TITLE: Thiamin-Diphosphate Enzymes Are an Ancient Family of Repeat Proteins

AUTHORS: Matthew Merski^{1,*} (ORCID 0000-0002-1844-6997) & Maria Górna^{1,*} (ORCID 0000-0002-1769-4542)

AUTHOR AFFILIATION: ¹Structural Biology Group, Biological and Chemical Research Centre, Department of Chemistry, University of Warsaw, Żwirki i Wigury 101, 02-089 Warsaw, Poland

CORRESPONDING AUTHOR: * to whom correspondence should be addressed (mailing address above): merski@gmail.com (+48 57 59 06 279); mgorna@chem.uw.edu.pl (+48 22 55 26 685)

ABSTRACT:

A repeating sequence and structure pattern that is highly similar to the canonical cofactor binding motif has been identified in the thiamin-diphosphate dependent (ThDP) enzyme family. We have identified more than a thousand of these repeats in a non-redundant set (N = 58) of ThDP enzyme structures. The repeating element has a helix-turn-strand secondary structure which typically begins with an [G/A]{X(1,2)}[G/A] sequence motif with a typical length of 29 residues. The catalytically important diphosphate and aminopyrimidine interacting domains are comprised of a set of six of these repeats in a conserved architecture with a flavodoxin-like 213465 strand order. The canonical ThDP binding motif is the fourth repeat in the ThDP binding domain, while the conserved aminopyrimidine interacting glutamate is part of the second repeat in its domain. The third and fourth repeats form a contact between the functional domains, while the fifth repeat in the N-terminal domain forms an inter-chain contact. The conservation of these functional properties highlights the role of these repeats in the function and structure of this well-studied enzyme family and agrees with the principle of modular assembly in protein ancestry.

INTRODUCTION: The origin of life on Earth at or before the beginning of the Archean period 4 billion years ago (BYA) required several events to occur: the organization of genetic information into a storage system, the formation of semi-permeable membranes, and the development of stable protein architectures¹. It has been theorized, first by Dayhoff², that small peptides formed spontaneously in the primordial Archean sea collected into oligomeric structures, possibly bound to nucleotides or similar compounds, which in turn then catalyzed additional synthesis of these oligomeric peptides, functioning as an early ribosome³. Fusion of these oligomeric peptide complexes into larger proteins could have given rise structures comprised of repeating sequences, which through genetic drift over a long enough time would eventually lead to complex, globular proteins⁴⁻⁶. These ancestral proteins may have had modular architectures, similar to modern repeat proteins in which a short (20-40 residue) sequence motif occurs several times within the protein chain⁷. Given that there are a number of globular proteins that are ubiquitous in all three kingdoms of life, the transition from the earliest repeat proteins to globular proteins likely preceded the last common universal ancestor (LUCA)⁸ that lived during the Archaean eon at least 2.5 BYA. The extreme length of time that has passed since then makes it likely that any repeating sequences have at best only weak similarity due to inevitable sequence drift. However, modern mammalian repeat proteins are quite well conserved despite 300 million years passing since the origin of that group suggesting that this divergence might occur quite slowly⁹. On the other hand, repeat proteins are more common in eukaryotic than bacterial genomes¹⁰ suggesting that many if not most date from the eukaryotic origin during the Proterozoic eon about 1-2 BYA¹¹ although the repetitive nature of repeat proteins makes untangling the phylogenic relationships between these proteins difficult¹². While as many as 25% of known, modern proteins can be classified as repeat proteins¹³⁻¹⁴, identification and analysis of the repeats themselves is complicated by several issues. Repeat sequences are essentially circular: there is no true start or end position¹⁴, insertions and deletions can occur within or between the repeats changing their apparent lengths, some repeats have a naturally variable length^{7, 15}, and repeats can also occur a non-integer number of times within a protein⁷. Fortunately, a number of computational repeat detection methods exist, each focused around a different methodology and each with its own strengths and weaknesses. Unfortunately, a comparison of four distinct identification methods found little (0.2%) consensus agreement between them¹⁶. Despite these difficulties, a general consensus has emerged about the sequence properties of repeat proteins, namely that they are 1) self-homologous regions within a protein, 2) which have amino acid compositional biases, and 3) repeating sequence and structure patterns¹⁷. Awareness of these difficulties led

to the development of a curated database of repeat protein structures (RepeatsDB) which uses a functional definition of a repeat protein as one that contains at least three copies of a repeating sequence with a clearly defined structure¹⁸. Recently, we developed a method to cluster related repeat protein sequences based on patterns of self-homology¹⁹ and we used this method to analyze the set of non-redundant known protein sequences (UniRef90)²⁰. Upon examining our results we were surprised to find that a significant number of these related repeat protein clusters were comprised exclusively of thiamin-diphosphate (ThDP) dependent enzymes.

The ThDP enzymes are widely dispersed throughout the kingdoms of life and catalyze a wide range of reactions utilizing their nucleotide-like cofactor ThDP, the biologically active form of vitamin B₁²¹. The reactions catalyzed by these enzymes have been the subject of 8 decades of investigation and the first crystal structures were reported nearly 30 years ago²² and since then several dozen genetically unique ThDP enzymes have been structurally characterized. A typical ThDP enzyme monomer is comprised of three similar domains each containing a core sheet of six β -strands surrounded by α -helices²³⁻²⁴ with the β -strands arranged in a parallel 213465 order that distinguishes it from other 3-layer alpha/beta/alpha folds such as the 321456 strand arrangement present in Rossmann domains²⁵⁻²⁶. One of these three domains (PP) interacts with the diphosphate moiety of the cofactor while a second domain (PYR) interacts with the cofactor's aminopyrimidine ring through an almost universally conserved glutamate²⁷⁻²⁸. All ThDP enzymes have at least one additional domain (here CFX, for cryptic function) which contain more sequence variation than the functional domains and have been previously sub-divided into several groups²⁹. Some CFX domains are known to bind the nucleotide-like cofactor flavin adenine dinucleotide (FADH) or the nucleotide adenosine triphosphate (ATP)³⁰⁻³¹. The ThDP family has been divided into two major groups based on the order of these domains in the protein²³. In the first group (DH), which is comprised of the pyruvate dehydrogenases³² and the transketolases³³, the domains are arranged PP, PYR, CFX from N to C. The second group (DC) which includes the acetohydroxyacid synthases³¹ and several decarboxylases³⁴⁻³⁵ is a circular permutation of the first group in which the domains are ordered PYR, CFX, followed by PP^{23, 36}. While the ThDP family has low shared sequence identity between its members³⁷⁻³⁸ they all can be easily recognized by a canonical thiamin-binding motif, identified three decades ago which consists of a GDG sequence followed by about 30 variable residues and finally a conserved glutamine³⁹. This canonical motif has been strictly maintained throughout the evolutionary history of this family, which is quite ancient and likely predates LUCA^{8, 40-43}, allowing ready

identification of ThDP enzymes from genomic sequences. Taking advantage of this recognizable motif, the TEED database of ThDP enzymes was assembled to aid bioengineering efforts and as of this writing it contains almost 120,000 protein sequences²⁹.

Upon identifying the presence of protein repeats within the ThDP family we quickly noted that the repeating units themselves were homologous to the canonical ThDP binding motif. (Fig. 1). We were then able to analyze these protein repeats which underlie of all the members of the ThDP enzyme family revealing that the complex architecture of this diverse family of globular enzymes is easily built up from a set of simple, repeated oligopeptides.

RESULTS:

Identification of ThDP sequence repeats: After identifying the presence of a notable amount of self-homology in the ThDP enzymes present in UniRef90¹⁹⁻²⁰, we engaged in a more thorough analysis of those proteins present in the PDB⁴⁴. Sequence redundant proteins were removed and the remaining ones were organized phylogenetically (Fig. 2). Reverse calculation using DOTTER⁴⁵ identified the canonical ThDP binding motif as one of the regions of self-homology and a (G/A){X(1,2)}(G/A) motif (where G/A indicated either a glycine or alanine residue and X(1,2) indicates one or two variable residues) at the beginning of a helix-turn-strand structural motif was quickly identified (Fig. 3). While the choice of the beginning of a repeat is largely arbitrary^{14, 46}, we identified this motif as the start of the repeat for historical reasons³⁹. Using these basic features, we were able to identify 1003 repeats from a set of 58 non-redundant ThDP proteins (Figure 3, SI Fig. S1, SI Table S1, S2, & S3) present in the Protein Data Bank in October 2019⁴⁷ by manually examination (Materials and Methods). At least three repeat sequences were identified in each protein (*i.e.* one of the basic identifiers of protein repeats¹⁶). All the repeats had an average (mean) length of 27 residues (median = 26), although the repeats from the functional PP and PYR domains tended to be slightly longer than those from the CFX domain (mean = 28.7, 27.9, & 24.0 respectively) (Fig. 4, SI Table S1). 663 (66%) of all the identified repeats were 22-32 residues long with a mean length of 26.8 residues, suggesting a relatively standard repeat length (Figure 4) and these comprised the “best repeats” subset (SI Tables S1 & S2). While the strongest sequence conservation currently detectable was the (G/A){X(1,2)}(G/A) motif, some other sequence tendencies were also identified (SI Table S4). The helix itself is somewhat enriched in alanine residues, although this is typical of helical regions in the PDB in general. There is often a glycine or proline in the first or second position after the helix (37.4% of the repeats between 22-32 residues long and 36.6% of all the identified repeats) and a proline or a glycine in the

position before the start of the strand (26.5% of the repeats between 22-32 residues long and 19.5% of the full set, see SI Tables S1-S4). Both of these proline/glycine residues were present in 7.8% of the repeats that were 22-32 residues long (5.4% among all the identified repeats). The strand itself is rich in valine, leucine, and isoleucine residues and the strand often ends on a hydrogen bond donating residue. These basic components were also noticed when the canonical ThDP binding motif was first identified³⁹ and at least one later analysis of the canonical ThDP motif also noted a conserved valine residue and other sequence features preceding the canonical GDG motif, exactly as would be expected to occur due to the preference for hydrophobic residues in the strand at the C-terminus of the preceding repeat⁴⁸. Several conserved functional positions were also associated with specific repeats (Figure 1C). The canonical ThDP binding sequence³⁹ is always the fourth repeat in the PP domain. The conserved catalytic glutamate²⁷ is present at the end of repeat two in the PYR domain. Repeat two in the CFX domain is associated with interaction with an additional nucleotide cofactor in those enzymes which also bind FADH or ATP³⁰⁻³¹. With these repeats identified, we expanded our analysis to the entirety of the TEED database²⁹. Sequences in the TEED database that had significant (>30%) pairwise identity to each of the 58 PDB structure sequences were used to generate matrices of consensus sequences. An increasing bias towards the amino acids alanine, glycine, and proline and a bias away from lysine and cysteine, which became more evident as the threshold for conservation (Simpson)⁴⁹⁻⁵⁰ increased and the sequence identity threshold decreased, was identified within the repeat regions (SI Figure S2). This bias became more and more pronounced as the overall conservation of the proteins decreased and the observed diversity at that specific position decreased. We also attempted to identify more specific sequence requirements within the repeats but this was stymied by the bias of the TEED database towards transketolases (TK), acetohydroxyacid synthases (AHAS) and pyruvate dehydrogenases (PDH) (SI Table S5). This largely restricted our analyses to these functional classes; we provide one analyzed example each from the DH class (the transketolases) (SI Figure S3) and one from the DC class (the acetohydroxyacid synthases, SI Figure S4). While there was no ubiquitous sequence pattern (other than the initial (G/A){X(1,2)}(G/A) motif), there was notable self-similarity between the copy of the repeat found in the PP and that found in the PYR domain for every repeat position in the representatives from both the DH and DC functional classes. The CFX domain repeats were much more diverged from the functional domain repeats and the domain topology is also different (see below) making it unclear if the repeats are present in the same order in the functional and CFX domains. While self-homologous residues were easier to detect near the

edges of the repeats it was possible to identify a number of positions conserved in both classes of ThDP enzyme (*e.g.* “AATF” or “DKPT $\phi\phi$ ” toward the C-terminus of the third and sixth DH repeats (SI Figure S3) or “PGPVLLD ϕ ” near the C-terminus of the sixth DC repeat (SI Figure S4) where ϕ is a hydrophobic residue). In several cases it was possible to detect the same conserved residue patterns from both the N and C termini allowing an additional estimate of the overall repeat length. Additionally, in both the DH and DC classes a conserved alanine was located at position 10 in repeat 3 and a conserved pair of residues was also identified at positions 5-6 in repeat 5 in both classes. These conserved residues were linked to conserved contacts in the ThDP enzymes (see below).

Our DOTTER-based self-homology method¹⁹ was not the only analysis method which could detect repeats in ThDP enzymes as a weak but noticeable positive signal was detected by other computational repeat identification methods (Supplemental Data S1). While not every repeat identification method could find the ThDP repeats (which is not unusual for repeat identification¹⁶) the fact that these repeats could be identified by multiple methodologies strengthens the assertion that these are protein repeats. That the divergence of this family occurred a remarkably long time ago suggests that ThDP-like repeats may exist in other protein families, either through gene transfer or due to their relative simplicity by having arisen multiple times during the existence of life on Earth. A search of the PDB using the ThDP repeat identified 881 non-ThDP enzyme structures with at least 3 copies of this repeat. Furthermore, these proteins tend to be enzymes that bind nucleotide-like molecules (Supplemental Data S2), as had been previously noted for a “transketolase motif” which also fits comfortably within our definition of the ThDP repeats³³. Given the likely ancient origin of the ThDP enzymes^{8, 41} and the significant amount of time that has passed since divergence of the family, it is not surprising that these repeats are generally weakly conserved, although variation similar to what is seen in the ThDP enzymes is also not unusual for protein repeats¹⁵.

Structural features of the repeats in the ThDP enzymes: The tandem repeat pattern can be easily identified in the structures of the individual domains in the ThDP enzymes and the arrangement of the repeats is highly conserved within the branches of ThDP family with a few notable exceptions (SI Table S6, SI Figure S5). The DH enzyme have a conserved insertion containing a secondary structure element after the first and fourth repeats in the PP domain (the insertion after the fourth repeat is missing in the phosphoketolase (PhK) enzymes) while the DC enzymes also have a conserved insertion after the fourth repeat in the PP domain. This insertion is also present in the PFOR enzymes, but is not associated with a secondary structure

element. The spatial arrangement of the repeats is also highly conserved. The functional PP and PYR domains have a 213465 parallel strand arrangement with the 213 and 465 sets of helices partitioned above and below their respective strand sets in the DH, DC and PFOR enzymes (Figure 1 C, SI Figure S6A). The CFX domains of the DC class also contain 6 repeats but in a Rossmann-like 321456 parallel arrangement with the 321 and 456 sets of helices arranged exclusively above or below the plane of the strands (Figure 1D, SI Figure S6B). This domain is not a true Rossmann domain as it does not contain the conserved ligand binding geometry of the true Rossmann domains²⁶ (SI Table S7). The CFX domains of the DH class are shorter, typically only containing 4 repeats with the first strand (Figure 1E, SI Table S6, SI Figure S6C). An additional two types of CFX domain with antiparallel strand arrangements are found in the PFOR enzymes (SI Figure S6D & S6E).

The repeat residues that are conserved in the both the DH and DC enzyme classes reflect the architecture of the ThDP enzymes. While it has long been known that the residues associated with repeat 4 in the PP domain bind the ThDP phosphate, the N-terminus of the helix in repeat 2 and the conserved sequences in repeat 5 (or in the equivalent insert region in the DC enzymes) also interact with the cofactor phosphate through the chelating magnesium. The conserved alanine at position 10 in repeat 3 is associated with the intra chain contact between the two functional domains although this alanine faces away from the contact surface (SI Figure S7). Likewise, the conserved residues in repeat 5 form an inter chain contact along with repeat 4 (SI Figure S8) and while there are no obvious hydrogen bonding interactions between these two helices, it should be noted that repeat 3 is in a central position in the functional domains and as this is an intra chain interaction and it is suspected that hydrophobic burial drives this interaction. Much like repeat 3, these residues are somewhat poorly oriented to directly participate in the interaction, they are difficult to identify from the sequences (SI Figure S1). While these residues can be readily identified from the structure, and there is a tendency for the sequence to be a hydrogen bonding residue followed by small one. Additionally, in the structure of a BFD enzyme (PDBID 6A50), there is a magnesium binding site formed by this inter chain contact through their backbone carbonyls in repeat 5 of the PYR domain.

DISCUSSION: The first structure of a ThDP enzyme was solved in the early 1990s⁵¹, and dozens of additional structures have been reported since then, revealing a number of nearly perfectly conserved features of this family, including both the canonical ThDP binding motif³⁹ and the conserved catalytic glutamate²⁷. Yet numerous questions about this family of enzymes

have remained unanswered, notably that the catalytic glutamate is not universally conserved²⁸, there is significant sequence diversity among the members of the family³⁷⁻³⁸, and that even the domain arrangements within the ThDP family can be variable³⁶, raising the question as to what is being fundamentally maintained by evolution within this family of enzymes, other than the canonical ThDP motif itself. But now the identification of the protein repeat architecture in the ThDP enzymes reveals an elegant organizing principle for this entire family of enzymes (Fig. 1).

Even though the ThDP enzymes diverged from their common ancestor in the distant past (even before LUCA), and the overall similarity between the repeats is quite weak, there are still substantial details to identify these repeats. Repeat identification is difficult, even under optimal conditions^{13, 52}, and there is little consensus agreement between methods¹⁶ however a consensus has emerged for identification of repeats¹⁷, namely that repeats have 1) evident self-homology, 2) compositional biases, and 3) a repeating pattern that is identifiable in the sequence and structure of the repeats. All of these are present in the ThDP enzymes. First, self-homology is quite evident from self-alignment of the ThDP sequences^{19, 45}. While there are not long stretches of regular patterns evident in the self-alignment plots, this is also not uncommon for many recognized repeat proteins and the previous analysis found that the amount of self-homology signal detected was well above the threshold amount expected as determined by comparison to other known repeat proteins¹⁸⁻¹⁹. Second, the ThDP repeats do have biased amino acid distributions (SI Figure S2). They show a preference for alanine, glycine, and proline and a bias against cysteine and lysine. While these preferences cannot be fully separated from the secondary structure propensities of these repeats (SI Table S8), the fact that these preferences were more evident as global sequence identity decreased suggests that they are an integral part of the architecture of these proteins as these preferences change more slowly than sequence does overall (SI Fig. S2). The fact that alanine occurs more often in helical sequences does not preclude it from being recognized as part of an identifying pattern in the ThDP repeats as there are conserved alanine residues identified in fully helical repeats⁵³. Third, repeating patterns, in both sequence and structure, are readily identifiable in the ThDP enzymes. While the most obvious of these is the (G/A){X(1,2)}(G/A) motif, there are definite preferences at other positions within the repeats as well. This includes the enrichment of isoleucine, leucine, and valine residues at the end (strand portion) of the repeat (as previously recognized⁴⁸) and the tendency of the turn portion of the repeat to be capped with proline and glycine residues (SI Table S4). It cannot be completely ruled out that these preferences are due to secondary structure constraints, however. And while no other

universally conserved position was detected, when the individual repeats are compared between the two functional domains, patterns emerge both within functional classes (SI Figures S3 & S4) and throughout the ThDP family of repeated sequence patterns within individual proteins. By measuring the distances from both ends of the repeats (N and C) we were able to account for differences in repeat lengths and identify conserved features such as the conserved alanine at position 10 in repeat 3 that is associated with the intra-chain contact between the PP and PYR domains (SI Figure S7) and the conserved residues at position 5 & 6 in repeat 5 that is associated with the ThDP phosphate/magnesium and an inter-chain interface (SI Figure S8). To the best of our knowledge, none of these conserved sequences have been previously noted, demonstrating the utility of recognizing the repeat architecture in the ThDP enzymes, given that these residues have orientations which are not obviously involved in the interactions but these sequences are strongly conserved in both the DH and DC branches of the family. While the conserved section of repeat 5 is easily identified in the DH enzymes, it occurs within what was identified as the insertion region between repeat 4 & 5 in the DC enzymes highlights some of the limitations of our ability to identify the limits of the repeats; even now it is unclear if the insertion after repeat 4 in the DC PP domain should be included as part of repeat 5 or not as there is no clear initiation sequence in these insertions near the conserved sequence (SI Figure S1) even though they appear to interact with the cofactor diphosphate. The conserved residues in repeat 5 also appear to be possibly involved in a cryptic magnesium binding site⁵⁴ (SI Figure 9) in the PYR domain, further suggesting the utility of identifying the repeat architecture of these enzymes, and perhaps the original, primitive function of these repeats. So, while the repeats themselves are quite diverged and sequence similarity is rather low, the identifying marks of self-homology, sequence bias, and repeating sequence and structure patterns are still notable suggesting that the modern ThDP enzymes still maintain a repeat architecture.

Identifying this repeat architecture also helps to simplify structural analysis of these enzymes, which has been previously quite complex²³. This complexity is partially due to breaks in the secondary structure, either due to natural structural drift or crystallographic assignments (Supplemental Figure S5), but recognition of the overarching repeat patterns alleviates this difficulty. The helix-turn-strand secondary structure pattern of the ThDP repeats (Supplemental Figure S6) also highlights many of the functions ThDP enzymes. Each of the three domains (PP, PYR, & CFX) has a unique function that is dependent on a specific repeat. The canonical repeat is always the fourth repeat in the PP domain (Fig. 1C) but it also cooperates with the conserved residues in repeat 5 and repeat 2 (at least in the DC branch and

the transketolases). This repeat also typically overlaps with the repeat preceding it (*e.g.* the GDG aspartate also functions as the terminal carbonyl residue of the previous repeat). The conserved, catalytic glutamate that mediates the protein interaction with the aminopyrimidine nitrogen³⁵ occurs just before the beginning of the third repeat. It bears noting that this repeat structure occurs even in the glyoxylate carboligases which are unique among the ThDP family in lacking the conserved glutamate that interacts with the aminopyrimidine moiety of the cofactor²⁸. Residues from these two functional repeats have also recently been shown to mediate catalysis and communicate through a hydrogen bond between two the paired monomers in transketolase which emphasizes their functional significance⁵⁵.

Both the PP and PYR domain have a universally conserved 213465 strand arrangement (Fig. 1, SI Fig. S6, SI Table S6) clearly indicating a common ancestral origin for these catalytic domains⁵⁶⁻⁵⁷. The CFX domains have a more diverse set of domain architectures with the members of the decarboxylase class having a Rossmann-like parallel 321456 strand arrangement and the dehydrogenases an antiparallel 1(-2)(-3)(-4) arrangement (Fig. 1, SI Fig. S6)^{23, 36, 38}. And while the functional ThDP domains have a similar strand arrangement to the Rossmann-like flavodoxins, the (G/A){X(1,2)}(G/A) motif that starts the ThDP repeats is similar but not identical to the motifs that start either the Rossmann (GxGxxG) or the Walker A (GxxGxG) proteins suggesting convergent evolution rather than common descent⁵⁸⁻⁵⁹. Additionally, some ThDP enzymes are known to bind other nucleotide-like ligands such as FAD³⁰ (which does not always have an obvious catalytic role⁶⁰) or bind ADP in a functional role (which was suggested to be derived from the FAD binding site³¹). The diphosphate of these ligands is bound by the second repeat in the CFX domain. FAD and ADP binding domains confined to the decarboxylase branch of the family with its Rossmann-like CFX domain, however they do not display the conserved geometry of the true Rossmann domains further indicating convergent evolution rather than common descent (SI Table S7)²⁶. More curious, however, are the two kinds of CFX domains which are exclusively present in the pyruvate-ferredoxin synthases/pyruvate synthases (PFOR) which have a PYR-CFX-PP domain arrangement similar to the decarboxylase branch, but domain architectures found in the dehydrogenase branch including a domain which differs from the dehydrogenase CFX domain by the addition/deletion of a single terminal repeat (SI Fig. S6, SI Table S6). These observations suggest that the PFORs may be the closest extant group to the root of the ThDP family, although we cannot rule out a polyphyletic origin⁶¹.

It bears noting here that the ThDP repeats are not necessary to bind thiamin diphosphate⁶²⁻⁶³ and some proteins involved in ThDP biosynthesis are apparent true

Rossmann proteins⁵⁶ indicating that the use of the ThDP repeat architecture was likely a stochastic rediscovery of a suitable phosphate binding structure⁴²⁻⁴³. That both the pyruvate dehydrogenases and the transketolases are widely dispersed among all three kingdoms of life, and both of these ThDP enzymes, as well as the possibly ancestral pyruvate synthases (PFOR) catalyze reactions believed to have been essential to life during the Archean eon^{8, 40-41} evidence the extreme age of this family. It is even more noteworthy in a family of enzymes that generally has low shared sequence homology³⁸ in which even catalytically essential residues can be mutated without major losses in enzymatic activity^{28, 64}, that we have been able to identify the universally conserved residues in repeats 3 and 5 by recognizing this architecture. That the ThDP repeats were conserved enough to be recognizable in addition to the observation of the conserved 213465 β -strand arrangement in the catalytic PP & PYR domains suggests that the repeat architecture is fundamental to the function and structure of these enzymes. All of these observations are consistent with the hypothesis that early proteins, such as the ancestor of the ThDP enzyme family, were repeat proteins derived from an assembly of short peptides^{2, 4-6}.

MATERIALS AND METHODS: Briefly, as described elsewhere¹⁹, protein chains in UniRef90²⁰ were analyzed using DOTTER⁴⁵. Protein chains with a threshold repeat signal from DOTTER were collated and compared pairwise using a Jaccard metric and then clustered using MCL⁶⁵. Clusters containing between 5 and 200 members (N=8569) were analyzed manually and 55 (0.64 %) of these were comprised exclusively of ThDP enzymes.

ThDP enzyme sequences were collected from the PDB⁴⁴ and redundancy was removed by CD-HIT⁶⁶ at 90% sequence identity. A multiple sequence alignment was generated from this set (N_{chains} = 58) using MUSCLE⁶⁷ and a phylogenetic tree was generated using MrBayes⁶⁸ (Fig. 2). Branches containing at least three leaves (N_{branches} = 10) were collected and then individual multiple sequence alignments were generated again using MUSCLE. The branches were named based on traditional functional groupings of ThDP enzymes. The members of these branches were analyzed by DOTTER and the DOTTER plots were decomposed to identify sections of sequence that generated self-similarity signal. The self-similarity score for any residue was defined as the maximum score in the in the row or column associated with that position. Plots of these DOTTER scores with sequence position and the multiple sequence alignments of the branches were used to identify the repeating sequence by looking for regions of conserved sequence identity. Putative repeats were then confirmed structurally using the protein structures as defined by PyMol⁶⁹. This lead to the detection of the

[G/A]{X(1,2)}[G/A] motif near the N terminus of a helix to start the repeat (a looser criterion than the “GDG” of the ThDP binding motif) and the set of non-redundant ThDP proteins was then manually analyzed. Because the repeats largely occurred in tandem, they were generally terminated by the appearance of the either the next [G/A]{X(1,2)}[G/A] motif or the start of another helix. When this was not possible, the borders were determined using the information available from the multiple sequence alignment and the protein structures. This identified a set of 1003 repeats which were then analyzed statistically to generate a “typical” repeat (Fig. S1, SI Table S1). The central 65% (from 20-85% of the distribution, Figure 4) was used to identify the repeat length. Sequence and structure propensities were calculated from the set of all the identified repeats. Domain architecture was identified manually using PyMol⁶⁹. The ligand binding geometry for identification of true Rossmann domains was analyzed as described elsewhere²⁶.

The non-redundant set of ThDP enzymes of known structure was compared to the TEED database²⁹ by pairwise global sequence alignment using the BioConductor suite in R⁷⁰. Sequences from the database were sorted based on their identity to each of the ThDP enzymes at cutoffs of 80, 70, 62, 50, 40, & 30 % identity. Within each set, for each repeat, the matching sequence was identified and recorded to produce a 21 member (all 20 amino acids and gap) matrix for each position in the repeat. At each position the degree of conservation was quantitated using the Simpson metric⁴⁹⁻⁵⁰. The amino acid distribution for each identity and Simpson threshold was also calculated. The distribution of amino acids for each position in the TEED set was calculated and used to formulate consensus repeats. The repeats from the functional PP and PYR domains were compared, with each position analyzed relative to its distance from either terminus to account for the variation in repeat lengths. Due to the bias in the TEED database towards transketolases, the repeats were generally analyzed within their functional subgroups.

The non-redundant PDB chains were also subject to other methods of *de novo* repeat identification with RADAR⁷¹, T-REKS⁷², HHREP⁵², XSTREAM⁷³, and TRUST⁷⁴ which gave weak results that were nevertheless consistent with our method (Supplemental Data S1). The ThDP repeat was used to search the set of all known protein structures using the sequence and secondary structure as defined by the PDB in Nov. 2019. The repeat was defined as [G/A][X][G/A] where X could be any one or two amino acids and [G/A] was either a glycine or alanine residue followed by a helical segment and then a beta strand segment where the last residue of the strand or the residue after the end of the strand was a glutamate, glutamine, aspartate, or asparagine. The helical region was allowed to include secondary structure

“H”, “G” or “I” while the strand was “B” or “E”. In order to match the sequence preferences, helical regions were also required to have at least 9% alanine content and strand regions must have contained at least a 20% total of valine, leucine or isoleucine residues. The length of the repeat must also have been 14-36 residues long and at least three had to occur within one protein chain for the chain to be identified as repeat containing. This identified 953 PDB structures (72 ThDP structures and 881 non-ThDP structures) which represented 0.56% of the PDB at the time (Supplemental Data S2).

ACKNOWLEDGEMENTS: The work was supported by the National Science Centre, Poland [grant agreement 2014/15/D/NZ1/00968] and the European Union’s Horizon 2020 research and innovation programme [Marie Skłodowska-Curie grant agreement No 655075]. The authors wish to thank Profs. Ben Luisi & Vikas Nanda for critical readings of the manuscript. Computer code for this work is available online at <https://gorna.uw.edu.pl/en/research/software>. The authors declare that they have no competing financial interests regarding this work.

REFERENCES:

- Goldman, A. D.; Beatty, J. T.; Landweber, L. F., The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J Mol Evol* **2016**, *82* (1), 17-26.
- Romero, M. L. R.; Rabin, A.; Tawfik, D. S., Functional Proteins from Short Peptides: Dayhoff's Hypothesis Turns 50. *Angew Chem Int Edit* **2016**, *55* (52), 15966-15971.
- Lohse, P. A.; Szostak, J. W., Ribozyme-catalysed amino-acid transfer reactions. *Nature* **1996**, *381* (6581), 442-444.
- Soding, J.; Lupas, A. N., More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **2003**, *25* (9), 837-846.
- Alva, V.; Soding, J.; Lupas, A. N., A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **2015**, *4*.
- Broom, A.; Doxey, A. C.; Lobsanov, Y. D.; Berthin, L. G.; Rose, D. R.; Howell, P. L.; McConkey, B. J.; Meiering, E. M., Modular Evolution and the Origins of Symmetry: Reconstruction of a Three-Fold Symmetric Globular Protein. *Structure* **2012**, *20* (1), 161-171.
- Andrade, M. A.; Perez-Iratxeta, C.; Ponting, C. P., Protein repeats: Structures, functions, and evolution. *J Struct Biol* **2001**, *134* (2-3), 117-131.
- Weiss, M. C.; Sousa, F. L.; Mrnjavac, N.; Neukirchen, S.; Roettger, M.; Nelson-Sathi, S.; Martin, W. F., The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **2016**, *1* (9).
- Schaper, E.; Gascuel, O.; Anisimova, M., Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes. *Mol Biol Evol* **2014**, *31* (5), 1132-1148.
- Kajava, A. V., Review: Proteins with repeated sequence - Structural prediction and modeling. *J Struct Biol* **2001**, *134* (2-3), 132-144.
- Knoll, A. H.; Javaux, E. J.; Hewitt, D.; Cohen, P., Eukaryotic organisms in Proterozoic oceans. *Philos T R Soc B* **2006**, *361* (1470), 1023-1038.
- Smith, T. F.; Gaitatzes, C.; Saxena, K.; Neer, E. J., The WD repeat: a common architecture for

- diverse functions. *Trends Biochem Sci* **1999**, *24* (5), 181-185.
13. Pellegrini, M.; Renda, M. E.; Vecchio, A., Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. *Bmc Bioinformatics* **2012**, *13*.
14. Marcotte, E. M.; Pellegrini, M.; Yeates, T. O.; Eisenberg, D., A census of protein repeats. *J Mol Biol* **1999**, *293* (1), 151-160.
15. Marold, J. D.; Kavran, J. M.; Bowman, G. D.; Barrick, D., A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. *Structure* **2015**, *23* (11), 2055-2065.
16. Schaper, E.; Kajava, A. V.; Hauser, A.; Anisimova, M., Repeat or not repeat?-Statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Res* **2012**, *40* (20), 10005-10017.
17. Luo, H.; Nijveen, H., Understanding and identifying amino acid repeats. *Brief Bioinform* **2014**, *15* (4), 582-591.
18. Paladin, L.; Hirsh, L.; Piovesan, D.; Andrade-Navarro, M. A.; Kajava, A. V.; Tosatto, S. C. E., RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures (vol 45, pg D308, 2016). *Nucleic Acids Res* **2017**, *45* (6), 3613-3613.
19. Merski, M.; Mlynarczyk, K.; Ludwiczak, J.; Skrzeczkowski, J.; Dunin-Horkawicz, S.; Gorna, M. W., Self-analysis of repeat proteins reveals evolutionarily conserved patterns. *Bmc Bioinformatics* **2020**, *21* (1).
20. Bateman, A.; Martin, M. J.; Orchard, S.; Magrane, M.; Alpi, E.; Bely, B.; Bingley, M.; Britto, R.; Bursteinas, B.; Busiello, G.; Bye-A-Jee, H.; Da Silva, A.; De Giorgi, M.; Dogan, T.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, D.; Gonzales, L.; Hatton-Ellis, E.; Ignatchenko, A.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lopez, R.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Nightingale, A.; Onwubiko, J.; Palka, B.; Pichler, K.; Pundir, S.; Qi, G. Y.; Raj, S.; Renaux, A.; Lopez, M. R.; Saidi, R.; Sawford, T.; Shypitsyna, A.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Wardell, T.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Xenarios, I.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M. C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Coudert, E.; Cucho, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Lemercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C. M.; Chen, Y. X.; Cowart, J.; Garavelli, J. S.; Huang, H. Z.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q. H.; Wang, Y. Q.; Yeh, L. S.; Zhang, J.; Consortium, U., UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **2019**, *47* (D1), D506-D515.
21. Jordan, F.; Nemeria, N. S., Experimental observation of thiamin diphosphate-bound intermediates on enzymes and mechanistic information derived from these observations. *Bioorg Chem* **2005**, *33* (3), 190-215.
22. Schellenberger, A., Sixty years of thiamin diphosphate biochemistry. *Bba-Protein Struct M* **1998**, *1385* (2), 177-186.
23. Duggleby, R. G., Domain relationships in thiamine diphosphate-dependent enzymes. *Accounts Chem Res* **2006**, *39* (8), 550-557.
24. Frank, R. A. W.; Leeper, F. J.; Luisi, B. F., Structure, mechanism and catalytic duality of thiamine-dependent enzymes. *Cell Mol Life Sci* **2007**, *64* (7-8), 892-905.
25. Fox, N. K.; Brenner, S. E.; Chandonia, J. M., SCOPe: Structural Classification of Proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* **2014**, *42* (D1), D304-D309.
26. Laurino, P.; Toth-Petroczy, A.; Meana-Paneda, R.; Lin, W.; Truhlar, D. G.; Tawfik, D. S., An Ancient Fingerprint Indicates the Common Ancestry of Rossmann-Fold Enzymes Utilizing Different Ribose-Based Cofactors. *Plos Biol* **2016**, *14* (3).
27. Jordan, F., Current mechanistic understanding of thiamin diphosphate-dependent enzymatic

- reactions. *Nat Prod Rep* **2003**, 20 (2), 184-201.
28. Kaplun, A.; Binshtein, E.; Vyazmensky, M.; Steinmetz, A.; Barak, Z.; Chipman, D. M.; Tittmann, K.; Shaanan, B., Glyoxylate carboligase lacks the canonical active site glutamate of thiamine-dependent enzymes. *Nat Chem Biol* **2008**, 4 (2), 113-118.
29. Widmann, M.; Radloff, R.; Pleiss, J., The Thiamine diphosphate dependent Enzyme Engineering Database: A tool for the systematic analysis of sequence and structure relations. *Bmc Biochem* **2010**, 11.
30. Berthold, C. L.; Moussatche, P.; Richards, N. G. J.; Lindqvist, Y., Structural basis for activation of the thiamin diphosphate-dependent enzyme oxalyl-CoA decarboxylase by adenosine diphosphate. *J Biol Chem* **2005**, 280 (50), 41645-41654.
31. Chipman, D. M.; Duggleby, R. G.; Tittmann, K., Mechanisms of acetohydroxyacid synthases. *Curr Opin Chem Biol* **2005**, 9 (5), 475-481.
32. Whitley, M. J.; Arjunan, P.; Nemeria, N. S.; Korotchikina, L. G.; Park, Y. H.; Patel, M. S.; Jordan, F.; Furey, W., Pyruvate dehydrogenase complex deficiency is linked to regulatory loop disorder in the alpha V138M variant of human pyruvate dehydrogenase. *J Biol Chem* **2018**, 293 (34), 13204-13213.
33. Schenk, G.; Layfield, R.; Candy, J. M.; Duggleby, R. G.; Nixon, P. F., Molecular evolutionary analysis of the thiamine-diphosphate-dependent enzyme, transketolase. *J Mol Evol* **1997**, 44 (5), 552-572.
34. Hasson, M. S.; Muscate, A.; McLeish, M. J.; Polovnikova, L. S.; Gerlt, J. A.; Kenyon, G. L.; Petsko, G. A.; Ringe, D., The crystal structure of benzoylformate decarboxylase at 1.6 angstrom resolution: Diversity of catalytic residues in thiamin diphosphate-dependent enzymes. *Biochemistry-U S* **1998**, 37 (28), 9918-9930.
35. Meyer, D.; Neumann, P.; Parthier, C.; Friedemann, R.; Nemeria, N.; Jordan, F.; Tittmann, K., Double Duty for a Conserved Glutamate in Pyruvate Decarboxylase: Evidence of the Participation in Stereoelectronically Controlled Decarboxylation and in Protonation of the Nascent Carbanion/Enamine Intermediate. *Biochemistry-U S* **2010**, 49 (37), 8197-8212.
36. Vogel, C.; Pleiss, J., The modular structure of ThDP-dependent enzymes. *Proteins* **2014**, 82 (10), 2523-2537.
37. Fullam, E.; Pojer, F.; Bergfors, T.; Jones, T. A.; Cole, S. T., Structure and function of the transketolase from Mycobacterium tuberculosis and comparison with the human enzyme. *Open Biol* **2012**, 2.
38. Vogel, C.; Widmann, M.; Pohl, M.; Pleiss, J., A standard numbering scheme for thiamine diphosphate-dependent decarboxylases. *Bmc Biochem* **2012**, 13.
39. Hawkins, C. F.; Borges, A.; Perham, R. N., A Common Structural Motif in Thiamin Pyrophosphate-Binding Enzymes. *Febs Lett* **1989**, 255 (1), 77-82.
40. Keller, M. A.; Turchyn, A. V.; Ralser, M., Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. *Mol Syst Biol* **2014**, 10 (4).
41. Fuchs, G., Alternative Pathways of Carbon Dioxide Fixation: Insights into the Early Evolution of Life? *Annu Rev Microbiol* **2011**, 65, 631-+.
42. Mande, S. S.; Sarfaty, S.; Allen, M. D.; Perham, R. N.; Hol, W. G. J., Protein-protein interactions in the pyruvate dehydrogenase multienzyme complex: Dihydrolipoamide dehydrogenase complexed with the binding domain of dihydrolipoamide acetyltransferase. *Structure* **1996**, 4 (3), 277-286.
43. Schenk, G.; Duggleby, R. G.; Nixon, P. F., Properties and functions of the thiamin diphosphate dependent enzyme transketolase. *Int J Biochem Cell B* **1998**, 30 (12), 1297-1318.
44. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000**, 28 (1), 235-242.
45. Kaisers, W. *seqTools: Analysis of nucleotide, sequence and quality content on fastq files*, 2019.
46. Wall, M. A.; Coleman, D. E.; Lee, E.; Iniguezllui, J. A.; Posner, B. A.; Gilman, A. G.; Sprang, S. R., The Structure of the G-Protein Heterotrimer G(I-Alpha-1)Beta(1)Gamma(2). *Cell* **1995**, 83 (6), 1047-1058.
47. Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C. X.; Chen, L.; Di Costanzo, L.; Christie, C.;

- Dalénberg, K.; Duarte, J. M.; Dutta, S.; Feng, Z. K.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Kalro, T.; Liang, Y. H.; Lowe, R.; Namkoong, H.; Peisach, E.; Periskova, I.; Prlic, A.; Randle, C.; Rose, A.; Rose, P.; Sala, R.; Sekharan, M.; Shao, C. H.; Tan, L. H.; Tao, Y. P.; Valasatava, Y.; Voigt, M.; Westbrook, J.; Woo, J.; Yang, H. W.; Young, J.; Zhuravleva, M.; Zardecki, C., RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* **2019**, *47* (D1), D464-D474.
48. Zhang, G. F.; Dai, J. Y.; Lu, Z. B.; Dunaway-Mariano, D., The phosphonopyruvate decarboxylase from *Bacteroides fragilis*. *J Biol Chem* **2003**, *278* (42), 41302-41308.
49. Beaugrand, G.; Edwards, M., Differences in performance among four indices used to evaluate diversity in planktonic ecosystems. *Oceanol Acta* **2001**, *24* (5), 467-477.
50. Lande, R., Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* **1996**, *76* (1), 5-13.
51. Dyda, F.; Furey, W.; Swaminathan, S.; Sax, M.; Farrenkopf, B.; Jordan, F., Preliminary Crystallographic Data for the Thiamin Diphosphate-Dependent Enzyme Pyruvate Decarboxylase from *Brewers-Yeast*. *J Biol Chem* **1990**, *265* (29), 17413-17415.
52. Biegert, A.; Soding, J., De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* **2008**, *24* (6), 807-814.
53. D'Andrea, L. D.; Regan, L., TPR proteins: the versatile helix. *Trends Biochem Sci* **2003**, *28* (12), 655-662.
54. Lu, X. Y.; Liu, Y. W.; Yang, Y. Q.; Wang, S. S.; Wang, Q.; Wang, X. Y.; Yan, Z. H.; Cheng, J.; Liu, C.; Yang, X.; Luo, H.; Yang, S.; Gou, J. R.; Ye, L. Z.; Lu, L. N.; Zhang, Z. D.; Guo, Y.; Nie, Y.; Lin, J. P.; Li, S.; Tian, C. G.; Cai, T.; Zhuo, B. Z.; Ma, H. W.; Wang, W.; Ma, Y. H.; Liu, Y. J.; Li, Y.; Jiang, H. F., Constructing a synthetic pathway for acetylcoenzyme A from one-carbon through enzyme design. *Nat Commun* **2019**, *10*.
55. Dai, S.; Funk, L. M.; von Pappenheim, F. R.; Sautner, V.; Paulikat, M.; Schroder, B.; Uranga, J.; Mata, R. A.; Tittmann, K., Low-barrier hydrogen bonds in enzyme cooperativity. *Nature* **2019**, *573* (7775), 609-+.
56. Medvedev, K. E.; Kinch, L. N.; Schaeffer, R. D.; Grishin, N. V., Functional analysis of Rossmann-like domains reveals convergent evolution of topology and reaction pathways. *Plos Comput Biol* **2019**, *15* (12).
57. Burroughs, A. M.; Allen, K. N.; Dunaway-Mariano, D.; Aravind, L., Evolutionary genomics of the HAD superfamily: Understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J Mol Biol* **2006**, *361* (5), 1003-1034.
58. Aravind, L.; Mazumder, R.; Vasudevan, S.; Koonin, E. V., Trends in protein evolution inferred from sequence and structure analysis. *Curr Opin Struc Biol* **2002**, *12* (3), 392-399.
59. Longo, L. M.; Jablonska, J.; Vyas, P.; Kanade, M.; Kolodny, R.; Ben-Tal, N.; Tawfik, D. S., On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. *Elife* **2020**, *9*.
60. Tittmann, K.; Schroder, K.; Golbik, R.; McCourt, J.; Kaplun, A.; Duggleby, R. G.; Barak, Z.; Chipman, D. M.; Hubner, G., Electron transfer in acetohydroxy acid synthase as a side reaction of catalysis. implications for the reactivity and partitioning of the carbanion/enamine form of (alpha-hydroxyethyl)thiamin diphosphate in a "nonredox" flavoenzyme. *Biochemistry-Us* **2004**, *43* (27), 8652-8661.
61. Lupas, A. N.; Ponting, C. P.; Russell, R. B., On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol* **2001**, *134* (2-3), 191-203.
62. Soriano, E. V.; Rajashankar, K. R.; Hanes, J. W.; Bale, S.; Begley, T. P.; Ealick, S. E., Structural similarities between thiamin-binding protein and thiaminase-I suggest a common ancestor. *Biochemistry-Us* **2008**, *47* (5), 1346-1357.
63. Sippel, K. H.; Robbins, A. H.; Reutzel, R.; Boehlein, S. K.; Namiki, K.; Goodison, S.; Agbandje-McKenna, M.; Rosser, C. J.; McKenna, R., Structural Insights into the Extracytoplasmic Thiamine-

Binding Lipoprotein p37 of *Mycoplasma hyorhinis*. *J Bacteriol* **2009**, *191* (8), 2585-2592.

64. Yep, A.; Kenyon, G. L.; McLeish, M. J., Saturation mutagenesis of putative catalytic residues of benzoylformate decarboxylase provides a challenge to the accepted mechanism. *P Natl Acad Sci USA* **2008**, *105* (15), 5733-5738.

65. Azad, A.; Pavlopoulos, G. A.; Ouzounis, C. A.; Kyrpides, N. C.; Buluc, A., HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic Acids Res* **2018**, *46* (6).

66. Huang, Y.; Niu, B. F.; Gao, Y.; Fu, L. M.; Li, W. Z., CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26* (5), 680-682.

67. Edgar, R. C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **2004**, *32* (5), 1792-1797.

68. Ronquist, F.; Huelsenbeck, J. P., MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **2003**, *19* (12), 1572-1574.

69. Schrodinger, L. *The PyMOL Molecular Graphics System*.

70. Huber, W.; Carey, V. J.; Gentleman, R.; Anders, S.; Carlson, M.; Carvalho, B. S.; Bravo, H. C.; Davis, S.; Gatto, L.; Girke, T.; Gottardo, R.; Hahne, F.; Hansen, K. D.; Irizarry, R. A.; Lawrence, M.; Love, M. I.; MacDonald, J.; Obenchain, V.; Oles, A. K.; Pages, H.; Reyes, A.; Shannon, P.; Smyth, G. K.; Tenenbaum, D.; Waldron, L.; Morgan, M., Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **2015**, *12* (2), 115-121.

71. Madeira, F.; Park, Y. M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A. R. N.; Potter, S. C.; Finn, R. D.; Lopez, R., The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* **2019**, *47* (W1), W636-W641.

72. Jorda, J.; Kajava, A. V., T-REKS: identification of Tandem REpeats in sequences with a K-means based algorithm. *Bioinformatics* **2009**, *25* (20), 2632-2638.

73. Newman, A. M.; Cooper, J. B., XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *Bmc Bioinformatics* **2007**, *8*.

74. Szklarczyk, R.; Heringa, J., Tracking repeats using significance and transitivity. *Bioinformatics* **2004**, *20*, 311-317.

FIGURES:

FIGURE 1:

Illustration of the repeat architecture of ThDP enzymes. **A)** The canonical ThDP repeat from a pyruvate decarboxylase (PDB 2nxw) showing the helix-turn-strand structure along with the ThDP cofactor. The canonical N-terminal GDG and C-terminal “NN” residues are shown as sticks along with the conserved glycine and proline identified in the ThDP repeats. **B)** The repeats that make up the PP domain in a pyruvate dehydrogenase (PDB 1umb) colored in order from N to C red, yellow, green, cyan, blue and magenta. The ThDP cofactor is shown as sticks and colored as in panel A. Likewise the canonical ThDP binding residues (the GDG aspartate (D175) and the asparagine (N203)) are shown as sticks in the cyan color of their associated repeat. **C)** The arrangement of secondary structures in an idealized ThDP domain

644 showing its tandem repeat structure. The repeats are colored in the same order as before and
 645 the position of the ligand interacting repeat in each of the domains is identified along with the
 646 positions of the canonical ThDP motif sequence in repeat four and the conserved catalytic
 647 glutamate in repeat three. Identically colored cartoon diagrams of the domain architectures for
 648 the **D)** PP & PYR domains, **E)** decarboxylase branch ThDP enzymes CFX domain and **F)**
 649 dehydrogenase branch ThDP enzymes CFX domain. A larger, more detailed version of this is
 650 given as SI Figure S6.

FIGURE 2:

Phylogenetic tree of the ThDP family enzymes. A phylogenetic tree of the ThDP enzymes of known structure. Groups of more than 3 members were separated into different groupings indicated by color as follows: AHAS (red), ALS (dark blue), BFD (magenta), OxCDC (dark green), PDC (cyan), PDH (green), PhK (light blue), PFOR (grey), POX (beige), & TK (mustard). The PDH group of proteins consists of two separately expressed chains which do not co-cluster in this tree but are both indicated by the same color here.

FIGURE 3A:

FIGURE 3B:

Illustration of the protein repeats present in the pyruvate oxidases cluster of ThDP enzymes. A) The repeats are highlighted in the order red, yellow, green, cyan, blue, and magenta for each of the three domains respectively. The three state secondary structure (α -helix (light blue), β -strand (plum), coil (dash)) for the repeat regions in PDB 3ey9 is shown above the multiple sequence alignment for reference. The conserved alanine in position 10 of repeat 3 and the conserved sequence in position 5-6 in repeat 5 is shown by a red box. In the DC enzymes, the conserved region in repeat 5 of the PP domain appears within the insertion between repeat 4 and 5. **B)** Cartoon illustration of the protein repeats present in the POX group of enzymes using the representative PDB 1y9d. The repeats are numbered and colored in the order red, yellow, green, cyan, blue, and magenta from N to C as in figure 1. Helical regions are indicated by a cylindrical tube and strand regions by an arrow. The number of the first helical residue in the helical region and the last beta residue in the strand region are also indicated. Insertions that contain secondary structure are indicated in grey while missing repeats are indicated by missing cartoon images. The three domains are shown in the order they occur in the protein. Secondary structure indication is derived from PyMol. Note that the helical region in the fifth repeat of the CFX domain is not resolved in the structure so it is colored grey here. Equivalent figures for a representative example from each of the other enzyme groups is given as SI Figure S5.

FIGURE 4:

Size distribution of the ThDP repeats identified from the filtered PDB set. The number of repeats of each sequence length is shown for all the domains (black circles, N = 37), the diphosphate binding PP domain (red diamonds, N = 36), aminopyrimidine interacting PYR (green squares, N = 27), and extra third domain (cyan triangles, N = 25). The typical (66%) repeat length of 22-32 residues is indicated by a set of vertical, dashed black lines.