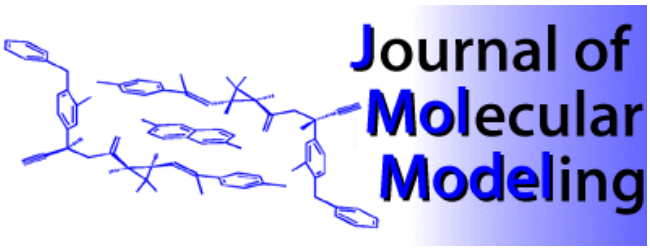
J. Mol. Model. **1998**, *4*, 83 – 93



FULL PAPER

Molecular Modeling of the Domain Shared Between CED-4 and its Mammalian Homologue Apaf-1: A Structural Relationship to the G-proteins

Timothy J. Cardozo and Ruben Abagyan

Biocomputing Laboratory, Skirball Institute of Biomolecular Medicine, NYU Medical Center, New York , NY 10016, USA. E-mail: abagyan@earth.med.nyu.edu

Received: 22 October 1997 / Accepted: 28 January 1998 / Published: 17 February 1998

Abstract Apoptosis (programmed cell death, PCD) is a characteristic type of cell death in which a regulated cellular response pathway mediated by cysteine proteases of the caspase family and Bcl-2 family proteins results in ordered and non-inflammatory involution of the cell. The CED-4 protein and its recently identified mammalian homologue Apaf-1 are critical but functionally uncharacterized components of the cell death machinery. We present here a three-dimensional molecular model for the central domain of CED-4, its alternatively spliced transcript (CED-4l) and Apaf-1. A novel protein family is identified and structure prediction for the family identifies a G-protein-like fold with high reliability. The three-dimensional model provides a potential structural explanation for the alternatively spliced variant as well as the known point mutations in CED-4. Regions of the CED-4 and Apaf-1 sequences which may interact with caspases and the Bcl-2 family are proposed. This new information provides a structural molecular framework for the interaction of CED-4-like proteins with the caspases and the Bcl-2 family in the regulation of apoptosis which is analogous to G-protein mediated interactions in well-defined signal transduction pathways.

Keywords Structure prediction, Apoptosis, Proto-oncogene proteins, G-proteins, Caspase

Abbreviations CED-4l: the alternatively spliced CED-4 transcript which contains an additional 24 residue insertion and is death inhibiting; ICE: Interleukin beta converting enzyme; P-NTPH: P-loop containing nucleotide triphosphate hydrolase; ZEGA: zero end-gap global alignment; PCD: programmed cell death; Apaf-Apoptosis Activating Factor

Introduction

Core molecular components of the cell death machinery – CED-3, CED-4 and CED-9 – were originally identified in the model genetic system of *Caenorhabditis elegans*[1]. CED-3 is one of the pro-apoptotic cysteine proteases[2] (caspases) of which there are several mammalian representatives (ICE, FLICE, etc.), and CED-9 is a member of the apoptosis-inibiting Bcl-2 family of proteins[3] which localizes to mitochondrial, ER and nuclear membranes. A wealth of data has recently been published on Bcl-2 (reviewed in Reed, 1997[4]) including its three-dimensional structure and the structure of its complex with a minimal peptide helix

Correspondence to: R. Abagyan

from Bax, a pro-apoptotic Bcl-2 family member[5]. The Bax peptide/Bcl-2 complex identifies what may be a key binding surface on the Bcl-2 protein. The three dimensional structure of interleukin-1 beta-converting enzyme (ICE), a caspase has also been solved by X-ray crystallography[6].

Since its cloning however, the biochemical role of CED-4 has not been elucidated and, until very recently, no vertebrate homologue for the protein had been identified. In the past year, interest in CED-4 has intensified with the demonstration that it interacts biochemically with both the caspases (effector arm) and the Bcl-2 family members (regulatory arm) and may therefore be the central player in the cell death machinery [7]. Despite this interest however, no functional data has accumulated about CED-4. Part of the reason for this is that the CED-4 amino acid sequence lies on the outer fringes of known sequence space, having no readily noticeable sequence similarity to any other protein or genomic sequence currently known [1,8]. There is some biological information related to the function of CED-4 however, namely two mutation sites that abolish its death-inducing function and an alternatively spliced version of the protein that inserts 24 residues between exons 3 and 4 of the protein's transcript and converts it from a death-inducing to a death-protective protein[9]. The first point mutation site is a putative ATP/GTP-binding P-loop at residues 159-165, and CED-3 proteolytic processing and subsequent apoptosis were shown to be dependent upon the lysine in this P-loop. The second point mutation site is at a structurally and functionally uncharacterized site at residues 250 and 251 of CED-4. Recently, the cross-linking of ATP analogues to the CED-4 P-loop has been reported and interpreted as evidence of ATP binding[10]. At about the same time, a mammalian protein, termed Apaf-1, was identified which shares a central 320 residue domain with CED-4 and was shown to be a required factor, along with dATP, in an *in vitro* apoptosis assay[11]. This protein is now considered to be the first vertebrate homologue of CED-4.

The existing experimental and sequence information on CED-4 and Apaf-1 therefore suggests nucleotide-binding, but which nucleotide and which structural fold? According to the structural classification of proteins (SCOP) database[12], at least two different folds –the P-NTPH fold and the "ATP pyrophosphatases"– contain P-loop motifs which actually bind ATP. Furthermore, a P-loop motif which is not involved in nucelotide binding occurs in other proteins with vastly different topologies, such as the actin-like heat shock-70 cognate ATPase (which binds ATP in a different location) and chymotrypsin (which does not bind nucleotides at all) (PROSITE documentation[13] and personal observations). Indeed, the P-loop containing proteins which *do* bind nucleotides may bind GTP and UTP instead of ATP, an observation which, given the relatively non-standard way in which ATP analogues, but not ATP itself, were used in CED-4 binding studies[10], leaves open the possibility that these nucleotides, or others such as dATP, may be the in vivo ligands of CED-4. (CTP but not GTP or UTP was used as a control in the binding studies[10]). In addition, ADP-containing di-nucleotides such as NAD, NADPH and FMN bind to additional folds such as TIM barrels and Rossman folds using similar atomic determinants as the classic P-loop containing nucleotide binding folds[14]. Which of these nucleotide binding structures and the many functions associated with them might be associated with CED-4/Apaf-1?

We applied a newly developed protein structural fold recognition method to address these questions by identifying the most likely structural fold and building a three-dimensional model of the common CED-4/Apaf-1 domain. The procedure, which has no bias towards the previously established P-loop motif and *in vitro* experimental information, assigns a probability of structural relationship (P-value) to two globally aligned sequences. These P-values are based on the statistical distribution of zero-end gap global alignment (ZEGA) scores derived from exhaustive comparison of all sequences of protein domains with known 3D folds[15]. ZEGA statistics were recently used to evaluate a structural relationship between the breast cancer gene TSG101 and yeast ubiquitin conjugating enzyme[16]. Our results here confirm a structural topology, suggest a new functional framework for CED-4/Apaf-1 and provide a three-dimensional structural explanation for the segment inserted by alternative splicing and for the known functional mutations in CED-4.

Results and discussion

CED-4and Apaf-1are related to plant disease resistance proteins and have a P-loop containing nucleotide tri-phosphate hydrolase(P-NTPH) fold

We ranked the entire NCBI non-redundant database of known protein sequences by their P-value of global structural relationship to the common CED-4/Apaf-1 domain. Several domain sequences emerged from a this search with alignment P-values of less than 0.0001– a highly significant P-value level[15]. These sequences are exclusively plant genes and most of them are demonstrated "disease resistance genes"involved in apoptosis[17] (Table 1). Furthermore, a number of these protein sequences were previously detected by an independent local sequence alignment algorithm[10].

A multiple alignment of this group of sequences identified the consensus residue positions (Figure 1). The only two conserved local patterns in the family correspond to the functionally significant mutation sites in CED-4. The first conserved box corresponds to the putative ATP binding site (P-loop) in CED-4 which abolishes apoptotic activity if mutated[18]. The second is centered on two aspartates in the family which abolished apoptosis when mutated in CED-4 without a reduction in protein stability[7]. The alternatively spliced insertion of 24 residues which converts CED-4 from a death-inducing to a death-inhibiting protein maps into the center of this motif-rich region as well (Figure 1, vertical line).

To provide a structural framework for CED-4/Apaf-1 and to understand the existing mutants and splicing variants, we predicted the structure of CED-4/Apaf-1 and the plant dis-

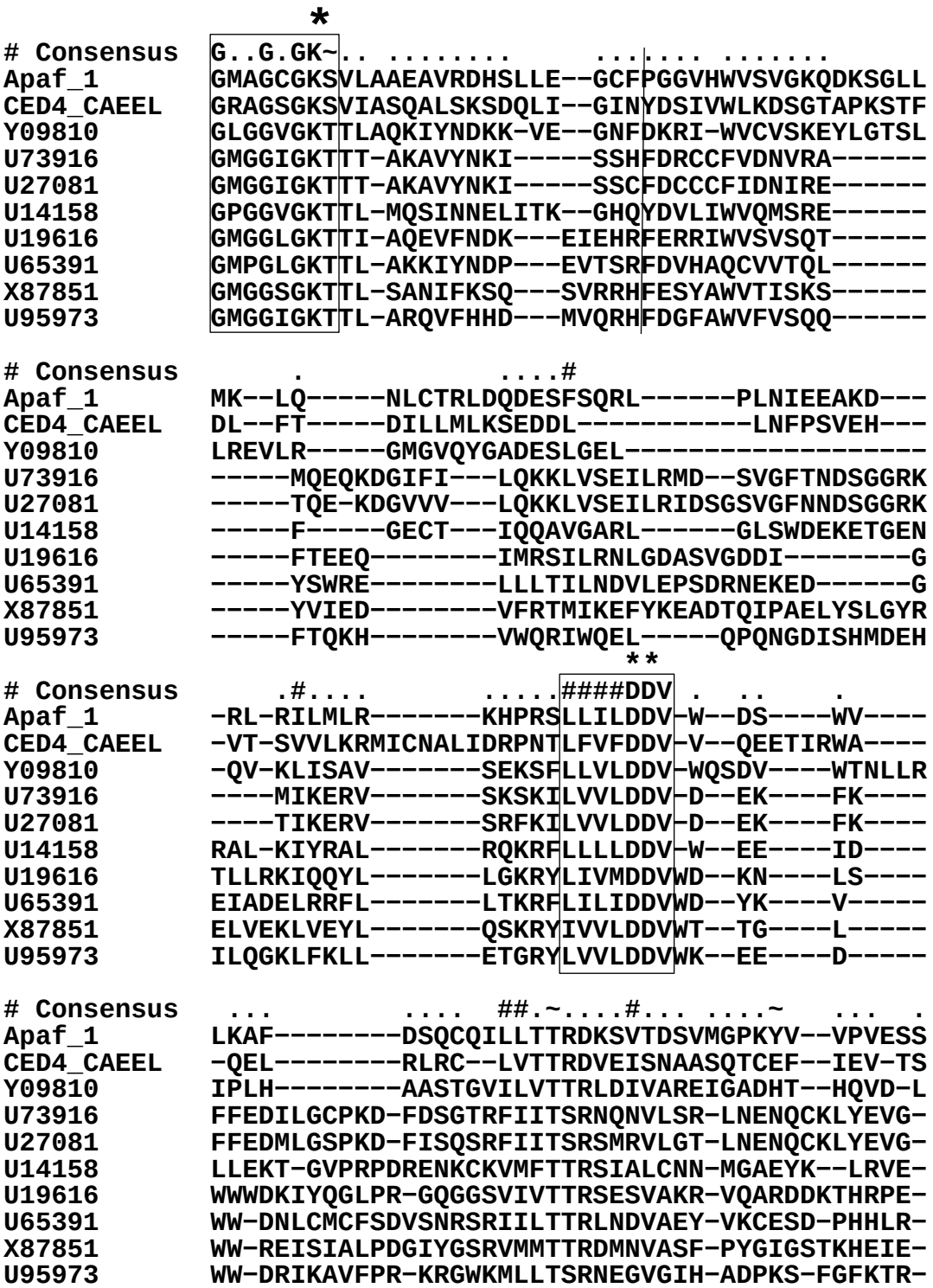


Figure 1 *CED-4 and Apaf-1 form a novel protein family with several plant disease resistance genes. Multiple alignment of CED-4, Apaf-1 and the plant protein sequences identified in Table I. The consensus pattern shows only two conserved "boxes". * marks functionally mutated residues. The first is the known ATP binding site and the second coincides with a functional mutation in CED-4. The red line marks the position of the alternative splicing insertion. Genebank accession numbers are shown in the column at the ends of the sequences. Consensus symbols: [Letter]= amino acid identity; #=hydrophobic; ~=polar; –=negative charge; +=positive charge; ^=small volume residue.*

Table 1 *Genomic sequences ranked by the structural significance of their relationship to CED-4/Apaf-1*

Sequence Description (Organism)	NCBI Identifier (Genebank)	P-value
disease resistance gene (O. sativa)	1773004 (Y09810)	0.0000001
RPS2 disease res. protein (A. thaliana)	625973 (U14158)	0.0000003
myosin HC homologue (A. thaliana)	699495 (U19616)	0.0000006
RPMI disease res. protein (A. thaliana)	1361985 (X87851)	0.000002
Rust resistance protein M (L. usitat.)	1842251 (U73916)	0.00002
PRF (L. esculentum)	1513144 (U65391)	0.00003
L6 (L. usitatissimum)	862905 (U27081)	0.00004

ease resistance proteins by ranking the P-values of their alignments to the sequences of all experimentally solved domain structures in the SCOP and PDB databases[12,19]. All the top ranked structures with highly significant P-values of relationship (<0.00001) to the family belong to the same SCOP fold classification—the parallel beta sheet a/ß fold of the P-loop containing nucleotide triphosphate hydrolases(P-NTPH's). To put this P-value in context, it is approximately the same P-value associated with an alignment between two ~100 residue immunoglobulin domains sharing 40% identical residues[15].

Our P-value measures the structural significance of a particular pairwise global sequence alignment and is not more sensitive to the motif than the rest of the sequence, so the existence of the P-loop, the known nucleotide binding and the high P-value for our top-ranked alignment are essentially convergent pieces of evidence confirming this structural relationship for the CED-4/Apaf-1 domain. Furthermore, a previous structural analysis of members of this fold family demonstrated that in addition to the P-loop, a sequence corresponding to the third beta-alpha turn in the fold is also conserved by its interaction with the nucleotide[20]. This sequence corresponds to the second conserved box of the CED-4/Apaf-1 family of proteins. Interestingly, the only functionally annotated sequence in the initial family (Table 1) which is not a plant disease resistance gene is NCBI ID 699495, a myosin homolog from A. thaliana. The myosin motor domain also has the P-NTPH fold. Thus, despite the distant sequence relationship, this confluence of evidence confirms this common topology between the CED-4/Apaf-1 family and the known P-NTPH structures.

CED-4, Apaf-1 and the related plant proteins are most closely related to the G-proteins

The P-NTPH fold classification contains the protein families shown in Table 2 which all bind and/or hydrolyse ATP, GTP or UTP and have the same core topology. Representative proteins with this fold include Ras, the myosin and kinesin motor domains, the F1-ATP synthase, and guanylate kinase. The topological differences between these families all concern the order and orientation of the edge beta strands 2 and 3 in the fold, a region which corresponds to the so-called "switch" regions in G-proteins. G-proteins utilize GTP to organize a binding surface at this edge of the parallel beta sheet for downstream effectors, whereas the motor domains and the F1-ATP synthase domains utilize the energy from ATP hydrolysis for the translation of whole domains relative to surrounding domains. Which of these might be the mechanism by which CED-4/Apaf-1 uses bound nucleotide? Of all the P-NTPH sequences, the G-protein Ras (PDB code 1q21) has the highest P-value for its alignment with any member of the CED-4/Apaf-1 family (with Apaf-1, Table 2). The known behaviour of CED-4 as a regulatory linker[7] between CED-3 (caspase) and CED-9 (Bcl-2) as well as the presence of a C-terminal WD repeat domain in Apaf-1 circumstantially support our prediction of a G-protein-like function. WD-repeats are involved in protein-protein interactions, occur exclusively in regulatory proteins rather than enzymes, and form a 7-bladed beta-propeller structure in the heterotrimeric G-protein beta subunit which links the components of the complex[11]. Furthermore, CED-4 binding to CED-3 promotes CED-3 autoprocessing (activation)[10,21] – a clear analogy to Ras function in which the downstream kinase is activated by binding to Ras. Conventional thinking envisions this interaction as one in which CED-4 uses nucleotide hydrolysis energy to force CED-3 into an auto-proteolytic conformation[22], but none of the known ATP-binding P-NTPH folds induces an *intra*-domain re-arrangement in a bound effector. All use the energy for domain translations *en masse*, whereas G-proteins induce effector structural responses more akin to what is expected for the CED-4/CED-3 interaction.

dATP was required for the in vitro apoptosis assay used to identify Apaf-1[11] and binding of complex ATP analogues to CED-4 was demonstrated using non-standard methods[10]. Thus, although the relevant *in vivo* nucleotide for CED-4 has not been definitively established experimentally (the strong-

Table 2 *Protein families with the P-loop containing nucleotide triphosphate hydrolase fold ranked by the structural significance of their relationship to CED-4/Apaf-1*

Family Name*	Found Protein (PDB Code)	P-value to CED-4/Apaf-1¥
G Proteins	Ras (1q21)	0.0001
Rec A-like	Central domain, F1 ATPase (1efr)	0.0002
Nucleotide/side kinases	Guanylate/Uridylate kinase (1gky)	0.01
Motor Proteins	kinesin/myosin ATPase domain(1vom)	0.01
6-P-fructo-2/fructose-2,6-bis kinase	bifunctional enzyme of same name	1.
Nitrogenase iron protein	dethiobiotin synthetase	1.

**The SCOP database of protein structure classifications is available at http://scop.mrc-lmb.cam.ac.uk/scop*
¥P-value for best alignment to any member of the GRAD family (see text).

est evidence involves loss of function in the P-loop mutant which would be equally applicable for any of the nucleotides), our results and the current *in vitro* experimental evidence suggest that it may be GTP, dATP or ATP. Since dATP is relatively scarce in the cytoplasm and may in fact be an apoptosis signal itself, the most likely scenario is that CED-4 uses dATP for a G-protein-like binding surface transition, a novel but structurally reasonable mechanism given the close structural relationship between ATP-binding members of this fold family and the G-proteins. We propose that the family of proteins sharing the common structural domain of CED-4 and Apaf-1 are therefore G-protein related apoptosis (GRAD) domains.

A molecular model of the CED-4 GRAD domain

We built a structural model for the CED-4 central domain based on its high P-value global sequence alignment to Ras (PDB code 1q21). In support of the model and the prediction, insertions and deletions in CED-4 relative to this template structure map to surface loops rather than the core of the structure (yellow ribbon in Figure 2a). Furthermore, the alignment shows a close match of the predicted secondary structure by two different methods [23,24] for CED-4 and Apaf-1 with the actual secondary structure of the PDB structure (Figure 2b). Note that, although selected sections of the alignment are highlighted, the whole, global alignment generated the top ranked probability and therefore many other residue comparisons underlie the model, not just the emphasized regions.

In the model, the known mutation sites at residues 159-165 and residues 250-251 in CED-4 map closely together in 3D space on the surface of the protein despite significant separation in sequence (Figure 2 purple ribbon). Both are in the surface pocket of the bound NTP surrounding the expected location of the gamma-phosphate indicating, as expected from the motif analysis, that the probable mechanism by which these mutations destroy CED-4 function, is through inhibition of the binding of NTP's. A mutation at residue 258 although it destroys CED-4 function is associated with reduced protein expression indicating that it operates through a reduction in protein stability rather than through a functionally significant mechanism[8]. This mutation maps to the core of the structure as expected (not shown). The 24 residue segment inserted by alternative splicing maps just C-terminal to the tip of strand 2, also on the protein surface. An insertion at this location would therefore not disrupt the core of the fold, but would alter the local surface of the protein. Interestingly, the insertion is located at the "business end" of the G-protein, the surface formed by the switch I and switch II regions which reorganizes depending on the whether GTP or GDP is bound.

Model for CED-4l

An independent fold recognition search with CED-4l, the alternatively spliced transcript of CED-4, also predicts the P-NTPH fold. However, the highest P-value P-NTPH sequence relative to CED-4l is now guanylate kinase (PDB structure 1gky, P=0.002). Guanylate kinase differs from the G-proteins in the insertion of a small subdomain at the previously noted, functionally important variable edge of the P-NTPH beta sheet. The alignment predicted between CED-4l and guanylate kinase offers an interpretation of the nature of the splicing insertion: it maps to a surface alpha-helix (green ribbon in Figure 3). In a model of CED-4l based on this alignment, the

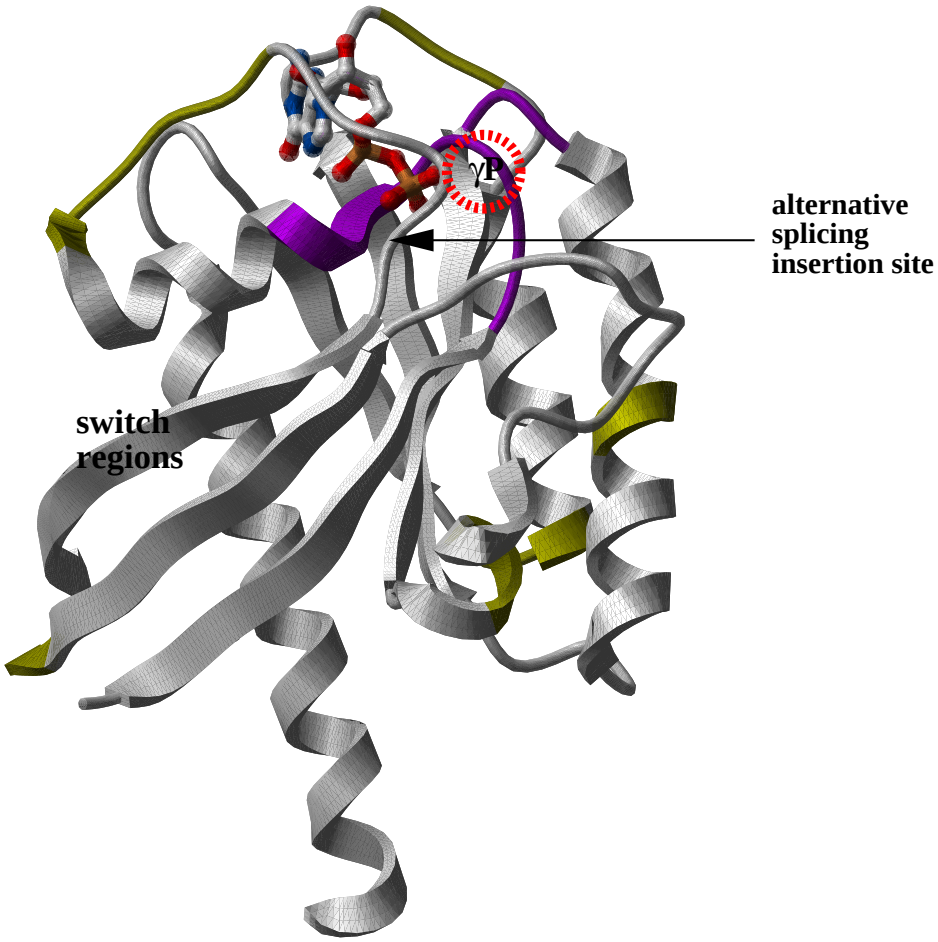


Figure 2a *CED-4/Apaf-1 adopts the 3D fold of the P-loop containing nucleotide triphosphate hydrolases. Views of the model of CED-4/Apaf-1 based on its similarity to Ras (PDB 1q21) looking down onto the nucleotide binding surface(left) and orthogonal to the parallel beta sheet (right). The locations in the model of function-destroying mutations at positions 159-165 and 250-251 are colored magenta. The site to which the alternatively spliced insertion maps is indicated. Insertions and deletions in 1q21 relative to CED-4 are colored yellow. The putative switch regions are indicated. The model includes the gdp molecule solved with 1q21 as a representative of a bound nucleotide. The approximate position of the g-phosphate is extrapolated with a red circle in the left view.*

ends of the segment to which the alternatively spliced sequence maps are very close together in space. The importance of this arrangment is that this segment may be inserted or deleted while leaving the other molecular determinants of nucleotide binding and hydrolysis intact. A localized surface rearrangement, rather than large-scale rearrangement of the protein, is therefore likely to be the structural mechanism by which the opposite apoptotic activity is introduced by alternative splicing. Given the close topological relationship between the G-proteins and guanylate kinase, it is remotely possible that the inserted helix represents a gain of kinase function. Much more likely however is that the insertion, which is directly in the G protein switch regions, alters the binding specificity or a necessary surface rearrangment of CED-4/Apaf-1. The convergence of predictions for CED-4/Apaf-1 and CED-4l to the same fold classification but to different three-dimensional structures which offer a rationale for the functional difference between them is a further validation of the individual three dimensional models.

CED-4/Apaf-1 may operate as G-proteins in well-defined signal transduction pathways

We have established that the GRAD domain has a G-protein-like fold. This relationship along with known data on CED-4 thus suggests that nucleotide binding or exchange in the GRAD domain organizes a binding surface as in the Ras system of signal transduction. In the case of CED-4/Apaf-1, the bound and activated downstream element might be a caspase (e.g. ICE) and/or a Bcl-2 family member whereas in Ras, it is a signal transduction kinase. The nucleotide-dependent binding surface for downstream kinases in G proteins—the so-called switch I and II regions—maps in our model to the region of the alternatively spliced inserted helix at the beta strand 3 edge of the parallel beta sheet. Thus, our model offers a possible explanation both for the functional mutations in CED-4 and for the opposite biological activities of CED-4 and CED-4l: both strongly influence a critical downstream effector binding surface (Figure 4).

Which domains, then, might bind to the downstream effector regions of CED-4/Apaf-1? Candidate domain interac-

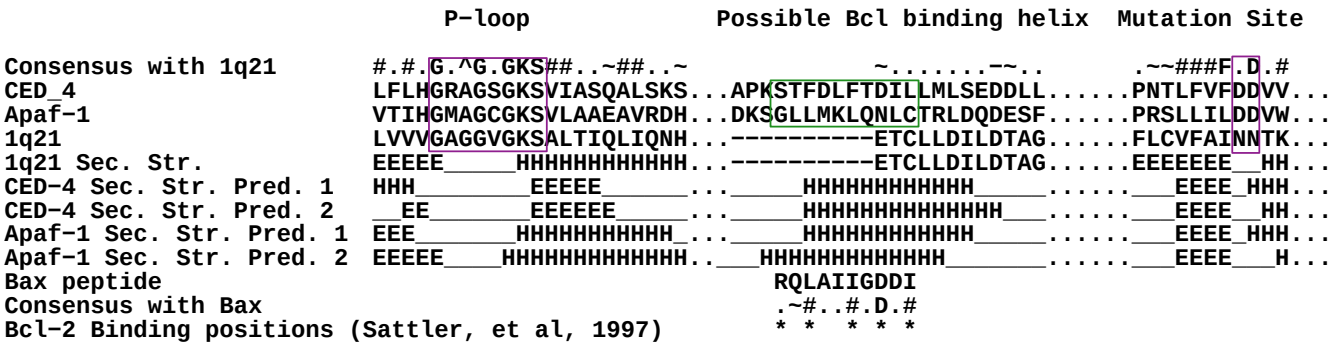


Figure 2b *Selected portions of the alignment of CED-4 sequence to Ras. Boxes correspond to the P-loop at residues 159-165, a potential Bcl-2-binding helix at residues 199-208, and functional mutations at residues 250-251 in CED-4. "1q21 Sec. Str." is the crystallographically determined secondary structure Ras. "CED-4 Sec. Str. Pred. 1" and "CED-4 Sec. Str. Pred. 2" are the secondary structure strings of CED-4 predicted by the Frishman and Argos and PHD methods respectively. "Apaf-1 Sec. Str. Pred. 1" and "Apaf-1 Sec. Str. Pred. 2" are the secondary structure strings of Apaf-1 predicted by the Frishman and Argos and PHD methods respectively. Bax peptide is the sequence of the minimal Bcl-X_L binding peptide in Bax. *'s in the Bcl-2 binding row indicate residues in the Bax peptide which were shown by binding studies to be necessary for full Bcl-2 binding.*

tions which would make sense at this binding surface since their presence is required in the *in vitro* experiments of Wang et. al. [11] include the caspases, the Bcl-2 family, and the flanking N-terminal domain of CED-4/Apaf-1 which is thought to be homologous to the pro-domain of CED-3 (proposed "CARD" domain[25]). Actually, a model for the cooperative binding of both Bcl-2 and the caspases/CARD domains at or near the switch site makes the most sense based on the currently available data. For one, CED-4 may form a complex with both CED-9 and CED-3[4],[26]. For another, in the known structural complex of a G-protein and its downstream kinase, the Rap-Raf complex [27], the binding interaction is an edge to edge association of the single beta pleated sheets in Rap and Raf virtually into one long extended beta sheet. Based on the crystal structure of ICE, the caspase or CARD topology – a single mixed beta sheet a/ß fold – would be structurally compatible with this interaction[6]. The homology between the CARD domains in caspases and CED-4/Apaf-1 does not necessarily mean that CARD domains bind to each other as has been suggested[28]. However, if the intra-chain CARD domain binds to the switch region, the purpose of the interaction may be to orient a surface for the binding of another CARD domain. Thus, the caspases and/or the CARD domains seem the best suited candidate molecules for downstream effector interactions.

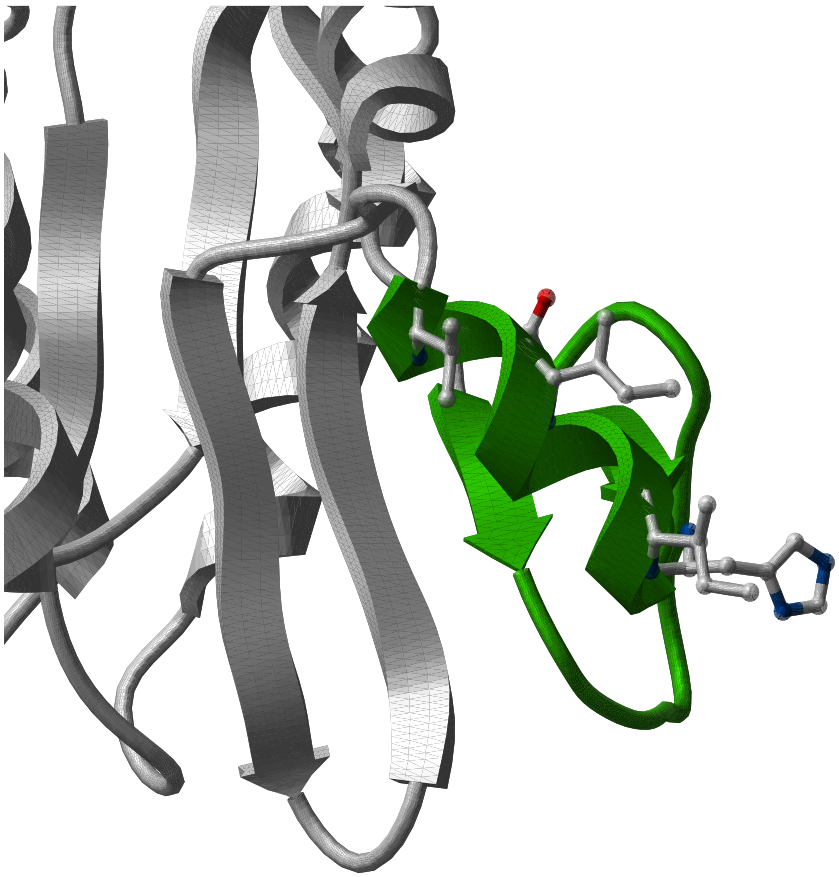
Separately, direct CED-4 binding to CED-9 has already been demonstrated[26], and the C-terminal half of the inserted helix in CED-4l presents a loosely similar pattern to a minimal Bcl-X_L binding Bax helix [5] as does the helix immediately N-terminal to it (Figures 2 and 3). Both these peptide segments are in or near the "switch" region of our model. Binding to Bcl-2/CED-9 may therefore present a conformational restriction nucleotide mediated rearrangement of the switch I and II regions (the bound helix "holds" the protein in a certain conformation).

Thus, an overall two-step apoptosis activation scheme (Figure 4) can be proposed which has Bcl-2 family members binding to CED-4 and preventing the rearrangement necessary for the activation of a bound caspase. Upon release of CED-4 into the cytoplasm, by competition with other Bcl-2 binding proteins such as Bax or other means, NDP may exchange for NTP and the apoptotic proteolytic cascade may proceed. Hydrolysis of the nucleotide, as in the G-protein system, turns off the caspase activation. Based on the available data, there is no way to tell which of the two steps – Bcl release or NTP induced surface reorganization – occurs first or whether they occur cooperatively, however, this view of the sequence of interactions between core cell death components is consistent with experiments showing the localization of CED-4 to Bcl-2 laden membranes in death protected cells, and direct experimental observation of CED-4 binding to CED-9 and CED-3 and promoting auto-processing of CED-3. Furthermore, this arrangement suggests an explanation for the parallel pro-apoptotic activities of Bax and CED-4.

The similarity to the Ras framework also implies the possibility that there exist as yet unidentified GAP-like or GEF-like proteins which regulate the CED-4/caspase interaction by promoting the hydrolysis or exchange of nucleotide (Figure 4). It is tempting to speculate that Apaf-3, an unknown and required mammalian protein isolated with Apaf-1[11], may serve this purpose.

Biological Implications

Apoptosis (programmed cell death, PCD) is a type of cell death resulting from a biochemical cellular program which systematically involutes the cell when activated and cleaves DNA into characteristic fixed-size fragments [29,30]. The discoveries of the roles of key oncogenes such as p53 and



```
                                P-loop       Alternatively Spliced Insert

Consensus with 1gky   S.####~G.^G~GKS       ......~~.D ~~#. .#~.#.S.
CED-4l                SFFLFLHGRAGSGKSVIA...RARVVSDTDDSHSITDFINRVLSR...
1gky                  SRPIVISGPSGTGKS---...PRAGEVNGKD-YNFV-SVDEFKSM...
1gky Sec. Str.        ___EEEE______HH   ...______B___ BEE_ _HHHHHHH...
CED-4l Sec.Str. Pred1                   ..._EEEEE_______HHHHHHHHHH_...
CED-4l Sec.Str. Pred2                   ...EEEEEE_______HHHHHHHHHHH...
Bax peptide                                           RQLAIIGDDI
Consensus with Bax                                    +~#..#.~.#
Bcl Binding positions (Sattler, et al, 1997)          * *  * * *
```

Figure 3 *Model for the inserted segment of CED-4l.* Upper panel*: View of the CED-4l model zoomed on the local area to which the alternatively spliced segment maps. The whole segment is colored green. Residues corresponding to those matching the Bcl-2 binding pattern are displayed in stick representation.* Lower panel*: Selected portions of the alignment to guanylate kinase (PDB code 1gky) based on which the model of CED-4l was built. Row notations are as in Figure 3. The green box indicates a potential Bcl-2 binding helix in CED-4 and corresponds to the helix in the green ribbon in the upper panel*

Bcl-2 in apoptosis, along with studies showing that apoptosis was a critical mechanism of tissue regression and modeling in development, helped to establish the fundamental importance of the cell death machinery in multicellular organisms particularly with respect to disease processes like cancer. A variety of stimuli may invoke the cell death program including free radicals, drugs such as staurosporine, hormone analogs such as dexamethasone, and either the withdrawal (nerve growth factor) or presentation (TNF, Fas-ligand) of cytokines [31]. The cytokine-induction of apoptosis implies a defined signalling pathway the main components of which would be of great interest to biology and medicine.

The three dimensional structural model we have presented here for CED-4 and Apaf-1, which are perhaps the central proteins in the cell death machinery, relies on a quantifiable, and highly reliable relationship to the G-proteins. The model suggests that the apoptotic signalling pathway may be very much like existing signal transduction pathways which utilize G-proteins as key intermediaries, specifically the growth signalling pathways which utilize Ras and the many pathways which utilize heterotrimeric G proteins. The pathways may be regulated in the same way, but a proteolytic cascade proceeding from the critical CED-4 regulatory interaction would be irreversible (a point of no return) as opposed to a kinase cascade and dependent on the rarely exposed dATP, a feature which would particularly make sense given the biological role of apoptosis in the multicellular organism. With these parallel conceptual frameworks our rough structural

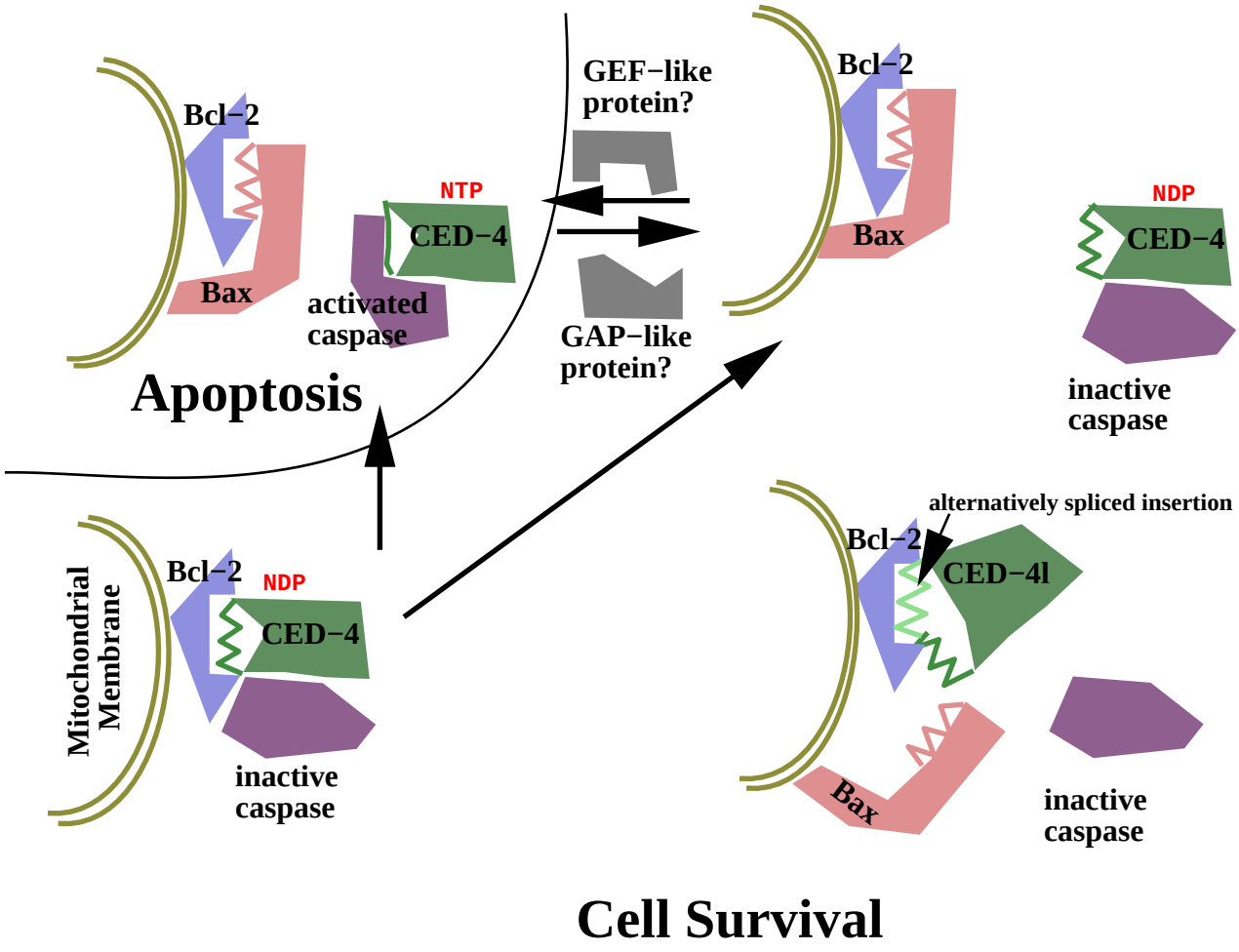


Figure 4 *Schematic of potential molecular interactions between CED-4/Apaf-1 and Bcl-2, Bax and the caspases. CED-4 is depicted for clarity, but the schematic applies to Apaf-1 as well.* Lower left: *CED-4 is initially in a complex with Bcl-2 and CED-3.* Lower right: *Under conditions of CED-4l overexpression (either bound to Bcl-2 as shown or not), the caspase cannot bind to CED-4l and be activated.* Upper left: *Either due to nucleotide tri-phosphate (NTP) binding, or in response to other effectors such as Bax competing for the Bcl-2 binding site followed by NTP binding, CED-4 is released from Bcl-2 due to a conformational change in its key binding surface and activates auto-proteolysis of the caspase. A proteolytic cascade leading to apoptosis results.* Upper right: *the G-protein model suggests the presence of an inactive hydrolized nucleotide (NDP) state in the cytoplasm which may potentially be regulated by GEF and GAP like proteins.*

model of CED-4 in hand, the specific areas of the CED-4 or Apaf-1 protein likely to be involved in binding to downstream effectors and binding to potential regulatory proteins such as GTPase activating proteins (GAP's) and guanine nucleotide exchange factors (GEF's) may be delineated. The model provides a structural and mechanistic interpretation of the previously observed biochemical behaviour of CED-4 as a regulatory linker protein of the (downstream) activities of the cysteine protease caspases and the regulatory Bcl-2 family members. The G-proteins, in their pleiotropic cellular response signaling pathways, do exactly this between receptor tyrosine kinase adaptor molecules (e.g. Grb2, SOS) and downstream kinases or between G-protein coupled receptors and their effectors.

Materials and Methods

Zero end gap global pairwise alignment

A zero end gap global alignment (ZEGA) is the Needleman and Wunsch [32] sequence alignment algorithm with the modification that end-gap penalties are not assigned. As opposed to commonly used local alignment algorithms (BLAST, FASTA) which find only the best matching continuous fragments of two compared sequences, global alignment algorithms find the best fit over the entire lengths of two sequences detecting therefore sequence similarities which are distrib-

uted over the whole sequence or several interrupted clusters of sequence similarity. The Needleman and Wunsch algorithm utilizes weights for residue comparisons provided by an input residue substitution matrix of which there are several choices. Furthermore, in order to align two sequences from end to end, the introduction of gaps into one or the other sequence is often necessary for example to account for different loop lengths between secondary structure elements. The numeric penalty assigned for the introduction or extension of such a gap is offset by improved local alignment downstream of the gap and is a variable parameter for the algorithm. Pairwise alignments in this work used the gonnet substitution matrix [33], a gap opening penalty of 2.4 and gap extension penalty of 0.15.

Multiple sequence alignment

The multiple alignment algorithm used was CLUSTAL [34] as implemented in the ICM program [35] with the same matrix and gap parameters as used for pairwise alignment. While a pairwise global alignment measures the similarity of two sequences, the multiple alignment algorithm is needed to extract the relative relationships of all the pairwise alignments between a set of sequences.

Probabilities and statistics for the structural significance of sequence alignments

The ZEGA P-value measures whether the alignment between two sequences indicates that the two sequences share the same structural fold. The P-value function is derived from the statistics of an exhaustive cross-comparison of all experimentally solved protein domain structures. By globally aligning all the structural domains of a certain length and plotting the distribution of global alignments between unrelated domains one derives a quantitative measure of the odds that a certain global alignment (e.g. a 20% identity alignment) could occur by chance. The quantitative measure has the following form for the distribution of alignment scores (A)[15]:

$$P_A(>t) = 1 - e^{-e^{-\xi y}}, \text{ where } y = \frac{t - m_A(L)}{\sigma_A(L)} \text{ and } \xi = 1.618$$

where P_A is the probability that the alignment score is higher than some score threshold *t*, L is the length of the shorter sequence of the alignment and $m_A(L)$ and $s_A(L)$ are derived coefficients[15]. Several measures of the alignment may be plotted including sequence identity, sequence similarity and alignment score (the sum of the diagonal of the residue comparison matix). However, in the derivation of these P-values it was demonstrated that those from the alignment score were vastly superior to sequence identity in detecting remote structural relationships (i.e. there are several false positive 40% sequence identity alignments in the database even though 40% is routinely considered a "safe" level of sequence identity for structure comparison purposes).

Molecular Models

The models of CED-4, Apaf-1 and CED-4l were built using the homology modeling methods described in Cardozo, et. al. [36] and are currently being deposited in the Protein Data Bank. The modeling procedure utilizes the global alignment resulting from the search to thread the query sequences onto the found structural template (PDB 1q21 Ras for CED-4 and Apaf-1; PDB 1gky Guanylate Kinase for CED-4l) This essentially places the residues of the query sequences on the backbone of the template. Substituted side chains loops containing inserted or deleted residues are then predicted by global optimization using a previously described free energy function and a Monte-Carlo conformational search procedure to sample different side-chain and loop conformations[36].

Acknowledgements We thank Rashmi Hegde and Arturo Zychlinsky for critical reading of the manuscript and Sergei Batalov for technical contributions. This work was supported by a grant from the Department of Energy (DoE grant DE-FG02-96ER62268). DoE's support does not constitute an endorsement by DoE of the views expressed in this article. TJC is supported by the Medical Scientist Training Program at New York University School of Medicine.

References

1. Ellis, H.; Horvitz, H. *Cell* **1986,** *44*, 817.
2. Yuan, J.; Shaham, S.; Ledoux, S.; Ellis, H.M.; Horvitz, H.R. *Cell* **1993,** *75*, 641.
3. Hengartner, M.O.; Horvitz, H.R. *Cell* **1994**, *76*, 665.
4. Reed, J.C. *Nature* **1997,** *387*, 773.
5. Sattler M.; Liang H.; Nettesheim D.; Meadows R. P.; Harlan J. E.; Eberstadt M.; Yoon H. S.; Shuker S. B.; Chang B. S.; Minn A. J.; Thompson C. B.; Fesik S. W. *Science* **1997,** *275*, 983.
6. Walker, N. et al. *Cell* **1994,** *78*, 343.
7. Chinnaiyan, A.; O'Rourke, K.; Lane, B.; Dixit, V. *Science* **1997,** *275*, 1122.
8. Yuan, J.; Horvitz, H. *Development* **1992**, *116*, 309.
9. Shaham, S.; Horvitz, H.R. *Cell* **1996,** *86*, 201.
10. Chinnaiyan, A.; Chaudhary, D.; O'Rourke, K.; Koonin, E.; Dixit, V.M. *Nature* **1997,** *388*, 728.
11. Zou, H.; Henzel, W.J.; Liu, X.; Lutschg, A.; Wang, X. *Cell* **1997**, *90*, 405.
12. Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536.
13. Bairoch, A. *Nucleic Acids Res* **1993**, *20*, 2013.
14. Branden, C.; Tooze, J. *Dehydrogenases*, Birkhauser, Cambridge, MA: 1980.
15. Abagyan, R.; Batalov, S. *J. Mol. Bio.* **1997**, *273*, 355.

16. Koonin, E.V.; Abagyan, R.A. *Nature Genetics* **1997**, *16*, 330.
17. Levine, A.; Pennell, R.I.; Alvarez, M.E.; Palmer, R.; Lamb, C. *Current Biology* **1996,** *6*, 427.
18. Claerwen, J.; Gschmeissner, S.; Fraser, A.; Evan, G.I. *Current Biology* **1997**, *7*, 246.
19. Abola, E.E.; Bernstein, F.C.; Bryant, S.H.; Koetzle, T.F.; Weng, J. In *Data Comission of the International Union of Crystallography;* Allen, F.H.; Bergerhoff and Sievers, R., Eds.: **1987**, p. 107.
20. Walker, J.E.; Saraste, M.; Runswick, M.J.; Gay, N.J. *EMBO J.* **1982,** *1*, 945.
21. Seshagiri, S.; Miller, L.K. *Current Biology* **1997,** *7*, 455.
22. Hengartner, M.O. *Nature* **1997,** *388*, 714.
23. Frishman, D.; Argos, P. *Proteins: Struct. Func. Gen.* **1997,** *27*, 329.
24. Rost, B.; Sander, C.; Schneider, R. *Computer Applications in the Biosciences* **1994,** *10*, 53.
25. Hofmann, K.; Bucher, P.; Tschopp, J. *TIBS* **1997,** *22*, 155.
26. Spector, M.S.; Desnoyers, S.; Hoeppner, D.J.; Hengartner, M.O. *Nature* **1997,** *385*, 653.
27. Nassar, N.; Horn, G.; Herrmann, C.; Scherer, A.; McCormick, F.; Wittinghofer, A. *Nature* **1995,** *375*, 554.
28. Vaux, D.L. *Cell* **1997**, *90*, 389.
29. Kerr, J.F.R.; Wyllie, A.H.; Currie, A.R. *Br. J. Cancer* **1972**, *26*, 239.
30. Wyllie, A.H. *Nature* **1980,** *284*, 555.
31. Jacobson, M.D.; Weil, M.; Raff, M.C. *Cell* **1997,** *88*, 347.
32. Needleman, S.B.; Wunsch, C.D. *J. Mol. Biol.* **1970,** *48*, 443.
33. Gonnet, G.H.; Cohen, M.A.; Benner, S.A. *Science* **1992** *256*, 1433.
34. Thompson, J.D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994,** *22*, 4673.
35. Abagyan, R.A.; Totrov, M.M.; Kuznetsov, D.A. *J. Comp. Chem.* **1994**, *15*, 488.
36. Cardozo, T.; Totrov, M.; Abagyan, R *Proteins: Structure, Function, Genetics* **1995**, *23*, 403.