Critical Assessment of Protein Intrinsic Disorder Prediction

Marco Necci[1,], Damiano Piovesan[1,*], CAID Predictors[#], DisProt Curators[#], Silvio C.E. Tosatto[1,+]*
[1] Dept. of Biomedical Sciences, University of Padua, Italy
[*]*contributed equally*
[#]*see list of co-authors at the end of the manuscript*
[+]*corresponding author*

Abstract

Intrinsically disordered proteins defying the traditional protein structure-function paradigm represent a challenge to study experimentally. As a large part of our knowledge rests on computational predictions, it is crucial for their accuracy to be high. The Critical Assessment of protein Intrinsic Disorder prediction (CAID) experiment was established as a community-based blind test to determine the state of the art in predicting intrinsically disordered regions in proteins and the subset of disordered residues involved in binding other molecules. A total of 43 methods, 32 for disorder and 11 for binding regions, were evaluated on a dataset of 646 proteins. The best methods use modern machine learning techniques and significantly outperform widely used first-generation methods across different types of targets. Disordered binding regions remain hard to predict correctly, which offers significant potential for improvement. Intriguingly, some of the top performing methods are also among the fastest.

Introduction

The existence of intrinsically disordered proteins (IDPs) or protein regions (IDRs) is now well-established in structural biology[1]. In the last decades, there has been increasing evidence for the involvement of IDPs and IDRs in a variety of biological processes essential to the life in diverse organisms[2,3]. As they are involved in essentially all major human diseases, they also represent promising novel targets for drug discovery[4,5]. Experimental methods to characterize IDPs and IDRs include X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), small-angle X-ray scattering (SAXS), circular dichroism (CD) and Förster resonance energy transfer (FRET)[6–9]. Each technique provides a unique point of view on the phenomenon of intrinsic disorder (ID) and multiple experimental evidence gives researchers insights over the functioning mechanisms of IDPs.
An accumulation of experimental evidence has corroborated the early notion that ID can be inferred from sequence features, first suggested by R J Williams[10]. Since then, dozens of ID prediction methods based on different principles and computing techniques have been published[11], such as VSL2B[12], DisEMBL[13], DISOPRED[14], IUPred[15], Espritz[16] and many others. Annotations related to ID both predicted and derived from experimental evidence are stored in a variety of dedicated databases, each focusing on particular aspects of the ID spectrum, namely DisProt[17], MobiDB[18], IDEAL[19], DIBS[20] and MFIB[21]. Over the last few years, ID annotations are also provided in some of the most important core data resources like InterPro[22], UniProt[23] and PDBe[24].
The large variety of ID predictors available makes it difficult to compare them, confounding biologists wanting to make an informed choice to find the best performers. In this report, we describe the first edition of CAID, an experiment inspired by the Critical Assessment of protein Structure Prediction (CASP) for the benchmarking of ID predictors on community curated datasets. CAID is the first experiment of its kind and is bound to set a new quality standard in the field.

Results

Intrinsically disordered proteins/regions (IDPs/IDRs) provide the cell with a range of different molecular functions which complement those of globular domains[25,26]. Despite their importance, IDPs/IDRs are

understudied due to the difficulties in measuring their dynamic behavior directly and because some of them are disordered only under particular environmental conditions, i.e. their structural disorder is context dependent[27] (such as pH, PTMs, localization, and binding). Given a new protein sequence, the task of an ID prediction software is to assign a score to each residue describing its tendency of being intrinsically disordered at any stage of the protein life. In CAID, we evaluate both the accuracy of prediction methods and technical aspects related to software implementation like their speed, which has a direct impact on carrying out large-scale analyses.

CAID was organized as follows (**Fig. 1**). Participants submitted their implemented prediction software to the assessors and during the submission phase, the authors provided support to install and test execution on MobiDB servers. After the submission deadline, the assessors ran the packages and generated predictions for a set of proteins whose disorder annotations were not available at the training time.

Structural properties of proteins can be studied by a number of different experimental techniques, some giving direct, whereas others indirect evidence of disorder. Different techniques are biased in different ways, for example, IDRs inferred from missing residues in X-ray experiments are generally shorter because longer non-crystallizable ID fragments are either excised when preparing the construct or are detrimental to the crystallization of the protein. On the contrary, circular dichroism can detect the absence of fixed structure in the full protein but does not provide the resolution offered by methods like X-ray or NMR. IDR annotations are more reliable when confirmed by multiple lines of independent and different experimental evidence.

In the first round of CAID, we selected the DisProt database as the reference for structural disorder because it merges manually curated disorder annotations at the protein level, with the majority of residues annotated with more than one experiment[17]. DisProt annotates disordered regions of at least 10 residues in length to evaluate IDRs likely to be associated with a biological function by excluding short loops, i.e. segments connecting secondary structure elements. In general, functional IDRs are long and cover a significant portion of the sequence (**Fig. 1**). DisProt was also used as a reference for regions involved in binding events that take place in a disordered environment. Binding regions were retrieved leveraging the functional annotations that accompany the structural-state annotations in DisProt.

In an ideal situation, all DisProt annotations would be complete, i.e. each entry in DisProt would be annotated with all the disordered (or binding) regions it contains under physiological conditions. If this was true, we could simply consider all residues as structured (negatives) that are not annotated as disordered (positives). However, annotations in DisProt might be incomplete for some entries, therefore we created an ancillary dataset, DisProt-PDB, in which negatives are restricted by PDB observed residues (**Fig. 1**). This dataset is more conservative but can be considered more reliable as it excludes “uncertain” residues, which have neither structural nor disorder annotation. Compared to DisProt, DisProt-PDB is more similar to datasets used to train some of the disorder predictors (e.g. in[13,14,16]) and for CASP disorder challenges[28].

The distribution of organisms reflects what is known from other studies[29,30] with the majority of ID targets coming from eukaryotes with a good representation of viruses and bacteria, and much fewer from archaea (**Fig. 1**). At the species level, annotations are strongly biased in favor of model organisms, with the majority of cases coming from human, mouse, rat, *Escherichia coli* and a few other common model organisms (see Supplementary Material). Target proteins are not redundant at the sequence level and are different from known examples available in the previous DisProt release, with a median sequence identity of 8.0% within the dataset and of 10.7% against the previous release of DisProt (see Supplementary Material). CAID has two main categories: the prediction of intrinsic structural disorder and the predictions of binding sites found in IDRs.

IDR prediction performance

The quality of IDR prediction can be evaluated in different ways depending on the scope of studying structural disorder. In some cases, it is relevant to know disorder content; in other cases, it is more important to know the exact position of the IDR in the sequence. Since disorder can be used as a proxy to estimate the complexity of an organism or to complement a sequence search, it is also important for a predictor to be fast enough to be applied at the genome-scale. For CAID, we report a simple metric, the maximum F-score (F_{max}; Online Methods), which takes into account predictions across the entire spectrum of sensitivity, from high to low[50].

The performance of top methods, based on F_{max} and calculated over all targets, are shown in **Figure 2** for the DisProt and DisProt-PDB datasets. The F-score, which has the advantage of being insensitive to dataset imbalance (**Fig. 1**), provides a ranking almost identical to the one obtained with the Matthew's Correlation Coefficient (MCC). See the Supplementary Material for a full comparison and the dependence of F-score and MCC on the predictor confidence score, along with the software default confidence threshold. All methods were compared with various baseline tools described in the "Online Methods": two naive techniques and two random predictors. In some applications, the objective was to predict which protein fragments of the protein are disordered, based on known examples in the PDB. However, this was a different problem than predicting functional IDRs, for instance, aiming to evaluate their biophysical properties. The two naive baselines help us understand this difference and assess how effective the transfer-by-homology of structural information can be for IDR prediction (see Discussion). In the Naive-PDB baseline, all residues not covered by a PDB structure are labelled as disordered. Similarly, in Naive-Gene3D, residues are considered disordered if they do not match any Gene3D HMM model. In the Shuffled-Dataset baseline, the reference is randomly shuffled at the dataset level, while Random is an actual random predictor which does not use any prior knowledge.

F_{max} (**Fig. 2**) and Area under the ROC curve (AUC) (**Fig. 3**) measures are substantially different when predictors are tested on the DisProt dataset, which contains "uncertain" residues, and when tested on the DisProt-PDB dataset. Naive-PDB, by definition, cannot predict negative residues outside PDB regions. Therefore, it mainly generates type I errors (56.5%, false positives), which drop to zero when considering the DisProt-PDB dataset, where the "uncertain" residues are completely filtered out. IDRs overlapping with PDB regions, usually corresponding to residues that take part in folding-upon-binding events, instead generate type II errors. These are far less common (20.4%, false negatives) and remain the same for the two datasets. The Naive-Gene3D baseline typically increases PDB coverage (negatives) exploiting the transfer-by-homology principle. As a consequence, the probability of type I errors is lower (48.6% false positives), and type II errors are only marginally more frequent (20.9% false negatives). For the DisProt dataset, Naive-Gene3D slightly outperforms Naive-PDB both in terms of F_{max} (**Fig. 2**) and AUC (**Fig. 3**). Instead, for the DisProt-PDB dataset, Naive-PDB is significantly superior to all methods with only 6.3% of mispredicted residues, all false negatives.

Looking at the Precision-Recall curves (**Fig. 2**), we notice that the top five methods (excluding FIDPnn/lr in the DisProt dataset and AUCpred-np in the DisProt-PDB dataset) leverage evolutionary information, introducing a database search as a preliminary step. The performance gain, which is on average 4.5% in terms of F_{Max}, comes at the cost of slowing down the prediction of two to four (2-4) orders of magnitude (**Fig. 2**, see also Supplementary Material).

Fully disordered proteins

Within the challenge of disorder prediction, we separately considered the special category of fully disordered proteins (IDPs). These proteins are interesting as they are particularly challenging to investigate experimentally, e.g. they cannot be probed with X-ray crystallography, yet they are of great interest as they fulfil unique biological functions[30,31]. We therefore designed another challenge, in which the task is to tell apart fully from not-fully disordered proteins given their sequences. Each protein was classified on the basis of the fraction of its residues predicted as disordered. We loosen the "fully" definition by adopting a 95% threshold for both predictions and reference IDRs. According to this

definition, the number of fully disorder proteins in the DisProt dataset is 40 out of 646. Different threshold values did not significantly affect the ranking (see Supplementary Material). In **Table 1** all methods are sorted based on MCC. False positives are limited for many methods although correct IDP predictions are generally made for less than half of the dataset. This suggests room for improvement, as fully disordered proteins should be easier to predict from sequence.

Prediction of disordered binding sites

As a second major challenge, CAID also evaluates the prediction of binding sites within IDRs, commonly referenced to as MoRFs, SLIMs or LIPs[18,32], leveraging DisProt annotations of binding regions (see Supplementary Material for dataset composition and overlap to other databases). In DisProt, binding annotations retrieved from literature are fraught with more ambiguity than disorder ones. In addition, the experimental evidence for the exact position of a binding region is often not accurate, as binding is annotated as a feature of an IDR. Our reference includes all entries in the DisProt dataset even if they were not annotated with binding regions. This translates to a dataset where the majority of targets (414 out of 646) have no positives. In this challenge, we kept the Naive-PDB and Naive-Gene3D baselines even if they are not designed to detect binding regions. Target binding regions in DisProt are found within IDRs. Therefore, the baselines are expected to attain high recall and low precision. All models perform poorly as do the naive baselines (**Fig. 4**). At F_{Max}, their recall is higher than their precision as for the baselines (**Fig. 4C**). However, the top 5 methods, ANCHOR-2[15], DisoRDPbind[33], MoRFchibi (light and web)[34], and OPAL[35], perform better than the baselines (**Fig. 4A**), which trade-off much more precision due to an abundant over-prediction. The execution time of the top five methods have very different scales and are inversely proportional to their performance, with the best methods requiring less CPU time.

Software implementation

We also evaluated for CAID technical aspects related to software implementation, i.e. speed and usability, which have a direct impact on the application of large-scale analyses (see Supplementary Material). In general, all methods incorporate a mix of different scripts and programming languages. Some software configuration scripts contain errors. In many cases data paths and file names are hardcoded inside the program, e.g. the path to the sequence database or the output file. Only a few programs allow specification of a temporary folder, which is important for parallel execution. It is possible to provide precalculated sequence searches only for a few methods. Several methods are implemented relying on dependencies, sometimes on specific software versions or CPUs with a modern instruction set. Some programs are particularly eager for RAM, crashing with longer input sequences, or do not have a timeout control and execute forever. Output formats differ, with some not adequately documented. Only a few software support multi-threading and only one was submitted as a Docker container. In summary, the software implementation for disorder predictors has considerable room for improvement in order to allow their use in practice.

Discussion

The problem of predicting protein ID is challenging for several reasons. The first is in the definition of ID, which is a term that indicates that the sequence of a protein does not encode a stable structural state that is ordered. Defining ID as a property that a protein does not have (i.e. order) implies that many conformational states fit the definition, covering a continuum between fully disordered states and folded states with long dynamic regions[36,37]. The second problem, which follows from the first, is that we do not yet have a consensus reference experimental method, or set of experimental methods, to characterize ID (as we have X-ray crystallography to define ordered structures), so we lack a clear operational definition of ID. The third problem is that disorder is more context-dependent than structure, so it is

strictly speaking not intrinsic, since it is not determined by the sequence alone. Some proteins remain unfolded until they bind a partner[38], others are disordered as long as they are in a specific cellular compartment and fold upon translocation[27], and some enzymes undergo order-to-disorder-to-order transition as part of their catalytic cycle[39].

Given these challenges, the CAID project represents a community effort to develop and implement evaluation strategies to assess: (1) clear definitions of ID and (2) the performance of methods to predict ID. Concerning (1), in its first round, CAID leverages the community-driven DisProt database[17], a repository of manually curated experimental evidence of disorder, to blind test disorder predictors. We anticipate that future rounds may include reference data coming from ever-improving consensus operational definitions, as for example those coming from NMR measurements, which are particularly powerful in characterising experimentally protein disorder. For example, one could define disordered regions as those that exhibit high conformational variability under physiological conditions using multiple orthogonal measures. Concerning (2), this kind of challenge was tried from 5th to 10th editions of CASP but was abandoned due to the lack of good reference data. Our intent with CAID is to prompt experimentalists to generate ID measurements more systematically.

One of the main properties of IDPs is their ability to form many low-affinity and high-specificity interactions[40]. It is, however, challenging to predict the interacting residues of an IDPs from its sequence. Presently, multiple high-throughput experiments for the detection of interactions capable of resolving the interacting regions exist[41]. However, binding sites obtained from older experiments (e.g. CoIP, Y2H) and reported in literature often lack this grade of resolution. Furthermore, while some attempts have been made to mitigate this problem[42], a high rate of false positives plagues all experimental methods to identify binding: proteins interacting in experimental conditions do not necessarily interact in the cell under physiological physico-chemical conditions or simply due to spatio-temporal segregation[43]. DisProt annotates binding partners and interaction regions of IDPs, which we use in CAID to attempt the first assessment of binding predictors. Furthermore, there are also many binding sites that are currently uncharacterized[44].

One of the major challenges we encountered in CAID is the identification of residues that are not disordered or do not bind, in other words, the definition of negatives. Knowledge about negative results is a long-standing problem in biology[45] and is especially relevant for our assessment. Consider the example of a 100-residue long protein with an IDR spanning 20 residues. We know that 20 residues are disordered but what can we say about the remaining 80 residues? If we assume that the annotation of this protein is complete, in other words, that all disordered regions that exist under physiological conditions have been annotated, then the remaining 80 residues are structured. However, if the annotation of this protein is not complete, how do we know which regions of the protein are structured? This is even more relevant for binding regions, since it is close to impossible to map all binding partners of a protein under all possible cellular settings. To overcome this problem intrinsic to how we detect and store data, we tested the performance of ID predictors in two scenarios. In the first scenario, we assumed that all annotations were complete, thus considering all the residues outside of annotated regions as structured. In the second scenario, we used resolved residues from PDBs to annotate structure and filtered out all residues that were neither covered by disorder nor structure annotation. Binding site predictors were instead only tested on a dataset where all residues outside of binding regions are considered not-binding.

Despite these challenges, CAID revealed progress in the detection of ID from sequence and highlighted that there is still scope for improving both disorder- and binding-site predictors. One of the primary goals of CAID was to determine whether automated algorithms perform better than naive assumptions on the structural state based on indirect sources, such as sequence conservation or three-dimensional structure. As far as ID is concerned, the performance of predictors in comparison to naive baselines largely depends on the assumption made on non-disordered residues. On the *DisProt-PDB* dataset, where disorder is inferred from DisProt annotation and order inferred from the presence of a PDB

structure and all other residues are filtered out, naive baselines outperform predictors. However, when only DisProt annotations are considered (*DisProt* dataset) tables are turned, and predictors, while obtaining lower overall scores, outperform naive baselines (Fig. 2 & 3). When “uncertain” residues are kept in the analysis (*DisProt* dataset), the number of false positives increases and, as a result, precision plunges, lowering the F-score. This means that predictors detect ID in the “uncertain” residues, indicating that either the DisProt annotation is incomplete or the predictors over-predict. Naive baselines are outperformed by predictors since they predict all “uncertain” residues as disordered, which are all counted as false positives. This suggests that whenever structural information is present, it is more reliable than predictors, which however do a good job at discerning disordered from structured residues.
An interesting particular case of disorder prediction is how predictors behave with DisProt targets whose annotations cover all (or almost all) residues, i.e. fully disordered proteins. This case is compelling because usually predictors are not trained on these examples. Predictors vastly outperform naive baselines in these cases due to their large over-prediction. The count of false positives puts baselines at a disadvantage, compensating for their low count of false negatives. Naive PDB classifies a protein as fully disordered whenever no structure is available for that protein. However, the absence of a protein from PDB may be simply due to the lack of studies on that protein. Naive Gene3D performs better since Gene3D models generalize from existing structures but still tends to over-predict disorder (or under-predict order). At the opposite side of the spectrum, methods that are too conservative in their disorder classification perform worse than expected on fully disordered proteins, for example, MobiDB-lite, which is conservative by design. While top-performing methods are able to achieve a certain balance between under- and over-prediction, it is interesting to notice how they are not able to identify all fully disordered targets: not even methods that trade-off specificity to increase their detection of relevant cases are able to attain full sensitivity. This confirms that predictors are not trained on this particular class of proteins and suggests that they have room for improvement in this direction.
The challenge of binding site identification is the first attempt at assessing binding predictors. As discussed above, this is intrinsically difficult due to the complex nature of this phenomenon and how it is detected and stored. While we are aware of these difficulties, we still think that an assessment is useful for researchers who either use or develop binding predictors. Furthermore, while it is arguable that this evaluation has limitations, its publication helps highlight such constraints and thus exposing this problem to the rest of the scientific community. We compared predictors to the same baselines used for the disorder challenge but, while their design remains unchanged, their underlying naive assumption changes slightly. The NaivePDB baseline assumes that whatever is not covered by a structural annotation in PDB, is not only disordered but also involved in one or more interactions. When considering all targets in the CAID dataset, including those which are not annotated as binders, predictors slightly outperform the baselines, but have limited performance overall. Figure 4 shows a disagreement with the *Binding-all* reference in both the positive and the negative classification, highlighting the potential for improvement of binding predictors. We have to consider that the dataset used is strongly unbalanced. Although a prominent function of IDPs/IDRs is to mediate protein-protein interactions, most of the targets (414 of 646) do not contain an identified binding region and targets that do include binding regions often have them spanning the whole disordered region in which they are found. This strong bias is due to how DisProt was annotated in the past, with the label “binding” being associated with an entire IDR. In the latest version of DisProt this annotation style has been dropped in favor of a more detailed one, resulting in future editions of CAID being less biased towards long binding regions. The improved definition of the boundaries of disordered binding regions could favour methods that were trained specifically to recognize shorter binding regions. In all, this indicates that there is large growth potential in both the models and the reference building for this challenge.
In conclusion, the CAID experiment has provided the first fully blind assessment of ID predictors in almost a decade and the first-ever of ID binding regions. The results are encouraging, showing that the

methods are mature enough to be useful but significant room for improvement remains. As the quality of ID data improves, we expect predictors to become more accurate and reliable.

Author contributions

MN and DP collected the predictions, produced the data, carried out the assessment and wrote the initial manuscript. SCET designed the experiment, guided the overall project and edited the manuscript. The CAID Predictors contributed prediction methods for assessment. All authors contributed to the discussions and writing of the manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

The development of the predictors was supported in part by National Science Foundation (grant 1617369), Natural Sciences and Engineering Research Council of Canada (grant 298328), Tianjin Municipal Science and Technology Commission (grant 13ZCZDGX01099), National Natural Science Foundation of China (grants 31970649 and 11701296), Natural Science Foundation of Tianjin (grant 18JCYBJC24900), Japan Agency for Medical Research and Development (grant 16cm0106320h0001), Australian Research Council (ARC) [DP180102060], Research Foundation Flanders (FWO) - project nr. G.0328.16N, Agence Nationale de la Recherche (ANR-14-CE10-0021, ANR-17-CE12-0016). The state task “Bioinformatics and proteomics studies of proteins and their complexes” № 0115-2019-004 for OG and ML. ZD acknowledges funding from ELTE Thematic Excellence Programme (ED-18-1-2019-0030) supported by the Hungarian Ministry for Innovation and Technology and the “Lendület” grant from the Hungarian Academy of Sciences (LP2014-18). PT acknowledges grants K124670 and K131702 from the Hungarian Scientific Research Fund (OTKA). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 778247, the Italian Ministry of University and Research PRIN 2017 grant 2017483NH8 and ELIXIR, the European infrastructure for biological data.

CAID Predictors:

Tamjidul Hoque; Computer Science, University of New Orleans, USA

Ian Walsh; Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), Singapore

Sumaiya Iqbal; Broad Institute of MIT and Harvard, USA

Michele Vendruscolo, Pietro Sormanni; Department of Chemistry, University of Cambridge, UK

Chen Wang; Columbia University, USA

Daniele Raimondi; ESAT-STADIUS, KU Leuven, Belgium

Ronesh Sharma; Fiji National University, Suva, Fiji

Yaoqi Zhou, Thomas Litfin; Institute for Glycomics and School of Information and Communication Technology, Griffith University, Australia

Oxana V. Galzitskaya, Michail Yu. Lobanov; Institute of Protein Research; Russian Academy of Sciences, Russia

Wim Vranken; Interuniversity Institute of Bioinformatics in Brussels, Vrije Universiteit Brussel, Belgium

Björn Wallner, Claudio Mirabello; Linköping University

Nawar Malhis; Michael Smith Laboratories, University of British Columbia, Canada

Zsuzsanna Dosztányi, Gábor Erdős, Bálint Mészáros; Department of Biochemistry, Eötvös Loránd University, Hungary

Jianzhao Gao, Kui Wang, Gang Hu, Zhonghua Wu; Nankai University, China

Alok Sharma; RIKEN Center for Integrative Medical Sciences, Japan; Griffith University, Australia

Jack Hanson, Kuldip Paliwal; School of Engineering and Built Environment, Griffith University, Australia

Isabelle Callebaut, Tristan Bitard-Feildel; Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, Paris, France

Gabriele Orlando; VIB-KU Leuven, 3000 Leuven, Belgium

Zhenling Peng; Tianjin University, China

Jinbo Xu, Sheng Wang; Toyota Technological Institute at Chicago, USA

David T. Jones, Domenico Cozzetto; University College London, UK

Fanchi Meng, Jing Yan; University of Alberta, Canada

Joerg Gsponer; University of British Columbia

Jianlin Cheng, Tianqi Wu; University of Missouri, Columbia

Lukasz Kurgan; Virginia Commonwealth University, USA

DisProt Curators:

Vasilis J Promponas, Stella Tamana; Department of Biological Sciences, University of Cyprus, Cyprus

Cristina Marino-Buslje, Elizabeth Martínez-Pérez; Fundación Instituto Leloir, Argentina

Anastasia Chasapi, Christos Ouzounis; Chemical Process & Energy Resources Institute, Centre for Research & Technology Hellas, Greece

A. Keith Dunker; Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, USA

Andrey V Kajava, Jeremy Y Leclercq; Centre de Recherche en Biologie cellulaire de Montpellier (CRBM), UMR 5237 CNRS, Université Montpellier, France

Burcu Aykac-Fas, Matteo Lambrughi, Emiliano Maiani, Elena Papaleo; Danish Cancer Society Research Center, Denmark

Lucia Beatriz Chemes, Lucía Álvarez, Nicolás S González Foutel; Consejo Nacional de Investigaciones Científicas y Técnicas. Instituto de Investigaciones Biotecnológicas IIBIO, Universidad Nacional de San Martín, Argentina

Valentin Iglesias, Jordi Pujols, Salvador Ventura; Departament de Bioquimica i Biologia Molecular and Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Spain

Nicolás Palopoli, Guillermo Ignacio Benítez, Gustavo Parisi; Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes - CONICET, Argentina

Claudio Bassot, Arne Elofsson, Sudha Govindarajan, John Lamb, Marco Salvatore; Department of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, Sweden

András Hatos, Alexander Miguel Monzon, Martina Bevilacqua, Ivan Mičetić, Giovanni Minervini, Lisanna Paladin, Federica Quaglia; Department of Biomedical Sciences, University of Padova, Italy

Emanuela Leonardi; Department of Woman and Child Health, University of Padova, Italy

Norman Davey; Division of Cancer Biology, The Institute of Cancer Research, UK

Tamas Horvath, Orsolya Panna Kovacs, Nikoletta Murvai, Rita Pancsa, Eva Schad, Beata Szabo, Agnes Tantos; Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Hungary

Sandra Macedo-Ribeiro, Jose A Manso, Pedro José Barbosa Pereira; Instituto de Biologia Molecular e Celular (IBMC) and Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Portugal

Radoslav Davidović, Nevena Veljkovic; Institute of Nuclear Sciences Vinca, University of Belgrade, Serbia

Borbála Hajdu-Soltész, Mátyás Pajkos, Tamás Szaniszló; Department of Biochemistry, Eötvös Loránd University, Hungary

Mainak Guharoy, Tamas Lazar, Mauricio Macossay-Castillo, Peter Tompa; Structural Biology Brussels, Vrije Universiteit Brussel (VUB), Belgium

References

1. Tompa, P. & Fersht, A. *Structure and Function of Intrinsically Disordered Proteins*. (CRC Press, 2009).
2. Dunker, A. K., Bondos, S. E., Huang, F. & Oldfield, C. J. Intrinsically disordered proteins and multicellular organisms. *Cell Signal. Div. Organ. Mediat. Intrinsically Disord. Proteins Protein Tyrosine Phosphatases* **37**, 44–55 (2015).
3. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
4. Cheng, Y. *et al.* Rational drug design via intrinsically disordered protein. *Trends Biotechnol.* **24**, 435–442 (2006).
5. Uversky, V. N. Intrinsically disordered proteins and novel strategies for drug discovery. *Expert Opin. Drug Discov.* **7**, 475–488 (2012).
6. Plitzko, J. M., Schuler, B. & Selenko, P. Structural Biology outside the box-inside the cell. *Curr. Opin. Struct. Biol.* **46**, 110–121 (2017).
7. Tompa, P. Unstructural biology coming of age. *Curr. Opin. Struct. Biol.* **21**, 419–425 (2011).
8. Holmstrom, E. D., Nettels, D. & Schuler, B. Conformational Plasticity of Hepatitis C Virus Core Protein

Enables RNA-Induced Formation of Nucleocapsid-like Particles. *J. Mol. Biol.* **430**, 2453–2467 (2018).
9. Felli, I. C. & Pierattelli, R. *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*. (Springer, 2015).
10. Williams, R. J. The conformational mobility of proteins and its functional significance. *Biochem. Soc. Trans.* **6**, 1123–1126 (1978).
11. Liu, Y., Wang, X. & Liu, B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* **20**, 330–346 (2019).
12. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208 (2006).
13. Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).
14. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).
15. Mészáros, B., Erdős, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* doi:10.1093/nar/gky384.
16. Walsh, I., Martin, A. J. M., Di Domenico, T. & Tosatto, S. C. E. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509 (2012).
17. Hatos, A. *et al.* DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* doi:10.1093/nar/gkz975.
18. Piovesan, D. *et al.* MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.* **46**, D471–D476 (2018).
19. Fukuchi, S. *et al.* IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.* **42**, D320-325 (2014).
20. Schad, E. *et al.* DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinforma. Oxf. Engl.* **34**, 535–537 (2018).
21. Fichó, E., Reményi, I., Simon, I. & Mészáros, B. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinforma. Oxf. Engl.* **33**, 3682–3684 (2017).
22. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
23. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
24. Velankar, S. *et al.* PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.* **44**, D385–D395 (2016).
25. van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
26. Xie, H. *et al.* Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* **6**, 1882–1898 (2007).
27. Jakob, U., Kriwacki, R. & Uversky, V. N. Conditionally and Transiently Disordered Proteins: Awakening Cryptic Disorder To Regulate Protein Function. *Chem. Rev.* (2014) doi:10.1021/cr400459c.
28. Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A. & Fidelis, K. Assessment of protein disorder region predictions in CASP10. *Proteins* **82 Suppl 2**, 127–137 (2014).
29. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).
30. Necci, M., Piovesan, D. & Tosatto, S. C. E. Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. *Protein Sci. Publ. Protein Soc.* **25**, 2164–2174 (2016).
31. Deiana, A., Forcelloni, S., Porrello, A. & Giansanti, A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PloS One* **14**, e0217889 (2019).
32. Kumar, M. *et al.* ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* doi:10.1093/nar/gkz1030.
33. Peng, Z. & Kurgan, L. High-throughput prediction of RNA, DNA and protein binding regions

mediated by intrinsic disorder. *Nucleic Acids Res.* **43**, e121 (2015).
34. Malhis, N., Jacobson, M. & Gsponer, J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res.* **44**, W488-493 (2016).
35. Sharma, R., Raicar, G., Tsunoda, T., Patil, A. & Sharma, A. OPAL: Prediction of MoRF regions in intrinsically disordered protein sequences. *Bioinformatics* doi:10.1093/bioinformatics/bty032.
36. Forman-Kay, J. D. & Mittag, T. From Sequence and Forces to Structure, Function, and Evolution of Intrinsically Disordered Proteins. *Structure* **21**, 1492–1499 (2013).
37. Sormanni, P. *et al.* Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.* **13**, 339–342 (2017).
38. Dyson, H. J. & Wright, P. E. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60 (2002).
39. Bahar, I., Chennubhotla, C. & Tobi, D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr. Opin. Struct. Biol.* **17**, 633–640 (2007).
40. Van Roey, K. *et al.* Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chem. Rev.* **114**, 6733–6778 (2014).
41. Blikstad, C. & Ivarsson, Y. High-throughput methods for identification of protein-protein interactions involving short linear motifs. *Cell Commun. Signal.* **13**, 38 (2015).
42. Vidalain, P.-O., Boxem, M., Ge, H., Li, S. & Vidal, M. Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* **32**, 363–370 (2004).
43. Scott, J. D. & Pawson, T. Cell signaling in space and time: where proteins come together and when they're apart. *Science* **326**, 1220–1224 (2009).
44. Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A million peptide motifs for the molecular biologist. *Mol. Cell* **55**, 161–169 (2014).
45. Mehta, D. Highlight negative results to improve science. *Nature* (2019) doi:10.1038/d41586-019-02960-3.
46. Zhou, J., Oldfield, C. J., Yan, W., Shen, B. & Dunker, A. K. Identification of Intrinsic Disorder in Complexes from the Protein Data Bank. *ACS Omega* **5**, 17883–17891 (2020).
47. Monzon, A. M. *et al.* Experimentally Determined Long Intrinsically Disordered Protein Regions Are Now Abundant in the Protein Data Bank. *Int. J. Mol. Sci.* **21**, 4496 (2020).
48. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).
49. Piovesan, D. & Tosatto, S. C. E. Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures. *Bioinformatics* doi:10.1093/bioinformatics/btx592.
50. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).

Figures

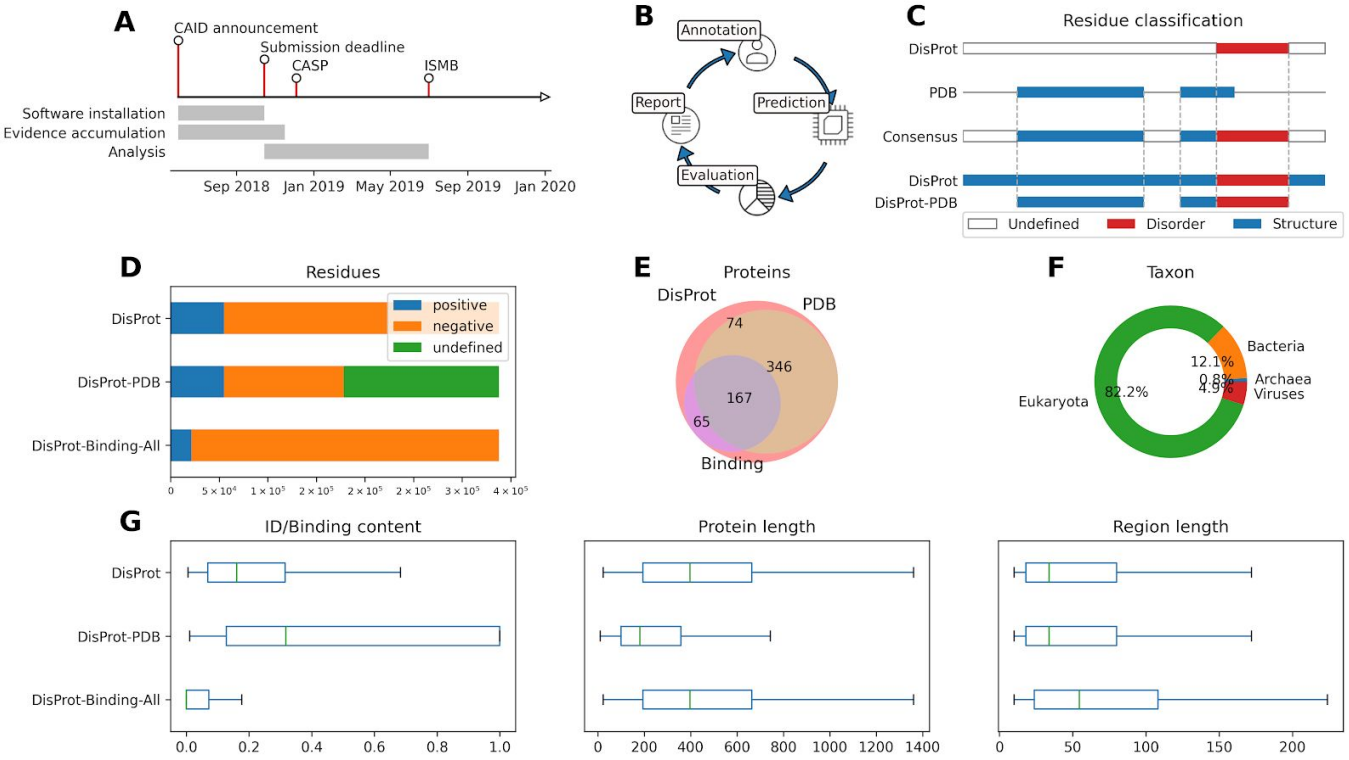

Figure 1 - (A) CAID timeline: phases of CAID from June 2018 to today. (B) CAID process: Iterative process of the CAID experiment (C) Residue classification strategy for the *DisProt* and *DisProt-PDB* references. (D) The number of residues for each class in different references. (E) The number of proteins for each set of annotations that they contain. (F) The numerosity of the representatives for each taxon (G) Positive label (disorder/binding) fraction in the dataset (left), distribution of protein length (middle) and region length (right) in the different reference sets.

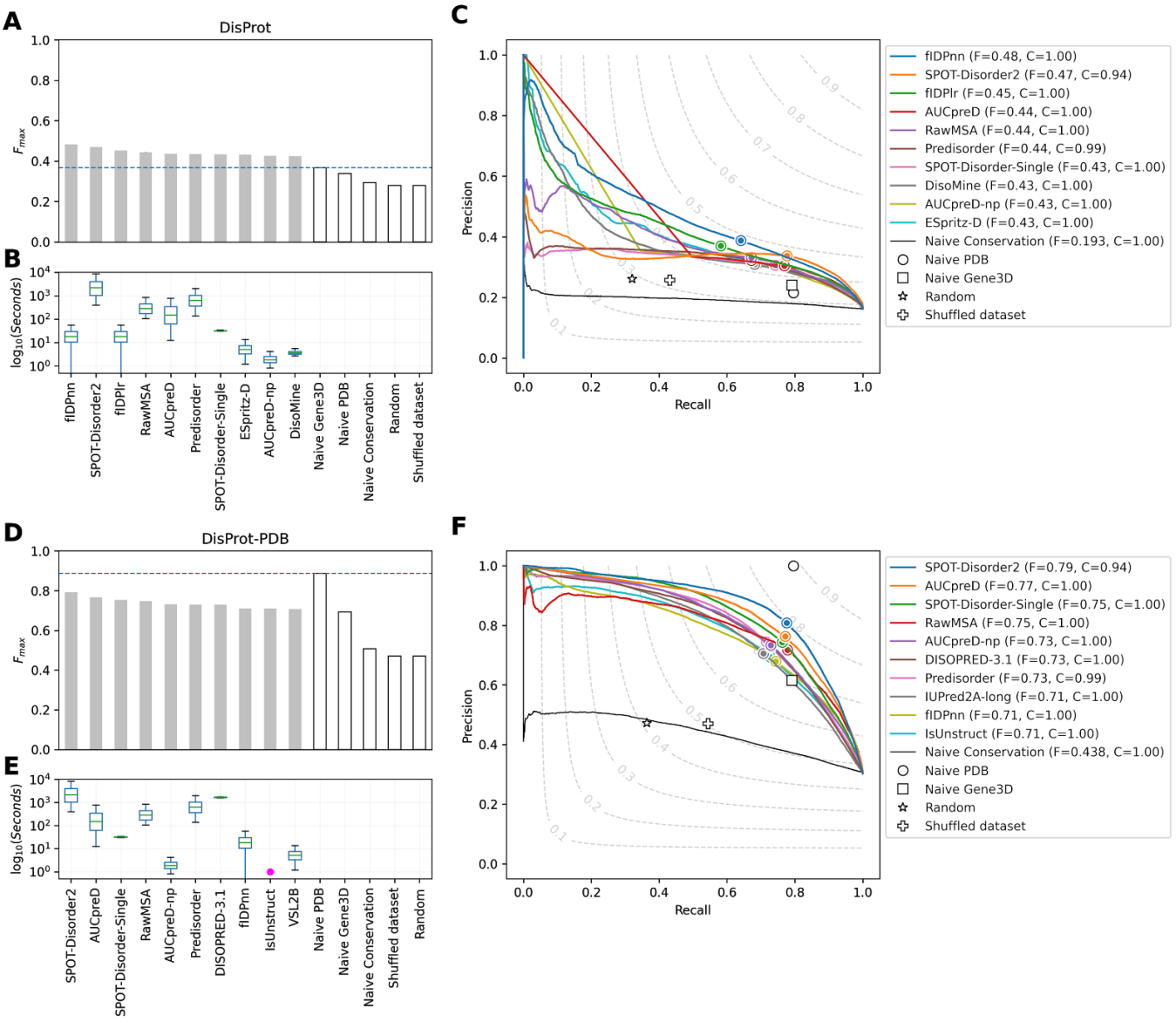

Figure 2 - Performance and CPU times for the ten top-ranking methods. Performance of predictors expressed as F_{max} (A, D), calculated at the dataset level, for the top ten best ranking methods (light gray) and baselines (dark gray) and the distribution of execution time per-target (B, E) when using *DisProt* (A, B) and *DisProt-PDB* datasets (D, E). The horizontal line in (A, D) indicates the F_{max} of the best baseline. Precision-recall curves of ten top-ranking methods and baselines using *DisProt* (C) and DisProt-PDB (F) dataset, with level curves of the F1-scores. Magenta dots on execution time plots (B, E) indicate an execution time below 1 second.

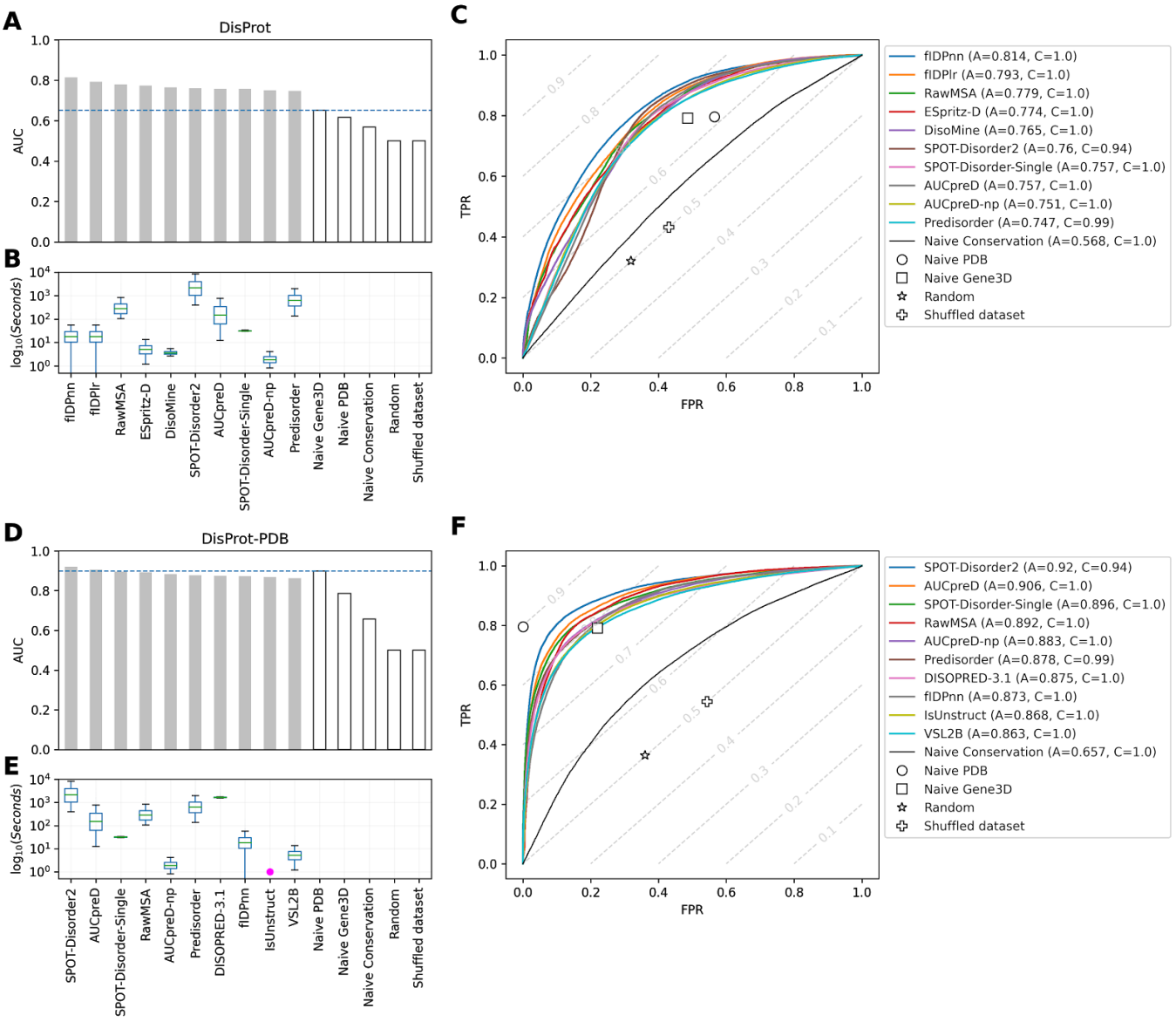

Figure 3 - Performance and CPU times for the ten top-ranking methods. Performance of predictors expressed as AUC (A, D) for the top ten best ranking methods (light gray) and baselines (dark gray) and the distribution of execution time per-target (B, E) when using *DisProt* (A, B) and *DisProt-PDB* datasets (D, E). The horizontal line in (A, D) indicates the AUC of the best baseline. ROC curves of ten top-ranking methods and baselines using *DisProt* (C) and DisProt-PDB (F) dataset, with level curves of the Balanced accuracy. Magenta dots on execution time plots (B, E) indicate an execution time below 1 second.

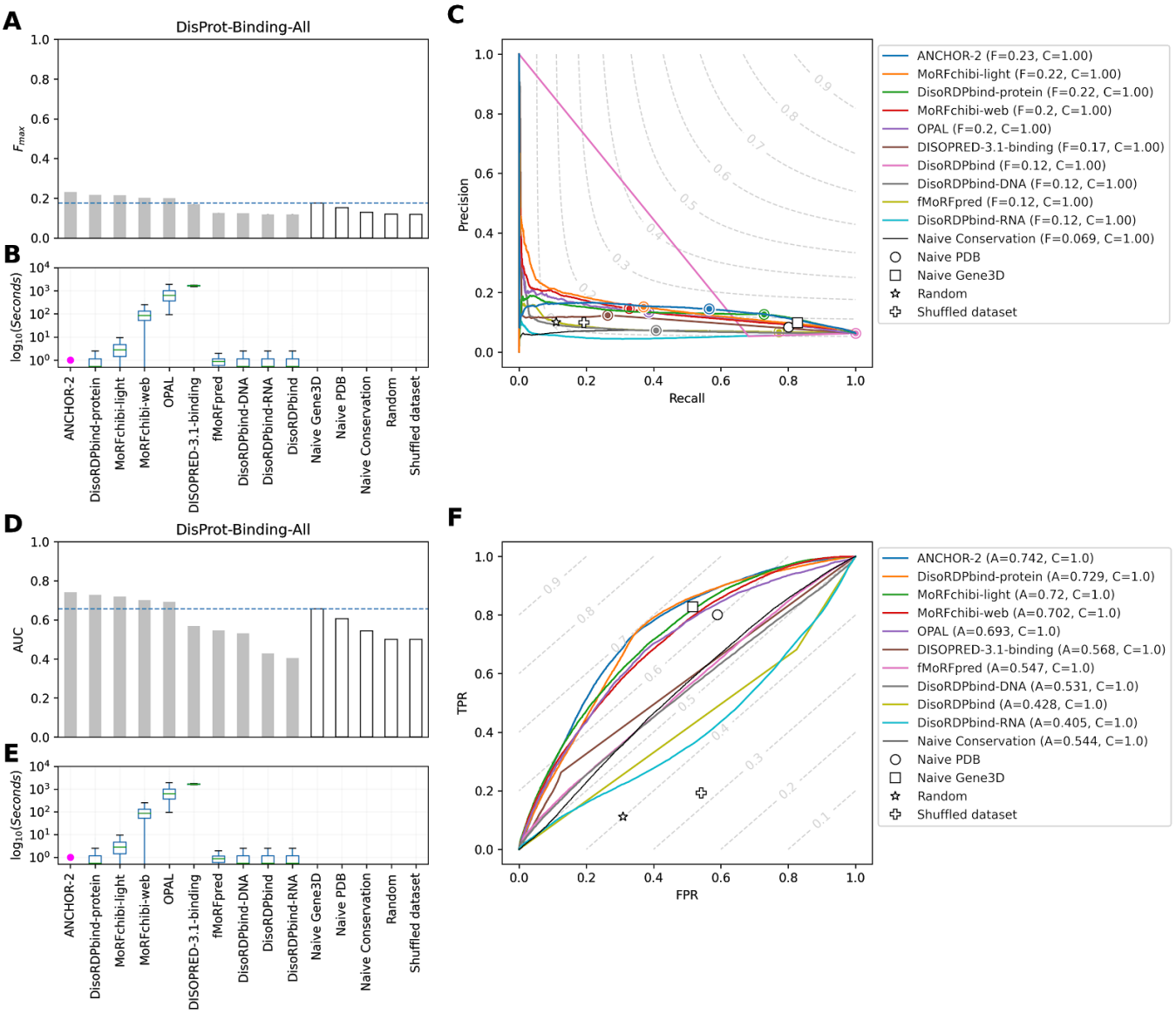

Figure 4 - Performance and CPU times for the ten top-ranking methods. Performance of predictors expressed as F_{max} (A) and AUC (D) for the top ten best ranking methods (light gray) and baselines (dark gray) and the distribution of execution time per-target (B, E) using *DisProt-Binding-All* dataset. The horizontal line in (A, D) indicates the F_{max} and AUC of the best baseline, respectively. Precision-Recall (C) and ROC curves (F) of ten top-ranking methods and baselines using *DisProt-Binding-All* dataset, with level curves of the F1-score and Balanced accuracy, respectively. Magenta dots on execution time plots (B, E) indicate an execution time below 1 second.

Tables

	DisProt-Full-ID (IDC>95%)									
	tn	**fp**	**fn**	**tp**	**mcc**	**f1s**	**tnr**	**tpr**	**ppv**	**bac**
flDPnn	527	16	15	26	0.598	0.627	0.971	0.634	0.619	0.802
RawMSA	524	19	15	26	0.574	0.605	0.965	0.634	0.578	0.800
VSL2B	520	23	18	23	0.492	0.529	0.958	0.561	0.500	0.759
Predisorder	531	12	22	19	0.503	0.528	0.978	0.463	0.613	0.721
flDPlr	508	35	14	27	0.493	0.524	0.936	0.659	0.435	0.797
IUPred-long	537	6	26	15	0.487	0.484	0.989	0.366	0.714	0.677
AUCpreD	530	13	24	17	0.452	0.479	0.976	0.415	0.567	0.695
SPOT-Disorder1	514	29	19	22	0.437	0.478	0.947	0.537	0.431	0.742
SPOT-Disorder-Single	536	7	26	15	0.474	0.476	0.987	0.366	0.682	0.676
SPOT-Disorder2	516	27	20	21	0.430	0.472	0.950	0.512	0.438	0.731
DisoMine	493	50	13	28	0.444	0.471	0.908	0.683	0.359	0.795
IsUnstruct	530	13	25	16	0.431	0.457	0.976	0.390	0.552	0.683
IUPred2A-long	537	6	28	13	0.441	0.433	0.989	0.317	0.684	0.653
Naive Gene3D	462	81	8	33	0.423	0.426	0.851	0.805	0.289	0.828
ESpritz-N	539	4	29	12	0.447	0.421	0.993	0.293	0.750	0.643
PyHCA	538	5	29	12	0.431	0.414	0.991	0.293	0.706	0.642
MobiDB-lite	541	2	30	11	0.458	0.407	0.996	0.268	0.846	0.632
JRONN	537	6	29	12	0.416	0.407	0.989	0.293	0.667	0.641
ESpritz-D	497	46	19	22	0.360	0.404	0.915	0.537	0.324	0.726
IUPred-short	542	1	31	10	0.455	0.385	0.998	0.244	0.909	0.621
IUPred2A-short	541	2	31	10	0.433	0.377	0.996	0.244	0.833	0.620
DisPredict-2	528	15	28	13	0.346	0.377	0.972	0.317	0.464	0.645
S2D-2	514	29	26	15	0.302	0.353	0.947	0.366	0.341	0.656
Naive PDB	428	115	10	31	0.321	0.332	0.788	0.756	0.212	0.772
AUCpreD-np	532	11	31	10	0.307	0.323	0.980	0.244	0.476	0.612
ESpritz-X	537	6	32	9	0.337	0.321	0.989	0.220	0.600	0.604
S2D-1	491	52	25	16	0.235	0.294	0.904	0.390	0.235	0.647
FoldUnfold	398	145	10	31	0.272	0.286	0.733	0.756	0.176	0.745
DISOPRED-3.1	538	5	35	6	0.258	0.231	0.991	0.146	0.545	0.569
DisEMBL-HL	543	0	37	4	0.302	0.178	1.000	0.098	1.000	0.549
DisEMBL-465	543	0	39	2	0.213	0.093	1.000	0.049	1.000	0.524
Naive Conservation	407	136	35	6	-0.062	0.066	0.750	0.146	0.042	0.448
DFLpred	543	0	41	0	0.000	0.000	1.000	0.000	0.000	0.500

GlobPlot	543	0	41	0	0.000	0.000	1.000	0.000	0.000	0.500
DynaMine	543	0	41	0	0.000	0.000	1.000	0.000	0.000	0.500

Table 1: Confusion matrix and metrics for the prediction of fully disordered proteins on *DisProt* dataset. Proteins with disorder prediction covering at least 95% of the sequence are considered fully disordered. Proteins in DisProt with disorder annotation covering at least 95% of the sequence are considered fully disordered. Predictors are sorted by their F1-Score.

Methods

All software programs were executed using a homogeneous cluster of nodes running Ubuntu 16.04 on Intel 8 core processors with 16 GB of RAM and a mechanical harddisk. In the text we refer to proteins as targets, to disordered residues as positive labels and structured/ordered residues as negative labels.

Reference sets

In CAID different reference sets were built, differing in the subset of DisProt used to define positive labels and in the definition of negatives labels.

For the disorder challenge, we generated two reference sets called *DisProt* and *DisProt-PDB*. Both references are composed of a set of 646 targets, annotated between June 2018 and November 2018 (DisProt release 2018_11). Positive labels in both reference sets are those residues annotated as disordered in the DisProt database. In the *DisProt* reference set, all labels that are not positive are assigned as negatives. In the *DisProt-PDB* set, PDB structures mapping on the protein sequence define negative labels. All residues that are not covered by either DisProt annotation or PDB structures, are masked and excluded from the analysis. It should be noted that a fraction of resolved structures in the PDB have been annotated as disordered[46,47]. While in this edition we decided to consider any resolved residue as structured, we plan to apply a filtering on the next editions of CAID. This problem will become less and less relevant as DisProt annotations will become more complete, since disorder always overwrites structure.

For the binding challenge we generated a reference sets that we called *DisProt-Binding-All.* Positive labels are those residues annotated as binding in the DisProt database, whereas all labels that are not positive are assigned as negatives. Notice that 232 targets have at least one annotation of binding in the DisProt database. *DisProt-Binding-All* is composed of all 646 targets considered in the analysis, hence the majority of targets (i.e. 646 - 232 = 414) do not contain positive labels.

Predictions

Most predictors output a series of score and state pairs per residue of the input sequence. Scores are floating point numbers normalized to the range [0,..,1], while states are binary labels predicting if a residue is in a disordered or structured state. If scores are missing, states will be used as scores. If states are missing, they are generated by applying a threshold to scores. When a threshold is not available, it defaults to 0.5. The rationale is that if only scores are provided without a threshold, they are interpreted as a probability density function with normal distribution between 0 and 1 and average on 0.5. This can be done since: if states are provided, scores values associated with positive and negative classifications are indicated by the states. Default threshold is always inferred from states. This ensures correct default threshold estimates for any distribution of scores. Prediction scores are rounded to the third decimal figure. This sets the number of possible thresholds to 1000. Bootstrapping samples the whole dataset with replacements 1000 times. Re-sampling is done at the label (residue) level. Confidence intervals are calculated on Student's T distribution at alpha set to 0.05.

Baselines

A number of baseline predictors have been built in order to be compared with actual predictors. Two are based on the reshuffling of a reference (*Shuffled targets* and *Shuffled dataset*), one is completely random (*Random*), one is based on an estimate of residue conservation through evolution (*Naive Conservation*), two are based on taking the reverse of the annotation of structure as positive labels (*Naive PDB* and Naive *Gene3D*).

Shuffled baselines are built by reshuffling a reference. A reference is reshuffled at the dataset level (*Shuffled dataset*) or at the target level (*Shuffled targets*). The former is a random classification that preserves the proportion of positive labels in the dataset. The latter is a random classification that preserves the proportion of positive labels in each target.

Random baseline is a random classifier in which the prediction score of each label is assigned randomly. It is built by randomly drawing floating point numbers out of a uniform distribution [0,..,1] and applying a threshold of 0.5.

Naive Conservation baseline is built by calculating the distance between the PSSM of each target in DisProt and the residue distribution of the BLOSUM62 substitution matrix. Target PSSMs are obtained by launching a 3 cycle PsiBLAST against UniRef90. The distance is calculated as the Jensens-Shannon divergence[48] of the two frequencies.

Naive baselines are based on the assumption that whatever is not annotated as structure is disorder. Two naive baselines are following this principle, *Naive PDB* has the structure annotation defined by PDB structures (October 2019), *Naive Gene3D* has structure annotations defined by Gene3D predictions (Gene3D version 4.2.0, calculated with InterProScan-5.38-76.0). The former uses PDB data mapped on UniProt sequences by Mobi 2.0[49], the latter uses Gene3D predictions generated by InterProScan.

Target and Dataset metrics

Metrics were calculated following two strategies: dataset and target. In the dataset strategy, all targets (proteins) reference classifications and prediction classifications are concatenated in two single arrays. Confusion matrix and subsequent evaluation metrics are calculated once comparing these arrays. In the target strategy confusion matrix and subsequent evaluation metrics are calculated for each target (protein) and the mean value of the evaluation metrics is taken. The former strategy is equivalent to summing together the confusion matrices for each target and computing evaluation metrics on the resulting confusion matrix, while the latter strategy is equivalent to calculating the evaluation metrics on the average of the confusion matrices of the targets.

Notes on evaluation metrics calculation

Throughout the manuscript, F_{max} and AUC are the main assessment criteria used. F_{max} is the maximum point in the precision/recall curve. AUC is the area under the ROC curve. Additional metrics are used for comparison and all follow standard definitions as described in the Supplementary Material. Fbeta (0.5, 1, 2) and MCC are set to 0 if the denominator is 0. Since the MCC denominator is a multiplication of the number of positive classifications, negative classifications, positive labels in the reference and negative labels in the reference, if any of these classes amounts to 0, we set MCC to 0. This means that for fully disordered proteins and for proteins predicted to be fully disordered or fully ordered, MCC is 0. This situation is very likely in target-strategy with DisProt-PDB dataset and it explains why the MCC for target-strategy is much lower than that for dataset-strategy (see Supplementary Material). This effect can also be seen in the heatmap of targets MCC where a large number of targets have MCC = 0.

Assessors policy

Prediction methods published by the assessors did not take part in the challenges. Their methods may be included for reference only.