THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOLOGY

USING GDDA BLAST AS A STATISTICAL METHOD TO STUDY THE HAD SUPERFAMILY

EVAN J. SMITH
Spring 2010

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Biology
with honors in Biology

Reviewed and approved* by the following:

Randen Patterson
Assistant Professor of Biology
Thesis Supervisor

Richard Cyr
Professor of Biology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

The Haloacid Dehalogenase (HAD) superfamily contains a diverse group of phosphoryl and carbonyl transfer enzymes. The family spans all superkingdoms of life and is present in many of the cellular compartments catalyzing reactions on very diverse substrates from nucleotides to saccharides to proteins. Despite the exploration of many different biochemical spaces, all of the members of this superfamily have two unique structural similarities: a squiggle and a flap motif located around the active site. Structural modifications called caps around these structural motifs allow more diverse interactions with substrates and increases reaction efficiency. Additionally, they provide a base for further classification. However, classifying, and establishing evolutionary trees for HAD is difficult because there exists a large amount of sequence divergence within the superfamily, even between families with similar cap structures. A research group led by Burroughs performed a phylogenetic analysis identifying 5 distinct lineages, and providing evidence for convergent structures. Their evolutionary relationships were mainly inferred from structural and functional similarities not statistical algorithms. Most phylogenetic studies are performed using statistical tools like multiple sequence alignments (MSA). Unfortunately multiple sequence alignment algorithms have failed in the past when presented with highly divergent families like HAD. However, using a novel statistical method, GDDA BLAST, which was created with specific modifications to identify sequence similarities between related but diverged sequences we sought to better resolve the evolutionary relationships in the superfamily. As a null hypothesis we used GDDA BLAST assuming a single lineage for HAD. Evolutionary contradictions were found within the lineage such as lack of monophyly in the groups and inconsistent speciation. This evidence led us to further investigate an evolutionary tree with multiple lineages like that suggested by the Burroughs group. Using a hierarchical clustering method we grouped the families into 4 distinct lineages before continuing with a neighbor joining algorithm. The resulting phylogram contained few evolutionary inconsistencies and those it did

could be explained through lateral transfer or other evolutionary mechanisms. It has become clear that GDDA BLAST can be an effective tool when used with clustering methods to generate robust phylogenetic, evolutionary relationships between divergent families with convergent structural elements.

TABLE OF CONTENTS

List of Figures

Acknowledgements

Professor Randen Patterson

Professor Damian Van Rossum

Graduate Student Gaurav Bhardwaj

Introduction

The Haloacid Dehalogenase (HAD) superfamily is broad and encompasses 33 different enzyme groups and spans all three superkingdoms of life. The many enzymes contained in this superfamily occupy slightly different biochemical niches within the cell and vary in their essentiality to the cell. However, they all perform some sort of carbon or phosphoryl group transfer (1). HAD enzymes can be found in many different compartments of the cell, essentially wherever phosphate groups are found. This superfamily is particularly interesting because its families show many significant structural similarities while varying greatly in sequence similarity (1)

The superfamily is a collection of enzymes unified by the presence of two structural features. The structures are squiggle and flap motifs suggested to be integral to the biochemical necessities of the enzyme families. The superfamily is named after the specific Haloacid dehalogenase (HAD) enzyme which contains the archetypal squiggle and flap structures (1).The HAD superfamily consists of phosphoesterases, ATPases, phosphonatases, dehalogenases, and sugar phosphomutases.

Good examples of the more ubiquitous members of the HAD superfamily are the phosphatases and the ATPases. These enzymes catalyze similar reactions removing phosphate groups. The ATPases decompose a high energy adenosine triphosphate (ATP) to adenosine diphosphate and eventually adenosine monophosphate (AMP). The removal of a high energy phosphate group is accompanied by a large energy release which is harnessed by ATPases to perform cellular functions (2). Resetting the chemical and electrical potentials in an active nerve requires sequestering of potassium and sodium in separate compartments against their chemical gradients. This process is powered by ATPase activity (2). Similarly a phosphatase removes phosphate groups but usually from proteins. Many proteins involved in cell signaling and

metabolism bind phosphates which act as a sort of switch rendering the intermediate, channel, enzyme, etc… active or inactive. A good example of this is the dephosphorylation of phosphofructokinase (PFK) by phosphatases which reduce its activity. The PFK enzyme controls the committed step in glycolysis (energy production); therefore the PFK phosphatases inhibit the breakdown of sugars and production of energy (2).

While much is understood about the function of the HAD enzymes, understanding the evolutionary history of this family can be challenging. There exists a large amount of protein sequence divergence within the super family which can be difficult to understand using traditional sequence comparison techniques called multiple sequence alignments (MSA). MSAs compare protein sequences and based on statistical analyses they can develop evolutionary distances. A previous research group had tried to infer a phylogenetic history based mainly on structural similarities and MSAs. The MSAs were largely ineffective. Most of their phylogenetic tree was generated from structural features.

Within the HAD superfamily one of the main structural identifiers is the presence of insertions. The folded enzymes are characterized by their elaborations which have the resemblance of caps. The distinction based on cap elaborations was developed by the Burroughs research group and is the basis for much of their phylogenetic analysis. They identified two locations where insertions are very common. The Burroughs research group splits the superfamily into three broad categories. The first group called C0 has either no insertion or minimal inserts. The final two groups, C1 and C2, possess inserts at location 1 or location 2, respectively. However, there remains a significant amount of gray area: for instance enzymes with one cap module and a minimal second flap, or enzymes with no cap modules but extensive insertions. This continuum between no caps and two caps would be expected in a divergent superfamily, the ambiguous caps serving functionally effective transition stages.

In fact, Burroughs proposes that the cap structures confer specific catalytic abilities. They suggest that squiggle and flap motifs and eventually the caps allow the active site to exclude solvent by sealing it off. The most simplistic approach to this concept is the rudimentary C0 cap. Not all members in the C0 group contain inserts but those that do show some basic solvent exclusion characteristics. A C1 cap configuration found in the majority of HAD enzymes shows similar catalytic cycles, the enzymes assume open and closed conformational states (1). Additionally, the added inserts function to better separate the solvent from the active site so that it does not interact with reaction efficiency(1). Little is known about the function of C2 caps but according to the Burroughs research group they are thought to function along similar principles.

The research group explains the divergence of the HAD superfamily as the rapid exploration of a diverse "workspace." What they mean by this is that structural modifications and insertions provided a greater surface interaction with substrates. This allowed for a more specific interaction targeting the substrate of choice and excluding others. Additionally, this allowed the enzymes to interact with a wider range of substrates in general leading to a rapid diversification (1). The elaboration and emergence of diverse cap structures improved solvent exclusion, substrate recognition and reaction efficiency.

Despite some uncertainty, the Burroughs group generated an inferred evolutionary tree. Assuming a presence in the last universal common ancestor(LUCA) the group was able to group/infer a tree with monophyletic enzyme families based on structural similarities between the HAD family subclasses and subgroups. This tree can be seen in Figure 1. Most of this was done through cap analysis and comparison as previously described and the tree denotes location of structural elaborations with O or X marks (Figure 1). Preliminary clustering was performed through a program called BLASTCLUST. Beyond that they performed a clustering based on structural signatures. For instance, they inferred a bifurcation between a C0 group with a five-stranded core sheet and a group with a 6-stranded core sheet (Figure 1). All the remaining cap

assemblages are thought to have arisen after this split. Additionally, the research group broke down the constituent assemblages based on cellular location and biochemical niches which are color coordinated in the tree (Figure 1). In the internal relationships between diverged subgroups, conventional multiple sequence alignment (MSA, statistical) methods were employed to differentiate lineages. This allowed them to have some statistical significance to their conclusions.

Cap subclasses

Looking more closely, the C0 group can be subdivided based on insertions. The simplest C0 enzymes contain few or no additional residues at either position. This elaboration includes certain phosphatases (KDO 8-P). Additional C0 phosphatases have longer elaborations at these positions yet are simplistic enough to not be categorized as being cap-worthy. The polynucleotide-phosphatases are included in this subgroup. In reality there is a continuum between simple C0 subgroups and more elaborate subgroups.

The C1 group of HAD enzymes is broken into two different subclasses. The first class termed the alpha-helical C1 caps by the Burroughs research group possesses different levels of complexity. The most simple insertions are observed in the acid phosphatase and cN-1 nucleotidase families. The remaining alpha-helical C1 enzymes comprise the majority of HAD enzymes, and can be broken down into three more subgroups. The Burroughs research group viewed the C1 caps as an evolutionary series of increasingly complex alpha-helices. Structurally, they differ mainly in the contacts that the additional inserted residues make. Additionally, they differ in the extent to which the helical caps fold, and the size of these caps, they differ significantly in their sequence data. Nevertheless there is incredible similarity in their overall topology and therefore they are still classed together.

The second subclass of C1 is called the P-type ATPase caps. This subclass describes a group of structures that are determined by Burroughs. to be evolutionarily unrelated to the previously described alpha-helical C1 caps. This is because the P-Type ATPase caps contain a structural pattern formed by internal duplications, repetitive structures. In fact the research group suggests that this group arose from C0 enzymes independent of alpha-helical C1 caps despite their single cap similarities. But, the group also identifies the differences within this subclass. Furthermore, within the P-type ATPase family there is diversity in primary sequence and insertions and deletions.

The HAD enzymes designated by a C2 distinction are further broken into two groups. And like the C1 caps the C2 group contains unrelated, independently developed subclasses. The larger groups are titled Cof-type phosphatases, NagD phosphatases and the smaller group is histidinol-phosphatase family. The major classes of C2 domains were determined to be unrelated despite having structurally similar motifs. There topologies are unique. Further it was suggested that the Cof-type phosphatases display repeating structural units and are thought to have arisen from serial duplications limited to the end region of the cap. The NagD subclass was similarly thought to have developed from serial duplications, but unique to other C2 caps. Nevertheless, the Burroughs group concedes that there is still an incredible amount of sequence diversity even within the subclasses despite topological similarities. The histidinol-phosphatase family was also given the distinction of having arisen independently. This small group contained unique, repetitive, but conserved domains.

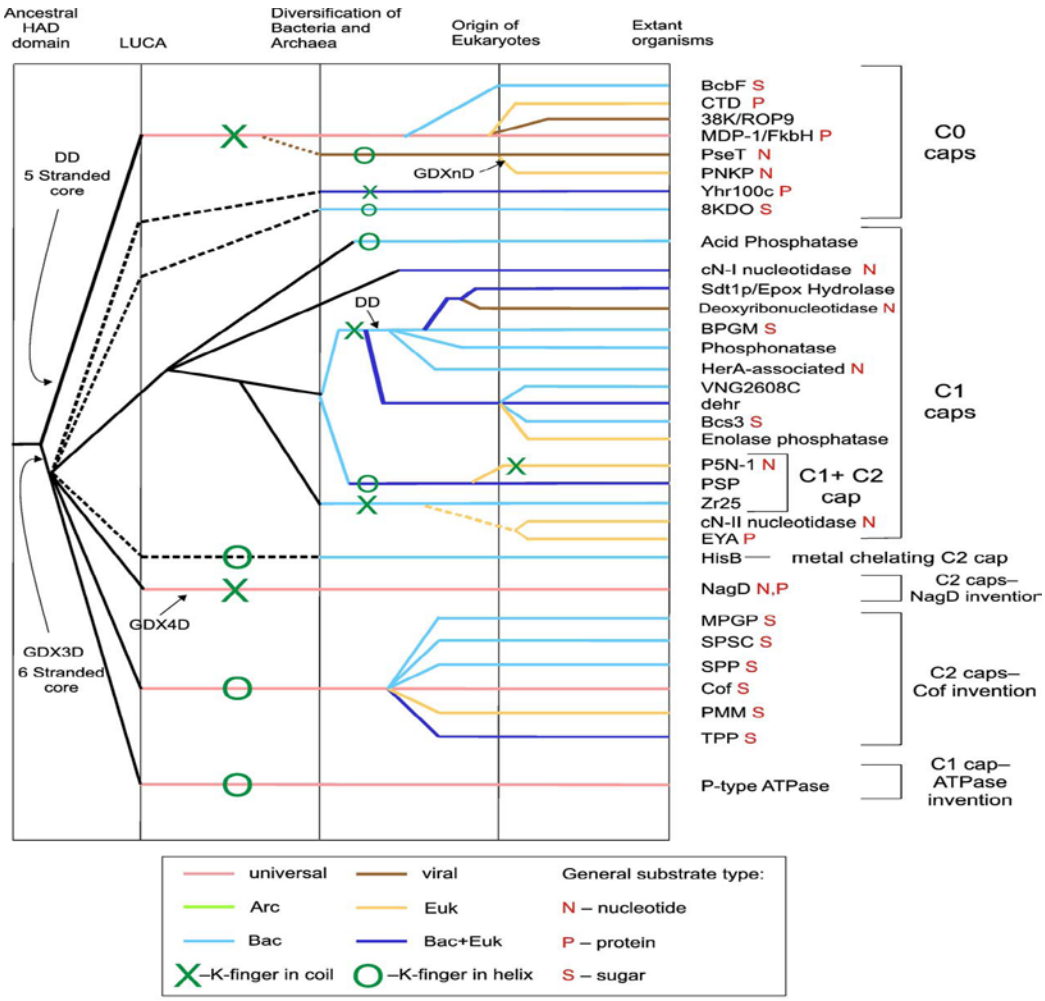


Figure 1: Inferred Phylogeny Tree of HAD Superfamily by Burroughs et al.:

Phylogenetic tree derived from examining structural, functional, and sequence similarities in HAD superfamily. There are 5 distinct lineages grouped by the brackets on the right. The cap designations and family names are also included in the grouping. Rough relation of the lineage's evolutionary time is included with major events (origin of eukaryotes, diversification of bacteria and archaea, LUCA) transecting. Arm length correlates to evolutionary time. The key at the bottom describes how the lineages are color coordinated based on presence in each superkingdom. Distinctions based on structural modifications are also included as per the key.

Examples of Convergence in the HAD superfamily

Convergence is one of the unique features of the HAD superfamily which has been uncovered by this phylogenetic analysis. There exist structural elements that independently arose throughout history and are seen in multiple lineages. Convergent structures can complicate a phylogenetic comparison when using superficial criteria for the analysis, like cap structures. The Burroughs group identifies many areas of convergence throughout its analyses.

Specifically, convergence refers to an evolutionary concept of unrelated lineages developing structures for similar functions. A good example of convergent evolution can be seen in wings. For instance, the bird and the butterfly's common ancestor did not have wing structures but both developed similar machinery to fly. The environmental pressures made it advantageous for a species to take to the air and control a new biological niche (3). Selection invented the wing. Therefore it is said that evolution led to a convergence between the lineages, and that the wing structure independently arose on at least two occasions. Convergence can also be defined as a trait. For instance the ability to fly is seen as convergence as well as the actual wing structure. Similarly, enzyme structures and sequences follow convergent principles in discovering novel functions.

Burroughs identifies 5 distinct lineages arising with similar structural and functional elements (Figure 1). For example, the evolution of structures capable of identifying nucleotides as substrates is seen in several families. These families include nucleotidases catalyzing the removal of a phosphate group from a nucleotide mono/di/triphosphate (e.g. ATP, GMP, and UDP). This catalytic ability arose on five different occasions. The cN-I, cN-II, Std1p, deoxyribonucleotidase, pyrimidine 5-nucleotidase families all possess nucleotidase activity despite being in the C1 or C2 structural assemblage. Similarly, phosphosugarmutase activity arose in separate lineages.

The elaboration of a C1 cap also arose independently on two occasions. The C1 assemblage includes the alpha-helical and P-Type ATPase families as previously discussed. The emergence of the insertions in both families was concurrent. The similar folded structures contain little sequence and motif similarity despite arising the similar biochemical need for solvent exclusion. Therefore, the two families share little ancestral similarity with regard to the C1 cap assemblage. They converged via similar functional principles. The Burroughs group placed the ATPase family in its own separate lineage at the bottom of the tree (Figure 1).

In fact, the principle of excluding solvent to improve reaction efficiency is a basic biochemical principle that could have been discovered independently. The method of solvent exclusion is seen via different strategies in C0 assemblage. Unlike other C0 families, the 8KDO enzyme forms a tetramer where solvent exclusion is a result of cooperation between elaborations in neighboring monomers (1). Each subunit does not have a strand worthy of being classified as a cap but when paired with other subunits the strand performs this basic solvent exclusion for another subunit (1).

Even within structurally similar groups there remains a large amount of sequence dissimilarity which may suggest convergence. For example the three categories of tetra-helical C1 caps share little sequence conservation despite sharing a similar structural topology (1). From this information alone one could draw different conclusions. On one hand it is possible that the families are related and possess a common ancestor with similar topology, and that over time neutral amino acid substitutions have taken place. This includes any substitution that does not affect the structure and function of the enzyme. For instance the replacement of a hydrophobic residue with another hydrophobic residue would not affect membrane spanning or internal interactions.

Alternatively, this relationship could suggest that the families are not related but independently developed similar C1 caps under similar biochemical selective pressures. Given

this second scenario it is not known whether the families are even ancestrally related without C1 caps, or completely unrelated. In the latter case the squiggle and flap motifs which unify the family, would also be convergent.

In the case of the HAD superfamily, functionally similar families should have similar sequences. However, they possess a large amount of protein sequence divergence. This idea is counterintuitive given parsimonious principles which state that a diverged protein should possess the least possible changes to discover a novel function (4). Furthermore, if an enzyme filling a specific biochemical niche undergoes some alteration to accommodate a slightly different niche then there should not be a plethora of sequence changes. Having said that, there also exist many neutral mutations in related proteins. Neutral mutations are largely the reason for highly diverged sequences with similar folding patterns. A thorough statistical study can attempt to resolve the confusing relationships and evolutionary histories of a diverged/converged protein family like HAD.

The superfamily presents problems for straight statistical analyses. There exist few MSA algorithms capable of generating meaningful relationships between highly diverged families. Additionally, the elements of convergence within the family would further complicate the phylogenetic analysis. The Burroughs group put together a very thorough analysis based mainly on structural elements but are lacking statistical significance. Fortunately, a novel MSA method called GDDA BLAST has been developed to tease out the evolutionary relationships between diverged proteins.

GDDA BLAST as a statistical method

Because sequences are divergent does not mean that they are unrelated. When sequences show less than 25% similarity they are said to be in the “twilight zone” of sequence similarity (5).

This zone is where many MSA algorithms are unable to generate positive alignments, it is likely why other MSAs failed with HAD. But, studies have shown that proteins can contain as little as 8% sequence similarity and still possess similar functions(similar folding) while proteins with 88% sequence similarity have proven to be functionally dissimilar(dissimilar folding)(5). Neutral mutations can arise reducing sequence similarity but not functional similarity. GDDA BLAST has shown to be effective in the "twilight zone" of sequence similarity, and to resolve some of the relationships that are very diverged (4). It is a good candidate to study the HAD superfamily.

Our approach with the HAD superfamily is purely statistical. We are studying the sequence data only, but we will address our results in light of the Burroughs research group. As mentioned, this method was established to build stronger evolutionary relationships between evolutionarily related proteins lacking sequence similarity. Exactly the type of relationships found between HAD families. The Burroughs research group studying the HAD superfamily encountered multiple instances of structurally related families with little sequence similarity. In fact they made assumptions based on structures and biochemical niches to resolve many of the evolutionary relationships. They did use some basic statistical analyses; however, as previously mentioned, many of these divergent families are difficult to order using older MSA methods for identifying evolutionary relationships.

GDDA BLAST has a few innovations to increase its effectiveness on divergent sequences. First, GDDA BLAST uses a seeding method to artificially improve sequence similarity, and thus identify more positive alignments. This methodology can be seen in Figure 2. This is done by modifying the original query sequence with a portion or "seed" of a profile sequence. The large set of query sequences (e.g. HAD superfamily) are all aligned/compared to a known group of similar profiles gathered from a protein database, in this case the NCBI non-redundant database.. Seeds are generated by taking some percentage of the profile sequence. Next, the modified query is aligned with the profile using rps-BLAST to score/identify positive

alignments. The “seeds” are positioned at every amino acid position along the query sequence, and rps-BLAST aligns each seeded query against the profiles, for every single position. Therefore, if a query is N number of residues long then there will be N number of seed positions and N number of alignments. Normally when aligning divergent proteins with little sequence similarity to known profiles the rps-BLAST algorithm cannot discover positive alignments. However, by using this “sliding seed,” the algorithm is given a starting point which should amplify the number of positive alignments. Phylogenetic profiles are generated from these alignments.

The phylogenetic profiles contain the alignment information from the positive alignments. Each query that is aligned with the profiles is scored on hits, identity, and coverage. A query with a very similar sequence to a profile will have a higher score than an unrelated sequence. The query is aligned with many different profiles each being scored. Only the alignments above a certain threshold of % coverage and identity are used in the phylogenetic profiles. The data collection process in Figure 2 shows this score. Coverage refers to the alignment length in relation to profile length. The percent identity value refers to a score given based on whether the profile residues match the query sequence, but can be affected by the type of residue dissimilarity. For instance, a mutation from hydrophobic amino acid to another hydrophobic amino acid (e.g. valine to alanine) would score as similar as compared to a mutation to arginine which would potentially have a greater effect on protein folding, and function. Lastly, the number of hits (alignments) is taken into consideration. It is necessary to remove poorly aligned sequences that give a false positive and can complicate the next step of generating a polygenetic tree.

The positive scores are entered into a matrix, the queries being on the y axis and the profiles on the x axis. From the query/profile scores a matrix comparing the query scores can be generated. Direct evolutionary distance is contained within these matrices. In effect what is being

done is not a direct comparison between all of the queries, but a comparison of the queries to a third party, the profiles. Euclidean distance can be inferred from this comparison assuming that the overall distance between query alignments translates to evolutionary distance (5). From the Euclidean distance a phylogram can be generated visualizing the phylogenetic history as seen in Figure 2.

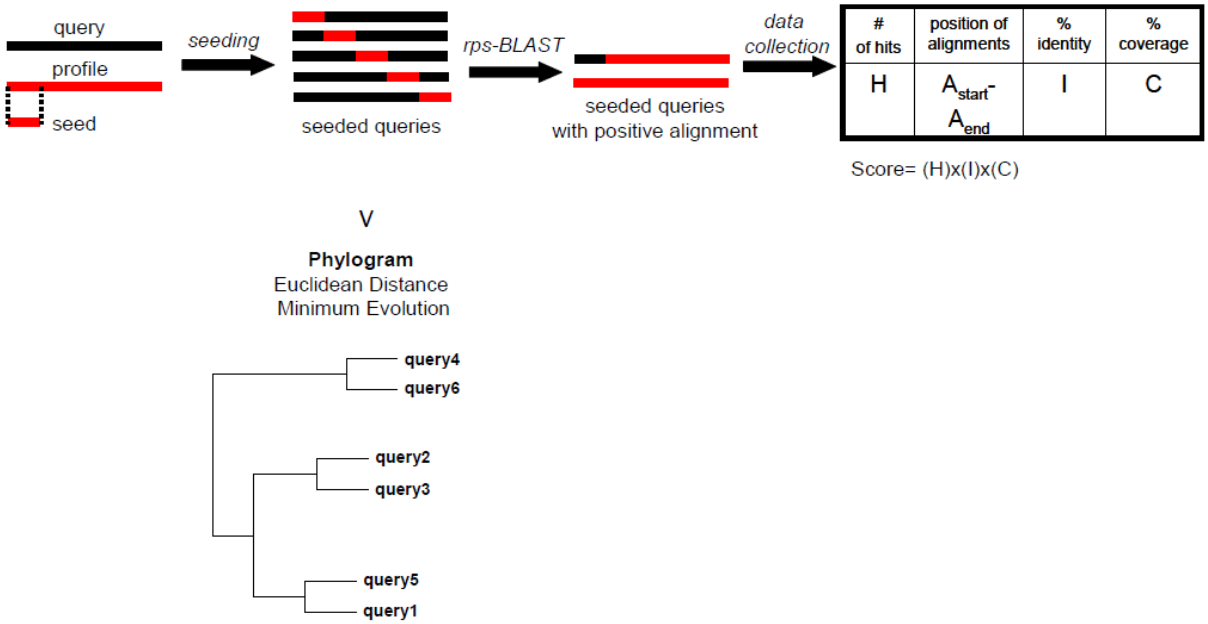


Figure 2: Schematic of GDDA BLAST computational method.

From top left: Basic seeding principle idea, excising "seed" from profile and creating seeded queries at every position. Rps-Blast generates alignments of seeded queries against profiles. Alignments are then filtered based on values for hits, identity, and % coverage. The scores from these profiles can be used to determine Euclidean distance between all of the queries and visualized in a phylogram using algorithms like neighbor joining.

GDDA BLAST has not before been used with a convergent family of enzymes. There are no provisions in the method to account for convergence like there are for divergence. In fact, the HAD superfamily serves as a good model for testing the ability of GDDA BLAST to study convergence. Convergence can be particularly confusing to MSA algorithms, and we may have to further modify our methods to obtain an evolutionarily acceptable tree. We know that we can study divergence, but the main question being asked is: can convergent families be studied using statistical analyses like GDDA BLAST?

Methodology and Results

Single Lineage Approach

First we looked to see if the HAD superfamily could be resolved into one robust tree. This first attempt could be viewed as a null hypothesis: assuming that all HAD members belong to a single lineage. We ran the 1561 HAD query sequences from the Burroughs research group against the profiles gathered from the NCBI non-redundant database. After generating the matrices from the previously described GDDA BLAST/rps-BLAST methodologies we used a Neighbor Joining (NJ) algorithm to construct the tree. NJ functions by identifying the closest taxa, pairing them, then identifying the next closest taxon to this pair and pairing them etc... Eventually this iterative process results in an unrooted tree with the last taxon as an outgroup. NJ is useful with HAD because it is simple, fast and can be used with large sets of sequences.

Unfortunately, this first tree was filled with inconsistencies (Figure 3). One of the most apparent is the incorrect speciation patterns down the branches of the tree. Incorrect speciation refers to a speciation pattern in the neighbor joining tree that is not consistent with assumed speciation events in our evolutionary history. For example, it is theorized that speciation progressed from prokaryotes to eukaryotes beginning with kinetoplastids, alveolates to plants, fungi and animaliae (6). A speciation sequence contradicting this sequence would be incorrect. Likewise, a speciation pattern progressing from plant to fungi to archaea would be incorrect. This pattern would suggest that the archaea and fungi have the evolutionarily closest forms of the enzyme with the plant species possessing an older more ancestral form. Given that the archaea and eukarya lineages split billions of years ago this pattern is unlikely (6). Additionally there are examples of enzymes becoming extinct in one species before reemerging in closely related

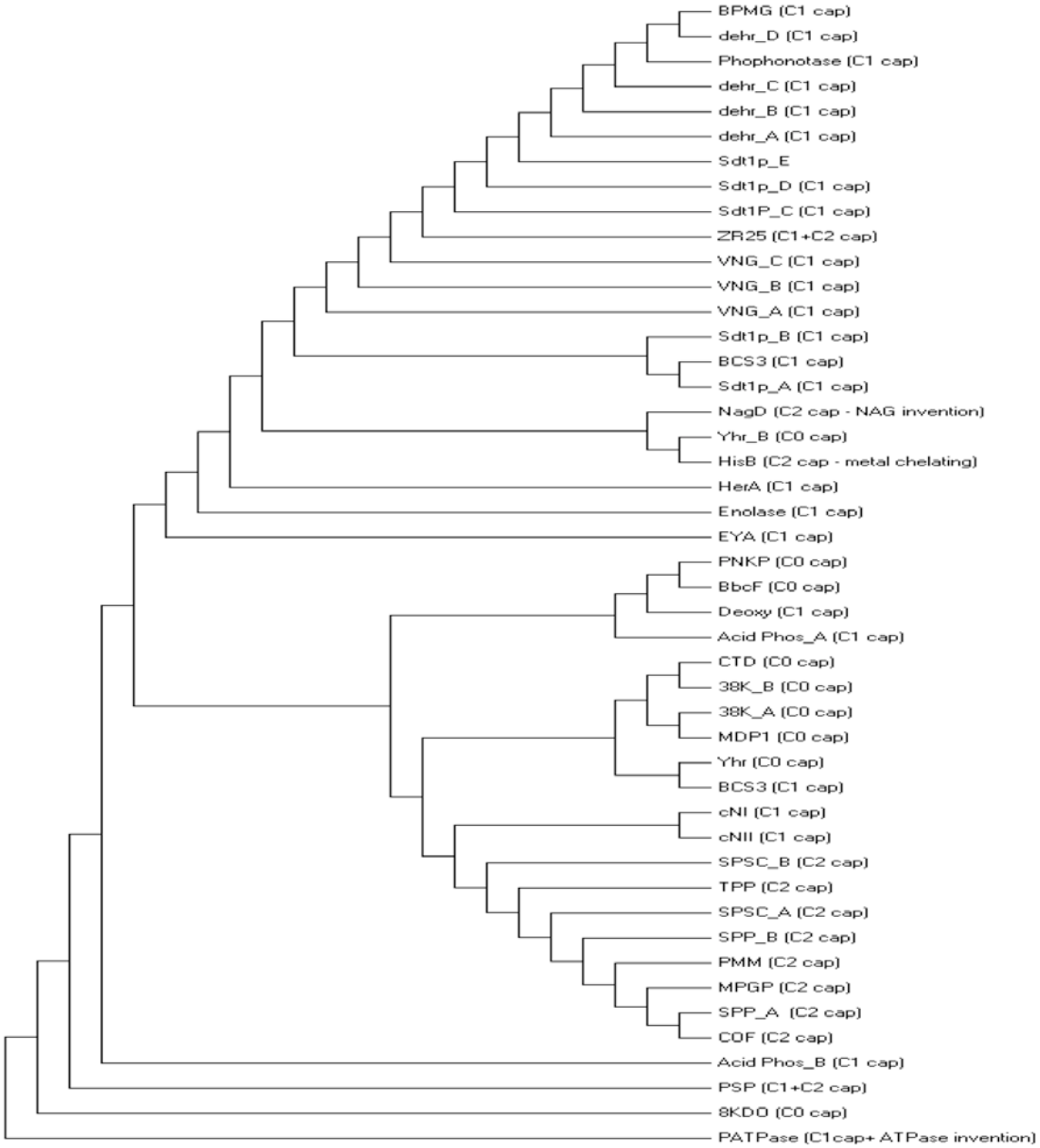


Figure 3: Phylogenetic tree of single lineage assumption:

Phylogenetic tree using GDDA BLAST, rps-BLAST, and neigbor joining algorithms, assuming a single lineage. The most derived groups/families lie at the ends of the branches with the more ancestral groups having the long branches. Several families are not monophyletic, and arise multiple times throughout the lineage. This tree was strictly a statistical analysis.

species. Within our single HAD tree a good example of this is Stdp1 family which reemerges throughout the tree. Another problem with the original tree is the lack of monophyly or cohesive groups. This also results in a family being dispersed throughout the tree like Sdt1p (Figure 3). Similarly, an enzyme family like the acid phosphatases can be found in various places throughout the tree (Figure 3). This is indeed the result of much sequence divergence even within an enzyme family.

A possible way to resolve a tree with speciation inconsistencies is pruning. This is done by removing sequences that consistently disrupt correct speciation patterns. Often there exist query sequences that are incorrectly identified or sequenced in the database. Understandably, these errors can greatly affect the evolutionary relationships. This would lead to the previously described scenario of an enzyme family disappearing and then reemerging. When pruning, as long as these sequences are within another well-defined, large group they can usually be removed without much problem. Trying to prune the first tree was unsuccessful. Excising a sequence here or there did not resolve the tree. In fact, extensive pruning and sequence removal was required to generate an evolutionarily consistent tree. However, at this point too many sequences were removed, overstepping the role of a minor pruning procedure. Furthermore, if you are able to remove sequences from a tree and maintain the same lineages and monophyly then a tree is said to be robust. Strong phylogenetic relationships should be robust.

There are two main explanations for incorrect speciation. One, incorrect speciation is the result of lateral gene transfer (or horizontal gene transfer). In the HAD superfamily this can manifest itself in a few ways. First, viral elements can crosscut species, inadvertently transferring genetic elements, resulting in the movement of HAD enzyme encoding genetic information (7). A possible example within this tree is the speciation pattern from Acid Phosphatase (Bacteria) to Deoxyribonucleotidase (viral) to BCBF (bacteria). The evolution of these enzyme families progressed likely through a viral intermediate which then passed the genetic information back

into another bacterial group. Likewise, prokaryotes can pass genetic information by direct interaction. This process is often seen between species that may share environmental niches (7). Additionally, lateral transfer could also have occurred because of plastid genome introduction during symbiosis. The introduction of mitochondria and chloroplasts from cyanobacteria and proteobacteria into archaea and eukarya lineages resulted in a genetic transfer and could explain incorrect speciation patterns (7). Furthermore, this transfer may have facilitated the opening of new biochemical and metabolic pathways for archaeal and eukaryotic species (1).

Two, the tree is incorrect, the phylogeny is incorrect, the single lineage assumption is incorrect. In this instance a new tree would need to be generated using a different clustering method or algorithm. These include maximum parsimony, UPGMA, and maximum likelihood to name a few. Given the speciation inconsistencies, the lack of monophyletic families and problems with simple pruning, it was decided that the single lineage hypothesis was impossible. Because the Burroughs group identified 5 distinct lineages it was expected that using GDDA BLAST would not resolve the family with one lineage. The next step was to employ a better preliminary clustering method to identify how many lineages are contained within this extensive family.

Hierarchical Clustering Approach

To improve upon this first tree we employed a different clustering method. After gathering our alignment information using rps-BLAST we scored the matrices in the normal N x M (query x profile) fashion. However, at this point instead of proceeding with a query-query comparison and using simple clustering method like NJ we used a broader clustering algorithm. This methodology can be followed in Figure 5. We created a matrix comparing the enzyme families. We scored this in a similar fashion to the N x N matrix, but as an average group-wise composite matrix (Figure 5). From this matrix we used a Pearson correlation to generate

statistical distance relationships between the composite groups. Using this information we were able to develop basic group vs. group evolutionary relationships. From this point we continued with a normal neighbor joining algorithm resolving the specific relationships within these larger clusters.

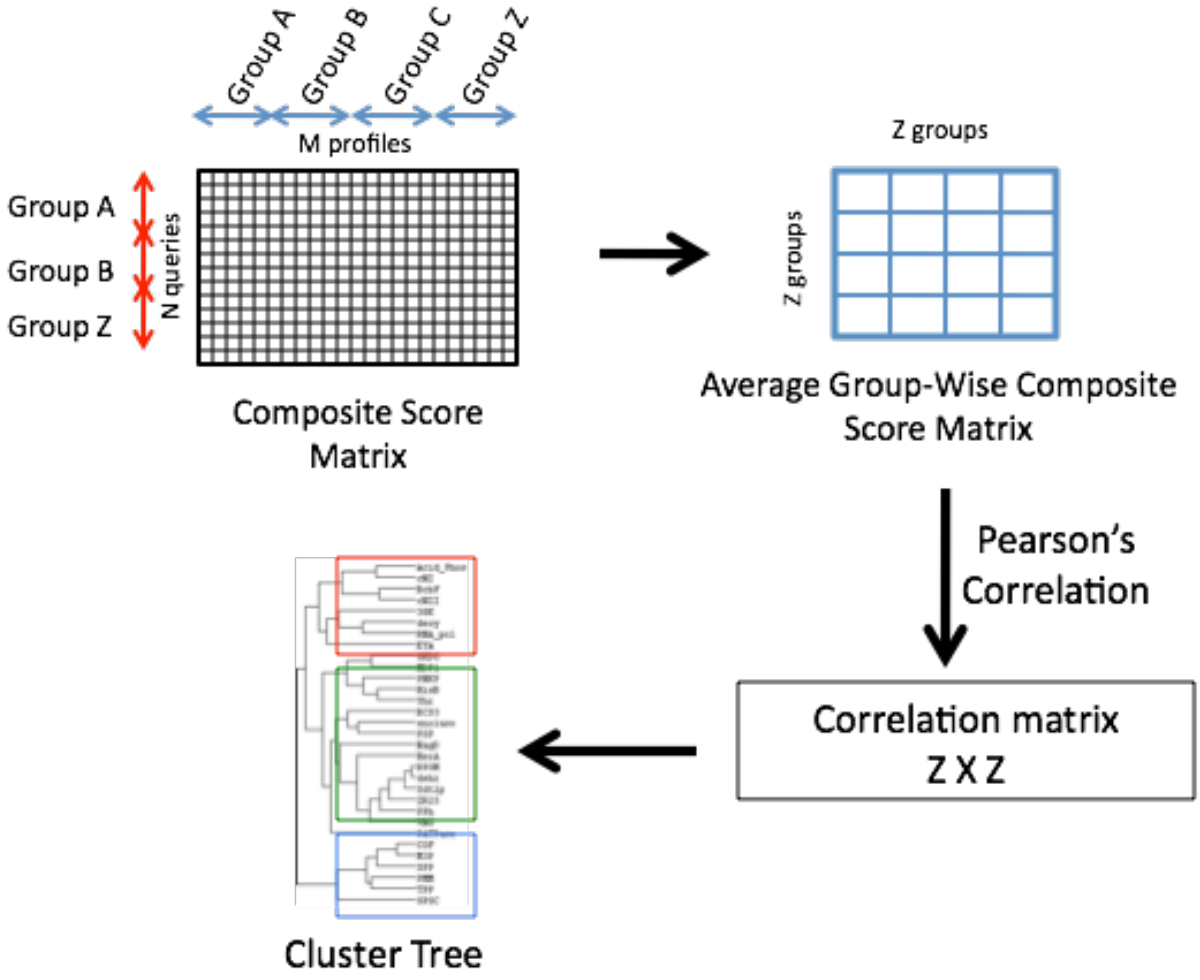


Figure 4: Schematic of GDDA BLAST with hierarchical clustering.

After initial query/profile alignments are generated and compared in the composite score matrix, the HAD enzyme groups' scores are compared in a matrix called the "average group-wise composite score matrix." From this clustering/grouping step a Pearson's correlation generates matrices. Then an algorithm like neighbor joining is then used to generate a phylogram as normal. The result is a phylogenetic tree with preliminary clustering as a rough separation of the families. This reduces the inconsistencies found in the single lineage approach and generates multiple lineages.

This method dealt with the assumption that members of the same group/family did not need to be compared against each other but instead compared as a whole against other groups. Using statistical alignments, a divergent (and convergent) superfamily can sometimes generate stronger relationships between clearly unrelated enzyme families. This hierarchical clustering allowed us to remove some dependence on the purely statistical methods instead using educated guesses. The clustering method resulted in a better tree than the single lineage assumption.

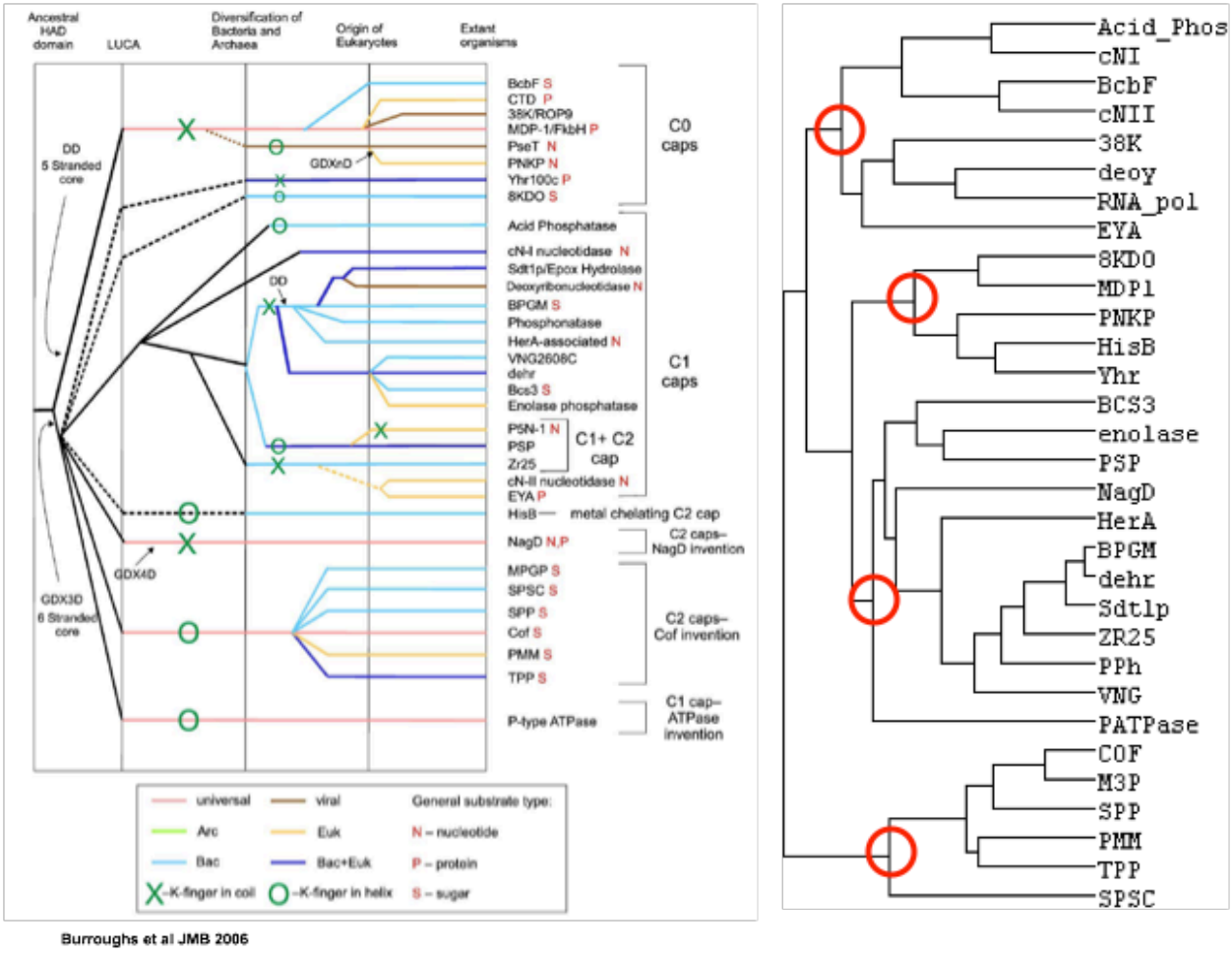
We identified 4lineages and generated a better tree than the single lineage approach. This is similar to the 5 lineages proposed by Burroughs. The direct comparisons of the trees can be seen in Figure 6. The 4 lineages originate from these families: cN-II nucleotidase, 8KDO, SPSC, PATPase/PSP. The origins of these lineages can be seen as red circles in Figure 6. All of these enzyme groups are either of bacterial origin or found universally in all superkingdoms of life. The lineages contain correct speciation patterns and progress from evolutionarily older organisms like bacteria to younger organisms like animals. Most families are monophyletic and the lineages are also robust when sequences are removed. None of these things could be said about the single lineage approach.

While there exist 4 lineages we only investigate one in detail. The lineage we specifically investigate here is that one rooted by the CNII nucleotidase family. This lineage, as seen unrooted in Figure 7, contains the families RNA polymerase A, 38k, deoxyribonucleotidase, Bcbf, Acid Phosphatase, CN-1 nucleotidase, EYA and CN-II nucleotidase. The CN-II nucleotidase family is universally contained in all kingdoms and fittingly serves as the origin for this lineage because of its long branch length. From this root come the EYA lineage which shares the multihelical C1 cap assemblage with the CN-II group and is also found in eukaryotes. The next lineage is the CN-1 nucleotidase family which is identified by a bihelical C1 cap and is found in both bacteria and eukarya. This contradicts the phylogeny proposed by the Burroughs group who suggested that the CN-I and CN-II families converged on nucleotide substrate interaction. Also, the evolution of a

bihelical cap from a more ancestral multihelical cap may seem counterintuitive, but many mutations can occur that result in the deletion of genetic material. Additionally, movement into bacterial genomes is likely via a lateral transfer event as previously described. It is also evolutionarily consistent that the older enzymes families are nucleotide specific and were progenitors to the diversification from an RNA world (8).

From here there is a closely related cluster of families containing the Acid Phosphatases, the archaeal 38K, and the bcbf lineages. All of these families contain rudimentary cap structures with acid phosphatase family being previously grouped with C1 cap containing families by Burroughs et al and found only in bacteria. The Bcbf family similarly is found only is bacteria but contains what was identified as a rudimentary C2 cap instead. Lastly, the 38k group contains a rudimentary C2 cap but is found in archaeal lineages only. Because the closely related families are found primarily in bacteria this is a possible example of some sort of lateral gene transfer.

Still more diverged families include the viral clade of deoxyribonucleotidase, the viral 38k clade and lastly the RNA polymerase clade found in eukaryotes. It is very likely that the archaeal 38k viral was transferred to a basal eukaryotic family and evolved into RNA polymerase which is now contained within many genomes: plants, animals, slime molds, fungi, and kinetoplastids. It is proposed that these very old 38k progenitors gave rise to the previously discussed deoxyribonucleotidases, RNA polymerases, and 38k groups in higher organisms.



Burroughs et al JMB 2006

Figure 5: Comparison of Burroughs et al HAD tree and tree using GDDA BLAST with hierarchical clustering.

Burroughs tree (left) same as figure 1 with 5 distinct lineages. Phylogram of HAD superfamily using GDDA BLAST with additional step of group-wise clustering/comparison resulting in the identification of 4 lineages marked by red circles. No family is found in more than one lineage and most are monophyletic within those lineages. This phylogenetic tree was constructed from purely statistical considerations.

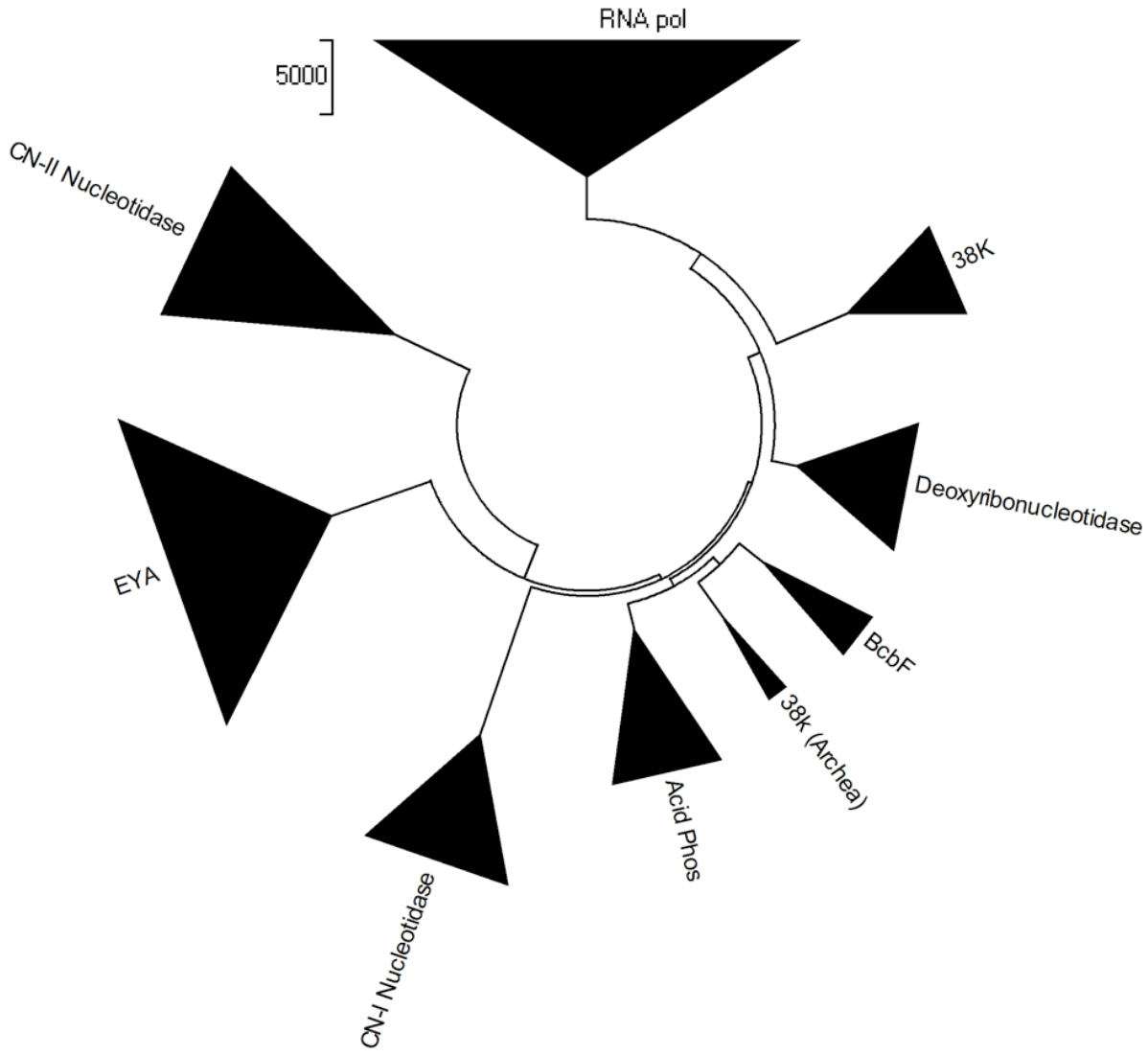


Figure 6: Lineage with CN-II Nucleotidase as the ancestral group.

Unrooted tree of the first lineage from figure 6 from the top containing the groups: CN-II Nucleotidase, EYA, CN-I Nucleotidase, Acid Phosp, 38K, Bcbf, deoxyribonucleotidase and RNA pol. Size of the family denoted by the size of the black triangle at branch end. Evolutionary distance is shown by the length of the branch. The 38k group is the only one lacking monophyly.

Discussion

We have used the HAD superfamily of enzymes as a case study to determine the effectiveness of GDDA BLAST in resolving highly divergent protein families. The divergent nature of this family makes it a good test for GDDA BLAST which was designed with provisions and modifications to identify statistically significant relationships between sequences with less than 25% sequence similarity. Other MSA algorithms have failed with a broad family with little sequence similarity like HAD.

We also showed that convergent structural and functional elements within a divergent enzyme family can be resolved using this methodology. Previous studies performed by Burroughs et al using the HAD superfamily had to use mostly non-statistical methods and grouped the various families based on biochemical and structural similarities, like the grouping based on similar cap motifs. The previous research group identified 5 distinct lineages and they found a good bit of convergence both in structural elaborations called caps and substrates, like nucleotides. Convergent lineages can be particularly difficult to resolve; however, with GDDA BLAST coupled with a hierarchical clustering method using a Pearson correlation more evolutionarily accurate relationships were generated.

We first rejected the idea that there could exist a single lineage unifying this broad and diverse super family. Our attempt to produce a single lineage led to many evolutionary contradictions in speciation patterns and monophyly. Related families were spread throughout the phylogenetic tree. These symptoms are common in divergent lineages and can be expected in convergent lineages. Because the previous research group identified multiple lineages we used this first single lineage approach as a sort of null hypothesis.

From here we used a better clustering method and were able to generate a tree with monophyletic clades. We used a group-composite score to cluster the families before proceeding with a query by query analysis within those lineages. We identified 4 distinct lineages as opposed to the 5 identified by Burroughs. The tree contained more evolutionarily accurate lineages than the first approach. The speciation patterns within the tree generally progressed from a bacterial or universal ancestral group to higher species which is expected but in contrast to the single lineage. The origin families of each lineage were either universal or found in bacteria. While there were mostly monophyletic groups there are a few examples of paraphyletic groups as was discussed previously with the 38K family, the older groups being archaeal and possibly progenitors.

Despite our positive results, there remains room for improvement. In future analyses a more well-defined tree can be established by inclusion of additional HAD sequences and even additional clustering. More sequences naturally add to the robustness of a tree and can be found in different databases. Also, because of the effectiveness of the Pearson correlation in producing distinct clusters it can be suggested that further clustering even within the lineages could resolve some of the families and groups lacking monophyly. A combination of the GDDA BLAST statistical method and some structural analysis method could provide a better phylogenetic tree. We did discuss the GDDA BLAST tree in light of the Burroughs research but we did not generate a tree based on this work.

While previous analyses using GDDA BLAST have concerned with rapidly evolving divergent families it is also apparent that coupled with a clustering method like Pearson's correlation the methodology can resolve convergent elements. In a broader context the ability to understand and generate evolutionary relationships allows us to untangle the questions about the origin of life and the rapid diversification of species and concomitant enzymatic machinery.

References

1. Burroughs, Maxwell A., Allen, Karen N., Dunaway-Mariano, Debra., Aravind, L. Evolutionary Genomics of the HAD Superfamily: Understanding the Structural Adaptations and Catalytic Diversity in a Superfamily of Phosphoesterases and Allied Enzymes. Journal of Molecular Biology (2006) 361, 1003-1034.
2. Silverthorn, Dee Unglaub. Human Physiology: An Integrated Approach. Fourth Edition. Pearson Educational Inc. 2009.
3. Conway Morris, Simon. *Life's solution: inevitable humans in a lonely universe*, Cambridge, UK: Cambridge University Press(2005).
4. Poling, Jeff. “What is Parsimony Anyway.” April 8, 2010. <http://www.dinosauria.com/jdp/misc/parsimony.htm>
5. Dae Ko, Kyung., Hong, Yoojin., Chang, Gue Su., Bharwaj, Guarav., Van Rossum, Damien B., Patterson, Randen L. Phylogenetic Profiles as a Unified Framework for Measuring Protein Structure, Function and Evolution. Physics Archives. June 2008
6. Bowler, Peter J. *Evolution: The History of an Idea*. University of California Press. (2003).
7. Doolittle, Ford W. (February 2000). "Uprooting the Tree of Life". *Scientific American*: 72–7.
8. Gilbert, Walter. The RNA World. *Nature* (February 1986). **319**: 618.

Appendix A
HAD Subgroups and families as per Burroughs (1).

Within the C0 cap assemblage they described two main lineages. The C0 caps can be split by either having a five or six stranded core. The least diverged five stranded enzymes include the MDP-1/FkbH families, which are unified by their biochemical similarities. This group is present in all three superkingdoms of life suggesting their presence in a LUCA as one of the oldest lineages. From this lineage Burroughs anchors the emergence of the 38K/ROP9, CTD, and the Bcbf families. The 38K/ROP9 and Bcbf families are small and specific to baculoviruses, apicomplexa and proteobacteria/proteobacteria viruses respectively. Additionally, the enzymes have different compartmental locations and functions. The Bcbf family has extensive structural modifications but lacks a definitive cap resulting in it being classified as C0. The CTD family is a large group of enzymes required for dephosphorylation of serine residues in the catalytic subunit of RNA polymerase. From there the family broadens including enzymes from many compartments. The most unrelated five stranded C0 cap family is the PNPK family. The PNPK(Polynucleotide kinase phosphatase) family includes enzymes which play a role in RNA and DNA repair by removing 3' phosphate groups. This family is present in both bacteria and eukarya with additional subgroups in animals.

The six stranded C0 cap assemblage lineages include the 8KDO, and Yhr100c lineages. These groups are thought to be more closely related to the other lineages of C1 and C2. In fact the Yhr100c family which functions as protein phosphatases during cell division has a great amount of core sequence similarity with more elaborate C2 families like NagD. Perhaps it served as an ancestral group. The 8KDO family is involved in the biosynthesis of polysaccharide chains. However, beyond this little is known about this C0 family and there do not appear to be many structurally or functionally related families.

The C1 cap assemblage is separated into two independently evolved lineages. As previously discussed the C1 containing caps are classified by structural motifs as either alpha-helical or P-type ATPase. Further, the alpha-helical cap lineage can be broken down into three groups distinguished by cap morphologies: bihelical, tetrahelical and multihelical. The tetrahelical assemblage includes the majority of C1 cap families.

The bihelical cap is seen in the acid phosphatases and the cN-1 nucleotidase family. However, there exists little evidence linking the two groups. Instead, bihelical caps are considered to be a simple ancestral structure and the two families split concurrently.

Tetrahelical assemblages also can be further classified by specific structural motifs. The first group classified by Burroughs is called the motif IV DD assemblage. This group includes the deoxyribonucleotidase, phosphonatase, STDI-epoxide hydrolase, HerA-associated, and beta-phosphoglucomutase(BPMG) families. Within this classification there are a few interesting relationships. For instance, according to the research group, the eukaryotic forms of the deoxyribonucleotidases do not group together in phylogenetic analysis suggesting introduction by phage/viral source. The second tetrahelical cap grouping includes the dehalogenase (dehr) and the enolase-phosphatase families. This assemblage is unified by specific sequence similarities. The dehr family while widespread is seemingly segmented and inconsistently represented in many lineages of life. The PSP, and P5N-1 families comprise the last tetrahelical C1 cap subgroup. A small ‘secondary’ C2 cap is the unifying structure in this subgroup.

The multi-helical cap assemblage is seen in three families of enzymes. This includes the cN-II a family of nucleotidases that Burroughs recognizes as having convergently evolved their catalytic ability. This ability is also found in cN-I, a bihelical assemblage. The other two families include the EYA(eyes absent) enzyme/transcription factor family, and the Zr25 family.

The P-type ATPase family having independently evolved a C1 cap is a basal lineage. Assuming convergence this whole family is completely unrelated to any alpha-helical

family. This family is ubiquitous throughout all three superkingdoms but there are instances of apparent lateral gene transfer according to Burroughs.

The C2 cap assemblage is broken into three lineages which all are thought to arise independently. The first lineage contains the HisB family. This family of enzymes displays the simplest version of a C2 cap and a very unique motif. They are specific to bacteria as a metal chelating cap. The NagD family contains a C2 unrelated to any other cap in the HAD superfamily. All members of this family contain a conserved aspartate in the C2 cap which likely acts as a substrate recognition factor. The NagD family is found throughout all kingdoms of life.

The last C2 lineage called the Cof phosphatase assemblage includes six diverse families. The Cof family which spans all kingdoms of life is the archetypal enzyme of this lineage. The subgroups in this lineage include enzymes that are diverse yet share a common sheet topology. The trehalose phosphate phosphatase(TPP) family catalyzes the formation of trehalose from other mono and diphosphates. The mannosyl-3-phosphoglycerate phosphatase (MPGP) family is found in both archaea and bacteria catalyzing the formation of phosphoglycerate. The phosphomannomutase (PMM) family of enzymes is specific to eukaryotes. The sucrose phosphate synthase C-terminal domain family(SPSC) and the sucrose phosphate phosphatase (SPP) family are the final two members of this large family and they are closely related in function as well as structure. Both of these families contain conserved C-terminal domains. .

VITA

Evan J. Smith
Bachelor of Science in Biology
The Pennsylvania State University
GPA: 3.87
Course Background: Completed biology coursework with a focus on neuroscience, as well as language coursework: proficient in both German and Spanish.

*Relevant Employment
8/08-present Lab Assistant and Technician
The Randen Patterson Lab, University Park, PA
-Performed and designed experiments identifying viral protein binding domains from designing primers to running yeast 2 hybrid(bait and fish) assays. Bulk of work was concerned with calcium signaling pathways with an Emphasis on TRPC5
-Served to upkeep lab stocks was responsible for lab orders/receipts of up to $1000.
-Worked in bioinformatics with our novel multiple sequence alignment program, GDDA BLAST, generating and analyzing phylogenetic trees.

-Listed as a coauthor: "Robust Phylogenetic Inferences in the 'Twilight-Zone' of Sequence Similarity," by first author Gaurav Bhardwaj and corresponding author Randen Patterson. In Review for Science Magazine.

4/08-7/08 Lab Assistant
The Claude DePamphillis Lab, University Park, PA
-Learned basic lab preparatory skills: creating solutions, gels, stocking, gene databasing.
-Gained experience in RNA and DNA isolations as well as PCR, RT-PCR.
-Worked in collaboration with researchers on generating phylogenetic trees of basal angiosperms based on single copy gene divergence in specific gene families

*Other Employment
6/07-8/07 Server
The Olive Garden, North Wales, PA
-Interacted daily with a diverse clientele, with an emphasis on quality service, efficiency.
-Was responsible for up to $2,000 in guest receipts per week.
-Dealt with a variety of personalities, wants and needs on the fly.

6/04-8/07 Electrical Assistant
Reliance Electric Inc, Ambler, PA
-Learned basic and advanced wiring in residential settings.
-Worked in team of up to 4 electricians

*Activities
Active member of Sigma Pi fraternity
Volunteered for Chestnut Hill Hospital
Volunteered for Food for the Hungry in Guatemala
Graduate of "La Escuela Minerva" in Xelaju, Guatemala