W350–W355 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl159

DILIMOT: discovery of linear motifs in proteins

Victor Neduva and Robert B. Russell*

EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Received February 14, 2006; Revised March 9, 2006; Accepted March 20, 2006

ABSTRACT

Discovery of protein functional motifs is critical in modern biology. Small segments of 3–10 residues play critical roles in protein interactions, post-translational modifications and trafficking. DILIMOT (DIscovery of LInear MOTifs) is a server for the prediction of these short linear motifs within a set of proteins. Given a set of sequences sharing a common functional feature (e.g. interaction partner or localization) the method finds statistically over-represented motifs likely to be responsible for it. The input sequences are first passed through a set of filters to remove regions unlikely to contain instances of linear motifs. Motifs are then found in the remaining sequence and ranked according to a statistic that measure over-representation and conservation across homologues in related species. The results are displayed via a visual interface for easy perusal. The server is available at http://dilimot.embl.de

INTRODUCTION

The modular nature of proteins is well established. Resources like SMART (1) and Pfam (2) contain thousands of sequence modules, most of which are globular protein domains of 30 or more residues, which cover a variety of functions. These databases have long been complemented by dozens of sequence comparison tools that have enabled the detection of both new domains, and instances of those already known [e.g. (3–6)].

Though very useful, these resources and tools are limited in the kinds of sequence features they cover: specifically they are normally restricted to relatively long stretches of sequence that normally fold into compact globular entities. They are not well suited to study the many hundreds of known short linear segments (usually shorter than 10 residues) known to participate in protein interactions, localization and post-translational modifications throughout many biological processes. Segments performing similar molecular functions often show a conserved sequence pattern, or linear motif that captures the key features important for their function. In contrast to most domains, linear motifs very often reside in disordered or non-globular regions of proteins (7) and are usually only conserved between closely related species (8). This, together with their short length, has to date made them difficult to detect either computationally or experimentally.

Efforts are underway to catalogue linear motifs and particular instances that are known to be functional. These include the eukaryotic linear motif (ELM) resource (7), Phospho.ELM (9), NetPhos (10), Prosite (11) and ScanSite (12). Many of these resources also allow users to find instances of known motifs within proteins of interest. However, tools to uncover new linear motifs are not generally available. Estimates of hundreds of still to be discovered motifs mediating protein interactions (13) make applications to do so very timely. DILIMOT (DIscovery of LInear MOTifs; http://dilimot.embl.de/) is a tool for doing just this. Users with a set of proteins sharing some common attribute can use the server to find candidate motifs likely to be responsible for it. The method is robust and sensitive: it performed well in a benchmark, and was previously able to rediscover dozens of previously known motifs from very noisy high-throughput protein interaction data (13).

METHODOLOGY OVERVIEW

The general principle behind the approach is that proteins with a common function will share a feature that mediates it, either a domain or a linear motif. In the absence of a domain, the linear motif is often the only common sequence feature, and is often detectable, simply by virtue of over-representation (13). The methodology is shown schematically in Figure 1. In the first stage of the procedure, parts of a given set of protein sequences least likely to contain linear motifs are removed, and the set is made non-redundant in terms of sequence similarity. Then all over-represented motifs in the remaining sequences are identified and ranked according to scores that consider the background probability of the motif been found in randomly selected sequences, the number of sequences containing the motif, the size of the set of sequences, and the degree to which instances of the motif are preserved in available orthologous proteins.

Filtering sequences and motif detection

Protein regions known to be depleted for linear motifs, such as globular domains, signal peptides, trans-membrane and

*To whom correspondence should be addressed. Tel: +49 6221 387 8473; Fax: +49 6221 387 8517; Email: russell@embl.de

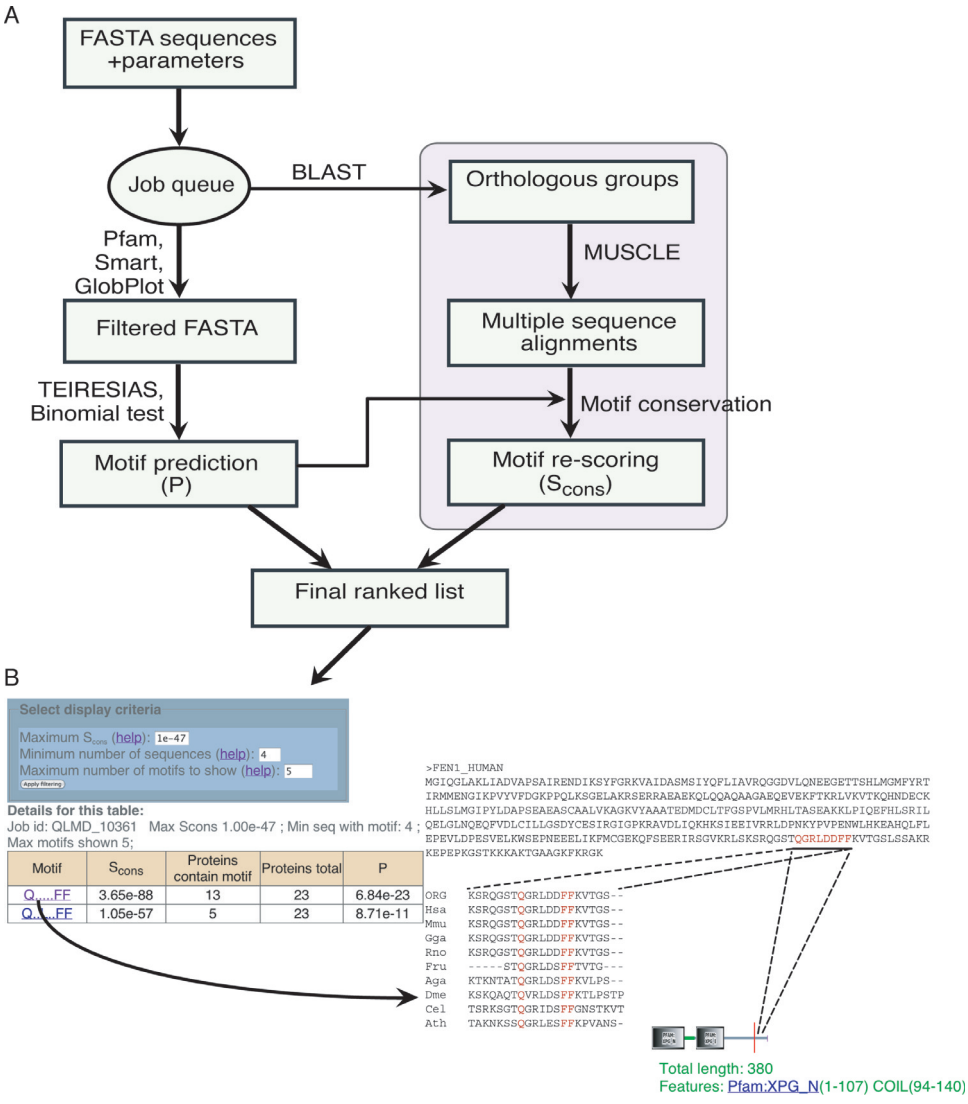


Figure 1. The server process and output. (**A**) Schematic showing how submitted sequences are filtered, motifs found and arranged into a ranked list sorted by P (left). When the species is provided, sequences are assigned to the orthologous groups, species–specific probabilities for over-represented motifs are calculated (coloured box) the list resorted by S_{CONS} (right). (**B**) Example of server output. A list of putative motifs is reported in an interactive table (left), which gives general details for each of them. Clicking on each motif launches an additional page (right) showing sequences containing the motif, where the motif is found in them and the degree to which the motif is conserved in related species. Motif locations (red bars) and other features found in the sequences, such as domains, are shown graphically and detailed below each image.

coiled-coil regions are defined using SMART (1) and Pfam (2) with default parameters. Globular regions not found in these resources can also be predicted using GlobPlot (14). Note that removal of low-complexity sequence regions is not sensible as linear motifs often reside in them. Users can specify the combination of particular features to be removed, and can moreover filter sequences themselves down to particular parts they know to be important for binding (e.g. deduced from deletion experiments). In this way, they can avoid spurious removal of motifs by overzealous automatic filtering, and can greatly improve the sensitivity by removing potential false motifs found in non-interacting parts of the sequence. Each set of sequences is also made non-redundant: just one representative of long sequence similar stretches, as defined by BLAST (6) ($E < 0.001$), is kept in the set. This avoids the discovery of motifs owing to multiple occurrences of long homologous segments that are not defined as domains in SMART or Pfam. Sequences with >97% of their residues filtered are removed from the set. Motifs are then uncovered in the remaining sequences using the TEIRESIAS pattern-matching algorithm (15), with parameters that can be specified by the user. Currently we do not allow for conservative substitutions (R/K, D/E, etc.), though we will provide these once we have tested their performance in a benchmark.

Scoring and ranking motifs

Short protein motifs follow a binomial distribution in randomized protein sequences (data not shown). We exploit this observation to compute the probability (P) of observing a motif a certain number of times in a set of a particular size. This calculation requires a background probability (P) of

finding the motif within similarly filtered, randomly selected sequences, which we compute by selecting random proteins from SWISS-PROT (16). The calculation does not currently explicitly account for protein length: we assume that sets will consist of proteins with a typical length distribution. This can lead to motifs within sets of very long sequences having lower probabilities (i.e. more significant). However the relative ranking of the motifs for a particular set would not change as all are affected equally. We are currently revising the calculations to correct for length.

As true instances of known motifs are often conserved over short evolutionary distances (8), it is also possible to increase sensitivity by combining multiple *P*-values from different species. We combine these as a product of separate *P*-values for all species (13). We refer to this product as S_{CONS}. The server currently contains orthologous groups from some nine complete eukaryotic genomes (Hsa, *Homo sapiens*; Mmu, *Mus musculus*; Rno, *Rattus norwegicus*; Gga, *Gallus gallus*; Fru, *Fugu rubripes*; Aga, *Anopheles gambiae*; Cel. *Caenorhabditis elegans*; Dme, *Drosophila melanogaster*; Ath, *Arabidopsis thaliana*) as defined by STRING (17) and aligned using MUSCLE (18); we will add reliable, completed eukaryotic genomes as they arrive. The user must merely provide the species to which the set of proteins belongs or in principle the closest relative among the complete genomes (e.g. one might select *H.sapiens* if working with Chimpanzee proteins). Motifs are ranked by S_{CONS} or by the single species *P*-value if conservation is not possible to compute, or not selected.

Performing many different queries at the same time creates problems because of multiple-testing: certain values of S_{CONS}/*P*-values will necessarily be smaller when results from many sets are compared. Note, however, that our thresholds, reported below and on the server, were derived under a multiple-testing scenario: many motifs derived from dozens of sets.

SERVER DETAILS

The site comprises a Perl-CGI front end, which calls the various parts of the underlying method and parses the results to produce user-specific display pages. Jobs are assigned an identifier; this or link, which is sent to the users e-mail address can be used to access the results for up to two weeks after the job is completed. Computationally, jobs take a few seconds or minutes to process, though actual response time depends on the number of sequences in the set and on the number of jobs in the queuing system.

Users are required to input at least three sequences in FASTA format, corresponding to the set of proteins in which to search for motif candidates. We have also developed an interface to the STRING (17) interaction server to aid the construction of sequence sets for motif hunting (see below). It is also possible to adjust several parameters to alter the output, including the type of filtering to be performed (see above), the species the sequences come from, the length limits for motif candidates, the number of fixed (i.e. non-'x') positions, and the minimum number of sequences in the set required to contain a reported motif. Results, when ready, are displayed in the current browser window, and a link to the results pages is sent to the users e-mail address if it is supplied.

Results are provided as a table of motifs ranked by the P or S_{CONS} value. The number of results displayed can be changed according to user-defined criteria, such as the number of sequences containing each candidate motif, or the maximum P or S_{CONS} value. When clicked, each motif is displayed individually (Figure 1), showing the location of candidate motifs in the sequence, the location of other features identified by SMART and Pfam, and how well the motif is conserved across orthologues from related species (in the form of a sub-alignment showing only the motif-containing region). Note that additional instances of the motifs within filtered regions are also reported to the user. Although these are not used in the scoring scheme, they might still be functional, perhaps residing in loop regions inside globular domains, or being homologues of other motif instances reported.

EXAMPLE APPLICATIONS OF THE SERVER

In previous work we applied the method on a set of proteins from high-throughput Yeast two-hybrid studies and other sources (13). However, the method can be applied to any other set of proteins with a common attribute. For example, sets of proteins sharing an interaction partner from biochemical experiments (e.g. literature derived interaction data, kinase substrates, etc.), or a common cellular localization [e.g. as determined by high-throughput experiments (19)]. Below we discuss some examples: the diversity of applications highlights both the methods utility and the general importance of short linear motifs in biological systems. Summary details for the motifs mentioned are given in Table 1.

Binding site prediction for end-binding protein 1 (EB1)

EB1 is one of a large family serving pivotal roles in eukaryotic microtubule dynamics. Several proteins that are proposed to interact with EB1 or its Yeast homologues have been identified in human [e.g. adenomatous polyposis coli protein, APC (20); cytoplasmic linker protein, CLIP-170 (21)] and in various species of Yeast [e.g. Karyogamy protein, Kar9 (22); temperature shock-inducible protein, Tip1 (21)] (Figure 2). For this example, there are not sufficient numbers of interaction partners in a single species to apply the method, so we tried combining sequences from different species into a single set. om different species into a single set. Although this prevented the strict use of motif conservation, the method still returned some interesting candidate motifs. The best scoring motif was

Table 1. Identification of linear motifs in various protein sets

Domain→motif	Source	1st correct motif (rank)	S_{CONS}	N/M
EB1→IP	Manual	SxIP(1)	N/A*	7/9
Nuclear localization	LifeBD	KxxKxK(1)	9.4×10^{-34}	9/27
PKB→RxRxx(ST)	Phospho.ELM	RxRxxS(1)	1.7×10^{-64}	17/28
CDK→(ST)Px(KR)	Phospho.ELM	SPxR(2)	8.5×10^{-31}	13/42
PKA→(KR)(KR)x(ST)	Phospho.ELM	RRxS(1)	0.0	36/77
CK-2→(ST)xxE	Phospho.ELM	SDxE(4)	1.5×10^{-65}	19/70

Source indicates where the set of proteins came from (see text). N is the number of proteins in the set containing the motif, M is the number of sequences in the set.

*S_{CONS} could not be calculated for EB1 as the proteins came from different species. The corresponding *P*-value is 2.1×10^{-10}.

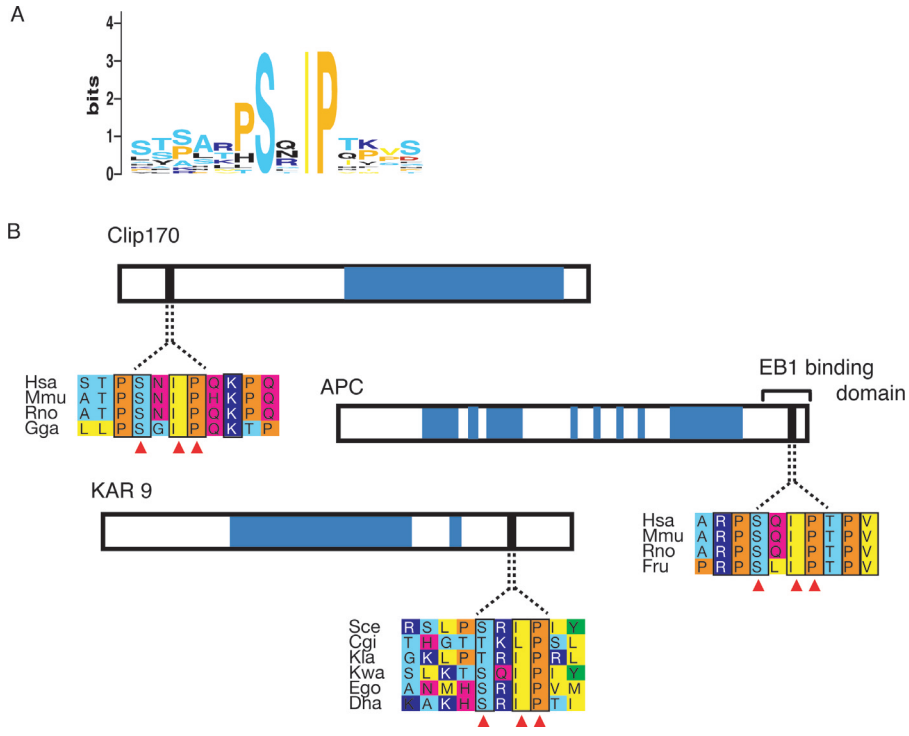


Figure 2. The EB1 motif SxIP detected by the server. (**A**) A sequence logo (27) for the EB1 binding motif, generated using all instances of the motif in the input set. (**B**) Examples of EB1 binding proteins from the input set (represented as boxes) and multiple alignments of putative motif containing regions. Dark blue regions in the boxes denote those removed by the domain and redundancy filters. A known EB1 binding region (in APC) lies at the C-terminus of a Pfam domain. To avoid its removal, we simply cut the sequence down to this region alone (switching the Pfam filter off will have similar effect). Sequences for the motif-containing region are shown aligned to the best homologues in closely related species. Amino acids in the alignments are coloured according to residue type: blue, positive; red, negative; light-blue, small; yellow, hydrophobic; green, aromatic; magenta, polar; Proline, orange. Positions within the predicted motif are denoted by red triangles. Species abbreviations: Hsa, *H.sapiens*; Mmu, *M.musculus*; Rno, *R.norwegicus*; Gga, *G.gallus*; Fru, *F.rubripes*; Cgi, *Candida glabrata*; Kla, *Kluyveromyces lactis*; Kwa, *Kluyveromyces waltii*; Ego, *Eremothecium gossypii*; Sce, *Saccharomyces cerevisiae*; Dha, *Debaryomyces hansenii*.

SxIP, and inspection of others suggested a longer variant P(T/S)x(L/V/I)P (i.e. this was constructed manually by combining different motifs). The best motif is well conserved in orthologues of the human and Yeast interactors. Interestingly, this motif agrees very well with a recently published crystal structure of EB1 bound to a peptide segment and with kinetic assays involving peptides containing the motif (23), which the authors propose to be centred around an IP pattern.

Cellular localization signals

We applied the approach to the detection of potential cellular transit signals by searching for patterns in groups of proteins sharing a common localization as described in LifeDB, a database of mammalian proteins with experimentally verified cellular locations (19). There were 19 localization sets with at least four sequence dissimilar proteins in them (sets of three or smaller only rarely give statistically significant signals). The sets include all major cellular compartments (e.g. nucleolus, peroxisome, etc.). Among the motifs we detected were several typical basic nuclear localization signals (NLS), which are clearly over-represented in nuclear/nucleolar localized proteins (Table 1).

Kinase phosphorylation sites

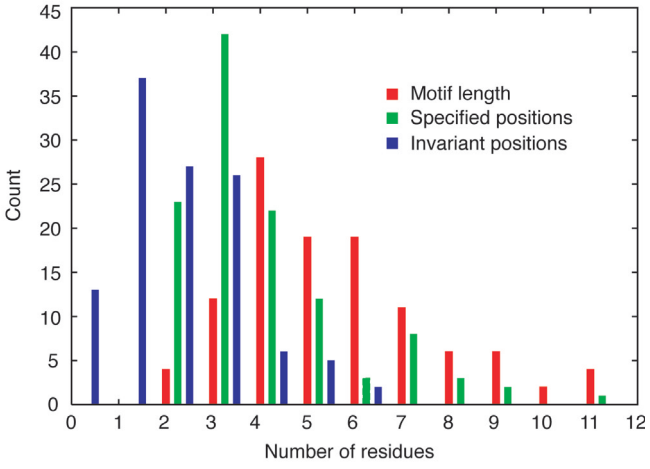
Phosphorylation sites for particular kinases often show very weak patterns, with little more than the phosphorylated Serine, Threonine or Tyrosine common to the substrates. Nevertheless, our approach correctly identifies four phosphorylation sites when presented with the nine datasets specific to kinases taken from the Phospho.ELM resource (9) that had regular expressions with two or more non-wildcard ('x') residues in the ELM database (7). These are detailed in Table 1.

PRACTICAL HINTS FOR FINDING TRUE MOTIFS

Our previous work (8,13) gave us considerable experience in finding both previously known, and new motifs, some of which we subsequently verified experimentally. What follows are some pointers as to how to distil true motifs from noise. This is particularly important when one is dealing with sets of sequences where a common function is open to question, due to experimental limitations (e.g. sets from error-prone two-hybrid interaction datasets). It is also sometimes difficult to find motifs that are highly variable in nature (e.g. containing only two fixed positions), as many known motifs are.

It is first useful to study the nature of previously determined linear motifs. The majority of these are between 4–8 residues in length, have 2–4 specified (i.e. non-'x') positions, of which 1–3 are a single invariant amino acid (Figure 3A). They also have different preferences for particular amino acids relative to globular proteins (Figure 3B). These preferences also vary

A



B

	ALL	LIG	TRG	MOD
A	0.48	0.62	0.22	0.36
C	2.17	0.38	0.39*	8.07
G	0.55	0.55	0.16*	0.76
P	2.33	3.54	0.76	0.73
S	0.79	0.50	0.53	1.80
T	0.73	0.56	0.41	1.44
K	0.87	0.60	1.13	0.87
R	1.66	1.02	1.75	1.56
Q	0.72	0.64	0.80*	0.80
N	0.45	0.38	0.42*	0.76
D	1.09	1.27	1.30	0.47
E	0.88	0.87	1.37	0.81
H	0.79	0.86	1.95	0.10*
I	0.79	0.85	0.81	0.67
L	1.02	1.07	2.32	0.38
M	1.37	1.31	1.76	1.45
V	0.68	0.85	0.40	0.37
F	1.32	1.66	1.48	0.41*
W	2.30	2.30	3.22*	2.23*
Y	1.71	1.85	0.80*	2.20

Figure 3. Features of known linear motifs. (**A**) Distributions of length (red), number of specified (i.e. non-'x'; green) and invariant (i.e. a single specific residue; blue) positions for 120 known linear motifs extracted from the ELM database (7). Note that four motifs with lengths of 13–18 are not shown in the first (red) plot for clarity. (**B**) Degree to which residues are over-represented in known motifs. Numbers show the ratio of the abundance of the residue within the 120 motifs from ELM to the abundance in globular domains as computed from the protein databank [PDB; (28)]. 'ALL' includes all 120, 'LIG' are the 66 ligand binding, 'TRG' the 16 targeting and 'MOD' the 30 modification site motifs. For 7 of 40 residues in the latter two categories there were too few counts to obtain a confident measurement (i.e. <5); these are denoted by an asterix. Note that we have not included a fourth ELM category CLV, which includes protein cleavage sites, as there were too few examples to compute meaningful numbers. Colour scheme: red, strongly favoured in linear motifs compared to globular proteins; orange, moderately favoured; light-blue moderately disfavoured; blue strongly disfavoured.

across different types of motifs (ligands, targeting signals or post-translational modifications). Proline features in many ligands, which is not surprising as many of the best known motifs contain it (e.g. SH3→PxxP and WW→PPxY). As would be expected, post-translational modification motifs contain more Cysteine, Serine, Threonine and Tyrosine residues than average sequences. Aromatic residues are also popular in all classes of motif. There are some curious differences between chemically similar amino acids. For example, Leucine and Methionine are common, but the β-branched aliphatic residues Valine and Isoleucine are not; Arginine is favoured, but not Lysine. The range of P and S_{CONS} corresponding to real motifs is dependent on the particular species being considered. In our previous study we calculated maximum S_{CONS} confidence values ($P < 0.001$) of 3.0×10^{-17} for Yeast, 7.5×10^{-14} for Nematode, 8.0×10^{-15} for Fly and 7.0×10^{-38} Human (13). These are reasonable starting points for motif hunting when looking in these species or their relatives, but are only a rough guide in practice, as real motifs can still occur with higher (less significant) values.

In practice, one should be prepared for the situation that a true motif might not be ranked first according to our metrics. This was the case even in a benchmark using sets of experimentally validated motifs (13). It tends to happen when motifs are very simple in nature, containing few residues, or those that are naturally abundant (e.g. the endosome sorting signal ExxxLL).

For more tentative motifs, we found it also useful to consider the number of times a motif is observed relative to the size of the set (beside the P and S_{CONS} values). In our hands, spurious motifs scored better when they occurred in a few members of a large set of proteins (e.g. 4 out of 30; most sets considered were 10 or fewer in size). This could be because large sets (e.g. 20 or more sequences) contain more putative false positives, and might also begin to deviate from the binomial distribution.

It is also always important to question the integrity of the set itself—i.e. do all the proteins really belong together, or are some more weak members than others? This is particularly so when one is dealing with interaction data from high-throughput techniques like the two-hybrid system [e.g. (13,24,25)], or protein chips (26).

The complexity of available interaction data can make it difficult to provide sets of sequences to the server, particularly if the user is not used to searching interaction databases. For easier access to interactions, and to provide suitable sets of query sequences, we developed an interface to the STRING (17) database. STRING allows users to look at interaction partners from a variety of sources for any protein of interest, to navigate around the network of interactions, and to select the type of interaction data displayed. STRING now allows users to run DILIMOT directly from any set of interactions. By navigating around the network centered on a single protein, it is possible to define sets suitable for finding either putative motifs that the protein might interact with or contain.

OUTLOOK

There are likely dozens or even hundreds of additional linear motifs still to be discovered, and we anticipate that DILIMOT will be central in uncovering these critically important functional sites. The method is applicable any time a common feature is sought within a set of protein sequences. We are constantly looking for new ways to apply the approach, and to improve performance and sensitivity. We hope that use by the computational and molecular biology community will improve the service, and welcome any comments from our users.

ACKNOWLEDGEMENTS

The authors are grateful to Damian Brunner (EMBL) and Andreia Feijao (EMBL) for advice on the EB1 binding motif, Christian von Mering (EMBL) for help defining orthologues and Francesca Diella (EMBL) for help with Phospho.ELM. The authors thank Toby Gibson (EMBL) for the critical reading of the manuscript. Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*., **32**, D142–D144.
2. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res*., **30**, 276–280.
3. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
4. Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res*., **30**, 4321–4328.
5. Bork,P. and Gibson,T.J. (1996) Applying motif and profile searches. *Meth. Enzymol*., **266**, 162–184.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
7. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M., Ausiello,G., Brannetti,B., Costantini,A. *et al*. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*., **31**, 3625–3630.
8. Neduva,V. and Russell,R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett*., **579**, 3342–3345.
9. Diella,F., Cameron,S., Gemund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Ponten,T., Blom,N. and Gibson,T.J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.
10. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol*., **294**, 1351–1362.
11. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res*., **32**, D134–D137.
12. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, **31**, 3635–3641.
13. Neduva,V., Linding,R., Su-Angrand,I., Stark,A., de Masi,F., Gibson,T.J., Lewis,J., Serrano,L. and Russell,R.B. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*., **3**, e405.
14. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res*., **31**, 3701–3708.
15. Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
17. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.
18. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
19. Simpson,J.C., Wellenreuther,R., Poustka,A., Pepperkok,R. and Wiemann,S. (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep*., **1**, 287–292.
20. Su,L.K., Burrell,M., Hill,D.E., Gyuris,J., Brent,R., Wiltshire,R., Trent,J., Vogelstein,B. and Kinzler,K.W. (1995) APC binds to the novel protein EB1. *Cancer Res*., **55**, 2972–2977.
21. Busch,K.E., Hayles,J., Nurse,P. and Brunner,D. (2004) Tea2p kinesin is involved in spatial microtubule organization by transporting tip1p on microtubules. *Dev. Cell*, **6**, 831–843.
22. Miller,R.K., Cheng,S.C. and Rose,M.D. (2000) Bim1p/Yeb1p mediates the Kar9p-dependent cortical attachment of cytoplasmic microtubules. *Mol. Biol. Cell*, **11**, 2949–2959.
23. Honnappa,S., John,C.M., Kostrewa,D., Winkler,F.K. and Steinmetz,M.O. (2005) Structural insights into the EB1-APC interaction. *EMBO J*., **24**, 261–269.
24. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
25. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.
26. Zhu,H., Bilgin,M., Bangham,R., Hall,D., Casamayor,A., Bertone,P., Lan,N., Jansen,R., Bidlingmaier,S., Houfek,T. *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.
27. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.
28. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.