doi:10.1016/S0022-2836(02)00885-9 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 323, 69–76

JMB

GXXXG and GXXXA Motifs Stabilize FAD and NAD(P)-binding Rossmann Folds Through C^{α}–H···O Hydrogen Bonds and van der Waals Interactions

Gary Kleiger and David Eisenberg*

Howard Hughes Medical Institute, UCLA-DOE Center for Genomics and Proteomics Molecular Biology Institute UCLA, Box 951570, Los Angeles, CA 90095-1570, USA

*Corresponding author

Here we present evidence that domains in soluble proteins containing either the GXXXG or GXXXA motif are stabilized by the interaction of a β-strand with the following α-helix. As an example, we characterized a β-strand–helix interaction from the FAD or NAD(P)-binding Rossmann fold. The Rossmann fold is one of the three most highly represented folds in the Protein Data Bank (PDB). A subset of the proteins that adopt the Rossmann fold also bind to nucleotide cofactors such as FAD and NAD(P) and function as oxidoreductases. These Rossmann folds can often be identified by the short amino acid sequence motif, $GX_{1-2}GXXG$. Here, we present evidence that in addition to this sequence motif, Rossmann folds that bind FAD and NAD(P) also typically contain either GXXXG or GXXXA motifs, where the first glycyl residue of these motifs and the third glycyl residue of the $GX_{1-2}GXXG$ motif are the same residue. These two motifs appear to stabilize the Rossmann fold: the first glycyl residue of either the GXXXG or GXXXA motif contacts the carbonyl oxygen atom from the first glycyl residue of the $GX_{1-2}GXXG$ motif consistent with the formation of a C^{α}–H···O hydrogen bond. In addition, both the glycyl and alanyl residues of the GXXXG or GXXXA motifs form van der Waals interactions with either a valine or isoleucine residue located either seven or eight residues further back along the polypeptide chain from the first glycine of the GXXXG or GXXXA motifs. Therefore, we combine both the $GX_{1-2}GXXG$ and GXXXG/A motifs into an extended motif, $V/IXGX_{1-2}GXXGXXXG/A$, that is more strongly indicative than previously described motifs of Rossmann folds that bind FAD or NAD(P). The $V/IXGX_{1-2}GXXGXXXG/A$ motif can be used to search genomic sequence data and to annotate the function of proteins containing the motif as oxidoreductases, including proteins of previously unknown function.

Keywords: sequence motif; Rossmann fold; hydrogen bond; FAD-binding; NAD(P)-binding

Introduction

GXXXG is an amino acid sequence motif that stabilizes helix–helix interactions in both membrane and soluble proteins. The GXXXG sequence motif has been found in transmembrane α-helices, some 32% above expectation,[1] and has been determined to stabilize the oligomerization of several membrane proteins, such as glycophorin A,[2] human carbonic anhydrase,[3] and members of the epidermal growth factor receptor family.[4]

Recently, we discovered that the GXXXG motif stabilizes helix–helix interactions in soluble proteins, where occurrences of the GXXXG motif in α-helices are observed 41% above expectation in known protein structures.[5] Some 26 protein structures from the non-redundant Protein Data Bank (PDB) contain a helix–helix interaction stabilized by the GXXXG motif.

Abbreviations used: PDB, Protein Data Bank; FSSP, fold classification based on structure–structure alignment of proteins (http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html); ORF, open reading frame; CATH, class, architecture, topology and homologous superfamily protein structure classification (http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html).

E-mail address of the corresponding author: david@mbi.ucla.edu

0022-2836/02/$ - see front matter

The stabilization of helix–helix interactions containing the GXXXG motif is based on the four-residue separation of the glycyl residues in the motif. This separation serves to align the glycyl residues on one face of the helix, providing a flat platform for the other helix and permits the helical axes to approach as close as 6.0 Å.[6] In this close configuration, the two glycyl C^{α} atoms are able to donate their hydrogen atoms to form hydrogen bonds with carbonyl oxygen atoms on the adjacent helix.[7] This C^{α}–H···O type hydrogen bond has been estimated to have an energy of 2.5–3.0 kcal/mol *in vacuo*, or approximately one-half the energy of an N–H···O or O–H···O hydrogen bond.[8] In addition, the close binding of both helices promotes van der Waals interactions, further stabilizing the helix–helix interaction.[9]

The term fold was introduced by Rossmann[10–12] to describe a nucleotide-binding domain found in families of oxidoreductases such as lactate dehydrogenase and flavodoxin. This fold begins with a β-strand connected by a short loop to an α-helix.[13] Here, we present evidence that the interaction of this β-strand and α-helix is frequently stabilized by either the GXXXG or GXXXA motif. Although not all Rossmann folds bind to the nucleotides FAD or NAD(P), those that do typically contain a conserved sequence motif, $GX_{1-2}GXXG$,[14] where the glycyl residues are located on the ligand-binding loop in between the β-strand and α-helix. The importance of the glycyl residues has been previously explained:[15] the first glycine allows a tight turn of the main-chain from the β-strand into the loop, and the second glycine permits close contact of the main-chain to the pyrophosphate of the nucleotide. The third glycine allows close packing of the helix with the β-strand. The third glycine of the $GX_{1-2}GXXG$ motif is also the first residue of a newly discovered motif, GXXXG/A, located on the α-helix. Together these two motifs stabilize FAD or NAD(P)-binding Rossmann folds and binding of the nucleotide cofactor to the domain.

Results

We counted 734 occurrences of the GXXXG/A motifs in all α-helices from a non-redundant PDB (172 occurrences of GXXXG and 562 occurrences of GXXXA). Occurrences of the GXXXG motif in α-helices are enriched above expectation by 41(±9)%.[5] We found that occurrences of the GXXXA motif in α-helices are enriched 53(±5)% above expectation. The GXXXA motif occurs 562 times from 12,802 total α-helices. We calculated that only 368(±17) occurrences of the GXXXA motif are expected if glycine and alanine residues were uniformly distributed in those helices. Like GXXXG, the observed enrichment suggests that at least some of the α-helices containing the GXXXA motif may stabilize structure. The AXXXG motif is not enriched above expectation (data not shown).

Approximately half of the GXXXG/A containing helices are involved in helix–helix interactions; however, 121 of the α-helices containing the GXXXG/A motif interact with at least one β-strand (20 for GXXXG and 101 for GXXXA). Some 22 helix–β-strand interactions were found in FAD or NAD(P) Rossmann folds and contain either the GXXXG (two cases) or GXXXA (20 cases) sequence motifs.

Helix–β-strand interactions containing the GXXXG or GXXXA motifs in Rossmann folds

As an example of a helix–β-strand interaction stabilized by the GXXXG motif, consider sarcosine oxidase (1el5[16]). The N-terminal domain of sarcosine oxidase adopts the Rossmann fold and binds FAD. The first α-helix (α-helix 1) contains a GXXXG motif that stabilizes an interaction between the α-helix and the first β-strand (β-strand 1) of the fold.

The helix–β-strand interaction from sarcosine oxidase appears to be stabilized by a C^{α}–H···O hydrogen bond and van der Waals interactions. The C^{α} hydrogen atom of the first glycyl residue of the GXXXG motif (Gly15) contacts a carbonyl oxygen atom on Gly10 located on the loop connecting the α-helix and the β-strand (Figure 1(a)). The distance from the C^{α} hydrogen atom to the carbonyl oxygen atom is 2.5 Å, and the distance from the C^{α} atom to the oxygen atom is 3.4 Å. In addition, the C^{α}–H···O angle is 145°. Together, these values define the formation of a C^{α}–H···O hydrogen bond between Gly15 and Gly10 (Table 1).[17] Supplementing the stabilization of the hydrogen bond is Val8, located on β-strand 1, forming van der Waals interactions with both glycyl residues of the GXXXG motif. The desolvation of Val8, Gly15 and Gly19 upon folding results in an enhanced binding energy of 1.1 kcal/mol,[18] stabilizing the interaction of α-helix 1 and β-strand 1 in the Rossmann fold of sarcosine oxidase (Figure 1(a)).

A similar helix–β-strand interaction is found in the structure of phenol hydroxylase (1foh[19]). Like sarcosine oxidase, this enzyme also contains a domain adopting the Rossmann fold. The first α-helix of the Rossmann fold in phenol hydroxylase contains a GXXXA motif that contributes to the stability of the fold in a structurally equivalent manner to GXXXG in sarcosine oxidase. The contact between the C^{α} hydrogen of the first glycyl residue of the GXXXA motif (Gly19) and the carbonyl oxygen atom of Gly14 is consistent with hydrogen bond formation (Figure 1(b); Table 1). In addition, Ile12 on β-strand 1 forms van der Waals interactions with both the glycyl and alanyl residues of the GXXXA motif, resulting in an enhanced binding energy of 1.5 kcal/mol upon desolvation of these residues. The helix–β-strand interactions of phenol hydroxylase and sarcosine oxidase are structurally equivalent, yet the

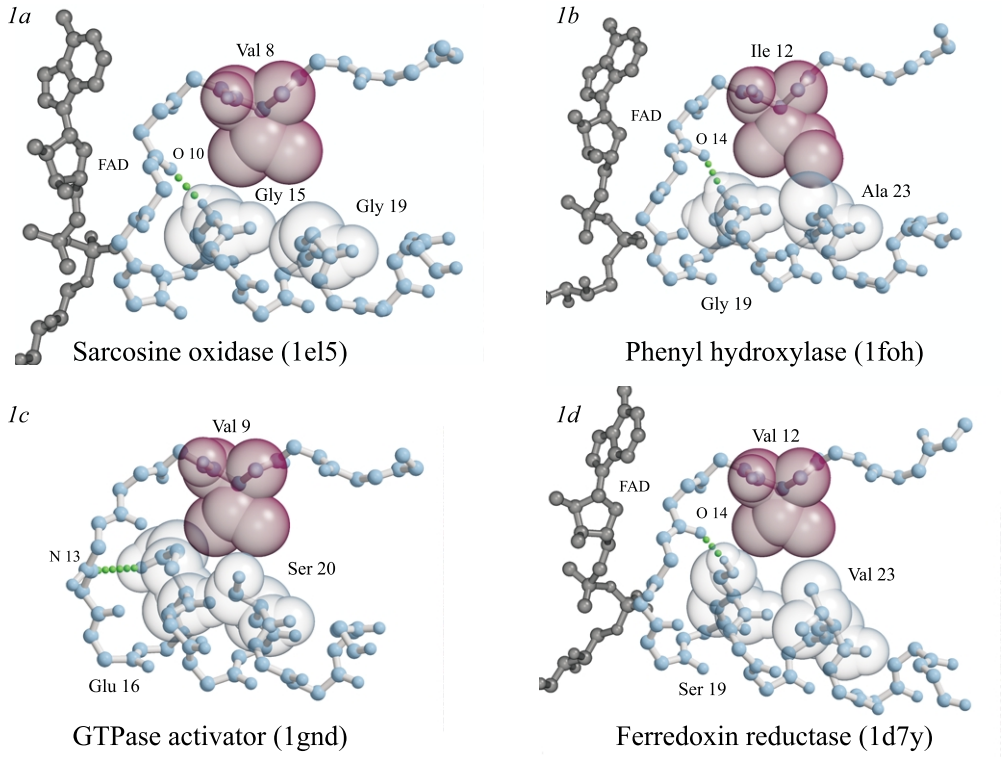

Figure 1. The FAD or NAD(P)-binding Rossmann fold often contains a C^{α}–H···O hydrogen bond that stabilizes its structure. (a) Ball-and-stick representation of β-strand 1 (horizontal on top), the ligand-binding loop (vertical on left center, containing O10) and α-helix 1 (horizontal on bottom, containing Gly15 and Gly19) of the Rossmann fold in sarcosine oxidase (1el5[16]). Gly15 and Gly19 (blue spheres) form a GXXXG motif in the α-helix. The H1A hydrogen of Gly15 and carbonyl oxygen atom of Gly10 show the stereochemical hallmarks of C^{α}–H···O hydrogen bond formation (green dots) and explain why Gly residues at these two position are almost always present in protein sequences that adopt the Rossmann fold and bind to FAD or NAD(P). Further stability to the fold is contributed by van der Waals interactions of Val8 (magenta spheres) with both glycyl residues of the GXXXG motif. (b) Ball-and-stick representation of phenol hydroxylase (1foh[19]); β-strand 1, the ligand-binding loop and α-helix 1 of the Rossmann fold are shown. The Rossmann fold in phenol hydroxylase is stabilized in a manner similar to sarcosine oxidase; a potential C^{α}–H···O hydrogen bond between the glycyl H1A atom of Gly19 and the carbonyl oxygen atom of Gly14 is observed (green dots), as are the van der Waals interactions between Ile12 (magenta spheres) and both residues of the GXXXA motif. (c) Ball-and-stick representation of the Rossmann fold for the GTPase activator (1gnd[21]). Notice that the GTPase activator contains an EXXXS sequence on α-helix 1 rather than GXXXG or GXXXA. However, Glu16 of the motif still forms a hydrogen bond to the backbone amide of Gly13 on the ligand-binding loop. van der Waals interactions between Val9 and both Glu16 and Ser20 are also observed. (d) Ball-and-stick representation of the Rossmann fold for ferredoxin reductase (1d7y[22]). Like the GTPase activator, ferredoxin reductase contains a different amino acid sequence, SXXXV, on α-helix 1. However, the OG atom of Ser19 still forms a hydrogen bond to the backbone carbonyl of Gly14 (green dots). The Rossmann folds of both the GTPase activator and ferredoxin reductase demonstrate how variation of the GXXXG motif is tolerated without compromising stability of the structure.

domains adopting the Rossmann fold share only 15% amino acid sequence identity.

Of the 22 Rossmann folds containing either the GXXXG or GXXXA motifs, 20 form contacts consistent with C^{α}–H···O hydrogen bond formation as well as van der Waals interactions that stabilize the helix–β-strand interaction of the Rossmann fold in a structurally equivalent manner (Figure 2 and Table 1). This conserved structural feature is remarkable considering that 230 of the 231 pairs of these 22 proteins have less than 30% sequence identity over their Rossmann folds. In all cases, we would anticipate that from both hydrogen bond formation and van der Waals interactions the energetic stabilization would be modest but significant, as indicated by the two examples discussed above.

Additional examples of helix–β-strand interactions in the Rossmann fold

We initially focused on the GXXXG and GXXXA motifs because of their over-representation in α-helices. To expand our analysis and include other motifs that may be structurally equivalent to GXXXG/A, we analyzed the fold classification based on structure–structure alignment of proteins (FSSP) database, which contains families of related protein structures.[20] Using the structure of the Rossmann fold of sarcosine oxidase as a probe,

Table 1. Common structural characteristics of PDB proteins containing the GXXXG and GXXXA sequence motifs

PDB	CATHid	α-Helical sequence motif	Function	$d_{C^{\alpha}\cdots O}$ (Å)	$d_{H\cdots O}$ (Å)	C^{α}–H···O ∠ (deg.)
1el5A1	3.50.50.60	GXXXG	Sarcosine oxidase	3.4	2.5	145
1pbe01	3.50.50.60	GXXXG	*p*-Hydroxybenzoate hydroxylase	3.4	2.4	160
1fohA1	3.50.50.60	GXXXA	Phenol hydroxylase	3.2	2.2	145
1an9A1	3.40.50.1140	GXXXA	D-Amino acid oxidase	3.3	2.2	124
1gpeA1	3.50.50.60	GXXXA	Glucose oxidase	3.4	2.5	148
1b37A1	3.50.50.60	GXXXA	Polyamine oxidase	3.5	2.5	141
1i8t	–	GXXXA	UDP-galactopyranose mutase	3.6	2.6	151
1trb02	3.50.50.60	GXXXA	Thioredoxin reductase	3.4	2.4	127
1chuA1	3.50.50.60	GXXXA	L-Aspartate oxidase	–	–	–
1lvl01	3.50.50.60	GXXXA	Dihydrolipoamide dehydrogenase	3.3	2.4	147
1hyu	–	GXXXA	Hydroperoxide reductase	3.6	2.6	145
1qjdA3	3.50.50.60	GXXXA	Flavocytochrome C3	3.2	2.2	155
3grs01	3.50.50.60	GXXXA	Glutathione reductase	3.4	2.5	158
1qlaA1	3.50.50.60	GXXXA	Fumarate reductase	3.4	2.4	140
1fcdA2	3.50.50.60	GXXXA	Flavocytochrome C sulfide dehydrogenase	3.3	2.4	124
2tmdA2	3.40.50.1140	GXXXA	Trimethylamine dehydrogenase	3.4	2.4	132
1cjc	3.40.50.1140	GXXXA	Adrenodoxin reductase	3.4	2.4	150
1f8rA2	3.50.50.60	GXXXA	L-Amino acid oxidase	3.4	2.4	154
1ybvA0	3.40.50.720	GXXXA	Trihydronaphthalene reductase	3.7	2.8	129
1fds00	3.40.50.720	GXXXA	17-Beta-hydroxysteroid dehydrogenase	3.7	2.7	118
1qrrA0	3.40.50.720	GXXXA	Sulfolipid biosynthesis protein	3.6	2.7	117
1e6w	3.40.50.720	GXXXA	3-Hydroxyacyl-coA dehydrogenase	–	–	–
1he2A0	3.40.50.720	GXXXL	Biliverdin IX beta reductase	3.3	2.4	130
1xel	3.40.50.720	GXXXC	UDP-galactose 4-epimerase	–	–	–
1cydA0	3.40.50.720	GXXXV	Carbonyl reductase	3.5	2.5	123
1nhp01	3.50.50.60	GXXXV	NADH peroxidase	3.3	2.4	134
1bxk	3.40.50.720	GXXXV	DTDP-glucose 4,6-dehydratase	3.6	2.6	124
1bl6A0	3.40.50.720	GXXXS	Alcohol dehydrogenase	–	–	–
1gnd04	3.50.50.60	EXXXS	GTPase activator	–	–	–
1d7y	–	SXXXV	Ferredoxin reductase	–	–	–
b18sA1	3.50.50.60	AXXXA	Cholesterol oxidase	–	–	–

Proteins with structural similarity (Dali *Z*-score greater than or equal to 4.0) to sarcosine oxidase and the amino acid sequences for structurally equivalent positions on α-helix 1. More than 75% of these proteins share a structurally conserved hydrogen bond between the first glycyl H1A atom of the GXXXG/A sequence motif and a carbonyl oxygen atom either five or six positions further back the polypeptide chain. Distances from both the H1A atom to the carbonyl oxygen atom and from the C^{α} atom to the carbonyl oxygen atom are given, as is the C^{α}–H···O bond angle. Conservation of this hydrogen bond suggests that it is important for stabilizing the FAD or NAD(P)-binding Rossmann fold. The list is sorted by motif type.

30 protein domains were identified that are structurally similar and bind to FAD or NAD(P) (Table 1). Of the 31 domains, 22 are those previously described as containing either the GXXXG or GXXXA motifs. However, an additional nine domains also adopt the Rossmann fold, yet α-helix 1 of these domains does not contain either the GXXXG or the GXXXA motif, although six of nine domains conserve the first glycyl residue of the motifs at structurally equivalent positions. Nevertheless the residue located at the position structurally equivalent to the first glycine of the GXXXG motif from sarcosine oxidase is also glycine in 28 of the 31 examples (Table 1). Of these 28 structures, 24 preserve the contact between the glycyl hydrogen atom of this residue and the carbonyl oxygen on the ligand-binding loop, suggesting the formation of a C^{α}–H···O hydrogen bond in these structures.

Although three of the 31 Rossmann folds do not conserve the glycyl residue at the first position of the sequence motif, the residues located at this position from two of these three folds still form either N–H···O or O–H···O contacts with the ligand-binding loop, consistent with hydrogen bond formation. For example, the GTPase activator (1gnd[21]) contains the sequence EXXXS on α-helix 1 at the position normally occupied by GXXXG/A. However, the hydrogen bond between α-helix 1 and the connecting loop is still conserved; the OE1 atom of Glu16 forms a hydrogen bond to the backbone amide of Gly13 (Figure 1(c)). The other example is ferredoxin reductase (1d7y[22]), which contains the sequence SXXXV on α-helix 1. Here the OG atom of Ser19 forms a hydrogen bond to the backbone carbonyl of Gly14 (Figure 1(d)).

In summary, of the 31 examples of Rossmann folds that bind FAD or NAD(P), 22 contain either the GXXXG or GXXXA motif on α-helix 1. In 20 of these 22 domains, the first glycyl residue of the GXXXG or GXXXA motif forms contacts with a glycyl residue located either five or six positions further back along the polypeptide chain, consistent with C^{α}–H···O hydrogen bond formation (Figure 2). In addition, structurally equivalent residues, located either seven or eight positions further back along the polypeptide chain on β-strand 1, form van der Waals interactions with the glycyl and alanyl residues of the GXXXG or GXXXA motifs. The identity of these residues on β-strand 1 is either valine or isoleucine in all but one of the 31 examples analyzed. Therefore, we

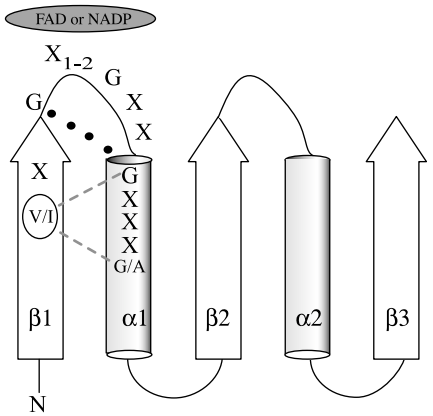

Figure 2. Secondary structure and sequence motifs of the FAD or NAD(P)-binding Rossmann fold. Cylinders represent α-helices and arrows are β-strands. The loop connecting β-strand 1 and α-helix 1 is referred to as the ligand-binding loop. The ligand is either FAD or NAD(P) and is depicted as a gray ellipse above the loop. This loop contains the previously discovered $GX_{1-2}GXXG$ motif, which is important for ligand binding. The newly discovered GXXXG/A motif is located on α-helix 1 and stabilizes the interaction between β-strand 1 and α-helix 1. The first glycyl residue of the motif contacts the carbonyl oxygen atom of a glycyl residue located five or six positions further back along the polypeptide chain; note that these glycine residues also correspond to the third and first glycine of the $GX_{1-2}GXXG$ motif, respectively. This contact shows the geometric hallmarks of a C^{α}–H···O hydrogen bond (black-dotted line). In addition, both residues of the GXXXG/A motif form van der Waals interactions (gray-broken lines) with a residue located on β-strand 1; this residue is either valine or isoleucine in all but one of the 31 examples of the fold. Combining both the $GX_{1-2}GXXG$ and GXXXG/A motifs results in a new motif with improved selectivity for Rossmann folds that bind FAD or NAD(P): $V/IXGX_{1-2}GXXGXXXG/A$.

infer a new sequence motif: $V/IXGX_{1-2}GXXGXXXG/A$, in which the first glycine of the GXXXG or GXXXA motifs is the third glycine of the previously known $GX_{1-2}GXXG$ motif.

Comparison of the predictive power of the $V/IXGX_{1-2}GXXGXXXG/A$ sequence motif for FAD or NAD(P)-binding Rossmann folds with previously known motifs

The $V/IXGX_{1-2}GXXGXXXG/A$ motif is more effective at predicting whether a given sequence adopts the Rossmann fold and binds to FAD or NAD(P) than previously determined sequence motifs. We calculated the mutual information value, where 0 indicates that we are no more certain about the fold type given the motif, and 0.18 is the maximum possible value, given the frequency of FAD or NAD(P)-binding Rossmann folds in the class, architecture, topology and homologous superfamily protein structure classification (CATH) (see Materials and Methods). The $GX_{1-2}GXXG$ motif[15] has a mutual information of 0.03. The $GX_{1-2}GXXGX_{1719}E/D$ motif, also previously described,[15] has an increased mutual information of 0.036, indicating that this motif is better than $GX_{1-2}GXXG$ for distinguishing the Rossmann fold over other folds. The motif described here, $V/IXGX_{1-2}GXXGXXXG/A$, has a mutual information of 0.043, corresponding to a 40% increase in information over the $GX_{1-2}GXXG$ motif. Therefore, comparing all three motifs, the $V/IXGX_{1-2}GXXGXXXG/A$ motif is best for predicting whether a protein sequence adopts the Rossmann fold and binds to FAD or NAD(P).

Table 2. The number of putative ORFs containing the $V/IXGX_{1-2}GXXGXXXG/A$ motif from selected genomes

Genome	ORFs with motif[a]	Total ORFs in genome	Fraction
Aeropyrum pernix	29	2694	0.011
Aquifex aeolicus	34	1522	0.022
Archaeoglobus fulgidus	52	2407	0.022
Bacillus subtilis	65	4100	0.016
Borrelia burgdorferi	9	849	0.011
Campylobacter jejuni	24	1634	0.015
Chlamydia trachomatis	12	894	0.013
Clostridium perfringens	44	2660	0.017
Deinococcus radiodurans chromosome1	58	2579	0.022
Deinococcus radiodurans chromosome2	9	357	0.025
Escherichia coli k12	69	4289	0.016
Haemophilus influenzae	26	1709	0.015
Helicobacter pylori 26695	15	1566	0.010
Methanococcus jannaschii	29	1715	0.017
Methanobacterium thermoautotrophicum	32	1869	0.017
Mycobacterium tuberculosis h37rv	99	3918	0.025
Mycoplasma genitalium	10	480	0.021
Neisseria meningitidis mc58	31	2025	0.015
P. aerophilum	51	2605	0.020
Rickettsia prowazekii	12	834	0.014
Salmonella typhimurium lt2	69	4451	0.016
Streptococcus pneumoniae r6	26	2035	0.013
Thermoplasma acidophilum	31	1478	0.021
Thermatoga maritima	29	1846	0.016
Xylella fastidiosa	30	2766	0.011
Yersinia pestis	55	3885	0.014

The total number of predicted ORFs for each genome is given, as is the fraction of those ORFs containing the $V/IXGX_{1-2}GXXGXXXG/A$ motif.

[a] See http://www.doe-mbi.ucla.edu/~kleiger/GenomeSeqs.html, which contains sequence and functional annotation for each ORF containing the $V/IXGX_{1-2}GXXGXXXG/A$ motif from the above 25 bacterial genomes.

Searching the ORFs from 25 fully sequenced genomes for the $V/IXGX_{1-2}GXXGXXXG/A$ motif

We searched the predicted open reading frames (ORFs) from 25 fully sequenced bacterial genomes for the $V/IXGX_{1-2}GXXGXXXG/A$ motif (Table 2).

Approximately 1–2% of the ORFs from each genome contain the V/IXGX$_{1-2}$GXXGXXXG/A motif. Those ORFs containing the motif are predicted to bind FAD or NAD(P) and function as oxidoreductases (see Discussion).

Discussion

By examining 31 representative structures of Rossmann folds that bind either FAD or NAD(P), we find a common hydrogen-bonded structural motif that appears to enhance the stability of these folds. In all, 28 of the 31 structures contain the previously known GX$_{1-2}$GXXG motif.[15] Some 24 of these 28 structures have atomic contacts between the third and the first glycyl residue of the motif consistent with the formation of C$^{\alpha}$–H···O hydrogen bonds. In addition, two of the three structures that do not contain the GX$_{1-2}$GXXG motif, but which contain residues at structurally equivalent positions, also appear to form a hydrogen bond, supporting the premise that the hydrogen bond is important for stabilizing FAD or NAD(P)-binding Rossmann folds. This hydrogen bond stabilizes the ligand-binding loop connecting α-helix 1 with β-strand 1. The backbone amides of this loop form hydrogen bonds with the pyrophosphate moiety of the nucleotide cofactor, explaining why stability of the loop is important.

If the hydrogen bond between α-helix 1 and the ligand-binding loop is important for stabilizing the Rossmann fold, why do five of the 31 proteins fail to make this contact? Some four of these five proteins conserve the glycyl residue at the last position of the GX$_{1-2}$GXXG sequence motif on α-helix 1, yet the atomic contacts between the glycyl C$^{\alpha}$ and carbonyl oxygen are slightly long for our definition of hydrogen bond formation. It is possible that uncertainty in the atomic coordinates of these examples may explain why at least some of these five proteins appear not to conserve the hydrogen bond.

Sequence motifs and oxidoreductase function

We used the V/IXGX$_{1-2}$GXXGXXXG/A sequence motif, as well as two previously described motifs (GX$_{1-2}$GXXG and GX$_{1-2}$GXXGX$_{17-19}$E/D), to distinguish sequences that adopt the Rossmann fold and bind to FAD or NAD(P) from those that do not. The V/IXGX$_{1-2}$GXXGXXXG/A motif performed the best of the three, with a mutual information value 40% greater than that for the GX$_{1-2}$GXXG motif. This increase in information is a consequence of fewer false positives. For example, the GX$_{1-2}$GXXG motif occurs in 816 out of 5928 possible sequences that do not adopt the Rossmann fold, whereas the V/IXGX$_{1-2}$GXXGXXXG/A motif occurs in only 29 of the same sequences. However, the increased selectivity comes at the expense of coverage. The GX$_{1-2}$GXXG motif occurs in 113 out of 161 possible sequences that do adopt the Rossmann fold, whereas the V/IXGX$_{1-2}$GXXGXXXG/A motif occurs in 62 of those sequences.

The V/IXGX$_{1-2}$GXXGXXXG/A motif is useful for predicting the biological function of proteins from fully sequenced genomes, including proteins of previously unknown function. For example, the V/IXGX$_{1-2}$GXXGXXXG/A motif occurs in 51 out of a total of 2605 ORFs from the hyperthermophilic bacterium *Pyrobaculum aerophilum*. The biological function of 16 of the 51 ORFs identified had been annotated only as "hypothetical protein". Because the proteins corresponding to these 16 ORFs contain the V/IXGX$_{1-2}$GXXGXXXG/A motif, and therefore are likely to bind either FAD or NAD(P) as a cofactor, the proteins can be annotated as oxidoreductases. The biological functions for proteins corresponding to ORFs from the other fully sequenced genomes can be assigned in a similar manner. While traditional annotation of the function of proteins relies on detecting sequence similarity to a protein of known function, our method is based only on the detection of the sequence motif and may complement current methods for the annotation of genomic sequence data. In fact, large databases of amino acid sequence motifs, such as the Prosite and Blocks databases,[23,24] already exist and could be used in a similar manner to annotate genomic protein sequences.

In conclusion, we find an expanded sequence motif, V/IXGX$_{1-2}$GXXGXXXG/A, associated with protein sequences that adopt the Rossmann fold and bind to FAD or NAD(P). This sequence motif is useful for annotating the function of proteins from the vast amount of genome sequence data with fewer false positives.

Materials and Methods

Databases

To account for the over-representation of certain protein families in the PDB, we used a non-redundant PDB set.†[25] In this set, only X-ray crystal structures of greater than 2.5 Å resolution are included. In addition, no two sequences in the non-redundant set share greater than 30% sequence identity to each other. The non-redundant PDB set (March, 2001) contains 1731 chains corresponding to 1611 proteins.

Counting the occurrences of amino acid sequence motifs

The method used to calculate the expected value for the GXXXA amino acid sequence motif has been described elsewhere.[1] Briefly, the expectation for any asymmetrical amino acid sequence motif (e.g. GXXXA) was calculated for each α-helix in our database. For a helix of known length and composition, we used ternary

† www.fccc.edu/research/-labs/dunbrack/culledpdb.html

bit strings to represent all possible sequences (e.g. 1 for glycine residues, 2 for alanine, 0 for non-glycine). Assuming that residues are uniformly distributed in the helix, the probability for the occurrence of a given number of motifs is computed by counting the occurrences of the motif in each enumerated sequence. The expected number of motifs is then given by:

$$\langle X \rangle = \sum_x xp(x) \quad (1)$$

where X is a random variable specifying the number of occurrences of the amino acid sequence motif in a given helix and $p(x)$ is the probability of observing x occurrences. The expectation of each α-helix was then summed over all helices to calculate the total expectation for a particular amino acid sequence motif for the entire database. The variance was calculated in a similar manner.

Helix–β-strand interactions

An α-helix and a β-strand were considered to interact if two or more contacts between the glycyl and/or alanyl residues of the GXXXG/A motif and at least one residue on the β-strand were found. Using the definition of Chothia *et al.*,[26] a contact is defined as two atoms within 0.6 Å of the sum of their van der Waals radii. Helix–β-strand interactions containing either the GXXXG motif or the GXXXA motif were identified using custom PERL software.

Calculating the coordinates for hydrogen atoms and structural analysis

The coordinates for hydrogen atoms were calculated using the CCP4 program HGEN, which uses standard geometry and a bond length of 1.0 Å. The structures of helix–β-strand interactions containing either the GXXXG or GXXXA motifs were analyzed and searched for atomic contacts of the C^{α}–H···O type using the program "O".

Calculating mutual information

Mutual information was calculated for each sequence motif using the standard equation:

$$I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (2)$$

where X is the binary random variable representing the fold type (i.e. Rossmann fold or other) and Y is the binary random variable representing whether a given sequence motif is present. Domains from the PDB were classified as FAD or NAD(P)-binding Rossmann folds using CATH (domains with either the 3.50.50.60, 3.40.50.720 or 3.40.50.1140 identifiers).[27] We found 161 domains that share less than 50% sequence identity with any other member of the set and adopt the Rossmann fold. Some 5928 entries are not FAD or NAD(P)-binding Rossmann folds which we call "other". The joint probabilities, $p(x, y)$, were estimated using the following frequencies: 113 of the 161 Rossmann folds contained the $GX_{1-2}GXXG$ sequence motif, and 816 of the 5931 other domains also contained the $GX_{1-2}GXXG$ sequence motif. Some 86 of the 161 Rossmann folds contained the $GX_{1-2}GXXGX_{17-19}E/D$ motif, and 259 of the 5928 other domains also contained the $GX_{1-2}GXXGX_{17-19}E/D$ motif. Finally, 62 of the 161 Rossmann folds contained the $V/IXGX_{1-2}GXXGXXXG/A$ motif, and 29 of the 5928 other domains also contained the $V/IXGX_{1-2}GXXGXXXG/A$ motif.

Genomic analysis

The putative ORFs from 25 fully sequenced bacterial genomes were downloaded from the entrez genome search and retrieval system of the Nation Center for Biotechnology Information (NCBI)† and searched for the $V/IXGX_{1-2}GXXGXXXG/A$ motif using custom PERL software. This routine is available for the analysis of individual sequences‡.

Acknowledgements

We thank Magdalena Ivanova and Tom Graeber for helpful discussion. We thank Parag Mallick for computational and database assistance, and Rob Grothe for statistical analysis. We thank DOE, NIH, and HHMI for support.

References

1. Senes, A., Gerstein, M. & Engelman, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GXXXG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.* **296**, 921–936.
2. Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. & Engelman, D. M. (1992). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, **31**, 12719–12725.
3. Whittington, D. A., Waheed, A., Ulmasov, B., Shah, G. N., Grubb, J. H., Sly, W. S. & Christianson, D. W. (2001). Crystal structure of the dimeric extracellular domain of human carbonic anhydrase XII, a bitopic membrane protein overexpressed in certain cancer tumor cells. *Proc. Natl Acad. Sci. USA*, **98**, 9545–9550.
4. Mendrola, J. M., Berger, M. B., King, M. C. & Lemmon, M. A. (2002). The single transmembrane domains of ErbB receptors self-associate in cell membranes. *J. Biol. Chem.* **277**, 4704–4712.
5. Kleiger, G., Grothe, R., Mallick, P. & Eisenberg, D. (2002). GXXXG and AXXXA: common alpha-helical interaction motifs in proteins, particularly in extremophiles. *Biochemistry*, **41**, 5990–5997.
6. MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science*, **276**, 131–133.
7. Senes, A., Ubarretxena-Belandia, I. & Engelman, D. M. (2001). The Calpha–H···O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc. Natl Acad. Sci. USA*, **98**, 9056–9061.
8. Scheiner, S., Kar, T. & Gu, Y. (2001). Strength of the Calpha–H···O hydrogen bond of amino acid residues. *J. Biol. Chem.* **276**, 9832–9837.

† www.ncbi.nlm.nih.gov
‡ http://www.doe-mbi.ucla.edu/cgi/kleiger/Rossmann.cgi

9. MacKenzie, K. R. & Engelman, D. M. (1998). Structure-based prediction of the stability of transmembrane helix–helix interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl Acad. Sci. USA*, **95**, 3583–3590.
10. Rossmann, M. G. & Argos, P. (1976). Exploring structural homology of proteins. *J. Mol. Biol.* **105**, 75–95.
11. Rao, S. T. & Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **76**, 241–256.
12. Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advan. Protein Chem.* **34**, 167–339.
13. Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of nucleotide-binding protein. *Nature*, **250**, 194–199.
14. Dym, O. & Eisenberg, D. (2001). Sequence-structure analysis of FAD-containing proteins. *Protein Sci.* **10**, 1712–1728.
15. Wierenga, R. K., Terpstra, P. & Hol, W. G. (1986). Prediction of the occurrence of the ADP-binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint. *J. Mol. Biol.* **187**, 101–107.
16. Wagner, M. A., Trickey, P., Chen, Z. W., Mathews, F. S. & Jorns, M. S. (2000). Monomeric sarcosine oxidase: 1. Flavin reactivity and active site binding determinants. *Biochemistry*, **39**, 8813–8824.
17. Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). The occurrence of C–H···O hydrogen bonds in proteins. *J. Mol. Biol.* **252**, 248–262.
18. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
19. Enroth, C., Neujahr, H., Schneider, G. & Lindqvist, Y. (1998). The crystal structure of phenol hydroxylase in complex with FAD and phenol provides evidence for a concerted conformational change in the enzyme and its cofactor during catalysis. *Structure*, **6**, 605–617.
20. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–603.
21. Schalk, I., Zeng, K., Wu, S. K., Stura, E. A., Matteson, J., Huang, M. *et al.* (1996). Structure and mutational analysis of Rab GDP-dissociation inhibitor. *Nature*, **381**, 42–48.
22. Senda, T., Yamada, T., Sakurai, N., Kubota, M., Nishizaki, T., Masai, E. *et al.* (2000). Crystal structure of NADH-dependent ferredoxin reductase component in biphenyl dioxygenase. *J. Mol. Biol.* **304**, 397–410.
23. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucl. Acids Res.* **30**, 235–238.
24. Henikoff, S., Henikoff, J. G. & Pietrokovski, S. (1999). Blocksredundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
25. Hobohm, U., Scharf, M. & Schneider, R. (1993). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.
26. Chothia, C., Levitt, M. & Richardson, D. (1981). Helix to helix packing in proteins. *J. Mol. Biol.* **25**, 215–250.
27. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Edited by P. Wright

(Received 29 May 2002; accepted 9 August 2002)