

The Blocks Database - A system for protein classification

Shmuel Pietrokovski¹, Jorja G. Henikoff² and Steven Henikoff³

¹ Fred Hutchinson Cancer Research Center, Seattle, WA 98104, USA
pietro@sparky.fhcrc.org

² Fred Hutchinson Cancer Research Center, Seattle, WA 98104, USA
jorja@howard.fhcrc.org

³ Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, WA 98104, USA
steveh@howard.fhcrc.org

ABSTRACT

The Blocks Database contains multiple alignments of conserved regions in protein families. The database can be searched by e-mail and World Wide Web (WWW) servers to classify protein and nucleotide sequences.

Introduction

Many known proteins can be grouped into families according to functional and sequence similarities. The similarity of the proteins across the sequences in each family is far from uniform. While some regions are clearly conserved, others display little sequence similarity. Often the conserved regions are crucial to the protein's function, for example enzymatic catalytic sites. Such conserved regions can be used to probe an uncharacterized sequence to indicate its function (1).

The description of a protein family by its conserved regions focuses on the family's characteristic and distinctive sequence features, thus reducing noise. Databases of conserved features of protein families can be utilized to classify sequences from proteins, cDNAs and genomic DNA (2-5). An example is the Blocks Database (3), which consists of ungapped multiple alignments of short regions, called "blocks" (6). The database was constructed from sequences of protein families using a fully automated method. Searching the Blocks Database with a sequence query allows detection of one or more blocks representing a family.

Block determination

A best set of blocks representing each protein group is found automatically by the two-step PROTOMAT system (3). The first step incorporates a motif finder. Currently we use the MOTIF algorithm (7): MOTIF exhaustively evaluates spaced triplets of amino acids that are common to multiple sequences. We have also implemented a Gibbs sampling motif finder that iteratively optimizes random "seeds" for

blocks (8). The MOTIF and Gibbs algorithms generate similar block sets for the sequences used in the Blocks Database (9). The second step of the PROTOMAT system combines and refines the original blocks and assembles a best set of blocks that is consistently found in most of the sequences in the group. An example of a best set of blocks for the iron-containing alcohol dehydrogenase family is presented in Figure 1. The two-step procedure is repeated for each protein group and the results are concatenated to make a database of blocks.

Current database version

Version 8.0 of the Blocks Database consists of 2884 blocks based on 770 protein families documented in PROSITE 12.0 (5), which is keyed to Swiss-Prot 29 (10). PROSITE also supplies the documentation for each family. The distributions of number of blocks and number of sequences per family are shown in Figure 2 for BLOCKS 8.0.

Searching the Blocks Database

The BLIMPS (Blocks IMProved Searcher) program searches the Blocks Database (9). BLIMPS transforms each block into a position specific scoring matrix (PSSM), sometimes called a profile (11). Each PSSM column corresponds to a block position and contains values based on the amino acid frequencies in each position.

To prevent domination of the PSSM by a large subgroup of related sequences, each sequence segment in a block is weighted using position-based sequence weights (12). To reduce the effect of small sequence

samples, the amino acid frequencies in each PSSM position (observed counts) are supplemented with artificial “pseudo-counts”. Currently we model pseudo-counts on amino acid substitution probabilities (13, SH and JGH, unpublished results).

BLIMPS compares a query sequence with a block by sliding the PSSM over the sequence (nucleotide sequences are translated in all the frames into six amino acid sequences). For every alignment, each sequence position receives the value of its amino acid in the aligned PSSM column. These scores are summed to obtain the score of the sequence segment. This is repeated with all blocks in the database, and the top scores are saved. In addition to searching a sequence against a database of blocks, BLIMPS can search a block against a database of sequences.

Block calibration

In order to recognize scores representing genuine relationships, it is necessary to know what scores are expected by chance alone. To accomplish this, each block is calibrated by searching it against the Swiss-Prot sequence database. Two scores specific to the block are noted - the score at the 99.5% level of the true negative scores and the median of the true positive scores (14). True positive scores are scores of blocks correctly aligned with their known family members and all other scores are assumed to be true negatives.

Blocks vary in width and conservation and hence their search scores are variable too. In order to compare scores from different blocks the scores need to be normalized. The 99.5% scores are used to standardize the raw search scores. Each raw score is divided by the 99.5% score of the

blocks and multiplied by 1000. Therefore, any standardized score above 1000 is a result better than all but the top 0.5% of the true negatives.

The median of standardized scores for true positive alignments is termed “strength”. Strong blocks are more effective than weak blocks (standardized scores < 1100) at separating true positives from true negatives.

Interpreting a search result

The Blocks Database can be searched with a sequence query using the BLIMPS program on our e-mail and WWW servers. As an example, BLOCKS 8.0 was searched with a bacterial dichlorocatechol oxidase (Swiss-Prot TFDF_ALCEU) as a protein query sequence. The search output for the first three hits is shown in Figure 3.

The three best alignments in the entire search are with the blocks of the iron-containing alcohol dehydrogenase family. All three blocks align with the query sequence in the same order as the sequences represented in the blocks, that is, A->B->C. This is most easily seen in the block map. This map also shows that the distances between the three blocks representing this family fit the distances between the segments of the query that align with these blocks. For example, the distance between A and B varies from 43 to 73 in known members of this family and is 41 for the query. Therefore, the query might be a member of this family. Additional evidence concerning a family relationship comes from examination of the alignment of each query segment with the closest single member of the family. The segment aligning with Block A is closest to the segment of ADHE_ECOLI in the block. The other two segments align

best with a different member of this family (ADH2_ZYMMO).

Intuitively, it seems unlikely that three high scoring blocks would align with correct distances in between by chance alone. But how unlikely? First, the alignment with the top ranking Block C (scoring 1171) probably did not occur by chance, because such a score was seen at the 99.33 percentile level of searches with randomized queries (15). Second, Blocks A and B were detected independently of the C (anchor) block. The probability of detection of these two additional blocks by chance can be estimated based on the rank of each block alignment, the sizes of the query sequence and the database, and the observed distances between blocks [see (15) for further details]. This estimate is about 3 in ten million ("P<2.7e-07 for BL00913A BL00913B in support of BL00913C"). The two independent measures, percentile and P estimate, can be combined to provide a confidence level of less than once in 7000 searches. We conclude that the query is a member of the iron-containing alcohol dehydrogenase family.

Examining the blocks and the PROSITE documentation of the family we see that Block C contains histidine residues that are probably important for binding the ferrous ion(s) required for the enzyme activity (16). Blocks can be viewed graphically on our WWW server (9) as sequence logos (17). Logos display the different amino acids in each position, the conservation of each position and of each amino acid. In the logo for Block C (Figure 4), conserved residues are easily seen. Note that the invariant glycine in position 17 in the block is substituted by alanine in the query sequence; this illustrates the flexibility of the search system.

The second and third hits illustrate chance alignments. Both hits rank below the 60th percentile. The second hit is a marginal multiple

block hit. Even though the top ten block hits in a search are reported, one should be increasingly cautious about block alignments with low percentiles. Note also that the P estimates for blocks in support becomes less meaningful as one goes down the list, and that no P estimates are reported for single block hits.

Other uses of the Blocks Database

The automated construction and extensive data in the Blocks Database make it suitable for uses other than protein classification. The local alignments of sequence segments provided data for the BLOSUM series of amino acid substitution matrices (18). These matrices performed very well in sequence database searches (19,20). The Blocks Database was also used to test and compare different methods for weighting sequences to reduce redundancy (12).

Many blocks are made up of sequence segments with known functions such as ligand binding regions, catalytic domains and transmembranal domains (SP unpublished observations). This can be a resource for research on specific domains. For example, in studying protein nucleotide binding sites one can search for block families annotated as having such sites or for blocks containing the known signature of the sites. The blocks found can help refine the signature and even reveal unannotated sites.

Other searchable databases of protein families

PROSITE is a compilation of specific sites, patterns and profiles found in protein sequences (5). The PRINTS (4), ProDom (21) and SBASE

(22) are databases of protein motifs and domains. PRINTS and SBASE have cross references to the Blocks Database. All these databases find conserved regions by different methods and may include different groups of proteins. Therefore, different databases can provide complementary information.

Access

Anonymous FTP

<u>Location</u>	<u>Address</u>	<u>Directory</u>
USA -	ncbi.nlm.nih.gov	/repository/blocks
UK -	ftp.ebi.ac.uk	/pub/databases/blocks
Israel -	bioinformatics.weizmann.ac.il	/pub/databases/blocks
Japan -	ftp.nig.ac.jp	/pub/db/blocks

The Blocks Database is distributed as a flat text file containing the individual block entries.

The NCBI site also includes the software that we developed to construct and utilize the Blocks Database, including the BLIMPS search program.

The BlockSearch program developed by R. Fuchs for fast block searches (23) can be found at the EBI site in directories pub/software/unix and pub/software/vax.

E-mail servers

“blocks@howard.fhcrc.org” (for searching the Blocks Database)

“blockmaker@howard.fhcrc.org” (for making blocks from user supplied protein sequences)

Send the word “help” in the subject line or as the only word in the message body to obtain help files from both servers.

WWW

[“http://blocks.fhcrc.org”](http://blocks.fhcrc.org)

This site offer Blocks Database searches, block retrievals, block logos, block construction, help files and related bibliography.

Acknowledgements

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

References

1. Bork, P., Ouzounis, C. and Sander, C. (1994) *Current Opinions in Structural Biology*, **4**, 393-403.
2. Smith, R. and Smith, T. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 118-122.
3. Henikoff, S. and Henikoff, J. G. (1991) *Nucleic Acids Res*, **19**, 6565-6572.
4. Attwood, T., Beck, M., Bleasby, A. and Parry-Smith, D. (1994) *Nucleic Acids Res*, **22**, 3590-3596.
5. Bairoch, A. and Bucher, P. (1994) *Nucleic Acids Research*, **22**(17), 3583-9.
6. Posfai, J., Bhagwat, A. S., Posfai, G. and Roberts, R. J. (1989) *Nucleic Acids Res*, **17**, 2421-2435.
7. Smith, H. O., Annau, T. M. and Chandrasegaran, S. (1990) *Proc Natl Acad Sci USA*, **87**, 826-830.
8. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. (1993) *Science*, **262**, 208-214.
9. Henikoff, S., Henikoff, J. G., Alford, W. J. and Pietrokovski, S. (1995) *Gene*, **163**, GC 17-26.
10. Bairoch, A. and Boeckmann, B. (1994) *Nucleic Acids Research*, **22**(17), 3578-80.
11. Gribskov, M., McLachlan, A. and Eisenberg, D. (1987) *Proc Natl Acad Sci USA*, **84**, 4355-4358.
12. Henikoff, S. and Henikoff, J. G. (1994) *J Mol Biol*, **243**, 574-578.
13. Tatusov, R., Altschul, S. and Koonin, E. (1994) *Proceedings of the National Academy of Sciences of the United States of America.*, **91**(25), 12091-12095.
14. Henikoff, S. and Henikoff, J. G. (1995) *Methods Enzymol*, **Accepted for publication**.
15. Henikoff, S. and Henikoff, J. G. (1994) *Genomics*, **19**, 97-107.
16. Cabiscol, E., Aguilar, J. and Ros, J. (1994) *J. Biol. Chem.*, **269**, 6592-6597.
17. Schneider, T. D. and Stephens, R. M. (1990) *Nucleic Acids Res*, **18**, 6097-6100.
18. Henikoff, S. and Henikoff, J. G. (1992) *Proc Natl Acad Sci USA*, **89**, 10915-10919.
19. Henikoff, S. and Henikoff, J. G. (1993) *Proteins*, **17**, 49-61.
20. Pearson, W. (1995) *Protein Science*, **4**(6), 1145-1160.
21. Sonnhammer, E. and Kahn, D. (1994) *Protein Sci.*, **3**, 482-492.
22. Pongor, S., Hatsagi, Z., Degtyarenko, K., Fabian, P., Skerl, V., Hegyi, H., Murvai, J. and Bevilacqua, V. (1994) *Nucleic Acids Res*, **22**, 3610-3615.
23. Fuchs, R. (1994) *Computer Applications in the Biosciences*, **10**(1), 79-80.

Figure legends

Figure 1. Blocks Database format.

Each block entry is divided into header and sequences parts. The header part consists of four lines. The ID (identification) line contains the block's family short description and identifies the entry as a block type. The AC (accession) line gives the block's accession code and the minimal and maximal distances of the block from the previous block or the protein N' end. The block accession code is made up of the letters "BL" followed by the family PROSITE accession number and the individual block's letter code (A for first block, B for second etc.; blocks from single block families have no letter suffix). The DE (description) line contains the long description of the family. The short and long descriptions are taken from PROSITE. The BL (block) line gives the spaced triplet motif of the block, the block's width, number of sequences, 99.5%-level raw score and strength score (median standardized score of known true positive sequences). Each sequence line contains the sequence Swiss-Prot name, the start position of the segment, the sequence segment and the sequence weight (100 being most distant). Segments that are less than 80% similar are separated by blank lines. Each block entry ends with a "/" line. The block entries are sorted by their accession codes - each family's blocks grouped together and ordered.

The figure shows the three blocks of the iron-containing alcohol dehydrogenase family from BLOCKS 8.0 .

Figure 2. Statistics of Blocks Database version 8.0.

a. Number of blocks per family. b. Number of sequences per family.

Figure 3. Block Search output, showing the first three hits. See discussion in text.

Figure 4. A sequence logo.

Block BL00913C, shown in Figure 1, was converted to a PSSM. For every column of the PSSM, each amino acid value was represented as a letter in the stack. The vertical scale shows the conservation, in bits, of the amino acids, which are shaded according to their properties.

```

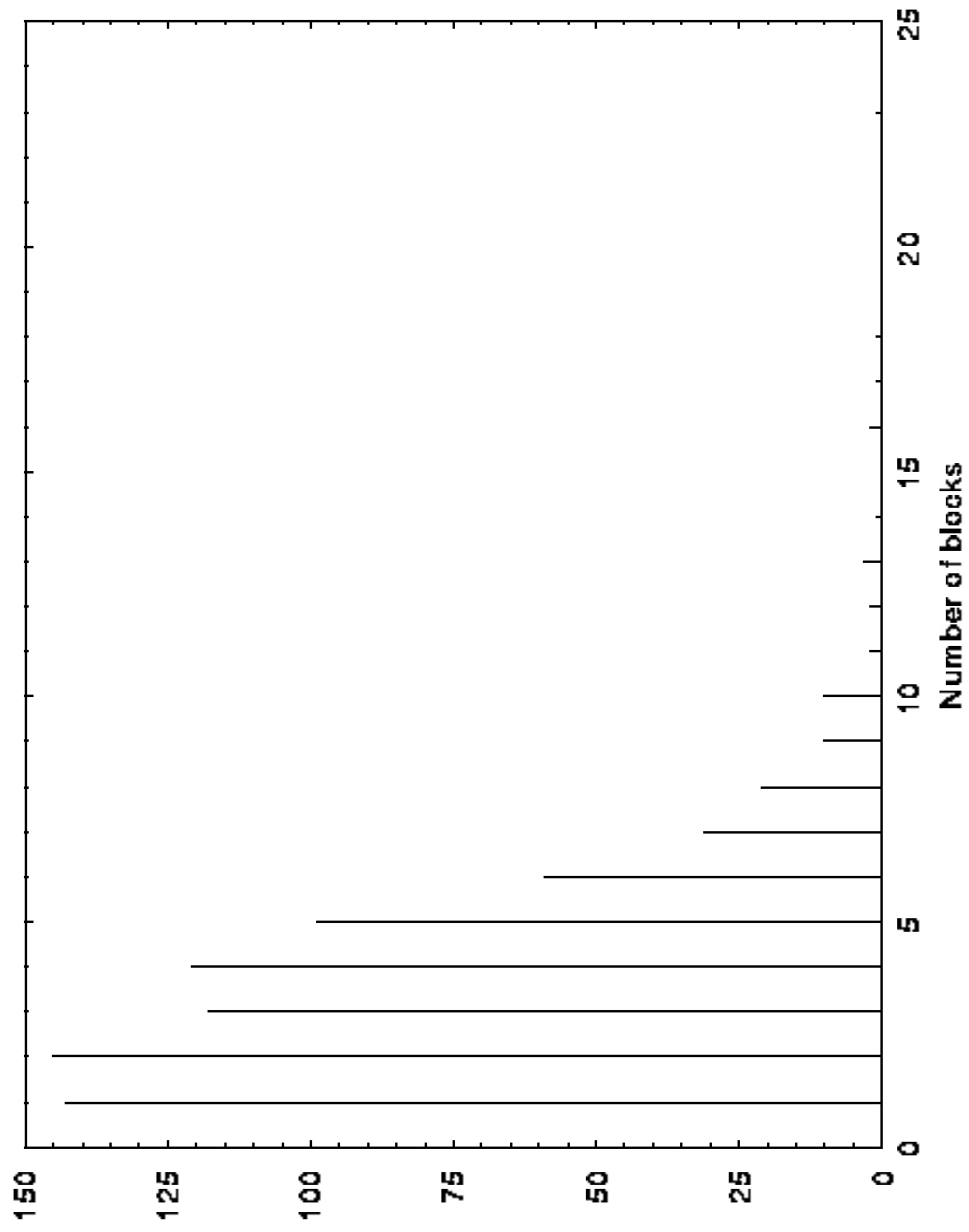
ID   ADH_IRON_1; BLOCK
AC   BL00913A; distance from previous block=(64,516)
DE   Iron-containing alcohol dehydrogenases proteins.
BL   GDK motif; width=38; seqs=11; 99.5%=703; strength=1474
ADH1_CLOAB ( 65) PDPSVETVFKGAELMRQFEPDWIIAMGGGSPIDAAKAM 50

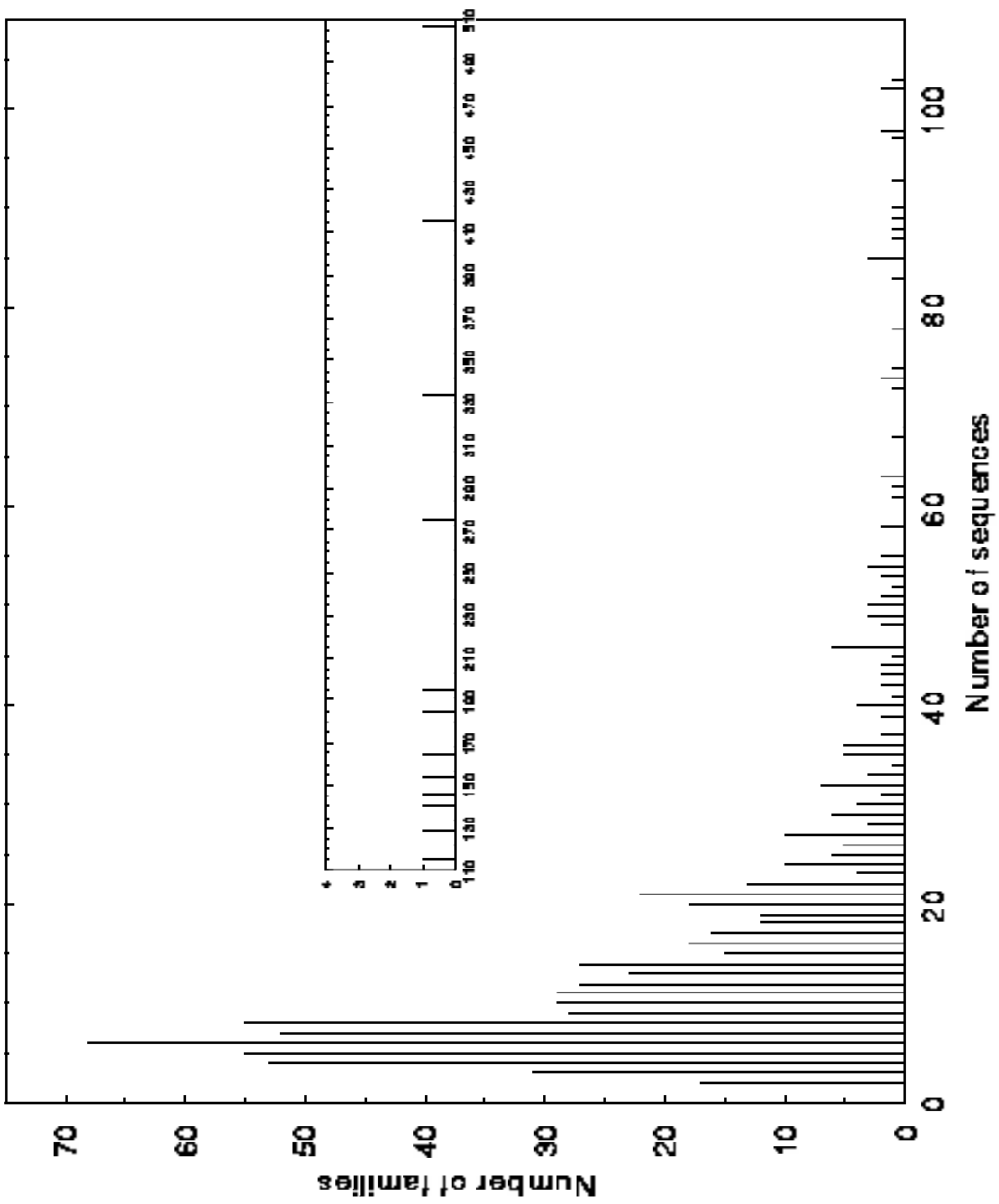
ADH2_ZYMMO ( 69) PNPTVTAVLEGLKILKDNNSDFVISLGGGSPHDCAKAI 46
ADH4_YEAST ( 71) PNPNIANVTAGLKVLEENSEIVVSIIGGSAHDNAKAI 58
ADHE_CLOAB ( 516) READLKTIKKATEEMSSFPDTIIALGGTPEMSSAKLM 100
ADHE_ECOLI ( 517) ADPTLSIVRKGAELANSFKPDVIIALGGGSPMDAAKIM 60
FUCO_ECOLI ( 70) PNPTITVVKEGLGVFQNSGADYLIAIGGGSPQDTCKAI 59
GLDA_BACST ( 68) GEASRNEVERIANIARKAEAAIVIGVGGKTLDTAKAV 73
GLDA_ECOLI ( 79) GECSQNEIDRLRGIETAQCGAILGIGGGKTLDTAKAL 82
MEDH_BACMT ( 67) PDPADTQVHEGVDFVKQENC DALVSIGGGSSHDTAKAI 58
ADHA_CLOAB ( 70) PNPRIITVVKKGIEICRENNVDLVLAIGGSAIDCSKVI 49
ADHB_CLOAB ( 70) PNPRVTVEKGVKICRENGVEVFLAIGGSAIDCAKVI 44
//
ID   ADH_IRON_1; BLOCK
AC   BL00913B; distance from previous block=(43,73)
DE   Iron-containing alcohol dehydrogenases proteins.
BL   PPD motif; width=34; seqs=11; 99.5%=679; strength=1484
ADH1_CLOAB ( 168) PDVAVVDSELAETMPKLTHTGMDALTHAIEAY 66

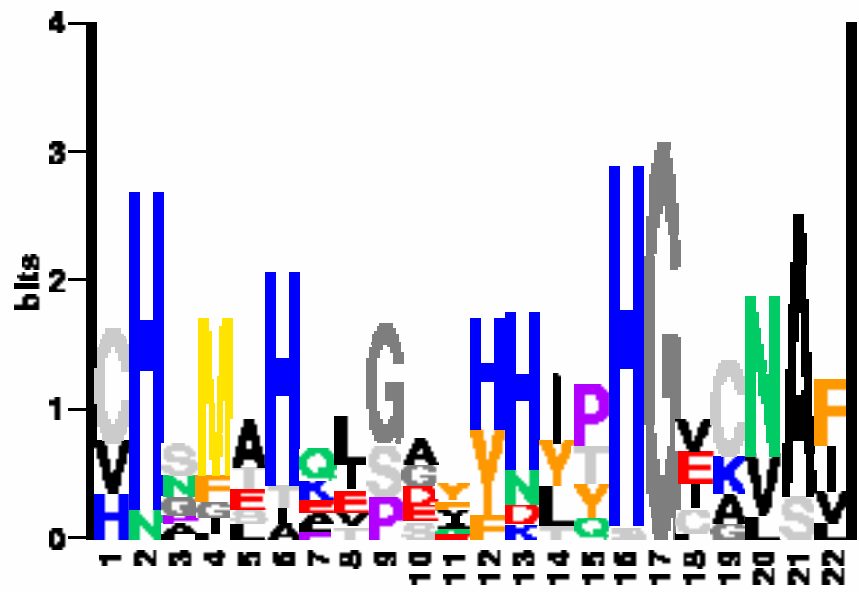
ADH2_ZYMMO ( 169) PMVSVNDPLLMVGMKGLTAATGMDALTHAFEAY 56
ADH4_YEAST ( 171) PAVAVNDPSTMFGLPPALTAATGLDALTHCIEAY 67
ADHE_CLOAB ( 627) PNMAIVDAELMMKMPKGLTAYSGIDALVNSIEAY 80
ADHE_ECOLI ( 628) PDMAIVDANLVMDMPKSLCAFGLDAVTHAMEAY 82
FUCO_ECOLI ( 172) PQVAFIDADMMDGMPALKAAATGVDALTHAIEGY 83
GLDA_BACST ( 149) PDLVLVDTKIIANAPPRLLASGIADALATWFEAR 100
GLDA_ECOLI ( 160) PNMVIVDTKIVAGAPARLLAAGIGDALATWFEAR 80
MEDH_BACMT ( 167) PTVAIVDPELMVKKPAGLTIATGMDALSHAIEAY 69
ADHA_CLOAB ( 170) PKFSVLDPTYTFTVPKNQTAAGTADIMSHTFESY 86
ADHB_CLOAB ( 170) PKFSILDPTYTYTVPNTQTAAGTADIMSHIFEVY 92
//
ID   ADH_IRON_1; BLOCK
AC   BL00913C; distance from previous block=(56,76)
DE   Iron-containing alcohol dehydrogenases proteins.
BL   HHG motif; width=22; seqs=11; 99.5%=492; strength=1428
ADHE_CLOAB ( 720) CHSMAIKLSSEHNIPSGIANAL 66

FUCO_ECOLI ( 262) VHGMHPLGAFYNTPHGVANAI 44
GLDA_BACST ( 259) HNGFTALEGEIHHLTHGEKVAF 100
GLDA_ECOLI ( 269) VHNGLTAIPDAHYYHGEKVAF 100
MEDH_BACMT ( 259) VHSISHQVGGVYKLGHCNSV 78
ADH1_CLOAB ( 258) CHSMAHKTGAVFHIPHCANAI 47
ADHE_ECOLI ( 721) CHSMAHKLGSQFHIPHGLANAL 47
ADH2_ZYMMO ( 261) VHAMAHLGGYYNLPHGVCNAV 36
ADH4_YEAST ( 263) VHALAHQLGGFYHLPHGVCNAV 41
ADHA_CLOAB ( 266) CHPMEHELSSAYDITHGVGLAI 50
ADHB_CLOAB ( 266) VHLMEHELSSAYDITHGVGLAI 49
//

```







PSSM of BL00913C (ADH_IRON_1) 11 sequences.