D178–D182 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki060

eBLOCKs: enumerating conserved protein blocks to achieve maximal sensitivity and specificity

Qiaojuan Jane Su, Lin Lu[1], Serge Saxonov[2,3] and Douglas L. Brutlag[2,*]

Abgenix, Inc., 6701 Kaiser Drive, MS 11, Fremont, CA 94555, USA, [1]Oracle Corporation, 400 Oracle Parkway, Redwood shores, CA 94065, USA, [2]Department of Biochemistry and [3]Biomedical Informatics, Stanford University, Stanford, CA 94305, USA

Received August 16, 2004; Revised and Accepted October 4, 2004

ABSTRACT

Classifying proteins into families and superfamilies allows identification of functionally important conserved domains. The motifs and scoring matrices derived from such conserved regions provide computational tools that recognize similar patterns in novel sequences, and thus enable the prediction of protein function for genomes. The eBLOCKs database enumerates a cascade of protein blocks with varied conservation levels for each functional domain. A biologically important region is most stringently conserved among a smaller family of highly similar proteins. The same region is often found in a larger group of more remotely related proteins with a reduced stringency. Through enumeration, highly specific signatures can be generated from blocks with more columns and fewer family members, while highly sensitive signatures can be derived from blocks with fewer columns and more members as in a superfamily. By applying PSI-BLAST and a modified *K*-means clustering algorithm, eBLOCKs automatically groups protein sequences according to different levels of similarity. Multiple sequence alignments are made and trimmed into a series of ungapped blocks. Motifs and position-specific scoring matrices were derived from eBLOCKs and made available for sequence search and annotation. The eBLOCKs database provides a tool for high-throughput genome annotation with maximal specificity and sensitivity. The eBLOCKs database is freely available on the World Wide Web at http://motif.stanford.edu/eblocks/ to all users for online usage. Academic and not-for-profit institutions wishing copies of the program may contact Douglas L. Brutlag (brutlag@stanford.edu). Commercial firms wishing copies of the program for internal installation may contact Jacqueline Tay at the Stanford Office of Technology Licensing (jacqueline.tay@stanford.edu; http://otl.stanford.edu/).

INTRODUCTION

During the last two decades, the successful scale-up of automated high-throughput DNA sequencing technologies has made a dramatic change to the biological discoveries in biology and biomedical sciences. An increasing number of complete genome sequences from various organisms have been determined, and the first draft of the full human genome sequence is now available. A major new goal of the Human Genome Project is functional genomics, which uses experimental and computational techniques to elucidate the function and structure of the encoded gene products. Computer-aided sequence analysis has become an increasingly important method for identifying the function of uncharacterized proteins translated from genomic or cDNA sequences. The primary method for sequence analysis is similarity search, such as BLAST (1,2), FASTA (3) and Smith–Waterman (4) programs. However, in many cases, the sequence of an unknown protein is too distantly related to any protein of known function to detect its resemblance by overall, or even local, sequence alignment. The biological function of such a sequence can often be revealed by detection within its sequence of patterns conserved among a family of proteins. Sequence patterns emphasize specific residues that are essential for a biological function ignoring other positions that are not as important for function. The conserved patterns of a protein family usually correspond to important functions such as ligand binding, catalysis, protein interaction, etc.

A conserved pattern or motif is derived from the multiple sequence alignment of a group of related proteins. Therefore, compilation of alignments of conserved protein regions is the basis of pattern recognition for protein identification. A number of protein family and superfamily databases have been built to archive the conserved alignments and searching

*To whom correspondence should be addressed. Tel: +1 650 723 6593; Fax: +1 650 723 6783; Email: brutlag@stanford.edu

tools that have been developed to link a query sequence to the related family or superfamily through different pattern matching algorithms. Widely used protein family and superfamily databases include Pfam (5), PROSITE (6), SMART (7), PRINTS (8), ProDom (9), Domo (10,11), BLOCKS+ (12), InterPro (13), SYSTERS (14), ProtoMap (15), CluSTr (16), SBASE (17), ProClass (18) and ProtoNet (19). Structural classification databases cluster proteins at the three-dimensional structure level. Structure classes are defined in databases such as SCOP (20), CATH (21) and FSSP (22).

A biologically important region is most stringently conserved among a smaller family of highly similar proteins. The same region is often found in a larger group of more remotely related proteins with a reduced stringency. The eBLOCKs database has been designed to enumerate a cascade of protein blocks with varied conservation levels for each functional domain. Through enumeration, highly specific signatures can be generated from blocks with more columns and fewer family members, while highly sensitive signatures can be derived from blocks with fewer columns and more members as in a superfamily. We have generated the eBLOCKs database to compile ungapped conserved regions, or blocks, directly from an unclassified sequence database in a generic and fully automated way. eBLOCKs builds protein groups from sequences based on PSI-BLAST searches (23). eBLOCKs clusters PSI-BLAST hit sequences into groups of many different conservation levels: subfamilies, families and superfamilies. Each group represents one distinct level of conservation, which can then be used to build patterns of a particular specificity. The enumeration of blocks with an array of different specificities determines the basis for generating motifs or position-specific scoring matrices (PSSMs) over a wide range of sensitivity and specificity. This feature of eBLOCKs allows recognition of a given query sequence by matching with blocks of all family levels, providing a solution to the dilemma of sensitivity and specificity in pattern recognition.

METHODS

eBLOCKs first uses PSI-BLAST to find all the sequences that share various similarities to a seed sequence. Sequences that are reported in a PSI-BLAST result usually fall into groups that share distinct regions or domains. A typical PSI-BLAST result is illustrated in Figure 1. Group 1 sequences share global similarity to the query sequence (seed sequence); Group 2 sequences share a domain close to the N-terminus of the seed sequence; Group 3 sequences share the C-terminal region of the seed sequence. Group 1 could form a protein family, while Groups 2 and 3 could be superfamilies sharing more distant similarities. Although the region shared by Group 2 is also included in Group 1, the same region is less conserved in Group 2 than in Group 1. Therefore, eBLOCKs not only builds high specificity blocks from Group 1, but also generates sets of high sensitivity blocks from Groups 2 and 3. The eBLOCKs database was built with three major steps: (i) cluster a PSI-BLAST result into individual groups representing distinct similarity modules; (ii) make gapped multiple sequence alignment for sequences contained in each group; and (iii) trim each gapped multiple alignment into ungapped subregions, or blocks.

Using a modified *K*-means clustering method, the sequences returned from the PSI-BLAST search are classified into *K* clusters, where *K* is automatically determined by a heuristic method. The individual cluster thus obtained represents a subgroup that aligns to a distinct region of the query sequence. The grouping of clusters is illustrated in Figure 2. In this example, Cluster 8 represents a group of closely related sequences that are globally similar. Cluster 2 represents a group of sequences that are almost globally similar but differ in the N-terminus. Cluster 9 represents a group of sequences that can only align at a region closer to the C-terminal end. Each cluster is further organized together with all of its 'covering' clusters to form a group, where the 'covering clusters' are the other clusters sharing a longer region that fully covers the region shared by the cluster. As shown in Figure 2, Cluster 8 forms Group 8, and Clusters 2 and 8 form Group 2, while all the three clusters form Group 9. Representing the same region by multiple groups with different levels of conservation allows eBLOCKs to annotate a novel sequence with maximal specificity and sensitivity. Models built from these different groups are able to extract patterns that are conserved within a subfamily, a family or a superfamily. The group assembly is done for each cluster, and thus the total number of groups is equal to the number of clusters found by *K*-means.

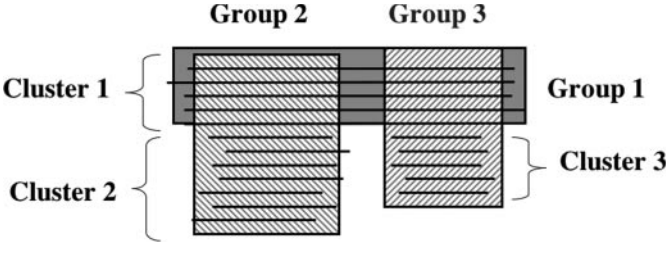

Figure 1. A typical PSI-BLAST result have multiple similarity modules. Group 1 contains sequences in Cluster 1; Group 2 contains sequences in Clusters 1 and 2; and Group 3 contains sequences in Clusters 1 and 3.

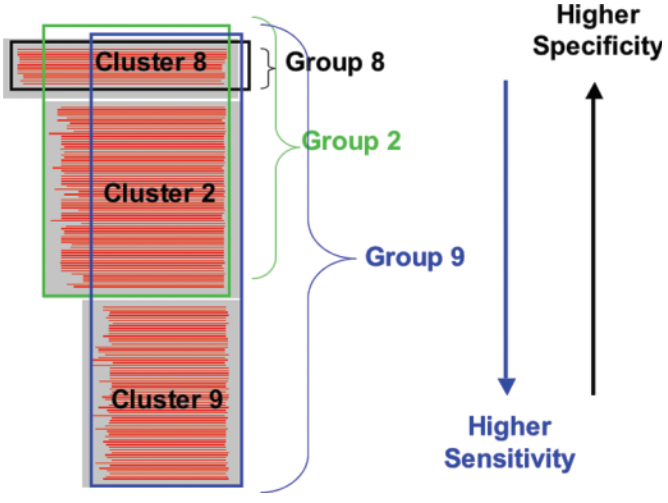

Figure 2. Clusters defined by *K*-means clustering are organized into groups. A typical conservation region is represented by multiple groups with different similarity levels, so as to maximize specificity and sensitivity. Group 8 contains sequences in Cluster 8; Group 2 contains sequences in Clusters 8 and 2; Group 9 contains sequences in Clusters 8, 2 and 9.

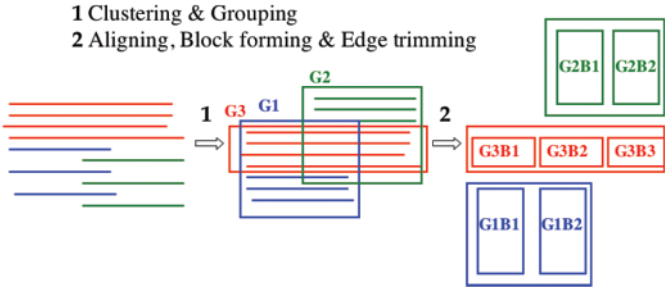

Figure 3. A flowchart for the eBLOCKs algorithm. Similarity groups that represent shared modules at different conservation levels are formed by the clustering and grouping of all the subject sequences returned by a PSI-BLAST search. Sequences in each group are aligned and the ungapped regions are excised to form several blocks. An eBLOCK accession number is composed of three parts: the SWISS-PROT accession number of the seed sequence, the group number as assigned by *K*-means clustering and the block number as the sequential number of trimmed blocks from the multiple sequence alignment for the group.

After the PSI-BLAST result has been divided into groups that represent distinct conservation modules, sequences in each group are aligned together. One multiple sequence alignment is generated for each group. The alignment is derived from the PSI-BLAST alignments. Such alignments contain gaps.

We define eBLOCKs as ungapped conserved regions. The block widths directly affect the specificities of the derived patterns, either as regular expressions or probability matrices. To ensure that the blocks provide sufficient specificity, we set a minimum width of 10 positions for eBLOCKs.

From each multiple alignment generated for each group, all the subregions with at least 10 consecutive non-gapped positions are trimmed out as raw blocks. Both the front and back edges of each raw block are examined to allow refinement and extension of the edges when the conservation level is high.

Figure 3 summarizes the generation of eBLOCKs from one PSI-BLAST result. Similarity groups representing shared modules at different conservation levels, including subfamilies, families, superfamilies, are formed by the clustering and grouping of all the subject sequences returned by a PSI-BLAST search. Sequences in each group are aligned and the ungapped regions are excised to form several blocks. An eBLOCKs accession number is composed of three parts: the SWISS-PROT accession number of the seed sequence, the group number as assigned by *K*-means clustering and the block number as the sequential number of trimmed blocks from the multiple sequence alignment for the group.

RESULTS

The current eBLOCKs database was built with Swiss-Prot Release 40, resulting in a total of 159 974 blocks. The distribution of the average information content is shown in Figure 4a. The distribution of the block width in the eBLOCKs database is shown in Figure 4b. The distribution of the number of member sequences is shown in Figure 4c.

Blocks can be used to detect related sequences through pattern matching. Two kinds of patterns can be computed from blocks: discrete motifs [regular expressions (24)] and PSSMs (25,26). Discrete motifs were generated for the blocks in eBLOCKs database using the eMOTIF package (27,28). PSSMs were computed for the blocks in eBLOCKs using eMATRIX package (26,29). Expectation values (*E*-value) were calculated as described below. For each motif generated by eMOTIF, an *E*-value is calculated from its specificity (S_i) by summing up all other equal or better specificities (S_j) in the database:

$$E\text{-value}(S_i) = \sum_{S_j \leqslant S_i} S_j \qquad \mathbf{1}$$

For each eMATRIX hit, the specificity can be converted into an *E*-value by multiplying by N, the number of tests performed, which is equal to $B * (L - W + 1)$, where B is the total number of blocks (29):

$$E\text{-value} = N \times S \qquad \mathbf{2}$$

The eBLOCKs database is available on the Web at http://motif.stanford.edu/eblocks/. Users can submit query sequences in 'Search A Sequence' page, and select either eMotif or eMatrix as the tool to search against eMOTIF or eMATRIX databases derived from the current eBLOCKs database. In the result page, eBLOCKs hits are ranked by *E*-values and each hit has a pointer to the eBLOCK record page. The eBLOCK record page displays the accession number of the block, the block alignment, the sequence Logo and provides commands to use the corresponding PSSM to scan a number of databases to retrieve matching sequences. The eBLOCKs database is also searchable by accession number and by keywords as provided in 'Search By Accession' and 'Search By Keyword' pages.

DISCUSSION

We have built an eBLOCKs database automatically from protein sequences. eBLOCKs represents a similarity region multiple times by enumerating groups with different levels of conservations (Figure 2). This important feature of eBLOCKs maximizes its sensitivity and specificity when used to annotate a query sequence. When a region is represented at the superfamily level, more remotely related sequences are included in the block, which is consequently narrower and allows more substitutions for the conserved residues (Group 9 in Figure 2). Conversely, a family or subfamily level block contains more closely related sequences, and therefore is wider and more restricted in residue substitutions (Group 8 in Figure 2). Thus, eBLOCKs actually archives family trees for each conservation region. Specificity increases when going up the tree to the subfamily level, and sensitivity increases when going down the tree to the superfamily level. By enumerating all family levels, eBLOCKs forms the basis for highly sensitive and highly specific pattern matching and enables pattern discovery with optimal sensitivity and specificity for a given query sequence.

The eBLOCKs building process is generic and can be applied to any set of protein sequences. The characterized proteins in SWISS-PROT are selected as the building set for this work since these are sources of extensively validated annotation and are therefore useful when applied to identify an unknown sequence. Nonetheless, TrEMBL or other protein

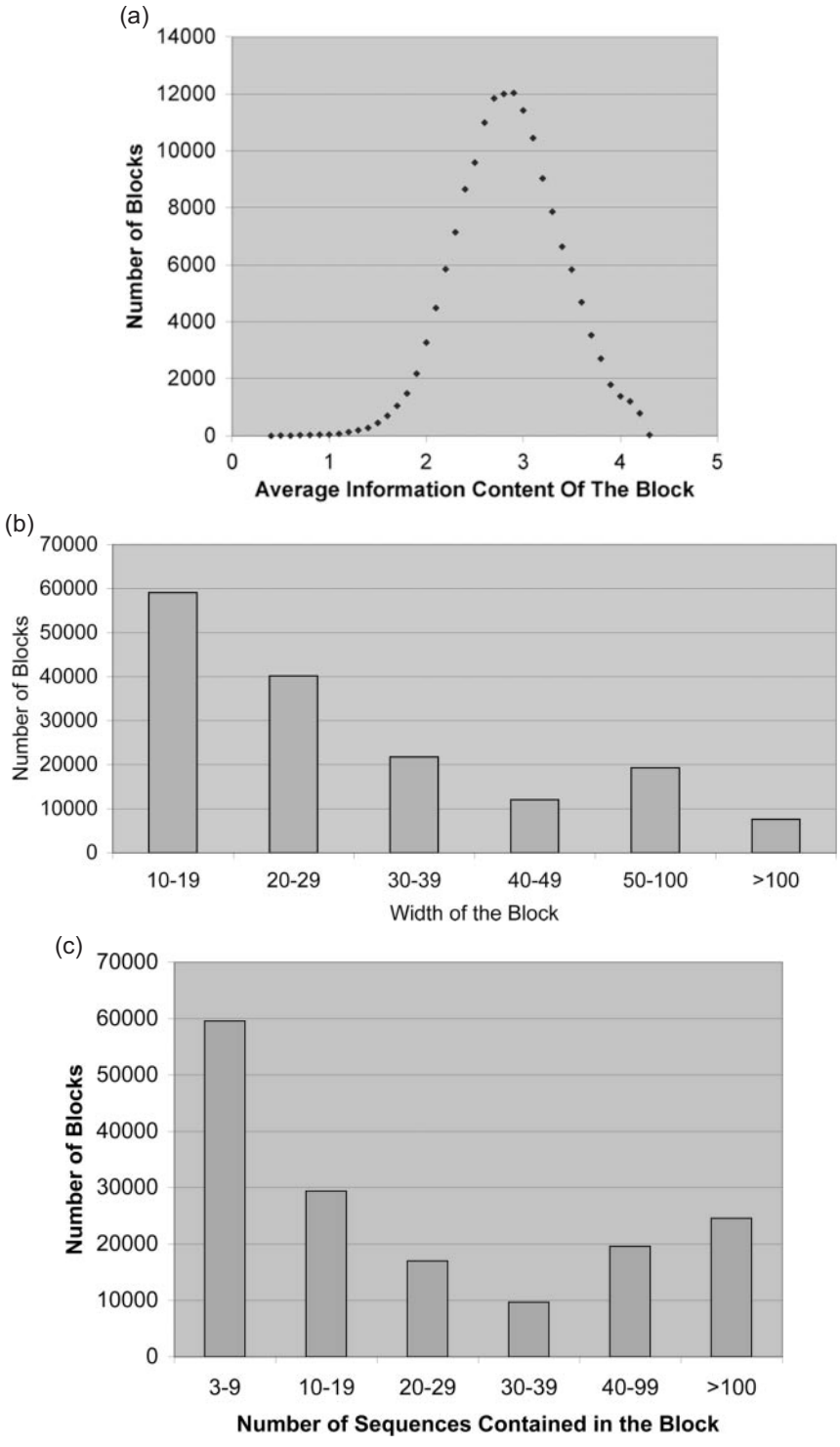

Figure 4. Statistics of the current eBLOCKs database. (**a**) The distribution of the average information content for the blocks. (**b**) The distribution of block width. (**c**) The distribution of the number of sequences contained in the blocks.

databases can be used as the building set if we desire to incorporate the most recent information from the large scale sequencing projects. Alternatively, a more restricted collection of sequences could be used as the source sequences in order to focus on a particular subset of sequences. In fact, eBLOCKs has been successfully applied to a set of signal transduction proteins and has generated a database called eSIGNAL (J. Alexander, unpublished data). In this work, the eMATRIX and eMOTIF algorithms were used to derive PSSMs and motifs from the eBLOCKs database and the corresponding eMATRIX search and eMOTIF search tools were used in sequence searches. Alternatively, one can use the IMPALA package (30) to derive PSSMs from the eBLOCKs database and to search sequences.

ACKNOWLEDGEMENTS

We thank Professor Trevor Hastie for helpful discussions about *K*-means algorithm. We thank Jessica Shapiro for careful review of the manuscript. We would like to thank TimeLogic for their donation of time and equipment that made this research possible. This work was supported by a grant from NHGRI HG02235-07.

REFERENCES

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
2. Altschul,S.F. and Gish,W. (1996) local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
3. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
4. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
5. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L., Studholme,D.J., Yeats,C. and Eddy,S.R. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
6. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
7. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
8. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P., Uddin,A. and Zygouri,C. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
9. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
10. Gracy,J. and Argos,P. (1998) Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics*, **14**, 174–187.
11. Gracy,J. and Argos,P. (1998) Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities. *Bioinformatics*, **14**, 164–173.
12. 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.
13. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P., Copley,R.R., Courcelle,E., Das,U., Durbin,R., Falquet,L., Fleischmann,W., Griffiths-Jones,S., Haft,D., Harte,N., Hulo,N., Kahn,D., Kanapin,A., Krestyaninova,M., Lopez,R., Letunic,I., Lonsdale,D., Silventoinen,V., Orchard,S.E., Pagni,M., Peyruc,D., Ponting,C.P., Selengut,J.D., Servant,F., Sigrist,C.J., Vaughan,R. and Zdobnov,E.M. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
14. Krause,A., Stoye,J. and Vingron,M. (2000) The SYSTERS protein sequence cluster set. *Nucleic Acids Res.*, **28**, 270–272.
15. Yona,G., Linial,N. and Linial,M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **28**, 49–55.
16. Kriventseva,E.V., Servant,F. and Apweiler,R. (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Res.*, **31**, 388–389.
17. Vlahovicek,K., Kajan,L., Murvai,J., Hegedus,Z. and Pongor,S. (2003) The SBASE domain sequence library, release 10: domain architecture prediction. *Nucleic Acids Res.*, **31**, 403–405.
18. Huang,H., Xiao,C. and Wu,C.H. (2000) ProClass protein family database. *Nucleic Acids Res.*, **28**, 273–276.
19. Sasson,O., Vaaknin,A., Fleischer,H., Portugaly,E., Bilu,Y., Linial,N. and Linial,M. (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res.*, **31**, 348–352.
20. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
21. Orengo,C.A., Pearl,F.M. and Thornton,J.M. (2003) The CATH domain structure database. *Methods Biochem. Anal.*, **44**, 249–271.
22. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
23. 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.
24. Abarbanel,R.M., Wieneke,P.R., Mansfield,E., Jaffe,D.A. and Brutlag,D.L. (1984) Rapid searches for complex patterns in biological molecules. *Nucleic Acids Res.*, **12**, 263–280.
25. Henikoff,S. (1996) Scores for sequence searches and alignments. *Curr. Opin. Struct. Biol.*, **6**, 353–360.
26. Wu,T.D., Nevill-Manning,C.G. and Brutlag,D.L. (1999) Minimal-risk scoring matrices for sequence analysis. *J. Comput. Biol.*, **6**, 219–235.
27. Nevill-Manning,C.G., Sethi,K.S., Wu,T.D. and Brutlag,D.L. (1997) Enumerating and ranking discrete motifs. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 202–209.
28. Nevill-Manning,C.G., Wu,T.D. and Brutlag,D.L. (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.
29. Wu,T.D., Nevill-Manning,C.G. and Brutlag,D.L. (2000) Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics*, **16**, 233–244.
30. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1111.