# Apoptotic Molecular Machinery: Vastly Increased Complexity in Vertebrates Revealed by Genome Comparisons

## L. Aravind,[1] Vishva M. Dixit,[2] Eugene V. Koonin[1]*

A comparison of the proteins encoded in the recently (nearly) completed human genome to those from the fly and nematode genomes reveals a major increase in the complexity of the apoptotic molecular machinery in vertebrates, in terms of both the number of proteins involved and their domain architecture. Several components of the apoptotic system are shared by humans and flies, to the exclusion of nematodes, which seems to support the existence of a coelomate clade in animal evolution. A considerable repertoire of apoptotic protein domains was detected in Actinomycetes and Cyanobacteria, which suggests a major contribution of horizontal gene transfer to the early evolution of apoptosis.

Comparison of genome sequences—or more precisely, of the protein sequences encoded in genomes—is a potentially powerful tool for identifying the components of functional systems and reconstructing their evolution. Such comparisons allow researchers to transfer information from well-studied model organisms to poorly characterized ones and to draw functional and evolutionary inferences from the presence, absence, and relative abundance of genes coding for different types of proteins in the compared genomes. Programmed cell death (apoptosis) is one of the central cellular processes in development, the stress response, aging, and disease in multicellular eukaryotes (1). Comparative analysis of the components of the apoptotic machinery have shown that many of the protein domains that perform critical roles in this system were already present in the common ancestor of animals, plants, and fungi (2). From the functions of the extant proteins containing these conserved domains, it can be extrapolated that they participated in ancestral signaling pathways, including those for pathogen and stress responses. The evolution of the programmed cell death system from such signaling pathways was probably driven by general kin selection during the emergence of multicellularity.

Here, we briefly discuss the results of a comparative analysis of the nearly complete protein sets of *Homo sapiens*, *Drosophila*

*melanogaster*, and *Caenorhabditis elegans* (3). It is only with the near-completion of the human genome sequence that such a comparison is poised to present an accurate picture of the relationships between the programmed cell death systems in vertebrates and invertebrates, and the results show a strikingly increased complexity of the apoptosis machinery in the former.

The evolutionary engineering of the apoptotic system followed the same pattern as seen in other signal transduction and regulatory systems, particularly in eukaryotes, namely the formation of a wide variety of protein domain architectures from a relatively small set of ancient conserved domains (4). Therefore, we applied a domain-centered approach to the comparative study of this system in animals. First, the occurrences of the individual domains in apoptotic proteins were enumerated as accurately as possible by using a sensitive sequence analysis method based on the information contained in the multiple alignments of the corresponding protein sequences (4). Second, the domain architectures of the apoptotic proteins identified in humans, flies, and nematodes (and, if applicable, other organisms) were systematically compared. A panoply of proteins with functions in almost every basic cellular process have been directly or indirectly linked to apoptosis, which is not too surprising because programmed cell death is a complicated series of events involving various cellular subsystems. Nevertheless, here we restrict the discussion to the central participants of cell death signaling and execution and their homologs that might shed light on the origin and evolution of apoptotic mechanisms.

Complete lists of the Gene Identifiers (GI numbers) for all detected components of the apoptotic machinery, including a brief anno-

tation of the domain architecture for each of the proteins, are available at ftp://ncbi.nlm.nih.gov/pub/koonin/PCD.

Examination of the number of occurrences of several domains that perform central functions in apoptosis shows a marked expansion in vertebrates relative to insects and nematodes (Table 1). The growth in the number of these domains detectable in humans was noticed in all functional categories of proteins that contribute to programmed cell death, but was particularly striking among the extracellular components of the apoptotic system (ligands and receptors), the intracellular adaptor domains that transfer the signal from the receptors to the executors of apoptosis (such as caspases), the BCL2 family of apoptosis regulators, and the NACHT family of nucleoside triphosphatases (NTPases).

In vertebrates, several secreted ligands, primarily members of the tumor necrosis factor (TNF) family, directly induce apoptosis (5). A single, previously undetected member of the TNF family was identified in *Drosophila*, which suggests that this ligand was already present before the divergence of the coelomates (6). The TNF family proteins function through specific receptors (TNFRs) that contain multiple repeats of an extracellular cysteine-rich domain and an intracellular Death domain (DD). Predicted receptors with a single copy of the cysteine-rich domain are present in *Drosophila*, *C. elegans*, and plants (6), but none of them has the same architecture as the vertebrate TNFRs, and accordingly these proteins cannot be considered TNFR orthologs (direct evolutionary counterparts).

The transmission of the external cell death stimuli to the executors of apoptosis, such as the caspases, is largely mediated by several specialized adaptor domains. The most prominent apoptotic adaptors are the CARD, DED, pyrin, and Death domains that have a common fold with six $\alpha$ helices and probably evolved from a common ancestor before the divergence of the extant animal lineages (2, 7). A conspicuous expansion in the number of distinct proteins that contain these domains, particularly the CARD domain, is seen in humans (Table 1). Furthermore, a previously undetected version of the $\alpha$-helical adaptor module, the pyrin domain, which was predicted during the present protein sequence analysis, appears to be vertebrate-specific. The pyrin domain was identified in pyrin, Asc (a CARD-domain protein), the interferon-induced protein 16, the AIM2 protein, a caspase from zebrafish, and some uncharacterized proteins that also contain the NACHT NTPase domain (8).

The BCL2 family proteins are conserved in all animals and have been implicated in alteration of mitochondrial permeability resulting in leakage of cytochrome c and the

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. [2]Department of Molecular Oncology, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080, USA.

*To whom correspondence should be addressed. E-mail: koonin@ncbi.nlm.nih.gov

triggering of apoptosis (9). The antiapoptotic members of this family interact with the Ced4-like apoptotic adenosine triphosphatases (AP-ATPases) and inhibit their function in caspase activation (10), whereas the proapoptotic family members (for example, BAK) apparently interact antagonistically with antiapoptotic forms. Consistent with the experimental data, Ced-9 from *C. elegans* and the poriferan BCL2 homolog cluster with the antiapoptotic members of the family in phylogenetic analyses, whereas the fly BCL2s cluster with the proapoptotic versions (11). Thus, the differentiation between the proapoptotic and antiapoptotic members of the BCL2 family might have been established

in the coelomates, whereas the ancestral form apparently functioned only in the antiapoptotic capacity. The vertebrates show a proliferation of both these versions, with extreme sequence divergence observed in several proapoptotic members such as Bid and Mil1 (10).

Another group of proapoptotic proteins—the so-called BH3-only proteins, which share only a region of limited sequence similarity (the BH3 motif) with the BCL2 family proteins—have recently attracted considerable attention (12). The BH3-only proteins interact with antiapoptotic members of the BCL2 family via an amphipathic helix formed by the BH3 motif and inactivate them. The trend

toward diversification in vertebrates seems to hold among the BH3-only proteins because several proteins of this group have been identified in mammals, as opposed to only one found thus far in *C. elegans* (EGL-1) (12). However, there is no statistically significant similarity between diverse BH3 proteins, nor is the motif itself prominent enough to allow reliable sequence-based predictions in genome-wide searches. It remains unclear whether all reported occurrences of the BH3 motif are functionally relevant and whether the BH3-only proteins share a common ancestry.

The NACHT NTPases that appear to be a sister group of the well-characterized AP-

**Table 1.** Domains and proteins involved in apoptosis and related pathways. The number of detected proteins containing each domain is indicated for each organism. H indicates the presence of homologous domains, but not orthologs.

| Protein/domain family | Vertebrates (human) | Arthropods (*Drosophila*) | Nematodes (*C. elegans*) | Others |
|---|---|---|---|---|
| *Receptors* | | | | |
| TNFR | 8 | H | H | H (plants) |
| IL1-like | 8 | 0 | 0 | 0 |
| Toll-like | 10 | 8 | 0 | |
| *Ligands* | | | | |
| TNF | 17 | 1 | 0 | 0 |
| Cysteine knots | TGF-like: 12; NGF-like: 3 | TGF-like: 3; Spaetzle-like: 3; NGF-like: 1 | 0 | 0 |
| Adaptors (six-α-helix domains) | | | | |
| Death | 30 | 9 | 6 | 0 |
| DED | 7 | 1 | 0 | 0 |
| CARD | 20 | 1 | 2 | 0 |
| Pyrin | 8 | 0 | 0 | 0 |
| Adaptors (other) | | | | |
| TIR | 22 | 10 | 1 | *Arabidopsis* (~135); *Streptomyces* (4); 1 each in a number of other bacteria |
| MATH (TRAF-like) | 6 | 3 | 1 | *Dictyostelium* (3); MATH domains found in all other eukaryotes |
| BCL2-family | 11 | 2 | 1 | |
| *Enzymes* | | | | |
| Caspases | Classic caspase: 14 (one inactive); paracaspase: 1 | Classic caspase: 7 | Classic caspase: 4; paracaspase: 1 | *Arabidopsis* (metacaspase: ≥10); yeast, *Plasmodium*, *Leishmania* (metacaspase: ≥1); *Anabaena* (metacaspase: ≥6); one metacaspase in several other bacteria; *Dictyostelium* (paracaspase: ≥1) |
| A20 | 3 | 1 | 1 | |
| Kinases | IKK: 4; DAP: 1; NIK: 1; IRAK: 4 | IKK: 2; NIK: 1; IRAK: 1 | DAP: 1; IRAK: 1 | H |
| NTPases | | | | |
| AP-ATPase | 1 | 1 | 1 | *Arabidopsis* (~173); *Streptomyces* (8); *M. tuberculosis* (6); 1 each in a number of other prokaryotes |
| NACHT | 18 (NAIP-like: 17; TP1-like: 1) | 2 (TP1-like: 1; a distinct form: 1) | 1 (TP1-like) | *Streptomyces* (3); *Synechocystis* (2); *Anabaena* (6) |
| D-GTPase | 2 | 1 | 2 | *Arabidopsis* (1) |
| *Nuclear factors* | | | | |
| NFκB | 5 | 3 | H | H (plants, fungi) |
| NFAT | 6 | 1 | H | H (plants, fungi) |
| P53 | 3 | 1 | 0 | 0 |
| E2F | 8 | 2 | 3 | *Arabidopsis* (6) |
| DP1 | 5 | 1 | 1 | *Arabidopsis* (2) |
| STAT | 6 | 1 | 4 | *Dictyostelium* (1) |
| RB | 3 | 1 | 1 | *Arabidopsis* (1) |
| CAD | 5 | 4 | 0 | 0 |
| BIR | 8 | 4 | 2 | Yeast (1) |

ATPases, such as Apaf1 and Ced4, have been identified as participants in diverse regulatory interactions including activation of the transcription factor NFκB and apoptosis regulation (CARD4/NOD1 and NAIP), transcription regulation (CIITA), and telomerase function (TP1) (13). All animals encode a TP1 ortholog that probably regulates the telomerase function and may not be involved in apoptosis (14). In addition, analysis of the human protein set shows a major, previously undetected expansion of the NACHT NTPase family; all newly detected members are

closely related to CIITA, NAIP, and CARD4 rather than to TP1. In mice, a locus containing several highly conserved NAIP paralogs affects the survival of the pathogenic bacterium *Legionella* in macrophages (15). Thus, the additional human NACHT NTPases might function in the regulation of immune response and related apoptotic processes.

The increase in the number of proteins containing apoptosis-associated domains is accompanied by diversification of their domain architectures, including the emergence of a considerable number of lineage-specific

architectures, particularly in vertebrates (Figs. 1 and 2). Together with the numerical expansion, this amounts to a major increase in the complexity of the apoptotic system (Fig. 2). The diversification of domain architectures and increase in overall complexity are particularly remarkable in the case of the DD-fold adaptor domains that contribute to many domain combinations unique to the vertebrates, in addition to those formed by the vertebrate-specific offshoot of this fold, the pyrin domain (Figs. 1 and 2). The NACHT NTPases also show diversification
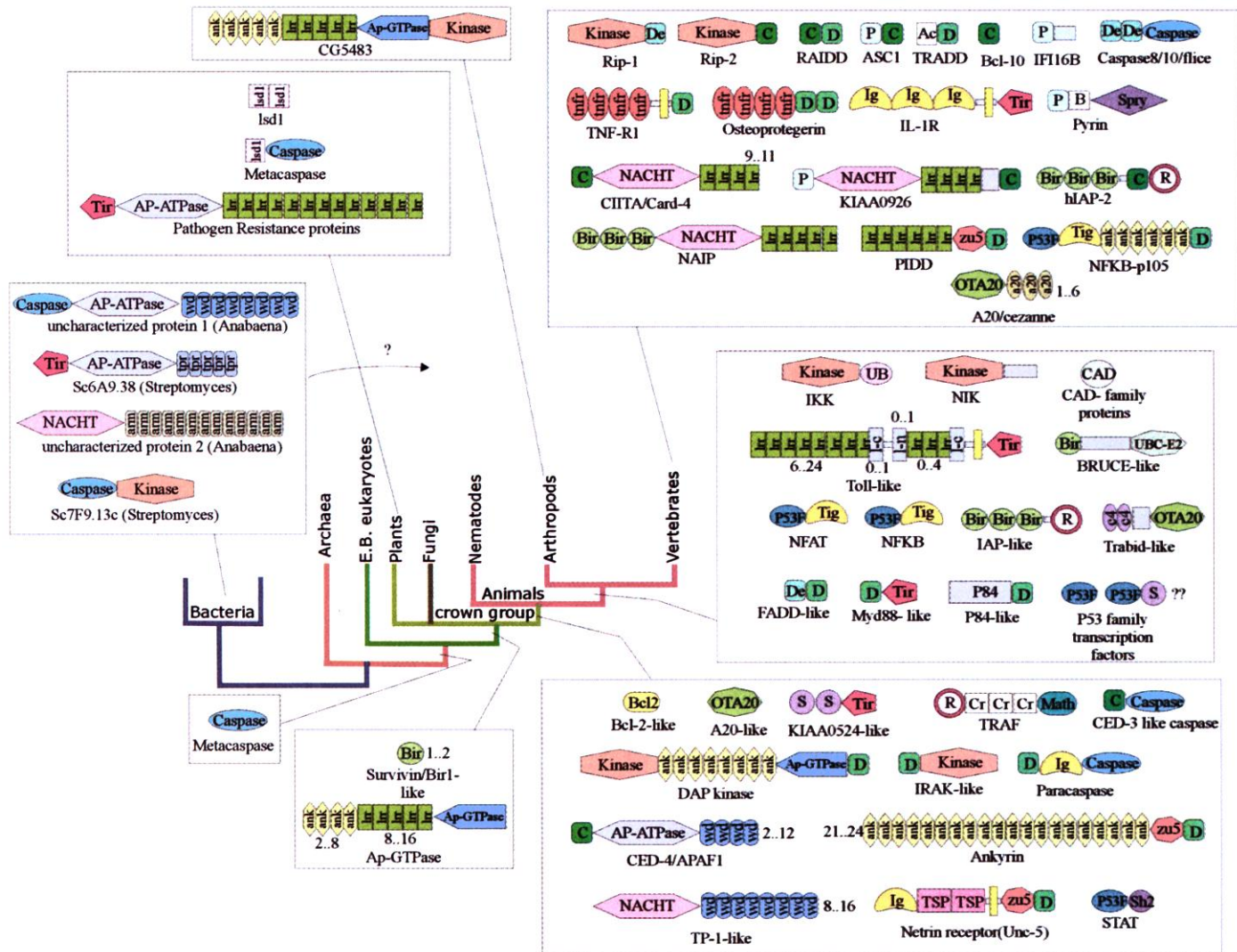


**Fig. 1.** Domain architectures of apoptotic proteins and their advent in evolution. The evolutionary tree for the major divisions of life is shown under the assumptions of an archaeal-eukaryotic clade (28) and a coelomate clade (see text). Each box shows the domain architectures of proteins that are either specific to a particular lineage (for example, vertebrates) or are shared by the two lineages coming out of a given internal node (for example, vertebrates and arthropods) and therefore inferred to have been present in their common ancestor. The direction of probable horizontal transfer of genes encoding homologs of apoptotic proteins is tentatively shown by an arrow pointing from bacteria to eukaryotes. E.B., early branching (eukaryotes). Domain name abbreviations: De, Death effector domain; D, Death domain; C, Card domain; P, pyrin domain; Ig, immunoglobulin domain; Tig, transcription factor immunoglobulin domain (as in NFκB); P53F, P53 fold all–β-strand domain; a20, A20-like Zn-finger; OTA20, OTU-A20–like predicted protease domain; S, SAM domain; R, RING finger; Cr, cysteine-rich domain; Math, meprin-associated Traf homology domain; Tir, Toll–interleukin 1 (IL1) receptor domain; ank, ankyrin repeat domain; TSP, thrombospondin domain; wd, WD40 propeller domain; zu5, zona pellucida Unc-5 domain; P84, conserved domain in the human P84 protein; c4, C4 "little" finger domain; CAD, common domain found in CAD and ICAD; UB, ubiquitin domain; UBC-E2, ubiquitin-conjugating E2 enzyme; tnfr, cysteine-rich domain in TNFR; Spry, SplA-ryanodine receptor domain; B, B-box domain; lrr, leucine-rich domain; arm, armadillo repeat; tpr, tetratricopeptide repeat; lsd1, plant hypersensitive response protein LSD1-like Zn-finger domain.

of the domain architectures in vertebrates through the addition of a variety of domains, such as BIR repeats, CARD, and pyrin, to an ancestral core that consists of a NACHT domain and leucine-rich repeats (Fig. 1).

In addition to the expansion and diversification of proteins containing evolutionarily conserved domains, several proteins with no detectable homologs outside the respective lineages have been implicated in apoptosis. These include the proapoptotic protein SMAC (Diablo) that is specific to vertebrates (*16*) and three small proteins with a similar hydrophobic NH$_2$-terminal peptide—Reaper, Grim, and Hid (*17*)—in *Drosophila*. These appear to have evolved largely from compositionally biased, nonglobular, or predominantly α-helical proteins through selection of specific peptides for interactions with other proteins.

From the wealth of genomic information, the evolution of the cell death pathways in animals can now be reconstructed in some detail (Fig. 1). As noticed previously, several of the key domains of this system were apparently present in the common ancestor of the eukaryotic crown group (*2, 18*). These ancient homologs of apoptotic proteins include enzymes such as the caspases [which were probably represented by an ancestral form resembling the extant plant and fungal metacaspases (*19*)], the predicted A20-like protease (*20*), AP-ATPase, NACHT NTPase, and the previously undetected apoptotic guanosine triphosphatase (AP-GTPase) (*21*); adaptors such as TIR, BIR, and MATH; and nuclear factors such as E2F, Rb, and signal transducers and activators of transcription (STATs). Some of these, such as the BIR-

domain protein survivin, appear to have an ancient function related to mitosis and cell cycle regulation rather than to apoptosis (*22*); this protein probably has been recruited for its antiapoptotic function only in the coelomates. Others, including proteins containing the TIR, caspase, and AP-ATPase domains, possibly interacted to form one or more pathways related to apoptosis and involved in pathogen or stress response. Participation of the MATH domain [present in the COOH-terminal ubiquitin hydrolase and the TNFR-associated factors (TRAFs)], the RING finger (the other domain of the TRAFs), and possibly A20-like proteases (*20*) in the ubiquitin signaling system suggests that these proteins as well as the ubiquitin-containing protein kinase IKK have been recruited for their functions in apoptosis from the ubiquitin-based pathways.

An apoptotic system resembling the core of the extant one appears to have emerged concomitantly with the origin of the metazoans. This event apparently was marked by the rapid divergence of the caspase-paracaspase protease family from a metacaspase-like ancestor, followed by the divergence of classical caspases and paracaspases (*19*). Another key early event in animal evolution was the emergence of the six–α-helical adaptor domain, which was soon followed by its diversification, the earliest split probably being between the DD and CARD domains, which are the only two domains of this class that apparently are present in all animals (Table 1). The direct apoptotic function of these domains in early animals remains to be ascertained, but even if they originally played a different role, the presence of DD and the

previously undetected ZU5 domain (*23*) in the netrin receptors (Unc-5) (*24*) and ankyrins in all animals suggests that these domains were already used in cytoskeleton- and receptor-mediated signaling. Other components of the apoptotic system and related molecules that were probably present in the common ancestor of all animals include the BCL-2 family proteins, certain adaptors such as TRAFs and Tollip, the A20-like protease, the AP-ATPase, and the IRAK and DAP protein kinases. The conservation of all these apoptotic components in animals suggests that a relatively simple, but (in its main features of execution and regulation) complete, molecular machinery for programmed cell death had evolved before the divergence of the major animal lineages.

Relative to the number of orthologs between nematodes and arthropods to the exclusion of vertebrates, vertebrates and arthropods share more orthologs of the apoptotic system components—and, notably, more domain architectures—to the exclusion of nematodes (Fig. 1). The group of apoptosis-related proteins specifically shared by vertebrates and arthropods includes the transcription factors NFAT and NFκB that apparently have evolved from ancestral immunoglobulin (Ig) domain–containing transcription factors, such as OLF-1 (SPT23) or Su(H), and the signaling cascade associated with NFκB. This cascade minimally consists of the Toll-like receptors, adaptors (MYD88 and FADD), and protein kinases including NIK (NFκB-inducing kinase) and two paralogs of IKK (Fig. 1). The presence of TNF but the apparent absence of a TNFR in the common ancestor of insects and vertebrates suggests

**Fig. 2.** Protein complexity plot for apoptotic domains. The "complexity quotient" of a given protein domain was defined as the product of two values: the number of different types of domains with which it co-occurs in proteins, and the average number of domains detected in these proteins (*4*). The complexity quotient is plotted against the total number of proteins that contain the respective domain in the protein set from a given organism. This plot allows a simultaneous assessment of the numerical and architectural contributions to the complexity of a functional system. The data points for the three animals are color-coded as indicated. The average values over all domains for each of the three organisms are also shown. The data points are for the apoptotic domains from Table 1; the points for selected individual domains are labeled (for abbreviations, see Fig. 1).
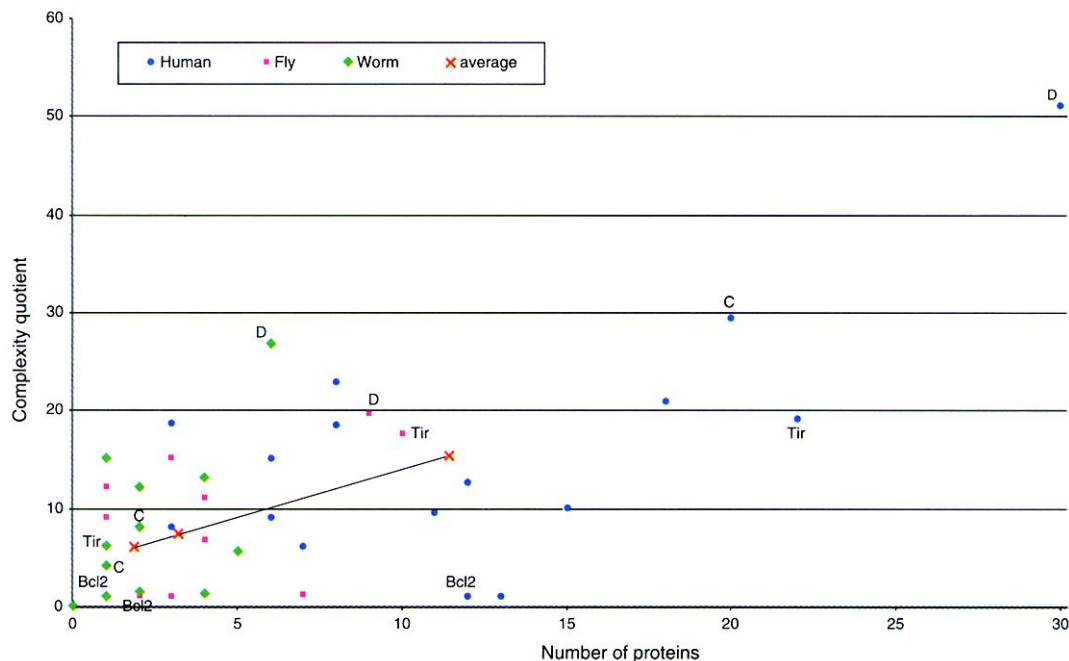
differences in the upstream portion of this apoptotic pathway. Another group of apoptosis-associated proteins that are shared by vertebrates and arthropods to the exclusion of nematodes is the CAD family, whose members regulate "post mortem" DNA degradation (25).

The most straightforward interpretation of these observations, with implications beyond apoptosis, is that the domains and domain architectures present in vertebrates and insects but not in nematodes are indeed shared derived characters (synapomorphies) of the coelomate clade. This is compatible with the traditional view of animal evolution but not with the currently popular ecdysozoa model, which argues for a clade of molting animals including arthropods and nematodes (26). However, the alternative explanation—coordinated loss, in the nematode lineage, of multiple genes coding for proteins involved in several apoptotic pathways—cannot be entirely ruled out (27).

As mentioned above, the prevailing theme in the evolution of the apoptosis-associated domains in the vertebrate lineage is the growth of complexity that is detectable across the entire range of the apoptosis-associated proteins and domains (Table 1, Figs. 1 and 2). In some cases, such as the BCL2 family, this is achieved primarily through duplication with limited diversification; on other occasions, the emergence of new domains (such as the pyrin domain) through a more radical modification of preexisting ones, and reorganization of protein domain architectures (for example, in NACHT NTPases), may be equally important. To a large extent, the innovations in apoptotic and related cytokine signaling in vertebrates could have been linked to the evolution of the vertebrate immune system, with its several new cell types that require highly specialized regulatory pathways.

Perhaps the greatest mystery in the evolution of apoptosis is the presence of homologs of several components of the eukaryotic apoptotic machinery in bacteria. At least two bacterial lineages, Actinomycetes and Cyanobacteria, encode a considerable repertoire of apoptosis-associated domains, including AP-ATPases, metacaspase-like proteases, NACHT NTPases, and TIR domains (Fig. 1). Some of the bacterial AP-ATPases are involved in transcription regulation and signaling (2), whereas the functions of the rest of these proteins remain unclear. However, it is almost certain that they are functionally connected, given the fusions of the metacaspase-like domain and the TIR domain with AP-ATPases (Fig. 1). The presence of the apoptosis-associated domains in the crown-group eukaryotes and in specific divisions of developmentally complex bacteria contrasts with their (thus far) complete absence in archaea and in other bacteria and suggests a history of concerted horizontal gene transfer. The direction of this transfer, however, is uncertain, and although acquisition of the corresponding genes from bacteria by early eukaryotes seems more likely—because the bacterial lineages probably had been fully established by the time of the emergence of the crown-group eukaryotes— the opposite model of a relatively late dissemination from eukaryotes to the bacteria cannot be dismissed.

The principal conclusion from the comparison of the apoptotic system components and their homologs encoded in the sequenced eukaryotic genomes is the major increase in complexity in vertebrates relative to insects and nematodes. This is manifest both in a numerical increase of apoptosis-related proteins (due to gene duplication) and in domain accretion, which leads to increasingly elaborate domain architectures within orthologous protein sets (Fig. 2).

What, if anything, is the unique contribution of the (nearly) complete genome sequences to our understanding of this system? At a qualitative level, most of the observations discussed here and the above conclusions do not depend on such sequences and, in fact, have been considered previously. However, only the genome sequences allow for a reasonably accurate quantitative comparison of the complexity of functional systems, including the apoptotic machinery, in different organisms and for a reasonably confident reconstruction of the ancestral systems. Moreover, the expansion of certain protein and domain families, such as the NACHT NTPases and the pyrin domain in vertebrates, became apparent only from the analysis of the nearly complete sequence of the human genome. And, of course, any statements that a particular protein or domain is lineage-specific—that is, missing in other lineages (for example, the vertebrate-specific pyrin domain)—rely both on the completeness of a representative genome sequence(s) from each of the lineages and on the assumption that they accurately reflect the gene complement of the entire lineage.

With the completion of several eukaryotic genomes, the study of the functional systems of these organisms, including apoptosis, is entering the postgenomic era. However, to understand the origin and evolution of apoptosis at a more satisfactory level, we need more genomes from diverse branches of life. Additional genome sequences of complex bacteria (such as Myxococcus, Cyanobacteria, and Actinomycetes), early-branching eukaryotes, and diverse animals such as primitive chordates will help to piece together the details of various steps in the evolution of cell death.

**References and Notes**

1. M. O. Hengartner, *Nature* **407**, 770 (2000); T. Rich, R. L. Allen, A. H. Wyllie, *Nature* **407**, 777 (2000); P. Meier, A. Finch, G. Evan, *Nature* **407**, 796 (2000); J. Yuan, B. A. Yankner, *Nature* **407**, 802 (2000).
2. L. Aravind, V. M. Dixit, E. V. Koonin, *Trends Biochem. Sci.* **24**, 47 (1999).
3. A preliminary version of the human Integrated Protein Index [5 International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001)] was used for this analysis. The analyzed protein set is available at ftp://ncbi.nlm.nih.gov/pub/koonin/PCD. The *C. elegans* predicted protein set was from the WormPep38 database [C. elegans Sequencing Consortium, *Science* **282**, 2012 (1998); www.sanger.ac.uk/Projects/C_elegans/wormpep]. The *D. melanogaster* predicted protein set was from the Genome Division of the Entrez retrieval system [ftp://ncbi.nlm.nih.gov/genbank/genomes/D_melanogaster/Scaffolds/; M. D. Adams *et al.*, *Science* **287**, 2185 (2000)].
4. For the purpose of this discussion, we define a domain as a distinct portion of protein sequence that shows detectable evolutionary conservation and is, to some extent, evolutionarily independent; that is, it appears in proteins with two or more domain arrangements, and possibly also as a stand-alone protein. Very often, but not always, domains defined in this fashion correspond to experimentally identified structural domains when the latter are known. A domain architecture of a protein is defined as a unique linear combination of domains. For domain detection, domain-specific, multiple alignment–based sequence profiles were constructed and run against the nonredundant protein sequence database (National Center for Biotechnology Information, NIH, Bethesda) or against protein sets from individual genomes using the PSI-BLAST program [S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997); S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998)]. Typically, multiple profiles were generated for a given domain to ensure complete recovery of the respective proteins [L. Aravind, E. V. Koonin, *J. Mol. Biol.* **287**, 1023 (1999)]. The library of profiles used for the detection of apoptosis-associated domains is available at ftp://ncbi.nlm.nih.gov/pub/koonin/PCD.
5. P. C. Rath, B. B. Aggarwal, *J. Clin. Immunol.* **19**, 350 (1999).
6. The CG12919 protein was identified here as the previously undetected TNF ortholog in *Drosophila* by using the TNF domain profile. The *Drosophila* protein CG6531 and *C. elegans* protein T02C5.1 are predicted receptors containing an extracellular cysteine-rich domain homologous to that of TNFR.
7. K. Hofmann, *Cell. Mol. Life Sci.* **55**, 1113 (1999).
8. The pyrin domain was initially identified in the database searches with the ASC1, AIM2, and pyrin protein sequences used as queries in PSI-BLAST–dependent profile searches. Further iterations of the database search showed moderate but statistically significant similarity between pyrin and DD. Secondary structure prediction [B. Rost, C. Sander, *Proteins* **19**, 55 (1994)] and threading [B. Rost, R. Schneider, C. Sander, *J. Mol. Biol.* **270**, 471 (1997)] strongly supported the presence of six α helices in the pyrin domain, indicating that it probably forms a DD-like fold. This domain has been independently described as the "pyrin-like motif," although the fold assignment has not been reported (T. Hlaing *et al.*, *J. Biol. Chem.*, in press).
9. R. J. Lutz, *Biochem. Soc. Trans.* **28**, 51 (2000).
10. A. M. Chinnaiyan, *Neoplasia* **1**, 5 (1999); J. M. McDonnell *et al.*, *Cell* **96**, 625 (1999).
11. A sequence alignment of the Bcl-2 family members was constructed using ClustalW [J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994)] followed by a phylogenetic analysis performed using the neighbor-joining method as implemented in the PHYLIP package [J. Felsenstein, *Methods Enzymol.* **266**, 418 (1996)].
12. D. C. S. Huang, A. Strasser, *Cell* **103**, 839 (2000); S. W. Fesik, *Cell* **103**, 273 (2000).
13. The acronym NACHT comes from the four functionally characterized NTPases that were originally identified as members of this family: NAIP, CIITA, HET-E, and TP1. Members of this NTPase family are most

likely GTPases, as indicated by the activity of CIITA and HET-E [E. V. Koonin, L. Aravind, *Trends Biochem. Sci.* **25**, 223 (2000)].

14. T. L. Beattie, W. Zhou, M. O. Robinson, L. Harrington, *Curr. Biol.* **8**, 177 (1998).
15. E. Diez, Z. Yaraghi, A. MacKenzie, P. Gros, *J. Immunol.* **164**, 1470 (2000).
16. A. M. Verhagen et al., *Cell* **102**, 43 (2000).
17. L. Goyal, K. McCall, J. Agapite, E. Hartwieg, H. Steller, *EMBO J.* **19**, 589 (2000).
18. The eukaryotic crown group is the assemblage of relatively late-diverging, major eukaryotic taxa whose exact order of radiation is difficult to determine with confidence. The crown group includes the multicellular eukaryotes (animals, fungi, and plants) and some unicellular eukaryotic lineages such as slime molds and Acanthamoebae [A. H. Knoll, *Science* **256**, 622 (1992); S. Kumar, A. Rzhetsky, *J. Mol. Evol.* **42**, 183 (1996)].
19. The sister group of the classic animal caspase family of thiol proteases are the paracaspases that thus far have been identified only in animals and *Dictyostelium*; together, these two families constitute the sister group of the metacaspases that have been detected in plants, protists, and bacteria [A. G. Uren et al., *Mol. Cell* **6**, 961 (2000)]. On the basis of conserved structural features, Uren et al. showed that the paracaspases and metacaspases are specifically related to the caspases, to the exclusion of other members of the caspase-gingipain fold [A. Eichinger et al., *EMBO J.* **18**, 5453 (1999)].
20. The A20 protein is a regulator of apoptosis that appears to be involved in the NFκB pathway and interactions with the TRAFs [R. Beyaert, K. Heyninck, S. Van Huffel, *Biochem. Pharmacol.* **60**, 1143 (2000)]. A20 belongs to a distinct family of predicted thiol proteases that is conserved in all crown-group eukaryotes and many viruses. None of the members of this family has a known biochemical function, but they share two conserved motifs with the cysteine proteases of arteriviruses, which led to the prediction of the protease activity. A20 and another protein of this family, cezanne, contain a specialized finger module that is also found in some proteins of the ubiquitin pathway. Together with a fusion of an A20-like protease domain with a ubiquitin hydrolase that has been detected in *C. elegans*, this suggests a functional connection between these predicted proteases and the ubiquitin system [K. S. Makarova, L. Aravind, E. V. Koonin, *Trends Biochem. Sci.* **25**, 50 (2000)]. An additional connection between apoptosis and the ubiquitin system is indicated by the demonstration that, similar to other RING fingers, the one in TRAF6 is an E3-like ubiquitin ligase pathway [L. Deng et al., *Cell* **103**, 351 (2000)].
21. The AP-GTPase is a previously undetected predicted GTPase typified by the COOH-terminal domain of the conserved apoptosis regulator, the DAP protein kinase [B. Inbal et al., *Nature* **390**, 180 (1997)]. This predicted GTPase family appears to be the sister group of the RAS/ARF family GTPases, but differs from them in having a divergent P-loop motif and a THXD instead of the NKXD signature motif. Additional AP-GTPases are found in plants and animals as multidomain proteins that also contain ankyrin, Lrr, and kinase domains. This domain architecture suggests that AP-GTPases participate in GTP-dependent assembly of signaling complexes.
22. A. G. Uren et al., *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10170 (1999).
23. The ZU5 domain is a previously undetected conserved domain that is present in receptors (such as netrin receptors and vertebrate zona pellucida proteins) and cytoskeletal proteins (such as ankyrins) and is predicted to be involved in anchoring receptors to the cytoskeleton.
24. S. L. Ackerman, B. B. Knowles, *Genomics* **52**, 205 (1998).
25. H. Sakahira, M. Enari, S. Nagata, *Nature* **391**, 96 (1998).
26. A. M. Aguinaldo et al., *Nature* **387**, 489 (1997).
27. L. Aravind, H. Watanabe, D. J. Lipman, E. V. Koonin, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11319 (2000).
28. J. R. Brown, W. F. Doolittle, *Microbiol. Mol. Biol. Rev.* **61**, 456 (1997).
29. We thank E. Birney and A. Bateman (The Sanger Center, Hinxton, UK) for kindly providing the preliminary version of the Integrated Protein Index and A. Uren for critical reading of the manuscript and useful comments. The release of the unpublished WormPep data set by The Sanger Center is acknowledged and greatly appreciated.

25 October 2000; accepted 18 January 2001

# Human DNA Repair Genes

Richard D. Wood,[1]* Michael Mitchell,[2] John Sgouros,[2]
Tomas Lindahl[1]

Cellular DNA is subjected to continual attack, both by reactive species inside cells and by environmental agents. Toxic and mutagenic consequences are minimized by distinct pathways of repair, and 130 known human DNA repair genes are described here. Notable features presently include four enzymes that can remove uracil from DNA, seven recombination genes related to RAD51, and many recently discovered DNA polymerases that bypass damage, but only one system to remove the main DNA lesions induced by ultraviolet light. More human DNA repair genes will be found by comparison with model organisms and as common folds in three-dimensional protein structures are determined. Modulation of DNA repair should lead to clinical applications including improvement of radiotherapy and treatment with anticancer drugs and an advanced understanding of the cellular aging process.

The human genome, like other genomes, encodes information to protect its own integrity (*1*). DNA repair enzymes continuously monitor chromosomes to correct damaged nucleotide residues generated by exposure to carcinogens and cytotoxic compounds. The damage is partly a consequence of environmental agents such as ultraviolet (UV) light from the sun, inhaled cigarette smoke, or incompletely defined dietary factors. However, a large proportion of DNA alterations are caused unavoidably by endogenous weak mutagens including water, reactive oxygen species, and metabolites that can act as alkylating agents. Very slow turnover of DNA consequently occurs even in cells that do not proliferate. Genome instability caused by the great variety of DNA-damaging agents would be an overwhelming problem for cells and organisms if it were not for DNA repair.

On the basis of searches of the current draft of the human genome sequence (*2*), we compiled a comprehensive list of DNA repair genes (Table 1). This inventory focuses on genes whose products have been functionally linked to the recognition and repair of damaged DNA as well as those showing strong sequence homology to repair genes in other organisms. Readers desiring further information on specific genes should consult the primary references and links available through the accession numbers. Recent review articles on the evolutionary relationships of DNA repair genes (*3*) and common sequence motifs in DNA repair genes (*4*) may also be helpful.

The functions required for the three distinct forms of excision repair are described separately. These are base excision repair (BER), nucleotide excision repair (NER), and mismatch repair (MMR). Additional sections discuss direct reversal of DNA damage, recombination and rejoining pathways for repair of DNA strand breaks, and DNA polymerases that can bypass DNA damage.

The BER proteins excise and replace damaged DNA bases, mainly those arising from endogenous oxidative and hydrolytic decay of DNA (*1*). DNA glycosylases initiate this process by releasing the modified base. This is followed by cleavage of the sugar-phosphate chain, excision of the abasic residue, and local DNA synthesis and ligation. Cell nuclei and mitochondria contain several related but nonidentical DNA glycosylases obtained through alternative splicing of transcripts. Three different nuclear DNA glycosylases counteract oxidative damage, and a fourth mainly excises alkylated purines. Remarkably, four of the eight identified DNA glycosylases can remove uracil from DNA. Each of them has a specialized function, however. UNG, which is homologous to the *Escherichia coli* Ung enzyme, is associated with DNA replication forks and corrects uracil misincorporated opposite adenine. SMUG1, which is unique to higher eukaryotes, probably removes the uracil that arises in DNA by deamination of cytosine. MBD4 excises uracil and thymine specifical-

[1]Imperial Cancer Research Fund, Clare Hall Laboratories, Blanche Lane, South Mimms, Herts EN6 3LD, UK. [2]Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK.

*Present address: University of Pittsburgh Cancer Institute, S867 Scaife Hall, 3550 Terrace Street, Pittsburgh, PA 15261, USA.