Front Biosci. 2010 Jun 1;15:801-25. PMID: 20515727

Computational identification and analysis of protein short linear motifs

Norman E. Davey [1,2,3,4] , Richard J. Edwards[5], Denis C. Shields[1,2,3]

[1]UCD Complex and Adaptive Systems Laboratory, University College Dublin, Dublin, Ireland, [2]UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland, [3]UCD School of Medicine and Medical Sciences, University College Dublin, Dublin, Ireland, [4] EMBL Structural and Computational Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany, [5] School of Biological Sciences, University of Southampton, Southampton, United Kingdom

TABLE OF CONTENTS

1. ABSTRACT

Short linear motifs (SLiMs) in proteins can act as targets for proteolytic cleavage, sites of post-translational modification, determinants of sub-cellular localization, and mediators of protein-protein interactions. Computational discovery of SLiMs involves assembling a group of proteins postulated to share a potential motif, masking out residues less likely to contain such a motif, down-weighting shared motifs arising through common evolutionary descent, and calculation of statistical probabilities allowing for the multiple testing of all possible motifs. Much of the challenge for motif discovery lies in the assembly and masking of datasets of proteins likely to share motifs, since the motifs are typically short (between 3 and 10 amino acids in length), so that potential signals can be easily swamped by the noise of stochastically recurring motifs. Focusing on disordered regions of proteins, where SLiMs are predominantly found, and masking out non-conserved residues can reduce the level of noise but more work is required to improve the quality of high-throughput experimental datasets (e.g. of physical protein interactions) as input for computational discovery.

2. INTRODUCTION

Short, linear motifs (SLiMs) are abundant and ubiquitous protein microdomains that play a central role in cell regulation (1). A defining feature of SLiMs is their length; which is generally between 2 and 10 residues (2), often only a subset of these residues mediate binding (70% of known instances have 4 defined positions or less). Many of the defined positions are degenerate, meaning that a functional residue does not need a specific amino acid at that position for functionality, rather one from a particular set of amino acids (this set is usually a grouping of physicochemically similar amino acids). By definition, SLiMs are also linear, in that residues are adjacent in the primary sequence of the protein as opposed to being in close proximity in the tertiary structure of the protein.

SLiMs, also referred to as linear motifs or minimotifs, typically act as protein ligands and mediate a plethora of biological processes including cell signaling, post-translational modification (PTM) and trafficking target proteins to specific subcellular localizations (2). SLiMs can control gene expression; recruitment of the transcriptional co-repressor Groucho/transducin-like enhancer-of-split (TLE) family, for example, is mediated by the WRPW C-terminal motif (3). (4)They are particularly important for intracellular signaling; examples include the tumor necrosis factor receptor (TNFR) superfamily, which signals by recruiting TNFR-associated factors (TRAFs) through a SLiM in their cytoplasmic tails (5), or the canonical regulatory signaling interactions of the SH3 binding motif PxxP (6). SLiMs can also have important extracellular activity, such as the binding to integrins via the charged RGD motif (2). They can act as molecular switches, causing activation or deactivation of proteins through ubiquitination, phosphorylation or the addition of some other PTMs (4). Modified SLiMs, can direct tasks as diverse as binding proteins to the bilayer lipid

membrane, through the addition of Palmitate group to S-palmitoylation sites (7), or assisting proper protein folding and tethering of adjacent cells, by acting as attachment sites for saccharide chains by glycosylation (8). Other important SLiM-mediated PTMs include cleavage sites, such as those for cleavage by Furin (9) and Taspase (10); many neuropeptides and peptide hormones are created through proteolytic cleavage of their protein precursors at SLiM cleavage sites (11).

SLiMs also play important roles in disease, either by mutation of native motifs or through nefarious use by an external pathogen or predator. Viruses often mimic human SLiMs to hijack a host's cellular machinery, thereby adding functionality to their compact genomes without necessitating new virally encoded proteins (12): Src binding motif PxxP in HIV *Nef* protein modulates replication (13); WW domain binding PPxY mediates budding in Ebola virus (14); FMDV targets cells via RGD-mediated integrin interactions (15), and a Dynein Light Chain binding motif in Rabies virus is vital for host infection (16). In the bacterial world, toxins from both Pseudomonas (17) and Cholera (18) are imported using KDEL-like signals. Similarly, several proteins involved in erythrocyte targeting by the malaria pathogen Plasmodium falciparum contain an import motif RxLxE/Q (19). Many Metazoan predators also use SLiMs to their advantage: the snake venom platelet aggregation activation inhibitors arastatin and albolabrin, contain the integrin binding RGD motif (20, 21). Finally, mutation to functional SLiM residues are implicated as the cause of many diseases including *Noonan syndrome* (22) and *Liddle's Syndrome* (see (12) for review).

Increased knowledge of SLiMs has increased interest in the therapeutic use of SLiMs as potential lead compounds. Encouraging studies have established the ability of small peptides to competitively bind proteins and the ability to target drugs to SLiM interactions (23, 24). In cancer therapeutics, the angiogenesis inhibitor Cilengitide (25), (an inhibitor of integrin-RGD motif interaction), and P53 reactivating Nutlin-3a (26, 27) (disruptor of MDM2-mediated ubiquitination and destruction of P53 returning the ability of cancer cells to apoptose) have provided promising results (28). SLiMs can also be used to target oncolytic viruses, while leaving normal cells unharmed: Davydova *et al.* specifically targeted integrins frequently over-expressed in Oesophageal Adenocarcinoma by genetically modifying an integrin-binding RGD motif of adenoviral coat proteins to alter its specificity (29).

2.1. Biological attributes of SLiMs

Discovery of novel classes of shared SLiMs among proteins with a common function (e.g. sharing an interaction partner) is difficult, since the signal is very weak and occurs against a background of many potential false positives. For this reason, searches are more likely to be successful if the search space is reduced as much as possible. One approach to this is to concentrate on motifs that share features with previously discovered motifs. Knowledge of typical motif attributes gained from known motifs has been used to create rules to classify and discover novel motifs.

2.1.1. Structural disorder

SLiMs tend to occur in "intrinsically unfolded" or "natively disordered" segments of proteins (30); it is estimated ~85% of known functional motifs occur in these regions (31). This bias can be clearly seen as a shift in predicted disorder scores between SLiM and non-SLiM residues (Figure 1). Disordered regions/proteins lack a well-defined three dimensional structure and show distinct amino acid biases, tending to be enriched in P, E, K, S, G and Q, whilst being depleted in W, Y, F, C, I, L and V (32). Computational prediction of disorder does not rely alone on composition, however, since the interactions among residues (e.g. enrichment for residues of shared charge) contribute to disorder (33). Disordered regions of the proteome were originally thought to act solely as linkers between the "real" functional units of proteins (34). However, the discovery of functional modules, such as SLiMs, in these regions has increased awareness of their importance.

Brown *et al* (35) studied 28 protein families with ordered and disordered regions, finding that disordered regions evolve significantly more rapidly than ordered regions. This rapid evolution means residues are less constrained and therefore more likely to convergently evolve a short linear motif (SLiM), which if advantageous to the protein may be retained by purifying selection. The lack of structure also plays a major role in the enrichment of SLiMs in disordered regions by allowing the disorder-to-order transition often necessary for ligand binding (1). Extensive tracts of disorder surrounding interacting SLiMs have been hypothesized to protect the proteins against unwanted aggregation (36), which is consistent with the observation that SLiM residues are often themselves less disorder-promoting than the flanking regions (30).

2.1.2. Sequence conservation

The short length (typically between three and ten amino acids in length) and degeneracy (positions are often flexible in terms of possible amino acids) of SLiMs impart an evolutionary plasticity which is unavailable to globular protein domains, meaning that *de novo* motifs can evolve convergently, appearing by point mutation to add new functionality to proteins (4). It has been hypothesized that such evolutionary transience contributes evolutionary flexibility to SLiM mediated pathways, allowing for species to quickly rewire pathways by removing or adding interactions through point mutation to a few key residues (4). Their discovery by conservation-based methods is difficult as their level of conservation is not as high as domains: many functional motifs are often not conserved beyond

vertebrates (37). Despite this, SLiMs are more conserved than surrounding non-functional residues (Figure 2), due to purifying selection (4).

2.1.3. Specificity

It has been observed that the defined residues of a SLiM alone are generally insufficient to interact with high specificity and that additional residues surrounding the core motif play an important role by either increasing the affinity of the interaction or hindering interaction with non-specific targets (39). For example, Pbs2, using the canonical PxxP motif, binds only the SH3 domain of Sho1 out of the 27 different SH3 domains known in yeast (40), suggesting highly constrained secondary information encoded in the context of the peptide not apparent in the core functional interacting Prolines. Further, the same analysis showed that the specificity of yeast Pbs2 peptides for SH3 domains in other species was not as high, suggesting that evolutionary pressures had tuned the specificity for Sho1 in the yeast proteome.

2.1.4. Affinity

One of the key differences between SLiM-domain and domain-domain interactions is the affinity of binding. Domains, when they bind to each other, tend to do so with relatively strong affinities: low-nanomolar or even picomolar affinities are known. The short length of SLiMs means that they rarely have such strong affinities, usually ranging between 1 and 150 μM (1). For example, the affinity of the Cyclin-binding motif has been measured as 0.19 μM (41) and the 14-3-3 binding motif at 0.15 μM (42). This low affinity is ideal for transient interactions in signal transduction or for quickly responding to a stimulus. Co-operative binding can increase affinity and many examples exist of several short linear motifs in a disordered region binding to a target protein co-operatively, with affinities rivaling that of a single domain (43).

2.1.5. Structure

Although SLiMs are enriched in disordered regions they should not be considered unstructured, as binding often involves a mechanism known as *induced fit* where an unstructured/disordered region is induced to form a structure when binding a globular region (1). Information from SLiMs in their bound state is leading to a better understanding of the structural constraints placed on these disordered regions (39). An analysis of the known bound SLiM suggested that a third of known motifs form helical structures when bound, a third of interactions form by beta augmentation (adding an extra strand to a beta sheet in the interacting partner) (44) and the rest form various structures specific for a particular SLiM (1, 45).

2.1.6. Amino acid preference

Certain residues are preferentially used by SLiMs (Figure 3). Alanine and Glycine are largely avoided, and there is an unusual (and unexplained) preference for Arginine over the chemically similar Lysine. The high likelihood of hydrophobic residues Isoleucine, Methionine and Valine being in an ambiguous position mirrors their chemical similarity and interchangeability in many known SLiMs. Proline, Tryptophan, Threonine and Cysteine are all highly enriched, reflecting their importance to many binding events (2). Weighted models could take advantage of this information to improve motif discovery tools.

2.2. Potential for novel SLiM discovery

It has been estimated that only a small proportion of functional motifs have been discovered to date and several observations allude to the immense potential for novel discovery. SLiMs are enriched in regions of disorder (with approximately 85% of instances adhering to this rule (2)) and with 17% of proteins predicted to be totally disordered and 20-50% to contain at least some disorder (31) there is large potential for convergent evolution of functional sites. Secondly, the inequality between the number of known domains types (~10,000) and SLiMs types (~200 (2, 47)) advocates research in the area, especially as a recent study suggested only 3-19% of known interactions can be explained in terms of domain-domain interactions (48). Although this number will undoubtedly grow as further complex structures are analyzed experimentally, it still leaves a large scope for SLiM mediated interactions. Finally, it has been estimated that 15-40% of protein-protein interactions are mediated by SLiMs (4), yet only 5% of interactions contained in HPRD are SLiM-mediated and with as few as 1% for human yeast-two hybrid interaction analyses (49).

2.3. Sources of SLiM information

Several projects have been attempting to collate information about known SLiMs through extensive literature-based curation, experimental discovery and high-throughput computational analyses. There are many sources of SLiM data available. ELM (2) and MnM (47) have curated classical SLiMs with an emphasis on ligand binding motifs. Phospho.ELM (50) and Phosphosite (10) have focused on curating phosphorylation sites, dbPTM (51) is a general repository for functional group modification, and MEROPS (52) and CutDb (53) are general repositories of cleavage sites. These sources have allowed the deduction of many of the attributes of SLiMs, such as the propensity of SLiMs to occur in disordered regions, and are being used increasingly by bioinformaticians creating SLiM discovery tools. Much of the filtering and masking techniques discussed in this paper were created, trained and tested based on observations from these data repositories.

2.3.1. Classical motifs

ELM (2) and MnM (47) are manually curated databases of all manner of SLiMs, including ligands, targeting motifs and PTMs (including cleavage), and are currently the most complete sources of non-modification SLiMs. ELM contains 132 motif types, each with a regular expression definition, and approximately 900 instances with high quality annotation including gene ontology data for cellular component and biological process, data on the source of the curation and structural data when possible. Whereas ELM is explicitly restricted to eukaryotes, MnM 2.0 (47) is a repository of 858 motif "consensus sequences" and 4,229 instances across taxa, including cellular localization information. It is not curated to the level of ELM but does offer links to the source of the data. PROSITE (54) still contains many SLiMs, however the focus of their curation has moved to large domain descriptors in recent years diminishing its use as a source for SLiMs.

2.3.2. Modification motifs

The most studied PTM is phosphorylation and this is reflected in the amount of phosphorylation data and resources available. Phospho.ELM (50) annotates 4,110 metazoan (predominantly human and mouse) substrate proteins with 2,103 tyrosine, 12,435 serine and 2,503 Threonine phosphorylation sites. Phosphosite (55) contains information for 49,016 phosphorylation sites in 8,327 proteins. Phosphosite also curates several other modification sites including acetylation, di-methylation, sumoylation and ubiquitination. Other PTMs databases are available, including OGlycBase (56) (242 glycoproteins, 2,413 verified O-glyc sites and 49 verified C-glyc sites (release 6.0)) and Ubiprot (57) (417 proteins modified ubiquitin attachment and 165 ubiquitinated sites). MEROPS and CutDb are two sources of cleavage sites for proteases. CutDb (53) contains 6293 cleavages sites for 549 proteases acting on 2246 substrates and MEROPS (52) is a highly annotated database of 100,807 peptidases (including orthologue information) grouped into 2627 families for which more than 7000 cleavages sites have been defined. UniProt (46) also has a large amount of modification data including 106,570 experimentally validated and predicted modification sites in 37,828 proteins (release 56). The dbPTM (51) database is a general repository for both experimental (~36,000) and predicted (2,860,047) PTM sites, collecting data from several other modification databases together along with various other sources of protein information such as solvent accessibility, orthologous protein clusters and secondary structure.

3. SLiM DISCOVERY

3.1. *A priori* motif discovery

Several web-based methods to discover novel instances of known SLiMs are available such as ELM (2), MnM (58) and Quasimotifinder (59). Proteins can be searched using these methods to return putatively functional sites. The majority of known motifs are highly likely to occur in a protein by chance and a protein of average length will have several positions where amino acids match the regular expressions of known functional sites. To increase confidence in returned putatively functional sites these methods use various context- and attribute-based measures.

3.1.1. Primary sequence

The ELM server uses the ELM database to search for regular expression matches to known functional motifs. Returned motifs are filtered to exclude motifs occurring in globular regions of proteins using information from Pfam (60) and SMART (61) (though accessible region are also included if a structure is available for filtered globular regions). Results can also be filtered based on the species and localization of the protein. Curated SLiM instances, and motifs matching known functional motifs in the corresponding position of homologous proteins, are also identified.

The Minimotif Miner (MnM) (47) searches an input protein for matches to the MnM dataset 2.0 (an extended version of the publicly accessible MnM 1.0 dataset), scoring motifs based on their surface accessibility and fold enrichment (based on the ratio of observed motifs to expected motifs). Motifs are also scored by their conservation in homologues taken from the Homologene clusters of which the input protein has membership.

Quasimotifinder (59) searches for conserved motifs that match signatures curated in the PROSITE database (54). The method uses physicochemical information to search for fuzzy matches. Motifs are scored using a Pythagorean-based function to consider both the physicochemical information and the conservation level of the motifs. The one major drawback of the method is the source of motifs searched, although PROSITE contains a high quality annotation, it is missing many of the motifs available to the other methods.

SLiMSearch (http://www.southampton.ac.uk/~re1u06/software/slimsearch/) is another regular-expression motif search tool, suitable for local high-throughput analyses. The method takes as input a dataset of proteins and a set of motifs, which could be from known databases or defined by the user. SLiMSearch uses the same input masking as SLiMFinder (62) (including UniProt features, IUPred (33) based disorder prediction, low complexity regions, user-selected residues/motifs and relative local conservation-based masking (38)). Motif probabilities are calculated to assess motifs for statistical over-representation (or under-representation), adjusting for evolutionary relationships between the sequences, using the SLiMChance statistical framework employed by SLiMFinder (62).

3.1.2. Structural information

An interesting direction for novel instance discovery is the incorporation of structural information (63, 64). These tools use information from bound SLiMs that look for variations of known peptides capable of binding to the peptide binding region specifically, avoiding peptides which have residues incompatible with binding. These techniques need to be trained on at least one bound structure and several peptides known to bind to the domain of interest; currently several SLiM/Domain pairs have sufficient information for such analyses. These techniques are powerful tools to discover novel instance of SLiMs as well as novel protein interactions.

D-MIST (63) is a method that uses information from domain bound SLiM complexes and interaction datasets to predict protein interactions and SLiMs by learned binding profiles. The method calculates motifs/profiles with high specificity by searching for interactors of a known domain for motifs similar to known binding SLiMs from structural studies and peptide-based approaches. Interactors of the domain containing the protein are then searched for motifs resembling the known domain binding SLiM and matches are used to create a profile which can be used to search for proteins with a similar binding interface.

iSPOT (64) uses known structures of bound SLiMs to a domain to create a matrix of probabilities that a residue in the domain forms a contact with a residue in the SLiM. This matrix can then be used to predict whether or not the SLiM has specificity for a particular domain and motifs can be scored for their predicted likelihood of binding to the domain. Although currently the method is more suited to classifying motif specificity, the application could be placed in a framework similar to D-MIST to discover novel motifs.

3.1.3. Keyword searches

SIRW (65) is a web-based system to retrieve proteins with a particular keyword or Gene Ontology (GO) term. The system allows the input of a motif that can be searched against those proteins. Significance of association of the motif with the keyword can then be assessed using Fisher's exact test. Such analyses have proved successful in the past with new instances of EH1 transcriptional repressor motifs discovered through enrichment in transcriptional keywords (66) and new instances of KEN box APCC-binding Destruction motifs identified from cell cycle keywords (67).

3.2. Post-translational modification prediction

High throughput mass spectrometry analyses in recent years have enriched data for many PTMs (68). For example, many analyses have created kinase-specific phosphorylation data (69), defining a particular region of the phosphorylated protein modified. With limited residues possible for modification and the degenerate nature of these sites, specific modification site discovery tools provide a more successful method for their discovery than generic SLiM discovery tools (69). These numerous experimental data can be used to create profiles to discover novel instances, increasing specificity by using contextual information. This class of SLiM discovery will not be considered explicitly in this discussion, however many of the techniques described in this paper will have applications to such analyses.

Many tools are available to predict functional group addition PTMs. Scansite (70) creates experimentally derived position-specific scoring matrices (PSSM) using oriented peptide library and phage display experiments for multiple kinases and several binding events of high interest such as PDZ, SH2, and 14-3-3 binding. Proteins can be searched for sites matching these PSSMs as well as user defined motifs and profiles. AutoMotif (71) predicts several classes of PTM sites in proteins using support vector machines (SVM). SVMs for each class of PTM are trained separately using positives, annotated in the Swiss-Prot database, as well as negatives sites. Many other predictors for various functional group addition PTMs are available, examples include C-mannosylation (72), N-terminal myristoylation (73) and sulfation (74). Several cleavage site predictors such as PeptideCutter (75), SignalP (76), ProP (77) are also available. PeptideCutter predicts sites for multiple proteases and chemicals; SignalP discovers cleavage sites for signal peptide; ProP also discovers signal peptides but focuses on Arginine and Lysine, in particular Furin cleavage sites. Recent methods for high-throughput discovery of cleavage site specificity (78) will no doubt further enhance the future ability of prediction methods to discover novel cleavage motifs.

3.3. *De novo* motif discovery

The concept of over-representation as an indicator of functionality is currently the most powerful and widely used approach for discovering *de novo* SLiMs computationally (62, 79). Any set of proteins where there is a strong hypothesis for a SLiM mediated functionality, such as targeting protein localization, mediating protein binding or acting as a recognition site for a post-translational modification, can be analyzed for SLiMs. Under this hypothesis, the function-mediating SLiM would occur more often than expected by chance. Typically, such over-representation has arisen because of selection for the motif in the proteins. The hypothesis that functional motifs will be over represented due to purifying selection is simple yet powerful. For example, a motif matching a functional site will evolve convergently by point mutation adding functionality through binding, localization or modification. If this functionality

is advantageous then the motif will be maintained under an evolutionary constraint. Secondly, if the motif is damaging (e.g. such as a localization signal for the wrong cellular compartment) there will be a selection pressure to remove the motif or thirdly further mutation and genetic drift will slowly wipe out the instance matching the functional motif, if it has no functional effect.

Neduva *et al.* (37) clearly demonstrated the potential of models based on convergent evolution when they applied Dilimot to discover SLiMs in multiple HPRD datasets. They were able to verify two of their predictions with direct-binding assays, a protein phosphatase 1 binding motif (DxxDxxxD) and a motif that binds Translin (VxxxRxYS) (37). As previously discussed, keyword enrichment has aided in the discovery of novel KEN box (67), KEPE (80) and EH1 motifs (66). Pattern matching discovery was used to discover functional 14-3-3 motifs. Loss of function information for EFF-1 based on truncation of the C-terminal mutants led to the discovery of 23 potential motifs matching known functional motifs from which two 14-3-3 motifs were experimentally validated as vital for function (58).

Several approaches for novel motif discovery are available. Algorithmic motif discovery uses solely the over-representation hypothesis to discovery putatively functional motifs. More successful approaches build biological models on top of algorithmic motif discovery using techniques such as masking and attribute-based inference to discover biologically relevant motifs.

3.3.1. Algorithmic motif discovery

Several approaches are available to discover raw motifs, which can be broadly classed as alignment-based or alignment-free. In both cases, results will tend to be dominated by longer regions of conservation or homology (*e.g.* globular domains) at the cost of SLiM detection and so corresponding care must be taken where this might be a problem.

TEIRESIAS (81) is an alignment-free algorithm that efficiently returns motifs occurring in greater than a user-defined number of proteins by avoiding the enumeration of the entire pattern space. The method can return rigid ambiguous motifs using a predefined set of possible ambiguities. SLiMBuild (62) (one of the algorithms employed by SLiMFinder) identifies convergently evolved, short motifs in a dataset, reducing search times by explicitly screening out motifs that do not occur in enough *unrelated* proteins; this screening overcomes the problem of shared protein domains swamping the signal of SLiMs. The method allows flexibility (wildcard spacers of variable lengths) and ambiguity (in a similar fashion to TEIRESIAS).
GLAM2 (82) is a generalization of the alignment-based Gibbs Sampling method of MEME (83) with the additional ability to discover flexible length motif by allowing insertions and deletions. D-motif (84) uses a correlated motif approach to find pairs of interfaces (without flexible length wildcards or ambiguity) that could mediate interactions within a PPI network; it is not yet clear if this method has any practical application, however, since known examples of such correlated motifs typically include homologous domains which are best analyzed by other methods. PRATT (85) allows both flexibility and ambiguity but is more suited to domain descriptor discovery. Several other methods are available such as MEME (83) and ASSET (86) (a review of methods can be found here (87)).

3.3.2. Biological models

Dilimot (37, 79) was the first method to explicitly attempt *de novo* computational discovery of SLiMs in datasets of proteins. The enrichment of motifs in disordered regions is utilized by removing globular regions and coiled coil regions, using information from SMART (61) and Pfam (60) and using the globular region prediction tool Globplot (88). Regions of strong homology are removed, leaving only one representative homologous region for motif discovery and thereby enriching for motifs that have evolved convergently. Raw motifs are then returned by TEIRESIAS (81) and scored using a binomial scoring scheme introduced by ASSET (86). Conservation of the motif for several closely related species is calculated and incorporated into the scoring scheme, under the assumption that true motifs are usually conserved across closely related species. The tool was benchmarked on ELM datasets and on protein interaction datasets from the Human Protein Reference Database (HPRD) (89) returning many previously known functional motifs as well as several potential novel motifs, of which two were experimentally validated (37).

SLiMDisc (90, 91) is also built on the basic pattern discovery abilities of the TEIRESIAS algorithm (81). Motifs are scored using an information content-based scoring scheme which use evolutionary weighted support (those SLiMs present in evolutionarily distant sequences are up-weighted and those primarily arising due to common evolutionary descent are down-weighted). A number of filtering options are provided, (disorder, globular regions etc) and user-defined masking is also possible, allowing experimental/topological information to be incorporated. This gives the user a great deal of control over the type of motif returned. SLiMDisc can be considered an empirical motif discovery tool as the scoring scheme is not based on a statistical scoring scheme. Rather, it is based on the observation that the scoring scheme performs well on benchmarking datasets and because of this can be seen as a complementary to the probabilistic methods of Dilimot (79) and SLiMFinder (62).

SLiMFinder (62) is a probabilistic SLiM discovery program building on the principles of the SLiMDisc algorithm (91). The TEIRESIAS raw motif discovery tool is (81) replaced by SLiMBuild (62) allowing flexible and ambiguous motifs to be returned. Proteins can be masked to exclude under-conserved residues (38), non-disordered regions predicted using IUPred (33), low complexity regions, specific amino acids or motifs, and annotated features including domains or user-annotated regions to allow any contextual information to be included in the analyses. Statistics are implemented in the SLiMChance algorithm (62), which is based on the binomial statistics introduced by ASSET (86) (also used by Dilimot (79)) with two major extensions: (1) homologous proteins are weighted (as in SLiMDisc) to account for the dependencies introduced into the probabilistic framework by homologous proteins; (2) introduction of significance scores, *i.e.* the probability that any motif considered would reach a binomial *p*-value by chance is calculated and used to rank motifs.

3.3.3. Structural models

Recently, methods have used structural data to attempt novel SLiM discovery from protein primary sequence. Alpha-MoRF pred (92) and ANCHOR (93) use the observation that many motifs have an inherent propensity to form a secondary structure. Alpha-MoRFpred discovers deviation towards order in the primary sequence using PONDR VL-XT methods and filters these region based to remove false positives using neural networks. ANCHOR is based on the scoring scheme introduced by IUPred (33) and predicts regions that are likely to undergo disorder to order transition on binding. Both methods have their drawbacks and neither is specifically trained to discover motifs of the length typical of SLiMs concentrating more on larger disordered interaction regions. However, it is an area that will undoubtedly prove fruitful in the discovery of novel disorder regions involved in protein-protein interactions.

4. DATASET DESIGN FOR SLiM DISCOVERY

Recent advances in methodology have caused dataset design to be the major limiting factor in motif discovery and we believe that the next major advance in computational discovery of functional SLiMs will come from improvements in this area. Because of the challenges raised by the short and degenerate nature of SLiMs, maximizing the signal to noise ratio is crucial. Strong hypothesis-driven dataset design will always be the most important driver of success in a motif discovery analysis.

4.1. Data sources

The majority of SLiM discovery analyses use PPI data and these form the focus of this section. Two other very interesting sources of data for SLiM discovery are localization data and Gene Ontology (GO) data, and many of the same issues and solutions are relevant to these analyses as well.

4.1.1. Gene ontology

The GO (94) is a maze for the uninitiated user (to understand how GO is annotated, see (95)) but it is also a good source of protein groupings, many of which are candidates for potential SLiM-mediated interactions. The data is continuously updated with information from multiple sources reported in the literature, clustering proteins into logical groups that are formalized descriptions of a shared underlying biology. Many GO terms, especially larger high level ontologies, will have such a diverse focus that they are unlikely to be enriched for any one motif. To maximize chances of success, datasets with a strong hypothesis that a SLiM is responsible for the grouping of proteins should be analyzed. For example, there is a strong likelihood that a shared motif involved in binding PDZ domains could be discovered using the GO term 'PDZ domain binding' (GO:0003684) which contains 19 proteins, however, the level above 'protein domain specific binding' (GO:0019904), which has 582 gene products, is likely to have drawn together too much noise for any one signal to be discovered.

4.1.2. Localization

Eukaryotic targeting to sub-cellular locations involves multiple pathways, many of them mediated by SLiMs (2). For example, the peroxisomal targeting motif that tags proteins for import into the lumen of the peroxisome (96) or the KDEL endoplasmic reticulum (ER) retrieval signals that return ER proteins secreted while trafficking exported proteins to the ER (97). As fluorescence-based methods add to our knowledge of protein localization many more SLiMs are expected to be discovered (98). Localisation data from datasets such as Locate (99) and GO (which also includes cellular component annotation) (94) or high-throughput analyses (100) can be used to search for over-represented targeting motifs.

4.1.3. Protein-protein interaction data

Ideally, all protein interaction interfaces would be solved with 3D structures of the interaction, however in reality only a small number of known proteins have been solved in complex (101) and even the best available interactomes are incomplete (102). Many proteins' functions rely on interaction with other proteins; to gain a true understanding of that functionality it is necessary to understand the method of binding in particular the residues mediating that binding. High-throughput analyses of protein-protein interactions have amassed large quantities of interaction data of varying quality. Lack of overlap (103), irreproducibility and high error rates have lowered

confidence in any one interaction returned from a single source (104), but in general the overall quality of the data as a source of information to infer novel SLiMs is unrivalled (105).

High-throughput experiments (105, 106) and literature curation in sources such as HPRD (89), STRING (107), Bind (108), DIP (109), IntAct (110), MIPS (111), MINT (112) and Reactome (113) (see for review (114)) have amassed large amounts of protein interaction data. PPI networks can be split into sub-networks on the hub-spoke model, where a central protein (hub) interacts with several interactors (spokes). The hypothesis for motif discovery is that the hub protein contains a module (for example a domain) that interacts through a SLiM in a subset of the interactors. Detecting such a motif, however, relies on the signal from the true SLiM-mediated interactors being strong enough to overcome the noise of proteins that interact via another mechanism and/or have been falsely included in the dataset of direct interactors with the hub (*e.g.* they may interact indirectly in complex or via shared intermediates). These issues are explored in the following sections.

4.2. Working with PPI data

Current interaction data is typically organized on a protein level as binary graphs, where an edge is indicative of an interaction between the two proteins signified by nodes. These interactions can be, but do not have to be, a physical interaction where the two proteins share an interface or co-occurrence in a more complicated multi-protein complex (not to be confused with transient complexes). This level of abstraction allows proteins to be described as simple networks allowing easy manipulation of data and rapid integration of the various data sources; however it ignores much of the information available that is valuable to the motif discovery process. Instead, PPI data can be conceptualized on 4 defined levels of information (Binary, Protein complex, Atomic and Topological) to aid motif discovery (Figure 4), the level of an interaction should, if possible, be considered during dataset construction and the interpretation of results.

4.2.1. Binary interaction

The binary level (Figure 4A) describes proteins that are known to have a physical interaction. Information from some small scale binding experiments and Y2H data consists of binary physical interactions that share an interaction interface. This data is suitable for inferring SLiMs mediating protein binding without adding noise from complex partners that do not share a physical interaction.

4.2.2. Protein complex interaction

This level (Figure 4B) describes the interactions of true functional units, complex-complex and complex-protein interactions. A complex interaction is not indicative of any physical interaction between any two of the proteins in the complex and therefore contains large amounts of noise for use in SLiM discovery. Tandem affinity purification and co-immunoprecipitation provides *in vivo* information on a complex level, with some dependency on experimental conditions affecting stringency of protein dissociation during protein separation. While a complete binary map is likely to be superior to complex information for SLiM discovery, a very sparse binary map may well be improved by the addition of the noisier complex information, as it will also include some binary interactions.

4.2.3. Atomic interaction

The lowest, and most interesting, level (Figure 4C) is the domain/SLiM/atomic level. The basic aim of *in silic*o motif discovery is to aid experimental methods in deciphering the atomic interaction level for given proteins from binary and complex level data. Information from this level is usually of higher quality, literature-curated information from sources such as truncation and mutagenesis studies, or structural data of bound proteins by NMR and X-ray. Typically motif-domain interactions are fairly well defined prior to proceeding to structural characterization, so the scope for novel motif discovery from structural data is relatively limited.

4.2.4. Topology specific interaction

Topology specific information (Figure 4D) is an extension of the atomic level to consider the separation of interactions through space. This information can be particularly useful for eliminating “biological false positives”, where the proposed interactions can never actually occur in nature due to the physical separation of the proposed interactors (i.e. due to occurrence separate cellular compartments). As with atomic level data, however, the increased quality comes at the cost of a reduction in quantity and whether such data routinely delivers enough signal for motif discovery is yet to be determined.

4.3. Issues with PPI data

4.3.1. Comparability of sources

Yeast 2-hybrid (Y2H) and protein Tandem Affinity Purification (TAP) are the largest sources of interaction data and both have well known biases. For Y2H, often proteins do not interact in the yeast nucleus and the proportion of interactions not detectable in Y2H has been shown to be high (115). Many modifications are not available in the assay host, this is an issue for regulated proteins were post-translational modifications are necessary for interaction or when glycosylation is necessary for folding (8). Many interactions are highly regulated requiring the presence or the

absence of a modification acting as a regulatory switch for interaction (115). For example, many SLiMs function only when phosphorylated (e.g. 14-3-3 (116) and Grb2-like Src Homology 2 (SH2) domain binding motifs (117)), an experimental bias against such motifs would obviously affect the likelihood of a SLiM being returned from an analysis.

Experimental information from Y2H analyses and affinity purification mass spectrometry (AP-MS) are the two largest contributors to PPI databases (111), however these two data sources offer a very different aspect on protein interactions. Sources of PPI data play an important role in the quality of an interaction network for motif discovery. Each source has different advantages and disadvantages that should be considered. In practice, with high-throughput analyses this is not always possible, however for small-scale analyses every effort should be made to include all ancillary information available.

4.3.2. High affinity bias

Binary data implies that a pair of proteins interact or they do not but, in reality, there are actually a variety of flavors of protein interactions and binding affinities. Many of these cause problems for high-throughput analyses. Many SLiM interactions are transient, making them extremely difficult to capture due to their short half-life (often less than a second) and low affinities (1). Experimental PPI discovery may preferentially discover domain-domain interactions that are usually in the picomolar range of affinity compared to SLiMs interactions that are usually between 1 and 150 micromolar (1). Yeast-2-Hybrid (Y2H) data has been observed to be impoverished for SLiM interactions when compared to manually curated low-throughput interaction data (although other factors mentioned above may influence this) (49) and Tandem Affinity Purification (TAP)-tagging often misses low affinity interactions due to the experiment procedures.

4.3.3. Ascertainment bias

As with all biological data, PPI data often suffer from ascertainment bias and is not of equal quality across the genome (118); data is often more complete and of higher quality for more easily studied proteins (e.g. proteins capable of interacting in the yeast nucleus) and for some proteins of high interest like disease causing genes (e.g. the "guardian of the genome" oncogene P53 (119, 120) which even has its own dedicated website (p53.free.fr)). Often high interest genes have large amounts of small-scale experimentation that means that not only do these genes have more data, but that data is often low-throughput and has higher quality annotation.

4.3.4. Incomplete data

Many motifs are involved in membrane-associated interactions. However, data of this type is under-represented in PPI datasets due to their biochemical properties (121) and the predominance of complex post-translational modification. Co-operative binding is well known: for example the nuclear receptors require co-activation, one interactor stabilizing the binding site for another (122). Most of the assays, however, are not set up for such complicated 3 partner binding events. Conversely, there are detectable interactions *in vitro* that never occur *in vivo* (123), so-called biological false positives. Often proteins when brought together will interact, however the observation that binding occurs *in vitro* does not signify that the interaction will occur *in vivo* as, for example, they may never co-localize (In an extreme case, they may not even be from the same organism).

4.4. Reducing noise in datasets

A high ratio of noise (proteins without an instance of the motif) to signal (proteins containing the motif) in a dataset has a negative effect on the ability of discovery methods to rank true positive motifs highly. There are two main techniques for reducing dataset noise; network pruning and motif enrichment. **Network pruning** is the removal of proteins that are unlikely to interact or mediate their function through SLiMs. **Motif enrichment** is the removal of regions of a protein that are unlikely to contain a functional motif. Each noise reduction technique has many approaches, a few of which will be discussed here.

Functional motifs have been observed to be enriched in certain regions and impoverished in others, however, motifs that are over-represented due to chance occur at random within a protein/network. Protein masking and network pruning when carried out correctly will preferentially remove proteins and regions of proteins that are impoverished for functional motifs. For example, removing 50% of proteins known to be mediated by domain-domain interactions from the datasets through network pruning should remove 50% of background randomly occurring motif instances, yet should remove no functional SLiMs. This provides enrichment for the functional SLiM increasing the statistical power (the probability of seeing a SLiM 5 times in 10 proteins is much less likely than seeing that motif 5 times in 20 proteins).

4.4.1. Network pruning

4.4.1.1. Domain-domain interactions

Proteins often bind through the same mechanism (48) and interacting proteins containing the same domains as homologues known to bind through a domain-domain interaction will often reuse that mechanism of binding; although not certain to bind through a previously known mechanism from a different protein pair, the hypothesis is that

reuse of such interfaces is more likely than a novel SLiM-mediated binding mechanism. Removal of these interactions from the dataset may therefore enrich for SLiM mediated binding. Pruning data can be taken from DIMA (124), 3did (125) and iPfam (126), datasets of known domain-domain interactions can be taken from experimentally discovered complex structures for a stricter scheme, several methods are available to infer interacting domains from PPI networks (127).

4.4.1.2. Multidomain proteins

Multi-module proteins draw together several sub-networks interacting through different interfaces (SLiM or domain) of the hub protein into a single network. Extracting these sub-networks may allow a signal to be discovered. As 65% of Eukaryotic proteins are multidomain (128), this problem is of major interest for motif discovery analyses. By analyzing the domain architecture of their interactors, spoke proteins can often be classified into sub-networks that can be analyzed separately. Neduva *et al* (37) attempted to enrich datasets for domain-SLiM interactions for a particular domain by grouping together proteins containing that domain, and pooling the interaction partners thereby increasing the signal. PIANA (129) observed that proteins with common interaction partners tend to interact through common interaction interfaces that they termed iMotifs. By grouping proteins' common interaction partners, they discovered pairs of interacting interfaces.

4.4.1.3. Physical contact

As discussed above, often binary interaction data signifies interaction with a complex containing a protein rather that an actual physical interface with the protein. It is possible to infer proteins in the dataset that have no physical contact by analyzing the source of the interaction data. If there is both Y2H and TAP data for a complex, removing any complex members that have no Y2H data may enrich for these physical interactions. MPCDB (130) a database of known protein complexes provides useful data for this type of network pruning. As the interaction data becomes more complete such tasks will be simplified and accuracy may improve. At present, however, caution must be taken as direct interaction data might not be available for all complexes of interest.

4.4.1.4. Topology

It has been estimated by computational methods that between 15% and 39% of human proteins contain a trans-membrane region (131). Membranes act as a separator for distinct subsets of interactors, since extracellular (EC) proteins and intracellular (IC) proteins can't interact with the same set of protein regions when membrane bound (see Figure 4D). Splitting data across a membrane enriches the dataset for proteins interacting with a particular region of the proteins. This approach would consider the intersection of PPI data and localization data, however as both sources are incomplete, such an analysis would prove intractable yet not impossible. Despite these problems, topology filtering is highly recommended for low throughput motif discovery in systems for which reliable data is available.

4.4.2. Motif enrichment

Much of the information described in the "Biological attributes of SLiMs" section can be used for masking and filtering; masking removes the regions pre-analyses to improve speed and is mainly used for searches with high computation load such as Dilimot and SLiMfinder. Filtering is the post-processing of motifs to remove or flag data that is unlikely to be functional; the ELM server uses such an approach.

4.4.2.1. Domains/globular regions

Globular regions are often impoverished for SLiMs due to a combination of evolutionary and structural constraints (132) and approximately 85% of known SLiMs occur in disordered regions (2). It has been shown that using domain masking improves the ability of motif discovery tools to return functional motifs (91). Domain prediction is a highly mature field of bioinformatics with resources such as Pfam (60), Interpro (133) and SMART (61) providing detailed domain signatures and with high predictive power which can be used to remove known globular regions. It should be noted, however, that not all annotated domains are globular (134) and care must be taken with such filtering.

Disordered regions may directly predicted in order to focus searches; typically this will mask out globular regions. A large number of disorder analysis algorithms, using various methods such as simple amino acid biases in a given window (88), probabilistic models (135), complicated amino acid interaction data (33) and incorporating conservation information (136) are now available (for review, see (1)). Also, recent interest in the functional role of protein disorder has led to a large increase in experimental data for intrinsically disordered regions. This data has been curated in the Disprot database (137) providing motif discovery with high quality data for masking of a subset of proteins with known disorder as well as providing an excellent resource for benchmarking and training of disorder prediction tools.

4.4.2.2. Evolutionarily under-constrained residues

Although disordered regions tend to be less conserved than globular domains (35), it has been observed that disordered residues with functional constraint are more conserved than average (38). Using conservation information from orthologues to mask residues based on their level of evolutionary constraint in relation to their local sequence

context leads to enrichment for functional residues. A scoring scheme using relative conservation increased 4-fold the ability of SLiMFinder to discover known ELMs from HPRD interaction datasets, where a subset of interactions were known to be mediated by SLiMs (38). One of the main decisions of conservation measures is the choice of proteins for the alignment. Use of all homologues potentially offers more information and conservation of a motif in paralogues (products of gene duplications) as well as their orthologues (products of speciation) is a strong indicator of functionality. However, use of paralogues may increase the difficulty to create a quality alignment and post-duplication functional diversity might mean that paralogues do not have the SLiM despite its functional relevance in the protein of interest. Use of only orthologues allows a simpler model of conservation and cleaner alignments, without the need to consider the pressure to diversify functionality on paralogues. However, definition of orthologues can often be difficult.

The degree of sequence divergence in the considered alignment also has an important influence; sequences which are closely related will have very little change at any residues, while distantly related sequences will typically have unreliable alignments (for disordered regions in particular), therefore, there is an optimal degree of sequence divergence. Finally, aligning disordered regions remains a difficult problem. Recent work defined the problem, pointing out none of the programs currently available is capable of reliably aligning SLiMs in distantly related sequences, with no tool correctly aligning more than 73% of SLiMs (138). For these reasons, a relative conservation score – comparing the conservation of a given residue to that of its neighbors – is generally more powerful than an absolute score, which is highly dependent on homology levels and alignment quality (38). Hopefully, the addition of set of benchmarking alignments to the BAliBASE (138) will provide the impetus for research into the field of disorder alignment.

4.4.2.3. Topology

The use of topology, for network pruning, to split interaction datasets of membrane bound proteins into EC and IC interactors has been discussed. Similar logic can be used for proteins interacting with membrane bound proteins to mask regions which are inaccessible to the hub protein for interaction. For example, an IC hub protein may only interact with IC regions of a transmembrane spoke protein once it has localized to the membrane. Analysis of the dataset should be considered compartment specific and the proteins masked accordingly. Again, reliable data is required to do this accurately.

4.4.2.4. Surface accessibility

15% of known SLiMs occur in accessible portions, such as loops, in globular regions of proteins. Analysis of these regions is often desirable but for these analyses it is advisable to consider only accessible residues. Tools such as DSSP (139) can be used to return surface accessibility scores for proteins with experimentally derived 3D structures; a recently added ELM filter takes advantage of such methods. Often the structure of the protein of interest is not available, however coverage for homologous protein structures is improving, allowing homology modeling to predict a structure (see for review (140)) from which accessibility scores can be calculated. Several techniques are also available which attempt to calculate surface accessibility from primary sequence (see (141)).

5. MOTIF STATISTICS

5.1. Motif-based metrics

It is often desirable to compare motifs, ranking based on their level of degeneracy or their likelihood of occurence (for example, the Dynein binding motif KxTQT occurs considerably less frequently than the highly degenerate CK1 phosphorylation site Sxx (ST)). The most useful score for motif comparison is the probability that a motif of interest will occur by chance at a single position in a protein (this can also be thought of as the probability that a motif chosen at random from a sequence, with the same length and number of wildcards, will be a particular motif). This calculation (*eq. 1*), which is the basis of the SLiMChance algorithm (employed by SLiMFinder and SLiMSearch), is straightforward once an appropriate amino acid background probability is chosen. This score is particularly useful for ranking motifs discovered, based on level of interest, in a search of known motifs against a single protein.

Information Content (143) is a measure of randomness, which can be used to describe the degeneracy of a motif (*eq. 2*) (85); highly ambiguous motifs have high levels of degeneracy and therefore randomness. Although motifs can be compared based on the level of randomness, the Information Content of a motif does not correlate uniformly with the likelihood of a motif to occur by chance. However, Information Content has been used in motif discovery, SLiMDisc (91) and PRATT (85) both used the flexibility of the scoring scheme to avoid the strict dependency rules of probability based scoring schemes allowing the manipulation of homology in input datasets by weighting (144). Also, Information Content is widely used in conservation measures for scoring columns of a multiple alignments (see review (145)). Finally, CompariMotif (146), a tool for comparison of motifs, scores motifs similarity based on their normalized shared information content.

5.2. Protein-based metrics

When a motif is known and proteins are being searched for novel instances, matches can occur regularly by chance (Figure 5). When searching for novel instances of a motif in a protein it is useful to quantify how unlikely it is that the motif would occur in that protein by chance, particularly as the probability of the protein containing a particular motif is highly correlated with its length. These probabilities can be calculated both probabilistically and empirically.

5.2.1. Probabilistic calculation

With knowledge of amino acid frequencies it is possible to calculate the probability of a motif occurring at any position in the protein by chance (*eq. 1*). From this it is possible to calculate the probability, using the binomial distribution, of the motif occurring in the protein *k* times (*eq. 3*) or, the highly useful metric, 1 or more times (*eq. 4)* (62). Probabilistic methods will give unique probabilities for each protein considering the effect of protein length, however it does not explicitly account for compositional biases (although when tested no effect was discovered (62)). SLiMChance (SLiMFinder and SLiMSearch) is based on these calculations and MnM (47) uses similar methods to calculate highly intuitive fold enrichment scores.

5.2.2. Empirical calculations

In motif count based metrics such as those used in Dilimot (37), the support of a given motif is counted in a background dataset, often the entire proteome, and the probability of a motif occurring in a given protein is estimated as the proportion of proteins in the background dataset containing that motif (*eq. 5)*. Motif count methods are based on empirical data and consider possible compositional biases present in many proteins. It does not consider the differing probabilities for proteins of differing lengths.

5.2.3. Background sampling

The background data sampled for amino acid frequencies/motif counts will have a strong effect on the calculated probabilities. For example, a test set of extracellular proteins or highly disordered proteins will have very different amino acid frequencies, and therefore motif counts, than the whole proteome and this should, if possible, be considered. Empirical calculations can employ dataset matching, *i.e.* selecting background datasets with similar attributes to the test dataset is also possible, but over-fitting is a problem. For probabilistic methods, it is preferable to sample amino acid frequencies from the dataset of interest.

5.3. Dataset-based motif probability

The utility of the over-representation hypothesis (that over-representation of convergently evolved motifs is a pointer to purifying selection), as a tool for the discovery of putatively functional motifs in a dataset has been proven in analyses (49, 62). The major task of motif scoring for motif discovery is to separate motifs that are over-represented due to purifying selection (true positives) from those which are over represented due to chance (false positives). Several scoring schemes have been applied to tackle this problem.

Empirical schemes, such as the Information Content based metric used by SLiMDisc and PRATT, are based on the observation that the methods work well on benchmarking datasets (*eq. 7*) but have problems of false positives in dataset that do not contain any true motifs. Probabilistic binomial scoring schemes, such as SLiMChance (SLiMFinder and SLiMSearch) and Dilimot represent null hypotheses, the background distribution defining the probability that a motif will occur with a given support if there were no evolutionary pressures selecting for the motif. By comparing supports of motifs with this distribution it is possible to calculate how unlikely a motif is to occur with a given support by chance (*eq. 6*). The Fisher's Exact test, based on the hypergeometric distribution, is often used to test for enrichment in motif rediscovery analyses (66, 67, 80). A dataset will be tested for enrichment of a motif against a background dataset which is considered as a control (*eq. 8*).

5.3.1. Achieving independence

A probabilistic scoring scheme is suitable for analyses when data is independent, or data can be organized in such a way that it can be assumed independent. With the inclusion of non-independent (evolutionarily divergent) proteins in datasets, motifs are often shared due to the lack of evolutionary distance to accumulate mutations rather than due to a purifying selection to keep functional motifs. For instance, if a motif occurs once, the probability of reoccurrence increases in homologous proteins. In such cases, simpler scoring schemes based on Information Content, such as those used in PRATT (85) and SLiMDisc (91) are cleaner and more intuitive. However, advances in dataset modeling have allowed statistical models to accurately calculate probabilities considering the dependencies of proteins of common descent (62). Dilimot (79) removes regions of homology (BLAST E> 0.001), keeping only a single instance. SLiMFinder (62) groups proteins by homology and weights success probability based on a framework introduced by SLiMDisc (91).

5.4. Dataset-based motif significance

Dataset based motif over-representation scores (*eq. 6&8*) are motif centric, returning the probability that **a given motif** will reach its support by chance. This score allows useful ranking of returned motifs yet has two major drawbacks; (1) Motifs scores can only be compared against scores for similar motifs (same number of non-wildcard

positions (*i.e.* 3-mer scores and 5-mer scores are not comparable)) in the same dataset and (2) motif scores do not offer any indication that any motif in the dataset will achieve that score by chance (62). Each dataset will have a proportion of motifs which are extremely unlikely yet are at the tail of a distribution of expected binomial *p*-values for a dataset (This observation, a multiple testing problem, is similar to the extreme value distribution of BLAST (147) where although a hit between two proteins may be unlikely, when searching against a proteome the chances of two sub-sequences matching by chance increases rapidly).

A dataset significance score (*eq. 9)* calculates the probability that **any motif** in the dataset will reach a given binomial p-value score by chance. The score allows motifs to be compared across motif lengths and between datasets which is of huge benefit to high-throughput analyses allowing the ranking of datasets based on their level of interest. Dilimot (37) introduced a simple confidence threshold cut-off by random sampling of datasets, allowing motifs which are in the tail of an expected score distribution to be discovered. SLiMFinder (62) heuristically calculates a significance score approximating the probability that the dataset would return any motifs with a given *p*-value by chance.

5.5. Outstanding issues for motif statistics

5.5.1. Selection against motif occurrences

It has been hypothesized that there may be selection against groups of close (sub-cellular location wise) proteins evolving motifs that compete for binding with another protein. In such a case, there would be strong pressure on a motif to be removed from the protein. Via *et al.* (148) surveyed PROSITE motifs (54) finding some evidence of selection against novel instances of functional motifs convergently evolving. A survey analyzing this hypothesis on less specific motifs from SLiM databases such as ELM or MnM is difficult, as often instances of matches to a motif regular expression have not been differentiated as either true or false positives. However, if this effect is true and widespread, it will have implications for over-representation based motif discovery methods.

5.5.2. Classification of motifs

One of the major misconceptions about over-representation based motif discovery is that all instance of an over-represented motif are equally interesting. In fact, over-represented motifs are discovered when a probabilistically defined number of randomly occurring background instances of a motif occur as well as 1 or more functional instances of the motif. These functional instances cause the support for a given motif to be statistically unlikely and therefore discoverable. In the situations when the number of background instances of a motif is less than the expected support it can be difficult for a set of true positives and false positives matching that regular expression to reach a significance threshold. This also works advantageously for motifs when the number of true positives for a functional motif is low and when the numbers of false positives is above the expected support.

5.5.3. Significance of ambiguous motifs

Although the method for calculation of the binomial p-value of an ambiguous motif is well known (62) and significance values for fixed motifs have also been defined the significance of ambiguous motifs still has not been fully explored. Currently, a heuristic approach is used comparing the ambiguous motifs against significance distributions for fixed motifs (62). The complicated nature of motif ambiguity, being made up of multiple possible combinations of support for fixed position motifs, makes the calculations extremely difficult and any exact solution will undoubtedly be computationally expensive. In such situations, permutation tests may provide robust computationally expensive, but feasible, estimates of significance.

5.5.4. Non-independence of datasets

High-throughput analyses on GO and protein-protein interaction datasets introduce a difficult multiple testing problem (149). If the datasets were independent, having no overlap, the significance statistics described above would be able to account for this multiple testing problem as they quantify the number of datasets which would need to be analyzed to see such an over-represented motif by chance. However, the fact that GO and PPI datasets are highly overlapping causes a number of complex dependencies. Normal statistical measures are insufficient to deal with the highly dependent and overlapping data produced by these analyses. A large amount of research is available in the field of multiple testing for GO term enrichment (150), which could be modified in future for high throughput SLiM discovery.

6. MOTIF ANALYSIS

Classification of potential functional motifs is a difficult procedure; the following four steps may help increase confidence and make the process more empirical.

6.1. Matching known motifs

One of the first tasks when analyzing putatively functional motifs is comparison against datasets of known motifs. CompariMotif (146) is a tool for making motif–motif comparisons, identifying and describing similarities between regular expression motifs. Motif relationships are scored using shared Information Content, allowing the best

matches to be easily identified in large comparisons. Motifs can be searched against the datasets from the ELM (2) and MnM servers (47), as well as the PhosphoMotif Finder (151) phosphorylation site database, to find matches to known motifs as well as "fuzzy" matches to the regular expression of known motifs.

A literature search may yield motifs that have not yet been added to these SLiM repositories. Motif databases are not exhaustive, mainly due to the difficulty in motif curation from the literature. Although standards have been suggested (152, 153), they are sadly not widely adopted by the scientific community and motifs are described in several different formats, for example the canonical SH2 Grb2 (154) binding motif YxN has been described as YxN, pYxN, Y.N, Tyr-x-Asn as well as being described in relation to its surrounding residues Tyr-Gly-Asn-Gly. Additionally, abstracts often fail to mention the word "motif" at all, perhaps failing to differentiate motifs as a class from active sites in a domain, or instead using one of the many alternative (and often more specific) terms, such as peptide, interface, interaction site, SLiM, ELM, LM, minimotif or mOrf.

Proteins returning a putatively functional motif that contains a phosphorylatable residue can be cross-referenced with a dataset of phosphorylation sites such as the Phospho.ELM (50) database for information on whether phosphorylation of that residue has been seen experimentally. Much work has been carried out on the Kinome and multiple specialized tools are available to both discover novel phosphorylation sites and to predict particular kinases for a given site (69). Care should be taken with interpretation of hits, as the degeneracy of many known functional motif regular expression will often cause non-biologically meaningful matches.

6.2. Conservation

Although functional SLiMs in disordered regions are not as conserved as domains (4), mainly due to the lack of strong structural constraints, they are more conserved than surrounding residues and more importantly, more conserved than non-functional instances matching the regular expression for the motifs (155). Expanding this observation, conservation can be used to distinguish between true and false positive by classifying based on presence or absence in homologues. Dilimot (37) incorporates motif conservation into the scoring scheme and SLiMDisc (91) and SLiMFinder (62) provide conservation metrics that allow the user to gauge putative motif functionality.

Although many conservation scoring schemes have been suggested no consensus has been agreed amongst the community as to which method should be used (see for reviews (145, 156)). Recent interest in the field of motif discovery has led to the development of conservation measures specifically for describing SLiM conservation, these methods use an Information Content based scoring scheme which incorporated phylogeny information to weight sequences (155) and a probabilistic method (157). More recently, relative conservation has been introduced to allow the quantification of conservation of residues compared to their surrounding residues (38). The method also advocated splitting the data into two states to consider the differing levels of conservation for globular and disordered regions.

6.3. Confidence through context

One of the conundrums of SLiMs is that the multitudes of false positives, often indistinguishable from true positives, are easily avoidable by the binding partner. However, clues lie in the observation that the number of residues involved in the binding seems inadequate to provide the specificity observed in these interactions (39). Contextual information such as propensity to form secondary structure, surface accessibility, residue conservation data and information about known motifs can be very important in further investigation of a motif and can be decisive in the rejection or selection of a motif for further experimental analysis. For example, propensity to form a secondary structure (shown by the dip in IUPred disorder score) and high conservation of the intrinsic residues and neighboring context of the HP1 binding motif of Chromatin assembly factor 1 subunit A is clearly illustrative of a functional motif (Figure 6).

6.3.1. Structural information

Motifs that bind a common interface will, in general, bind with a similar secondary structure (for example the NRBOX motif LxxLL when bound forms a short alpha helix (158)). Using this information, it may be possible to differentiate true and false positives based on their propensity to form a particular secondary structure. A simple example would be the presence of a proline, a potent alpha helix breaker (159), in a putative motif for which all functional instances form an alpha helix when bound, which would be indicative of a false positive.

Often a motif can offer hints to its bound structure based on the residue spacing, for example the PCNA binding motif Qxx (ILM)xx (DHFM) (FMY) (160) or the MDM2 binding motif FxxxWxx (LIV) (161), both of which are natively alpha-helical and have defined residues matching the helical moment of 3.6 (162), signifying that these residues of the motif are adjacent on one side of the helix. Motifs with such a residue spacing, where residues with aligned side chains are more highly conserved and the region looks like it has a propensity to form a helix has a large body of contextual information to suggest that it is a functional helical motif.

Motifs obviously need to be surface accessible in order to be available for intermolecular interactions, and visualizing the position of a SLiM in a 3D structure, using visualization tools such as seeMotif (142), can give additional confidence, or otherwise, of a motif prediction. This is obviously limited by the availability of 3D structures, however, and an additional confounding factor in the case of SLiMs is their preponderance for occurring in structurally disordered regions of proteins, which are notoriously difficult to solve structurally due to their dynamic nature. (134). Secondly, the possibility must be borne in mind that a region may be buried in the typical conformation of a protein, but become accessible after a structural rearrangement.

6.4. Off-target motifs

A common occurrence in novel dataset-driven over-representation based motif discovery is "off-target" motifs, *i.e.* the discovery of a known functional motif that obviously is not mediating the function hypothesized for the analysis. In general, it is advisable to consider this possibility thoroughly when deciding on the next step of analysis such as validation.

6.4.1. Modification

Many proteins are regulated by protein modifications such as phosphorylation, ubiquitination or sumoylation; therefore motifs modified by these post-translational modifications are omnipresent. In biology, modification sites are the most commonly occurring SLiMs with many proteins known to have multiple modification sites (*e.g.* P53 has at least 14 phosphorylation sites (31)). As a result, there is a reasonable chance that **any** set of proteins will have over-represented PTM motifs. To help combat this problem, SLiMFinder provides options for masking user-defined motifs, or even specific amino acids (such as serines to reduce phosphorylation motifs) but it remains to be seen whether these options will significantly enhance motif discovery.

6.4.2. Localization

Sets of interacting proteins often co-localize in particular cellular regions (*e.g.* proteins involved in transcription will be present in the nucleus). The hypothesis of novel motif discovery is that a motif mediating the interaction would be over-represented, though often motifs returned from interaction datasets of these location-specific proteins can be localization signals (for example, commonly occurring nuclear localization signals (163) are often returned). This problem will obviously be exacerbated by the use of topological pruning of datasets. Cross-referencing motifs against a dataset of known localization signals (2, 47), or even masking these motifs, will aid in such situations.

6.4.3. Indirect binding

Due to the nature of interaction datasets (discussed in Dataset design), often the binary interaction network for a motif will include many proteins that interact with a complex involving the hub protein but have no direct interaction with the hub itself. This can cause hub-centric datasets to return a motif with which it does not interact with an interface on that hub, but with an interface on a binding partner in the same complex. Such a motif could prove an expensive false positive if brought to the experimental stage. Complex data for the hub, from sources such as immunoprecipitation, TAP or NMR or the MPCDB database (130) will provide information about the likelihood that a motif is an off target motif of this type. Such motifs can often be recognized by the fact that the interactomes of several members of the same complex are all returning the same motif.

6.4.4. Multi-functionality

Several motifs are known to be widely over-represented due to re-use of the motif by the proteome for multiple functions, for example arginine and lysine rich motifs such as KRK are involved in cleavage (11), localization (164) and modification (102). An unusual example is the N-terminus motif M (AGS) (AGS) which even has a genomic component. The mammalian translation initiation Kozak sequence GCCRCCaugG that binds mRNA to the small subunit of the ribosome also has an effect on the +2 residue as it favors G in the first position of the codon (165) enriching for Ala and Gly. The other components are, for methionine aminopeptidase that cleaves Methionine only when small amino acid occurs downstream (166) and myristoylation sites, a common N-terminal modification of Glycine (73). For this reason, SLiMFinder provides options to mask out these common motifs prior to searching.

7. CONCLUSION

The vast repertoire of activities mediated by SLiMs underlines the importance of their study and the vital part they play in cell functionality. Due to their elusiveness, both experimentally and computationally, many of SLiMs are still to be discovered. This, in conjunction with the recent expansion of experimentally derived examples, has made SLiM discovery a fruitful field of research that has expanded rapidly and is on the cusp of taking a place alongside domain-based tools as a primary source of protein function inference.

The potential of computational methods for motif discovery has been demonstrated and, although the deluge of motifs expected from these methods has yet to appear, methods have improved to a point where they can enrich experimental analysis. Advances continue in areas such as motif occurrence statistics, motif discovery algorithm

design, motif enrichment methods and motif classification strategies leaving the field primed for the inundation of experimental data expected over the coming years. However, many hurdles still remain; computationally, the statistics of ambiguous occurrences are still ill-defined, and the field of *de novo* motifs discovery from primary sequence is still in its infancy. Experimentally, methods for the high throughput discovery of SLiMs and SLiM-mediated interactions are still to be fully explored. Advances will be needed to design experiments to discover modules that are low affinity, highly regulated, and often temporal or reactionary to stimuli.

Our current knowledge is just the tip of the iceberg with regards to the importance of SLiMs and over the next decade we should see an explosion in the recognition of their significance by the wider biological community. Research in this area will at the very least enrich our understanding of cellular biology, but, it is not unfounded optimism to believe that they may play a central role in future therapeutic advances against a range of important human diseases.

ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland.

REFERENCES

1. F Diella, N Haslam, C Chica, A Budd, S Michael, NP Brown, G Trave & TJ Gibson: Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci* 13, 603, (2008)
2. P Puntervoll, R Linding, C Gemund, S Chabanis-Davidson, M Mattingsdal, S Cameron, DM Martin, G Ausiello, B Brannetti, A Costantini, F Ferre, V Maselli, A Via, G Cesareni, F Diella, G Superti-Furga, L Wyrwicz, C Ramu, C McGuigan, R Gudavalli, I Letunic, P Bork, L Rychlewski, B Kuster, M Helmer-Citterich, WN Hunter, R Aasland & TJ Gibson: ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 31, 30, (2003)
3. G Jimenez, CP Verrijzer & D Ish-Horowicz: A conserved motif in goosecoid mediates groucho-dependent repression in Drosophila embryos. *Mol Cell Biol* 19, 7, (1999)
4. V Neduva & RB Russell: Linear motifs: evolutionary interaction switches. *FEBS Lett* 579, 5, (2005)
5. H Ye, YC Park, M Kreishman, E Kieff & H Wu: The structural basis for the recognition of diverse receptor sequences by TRAF2. *Mol Cell* 4, 30, (1999)
6. SS Li & SS-C Li: Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. *Biochem J* 390, 653, (2005)
7. J Ren, L Wen, X Gao, C Jin, Y Xue & X Yao: CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng Des Sel* (2008)
8. D Shental-Bechor & Y Levy: Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. *Proc Natl Acad Sci U S A* 105, 8261, (2008)
9. SS Molloy, PA Bresnahan, SH Leppla, KR Klimpel & G Thomas: Human furin is a calcium-dependent serine endoprotease that recognizes the sequence Arg-X-X-Arg and efficiently cleaves anthrax toxin protective antigen. *J Biol Chem* 267, 16402, (1992)
10. JJ-D Hsieh, EH-Y Cheng & SJ Korsmeyer: Taspase1: a threonine aspartase required for cleavage of MLL and proper HOX gene expression. *Cell* 115, 303, (2003)
11. C Brenner & RS Fuller: Structural and enzymatic characterization of a purified prohormone-processing enzyme: secreted, soluble Kex2 protease. *Proc Natl Acad Sci U S A* 89, 926, (1992)
12. K Kadaveru, J Vyas & MR Schiller: Viral infection and human disease--insights from minimotifs. *Front Biosci* 13, 6471, (2008)
13. K Saksela, G Cheng & D Baltimore: Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef+ viruses but not for down-regulation of CD4. *EMBO J* 14, 491, (1995)
14. RN Harty, ME Brown, G Wang, J Huibregtse & FP Hayes: A PPxY motif within the VP40 protein of Ebola virus interacts physically and functionally with a ubiquitin ligase: implications for filovirus budding. *Proc Natl Acad Sci U S A* 97, 6, (2000)
15. G Fox, NR Parry, PV Barnett, B McGinn, DJ Rowlands & F Brown: The cell attachment site on foot-and-mouth disease virus includes the amino acid sequence RGD (arginine-glycine-aspartic acid). *J Gen Virol* 70, 637, (1989)
16. GS Tan, MA Preuss, JC Williams & MJ Schnell: The dynein light chain 8 binding motif of rabies virus phosphoprotein promotes efficient viral transcription. *Proc Natl Acad Sci U S A* 104, 34, (2007)
17. ME Jackson, JC Simpson, A Girod, R Pepperkok, LM Roberts & JM Lord: The KDEL retrieval system is exploited by Pseudomonas exotoxin A, but not by Shiga-like toxin-1, during retrograde transport from the Golgi complex to the endoplasmic reticulum. *J Cell Sci* 112, 475, (1999)

18. I Majoul, K Sohn, FT Wieland, R Pepperkok, M Pizza, J Hillemann & HD Soling: KDEL receptor (Erd2p)-mediated retrograde transport of the cholera toxin A subunit from the Golgi involves COPI, p23, and the COOH terminus of Erd2p. *J Cell Biol* 143, 612, (1998)
19. M Marti, RT Good, M Rug, E Knuepfer & AF Cowman: Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* 306, 1933, (2004)
20. MA McLane, J Gabbeta, AK Rao, L Beviglia, RA Lazarus & S Niewiarowski: A comparison of the effect of decorsin and two disintegrins, albolabrin and eristostatin, on platelet function. *Thromb Haemost* 74, 1322, (1995)
21. S Swenson, S Ramu & FS Markland: Anti-angiogenesis and RGD-containing snake venom disintegrins. *Curr Pharm Des* 13, 2871, (2007)
22. B Pandit, A Sarkozy, LA Pennacchio, C Carta, K Oishi, S Martinelli, EA Pogna, W Schackwitz, A Ustaszewska, A Landstrom, JM Bos, SR Ommen, G Esposito, F Lepri, C Faul, P Mundel, S Lopez, Juan P, R Tenconi, A Selicorni, C Rossi, L Mazzanti, I Torrente, B Marino, MC Digilio, G Zampino, MJ Ackerman, B Dallapiccola, M Tartaglia & BD Gelb: Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nat Genet* 39, 1012, (2007)
23. I Baran, RS Varekova, L Parthasarathi, S Suchomel, F Casey & DC Shields: Identification of potential small molecule peptidomimetics similar to motifs in proteins. *J Chem Inf Model* 47, 474, (2007)
24. Y Cheng, T LeGall, CJ Oldfield, JP Mueller, Y-YJ Van, P Romero, MS Cortese, VN Uversky & AK Dunker: Rational drug design via intrinsically disordered protein. *Trends Biotechnol* 24, 442, (2006)
25. PA Burke, SJ DeNardo, LA Miers, KR Lamborn, S Matzku & GL DeNardo: Cilengitide targeting of alpha(v)beta(3) integrin receptor synergizes with radioimmunotherapy to increase efficacy and apoptosis in breast cancer xenografts. *Cancer Res* 62, 4272, (2002)
26. C Tovar, J Rosinski, Z Filipovic, B Higgins, K Kolinsky, H Hilton, X Zhao, BT Vu, W Qing, K Packman, O Myklebost, DC Heimbrook & LT Vassilev: Small-molecule MDM2 antagonists reveal aberrant p53 signaling in cancer: implications for therapy. *Proc Natl Acad Sci U S A* 103, 1893, (2006)
27. LT Vassilev, BT Vu, B Graves, D Carvajal, F Podlaski, Z Filipovic, N Kong, U Kammlott, C Lukacs, C Klein, N Fotouhi & EA Liu: In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* 303, 848, (2004)
28. NA Laurie, SL Donovan, C-S Shih, J Zhang, N Mills, C Fuller, A Teunisse, S Lam, Y Ramos, A Mohan, D Johnson, M Wilson, C Rodriguez-Galindo, M Quarto, S Francoz, SM Mendrysa, RK Guy, J-C Marine, AG Jochemsen & MA Dyer: Inactivation of the p53 pathway in retinoblastoma. *Nature* 444, 66, (2006)
29. J Davydova, LP Le, T Gavrikova, M Wang, V Krasnykh & M Yamamoto: Infectivity-enhanced cyclooxygenase-2-based conditionally replicative adenoviruses for esophageal adenocarcinoma treatment. *Cancer Res* 64, 4327, (2004)
30. M Fuxreiter, P Tompa & I Simon: Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23, 956, (2007)
31. RB Russell & TJ Gibson: A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS Lett* 582, 1275, (2008)
32. P Romero, Z Obradovic, X Li, EC Garner, CJ Brown & AK Dunker: Sequence complexity of disordered protein. *Proteins* 42, 48, (2001)
33. Z Dosztanyi, V Csizmok, P Tompa & I Simon: IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 4, (2005)
34. PE Wright & HJ Dyson: Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293, 331, (1999)
35. CJ Brown, S Takayama, AM Campen, P Vise, TW Marshall, CJ Oldfield, CJ Williams & AK Dunker: Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* 55, 110, (2002)
36. S Abeln & D Frenkel: Disordered flanks prevent peptide aggregation. *PLoS Comput Biol* 4, (2008)
37. V Neduva, R Linding, I Su-Angrand, A Stark, M de, Federico, TJ Gibson, J Lewis, L Serrano & RB Russell: Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol* 3, (2005)
38. NE Davey, DC Shields & RJ Edwards: Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics* 25, 450, (2009)
39. A Stein & P Aloy: Contextual specificity in peptide-mediated protein interactions. *PLoS ONE* 3, (2008)
40. A Zarrinpar, S-H Park & WA Lim: Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* 426, 680, (2003)
41. C Lee, JH Chang, HS Lee & Y Cho: Structural basis for the recognition of the E2F transactivation domain by the retinoblastoma tumor suppressor. *Genes Dev* 16, 3212, (2002)
42. FC Stomski, M Dottore, W Winnall, MA Guthridge, J Woodcock, CJ Bagley, DT Thomas, RK Andrews, MC Berndt & AF Lopez: Identification of a 14-3-3 binding sequence in the common beta chain of the granulocyte-macrophage colony-stimulating factor (GM-CSF), interleukin-3 (IL-3), and IL-5 receptors that is serine-phosphorylated by GM-CSF. *Blood* 94, 1942, (1999)
43. MT Drake, MA Downs & LM Traub: Epsin binds to clathrin by associating directly with the clathrin-terminal domain. Evidence for cooperative binding through two discrete sites. *J Biol Chem* 275, 89, (2000)

44. H Remaut & G Waksman: Protein-protein interaction through beta-strand addition. *Trends Biochem Sci* 31, 444, (2006)
45. V Vacic, CJ Oldfield, A Mohan, P Radivojac, MS Cortese, VN Uversky & AK Dunker: Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* 6, 2366, (2007)
46. A Bairoch, R Apweiler, CH Wu, WC Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, MJ Martin, DA Natale, C O'Donovan, N Redaschi & LS Yeh: The Universal Protein Resource (UniProt). *Nucleic Acids Res* 33, 9, (2005)
47. S Rajasekaran, S Balla, P Gradie, MR Gryk, K Kadaveru, V Kundeti, MW Maciejewski, T Mi, N Rubino, J Vyas & MR Schiller: Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res* 37, 90, (2009)
48. B Schuster-Bockler & A Bateman: Reuse of structural domain-domain interactions in protein networks. *BMC Bioinformatics* 8, (2007)
49. V Neduva & RB Russell: Peptides mediating interaction networks: new leads at last. *Curr Opin Biotechnol* 17, 71, (2006)
50. F Diella, CM Gould, C Chica, A Via & TJ Gibson: Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res* 36, 4, (2008)
51. T-Y Lee, H-D Huang, J-H Hung, H-Y Huang, Y-S Yang & T-H Wang: dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res* 34, 7, (2006)
52. ND Rawlings, FR Morton, CY Kok, J Kong & AJ Barrett: MEROPS: the peptidase database. *Nucleic Acids Res* 36, 5, (2008)
53. Y Igarashi, A Eroshkin, S Gramatikova, K Gramatikoff, Y Zhang, JW Smith, AL Osterman & A Godzik: CutDB: a proteolytic event database. *Nucleic Acids Res* 35, 9, (2007)
54. N Hulo, CJ Sigrist, S Le, V., PS Langendijk-Genevaux, L Bordoli, A Gattiker, C De, E., P Bucher & A Bairoch: Recent improvements to the PROSITE database. *Nucleic Acids Res* 32, 7, (2004)
55. PV Hornbeck, I Chabra, JM Kornhauser, E Skrzypek & B Zhang: PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 4, 1561, (2004)
56. R Gupta, H Birch, K Rapacki, S Brunak & JE Hansen: O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* 27, 372, (1999)
57. AL Chernorudskiy, A Garcia, EV Eremin, AS Shorina, EV Kondratieva & MR Gainullin: UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics* 8, (2007)
58. S Balla, V Thapar, S Verma, T Luong, T Faghri, CH Huang, S Rajasekaran, C del, J. J., JH Shinn, WA Mohler, MW Maciejewski, MR Gryk, B Piccirillo, SR Schiller & MR Schiller: Minimotif Miner: a tool for investigating protein function. *Nat Methods* 3, 7, (2006)
59. R Gutman, C Berezin, R Wollman, Y Rosenberg & N Ben-Tal: QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. *Nucleic Acids Res* 33, 61, (2005)
60. A Bateman, L Coin, R Durbin, RD Finn, V Hollich, S Griffiths-Jones, A Khanna, M Marshall, S Moxon, EL Sonnhammer, DJ Studholme, C Yeats & SR Eddy: The Pfam protein families database. *Nucleic Acids Res* 32, 41, (2004)
61. I Letunic, T Doerks & P Bork: SMART 6: recent updates and new developments. *Nucleic Acids Res* 37, 32, (2009)
62. RJ Edwards, NE Davey & DC Shields: SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE* 2, (2007)
63. D Betel, KE Breitkreuz, R Isserlin, D Dewar-Darch, M Tyers & CWV Hogue: Structure-templated predictions of novel protein interactions from sequence information. *PLoS Comput Biol* 3, 1789, (2007)
64. B Brannetti & M Helmer-Citterich: iSPOT: A web tool to infer the interaction specificity of families of protein modules. *Nucleic Acids Res* 31, 3711, (2003)
65. C Ramu: SIRW: A web server for the Simple Indexing and Retrieval System that combines sequence motif searches with keyword searches. *Nucleic Acids Res* 31, 3774, (2003)
66. RR Copley: The EH1 motif in metazoan transcription factors. *BMC Genomics* 6, (2005)
67. S Michael, G Trave, C Ramu, C Chica & TJ Gibson: Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. *Bioinformatics* 24, 7, (2008)
68. TB Schreiber, N Mausbacher, SB Breitkopf, K Grundner-Culemann & H Daub: Quantitative phosphoproteomics - an emerging key technology in signal-transduction research. *Proteomics* (2008)
69. R Linding, LJ Jensen, GJ Ostheimer, MATM van Vugt, C Jorgensen, IM Miron, F Diella, K Colwill, L Taylor, K Elder, P Metalnikov, V Nguyen, A Pasculescu, J Jin, JG Park, LD Samson, JR Woodgett, RB Russell, P Bork, MB Yaffe & T Pawson: Systematic discovery of in vivo phosphorylation networks. *Cell* 129, 1426, (2007)
70. JC Obenauer, LC Cantley & MB Yaffe: Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res* 31, 3641, (2003)
71. D Plewczynski, A Tkacz, LS Wyrwicz, L Rychlewski & K Ginalski: AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. *J Mol Model* 14, 76, (2008)

72. K Julenius & K Julenius: NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. *Glycobiology* 17, 876, (2007)
73. G Bologna, C Yvon, S Duvaud & A-L Veuthey: N-Terminal myristoylation predictions by ensembles of neural networks. *Proteomics* 4, 1632, (2004)
74. F Monigatti, E Gasteiger, A Bairoch & E Jung: The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics* 18, 770, (2002)
75. MR Wilkins, E Gasteiger, A Bairoch, JC Sanchez, KL Williams, RD Appel & DF Hochstrasser: Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol* 112, 552, (1999)
76. O Emanuelsson, S Brunak, H von, Gunnar & H Nielsen: Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2, 971, (2007)
77. P Duckert, S Brunak & N Blom: Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel* 17, 112, (2004)
78. O Schilling & CM Overall: Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat Biotechnol* 26, 694, (2008)
79. V Neduva & RB Russell: DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res* 34, 5, (2006)
80. F Diella, S Chabanis, K Luck, C Chica, C Ramu, C Nerlov & TJ Gibson: KEPE--a motif frequently superimposed on sumoylation sites in metazoan chromatin proteins and transcription factors. *Bioinformatics* 25, 5, (2009)
81. I Rigoutsos & A Floratos: Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* 14, 67, (1998)
82. MC Frith, NFW Saunders, B Kobe & TL Bailey: Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol* 4, (2008)
83. TL Bailey, N Williams, C Misleh & WW Li: MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34, 73, (2006)
84. S-H Tan, W Hugo, W-K Sung & S-K Ng: A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics* 7, (2006)
85. I Jonassen, JF Collins & DG Higgins: Finding flexible patterns in unaligned protein sequences. *Protein Sci* 4, 95, (1995)
86. AF Neuwald & P Green: Detecting patterns in protein sequences. *J Mol Biol* 239, 712, (1994)
87. A Brazma, I Jonassen, I Eidhammer & D Gilbert: Approaches to the automatic discovery of patterns in biosequences. *J Comput Biol* 5, 305, (1998)
88. R Linding, RB Russell, V Neduva & TJ Gibson: GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31, 3708, (2003)
89. GR Mishra, M Suresh, K Kumaran, N Kannabiran, S Suresh, P Bala, K Shivakumar, N Anuradha, R Reddy, TM Raghavan, S Menon, G Hanumanthu, M Gupta, S Upendran, S Gupta, M Mahesh, B Jacob, P Mathew, P Chatterjee, KS Arun, S Sharma, KN Chandrika, N Deshpande, K Palvankar, R Raghavnath, R Krishnakanth, H Karathia, B Rekha, R Nayak, G Vishnupriya, HG Kumar, M Nagini, GS Kumar, R Jose, P Deepthi, SS Mohan, TK Gandhi, HC Harsha, KS Deshpande, M Sarker, TS Prasad & A Pandey: Human protein reference database--2006 update. *Nucleic Acids Res* 34, 4, (2006)
90. NE Davey, RJ Edwards & DC Shields: The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res* 35, 9, (2007)
91. NE Davey, DC Shields & RJ Edwards: SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res* 34, 54, (2006)
92. Y Cheng, CJ Oldfield, J Meng, P Romero, VN Uversky & AK Dunker: Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* 46, 13477, (2007)
93. B Meszaros, I Simon & Z Dosztanyi: Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* 5, (2009)
94. M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin & G Sherlock: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 29, (2000)
95. DP Hill, B Smith, MS McAndrews-Hill & JA Blake: Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics* 9 Suppl 5, (2008)
96. SJ Gould, GA Keller, N Hosken, J Wilkinson & S Subramani: A conserved tripeptide sorts proteins to peroxisomes. *J Cell Biol* 108, 64, (1989)
97. LV Lotti, G Mottola, MR Torrisi & S Bonatti: A different intracellular distribution of a single reporter protein is determined at steady state by KKXX or KDEL retrieval signals. *J Biol Chem* 274, 10420, (1999)
98. C Conrad, H Erfle, P Warnat, N Daigle, T Lorch, J Ellenberg, R Pepperkok & R Eils: Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res* 14, 1136, (2004)
99. J Sprenger, F Lynn, J, S Karunaratne, K Hanson, NA Hamilton & RD Teasdale: LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res* 36, 3, (2008)

100. JC Simpson, R Wellenreuther, A Poustka, R Pepperkok & S Wiemann: Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep* 1, 292, (2000)
101. HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov & PE Bourne: The Protein Data Bank. *Nucleic Acids Res* 28, 42, (2000)
102. MPH Stumpf, T Thorne, S de, Eric, R Stewart, HJ An, M Lappe & C Wiuf: Estimating the size of the human interactome. *Proc Natl Acad Sci U S A* 105, 6964, (2008)
103. P Bork: Comparative analysis of protein interaction networks. *Bioinformatics* 18 Suppl 2, (2002)
104. ME Futschik, G Chaurasia & H Herzel: Comparison of human protein-protein interaction maps. *Bioinformatics* 23, 611, (2007)
105. H Yu, P Braun, MA Yildirim, I Lemmens, K Venkatesan, J Sahalie, T Hirozane-Kishikawa, F Gebreab, N Li, N Simonis, T Hao, J-F Rual, A Dricot, A Vazquez, RR Murray, C Simon, L Tardivo, S Tam, N Svrzikapa, C Fan, A-S de Smet, A Motyl, ME Hudson, J Park, X Xin, ME Cusick, T Moore, C Boone, M Snyder, FP Roth, A-L Barabasi, J Tavernier, DE Hill & M Vidal: High-quality binary protein interaction map of the yeast interactome network. *Science* 322, 110, (2008)
106. P Uetz, L Giot, G Cagney, TA Mansfield, RS Judson, JR Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields & JM Rothberg: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403, 627, (2000)
107. LJ Jensen, M Kuhn, S Chaffron, T Doerks, B Kruger, B Snel & P Bork: STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 35, 62, (2007)
108. C Alfarano, CE Andrade, K Anthony, N Bahroos, M Bajec, K Bantoft, D Betel, B Bobechko, K Boutilier, E Burgess, K Buzadzija, R Cavero, C D'Abreo, I Donaldson, D Dorairajoo, MJ Dumontier, MR Dumontier, V Earles, R Farrall, H Feldman, E Garderman, Y Gong, R Gonzaga, V Grytsan, E Gryz, V Gu, E Haldorsen, A Halupa, R Haw, A Hrvojic, L Hurrell, R Isserlin, F Jack, F Juma, A Khan, T Kon, S Konopinsky, V Le, E Lee, S Ling, M Magidin, J Moniakis, J Montojo, S Moore, B Muskat, I Ng, JP Paraiso, B Parker, G Pintilie, R Pirone, JJ Salama, S Sgro, T Shan, Y Shu, J Siew, D Skinner, K Snyder, R Stasiuk, D Strumpf, B Tuekam, S Tao, Z Wang, M White, R Willis, C Wolting, S Wong, A Wrong, C Xin, R Yao, B Yates, S Zhang, K Zheng, T Pawson, BFF Ouellette & CWV Hogue: The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* 33, 24, (2005)
109. L Salwinski, CS Miller, AJ Smith, FK Pettit, JU Bowie & D Eisenberg: The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32, 51, (2004)
110. S Kerrien, Y Alam-Faruque, B Aranda, I Bancarz, A Bridge, C Derow, E Dimmer, M Feuermann, A Friedrichsen, R Huntley, C Kohler, J Khadake, C Leroy, A Liban, C Lieftink, L Montecchi-Palazzi, S Orchard, J Risse, K Robbe, B Roechert, D Thorneycroft, Y Zhang, R Apweiler & H Hermjakob: IntAct--open source resource for molecular interaction data. *Nucleic Acids Res* 35, 5, (2007)
111. P Pagel, S Kovac, M Oesterheld, B Brauner, I Dunger-Kaltenbach, G Frishman, C Montrone, P Mark, V Stumpflen, H-W Mewes, A Ruepp & D Frishman: The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21, 834, (2005)
112. A Chatr-aryamontri, A Ceol, LM Palazzi, G Nardelli, MV Schneider, L Castagnoli & G Cesareni: MINT: the Molecular INTeraction database. *Nucleic Acids Res* 35, 4, (2007)
113. I Vastrik, P D'Eustachio, E Schmidt, G Gopinath, D Croft, B de, Bernard, M Gillespie, B Jassal, S Lewis, L Matthews, G Wu, E Birney & L Stein: Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 8, (2007)
114. S Mathivanan, B Periaswamy, TKB Gandhi, K Kandasamy, S Suresh, R Mohmood, YL Ramachandra & A Pandey: An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* 7 Suppl 5, (2006)
115. M Koegl & P Uetz: Improving yeast two-hybrid screening systems. *Brief Funct Genomic Proteomic* 6, 312, (2007)
116. J Zha, H Harada, E Yang, J Jockel & SJ Korsmeyer: Serine phosphorylation of death agonist BAD in response to survival factor results in binding to 14-3-3 not BCL-X(L). *Cell* 87, 628, (1996)
117. M Colledge & SC Froehner: Tyrosine phosphorylation of nicotinic acetylcholine receptor mediates Grb2 binding. *J Neurosci* 17, 5045, (1997)
118. L Hakes, JW Pinney, DL Robertson & SC Lovell: Protein-protein interaction networks and biology--what's the connection? *Nat Biotechnol* 26, 72, (2008)
119. CC Harris: p53: at the crossroads of molecular carcinogenesis and risk assessment. *Science* 262, 1981, (1993)
120. DP Lane: Cancer. p53, guardian of the genome. *Nature* 358, 16, (1992)
121. I Stagljar & S Fields: Analysis of membrane protein interactions using yeast-based technologies. *Trends Biochem Sci* 27, 563, (2002)
122. RS Savkur & TP Burris: The coactivator LXXLL nuclear receptor recognition motif. *J Pept Res* 63, 212, (2004)
123. ME Cusick, N Klitgord, M Vidal & DE Hill: Interactome: gateway into systems biology. *Hum Mol Genet* 14 Spec No. 2, 81, (2005)

124. P Pagel, M Oesterheld, V Stumpflen & D Frishman: The DIMA web resource--exploring the protein domain network. *Bioinformatics* 22, 998, (2006)
125. A Stein, A Panjkovich & P Aloy: 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res* 37, 4, (2009)
126. RD Finn, M Marshall & A Bateman: iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 21, 2, (2005)
127. BA Shoemaker & AR Panchenko: Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol* 3, (2007)
128. D Ekman, AK Bjorklund, J Frey-Skott & A Elofsson: Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol* 348, 43, (2005)
129. R Aragues, D Jaeggi & B Oliva: PIANA: protein interactions and network analysis. *Bioinformatics* 22, 1017, (2006)
130. HW Mewes, S Dietmann, D Frishman, R Gregory, G Mannhaupt, KFX Mayer, M Munsterkotter, A Ruepp, M Spannagl, V Stumpflen & T Rattei: MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res* 36, 201, (2008)
131. M Ahram, ZI Litou, R Fang & G Al-Tawallbeh: Estimation of membrane proteins in the human proteome. *In Silico Biol* 6, 386, (2006)
132. S Ren, VN Uversky, Z Chen, AK Dunker & Z Obradovic: Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. *BMC Genomics* 9 Suppl 2, (2008)
133. S Hunter, R Apweiler, TK Attwood, A Bairoch, A Bateman, D Binns, P Bork, U Das, L Daugherty, L Duquenne, RD Finn, J Gough, D Haft, N Hulo, D Kahn, E Kelly, A Laugraud, I Letunic, D Lonsdale, R Lopez, M Madera, J Maslen, C McAnulla, J McDowall, J Mistry, A Mitchell, N Mulder, D Natale, C Orengo, AF Quinn, JD Selengut, CJA Sigrist, M Thimma, PD Thomas, F Valentin, D Wilson, CH Wu & C Yeats: InterPro: the integrative protein signature database. *Nucleic Acids Res* 37, 5, (2009)
134. P Tompa, M Fuxreiter, CJ Oldfield, I Simon, AK Dunker & VN Uversky: Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* 31, 335, (2009)
135. A Bulashevska & R Eils: Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered. *J Theor Biol* (2008)
136. ZR Yang, R Thomson, P McNeil & RM Esnouf: RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21, 3376, (2005)
137. M Sickmeier, JA Hamilton, T LeGall, V Vacic, MS Cortese, A Tantos, B Szabo, P Tompa, J Chen, VN Uversky, Z Obradovic & AK Dunker: DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 35, 93, (2007)
138. E Perrodou, C Chica, O Poch, TJ Gibson & JD Thompson: A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics* 9, (2008)
139. W Kabsch & C Sander: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 637, (1983)
140. Y Zhang & Y Zhang: Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* 18, 348, (2008)
141. H Chen & H-X Zhou: Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins* 61, 35, (2005)
142. CE Shannon: The mathematical theory of communication. 1963. *MD Comput* 14, 17, (1997)
143. I Jonassen, JF Collins & DG Higgins: Scoring function for pattern discovery programs taking into account sequence diversity. *Reports in Informatics* (1996)
144. JA Capra & M Singh: Predicting functionally important residues from sequence conservation. *Bioinformatics* 23, 82, (2007)
145. RJ Edwards, NE Davey & DC Shields: CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics* 24, 1309, (2008)
146. SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller & DJ Lipman: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 402, (1997)
147. A Via, PF Gherardini, E Ferraro, G Ausiello, T Scalia, Gianpaolo & M Helmer-Citterich: False occurrences of functional motifs in protein sequences highlight evolutionary constraints. *BMC Bioinformatics* 8, (2007)
148. R Rosenfeld, I Simon, GJ Nau & Z Bar-Joseph: A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Res* 36, (2008)
149. JJ Goeman & U Mansmann: Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 24, 544, (2008)
150. R Amanchy, B Periaswamy, S Mathivanan, R Reddy, SG Tattikota & A Pandey: A curated compendium of phosphorylation motifs. *Nat Biotechnol* 25, 286, (2007)
151. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and symbolism for amino acids and peptides. Corrections to recommendations 1983. *Eur J Biochem* 213, (1993)

152. R Aasland, C Abrams, C Ampe, LJ Ball, MT Bedford, G Cesareni, M Gimona, JH Hurley, T Jarchau, VP Lehto, MA Lemmon, R Linding, BJ Mayer, M Nagai, M Sudol, U Walter & SJ Winder: Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett* 513, 144, (2002)
153. HWHG Kessels, AC Ward & TNM Schumacher: Specificity and affinity motifs for Grb2 SH2-ligand interactions. *Proc Natl Acad Sci U S A* 99, 8529, (2002)
154. C Chica, A Labarga, CM Gould, R Lopez & TJ Gibson: A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bio* (2008)
155. WS Valdar: Scoring residue conservation. *Proteins* 48, 41, (2002)
156. H Dinkel & H Sticht: A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics* 23, 303, (2007)
157. BD Darimont, RL Wagner, JW Apriletti, MR Stallcup, PJ Kushner, JD Baxter, RJ Fletterick & KR Yamamoto: Structure and specificity of nuclear receptor-coactivator interactions. *Genes Dev* 12, 3356, (1998)
158. PY Chou & GD Fasman: Prediction of protein conformation. *Biochemistry* 13, 245, (1974)
159. E Warbrick: The puzzle of PCNA's many partners. *Bioessays* 22, 1006, (2000)
160. M Uesugi & GL Verdine: The alpha-helical FXXPhiPhi motif in p53: TAF interaction and discrimination by MDM2. *Proc Natl Acad Sci U S A* 96, 14806, (1999)
161. C Cohen & DA Parry: Alpha-helical coiled coils and bundles: how to design an alpha-helical protein. *Proteins* 7, 15, (1990)
162. DT Chang, TY Chien & CY Chen: seeMotif: exploring and visualizing sequence motifs in 3D structures. *Nucleic Acids Res* (2009)
163. SH Liang & MF Clarke: A bipartite nuclear localization signal is required for p53 nuclear import regulated by a carboxyl-terminal domain. *J Biol Chem* 274, 32703, (1999)
164. G Craggs & S Kellie: A functional nuclear localization sequence in the C-terminal domain of SHP-1. *J Biol Chem* 276, 23725, (2001)
165. M Kozak: Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44, 92, (1986)
166. F Sherman, JW Stewart & S Tsunasawa: Methionine or not methionine at the beginning of a protein. *Bioessays* 3, 31, (1985)

Abbreviations: EC: Extracellular, ELM: Eukaryotic Linear Motif, GO: Gene Ontology, HPRD: Human Protein Reference Database, IC: Intracellular, PSSM: Position-Specific Scoring Matrices., PTM: Post-Translational Modification, SLiM: Short, Linear Motif, SVM: Support Vector Machine, TAP: Tandem Affinity Purification, Y2H: Yeast-2-Hybrid

Key Words: Protein Linear Motif, Motif, Protein, Short Linear Motif, Evolution, Bioinformatics, Review

Send correspondence to: Denis Shields, UCD Complex and Adaptive Systems Laboratory, University College Dublin, Belfield, Dublin 4, Tel: 00-353-17165344, Fax: 00-353-17165396, E-mail: denis.shields@ucd.ie

Table 1. List of motifs statistics

Score type	Description	Equation	#
Motif-based metrics	Probability a motif chosen at random will be motif *m*	$p_m = \prod_{i=1}^{l} \sum_{k=1}^{x} f(m_{ik})$	***(eq. 1)***
	Information Content based measure of randomness of the motif *m*	$IC_m = \sum_{i=1}^{l} -\log_{20}(\sum_{k=1}^{x} f(m_{ik}))$	***(eq. 2)***
Protein-based metrics	Probability that the motif *m* will occur c times in a protein	$p_c = \frac{N!}{(N-c)!c!} {p_m}^c (1-p_m)^{N-c}$	***(eq. 3)***
	Probability that the motif *m* will occur 1 or more times in a protein	$p_{1+} = 1-(1-p_m)^N$	***(eq. 4)***
	Count-based probability the motif *m* occurring 1 or mores times in a protein	$p_{1+} = C/B$	***(eq. 5)***
Dataset-based metrics	Probability a given motif will occur with a support of *k* or more in a dataset	$p = \sum_{j=k}^{n} \frac{n!}{(n-k)!k!} {p_{1+\mu}}^k (1-p_{1+\mu})^{n-k}$	***(eq. 6)***
	Information Content based empirical score for the motif *m* in n proteins	$IC = n_w * IC_m$	***(eq. 7)***
	Count-based probability the motif *m* will occur support of k times in a dataset	$p = \frac{\binom{C}{k}\binom{B-C}{n-k}}{\binom{B}{n}}$	***(eq. 8)***
Dataset-based Significance	Estimated probability any motif in a dataset will reach the *p* of motif *m*	$Sig = 1-(1-p)^{Rl}$	***(eq. 9)***

m is the motif of interest, *l* is the number of non-wildcard positions in the motif, *x* is the maximum length of a wildcard region allowed, m_i is position *i* in the motif, *x* is the number of ambiguous possibilities at position *i*, m_{ik} is the k^{th} ambiguous possibility in m_i , $f(m_{ik})$ is the background frequency of the amino acid m_{ik}. *N* is the number of positions in the protein that the motif *m* can occur, *C* is the count of proteins containing 1 or more occurrence of the motif in a background dataset, *B* is the size of the background dataset. *n* is the number of proteins or protein clusters in the dataset, *k* is the support of the motif (i.e. the number of proteins containing it), $p_{1+\mu}$ is the mean success probability of a motif occurring in any protein in the dataset, n_w is the support weighted based on the homology of the proteins in the dataset. R is calculated as $20^l (x+1)^{l-1}$.

Figure 1. Comparison of the IUPred scores (33) for disorder (0 is highly globular and 1 is strongly disordered) of ELM and Non-ELM residues for ELM containing proteins in the ELM database (2).

Figure 2. Comparison of the relative local conservation (RLC) scores (38) of ELM (motif) and Non-ELM (all other) residues for ELM-containing proteins in the ELM database.

Figure 3. The difference in the proportion of residues contained in known functional motifs, from the ELM database (2), compared to background amino acid frequencies from UNIPROT human (46). Fixed refers to motif positions that are non-ambiguous, ambiguous to motif positions that are ambiguous, and full to the combination of both fixed and ambiguous positions.

Figure 4. Representation of the different levels of information available for protein-protein interaction data. A: The binary level , B: the protein complex level, C: the atomic level and D the topology level.

Figure 5. Expectation of known ELMs. The distribution of counts, for all 132 motif classes in the ELM database, against the number of residues inspected before 1 instance of a given ELM would be expected by chance.

Figure 6. HP1 binding motif of Chromatin assembly factor 1 subunit A. (A) IUPred plot for the 100 residues window surrounding the motif showing the decreased disorder surrounding the motif. (B) The beta strand structure of the bound ligand binding by beta augmentation. (C) Conservation of the neighboring region shows the constraints surrounding the core PxVxL residues.

Running title: Computational discovery of short linear motifs