Prediction of short linear protein binding regions.

Catherine Mooney [1,2,4], Gianluca Pollastri [1,3], Denis C. Shields [1,2,4*] and Niall J. Haslam [1,2,4]

[1] *Complex and Adaptive Systems Laboratory,* [2] *Conway Institute of Biomolecular and Biomedical Science,* [3] *School of Computer Science and Informatics, and* [4] *School of Medicine and Medical Science, University College Dublin*

---

*To whom correspondence should be addressed - denis.shields@ucd.ie +35317165344

Prediction of short linear protein binding regions.

Catherine Mooney [1,2,4], Gianluca Pollastri [1,3], Denis C. Shields [1,2,41] and Niall J. Haslam [1,2,4]

[1] *Complex and Adaptive Systems Laboratory,* [2] *Conway Institute of Biomolecular and Biomedical Science,* [3] *School of Computer Science and Informatics, and* [4] *School of Medicine and Medical Science, University College Dublin*

Abstract

Short linear motifs in proteins, typically of 3-12 residues in length, play key roles in protein-protein interactions, frequently binding specifically to peptide-binding domains within interacting proteins. Their tendency to be found in disordered segments of proteins has meant that they have often been overlooked. Here we present SLiMPred (Short Linear Motif Predictor), the first general *de novo* method to computationally predict such regions in protein primary sequences independent of experimentally defined homologs and interactors. The method applies machine learning techniques to predict new motifs based on annotated instances from the Eukaryotic Linear Motif database as well as structural, biophysical and biochemical features derived from the protein primary sequence. We have integrated these data sources and benchmarked the predictive accuracy of the method finding that it performs equivalently to a predictor of protein binding regions in disordered regions in addition to having predictive power for other classes of motifs sites such as polyproline II helix motifs and short linear motifs lying in ordered regions. It will be useful to predict peptides involved in potential protein associations and aid in the functional characterisation of proteins, especially in proteins lacking experimental knowledge of structure and interactions. We conclude that, despite the diversity of motif sequences and structures, SLiMPred is a valuable tool to prioritise potential interaction motifs in proteins.

Keywords: intrinsically unstructured proteins; molecular recognition;

[1] To whom correspondence should be addressed - denis.shields@ucd.ie +35317165344

protein-protein interface; linear motif; mini-motif; functional prediction; peptide binding;

Introduction

Many proteins or protein regions fail to fold into fixed tertiary structures. Over the last ten years these intrinsically unstructured/disordered proteins have been shown to be important functionally leading to an alternative view of protein function to the traditional sequence structure function paradigm (1). Short linear motifs (SLiMs), also referred to as linear motifs, minimotifs or Eukaryotic Linear Motifs (ELMs) are short functional sites typically found in disordered regions (2) and play a central role in cell regulation by acting as protein ligands, mediating many biological processes including cell signalling, post-translational modification and protein trafficking (3). Due to their short length (typically 3-12 residues) and the tendency to evolve convergently, their discovery is both an experimental and computational challenge.

Several databases exist to capture instances of short functional sites in proteins, described using regular expressions as short linear motifs e.g. MiniMotif and ELM (4, 5). These functional sites are described using regular expressions, which denote the functional important residues that have usually been experimentally determined using, for example, deletion or other mutagenesis experiments (6). The ELM database provides a classification scheme consisting of four classes, ligand (LIG), targeting (TRG), cleavage (CLV) and modification (MOD), and in the absence of an ontology describing linear motifs we will follow their classification scheme. Other databases have a specific focus, for example, PHOSIDA is a database of post-translational modification sites (7) and the phospho.ELM database (8) provides a resource for cataloguing the many protein phosphorylation motifs and has been successfully used to create tools for the prediction of kinase specificities, for example, NetPhorest (9).

Linear motif discovery, i.e. the process of identifying biologically functional instances of such sites is challenging due to the likelihood that a motif will be over-represented in a protein by chance (10, 11). To avoid this many methods employ evolutionary information (12) at the local and global level to filter potential SLiM instances (13). Such approaches have been successful in the discovery of new instances of phosphorylation sites in disordered proteins, for example, in Src homology kinases (14). Others such as MEME

(15) and NestedMICA (16) use expectation maximisation techniques. These approaches require information about the query protein, such as its known interactors, that precludes *de novo* prediction and are incapable of identifying instances that are not present in large numbers in the query set. Even the identification of novel instances of known motifs is problematic as the frequency of occurrence of SLiM definitions with low information content is high (17). This is evidenced by the number of potential SLiMs that are filtered by the ELM server when browsing using simple regular expression matches (5). Profile based methods may be able to improve on this but are still unable to identify motifs that occur with low frequency or as single occurrences. The current methods (10, 13) are best at exploring PPI (protein-protein interaction) sets and other groupings of proteins where the user has an expectation that the interaction is mediated by a motif. Whilst there is room for improvement in these methods there is also a requirement for *de novo* prediction methods that are not dependent on evolutionary information, in order to allow genome wide scanning for novel motifs. In sequenced organisms where there is little or no PPI data available *de novo* motifs prediction tools enable exploration of potential interactions without the availability of experimental data (18). To date *de novo* prediction of ligand recognition motifs, the largest category in ELM, has been neglected (19).

Although SLiMs are found in accessible parts of structured domains, they are overrepresented in intrinsically disordered regions of protein sequences, and an occurrence of a match to a SLiM is more likely to be functional if it is found in a region of intrinsic disorder (20). Protein disorder prediction programs such as IUPred (21), PONDR (22), DISOPRED (23) and others have shown their utility in the masking of disorder prior to a SLiM discovery search (24). The occurrence of SLiMs in disordered regions is associated with higher relative local conservation (RLC) of the SLiM compared to the rest of the surrounding region as disordered regions of proteins are under less evolutionary pressure than structured regions of proteins (13). The degree of conservation of the motif instance relative to the background local sequence conservation can be used in SLiM discovery and has been successfully implemented (12, 13). It is also possible that the motif may have evolved only once, making the job of identification by recurrence in unrelated proteins impossible.

It is more than ten years since it was first shown that it was possible to identify functional regions in disordered proteins that correspond to experimentally determined disorder-to-order transitions upon binding (25). Func-

tional sites in proteins can vary in their length from protein modification sites of two residues through to much longer recognition features of up to 70 amino acids (26). These longer functional sites responsible for mediating protein interactions are known are Molecular Recognition Features (MoRFs) (27) or protein binding regions (26). These binding regions are crucial for interaction with binding partners and have been classified as α-helical, β-structural, irregular, PPII, and complex MoRFs (28). Despite the discrepancy in length it has been shown that SLiMs have a high overlap with disordered binding regions such as MoRFs, and that in many cases they represent an equivalent set of motifs, with the exception of cases where the SLiMs occurs in an ordered domain (2). The identification and classification of residues likely to undergo such an order to disorder transition has led to the successful development of several tools for the prediction of such binding regions in disordered proteins. ANCHOR (29, 30) and α-MoRF-PredII (31, 32) are two methods for the prediction of these binding regions, however neither is specifically trained to discover short motifs which are typical of SLiMs.

The usual definition of a motif is a consensus sequence that is shared among proteins; more specifically the working definition of a short linear motif is typically one that is shared among two or more unrelated proteins, having evolved convergently in two different contexts. However, the ELM database nonetheless does include some instances of motifs for which the only documented instances have only evolved once (e.g. only occurring at homologous sites within one protein family, e.g. the Alix binding motif (LYPxL) in HIV and SIV gag proteins). Thus, the concept of what a SLiM is, while typically representing sequences that can and do evolve in multiple unrelated proteins to perform the same function, is sometimes used to consider short linear motifs which may have the potential to do this, but have only been observed in a single protein family. For the purpose of this study, we use the word SLiM in this more general sense of "a short protein region with SLiM-like properties". These regions share some properties in common with the ANCHOR and α-MoRF-PredII predicted binding regions, with a particular emphasis on trying to identify regions involved in short motifs, and with the aim of identifying such motifs in a wider range of structural contexts compared to those two methods. Thus, for the purposes of this manuscript, we will use the word motif or SLiM to indicate not only recurrently evolving motifs, but also those which have evolved only once in order to convey their short, linear nature.

Here we describe SLiMPred, a *de novo* method for the discovery of short

linear motifs involved in ligand recognition from the protein primary sequence. Using manually curated and computational predictions of motifs we have trained and tested a new bidirectional recurrent neural network (BRNN) in five-fold cross-validation. We include information about the structural context of the motif by using predicted secondary structure, structural motifs, solvent accessibility and disorder. Importantly no previous knowledge about the protein, i.e. no evolutionary, structural or experimental information, is required. Therefore our method is useful not only in the context of our current understanding of protein-protein interaction networks but also when there is little or no PPI or other experimental data available. SLiMPred has been tested and compared to ANCHOR and performs well, although these two methods have a different focus. SLiMPred predicts short linear motifs in ordered and disordered protein sequences, whereas ANCHOR predicts long and short disordered binding regions. We have attempted to clarify where there is overlap between the two methods, under what conditions one might chose to use one method as opposed to the other, and where a consensus prediction might be beneficial. This is an important addition to the repertoire of tools available for protein motif discovery, which we hope will prove useful for guiding future experiments where the aim is to uncover new associations on a genomic scale or to analyse individual proteins and complexes.

Results

Five-fold cross-validation training and testing is performed on a subset of the full sequence i.e. domain/non-domain sections which have at least one residue labelled as a SLiM (see Materials and Methods). For every residue SLiMPred predicts the probability between 0 and 1 of that residue being part of a SLiM. The closer the probability is to 1, the more confident SLiMPred is that the residue is a SLiM residue. The results in five-fold cross-validation for the training dataset, ELM300, are shown in Figure 1(a) as a ROC curve with thresholds from 0 to 1, i.e. the cut-off above which a residue is considered to be predicted as a SLiM residue. Figure 1(a) demonstrates the effectiveness of the method, showing that it is possible to retain a high true positive rate (TPR) with a low false positive rate (FPR) at a threshold of 0.1. In Figure 1(b) we show the results for the training dataset full protein sequences, again in five-fold cross-validation, in comparison to ANCHOR. The effect of the increased number of non-SLiM residues in the dataset can be seen as an increase in the FPR, since the full sequences training dataset has over 98%

of the residues labelled as non-SLiM. This is an important consideration when deciding what threshold to choose and consequently the confidence that can be put in a positive prediction. For example, if we take a sequence of length 300 residues with a SLiM of length 7 – from the ROC curve in Figure 1(b) at a threshold of 0.1 we can expect a TPR of 44% and a FPR of 22% i.e. 3 out of the 7 SLiM residues will be correctly predicted as positive and 66 of the total 300 residues will be incorrectly predicted as positive. At a 0.5 threshold 13% of the 7 SLiM residues will be correctly predicted (1 residue) and 12 other residues (4%) will be incorrectly predicted. Increasing the threshold will decrease both the false and true positive rate, however we believe a threshold of 0.1 offers a good balance between the two for initial scanning to explore candidate regions. In Table 3 we show the TPR and FPR at a threshold of 0.5 for ANCHOR and 0.1 for SLiMPred and the area under the curve (AUC) for both.

ELM300 Structural Classes

We divided the ELM300 training set into three structural classes. These were: α-Helix, β-Sheet and Polyproline II Helix (Table 6). These three classes roughly correspond to the three main structural classes that SLiMs form upon binding to their interacting domain and are based on the available structural data (33). This classification is not definitive since peptide binding conformations are not absolute, nevertheless it provides a useful guide. The composition of each class, in terms of the number of unique motifs, sequences and SLiM and non-SLiM residues are shown in Table 2. Any sequences that did not fall into one of these classes was put into the "others" class. This class includes SLiMs which adopt multiple or "complex" secondary structures when bound to different partners and those containing "irregular" secondary structures (34). The results in Figure 2 show the performance of ANCHOR and SLiMPred in the four categories and Table 4 shows TPR and FPR at a 0.5 threshold for Anchor and a 0.1 threshold for SLiMPred and the AUC for both. ANCHOR performs similarly to SLiMPred in the β-Sheet motifs and the "others" classes and better for α-Helix motifs while SLiMPred is noticeably better on the Polyproline II class of motifs.

Benchmarking against ANCHOR

SLiMPred aims to find residues that share properties with SLiM residues, one feature of which is that they are typically in interaction sites of less than 12 residues. However, it is likely that such properties may also be shared by

a subset of residues in longer binding regions. We were therefore interested to determine if it is showing substantial overlap with methods designed to predict disordered binding regions.

ANCHOR is a tool to identify residues in disordered regions of proteins that are likely to form binding interactions with protein domains (30). In order to predict disordered binding regions, ANCHOR identifies segments that reside in disordered regions, cannot form enough favourable intra-chain interactions to fold on their own and are likely to gain stabilizing energy by interacting with a globular protein partner. The approach relies on the pairwise energy estimation approach that is the basis for IUPred, a general disorder prediction method (21). It is intended to specifically identify regions of disordered proteins that are more likely to undergo disorder-to-order transitions upon binding. Such regions are similar to, but not limited to, the ligand binding motifs that SLiMPred is designed to identify. However they tend to be longer. SLiMs are typically 3-12 residues in length, disordered binding regions range from 5 to approximately 70 residues. However, for the purposes of comparison, and to determine if SLiMPred can identify features of binding regions of various lengths, in addition to shorter motifs, SLiMPred has been compared against ANCHOR using the ANCHOR long and short binding region datasets.

Prior to testing we redundancy reduced both datasets to $< 30\%$ sequence similarity with respect to the SLiMPred training set. The results shown are ensemble results from all models across the five cross-validation folds. Figure 3 shows the results for SLiMPred and ANCHOR on the ANCHOR long (18 sequences) and short (28 sequences) binding region datasets and Table 5 shows the TPR and FPR at a 0.5 threshold for Anchor and a 0.1 threshold for SLiMPred and AUC.

Both ANCHOR and SLiMPred perform better at identifying short binding regions (Figure 3(a)). This demonstrates that SLiMPred can accurately classify residues which are likely to undergo disorder-to-order transitions upon binding in short binding regions (5-30 residues in length) with a similar accuracy to ANCHOR, but that correct prediction of long binding regions (> 30 residues in length) is beyond the scope of SLiMPred at present.

SteinAloy and ELM63 Datasets

Figure 4 summarises the performance of both SLiMPred and ANCHOR on the combined SteinAloy and ELM63 independent test datasets described in Table 1. In order to further determine the ability of ANCHOR and

SLiMPred to predict motifs in ordered or disordered regions we split the dataset into "ordered" and "disordered" sequences based on the average disorder (predicted by IUPred (21) measured from 25 residues before the first SLiM residue (or first residue of the sequence) and continuing another 25 residues (or to the end of the sequence). See Figure 4(a) for the distribution of disorder around the SLiM. Sequences were placed in the "ordered" dataset if the average disorder around the SLiM was < 0.5 otherwise they were included in the "disordered" dataset resulting in 86 "disordered" sequences (Figure 4(b)) and 56 "ordered" sequences (Figure 4(c)). Whilst this is again a small dataset it is substantially bigger than the ANCHOR dataset (after redundancy reduction) offering a more balanced comparison of the effectiveness of the two methods. SLiMPred and ANCHOR have comparable effectiveness at identifying SLiM residues in disordered regions as shown in Figure 4(b). However SLiMPred is significantly better than ANCHOR at identifying SLiM residues in ordered regions as shown in Figure 4(c). This is not surprising as ANCHOR is designed specifically to address the identification of binding regions in disordered proteins. This suggests that SLiMPred is more generically applicable and has a role to play in the identification of ligand binding residues in structured proteins. SLiMPred is suited to the discovery of short motif like binding regions like the classic RGD integrin binding peptide through to longer nuclear localisation signal recognised by importins which extends to 25 residues in length.

Figure 5 shows the output for an example sequence (EPS15_MOUSE) from the ELM63 validation dataset (i.e., a protein that was excluded from the training set used to develop the method). The region in the diagram (residues 614 to 653 of the full protein sequence) is composed of a number of DPF tripeptide repeats, however only one of those has been experimentally determined to bind α- and β2-adaptin, albeit binding with different affinities (35). Since neither the motif description nor the protein sequence where included in the training set this example is shown to demonstrate the ability of SLiMPred to predict novel ligand binding regions. The disorder profile for this segment indicates it is moderately disordered. The region is predicted to be coiled and solvent accessible using Porter (data not shown) (36), indicating that it is likely to be a ligand binding region. Furthermore, the RLC scores for the high scoring SLiMPred residues are also high, indicating a high degree of conservation. This example shows that information presented to the user is valuable for making decisions about the likely correctness of the algorithm. The algorithm is not simply making predictions about the spe-

cific residues, otherwise all DPF motifs in that region would be high scoring. The output in Figure 5 shows how SLiMPred can be used in practice to help make contextual decisions about the scoring of putative motifs.

Webserver Implementation

The webserver implementation of the algorithm requires as input a UniProt accession number which the server uses to retrieve the sequence from the UniProt webservice (37). The server implements a queueing system to facilitate the retrieval of previous jobs, which are held for 21 days. A graphical representation of the score as a function of the residue at a given sequence position is included in the results page. In addition the IUPred score for each residue is displayed. The server optionally attempts to build an alignment of the orthologues of the protein sequence in order to provide the relative local conservation (RLC) score (13) as an additional plot in the output. The currently available predictions of RLC are for metazoan sequences. The SLiMPred score is unaffected by the presentation of the RLC score for the protein; however, this is instructive in the interpretation of results.

Genome Analysis

We ran the SLiMPred program on the Human proteome to identify *de novo* putative peptide motifs. The results in Table 7 show a number of putative ELM instances that resemble known ELMs found using CompariMotif (38), and another experimentally validated motif which is not part of the training set. These results are for the most stringent settings: eight consecutive residues with a SLiMPred score over 0.8. They demonstrate that the program is capable of finding novel motifs and putative new instances of known motifs. However, the motif descriptors are not very informative and the regions may therefore have different functionality. Figure 6 shows the SLiMPred prediction for a subset of the uncharacterised protein C2orf55, CB055 human (UniProt accession No.: Q6NV74). We also show disorder and RLC in addition to SLiMPred as these are scores that have been previously shown to enrich true positive instances of linear motifs. As can be seen from the alignment below the plot, the region predicted by SLiMPred is well conserved relative to the surrounding region as indicated by the high RLC scores. The region also has a moderate propensity for disorder as predicted by IUPred. Taken together, these three measures provide corroborative evidence that this is a true positive. This information would be useful for experimentalists trying to design mutational studies of a protein believed to

be involved in a protein-protein interaction. The method offers a way to prioritise likely regions that might mediate the interaction.

Clearly, it would be possible to train a predictor which incorporated conservation information. However, we have chosen not to, as the evolutionary context of training and application datasets may be very different. Users can combine these independent strands of information (SLiMPred, disorder prediction from IUPred and RLC) themselves as interpretation of the results based on an understanding of the target protein is required. Ensemble meta-predictions integrating ANCHOR, SLiMPred and conservation based prediction may have increased power for certain applications. The challenge is in developing good training test datasets to fully evaluate these.

Conclusions

The amount of proteomic information generated by high-throughput biological experiments is quickly outpacing our ability to annotate the proteome. It is crucial to develop and make available fast and accurate computational methods of sequence annotation. We have developed the first *de novo* method for SLiM prediction (SLiMPred) based on machine learning methods. No previous knowledge about the protein sequence is required, making our method useful where there is little or no PPI or other experimental data available. We have trained SLiMPred in five-fold cross-validation on a non-redundant set of annotated proteins from ELM and benchmarked the result against ANCHOR.

SLiMPred and ANCHOR perform equally well in disordered contexts and clearly there is some overlap here between both methods in their ability to predict SLiMs or binding regions. SLiMPred however is limited to predicting short binding regions and is unable to identify long (> 30 residues in length) binding regions. ANCHOR has been specifically developed for the prediction of binding regions in disorder proteins and as such does not perform well in an ordered context. We have shown that SLiMPred can predict SLiMs in structured regions and we see that it has an important role to play in this context. We believe that SLiMPred is an important contribution to the field of SLiM, MoRF or protein binding region prediction. Using SLiMPred in conjunction with ANCHOR will allow users to place greater confidence in SLiM/binding region prediction in disordered regions when there is a consensus between both predictors. SLiMPred predictions are given per residue, and it is then up to a user to interpret whether such residue predictions com-

bine into short segments that may form motifs. This is best interpreted in the light of other information, including evolutionary conservation. The scores could also potentially be used to identify longer regions, but given the low predictive power seen for longer disorder binding regions, we do not favour such use of the output. For identifying longer disordered binding regions, ANCHOR shows superior performance and would be the tool of choice. We would suggest the use of SLiMPred for the prediction of SLiMs in structured domains and ANCHOR for the prediction of long binding regions in disordered proteins.

SLiMPred is available as part of our webserver for SLiM discovery and annotation. Our server is designed to allow fast and reliable annotation of protein sequences. It is freely available for academic users at `http://bioware.ucd.ie/`. Linux binaries and the benchmarking sets are freely available for academic users upon request.

Materials and Methods

ELM300 – The SLiMPred Training Set

The ELM database is a manually annotated set of instances of short linear motifs (SLiMs) that have been experimentally demonstrated to interact with protein domains (5). The true positives are taken from the ELM database ligand class, which is comprised of 556 unique sequences with at least one instance of the 74 SLiMs in the ELM database on February 1st, 2010, as the source of our training set. This dataset was internally redundancy reduced to $< 30\%$ sequence similarity leaving 300 sequences with at least one instance of 65 of the 74 possible SLiMs. This training dataset is referred to as the ELM300 dataset.

Each residue of every sequence in this set was labelled as either being part of a SLiM or non-SLiM (see Table 1). Just over 1% of residues are labelled as being in a SLiM, making this a very difficult classification problem. Due to this very unbalanced nature of the dataset it was decided to use SMART (39) to split the sequence into domain/non-domain sections. During training, only domain or non-domain sections which included at least one residue labelled as a SLiM were included. We also excluded any sections which were less than ten residues in length. This reduced the total number of residues in the training set to 43,412 and increased the proportion of SLiM residues to 5.7%. The secondary structure (36), solvent accessibility (40), structural motifs (41) and disorder (21) were predicted for the full

sequence. Each domain section is then appropriately labelled with these features, and each section may contain multiple SLiM regions. While an ideal training set would comprise both positively labelled SLiM residues, and negatively labelled residues that are definitively proven not to be involved in SLiMs, in practice there is little experimental data regarding the true negative features of this dataset. Therefore, our negative residues comprise a combination of true negatives and possibly a number of false negatives. While this reduces the power of the data to make predictions, it is the only feasible approach given the current state of the data. The input to the BRNN for each domain/non-domain section is the length of the section, the three predicted structural features and predicted disorder per residue, the SLiM/non-SLiM labels per residue and a single extra input representing if the section is, or is not, predicted to be a domain by SMART.

Independent Test Datasets

Anchor Long and Short Datasets

ANCHOR is a tool to identify sequences in disordered regions of proteins that are likely to form ligand interactions with structured protein domains (30). The datasets used to train ANCHOR (short binding regions data and long binding regions data) were used here to ascertain the overlap between the two predictors, and to determine if SLiMPred can accurately predict disorder-to-order transition interactions (see (26) for more details on the datasets). Before testing, both ANCHOR datasets were redundancy reduced to less than 30% sequence similarity with respect to the SLiMPred training dataset (ELM300). After redundancy reduction, 28 out of the original 40 short binding region sequences and 18 out of the original 26 long binding region sequences remained. One of the major challenges in this work is the construction of a validation dataset, since so few instances have been characterised. Although the redundancy reduction results in a small dataset, it is nonetheless crucial for benchmarking the performance of SLiMPred.

SteinAloy Dataset

Recently Stein and Aloy have developed a method that attempts to identify SLiM mediated recognition of peptides by structured domains (42). This analysis revealed 63 novel SLiMs that are not part of our training dataset. To create our test dataset we searched the supplemental data provided with (42) for SLiMs labelled as "PLoS_CB_2010" i.e. novel SLiMs of which we found 63. We linked the PDBIDs (43) to their UniProtKB accession number (37),

and then used the regular expression provided in the supplemental data to find the SLiM in the sequence. We redundancy reduced this set with respect to the ELM300 set resulting in an independent test set of 79 sequences with at least one instances of 35 novel motifs.

ELM63 Dataset

Since creating the ELM300 training datasets a number of new SLiMs have been added to the ELM database (5). A search of the ELM database found 13 new SLiMs and 68 new sequences with these SLiMs *in situ*. Redundancy reduction of this set to less than 30% sequence similarity with respect to the ELM300 dataset results in an independent test set of 63 sequences with 12 novel SLiMs.

Predictive Algorithms and Implementation

To learn the mapping between inputs $\mathcal{I}$ and outputs $\mathcal{O}$ (sequence to short linear motif) an architecture composed of BRNN (44) with an input window of size 26 was used. BRNN have been used successfully for the prediction of secondary structure (36), solvent accessibility (40) and structural motifs (41) amongst other things. BRNN have the advantage over standard Feed-forward neural networks that they can automatically find the optimal context on which to base a prediction, i.e. the number of residues that are informative to determine a property. Because of their recursive nature, BRNN also have a relatively low number of free parameters compared to other neural networks with similar input size. This is particularly useful when the number of examples is small, as in this work.

These networks take the form:

$$\begin{aligned} o_j &= \mathcal{N}^{(O)}\left(i_j, h_j^{(F)}, h_j^{(B)}\right) \\ h_j^{(F)} &= \mathcal{N}^{(F)}\left(i_j, h_{j-1}^{(F)}\right) \\ h_j^{(B)} &= \mathcal{N}^{(B)}\left(i_j, h_{j+1}^{(B)}\right) \\ & \quad j = 1, \ldots, N \end{aligned}$$

where i_j (respectively o_j) is the input (respectively output) of the network in position j, and $h_j^{(F)}$ and $h_j^{(B)}$ are forward and backward chains of hidden vectors with $h_0^{(F)} = h_{N+1}^{(B)} = 0$. We parametrise the output update, forward update and backward update functions (respectively $\mathcal{N}^{(O)}$, $\mathcal{N}^{(F)}$ and $\mathcal{N}^{(B)}$) using three two-layered feed-forward neural networks.

Encoding sequence and structural information

Input i_j associated with the j-th residue contains primary sequence information and predicted structural information:

$$i_j = (i_j^{(E)}, i_j^{(T)}) \tag{1}$$

where, assuming that e units are devoted to sequence, and t to structural information:

$$i_j^{(E)} = (i_{j,1}^{(E)}, \ldots, i_{j,e}^{(E)}) \tag{2}$$

and:

$$i_j^{(T)} = (i_{j,1}^{(T)}, i_{j,t}^{(T)}) \tag{3}$$

Hence i_j contains a total of $e + t$ components.

As in (36) $e = 26$: beside the 20 standard amino acids, B (aspartic acid or asparagine), U (selenocysteine), X (unknown), Z (glutamic acid or glutamine) and . (gap) are considered, while the 26th input encodes the length of the sequence. We use $t = 23$ for representing structural information. Hence the total number of inputs for a given residue is $e + t = 49$. The first three structural input units contain the predicted three-class (DSSP style) secondary structure (SS) representing the predicted probability of the j-th residue belonging to either helix, strand or coil. The next four input units contain the predicted probability of the j-th residue belonging to one of four solvent accessibility (SA) classes. The following 14 input units contain the predicted probability of the j-th residue belonging to one of 14 structural motifs (SM), and the final two inputs are the predicted probability of the residue occurring in a disordered region or a predicted domain. SS, SA and SM are predicted using the Distill Server (45). Disorder is predicted using IUPred (46) and domains are predicted using SMART (39).

Training, Ensembling

Training is conducted by five-fold cross-validation, i.e. five different sets of training runs are performed in which a different fifth of the overall set is reserved for testing. The five fifths are roughly equally sized, disjoint, and their union covers the whole set. The training set is used to learn the free parameters of the network by gradient descent. Five models are trained independently for each fold and ensemble averaged to build the final predictor. Differences among models are introduced by varying the size of the window considered by the output network. 250 epochs of training are performed for

each model and the learning rate is halved every time we do not observe a reduction of the error for more than 50 epochs. When testing on an entirely different set from the one used during training we ensemble-combine all the models from all cross-validation folds.

Evaluating performance

To evaluate the performance of SLiMPred we measure the sensitivity or true positive rate (TPR) and specificity or false positive rate (FPR) as we increase the discrimination threshold from 0 to 1. The results are shown as a Receiver Operating Characteristic (ROC) curve where TPR is plotted against FPR, which are calculated as follows:

$$\begin{aligned} TPR &= \frac{TP}{TP+FN} \\ FPR &= \frac{FP}{FP+TN} \end{aligned}$$

where:

- True positives (TP): the number of residues predicted in a class that are observed in that class.
- False positives (FP): the number of residues predicted in a class that are not observed in that class.
- True negatives (TN): the number of residues predicted not to be in a class that are not observed in that class.
- False negatives (FN): the number of residues predicted not to be in a class that are observed in that class.

Acknowledgements

The authors acknowledge the Research IT Service at University College Dublin for providing HPC resources that have contributed to the research results reported within this paper. The work was funded through a Science Foundation Ireland (SFI) PI grant (Grant No. 08/IN.1/B1864) to Denis Shields and SFI Research Frontiers Grant 10/RFP/GEN2749 to Gianluca Pollastri.

[1] H. Dyson, P. Wright, Intrinsically unstructured proteins and their functions, Nature Reviews Molecular Cell Biology 6 (3) (2005) 197–208.

[2] M. Fuxreiter, P. Tompa, I. Simon, Local structural disorder imparts plasticity on linear motifs, Bioinformatics 23 (8) (2007) 950.

[3] N. Davey, R. Edwards, D. Shields, Computational identification and analysis of protein short linear motifs, Frontiers in Bioscience 15 (2010) 801–825.

[4] S. Rajasekaran, S. Balla, P. Gradie, M. R. R. Gryk, K. Kadaveru, V. Kundeti, M. W. W. Maciejewski, T. Mi, N. Rubino, J. Vyas, M. R. R. Schiller, Minimotif miner 2nd release: a database and web system for motif search., Nucl. Acids Res. 37 (Database Issue) (2008) D185–D190.

[5] C. M. Gould, F. Diella, A. Via, P. l. Puntervoll, C. Gemünd, S. Chabanis-Davidson, S. Michael, A. Sayadi, J. C. Bryne, C. Chica, M. Seiler, N. E. Davey, N. J. Haslam, R. J. Weatheritt, A. Budd, T. Hughes, J. Pas, L. Rychlewski, G. Travé, R. Aasland, M. Helmer-Citterich, R. Linding, T. J. Gibson, ELM: the status of the 2010 eukaryotic linear motif resource., Nucl. Acids Res. 38 (Database issue) (2010) D167–80.

[6] K. Ngoei, B. Catimel, N. Church, D. Lio, C. Dogovski, M. Perugini, P. Watt, H.-C. Cheng, D. Ng, M. Bogoyevitch, Characterization of a novel JNK (c-Jun N-terminal kinase) inhibitory peptide, Biochem. J. 434 (2011) 399413.

[7] F. Gnad, J. Gunawardena, M. Mann, PHOSIDA 2011: the posttranslational modification database., Nucl. Acids Res. 39 (Database Issue) (2010) D253–260.

[8] H. Dinkel, C. Chica, A. Via, C. M. Gould, L. J. Jensen, T. J. Gibson, F. Diella, Phospho.ELM: a database of phosphorylation sites–update 2011., Nucl. Acids Res. 39 (Database Issue) (2010) D261–7.

[9] M. L. Miller, L. J. Jensen, F. Diella, C. Jorgensen, M. Tinti, L. Li, M. Hsiung, S. A. Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovsky, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li,

G. Cesareni, T. Pawson, B. E. Turk, M. B. Yaffe, S. Brunak, R. Linding, Linear motif atlas for phosphorylation-dependent signaling., Science Signaling 1 (35).

[10] V. Neduva, R. B. Russell, DILIMOT: discovery of linear motifs in proteins., Nucl. Acids Res. 34 (Web Server issue) (2006) W350–5.

[11] R. Gutman, C. Berezin, R. Wollman, Y. Rosenberg, N. Ben-Tal, QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns, Nucl. Acids Res. 33 (suppl_2) (2005) W255–261.

[12] C. Chica, F. Diella, T. J. Gibson, Evidence for the concerted evolution between short linear protein motifs and their flanking regions., PLoS ONE 4 (7) (2009) e6052.

[13] N. E. Davey, D. C. Shields, R. J. Edwards, Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery., Bioinformatics 25 (4) (2009) 443–50.

[14] S. Ren, V. N. Uversky, Z. Chen, K. K. Dunker, Z. Obradovic, Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions., BMC genomics 9 Suppl 2.

[15] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, W. S. Noble, MEME SUITE: tools for motif discovery and searching., Nucl. Acids Res. 37 (Web Server issue) (2009) W202–208.

[16] M. Dogruel, T. A. Down, T. J. Hubbard, NestedMICA as an ab initio protein motif discovery tool., BMC Bioinformatics 9 (2008) 19.

[17] N. E. Davey, N. J. Haslam, D. C. Shields, R. J. Edwards, SLiMSearch 2.0 biological context for short linear motifs in proteins, Nucl. Acids Res. Webserver Issue (2011) in press.

[18] C. Kesmir, A. K. Nussbaum, H. Schild, V. Detours, S. Brunak, Prediction of proteasome cleavage motifs by neural networks., Protein Eng. 15 (4) (2002) 287–96.

[19] R. Malik, K. Dulla, E. A. Nigg, R. Koerner, From proteome lists to biological impact–tools and strategies for the analysis of large MS data sets., Proteomics 10 (6) (2010) 1270–1283.

[20] A. Via, C. M. Gould, C. Gemünd, T. J. Gibson, M. Helmer-Citterich, A structure filter for the Eukaryotic Linear Motif Resource., BMC Bioinformatics 10 (1) (2009) 351.

[21] Z. Dosztányi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content., Bioinformatics 21 (16) (2005) 3433–4.

[22] B. Xue, R. L. Dunbrack, R. W. Williams, a. K. Dunker, V. N. Uversky, PONDR-FIT: a meta-predictor of intrinsically disordered amino acids., Biochimica et biophysica acta 1804 (4) (2010) 996–1010.

[23] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, D. T. Jones, The DISOPRED server for the prediction of protein disorder., Bioinformatics 20 (13) (2004) 2138–9.

[24] F. Diella, S. Chabanis, K. Luck, C. Chica, C. Ramu, C. Nerlov, T. J. Gibson, KEPE–a motif frequently superimposed on sumoylation sites in metazoan chromatin proteins and transcription factors., Bioinformatics 25 (1) (2009) 1–5.

[25] E. Garner, P. Romero, A. Dunker, C. Brown, Z. Obradovic, Predicting binding regions within disordered proteins, Genome Informatics Series (1999) 41–50.

[26] B. Mészáros, I. Simon, Z. Dosztányi, Prediction of protein binding regions in disordered proteins, PLoS Comput. Biol 5 (2009) e1000376.

[27] A. Mohan, C. Oldfield, P. Radivojac, V. Vacic, M. Cortese, A. Dunker, V. Uversky, Analysis of molecular recognition features (MoRFs), Journal of molecular biology 362 (5) (2006) 1043–1059.

[28] V. Vacic, C. Oldfield, A. Mohan, P. Radivojac, M. Cortese, V. Uversky, A. Dunker, Characterization of molecular recognition features, MoRFs, and their binding partners, Journal of proteome research 6 (6) (2007) 2351–2366.

[29] B. Meszaros, I. Simon, Z. Dosztanyi, Prediction of Protein Binding Regions in Disordered Proteins, PLoS Computational Biology 5 (5) (2009) e1000376.

[30] Z. Dosztányi, B. Mészáros, I. Simon, ANCHOR: web server for predicting protein binding regions in disordered proteins., Bioinformatics 25 (20) (2009) 2745–6.

[31] C. Oldfield, Y. Cheng, M. Cortese, P. Romero, V. Uversky, A. Dunker, Coupled folding and binding with α-helix-forming molecular recognition elements, Biochemistry 44 (37) (2005) 12454–12470.

[32] Y. Cheng, C. Oldfield, J. Meng, P. Romero, V. Uversky, A. Dunker, Mining α-helix-forming molecular recognition features with cross species sequence alignments, Biochemistry 46 (47) (2007) 13468–13477.

[33] A. K. Dunker, C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang, J. W. Chen, V. Vacic, Z. Obradovic, V. N. Uversky, The unfoldomics decade: an update on intrinsically disordered proteins., BMC genomics 9 Suppl 2 (2008) S1.

[34] C. Oldfield, J. Meng, J. Yang, M. Yang, V. Uversky, A. Dunker, Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners, BMC genomics 9 (Suppl 1) (2008) S1.

[35] D. Owen, Y. Vallis, B. Pearse, H. McMahon, P. Evans, The structure and function of the β2-adaptin appendage domain, The EMBO Journal 19 (16) (2000) 4216–4227.

[36] G. Pollastri, A. McLysaght, Porter: a new, accurate server for protein secondary structure prediction., Bioinformatics 21 (8) (2005) 1719–20.

[37] T. U. Consortium, The Universal Protein Resource (UniProt) in 2010., Nucl. Acids Res. 38 (Database issue) (2010) D142–8.

[38] R. J. Edwards, N. E. Davey, D. C. Shields, CompariMotif: quick and easy comparisons of sequence motifs., Bioinformatics 24 (10) (2008) 1307–9.

[39] I. Letunic, T. Doerks, P. Bork, SMART 6: recent updates and new developments., Nucl. Acids Res. 37 (Database issue) (2009) D229–32.

[40] G. Pollastri, A. J. M. Martin, C. Mooney, A. Vullo, Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information., BMC Bioinformatics 8 (2007) 201.

[41] C. Mooney, A. Vullo, G. Pollastri, Protein Structural Motif Prediction in Multidimensional ϕ-ψ Space Leads to Improved Secondary Structure Prediction, Journal of Computational Biology 13 (8) (2006) 1489–1502.

[42] A. Stein, P. Aloy, Novel Peptide-Mediated Interactions Derived from High-Resolution 3-Dimensional Structures, PLoS Computational Biology 6 (5) (2010) e1000789.

[43] H. Berman, T. Battistuz, T. Bhat, W. Bluhm, P. Bourne, K. Burkhardt, Z. Feng, G. Gilliland, L. Iype, S. Jain, et al., The protein data bank, Acta Crystallographica Section D: Biological Crystallography 58 (6) (2002) 899–907.

[44] P. Baldi, S. Brunak, P. Frasconi, G. Soda, G. Pollastri, Exploiting the past and the future in protein secondary structure prediction, Bioinformatics 15 (11) (1999) 937.

[45] D. Baú, A. J. M. Martin, C. Mooney, A. Vullo, I. Walsh, G. Pollastri, Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins., BMC Bioinformatics 7 (2006) 402.

[46] Z. Dosztányi, V. Csizmók, P. Tompa, I. Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins., J. Mol. Biol. 347 (4) (2005) 827–839.

	Seq	Motifs	SLiM	non-SLiM	Total
ELM300	300	377	2461	223437	225898
Anchor_Long	18	29	864	9125	9,989
Anchor_Short	28	32	506	12987	13493
SteinAloy	79	82	490	35193	35683
ELM63	63	74	405	39751	40156

Table 1: Composition of each of the datasets: the number of sequences (Seq), the number of motif instances, the number of positive (SLiM) and negative (non-SLiM) residues and the total number of residues in each dataset. ELM300 is the SLiMPred training dataset. Anchor_Long, Anchor_Short, SteinAloy and ELM63 are independent test datasets (see Datasets section for more details).

	α-Helix	β-Sheet	Polyproline II	Other
Number of Sequences	30	49	30	141
Unique ELM Motifs	7	11	8	31
SLiM Residues	387	324	218	1239
Non-SLiM Residues	24702	34194	16642	121872

Table 2: Composition of the ELM300 dataset. The "other" class includes ELMs which didn't fall into any of the previous three classes such as binding regions that can have different geometries when bound to different partners.

		ELM300			ELM300–Full Seq.		
	Threshold	TPR	FPR	AUC	TPR	FPR	AUC
Anchor	0.5				0.57	0.29	0.69
SliMPred	0.1	0.43	0.12	0.74	0.44	0.22	0.67

Table 3: AUC, TPR and FPR, at 0.1 and 0.5 thresholds for Anchor and SLiMPred, on the ELM300 training dataset.

		α-Helix			β-Sheet			Polyproline II Helix			Other		
	Thresh	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
Anchor	0.5	0.79	0.43	0.77	0.51	0.27	0.68	0.24	0.27	0.47	0.56	0.30	0.69
SliMPred	0.1	0.52	0.24	0.69	0.31	0.23	0.65	0.54	0.22	0.75	0.41	0.23	0.66

Table 4: AUC, TPR and FPR, at 0.1 and 0.5 thresholds for Anchor and SLiMPred, on the structural subsets of ELM300.

		Short Anchor Dataset			Long Anchor Dataset			"Disordered" Set			"Ordered" Set		
	Thresh	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
Anchor	0.5	0.58	0.26	0.67	0.44	0.30	0.59	0.65	0.31	0.74	0.11	0.13	0.50
SliMPred	0.1	0.42	0.20	0.69	0.29	0.25	0.56	0.57	0.26	0.72	0.37	0.17	0.69

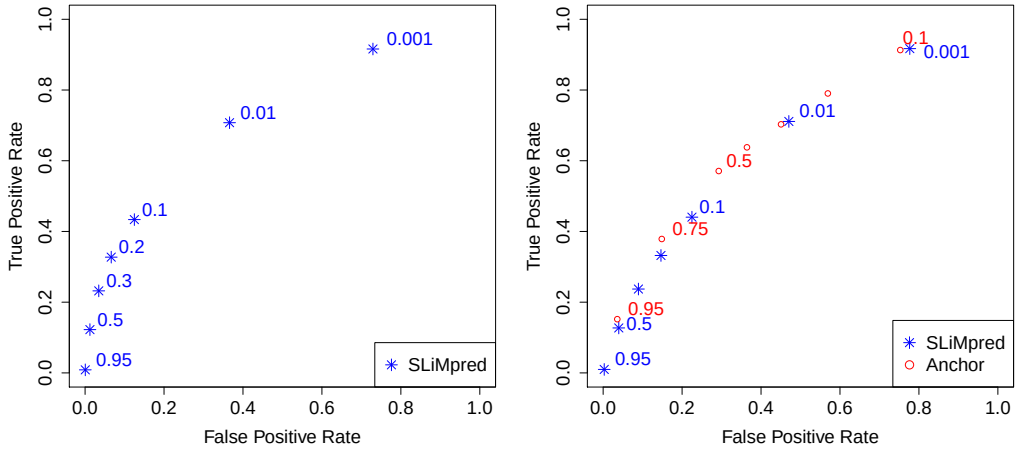
Table 5: AUC, TPR and FPR, at 0.1 and 0.5 thresholds for Anchor and SLiMPred, on the two ANCHOR test sets, Anchor long and short, and the combined SteinAloy/ELM63 datasets split into "Disordered" and "Ordered" Sets.

α-Helix	β-Sheet	Polyproline II
LIG_EH1_1	LIG_Dynein_DLC8_1	LIG_CAP-Gly_1
LIG_GLEBS_BUB3_1	LIG_PDZ_1	LIG_SH3_1
LIG_IQ	LIG_PP1	LIG_SH3_2
LIG_MDM2	LIG_PP2B_1	LIG_SH3_3
LIG_NRBOX	LIG_SH2_GRB2	LIG_SH3_5
LIG_Sin3_1	LIG_SH2_SRC	LIG_TRAF2_1
LIG_Sin3_3	LIG_SH2_STAT3	LIG_TRAF6
	LIG_SH2_STAT5	LIG_WW_1
	LIG_SIAH_1	
	LIG_TRFH_1	
	LIG_WRPW_1	

Table 6: Classification of the ELM motifs into three structural classes: α-Helix, β-Sheet, Polyproline II Helix.

UniProt ID	Gene Name	Motif	Position	Comment	Evidence
APC_HUMAN	Adenomatous polyposis coli protein	rps-QIPTPVNN-ntk	2801 - 2811	Microtubule tip localization signal (Known)	PMID:19632184
P85A_HUMAN	Phosphatidylinositol 3-kinase regulatory subunit alpha	pap-ALPPKPPK-ptt	306 - 313	LIG_SH3_2 (Known)	PMID:12186904
CB055_HUMAN	Uncharacterized C2orf55 protein	ksk-FKTFKKFF-gkk	30 - 37	CLV_PCSK_SKI1_1 TRG_NLS_Bipartite_1 similarity?	Uncharacterised
ZN837_HUMAN	Zinc finger protein 837	rek-RPEEPRPL-eed	28 - 35	LIG_SH3_3 similarity?	Uncharacterised
ZN281_HUMAN	Zinc finger protein 281	aap-AAEPPPPP-apd	87 - 94	LIG_SH3_3 similarity?	Uncharacterised
HMX3_HUMAN	Homeobox protein HMX3	pqp-PPPP-ppap	18 - 22	LIG_SH3_3 similarity?	Uncharacterised
PDZD2_HUMAN	PDZ domain containing protein 2	ese-EEQIE-ics	1596 - 1600	LIG_PDZ_3 similarity?	Uncharacterised
BCAS1_HUMAN	Breast carcinoma-amplified sequence 1	kti-TPPEP-ept	482 - 486	LIG_SH3_3 LIG_FHA_2 similarity	Uncharacterised
KI26B_HUMAN	Kinesin-like protein KIF26B	eps-SFPFE-elp	944 - 948	LIG_BRCT_BRCA1_1 similarity?	Uncharacterised
MAVS_HUMAN	Mitochondrial antiviral signaling protein	drp-PDPLE-pps	104 - 108	LIG_PDZ_3 similarity?	Uncharacterised

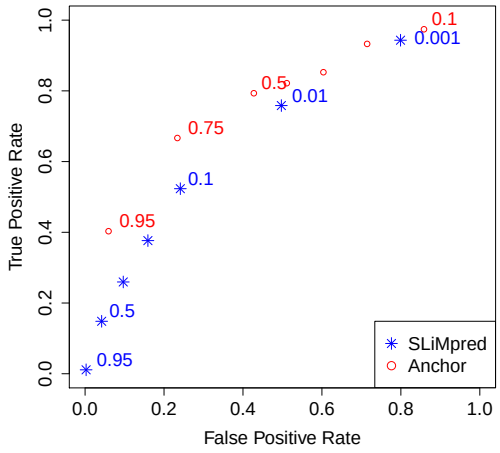
Table 7: Top hits in the Human SLiMPred analysis. The gene name and description are shown in the first two columns. The motif identified by SLiMPred is indicated in capital letters, with the flanking three residues either side shown in lowercase and the motif matching to the ELM instance in bold. The position and matches to known motifs are shown in the next columns. Putative matches to regular expressions from CompariMotif are shown where found, and known motifs are indicated. Finally any evidence, and a reference to the PubMed Identifier (PMID) is indicated in the final column, where uncharacterised indicates a lack of annotation for the protein generally.



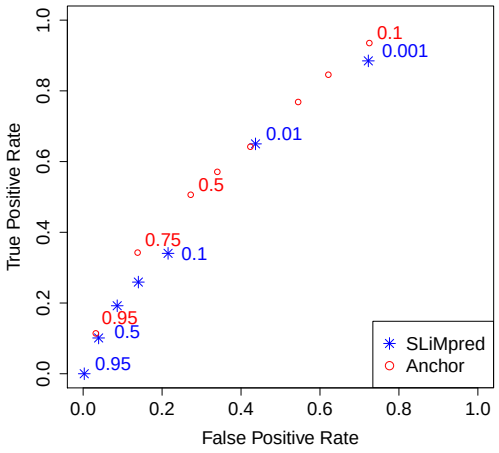
(a) ELM300

(b) ELM300–Full Sequence

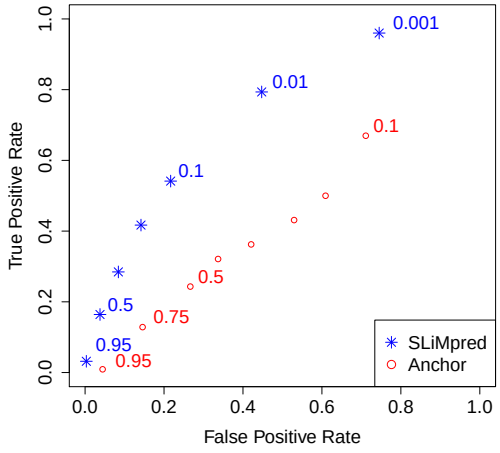
Figure 1: ROC curves (true positive rate as a function of false positive rate) for the ELM300 training dataset (a) SLiMPred trained in five-fold cross-validation. Values for residues used in training only i.e. domain/non-domain sections with at least one SLiM residue (b) SLiMPred and ANCHOR tested on the full length of the training set sequences.



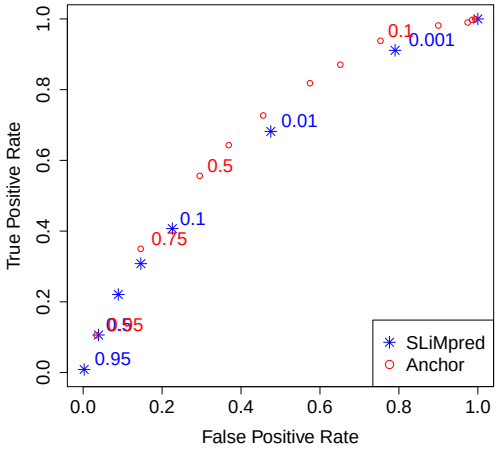
(a) α-Helix



(b) β-Sheet

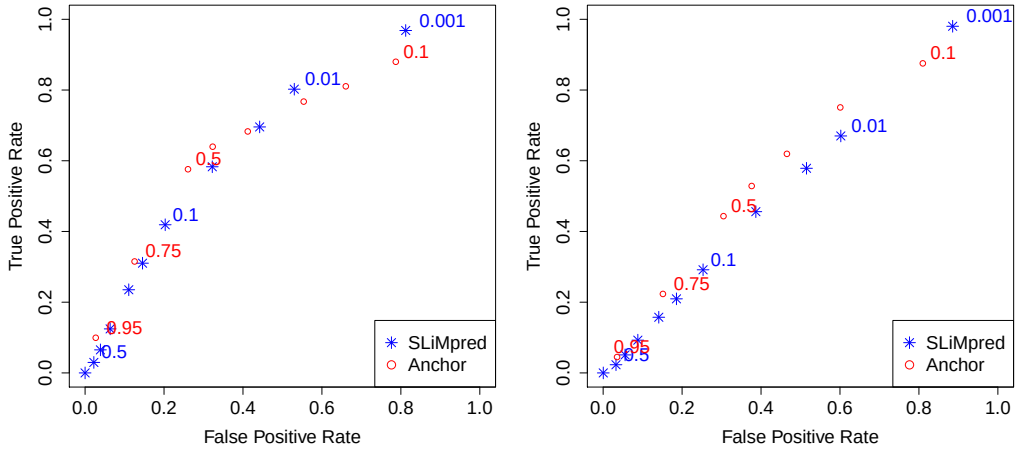


(c) Polyproline II Helix



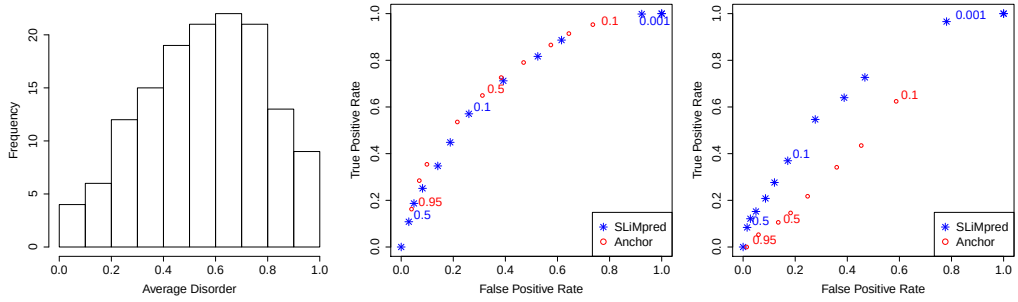
(d) Other

Figure 2: ROC curves (true positive rate as a function of false positive rate) for SLiMPred and Anchor tested on structural subsets defined as α-Helix, β-Sheet and Polyproline II Helix (see Table 6) from the ELM300 training dataset. The "other" class includes ELMs which didn't fall into any of the previous three classes such as binding regions that can have different geometries when bound to different partners.



(a) Short Anchor (b) Long Anchor

Figure 3: ROC curves (true positive rate as a function of false positive rate) for SLiMPred and ANCHOR tested on (a) Short ANCHOR dataset and (b) Long ANCHOR dataset.



(a) Average Disorder around SLiM (b) "Disordered" Set (c) "Ordered" Set

Figure 4: Combined SteinAloy/ELM63 datasets. **(a)** Histogram of disorder distribution. ROC curves (true positive rate as a function of false positive rate) for SLiMPred and ANCHOR for **(b)** "Disordered" motifs and for **(c)** "Ordered" motifs.

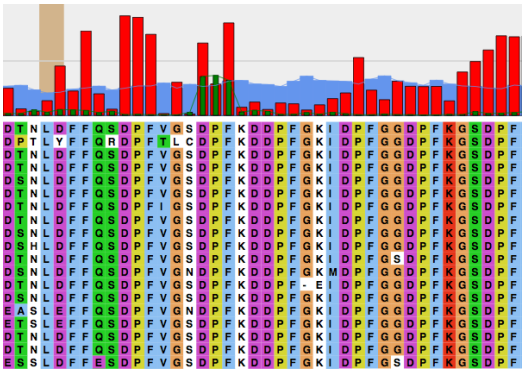


Figure 5: Output for the SLiMPred server for protein accession number: P42567 EPS15_MOUSE. The SLiMPred score is shown in green. The IUPred disorder propensity score is shown in Blue. The relative local conservation (RLC) score is shown in Red. The input sequence is the first sequence in the alignment, containing a subset of the orthologues, beneath the chart. The light grey horizonal line indicates the maximum score for SLiMPred (1.0).

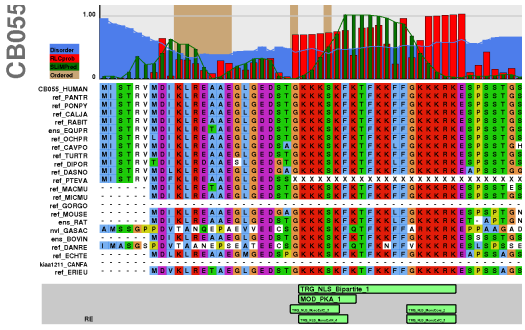


Figure 6: Sample Output for the SLiMPred server for protein accession number: Q6NV74 CB055_Human. The SLiMPred score is shown in green. The IUPred disorder propensity score is shown in Blue. The relative local conservation (RLC) score is shown in Red. The input sequence is the first sequence in the alignment, containing a subset of the orthologues, beneath the chart. Regular expression matches to known SLiMs from the ELM database are shown underneath.