

Computational discovery and experimental validation of novel Bcl-2 homologs in Zebrafish (*Danio rerio*)

Kiran Mukhyala¹, Erica Kratz², Avi Ashkenazi², Reece K. Hart^{1,3§}

¹Department of Bioinformatics, Genentech Inc., South San Francisco, CA 94080, USA

²Department of Molecular Oncology, Genentech Inc., South San Francisco, CA 94080, USA

³Department of Protein Engineering, Genentech Inc., South San Francisco, CA 94080, USA

Email:

KM: mukhyala.kiran@gene.com

EK: kratz.eric@gene.com

AA: askenazi.avi@gene.com

RKH: hart.reece@gene.com

§Corresponding author

author revision: #712 (2007-03-08 06:52)

Note from Reece: This manuscript was originally submitted in 2005 and was rejected. In 2006, our follow-on experimental paper (Kratz et al, *Cell Death Differ.* **13(10)**:1631-40) subsumed many of the results herein. Although we eventually resubmitted this manuscript after the publishing of Kratz (2006), it was deemed by reviewers to overlap too much with the experimental paper to be accepted. We scooped ourselves. By then, the genes discovered herein and presented in the experimental paper were already in databases. All authors agreed that the paper had become too stale to resubmit, but it remains the only record of how the genes were discovered and provides novel discussion that has not been presented elsewhere. It is offered here for completeness.

Supplemental materials are available at <http://harts.net/reece/pubs/> .

Abstract

Background

The intrinsic apoptotic pathway integrates numerous intracellular stimuli that determine the life-or-death fate of cells. This pathway is a crucial regulator of development and homeostasis in many tissues, and dysregulation of apoptosis is implicated in severe diseases including cancer. A balance of interactions between the pro-survival and pro-apoptotic members of the Bcl-2 family of proteins is the primary regulatory mechanism of the intrinsic apoptotic pathway. As part of an effort to establish zebrafish as a model system with which to study the intrinsic apoptotic pathway, we sought to discover and characterize the full complement of Bcl-2 proteins in zebrafish.

Results

Here we describe the computational discovery of four novel members of the zebrafish Bcl-2 family – orthologs (or co-orthologs) of human Bax, Bik, Bmf and BBC3 – from publicly available sequences using a combination of BLAST, PROSITE patterns, and custom Hidden Markov Models (HMM). We demonstrate that these candidates induce apoptosis in zebrafish embryos as effectively as other established members of the family using both light microscopy and staining of active caspase-3.

Conclusions

The discovery of these genes establishes that zebrafish possess an intrinsic apoptotic pathway that is nearly identical in composition – and presumably function – to that in humans, and therefore provides new tools to study the importance of apoptosis in vertebrate disease and development. We highlight a few important distinctions between the human and zebrafish pathways that may confound the use of zebrafish for genetic screens.

Background

Apoptosis is the preprogrammed mechanism by which cells die in order to regulate growth, adapt to the environment and protect an organism from cellular aberrations. Robust control of apoptosis is required for development and homeostasis, and dysregulation of apoptotic pathways often leads to serious diseases. For example, many cancers evolve somatic changes that selectively inhibit pro-apoptotic pathways and/or enhance the pro-survival pathways in order to evade host responses intended to regulate growth [1][2][3]. Therefore, an understanding of cancer progression and treatment necessitates an understanding of apoptosis.

Two apoptotic pathways have been described in mammals: the extrinsic pathway, typically initiated by the extracellular death-inducing ligands of the TNF family, and the intrinsic pathway, primarily responding to intracellular stimuli and mediated by the Bcl-2 family, but also activated by components of the extrinsic pathway [4][5][6][7]. Bcl-2 proteins – named for the role of its eponymous member in B cell lymphomas [8] – contain one or more of four α -helical protein interaction motifs known as the Bcl-2 homology (BH) domains (BH1, BH2, BH3 and BH4). Bcl2 proteins often share little sequence similarity outside these domains. The Bcl-2 family may be divided into three groups: (i) the pro-survival Bcl-2-like proteins, most of which contain all four BH motifs, (ii) the pro-apoptotic multidomain subfamily members, which typically lack a BH4 motif, and (iii) the pro-apoptotic BH3-only proteins. Some members of the family also have a C-terminal transmembrane region that localizes the proteins to the cytoplasmic face of the outer mitochondrial membrane, nuclear envelope, or endoplasmic reticulum.

Because aberrations in or deletions of Bcl-2 proteins often cause pathologies that prevent the characterization of their functional importance, a genetically tractable and easily maintained model organism is desirable. To the extent that zebrafish can be shown to possess an intrinsic pathway similar to that of mammals, the organism would provide a platform for elucidating the roles of Bcl-2 proteins in apoptosis. The prerequisites for the use of zebrafish as a model for apoptotic signaling are 1) a demonstration

that the major components of the intrinsic pathway are present, and 2) an understanding of the functional correspondence of those components. Several members of the zebrafish Bcl-2 family have been identified by homology to human Bcl-2 genes [9], but more recent efforts using PSI-BLAST and translated BLAST were unable to identify new members of the family among common sequence sources [10][11].

We suspected the existence of novel and divergent Bcl-2 genes in zebrafish that would not be readily discoverable using BLAST or similar strategies. Therefore, we undertook a comprehensive search for these sequences using custom Hidden Markov Models (HMMs) and our publicly available Unison database [12] for feature-based mining. Here we report the discovery of four novel zebrafish Bcl-2 orthologs of human Bax, Bik, Bmf and BBC3, and the experimental demonstration that these candidates induce apoptosis in zebrafish embryos.

Results

Per the Zebrafish Information Network (ZFIN) nomenclature guidelines [13], zebrafish gene names are shown in lowercase italicized type (eg, *bax2*) and proteins names are shown with in non-italic type with an initial capital (eg, Bax2). The sequences described herein are available in supplementary materials and from ZFIN using the gene names *bax2*, *bik*, *bmf2* and *bbc3*.

Construction of BH domain HMMs

We built Hidden Markov Models (HMMs) corresponding to each of the six Bcl-2 family PROSITE entries as described in Methods. In order to determine appropriate HMM score thresholds, we aligned these HMMs to all sequences in UniProtKB/Swiss-Prot release 47 and computed precision and recall with respect to a curated list of known Bcl-2 family members and related sequences. From the precision-recall plots shown in Figure 1, we chose HMM alignment score thresholds as follows: BH1 (pattern), 21; BH2 (pattern), 13; BH3 (pattern), 12; BH4 (pattern), 22; BH4 (matrix), 20; Bcl2 (matrix), 20. These HMMs and thresholds improve recall while slightly reducing precision compared to the PROSITE pat-

terns and matrices from which they were derived. The curated list of Bcl-2 family members, sequence alignments and HMMs are available in supplemental materials.

We used HMM Logo [14] to visualize the information content and contribution of each state of the derived BH3 domain HMM, and to compare these to the original BH3 PROSITE pattern, [LIVAT] . . . L [KARQ] . [IVAL] GD [DESG] [LIMFV] [DENSHQ] [LVSHRQ] [NSR] (Figure 2). The increased sensitivity and decreased specificity appears to derive primarily from a combination of Met and Tyr at position 1, Leu at position 6, Met and Ser at position 8, Ala and Ser at position 9, Glu at position 10, Gln, Lys, and Arg at position 11, and Glu at position 15. As discussed below, only the BH3 HMM was useful in identifying the new candidates presented herein; for this reason, we did not perform similar comparisons of the other PROSITE patterns and matrices with the derivative HMMs.

BLAST Identification of Zebrafish *bax2*

Before undertaking feature-based mining for zebrafish Bcl-2 family members using HMMs, we sought to identify new members with BLAST or PROSITE patterns. We built a BLAST database of the 136655 zebrafish protein sequences described in Methods and queried this database with known human, mouse, and chicken Bcl-2 proteins sequences using default parameters. This search identified Ensembl:ENS-DARP00000040899 as 33% identical to human Bax with an e-value of $1e-14$. Neither this sequence nor any genomically overlapping sequence appeared in any other database at the time. Ensembl annotated this gene as hypothetical. We dubbed this gene *bax2* based on the evidence below.

None of the PROSITE BH patterns aligned to the Bax2 protein sequence, but the Bcl-2, BH1, BH2 and BH3 HMMs aligned with scores/e-values of 33/7.5E-7, 14/0.031, 6.1/110, 2.4/1300. Figure 3 shows an alignment of the BH3 domains from Bax2, other candidates presented later, and known Bcl-2 family members. Despite the poor score of the BH3 HMM alignment, Bax2 shows qualitatively good similarity with the BH3 domains of other sequences and plausible amino acid substitutions at all other positions within the domain. TMHMM [15] predicted that the sequence contains a transmembrane domain, as does human Bax.

We aligned candidate zebrafish Bcl-2 family members against human Bcl-2 sequences, and vice versa, using BLASTP to identify reciprocal BLAST alignments and corresponding pairwise identity and alignment coverage (Table 1). We also assessed the extent of conserved synteny, and therefore the implied orthology, between zebrafish and humans using Ensembl's contigview. Human Bax shares conserved synteny with zebrafish Bax2, which implies that it derives from the same ancestral sequence. Reciprocal best BLAST hits show that human Bax is more similar to the previously known zebrafish Bax than it is to Bax2.

Although the BH2 and BH3 HMM alignments do not score significantly, the alignment of these domains in the context of significant scores for Bcl-2 and BH1, the absence of a predicted BH4 domain, the similarity of the BH3 domain with those of other members, the prediction of a TM domain, the conserved synteny, and the reciprocal BLAST analysis led us to conclude that zebrafish Bax2 is a proapoptotic, multidomain member of the Bcl-2 family and orthologous to human Bax. The poor alignment of BH2 and BH3 HMMs to Bax2 most likely reflects that these HMMs are biased toward the distantly related sequences from which they were constructed, but we did not test this conjecture explicitly.

We were unable to identify additional Bcl-2 family members using BLAST. In addition, PROSITE patterns did not match any of the currently known zebrafish Bcl-2 family members and did not identify any new zebrafish sequences.

Feature-based Mining identification of Zebrafish Bik, Bmf2, and BBC3

We loaded alignments of the custom HMMs to all zebrafish sequences using Unison's automated update facility (see Methods). Once the results were loaded, feature-based mining involved framing an SQL statement [16] that represents a proteomic query of the precomputed sequence features. We composed a query to identify and rank all zebrafish sequences which aligned to BH HMMs using the score criteria defined above, optionally followed by a TMHMM-predicted transmembrane domain, and excluded sequences that overlapped genomically with a known Bcl-2 family member. This query and its

results are available in supplemental materials. We pursued the most promising proteins and dubbed them Bik, Bmf2, and BBC3 based on evidence presented below.

The zebrafish Bik, Bmf2, and BBC3 protein sequences aligned to the BH3 HMM with scores/e-values of 18/0.023, 23/0.00053, 18.6/0.02, and 13/0.77 respectively. These candidates did not align to the BH1, BH2 and BH4 HMMs with E-values less than 10 (the HMMer default), indicating that these proteins might be members of the pro-apoptotic BH3-only subfamily of Bcl-2 proteins. We compared the BH3 regions of these genes with those of previously known human and zebrafish Bcl-2 genes and observed that the conservation exhibited by these alignments do not concur with our expectations based on the conserved synteny presented in Table 1. For example, the BH3 domain of Bik is more similar to that of human Bid despite the orthology to, and BLAST similarity with, human Bik; similar incongruity is seen for other members as well. Human and zebrafish Bik contain transmembrane domains as predicted by TMHMM; TMHMM does not predict transmembrane domains for human or zebrafish Bmf or BBC3 proteins.

Bmf2 appears to be a second teleost ortholog of mammalian Bmf. For reasons presented in the discussion, duplication of some Bcl-2 family members is expected. Although Bmf2 shares conserved synteny with its human counterpart, Bmf is the homolog identified by reciprocal best BLAST hits. Zebrafish BBC3 was remotely similar to human BBC3 (Table 1) and to mouse BBC3 (25% identity, 45% similarity over 66 residues), based on BLAST alignments. It also shares conserved synteny with its human homolog.

Based on the presence of the BH3 domains, the absence of other BH domains, TM domain predictions, sequence similarity, and conserved synteny, we inferred these candidates to be zebrafish orthologs of human Bik, Bmf, and BBC3.

Experimental Validation of Bcl-2 Candidate Genes

To evaluate the ability of the candidate Bcl-2 proteins to activate the intrinsic apoptotic pathway, we injected zebrafish embryos with synthetic RNA encoding zebrafish Bax2, Bik, Bmf2, or BBC3. Ectopic

expression of Bax2, Bik, Bmf2, and BBC3, but not GFP as a negative control, decreased embryonic viability in a dose-dependent manner (Figure 4). Figure 5 shows zebrafish embryos exhibiting hallmarks of apoptosis after injection with minimally lethal doses of Bax2, Bik, Bmf2, or BBC3 RNA. The disintegration of the yolk cell and blastomere and the activation of caspase-3 demonstrates that expression of these genes kills embryos by activating the apoptotic pathway. Activated caspase-3 is a specific, well-defined, and sensitive marker of apoptosis and we prefer this assay to TUNEL staining [17]. These results are consistent with our predictions that these proteins are zebrafish Bcl-2 family members.

Discussion

Orthology, Sequence Similarity and Functional Similarity

Inference of orthology by pairwise sequence comparison can be misleading when homologs are extremely diverged, genes have been lost in one or both clades, or the gene sets are incomplete [18][19][20]. Because we were able to identify regions of conserved synteny for the Bcl-2 genes identified herein, we believe these genes to be legitimate orthologs [21] of their human counterparts – that is, derived from a common ancestral sequence as a result of speciation. The *bax2* and *bmf2* genes appear to be in-paralogs of previously known *bax* and *bmf* genes, respectively, and therefore co-orthologous pairs to the single Bax and Bmf genes in humans. The existence of many-to-one orthologs between genes of the teleost lineage and humans has become apparent for many protein families [22][23] and was already known for two members of the Bcl-2 family, Mcl-1 and Bok [17].

Orthology does not imply conservation of function. Upon gene duplication, such as occurred in teleosts after divergence with the (eventually) human lineage, the Bcl-2 repertoire in both species likely underwent a combination of nonfunctionalization, neofunctionalization and subfunctionalization in which genes lost function, gained new function, or refined an existing function [24]. Shakhnovich and Koonin have recently suggested that paralogous families of essential genes are able to diverge rapidly and evolve by subfunctionalization [25]. Furthermore, the subfunctionalization of in-paralogs – that is, para-

logs that arose as a result of duplication *after* speciation – is a means by which species adapt and evolve [21]. The evolutionary causes of the extreme divergence of human Bik and BBC3 with their zebrafish counterparts are especially intriguing. Duplicate genes and subfunctionalization within the Bcl-2 family may hinder the interpretation of genetic screens in zebrafish. The utility of zebrafish as a model system for studying apoptosis will depend on understanding the functional and phylogenetic nuances of zebrafish and human Bcl-2 family members.

Bax and Bak provide an example of the importance of distinguishing between orthology and function. The Bax2-induced killing of zebrafish embryos can be rescued by co-overexpressing Blp1 but not Blp2, which are functionally similar to human Bcl-xL and Bcl-2 respectively [17]. By analogy with known interactions of hBak with hBcl-xL but not hBcl-2 [26], the rescue experiments are more consistent with Bax2 being functionally similar to human Bak despite being orthologous to human Bax. As mentioned above, we have named genes according to orthology rather than function. We have recently published a preliminary investigation of the function and expression of Bcl-2 genes in zebrafish, including the genes presented here [17].

We also note that zebrafish Bik and BBC3 appear to have diverged from their human counterparts more than the orthologous pairs of other Bcl-2 family members (Table 1). The dissimilarity of the orthologous sequences of each pair is much greater than we would have expected or for typical orthologous pairs of this family. This divergence likely reflects a combination of several factors: neofunctionalization, little sequence conservation pressure outside of the BH domains, BLAST alignment difficulties at low identity (with the default parameters used here), yet undiscovered in-paralogy with other sequences, and possible compensatory changes in other genes.

Zebrafish Bcl-2 Family Completeness

At least 13 of 20 human Bcl-2 proteins were known to have functional homologs in zebrafish prior to the present investigation, to which we have added four orthologs that were not previously known. The

correspondence of known human and zebrafish Bcl-2 proteins is summarized in Figure 6. We are unaware of zebrafish homologs of A1, Bcl-W, Bak, Bcl-xS or HRK.

In this study we identified a fragment of a zebrafish sequence that we believe is an ortholog of human Bim based on reciprocal BLAST hits and tentative conserved synteny with at least one gene. The C-terminal fragment of this gene shows 32% identity with developmentally regulated RNA-binding protein-1 (drbp-1), a mouse and rat RNA binding protein with putative involvement in neural development. The two most likely explanations for the the fragmentary Bim similarity are nonfunctionalization and genome misassembly. We have been unable to confirm either speculation, but we suspect that the abrupt concatenation with a fragment of a protein of unrelated function is more likely to result from genome misassembly. We were unable to clone this sequence but are pursuing this fragment using new zebrafish assemblies as they become available.

Feature-based sequence mining

We – and surely many other groups – had attempted to identify new Bcl-2 family members in zebrafish using commonplace search tools with common data sources. The HMM techniques we used here were not novel and have long been known to be more sensitive than BLAST or PROSITE queries alone. Similarly, the Ensembl *ab initio* (predicted) sequence sources we used were publicly available but, we suspect, not often used for database searching. Thus, we infer that the combination of carefully-constructed sequence alignments and HMM profiles had not been applied to less common sequence sources such as Ensembl *ab initio* sequences. We were surprised by this lesson.

Mining for sequences that “match” diverse feature types is conceptually simple and practically burdensome. Tasks such as eliminating sequence redundancy among many source databases, running and verifying the output of sequence analysis programs, and joining disparate prediction types are often perceived to require more effort than a question deserves. As a result, queries are not answered completely and discovery opportunities are missed. Maintaining currency by repeating this process is tedious.

Our efforts have been greatly facilitated by Unison, a freely available database of essentially all available sequences and diverse predictions on many of those sequences [12]. This infrastructure enables rapid holistic feature-based mining for sequences and the generation of new hypotheses. Unison facilitates these tasks by maintaining a comprehensive, non-redundant set of protein sequences, automating the process of running and loading proteomic predictions, maintaining “run histories” that enable incremental updates with respect to new sequences and new models (eg, Pfam, CDD, or PROSITE), and coordinating the execution and loading of results on a computational cluster. In the present study, nearly 200 000 sequence entries from eight source databases were reduced to 136655 distinct sequences, thus eliminating redundant computation and analysis; HMM, TMHMM, regular expression (PROSITE) other analyses for the distinct sequences were automatically queued to a compute cluster; data integration and exploration was facilitated by SQL queries; and results were easily browsable through the Unison web interface.

Conclusions

We have presented the computational discovery and experimental validation of four novel zebrafish Bcl-2 family members – *bax2*, *bik*, *bmf2* and *bbc3*. We have named these genes according to orthology with their human counterparts using conventions of the Zebrafish Nomenclature Committee [13]. A fifth candidate, putatively an ortholog of human Bim, occurs in a region of the draft zebrafish genome which appears to be misassembled and therefore this candidate could not be cloned or experimentally validated. The identities of these genes are based on sequence homology, conserved synteny and functional similarity. The precise evolutionary and functional relationships of these candidates to their human counterparts is unclear and is currently under investigation. The discovery of these genes expands our understanding of the evolutionary conservation of apoptosis and bolsters the use of zebrafish for elucidating the functions of Bcl-2 family members. We will continue to search for the remaining members of this family.

Methods

Feature-Based Mining Sequence Sources

Feature-based mining was performed in Unison [12]. Briefly, Unison comprises a comprehensive, non-redundant compendium of sequences from many source databases and diverse precomputed proteomic predictions within a relational database. This infrastructure enables rapid, holistic querying of proteins and their precomputed features (“feature-based mining”) by crafting appropriate Structured Query Language statements [16]. At the time of this study, Unison included 6.5M distinct sequences, including 136655 distinct zebrafish sequences from RefSeq as of August 2005 [27], UniProtKB/Swiss-Prot and UniProtKB/TrEMBL release 47 [28], and known, novel and *ab initio* zebrafish sequences from Ensembl Release 35 [29]. The Unison schema, non-proprietary data and predictions, tools, and web interface are released under the Academic Free License and are available for use and download at <http://unison-db.org/>.

Bcl-2 and BH Domain Hidden Markov Models

Multiple sequence alignments were obtained from PROSITE patterns and matrices [30] via the PROSITE web interface. For the patterns PS01080 (BH1), PS01258 (BH2), PS01259 (BH3), and PS01260 (BH4), and for the matrix PS50063 (BH4), the false negative (FN) sequences were manually incorporated into the alignment of true positive (TP) sequences. For the PS50062 (Bcl-2) matrix, manual alignment of the three false negatives was not obvious and only the true positives were used. These six sequence alignments were used to build “global” HMMs using the hmmbuild and hmmcalibrate programs from HMMER v2.3.2 [31] with default arguments. Unison's update framework was used to run and load hmmsearch results for all HMMs aligned to all 136655 zebrafish sequences in Unison. The multiple sequence alignments and resulting HMMs are available as supplementary materials.

Zebrafish Care and Injections

Adult Tubingen Long-fin fish were obtained from the zebrafish International Resource Center. Fish were maintained according to The Zebrafish Book [13]. Synthetic RNA was generated by Ambion mMessage mMachine according to manufacturer's directions, and diluted to the appropriate concentration in 1X Danieus's solution + 0.2% phenol red. 1-2 cell stage embryos were injected with a volume of 4.6 nL via a Nanoliter 2000 (World Precision Instruments) microinjector.

Anti-active Caspase-3 staining

Embryos were fixed in 4% PFA and dehydrated in methanol for a minimum of two hours. After rehydration, embryos were washed with water, permeabilized in acetone for 7 minutes at -20°C , and washed again in water. Embryos were washed several times with PBS+0.5%Tween (PBST), then blocked for two hours at room temperature in 5% fetal bovine serum, 2 mg/mL BSA in PBST. Embryos were incubated in rabbit anti-active Caspase-3 (Pharmingen #559565) overnight at 4°C , washed several times in PBST, and then incubated with secondary antibody, Cy3 goat anti-rabbit (Jackson Immunology #111-166-003), at room temperature for two hours. Both antibodies were diluted 1:500 in blocking solution. Embryos were washed again with PBST before visualization with a Leica MZFL3 fluorescence microscope.

Authors' contributions

KM and RH conceived of the computational search strategy. KM built the hidden Markov models, implemented the search strategy, and analyzed the initial prediction results. EK was responsible for experimental validation of the Bcl-2 genes and for providing experimental data. AA initiated the search for zebrafish orthologs of Bcl-2 family members. RH critically evaluated this study, authored the manuscript, and developed the Unison mining platform. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Andreas Strasser for providing zebrafish Noxa, William Wood for support, and members of the Bioinformatics, Molecular Oncology and Protein Engineering departments for helpful comments. We appreciate the curation and consultation efforts of the Zebrafish Information Network regarding orthology evidence and nomenclature. We are indebted to the authors of the high-quality and freely-available tools and databases that made this work possible, including BioPerl, Ensembl, HMMer, PostgreSQL, PROSITE, UniProt, and numerous Open Source packages.

References

1. Cory S, Adams JM: **The Bcl2 family: regulators of the cellular life-or-death switch.** *Nat Rev Cancer* 2002, **2**:647-656.
2. Kirkin V, Joos S, Zornig M: **The role of Bcl-2 family members in tumorigenesis.** *Biochim Biophys Acta* 2004, **1644**:229-249.
3. LeBlanc H, Lawrence D, Varfolomeev E, Totpal K, Morlan J, Schow P, Fong S, Schwall R, Sinicropi D, Ashkenazi A: **Tumor-cell resistance to death receptor--induced apoptosis through mutational inactivation of the proapoptotic Bcl-2 homolog Bax.** *Nat Med* 2002, **8**:274-281.
4. Andersen MH, Becker JC, Straten P: **Regulators of apoptosis: suitable targets for immune therapy of cancer.** *Nat Rev Drug Discov* 2005, **4**:399-409.
5. Borner C: **The Bcl-2 protein family: sensors and checkpoints for life-or-death decisions.** *Mol Immunol* 2003, **39**:615-647.
6. Strasser A: **The role of BH3-only proteins in the immune system.** *Nat Rev Drug Discov* 2005, **5**:189-200.
7. van Delft MF, Huang DC: **How the Bcl-2 family of proteins interact to regulate apoptosis.** *Cell Res* 2006, **16**:203-213.
8. Tsujimoto Y, Cossman J, Jaffe E, Croce CM: **Involvement of the bcl-2 gene in human follicular lymphoma.** *Science* 1985, **228**:1440-1443.
9. Inohara N, Nunez G: **Genes with homology to mammalian apoptosis regulators identified in zebrafish.** *Cell Death Differ* 2000, **7**:509-510.
10. Aouacheria A, Brunet F, Gouy M: **Phylogenomics of life-or-death switches in multicellular animals: Bcl-2, BH3-Only, and BNip families of apoptotic regulators.** *Mol Biol Evol* 2005, **22**:2395-2416.
11. Coultas L, Huang DC, Adams JM, Strasser A: **Pro-apoptotic BH3-only Bcl-2 family members in vertebrate model organisms suitable for genetic experimentation.** *Cell Death Differ* 2002, **9**:1163-1166.
12. *Unison Database.* <http://unison-db.org/>
13. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, Frazer K, Haendel M, Howe DG, Mani P, Ramachandran S, Schaper K, Segerdell E, Song P, Sprunger B, Taylor S, Van Slyke CE, Westerfield M: **The Zebrafish Information Network: the zebrafish model organism database.** *Nucleic Acids Res* 2006, **34**:D581-5.
14. Schuster-Bockler B, Schultz J, Rahmann S: **HMM Logos for visualization of protein families.** *BMC Bioinformatics* 2004, **5**:7.
15. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**:567-580.
16. Kline K: **SQL in a Nutshell.** ,2001
17. Kratz E, Eimon PM, Mukhyala K, Stern H, Zha J, Strasser A, Hart R, Ashkenazi A: **Functional characterization of the Bcl-2 gene family in the zebrafish.** *Cell Death Differ* 2006, **13**:1631-1640.
18. Dehal PS, Boore JL: **A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database.** *BMC Bioinformatics* 2006, **7**:201.

19. Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, Brinkman FS: **Improving the specificity of high-throughput ortholog prediction.** *BMC Bioinformatics* 2006, **7**:270.
20. Goodstadt L, Ponting CP: **Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human.** *PLoS Comput Biol* 2006, **2**:e133.
21. Sonnhammer EL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends Genet* 2002, **18**:619-620.
22. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH: **Zebrafish hox clusters and vertebrate genome evolution.** *Science* 1998, **282**:1711-1714.
23. Aparicio S: **Exploding vertebrate genomes.** *Nat Genet* 1998, **18**:301-303.
24. Lynch M, Katju V: **The altered evolutionary trajectories of gene duplicates.** *Trends Genet* 2004, **20**:544-549.
25. Shakhnovich BE, Koonin EV: **Origins and impact of constraints in evolution of gene families.** *Genome Res* 2006, **16**:1529-1536.
26. Willis SN, Chen L, Dewson G, Wei A, Naik E, Fletcher JI, Adams JM, Huang DC: **Proapoptotic Bak is sequestered by Mcl-1 and Bcl-xL, but not Bcl-2, until displaced by BH3-only proteins.** *Genes Dev* 2005, **19**:1294-1305.
27. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2007, **35**:D5-12.
28. UniProt Consortium: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007, **35**:D193-7.
29. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35**:D610-7.
30. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database.** *Nucleic Acids Res* 2006, **34**:D227-30.
31. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
32. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor.** *Bioinformatics* 2004, **20**:426-427.
33. *Zebrafish Information Network.* <http://www.zfin.org/>
34. Nadeau JH, Sankoff D: **Counting on comparative maps.** *Trends Genet* 1998, **14**:495-501.

Figures

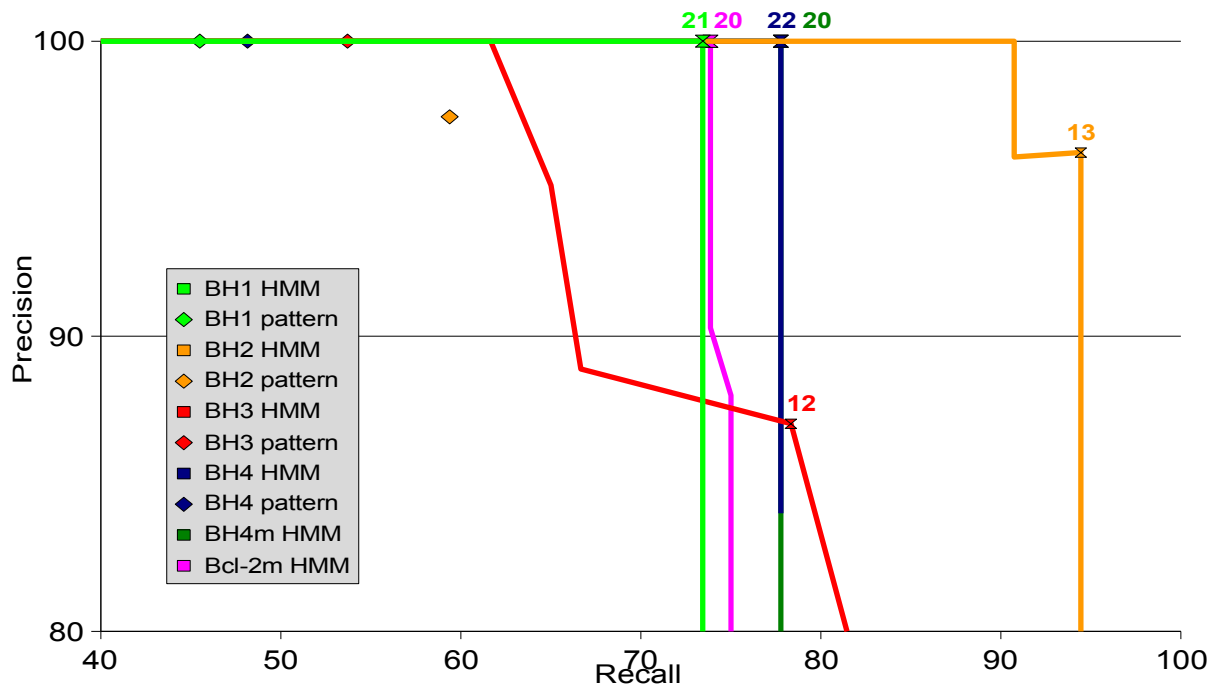


Figure 1. Precision-Recall plots for the Hidden Markov Models used in this study.

Precision and recall were measured using the UniProtKB/Swiss-Prot database and a curated subset of Bcl-2 family members therein. HMM precision and recall at integral score thresholds are denoted by lines. BH4m HMM and Bcl-2m HMM are HMMs built from alignments on PROSITE matrices; the remaining HMMs were built from alignments on PROSITE patterns (see Methods). Hourglass glyphs and numbers indicate HMM score thresholds selected for this study. Diamonds denote the precision and recall of the PROSITE patterns from which the HMMs were derived.

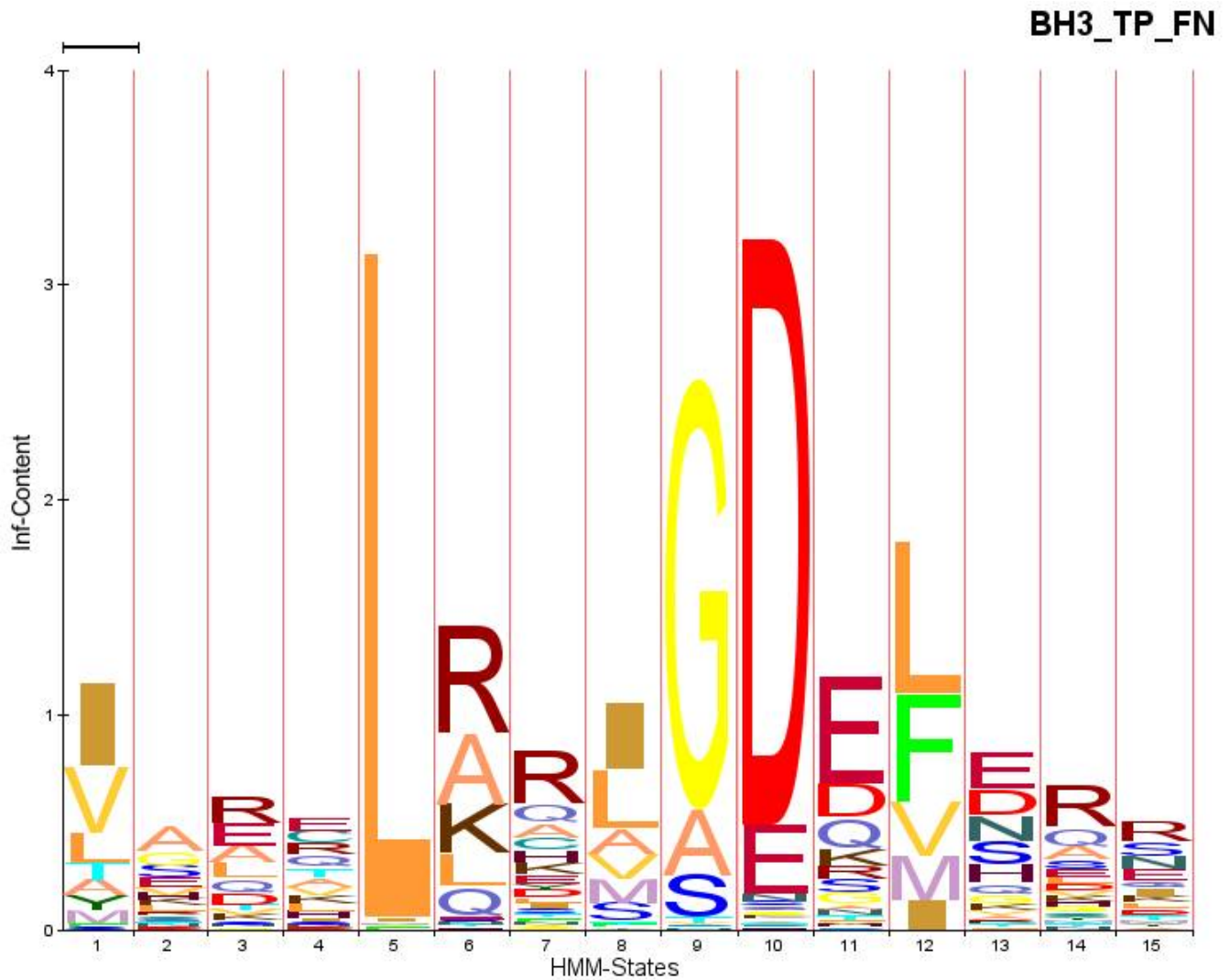


Figure 2. Information content of states within the BH3 HMM.

The information content of each amino acid at all 15 positions of the HMM are indicated by the letter size. Each position of the PROSITE BH3 regular expression, $[LIVAT] \dots L[KARQ]$.

$[IVAL]GD[DESG][LIMFV][DENSHQ][LVSHRQ][NSR]$, is a subset of those emitted by the corresponding HMM states. The figure was generated by HMM Logo [14].

zBAD	Y	G	Q	Q	L	R	R	M	S	D	E	F	D	K
hBAD	Y	G	R	E	L	R	R	M	S	D	E	F	V	D
mBAD	Y	G	R	E	L	R	R	M	S	D	E	F	E	G
hBAK	V	G	R	Q	L	A	I	I	G	D	D	I	N	R
mBAK	V	G	R	Q	L	A	L	I	G	D	D	I	N	R
zBAX	L	A	Q	C	L	Q	Q	I	G	D	E	L	D	G
zBAX2	L	A	N	T	I	K	V	I	G	D	K	L	D	Q
hBAX	L	S	E	C	L	K	R	I	G	D	E	L	D	S
mBAX	L	S	E	C	L	R	R	I	G	D	E	L	D	S
zBLP2	L	Y	R	V	L	R	D	A	G	D	E	I	E	R
hBCL-2	V	H	L	T	L	R	Q	A	G	D	D	F	S	R
mBCL-2	V	H	L	T	L	R	R	A	G	D	D	F	S	R
hBCLW	L	H	Q	A	M	R	A	A	G	D	E	F	E	T
mBCLW	L	H	Q	A	M	R	A	A	G	D	E	F	E	T
zBLP1	V	K	E	A	L	R	D	S	A	N	E	F	E	L
hBCLxL	V	K	Q	A	L	R	E	A	G	D	E	F	E	L
mBCLxL	V	K	Q	A	L	R	E	A	G	D	E	F	E	L
hBCLxS	V	K	Q	A	L	R	E	A	G	D	E	F	E	L
zBID	M	A	A	E	L	I	R	I	A	D	L	L	E	Q
hBID	I	A	R	H	L	A	Q	V	G	D	S	M	D	R
mBID	I	A	R	H	L	A	Q	I	G	D	E	M	D	H
zBIK	I	G	R	Q	L	A	Q	I	G	D	E	M	D	N
hBIK	L	A	L	R	L	A	C	I	G	D	E	M	D	V
zBIM	V	A	R	E	L	R	R	I	G	D	E	F	N	R
hBIM	I	A	Q	E	L	R	R	I	G	D	E	F	N	A
mBIM	I	A	Q	E	L	R	R	I	G	D	E	F	N	E
zBMF	I	G	Q	K	L	Q	L	I	G	D	Q	F	Y	Q
zBMF2	I	G	R	K	L	R	E	M	G	D	Q	F	Q	Q
hBMF	I	A	R	K	L	Q	C	I	A	D	Q	F	H	R
mBMF	I	A	R	K	L	Q	C	I	A	D	Q	F	H	R
zBOK1	V	S	S	V	L	L	W	L	G	D	E	L	E	Y
zBOK2	V	S	V	V	L	L	K	L	G	D	E	L	E	C
hBOK	V	C	A	V	L	L	R	L	G	D	E	L	E	M
mBOK	V	C	T	V	L	L	R	L	G	D	E	L	E	Q
hHRK	T	A	A	R	L	K	A	L	G	D	E	L	H	Q
mHRK	T	A	L	R	L	Q	A	L	G	D	E	L	H	R
hMCL1	A	L	E	T	L	R	R	V	G	D	G	V	Q	R
mMCL1	A	L	E	T	L	R	R	V	G	D	G	V	Q	R
zNOXA	C	A	Q	Q	L	R	N	I	G	D	L	L	N	W
hNOXA	C	A	T	Q	L	R	R	F	G	D	K	L	N	F
zBBC3	V	A	V	Q	L	R	T	I	G	D	E	M	N	A
hBBC3	I	G	A	Q	L	R	R	M	A	D	D	L	N	A

Figure 3. Alignment of BH3 domains from known and candidate Bcl-2 family members.

The sequences were aligned on the BH3 domain as predicted by the BH3 HMM. Grouping and rendering was performed with Jalview [32] using average distance joining on percent identity. Human, mouse, and zebrafish proteins are prefixed with h, m, and z respectively. Bcl-2 candidates are identified by bold italic gene names.

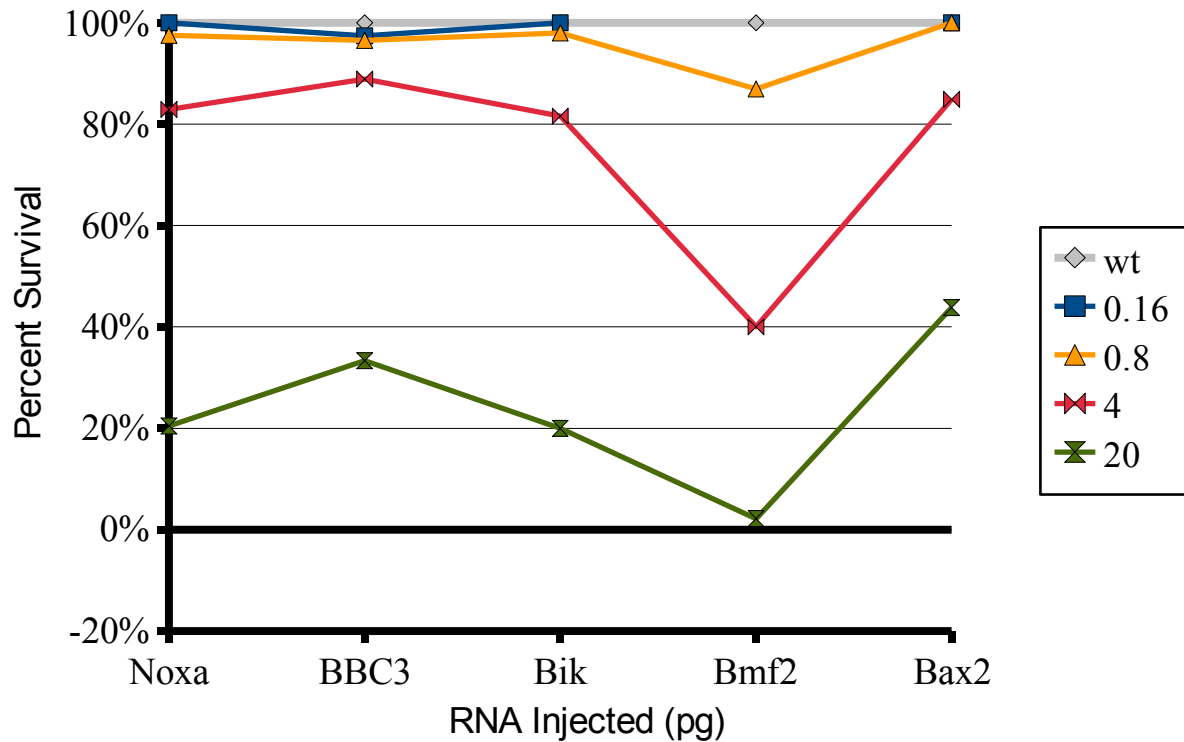


Figure 4. Percent survival of zebrafish embryos eight hours after injection with Bcl-2 candidates.

Cell death was determined visually by embryonic morphology. Zebrafish Noxa, a known Bcl-2 family member, is provided as a positive control. Injections of 500pg of GFP RNA as a negative control result in no killing (not shown).

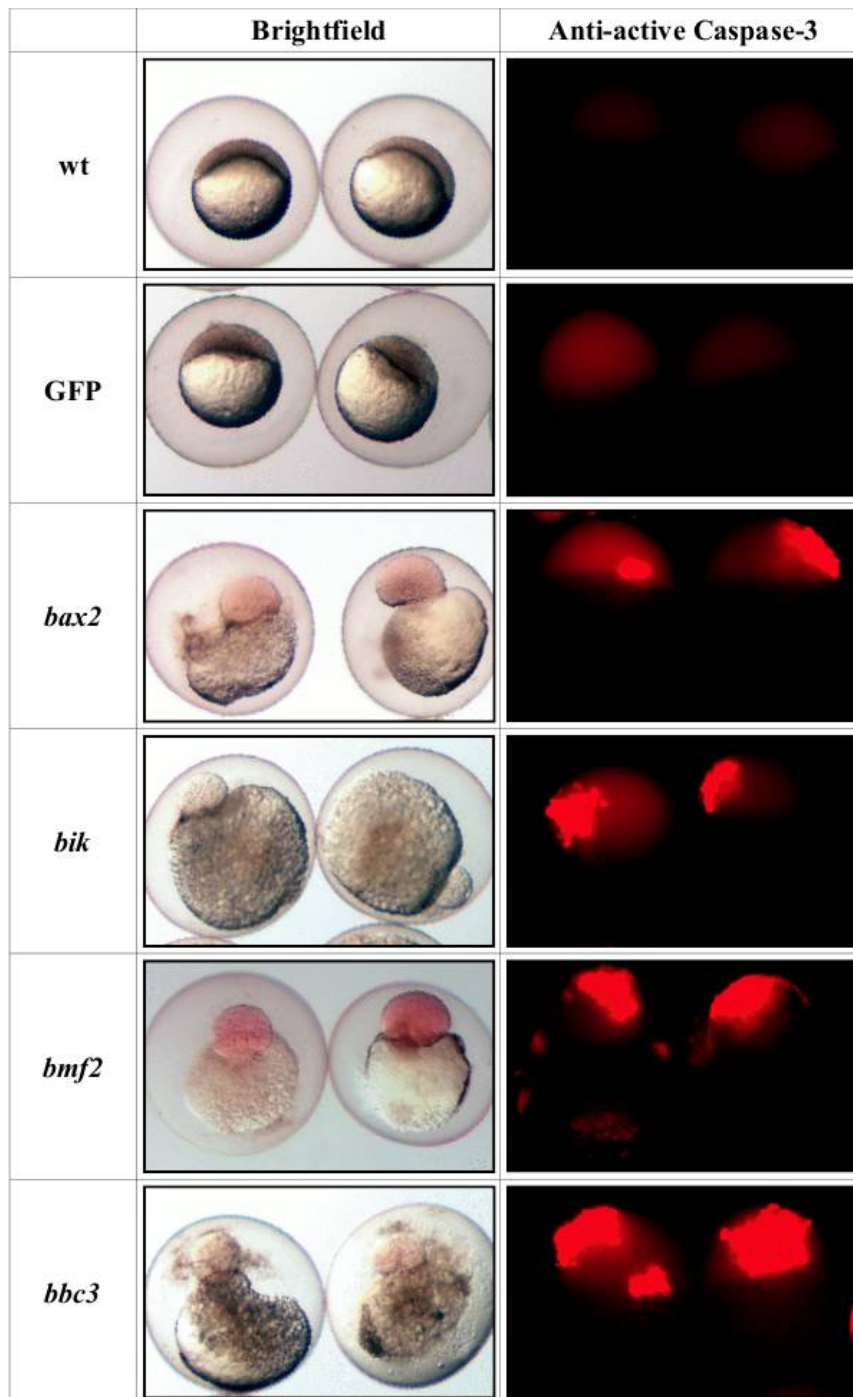


Figure 5. Visualization of common hallmarks of apoptosis induced by Bcl-2 candidates.

Embryos were injected with 500pg of GFP or a minimally lethal dose of proapoptotic candidate mRNA as described in [17]. Left panel: Yolk cell and blastomere fragmentation using brightfield microscopy.

Right panel: Immunofluorescent staining of active caspase-3.

Human				Zebrafish			Percent Ide/Sim	Human Domain Structure
Protein	NCBI Accession	UniProt	Locus	Protein	NCBI Accession	UniProt		
prosurvival								
A1	NP_004040	B2LA1_HUMAN	15q25.1	<i>unidentified</i>				
BCL-2	NP_000624	BCL2_HUMAN	18q21.33	Bcl	NP_001025424	Q564A4_BRARE	47/ 59	
BCLW	NP_004041	BCLW_HUMAN	14q11.2	<i>unidentified</i>				
BCLxL	NP_612815	BCLX_HUMAN	20q11.21	Bcl2l	NP_571882	Q90Z98_BRARE	52/ 67	
BOO	gi23396469	B2L10_HUMAN	15q21.2	Bcl2l10	NP_919379	Q8UWD5_BRARE	29/ 41	
MCL1	NP_068779	MCL1_HUMAN	1q21.2	Mcl1a	NP_571674	Q9I9N3_BRARE	41/ 61	
				Mcl1b	AAL32470	Q8UWD6_BRARE	35/ 57	
proapoptotic multidomain								
BAK	NP_001179	BAK_HUMAN	11q24.3	<i>unidentified</i>				
BAX	NP_620116	BAXA_HUMAN	19q13.33	Bax	NP_571637	Q9I9N4_BRARE	52/73	
				Bax2	NP_001013314	Q5EAR7_BRARE	30/52	
BCLxS	NP_001182	Q5TE63_HUMAN	20q11.21	<i>unidentified</i>				
BOK	NP_115904	BOK_HUMAN	2q37.3	Bok1	NP_001003612	Q6DC66_BRARE	67/81	
				Bok2	NP_957479	Q0GKC3_BRARE	57/74	
proapoptotic BH3-only								
BAD	NP_004313	BAD_HUMAN	11q13.1	Bad	NP_571654	Q9I9N2_BRARE	48/59	
BID	NP_001187	BID_HUMAN	22q11.21	Bid		Q0GKC5_BRARE	25/44	
BIK	NP_001188	BIK_HUMAN	22q13.2	Bik	NP_001038503	Q5R6V6_BRARE	43/56	
BIM	NP_619527	BIM_HUMAN	2q13	Bim	XP_685676		32/43	
				Bmf	NP_001038689	Q0GKC7_BRARE	33/47	
BMF	NP_001003940	BMF_HUMAN	15q15.1	Bmf2	NP_001038938	Q0GKC4_BRARE	50/64	
HRK	NP_003797	HRK_HUMAN	12q24.22	<i>unidentified</i>				
NOXA	NP_066950	APR_HUMAN	18q21.32	Noxa	NP_001038939	Q0GKC8_BRARE	51/60	

Figure 6. Correspondence between human and zebrafish Bcl-2 family members.

Human-Zebrafish alignments were performed with bl2seq using the BLOSUM45 matrix, gap open cost 12, gap extension cost 2. Domain structures are those of the human homolog, aligned on the BH3 domain if present, otherwise on the BH1 domain. Color legend: blue, BH1; yellow, BH2; red, BH3; green, BH4; black, transmembrane. The Bcl-2 candidates identified herein are denoted by bold gene names. ZFIN [33] protein names are used where available.

Tables

Z'fish Protein	Source Database and Accession	Entrez accession	Human Protein	E-value	Score	% Identity	% Coverage	z→h rank	h→z rank	Cons. Syn.?
Bax	RefSeq:NP_571637	NP_571637	BAX	2.00E-47	189	51	98	1	1	N
Bax2	E35:ENSDARP00000040899	ABI18121		1.00E-14	81	33	51	1	2	Y
Bik	UP:Q5RGV6_BRARE	ABI18125	BIK	1.41E+04	20	47	12	-	-	Y
Bmf	RefSeq:NP_001038689	NP_001038689	BMF	1.00E-05	50	32	91	1	1	N
Bmf2	E35:FGENESH00000082230	ABI18127		1.10E-02	42	41	42	1	2	Y
BBC3	E35:FGENESH00000078270	ABI18122	PUMA	2.10E+01	30	25	49	-	-	Y

Table 1. Novel zebrafish Bcl-2 proteins and comparison with with human homologs.

The e-value, score, percent identity, and coverage (length of alignment / length of shorter sequence) were determined by BLAST; z→h rank is the rank by BLAST alignment of the human homolog among all human sequences in Unison using the zebrafish gene as a query with the BLOSUM45 substitution matrix and default parameters, and conversely for h→z rank. Bcl-2 candidates are denoted by bold italic gene names. Human and Zebrafish Bik, and Human and Zebrafish BBC3, do not align by BLAST with permissive parameters. Conserved synteny was assessed manually using Ensembl contigview using criteria outlined in [34]. Genes that define such conserved synteny are: Bax2, FTL; Bik, MCAT; Bmf2, RAD51 and FAM82C; BBC3, SUMO1. Zebrafish Bax and Bmf were previously known Bcl-2 family members and are shown for comparison of reciprocal BLAST results and co-orthology. Source database abbreviations: E35, Ensembl release 35; UP, UniProtKB; RefSeq, NCBI RefSeq. ZFIN [33] protein names are used where available.

Additional files

These are available at <http://harts.net/reece/pubs/> .

Additional file 1 **bcl2-list.txt** [text/UniProt IDs]

Curated list of UniProt sequences identifiers of known and predicted Bcl-2 family members

Additional file 2 **BH1.aln** [text/ClustalW]

Alignment of PROSITE BH1 pattern true positive (TP) and false negative (FN) sequences

Additional file 3 **BH1.hmm** [text/HMMer]

HMM built from BH1 alignment

Additional file 4 **BH2.aln** [text/ClustalW]

Alignment of PROSITE BH2 pattern TP and FN sequences

Additional file 5 **BH2.hmm** [text/HMMer]

HMM built from BH2 pattern alignment

Additional file 6 **BH3.aln**[text/ClustalW]

Alignment of PROSITE BH3 pattern TP and FN sequences

Additional file 7 **BH3.hmm** [text/HMMer]

HMM built from BH3 pattern alignment

Additional file 8 **BH4.aln** [text/ClustalW]

Alignment of PROSITE BH4 pattern TP and FN sequences

Additional file 9 **BH4.hmm** [text/HMMer]

HMM built from BH4 pattern alignment

Additional file 10 **BH4m.aln** [text/ClustalW]

Alignment of PROSITE BH4 matrix TP and FN sequences

Additional file 11 **BH4m.hmm** [text/HMMer]

HMM built from BH4 matrix alignment

Additional file 12 **Bcl2m.aln** [text/ClustalW]

Alignment of PROSITE Bcl2 matrix TP sequences

Additional file 13 **Bcl2m.hmm** [text/HMMer]

HMM built from Bcl2 matrix alignment

Additional file 14 **mining-view.txt** [text/SQL]

SQL view definition for sequence mining

Additional file 15 **mining-results.txt** [text]

Mining results obtained from the mining view

Additional file 16 **novels.fa** [text/FASTA]

Candidate Bcl-2 sequences