On the Emergence of P-Loop NTPase and Rossmann Enzymes from a Beta-Alpha-Beta Ancestral Fragment

- This article is dedicated to the memory of Michael G. Rossmann

Liam M. Longo[1,2,3], Jagoda Jabłońska[1], Pratik Vyas[1], Manil Kanade[1,4], Rachel Kolodny[5,*], Nir Ben-Tal[6,*], Dan S. Tawfik[1,*]

[1]Weizmann Institute of Science, Department of Biomolecular Sciences, Rehovot, Israel

[2]Current address: Tokyo Institute of Technology, Earth-Life Science Institute, Tokyo, Japan

[3]Current address: Blue Marble Space Institute of Science, Seattle, USA

[4]Current address: Sincrotrone Trieste S.C.p.A., Trieste, Italy

[5]University of Haifa, Department of Computer Science, Haifa, Israel

[6]Tel Aviv University, George S. Wise Faculty of Life Sciences, Department of Biochemistry and Molecular Biology, Tel Aviv, Israel

[*]To whom correspondence should be addressed: trachel@cs.haifa.ac.il, bental@tauex.tau.ac.il, dan.tawfik@weizmann.ac.il

Abstract

Dating back to the last universal common ancestor (LUCA), the P-loop NTPases and Rossmanns now comprise the most ubiquitous and diverse enzyme lineages. Intriguing similarities in their overall architecture and phosphate binding motifs suggest common ancestry; however, due to a lack of sequence identity and some fundamental structural differences, these families are considered independent emergences. To address this longstanding dichotomy, we systematically searched for 'bridge proteins' with structure and sequence elements shared by both lineages. We detected homologous segments that span the first βαβ segment of both lineages and include two key functional motifs: (*i*) a phosphate binding loop – the 'Walker A' motif of P-loop NTPases or the Rossmann equivalent, both residing at the N-terminus of α1; and (*ii*) an Asp at the tip of β2. The latter comprises the 'Walker B' aspartate that chelates the catalytic metal in P-loop NTPases, or the canonical Rossmann β2-Asp that binds the cofactor's ribose moiety. Tubulin, a Rossmann GTPase, demonstrates the potential of the β2-Asp to take either one of these two roles. We conclude that common P-loops/Rossmann ancestry is plausible, although convergence cannot be completely ruled out. Regardless, both lineages most likely emerged from a polypeptide comprising a βαβ segment carrying the above two functional motifs, a segment that comprises the core of both enzyme families to this very day.

Introduction

In 1970 Michael Rossmann reported the structure of the first αβα sandwich protein, lactate dehydrogenase[1]. This NAD-utilizing enzyme would later become representative of what is now known as the '*Rossmann fold*'[2]. About a decade later, on the basis of a sequence analysis, another major αβα sandwich domain that utilizes phosphorylated nucleosides was proposed[3], which is now known as the P-loop NTPase, or '*P-loop*' for short. The importance of these two evolutionary lineages, Rossmanns and P-loops, cannot be overstated: Both lineages have diversified extensively, and each is individually associated with more than 120 families and 75 different enzymatic reactions (see *Methods*). Furthermore, these two lineages are ubiquitous across the tree of life[4]. Accordingly, essentially all studies aimed at unraveling the history of protein evolution concluded that these enzymes emerged well before the last universal common ancestor (LUCA), and were among the very first, if not the first, enzyme families[4–9]. Indeed, both P-loops and Rossmanns are dubbed nucleotide binding domains because they both make use of phosphorylated ribonucleosides such as ATP or NAD, as well as of other pre-LUCA cofactors such as SAM[10].

As elaborated in the next section, the P-loop and Rossmann domains share a number of similar features, but also some distinct differences. Given their pre-LUCA origin, a common P-loop/Rossmann ancestor – even if it did exist at some point – is surely lost to time. Both lineages emerged during the so-called "big bang" of protein evolution, an event that marks the birth of the major protein classes, yet occurred too early to be reconstructed by phylogenetic means[8]. Thus, a fundamental enigma surrounding the birth of the first enzymes is whether the Rossmann and the P-loop lineages diverged from a common ancestor, or perhaps, given that they both make use of phosphorylated ribonucleosides, have converged to similar structural and functional features. The former is the more evolutionarily appealing scenario, yet the latter is as common and tangible[11,12].

To address this longstanding question, we performed a detailed analysis looking for indications of common ancestry with respect to the core elements of these two classes, namely their most conserved and functionally critical structural elements. Indeed, global sequence homology between these lineages, or even shared short sequence motifs, cannot be detected. As such, large-scale analyses of protein homology[7], including SCOPe[13] and the Evolutionary Classifications of Protein Domains (ECOD) database[14] classify P-loop NTPases and Rossmanns as independent evolutionary emergences. However, loss of detectable sequence homology would be expected between lineages that split in the distant past, especially if both have diverged extensively, as is the case for P-loops and Rossmanns. Nonetheless, structural anatomy[10] and sophisticated ways of detecting sequence homology may assign common ancestry in highly diverged lineages on the basis of a few common sequence-structure features[11,12,15,16]. Further, parallel evolution may operate, with relics of an ancient common ancestor surfacing sporadically in contemporary proteins, thus resulting in detectable sequence and/or structural homology. Thus, if P-loops and Rossmanns do share common ancestry, we might expect the existence of 'bridge proteins'; that is, proteins belonging to one lineage with features that are distinct for the other lineage.

Here, we report the detection and analysis of common features and bridge proteins between P-loops and Rossmanns. The existence of such common features and bridge proteins supports common ancestry, though it does not rule out convergence. Nonetheless, our results suggest what the key features of the ancestor(s) might have been, and indicate that even if these lineages emerged independently, their ancestors shared the very same features. To best frame this analysis, however, we must first dissect the canonical features of Rossmann and P-loop proteins.

P-loop and Rossmann – similar but distinctly different

Both P-loops and Rossmanns adopt the αβα 3-layer sandwich fold (**Figure 1**). This fold, which comprises a parallel β-sheet sandwiched between two layers of α-helices, is among the most ancient, if not the most ancient, protein folds known[5,8,17,18]. In essence, αβα sandwich proteins consist of a tandem repeat of β-loop-α elements, where the loops form the active-site (hereafter referred to as the "functional" or "top" loops; **Figure 1A**). The minimal P-loop or Rossmann domain comprises five β-loop-α elements linked via short "connecting" or "bottom" loops that generally have no functional role. Although many domains have six strands, and sometimes more, we will hereafter consider the minimal 5-stranded core domain for simplicity.

While the overall fold is conserved, the topology – specifically, the strand order of the interior β-sheet – differs between Rossmanns and P-loops. The Rossmann topology (β3-β2-β1-β4-β5) has a pseudo-2-fold axis of symmetry between β1-β3 and β4-β5 (**Figure 1B**; or β1-β3 and β4-β6 in the common 6-stranded domains). However, in the P-loop topology, at least two strands are swapped (**Figure 1C)**. The most common P-loop topology is β2-β3-β1-β4-β5 (ref. [4]); although, as discussed below, P-loops can adopt several different strand topologies.

The second shared hallmark is that both P-loops and Rossmans bind phosphorylated ribonucleoside ligands as substrates, co-substrates or cofactors (hereafter, phospho-ligands). While the overall binding mode of phospho-ligands differs, the binding modes of their phosphate moieties share a few similarities (**Figure 2A**,**B**): (*i*) The phosphate is bound by the first β-loop-α element which resides in the center of the domain (hereafter, β1-(phosphate binding loop)-α1); (*ii*) both phosphate binding loops mediate binding via a "nest" of hydrogen bonds formed by backbone amides at the N-terminus of the first canonical α-helix (α1) as well as via residues from the loop itself; and (*iii*) both phosphate binding loops are glycine-rich sequences with similar patterns: the canonical Rossmann motif is GxGxxG,

while the canonical P-loop motif, dubbed Walker A, is GxxGxGK[T/S]. To avoid confusion, P-loop is used here to refer to the evolutionary lineage of P-loop NTPases only. When referring to the phosphate binding element of a protein, with no relation to a specific protein lineage, phosphate binding loop (or PBL) is used. Hence, P-loop PBL relates to the phosphate binding loop of P-loop NTPases (the Walker A motif), and Rossmann PBL to the Rossmann's phosphate binding loop. The structural segment in which the phosphate binding loop resides is accordingly dubbed β1-PBL-α1, or β-PBL-α for simplicity.

However, despite similar phosphate binding elements, the mode of phospho-ligand binding by Rossmanns and P-loops is fundamentally different, and this difference relates to important functional differences between the two lineages. Although Rossmann and P-loop proteins both utilize phosphorylated nucleosides, the phosphate groups of these metabolites play a fundamentally different role. P-loops primarily catalyze phosphoryl transfer (including to water, *i.e.*, hydrolysis) and thus most often operate on ATP and GTP with the help of a metal dication. Rossmanns, on the other hand, primarily use NAD(P), with the phosphate moieties serving only as a handle for binding, while the redox chemistry occurs elsewhere (*e.g.*, the nicotinamide base in NAD^+). These functional differences are accompanied by a number of structural differences in the mode of phosphate binding: The P-loop Walker A is a relatively long, surface-exposed loop that extends beyond the protein's core and wraps, like the palm of a hand, around the phosphate moieties of the ligand (**Figure 2A**). The Rossmann PBL, however, is short and forms a flat interaction surface, with the phosphate groups interacting mostly with the N-terminus of α1 via a highly conserved and ordered water molecule (**Figure 2B**)[19]. Foremost, the orientation of the phospho-ligand being bound is different: The nucleoside moiety in Rossmanns is oriented "inside", *i.e.*, in the direction of the β-sheet core, whereas in P-loops it points "outside", *i.e.*, away from the protein interior – an approximately 180-degree rotation compared to Rossmanns.

The above difference in orientation relates to differences in the interactions that Rossmanns and P-loops make with parts of the ribonucleotide ligands other than their phosphate moieties. In the canonical Rossmann binding site, both the phosphate moiety and the ribose moiety are bound. The phosphate interacts with the Rossmann PBL at the N-terminus of α1 while the ribose moiety is held in place by an Asp/Glu residue at the tip of β2 (**Figure 2B**). This acidic residue forms a unique bidentate interaction with the 2' and 3' hydroxyls of the ribose moiety, and was shown to be present as Asp in the earliest Rossmann ancestor (hereafter β2-Asp)[10]. In P-loops, on the other hand, the core of the αβα domain does not interact with the ribose, instead making more extensive, catalytically-oriented interactions with the phosphate moieties (via the Walker A P-loop, **Figure 2A**, as well as other key residues). Foremost, phospho-ligand binding also involves a metal cation, mostly Mg^{2+}, but also Ca^{2+}, by two key conserved residues: the hydroxyl of the canonical serine or threonine of the Walker A motif (GxxGxGK[**S/T**]) and an Asp/Glu residing on the tip of an adjacent β-strand. This Asp/Glu residue is the crux of the so-called "Walker B" motif, which is typically located at the tip of either β3 or β4 (see Ref. [20] for a detailed analysis).

A shared β-(phosphate binding loop)-α evolutionary seed?

Individually, any one of the shared features described above may relate to convergence. The αβα sandwich fold has likely emerged multiple times, independently[21]. The key shared functional feature, namely the phosphate binding site at the N-terminus of an α-helix, and the Gly-rich phosphate binding motifs, were likely favored at the early stages protein evolution because they effectively comprise the only mode of phosphate binding that can be realized with short and simple peptides[22]. Thus, that Rossmanns and P-loops share Gly-rich loops, and the same mode of phosphate biding, may also be the outcome of convergence, especially

because the overall mode of binding of their phospho-ligands fundamentally differs (**Figure 2A**, **B**).

Curiously, however, the phosphate binding site is located in the first β-loop-α element of both Rossmanns and P-loops. In fact, the β1-α1 location of the PBL is seen not only in P-loop and Rossmann proteins, but also in Rossmann-like protein classes such as flavodoxin and HUP. However, a closer examination reveals that, although rare, phosphate binding in αβα sandwich folds can occur at alternative locations, suggesting that there is no inherent, physical constraint on its location. An illustrative example can be found in the HUP lineage (ECOD X-group 2005) – a monophyletic group of 3-layer αβα sandwich, Rossmann-like proteins that includes Class I aminoacyl tRNA synthetases[23]. Most families within this lineage achieve phosphate binding at the tip of α1, as do Rossmann and P-loop proteins. However, two families, the universal stress protein (Usp) family (F-group 2005.1.1.145) and electron transport flavoprotein (ETF; F-group 2005.1.1.132) both use the tip of α4 (**Figure 3**). Intriguingly, α4, resides on the other side of the β-sheet, just opposite to α1. Accordingly, this change in the PBL's location results in a flip of ATP's phosphate groups, while preserving all other features of the binding site, including the adenine's location and the anchoring of the ribose moiety to β1 and β4 (**Figure 3** mid panel). Thus, from a purely biophysical point of view, α1 and α4 are equivalent locations for phosphate binding. Nonetheless, the ancestral phosphate binding site in both P-loops and Rossmanns resides at the tip of α1 (as judged by α4 being a rare exception). This suggests that the positioning of the PBL at the tip of α1 in both Rossmann and P-loop proteins is a signal of shared ancestry rather than convergence. As a minimum, the identification of α4 as a feasible alternative supports a model of emergence of both lineages from a seed β–PBL–αβ fragment, as outlined further below. By this scenario, α4 only emerged at a later stage, well after phosphate binding had been established at α1.

A shared β2-Asp motif?

As outlined above, the β1-PBL-α1 segment of P-loops and Rossmanns likely represents a primordial polypeptide that could later be extended to give the modern αβα sandwich domains[7,24]. However, there are several indications that the ancestral, seeding peptide(s) of both P-loops and Rossmanns also contained β2[7]. In the case of the Rossmann, β2 of the seeding primordial peptide plays a functional role: an Asp at its tip forms a bidentate interaction with the hydroxyls of the nucleotide's ribose (**Figure 2B**)[10]. The putative Rossmann common ancestor thus comprises a β-PBL-α-β-Asp fragment. Might such a fragment also be the P-loop ancestor?

In fact, both families make use of an Asp residue at the tip of the β-strand just next to β1 – in P-loops this residue is the above-mentioned Walker B motif (**Figure 2A**). Is this feature also a sought-after signature of shared ancestry? In the simplest P-loop topology, the Walker B-Asp resides at the tip of the β-strand which is adjacent to β1, as it is in Rossmanns. Thus, putting aside the connectivity of strands, both P-loop and Rossmann possess a functional core of two adjacent strands, one from which the PBL extends and the other with an Asp at its tip (**Figure 2A**, **B**). However, because in P-loops the strand topology is swapped, in the primary sequence, the Walker B-Asp resides at the tip of β3 (in the simplest topology described in **Figure 1C**, and at the tip of β4 in another common topology as detailed below). As elaborated later, variations in topology of P-loops support a model by which additional β-loop-α elements got inserted into the ancestral β-PBL-α-β-Asp seed fragment such that what was initially β2 became β3 (or even β4 in other P-loop families).

However, even if we put the question of topology aside for the moment, there remains the fundamental functional difference between the P-loop Walker B-Asp (binding of a catalytic dication) and the Rossmann β2-Asp (ribose binding; **Figure 2A** and **2B**,

respectively). Can this difference be reconciled? This question might be answered by identifying cases of parallel evolution, or 'bridging proteins'. Specifically, we searched for examples of Rossmanns acting as NTPases, and examined whether they use a catalytic metal and, if so, whether this metal cation is bound by the β2-Asp.

Tubulin - a parallelly evolved Rossmann NTPase

As explained above, Rossmanns typically use the ligand's phosphate moiety as a binding handle, whereas P-loops perform chemistry on the ligands' phosphate groups. Thus, to discover bridging proteins, we looked at the minority of Rossmann families that do act as NTPases. In all but one of these, the NTP is bound in the canonical Rossmann mode, namely with the NTP's ribose moiety bound to the β2-Asp (**Figure S1**). However, one family, tubulin, is an outlier. Tubulin is a GTPase first discovered in eukaryotes. With time, bacterial and archaeal tubulins were discovered, indicating that this lineage originated in the LUCA[25,26]. Tubulin has undisputable hallmarks of a Rossmann[27], as noted originally[28], and is categorized as such (ECOD family: 2003.1.6.1). The strand topology is distinctly Rossmann (3-2-1-4-5), with a phosphate binding loop located between β1 and α1. We further note that binding of GTP's phosphate groups is mediated by a water molecule bound to the N-terminus of α1 (**Figure 2C**), as in canonical Rossmanns[19] (**Figure 2B**) and in contrast to P-loops. However, as noticed by those who solved the first tubulin structures, GTP is oriented differently compared to the nucleotide cofactors bound by other Rossmanns[27]. Our examination reveals that tubulin binds GTP in the P-loop NTPase mode – namely, with the nucleoside pointing away from the domain's core (**Figure 2C**). Indeed, tubulin's phosphate binding loop is truncated relative to other Rossmanns and adopts a conformation akin to a tight hairpin (**Figure 2**, **Figure S2A**). In fact, tubulin has a second phosphate binding loop

that resides at the tip of α4 and has a critical role in catalysis, indicating that α4 can readily take the role of phosphate binding as seen in the HUP families described above (**Figure 3**).

Foremost, the β2-Asp interaction with the ribose, a hallmark of Rossmanns, is absent in tubulin (**Figure 2C**). Rather, the canonical Asp at the tip of β2 ligates a catalytic dication (**Figure 2C**). Further, tubulin's binding of the dication adopts a P-loops like octahedral geometry[29] with the β2-Asp interacting with a water that occupies an equivalent site as the water molecule that the Walker B-Asp in P-loops interacts with (**Figure S3**). The β2-Asp is essential for tubulin's catalytic activity[30], and its Walker B like mode of action is seen across multiple tubulin structures (in many tubulins Asn is seen at the β2 tip position, though this β2-Asn also ligates the dication, either directly or via a water molecule (**Figure S2B**; **Table S1**).

Tubulin therefore comprises an intriguing case of a Rossmann that evolved an NTPase function by reorienting the NTP substrate to bind in the P-loop NTPases mode, thereby repurposing the canonical Rossmann β2-Asp to ligate a catalytic metal cation. Put differently, tubulin shows that the functional differences between the P-loop Walker B and the Rossmann β2-Asp can be reconciled.

Shared *theme*s between Rossmanns and P-loops

Encouraged by tubulin, we endeavored to look for additional evidence for bridging proteins, ideally with respect to not only structure but also sequence. To this end, we employed the concept of an *agile theme* – short stretches for which alignments are statistically significant (≥ 20 residues; HHSearch E-value $<10^{-3}$) yet with the flanking regions showing no detectable sequence homology[31]. In the context of this work, we specifically searched for shared themes in structures that belong to Rossmanns (X-Group 2003) on the one hand and P-loops (X-group 2004) on the other. By focusing the sequence homology search on evolutionarily-

distinct domains, and by using bait sequences derived from validated sequence themes[16], the sensitivity and accuracy of this approach exceeds that of standard HMM-based searches (further details about themes detected between other X-groups are described in a forthcoming manuscript[31]). Given a stringent statistical threshold, only a few shared themes were detected, all involving the P-loop enzyme HPr kinase/phosphatase (F-Group 2004.1.2.1; PDB: 1ko7; **Figure 4A**). A few different Rossmann F-groups share a theme with this P-loop, with sorbitol dehydrogenase (F-Group 2003.1.1.417; PDB: 1k2w) and short chain dehydrogenase (F-Group 2003.1.1.332; PDB: 3tjr) showing the highest overlap (**Figure 4**).

HPr kinase/phosphatase is a bifunctional bacterial enzyme that catalyzes the phosphorylation of a signaling protein (HPr) and its dephosphorylation[32]. The P-loop domain comprises its C-terminal domain and carries the kinase function (hereafter Hpr kinase). Remarkably, the Walker B-Asp of Hpr kinase resides at the tip of β2, rather than at β3 or β4 as in the canonical P-loops. Consequently, although no such constraint or steering was applied to the search algorithm, the detected shared theme encompasses an intact β1-PBL-α1-β2-Asp element in the Rossmann proteins (where this element is canonical) as well as in this unique P-loop family (**Figure 4A**). As expected, this element is conserved in both the P-loop Hpr kinase and in the Rossmann families, with the Gly residues of the phosphate binding motifs, and the β2-Asp's being almost entirely conserved (**Figure 4B**). This result underscores the significance of the β1-PBL-α1-β2-Asp motif as the shared evolutionary seed of both Rossmanns and P-loops (detailed in the next section).

Consistent with the idea of parallel evolution, these bridging P-loop and Rossmann proteins seem to be at the fringes of their respective lineages. In the case of HPr kinase, the active site is characterized by a canonical Walker A motif but the Walker B-Asp is uncharacteristically situated at the tip of β2 (**Figure 4D**). Further, in P-loop families with the simplest topology, the Walker B-Asp resides at the tip of a β-strand that structurally resides

next to β1 (β3, **Figure 2A**). However, in most P-loops, another strand, typically β4, is inserted between β1 and the strand carrying the Walker B-Asp (**Figure S4**; Ref. [4]). Hpr kinase belongs to this second category; however, its intervening strand is highly unusual – an anti-parallel β-strand inserted between β1 and β2. In the primary sequence, the intervening strand is an N-terminal extension, and thus upstream to β1 (**Figure 4C**; annotated as β-1). Indeed, HPr kinase is classified as an outlier with respect to the greater space of P-loop proteins. The P-loop X-group in ECOD (X-group 2004) is split into two topology groups (T-groups): *P-loop containing nucleoside triphosphate hydrolases*, which includes 196 F-groups that represent the abundant, canonical P-loop proteins, and *PEP carboxykinase catalytic C-terminal domain*, which is comprised of just three F-groups. HPr kinase is classified as part of the latter. As discussed below, the variation in topology of HPr also highlights the structural plasticity of the P-loop fold with respect to insertions.

The sorbitol dehydrogenase and short chain dehydrogenase both have the canonical Rossmann strand topology and β2-Asp. Homology modeling of the enzyme-NAD^+ complex, and inspection of closely related structures, suggest that binding of the NAD^+ cofactor is also canonical (**Figure 4D**). However, the PBL of these three Rossmann proteins is nonstandard (GxxxGxG instead of the canonical Rossmann which is GxGxxG; **Figure 4A**). Further, although the structural positioning of the last two glycine residues is rather similar to that in canonical Rossmann proteins, the extended GxxxGxG motif results in an extended PBL with higher resemblance to the P-loop (**Figure 4F**). Indeed, the sequence alignment reveals that Hpr's P-loop (which is canonical) and these nonstandard Rossmann PBLs are only few mutations away from each other (**Figure 4A**, and **4F**, overlay in right panel).

An ancestral βαβ seed of both Rossmanns and P-loops

The above findings support the notion of a common Rossmann/P-loop ancestor the minimal structure of which is βαβ. This ancestral polypeptide includes just two functional motifs: a phosphate binding loop and an Asp, which could play a dual role of either binding the ribose moiety of various nucleotides or of ligating a dication such as Mg^{2+} or Ca^{2+}. Previously, such a polypeptide (*i.e.*, β-PBL-α-β-Asp) has been proposed as the seed from which Rossmann enzymes emerged (Refs.[7,33] and references therein). In contrast, a βα element was assigned as the P-loop ancestral seed (*i.e.*, a β-P-loop-α segment; Refs.[7,33], and references therein). Here, we argue that ancestral polypeptide(s) comprising a βαβ element gave rise to both lineages, and possibly that a single polypeptide served as a common ancestor of both lineages.

From an ancestral seed to intact domains

We further hypothesize that the above seed fragment was subsequently expanded by addition of β-loop-α elements. Expansion has also enabled a functional split, or sub-specialization, of the two separate lineages – Rossmann and P-loop, that further evolved and massively diversified. In essence, this split regards two key elements – phospho-ligand binding and β-strand topology. Our analysis indicates the feasibility of both.

The plausibility of common descent of the P-loop Walker A and the Rossmann phosphate binding loops is indicated by the detected shared theme described above (**Figure 4**). Although the canonical motifs of both lineages differ, there still exists – particularly among Rossmann proteins – alternative motifs that could diverge via a few mutations to a Walker A P-loop. Other Rossmanns possess a GxGGxG motif that also represents a potential jumping board to a Walker A P-loop[24]. Given that it mediates binding rather than catalysis, and owing to its lower conservation, we speculate that a Rossmann PBL could be replaced by a Walker A-like P-loop. However, at present, how permissive the Rossmann PBL is to

sequence changes that will render it Walker A-like is unclear. Future experimental work, possibly using the Rossmann enzymes indicated here (**Figure 4**), might lend support to the hypothesis that these two PBLs are indeed evolutionarily related. The identified shared themes also indicate that the β2-Asp of the presumed ancestral fragment could not only bind the ribose moiety as in Rossmanns, but also serve as a Walker B, as in P-loops. Tubulin lends further support, indicating that a β2-Asp can indeed play a dual role.

Expansion of the ancestral βαβ fragment would enable not only the sub-functionalization of the two functional elements described above, but also the fixation of two separate β-strand topologies – the sequential Rossmann topology *versus* the swapped P-loop one. Both folds are in essence a tandem repeat of β-loop-α elements (**Figure 1**). The evolutionary history of other repeat folds tells us that expansion would typically occur by duplication of the ancestral fragment[34–37], or parts of it, but also from fusion with independently emerging fragments[38,39]. Regardless of the origin of the extending fragment(s), given a βαβ ancestor, sequential fusions of αβ elements (or larger elements) could give rise to either one of these two folds. As summarized in **Figure 5**, a newly added β-strand can align at the edge, next to the existing two strands, leading to a Rossmann like topology. Alternatively, a βα element inserted in between the two ancestral strands would result in P-loop topology (*Step 2*; note that insertion results in the ancestral β2 that carries the Walker B-Asp becoming β3). In next extension, the newly added 4th strand aligns at the other side of β1 (*Step 3*; β4 added with its preceding helix, α3). The subsequent strand(s) could in principle be added sequentially, one next to the other, to yield the intact domains as we know them today. Indeed, that extensions at the edges of the core β-sheet happen in both folds is evident as both include proteins with either 5 or 6-strands. Further, circular permutations are common

in Rossmanns, as are non-canonical additions of a 7th strand, including anti-parallel strands, at either end of the β-sheet[38].

Also in support of the above scenario, transitions in the topology of β-sheets that result from strand swaps, or strand invasions, have been documented[38]. In particular, the P-loop lineage seems to have undergone various strand swaps and insertions that gave rise to a variety of topologies[4], including the noncanonical one, with an antiparallel strand, seen in Hpr kinases (**Figure 4C**). Indeed, a survey of P-loop F-groups reveals multiple strand topologies (**Figure S2**; see also Ref. [4]). In general, families that catalyze phosphoryl transfer, namely kinases such as thymidylate kinase (F-group 2004.1.1.166), but also GTPases such as elongation factor Tu (F-group 2004.1.1.258), tend to have the simplest 2-3-1-4-5 topology illustrated in **Figure 1C** (see Ref. 4). In these proteins, the Walker A P-loop and the Walker B-Asp reside on adjacent strands, with the Walker B motif on the tip of β3 (as illustrated in **Figure 2A**). On the other hand, "motor proteins", in which ATPase activity drives a large conformational change that turns into some further action, such as helicases or the ATP cassette of ABC transporters, tend to have a strand inserted between β1 and β3, to yield a 2-3-4-1-5 topology. Here, the Walker B-Asp is also situated on the tip of β3, yet with an intervening strand (β4) between the Walker A and Walker B motifs. The split between these topologies is ancient, likely predating the LUCA[4], and supports the hypothesis that early events of fusions as well as insertions were associated with the functional radiation of the P-loops lineage.

In contrast to the P-loops variable topologies, the pseudo-symmetrical Rossmann topology seems highly conserved (in a previous analysis of the Rossmann fold, we did not detect a single structure annotated as Rossmann with swapped strand topology[10]). Further, the very same topology appears in other domains, so-called Rossmann-like, or Rossmannoid domains (foremost, Flavodoxin, 2-1-3-4-5, and HUP, 3-2-1-4-5). The latter two also

represent ancient, pre-LUCA phospho-ligand binding domains that likely evolved independently of the Rossmann[21] and converged to the same topology. That convergence to the Rossmann topology occurred frequently and may relate to the higher thermodynamic and/or kinetic stability of the symmetric strand topology. Indeed, the design of Rossmann-like proteins is readily realized compared to P-loops-like proteins with the swapped strand topology[35]. Furthermore, systematic assays of refoldability of the *E. coli* proteome, and a comparison of the folds to which proteins belong, indicated that Rossmann is among the most refoldable folds while P-loop NTPases are among the poorly refolding ones[40].

Conclusions

Protein evolution spans nearly 4 billion years, with the founding events occurring pre-LUCA. As such, for many protein families, definitive assignment of homologous versus analogous relationships (shared ancestry versus convergent evolution) may never be possible[8]. Confounding matters further, early constraints on protein sequence and structure have further limited the number of possible solutions to a subset of structures and binding motifs[22], making convergence a more likely scenario, particularly in the most ancient proteins. Thus, although discovered several decades ago, whether the P-Loop NTPase and Rossmann lineages diverged or converged remains an open question. The availability of thousands of structures, highly curated databases that catalogue them[13,14], and sensitive search methods[41] and algorithms[42] allows this question to be reexamined. Here, evidence in favor of common ancestry between these lineages is provided, though convergence cannot and should not be entirely ruled out. Whether it was convergence or divergence, our analysis suggests that both lineages emerged from a polypeptide comprising a β-PBL-α-β-Asp fragment. Such a polypeptide was likely the ancestor of both P-loops and Rossmanns–-be it the same polypeptide, or two (or more) independently emerged ones. Reconstruction of ancestral

polypeptides[43], including 40 residue polypeptides that relate to the P-loop NTPase ancestor[44], may allow us to further examine the common *versus* independent emergence scenarios.

Methods

The functional diversity of the P-loop and Rossmann lineages

In total, three X-groups comprising 663 ECOD F-groups were analyzed (**Supplemental File 1**): P-loop-like (X-group 2004; 157 F-groups), Rossmann-like (X-group 2003; 168 F-groups) and Rossmann-like structures with the crossover (X-group 2111; 338 F-groups). For this analysis, X-groups 2003 and 2111 were merged. The sequences of each F-group (70% identity cutoff) were mapped to a SUPERFAMILY[45] entry with HMMsearch[41] using the HMM profiles provided by the SUPERFAMILY database. The SUPERFAMILY EC2Domain mapping file was used to collect the Enzyme Commission (EC) classes associated with each family. In total, we identified 75 EC classes associated with P-loops (X-group 2004) and 727 with Rossmanns (X-groups 2003 and 2111). Within all three X-groups, the majority of families exhibit transferase activity (2.-.-.-). Within the Rossmann-like X-group, oxidoreductases (1.-.-.-) are also common. For both P-loop and Rossmann-like structures with the crossover, the second most common enzyme activity is hydrolase (3.-.-.-), while for Rossmann-like families, hydrolase activity is the least common.

Identification of shared themes between P-loops and Rossmanns

We used HHSearch (version 3.0.0)[15] to compare a set of previously curated themes[16] to a 70% non-redundant set of ECOD domains (version develop210). Using an E-value threshold of 10^{-3}, a coverage threshold of 85% (for the local alignment), and a minimal length of 20 residues, we identified 267 themes with significant hits to proteins belonging to both ECOD X-groups 2004.1 (P-loop domains-like) and 2003.1 (Rossmann-like). All of these themes matched the same P-loop domain (e1ko7A1) with various Rossmann domains. To reduce the extensive redundancy among the themes, which in turn leads to redundancy in the detected proteins, we kept only two representatives Rossmanns per theme. To identify the

representative domains, we re-aligned the parts matching the theme using a Smith-Waterman (SW) or Needleman-Wunsch (NW) alignment, and the parts before and after the theme using an SW alignment. The representative domains are the ones with the most similar matching parts and the most dissimilar flanking parts. A p-value for the aligned parts was calculated from the significance of the alignment score relative to scores from alignments of random segments. Here, we estimated the parameters of the extreme value distribution (EVD) from the scores of the alignments between one of the two well-aligned segments and 1000 randomly chosen segments drawn from a multinomial distribution estimated from the other of the two well-aligned segments. We kept only cases where the matching parts have a score with a p-value lower than 0.05. This procedure resulted in a set of 57 Rossmann domains, each aligned to the Hpr kinase (PDB: 1ko7). For 50 of these 57 hits, the matching parts were aligned with the SW local alignment and the rest were aligned by NW global alignment. Here, we report the 51 cases that match the β1-β1-β2 regions (26 from the ECOD family 2003.1.1.417, 18 from 2003.1.1.332, 5 from 2003.1.1.410, and 2 from 2003.1.1.11; **Table S2**) The alignments presented in the manuscript are the global alignments recalculated for the β-PBL-α-β elements in themes that bridge the two evolutionary lineages.

Modeling ligand placements in unliganded structures

HrP kinase (e1ko7A1) does not have a ligand bound. The conformation of the PBL, however, is canonical. Thus, despite no structure from the same F-group having a relevant ligand bound, the overall positioning of the ligand can be estimated by overlaying a canonical PBL with a bound ligand from a different P-loop F-group. To generate **Figure 4C**, the PBL and ligand from ECOD domain e6at2A2, corresponding to residues 247-257 was aligned to residues 150-160 in chain A of 1ko7. This structure was chosen because it is in the same T-group as HrP kinase (2004.1.2.-) and, although the two domains have nearly undetectable

sequence identity, they share the same general topology, including the inserted anti-parallel β-strand adjacent to β1. The sequences of the PBLs (FGLSGTGKTTL and VGPNGSGKSTV for 6at2 and 1ko7, respectively; identical residues underlined) show high similarity as do the the structures of the PBLs that were aligned to generate the modelled ligand (Cα RMSD of 0.49 Å).

Calculating Consensus Sequences and Residue Conservation Scores

The relevant ECOD F-groups (**Figure 4**) were mapped to the corresponding Pfam families. Since 2003.1.1.417 and 2003.1.1.332 are associated with one Pfam family, they were analyzed jointly. Seed alignments were extracted from Pfam, clustered at 70% sequence redundancy using CD-HIT[46], and the consensus sequence and conservation scores were calculated for the shared region (theme) using JalView.

Figure Legends

Figure 1. The 3-layer αβα sandwich. **A**. The αβα sandwich is a modular fold comprised of repeating β-loop-α elements. This side-view shows two tandem βα elements: the functional loops are situated on the "top" of the fold (thick lines) and the β-loop-α element are linked via short, bottom loops (thin, dashed lines). Shown here are the first two elements with a Rossmann topology, beginning with β1 at the N-terminus, and the first two helices (α1, α2) that in this cartoon comprise one external layer of the sandwich. **B**. A view from the top reveals the αβα sandwich architecture with its three layers: a parallel β-sheet flanked on both sides by α-helices. The top active site loops face the reader and the N- and C-termini and bottom connecting loop are at the back. The order of the β-strands in the interior β-sheet follows the canonical Rossmann topology. **C**. The most common, core P-loop NTPase (P-loops) topology. Noted in red are the differences from the Rossmann topology—migration of β3 from the edge to the center, and of α2 and α5 from one external layer to another.

Figure 2. The Ligand Binding Modes of Rossman and P-Loop Proteins. The phosphate binding loops (PBLs) of both lineages connect the C-terminus of β1 to the N-terminus of α1 (conserved glycine residues are colored magenta). The Rossmann β2-Asp, and the P-loop Walker B-Asp, are in green sticks. Water molecules are denoted by red spheres, and metal dications by green spheres. **A**. The canonical P-loop NTPase binding mode. The phosphate binding loop (the P-loop Walker A motif; GxxxxGK(T/S)) begins with the first conserved glycine residue at the tip of β1 and ends with T/S within α1. The Walker B-Asp, located at the tip of β3, interacts with the catalytic Mg^{2+}, either directly, or via a water molecule as seen here. **B**. The canonical Rossmann binding mode. The phosphate binding site includes a canonical water molecule (α1 has been rendered transparent so that the conserved water is visible). The Asp sidechain at the tip of β2 (β2-Asp) forms a bidentate interaction with both

hydroxyls of the ribose. Note also the opposite directions of the ribose and adenine moieties in P-loops (pointing to the front) *versus* Rossmann (pointing back). **C**. Tubulin is a GTPase that belongs to the Rossmann lineage. It possesses the canonical Rossmann strand topology, phosphate binding loop (including the mediating water), and β2-Asp. However, the ligand, GTP, is bound in the P-loop NTPase mode (as in **A**). Accordingly, the β2-Asp makes a water mediated interaction with the catalytic metal cation (Ca^{2+}or Mg^{2+}) thus acting in effect as a Walker B-Asp (the metal cation's coordination schemes are also identical, see **Figure S3**). ECOD domains used in this figure, from left to right, are e1yrbA1, e1lssA1, and e5j2tB1.

Figure 3. Alternative phosphate binding sites in αβα sandwich enzymes. HUP proteins are αβα sandwich proteins with Rossmann-like strand topology. The canonical HUP phosphate binding loop is located at the tip of α1 and colored magenta (left panels; NAD^+ synthase; ECOD F-group 2005.1.1.13; shown is domain e1xngA1) as in Rossmann and P-loop NTPases (**Figure 2**). However, Usp (universal stress proteins) is a HUP family that exhibits kinase activity wherein phosphate binding migrated to the tip of α4 (right panels; ECOD F-group 2005.1.1.145; shown is domain e2z08A1; residues interacting with phosphate groups are colored magenta). As shown in the overlay (middle panels), despite the variation in the phosphate binding site, the ribose and adenine binding modes are identical.

Figure 4. Theme sharing between Rossmann and P-loop enzymes. A. Sequence alignment of the shared themes. PDB codes are shown on the right, and the ECOD F-group to which they belong are on the left. The identified themes involve a segment of a single P-loop NTPase, Hpr kinase (top line, ECOD domain e1ko7A1), that aligns to a variety of Rossmanns that belong to four different F-groups (representatives shown here; see **Table S2** for the complete list of agile themes). **B**. The consensus sequence of each F-group (see

Methods) is shaded according to the degree of conservation. The individual sequences identified by the theme search show higher similarity by default, yet nonetheless, the family consensus sequences also align well, and the identical residues tend to be conserved. **C-D**. Although detection of the shared theme was based on sequence only, structurally, the shared theme encompasses the β1-PBL-α1-β2-Asp element in both the P-loop protein (panel C; Hpr kinase, ECOD domain e1ko7A1) and the theme-related Rossmanns (panel D; ECOD domains e3gedA1, e1kvtA1, e3ondA1, and e3tjrA1; the ligand is bound by domain e3tjrA1). Note that only the pyrophosphate and ribose moieties of the ligand are shown for clarity. The conserved phosphate binding loop glycine residues are colored magenta and the β2-Asp is colored green. For panel C, the ligand binding mode was modelled using the structure of a liganded P-loop protein (see *Methods*). **E**. An overlay of the β1-PBL-α1-β2-Asp element of the Hpr Kinase (cyan; ECOD domain e1ko7A1) and one of the theme-related Rossmann dehydrogenases (yellow; ECOD domain e3tjrA1). **F**. Structural details of the phosphate binding loops: The Walker A binding loop of Hrp kinase (left panel; ECOD domain e1ko7A1); the phosphate binding loop of sorbitol dehydrogenase (middle panel; ECOD domain e1k2wA1); and an overlay of both loops (right panel).

Figure 5. Divergence of the Rossmann and P-loop NTPase folds from a common ancestral polypeptide. Emergence begins with the presumed β-PBL-α-β-Asp ancestor, that could act as either Rossmann, or a P-loop NTPase, depending on how the phospho-ligands bind and the role taken by the β2-Asp (**Figure 2A** and **C**). In the second step, the ancestral fragment is either extended at its C-terminus by fusion of an αβ fragment to generate a Rossmann-like domain (top row). Alternatively, a βα fragment is inserted between α1 and β2 to yield a P-loop-like domain (bottom row). Note that insertion results in the ancestral β2 that carries the Walker B-Asp becoming β3. Note also that the location of the added helix, α2,

differs. It may locate next to α1, namely on the same layer of what will become a Rossmann αβα sandwich, or, on the opposite side, as in P-loop NTPases.

Acknowledgments

This research has been supported by Grant 94747 by the Volkswagen Foundation. N.B.-T.'s research is supported in part by the Abraham E. Kazan Chair in Structural Biology, Tel Aviv University. D.S.T. is the Nella and Leon Benoziyo Professor of Biochemistry. We are grateful to Ita Gruic-Sovulj for her role in the analysis of the HUP domain that led to Figure 3, and to Andrei Lupas for insightful and critical comments.

References

1. Adams, M. J. *et al.* Structure of lactate dehydrogenase at 2.8 Å resolution. *Nature* (1970). doi:10.1038/2271098a0
2. Rossmann, M. G., Moras, D. & Olsen, K. W. Chemical and biological evolution of a nucleotide-binding protein. *Nature* (1974). doi:10.1038/250194a0
3. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* (1982). doi:10.1002/j.1460-2075.1982.tb01276.x
4. Leipe, D. D., Koonin, E. V. & Aravind, L. Evolution and classification of P-loop kinases and related proteins. *J. Mol. Biol.* (2003). doi:10.1016/j.jmb.2003.08.040
5. Ma, B. G. *et al.* Characters of very ancient proteins. *Biochemical and Biophysical Research Communications* (2008). doi:10.1016/j.bbrc.2007.12.014
6. Edwards, H., Abeln, S. & Deane, C. M. Exploring Fold Space Preferences of New-born and Ancient Protein Superfamilies. *PLoS Comput. Biol.* (2013).

doi:10.1371/journal.pcbi.1003325

7. Alva, V., Söding, J. & Lupas, A. N. A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**, e09410 (2015).

8. Aravind, L., Mazumder, R., Vasudevan, S. & Koonin, E. V. Trends in protein evolution inferred from sequence and structure analysis. *Current Opinion in Structural Biology* (2002). doi:10.1016/S0959-440X(02)00334-2

9. Goncearenco, A. & Berezovsky, I. N. Protein function from its emergence to diversity in contemporary proteins. *Phys. Biol.* (2015). doi:10.1088/1478-3975/12/4/045002

10. Laurino, P. *et al.* An ancient fingerprint indicates the common ancestry of Rossmann fold enzymes utilizing different ribose based cofactors. *PLOS Biol.* (2016). doi:10.1371/journal.pbio.1002396

11. Galperin, M. Y. & Koonin, E. V. Divergence and convergence in enzyme evolution. *Journal of Biological Chemistry* (2012). doi:10.1074/jbc.R111.241976

12. Elias, M. & Tawfik, D. S. Divergence and convergence in enzyme evolution: Parallel evolution of paraoxonases from quorum-quenching lactonases. *Journal of Biological Chemistry* (2012). doi:10.1074/jbc.R111.257329

13. Chandonia, J. M., Fox, N. K. & Brenner, S. E. SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins – extended Database. *J. Mol. Biol.* (2017). doi:10.1016/j.jmb.2016.11.023

14. Cheng, H. *et al.* ECOD: An Evolutionary Classification of Protein Domains. *PLoS Comput. Biol.* (2014). doi:10.1371/journal.pcbi.1003926

15. Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. Fast and accurate automatic structure prediction with HHpred. *Proteins Struct. Funct. Bioinforma.* (2009). doi:10.1002/prot.22499

16. Nepomnyachiy, S., Ben-Tal, N. & Kolodny, R. Complex evolutionary footprints

revealed in an analysis of reused protein segments of diverse lengths. *Proc. Natl. Acad. Sci. U. S. A.* (2017). doi:10.1073/pnas.1707642114

17. Bukhari, S. A. & Caetano-Anollés, G. Origin and Evolution of Protein Fold Designs Inferred from Phylogenomic Analysis of CATH Domain Structures in Proteomes. *PLoS Comput. Biol.* (2013). doi:10.1371/journal.pcbi.1003009

18. Winstanley, H. F., Abeln, S. & Deane, C. M. How old is your fold? *Bioinformatics* (2005). doi:10.1093/bioinformatics/bti1008

19. Bottoms, C. A., Smith, P. E. & Tanner, J. J. A structurally conserved water molecule in Rossmann dinucleotide-binding domains. *Protein Sci.* (2002). doi:10.1110/ps.0213502

20. Shalaeva, D. N., Cherepanov, D. A., Galperin, M. Y., Golovin, A. V. & Mulkidjanian, A. Y. Evolution of cation binding in the active sites of P-loop nucleoside triphosphatases in relation to the basic catalytic mechanism. *Elife* (2018). doi:10.7554/eLife.37373

21. 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). doi:10.1371/journal.pcbi.1007569

22. Longo LM, Petrović D, Kamerlin SCL, T. D. Short and simple sequences favored the emergence of N-helix phospho-ligand binding sites in the first enzymes. *Proc Natl Acad Sci U S A* (2020).

23. Aravind, L., Anantharaman, V. & Koonin, E. V. Monophyly of Class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: Implications for protein evolution in the RNA world. *Proteins Struct. Funct. Genet.* (2002). doi:10.1002/prot.10064

24. Zheng, Z., Goncearenco, A. & Berezovsky, I. N. Nucleotide binding database NBDB -

A collection of sequence motifs with specific protein-ligand interactions. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkv1124

25. Yutin, N. & Koonin, E. V. Archaeal origin of tubulin. *Biol. Direct* (2012). doi:10.1186/1745-6150-7-10

26. Margolin, W., Wang, R. & Kumar, M. Isolation of an ftsZ homolog from the archaebacterium Halobacterium salinarium: Implications for the evolution of FtsZ and tubulin. *J. Bacteriol.* (1996). doi:10.1128/jb.178.5.1320-1327.1996

27. Nogales, E., Downing, K. H., Amos, L. A. & Löwe, J. Tubulin and FtsZ form a distinct family of GTPases. *Nat. Struct. Biol.* (1998). doi:10.1038/nsb0698-451

28. Nogales, E., Wolf, S. G. & Downing, K. H. Structure of the αβ tubulin dimer by electron crystallography. *Nature* (1998). doi:10.1038/34465

29. Kanade, M., Chakraborty, S., Shelke, S. S. & Gayathri, P. A Distinct Motif in a Prokaryotic Small Ras-Like GTPase Highlights Unifying Features of Walker B Motifs in P-Loop NTPases. *J. Mol. Biol.* (2020). doi:10.1016/j.jmb.2020.07.024

30. Farr, G. W. & Sternlicht, H. Site-directed mutagenesis of the GTP-binding domain of β-tubulin. *J. Mol. Biol.* (1992). doi:10.1016/0022-2836(92)90700-T

31. Kolodny, R., Nepomnyachiy, S., Tawfik, D. S. & Ben-Tal, N. Agile themes: short protein segments found in different architectures. *Submitted* (2020).

32. Márquez, J. A. *et al.* Structure of the full-length HPr kinase/phosphatase from Staphylococcus xylosus at 1.95 Å resolution: Mimicking the product/substrate of the phospho transfer reactions. *Proc. Natl. Acad. Sci. U. S. A.* (2002). doi:10.1073/pnas.052461499

33. Laurino, P. *et al.* An Ancient Fingerprint Indicates the Common Ancestry of Rossmann-Fold Enzymes Utilizing Different Ribose-Based Cofactors. *PLoS Biol.* (2016). doi:10.1371/journal.pbio.1002396

34. Eck, R. V. & Dayhoff, M. O. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science (80-.).* (1966). doi:10.1126/science.152.3720.363

35. Romero Romero, M. L. *et al.* Simple yet functional phosphate-loop proteins. *Proc. Natl. Acad. Sci.* (2018). doi:10.1073/pnas.1812400115

36. Zhu, H. *et al.* Origin of a folded repeat protein from an intrinsically disordered ancestor. *Elife* (2016). doi:10.7554/eLife.16761

37. Longo, L. M. *et al.* Primordial emergence of a nucleic acid binding protein via phase separation and statistical ornithine to arginine conversion. *bioRxiv* (2020). doi:10.1101/2020.01.18.911073

38. Grishin, N. V. Fold change in evolution of protein structures. *J. Struct. Biol.* (2001). doi:10.1006/jsbi.2001.4335

39. Setiyaputra, S., MacKay, J. P. & Patrick, W. M. The structure of a truncated phosphoribosylanthranilate isomerase suggests a unified model for evolution of the (βα)8 barrel fold. *J. Mol. Biol.* (2011). doi:10.1016/j.jmb.2011.02.048

40. To, P., Whitehead, B., Tarbox, H. E. & Fried, S. D. Non-Refoldability is Pervasive Across the E. coli Proteome. *bioRxiv* (2020).

41. Hancock, J. M., Zvelebil, M. J., Hancock, J. M. & Bishop, M. J. HMMer. in *Dictionary of Bioinformatics and Computational Biology* (2004). doi:10.1002/9780471650126.dob0323.pub2

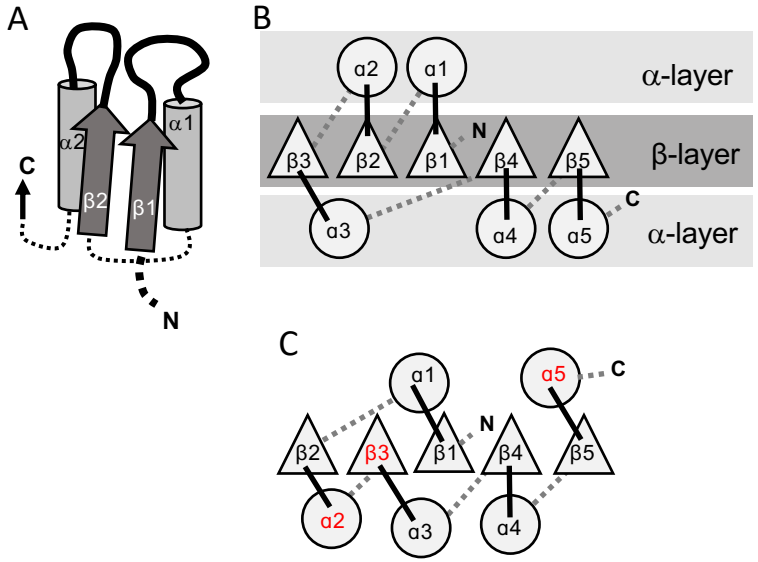
42. Nepomnyachiy, S., Ben-Tal, N. & Kolodny, R. Global view of the protein universe. *Proc. Natl. Acad. Sci.* (2014). doi:10.1073/pnas.1403395111

43. Longo, L. *et al.* Primordial emergence of a nucleic acid binding protein via phase separation and statistical ornithine to arginine conversion. *Proc. Natl. Acad. Sci.* (2020). doi:10.1101/2020.01.18.911073

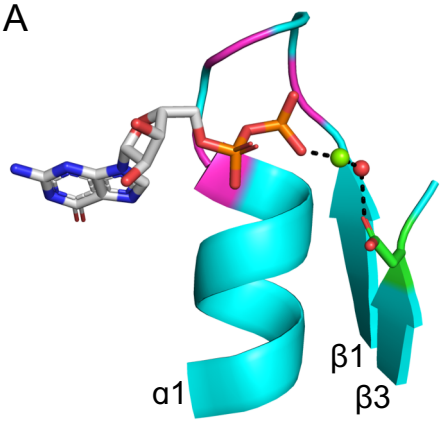
44. Vyas, P. *et al*. Helicase-Like Functions in Phosphate Loop Containing Beta-Alpha Polypeptides. *Submitt. Publ. bioRxiv* (2020).

45. Wilson, D. *et al*. SUPERFAMILY - Sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* (2009). doi:10.1093/nar/gkn762

46. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts565

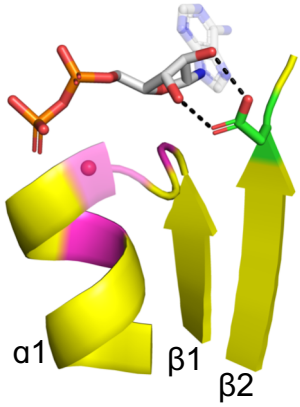

A
α2
α1
C
β2
β1
N
B
α2
α1
α-layer
β3
β2
β1
N
β4
β5
β-layer
C
α3
α4
α5
α-layer
C
α1
α5
C
β2
β3
β1
N
β4
β5
α2
α3
α4

A



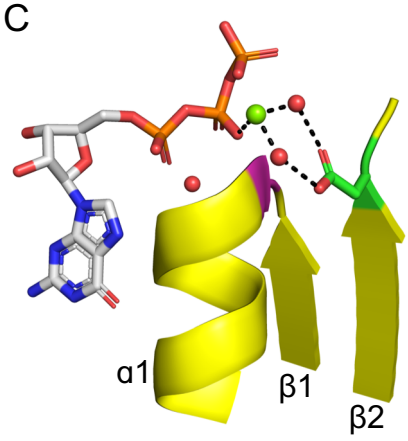
Canonical P-Loop

B

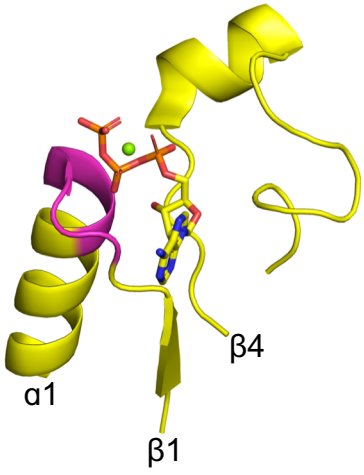


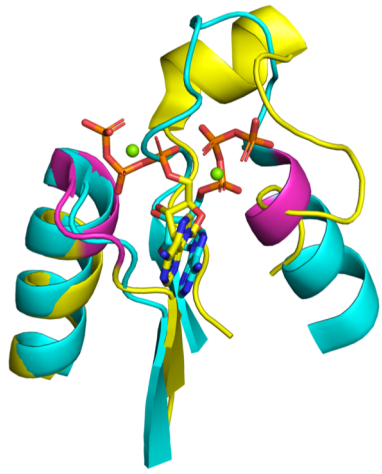
Canonical Rossmann

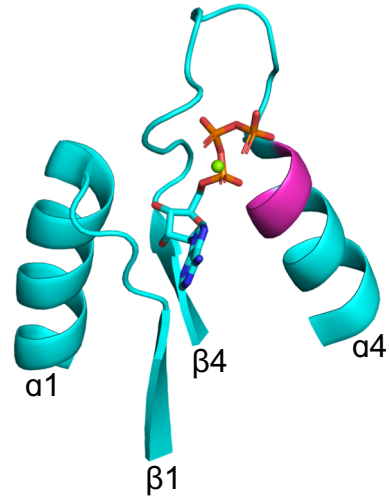
C

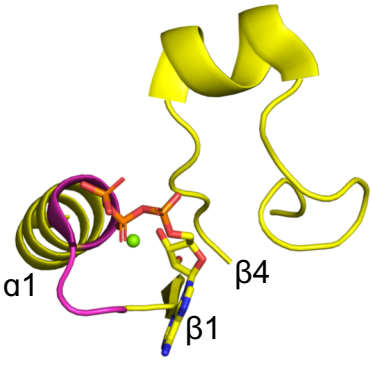


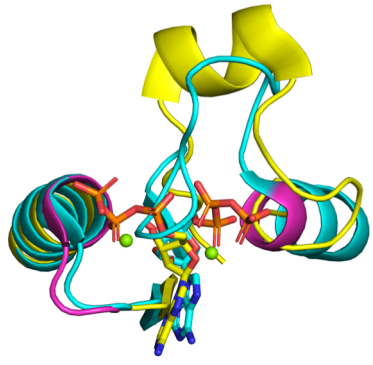
Tubulin
(Rossmann GTPase)

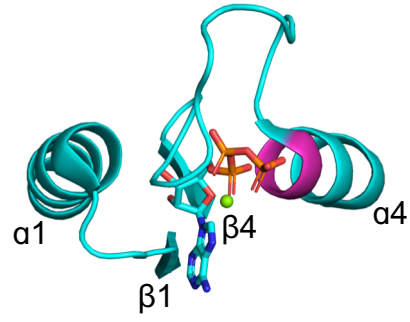








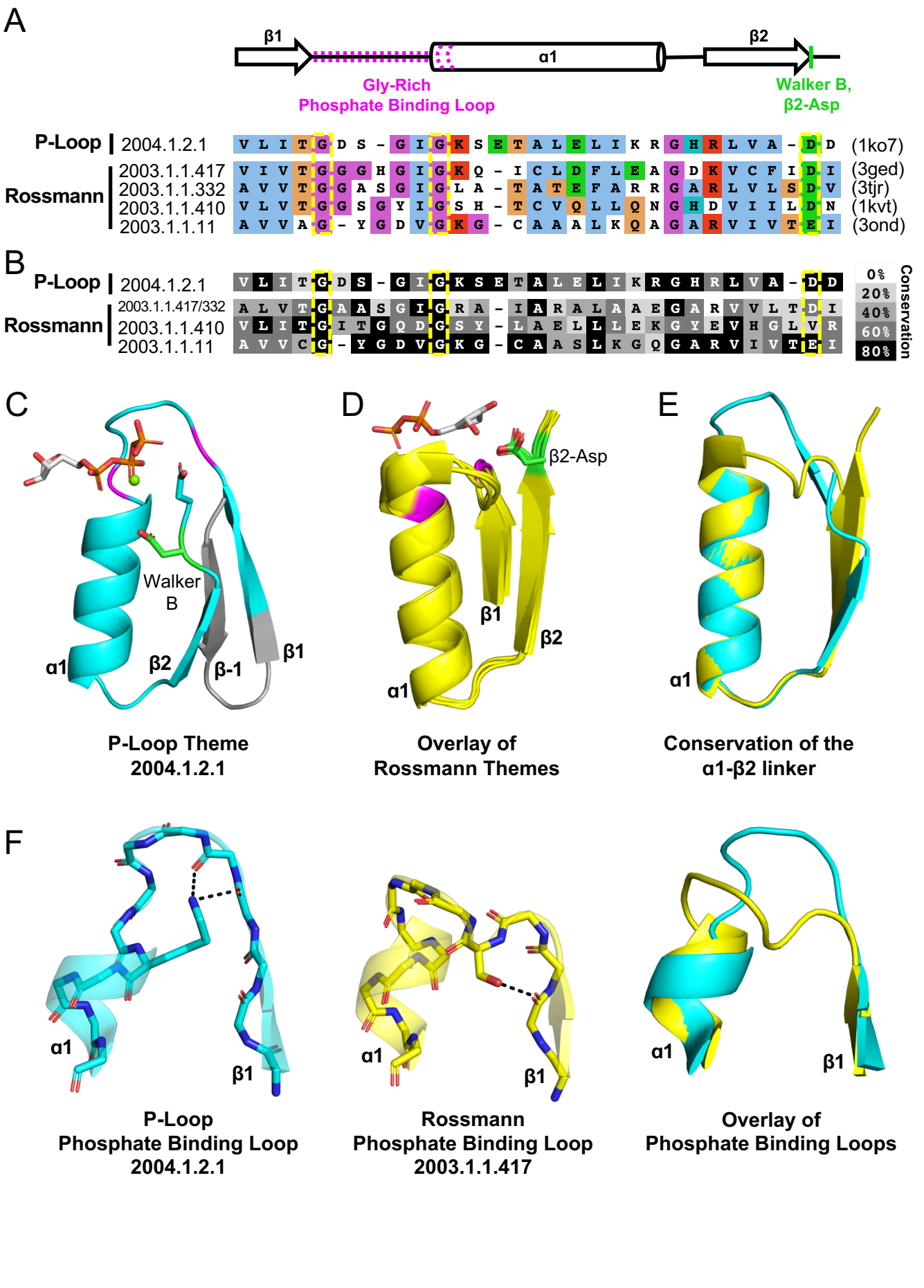


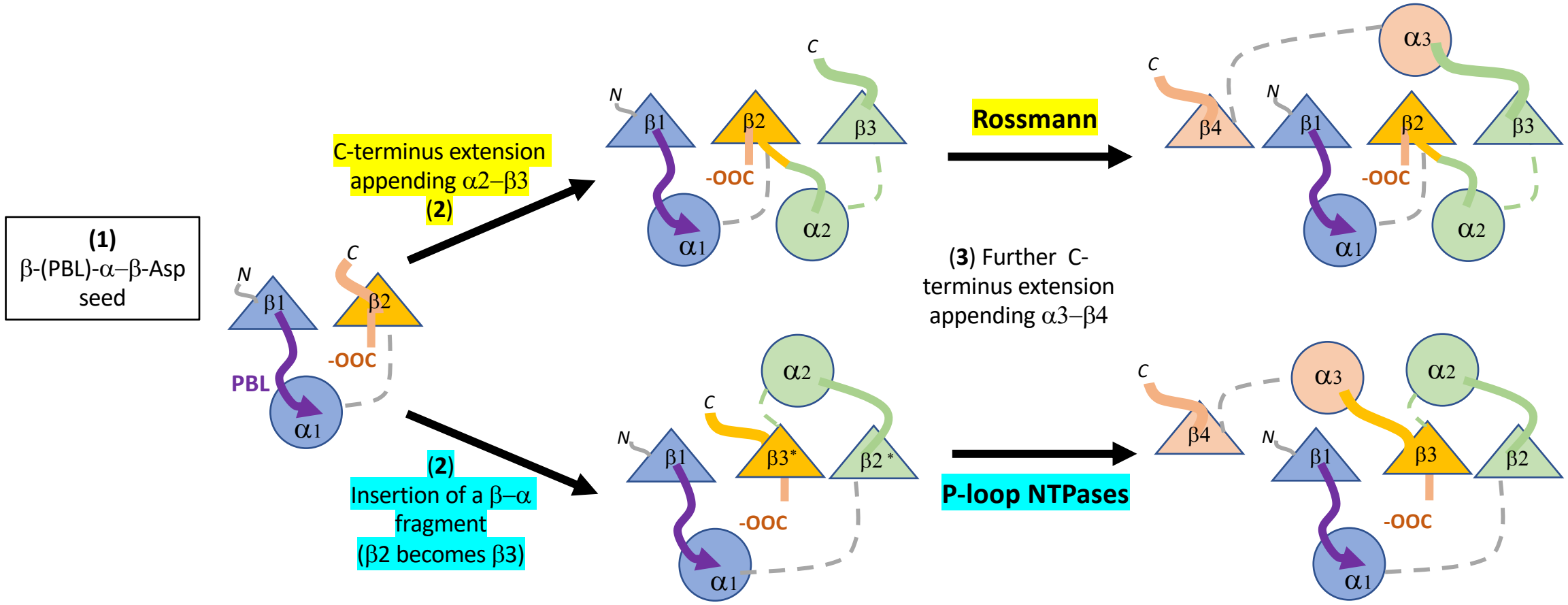


NAD+ synthase,
Canonical α1
Binding Mode

Overlay

Usp,
Non-canonical α4
Binding Mode


A
β1
α1
β2
Gly-Rich
Phosphate Binding Loop
Walker B,
β2-Asp
P-Loop
2004.1.2.1
(1ko7)
Rossmann
2003.1.1.417
(3ged)
2003.1.1.332
(3tjr)
2003.1.1.410
(1kvt)
2003.1.1.11
(3ond)
B
P-Loop
2004.1.2.1
Rossmann
2003.1.1.417/332
2003.1.1.410
2003.1.1.11
0%
20%
40%
60%
80%
Conservation
C
Walker
B
α1
β2
β-1
β1
P-Loop Theme
2004.1.2.1
D
β2-Asp
β1
β2
α1
Overlay of
Rossmann Themes
E
α1
Conservation of the
α1-β2 linker
F
α1
β1
P-Loop
Phosphate Binding Loop
2004.1.2.1
α1
β1
Rossmann
Phosphate Binding Loop
2003.1.1.417
α1
β1
Overlay of
Phosphate Binding Loops


(1)
β-(PBL)-α–β-Asp
seed
C-terminus extension
appending α2–β3
(2)
(2)
Insertion of a β–α
fragment
(β2 becomes β3)
(3) Further C-
terminus extension
appending α3–β4
Rossmann
P-loop NTPases
N
C
β1
β2
β3
α1
α2
α3
β4
β3*
β2*
PBL
-OOC