FLIPPER: predicting and characterizing linear interacting peptides in the Protein Data Bank

Alexander Miguel Monzon[1], Paolo Bonato[1], Marco Necci[1], Silvio C.E. Tosatto[1,*], Damiano Piovesan[1]

[1]Dept. of Biomedical Sciences, University of Padua, Via Ugo Bassi 58/B, Padua, 35121, Italy

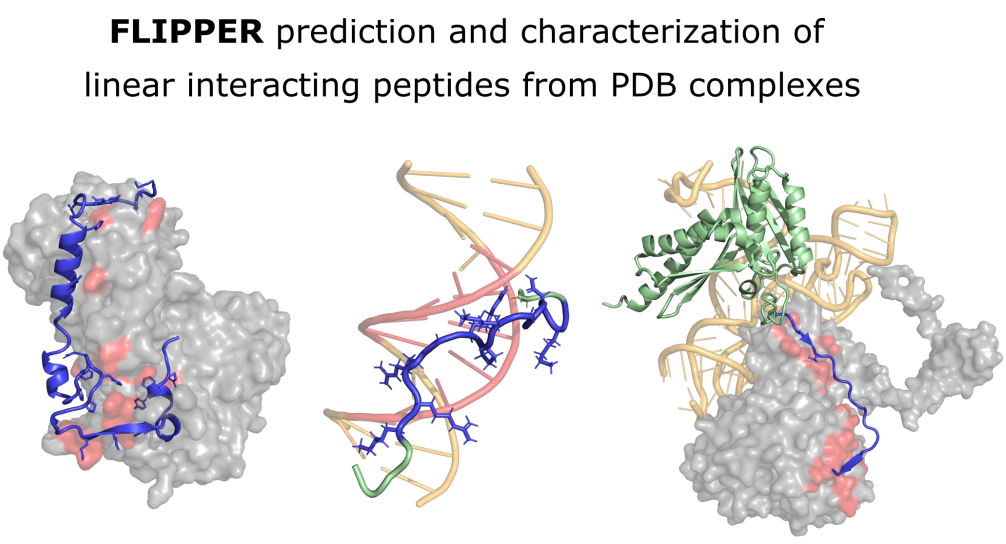
*To whom correspondence should be addressed. E-mail silvio.tosatto@unipd.it; Tel: +39 049 827 6269

Abstract

A large fraction of peptides or protein regions are disordered in isolation and fold upon binding. These regions, also called MoRFs, SLiMs or LIPs, are often associated with signaling and regulation processes. However, despite their importance, only a limited number of examples are available in public databases and their automatic detection at the proteome level is problematic. Here we present FLIPPER, an automatic method for the detection of structurally linear sub-regions or peptides that interact with another chain in a protein complex. FLIPPER is a random forest classification that takes the protein structure as input and provides the propensity of each amino acid to be part of a LIP region. Models are built taking into consideration structural features such as intra- and inter-chain contacts, secondary structure, solvent accessibility in both bound and unbound state, structural linearity and chain length. FLIPPER is accurate when evaluated on non-redundant independent datasets, 99% precision and 99% sensitivity on PixelDB-25 and 87% precision and 88% sensitivity on DIBS-25. Finally, we used FLIPPER to process the entire Protein Data Bank and identified different classes of LIPs based on different binding modes and partner molecules. We provide a detailed description of these LIP categories and show that a large fraction of these regions are not detected by disorder predictors. All FLIPPER predictions are integrated in the MobiDB 4.0 database.

FLIPPER software is available at URL: https://github.com/BioComputingUP/FLIPPER.

Graphical Abstract



Keywords

Linear interacting peptides; intrinsic disorder; protein structure; binding modes prediction; machine learning.

Abbreviations

PDB, Protein Data Bank; LIP, Linear Interacting Peptide; IDP, Intrinsically Disordered Protein; IDR, Intrinsically Disordered Region; RSA, Relative Solvent accessibility; TPR, True Positive Rate, or recall, or sensitivity; TNR, True Negative Rate, or specificity; FPR, False Positive Rate; BAC, Balanced Accuracy; MCC, Matthews' Correlation Coefficient; PPV, Positive Predictive Value, or precision.

Highlights

- IDPs/IDRs exhibit a wide diversity of binding modes
- LIPs are functional regions with specific structural features
- FLIPPER is an accurate and fast LIP predictor from PDB structure complexes
- Different types of LIPs can be identified depending on the interaction partners
- FLIPPER provides the largest high quality data set of LIPs

Introduction

Intrinsically disordered proteins and regions (IDPs/IDRs) do not fold into a well-defined native structure but rather populate functional states defined by heterogeneous ensembles of rapidly interchanging conformations [1,2]. They play a key role in many biological processes as DNA and RNA binding, transcription, translation, cell-cycle regulation and signaling thanks to their unique binding modes, remaining unfolded or partially unfolded during interactions [3]. IDPs/IDRs binding events imply structural transitions from dynamic unbounded states to more constrained protein states [4]. In some cases, the IDPs retain a high level of flexibility and dynamism even in the bound state, in a phenomenon known as fuzziness [5–7]. Despite the same IDR can exhibit a variety of binding modes depending on the interaction partner [7,8], the majority of bound IDRs tends to adopt defined or partially defined secondary structure [1,9].

Short and structurally linear binding interfaces in IDRs are known as molecular recognition features (MoRFs), protean segments (ProS), short linear motifs (SLiMs) or linear interacting peptides (LIPs) [10–13]. These regions are crucial to cell physiology [9], but different definitions and names correspond to slightly different subtypes. MoRFs and LIPs are generally used to indicate the structural properties of these regions, thus embracing the vast majority of binding modes adopted by IDPs/IDRs. On the other end, SLiMs, also called miniMotifs [10], are strictly associated to well defined functional roles, e.g. cell signaling regulation. SLiMs are shorter (3-10 amino acids) and well conserved by convergent evolution [9,14,15]. ProS is a term only used by the IDEAL database [16] and it refers to those protein segments able to switch from disordered to ordered state upon binding.

Another well defined class of IDRs are those involved in domain-swapping and intertwined associations [17,18]. They are found in homo-oligomeric protein complexes where two or more identical chains exchange structural elements. Domain-swapping has important biological consequences such as formation of large protein aggregates and protein function modulation [18].

Different (or similar) types of LIPs are available in specialized manually curated databases. DIBS and MFIB provide examples of LIPs folding when binding a globular protein or another LIP, respectively [19,20], and collect examples directly from the Protein Data Bank (PDB) [21]. ELM collects SLiMs from different types of experiments (not only PDBs) and identifies key positions linked to function [10]. FuzDB provides examples of fuzzy interactions, e.g. binding IDRs, which preserve a disordered behaviour in the bound state [22] from the literature. DisProt and IDEAL collect information about binding regions and folding upon binding regions from the literature but they do not provide information about binding modes [16,23]. All these resources are of high quality and provide a valuable source of information for the implementation of sequence-based LIP prediction methods [24–27] [28]. However, training data cover a tiny fraction of LIPs available in the PDB and sequence predictors suffer a low sensitivity when evaluated on independent datasets [29].

In this work we present the Fast Linear Interacting Peptides Predictor (FLIPPER), to accurately detect LIPs in protein structure complexes. FLIPPER is a random forest classification that takes the protein structure as input and provides the propensity of each amino acid to be part of a LIP region. Models are built taking into consideration structural features such as intra- and inter-chain contacts, secondary structure, solvent accessibility in both bound and unbound state, structural linearity and chain length. FLIPPER provides accurate predictions and outperforms the Mobi 2.0 method [12] when evaluated on independent datasets. In this work we also used FLIPPER to scan the entire PDB. We identified and classified different LIP

flavours (types) on the basis of their interaction partners. All FLIPPER predictions are available from the MobiDB 4.0 database [30].

Results

LIP structural features

FLIPPER was trained on a set of structural features (see Methods) which are effective in discriminating LIP from non-LIP residues. Figure 1 shows the feature value distributions for LIP and non-LIP residues in the training set. Despite all distributions being significantly different when comparing positives and negatives (Kolmogorov-Smirnov P-values < 0.001) some features provide a stronger separation. Indeed, these features are intrinsically similar and correlate strongly, namely inter-chain contacts and relative solvent accessibility (Suppl. Figure S1). Feature importance from the final model is reported in Suppl. Figure S2.

LIP residues have a higher content of inter-chain over intra-chain contacts and a higher RSA and delta RSA compared to non-LIP residues. LIPs display higher structural linearity (see definition in methods), while secondary structure elements like helices and sheets are more equally distributed between LIP and non-LIP residues. As expected, LIPs show an elevated prevalence of coil due to their intrinsic flexibility and propensity to be unstructured.

FLIPPER evaluation

The evaluation on the training set and on the two independent datasets (PixelDB-25 and DIBS-25) and a comparison with the Mobi 2.0 software is provided in Table 2. Additional data about cross-validation is provided in Suppl. Table S2. Mobi 2.0 contains a module to detect LIPs in PDB structure complexes by simply evaluating the intra-/inter-chain contacts ratio [12]. FLIPPER outperforms Mobi 2.0, showing significantly higher score for all metrics except specificity. Mobi 2.0 generally provides a limited number of false positives but a lot of false negatives. The evaluation dataset is unbalanced toward negative examples and since the high number of predicted true negatives the specificity is similar to the one obtained by FLIPPER (Table 1). FLIPPER reaches an MCC of 0.99 for PixelDB-25 and 0.86 for DIBS-25, which is comparable with the one obtained in cross-validation. The overall performance on PixelDB-25 is slightly better compared to DIBS-25. This could be explained by the fact that DIBS is focused on a special class of LIPs, those within intrinsically disordered regions and that fold upon binding. Meanwhile PixelDB contains LIPs that represent a broader range of binding modes. FLIPPER and Mobi 2.0 present high and similar values of specificity, however FLIPPER shows significantly lower values of false negative rate and a better sensitivity. This indicates FLIPPER is able to spot a larger fraction of LIPs residues which remained undetected by Mobi 2.0. All performance results on filtered datasets are comparable when considering the full PixelDB and DIBS (see Suppl. Table S3).

Manually inspecting some of the obtained predictions, we saw that most of the errors are concentrated on residues flanking LIPs. There is one case that brings down the performance in cross-validation which is the PDB 1RF8, chain B (Figure 2) from the ANCHOR data set. That chain contains a LIP that completely wraps the receptor. FLIPPER correctly classifies half of the residues but misses those LIP residues which have a high number of intra-chain contacts and low linearity.

LIP flavours

FLIPPER was used to scan the entire PDB (version July 2020) and depending on the different interaction types and partners, we defined types, or flavours, of LIPs (Figure 3). The LIP flavours can be grouped in homo- and hetero-LIPs. Homo-LIPs (which correspond to the Homo flavour) interact with identical chains forming homo oligomers. Hetero-LIPs are more heterogeneous and interact with one or more different proteins, and/or nucleic acid polymers.

In order to have a better overview of the landscape of the different types of LIPs, we combined PDB predictions at the protein level. Each PDB chain was mapped to the corresponding UniProt sequence [31] by using the SIFTS service [32]. The LIP consensus definition corresponds to the union of all LIP regions detected in different PDB chains mapping to the same protein sequence.

LIPs predicted in the PDB and consensus statistics are provided in Table 3. FLIPPER identified a total of 20,009 different PDBs complexes with 65,020 PDB chains containing at least one LIP. FLIPPER predicted a total of 77,579 LIPs with a median length of 14 residues and minimum length of five residues (PDB, Table 3). A total of 12,910 LIPs in 8,661 different proteins were identified using the consensus strategy (Consensus, Table 3). Around 26.4% (2,284) of the proteins have more than one LIP and 89.5% (7,757) proteins have LIPs of only one flavour (Suppl. Figure S3). The most abundant flavours are Proteins and Homo. A few LIPs interact with DNA/RNA (Nucleic Acids) and another fraction of LIPs interact with both DNA/RNA and proteins (Mixed), for example those PDB chains which are part of ribosomes and transcription factors. The Homo and Protein flavours represent 53.8% and 30.6% of the entire dataset respectively, while only 843 (9.7%) proteins have LIPs interacting with nucleic acids. The longest consensus LIPs are in the Mixed class, followed by Protein, Homo and Nucleic Acids classes (Consensus, Table 3 and Suppl. Figure S4).

Structural properties of different LIP flavours

In order to analyze and compare the properties of different LIPs classes a series of general statistics at the protein level (consensus) are provided. The secondary structure content, RSA, delta RSA, chain linearity, intra- and inter-chain contacts and FLIPPER score distributions are shown in Figure 4 for the different LIP classes. For that analysis, since structural features are derived from PDB data, when multiple LIPs map to the same protein and overlap, only the longest one is considered as representative of that LIP. This operation removed the redundancy of identical sequences and reduced the number of chains from ca. 65 thousands to 8,661 (Table 3). All statistics have been performed considering only representative chains. Figure 4 also shows the Non-LIP category as a control, which includes all non-LIP residues in the selected PDB chains.

Homo LIPs

As expected, Homo LIPs have a statistically significant higher fraction of alpha-helix and beta-strand (Kolmogorov-smirnov test p-value < 0.01) compared to the other flavours (Figure 4a). This flavour has a structural role, they are commonly found in intertwined homo-oligomers as small segments or domains exchanged between different subunits of a protein complex [17]. Also, they can participate in domain swapping as “swapped” regions, or be organized in large protein aggregates [18,33].

Figure 3 shows three examples of Homo LIPs. The structure of the p53 oligomerization domain [34] (UniProt: P04637, PDB: 1SAF) is composed of four identical protein chains (chains A, B, C and D; LIP residues 319-359). The tetramerization domain is located in the intrinsic disordered C-terminus of p53 [35] and folds upon oligomerization. This LIP has the highest content of alpha helix (about 50%) compared to the other examples and covers the 10% of the UniProt protein sequence (Suppl. Table S4).

The RNA-binding protein FUS (UniProt: P35637, PDB: 5W3N) is a disordered protein which participates in liquid-liquid phase separation. The structure (chains A, B, C, D ,F ,G ,H and I; LIP residues 37-97) is a self-assembly fibril formed by the low-complexity domain of FUS while the rest of the protein remains disordered [36]. Protein chains with the LIPs cover the whole protein sequence, presenting large average amounts of inter-contacts and delta RSA due to the characteristic fibrillar organization (Suppl. Table S4).

The chains of the receptor-binding tip of the bacteriophage T4 long-tail fiber (UniProt:P03744, PDB:2XGF) are the longest LIPs identified by FLIPPER (PDB residues 864-1021) in the Homo subset. The structure is composed of three identical protein chains that exhibit an elongated conformation which intertwine at the end of the tip [37]. As expected, this LIP presents the highest values of beta sheet content (48%) and linearity compared to the general trend. The average FLIPPER score for this example is 0.83, slightly lower than the other two examples of Homo LIPs (Suppl. Table S4).

Nucleic Acids LIPs

LIP flavours have statistically significant different numbers of inter- and intra-chain contacts (Kolmogorov-smirnov test p-value < 0.01), with the exception of Homo and Mixed flavours (Figure 4h-i). Interestingly, Nucleic Acids LIPs have more intra-chain contacts while lacking secondary structure (Figure 4c) and they are shorter with a median of 10 residues (Table 3). Delta RSA is particularly low for this flavour but it is an artifact as DNA and RNA molecules are not taken into consideration by DSSP when it calculates solvent accessibility (Figure 4e). Most of these LIPs are part of DNA binding proteins, which non-covalently bind and release nucleic acids exploring different conformations [38].

The structure of a truncated form of HMG-I/HMG-Y protein bound to DNA (UniProt: P17096, PDB: 2EZD chain A) is another example of folding upon binding, where the LIP is identified between the residues 7-20 in the PDB (Figure 3). In absence of DNA, this protein is fully disordered [39] and the LIP identified by FLIPPER is annotated as a structural transition region in the DisProt database [23] (identifier of a region evidence in DisProt: DP00040r006). It can be also observed that the LIP is partially unfolded, maintaining its flexibility after binding. Indeed, all its residues are defined as coil (Suppl. Table S4).

Protein LIPs

Protein LIPs have a higher median for RSA, delta RSA, chain linearity (Figure 4 d-f) and lower number of intra-chain contacts compared to the other LIP flavours (Figure 4h). They better fit FLIPPER definition of a linear peptide and, accordingly, FLIPPER score for this flavour is consistently higher indicating that they are easier to predict. They are more heterogeneous than Homo LIPs since they interact with one or more different proteins or nucleic acid polymers. The median alpha-helix content is comparable to Homo LIPs, however, they have less beta-strands and more coiled elements (Kolmogorov-smirnov test p-value < 0.01). These regions exist in different conformations depending on the context and can adapt their conformation to different interaction partners [28,40].

The p27Kip1 kinase inhibitory domain (UniProt: P46527, PDB: 1JSU chain C) is bound to the phosphorylated cyclin A-cyclin-dependent kinase 2 (Cdk2) and cyclin A [41]. The extended conformation of p27Kip1 is identified as a LIP from residues 25-93 (Figure 3). The extended conformation of the p27Kip1 domain is detected by FLIPPER as a LIP from PDB residues 25-93 (PDB code: 1JSU, chain: C). Since the unbound form of p27Kip1 is intrinsically disordered [42], the LIP folds upon binding (DisProt evidence code: DP00040r006) covering the full protein sequence (Suppl. Table S4). Moreover, this LIP region contains a Short Linear Motif (SLiM) from residues 27-37 identified within the function class Cyclin N-terminal Domain Docking Motifs in the ELM database (ELM identifier: DOC_CYCLIN_RXL_1) [10].

Mixed LIPs

LIP lengths statistically differ across subsets (Kolmogorov-smirnov test p-value < 0.01), being the Mixed class the one with longest LIPs with a median of 20 residues, and Nucleic Acids those with the shortest with a median of 10 residues (Table 3). Mixed LIPs are found on protein structures which are mostly part of big macromolecular complexes namely, ribosomes, nucleosomes and polymerases.

The 40S ribosomal protein S3 structure (UniProt: P23396, PDB: 5A2Q chain D) is part of the structure complex of the Hepatitis C Virus bound to the Human Ribosome. The LIP is detected in residues 206-227 interacting with ribosomal protein S17, the receptor of activated protein C kinase 1 (RACK1) and the DNA. The LIP is located at the C-terminal, in a predicted disordered region by MobiDB-Lite [43]. Its high content of coil (75%) as well as the disorder prediction indicates a strong propensity of this region to be disordered and highly flexible in the unbound state (Suppl. Table S4).

LIPs sequence composition and disorder prediction

Suppl. Figure S5 shows FLIPPER predicted LIPs, PDB observed and DisProt amino acid enrichment by using TrEMBL as a background frequency distribution. DisProt represents the reference for IDPs/IDRs amino acid composition and PDB observed (residues with available coordinates) represents well structured / globular domains. Pearson's correlation coefficient is 0.59 and 0.40 for FLIPPER/DisProt and FLIPPER/PDB comparisons, respectively. Compared to DisProt, LIPs are less enriched in disordered promoting residues (Glutamic Acid E, Lysine K, Proline P, Glutamine Q) and less depleted in hydrophobic residues. The most significantly depleted LIPs residues (Glycine G, Alanine A, Valine V and Tryptophan W) have been shown to be depleted also in DIBS, MFIB, ELM and FuzDB databases [44]. LIPs seem to have a composition in between disordered and ordered proteins, but with some specificities

like the enriched Arginine (R) and depleted Glycine (G) and Alanine (A). This information could be important for sequence based prediction methods.

To evaluate the propensity of LIPs to be disordered, five different methods were used to predict protein disorder. Suppl. Figure S6 shows the fraction of disordered LIPs found in each subset by the different methods. A LIP was classified as disordered when more than 50% of its residues are disordered. It can be seen that most of the LIPs are predicted as ordered. Only the Nucleic Acids subset shows a high tendency to disordered LIPs, supported by most predictors. This is related with the fact that disordered predictors have their intrinsic bias to particular flavours of protein disorder [45], depending on how they were trained. Particularly, VSL2b is one of the methods with the highest False Positive Rates, however it does not over-predict disorder [46]. VSL2b is the unique method that predicts most of the LIPs as disordered in the different subsets. This analysis also highlights the complexity at sequence level of these regions for the current state-of-the-art disorder predictors, since LIPs don't have the classic sequence signature of IDPs/IDRs and as it was observed before, their composition is in between structured and disordered proteins.

Comparison with curated databases

High-quality data about intrinsically disordered binding regions are stored in different manually curated databases which cover a wide range of different binding modes such as folding upon binding, short linear motifs, fuzzy interactions and mutual folding induced by binding. The amount of proteins included in curated databases is way lower compared to FLIPPER dataset and it is interesting to evaluate whether FLIPPER predictions can be used to guide biocuration. In Table 4 FLIPPER predictions are compared with curated LIPs. Overlapping (true positives), missed (false negatives) and overpredicted (false positives) residues are provided as percentages considering the union of predicted and curated LIP residues. Underpredicted residues can be explained by the fact that a fraction of curated LIPs are inaccessible to FLIPPER since not covered by any PDB experiment. This is particularly true for DisProt and IDEAL which collects experimental evidence from a large number of different techniques. Other differences arise from the different types and definition of the regions captured in the databases. ELM for example concentrates on short functional motifs and FuzDB on regions which perform transient interactions which are difficult to capture in PDB structures, often defined in regions of missing residues. MFIB and DIBS are the most similar databases with ca. 50% of common residues. The similarity can be explained by the fact the starting point for curating new entries is the PDB. Underprediction, instead, can be explained by the fact that biocurators tend to define the full chain as a LIP, whereas FLIPPER focuses on interacting positions, e.g. excluding N- / C-terminal tails. Despite all these differences the overpredicted fraction (22.2% / 18,496 residues) and all new predicted proteins (7,487) are worth inspecting to integrate novel functional evidence in curated databases.

Repeat domains are poorly represented by the LIP definition. RepeatsDB [47] provides annotations and classification of tandem repeat structures from the PDB. We found that only the 5.0% (704) of the RepeatsDB PDB chains are predicted by FLIPPER and that only in the 3.5% (513) of those chains contain repeated regions overlapping LIP predictions. We think that the structural characteristics of tandem repeat protein structures are substantially different from LIPs, since they have a regular well-folded structure with high density of intra-chain contacts.

Discussion and conclusion

The majority of intrinsically disordered proteins and regions (IDP/IDRs) are involved in key cellular processes thanks to their binding modes. In order to better understand how these interactions occur at structural level we developed FLIPPER to automatically identify the structurally linear sub-regions or peptides in protein complexes. The method is a random forest classifier trained on eight different structural features. Linearity, fraction of coil regions, intra- and inter-chain contacts, have been shown to discriminate well LIPs from non-LIPs residues (Figure 1). FLIPPER can accurately identify LIPs on structure complexes composed of proteins and nucleic acid chains. It shows high precision and sensitivity, a low false positive rate and good generalization on unseen examples (Table 2).

FLIPPER has been used to explore the landscape of LIPs flavours, or types, by processing the entire PDB. Homo LIPs, found in homo-oligomers, are mainly involved in domain swapping and intertwined homo complexes. Hetero LIPs, depending on the type of interacting partner, have a different fraction of secondary structures and solvent accessibility, while sharing comparable linearity and inter- and intra- contacts values (Figure 4). Indeed, the latter strongly explain the difference between homo and hetero LIPs, no matter the kind of binding partner involved.

LIPs show an amino acid composition in between ordered and disordered. This is also reflected by state-of-the-art disorder predictors that classify most of the LIPs as not disordered.

In conclusion, FLIPPER identifies LIPs in protein structure complexes, providing a residue by residue LIP propensity. FLIPPER captures the wide conformational variability of LIPs available in the entire PDB. FLIPPER predictions can be used to build new training sets for sequence-based methods able to identify different LIPs flavours. Moreover, FLIPPER predictions can guide the selection of targets to be curated in IDP databases. The FLIPPER software is available at URL: https://github.com/BioComputingUP/FLIPPER. All FLIPPER predictions are available in MobiDB 4.0.

Material and Methods

Algorithm

FLIPPER is a random forest classifier based on the Scikit-learn library [48]. FLIPPER considers an ensemble of 20 different decision trees, with no fixed maximum depth. Every leaf of a tree represents a split of the dataset created applying a cutoff on a residue feature (see features paragraph). To reduce overfitting, FLIPPER was trained with an early stopping procedure in which a new leaf is created when one split covers at least 0.05% of the training residues. The algorithm optimizes the split of positive (LIP) and negative (non-LIP) residues finding the best combination of leaves and thresholds.

Split quality is estimated with Gini impurity [49], given by the formula:

$$G = \sum_{i \in C} p(i) \cdot [1 - p(i)]$$

Where C is the set of classes, LIP and no-LIP, and $p(i)$ the probability of a target to belong to class *i*. When G equals to zero the separation of the classes is perfect. A node split is accepted if it lowers the impurity. A change in impurity is weighted by the total number of examples as follows:

$$n/N * G(CurrentNode) \; - \; n_r/n * G(RightNode) - n_l/n * G(LeftNode)$$

Where N is the total number of targets, *n* is the number of targets represented by the current node and n_r and n_l are the number of targets in the right and left nodes.

FLIPPER applies two postprocessing steps to get the final binary classification of the residues starting from the LIP probability provided by the random forest classification. A smoothing step is applied using a sliding window of size $w = 2 * c + 1$centered on a target residue. The new probability is obtained as follows:

$$s(x) \; = \frac{1}{1 \; + \; min(l \; - \; 1, x + c) \; - \; max \; (0, x - c)} \sum_{i=max(0,x-c)}^{min(l-1,x+c)} s(i)$$

with $x \; \in \; [0, l - 1]$, being l the length of the signal. By default FLIPPER classifies as LIP, those residues with a score higher than 0.5. The second step is a gap filling pass which removes all discontinuities.

Other types of classifiers (MLP and SVM) were tested obtaining similar results, the random forest was chosen because it generates models that can be interpreted.

FLIPPER on average processes one PDB file in less than a second, the computational cost is mainly taken by Input/Output operations.

Features

FLIPPER models are trained considering a set of features which can be calculated from the protein structure. For each residue structural local properties are captured by calculating the following features:

- Inter- and intra-chain contacts. Two residues are considered in contact when the distance between two atoms is lower than 3.8 Å. Only long range intra-chain contacts were considered with a sequence separation of at least 7 residues. Both the count and the average over a window are considered.
- Three state secondary structure content considering alpha-helix (“G,” “I,” “H”), beta-strand (“E”, “B”) and irregular/coil (“T”, “S”, unassigned residues) as assigned by [50] DSSP and averaged over a window.
- Relative Solvent Accessibility (RSA) calculated as ASA/MaxASA, where the ASA is provided by DSSP and MaxASA is taken from [51], averaged over a window.

- Delta RSA calculated considering the RSA difference between the chain in isolation and in complex, averaged over a window. Since DSSP does not consider non-standard peptidic molecules, the delta RSA is always zero for LIP residues interacting with DNA, RNA or other ligands.

- Structural linearity, the spatial distance between C_α of the first and last residues of a fixed sequence separation, divided by the window length. The theoretical maximum linearity is 3.8 Å (distance of two consecutive Cα without torsions). In practice, a value larger than 1.0 Å indicates linearity.

- Chain length. The length of chain capped at 100 residues and divided by 100.

For contacts and secondary structure features, the window size was set to 11 residues. For RSA, delta RSA and linearity the window size was 41 residues.

Datasets

The FLIPPER training set is composed of 70 different PDB structures which include 123 protein chains, 53 of which do not contain any LIPs (Table 1). PDB chains are not redundant at the sequence level with a maximum of 35% sequence identity (Suppl. Table S1). Most of the PDB examples come from the ANCHOR training datasets [27]. Since FLIPPER is implemented to detect linear regions in close contact with the partner, we visually inspected each structure and revised the ANCHOR dataset definition by narrowing LIPs region boundaries where necessary. The training set was integrated to include 27 chains (PDB IDs and chain IDs: 1dev_A, 1dev_B, 1t08_C, 1ee5_A, 1ycq_A, 1ycq_B, 2fym_A, 2fym_B, 1iwq_A, 1iwq_B, 1fv1_C, 2nl9_A, 2nl9_B, 1nx1_A, 1nx1_C, 1p4b_H, 1p4b_B, 2d1x_P, 2d1x_A, 1ee5_B, 5c2v_A, 1mnm_C, 4gkh_I, 2d1x_P, 2d1x_A, 4n4c_A, 1bh5_A) with both LIPs and globular domains. These cases were manually chosen from MobiDB 3.0 [52] which contains LIPs annotations provided by the Mobi 2.0 method [12].

Two different validation sets were built from the PixelDB [53] and DIBS [20] databases (Table 1). PixelDB contains peptide ligands bound with one or more receptors, clustered by structural similarity of the peptide-binding protein. DIBS contains examples of IDP chains which form complexes with globular chains. PDB chains with more than 25% sequence similarity with the training set were removed by using BLASTclust [54]. The resulting PixelDB-25 dataset contains 1,244 proteins and DIBS-25 828 proteins (Table 1).

Curated LIP (or IDR) annotations provided by DisProt, IDEAL, MFIB, DIBS, FuzDB and ELM databases were downloaded from MobiDB (version 2020_09) [30].

Training and evaluation

FLIPPER was trained with a 10-fold cross-validation and the output is generated averaging over all 10 models. Hyper parameters were manually set based on a grid search evaluated on the whole training set. FLIPPER performance on 10-fold cross-validation on training set is provided in the Suppl. Table S2. In the same table is also reported the performance in cross-validation shuffling labels in the LIP70 database before performing the cross-validation. The comparison with other methods, Mobi 2.0, was performed against PixelDB-25 and DIBS-25 datasets. All evaluations are performed by residue, considering as positives those belonging to PDB chains classified as peptides in PixelDB and as disordered in DIBS. The following performance measures were calculated: balanced accuracy (BAC), F1-score, Matthews'

correlation coefficient (MCC), positive predictive value (PPV) or precision, true negative rate (TNR) or specificity and true positive rate (TPR) or recall.

Disorder prediction

Disorder predictions were downloaded from MobiDB (version 2020_09) [30]. The following methods were considered: MobiDB-Lite (single consensus-based prediction; [43]), ESpritz X-ray [55], IUpred-long [56] and VSL2b [57]. These methods have different performance regarding disorder predictions, being MobiDB-lite the most restrictive with the lower rate of false positives. Meanwhile, VSL2b is more permissive to predict disorder and ESpritz has the best average prediction performance. Consensus-50 was also obtained from MobiDB, which is less restrictive than MobiDB-Lite (agreement on a given residue being disordered by more than 50% of the predictors) and shows a better performance than individual disorder predictors [58,59].

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 778247. This work was funded by ELIXIR, the research infrastructure for life-science data. Additional funding from the Italian Ministry of University and Research (MIUR), PRIN (Grant No. 2017483NH8). A.M.M. is funded by the research programme "MSCA Seal of Excellence @UniPD".

References

[1] N.E. Davey, The functional importance of structure in unstructured protein regions, Curr. Opin. Struct. Biol. 56 (2019) 155–163. https://doi.org/10.1016/j.sbi.2019.03.009.
[2] R. van der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, P.M. Kim, R.W. Kriwacki, C.J. Oldfield, R.V. Pappu, P. Tompa, V.N. Uversky, P.E. Wright, M.M. Babu, Classification of Intrinsically Disordered Regions and Proteins, Chem. Rev. 114 (2014) 6589–6631. https://doi.org/10.1021/cr400525m.
[3] R. Pancsa, M. Fuxreiter, Interactions via intrinsically disordered regions: What kind of motifs?, IUBMB Life. 64 (2012) 513–520. https://doi.org/10.1002/iub.1034.
[4] P.E. Wright, H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, J. Mol. Biol. 293 (1999) 321–331. https://doi.org/10.1006/jmbi.1999.3110.
[5] A. Borgia, M.B. Borgia, K. Bugge, V.M. Kissling, P.O. Heidarsson, C.B. Fernandes, A. Sottini, A. Soranno, K.J. Buholzer, D. Nettels, B.B. Kragelund, R.B. Best, B. Schuler, Extreme disorder in an ultrahigh-affinity protein complex, Nature. 555 (2018) 61–66. https://doi.org/10.1038/nature25762.
[6] M. Fuxreiter, Fold or not to fold upon binding — does it really matter?, Curr. Opin. Struct. Biol. 54 (2019) 19–25. https://doi.org/10.1016/j.sbi.2018.09.008.
[7] M. Fuxreiter, Fuzziness in Protein Interactions-A Historical Perspective, J. Mol. Biol. 430 (2018) 2278–2287. https://doi.org/10.1016/j.jmb.2018.02.015.
[8] P. Tompa, M. Fuxreiter, Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions, Trends Biochem. Sci. 33 (2008) 2–8. https://doi.org/10.1016/j.tibs.2007.10.003.
[9] N.E. Davey, K. Van Roey, R.J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, T.J. Gibson, Attributes of short linear motifs, Mol BioSyst. 8 (2012) 268–281. https://doi.org/10.1039/C1MB05231D.
[10] M. Kumar, M. Gouw, S. Michael, H. Sámano-Sánchez, R. Pancsa, J. Glavina, A.

Diakogianni, J.A. Valverde, D. Bukirova, J. Čalyševa, N. Palopoli, N.E. Davey, L.B. Chemes, T.J. Gibson, ELM—the eukaryotic linear motif resource in 2020, Nucleic Acids Res. 48 (2020) D296–D306. https://doi.org/10.1093/nar/gkz1030.
[11] A. Mohan, C.J. Oldfield, P. Radivojac, V. Vacic, M.S. Cortese, A.K. Dunker, V.N. Uversky, Analysis of Molecular Recognition Features (MoRFs), J. Mol. Biol. 362 (2006) 1043–1059. https://doi.org/10.1016/j.jmb.2006.07.087.
[12] D. Piovesan, S.C.E. Tosatto, Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures, Bioinforma. Oxf. Engl. 34 (2018) 122–123. https://doi.org/10.1093/bioinformatics/btx592.
[13] D. Shaji, T. Amemiya, R. Koike, M. Ota, Interface property responsible for effective interactions of protean segments: Intrinsically disordered regions that undergo disorder-to-order transitions upon binding, Biochem. Biophys. Res. Commun. 478 (2016) 123–127. https://doi.org/10.1016/j.bbrc.2016.07.082.
[14] F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N.P. Brown, G. Trave, T.J. Gibson, Understanding eukaryotic linear motifs and their role in cell signaling and regulation, Front. Biosci. J. Virtual Libr. 13 (2008) 6580–6603.
[15] K. Van Roey, B. Uyar, R.J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T.J. Gibson, N.E. Davey, Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation, Chem. Rev. 114 (2014) 6733–6778. https://doi.org/10.1021/cr400585q.
[16] S. Fukuchi, T. Amemiya, S. Sakamoto, Y. Nobe, K. Hosoda, Y. Kado, S.D. Murakami, R. Koike, H. Hiroaki, M. Ota, IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners, Nucleic Acids Res. 42 (2014) D320-325. https://doi.org/10.1093/nar/gkt1010.
[17] S.S. MacKinnon, A. Malevanets, S.J. Wodak, Intertwined Associations in Structures of Homooligomeric Proteins, Structure. 21 (2013) 638–649. https://doi.org/10.1016/j.str.2013.01.019.
[18] N.M. Mascarenhas, S. Gosavi, Understanding protein domain-swapping using structure-based models of protein folding, Prog. Biophys. Mol. Biol. 128 (2017) 113–120. https://doi.org/10.1016/j.pbiomolbio.2016.09.013.
[19] E. Fichó, I. Reményi, I. Simon, B. Mészáros, MFIB: a repository of protein complexes with mutual folding induced by binding, Bioinforma. Oxf. Engl. 33 (2017) 3682–3684. https://doi.org/10.1093/bioinformatics/btx486.
[20] E. Schad, E. Fichó, R. Pancsa, I. Simon, Z. Dosztányi, B. Mészáros, DIBS: a repository of disordered binding sites mediating interactions with ordered proteins, Bioinforma. Oxf. Engl. 34 (2018) 535–537. https://doi.org/10.1093/bioinformatics/btx640.
[21] S.K. Burley, H.M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J.M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D.S. Goodsell, R.K. Green, V. Guranović, D. Guzenko, B.P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prlić, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva, C. Zardecki, RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy, Nucleic Acids Res. 47 (2019) D464–D474. https://doi.org/10.1093/nar/gky1004.
[22] M. Miskei, C. Antal, M. Fuxreiter, FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies, Nucleic Acids Res. 45 (2017) D228–D235. https://doi.org/10.1093/nar/gkw1019.
[23] A. Hatos, B. Hajdu-Soltész, A.M. Monzon, N. Palopoli, L. Álvarez, B. Aykac-Fas, C. Bassot, G.I. Benítez, M. Bevilacqua, A. Chasapi, L. Chemes, N.E. Davey, R. Davidović, A.K. Dunker, A. Elofsson, J. Gobeill, N.S.G. Foutel, G. Sudha, M. Guharoy, T. Horvath, V. Iglesias, A.V. Kajava, O.P. Kovacs, J. Lamb, M. Lambrughi, T. Lazar, J.Y. Leclercq, E. Leonardi, S. Macedo-Ribeiro, M. Macossay-Castillo, E. Maiani, J.A. Manso, C. Marino-Buslje, E. Martínez-Pérez, B. Mészáros, I. Mičetić, G. Minervini, N. Murvai, M. Necci, C.A. Ouzounis, M. Pajkos, L. Paladin, R. Pancsa, E. Papaleo, G. Parisi, E. Pasche, P.J. Barbosa Pereira, V.J. Promponas, J. Pujols, F. Quaglia, P. Ruch, M. Salvatore, E. Schad, B. Szabo, T. Szaniszló, S. Tamana, A. Tantos, N. Veljkovic, S.

Ventura, W. Vranken, Z. Dosztányi, P. Tompa, S.C.E. Tosatto, D. Piovesan, DisProt: intrinsic protein disorder annotation in 2020, Nucleic Acids Res. (2019) gkz975. https://doi.org/10.1093/nar/gkz975.
[24] D.T. Jones, D. Cozzetto, DISOPRED3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics. 31 (2015) 857–863. https://doi.org/10.1093/bioinformatics/btu744.
[25] N. Malhis, M. Jacobson, J. Gsponer, MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences, Nucleic Acids Res. 44 (2016) W488–W493. https://doi.org/10.1093/nar/gkw409.
[26] N. Malhis, J. Gsponer, Computational identification of MoRFs in protein sequences, Bioinformatics. 31 (2015) 1738–1744. https://doi.org/10.1093/bioinformatics/btv060.
[27] B. Mészáros, I. Simon, Z. Dosztányi, Prediction of Protein Binding Regions in Disordered Proteins, PLoS Comput. Biol. 5 (2009). https://doi.org/10.1371/journal.pcbi.1000376.
[28] M. Miskei, A. Horvath, M. Vendruscolo, M. Fuxreiter, Sequence-Based Prediction of Fuzzy Protein Interactions, J. Mol. Biol. 432 (2020) 2289–2303. https://doi.org/10.1016/j.jmb.2020.02.017.
[29] M. Necci, D. Piovesan, C. Predictors, D. Curators, S.C.E. Tosatto, Critical Assessment of Protein Intrinsic Disorder Prediction, BioRxiv. (2020) 2020.08.11.245852. https://doi.org/10.1101/2020.08.11.245852.
[30] D. Piovesan, M. Necci, N. Escobedo, A.M. Monzon, A. Hatos, I. Mičetić, F. Quaglia, L. Paladin, P. Ramasamy, Z. Dosztányi, W.F. Vranken, N.E. Davey, G. Parisi, M. Fuxreiter, S.C.E. Tosatto, MobiDB: intrinsically disordered proteins in 2021, Nucleic Acids Res. (2020). https://doi.org/10.1093/nar/gkaa1058.
[31] UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515. https://doi.org/10.1093/nar/gky1049.
[32] S. Velankar, J.M. Dana, J. Jacobsen, G. van Ginkel, P.J. Gane, J. Luo, T.J. Oldfield, C. O'Donovan, M.-J. Martin, G.J. Kleywegt, SIFTS: Structure Integration with Function, Taxonomy and Sequences resource, Nucleic Acids Res. 41 (2013) D483-489. https://doi.org/10.1093/nar/gks1258.
[33] V. Esposito, F. Guglielmi, S.R. Martin, K. Pauwels, A. Pastore, R. Piccoli, P.A. Temussi, Aggregation Mechanisms of Cystatins: A Comparative Study of Monellin and Oryzacystatin, Biochemistry. 49 (2010) 2805–2810. https://doi.org/10.1021/bi902039s.
[34] G.M. Clore, J. Ernst, R. Clubb, J.G. Omichinski, W.M.P. Kennedy, K. Sakaguchi, E. Appella, A.M. Gronenborn, Refined solution structure of the oligomerization domain of the tumour suppressor p53, Nat. Struct. Biol. 2 (1995) 321–333. https://doi.org/10.1038/nsb0495-321.
[35] S. Kannan, D.P. Lane, C.S. Verma, Long range recognition and selection in IDPs: the interactions of the C-terminus of p53, Sci. Rep. 6 (2016) 23750. https://doi.org/10.1038/srep23750.
[36] D.T. Murray, M. Kato, Y. Lin, K.R. Thurber, I. Hung, S.L. McKnight, R. Tycko, Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains, Cell. 171 (2017) 615-627.e16. https://doi.org/10.1016/j.cell.2017.08.048.
[37] S.G. Bartual, J.M. Otero, C. Garcia-Doval, A.L. Llamas-Saiz, R. Kahn, G.C. Fox, M.J. van Raaij, Structure of the bacteriophage T4 long tail fiber receptor-binding tip, Proc. Natl. Acad. Sci. 107 (2010) 20287–20292. https://doi.org/10.1073/pnas.1011218107.
[38] M. Ganji, M. Docter, S.F.J. Le Grice, E.A. Abbondanzieri, DNA binding proteins explore multiple local configurations during docking via rapid rebinding, Nucleic Acids Res. 44 (2016) 8376–8384. https://doi.org/10.1093/nar/gkw666.
[39] J.R. Huth, C.A. Bewley, M.S. Nissen, J.N.S. Evans, R. Reeves, A.M. Gronenborn, G.M. Clore, The solution structure of an HMG-I(Y)–DNA complex defines a new architectural minor groove binding motif, Nat. Struct. Biol. 4 (1997) 657–665. https://doi.org/10.1038/nsb0897-657.
[40] B. Mészáros, L. Dobson, E. Fichó, G.E. Tusnády, Z. Dosztányi, I. Simon, Sequential, Structural and Functional Properties of Protein Complexes Are Defined by How Folding and Binding Intertwine, J. Mol. Biol. 431 (2019) 4408–4428. https://doi.org/10.1016/j.jmb.2019.07.034.

[41] A.A. Russo, P.D. Jeffrey, A.K. Patten, J. Massagué, N.P. Pavletich, Crystal structure of the p27 Kip1 cyclin-dependent-kinase inibitor bound to the cyclin A–Cdk2 complex, Nature. 382 (1996) 325–331. https://doi.org/10.1038/382325a0.
[42] E.A. Bienkiewicz, J.N. Adkins, K.J. Lumb, Functional Consequences of Preorganized Helical Structure in the Intrinsically Disordered Cell-Cycle Inhibitor p27 ^{Kip1} †, Biochemistry. 41 (2002) 752–759. https://doi.org/10.1021/bi015763t.
[43] M. Necci, D. Piovesan, Z. Dosztányi, S.C.E. Tosatto, MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins, Bioinformatics. 33 (2017) 1402–1404. https://doi.org/10.1093/bioinformatics/btx015.
[44] M. Necci, D. Piovesan, S.C.E. Tosatto, Where differences resemble: sequence-feature analysis in curated databases of intrinsically disordered proteins, Database. 2018 (2018). https://doi.org/10.1093/database/bay127.
[45] M. Necci, D. Piovesan, S.C.E. Tosatto, Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe, Protein Sci. Publ. Protein Soc. 25 (2016) 2164–2174. https://doi.org/10.1002/pro.3041.
[46] J.T. Nielsen, F.A.A. Mulder, Quality and bias of protein disorder predictors, Sci. Rep. 9 (2019) 5137. https://doi.org/10.1038/s41598-019-41644-w.
[47] L. Paladin, M. Bevilacqua, S. Errigo, D. Piovesan, I. Mičetić, M. Necci, A.M. Monzon, M.L. Fabre, J.L. Lopez, J.F. Nilsson, J. Rios, P.L. Menna, M. Cabrera, M.G. Buitron, M.G. Kulik, S. Fernandez-Alberti, M.S. Fornasari, G. Parisi, A. Lagares, L. Hirsh, M.A. Andrade-Navarro, A.V. Kajava, S.C.E. Tosatto, RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures, Nucleic Acids Res. 49 (2021) D452–D457. https://doi.org/10.1093/nar/gkaa1097.
[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, (2011).
[49] A. D'Ambrosio, V.A. Tutore, Conditional Classification Trees by Weighting the Gini Impurity Measure, in: S. Ingrassia, R. Rocci, M. Vichi (Eds.), New Perspect. Stat. Model. Data Anal., Springer, Berlin, Heidelberg, 2011: pp. 273–280. https://doi.org/10.1007/978-3-642-11363-5_31.
[50] W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers. 22 (1983) 2577–2637. https://doi.org/10.1002/bip.360221211.
[51] B. Rost, C. Sander, Conservation and prediction of solvent accessibility in protein families, Proteins Struct. Funct. Bioinforma. 20 (1994) 216–226. https://doi.org/10.1002/prot.340200303.
[52] D. Piovesan, F. Tabaro, L. Paladin, M. Necci, I. Micetic, C. Camilloni, N. Davey, Z. Dosztányi, B. Mészáros, A.M. Monzon, G. Parisi, E. Schad, P. Sormanni, P. Tompa, M. Vendruscolo, W.F. Vranken, S.C.E. Tosatto, MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins, Nucleic Acids Res. 46 (2018) D471–D476. https://doi.org/10.1093/nar/gkx1071.
[53] V. Frappier, M. Duran, A.E. Keating, PixelDB: Protein–peptide complexes annotated with structural conservation of the peptide binding mode, Protein Sci. 27 (2018) 1535–1537. https://doi.org/10.1002/pro.3431.
[54] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
[55] I. Walsh, A.J.M. Martin, T. Di Domenico, S.C.E. Tosatto, ESpritz: accurate and fast prediction of protein disorder, Bioinformatics. 28 (2012) 503–509. https://doi.org/10.1093/bioinformatics/btr682.
[56] B. Mészáros, G. Erdős, Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res. 46 (2018) W329–W337. https://doi.org/10.1093/nar/gky384.
[57] P. Radivojac, Z. Obradović, C.J. Brown, A.K. Dunker, Prediction of boundaries between intrinsically ordered and disordered protein regions, Pac. Symp. Biocomput. Pac. Symp. Biocomput. (2003) 216–227.
[58] M. Necci, D. Piovesan, Z. Dosztányi, P. Tompa, S.C.E. Tosatto, A comprehensive assessment of long intrinsic protein disorder from the DisProt database, Bioinformatics.

34 (2018) 445–452. https://doi.org/10.1093/bioinformatics/btx590.
[59] I. Walsh, M. Giollo, T. Di Domenico, C. Ferrari, O. Zimmermann, S.C.E. Tosatto, Comprehensive large-scale assessment of intrinsic protein disorder, Bioinformatics. 31 (2015) 201–208. https://doi.org/10.1093/bioinformatics/btu625.
[60] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2009. https://doi.org/10.1007/978-0-387-98141-3.

Tables

Dataset	Proteins	PDB chains	LIPs	non-LIPs	LIP residues	non-LIP residues
Training	113	123	70	53	1,725	12,922 (1,203)*
PixelDB	1,313	3,465	1,952	1,513	25,575	395,712
PixelDB-25	1,244	2,947	1,436	1,511	24,836	329,945
DIBS	906	1,639	765	874	13,279	155,267
DIBS-25	828	1,409	706	703	11,600	115,824

Table 1. Datasets composition. "Proteins" corresponds to the number of different UniProtKB accessions mapping to PDB chains. "LIPs" and "non-LIPs" contain the number of PDB chains with LIP and non-LIP regions. (*) Non-LIP residues in the same chain containing a LIP. For both PixelDB and DIBS, LIPs cover the full PDB chain.

Dataset	Method	MCC	Accuracy	F1 Score	False Negative Rate (FNR)	Precision (PPV)	Specificity (TNR)	Sensitivity (TPR)
Training	FLIPPER	0.949	0.975	0.955	0.043	0.954	0.994	0.957
	Mobi 2.0	0.556	0.729	0.587	0.518	0.750	0.976	0.482
PixelDB-25	FLIPPER	0.985	0.992	0.986	0.015	0.987	0.999	0.985
	Mobi 2.0	0.704	0.815	0.721	0.357	0.820	0.987	0.643
DIBS-25	FLIPPER	0.857	0.931	0.87	0.124	0.865	0.986	0.876
	Mobi 2.0	0.634	0.779	0.661	0.422	0.771	0.98	0.578

Table 2. FLIPPER prediction evaluation. FLIPPER performance is compared with Mobi2 on the FLIPPER training set, PixelDB-25 and DIBS-25.

Type	PDB					Consensus			
	PDBs	PDB chains	LIPs	LIP residues	Median LIP length	Proteins	LIPs	LIP residues	Median LIP length
Homo	8,541	27,365	32,208	610,740	14	3,239	4,399	88,049	14
Protein	11,735	34,436	39,216	790,117	14	5,423	7,369	155,247	15
Nucleic Acids	583	1,503	1,704	20,873	8	307	355	4,818	10
Mixed	1,175	4,012	4,451	96,870	17	664	787	19,355	20
Total	20,009*	65,020*	77,579	1,518,600	14	8,661*	12,910	267,469	15

Table 3. **Dataset composition of the LIPs classes.** At the PDB level (PDB) LIPs are considered independently in each PDB chain. At the consensus level (Consensus) LIPs found on PDBs mapping to the same UniProt accession are merged together. (*) The total is less than the sum of the column since the same protein can have LIPs of different types.

Database	Proteins*	LIPs°			Residues^ (%)		
	Common	Common	Under	Over	Common	Under	Over
MFIB	172 (69.9%)	174	5	44	51.9	40.5	7.6
DIBS	473 (94.8%)	513	9	236	49.5	20.5	30.0
IDEAL	166 (80.6%)	151	46	150	29.9	42.0	28.1
DisProt	260 (53.3%)	216	76	186	18.6	67.2	14.2
ELM	692 (33.2%)	501	699	692	14.2	18.9	66.9
FuzDB	52 (50.5%)	39	43	73	5.7	75.2	19.1
Total	1,174 (40.4%)	1,062	723	765	26.8	51.1	22.2

Table 4: FLIPPER consensus performance on manually curated databases. (*) The number of database proteins predicted by FLIPPER. In parenthesis the database percentage. (°) Common is the number of database regions that overlap by at least one residue with FLIPPER predictions. Under and Over represent the number of regions not predicted by FLIPPER and those predicted but not overlapping with any database regions, respectively. (^) The percentage of overlapping (Common), underpredicted/missed (Under) and overpredicted (Over). Percentages are calculated over the union of database and predicted residues. Both FLIPPER and database regions are merged at the protein level as in Table 3. For the total line all databases regions were merged together.

Figures

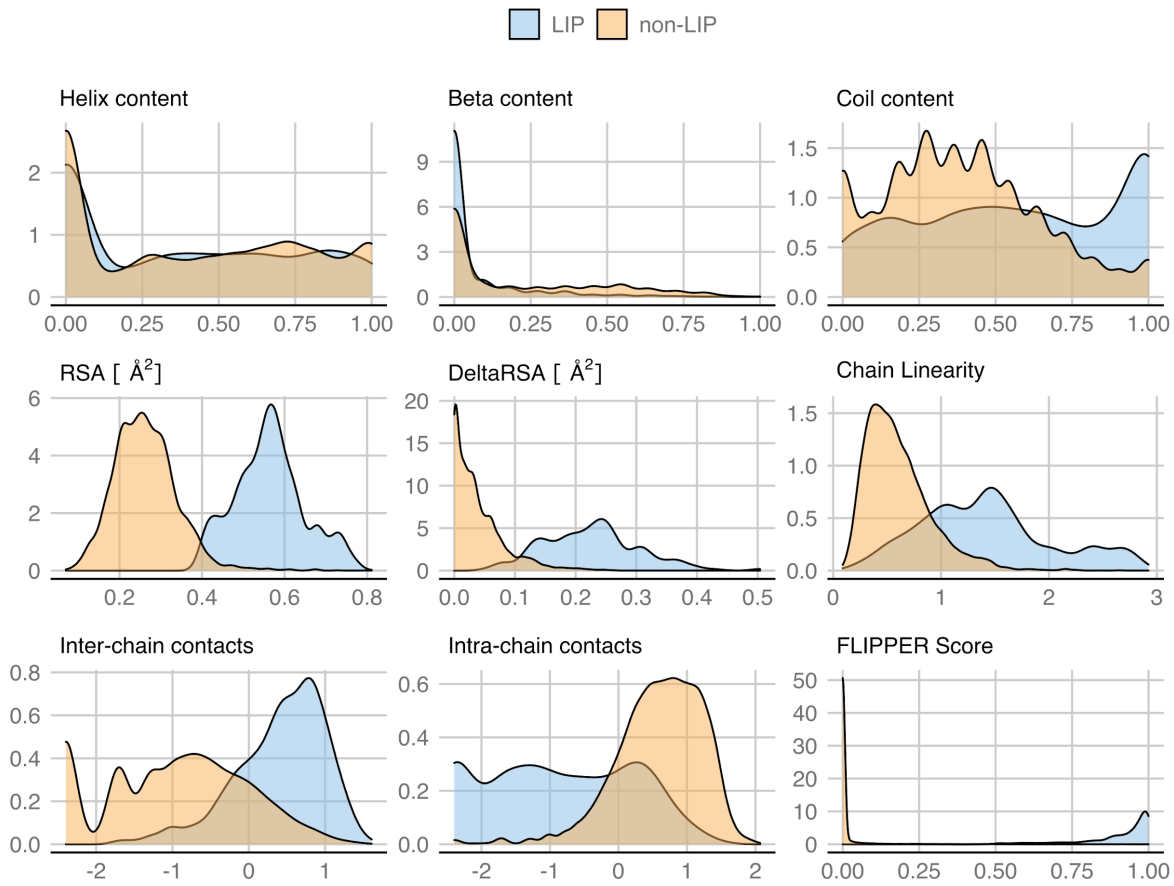


Figure 1. **Features distributions in the training set.** The density plots represent feature values for LIP and non-LIP residues in the training set. Density values are automatically calculated with the geom_density() function of the ggplot2 R package [60]. Intra- and Inter-chains contacts are expressed in a logarithmic scale. Distributions of positive and negative classes, to a lesser extent for helix and beta, are statistically different for all features (Kolmogorov-Smirnov P-value < 0.001).

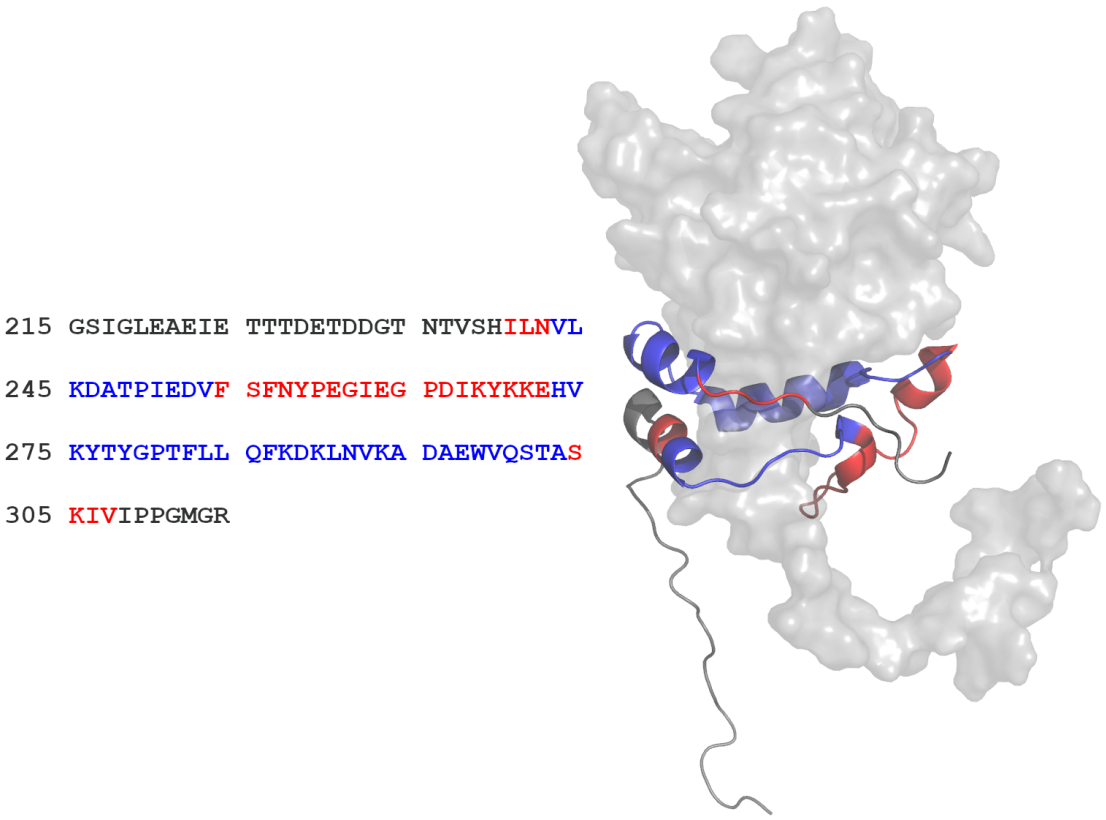


Figure 2. FLIPPER worst prediction on the training set. A fragment of the Eukaryotic initiation factor 4F subunit p150, component of the eIF4F complex of *Saccharomyces cerevisiae*, is shown in cartoon representation (PDB 1RF8, chain B). The LIP in the training set is defined from residue 240 to 307. FLIPPER underpredicted LIP residues but did not make any false positive prediction. True positives residues are 42 (blue) and false positives are 26 (red). The large central mispredicted region (red) is apparently more distant from the partner chain compared to the rest of LIP residues.

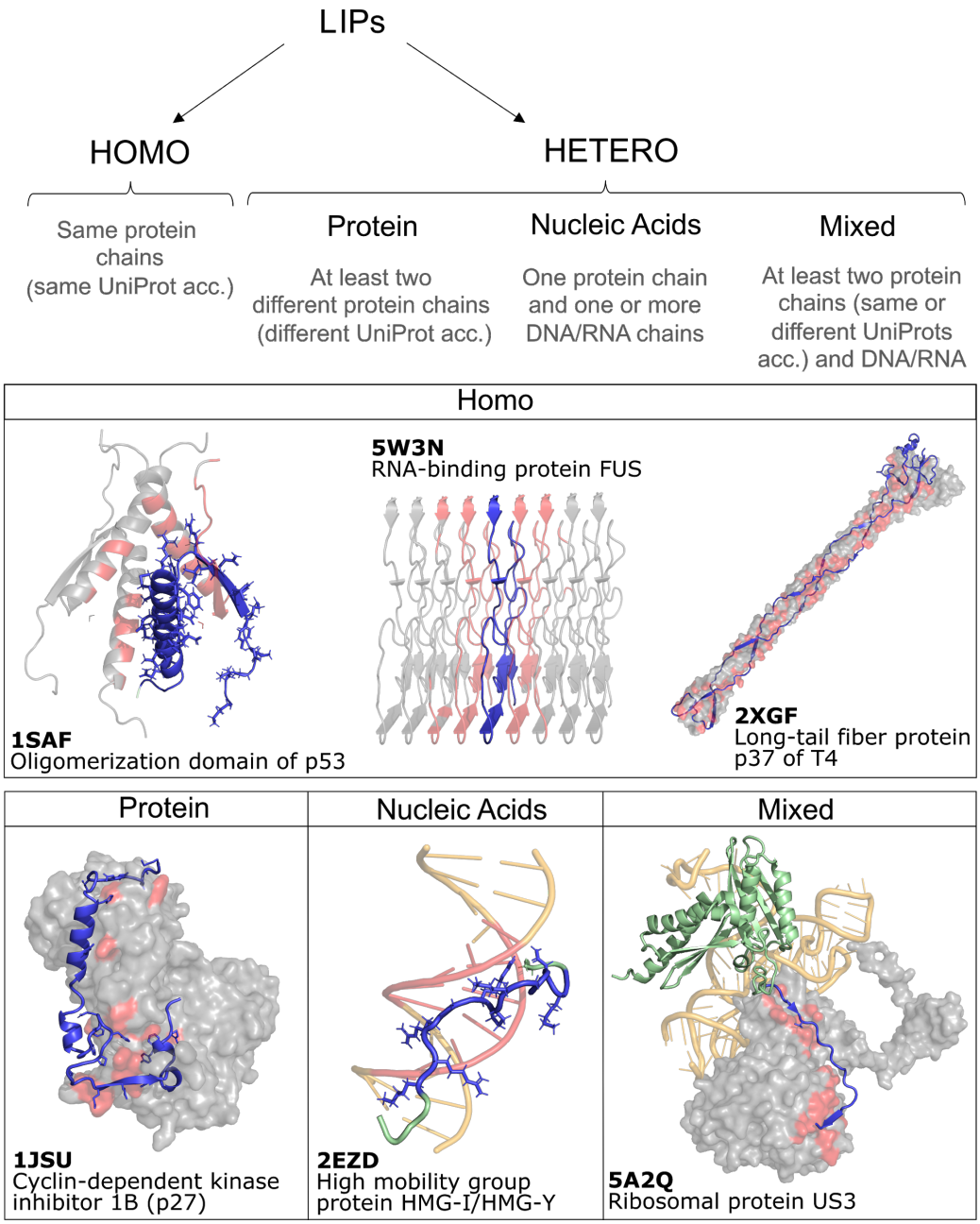


Figure 3. LIPs classification schema. LIPs classification and one example per category are shown. The color representation is as follows: blue LIP residues, green the PDB chain that contains the LIP, grey the protein chain/s interacting with the LIP and red the residues involved in the interaction with the LIP. Since FLIPPER identifies all protein chains of the Homo subset as LIP, only one is colored in blue for clarity.

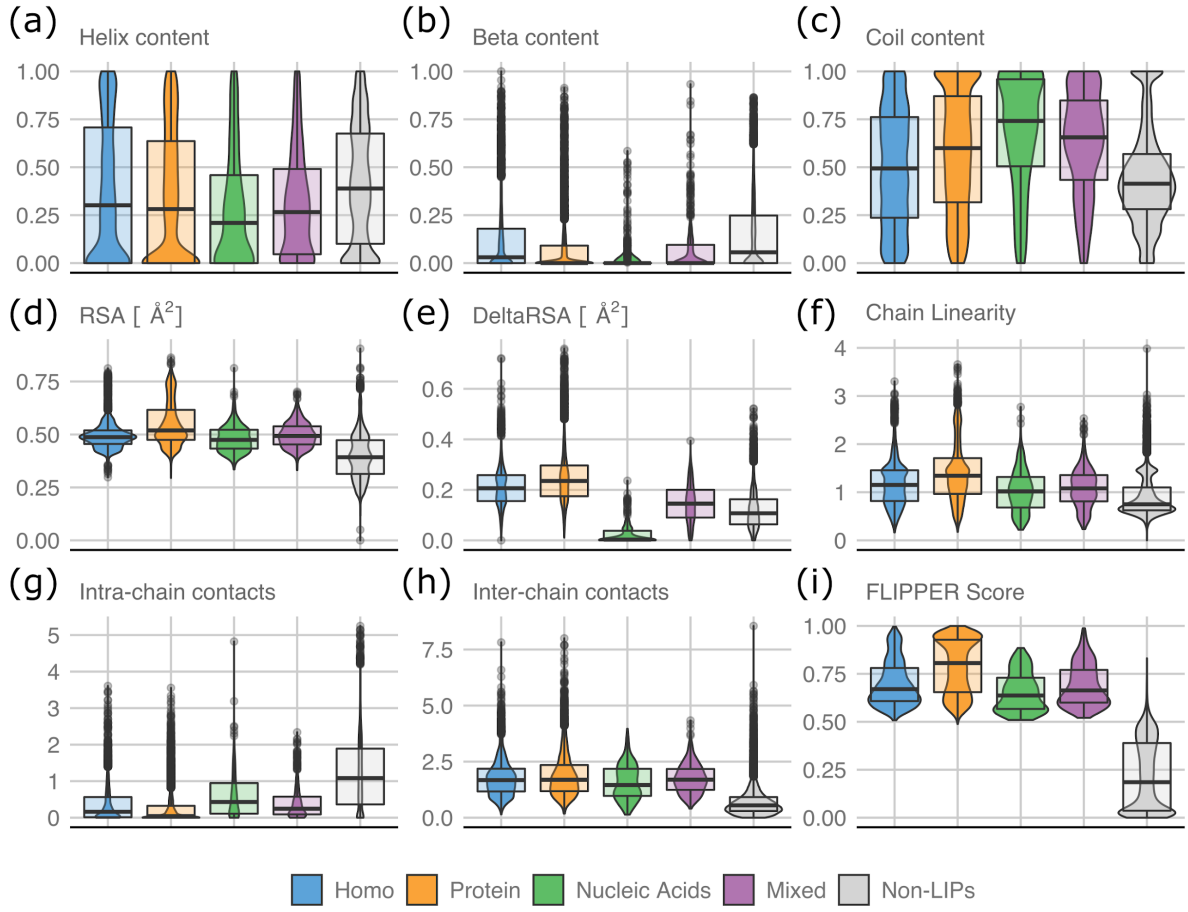


Figure 4. **Features distributions in LIPs classes.** Average feature values were considered for each consensus LIP in a particular subset. Non-LIP residues were taken from chains containing at least one LIP. A representative PDB and chain was chosen for those consensus LIPs with more than one PDB and chain associated.

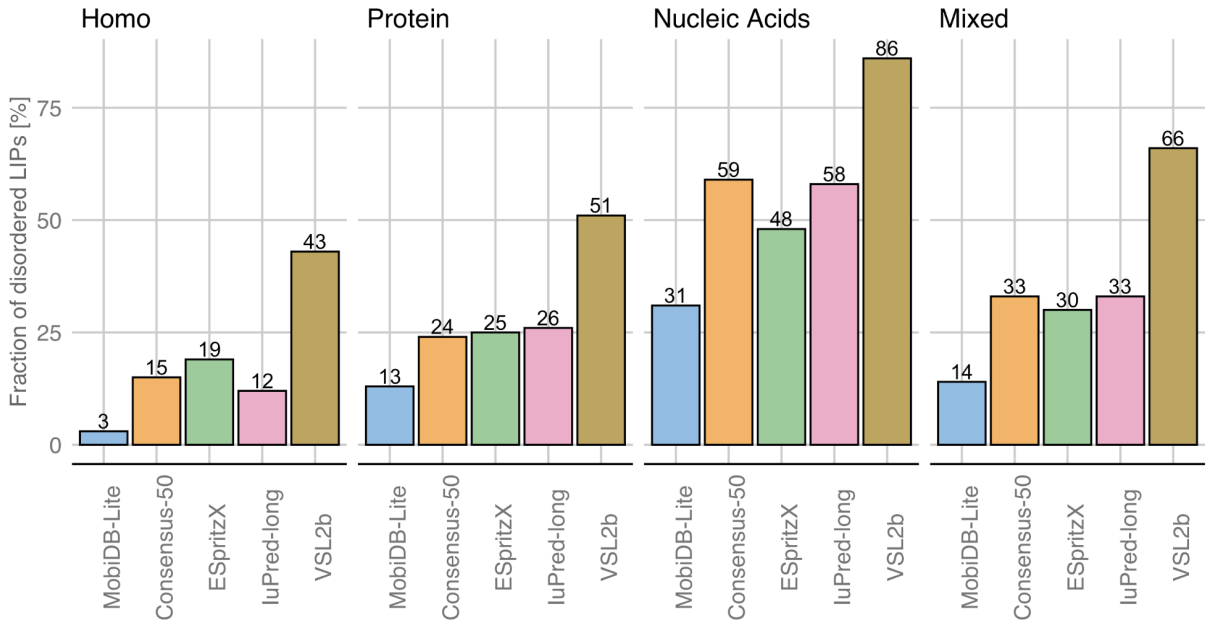


Figure 5. **Fraction of disordered LIPs in the classes.** A LIP is considered disordered if more than 50% of its residues are predicted to be disordered by the corresponding method. MobiDB-lite is a consensus which considers predictions from 8 different methods. Consensus-50 corresponds to a majority vote on the same set of predictions considered by MobiDB-lite.