

Jorja G. Henikoff¹
 Shmuel Pietrokovski²
 Claire M. McCallum¹
 Steven Henikoff¹

¹Fred Hutchinson Cancer
 Research Center,
 Seattle, WA, USA

²Department of Molecular
 Genetics, Weizmann
 Institute of Science,
 Rehovot, Israel

Blocks-based methods for detecting protein homology

The most highly conserved regions of proteins can be represented as blocks of aligned sequence segments, typically with multiple blocks for a given protein family. The Blocks Database World Wide Web (<http://blocks.fhcrc.org>) and e-mail (blocks@blocks.fhcrc.org) servers provide tools to search DNA and protein queries against the Blocks+ Database of multiple alignments. We describe features for detection of distant relationships using blocks. Blocks+ includes protein families from the PROSITE, Prints, Pfam-A, ProDom and Domo databases. Other features include searching Blocks+ with the BLIMPS and NCBI's IMPALA programs, sequence logos, phylogenetic trees, three-dimensional display of blocks on PDB structures, and a polymerase chain reaction (PCR) primer design strategy based on blocks.

Keywords: Protein homology / Database searching / Sequence analysis / Chromomethylase / Polymerase chain reaction primers
 EL 3958

1 Introduction

Progress in sequencing technology has fueled exponential increases in the number of protein-coding sequences present in sequence databanks. Because new genes are typically related to previously sequenced genes, the number of documented protein families has been increasing much less rapidly. We may look forward to a time when nearly all protein-coding genes can be assigned to a protein family on the basis of database search results, providing insight into their structure and function. However, searches of sequence databanks can yield voluminous results, making it a challenge to deduce family membership from hits to many related sequences. Furthermore, a large percentage of proteins consist of multiple modules which can lead to confusion in interpreting the results of searching databases of sequences. Searching databases of protein families, in which an entry is a single family, domain or module, can minimize these problems and also provide direct links to information about families as a whole, not just about individual members.

2 The Blocks Database

The Blocks Database of conserved regions of proteins was introduced in 1991 as a method for classifying new

sequences [1]. A fully automated algorithm is used on a collection of related proteins, yielding a set of blocks, where blocks are ungapped multiple alignments of conserved regions of related sequences. When applied to the current collection of 994 unique protein families documented in PROSITE Version 14.0 [2], a total of 4034 blocks are produced. The Blocks Database is searched with either a protein or (translated) DNA sequence query. High-scoring hits for single or multiple blocks representing a family are reported along with estimates that these hits occurred by chance. The utility of blocks extends beyond sequence classification. For example, the Blocks Database is the source of the widely used BLOSUM series of amino acid substitution matrices [3]. In more recent years, other multiple sequence alignment and searching tools have been introduced (Fig. 1). For example, Block Maker discovers blocks in related sequences provided by the user [4]. The use of two unrelated block-finding algorithms, Motif [5] and Gibbs sampling [6], 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 Blocks databases or multiple alignments submitted by the user, can be used as queries in searches of sequence databanks. Choosing the MAST (Motif Alignment and Search Tool) button sends a set of blocks to the San Diego Supercomputer server [7] to search the up-to-date protein sequence databases. Similarly, choosing BLAST or PSI-BLAST sends the COBBLER (for Consensus Biasing By Locally Embedding Residues [8]) sequence derived from blocks to these popular NCBI searchers [9]. The COBBLER sequence is designed to efficiently represent protein family information in a single artificial sequence. LAMA (for Local Alignment of Multiple Alignments) [10] searches blocks against Blocks databases. LAMA uses concentrated multiple alignment information

Correspondence: Dr. Steven Henikoff, Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, P.O. Box 19024, Seattle, WA 98109, USA
E-mail: steveh@fhcrc.org
Fax: +206-667-5889

Abbreviations: aa, amino acid; CMT, chromomethylase; COBBLER, consensus biasing by locally embedding residues; CODEHOP, consensus-degenerate hybrid oligonucleotide primer; LAMA, local alignment of multiple alignments; PDB, protein data bank; PSSM, position-specific scoring matrix; WWW, World Wide Web

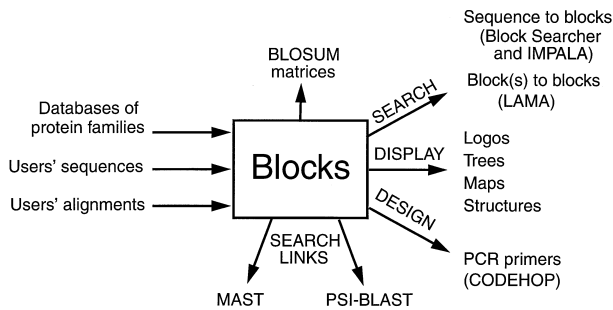


Figure 1. Overview of Blocks server functions.

in both query and database. This makes the detection of weak relationships possible.

2.1 Blocks+

The collection of protein families in PROSITE is incomplete, and other compilations are needed to expand coverage of the Blocks Database. To construct Blocks+ [11], lists of protein families from PROSITE are supplemented with additional families from the Prints [12], Pfam-A [13], ProDom [14] and Domo [15] protein family databases. Blocks for these families are computed by extracting SWISS-PROT [16] sequences documented in the source protein family databases and presenting them to the automated PROTOMAT system [1, 17]. However, to minimize redundancy, the resulting blocks for a family are added to Blocks+ only if a LAMA blocks-*versus*-blocks search [10] of them against the current database results in no significant hits. This recursive procedure yields sets of blocks extracted from Pfam-A families not found in either PROSITE or Prints, blocks from ProDom not found in the previous three databases, and blocks from Domo not found in any of the other databases. The Blocks+ Database consists of 10 070 blocks from 2235 different protein families as of October 10, 1999 (Fig. 2). Since the multiple alignments in the source family databases are not used, the alignments in Blocks+ may not coincide with them. Therefore, the LAMA blocks-*versus*-blocks searching program [10] is used to search each set of blocks in Blocks+ against blocks carved out of these source alignments [18], and World Wide Web (WWW) links are made when hits are found.

2.2 Searching Blocks Databases

The Block Searcher uses the BLIMPS searching program [4] to compare a DNA or protein query sequence with each block in the database of blocks being searched. Each block is converted to a numerical position-specific scoring matrix (PSSM) for the comparison [19]. The results for individual blocks are then analyzed to combine hits for blocks belonging to the same protein family. *E*-val-

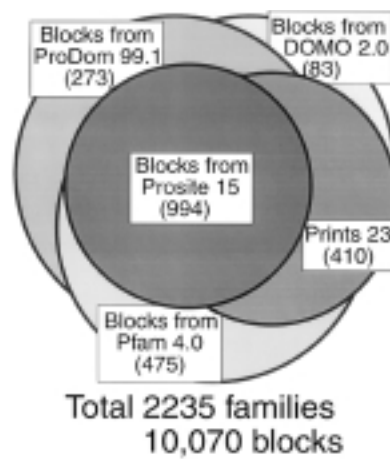


Figure 2. Composition of the Blocks+ Database (as of October 10, 1999).

ues are computed for multiple block hits using methods developed for searches of block queries against sequence databases with the MAST searching tool [7, 20]. This method requires computing the score distribution for each block, which can be done explicitly when the PSSM derived from a block contains only integers [21]. The probability of obtaining the score for the alignment with the query sequence can then simply be looked up in the score distribution. Figure 3a displays an example of the Block Searcher applied to a putative “chromomethylase”, a DNA methyltransferase homolog with a chromodomain [22]. An alternative to the Block Searcher for protein queries is the IMPALA Searcher, which has been made available for the Blocks WWW server by the BLAST group at NCBI [23]. IMPALA searches a database of PSI-BLAST PSSMs [9] which are constructed for each family in Blocks+ by searching with the COBBLER sequence to query the SWISS-PROT sequences known to belong to the family with PSI-BLAST. The COBBLER sequence is a representative sequence stretching from 10 amino acids (aa) upstream of the first block to 10 aa downstream of the last block, into which consensus residues deduced from block regions are embedded. PSI-BLAST searching is iterated until convergence, yielding a database of one PSI-BLAST PSSM for each family in Blocks+. Figure 3b shows an example of IMPALA output, which consists of the familiar BLAST output and *E*-value statistics, and includes links to the Blocks+ families hit. Unlike the Block Searcher, IMPALA may insert gaps in the alignment of the query with the blocks and may also align regions between blocks. Since the Block and IMPALA Searchers tend to report the same true positive hits but different false positives (*e.g.*, compare Fig. 3a to 3b), users who search with both and compare the results may be better able to distinguish true from false hits for challenging queries. In Fig. 3, the chromodomain is detected at *E* =

0.11 with IMPALA and $E = 0.012$ with Block Searcher, confirming that this protein, designated CMT2, is a chromomethylase. This example illustrates the importance of using a protein family database for sequence classification: CMT chromodomains are almost undetectable using standard sequence searching methods.

2.3 Visualizing blocks

Once a query sequence hits a family in the Blocks+ Database, family information including curated annotations, sequence logos, phylogenetic trees and three-dimensional structures is immediately available. Links are provided to protein family databases and to other family sites. Sequence logos [24] display the block alignments graphically and are useful for examining sequence conservation (Fig. 4a). Phylogenetic trees are made from the block alignments by ClustalW [25] and can be useful for delineating subfamily relationships (Fig. 4b). In each block position the total height is proportional to the information content of the position, and each amino acid's height is proportional to its frequency in the position. Resi-

dues are colored according to their chemical and physical properties. An increasing number of protein families are represented by one or more three-dimensional structures in the PDB database (<http://www.rcsb.org/pdb>). To map blocks onto a structure in PDB, MAST [20] is used to search PSSMs against the database of PDB sequences. Segments within corresponding PDB structures are color-coded to indicate the block that they represent. The 3-D Blocks representation can be viewed by WWW browsers with helper software that can process Rasmol [26] commands, such as Chime (<http://www.mdl.com/chemscape/chime>).

2.4 Searching databases 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. How the multiple alignment information is represented is

```
a) Block Searcher
Family Strand Blocks Anchor Combined
BL00094 C-5 cytosine-specific DNA methylase 1 6 of 6 7.8e-08 1e-30
BL00598 Chromo domain proteins. 1 1 of 1 0.013 0.012
BL01087 Radical activating enzymes proteins 1 1 of 3 1 1
BL00799 Granulins proteins. 1 1 of 8 1.5 1.5
PF01507 Phosphoadenosine phosphosulfate red 1 1 of 5 3.1 3
PF00426 Outer Capsid protein VP4 (Hemagglut 1 1 of 24 3.1 3.2
BL01230 RNA methyltransferase trmA family p 1 1 of 5 3.7 3.7
BL00439 Acyltransferases ChoActase / COT / 1 2 of 8 32 4
PR00499 Neutrophil cytosol factor 2 signatu 1 1 of 6 4.3 4.3
BL01200 Tub family proteins. 1 1 of 5 4.7 4.7
=====
>BL00094 6/6 blocks Combined E-value= 1e-30: C-5 cytosine-specific DNA methylases
proteins.
Block Frame Location (aa) Block E-value
BL00094A 0 676-696 0.0095
BL00094B 0 881-896 7.6e-05
BL00094C 0 925-936 1.2
BL00094D 0 963-982 0.00074
BL00094E 0 1091-1106 7.8e-08
BL00094F 0 1118-1127 0.032

|--- 252 amino acids---|
BL00094 AA:.....BB:..C:..DD:.....EE:..F
gi|2832630|emb|CAA <:.....AA:.....BB:..C:..DD:.....EE:F

BL00094A <->A (0,1141):675 A<->B (40,100):184
MTB1_BREEP|P10283 1 MKVLSLFSGCGGMDLGLGGF 1190 DFINGGPPCQGFSGMN
gi|2832630|emb|CAA 676 LpVLDLlySGCGGMStGLSLGA 881 gVICGGPPCQGISGYN

BL00094C B<->C (22,57):28 C<->D (21,47):26
MTSI_SPISQ|P15840 180 PKYLLMENVGAL 1270 ILEAGYGVQSQRKRAFIWA
gi|2832630|emb|CAA 925 PsYVLMENVvdI 963 IMtAGcYGLSQfRSRVFMWG

BL00094E D<->E (70,274):108 E<->F (10,31):11
MTD5_DACSA|P50185 259 RVLTPRECARLQGFPE 382 YKQIGNAVPV
gi|2832630|emb|CAA 1091 RVLTIRESARLQGFPPD 1118 YcQIGNAVAV

-----
>BL00598 1/1 blocks Combined E-value= 0.012: Chromo domain proteins.
Block Frame Location (aa) Block E-value
BL00598 0 808-829 0.013
Up to 1 repeats expected:
Other reported alignments:

BL00598 <->
YNZ8_CAEEL|P45968 40 VLQVRWRLGYGADEDTWEPEEDL
gi|2832630|emb|CAA 808 kFkVhWKGYrSDEDTWElaEeL
```

b) IMPALA Searcher

Sequences producing significant alignments:	Score (bits)	E Value
BL00094 C-5 cytosine-specific DNA methylases proteins.	126	4e-30
BL00598 Chromo domain proteins.	32	0.11
BL00242 Integrins alpha chain proteins.	32	0.18
PR00849 GLYCOSYL HYDROLASE FAMILY 58 SIGNATURE	27	3.0
PF00094 von Willebrand factor type D domain	27	3.7
PR00170 SODIUM CHANNEL SIGNATURE	27	4.0
BP02799 PROTEIN REPEAT SUBSTRATE E	27	4.1

>BL00094 C-5 cytosine-specific DNA methylases proteins.
Length = 410

Score = 126 bits (316), Expect = 4e-30
Identities = 74/370 (20%), Positives = 117/370 (31%), Gaps = 120/370 (32%)

Query: 871 SKILPLPGRVGVICGGPPCCGGISGYNRHRNVDSPLNDRNQIIVFMDIVEYLKPSYVLM 930
++I P ++ GG PCQ S + + D R I++ +P + +M

Sbjct: 58 TQIQDFPS-FDILIGGFPCQDFXSAGKQKGFG---DTRGTLFFEIERILKAYRPFKFFIM 112

Query: 931 ENVVDILRMDKGLGRYALSRLVNMRYQARLGIMTAGCYGLSQFRSRVFMGAVPNKNLP 990
ENV + DKG + L +L + Y I+ A YG+ Q R RVF+ G +++ P

Sbjct: 113 ENVKGLTTHDKGRFTKILQKHELNYGV-YLILNASDYGVFQNRERVFIVGL--DQSQP 169

Query: 991 PFPLPTH-----DVI-----VRYGLPLEFERNVVAYA 1017
+ +H D++ +Y +F ++A+

Sbjct: 170 ELTITSHIGATDSHKFKQLSNQASLFDTNKIMLVRDILEHPLDKYNCSTDFVNKLLAF- 228

Query: 1018 EGQPRKLE-----KALVLKDAISDLPHVCFISTRICNS----- 1051
G P KL + C N+

Sbjct: 229 IGHPIKLNKRLIDYRNGNSIHSWELGIKGETSDEIQFMNALIANRRKKHFGAHQDGKK 288

Query: 1052 -----GPFARLWWDTEVPTVL 1067
F L D T++

Sbjct: 289 LTIEQIKTFFEHDLLDSIMQSLITKGYLQEVNGRFNPVAGNMSFEVFKFLDPDVSVITLV 348

Query: 1068 TVPTCHSQVPSKPFQALLHPEQDRVLTIRE SARLQGFDPDYFCGCTIKERYCQIGNAVAV 1127
+ ++H + R LT RE AR+QGFPD PQF Y QIGN+V V

Sbjct: 349 SSD-----AHKIGVVHQNRIRKLTTPRECARIQGFDPDFQFHPKDSLAYKQIGNSVVP 400

Query: 1128 SVSRALGYSL 1137
V +A+ L

Sbjct: 401 PVVKAVILDL 410

Score = 49.8 bits (118), Expect = 6e-07
Identities = 11/53 (20%), Positives = 22/53 (40%)

Query: 676 LPVLDLYSGCGGMSTGLSLGAKISGVDVVTKWAVDQNTAAACKSLKLNHPNTQV 728
L V+DL++G GG+ G G++ + + + A + +N

Sbjct: 2 LKVIDLFAGVGGRLGFEQAGHELGIETACVLSSEIDKHAQTTYAMNPFHEQSQ 54

>BL00598 Chromo domain proteins.
Length = 42

Score = 32.3 bits (73), Expect = 0.11
Identities = 9/26 (34%), Positives = 15/26 (57%)

Query: 804 KNGLKFKVHWKGYRSDDEDTWELAEEL 829
K + + + + WKG+ +TWE E L

Sbjct: 7 KGQVEYMIKWKGNEMHNTWEPENL 32

Figure 3. Block Searcher and IMPALA Sercher outputs. A hypothetical *Arabidopsis thaliana* protein sequence translated from predicted exons in Genpept entry gil2832630lemb|CAA-16759.1 was used to query Blocks+ with an expected cutoff value of 5. Known true positive hits for this query sequence are BL00094 (cytosine DNA methyltransferases) and BL00598 (chromodomains), which are the top two hits for both Block Searcher and IMPALA Searcher. Alignments are shown for the top two hits. (a) Block Searcher output. Each hit consists of one or more blocks from a protein group found in the query sequence. One set of the highest-scoring blocks that are in the correct order and separated by distances comparable to the Blocks database family is selected for analysis. If this set includes multiple blocks the probability that the lower scoring blocks support the highest scoring block is reported. Maps of the database blocks and query sequence are shown: (AAA) block roughly in proportion to its width; (:) minimum distance between blocks in the database; (.) maximum distance between blocks in the database; (< >) the sequence has been truncated to fit the page. The query map is aligned on the highest scoring block. Multiple block hits that are consistent with the highest

scoring block are separated by colons. The alignment of the query sequence with the sequence closest to it in the Blocks database is shown. The distance between detected blocks is listed as (min, max): for the database entry followed by the distance in the query. Upper case in the query indicates at least one occurrence of the residue in that column of the block. (b) IMPALA Searcher output. The IMPALA alignment detects the region corresponding to BL00094A in the query sequence as a separate high scoring segment, which lies 143 amino acids upstream of BL00094B. The query sequence is aligned with the COBBLER sequence used to make the PSI-BLAST PSSM. In the two alignments shown, no gaps have been inserted within the block regions.

important, and PSSMs from Blocks using position-based sequence weights [27] and position-based pseudocounts [19] substantially outperformed BLAST and Smith-Waterman searching using single-sequence representatives [8]. Consensus-embedded single sequences using the COBBLER method also outperformed single sequence representatives in comprehensive evaluations. The Blocks Database servers provide direct links for search-

ing the current sequence databanks. In addition, the LAMA program [10] allows the users to search entries from the Blocks+ Database, or blocks derived from their own sequences, against the Blocks+ Database. This method compares multiply aligned sequences in both query and target entries. Therefore, all compared positions are explicitly represented by distributions of amino acids, rather than single residues. This increases the sen-

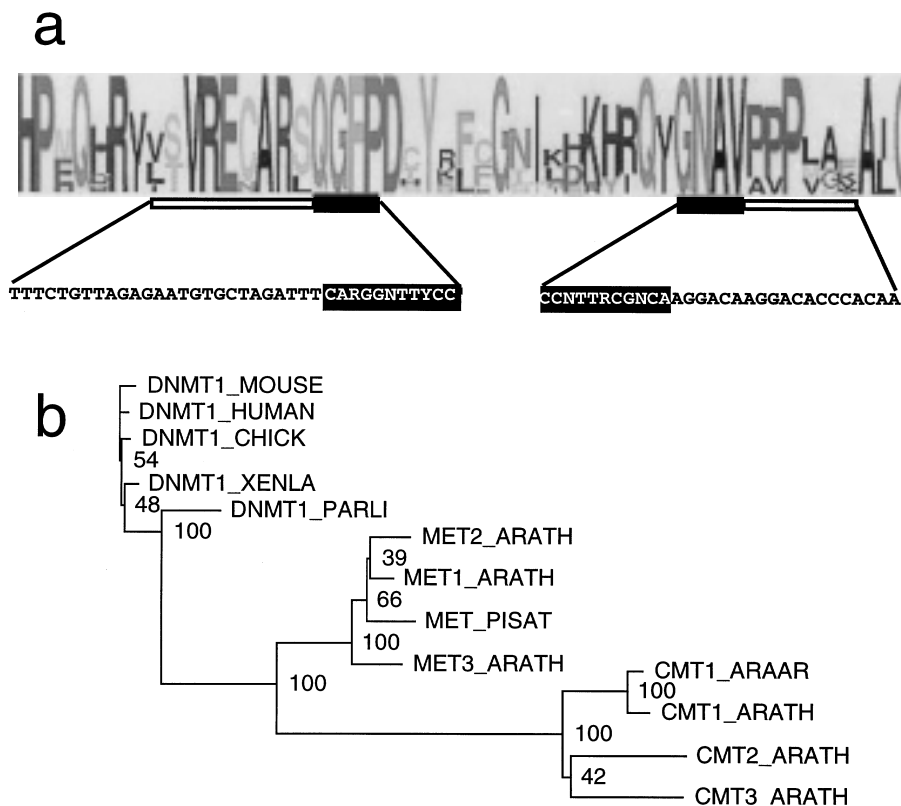


Figure 4. Block-based tools applied to chromomethylases. Eight Dnmt1-like DNA methyltransferases and two CMT1s were submitted to Block Maker. (a) Sequence logo of Motif Block H and location of CODEHOP-designed primers (see Fig. 5). In the logo, the height of each amino acid is scaled in bits of information 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). The *A. thaliana*

codon usage table was chosen. The weights of CMT1s were increased 4-fold to favor the recovery of CMTs. A new CMT, CMT3 [36] was discovered using these primers. (b) Phylogenetic bootstrap tree of Dnmt1s and CMTs, including CMT2 and CMT3.

sitivity of the method, allowing the detection of subtle relations between corresponding positions in long-diverged protein families. Selectivity is achieved by completely avoiding the use of interblock sequence regions. These are not conserved between all family members and can lead to chance similarity. LAMA was shown to detect sequence similarities beyond the range of sequence-to-sequence and sequence-to-alignment methods [10]. Several predictions made by LAMA [28–30] were experimentally confirmed [31–34].

2.5 Consensus-degenerate hybrid oligonucleotide primer (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 nondegenerate 5' "clamp" stabilizes annealing of the core to the starting template (Fig. 5) [35]. 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 nondegenerate 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 method has been validated by the successful amplification of products that have proven challenging using conventional methods [35].

CODEHOPs are designed automatically by a program that predicts optimal primers given a set of blocks. Hyper-text links to the CODEHOP designer are provided from

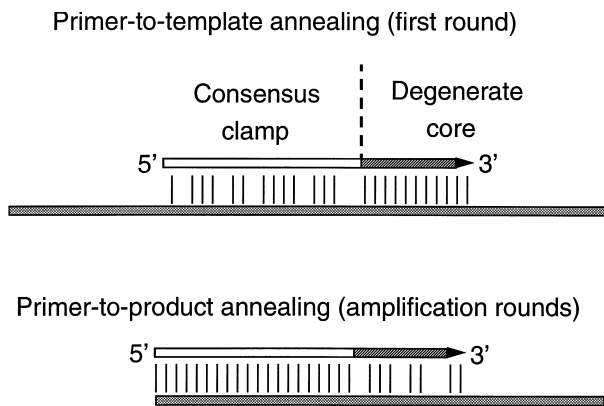


Figure 5. CODEHOP primer design strategy (adapted from [35]). Degeneracy is confined to an 11–12 aa 3' "core", and a single consensus sequence provides a 5' "clamp". The core is too short to anneal stably to the target template sequence, but the consensus clamp stabilizes annealing because it partially matches the template. After a primer is incorporated, the consensus clamp precisely matches the other primers in the pool, leading to stable priming for all subsequent PCR cycles. This differs from degenerate PCR, because degeneracy throughout the molecule leads to mismatching of each incorporated primer with other primers in the pool, requiring lower annealing temperatures. CODEHOP primers use stringent annealing temperatures, and the short length of the degenerate region makes it possible to design primers to shorter regions of conservation. A typical pair of CODEHOP primers is shown in Fig. 4a.

the Blocks+ and Prints Databases, Block Maker 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 in Fig. 4 for CMTs. At the time this work was performed, only CMT1s from closely related *Arabidopsis* species were known. Block Maker was used to make blocks from Dnmt1-like proteins and CMTs, and CODEHOP primers were designed, giving CMTs 4-fold higher weights in order to favor their recovery (Fig. 4a). These primers were used to successfully amplify CMT3, a new CMT that has only 40% amino acid sequence identity to CMT1 [22] and CMT2 (see Fig. 3). Figure 4b shows a phylogenetic tree of currently known Dnmt1s and CMTs obtained from Block Maker Motif blocks. Interestingly, a trichotomy is seen, with plant-specific Dnmt1s lacking chromodomains showing more affinity than their animal-specific counterparts with CMTs.

3 Conclusion

As an ever-increasing fraction of protein-coding sequences are found to fall into families, tools such as those described here become more important for interpreting the large volume of sequence data. Just as the functional inferences from protein sequence comparison have driven genomics during the final years of the previous millennium, we anticipate that blocks-based tools will play a role in the emerging field of proteomics at the beginning of the new millennium.

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

Received December 8, 1999

4 References

- [1] Henikoff, S., Henikoff, J. G., *Nucleic Acids Res.* 1991, *19*, 6565–6572.
- [2] Hofmann, K., Bucher, P., Falquet, L., Bairoch, A., *Nucleic Acids Res.* 1999, *27*, 215–219.
- [3] Henikoff, S., Henikoff, J. G., *Proc. Natl. Acad. Sci. USA* 1992, *89*, 10915–10919.
- [4] Henikoff, S., Henikoff, J. G., Alford, W. J., Pietrovski, S., *Gene* 1995, *163*, GC17–GC26.
- [5] Smith, H. O., Annau, T. M., Chandrasegaran, S., *Proc. Natl. Acad. Sci. USA* 1990, *87*, 826–830.
- [6] Neuwald, A. F., Liu, J. S., Lawrence, C. E., *Prot. Sci.* 1995, *4*, 1618–1632.
- [7] Bailey, T. L., Gribskov, M., *Bioinformatics* 1998, *14*, 48–54.
- [8] Henikoff, S., Henikoff, J. G., *Prot. Sci.* 1997, *6*, 698–705.
- [9] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., *Nucleic Acids Res.* 1997, *25*, 3389–3402.
- [10] Pietrovski, S., *Nucleic Acids Res.* 1996, *24*, 3836–3845.
- [11] Henikoff, S., Henikoff, J. G., Pietrovski, S., *Bioinformatics* 1999, *15*, 471–479.
- [12] Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. N., Wright, W., *Nucleic Acids Res.* 1999, *27*, 220–225.
- [13] Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D., Sonnhammer, E. L. L., *Nucleic Acids Res.* 1999, *27*, 260–262.
- [14] Corpet, F., Gouzy, J., Kahn, D., *Nucleic Acids Res.* 1999, *27*, 263–267.
- [15] Gracy, J., Argos, P., *Bioinformatics* 1998, *14*, 164–173.
- [16] Bairoch, A., Boeckmann, B., *Nucleic Acids Res.* 1992, *20*, 2019–2022.
- [17] Posfai, J., Bhagwat, A. S., Posfai, G., Roberts, R. J., *Nucleic Acids Res.* 1989, *17*, 2421–2435.
- [18] Henikoff, J. G., Henikoff, S., Pietrovski, S., *Nucleic Acids Res.* 1999, *27*, 226–228.
- [19] Henikoff, J. G., Henikoff, S., *Comput. Appl. Biosci.* 1996, *12*, 135–143.
- [20] Bailey, T. L., Gribskov, M., *J. Comput. Biol.* 1997, *4*, 45–59.

- [21] Tatusov, R. L., Altschul, S. F., Koonin, E. V., *Proc. Natl. Acad. Sci. USA* 1994, 91, 12091–12095.
- [22] Henikoff, S., Comai, L., *Genetics* 1998, 149, 307–318.
- [23] Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Loonin, E. V., Aravind, L., Altschul, S. F., *Bioinformatics* 1999, 15, 1000–1011.
- [24] Schneider, T. D., Stephens, R. M., *Nucleic Acids Res.* 1990, 18, 6097–6100.
- [25] Henikoff, J. G., Pietrokovski, S., Henikoff, S., *Nucleic Acids Res.* 1997, 25, 222–225.
- [26] Sayle, R. A., Milner-White, E. F., *Trends Biochem. Sci.* 1995, 20, 374.
- [27] Henikoff, S., Henikoff, J. G., *J. Mol. Biol.* 1994, 243, 574–578.
- [28] Pietrokovski, S., Henikoff, S., *Mol. Gen. Genet.* 1997, 254, 689–695.
- [29] Pietrokovski, S., *Prot. Sci.* 1998, 7, 64–71.
- [30] Herbert, A., Alfken, J., Kim, Y. G., Mian, I. S., Nishikura, K., Rich, A., *Proc. Natl. Acad. Sci. USA* 1997, 94, 8421–8426.
- [31] van Pouderooyen, G., Ketting, R. F., Perrakis, A., Plasterk, R. H., Sixma, T. K., *EMBO J.* 1997, 16, 6044–6054.
- [32] Wang, H., Hartswood, E., Finnegan, D. J., *Nucleic Acids Res.* 1999, 15, 455–461.
- [33] Hall, T. M., Porter, J. A., Young, K. E., Koonin, E. V., Beachy, P. A., Leahy, D. J., *Cell* 1997, 91, 85–97.
- [34] Schwartz, T., Rould, M. A., Lowenhaupt, K., Herbert, A., Rich, A., *Science* 1999, 284, 1841–1845.
- [35] Rose, T. M., Schultz, E. R., Henikoff, J. G., Pietrokovski, S., McCallum, C. M., Henikoff, S., *Nucleic Acids Res.* 1998, 26, 1628–1635.
- [36] McCallum, C. M., Comai, L., Greene, E. A., Henikoff, S., *Nature Biotechnol.* 2000, 18, 455–457.