Hi Dr. Eisen,

Thank you for taking your time to read this. My name is Robert Grinshpon, I’m a biochemistry Phd candidate at NCSU, and I’m doing my research on caspases under A. Clay Clark at UT - Arlington. Just to get you refreshed on the caspase family, they are Cysteinyl aspartate-specific proteases that have representative members in every kingdom of life. Although their ancestral function is debatable, their role in programmed cell death was a prerequisite for multi-cellularity. Caspases are difficult to classify, which is one problem I’d like to solve with this research. There are eleven human caspases (twelve in some sub-saharan populations). It has been reported that three ancestral caspases existed when vertebrates diverged. Caspase-(1/4/5) play a role in inflammation, caspase-(2/8/9/10) are classified as ‘initiators,’ and caspase-(3/6/7) are the ‘executioners’ of apoptosis. Here’s where it starts to gets difficult, some caspases are monomers and they must form a dimer to become active. Upon a death signal, the monomeric caspases use their recruitment domains to induce dimerization by increasing their localized concentration. The initiators (-2, -9) and the inflammatory (-1, -4, -5) have CARD domains, and the initiators (-8, -10) have DED domains, shown in the images below. The executioner caspases (-3, -6, -7) don’t have a long pro-domain, and neither does caspase-14; (which has a discrete role in keratinocyte terminal differentiation). The short pro-domain caspases evolved a mechanism that allowed them to exist as dimers while simultaneously inhibiting their activity. (This mechanism is one feature my PI is interested in explaining). Caspases are also classified by their substrate preference motif; (-1, -4, -5, -14) are group I, (-6, -8, -9, -10) are in group II, and (-2, -3, -7) are group III. However, only aspartate at the P1 position is inflexible (at least in vertebrates).

Active-site targeted drugs exhibit cross-reactivity among caspases due to the conservation of the active-site structure. So our lab is trying to learn about the innate allosteric regulatory mechanisms, and assess them as potential drug targets. My preliminary project proposal was based on an allosteric mechanism described in caspase-6. It is the only caspase inhibited at physiologically relevant concentrations of zinc (and only zinc) when bound 35 angstroms away to an unconserved exosite. The crystal structure of caspase-6 revealed that the mechanism of inhibition involved a strand-to-helix conformational change of helix-3 that misaligns the active-site, whereas the other caspases have helix breaking glycine and/or prolines on their helix-3 I used AGADIR, a helical propensity prediction algorithm, to redesign caspase-6 helix-3 to mimic the propensity of caspase-3 helix-3. The mechanism appears to have evolved in the mammalian lineage at first glance, because the exosite is not conserved in fish, nor does the fish’ caspase-6 helix-3 have a propensity to change conformations. What I found was that only 2 of 18 residues along the extended helix we’re conserved between caspase-3 and -6, and restoring E135D had a drastic effect on the helical propensity (Image below). After doing some research I learned, although E and D are functionally similar, their participation in helix formation is drastically different. Aspartate is considered a strong helix capping residue, while glutamate is not.

This discovery is what made me want to learn more about how the mechanism evolved. I’m not a trained evolutionary biologist, nor a phylogeneticist, so I was taking on an ambitious task. I became interested in doing an ancestral state reconstruction (ASR), and resurrecting the caspase-(3/6) ancestor protein in the lab, then characterize the substrate preference motif, or do N-terminomics (sorry ☺). I have also read that phylogenetic algorithms can potentially predict every mutation along any given branch. I learned that the quality of the MSA determines the quality of the tree, and the ASR in turn. Then I discovered the state of the databases. I don’t know about all the tools available to me, but there are over 40,000 hits for ‘caspase’ on NCBI, and I was only able to extract roughly 1200 good sequences of the 25,000+ animal caspases. After generating the MSAs, I found that a percentage of the sequences were not so good after all, and needed to be removed from the data set. I saw this as a problem, and that it is likely to become a bigger problem as we continue to produce more data than we can process.

Then, my PI wanted me to begin a collaboration with a professor working on coral and sea anemone, and caspases are involved in coral bleaching. I was given 12 coral caspase sequences by the collaborating student, and she named them according to COG. I am very familiar with the human caspases, and at a quick glance I could tell the names were not a reflection of how the protein would function. Then it dawned on me that if COG naming is a common practice, then there are potentially thousands of genes that were classified and named incorrectly, and submitted into the chasm of sequence databases (maybe I’m just not experienced enough in navigating databases). Apparently, fewer than 0.5% of current UniProt database protein annotations come from experiments. This becomes a problem when bioinfomaticists, that are not familiar with the structure-function elements, must rely on the quality control of the submitter of the data. If the sequences used in subsequent computation are related by peripheral consensus and not functional homology, then the quality of the MSA will be compromised, and so on down the pipeline.

One of the coral caspases sequences provided to me was named caspase-2, but when I manually scrutinized the sequence, it appeared to be a hybrid between caspase (-2, -3, -9). It had highest consensus for caspase-2, caspase-9’s dimer interface, and the highest similarity to capase-3’s active site binding loops. Which has led me to think very hard about this, and led me to the hypothesis that protein families are modulations of a conserved structural scaffold. If some degree of caspase (-2, -3, -9) character was conserved, then the hybrid will have some degree of their respective functions. I see a problem in the future if we don’t re-correct our course. It’s too bad that your plos one biodiversity hub didn’t pan out. I was thinking something like that needs to exist.

In my research, I’ve found two prime examples of scientific articles that drew incorrect conclusions based on a poor understanding of evolution, which lead to years of dead end thinking. I believe the misunderstanding comes from the common myth that evolution implies increased complexity, yet that fact has been completely ignored when it comes to making scientific claims from the data presented. The absence of knowledge does not imply knowledge, and opinions without caveats are dangerous. The two model organisms for programmed cell death (PCD) are *c. elegans* and *d. melanogaster*, and by coincidence, neither of them have an extrinsic pathway for apoptosis. It was concluded that extrinsic apoptosis was a mammalian invention, until genes involved in extrinsic apoptosis were discovered when the sea anemone genome was submitted in 2006. Another example, only one active caspase was found in *c. elegans*, CED-3. It was then concluded that since nematodes are evolutionarily older, that CED-3 must represent the common ancestor that all human caspases evolved from. It was later shown that there were two caspase gene deletion events in the nematode lineage, of the three ancestral to vertebrate caspases. My point is, that drawing hard conclusions from small snippets of data is bad science, and bad for science.

The complexity of the caspase family before vertebrates becomes very blurry very fast, therefore I don’t believe hard conclusions can be drawn until we have a much more complete picture of evolutionary events (I explain my idea for further down). I determined that using the arbitrary human naming scheme to represent the functions of all orthologs is inherently a flawed idea. There are many non-homologous to human caspases that I have in my database, but I don’t know where to group them because their name has nothing to do with their evolutionary relationships. And I’d imagine that some of the caspases in my database, are not directly related to the gene-tree they are aligned into. So I began hatching out a method to annotate unknown caspases that are older than fish, and predict their function. I also proposed changing the naming scheme from being based off the arbitrary names in humans, that we group caspases based on their homology to the three ancestral caspases that pre-date vertebrates. For example, caspase (1, 4, 5, 14), (2, 8, 9, 10) and (3, 6, 7) should be changed to – caspase-(A1, A2, A3, A4), (B1, B2, B3, B4) ect. Where A1 would be the most related to the ancestral state, and its duplications named subsequently. I also propose using a 3-4 letter combination of upper and lower case letters and numbers that designates the species and phylum of the gene.

My idea is to partition the identifying characteristics of each set of orthologs into as many possible categories, and assign each category a weight based on how much it contributes to the identity of that group. The algorithm will assess a score for each partition, and then aggregate a total score for any unknown caspase. If caspases are modular like I suspect, then a percent identity for caspase (-2, -3, -9) should be reported for the unknown coral caspase, and the known conserved functions could be tested based on the results of the analysis.

The first step is to remove the positions from the identification template that are conserved among all paralogs and orthologs. I would do this by making a series of consensus sequences (based on timescale, and %conserved cutoff) for each set of orthologous-to-human caspases. Peripheral elements of secondary structure that differ in size and conformation make up as much as fifty percent of a given sequence, and this is why BLAST E-scores are not the best indicators of homology. Removing the completely conserve residues from pairwise consensus will automatically increase the significance of the residues that are in consensus, but not removed from the identification template. The resulting E-score will be just one of the variables that contribute to the identification. Every residue that remains will be weighted based on their level of conservation within each orthologs’ MSA. Sites of experimentally determined function will annotated onto the template sequence, and those residues will be given a higher weight. Important positions will include post-translational modifications, and functions that require groups of interacting residues such as; allosteric networks, protein-protein interaction interfaces, substrate binding residues, metal coordination residues, and regulatory cleavage-sites.

I’m also wondering if the length of domains has any evolutionary significance. For example, we know that the length of the inter-subunit linker does have a functional role, so what about peculiar extensions of functionally important turns, and loops. Ultimately, multiple templates (or two) for each set of orthologs, and then an unknown sequence can be compared to each template. The output will be %identity to each template.

After considerable thought, I checked the literature again to see if someone had already done my idea. I’ve found one available annotation method that is used to predict functionally important regions, called evolutionary trace annotation (ETA). I have a lot more reading to do on it, but I think I could use this method to determine the important residues for my identity template. To my understanding, ETA can take a known crystal structure, use it to predict a crystal structure for every ortholog in a given dataset, then compile and superimpose the structures data into a structure template. From there I can determine the hydrogen bonds of every functional residue. The changes in hydrogen bonding networks should be important for the functional prediction of unknown sequences. Some other attributes that I considered annotating are %likelihood of secondary structure, chemical properties of the amino acid, probability of mutation based on a distance matrix, and buried or solvent exposed. I still have a lot more thinking to do.

Then I was thinking… even if I determine the unknown coral caspase to be a hybrid (-2, -3, -9), then how will I know where this protein fits into the bigger picture of evolution… I won’t. Ill he to rely on the same speculation that lead to years of perpetuating myths of evolution. Then I came up with the following ambitious idea.

The idea I’m about to present is not part of my actual PhD project. I just want to write my ideas down, make a case for it, and see what you think. I trust that you’ve thought about this stuff a lot more than I have. First, I think that we need to create a species tree using all the available full genomes that can be appended as new genomes are sequenced. It is my understanding that a way to improve ASR certainty, is to reconcile the gene tree to a known species tree. This will define true distance relationships since the length of the branch is directly related to the number of mutations that occurred. Having true branch lengths should improve our understanding of protein evolution. Then we can hypothetically extract every member of every protein family that exists in that species, and list repertoire at every tip of the tree. We could then do a modified ASR to determine the set of proteins of any given family, at every node on the tree. The we can deduce the evolutionary events; such as duplication, deletion, neofunctionalization, and/or subfunctionalization, based on the changes in the repertoire (or tested function) from node to node. This potentially be accomplished for every gene family. Then way we can look at the presence or absence of interacting partners from other families at any ancestral node, and better predictions could be made about the global function, not just molecular function.

Im not so sure my next idea will work, but I think it would be awesome if it did. Now that the evolutionary events have been mapped onto the species-tree, we can focus in on the neofunctionalization events. Using ASR, we can theoretically re-trace the evolutionary path of mutations along branches of interest. We could then add the mutations one at a time to a molecular dynamics simulation. My hypothesis is that we will be able to visualize the structural changes over time that confer new protein functions. We could also test those predictions in the lab, hopefully find the exact mutation that allowed for new function. Thanks for reading this if you made it all the way. Any advice is greatly appreaciated. If you think Im wasting my time or someone has already done this, please tell me.







