Published in final edited form as:
Proteins. 2014 October ; 82(10): 2597–2608. doi:10.1002/prot.24624.

Functional Evolution of PLP-dependent Enzymes based on Active-Site Structural Similarities

Jonathan Catazaro[1], **Adam Caprez**[2], **Ashu Guru**[2], **David Swanson**[2], and **Robert Powers**[1,*]

[1]Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588-0304 USA

[2]Holland Computing Center, University of Nebraska-Lincoln, Lincoln, NE 68588-0150 USA

Abstract

Families of distantly related proteins typically have very low sequence identity, which hinders evolutionary analysis and functional annotation. Slowly evolving features of proteins, such as an active site, are therefore valuable for annotating putative and distantly related proteins. To date, a complete evolutionary analysis of the functional relationship of an entire enzyme family based on active-site structural similarities has not yet been undertaken. Pyridoxal-5'-phosphate (PLP) dependent enzymes are primordial enzymes that diversified in the last universal ancestor. Using the Comparison of Protein Active Site Structures (CPASS) software and database, we show that the active site structures of PLP-dependent enzymes can be used to infer evolutionary relationships based on functional similarity. The enzymes successfully clustered together based on substrate specificity, function, and three-dimensional fold. This study demonstrates the value of using active site structures for functional evolutionary analysis and the effectiveness of CPASS.

Keywords

Functional Evolution; PLP-dependent Enzymes; CPASS; ligand binding sites

Introduction

Sequence comparison methods and structural homology models are often used to imply evolutionary relationships between proteins in the same family or of the same function.[1,2] These analyses are additionally used to infer the function of uncharacterized proteins.[3,4] Enhancing our understanding of protein evolution furthers our insights into cellular processes, enables the discovery of drug targets and disease markers, while validating or establishing phylogenies of organisms, among numerous other beneficial impacts.[5] However, evolutionary studies are limited for distantly related proteins in which sequence and structural similarity may not be easily deduced. Sequence comparison methods assume a constant rate of amino acid substitution and are not suitable for all phylogenetic reconstructions of ancestral sequences. Correspondingly, for sequences with minimal identity local conformations can vary greatly and may inhibit structural studies.[6]

*To whom correspondence should be addressed, Robert Powers, University of Nebraska-Lincoln, Department of Chemistry, 722 Hamilton Hall, Lincoln, NE 68588-0304, rpowers3@unl.edu, Phone: (402) 472-3039, Fax: (402) 472-9402.

It is understood that although the rate of sequence evolution is highly variable, residues essential to function and active site chemistry will remain highly conserved over time.[7] Slight changes in the position and identity of an active site residue can lead to changes in both substrate and reaction specificity.[1,8] In fact, it has been proposed that catalytic residues maintain their position and identity longer than the configuration of some secondary structures.[9] The importance of these residues has spawned many computational approaches to assist with the prediction of active site residues,[10–12] the functional annotation of proteins based on active sites,[13–15] as well as gene functional annotation.[16,17] Additionally, the evolution of homologous proteins sharing low sequence identities has been inferred using active site structure geometries.[9,18,19] These observations are consistent with the premise that molecular evolution and functional evolution proceed through distinct processes. Molecular evolution is driven by random mutational events where maintaining function is paramount.[1,2] Conversely, protein functional evolution requires a new gene through duplication, acquisition or creation that enables protein functional drift while avoiding a negative impact on cell fitness.[20–22] Correspondingly, analyzing functional evolution may require a different approach from the global sequence comparisons routinely used for molecular evolution.[1,2] In this manner, active site structures may be a suitable alternative for inferring an evolutionary relationship based on function.[23]

Some of the most diverse and versatile classes of enzymes are those that utilize pyridoxal-5'-phosphate (PLP) as a cofactor.[24] These enzymes are found in five of the six classes defined by the Enzyme Commission and catalyze a multitude of different reactions.[25] The diversity of the reactions can be attributed to the electron sink properties of the cofactor[26] and the stereochemical restraints in the active site of the enzyme.[27–29] PLP-dependent enzymes acting on amino acid substrates can be separated into four families of paralogous proteins annotated as fold-types I-IV.[26,30] Interestingly, each of the fold-types are highly promiscuous in terms of function, and multiple examples of convergent evolution have been discovered.[24] These observations have led to the hypothesis that the four lineages of PLP-dependent enzymes developed in the last common universal ancestor.

Extensive work has been done with sequence-based methods to establish a phylogenetic relationship among PLP-dependent enzymes.[24,31,32] These studies, however, have been limited to small subgroups due to the low sequence identity within each fold-type and across all PLP-dependent enzymes. Remarkably, the functions of these enzymes are highly dependent on the structure of their active sites. Although these enzymes lend themselves to an evolutionary study based on active site structures, one has not yet been undertaken.

Here, we use a novel method to compare the active site structures of 204 PLP-dependent enzymes from the four different fold-types that are specifically involved in amino acid metabolism. The Comparison of Protein Active Site Structures (CPASS) software and database was developed to aid in the functional annotation of uncharacterized proteins by comparing ligand defined active sites.[15,33] As a sensitive measure of active site geometry, it was proposed that CPASS could also be used to model families of divergent proteins. In this work we present the functional evolution of PLP-dependent enzymes based on active site structures. To the authors' knowledge, this is the first study of its kind that models the functional evolution of a superfamily of proteins by active site comparisons, and the first to

model all PLP-dependent fold-types together on one network. Results were compared to previous studies using both sequence and structure-guided methods. We find that the active site comparisons successfully cluster the enzymes based on both function and three-dimensional fold-type.

Materials and Methods

Active Site Structure Comparison

Three-dimensional structures with ligand-defined active sites were gathered from the Protein Data Bank on May 9, 2014.[34] An all versus all comparison was carried out between the active site structures using the Comparison of Protein Active Site Structures (CPASS) software. CPASS finds the optimum structural alignment between two active sites by maximizing a scoring function based on structure (Cα,Cβ and Ligand RMSD, Solvent accessible surface area) and sequence (BLOSUM62) similarity.[35,36] The scoring function is the summation of an ideal pairing of residues between each of the active sites rather than a summation of all possible pairings. Therefore, it is possible that active site residues remain unpaired during analysis. Since the CPASS algorithm is dependent upon the size of the active site query, an asymmetrical similarity matrix was generated from the all versus all comparison. The lower score of the two pairwise comparisons was extracted and converted to a distance by subtracting from 100% similarity. From the resulting distance matrix, only the PLP bound proteins that acted on amino acid substrates were extracted. Redundancies in the PLP-distance matrix were removed. A total of 204 X-ray crystal structures and corresponding active sites were analyzed in this study.

Protein Annotation

A total of 204 PLP-dependent proteins were functionally annotated with Enzyme Commission (EC) numbers using the UniProt database[37] and in-house software. Enzymes that did not have an EC number in the UniProt database were assigned manually using the PDB. For enzymes with more than one number or classification, only the first was taken. Species, phylogenetic kingdom, and fold-type annotations were done manually using the PDB and UniProt databases as well as literature sources.

Phylogenetic Analysis

A phylogenetic network using all of the PLP-dependent enzymes was generated with Splitstree4[38] using the Neighbor-Net algorithm.[39] The network represents a deterministic model of the distances obtained from the CPASS algorithm and software. Identification of the four fold-type clusters was accomplished by visual inspection of the network. Each of the clusters is continuous and does not contain enzymes that are not of the same fold-type. Phylogenetic trees were created for each of the fold-types using Dendroscope,[40] CPASS distances, and the Neighbor-join algorithm. Neighbor joining is an agglomerative clustering method in which a tree is built from a distance matrix. Briefly, the two closest taxa in the distance matrix are joined to create a new node. Distances from each of the remaining taxa to the new node are recalculated and the matrix is updated. As this process is repeated, the distance matrix shrinks until only a pair of nodes remains.

Sequence Identity

The PDB IDs for the 204 PLP-dependent enzymes in this study were submitted to the UniProt database for primary sequence retrieval. The sequences were subsequently submitted to Clustal Omega[41] for multiple sequence alignment (MSA) using the default parameters. The MSA and percent identity matrix were retrieved upon completion and visualized in the R software environment. A histogram was generated with intervals of 0.2 percent identity and represents the entire, non-condensed percent identity matrix. A heat map of the entire matrix was also generated with a white to red gradient corresponding to low and high percent identity. The ordering of the sequences in the heat map corresponds to the ordering in Table S1.

Structure Alignment

The 204 PLP-dependent enzymes were further subjected to an all versus all structural alignment using TM-align.[42] Results from the pairwise comparisons were organized into a matrix and were visualized in the R software environment. A histogram was generated with TM-align score intervals of 0.001 and represents the entire symmetrical matrix. Correspondingly, a heat map of the matrix was generated with a white to red gradient representing low and high TM-align scores, respectively. Ordering of the heat map is the same as the sequence identity ordering and follows the order in Table S1.

Comparison of Type-I Transaminase Active Sites

From the 204 PLP-dependent enzymes in the CPASS dataset, 4 enzymes were selected to elucidate the active site differences that encode substrate specificity. An acetylornithine (ACO) transaminase (PDB ID: 2ORD), a γ-aminobutyric acid (GABA) transaminase (PDB ID: 1OHW), an aspartate (Asp) transaminase (PDB ID: 1ARG), and a tyrosine (Tyr) transaminase (PDB ID: 3DYD) were selected from functional clusters 1, 1, 4b, and 4a, respectively. Active site residues for each of the enzymes were selected based on their proximity to the PLP ligand (6 Å) as defined by CPASS. Active sites were overlaid in Chimera[43] and visually inspected for the changes resulting in different substrate specificity. Additional transaminases with the same function as the four in question were subjected to individual multiple sequence alignments in Clustal Omega. The additional enzymes were manually selected from the PDB and were not present in our original dataset because a bound ligand was not present in the structure.

Results

Phylogenetic Network Analysis

Phylogenetic analyses were carried out on 204 PLP-dependent enzymes with bound cofactor (Table S1) from each of the four fold-types previously defined in the literature.[26] It is well known that each of the fold-types originated from separate evolutionary lineages. Using sequence-based phylogenetic methods,[24] each of the lineages has been visualized with a molecular evolutionary tree structure. However, due to their low sequence identity visualizing all of the fold-types together has not yet been accomplished. Using active site similarity comparisons, it is possible to display all of the PLP-dependent enzymes in one

phylogenetic network that demonstrates an evolutionary relationship based on function (Figure 1). Type-I enzymes make up the largest portion of the PLP enzymes and encompasses most of the enzymes that act at the Cα position of the substrate. Type-II enzymes catalyze reactions at the Cα and Cβ positions and contain fewer members compared to Type-I. The final two folds, Types III and IV, have the fewest members of PLP-dependent enzymes. The function-based phylogenetic network suggests that each of the fold-types originated from an independent lineage. The main branches of the network all spawn from the center, indicating a universal organism contained ancestral enzymes for each of the fold-types. Also evident are clear separations within each of the groups that form function-specific clusters (numbered 1 through 13). The functional network also highlights the advantages of using an enzyme's active site to elucidate an evolutionary relationship. Each of the PLP-dependent fold-types has a defined active site geometry, and within each type, each function-specific cluster has a unique variation in the active-site geometry. However, the functional network only gives a general overview of the relatedness between each of the functions in the fold-types and between the fold-types. To make valid inferences of functional evolutions, further analysis is still necessary.

Phylogenetic Analysis of Type-I Enzymes

Active site similarity scores from the enzymes in the Type-I fold-type were analyzed in detail to find a phylogenetic relationship among functional clusters 1–7 (Figure 2). The dominating function of fold-type I involves transaminase activity (EC 2.6.1), which is found in clusters 1, 2, 4 and 7. Clusters 3, 5 and 6 are primarily comprised of carbon-sulfur lyases (EC 4.4.1), decarboxylases (EC 4.1.1), and glycine hydroxymethyltransferases (EC 2.1.2.1), respectively. Although clusters 1, 2, 4 and 7 share a broad transaminase activity, these clusters are punctuated in the phylogenetic tree by the remaining functional clusters. This observation suggests multiple Type-I enzymes existed in a common ancestor with specialized function. Conversely, a continuous grouping of all the transaminases would have implied the existence of a single transaminase with broad substrate specificity. Previous studies have recognized similar divisions within the Type-I fold using sequence homology methods and structural alignment tools; and came to the same conclusion.[44–46] The punctuation in the active site similarity tree highlights the sensitivity of CPASS for comparing residues and geometries of active sites; and supports its use as a phylogenetic tool.

There are multiple examples of bacterial and eukaryotic species with homologous enzymes within the Type-I functional clusters (Table S1). A general idea of the extent of enzyme divergence from a universal ancestor can be elucidated based on how these homologous enzymes group within the Type-I functional clusters. In cluster 3, the methionine gamma-lyases (EC 4.4.1.11) belonging to eukaryotic *T. vaginalis* and bacterial *P. putida* were found to be closely related. Directly following in the phylogentic tree are the cystathionine gamma-lyases (EC 4.4.1.1) for both eukaryotes and bacteria. This indicates that a methionine gamma-lyase and a cystathionine gamma-lyase existed in a universal ancestor. Similar examples are also found in clusters 2, 6, and 7 for phosphoserine transaminases (EC 2.6.1.52), kynureninases (EC 3.7.1.3) and glycine hydroxymethyltransferases (EC 2.1.2.1), respectively.

Aspartate transaminases are present in both clusters 4a and 4b, but a clear separation between these two clusters is apparent in the phylogenetic tree shown in Figure 1. This is due to the fact that eukaryotic aspartate transaminases are only found in cluster 4b, while bacterial aspartate transaminases are found in both clusters 4a and 4b. The eukaryotic enzymes in cluster 4a are comprised of carboxylate synthases (EC 4.4.1) from the kingdom *Plantae*, and tyrosine and alanine transaminases (EC 2.6.1.5 and 2.6.1.2) from the kingdom *Animalia*. This finding does not support the idea that all Type-I enzymes became highly specialized in the last universal ancestor. Instead, our results suggest that a promiscuous aspartate transaminase was passed on to emergent species and, through subsequent speciation and gene duplication events, the enzymes in clusters 4a and 4b became more functionally specialized and more substrate specific.

Phylogenetic Analysis of Type-II Enzymes

A detailed phylogenetic analysis was also carried out on the Type-II PLP-dependent enzymes using CPASS similarity scores and the Neighbor-Join algorithm (Figure 3a). Three functional clusters based on active site similarity were identified for this fold-type and are labeled as clusters 8, 9 and 10. Enzymes catalyzing deaminase reactions on 1-aminocyclopropane-1-carboxylate (EC 3.5.99.7) are found exclusively in cluster 8. These enzymes belong to species in archaeal and eukaryal domains indicating that their function was specialized in a common ancestor. The same observation was made for the enzymes in cluster 10. Threonine synthase (EC 4.2.3.1), tryptophan synthase (EC 4.2.1.20), and serine dehydratase/threonine deaminase (EC 4.3.1.17/4.3.1.19) enzymes from archaeal, bacterial, and eukaryal species cluster by function rather than by domain. Cluster 9 contains cysteine synthases (EC 2.5.1.47) from the bacterial, archaeal, and eukaryal domains and cystathionine β-synthases (EC 4.2.1.22) from the eukaryal domain. The clustering of the cysteine synthases by function rather than by domain once again indicates specialization in a common ancestor; however, an insufficient number of cystathionine β-synthases are represented in the dataset to elucidate its functional evolution.

Phylogenetic Analysis of Type-III Enzymes

Phylogenetic analysis was also undertaken to explain the emergence of enzymes belonging to the Type-III fold-type (Figure 3b). Clusters 11 and 12 were found to be discernibly different from one another in the phylogenetic network (Figure 1) and, correspondingly, separate from one another on the phylogenetic tree created with the Neighbor-Join algorithm. Alanine racemases (EC 5.1.1.1) exclusively populate functional cluster 12; whereas, functional cluster 11 is made up of three comparable decarboxylases. Unlike most of the enzymes in fold-types I and II, it does not appear the Type-III enzymes specialized in the last common ancestor. In our dataset, diaminopimelate decarboxylases (DAPDC, EC 4.1.1.20) and arginine decarboxylases (ADC, EC 4.1.1.19) from bacterial and archaeal species are present while ornithine decarboxylases (OCD, EC 4.1.1.17) appear only from eukaryal species. However, function specific clustering of the enzymes in cluster 11 is not observed. It is likely that a decarboxylase with a very generic active site existed in a common ancestor due to the kingdom specific characteristics of the current enzymes.[47,48] Additionally, ODCs for bacterial species can be found in functional cluster 5 of the Type-I enzymes. This example of convergent evolution further supports the idea of a broad-based

decarboxylase belonging to the Type-III fold. A viral ADC (PDB ID: 2NV9) is also present within functional cluster 11, but does not cluster with the bacterial ADCs. It was previously hypothesized that this enzyme was acquired from a bacterial ornithine decarboxylase and subsequently mutated toward arginine specificity.[49] CPASS analysis supports this proposal as the enzyme clusters with eukaryal ODCs and a bacterial DAPDC instead of the bacterial ADCs. Specifically, this finding represents an example of convergent evolution within the Type-III fold.

Phylogenetic Analysis of Type-IV Enzymes

Type-IV enzymes occupy the smallest portion of all PLP-dependent enzymes and form only one functional cluster (Figure 3c). Cluster 13 contains aminodeoxychorismate lyases (EC 4.1.3.38), branched chain amino acid transaminases (EC 2.6.1.42) and a D-amino acid transaminase (EC 2.6.1.21). Similar to the other three fold-types, the function specific enzymes in this family group together. However, because there is only one D-amino acid transaminase represented in the dataset, the enzyme is grouped with the branched chain amino acid transaminases. Interestingly, D-amino acid transaminases share similar active site characteristics with the Type-I transaminases.[50] The geometry and residue identity is different enough, however, to distinguish between Types I and IV enzymes without having to look at the three-dimensional structure.

Sequence Identity and Structural Similarity

A heat map and corresponding histogram was generated from a multiple sequence alignment (MSA) of the 204 PLP-dependent enzymes (Figure 4a–b). Values above 20% identity belong to self-comparisons and comparisons of closely related homologous enzymes within the same organism. This can be seen visually as the diagonal line in the corresponding heat map. As expected, the histogram shows that the majority of the pairwise comparisons are well-below 20% sequence identity. This is consistent with prior observations that PLP-dependent enzymes have very low sequence identity.[24,31,32] Importantly, establishing a protein function based solely on sequence homology significantly decreases in reliability as sequence identity falls below 50%.[51] Presumably, a similar high level of uncertainty in establishing a functional evolutionary relationship would occur for proteins with sequence identity well-below 20%.

The 3D structures of PLP-dependent enzymes were also analyzed to ascertain the reliability of using structures of distantly related proteins to generate a functional evolutionary model. Figure 5a–b illustrates the resulting histogram and heat map of the pairwise structural alignments of the 204 PLP-dependent enzymes. The histogram shows a bimodal distribution of TM-align scores. TM-align scores greater than 0.5 are consistent with two proteins sharing the same CATH or SCOP fold. Correspondingly, the ~0.6 TM-align scores occur for PLP-dependent enzymes within the same fold-type. Conversely, the lower TM-align scores result from comparisons between fold-types. The associated heat map of TM-align scores clarifies this point; high TM-align scores are clustered by fold-type. Although structural alignments are able to separate the fold-types and determine proper PLP-dependent enzyme membership within a fold-type, the global structural information is not sensitive enough to cluster the PLP-dependent enzymes based on function. In essence, the global structural

similarity between enzymes within a fold-type over-whelms any subtle differences related to functional divergence. Similarly, the large structural differences between fold-types negate any local similarities that may be present due to a common functional ancestor.

Comparison of Type-I Transaminase Active Sites

The evolutionary network based on active-site similarity depicted in Figure 1 identified acetylornithine (ACO) transaminase (EC 2.6.1.11, PDB ID: 2ORD) and a γ-aminobutyric acid (GABA) transaminase (EC 2.6.1.19, PDB ID: 1OHW) as nearest neighbors in the network. Both enzymes correspond to fold-type I and are located in cluster 1, but the enzymes have different EC numbers and substrate specificity. Thus, the active sites were compared in order to reveal structural change(s) that contribute to the different substrate specificities that led to a functional divergence (Figures 6a–b). The goal was to further explore the validity of the phylogenetic network based on CPASS similarity scores to provide insights on the pathway of functional evolution. Further pairwise comparisons were also made with enzymes from functional clusters 4a and 4b in order to show a progressive, step-wise transition of one active site structure to the next. An aspartate (Asp) transaminase belonging to functional cluster 4b (EC 2.6.1.1, PDB ID: 1ARG) was overlaid with the GABA transaminase to show the active site structural change between functional clusters (Figures 6b, d). The final comparison between the Asp transaminase and a tyrosine (Tyr) transaminase (EC 2.6.1.5, PDB ID: 3DYD) was undertaken to show the active site differences between functional clusters 4b and 4a as well as the structural changes necessary to accommodate the larger substrate (Figures 6c–d).

Discussion

Sequence and structural homology models are typically used to infer descent from a common ancestor – *molecular evolution*. Many examples of protein evolution have been constructed based solely on sequence information. The advent of computational algorithms based on probability theory has provided a powerful means for inferring sequence homology and evolutionary relationships.[52,53] However, these methods encounter difficulties when dealing with very divergent sequences. At low sequence identity, the incorporation of gaps for a suitable alignment may produce results that are not genuine.[54] The spatial arrangement of local structures begins to define the function of a protein and represents an important, but distinct, evolutionary signal – *functional evolution*.[23] Several methods have been developed that use structural templates to help guide and ultimately improve the alignment of divergent sequences.[55,56] Conversely, when both sequence and structure are divergent these methods become less reliable.[51] Critically, while molecular evolution and functional evolution are related, they proceed through distinctly different mechanisms.[1,2,20] Correspondingly, different approaches are needed to measure the evolution of homologous proteins across multiple species compared to investigating the evolutionary relationship of various protein functions. Thus, instead of applying global sequence similarities to infer an evolutionary relationship between distantly related proteins, investigating functional evolution requires a targeted approach by analyzing the sequence and structural characteristics of functional epitopes or protein active-sites.

The CPASS software and database was initially designed to functionally annotate novel proteins based on ligand defined active site structures and to act as a complement to drug discovery programs.[15,33] CPASS finds the optimal sequence and structural alignment between two active sites, exclusive of sequence connectivity, and returns a similarity score. It was found that CPASS could effectively group and annotate proteins that bound the same ligand based on specific function.[15] Further analysis of the results led to the hypothesis that CPASS may also infer an evolutionary relationship based on function since active site structure and geometry is primarily functionally dependent. CPASS does, however, have an inherent disadvantage since it requires a 3D structure of the enzyme with a bound ligand.

PLP-dependent enzymes encompass a broad-range of catalytic functions and represent a significant percentage of an organism's genome.[57] Correspondingly, PLP-dependent enzymes are a common target for investigating protein evolution.[24,31,32] Complicating these analyses is the fact that PLP-dependent enzymes suffer from low sequence identity within and between the four different fold-types.[45,58] For these very distantly related proteins, any sequence similarity that still remains will only be found in the regions of the proteins that evolve at a significantly slower rate (functionally important) rather than over the entire sequence length. The incorporation of additional sequence information, which is not associated with these functionally conserved regions, is likely to negatively impact a global analysis. Instead, CPASS presents an alternative approach to investigate the functional evolution of PLP-dependent enzymes by restricting the analysis to the conserved, functionally relevant active-site sequence and structure. A total of 204 PLP-dependent enzymes had a structure deposited in the RCSB PDB[34] that contained a bound PLP and could be used for a CPASS analysis of functional evolution.

The results from the CPASS analysis of the PLP-dependent enzymes agree with previous findings that the enzymes developed from four different fold-types in the last universal ancestor (Figure 1). CPASS further clustered the enzymes together by specific function within each of the four fold-types. Again, this is consistent with the existing functional annotation for the PLP-dependent enzymes. It is critical to point out that the CPASS analysis did not utilize the fold-type classification or the available functional annotation to generate the phylogenetic network depicted in Figure 1. The network was generated strictly based on the CPASS active-site similarity scores. The fold-type classification and functional cluster labels were simply added to the network after the fact. Correspondingly, these results demonstrate the reliability of using enzyme active sites for comparing distantly related proteins to infer an evolutionary and functional relationship. Conversely, it was not possible to generate a similar function-based network for PLP-dependent enzymes for the four fold types using either global sequence or structure similarity. A network generated from sequence alignment data results in an essentially random, nonsensical map because of the extremely low sequence alignment within and between the fold-types (Figure S1). Similarly, a network generated from structure homology data groups the PLP-dependent enzymes into the four fold-types, which of course is not surprising since the fold types were defined by structure homology (Figure S2). But, any further refinement into functional clusters or any association between the four fold-types is, again, essentially random and irrelevant.

In order to maintain function, the active site of a protein remains highly conserved. Similarly, as the function of an enzyme evolves to accommodate a different substrate or catalyze a different reaction, only subtle changes in the geometry of the active-site or moderate sequence substitutions, deletions or insertions are necessary. These active site changes are likely to occur in a slow and step-wise process utilizing the same mechanisms that drive overall molecular evolution. Maintaining an active-site change will depend on how the new (if any) enzymatic activity is contributing to cell fitness. In essence, a sequential path from one active-site to another is expected where the observed sequence and structural changes directly contribute to the functional differences between the enzymes. This is exactly what was observed in the CPASS derived phylogenetic network. Nearest neighbors identifies subtle differences in active site sequences and structures that occur as the enzymatic activity and substrate specificity diverges. In effect, these changes identify an evolutionary path for the divergence of enzyme function. As an example, ACO transaminase, GABA transaminase, L-tyrosine transaminase, and aspartate transaminase are fold-type I enzymes belonging to clusters 1, 1, 4a, and 4b, respectively, in our CPASS-based network (Figures 1 and 2).

A direct comparison between the ACO and GABA transaminase active site structures identified a single amino acid substitution from a threonine to an asparagine, respectively, at the entrance to the active site (Figures 6a–b). A MSA of several ACO transaminase enzymes further verifies that a serine/threonine is conserved at this position (Figure 7a). The serine/threonine is most likely necessary to stabilize the larger ACO ligand through the formation of additional hydrogen bonds. Previous work has recognized the function of the conserved serine/threonine,[59,60] but has not identified the residue as being a primary difference between ACO and GABA transaminases. Conversely, a MSA of GABA transaminases confirms the presence of a conserved glycine/asparagine at the entrance to the binding pocket (Figure 7b). A small amino-acid at this location may open-up the binding pocket and/or increase flexibility, which, in turn, may lead to substrate promiscuity.[57] Thus, the glycine/asparagine to serine/threonine substitution may simply reduce substrate promiscuity and increase the selectivity to ACO. Inspection of the differences between the GABA and Asp active sites reveals the addition of a Tyr side chain in the active site of the Asp transaminase (Figures 6b, d). The hydroxyl group on the Tyr side chain is known to stabilize the internal aldimine between the PLP cofactor and the catalytic lysine at physiological pH and to increase the k_{cat} for the reaction in Asp transaminases.[61] Binding pocket hydrophobicity is also increased by the presence of the Tyr side chain, which may explain the moderate affinity of Asp transaminases for aromatic amino acids.[62] Importantly, ACO and GABA transaminases substitute a smaller glutamine residue to hydrogen bond with the cofactor.[59,63] A MSA of several Asp transaminases from functional clusters 4a and 4b shows that the Tyr residue is highly conserved (Figure 7c). Lastly, the progression from an Asp to Tyr transaminase follows a spatial substitution of Asn to Tyr as opposed to a sequence substitution. The Tyr at the entrance of the binding pocket fulfills the same role as the Asn by forming hydrogen bonds with the PLP phosphate group. However, the effects of the spatial rearrangement are twofold. Mutation of the Asn in Asp transaminases removes a coordinated water molecule from the binding pocket to make room for aromatic substrates,[64] and the presence of the Tyr side chain increases the hydrophobicity of the

active site.[62] Therefore, a single spatial substitution can greatly change the affinity for certain substrates in Type-I enzymes. An MSA of five Tyr transaminases confirms the conservation of the Tyr residue in the active site (Figure 7d)

CPASS recognized these subtle, sequential changes that occurred between these four, nearest-neighbor active sites while also acknowledging the overall active-site similarity. This enabled the close grouping of the ACO transaminase, GABA transaminase, L-tyrosine decarboxylase, and aspartate aminotransferase within the same functional cluster; and, importantly, the identification of a clear evolutionary path to achieve the observed substrate and functional divergence.

Conclusion

PLP-dependent enzymes have been a common target of molecular evolution studies, but prior efforts have been limited to small subgroups due to the low sequence identity within each fold-type and across all PLP-dependent enzymes.[2424,31,32] Instead, we have demonstrated the first phylogenetic network based on functional evolution to model all PLP-dependent fold-types together on a *single* network. We were able to correctly cluster all the PLP enzymes into their previously assigned fold-type simply based on the sequence and structure similarity of the PLP-binding site. Furthermore, the enzymes clustered based on their EC numbers within each fold-type cluster. Our CPASS-derived functional network allowed us to follow the step-wise evolution of substrate specificity and catalytic activity. Effectively, nearest neighbors in the CPASS-derived functional network were related by single-amino-acid changes for highly-conserved active-site residues. It was not possible to achieve a similar network using either global sequence or structure similarity. Thus, the study herein successfully determined an evolutionary relationship based on function between many diverse and highly divergent PLP-dependent enzymes using the available active site structures. Also, our results clearly verifies that CPASS is an effective approach for identifying an evolutionary relationship based on function using similarities in active site structure and sequence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute of Allergy and Infectious Diseases [grant number R21AI081154] as well as by a grant from the Nebraska Tobacco Settlement Biomedical Research Development Funds to RP. The research was performed in facilities renovated with support from the National Institutes of Health [grant number RR015468-01]. This work was completed utilizing the Holland Computing Center of the University of Nebraska.

References

1. Todd AE, Orengo CA, Thornton JM. Evolution of Function in Protein Superfamilies, from a Structural Perspective. J. Mol. Biol. 2001; 307:1113–1143. [PubMed: 11286560]
2. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. Evolving Genes Proteins, Symp Rutgers, State Univ. 1965; 1964:97–166.

3. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008; 92(5):255–264. [PubMed: 18703132]
4. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. Structural genomics: beyond the human genome project. Nat Genet. 1999; 23(2):151–157. [PubMed: 10508510]
5. Pál C, Papp B, Lercher MJ. An integrated view of protein evolution. Nat. Rev. Genet. 2006; 7(5): 337–348. [PubMed: 16619049]
6. Chothia C, Gough J. Genomic and structural aspects of protein evolution. Biochemical Journal. 2009; 419(1):15. [PubMed: 19272021]
7. Martincorena I, Seshasayee ASN, Luscombe NM. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature (London, U K). 2012; 485(7396):95–98. [PubMed: 22522932]
8. Eliot AC, Kirsch JF. Pyridoxal Phosphate Enzymes: Mechanistic, Structural, and Evolutionary Considerations. Annu. Rev. Biochem. 2004; 73(1):383–415. [PubMed: 15189147]
9. Hasson MS, Schlichting I, Moulai J, Taylor K, Barrett W, Kenyon GL, Babbitt PC, Gerlt JA, Petsko GA, Ringe D. Evolution of an enzyme active site: The structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. Proceedings of the National Academy of Sciences. 1998; 95:10396–10401.
10. La D, Sutch B, Livesay DR. Predicting protein functional sites with phylogenetic motifs. Proteins: Struct., Funct., Bioinf. 2004; 58(2):309–320.
11. Georgi B, Schultz J, Schliep A. Partially-supervised protein subclass discovery with simultaneous annotation of functional residues. BMC Struct. Biol. 2009; 9(1):68. [PubMed: 19857261]
12. Gutteridge A, Bartlett GJ, Thornton JM. Using A Neural Network and Spatial Clustering to Predict the Location of Active Sites in Enzymes. J. Mol. Biol. 2003; 330(4):719–734. [PubMed: 12850142]
13. Russell RB, Sasieni PD, Sternberg MJE. Supersites Withing Superfolds. Binding Site Similarity in the Absence of Homology. J. Mol. Biol. 1998; 282:903–918. [PubMed: 9743635]
14. Schmitt S, Kuhn D, Klebe G. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. J. Mol. Biol. 2002; 323(2):387–406. [PubMed: 12381328]
15. Powers R, Copeland JC, Germer K, Mercier KA, Ramanathan V, Revesz P. Comparison of protein active site structures for functional annotation of proteins and drug design. Proteins: Struct., Funct., Bioinf. 2006; 65(1):124–135.
16. Fetrow JS, Godzik A, Skolnick J. Functional Analysis of the Escherichia coli Genome Using the Sequence-to-Structure-to-Function Paradigm: Identification of Proteins Exhibiting the Glutaredoxin/Thioredoxin Disulfide Oxidoreductase Activity. J. Mol. Biol. 1998; 282:703–711. [PubMed: 9743619]
17. Kolker E, Picone AF, Galperin MY, Romine MF, Higdon R, Makarova KS, Kolker N, Anderson GA, Qiu X, Auberry KJ, Babnigg G, Beliaev AS, Edlefsen P, Elias DA, Gorby YA, Holzman T, Klappenbach JA, Konstantinidis KT, Land ML, Lipton MS, McCue LA, Monroe M, Pasa-Tolic L, Pinchuk G, Purvine S, Serres MH, Tsapin S, Zakrajsek BA, Zhu W, Zhou J, Larimer FW, Lawrence CE, Riley M, Collart FR, Yates JR, Smith RD, Giometti CS, Nealson KH, Fredrickson JK, Tiedje JM. Global profiling of Shewanella oneidensis MR-1: Expression of hypothetical genes and improved functional annotations. Proceedings of the National Academy of Sciences. 2005; 102(6):2099–2104.
18. Hemrika W, Renirie R, Dekker HL, Barnett P, Wever R. From phosphatases to vanadium peroxidases: A similar architecture of the active site. Proceedings of the National Academy of Sciences. 1997; 94:2145–2149.
19. Kull FJ, Vale RD, Fletterick RJ. The case for a common ancestor: kinesin and myosin motor proteins and G proteins. Journal of Muscle Research and Cell Motility. 1998; 19:877–886. [PubMed: 10047987]
20. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010; 20(10):1313–1326. [PubMed: 20651121]

21. Hughes AL. The evolution of functionally novel proteins after gene duplication. Proc R Soc London, Ser B. 1994; 256(1346):119–124.
22. Prince VE, Pickett FB. Splitting pairs: The diverging fates of duplicated genes. Nat Rev Genet. 2002; 3(11):827–837. [PubMed: 12415313]
23. Jones S, Thornton JM. Searching for functional sites in protein structures. Curr Opin Chem Biol. 2004; 8(1):3–7. [PubMed: 15036149]
24. Christen P, Mehta PK. From Cofactor to Enzymes. The Molecular Evolution of Pyridoxal-5'-Phosphate-Dependent Enyzmes. The Chemical Record. 2001; 1:436–447.
25. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res. 2000; 28(1):304–305. [PubMed: 10592255]
26. Jansonius JN. Structure, evolution and action of vitamin B_6-dependent enzymes. Curr. Opin. Struct. Biol. 1998; 8:759–769. [PubMed: 9914259]
27. Dunathan HC. Conformation and Reaction Specificity in Pyridoxal Phosphate Enzymes. Proceedings of the National Academy of Sciences. 1966; 55:712–716.
28. Hayashi H. Pyridoxal Enzymes: Mechanistic Diversity and Uniformity. Journal of Biochemistry. 1995; 118:463–473. [PubMed: 8690703]
29. Christen P, Kasper P, Gehring H, Sterk M. Stereochemical constraint in the evolution of pyridoxal-5'-phosphate-dependent enzymes. A hypothesis. FEBS Lett. 1996; 389:12–14.
30. Schneider G, Kack H, Lindqvist Y. The manifold of vitamin B_6 dependent enzymes. Structure. 2000; 8:1–6. [PubMed: 10673423]
31. Alexander FW, Sandmeier E, Mehta PK, Christen P. Evolutionary relationships among pyridoxal-5'-phosphate-dependent enzymes. Regio-specific α,β and γ families. Eur J Biochem. 1994; 219(3):953–960. [PubMed: 8112347]
32. Mehta PK, Christen P. The molecular evolution of pyridoxal-5'-phosphate-dependent enzymes. Adv Enzymol Relat Areas Mol Biol. 2000; 74:129–184. [PubMed: 10800595]
33. Powers R, Copeland JC, Stark JL, Caprez A, Guru A, Swanson D. Searching the protein structure database for ligand-binding site similarities using CPASS v.2. BMC Research Notes. 2011; 4(1): 17. [PubMed: 21269480]
34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–242. [PubMed: 10592235]
35. Henikoff S, Henikoff JG. Amino acid substitution matrixes from protein blocks. Proc Natl Acad Sci U S A. 1992; 89(22):10915–10919. [PubMed: 1438297]
36. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrixes. Proteins: Structure, Function, and Genetics. 1993; 17(1):49–61.
37. Consortium TU. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2007; 35:D193–D197. (Database). [PubMed: 17142230]
38. Huson DH. Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. 2005; 23(2):254–267. [PubMed: 16221896]
39. Bryant D. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Mol. Biol. Evol. 2003; 21(2):255–265. [PubMed: 14660700]
40. Huson DH, Scornavacca C. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Systematic Biology. 2012; 61(6):1061–1067. [PubMed: 22780991]
41. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 2011; 7
42. Zhang Y. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005; 33(7):2302–2309. [PubMed: 15849316]
43. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera-A Visualization System for Exploratory Research and Analysis. J. Comput. Chem. 2004; 25(13):1605–1612. [PubMed: 15264254]
44. Jensen RA, Gu W. Evolutionary Recruitment of Biochemically Specialized Subdivisions of Family I within the Protein Superfamily of Aminotransferases. J. Bacteriol. 1996; 178(8):2161–2171. [PubMed: 8636014]

45. Grishin NV, Phillips MA, Goldsmith EJ. Modeling of the spatial structure of eukaryotic ornithine decarboxylases. Prot. Sci. 1995; 4:1291–1304.
46. Rausch C, Lerchner A, Schiefner A, Skerra A. Crystal structure of the ω-aminotransferase from Paracoccus denitrificans and its phylogenetic relationship with other class III aminotransferases that have biotechnological potential. Proteins: Struct., Funct., Bioinf. 2013; 81(5):774–787.
47. Liu L, Iwata K, Yohda M, Miki K. Structural insight into gene duplication, gene fusion and domain swapping in the evolution of PLP-independent amino acid racemases. FEBS Lett. 2002; 528:114–118. [PubMed: 12297289]
48. Kidron H, Repo S, Johnson MS, Salminen TA. Functional Classification of Amino Acid Decarboxylases from the Alanine Racemase Structural Family by Phylogenetic Studies. Mol. Biol. Evol. 2006; 24(1):79–89. [PubMed: 16997906]
49. Shah R, Coleman CS, Mir K, Baldwin J, Van Etten JL, Grishin NV, Pegg AE, Stanley BA, Phillips MA. Paramecium bursaria Chlorella Virus-1 Encodes an Unusual Arginine Decarboxylase That Is a Close Homolog of Eukaryotic Ornithine Decarboxylases. J Biol Chem. 2004; 279(34):35760–35767. [PubMed: 15190062]
50. Sugio S, Petsko GA, Manning JM, Soda K, Ringe D. Crystal Structure of a D-Amino Acid Aminotransferase: How the Protein Controls Stereoselectivity. Biochemistry. 1995; 34:9661–9669. [PubMed: 7626635]
51. Rost B. Enzyme function less conserved than anticipated. J Mol Biol. 2002; 318(2):595–608. [PubMed: 12051862]
52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17):3389–3402. [PubMed: 9254694]
53. Karplus K, Barrett W, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics. 1998; 14(10):846–856. [PubMed: 9927713]
54. Doolittle, RF. Of URFs and ORFs: A primer on how to analyze derived amino acid sequences. University Science Books; 1986.
55. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008; 36(7):2295–2300. [PubMed: 18287115]
56. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 2006; 34:W604–W608. (Web Server). [PubMed: 16845081]
57. Percudani R, Peracchi A. A genomic overview of pyridoxal-phosphate-dependent enzymes. EMBO reports. 2003; 4(9):850–854. [PubMed: 12949584]
58. Percudani R, Peracchi A. The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families. BMC Bioinf. 2009; 10(1):273.
59. Rajaram V, Ratna Prasuna P, Savithri HS, Murthy MRN. Structure of biosynthetic N-acetylornithine aminotransferase from Salmonella typhimurium: Studies on substrate specificity and inhibitor binding. Proteins: Struct.,Funct.,Bioinf. 2007; 70(2):429–441.
60. Kack H, Sandmark J, Gibson K, Schneider G, Lindqvist Y. Crystal Structure of Diaminopelargonic Acid Synthase: Evolutionary Relationships between Pyridoxal-5'-phosphatedependent Enzymes. J. Mol. Biol. 1999; 291:857–876. [PubMed: 10452893]
61. Goldberg JM, Swanson RV, Goodman HS, Kirsch JF. The Tyrosine-225 to Phenylalanine Mutation of Escherichia coli Aspartate Aminotransferase Results in an Alkaline Transition in the Spectrophotometric and Kinetic pK_a Values and Reduced Values of both k_{cat} and K_m. Biochemistry. 1991; 30:305–312. [PubMed: 1988027]
62. Kamitori S, Okamoto A, Hirotsu K, Higuchi T, Kuramitsu S, Kagamiyama H, Matsuura Y, Katsube Y. Three-Dimensional Structures of Aspartate Aminotransferase from *Escherichia coli* and Its Mutant Enzyme at 2.5A Resolution. Journal of Biochemistry. 1990; 108:175–184. [PubMed: 2121725]
63. Liu W, Peterson PE, Carter RJ, Zhou X, Langston JA, Fisher AJ, Toney MD. Crystal Structures of Unbound and Aminooxyacetate-Bound *Escherichia coli* y-Aminobutyrate Aminotransferase. Biochemistry. 2004; 43:10896–10905. [PubMed: 15323550]

64. Malashkevich VN, Onuffer JJ, Kirsch JF, Jansonius JN. Alternating arginie-modulated substrate specificity in an engineered tyrosine aminotransferase. Nat. Struct. Biol. 1995; 2(7):548–553. [PubMed: 7664122]

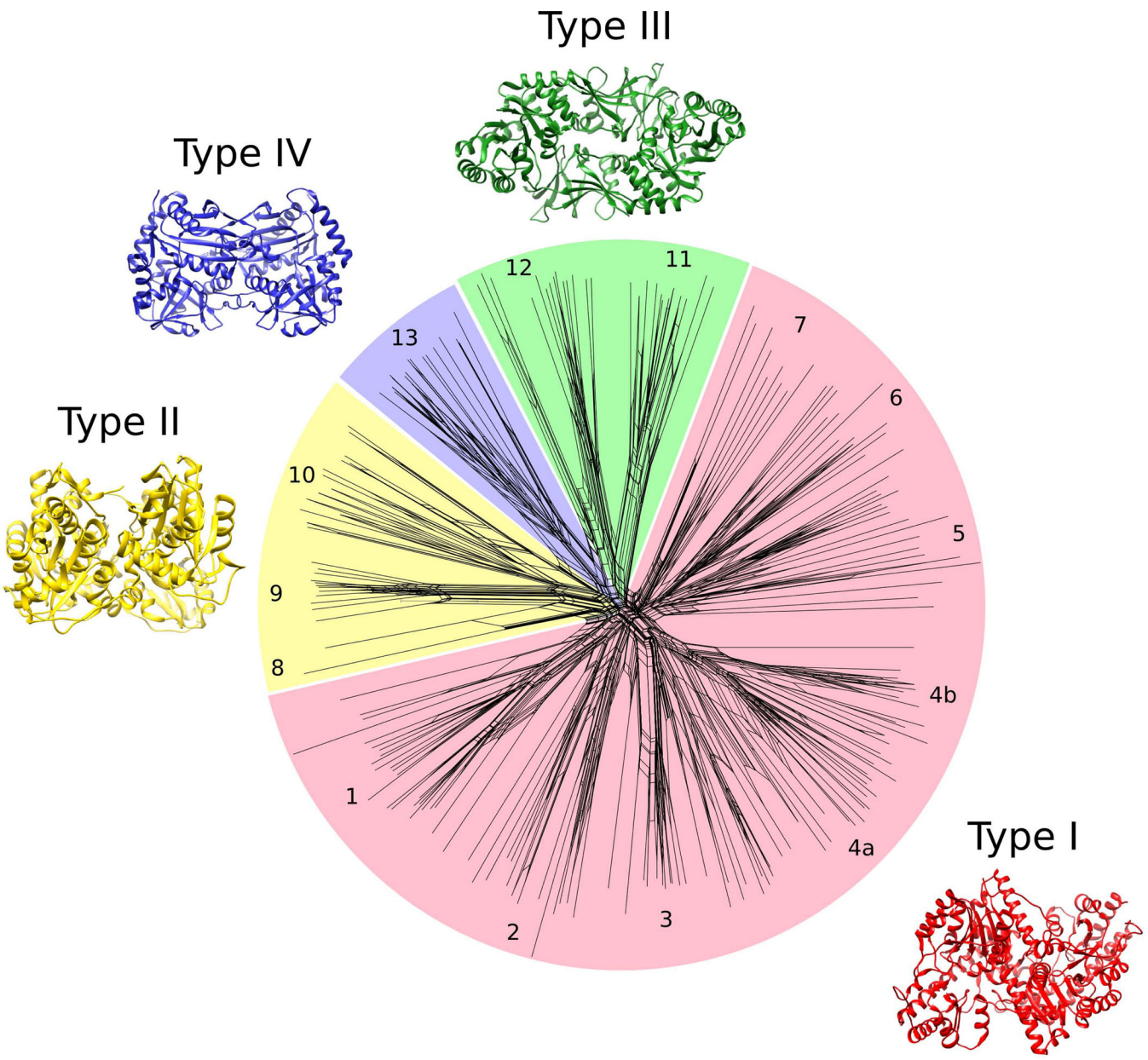

Figure 1. Active site network analysis of PLP-dependent enzymes

A phylogenetic network analysis of 204 PLP-dependent enzymes with ligand defined active sites in the PDB. The red, yellow, green and blue colored regions and protein structures correspond to fold-types I, II, III and IV, respectively. Individual clusters found within each of the folds represent groups of function specific enzymes and are numbered 1–13. The network was generated using CPASS similarity scores and the Neighbor-net method.

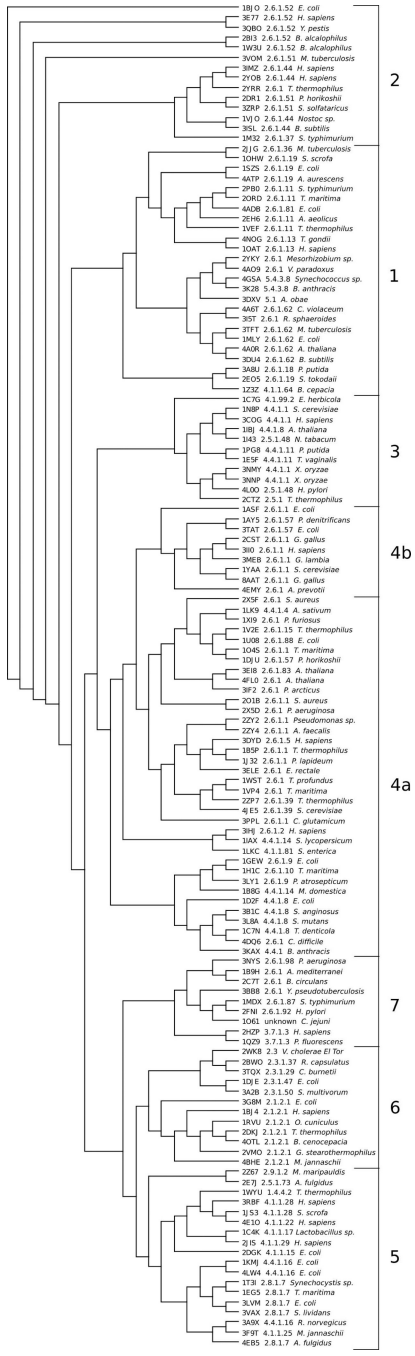

Figure 2. Type-I PLP-dependent enzyme phylogenetic analysis

The phylogenetic tree of PLP-dependent enzymes belonging to the Type-I fold. Each leaf of the tree is annotated with PDB ID, EC number and species. Branch lengths represent the evolutionary distance between enzymes based on active site structure comparisons. Functional clusters 1–7 correspond to the numbers in the network analysis of all PLP-dependent enzymes (Figure 1). The tree was constructed using CPASS similarity scores and the Neighbor-join algorithm.

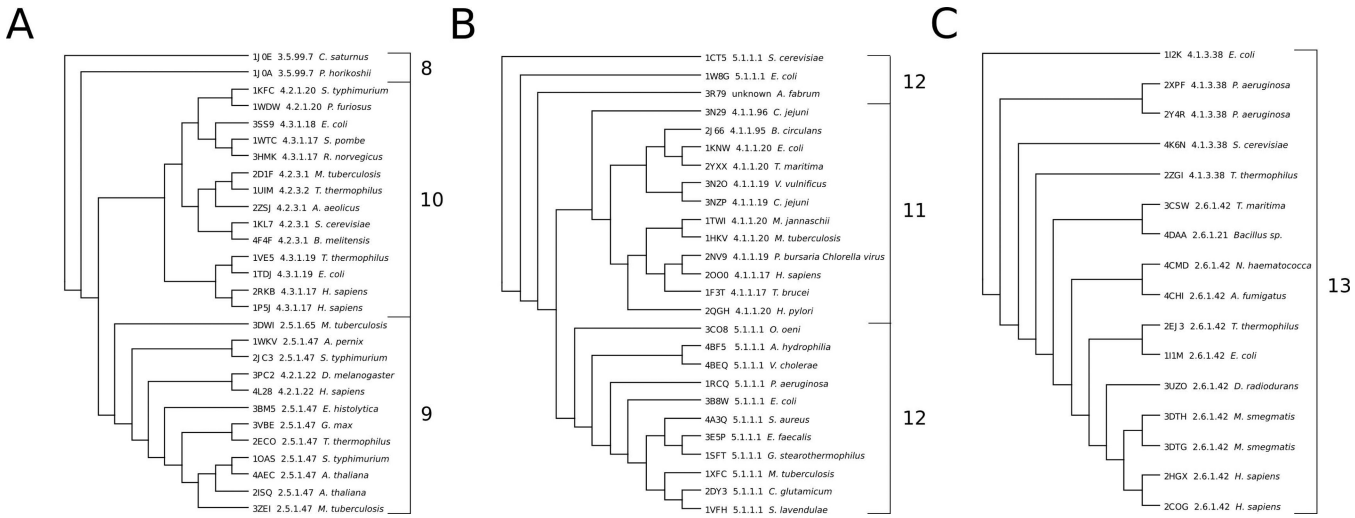

Figure 3. Types-II, -III, and -IV PLP-dependent enzyme phylogenetic analysis

Phylogenetic reconstructions of PLP-dependent enzymes from Types-II (A), -III (B), and -IV (C) folds were built using CPASS similarity scores and the Neighbor-join algorithm. Enzymes are labeled with their PDB ID, an EC number and organism. Tree branch lengths correspond to the evolutionary distance between the active sites of the enzymes. Numbers 8–13 represent the functional clusters found in the network analysis (Figure 1).

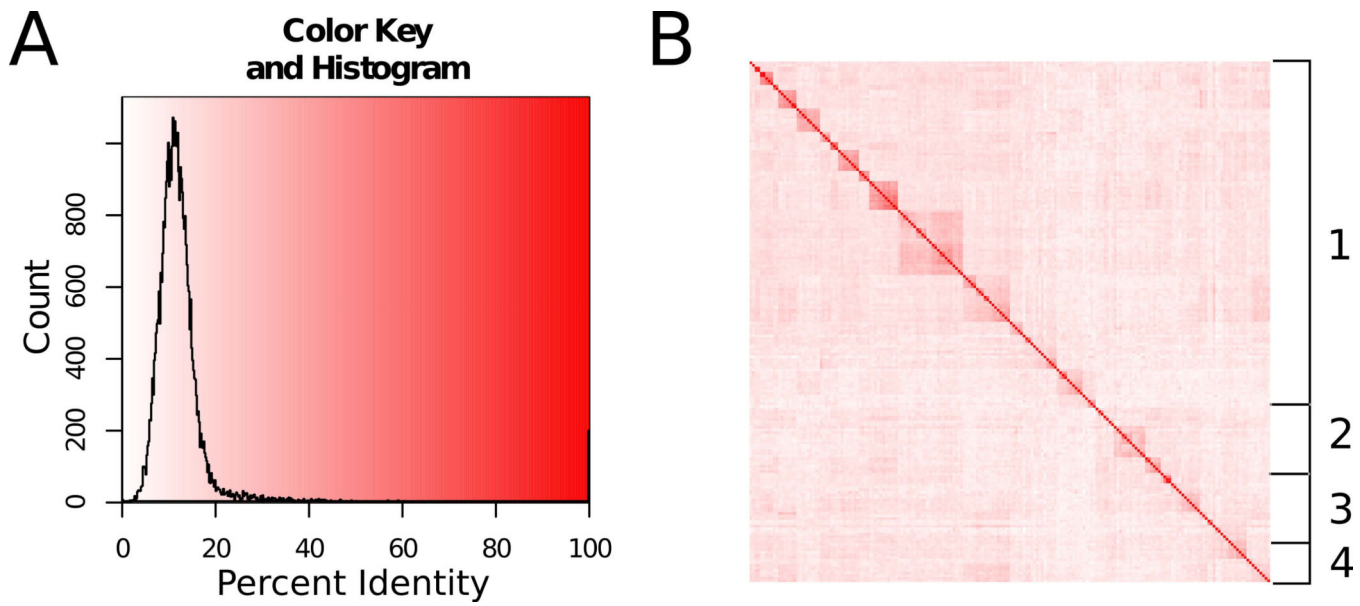

Figure 4. Sequence identity histogram and heat map

(A) A histogram of the percent similarity matrix representing 204 PLP-dependent protein sequences. The entire similarity matrix was analyzed in 0.2% intervals to generate the plot. (B) A heat map of the percent similarity matrix used in (A). The white to red coloring of the heat map corresponds to the percent similarity and color gradient in (A). The order of the sequences is the same as in Table S1 and the associated network map is illustrated in Figure S1.

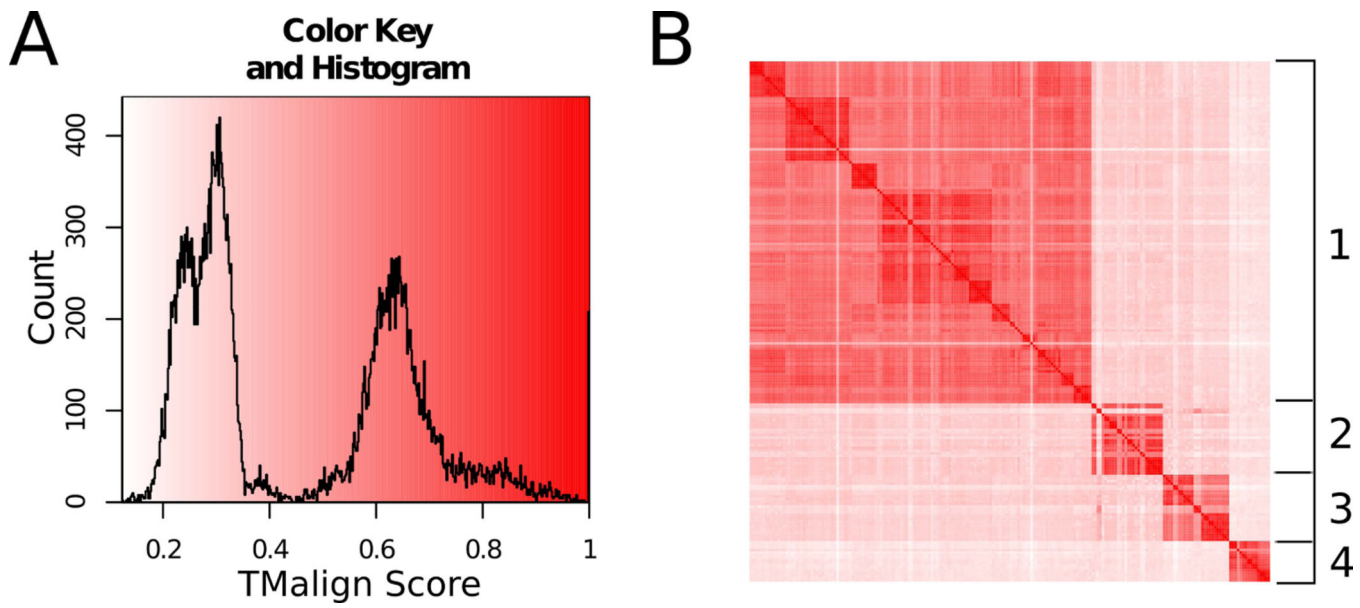

Figure 5. Structural alignment histogram and heat map

(A) A histogram of the pairwise TMalign scores for the 204 PLP-dependent proteins in the dataset. The entire symmetrical score matrix was evaluated in 0.002 intervals to produce the plot. (B) A heat map of the corresponding TMalign score matrix used in (A). The white to red coloring scheme of the heat map corresponds to the score and gradient in the respective histogram. The order of the sequences is the same as in Table S1 and the associated network map is illustrated in Figure S2.

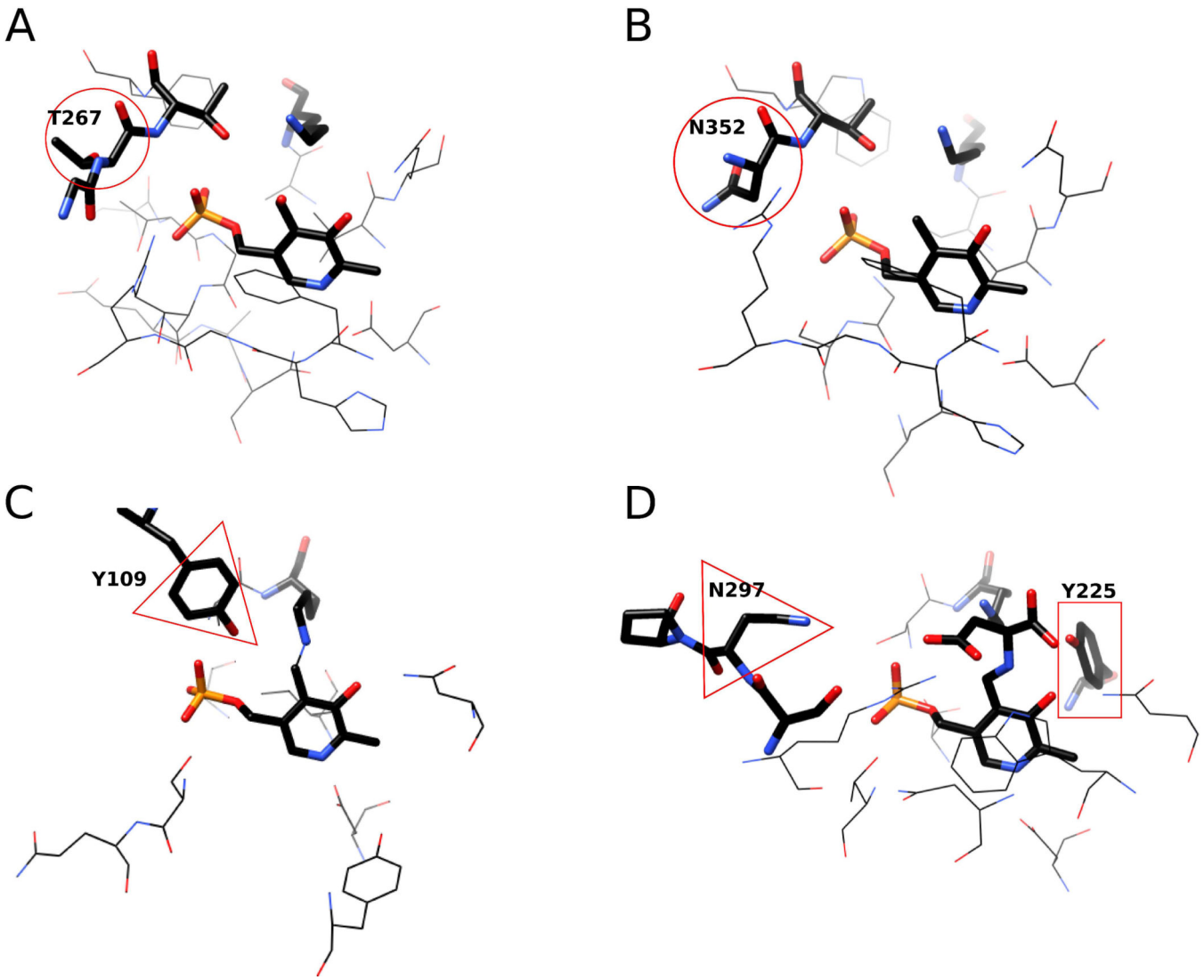

Figure 6. Side by side alignment of ACO, GABA, Asp, and Tyr transaminase active sites

(A) The active site of 2ORD, an ACO transaminase. Circled in red is the threonine residue that stabilizes the substrate (not shown) in the active site. (B) The active site of 1OHW, a GABA transaminase. The red circle indicates the position of the asparagine residue relative to the position of the threonine residue from (A). (C) The active site of 3DYD, a Tyr transaminase. Indicated in the red triangle is the conserved tyrosine residue relative to the position of the asparagine residue in (D). (D) The active site structure of 1ARG, an Asp transaminase. The red box indicates the addition of a tyrosine residue in the active site relative to the GABA transaminase active site in (B).

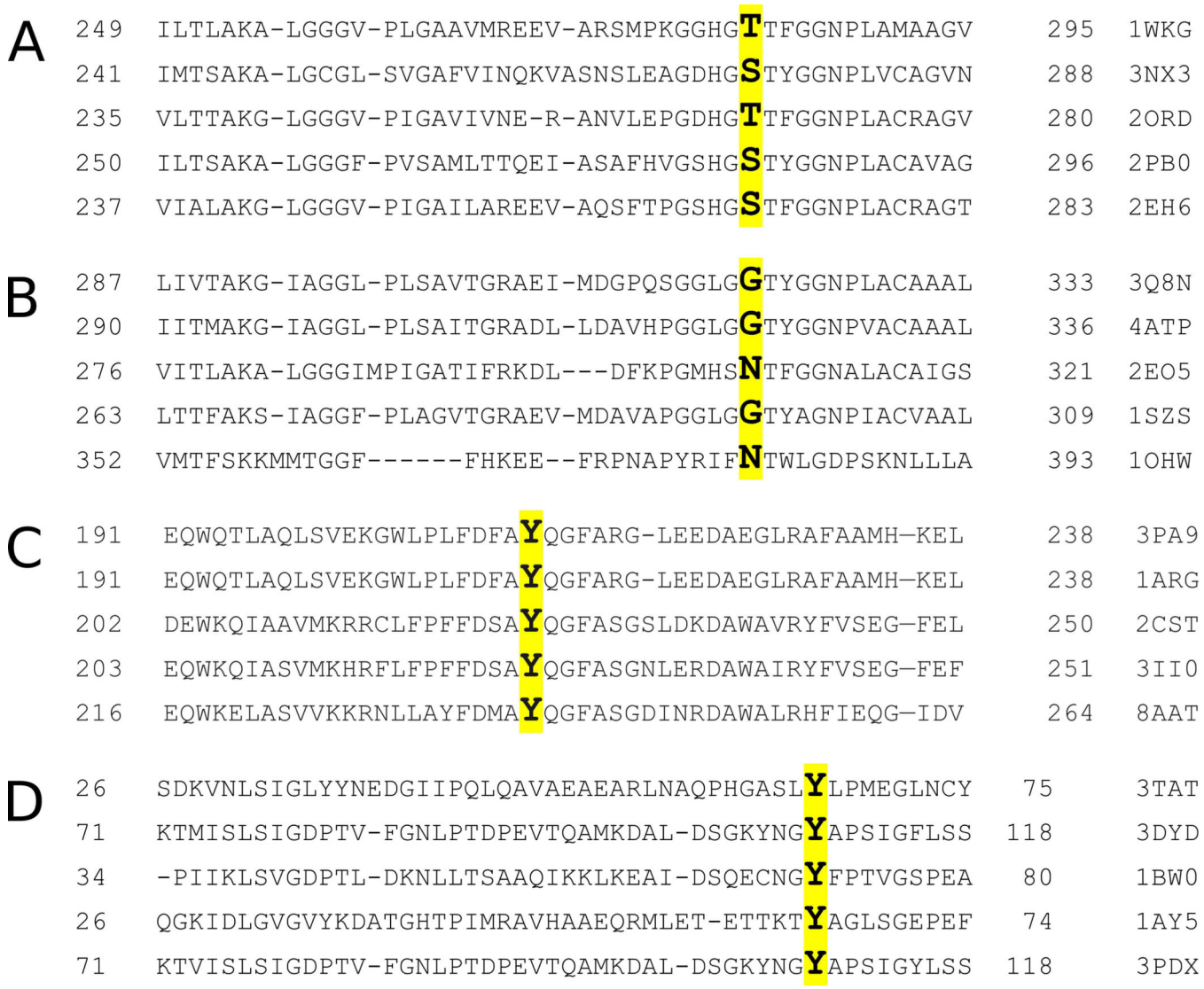

Figure 7. MSAs of ACO, GABA, Asp, and Tyr transaminases

Segments of individual MSAs from ACO (A), GABA (B), Asp (C), and Tyr (D) transaminases. Each segment highlights a conserved residue corresponding to the active sites in Figure 7. PDB IDs of the selected enzymes are on the far right of the alignment.