W356–W361 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl309

MAGIIC-PRO: detecting functional signatures by efficient discovery of long patterns in protein sequences

Chen-Ming Hsu, Chien-Yu Chen[1,*] and Baw-Jhiune Liu

Department of Computer Science and Engineering, Yuan Ze University, Chung-Li, 320, Taiwan, Republic Of China and [1]Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei, 106, Taiwan, Republic Of China

Received February 14, 2006; Revised March 2, 2006; Accepted April 11, 2006

ABSTRACT

This paper presents a web service named MAGIIC-PRO, which aims to discover functional signatures of a query protein by sequential pattern mining. Automatic discovery of patterns from unaligned biological sequences is an important problem in molecular biology. MAGIIC-PRO is different from several previously established methods performing similar tasks in two major ways. The first remarkable feature of MAGIIC-PRO is its efficiency in delivering long patterns. With incorporating a new type of gap constraints and some of the state-of-the-art data mining techniques, MAGIIC-PRO usually identifies satisfied patterns within an acceptable response time. The efficiency of MAGIIC-PRO enables the users to quickly discover functional signatures of which the residues are not from only one region of the protein sequences or are only conserved in few members of a protein family. The second remarkable feature of MAGIIC-PRO is its effort in refining the mining results. Considering large flexible gaps improves the completeness of the derived functional signatures. The users can be directly guided to the patterns with as many blocks as that are conserved simultaneously. In this paper, we show by experiments that MAGIIC-PRO is efficient and effective in identifying ligand-binding sites and hot regions in protein–protein interactions directly from sequences. The web service is available at http://biominer.bime.ntu.edu.tw/magiicpro and a mirror site at http://biominer.cse.yzu.edu.tw/magiicpro.

INTRODUCTION

As more and more protein sequences become available with their structures undetermined, recognizing functional signatures directly from sequences is particularly desirable in functional proteomics (1–3). Automatic discovery of patterns in unaligned biological sequences is an important problem in molecular biology (5–9). For a good review on the mining algorithms, the readers can refer to Refs (10–12). When compared with the approaches based on multiple sequence alignment in identifying functional regions, pattern mining algorithms have the advantage of automatically determining the subset of sequences involved in the final mining results (13). The derived patterns are useful in many research issues in Bioinformatics, including automatic functional annotation of sequences, database search of homologues, detection of functional sites and prediction of hot regions in protein–protein interactions (2,14–16).

Pattern mining algorithms can be categorized by the description models they employ. Frequently used models include regular expressions, profiles and hidden Markov models (HMMs) (11). This paper focuses on discovery of patterns expressed in regular expression and considers only exact components in a pattern. An exact component permits only one specific amino acid in one position, such as the capital letters in the pattern N-R-x(5,19)-Y-x-G-x(3)-D. In this example, 'x' stands for a wildcard that matches any amino acid. Both 'x' and 'x(3)' are called rigid gaps, a gap of fixed length, which are composed of one and three wildcards, respectively. On the other hand, x(5,19) is a flexible gap, a gap of flexible length, which admits at least five successive wildcards and at most 19 successive wildcards in between the exact components 'R' and 'Y'. A flexible gap handles the don't-care regions where large insertions and deletions might happen during evolution, while a rigid gap deals with the conservative substitutions allowed in biological sequences.

*To whom correspondence should be addressed. Tel: +886 2 33665334; Fax: +886 2 23627620; Email: cychen@mars.csie.ntu.edu.tw

When only exact elements are considered in mining process, the derived patterns are usually very sparse, in which the pattern elements are interleaved with a large amount of gaps. Patterns of this type are hard to detect but are greatly appreciated because they concisely highlight the important residues associated with protein functional sites. In proteins, the conserved residues usually appear as clusters (it is called a block in this paper), and multiple clusters together constitute an important substructure. The conserved regions that strongly correlate with each other and conserved simultaneously are usually interleaved with large irregular gaps (16,17). In other words, the residues associated with a functional motif are not necessarily found in one region of the sequence (2,16–20). This complicates the mining process and often confuses the approaches based on multiple sequence alignment.

Regular expression is considered as a deterministic model contrary to the probabilistic models such as profiles and HMMs (11). A deterministic pattern can be matched or not matched by a sequence. In the mining process, a pattern will be reported as long as it matches more than a user-specified percentage of the input sequence set. This is the so-called minimum support constraint (21). A pattern is said to be diagnostic for a family if it matches all the known sequences in the family, and no other known sequence (10). However, a diagnostic pattern does not always correspond to a functional signature. By setting the minimum support constraint as a lower value, MAGIIC-PRO can discover patterns that really present as functional signatures but are only conserved in a subset of input sequences. Such patterns are more informative and useful in predicting ligand binding or protein interaction.

Discovering sparse and flexible patterns which are conserved in only a subset of input sequences is a time-consuming task due to the large search space of solutions. So many related studies employ other constraints in addition to the minimum support constraint to expedite the mining process. Mining algorithms that consider only short conserved words (5,6,18,19) or rigid gaps (7–9,13), such as web service Teiresias (13), are efficient and effective in identifying short motifs. On the other hand, the Pratt (22) algorithm introduced the concept of gap flexibility to enlarge the search space. However, allowing large flexible gaps might derive patterns with the conserved residues scattered. Furthermore, it has been shown by experiments in our recent work that considering large flexibilities causes the failure of Pratt to deliver satisfied results within an acceptable time (17). Different from the previous works, our approach considers two types of gaps to improve the mining efficiency, where the gaps within a conserved region are called an intra-block gap and the gaps in between two adjacent conserved regions are called inter-block gaps (17). Using two types of gap constraints for different purposes improves the efficiency of mining process while keeping high accuracy of mining results. The server MAGIIC-PRO further employs rigid intra-block gaps instead of the flexible ones proposed in Ref. (17) since it has been observed in protein sequences that insertions and deletions are seldom present in highly conserved regions (2,13). Our experimental results also reveal that considering only rigid gaps within a block is useful in eliminating noisy patterns.

MAGIIC-PRO provides many useful tools for examining and visualizing the derived patterns, which will be described in detail later. After that, we will show by experiments that MAGIIC-PRO is efficient and effective in identifying functional sites and predicting hot regions in protein–protein interactions.

METHOD

The web service MAGIIC-PRO is in particular designed for mining protein sequences, where the kernel algorithm executing sequential pattern mining is based on our previously developed algorithm MAGIIC (17) incorporated with several state of the art data mining techniques. MAGIIC-PRO first quickly identifies rigid gapped blocks by bounded-prefix growth technique of MAGIIC. After that, the candidate blocks are concatenated into patterns with large irregular gaps by exploiting the antimonotone characteristic of this problem (20,21). Finally, a newly proposed bounded-gap closure checking scheme developed based on Ref. (23) is executed to eliminate patterns that can be covered by other super patterns with the same occurrences.

After the mining process terminates, MAGIIC-PRO generates a pattern snapshot that shows all the derived patterns in alignment with the query protein. The residues present in different patterns are combined together to create a conservation plot, where the conservation level of each residue is determined by the percentage of total number of supporting proteins merged from different patterns. The conservation plot provides a whole picture about the conserved residues of a query protein.

Input

We assume that every user of MAGIIC-PRO has a protein sequence of interest at hand. MAGIIC-PRO takes a protein sequence as input, and helps the users to prepare the training data for pattern mining. The task of collecting relative sequences of the query protein can be achieved by using Swiss-Prot annotations or executing the PSI-BLAST program. Once the query protein and the training data have been determined, the mining process is executed using the parameters described in the following subsection.

Parameters

The most important parameter of MAGIIC-PRO is the minimum support constraint. A pattern will be reported as long as it matches at least a certain number of sequences. The support constraint is critical to the mining results, but it might not be possible to know in advance by what percentage level a satisfied pattern can be discovered. Since lower values bring more patterns, the users are suggested to start with a large support constraint, e.g. 90%, and MAGIIC-PRO will decrease it gradually until a desired number of patterns have been found.

In addition to the minimum support constraint, MAGIIC-PRO has some other parameters for advanced users. Before going into the details, we first give a formal definition of a pattern block. Assume that a pattern is consisted of pattern elements as a sequence, and each successive pair of elements is either interleaved with a gap or not. In this work, small and

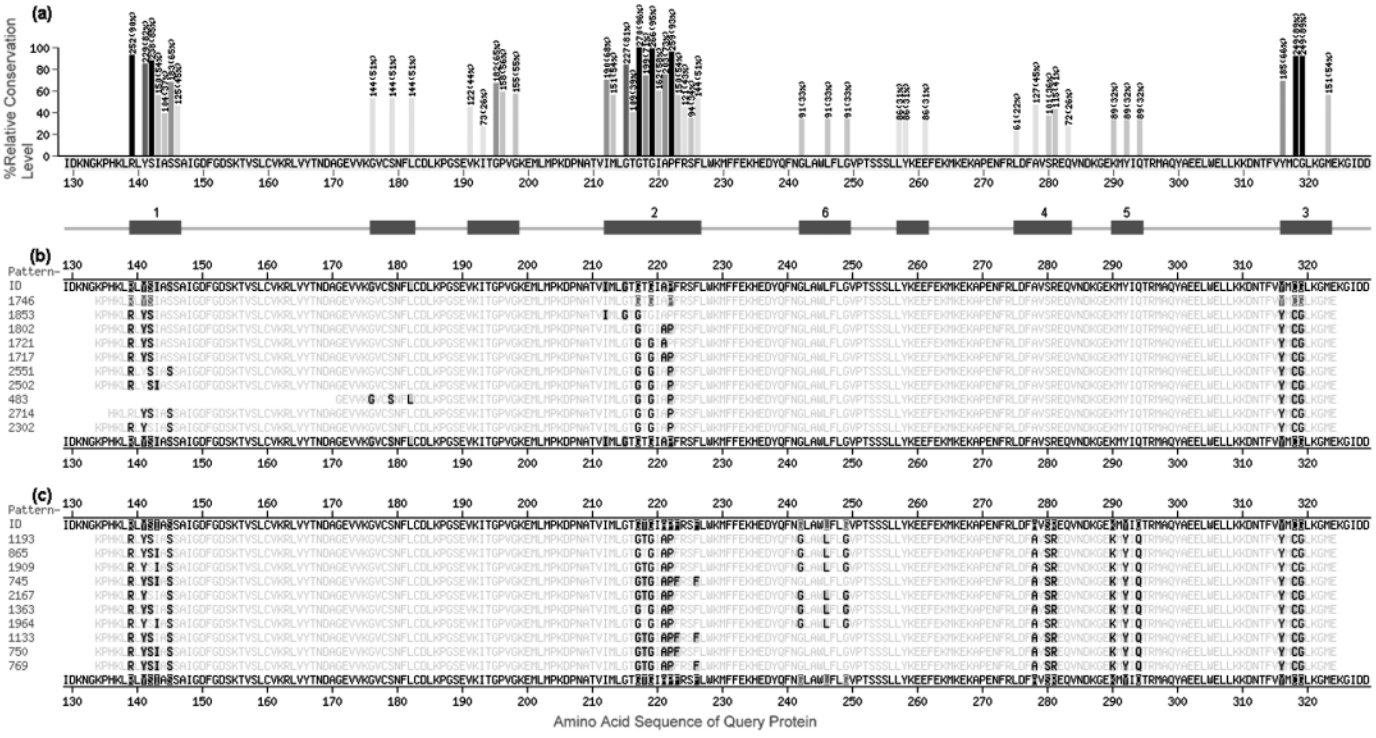

Figure 1. Examples of pattern snapshots and the conservation plot provided by MAGIIC-PRO. (**a**) The complete conservation plot derived from all the patterns. (**b**) Top 10 high-support patterns with three or more blocks. (**c**) Top 10 large-size patterns with three or more blocks.

rigid gaps are considered as intra-block gaps, while large and flexible gaps are treated as inter-block gaps. This thus defines the boundaries of the blocks. The notation x(a,b), $a < b$, is used for a flexible gap with minimum length of a and maximum length of b, and x(a) stands for a rigid gap with a fixed length of a. The wildcard x(a) is omitted if $a = 0$, and is written as x if $a = 1$, i.e. $x = x(1)$. The first group of the advanced parameters specifies the gap constraints.

(i) The maximum length of an intra-block gap (default value = 3);
(ii) The maximum relative flexibility of an inter-block gap with respect to the length of the inter-block gap present in the query protein (default value = 30%);
(iii) The second group of the advanced parameters specifies the size or length constraints.
(iv) The minimum number of elements in a block (default value = 3);
(v) The minimum number of blocks in a pattern (default value = 2);

We argue that a pattern should have at least two blocks to be meaningful, because an important region is seldom to be conserved singly either from structural or functional aspects. In this way, the users can be directly guided to the important discoveries.

Output

After the mining process finishes, the users can first take a look on the conservation plot and pattern snapshot. As shown in the Figure 1a, the locations of the conserved regions are summarized in the complete conservation plot derived from all the patterns. It can be observed in Figure 1a that there are nine conserved regions in the query protein. In the same web page, the users are provided with an interactive interface to collect patterns of interest in a pattern snapshot.

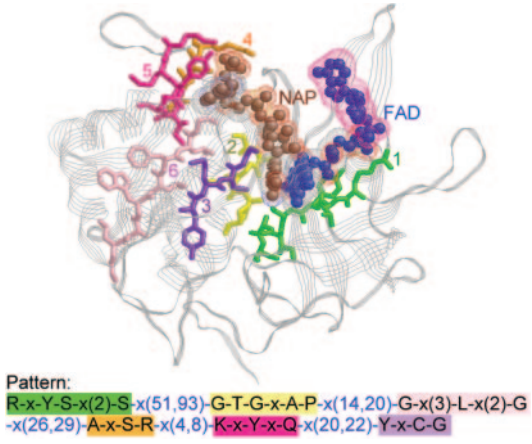

Figure 2. A pattern plotted with an available structure of the Oxidoreducatase FAD/NAD(P)-binding protein. The blocks are numbered in the same way as in Figure1a. Structures are shown with the conserved pattern blocks plotted with sticks in different colors, block 'R-x-Y-S-x(2)-S' highlighted in green, block 'G-T-G-x-A-P' in yellow, block 'G-x(3)-L-x(2)-G' in pink, block 'A-x-S-R' in orange, block 'K-x-Y-x-Q' in deep pink, and block 'Y-x-C-G' in purple, the ligand FAD plotted with ball-and-stick representation in blue, and the ligand NAP with ball-and-stick representation in brown. (PDB code1QFY:A, query protein: P10933, FENR1_PEA).

Different from the conservation plot, a pattern snapshot in addition tells which pattern blocks are simultaneously conserved during evolution. The users are suggested to browse the lists of the top 10 high-support and top 10 large-size patterns. The size of a pattern is defined as the number of exact components it contains. A pattern with a high support usually highlights the most highly conserved residues that are related to a functional region, while a longer pattern with a lower support in general provides a complete signature with respect to a functional site.

Here we use the same example from Figure 1a to explain how the interactive snapshot can facilitate examining the mining results. In Figure 1b, we first examine the top 10 high-support patterns with 3 or more blocks. Similar patterns can be considered as being associated with the same functional site, but each of them is distinct from the others because the sets of supporting sequences are different. The top one pattern in Figure 1b identifies the most three important regions of this query protein, which are related to the binding sites of the ligands FAD and NAP, denoted as the blocks 1, 2 and 3 in both Figures 1a and 2. Next, we can request the top ten large-size patterns with three or more blocks. It is observed in Figure 1c that blocks 4 and 5 are the next most conserved blocks that are simultaneously conserved with blocks 1, 2 and 3, and the further next is the block 6. The top one large-size pattern in Figure 1c is plotted with an available PDB structure in Figure 2, showing that these six blocks together constitute a complete signature regarding the binding sites of ligands FAD and NAP.

To facilitate studying the patterns of interest, we provide five useful links for each pattern. First, the web page highlights the locations of the pattern in its supporting sequences. Second, the derived pattern can be plotted with a protein structure if there are PDB entries available for any of the supporting sequences. Third, the derived pattern can be fed to the ScanProsite web service to check its selectivity, the ability to reject false positive matches. Fourth, the users can perform a multiple sequence alignment on the segments of supporting sequences that are associated with the selected pattern. This helps the user to construct a more generalized pattern with amino acid substitutions considered. Fifth, MAGIIC-PRO aligns each excluded sequence with the segment of the query protein. This helps to tell why a particular sequence does not match the pattern.

RESULTS AND DISCUSSIONS

In this section, we first demonstrate the efficiency of MAGIIC-PRO in identifying long patterns based on the 13 datasets with different levels of similarities listed in Table 1. With the default settings of the advanced parameters, MAGIIC-PRO starts the search by setting the minimum support constraint as 90%, and decreases this constraint step by step until at least one pattern have been found. At this stage, we observed that most of the patterns with the maximum support are related to a functional site of the query protein but do not serve as a complete signature of a functional site. In order to find patterns with more conserved blocks involved, we continued decreasing the minimum support constraint and stopped the process when the calculation time of a single mining task is longer than 60 s. Table 1 reports the minimum support where we stopped for each dataset, as well as the searching time used in the latest search. Table 1 also shows the number of blocks generated in the first stage of the mining process and the number of derived patterns with at least two blocks. It is observed in Table 1 that a large amount of single blocks do not collaborate with other blocks to form a longer pattern. The number of patterns converges rapidly when the number of blocks in a pattern increases. The patterns found in the top 10 high-support and the top 10 large-size lists of each dataset demonstrate the potential of MAGIIC-PRO in identifying functional sites and hot regions in protein–protein interactions. Owing to the limited space, we only show one experimental result in the following paragraph, while the others are provided on the web page of MAGIIC-PRO.

Here we use the case of query protein P00730 to illustrate that the long patterns found by MAGIIC-PRO are biologically meaningful. The pattern in Figure 3a constitutes the pocket for INF (N-(Hydroxyaminocarbonyl) Phenylalanine) and the zinc ions. This pattern matches 42 sequences in the training data. A longer pattern with a lower support

Table 1. Analysis of the efficiency of MAGIIC-PRO

Query protein	Size of training data	Setting of minimum support (%)	No. of blocks derived in the first stage	No. of the derived patterns with different number of blocks					Time used (in s)
				2	3	4	5	6	
O14965 (STK6_HUMAN) Serine/threonine-protein kinase 6	1910	60	23	4	—	—	—	—	27
P51656 (DHB1_MOUSE) Estradiol 17-beta-dehydrogenase 1	494	20	211	462	34	—	—	—	9
P19120 (HSP7C_BOVIN) Heat shock cognate 71 kDa protein	473	90	115	3	—	—	—	—	12
P00962 (SYQ_ECOLI) Glutaminyl-tRNA synthetase	346	70	75	12	4	—	—	—	4
P10933 (FENR1_PEA) Ferredoxin—NADP reductase	280	20	1490	1112	948	618	192	7	23
P08622 (DNAJ_ECOLI) Chaperone protein dnaJ	275	80	79	86	78	66	5	—	5
P25910 (BLAB_BACFR) Beta-lactamase type II precursor	267	5	2712	860	139	26	—	—	24
P27142 (KAD_BACST) Adenylate kinase	243	80	146	178	33	—	—	—	2
P22887 (NDKC_DICDI) Nucleoside diphosphate kinase	233	70	78	187	169	37	—	—	1
P09372 (GRPE_ECOLI) Protein grpE	195	30	479	1533	1836	599	23	—	9
P00730 (CBPA1_BOVIN) Carboxypeptidase A1 precursor	57	50	141	49	72	9	3	—	1
P08692 (ARSC1_ECOLI) Arsenate reductase	51	70	18	7	—	—	—	—	<1
P35568 (IRS1_HUMAN) Insulin receptor substrate 1	25	80	20	2	—	—	—	—	<1

The symbol hyphen stands for 'no patterns found.'

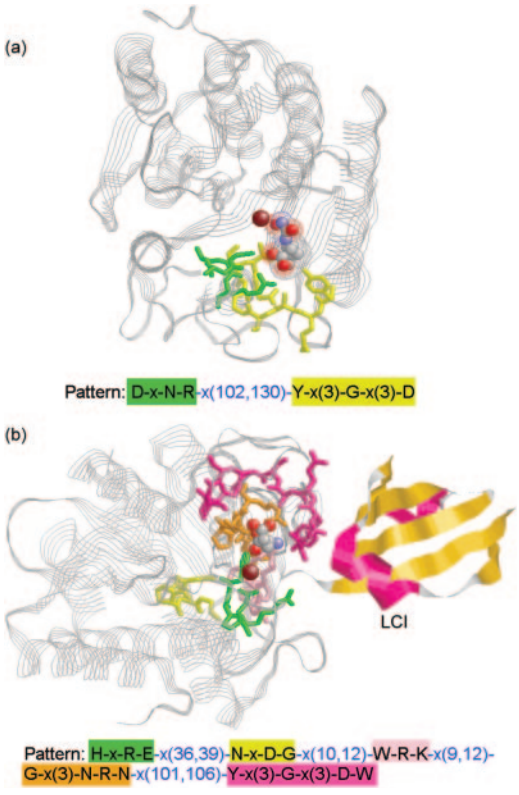

Figure 3. The patterns discovered for P00730. Patterns are shown in sticks with different blocks plotted by distinct colors, LCI protein in ribbons and zinc ions in crimson spheres. (**a**) The pattern with a high support is plotted with the structure of the bovine pancreatic carboxypeptidase A complexed with the ligand INF, which is plotted in ball-and-sticks representation and colored in CPK (1hdq.pdb, P00730) (**b**) A longer pattern with a lower support provides the contact regions when interacting with the protein LCI, shown with the structure of another protein P48052 in complex with LCI, where the ligand GLU is plotted in ball-and-sticks representation and colored in CPK. (1DTD.pdb, P48052).

(28 sequences) is plotted in Figure 3b. It is of interest that this pattern constitutes the substructure which presents its importance from another protein (P04852) in the complex with the LCI protein. This small example shows the necessity of finding motifs with different conservation levels that match different subsets of sequences in the training data. On the other hand, the diagnostic patterns provided in the PROSITE database simply capture the signature regarding the zinc-binding site.

Limitation of MAGIIC-PRO

Since the minimum number of the elements in a block is suggested to be set as 3, it might happen that some residue or two of the residues are conserved but cannot be found by MAGIIC-PRO in its primitive results. In this case, the users are suggested to perform a multiple sequence alignment for an interested pattern on the matched segments of supporting sequences through the link provided by MAGIIC-PRO. By this way, the derived patterns can be enhanced with multiple sequence alignment to have both singly conserved residues and conservative substitutions well considered.

CONCLUSION

Detecting functional signatures directly from primary information is a challenging task. The mining process is tedious especially when the users have no prior knowledge about the query protein that can be used to judge how the mining results are. MAGIIC-PRO quickly guides the biologists directly to the most highly conserved regions, and after that the users can extend the derived patterns by using the advanced parameters to refine the mining results. The derived patterns are useful in prediction of protein functions and structures, protein–ligand interactions and protein–protein interactions.

ACKNOWLEDGEMENTS

The authors would like to thank National Science Council of Republic of China, Taiwan, for the financial support under the contract NSC 94-2213-E-002-125. Funding to pay the Open Access publication charges for this article was provided by Yuan Ze University and National Taiwan University.

Conflict of interest statement. None declared

REFERENCES

1. 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.
2. Su,Q.J., Lu,L., Saxonov,S. and Brutlag,D.L. (2005) eBLOCKs: enumerating conserved protein blocks to achieve maximal sensitivity and specificity. *Nucleic Acids Res*., **33**, D178–D182.
3. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J. (2006) The PROSITE database. *Nucleic Acids Res*., **34**, D227–D230.
4. Ogiwara,A., Uchiyama,I., Yasuhiko,S. and Kanehisa,M. (1992) Construction of a dictionary of sequence motifs that characterize groups of related proteins. *Protein Eng*., **5**, 479–488.
5. Saqi,M.A.S. and Sternberg,M.J.E. (1994) Identification of sequence motifs from a set of proteins with related function. *Protein Eng*., **7**, 165–171.
6. Blekas,K., Fotiadis,D.I. and Likas,A. (2003) Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics*, **19**, 607–617.
7. Califano,A. (2000) SPLASH: structural pattern localization analysis by sequential histograms. *Bioinformatics*, **16**, 341–347.
8. Narasimhan,G., Bu,C., Gao,Y., Wang,X., Xu,N. and Mathee,K. (2002) Mining protein sequences for motifs. *J. Comput. Biol*., **9**, 707–720.
9. Brazma,A., Jonassen,I., Eidhammer,I. and Gilbert,D. (1998) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol*., **5**, 277–305.
10. Eidhammer,I., Jonassen,I. and And Taylor,W.R. (2004) *Protein Bioinformatics: An Algorithmic Approach To Sequence And Structure Analysis*. Hoboken, NJ: John Wiley & Sons.
11. Gusfield,D. (1997) *Algorithms On Strings, Trees And Sequences: Computer Science And Computational Biology*. Cambridge, UK: Cambridge University Press.
12. Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the Teiresias algorithm. *Bioinformatics*, **14**, 55–67.
13. Ogmen,U., Keskin,O., Aytuna,A.S., Nussinov,N. and Gursoy,A. (2005) PRISM: protein interactions by structural matching. *Nucleic Acids Res*., **33**, W331–W336.
14. Keskin,O., Ma,B. and Nussinov,R. (2005) Hot regions in protein-protein interactions: the organization and contribution of structurally conserved host spot residues. *J. Mol. Biol*., **245**, 1281–1294.

15. Chakrabarti,S., Anand,A.P., Bhardwaj,N., Pugalenthi,G. and Sowdhamini,R. (2005) SCANMOT: searching for similar sequences using a simultaneous scan of multiple sequence motifs. *Nucleic Acids Res*., **33**, W274–W276.
16. Hsu,C.-M., Chen,C.-Y., Hsu,C.-C. and Liu,B.-J. Efficient discovery of structural motifs from protein sequences with combination of flexible intra- and inter-block gap constraints. LNAI 3918, Springer-Verlag, Singapore*Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-06)*, 9–12 April, 530–539.
17. Neuwald,A.F. and Green,P. (1994) Detecting patterns in protein sequences. *J. Mol. Biol*., **239**, 698–712.
18. Wang,J.T.L., Marr,T.G., Shasha,D., Shapiro,B.A. and Chirn,G.W. (1994) Discovering active motifs in sets of related protein sequences and using them for classification. *Nucleic Acids Res*., **22**, 2769–2775.
19. Ke,W., Yabo,X. and Jeffery,X.X. (2004) Scalable sequential pattern mining for biological sequences. In *Proceedings of The 13th International Conference on Information and Knowledge Management (CIKM-04),* Washington, D.C., USA, November 8–13, USA, NY: ACM Press, pp. 178–187.
20. Pei,J., Han,J. and Wang,W. (2002) Mining sequential patterns with constraints in large databases. *In Proceedings of The 11th International Conference on Information and Knowledge Management (CIKM-02),* McLean, Virginia, USA, November 4–9, USA, NY: ACM Press, pp. 18–25.
21. Jonassen,I. (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci*., **13**, 509–522.
22. Wang,J. and Han,J. BIDE: efficient mining of frequent closed sequences. In *Proceedings of The 20th International Conference on Data Engineering (ICDE-04),* Boston, MA, USA, 30 March-2 April, IEEE Computer Society Press, pp. 79–90.