

CAID

Critical Assessment of Intrinsic protein Disorder

Silvio Tosatto
University of Padua

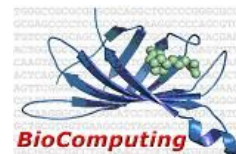
1222-2022
800
ANNI



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



IDP
 $f_{(un)}$



Intrinsically Disordered Proteins (IDPs) and Regions (IDRs)



Disordered
Dynamic



Disordered
Compact

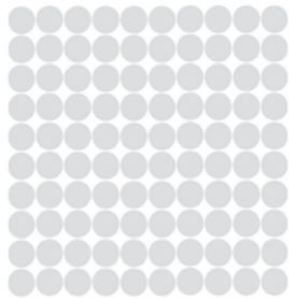


Disordered
Extended

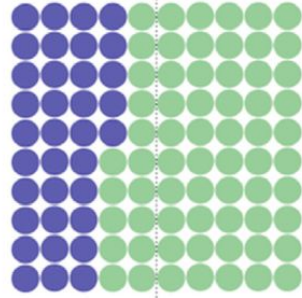


Disordered
Residual Structure

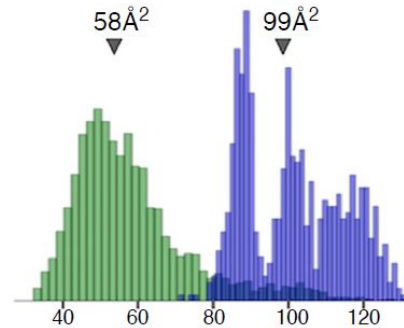
Primary isoforms of ca. 21k human proteins



11 million residues



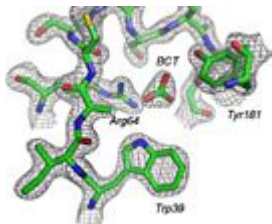
3-4 million residues in IDRs



Å² accessible surface per residue



IDP Assessments



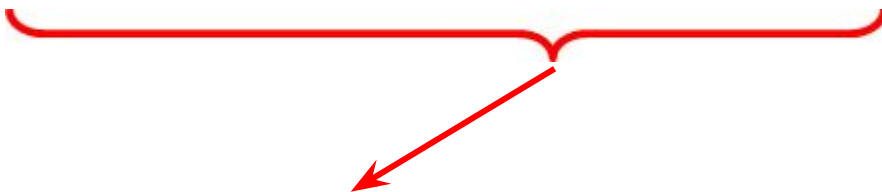
Missing residues
X-ray



Mobile residues
NMR



Functional disorder
“Secondary methods”



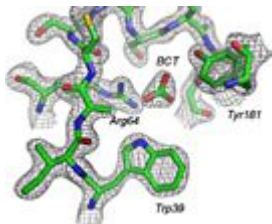
CASP5 (2002) - CASP10 (2012)

(Monastyrskyy et al, *Proteins*, 2011)

(Monastyrskyy et al, *Proteins*, 2013)

- Few targets (ca. 100)
- 7:1 short / long regions
- 10% ID residues
- Regions at the C- / N-termini

IDP Assessments



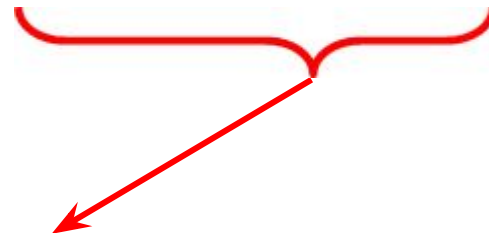
Missing residues
X-ray



Mobile residues
NMR



Functional disorder
“Secondary methods”



New DisProt 2020 (unpublished)

- 1,567 proteins (**764 new**)
- 646 non-redundant proteins used for CAID-1.

DisProt: intrinsic disorder evidence from the literature



Manually curated repository of disordered proteins

- first publication in 2005
- previous release in **2017**
 - completely re-annotated
 - ca. 200 new entries
- last update (**01/2020**)
 - quantitative and qualitative improvements
 - ca. 800 new entries

DisProt

Version: 8.0
Release: 2019_09

Intrinsically disordered proteins

DisProt is a database of intrinsically disordered proteins. Disordered regions are manually curated from literature. DisProt annotations cover both structural and functional aspects of disorder detected by specific experimental methods. Annotation concepts and detection methods are encoded in the Disorder Ontology. Read more about DisProt

Proteins per organism	Statistics
H. sapiens	Total 1.8k
M. musculus: 88	Not ambiguous 1.4k
R. norvegicus: 50	Proteins 1.8k
S. cerevisiae: 128	Regions 3.5k
V. fischeri: 126	Residues 105.2k
E. coli: 7	Disorder extent 19.7%
A. thaliana	
D. melanogaster: 30	
C. elegans: 13	
Fungi: 198	

How to cite: Povesan D et al. DisProt 7.0: a major update of the database of disordered proteins Nucleic Acids Res., 2016, PMID 27699601

API: REST API documentation here

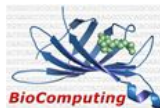
Disorder ontology: You can download the ontology from the Download page or explore it from the Release notes page

Integrated resources: UniProt, Pfam, BITEM, Europe PMC

BioComputingUP - Department of Biomedical Sciences - University of Padova, Italy - 2019

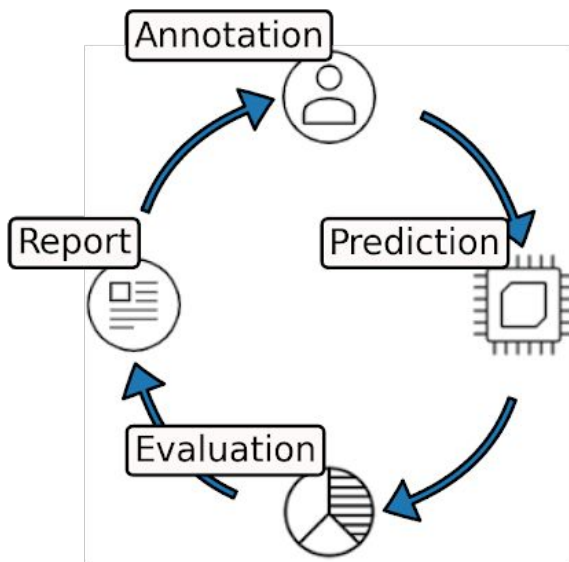
www.disprot.org

(Hatos et al., Nucleic Acids Research Database Issue 2020)



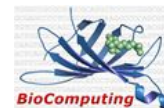


CAID cycle



- **Ground truth generation**
 - Literature curation (DisProt)
- **Prediction**
 - Stand-alone software (MobiDB servers)
- **Assessment**
 - Accuracy
 - Technical evaluation

idpcentral.org/caid

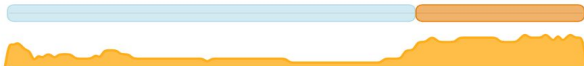


Prediction Setting

>P37840 Alpha-synuclein

```
MDVFMKGLSKAKEGVVAAAEEKTKQGVAAEAGKTEGVLVYVGSKTKEGVVH  
GVATVAEKTKEQVTNVTGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQL  
GKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA
```

+
[PSSM, MSA]



Stand-alone software (no web servers)

Installed on local machines @ University of Padua

- SGE cluster (nodes with different hardware, no GPU)
- Ubuntu (14.04/16.04), 64 bit
- 16Gb RAM, 8 cores (per node)

Predictors

27 software packages

- 8 unpublished
- 5 recently published (since 2017)

produced

43 outputs

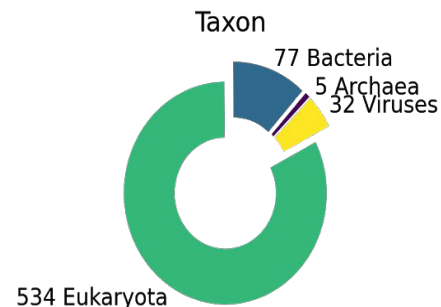
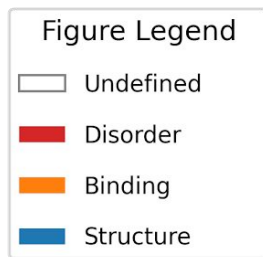
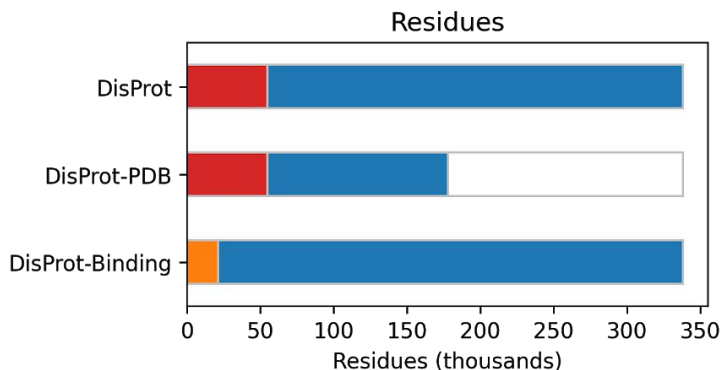
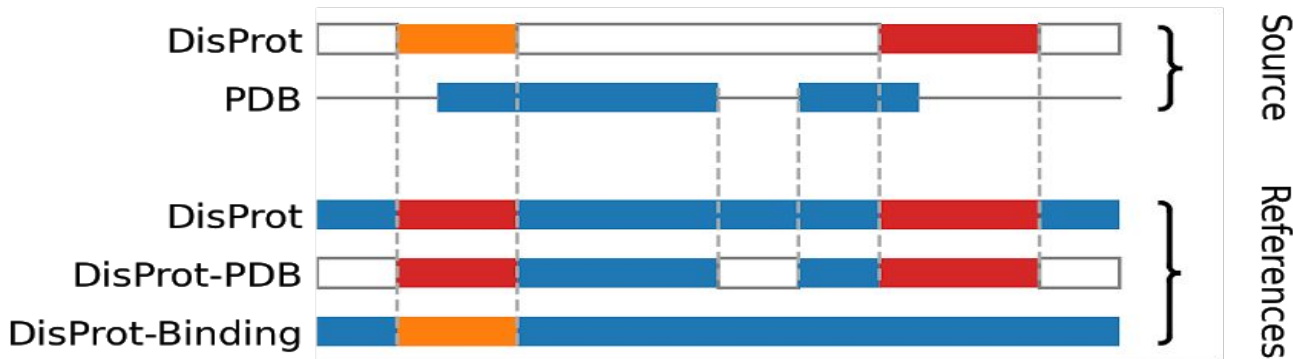
- 32 disorder
- 12 binding

gave rise to

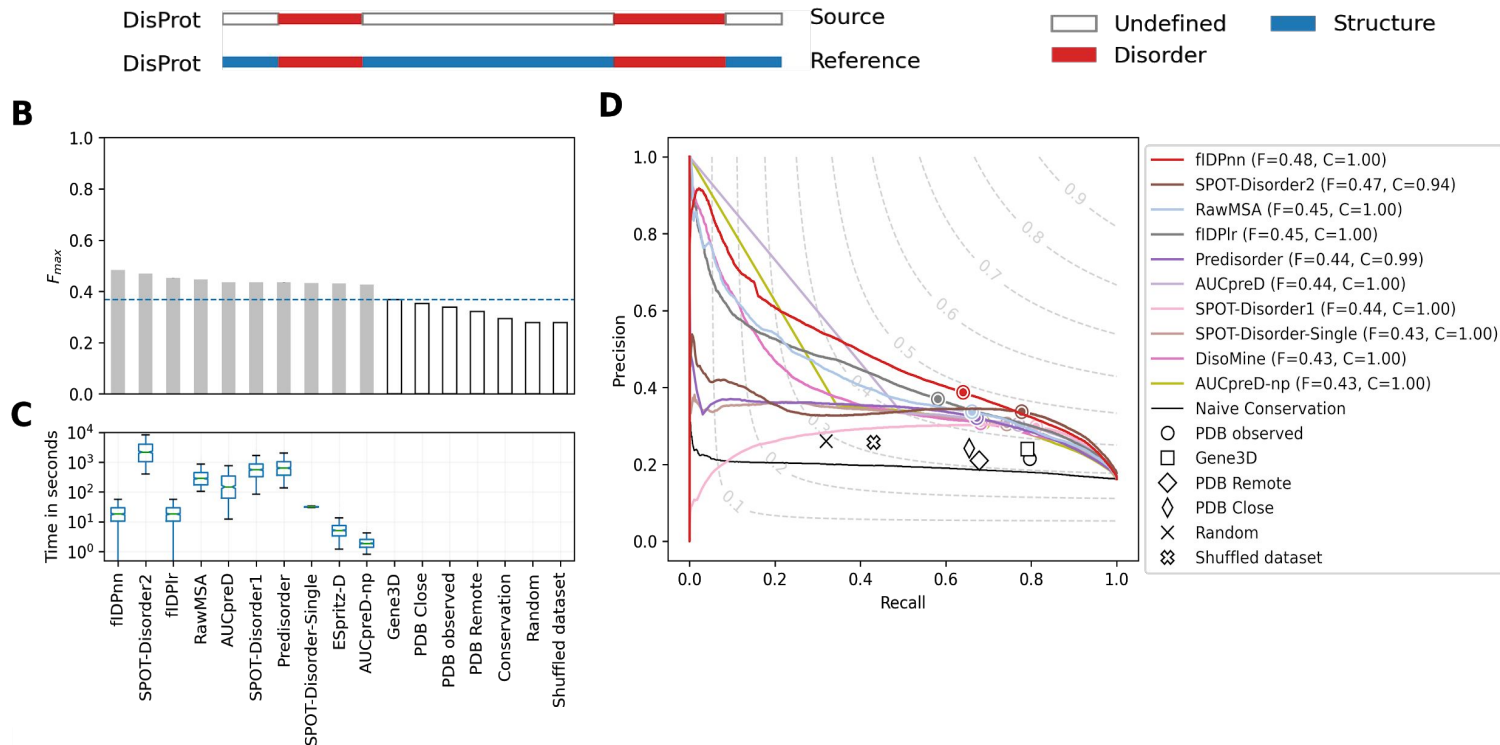
Many Issues

- Different output format
- Large (up to 60Gb)
- No control on output paths
- Specific version or obsolete system libraries
- Compile on new hardware (TensorFlow)
- Only one is provided as container/VM

Residue classification for *DisProt* and *DisProt-PDB*



Prediction success and CPU times, top ten **disorder** predictors *DisProt* dataset



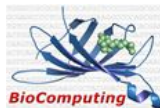
Performance of predictors expressed as maximum F1-Score across all thresholds (F_{max}) (panel B) for the top ten best ranking methods (light gray) and baselines (white) and the distribution of execution time per-target (panel C) using the *DisProt* dataset. The horizontal line in panel B indicates the F_{max} of the best baseline. Precision-Recall (panel D) of ten top-ranking methods and baselines using the *DisProt* dataset, with level curves of the F1-Score. Magenta dots on panel C indicate that the whole distribution of execution-times is lower than 1 second.

Prediction success, top 10 **disorder** predictors *fully disordered proteins* dataset

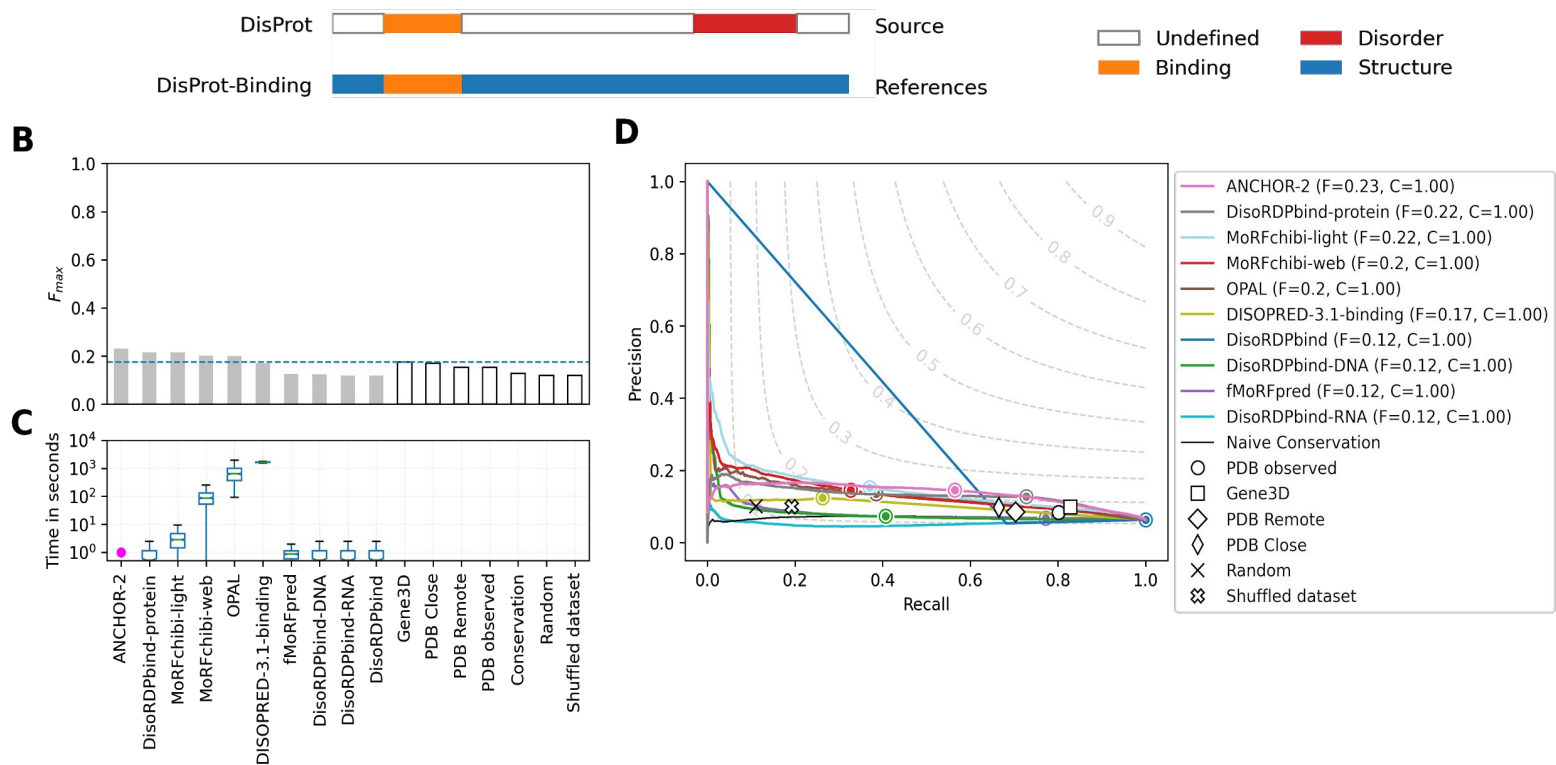
	TN	FP	FN	TP	MCC	F1-S	TNR	TPR	PPV	BAC
fIDPnn	585	16	19	26	0.569	0.598	0.973	0.578	0.619	0.776
RawMSA	582	19	19	26	0.546	0.578	0.968	0.578	0.578	0.773
VSL2B	578	23	22	23	0.468	0.505	0.962	0.511	0.500	0.736
fIDPIr	566	35	18	27	0.468	0.505	0.942	0.600	0.435	0.771
Predisorder	589	12	26	19	0.479	0.500	0.980	0.422	0.613	0.701
SPOT-Disorder1	572	29	23	22	0.416	0.458	0.952	0.489	0.431	0.720
DisoMine	551	50	17	28	0.421	0.455	0.917	0.622	0.359	0.770
AUCpreD	588	13	28	17	0.431	0.453	0.978	0.378	0.567	0.678
SPOT-Disorder2	574	27	24	21	0.409	0.452	0.955	0.467	0.438	0.711
SPOT-Disorder-Single	594	7	30	15	0.452	0.448	0.988	0.333	0.682	0.661
IsUnstruct	588	13	29	16	0.411	0.432	0.978	0.356	0.552	0.667
IUPred2A-long	595	6	32	13	0.420	0.406	0.990	0.289	0.684	0.639
Gene3D	505	96	10	35	0.391	0.398	0.840	0.778	0.267	0.809

Proteins with disorder prediction or disorder annotation covering at least 95% of the sequence are considered fully disordered. Predictors are sorted by their F1-Score.

The Gene3D baseline is shown for comparison.



Prediction success and CPU times, top ten **binding** predictors *DisProt-Binding* dataset



Performance of predictors expressed as maximum F1-Score across all thresholds (F_{max}) (panel B) for the top ten best ranking methods (light gray) and baselines (white) and the distribution of execution time per-target (panel C) using the *DisProt-Binding* dataset. The horizontal line in panel B indicates the F_{max} of the best baseline. Precision-Recall (panel D) of ten top-ranking methods and baselines using *DisProt-Binding* dataset, with level curves of the F1-Score. Magenta dots on panel C indicate that the whole distribution of execution-times is lower than 1 second.

Conclusions

- Novel predictors outperform old ones
- Evolutionary information (MSA, PSSM) is useful
- Disorder is different from missing residues
- A number of disorder predictors are difficult to use (configuration, dependencies, different output formats)
- Disordered binding regions are more difficult to predict
- MCC and F-score are recommended to evaluate predictors. F-score is better as the MCC is undefined when some classes are not represented (e.g. for fully disordered proteins)
- Fully disordered proteins represent a significant number and cannot be evaluated using PDB missing residues

CAID bioRxiv preprint: <https://doi.org/10.1101/2020.08.11.245852>



CAID-2

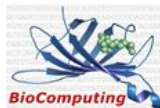
- **Software submission** **30 April 2021**
- Software successfully installed 15 May 2021
- Preliminary results July 2021
- Publication of results 2022



idpcentral.org/caid



[@BioComputingUP](https://twitter.com/BioComputingUP)



CAID meeting (online)

Thursday, **25 February 2021**, 2-6PM CET

- Methods presentations
- Town-hall discussion



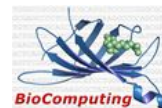
...more info coming soon



idpcentral.org/caid



[@BioComputingUP](https://twitter.com/BioComputingUP)





biocomputingup.it

