

# New features of the Blocks Database servers

Jorja G. Henikoff, Steven Henikoff\* and Shmuel Pietrokovski<sup>†</sup>

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

Received September 22, 1998; Accepted September 23, 1998

## ABSTRACT

**Blocks are ungapped multiple sequence alignments representing conserved protein regions, and the Blocks Database consists of blocks from documented protein families. World Wide Web (<http://www.blocks.fhrc.org>) and Email ([blocks@blocks.fhrc.org](mailto:blocks@blocks.fhrc.org)) servers provide tools for homology searching and for analyzing protein family relationships. New enhancements include a multiple alignment processor that extends the use of these tools to imported multiple alignments of families not present in the database and a PCR primer designer that implements a new strategy for gene isolation.**

## INTRODUCTION

The Blocks Database was originally introduced to aid in the family classification of proteins (1). Blocks are ungapped multiple alignments corresponding to the most conserved regions of proteins. To construct the Blocks Database, lists of family members obtained from Prosite (2) are used to find representative sets of blocks, employing a fully automated motif-finding method, which does not use Prosite patterns. Blocks v. 11.0 contains 4034 blocks representing 994 protein families. The Blocks Database can be searched for sequence similarities using protein or nucleic acid queries. For searching, the Blocks Database is augmented with 1781 blocks that correspond to the ungapped multiple alignments for 287 families from the PRINTS fingerprint database (3) that are absent from Prosite. In recent years, the Blocks Database has been enhanced with the addition of searching and analysis tools (summarized in Fig. 1), which are utilized via the World Wide Web (WWW). Previously, we described several of these enhancements, including scoring improvements, consensus and PSSM (position-specific scoring matrix)-based searching of sequence databanks, blocks-versus-blocks searching, and sequence logo and tree representations of multiple alignments (4,5). During the past year, new enhancements have been added to the Blocks Database and its companion BlockMaker server (6), including a multiple alignment processor, the novel CODEHOP PCR primer designer and links to 3D structural displays.

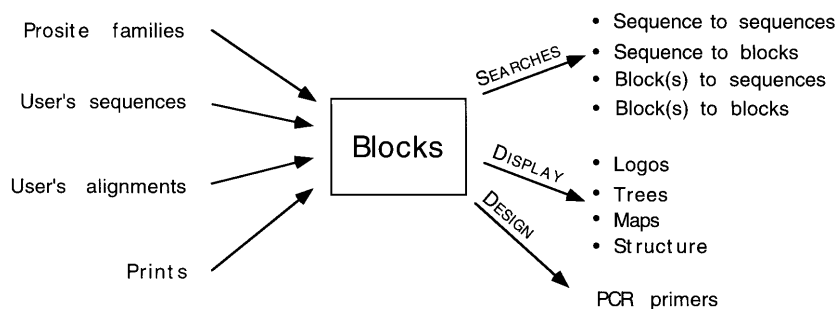
## BLOCKS-BASED ANALYSIS OF FAMILIES NOT IN THE BLOCKS DATABASE

Currently, the Blocks Database is keyed to curated catalogs of protein families: the Prosite and PRINTS databases. The families chosen for inclusion in these catalogs are chosen manually, and many known protein families are absent because of the rigors of curation. Often a biologist is interested in a protein family that has not been included in the Blocks Database. Undoubtedly, the recent explosion of data resulting from sequencing of whole genomes has contributed to the difficulty in curation. Yet not all of the missing families are newly-discovered ones: some families that have been studied and documented for many years have inexplicably been omitted from these compendiums. For example, the enzyme catalyzing the fifth step in the universal pathway for purine *de novo* biosynthesis, aminoimidazole ribonucleotide synthetase (AIRS, EC 6.3.4.13), is not represented, even though other enzymes in the pathway that are often present on multi-enzyme proteins containing AIRS are represented in the Blocks Database. Soon we expect to expand the Blocks Database by inclusion of uncurated families obtained using fully automated methods. Meanwhile, the Blocks Multiple Alignment Processor can provide tools for analysis and searching of families that are not represented in the Blocks Database. User-provided multiple alignments generated by any means are converted by the Processor to blocks for analysis.

We use AIRS as an example of a protein family that can be analyzed with blocks-based tools available on the Blocks WWW site. Protein family databases that are constructed using fully automated methods, such as ProDom (7) and Domo (8), provide multiple sequence alignments representing the AIRS domain; these alignments can be accessed via hypertext links from database entries of sequences with the AIRS domain, for example, the SwissProt sequence entry for *Escherichia coli* AIRS: PUR5\_ECOLI. Pasting the AIRS multiple alignment from Domo into the Blocks Multiple Alignment Processor window and submitting the alignment, six blocks are returned. The blocks have been carved out from the Domo multiple alignment by removal of all alignment columns with a gap (–) character in one or more sequences. The resulting blocks are retained if they are at least 10 columns wide. These blocks can be examined directly in their conventional text representation, or more informatively by displaying them as sequence logos (Fig. 2). In a logo, each position in the alignment is converted to a stack of color-coded letters representing amino acid residues, where the height of each

\*To whom correspondence should be addressed. Tel: +1 206 667 4515; Fax: +1 206 667 5889; Email: [stevh@muller.fhrc.org](mailto:stevh@muller.fhrc.org)

<sup>†</sup>Present address: Department of Molecular Genetics, The Weizmann Institute, Rehovot 76100, Israel



**Figure 1.** Overview of the Blocks Database. Input sources (left) and applications (right) of the Blocks Database.



**Figure 2.** Sequence logo of AIRS Block B and location of CODEHOP-designed primers. In the logo, the height of each amino acid is scaled in bits of information and is proportional to its degree of conservation. A pair of primers is schematically aligned with the two block segments from which they were designed. For each primer, the 5' consensus clamp is depicted as an open line (corresponding to the sequence in standard text), and the 3' degenerate core is depicted as a solid line (corresponding to the sequence in white on black letters, using the IUB-PAC code for degenerate positions). These primers were found for regions in the second AIRS block corresponding to positions 59–67 and 98–109 in SwissProt entry PUR5\_ECOLI. Default parameters were used for maximum degeneracy (128) and melting temperature of the consensus clamp (60°C). A core strictness value of 0.05 was chosen, where the range is 0 to 1 (a setting of 0 stipulates that all possible coding sequences for residues in the core region of the block are present in the pool of primers). The zebrafish codon usage table was chosen.

letter and the total height of the stack reflects the degree of conservation (6,9). Other display options are a block map, displaying the position of the block regions in each sequence, and a neighbor-joining bootstrap tree, which shows the relationships between all of the sequences.

Searching options that are available for the Blocks Database are also available for the user-submitted families, including COBBLER [Consensus Biasing By Locally Embedding Residues (10)], MAST [Multiple Alignment Searching Tool (11)] and LAMA [Local Alignment of Multiple Alignments (12)]. COBBLER-based searches are carried out by automatically embedding the blocks into a single sequence and then sending the resulting sequence to either the BLAST or PSI-BLAST server. In the case of AIRS, the six Domo-derived blocks were embedded into PUR5\_ECOLI and sent to PSI-BLAST. In addition to detection of several more AIRS-containing proteins not among the 10 bacterial and eukaryotic members present in Domo, PSI-BLAST finds likely AIRS domains in many other organisms, including archaea, and in multifunctional proteins found in eukaryotes, pinpointing the location of single or tandem AIRS domains in each one.

PSI-BLAST also identifies statistically significant sequence similarities between the AIRS domain and hydrogenases in bacteria and archaea, between the first half of the AIRS domain and bacterial selenophosphate synthetases and between AIRS and part of the previous enzyme in the pathway, FGAR synthetase. Are these true homology relationships or chance similarities? One way of deciding is to choose the MAST option, which sends position-specific scoring matrices corresponding to each of the six AIRS blocks to the MAST server (<http://www.sdsc.edu/meme/meme.2.0/website/mast.html>). In general, searching a sequence database with such a block-based query provides better separation of true from false positives than searching with a query that consists of a single sequence representing the family (10,13). MAST detects the same AIRS homologs as PSI-BLAST with expected values  $E < 10^{-11}$ , but the highest-scoring dehydrogenase is detected at only  $E = 0.4$ , the highest-scoring selenophosphate synthetase at  $E = 6.6$  and FGAR synthetase is not detected at all. The excellent separation of AIRS from non-AIRS by MAST is useful for identifying homologous proteins with the same enzymatic activities. Because the MAST results do not confirm the COBBLER/PSI-BLAST hits to non-AIRS enzymes, these

hits should be viewed with extreme caution, although it is possible that further evidence will provide confirmation of structural similarity.

### CODEHOP PCR PRIMER DESIGN

Short regions of proteins with high conservation are frequently used for the isolation of homologs in genomes of interest by designing PCR primers from blocks. Over the years, various rules of thumb have been applied to the design of degenerate primers for this purpose; however, development of systematic methods have been stymied by unknown factors, such as the unknown effect of mismatches in various positions of a primer on annealing temperature. Recently, our group introduced a new method for PCR primer design in which degeneracy is confined to the short 3' 'core', while a non-degenerate 5' 'clamp' stabilizes annealing of the core to the starting template (14). To maximize stabilization, the clamp consists of a consensus sequence that is designed from the region of the block immediately upstream of the region used to design the core. In subsequent rounds, when primer must anneal to product molecules that have incorporated the primer, high stringency priming will occur because the clamp is a single non-degenerate sequence. This differs from degenerate PCR, where low annealing temperatures are utilized to involve all of the primers in annealing to product templates that have incorporated different degenerate primers. Moreover, the use of a short degenerate core of only 11–12 bp minimizes the length of conservation needed for successful amplification, thus permitting the design of primers from blocks that are too diverged for the practical design of conventional primers. The CODEHOP (for Consensus-DEgenerate Hybrid Oligonucleotide Primers) method has been validated by the successful amplification of products that have proven challenging using conventional methods (14).

CODEHOPs are designed automatically by a program that predicts optimal primers given a set of blocks. Hypertext links to the CODEHOP designer are provided from the Blocks and PRINTS Databases, BlockMaker and the Multiple Alignment Processor. Several options are available for customizing primer design, including choice of codon usage table for the target

genome, choice of annealing temperature, which determines the length of the clamp, choices for the degree of degeneracy and stringency of matches to the block in the core region, and the ability to change the weights of input sequences to favor a subset of interest.

The use of the CODEHOP designer is illustrated for the AIRS blocks described above. Optimal CODEHOP primers are found for what appear to be the most highly conserved regions, which are found in the second block (Fig. 2).

### ACCESS

The Blocks WWW server at <http://blocks.fhcrc.org/> implements all of the routines described in this article, which should be cited when the Blocks Database servers are used. The Blocks Database can also be searched via Email by submitting a DNA or protein sequence in FASTA or other common formats to [blocks@blocks.fhcrc.org](mailto:blocks@blocks.fhcrc.org).

### ACKNOWLEDGEMENT

This work is supported by a grant from the NIH (GM29009).

### REFERENCES

- 1 Henikoff, S. and Henikoff, J.G. (1991) *Nucleic Acids Res.*, **19**, 6565–6572.
- 2 Bairoch, A., Bucher, P. and Hovmann, K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
- 3 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.
- 4 Henikoff, J.G., Pietrovski, S. and Henikoff, S. (1997) *Nucleic Acids Res.*, **25**, 222–225.
- 5 Henikoff, S., Pietrovski, S. and Henikoff, J.G. (1998) *Nucleic Acids Res.*, **26**, 309–312.
- 6 Henikoff, S., Henikoff, J.G., Alford, W.J. and Pietrovski, S. (1995) *Gene*, **163**, GC17–GC26.
- 7 Sonnhammer, E.L.L. and Kahn, D. (1994) *Protein Sci.*, **3**, 482–492.
- 8 Gracy, J. and Argos, P. (1998) *Bioinformatics*, **14**, 164–173.
- 9 Schneider, T.D. and Stephens, R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
- 10 Henikoff, S. and Henikoff, J.G. (1997) *Protein Sci.*, **6**, 698–705.
- 11 Bailey, T.L. and Gribskov, M. (1997) *J. Comput. Biol.*, **4**, 45–59.
- 12 Pietrovski, S. (1996) *Nucleic Acids Res.*, **24**, 3836–3845.
- 13 Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
- 14 Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrovski, S., McCallum, C.M. and Henikoff, S. (1998) *Nucleic Acids Res.*, **26**, 1628–1635.