PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Searching for protein signatures using a multilevel alphabet

Ronit Hod, Refael Kohen, and Yael Mandel-Gutfreund*

Faculty of Biology, Technion – Israel Institute of Technology, Haifa 32000, Israel

ABSTRACT

Short motifs are known to play diverse roles in proteins, such as in mediating the interactions with other molecules, binding to membranes, or conducting a specific biological function. Standard approaches currently employed to detect short motifs in proteins search for enrichment of amino acid motifs considering mostly the sequence information. Here, we presented a new approach to search for common motifs (protein signatures) which share both physicochemical and structural properties, looking simultaneously at different features. Our method takes as an input an amino acid sequence and translates it to a new alphabet that reflects its intrinsic structural and chemical properties. Using the MEME search algorithm, we identified the proteins signatures within subsets of protein which encompass common sequence and structural information. We demonstrated that we can detect enriched structural motifs, such as the amphipathic helix, from large datasets of linear sequences, as well as predicting common structural properties (such as disorder, surface accessibility, or secondary structures) of known functional-motifs. Finally, we applied the method to the yeast protein interactome and identified novel putative interacting motifs. We propose that our approach can be applied for *de novo* protein function prediction given either sequence or structural information.

Proteins 2013; 00:000–000.

Key words: motif search; multilevel alphabet; protein disorder; secondary structure; surface accessibility.

INTRODUCTION

Proteins are built from modular units known as domains. Protein domains fold into independent three-dimensional (3D) structures, usually performing a specific biological function. To date, more than 12,000 families of domains have been characterized.[1] While the domains cover the majority of the protein landscape, the remaining protein segments, which are usually less characterized, may contribute to important biological functions, for example, domain linkers and protein tails involved in protein–protein interactions.[2] These latter regions tend to be of low amino acid complexity and are intrinsically unstructured.[3] In many cases, these regions contain short linear motifs that confer their biological functions.[4] Protein short linear motifs (SLIMs) are short amino acid segments (3–10 long) that can be involved in mediating protein interactions with nucleic acids[5] or with other proteins, like in the case of the Dynein light chain binding motif that is important for cell trafficking.[2] Linear motifs can also represent phosphorylation sites or tagging sequences like the nuclear localization signal (NLS).[6] Linear motifs are often characterized by an additional structural property and are found preferentially in disordered regions.[7] For example, post-translational modifications (PTM), such as methylation, acetylation, and phosphorylation, are believed to be located preferentially in unstructured (disordered) regions.[8] The advantages of disordered regions for molecular interactions have been studied extensively; the general assumption is that high specificity coupled with low affinity required to form the molecular complex is accomplished by lowering the free energy needed to adapt the unstructured interface to their interacting partner. Furthermore, unordered regions can enable a large interaction surface and reduce the requirement of the two partners to be in the exact orientation with respect to each other. Disordered regions in proteins can fold to bind a unique binding site, or, on the contrary, allow one protein to bind multiple partners.[9] Indeed, proteins characterized as hubs in protein–protein interac-

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Israeli Science Foundation, ISF; Grant number: 1297/09.
*Correspondence to: Yael Mandel-Gutfreund, Faculty of Biology, Technion – Israel Institute of Technology, Haifa 32000, Israel. E-mail: yaelmg@tx.technion.ac.il
Received 7 August 2012; Revised 9 January 2013; Accepted 11 January 2013
Published online 5 February 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24261

tion networks have been found to be significantly more disordered than proteins with relatively few interacting partners, and so are their short sequence motifs that are presumably involved in binding.[10] It has been shown recently that the C terminal tail of p53 that is believed to be disordered mediates fast sliding along the DNA, allowing rapid scanning of long DNA regions.[11] The role of intrinsic disordered regions in DNA recognition has been suggested to be a common mechanism for DNA-binding proteins.[12] In addition to disorder, other features can contribute to the functionality of short linear motifs, such as the secondary structure in which the motif is embedded or its accessible surface area. As expected, functional motifs tend to appear on the protein surface and are abundant in loops.[13,14] In many cases, functional motifs are relatively more conserved compared to their neighboring regions.[15] Moreover, in a recent study comparing functional vs. non-functional short motifs, it was shown that in true positive linear motifs, the defined positions (fixed and degenerated residues) show higher conservation than in their false positive matches.[7] Indeed, evolutionary weighting has been shown to improve the performance of several motif search approaches.[16]

Motif search is a highly challenging problem in biology, attempting to find short common patterns in extremely large datasets.[4] Several databases of short linear motifs have been established[17–19] and it is estimated that hundreds of linear motifs remain to be discovered.[2] Quite impressively, the Minimotif Miner 2nd release has expanded the motif database from 462 to over 5000 motifs in two years,[20] and the ELM database increased to 1800 annotated motif instances as of 2011.[21] Furthermore, many bioinformatics tools have been developed to date for detecting common short linear motifs overrepresented among proteins that do not necessarily share sequence similarity but are known to interact with particular domains or proteins, for example, SLiMDisc,[22] DILIMOT,[23] SLiMFinder,[24] DRIMUST,[25] and DLocalMotif.[26] These methods are usually referred to as *occurrence based* methods. The MEME program[27] uses the Expectation Maximization (EM) algorithm[28] to discover enriched *de novo* motifs. MEME motifs are represented by position-specific probability matrices (PSPM) and are scored by an *E*-value that represents their statistical significance. In addition, specialized methods such as D-STAR,[29] DSLIMMER,[30] and iELM[31] were developed to find correlated short linear motifs within protein–protein interaction (PPI) data. Generally, these methods rely on finding statistically overrepresented, co-occurring similar substring pairs in PPI data (correlated motifs). Furthermore, several machine learning techniques were also applied to discover short linear motifs in proteins (e.g., see Ref. 32).

Here, we present a new approach for identifying common linear motifs that reflect both the common sequence and structural properties. Over the years, many studies have employed modified amino acid alphabets. Most of the studies concentrated on seeking the best reduced alphabet required to fold a protein.[33–35] More recently, different studies have investigated the use of a reduced alphabet for efficient homology search,[36] for protein structure prediction and for fold recognition.[37–39] Weathers *et al.*[40] showed that a reduced amino acid alphabet based on chemical similarly can be employed successfully for accurate recognition of intrinsically disorder proteins. Here, we use a modified amino acid alphabet that captures different properties of the protein residue (such as charge, disorder, secondary structure, etc.) to search for enriched, short multilevel motifs in sets of proteins having a common function. As a first step, we take an amino acid sequence as an input and translate each residue into a single letter that reflects a combination of its intrinsic physicochemical and structural properties. Subsequently, we use the MEME search algorithm to search for enriched short motifs in the data, simultaneously examining different levels of information imbedded into the protein sequence.

MATERIALS AND METHODS

Data construction

Amphipathic helices

Twenty protein domain structures confirmed to possess an amphipathic helix (AH) based on data from Ref. [41] were extracted from the Protein Data Bank (PDB).[42] In addition, 10 control sets were extracted, each including 20 protein domains from the PDB annotated in SCOP (Version 1.75)[43] as "all beta" domains. Furthermore, a total of 305 proteins including AH were extracted from the UNIPROT database.[44] The original set was cleaned for redundancy to include only proteins with less than 35% sequence identity using PISCES.[45] Sequences shorter than 50 amino acids were removed, resulting in a final set of 55 proteins.

PCNA interacting protein

Human proteins annotated to interact with PCNA were extracted from the UNIPROT database. The dataset was cleaned for redundancy (removing sequences with higher than 35% identity) using PISCES.[45] The final set included 23 proteins. To extract structures from the PDB, the keyword "PCNA" was used, including all eukaryote complexes. The dataset was further cleaned for redundancy. Since the data included both full-length protein domains and short peptides, we used the "needle" pairwise alignment tool[46] and calculated the percent identity as the number of aligned amino acids divided by the minimal chain length. Due to the limited number of structures in the PDB, we included sequences sharing up to 50% sequence identity. Finally, the structural PCNA dataset included 13 structures of proteins/peptides.

TRAF6 substrates

Proteins were extracted from the Human Protein Reference Dataset (HPRD).[47] A total of 87 proteins reported to interact directly with TRAF1 (sharing up to 50% sequence identity) were analyzed.

Ca+2-binding proteins

Nineteen protein structures annotated to contain a calcium-binding loop[48] were extracted from the PDB.

The yeast protein–protein interaction dataset

The five most connected proteins were selected from CCSB-YI11 database,[49] including YLR291C with 82 partners, YLR423C with 58 partners, YIR038C with 46 partners, YDR510W with 40 partners, and YBR261C with 39 partners.

Motif search algorithm

An in-house modified version of MEME 4.4[27] was used as the motif search algorithm. The software was modified to adjust to the multilevel alphabets, supporting the new alphabet length (correcting for alphabet size). The following flags were used: –protein –mod oops –minw 6 –maxw 10 –spmap uni –prior dirichlet.

Predicting amino acid properties

The disorder property was calculated from the primary sequence employing the VSL2 predictor package.[50] The protein's secondary structure and surface accessibility were calculated using SSpro4.1.[51] For proteins for which structural data were available, the properties were extracted directly using DSSP, as inferred from PDBfinder.[52]

Information content calculation

Information content (IC) was calculated as described in MEME.[27] In summary, the IC at position i for each motif was calculated as in Formula (1):

$$\mathrm{IC}(i) = \log_2(N) - H(i) \quad (1)$$

where $H(i)$ is the entropy at position i and N is the number of letters in the alphabet. The total entropy (H) for each position was calculated as in Formula (2):

$$H = -(\mathrm{sum}\ f(\mathrm{aa}, \mathrm{i}) \times \log_2[f(\mathrm{aa},\ i)]) \quad (2)$$

where f(aa, i) is the frequency of amino acid (aa) at position i. The relative IC (RIC) at position i was calculated as in Formula (3):

$$\mathrm{RIC} = \mathrm{IC}(i)/\log_2(N) \quad (3)$$

RESULTS AND DISCUSSION

Outline of the approach

As a first step, we translated a polypeptide string to the new alphabet representing the protein sequence as a multilevel property string rather than a primary amino acid sequence [Fig. 1(A)]. To this end, each amino acid was classified to one of five groups based on its physiochemical properties: positive, negative, hydrophobic/aromatic, mild hydrophobic, and polar (see Table I). In addition, we characterized each residue based on different structural properties, such as the secondary structure in which it resides (alpha helix, beta strand, and coil), its propensity to be disordered vs. ordered, or the accessibility of the residue to the surface (buried vs. exposed). Structural properties were either extracted from the protein structures (when the structures were available in the PDB[42]) or predicted from the sequence using state-of-the-art prediction methods (as described in detail in the Materials and Methods section). Consequently, we combined the physiochemical classification with the structural classification (including either one or two structural properties) to a mutual classification which is assigned a letter from the English alphabet. Overall, the number of multilevel classifications (letters) cannot exceed 20 (the number of different letters in the original amino acid alphabet). For each combination of physicochemical and structural properties, we derived a unique alphabet: *ab_SS*, *ab_DIS*, *ab_SA*, *ab_SA*, and *ab_DIS* for secondary structure, disorder/order, surface accessibility, and a combination of surface accessibility and disorder, respectively. The letter assignments that are unique to each alphabet are presented in Table I and Supporting Information Table S1. For example, a glutamic acid in a given position in a protein that is predicted to be in an alpha helical structure will be assigned the letter "*F*" in *ab_SS*, while the same position can be assigned "*I*" in *ab_DIS* (see Table I). In addition, we translated the proteins in the PDB and UNIPROT databases into the three different multilevel alphabets (*ab_SS*, *ab_DIS*, and *ab_SA*), each reflecting a unique combination of physicochemical and structural properties. The translated databases were then used to execute the motif search algorithm (see Materials and Methods section) in an attempt to detect significant linear motifs that represent common properties shared by the proteins in any dataset of interest [Fig. 1(B)].

Detection of combined structural and sequence motifs from solved structures

To explore the competence of the new alphabet in finding common property motifs, we applied it to search for a well-known structural motif, namely, the AH. Several algorithms are available for screening sequences for AHs. The majority of these algorithms such as the drawing program Helical Wheel,[53] MPEX,[54] and Amphipaseek[55] are able to characterize AH patterns given a sequence of interest. HELIQUEST is a web server that searches for unique helical structures in a large database and enables searching for similar patterns in large test datasets.[56] Our method was designed for the *de novo*

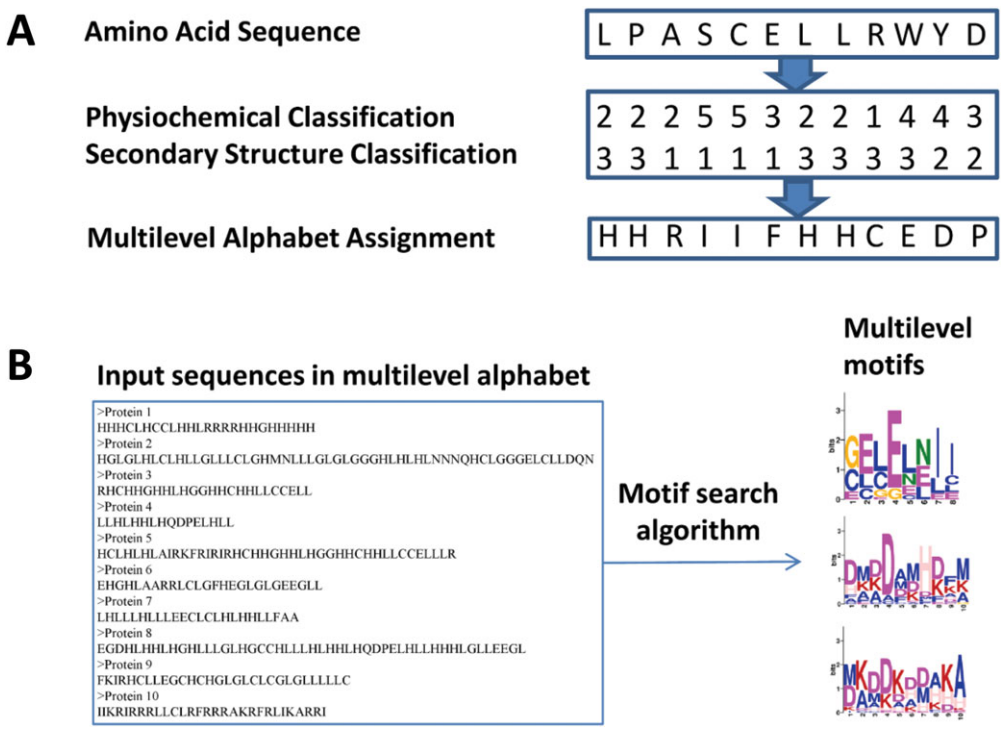

Figure 1

Flowchart representing the translation from an amino acid sequence into the new property alphabet. **A**: To translate a given sequence into a multilevel alphabet, each residue in the sequence is classified into one of five physiochemical groups: positive (+) = 1, negative (−) = 2, mild hydrophobic (MH) = 3, hydrophobic/aromatic (HA) = 4, and polar = 5. In addition, for each residue in the polypeptide string, different structural properties (such as secondary structure or surface accessibility) are calculated, and the residue is assigned a score per property (e.g., alpha helix = 1, beta strand = 2, coil = 3). Next, each residue is assigned a unique letter belonging to the English alphabet that represents the unique property combination (e.g., *F* in the multilevel alphabet represents a negative charged amino acid found in an alpha helical structure). The string representing multilevel properties encoded by the new alphabet is shown in the lower row. **B**: The set of sequences represented by the multilevel alphabet are used as input for the motif search algorithm in order to search for common multilevel motifs.

detection of combined structural and sequence motifs that are enriched in a given dataset. To test whether the approach can detect AH patterns enriched within a given dataset, we extracted a set of proteins that possess AHs from a dataset of proteins with known structures.[41] The sequences were translated into the new alphabet, which combines physiochemical and SS features, defined as *ab_SS* (Table I). Using a local version of MEME modified to support the multilevel alphabet, we detected an enrichment of the AH motif with an *E*-value of 3.4 $e-5$, which was the most significant motif [Fig. 2(A)]. Supporting Information Table S2 summarizes the results of AH predictions in 20 protein domains retrieved from PDB in comparison to the known AHs extracted from experimentally solved structures. As demonstrated in Supporting Information Table S2, in 19 of the 20 domains, we detected the significant motif that fell precisely within a known helix structure. As observed from the helical wheel presentations, which we applied to each predicted motif, the majority of the motifs (17/20) had a

Table I
The Multilevel Alphabets

	AA classifications	+: R, K	−: E, D	HA: F, W, Y	MH: A, I, L, V,P, M	P: P, S, T, N, Q, H, C, G
Secondary Structure (*ab_SS*)	α helix	*A*	*F*	*K*	*R*	*I*
	β sheet	*M*	*P*	*D*	*N*	*Q*
	Coil	*C*	*G*	*E*	*H*	*L*
Disorder (*ab_DIS*)	Order	*C*	*I*	*G*	*H*	*L*
	Disorder	*A*	*F*	*E*	*D*	*K*
Surface Accessibility (*ab_SA*)	Exposed	*C*	*I*	*G*	*H*	*L*
	Buried	*A*	*F*	*E*	*D*	*K*

+, positive aa; −, negative aa; HA, hydrophobic; MH, mild hydrophobic; P, polar.

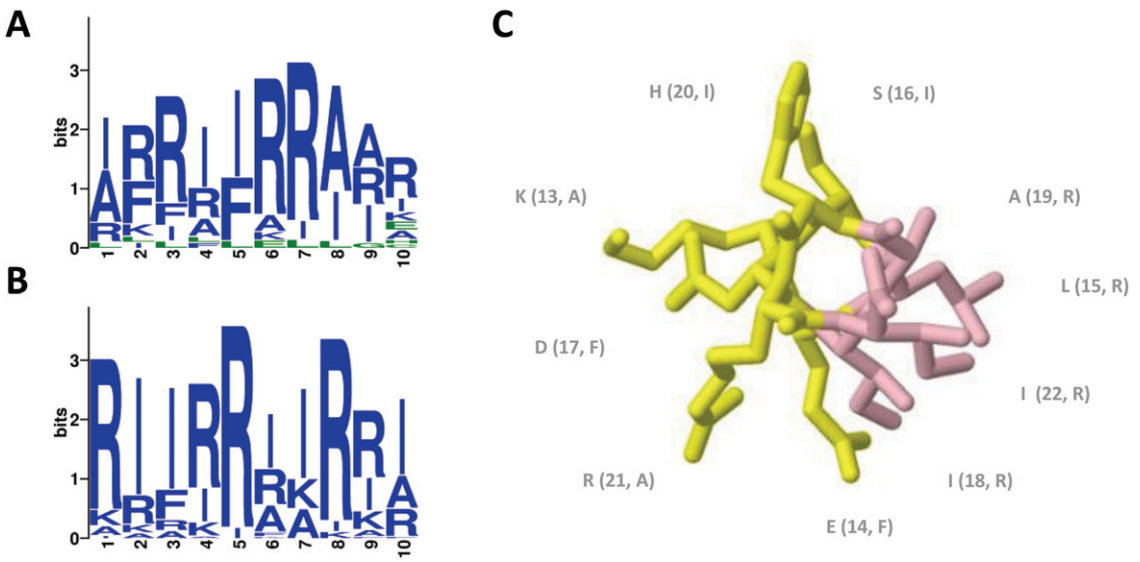

Figure 2

Predicting a common AH motif employing the multilevel alphabet. **A**: The AH motif detected from structural data using the multilevel *ab_SS* alphabet. As demonstrated, hydrophobic (R) residues are separated by polar (I) and charged (A) residues, which are all found in an α helix structure background (blue color). **B**: The AH motif detected from primary amino acid sequences extracted from UNIPROT using the multilevel *ab_SS* alphabet. **C**: A 3D structure of the human aldolase (PDB ID 1ALD positions 13–22) AH, which was predicted correctly by the multilevel *ab_SS* alphabet. Residues are colored based on physicochemical properties: hydrophobic in pink, and polar and charged amino acids in yellow. The amino acid positions are shown in brackets followed by the letter representing the residue in the multilevel *ab_SS* alphabet (see Table I).

perfect amphipathic pattern; for the remaining three, we noticed that while the distribution of polar vs. hydrophobic residues was unisotropic, they were not purely amphipathic. In Figure 2(C), we demonstrate a 3D sketch of one of the predicted AHs found in human aldolases proteins (PDB ID 1ALD). To further evaluate the statistical power of our method, we selected ten random control sets of 20 protein domains from the PDB that were defined in SCOP as "all beta" and were thus not expected to contain AHs. We first ran the MEME program on each of the control sets and, as anticipated, could not detect any motif that resembled the AH pattern (Supporting Information Table S3). We further combined the sequences of the 20 protein domains containing known AHs with a randomly selected set of 20 domains from the "all beta" set. Consequently, we ran MEME using the -zoops flag (which allows the detection of zero or one occurrence per motif). In each case, we selected the best motif that possessed an AH pattern and counted the number of sequences from each group (the original and the control sets) in which the motif was detected. We repeated this test ten times, each time selecting an independent control set. Finally, we calculated the sensitivity and specificity for each test. The results of the latter tests are summarized in Supporting Information Table S4. As shown, the sensitivity in all cases was 1, meaning that in all tests, the motif was detected in all of the 20 sequences known to possess the AH. Nevertheless, the specificity had a wider range, spanning from 0.4 to 0.75 with an average False Detection Rate (FDR) of 0.27. To try and decipher the relatively high FDR, we analyzed the structures of the proteins in the control set in which we detected the AH motifs and surprisingly found that 57 of the 77 predicted motifs were actually true AHs that were found in domains that are mostly beta (though they were annotated in SCOP as "all beta"). Thus, we concluded that our approach is highly sensitive in detecting motifs that are composed of sequence and structural elements.

Detection of combined structural and sequence motifs from the primary sequence

Next we sought to examine whether the motif can be detected given merely the primary sequence. As a first step, we extracted the primary sequences of the proteins for which we had structures available and repeated the procedure described above, this time relying on sequence information alone. We further predicted the secondary structure of the domains and consequently translated it into the multilevel *ab_SS* alphabet combining physiochemical and SS features (see Table I). As shown in the last column in Supporting Information Table S2, 18 of the 20 sequences were verified as AH (as validated by the secondary structure prediction and the helical wheel program). Furthermore, for 16 of the motifs that were predicted from the structural dataset, we could match a motif that was detected given sequence data only (ten of which were predicted as the most significant motifs).

While the results obtained based on information from the primary sequence alone were encouraging, one can still argue that in cases in which the structure is available, secondary structure predictions are much more reliable in comparison to proteins for which the 3D structure is not available in the database. Thus, to validate the method given only sequence data, we extracted a large dataset of protein sequences from the UNIPROT database that were annotated as possessing AHs (see Materials and Methods section). Here, again, we first predicted the secondary structure of the full-length protein and then translated it into the multilevel *ab_SS* alphabet, which combines physiochemical properties and SS (see Table I). As illustrated in Figure 2(B), when applying our method on a set of 55 proteins, we could clearly see that the best detected motif (*E*-value of 1.3 $e-138$) had the characteristic signature of AH, that is, the expected periodicity of the hydrophobic amino acid found in a α helix background. The list of motifs predicted in each sequence is given in Supporting Information Table S5. Notably, when applying MEME on the primary amino acid sequences (with no added information), we couldn't detect any significant motif that resembled an AH pattern.

To demonstrate the applicability of the multilevel alphabet to pattern searching, we translated all *Saccharomyces cerevisiae* nuclear proteins from the SWISSPROT database into the new *ab_SS* alphabet (Table I). Applying the PROSITE search algorithm,[57] we searched for the AH pattern in the translated data using a regular expression that we defined based on the previously identified motif [RK][AIF][AIF][RK][RK][AIF][AIF][RK]X[AIF] (X represents any letter, RK represents hydrophobic and hydrophobic/aromatic groups in α helical structure, and AIF represents polar and charged groups in α helical structure). Overall, of the 1595 sequences, we identified 614 proteins with at least one AH motif (of length 10). To validate our results, we selected all proteins from the latter group for which a 3D structure was available in the PDB. Overall, 57 of the 614 proteins had structures, and of this number, we predicted 71 occurrences of an AH motif. We further identified the location of the predicted motif within the structure and defined the secondary structure of each residue in the solved structure. As shown in Supporting Information Table S6, in 52 of the 71 predicted motifs, all residues of the predicted motif were found within a helical structure. The rest of the predicted motifs were located mainly within helices, excluding one or two amino acids, usually located at the edges of the motifs, and were predicted to be in a turn or coil. Only in two cases (no. 35 and no. 56 in Supporting Information Table S6), the predicted motif was not found within a helix structure. We further tested each of the predicted motifs using the helical wheel drawing tool, and confirmed that 69 of the 71 motifs had a perfect amphipathic pattern (Supporting Information Table S6). Overall, these results strongly reinforce that our method is powerful in predicting multilevel motifs from large-scale data.

Comparison to different search methods

As described above, we suggest a novel way of detecting functional motifs that reflect both physicochemical and structural information. The algorithm we chose for this purpose was MEME,[27] as it rests on a solid statistical foundation and has been used for many years for motif searching. Nevertheless, we also compared the output of our encoding method with other motif search methods, including PRATT,[57] SlimFinder,[24] and DILIMOT,[23] using the dataset of the 20 characterized AHs extracted from the PDB.[41] In Supporting Information Table S7, we present the results of the best enriched motif detected by each motif search algorithm using the original amino acid sequence and the multilevel *ab_SS* alphabet. As shown, all three methods tested failed to uncover the AH pattern from the primary amino acid sequence. Nevertheless, the AH characteristic (periodicity of the hydrophobic amino acid in a α helix structure) was detected by SLIMFinder and was implied by the other methods when using the multilevel alphabet as an input (combining the physicochemical properties with the SS structural features). However, while the different methods were able to weakly capture the multilevel motif, MEME clearly outperformed the other methods (Supporting Information Table S7) and was thus selected by us as the method of choice.

Revealing additional structural information for previously known motifs

PCNA-binding motif

Following the success of our approach in identifying the well-characterized AH motif, we were interested in examining whether we could employ different multilevel alphabets for detecting common structural features of previously known functional motifs. The proliferating cell nuclear antigen (PCNA) protein plays an important role in DNA replication. PCNA can interact with several different partners to regulate different reactions.[58] It has been shown previously that PCNA partners bind PCNA via a conserved binding motif QXX(h)XX(a)(a), where h and a represent moderately hydrophobic and hydrophobic amino acids, respectively.[59] We applied MEME to the UNIPROT PCNA dataset (see Materials and Methods section), which was translated into the multilevel *ab_DIS* alphabet, in which information regarding the order/disorder state of each residue in the sequence is incorporated (Table I). Here, we found that the best detected motif (*E*-value 1.5 $e-21$) encapsulated the known PCNA binding motif,[60] suggesting that the conserved binding site is structurally disordered [Fig. 3(A)]. A full list of predicted

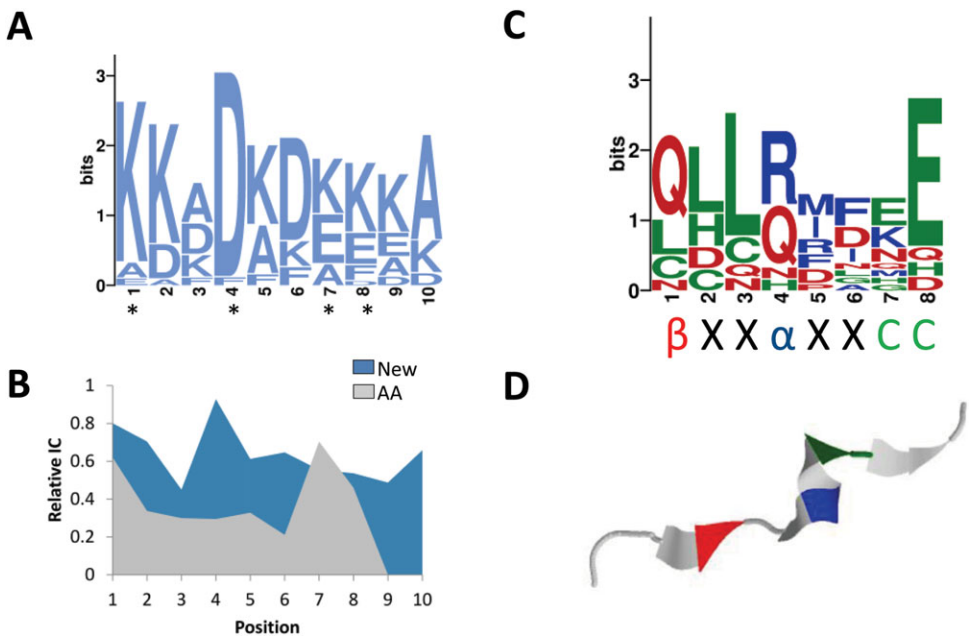

Figure 3

A: The PCNA-binding motif predicted from the sequence using the multilevel *ab_DIS* alphabet. The known conserved positions are marked with asterisks. The expected motif KXXDXXEE indicates that the contact site with the PCNA that we predict is disordered (alphabet given in Table I). **B**: The RIC calculated for the predicted PCNA-binding motif. In light blue is the RIC calculated from the PSPM representing the motif derived from the *ab_DIS* alphabet, and in light gray is the RIC calculated from the PSPM derived from the primary amino acid *ab_AA* alphabet. **C**: The PCNA-binding motif (*E*-value 2.1 *e*−9) predicted from the structural data translates into the multilevel *ab_SS* alphabet. Under the logo is the consensus secondary structure of the PCNA-binding motif, as predicted by our method. In red, blue, and green are the β sheet, α helix, and coiled coil, respectively. **D**: A 3D structure representation of one of the structures in our PCNA-binding peptides (PDB ID 1VYJ).

PCNA motifs is given in Supporting Information Table S8. Interestingly, when applying MEME to the primary sequences, we detected the PCNA-like motif only with a non-significant *E*-value of 2.3 *e*+13, which is beyond the threshold of MEME. Notably, the motif extracted from the primary amino acid sequence lacked the conserved amino acid in the 4th position. Furthermore, the multilevel motif detected with the *ab_DIS* alphabet had significantly higher IC relative to the motif detected from the primary sequences [Fig. 3(B)]. Finally, when applying the PROSITE search algorithm to the translated sequences with the regular expression found by MEME (which is similar to the PCNA-binding motif but lacks the final hydrophobic positions), the motif was detected in 21 of the 23 proteins. To evaluate the sensitivity of the method to detect the real binding motif, we divided the proteins in the UNIPROT PCNA database into non-overlapping segments, including 21 regions containing the known PCNA motif and 703 regions lacking the motif. Overall, while the specificity of our method to detect the disordered PCNA motif was very high (0.96), the sensitivity (recall rate) was relatively low (0.38). These results were comparable to the results we obtained employing ANCHOR methods[61,62] on the same dataset searching for the PCNA PROSITE motif in regions predicted to be involved in binding by the ANCHOR algorithm.[61,62] Overall, the specificity and sensitivity of the latter method were 0.84 and 0.43, respectively.

In an attempt to find additional features characterizing the PCNA-binding site, we ran MEME on a subset of PCNA-binding proteins (see Materials and Methods section) that was translated to *ab_SS* (see Table I). As shown in Figure 3(C), the motif detected when using the multilevel alphabet with the secondary structure as the structural feature suggested a preference for β strand in the first position, α helix in the middle position, and a coil unstructured region in the two last positions [Fig. 3(C)]. A full list of motif hits in the entire dataset is given in Supporting Information Table S9. The structural preference detected by our method is in accordance with the finding of Sakurai *et al.*,[63] who found that two β sheet regions connected to a short α helix via a short coil unstructured region represents the main interacting interface with PCNA. To further validate our results, we examined a small engineered peptide that binds PCNA (PDB ID 1VYJ: B), and indeed the secondary structure pattern of the motif was consistent with the motif we detected using our new multilevel alphabet [Fig. 3(D)].

When employing our method to the PCNA-binding regions (instances) from the ELM database[21] that were verified experimentally to include PCNA-binding motifs, we clearly identified the known motif as being

intrinsically disordered. To assess whether the motif is predicted to be accessible, we employed the multilevel *ab_SA* alphabet. Interestingly, we found that the motif was predicted to be partially exposed in the N terminal side and buried in the C terminal side (Supporting Information Fig. S1). While these results could be a result of erroneous SA predictions that could arise from the hydrophobic nature of the motif they may imply that the binding motif is preferentially found in partially exposed regions. This notion is consistent with the hypothesis that intrinsically disordered regions are prone to protease digestion and are thus expected to be fully or partially protected in their unbound state.[9] To further test this hypothesis, we selected three other motifs from the ELM database having a highly conserved motif that are natively disordered, but due to their conserved hydrophobic nature, are predicted to be buried and misidentified. In all three cases tested (LIG_NRBOX, LIG_PPI, LIG_EH1), we clearly identified the sequential motif using the primary amino acid sequence *ab_AA* and found that the motif was significantly disordered (using the multilevel *ab_DIS* alphabet). Nevertheless, all motifs were predicted to be mostly or partially buried. When employing the combined *ab_DIS* and *ab_SA* alphabet, we could clearly see that the motif is partially disordered and buried, and partially disordered and exposed, as in the case of the PCNA motif (Supporting Information Fig. S1).

Tumor Necrosis Factor Receptor 6 (TRAF6)

TRAF6 is an E3 ubiquitin ligase that mediates the synthesis of "Lys-63"-linked-polyubiquitin chains conjugated to proteins. The ubiquitination process can be mediated by numerous binding partners that bind to TRAF6, such as p62, TAB2, TAB3, and so forth, altering the substrate protein set. The TRAF6/p62 ubiquitination motif was identified by Jadhav *et al.*[13,64] who suggested that approximately 50% of ubiquitin high-confident predicted sites were located in regions predicted to be in loops or in structurally disordered regions. Moreover, 84% of the high-confident sites were proposed to be exposed. To test whether our method could capture the unique properties of TRAF6-binding sites, we extracted a set of sequences from the HPRD database that were confirmed experimentally to interact directly with TRAF6 but not necessarily mediated by p62 (see Materials and Methods section). Subsequently, we searched for an enriched motif in the original amino acid sequence, as well as in sequences that were translated into the different multilevel *ab_SS*, *ab_SA* and *ab_DIS* alphabets. Analyzing the amino acid sequence alone, we managed to identify a significant motif with lysine in the second position. This motif was detected in 47% of the proteins in our dataset (Supporting Information Fig. S2). Nevertheless, employing the motif search algorithm on the sequences translated into *ab_SA* (Table I), we found the motif DADX-X[ED]KD, where A represents the ubiquitinized lysine surrounded by the hydrophobic environment with a hydrophobic–polar hydrophobic tail (Supporting Information Fig. S2 and Table S10). This motif matches the ubiquitination-binding site that was suggested previously by Jadhav *et al.*[13,64] To test the significance of this motif, we ran MEME on 10 randomly selected human datasets extracted from UNIPROT (87 proteins each) translated into the multilevel *ab_SA* alphabet (Table I). Employing the CompareMotif tool,[65] we tested the regular expressions extracted from the random datasets and the TRAF6 motif detected with the multilevel alphabet, and compared it to the regular expression built according to Jadhav *et al.*[13] Overall, none of the motifs detected in the random datasets contained the known TRAF6 motifs. Interestingly, the motif predicted in the positive set that was translated into *ab_SA* was predicted to be buried and not exposed, as suggested by Jadhav *et al.*[13,64] This discrepancy could be explained by the different datasets used in both cases, keeping in mind that Jadhav and coworkers searched only for TRAF6/p62-binding sites, while we examined all known TRAF6 partners. Notably, in Jadhav *et al.*,[13,64] 16% of the predicted TRAF6-binding sites were predicted to be buried, as we had anticipated. It will be interesting to explore further whether the new alphabet actually refines motif detection and extracts false positive sequences that lack the unique property, or whether this property is not essential.

Detecting structural features that coincide with known sequence motifs

Calcium-binding proteins

Calcium-binding proteins (CaBPs) are important cell regulators that participate in many processes, such as cell division, differentiation, motility, and programmed cell death. Many different CaBPs and different Ca+2-binding sites possess the well-characterized Ca+2-binding motif DX[DN]XDG. This motif is known to be located in loops and it shares a common binding mode. In an attempt to highlight the structural properties common to Ca+2-binding sites, we extracted CaBPs from the PDB based on the dataset from Ref.[48]. We searched for enriched motifs in the primary amino acid sequence and the translated datasets (based on the alphabet using disorder, SS and SA as structural features). As expected, when using the amino acid sequence alone, we found that the most significant motif (*E*-value 8.5 *e*−4) DX[DN]XDG exactly matched the known Ca+2 motif,[48] overlapping 88% of the verified motifs (Supporting Information Fig. S3). When we used the multilevel *ab_SS* alphabet, we identified an enriched motif (overlapping 35% of the verified motifs) in which all residues were predicted to be in a coil structure (Supporting Information Fig. S3).

In addition, when searching for motifs within the translated dataset using *ab_DIS*, we indeed found that the Ca+2-binding motifs (which overlapped 59% of the verified motifs) were located in ordered regions as previously shown[48] (Supporting Information Fig. S3). Finally, we repeated the procedure with *ab_SA* and again found the significant Ca+2-binding motif in a buried environment (Supporting Information Fig. S3); these latter motifs matched 53% of the verified motifs. Based on these results, we concluded that the majority of Ca+2-binding motifs are found in coil structures that are mostly buried and relatively rigid (ordered). While these results are non-intuitive, they are in high accordance with previous structural studies of the Ca+2-binding site, suggesting that the DX[DN]XDG motif, unlike many other linear motifs,[7] appears in relatively rigid coil regions.[48] These results again exemplify the promise of our multilevel motif search approach to identify common structural properties of functional motifs in large datasets.

Searching for common motifs in the yeast interactome

It has been estimated that 15–40% of human domain-motif interactions are mediated by short linear motifs, however, to date only a small fraction of these motifs has been discovered.[66] The role of linear motifs as mediators of PPIs was demonstrated by Neduva *et al.*,[2] who rediscovered known motifs and predicted many novel ones. We applied our approach to search for new unknown motifs in the yeast *S. cerevisiae* PPI network. To this end, we extracted the yeast CCSB-YI11 interactome from the CCSB interactome DB,[49] and translated the amino acid sequences of the partners into our new alphabet using either SS, SA, or disorder as the structural feature. Here, we present the results for the five most connected proteins (Supporting Information Table S11). The motif search results using the multilevel alphabets were compared to the results using the primary amino acid alphabet. Overall, we found that the motifs detected using the new alphabet had higher IC compared to the amino acid sequence motifs (Supporting Information Fig. S4). Specifically, the motif detected among the partners of the autophagy-related protein 17 (YLR423C) when employing the *ab_DIS* alphabet K[KA]D[KA][KA][DE][DK][KF]K[KDF] showed a periodicity of positive charged/polar residues and mild hydrophobic amino acids, all predicted to be disordered. To test the significance of the latter motif, we calculated the occurrence of the motif in all *S. cerevisiae* proteins downloaded from the UNIPROT database, which were further translated into the multilevel *ab_DIS* alphabet. The motif detected was present in 19 of the 58 proteins in the autophagy-related protein 17 dataset (32.7%) compared to 817 of the 7763 proteins in the entire *S. cerevisiae* dataset from SwissProt (10.5%). We further confirmed that this enrichment is indeed significant only among the partners of the autophagy-related protein 17 based on the hyper geometric tail distribution (P-value = 1.5 $e-7$ using the Fisher exact test). We suggest that there may be signals in PPIs that are not characterized by the amino acid sequence, rather by a cluster of physiochemical properties combined with other properties that mediate the interactions. Such examples are well-known in protein–RNA interactions, where the secondary structure of the RNA is critical for recognition.[67] Based on our analysis, we suggest that novel motifs combining sequence and structural information are possibly involved in mediating many other PPIs. While these motifs could not be detected at the amino acid sequence level, they may play an important functional role in mediating the interactions. Clearly, further experimental testing will be required to confirm the importance of these motifs.

CONCLUSIONS

The involvement of short linear motifs in different protein functions and specifically in protein interactions has been broadly demonstrated.[2,4] While some linear motifs have been well-characterized,[17,20] many of them are less understood[59,64,68] and some are yet to be confirmed.[2] The preference of the short linear motif to be located within a specific protein environment (i.e., a certain secondary structure, surface accessibility, disordered regions) has been widely demonstrated.[4,8,14,22,69] Some of the available motif search algorithms use filtering procedures to account for these preferences.[17,22,23] In this study, we present a novel approach for the *de novo* detection of short linear motifs that simultaneously take into account the common physiochemical properties as well as other structural properties. To this end, we used a modified version of the MEME software and searched for common motifs in the translated protein sequence rather than the primary amino acid sequence. The motifs are represented by the graphical logo representation, demonstrating the combined sequence and property motif.

Employing our approach in several case studies, we were able to extract known structural motifs from linear sequences such as the AH, which represents a structural motif with no sequence conservation. Moreover, we demonstrated the ability of our method to reveal additional properties that characterized known sequence motifs, such as in the case of PCNA- and Ca+2-binding motifs. We further demonstrated the applicability of the method to screen for true positive motifs that, in addition to the consensus sequence, require a certain structural feature in order to be functional. Finally, we demonstrated the potential of the method to search for novel common motifs in yeast PPI networks. These predicted motifs can be further implemented in machine learning algorithm to search for new partners of a given protein or its

homologue within the same organism or in other related species. The advantage of our multilevel motif search approach was demonstrated in datasets of proteins whose 3D structure is known, as well as for protein sets for which structural information was predicted from the sequence. Our multilevel motif search approach has a great advantage over other methods as it is fast and simple, and can be applied to very large datasets. The novelty of the approach lies in the encoding and in the purported ability of the encoding to identify functional domains that might otherwise remain hidden when considering the primary sequence alone. Clearly, further investigation is required to reveal the biological functions of these motifs.

ACKNOWLEDGMENTS

Authors thank Zohar Yakhini and Rachel Kolodny for their helpful suggestions.

REFERENCES

1. Finn R, Mistry J, Tate J, Coggill P, Heger A, Pollington J, Gavin O, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer E, Eddy S, Bateman A. The Pfam protein families database. Nucleic Acids Res 2010;38 (Database issue):D211–D222.
2. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB. Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 2005;3:e405.
3. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
4. Neduva V, Russell RB. Linear motifs: evolutionary interaction switches. FEBS Lett 2005;579:3342–3345.
5. Dow LK, Jones DN, Wolfe SA, Verdine GL, Churchill ME. Structural studies of the high mobility group globular domain and basic tail of HMG-D bound to disulfide cross-linked DNA. Biochemistry 2000;39:9725–9736.
6. Cokol M, Nair R, Rost B. Finding nuclear localization signals. EMBO Rep 2000;1:411–415.
7. Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ. Attributes of short linear motifs. Mol Biosyst 2012;8:268–281.
8. Gsponer J, Babu M. The rules of disorder or why disorder rules. Prog Biophys Mol Biol 2009;99:94–103.
9. Dunker A, Brown C, Lawson J, Iakoucheva L, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–6582.
10. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2006;2:e100.
11. Tafvizi A, Huang F, Fersht AR, Mirny LA, van Oijen AM. A single-molecule characterization of p53 search on DNA. Proc Natl Acad Sci USA 2011;108:563–568.
12. Vuzman D, Levy Y. Intrinsically disordered regions as affinity tuners in protein–DNA interactions. Mol Biosyst 2011;8:47–57.
13. Jadhav TS, Wooten MW, Wooten MC. Mining the TRAF6/p62 interactome for a selective ubiquitination motif. BMC Proc 2011;5 (Suppl 2):S4.
14. Via A, Gould CM, Gemünd C, Gibson TJ, Helmer-Citterich M. A structure filter for the Eukaryotic Linear Motif Resource. BMC Bioinformatics 2009;10:351.
15. Davey NE, Shields DC, Edwards RJ. Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics (Oxford, England) 2009;25: 443–450.
16. Haslam NJ, Shields DC. Profile-based short linear protein motif discovery. BMC Bioinformatics 2012;13:104.
17. Puntervoll P, Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferrè F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Küster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31:3625–3630.
18. Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 2003;31:3635–3641.
19. Balla S, Thapar V, Verma S, Luong T, Faghri T, Huang CH, Rajasekaran S, del Campo JJ, Shinn JH, Mohler WA, Maciejewski MW, Gryk MR, Piccirillo B, Schiller SR, Schiller MR. Minimotif Miner: a tool for investigating protein function. Nat Methods 2006;3:175–177.
20. Rajasekaran S, Balla S, Gradie P, Gryk MR, Kadaveru K, Kundeti V, Maciejewski MW, Mi T, Rubino N, Vyas J, Schiller MR. Minimotif miner 2nd release: a database and web system for motif search. Nucleic Acids Res 2009;37 (Database issue):D185–D190.
21. Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, Jödicke L, Dammert MA, Schroeter C, Hammer M, Schmidt T, Jehl P, McGuigan C, Dymecka M, Chica C, Luck K, Via A, Chatr-Aryamontri A, Haslam N, Grebnev G, Edwards RJ, Steinmetz MO, Meiselbach H, Diella F, Gibson TJ. ELM—the database of eukaryotic linear motifs. Nucleic Acids Res 2011;40(D1):D242–D251.
22. Davey NE, Edwards RJ, Shields DC. The SLiMDisc server: short, linear motif discovery in proteins. Nucleic Acids Res 2007;35 (Web Server issue):W455–W459.
23. Neduva V, Russell RB. DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res 2006;34(Web Server issue):W350–W355.
24. Davey NE, Haslam NJ, Shields DC, Edwards RJ. SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. Nucleic Acids Res 2010;38 (Web Server issue):W534–W539.
25. Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol 2007;3:e39.
26. Mehdi AM, Sehgal MS, Kobe B, Bailey TL, Boden M. DLocalMotif: a discriminative approach for discovering local motifs in protein sequences. Bioinformatics (Oxford, England), Bioinformatics. 2013; 29:39-46.
27. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 1994;2:28–36.
28. Do CB, Batzoglou S. What is the expectation maximization algorithm? Nat Biotechnol 2008;26:897–899.
29. Tan SH, Hugo W, Sung WK, Ng SK. A correlated motif approach for finding short linear motifs from protein interaction networks. BMC Bioinformatics 2006;7:502.
30. Hugo W, Ng SK, Sung WK. D-SLIMMER: domain-SLiM interaction motifs miner for sequence based protein-protein interaction data. J Proteome Res 2011;10:5285–5295.
31. Weatheritt R, Luck K, Petsalaki E, Davey N, Gibson T. The identification of short linear motif-mediated interfaces within the human interactome. Bioinformatics (Oxford, England) 2012;28:976–982.
32. Mooney C, Pollastri G, Shields DC, Haslam NJ. Prediction of short linear protein binding regions. J Mol Biol 2011;415:193–204.
33. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–809.
34. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6:1033–1038.
35. Li T, Fan K, Wang J, Wang W. Reduction of protein sequence complexity by residue grouping. Protein Eng 2003;16:323–330.

36. Edgar RC. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. Nucleic Acids Res 2004;32:380–385.
37. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000;13:149–152.
38. Zhang J, Chen R, Liang J. Empirical potential function for simplified protein models: combining contact and local sequence-structure descriptors. Proteins 2006;63:949–960.
39. Melo F, Marti-Renom MA. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. Proteins 2006; 63:986–995.
40. Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBS Lett 2004;576:348–352.
41. Bajaj M. Development of a prediciton method for amphipathic α-helices from protein primary structure. University of Nebraska; 2005.
42. Berman H, Battistuz T, Bhat T, Bluhm W, Bourne P, Burkhardt K, Feng Z, Gilliland G, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook J, Zardecki C. The protein data bank. Acta Crystallogr D Biol Crystallogr 2002;58 (Part 6):899–907.
43. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
44. Apweiler R, Bairoch A, Wu C, Barker W, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin M, Natale D, O'Donovan C, Redaschi N, Yeh L. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004;32 (Database issue):D115–D119.
45. Wang G, Dunbrack RJ. PISCES: a protein sequence culling server. Bioinformatics (Oxford, England) 2003;19:1589–1591.
46. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000;16:276–277.
47. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database—2009 update. Nucleic Acids Res 2009;37(Database issue):D767–D772.
48. Rigden DJ, Woodhead DD, Wong PW, Galperin MY. New structural and functional contexts of the Dx[DN]xDG linear motif: insights into evolution of calcium-binding proteins. PLoS One 2011;6:e21507.
49. CCSB Interactome Database. http://interactome.dfci.harvard.edu/index.php?page=home.
50. Vucetic S, Brown C, Dunker A, Obradovic Z. Flavors of protein disorder. Proteins 2003;52:573–584.
51. Cheng J, Randall A, Sweredoski M, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005;33 (Web Server issue):W72–W76.
52. Hooft RW, Sander C, Scharf M, Vriend G. The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. Comput Appl Biosci 1996;12:525–529.
53. http://cti.itc.virginia.edu/∼ cmg/Demo/wheel/wheelApp.html. Helical wheel.
54. Snider C, Jayasinghe S, Hristova K, White SH. MPEx: a tool for exploring membrane proteins. Protein Sci 2009;18:2624–2628.
54. Sapay N, Guermeur Y, Deléage G. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. BMC Bioinformatics 2006;7:255.
56. Gautier R, Douguet D, Antonny B, Drin G. HELIQUEST: a web server to screen sequences with specific alpha-helical properties. Bioinformatics (Oxford, England) 2008;24:2101–2102.
57. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 2006;34 (Web Server issue):W362–W365.
58. Maga G, Hubscher U. Proliferating cell nuclear antigen (PCNA): a dancer with many partners. J Cell Sci 2003;116(Part 15):3051–3060.
59. Warbrick E. PCNA binding through a conserved motif. Bioessays 1998;20:195–199.
60. Xu H, Zhang P, Liu L, Lee MY. A novel PCNA-binding motif identified by the panning of a random peptide display library. Biochemistry 2001;40:4512–4520.
61. Dosztanyi Z, Meszaros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics (Oxford, England) 2009;25:2745–2746.
62. Meszaros B, Simon I, Dosztanyi Z. Prediction of protein binding regions in disordered proteins. PLoS Computat Biol 2009;5: e1000376.
63. Sakurai S, Kitano K, Yamaguchi H, Hamada K, Okada K, Fukuda K, Uchida M, Ohtsuka E, Morioka H, Hakoshima T. Structural basis for recruitment of human flap endonuclease 1 to PCNA. EMBO J 2005;24:683–693.
64. Jadhav T, Geetha T, Jiang J, Wooten MW. Identification of a consensus site for TRAF6/p62 polyubiquitination. Biochem Biophys Res Commun 2008;371:521–524.
65. Edwards RJ, Davey NE, Shields DC. CompariMotif: quick and easy comparisons of sequence motifs. Bioinformatics (Oxford, England) 2008;24:1307–1309.
66. Neduva V, Russell RB. Peptides mediating interaction networks: new leads at last. Curr Opin Biotechnol 2006;17:465–471.
67. Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA 2010;16:1096–1107.
68. Rüdiger S, Germeroth L, Schneider-Mergener J, Bukau B. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. EMBO J 1997;16:1501–1507.
69. Dasgupta B, Nakamura H, Kinjo AR. distinct roles of overlapping and non-overlapping regions of hub protein interfaces in recognition of multiple partners. J Mol Biol 2011;411:713–727.