Chapter 9
Bioinformatics Approaches for Predicting Disordered Protein Motifs

Pallab Bhowmick, Mainak Guharoy and Peter Tompa

Abstract Short, linear motifs (SLiMs) in proteins are functional microdomains consisting of contiguous residue segments along the protein sequence, typically not more than 10 consecutive amino acids in length with less than 5 defined positions. Many positions are 'degenerate' thus offering flexibility in terms of the amino acid types allowed at those positions. Their short length and degenerate nature confers evolutionary plasticity meaning that SLiMs often evolve convergently. Further, SLiMs have a propensity to occur within intrinsically unstructured protein segments and this confers versatile functionality to unstructured regions of the proteome. SLiMs mediate multiple types of protein interactions based on domain-peptide recognition and guide functions including posttranslational modifications, subcellular localization of proteins, and ligand binding. SLiMs thus behave as modular interaction units that confer versatility to protein function and SLiM-mediated interactions are increasingly being recognized as therapeutic targets. In this chapter we start with a brief description about the properties of SLiMs and their interactions and then move on to discuss algorithms and tools including several web-based methods that enable the discovery of novel SLiMs (*de novo* motif discovery) as well as the prediction of novel occurrences of known SLiMs. Both individual amino acid sequences as well as sets of protein sequences can be scanned using these methods to obtain statistically overrepresented sequence patterns. Lists of putatively functional SLiMs are then assembled based on parameters such as evolutionary sequence conservation, disorder scores, structural data, gene ontology terms and other contextual information that helps to assess the functional credibility or significance of these motifs. These bioinformatics methods should certainly guide experiments aimed at motif discovery.

P. Tompa (✉) · M. Guharoy · P. Bhowmick
VIB Department of Structural Biology, Vrije Universiteit Brussel (VUB), Building E, Pleinlaan 2, 1050 Brussels, Belgium
e-mail: mainak.guharoy@vib-vub.be

P. Tompa
Institute of Enzymology, Research Center of Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
e-mail: ptompa@vub.ac.be

I. C. Felli, R. Pierattelli (eds.), *Intrinsically Disordered Proteins Studied by NMR Spectroscopy,* Advances in Experimental Medicine and Biology,
DOI 10.1007/978-3-319-20164-1_9

Keywords Protein sequence · Short linear motifs · Motif prediction · Intrinsic disorder · Post-translational modification · PTM · Multiple sequence alignments · Position-specific weight matrix · PWM · Protein interaction · Evolutionary conservation

1 Introduction

Protein modularity is a central and recurrent theme in our understanding of protein function. The basic functioning of almost all proteins occurs by the interaction of its modules with various other partners (proteins, nucleic acids, small molecules, etc.). Each module has a defined set of function(s) (eg, interactions with specific partners) that is linked to its surface characteristics, shape and structural dynamics and the variety of functions that a protein can carry out is closely linked to the number and types of modules it contains (Bhattacharyya et al. 2006). These modules include globular domains, *S*hort *Li*near *M*otifs (SLiMs) or other *Mo*lecular *R*ecognition *F*eatures (MoRFs). The presence of these elements in a given protein will determine its function by specifying its set of interaction partners.

Protein domains possess well-defined three dimensional structures with the members of any given domain family sharing strong and clearly visible evolutionary relationships; domain signatures are therefore comparatively easy to detect from protein primary sequence using information contained in databases such as Pfam (Finn et al. 2014) and Prosite (Sigrist et al. 2013). Domain structures can also be predicted reliably using *in silico* methods such as homology modelling based on sequence-structure alignments and this is now done routinely in protein structure prediction competitions like CASP (*C*ritical *A*ssessment of protein *S*tructure *P*rediction) (Moult et al. 2014). The Protein Data Bank (PDB) currently has more than 100,000 deposited structures that have accumulated rapidly over the past few decades (Berman et al. 2013), and most of the domain types are now thought to have been discovered.

At present, scientists are focusing not only on structured regions of the proteome but also on the disordered regions in search of functional modules (Tompa 2012; Habchi et al. 2014). In eukaryotes, up to 33% of the proteome may have putative long disordered segments (defined as >30 consecutive disordered residues) (Ward et al. 2004). Contained within these disordered regions, there may be a million or more estimated peptide motifs (SLiMs) existing in the proteome (Tompa et al. 2014) although relatively few of them have been discovered and experimentally validated so far. Work over the past decade has brought to the forefront the importance of sequence (peptide) motifs in protein function. These motifs are typically found at functional sites of proteins like cleavage sites, binding sites, sites for post-translational modifications and sub-cellular targeting sites. Some of the functions mediated by peptide motifs include specific protein-protein interactions, regulatory functions and signal transduction (Van Roey et al. 2014). The large number of annotated motifs in the *E*ukaryotic *L*inear *M*otif (ELM) database (Dinkel et al. 2014) provide overwhelming evidence of the fact that linear motifs are a ubiquitous and essential part of cellular biology.

Although clearly very abundant, true positive (ie, functional) linear motif instances are difficult to predict *de novo* from protein sequences due to the difficulty associated with obtaining robust statistical assessments (Gould et al. 2010). It is therefore of great interest to discover (using both computational and experimental techniques) new functional motifs that may form the basis of future drug discovery, by disrupting or regulating important interactions.

2 Short Linear Motifs (SLiMs) and Molecular Recognition Features (MoRFs)

In this chapter we focus on the characteristic features of SLiMs and on the various algorithms that have been developed to aid in their identification. Protein sequence motifs (SLiMs) have been described as functional microdomains that are short and flexible in length (between 2 to 11 consecutive residues). These are thought to arise by convergent evolution (Davey et al. 2009; Dinkel et al. 2014), thus the same SLiM may be found within otherwise unrelated proteins. They form compact functional modules and mainly occur within intrinsically disordered regions and surface accessible regions of proteins (Fuxreiter et al. 2007). Of the residues that constitute a SLiM, only a certain fraction are invariant (ie, fully conserved) across multiple instances of the motif. Usually these residues confer functional specificity, for binding interactions and/or undergo posttranslational modifications (PTMs). Other positions may tolerate conservative substitutions (eg, residues with similar size and/or physicochemical characteristics may be used interchangeably). Finally, some positions are not under selective constraints (wildcard positions). Thus, SLiMs have well-defined sequence patterns that are usually represented graphically using sequence logos (Schneider and Stephens 1990) or by machine-readable regular expressions (REs), that constitute position-specific definitions of allowed residue types and/or certain wildcard or ambiguous positions. Regular Expressions (REs) will be explained and elaborated upon later in the chapter.

*Mo*lecular *R*ecognition *F*eatures (MoRFs) are so-called because these protein segments form a specific class of intrinsically disordered regions (IDRs) that exhibit specific molecular recognition and binding functions. MoRFs are short (usually 20 residues or fewer) segments that are located within longer IDRs and are very interaction-prone (Vacic et al. 2007). MoRFs undergo characteristic disorder-to-order transitions upon binding to their partners (Mohan et al. 2006); based upon their bound state structures, they have been classified into α-MoRFs, β-MoRFs and ι-MoRFs (the latter class forms non-regular structures without regular backbone hydrogen bonding patterns). Unlike SLiMs, MoRFs are not defined on the basis of a sequence pattern (RE), but as interaction-prone disordered segments that (are predicted to) form ordered secondary structures upon binding to a protein partner. However, MoRF segments may themselves contain SLiMs, such as demonstrated in Fig. 9.1a (see next section).

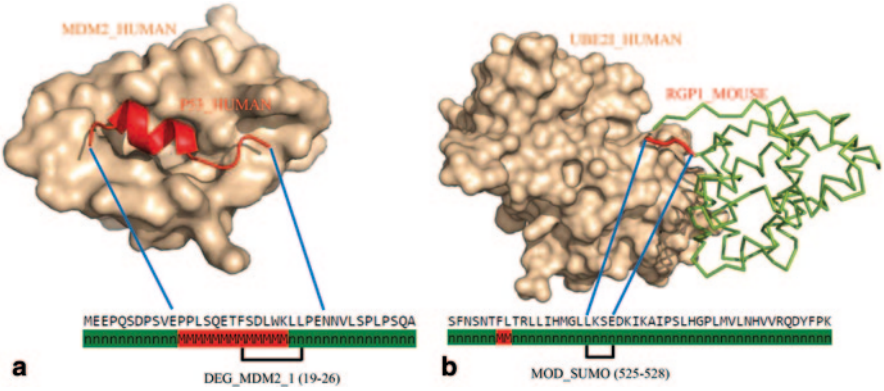

Fig. 9.1 Examples of SLiM-mediated interactions. **a** The p53 peptide (*red* cartoon) that is recognized by the folded SWIB domain (surface representation) of MDM2 (PDB code: 1YCR) is a MoRF that attains a helical bound state conformation. This MoRF region also contains a SLiM (degron) as indicated on the figure. **b** Interaction between the mammalian SUMO E2 enzyme (UBE2I, in surface representation) and its SUMOylation substrate RanGAP1 (*green* ribbon) mediated by a modification motif (shown in *red*) (PDB code: 1KPS). In both the figures, the amino acid sequence of the peptide motif segments and their sequence neighborhood are shown below their respective molecular diagrams along with the MoRFPred predictions (the letter 'M' on a *red* background indicates the segments that are predicted to be a MoRF, whereas 'n' against a *green* background indicates non-MoRF residues). The SLiM segments and their corresponding ELM identifiers are also indicated

3 Motif (SLiM)-Mediated Interactions and Their Biological Importance

Our current understanding of protein-protein interactions has changed significantly with the knowledge of how IDRs play crucial roles in enabling protein interactions ('domain-peptide' interactions) (Dinkel et al. 2014; Petsalaki and Russell 2008; Edwards et al. 2012). Interactions mediated by SLiMs have been shown to function in diverse processes, such as in the control of cell cycle progression, substrate selection for proteasomal degradation, targeting proteins to specific subcellular locations and for stabilizing scaffolding complexes. Figure 9.1a shows an example of a motif-mediated interaction (a p53 peptide bound to the folded SWIB domain of MDM2) (Schon et al. 2002). The region of p53 present in the crystal structure contains an 8-residue SLiM (the ELM degradation motif 'DEG_MDM2_1'). The motif is disordered in the unbound state, but forms an α-helical secondary structure in the complex with MDM2, thus conforming to the classical definition of a MoRF. In this example, the SLiM overlaps with a larger MoRF segment that can be detected by the MoRFPred predictor (Disfani et al. 2012). Figure 9.1b illustrates recognition of the ELM SUMOylation motif 'MOD_SUMO' present on the C-terminal domain of RanGAP1 by the mammalian SUMO E2 enzyme UBE2I (Bernier-Villamor et al. 2002). Note that in this case the peptide motif is not classified as a MoRF by the predictor.

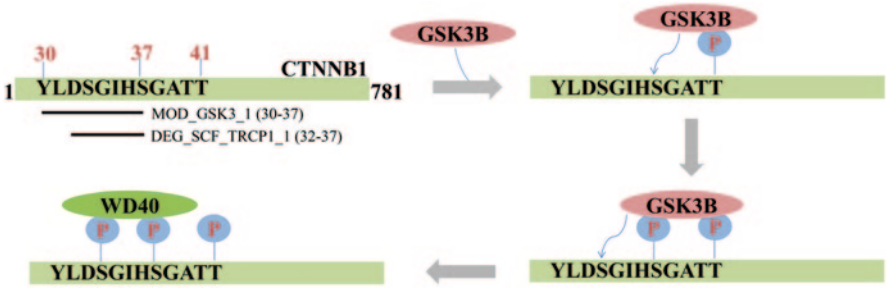

Fig. 9.2 Schematic illustration of the use of multiple overlapping SLiMs (ELM identifiers MOD_GSK3_1 and DEG_SCF_TRCP1) in beta-catenin (CTNNB1) that allows the recognition and relay (sequential) phosphorylation of beta-catenin by glycogen synthase kinase-3 beta (GSK3B) resulting in the activation of a degradation motif (degron) that is recognized by the WD40 repeat domain of the substrate adaptor subunit of a multi-subunit E3 ubiquitin ligase, resulting in the ubiquitination of beta-catenin and its 26 S proteasome-mediated degradation. Phospho groups are shown in *blue circles* and 'P' written in *red*

Interface areas in peptide-protein complexes observed in the PDB average about 500 Å^2 (London et al. 2012), significantly smaller than the size of an average protein-protein hetero-interface (1900 Å^2) or homodimer interface (3900 Å^2) (Janin et al. 2008). The limited size of SLiM-mediated interfaces often results in micromolar binding affinity for these interactions, whereas globular protein-protein complexes formed via domain-domain interactions can be much stronger (nano-molar or lower Kd). This permits transient and reversible interactions that are necessary for many dynamic cellular binding events, such as those required for the rapid transmission of intracellular signals (Neduva and Russell 2005; Gibson 2009).

A further advantage is the 'switching' behaviour that can be achieved by the use of PTMs within SLiMs to regulate interactions. Phosphorylation/dephosphorylation is widely used to enhance (or disrupt) interactions for example, and this enables direct cross-talk between multiple signaling pathways (Akiva et al. 2012). Multiple SLiMs can also form more complex switches by co-operating with each other and acting in synergy with post-translational modifications to assist switching between different functional states of proteins (Dinkel et al. 2014). In the example illustrated in Fig. 9.2, the phosphorylation of beta-catenin (CTNNB1) at Thr41 generates a docking site for Glycogen synthase kinase-3 beta (GSK3B) which phosphorylates Ser37 and generates a new docking site for GSK3B. Subsequent phosphorylation of Ser33 by GSK3B switches CTNNB1 binding specificity to the F-box/WD40 repeat containing protein BTRC which functions as a substrate recognition component of a SCF (SKP1-CUL1-F-box protein) multi-subunit E3 ubiquitin-protein ligase. This results in the recruitment of β-catenin to the SCF E3 ligase complex followed by ubiquitination and proteasome-dependent degradation of β-catenin (Wu et al. 2003; Hagen and Vidal-Puig 2002; Van Roey et al. 2013).

SLiMs represent an important target for diseases, both in terms of causal mutations and potential therapeutics (Uyar et al. 2014). Further, many pathogens have taken advantage of the plasticity of SLiMs by mimicking host motifs to dysregulate

and rewire cellular pathways of the host to their own advantage (Davey et al. 2011b; Kadaveru et al. 2008). Our growing appreciation of the importance of motif-mediated protein functions is evidenced by the recent growth of motif databases. The eukaryotic linear motif (ELM) resource maintains curated data on protein SLiMs whose functional validity has been demonstrated experimentally (Dinkel et al. 2014). MiniMotifMiner (MnM) (Mi et al. 2012) is another resource dedicated to the annotation and detection of a broad spectrum of motifs from a large number of species and currently contains 880 consensus minimotifs and 294,053 instances. Similar to SLiM, minimotif is another term used to define short contiguous peptide sequences that possess a demonstrated function (including post translation modifications, binding to a target protein or molecule and protein trafficking) in at least one protein. Another database ScanSite (Obenauer et al. 2003) stores data for 65 motifs in 12 different groups (functionally similar motifs have been grouped together). Similarly, Prosite (Sigrist et al. 2013) contains data for 1308 patterns or regular expressions although it contains domain signatures in addition to SLiMs. However, in spite of their immense functional importance in eukaryotic cell regulation, detailed information regarding the majority of SLiMs are still limited, and at present only a small proportion of human motifs have been discovered (Tompa et al. 2014). This highlights the pressing need to develop and further enhance computational methods that can efficiently predict novel SLiMs in protein sequences and thereby serve as a useful guide for experimental motif discovery efforts.

4 Representing Motifs: *R*egular *E*xpressions (REs), *P*osition *W*eighted *M*atrices (PWMs) and *P*osition-*S*pecific *S*coring *M*atrices (PSSMs)

SLiMs are commonly represented by RE-patterns and PWMs. SLiMs are comprised of both defined amino acid positions as well as wildcard positions which may be occupied by any amino acid type. Defined positions may be (i) *fixed* or *invariant*, in which only a single amino acid type is permitted at that position, or (ii) *ambiguous*, in which case multiple amino acids (often of similar size and/or physicochemical properties) may occupy that site and still result in a functional SLiM. Thus, a RE describes a sequence of letters that may match at each position in a given motif. The simplest RE is just a string of letters, such as the "RGD" motif present in extracellular matrix proteins that is recognised by different members of the integrin family (Corti and Curnis 2011). This regular expression matches only one defined amino acid sequence: Arg-Gly-Asp (RGD). To allow variable positions in a RE, additional symbols are used. For example, [KR] specifies that either K or R may be present; {min, max} specifies a range of minimum and maximum numbers of residues allowed (eg. M{0,1}) indicates that Met can either be absent (0) or can be present but only once (1); the '.' (dot symbol) at a given position indicates that any amino acid is allowed at that position. One disadvantage of REs is that residue-specific frequency information is lost: [KR] does not indicate the relative occurrence frequency

Table 9.1 Description of the different types of symbols used to construct Regular Expressions (REs) for peptide motif representation

Character	Name	Description
.	Dot	Any amino acid allowed
[…]	Allowed character class	Amino acids listed are allowed
[^..]	Disallowed character class	Amino acids listed are not allowed
X{min, max}	Allowed range (number) of consecutive specified character 'X'	Min required, max allowed
^	Caret	Matches the amino terminal
$	Dollar	Matches the carboxy terminal
?	Question	One amino acid is allowed but is optional
*	Star	Any number of amino acids are allowed but are optional
+	Plus	One amino acid is allowed, additional are optional
|	Alternation	Matches either expression it separates

of K vs R. Table 9.1 provides an overview of how regular expressions are used to represent sequence motifs.

Unlike REs, PWMs indicate the probability of each residue type occurring at each position in a motif. PWMs are widely used for characterizing and predicting sequence motifs (Bailey 2008). A PWM is an 'n' by 'w' matrix where 'n' is the number of letters in the sequence alphabet (20 amino acids for proteins) and 'w' is the number of motif positions. $P_{a,i}$ represents the probability of letter 'a' at the i^{th} position in the motif. A PWM can be used to define an occurrence probability for any possible sequence containing 'w' characters (calculated as the product of the corresponding entries in the PWM), based on the assumption that each motif position is statistically independent. The relationship between a RE and the corresponding PWM is shown in Fig. 9.3 for the KEN-box motif. The 16 validated occurrences (sites) from which this motif was constructed (data from ELM entry DEG_APCC_KENBOX_2) are shown aligned with each other on the left-hand panel. The corresponding RE is shown in the middle panel along with the observed counts of each letter in the corresponding alignment columns (frequency table). The PWM is shown on the right-hand panel. Finally, the figure represents the KEN-box sequence logo (Schneider and Stephens 1990).

Motif discovery algorithms also output a position-specific scoring matrix (PSSM) which takes the background probabilities of different letters into account (Bailey 2008). The PSSM entries are calculated as a log likelihood: $S_{a,j} = \log_2 (P_{a,j}/f_a)$, where f_a is the overall (background) probability of letter '*a*' in the set of input sequences that will be scanned for motif occurrences, and $P_{a,j}$ represents the frequency of letter 'a' at the jth position as explained earlier. Sequences are assigned scores by summing up (rather than multiplying position specific probabilities as with a PWM) the appropriate numbers from the PSSM table. PSSM scores are more useful for scanning sequences as compared to PWM probabilities because they allow scaling by background probability: this reduces false positive rates caused by non-uniform distribution of letters in sequences (Xia 2012; Bailey 2008).

Aligned sites					
Site1	N	K	E	N	E
Site2	D	K	E	N	G
Site3	D	K	E	N	E
Site4	D	K	E	N	Q
Site5	E	K	E	N	E
Site6	K	K	E	N	G
Site7	Q	K	E	N	S
Site8	S	K	E	N	V
Site9	D	K	E	N	M
Site10	N	K	E	N	Y
Site11	V	K	E	N	L
Site12	N	K	E	N	D
Site13	S	K	E	N	Q
Site14	N	K	E	N	A
Site15	N	K	E	N	Y
Site16	A	K	E	N	E

Regular Expression (RE)					
XKENX					
G	0	0	0	0	2
A	1	0	0	0	1
V	1	0	0	0	1
S	2	0	0	0	1
T	0	0	0	0	0
C	0	0	0	0	0
M	0	0	0	0	1
L	0	0	0	0	1
I	0	0	0	0	0
K	1	16	0	0	0
R	0	0	0	0	0
E	1	0	16	0	4
D	4	0	0	0	1
Q	1	0	0	0	2
N	5	0	0	16	0
F	0	0	0	0	0
Y	0	0	0	0	2
W	0	0	0	0	0
P	0	0	0	0	0
H	0	0	0	0	0

Position-specific Weight Matrix(PWM)					
G	0	0	0	0	0.125
A	0.062	0	0	0	0.062
V	0.062	0	0	0	0.062
S	0.125	0	0	0	0.062
T	0	0	0	0	0
C	0	0	0	0	0
M	0	0	0	0	0.062
L	0	0	0	0	0.062
I	0	0	0	0	0
K	0.062	1	0	0	0
R	0	0	0	0	0
E	0.062	0	1	0	0.25
D	0.25	0	0	0	0.062
Q	0.062	0	0	0	0.125
N	0.312	0	0	1	0
F	0	0	0	0	0
Y	0	0	0	0	0.125
W	0	0	0	0	0
P	0	0	0	0	0
H	0	0	0	0	0

Fig. 9.3 Converting a multiple sequence alignment of known motif instances into a RE and PWM. The alignment of motif sites (validated instances of the KEN-box (Dinkel et al. 2014)) is shown on the *left*. The RE is shown at the *top* of the *middle* panel. The counts of each amino acid type in each alignment column (the position specific count matrix, PSCM) are shown beneath the RE. The PWM is shown on the *right* hand side. The last figure shows the information content sequence logo for the motif (generated by http://weblogo.berkeley.edu/logo.cgi)

5 Overview of Functionally Specialized SLiM Categories in ELM

The latest published ELM release contained 197 classes and 2404 instances (Dinkel et al. 2014). SLiMs in ELM have been classified into six categories based on their function: proteolytic cleavage sites ('CLV'), sub-cellular targeting sites ('TRG'), ligand binding sites ('LIG'), post-translational modification sites ('MOD'), destruction motifs or degrons ('DEG') and finally, docking sites ('DOC') (Table 9.2). Figure 9.4 shows representative examples of SLiM-mediated interactions from each ELM class (except 'CLV' sites for which none of the entries had a corresponding PDB entry).

Table 9.2 Summary of data stored in the ELM database (as of September 2013) (reprinted with permission from Dinkel et al. 2014). Breakup of ELM data according to (1) the six ELM class types (LIG, MOD, TRG, DEG, DOC and CLV motifs) and the number of ELM classes corresponding to each class, (2) ELM instances by organism type, (3) the number of ELMs that are represented in the PDB, and finally, (4) the number of GO terms associated with the data in ELM

Functional sites	ELM classes		ELM instances		PDB structures	GO terms	
Total	197			2404	290	419	
By category	LIG	103	Human	1391		Biological process	217
	MOD	30	Mouse	211			
	TRG	23	Rat	115		Cell compartment	95
	DEG	15	Yeast	86			
	DOC	15	Fly	77		Molecular function	107
	CLV	11	Other	524			

Cleavage ‘CLV’ sites are recognised by proteases for the processing of predecessor proteins into their active biological products (eg, N-arginine dibasic convertase is an endopeptidase that recognizes (.RK)|(RR[^KR]) dibasic cleavage sites for processing secreted proteins (Hospital et al. 2000)). ‘TRG’ motifs are used for protein recognition and targeting to diverse sub-cellular compartments: for example, the ‘tyrosine-based sorting signal’ (Y..[LMVIF] motif) is found in the cytosolic tails of some membrane proteins and is responsible for deciding the traffic flow in endosomal and secretory pathways (Fig. 9.4a). Motifs that mediate binding to globular protein domains form the ‘LIG’ class: for example, the AP2 (Adaptor Protein) α subunit recognizes and binds to accessory endocytic proteins such as amphiphysin, AP180 and synaptojanin170 via their F.D.F motifs resulting in their recruitment to the site of clathrin coated vesicle formation and thereby assists and regulates vesicle assembly (Brett et al. 2002) (Fig. 9.4b). SLiMs located at post-translational modification sites constitute the ‘MOD’ class (eg, the Protein kinase B substrate phosphorylation site has residue preferences as shown in Fig. 9.4c).

Earlier ELM versions contained only these four motif categories (‘CLV’, ‘TRG’, ‘LIG’, and ’MOD’) (Gould et al. 2010). Recently however with the increase in the number of ELM classes, two additional but functionally specialized ‘LIG’ (ligand-binding) categories were introduced—‘DEG’ (degron) motifs and ‘DOC’ (docking) motifs. Degrons are motif sequences embedded within proteins that enable their specific recognition by E3 ubiquitin ligases, normally resulting in the channeling of these substrates into the ubiquitin-proteasomal degradation pathway (Glickman and Ciechanover 2002). For example, the [IL]A(P).{6,8}[FLIVM].[FLIVM] motif present in the α subunit of the heterodimeric transcription factor Hif-1 (hypoxia-inducible factor 1) is an oxygen-dependent degron that is hydroxylated by prolyl hydroxylases under conditions of normal oxygen availability (Masson and Ratcliffe 2003). Prolyl hydroxylation confers degron recognition and binding by the von Hippel-Lindau tumor suppressor protein (pVHL) (Fig. 9.4d) which forms a multi-subunit E3 ubiquitin ligase complex with elongin C, elongin B, Cul-2, and Rbx1 leading to the ubiquitination and proteasomal degradation of Hif-1α (Min et al. 2002).

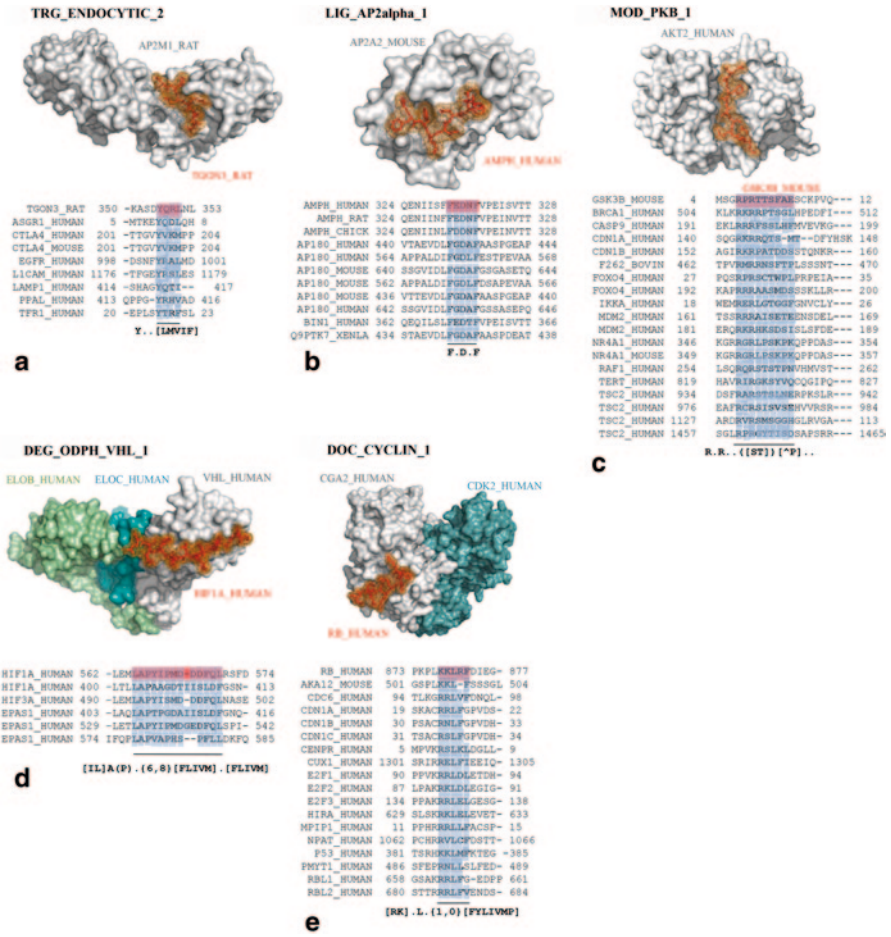

Fig. 9.4 PDB structures corresponding to representative examples from each ELM class showing the SLiM peptide (drawn using stick representation, *colored red* and surrounded by a surface *mesh*) in complex with their globular protein partners (displayed using *light grey* surface representation). SLiM-containing sequence segments of all the experimentally validated vertebrate instances (data from ELM) are shown in the multiple sequence alignments. The first sequence in each alignment corresponds to the SLiM-containing protein shown in the PDB structure (SLiM residues are shown in *red*). SLiM residues for the other instances are highlighted using *light blue color*. Consensus motif patterns are shown in *bold* under each alignment. **a** Targeting motif derived from the trans-Golgi network integral membrane protein (TGN38) interacting with the mu subunit of the adaptor protein complex 2, Ap2m1 (PDB code: 1BXX). **b** Ligand binding motif from human Amphiphysin interacting with the alpha-2 subunit of the adaptor protein complex 2, Ap2a2 (PDB code: 1KY7). **c** Modification motif from Glycogen synthase kinase-3 beta (Gsk3b) in complex with the kinase domain from RAC-beta serine/threonine-protein kinase (AKT2) (PDB code: 1O6K). **d** Degradation (*degron*) motif of human hypoxia-inducible factor 1-α protein (HIF1A) interacting with the Von Hippel-Lindau (VHL) component of the multi-subunit VHL ubiquitination complex (PDB code: 1LM8). **e** Docking motif derived from human Retinoblastoma-associated protein, RB1 interacting with the cyclin A2/CDK2 complex (PDB code: 1H25). Figures were drawn using PyMol

Finally, docking('DOC')motifs are used to recruit modifying enzymes onto their target substrates. However, 'DOC' sites are distinct from 'MOD' sites that are targeted for the actual enzymatic modification; initial binding to docking motifs on the substrate helps to direct and enhance enzyme specificity for the modification site (the two motifs together can be considered to possess a bi-partite architecture). For example, the docking motif DOC_CYCLIN_1 ([RK].L.{0,1}[FYLIVMP]) initiates substrate interactions with cyclin (Fig. 9.4e) resulting in increased specificity of phosphorylation (at the associated MOD_CDK_1 phosphorylation sites) by cyclin/Cdk complexes (Takeda et al. 2001).

6 Motif Discovery Algorithms and Tools

Given the diverse gamut of functions that are mediated by SLiMs, the development of methods and algorithms that will aid in (1) the discovery of new motifs (*de novo* motif prediction), and (2) filtering functional motif instances from the background of stochastic occurrences, is expected to be useful for identifying functional sites in proteins, especially within the unstructured segments. Usually motif discovery algorithms fall into three categories: enumeration, deterministic optimization and probabilistic optimization (D'Haeseleer 2006).

Enumeration is an exhaustive search based word counting method. The target sequences are broken up into shorter fragments (words of length 'n') and by counting the occurrence frequencies of all 'n-mers', the method attempts to identify statistically overrepresented short motifs. The highest occurrence frequency within the target sequences does not necessarily indicate a specific motif; statistical overrepresentation can be more reliably estimated by searching for motif patterns that appear more frequently than the random expectation (this random expectation is based on a background model that takes into account compositional biases). These steps need to be repeated several times until it finds statistically significant motifs. Further, by allowing mismatches and degeneracy in certain positions, consensus motifs can be defined in a more flexible and realistic manner. Alternatively, multiple overrepresented motifs that exhibit similarity may be combined into a single, more flexible motif. However, this method is computationally expensive because it requires the generation and storage of large numbers of short segments in memory.

Deterministic optimization is based on Expectation Maximization (EM). In the first step of EM, a PWM is initialized with a single n-mer segment of user-defined length ('n') along with some amount of background frequencies (nucleotides or amino acids). Next the input sequences are split into substrings (n-mers) and each substring then matched against the PWM. A probability value is calculated that indicates whether the substring was generated by the motif (PWM) model or by the background sequence distribution. Taking a weighted average of the current probabilities for each substring, the PWM is refined and the probabilities for the

substrings then recalculated based on the updated PWM. The steps are repeated iteratively until a maximum likelihood motif model (PWM) is obtained. A well-known implementation of EM is the Multiple EM for Motif Elicitation (MEME) software (Bailey et al. 2006).

Finally, probabilistic optimization is based on Gibbs sampling. Briefly one motif from each input sequence is randomly selected to determine an initial model and a PSSM is built from those sub-strings. Then the PSSM is used to scan each input sequence to find a motif that better contributes to improve the PSSM quality; this new motif with higher PSSM score is then added to the model and the old motif is removed. This process is repeated until the PSSM reaches convergence. The algorithm assumes that most of the target sequences will contain the motif. *Align*s Nucleic *A*cid *C*onserved *E*lements (AlignACE) (Chen et al. 2008) is a program based on the Gibbs sampling approach and is used to discover motifs from sets of DNA sequences.

Many *de novo* motif discovery tools are currently available that are dedicated to discover motifs present in disordered protein regions. De novo discovery methods take as input the protein primary sequence and utilize features such as disordered structural environment and evolutionary context as pointers to reduce false positive matches (Davey et al. 2012b). Functional SLiMs have been characterized to be enriched within disordered regions of the proteome, motif residues can be distinguished from their sequence neighborhood on the basis of higher evolutionary conservation, and furthermore, SLiMs often exhibit a propensity to form ordered secondary structures upon partner binding (Davey et al. 2012b). These additional layers of information are therefore used to enhance the filtering and removal of false positive hits.

Additional strategies to improve true positive motif detection include: removal prior to input of sequence segments that are spurious for motif discovery (eg, masking repeat sequences and low complexity regions), and sequence regions that are poorly represented in SLiMs (such as well structured domains, transmembrane segments and poorly conserved segments). Furthermore, the use of multiple motif predictors that cover a range of motif descriptions and search algorithms, followed by a comparison of results is always recommended. Optimizing the runtime details such as motif width, expected number of motif occurrences, deciding cutoffs for various parameters also require careful consideration. Sometimes it may be useful to combine similar motifs into a smaller set of (more) flexible motif descriptions. Users should also consider multiple high scoring motifs as the top hit may not necessarily be the most biologically relevant. Finally, the chances of detecting a true functional motif are also maximized if one can reduce (based on available evidence) the number of sequences that are not likely to possess that functionality (“noise”).

The Discovery@Bioware portal (http://bioware.ucd.ie/~compass/biowareweb/) and MEME Suite (http://meme.nbcr.net) contain a host of useful resources pertaining to the discovery, characterization and analysis of SLiMs (Table 9.3). The Eukaryotic Linear Motif (ELM) resource (http://elm.eu.org) has an extensive collection of curated SLiM instances, and is a useful tool for sequence annotation to identify protein segments that match known functional SLiMs. Regular expressions

Table 9.3 A list of commonly used motif discovery resources that enable motif prediction, discovery and analysis

Name	Description
SLiMProb	Searches for occurrences of pre-defined motifs (REs) in protein sequences (http://bioware.ucd.ie/~compass/biowareweb/)
SLiMSearch 3	Searches for occurrences of pre-defined motifs proteome wide (http://bioware.ucd.ie/~compass/biowareweb/)
SLiMPred	Predicts potential SLiMs in a protein sequence (http://bioware.ucd.ie/~compass/biowareweb/)
SLiMPrints	Predicts potential motifs by searching for clusters of locally conserved residues present in intrinsically disordered regions (http://bioware.ucd.ie/~compass/biowareweb/)
SLiMFinder	Identify SLiMs in a group of proteins (http://bioware.ucd.ie/~compass/biowareweb/)
GLAM2	Identify DNA or protein motifs using gapped local alignment (http://meme.nbcr.net)
MEME	Identify DNA or protein motifs using EM (http://meme.nbcr.net)
ELM	Database of experimentally validated SLiMs in eukaryotic proteins and a resource for investigating candidate functional SLiMs (http://elm.eu.org/)
MnM	Examines query protein for presence of short contiguous peptide sequences that have a known function in at least one protein (http://mnm.engr.uconn.edu/MNM/SMSSearchServlet)

representing the ELM classes are used by ELM's motif detection pipeline to scan proteins for putative SLiM instances (Davey et al. 2012a; Dinkel et al. 2012). Minimotif Miner (MnM, http://mnm.engr.uconn.edu/MNM/SMSSearchServlet) is also widely used for motif searches and analysis.

7 Details of Usage and Functionality of Some Selected Motif Discovery Tools

SLiMPrints (short linear motif fingerprints, currently at version 3.0) attempts to identify putative functional motifs from the input amino acid sequence on the basis of evolutionary conservation as a discriminatory feature for SLiM discovery (Davey et al. 2012a). Residue conservation statistics are analyzed and their significance estimated by comparison against the background conservation of neighboring residues. The method identifies relatively conserved (overconstrained) proximal residue clusters present within disordered regions; such "islands of conservation" located inside structurally unconstrained and mutation-prone disordered regions have been shown to be indicative of putatively functional SLiMs. The reader is referred to the original publication for a detailed description of the methodology (Davey et al. 2012a).

We demonstrate here how the user can provide input to the SLiMPrints web application (http://bioware.ucd.ie/~compass/biowareweb/Server_pages/slimprints.php),

provide a brief overview of the methodology involved and finally, describe how the output is displayed and its contents. The user can analyse a protein of interest by providing the UniProt Accession of the protein into the search box (Fig. 9.5, "Query protein"). SLiMPrints contains pre-computed multiple sequence alignments of least divergent orthologs selected using the GOPHER algorithm, following a BLAST search for homologs against a database of EnsEMBL metazoan (plus *Saccharomyces cerevisiae*) genomes (Flicek et al. 2011). The alignments have been processed to increase their quality by the removal of potential biases (for example, low complexity regions in highly divergent proteins were removed from the alignments and alignments with identified orthologs in <10 metazoan species were not considered further) (Davey et al. 2012a). Further, regions shown to be deficient in motifs (annotated domains, transmembrane segments, extracellular regions and highly structured residues) are masked before the motif discovery step. Because the algorithm aims to identify regions of functional constraint (proximal clusters of strongly conserved residues) against a backdrop of evolutionary drift especially within disordered segments, relative local conservation (RLC) statistics (that measures residue conservation against the background conservation of a neighboring sequence window) are employed to obtain better information about the putative functionality of a motif region. SLiMPrints combines RLC and disorder predictions to identify putative SLiMs in the input sequence. Figure 9.5 illustrates an example SLiMPrints output using human p53 as the input sequence. The output contains the identified motifs ranked by their significance score (Sig_{motif} is a metric that represents the likelihood/significance of the observed grouping of highly conserved residues that form a putative sequence motif (Davey et al. 2012a)). The underlying alignment(s) corresponding to the respective motif regions can be visualized by clicking on the "view" links. The RE of the obtained motifs and their sequence context (with the motif start and end residue positions in the input sequence) are also printed. The average IUPred (Dosztanyi et al. 2005) disorder score of the motif is also output. Finally, if the obtained motif matches an annotated ELM identifier, the ELM entry is also shown.

SLiMFinder (Short, Linear Motif Finder) software/web server (http://bioware.ucd.ie/~compass/biowareweb/Server_pages/slimfinder.php) is intended to allow researchers to *de novo* discover novel SLiMs from a set of input sequences (Davey et al. 2010). The purpose is to identify shared motifs among a set of unrelated proteins that possess a common function suspected to be SLiM-mediated (eg, binding to a common protein partner). SLiMFinder accounts for evolutionary relationships amongst the input sequences by clustering them into unrelated protein clusters (UPCs), such that proteins separated into different clusters do not share any BLAST-detectable similarity (Altschul et al. 1990). An explicit model of convergent evolution is used whereby the method searches for SLiMs that are statistically overrepresented in a maximum number of proteins from the different UPCs. SLiMFinder combines two algorithms: (i) SLiMBuild, which performs the actual task of identifying recurring motifs, and, (ii) SLiMChance estimates the statistical significance of returned motifs. We refer the reader to the original publication for full details of the methodology involved (Edwards et al. 2007).

Fig. 9.5 SLiMPrints input and output. Input options: SLiMPrints takes as input a UniProt accession number (shown on the *top* panel). Output options: summarized results of SLiMPrint hits are initially displayed as shown below the input options panel. This section provides a summary of the identified motifs along with their main features (highlighted using the *red* ovals and the *red arrows*). The results specify the motif rank, a "Visualize" option (link to visualize alignment of orthologs, an example of which is shown in the *bottom* panel), "Sig_{motif}" (Significance score of the identified motif), "Motif" (Regular Expression of the observed motif), "Context" (motif containing sub-sequence), "IUPred" (average disorder score of the motif) and "Annotated ELM" (if the motif is found in ELM)

Figure 9.6 shows the SLiMFinder web server input page. The input may be a list of UniProt IDs or user-built sequence files in UniProt or FASTA format. Next to the input box are the lists of options (separately for 'Masking', 'SLiMBuild', 'SLiMChance' and 'Output') that the user can employ to fine tune searches. First, there are multiple options to mask out regions (from the input sequences) known to be depleted in SLiMs: users can exclude from the motif search unconserved residues, ordered regions (based on IUPred predictions) such as Pfam domains, low complexity regions as well as certain amino acid types. Next, SLiMBuild has options that specify the minimum and maximum number of consecutive wildcard positions that are to be permitted, the total number of allowed wildcard positions and the minimum number of input sequences that must contain each generated motif for it to be returned as a putative SLiM. Users will also find settings to modify residue groupings based on physicochemical or other parameters: these groupings are used to define ambiguous SLiM positions. Once a set of motifs is generated by SLiMBuild, the SLiMChance algorithm assigns a statistical significance score (P-value) to each motif (the user can select the significance cutoff for returning motifs). Although the default behaviour is to return upto 100 motifs at P-value <=0.99, the most significant motifs are those with P<=0.05 (the stricter the significance cutoff, the smaller the proportion of false positive hits).

SLiMFinder output provides rich visualization and a host of options for data analysis (Fig. 9.6). In the main output page, a summary of the returned (predicted) motifs are shown ranked by significance score. With each motif hit there are associated hyperlinks: under the "Aligned" column, the 'M' and 'A' alignment links will allow the visualization of the motif region in the input sequences ('masked' and 'unmasked', respectively). Clicking the red links under the "Proteins" column shows those proteins in which the motif was found and their position in the sequence. The small thumbnail figure under "Plot" will direct the user to alignments for the corresponding protein and its GOPHER orthologs around the region of the generated motif. Finally, for each putatively returned motif there are links to run CompariMotif (Edwards et al. 2008) and SLiMSearch (Davey et al. 2011a): the former compares the motif to known, literature-derived motifs, whereas the latter searches for all UniProt entries that contain this motif alongwith statistical estimates about the validity of the observed occurrence.

GLAM2 (*G*apped *L*ocal *A*lignment of *M*otifs) is a software for finding motifs in input (protein or DNA) sequences (Frith et al. 2008). The web version is located at http://meme.nbcr.net/meme/cgi-bin/glam2.cgi. GLAM2 examines the set of input sequences for common motifs and finds a motif alignment with maximum score. GLAM2 enables the detection of gapped (*ie*, with indels) motifs. The algorithm starts from an initial random alignment constructed from the input sequences and uses simulated annealing to make repetitive changes to it. These changes are random and they affect the motif score (which can either increase or decrease), the idea being to prevent the system from being trapped in local optima. The changes are applied iteratively until the score fails to improve further even after 'n' successive changes (n=10,000 by default). The types of changes that are possible and their details are beyond the scope of this chapter and the reader is referred to the original

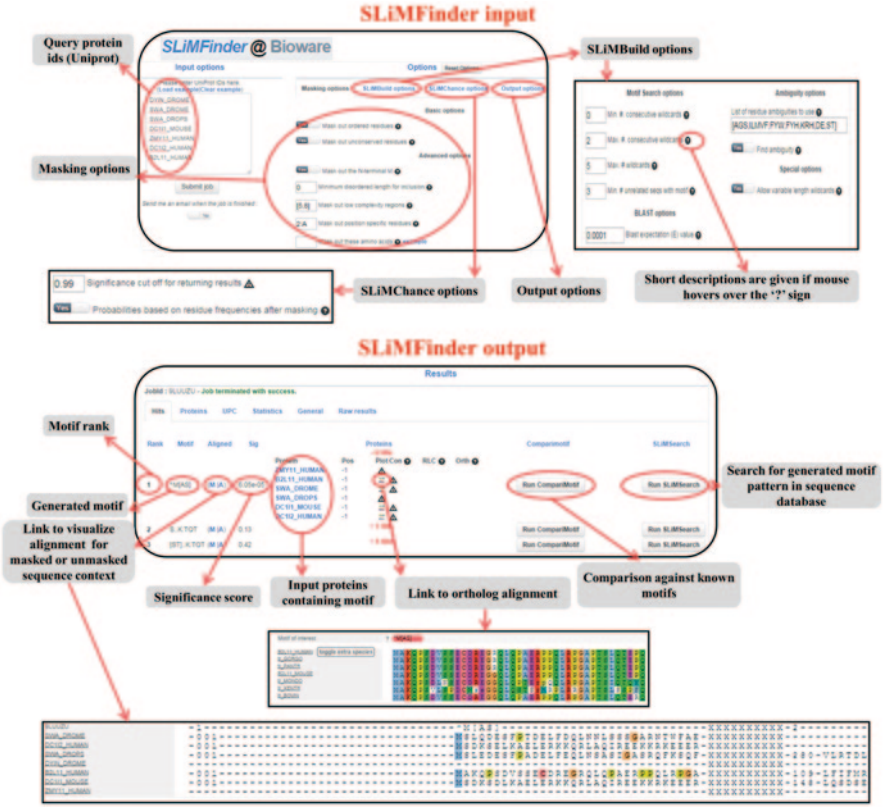

Fig. 9.6 SLiMFinder input and output. Input options are shown on the *top*. Input is a list of UniProt identifiers corresponding to the set of proteins in which we want to discover common (shared) motifs. Options are categorized into the following sub-sections: "Masking", "SLiMBuild", "SLiMChance" and "Output" options (shown using the red ovals and *arrows*). The web server provides short descriptions for each option if the user hovers the mouse over the '?' sign next to each option. Output: summarized results are initially displayed (shown in the panel below the input options). This section outputs the "Rank" (motif rank), "Motif" (RE of the generated motif), "Aligned (M|A)" (links to visualize motif alignments for masked or unmasked sequence context, an example is shown on the bottom most panel), "Sig" (motif significance score), "Proteins" (list of input proteins that contain the motif). Under the "Proteins" header, the user will see in red the number of proteins containing each predicted motif. By clicking on the number of hits, the output will expand to show the names of those proteins from the input list that contain the motif in question. Each protein can then be further analyzed for that motif based on the conservation statistics (for example by clicking on "Link to ortholog alignment"). Finally, "Run CompariMotif" (comparison against known motifs) and "Run SLiMSearch" (search for generated motif pattern in sequence databases) functions are also available for each predicted motif

publication (Frith et al. 2008). Essentially, GLAM2 builds on the idea that motifs contain a certain number of "key positions" defined by strict residue preferences at highly conserved and therefore presumed to be functional sites. The algorithm optimizes the number of key positions and then searches for an alignment of substrings

(one from each input sequence) to match a series of key positions. Thus in the scoring scheme, the alignments of identical or similar residues in the same key positions are rewarded, whereas insertions and deletions are penalised. Ultimately with the simulated annealing approach GLAM2 attempts to find a motif alignment with maximum score. To cross-check that a reproducible, high-scoring motif has been identified, the steps are repeated multiple (by default 10) times using different starting alignments selected randomly by the program. The algorithm then checks whether similar (but not necessarily identical) alignments recur. This is suggestive that the optimal motif has been found.

Figure 9.7 shows the input page on the GLAM2 server and an example output. Input can be either in the form of a text file containing the input sequences or by pasting the sequences into the box provided. The user can check details about the input formatting by clicking on the links (colored cyan) just above the input box. There are several parameters that can be customized (Fig. 9.7). The allowed alignments can be constrained by specifying variables such as: minimum number of input sequences to be used in building the motif alignment, minimum and maximum number of aligned columns (*ie*, key positions), and the initial number of aligned columns. The user can also modify the scores for tolerating insertions and deletions, and turn off/on shuffling of original sequence (used as a control to compare with the score of original sequence). Running GLAM2 is computationally heavy and the analysis time depends on sequence length and the size of the input dataset. One feature of this method is that it can detect only a single motif at a given time (by default 10 variants/replicates of the highest scoring motif are generated) and it does not model alternative binding motifs simultaneously (Tran and Huang 2014; Frith et al. 2008). However, more advanced users can use the command line installation to detect alternate (weaker) motifs, by first masking the strongest identified motif region (using the program 'glam2mask') and then re-running GLAM2.

The output is provided in three different formats: html, text and MEME text format. Figure 9.7 (*bottom*) shows a screenshot from the html output page. Because GLAM2 attempts to find the strongest motif in the set of input sequences using a 'replication strategy', if the top ranking motifs are very similar to each other, it is an indication that a successful replication has been achieved. Thus, by default GLAM2 outputs 10 variations of the strongest motif shared by the input sequences (this value "number of alignment replicates" can be changed by the user). Thus the topmost/first alignment is the interesting one: the purpose of the others is to indicate the reproducibility of the first motif. The output contains the list of motifs with maximum score and corresponding alignments of the motif containing segments (only the first one is shown in Fig. 9.7), their start and end positions, marginal score for each motif segment (this reflects the amount by which the total alignment score would decrease if that segment were to be removed from the alignment; thus, higher scores reflect better matches to the motif), and finally, the motif sequence logo. For each candidate motif, GLAM2 has additional options including, for example, scanning the motif against sequence databases (using GLAM2SCAN). The HTML output page also provides a link to view the *P*osition *S*pecific *P*robability *M*atrix (PSPM).

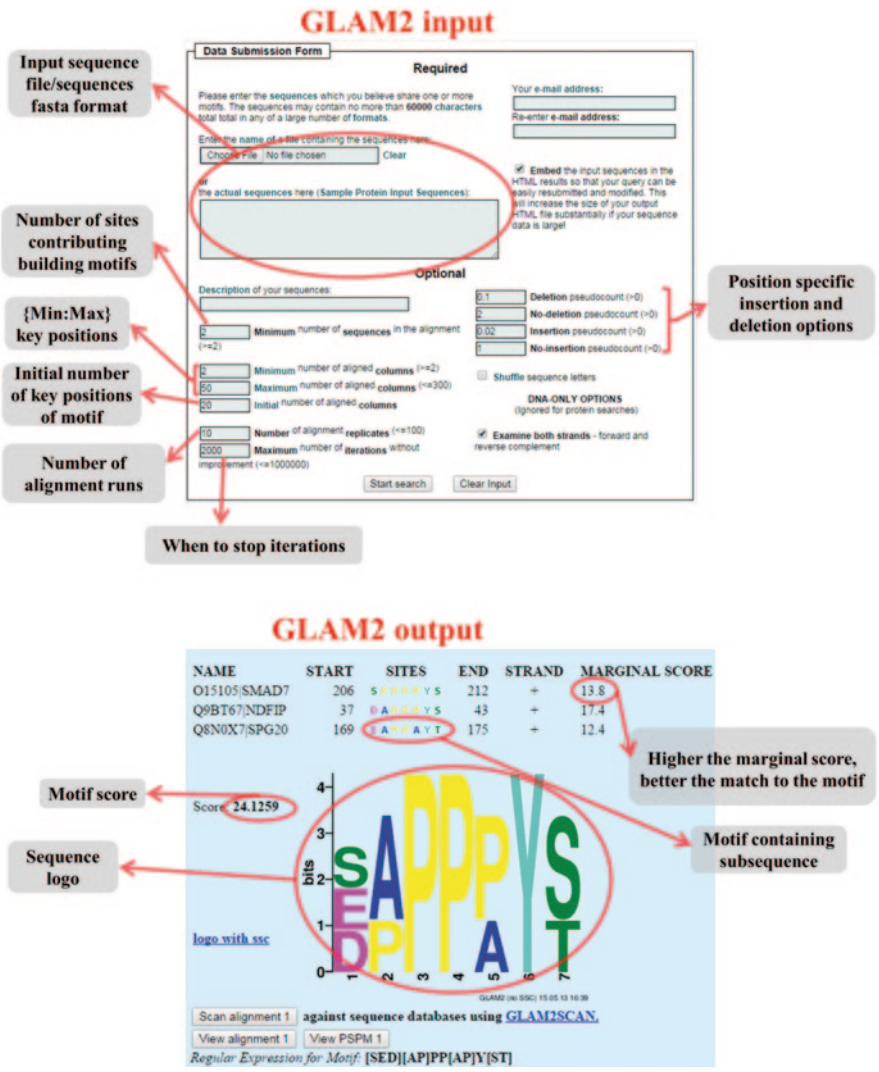

Fig. 9.7 GLAM2 input and output. Input (*top* panel) is accepted in fasta format. The available input options are shown using *red arrows*. These include options to specify the number of sites contributing to the motif (if known), number of key positions (maximum, minimum and initial number), maximum number of iterations and position specific insertion and deletion penalty scores. Output (*bottom* panel) showing the best statistically significant motif and a list of motif occurrences in the input dataset, their start and end positions, and marginal score followed by the motif logo. Hyperlinked buttons ("Scan alignment", "View alignment" and "View PSPM") that allow the motif to be analysed are shown at the bottom

MEME *M*ultiple *E*M for *M*otif *E*licitation (MEME) is a widely used tool for searching novel 'signals' in sets of biological sequences (Bailey et al. 2006); the webserver version is available on MEME suite (http://meme.nbcr.net/meme/cgi-bin/meme.cgi).

MEME has been used previously to discover common transcription factor binding sites in promoter sequences of similarly regulated genes (Lyons et al. 2000) and to identify novel sequence signatures in proteins with common interaction partners identified from large scale protein interaction data in *Saccharomyces cerevisiae* (Fang et al. 2005). MEME is based on the expectation maximization (EM) algorithm and it looks for ungapped, shared sequence patterns within the input (DNA or protein) sequences. One drawback is its inability to discover motifs containing indels as it does not allow gaps. To increase the chances of finding statistically significant motifs, it is recommended to keep the input sequences as short as possible (eg, by deleting repetitive regions and low complexity regions that do not generally contain functional motifs) and to curate the input sequence list to reduce as much as possible those sequences that are not likely to contain the motif. Although only a single motif can be modeled at a time, MEME erases previously discovered motifs and repeats the search, this enables new patterns to be extracted (Tran and Huang 2014; Hu et al. 2005; Bailey et al. 2006; Bailey et al. 2009).

For web server use, one has to provide a set of FASTA format sequences by either uploading a text file or by pasting the sequence information into the box as shown in Fig. 9.8. The other required input is an email address where the results will be sent. MEME searches for motifs ranging from 6 and 50 residues in length by default, although the user can specify other values between {2,300}. There is an option to specify the estimated number of motif sites per input sequence, particularly if there is any prior knowledge about the distribution of motif occurrences within the dataset. These options for setting the distribution of motif occurrences are called OOPS (*O*ne *O*ccurrence *P*er input *S*equence), ZOOPS (*Z*ero or *O*ne *O*ccurrence *P*er input *S*equence) and ANR (*A*ny *N*umber of *R*epetitions) modes. 'OOPS' assumes that each input sequence contains exactly one occurrence of each returned motif, whereas 'ZOOPS' assumes that each input sequence may contain at most one occurrence of each returned motif; the latter option is useful when certain of the input sequences may be missing some of the motifs. The ANR option can be used to explore multiple occurrences of a given motif within one or more sequences. MEME uses the ZOOPS option by default.

The output is generated in three different formats: HTML, TEXT and XML. Figure 9.8 shows part of the HTML output. MEME generates up to three top-ranking motifs by default, and each of the generated motifs may be present in either a subset of sequences or in all the input sequences (this refers to the number of occurrences). Every output motif is assigned an 'E-value'. The E-value refers to the probability of finding an equally well-conserved sequence pattern in random sequences; thus, the lower the E-value, the greater the statistical significance of the observed motif. The output overview shows the rank of the motif, its E-value and number of occurrences (sites) and the sequence logo for the motif. Below the "Motif Overview" section, further details about each of the identified motifs are available. This includes the multiple alignments showing the identified motif region in the

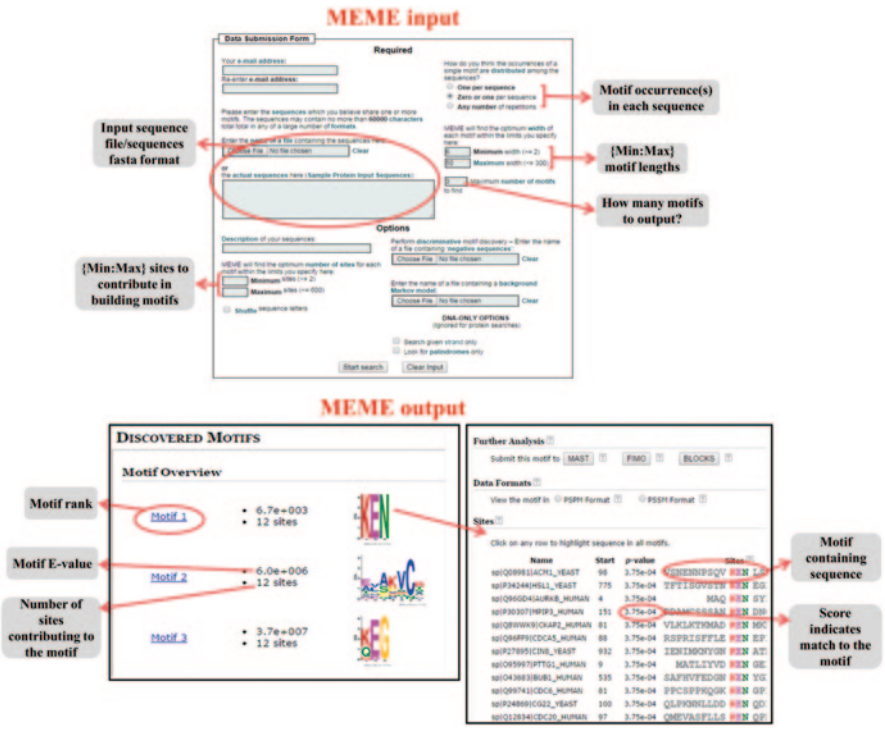

Fig. 9.8 MEME input and output. Input options are shown in the *top* panel. There are options to include the number of sites for each motif (if there is prior knowledge about the number of occurrences), and options to specify motif length. Output (*bottom left*) showing a list of protein motifs (by default 3 motifs) that MEME has discovered in the input sequences. Some of the hyperlinked buttons that allow the motif to be analysed further are shown at the *bottom right*

input sequences (Fig. 9.8, bottom right panel). Below the alignments are so-called "Block diagrams" showing the relative positions of the motifs within the input sequences (not displayed in the figure). Clickable buttons allow each motif to be analysed by other programs. Clicking on the 'MAST' (*Motif Alignment* and *Search Tool*) button will send the motif to the MAST web server where various sequence databases (or sets of user-uploaded sequences) can be searched for sequences that contain matches to that motif. Similarly, the button 'FIMO' (*Find Individual Motif Occurrences*) (Grant et al. 2011) will also trigger searches of sequence databases for hits to the motif patterns. Finally, these motifs may be compared against entries in the BLOCKS database of protein motifs (Henikoff et al. 1999) by clicking on the 'BLOCKS' button.

8 Prediction Performance on Disordered Motifs: Case Study on the KEN-box Motif

KEN-box mediated target selection is one of the mechanisms used in proteasomal destruction of mitotic cell cycle regulatory proteins via the Anaphase-promoting complex (APC/C complex) (Peters 2006; Michael et al. 2008; Pfleger and Kirschner 2000). 'KEN' motifs are significantly enriched in proteins with cell cycle keywords and further the KEN-box is significantly conserved throughout the eukaryotic taxon (Michael et al. 2008). Cdh1 and Cdc20 act as APC/C co-activators at distinct stages of the cell cycle. Cdc20 interacts with the APC complex during the M phase and is later replaced by Cdh1 (late M/G1 transition). Whereas both Cdh1 and Cdc20 can recognise target proteins via the Destruction Box (D-box) motif, the KEN-box is only recognised by Cdh1. Interestingly Cdc20 itself contains a KEN-box that is identified by Cdh1 and undergoes temporal degradation; Cdh1 then replaces Cdc20 as the adaptor of the APC complex. However Cdh1 contains two D-box motifs that ensure self-degradation of Cdh1 via APC/C in an auto-regulatory feedback mechanism; this is important for tuning the levels of active Cdh1 throughout G1 (Listovsky et al. 2004).

Motif discovery algorithms have to deal with the problem of spurious (stochastic) pattern matches that turn out to be non-functional (false positive) instances. In other words, merely observing a KEN pattern within a protein sequence does not necessarily indicate a functional degradation targeting motif. Many factors including protein cellular compartmentalization, tertiary structure and motif accessibility, etc regulate interaction of the KEN-containing protein with APC/C. All the functional KEN-box motifs discovered so far have been found within natively unfolded (disordered) regions of proteins; however, certain proteins (eg, HIPK4) carry a KEN-motif within a globular domain although their role in proteasomal degradation is unknown (Michael et al. 2008).

KEN-box instances were collected from the ELM database: 16 instances from 14 proteins were found classified as true positives (Dinkel et al. 2014). Table 9.4 shows their prediction performance using the 4 motif discovery algorithms discussed in the previous section. Whereas SLiMPrints analyzes every protein individually, the other methods (SLiMFinder, GLAM2 and MEME) take a set of sequences as input. Thus the complete set of 14 sequences carrying validated KEN motifs were supplied as input. With each method, we always tried the default settings first to evaluate how well these parameters performed. Any modifications that were necessary are mentioned at the appropriate places in the following description.

Of the 16 known instances, *SLiMPrints* returned 9 instances as significant hits ($P < 0.05$) that either completely or partially overlapped with the known KEN box and were recognized as being similar to the ELM entry LIG_APCC_KENbox_2. For two proteins ('CIN8_YEAST' and 'VE1_BPV1') it completely failed to predict the KEN-boxes. In case of the viral protein 'VE1_BPV1', this failure may have been due to the fact that SLiMPrints has been trained on the EnsEMBL (Flicek et al. 2014) metazoan and *Saccharomyces cerevisiae* genomes, and therefore it is unable to predict for viral proteins. For 'CIN8_YEAST' the program resulted in an error message.

Table 9.4 Prediction accuracy on the KEN-box (.KEN.) motif using four motif discovery algorithms ('Yes' indicates that the motif was successfully identified, 'No' that the method failed to identify the motif; '*' indicates that the KEN motif was returned by the algorithm as a significant hit; (Number) indicates the rank obtained for the predicted motif)

KEN-box containing proteins			Motif discovery methods used			
Protein name	Gene name	Start,End	SLiMPrints[a]	SLiM-Finder[b]	GLAM2[c]	MEME[c]
ACM1_YEAST	ACM1	97,101	Yes*(2)	Yes*(10)	Yes(1)	Yes(1)
AURKB_HUMAN	AURKB	3,7	Yes(5)	Yes*(10)	Yes(1)	Yes(1)
BUB1_HUMAN	BUB1	534,538	Yes*(3)	Yes*(10)	No	Yes(1)
BUB1_HUMAN	BUB1	624,628	Yes(16)	Yes*(10)	Yes(1)	No
BUB1B_HUMAN	BUB1B	25,29	Yes(25)	Yes*(10)	Yes(1)	Yes(1)
BUB1B_HUMAN	BUB1B	303,307	Yes*(9)	Yes*(10)	No	No
CDC20_HUMAN	CDC20	96,100	Yes(20)	Yes*(10)	Yes(1)	Yes(1)
CDC6_HUMAN	CDC6	80,84	Yes*(1)	Yes*(10)	Yes(1)	Yes(1)
CDCA5_HUMAN	CDCA5	87,91	Yes*(2)	Yes*(10)	Yes(1)	Yes(1)
CG22_YEAST	CLB2	99,103	Yes*(1)	Yes*(10)	Yes(1)	Yes(1)
CIN8_YEAST	CIN8	931,935	No	Yes*(10)	Yes(1)	Yes(1)
CKAP2_HUMAN	CKAP2	80,84	Yes*(1)	Yes*(10)	Yes(1)	Yes(1)
HSL1_YEAST	HSL1	774,778	Yes*(8)	Yes*(10)	Yes(1)	Yes(1)
MPIP3_HUMAN	CDC25C	150,154	Yes(19)	Yes*(10)	Yes(1)	Yes(1)
PTTG1_HUMAN	PTTG1	8,12	Yes*(3)	Yes*(10)	Yes(1)	Yes(1)
VE1_BPV1	E1	27,31	No	Yes*(10)	Yes(1)	Yes(1)

[a] SLiMPrints accepts a single protein sequence at a time and provides the score for the identified motif

[b] SLiMFinder can take multiple sequences simultaneously as input. SLiMFinder can either use the complete set of input sequences or automatically selects a subset thereof such that a high confidence motif can be generated. For this example SLiMFinder returned a list of 11 significant motifs, KEN motif was found in 10th position. Two similar motifs ('KEN..D' and 'KEN.{1,2}P') were ranked at 4th and 6th positions respectively

[c] Although GLAM2 and MEME can optimize how many sequences to use in order to obtain significant candidate motifs, in this case study both methods were controlled to use all 14 input sequences simultaneously. This was meaningful because in this particular example we knew beforehand that all the input sequences contained a true positive KEN motif

SLiMFinder performed significantly well on the dataset using default parameters. SLiMFinder outputs a list of candidate motifs identified from the set of input sequences ranked by their significance score. We found a KEN motif (with a significance score of 0.002) at rank 10 that contained all 16 KEN instances. Interestingly, two higher ranking motifs that closely resembled the KEN were also found: KEN.P ranked #4 (Sigscore = 6.96E-5) and KEN.{1,2}P ranked #6 (Sigscore = 9.53E-5). These two motifs contained 9 and 10 respectively of the total KEN instances present in the input dataset.

GLAM2 initially failed to detect the KEN-motif in the input set. The following parameters were used (all default settings, except for the number of motif containing sequences, which we knew beforehand to be 14): –z 14 (number of sequences), –a 2 (minimum width of motif), –b 50 (maximum width of motif), –w 20 (initial number of 'key positions'), and –n 2000 (number of iterations). On reflection, we felt that there was a mismatch between the length of the KEN motif and the value used for the "initial number of key positions" parameter; accordingly, we modified this to a low value consistent with the length of the motif being searched (*ie*, w = 2). This enabled GLAM2 to successfully identify 14 out of the 16 motif instances (Table 9.4). BUB1_HUMAN and BUB1B_HUMAN each contain 2 validated KEN-boxes, however only one from each protein was identified (since GLAM2 assumes that every input sequence may contain at most one occurrence of each motif). Further, we tested different values of 'w', and all values in the range [2, 15] were successfully able to recover 14 instances (one from each input sequence).

MEME also did an excellent job of discovering KEN-box motifs in the ELM benchmark dataset. It successfully identified 14 of the 16 instances using the following parameters: –minw 6 (minimum width of motif), –maxw 50 (maximum width of motif), –minsites 14 (minimum number of motifs), –maxsites 14 (maximum number of motifs), and –mod zoops (zero or one occurrences). The 'minsites' and 'maxsites' values were set to 14 since the number of motif occurrences in the dataset were already known (default values were used for all the other parameters). However, MEME failed to identify the second motifs of 'BUB1_HUMAN' (624, 628) and 'BUB1B_HUMAN' (303, 307) because the 'zoops' mode assumes that each input sequence may contain at most one occurrence of each motif. Although we knew that these two sequences contained 2 KEN-boxes each, there is no parameter setting on the input page where we could set the number of motif occurrences exactly to 2. We did however use the ANR (*A*ny *N*umber of *R*epetitions) option to try and detect the multiple motifs. However, this option resulted in a large number of false positive hits and even so the multiple KEN's in both BUB1 and BUB1B remained unidentified.

9 Limitations of Motif Discovery Algorithms

Although motif discovery algorithms have improved considerably over the past years, considerable challenges remain. For example, since a large majority of motif types have been characterized to be preferentially located in disordered protein

segments, one main challenge will be to design effective multiple sequence alignment tools that can efficiently align intrinsically disordered regions. However, it can also be argued that by focusing mostly on IDRs and by routinely masking out structured domains we might miss finding (some) novel SLiMs. On the other hand, another level of complexity is introduced if we include domain sequences in the alignments used for motif discovery. The strong similarities between domain sequences would hide the weak SLiM signals. Although it is difficult to estimate how frequently functional SLiMs may occur within domains (eg, on their surface regions), this might be an avenue to explore in the future. Another limitation of motif discovery algorithms is their unsuitability to take entire genomes as input to discover motifs. Especially with short length motifs, their statistical significance in the context of the entire proteome is difficult to establish. Therefore, motif discovery tools need to be improved further to be able to discover the full complement of short linear motifs in the proteome.

References

Akiva E, Friedlander G, Itzhaki Z et al (2012) A dynamic view of domain-motif interactions. PLoS Comput Biol 8(1):e1002341. doi:10.1371/journal.pcbi.1002341

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2

Bailey TL (2008) Discovering sequence motifs. Methods in Mol Biol 452:231–251. doi:10.1007/978-1-60327-159-212

Bailey TL, Williams N, Misleh C et al (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34(Web Server issue):W369–373. doi:10.1093/nar/gkl198

Bailey TL, Boden M, Buske FA et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37(Web Server issue):W202–208. doi:10.1093/nar/gkp335

Berman HM, Kleywegt GJ, Nakamura H et al (2013) The future of the protein data bank. Biopolymers 99(3):218–222. doi:10.1002/bip.22132

Bernier-Villamor V, Sampson DA, Matunis MJ et al (2002) Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. Cell 108(3):345–356

Bhattacharyya RP, Remenyi A, Yeh BJ et al (2006) Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev Biochem 75:655–680. doi:10.1146/annurev.biochem.75.103004.142710

Brett TJ, Traub LM, Fremont DH (2002) Accessory protein recruitment motifs in clathrin-mediated endocytosis. Structure 10(6):797–809

Chen X, Guo L, Fan Z et al (2008) W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. Bioinformatics 24(9):1121–1128. doi:10.1093/bioinformatics/btn088

Corti A, Curnis F (2011) Isoaspartate-dependent molecular switches for integrin-ligand recognition. J Cell Sci 124(Pt 4):515–522. doi:10.1242/jcs.077172

Davey NE, Shields DC, Edwards RJ (2009) Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics 25(4):443–450. doi:10.1093/bioinformatics/btn664

Davey NE, Haslam NJ, Shields DC et al (2010) SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. Nucleic Acids Res 38(Web Server issue):W534–539. doi:10.1093/nar/gkq440

Davey NE, Haslam NJ, Shields DC et al (2011a) SLiMSearch 2.0: biological context for short linear motifs in proteins. Nucleic Acids Res 39(Web Server issue):W56–60. doi:10.1093/nar/gkr402

Davey NE, Trave G, Gibson TJ (2011b) How viruses hijack cell regulation. Trends Biochem Sci 36(3):159–169. doi:10.1016/j.tibs.2010.10.002

Davey NE, Cowan JL, Shields DC et al (2012a) SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Res 40(21):10628–10641. doi:10.1093/nar/gks854

Davey NE, Van Roey K, Weatheritt RJ et al (2012b) Attributes of short linear motifs. Mol Biosyst 8(1):268–281. doi:10.1039/c1mb05231d

D'Haeseleer P (2006) How does DNA sequence motif discovery work? Nat Biotechnol 24(8):959–961. doi:10.1038/nbt0806-959

Dinkel H, Michael S, Weatheritt RJ et al (2012) ELM—the database of eukaryotic linear motifs. Nucleic Acids Res 40(Database issue):D242–D251. doi:10.1093/nar/gkr1064

Dinkel H, Van Roey K, Michael S et al (2014) The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res 42(Database issue):D259–D266. doi:10.1093/nar/gkt1047

Disfani FM, Hsu WL, Mizianty MJ et al (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28(12):i75–i83. doi:10.1093/bioinformatics/bts209

Dosztanyi Z, Csizmok V, Tompa P et al (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21(16):3433–3434. doi:10.1093/bioinformatics/bti541

Edwards RJ, Davey NE, Shields DC (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PloS ONE 2(10):e967. doi:10.1371/journal.pone.0000967

Edwards RJ, Davey NE, Shields DC (2008) CompariMotif: quick and easy comparisons of sequence motifs. Bioinformatics 24(10):1307–1309. doi:10.1093/bioinformatics/btn105

Edwards RJ, Davey NE, O'Brien K et al (2012) Interactome-wide prediction of short, disordered protein interaction motifs in humans. Mol Biosyst 8(1):282–295. doi:10.1039/c1mb05212h

Fang J, Haasl RJ, Dong Y et al (2005) Discover protein sequence signatures from protein-protein interaction data. BMC Bioinformatics 6:277. doi:10.1186/1471-2105-6-277

Finn RD, Bateman A, Clements J et al (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database issue):D222–D230. doi:10.1093/nar/gkt1223

Flicek P, Amode MR, Barrell D et al (2011) Ensembl 2011. Nucleic Acids Res 39(Database issue):D800–D806. doi:10.1093/nar/gkq1064

Flicek P, Amode MR, Barrell D et al (2014) Ensembl 2014. Nucleic Acids Res 42(Database issue):D749–D755. doi:10.1093/nar/gkt1196

Frith MC, Saunders NF, Kobe B et al (2008) Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol 4(4):e1000071. doi:10.1371/journal.pcbi.1000071

Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23(8):950–956. doi:10.1093/bioinformatics/btm035

Gibson TJ (2009) Cell regulation: determined to signal discrete cooperation. Trends Biochem Sci 34(10):471–482. doi:10.1016/j.tibs.2009.06.007

Glickman MH, Ciechanover A (2002) The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. Physiol Rev 82(2):373–428. doi:10.1152/physrev.00027.2001

Gould CM, Diella F, Via A et al (2010) ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res 38(Database issue):D167–D180. doi:10.1093/nar/gkp1016

Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27(7):1017–1018. doi:10.1093/bioinformatics/btr064

Habchi J, Tompa P, Longhi S et al (2014) Introducing protein intrinsic disorder. Chem Rev 114(13):6561–6588. doi:10.1021/cr400514h

Hagen T, Vidal-Puig A (2002) Characterisation of the phosphorylation of beta-catenin at the GSK-3 priming site Ser45. Biochem Biophys Res Commun 294(2):324–328. doi:10.1016/S0006-291×(02)00485-0

Henikoff JG, Henikoff S, Pietrokovski S (1999) New features of the Blocks Database servers. Nucleic Acids Res 27(1):226–228

Hospital V, Chesneau V, Balogh A et al (2000) N-arginine dibasic convertase (nardilysin) isoforms are soluble dibasic-specific metalloendopeptidases that localize in the cytoplasm and at the cell surface. Biochem J 349(Pt 2):587–597

Hu J, Li B, Kihara D (2005) Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res 33(15):4899–4913. doi:10.1093/nar/gki791

Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41(2):133–180. doi:10.1017/S0033583508004708

Kadaveru K, Vyas J, Schiller MR (2008) Viral infection and human disease–insights from minimotifs. Front Biosci: A J Virt Lib 13:6455–6471

Listovsky T, Oren YS, Yudkovsky Y et al (2004) Mammalian Cdh1/Fzr mediates its own degradation. EMBO J 23(7):1619–1626. doi:10.1038/sj.emboj.7600149

London N, Raveh B, Schueler-Furman O (2012) Modeling peptide-protein interactions. Methods Mol Biol 857:375–398. doi:10.1007/978-1-61779-588-617

Lyons TJ, Gasch AP, Gaither LA et al (2000) Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. Proc Natl Acad Sci U S A 97(14):7957–7962

Masson N, Ratcliffe PJ (2003) HIF prolyl and asparaginyl hydroxylases in the biological response to intracellular O(2) levels. J Cell Sci 116(Pt 15):3041–3049. doi:10.1242/jcs.00655

Mi T, Merlin JC, Deverasetty S et al (2012) Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. Nucleic Acids Res 40(Database issue):D252–D260. doi:10.1093/nar/gkr1189

Michael S, Trave G, Ramu C et al (2008) Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. Bioinformatics 24(4):453–457. doi:10.1093/bioinformatics/btm624

Min JH, Yang H, Ivan M et al (2002) Structure of an HIF-1alpha -pVHL complex: hydroxyproline recognition in signaling. Science 296(5574):1886–1889. doi:10.1126/science.1073440

Mohan A, Oldfield CJ, Radivojac P et al (2006) Analysis of molecular recognition features (MoRFs). J Mol Biol 362(5):1043–1059. doi:10.1016/j.jmb.2006.07.087

Moult J, Fidelis K, Kryshtafovych A et al (2014) Critical assessment of methods of protein structure prediction (CASP)–round x. Proteins 82(Suppl 2):1–6. doi:10.1002/prot.24452

Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. FEBS lett 579 (15):3342–3345. doi:10.1016/j.febslet.2005.04.005

Obenauer JC, Cantley LC, Yaffe MB (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31(13):3635–3641

Peters JM (2006) The anaphase promoting complex/cyclosome: a machine designed to destroy. Nat Rev Mol Cell Biol 7 (9):644–656. doi:10.1038/nrm1988

Petsalaki E, Russell RB (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 19(4):344–350. doi:10.1016/j.copbio.2008.06.004

Pfleger CM, Kirschner MW (2000) The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. Genes Dev 14 (6):655–665

Van Roey K, Dinkel H, Weatheritt RJ et al (2013) The switches. ELM resource: a compendium of conditional regulatory interaction interfaces. Sci Signal 6(269):rs7. doi:10.1126/scisignal.2003345

Van Roey K, Uyar B, Weatheritt RJ et al (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev 114(13):6733–6778. doi:10.1021/cr400585q

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18(20):6097–6100

Schon O, Friedler A, Bycroft M et al (2002) Molecular mechanism of the interaction between MDM2 and p53. J Mol Biol 323(3):491–501

Sigrist CJ, de Castro E, Cerutti L et al (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41(Database issue):D344–D347. doi:10.1093/nar/gks1067

Takeda DY, Wohlschlegel JA, Dutta A (2001) A bipartite substrate recognition motif for cyclin-dependent kinases. J Biol Chem 276(3):1993–1997. doi:10.1074/jbc.M005719200

Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 37(12):509–516. doi:10.1016/j.tibs.2012.08.004

Tompa P, Davey NE, Gibson TJ et al (2014) A million peptide motifs for the molecular biologist. Mol Cell 55(2):161–169. doi:10.1016/j.molcel.2014.05.032

Tran NT, Huang CH (2014) A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. Biology Direct 9:4. doi:10.1186/1745-6150-9-4

Uyar B, Weatheritt RJ, Dinkel H et al (2014) Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? Mol Biosyst 10(10):2626–2642. doi:10.1039/c4mb00290c

Vacic V, Oldfield CJ, Mohan A et al (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res 6(6):2351–2366. doi:10.1021/pr0701411

Ward JJ, Sodhi JS, McGuffin LJ et al (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337(3):635–645. doi:10.1016/j.jmb.2004.02.002

Wu G, Xu G, Schulman BA et al (2003) Structure of a beta-TrCP1-Skp1-beta-catenin complex: destruction motif binding and lysine specificity of the SCF(beta-TrCP1) ubiquitin ligase. Mol Cell 11(6):1445–1456

Xia X (2012) Position weight matrix, Gibbs sampler, and the associated significance tests in motif characterization and prediction. Scientifica 2012:917540. doi:10.6064/2012/917540