Functional Divergence in the Caspase Gene Family and Altered Functional Constraints: Statistical Analysis and Prediction

Yufeng Wang and Xun Gu

Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 50011

Manuscript received September 1, 2000
Accepted for publication March 29, 2001

ABSTRACT

In this article, we explore the pattern of type I functional divergence (*i.e.*, altered functional constraints or site-specific rate difference) in the caspase gene family that is important for apoptosis (programmed cell death) and cytokine maturation. By taking advantage of substantial experimental data from caspases, the functional/structural basis of our posterior predictions from sequence analysis was extensively studied. Our results are as follows: (1) Phylogenetic analysis shows that the evolution of major caspase-mediated pathways has been facilitated by gene duplications, (2) type I functional divergence (altered functional constraints) is statistically significant between two major subfamilies, CED-3 and ICE, (3) 4 of 21 predicted amino acid residues (for site-specific rate difference between CED-3 and ICE) have been verified by experimental evidence, and (4) we found that some CED-3 caspases may inherit more ancestral functions, whereas other members may employ some recently derived functions. Our approach can be cost effective in functional genomics to make statistically sound predictions from amino acid sequences.

GENE family proliferation provides the raw material for functional innovation in higher eukaryotes. After gene duplication, the classical model (Ohno 1970) suggests that one gene copy maintains the original function, while the other copy is free to accumulate amino acid changes toward functional divergence. Since then, many specific models have been proposed (*e.g.*, Li 1983; Clark 1994; Force *et al.* 1999). However, the details of functional divergence between duplicate genes remain largely unexplored. Gu (1999) developed a method to detect amino acid residues that contribute to functional divergence after gene duplication, which can be considered as candidates for further experimentation. Certainly, its effectiveness for functional genomics needs to be verified by using gene families with substantial biological/structural information.

Apoptosis, or programmed cell death, is an ordered process in which cells commit suicide when they are not needed or are potentially harmful. The key component in the apoptotic machinery is a cascade of *c*ysteine *asp*artyl prote*ase*s (caspases). All caspases, which are initially inactive proenzymes, share the same processing scheme to achieve mature forms after cleavage(s) at specific Asp sites (Kumar 1995; Thornberry and Lazebnik 1998). To date, at least 14 members of the caspase gene family have been identified in mammals, which can be further classified into two major subfamilies, CED-3 (including caspase-2, -3, -6, -7, -8, -9, -10, and -14) and ICE (including caspase-1, -4, -5, -11, -12, and -13; Nicholson and Thornberry 1997). Substantial evidence has shown that the CED-3-type caspases are essential for most apoptotic pathways (Yuan *et al.* 1993; Kuida *et al.* 1996). In contrast, the major function of the ICE-type caspases is to mediate immune response, although some members may play a role in cell death in some circumstances (Yuan and Horvitz 1990; Wang *et al.* 1998). X-ray crystallography has also shown a significant structural difference between these two types of caspases (*e.g.*, Wilson *et al.* 1994; Rotonda *et al.* 1996).

In this article, we take advantage of experimental evidence of caspases to study the functional-structural basis of statistical predictions from Gu's (1999) method. We statistically evaluate the functional divergence between CED-3 and ICE subfamilies and then show that our predictions are consistent with the observations from structural or functional assay. Our analysis shows the potential of evolutionary analysis for functional genomics.

METHODS

The data set: We conducted an exhaustive search (*e.g.*, the gapped BLAST and PSI-BLAST) in several major databases to find all available sequences that are homologous to the *Caenorhabditis elegans* CED-3 gene. After synthetic peptides, expressed sequence tags, partial sequences, and redundant sequences were removed, the final data set includes 42 CED-3 homologous sequences, whose accession numbers are listed in the Figure 3 legend.

Multiple alignment and phylogenetic analysis: The

Corresponding author: Xun Gu, Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, 332 Science II Hall, Iowa State University, Ames, IA 50011. E-mail: xgu@iastate.edu

multiple alignment of 42 caspase amino acid sequences was obtained by the program CLUSTALX (Thompson *et al.* 1997), followed by manual editing according to the structure information (Nicholson and Thornberry 1997). A phylogenetic tree was inferred by the neighbor-joining method (Saitou and Nei 1987) using MEGA2.0 (http://www.megasoftware.net/). PAUP4.0 and PHYLIP were used to examine whether the inferred phylogeny is sensitive to any tree-making method. To evaluate the intensity of functional constraints in each caspase, we calculated the ratio of nonsynonymous to synonymous rates between human/mouse orthologs using Li (1993) and modified Nei and Gojobori (in MEGA2.0) methods.

Type I functional divergence (altered functional constraint) analysis: *Types of amino acid configurations:* Consider a multiple alignment of a gene family with two homologous genes *A* and *B* (Figure 1A). Although different classifications were put forward (*e.g.*, Livingstone and Barton 1996), we adopt the following schemes: (i) Type 0 represents amino acid configurations that are universally conserved through the whole gene family, implying that these residues may be important for the common function shared by all member genes, (ii) type I represents amino acid configurations that are very conserved in gene *A* but highly variable in gene *B*, or vice versa, implying that these residues may have experienced altered functional constraints resulting in site-specific rate difference, (iii) type II represents amino acid configurations that are very conserved in both genes but their biochemical properties are very different, *e.g.*, charge positive *vs.* negative, implying that these residues may be responsible for functional specification in the different subfamilies, and (iv) amino acid configurations at many residues are not so clear-cut that they have to be regarded as unclassified (type U).

Several algorithms were proposed to define these types of amino acid configurations automatically (*e.g.*, Casari *et al.* 1995; Lichtarge *et al.* 1996; Landgraf *et al.* 1999). However, these methods are subject to various problems, *e.g.*, negligence of phylogenetic tree, unclear statistical basis, or arbitrary cutoff for classification. To deal with these problems, a statistical model is needed.

Functional divergence and altered functional constraint: After gene duplication, two duplicates can undergo substantial functional divergence. It seems that only a small portion of residues are involved in functional divergence (Golding and Dean 1998). The trajectories of differentiation can affect the evolutionary pattern of the gene family divergence in several ways. According to Gu (1999), *type I functional divergence refers to the evolutionary process that results in altered functional constraints (or site-specific rate difference) between two duplicate genes, regardless of the underlying evolutionary mechanisms.* Intuitively, type I amino acid configuration is likely observed at a residue with different evolutionary rates between duplicate genes. However, because of the stochastic nature of molecular evolution, each site, no matter whether it is related to functional divergence, has a nonzero probability of becoming any type of amino acid configuration (Figure 1B). Therefore, instead of classifying *ad hoc* type I amino acid configuration, Gu's (1999) method is to compute the (posterior) probability of type I functional divergence for each amino acid site. Type II functional divergence can be defined in the same manner (results not shown).

Statistical modeling for type I functional divergence: It is conceptually convenient to use the ancestral gene (before duplication) as a reference. For each duplicate gene cluster, the evolutionary rate at a site may differ from the ancestral gene, which is called the F_1 site (functional divergence related); otherwise it is called F_0 site (functional divergence unrelated). As shown in Figure 1B, different evolutionary rates between duplicate genes are expected only when a site is F_1 in at least one cluster (*e.g.*, sites 3, 4, and 5), a status denoted by S_1. *The coefficient of type I functional divergence* (θ) between two gene clusters is defined as the probability of a site being status S_1, *i.e.*, $\theta = P(S_1)$. The alternative status is S_0, which means a site being F_0 in both clusters (*i.e.*, the evolutionary rate of each duplicate gene is the same as the ancestral gene, *e.g.*, sites 1 and 2 in Figure 1B). Obviously, $P(S_0) = 1 - \theta$. The null hypothesis is $\theta = 0$, which means that the evolutionary rate is virtually the same between duplicate genes (as well as the ancestral gene) at each site. In this case, the model is reduced to the conventional rate variation among sites (*e.g.*, Gu and Zhang 1997).

Let λ_A and λ_B be the evolutionary rates of a site in clusters A and B, respectively, which vary among sites. For a site being F_0 in both clusters (status S_0) with a probability of $1 - \theta$, we can assume $\lambda_A = \lambda_B$ without loss of generality. However, for a site being S_1 (*i.e.*, being F_1 in at least one cluster) with a probability of θ, we have $\lambda_A \neq \lambda_B$. To avoid too many parameters, Gu (1999) made the following simplification: Under S_1, although $\lambda_A > \lambda_B$ at some sites or vice versa at others, over all sites λ_A and λ_B are statistically independent. Figure 1C outlines the statistical procedure on how to estimate θ from sequences.

Prediction of critical amino acid residues: If $\theta > 0$ significantly, it provides statistical evidence that type I functional divergence (site-specific rate difference) may have occurred after gene duplication. If so, it is of interest to predict which residues are responsible, which can be achieved by posterior analysis (Figure 1C). Let $P(S_1|X)$ be the posterior probability of a site being S_1 when the amino acid configuration (X) is observed. Since the alternative status S_0, with posterior probability $P(S_0|X) = 1 - P(S_1|X)$, means no altered functional constraint, the predicted residues are meaningful only when $P(S_1|X) > 0.5$ such that the posterior odd ratio $R(S_1/S_0) = P(S_1|X)/P(S_0|X) > 1$. A more stringent cutoff may be $P(S_1|X) > 0.67$ or $R(S_1/S_0) > 2$.

Cluster-specific type I functional divergence: func-

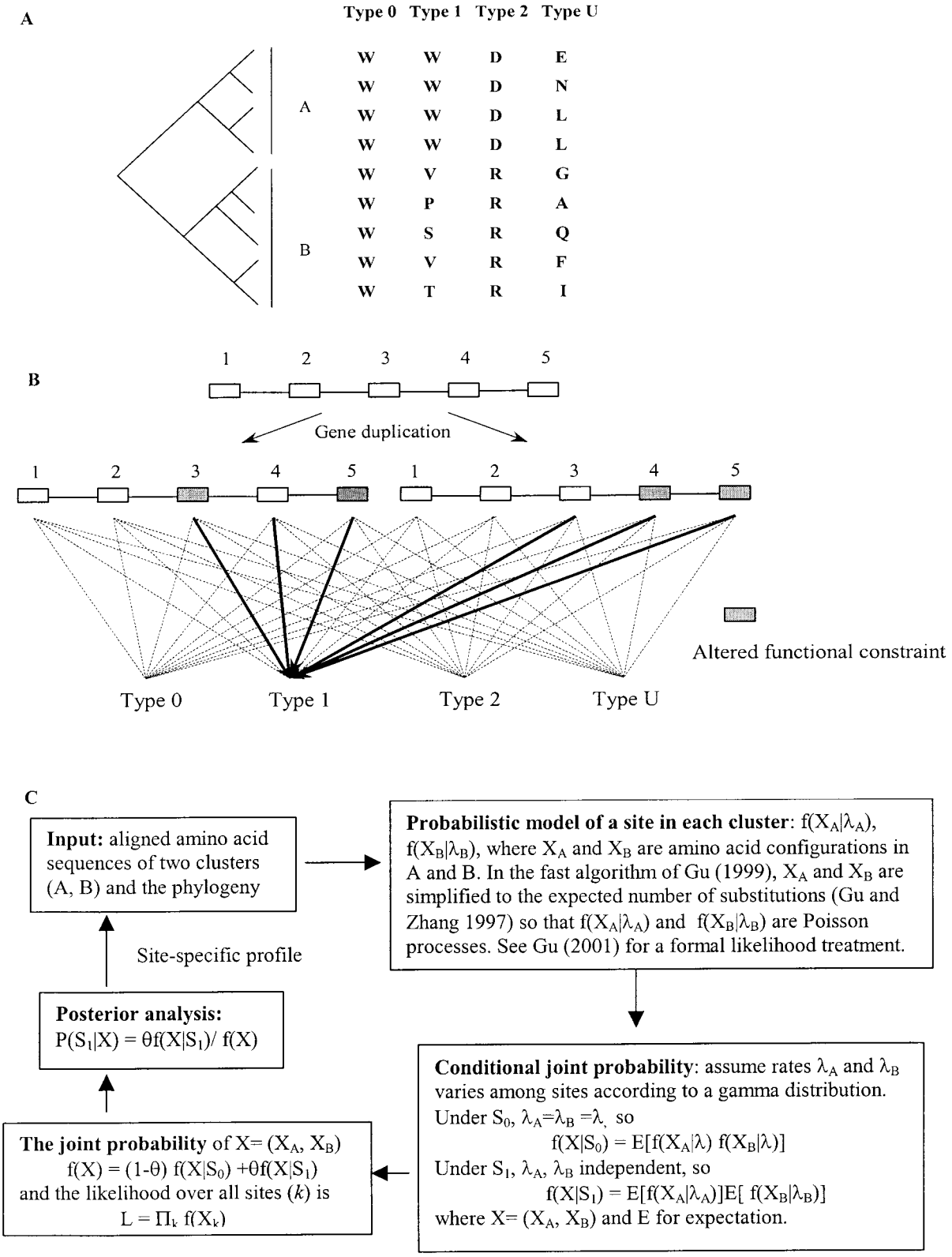


FIGURE 1.—(A) Classification of amino acid configurations for two duplicate gene clusters. Type 0 sites are universally conserved through the whole gene family. Type I sites are very conserved in one cluster but highly variable in the other. Type II sites are very conserved in both clusters but with very different biochemical properties. Type U sites are unclassifiable. (B) A diagram shows the stochastic nature of molecular evolution. Each site (represented as a box) has a nonzero probability for any type of amino acid configuration. At site 1 or 2, no altered functional constraint occurs in either cluster, a status defined as $S_0 = (F_0, F_0)$. At site 3, 4, or 5, altered functional constraint occurs in at least one cluster, a status defined as $S_0 = (F_1, F_0)$ or (F_0, F_1) or (F_1, F_1) (see METHODS for details). (C) A flow chart to illustrate GU's (1999) method.

tional distance analysis: The two-cluster analysis described above cannot tell in which gene cluster the altered functional constraint took place after gene duplication. This problem can be solved by a simple method when at least three homologous gene clusters are available. For any cluster i, let $\theta_i = P_i(F_1)$ be the probability of a site having a different rate from the ancestral gene, and $P_i(F_0) = 1 - \theta_i$ be the probability of having the same rate. Consider two clusters i and j in which the coefficient of type I functional divergence is denoted by $\theta_{ij} = P_{ij}(S_1) = 1 - P_{ij}(S_0)$. If a site being F_1 or F_0 is independent between clusters, we have the relation $P_{ij}(S_0) = P_i(F_0) \times P_j(F_0)$ or $1 - \theta_{ij} = (1 - \theta_i)(1 - \theta_j)$. Therefore, we define *type I functional distance* between clusters i and j as $d_F(i, j) = -\ln(1 - \theta_{ij})$ and *functional branch length* for cluster i or j as $b_F(i) = -\ln(1 - \theta_i)$ and $b_F(j) = -\ln(1 - \theta_j)$, respectively. Obviously, $d_F(i, j)$ is additive, *i.e.*,

$$d_F(i, j) = b_F(i) + b_F(j). \quad (1)$$

When the coefficient of type I functional divergence (θ_{ij}) for each pair of clusters is estimated, the matrix of $d_F(i, j)$ can be computed easily. Then, a standard least-squares method is implemented on the basis of Equation 1 for estimating all b_F's. A large b_F indicates substantial altered functional constraints in this gene cluster, while $b_F = 0$ indicates that the evolutionary rate of each site in this duplicate gene is almost identical to the ancestral gene. In other words, a duplicate gene cluster with $b_F = 0$ may contain a larger component of *ancestral function* compared to other gene clusters.

RESULTS

Evolution of caspase-mediated molecular pathways: *The phylogenetic tree:* The evolutionary tree (Figure 2) of the caspase gene family was inferred by the neighbor-joining (NJ) method (Saitou and Nei 1987). The parsimony (PAUP4.0) and likelihood (PHYLIP) methods give virtually the same topology (data not shown). The presence of caspases in vertebrates, arthropods, and nematodes suggests that the emergence of the caspase gene family might be close to or even earlier than the origin of the animal kingdom.

Although Aravind *et al.* (1999) suggested that caspase may evolve from an ancient protease supergene family, the root of the inferred tree (Figure 2) remains unclear. The evolutionary pattern of caspases can be generally described as follows. On the basin of the tree (see A in Figure 3), there were at least four duplication events that had occurred during a very short time period, resulting in five major lineages: (i) the ICE subfamily, consisting of caspase-1, -4, -5, -13, -11, and -12; (ii) caspase-14; (iii) caspase-2; (iv) caspase-9; and (v) the common ancestor of caspase-8/-10 and caspase-3/-6/-7. In addition, the effector caspases (E-casp-3/-7/-6) and the ancestor of caspase-8 and -10 were generated before the emergence of arthropods. Interestingly, in contrast to the major (ancient) lineages in CED-3-type caspases, ICE-type caspases diversified recently after the divergence of amphibians and mammals, and some of them (*e.g.*, caspase-4 and -5) arose even after the mammalian radiation.

Evolutionary innovations of the caspase-mediated apoptosis pathway by gene duplications: To understand the origin of different caspase-mediated biochemical pathways in apoptosis, we compared the evolutionary relationship of (CED-3-type) caspases with apoptotic pathways (Figure 3). Our major finding is that major evolutionary lineages of caspases may coincide with different caspase-mediated apoptotic pathways triggered by specific death signals. That is, (i) caspase-9 is a key component in the mitochondrial initiated pathway, which is initiated by the intracellular stimuli, upstream Bcl-2, and Apaf-1 proteins (Budihardjo *et al.* 1999); (ii) caspase-2 initiates the apoptosis induced by negative signaling after B cell Ag receptor (BCR) ligation (Chen *et al.* 1999); (iii) apoptoses mediated by caspase-8 and -10 are similar, both initiated by responding to the death receptors (DRs), which contain the death effector domain (DED); and (iv) uniquely, caspase-14 is not processed by any known death stimuli (Van De Craen *et al.* 1998). In summary, since ancient origins, these caspases may evolve through different avenues and provide cells with a potential to initiate apoptosis in response to a variety of intracellular or intercellular stimuli.

Interestingly, although upstream initiator caspases (I-casps, *e.g.*, casp-2, -9, -8/-10) are recruited by different receptors under different physiological or pathological stimuli, they all eventually catalyze the same set of downstream effector caspases (caspase-3, -6, -7), which are the real killers that commit the cell suicide (Figure 3). Our results suggest that (1) gene duplication followed by functional divergence is one major mechanism to generate the complexity of the apoptotic network and (2) such a process is constrained by coordinated regulation. Indeed, in the last step, effector caspases as real killers remain unchanged when more initial death signals are continuously recruited at different levels during the evolution of apoptotic pathways.

Predicting critical residues for type I functional divergence (altered functional constraints) between CED-3 and ICE subfamilies: We estimated that the coefficient of functional divergence between ICE and CED-3 subfamilies is $\theta = 0.29 \pm 0.05$ [the ML option in Gu's (1999) method], implying that the altered functional constraint between them is statistically significant. Further, we use the posterior probability $P(S_1|X)$ to predict critical amino acid residues responsible for type I functional divergence (site-specific rate difference) between CED-3 and ICE subfamilies (Figure 1C). The baseline of the site-specific profile measured by $P(S_1|X)$ is ~0.2–0.3 (Figure 4A). Thirty-two sites (16% of total sites) have $P(S_1|X) > 0.5$. The fact that most sites have scores <50% indicates their similar functional roles between CED-3 and ICE.

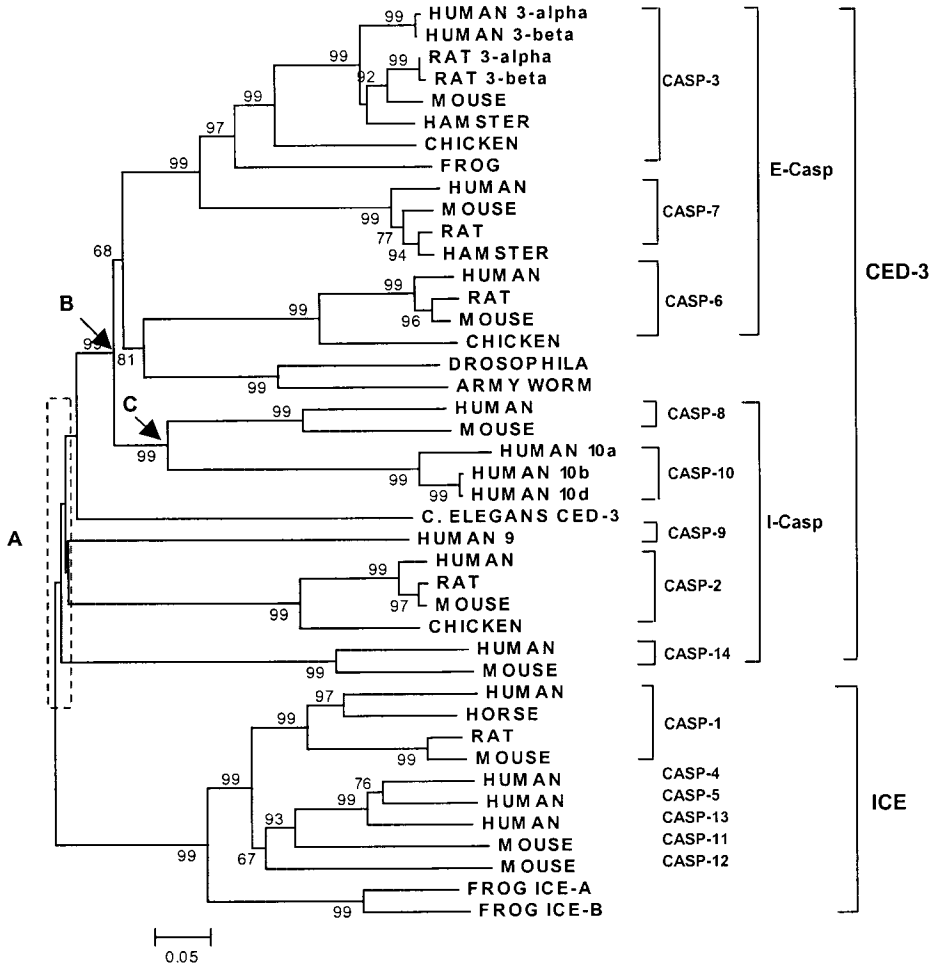


FIGURE 2.—The phylogenetic tree of the caspase gene family, inferred by the neighbor-joining method on the basis of the amino acid sequence with Poisson correction. Bootstrap values >50% are presented. Initiator caspases (I-casps) are involved in upstream regulatory events, and effector caspases (E-casps) directly lead to cell disassembly. The accession numbers for protein sequences are (1) casp-3, U13737 (human 3-α), U13738 (human 3-β), U49930 (rat 3-α), U58656 (rat 3-β), Y13086 (mouse), U27463 (hamster), AF083029 (chicken), D89784 (frog); (2) casp-7, U37448 (human), Y13088 (mouse), AF072124 (rat), U47332 (hamster); (3) casp-6, U20536 (human), AF025670 (rat), Y13087 (mouse), AF082329 (chicken); (4) casp-8, AF102146 (human), AF067841 (mouse); (5) casp-10, U60519 (human 10a), U86214 (human 10/b), AF111345 (human 10/d); (6) casp-9, U60521 (human); (7) casp-2, U13021 (human), U77933 (rat), Y13085 (mouse), U64963 (chicken); (8) casp-14, AF097874 (human), AJ007750 (mouse); (9) casp-1, X65019 (human), AF090119 (horse), L28095 (mouse), U14647 (rat), D89783 (frog ICE-A), D89785 (frog ICE-B); (10) casp-4, Z48810 S78281 (human); (11) casp-5, X94993 (human); (12) casp-13, AF078533 (human); (13) casp-11, Y13089 (mouse); (14) casp-12, Y13090 (mouse); (15) invertebrate caspase, P42573 (*C. elegans* CED-3), Y12261 (*Drosophila melanogaster*), U81510 (armyworm, *Spodoptera frugiperda*).

Although posterior analysis is widely used in bioinformatics, the cutoff value for residue selection is usually empirical. We found that when the first 21 highest-scored residues are removed from the multiple alignment, the estimate of θ is virtually 0. These 21 amino acid residues (among 198 residues) corresponding to the cutoff value $P(S_1|X) > 0.61$ are then chosen for further analysis. Of course, this procedure is meaningful only when $\theta > 0$ significantly.

The functional-structural basis of altered functional constraints: We mapped these 21 predicted sites onto the 3-D structure of caspases. The resolved X-ray crystal structures of human caspase-1 and -3 (WILSON *et al.* 1994; ROTONDA *et al.* 1996) were used to illustrate the structural features of ICE and CED-3 subfamilies, respectively. From the literature, we found experimental evidence for four predicted residues that are involved in the functional-structural divergence between CED-3 and ICE subfamilies (Figure 4B):

1. *Residue 161(348) (In the literature, this site is numbered as W348, according to the protein sequence of human caspase-1) is critical for CED-3 caspase substrate specificity by interacting with a unique surface loop in 3-D structure* [$P(S_1|X) = 0.999$] (ROTONDA *et al.* 1996). At this position, all 22 sequences from the CED-3 subfamily contain an invariant tryptophan (W), whereas a variety of residues are present in the ICE subfamily (Figure 5). Crystal structural analysis reveals that W348 is a key determinant for the caspase-3 (CED-3)-type specificity. First, W348 forms a narrow pocket with the surface loop that is highly conserved in the CED-3 subfamily; see the boxed region in Figure 5. The steric constriction due to this pocket determines the preference of caspase-3 to the substrates with small hydrophilic side chains. Second, W348 along with a group of residues forms a hydrogen bond network, which affects the interaction with the substrate. In contrast, the surface loop shared with

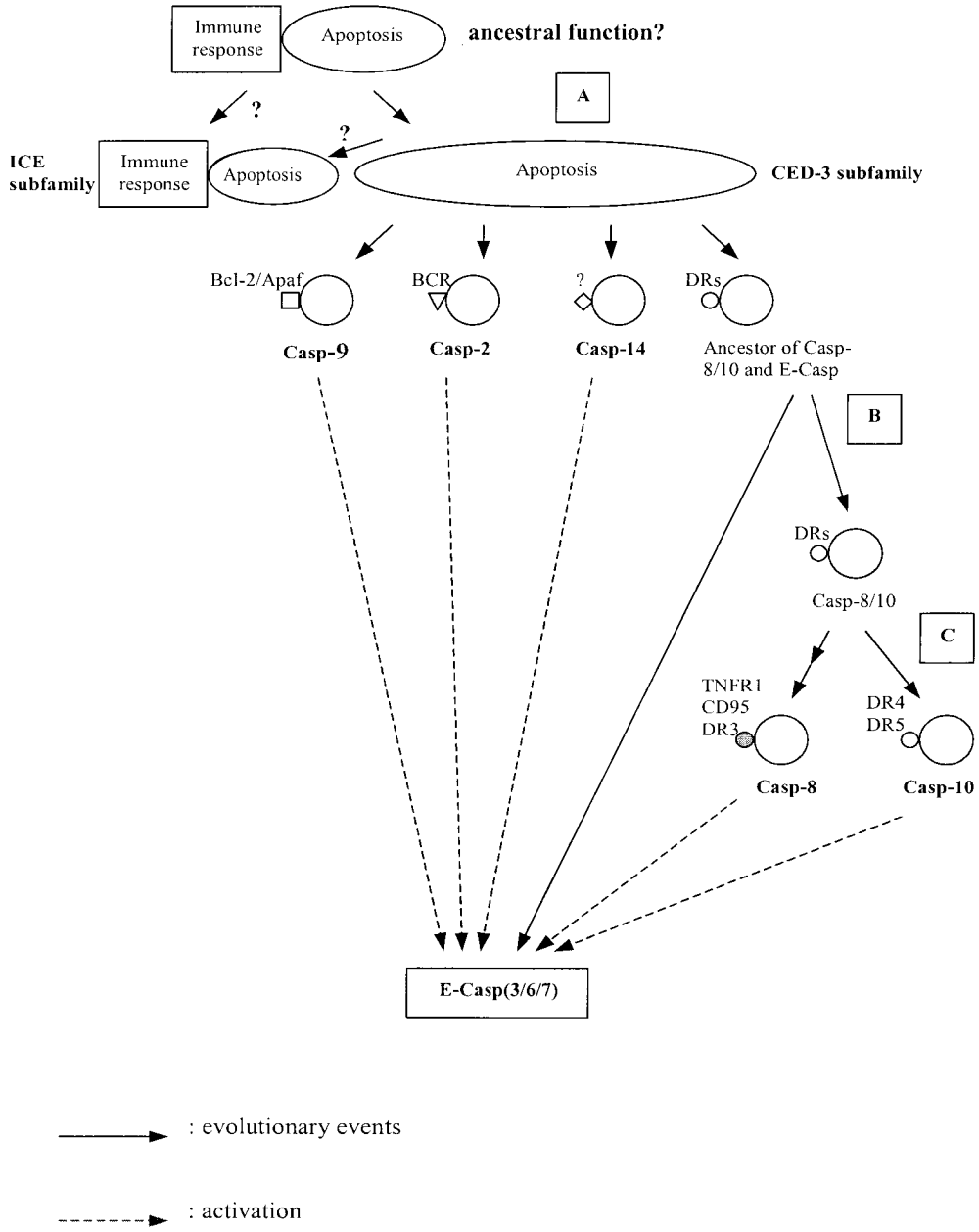


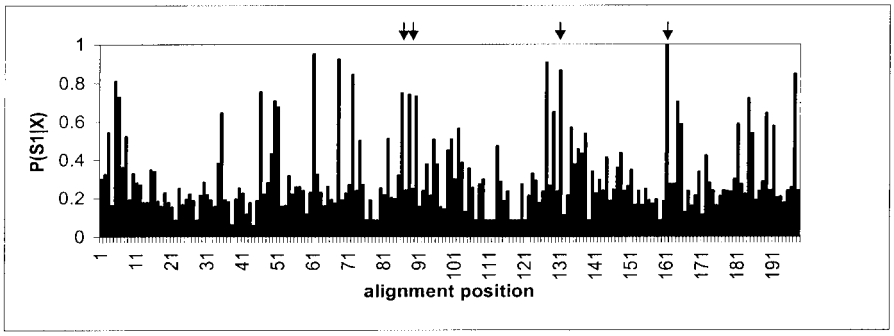
FIGURE 3.—A schematic of evolution of caspase-mediated pathways. Note that the ancestral function of caspases (as well as the origin of ICE-type caspases) is uncertain. A–C correspond to ancestral nodes in Figure 1. Bcl-2/Apaf, BCR, death receptors (DRs), TNFR1, and CD95 are death signals for specific apoptotic pathways. Caspase-3/-6/-7 are effector caspases (E-casps), which are the real killer proteins in programmed cell death.

CED-3 caspases seems to be deleted in all ICE-type caspases, as shown in the boxed region in Figure 5. Hence, the relaxed evolutionary constraint observed at this position in the ICE subfamily is likely to be caused by the 3-D structural difference.

2. *Residues 86* [$P(S_1|X) = 0.75$] *and 88*[$P(S_1|X) = 0.74$] *are responsible for 3-D difference with an unknown functional role.* Indeed, in human caspase-1 (ICE), these two residues appear to lie in a small loop that is not found in the CED-3 subfamily.
3. *Residue 131* [$P(S_1|X) = 0.866$] *is proteolytic site specific to the ICE subfamily.* All caspases are synthesized as inactive proenzymes that need to be processed to the mature forms (NICHOLSON *et al.* 1995). However, distinct cleavage sites within the precursors are found for two subfamilies. D131 is known as a cleavage site in human caspase-1 (ICE type; THORNBERRY *et al.* 1992). All ICE-type caspases preserve an Asp (D) at this position, except for mouse caspase-12 (Asn, E). However, human caspase-3 (CED-3 type) utilizes two other Asn sites for cleavage (ROTONDA *et al.* 1996) so that the functional role of position 131 in CED-3 caspases is no longer important. Therefore, the altered evolutionary constraints at this position can be well explained by the different utilization of cleavage sites for the precursor processing between CED-3 and ICE subfamilies.

Pattern of type I functional divergence among CED-3-type caspases: The CED-3 subfamily consists of a spe-

A



B

site		CED-3	ICE
161	Sequence conservation	An invariant Trp (W)	Highly variable
	Structural features	Form a narrow pocket with an extra loop; form a H-bond	No extra loop; a shallow depression found
	Substrate specificity	Network with a group of amino acids; Hydrophilic side chains	Hydrophobic side chains
86/88	Structural features	No surface loop	Lie in an extra surface loop
131	Sequence conservation	Highly variable	Highly conserved
	Structural features	Not a cleavage site	Cleavage site for proenzyme processing

FIGURE 4.—(A) The site-specific profile for predicting critical amino acid residues responsible for the functional divergence between CED-3 and the ICE subfamilies, measured by the posterior probability of being functionally divergence related at each site [$P(S_1|X)$]. The arrows point to four amino acid residues at which functional divergence between two subfamilies has been verified by experimentation. (B) Four predicted sites that have been verified by experimentation.

cific group of caspases that mediate the programmed cell death in a well-regulated proteolytic cascade and employ related but distinct functions. Here we address an interesting problem, *i.e.*, to infer the trend of altered functional constraint of each cluster.

We study five gene clusters: caspase-3, -7, -6, -8/-10, and -2. Due to insufficient data, caspase-9 was excluded, and caspase-8 and -10 are grouped for their closely related function (FERNANDES-ALNEMRI *et al.* 1996). The upper diagonal of Table 1 shows pairwise coefficients of type I functional divergence (θ) between them; all of them are significantly >0 ($P < 0.05$), with only one exception; *i.e.*, $\theta = 0.006$ between caspase-7 and cluster-8/-10.

To explore the pattern of type I functional divergence in each cluster, we performed functional distance analysis (see METHODS). The pairwise functional distances (d_F) between clusters are shown in the lower diagonal of Table 1. The star-like tree presented in Figure 6 shows the type I functional branch length (b_F) of each cluster, estimated by the least-squares method. The null hypothesis of equal b_F value for each cluster was statistically rejected ($P < 0.05$).

Long functional branch lengths (b_F) of caspase-3, -6, and -2 suggest that these genes may have undergone extensive altered functional constraints as a result of specialized functional roles in apoptosis (Figure 6). Supportive experimental evidence is summarized as follows: (i) The nonredundant functional role of caspase-3 in neurological apoptosis is confirmed by caspase-3 $^{-}/^{-}$ knockout mice (KUIDA *et al.* 1996), (ii) caspase-6 and -3 have different substrate specificity, but both participate in the protease amplification cycle by activating each other, which triggers a series of apoptotic interactions (LAZEBNIK *et al.* 1995; SRINIVASULA *et al.* 1996), and (iii) caspase-2 has its unique dual-role position in positive and negative regulation in apoptosis by differential expression of two alternative splicing isoforms (2_L and -2_S; WANG *et al.* 1994). This dual-role property is also confirmed by knockout mice: Caspase-2 deficiency causes one defective apoptotic pathway (mediated by granzyme B and perforin) but accelerates another pathway (cell death of motor neurons; BERGERON *et al.* 1998).

In contrast, virtually zero b_F values of caspase-7 and -8/-10 indicate that the evolutionary rate of each site in these genes is almost identical to that of the ancestral gene. In this regard, these caspases may inherit a large component of ancestral function during caspase gene family evolution.

For each duplicate gene, the average intensity of functional constraints can be approximately measured by the d_N/d_S ratio between appropriate orthologous sequences (*e.g.*, human-mouse). Interestingly, caspase-3, -6, and -2 (long b_F) have lower d_N/d_S ratios than caspase-7 and

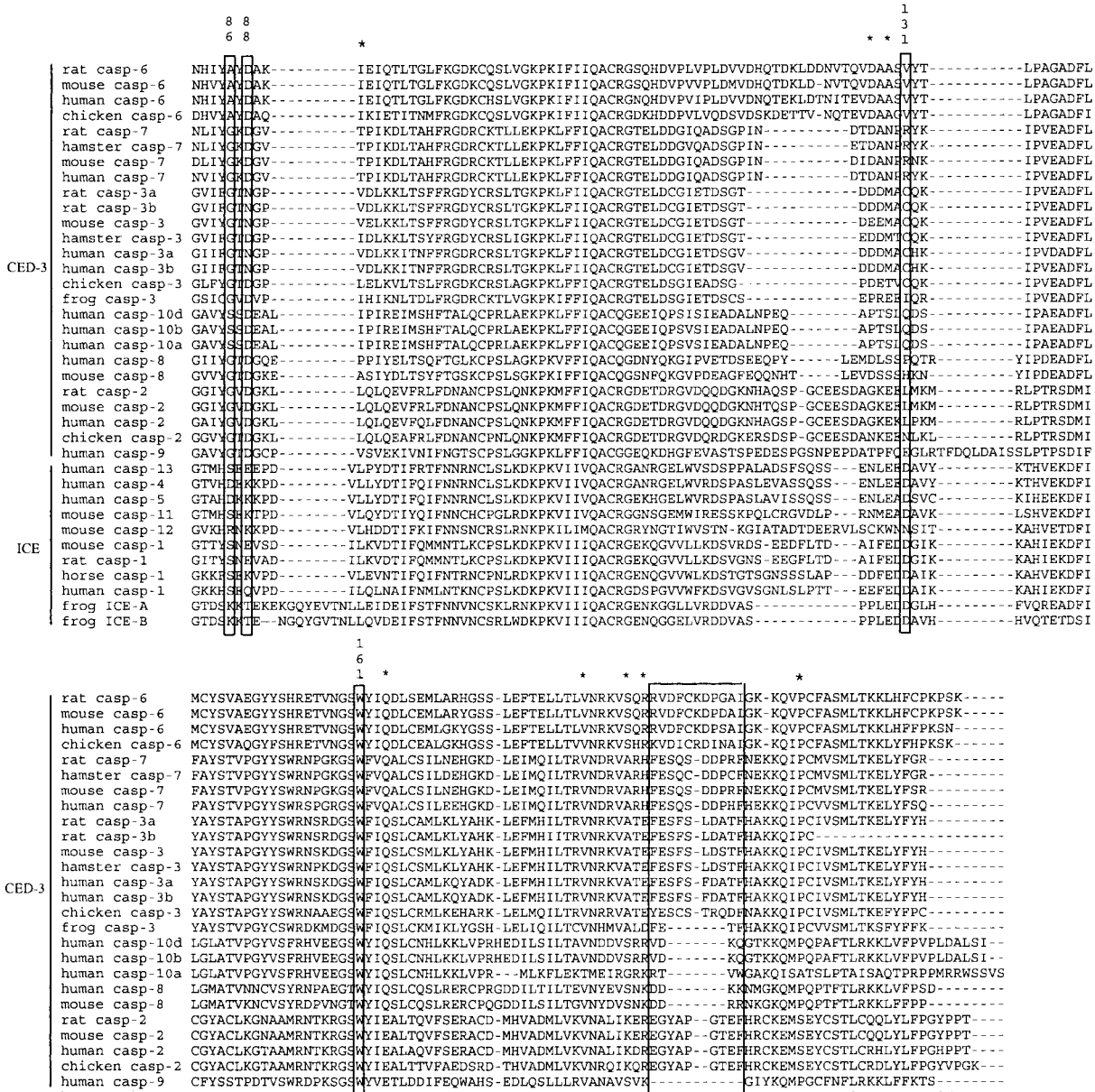


FIGURE 5.—Alignment of predicted regions of caspases. Four predicted sites with experimental evidence are highlighted. The sites with asterisks are predicted residues within this region. The boxed region in the C terminus is the critical region for CED-3 substrate specificity: Most CED-3-type caspases form a surface loop, whereas a shallow depression is found in ICE-type caspases.

-8/-10 (zero b_F), indicating that type I functional divergence in caspases may result in a stronger functional constraint (Figure 6B).

DISCUSSION

The significance of this functional divergence study is twofold: First, we showed that altered functional constraint after gene duplication may play an important role in evolutionary novelties after gene duplication. Second, the site-specific profile based on posterior analysis is useful not only for understanding the functional-structural basis of protein family evolution but also for designing a cost-effective approach in functional genomics, *e.g.*, the strategy for a large-scale mutagenesis.

Predicted sites for type I functional divergence (site-

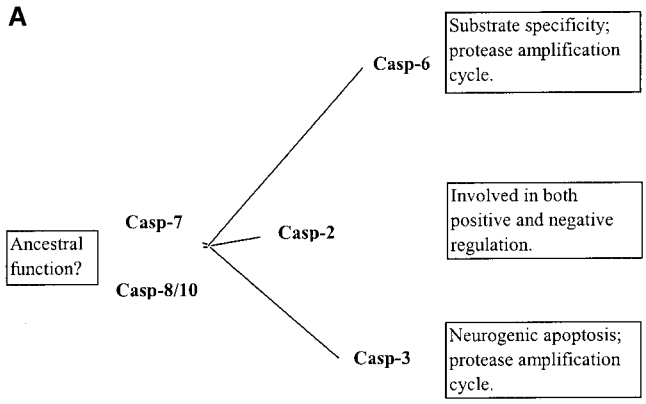
TABLE 1

θ values and d_F values from pairwise comparisons in the CED-3 subfamily

	($\theta_{ij} \pm$ SE)				
$d_F(i, j) \pm$ SE	Caspase-3	Caspase-7	Caspase-6	Caspase-8/-10	Caspase-2
Caspase-3		0.437 ± 0.178	0.844 ± 0.113	0.467 ± 0.113	0.540 ± 0.110
Caspase-7	0.574 ± 0.257		0.579 ± 0.198	0.006 ± 0.022	0.198 ± 0.184
Caspase-6	1.858 ± 0.724	0.865 ± 0.470		0.527 ± 0.190	0.627 ± 0.125
Caspase-8/10	0.629 ± 0.212	0.006 ± 0.022	0.749 ± 0.401		0.306 ± 0.180
Caspase-2	0.777 ± 0.239	0.221 ± 0.229	0.986 ± 0.335	0.365 ± 0.259	

θ_{ij}, the coefficient of the functional divergence between clusters i and j; $d_F(i, j)$, the distance of the functional divergence between clusters i and j; $d_F(i, j) = -\ln(1 - \theta_{ij})$, where i and j are the row and column designation numbers, respectively, and $i \neq j$; SE, standard error.

specific rate difference) without evidence could be either lacking experimental data or due to statistical artifacts (*e.g.*, cutoff value). On the other hand, experimentally verified critical sites that were missed by our analysis may indicate other types of functional divergence (*e.g.*, type II). Clearly, the accuracy of our prediction depends on how strong the association is between functional divergence and site-specific rate difference. To avoid overinterpretation, we should adopt the posterior-based analysis (site-specific profile) in practice only when $\theta > 0$ significantly, and the cutoff value should be weighted by other biological information.



B

cluster	b_F	d_N/d_S
Casp-3	0.694	0.085
Casp-7	-0.030	0.118
Casp-6	0.901	0.071
Casp-8/10	-0.002	0.284
Casp-2	0.198	0.086

FIGURE 6.—(A) A star-like topology of the CED-3 caspases in terms of type I functional branch length b_F. Biological evidence of functional specification for each caspase cluster is shown in the stacked boxes. (B) Functional branch length (b_F) and the ratio of nonsynonymous to synonymous rates (d_N/d_S) for each gene cluster, which were computed by using human-mouse sequences.

Many other methods are available for functional prediction from molecular evolutionary analysis (*e.g.*, see GOLDING and DEAN 1998 as a review; POLLOCK *et al.* 1999; SUZUKI and GOJOBORI 1999; NAYLOR and GERSTEIN 2000; DERMITZAKIS and CLARK 2001; GAUCHER *et al.* 2001; GU 2001). For example, the method of SUZUKI and GOJOBORI (1999) for detecting positive selection on single sites could make an effective prediction when natural selection is the major force for functional diversity. However, its application is unfeasible for many ancient gene families (*e.g.*, WANG and GU 2000, as synonymous distance is saturated). Since all these approaches have their own limitations but complement in some aspects, appropriate combination is strongly recommended.

In many models (*e.g.*, LI 1983; CLARK 1994; FORCE *et al.* 1999), ancestral function of a gene family is conceptual rather than measurable. Functional distance analysis provides a quantitative measure for the altered functional constraints between the ancestral gene and one duplicate gene. This evolutionary measure has raised an interesting hypothesis that caspase-7 and -8/-10 may represent the function of the common ancestor of the CED-3 subfamily since their respective functional branch lengths are virtually zero. We hope this hypothesis can be tested by experimentation.

Similar to any site-specific analysis, our prediction is sensitive to the quality of the multiple alignment. We examined the multiple alignment of the caspase family, particularly in the surrounding regions of the four verified predicted sites (Figure 5). To the best of our knowledge, the alignment can be considered "nearly optimized." For example, the alignment of position 161 (W348) is almost indisputable.

In conclusion, we conducted a case study to show

the capability of predicting type I functional divergence (*i.e.*, altered functional constraints that are site specific) from sequence evolution. Moreover, our analysis showed that a comprehensive approach including various computational methods and multilevel information (from sequence to experimental data) is beneficial for understanding functional diversity of a large gene family in the postgenomics era.

We are grateful to Drs. C.-I Wu and Galvin Naylor for constructive comments, which have improved the manuscript significantly. Thanks go to Jianying Gu for assistance.This study is supported by National Institutes of Health grant RO1 GM62118 to X.G.

LITERATURE CITED

Aravind, L., V. M. Dixit and E. V. Koonin, 1999 The domains of death: evolution of the apoptosis machinery. Trends Biochem. Sci. **24:** 47–53.

Bergeron, L., G. I. Perez, G. MacDonald, L. Shi, Y. Sun *et al.*, 1998 Defects in regulation of apoptosis in caspase-2-deficient mice. Genes Dev. **12:** 1304–1314.

Budihardjo, I., H. Oliver, M. Lutter, X. Luo and X. Wang, 1999 Biochemical pathways of caspase activation during apoptosis. Annu. Rev. Cell. Dev. Biol. **15:** 269–290.

Casari, G., C. Sander and A. Valancia, 1995 A method to predict functional residues in proteins. Nat. Struct. Biol. **2:** 171–178.

Chen, W., H. G. Wang, S. M. Srinivasula, E. S. Alnemri and N. R. Cooper, 1999 B cell apoptosis triggered by antigen receptor ligation proceeds via a novel caspase-dependent pathway. J. Immunol. **163:** 2483–2491.

Clark, A. G., 1994 Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA **91:** 2950–2954.

Dermitzakis, E. T., and A. G. Clark, 2001 Non-neutral diversification after duplication in mammalian developmental genes. Mol. Biol. Evol. **18:** 557–562.

Fernandes-Alnemri, T., R. C. Armstrong, J. Krebs, S. M. Srinivasula, L. Wang *et al.*, 1996 In vitro activation of CPP32 and Mch3 by Mch4, a novel human apoptotic cysteine protease containing two FADD-like domains. Proc. Natl. Acad. Sci. USA **93:** 7464–7469.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531–1545.

Gaucher, E. A., M. M. Miyamoto and S. A. Benner, 2001 Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc. Natl. Acad. Sci. USA **98:** 548–552.

Golding, G. B., and A. M. Dean, 1998 The structural basis of molecular adaptation. Mol. Biol. Evol. **15:** 355–369.

Gu, X., 1999 Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16:** 1664–1674.

Gu, X., 2001 Maximum likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. **18:** 453–464.

Gu, X., and J. Zhang, 1997 A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. **14:** 1106–1113.

Kuida, K., T. S. Zheng, S. Na, C. Kuan, D. Yang *et al.*, 1996 Decreased apoptosis in the brain and premature lethality in CPP32-deficient mice. Nature **384:** 368–372.

Kumar, S., 1995 ICE-like proteases in apoptosis. Trends Biochem. Sci. **20:** 198–202.

Landgraf, R., D. Fischer and D. Eisenberg, 1999 Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng. **12:** 943–951.

Lazebnik, Y. A., A. Takahashi, R. D. Moir, R. D. Goldman, G. G. Poirier *et al.*, 1995 Studies of the lamin proteinase reveal multiple parallel biochemical pathways during apoptotic execution. Proc. Natl. Acad. Sci. USA **92:** 9042–9046.

Li, W.-H., 1983 Evolution of duplicate genes and pseudogenes, pp. 14–37 in *Evolution of Genes and Proteins*, edited by M. Nei and R. K. Keohn. Sinauer Associates, Sunderland, MA.

Li, W.-H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

Lichtarge, O., H. R. Bourne and F. E. Cohen, 1996 An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. **257:** 342–358.

Livingstone, C. D., and G. J. Barton, 1996 Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol. **266:** 497–512.

Naylor, G. J., and M. Gerstein, 2000 Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. J. Mol. Evol. **51:** 223–233.

Nicholson, D. W., and N. A. Thornberry, 1997 Caspases: killer proteases. Trends Biochem. Sci. **22:** 299–306.

Nicholson, D. W., A. Ali, N. A. Thornberry, J. P. Vaillancourt, C. K. Ding *et al.*, 1995 Identification and inhibition of the ICE/CED-3 protease necessary for mammalian apoptosis. Nature **376:** 37–43.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

Pollock, D., W. R. Taylor and N. Goldman, 1999 Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287:** 187–198.

Rotonda, J., D. W. Nicholson, K. M. Fazil, M. Gallant, Y. Gateau *et al.*, 1996 The three-dimensional structure of apopain/CPP32, a key mediator of apoptosis. Nat. Struct. Biol. **7:** 619–625.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Srinivasula, S. M., T. Fernandes-Alnemri, J. Zangarilli, N. Robertson, R. C. Armstrong *et al.*, 1996 The Ced-3/interleukin 1beta converting enzyme-like homolog Mch6 and the lamin-cleaving enzyme Mch2alpha are substrates for the apoptotic mediator CPP32. J. Biol. Chem. **271:** 27099–27106.

Suzuki, Y., and T. Gojobori, 1999 A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16:** 1315–1328.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. J. Higgins, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.

Thornberry, N. A., and Y. Lazebnik, 1998 Caspases: enemies within. Science **281:** 1312–1316.

Thornberry, N. A., H. G. Bull, J. R. Calaycay, K. T. Chapman, A. D. Howard *et al.*, 1992 A novel heterodimeric cysteine protease is required for interleukin-1 beta processing in monocytes. Nature **356:** 768–774.

Van De Craen, M., G. Van Loo, S. Pype, W. Van Criekinge, I. Van Den Brande *et al.*, 1998 Identification of a new caspase homologue: caspase-14. Cell Death Differ. **5:** 838–846.

Wang, Y., and X. Gu, 2000 Evolutionary patterns of gene families generated in the early stage of vertebrates. J. Mol. Evol. **51:** 88–96.

Wang, L., M. Miura, L. Bergeron, H. Zhu and J. Yuan, 1994 Ich-1, an Ice/ced-3-related gene, encodes both positive and negative regulators of programmed cell death. Cell **78:** 739–750.

Wang, S., M. Miura, Y. K. Jung, H. Zhu, E. Li *et al.*, 1998 Murine caspase-11, an ICE-interacting protease, is essential for the activation of ICE. Cell **92:** 501–509.

Wilson, K. P., J. A. Black, J. A. Thomson, E. E. Kim, J. P. Griffith *et al.*, 1994 Structure and mechanism of interleukin-1 beta converting enzyme. Nature **370:** 270–275.

Yuan, J. Y., and H. R. Horvitz, 1990 The Caenorhabditis elegans genes ced-3 and ced-4 act cell autonomously to cause programmed cell death. Dev. Biol. **138:** 33–41.

Yuan, J., S. Shaham, S. Ledoux, H. M. Ellis and H. R. Horvitz, 1993 The C. elegans cell death gene ced-3 encodes a protein similar to mammalian interleukin-1 beta-converting enzyme. Cell **75:** 641–652.

Communicating editor: C.-I Wu