Adaptations of the Helix-Grip Fold for Ligand Binding and Catalysis in the START Domain Superfamily

Lakshminarayan M. Iyer, Eugene V. Koonin, and L. Aravind*
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

ABSTRACT **With a protein structure comparison, an iterative database search with sequence profiles, and a multiple-alignment analysis, we show that two domains with the helix-grip fold, the star-related lipid-transfer (START) domain of the MLN64 protein and the birch allergen, are homologous. They define a large, previously underappreciated superfamily that we call the START superfamily. In addition to the classical START domains that are primarily involved in eukaryotic signaling mediated by lipid binding and the birch antigen family that consists of plant proteins implicated in stress/pathogen response, the START superfamily includes bacterial polyketide cyclases/aromatases (e.g., TcmN and WhiE VI) and two families of previously uncharacterized proteins. The identification of this domain provides a structural prediction of an important class of enzymes involved in polyketide antibiotic synthesis and allows the prediction of their active site. It is predicted that all START domains contain a similar ligand-binding pocket. Modifications of this pocket determine the ligand-binding specificity and may also be the basis for at least two distinct enzymatic activities, those of a cyclase/aromatase and an RNase. Thus, the START domain superfamily is a rare case of the adaptation of a protein fold with a conserved ligand-binding mode for both a broad variety of catalytic activities and noncatalytic regulatory functions. Proteins 2001;43:134–144. Published 2001 Wiley-Liss, Inc.[†]**

Key words: START; aromatase; cyclase; lipid binding; birch allergen

INTRODUCTION

The interaction of proteins with small molecules is the basis of enzymatic activity and its regulation. The recognition of small molecules by proteins involves a variety of globular domains that catalyze specific reactions on them, use them as cofactors in other reactions, and bind them and undergo a conformational change in the process. Comparative analyses of protein sequences and structures have resulted in the elucidation of diverse modes of regulatory interactions between distinct protein domains and small molecules.[1–7] Several domains such as Per–Arnt–Sim domain (PAS),[4] cGMP Phosphodiesterase–adenylyl cyclase–FhlA domain (GAF),[5] Aspartokinase–chorismate mutase–TyrR domain (ACT),[3] ATP-cone,[8] Pleckstrin Homology domain (PH),[9] and Sec14p[10] are specialized small-molecule-binding domains that regulate metabolic or signal transduction pathways based on interactions with specific ligands. Alternatively, some of these domains use the bound ligands as sensors for stimuli such as redox potential and light. Most of these domains have diversified in evolution to accommodate different ligands but retain some shared structural features, such as various amino acids for the ACT domain.[3] On a very small number of occasions, variants of the same domain, such as the double-stranded β-helix[11] and USPA[12] domains, are used both for catalysis and for ligand binding. These cases are of particular interest for understanding the general principles of protein/small-molecule interactions and the evolutionary diversification of protein families for performing distinct functions. Here, we describe the star-related lipid-transfer (START) domain superfamily, which represents a case of exaptation of the same fold for the simple binding of different molecules and catalytic activity.

The START domain was initially identified as a widespread lipid-binding domain present in multicellular eukaryotes (plants and animals) with a regulatory role in signal transduction,[13] analogous to the PH[9] and Sec14p[10] domains. Representatives of the START domain family have been shown to bind different ligands such as sterols (steroidogenic acute phase response protein) and phosphatidylcholine (PPCT).[14,15] Ligand binding by the START domain also can regulate the activities of other domains that co-occur with the START domain in multidomain proteins such as Rho-gap, the homeodomain, and the thioesterase domain.[13] The subsequent solution of the crystal structure of the START domain from the MLN64 protein[16] revealed that this domain adopts an α/β structure of the helix-grip [TATA binding protein (TBP)-like] fold, as classified in the Structural Classification of Proteins (SCOP) database.[17] The examination of the START domain structure has also suggested the presence of a binding pocket that has been implicated in the accommodation of the lipid ligand. A similarity between the structures

A list of the Genbank identifier numbers of all the identified START domains, including the 59 plant proteins, is available at ftp://ncbi.nlm.nih.gov/pub/aravind/.

*Correspondence to: L. Aravind, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894. E-mail: aravind@ncbi.nlm.nih.gov

Received 26 September 2000; Accepted 14 December 2000

of the START domain and the birch allergen (BA)[18] has been noticed, but there is no certainty regarding the homology of the structures, and the functional implications of this similarity remain unclear.[16]

Here, by using sequence profile searches and structure comparison and prediction, we show that the START domain and BA belong to a large, hitherto incompletely described superfamily of ligand-binding proteins predicted to possess the same fold and similar ligand-binding features. The START domain superfamily includes bacterial, archaeal, and eukaryotic proteins with diverse predicted roles in metabolism, RNA degradation, and signal transduction. Many of these proteins appear to bind ligands without catalyzing a chemical reaction, but we show that polyketide cyclases (aromatases) also belong to this superfamily, and we propose the ligand-binding site and possible catalytic mechanism of these enzymes. Additionally, we identify numerous hitherto unnoticed cyclases encoded by the genomes of *Actinomycetes* that could be involved in the synthesis of polyketide antibiotics and toxins.

MATERIALS AND METHODS

The initial characterization of the phyletic distribution of each protein, the delineation of likely orthologous relationships, and the detection of other obvious homologs were carried out through a search of the nonredundant protein sequence database at the National Center for Biotechnology Information (National Institutes of Health, Bethesda, MD) with the gapped BLAST program.[19] An in-depth analysis of protein sequence similarities was performed with iterative PSI-BLAST searches with a profile inclusion cutoff of E (expectation) = 0.01, with several different sequences extracted from the results of the first-pass search employed as seeds.[19] The significance of the matches was assessed in terms of the E value obtained on the first detection of the given sequence over the 0.01 threshold in the course of iterative searches. In addition, to eliminate false positives that might have emerged because of the compositional bias of a particular query, we determined if, in the iterative searches, different queries from a domain family consistently retrieved from the database approximately the same set of proteins. Multiple alignments of protein sequences were constructed by the parsing of pairwise alignments generated by PSI-BLAST and their realignment with the CLUSTALX program[20] followed by manual refinement. The final data on the phyletic distribution of the START superfamily were obtained through the construction of a profile from the multiple alignment of representatives of all families in the START superfamily[21] and a search of the protein sets from individual complete genomes and those from the nearly complete genomes of *Arabidopsis thaliana* and *Schizosaccharomyces pombe* (obtained from the Genome division of the Entrez system).

Protein secondary structure prediction was performed with the PHD[22] and PSIPRED[23] programs. Manipulations with protein structures and visualization were carried out with the Swiss PDB viewer (version 3.51) program,[24] and the ribbon diagrams were rendered with the Molscript program.[25] The structural classification of proteins was based on the SCOP database;[17] tertiary structure alignments and comparisons were carried out with the DALI program and FSSP database[26] and the VAST program.[27]

RESULTS AND DISCUSSION

Helix-Grip Fold

The availability of the MLN64 START domain structure provided a means of exploring the deep evolutionary relationships of this superfamily. As previously noted, structural comparisons showed the BA, C-terminal domain of phosphoglucomutase (PGM), and TBP to be the closest neighbors of the START domain.[16] Using the VAST and DALI programs, we further performed transitive structural comparisons with each of the neighbors of the START domain that were initially detected. This helped us to identify seven distinct structures that shared this common fold, namely, the TBP, adaptin appendage module C-terminal domain, DNA glycosylase II N-terminal domain, C-terminal domain of PGM, BA, naphthalene dioxygenase, and START domain (Fig. 1). The SCOP database[17] classifies most of these structures in the TBP-like fold; here, we use the term *helix-grip fold*, which was proposed in the description of the naphthalene dioxygenase structure.[28] The helix-grip fold consists of a central β-sheet with two helical units near the N- and C-termini, with the C-terminal helix packing tightly against the sheet (Fig. 1). The members of this fold can be further subdivided into specific categories on the basis of diagnostic structural features (Fig. 1). The version of this fold present in TBP, adaptin appendage C domain and DNA glycosylase II N-terminal domain, contains a central core of five strands and an N-terminal helical region that consists of a single helix. The category defined by the C-terminal domain of PGM is similar but contains six strands in the core. The naphthalene dioxygenase contains a seven-stranded core with several helical inserts and a significantly distorted C-terminal helix. BA and the classic START domain (CSD) have the same topology, with a seven-stranded core sheet and two helices in the N-terminal helical segment that are positioned at very similar angles (Fig. 1). Furthermore, a tertiary structure-based alignment of the sequences of the two domains shows noticeable primary structure patterns corresponding to the similarity in the structural elements.

Furthermore, BA and CSD domains share a deep pocket formed by the interface between the core sheet and the C-terminal helix that differentiates them from the five-stranded TBP-like forms of the helix-grip fold that have a flat interface without a pocket (Figs. 2 and 3). This is consistent with the sideways binding of DNA by the core sheet in TBP,[29] in contrast to the predicted docking of cholesterol in the pocket of the START domain.[16] The other seven-stranded version of the helix-grip fold, the naphthalene dioxygenase α-subunit, binds its substrate in a pocket similar to that present in CSD and BA,[28] and they have been classified in the SCOP database in the same superfamily. However, CSD and BA are structurally distinct from the naphthalene dioxygenase family in that they lack the α-helical elaborations that are typical of the

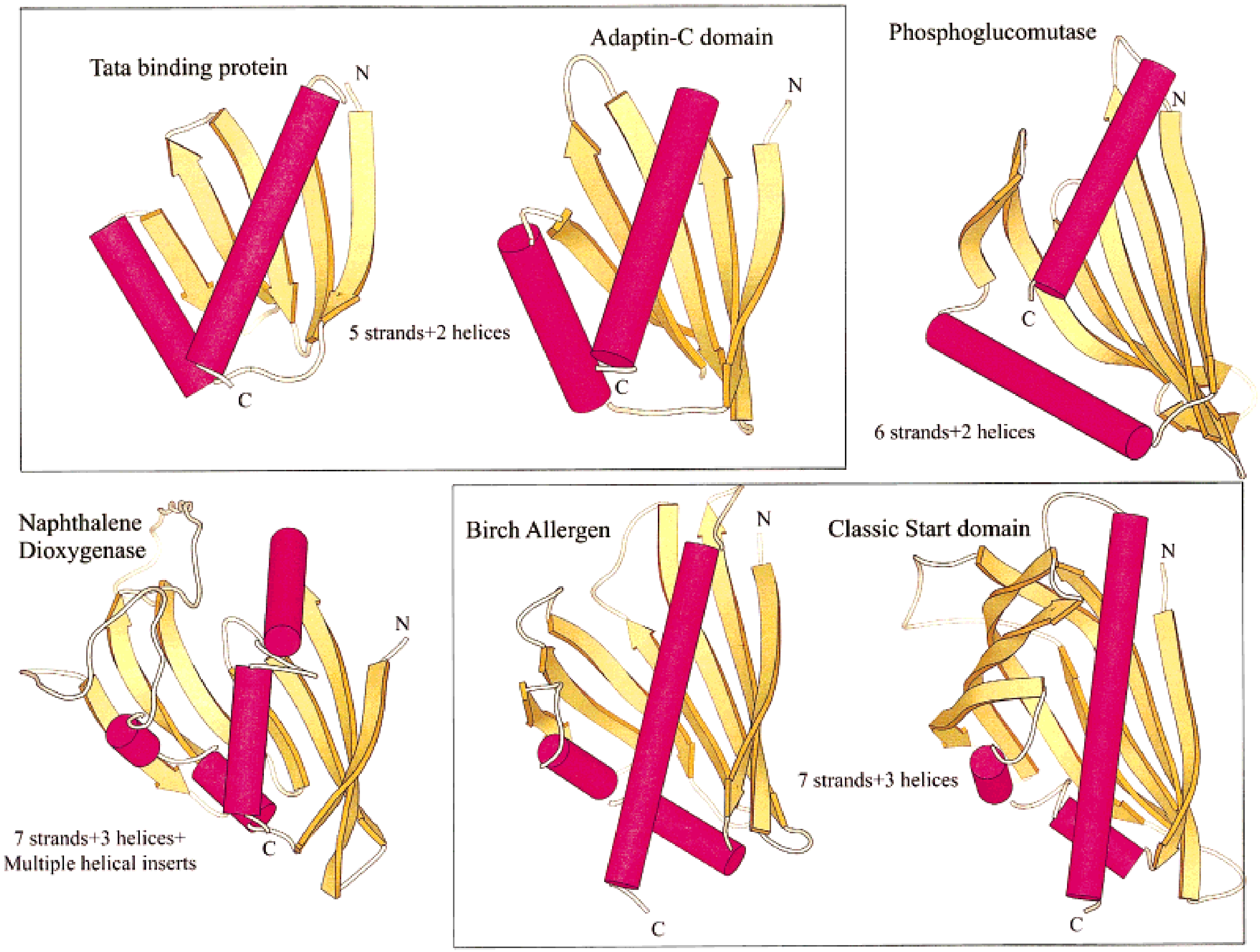

Fig. 1. Comparison of representative structures of the helix-grip fold. Note the common features shared by the BA and the MLN64 START domain to the exclusion of the other members of this family. The multiple helical inserts in the naphthalene dioxygenase α-subunit structure have been smoothed to highlight the core helix-grip fold.

latter (Fig. 1), including the specific conserved histidines involved in iron binding.[28] Naphthalene dioxygenases do not possess counterparts to the prominent sequence patterns that are shared by CSD and BA, as determined by an examination of structure-based alignments of all these proteins. Hence, we believe that CSD and BA should be considered representatives of a distinct protein superfamily within the helix-grip fold.

Sequence–Structure Analysis of the START Domain Superfamily

Given the ability of sequence profile searches to detect subtle relationships that originally have been thought to be identifiable only by structural comparisons,[30] we sought to investigate the relationship between the CSD and BA and identify potential additional homologs of these proteins with multiple queries to initiate searches as described previously.[3] A PSI-BLAST search (profile inclusion threshold = 0.01) using the sequence of the helix-grip domain of BA (PDB, 1BV1) as the query retrieved from the database previously identified homologs of BA, such as plant pathogenesis and stress-induced proteins (e.g., PR10), cytokinin-binding proteins and latex proteins,[31] and, in subsequent iterations, polyketide cyclases from the *Actinomycetes* and several uncharacterized bacterial proteins. Iterative searches begun with the polyketide cyclase sequences as queries recovered not only the plant BA-related proteins but also previously identified CSD-containing proteins, such as the steroidogenic acute response protein. These searches additionally identified the START domain in a group of proteins that is conserved in all eukaryotes and is typified by the SPCC16A11.07 protein from *S. pombe*. A PSI-BLAST search seeded with this protein sequence retrieved after 10 iterations the CSD-containing proteins ($e = 10^{-3}$ to 10^{-4}) and, within 15 iterations, the plant proteins of the BA family ($e = 10^{-2}$ to 10^{-3}). In contrast, none of the other members of the helix-grip fold (according to SCOP and as defined previously), including the naphthalene dioxygenases, could be detected in these

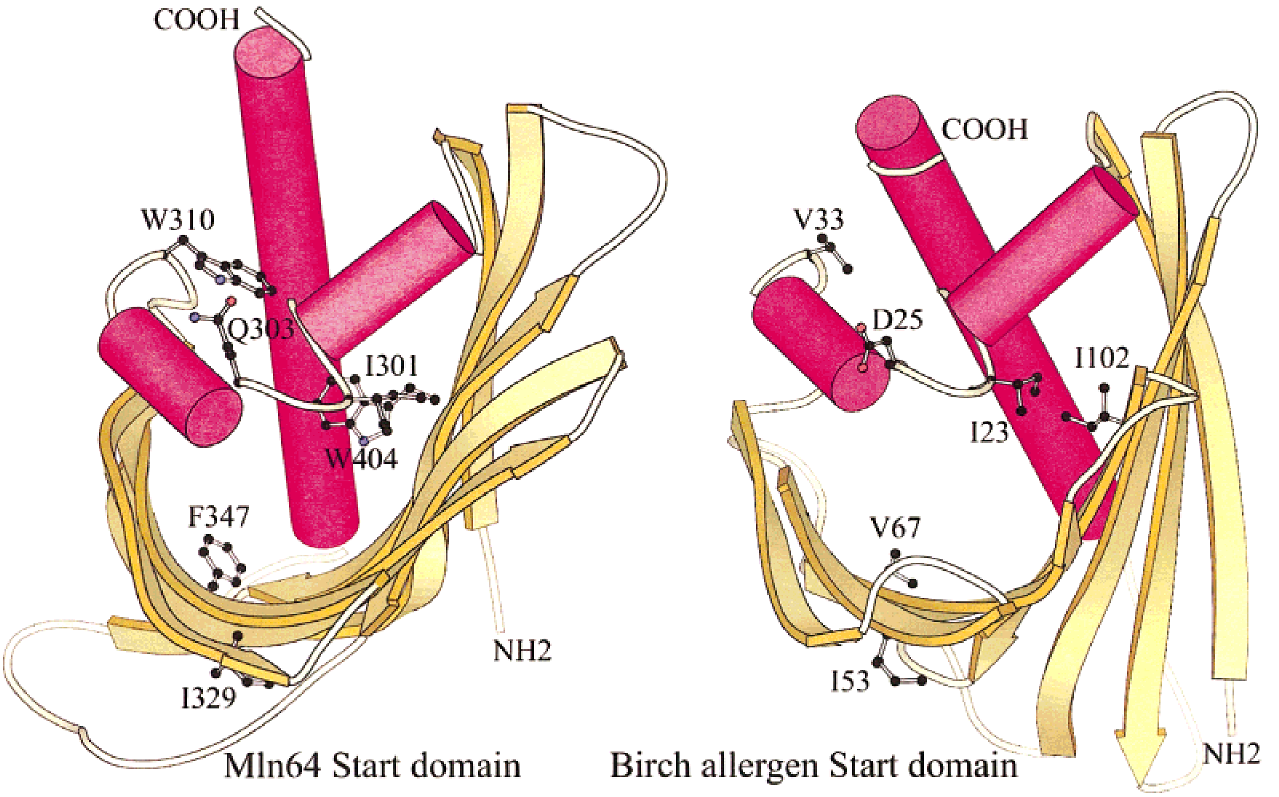

Fig. 2. Comparison of the two START domain structures to highlight the aperture of the lipid binding pocket and the differences in its width and accessibility. Some conserved residues characteristic of the superfamily are shown. Note the conserved polar residue on the upper rim of the pocket aperture prior to helix-2 (Q303 and D25 in MLN64 and the birch antigen, respectively).

exhaustive searches. Thus, the CSD and BA families, the polyketide cyclases, and a large number of uncharacterized proteins form a distinct single-sequence superfamily (hereafter, the START domain superfamily) with the helix-grip fold.

To investigate the structural features and conservation pattern of the START domain superfamily, we constructed a multiple alignment that was refined with the structural alignment of the Mln64 START and BA domains as a guide (Fig. 4). The secondary structures predicted for subsets of sequences that had no close affinity with any of the known structures were fully consistent with the seven-stranded, three-helical helix-grip fold of the START family (Fig. 4). The conservation pattern of this family was chiefly centered around the hydrophobic residues in different strands, with the maximum conservation associated with the distal strand 7. Strands 3 and 4 are particularly variable in size and sequence (Fig. 4). Helices 1 and 2 show notable conservation, whereas the strongly predicted C-terminal helix 3 is typically conserved only within individual families (Fig. 4). A notable feature of the START domain is the conservation of a polar residue, most often aspartate, in the loop immediately upstream of helix 2. As indicated previously, the N-terminal helices contribute to the formation of the (predicted) substrate-binding pocket in both structurally characterized proteins of the START superfamily, BA and MLN64, and the conserved polar residue is associated with the rim of this pocket (Figs. 2 and 4).

A comparison of the pocket in the BA and CSD domains reveals differences in the aperture size and accessibility. The CSD domain has a closed pocket with a narrow aperture in contrast to the BA domain, which has a broader aperture with a tunnel penetrating through the entire structure (Fig. 3). Mapping the sequence alignments onto the structure shows that these differences in pocket size are largely attributable to the differences in the lengths of strands 2, 3, and 4 (Figs. 1 and 4). In particular, in the CSDs the region between the cores of strands 3 and 4 is longer than in the BAs and curves strongly to form a closed pocket (Fig. 1). Thus, the length of this variable region can be used to predict the tunnel width in the START domain. Most of the conserved residues map to the core of the strands that form the surface of the tunnel or the pocket, which suggests a similar function for these structures in different members of the START superfamily (Figs. 2 and 4). The inner surface of the pocket in all members of the superfamily contains conserved hydrophobic residues and polar residues (Figs. 2 and 4), indicating that this domain could accommodate both hydrophobic and polar surfaces of molecules such as sterols and PPCT.[16]

The sequence and structure conservation between the CSD and BA families and, in particular, the conservation

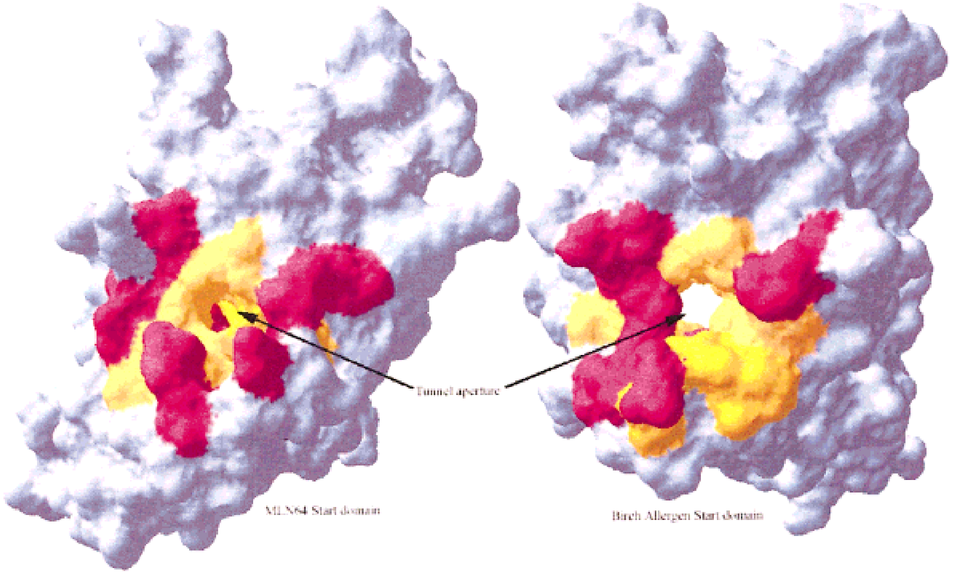

Fig. 3. Surface view of the predicted ligand-binding tunnel and its aperture in the BA and CSDs. Note the differences in the tunnel width and accessibility. The predicted functionally important residues associated with the tunnel are shown; hydrophobic residues are yellow, and polar residues are red.

of the predicted binding pocket suggest that like CSDs, the BA-family proteins also bind specific ligands. This is consistent with the high-affinity binding of cytokinins and their analogs by a close relative of the BA protein, the CSBP protein from Mung bean.[32] Some of the proteins of the BA family, including BA itself and its homolog from ginseng, have been reported to possess ribonuclease activity.[33,34] This activity could be accounted for by RNA binding in the wider pocket of the BA-type START domains. The conserved helix-capping polar residue preceding helix-2 (typically an acidic residue in this family) and another polar residue (usually aspartate or asparagine) in the loop between strands 2 and 3, which is unique to the BA family, are located on the rim of the tunnel (Fig. 4). These two residues could potentially direct water for an attack on the polynucleotide. Furthermore, cytokinins and RNA share a common determinant in the form of a purine, which could be the principal feature recognized by the BA-type START domains. The functions of the plant BA-family proteins in cytokinin binding and RNA degradation are compatible with their reported role in pathogen and stress response.[31,35]

Goodpasture antigen-binding protein, a CSD protein, contains a PH domain and a START domain and has been shown to phosphorylate the goodpasture antigen involved in a human autoimmune disease.[36] Further experimental evidence is required to confidently implicate the START domain in this activity, but if it is confirmed, it may suggest additional catalytic capabilities of the START domain.

In addition to the CSD and BA families, another large family of proteins containing the START domain includes cyclases or aromatases involved in the biosynthesis of polyketide antibiotics in *Actinomycetes* and their homologs from other bacteria, creanarchaea, and eukaryotes. The prototype cyclase/aromatase, TcmN, participates in the biosynthesis of tetracenomycin in *Streptomyces glaucescens*, and its homolog WhiE-VI is involved in the synthesis of the polyketide spore pigment in *S. coelicolor*.[37,38] The

Fig. 4. Multiple alignment of the START domain superfamily. The coloring reflects the amino acid conservation profile above an 80% consensus at a given position in the alignment. The consensus profile was calculated with the following classes of residues: hydrophobic (h: LIYFMWACV), shaded yellow; charged (c: KERDH); polar (p: STEDRKHNQ), colored purple; small (s: SAGDNPVT), colored green; and big (b: LIFMWYERKQ). The conserved polar residues that may have a role in catalysis or substrate binding in the cyclases/aromatases and the BA family RNases are shown with blue shading. The consensus secondary structure shown above the alignment represents the shared core of the known (PDB codes: 1bv1, BA; 1em2, CSD) and predicted structures. Additionally, the secondary structures that are derived from the two known structures and the structure predicted from the alignment that excludes the representatives with known structures are shown below it. The proteins with known crystal structures are boxed. The five families comprising the START superfamily are indicated on the right side of the alignment: (1) the CSD family, (2) the BA family, (3) the CAS family, (4) *Prochlorococcus*-plant specific START domains, and (5) Ydr214w-like START domains. Ao = *Asparagus officinalis*; Ap = *Aeropyrum pernix*; At = *Arabidopsis thaliana*; Bp = *Betula pendula*; Bs = *Bacillus subtilis*; Bt = *Bos taurus*; Ce = *Caenorrhabditis elegans*; Dd = *Dictyostelium discoideum*; Dm = *Drosophila melanogaster*; Dr = *Deinococcus radiodurans*; Ec = *Escherichia coli*; Gm = *Glycine max*; Ha = *Helianthus annuus*; Hs = *Homo sapiens*; Ll = *Lilium longiflorum*; Me, *Methylobacterium extorquens*; Ml = *Mycobacterium leprae*; Mt = *Mycobacterium tuberculosis*; Nm = *Neisseria meningitides*; Nt = *Nicotiana tabacum*; Oc = *Oligotropha carboxidovorans*; Os = *Oryza sativa*; Pa = *Pseudomonas aeruginosa*; Pd = *Paracoccus denitrificans*; Pgi = *Panax ginseng*; Pm = *Prochlorococcus marinus*; PPS = *Papaver somniferum*; Ps = *Pseudomonas syringae*; Psem = *Pseudotsuga menziesii*; Rg = *Rubrivivax gelatinosus*; Rp = *Rickettsia prowazekii*; Sc = *Saccharomyces cerevisiae*; Sp = *Schizosaccharomyces pombe*; Sso, *Sulfolobus solfataricus*; STRCO = *Streptomyces coelicolor*, STRGA = STRGRI, STRCM, STRVN, and STRARG = *Streptomyces sp.*; Syn = *Synechocystis sp.* (strain PCC 6803); Vr = *Vigna radiata*; Vv = *Vitis vinifera*; Xf = Xylella *fastidiosa*; Zm = *Zymomonas mobilis*.

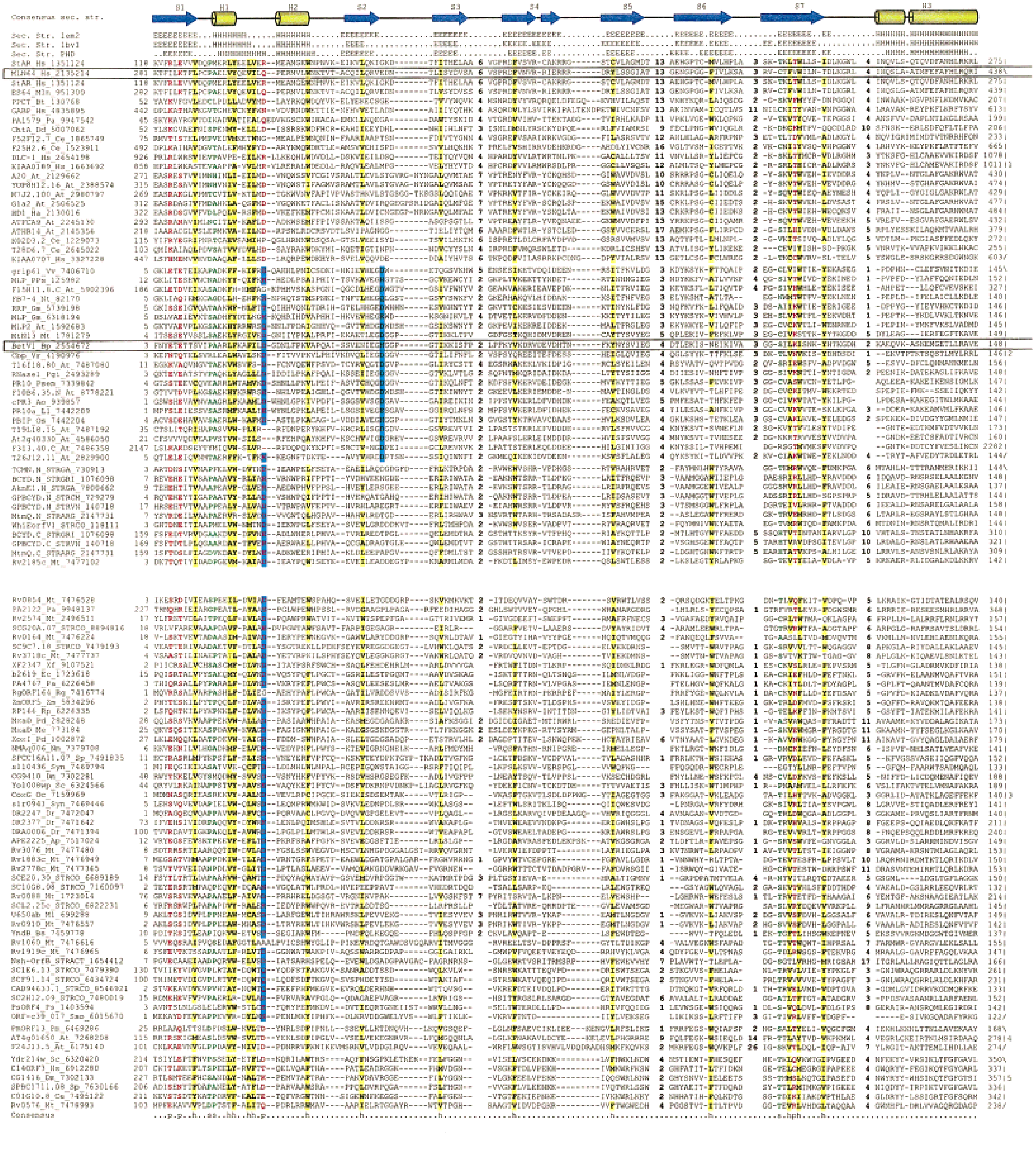

Figure 4.

linear polyketide substrates for these enzymes are synthesized via the condensation of malonyl-CoA and acetyl-CoA units. The cyclases/aromatases then catalyze the cyclization of the linear ketides with the elimination of water and the formation of aromatic rings in the cyclized ketides.[38,39] Because the START domain occupies most of the protein's length in the cyclases/aromatases and is their only globular domain, it should form the active site of these enzymes. The START domains of cyclases/aromatases are predicted to contain a wide ligand-binding pocket similar to that in the BA family, as indicated by the short spacer between strands 3 and 4 (Figs. 2–4). Thus, this predicted ligand-binding tunnel is sufficient in its dimensions to accommodate the polyketide substrate and is likely to sterically

TABLE I. Distribution of START Domain Proteins in Completely Sequenced Genomes[†]

Species	START proteins	Gene name[a]
Bacteria		
Bacillus subtilis	1	YndB
Deinococcus radiodurans	4	DRA0006, DR2377, DR2247, DR0721
Synechocystis PCC6803	2	slr1634, sll0436
Mycobacterium tuberculosis	18	Rv0088, Rv0164, Rv0369c, Rv0576, Rv0854, Rv0856, Rv0857, Rv0910, Rv3076, Rv1060, Rv1546, Rv1883c, Rv1919c, Rv2185c, Rv2186c, Rv2574, Rv2778c, Rv3718c
Streptomyces coelicolor	13	WhiE orf VI, ActVII, SC6G10.02c, 2SCG18.30c, SC9C7.18, SC7C7.14, SCG20A.07, SCE20.30, SCL2.25c, SCH10.30c, SC10G8.08, SC1E6.13, SCH35.36c
Other *Streptomycetes*	6	TCMN, AknE1, MtmQ, gis: 1076098, 729279, 140718
Pseudomonas aeruginosa	3	PA1579, PA2122, PA4767
Rickettsia prowazekii	1	RP166
Neisseria meningitidis MC58	1	NMB0797
Neisseria meningitidis Z2491	1	NMA1006
Escherichia coli	1	B2619/YfjG
Vibrio cholerae	1	YfjG
Xylella fastidiosa	1	XF2347
Archaea (Crenarchaeota)		
Aeropyrum pernix	1	APE2225
Sulfolobus	1	ORF-c39_017
Eukaryotes		
Caenorrhabditis elegans	10	C01G10.8, C06H2.2, F25H2.6, F26F4.4, F45H7.2, F45H7.3, F52F12.7, K02D3.2, R144.3, T28D6.7
Drosophila melanogaster	7	CG1416, CG3522, CG6310, CG6565, CG8480, CG9410, GH07688
Homo sapiens	15	C14ORF3, StAR, GABP, DLC-1, KIAA0189, KIAA0707, MLN64, KIAA1300, PPCT, CGI-52, GTT1, DKFZp434G241; gi: 4902678 + 2 from unannotated human genome sequence
Saccharomyces cerevisiae	2	Ydr214wp, Yol008wp
Arabidopsis thaliana	59	For a complete list, see the supplementary material.
Plasmodium	1	PfC0360w

[†] Only those genomes that encode START domains are included.

[a] The gene name is the systematic name assigned by the respective sequencing projects; those proteins that lack a systematic gene name are indicated with Genbank identifiers (GI numbers).

[b] CAS = cyclase/aromatase-type START domain.

[c] Hypothetical operons are defined as clusters of genes separated by less than 150 base pairs.

[d] YS = Ydr214wp-like START domain.

[e] CSD = classic START domain.

[f] PS = *Prochlorococcus marinus* Orf13-like START domain.

[g] BA = birch allergen-like START domain.

Family classification and other comments	Fusion of START with other domains
CAS[b]	
All are representatives of the CAS family. DR2247 is in an operon[c] with an α/β hydrolase. DRA0006 is in an operon with a dehydrogenase.	
CAS	
Rv0576 is of the YS[d] family. All the rest belong to the CAS family. Rv0854, Rv0856, Rv0857-single operon. Rv0164-operon with acyl carrier protein. Rv0576, Rv0910-operons with a glyoxylase I/dioxygenase family protein. Rv1060-operon with a fatty acid synthetase. Rv1883c-operon with a sterol/straight chain oxidoreductase. Rv1919c-operon with a fatty acid CoA ligase. Rv2574-operon with a metalloprotease and ApbA (implicated in thiamine biosynthesis). Rv2778c-operon with a ferredoxin reductase. Rv0088-operon with a methyltransferase. Rv3718c-operon with dimunito-like oxidoreductase and cyclopropane-fatty-acyl-phospholipid synthase.	Rv0576: HTH + S + X; S: Start domain; X: A hydrolase-like domain found in Actinomycetes with a conserved cysteine, histidine and asparate reminiscent of papain-like proteases; Rv0369c: S + transmembrane
All proteins belong to the CAS family.	TCMN: 2*S + Methyltransferase; MtmQ: 2*S + small subunit of aromatic oxygenase; AknE1: 2*S; 1076098: 2*S; 729279: 2*S; 140718: 2*S; ActVII: 2*S
PA1579 belongs to the CSD[e] family; the rest belong to the CAS family. PA4767 is the ortholog of YfjG and forms an operon with the ortholog of YfjF.	PA2122: 4* Transmembrane + S
CAS, ortholog of YfjG	
Ortholog of YfjG, forms an operon with the ortholog of YfjF	
Same as above	
Same as above	
Same as above	
Same as above	
CAS	
CAS	
C01G10.8: YS family, R144.3: CAS family. The rest are in the CSD family.	C01G10.8: HCH1 + S; R144.3: X + C2H2 + C2H2 + S (possible artificial fusion); F25H2.6: PH + S; F45H7.2: Rhogap + S; F45H7.3: Rhogap + S; F26F4.4: Transmembrane + S; HCH1: uncharacterized, conserved globular domain
CG1416: YS family, CG9410 and CG6310: CAS family. The rest are in the CSD family.	CG1416: HCH1 + S; GH07688: PH + S; CG8480: Rhogap + S; CG3522: 2* S
C14ORF3: YS family. The others are in the CSD family.Of the the two sequences obtained from the unannotated human genome sequence, one belongs to the CAS family.	C14ORF3: HCH1 + S; GABP: PH + S; DLC-1: Sam + Rhogap + S; KIAA0189: Rhogap + S; KIAA0707: 2*Thioesterase + S; RhoGAP: Rhogap + S
Yol008wp: CAS family, Ydr214wp: YS family	Ydr214wp: HCH1 + S
T21B14.13: YS family, AT4g17650: CAS family. Seventeen are of the PS[f] family, 20 are in the BA[g] family, and 20 are in the CSD family, 16 of which are fused to a homeodomain. T28J14_200, AT4g26920 and AT4g14500 are stand-alone CSD proteins.	T21B14.13: HCH1 + S; T16I18.90: HD + bZIP + S; F5F19.21: HD + bZIP + S; MXC7.11: Transmembrane + S; F15H11.8: 2*S; F10B6.35: 4*S; F3I3.40: Possible artificial fusion with an RNA helicase
YS family	HCH1 + S

affect the substrate in favor of the cyclization and aromatization reactions. The roof of the tunnel, mainly consisting of the C-terminal helix, is predicted to be strongly hydrophobic and probably binds the surface of the polyketide that lacks polar groups; this interaction could induce the necessary conformation for the cyclization of the substrate.

Like the CSD–BA superfamily, most members of the cyclase/aromatase family contain a helix-capping polar, typically acidic residue immediately upstream of helix-2 (Fig. 4). It might also play a critical role in catalysis and/or anchoring the polar part of the substrate through hydrogen bonding. The cyclases/aromatases either contain a single START domain and act as dimers (TcmN) or consist of two START domains repeated in tandem (actinorhodin and griseusin Aro/Cyc).[39] The two-domain cyclases/aromatases undergo functional diversification of their tandem START domains, with each of them probably binding different reaction intermediates.[39] The functions of the cyclase/aromatase-type START (CAS) domains outside the *Actinomycetes* are not clearly understood. Of particular interest is the group of CAS proteins conserved in all proteobacteria, including *Rickettsia*, and in eukaryotes (Fig. 4). This pattern of phyletic distribution is typical of mitochondrial proteins, and the degree of conservation in eukaryotes suggests that it is likely to be a previously uncharacterized active cyclase/aromatase.

In addition to the CSD, BA, and CAS families, we identified two other less widespread and poorly characterized families of proteins containing the START domain. One of these is so far present only in plants and the cyanobacterium *Prochlorococcus*; some members of this family contain large inserts between strands 4 and 5 and between strands 6 and 7. The other family is well conserved in eukaryotes and is typified by the yeast Ydr214wp protein and its orthologs (Fig. 4).

This analysis shows that a similar substrate-binding tunnel in the START domain is used for very different enzymatic reactions, such as polyketide cyclization/aromatization and RNA degradation. In contrast, numerous members of the START superfamily appear not to possess any enzymatic activity but only to bind diverse ligands. This combination of enzymatic and nonenzymatic functions in the START superfamily may suggest that the catalytic activities of the START domain to a large extent depend on steric effects of the ligand being fit in the binding pocket and are secondarily grafted on a fold whose primary role is small-molecule binding.

Evolutionary Trends and Phyletic Distribution of START Domain Proteins

We analyzed the distribution of START domains in the genomes of various organisms with multiple-alignment-derived sequence profiles.[21] On the basis of sequence similarity clustering and reciprocal retrieval in BLAST searches, the START superfamily was subdivided into the five families previously mentioned (Fig. 4 and Table I). The CSD family proteins were originally detected only in plants and animals. Here, we also identified a CSD in the *Dictyostelium discoideum* protein CheaterA, mutations in which induce the preferential formation of spores rather than stalk.[40] In addition to the START domain, this protein contains an F-box and a WD40 β-propeller domain and is likely to function in a ubiquitin-linked pathway.[40] The presence of the START domain in this protein is of particular interest because it may mediate signaling by binding specific ligands such as the chlorinated hexanone derivative DIF-1, a *Dictyostelium* morphogen.[41] The presence of CSD in *Dictyostelium* also suggests that this domain was already present in the common ancestor of the multicellular eukaryotes and combined with a variety of other domains in each of the lineages. In plants, there is a specific expansion of proteins containing CSD and a Glabra-2-like homeodomain. A CSD was also detected in the bacterium *Pseudomonas aeruginosa* (gene product PA1579); this protein is closely related to the animal phospatidylcholine-binding proteins, which suggests horizontal acquisition from animals. It seems likely that the START domain of this *P. aeruginosa* protein plays a role in the interaction of the bacterium with animal cells.

The BA family appears to be specific to plants in which it has undergone considerable proliferation (Table I). This proliferation is reminiscent of some of the plant pathogenesis-related loci[42] and is compatible with stress-response-related functions of the BA family proteins, including transport of cytokinins and other small molecules and RNA degradation. The other distinct START domain family that thus far has been found only in plants and the cyanobacterium *Prochlorococcus* also shows a specific expansion in the former. The majority of these proteins appear to have arisen from tandem gene duplications at a single locus in *A. thaliana*. (Table I).

The CAS family is the most widespread group of START domains represented in bacteria, archaea, and eukaryotes (Table I). The proteobacterial/eukaryotic form typified by *Escherischia coli* YfjG (previously discussed) is accompanied in most proteobacteria by another small, uncharacterized gene in a conserved operon. In addition to this protein, *P. aeruginosa* encodes a unique form of the CAS domain that is associated with four transmembrane helices. CAS domains are absent in euryarchaea but are present in both sequenced crenarchaeal genomes, those of *Sufolobus solfataricum* and *Aeropyrum pernix*, which suggests an early invasion of the crenarcheal lineage via horizontal gene transfer from bacteria. We identified at least 18 members of the CAS family in *M. tuberculosis* and at least 10 in *S. coelicolor*, which correlates with the extensive production of polyketide metabolites by these organisms.[43] Recently, a polyketide toxin was identified in *Mycobacterium ulcerans*, and a role for similar metabolites in *M. leprae* and *M. tuberculosis* pathogenesis was proposed.[44] Furthermore, polyketides also function as siderophores and cell-wall components in mycobacteria.[43] The detection of the CAS family proteins is expected to aid in identifying the enzymes involved in the synthesis of these compounds. An analysis of the gene neighborhood of the *M. tuberculosis* CAS family proteins showed that some of these proteins are encoded in a cluster duplicated in tandem, whereas the genes for others are associated with genes for various enzymes such as oxidoreductases, methylases, and dioxygenases that are also involved in the biosynthesis of antibiotics in *Streptomyces* (Table I). Thus, the cyclase genes probably have expanded through tandem duplication followed by dissemination, eventually resulting in the association with the genes for other enzymes to which they are functionally linked. Studies in *Streptomyces* have shown the potential role of different cyclases in generating the diversity of the polyketide products.[37,39] The duplication and diversification of cyclases in *Actinomycetes* appear to represent an evolutionary mechanism for the generation of a diverse set of metabolites without much diversification of other enzymes in the pathway. Our prediction of numerous polyketide cyclases could aid in the combinatorial generation of biologically active polyketides, including antibiotics.

Finally, the START domains that are represented by a single copy in all sequenced eukaryotic genomes from *P. falciparum* to vertebrates (the Ydr214w-like family) are always found in combination with another globular domain (Table I). This domain contains a conserved histidine and also occurs as a stand-alone version in the yeast HCH1 protein, which appears to be functionally linked to the Hsp90 chaperone.[45] The degree of conservation of this protein across the wide range of eukaryotes is generally consistent with a specific enzymatic function. This version of the START domain was also found in *Mycobacterium*, where it is fused to a DNA-binding helix-turn-helix domain, suggesting a role in transcription regulation, and to another uncharacterized domain with a likely enzymatic function (Table I and data not shown); the presence of this domain in *Mycobacteria* is probably due to horizontal transfer from eukaryotes.

CONCLUSIONS

We showed that START domains belong to an ancient superfamily of helix-grip-fold proteins with a far greater diversity and phyletic spread than previously suspected. The phyletic distribution of different families within the START superfamily, with a widespread representation in bacteria, suggests that the domain itself emerged early in bacterial evolution. The ancestral START domain probably functioned as a ligand-binding domain but also could have evolved the cyclase/aromatase activity at an early stage of evolution. The extensive utility of this activity in metabolic diversification provided the niche for the large-scale expansion of this domain in *Actinomycetes*. From bacteria, the START domains probably were disseminated to some archaeal and eukaryotic lineages, on more than one occasion in the latter case. In the early eukaryotes, as in most bacteria, a small number of START domains probably retained their enzymatic activity. The emergence of multicellularity in eukaryotes brought about the extensive recruitment of two versions of the START domain in different capacities, which was accompanied by their proliferation. The first of these new eukaryotic functions, which involves the CSDs, is signaling mediated by lipid binding. In this case, the START domain underwent multiple lineage-specific fusions to other effector domains in a fashion that is typical of multidomain eukaryotic signaling proteins.[13] The second new function appears to be generally related to stress response and involves the BA family,[31,35] which emerged in the plant lineage and underwent proliferation along with the acquisition of multiple activities that include ligand binding and RNAse activity.

We predict that the entire START superfamily shares a common ligand-binding mode that appears to be the common denominator of the diverse functions of these domains. By showing that polyketide cyclases belong to the START-domain superfamily, we predict their three-dimensional structure and the catalytic pocket. The START domain is a relatively rare case of adaptation of the same protein fold for binding highly diverse ligands and catalyzing at least two distinct reactions via subtle modifications of the ligand-binding pocket. A detailed analysis of the interaction of START domains with their specific ligands/substrates could help in understanding the general principles of protein/small-molecule interactions and the evolution of catalysis.

REFERENCES

1. Morais Cabral JH, Lee A, Cohen SL, Chait BT, Li M, Mackinnon R. Crystal structure and functional analysis of the HERG potassium channel N terminus: a eukaryotic PAS domain. Cell 1998;95: 649–655.
2. Ponting CP, Aravind L. PAS: a multifunctional domain family comes to light. Curr Biol 1997;7:R674–677.
3. Aravind L, Koonin EV. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J Mol Biol 1999;287:1023–1040.
4. Taylor BL, Zhulin IB. PAS domains: internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol Rev 1999;63:479–506.
5. Aravind L, Ponting CP. The GAF domain: an evolutionary link between diverse phototransducing proteins. Trends Biochem Sci 1997;22:458–459.
6. Aravind L, Koonin EV. The STAS domain—an unexpected structural and functional link between human disease-associated anion transporters and bacterial antisigma-factor antagonists. Curr Biol 2000;10:R53–55.
7. Hurley JH, Misra S. Signaling and subcellular targeting by membrane-binding domains. Annu Rev Biophys Biomol Struct 2000;29:49–79.
8. Aravind L, Wolf YI, Koonin EV. The ATP-cone: an evolutionarily mobile, ATP-binding regulatory domain. J Mol Microbiol Biotechnol 2000;2:191–194.
9. Saraste M, Hyvonen M. Pleckstrin homology domains: a fact file. Curr Opin Struct Biol 1995;5:403–408.
10. Aravind L, Neuwald AF, Ponting CP. Sec14p-like domains in NF1 and Dbl-like proteins indicate lipid regulation of Ras and Rho signaling. Curr Biol 1999;9:R195–197.
11. Gane PJ, Dunwell JM, Warwicker J. Modeling based on the structure of vicilins predicts a histidine cluster in the active site of oxalate oxidase. J Mol Evol 1998;46:488–493.
12. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 1999;9:608–628.
13. Ponting CP, Aravind L. START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. Trends Biochem Sci 1999;24:130–132.
14. Kallen CB, Billheimer JT, Summers SA, Stayrook SE, Lewis M, Strauss III JF. Steroidogenic acute regulatory protein (StAR) is a sterol transfer protein. J Biol Chem 1998;273:26285–26288.
15. Akeroyd R, Moonen P, Westerman J, Puyk WC, Wirtz KW. The complete primary structure of the phosphatidylcholine-transfer protein from bovine liver. Isolation and characterization of the cyanogen bromide peptides. Eur J Biochem 1981;114:385–391.

16. Tsujishita Y, Hurley JH. Structure and lipid transport mechanism of a StAR-related domain. Nat Struct Biol 2000;7:408–414.
17. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
18. Gajhede M, Osmark P, Poulsen FM, Ipsen H, Larsen JN, Joost van Neerven RJ, Schou C, Lowenstein H, Spangfort MD. X-ray and NMR structure of Bet v 1, the origin of birch pollen allergy. Nat Struct Biol 1996;3:1040–1045.
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
20. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876–4882.
21. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 1999;15:1000–1011.
22. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
23. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
24. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
25. Kraulis PJ. Molscript. J Appl Crystallogr 1991;24:946–950.
26. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
27. Wang Y, Addess KJ, Geer L, Madej T, Marchler-Bauer A, Zimmerman D, Bryant SH. MMDB: 3D structure data in Entrez. Nucleic Acids Res 2000;28:243–245.
28. Carredano E, et al. Substrate binding site of naphthalene 1,2-dioxygenase: functional implications of indole binding. J Mol Biol 2000;296:701–712.
29. Nikolov DB, Chen H, Halay ED, Usheva AA, Hisatake K, Lee DK, Roeder RG, Burley SK. Crystal structure of a TFIIB-TBP-TATA-element ternary complex. Nature 1995;377:119–128.
30. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem Sci 1998;23:444–447.
31. Osmark P, Boyle B, Brisson N. Sequential and structural homology between intracellular pathogenesis-related proteins and a group of latex proteins. Plant Mol Biol 1998;38:1243–1246.
32. Fujimoto Y, Nagata R, Fukasawa H, Yano K, Azuma M, Iida A, Sugimoto S, Shudo K, Hashimoto Y. Purification and cDNA cloning of cytokinin-specific binding protein from mung bean (Vigna radiata). Eur J Biochem 1998;258:794–802.
33. Moiseyev GP, Fedoreyeva LI, Zhuravlev YN, Yasnetskaya E, Jekel PA, Beintema JJ. Primary structures of two ribonucleases from ginseng calluses. New members of the PR-10 family of intracellular pathogenesis-related plant proteins. FEBS Lett 1997; 407:207–210.
34. Bufe A, Spangfort MD, Kahlert H, Schlaak M, Becker WM. The major birch pollen allergen, Bet v 1, shows ribonuclease activity. Planta 1996;199:413–415
35. Gamas P, de Billy F, Truchet G. Symbiosis-specific expression of two Medicago truncatula nodulin genes, MtN1 and MtN13, encoding products homologous to plant defense proteins. Mol Plant Microbe Interact 1998;11:393–403.
36. Raya A, Revert F, Navarro S, Saus J. Characterization of a novel type of serine/threonine kinase that specifically phosphorylates the human goodpasture antigen. J Biol Chem 1999;274:12642–12649.
37. Alvarez MA, Fu H, Khosla C, Hopwood DA, Bailey JE. Engineered biosynthesis of novel polyketides: properties of the whiE aromatase/cyclase. Nat Biotechnol 1996;14:335–338.
38. Shen B, Hutchinson CR. Deciphering the mechanism for the assembly of aromatic polyketides by a bacterial polyketide synthase. Proc Natl Acad Sci U S A 1996;93:6600–6604.
39. Zawada RJ, Khosla C. Domain analysis of the molecular recognition features of aromatic polyketide synthase subunits. J Biol Chem 1997;272:16184–16188.
40. Ennis HL, Dao DN, Pukatzki SU, Kessin RH. Dictyostelium amoebae lacking an F-box protein form spores rather than stalk in chimeras with wild type. Proc Natl Acad Sci U S A 2000;97:3292–3297.
41. Morris HR, Taylor GW, Masento MS, Jermyn KA, Kay RR. Chemical structure of the morphogen differentiation inducing factor from Dictyostelium discoideum. Nature 1987;328:811–814.
42. Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BB, Jones JD. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. Cell 1997;91:821–832.
43. O'Hagan D. The polyketide metabolites. Chichester, UK: Horwood; 1991.
44. George KM, Chatterjee D, Gunawardana G, Welty D, Hayman J, Lee R, Small PL. Mycolactone: a polyketide toxin from Mycobacterium ulcerans required for virulence. Science 1999;283:854–857.
45. Nathan DF, Vos MH, Lindquist S. Identification of SSF1, CNS1, and HCH1 as multicopy suppressors of a Saccharomyces cerevisiae Hsp90 loss-of-function mutation. Proc Natl Acad Sci U S A 1999;96:1409–1414.