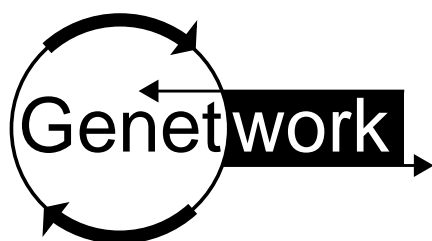


REVIEWS

- 48 Cohen, B. *et al.* (1992) *Genes Dev.* 6, 715–729
 49 Bourgooin, C., Lundgren, S.E. and Thomas, J.B. (1992) *Neuron* 9, 549–561
 50 Thor, S. and Thomas, J.B. (1997) *Neuron* 18, 397–409
 51 Curtiss, J. and Heilig, J.S. (1997) *Dev. Biol.* 190, 129–141
 52 Taira, M., Jamrich, M., Good, P.J. and Dawid, I.B. (1992) *Genes Dev.* 6, 356–366
 53 Toyama, R. *et al.* (1995) *Development* 121, 383–391
 54 Barnes, J.D. *et al.* (1994) *Dev. Biol.* 161, 168–178
 55 Taira, M., Saint-Jeannet, J-P. and Dawid, I.B. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 895–900
 56 Shawlot, W. and Behringer, R.R. (1995) *Nature* 374, 425–430
 57 Toyama, R. *et al.* (1995) *Dev. Biol.* 170, 583–593
 58 Taira, M., Otani, H., Jamrich, M. and Dawid, I.B. (1994) *Development* 120, 1525–1536
 58 Fujii, T. *et al.* (1994) *Dev. Dyn.* 199, 73–83
 60 Karavanov, A.A. *et al.* (1996) *Int. J. Dev. Biol.* 40, 453–461
 61 Toyama, R. and Dawid, I.B. (1997) *Dev. Dyn.* 209, 406–417
 62 Pfaff, S.L. *et al.* (1996) *Cell* 84, 309–320
 63 Porter, F.D. *et al.* (1997) *Development* 124, 2935–2944
 64 Sheng, H.Z. *et al.* (1996) *Science* 272, 1004–1007
 65 Zhadanov, A.B. *et al.* (1995) *Dev. Dyn.* 202, 354–364
 66 Tsuchida, T. *et al.* (1994) *Cell* 79, 957–970
 67 Appel, B. *et al.* (1995) *Development* 121, 4117–4125
 68 Glasgow, E., Karavanov, A.A. and Dawid, I.B. (1997) *Dev. Biol.* 192, 405–419
 69 Li, H. *et al.* (1994) *EMBO J.* 13, 2876–2885
 70 Sheng, H.Z. *et al.* (1997) *Science* 278, 1809–1812
 71 Johnson, R.L. and Tabin, C.J. (1997) *Cell* 90, 979–990
 72 Warren, A.J. *et al.* (1994) *Cell* 78, 45–57
 73 Arber, S. *et al.* (1997) *Cell* 88, 393–403
 74 Frangiskakis, J.M. *et al.* (1996) *Cell* 86, 59–69
 75 Higuchi, O., Amano, T., Yang, N. and Mizuno, K. (1997) *Oncogene* 14, 1819–1825
 76 Schmeichel, K.L. and Beckerle, M.C. (1997) *Mol. Biol. Cell* 8, 219–230
 77 Bach, I. *et al.* (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 2720–2724
 78 Wadman, I. *et al.* (1994) *EMBO J.* 13, 4831–4839
 79 Osada, H. *et al.* (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 9585–9589
 80 Mao, S., Neale, G.A. and Goorha, R.M. (1997) *Oncogene* 14, 1531–1539
 81 Pomiès, P., Louis, H.A. and Beckerle, M.C. (1997) *J. Cell Biol.* 139, 157–168
 82 Louis, H.A. *et al.* (1997) *J. Biol. Chem.* 272, 27484–27491
 83 Kotake, K. *et al.* (1997) *J. Biol. Chem.* 272, 29407–29410
 84 Hiraoka, J. *et al.* (1996) *FEBS Lett.* 399, 117–121
 85 Crawford, A.W., Pino, J.D. and Beckerle, M.C. (1994) *J. Cell Biol.* 124, 117–127

I.B. Dawid, J.J. Breen and R. Toyama are in the Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA.

GENETWORK



Exploring protein homology with the Blocks server

Proteins typically contain regions of conserved sequence and structure punctuated by regions showing little conservation. These more-conserved regions can be revealed as ungapped 'blocks' in alignments of related sequences. Blocks describe characteristic regions of protein families and the range of sequences they can adopt. Consequently, the identification and analysis of blocks can be valuable for understanding protein function, and computational tools based on them

have been developed and refined as sequence databanks have grown.

The Blocks Database¹ was introduced in 1991 as a method for classifying newly determined sequences. A fully automated system finds and extends motif alignments for a collection of related proteins and chooses the highest-scoring set of blocks in which the blocks are in the same order along the sequences. When applied to the current collection of 932 unique protein families documented in Prosite² (v. 9.3), a total of 3417 blocks are produced. The median block is 34 amino acids wide and includes 11 sequences. The Blocks Database is searched with either a protein or (translated) DNA sequence query. In a search, the alignment information available in a block is extracted by using recently improved methodology to convert the block to a position-specific scoring matrix (PSSM)³. High-scoring hits for single or multiple blocks representing a family are reported along with estimates that these hits occurred by chance, based on rank statistics. Searching can also be performed on the Prints Database⁴, which currently represents 800 protein families. Prints are conceptually similar to Blocks, but they are constructed using a semi-manual alignment method. The current default for searching is a composite database consisting of

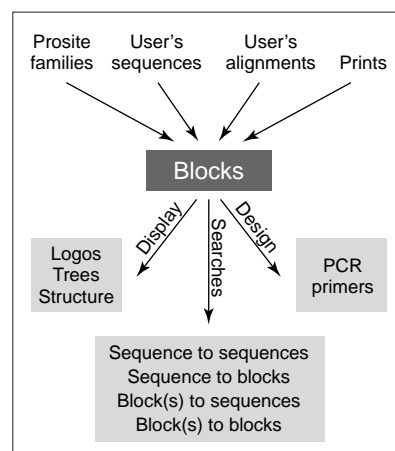


FIGURE 1. Overview of the Blocks WWW site.

Blocks supplemented with the 231 protein families in Prints that are not represented in Blocks. Both Prosite (on which Blocks is based) and Prints provide key documentation and references for each protein family. Search results can be conveniently explored via WWW links from Blocks output to Prosite, Prints, other sources of protein family documentation⁵ and individual sequence entries via Entrez⁶ and beyond.

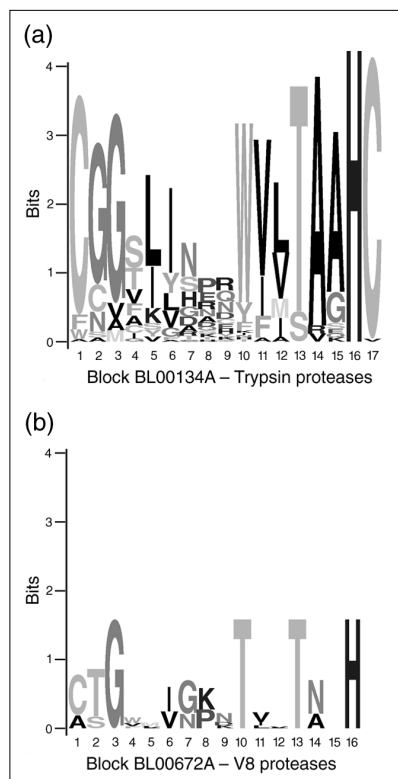


FIGURE 2. Example of LAMA output. The block containing the histidine active site residue of the trypsin family of serine proteases [represented by the logo in (a)] was used to query the Blocks Database. The hit is to the block containing the histidine active site residue of the V8 family of serine proteases (b). The alignment is across the first 16 positions of each block. The score for this hit is highly significant: it would occur by chance less than once in one million searches. In a logo, the height of each stack of residues is calculated in bits of information, which reflects conservation, with a correction made for the number of sequences in each block: because the V8 block contains only five sequences and the trypsin block contains 170 sequences, the height of the V8 stack is lower.

In recent years, the Blocks Database Searcher has been augmented with other multiple sequence alignment and searching tools (Fig. 1). Block Maker discovers blocks in related sequences provided by the user. The use of two different methods, Motif and Gibbs sampling, provides a 'reality check': if sequences are not truly related, it is unlikely that both methods will find the same block alignments. Block-Maker-generated blocks, blocks retrieved from the Blocks or Prints databases, or multiple alignments submitted by the user can be used as queries in searches of sequence databanks. Choosing the MAST (multiple alignment searching tool⁷) button sends a set of PSSMs to the San Diego Supercomputer server to search the up-to-date protein sequence databanks. Similarly, choosing BLAST or PSI-BLAST sends

the COBBLER (for consensus biasing by locally embedding residues⁸) sequence derived from blocks to these popular NCBI searchers⁹. The COBBLER sequence is designed to represent protein family information efficiently in a single artificial sequence and should be a more sensitive query for BLAST or PSI-BLAST than any real family member. LAMA (for Local Alignment of Multiple Alignments¹⁰) searches blocks against the Blocks/Prints databases, or against blocks from a user's own set of sequences. LAMA's use of concentrated multiple alignment information in both query and database makes possible the detection of subtle relationships that are beyond the range of sequence-to-sequence and sequence-to-alignment methods (Fig. 2).

The Blocks WWW site also provides options for display of multiple alignments. Blocks can be viewed as sequence logos¹¹, which are vivid summaries of multiple alignment information (Fig. 2). The use of sequence-weighted PSSMs for constructing logos avoids biasing logos in favor of redundant sequences in the alignment. A set of blocks can also be used to construct and display a neighbor-joining tree¹² for examination of possible subfamily relationships. Because blocks represent the most highly conserved regions of proteins, misaligned regions are avoided, and so trees from blocks should be of high quality. Bootstrap resampling percentages are supplied to aid in evaluating the significance of each branch. A new display option is provided by a link to the eMOTIF server¹³: if the three-dimensional structure of a protein is known, eMOTIF indicates block regions on the structure.

The utility of blocks extends beyond analysis of protein homology. The Blocks Database is the source of the widely used BLOSUM series of amino acid substitution matrices¹⁴. Blocks are also useful for designing degenerate PCR primers¹⁵ to isolate distantly related genes. A novel PCR primer design tool, called CODEHOP (consensus-degenerate hybrid oligonucleotide primer¹⁶) is the latest addition to the Blocks WWW site. CODEHOPs are designed from input blocks for amplification of unknown sequences encoding distantly related proteins. Our laboratory and others have found that CODEHOPs can efficiently amplify sequences that are too dissimilar from input sequences for use of conventional degenerate primers. The CODEHOP designer can utilize multiple alignments that are either generated at the site or are provided by the user.

The Blocks database and tools are available on the WWW at <http://blocks.fhcr.org> and by e-mail at blocks@blocks.fhcr.org (send a blank message with the word 'help').

References and URLs

1 Henikoff, S. and Henikoff, J.G. (1991) *Nucleic Acids Res.* 19, 6565-6572
2 Bairoch, A., Bucher, P. and Hofmann, K. (1997) *Nucleic Acids*

Res. 25, 217-221
3 Henikoff, J.G. and Henikoff, S. (1996) *CABIOS* 12, 135-143
4 <http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html>
5 <http://proweb.org/>
6 <http://www.ncbi.nlm.nih.gov/Entrez/>
7 <http://www.sdsc.edu/MEME/meme/website/mast.html>
8 Henikoff, S. and Henikoff, J.G. (1997) *Prot. Sci.* 6, 698-705
9 <http://www.ncbi.nlm.nih.gov/BLAST/>
10 Pietrokovski, S. (1996) *Nucleic Acids Res.* 24, 3836-3845
11 Schneider, T.D. and Stephens, R.M. (1990) *Nucleic Acids Res.* 18, 6097-6100
12 Saitou, N. and Nei, M. (1987) *Mol. Biol. Evol.* 4, 406-425
13 <http://dna.Stanford.EDU/emotif/>
14 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915-10919
15 D'Esposito, M., Pilia, G. and Schlessinger, D. (1994) *Hum. Mol. Genet.* 3, 735-740
16 <http://www.blocks.fhcr.org/codehop.html>

Shmuel Pietrokovski
pietro@muller.fhcr.org

Jorja G. Henikoff
jorja@muller.fhcr.org

Steven Henikoff*
stevh@muller.fhcr.org

**Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109-1024, USA.*

Letters to the Editor

We welcome letters on any topic of interest to geneticists and developmental biologists.

Write to:

Mark Patterson
TIG@elsevier.co.uk
 Fax 01223 464430

Trends in Genetics,
 Elsevier Trends Journals,
 68 Hills Road,
 Cambridge, UK CB2 1LA.