, as is well known.[34]

Table 1. Performances of widely used disorder predictors in recognizing IDPs over the ProtSel dataset, sorted by decreasing precision, before and after retuning.

	S_N	S_P	f_P	ACC	Pr	n_d/N_d
			Before retuning			
DISOPRED2	0.68	0.94	0.06	0.81	0.66	1.03
IUPred	0.58	0.95	0.05	0.76	0.62	0.94
ESpritz	0.79	0.75	0.25	0.77	0.33	2.39
PONDR VSL2B	0.88	0.68	0.32	0.78	0.30	2.90
			After retuning			
DISOPRED2	0.66	0.94	0.06	0.80	0.66	1.00
PONDR VSL2B	0.62	0.94	0.06	0.78	0.65	1.00
IUPred	0.59	0.91	0.09	0.75	0.56	1.00
ESpritz	0.59	0.92	0.08	0.76	0.58	1.00

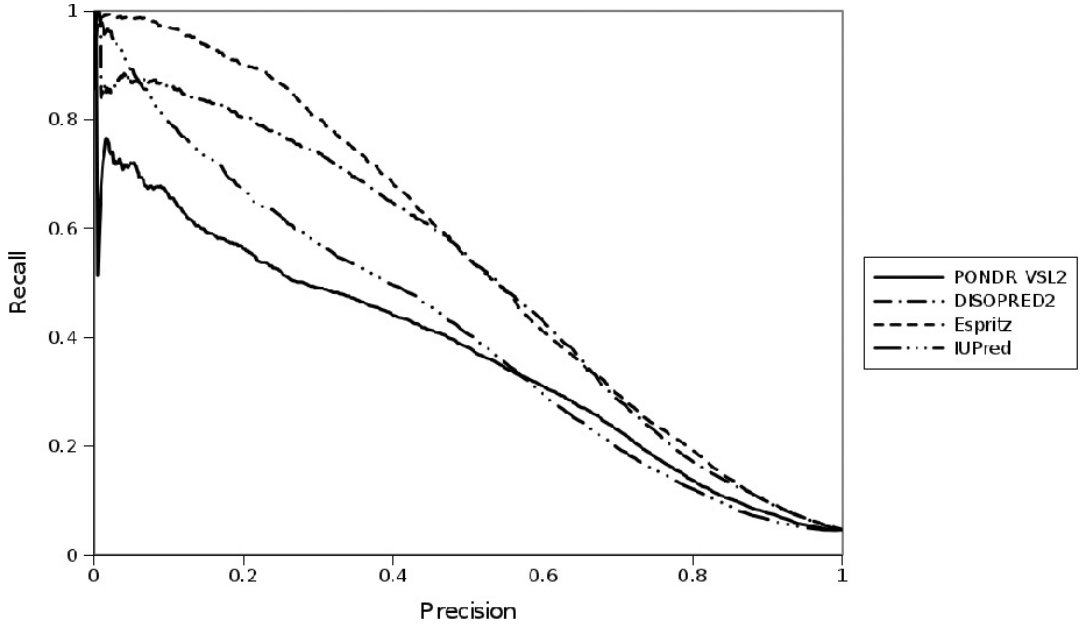

Fig. 1. Precision-recall curves.

Table 2. Thresholds to discriminate disordered residues from ordered ones.

Predictor	Default thresholds	Retuned thresholds
DISOPRED2	0.05	0.051
PONDR VSL2B	0.5	0.692
ESpritz	0.063	0.122
IUPred	0.5	0.485

The average discriminative thresholds that guarantee in each predictor a precision close to the sensitivity varied considerably, and they are reported in Table 2.

With these thresholds all predictors displayed similar sensitivity and precision, close to 0.6 (see Table 1).

3.2. *How many disordered proteins in the human genome?*

As case studies, we evaluated the occurrence of IDPs in the two datasets HS and HSCP (*Homo sapiens* proteome from Swiss-Prot and human proteins associated with cancer, respectively), before and after re-tuning of the thresholds, and the disagreement among the predictors, evaluated through the disagreement index Δ, defined in the first paragraph of Methods (Tables 3 and 4).

The first observation is that predictors gave quite different estimates for the number of IDPs in both HS_SP and HSCP before re-tuning, with default thresholds (Table 3). The VSL2B estimate in HS (61%) was significantly higher than those of DISOPRED2 (46%), ESpritz (42%) and IUPred (26%). Since VSL2B and ESpritz

Table 3. Number of intrinsically disordered proteins in HS (human proteome) and in HSPC (human cancer-associated proteins) predicted by PONDR VSL2B, DISOPRED2, ESpritz and IUPred, before and after re-tuning. In parentheses relative percentages.

	Before re-tuning	After re-tuning	Before re-tuning	After re-tuning
PONDR VSL2B	12,274 (61%)	7,351 (36%)	2,154 (51%)	1,382 (44%)
DISOPRED2	9,362 (46%)	9,141 (45%)	1,769 (56%)	1,732 (54%)
ESpritz	8,529 (42%)	5,863 (29%)	1,552 (37%)	1,118 (35%)
IUPred	5,346 (26%)	5,726 (28%)	1,022 (32%)	1,098 (34%)
	HS		HSCP	

Table 4. Disagreement index between PONDR VSL2B, ESpritz, IUPred and DISOPRED2 in HS (human proteome) and in HSPC (human cancer-associated proteins), before and after the re-tuning procedure.

HS	VSL2B	ESpritz	IUPred	DISOPRED2	HSCP	VSL2B	ESpritz	IUPred	DISOPRED2
				Before re-tuning					
VSL2B	0	0.35	0.34	0.19	VSL2B	0	0.21	0.36	0.17
ESpritz		0	0.09	0.23	ESpritz		0	0.18	0.17
IUPred			0	0.22	IUPred			0	0.24
				After re-tuning					
VSL2B	0	0.13	0.11	0.16	VSL2B	0	0.14	0.12	0.16
ESpritz		0	0.09	0.21	ESpritz		0	0.10	0.23
IUPred			0	0.16	IUPred			0	0.22

have a lower precision, their estimates can well be biased toward excess. After re-tuning, both VSL2B and ESpritz showed a decrease in the number of identified IDPs, and the percentages of IDPs, respectively, identified by VSL2B and ESpritz were similar, close to 30%. DISOPRED2, which has a default precision of 0.66 and a ratio n_d/N_d close to 1, is less affected by the retuning procedure, as expected. Also in the HSCP dataset, VSL2B and ESpritz showed, after retuning, a reduction of predicted IDPs, however the frequency of IDPs is higher in HSCP than in HS proteins (see Table 3). From Table 3, we can conclude that the frequency of IDPs does not exceed 45% in HS and 54% in HSCP.

The second observation is that, re-tuning the precision of the two predictors generally decreases the disagreement index between them (Table 4), with the exception of DISOPRED2 versus ESpritz in the HSCP dataset. This result clearly indicates that finding a good compromise between sensitivity and precision improves the agreement of predictors over single proteins.

4. Discussion and Conclusions

The problem raised in the present paper originated by the observation that disorder predictors are generally tested and used in two quite different kind of contexts: (i) they are tested to search for short, rare structural disorder, as in the CASP

experiments; (ii) they are used to search for IDPs in large databases. However, we have shown that a predictor with its default settings, tuned to have an optimum sensitivity and specificity in context (i), can overestimate disorder in context (ii) if its n_d/N_d ratio is higher than 1. So, when one wants to use disorder predictors for genome-wide, large scale surveys it is important to fine-tune their performances to get a good compromise between sensitivity and precision that allows to keep the overestimate of the number of IDPs under control and, possibly, also the overestimate of disordered residues at a reasonable level. In this paper, after having shown that some predictors in their default settings tend to have low precision in identifying IDPs, we have proposed a re-tuning of the predictor settings so to enhance their precision and obtain a number of predicted IDPs reasonably similar to the number of IDPs effectively present in the dataset.

In the case studies, we have shown that the predictors returned quite different occurrences of IDPs among human proteins, when used with default settings. The re-tuning procedure generally reduces the disagreement among predictors, as indicated by the disagreement index (Table 4). The percentage of putative IDPs found by PONDR VSL2 and ESpritz in HS significantly decreases after re-tuning, and it is lower than 45%. Also in HSCP, we observed a decrease in the disagreement index, in particular between PONDR VSL2, ESpritz and IUPred, and a slight reduction in the percentage of IDPs found by the predictors (less than 54%).

It has been reported that 79% of cancer-associated proteins in *Homo sapiens* are intrinsically disordered.[6] That estimate was based on the use of PONDR VL-XT and on a different operational definition of a disordered protein, as containing at least one long (>30 residue) disordered segment. We checked that, if one uses this criterion, the number of IDPs predicted by the predictors with default settings is remarkably higher than when one adopts the criterion we follow here, based on the occurrence of at least 30% of disordered residues (compare Table 5 with Table 3). We believe that our criterion is sound because the presence of a segment of 30 residues in proteins of about 600 residues (average length of human proteins) is less significant than the presence of at least 30% of disordered residues, i.e. about 180 disordered residues, to classify that protein as disordered.

Table 5. Number of proteins with al least one long disordered segment (>30 residues) in HS (human proteome) and in HSPC (human cancer-associated proteins) predicted by PONDR VSL2B, DISOPRED2, ESpritz and IUPred, with default settings.

	HS	HSCP
PONDR VSL2	15,481 (77%)	2,638 (83%)
DISOPRED2	13,244 (65%)	2,364 (74%)
ESpritz	10,888 (54%)	2,341 (74%)
IUPred	10,027 (50%)	1,904 (60%)

Our upper-bound estimate of 54%, based on the DISOPRED2, indicates that the occurrence of IDPs in cancer-associated proteins had been overestimated in the past.

As a general conclusion, we believe to have shown that the use of disorder predictors in large scale, genome-wide surveys, should be complemented by a preliminary analysis of their precision over reliable experimental test sets, such as the ProtSel used here. In this way, we can limit the overestimation of disorder observed in some predictors. Nevertheless, looking at Table 3, it is evident that, even after our proposed retuning, the estimates in the number of predicted IDPs vary considerably among different predictors, and the understanding of this variance is a major issue in the field. Genomic estimates on how large is the unfoldome should consider the observation by Orengo and Thornton that it is possible to assign about two thirds of the sequences from completed genomes to as few as 1,400 domain families for which structures are known.[35] Since among the sequences that are hard to structurally classify with family domains there are membrane proteins (folded even if hard to crystallize) and structural singletons, this observation should be used to tune the output of disorder predictors. Also a detailed and still missing study of how much of the predicted disorder is covered by protein domain databases would be very relevant.

Acknowledgments

This paper was funded by CASPUR (Inter-University Consortium for the Application of Super-Computing for Universities and Research, via dei Tizii 6b 00185 Rome, Italy) HPC grant 2010.

References

1. Uversky VN, Dunker AK, Understanding protein non-folding, *Biochimica et Biophysica Acta* **1804**:1231−1264, 2010.
2. Tompa P, *Structure and Function of Intrinsically Disordered Proteins*, CRC Press, 2010.
3. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN, The unfoldomics decade: An update on intrinsically disordered proteins, *BMC Genomics* **9**:S1−S26, 2008.
4. Uversky VN, Oldfield CJ, Dunker AK, Intrinsically disordered proteins in human diseases: Introducing the D2 concepts, *Annu Rev Biophys* **37**:215−246, 2008.
5. Uversky VN, Oldfield CJ, Midic U, Xie H, Vucetic S, Xue B, Iakoucheva LM, Obradovic Z, Dunker AK, Unfoldomics of human diseases: Linking protein intrinsic disorder with diseases, *BMC Genomics* **10**:S7, 2009.
6. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK, Intrinsic disorder in cell-signalling and cancer-associated proteins, *J Mol Biol* **323**:573−584, 2002.
7. Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN, Abundance of intrinsic disorder in protein associated with cardiovascular disease, *Biochemistry* **45**:10448−10460, 2006.
8. Raychaudhuri S, Dey S, Bhattacharyya NP, Mukhopadhyay D, The role of intrinsically unstructured proteins in neurodegenerative diseases, *PLoS ONE* **4**:e5566, 2009.

9. Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN, Protein disorder in the human diseasome: Unfoldomics of human genetic diseases, *BMC Genomics* **10**:S12−S36, 2009.
10. Uversky VN, Amyloidogenesis of natively unfolded proteins, *Curr Alzheimer Res* **5**:260−287, 2008.
11. Daughdrill G, Pielak G, Uversky VN, Cortese M, Dunker AK, Natively unfolded proteins, in Buchner J and Kiefhaber T (eds.), *Protein Folding Handbook*, Weinheim, Wiley-VCH, pp. 275−337, 2005.
12. Ferron F, Longhi S, Canard B, Karlin D, A practical overview of protein disorder prediction methods, *Proteins* **65**:1−14, 2006.
13. Dosztanyi Z, Meszaros B, Istvan S, Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins, *Brief Bioinform* **2**:225−243, 2009.
14. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK, Predicting intrinsic disorder in proteins: An overview, *Cell Res* **19**:929−949, 2009.
15. Deng X, Eickholt J, Cheng J, A comprehensive overview of computational protein disorder prediction method, *Mol BioSyst* **8**:114−121, 2012.
16. Ward JJ, Sodhi JS, McGuffic LJ, Buxton BF, Jones DT, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J Mol Biol* **337**:635−645, 2004.
17. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK, Exploiting heterogeneous sequence properties improves prediction of protein disorder, *Proteins* **7**:176−182, 2005.
18. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z, Length-dependent prediction of protein intrinsic disorder, *BMC Bioinformatics* **7**:208−225, 2006.
19. Dosztanyi Z, Csimok V, Tompa P, Simon I, The pairwise energy content estimated from amino acid composition discriminate between folded and intrinsically unstructured proteins, *J Mol Biol* **347**:627−639, 2005.
20. Walsh I, Martin AJM, Di Domenico, Tosatto SCE, ESpritz: Accurate and fast prediction of protein disorder, *Bioinformatics* **28**(4):503−509, 2012.
21. Bordoli L, Kiefer F, Schwede T, Assessment of disorder predictions in CASP7, *Proteins* **69**:129−136, 2007.
22. Noivirt-Brik O, Prilusky J, Sussman JL, Assessment of disorder predictions in CASP8, *Proteins* **77**:210−216, 2009.
23. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A, Evaluation of disorder predictions in CASP9, *Proteins* **79**(S10):107−118, 2011.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov JN, Bourne PE: The protein data bank, *Nucleic Acids Res* **28**:235−242, 2000.
25. Sickmeier M, Hamilton J, LeGall T, Vacic V, Cortese M, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK, DisProt: The database of disordered proteins, *Nucl Acid Res* **35**:D786−D793, 2007.
26. Vucetic S, Brown CJ, Dunker AK, Obradovic Z, Flavors of protein disorder, *Proteins* **52**:573−584, 2003.
27. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B, Protein disorder — a breakthrough invention of evolution? *Curr Opin Struct Biol* **21**:412−418, 2011.
28. Gsponer J, Futschik ME, Teichmann SA, Babu MM, Tight regulation of unstructured proteins: From transcript synthesis to protein degradation, *Science* **322**:1365−1368, 2008.
29. Hobohm U, Scharf M, Schneider R, Sander C, Selection of a representative set of structures from the Brookhaven Protein Data Bank, *Protein Sci* **1**:409−417, 1992.

30. Hobohm U, Sander C, Enlarged representative set of protein structure, *Protein Sci* **3**:522−524, 1994.
31. Griep S, Hobohm U, PDBSelect 1992−2009 and PDBfilter-select, *Nucleic Acids Res* **38**: D318−D319, 2010.
32. The Uniprot consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Res* **40**:D71−D75, 2012.
33. Efron B, Tibshirani RJ, *An Introduction to the Bootstrap*, Chapman & Hall, CRC, 1993.
34. Davis J, Goadrich M, The relationship between precision-recall and ROC curves, *Proc 23rd Int Conf Machine Learning*, Pittsburgh, 2006.
35. Orengo CA, Thornton JM, Protein families and their evolution. A structural perspective, *Annu Rev Biochem* **74**:867−900, 2005.

Antonio Deiana received his M.Sc. degree (laurea) in Physics and Ph.D. in Biophysics from Sapienza University of Rome in 2007 and 2011, respectively. He is a post-doc research associate at the Department of Physics of Sapienza University of Rome since 2011, concentrating on bioinformatics and computational biophysics.

Andrea Giansanti received his laurea in Physics from Rome University in 1977. He was hired as an Assistant Professor in 1981 and since then he has served in the Sapienza University of Rome, apart from sabbatical leaves at the Rockefeller University in New York, and in other institutions in Europe and Italy. He has been working both theoretically and experimentally on subjects ranging from molecular and computational biophysics to nonlinear dynamics and the statistical mechanics of Hamiltonian systems. He has taught several courses in experimental physics and in condensed matter and computational biophysics. He is a fellow of the Royal Society of Arts of London (RSA).