

**Superior performance in protein homology detection
with the Blocks Database servers**

Steven Henikoff^{1,*}, Shmuel Pietrokovski and Jorja G. Henikoff

¹Howard Hughes Medical Institute
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North
Seattle, WA 98109-1024, USA
Phone: (206) 667-4515
FAX: (206) 667-5889
E-mail: steveh@muller.fhcrc.org

*To whom correspondence should be addressed

ABSTRACT

The Blocks Database World Wide Web (<http://www.blocks.fhcrc.org>) and e-mail (blocks@blocks.fhcrc.org) servers provide tools for the detection and analysis of protein homology based on alignment blocks representing conserved regions of proteins. During the past year, searching has been augmented by supplementation of the Blocks Database with blocks from the Prints Database, for a total of 4754 blocks from 1163 families. Blocks from both the Blocks and Prints Databases and blocks that are constructed from sequences submitted to Block Maker can be used for blocks-versus-blocks searching of these databases with LAMA, and for viewing logos and bootstrap trees. Sensitive searches of up-to-date protein sequence databanks are carried out via direct links to the MAST server using position-specific scoring matrices and to the BLAST and PSI-BLAST servers using consensus-embedded sequence queries. Utilizing the trypsin family to evaluate performance, we illustrate the superiority of blocks-based tools over expert pairwise searching or Hidden Markov Models.

INTRODUCTION

The Blocks Database contains multiple alignments that represent conserved regions of proteins. To construct the database, a two-step procedure is applied to successive sets of related protein sequences: the spaced-triplet algorithm of Smith *et al.* (1) detects candidate block alignments and these are assembled by the MOTOMAT algorithm (2). Lists of related sequences are obtained from Prosite, which documents the relationships (3). The procedure is fully automated, and so related sequences submitted by users can be similarly used to construct blocks via BlockMaker, which additionally detects blocks with a Gibbs sampling algorithm (4).

The Blocks Database has been used for classification of unknown protein sequences and for detection of distant relationships. Protein or nucleotide query sequences are searched against blocks, which are converted to position-specific scoring matrices (PSSMs) for this purpose. Advances in PSSM construction (5, 6) have been incorporated into the Blocks Searcher, which reports high-scoring hits based on both a local measure of similarity for single blocks and a global measure for multiple blocks (7). LAMA (Local Alignment of Multiple Alignment) searches the Blocks and Prints Databases with BlockMaker-generated blocks or user-supplied multiple alignments to detect more subtle similarities between families (8). The Blocks World Wide Web (WWW) server (Fig. 1) includes links to Prosite, Prints, Swiss-Prot and protein family sites to aid in the interpretation of results.

The Blocks Database WWW server also provides tools for analyzing block alignments. The position of a block within each sequence is now shown on a block map. Blocks are displayed as sequence logos (9), which are vivid information-based representations of multiple alignments. Block alignments are also used to construct protein family trees, which can now be visualized with bootstrap resampling percentages, a widely-used measure for ascertaining the statistical significance of nodes. With the increasing use of trees for delineating subfamily relationships, the need for statistical support intensifies.

Previous annual reports describe how the Blocks Database is searched (10) and outline

previous enhancements to the servers (11). Here we describe additional major enhancements that were introduced during 1997 and illustrate how the use of blocks-based tools and links can provide superior performance to that obtained using other methods.

SEARCHING BLOCKS/PRINTS

Blocks are constructed automatically from protein families represented in Prosite. Prints fingerprints are very similar to blocks in that they also consist of ungapped multiple alignments representing conserved regions of proteins. However, Prints fingerprints are excised semi-manually from sequence alignments of related proteins, and additional family members are added by scanning a protein sequence databank with these fingerprints (12). Because not all families represented in Prints are present in Prosite, sequences submitted to the Blocks servers are used to search a combined Blocks/Prints database by default. Currently, the Blocks Database (version 9.3) contains 932 families from Prosite, supplemented with 231 families from Prints that are not present in Prosite, for a total of 4754 blocks from 1163 families. Optionally, the entire Prints Database, consisting of fingerprints from 800 families (in version 17) is searched. Searches are carried out similarly in either case. For the WWW server, Block hits reported in the results list are linked to Prosite (<http://expasy.hcuge.ch/sprot/prosite.html>) and other WWW documentation (13) and Prints hits are linked to the Prints WWW server (14; <http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html>) for documentation.

SEARCHING SEQUENCE DATABANKS USING BLOCK-BASED ALIGNMENTS

As sequence databanks grow, so does the background of chance similarities, making it more difficult to detect or confirm an interesting homologous relationship using a single sequence query. Multiple alignment information present in blocks can potentially improve the detection of sequence similarity in searches of sequence databanks. This is not the case for highly simplified representations of motifs, such as Prosite patterns, which performed worse overall than single sequence searches using BLAST or Smith-Waterman in comprehensive tests (15). How the multiple alignment information is represented is important, and PSSMs from Blocks using position-based sequence weights (5) and position-based pseudocounts (6) substantially outperformed BLAST and Smith-Waterman searching using single-sequence representatives (15). Consensus-embedded single sequences using the COBBLER method also outperformed single sequence representatives in comprehensive evaluations.

The Blocks Database servers provide direct links for searching the current non-redundant (nr) protein sequence databank using blocks from the Blocks or Prints database, or blocks from user-submitted sequences via BlockMaker. By selecting a WWW link, PSSMs computed from the blocks are sent to the San Diego Supercomputer Center MAST (16) server or COBBLER-embedded sequences are sent to either the NCBI BLAST (17) or PSI-BLAST (18) servers. MAST exhaustively searches multiple PSSMs against each sequence in the database. PSI-BLAST initially constructs a global PSSM from a single sequence, searches it against the database, then members detected above a threshold are added to the PSSM and the search repeated (18). Using the COBBLER-embedded consensus sequence rather than a

single sequence representative from a family may be advantageous for PSI-BLAST. Both MAST and PSI-BLAST return results with expected (E) values, maps and alignments, from PSI-BLAST in a browser window and from MAST by e-mail.

HOW EFFECTIVE ARE BLOCK-BASED METHODS FOR GLOBALLY ALIGNABLE FAMILIES?

Block-based methods were developed initially for motif detection, however, we have found that they perform well even when proteins are homologous over their full lengths. In a recent essay, Pearson (19) described a database searching strategy for sensitive detection of proteins that align globally: potentially interesting database hits below the level of statistical significance are used to search the database again, looking for significant hits to known members of the family. Pearson demonstrated this strategy for the trypsin family of serine proteases. Because this globally alignable example should be an especially challenging one for blocks-based methods, we carried out comparable searches using the four trypsin blocks from the current Blocks Database (v. 9.3 BL00134A-D) and compared the results with those of Pearson (Table 1). Best performance was obtained using MAST, which detected all true positives at higher significance levels than Pearson detected in his initial search and all but one that Pearson detected in successive searches, well above the level of significance of the first false positive. MULTIMAT (20), which uses the same PSSMs as MAST but orders results differently, scored all true positives detected by Pearson (including SP1_RARFA) above the first false positive (data not shown). Other multiple alignment-based methods available from the Blocks servers also performed better than Pearson's first search, but missed more of the ones Pearson picked out in successive searches. As might be expected, PSI-BLAST using the COBBLER-embedded sequence performed better than the COBBLER sequence using BLAST, and subsequent PSI-BLAST iterations until convergence detected more new members above the first false positive hit.

A LAMA search of the Blocks Database using the trypsin family blocks detects the V8 family of five serine proteases. One member (MPR_BACSU) was reported by Pearson not to have any statistically significant match to the three queries he tried, and two others (ETA_STAAU and ETB_STAAU) were not detected at all. Submission of the COBBLER sequence representing the V8 family (BL00672) to PSI-BLAST led to the detection of members of the S2C (or HTRA/HHOA/HHOB) family of serine proteases with $E=10^{-37}$ to 10^{-10} after the first iteration; none of the S2C proteases were detected in any of Pearson's searches. Note that these methods are not mutually exclusive, and so it may be worthwhile to use Pearson's strategy after carrying out sequence database searches with MAST and COBBLER/PSI-BLAST.

Although only comprehensive tests can be used to assess overall performance, it is interesting that best performance for the globally-alignable trypsin family was for a purely ungapped local method (MAST with the set of four PSSMs from the Blocks Database). PSSM-SWAT did less well than MAST, even though it uses the same PSSMs (embedded in a single sequence) but carries out Smith-Waterman alignments. PSI-BLAST also did less well than MAST, even though it uses essentially the same PSSM construction strategy (18) but seeks gapped alignments like SWAT and can use multiple alignment information over the

whole length of the query. It is sometimes assumed that global methods, such as Hidden Markov Models (HMMs) (21), are necessary in order to efficiently capture multiple alignment information over the full length of a protein domain; however, this assumption is clearly incorrect. Comparison of the ability of the Blocks Database to the HMM-based Pfam database (22) to classify each of the sequences in Table 1 not present in the trypsin blocks or HMMs shows that all were correctly classified with high confidence by Block Searcher, yet none were classified as trypsin-related proteins by the HMMer searching program on the Pfam server (Table 2). HMMer even failed to classify a sequence present in the Pfam HMM itself (PRTZ_BOVIN). The poor performance of Pfam-HMMer for this family is not attributable to insufficient alignment information, because 273 sequences were used to construct the HMM. Rather, poor performance is attributed to the presence of extensive misaligned or unalignable regions within the HMM's global alignment. Misaligned sequences reduce specificity of the resulting model, and so searching performance suffers (15). Such extensive misalignments are not present in the blocks used to construct trypsin PSSMs.

We anticipate that as sequence databanks grow and protein families expand, the well-established efficiency of Block-based methods for capturing alignment information will continue to increase in popularity.

ACCESS

The Blocks Database is distributed as a flat text file containing the individual block entries via anonymous ftp from: <ftp://ncbi.nlm.nih.gov/repository/blocks>. Blocks Database searches are performed via e-mail by submitting a DNA or protein sequence in FASTA or other common format to blocks@blocks.fhrc.org. BlockMaker may be used via e-mail by submitting a set of related sequences in a common format to blockmaker@blocks.fhrc.org. For either server, send the word 'help' in the subject line or as the only word in the message body. The Blocks WWW server at <http://blocks.fhrc.org/> implements all of the routines described in this article, which should be cited when the Blocks Database servers are used.

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. Smith, H. O., Annau, T. M. and Chandrasegaran, S. (1990) *Proc. Natl. Acad. Sci. USA*, **87**,826-830.
2. Henikoff, S. and Henikoff, J. G. (1991) *Nucleic Acids Res.*, **19**,6565-6572.
3. Bairoch, A., Bucher, P. and Hovmann, K. (1997) *Nucleic Acids Res.*, **25**,217-221.
4. Neuwald, A. F., Liu, J. S. and Lawrence, C. E. (1995) *Prot. Sci.*, **4**,1618-1632.

5. Henikoff, S. and Henikoff, J. G. (1994) *J. Mol. Biol.*, **243**,574-578.
6. Henikoff, J. G. and Henikoff, S. (1996) *CABIOS*, **12**,135-143.
7. Henikoff, J. G. and Henikoff, S. (1996) *Meth. Enzymol.*, **266**,88-105.
8. Pietrokovski, S. (1996) *Nucleic Acids Res.*, **24**,3836-3845.
9. Schneider, T. D. and Stephens, R. M. (1990) *Nucleic Acids Res.*, **18**,6097-6100.
10. Pietrokovski, S., Henikoff, J. G. and Henikoff, S. (1996) *Nucleic Acids Res.*, **24**,197-201.
11. Henikoff, J. G., Pietrokovski, S. and Henikoff, S. (1997) *Nucleic Acids Res.*, **25**,222-225.
12. Attwood, T. K. and Beck, M. E. (1994) *Protein Engineering*, **7**,841-848.
13. Henikoff, S., Endow, S. A. and Greene, E. A. (1996) *Trends Biochem. Sci.*, **21**,444-445.
14. Attwood, T. K., Beck, M. E., Bleasby, A. J., Degtyarenko, K., Michie, A. D. and Parry Smith, D. J. (1997) *Nucleic Acids Res.*, **25**,212-217.
15. Henikoff, S. and Henikoff, J. G. (1997) *Prot. Sci.*, **6**,698-705.
16. Bailey, T. L. and Gribskov, M. (1997) *J. Comput. Biol.*, **4**,45-59.
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.*, **215**,403-410.
18. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Aneng, Z., Miller, W. and Lipman, D. J. (1997) *Nucleic Acids Res.*, **25**,3389-3402.
19. Pearson, W. R. (1997) *CABIOS*, **13**,325-332.
20. Henikoff, S., Henikoff, J. G., Alford, W. J. and Pietrokovski, S. (1995) *Gene*, **163**,GC17-GC26.
21. Eddy, S. R. (1996) *Curr. Opin. Struct. Biol.*, **6**,361-365.
22. Sonnhammer, E. L., Eddy, S. R. and Durbin, R. (1997) *Proteins: Struct. Funct. Genet.*, **28**,405-420.

Table 1. Detection of trypsin family members^a

Swiss-Prot ID ^b	% Identity	SSEARCH	Available from the Blocks WWW server				
			PSSM/MAST	COB/BLAST	PSI-BLA ^c	PSI-BLA ^d	PSSM/SWAT
TRYA_DROME	42.1	10 ⁻³¹	[10 ⁻³¹]	10 ⁻²⁹	[10 ⁻⁶⁶]	[10 ⁻⁷⁴]	[10 ⁻³⁷]
ACRO_PIG	35.7	10 ⁻²⁶	[10 ⁻³⁴]	10 ⁻³⁶	[10 ⁻⁶⁸]	[10 ⁻⁷⁴]	[10 ⁻⁴⁰]
HGF_HUMAN	31.6	10 ⁻¹⁸	10 ⁻²⁰	10 ⁻²⁵	[10 ⁻⁵¹]	[10 ⁻⁵⁶]	10 ⁻²⁵
ACH1_LONAC	33.5	10 ⁻¹⁶	10 ⁻²³	10 ⁻¹⁵	[10 ⁻⁴⁶]	[10 ⁻⁵²]	10 ⁻²³
CERC_SCHMA	26.9	10 ⁻⁶	10 ⁻⁷	10 ⁻¹²	[10 ⁻²³]	[10 ⁻²⁷]	10 ⁻¹²
CO2_HUMAN	26.1	10 ⁻⁵	10 ⁻⁹	10 ⁻¹³	[10 ⁻³³]	[10 ⁻³⁵]	10 ⁻¹⁷
CFAB_MOUSE	24.0	10 ⁻³	10 ⁻¹⁰	10 ⁻⁹	[10 ⁻³²]	[10 ⁻³⁵]	10 ⁻¹²
PRTZ_BOVIN	25.2	.01	.009	.0006	[10 ⁻²⁷]	[10 ⁻³²]	10 ⁻⁷
GSEP_BACLI	20.6	-	.0002	-	[10 ⁻⁷]	[10 ⁻¹⁹]	-
PRLA_LYSEN	21.5	-	.04	-	-	-	-
PRTB_STRGR	24.0	-	10 ⁻⁷	-	1.3	.17	.04
PRTA_STRGR	23.4	-	10 ⁻⁵	-	-	.17	-
MPR_BACSU ^e		-	10 ⁻⁵	-	[10 ⁻⁵]	[10 ⁻¹³]	.003
GLUP_STRGR ^e		-	[10 ⁻¹⁴]	-	-	-	[.007]
SFA1_STRFR ^e			10 ⁻¹⁰	-	-	.1	.07
SFA2_STRFR ^e			.09	-	-	-	.09
SP1_RARFA ^e			-	-	-	-	-
<u>First false positives:</u>							
LORI_MOUSE		.24					
RFE_MYCLE			1.3				
APB_HUMAN				.52			
P2X1_HUMAN					2.9		
PR1C_HORVU						1.0	
VE2_HP39							1.1

^aSearches using TRYP_BOVIN and multiple alignment representations were performed on Swiss-Prot 33 using default parameters, except for PSI-BLAST, which searched the 9/19/97 version of Swiss-Prot (with default threshold set at $E=0.01$). E-values are reported only for true positive members of the trypsin family that scored above the first non-trypsin. Smith-Waterman SSEARCH results and % identities are from Pearson (19). For clarity, search results for only 3 of the highest-scoring 18 sequences reported by Pearson (with $E < 10^{-20}$) are shown. PSSM/MAST was performed with trypsin family blocks (BL00134A-D) from Blocks v. 9.3 converted to PSSMs and searched with MAST, COB/BLAST was performed using the BL00134 COBBLER-embedded sequence (derived from MCP6_MOUSE) searched with BLASTP v. 1.4, PSI-BLAST was performed with the COBBLER-embedded sequence, and PSSM/SWAT was performed using the PSSM-embedded sequence with SWAT, a Smith-Waterman searching program from Phil Green. Brackets indicate that the sequence detected contributed to the PSSM used for searching.

^bDetected by Pearson using TRYP_BOVIN, except as noted.

^cStarting with the COBBLER sequence, first iteration.

^dStarting with the COBBLER sequence, 3 iterations to convergence.

^eNot detected by Pearson using TRYP_BOVIN, but found in successive SSEARCH trials.

Table 2. Classification of trypsin family members by Blocks and Pfam

<u>Swiss-Prot ID</u>	<u>Block Searcher</u>	<u>Pfam-HMMer</u>
TRYA_DROME	[100% 10 ⁻⁹]	[+]
ACRO_PIG	[100% 10 ⁻⁷]	[+]
HGF_HUMAN	100% 10 ⁻⁹	[+]
ACH1_LONAC	100% 10 ⁻⁵	[+]
CERC_SCHMA	100% .0004	[+]
CO2_HUMAN	100% .0008	[+]
CFAB_MOUSE	100% .0004	[+]
PRTZ_BOVIN	99.8% .043	[-]
GSEP_BACLI	100% .0002	-
PRLA_LYSEN	99% .001	-
PRTB_STRGR	99.6% 10 ⁻⁶	-
PRTA_STRGR	98% 10 ⁻⁶	-
MPR_BACSU	99.9% .05	-
GLUP_STRGR	[99.9% 10 ⁻⁹]	-
SFA1_STRFR	99.7% 10 ⁻⁷	-
SFA2_STRFR	91% .001	-
SP1_RARFA	75% .01	-
ETA_STAAU (V8)	90%	-
ETB_STAAU (V8)	94%	-
STSP_STAAU (V8)	67% .019	-

Searches using single sequence queries versus the Blocks Database were performed by the WWW Block Searcher. The percentile is for the best-scoring trypsin block, and for multiple-block hits, the E-value is an independent measure that the other blocks were detected by chance. For each search performed, the combination of these two measures exceeded all false positive hits. HMMer searches were performed by the Sanger Center Pfam server (<http://www.sanger.ac.uk/Software/Pfam/>). (+) indicates a correct classification as a member of the trypsin family based on a score above the default threshold, and (-) indicates that the trypsins were not found by HMMer. Brackets indicate that the sequence detected was present in the blocks or Pfam alignment.

Figure 1. The Blocks WWW server home page.

<http://blocks.fhcrc.org/>