Published in final edited form as:

J Mol Biol. 2010 November 26; 404(2): 232–246. doi:10.1016/j.jmb.2010.09.034.

Evolutionary origin of a secondary structure: π -helices as cryptic but widespread insertional variations of α -helices enhancing protein functionality

Richard B. Cooley^a, Daniel J. Arp^b, and P. Andrew Karplus^a

^a Department of Biochemistry and Biophysics, 2011 Ag & Life Sciences Bldg, Oregon State University, Corvallis, OR 97331

^b Department of Botany and Plant Pathology, 2082 Cordley Hall, Oregon State University, Corvallis OR 97331

Abstract

Formally annotated π -helices are rare in protein structure but have been correlated with functional sites. Here, we analyze protein structures to show that π -helices are the same as structures known as α -bulges, α -aneurisms, π -bulges, and looping-outs and are evolutionarily derived by the insertion of a single residue into an α -helix. This newly discovered evolutionary origin explains both why π -helices are cryptic, being rarely annotated despite occurring in 15% of known proteins, and why they tend to be associated with function. An analysis of the π -helices in the diverse ferritin-like superfamily illustrates their tendency to be conserved in protein families, and identifies a putative π -helix-containing primordial precursor, a “missing link” intermediary form of the ribonucleotide reductase family, vestigial π -helices, and a novel function for π -helices we term peristaltic-like shifts. This new understanding of π -helices paves the way for this generally overlooked motif to become a noteworthy feature that will aid tracing the evolution of many protein families, guide investigations of protein and π -helix functionality, and contribute additional tools to the protein engineering toolkit.

Keywords

π -helix; α -aneurism; protein evolution; ferritin-like superfamily; secondary structure

For editorial correspondence: P. Andrew Karplus, Department of Biochemistry and Biophysics, 2011 Ag & Life Sciences Bldg, Oregon State University, Corvallis, OR 97331, Phone: (541) 737-3200; Fax (541) 737-0481, karplusp@science.oregonstate.edu.

Supplemental Information

Supplemental Figures 1–4

Supplemental Tables 1–5

Supplemental Data Set: Results of PDB survey for π -helices

Author Contributions R.B.C. performed all structural and evolutionary analyses, and designed Perl scripts to survey the Protein Data Bank. D.J.A. contributed to the study design, guided analyses of the ferritin-like superfamily and helped edit the paper. P.A.K. designed and supervised all aspects of this study. R.B.C. and P.A.K. wrote the paper. All authors discussed the results and commented on the manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

The diversification of protein function is an essential process that enables organisms from all kingdoms of life to thrive in an extraordinary range of environments. The goal of what has been called the functional synthesis of evolution is to gain a detailed understanding of how mutational changes in proteins give rise to novel functionality during this process^{1–3}. However, how the first protein folds arose and concrete mechanisms by which protein secondary structure and protein folds evolve have been difficult to define.

While α -helices, a dominant secondary structure in proteins, are defined by main-chain hydrogen bonds (H-bonds) between residues four apart in sequence, π -helices are a protein secondary structure with main-chain H-bonds between residues five apart in sequence (referred to as π -type H-bonds). They were predicted in the 1950s⁴, and although extended π -helices of regular conformation do not occur in proteins⁵, short π -helical segments do. The empirical definition of a π -helix is the occurrence of at least two sequential π -type H-bonds⁶. The literature for these short motifs is rather muddy, evolving from early work indicating that π -helices do not occur⁷, to a report documenting π -helices as rare but with special functional roles⁸, and continuing with the work of Fodje & Al-Karadaghi⁶ who showed that π -helices are more common than suspected, occurring in “one of every 10 proteins.” Still, π -helices tend to be overlooked in protein structures. π -helical conformations have been noted to occur in protein simulations^{9–11}, although it is controversial whether these occurrences give informative insight into the true occurrence of such helices or reflect a shortcoming of the molecular mechanics force-fields¹².

The large majority of naturally occurring π -helices consist of seven-residue segments with the minimal two π -type H-bonds⁶. What has not been previously noted is that these π -helices have the same conformation and H-bonding pattern seen in protein engineering studies when a single amino acid is inserted into an existing α -helix to create what was called an α -aneurism¹³ or looping out¹⁴ (Fig. 1). Also matching this conformation are α -helical distortions seen in some natural proteins that have been called π -bulges¹⁵ or α -bulges¹⁶. This striking similarity led us to hypothesize that natural π -helices arise during evolution via single residue insertions into α -helices. Here, through an analysis of known protein structures, we provide compelling evidence that naturally occurring π -helices are indeed evolutionarily related to α -helices. We further show that this often overlooked secondary structure is a highly informative marker with important evolutionary and functional implications, giving new insights into the origin, evolution and diversification of protein functionality, even for well characterized protein families.

Results and Discussion

The evolutionary relationship between α - and π -helices

π -helices are associated with α -helices—To explore our hypothesis that naturally occurring π -helices arose from the insertion of a single amino acid into an α -helix, we first investigated the π -helices identified by Fodje & Al-Karadaghi⁶ more closely. Reclassification of their list of π -helices (see Fig. 2 and *Material and Methods*) yielded 106 π -helical segments from 79 protein chains, with the longest ones having five consecutive π -type H-bonds (Table 1). Of these 106 π -helices, 102 were present in the midst of an α -helix, consistent with the insertion hypothesis. Interestingly, this observation completely explains why π -helices continue to be overlooked. When a π -helix is sandwiched in an α -helix, the first few and last few residues of the π -helix are also part of the bordering α -helices, and because automated secondary structure assignment algorithms such as DSSP¹⁷ and STRIDE¹⁸ give precedence to the α -helical assignment, these residues are designated as α -helical. Then, the remaining (if any) π -helix residues are too few to meet the minimal length

of a π -helix*, and so are designated as “turns”, leaving the π -helix completely unrecognized (e.g. Fig. 2). In the case of the 106 π -helices in this dataset, 105 are cryptic in this sense and only one is formally annotated by DSSP as a π -helix.

π -helices can be derived from α -helices by a one-residue insertion—For more direct evidence of an evolutionary relationship between α - and π -helices, we used the FSSP database¹⁹ to survey all structurally known homologs of each π -helix-containing protein. Based on the FSSP survey, 88 of the 106 π -helices occurred in proteins with a structurally known homolog for which the equivalent segment is a pure α -helix with one less residue (Supplemental Tables 1 and 2). All of the identified α -/ π -helical homolog pairs were members of the same superfamily, had structural alignments with Z-scores ranging from 6.4 to 56.0 and had sequence identities ranging from 5 to 70%. For the pairs with high levels of sequence similarity, the sequences themselves provide unambiguous evidence for the insertion of a single residue.

Considering that the Protein Data Bank (PDB) is far from a comprehensive set of protein structures and that all 18 π -helices without known α -helical homologs are embedded within α -helices (Supplemental Fig. 1), this remarkable 83% identification rate implies that the overwhelming majority of π -helices are indeed evolutionarily related to α -helices. It is noteworthy that the insertion of a single residue is able to account not for just π -helices with two H-bonds, but also those with three, four, and five H-bonds (Table 1, Fig. 3). We suggest the difference between these outcomes is how the helix adjusts its conformation to accommodate the extra amino acid (Supplemental Fig. 2); this would, of course, depend on the local environment in which the insertion takes place. Also noteworthy is that when multiple, overlapping π -helices are present within a long α -helix, a set of homologous proteins exist that document a potential stepwise pathway of accumulation of insertions (Supplemental Table 2; Fig. 4).

Taken all together, these observations show that the insertion of a single residue into an α -helix is a sufficient explanation for generating the whole spectrum of π -helices observed in this dataset, and unifies π -helices, π -bulges, α -aneurisms and α -bulges as a single phenomenon. These structures encompass a breadth of conformations and H-bonding lengths and it is a matter of definition whether one considers them a perturbation of an α -helix or a π -helix. We propose that π -helix is the best designation for all these structures because they all result from a single type of event and “ π -helix” is the only name that encompasses both the locally bulged structures and the more extended helical structures. In terms of how the α -helix is perturbed, the π -helices in this set demonstrate they sometimes minimally perturb the helix they enter just causing a local bulge or a bulge with a slight bend (e.g. Figs. 1a and 5c) or a more major kink and distortion (e.g. Fig. 5d). π -helices, as so defined, encompass a rather broad range of (ϕ, Ψ) -angles (Supplementary Fig. 2 and Fig. 2a from Fodje & Al-Karadaghi⁶) and this makes them more like 3_{10} -helices, which are much more conformationally diverse than α -helices²⁰.

π -helices occur in ~15% of all proteins—The analysis of the reclassified list of π -helices from Fodje & Al-Karadaghi⁶ conclusively supports the hypothesis for the relatedness of α - and π -helices and therefore it becomes of great interest to know the abundance of these motifs in nature. Using just our standard criteria and some broad (ϕ, Ψ) restrictions (see *Materials and Methods*), a survey of the current PDB (Table 1) shows that of the 5620 diverse protein chains, 803 chains (~15%) contain a total of 1010 π -helices, with the longest being seven H-bonds in length; proportionately similar results were found with the use of

*For 3_{10} -, α - and π -helices, DSSP and STRIDE generally do not assign the N- and C-capping residues as part of the helix. Thus, what we describe here as a seven residue π -helix is defined by DSSP to encompass only the central five residues.

more or less stringent hydrogen bonding criteria or the surveying of a larger subset of the PDB (Supplemental Table 3). Removing the (ϕ, Ψ) restrictions in our search resulted in relatively few additional hits (~5% more), showing that aside from π -helices, very few turn structures satisfy the H-bonding criteria. The results show π -helices are widespread, occurring in one in six protein chains, yet infrequent – occurring only once per chain in 84% of cases (Supplemental Table 4). The greatest number of π -helices within a single chain was eight, occurring in the bacterial homologue of a Na^+/Cl^- -dependent neurotransmitter transporter²¹ (PDB 2a65). All eight of these π -helices go unrecognized by DSSP annotation, yet two of them, which were identified in the original report as “helical disruptions”²¹, are at the substrate binding site and intimately connected with function.

Although we have not inspected each individual π -helix resulting from this broader search, they too are strongly associated with α -helices. More than 95% of the π -helices have at least 3 residues within them that are designated as α -helical by DSSP and, consistent with this, they are nearly all “cryptic”: compared with the 1010 π -helices we identify in this dataset, DSSP only formally annotates 44 π -helices (Supplemental Table 3). Also, checking the six and seven H-bonded π -helices seen in the PDB survey shows that they too have homologs with a pure α -helix that differs only by a single residue (Fig. 3), showing that they too can arise via the same insertional phenomenon as π -helices with two, three, four and five H-bonds seen in the smaller dataset. An interesting detail consistent with previous work on π -helices⁶ and π -bulges¹⁵ is that a proline residue is commonly present where the π -helix transitions back to an α -helix as 36% of seven-residue π -helices have a proline immediately following them.

The association of π -helices with protein function

The unification of π -helices and π -bulges as a single phenomenon strengthens their association with the evolution of protein function because both π -helices⁸ and π -bulges¹⁵ have also been shown to be enriched at protein functional sites. Illustrating this point are three examples of enzymes from branches of well-studied, large superfamilies with an α -helix-derived π -helix critically involved in the enzyme’s active site (Fig. 5). Because these π -helices (and most of those listed in Supplemental Table 1) are present in discrete subgroups of larger protein superfamilies, we infer that they were evolutionarily derived from α -helix-containing ancestors (Fig. 5a). In each of these cases, the π -helix is conserved within the protein family described. The first example involves the phosphorylase branch of the ancient²² UDP-glucose-dependent glycosyl transferases (Fig. 5b). In this family, a previously unrecognized α - to π -helix conversion places a Trp residue at the pyridoxal phosphate binding site and its occurrence in the superfamily correlates perfectly with the transition from UDP-glucose dependence to pyridoxal phosphate dependence. In mercuric ion reductases (Fig. 5c), an α - to π -helix conversion that is clearly due to the insertion of a single residue compared to more ancient members of the pyridine nucleotide linked disulfide reductases places a key catalytic Tyr residue into the mercury binding site²³. In acetylcholine esterase (Fig. 5d), a previously unrecognized α - to π -helix conversion unique to this subgroup of the α/β hydrolase superfamily is associated with a bend in a helix that changes the shape of the active site pocket and allows a new chain segment to supply the glutamate of the Glu-His-Ser catalytic triad. In addition to these three examples, the previously described functionally important α -aneurisms and α -bulges (which can now all be called π -helices) from heat shock transcription factor¹⁶, halo- and bacteriorhodopsin^{24; 25}, and a subgroup of peroxidoredoxins²⁶ are also present and conserved within all members of a branch of broader protein families or superfamilies.

A rationale for π -helix association with function—This novel explanation for the origin of π -helices provides a rationale for their association with functional sites in proteins.

Because π helix formation via the insertion of a residue has been observed to destabilize a protein by $\sim 3\text{--}6$ kcal/mol^{13; 14}, corresponding to a 100- to 10,000-fold decrease in the population of the folded state, π -helices would tend to be selected against unless they were associated with a functional advantage. This association could be direct, in that the π -helix forming mutation is directly adaptive, or it could be indirect in that the π -helix forming mutation is a non-adaptive change resulting from genetic drift in a sufficiently stable protein¹ but which sets the stage for a subsequent adaptive mutation². We are aware of no other such motif or class of mutation that as a group is so consistently associated with function, making π -helices powerful novel markers for mapping the evolution of and identifying unique functionalities associated with particular branches of protein families. As such, π -helices (e.g. Supplemental Tables 1 and 2, and the Supplemental PDB Survey for π -helices) provide fertile ground for evolutionary and functional analyses of many protein families.

Peristaltic-like shifts as a novel functionality of the π -helix—Our refined definition of π helices also brings to light a unique structural versatility that can contribute to protein function. The hydroxylase component of soluble methane monooxygenase (MMOH), a member of the bacterial multicomponent monooxygenases (BMM), has been reported to have an active site π helix that extends upon the binding of a product analog²⁷. Closer inspection of this active site segment shows that actually no π -helical extension occurs. Instead what was thought to be one long π -helix consists of two overlapping π -helices (designated π B and π D), and during ligand binding one of them (π D) is shifted six residues in the C-terminal direction in a manner reminiscent of esophageal peristalsis (Fig. 6a). Such a peristaltic shift of π -helices has not been noted before. Interestingly, our inspection of the related toluene-4-monooxygenase hydroxylase structure²⁸ shows that the binding of the regulatory subunit (common to all BMMs) causes an equivalent active site π -helical shift (Fig. 6b) that enlarges the buried active site cavity, presumably preparing it for substrate binding. This π -helical shift provides a plausible mechanism for how BMMs are activated by their regulatory subunits²⁹. Of further note, the active site π -helical shift in toluene-4-monooxygenase is coordinated with a shift in a neighboring π -helix (designated π E and conserved in all BMM enzymes) that appears to transmit the effect of regulatory subunit binding to the active site (Fig. 6b). This controlled pushing and pulling of π -helices in and around the buried active site of BMM enzymes provides an example of how a π -helix not directly in an active site can still specifically contribute to function.

The power of π -helices as evolutionary markers: the ferritin-like superfamily

To illustrate the kinds of insights that can come from this awareness of π -helices, we explore here the structure-function relations and evolutionary history of the BMM family for which the best studied representative, MMOH, has a remarkable twelve π -helices within its homologous α - and β -subunits. The BMM enzymes belong to the ferritin-like superfamily, which is characterized by a conserved four-helix bundle core (referred to as helices A, B, C and D) with a carboxylate-bridged di-nuclear center usually coordinated by six protein side chains³⁰. This superfamily includes the more ancient rubrerythrins, Δ^9 -desaturases, ferritins and bacterioferritins, as well as the younger class I ribonucleotide reductase R2 subunits (RNR) and BMMs^{31–34}. Using a structure-based sequence alignment of proteins from each of the major families within this broad superfamily, we generated a maximum likelihood phylogenetic tree onto which we mapped a minimalist set of gains and losses of π -helices (Figs. 7a and 7b).

Origin of the ferritin-like superfamily—A first major insight derived by considering π helices is a concrete proposal for the peptide-based origin of this superfamily. The root of the tree is placed near the rubrerythrins because internal symmetry in the sequence of a

rubrerythrin-like protein, erythrin, from the primitive eukaryote *Cyanophora paradoxa* (31% sequence identity between homologous helices) implies that the core four-helix bundle of the superfamily developed via the duplication and fusion of an ancient gene encoding a two-helix chain that formed a dimer³¹ (Fig. 8). Here, we extend this inference by noting that rubrerythrin's unique seventh iron ligating residue (a Glu in core helix C) is located on a π -helix (called πA_2), and furthermore, that the sequence of erythrin not only conserves this seventh ligating residue but also has an additional internal symmetry-related Glu residue inserted in core helix A (Fig. 8b). This previously unrecognized feature of erythrin strongly supports the conclusion that this original ancestor of the ferritin-like superfamily had a higher internal symmetry involving two π helices and an unprecedented eight metal ligating residues. These two π -helices would have been generated from a single insertion into the primordial 2-helix protein (πA Fig. 8a) which upon gene duplication and fusion became helices πA_1 in helix A and πA_2 in helix C. An —insertion at the level of such a peptide would not be distinct from a conformational adjustment that creates the π -helix during dimerization and iron binding. Given that the dominant pattern for characterized superfamily members is to have six ligating residues, with rubrerythrin being considered unusual due to its seventh ligating residue, this proposal for an ancestral di-iron center containing eight ligands is unanticipated and demonstrates the power that considering π helices can have in evaluating evolutionary phenomena. The existence of erythrin as a modern day protein is remarkable and suggests it has experienced a slow rate of change making it the protein-equivalent of a living fossil, which is further supported by the observation that it is found only organisms that themselves are a kind of living fossil (*C. paradoxa* and *Gloeobacter violaceus*). This concrete hypothesis for a novel metallocenter structure in erythrin not only defines a peptide-based origin of this protein fold, but also provides clues to the novel chemistry associated with the development of oxygenic photosynthesis on earth.

Correlation between π -helices and metallocenter geometry—A second major insight derived from considering π -helices has to do with the development of the distinct chemistry of the different families within this superfamily. With the root of the tree placed near the erythrins, one can follow from there the formation/deletion of π -helices through the superfamily (Fig. 7a). This history shows mostly additions of π -helices, but in three cases – πA_1 , πA_2 and πC – π helical insertions are lost. Remarkably, every one of the π -helices in the core of the tree (πA - πD) are at the metallocenter where their addition or loss would directly influence metallocenter geometry and thus active site chemistry (Fig. 7b; Fig. 9). πA_2 contributes the seventh ligating residue characteristic to rubrerythrins and presumably erythrin but is not present in any other branches. πB from core-helix C contributes a ligating Glu residue (E193 in Fig. 9) which displays oxidation-state-dependent coordination in RNR Ia/b enzymes: in the oxidized (diferric) state this iron-Glu interaction is monodentate, but in the reduced (diferrous) state the interaction is bidentate. However, in the more ancient enzymes that do not have πB , this Glu residue coordinates through bidentate interactions regardless of the metallocenter oxidation state. Insertion πC correlates with a residue change in the nearby ligand from Glu in more ancient enzymes to Asp in RNR Ia/b enzymes (red text in Fig. 9). Then, upon the subsequent insertion of πD and a shift in πC , this Asp ligand converts back to Glu in RNR Ic (E89) and BMM enzymes. πD also correlates with a further transition in the coordinating properties of the Glu ligand in core-helix C (E193 in RNR Ic) from oxidation-dependent in RNR Ia/b enzymes to universally monodentate regardless of the oxidation state in the BMM enzymes (blue text in Fig. 9). This monodentate interaction is critical for BMM enzymes³⁵. Interestingly, however, the function of πB and πD have extended beyond just ligand coordination chemistry once the BMM family diverged from the RNR Ic family as both these π -helices in BMMs play roles in substrate binding as well, as shown in Fig. 6.

Evolutionary steps from RNR to BMMs—A particularly interesting case involves insertion π D whose appearance is correlated with a change in both metallocenter geometry and chemistry in the RNR family. Interestingly, surveying RNR Ic-like sequences while considering the π D helix led us to discover a group of sequences (designated Ia* in Fig. 7a) which has the π D insertion yet still possesses the Tyr residue that is the defining characteristics of class Ia/b enzymes. This Ia* group represents an evolutionary link between RNR Ia and Ic enzymes, proving that the π -helical insertion preceded and thus could have played a role in the changes in active site chemistry of RNR Ic compared to the RNR Ia/b and BMM enzymes.

Due to the metallocenter similarity of BMMs and RNR class Ic, it has been hypothesized that the BMMs diverged from RNR class Ic^{36; 37}. Here, we note that the presence of π D in both the BMMs and RNR class Ic reinforces this relationship and supports this more parsimonious path for the evolution of these enzymes (see dotted line in Fig. 7a), even though the branching pattern based purely on global sequence similarities suggests the independent generation of both π D and the accompanying active site geometry features common to BMMs and RNR Ic. The low bootstrap value of 53 for the BMM/RNR branching point indicates the relationships are sufficiently distant that the branching is not well defined, so alternate proposals may have validity. If the more parsimonious path is correct, it would imply that after the insertion of π D, the RNR Ic sequences experienced a slow rate of evolutionary change due to constraints to conserve RNR function, whereas the BMM branch experienced much more extensive sequence change associated with optimizing its new monooxygenase function. We suggest that such a conservation of an existing function also accounts for why the erythrins have diverged very little from the symmetric precursor of the superfamily while the other families — which developed novel functions — diverged much more over this same time period.

Vestigial π -helices in BMM β -subunits—Considering the roles of π -helices in BMM evolution, it is also interesting that this evolutionary history shows that the BMM β -subunits, even though they no longer bind metals, have retained π -helices π B and π D as vestigial features. This suggests that although the insertion of a residue to create a π -helix is energetically unfavorable, once a π -helix has been accepted and the protein has evolved to accommodate it, then its removal becomes unfavorable. This was experimentally documented for the π -helix (α bulge) in heat shock factor, for which deleting a residue did create an α -helix but significantly destabilized the protein¹⁶. Finally, we note that this tree provides concrete examples of two non-active site π -helices that play functional roles: π E in the BMM enzymes, whose function in regulating intersubunit interactions is described above, and the previously unrecognized π M in eukaryotic ferritins, which occurs at a three-fold subunit interface and plays a role in creating an iron deposition channel unique to those ferritins³¹.

Outlook

The examples highlighted in this analysis provide compelling evidence that the overwhelming majority of π -helices are evolutionarily related to α -helices; the examples also illustrate the power that an awareness of this relationship brings to enhancing our understanding of protein structure, function and evolution, even for protein families such as glycogen phosphorylase and ribonucleotide reductase that have been extremely well studied. In the case study of the ferritin-like superfamily, the explicit consideration of π -helices provides many novel insights from specific predictions of a π -helix-containing dimeric primordial precursor of the family to the involvement of π -helices in key steps of functional diversification and the characterization of peristaltic-like shifts in the activation and substrate binding of the BMM family. This case study also underscores that α - and π -helices

interconvert both directions during evolution, with the predominant direction being the creation of π -helices, and that, as can be rationalized from first principles, their gain and loss tends to be correlated both temporally (in the evolutionary process) and spatially (in the protein structure) with changes in functionality.

This analysis demonstrates the insertional origin of π -helices as a general phenomenon of protein evolution. It also resolves long standing debates about the existence of π -helices in proteins and paves the way for this generally overlooked motif to become a noteworthy feature that will aid tracing the evolution of many protein families, guide investigations of protein and π -helix functionality, and contribute additional tools to the protein engineering toolkit. Given the functional and evolutionary importance of π -helices, we recommend that contrary to current practice, π -helices should be given precedence over α -helices in secondary structure assignment. This will allow them to be identified rather than overlooked and for those applications in which π -helices are not of interest, all π -helical residues could be counted as α -helical. Until this change happens, our π -HUNT script can be used to identify π -helices based on DSSP output. Toward the goal of identifying these cryptic and functionally relevant structures, we provide an initial framework in the form of a comprehensive list of 2967 π -helices present in over 2400 chains (see Supplemental Data Set) that was derived from our analysis of all chains in the PDB with <90% sequence identity.

Materials and Methods

The initial π -helix dataset

We have adopted the well accepted definition that helices (α , 3_{10} and π) are fundamentally defined by the presence of two sequential main chain hydrogen bonds of the correct geometry³⁸. This is the criteria used by the common automated secondary structure assignment programs DSSP¹⁷ and STRIDE¹⁸, and is also the criterion used by Fodje & Al-Karadaghi⁶. For the identification of π -helices, we use here the standard criteria that these consecutive main chain hydrogen bonds be between residues five-apart in sequence (called π -type H-bonds) and for residues internal to the π -helix, the (ϕ, Ψ) -values lie within the broadly defined α -helical region of the Ramachandran plot.

The 104 naturally occurring π -helices previously identified by Fodje & Al-Karadaghi⁶ were reclassified using the precise definitions based on backbone H-bond information as annotated by DSSP¹⁷. As backbone NH groups can be involved in H-bonded interactions with more than one acceptor, we only considered as —qualifying those π -type H-bonds for which the π -type H-bond was the strongest of the H-bonds from that donor. Qualifying π -type H-bonds were further described as strong (≤ -2.0 kcal/mol), medium (-1.9 to -1.0 kcal/mol) and weak (-0.9 to -0.5 kcal/mol). A π -helix was initiated by two sequential qualifying π -type H-bonds where one of them had to be of medium strength or stronger. Any break in the π -type H-bonding ended a π -helix. Figure 2 provides an example of this classification process.

PDB searches using π -HUNT

A non-redundant list of PDB protein chains was generated using the CullPDB server³⁹ such that (i) no two protein chains had greater than 25% sequence similarity, (ii) every structure was determined at or better than 2.5 Å using x-ray crystallography, and (iii) the associated R-factor for each structure was <0.25. Another non-redundant list of PDB chains was generated in the same manner except that no two chains had greater than 90% sequence identity. A Perl script was written to analyze the DSSP file of each structure in these lists (files were obtained from <ftp://ftp.cmbi.kun.nl/pub/molbio/data/dssp/>) and return π -helices

based on π -type hydrogen bonding patterns and (ϕ, Ψ) -angles. In terms of π -type H-bonding patterns, three different criteria were used: WW (at least two sequential π -type hydrogen bonds with strengths ≤ -0.5 kcal/mol), MW (at least two sequential π -type hydrogen bonds where the strength of at least one is ≤ -1.0 kcal/mol and the other is at least ≤ -0.5 kcal/mol) and MM (at least two sequential π -type hydrogen bonds with strengths ≤ -1.0 kcal/mol). In terms of the (ϕ, Ψ) -criteria, torsion angles of residues internal to the π -helix were restricted to the broad helical region as defined by STRIDE¹⁸ ($-180^\circ < \phi < 0^\circ$, $-120^\circ < \Psi < 45^\circ$). This script, called π -HUNT, is available from the authors upon request. Searches were also performed that returned only those residues formally annotated by DSSP as π -helical. Comparison of the results from these searches was used to assess the robustness of each criterion.

Identification of α -/ π -helical homologous pairs

To explore the evolutionary relationships between naturally occurring π -helices and α -helices, homologs of π -helix-containing proteins with one less residue and an α -helix in the region homologous to the π -helix were initially identified by visual inspection of the structure-based sequence alignments provided by the Families of Structurally Similar Proteins (FSSP) database¹⁹. Homology between identified α -/ π -helix pairs was then confirmed by SCOP classification³⁰, visual inspection, and consideration of similarity in protein function.

Phylogenetic analysis of the ferritin-like superfamily

Representative protein structures for each of the major families within the ferritin-like superfamily³² were selected from the Protein Data Bank⁴⁰. Structure-based sequence alignments for each of these proteins were obtained from the FSSP database¹⁹ and manually adjusted if needed to provide consistency in the multiple sequence alignment. For the proteins in Figure 7a without a known structure, sequences were added to the structure-based sequence alignment profile using the CLUSTALW algorithm⁴¹. The phylogenetic tree was then made as previously reported³³ using maximum likelihood methods with the JTT model for amino acid substitutions⁴².

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Joe Thornton, Jacque Fetrow and Brian Matthews for critique of an earlier manuscript. This work was supported in part by the General Medical Institute, NIH grant GM R01-083136 and the Oregon Agricultural Experiment Station.

Abbreviations

H-bonds	hydrogen bonds
MMOH	soluble methane monooxygenase hydroxylase
PH	phenol hydroxylase
ToMO	toluene monooxygenase
BMM	bacterial multi-component monooxygenases
RNR	ribonucleotide reductase

DSSP	Definition of Secondary Structure in Proteins
FSSP	Families of Structurally Similar Proteins

References

1. Bloom JD, Arnold FH. In the light of directed evolution: pathways of adaptive protein evolution. *Proc Natl Acad Sci U S A*. 2009; 106(Suppl 1):9995–10000. [PubMed: 19528653]
2. Dean AM, Thornton JW. Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet*. 2007; 8:675–88. [PubMed: 17703238]
3. Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol*. 2009; 10:709–20. [PubMed: 19756040]
4. Low BW, Baybutt RB. The π -helix- A Hydrogen Bonded Configuration of the Polypeptide Chain. *J Am Chem Soc*. 1952; 74:5806–5807.
5. Hollingsworth SA, Berkholz DS, Karplus PA. On the occurrence of linear groups in proteins. *Protein Sci*. 2009; 18:1321–5. [PubMed: 19472372]
6. Fodje MN, Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the π -helix. *Protein Eng*. 2002; 15:353–8. [PubMed: 12034854]
7. Creighton, TE. *Proteins: Structures and Molecular Properties*. Freeman and Company; New York: 1993. p. 182-186.
8. Weaver TM. The π -helix translates structure into function. *Protein Sci*. 2000; 9:201–6. [PubMed: 10739264]
9. Armen R, Alonso DO, Daggett V. The role of α -, 3_{10} -, and π -helix in helix $\rightarrow$ coil transitions. *Protein Sci*. 2003; 12:1145–57. [PubMed: 12761385]
10. Mahadevan J, Lee KH, Kuczera K. Conformational Free Energy Surfaces of Ala₁₀ and Aib₁₀ Peptide Helices in Solution. *J Phys Chem B*. 2001; 105:1863–1876.
11. Lee KH, Benson DR, Kuczera K. Transitions from α to π helix observed in molecular dynamics simulations of synthetic peptides. *Biochemistry*. 2000; 39:13737–47. [PubMed: 11076513]
12. Feig M, MacKerell AD Jr, Brooks CL III. Force field influence on the observation of π -helical protein structures in molecular dynamics simulations. *J Phys Chem B*. 2003; 107:2831–2836.
13. Keefe LJ, Sondek J, Shortle D, Lattman EE. The α -aneurism: a structural motif revealed in an insertion mutant of *Staphylococcal nuclease*. *Proc Natl Acad Sci U S A*. 1993; 90:3275–9. [PubMed: 8475069]
14. Heinz DW, Baase WA, Dahlquist FW, Matthews BW. How aminoacid insertions are allowed in an α -helix of T4 lysozyme. *Nature*. 1993; 361:561–4. [PubMed: 8429913]
15. Cartailler JP, Luecke H. Structural and functional characterization of π -bulges and other short intrahelical deformations. *Structure*. 2004; 12:133–44. [PubMed: 14725773]
16. Hardy JA, Walsh ST, Nelson HC. Role of an α -helical bulge in the yeast heat shock transcription factor. *J Mol Biol*. 2000; 295:393–409. [PubMed: 10623534]
17. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983; 22:2577–637. [PubMed: 6667333]
18. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins*. 1995; 23:566–79. [PubMed: 8749853]
19. Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v.3. *Bioinformatics*. 2008; 24:2780–1. [PubMed: 18818215]
20. Barlow DJ, Thornton JM. Helix geometry in proteins. *J Mol Biol*. 1988; 201:601–19. [PubMed: 3418712]
21. Yamashita A, Singh SK, Kawate T, Jin Y, Gouaux E. Crystal structure of a bacterial homologue of Na⁺/Cl⁻-dependent neurotransmitter transporters. *Nature*. 2005; 437:215–23. [PubMed: 16041361]

22. Gibson RP, Turkenburg JP, Charnock SJ, Lloyd R, Davies GJ. Insights into trehalose synthesis provided by the structure of the retaining glucosyltransferase OtsA. *Chem Biol.* 2002; 9:1337–46. [PubMed: 12498887]
23. Rennex D, Cummings RT, Pickett M, Walsh CT, Bradley M. Role of tyrosine residues in Hg(II) detoxification by mercuric reductase from *Bacillus* sp. strain RC607. *Biochemistry.* 1993; 32:7475–8. [PubMed: 8338845]
24. Luecke H, Schobert B, Richter HT, Cartailler JP, Lanyi JK. Structure of bacteriorhodopsin at 1.55 Å resolution. *J Mol Biol.* 1999; 291:899–911. [PubMed: 10452895]
25. Kolbe M, Besir H, Essen LO, Oesterhelt D. Structure of the light-driven chloride pump halorhodopsin at 1.8 Å resolution. *Science.* 2000; 288:1390–6. [PubMed: 10827943]
26. Sarma GN, Nickel C, Rahlfs S, Fischer M, Becker K, Karplus PA. Crystal structure of a novel *Plasmodium falciparum* 1-Cys peroxiredoxin. *J Mol Biol.* 2005; 346:1021–34. [PubMed: 15701514]
27. Sazinsky MH, Lippard SJ. Product bound structures of the soluble methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath): protein motion in the α -subunit. *J Am Chem Soc.* 2005; 127:5814–25. [PubMed: 15839679]
28. Bailey LJ, McCoy JG, Phillips GN Jr, Fox BG. Structural consequences of effector protein complex formation in a diiron hydroxylase. *Proc Natl Acad Sci U S A.* 2008; 105:19194–8. [PubMed: 19033467]
29. Cooley RB, Dubbels BL, Sayavedra-Soto LA, Bottomley PJ, Arp DJ. Kinetic characterization of the soluble butane monooxygenase from *Thauera butanivorans*, formerly '*Pseudomonas butanovora*'. *Microbiology.* 2009; 155:2086–96. [PubMed: 19383682]
30. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.* 1995; 247:536–40. [PubMed: 7723011]
31. Andrews, S. Iron Storage in Bacteria. In: Poole, RK., editor. *Advances in Microbial Physiology.* Vol. 40. Academic Press; San Diego: 1998. p. 281–349.
32. Gomes CM, Le Gall J, Xavier AV, Teixeira M. Could a diiron-containing four-helix-bundle protein have been a primitive oxygen reductase? *Chembiochem.* 2001; 2:583–7. [PubMed: 11828492]
33. Leahy JG, Batchelor PJ, Morcomb SM. Evolution of the soluble diiron monooxygenases. *FEMS Microbiol Rev.* 2003; 27:449–79. [PubMed: 14550940]
34. Wiedenheft B, Mosolf J, Willits D, Yeager M, Dryden KA, Young M, Douglas T. An archaeal antioxidant: characterization of a Dps-like protein from *Sulfolobus solfataricus*. *Proc Natl Acad Sci U S A.* 2005; 102:10551–6. [PubMed: 16024730]
35. Schwartz JK, Wei PP, Mitchell KH, Fox BG, Solomon EI. Geometric and electronic structure studies of the binuclear nonheme ferrous active site of toluene-4-monooxygenase: parallels with methane monooxygenase and insight into the role of the effector proteins in O₂ activation. *J Am Chem Soc.* 2008; 130:7098–109. [PubMed: 18479085]
36. Andersson CS, Hogbom M. A *Mycobacterium tuberculosis* ligand-binding Mn/Fe protein reveals a new cofactor in a remodeled R2-protein scaffold. *Proc Natl Acad Sci U S A.* 2009; 106:5633–8. [PubMed: 19321420]
37. Hogbom M, Stenmark P, Voevodskaya N, McClarty G, Graslund A, Nordlund P. The radical site in chlamydial ribonucleotide reductase defines a new R2 subclass. *Science.* 2004; 305:245–8. [PubMed: 15247479]
38. IUPAC. IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969). *Biochemistry.* 1970; 9:3471–9. [PubMed: 5509841]
39. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics.* 2003; 19:1589–91. [PubMed: 12912846]
40. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.* 2000; 28:235–42. [PubMed: 10592235]

41. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994; 22:4673–80. [PubMed: 7984417]
42. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 1992; 8:275–82. [PubMed: 1633570]
43. Kurtz DM Jr, Prickril BC. Intrapeptide sequence homology in rubrerythrin from *Desulfovibrio vulgaris*: identification of potential ligands to the diiron site. *Biochem Biophys Res Commun.* 1991; 181:337–41. [PubMed: 1958203]

Fig. 1.

The common two-H-bond π -helix is the same motif as an engineered α -aneurism. a) An α -helix in *Staphylococcus aureus* nuclease (left, green cartoon; PDB 1eyd) develops an α -aneurism (right, purple cartoon; PDB 1sty) when a single amino acid (highlighted blue) is inserted into it¹³. b) An overlay of the π -helix from fumarase C of *Escherichia coli* (residues 155–161, orange carbon atoms; PDB 1fur) and the same engineered α -aneurism from panel a (purple carbon atoms) demonstrates their equivalence. These α -aneurism-type helical distortions have been independently characterized as looping outs¹⁴, π -bulges¹⁵ and α -bulges¹⁶, but not as π -helices. Black-dashed lines represent the π -type H-bonds. (ϕ , Ψ) torsion angles are written beside each residue. In panel b, nitrogen (blue) and oxygen (red) atoms are indicated.

Fig. 2. Defining π -helices based on (i+5,i) π -type H-bonds. a) Energy cutoffs (in kcal/mol, calculated by DSSP¹⁷) for strong (s), medium (m), and weak (w) H-bonds. b) Excerpted DSSP output for residues 245–267 of methane monooxygenase hydroxylase (MMOH) (PDB 1mtv, β chain), highlighting the π -type H-bonding patterns. Green boxes highlight π -type donor interactions and red boxes highlight the acceptor interactions. Solid outlines indicate the π -type interaction is the strongest H-bond for that residue, and dotted outlines indicate those that are the second strongest. The column contents are labeled. For each of the four H-bond interaction columns, the two numbers given are relative positions of the partner residue and the energy of the interaction (in kcal/mol). The first two of these columns are the strongest interactions and the second two columns are the second strongest interactions. c) Short hand summary of the π -type H-bonding and the π -helix assignments. For each residue acting as a π -type donor or acceptor, the strength of the interaction is given in the appropriate column. The letter is in parentheses if the π -type interaction is not the strongest H-bond for that residue. Based on the qualifying criteria described in the *Materials and Methods*, this segment in MMOH that was previously classified as a single 20 residue π -helix with 5-(i+5,i) H-bonds⁶ is seen to be three distinct, but overlapping π -helices with 2-, 2- and 4-H-bonds (see rectangles in panel c). Note that these three π -helices are cryptic in that none are assigned as π -helical (I) in the DSSP summary (fourth column).

Fig. 3.

A single insertion can generate all occurring lengths of π -helices. Panels a, b, c, d, e and f show one example each of a 2-, 3-, 4-, 5-, 6- and 7-H-bond π -helix, respectively. In each panel, the equivalent α -helix (left) and π -helix (right) from a pair of homologous proteins are shown with their PDB codes. Above the black arrow that represents the insertion process is the sequence alignment from the FSSP database illustrating the single insertion. The π -helices from panels a (formaldehyde ferredoxin oxidoreductase), b (glycogen phosphorylase), and d (MMOH α -subunit; insertion π D from Fig. 7) are all located within the active site of their respective protein, while that shown in panel c (MMOH α -subunit) and panel e (α -subunit of Toluene-4-Monooxygenase; insertion π E from Fig. 7) are involved in intersubunit interactions, suggesting a functional role for these. All atoms in the α -helices are colored green, while carbon, nitrogen and oxygen atoms in the π -helices are colored orange, blue and red, respectively. Black dashed lines in the π -helices indicate the (i+5,i) π -type hydrogen bonds.

Protein	PDB	Residues	Alignment	Structure
Bacterioferritin	1bfr	86-105	MLRSD.LALEL.DGAKNL.REAI
Class Ib RNR	1uzr	155-175	RKVAS.TLLESFLFYSGF.YLPM
Class Ic RNR	1syy	184-205	NLVGYIIMEGIFFYSGF.VMIL
MMOH – β -subunit	1mty	244-266	SAFSVHAVYDALFGQFVRREFFQ

Fig. 4.

Evidence that overlapping π -helices result from stepwise insertions of single amino acids. In the ferritin-like superfamily, putative ancestral forms in the evolution of the three, overlapping π -helices in the β -subunit of MMOH (shown in Fig. 2) (bottom) are represented by the structurally equivalent helices from bacterioferritin (top), which is a pure α -helix, the R2 subunit of class Ib ribonucleotide reductase (second from top), which includes one π -helix, and the R2 subunit of class Ic ribonucleotide reductase (third from top), which includes overlapping π helices. Insertion events represented are designated π B, π D and π J in Figure 7. Amino acid alignments indicate the effective points of insertion. Lines below each sequence indicate the π helical regions. π -helices are colored orange, α -helices are colored green, and the inserted amino acid, as defined by structural alignment, is colored in blue.

Fig. 5.

Three examples of α -helix derived active site π -helices. a) Evolutionary path by which π -helices are derived from α -helices via a single residue insertion (orange circle). b) Shown is the π -helix-containing PLP-bound active site of PLP-dependent glycogen phosphorylase (orange carbon atoms; PDB 3gpb) overlaid onto the α -helix-containing active site of trehalose-6-phosphate synthase (semitransparent green carbon atoms; PDB 1uqu). The structurally derived sequence alignments are shown below the structure (orange and green letters correspond to the π - and α -helix-containing sequences, respectively), which demonstrate the presence of the inserted residue that forms the π -helical segment (orange bar). c) The active site π -helix of mercuric ion reductase (orange carbon atoms; PDB 1zk7) is overlaid on the homologous α -helix of dihydrolipoamide dehydrogenase (semitransparent green carbon atoms; PDB 1ebd). d) A π -helix of acetylcholine esterase (orange carbon atoms; PDB 1ea5) is overlaid onto the homologous α -helix of mycolyl transferase (green transparent carbon atoms; PDB 1f0p). In all panels, dashed-black lines show the π -type H-bonds and red dashed lines show interactions within the active sites. Nitrogen (blue), oxygen (red), sulfur (yellow) and phosphorus (purple) atoms are indicated.

Fig. 6.

π -helical peristalsis in the active site of BMM enzymes. A) Two π -helices (labeled π B and π D) in the active site of MMOH are overlapping in the resting state (left). Upon binding the product analogue 6-bromohexan-1-ol (blue spheres), these two π -helices no longer overlap due to a shift in π B (right). B) The resting state structure of the toluene-4-monooxygenase hydroxylase also shows two overlapping π -helices at the active site (left). Upon binding of the regulatory component (ToMOD, semitransparent yellow), π B shifts identically to that of product-bound MMOH. As π B moves downward, an adjacent π -helix (labeled π E), which is sandwiched between the active site π -helix and the regulatory subunit, simultaneously elongates and shifts upward. For details of the π -type H-bonds, see Supplemental Figure 3. Orange bars mark individual π -helices and blue arrows the direction of π -helix movement. π -type H-bonds are shown by black dashed lines. π -helix (orange) and α -helix carbon atoms are indicated, as are nitrogen (blue), oxygen (red), iron (purple) and bromine (brown). PDB codes are 1mty, 1xvb, 3dhg and 3dhi.

Fig. 7.

π -helices and the evolution of the ferritin-like superfamily. a) Evolutionary tree showing that, from a symmetrical erythrin-like ancestor containing two π -helices (Fig. 8), a simple set of insertion (orange circles) and deletion (crossed-out orange circles) events can account for all π -helices seen in modern-day family members (Supplemental Table 5). The branching shown is based on maximum likelihood analysis as previously described³³ from structurally-derived sequence alignments (bootstrap values shown). Organism names, omitted for clarity, are shown in Supplemental Figure 4. The length of the bar (bottom right) corresponds to 0.5 substitutions/site. New abbreviations are PH: phenol hydroxylase family and ArH: aromatic monooxygenase hydroxylase family. b) Ribbon diagram of the α -subunit of MMOH (PDB 1mtj, chain D) showing the locations of its six π -helices (π B, π D, π E- π H) and also showing the locations of the remaining eight π -helices seen in other members of the ferritin-like superfamily (π A₁, π A₂, π C, π I- π M). Indicated are the core four-helix bundle (green), the surrounding helices (transparent gray), and the π -helices π A₁ (yellow), π A₂ (light blue), π C (cyan), π B and π D (blue), π E and π G- π L (orange), π F (purple) and π M (brown). Iron atoms (pink spheres) and the N- and C-termini are shown.

Fig. 8.

Model for the origins of the ferritin-like superfamily. a) Scheme showing the formation of the four-helix bundle characteristic of the ferritin-like superfamily via gene duplication and fusion of a primordial two-helix precursor that, through a single insertion π A (orange circle), gave rise to two symmetry related π -helices (πA_1 and πA_2) and eight ligating residue in the erythrin-like precursor of this superfamily. b) Sequence alignments showing the residual internal sequence similarity in erythrins and, to a lesser degree, rubrerythrin⁴³. Shown are the alignments of helix A with helix C and helix B with helix D in erythrin from *Cyanophora paradoxa* (CP Ery), erythrin from *Gloeobacter violaceus* (GV Ery), rubrerythrin from *Desulfovibrio vulgaris* (Rub) and bacterioferritin from *Escherichia coli* (Bfr) included as an example of a canonical superfamily member having only six ligating residues and less recognized internal symmetry. Red highlighting indicates residues involved (or proposed to be involved in the case of erythrin) in iron-ligation. Green highlighting indicates additional residues that are identical between homologous helices within the same protein. Orange bars indicate π -helical residues in rubrerythrin, and potential π -helical residues in erythrin.

Fig. 9.

π -helices πA_2 - πD correlate with changes in metallocenter geometry. Shown is an annotated view of the di-nuclear site in RNR Ic (PDB 1syy) with the backbone ribbon of core-helix B of the four-helix bundle, contributing H123 and E120, omitted for clarity. Core helices A, C and D and their direction from N- to C-terminus are labeled on the left, and π - (orange) and α -helices (green) are indicated. As described in the text, the proximity of πA_2 - πD to their respective di-nuclear centers are shown and suggest they play a role in influencing metallocenter chemistry.

Table 1Classification of π -helices in the representative dataset and the current Protein Data Bank

H-bonds in π -helix ^a	Number of π -helices ^b	Number with α -helical homologs ^c	Protein Data Bank ^d
2	82	70	822
3	13	9	138
4	7	5	26
5	4	4	20
6	0	0	3
7	0	0	1
Total	106	88	1010

^aThe number of (i+5) $\rightarrow$ i hydrogen bonds in the π -helix. To qualify as a π -helix, at least two sequential (i+5) $\rightarrow$ i hydrogen bonds must exist where one is ≤ -0.5 kcal/mol and the other is ≤ -1.0 kcal/mol

^bFrom the reclassified Fodje & Al-Karadaghi⁶ dataset

^cPutative homologs of the π -helix-containing proteins from the Fodje & Al-Karadaghi⁶ dataset with a pure α -helix at the equivalent position.

^dNumber of π -helices in 5620 diverse ($<25\%$ sequence identity with one another) protein chains determined at ≤ 2.5 Å resolution.