222–225 Nucleic Acids Research, 1997, Vol. 25, No. 1

Recent enhancements to the Blocks Database servers

Jorja G. Henikoff, Shmuel Pietrokovski and Steven Henikoff[1,*]

Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, USA and [1]Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Al-162, Seattle, WA 98104, USA

Received September 27, 1996; Accepted October 8, 1996

ABSTRACT

The Blocks Database contains multiple alignments of conserved regions in protein families which can be searched by e-mail (blocks@blocks.fhcrc.org) and World Wide Web (http://blocks.fhcrc.org/) servers to classify protein and nucleotide sequences. Recent enhancements to the servers include: (i) improved calculation of position-specific scoring matrices from blocks; (ii) availability of the Prints protein fingerprint database for searching in Blocks format; (iii) a representative sequence biased towards the Blocks of a protein family; (iv) a tree constructed from the Blocks of a protein family; (v) links to related World Wide Web pages for a family; and (vi) the new Local Alignment of Multiple Alignments (LAMA) method to search a block against a database of blocks.

INTRODUCTION

The Blocks Database (1,2) represents documented families of protein sequences by ungapped multiple alignments of their conserved regions called 'blocks' (3). A family may be represented by a single block, or by several, the average is between three and four (Fig. 1A). The Blocks Database is constructed automatically from groups of proteins known to be related. The spaced triplet algorithm of Smith *et al.* (4) is used to generate motifs which are then merged, extended, and assembled into a representative set by the MOTOMAT algorithm (1). Other motif-finders can be used with MOTOMAT. For instance, the Block Maker server (5) makes one set of blocks using this procedure and a second set using motifs generated by Gibbs sampling (6). Similarly, while the protein families documented in PROSITE (7) provide the basis for the Blocks Database, only the list of sequences belonging to each family is used, so that other sources of such lists could also be used.

The Blocks Database was designed to classify biological sequences. Protein or nucleotide sequences can be searched against blocks, which are converted to position-specific scoring matrices (PSSMs) for this purpose. Comparison takes into account both local similarity between the query and family, represented by a single block, and global similarity, represented by the entire set of blocks for families with more than one conserved region (8). Intervening non-conserved regions are ignored, so that the comparison concentrates on the regions characteristic of each documented family. Searching the Blocks Database can be more sensitive and selective than searching the sequence databases when the query belongs to a family that has only a few conserved regions in common (3,8).

IMPROVED POSITION-SPECIFIC SCORING MATRICES FROM BLOCKS

When a query sequence is compared with a database of blocks, it is slid along each block and every possible alignment with each block is scored. A PSSM, sometimes called a profile (9), is computed from each block for this purpose. A PSSM has 20 rows, one for each amino acid, and as many columns as there are positions in the block. Each column and row entry in a PSSM contains a numerical score for the alignment of that block column with that amino acid in the query sequence. The total alignment score is then summed over all block columns.

Computing PSSM scores is an area of active research (10–13), and those used by the Blocks Searcher were improved in early 1996 to take advantage of this work. The current scores are log-odds ratios which incorporate position-based pseudo-counts computed from the substitution probabilities underlying the BLOSUM series of scoring matrices (14). The addition of pseudo-counts to the observed counts of amino acids in blocks helps compensate for possible inadequate sampling of family members. We have shown these scores to be an improvement over the former scores, which were based on odds ratios, as well as over other methods in general use (12).

SEARCHING THE PRINTS DATABASE IN BLOCKS FORMAT

The Blocks Database depends on documented groups of protein families. While we continue to depend on PROSITE (7) for this documentation, we have begun to supplement it. The Prints protein fingerprint database (15) is very similar in concept to the Blocks Database, although it is constructed differently (16). Groups of conserved motifs for a protein family are excised semi-manually from sequence alignments and used as fingerprints. Additional family members are sought by scanning the protein databases with these fingerprints.

A copy of the Prints Database in Blocks Database format can now be searched by the Blocks WWW server with links to the

*To whom correspondence should be addressed. Tel: +1 206 667 4515; Fax: +1 206 667 5889; Email: steveh@howard.fhcrc.org

A

```
ID   LIPASE_GDXG_HIS; BLOCK
AC   BL01173A; distance from previous block=(66,343)
DE   Lipolytic enzymes "G-D-X-G" family,  histidine.
BL   HGG motif; width=15; seqs=6; 99.5%=698; strength=1419
 BAH_STRHY (  67) RTLLYLHGGSYALGS 100
 EST_ACICA (  73) QLIFHIHGGAFFLGS  83
LIP2_MORSP ( 159) AAMLFFHGGGFCIGD 100
YBAC_ECOLI (  85) ATLFYLHGGGFILGN  76
LIPS_HUMAN ( 344) SLIVHFHGGGFVAQT  77
  LIPS_RAT ( 343) ALVVHIHGGGFVAQT  75
//
ID   LIPASE_GDXG_HIS; BLOCK
AC   BL01173B; distance from previous block=(18,18)
DE   Lipolytic enzymes "G-D-X-G" family,  histidine.
BL   YEA motif; width=25; seqs=6; 99.5%=1208; strength=1527
 BAH_STRHY ( 100) VLALHYRRPPESPFPAAVEDAVAAY 100
 EST_ACICA ( 106) VIHVDYPLAPEHPYPEAIDAIFDVY  94
LIP2_MORSP ( 192) VVSVDYRMAPEYPAPTALKDCLAAY  85
YBAC_ECOLI ( 118) VIGIDYTLSPEARFPQAIEEIVAAC  86
LIPS_HUMAN ( 377) IISIDYSLAPEAPFPRALEECFFAY  58
  LIPS_RAT ( 376) IISIDYSLAPEAPFPRALEECFFAY  58
//
ID   LIPASE_GDXG_HIS; BLOCK
AC   BL01173C; distance from previous block=(6,10)
DE   Lipolytic enzymes "G-D-X-G" family,  histidine.
BL   GSG motif; width=19; seqs=6; 99.5%=1007; strength=1483
 BAH_STRHY ( 131) GCPPGRVTLAGDSAGAGLA  88
 EST_ACICA ( 137) GIKPKDIIISGDSCGANLA 100
LIP2_MORSP ( 227) GASPSRIVLSGDSAGGCLA  77
YBAC_ECOLI ( 153) QINMSRIGFAGDSAGAMLA  93
LIPS_HUMAN ( 412) GSTGERICLAGDSAGGNLC  66
  LIPS_RAT ( 411) GSTGERICLAGDSAGGNLC  66
//
```

B

```
>BL01173 LIPS_HUMAN with embedded consensus blocks
mdlrtmtqslvtlaedniaffssqgpgetaqrlsgvfagvreqalglepalgrllgvahlfdldpetpangyrslvhtarcclahllhksryvasnrrs
iffrtshnlaeleaylaaltqlralvyyaqrllvtnrpgvlffegdegltadflreyvtlhkgcfygrclgfqftpairpflqtisiglvsfgehykrn
etglsvaasslftsgrfaidpelrgaeferitqnldvhfwkafwnitemevlsslanmasatvrvsrllslppeafempltadptltvtispplahtgp
gpvlvrlisydlregqdseelssliksngqrslelwprpqqaprsrALILHFHGGGFVLGSsrshepylkswaqelgapVISIDYRLAPEAPFPQAIED
CFAAYcwaikhcallGCTPSRICLAGDSAGGNLAftvalraaaygvrvpdgimaaypatmlqpaaspsrllslmdpllplsvlskcvsayagaktedhs
nsdqkalgmmglvrrdtalllrdfrlgasswlnsflelsgrksqkmsepiaepmrrsvseaalaqpqgplgtdslknltlrdlslrgnsetssdtpems
lsaetlspstpsdvnfllppedageeaeaknelspmdrglgvraafpegfhprrssqgatqmplysspivknpfmspllapdsmlkslppvhivacald
pmlddsvmlarrlrnlgqpvtlrvvedlphgfltlaalcretrqaaelcverirlvltppagagpsgetgaagvdggcggrh
```

C

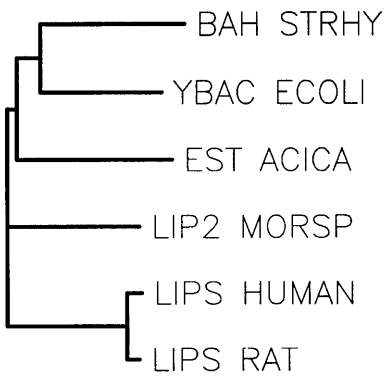

Figure 1. Blocks from v. 9.1 of the Blocks Database representing the lipolytic 'G-D-X-G' enzyme family. (**A**) The blocks start with descriptive ID, AC and DE lines modified from PROSITE entry PS01173. The AC line contains the block name (BL01173A, B, C) and indicates the minimum and maximum distances among all the sequences in the block from the preceding block. In this example, the A block starts between residues 66 and 343 for the six sequences, the B block is separated from the A block by 18 residues for each sequence, and the C block from the B block by from 6–10 residues. The BL line contains information specific to the Blocks Database, including the block width and number of sequences. The 99.5% number is a calibration score used by the BLIMPS searching program for normalizing search scores. The strength is similarly derived from a calibration search and represents the median normalized true positive score in a search of the block against SWISS-PROT. The multiply aligned sequence segments follow the BL line. Each alignment line lists the sequence name, starting position, amino acids and a sequence weight computed using the position-based method (29). The sequence weights have been normalized so that the most distant sequence in each block is assigned weight 100. (**B**) The COBBLER sequence derived from the blocks in (A). Consensus residues for each block were embedded in LIPS_HUMAN, a sequence closest to the consensus. The embedded residues are shown in upper case and underlined. (**C**) Neighbor-joining tree derived from the blocks in (A).

Prints WWW server. While many of the families represented in Prints are also in the Blocks Database, the Prints blocks may differ. Prints also contains families not represented in Blocks, so users are encouraged to search both databases.

AN EMBEDDED CONSENSUS SEQUENCE FOR SEARCHING

Multiple alignment information represented by the blocks for a family provides a simple strategy for improving similarity searches of sequence databases. Consensus residues obtained from the blocks are embedded into a family member to bias it towards the consensus in the family's conserved regions (17). These COBBLER (COnsensus Biasing By Locally Embedding Residues) sequences are provided for each family in the Blocks Database as part of the 'Get Block' request and can be used with standard searching programs such as BLAST (18) and FASTA (19) to improve detection of distant family members in the sequence databases. Figure 1B shows the COBBLER sequence for the lipolytic 'G-D-X-G' enzyme family (BL01173).

A FAMILY TREE FROM BLOCKS

The 'Get Block' request now returns a tree made from the block alignments for the family. The tree is made by the CLUSTALW program (20) from the alignments in the blocks using the neighbor-joining method (21), and is useful for inferring similar function. It is based on a matrix of distances between all pairs of block sequence segments. For requests from the Blocks e-mail server, the text 'treefile' is returned. It can be displayed by programs such as 'drawgram' from the PHYLIP (22) package, or by other programs that recognize the tree format used by

A

block 1	from:to		block 2	from:to	alignment length		Z-score	expected number
BL00122D	19 : 41	and	BL00708B	10 : 32	(23)	score	37 (6.0	1.2e-01)
BL00122D	23 : 39	and	BL01173C	1 : 17	(17)	score	49 (8.3	0.0e+00)
BL00708B	15 : 31	and	BL01173C	2 : 18	(17)	score	43 (6.8	1.8e-02)

B

```
BL00708B: DAP1_YEAST  774  IDESKIAIWGWSYGGFT
          DAP2_YEAST  668  VDPQKISLFGWSYGGYL
          PPCE_AERHY  526  TRTDRLAIRGGSNGGLL
          PPCE_PIG    543  TSPKRLTINGGSNGGLL
          PTRB_ECOLI  521  GSPSLCYAMGGSAGGML
          YL31_CAEEL  526  ANRSEVAVMGGSYGGYE
          ACPH_PIG    576  FDAGRVALMGGSHGGFL
          ACPH_RAT    576  FDARRVALMGGSHGGFL
          DPP4_HUMAN  619  VDNKRIAIWGWSYGGYV
          DPP4_MOUSE  613  VDSKRVAIWGWSYGGYV
          DPP4_RAT    620  VDSKRVAIWGWSYGGYV
          PPCE_FLAME  545  TSKEYMALSGRSNGGLL
          PPCF_FLAME  545  TSKDYMALSGRSNGGLL

BL01173C: BAH_STRHY   132  CPPGRVTLAGDSAGAGL
          EST_ACICA   138  IKPKDIIISGDSCGANL
          LIP2_MORSP  228  ASPSRIVLSGDSAGGCL
          YBAC_ECOLI  154  INMSRIGFAGDSAGAML
          LIPS_HUMAN  413  STGERICLAGDSAGGNL
          LIPS_RAT    412  STGERICLAGDSAGGNL
```

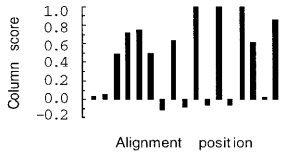

Figure 2. Sequence similarity between serine catalytic sites from different enzymes. (**A**) LAMA output for block-to-block search of the Blocks Database v9.1. BL00122: carboxylesterases type-B; BL00708: prolyl oligopeptidases; BL01173: lipolytic 'G-D-X-G' enzymes. The expected number is the number of times a score is expected to occur by chance for a blocks database of this size (3300 entries). (**B**) The aligned block regions of the prolyl oligopeptidases and lipolytic 'G-D-X-G' enzymes. Each sequence segment is labeled by its SWISS-PROT entry name and start position. The catalytic serines are outlined. The scores between the amino acid distributions are shown below the corresponding columns.

PHYLIP. For requests from the Blocks WWW server, there are options to display the tree. The tree made from the lipolytic 'G-D-X-G' enzyme family blocks is shown in Figure 1C.

PROTEIN FAMILY WWW LINKS

One of the advantages of basing the Blocks Database on PROSITE has been the link to PROSITE's documentation. This documentation can be immediately accessed when a hit is detected in the Blocks Database. With the expansion of the WWW, some researchers are building WWW pages that contain a more comprehensive set of documentation for their family of interest, including graphics, meeting notifications, etc. We have started an effort to promote the development of more WWW resources (23) of this type. The Blocks Searcher now includes links to related WWW pages with the 'Get Block' request.

SEARCHING THE BLOCKS DATABASE WITH BLOCKS

Comparing sequences with blocks and other types of multiple alignments can be more sensitive than sequence-to-sequence comparisons (9,17). We have advanced one step further, developing a method for comparing multiple alignments with multiple alignments (24). Each multiple alignment is treated as a sequence of amino acid distributions. Multiple alignments can then be aligned with each other by using an appropriate measure for scoring the similarity between amino acid distributions (analogous to the use of an amino acid substitution matrix in sequence-to-sequence alignment). We termed the method LAMA (Local Alignment of Multiple Alignments) and tested it by searching blocks against the Blocks and other multiple alignment databases. LAMA identified genuine relations between protein families beyond the range of sequence-to-sequence and sequence-to-block search methods (24). LAMA can be used at our WWW server for searching the Blocks and Prints Databases and for comparing individual blocks. The WWW server also provides a tool for converting user provided multiple alignments to the Blocks format required by LAMA, as well as computing the PSSM of the resulting block and a graphical logo representation of it (5,25).

The following example illustrates the method. LAMA searches of the Blocks Database identified sequence similarity between serine active sites from carboxylesterase, lipase and endopeptidase families (Fig. 2). All of these enzymes apparently cleave their substrates using a catalytic triad charge relay system. These relations were very difficult or impossible to detect with sequence-to-sequence searches, and sequence-to-blocks searches only detected the relation between the type-B carboxylesterases and lipolytic 'G-D-X-G' enzymes (Table 1).

Table 1. Sequence searching results for three related families of enzymes

Queries:	Searching SWISSPROT[a]			Searching Blocks[b]		
(no. of sequences)	carboxylesterases type-B	lipolytic'GDXG' enzymes	prolyl oligopeptidases	carboxylesterases type-B	lipolytic'GDXG' enzymes	prolyl oligopeptidases
carboxylesterases type-B (54)	–	29/432[c]	0	–	49/54	0
lipolytic 'GDXG' enzymes (8)	30/432	–	0	1/8	–	0
prolyloligopeptidases (14)	0	0	–	0	0	–

[a]Searching with the BLASTP (18) program. Identified hits had *P* values < 0.09.
[b]Searching with the BLIMPS (27) program. Only multiple hits with expectant values < 0.0054 were identified.
[c]The fraction represents the number of hits/number of all possible hits.

CURRENT VERSIONS AND USAGE

The Blocks Database searching program (26,27) were first made available by anonymous ftp in 1991. Version 9.1 of the Blocks Database contains 3300 blocks representing 906 different protein families documented in version 12 of PROSITE, and is complemented by version 12.0 of the Prints Database (15) with 2875 blocks from 550 families for a total of >1000 unique families. With a median of 12 (mean 23) sequences per family, ~40% of SWISS-PROT (28) is represented in Blocks 9.1. The BLIMPS searching program (v. 3.1) is currently used to query these databases. Both BLIMPS and the Blocks Database are freely available by anonymous ftp, however, most biologists prefer to use the Blocks Searcher e-mail service, which was initiated in the summer of 1992, or WWW service, initiated in 1994. The Blocks Searcher averages ~100 searches per day with the e-mail server handling slightly more than half of the requests.

ACCESS

Anonymous FTP site

ftp://ncbi.nlm.nih.gov/repository/blocks

E-mail server

blocks@blocks.fhcrc.org
Send the word 'help' in the subject line or as the only word in the message body.

WWW server

http://blocks.fhcrc.org/
Additional information or assistance may be obtained by sending an e-mail message to webmaster@blocks.fhcrc.org.

ACKNOWLEDGEMENTS

This work is supported by a grant from the NIH (GM29009). S.P. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation.

REFERENCES

1 Henikoff,S. and Henikoff,J.G. (1991) *Nucleic Acids Res.*, **19**, 6565–6572.
2 Henikoff,J.G. and Henikoff,S. (1996) *Methods Enzymol.*, **266**, 88–105.
3 Posfai,J., Bhagwat,A.S., Posfai,G. and Roberts,R.J. (1989) *Nucleic Acids Res.*, **17**, 2421–2435.
4 Smith,H.O., Annau,T.M. and Chandrasegaran,S. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 826–830.
5 Henikoff,S., Henikoff,J.G., Alford,W.J. and Pietrokovski,S. (1995) *Gene*, **163**, GC17–GC26.
6 Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) *Science*, **262**, 208–214.
7 Bairoch,A. (1992) *Nucleic Acids Res.*, **20**, 2013–2018.
8 Henikoff,S. and Henikoff,J.G. (1994) *Genomics*, **19**, 97–107.
9 Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.
10 Brown,M.P., Hughey,R., Krogh,A., Mian,I.S., Sjolander,K. and Haussler,D. (1993) In Hunter,L., Searls,D. and Shavlik,J. (eds), *Proc. First Int. Conf. on Intelligent Systems for Molecular Biology.* AAAI Press, Washington DC, pp. 47–55.
11 Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) *Proc. Natl. Acad. Sci. USA*, **91**,12091–12095.
12 Henikoff,J.G. and Henikoff,S. (1996) *CABIOS*, **12**, 135–143.
13 Bailey,T.L. and Gribskov,M. (1996) In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 15–24.
14 Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.
15 Attwood,T.K. and Beck,M.E. (1994) *Protein Engng*, **7**, 841–848.
16 Parry-Smith,D.J. and Attwood,T.K. (1992) *CABIOS*, **8**, 451–459.
17 Henikoff,S. and Henikoff,J.G. (1996), submitted.
18 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
19 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
20 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
21 Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.
22 Felsenstein,J. (1988) *Annu. Rev. Genet.*, **22**, 521–565.
23 Henikoff,S., Endow,S.A. and Greene,E.A. (1996) *Trends Biochem. Sci.*, **21**, in press.
24 Pietrokovski,S. (1996) *Nucleic Acids Res.*, **24**, 3836–3895.
25 Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
26 Henikoff,S., Wallace,J.C. and Brown,J.P. (1990) *Methods Enzymol.*, **183**, 111–132.
27 Wallace,J.C. and Henikoff,S. (1992) *CABIOS*, **8**, 249–254.
28 Bairoch,A. and Boeckmann,B. (1992) *Nucleic Acids Res.*, **20**, 2019–2022.
29 Henikoff,S. and Henikoff,J.G. (1994) *J. Mol. Biol.*, **243**, 574–578.