BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 4 2009, pages 443–450
doi:10.1093/bioinformatics/btn664

Sequence analysis

Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery

Norman E. Davey[1], Denis C. Shields[1] and Richard J. Edwards[2,*]

[1]UCD Complex and Adaptive Systems Laboratory, UCD Conway Institute of Biomolecular and Biomedical Sciences, University College Dublin, Dublin 4, Ireland and [2]School of Biological Sciences, University of Southampton, Boldrewood Campus, Southampton SO16 7PX, UK

Received on September 4, 2008; revised on December 1, 2008; accepted on December 27, 2008

Advance Access publication January 9, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Short linear motifs (SLiMs) are important mediators of protein–protein interactions. Their short and degenerate nature presents a challenge for computational discovery. We sought to improve SLiM discovery by incorporating evolutionary information, since SLiMs are more conserved than surrounding residues.

Results: We have developed a new method that assesses the evolutionary signal of a residue in its sequence and structural context. Under-conserved residues are masked out prior to SLiM discovery, allowing incorporation into the existing statistical model employed by SLiMFinder. The method shows considerable robustness in terms of both the conservation score used for individual residues and the size of the sequence neighbourhood. Optimal parameters significantly improve return of known functional motifs from benchmarking data, raising the return of significant validated SLiMs from typical human interaction datasets from 20% to 60%, while retaining the high level of stringency needed for application to real biological data. The success of this regime indicates that it could be of general benefit to computational annotation and prediction of protein function at the sequence level.

Availability: All data and tools in this article are available at http://bioware.ucd.ie/~slimdisc/slimfinder/conmasking/.

Contact: r.edwards@southampton.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Short linear motifs (SLiMs) are functional microdomains that ubiquitously play a central role in biological pathways (Diella *et al*., 2008; Neduva and Russell, 2005). SLiM-mediated tasks include recognition sites for protein modification, protein cleavage sites and targeting motifs for subcellular localization (Puntervoll *et al*., 2003). SLiMs are typically less than 10 amino acids long and have less than five defined positions, many of which will be 'degenerate' and incorporate some degree of flexibility in terms of the amino acid at that position. Their length and degeneracy gives them an evolutionary plasticity which is unavailable to domains meaning that they will often evolve convergently, adding new functionality to proteins (Diella *et al*., 2008). The potential of SLiMs as future therapeutic targets makes their discovery of great interest (Kadaveru *et al*., 2008; Neduva and Russell, 2006b).

Motif discovery algorithms can be broadly divided into two types. Those that look for motifs in *related* sequences, using alignment-based approaches, are generally optimal for discovery of very strong (or long) 'family descriptor' motifs [e.g. MEME (Bailey and Elkan, 1995)] or for improving definitions of known motifs [e.g. GLAM2 (Frith *et al*., 2008)]. SLiMs represent a special challenge, as their short nature means that they can be swamped by stronger signals from evolutionary relationships. More successful SLiM discovery methods (Davey *et al*., 2006; Edwards *et al*., 2007; Neduva and Russell, 2006) are built on the model that explicitly invokes convergent evolution, using over-representation of motifs in *unrelated* proteins that share a common attribute (such as interacting with a common protein binding domain, having a common post-translational modification or localizing to the same sub-cellular location). DILIMOT (Neduva and Russell, 2006a) and SLiMDisc (Davey *et al*., 2006) combine these techniques with heuristic scoring schemes to successfully discover new functional motifs and rediscover known motifs. More recently, SLiMFinder (Edwards *et al*., 2007) built upon the work of DILIMOT and SLiMDisc, introducing improved motif definition and a significance-based motif scoring scheme.

Alignment-free methods can still benefit from the evolutionary signal provided by related proteins by analysing patterns of conservation. DILIMOT can use motif occurrences in other species as part of its scoring scheme (Neduva and Russell, 2006a), while both SLiMDisc and SLiMFinder can make use of a number of conservation metrics to improve confidence in motif predictions (Davey *et al*., 2007). Functional instances of motifs are more highly conserved than non-functional instance of the same motif, a fact that has been used (in association with motif enrichment) to discover candidate functional KEN-box motifs (Michael *et al*., 2008) and to improve the classification of true and false positives occurrences of known SLiMs (Chica *et al*., 2008; Dinkel and Sticht, 2007). These techniques score a motif based on its conservation in homologues of the protein containing the motif, weighting the conservation in homologues based on their divergence from the query protein. In addition to the motif itself, however, there is important context information in the conservation signal of the region surrounding the motif, which is often important for specificity and for the ability of the region to adopt a particular conformation in the

*To whom correspondence should be addressed.

bound state (Stein and Aloy, 2008). Removing residues which seem under conserved relative to the background level of conservation should enrich a dataset for functional residues and consequentially, functional motifs (Valdar, 2002). Furthermore, it is important to consider the structural context of the motif, particularly as SLiMs tend to occur in disordered regions (Fuxreiter *et al*., 2007), which are less conserved and more difficult to align than ordered regions (Perrodou *et al*., 2008). Improvements can also be made in terms of how the evolutionary data are incorporated into the models of SLiM discovery. Computational masking of residues which are unlikely to be of functional significance, prior to analysis, leads to a functional enrichment in the dataset. Many analyses have used the technique of masking regions such as domains, ordered regions, transmembrane helices and regions not commonly involved in binding (Edwards *et al*., 2007; Michael *et al*., 2008; Neduva *et al*., 2005). Masking aims to increase the ratio of signal to noise in the dataset by decreasing the number of positions where a motif can occur by chance, which in turn decreases the likelihood of seeing a given motif multiple times, making motifs of interest more easily identifiable. While raw conservation scores are hard to incorporate into over-representation statistics, masking of residues that are less conserved than their local context allows existing statistical models to be used.

Here, we describe a method for the masking of locally under-conserved residues and its capacity to improve the ability of SLiMFinder (Edwards *et al*., 2007) to return known functional motifs from benchmarking datasets. We test the method on a biologically relevant benchmarking dataset, consisting of interaction datasets where a subset of the proteins interact with the hub protein through a known SLiM, and demonstrate a marked improvement in SLiM discovery while retaining the high levels of precision of SLiMFinder.

2 METHODS

2.1 Relative local conservation

The relative local conservation (RLC) for a residue is calculated based on a multiple sequence alignment of orthologous proteins. In general, alignment of disordered regions—where SLiMs are generally found (Fuxreiter *et al*., 2007)—is not as good as for globular domains, especially in highly divergent proteins (Perrodou *et al*., 2008). RLC mitigates this problem in two ways. First, because the method is based upon *relative* conservation, it is only necessary for residues to be more conserved than their surrounding residues. Most of the information for masking comes therefore from columns which lack conservation in closely related sequences and have high levels of randomness indicating no functional constraint. Second, residues are first defined as being 'ordered' or 'disordered' and conservation is only compared across residues within the same disorder class.

The conservation score C_i for each column i of the alignment can be calculated by any suitable scoring metric with high values for conserved residues and low values for more variable residues (see below). This is then used to calculate the mean background conservation score b_i across a window surrounding the residue i [Equation (1)].

$$b_i = \frac{1}{2N+1} \sum_{j=i-N}^{i+N} C_j \tag{1}$$

where N is number of residues either side of residue i that together with i make the window for the background conservation comparison. The mean conservation score for the window should be calculated considering the modularity of proteins and only consider residues of the same state of order. Any residue within the window which is predicted to be in a different state from the residue i is not considered [and the denominator $(2N+1)$ is reduced accordingly]. The continuity and size of regions of order and disorder is not considered and the background conservation window may span several such regions as a result.

Raw conservation scores are converted to the RLC score by subtracting the background mean conservation across the appropriate window and normalizing by dividing by the SD of the window [Equation (2)]. Values above 0 are more conserved than average and below 0 are less conserved than average and likely to be less constrained. As all scores are relative to the mean, the distribution of conservation is approximately evenly distributed around the mean of 0. Normalizing the score to SDs permits the comparison of residues in different proteins, which may have different divergence patterns and thus different magnitudes of deviation from the mean. (If a protein has no identifiable orthologues, all residues have an RLC score of 0.)

$$\mathrm{RLC}_i = \frac{C_i - b_i}{\sigma_i} \tag{2}$$

where σ_i is the SD of the RLC values across the same residues used to generate the background score b_i.

2.2 RLC masking

Any residues with an RLC value below the chosen threshold are masked out. Human Epsin 2 (EPN2, UniProt O95208) provides a good example of how the masking works as it contains a SLiM-rich region involved in binding to AP-2 complex subunit alpha-1 (AP2A1, UniProt O95782) and the clathrin heavy chain (CLTC UniProt Q00610) (Supplementary Fig. 1). This region contains six SLiMs thought to be involved in AP-2 binding: a clathrin box motif (336–340), four DPW AP-2 binding motifs (353–355, 378–380, 392–394 and 410–412) and an FxxFxxL motif that stabilizes AP-2/clathrin binding (459–465) (Praefcke *et al*., 2004). These motifs can be clearly seen as peaks in RLC compared with their neighbouring residues, many of which would be masked out with an RLC < 0. Three other interesting RLC peaks of note in this region are the PW at 376–377, the DAW at 428–430 and the ELF at 446–448, which are all functional DP[FW] AP-2 binding motifs in the closely related protein, Epsin 1 (Praefcke *et al*., 2004). (Epsin-2 has a second FxxFxxL motif at 445–451, which we speculate could have evolved to stabilize the interaction in place of, or leading to, the two lost DPW motifs).

2.3 Residue conservation methods

Several residue scoring methods for alignments of orthologues were tested [see (Capra and Singh, 2007) for review and details]: Shannon entropy (Cover and Thomas, 1991), Shannon entropy of residue properties (Mirny and Shakhnovich, 1999), von Neumann entropy (VE) (Caffrey *et al*., 2004), Relative Entropy (Cover and Thomas, 1991) and Jensen–Shannon divergence score (Lin, 1991). These methods consider each column of the alignment independently. Code for calculating these scores was obtained from the supplementary material of Capra and Singh (2007).

2.4 Disorder prediction

Residue disorder was predicted using IUPRED (Dosztanyi *et al*., 2005). Residues with a disorder score under 0.2 were defined as 'ordered', while all others were defined as 'disordered'. Whereas the default threshold of 0.5 correctly identifies 95% of ordered residues in the DisProt database (Sickmeier *et al*., 2007), but has a tendency to incorrectly predict disordered residues as ordered, a threshold of 0.2 correctly identifies 95% of disordered residues (data not shown), which is more appropriate for this application. There was no minimum length for ordered or disordered regions in this analysis.

2.5 Benchmarking datasets

Three benchmarking datasets were used in this study. For optimizing the RLC method, the collection of validated SLiMs from the ELM database

(Puntervoll *et al*., 2003) was used. All residues in these proteins were then defined as 'ELM residues', if they corresponded to a defined position (fixed or degenerate but not wildcard) in an annotated ELM occurrence, or 'Non-ELM residues' if they did not. Any proteins with less than two orthologues identified by GOPHER (below) were excluded from the analysis (i.e. the proteins were removed from the dataset). This provided 633 occurrences of 49 motifs in 297 proteins. A second ELM benchmarking dataset introduced by Neduva *et al*. (2005) was used to analyse the effects of RLC masking on SLiM discovery. This dataset consists of sets of proteins for 16 different ELMs, where each protein contains a known occurrence of the ELM and at least three proteins in each set have no detectable homology. This dataset has previously been used for benchmarking DILIMOT, SLiMDisc and SLiMFinder. The final dataset consists of data from the Human Protein Reference Database (HPRD) (Mishra *et al*., 2006) (release 6, January 2007). HPRD provides human protein–protein interactions, many of which are known to be mediated by SLiMs in one of the interacting proteins. We have created a small benchmarking set of 20 hub proteins, identified by ELM as interacting through SLiMs, as a biological realistic dataset that is directly representative of a typical analysis for which these SLiM discovery methods are designed. Datasets are available at http://bioware.ucd.ie/∼slimdisc/slimfinder/conmasking/.

2.6 Orthologous alignments for human proteins

Multiple sequence alignments of metazoan orthologues were constructed for each protein in the benchmarking datasets using the GOPHER algorithm (Davey *et al*., 2007). Homologues for each sequence were identified using a BLAST (Altschul *et al*., 1997) search against a database of EnsEMBL (Birney *et al*., 2006) proteomes and orthologues predicted using default GOPHER parameters. Multiple sequence alignments were generated with MUSCLE (Edgar, 2004). Alignments are available at http://bioware.ucd.ie/∼slimdisc/slimfinder/conmasking/.

2.7 SLiMFinder analysis

Datasets were analysed using SLiMFinder version 3.0 with default settings unless otherwise stated; to be returned, motifs must therefore occur in at least three proteins that do not share BLAST-detectable homology. RLC masking was performed using VE (Caffrey *et al*., 2004) and a flanking window size, N, of 30 amino acids. Residues that did not have a RLC score greater than zero were masked out. Any proteins with less than two orthologues identified by GOPHER were not subject to RLC masking (i.e. the proteins were still in the dataset but none of their residues were masked by the RLC algorithm). All P-values mentioned refer to SLiMFinder 'Sig' values (Edwards *et al*., 2007), which estimates the probability of that dataset returning any motifs that are over-represented to the same (or greater) extent, as defined by the probability of the observed support (number of unrelated proteins) for the motif given the composition of the dataset. These values are corrected for the size of the 'motif space' searched, accounting for the number of defined positions and possible wildcard spacers between defined residues.

3 RESULTS

3.1 Optimizing RLC parameters

The RLC calculation can be used with any column-based conservation scoring metric. Capra and Singh (2007) have recently reviewed several of these and so we sought to determine which was most appropriate for use in SLiM discovery by assessing their performance using known ELM occurrences. Performance was assessed in terms of the proportion of ELM residues that were successfully recalled using an RLC score greater than zero, i.e. residues marked as more conserved than expected given their local context. Selection of the window length N can be seen as a balance between maintaining the focus on the local region of interest and including enough residues to overcome stochastic fluctuations in individual residue scores. Window lengths between 5 and 75 were examined and, for all scoring schemes, the recall climbed steadily from a window length of 5 to 15, was approximately equivalent between 15 and 30, and then decreased at a steady rate for $N > 30$ (Supplementary Fig. 2A). This non-uniform distribution of the recall for functional residues for different window lengths advocates the use of local conservation in these situations. In all cases, VE (Caffrey *et al*., 2004) performed best, although recall was high for all five methods (Supplementary Fig. 2B). Very similar results were obtained if the ELMs were divided into types (Ligand/Modification/Targeting) (data not shown). VE was therefore selected for all further analyses.

3.2 Optimizing the RLC threshold for masking

The distribution of RLC scores for all disordered residues (IUPred score of >0.2) in all ELM-containing proteins in the ELM database were calculated (Fig. 1). Residues were defined as 'ELM' or 'Non-ELM' and the RLC distribution of the two groups was compared. The two datasets had significantly different distributions (Welch two-sided t-test $P = 2.2 \times 10^{-16}$). Because the RLC values are all relative to mean conservation, the expected mean RLC score in the absence of any bias is inherently zero. The mean RLC for non-ELM residues was 0.008, which is very close to this theoretical expectation, compared with means for ELM residues of 0.588 for ambiguous positions and 0.788 for fixed positions. For non-ELM residues, the split of residue scores around the (random expectation) mean of 0 was 48.4%/51.6%, while for ambiguous ELM positions the split was 21.5%/78.5% (142/518) and for fixed ELM positions it was 12.5%/87.5% (96/671). ELM residues therefore show a marked increase in RLC scores over their potentially non-functional background.

To analyse the RLC enrichment of ELM against non-ELM residues in more detail, the empirical cumulative RLC distributions of the two groups were compared (Fig. 1B). The optimal RLC threshold for masking should maximize the number of non-functional residues masked, while minimizing the number of known functional residues masked. Figure 1B shows the level of enrichment that is available at each value of the conservation score (where enrichment is the proportion of potential false positives masked minus the proportion of true positives removed). At RLC$\geq$0 we get the maximum enrichment of 31.7% by masking 16.7% of the ELM residues (both ambiguous and fixed) and 48.4% of the non-ELM residues. Depending on the desired outcome in terms of enrichment versus recall, other RLC thresholds may also be appropriate. For example, at RLC≥ -1 only 4.1% of the functional residues are masked, while four times as many of potential false positive residues are removed (16.4%), although the level of enrichment is not as high (12.3%). A threshold of zero, however, has the additional benefit of being intuitively easy to understand (anything less conserved than the mean of its local context is masked out) and independent of normalization.

Interestingly, an RLC threshold of zero differentially masks ELM residues of different motif types (ligand, modification or targeting), with residues in modification motifs having significantly lower RLC than either targeting or ligand binding motifs (Supplementary Table 1). Modification sites tend to be more degenerate than

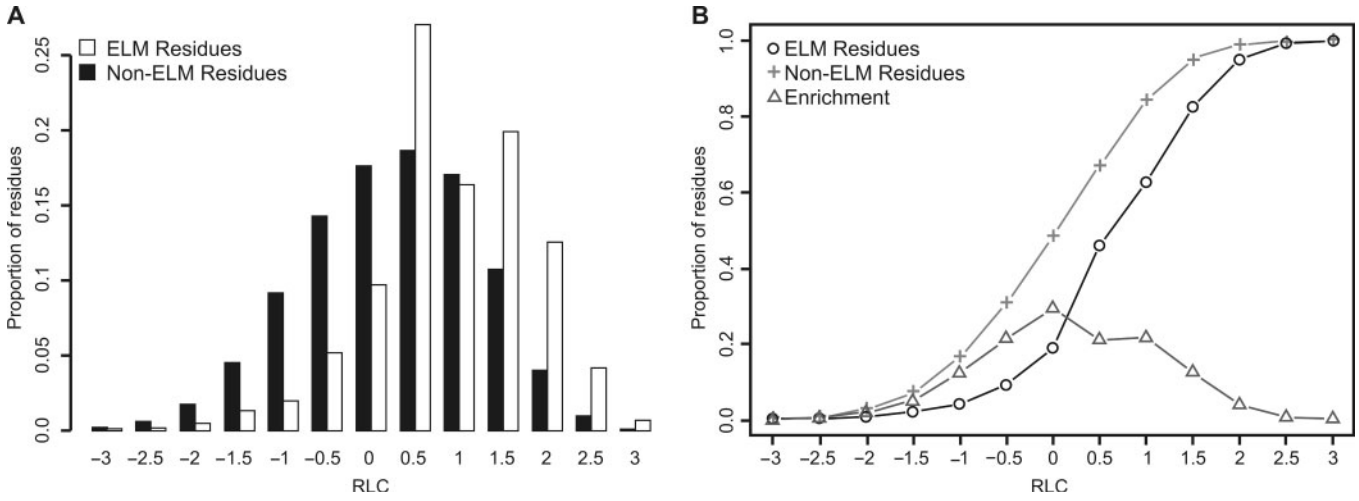

Fig. 1. Comparison of RLC for ELM and non-ELM residues. (**A**) Binned distribution of RLC scores for annotated ELM residue (white) and for non-ELM residues (black). (**B**) Cumulative distribution of RLC scores for annotated ELM residues (circles), cumulative distribution of RLC scores for non-ELM residues (plus signs) and the enrichment (difference between the proportion of ELM and non-ELM residues) available at each RLC (triangles).

other SLiMs and often proteins have several such sites working cooperatively to regulate activity. This combination of attributes may serve to increase plasticity by increasing the likelihood of new sites evolving convergently by chance, while simultaneously reducing the potential impact of losing any given motif occurrence. Unsurprisingly, the conservation of ambiguous positions is weaker than of fixed positions and yet, with the exception of modification motifs, it still provides a huge enrichment over the ∼50% of non-ELM residues masked. For 22% of the 633 motifs in the test set at least one of the residues will be masked using the default settings. However, only 9% of the motifs will have more than one residue masked. In general, these masked residues will be confined to the ambiguous positions and to similar non-critical positions in each motif, which should minimize the impact on motif discovery tools.

3.3 Alignment quality and disorder prediction

Alignment-based conservation methods obviously rely on the generation of a multiple sequence alignment, which is more difficult to do accurately for SLiMs in disordered regions than globular domains (Perrodou *et al*., 2008). By considering the relative conservation of residues in their local context, RLC remains reasonably robust to overall alignment quality, as important residues will still tend to be more conserved—and better aligned as a result—even in poorly aligned regions (Supplementary Fig. 3). Indeed, when overall conservation is very high, RLC is not so good at discriminating ELMs from non-ELMs, because the non-ELM residues themselves tend to get higher RLC scores (Supplementary Fig. 3).

These effects of alignment quality are reflected by the observation that, while RLC is reasonably robust to the IUPred threshold used for disorder prediction, it actually achieves a better recall of ELM residues in disordered regions, irrespective of the threshold used (Supplementary Fig. 4A and B). Indeed, there is even a tendency for the discriminatory power of RLC masking to improve as the IUPred threshold (i.e. confidence of disorder) is increased. This needs to be balanced, however, against the loss of ELM residues by masking disorder using a higher IUPred threshold; when disorder and RLC masking are considered together, the default IUPred threshold of 0.2 chosen for SLiMFinder masking (Edwards *et al*., 2007) gives a good compromise that maximizes the benefit of masking, while avoiding the elimination of too many residues from analysis (Supplementary Fig. 4C). Improvements in both alignment quality and disorder prediction would obviously be expected to improve RLC masking, but these are complex issues beyond the scope of this particular study.

3.4 Benchmarking

The ELM benchmarking dataset introduced by Neduva *et al*. (2005) contains 16 datasets of proteins where each protein contains a known ELM and at least three proteins have no other detectable homology. This dataset, although not biologically realistic, provides a solid platform for comparison of methods and allows the limitations of these methods to be shown. We compared the use of conservation masking with a RLC cut-off of 0 to the results of the same dataset without conservation masking (Table 1). Both datasets also have ordered regions masked using IUPRED with a cut-off of 0.2. For each of the 16 datasets the top ranking motif that matched the relevant ELM was recorded. When RLC masking is used, 13 out of the 16 ELM datasets returned (a variant of) the motif of interest as the highest ranking motif, 12 of which were significant ($P < 0.01$). Two other datasets returned a significant motif ($P < 0.05$) that, while not matching the ELM used to generate the data, matched a different known ELM: the LIG_NRBOX dataset returned a (MOD_SUMO) sumoylation motif and the LIG_14_3_3 dataset returned a MAPK binding site (MOD_ProDKin_1) that also matches the core SH3 recognition sequence PxxP. These results are consistent with known biology: the LIG_NRBOX and MOD_SUMO are both enriched in nuclear proteins, while phosphorylation is important

Table 1. Effects of RLC masking on the ELM benchmarking dataset

ELM[a]	N[b]	True motif[a]	RLC[c]	Sig[c]	No RLC[d]	Sig[d]
TRG_ER_KDEL_1	12	[KRHQSAP][DENQT]EL$	**DEL$**	**0.00e+00**	**K.{0,2}DEL$**	**4.26e-26**
LIG_CtBP	26	P.[DEN]L[VAST]	**P[ILM]DL**	**2.61e-09**	**P[ILM]DL**	**0.009**
LIG_PCNA	13	Q..[ILM]..[FHM][FHM]	**Q..[IL].SFF**	**4.93e-07**	**[IL].S[FH]F**	**1.62e-6**
LIG_SH3_2	9	P..P.[KR]	**P..P.R.{0,1}P**	**7.42e-06**	**P..P.R.{0,1}P**	**0.002**
LIG_RGD	15	RGD	**RGD**	**1.11e-05**	R.D[ST]V	1 (5)
LIG_Dynein_DLC8_1	4	[KR].TQT	**TQT**	**4.34e-05**	**S..K.TQT**	**0.003**
LIG_CYCLIN_1	22	[RK].L.{0,1}[FYLIVMP]	**RR.{0,1}L.{0,1}F**	**5.34e-04**	**RR.{0,1}L.{0,1}F**	**0.002**
LIG_Clathr_ClatBox_1	15	L[ILM].[ILMF][DE]	**L.D.{0,1}L**	**6.95e-04**	L.{1,2}DL.{0,2}D	1 (22)
LIG_AP_GAE_1	8	[DE][DES].F.[DE][LVIMFD]	**F.DF.S**	**0.001**	D.F..F.S.P[f]	0.097
LIG_RB	25	[LI].C.[DE]	**L.C.[DE]**	**0.002**	L.C.E	1 (11)
MOD_SUMO	29	[VILAFP]K.[EDNGP]	**[LV].[IV]K.E**	**0.003**	**[FIV]K.E**	**1.09e-5**
LIG_14-3-3_1	4	R[SFYW].S.P	**RS.S.P**	**0.005**	RS.S.P	0.53 (3)
LIG_NRBOX	9	L..LL	***IK.E..D***[e]	**0.010**		
LIG_14-3-3_3	6	[RHK][STALV].[ST].[PESRDIF]	***PP.TP..R***[e]	**0.011**	S.P.S.T.P	1 (5)
TRG_LysEnd_APsAcLL_1	10	[DER]...L[LVI]	K..L[LV][f]	0.56		
MOD_N-GLC_2	5	N[^P][ST]				

[a]The ELM identifier and regular expression definition.
[b]The number of proteins in the dataset.
[c]The top ranking motif (and significance) that matches a known ELM when RLC masking is used. Significant motifs ($P < 0.05$) are shown in bold.
[d]The top ranking motif (and significance) that matches a known ELM without RLC masking. All motifs are the top ranked motif returned by SLiMFinder unless otherwise indicated by a number in brackets after the significance. Significant motifs ($P < 0.05$) are shown in bold.
[e]Signifies that the top ranking significant motif matches a true functional motif but not the motif used to create the dataset.
[f]Signifies that the top ranking motif returned matches the true motif for the dataset but has not reached the significance cut-off.

in 14-3-3 mediated signal transduction (Puntervoll *et al*., 2003). None of the datasets, with or without RLC masking, returned any significant motifs that did not match a variant of a known ELM in Table 1.

RLC masking returns five more correct significant motifs than the disorder masking alone (12 compared with 7) and 8 more true motifs as the top ranking motif (15 compared with 7). This is a significant ($P = 0.003$, Fisher's exact test) increase in the recall of the method. Six out of the seven motifs returned using disorder masking alone have their significance increased by RLC masking. Neither the TRG_LysEnd_APsAcLL_1 nor the MOD_N-GLC_2 dataset returned a significant motif. The TRG_LysEnd_APsAcLL_1 returns a variant of the true binding motif found in 3 of the 10 functional instances in the dataset as the top ranking motif, but this was not significant ($P = 0.56$). MOD_N-GLC_2 is short and therefore has a high likelihood of occurring by chance, making it difficult for the dataset to have enough statistical power. More importantly, none of the instances are conserved, perhaps due to the fact that multiple instances of the motif occur in several proteins in the dataset possibly easing the constraint on any one instance/occurrence of the motif. Neither of these datasets returned a significant motif with disorder masking alone.

3.5 Interaction datasets

The concept of over-representation as an indicator of functionality is easily applied to interaction datasets. Indeed, DILIMOT was able to successfully identify new binding motifs from interaction data (Neduva *et al*., 2005), while SLiMFinder has been demonstrated to return known motifs from 14-3-3 interactors and Hepatitis B virus phage display results (Edwards *et al*., 2007). HPRD (Mishra *et al*., 2006) provides a dataset of human protein–protein interactions, many of which are known to be mediated by SLiMs in one of the interacting proteins, and we have created a small benchmarking set of 20 hub proteins from these SLiM-mediated interaction datasets as a realistic benchmarking dataset that is representative of typical analyses for which SLiM discovery methods are designed. Each dataset contains all known HPRD interactors for a given hub protein, a subset of which are known to be mediated by a SLiM.

These datasets were analysed using SLiMFinder with an IUPred disorder cut-off of 0.2. RLC masking (VE method, $N = 30$) gives a dramatic improvement both in terms of the number and significance of SLiMs returned from these data (Table 2). Without RLC masking, the correct motif is returned for only four of the datasets. When RLC masking is used, each of these is returned with increased significance, along with an additional eight datasets. This represents a significant increase in recall ($P = 0.002$, Fisher's exact test), which is especially encouraging given the biological relevance of the datasets. An additional three motifs are returned with non-significant P-values ($0.05 < P < 0.99$), which implies that improving the quality of these datasets slightly might be sufficient for them to also recall the expected motif. (HPRD interactions include some indirect interactions that are mediated through a shared interaction partner, e.g. in a protein complex.)

Five of the 20 datasets returned unexpected significant motifs. Four of these match other known ELMs, however. Datasets for PCNA, RB1 and UBE21 all returned lysine/arginine-rich motifs, which are extremely common as they have been linked to several functions (protein cleavage, localization and post-translational modification) making them enriched in many datasets. As these three hub proteins are nuclear proteins, this motif is most likely over-represented due to the presence of functional nuclear localization (KRK) motifs in several of their interactors. (PCNA and UBE2I also returned the motif of interest for the dataset as a significant motif.)

Table 2. Results for the 20 protein benchmarking dataset from HPRD where a subset of the interactors for a hub protein are known to be SLiM mediated

HPRD[a]	Symbol[a]	Hub[a]	ELM[b]	ELM regular expression[b]	RLC[c]	S/N[d]	No RLC[c]
00150	GRB2	Grb2	LIG_SH3	P..P	P.P..P****	34/164	S.{1,2}Y.N*
			LIG_SH2_GRB2	Y.N	YEN**	9/164	
00215	YWHAH	14-3-3 Eta	LIG_14-3-3_1	R[SFYW].S.P	[KR]S.S.P**	9/47	
00350	CLTC	Clathrin, heavy polypeptide	LIG_Clathr_ClatBox_1	L[ILM].[ILMF][DE]	LLDL****	8/35	DL[LM]D**
00453	CCNA2	Cyclin A2	LIG_CYCLIN_1	[RK].L.{0,1}[FYLIVMP]			
00607	FNTA	Farnesyltransferase α subunit	MOD_ASX_betaOH_EGF	C[^DENQ][LIVM]..$	C..S$***	3/10	
00627	ITGA5	Integrin alpha 5	LIG_RGD	RGD			
01456	PCNA	Proliferating cell nuclear antigen	LIG_PCNA	Q..[ILM]..[FHM][FHM]	[IL]..FF****	12/65	
01574	RB1	Retinoblastoma 1	LIG_RB	[LI].C.[DE]			
03288	PPARG	Peroxisome proliferator AR Υ	LIG_NRBOX	L..LL	L.RLL	5/22	
03334	DYNLL1	Dynein light chain 1	LIG_Dynein_DLC8_1	[KR].TQT	[KR].TQ	10/52	
03786	NEDD4	NEDD4	LIG_WW_1	PP.Y	PP.Y****	10/28	
03822	TRAF6	TRAF6	LIG_TRAF6	.P.E..[FYWHDE].			
04015	CTBP1	C-terminal binding protein	LIG_CtBP	P.[DEN]L[VAST]	P[IL]D.S**	6/26	
04946	CCNA1	Cyclin A1	LIG_CYCLIN_1	[RK].L.{0,1}[FYLIVMP]			
05462	GIPC1	GIPC1	LIG_PDZ_1	.[ST].[VIL]$	[DE]S.V$****	4/26	S.V$****
05639	YWHAG	14-3-3 gamma	LIG_14-3-3_1	R[SFYW].S.P	R.RS.S..P****	9/206	
08968	EPS15	Eps15	LIG_EH	NPF	TNPF****	6/24	
09045	UBE2I	Ubiquitin conjugating enzy. E2I	MOD_SUMO	[VILAFP]K.[EDNGP]	VK.E****	26/87	VK.E*
09347	GGA2	GGA2	LIG_AP_GAE_1	[DE][DES].F.[DE][LVIMFD]	DDF..F..A****	3/19	
09424	YAP1	Yes associated protein	LIG_WW_1	PP.Y	PP.Y	4/15	

[a]HPRD ID, HGNC gene symbol and description for the hub protein.
[b]The ELM expected to be enriched for the dataset. In some cases the ELM regular expression has been simplified for space reasons.
[c]The top ranking motifs matching the ELM consensus for the dataset ($P < 0.99$).
[d]S is the motif's support (the number of proteins a motif occurs in) in the dataset and N is the number of proteins in the dataset.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$.

As was also observed for the LIG_14-3-3_3 ELM benchmarking dataset, the dataset for 14-3-3η (YWHAH) returned a proline-rich motif (P..[ST]P.P), which resembles a putative MAP kinase binding site. Again, YWHAH also returned the anticipated archetype 14-3-3 binding motif. The remaining motif of the five, the unknown significant motif PRGxxL, was returned as the fourth-ranked motif from the GIPC1 dataset, along with the expected PDZ ligand motif. This motif exhibits lower level of conservation than the known true positives (data not shown), and may therefore represent a false positive. Alternatively, it could represent a possible new discovery.

4 DISCUSSION

Neduva *et al*. (2005) clearly demonstrated the potential of models based on convergent evolution when they applied DILIMOT to discover SLiMs in multiple HPRD datasets and were able to verify two of their predictions with direct-binding assays. However, DILIMOT does not specify results in terms of statistical significance, making the interpretation of large-scale analyses difficult. SLiMFinder (Edwards *et al*., 2007) was able to improve predictions by refining motif definitions with the incorporation of amino acid ambiguities and, most importantly, constructing a statistical framework that was able to return known motifs with high precision and improve confidence in results. This precision came at a cost, however, and many datasets failed to yield any results at all due to the poor signal-to-noise qualities of the data. The RLC masking method goes a long way to overcoming this problem by reducing the noise present in the dataset without severely impacting on the signal, which in turn significantly improves the re-discovery of known motifs from biologically meaningful datasets. Future work will apply this technique on a larger scale in an attempt to make new discoveries, but the potential looks very promising indeed.

The statistical calculations of SLiMChance assume independence between sequences (Edwards *et al*., 2007). For evolutionarily related sequences, this is obviously not true and SLiMFinder attempts to compensate for this by grouping such sequences together into a single cluster, which is scaled in size according to the degree of relatedness using the minimum spanning tree (MST) method of SLiMDisc (Davey *et al*., 2006). This scaling maintains the assumption of independence provided the sequences in a cluster are of similar lengths and/or all share a similar degree of sequence identity with each other. Masking of sequences can complicate this further because more similar sequences are inherently likely to be masked in a similar fashion. This is obviously true for alignment-based methods, such as RLC because closely related sequences may share orthologues and might be expected to have many of the same residues masked. The same also applies, however, for disorder and domain-based masking, which are also dependent on sequence. MST calculations are based on unmasked sequences, which can therefore over- or under-estimate the similarity of the masked sequences if masking is systematically biased toward or away from homologous regions. In reality, this seems to have a reasonably small effect as SLiMFinder runs with RLC masking maintain very high stringency (low false positive) levels in benchmarking, and return true positives with greater sensitivity. If deviations from independence are a concern, however, a similar strategy to Neduva *et al*. could be employed, removing all but one related sequence from the dataset prior to SLiMFinder analysis.

Multiple sequence alignment residue scoring schemes can calculate residue conservation in several ways. The level of constraint on a position can be calculated defining the residues that are generally important for the protein (Capra and Singh, 2007) or the conservation of the residues from a particular query protein compared with its homologues can be used. Also, phylogenetic trees can be used to weight the contribution of residues based on the divergence of the proteins (Chica *et al*., 2008). Other scoring schemes to quantify residue conservation may also be worth investigating. The fact that all five methods tested in this article did very well in terms of ELM recall demonstrates that the most important aspect of RLC masking is the innovative use of local relative conservation, rather than the specific choice of residues conservation score. At the same time, the fact that VE outperformed the other scores while still not achieving 100% recall implies that there is still room for improvement. Nevertheless, the potential for improvement here seems marginal next to the improvements that can be made in terms of interaction dataset coverage and quality, which is likely to have a much more significant impact on the quality of the results. (Of course, we must also consider that there will probably be a subset of true functional motifs that, for various reasons including recent convergent evolution, do not show an enrichment in terms of conservation and will therefore need alternative discovery methods.)

RLC masking should not be considered as solely a tool to improve novel SLiM discovery. Enrichment-based methods searching for novel instances of previously known motifs, such as the KEN box analysis previously mentioned (Michael *et al*., 2008), could also benefit from its use. Indeed, motif conservation has proved an appropriate method for the classification of instances of known SLiMs as true or false positives for single proteins (Chica *et al*., 2008; Dinkel and Sticht, 2007). Whether the best approach in such cases is to mask residues or simply record their conservation scores for ranking purposes remains to be seen and is likely to depend on the goal of a given analysis. The rejection of motifs as false positives based on RLC masking might encounter problems at the threshold of RLC < 0 as we know that we have removed around 16% of true positive residues, which corresponds to losing over 1/5 instances of ELMs (Supplementary Table 1). However, as demonstrated, using the normalized RLC score and adjusting the threshold to remove only the least conserved residues can raise the recall of ELM residues to over 95%, while still removing nearly 1/5 of all non-ELM residues. Furthermore, if the goal is to determine whether a given known motif is enriched in a given protein dataset, the same benefits of RLC masking to SLiM discovery will apply, removing false positives, shrinking of the motif search space and boosting motif significance. This could allow for discovery of enriched motifs in datasets which previously returned no enrichment due to a poor signal-to-noise ratio. To allow for either type of analysis, we have added RLC scoring to SLiMSearch, a tool for searching a sequence database with a known SLiM. SLiMSearch (http://bioinformatics.ucd.ie/shields/software/slimsearch/) incorporates the same masking options and evolutionarily weighted enrichment calculations as SLiMFinder, enabling ease of further analysis using the same settings as for SLiM discovery. Alternatively, SLiMSearch (and SLiMFinder) can simply return the RLC score for motif instances without any masking, allowing the information to be incorporated into a ranking scheme.

SLiM discovery is like looking for a needle in a haystack, but with intelligent dataset design the task becomes significantly easier. This has been shown previously in analyses which have returned known functional SLiMs from noisy datasets where only small subsets of the proteins in the dataset contain the SLiM (Davey *et al*., 2006; Edwards *et al*., 2007; Neduva *et al*., 2005). Masking has been instrumental to the rediscovery of these motifs and to the discovery of novel motifs. The analysis of motif conservation can also aid considerably in the classification of over-represented SLiMs allowing users to quickly define which instances of a motif are of interest for further analysis. Here, we have shown that 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. This enrichment increases the ability of current methods to return functional motifs and analysis of benchmarking datasets using this masking scheme has shown a significant improvement in recall of known functional motifs compared with current masking schemes. This opens the door for systematic SLiM discovery to identify discrete points of protein–protein interaction that may be particularly suitable for drug targeting.

Furthermore, the magnitude of the improvement indicates that RLC masking has the potential to positively impact on computational predictions of protein function beyond the field of SLiM discovery. While the methodology developed is focused specifically on the identification of recurring motifs within disordered regions, it is clear that the approach is likely to prove very successful in identifying other functional features of disordered protein regions. Examples of these include single conserved residues, non-recurring binding sites and motifs that are longer than the typical SLiM length of 2–11 residues for which the current discovery methods are optimized. Evaluating the function of disordered regions is a field that is relatively new, in contrast to the study of function in ordered regions of proteins. Nonetheless, it is clear that disordered regions play a crucial role in biology (Dunker and Obradovic, 2001; Dunker *et al*., 2005; Tompa, 2005) and provide excellent targets for future therapies (Russell and Gibson, 2008). We anticipate that the systematic application of relative conservation analysis employing the methodology outlined in this article will result in significant advances in our understanding of the functional role of disordered regions of proteins, not only for convergently evolved recurrent motifs, but also for many other functional features of disordered regions.

All data and tools in this article are available at http://bioware.ucd.ie/~slimdisc/slimfinder/conmasking/.

ACKNOWLEDGEMENTS

The authors would like to thank Claudia Chica for useful discussions during the early stages of this project.

Funding: Science Foundation Ireland.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.

Birney,E. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.

Caffrey,D.R. *et al*. (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*., **13**, 190–202.

Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Chica,C. *et al*. (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, **9**, 229.

Cover,T. and Thomas,J. (1991) *Elements of Information Theory*. John Wiley and Sons, New York, USA.

Davey,N.E. *et al*. (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res*., **34**, 3546–3554.

Davey,N.E. *et al*. (2007) The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res.*, **35**, W455–W459.

Diella,F. *et al*. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci*., **13**, 6580–6603.

Dinkel,H. and Sticht,H. (2007) A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics*, **23**, 3297–3303.

Dosztanyi,Z. *et al*. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Dunker,A.K. and Obradovic,Z. (2001) The protein trinity–linking function and disorder. *Nat. Biotechnol.*, **19**, 805–806.

Dunker,A.K. *et al*. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.*, **272**, 5129–5148.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.

Edwards,R.J. *et al*. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, **2**, e967.

Frith,M.C. *et al*. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol*., **4**, e1000071.

Fuxreiter,M. *et al*. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.

Kadaveru,K. *et al*. (2008) Viral infection and human disease–insights from minimotifs. *Front Biosci*., **13**, 6455–6471.

Lin,J. (1991) Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, **37**, 145–151.

Michael,S. *et al*. (2008) Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. *Bioinformatics*, **24**, 453–457.

Mirny,L.A. and Shakhnovich,E.I. (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.*, **291**, 177–196.

Mishra,G.R. *et al*. (2006) Human protein reference database–2006 update. *Nucleic Acids Res*., **34**, D411–D414.

Neduva,V. and Russell,R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett*., **579**, 3342–3345.

Neduva,V. and Russell,R.B. (2006a) DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res*., **34**, W350–W355.

Neduva,V. and Russell,R.B. (2006b) Peptides mediating interaction networks: new leads at last. *Curr. Opin. Biotechnol.*, **17**, 465–471.

Neduva,V. *et al*. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*., **3**, e405.

Perrodou,E. *et al*. (2008) A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics*, **9**, 213.

Praefcke,G.J. *et al*. (2004) Evolving nature of the AP2 alpha-appendage hub during clathrin-coated vesicle endocytosis. *EMBO J.*, **23**, 4371–4383.

Puntervoll,P. *et al*. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*., **31**, 3625–3630.

Russell,R.B. and Gibson,T.J. (2008) A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS Lett*., **582**, 1271–1275.

Sickmeier,M. *et al*. (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res*., **35**, D786–D793.

Stein,A. and Aloy,P. (2008) Contextual specificity in peptide-mediated protein interactions. *PLoS ONE*, **3**, e2524.

Tompa,P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*., **579**, 3346–3354.

Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.