

## Review

# The apoptosis database

KS Doctor<sup>1</sup>, JC Reed<sup>1</sup>, A Godzik<sup>1</sup> and PE Bourne<sup>\*,1,2</sup>

<sup>1</sup> The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

<sup>2</sup> San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA

\* Corresponding author: PE Bourne, Tel: 858-534-8301; Fax: 858-822-0873, E-mail: bourne@spsc.edu

Received 10.9.02; revised 3.12.02; accepted 10.12.02

Edited by Dr Green

## Abstract

The apoptosis database is a public resource for researchers and students interested in the molecular biology of apoptosis. The resource provides functional annotation, literature references, diagrams/images, and alternative nomenclatures on a set of proteins having 'apoptotic domains'. These are the distinctive domains that are often, if not exclusively, found in proteins involved in apoptosis. The initial choice of proteins to be included is defined by apoptosis experts and bioinformatics tools. Users can browse through the web accessible lists of domains, proteins containing these domains and their associated homologs. The database can also be searched by sequence homology using basic local alignment search tool, text word matches of the annotation, and identifiers for specific records. The resource is available at <http://www.apoptosis-db.org> and is updated on a regular basis.

*Cell Death and Differentiation* (2003) 10, 621–633. doi:10.1038/sj.cdd.4401230

**Keywords:** apoptosis; programmed cell death; relational database; bioinformatics; homology; protein families; protein domains; domain classification

**Abbreviations:** BLAST, basic local alignment search tool; PSI-BLAST, position, specific iterative BLAST; FPS, family pairwise search; HMM, hidden Markov model; SMART, simple modular architecture research tool; DART, domain architecture retrieval tool; SCOP, structural classification of proteins; AIF, apoptosis-inducing factor; BAG, Bcl-2-binding athanogene; BIR, bacterial IAP repeats; IAP, inhibitor of apoptosis protein; BOK, Bcl-2-related ovarian killer; CARD, caspase recruitment domain; CIDE, cell death-inducing DNA fragmentation factor, alpha subunit, like effector; DED, death effector domain; NB-ARC, nucleotide-binding-Apaf-1, R-gene, CED-4; PCD, programmed cell death; Smac, second mitochondria-derived activator of caspase; TRAF, TNF receptor-associated factor; MATH, meprip and TFAF homology; XIAP, X-chromosome-linked inhibitor of apoptosis; TMHMM, transmembrane hidden Markov model.

## Introduction

The set of known proteins that directly regulate apoptosis has grown rapidly over the last 15 years. This growth will continue until all the proteins directly involved in the cell death signaling process are known. This assortment of proteins with wide ranging biochemical functions is linked together conceptually in the minds of apoptosis researchers. Assembling an up-to-date view of this conceptual collection of proteins within the context of apoptosis requires a considerable effort, or more specifically, complete immersion into the field. One principal goal of an apoptosis review article is to assemble such a collection of protein annotations as an educational and research resource. The apoptosis database described here is designed to fulfil the same goal, but to immediately allow the user to dig deeper using local and remote information and to always remain current with respect to the proteins known to be involved in apoptosis.

The foundation of the database is a set of proteins and their distinctive structural domains that are often, if not exclusively, involved in the apoptosis signaling pathway. We refer to these domains as apoptotic domains. The goal is to use these proteins and their associated domains as a framework for functional annotation that is generated automatically or added by apoptosis researchers through the database curator. Examples of annotations present at the various levels of organization are: (a), broad functional information on structural domains conserved across protein families; (b), groups of homologous proteins that contain a recognized domain yet may or may not have an apoptotic function; (c), proteins that have the same set of domains but which differ markedly in their roles in apoptosis.

Additional information such as protein–protein interactions, domain folds from structural classification of proteins (SCOP), protein modifications, genomic information, and literature references are not presently within the database, but are linked via external resources. Some of these features will be incorporated into the database at later stages of development, so they can be queried directly.

The apoptosis database provides a depth of information and a perspective that is not presently available in more general molecular biology resources that are broad but shallow containing limited information on a large number of proteins. Examples of general purpose curated databases are the National Center for Biological Information's RefSeq,<sup>1</sup> the Weizmann Institute's GeneCards,<sup>2</sup> the Swiss Institute of Bioinformatics's SwissPro,<sup>3</sup> and the University of Washington's Pfam.<sup>4</sup> Consider a specific example of the limitation to be found in a general purpose database. NF- $\kappa$ B is a protein with a well-established role before the discovery of its role in apoptosis. The GeneCards and SwissProt databases indicate the initial role of NF- $\kappa$ B in inflammation response, but neglect to mention its role in regulating expression of apoptosis

genes. The apoptosis database highlights the apoptotic role, which would otherwise go unobserved.

The same broad but shallow argument can be made at the level of protein domains. The Pfam database is a powerful and popular tool for finding domains in proteins based on HMM profiles. However, useful functional categorization of proteins requires that the domains be assembled. Other online resources like simple modular architecture research tool (SMART)<sup>5</sup> and domain architecture retrieval tool (DART)<sup>6</sup> rely on domain architecture (the presence and order of domains) to differentiate between homologs containing a domain. We have found it necessary to go beyond domain architecture to provide the most useful annotation. The annotation is based upon (see Materials and Methods and Table 1):(a), the structure of individual domains as found in SCOP;(b), sequence profile-based homology based on family pair-wise search; (c), homology to identified orthologs; (d), specific proteins.

Ideally, we would manually annotate each protein. However, considering the number of proteins involved in apoptosis and the constantly expanding and changing functional annotation, this approach, by itself, is impractical. At the other end of the spectrum are fully automated resources that rely on the detection of remote homology, but which suffer from high error rates and propagation errors and fail to have a good measure of selectivity *versus* sensitivity.

Here, we use an intermediate approach for the apoptotic pathway – utilizing expert knowledge to define model proteins within the pathway and extending by inference using bioinformatics approaches, the functional annotation to other proteins. This approach works as a result of evolutionary constraints applied to apoptotic domains. In practice, often the human and mouse orthologs with published functional annotation will be manually selected to represent one group via their apoptotic domains. If any sequence (from a vertebrate) has a domain most similar to that in this group, it will be associated with the family's functional annotation. Our analysis thus creates families consisting of orthologs, and possibly paralogs, which have differentiated since approximately the time of vertebrate radiation (~500 million years ago) based on their apoptotic domains. We have observed that paralogs having differentiated after vertebrate radiation still have a related function in the same branch of the pathway.

## Results

The apoptotic domains maintained in the current version of the database are listed in Table 2. Of the 13 domains, 12 have representative structures in the PDB.<sup>6</sup> The first three – death effector domain (DED), Caspase recruitment domain (CARD), and death domains (DDs) – share the same SCOP superfamily and fold.<sup>7</sup> Although the domains may be in several proteins involved in apoptosis, the mere presence of the domain does not guarantee an apoptotic function. Nine of the 13 domains are also found in proteins that do not have direct involvement in apoptosis, as indicated by their Swiss-Prot annotation. Thus, the presence of a domain cannot be used to automatically define a protein's involvement in apoptosis. We have compared several alternative means of categorizing proteins with respect to their function in apoptosis. These have been combined as described in the Materials and Methods section.

One potential means to differentiate functionally between homologs that share at least one putative apoptotic domain is to use domain architecture – the set and order of domains. Examples of caspase domain architectures are shown in Figure 1. As illustrated, the function of a protein is not distinct between domain architectures. Although caspases are essential to apoptosis, some caspases are instead involved in cytokine activation rather than apoptosis. For example, human caspase-1 is involved in the inflammation response pathway by virtue of its ability to process proinflammatory (pro-IL) cytokines, pro-IL-1beta and pro-IL-18.<sup>8</sup> Mammalian caspase-1 proteins contain three domains: the CARD domain involved in protein–protein binding, and thus, regulation; caspase large subunit; and caspase small subunits (both subunits, together, form the catalytic site). Mammalian caspase-2 and caspase-9 proteins also share this combination of domains, yet they are both involved in different parts of the apoptosis signaling pathway. Thus, in this instance the domain architecture does not help in differentiating biological function. The details of the specific domains in each protein must be explored in detail.

Consider further our example of the caspases. Caspases have been categorized and renamed by a committee.<sup>9</sup> This standard nomenclature based on individual genes is very useful in both conceptualizing and discussing these proteins and is captured in the database. Phylogenetic analysis of the individual domains shared among caspases leads to cate-



**Table 1** Example of the levels of functional annotation within the apoptosis database

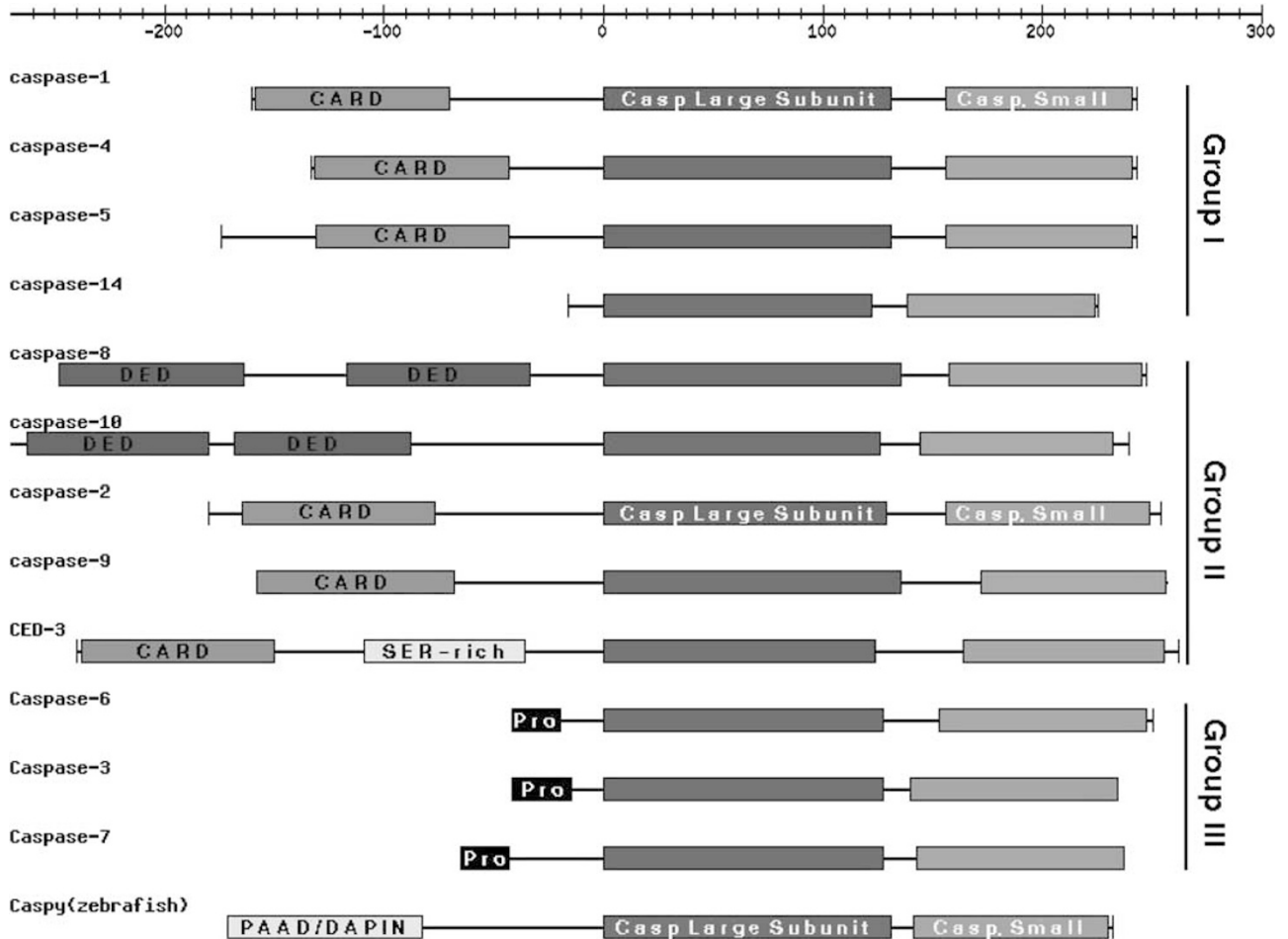
Grouping level	Example grouping	Associated annotation
Structure	Six helix bundle, Greek key CARD/DED/DD	Protein–protein binding domains
Profile	CARD domain	Adapter domain present in a variety of protein families and involved in apoptosis as well as procytokine processing
Family	CARD of RAIDD	RAIDD is an adapter protein that reportedly recruits caspase-2 to the activated TNFR-1 complex via its CARD domain (binding to CARD of caspase-2)
Protein	Human RAIDD	Human RAIDD is an adapter molecule reportedly transferring the proapoptotic signal from TNFR-1 to the activation of caspase-2

**Table 2** Protein domains described in the apoptosis database.

Name and SCOP specification	Exclusively apoptotic	Structure	Generalized function	Assembly
Death effector domain (DED) SCOP: death domain	Yes (possible exception of PEA-15)		Protein–protein binding via large surface area	Homodimers with self and other DD-containing proteins
Caspase recruitment domain (CARD) SCOP: death domain	No (Caspase and NFκB induction)		Protein–protein binding via large surface area	Homodimers with self and other CARD-containing proteins
Death domain (DD) SCOP: death domain	No (Broad range of proteins have death domains)		Protein–protein binding via large surface area	Homodimers with self and other DD-containing proteins
Caspase large and small subunits SCOP: caspase-like (two structural domains)	No (Either apoptosis or pro-cytokine processing)		Protease zymogens with high specificity. Cleaving caspases and specific apoptosis inducing proteins	Combined large and small subunits form the active caspase (after zymogen is activated by cleavage)
TRAF domain (MATH domain) SCOP: TRAF	No (Meprin proteins and several proteins with TRAF-like domains have no role in apoptosis)		Adapter proteins. Peptide and protein binding	Homotrimers for TRAF-2 proteins and likely homo-trimers and possible heterotrimers for other
NB-ARC domain	No (Only in combination: NB-Arc+ CARD)	Not solved	Nucleotide-binding protein oligomerization domain	Oligomers
BIR SCOP: IAP repeats	No (Exceptions like Ac-IAP)		Zn binding interaction via binding other proteins in the apoptosis pathway	Usually 2–3 BIR domains per protein
BAG domain SCOP: BAG domain	No		Hsp70/Hsc 70 binding	Monomers
Bcl-2 domain SCOP: Bcl-2	Yes (Except special case of BH4 only proteins)		Binds other Bcl-2 family proteins creating or indirectly opening pores in mitochondria	Homodimerization with self and other Bcl-2 family proteins
AIF C-terminal domain	Unknown		Unknown	Unknown

Table 2 (continued)

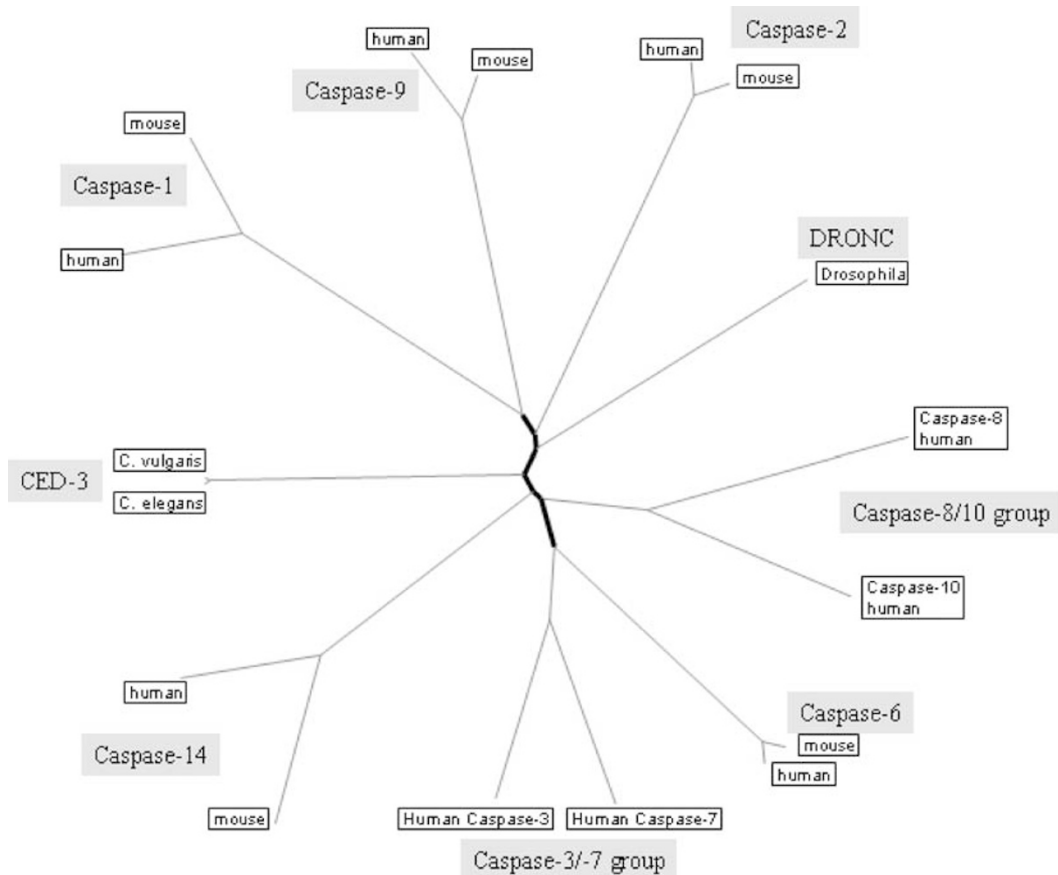
Name and SCOP specification	Exclusively apoptotic	Structure	Generalized function	Assembly
TNF receptor cysteine-rich repeat SCOP: TNF receptor-like	No (Other receptors involved in immunity)		Binds TNF family ligands via this extracellular domain	Trimerization induced by binding TNF-family ligands
CIDE N-terminal domain	Yes		Protein-protein binding	Homodimerization with self and other CIDE-N domain-containing proteins



**Figure 1** Human caspase proteins in the apoptosis database. Group I caspases are cytokine activators, group II proteins are initiators in the apoptotic pathway and group III proteins are apoptosis executioners. This is an example of a diagram maintained in the database with its text and caption subject to text search. Other bioinformatics tools that display the same or similar domain architectures (SMART,<sup>5</sup> DART) are linked to the resource

gories that closely follow the nomenclature. Figures 2 and 3 show dendrograms based on a ClustalW multiple sequence analysis of the caspase small subunit and the CARD domain,

respectively. In both cases, the caspase-1 proteins are distinct from caspase-2 and caspase-9. This evolutionary separation of domains is the principal criteria, which we use to



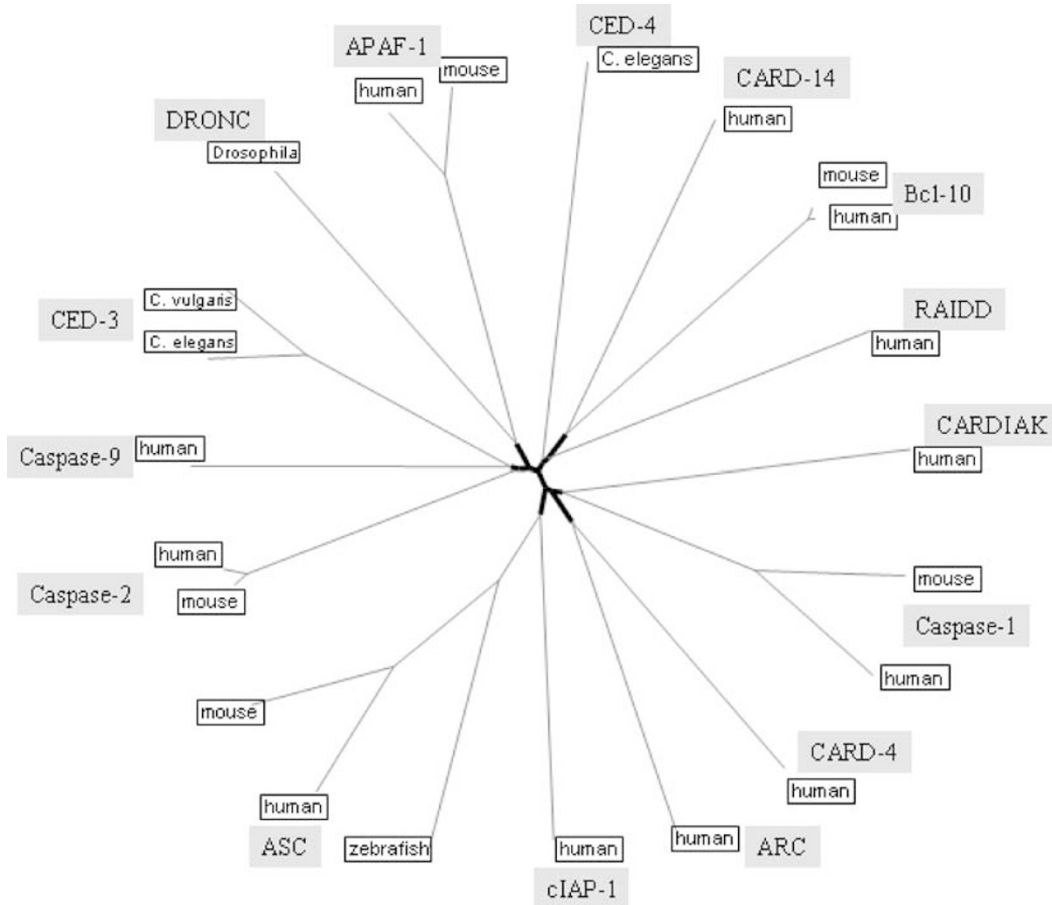
**Figure 2** Phylogeny dendrogram for the caspase small subunit of several caspases based on a ClustalW alignment. Only the proteins representing each family are shown. Branch lines of this unrooted tree that are not exclusive to a family are shown in bold. A simplified version of this diagram is maintained in the database

automate the functional categorization of these proteins. The phylogeny diagram created from the ClustalW alignment is only used as a visual tool to illustrate the separation. In characterizing proteins within the database we use the family pairwise search (FPS) algorithm<sup>10</sup> (see Materials and Methods).

Caspase-1, caspase-2, and caspase-9 are designated as separate families. The apoptotic domains (CARD, caspase small subunit and caspase large subunit) of the representative orthologs define the families for use in the FPS algorithm. Table 3 details the FPS-based classification for a few caspases. Human and mouse caspase-1 were used to establish the caspase-1 family. The FPS-based categorization of each of their domains unambiguously places them into the caspase-1 family. The caspase-9 family was founded by the human and mouse caspase-9 orthologs. The *Xenopus* protein labeled as caspase-9 is clearly categorized by FPS as a caspase-9 based on each of the three domains. Human caspase-2 does not fall into either caspase-1 or caspase-9 families based on rank and relative expectation value over each of its domains. Two additional *Xenopus* homologs of caspase also have the same combination of domains (CARD, caspase small subunit and caspase large subunit), but their role in apoptosis is not yet known. *Xenopus* ICE-A/B proteins have closer homology to

caspase-1, but the assignment is problematic since, although the rank is consistent across domains, some expectation values ( $E > 10^{-8}$ ) are inconclusive. This methodology has, by default, placed the *Xenopus* paralogs of mammalian caspase-1 proteins into the caspase-1 category (*Xenopus*/mammalian divergence occurred approximately 365 million years ago).<sup>11</sup> This includes several other mammalian caspases which seem to have emerged from a common caspase-1 ancestral gene during vertebrate radiation approximately 500 million years ago. However, as more functional information is published about these proteins, a choice can be made to leave them in the caspase-1 family or to create a new family.

The above indicates the challenge in providing a clear separation of apoptotic proteins. In summary, candidate proteins for inclusion in the database are defined based on the presence of homology to a known apoptotic domain. These homologs are then clustered based on the closest homology to a set of annotated orthologs. The groups created by these representative orthologs are called families. This homology to a family is based only on each apoptotic domain. For vertebrate proteins, these independent measures of family were consistent over each of the protein's apoptotic domains, when accepting homology with expectation values ( $E$ -values) better than  $10^{-8}$ .



**Figure 3** Phylogeny dendrogram for caspase recruitment domains (CARDs) of several caspases based on a ClustalW alignment. Only the proteins representing each family are shown. Branch lines of this unrooted tree that are not exclusive to a family are shown in bold. A simplified version of this diagram is maintained in the database

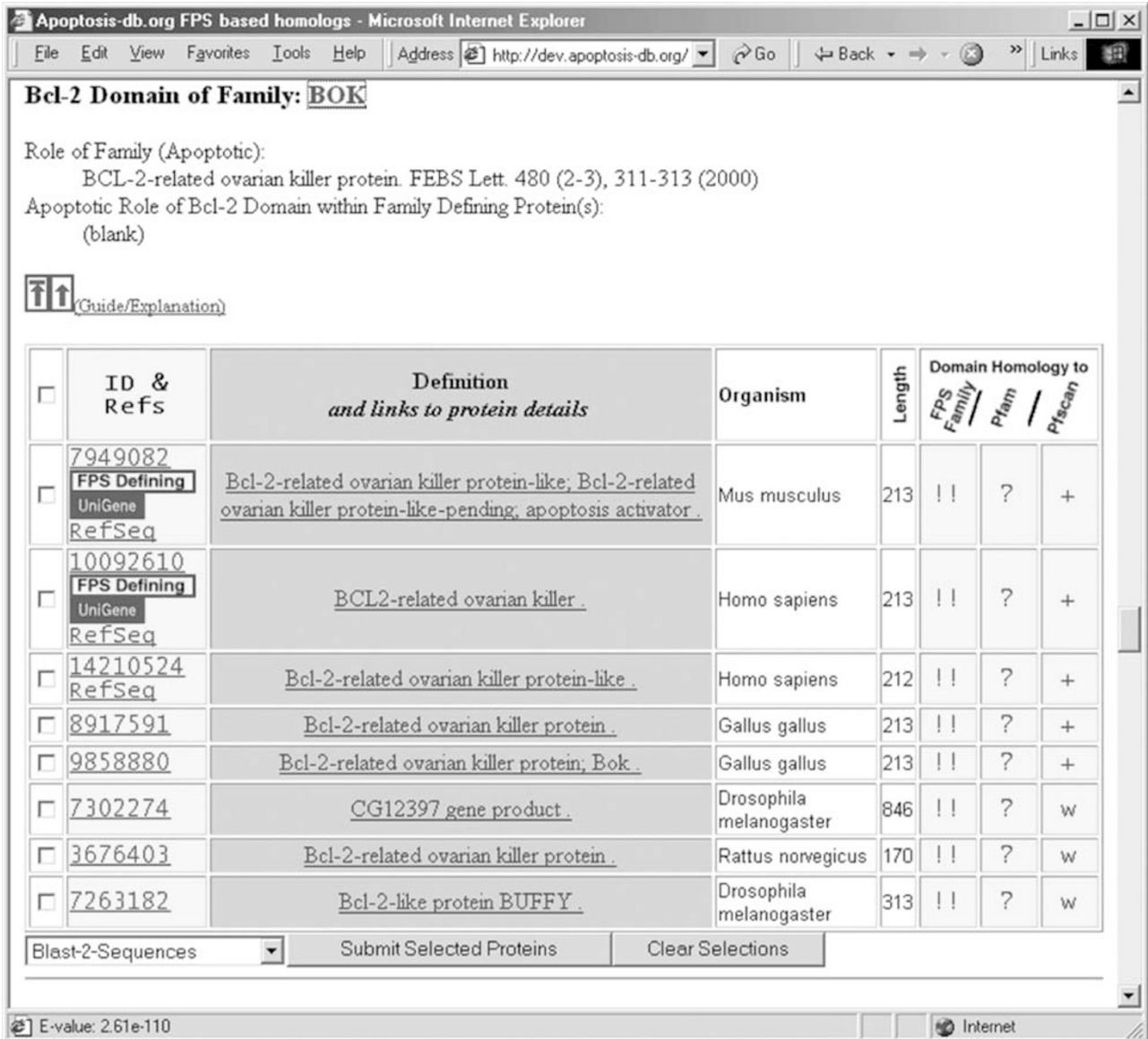
**Table 3** Rank and expectation values of some caspases (rows) to the caspase-1 and caspase-9 FPS families (columns)

Rank and expectation value	Caspase-1 CARD		Caspase-1 large subunit		Caspase-1 small subunit		Caspase-9 CARD		Caspase-9 large subunit		Caspase-9 small subunit	
Human caspase-1 [P29466]	1	$6 \times 10^{-30}$	1	$4 \times 10^{-127}$	1	$2 \times 10^{-70}$	—	—	9	$8 \times 10^{-13}$	—	—
Mouse caspase-1 [P29452]	1	$2 \times 10^{-28}$	1	$4 \times 10^{-122}$	1	$1 \times 10^{-67}$	—	—	8	$5 \times 10^{-14}$	5	$3 \times 10^{-04}$
Human caspase-9 [P55211]	*—	—	6	$2 \times 10^{-23}$	16	$6 \times 10^{-03}$	1	$2 \times 10^{-40}$	1	$9 \times 10^{-86}$	1	$3 \times 10^{-47}$
<i>Xenopus</i> caspase-9 [7619908]	—	—	6	$4 \times 10^{-27}$	15	$9 \times 10^{-09}$	1	$1 \times 10^{-18}$	1	$2 \times 10^{-53}$	1	$2 \times 10^{-34}$
Human caspase-2 [P42575]	—	—	3	$2 \times 10^{-45}$	—	—	—	—	6	$1 \times 10^{-24}$	7	$8 \times 10^{-07}$
<i>Xenopus</i> ICE-A [P55865]	1	$1 \times 10^{-06}$	1	$8 \times 10^{-81}$	1	$5 \times 10^{-42}$	—	—	12	$9 \times 10^{-12}$	10	$1 \times 10^{-01}$
<i>Xenopus</i> ICE-B [P55867]	1	$5 \times 10^{-05}$	1	$4 \times 10^{-82}$	1	$1 \times 10^{-34}$	—	—	13	$1 \times 10^{-13}$	6	$1 \times 10^{-05}$

These are based on the FPS algorithm used to sort proteins into their families. Rank and expectation values denoted as '—' are either above the minimum recording threshold of 0.1 or no alignment was possible.

This is superior to a resource that is reliant on Pfam or Pfscan for domain definitions. For example, the C-terminal motif of AIF proteins (or PCD-8, protein cell death-8) is an apoptotic domain whose apoptotic function is masked in the Pfam and Pfscan databases by another well-known functional role. AIF proteins have an additional well-recognized pyridine nucleotide-disulfide oxidoreductase domain, which is found in

a large number of nonapoptotic proteins. The structure of the whole AIF has recently been solved, confirming a C-terminal structural domain only present in these apoptotic homologs, which is similar to bacterial ferredoxin reductase.<sup>12</sup> This domain was used to define a number of apoptotic homologs, although it is yet not characterized as such in the Pfam or Prosite domain databases.



**Figure 4** Example of a Homolog Listing. The image shows part of the larger listing of all proteins (including minor variants or alternative splicing variants) with significant homology to the Bcl-2 domain. Only the proteins most similar to the bcl-2-related ovarian killer (BOK) family are listed. These include the three proteins used to represent the family (labeled on ID column), several orthologs from other species and possible paralogs

Attempts to use single domains that are repeated in a given protein to establish families can be problematic. Consider the bacterial IAP repeats (BIR) domains of inhibitor of apoptosis protein (IAP) proteins. IAP proteins have between one to three BIR domains. In the human cellular IAP-1 protein, the first repeat has 43–49% sequence identity to the second and third repeats, respectively. However, the first repeat of cellular IAP-1 is 49% identical to the first BIR repeat of human XIAP proteins. That is, percent sequence identities and expectation values do not distinguish between the two groups, c-IAP-1 and X-chromosome-linked inhibitors of apoptosis (XIAP), if single domains are compared. By combining each domain into an FPS family, each functionally distinct group was easily separated allowing homologs from other species, variants,

and alternatively spliced forms to automatically group together. Similarly, repeated DED domains in caspase-8, caspase-10 and FLIP are also distinguishable by this approach.

A recent application of the database was to provide a reference set of proteins for use in the annotation of apoptotic proteins in the RIKEN mouse cDNA collection. The database was used in the first phase of the work, cataloging which human and other vertebrate proteins should have orthologs in mouse. A starting set of 187 apoptotic domains out of final derived set of 294, including alternative splicing forms, were taken from the database and used to query the RIKEN database. While a number of assignments are putative, clearly some domains and motifs must be added to the

database in the future, for example, the BH3 only motif, TNF receptor extracellular cysteine-rich repeats, and the PAAD/PYRIN domain, to provide a more extended coverage of apoptotic domains.

## Database access

The database can be accessed by protein sequence using sequence homology searches with basic local alignment search tool (BLAST), by text string searches, by database ID searches, or by browsing lists of domains, families, or homologs. The lists of homologs are the primary means of accessing individual proteins, and several types of lists are found with alternative organization and filtering.

The list of homologs for an apoptotic domain is the most commonly used list. It shows all the homologs separated into families, including proteins with weak homology to the selected domain (Figure 4). The level of homology for the domain based on FPS, Pfam and Pfscan is symbolized by: !! (strong), + (reliable), w (weak), ? (unreliably weak) or an empty area for undetected homology. The expectation value for each of these is viewable upon 'mousing over' the symbol. The lists are ordered by decreasing homology to each of the families. This order gives important clues as to reliability of the functional annotation for that homolog. Any two proteins can be selected for a pairwise alignment and several members on the list can be selected for multiple sequence alignment using ClustalW. Individual or multiple sequences can be output in FASTA format. All sequence alignments can be performed on the domain or the full-length protein sequence. This list of homologs can also be filtered to removed alternative splice forms or minor variants.

An alternative list of homologs containing an apoptotic domain is organized by taxonomy. In this case, mammalian proteins are shown first, then other vertebrates, then other eukaryotes, then noneukaryotes. This list can also be viewed with either the graphical representation of domain architecture or textual annotation.

Each family has functional annotation describing its role in apoptosis. Also associated with each family is the representative protein and the domains used to establish the family, literature references associated with the family, and functional annotation of each of the separate domains, if known.

Computed details of each protein record are also available:

Homology for domains using Pfam, Pfscan, FPS, and PSI-BLAST.

Details of the variable PSI-BLAST profile/query used to retrieve the protein from the GenBank non-redundant (nr) database.

Possible transmembrane helices (TNF receptors, Bcl-2 family members, and others) calculated by trans-membrane hidden Markov model (TMHMM).

Possible signal cleavage sites calculated using SigCleave and as a peripheral output of TMHMM.

Predicted alternative splice forms or minor variant relations among proteins.

## Discussion

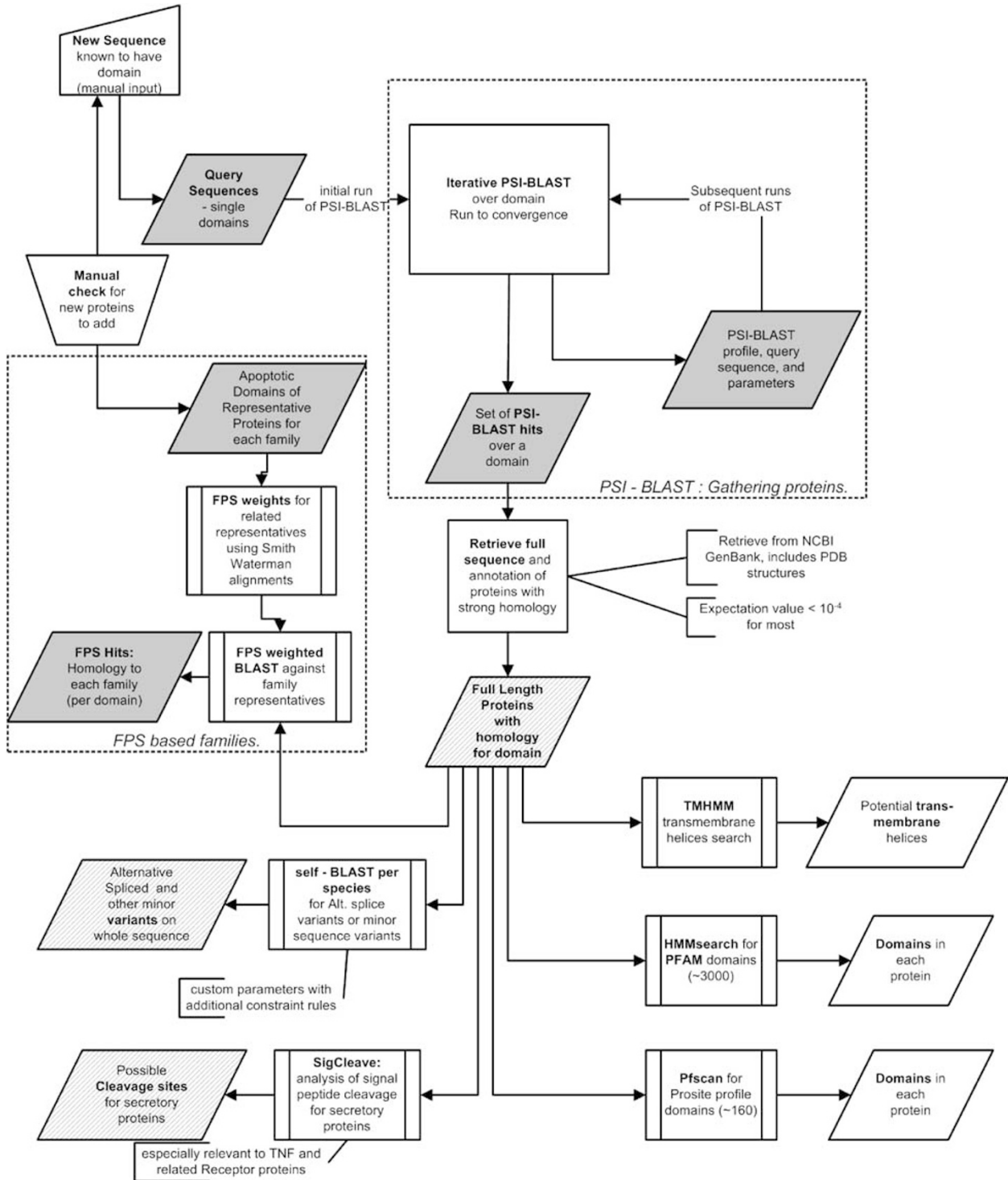
The field of apoptosis has become one of the fastest growing areas of biomedical research, as tracked by numbers of publications devoted to the topic, according to the ISI.<sup>13</sup> Consequently, we have established a database consisting of a set of proteins and their distinctive domains that are often, if not exclusively, involved in the apoptosis signaling pathway. The database provides a domain structure classification based on SCOP, a classification of domains based on sequence and a classification of protein families based on combinations of domains. Each classification has its own set of functional annotation. Starting with manually selected and curated apoptotic proteins, additional proteins are automatically added using bioinformatics techniques.

Future work includes the addition of new domains, for example, the PAAD domain<sup>14</sup> if sufficient evidence amasses to justify their inclusion as true apoptotic domains. Further, features other than structural domains will be used to identify and categorize apoptotic proteins. For example, second mitochondria-derived activator of caspase (Smac/DIABLO) and its functional counterparts in *Drosophila* share only a short N-terminal motif that is required for promoting apoptosis by binding IAPs. In Smac, the N-terminal four amino acids (after loss of a targeting domain) are required for the activity of the full-length protein and seven amino acids are sufficient for activity as a peptide.<sup>15</sup> Methods such as MEME<sup>16</sup> search and recognize biologically relevant short motifs of this nature and will expand the characterization methodology used in a future version of the database.

Domain structure requires further annotation within the database. First, domains for which structures exist and have apparent structure homology are not indicated. Moreover, many sequences for which structures do not exist could be modeled by comparative (homology) modeling, or fold recognition and structural annotation provided. Such structure annotation could be automated.

Currently, the detailed clustering based on orthologous sets of proteins is fully supervised as far as choosing clusters. In the future, we will use general trends observed for each domain or family to automatically suggest changes to the representative proteins of the families as the sequence database is updated. For instance, when a clear closest homolog is present between a single set of mammalian orthologs and *Xenopus*, the mammalian proteins could be tested for forming a family. The divergence rates of caspase orthologs (or most distinctively similar homologs) among vertebrate species has been studied (data not shown). The results showed that each orthologous group had independent divergence rates for vertebrate caspase large subunits (the 'P20' domain). Some sets of orthologs had similar divergence rates; however, those ortholog sets did not have in common either domain architecture, apoptotic functional roles, nor group specificity. Thus, as orthologs are found between distant species, they can be used to predict divergence rates, but only on a per-family, and not per-domain basis.





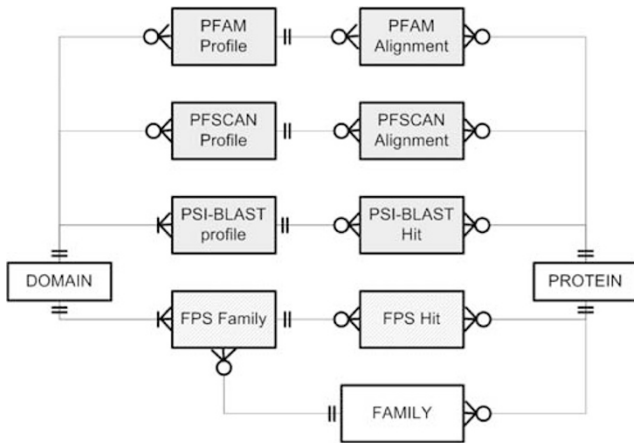
**Figure 5** Process diagram for the creation and update of the apoptosis database. Rectangles are processes, rectangles with additional vertical borders are process developed elsewhere and incorporated into the database with only parametric or configuration changes. Parallelograms denote data that are incorporated into the database. Stripes denote data over whole length proteins while gray filling denotes data on single domains. Asymmetric shapes (trapezoids) are manual curation processes

## Materials and Methods

Figure 5 outlines the process of gathering, annotating, and classifying apoptotic proteins into families.

## Gathering proteins into the database

The selection of proteins for inclusion is based on homology to apoptotic domains (Table 2). PSI-BLAST<sup>17</sup> version 2.1.3 running against the NCBI



**Figure 6** Relational database schema for profile- and FPS-based domain assignments. The schema is shown in standard 'crows foot' notation. Each rectangle represents a table (relation). The relation between records in two tables is represented by the symbols near each pair. Double crossing lines denotes one and only one record must be present in that table. A single crossing line with three forking lines denotes one or more corresponding records must be present. A circle with three forking lines denotes that zero or more corresponding records are allowed. Profile based assignments (gray) are made per domain onto full-length protein sequences (PROTEIN). An apoptotic domain (DOMAIN) can have 0 or more Pfam and Pfscan profiles representing it, but must have at least one PSI-BLAST and FPS family representation. The FPS families are defined over individual domains of the representative full-length sequences (FAMILY)

nr database is used to create dynamic profiles. The PSI-BLAST profiles are robust and dynamic enough to gather all known domains when several (3–17) profiles are used to represent each domain. The initial round of iterative profile building always starts with a single seed sequence. This seed comes from interaction with experts in the field, literature, standard curated profile analysis from Pfam or Prosite, or prior PSI-BLAST results. Each PSI-BLAST run is independent – allowing for cases when the sequence homology between seeds for the same domain is very poor. In cases where many closely related homologs can dominate a profile, a high-expectation filter for inclusion into the profile combined with appropriate choice of seeds allows for rarer homologs to maintain a signal in the profile. These parameters are manually set and fixed for all the PSI-BLAST iterations. Most often, PSI-BLAST iterations are taken to convergence – no changes occur in the sequences defining the profile between iterations.

The profiles, parameter sets, and seed sequences are stored in the database. Figure 6 illustrates part of the database schema detailing the relation between profiles, alignments and associated domains, proteins, and families. Updates of the GenBank nr protein sequence database are periodically used to update the apoptosis database. The stored profiles are used to determine what new homologs to add with a resultant incremental changes to the profiles. Parameter changes and new seeds determined by experts are also incorporated as needed during updates. The current set of seeds and PSI-BLAST parameters used to populate the database are included in the Appendix.

### Domain annotation

Proteins with apoptotic domains found using PSI-BLAST are retrieved from GenBank using Perl scripts that include the Boulder::Genbank module. The protein full-length sequence, organism (taxonomy) and other information are retrieved from the GenBank record at NCBI. These protein

sequences are then annotated for the presence of domains, transmembrane helices and possible signal peptide cleavage sites (Figure 5).

HMMsearch against the Pfam database (>3000 domains) and Pfscan (Not published. Available at URL: <http://www.isrec.isb-sib.ch/software/PFSCAN/form.html>) against the Prosite profiles database (161 domains) are used to assign domains. In most cases, the apoptotic domain is also profiled in Pfam and/or Prosite profiles database. These static, curated profiles help confirm assignments made using PSI-BLAST, but the PSI-BLAST profiles cover additional (often newer) homologs that are not yet represented by these standard profiles. In cases where Pfam or Prosite finds additional homologs for an apoptotic domain, new PSI-BLAST runs are created to cover this range of new homologs. The hits resulting from all domains in Pfam and Prosite profiles are stored in the database along with version information for the Pfam and Prosite databases.

The definition of structural superfamilies is taken from SCOP. Currently, any one sequence per domain that corresponds to a protein structure recorded in the PDB (as reported by NCBI GenBank sequence records) is checked for representation in SCOP.

### Protein features

Transmembrane helices are detected using the TMHMM program.<sup>18</sup> The output is parsed and stored into the database with individual sequence ranges labeled as intracellular, extracellular, or transmembrane.

UniGene records for the proteins in the database do not come directly from the retrieved GenBank entries. The entire UniGene database from NCBI is downloaded, parsed, and analyzed for relations to current or former GenBank proteins in the nr database. Matches that cannot be made directly via GI number are performed by matching both the sequence and source organism exactly.

Alternative splice forms and minor protein variants from the same species are established using the BLAST algorithm (Figure 5). All the proteins in the database are separated into species-specific databases. Each protein is tested for BLAST hits against its species with greater than 96% identity and allowing for mismatches at the ends of the alignments. The representative for a set of alternative splice forms/variants is the longest sequence taken from all SwissProt, UniGene, and RefSeq sequences.

### FPS clustering

The FPS algorithm calculates the expectation value of a query sequence against a set of FPS families. Each FPS family consists of one or more sequences and its normalization weight. In our case, the FPS families are the apoptotic domains of manually selected full-length representative sequences (usually the human and mouse orthologs, if available). The expectation value of a query sequence to a family is the weighted product of *P*-values from alignments (BLAST) to each sequence defining the FPS family. Normalization of *P*-values must be performed, since the sequences defining FPS families are related (thus, not independent measurements). The normalization weights are estimated by Smith–Waterman alignments against a random library of 995 sequences from the PDB database.

All proteins retrieved by PSI-BLAST searches for apoptotic domains are categorized using this FPS algorithm. Although the best family for each domain (with an expectation value cutoff of  $10^{-8}$ ) is sufficient for categorizing each protein, the expectation values to all families are stored in the database. This richer representation is useful for representing cases

where classification into a single family is ambiguous. These cases are later evaluated for the possible establishment of new FPS families.

References are related to several entities in the database: families, FPS- families, domains, proteins, and multiple sequence alignments, and diagrams. Several types of static diagrams (such as Figure 1 and phylogeny dendrograms similar to Figures 2 and 3) are linked to the domains, which they describe and are accompanied by the readable text in the images in the database, which allows for matching relevant diagrams via text searches over the database.

The system is implemented using the MySQL (MySQL AB) relational database management system, with users accessing both static and dynamic web pages created by Perl scripts and a Perl module (library) specific for the database. Other Perl modules used include XML::Twig (Michel Rodriguez, xmlltwig.com) for parsing XML data from PSI-BLAST; DBI:: and DBD::mysql (Tim Bunce, Jochen Wiedmann) for interfacing to the database; Boulder:: (Lincoln Stein) and Bio:: (bioperl.org) for sequence and literature references retrieval and parsing; and LWP:: and HTML::TokeParser (by Gisle Aas) for accessing and retrieving information on external web sites.

Two versions of the database, a development and a production version of the database are maintained at all times. The production version is static and accessible to all uses of the Internet from the URL <http://www.apoptosis-db.org>. The development database is dynamic with software upgrades and data added on a regular basis. Periodically, the development version is copied to create a prerelease version. Apoptosis experts review the prerelease version with emphasis on those proteins gathered automatically. When the experts are satisfied the prerelease version becomes the public production database.

## Acknowledgements

This research was supported by Grant NSF DBI 0078731 from the National Science Foundation, Division of Biological Infrastructure. We thank Professor Guy Salvesen for expertise in the field of caspases. We thank Dr. Darek Kedra and Dr. Greg Quinn for developing and maintaining the core bioinformatics computational infrastructure at the Burnham Institute and for providing programming advice and expertise.

## References

1. Pruitt KD and Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* 29: 137–140
2. Rebhan M, Chalifa-Caspi V, Prilusky J and Lancet D (1997) GeneCards: encyclopedia for genes, proteins and diseases. (Rehovot, Israel). Weizmann Institute of Science, Bioinformatics Unit and Genome Center, World Wide Web URL: <http://bioinformatics.weizmann.ac.il/cards>

3. Bairoch A and Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28: 45–48
4. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL and Sonnhammer EL (2000) The Pfam protein families database. *Nucleic Acids Res.* 28: 263–266
5. Schultz J, Copley RR, Doerks T, Ponting CP and Bork P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28: 231–234
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res.* 28: 235–242
7. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* 30: 264–267
8. Grutter MG (2000) Caspases: key players in programmed cell death. *Curr. Opin. Struct. Biol.* 10: 649–655
9. Alnemri ES, Livingston DJ, Nicholson DW, Salvesen G, Thornberry NA, Wong WW, Yuan J (1996) Human ICE/CED-3 protease nomenclature. *Cell* 87: 171
10. Grundy WN and Bailey TL (1999) Family pairwise search with embedded motif models. *Bioinformatics* 15: 463–470
11. Feng DF, Cho G, Doolittle RF (1997) Determining divergence times with a protein clock: update and reevaluation. *Proc. Natl. Acad. Sci. USA* 94: 13028–13033
12. Maté MJ, Ortiz-Lombardía M, Boitel B, Haouz A, Tello D, Susin SA, Penninger J, Kroemer G and Alzari PM (2002) The crystal structure of the mouse apoptosis-inducing factor AIF. *Nat. Struct. Biol.* 9: 442–446
13. Thomson ISI Essential Science Indicators, ESI Special Topics; <http://www.esi-topics.com>
14. Pawlowski K, Pio F, Chu Z, Reed JC and Godzik A (2001) PAAD – a new protein domain associated with apoptosis, cancer and autoimmune diseases. *Trends Biochem. Sci.* 26:85–87
15. Silke J, Verhagen AM, Ekert PG and Vaux DL (2000) Sequence as well as functional similarity for DIABLO/Smac and Grim, Reaper and Hid. *Cell Death Differ.* 7: 1275
16. Bailey TL and Gribskov M (1998) Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics* 14:48–54
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402
18. Krogh A, Larsson B, von Heijne G and Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305:567–580

## Appendix A

List of PSI-BLAST (run locally using blastpgp) parameters used to gather homologs for a domain is summated in Table 4. The following parameters common to all PSI-BLAST runs: maximum number of passes (iterations) – 10; filter type – ('m S'); output type – '7', XML output; processors – 4. All other parameters were defaults for blastpgp.

Table 4 List of PSI-BLAST parameters

Blast ID	Domain	NCBI gi	Expectation cutoff	Converged	Query start	Query end
157	AIF/PCD-8 C-term Motif	4757732	0.001	Yes	490	614
78	BAG domain	3523107	$1 \times 10^{-06}$	Yes	156	209
79	BAG domain	7304915	$1 \times 10^{-05}$	Yes	448	498
80	BAG domain	7504500	$1 \times 10^{-05}$	Yes	134	192
81	BAG domain	7504919	$1 \times 10^{-06}$	Yes	397	447
82	BAG domain	7491412	$1 \times 10^{-05}$	Yes	135	185
158	BAG domain	4757834	$1 \times 10^{-05}$	Yes	109	190
83	Bcl-2 domain	4757840	$1 \times 10^{-12}$	Yes	1	147
84	Bcl-2 domain	4502363	$1 \times 10^{-15}$	Yes	1	184
85	Bcl-2 domain	14760604	$1 \times 10^{-14}$	Yes	1	171
86	Bcl-2 domain	12741054	$1 \times 10^{-14}$	Yes	1	210
87	Bcl-2 domain	4757842	$1 \times 10^{-14}$	Yes	1	165
88	Bcl-2 domain	14772457	$1 \times 10^{-15}$	Yes	1	209
89	Bcl-2 domain	6980395	$1 \times 10^{-12}$	Yes	1	198
90	Bcl-2 domain	6708482	$1 \times 10^{-12}$	Yes	97	238
91	Bcl-2 domain	9966783	$1 \times 10^{-08}$	Yes	1	204
92	Bcl-2 domain	7662506	$1 \times 10^{-08}$	Yes	1	141
159	Bcl-2 domain	14161579	$1 \times 10^{-06}$	Yes	1	143
93	BIR domain	4502141	$1 \times 10^{-15}$	Yes	49	115
94	BIR domain	4502141	$1 \times 10^{-12}$	Yes	187	252
95	BIR domain	2497243	$1 \times 10^{-12}$	Yes	44	111
96	BIR domain	2497243	$1 \times 10^{-12}$	Yes	226	294
97	BIR domain	4502143	$1 \times 10^{-10}$	Yes	26	94
98	BIR domain	6322548	$1 \times 10^{-12}$	Yes	21	118
99	BIR domain	6322548	0.0001	Yes	153	242
120	Caspase large subunit	4502569	$1 \times 10^{-05}$	Yes	239	363
121	Caspase large subunit	266321	$1 \times 10^{-05}$	Yes	161	289
122	Caspase large subunit	14741608	0.0001	Yes	181	308
123	Caspase large subunit	14731648	0.0001	Yes	232	364
124	Caspase large subunit	12644321	0.001	Yes	159	292
125	Caspase large subunit	542466	0.0001	Yes	241	363
126	Caspase large subunit	3462860	0.0001	Yes	257	387
127	Caspase large subunit	4633107	0.0001	Yes	194	323
128	Caspase large subunit	5803078	0.001	Yes	331	469
100	CARD	4388927	0.0002	Yes	1	101
101	CARD	4505419	0.002	Yes	1	112
102	CARD	4502379	0.002	Yes	16	96
103	CARD	13488607	0.002	Yes	24	116
104	CARD	14165284	0.002	Yes	12	103
105	CARD	2506262	0.002	Yes	1	101
106	CARD	231729	0.002	Yes	1	91
107	CARD	5668628	0.002	Yes	1	101
108	CARD	4633107	0.002	Yes	11	110
109	CARD	11545922	0.002	Yes	111	201
110	CARD	11545922	0.002	Yes	111	201
111	CARD	14774647	0.002	Yes	420	496
112	Caspase small subunit	4502569	0.001	Yes	387	480
113	Caspase small subunit	14770290	0.001	Yes	317	404
114	Caspase small subunit	14741608	0.001	Yes	337	435
115	Caspase small subunit	1708949	0.001	Yes	196	294
116	Caspase small subunit	14731648	0.001	Yes	385	479
117	Caspase small subunit	1168878	0.001	Yes	405	504
118	Caspase small subunit	3462860	0.001	Yes	404	495
119	Caspase small subunit	4633107	0.001	Yes	357	451
129	CIDE N-terminal	6730242	0.001	Yes	1	116
130	CIDE N-terminal	6066229	0.001	Yes	1	97
131	CIDE N-terminal	6753628	0.001	Yes	1	105
132	Death domain	7505816	0.002	Yes	1306	1399
133	Death domain	585366	0.0002	Yes	39	121
134	Death domain	4507583	0.0001	Yes	218	336
135	Death domain	14768299	0.0002	Yes	19	103
136	Death domain	5706378	0.002	Yes	27	128
137	Death domain	4505297	0.0002	Yes	29	105
138	Death domain	10130019	0.001	Yes	789	873
139	Death domain	4506539	0.0001	Yes	583	669
140	Death domain	4507575	0.0002	Yes	356	441
141	Death domain	6678505	0.0001	Yes	847	932
142	DED	1082245	$2 \times 10^{-05}$	Yes	172	247
143	DED	3193167	$2 \times 10^{-05}$	Yes	100	184
144	DED	3193167	$2 \times 10^{-05}$	Yes	2	87
145	DED	1914847	$2 \times 10^{-06}$	Yes	2	72

Table 4 (Continued)

Blast ID	Domain	NCBI gi	Expectation cutoff	Converged	Query start	Query end
146	DED	4758144	$2 \times 10^{-05}$	Yes	24	110
147	DED	1167559	0.002	Yes	1	86
148	DED	2498752	0.0001	Yes	3	87
149	NB-ARC domain	4502123	$1 \times 10^{-12}$	Yes	93	405
150	NB-ARC domain	477515	$1 \times 10^{-10}$	Yes	119	430
151	NB-ARC domain	6456486	$1 \times 10^{-12}$	Yes	101	401
152	TRAF domain	13650033	$1 \times 10^{-07}$	Yes	265	414
153	TRAF domain	4959432	$1 \times 10^{-06}$	Yes	299	487
154	TRAF domain	5032193	$1 \times 10^{-06}$	Yes	239	416
155	TRAF domain	10863939	$1 \times 10^{-10}$	Yes	324	501
156	TRAF domain	2138180	$1 \times 10^{-10}$	Yes	356	539
160	TNF receptor cysteine-rich repeat	4507575	0.01	No	44	195