W552–W558 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 28 May 2009*
doi:10.1093/nar/gkp439

seeMotif: exploring and visualizing sequence motifs in 3D structures

Darby Tien-Hao Chang[1], Ting-Ying Chien[2] and Chien-Yu Chen[3,*]

[1]Department of Electrical Engineering, National Cheng Kung University, Tainan 70101, [2]Department of Computer Science and Information Engineering and [3]Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei, 10617, Taiwan, R.O.C.

Received March 4, 2009; Revised April 23, 2009; Accepted May 11, 2009

ABSTRACT

Sequence motifs are important in the study of molecular biology. Motif discovery tools efficiently deliver many function related signatures of proteins and largely facilitate sequence annotation. As increasing numbers of motifs are detected experimentally or predicted computationally, characterizing the functional roles of motifs and identifying the potential synergetic relationships between them are important next steps. A good way to investigate novel motifs is to utilize the abundant 3D structures that have also been accumulated at an astounding rate in recent years. This article reports the development of the web service seeMotif, which provides users with an interactive interface for visualizing sequence motifs on protein structures from the Protein Data Bank (PDB). Researchers can quickly see the locations and conformation of multiple motifs among a number of related structures simultaneously. Considering the fact that PDB sequences are usually shorter than those in sequence databases and/or may have missing residues, seeMotif has two complementary approaches for selecting structures and mapping motifs to protein chains in structures. As more and more structures belonging to previously uncharacterized protein families become available, combining sequence and structure information gives good opportunities to facilitate understanding of protein functions in large-scale genome projects. Available at: http://seemotif.csie.ntu.edu.tw, http://seemotif.ee.ncku.edu.tw or http://seemotif.csbb.ntu.edu.tw.

INTRODUCTION

In the past few years, an urgent demand for efficient derivation of sequence motifs has arisen, due to the large number of sequencing projects for various species (1–4). Motifs are frequently observed in biological sequences, such as transcription factor binding sites in DNA sequences and catalytic or protein–protein interaction sites in protein sequences. These motifs have been shown critical for biomolecules to perform their functions. While developing algorithms to discover motifs attracts more and more attention today, efficient determination of the biological roles of such motifs is considered as one of the next important steps (5).

In proteins, sequence motifs are largely used in automated function annotation and residue characterization. Examples of databases containing protein motifs include PROSITE (2), ELM (3), MnM (6), BLOCKS (7) and PRINTS (8). Protein sequence motifs can be roughly categorized into three distinct groups. Motifs of the first group are derived for the goal of function or domain classification, where PROSITE patterns serve as a famous example. The second group focuses more on functional sites of biopolymers, such as catalytic sites and protein–DNA or protein–protein interaction sites. Such motifs often serve as sequence signatures for predicting interacting regions and can be used to predict functionally important residues (9,10). Many motifs fall in both the first and the second groups, although they were first derived only for one of the two goals. The third group comprises motifs that are considerably shorter than those in the first two groups. They usually contain only 3–10 residues and are named short linear motifs (4,11). Many such particular sequence patterns are shown to provide pivotal functional roles, and some of them are found in intrinsically disordered regions. Both ELM and MnM contain hundreds of such motifs.

In addition to known motifs, novel motifs are continuing to be discovered daily by computational tools. Several pattern mining algorithms, including PRATT (12,13), TEIRESIAS (14), MEME (15), QuasiMotiFinder (16) and MAGIIC-PRO (17), have demonstrated the potential to discover valuable signatures using sequences only. In addition, many recent studies focus on developing effective methods for short linear motifs (18–22). Meanwhile, the task of detecting putative

*To whom correspondence should be addressed. Tel: +886 2 3366 5334; Fax: +886 2 2362 7620; Email: cychen@mars.csie.ntu.edu.tw

occurrences of these short motifs in sequences is as challenging as that of discovering novel linear motifs and has also attracted as much attention in recent years (3,6,23–25). Though motifs are detected quickly *in silico*, validating these novel motifs and associating them with biological functions through experimental methods is expensive both financially and in terms of person-hours. The CompariMotif method was recently proposed to identify and score similarities between linear motifs and provide a search facility to compare the newly discovered motifs with that in databases (5).

Along with analysis of sequence motifs in a systematic way, exploring sequence motifs in known 3D structures greatly helps to understand how they mediate protein–protein or protein–ligand interactions (26–28). As the number of entries in structure databases increases at an astounding rate, it is greatly desirable to have an easy-to-use exploration tool in which researchers can quickly see the location and conformation of newly discovered motifs among a number of related structures. Several tools have been developed to tackle this issue, including Motif3D (29), 3MATRIX and 3MOTIF (30) and MSDmotif (28), though most of them do not allow the users specify their own sequence motifs or select appropriate structures for visualization conveniently.

In this article, we present the service seeMotif, aimed at providing an easy-to-use web interface for visualizing and exploring sequence motifs in protein tertiary structures. Here a motif is defined as a pattern or a set of patterns in regular expression form. Entry IDs of PROSITE and ELM databases are also allowed. One of the most distinct features of seeMotif is it accepts multiple motifs in different styles simultaneously, providing an efficient and effective way to compare motifs. Moreover, viewing motifs in tertiary structures reveals the spatial relations of individual motifs, which provides more information about motif synergy than just considering overlapping relationships within sequences. By incorporating the structural data, seeMotif is capable of listing valuable information about sequence motifs, such as the proportion of surface residues and the distance to any ligands. As more and more structures belonging to previously uncharacterized protein families become available, combining sequence and structure information together gives good opportunities to facilitate understanding of protein functions in large-scale genome and proteome projects.

METHODS

seeMotif aims to provide a user-friendly environment for visualizing sequence motifs on 3D structures from the Protein Data Bank (PDB) (31). The visualizing procedures should be carefully handled in order to correctly present sequence motifs in protein structures. For long motifs, it is usually the case that none of the sequences from which the motifs are derived have had complete 3D structures experimentally determined. On the other hand, instances of short motifs are abundant in structure databases but many of them might simply happen by coincidence. In this regard, looking for a potential set of PDB structures on which the sequence motifs can be correctly plotted would be the first and the most important step in seeMotif. Conventional pattern matching techniques are not appropriate in handling this issue, because long motifs often miss potential true positives while short motifs usually induce many false positives. Furthermore, some patterns cannot be perfectly matched on PDB chains, due to natural or artificial position mutations or because of missing residues resulting from crystallography limitations. Therefore, seeMotif introduces two better ways for the structure selection procedure. In both ways, users are asked to specify a reference sequence that must match at least one of the motifs in the query. The first method is a three-step filtering procedure to identify potential occurrences of motifs in the homologous structure chains of the reference sequence. The second method looks for potential structures through a pre-constructed interaction network built from protein complexes in PDB. The second approach is based on the assumption that proteins having the same binding partner might contain the same motifs. In the following, we first describe how motifs are expressed and what motif blocks are. After that, the filtering constraints on block level are described, and finally the alternative procedures for structure selection are introduced.

Motif expression

Although there is no universal format for expressing sequence motifs, most motif databases and mining algorithms employ regular expression when describing patterns. Examples of POSIX (Portable Operating System Interface) regular expression are 'P..P', 'L.C.E' and '[RKY]..P..P', where the symbol '.' represents an arbitrary position. An uppercase letter matches the one-letter code of an amino acid defined by the standard IUPAC (The International Union of Pure and Applied Chemistry), and a paired square bracket '[]' matches a position where any letter in the bracket is accepted. For example: [STAIV] matches a position of Ser, Thr, Ala, Ile or Val. In seeMotif, the symbol 'x', which is widely used to represent wildcard in biological convention, is treated the same as '.'. Another adopted form to present motifs is in PROSITE style. An example (PROSITE ID: PS00136, aspartic acid active site of serine proteases) is '[STAIV]-{ERDL}-[LIVMF]-[LIVM]-D-[DSTA]-G-[LIVMFC]-x(2,3)-[DNH]'. Here, a paired curly brackets '{}' matches a position of any amino acid except for those letters denoted inside. For example: {ERDL} stands for any amino acid except Glu, Arg, Asp and Leu. The symbol '-' is used to separate any two elements described above. A symbol followed by a brackets '()' with one or two numbers inside provide a succinct notation of repeated elements. For example, x(3) equals to x-x-x; x(2,4) can be x-x, x-x-x or x-x-x-x; A(3) corresponds to A-A-A.

Motif blocks

A block is a sub-motif which contains no flexible wildcard regions, such as x(2,3), and no wildcard regions longer than four, such as x(5). In other words, we aim to partition

a long motif into several smaller blocks. Insertions or deletions of a large segment are allowed in between blocks, but not allowed within them. For example, the motif 'H-G-T-x(3)-G-x(77,101)-A-x-G-N-x(57,78)-G-T-S-x(3)-P' can be decomposed into three motif blocks: 'H-G-T-x(3)-G', 'A-x-G-N' and 'G-T-S-x(3)-P'. Each of these blocks can be treated as a small motif, but it should be remembered that these blocks are relevant to each other from the evolutionary point of view. For example, the motif blocks found by the mining server MAGIIC-PRO usually interact or cooperate with each other when proteins are folded or bind to other molecules (9,10). seeMotif can help to discover these relationships.

In seeMotif, a block match on the structure chains is assessed by two properties calculated from the local sequence alignment. An index named 'conservation score' is defined by the ratio of the number of exact matches and positive substitutions appearing on the non-wildcard positions to the number of all non-wildcard positions of a block in the consensus line of BLAST (32) output. In addition, a block is regarded as 'broken' if it is interrupted by a gap '−' on the reference sequence.

Structure selection by homology

seeMotif proposes a three-step procedure for structure selection through homology inference. The first step of the proposed approach involves pattern matching (described in the subsection 'Motif expression') to determine where the motifs are present in the reference sequence. The second step performs local sequence alignment using BLAST on the reference sequence against the set of protein chains in PDB structures. PDB structures containing at least one protein chain homologous to the reference sequence, defined by *e*-value <0.001, are then collected by seeMotif. Briefly speaking, for short motifs that might match too many unrelated proteins, this step helps to restrict the target list within homologues of the reference sequence. On the other hand, for longer or more specific motifs that match rare proteins, employing sequence alignment turns out to collect a larger set of potentially homologous structures for the next step than that obtained by performing conventional pattern matching.

With the results of the first and the second steps, a potential position mapping between a motif and a PDB chain can be constructed, even when the sequence is incomplete or has been mutated. The third step adopts two constraints to eliminate putative wrong position mappings: a block that has a conservation score <0.5 or is broken will be excluded in the result page of seeMotif. It should be noticed that the conservation score is calculated for each instance of a block, which means a motif or motif block might be present in one PDB chain but missing in another. This scenario increases the number of PDB structures selected by seeMotif while preserving the visualization quality.

Structure selection by interaction network

Investigations on linear motifs reveal that many of them occur in otherwise unrelated proteins (21). To avoid missing putative structures when visualizing linear motifs, seeMotif provides another way for structure selection that utilizes a pre-constructed protein–protein interaction (PPI) network based on 7459 PDB complex structures. All the UniProt (33) IDs present in PDB structures with more than one protein chain are collected as vertices of the network, and a link is constructed between two vertexes as long as they co-occur in the same complex. When a user submits a reference sequence, either via a UniProt or PDB ID, seeMotif extracts all of its binding partners in the PPI network and the associated complexes containing any of these binding partners. Afterward, pattern matching is performed on all the protein chains from the selected complexes, but not the binding partners of the reference protein. In this way, sequence motifs that have arisen convergently might have chances to be visualized among unrelated proteins for further sequence or structural characterization.

WEB INTERFACE

In the input page of seeMotif, users can specify the query motifs in a text field. Distinct motifs should be separated by a blank space. Various pattern formats, including POSIX regular expression, PROSITE patterns and the entry IDs of motif databases, can be used to specify a sequence motif. Moreover, seeMotif provides a simple format for pointing out a block by positions directly. For example, '165–214' indicates the segment of the reference sequence from its 165th to 214th positions. Motifs described by hidden Markov model (HMM) profiles such as that of Pfam (34) or consensus sequences derived from results of multiple sequence alignments should also be specified in this way. The reference sequence can be specified in three ways, using a UniProt accession number or entry name, providing a PDB ID along with a chain number, or typing in the protein sequence in FASTA format directly. Moreover, seeMotif accepts the search result pages of databases ELM and MnM or the mining results of MAGIIC-PRO and SLiMDisc as the input by using the technique called *bookmarklet*. Detailed information can be found on the web site of seeMotif.

Before the result page is presented, seeMotif uses an intermediate web page for reexamining the query and setting options. In this page, the users are requested to specify the approach of selecting structures. In the result page, seeMotif maps all the motifs onto the structures collected from PDB through the desired approach. There are two visualization modes in the result page: *multi-motif* and *single-motif*. In *multi-motif* mode, multiple motifs can be visualized simultaneously in distinct colors. It might happen that more than one motif can match the same residues of the reference sequence, resulting in conflicts when assigning colors. In this regard, overlaps between motifs should be detected before coloring the residues.

In *multi-motif* mode, only one matched position is considered for each motif block at a time. seeMotif enumerates all the combinations of motif matching and users can switch among them via a drop-down list. Once the positions of motif blocks are selected, motif blocks that

Table 1. Examples of motifs found for an interested protein (UniProt entry P35835)

Motif sources	Motifs	Database ID
MAGIIC-PRO	H-G-T-x(3)-G-x(77,101)-A-x-G-N-x(57,78)-G-T-S-x(3)-P	
GLAM2	165-214 or QDGSSHGTHVAGTIAALNNSIGVLGVAPSASLYAVKVLDSTGSGQYSWII	
PROSITE	[STAIV]-{ERDL}-[LIVMF]-[LIVM]-D-[DSTA]-G-[LIVMFC]-x(2,3)-[DNH]	PS00136
PROSITE	H-G-[STM]-x-[VIC]-[STAGC]-[GS]-x-[LIVMA]-[STAGCLV]-[SAGM]	PS00137
PROSITE	G-T-S-x-[SA]-x-P-x-{L}-[STAVC]-[AG]	PS00138
ELM	[RK].[AILMFV][LTKF]	CLV_PCSK_SKI1_1
ELM	Y[QDEVAIL][DENPYHI][IPVGAHS]	LIG_SH2_SRC

overlap on the reference sequence are assigned with the same color to show their relationships. On the other hand, a block without overlapping with the others can be colored in two ways: (i) the simpler one—blocks belonging to the same motif have a single color; (ii) the more complex one—each of the blocks has its own color. In *single-motif* mode, the adopted coloring scheme is simple. Each motif takes a distinct color, and for the selected motif all the matched positions are shown simultaneously. Multi-block motifs are excluded from the list, since it would be too complicated to show all the occurrences at the same time. The matched residues on the collected PDB chains are colored in agreement with the reference sequence to show their correspondences. Once the visualization method is selected, seeMotif generates a list of the collected protein structures for selection and equips a Jmol plug-in (available at http://www.jmol.org/) for rendering the selected protein structure. In addition, the domain information from Pfam is considered for further filtering unwanted matches. When users decide to see only motifs outside a domain, blocks having any positions inside a domain are excluded. Conversely, when users decide to see only motifs inside a domain, only blocks having all positions inside a domain are reserved. Users can just turn off this filter by selecting the option 'both are fine'. Screenshots and more detailed explanations of how to use these facilities for exploring and visualizing motifs can be found in the following section.

EXAMPLES

With a sequence of interest, researchers can conduct motif discovery through motif discovery web servers or standalone command-line tools. Usually a set of homologous sequences of the query protein is required during motif discovery process. This is often achieved by performing sequence alignment tools against existing sequence databases like UniProt. Alternatively, some motif discovery servers ask for the sequence set for computation directly from the users. In addition to finding potential motifs from scratch, users also frequently exploit pattern matching utilities provided by curated motif databases such as PROSITE and ELM. In this section, we use the protein Subtilisin NAT of *Bacillus* (UniProt entry P35835) as an example to demonstrate what seeMotif can do with the sequence motifs discovered by computational tools or extracted from motif databases described above.

In Table 1, we show seven sequence motifs, two discovered by mining algorithms and five from databases. seeMotif accepts patterns of different styles in a single query. We can send a query by using a combination of the three-block motif from MAGIIC-PRO along with the polypeptide QDGSSHGTHVAGTIAALNNSIGVLGVA PSASLYAVKVLDSTGSGQYSWII that GLAM2 (35) reports as the region of the best motif it found. This one-block motif can be specified alternatively by its range on the reference sequence, 165–214. It is shown in Figure 1 that both the first block of the first motif and the second motif match the segment 'HGTHVAG' of the reference sequence. Namely, these two blocks are *overlapping*. This is why they share the same color in *multi-motif* visualization mode. As shown in the structure view embedded in Figure 1, all the motif blocks used are related to the catalytic site, though some position mutations are observed in the longest block (area marked with the number '5' in Figure 1). There are more than 100 structures found in PDB for this query. Each contains at least one protein chain homologous to the reference sequence. A sequence alignment view showing the correspondences between the reference sequence and the collected PDB chains can be found by clicking an external link, named 'sequence alignment', provided above the structure panel.

Next, we attempt to examine more patterns listed in Table 1 in a single run of seeMotif, where the three PROSITE patterns can be quickly specified by using PROSITE IDs PS00136, PS00137 and PS00138. In the intermediate page, it is warned that only up to five motifs can be visualized in a single run owing to performance consideration. We unselected the PROSITE pattern PS00136, added the ELM motif 'LIG_SH2_SRC' and then submitted the query (subfigure marked with the number '5'). When visualizing the motifs in *multi-motif* mode, overlapping motif blocks are plotted in the same colors. In this example, as shown in Figure 2 (area marked with the number '2'), two combinations of position mappings are identified by seeMotif for *multi-motif* mode. These two mapping possibilities are owing to the multiple matches of the short motif from ELM database, 'Y[QDEVAIL][DENPYHI][IPVGAHS]'. Once the mapping combination is changed, all the motifs will be re-mapped on the reference sequence as well as in the structure panel. In the structure panel, structures are by default sorted by the *e*-values reported from execution of BLAST. seeMotif provides three additional strategies to reorder the structure list. If the users are interested in knowing whether the motifs play an important role

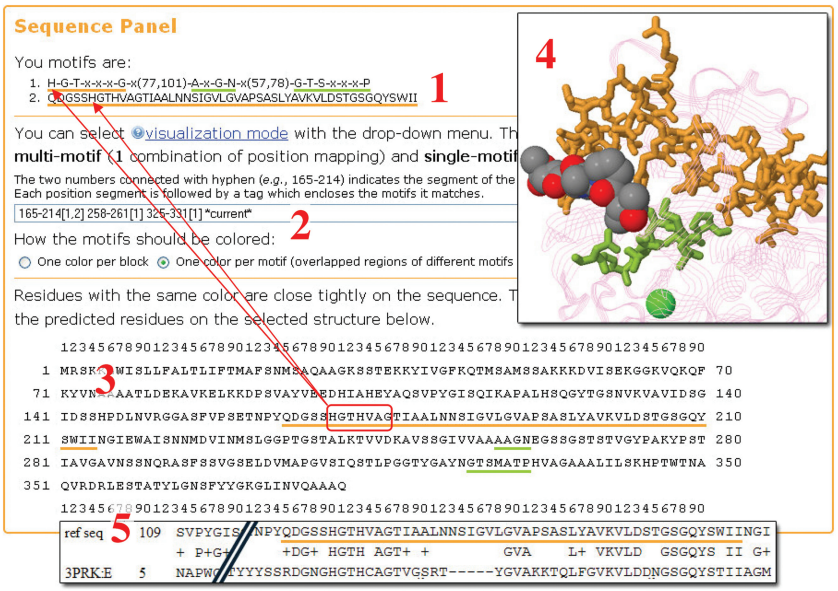

Figure 1. An example of using seeMotif by combining two motifs from different computational tools. This figure comprises five parts: (1) motifs; (2) position mappings between the motifs and the reference sequence; (3) reference sequence with the matched positions highlighted; (4) a snapshot of the structure panel; and (5) local alignment of the reference sequence and the selected protein chain. In this example, motifs are plotted on the PDB structure 3PRK:E. Protein chains are shown in *strands* style. The residues matched by any of the motifs are illustrated as *sticks* with distinct colors corresponding to their sequence expression form in the sequence panel. Ligands (the inhibitor in this structure) are displayed in *spacefill* and colored in CPK mode.

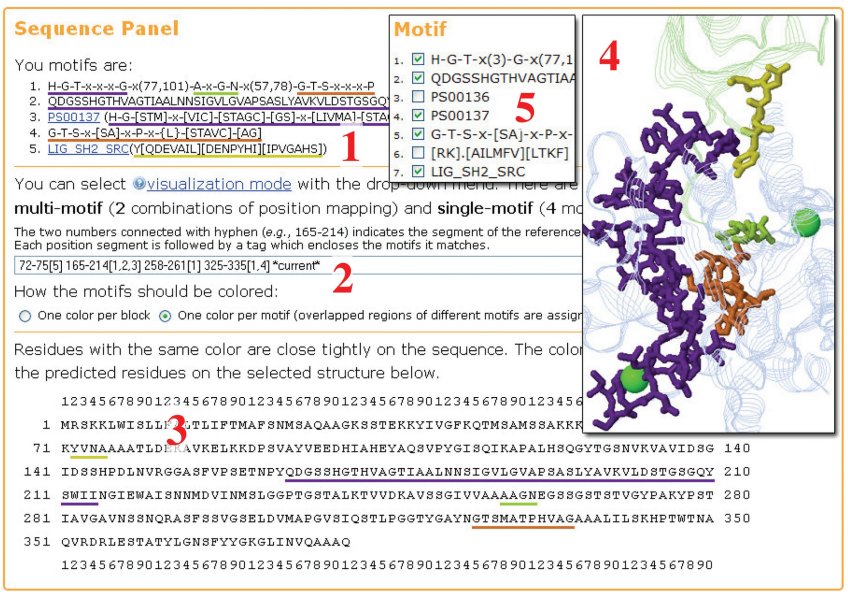

Figure 2. An example of using seeMotif by combining five motifs from both computational tools and existing motif databases. This figure comprises five parts: (1) motifs; (2) position mappings between the motifs and the reference sequence; (3) reference sequence with matched positions highlighted; (4) a snapshot of the structure panel; and (5) the motif selection panel in the intermediate page. In this example, motifs are plotted on the PDB structure 1SCJ. Protein chains are shown in *strands* style, where each chain has its own color. The residues matched by any of the motifs are illustrated as *sticks* with distinct colors corresponding to their sequence expression form in the sequence panel. CA ions are displayed in *spacefill*.

in binding other molecules, it is suggested to sort the structures by the nearest distance between any matched positions to any ligands or small ions found in the structures. Another way to explore the structures is sorting them by the total number of positions matched by the motifs (#MR as the column name). Finally, structures can also be sorted according to the proportion of surface residues among the set of matched residues (%Sur). How the motifs interact with each other in space can be viewed in the structure panel. Furthermore, it is shown in

Figure 2 that seeMotif can show the motifs on more than one chain in the same PDB structure. This facility is particularly useful when studying quaternary structures and can be easily extended to exploring protein–protein interactions.

In the last example, we demonstrate how to use the option *structure selection through interaction network*. In this example, the SV40 large T antigen (UniProt ID: P03070) is used as the reference sequence. With the reference sequence, seeMotif identifies its binding partner (P06400) and finds two complex structures (1GH6 and 1GUX) containing P06400. After that, the motif 'L.C.E' is used to scan the binding partners of P06400 in these two structures and successfully found two protein chains (1GH6:A and 1GUX:E) to visualize the motif.

CONCLUSION AND FUTURE PERSPECTIVES

In this article, we present the seeMotif server that aims to visualize sequence motifs on 3D structures related to a given reference protein. seeMotif provides an easy-to-use interface for submitting motifs in different formats. Visualizing the spatial characteristics of motifs in an interactive way and exploring the relationships between motifs in different structures largely help to understand how a cluster of particular residues affects protein functions. seeMotif will be regularly updated based on the newest release of UniProt, PROSITE, ELM and PDB databases. Furthermore, how to exploit seeMotif in detecting protein–protein and protein–DNA interactions deserves further studies in the future.

FUNDING

National Science Council of Republic of China, Taiwan [NSC 96-2320-B-006-027-MY2, NSC 96-2221-E-006-232-MY2, 97-2627-P-001-002 and 97-2221-E-002-184]. Funding for open access charge: National Science Council of Republic of China, Taiwan.

Conflict of interest statement. None declared.

REFERENCES

1. Sandve,G.K. and Drablos,F. (2006) A survey of motif discovery methods in an integrated framework. *Biol. Direct*, **1**, 11.
2. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J.A. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
3. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M.A., 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.
4. Neduva,V. and Russell,R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett.*, **579**, 3342–3345.
5. Edwards,R.J., Davey,N.E. and Shields,D.C. (2008) CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics*, **24**, 1307–1309.
6. Rajasekaran,S., Balla,S., Gradie,P., Gryk,M.R., Kadaveru,K., Kundeti,V., Maciejewski,M.W., Mi,T., Rubino,N., Vyas,J. *et al.* (2009) Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res.*, **37**, D185–D190.
7. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the Blocks Database servers. *Nucleic Acids Res.*, **28**, 228–230.
8. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
9. Hsu,C.M., Chen,C.Y., Liu,B.J., Huang,C.C., Laio,M.H., Lin,C.C. and Wu,T.L. (2007) Identification of hot regions in protein-protein interactions by sequential pattern mining. *BMC Bioinformatics*, **8(Suppl. 5)**, S8.
10. Chien,T.Y., Chang,D.T.H., Chen,C.Y., Weng,Y.Z. and Hsu,C.M. (2008) E1DS: catalytic site prediction based on 1D signatures of concurrent conservation. *Nucleic Acids Res.*, **36**, W291–W296.
11. Davey,N.E., Shields,D.C. and Edwards,R.J. (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res.*, **34**, 3546–3554.
12. Jonassen,I., Collins,J.F. and Higgins,D.G. (1995) Finding Flexible Patterns in Unaligned Protein Sequences. *Protein Sci.*, **4**, 1587–1595.
13. Jonassen,I. (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.*, **13**, 509–522.
14. Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.
15. 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.
16. Gutman,R., Berezin,C., Wollman,R., Rosenberg,Y. and Ben-Tal,N. (2005) QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. *Nucleic Acids Res.*, **33**, W255–W261.
17. Hsu,C.M., Chen,C.Y. and Liu,B.J. (2006) MAGIIC-PRO: detecting functional signatures by efficient discovery of long patterns in protein sequences. *Nucleic Acids Res.*, **34**, W356–W361.
18. Edwards,R.J., Moran,N., Devocelle,M., Kiernan,A., Meade,G., Signac,W., Foy,M., Park,S.D.E., Dunne,E., Kenny,D. *et al.* (2007) Bioinformatic discovery of novel bioactive peptides. *Nat. Chem. Biol.*, **3**, 108–112.
19. Davey,N.E., Edwards,R.J. and Shields,D.C. (2007) The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res.*, **35**, W455–W459.
20. Davey,N.E., Shields,D.C. and Edwards,R.J. (2009) Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics*, **25**, 443–450.
21. 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**, 2090–2099.
22. Edwards,R.J., Davey,N.E. and Shields,D.C. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, **2**, e967.
23. Chica,C., Labarga,A., Gould,C.M., Lopez,R. and Gibson,T.J. (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, **9**, 229.
24. Dinkel,H. and Sticht,H. (2007) A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics*, **23**, 3297–3303.
25. 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.
26. Stein,A. and Aloy,P. (2008) Contextual specificity in peptide-mediated protein interactions. *PLoS ONE*, **3**, e2524.
27. Stein,A., Panjkovich,A. and Aloy,P. (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.*, **37**, D300–D304.
28. Golovin,A. and Henrick,K. (2008) MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, **9**, 312.
29. Gaulton,A. and Attwood,T.K. (2003) Motif3D: relating protein sequence motifs to 3D structure. *Nucleic Acids Res.*, **31**, 3333–3336.

30. Bennett,S.P., Lu,L. and Brutlag,D.L. (2003) 3MATRIX and 3MOTIF: a protein structure visualization system for conserved sequence motifs. *Nucleic Acids Res.*, **31**, 3328–3332.
31. Kirchmair,J., Markt,P., Distinto,S., Schuster,D., Spitzer,G.M., Liedl,K.R., Langer,T. and Wolber,G. (2008) The Protein Data Bank (PDB), Its related services and software tools as key components for in silico guided drug discovery. *J. Medicinal Chem.*, **51**, 7021–7040.
32. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., 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.
33. Bairoch,A., Consortium,U., Bougueleret,L., Altairac,S., Amendolia,V., Auchincloss,A., Argoud-Puy,G., Axelsen,K., Baratin,D., Blatter,M.C. *et al.* (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
34. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
35. Frith,M.C., Saunders,N.F.W., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, **4**, e1000071.