doi:10.1006/jmbi.2001.5378 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 317, 41–72

JMB

Classification and Evolution of P-loop GTPases and Related ATPases

Detlef D. Leipe, Yuri I. Wolf, Eugene V. Koonin* and L. Aravind

National Center for
Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894, USA

**Corresponding author*

Sequences and available structures were compared for all the widely distributed representatives of the P-loop GTPases and GTPase-related proteins with the aim of constructing an evolutionary classification for this superclass of proteins and reconstructing the principal events in their evolution. The GTPase superclass can be divided into two large classes, each of which has a unique set of sequence and structural signatures (synapomorphies). The first class, designated TRAFAC (after translation factors) includes enzymes involved in translation (initiation, elongation, and release factors), signal transduction (in particular, the extended Ras-like family), cell motility, and intracellular transport. The second class, designated SIMIBI (after signal recognition particle, MinD, and BioD), consists of signal recognition particle (SRP) GTPases, the assemblage of MinD-like ATPases, which are involved in protein localization, chromosome partitioning, and membrane transport, and a group of metabolic enzymes with kinase or related phosphate transferase activity. These two classes together contain over 20 distinct families that are further subdivided into 57 subfamilies (ancient lineages) on the basis of conserved sequence motifs, shared structural features, and domain architectures. Ten subfamilies show a universal phyletic distribution compatible with presence in the last universal common ancestor of the extant life forms (LUCA). These include four translation factors, two OBG-like GTPases, the YawG/YlqF-like GTPases (these two subfamilies also consist of predicted translation factors), the two signal-recognition-associated GTPases, and the MRP subfamily of MinD-like ATPases. The distribution of nucleotide specificity among the proteins of the GTPase superclass indicates that the common ancestor of the entire superclass was a GTPase and that a secondary switch to ATPase activity has occurred on several independent occasions during evolution. The functions of most GTPases that are traceable to LUCA are associated with translation. However, in contrast to other superclasses of P-loop NTPases (RecA-F1/F0, AAA +, helicases, ABC), GTPases do not participate in NTP-dependent nucleic acid unwinding and reorganizing activities. Hence, we hypothesize that the ancestral GTPase was an enzyme with a generic regulatory role in translation, with subsequent diversification resulting in acquisition of diverse functions in transport, protein trafficking, and signaling. In addition to the classification of previously known families of GTPases and related ATPases, we introduce several previously undetected families and describe new functional predictions.

Keywords: molecular evolution; LUCA; GTPase; TRAFAC; SIMIBI

Abbreviations used: GAP, GTPase activating protein; GEF, GTPase exchange factor; TRAFAC, translation factor; SIMIBI, signal recognition particle, MinD and BioD; SRP, signal recognition particle; LUCA, last universal common ancestor; BioD, dethiobiotin synthetase; FTHFS, formyltetrahydrofolate synthetase; PyrG, CTP synthetase; CobB, cobyrinic a,c-diamide synthetase; CobQ, cobyric acid synthetase; PurA, adenylosuccinate synthetase; SR, signal receptor; HypB, hydrogenase expression protein B; ML, maximum likelihood.

E-mail address of the corresponding author: koonin@ncbi.nlm.nih.gov

0022-2836/02/010041–32 $35.00/0

Introduction

Proteins that bind and hydrolyze nucleoside triphosphates are crucial for almost all aspects of life. They belong to several chain folds, most notably, the dinucleotide-binding (Rossmann) fold and the related tubulin/FtsZ fold, the mononucleotide-binding fold (P-loop NTPases), the protein kinase fold, the histidine kinase/HSP90/TopoII fold, and the HSP70/RNAse H fold.[1–6] P-loop NTPases are the most populous protein fold in most cellular organisms and comprise 10 to 18% of all gene products.[7] Structurally, P-loop NTPases are α/β proteins that contain regularly recurring α-β units with the β strands forming a central, (mostly) parallel β-sheet surrounded on both sides by α-helices. At the sequence level, the P-loop NTPase fold is characterized by an N-terminal Walker A motif, which consists of a flexible loop occurring between a strand and a helix.[1,3,8] The loop typically adopts the sequence pattern GxxxxGK [ST], whose function is to properly position the triphosphate moiety of a bound nucleotide. The distal Walker B motif contains a conserved aspartate (or, less commonly, glutamate) residue, which is situated at the end of a strand and binds a water-bridged Mg ion.[8]

At least seven major monophyletic lineages within the P-loop NTPase fold can be defined on the basis of distinct sequence and structural features. These are: (i) RecA and F1/F0-related ATPases; (ii) nucleic acid-dependent ATPases (helicases, Swi2, and PhoH-like ATPases); (iii) AAA+ ATPases; (iv) MJ/PH/AP/NACHT NTPases; (v) ABC-PilT ATPases; (vi) nucleotide kinases; and (vii) the GTPases.[9–12] (L.A. and E.V.K., unpublished results) Structurally, P-loop NTPases can be divided into two groups. One group includes the nucleotide kinases and the GTPases where the strand leading to the P-loop and the Walker B strand are direct neighbors. The other group, including AAA+, ABC, SF1/2 helicases, and RecA/F1 ATPases, is characterized by an additional strand inserted between the P-loop strand and the Walker B strand.

Historically, the first biochemical encounters with GTPases resulted from the investigation of enzymes that utilized GTP as opposed to ATP as a substrate to regulate a variety of critical cellular processes. These included initiation, elongation and termination of translation,[13] and signaling through 7TM receptors involved in sensing a variety of stimuli ranging from hormones to light.[14] This was soon followed by the identification of the Ras oncogene and the cellular Ras genes, and their relatives as intracellular transducers of growth factor signals. The first three-dimensional structure of a GTPase was that of elongation factor Tu (EF-Tu)[15] and, by the mid-eighties, the significant similarities in structure and sequence between Ras and EF-Tu had been recognized.[16–18] The subsequent sequencing and functional analysis of a large number of the GTPases extended this understanding and uncovered the great diversity amidst the GTPases; the first large-scale synthesis of this information, along with a rough classification of then well-known GTPases, including translation factors, Ras-related and signal recognition proteins, was published in 1991.[9]

The last decade has seen an explosion of knowledge regarding the structure and function of GTPases, especially in eukaryotes. The extended Ras-like family (see below) has been shown to have a significant role in different forms of signal transduction, intracellular trafficking, and cytoskeletal re-organization.[19,20] The regulators of these GTPases (the GTPase activating proteins, GAPs, and the GTPase exchange factors, GEFs) have been characterized extensively in terms of their biology and structure.[21–24] Similarly, the functions of other GTPases, such as septins involved in eukaryotic cell division and vesicular fusion,[25,26] and dynamins involved in vesicular trafficking,[27,28] have been investigated extensively. Structural studies have revealed details regarding the functions of several GTPases in the translation machinery, including initiation, elongation, and release factors, and other accessory factors such as the ERA GTPases.[29–31] Biochemical studies revealed previously unknown roles for the IF2 and OBG GTPases in translation-related processes.[32,33] Structure determination also resulted in clarification of the biochemistry of several other GTPase superclass proteins, such as the signal recognition particle (SRP) GTPases.[34,35]

Several families of previously undetected, predicted GTPases were identified by sequence analysis methods during genome-comparison studies, and the close relationship between several translation factor-related (predicted) GTPases such as Obg, EngA, TrmE, and YihA has been recognized†.[36] Furthermore, progress has been made in linking isolated families into larger assemblages and, in particular, the relationship between the

P-loop GTPases and the mechanochemical ATPases myosin and kinesin has been recognized.[39,40] Conversely, several P-loop NTPases that do not belong to the GTPase superclass considered here, such as the NACHT proteins, that belong to a recently defined, distinct assemblage, and the McrB protein of the AAA+ superclass, have been shown to possess GTPase activity.[41,42]

These studies have illuminated specific aspects of the relationship between different families of GTPases; however, to our knowledge, a formal definition of the GTPase superclass and a comprehensive classification of this vast assemblage of proteins do not exist. Classifications of protein structures, such as SCOP,[43,44] CATH,[45,46] and FSSP,[47] provide the framework for understanding the relationships between the structurally characterized members of this superclass, but do not substitute for a comprehensive analysis aimed at

† http://www.ncbi.nlm.nih.gov/COG/[37,38]

understanding evolutionary relationships. We sought to develop a phylogenetic classification of GTPases and related ATPases in order to reconstruct some of the key events in their evolution, from very early stages antedating the last common ancestor (LUCA) of all modern life forms, to later stages, such as diversification of the major lineages of eukaryotes and prokaryotes. Other ramifications of such a classification would include generation of a template for a systematic investigation of unexplored branches of the GTPase superclass and functional predictions that could guide future experiments. With these objectives, we explored the GTPase superclass in detail by using computational methods for sequence and structure comparison. Similarity-based clustering, traditional phylogenetic tree construction, and a cladistic approach (identification of shared derived characters) were combined to extract evolutionary information at various levels and to develop the classification. Approximately 60 distinct, ancient, conserved groups of GTPases were detected and shown to belong to two large classes, the first one including classic GTPases such as translation factors and the Ras family, and second one including the SRP GTPases and several families of ATPases. On the basis of this evolutionary classification, we hypothesize that the emergence and considerable diversification (into seven to ten distinct forms) of the GTPases antedate LUCA. The original function of the GTPases was probably related to translation, but subsequently they were recruited for a diverse array of other functions.

Results

Building an evolutionary classification of the GTPase superclass

The data for this study were gathered in two steps. Initially, all GTPases and their homologs, for which 3D structures were available, were extracted and their sequences and structures were aligned using DALI, VAST and the FSSP database. These sequence and structural alignments were used for preliminary identification of the structural features and their corresponding sequence cognates that differentiate the GTPase superclass members from the rest of the P-loop NTPases. The majority of these proteins were known or predicted GTPases, but some ATPases and ATP-binding proteins with no reported GTPase activity, such as NifH, ArsA, BioD, kinesins and myosins, also showed a clear structural relationship with GTPases. Examination of the structural and corresponding sequence alignments showed that they belong to the GTPase superclass, despite the lack of GTPase activity and, accordingly, they were included in this study. At the next step, representative sequences of all previously known groups of GTPases and their homologs were extracted from the database and analyzed to delineate the set of sequence motifs that define the GTPase superclass and their structural cognates. Those proteins that conformed to these features were selected as *bona fide* GTPase superclass members and used as seeds in profile searches, which were iterated until non-GTPase P-loop NTPases were detected. The searches were conducted transitively to maximize the chances of detection of distant members of the GTPase superclass. At each step, the retrieved candidates were assessed for membership in the GTPase superclass through examination of multiple alignments to detect the signature motifs.

All extracted protein sequences were subjected to detailed comparative analysis with the goal of elucidating evolutionary relationships. At the top level, these relationships were identified using mainly structural comparisons and some most persistent sequence features; at the intermediate level, the analysis relied on sequence signatures and on sequence similarity detection in iterative database searches; finally, phylogenetic analysis was primarily used for classification at the bottom level. At both the intermediate and bottom levels, similarity-based clustering of protein sequences as implemented in the BLASTCLUST program was employed for initial identification of candidate protein groups. For constructing the final evolutionary classification of GTPases, we attempted to apply the cladistic approach as consistently as possible, at least at the top and intermediate levels. Specifically, the sequence and structural motifs were identified that are likely to constitute shared derived characters (synapomorphies) and hence support clades in the range where conventional phylogenetic trees do not yield sufficient resolution. At the bottom level, the COG database was used as the guide for identification of orthologous relationships, particularly among prokaryotic proteins. Typically, phylogenetic analysis was performed for a single COG, although in some cases, the analyzed family included two or more paralogous COGs. It should be noted that the phylogenetic analysis described here focussed largely on deciphering the relationships between the three primary kingdoms (Archaea, Bacteria, and Eukaryota) and therefore only regions that could be aligned unambiguously between proteins from all three kingdoms within the given family were selected for phylogenetic analysis. Therefore, some of the trees do not provide good resolution for the branching pattern within a lineage, e.g. within the Bacteria.

Inferences on the most likely point of origin of GTPase families were made by taking into account both phyletic distribution and tree topology. Thus, if a particular family is widely represented in all three primary kingdoms, this is evidence in favor of its presence in LUCA. In addition, this conclusion is reinforced when the phylogenetic tree for this family conforms to the "standard model" topology, with a bacterial and archaeo-eukaryotic primary clade.[48] In contrast, the derivation of the family in LUCA or earlier becomes suspect when a fundamentally different topology is observed, such as grouping of bacteria with eukaryotes. In such a case, a (pre)LUCA origin of the given family necessitates the additional postulate of displacement of the ancestral form with the bacterial one in eukaryotes, which makes a bacterial or archaeal origin with subsequent dissemination by horizontal gene transfer a viable alternative. Below, in the discussion of the evolution of individual GTPase families, we follow these principles of phylogenetic inference, in some cases without referring to them explicitly.

The core of the GTPase domain contains seven β-strands, which we designated strands 1 through 7 to facilitate structural comparisons (Figure 1). This convention deviates from the previous nomenclature; the correspondence with the conserved G-motifs described by Bourne and co-workers[9] is as follows: G1 is strand 1, followed by Walker A motif, G2 is the loop N-terminal of strand 2, G3 is strand 4 (Walker B motif), G4 is strand 6 along with the [NT]KxD motif, and G5 is strand 7 along with the SA[KL] motif.

The GTPase superclass is divided into two distinct classes. Over 20 major monophyletic lines of descent, or families, were identified within these classes on the basis of distinct conserved features. Some of these families formed distinct clusters, or superfamilies, within the corresponding class. However, for a number of families, no specific higher-level relationships beyond assignment to one of the two classes could be recognized. Within each family, smaller monophyletic clusters, or subfamilies, were identified. Typically, a subfamily is a cluster of orthologs or closely related paralogs whose origin is traceable at least to the early stages of evolution of one of the superkingdoms of life (Archaea, Bacteria or Eukaryota); many of the subfamilies correspond to individual COGs as defined in the COG database.

The GTPase superclass of the P-loop NTPase fold

The GTPases are defined as a monophyletic superclass within the P-loop NTPase fold by a number of synapomorphies. At the sequence level, these include the specific form of the Walker B motif with a conserved glycine residue (typically, within the signature hhhhDxxG, where h is a hydrophobic residue) and the distal [NT]KxD motif that is not found in other P-loop NTPases.[9] The conserved Walker B glycine residue makes a hydrogen bond to the terminal γ-phosphate oxygen atom, whereas the [NT]KxD motif is responsible for the specificity for guanine over other bases.[9] GTPases also differ, with respect to the NTP hydrolysis mechanism, from other P-loop NTPases. Helicases, RecA, AAA + , or ABC-type NTPases rely on a conserved glutamate residue to serve as a general base in abstracting a proton from the catalytic (attacking) water molecule.[49–56] In contrast, in GTPases, the γ-phosphate group itself acts as the general base in abstracting a water proton, and subsequently, the generated nucleophilic hydroxide ion attacks the protonated γ-phosphate group to generate the penta-covalent reaction intermediate.[57–59]

On the basis of shared structural and sequence features, the GTPase superclass can be divided into two large classes. The first class includes the majority of the well-known GTPases; namely, the universal and lineage- specific translation factors, signal-transducing, heterotrimeric G proteins (HTGPs), the extended Ras family, with members involved in signal transduction as well as cell motility in prokaryotes, septins involved in eukaryotic cell division, dynamins involved in vesicular trafficking, and some ATPases, such as kinesin and myosin, that function as motors in eukaryotic cell motility and intracellular transport. This class was designated TRAFAC (for translation factor-related) GTPases. The second class includes the signal recognition particle-associated GTPases, the G3E family, the XAB1 family, the CLP1 family, the MRP/MinD-related superfamily, and several metabolic enzymes, such as BioD-related enzymes, adenylosuccinate synthetase, and formyltetrahydrofolate ligase. This class was named SIMIBI for its three largest subgroups, the signal recognition GTPases, the MinD superfamily, and the BioD superfamily. In both GTPase classes, the NKxD motif that provides specificity for GTP can be secondarily disrupted or modified, which results in a loss of specificity as observed, for example, in myosin, kinesin, and the MRP/MinD/BioD proteins. A complete evolutionary classification of the GTPase superclass is given in Table 1, and the phyletic distribution of the delineated groups of GTPases is shown in Table 2. Figure 1 shows topology diagrams for selected GTPase families, Figures 2 and 3 show the alignments for the two GTPase classes, Figure 4 depicts the domain architectures of multidomain GTPases, Figure 5 shows the ML phylogenetic trees for GTPase families or subfamilies with broad phyletic distribution. Phylogenetic analysis was also reproduced using the neighbor-joining methods, but in no case were substantial differences in the tree topology observed compared to the ML trees, and therefore the neighbor-joining trees are not shown. Figure 6 shows the proposed evolutionary scenarios for the ancient conserved groups (individual subfamilies of GTPases) superimposed on a relative temporal scale. Below, using this information, we briefly discuss the salient features that support individual clades and some of the functional and evolutionary implications.

The TRAFAC class

In sequence terms, the TRAFAC GTPases are characterized by a highly conserved threonine or serine residue in the loop between strands 2 and 3, that makes a hydrogen bond to the Mg cation required for GTP hydrolysis[9] (Figures 1 and 2). In addition, many, if not all, TRAFAC GTPases have a conserved serine residue in strand 7 that is

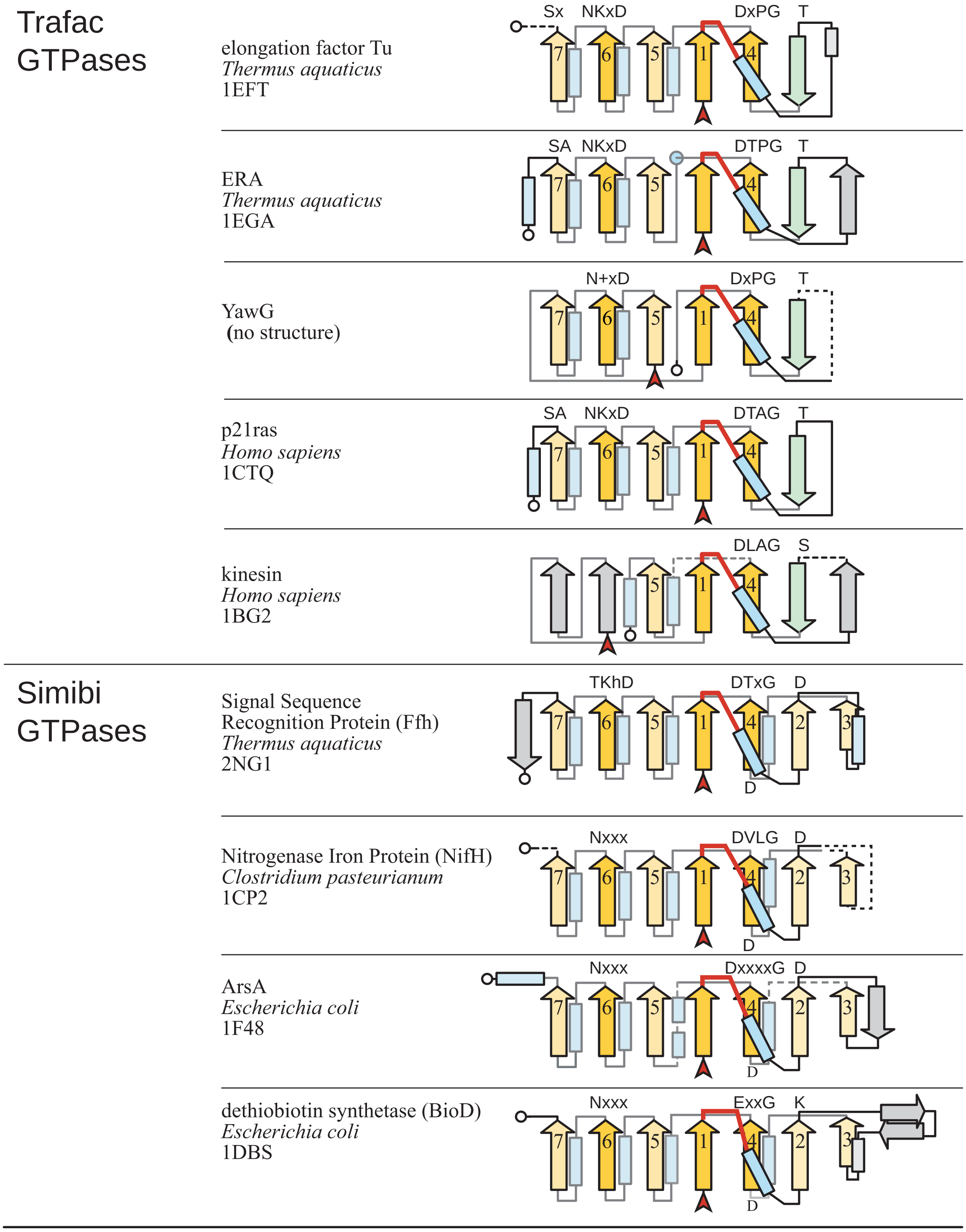

Figure 1. Topology diagrams of selected P-loop GTPases. Sequences are identified with protein and species names, and the PDB code. Strands are shown as arrows with the arrowhead on the C-terminal side and numbered 1 through 7. Strands 1, 4, and 7 that encompass the conserved sequence motifs GxxxGK[ST] (Walker A), DxxG (Walker B), and [NT]KxD (the GTP-specificity motif) are rendered in orange; strands 5 and 7 are in light orange; the antiparallel strand of the TRAFAC GTPases is shown in green, and other strands that presumably were absent from the ancestral GTPase domain are in gray. Helices are shown as blue rectangles when above the plane of the β-sheet and in faint blue when below the β-sheet. The P-loop is shown as a red line, a red arrowhead marks the N terminus of the GTPase domain and the C terminus is shown by a ring. Broken lines indicate secondary structure elements that have been left out for clarity. The diagram for the YawG/YlqF family is based on sequence comparison only (no structure available) to demonstrate the circular permutation of the GTPase domain (see the text). Highly conserved sequence signatures are shown above or below their respective strands with the family consensus (x for any amino acid, h for hydrophobic).

Table 1. Evolutionary classification of the GTPase superclass of P-loop NTPases

Main synapomorphies: DXXG motif in Walker B, NKXD motif after strand 6

A. *TRAFAC class*
Special features: strand 3 adjacent and anti-parallel to strand 4 (Walker B); conserved threonine preceding strand 3
Translation factor superfamily
Classic translation factor family
Special features: association of GTPase domain with β-EI domain
IF2/eIF5B subfamily
eIF2γ/SelB subfamily
EF-Tu/EF-1α subfamily
EF-G/EF-2 subfamily
Bms1-like family
Special features: associated with a C-terminal SelB domain
OBG-HflX-like superfamily
Special features: FxT motif N-terminal to strand 3
OBG family
Special features: typically associated with C-terminal TGS domain
OBG subfamily
YyaF/YchF subfamily
Ygr210 subfamily
Drg subfamily
Nog1 subfamily
HflX family
Special features: typically associated with a glycine-rich segment N-terminal to the GTPase domain
TrmE-Era-EngA-YihA-Septin like superfamily
TrmE family
Special features: contains a large conserved extension N-terminal to the GTPase domain
FeoB family
Special features: associated with membrane spanning segments
YihA(EngB) family
Special features: DxxG[FY]G type motif associated with Walker B
Era family
Special features: Typically associated with C-terminal pseudo-KH domain
YfgK(EngA) family
Special features: duplication of the GTPase domain
Septin family
Special features: loss of asparagine in the NKXD motif of several members of this family
Septin/Pnut subfamily
C18B2.5-like paraseptin subfamily
Aig1/Toc34/Toc159-like paraseptin subfamily
VC1806-like paraseptin subfamily
YfjP-like proteins
Ras-like superfamily
AP-GTPase family
dynein light intermediate chain 1/2 family
Extended Ras-like family
Ras/Rab subfamily
Rho/Rac/Ran subfamily
Arf subfamily
Heterotrimeric GTPase subfamily
MglA subfamily
Myosin-kinesin superfamily
Special features: Loss of the strand 6 and 7 and the presence of additional N-terminal strands that take their position in 3D structure
Myosin family
Kinesin family
Separate families of the TRAFAC class
YlqF/YawG Family
Special features: circularly permuted GTPases
YqlF subfamily
YawG subfamily
MJ1464 subfamily
YqeH subfamily
YjeQ subfamily
Dynamin/Fzo/YdjA family
Special features: presence of characteristic hhP motif N-terminal of P-loop
Dynamin subfamily
Fuzzy onions subfamily
YdjA subfamily
GB1/RHD3 family
Special features: Presence of modified hhhRD motif in place of the classic NKXD motif
RHD3 subfamily
BFP-GB1 subfamily

B. *SIMIBI class*

- Special features: strand 2 adjacent and parallel to Walker B strand; conserved asparate at the end of strand 2
- **MinD/Mrp-ETK superfamily**
- Special feature: typically contain additional aspartate in strand 2 and only the N of the NKXD motif is retained
  - **Mrp/MinD family**
  - Special features: presence of KGGh motif in the P-loop and retention of only the asparagine in the NKXD motif
    - Mrp/NBP35 subfamily
    - AF2380 subfamily
    - MinD subfamily
    - MotR/FlhG subfamily
    - ParA/Soj subfamily
    - NifH subfamily
    - ChlL/FrxC subfamily
    - ArsA subfamily
  - **ETK family**
  - Special features: bacterial tyrosine kinases Walker B of the form hhhhDTPP
- **BioD-FTHFS superfamily**
- Special features: loss of NKXD, substitution of D at the end of strand 2 by a basic residue, E in Walker B
  - **BioD family**
    - Dethiobiotin synthase (BioD proper) subfamily
    - PTA subfamily
    - PyrG subfamily
    - CobQ subfamily
    - CobB subfamily
  - **PurA family**
  - Special features: arginine at the end of strand 2 with a TKXD associated with strand 6
  - **Ta0025 family**
  - Special features: replacement of S/T in Walker A motif by a R or M
  - **FTHFS family**
  - Special features: Large insert between strand 2 and 3 of the NTPase domain and Walker A of the form TPXGEGK[TS]

Separate families of the SIMIBI class

- **Signal recognition associated GTPase family**
- Special features: unique N-terminal α-helical extension
  - SRP54/Ffh subfamily
  - SR/FtsY subfamily
  - FlhF subfamily
- **G3E family**
- Special features: intact NKXD combined with the E in Walker B signature
  - UreG subfamily
  - HypB subfamily
  - CobW subfamily
  - ArgK subfamily
- **Clp1/Gcr3 family**
- Special features: NKXD motif completely eroded, typical D in Walker B substituted by N
- **XAB1 family**
- Special features: canonical [NT]KXD motif with a GPNG signature associated with the third strand of the NTPase domain

involved in guanine base binding.[9,60] Structurally, the distinctive feature of the TRAFAC class is that the strand that flanks the Walker B-containing strand 4 on the right-hand side is antiparallel to it (Figure 1). In the SIMIBI class, the topology of the equivalent elements is completely different (see below).

The translation factor superfamily

This superfamily includes the classical translation factors with several closely related ancient subfamilies and the enigmatic BMS1 family that is restricted to the eukaryotic lineage.

The classic translation factor family

Four ancient conserved subfamilies of GTPases in the translation factor family are widespread, if not ubiquitous, in all three superkingdoms, namely IF2/eIF5B, eIF-2(/SelB, EF-Tu/EF-1α, and eEF2/EF-G (Table 2). Each of these translation factors probably has been vertically inherited from LUCA (Figure 6). Elongation factor EF-Tu/EF-1α is a three-domain protein (Figure 4) that forms a ternary complex with aminoacyl-tRNA and protects the aminoester bond against hydrolysis until a correct match between codon and anticodon is achieved. The aminoacyl-tRNA complex is released upon GTP hydrolysis to allow incorporation of the amino acid into the nascent protein chain.[60] GDP release in EF-Tu/EF-1α relies on a specific GEF, EF-Ts.[61] SelB is a four-domain selenocysteine-specific elongation factor [62,63] that so far has been detected only in several scattered bacterial species. The eIF-2(proteins are the archaeo-eukaryotic orthologs of SelB that function as initiation factors, rather than as elongation factors. The eIF-2(sequences contain a characteristic insertion of a Zn-ribbon N-terminal to the Walker B strand. The archaeo-eukaryotic eIF-2γ, along with other eIF2 subunits, forms a ternary complex with GTP and $\text{Met-tRNA}_\text{i}^\text{Met}$ and facilitates the binding of $\text{Met-tRNA}_\text{i}^\text{Met}$ to the ribosome to form a 43 S preinitia-

Table 2. Phyletic distribution of GTPase families and subfamilies

Protein	Spiro-chaet.	Chlamy-diales	Aqui-fex	Thermo-toga	Deino-coccus	Firmi-cutes	Cyano-bac.	Proteo-bac.	Eury-arch.	Cren-arch.	Met-azoa	Green plants	Fungi	Other euks
EFG/EF2	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RF-3	−	−	−	−	+	+	+	+	−	−	−	−	−	−
LepA	+	+	+	+	+	+	−	−	−	−	+	+	+	−
EFTu/1α	+	+	+	+	+	+	+	+	+	+	+	+	+	+
GP-1	−	−	−	−	−	−	−	−	+	+	+	−	−	−
CysN/NodQ	−	−	−	−	−	+	−	+	−	−	−	−	−	−
eRF2/HBS1	−	−	−	−	−	−	−	−	−	−	+	+	+	+
elF2/SeIB	−	−	+	−	−	−	−	+	+	+	+	+	+	+
IF2	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Era	+	−	+	+	+	+	+	+	−	−	+	+	−	+
ThdF	+	+	+	+	+	+	+	+	−	−	+	+	+	−
EngA	+	+	+	+	+	+	+	+	−	−	−	+	−	−
FeoB	+	+	+	+	+	+	+	+	+	−	−	−	−	−
HflX	−	+	+	−	+	+	+	+	+	+	+	+	−	−
YihA/EngB	+	−	+	+	−	+	−	+	+	−	+	+	+	+
DRG	−	−	−	−	−	−	−	−	+	+	+	+	+	+
NOG1	−	−	−	−	−	−	−	−	+	+	+	+	+	+
Ygr210	−	−	−	−	−	−	−	−	+	+	−	−	+	−
YyaF	+	+	+	+	+	+	+	+	−	−	+	+	+	+
obg	+	+	+	+	+	+	+	+	−	−	+	+	+	−
YawG	−	−	−	−	−	−	−	−	−	−	−	−	−	−
YIqF	+	−	−	+	−	+	+	+	−	−	+	+	+	−
YjeQ	+	−	−	+	−	+	+	+	−	−	−	+	−	−
YqeH	−	−	−	+	−	+	−	+	−	−	+	+	−	−
Ras/MgIA	−	−	+	−	+	−	−	+	+	−	+	+	+	+
sep/Toc33	−	−	−	−	−	−	+	+	−	−	+	+	+	−
dyn/YjdA	−	−	−	−	+	+	+	+	−	−	+	+	+	+
kin/myos	−	−	−	−	−	−	−	−	−	−	+	+	+	+
Bmsl	−	−	−	−	−	−	−	−	−	−	+	−	+	+
hGBP1/RHD3	−	−	−	−	−	−	−	−	−	−	+	+	+	−
AP-GTPase	−	−	−	−	−	−	−	−	−	−	+	−	−	−
SR54/Ffh	+	+	+	+	+	+	+	+	+	+	+	+	+	+
FtsY/SR	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RlhF	+	−	+	+	−	+	−	+	−	−	−	−	−	−
ArgK	−	−	−	−	+	+	−	+	+	+	+	−	−	+
CobW	−	−	+	−	+	+	+	+	−	−	+	+	+	−
HypB	−	−	+	−	+	+	+	+	+	−	−	−	−	−
UreG	+	+	−	+	+	+	+	+	−	−	−	+	+	−
XPAbp1	−	−	−	−	−	−	−	−	+	+	+	+	+	+
NifH	+	−	−	−	−	+	+	+	+	−	−	−	−	−
BchL	−	−	−	−	−	+	+	+	−	−	−	+	−	+
Mrp	−	−	+	+	+	+	+	+	+	+	+	+	+	+
MinD/MotR	−	−	+	+	+	+	+	+	+	−	−	+	−	−
ArsA	−	−	+	−	−	+	+	+	+	+	+	+	+	−
Soj	+	+	+	−	+	+	+	+	+	+	−	−	−	−
AF2380	−	−	−	−	−	−	−	−	+	−	−	−	−	−
ETK/Wzc	−	−	−	−	−	−	+	+	−	−	−	−	−	−
Clp1/Grc3	−	−	−	−	−	−	−	−	+	+	+	+	+	+
BioD	−	+	+	−	−	+	+	+	+	−	−	−	+	−
PyrG	+	+	+	+	+	+	+	+	+	+	+	+	+	+
CobB	−	−	−	−	−	+	+	+	+	+	−	−	−	−
CobQ	−	−	−	−	+	+	+	+	+	+	−	−	−	−
Fthfs	−	−	−	+	−	+	−	+	+	−	+	+	+	+
PurA	−	−	+	+	+	+	+	+	+	+	+	+	+	+
PTA	−	−	−	−	+	−	+	+	+	−	−	−	−	−
Ta0025	−	−	−	+	−	−	−	−	+	−	−	−	−	−

tion complex.[64] EF2/EF-G is a ubiquitous five-domain GTPase (Figure 4) that catalyses the translocation of tRNA on the ribosome. IF2/eIF5B is also ubiquitous in Bacteria, Archaea, and Eukaryota, and mediates the binding of Met-tRNA$_i^{Met}$ to the ribosome and, in eukaryotes, joining of the two ribosomal subunits.[30,32,65]

GTPases are involved in translation termination, but the origins of the release factors differ in the three superkingdoms. Bacterial release factor RF-3 is derived from the translocation factor EF2/EF-G subfamily, whereas the eukaryotic release factor eRF3/HBS1 is a paralog of the elongation factor EF-Tu/EF-1α; there is no corresponding release factor in Archaea. Thus, the release factors apparently have been recruited for translation termination from the pool of ancient translation-associated GTPases after the divergence of the three superkingdoms.[66,67]

Several other groups of GTPases also belong to these ancient translation factor sub-families and apparently were derived through lineage-specific duplication and divergence of translation factors such as Ef2/EF-G and EF-Tu/EF-1α. LepA is a derivative of the EF2/EF-G family and is ubiquitous in Bacteria and Eukaryota (e.g. yeast GUF1p), but is missing from Archaea (Figure 5(a)). This pattern of phyletic distribution suggests that LepA evolved through a duplication of the EF-G gene in bacteria, followed by early transfer into the eukaryotic lineage, most likely from the promitochondrial endosymbiont. TypA (tyrosine phosphorylated protein A)/BipA, whose exact function remains unknown, is another product of EF-G duplication, which is widespread in Bacteria and is present also in plants, which might have acquired this gene from the pro-chloroplast symbiont (Figure 5(a)). Another similar group of lineage-specific derivatives of the EF-G/EF2 subfamily includes the TetQ/TetM proteins that appear to have spread in bacteria *via* mobile elements owing to their role in reversing the action of tetracycline on the ribosome.[68] Similarly, the GP-1 proteins of Archaea and Eukaryota are a lineage-specific derivative of the EF-TU/EF-1α subfamily and have a distinct N-terminal domain. CysN/NodQ proteins are GTPases that function as regulatory subunits of ATP sulfurylase, which catalyses the first step of inorganic sulfate assimilation for the biosynthesis of sulfur, containing compounds in proteobacteria and Mycobacterium.[69,70] These might have evolved *via* a duplication of EF-Tu, with subsequent exaptation for a function unrelated to translation.

Phylogenetic analyses of individual translation factor subfamilies conform to the ''standard model'' of evolution,[48] by showing a clear bifurcation into a bacterial and an archaeo-eukaryotic branch (Figure 5(a) and data not shown). This evolutionary pattern suggests that all four of these subfamilies were already present in LUCA.[71,72] Phylogenetic analysis of the GTPase domain groups EF-Tu/EF-1α with eIF-2γ/SelB, whereas EF-G/EF2 clusters with IF2 (trees not shown). Therefore, a possible scenario for the evolution of translation factors involves a two-domain ancestor with an N-terminal GTPase domain and a C-terminal EI domain (the Elongation factor- Isomerase domain),[73] with a function in RNA/protein interaction, probably in the context of a primitive translation system. The first gene divergence event produced the common ancestor of EF-Tu/EF-1α and eIF2/SELB, which both rely on dedicated guanine-nucleotide exchange factors (GEFs), and a low-GDP-affinity GTPase, the common ancestor of the EF-G/EF2 and IF2 family, both of which rely on tRNA mimicry and do not use GEFs.

The Bms1 family

This family includes two closely related paralogs, typified by the *Saccharomyces cerevisiae* proteins Bms1p and Tsr1p, which are widely conserved in eukaryotes. They are not closely related to any particular GTPase family, but group with the classic translation factors and their relatives in similarity-based clustering. These proteins contain a previously unidentified N-terminal GTPase domain and a C-terminal domain that is found also in the SelB proteins. Neither Bms1p nor Tsr1p have been characterized biochemically, but their strong conservation in eukaryotes, including early-branching protists, such as Plasmodium, and the highly degraded *Guillardia theta* nucleomorph, point to an important function. The yeast Bms1p localizes mainly to the nucleolus and the nuclear pore complex,[74] which suggests that it might function as a critical regulator of RNA transport and/or in ribosomal assembly. Tsr1p is over-expressed, along with translation factors and ribosomal proteins, under several stress conditions and gene disruption causes a lethal phenotype in yeast.[75] The Bms1p orthologs have a D → E substitution in the Walker B motif, which is the only such case in the TRAFAC class, in contrast to the relatively common presence of this substitution in the SIMIBI class (Figures 2 and 3, and see below). Additionally, Bms1p and its orthologs have the THXD sequence replacing the canonical [NT]KXD motif, which is uncommon in the translation factor superfamily (Figure 2). Tsr1p and its orthologs appear to be inactive members of this family with disrupted Walker A and B motifs, and probably function as non-enzymatic regulators, perhaps of the Bms1p function.

The OBG-HflX superfamily

The OBG and HflX families form a distinct group in sequence-similarity-based clustering. An apparent synapomorphy, a phenylalanine residue in the loop N-terminal of the antiparallel strand, which is conserved in these families, but not in other GTPases (Figure 2), also seems to support the monophyly of this assemblage.

The HflX family

A distinct conserved domain with a glycine-rich segment N-terminal of the GTPase domain (Figure 4) characterizes the HflX family. The *Escherichia coli* HflX has been implicated in the control of the λ phage *c*II repressor proteolysis,[76] but the actual biological functions of these GTPases remain unclear. HflX is widespread, but not universally represented in all three superkingdoms (missing, for example, from Mycoplasma, the epsilon subdivision of Proteobacteria, spirochaetes, the archaeon Methanobacterium, and fungi). In phylogenetic analysis, most of the eukaryotic sequences group with the α-proteobacteria (Figure 5(b)), which indicates that HflX probably was acquired

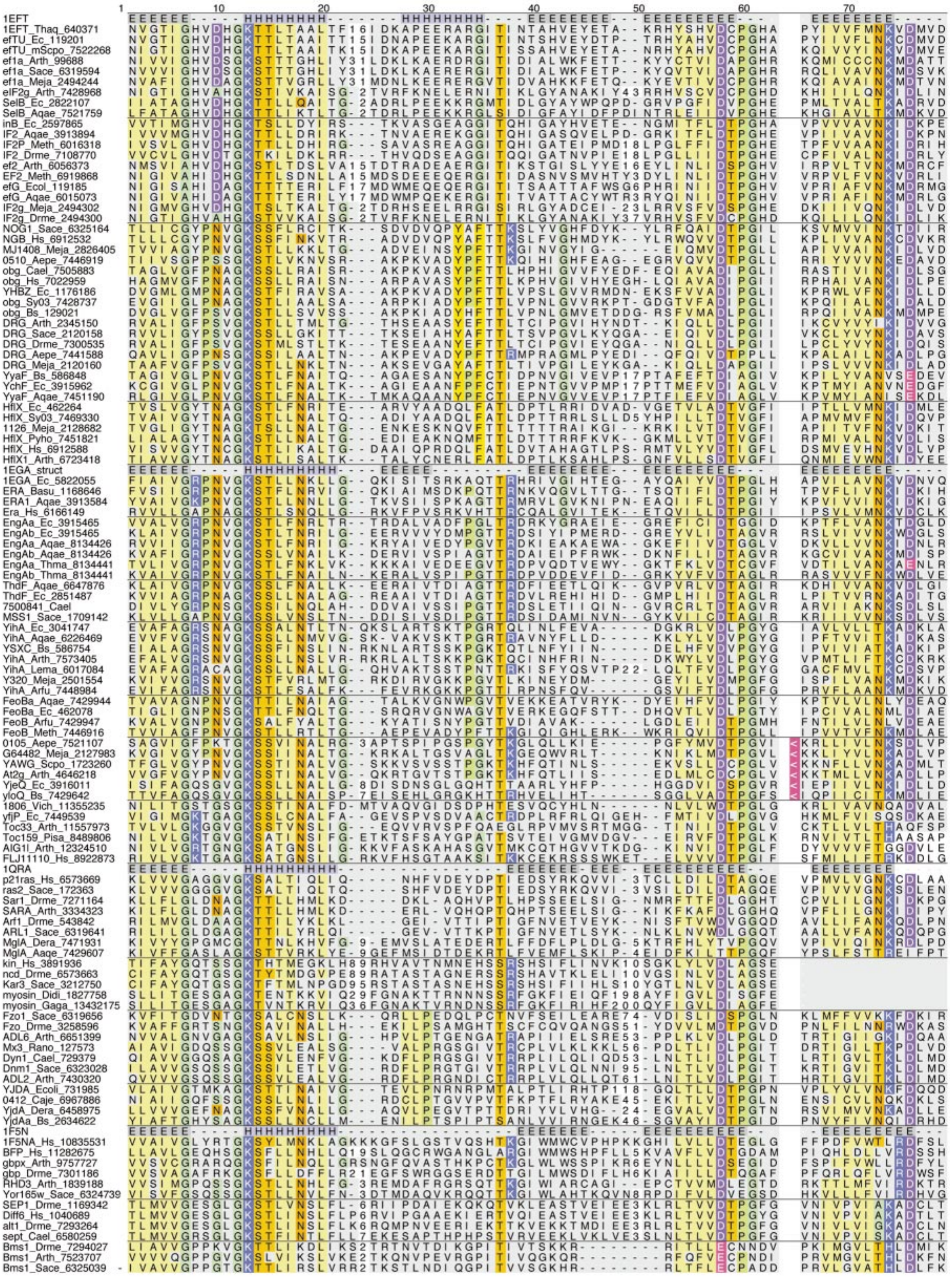

Figure 2 (*legend opposite*)

by eukaryotes *via* the mitochondrial route. If HflX was present in LUCA, one has to assume that, in Eukaryota, the ancestral version of HflX has been displaced by the α-proteobacterial ortholog following mitochondrial endosymbiosis. However, with the data available, we cannot discount an alternative scenario that assumes that HflX arose in Bacteria, and that the current broad phyletic distribution of the HflX family emerged secondarily *via* multiple horizontal transfers.

The Obg family

The Obg family consists of five well-delimited, ancient subfamilies, namely obg, DRG, YyaF/YchF, Ygr210, and NOG1. Four of these groups (Obg, DRG, YyaF/YchF, and Ygr210) are characterized by a distinct glycine-rich motif immediately following the Walker B motif (consensus GAxxGxGxGxxxl where l is one of the aliphatic residues I, L, or V; not shown). The NOG1 subfamily, while lacking this motif, shares the motif YxFTTxxxxxG in the second strand (G2) with the rest of the Obg family (Figure 2). Obg/CgtA is an essential gene that is involved in the initiation of sporulation and DNA replication in the bacteria Caulobacter and Bacillus, but its exact molecular role is unknown.[77,78] *Bacillus subtilis* Obg is associated with the ribosome and, specifically, with the ribosomal protein L13.[33] Furthermore, several OBG family members possess a C-terminal RNA-binding domain, the TGS domain (Figure 4), that is present also in threonyl-tRNA synthetase and in bacterial guanosine polyphosphatase SpoT.[79] These observations, taken together with the universal phyletic distribution of some of the OBG subfamilies (Table 2), suggest that these proteins are hitherto uncharacterized translation factors. Nog1 is a nucleolar protein that might function in ribosome assembly.[80] Walker A, Walker B and NKxD motifs are highly conserved in the Obg family, with the exception of the YyaF/YchF subfamily, in which the aspartate residue in the NKxD motif is replaced by a glutamate residue (Figure 2). The DRG and Nog1 subfamilies are ubiquitous in Archaea and Eukaryotes, the Ygr210 subfamily is present in Archaea and fungi, and the Obg and YyaF/YchF subfamilies are ubiquitous in Bacteria and Eukaryotes (Table 2 and Figure 5(c)). The Obg/Nog1 and DRG subfamilies appear to form one major branch of the Obg family and the Ygr210 and YyaF subfamilies form another branch, both of which appear to be traceable to LUCA (Figure 5(c)). Among eukaryotes, the Ygr210 subfamily is represented only in fungi. These fungal proteins form a tight cluster with their archaeal orthologs, which suggests the possibility of horizontal transfer from Archaea to fungi; alternatively, this subfamily might have been lost in other crown-group eukaryotic lineages (Figure 5(c)). The Nog1 subfamily probably emerged early in the archaeo-eukaryotic lineage through duplication and divergence from the DRG family, which might have involved a shift in function (Figure 5(c)). The ubiquitous presence of each of the two major branches of the Obg family (Ygr210-YyaF/YchF and obg-Drg(Nog1); Figure 5(c)) suggests that the ancestor of each branch was already present in LUCA and, accordingly, the original duplication and subsequent divergence in this family antedates LUCA (Figure 6). Both the Obg and YyaF/YchF subfamilies include well-defined bacterial-eukaryotic clusters, which is compatible with the possibility of horizontal transfer from bacteria to eukaryotes (Figure 5(c)). The bacterial/eukaryotic YyaF/YchF subfamily and the archaeo-eukaryotic Drg subfamily, both of which contain the RNA-binding TGS domain, might most closely resemble the ancestor of the Obg family that was present in LUCA.

The TrmE-Era-EngA-Septin-like superfamily

This heterogeneous assemblage supported by similarity-based clustering includes several distinct families that generally show sequence conservation in the region between the Walker A and B motifs, to the exclusion of other GTPases. Most lineages in this assemblage represent predominantly bacterial elaborations of the TRAFAC class.

Figure 2. Sequence alignment of the TRAFAC GTPase class. The alignment shows two conserved sequence regions, the region surrounding the Walker A and B motifs (first block), and the region around the NKxD motif and its derivatives (second block). From top to bottom, the alignment includes sequences from 20 families separated by horizontal lines: EF-Tu/1a, eIF2 g/SelB, IF2, EF-G/2, Obg/DRG, HflX, Era, EngA, TrmE, YihA, FeoB, YawG/YjeQ, yjfP/Toc33, Ras/Arf/MglA, kinesin/myosin, dynamins/Fuzzy onions, YjdA, hGBP1/RHD3, septins, Bms1. The NKxD motif is located at the N terminus of the YawG/YjeQ family proteins (red arrowheads), but the order of sequence elements is reversed in the Figure to show the alignment with other GTPases. Residues that are widely conserved or discussed in the text are color-coded in light yellow for hydrophobic residues (A,C,I,F,L,M,T,Y,W), green for small residues (G,A,S), orange for amides (N,Q), light orange for hydroxy residues (S,T), yellow for aromatic residues (F,Y,W), blue for basic residues (K,R), purple for aspartate, and red for glutamate. Secondary structure elements are shown above the respective sequence as E for strand and H for helix and identified with the PDB code. The protein name, species name abbreviation, and the GenBank GI number identify sequences. Species names are abbreviated as follows: Acam, *Acidianus ambivalens*; Aepe, *Aeropyrum pernix*; Aqae, *Aquifex aeolicus*, Arfu, *Archeoglobus fulgidus*; Arth, *Arabidopsis thaliana*; Azvi, *Azotobacter vinelandii*; Bs, *Bacillus subtilis*; Cael, *Caenorhabditis elegans*; Caje, *Campylobacter jejuni*; Clpa, *Clostridium pasteurianum*; Chvu, *Chlorella vulgaris*; Dera, *Deinococcus radiodurans*; Didi, *Dictyostelium discoideum*; Drme, *Drosophila melanogaster*; Ec, *Escherichia coli*; HaC1, *Halobacterium* sp. NRC-1; Hepy, *Helicobacter pylori*; Hs, *Homo sapiens*; Lema, *Leishmania major*; Meja, *Methanococcus jannaschii*; Meth, *Methanobacterium thermoautotrophicum*; Mytu, *Mycobacterium tuberculosis*; Pisa, *Pisum sativum*; Pyab, *Pyrococcus abyssi*; Pyho, *Pyrococcus horikoshii*; Rano, *Rattus norvegicus*; Rhrh, *Rhizobium rhizogenes*; Sace, *Saccharomyces cerevisiae*; Scpo, *Schizosaccharomyces pombe*; Stco, Sy01, *Synechococcus PCC8801*; Sy03, *Synechocystis PCC6803*; Thac, *Thermoplasma acidophilum*; Thaq, *Thermus aquaticus*; Thma, *Thermotoga maritima*; Vich, *Vibrio cholerae*. A lower-case c in front of the species abbreviation identifies a chloroplast sequence. In proteins with two GTPase domains, the protein name is suffixed with a or b for the N and C-terminal GTPase domain, respectively, e.g. EngAa and EngAb.

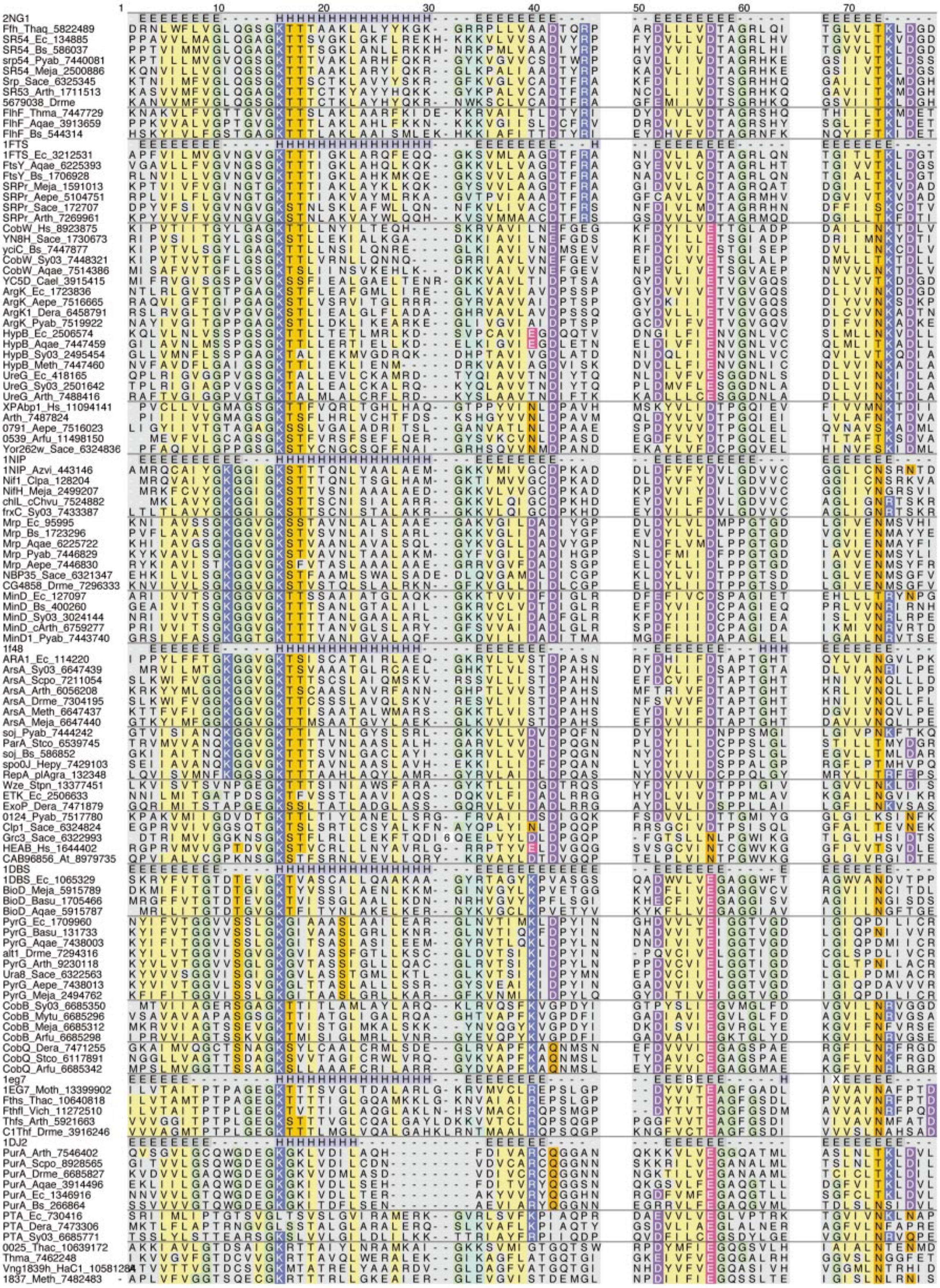

Figure 3. Sequence alignment of the SIMIBI GTPase class. The alignment shows the three most highly conserved regions, the P-loop and the two surrounding strands (block 1), the Walker B motif (block 2), and the NKxD motif and its derivatives (block 3). From top to bottom, the alignment includes sequences from 20 subfamilies and families separated by horizontal lines: SR54/Ffh, FlhF, SR/FtsY, G3E, XPAbp1, NifH/ChlL, Mrp, MinD, XPAbp1, ArsA, soj, Wze/ETK, Clp1, BioD, PyrG, CobB, CobQ, FTHFS, PurA, PTA, Ta0025. Organism name abbreviations and residue coloring are as for Figure 2.

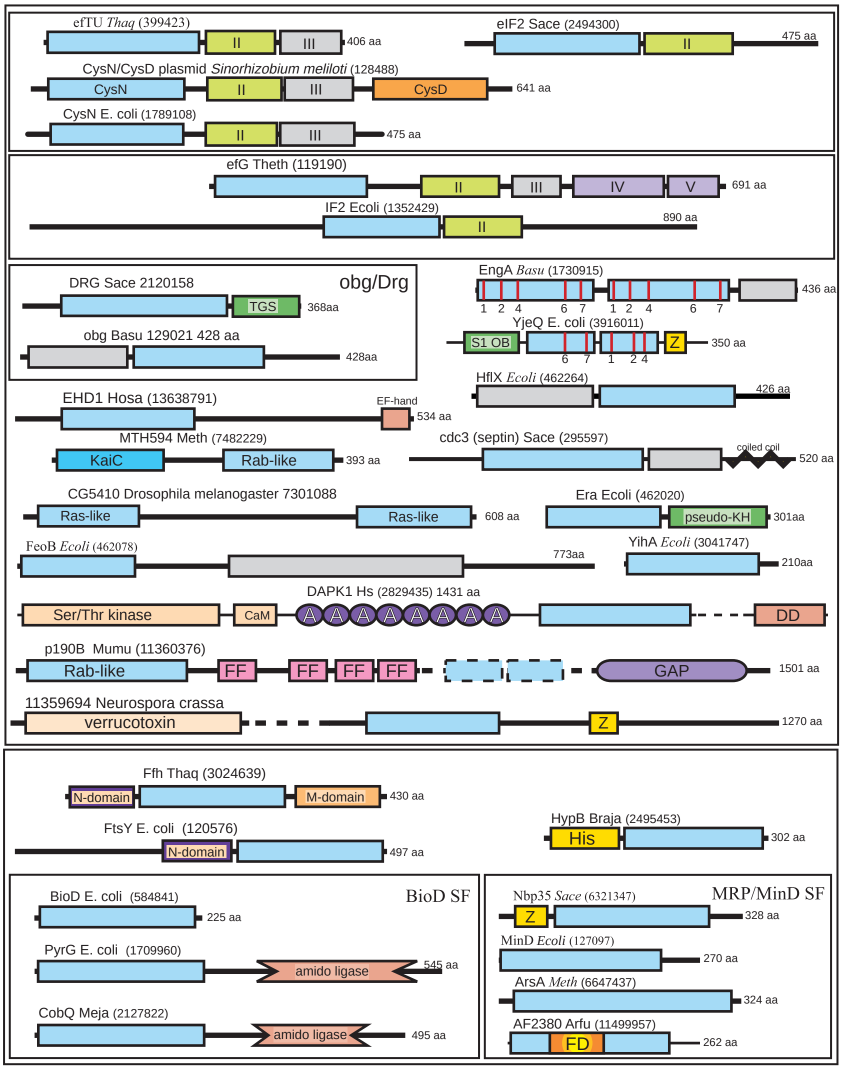

Figure 4. Domain architecture of selected proteins of the GTPase superclass. Proteins are represented as horizontal lines and rectangles or other geometric shapes indicate conserved domains. The GTPase domain is in light blue (enclosed by a broken line when inactivated), known or predicted RNA-binding domains are in green, and the tRNA-mimicry domains of EF-G/EF2 and IF2 are in faint purple. The organism name abbreviation, gene or protein name, and the GenBank identifier (in parentheses) identify proteins. EngA has a tandem repeat of GTPase domains and YjeQ has undergone circular permutation and the NKxD motif has been transferred to the N terminus of the protein; the GTPase motifs are indicated by red lines and labeled: 1, P-loop; 2, conserved threonine residue in strand 2 (G2); 4, Walker B motif; 6, NKxD motif; 7, SAx motif. Domain designations: A, ankyrin repeats; DD, death domain; FD, ferredoxin; FF, FF domain; S1 OB, S1-type OB(oligomer-binding)-fold domain; TGS, tRNA-binding domain present in tyrosyl-tRNA synthetase, GTPases, and SpoT; Z, (predicted) Zn-binding domains. The species name abbreviations are Basu, *Bacillus subtilis*; Braja, *Bradyrhizobium japonicum*; *E. coli*, *Escherichia coli*; Meja, *Methanococcus jannaschii*; Meth, *Methanobacterium thermoautotrophicum*; Mumu, *Mus musculus*; Psde, *Pseudomonas denitrificans*; Sace, *Saccharomyces cerevisiae*; Thaq, *Thermus aquaticus*; Theth, *Thermus thermophilus*; Trbr, *Trypanosoma brucei*.

TrmE/ThdF family

TrmE is ubiquitous in Bacteria and is a widespread mitochondrial protein in Eukaryotes (apparently missing from Drosophila), but is absent from Archaea. The yeast member of this family, MSS1, is involved in mitochondrial translation;[81] bacterial members are often present in translation-related operons.[82] This is consistent with experimental evidence showing that the *E. coli* homolog plays an important role in the incorporation of the modified base 5-methylaminomethyl-2-thiouridine in tRNA.[83] Given the bacterial-eukaryotic pattern of phyletic distribution and the mitochondrial function in eukaryotes, it appears likely that that this gene has been acquired by eukaryotes from the pro-mitochondrial endosymbiont.

The FeoB family

The FeoB family of GTPases is widespread, although not ubiquitous, in Bacteria and Archaea, but missing from Eukaryota. This family represents an unusual exaptation of GTPases for high-affinity

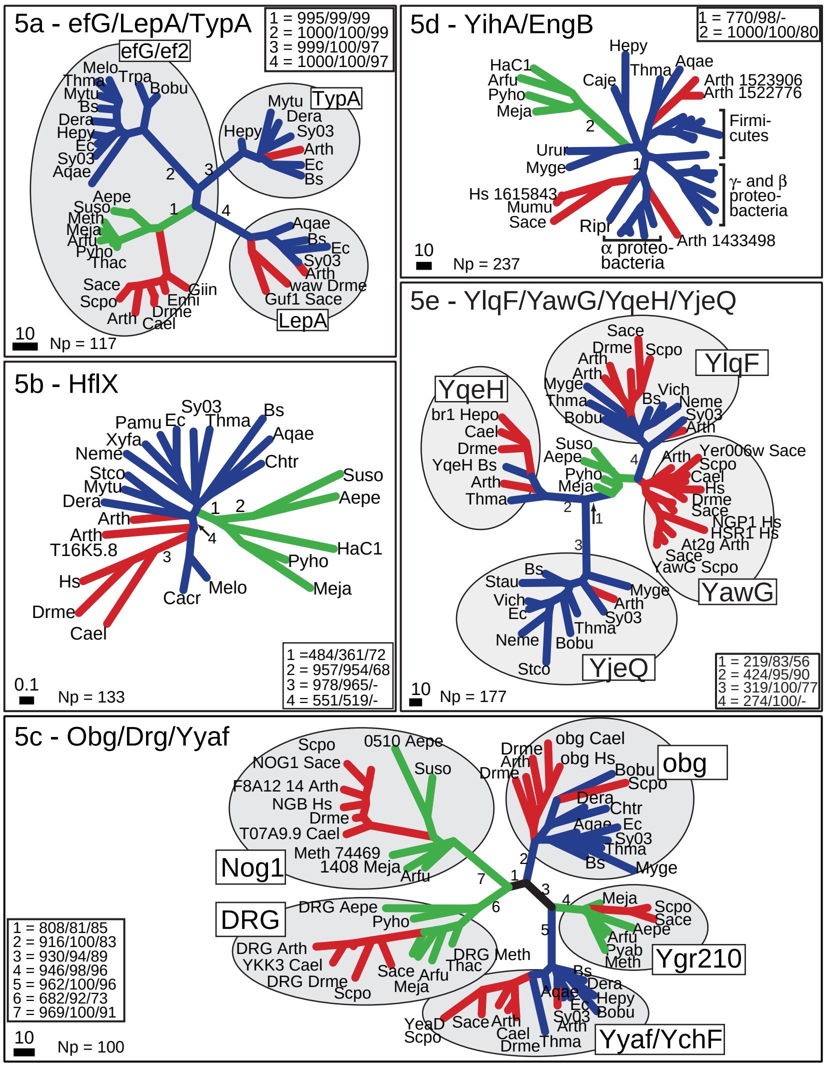

Figure 5 (*legend opposite*)

iron (II) transport.[84] These proteins contain an integral membrane domain, with 12 predicted transmembrane helices, C-terminal to the GTPase domain, and probably function as NTP-dependent pumps, analogously to the ABC ATPases, in the transport process. While the evolutionary connection between FeoB and other families of this GTPase superfamily is beyond doubt, changes in the NKxD motif suggest that the specificity toward GTP might have been lost in some of the members. The pattern of phyletic distribution of the FeoB family suggests an origin within one of the prokaryotic lineages, with subsequent horizontal dissemination. However, an ancient origin for this group, with a subsequent early loss in eukaryotes, cannot be dismissed on the basis of the available information.

The YihA (EngB) family

This family of GTPases is typified by the *E. coli* YihA, which is an essential protein involved in cell division control.[85,86] Cell division defects of *yihA* null mutants have been correlated with decreased levels of the cell division protein FtsZ.[86] YihA and its orthologs are small proteins that typically contain less than 200 amino acid residues and consist of the GTPase domain only (some of the eukaryotic homologs contain an N-terminal extension of about 120 residues that might be involved in organellar targeting). The YihA family is widespread, but not ubiquitous in all three superkingdoms (missing, for example, from the Crenarchaeota, Caenorhabditis, and Drosophila). In phylogenetic analysis, there is no common branch for Archaea and Eukaryota (Figure 5(d)). Instead, most eukaryotic sequences group with the α-proteobacteria (Figure 5(d)), which is indicative of a lateral transfer *via* the mitochondrial route. As in the cases of HflX and PurA (see below), if YihA was present in the LUCA, displacement of the ancestral eukaryotic version of this GTPase by the α-proteobacterial version upon mitochondrial endosymbiosis has to be postulated. However, the available data do not

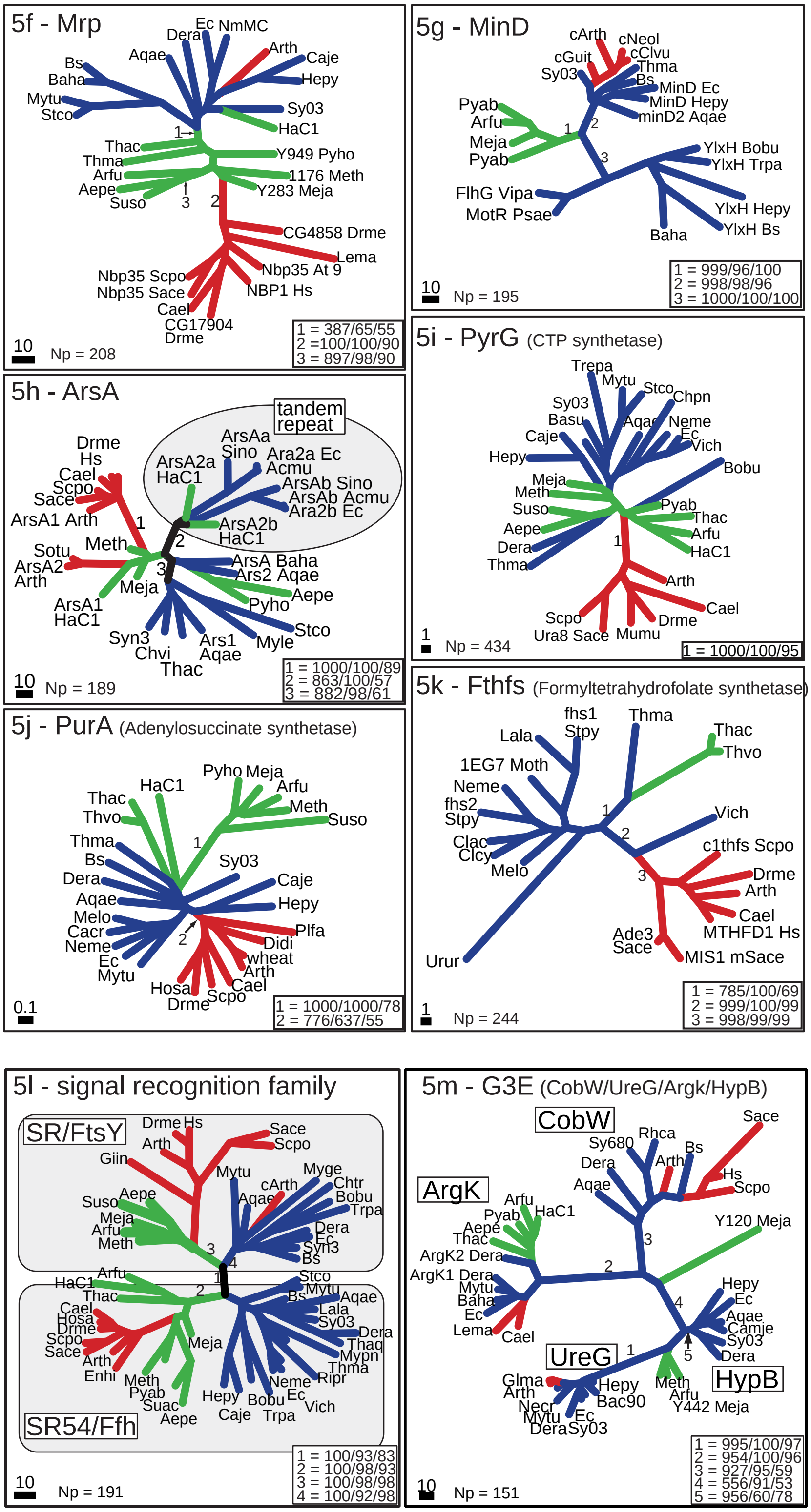

Figure 5. Unrooted phylogenetic trees of selected GTPase families. Tree topologies are from Molphy maximum-likelihood (ML) analysis. The scale bar indicates the number of inferred substitutions per 100 sites (amino acid residues). Support for major branches is indicated by bootstrap probabilities for PHYLIP Protdist/Fitch distance analysis (first number), RELL bootstrap probabilities for ML analysis (second number) and the reliability value computed by the TreePuzzle 5 software (third number). In (b) and (j), the second number is from 1000 neighbor-joining bootstrap replicates. Bacterial branches are in blue, eukaryotic branches in red and archaeal and shared archaeoeukaryotic branches are in green. Branches are black if present in LUCA or if they cannot be attributed unambiguously to either of the three major lineages. Np, number of alignment positions used for tree analysis. Species name abbreviations are as for Figures 2 and 3. Additional abbreviations are: Cacr, *Caulobacter crescentus*; Clac, *Clostridium acidurici*; Clcy, *Clostridium cylindrosporum*; Melo, *Mesorhizobium loti*; Moth, *Morella thermoacetica*; Pamu, *Pasteurella multocida*; Thvo, *Thermoplasma volcaAnium*; Xyfa, *Xylella fastidiosa*.

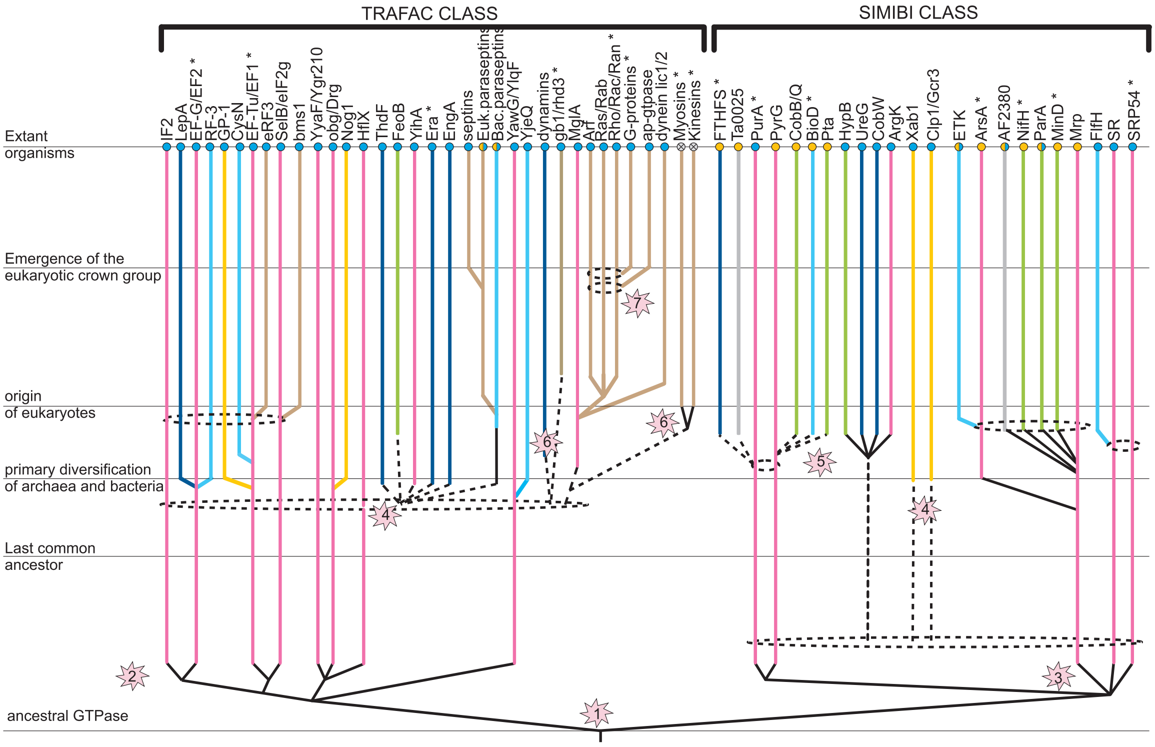

Figure 6. Inferred evolutionary history of GTPase families. The Figure shows several relative temporal epochs separated by the major evolutionary transitions that mark their boundaries. The filled colored bars indicate the maximum depth to which the GTPase lineages can be traced with respect to these temporal epochs. The broken lines indicate uncertainty in terms of the exact point of origin of a lineage. The broken-line ellipses bundle groups of lineages from within which a new lineage potentially could have emerged. The numbered stars indicate various evolutionary events associated with the diversification of the GTPases; namely: 1, separation of GTPases from the rest of the P-loop fold; 2, origin and diversification of the most ancient TRAFAC GTPases associated with translation machinery; 3, origin of the cell membrane and diversification of the SIMIBI class proteins into associated functional niches; 4, bacterial or archaeo-eukaryotic lineage-specific diversification of GTPases into functional niches conserved throughout these lineages; 5, diversification of metabolic enzymes in the SIMIBI class; 6, provenance of the eukaryotic cytoskeleton and intracellular transport apparatus; 7, diversification of the eukaryotic signal transduction apparatus. Color key: red/pink, universal; dark blue, Bacteria and Eukaryota; light blue, Bacteria; brown, Eukaryota; grey, Archaea and Thermotoga; green, Archaea and Bacteria; orange, Eukaryota and Archaea. An asterisk (*) identifies families with known 3D structures. Blue circles: families with conserved aspartate in the [NT]KxD motif that are known or predicted to possess specificity for GTP; yellow circles, families in which the [NT]KxD motif is absent or the aspartate residue not conserved and which are known or predicted not to have GTP specificity; blue-yellow circles, the [NT]KxD motif retained in some family members but not others; circle marked with a cross, the C-terminal part of the protein, which contains the [NT]KxD motif in other GTPases, is missing altogether.

rule out an alternative scenario whereby YihA evolved in bacteria and has been acquired independently by Eukaryota and Archaea (or at least the Euryarchaeota).

The Era family

The Era family is characterized by a distinct derivative of the KH domain (the pseudo-KH domain) which is located C-terminal to the GTPase domain (Figure 4).[31,87,88] Era is ubiquitous in Bacteria and found also in Metazoa and plants, but is missing thus far from fungi. Among archaea, only Methanococcus and Pyrococcus encode Era-like GTPases that might have been acquired from thermophilic bacteria and have lost the C-terminal extension. Given this phyletic distribution, it seems likely that the eukaryotic members of the Era family have been acquired from bacteria *via* horizontal transfer, apparently with subsequent loss in (some of the) fungi. Era is likely to be a translation factor, whose association with 16 S RNA, probably *via* the pseudo-KH domain, stimulates the GTPase activity.[89–91]

The EngA/YfgK family

The EngA family is named after essential neisserial GTPase A encoded by a gene transcriptionally linked in an operon with RdgC, which is thought to be involved in pilin antigenic variation in *Neisseria gonorrhoeae*.[92] EngA and its orthologs are composed of two GTPase domains (Figure 4) and, since the sequences of the two domains are more similar to each other than to other GTPases, it is likely that an ancient gene duplication, rather than a fusion of evolutionarily distinct GTPases, gave rise to this family. EngA is ubiquitous in Bacteria and is present also in Arabidopsis, but not in other Eukaryota or Archaea; thus, this family probably is of bacterial origin, with the plant member(s) acquired horizontally from the pro-chloroplast symbiont. The EngA family remains functionally uncharacterized, but its conservation in all Bacteria, including parasites with small genomes (e.g. Chlamydia, spirochaetes and Mycoplasma), that are not represented in many other protein families is suggestive of an essential function, such as translation regulation.

The Septin-like family: septins, paraseptins and related GTPases

Septins were first identified as GTPases participating in the late stages of the formation of the septum between dividing cells during mitosis and meiosis in animals and fungi.[93] Subsequent studies have shown that septins are involved also in the regulation of other cellular processes, such as vesicular fusion and tethering of vesicles to the exocyst complex.[94] Some septins bind phosphatidyl-inositol 4,5-bisphosphate, which precludes GTP-binding and thereby regulates the association of septins with cellular membranes.[95] Septins are so far limited to eukaryotes in their distribution; a large expansion of the septin subfamily is seen in animals and fungi, but so far not in any other eukaryote. This is compatible with the existence of an animal/fungi clade;[96] at least three distinct groups within the septin subfamily appear to have evolved in the common ancestor of this clade, followed by further duplications after the divergence of the animal and fungal lineages.

With a wider, if patchy, presence throughout the eukaryotic kingdom and some prokaryotic lineages, paraseptins and other septin-related proteins form the remaining subfamilies of this GTPase family. The paraseptins include a diverse assembly of proteins; one of the conserved groups among these shows a sporadic distribution and is typified by an uncharacterized protein from *Entamoeba histolytica* and its homologs from Neurospora (fused to a verrucotoxin domain and Zn-binding domains; Figure 4), and C18B2.5 and F43B10.2 (inactive) from *Caenorhabditis elegans*. Another conserved paraseptin branch is the Aig1-Toc34/Toc159 group, with at least two conserved subgroups; namely, the Aig1-like proteins represented in plants and vertebrates, and Toc34/Toc159-like proteins found in Dictyostelium (GtpA) and plants (Figure 5(f)). The former subgroup is involved in the regulation of anti-pathogen defense in both vertebrates and plants,[97,98] whereas the plant members of the latter subgroup are involved in the import of peptides into chloroplasts.[98,99] Bacteria encode their own group of septin-related proteins, typified by VC1806 from Vibrio, NMA0132 (Neisseria), HP0744 (Helicobacter), and slr1428 (Synechocystis). Septins and paraseptins show a distinct deviation in the NKXD motif (Figure 2). An even more distant set of relatives of the Septin-paraseptin subfamilies are the three lineage-specific *E. coli* GTPases, YeeP, YkfA, and YfjP, which probably have diverged from bacterial paraseptins.

This family of GTPases is more similar in sequence to the predominantly bacterial GTPase families, such as the Era, FeoB, TrmE, YihA, and EngA families, than to other TRAFAC class members. The sporadic presence of the septin-paraseptin family is in contrast to the wider distribution of these other GTPase families in bacteria. It seems likely that the septin-paraseptin family originally evolved in Bacteria, from within the bacterial GTPases of the TrmE-Era-EngA-like families, followed by horizontal transfer from the pro-mitochondrial endosymbiont to eukaryotes, where they were gave rise to the septins, which assumed essential functions in certain eukaryotic lineages. Extensive, lineage-specific gene loss probably additionally contributed to the patchy distribution of these proteins in eukaryotes.

The Ras-like superfamily

The large, extended Ras-like family and several smaller families of more distantly related GTPases form a higher-order assemblage that is detected readily by similarity-based clustering. The Ras-like superfamily is a predominantly eukaryotic offshoot of the TRAFAC class.

The extended Ras-like family(Ras/Rab, Rac/Rho/cdc42, trimeric G proteins/Arf/Sar1, MglA)

This family is found in all three superkingdoms, but shows a particularly spectacular expansion in eukaryotes. The most distinctive aspect of the distribution of Ras-like proteins in prokaryotes is that they are encoded most frequently in an operon with a small MglB protein which, in *Myxococcus xanthus,* is involved in gliding motility[100] and is a homolog of the eukaryotic dynein light chain 7 (roadblock).[101] This suggests that the prokaryotic small Ras-like GTPases (MglA proteins) co-evolved with the MglB family proteins and might function together with them in a GTP-dependent motor (that, however, may not be involved exclusively in gliding motility because mglAB operons are present also in bacteria that do not glide). In eukaryotes, the role of the GTPase cofactor of the roadblock-family proteins probably belongs to the dynein light/intermediate chain 1/2, which is a GTPase distantly related to the extended Ras-like family (see below). The MglA subfamily of prokaryotic GTPases appears to have been extensively disseminated horizontally along with the MglB gene.[101]

Eukaryotes show an extraordinary diversity of the Ras-like family, with at least three distinct branches, the Ras/Rab subfamily, the Rac/Rho subfamily, and the Arf subfamily, which emerged early in the evolution of eukaryotes. The heterotrimeric GTPases so far have been found only in the crown-group eukaryotes, and probably evolved prior to the divergence of the crown group, through insertion of an α-helical bundle into the GTPase domain of one of the ancient eukaryotic subfamilies. An important feature of the extended Ras-like family in eukaryotes is their association with a wide variety of GAPs and GEFs. These diverse GAPs and GEFs typically are multidomain proteins that enable recruitment of the GTPases in various functional contexts and represent one of the important, unique features of the eukaryotes. Additionally, some GTPase domains of the extended Ras-family have been incorporated into multidomain proteins. For example, the RhoGap p190 has an N-terminal active GTPase domain separated by multiple FF domains from two inactive Ras-like GTPase domains (Figure 4). The CG5410 group of proteins conserved throughout the eukaryotes contains a head-to-tail fusion of two GTPase domains of the Rac/Rho subfamily.

Distant relatives of the extended RAS-like GTPase family

At least two small families of GTPases are distantly related to the extended Ras-like GTPase family rather than to any of the other GTPase families. They are so far restricted to eukaryotes and thus may be fast-evolving members of the Ras-like family whose sister groups could not be identified because of the rapid divergence. One of these families is the Ap-GTPase, so named after the apoptotic DAP kinase, in which they were originally detected as distinct domains.[102] So far, Ap-GTPases have been detected only in plants and animals, and they occur always in multi-domain proteins associated with repetitive superstructure-forming domains such as leucine-rich repeats and ankyrin repeats (Figure 4). These GTPases are probably involved in GTP-dependent assembly of complexes through the repeat units.

The dynein light intermediate chains 1/2 comprise the other family of eukaryote-specific GTPases that are distantly related to the Ras-like family. They are regulatory subunits of the dynein motor and, given the presence of roadblock/MglB family proteins in the dynein complex, perhaps function together with them. However, these proteins show no specific relationship with the MglA-like GTPases or any other subfamily of the Ras-like family.

The myosin-kinesin superfamily

The two major families of eukaryotic cellular motor ATPases, kinesin and myosin, constitute the myosin-kinesin superfamily within the TRAFAC class of GTPases (the third eukaryotic motor ATPase, dynein, belongs to the AAA+ class of P-loop NTPases[11,103]). They mediate ATP-dependent movement of chromosomes, vesicles, and organelles along tubulin-microtubules in the case of the kinesins, and along actin filaments in the case of myosins.[104,105] Consistent with the loss of specificity towards GTP, kinesin and myosin lack the NKxD motif. Specifically, these proteins share a deletion of strands 6 and 7 of the ancestral core of the TRAFAC GTPase (Figure 1), resulting in the loss of the NKxD motif. However, their core sheet is extended to a configuration similar to that in the regular GTPase domains by the addition of two strands at the N terminus that take the place of the lost strands. These strands might have arisen through a circular permutation similar to that seen in the YlqF/YawG family (see below). In 3D structure comparisons, the myosin-kinesin superfamily shows closest similarity to the classic TRAFAC GTPases, such as Ras; more specifically, the Walker B strand is flanked by an anti-parallel strand with a conserved N-terminal alcoholic residue (serine) shared with all other TRAFAC GTPases (Figures 1 and 2). This synapomorphy establishes the membership of myosin and kinesin in the TRAFAC class.

The emergence of the motor ATPases was critical for the genesis of the organizational complexity of the eukaryotes. Myosin and kinesin families probably evolved, at the onset of eukaryotic evolution, from an ancestral protein that might have already had a motor function.[40,106] This hypothetical, ancestral motor ATPase apparently arose as an offshoot of one of the more ancient TRAFAC GTPase families, probably of the Ras-like superfamily. This process should have involved deletion of C-terminal strands and emergence of N-terminal ones, or a circular permutation. Given the association of MglA-like GTPases of the extended Ras-like family with bacterial motility (see above), the common ancestor of myosins and kinesins might have been derived from a prokaryotic Ras-like GTPase that already functioned as a motor or a regulator of motility. However, the sequences of the motor ATPases have diverged to such an extent that tracing their origin to a specific prokaryotic subfamily appears impossible. The subsequent divergence of myosins and kinesins probably reflects differential adaptation to their distinct cytoskeletal microfilament partners, those comprised of actin and tubulin, respectively.

The YawG/YlqF family

The YlqF/YawG family consists of five distinct sub-families, typified by the proteins YlqF, YqeH (both from *B. subtilis*), YjeQ (*E. coli*), MJ1464 (from *Methanococcus jannaschii*), and *Schizosaccharomyces pombe* YawG, respectively (Figure 5(g)). All these proteins show a circular permutation of the GTPase signature motifs so that the C-terminal strands 5, 6, and 7 (strand 6 contains the NKxD motif) are relocated to the N terminus (Figures 1, 2 and 4). One human YawG homolog, Ngp-1, has been shown to localize in the nucleolus and nucleolar organizers in all cell types analyzed,[107] which is indicative of a function in ribosomal assembly. Several members of this family contain the S1-OB fold RNA-binding domain[108] N-terminal to the GTPase domain, which supports such a function and/or a function in translation.

The YlqF subfamily is represented in a phylogenetically diverse array of Bacteria (including Gram-positive bacteria, two proteobacteria, Synechocystis, Borrelia, and Thermotoga) and in all eukaryotes. Apparently, genes coding for proteins of this subfamily have been transferred from bacteria to eukaryotes on two independent occasions, once at an early stage of eukaryotic evolution, probably from the proto-mitochondrion, and, for the second time, from chloroplasts to plants (Figure 5(e)). Eukaryotic YawG and its archaeal homologs share a C-terminal domain (the YlqF-C domain) with the YlqF subfamily (Figure 4). YqeH and YjeQ are more divergent, smaller subfamilies; the first of these is common in eukaryotes and sporadically present in bacteria, whereas the second one is predominantly bacterial, with probable acquisition by plants from chloroplasts (Figure 5(e)). In phylogenetic analysis, the location at which YqeH/YjeQ attaches to the rest of the tree is unstable: the attachment can be either within the Archaea (as shown in Figure 5e), on the branch connecting Archaea and Eukaryota, or as an unresolved trifurcation with YlqF and archaeoeukaryotic YawG. We hypothesize that the protein was already present in LUCA and that YqeH/YjeQ originated from a duplication in the Bacteria and diverged so rapidly that a common branch for the bacterial lineages (YqeH/YjeQ + YlqF) cannot be recovered reliably in phylogenetic analysis. However, given the lack of support for a distinct archaeo-eukaryotic branch (Figure 5(e)) and the absence of YawG in the completed archaeal genomes of Halobacterium, Archaeoglobus, and Methanothermobacter, it cannot be ruled out that the family originated in the Bacteria and was subsequently disseminated *via* horizontal transfer.

The circular permutation appears to have occurred in the common ancestor of this family and might have been preceded by a duplication of the GTPase domain. EngA is the only widespread TRAFAC GTPase family that contains a tandem repeat of the GTPase domain (see above). However, since EngA is largely limited to Bacteria and, by sequence comparison, is not closer to Ylqf/YawG family than to other TRAFAC GTPases, it is probably an independent duplication, rather than a derivative of the duplicated ancestor of the more widespread YlqF/YawG family. The circular permutation in the YlqF/YawG family is unusual from the point of view of the structure of the P-loop NTPase domain. Typically, circular permutation through duplication, with subsequent elimination of the termini, occurs in domains whose N and C termini are juxtaposed in space.[109,110] In many P-loop NTPases, the N and C termini are not in close proximity,. and circular permutation have not been encountered in any other member of this vast fold (Figure 1). However, structures of certain GTPases, such as Era show that the C-terminal-most helix of the GTPase can occur in close spatial proximity to the N-terminal (Walker A-associated) strand. It appears likely that such an initial conformation allowed the emergence of a permutation in the YlqF/YawG family without disrupting the integrity of the GTPase domain and its active site.

The Dynamin/YjdA/Fzo/YbpR family

Dynamins are involved in budding of clathrin-coated vesicles from the plasma membrane,[111] whereas their close relatives, Dnm1 and DRP-1 (from yeast and *C. elegans*, respectively), participate in mitochondrial division, presumably by severing the mitochondrial outer membrane.[28,112,113] Other close relatives of dynamins, the Mx proteins, are interferon-induced GTPases that inhibit multiplication of certain negative-strand RNA viruses in vertebrates.[114] These proteins form the dynamin-like subfamily that appears to regulate the dynamics of a variety of membrane structures.[115–117,118] The dynamin-like subfamily is widely represented throughout the part of the eukaryotic tree sampled thus far and probably emerged at a early stage during evolution of eukaryotes. The subfamily most closely related to the dynamins includes bacterial YjdA proteins, which are found sporadically in phylogenetically distant bacterial lineages. It seems likely that they are related to the eukaryotic dynamin-like subfamily *via* an early horizontal transfer between bacteria and eukaryotes. Fuzzy onions (Fzo) proteins form another specific, conserved group of transmembrane GTPases, thus far restricted to fungi and metazoa, which function in mitochondrial fusion.[27,119,120] The closest relatives of the Fzo subfamily belong to the bacterial YbpR (*B. subtilis*) subfamily found in several bacterial lineages, including Gram-positive bacteria, helicobacteria, cyanobacteria, and actinomycetes. Given the broad distribution of this subfamily in bacteria and the mitochondrial function of the eukaryotic members of this subfamily, it appears likely that the Fzo subfamily emerged through extensive divergence from a YbpR-like ancestor, present in the proto-mitochondrion. Eukaryotic sarcalumenins, which bind calcium in the endoplasmic reticulum, are found in plants and animals, and belong to the YbpR subfamily; probably, these GTPases have been transferred independently into the eukaryotic crown-group lineage from the Thermus-Deinococcus lineage of bacteria, with subsequent fusion of EF-HAND domains (EHD1 in Figure 4).

The GB1/RHD3 family

This family is characterized by a distinct motif, hhhRD, which is a derivative of the NKxD motif (Figure 2). Structural superposition shows that the N-glycosidic bond angle (between base and ribose) in the GTP-hGBP1 complex is different from that found in Ras-like proteins or EF-Tu, but the conserved aspartate residue makes a double hydrogen bond to the endocyclic and exocylic nitrogen atoms of guanine and thus provides the specificity towards GTP, just like the aspartate residue in the canonical guanosine-binding [NT]KxD motif.[121] This family is widespread in eukaryotes, but not detectable in bacteria or archaea. One conserved subfamily is typified by the Arabidopsis protein root hair defective 3 (RHD3),[122] whose orthologs are present in all crown-group eukaryotes. The other subfamily is typified by the interferon γ-induced antiviral GB1 protein that is conserved in animals. The other GTPases of this subfamily are the brain finger proteins (BFPs),[123] in which the GTPase domain is combined with an N-terminal RING finger domain, which implicates these proteins in ubiquitin-mediated signaling. Most members of this family have a large C-terminal, α-helical extension that probably participates in protein-protein interactions.

The SIMIBI class

The second class of GTPases and related NTPases includes several large and small families, a subset of which form two broader assemblages, the MinD/Mrp and the BioD-related superfamilies. All these proteins share structural and sequence similarities that clearly distinguish them from the TRAFAC GTPases. The distinction between a large subset of our TRAFAC class and the SRP GTPases was realized previously,[9,43,44] but the specific relationship between the SRP GTPases, the MinD/Mrp and BioD superfamilies and the other SIMIBI families, to our knowledge, has never been described. Typically, in structural similarity searches with DALI and VAST, the structurally characterized representatives of the SIMIBI class show higher scores with each other than with any of the TRAFAC GTPases. For example, in a VAST search with the structure of dethiobiotin synthetase (1BYI), the best hit was the signal sequence recognition protein Ffh from *Thermus aquaticus* with a VAST *P*-value of 10^{-9}. Conversely, in a VAST search with the *T. aquaticus* Ffh GTPase domain (NG1), dethiobiotin synthetase and nitrogenase iron protein from *Clostridium pasteurianum* (1CP2) score significantly higher than the structures of p21ras or translation factors. Closer inspection reveals a clear separation of these two classes on the basis of the design of the part of the β-sheet that is formed by the region N-terminal of the Walker B strand and flanks it in the 3D structure. This region forms two parallel β-strands flanking the Walker B strand in the SIMIBI proteins. In contrast, in the TRAFAC GTPases, the Walker B strand is flanked by a strand oriented anti-parallel to the rest of the sheet (Figure 1).

This division between the TRAFAC class and the SIMIBI class is apparent also at the sequence level. For example, in iterative searches initiated with SIMIBI proteins of the Clp1/Grc3 family (see below), SRP GTPases, Mrp, and BioD score consistently higher and are recovered earlier than TRAFAC class members. In addition to general sequence similarity, three distinct sequence synapomorphies for the SIMIBI clade can be established. First, there is a conserved aspartate residue at the C terminus of strand 2 (Figure 3) that has been shown to make a direct or water-mediated hydrogen bond to the bound Mg ion, e.g. in *Acidianus ambivalens* Ffh (a SRP superfamily member), and in Mrp/MinD superfamily proteins such as in *Pyrococcus furiosus* MinD, the arsenite pump ArsA and the nitrogenase iron protein NifH.[35,124–126] Second, the SIMIBI class proteins typically have a specific variation of the Walker A motif that includes a third conserved glycine residue (consensus GxxGxGK[ST]) that is found rarely in the TRAFAC GTPases (Figure 3). Third, the SIMIBI class proteins contain a conserved aspartate residue in the beginning of strand 4 (Walker B). The function of this residue is not known, but the location at the N-terminal side of the β-sheet (far removed from the nucleotide-binding site) suggests that it is required for structural stabilization of the strand.

Furthermore, the [NT]KxD motif that provides specificity for guanine is well conserved in most TRAFAC GTPases, but shows a great deal of sequence variation in the SIMIBI NTPases, many of which, accordingly, show no specificity for GTP (Figures 2 and 3). Similarly, there is strong conservation of the hhhhDxxG Walker B motif in the TRAFAC GTPases whereas, in the SIMIBI class, the aspartate residue is often substituted by glutamate; there is also length variation between the conserved acidic residue and the glycine residue of the Walker B motif in the SIMIBI class that is not found in the TRAFAC GTPases (Figures 2 and 3). In addition, SIMIBI NTPases tend to form dimers (homodimers, e.g. NifH, ArsA, BioD, and PurA, or heterodimers as in the case of the signal recognition particle and its receptor) as opposed to translation factors, Ras or other TRAFAC GTPase that typically function as monomers or in larger complexes. We discuss below the various distinct superfamilies and families of the Simibi class.

The MinD/Mrp-Etk-superfamily

The large MinD/MRP family and the small bacterial ETK family are unified into a superfamily on the basis of sequence similarity-based clustering. Mrp and its close relatives appear to form the ancient core of this superfamily, with the other branches emerging as diversifications of this core following ancient or recent duplications.

The Mrp/MinD family

The Mrp/MinD family is characterized by the synapomorphic KGG signature in the Walker A motif, which, in this family, has the consensus GKGGhGK[ST].[127] In the structure of the nitrogenase iron dimer, the conserved lysine residue in the KGG motif of the second monomer interacts across the interface with the terminal oxygen atom of the β-phosphate group of ATP.[125] The conservation of the lysine residue in the KGG signature suggests that all proteins of this family might function as dimers. Only the asparagine residue of the [NT]KxD motif is conserved in the Mrp/MinD family, with the exception of MinD and ChlL, in which this asparagine residue is followed by a positively charged residue (Figure 3). The conserved aspartate residue of the GTPases that provides specificity for guanine binding is almost always missing and, indeed, none of these proteins has been shown to have GTP specificity. The Mrp/MinD family consists of nine ancient conserved subfamilies.

The *Mrp/NBP35* subfamily is widespread in all three superkingdoms (Figure 5(f)) and is likely to perform an essential, conserved function. The close relationship between this subfamily and the well-characterized MinD subfamily (see below) suggests a role in chromosomal partitioning or related functions. This is supported by findings that yeast NBP35 is essential and localizes predominantly to the nucleus.[128,129] A distinct sequence feature of this family is a methionine residue following the asparagine residue corresponding to the NKxD motif (Figure 3). Eukaryotes typically encode two or more paralogs of the Mrp group, one of which contains a metal-binding domain with four conserved cysteine residues N-terminal to the NTPase domain (Figure 4).[128,130] Mrp is ubiquitous in Archaea and Eukaryota, and widespread in Bacteria (but missing from spirochaetes, Ureaplasma, Mycoplasma, and Chlamydia). The phylogenetic tree for the Mrp family is compatible with the standard model of vertical inheritance and, together with the near-ubiquity of this family, indicates that LUCA had a Mrp protein (Figure 5(f)). Additionally, a few sporadic horizontal transfers from bacteria to Halobacterium and the plant lineage, appear to have occurred (Figure 5(f)).

The *AF2380* subfamily is a small group of functionally uncharacterized proteins related to the MRP NTPases and is found only in Euryarchaeota and Thermotoga. This group is characterized by insertion of a ferredoxin domain into the SIMIBI NTPase domain just prior to strand 3 (Figure 4).

The MinD subfamily is typified by the membrane ATPase MinD, which is required for the correct placement of the division site in *E. coli*.[131] MinD differs from Mrp in having an arginine residue following the asparagine residue in the equivalent of the NKxD motif (Figure 3). MinD is nearly ubiquitous in Bacteria and widespread in Archaea, but missing from Crenarchaeota (Figure 5(g)). MinD is absent from nuclear eukaryotic genomes, but is present in chloroplasts of green plants, algae, and cryptomonads.[132,133] This phyletic distribution suggests that MinD is of bacterial origin and was transferred laterally to Euryarchaeota. MotR/FlhG is a small sub-group, with orthologs so far found only in spirochaetes, proteobacteria and Bacillus (Figure 5(g)), which probably was derived from a MinD duplication in bacteria. MotR is thought to have a regulatory role in bacterial motility, possibly by controlling the number and placement of flagella.[134]

The ParA/Soj subfamily is typified by the ParA ATPase, which, together with ParB, is involved in partitioning of newly replicated chromosomes and low-copy number plasmids.[135–138] The exact molecular role of ParA is not known, but it has been identified as a transcriptional repressor of Spo0A and other sporulation genes in *B. subtilis*.[138,139] The ParA family tree consists of a core of conserved sequences and a ‘‘halo’’ of fast-evolving sequences that are mostly plasmid-encoded (e.g. pRK2 in *E. coli*, and pNRC100 in Halobacterium, tree not shown). ParA is widespread in Bacteria (but missing from Aquifex, *E. coli*, Synechocystis, *Thermotoga maritima*, and Mycoplasma) and Archaea, but missing from Eukaryota. Given the common presence of genes coding for these proteins on plasmids, a bacterial origin of this family, with subsequent multiple horizontal exchanges between Bacteria and Archaea, appears likely.

NifH is a component of the nitrogenase complex, which fixes nitrogen into ammonia in an ATP-dependent process.[125] The NifH subfamily is found in a variety of nitrogen-assimilating bacteria, such as Cyanobacteria, Clostridium, and Rhizobium, and some Archaea, such as Methanobacterium and Methanococcus (tree not shown). The ChlL/FrxC subfamily is closely related to NifH and is hypothesized to have a role in light-independent conversion of protochlorophyllide to chlorophyllide.[140] Proteins of this family are present only in Cyanobacteria and chloroplasts.

The ArsA subfamily is represented by the arsenite-translocating ATPase ArsA, which is part of a multi-subunit pump that catalyses the extrusion of oxyanions, such as arsenite, antimonite and arsenate, from the cell.[141] ArsA contains either a single SIMIBI NTPase domain or a tandem repeat associated with membrane spanning segments (Figure 5(h)). ArsA is widespread in Eukaryota and Archaea, but shows sporadic distribution in Bacteria (present in Aquifex, *E. coli*, Mycobacterium and Synechocystis, but missing from the great majority of bacterial genomes sequenced so far). In phylogenetic trees, the proteobacterial sequences do not cluster with the other bacteria, but with the ArsA homologs from the archaeon *Halobacterium* sp. NRC-1 (Figure 5(h)) with which they share a tandem duplication of the NTPase domain. The ArsA homologs from *E. coli*, *Acidiphilium multivorum*, and Halobacterium NRC-1 are encoded on plasmids.[142–144] Plants have two ArsA homologs, one of which is of the eukaryotic type, whereas the other probably is the result of a recent transfer from Archaea (Figure 5(h)). Given the topology of the tree, the two types of domain organization and the poor representation in bacteria, ArsA was probably not represented in LUCA, but is of archaeal or archaeoeukaryotic provenance, and has spread to bacteria, on perhaps more than one occasion, owing to specific environmental selective advantages conferred by its possession. This is compatible with a comparative analysis of the bacterial *ars* operon (*arsRDABC*) that suggests that ArsA is a recent addition to the bacterial gene repertoire.[145]

The Etk family

This family is typified by the bacterial proteins Wzc and Etk, which have been shown to function as tyrosine kinases regulating the synthesis of extracellular polysaccharide structures.[146,147] The genes coding for these kinases typically occur in an operon with the low molecular mass phosphatase Wzb, indicating that the two enzymes function as an antagonistic pair in regulating the phosphorylation state of the polysaccharide synthesis/assembly apparatus.[148,149] Many proteobacterial members of this family have an N-terminal, membrane-spanning segment, whereas others occur as stand-alone NTPase domains, with the membrane-spanning part encoded by a separate gene. These proteins have a Walker B motif with a DTPP signature. Some members of this family have the canonical NKxD motif, but in most of them this motif is partially eroded, although the asparagine residue is conserved (Figure 3). The ETK-like proteins occur sporadically in bacteria, such as proteobacteria and some Gram-positive bacteria. This family is most closely related in sequence to the Mrp-MinD family members and, given its very limited phyletic horizon, probably emerged, relatively recently, from within the Mrp-MinD family through extensive divergence.

The BioD-FTHFS superfamily

The BioD (dethiobiotin synthetase) and FTHFS (formyltetrahydrofolate synthetase) families, along with some smaller families, such as adenylosuccinate synthetase and Ta0025, that are more distantly, but specifically related to the BioD family, constitute this large assemblage. A distinct feature of this superfamily is the presence of a glutamate residue in the Walker B motif in place of the aspartate residue found in most P-loop NTPases (Figure 3). Several families of this superfamily also possess a non-canonical form of the Walker A motif. Biochemically, the common feature of these proteins is that they typically catalyze a kinase-like reaction in various metabolic contexts.

The BioD family

This family has at least two distinct synapomorphies: (i) the two conserved glycine residues of the Walker A motif are spaced five or six residues apart rather than four as in the vast majority of P-loop NTPases; and (ii) the second strand contains a conserved lysine residue rather than the usual aspartate residue of the SIMIBI class (Figure 3); this lysine residue has been implicated in substrate binding.[150] Several ancient conserved subfamilies can be delineated within this family. BioD catalyzes the penultimate step in the biosynthesis of biotin, the formation of dethiobiotin from diaminopelargonic acid, CO_2 and ATP.[151,152] The NTPase domain of BioD essentially functions as a kinase in this reaction. The dethiobiotin synthetase subfamily (BioD proper) is widespread in Bacteria and present in *M. jannaschii* and *S. cerevisiae*, but not in any other archaea or eukaryotes studied so far (tree not shown).

PyrG, CobB, and CobQ proteins form subfamilies characterized by the fusion of an N-terminal SIMIBI NTPase domain with a C-terminal amidoligase domain (Figure 4).[153] PyrG (CTP synthetase) catalyzes ATP-dependent amidation of dUTP to form dCTP,[154,155] whereas CobB (cobyrinic a,c-diamide synthetase), and CobQ (cobyric acid synthetase) catalyze the amidation of cobalamin precursors.[156] PyrG has a variant of the Walker A motif, in which the conserved serine/threonine in the GK[ST] signature is replaced by glycine (Figure 3). While CobB and CobQ are limited in their distribution to Euryarchaeota and a few Bacteria, PyrG is nearly ubiquitous in Bacteria, Archaea, and Eukaryota (missing only from some Mycoplasma species). Phylogenetic trees support only the monophyly of the eukaryotic PyrG, whereas Archaea and the major bacterial lineages form an unresolved polytomy (Figure 5(i)). Thus, on the basis of its phyletic pattern, PyrG was probably present in LUCA, but the lack of clear phylogenetic signal could mean that its present distribution is more a result of multiple subsequent horizontal transfers.

Another subfamily of the BioD family includes phosphate acetyltransferases, typified by *E. coli* Pta. This enzyme catalyzes the synthesis of acetyl phosphate from inorganic phosphate and acetyl-CoA. Most PTAs have a non-canonical Walker A motif, with the K of the GK[ST] signature replaced by a hydrophobic residue (Figure 3). PTA is sporadic in Bacteria (proteobacteria, Synechocystis and Deinococcus), but is present in the archaea Archaeoglobus and Halobacterium, a distribution that suggests a history of horizontal transfers, and possibly lineage-specific losses.

Ancient conserved families distantly related to the BioD family

The adenylosuccinate synthetase (PurA) family

Despite atypical Walker A and Walker B motifs, adenylosuccinate synthetase (PurA) has been recognized as a P-loop GTPase *via* structural comparisons.[43] Several features suggest a specific distant relationship of PurA to the BioD family within the SIMIBI class. PurA catalyzes a kinase-like reaction similar to the BioD proteins and shares with them the D → E substitution in the Walker B motif, and a basic residue (in this case, arginine) at the end of strand 2 (Figure 3). However, PurA differs from the regular BioD family members in having a more canonical TKXD signature associated with strand 6, and a prominent deletion in the loop between helix 1 and strand 2 (Figure 3). PurA catalyzes the formation of adenylosuccinate from IMP and aspartate in the *de novo* biosynthesis of AMP.[157] This enzyme is nearly ubiquitous in Bacteria, Eukaryota and Archaea (Figure 5(j)) but missing in *Aeropyrum pernix* along with the rest of the *de novo* purine biosynthesis enzymes.[158] While the wide distribution of the enzyme is compatible with an origin in LUCA, the topology is suspect because Archaea and Eukaryota do not form a clade (Figure 5j). As in the case of HflX and YihA, placing the origin of this family in the LUCA would imply that the current eukaryotic PurA gene is of bacterial origin and that the original eukaryotic PurA ortholog has been lost. However, a bacterial origin of PurA, with independent acquisition by archaea and eukaryotes through lateral transfer, cannot be discounted on the basis of the available data.

The Ta0025 family

This small family consists of orthologous proteins present in the archaea Thermoplasma, Methanobacterium, and Halobacterium, and the hyperthermophilic bacterium Thermotoga. These proteins are characterized by the five-residue distance between the conserved glycine residues in the P-loop and the D → E substitution in the Walker B motif shared with the BioD and G3E families (a EGx[GAS] instead of the usual DxxG; Figure 3). Arginine or methionine replaces the alcoholic residue in the P-loop, and only the asparagine residue of the [NT]KxD motif is conserved (Figure 3). These features and the limited phyletic distribution suggest that this group was derived from within the BioD family. MTH1837 from Methanothermobacter is fused to a Rossmann-fold domain similar to those found in the NAD-dependent carbohydrate epimerase/dehydratase family. This might suggest a role for the Ta0025 family NTPases in an unknown pathway of carbohydrate metabolism that is specific for thermophiles.

The FTHFS family

FTHFS catalyzes the synthesis of formyltetrahydrofolate *via* ATP-dependent condensation of formate with tetrahydrofolate.[159,160] FTHFS consists of an N-terminal SIMIBI class NTPase domain combined with a unique C-terminal extension. This family is characterized by a unique, large, structured insert within the NTPase domain between strands 2 and 3, a distinctive signature in the Walker A motif (TPxGEGK[TS]) and, similarly to the BioD and G3E families, contains a glutamate residue in the Walker B motif (Figure 3). The asparagine residue of the NKxD motif is retained in this family and is followed by a conserved aspartate residue four residues downstream that may be the equivalent of the D seen in the canonical form of the motif (Figure 3). This family shows little sequence diversity; it is distributed widely in Bacteria and sporadically in Archaea (e.g. in several Gram-positive bacteria, such as Clostridium, Ureaplasma and Thermotoga, γ and α-proteobacteria, and Thermoplasma), and is conserved throughout the eukaryotic crown group (Figure 5(k)). This pattern is compatible with a bacterial origin and horizontal transfer to eukaryotes from the pro-mitochondrial symbiont and, independently, to the archaeon Thermoplasma.

Signal-recognition-associated GTPase family

The signal recognition particle GTPase (SRP54/Ffh) and the α-subunit of the signal receptor (SR; FtsY in Bacteria) form the two main subfamilies of the signal-recognition-associated GTPase family. The SRP is a ribonucleoprotein that transports specific proteins to cell or endoplasmic reticulum (ER) membranes for insertion or secretion. The SRP54/Ffh protein consists of three domains, N, G (GTPase), and M. The carboxy-terminal M-domain contains the SRP RNP-binding site and the signal-sequence-binding site.[161,162] The SRP receptor contains homologous equivalents of the N and G domains of SR54, but not of the C-terminal M domain. Instead, the SR/FtsY subfamily has an N-terminal extension that is important for membrane attachment.[163] The SRP and receptor GTPases are ubiquitous in the three superkingdoms and the two corresponding parts of the phylogenetic tree follow the standard model, with distinct archaeo-eukaryotic branches (Figure 5(l)). This indicates that both these GTPases were already present in the LUCA and that the original duplication leading to the divergence of SR and FtsY occurred at an even earlier stage of evolution.[164] In contrast to these ubiquitous GTPases, the FlhF subfamily appears to have emerged in Bacteria, possibly through a duplication of SRP54 (tree not shown), and has been recruited for flagellar assembly.[165] The SRP-associated GTPase family seems to bear some primitive functional features of what might have been the common ancestor of the SIMIBI clade, a GTPase that functioned as part of a ribonucleoprotein complex, but also had a membrane-associated function.

The G3E family

The presence of conserved sequence features, such as the GxxGxGK[ST] variant of the Walker A motif, the Mg^{2+}-binding aspartate residue in strand 2, and the other aspartate residue at the N terminus of the Walker B strand (Figure 3), firmly establishes this family in the SIMIBI class. This family is defined by a glutamate residue in the Walker B (G3) motif substituting for the conserved Mg^{2+}-binding aspartate residue,[166] combined with an intact NKXD motif. The G3E family contains four well-defined ancient conserved subfamilies, UreG, HypB, CobW, and ArgK (Figure 5(l)). UreG is an accessory urease subunit that appears to be involved in the assembly of the nickel metallocenter in metalloenzymes.[167] This subfamily is widespread in bacteria and also found in some eukaryotes (plants and the fungus *Neurospora crassa,* but not *S. cerevisiae* or animals). HypB (hydrogenase expression protein B) contains an N-terminal histidine-rich region implicated in metal binding, followed by a GTPase domain. This protein probably functions similarly to UreG, in incorporating Ni^{2+} into the hydrogenase,[168–171] and is found in both Archaea and Bacteria. CobW participates in cobalamin synthesis, but its exact role in this process in unclear;[172] CobW is widespread in Bacteria and Eukaryota, but missing from Archaea (Figure 5(m)). ArgK participates in the transport of positively charged amino acids (lysine, arginine, and ornithine) and has arginine kinase activity.[173] ArgK is found in a small, but phylogenetically diverse array of bacteria and archaea, and in Caenorhabditis and Leishmania among the eukaryotes. Although ArgK has been characterized as an ATPase,[173] the conservation of the aspartate residue in the [NT]KxD motif in ArgK and all other G3E family members suggests that their actual *in vivo* substrate is GTP rather than ATP. Since all four G3E subfamilies are widespread in Bacteria, but show limited distribution in Archaea and Eukaryotes (Figure 5(m)), we hypothesize that the G3E superfamily originated in bacteria and was disseminated subsequently through horizontal transfer.

The Clp1/Grc3 family

The Clp1/Grc3 family is represented by one or two members in all Eukaryota and Archaea, with the exception of Methanobacterium. Grc3 (YLL035w) of *S. cerevisiae* is an essential cell-cycle-regulated protein, whereas its paralog Clp1 is the 50 kDA subunit of the tetrameric CF1A protein complex that is required for mRNA cleavage and polyadenylation.[174,175] The phyletic pattern of this family is also consistent with a role in RNA processing, indicating that these proteins participated in the assembly of the poly(A) cleavage complex in the common ancestor of archaea and eukaryotes. The conserved aspartate residue at the base of the Walker B strand, typical of the SIMIBI class, is missing from most members of this family, and the NKXD motif is eroded completely, which indicates that these proteins are not specific for GTP (Figure 3). An unusual variant of the Walker B motif (NxxG) instead of (DxxG) is found in many eukaryotic and at least one archaeal member of this family (Figure 3).

The XAB1 family

This family is ubiquitous in eukaryotes and is present in several archaea; most eukaryotes have at least two paralogs of this family. Human XAB1 is a ubiquitously expressed cytoplasmic GTPase that binds the DNA-repair protein Xp-A and possibly regulates its translocation into the nucleus.[176] This family has a canonical [NT]KXD motif (Figure 3) consistent with the observed GTPase activity.[176] Additionally, this family has a distinguishing GPNG motif associated with strand 3 of the NTPase domain (not shown). The phyletic and expression profile of this family suggests a basic cellular role, such as assembly of certain cellular complexes or RNA metabolism.

Discussion

The major events in GTPase evolution

The comprehensive detection and analysis of the proteins of the GTPase superclass allows a fairly detailed reconstruction of their evolution at various levels and thereby provides a natural, evolutionary classification for this class of proteins. Between ten and 13 ancient conserved groups of the GTPase superclass are nearly ubiquitous in Bacteria, Archaea, and Eukaryota (Table 2). Of these, at least the translation factors IF2/Fun12, EF-Tu/EF-1α, and EFG-lepA/EF2, SelB/EIF2 g, two OBG lineages, YawG/YlqF, the SRP GTPase Ffh/SR54 and FtsY/SR receptor, and the Mrp-NBP35 family appear to have been distinct lineages already in LUCA (Figure 6). Additionally, the HflX family, the PurA family, and the PyrG family probably extend back to LUCA (Figure 6). A common biochemical theme in the majority, if not all, GTPase superclass members traceable to LUCA is their association with ribonucleoprotein complexes either in translation or in the protein secretion machinery. Thus, the Ur-GTPase probably was a ribonucleoprotein-associated enzyme that had been part of the translation system since an early stage of the latter's evolution; a similar conclusion has been reached in a recent study of a subset of bacterial GTPases.[38] This ancestral GTPase already differed from the rest of the P-loop NTPases through several distinct features, including, importantly, the specialized DxxG configuration of the Walker B motif and the NKXD motif associated with strand 6, which provides specificity towards GTP. From this ancestor, several diversification events occurred prior to the emergence of LUCA (Figure 6).

The first of these major events marked the divergence of the TRAFAC and SIMIBI GTPases. The association of two universal members of the SIMIBI class with translation-coupled secretion (SR54/SR) and the probable involvement of a third one in membrane-associated chromosome partitioning (Mrp/MinD) suggests that the separation of the two classes of GTPases might have been associated with a crucial event in the evolution of life, the origin of the lipid-based cellular membrane. The early diversification of the TRAFAC class must have been associated with the increasing complexity of the translation machinery. The separation of the

OBG family, the HflX family, and the origin of the YawG/YlqF family as a result of GTPase domain permutation apparently were among the earliest events along this path of evolution, followed by emergence of the other ancient translation factors. The diversification of the SIMIBI class probably was associated initially with the evolution of the secretion machinery and the emergence of MRP, which involved loss of GTP specificity, was associated with the origin of the chromosome-segregation mechanism and with the increase of chromosome size in early biological systems. Notably, none of these ancient GTPases, with the possible exception of MRP, interacts with DNA, as opposed to other major classes of P-loop NTPases such as RecA, ABC, AAA + , or helicases. Hence, we speculate that the early diversification events among GTPases occurred prior to the advent of DNA as the principal genetic material; the emergence of MRP might have been associated with the reverse transcription stage in genome replication, when DNA started coming to the fore.[177] In addition, even at a pre-LUCA stage of evolution, some metabolic enzymes, such as PyrG and perhaps PurA, probably evolved from the ancestral SIMIBI GTPase, losing the specificity for guanine in the process.

The next set of GTPase groups emerged at the base of the archaeo-eukaryotic and bacterial lineages, and involved several fundamental, lineage-specific innovations, many of which are associated with lineage-specific developments in the translation and RNA processing systems. Some examples include the diversification of a large assembly of GTPases, which includes the Era, EngA, and TrmE families in bacteria, and XAB1 and CLP1 in the archaeo-eukaryotic lineage (Figure 6). The SIMIBI class also spawned several groups of NTPases involved in metabolic functions, such as the G3E family and several families of the BioD-like superfamily. Some of these lineage-specific families attained a secondary broad phyletic distribution, apparently through horizontal gene transfer.

During the subsequent phase of evolution, proteins of the GTPase superclass continued to be recruited for many new functions. These include isolated cases of adaptation for the transport function, as in the case of FeoB of the TRAFAC class and ArsA of the SIMIBI class. The earlier diversification of the ABC ATPases that largely filled this functional niche probably prevented more extensive utilization of GTPases. The Mrp-MinD family, in addition to further diversifying in chromosome-partitioning-related functions, was utilized in other roles, such as nitrogen fixation. The extended Ras-like family, the Fzo/dynamin-family, and the septin-related family also arose from later divergence events in particular lineages. In prokaryotes, the members of these families remain functionally largely uncharacterized and show a sporadic distribution, suggesting that, unlike ancient GTPases, these proteins are not involved in core cellular functions. In contrast, in eukaryotes, these families started to occupy critical functional niches and underwent extensive diversification that involved their recruitment for almost all functional systems of the eukaryotic cell. The most striking case is the Ras-like family, which is relatively inconspicuous in prokaryotes and is not represented at all in many bacteria, but has vastly expanded in eukaryotes and became one of the foundations of eukaryotic signaling systems. Several families of GTPases were ''invented'' in eukaryotes, including the kinesin-myosin superfamily, the GBP1/RHD3 family, and the AP-GTPase family. In particular, the emergence of the kinesin-myosin superfamily was a pivotal event in the evolution of the eukaryote-specific cytoskeleton. More generally, it may be concluded that the early evolution of organizational complexity in eukaryotes and the subsequent elaboration of eukaryotic signaling systems substantially depended on the emergence of new GTPases.

A clear overall picture of the macro-events in GTPase evolution seems to emerge: (i) rounds of duplication of a GTPase with ancestral functions in translation and/or ribonucleoprotein complex assembly followed by colonization of several new functional niches, such as secretion and chromosomal dynamics, and probably some central metabolic functions, characterized the pre-LUCA phase of GTPase evolution; (ii) diversification within the original functional milieus gave rise to several bacteria-specific or archaeo-eukaryotic functions, for example, among translation factors; (iii) at the later stages of evolution, particularly in conjunction with the emergence of eukaryotic complexity, numerous additional duplications and diversification led to the colonization, by GTPases and related NTPases, of practically every conceivable functional niche, with particularly prominent roles in signal transduction. At least at these later stages of evolution, horizontal gene transfers had a major role in expanding the phyletic distribution of many GTPase families.

Origin and loss of GTPase activity

An inevitable corollary of the conclusion that the common ancestor of the GTPase superclass already had the NKxD motif associated with strand 6 is that the ATPases that belong to this superclass and its inactive members have all been derived secondarily through the loss of GTPase specificity or activity. In the TRAFAC class, the GTPase activity was replaced by ATPase activity in the kinesin-myosin clade *via* deletion of the NKxD-containing region. The SIMIBI class shows a GTP → ATP shift in the MinD family and in the BioD-related families through divergence in the region of strand 6 containing the NKxD motif. Disruption of the Walker A and/or B motifs, e.g. in Tsr1p of the Bms1p family or in some septin derivatives, suggests that GTPases have been inactivated independently on several occasions.

The more fundamental issue of the teleology of the initial emergence of GTPase specificity in the ancestor of the entire superclass remains uncertain. There could be at least three possible explanations. Alternative 1: the GTPase specificity initially evolved by chance and was fixed as many important functional systems coalesced around the GTPases (a frozen accident hypothesis). However, the preservation of GTPase activity over a tremendous temporal span, in several functions distinct from what appear to be the original functions of this superclass, taken together with multiple losses, suggests selection for this activity wherever it is required. This leads to alternative 2: the GTPase specificity evolved as an adaptation. The ability to use GTP with high specificity might have prevented competition with the more abundant ATPases, or the effects of fluctuations in ATP levels owing to the activity of these ATPases. Thus, GTPases were selected for critical functions, such as translation, that required steady performance. Alternative 3: the relatively slow GTP hydrolysis rates associated with the early GTPases allowed a special regulatory cycle through GAPs and GEFs, and this provided a unique functional role for the origin and retention of GTP specificity. This alternative is not well supported by the lack of GAPs or GEFs in a wide range of GTPases, suggesting that this form of regulation was not necessarily ancestral to the entire superclass. These alternatives are not entirely mutually exclusive, and a combination of them potentially could have contributed to the fixation of the GTP specificity of GTPases.

Conclusions

The present study shows that our understanding of the structure, functions, and evolution of GTPases has vastly improved in the past decade since the publication of the seminal paper by Bourne and colleagues,[9] but many shortcomings in our current knowledge are equally obvious. Given that multiple complete genomes from all three domains of life are currently available, it seems likely that the present classification includes most of the widespread families of GTPases and related ATPases. However, elucidation of the functional roles of many of these families has not kept up with the rapid progress of genome sequencing. The present analysis suggests the general functional properties that may be expected of the uncharacterized groups of GTPases and makes it possible to address specifically those areas of the GTPase tree where lacunae exist in the understanding of biological and biochemical functions. If experimental studies in the next decade adequately address the now apparent diversity of GTPases and equal in intensity those conducted on the extended Ras-family in the 1990s, there is little doubt that, in 2010, we will have a nearly comprehensive picture of the functional roles of P-loop GTPases.

Materials and Methods

Sequences of GTPases and GTPase-related proteins were extracted from the non-redundant (NR) protein sequence database (National Center for Biotechnology Information, NIH, Bethesda) by using the PSI-BLAST program,[178,179] with the sequences of various previously identified GTPases used as queries. The validity of the candidates detected by this procedure was verified by using multiple sequence alignments to detect signature motifs and 3D-structure-based alignments and comparisons whenever structures were available. Multiple alignments were constructed using the Clustal X program,[180] and corrected on the basis of PSI-BLAST results and structural alignments. For each family of GTPases, the phyletic distribution was evaluated in terms of the presence of homologues in a phylogenetically diverse sample of 32 sequenced genomes, including fungi (Saccharomyces), animals (Caenorhabditis, Drosophila, Homo), green plants (Arabidopsis), Crenarchaeota (Aeropyrum, Sulfolobus), Euryarchaeota (Archaeoglobus, Halobacterium, Methanococcus, Methanobacterium, Thermoplasma, Pyrococcus), Chlamydiales (Chlamydophila, Chlamydia), proteobacteria (*E. coli*, Caulobacter, Vibrio, Rickettsia, Neisseria, Campylobacter, Helicobacter), Cyanobacteria (Synechocystis), Firmicutes (Bacillus, Lactococcus, Clostridium, Mycobacterium, Mycoplasma/Ureaplasma, Streptococcus), the Thermus/Deinococcus group (Deinococcus), Spirochaetales (Borrelia, Treponema), Thermotogales (Thermotoga) and Aquificales (Aquifex). The database of clusters of orthologous groups of proteins (COGs) served as a guide for identifying these conserved groups, especially in the prokaryotic genomes.[181] For the preliminary delineation of relationships within the GTPase superclass, single-linkage sequence similarity clustering using the BLAST bit score/HSP length ratio was carried out using the BLASTCLUST program (I. Dondoshansky, Y. I. Wolf and E.V.K., unpublished†). For phylogenetic analysis, regions that contained gaps in the majority of sequences and some regions with uncertain alignments were excluded. Phylogenetic trees were constructed by using the PROTDIST and FITCH programs of the PHYLIP package with the default parameters‡, followed by optimization *via* local rearrangements conducted using the Maximum Likelihood (ML) method with the JTTF substitution model as implemented in the MOLPHY package.[182,183] In addition, trees were constructed with the TREEPUZZLE 5 software using the BLOSUM62 matrix and otherwise the default settings [184,185] and with the NEIGHBOR program of PHYLIP using the neighbor-joining method.[186] Support for selected tree branches was measured by bootstrapping (10,000 resamplings for ML, 1000 resamplings for PHYLIP analysis) and with the quartet puzzling support value of TREEPUZZLE 5. For structural comparisons, the VAST database at http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml and DALI and FSSP were used.[187–189]

† ftp://ncbi.nlm.nih.gov/blast/documents/README.bcl

‡ Felsenstein, J. (1993). PHYLIP 3.5c - computer package distributed by the author 3.5c edit. Department of Genetics SK-50, University of Washington, Seattle, Washington 98195.

References

1. Saraste, M., Sibbald, P. R. & Wittinghofer, A. (1990). The P-loop – a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**, 430-434.
2. Gorbalenya, A. E. & Koonin, E. V. (1990). Superfamily of UvrA-related NTP-binding proteins. Implications for rational classification of recombination/repair systems. *J. Mol. Biol.* **213**, 583-591.
3. Milner-White, E. J., Coggins, J. R. & Anton, I. A. (1991). Evidence for an ancestral core structure in nucleotide-binding proteins with the type A motif. *J. Mol. Biol.* **221**, 751-754.
4. Schulz, G. E. (1992). Binding of nucleotides by proteins. *Curr. Opin. Struct. Biol.* **2**, 61-67.
5. Schweins, T. & Wittinghofer, A. (1994). GTP-binding proteins. Structures, interactions and relationships. *Curr. Biol.* **4**, 547-550.
6. Vetter, I. R. & Wittinghofer, A. (1999). Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer. *Quart. Rev. Biophys.* **32**, 1-56.
7. Koonin, E. V., Wolf, Y. I. & Aravind, L. (2000). Protein fold recognition using sequence profiles and its application in structural genomics. *Advan. Protein Chem.* **54**, 245-275.
8. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. (1982). Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945-951.
9. Bourne, H. R., Sanders, D. A. & McCormick, F. (1991). The GTPase superfamily: conserved structure and molecular mechanism. *Nature,* **349**, 117-127.
10. Gorbalenya, A. E. & Koonin, E. V. (1993). Helicases: amino acid sequence comparisons and structure-function relationships. *Curr. Opin. Struct. Biol.* **3**, 419-429.
11. Neuwald, A. F., Aravind, L., Spouge, J. L. & Koonin, E. V. (1999). AAA + : a class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res.* **9**, 27-43.
12. Leipe, D. D., Aravind, L., Grishin, N. V. & Koonin, E. V. (2000). The bacterial replicative helicase DnaB evolved from a RecA duplication. *Genome Res.* **10**, 5-16.
13. Rodnina, M. V., Stark, H., Savelsbergh, A., Wieden, H. J., Mohr, D., Matassova, N. B. *et al.* (2000). GTPases mechanisms and functions of translation factors on the ribosome. *Biol. Chem.* **381**, 377-387.
14. Takai, Y., Sasaki, T. & Matozaki, T. (2001). Small GTP-binding proteins. *Physiol. Rev.* **81**, 153-208.
15. Kabsch, W., Gast, W. H., Schulz, G. E. & Leberman, R. (1977). Low resolution structure of partially trypsin-degraded polypeptide elongation factor, EF-Tu, from *Escherichia coli. J. Mol. Biol.* **117**, 999-1012.
16. Halliday, K. R. (1983). Regional homology in GTP-binding proto-oncogene products and elongation factors. *J. Cyclic Nucleotide Protein Phosphor. Res.* **9**, 435-448.
17. Jurnak, F. (1985). Structure of the GDP domain of EF-Tu and location of the amino acids homologous to ras oncogene proteins. *Science,* **230**, 32-36.
18. McCormick, F., Clark, B. F., la Cour, T. F., Kjeldgaard, M., Norskov-Lauritsen, L. & Nyborg, J. (1985). A model for the tertiary structure of p21, the product of the ras oncogene. *Science,* **230**, 78-82.
19. Macara, I. G., Lounsbury, K. M., Richards, S. A., McKiernan, C. & Bar-Sagi, D. (1996). The Ras superfamily of GTPases. *FASEB J.* **10**, 625-630.
20. Garcia-Ranea, J. A. & Valencia, A. (1998). Distribution and functional diversification of the ras superfamily in *Saccharomyces cerevisiae. FEBS Letters,* **434**, 219-225.
21. Yu, H. & Schreiber, S. L. (1995). Structure of guanine-nucleotide-exchange factor human Mss4 and identification of its Rab-interacting surface. *Nature,* **376**, 788-791.
22. Barrett, T., Xiao, B., Dodson, E. J., Dodson, G., Ludbrook, S. B., Nurmahomed, K. *et al.* (1997). The structure of the GTPase-activating domain from p50rhoGAP. *Nature,* **385**, 458-461.
23. Nassar, N., Hoffman, G. R., Manor, D., Clardy, J. C. & Cerione, R. A. (1998). Structures of Cdc42 bound to the active and catalytically compromised forms of Cdc42GAP. *Nature Struct. Biol.* **5**, 1047-1052.
24. Zhu, Z., Dumas, J. J., Lietzke, S. E. & Lambright, D. G. (2001). A helical turn motif in Mss4 is a critical determinant of Rab binding and nucleotide release. *Biochemistry,* **40**, 3027-3036.
25. Trimble, W. S. (1999). Septins: a highly conserved family of membrane-associated GTPases with functions in cell division and beyond. *J. Membr. Biol.* **169**, 75-81.
26. Field, C. M. & Kellogg, D. (1999). Septins: cytoskeletal polymers or signalling GTPases? *Trends Cell Biol.* **9**, 387-394.
27. Hales, K. G. & Fuller, M. T. (1997). Developmentally regulated mitochondrial fusion mediated by a conserved, novel, predicted GTPase. *Cell,* **90**, 121-129.
28. Marks, B., Stowell, M. H., Vallis, Y., Mills, I. G., Gibson, A., Hopkins, C. R. & McMahon, H. T. (2001). GTPase activity of dynamin and resulting conformation change are essential for endocytosis. *Nature,* **410**, 231-235.
29. Czworkowski, J., Wang, J., Steitz, T. A. & Moore, P. B. (1994). The crystal structure of elongation factor G complexed with GDP, at 2.7 Å resolution. *EMBO J.* **13**, 3661-3668.
30. Choi, S. K., Lee, J. H., Zoll, W. L., Merrick, W. C. & Dever, T. E. (1998). Promotion of met-tRNAiMet binding to ribosomes by yIF2, a bacterial IF2 homolog in yeast. *Science,* **280**, 1757-1760.
31. Chen, X., Court, D. L. & Ji, X. (1999). Crystal structure of ERA: a GTPase-dependent cell cycle regulator containing an RNA binding motif. *Proc. Natl Acad. Sci. USA,* **96**, 8396-8401.
32. Pestova, T. V., Lomakin, I. B., Lee, J. H., Choi, S. K., Dever, T. E. & Hellen, C. U. (2000). The joining of ribosomal subunits in eukaryotes requires eIF5B. *Nature,* **403**, 332-335.
33. Scott, J. M., Ju, J., Mitchell, T. & Haldenwang, W. G. (2000). The *Bacillus subtilis* GTP binding protein obg and regulators of the sigma(B) stress response transcription factor cofractionate with ribosomes. *J. Bacteriol.* **182**, 2771-2777.
34. Freymann, D. M., Keenan, R. J., Stroud, R. M. & Walter, P. (1999). Functional changes in the structure of the SRP GTPase on binding GDP and Mg^{2+}GDP. *Nature Struct. Biol.* **6**, 793-801.
35. Montoya, G., Kaat, K., Moll, R., Schafer, G. & Sinning, I. (2000). The crystal structure of the conserved GTPase of SRP54 from the archaeon *Acidianus ambivalens* and its comparison with related structures suggests a model for the SRP-SRP receptor complex. *Struct. Fold. Des.* **8**, 515-525.

36. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucl. Acids Res.* **28**, 33-36.
37. Mittenhuber, G. (2001). Comparative genomics of prokaryotic GTP-binding proteins (the Era, Obg, EngA, ThdF (TrmE), YchF and YihA families) and their relationship to eukaryotic GTP-binding proteins (the DRG, ARF, RAB, RAN, RAS and RHO families). *J. Mol. Microbiol. Biotechnol.* **3**, 21-35.
38. Caldon, C. E., Yoong, P. & March, P. E. (2001). Evolution of a molecular switch: universal bacterial GTPases regulate ribosome function. *Mol. Microbiol.* **41**, 289-297.
39. Smith, C. A. & Rayment, I. (1996). Active site comparisons highlight structural similarities between myosin and other P-loop proteins. *Biophys. J.* **70**, 1590-1602.
40. Kull, F. J., Vale, R. D. & Fletterick, R. J. (1998). The case for a common ancestor: kinesin and myosin motor proteins and G proteins. *J. Muscle Res. Cell Motil.* **19**, 877-886.
41. Harton, J. A., Cressman, D. E., Chin, K. C., Der, C. J. & Ting, J. P. (1999). GTP binding by class II transactivator: role in nuclear import. *Science,* **285**, 1402-1405.
42. Pieper, U., Brinkmann, T., Kruger, T., Noyer-Weidner, M. & Pingoud, A. (1997). Characterization of the interaction between the restriction endonuclease McrBC from *E. coli* and its cofactor GTP. *J. Mol. Biol.* **272**, 190-199.
43. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
44. Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **28**, 257-259.
45. 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.
46. Orengo, C. A., Todd, A. E. & Thornton, J. M. (1999). From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374-382.
47. Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316-319.
48. Doolittle, R. F. & Handy, J. (1998). Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr. Opin. Genet. Dev.* **8**, 630-636.
49. Story, R. M. & Steitz, T. A. (1992). Structure of the recA protein-ADP complex. *Nature,* **355**, 374-376.
50. Brosh, R. M. & Matson, S. W. (1995). Mutations in motif II of *Escherichia coli* DNA helicase II render the enzyme nonfunctional in both mismatch repair and excision repair with differential effects on the unwinding reaction. *J. Bacteriol.* **177**, 5612-5621.
51. Subramanya, H. S., Bird, L. E., Brannigan, J. A. & Wigley, D. B. (1996). Crystal structure of a DExx box DNA helicase. *Nature,* **384**, 379-383.
52. Hung, L. W., Wang, I. X., Nikaido, K., Liu, P. Q., Ames, G. F. & Kim, S. H. (1998). Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature,* **396**, 703-707.
53. Campbell, M. J. & Davis, R. W. (1999). On the *in vivo* function of the RecA ATPase. *J. Mol. Biol.* **286**, 437-445.
54. Soultanas, P., Dillingham, M. S., Velankar, S. S. & Wigley, D. B. (1999). DNA binding mediates conformational changes and metal ion coordination in the active site of PcrA helicase. *J. Mol. Biol.* **290**, 137-148.
55. Perkins, G. & Diffley, J. F. (1998). Nucleotide-dependent prereplicative complex assembly by Cdc6p, a homolog of eukaryotic and prokaryotic clamp-loaders. *Mol. Cell,* **2**, 23-32.
56. Herbig, U., Marlar, C. A. & Fanning, E. (1999). The Cdc6 nucleotide-binding site regulates its activity in DNA replication in human cells. *Mol. Biol. Cell,* **10**, 2631-2645.
57. Sondek, J., Lambright, D. G., Noel, J. P., Hamm, H. E. & Sigler, P. B. (1994). GTPase mechanism of Gproteins from the 1.7-Å crystal structure of transducin alpha-GDP-AIF-4. *Nature,* **372**, 276-279.
58. Schweins, T., Scheffzek, K., Assheuer, R. & Wittinghofer, A. (1997). The role of the metal ion in the p21ras catalysed GTP-hydrolysis: Mn^{2+} *versus* Mg^{2+}. *J. Mol. Biol.* **266**, 847-856.
59. Coleman, D. E., Berghuis, A. M., Lee, E., Linder, M. E., Gilman, A. G. & Sprang, S. R. (1994). Structures of active conformations of Gi alpha 1 and the mechanism of GTP hydrolysis. *Science,* **265**, 1405-1412.
60. Berchtold, H., Reshetnikova, L., Reiser, C. O., Schirmer, N. K., Sprinzl, M. & Hilgenfeld, R. (1993). Crystal structure of active elongation factor Tu reveals major domain rearrangements [published erratum appears in *Nature* 1993 Sep 23; **365**, 368]. *Nature,* **365**, 126-132.
61. Zhang, Y., Yu, N. J. & Spremulli, L. L. (1998). Mutational analysis of the roles of residues in *Escherichia coli* elongation factor Ts in the interaction with elongation factor Tu. *J. Biol. Chem.* **273**, 4556-4562.
62. Forchhammer, K., Leinfelder, W. & Bock, A. (1989). Identification of a novel translation factor necessary for the incorporation of selenocysteine into protein. *Nature,* **342**, 453-456.
63. Kromayer, M., Wilting, R., Tormay, P. & Bock, A. (1996). Domain structure of the prokaryotic selenocysteine-specific elongation factor SelB. *J. Mol. Biol.* **262**, 413-420.
64. Merrick, W. C. (1992). Mechanism and regulation of eukaryotic protein synthesis. *Microbiol. Rev.* **56**, 291-315.
65. Lee, J. H., Choi, S. K., Roll-Mecak, A., Burley, S. K. & Dever, T. E. (1999). Universal conservation in translation initiation revealed by human and archaeal homologs of bacterial translation initiation factor IF2. *Proc. Natl Acad. Sci. USA,* **96**, 4342-4347.
66. Wallrapp, C., Verrier, S. B., Zhouravleva, G., Philippe, H., Philippe, M., Gress, T. M. & Jean-Jean, O. (1998). The product of the mammalian orthologue of the *Saccharomyces cerevisiae* HBS1 gene is phylogenetically related to eukaryotic release factor 3 (eRF3) but does not carry eRF3-like activity. *FEBS Letters,* **440**, 387-392.
67. Inagaki, Y. & Ford Doolittle, W. (2000). Evolution of the eukaryotic translation termination system: origins of release factors. *Mol. Biol. Evol.* **17**, 882-889.
68. Shoemaker, N. B., Vlamakis, H., Hayes, K. & Salyers, A. A. (2001). Evidence for extensive resistance gene transfer among bacteroides sand among bacteroides and other genera in the human colon. *Appl. Environ. Microbiol.* **67**, 561-568.
69. Leyh, T. S., Taylor, J. C. & Markham, G. D. (1988). The sulfate activation locus of *Escherichia coli* K12:

cloning, genetic, and enzymatic characterization. *J. Biol. Chem.* **263**, 2409-2416.

70. Schwedock, J. & Long, S. R. (1990). ATP sulphurylase activity of the nodP and nodQ gene products of *Rhizobium meliloti*. *Nature,* **348**, 644-647.
71. Keeling, P. J. & Doolittle, W. F. (1995). Archaea: narrowing the gap between prokaryotes and eukaryotes. *Proc. Natl Acad. Sci. USA,* **92**, 5761-5764.
72. Keeling, P. J., Fast, N. M. & McFadden, G. I. (1998). Evolutionary relationship between translation initiation factor eIF-2gamma and selenocysteine-specific elongation factor SELB: change of function in translation factors. *J. Mol. Evol.* **47**, 649-655.
73. Avarsson, A. (1995). Structure-based sequence alignment of elongation factors Tu and G with related GTPases involved in translation. *J. Mol. Evol.* **41**, 1096-1104.
74. Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., Zhao, Y. & Chait, B. T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148**, 635-651.
75. Casalone, E., Barberio, C., Cavalieri, D., Ceccarelli, I., Riparbelli, M., Ugolini, S. & Polsinelli, M. (1999). Disruption and phenotypic analysis of six novel genes from chromosome IV of *Saccharomyces cerevisiae* reveal YDL060w as an essential gene for vegetative growth. *Yeast,* **15**, 1691-1701.
76. Noble, J. A., Innis, M. A., Koonin, E. V., Rudd, K. E., Banuett, F. & Herskowitz, I. (1993). The *Escherichia coli* hflA locus encodes a putative GTP-binding protein and two membrane proteins, one of which contains a protease-like domain. *Proc. Natl Acad. Sci. USA,* **90**, 10866-10870.
77. Maddock, J., Bhatt, A., Koch, M. & Skidmore, J. (1997). Identification of an essential *Caulobacter crescentus* gene encoding a member of the Obg family of GTP-binding proteins. *J. Bacteriol.* **179**, 6426-6431.
78. Okamoto, S., Itoh, M. & Ochi, K. (1997). Molecular cloning and characterization of the obg gene of *Streptomyces griseus* in relation to the onset of morphological differentiation. *J. Bacteriol.* **179**, 170-179.
79. Wolf, Y. I., Aravind, L., Grishin, N. V. & Koonin, E. V. (1999). Evolution of aminoacyl-tRNA synthetases – analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9**, 689-710.
80. Park, J. H., Jensen, B. C., Kifer, C. T. & Parsons, M. (2001). A novel nucleolar G-protein conserved in eukaryotes. *J. Cell Sci.* **114**, 173-185.
81. Decoster, E., Vassal, A. & Faye, G. (1993). MSS1, a nuclear-encoded mitochondrial GTPase involved in the expression of COX1 subunit of cytochrome *c* oxidase. *J. Mol. Biol.* **232**, 79-88.
82. Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. V. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* **11**, 356-372.
83. Cabedo, H., Macian, F., Villarroya, M., Escudero, J. C., Martinez-Vicente, M., Knecht, E. & Armengod, M. E. (1999). The *Escherichia coli* trmE (mnmE) gene, involved in tRNA modification, codes for an evolutionarily conserved GTPase with unusual biochemical properties. *EMBO J.* **18**, 7063-7076.
84. Kammler, M., Schon, C. & Hantke, K. (1993). Characterization of the ferrous iron uptake system of *Escherichia coli. J. Bacteriol.* **175**, 6212-6219.
85. Arigoni, F., Talabot, F., Peitsch, M., Edgerton, M. D., Meldrum, E., Allet, E. *et al.* (1998). A genome-based approach for the identification of essential bacterial genes. *Nature Biotechnol.* **16**, 851-856.
86. Dassain, M., Leroy, A., Colosetti, L., Carole, S. & Bouche, J. P. (1999). A new essential gene of the 'minimal genome' affecting cell division. *Biochimie,* **81**, 889-895.
87. Zuber, M., Hoover, T. A., Dertzbaugh, M. T. & Court, D. L. (1997). A *Francisella tularensis* DNA clone complements *Escherichia coli* defective for the production of Era, an essential Ras-like GTP-binding protein. *Gene,* **189**, 31-34.
88. Johnstone, B. H., Handler, A. A., Chao, D. K., Nguyen, V., Smith, M., Ryu, S. Y. *et al.* (1999). The widely conserved Era G-protein contains an RNA-binding domain required for Era function *in vivo. Mol. Microbiol.* **33**, 1118-1131.
89. Gollop, N. & March, P. E. (1991). A GTP-binding protein (Era) has an essential role in growth rate and cell cycle control in *Escherichia coli. J. Bacteriol.* **173**, 2265-2270.
90. Britton, R. A., Powell, B. S., Dasgupta, S., Sun, Q., Margolin, W., Lupski, J. R. & Court, D. L. (1998). Cell cycle arrest in Era GTPase mutants: a potential growth rate-regulated checkpoint in *Escherichia coli* [published erratum appears in *Mol. Microbiol.* 1998, **28**, 1391-1393]. *Mol. Microbiol.* **27**, 739-750.
91. Meier, T. I., Peery, R. B., McAllister, K. A. & Zhao, G. (2000). Era GTPase of *Escherichia coli*: binding to 16S rRNA and modulation of GTPase activity by RNA and carbohydrates. *Microbiology,* **146**, 1071-1083.
92. Mehr, I. J., Long, C. D., Serkin, C. D. & Seifert, H. S. (2000). A homologue of the recombination-dependent growth gene, rdgC, is involved in gonococcal pilin antigenic variation. *Genetics,* **154**, 523-532.
93. Neufeld, T. P. & Rubin, G. M. (1994). The Drosophila peanut gene is required for cytokinesis and encodes a protein similar to yeast putative bud neck filament proteins. *Cell,* **77**, 371-379.
94. Kartmann, B. & Roth, D. (2001). Novel roles for mammalian septins: from vesicle trafficking to oncogenesis. *J. Cell Sci.* **114**, 839-844.
95. Zhang, J., Kong, C., Xie, H., McPherson, P. S., Grinstein, S. & Trimble, W. S. (1999). Phosphatidylinositol polyphosphate binding to the mammalian septin H5 is modulated by GTP. *Curr. Biol.* **9**, 1458-1467.
96. Wainright, P. O., Hinkle, G., Sogin, M. L. & Stickel, S. K. (1993). Monophyletic origin of the metazoa: an evolutionary link with fungi. *Science,* **260**, 340-342.
97. Poirier, G. M., Anderson, G., Huvar, A., Wagaman, P. C., Shuttleworth, J., Jenkinson, E. *et al.* (1999). Immune-associated nucleotide-1 (IAN-1) is a thymic selection marker and defines a novel gene family conserved in plants. *J. Immunol.* **163**, 4960-4969.
98. Reuber, T. L. & Ausubel, F. M. (1996). Isolation of Arabidopsis genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes. *Plant Cell,* **8**, 241-249.
99. Gutensohn, M., Schulz, B., Nicolay, P. & Flugge, U. I. (2000). Functional analysis of the two Arabidopsis homologues of Toc34, a component of the chloroplast protein import apparatus. *Plant J.* **23**, 771-783.
100. Spormann, A. M. & Kaiser, D. (1999). Gliding mutants of *Myxococcus xanthus* with high reversal frequencies and small displacements. *J. Bacteriol.* **181**, 2593-2601.
101. Koonin, E. V. & Aravind, L. (2000). Dynein light chains of the Roadblock/LC7 group belong to an

ancient protein superfamily implicated in NTPase regulation. *Curr. Biol.* **10**, R774-R776.

102. Aravind, L., Dixit, V. M. & Koonin, E. V. (2001). Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science,* **291**, 1279-1284.
103. Mocz, G. & Gibbons, I. R. (2001). Model for the motor component of dynein heavy chain based on homology to the AAA family of oligomeric ATPases. *Structure,* **9**, 93-103.
104. Endow, S. A. (1999). Microtubule motors in spindle and chromosome motility. *Eur. J. Biochem.* **262**, 12-18.
105. Sack, S., Kull, F. J. & Mandelkow, E. (1999). Motor proteins of the kinesin family. Structures, variations, and nucleotide binding sites. *Eur. J. Biochem.* **262**, 1-11.
106. Kull, F. J., Sablin, E. P., Lau, R., Fletterick, R. J. & Vale, R. D. (1996). Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature,* **380**, 550-555.
107. Racevskis, J., Dill, A., Stockert, R. & Fineberg, S. A. (1996). Cloning of a novel nucleolar guanosine 5′-triphosphate binding protein autoantigen from a breast tumor. *Cell Growth Differ.* **7**, 271-280.
108. Draper, D. E. & Reynaldo, L. P. (1999). RNA binding strategies of ribosomal proteins. *Nucl. Acids Res.* **27**, 381-388.
109. Ponting, C. P. & Russell, R. B. (1995). Swaposins: circular permutations within genes encoding saposin homologues. *Trends Biochem. Sci.* **20**, 179-180.
110. Pan, T. & Uhlenbeck, O. C. (1993). Circularly permuted DNA, RNA and proteins – a review. *Gene,* **125**, 111-114.
111. Urrutia, R., Henley, J. R., Cook, T. & McNiven, M. A. (1997). The dynamins: redundant or distinct functions for an expanding family of related GTPases?. *Proc. Natl Acad. Sci. USA,* **94**, 377-384.
112. van Der Bliek, A. M. (2000). A mitochondrial division apparatus takes shape. *J. Cell Biol.* **151**, F1-F4.
113. Schmid, S. L., McNiven, M. A. & De Camilli, P. (1998). Dynamin and its partners: a progress report. *Curr. Opin. Cell Biol.* **10**, 504-512.
114. Pitossi, F., Blank, A., Schroder, A., Schwarz, A., Hussi, P., Schwemmle, M. *et al.* (1993). A functional GTP-binding motif is necessary for antiviral activity of Mx proteins. *J. Virol.* **67**, 6726-6732.
115. van der Bliek, A. M. (1999). Functional diversity in the dynamin family. *Trends Cell Biol.* **9**, 96-102.
116. Sever, S., Muhlberg, A. B. & Schmid, S. L. (1999). Impairment of dynamin's GAP domain stimulates receptor-mediated endocytosis. *Nature,* **398**, 481-486.
117. Stowell, M. H., Marks, B., Wigge, P. & McMahon, H. T. (1999). Nucleotide-dependent conformational changes in dynamin: evidence for a mechanochemical molecular spring. *Nature Cell Biol.* **1**, 27-32.
118. Sever, S., Damke, H. & Schmid, S. L. (2000). Dynamin: GTP controls the formation of constricted coated pits, the rate limiting step in clathrin-mediated endocytosis. *J. Cell Biol.* **150**, 1137-1148.
119. Hermann, G. J., Thatcher, J. W., Mills, J. P., Hales, K. G., Fuller, M. T., Nunnari, J. & Shaw, J. M. (1998). Mitochondrial fusion in yeast requires the transmembrane GTPase Fzo1p. *J. Cell Biol.* **143**, 359-373.
120. Sesaki, H. & Jensen, R. E. (1999). Division versus fusion: Dnm1p and Fzo1p antagonistically regulate mitochondrial shape. *J. Cell Biol.* **147**, 699-706.
121. Prakash, B., Renault, L., Praefcke, G. J., Herrmann, C. & Wittinghofer, A. (2000). Triphosphate structure of guanylate-binding protein 1 and implications for nucleotide binding and GTPase mechanism. *EMBO J.* **19**, 4555-4564.
122. Parker, J. S., Cavell, A. C., Dolan, L., Roberts, K. & Grierson, C. S. (2000). Genetic interactions during root hair morphogenesis in Arabidopsis. *Plant Cell,* **12**, 1961-1974.
123. Inoue, S., Orimo, A., Saito, T., Ikeda, K., Sakata, K., Hosoi, T. *et al.* (1997). A novel RING finger protein, BFP, predominantly expressed in the brain. *Biochem. Biophys. Res. Commun.* **240**, 8-14.
124. Zhou, T., Radaev, S., Rosen, B. P. & Gatti, D. L. (2000). Structure of the ArsA ATPase: the catalytic subunit of a heavy metal resistance pump. *EMBO J.* **19**, 4838-4845.
125. Schindelin, H., Kisker, C., Schlessman, J. L., Howard, J. B. & Rees, D. C. (1997). Structure of ADP × AIF4(−)-stabilized nitrogenase complex and its implications for signal transduction. *Nature,* **387**, 370-376.
126. Hayashi, I., Oyama, T. & Morikawa, K. (2001). Structural and functional studies of MinD ATPase: implications for the molecular recognition of the bacterial cell division apparatus. *EMBO J.* **20**, 1819-1828.
127. Koonin, E. V. (1993). A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif [published erratum appears in *J. Mol. Biol.* 1993, **232**, 1013]. *J. Mol. Biol.* **229**, 1165-1174.
128. Vitale, G., Fabre, E. & Hurt, E. C. (1996). NBP35 encodes an essential and evolutionary conserved protein in *Saccharomyces cerevisiae* with homology to a superfamily of bacterial ATPases. *Gene,* **178**, 97-106.
129. Gerard, E., Labedan, B. & Forterre, P. (1998). Isolation of a minD-like gene in the hyperthermophilic archaeon Pyrococcus AL585, and phylogenetic characterization of related proteins in the three domains of life. *Gene,* **222**, 99-106.
130. Nakashima, H., Grahovac, M. J., Mazzarella, R., Fujiwara, H., Kitchen, J. R., Threat, T. A. & Ko, M. S. (1999). Two novel mouse genes – Nubp2, mapped to the t-complex on chromosome 17, and Nubp1, mapped to chromosome 16 – establish a new gene family of nucleotide-binding proteins in eukaryotes. *Genomics,* **60**, 152-160.
131. de Boer, P. A., Crossley, R. E., Hand, A. R. & Rothfield, L. I. (1991). The MinD protein is a membrane ATPase required for the correct placement of the *Escherichia coli* division site. *EMBO J.* **10**, 4371-4380.
132. Wakasugi, T., Nagai, T., Kapoor, M., Sugita, M., Ito, M., Ito, S. *et al.* (1997). Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. *Proc. Natl Acad. Sci. USA,* **94**, 5967-5972.
133. Turmel, M., Otis, C. & Lemieux, C. (1999). The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proc. Natl Acad. Sci. USA,* **96**, 10248-10253.
134. Campos-Garcia, J., Najera, R., Camarena, L. & Soberon-Chavez, G. (2000). The pseudomonas aeruginosa motR gene involved in regulation of bacterial motility. *FEMS Microbiol. Letters,* **184**, 57-62.

135. Motallebi-Veshareh, M., Rouch, D. A. & Thomas, C. M. (1990). A family of ATPases involved in active partitioning of diverse bacterial plasmids. *Mol. Microbiol.* **4**, 1455-1463.
136. Mohl, D. A. & Gober, J. W. (1997). Cell cycle-dependent polar localization of chromosome partitioning proteins in *Caulobacter crescentus*. *Cell,* **88**, 675-684.
137. Kim, H. J., Calcutt, M. J., Schmidt, F. J. & Chater, K. F. (2000). Partitioning of the linear chromosome during sporulation of *Streptomyces coelicolor* A3(2) involves an oriC-linked parAB locus. *J. Bacteriol.* **182**, 1313-1320.
138. Quisel, J. D. & Grossman, A. D. (2000). Control of sporulation gene expression in *Bacillus subtilis* by the chromosome partitioning proteins Soj (ParA) and Spo0 J (ParB). *J. Bacteriol.* **182**, 3446-3451.
139. Ireton, K., Gunther, N. W. T. & Grossman, A. D. (1994). spo0 J is required for normal chromosome segregation as well as the initiation of sporulation in *Bacillus subtilis*. *J. Bacteriol.* **176**, 5320-5329.
140. Suzuki, J. Y. & Bauer, C. E. (1992). Light-independent chlorophyll biosynthesis: involvement of the chloroplast gene chlL (frxC). *Plant Cell,* **4**, 929-940.
141. Rosen, B. P. (1990). The plasmid-encoded arsenical resistance pump: an anion-translocating ATPase. *Res. Microbiol.* **141**, 336-341.
142. Chen, C. M., Misra, T. K., Silver, S. & Rosen, B. P. (1986). Nucleotide sequence of the structural genes for an anion pump. The plasmid-encoded arsenical resistance operon. *J. Biol. Chem.* **261**, 15030-15038.
143. Suzuki, K., Wakao, N., Kimura, T., Sakka, K. & Ohmiya, K. (1998). Expression and regulation of the arsenic resistance operon of *Acidiphilium multivorum* AIU 301 plasmid pKW301 in *Escherichia coli*. *Appl. Environ. Microbiol.* **64**, 411-418.
144. Ng, W. V., Ciufo, S. A., Smith, T. M., Bumgarner, R. E., Baskin, D., Faust, J. *et al.* (1998). Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome?. *Genome Res.* **8**, 1131-1141.
145. Rosen, B. P. (1999). Families of arsenic transporters. *Trends Microbiol.* **7**, 207-212.
146. Wugeditsch, T., Paiment, A., Hocking, J., Drummelsmith, J., Forrester, C. & Whitfield, C. (2001). Phosphorylation of Wzc, a tyrosine autokinase, is essential for assembly of group 1 capsular polysaccharides in *Escherichia coli*. *J. Biol. Chem.* **276**, 2361-2371.
147. Ilan, O., Bloch, Y., Frankel, G., Ullrich, H., Geider, K. & Rosenshine, I. (1999). Protein tyrosine kinases in bacterial pathogens are associated with virulence and production of exopolysaccharide. *EMBO J.* **18**, 3241-3248.
148. Vincent, C., Doublet, P., Grangeasse, C., Vaganay, E., Cozzone, A. J. & Duclos, B. (1999). Cells of *Escherichia coli* contain a protein-tyrosine kinase, Wzc, and a phosphotyrosine-protein phosphatase, Wzb. *J. Bacteriol.* **181**, 3472-3477.
149. Vincent, C., Duclos, B., Grangeasse, C., Vaganay, E., Riberty, M., Cozzone, A. J. & Doublet, P. (2000). Relationship between exopolysaccharide production and protein-tyrosine phosphorylation in Gram-negative bacteria. *J. Mol. Biol.* **304**, 311-321.
150. Yang, G., Sandalova, T., Lohman, K., Lindqvist, Y. & Rendina, A. R. (1997). Active site mutants of *Escherichia coli* dethiobiotin synthetase: effects of mutations on enzyme catalytic and structural properties. *Biochemistry,* **36**, 4751-4760.
151. Eisenberg, M. A. & Krell, K. (1979). Dethiobiotin synthetase. *Methods Enzymol.* **62**, 348-352.
152. Huang, W., Jia, J., Gibson, K. J., Taylor, W. S., Rendina, A. R., Schneider, G. & Lindqvist, Y. (1995). Mechanism of an ATP-dependent carboxylase, dethiobiotin synthetase, based on crystallographic studies of complexes with substrates and a reaction intermediate. *Biochemistry,* **34**, 10985-10995.
153. Galperin, M. Y. & Grishin, N. V. (2000). The synthetase domains of cobalamin biosynthesis amidotransferases cobB and cobQ belong to a new family of ATP-dependent amidoligases, related to dethiobiotin synthetase. *Proteins: Struct. Funct. Genet.* **41**, 238-247.
154. Weng, M., Makaroff, C. A. & Zalkin, H. (1986). Nucleotide sequence of *Escherichia coli* pyrG encoding CTP synthetase. *J. Biol. Chem.* **261**, 5568-5574.
155. Pappas, A., Park, T. S. & Carman, G. M. (1999). Characterization of a novel dUTP-dependent activity of CTP synthetase from *Saccharomyces cerevisiae*. *Biochemistry,* **38**, 16671-16677.
156. Blanche, F., Couder, M., Debussche, L., Thibaut, D., Cameron, B. & Crouzet, J. (1991). Biosynthesis of vitamin B12: stepwise amidation of carboxyl groups b, d, e, and g of cobyrinic acid a,c-diamide is catalyzed by one enzyme in *Pseudomonas denitrificans*. *J. Bacteriol.* **173**, 6046-6051.
157. Honzatko, R. B. & Fromm, H. J. (1999). Structure-function studies of adenylosuccinate synthetase from *Escherichia coli*. *Arch. Biochem. Biophys.* **370**, 1-8.
158. Natale, D. A., Shankavaram, U. T., Galperin, M. Y., Wolf, Y. I., Aravind, L. & Koonin, E. V. (2000). Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol.* **1**, RESEARCH0009.
159. Radfar, R., Shin, R., Sheldrick, G. M., Minor, W., Lovell, C. R.,Odom, J. D. *et al.* (2000). The crystal structure of N(10)-formyltetrahydrofolate synthetase from Moorella thermoacetica. *Biochemistry,* **39**, 3920-3926.
160. Kounga, K., Song, S., Haslam, G. C. & Himes, R. H. (1996). Site-directed mutagenesis of putative catalytic and nucleotide binding sites in N10-formyltetrahydrofolate synthetase. *Biochim. Biophys. Acta,* **1296**, 112-120.
161. Zopf, D., Bernstein, H. D., Johnson, A. E. & Walter, P. (1990). The methionine-rich domain of the 54 kd protein subunit of the signal recognition particle contains an RNA binding site and can be cross-linked to a signal sequence. *EMBO J.* **9**, 4511-4517.
162. Keenan, R. J., Freymann, D. M., Walter, P. & Stroud, R. M. (1998). Crystal structure of the signal sequence binding subunit of the signal recognition particle. *Cell,* **94**, 181-191.
163. de Leeuw, E., te Kaat, K., Moser, C., Menestrina, G., Demel, R., de Kruijff, B. *et al.* (2000). Anionic phospholipids are involved in membrane association of FtsY and stimulate its GTPase activity. *EMBO J.* **19**, 531-541.
164. Gribaldo, S. & Cammarano, P. (1998). The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. *J. Mol. Evol.* **47**, 508-516.
165. Carpenter, P. B., Hanlon, D. W. & Ordal, G. W. (1992). flhF, a *Bacillus subtilis* flagellar gene that encodes a putative GTP-binding protein. *Mol. Microbiol.* **6**, 2705-2713.

166. Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996). *Escherichia coli* - functional and evolutionary implications of genome scale computer-aided protein sequence analysis. In *Genomes of Plants and Animals* (Gustafson, J. P. & Flavell, R. B., eds), Plenum Press, New York and London.

167. Lee, M. H., Mulrooney, S. B., Renner, M. J., Markowicz, Y. & Hausinger, R. P. (1992). *Klebsiella aerogenes* urease gene cluster: sequence of ureD and demonstration that four accessory genes (ureD, ureE, ureF, and ureG) are involved in nickel metallocenter biosynthesis. *J. Bacteriol.* **174**, 4324-4330.

168. Lutz, S., Jacobi, A., Schlensog, V., Bohm, R., Sawers, G. & Bock, A. (1991). Molecular characterization of an operon (hyp) necessary for the activity of the three hydrogenase isoenzymes in *Escherichia coli. Mol. Microbiol.* **5**, 123-135.

169. Colbeau, A., Richaud, P., Toussaint, B., Caballero, F. J., Elster, C.,Delphin, C. *et al.* (1993). Organization of the genes necessary for hydrogenase expression in *Rhodobacter capsulatus*. Sequence analysis and identification of two hyp regulatory mutants. *Mol. Microbiol.* **8**, 15-29.

170. Maier, T., Jacobi, A., Sauter, M. & Bock, A. (1993). The product of the hypB gene, which is required for nickel incorporation into hydrogenases, is a novel guanine nucleotide-binding protein. *J. Bacteriol.* **175**, 630-635.

171. Fu, C., Olson, J. W. & Maier, R. J. (1995). HypB protein of Bradyrhizobium japonicum is a metal-binding GTPase capable of binding 18 divalent nickel ions per dimer. *Proc. Natl Acad. Sci. USA,* **92**, 2333-2337.

172. Crouzet, J., Levy-Schil, S., Cameron, B., Cauchois, L., Rigault, S., Rouyez, M. C. *et al.* (1991). Nucleotide sequence and genetic analysis of a 13.1-kilobase-pair *Pseudomonas denitrificans* DNA fragment containing five cob genes and identification of structural genes encoding Cob(I)alamin adenosyltransferase, cobyric acid synthase, and bifunctional cobinamide kinase-cobinamide phosphate guanylyltransferase. *J. Bacteriol.* **173**, 6074-6087.

173. Celis, R. T., Leadlay, P. F., Roy, I. & Hansen, A. (1998). Phosphorylation of the periplasmic binding protein in two transport systems for arginine incorporation in *Escherichia coli* K-12 is unrelated to the function of the transport system. *J. Bacteriol.* **180**, 4828-4833.

174. El-Moghazy, A. N., Zhang, N., Ismail, T., Wu, J., Butt, A., Ahmed, Khan S. *et al.* (2000). Functional analysis of six novel ORFs on the left arm of chromosome XII in *Saccharomyces cerevisiae* reveals two essential genes, one of which is under cell-cycle control. *Yeast,* **16**, 277-288.

175. Minvielle-Sebastia, L., Preker, P. J., Wiederkehr, T., Strahm, Y. & Keller, W. (1997). The major yeast poly(A)-binding protein is associated with cleavage factor IA and functions in premessenger RNA 3′-end formation. *Proc. Natl Acad. Sci. USA,* **94**, 7897-7902.

176. Nitta, M., Saijo, M., Kodo, N., Matsuda, T., Nakatsu, Y., Tamai, H. & Tanaka, K. (2000). A novel cytoplasmic GTPase XAB1 interacts with DNA repair protein XPA. *Nucl. Acids Res.* **28**, 4212-4218.

177. Leipe, D., Aravind, L. & Koonin, E. (1999). Did DNA replication evolve twice independently? *Nucl. Acids Res.* **27**, 3389-3401.

178. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

179. Altschul, S. F. & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444-447.

180. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**, 403-405.

181. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T. & Rao, B. S., *et al.* (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucl. Acids Res.* **29**, 22-28.

182. Adachi, J. & Hasegawa, M. (1996). *MOLPHY Version 2.3. Programs for Molecular Phylogenetics Based on Maximum Likelihood*, Institute of Statistical Mathematics, Tokyo.

183. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275-282.

184. Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964-969.

185. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.

186. Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425.

187. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377-385.

188. Madej, T., Gibrat, J. F. & Bryant, S. H. (1995). Threading a database of protein cores. *Proteins: Struct. Funct. Genet.* **23**, 356-369.

189. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science,* **273**, 595-603.

Edited by J. Thornton

(Received 6 August 2001; received in revised form 8 December 2001; accepted 8 December 2001)