

Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations

Steven Henikoff^{1,*}, Jorja G. Henikoff¹ and Shmuel Pietrokovski^{1,2}

¹Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, 100 Fairview Avenue North, Seattle, WA 98109-1024, USA

Received on January 18, 1999; revised on April 2, 1999; accepted on April 13, 1999

Abstract

Motivation: As databanks grow, sequence classification and prediction of function by searching protein family databases becomes increasingly valuable. The original Blocks Database, which contains ungapped multiple alignments for families documented in PROSITE, can be searched to classify new sequences. However, PROSITE is incomplete, and families from other databases are now available to expand coverage of the Blocks Database.

Results: To take advantage of protein family information present in several existing compilations, we have used five databases to construct Blocks+, a unified database that is built on the PROTOMAT/BLOSUM scoring model and that can be searched using a single algorithm for consistent sequence classification. The LAMA blocks-versus-blocks searching program identifies overlapping protein families, making possible a non-redundant hierarchical compilation. Blocks+ consists of all blocks derived from PROSITE, blocks from Prints not present in PROSITE, blocks from Pfam-A not present in PROSITE or Prints, and so on for ProDom and Domo, for a total of 1995 protein families represented by 8909 blocks, doubling the coverage of the original Blocks Database. A challenge for any procedure aimed at non-redundancy is to retain related but distinct families while discarding those that are duplicates. We illustrate how using multiple compilations can minimize this potential problem by examining the SNF2 family of ATPases, which is detectably similar to distinct families of helicases and ATPases.

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

Contact: steveh@fhcrc.org

Introduction

Continuing improvements in sequencing technology and the launching of large-scale sequencing projects have 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. Unfortunately, searches of sequence databanks can yield voluminous results, and it becomes increasingly challenging to deduce family membership from hits to many related sequences. Furthermore, a large percentage of proteins consist of multiple modules, and this 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.

The value of protein family databases has motivated the construction of several compilations. Currently, there are a few curated databases, among which PROSITE (Hofmann *et al.*, 1999), Prints (Attwood *et al.*, 1999) and Pfam-A (Bate-man *et al.*, 1999) are well established. In curated compilations, an expert judgment has been made concerning membership in a protein family, and this judgment is often supported by documentation and links to the literature. However, curation is labor-intensive, and none of these curated databases contains all of the currently known protein families.

Proteins are grouped and protein family information is represented and searched in different ways in different curated collections. A PROSITE entry corresponds to a set of protein sequences grouped by an expert using biological information which is provided as documentation. A protein family is represented in PROSITE by one or more simple patterns or weight matrices corresponding to shared modules or domains. These are derived from unpublished multiple

*To whom correspondence should be addressed. ²Present address: Molecular Genetics Department, Weizmann Institute of Science, PO Box 26, Rehovot 76100, Israel.

alignments of the sequences in the family. There are several programs for searching PROSITE patterns. Since early 1997, PROSITE has added only 18 new families with minimal documentation, for a total of 1015.

A Prints entry also corresponds to a set of protein sequences grouped by an expert with documentation. However, a Prints family is represented by a fingerprint consisting of a set of ungapped multiple alignments corresponding to the shared modules or domains, following the blocks model (Posfai *et al.*, 1989). Prints alignments can be used to derive patterns or weight matrices for searching with a variety of algorithms. Prints adds 50 to 100 entries annually and currently contains 1100 families.

A Pfam-A seed entry similarly corresponds to a set of protein sequences grouped by an expert, but no documentation beyond links to sequence databanks and PROSITE is generally included. The entry is initially represented by a gapped multiple alignment constructed semi-manually, and this is searched against sequence databases to add more sequences to the family. Addition of new sequences and adjustment of the seed alignment are performed in an automated, uncurated manner. Unlike the procedures used to construct PROSITE and Prints, which concentrate on the conserved regions of a family's sequences, Pfam's gapped alignments may include long regions of uncertain alignment between conserved regions. A curator must sometimes manually excise the conserved regions from the full-length sequences in order to obtain a reasonable seed alignment. Programs are available to search Pfam's gapped multiple alignments. Pfam has undergone a recent spurt of growth, expanding from 527 to 1408 entries during 1998.

In addition to curated compilations, protein family databases based solely on sequence similarity are available. These have been constructed by automated clustering of protein sequence databanks. There have been many attempts to automatically cluster protein databases based on sequence similarity (Smith and Smith, 1990; Sheridan and Venkataraghavan, 1992; Gonnet *et al.*, 1992; Harris *et al.*, 1992; Sonnhammer and Kahn, 1994; Gracy and Argos, 1998), both to avoid the manual effort of curation and to suggest new family groupings. The most developed and best maintained effort is the current version of ProDom (Corpet *et al.*, 1999), which uses PSI-BLAST (Altschul *et al.*, 1997) to cluster Swiss-Prot. Each ProDom entry corresponds to a single module or domain which is represented by a gapped multiple alignment in just the shared region, together with a consensus sequence derived from it. The consensus sequences or aligned segments can be searched as sequence databases. A few of the larger ProDom groups are now being curated. ProDom is re-constructed with each release of Swiss-Prot and in 1998 had 17 777 entries with at least two sequence segments in the alignment.

Domo is another automatically clustered collection of protein families which tries to group sequences that share multiple modules into a single entry (Gracy and Argos, 1998). Like ProDom, a Domo entry is represented by a gapped multiple alignment. Domo had 8877 entries in 1998. Although ProDom and Domo have many more entries than PROSITE, Prints or Pfam-A, their groupings are made solely from sequence similarity without the benefit of expert opinion, and many entries include only a few sequences. Moreover, family designations change with each version of a clustering database, rather than being fixed as in curated compilations.

There is considerable overlap between all of these collections, and most of the largest protein families are represented in all of them. Because the definition of an entry differs among them, however, it is difficult to directly compare the protein families they represent. Furthermore, using these collections to classify new sequences is problematic because different family representations and searching algorithms are utilized. It can be difficult even to know whether null search results stem from the absence of a family in the database or from a limitation of the searching method.

Primary compilations of protein families are valuable resources. But because they are so different and have been initiated and maintained independently, their integration is difficult (but see Wu *et al.*, 1999). To address this problem, the InterPro consortium plans to integrate PROSITE, Prints and Pfam (Attwood *et al.*, 1999). Here we describe an automated approach to integrating multiple diverse compilations by expansion of the Blocks Database, a collection of ungapped multiple alignments of conserved regions of protein families (Henikoff *et al.*, 1999). Its families have traditionally been drawn from those documented in PROSITE, but its block-making procedure, PROTOMAT (Henikoff and Henikoff, 1991), can be applied to any set of related protein sequences and will detect multiple conserved regions (Henikoff *et al.*, 1995). Previously, the Blocks Database was augmented with families from Prints that are absent from PROSITE (Henikoff *et al.*, 1997). Here we apply PROTOMAT successively to Pfam-A families and to the largest entries in ProDom and Domo. The result is a unified database of blocks representing twice as many protein families as originally while still favoring families from curated collections.

Methods

Overview

To construct Blocks+, we first process PROSITE entries with PROTOMAT in the original way; PROTOMAT makes blocks from all full-length sequences in an entry. Prints blocks are then added for entries without corresponding PROSITE entries. Next, PROTOMAT is used to make blocks from Pfam-A entries; the LAMA blocks-versus-

blocks searching program (Petrokovski, 1996) is used to compare the Pfam blocks with those from PROSITE and Prints already in Blocks+, adding blocks for families with no overlap to Blocks+. Next, ProDom entries are processed with PROTOMAT and LAMA in the same way, comparing ProDom blocks with Blocks, Prints and Pfam blocks already in Blocks+. Finally, Domo is processed with PROTOMAT and LAMA in the same way. This hierarchical procedure favors inclusion of annotated entries (PROSITE and Prints) over verified entries (Pfam-A) over unverified entries (ProDom and Domo) and inclusion of entries with stable designations over those that change with each version of a clustering database.

Blocks/Prints

The PROTOMAT system consists of a sequence extraction utility, a motif finding program and a block assembly program. Application of this automated system successively to entries from PROSITE and Swiss-Prot to create the Blocks Database has been previously described (Henikoff and Henikoff, 1991). Prints fingerprints are similar to blocks in representing the most highly conserved regions of a protein family by a series of ungapped multiple alignments, except that fingerprints are crafted from semi-manual methods (Attwood and Beck, 1994). Fingerprints were added to the Blocks Database simply by reformatting, omitting Prints entries that are annotated as corresponding to PROSITE entries, creating the Blocks/Prints Database (Henikoff *et al.*, 1997). The Prints accession number was retained. Prints-derived blocks were calibrated theoretically (Tatusov *et al.*, 1994) for searching with the BLIMPS searching program (Henikoff *et al.*, 1995).

The simple procedure used for Prints could not be used for the other databases because their gapped alignments do not conform to the Blocks model. Therefore, PROTOMAT was used to make blocks from the Swiss-Prot sequences listed in each Pfam-A, ProDom and Domo entry, and blocks from a subset of entries were added to Blocks/Prints to create Blocks+. Blocks+ was initialized to Blocks/Prints and the following procedure was used to determine which entries to add from the other databases.

Blocks corresponding to Pfam-A entries

PROTOMAT was run on each Pfam-A entry (version 3.4, February 1999), using single occurrences of each full-length Swiss-Prot sequence; an individual sequence may appear more than once in a Pfam-A alignment if it contains repeated modules. Trembl sequences in Pfam-A, preliminary entries that lack Swiss-Prot documentation, were not included. Entries that contained fewer than 3 different Swiss-Prot sequences were not processed. The PROTOMAT repeats parameter was derived from the number of repeats implied by the

Pfam-A alignment. The resulting blocks for the entry were searched against Blocks+ (so far) and added to it if no hit was found. A Pfam-derived set of blocks that did not report hits above the LAMA threshold was accepted and its blocks added to Blocks+. The Pfam-A accession number was adopted with a letter appended for each block (e.g. PF00176A, PF00176B, ...).

Blocks corresponding to ProDom and Domo entries

The same basic procedures described above for Pfam-A were used to add entries from ProDom not already represented in Blocks+ after Pfam-A was processed, and from Domo after ProDom was processed. However, because both ProDom and Domo contain a large number of entries compared with Pfam-A, a two-step procedure was used to reduce the number of entries from ProDom and Domo to be processed by PROTOMAT. Ungapped blocks at least 10 amino acids wide in all of the sequences were first 'carved out' from ProDom's and Domo's multiple alignments (Henikoff *et al.*, 1999), and LAMA was used to search them against Blocks/Prints. Entries that reported LAMA hits above a threshold selected to minimize false positive hits based on empirical tests ($Z = 8.2$, Petrokovski, 1996) were not processed further.

For ProDom (version 36, August 1998), gapped multiple alignments for entries 1–2999 were processed; higher-numbered ProDom entries generally do not include enough diverse sequences to be interesting. All full-length Swiss-Prot sequences were used and accession numbers were constructed from 'PD' plus the ProDom entry number with a letter appended for each block (e.g. PD00492A). For Domo (version 2.0, April 1998), gapped multiple alignments for entries 1–1999 were processed, all full-length Swiss-Prot (not PIR) sequences were used, and the Domo accession numbers were retained with a letter appended for each block (e.g. DM00547A).

To calibrate blocks made from Pfam-A, ProDom and Domo entries, the BLIMPS searching program (Henikoff *et al.*, 1995) was used to search each block against Swiss-Prot as is done for PROSITE-derived blocks (Henikoff and Henikoff, 1991). For WWW use, logos, bootstrap trees and COBBLER sequences were made from them (Henikoff *et al.*, 1997). Because blocks found by PROTOMAT will not necessarily correspond to aligned regions present in Pfam-A, ProDom and Domo entries, LAMA was used to search the blocks in Blocks+ derived from Pfam-A, ProDom and Domo against blocks at least 10 positions wide in all of the sequences carved out from their multiple alignments as described above. For hits above the LAMA threshold, links were then established from the Pfam-derived blocks in Blocks+ to Pfam-A entries, from the ProDom-derived blocks to ProDom entries, and from the Domo-derived

blocks to Domo entries. LAMA was also used to search Blocks and Prints versus the carved-out alignments from Pfam-A, ProDom and Domo, establishing links to entries in those collections considered redundant for Blocks+.

Other alternatives

We considered using the ungapped blocks carved out from the Pfam-A, ProDom and Domo multiple alignments to supplement Blocks/Prints instead of making blocks from the sequences represented in them. However, no ungapped blocks at least 10 amino acids wide in all of the sequences could be found for 30 Pfam-A seed or for 505 Pfam-A full alignments, for 493 of the first 2999 entries from ProDom 36, and for 89 of the first 1999 Domo entries, owing to the extensive use of gaps and to our requirement that all positions be ungapped in all sequences. This finding highlights the differences between the local Blocks/Prints model and the more global models used by the other collections. Blocks+ is designed for detecting conserved regions in query sequences, a task for which the ungapped block model works well, and does not attempt to construct multiple alignments across unconserved regions.

Another possible way to identify redundant groups between the five collections would be to compare the lists of sequences in each entry. However, this approach is difficult because sequences will inevitably be grouped differently by each method, and it is too indirect when the task is to collect conserved regions. In contrast, direct comparison of multiple alignments with LAMA compares the same entities, blocks, that we want to collect into Blocks+.

Availability

Blocks+ is the default Blocks Searcher database, and is available for WWW retrieval, searching and analysis at <http://blocks.fhrc.org>. Programs were implemented in the C programming language, and source code and SUN Solaris executable versions are available by anonymous ftp from the NCBI repository (<ftp.ncbi.nlm.nih.gov>, `cd repository/blocks/unix`).

Results

Blocks+ construction and testing

Construction of Blocks+ was performed by successive addition of blocks from Prints 22, Pfam 3.4, ProDom 36 and Domo 2.0 to Blocks 11.0, for a total of 8909 blocks representing 1995 protein families (Figure 1). This procedure approximately doubles the size of the Blocks 11.0 Database based on PROSITE alone.

To check whether the addition of protein families from diverse sources altered the consistency of block alignments in the Blocks Database, we computed a series of amino acid

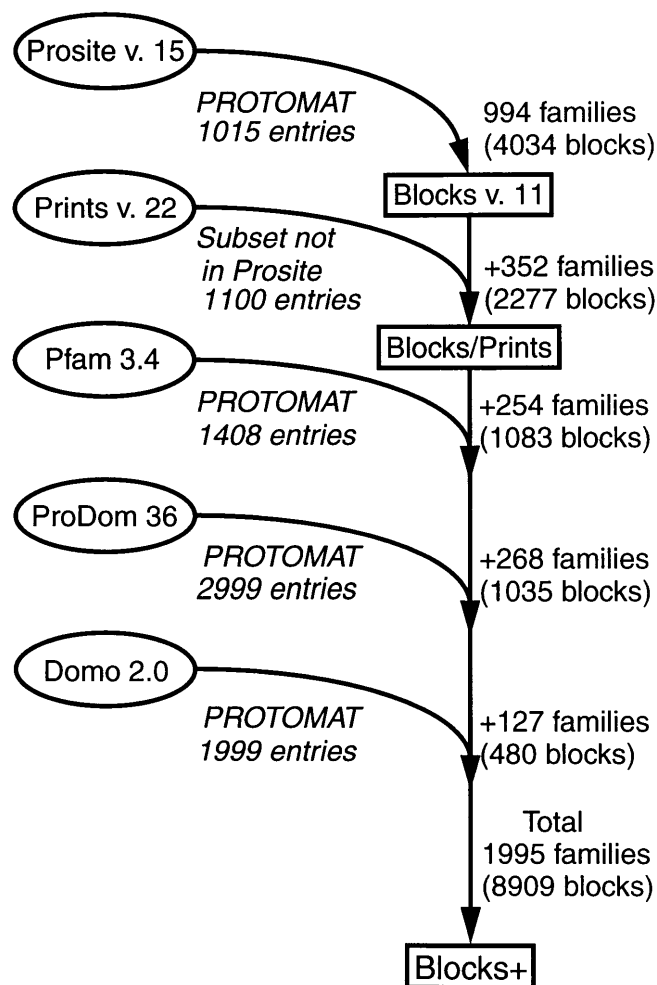


Fig. 1. Flow chart diagram of the hierarchical procedure used to build Blocks+ from constituent databases.

substitution matrices from Blocks+. The BLOSUM series was originally produced from the PROSITE-based Blocks 5.0 by iteratively running the PROTOMAT block maker, which requires a substitution matrix to refine conserved regions, and the BLOSUM matrix construction programs, starting with a unitary matrix consisting of a single positive score for all matches and a single negative score for all mismatches, and ending with convergence (Henikoff and Henikoff, 1992). Blocks+ contains five times as many amino acid residues as Blocks 5.0. Nevertheless, the Blocks+ equivalent of the popular BLOSUM 62 substitution matrix, with relative entropy 0.70, has identical half-bit values for 174/210 scores, and the other 36 scores differ by only ± 1 (data not shown). This finding suggests that the block alignments in Blocks+ are consistent with those found in a purely PROSITE-based version of the database and addresses the concern that the BLOSUM series is biased towards families fa-

vored by the curators of PROSITE. If any such bias exists, it is shared by curators of Prints and Pfam-A and by developers of ProDom and Domo. This result also confirms that there were sufficient alignment data in Blocks 5.0 to compute BLOSUM 62 with adequate accuracy. In comprehensive empirical tests of the BLOSUM series based on Blocks+, we have not detected performance improvements over the 1992 series (unpublished results).

PROTOMAT was not applied to Prints, because its manually curated fingerprints and individually documented block alignments conform to the model that underlies the Blocks Database. Rather, Prints documentation of overlaps between Prints and PROSITE entries was used to exclude redundant Prints entries. A total of 2277 blocks for 352 Prints entries were added to Blocks+. This provided a standard of comparison for our automated LAMA-based procedure for detecting family overlap. We asked how many LAMA hits were found by searching the Prints-derived Blocks+ blocks against Blocks 11. About one sixth (58/352) of these Prints families were classified as significantly similar to those already present in Blocks from PROSITE in that they shared at least one block. That is, compared to the standard, our automated procedure reports relationships that are not sufficiently compelling to be merged in Prints. This is not unexpected, because LAMA is capable of detecting structural motifs, such as families of helix-turn-helix DNA-binding proteins (Petrokovski and Henikoff, 1997), which are shared between families that are thought not to be ancestrally related. LAMA is also capable of detecting very distant ancestral relationships between distinct protein families. Other hits may represent conserved regions from unrelated protein families that share similar domains or compositional biases, for example, different families of proline-rich proteins. Based on this result, we expect that our LAMA-based procedure excludes distinct new protein families from Blocks+ ~1/6 of the time. However, by applying this procedure successively to different family databases, a family from one database that was excluded because of a marginal similarity to one already present might be added as an entry from another database if its entry contains a different subset of sequences.

Capturing the SNF2 family in Blocks+

To illustrate how our procedure performs on a challenging example, we choose the SNF2 family of proteins. This family of mostly DNA-stimulated ATPases is the subject of intense interest: various SNF2 family members perform the central catalytic role in eukaryotic chromatin remodeling complexes (Cote *et al.*, 1994; Tsukiyama *et al.*, 1995; Varga-Weisz *et al.*, 1997) and in regulation of the basal transcription machinery (Auble *et al.*, 1997). Numerous reviews have been written about members of this family, and an excellent

WWW site is maintained by Jonathan Eisen (<http://www.stanford.edu/~jeisen/SNF2/snf2.html>). It might seem surprising that the SNF2 family is absent from both PROSITE and Prints. However, shared similarity with several other families, suggesting common ancestry, may have led to confusion and is indicative of the difficulty of precisely delineating a protein family. Because these distantly related families of proteins are thought to be distinct (Tatusov *et al.*, 1994) yet share a series of 6–7 conserved ‘helicase motifs’ (so called because members of some families unwind nucleic acid duplexes), it has been difficult to unequivocally distinguish families based on sequence similarity alone. Should SNF2 sequences be classified separately or together with other proteins that share helicase motifs? The opinion of researchers in this field is that they form a separate family, as reflected in Eisen’s WWW site, which excludes other families that share helicase motifs, noting that helicase activity has never been detected for any SNF2 family member (e.g. Cote *et al.*, 1994). Nevertheless, there is little doubt that different families that share helicase motifs share a similar fold (Bork and Koonin, 1993).

Examination of Blocks+ reveals that it includes the SNF2 family as an entry derived from Domo (DM00547). PROTOMAT reported 6 blocks shared by 21 sequences in the Domo entry, discarding one sequence (VEF_GVHA) because it was too dissimilar from the other 21 sequences. VEF_GVHA is not a SNF2 family member, but rather is a viral glycoprotein that appears to have been included in the Domo cluster by chance. The SNF2 WWW site lists the 21 sequences (not the VEF_GVHA glycoprotein), which belong to 9 of 12 identified eukaryotic sub-families. Examination of Eisen’s manually adjusted multiple sequence alignment reveals that all 6 PROTOMAT-generated block alignments are present. Moreover, the bootstrap neighbor-joining tree provided for this set of blocks in Blocks+ reveals that four of Eisen’s five sub-families with multiple members are delineated from all other sub-families with at least 99% bootstrap support (the CHD1 sub-family has 73% support) (Figure 2). The remaining four sub-families identified by Eisen are single sequences with deep branches in the DM00547 tree, which is consistent with their belonging to separate sub-families. Therefore, the blocks representing the SNF2 family in Blocks+ provide detailed accurate alignment information for SNF2 conserved regions.

Because Domo is the last of the five family databases in the Blocks+ hierarchy, one wonders why no SNF2 family entries in Pfam and ProDom were found or retained. Indeed, all three databases contained multiple SNF2 family entries (Table 1). Pfam entry PF00176 (SNF2_N) consists exclusively of SNF2 members. In the course of constructing Blocks+, PROTOMAT was used to make blocks from PF00176, and these were used to search the developing Blocks+ database with LAMA. A hit above the LAMA

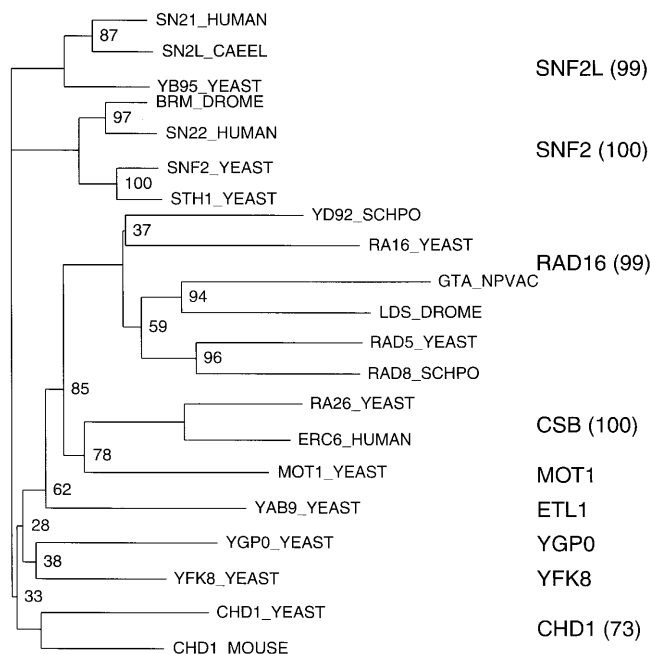


Fig. 2. Tree of SNF2 family members found in Blocks+ entry DM00547. The neighbor-joining tree with 100 bootstrap resamplings was taken from the Blocks+ database. Sub-families documented at the SNF2 web site are indicated on the right. Those with multiple members are shown with bootstrap percentages, and the others are singletons.

threshold was found for one of these blocks to the PROSITE-derived block, BL00039D, which is the most C-terminal block in the DEAD_ATP Helicase family. Therefore, PF00176 was discarded. A second Pfam entry (PF00271, Helicase_C), which includes 236 Swiss-Prot sequences, combines sequences from multiple distinct families containing helicase motifs. PROTOMAT detected a single 8 aa wide block. Using this narrow PF00271 block as LAMA query, a hit to BL00039D was detected, but below the threshold. Therefore, PF00271 was added to Blocks+.

ProDom version 36 contained two entries composed entirely of SNF2 family members, 492 and 593, which represent upstream and downstream conserved regions respectively (Figure 3). PROTOMAT was applied to entry 492 and block PD00492A detected Prints block PR00851D, a family of DNA repair helicases, above the LAMA threshold, and so PD00492 was excluded. PROTOMAT was also applied to entry 593 and one of the resulting blocks hit BL00039D so it was also excluded from Blocks+.

Domo contains three entries composed of SNF2 family members, DM00266, DM00465 and DM00547, representing different conserved regions (Figure 3). PROTOMAT-

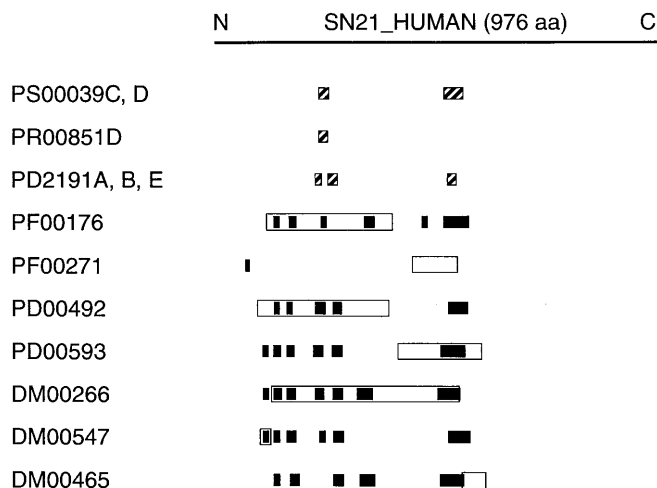


Fig. 3. PROTOMAT-generated blocks for members of the SNF2 family (filled boxes) and other blocks detected by LAMA as significantly similar to blocks in DM00547 (hatched boxes). Blocks are aligned with SN21_HUMAN, the Cobble sequence representative of DM00547. Open boxes depict the extent of alignment for each of the entries in the original databases from which Blocks+ is derived. Note that except for PF00271, which merges multiple families, mostly similar blocks were found by PROTOMAT for the different entries containing SNF2 family members.

generated blocks DM00266G and DM00465F each detected BL00039D above the LAMA threshold and so both DM00266 and DM00465 were excluded. DM00547D detected BL00039C and PR00851D (the 'DExx' region of the ATP binding domain, Figure 4) but below the LAMA threshold, and DM00547F detected BL00039D also below the LAMA threshold. Therefore, DM00547A–F were added to Blocks+. When LAMA was used to search DM00547A–F against carved-out Domo multiple alignments, hits were found to entry DM00266 with blocks B–E, to DM00547 with block A and to DM00465 with block F, and WWW links were automatically made to these entries. A WWW link was also made manually to the Eisen web site.

To ascertain whether the SNF2 blocks are adequate for classifying homologous sequences searched against Blocks+, each member of Eisen's SNF2 sub-families that are not represented in the DM00547-derived blocks was used as query. In every case, classification was successful, with identification of multiple blocks at levels of statistical significance that have never been seen for chance hits (Figure 5).

In summary, our procedure applied successively to multiple family databases eventually resulted in a satisfactory set of blocks representing the SNF2 family.

Blocks and the subset of Prints not in Blocks routinely using existing annotations. Because Pfam-A maintains family IDs, updating requires only adding new entries that survive LAMA searching of entries above it in the hierarchy and deleting redundant entries below. Full updates to ProDom and Domo would be onerous, because almost all of the family IDs change. A simple practical alternative is to use LAMA to establish links from existing Blocks+ ProDom or Domo blocks to the new constituent database, allowing users access to up-to-date information about a family without requiring a complete turnover of Blocks+ entries. This procedure was used to establish links to the most recent ProDom database (version 99.1).

Whereas ProDom and Domo were successful in providing high quality alignment information for the SNF2 family, automated documentation proved to be difficult. Domo named this family ‘CHROMO BROMODOMAIN SHADOW GLOBAL’ based on words and names found in Swiss-Prot documentation (Table 1). CHD1 proteins have two chromodomains (chromo), SNF2 and brahma proteins have bromodomains, one of the chromodomains in CHD1 is referred to as the ‘shadow’ chromodomain because it is difficult to detect, and MOT1, uniquely among SNF2 family members, is a ‘global’ repressor of transcription. In other words, 3 of the 4 words automatically extracted to describe this family refer to entirely unrelated modules, and the other term is inappropriate. ProDom uses different descriptors that are verbose and cryptic; for example, PD00492 is described as ‘PROTEIN HELICASE ATP-BINDING NUCLEAR DNA REPAIR DNA-BINDING TRANSCRIPTION ACTIVATOR REGULATION’ (Table 1). Such problems in obtaining useful automated annotations illustrates the desirability of manual documentation. In the case of the SNF2 family, manual documentation (for Pfam and Blocks+) consists of placing a

WWW link to a site dedicated to the family, which is maintained by an authority in the biological research community. There are dozens of protein families that currently have dedicated WWW sites (<http://www.proweb.org>), and these provide rich sources of information that can be exploited by family databases in lieu of detailed expert curation.

A unified database of consistent protein family representations

The various curated protein family compilations are assembled using very different methods for family definition and representation. In each case, manual methods are applied by experts who try to assure consistency. However, protein families have evolved over billions of years, constrained only by natural selection on protein structure and function, and the resulting diversity is challenging for database curators. For instance, Pfam splits the SNF2 family into two distinct parts: the single C-terminal conserved region merged with C-terminal regions of helicases, and the rest of the conserved regions identified as a SNF2-specific ‘domain’. However, the other conserved regions are also shared with families of helicases, for example, the DExx motif (Figure 4), and it is widely accepted that these families have diverged from one another as single enzymatic units, not as two separate modules (Bork and Koonin, 1993). Automated compilations are likewise confounded, as both ProDom and Domo split the SNF2 family into two or three separate domains. In contrast, the Blocks+ SNF2 entry includes the conserved regions as a single entry. This is a consequence of our procedure, which analyzes entire sequences in a protein family entry, finding the conserved regions shared by most or all sequences without reference to the alignment or pattern provided in the original entry.

Table 1. SNF2 family and distant relatives detected in LAMA searches of developing Blocks+

	Family	Description	#Sequences	In Blocks+?	LAMA detected
PROSITE	PS00039	DEAD-box subfamily ATP-dependent helicases	89	BL00039A-D	
Prints	PR00851	XERODERMA PIGMENTOSUM GROUP B	10	PR00851A-L	
Pfam	PF00176	SNF2 and others N-terminal domain	138	no	BL00039D
	PF00271	Helicases conserved C-terminal domain	236	PF00271	
ProDom	PD00492	PROTEIN HELICASE ATP-BINDING NUCLEAR DNA REPAIR DNA-BINDING TRANSCRIPTION ACTIVATOR REGULATION	42	no	PR00851D
	PD00593	PROTEIN HELICASE ATP-BINDING NUCLEAR DNA-BINDING DNA REPAIR TRANSCRIPTION ACTIVATOR REGULATION	35	no	BL00039D
Domo	DM00266	ATP NP-bind	29	no	BL00039D
	DM00547	CHROMO BROMODOMAIN SHADOW GLOBAL	22	DM00547A-F	
	DM00465	BROMODOMAIN	20	no	BL00039D

Although our procedure extracts families from diverse sources, consistency is imposed by application of the blocks data model which allows a single scoring method to be used for sequence classification. Other systems lack this degree of consistency: manual methods used in all of the curated compilations are inherently subjective, and gap penalties used to produce multiple alignments have no theoretical foundation (Altschul, 1991). Even probabilistic hidden Markov model methods used to produce full Pfam-A alignments are founded on seed alignments which are constructed with deterministic gap penalties and manual adjustments, including those needed to excise conserved regions from multi-domain proteins. In contrast, consistency of PROTOMAT-generated alignments is insured by the use of a single source of alignment scores, the BLOSUM 62 substitution matrix, which was expressly designed to work with PROTOMAT.

Acknowledgements

This work was supported by a grant from NIH. We thank Liz Greene for helpful discussions.

References

- Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.
- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Attwood,T.K. and Beck,M.E. (1994) PRINTS—a protein motif fingerprint database. *Protein Eng.*, **7**, 841–848.
- Attwood,T.K., Flower,D.R., Lewis,A.P., Mabey,J.E., Morgan,S.R., Scordis,P., Selley,J.N. and Wright,W. (1999) PRINTS prepares for the new millenium. *Nucleic Acids Res.*, **27**, 220–225.
- Auble,D.T., Wang,D., Post,K.W. and Hahn,S. (1997) Molecular analysis of the SNF2/SWI2 protein family member MOT1, an ATP-driven enzyme that dissociates TATA-binding protein from DNA. *Mol. Cell. Biol.*, **17**, 4842–4851.
- Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Finn,R.D. and Sonnhammer,E.L.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, **27**, 260–262.
- Bork,P. and Koonin,E.V. (1993) An expanding family of helicases within the ‘DEAD/H’ superfamily. *Nucleic Acids Res.*, **21**, 751–752.
- Corpet,F., Gouzy,J. and Kahn,D. (1999) Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.*, **27**, 263–267.
- Cote,J., Quinn,J., Workman,J.L. and Peterson,C.L. (1994) Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science*, **265**, 53–60.
- Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.
- Gracy,J. and Argos,P. (1998) Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics*, **14**, 164–173.
- Harris,N., Hunter,L. and States,D. (1992) Megaclassification: discovering motifs in massive datastreams. In *Tenth National Conference on Artificial Intelligence*, pp. 837–842. AAAI Press, San Jose.
- Henikoff,J.G., Pietrokovski,S. and Henikoff,S. (1997) Recent enhancements to the blocks database servers. *Nucleic Acids Res.*, **25**, 222–225.
- Henikoff,J.G., Henikoff,S. and Pietrokovski,S. (1999) New features of the Blocks Database servers. *Nucleic Acids Res.*, **27**, 226–228.
- Henikoff,S. and Henikoff,J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, **19**, 6565–6572.
- Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
- Henikoff,S., Henikoff,J.G., Alford,W.J. and Pietrokovski,S. (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, **163**, GC17–GC26.
- Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
- Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Res.*, **25**, 1665–1677.
- Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
- Pietrokovski,S. and Henikoff,S. (1997) A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. *Mol. Gen. Genet.*, **254**, 689–695.
- Posfai,J., Bhagwat,A.S., Posfai,G. and Roberts,R.J. (1989) Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Res.*, **17**, 2421–2435.
- Sheridan,R.P. and Venkataraghavan,R. (1992) A systematic search for protein signature sequences. *Proteins: Struct. Funct. Genet.*, **14**, 16–28.
- Smith,R.F. and Smith,T.F. (1990) Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl Acad. Sci. USA*, **87**, 118–122.
- Sonnhammer,E.L.L. and Kahn,D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*, **3**, 482–492.
- Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
- Tsukiyama,T., Daniel,C., Tamkun,J. and Wu,C. (1995) ISWI, a member of the SWI2/SNF2 ATPase family, encodes the 140 kDa subunit of the nucleosome remodeling factor. *Cell*, **83**, 1021–1026.
- Varga-Weisz,P.D., Wilm,M., Bonte,E., Dumas,K., Mann,M. and Becker,P.B. (1997) Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature*, **388**, 598–602.
- Wu,C.H., Shivakumar,S. and Huang,H. (1999) ProClass protein family database. *Nucleic Acids Res.*, **27**, 272–274.