W190–W196 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 6 June 2011*
doi:10.1093/nar/gkr411

CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs

Ian Walsh[1], Alberto J. M. Martin[1], Tomàs Di Domenico[1], Alessandro Vullo[2], Gianluca Pollastri[3] and Silvio C. E. Tosatto[1,*]

[1]Department of Biology, University of Padua, Padova 35131, Italy, [2]European Bioinformatics Institute, EMBL Outstation, Hinxton CB10 1SD, UK and [3]School of Computer Science and Informatics, University College Dublin, Dublin 4, Ireland

Received February 27, 2011; Revised May 3, 2011; Accepted May 7, 2011

ABSTRACT

CSpritz is a web server for the prediction of intrinsic protein disorder. It is a combination of previous Spritz with two novel orthogonal systems developed by our group (Punch and ESpritz). Punch is based on sequence and structural templates trained with support vector machines. ESpritz is an efficient single sequence method based on bidirectional recursive neural networks. Spritz was extended to filter predictions based on structural homologues. After extensive testing, predictions are combined by averaging their probabilities. The CSpritz website can elaborate single or multiple predictions for either short or long disorder. The server provides a global output page, for download and simultaneous statistics of all predictions. Links are provided to each individual protein where the amino acid sequence and disorder prediction are displayed along with statistics for the individual protein. As a novel feature, CSpritz provides information about structural homologues as well as secondary structure and short functional linear motifs in each disordered segment. Benchmarking was performed on the very recent CASP9 data, where CSpritz would have ranked consistently well with a Sw measure of 49.27 and AUC of 0.828. The server, together with help and methods pages including examples, are freely available at URL: http://protein.bio.unipd.it/cspritz/.

INTRODUCTION

The 3D native structure of proteins has been considered the major determinant of function for many years. Over the last decade there has been a growing realization of an alternative mechanism whereby non-folding regions are both widespread and also carry functional significance (1,2). These non-folding regions within a protein, coming in various guises ranging from fully extended to molten globule-like and partially folded structures (3), are collectively known as intrinsically disordered regions (4). Such regions often become structured upon binding to a target molecule and have been shown to be involved in various biological processes such as cell signaling or regulation (5), DNA binding (6) and molecular recognition in general (3,7). An interesting observation is that the amount of disorder within a proteome seems to correlate with complexity of the organism, with an apparent increase in disorder for eukaryotic organisms (8,9). The conservation of disorder (10,11) and specific amino acid patterns (12,13) (e.g. PxPxP) have also been studied. Indeed, there is a growing realization that intrinsically disordered regions are widely used as hubs for protein–protein interactions (14), for which structural data can be accessed in the ComSin database (15). Functional linear motifs (16,17), which are mostly hidden in disordered regions (18), have been characterized in resources such as ELM (19), an online repository of linear motifs.

The experimental determination of native disorder, once considered an anomaly, can be time consuming, difficult and expensive. As a result, computational approaches have largely driven our understanding of disorder over the last decade (14). The bi-yearly Critical Assessment of Techniques for protein Structure Prediction (CASP) experiment has included a disorder category since CASP5 in 2002 (20). Previously published methods can be roughly divided into biophysical and machine learning approaches. The former rely on the unique amino acid distribution associated with protein disorder (21–23). Machine learning methods use either neural networks (24–26) or support vector machines (9,27) and are commonly based on sequence profiles, predicted secondary structure and more recently template structures (28).

*To whom correspondence should be addressed. Tel: +390 498 276 958; Fax: +390 498 276 280; Email: silvio.tosatto@unipd.it

More recently, meta servers combining several biophysical and machine learning methods have been published (29–31). All these methods have shown promising results, possibly for two reasons: (i) as the amino acid sequence contains all the information to determine structure it is reasonable to assume that unstructured regions have specific amino acid propensities and (ii) disorder is important in many biological functions and therefore unstructured protein segments should be conserved by evolution. Knowing that disordered segments have a biased sequence, machine learning techniques should excel. In this paper we describe and benchmark CSpritz, an extension of our previous Spritz server (27) based on three distinct modules for the prediction of intrinsically disordered regions in proteins. The performance of the method will be benchmarked on the latest available data for short and long disordered segments. A novel addition to the CSpritz server is information about homologous structures found from PSI-BLAST searches, secondary structure and linear motifs contributing to the functional annotation of disordered segments.

MATERIALS AND METHODS

CSpritz predicts intrinsic disorder from protein sequences through a combination of three machine learning systems, which will be described in the following sections. Most methods consider short and long disorder separately, as they have different characteristics. Short disorder can be derived from residues missing backbone atoms in X-ray crystallographic structures deposited in the Protein Data Bank (PDB) (32). Long disorder is taken from the Disprot database (33) because it is largely missing from the PDB. All data sets used throughout training are appropriately redundancy reduced using UniqueProt (34) and in all cases contain only sequences available before May 2008 (i.e. the start of CASP8).

Spritz

The original Spritz (27) is based on PSI-BLAST (35) multiple sequence profiles and predicted secondary structure. Support Vector Machines (SVMs) were used on a local sequence window to train two specialized binary classifiers, for long and short regions of disorder. A description of the data sets can be found in the previous publication (27). In addition to the original *ab initio* version of Spritz, a filter removing PDB structural homologues from predicted disorder is implemented. This works by performing a PSI-BLAST search against a redundancy reduced sequence database. The generated sequence profile is then used in a final PSI-BLAST round against a filtered PDB. Residues matching a structural template are assigned a Spritz score below the disorder threshold.

Punch

Punch is a SVM based predictor extending Spritz. Sequence and structural homologues are detected as in Spritz. In addition, Porter secondary structure (36) and PaleAle relative solvent accessibility (37) are also included. Unlike Spritz, information about structural templates is encoded and fed directly to the SVM together with the other inputs. The two data sets used for learning (see Supplementary Data) are a large set of disordered X-ray chains derived from the PDB (December 2007) and a publicly available data set (24) based on disordered X-ray segments from the PDB (May 2004). The assignment of disorder is different in both data sets and does not necessarily intersect.

ESpritz

ESpritz is a fast predictor using bidirectional recursive neural networks (BRNNs) (38). BRNNs do not require contextual windows because they extract this information dynamically from the sequence. ESpritz consists of 20 inputs where each unit is allocated for one of the 20 amino acids. Although the method is very simple, the BRNN is useful for extracting relevant patterns required for disorder without the use of PSI-BLAST sequence alignments (results not shown). Like Spritz, two types of data based on long and short disorder types are designed (see Supplementary Material). The short disorder set is built from X-ray PDB structures (May 2008). Long disorder segments are extracted from Disprot (version 3.7) with identical sequences removed.

Linear motifs and secondary structure

It can be useful to unify the following information for disordered segments: (i) amino acids involved; (ii) secondary structure; and (iii) important linear motifs. CSpritz offers this predicted information in various forms (see output section). Secondary structure propensities are predicted from Porter (36). Linear motifs (LMs) are selected from ELM (19) as the ligand binding subset (names starting with LIG). ELM is a resource for predicting functional sites in eukaryotic proteins where functional sites are identified by patterns. These motifs are supposed to be representative of the more studied LM–protein binding examples. The selected LMs are returned when sub-sequences are matched by their regular expressions in ELM.

PERFORMANCE EVALUATION

Combination

Experiments were carried out for the best procedure to combine Punch, Spritz and ESpritz. After trying majority voting, unanimous votes and combination with neural networks, the simplest method of averaging the probabilities produced by each system was found to be the best (data not shown). The optimal decision threshold was determined on data independent from the benchmarking set by maximizing the Sw measure (39). CASP8 data (39) was used for short and Disprot (version 3.7) for long disorder. Regular expressions are incorporated to fill disordered regions separated by less than three residues. The Pearson correlation of the probabilities produced on CASP9 disorder targets was calculated to test how different the three predictors are. Table 1 shows this correlation and proves that the three systems

Table 1. Pearson correlation of the three systems on CASP9 targets

	ESpritz	Spritz	Punch
ESpritz	1.00	0.51	0.59
Spritz		1.00	0.42
Punch			1.00

The probabilities are produced by each component on all residues for 117 CASP9 targets. Since the correlations are low, combining the three systems improves performance over the individual systems.

are indeed sufficiently different. This is important for combining the three systems since it is well known that ensembling predictions which are different or uncorrelated improve generalization performance considerably (40). In particular, combination is especially beneficial when the wrongly predicted residues for each predictor do not correlate (i.e. their probabilities do not correlate) (41,42).

Benchmarking sets

Validation of short disorder segments is performed on the 117 CASP9 targets (URL: http://www.predictioncenter.org/casp9/), comparing with other groups taking part in the disorder category experiment according to their official CASP results. In order to validate the long disorder segments we choose DisProt entries enriched with PDB annotation from the SL data set defined in (43). Unfortunately, selecting sequences with <40% sequence identity to our training set leaves only 29 proteins. We also define a set of 569 X-ray sequences (Xray569) deposited in the PDB (resolution at most 2.5 Å and R-free <0.25) between May 2008 and September 2010 reduced by sequence identity using UniqueProt (34) to an HSSP value of 0 to our training data and among each other. Supplementary Table S1 shows the size and composition of the validation data sets. Note that to ensure a fair comparison to other methods on our benchmarking sets, CSpritz was in all cases run with sequence and PDB databases frozen prior to May 2008.

CASP short disorder

To assess the performance of our server for the short disorder option, we rank all groups participating in the CASP9 experiment. Table 2 shows the top 5 (out of 32) groups plus CSpritz and Spritz ranked by Sw, a commonly used measure at CASP. For Sw, as in the CASP8 assessment (39) the statistical significance of the evaluation scores was determined by bootstrapping: 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). For a full list of rankings see the online methods page. Our results suggest a consistently good performance of our server, especially when taking into account that some of the top five are meta-servers and some are not publicly available.

Table 2. Results for the top five performing groups at the CASP9 experiment, CSpritz and the original Spritz

GroupID: Name	Sw (±SE)	ACC	AUC
291: PRDOS2	50.44 (±1.08)	75.22	0.852
119: MULTICOM-REFINE	49.53 (±1.00)	74.77	0.818
000: CSpritz	49.27 (±1.02)	74.64	0.828
351: BIOMINE_DR_PDB	48.21 (±1.25)	74.11	0.818
374: GSMETADISORDERMD	47.13 (±0.96)	73.57	0.815
193: MASON	45.98 (±1.17)	73.00	0.740
000: Spritz	24.91 (±1.18)	62.46	0.716

Disordered segments of less than three residues were removed (results unchanged if included, see Supplementary Table S3). The standard error (SE) for Sw is shown in brackets. ACC is the accuracy, i.e. (sensitivity + specificity)/2, and AUC the area under the receiver operator curve. A total of 32 groups participated in CASP9 disorder prediction category.

Table 3. Comparison for DisProt disordered regions

Method	Sw (±SE)	ACC	AUC
CSpritz (short)	54.64 (±3.58)	77.32	0.837
CSpritz (long)	65.70 (±3.52)	82.85	0.891
Spritz (short)	12.12 (±6.16)	56.06	0.685
Spritz (long)	35.55 (±3.58)	67.78	0.734
PONDR-FIT	51.53 (±4.34)	75.77	0.817
Disopred2	46.20 (±4.00)	73.10	0.806
IUPred (short)	37.65 (±4.77)	68.83	0.814
IUPred (long)	42.57 (±4.75)	71.29	0.818

CSpritz is compared with the original Spritz, PONDR-FIT, Disopred and IUPred. Where applicable both short and long options are reported. The standard error (SE) for Sw is shown in brackets. ACC is the accuracy, i.e. (sensitivity + specificity)/2, and AUC the area under the receiver operator curve. The decision threshold and best Sw was found to be 0.26 and 51.85 on the training set.

DisProt long disorder

The long disorder type performance of CSpritz was benchmarked by comparing Sw, accuracy and AUC with the original Spritz and state-of-the-art predictors PONDR-FIT (30), Disopred (9) and IUPred (23). Table 3 shows CSpritz performing significantly better than the other predictors for this type of disorder. In addition CSpritz improves over the long disorder predictions made by our previous server Spritz.

Large-scale performance

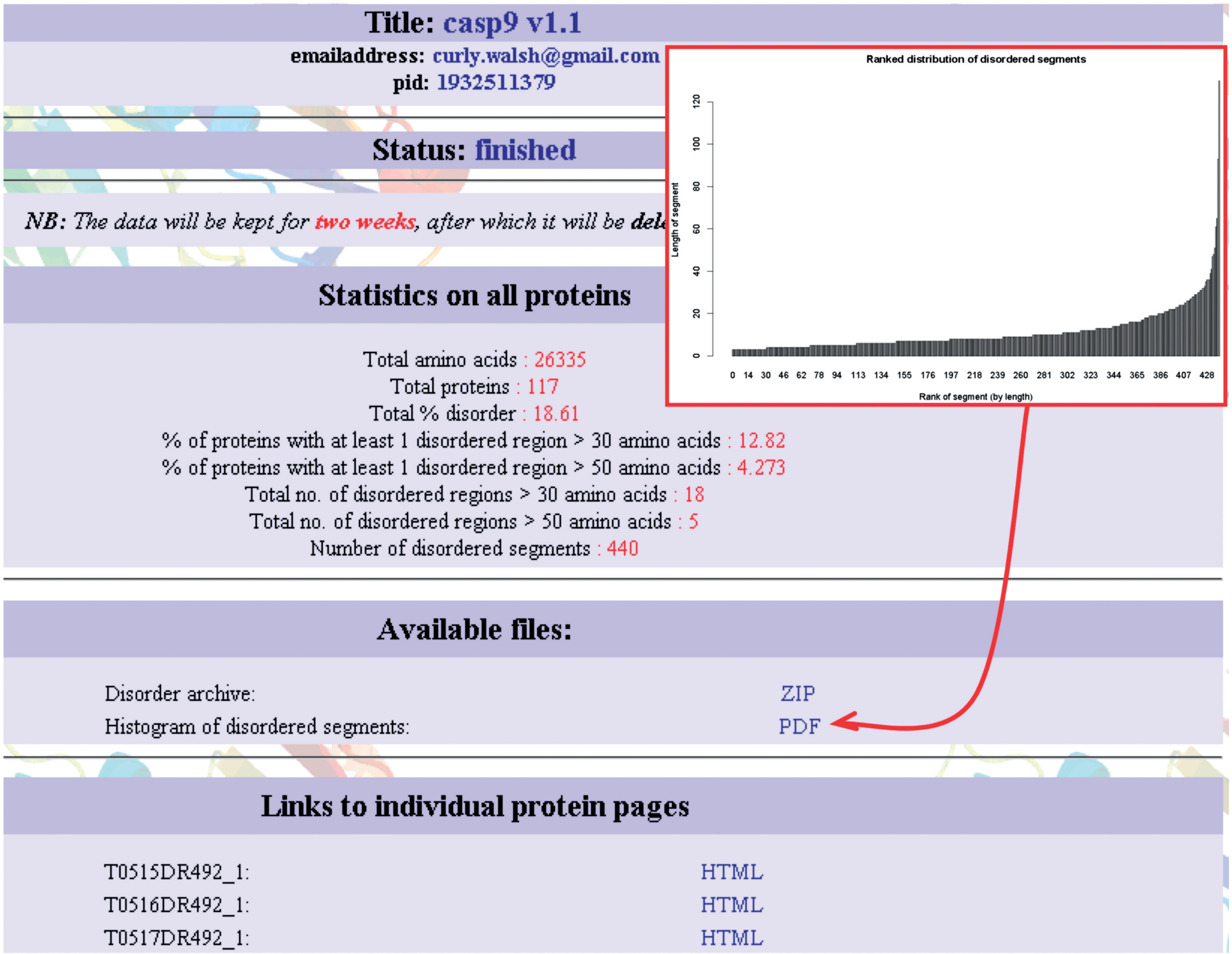
To estimate the run time of CSpritz compared to others and validate the predictions on a larger set of PDB structures we use the Xray569 set. The results (Supplementary Table S2) are similar to the DisProt set and confirm the performance of CSpritz compared to the other methods. As can be expected, all methods are better at predicting disorder at the N- and C-termini than in the central part of the protein sequences. The execution time for CSpritz is largely determined by the PSI-BLAST search and comparable to the original Spritz and Disopred2, with ca. 15 min for an average protein. When executing multiple predictions, the CSpritz web server will run up to five proteins in parallel, reducing the overall time significantly.

SERVER DESCRIPTION

The CSpritz input page is designed with simplicity in mind. A single or multiple sequences in FASTA format are the only input required and can be either pasted or uploaded as a file. Pasting is limited to 32 000 characters but uploading has no restrictions. User email address and a query title are optional. Either short (default) or long disorder options can be selected, with the appropriate decision thresholds determined on data not involved in the benchmarking. To facilitate navigation, help and methods pages are available at the top of the interface.

The CSpritz output is presented in two main pages. The first page, displaying statistics, links to individual pages and a downloadable archive for all user supplied proteins, is present only if more than one sequence was submitted. A histogram of disordered segments and an archive for download containing all generated data are also available. Figure 1 shows a sample global page for the 117 CASP9 targets.

The second output displays predicted disorder and annotation for individual proteins. In addition to showing the sequence with predicted secondary structure and disorder, several statistics regarding the distribution of disorder are presented. An extensive description of the output is available as part of the online help page. Two graphs plot the probability of disorder and the number of available structural templates versus disordered regions in homologous PDB structures. The last part of the output concerns the presence of putative linear motifs and secondary structure propensity for disordered segments. This can be a useful source of functional annotation, as shown in Figure 2 for *Drosophila melanogaster* Cryptochrome (dCRY). Following computational analysis, functional linear motifs were experimentally confirmed in the disordered C-terminus of dCRY (44). CSpritz aims to speed up this type of analysis by providing additional clues. In dCRY the putative linear motifs (Figure 2) match the disordered residues having a favorable alpha helical propensity. It is known that many such interactions



Title: casp9 v1.1

emailaddress: curly.walsh@gmail.com

pid: 1932511379

Status: finished

***NB:** The data will be kept for two weeks, after which it will be dele*

Statistics on all proteins

Total amino acids : 26335

Total proteins : 117

Total % disorder : 18.61

% of proteins with at least 1 disordered region > 30 amino acids : 12.82

% of proteins with at least 1 disordered region > 50 amino acids : 4.273

Total no. of disordered regions > 30 amino acids : 18

Total no. of disordered regions > 50 amino acids : 5

Number of disordered segments : 440

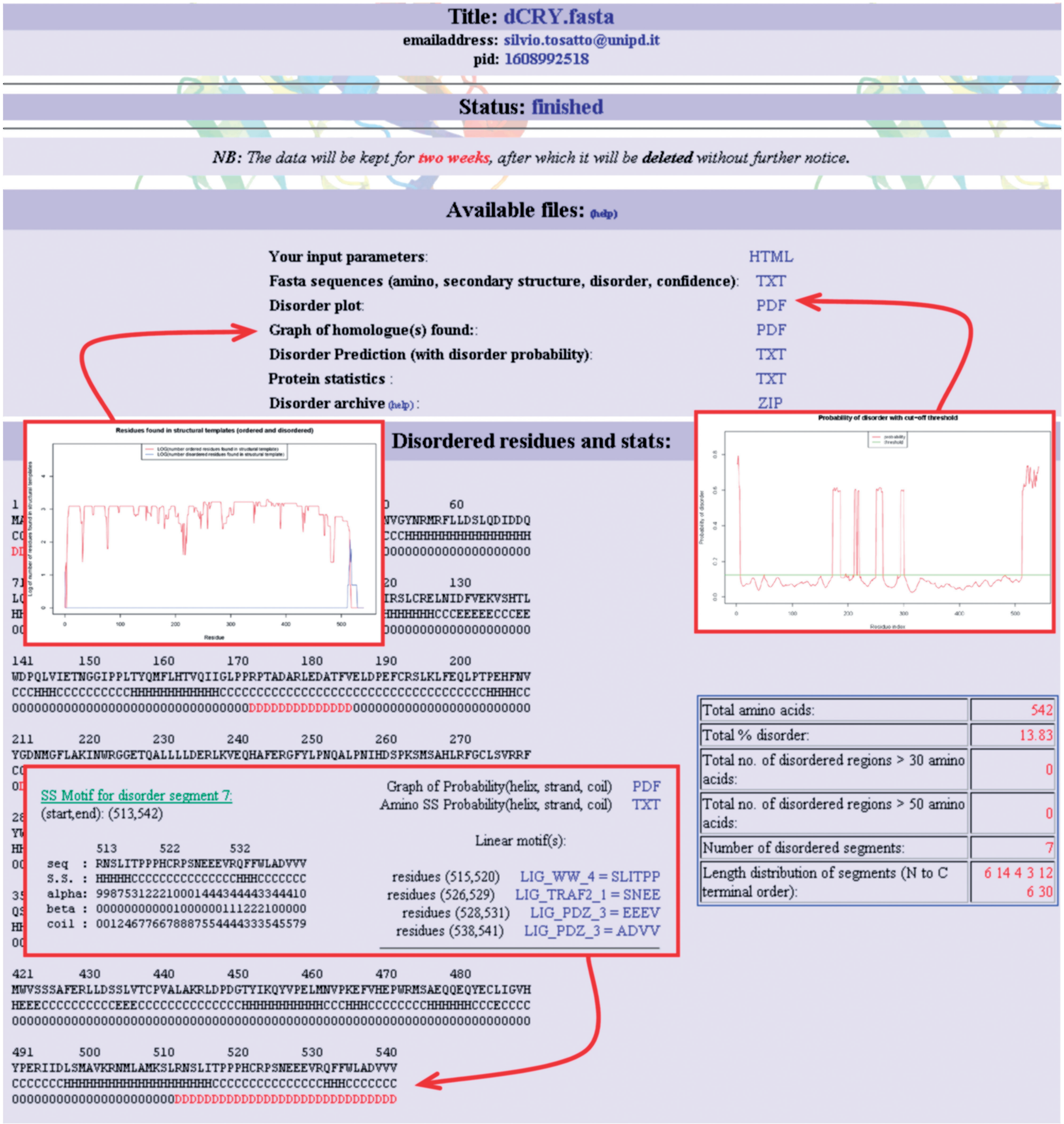
Available files:

Disorder archive:	ZIP
Histogram of disordered segments:	PDF

Links to individual protein pages

T0515DR492_1:	HTML
T0516DR492_1:	HTML
T0517DR492_1:	HTML

Figure 1. Global output page for multiple sequences. Summary statistics are displayed for some interesting values about the disorder segments of all query sequences. An archive is offered for download containing all disorder predictions, linear motifs and statistics for each protein the user supplied. The inset shows a graph displaying the length distribution of disorder segments among all proteins.



Title: dCRY.fasta

emailaddress: silvio.tosatto@unipd.it
pid: 1608992518

Status: finished

*NB: The data will be kept for **two weeks**, after which it will be **deleted** without further notice.*

Available files: (help)

Your input parameters:	HTML
Fasta sequences (amino, secondary structure, disorder, confidence):	TXT
Disorder plot:	PDF
Graph of homologue(s) found::	PDF
Disorder Prediction (with disorder probability):	TXT
Protein statistics :	TXT
Disorder archive (help) :	ZIP

Disordered residues and stats:

Residues found in structural templates (ordered and disordered)
LOG(number ordered residues found in structural template)
LOG(number disordered residues found in structural template)
Log of number of residues found in structural templates
0
1
2
3
4
0
100
200
300
400
500
Residue

```
1                                                    0       60
MA                                                   NVGYNRMRFLLDSLQDIDDQ
CC                                                   CCCHHHHHHHHHHHHHHHHH
DI                                                   00000000000000000000

71                                                   20      130
LQ                                                   IRSLCRELNIDFVEKVSHTL
HH                                                   HHHHHHHCCCEEEEECCCEE
00                                                   00000000000000000000

141      150       160       170       180       190       200
WDPQLVIETNGGIPPLTYQMFLHTVQIIGLPPRPTADARLEDATFVELDPEFCRSLKLFEQLPTPEHFNV
CCCHHHCCCCCCCCCCHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHCC
0000000000000000000000000000000DDDDDDDDDDDDDD0000000000000000000000000

211      220       230       240       250       260       270
YGDNMGFLAKINWRGGETQALLLLDERLKVEQHAFERGFYLPNQALPNIHDSPKSMSAHLRFGCLSVRRF
CC
0D

28
YW
HH
00

35
QS
HH
00
```

SS Motif for disorder segment 7:
(start,end): (513,542)

```
          513       522        532
 seq  : RNSLITPPPHCRPSNEEEVRQFFWLADVVV
 S.S. : HHHHHCCCCCCCCCCCCCCHHHCCCCCCC
 alpha: 998753122210001444344443344410
 beta : 000000000001000000111222100000
 coil : 001246776678887554444333545579
```

Graph of Probability(helix, strand, coil)	PDF
Amino SS Probability(helix, strand, coil)	TXT

Linear motif(s):

residues (515,520)	LIG_WW_4 = SLITPP
residues (526,529)	LIG_TRAF2_1 = SNEE
residues (528,531)	LIG_PDZ_3 = EEEV
residues (538,541)	LIG_PDZ_3 = ADVV

```
421      430       440       450       460       470       480
MWVSSSAFERLLDSSLVTCPVALAKRLDPDGTYIKQYVPELMNVPKEFVHEPWRMSAEQQEQYECLIGVH
HEEECCCCCCCCCCEEECCCCCCCCCCCCCCHHHHHHHHHHHCCCHHHCCCCCCCCHHHHHHCCCECCCC
0000000000000000000000000000000000000000000000000000000000000000000000

491      500       510       520       530       540
YPERIIDLSMAVKRNMLAMKSLRNSLITPPPHCRPSNEEEVRQFFWLADVVV
CCCCCCCHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCHHHCCCCCC
0000000000000000000000DDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
```

Total amino acids:	542
Total % disorder:	13.83
Total no. of disordered regions > 30 amino acids:	0
Total no. of disordered regions > 50 amino acids:	0
Number of disordered segments:	7
Length distribution of segments (N to C terminal order):	6 14 4 3 12 6 30

Figure 2. Individual output page for *D. melanogaster* Cryptochrome. The main figure shows the list of available files and actual disorder prediction. The latter is composed of the amino acid sequence, its predicted secondary structure and the CSpritz disorder classification, with disordered residues highlighted in red font. Disorder statistics about the protein is presented on the right. Two insets show the graphs for the disorder propensity plot (top right) and number of available structural coordinates versus disordered segments in homologous sequences. The inset on the bottom part shows the annotated disordered segment covering the C-terminus of Cryptochrome (residues 513–542). The propensities for secondary structure and location of putative functional motifs are shown. Links to the ELM description of the motif amino acids involved in the motif are supplied on the right. A graph and probabilities secondary structure propensity are also supplied.

involve disorder to secondary structure transitions upon binding (45).

CONCLUSIONS

We have described CSpritz, a novel web server for the prediction of intrinsically disordered protein segments from sequence. It allows the batch prediction of many sequences simultaneously, providing overview statistics. The single protein sequence is annotated with disorder and useful information regarding local secondary structure and possible interaction motifs, providing a first step towards the functional interpretation of disorder. Future work will concentrate on improving the functional description of disordered regions by including other types of related information such as repeats (46) and aggregation (47).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to members of the BioComputing UP lab for insightful discussions.

FUNDING

University of Padova (CPDA098382, CPDR097328 to S.T.); FIRB Futuro in Ricerca (RBFR08ZSXY to S.T.). Funding for open access charge: FIRB Futuro in Ricerca grant from the Italian Ministry of Education, University and Research (MIUR).

Conflict of interest statement. None declared.

REFERENCES

1. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
2. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
3. Tompa,P. and Fuxreiter,M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.*, **33**, 2–8.
4. Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
5. Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M. and Obradovic,Z. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.
6. Weiss,M.A., Ellenberger,T., Wobbe,C.R., Lee,J.P., Harrison,S.C. and Struhl,K. (1990) Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. *Nature*, **347**, 575–578.
7. Tompa,P., Fuxreiter,M., Oldfield,C.J., Simon,I., Dunker,A.K. and Uversky,V.N. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, **31**, 328–335.
8. Dunker,A.K., Obradovic,Z., Romero,P., Garner,E.C. and Brown,C.J. (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 161–171.
9. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
10. Schaefer,C., Schlessinger,A. and Rost,B. (2010) Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. *Bioinformatics*, **26**, 625–631.
11. Siltberg-Liberles,J. (2010) Evolution of structurally disordered proteins promotes neostructuralization. *Mol. Biol. Evol.*, **28**, 59–62.
12. Lise,S. and Jones,D.T. (2005) Sequence patterns associated with disordered regions in proteins. *Proteins*, **58**, 144–150.
13. Lobanov,M.Y., Furletova,E.I., Bogatyreva,N.S., Roytberg,M.A. and Galzitskaya,O.V. (2010) Library of disordered patterns in 3D protein structures. *PLoS Comput. Biol.*, **6**, e1000958.
14. Russell,R.B. and Gibson,T.J. (2008) A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS Lett.*, **582**, 1271–1275.
15. Lobanov,M.Y., Shoemaker,B.A., Garbuzynskiy,S.O., Fong,J.H., Panchenko,A.R. and Galzitskaya,O.V. (2010) ComSin: database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder. *Nucleic Acids Res.*, **38**, D283–D287.
16. Gibson,T.J. (2009) Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci.*, **34**, 471–482.
17. Diella,F., Haslam,N., Chica,C., Budd,A., Michael,S., Brown,N.P., Trave,G. and Gibson,T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci.*, **13**, 6580–6603.
18. Fuxreiter,M., Tompa,P. and Simon,I. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.
19. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemund,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.
20. Melamud,E. and Moult,J. (2003) Evaluation of disorder predictions in CASP5. *Proteins*, **53(Suppl. 6)**, 561–565.
21. Uversky,V.N. (2002) What does it mean to be natively unfolded? *Eur. J. Biochem.*, **269**, 2–12.
22. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P. and Dunker,A.K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61(Suppl. 7)**, 176–182.
23. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
24. Cheng,J., Sweredoski,M.J. and Baldi,P. (2005) Accurate prediction of protein disordered regions by mining protein structure data. *Data Min Knowl Disc*, **11**, 213–222.
25. Jones,D.T. and Ward,J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **53(Suppl. 6)**, 573–578.
26. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
27. Vullo,A., Bortolami,O., Pollastri,G. and Tosatto,S.C. (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.*, **34**, W164–W168.
28. McGuffin,L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, **24**, 1798–1804.
29. Mizianty,M.J., Stach,W., Chen,K., Kedarisetti,K.D., Disfani,F.M. and Kurgan,L. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.
30. Xue,B., Dunbrack,R.L., Williams,R.W., Dunker,A.K. and Uversky,V.N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*, **1804**, 996–1010.
31. Schlessinger,A., Punta,M., Yachdav,G., Kajan,L. and Rost,B. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **4**, e4433.
32. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
33. Sickmeier,M., Hamilton,J.A., LeGall,T., Vacic,V., Cortese,M.S., Tantos,A., Szabo,B., Tompa,P., Chen,J., Uversky,V.N. *et al.* (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res.*, **35**, D786–D793.
34. Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.
35. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
36. Pollastri,G. and McLysaght,A. (2005) Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21**, 1719–1720.

37. Pollastri,G., Martin,A.J., Mooney,C. and Vullo,A. (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, **8**, 201.
38. Baldi,P. and Pollastri,G. (2003) The principled design of large-scale recursive neural network rchitectures–dag-rnns and the protein structure prediction problem. *J. Mach. Learn.*, **4**, 575–602.
39. Noivirt-Brik,O., Prilusky,J. and Sussman,J.L. (2009) Assessment of disorder predictions in CASP8. *Proteins*, **77(Suppl. 9)**, 210–216.
40. Sollich,P. and Krogh,A. (1996) Learning with ensembles: how over-fitting can be useful. *Adv. Neural Inform. Processing Sys.*, **8**, 190–196.
41. Albrecht,M., Tosatto,S.C., Lengauer,T. and Valle,G. (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng.*, **16**, 459–462.
42. Ali,K.M. and Pazzani,M.J. (1996) Error reduction through learning multiple descriptions. *Mach. Learn.*, **24**, 173–202.
43. Sirota,F.L., Ooi,H.S., Gattermayer,T., Schneider,G., Eisenhaber,F. and Maurer-Stroh,S. (2010) Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC Genomics*, **11(Suppl. 1)**, S15.
44. Hemsley,M.J., Mazzotta,G.M., Mason,M., Dissel,S., Toppo,S., Pagano,M.A., Sandrelli,F., Meggio,F., Rosato,E., Costa,R. *et al.* (2007) Linear motifs in the C-terminus of D. melanogaster cryptochrome. *Biochem. Biophys. Res. Commun.*, **355**, 531–537.
45. Vanhee,P., Stricher,F., Baeten,L., Verschueren,E., Lenaerts,T., Serrano,L., Rousseau,F. and Schymkowitz,J. (2009) Protein-peptide interactions adopt the same structural motifs as monomeric protein folds. *Structure*, **17**, 1128–1136.
46. Marsella,L., Sirocco,F., Trovato,A., Seno,F. and Tosatto,S.C. (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*, **25**, i289–i295.
47. Trovato,A., Seno,F. and Tosatto,S.C. (2007) The PASTA server for protein aggregation prediction. *Protein Eng. Des. Sel.*, **20**, 521–523.