doi:10.1016/j.jmb.2004.08.023 *J. Mol. Biol.* (2004) **343**, 1–28

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

STAND, a Class of P-Loop NTPases Including Animal and Plant Regulators of Programmed Cell Death: Multiple, Complex Domain Architectures, Unusual Phyletic Patterns, and Evolution by Horizontal Gene Transfer

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

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

Using sequence profile analysis and sequence-based structure predictions, we define a previously unrecognized, widespread class of P-loop NTPases. The signal transduction ATPases with numerous domains (STAND) class includes the AP-ATPases (animal apoptosis regulators CED4/Apaf-1, plant disease resistance proteins, and bacterial AfsR-like transcription regulators) and NACHT NTPases (e.g. NAIP, TLP1, Het-E-1) that have been studied extensively in the context of apoptosis, pathogen response in animals and plants, and transcriptional regulation in bacteria. We show that, in addition to these well-characterized protein families, the STAND class includes several other groups of (predicted) NTPase domains from diverse signaling and transcription regulatory proteins from bacteria and eukaryotes, and three Archaea-specific families. We identified the STAND domain in several biologically well-characterized proteins that have not been suspected to have NTPase activity, including soluble adenylyl cyclases, nephrocystin 3 (implicated in polycystic kidney disease), and Rolling pebble (a regulator of muscle development); these findings are expected to facilitate elucidation of the functions of these proteins. The STAND class belongs to the additional strand, catalytic E division of P-loop NTPases together with the AAA+ ATPases, RecA/helicase-related ATPases, ABC-ATPases, and VirD4/PilT-like ATPases. The STAND proteins are distinguished from other P-loop NTPases by the presence of unique sequence motifs associated with the N-terminal helix and the core strand-4, as well as a C-terminal helical bundle that is fused to the NTPase domain. This helical module contains a signature GxP motif in the loop between the two distal helices. With the exception of the archaeal families, almost all STAND NTPases are multidomain proteins containing three or more domains. In addition to the NTPase domain, these proteins typically contain DNA-binding or protein-binding domains, superstructure-forming repeats, such as WD40 and TPR, and enzymatic domains involved in signal transduction, including adenylate cyclases and kinases. By analogy to the AAA+ ATPases, it can be predicted that STAND NTPases use the C-terminal helical bundle as a "lever" to transmit the conformational changes brought about by NTP hydrolysis to effector domains. STAND NTPases represent a novel paradigm in signal transduction, whereby adaptor, regulatory switch, scaffolding, and, in some cases, signal-generating moieties are combined into a single polypeptide. The STAND class consists of 14 distinct families, and the evolutionary history of most of these families is riddled with dramatic instances of lineage-specific expansion and apparent horizontal gene transfer. The STAND NTPases are most abundant in developmentally and organizationally complex

Abbreviations used: HGT, horizontal gene transfer; LUCA, last universal common ancestor.

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

0022-2836/$ - see front matter Published by Elsevier Ltd.

prokaryotes and eukaryotes. Transfer of genes for STAND NTPases from bacteria to eukaryotes on several occasions might have played a significant role in the evolution of eukaryotic signaling systems.

Published by Elsevier Ltd.

Keywords: molecular evolution; P-loop NTPases; AP-ATPases; signaling; multidomain proteins

*Corresponding author

Introduction

Utilization of nucleotides as energy intermediates, building blocks for nucleic acids, or regulatory signals is at the center of all fundamental processes in biochemistry. Of the several distinct nucleotide-binding protein folds, the P-loop NTPase fold is the most prevalent domain in proteins encoded in the genomes of most cellular life-forms.[1–4] P-loop NTPase domains have been detected in approximately 5–10% of the predicted gene products in the sequenced prokaryotic and eukaryotic genomes.[5] Analysis of phyletic patterns and phylogenetic relationships of P-loop NTPases from extant organisms indicates that the last universal common ancestor (LUCA) of all modern cellular life-forms already encoded multiple and diverse P-loop NTPases. Thus, this domain must have been among the first to emerge, and comparative analysis of P-loop NTPases has the potential to reveal important aspects of the earliest stages of cellular evolution.[6–9]

While there is a certain degree of diversity in the reactions catalyzed by enzymes of the P-loop NTPase fold, by far the most common one is the hydrolysis of the β–γ phosphate bond of a bound nucleoside triphosphate (NTP). The free energy of NTP hydrolysis is typically utilized to induce conformational changes in other molecules, which constitutes the basis of the biological functions of most P-loop NTPases. Typically, P-loop NTPases show substantial substrate preference for either ATP or GTP. Structurally, the P-loop fold adopts a three-layered α/β sandwich configuration that contains regularly recurring α–β units with the β-strands forming a central, mostly parallel sheet, which is sandwiched between α-helices on both sides[1] (see SCOP database†).[10] At the sequence level, P-loop NTPases are generally characterized by two strongly conserved sequence signatures, the Walker A and Walker B motifs which bind, respectively, the β and γ phosphate moieties of the bound NTP, and a Mg^{2+} cation.[11] The Walker A motif (the P-loop proper) forms a flexible loop between strand 1 and helix 1 of the P-loop domain and has the characteristic sequence pattern GxxxxGK [ST] (x indicates any amino acid residue, alternative residues are shown in brackets) or a variation thereof.[11,12] Side-chain and backbone atoms of the P-loop residues are critical for the positioning of the triphosphate moiety of the bound nucleotide that makes it susceptible to hydrolysis.[1,2] The Walker B motif is composed of a conserved aspartate (or, less often, glutamate) residue at the C terminus of a hydrophobic strand and provides a bond for the octahedral coordination of a Mg^{2+} cation, which, in turn, is coordinated to the β and γ-phosphate moieties of the substrate.[11,12] A hydrogen bond between the Walker B aspartate and the conserved threonine/serine of the P-loop secures the proper relative positioning of the two phosphate-binding motifs.

Comparative sequence and structure analyses suggest that all P-loop ATPase domains belong to one of the two major divisions. The kinase-GTPase (KG) division includes the kinases and GTPases, which share a number of structural similarities, such as the adjacent placement of the P-loop and Walker B strands.[9,13] The additional strand, catalytic E division (for ASCE) is characterized by an additional strand in the core sheet, which is located between the P-loop strand and the Walker B strand.[9,13] Most members of the ASCE division utilize ATP as the preferred substrate and, in contrast to the kinases and GTPases, contain a conserved proton-abstracting acidic residue (typically, glutamate) which primes a water molecule for the nucleophilic attack on the γ-phosphate group of ATP. The ASCE division includes AAA+, ABC, PilT, HerA-FtsK, superfamily 1/2 (SF1/2) helicases, and the RecA/ATP-synthase superfamilies of ATPases, along with several additional, less confidently classified lineages.[9,13–16]

In the past decade, a number of P-loop NTPases have been intensely studied with regard to their critical roles in a range of complex biological processes, such as programmed cell death, disease, and stress response in plants and animals, telomere biogenesis, and heterocaryon incompatibility in fungi. Sequence comparisons showed that the NTPases involved in these functions constitute two major families, the AP (apoptotic)-ATPases and the NACHT NTPases.[17–20] The AP-ATPase family includes the animal APAF1/CED4 ATPases that regulate apoptosis, the plant pathogen and stress resistance proteins, several bacterial transcription regulators, such as GutR and AfsR, and many uncharacterized bacterial proteins.[19,21] The NACHT family consists of the animal disease response NTPases such as CARD4, the NAIP proteins, the telomerase subunit TP1, the fungal heterocaryon incompatibility protein Het-E-1, and uncharacterized proteins from various Bacteria.[20,21]

In previous studies, we attempted to reconstruct the major aspects of the natural history of GTPases,

† http://scop.mrc-lmb.cam.ac.uk/scop/

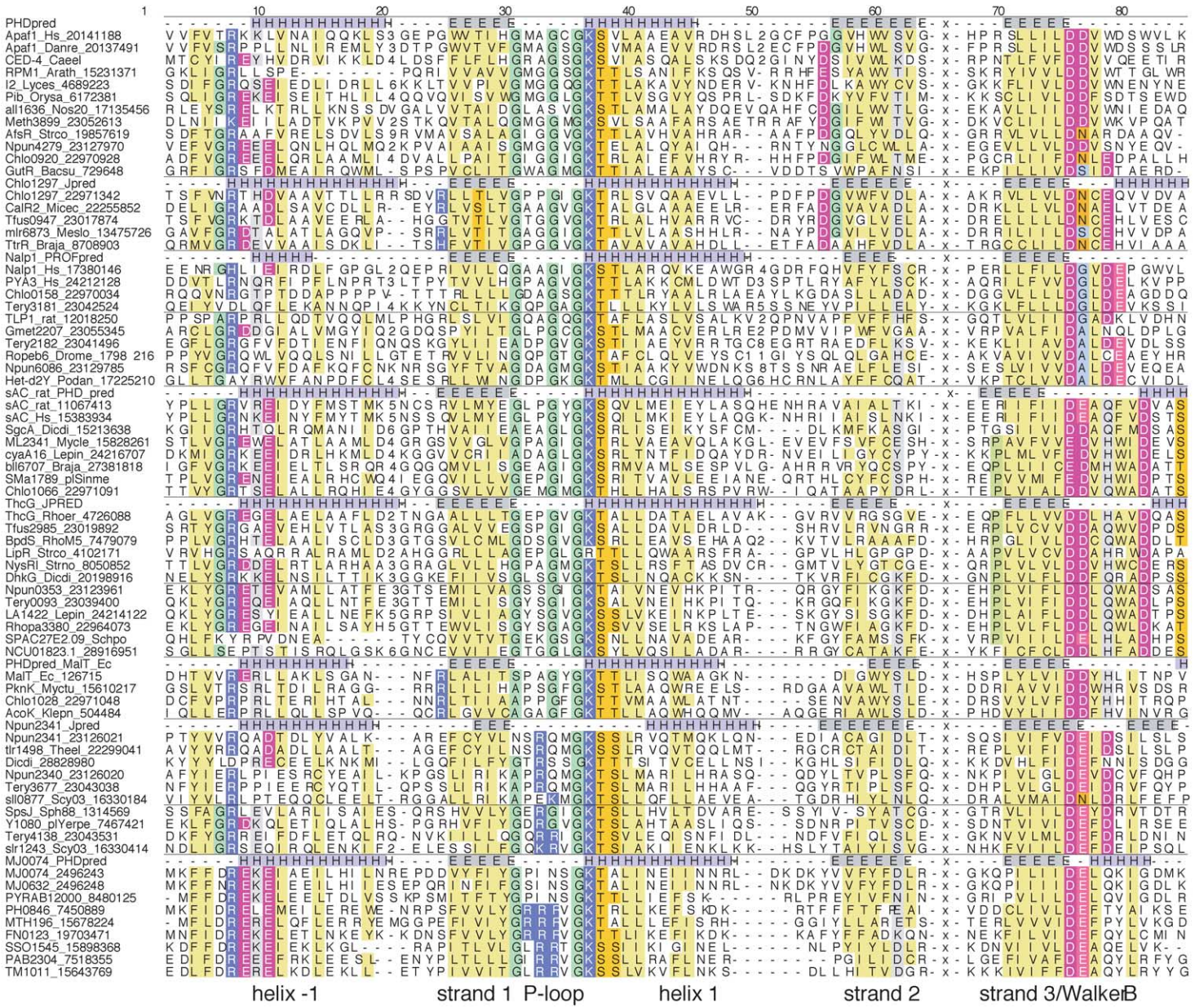

Figure 1a (*legend on p.5*)

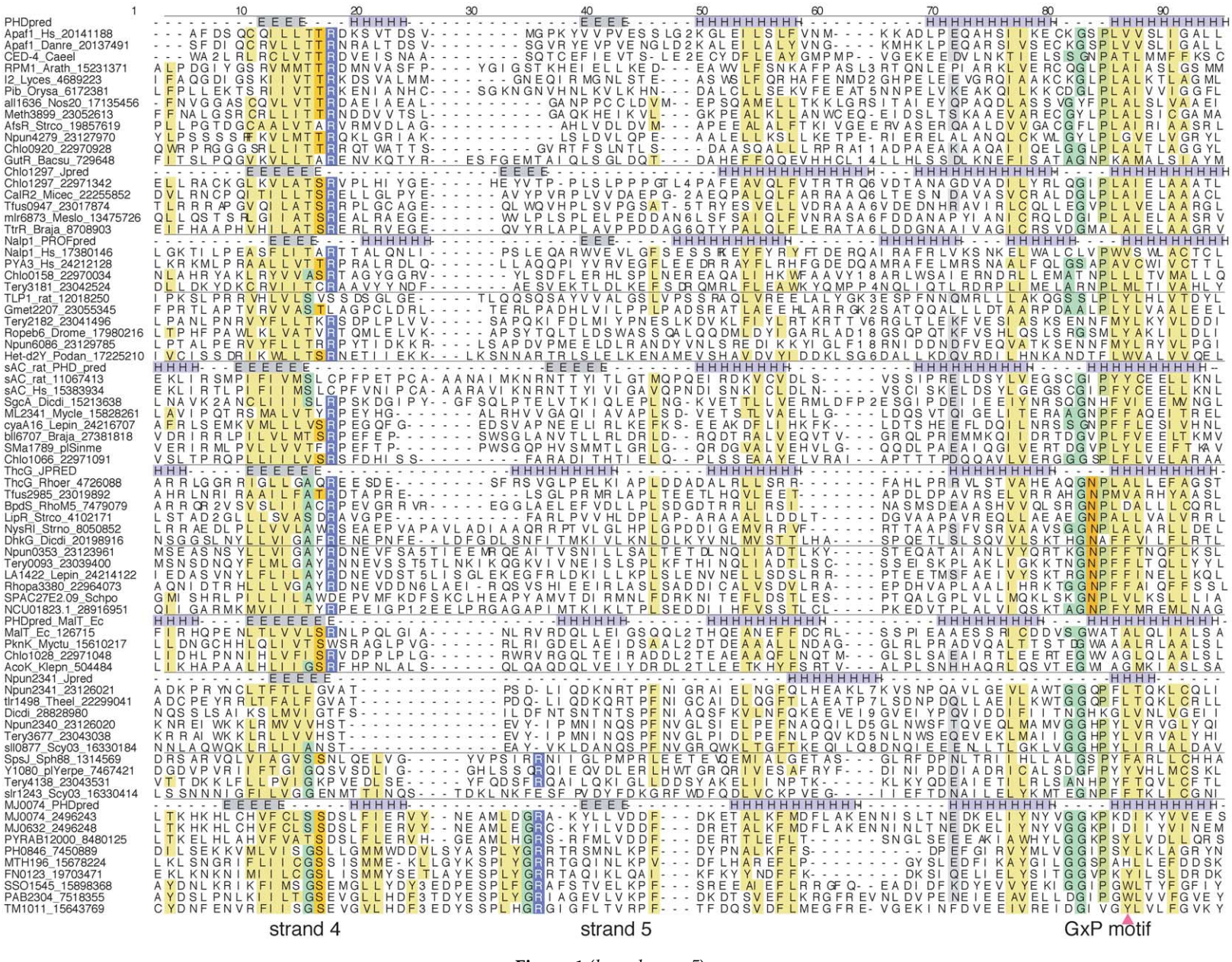

Figure 1 (*legend on p. 5*)

kinases, and AAA+ ATPases; in the process, several previously unnoticed families of (predicted) P-loop NTPases were identified, and many functional predictions were made.[9,13–16] Here, we employ sequence-profile searches, multiple alignment analysis, secondary and tertiary structure predictions, phylogenetic analysis, and contextual information derived from comparative genomics to identify and systematically investigate a major new class of P-loop NTPases within the ASCE division. This class includes the AP-ATPase and NACHT families along with the previously described families of predicted archaeal NTPases[22,23] and numerous other, uncharacterized proteins from diverse organisms. The majority of these (predicted) NTPase domains occur in large, multidomain proteins, which seem to represent a distinct, ancient architectural paradigm in signal transduction process shared by complex life-forms from the three superkingdoms. We named these proteins the signal transduction ATPases with numerous domains (STAND) class of NTPases.

Results and Discussion

STAND NTPases: identification and characterization of the defining sequence and structural features of the class

Previous analyses of the animal apoptotic proteins APAF/Ced4 and plant pathogen-resistance ATPases had defined the AP-ATPase family, with representatives from animals, plants, and several diverse bacterial lineages.[17–20] Likewise, animal proteins involved in inflammatory responses and innate cellular immunity to bacteria, such as NAIP and its vertebrate paralogs, the telomerase subunit TP1, and the fungal heterocaryon incompatibility proteins Het-E-1 defined the NACHT family of NTPases, which also includes uncharacterized bacterial proteins.[20,21] Certain key sequence features shared by these two families of NTPases could be identified through superposition of their respective multiple alignments (Figure 1). Notably, similar features were also detected in two other previously described families of predicted NTPases, the so-called MJ-type and PH-type NTPases, which show lineage-specific expansion in different archaeal species.[22,23] In particular, the NTPase domains of all these families contain a conserved C-terminal region with a predicted helical structure and the characteristic hhGRExE motif located N-terminally of the Walker A motif (Figure 1). The C-terminal regions of all these proteins contain a highly conserved sequence motif with a GxP or GxxP signature (Figure 1 and see below for details). The conservation of these motifs suggested that the respective families form a monophyletic group of P-loop NTPases. We further examined this possibility using sequence profile searches. Position-specific scoring matrices (PSSMs) for each of the aforementioned NTPase families were run against the non-redundant (NR) protein database (National

Figure 1. Multiple sequence alignment of the STAND ATPase and GxP domains. The alignment shows the NTPase domain (from start to strand 5) and the GxP domain that includes the last three helices and the GxP motif. Numbers or a column of the letter x indicate poorly conserved regions that were left out of the alignment. Residues that are widely conserved or discussed in the text are color-coded with light yellow for hydrophobic residues (A, C, I, F, L, M, T, Y, W), green for small residues (G, A, S), light orange for hydroxy residues (S, T), orange for amides (N, Q), blue for basic residues (K, R, H), purple for aspartate, and red for glutamate. Predicted secondary structure elements are shown above the respective sequence (E for strand and H for helix). The red arrowhead indicates the site of a tryptic digestion site in GutR.[32] Sequences are identified with protein name and an organism name abbreviation. Open reading frames have been labeled with the identifier from the /gene, /allele, or /locus_tag field in the GenBank sequence record (where present). In addition, for many sequences, a unique identifier, the GenBank GI number is provided. Organism name abbreviations are shown below: AN, *Aspergillus nidulans*; Arath, *Arabidopsis thaliana*; Avin, *Azotobacter vinelandii*; Bacsu, *Bacillus subtilis*; BL, *Bifidobacterium longum*; Braja, *Bradyrhizobium japonicum*; Bt, *Bacteroides thetaiotaomicron*; Burfu, *Burkholderia fungorum*; Caeel, *Caenorhabditis elegans*; Chlo, *Chloroflexus aurantiacus*; Cioin, *Ciona intestinalis*; Cloac, *Clostridium acetobutylicum*; Clote, *Clostridium tetani*; Chut, *Cytophaga hutchinsonii*; Corgl, *Corynebacterium glutamicum*; Danre, *Danio rerio*; Dicdi, *Dictyostelium discoideum*; Drome, *Drosophila melanogaster*; Ec, *Escherichia coli*; Faci, *Ferroplasma acidarmanus*; FN, *Fusobacterium nucleatum*; Gmet, *Geobacter metallireducens*; Hs, *Homo sapiens*; Klepn, *Klebsiella pneumoniae*; Lepin, *Leptospira interrogans*; Lgas, *Lactobacillus gasseri*; Lyces, *Lycopersicon esculentum*; MA, *Methanosarcina acetivorans*; Magn, *Magnetospirillum magnetotacticum*; Mmc, *Magnetococcus* sp. *MC-1*; Meslo, *Mesorhizobium loti*; Meth, *Methanosarcina barkeri*; MJ, *Methanocaldococcus jannaschii*; Metma, *Methanosarcina mazei*; Mg, *Magnaporthe grisea*; Micec, *Micromonospora echinospora*; MTH, *Methanothermobacter thermautotrophicus*; Mycle, *Mycobacterium leprae*; MYPE, *Mycoplasma penetrans*; Myctu, *Mycobacterium tuberculosis*; Nc, *Neurospora crassa*; Nicgl, *Nicotiana glutinosa*; Niteu, *Nitrosomonas europaea*; Nos20, *Nostoc sp. PCC 7120*; Npun, *Nostoc punctiforme*; Orysa, *Oryza sativa*; PAE, *Pyrobaculum aerophilum*; Pire1, *Pirellula sp. 1*; Podan, *Podospora anserina*; Pflu, *Pseudomonas fluorescens*; Pseae, *Pseudomonas aeruginosa*; PF, *Pyrococcus furiosus*; PH, *Pyrococcus horikoshii*; PAB or Pyrab- *Pyrococcus abyssi*; Reut, *Ralstonia metallidurans*; rat, *Rattus norvegicus*; Rhoer, *Rhodococcus erythropolis*; RhoM5, *Rhodococcus* sp. *M5*; Rhopa, *Rhodopseudomonas palustris*; Rc, *Rickettsia conorii*; Stral, *Streptomyces albus*; Strco or Sco, *Streptomyces coelicolor*; Strhy, *Streptomyces hygroscopicus*; Strno, *Streptomyces noursei*; Strve, *Streptomyces venezuelae*; Schpo, *Schizosaccharomyces pombe*; Sinme, *Sinorhizobium meliloti*; SSO, *Sulfolobus solfataricus*; Scy03, *Synechocystis* sp. *PCC 6803*; Sph88, *Sphingomonas* sp. *S88*; Tfus, *Thermobifida fusca*; Thermoc, *Thermococcus* sp.; Theel, *Thermosynechococcus elongatus*; TM, *Thermotoga maritima*; Tery, *Trichodesmium erythraeum*; Vibch, *Vibrio cholerae*; Yerpe, *Yersinia pestis*. If a gene is reported to be of plasmid origin, the organism name is prefixed by the letters pl, e.g. alr7190_plNos20 means that the corresponding GenBank record identifies the gene as being located on a plasmid of the cyanobacterium Nostoc sp. PCC 7120. A lower case c identifies a chloroplast sequence.

Center for Biotechnology Information, NIH, Bethesda) using the PSI-BLAST program and inclusion thresholds of 10^{-4} (the random expectation or *E*-value) or lower (see Materials and Methods). The Walker A motif was excluded from the alignments for the construction of these PSSMs to avoid the generic attraction of the searches toward large families. In these searches, the sequence of AP, NACHT, MJ, and PH-NTPases detected each other with significant *E*-values. For example, the PH-type NTPase PSSM detected the MJ-type NTPases in iteration 2 ($e=10^{-6}$), the AP-ATPases in iteration 3 ($e=10^{-5}$), and NACHTs in iteration 5 ($e=10^{-4}$).

Additionally, these searches retrieved from the database a variety of proteins, that have not been previously known to contain related NTPase domains. These newly detected proteins included the C-terminal domain of the soluble adenylyl cyclases, the N-terminal domains of nephrocystin-3 and Rolling pebble proteins, and several large, uncharacterized proteins from eukaryotes and Bacteria. The searches were terminated when representatives of previously defined classes of P-loop domains, such as AAA+, KAP, ABC, PilT or VirD, were detected with *E*-values above the cut-off. Reciprocal PSI-BLAST searches with the newly detected relatives of the above families used as queries allowed us to establish their affinities and eliminate representatives of the previously defined classes. For example, searches initiated with the NTPase domain of *Drosophila* rols6 protein (gi17980216) detect the homologs from *Anopheles* and mammals in the first iteration; in the third iteration, numerous HetDE proteins and one cyanobacterial NACHT homolog were detected with $e < 10^{-4}$ and no false positives. Typically, these searches retrieved a consistent set of proteins prior to the encroachment of members of previously defined classes of P-loop NTPases. Most of the functionally characterized members of this distinctive set of P-loop domains are parts of large, multidomain proteins that contain three or more globular domains and participate in diverse signaling processes (see details below). Thus, we named this newly delineated group of P-loop NTPase domains the signal transduction ATPases with numerous domains (STAND) class. In iterated database searches, the sequences of STAND NTPases consistently showed significant similarity to members of the ASCE division, such as AAA+ and ABC ATPases, but not to KG division NTPases. Therefore, we hypothesized that the STAND class belonged to the ASCE division (see also below).

The NTPase domain sequences of the STAND class were grouped into distinct clusters of proteins with highly significant sequence similarity, multiple alignments were generated for each cluster, and characteristic conserved motifs were identified (see Materials and Methods). The alignments of the individual clusters were then combined into a single multiple alignment using these conserved motifs and secondary structure predictions as guides. Secondary structure prediction suggested that the STAND domains have a five-stranded core with the Walker A (GxxxxGK[ST]) motif associated with strand 1 and the Walker B motif associated with the highly conserved strand 3 (Figure 1). In the majority of these domains, with the exception of the NACHT family, the Walker B motif contains two conserved, successive acidic residues. In the NACHT NTPases, the second acidic residue is missing, but another conserved aspartate is present three positions downstream of the first one (Figure 1). By analogy to other P-loop NTPases, the proximal aspartate is predicted to coordinate the Mg^{2+} cation. The second acidic residue (aspartate or glutamate in different families of the STAND class) is likely to function as a proton-abstracting moiety similarly to the conserved glutamate of NTPases of the ASCE divisions.[24–29] Strand 4 of the STAND class contains a conserved polar residue at the C terminus (Figure 1). An equivalent conserved residue is seen throughout the ASCE division and corresponds to the Sensor-I motif of the AAA+ superclass[14,15] and the [ST][AG][ST] motif of the superfamily I and II helicases.[30] These conserved features, along with the preferential retrieval of the AAA+ and ABC NTPases in searches with profiles for the STAND class, support the classification of the STAND NTPases in the ASCE division (Figure 2).

The STAND NTPases are defined by several sequence and architectural features that set them apart from other ASCE division NTPases (Figures 1 and 2). The most diagnostic ones are the aforementioned hhGRExE motif ahead of the Walker A motif and the GxP motif located C-terminally of the NTPase domain (Figure 1). A comparison of the predicted secondary structure of the STAND NTPase domain with the structures of known ASCE NTPases suggests that the GxP motif is associated with a helical bundle located to the C terminus of the core P-loop domain (Figure 2); we refer to this domain as the GxP module. The STAND NTPases also contain a less conserved sequence feature associated with strand 4, with the signature hhh[GST][ST]R seen in many sequences (Figure 1). In the NACHT family, two conserved motifs have been noticed at the C terminus of the GxP module (motifs VI and VII[20]) (Figures 1–4). We explored this region further and found that most STAND ATPases, with the exception of some members of the MJ- and PH-type families,[31] contain a region of ~200 amino acid residues (gray hexagon in Figure 3) between the GxP module and the C terminus of the protein or the N terminus of any additional domains that may be present (Figure 3). This region is predicted to adopt a globular fold with six helices that appear to be equivalent in all STAND NTPases (Figure 4). While there is little sequence conservation in this region throughout the STAND class, the sequences within individual families are notably conserved (Figure 4), suggesting that this domain might be important for family-specific functions. Consistent with these

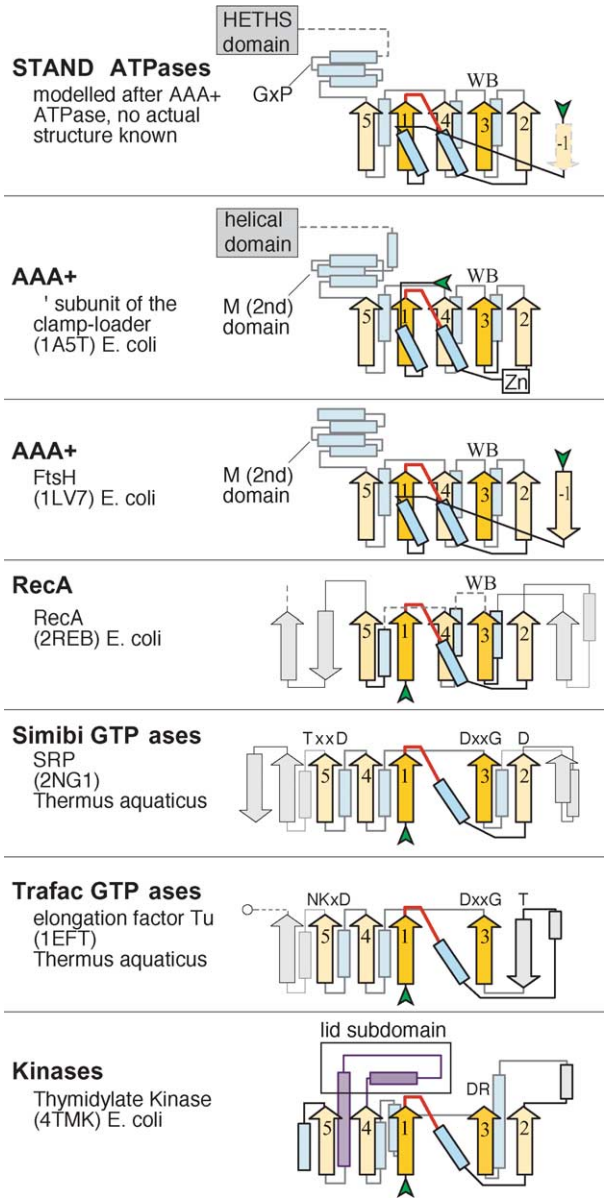

Figure 2. Topology diagrams of domains representative of the major divisions of the P-loop fold. Strands are shown as arrows with the arrowhead on the C-terminal side. Strands 1 and 3 that encompass the conserved sequence motifs GxxxxGK[ST] (Walker A) and hhhh[DE] (Walker B) are rendered in orange; the other core strands (2, 4, 5) are in light orange; non-conserved structural elements that might have been absent from the ancestral P-loop NTPase 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 green arrowhead marks the N terminus of the kinase domain, and the kinase lid subdomain is rendered in purple. Broken lines indicate secondary structure elements that are not present in the PDB file or that were left out for clarity. The gap between strands 1 and 3 in GTPases and kinases was introduced for presentation purposes only. No experimentally determined structure is available for any STAND NTPase; thus, the topology diagram is modeled after the AAA+ ATPases.

observations, the size of the products of limited proteolysis of two STAND ATPases, GutR and MalT, was compatible with cleavage occurring between the GxP module and the C-terminal (predicted) helical domain.[32,33] Thus, the GxP module in most STAND NTPases appears to be followed by a distinct helical domain, which we named the helical third domain of STAND proteins (HETHS) domain (Figure 4). Database searches with the HETHS domain PSSM did not detect significant similarity to any other known protein domains.

Evolutionary classification, phyletic patterns, and domain architectures of the STAND NTPases

We developed an evolutionary classification of the STAND NTPase domains by combining different types of information. Firstly, at the lowest level, conventional phylogenetic analysis was employed to reconstruct the evolutionary history of distinct groups of STAND domains that were identified by similarity-based clustering. This analysis helped in delineating orthologous groups and lineage-specific expansions of paralogs. Secondly, conserved sequence motifs, including those in the C-terminal HETHS domain, were treated as shared derived characters (synapomorphies) to establish higher order relationships between families. Finally, phyletic patterns of orthologous sets and families and domains architectures were compared to infer the likely evolutionary scenarios. This analysis resulted in identification of five major clades of STAND NTPases (Table 1). In this section, we briefly describe the reconstructed evolutionary history, domain organization, and (predicted) functions of STAND NTPases according to this classification.

AP-ATPase clade

The AP-ATPase clade is typified by a conserved aspartate N-terminal of strand 2, the hhhToR signature (o designates an alcoholic residue) in strand 4, and a conserved serine and the hxhHD motif in the HETHS domain (Figures 1 and 4). AP-ATPase domains are often associated with C-terminal superstructure-forming domains, such as WD40, LRR, or TPR repeats (Figure 3), and N-terminal DNA-binding HTH domains or protein–protein interaction domains, such as DEATH-like six-helix domains and TIR domains. This clade consists of two major families, the classic AP-ATPases and CalR2.

The animal members of the classic AP-ATPase family, including nematode CED4 and mammalian Apaf-1, are involved in cell-death signaling by activating caspases.[34] ATP-binding triggers Apaf-1 oligomerization and association with procaspase-9, resulting in the formation of the "apoptosome".[35,36] Most of the plant disease-resistance proteins, which consist of the AP-ATPase domain fused with

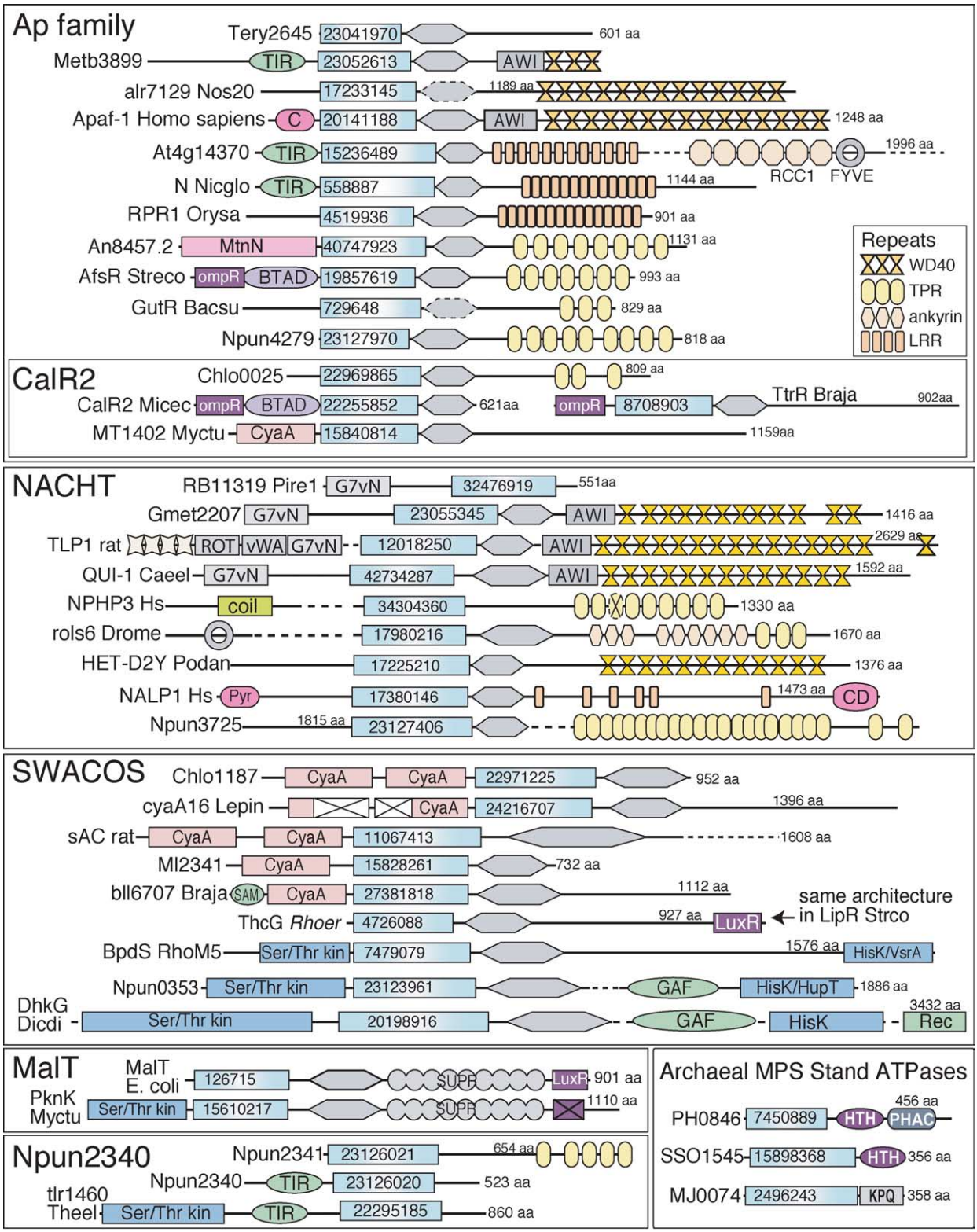

Figure 3. Domain architectures of selected STAND ATPases. Proteins are represented as horizontal lines, and rectangles or other geometric shapes indicate conserved domains. A blue rectangle shows the core ATPase together with the C-terminal GxP domain, and a gray stretched hexagon shows the HETHS domain. The HETHS domains in GutR and Nost7129 are highly diverged and accordingly shown with broken lines. LuxR and OmpRtype HTH domains and the HTH domain in the Archaea are shown in purple. Proteins are identified with the protein name, organism name abbreviation and the GenBank identifier in the ATPase box. Domain designations: AWI-ATPase-WD40 intervening domain, a small domain of unknown function that precedes the WD40 repeats in some AP and NACHT ATPases; Btad, Btad domain;[133] CyaA, adenylate cyclase; X1, a small domain of unknown function found in some NACHT ATPases; HisK, histidine kinase; LLR, leucine-rich repeat; LuxR, LuxR-type HTH domain; OmpR, OmpR-type HTH domains; MtnN, 5′-methylthioadenosine nucleosidase; TPR, tetratricopeptide repeat; Rec, receiver domain of response regulator;

Table 1. Classification of STAND NTPases

A. *AP-ATPase clade*
Aspartate at N terminus of strand 2, hxhhhT[ST]R in strand 4, conserved serine in HETHS domain, often with LRR, WD40 or TPR repeats at C terminus
  Several families in Eukaryota and Bacteria including animal Apaf-1 and CED-4, green plant disease-resistance proteins, and several bacterial families, such as AfsR and GutR, mostly represented in Cyano- and Actinobacteria
  CalR2 family, N-terminal OmpR-type HTH domain, in Actinobacteria, *Chloroflexus*, and some Proteobacteria
B. *NACHT clade*
Second acidic residue in Walker B replaced by a tiny residue (G, A, or S), with C-terminal WD40, TPR, LRR or ankyrin repeats
  NAIP-like family
    Caterpillar subfamily: N-terminal pyrin or CARD domains, includes CIITA, Nod1, Nod2, Nalp1, Nalp2 among others
    Npun3725/Chlo0158 subfamily (bacterial and two archaeal homologs)
  TLP1-like family
    TLP1 subfamily (TLP1, NPHP3, QUI-1; Metazoa and *Geobacter*)
    HetDE subfamily (some ascomycetes)
    Rolling pebbles (Metazoa)
    Npun6086 (Cyanobacteria)
In addition, this family includes several uncharacterized proteins from diverse bacteria, from e.g. *Ralstonia* and *Cytophaga*
C. *SWACOS clade*
Characteristic Walker B signature: PhhhhhDDh[HQ]hhDxxS, middle residue in GxP motif often asparagine, contain adenylyl cyclase or Ser/Thr kinase domain
  sAC/Chlo1187 family, N-terminal CyaA domain (Metazoa, *Dictyostelium*, *Chloroflexus*, α-proteobacteria)
  LipR/ThcG family (Actinobacteria)
  DhkG/Npun0353 family, N-terminal Ser/Thr kinase domain, C-terminal His kinase (Cyanobacteria, *Dictyostelium*, *Leptomonas*, and the Alphaproteobacteria *Magnetospirillum*, and *Rhodopseudomonas*)
D. *MalT clade*
G missing in hhGR motif, acidic second strand SlDxxD, SUPR and HTH/LuxR domain at C terminus, includes MalT, AcoK, AlkS, PknK (α-proteobacteria, Actinobacteria, *Chloroflexus*)
E. *MNS clade*
Arginine in P-loop, no arginine in strand 4, glutmate in Walker B, often two glycine residues in GxP motif
  MJ-type family, ATPase+KPQ domain
  PH-type family, ATPase+HTH+PHAC domain
  SSO-type family, ATPase+HTH domain
  Npun2340/2341 family, N-terminal TIR or C-terminal TPR domain (Cyanobacteria, *Clostridium*, *Dictyostelium*)
  SpsJ family, hhhhDE[YF]D in Walker B, no large additional domains (*Sphingomonas*, *Yersinia*, Cyanobacteria)
  BL0662 family, no additional domains (Actinobacteria and *Lactobacillus*)

This is an abbreviated representation of the evolutionary classification of the STAND NTPases; see the main text for more detailed descriptions.

C-terminal leucine-rich repeats (LRRs) and N-terminal TIR domains or coiled-coil regions (Figure 3), function as pathogen recognition proteins that trigger the death of pathogen-infected cells.[37–40] A small subgroup of plant AP-ATPases contain β-propeller-forming RCC1 repeats and lipid-binding FYVE domains[41] at their extreme C terminus (Figure 3). While most plant members of this family are implicated in pathogen response, at least one plant AP-ATPase, the maize PSiP protein, has a developmental function in pollen tube orientation.[42] PSiP has also been reported to have adenylate cyclase activity. Although some STAND NTPases contain adenylate cyclase domains (see below), we did not detect an adenylate cyclase domain in the published PSiP sequence (GI:15387663). Furthermore, this claim is questionable, because none of the known P-loop domains is known to have nucleotide cyclase activity and, moreover, this reaction appears to be inconsistent with the chemistry of the phosphohydrolase reactions typically catalyzed by P-loop-fold proteins. Among the bacterial AP-ATPases, AfsR is a transcription factor involved in both regulation of secondary metabolism and in morphological differentiation in *Streptomyces*,[43,44] whereas the *Bacillus subtilis* GutR protein regulates expression of the glucitol dehydrogenase gene GutB.[32,45]

Prokaryotic AP-ATPases are represented mostly in Actinobacteria and Cyanobacteria, and sporadically in other bacterial and archaeal lineages (Table 2). Among eukaryotes, APATPases are found in animals and plants, and in fungi with large genomes, such as *Neurospora*, *Aspergillus* and *Magnaporthe*. Animals have a single orthologous group of APATPases, which includes CED-4, DARK-1, and APAF-1. In plants, AP-ATPases have undergone extensive lineage-specific expansion, with more than 200 paralogs in *Arabidopsis* and ~800 in rice.[18,46,47] This proliferation appears to be related to the diversification of the C-terminal LRRs, which show specificity toward different pathogens.[48–50] As noticed,[21] animal and plant AP-ATPases form a well-supported clade in phylogenetic trees (Figure 5a). This clade clusters with a subset of prokaryotic AP-ATPases from *Methanosarcina*

Ser/Thr kin, serine/threonine kinase; SUPR, superhelical peptide repeats in MalT related to TPRs;[33] TIR, Toll/interleukin-1-like receptor. Organism name abbreviations are as in Figure 1.

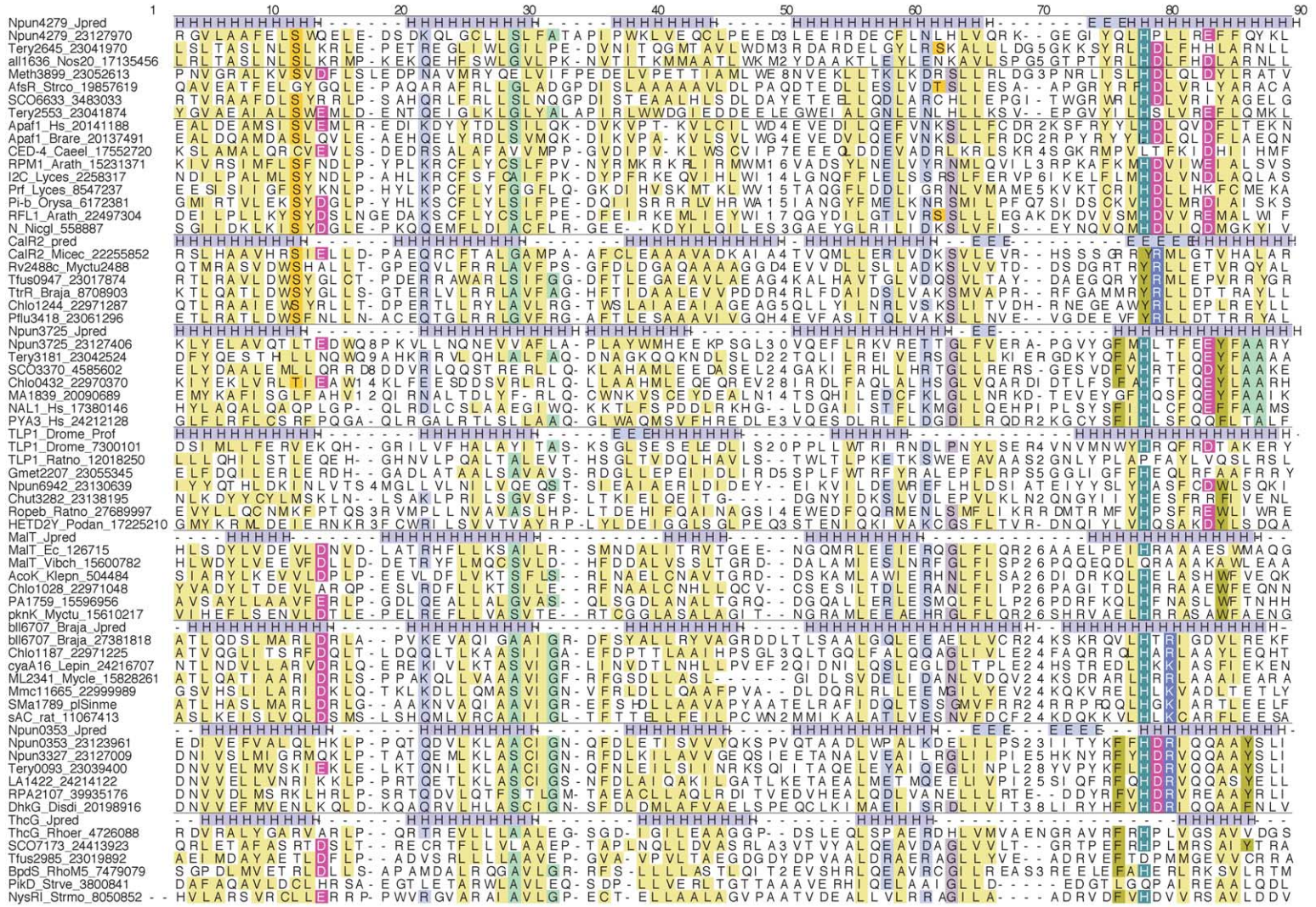

Figure 4. Multiple sequence alignment of the HETHS domains. Sequences are aligned on the basis of sequence similarity for four of the five major groups: Apaf-1/AfsR/CalR2, MalT, Swacos, and Npun2340/2341. There is no appreciable sequence conservation between the groups in this region, and the alignment largely follows predicted secondary structure elements.

Table 2. Phyletic distribution and lineage-specific expansion of STAND NTPases

Family	Actinobacteria	Cyanobacteria	Proteobacteria	Other bacteria	Archaea	Eukaryota
AP-ATPases AfsR/ GutR	Tfus:2, Myctu:1, Strco:5	Npun:6, Nos20:3, Tery:5, Crowa:1	–	Chlo:1, Bacsu:1	Metba:1, Pyrho:1	Some Metazoa, many green plants, Asco-mycetes: Mg:6, An:11, Nc:4
CalR2 family	Mycle:1, Myctu:5, Tfus:2, Strco:4, Rhoer:1	–	Braja:8	Chlo:3	–	–
NACHT–NALP-like	Myctu:0, Strco:2, Tfus:0	Npun:3, Nos20:9, Tery:3	Ricco:1, MagC1:1	Chlo:3, Cythu:1	Metba:1	In mammals (more than 20 paralogs in human) and the tunicate *Ciona*
NACHT–TLP1-like	–	Npun:4, Nos20:0, Tery:1, Crowa:2	Geome:1, Ralme:1	Chlo:0, Cythu:1	–	Podan:3, Neucr:12, Hs:2
SAC/Chlau1187 family	Tfus:0, Myctu:0, Mycle:2, Strco:0	–	Braja:3, Rhopa:1, Meslo:2, Sinme:4, MacC1:1	Chlo:6, Lepin:1	–	Mammals and Dicdi (no other eukaryotes)
LipR/ThcG	Myctu:1, Strco:9, Tfus:1, RhoM5:1	–	–	–	–	–
Npun0353 DhkG	–	Npun:13, Nos20:13, Tery:4, Scy03:0	Magma:1, Rhopa:2, Ralme:1	Lepin:3	–	Dicdi:1, Neucr:1, Schpo:2
MalT	Tfus:0, Myctu:1, Strco:1	–	Ralme:6, Pseae:4, Ecoli:1, Vibch:1	Chlo:4, Claab:1	–	–
Npun2340	–	Npun:8, Nos20:3, Tery:5, Theel:1, Scy03:1	–	Clote:2	–	Dicdi:1
Npun2341	–	Npun:4, Nos20:3, Tery:1, Theel:1, Scy03:1	–			
SpsJ	–	Npun:2, Nos20:1, Tery:6, Scy03:1	Sph88:1, Yerpe:1	–	–	–
Sso1545	–	–	–	Thema:1	Pyrab:9, Pyrho:4, Metac:1, Pyrfu:2, Sulso:8, Pyrae:2	–
Ph0846	–	–	–	Fusnu:1, Biflo:2, Geome:1	Pyrho:9, Pyrfu:8, Pyrab:6, Metth:1, Metac:2, Metja:1, Metma:1, Pyrae:2, Sulso:1	–
Mj0074	–	–	–	–	Metja:17, Pyrab:3, Pyrho:2	–

The number of detected members of the given family of STAND NTPases in species with completely sequenced genomes from the respective taxa is shown here; for eukaryotes, the range of taxa in which the respective family is represented is given instead or in addition. Species name abbreviations: Arath, *Arabidopsis thaliana*; Bacsu, *Bacillus subtilis*; Biflo, *Bifidobacterium longum*; Braja, *Bradyrhizobium japonicum*; Caeel, *Caenorhabditis elegans*; Chlo, *Chloroflexus aurantiacus*; Clote, *Clostridium tetani*; Crowa, *Crocosphaera watsonii*; Cythu, *Cytophaga hutchinsonii*; Danre, *Danio rerio*; Dicdi, *Dictyostelium discoideum*; Drome, Drosophila melanogaster, Ec, *Escherichia coli*; Fusnu, *Fusobacterium nucleatum*; Geome, *Geobacter metallireducens*; Hs, *Homo sapiens*; Meslo, *Mesorhizobium loti*; Mycle, *Mycobacterium leprae*; Myctu, *Mycobacterium tuberculosis*; Nos20, *Nostoc sp. PCC 7120*; Klepn, *Klebsiella pneumoniae*; Lepin, *Leptospira interrogans*; Lyces, *Lycopersicon esculentum*; MagC1, *Magnetococcus sp. MC-1*; Metba, *Methanosarcina barkeri*; Metth, *Methanothermobacter thermautotrophicus*; Metja, *Methanocaldococcus jannaschii*; Metma, *Methanosarcina mazei*; Micec, *Micromonospora echinospora*; Neucr, *Neurospora crassa*; Nicgl, *Nicotiana glutinosa*; Npun, *Nostoc punctiforme*; Orysa, *Oryza sativa*; Podan, *Podospora anserina*; Psefl, *Pseudomonas fluorescens*; Pseae, *Pseudomonas aeruginosa*; Pyrho, *Pyrococcus horikoshii*; Pyrab, *Pyrococcus abyssi*; Pyrae, *Pyrobaculum aerophilum*; Ralme, *Ralstonia metallidurans*; Rhoer, *Rhodococcus erythropolis*; RhoM5, *Rhodococcus sp. M5*; Rhopa, *Rhodopseudomonas palustris*; Ricco, *Rickettsia conorii*; Sph88, *Sphingomonas sp. S88*; Strco, *Streptomyces coelicolor*; Strno, *Streptomyces noursei*; Strve, *Streptomyces venezuelae*; Schpo, *Schizosaccharomyces pombe*; Sinme, *Sinorhizobium meliloti*; Sulso, *Sulfolobus solfataricus*; Scy03, *Synechocystis sp. PCC 6803*; Tfus, *Thermobifida fusca*; Theel, *Thermosynechococcus elongatus*; Trier, *Trichodesmium erythraeum*; Yerpe, *Yersinia pestis*; Vibch, *Vibrio cholerae*. If the gene is reported to be of plasmid origin, the organism is prefixed by the letters pl.

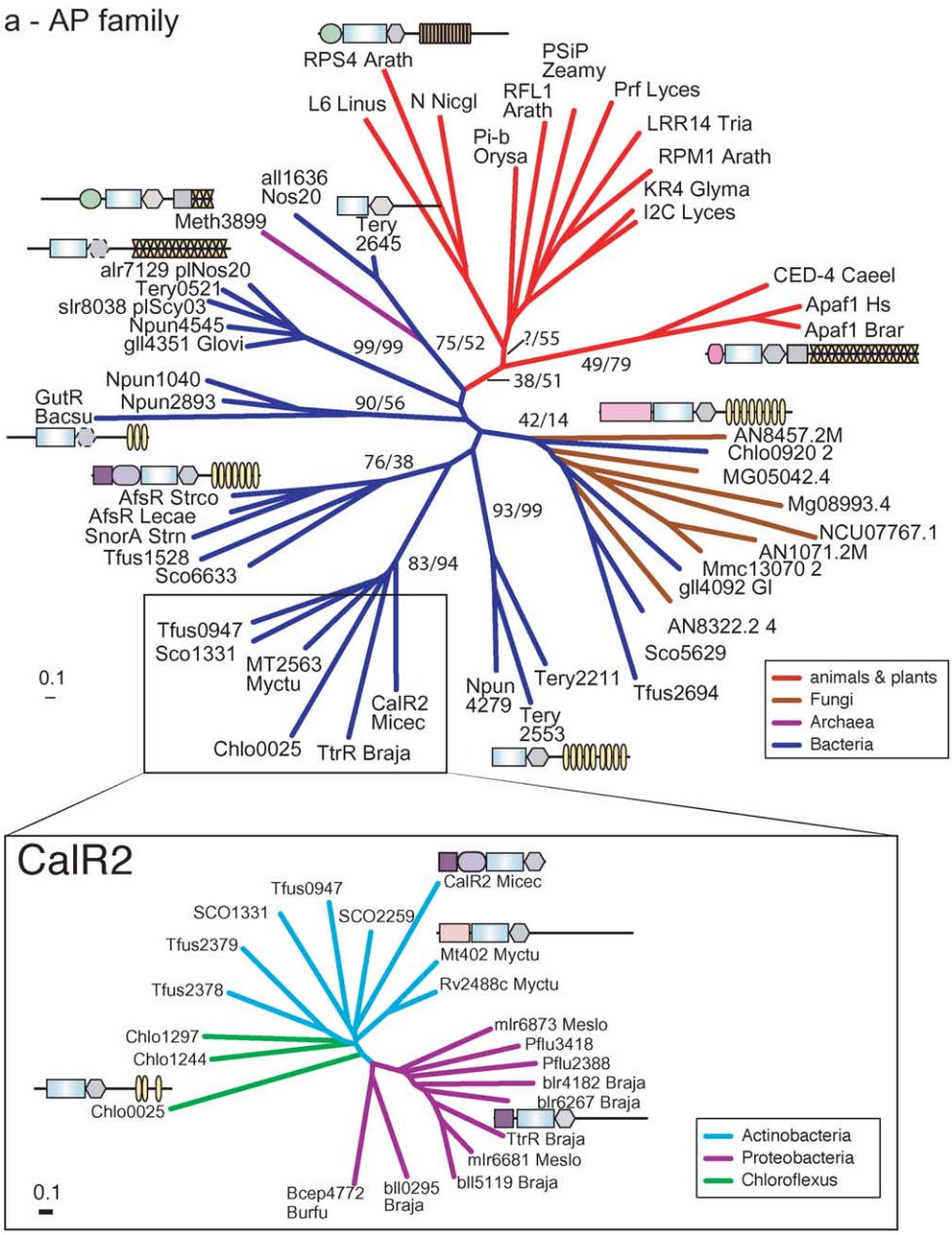

Figure 5a (*legend on p.15*)

barkeri (Meth3899, g23052613) and Cyanobacteria, but not with the fungal members of the family (Figure 5a). The fungal AP-ATPases show small lineage-specific expansions in various filamentous ascomycetes, with ~4–12 paralogs encoded in these genomes (Table 2). These fungal AP-ATPases are often fused to N-terminal domains of the purine nucleoside phosphorylase/S-adenosine homocysteine nucleosidase fold. In phylogenetic trees, fungal APNTPase domains are nested within a cluster with representatives from diverse Bacteria (Figure 5a). Phylogenetic analysis also identified several subfamilies of AP-ATPases that were present only in the Bacteria. These include the AfsR subfamily that is widespread in Actinobacteria and a subfamily that is found exclusively in Cyanobacteria (Figure 5a). These phyletic patterns and phylogenetic affinities suggest that the AP-ATPases attained their diversity in Bacteria and were sporadically acquired by eukaryotes and Archaea *via* horizontal gene transfer (HGT). The transfer to eukaryotes appears to have occurred on at least two occasions: to the common ancestor of the clade that includes animals and plants, giving rise to the animal and plant regulators of programmed cell death, and, independently, to an ancestral filamentous fungus. Given that current phylogenetic models favor an animal-fungal clade,[51] to the exclusion of plants, the representatives of the "CED-4 clade" probably have been lost in fungi.

The CalR2 family is characterized by the D[NST]XE consensus in the Walker B motif, by the presence of a conserved arginine and a hydroxy residue before the P-loop (consensus RxxoxxGxxxGko),

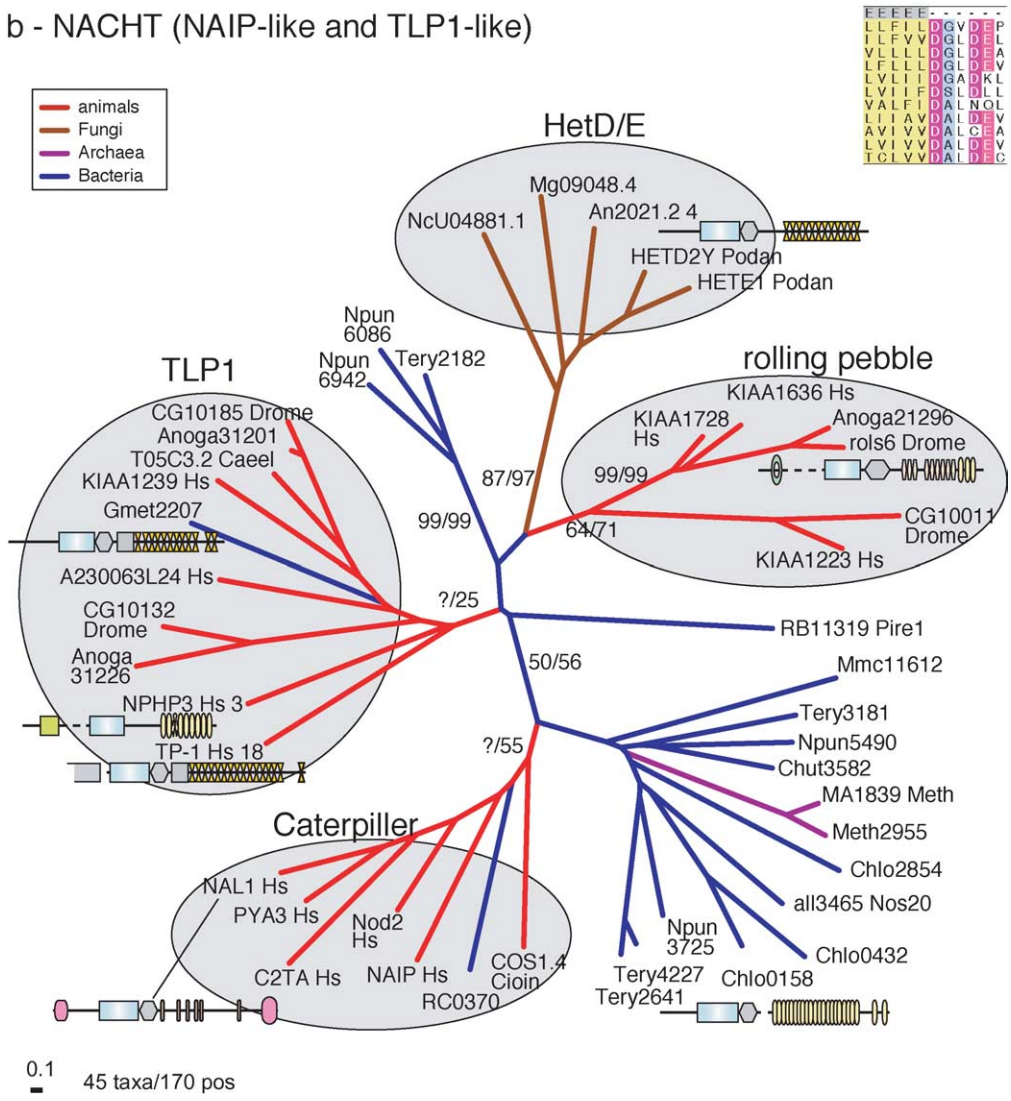

Figure 5b (*legend on p.15*)

and a conserved acidic residue in strand 2 (Figure 1 and Table 1). The family is typified by the CalR2 protein, a putative regulator of the *Micromonospora echinospora* calicheamicin metabolism locus.[52] This exclusively bacterial family is represented in Actinobacteria, *Chloroflexus*, and some Proteobacteria with larger genomes, such as *Bradyrhizobium*, *Mesorhizobium*, and *Pseudomonas*, but is thus far absent from the Cyanobacteria. Most of the genomes that encode predicted NTPases of this family have more than one paralog (Table 1), suggesting multiple, small lineage-specific expansions in various bacterial lineages. Most members of this family contain an OmpR-type HTH domain at the N terminus. Interestingly, one of the few members of this family that lack the HTH domain (*Mycobacterium tuberculosis*) contains an N-terminal adenylyl cyclase domain, a feature that is otherwise characteristic of another group of STAND ATPases, the sAC/Chlo1187 family (see below and Figure 3). The presence of the HTH domain suggests that most of the predicted NTPases of this family are transcriptional regulators.

The NACHT clade

This clade was originally named after its representatives, neuronal apoptosis inhibitor protein NAIP, MHC class II transcription activator CIIA, heterokaryon incompatibility factor HET-E, and telomerase-associated protein TLP1 and represent a second clade of STAND NTPases involved in animal apoptosis.[20] The NACHT family is defined by a "tiny" residue (glycine, alanine or serine) directly C-terminal of the Mg^{2+}-coordinating aspartate, and two additional acidic residues one position downstream (consensus hhhhD[GAS]hDE with slight variations)[20] (Figure 1). Similarly to the AP-ATPases, the C termini of NACHT NTPases often contain repeats, such as WD40, TPR, LRR or ankyrin, that form periodic superstructures (Figure 3). High sequence divergence within the NTPase and HETHS domains hamper phylogenetic analysis of the NACHT clade; nevertheless, distinct NAIP-like and TLP1-like families can be defined (Figure 5b). Human C2TA and *Podospora anserina* HET-E, each representing one of the two NACHT families,

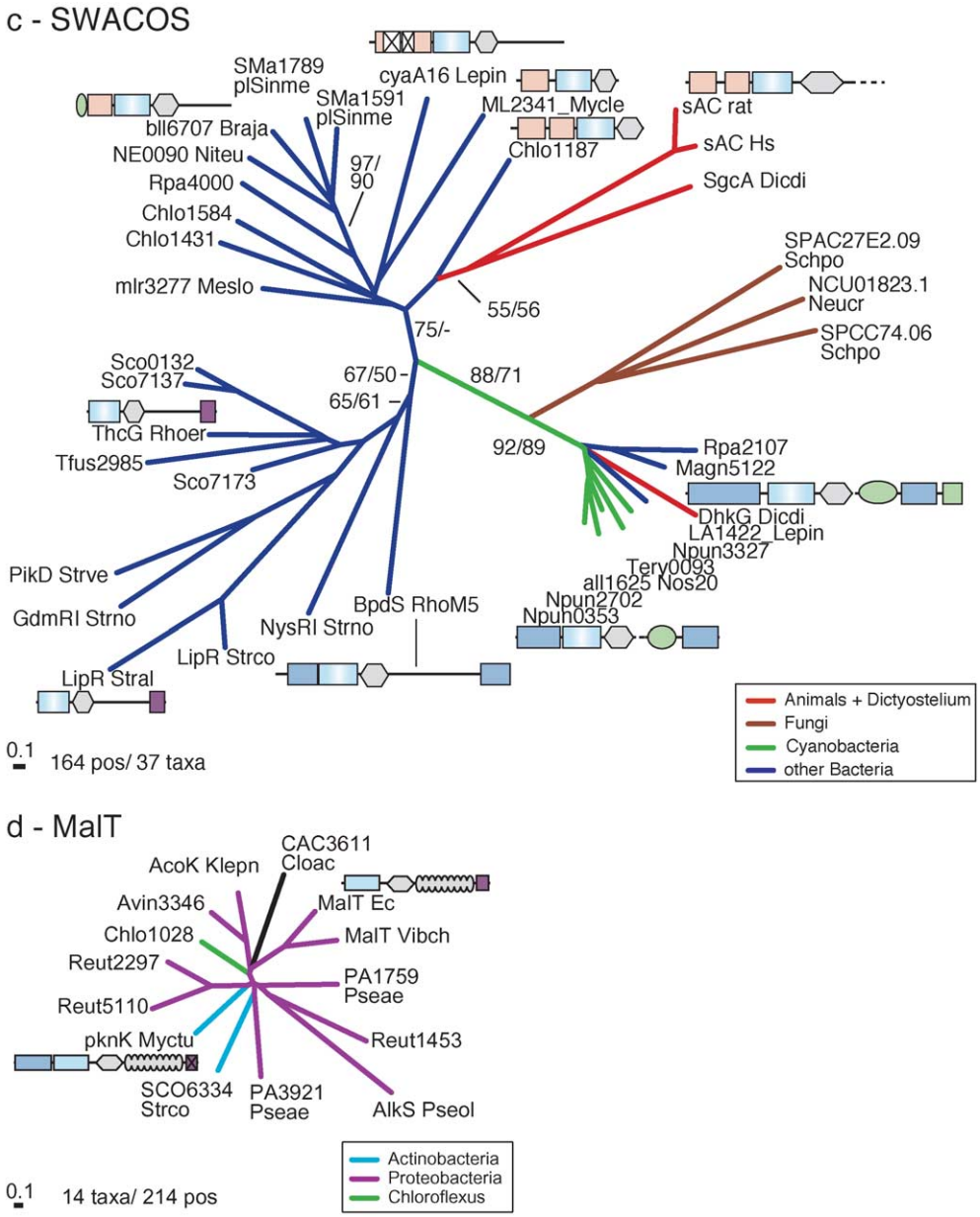

Figure 5c and d (*legend on p.15*)

show specificity for GTP over ATP.[53,54] This suggests that preference for GTP might be a widespread feature of NACHT NTPases, although functional implications of this specificity remain unclear.

The NAIP-like family is defined by a motif with the consensus FhHxxQE[YF]hxA that is found in the HETHS domain. The family consists of two distinct subfamilies. The first subfamily includes the so-called vertebrate caterpillar proteins (CIITA, Nod1/CARD4, Nod2/Card15, DEFCAP/CARD7/NALP1, CIAS1, and their close paralogs) and related proteins from the urochordate *Ciona intestinalis* (Figure 5b). The mammalian forms typically contain C-terminal LRR repeats and N-terminal pyrin or CARD domains, and perform diverse functions related to immune and inflammatory responses and possibly regulation of neuronal apoptosis (reviewed by Harton *et al.*[55]). Humans have 22 paralogous genes for caterpillar proteins, which appear to have evolved in a relatively recent lineage-specific expansion at some point during the diversification of vertebrates. In phylogenetic trees, a protein from the intracellular pathogenic bacterium *Rickettsia connori* that lacks the C-terminal LRRs is nested among the vertebrate NAIP-like proteins, which suggest a late HGT from the animal host to bacterial parasite. The second subfamily is typified by the Chlo0432 protein from *Chloroflexus*. This subfamily seems to be entirely prokaryotic in its distribution and is present in Cyanobacteria, Actinobacteria, *Chloroflexus*, a few other Bacteria (*Magnetococcus*, *Cytophaga*), and in a single archaeon, *Methanosarcina*. Most of these proteins contain numerous C-terminal TPR repeats (Figure 3).

The TLP1-like family shows some conservation of

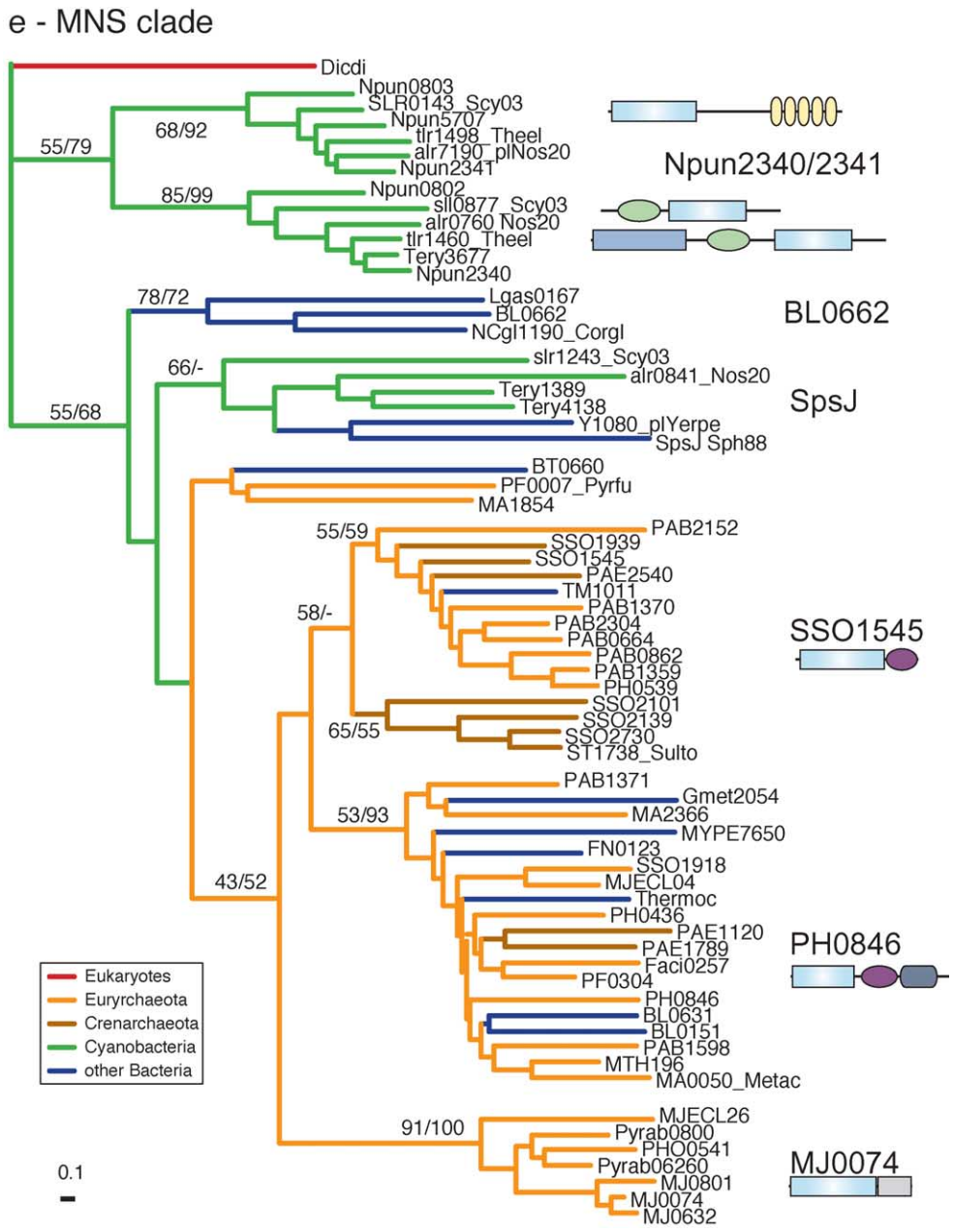

Figure 5. a–e, Unrooted phylogenetic trees of selected STAND NTPase families. The scale bar represents the number of inferred substitutions per 100 sites (amino acid residues). Support for major branches is indicated by percentage bootstrap probabilities for 1000 replications of PHYLIP Protdist/Fitch distance (numerator) and PAUP maximum parsimony analysis (denominator) with the exception of b, where the distance bootstrap number has been computed with the MEGA2 program. The tree branches for Cyanobacteria are in green, other bacterial branches are in blue, eukaryotic branches are in red, and archaeal branches are in purples. Branches are black if phylogenetic origin is uncertain. *N* is the number of alignment positions used for tree analysis/number of organisms in alignment. Organism name abbreviations are as in Figure 1. Domain diagrams are abbreviated, for color code explanation, GenBank identifiers, and domain names, see Figure 3.

the N-terminal hhGR motif that is not present in the NAIP-like family and a number of its members tend to associate with numerous C-terminal WD40 repeats that are predicted to form multiple stacks of β-propeller structures (Figures 1 and 3). Several distinct subfamilies can be identified within the TLP1-like family (Table 2 and Figure 5b). One of these subfamilies is typified by the mammalian telomerase subunit TLP1, which has a distinct N-terminal domain, which is also seen as a stand-alone protein in several Bacteria. In addition to TLP1, we identified three other paralogs of this subfamily in vertebrates and two in insects. One of the vertebrate paralogs is the recently characterized protein nephrocystin, which is mutated in polycystic kidney disease.[56,57] This protein combines an N-terminal NACHT NTPase domain with C-terminal TPR repeats, instead of the WD40 repeats, which are more common in this family. *Caenorhabditis elegans* QUI-1 also contains a C-terminal WD40

domain but has a divergent sequence of the NTPase domain, including the substitution of the tiny residue in the Walker B motif by a second aspartate. QUI-1 functions in the sensory pathway that mediates the worm's avoidance reaction to "bitter" substances like quinine and SDS.[58] The TLP1-like family also includes prokaryotic representatives from the α-proteobacterium *Geobacter metallireducens* and the planctomycete *Pirellula*.

A second subfamily includes members of lineage-specific expansions in the ascomycete fungi and is typified by the vegetative incompatibility proteins, HET-D and HET-E, of *Podospora anserina*.[59,60] These proteins contain a unique N-terminal globular domain so far detected only in filamentous fungi.

A third subfamily is typified by the animal-specific NACHT NTPase domain that we detected in the *Drosophila* Rolling pebbles gene product. This protein has a RING finger domain N-terminal to the NTPase domain and C-terminal ankyrin repeats (Figure 3). Rolling pebbles has been implicated in myoblast morphogenesis as a potential regulator of the fusion of muscle precursor cells.[61,62] A single representative of this subfamily is observed in nematodes, whereas arthropods and vertebrates have two and three paralogs, respectively. In addition, there is a small cyanobacterial group related to *Nostoc punctiforme* Npun6086 (GI:23129785), with several homologs in *Nostoc*, *Crocosphaera*, and *Trichodesmium* that we could not confidently place with one of the other TLP1-like subfamilies (Figure 5b).

Additionally, a small bacterial subfamily of NACHT NTPases that are not specifically related to neither of the two above families was detected in *Cytophaga* and *Ralstonia*. The most striking features of the NACHT NTPases are their widely scattered phyletic patterns, the major lineage-specific expansion of the caterpillar proteins in vertebrates, and frequent substitutions of the C-terminal repetitive domains. The simplest explanation for the phyletic patterns of NACHTs seems to be rampant horizontal mobility. In particular, there were probably multiple cross-kingdom transfers between Bacteria and eukaryotes[21] and see discussion below). Given that the C-terminal repeats are likely to mediate protein–protein interactions, their substitution might have resulted in diversification of the target specificity of these NTPases.

The SWACOS clade

The great majority of proteins in this clade display N-terminal fusions with either an adenylyl cyclase or a Ser/Thr protein kinase (Figure 3). Accordingly, we named this subdivision of the STAND class the SWACOS clade (for STAND with adenylyl cyclase or Ser/Thr protein kinase domains). Unlike most of the other STAND NTPases, those in the SWACOS clade do not contain C-terminal LRR, WD40, or TPR repeats (Figures 3 and 4). The SWACOS clade has a characteristic signature in the Walker B motif: PhhhhhDDh[HQ]hhDxxS (Figure 1). In addition, the variable residue of the GxP motif is typically an asparagine (consensus GNP), and the conserved Ser/Thr found at the C terminus of strand 4 in many other STAND domains is often replaced by an aromatic or hydrophobic residue (Figure 1). Based on sequence conservation, domain architectures, and phyletic patterns, the SWACOS clade can be subdivided into three major families: bacterial-eukaryotic sAC/Chlo1187, actinobacterial ThcG/BpdS, and the largely cyanobacterial dhkG/Npun0353 (Table 1 and Figure 5c).

The sAC/Chlo1187 family is defined by the fusion of the STAND ATPase domain with one or two N-terminal adenylyl cyclase domains (Figure 3). The mammalian members of this family are referred to as soluble adenylyl cyclases (sAC) as opposed to the great majority of adenylyl cyclases that are associated with membranes.[63] Mammalian sAC capacitates sperm cells through Ca^{2+} and bicarbonate-stimulated cAMP accumulation,[63–65] whereas *Dictyostelium* guanylyl cyclase has a role in chemotactic sensitivity and aggregation.[66] As with other families of the STAND class, this family has a patchy phyletic distribution, with several instances of lineage-specific expansion (Table 2). This family is represented in some bacteria (*Chloroflexus*, *Mycobacterium leprae*, several Proteobacteria) and, among the eukaryotes, in slime mold and mammals, but so far missing from all other animals, fungi, and plants (Figure 5c, Table 2). In addition to the eukaryotic sAC proteins, the combination of the NTPase domain of this family with N-terminal adenylyl cyclase is seen in Actinobacteria, α-proproteobacteria, and *Chloroflexus*. In phylogenetic trees, mammalian sACs group with the *Dictyostelium* homolog,[67,68] and this eukaryotic clade is nested within a larger bacterial cluster (Figure 5c). Thus, eukaryotic sACs were probably acquired *via* HGT from bacteria by the common ancestor of the slime mold-animal lineage, with subsequent loss of this gene in some of the animals.

The LipR/ThcG family is a small family of actinobacterial proteins, which consist of an N-terminal NTPase domain and a C-terminal LuxR-type helix-turn-helix domain[69] (Figure 3). There is considerable experimental data implicating members of this family in transcription regulation. In particular, LipR is a transcriptional activator of the LipA lipase in *Streptomyces avermitilis*,[70] NysRI is a regulator of the nystatin biosynthesis gene cluster in *Streptomyces noursei*,[71] PikD is a positive regulator of pikromycin biosynthesis in *Streptomyces venezuelae*,[72] and GdmRI is a putative regulator of the geldanamycin gene cluster in *Streptomyces hygroscopicus*.[73] The BpdS protein of the Actinobacteria *Rhodococcus* sp. *M5* is a member of the same family but contains two additional kinase domains, an N-terminal serine/threonine kinase and a C-terminal DegU-type histidine kinase domain (Figure 3). BpdS is the histidine kinase component of the *Rhodococcus* sp. *M5* BpdT/BpdS two-component signal transduction system

that regulates biphenyl/polychlorobiphenyl metabolism.[74]

The DhkG/Npun0353 family is characterized by a conserved lysine and phenylalanine residue in strand 2 (Figure 1). The domain architecture includes an N-terminal Ser/Thr kinase domain and a C-terminal HupT-type histidine kinase domain (Figure 3). This resembles the domain organization of BpdS, but the histidine kinase domains belong to distinct families, supporting the case for independent origin of this domain architecture in the LipR/ThcG and DhkG/Npun0353 families (Figure 3). The DhkG/Npun0353 family is represented in Cyanobacteria, the spirochaete *Leptospira*, the α-proteobacteria *Magnetospirillum* and *Rhodopseudomonas*, the slime mold *Dictyostelium* (DhkG), and the fungi *Schizosaccharomyces pombe* and *Neurospora crassa* (Figure 5c). Both the NTPase and S/T kinase domains of *Dictyostelium* DhkG are nested within bacterial clusters in phylogenetic trees, suggesting that the entire protein has been relatively recently acquired from Bacteria *via* HGT (Figure 5c and data not shown). The fungal members of this family form a monophyletic clade and were probably derived *via* an independent HGT event (Figure 5c).

The MalT clade

This is a small, bacteria-specific clade that is characterized by a conserved arginine at the N terminus of strand 1, substitution of the first P-loop glycine by serine or alanine, a conserved tryptophan in the second strand, accumulation of acidic residues C-terminal of the second strand (not shown), and a variant of the GxP motif, in which the proline is typically replaced by a hydrophobic residue, whereas the variable position is occupied by tryptophan (Figure 1). MalT is the regulator of the *Escherichia coli* maltose operon, and *Pseudomonas putida* AlkS is the regulator of the alkane-utilization alkBFGHJKL operon.[75,76] In addition to the STAND ATPase and the HETHS domain, MalT contains a LuxR-type helix-turn-helix motif at the very C terminus,[69] and a divergent variant of TPR repeats[33] (Figure 3). The mycobacterial MalT ortholog additionally contains an N-terminal Ser/Thr kinase domain (Figure 3). *E. coli* MalT is the central regulator of the maltose system and integrates multiple regulatory signals, including maltotriose as activator and cysthathionase MalY, Aes, and the maltose transporter MalK as repressors.[77–79] MalT is a monomer in solution but oligomerizes in the presence of the positive effectors, ATP and maltotriose.[80] Structural data suggest that the helical repeat domain is the maltotriose-binding site and that the oligomer is assembled *via* interaction between the HETHS domain of one MalT monomer and the helical repeat domain of the following monomer.[33] MalT is widespread in γ-proteobacteria and is also found in the β-proteobacterium *Ralstonia*, the Gram-positive bacterium *Clostridium*, *Chloroflexus*, and the Actinobacteria *Mycobacterium* and *Streptomyces*, suggesting extensive horizontal mobility among the Bacteria (Table 2 and Figure 5d).

The MNS clade

This clade (named after the constituent families; see below) includes the previously defined MJ and PH-NTPase families and several additional families of prokaryotic predicted NTPases that were identified as part of this work. Members of this clade show widespread conservation of one or more arginine residues in the P-loop, and the second acidic residue in the Walker B motif is a glutamate (Figure 1). In addition, the GxP motif is often preceded by a second glycine residue (Figure 1). Interestingly, the MNS clade lacks the conserved arginine in the motif associated with strand-4, which is present in all other STAND families (Figure 1), but contains a conserved arginine in the loop between strand 5 and the preceding helix. This arginine is equivalent in its position to the arginine finger seen in the AAA+ ATPases.[81–89] The MNS clade consists of several distinct families, which are typically represented by lineage-specific expansions, largely in Archaea, and, to a lesser extent, in bacteria. None of these proteins has been characterized biochemically or biologically. However, the strict conservation of the Walker A and B motifs and the presence of all characteristic structural elements of the P-loop domain leave little doubt that they are active NTPases.

The MJ-type family is characterized by a conserved glutamine in the Walker B motif (hhhhDExQ) and a set of histidine residues in strand 4.[22] This family is represented by 17 paralogs in *Methanocaldococcus jannaschii*, three in *Pyrococcus abyssi*, and one in *Pyrococcus horikoshii* (Table 2).

The PH-type family is characterized by an arginine triplet in the P-loop (consensus GRRRhGKT) (Figure 1). The NTPase domain is typically fused at the C terminus to a winged HTH domain and an endonuclease domain of the PHAC family.[90] This family shows a lineage-specific expansion in *Pyrococcus* and is represented in many other Euryarchaeota (*Ferroplasma*, *Methanobacterium*, *Methanocaldococcus*, *Methanosarcina*, *Thermococcus*), Crenarcheota (*Sulfolobus*, *Pyrobaculum*), and an assemblage of phylogenetically diverse Bacteria (e.g. *Bacteroides*, *Bifidobacterium*, *Fusobacterium*, *Geobacter*, *Mycoplasma penetrans* and *Thermotoga*; Table 2).

The SSO-type NTPase family is typified by SSO1545 of *Sulfolobus solfataricus*; a synapomorphy of this family is a doublet of arginine residues in the P-loop (Figure 1). Similarly to the PH-type family, members of this family have a winged HTH domain at the C terminus, but not the endonuclease domain (Figure 3). Like the previous two families, this is a predominantly archaeal family, with lineage-specific expansions of nine paralogs each in the crenarchaeon *Sulfolobus solfataricus* and the euryarchaeon *Pyrococcus abyssi*. Other Archaea, including *Pyrobaculum*, other *Sulfolobus* species,

Methanosarcina, other *Pyrococcus* species, and the hyperthermophilic bacterium *Thermotoga maritima* encode smaller number of predicted NTPases of this family (Table 2).

Clustering by sequence similarity and certain shared architectural features suggest that the above three families comprise a monophyletic group within the MNS clade. These proteins are the smallest in the STAND class and lack the C-terminal HETHS domain. Furthermore, they differ from most of the other STAND NTPases in lacking fusions to superstructure-forming repeats or enzymatic domains involved in signal transduction, such as kinases or adenylyl cyclases. Although this group of STAND NTPases is widespread in Archaea, phylogenetic trees of these proteins do not show the Crenarchaeota/Euryarchaeota split (Figure 5e). Furthermore, they are absent in several Archaea with sequenced genomes, such as *Halobacterium*, *Thermoplasma*, and *Aeropyrum*. Thus, although the phyletic patterns suggest an archaeal origin of this group, it remains unclear whether it was already present in the common ancestor of Archaea. The phylogenetic tree topology suggests wide horizontal dissemination of this family among Archaea, and, to a certain extant, among bacteria (Figure 5e and Table 2).

The Npun2340/2341 family is a small family (named after two members from *Nostoc punctiforme*) that consists predominantly of cyanobacterial proteins and is characterized by a conserved Arg/Gln motif and frequent substitution of the first glycine residue in the Walker A motif (consensus [AGNST]xRQhGK[ST][ST]), the hhhhDE signature in Walker B motif, a highly conserved [ST]PFNh motif next to the fifth strand, and glutamine or histidine as the middle residue in the GxP motif (Figure 1). The Npun2340/2341 family can be divided into two subfamilies that differ in domain composition and the degree of lineage-specific expansion (Figure 5e and Table 2). The NTPase domain in the Npun2340 subfamily is fused to an N-terminal TIR domain similar to the TIR domains present in some of the plant disease-resistance AP-ATPases (Figure 3). At the C terminus, these proteins contain a unique α-helical globular domain of ~100 amino acid residues, which is much shorter than the HETHS domain and does not show detectable sequence similarity to the latter. The Npun2341 subfamily lacks the N-terminal TIR domain but contains C-terminal TPRs (Figure 3), and an α-helical domain between the NTPase domain and the TPRs. No sequence similarity could be detected between this domain and the HETHS domain seen in other families. Two proteins of this family, one from each of the subfamilies, are encoded by adjacent genes in *Nostoc punctiforme* and in the pCC7120alpha plasmid of *Anabaena* sp. *PCC 7120*. Thus, it appears likely that the two subfamilies evolved through tandem duplication in Cyanobacteria. The Npun2340 subfamily protein Tlr1460 from *Thermosynechococcus elongatus* contains an N-terminal SpkB-type serine/threonine kinase (Figure 3). A stand-alone form of this specific version of the kinase domain is encoded by most cyanobacterial genomes suggesting that Npun2340/2341 family NTPases and SpkB-type serine/threonine kinases function synergistically in a conserved cyanobacterial signaling pathway. In addition to the cyanobacterial proteins, more divergent members of this family were also detected in *Clostridium tetani* and in *Dictyostelium*.

The SpsJ family is another small family so far represented only in the α-proteobacterium *Sphingomonas*, the γ-proteobacterium *Yersinia pestis*, and in the cyanobacterium *Trichodesmium*, which has a lineage-specific expansion (Table 2). The SpsJ family shows a hhhhDE[YF]D signature in the Walker B motif (Figure 1). SpsJ from *Sphingomonas* sp. *S88* and GelJ from *Sphingomonas elodea* are involved in synthesis or secretion of capsular polysaccharides[91] but their biochemical roles in these processes remain uncharacterized. These proteins consist of an NTPase module, with an α-helical C-terminal domain, which, despite occurring in a similar position, shows no detectable similarity to the HETHS domain of other STAND class members. The members of this family from *Trichodesmium* (e.g. g23043531) are larger proteins and are the only members of the STAND class that contain a duplication of the NTPase domain within the same polypeptide.

The Bl0662 family is typified by *Bifidobacterium longum* ORF BL0662 and characterized by a conserved arginine at the N terminus of strand 1 and a conserved threonine in the Walker B strand (data not shown). This is a very small family of proteins that currently consists of only about ten homologs found in the Actinobacteria *Bifidobacterium* and *Corynebacterium*, and the Gram-positive bacterium *Lactobacillus* (Figure 5e). The Bl0662 family proteins are all very short (between 250 and 380 residues) and appear to represent the only instances of a solo version of the STAND domain.

Domain architectures and their implications for the biochemical functions of STAND NTPases

STAND ATPases show a wide range of fusions to domains involved in protein–protein or protein–DNA interactions, small-molecule-binding domains, and catalytic domains involved in signal transduction (Figure 3). Many of these architectures are either unique to a particular lineage or are shared by a small set of phylogenetically distant organisms. Examination of the domain architectures of a diverse sample of STAND proteins from representative organisms, we found that they contain, on an average, three to four domains per protein (superstructure-forming repeats were counted as a single domain). Similar analysis of other classes of P-loop NTPases, such as AAA+ ATPases, helicases, kinases, and GTPases, indicated a lower level of complexity, with approximately one to 2.5 domains per protein. The remarkable diversity notwithstanding (Figure 3), the domain

architectures of the STAND NTPases seem to follow three major themes: (i) fusion of the STAND NTPase domain (along with the HETHS domain) to N or C-terminal catalytic domains; (ii) fusion of the NTPase domain with N-terminal DNA-binding or peptide-interaction domains and C-terminal superstructure-forming repeats; and (iii) fusion of the NTPases domain with C-terminal DNA-binding domains either directly or *via* intervening superstructure-forming repeats. Similar domain architectures of STAND-containing proteins appear to have been independently derived on multiple occasions during evolution (Figure 3), suggesting that there are strong functional constraints favoring the repeated emergence of these domain combinations. For example, the N-terminal HTH domains of the CalR2 family and AfsR are of the OmpR class, whereas the C-terminal HTH of the MalT and ThcG families belong to the LuxR type (Figure 3). Similarly, the C-terminal histidine kinase domains of BpdS and the cyanobacterial Npun0353 family belong to different subfamilies of histidine kinases. These architectural themes suggest that STAND NTPases act as regulatory nexuses involved in integration of multiple signals that are transmitted by various fused signaling domains. The structural similarity with the AAA+ ATPases, together with the presence of superstructure-forming repeats at their C termini, suggests that STAND NTPases might additionally act as scaffolds for NTP-dependent assembly of protein complexes on the periodic surfaces of these repeats. Movements of the GxP module and the HETHS domain (when present) in response to the bound nucleotide are likely to be central to these functions of the STAND NTPases.

Consistent with the inferences that can be drawn from domain architectures, all functionally characterized STAND NTPases have a role in signal transduction or transcription regulation. Many of the bacterial STAND NTPases respond to the availability of simple nutrients in the environment. In particular, *E. coli* MalT is a positive regulator of the maltose regulon,[92] AcoK is required for the expression of the acetoin operon,[93] *Bacillus* GutR is the transcriptional activator of the glucitol operon,[45] *Rhodococcus* sp. M5 BpdS regulates biphenyl/polychlorobiphenyl metabolism,[74] and *Streptomyces avermitilis* LipR activates the LipA lipase, which apparently is involved in the utilization of oils present in the medium.[70]

Other STAND NTPases, such as AfsR, GdmRI, NysRI, and PikD proteins from *Streptomyces* and related Actinobacteria, are regulators of the biosynthesis of geldanamycin, pikromycin, and other secondary metabolites.[43,44,71–73] However, even in the case of MalT or GutR, the regulatory role is not a simple feedback loop where binding of an inducer alone stimulates the transcription-activating or inhibiting activity of the transcription factor. Instead, STAND ATPases seem to be part of complex regulatory networks that integrate many different signals. For example, MalT is activated or inhibited by at least three other proteins (MalK, MalY, AES), in addition to monitoring the presence of the inducer (maltotriose) and ATP.[77,78] Similarly, some of the eukaryotic homologs are known to integrate several input signals. Successive binding of cytochrome *c* and ATP promotes human Apaf-1 to assemble into a heptameric platform and bind procaspase-9 in the so-called apoptosome.[35,36] Mammalian soluble adenylyl cyclase plays a role in the cAMP-mediated activation of spermatozoa and seems to be a sensor of Ca^{2+} and bicarbonate.[94] Furthermore, in this case, the conformational change mediated by the STAND ATPase domain favors proteolytic cleavage and release of an active, soluble, N-terminal adenylyl cyclase domain.[64,65]

Molecular studies on signal transduction systems revealed several paradigms that are relevant in both eukaryotes and Bacteria. These include the two-component relay between histidine kinases and receiver domains, the single-component systems, where a small-molecule-binding domain regulates a fused DNA-binding domain by sensing effectors, and regulation of substrate protein properties by post-translational modifications, e.g. phosphorylation, ubiquitination, and reversal of these modifications. Although the details of the mechanism of action of the STAND ATPases remain to be elucidated, we propose that these proteins represent a novel paradigm in signal transduction whereby roles of scaffold, adaptor, and regulatory switch are combined in a single protein. The STAND NTPases could function as signaling hubs, in which signals are received and relayed to the next component in the chain. In prokaryotes, this principle is utilized in various signaling contexts, whereas in eukaryotes, it applies more specifically to defense against pathogens, regulation of programmed cell death, and self/non-self-discrimination.

Horizontal gene transfer and lineage-specific expansion of paralogs: major forces in evolution of STAND NTPases

The STAND class is represented in all three superkingdoms of life but individual families show extremely patchy phyletic patterns. These proteins are particularly widespread in Actinobacteria, Cyanobacteria, Chloroflexus, and certain Alphaproteobacteria and Archaea, but are rare or absent in most other prokaryotic lineages (Tables 1 and 2, Figure 5a–e). Among eukaryotes, STAND NTPases are found in most representatives of the crown group (Tables 1 and 2, Figure 5a–e) but so far are missing in diverse unicellular eukaryotes with sequenced genomes, which include *Giardia*, trypanosomes, apicomplexans, microsporidians, and the yeast *Saccharomyces cerevisiae*. So far, STAND NTPases have not been detected in viral genomes. Phylogenetic trees of most families in the STAND class contain strongly supported clades that bring together proteins from phylogenetically diverse organisms, often from two superkingdoms (Figure 5a–e). The trees of eukaryotic STAND NTPases tend

to follow the higher order organismal phylogeny (thus, the animal CED4/Apaf1 ATPases and plant resistance proteins comprise the two principal eukaryotic branches of the AP-ATPase family), but there are major lacunae in the phyletic patterns. These features put STAND NTPases in stark contrast to other P-loop NTPases of the ASCE division, such as the AAA+, RecA-like, and ABC NTPases, which include many families that are highly conserved throughout the evolution of the major lineages of life and have phylogenetic trees that generally tend to follow the organismal phylogenies.[9,95–98] Thus, it appears that the STAND class has a far more prominent history of HGT and gene loss than most of the other P-loop NTPases.

The striking differences in the occurrence of most STAND NTPase families in different taxa, often even closely related ones, point to two other major evolutionary processes, lineage-specific gene loss and lineage-specific expansion of paralogs. In eukaryotes, a particularly notable case of extensive gene loss is the sAC family, which appears to have been eliminated on multiple occasions among animals.[68] The AP-ATPase and TLP1-like families also might have been lost in certain eukaryotic lineages; however, the currently available genomic data do not allow us to distinguish between this possibility and the alternative, HGT-based scenario. Lineage-specific expansions are seen in many families of the STAND class, the most dramatic cases being the disease-resistance AP-ATPases in plants, NACHT NTPases in vertebrates, and MJ-type NTPases in *Methanocaldococcus jannaschii* (Table 2). Lineage-specific expansions appear to be a major adaptation strategy in organisms, especially eukaryotes, which are confronted with multiple cues of the same general nature. In particular, interaction with different pathogens, detoxification or modification of multiple environmental compounds, production of diversified secondary metabolites, and transcriptional or signaling response to multiple environmental stimuli appear to be perpetuated through lineage-specific expansions.[99] The expansions in the STAND class seem to conform with this principle, as illustrated by the diversification of the plant AP-ATPases and vertebrate NACHT NTPases, which are involved in the response to numerous pathogens.[40,100–103] A variation on this theme is the lineage-specific expansion of the fungal NACHT NTPases, which appear to participate in self/non-self-discrimination during the fusion of vegetative mycelia to form heterokaryons.[59,60] Similarly, in prokaryotes, the lineage-specific expansion of STAND domains fused with DNA-binding HTH domains is analogous to similar expansions in other classes of transcriptional regulators.[31,104] These expansions might have allowed Actinobacteria to regulate the expression of the biosynthetic pathways for a wide range of secondary metabolites that evolved in this bacterial lineage. Based on this precedence, we suspect that similar expansions represent adaptations that have enabled complex signal transduction switches in the expanded biosynthetic and developmental pathways of α-proteobacteria, planctomycetes, and Cyanobacteria. The other contribution to lineage-specific expansions comes from selfish elements, such as transposons, that proliferate in various genomes. The presence of an endonuclease domain of the PHAC family in the PH-type ATPases raises the possibility that these genes could be such selfish genomic elements, with the nuclease domain mediating transposition.

Excluding the MJ/PH/SSO-type NTPase families, which have a simple architecture and lack the HETHS domain, the abundance of STAND NTPases in a genome clearly correlates with the developmental and organizational complexity of the organism. In particular, among prokaryotes, STAND NTPases are most diverse in filamentous, "multicellular" Bacteria, namely, Actinobacteria, Cyanobacteria, *Chloroflexus* and α-proteobacteria of the Rhizobiaceae group, and the "multicellular" archaeon *Methanosarcina*. Furthermore, among Cyanobacteria, the STAND proteins are more prevalent in those species that have a relatively large genome along with complex organization or development. Thus, the filamentous Cyanobacteria *Anabaena*, *Nostoc* and *Trichodesmium erythraeum* have numerous STAND NTPases, whereas the simpler forms, such as *Synechocystis* sp. *PCC 6803* and *Thermosynechococcus elongatus*, have fewer proteins of this class, and the "minimal" cyanobacterium *Prochlorococcus* has none (Table 2). Most of the complex Bacteria (e.g. *Streptomyces*, *Anabaena*, *Nostoc*, *Chloroflexus* and *Bradyrhizobium*) encode representatives of various families of the STAND class, suggesting that, in addition to the lineage-specific expansions, they accumulated diverse members of this class *via* HGT. Among eukaryotes, numerous paralogous STAND NTPases are encoded in the genomes of all filamentous fungi but are either absent or rare in yeasts (Tables 1 and 2). Similarly, the protist *Dictyostelium discoideum*, which belongs to the crown group and has a complex developmental cycle, encodes members of multiple families of the STAND class, whereas true unicellular protists, such as *Giardia*, *Cryptosporidium* and *Plasmodium*, have none.

STAND NTPases and organizational complexity

This distribution of the STAND NTPases in complex bacteria mimics, at least roughly, the distribution of several other signaling proteins that were initially considered to be typical of the eukaryotic signaling systems.[105] These include serine/threonine protein kinases, FHA-domain-containing proteins, adenylyl cyclases, caspase-like proteases, and proteins containing WD40 repeats and TIR domains. These domains often co-occur in different combinations in large polypeptides. This suggests that STAND NTPases, along with these additional, "eukaryote-type", signaling domains, comprise building blocks for

multidomain proteins and multisubunit complexes which are specifically involved in signaling cascades associated with development and differentiation.[105] Given a certain degree of functional interactions between these components, they might have a tendency to be horizontally transferred together, perhaps as operons encoding functionally linked sets of signaling proteins. Such transfers might have favored emergence of developmental and organizational complexity in those prokaryotic lineages that accumulated a certain number of these components. Even larger sets of signaling proteins might have been acquired *via* megaplasmids carrying several genes of this category. This scenario is consistent with the presence of genes encoding such proteins in megaplasmids from Cyanobacteria and α-proteobacteria (e.g. the *Nostoc* sp. PCC 7120 408.Kbp plasmid pCC7120-alpha, *Sinorhizobium meliloti* 1.35 Mbp pSymA megaplasmid; L. M. Iyer & L.A. unpublished results).

Extending the previously published hypothesis proposed specifically for the programmed cell death system,[21] we suggest that eukaryotic developmental complexity was more generally affected by HGT of signaling proteins from Bacteria. While some of these signaling proteins, such as serine/threonine kinases and Ras-like GTPases, might have been acquired from the α-proteobacterial precursor of the mitochondrion, the phylogenetic trees for the STAND class families suggest multiple transfers, some of these at much later points (Tables 1 and 2, Figure 5a–e). These transfers appear to have occurred during diversification of the eukaryotic crown group and could have involved both the α-proteobacterial symbionts and other, more transient, symbionts or even ingested Bacteria.[106] The shared habitats of some bacteria, protists, and multicellular eukaryotes, such as Actinobacteria, slime molds, and filamentous fungi in soil, might have facilitated some of the apparent more recent horizontal transfers of STAND NTPases observed in these organisms (Tables 1 and 2, Figure 5a–e).

The origin and evolution of the STAND class and its relationships with other classes of P-loop NTPases

In contrast to the other classes of P-loop NTPases, not a single family within the STAND class could be traced back to the LUCA of the extant life forms. This implies either of the following scenarios: (i) STAND NTPases were absent in LUCA and were derived later *via* rapid divergence, perhaps from a prokaryotic AAA+ ATPase or; (ii) STAND NTPases evolved prior to LUCA, which had at least one representative of this class, but subsequent gene losses, lateral transfers, and domain shuffling erased all phylogenetic information related to their origins. The marked distinctness of the STAND NTPase domain from all other ASCE P-loop domains suggests an early origin, perhaps favoring scenario (ii). Both sequence features (such as the arginine equivalent to the arginine finger of AAA+ ATPases, as opposed to the arginine in strand 4) and domain architectures indicate a fundamental split between the MNS clade and the rest of the STAND NTPases, which are unified by the presence of a C-terminal HETHS domain (Figure 6). This split might represent a basal divergence in the STAND class that accompanied the separation of the archaeal and bacterial lineages, but the lateral transfers and gene losses do not allow us to assess this scenario with greater clarity. Notably, the greatest diversity of the STAND NTPases is seen in Cyanobacteria, Actinobacteria, and *Chloroflexus*. Given that several phylogenetic analyses suggested that these Bacteria comprise a higher order clade,[107,108] the possibility exists that the STAND class underwent a major diversification in the common ancestor of the Actinobacteria–Cyanobacteria–*Chloroflexus* clade and was subsequently disseminated among other taxa *via* HGT (Figure 7). Although it is unlikely that we ever will be in a position to distinguish between the above two scenarios, additional bacterial genome sequences might help in determining whether a major diversification of STAND NTPases indeed took place in the ancestor of the Actinobacteria–Cyanobacteria–*Chloroflexus* clade.

Within the ASCE division, the STAND class appears to be most closely related to the AAA+ class. Unlike the RecA/ATP synthase, SFI/II helicase, PilT and HerA-FtsK classes, but similarly to the AAA+ class, the core sheet of the NTPase domain consists of only five strands (Figures 1 and 2). While this could be a primitive character of the ASCE division, there are additional similarities between STAND and AAA+ NTPases that are likely to comprise synapomorphies of a higher order clade. In particular, the two classes share a helix N-terminal of the P-loop. In both classes, this helix often contains a glycine and an acidic residue at its very N terminus (the most common versions of this motif are GR[DE] in the STAND class and GQ[DE] in the AAA+ class; see Figure 1 and the work done by Iyer *et al*.[15]). In addition, in both classes, a helical bundle is located immediately C-terminal of strand 5 of the NTPase core.[15,81–89] Furthermore, similar sequence signatures are associated with strand 4 of both classes (Figures 1 and 2), which in AAA+ NTPases comprises the sensor I motif. Some of these features are also shared with another, poorly characterized group of predicted membrane-associated P-loop NTPases, the KAP family[16]), suggesting an evolutionary connection between all three classes of NTPases.

By analogy with the AAA+ ATPases,[81,109] one could speculate that the GxP module of the STAND domain functions as an adaptor transmitting the conformational changes triggered by NTP hydrolysis to another domain that is fused or non-covalently bound to the NTPase module. In particular, the GxP motif might act as a hinge facilitating NTP-dependent movement of the flanking helices. The importance of the motif is emphasized by the fact that a G→E mutation of the glycine residue in the

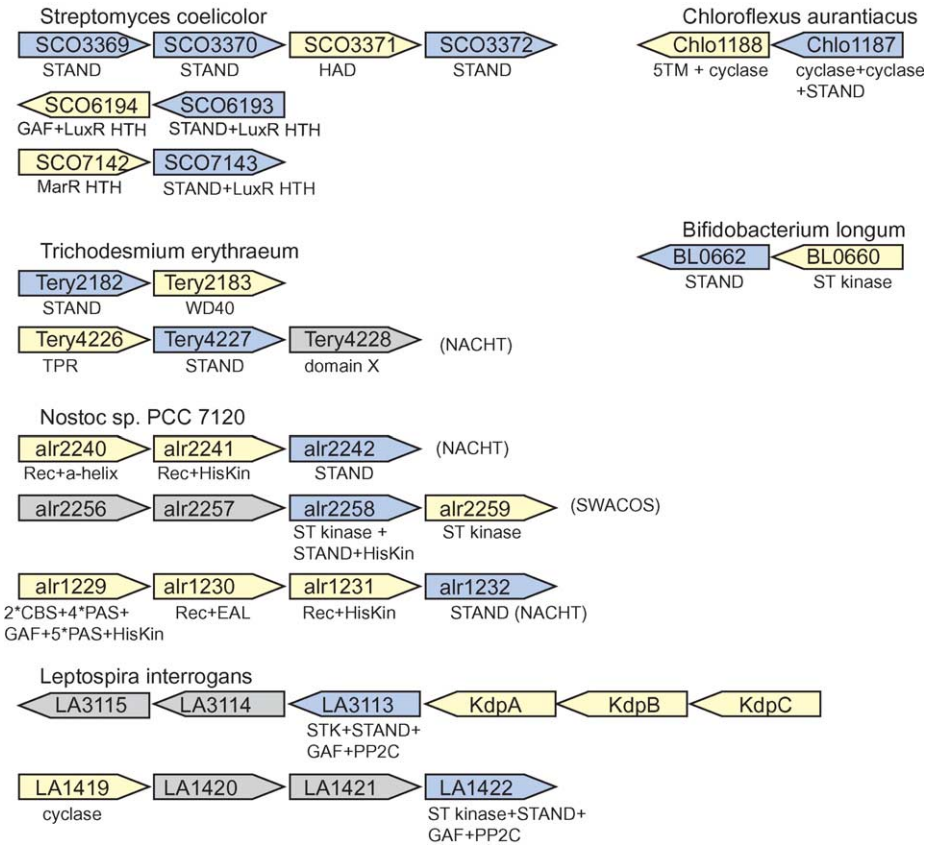

Figure 6. Operon organization of selected bacterial STAND ATPases. Abbreviations: ST kinase, serine/threonine kinase; HisKin, histidine kinase; 5TM, 5 transmembrane domain; PP2C, Sigma factor PP2C-like phosphatase; Rec, receiver domain of response regulator. The X symbolizes an uncharacterized conserved domain that is found fused to the STAND domain in some STAND ATPases (e.g. gis 23041966 and 23043628). The gene containing the STAND domain is in blue, other genes with characterized domains are colored yellow, and the remainder is gray.

GxP motif completely abolishes P2 rust-resistance in flax.[50] Similarly, an *Arabidopsis* RPP1 mutant that carries a deletion of the glycine residue and the N-terminal adjacent residue (which probably disrupts the entire helix that precedes the GxxP motif) has lost the resistance provided by the wild-type.[110]

Beyond these shared features, the AAA+ and STAND classes differ substantially. In the STAND domain, we were unable to detect equivalents of the conserved arginine finger motif that is located between strand 5 and the preceding helix (with the exception of the MNS clade, which appears to have a conserved arginine in a position equivalent to that of the AAA+ arginine finger; see Figure 1) and the sensor II arginine, which is located in the C-terminal helical bundle of the AAA+ ATPases.[81–89,111] The complementary presence of the conserved arginine in strand 4 of all HETHS-domain-containing STAND NTPases and a conserved arginine in the loop between strand 5 and the preceding helix of the MNS clade suggests that these residues play a functionally equivalent role in stabilizing the negative charge on the reaction intermediate during ATP hydrolysis (Figure 1). However, the position of the conserved strand 4 arginine in the HETHS-domain-containing STAND proteins precludes it from playing a role in facilitating ring formation, which is characteristic of the arginine fingers of the AAA+ class.[15,81,111–113] Oligomerization has been reported for several STAND proteins including MalT, Apaf-1, C2TA, and Nod1[35,36,80,114–116] but, with the exception of Apaf-1, there is no evidence that these oligomers are toroidal structures. However, even in Apaf-1, the main factor in ring-formation is the CARD domain rather than the ATPase domain.[116] Given these observations and the absence of an AAA+-like arginine finger, it remains doubtful whether the STAND ATPases (with the exception of the MNS clade) have an intrinsic propensity to form oligomeric rings similar to those formed by AAA+ ATPases.

General Conclusions

Our understanding of the structure, function, and evolutionary history of the P-loop NTPases has vastly improved in the two decades since the relationship between kinases, ATP synthetase, and several other P-loop proteins has been recognized for the first time.[11] In particular, several large, distinct classes of P-loop domains have been delineated and evolutionary relationships within

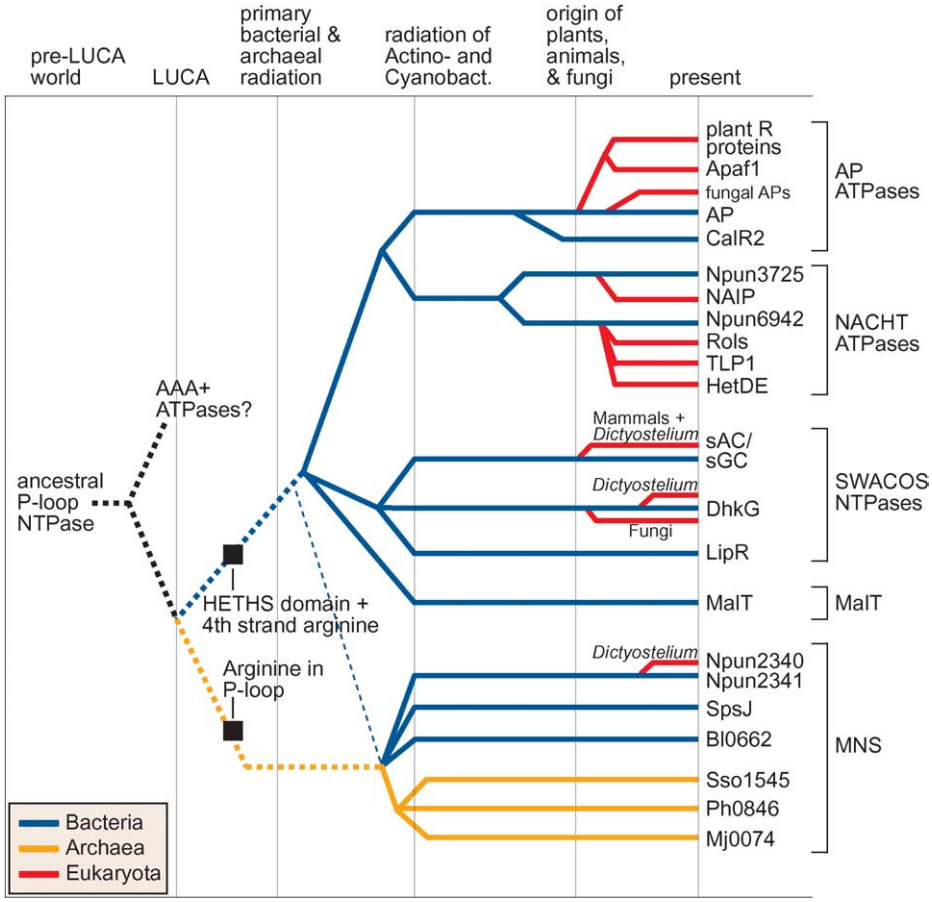

Figure 7. Inferred evolutionary history of STAND NTPases. The Figure shows relative temporal epochs and marks major evolutionary events by vertical lines. The evolution of the protein-coding gene is traced with horizontal colored lines. A broken line indicates uncertainty with respect to the exact point of origin. The Figure emphasizes horizontal gene transfer (HGT) from Bacteria to Eukaryota as indicated by red lines sprouting from blue lines. In addition, there are many inferred cases of HGT between Bacteria and Archaea and within the Bacteria that are not depicted here; see the text for details. Please also note that the color scheme for Bacteria (blue), Archaea (orange), and Eukaryota (red) is just an approximation for the phyletic distribution; there are several cases were Archaea are found within a predominantly bacterial lineage and *vice versa*.

each of these classes have been partly resolved. Here, using genomic sequence information and structure predictions, we identify the STAND class, which consists of several families of large, multidomain P-loop NTPases from Archaea, Bacteria, and eukaryotes, including the animal and plant regulators of pathogen defense and programmed cell death. We characterized the defining sequence features and domain organization of the STAND ATPases, identified the STAND NTPase domain in functionally important proteins implicated in human disease and development, delineated several previously uncharacterized families, and constructed an evolutionary classification. We show that evolution of the STAND class was dominated by numerous lineage-specific expansions, HGT, gene loss, and extensive domain shuffling, to an extent unprecedented in other NTPases. These events obscure the early evolutionary history of STAND NTPases such that none of the extant lineages can be traced back to LUCA. Among other P-loop NTPases, the STAND class appears to be most closely related to the AAA+ ATPases, and the two classes probably share an ancestral evolutionary relationship and some mechanistic features, such as transmission of conformational changes *via* a C-terminal helical bundle to effector domains or proteins. The STAND NTPases seem to represent a novel paradigm in signal transduction: signaling nexus proteins that integrate scaffolds, adaptors, signaling enzymes, and regulatory switches in a single, multidomain protein.

Supporting material

Complete lists of STAND NTPases (represented with GenBank GI numbers) from sequenced genomes and alignments used for phylogenetic tree construction are available†.

† ftp://ftp.ncbi.nih.gov/pub/aravind/STAND/

Materials and Methods

Sequences of STAND proteins and other relevant proteins were extracted from the non-redundant (NR) protein sequence database (National Center for Biotechnology Information, NIH, Bethesda) by using the PSI-BLAST program,[117,118] with the sequences of previously identified STAND ATPases employed as queries. Sequence similarity-based protein clustering was performed using the BLASTCLUST program†. Multiple alignments were constructed using the Clustal X or T-Coffee programs[119,120] and corrected on the basis of PSI-BLAST results. Alignments were rendered using the ALSCRIPT software.[121] For each family, the phyletic distribution was evaluated in terms of the presence of homologs in completed genomes from the three primary kingdoms, Bacteria, Archaea, and Eukaryota. Statistically significant conserved motifs were then identified in the NTPase domains using the Gibbs sampling algorithm as implemented in the Probe program.[122] Protein secondary structure prediction was performed using JPRED and the PHD program through the PredictProtein server.[123,124] Domain architectures were analyzed using the SMART, Pfam and CDD databases and software tools.[125–127]

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§.[128] Support for selected tree branches was measured by 1000 bootstrap resamplings with PHYLIP (protdist/fitch, randomized species input order, three jumbles) or the minimum evolution method as implemented in MEGA2.[129] Bootstrap values were also computed using maximum parsimony analysis as implemented in the PAUP software package[130] for 1000 replicates with the heuristic search type and random addition option set to ten reps. Phylogenetic trees were rendered with the TREEVIEW program.[131] Phylogenetic analysis described here focused largely on deciphering the relationships between the three primary kingdoms (Archaea, Bacteria, and Eukaryota) and, accordingly, only regions that could be unambiguously aligned between proteins from all three kingdoms within a given family were selected for phylogenetic analysis. Therefore, some of the trees do not provide good resolution of the branching pattern within a lineage, e.g. within the Bacteria. For evolutionary reconstructions, the "standard model" of early evolution, which postulates the original split between the bacterial and archaeo-eukaryotic lineages,[132] was employed as the null hypothesis.

References

1. 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.
2. 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.
3. Schulz, G. E. (1992). Binding of nucleotides by proteins. *Curr. Opin. Struct. Biol.* **2**, 61–67.
4. Vetter, I. R. & Wittinghofer, A. (1999). Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer. *Quart. Rev. Biophys.* **32**, 1–56.
5. 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.
6. Doolittle, R. F., Feng, D.-F., Tsang, S., Cho, G. & LIttle, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, **271**, 470–477.
7. Miyamoto, M. M. & Fitch, W. M. (1996). Constraints on protein evolution and the age of the eubacteria/eukaryote split. *Syst. Biol.* **45**, 568–575.
8. Anantharaman, V., Koonin, E. V. & Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucl. Acids Res.* **30**, 1427–1464.
9. Leipe, D. D., Wolf, Y. I., Koonin, E. V. & Aravind, L. (2002). Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.* **317**, 41–72.
10. 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.
11. 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.
12. Gorbalenya, A. E. & Koonin, E. V. (1989). Viral proteins containing the purine NTP-binding sequence pattern. *Nucl. Acids Res.* **17**, 8413–8440.
13. Leipe, D. D., Koonin, E. V. & Aravind, L. (2003). Evolution and classification of P-loop kinases and related proteins. *J. Mol. Biol.* **333**, 781–815.
14. 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.
15. Iyer, L. M., Leipe, D. D., Koonin, E. V. & Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *J. Struct. Biol.* **146**, 11–31.
16. Aravind, L., Iyer, L., Leipe, D. & Koonin, E. (2004). A novel family of P-loop NTPases with an unusual phyletic distribution and transmembrane segments inserted within the NTPase domain. *Genome Biol.* **5**, R30.1–R30.10.
17. Chinnaiyan, A. M., Chaudhary, D., O'Rourke, K., Koonin, E. V. & Dixit, V. M. (1997). Role of CED-4 in the activation of CED-3. *Nature*, **388**, 728–729.
18. van der Biezen, E. A. & Jones, J. D. (1998). The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr. Biol.* **8**, R226–R227.
19. Aravind, L., Dixit, V. M. & Koonin, E. V. (1999). The domains of death: evolution of the apoptosis machinery. *Trends Biochem Sci.* **24**, 47–53.
20. Koonin, E. V. & Aravind, L. (2000). The NACHT family – a new group of predicted NTPases implicated in apoptosis and MHC transcription activation. *Trends Biochem Sci.* **25**, 223–224.

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

‡ http://evolution.genetics.washington.edu/phylip.html

§ http://ftp.cse.sc.edu/bioinformatics/molphy/molphy-2.3b3/

21. Koonin, E. V. & Aravind, L. (2002). Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell Death Differ.* **9**, 394–404.
22. Koonin, E. V. (1997). Evidence for a family of archaeal ATPases. *Science*, **275**, 1489–1490.
23. Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I. & Koonin, E. V. (1999). Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* **9**, 608–628.
24. Story, R. M. & Steitz, T. A. (1992). Structure of the recA protein–ADP complex. *Nature*, **355**, 374–376.
25. Campbell, M. J. & Davis, R. W. (1999). On the *in vivo* function of the RecA ATPase. *J. Mol. Biol.* **286**, 437–445.
26. 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.
27. 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.
28. 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.
29. Sawaya, M. R., Guo, S., Tabor, S., Richardson, C. C. & Ellenberger, T. (1999). Crystal structure of the helicase domain from the replicative helicase-primase of bacteriophage T7. *Cell*, **99**, 167–177.
30. Gorbalenya, A. E. & Koonin, E. V. (1993). Helicases: amino acid sequence comparisons and structure–function relationships. *Curr. Opin. Struct. Biol.* **3**, 419–429.
31. Aravind, L. & Koonin, E. V. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucl. Acids Res.* **27**, 4658–4670.
32. Poon, K. K., Chu, J. C. & Wong, S. L. (2001). Roles of glucitol in the GutR-mediated transcription activation process in *Bacillus subtilis*: glucitol induces GutR to change its conformation and to bind ATP. *J. Biol. Chem.* **276**, 29819–29825.
33. Steegborn, C., Danot, O., Huber, R. & Clausen, T. (2001). Crystal structure of transcription factor MalT domain III: a novel helix repeat fold implicated in regulated oligomerization. *Structure (Camb)*, **9**, 1051–1060.
34. Chaudhary, D., O'Rourke, K., Chinnaiyan, A. M. & Dixit, V. M. (1998). The death inhibitory molecules CED-9 and CED-4L use a common mechanism to inhibit the CED-3 death protease. *J. Biol. Chem.* **273**, 17708–17712.
35. Saleh, A., Srinivasula, S. M., Acharya, S., Fishel, R. & Alnemri, E. S. (1999). Cytochrome c and dATP-mediated oligomerization of Apaf-1 is a prerequisite for procaspase-9 activation. *J. Biol. Chem.* **274**, 17941–17945.
36. Adams, J. M. & Cory, S. (2002). Apoptosomes: engines for caspase activation. *Curr. Opin. Cell Biol.* **14**, 715–720.
37. Heath, M. C. (2000). Hypersensitive response-related death. *Plant Mol. Biol.* **44**, 321–334.
38. Shirasu, K. & Schulze-Lefert, P. (2000). Regulators of cell death in disease resistance. *Plant Mol. Biol.* **44**, 371–385.
39. Whitham, S., Dinesh-Kumar, S. P., Choi, D., Hehl, R., Corr, C. & Baker, B. (1994). The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor. *Cell*, **78**, 1101–1115.
40. Meyers, B. C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R. W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell*, **15**, 809–834.
41. Stenmark, H., Aasland, R. & Driscoll, P. C. (2002). The phosphatidylinositol 3-phosphate-binding FYVE finger. *FEBS Letters*, **513**, 77–84.
42. Moutinho, A., Hussey, P. J., Trewavas, A. J. & Malho, R. (2001). cAMP acts as a second messenger in pollen tube growth and reorientation. *Proc. Natl Acad. Sci. USA*, **98**, 10481–10486.
43. Horinouchi, S., Kito, M., Nishiyama, M., Furuya, K., Hong, S. K., Miyake, K. & Beppu, T. (1990). Primary structure of AfsR, a global regulatory protein for secondary metabolite formation in *Streptomyces coelicolor* A3(2). *Gene*, **95**, 49–56.
44. Umeyama, T., Lee, P. C. & Horinouchi, S. (2002). Protein serine/threonine kinases in signal transduction for secondary metabolism and morphogenesis in *Streptomyces*. *Appl. Microbiol. Biotechnol.* **59**, 419–425.
45. Ye, R., Rehemtulla, S. N. & Wong, S. L. (1994). Glucitol induction in *Bacillus subtilis* is mediated by a regulatory factor, GutR. *J. Bacteriol.* **176**, 3321–3327.
46. Meyers, B. C., Dickerman, A. W., Michelmore, R. W., Sivaramakrishnan, S., Sobral, B. W. & Young, N. D. (1999). Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.* **20**, 317–332.
47. Sakamoto, K., Tada, Y., Yokozeki, Y., Akagi, H., Hayashi, N., Fujimura, T. & Ichikawa, N. (1999). Chemical induction of disease resistance in rice is correlated with the expression of a gene encoding a nucleotide binding site and leucine-rich repeats. *Plant Mol. Biol.* **40**, 847–855.
48. Parniske, M., Hammond-Kosack, K. E., Golstein, C., Thomas, C. M., Jones, D. A., Harrison, K. *et al*. (1997). Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell*, **91**, 821–832.
49. Thomas, C. M., Jones, D. A., Parniske, M., Harrison, K., Balint-Kurti, P. J., Hatzixanthis, K. & Jones, J. D. (1997). Characterization of the tomato Cf-4 gene for resistance to *Cladosporium fulvum* identifies sequences that determine recognitional specificity in Cf-4 and Cf-9. *Plant Cell*, **9**, 2209–2224.
50. Dodds, P. N., Lawrence, G. J. & Ellis, J. G. (2001). Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the P and P2 rust resistance specificities in flax. *Plant Cell*, **13**, 163–178.
51. 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.
52. Ahlert, J., Shepard, E., Lomovskaya, N., Zazopoulos, E., Staffa, A., Bachmann, B. O. *et al*. (2002). The calicheamicin gene cluster and its iterative type I enediyne PKS. *Science*, **297**, 1173–1176.
53. Espagne, E., Balhadere, P., Begueret, J. & Turcq, B. (1997). Reactivity in vegetative incompatibility of the HET-E protein of the fungus *Podospora anserina* is dependent on GTP-binding activity and a WD40 repeated domain. *Mol. Gen. Genet.* **256**, 620–627.
54. 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.
55. Harton, J. A., Linhoff, M. W., Zhang, J. & Ting, J. P. (2002). Cutting edge: CATERPILLER: a large family of mammalian genes containing CARD, pyrin, nucleotide-binding, and leucine-rich repeat domains. *J. Immunol.* **169**, 4088–4093.
56. Olbrich, H., Fliegauf, M., Hoefele, J., Kispert, A., Otto, E., Volz, A. *et al*. (2003). Mutations in a novel gene, NPHP3, cause adolescent nephronophthisis, tapeto-retinal degeneration and hepatic fibrosis. *Nature Genet.* **34**, 455–459.
57. Watnick, T. & Germino, G. (2003). From cilia to cyst. *Nature Genet.* **34**, 355–356.
58. Hilliard, M. A., Bergamasco, C., Arbucci, S., Plasterk, R. H. & Bazzicalupo, P. (2004). Worms taste bitter: ASH neurons, QUI-1, GPA-3 and ODR-3 mediate quinine avoidance in *Caenorhabditis elegans*. *EMBO J.* **23**, 1101–1111.
59. Saupe, S., Turcq, B. & Begueret, J. (1995). A gene responsible for vegetative incompatibility in the fungus *Podospora anserina* encodes a protein with a GTP-binding motif and G beta homologous domain. *Gene*, **162**, 135–139.
60. Espagne, E., Balhadere, P., Penin, M. L., Barreau, C. & Turcq, B. (2002). HET-E and HET-D belong to a new subfamily of WD40 proteins involved in vegetative incompatibility specificity in the fungus *Podospora anserina*. *Genetics*, **161**, 71–81.
61. Rau, A., Buttgereit, D., Holz, A., Fetter, R., Doberstein, S. K., Paululat, A. *et al*. (2001). rolling pebbles (rols) is required in *Drosophila* muscle precursors for recruitment of myoblasts for fusion. *Development*, **128**, 5061–5073.
62. Menon, S. D. & Chia, W. (2001). *Drosophila* rolling pebbles: a multidomain protein required for myoblast fusion that recruits D-Titin in response to the myoblast attractant Dumbfounded. *Dev. Cell*, **1**, 691–703.
63. Buck, J., Sinclair, M. L., Schapal, L., Cann, M. J. & Levin, L. R. (1999). Cytosolic adenylyl cyclase defines a unique signaling molecule in mammals. *Proc. Natl Acad. Sci. USA*, **96**, 79–84.
64. Chen, Y., Cann, M. J., Litvin, T. N., Iourgenko, V., Sinclair, M. L., Levin, L. R. & Buck, J. (2000). Soluble adenylyl cyclase as an evolutionarily conserved bicarbonate sensor. *Science*, **289**, 625–628.
65. Litvin, T. N., Kamenetsky, M., Zarifyan, A., Buck, J. & Levin, L. R. (2003). Kinetic properties of "soluble" adenylyl cyclase. Synergism between calcium and bicarbonate. *J. Biol. Chem.* **278**, 15922–15926.
66. Roelofs, J., Meima, M., Schaap, P. & Van Haastert, P. J. (2001). The *Dictyostelium* homologue of mammalian soluble adenylyl cyclase encodes a guanylyl cyclase. *EMBO J.* **20**, 4341–4348.
67. Roelofs, J. & Van Haastert, P. J. (2001). Genes lost during evolution. *Nature*, **411**, 1013–1014.
68. Roelofs, J. & Van Haastert, P. J. (2002). Deducing the origin of soluble adenylyl cyclase, a gene lost in multiple lineages. *Mol. Biol. Evol.* **19**, 2239–2246.
69. De Schrijver, A. & De Mot, R. (1999). A subfamily of MalT-related ATPdependent regulators in the LuxR family. *Microbiology*, **145**, 1287–1288.
70. Valdez, F., Gonzalez-Ceron, G., Kieser, H. M. & Servin-Gonzalez, L. (1999). The *Streptomyces coelicolor* A3(2) lipAR operon encodes an extracellular lipase and a new type of transcriptional regulator. *Microbiology*, **145**, 2365–2374.
71. Brautaset, T., Sekurova, O. N., Sletta, H., Ellingsen, T. E., StrLm, A. R., Valla, S. & Zotchev, S. B. (2000). Biosynthesis of the polyene antifungal antibiotic nystatin in *Streptomyces noursei* ATCC 11455: analysis of the gene cluster and deduction of the biosynthetic pathway. *Chem. Biol.* **7**, 395–403.
72. Wilson, D. J., Xue, Y., Reynolds, K. A. & Sherman, D. H. (2001). Characterization and analysis of the PikD regulatory factor in the pikromycin biosynthetic pathway of *Streptomyces venezuelae*. *J. Bacteriol.* **183**, 3468–3475.
73. Rascher, A., Hu, Z., Viswanathan, N., Schirmer, A., Reid, R., Nierman, W. C. *et al*. (2003). Cloning and characterization of a gene cluster for geldanamycin production in *Streptomyces hygroscopicus* NRRL 3602. *FEMS Microbiol. Letters*, **218**, 223–230.
74. Labbe, D., Garnon, J. & Lau, P. C. (1997). Characterization of the genes encoding a receptor-like histidine kinase and a cognate response regulator from a biphenyl/polychlorobiphenyl-degrading bacterium, *Rhodococcus* sp. strain M5. *J. Bacteriol.* **179**, 2772–2776.
75. Eggink, G., Engel, H., Meijer, W. G., Otten, J., Kingma, J. & Witholt, B. (1988). Alkane utilization in *Pseudomonas oleovorans*. Structure and function of the regulatory locus alkR. *J. Biol. Chem.* **263**, 13400–13405.
76. Panke, S., Meyer, A., Huber, C. M., Witholt, B. & Wubbolts, M. G. (1999). An alkane-responsive expression system for the production of fine chemicals. *Appl. Environ. Microbiol.* **65**, 2324–2332.
77. Boos, W. & Bohm, A. (2000). Learning new tricks from an old dog: MalT of the *Escherichia coli* maltose system is part of a complex regulatory network. *Trends Genet.* **16**, 404–409.
78. Danot, O. (2001). A complex signaling module governs the activity of MalT, the prototype of an emerging transactivator family. *Proc. Natl Acad. Sci. USA*, **98**, 435–440.
79. Joly, N., Danot, O., Schlegel, A., Boos, W. & Richet, E. (2002). The Aes protein directly controls the activity of MalT, the central transcriptional activator of the *Escherichia coli* maltose regulon. *J. Biol. Chem.* **277**, 16606–16613.
80. Schreiber, V. & Richet, E. (1999). Self-association of the *Escherichia coli* transcription activator MalT in the presence of maltotriose and ATP. *J. Biol. Chem.* **274**, 33220–33226.
81. Guenther, B., Onrust, R., Sali, A., O'Donnell, M. & Kuriyan, J. (1997). Crystal structure of the delta′ subunit of the clamp-loader complex of *E. coli* DNA polymerase III. *Cell*, **91**, 335–345.
82. Yu, R. C., Hanson, P. I., Jahn, R. & Brunger, A. T. (1998). Structure of the ATP-dependent oligomerization domain of N-ethylmaleimide sensitive factor complexed with ATP. *Nature Struct. Biol.* **5**, 803–811.
83. Neuwald, A. F. (1999). The hexamerization domain of N-ethylmaleimidesensitive factor: structural clues to chaperone function. *Struct. Fold. Des.* **7**, R19–R23.
84. Bochtler, M., Hartmann, C., Song, H. K., Bourenkov, G. P., Bartunik, H. D. & Huber, R. (2000). The structures of HsIU and the ATP-dependent protease HsIU-HsIV. *Nature*, **403**, 800–805.
85. Liu, J., Smith, C. L., DeRyckere, D., DeAngelis, K., Martin, G. S. & Berger, J. M. (2000). Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. *Mol. Cell*, **6**, 637–648.
86. Fodje, M. N., Hansson, A., Hansson, M., Olsen, J. G., Gough, S., Willows, R. D. & Al-Karadaghi, S. (2001).

Interplay between an AAA module and an integrin I domain may regulate the function of magnesium chelatase. *J. Mol. Biol.* **311**, 111–122.
87. Yamada, K., Kunishima, N., Mayanagi, K., Ohnishi, T., Nishino, T., Iwasaki, H. *et al*. (2001). Crystal structure of the Holliday junction migration motor protein RuvB from *Thermus thermophilus* HB8. *Proc. Natl Acad. Sci. USA*, **98**, 1442–1447.
88. Guo, F., Maurizi, M. R., Esser, L. & Xia, D. (2002). Crystal structure of ClpA, an Hsp100 chaperone and regulator of ClpAP protease. *J. Biol. Chem.* **277**, 46743–46752.
89. Krzywda, S., Brzozowski, A. M., Verma, C., Karata, K., Ogura, T. & Wilkinson, A. J. (2002). The crystal structure of the AAA domain of the ATP-dependent protease FtsH of *Escherichia coli* at 1.5 Å resolution. *Structure (Camb)*, **10**, 1073–1083.
90. Aravind, L., Makarova, K. S. & Koonin, E. V. (2000). SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucl. Acids Res.* **28**, 3417–3432.
91. Yamazaki, M., Thorne, L., Mikolajczak, M., Armentrout, R. W. & Pollock, T. J. (1996). Linkage of genes essential for synthesis of a polysaccharide capsule in *Sphingomonas* strain S88. *J. Bacteriol.* **178**, 2676–2687.
92. Cole, S. T. & Raibaud, O. (1986). The nucleotide sequence of the malT gene encoding the positive regulator of the *Escherichia coli* maltose regulon. *Gene*, **42**, 201–208.
93. Peng, H. L., Yang, Y. H., Deng, W. L. & Chang, H. Y. (1997). Identification and characterization of acoK, a regulatory gene of the *Klebsiella pneumoniae* acoABCD operon. *J. Bacteriol.* **179**, 1497–1504.
94. Jaiswal, B. S. & Conti, M. (2003). Calcium regulation of the soluble adenylyl cyclase expressed in mammalian spermatozoa. *Proc. Natl Acad. Sci. USA*, **100**, 10676–10681.
95. Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J. *et al*. (1989). Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl Acad. Sci. USA*, **86**, 6661–6665.
96. Baldauf, S. L., Palmer, J. D. & Doolittle, W. F. (1996). The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl Acad. Sci. USA*, **93**, 7749–7754.
97. Doolittle, R. F. & Handy, J. (1998). Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr. Opin. Genet. Dev.* **8**, 630–636.
98. 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.
99. Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12**, 1048–1059.
100. Pan, Q., Wendel, J. & Fluhr, R. (2000). Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J. Mol. Evol.* **50**, 203–213.
101. Bai, J., Pennill, L. A., Ning, J., Lee, S. W., Ramalingam, J., Webb, C. A. *et al*. (2002). Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. *Genome Res.* **12**, 1871–1884.
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. Tschopp, J., Martinon, F. & Burns, K. (2003). NALPs: a novel protein family involved in inflammation. *Nature Rev. Mol. Cell Biol.* **4**, 95–104.
104. Greenberg, D. B., Stulke, J. & Saier, M. H., Jr (2002). Domain analysis of transcriptional regulators bearing PTS regulatory domains. *Res. Microbiol.* **153**, 519–526.
105. Aravind, L., Anantharaman, V. & Iyer, L. M. (2003). Evolutionary connections between bacterial and eukaryotic signaling systems: a genomic perspective. *Curr. Opin. Microbiol.* **6**, 490–497.
106. Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307–311.
107. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1**, 8.
108. Iyer, L. M., Koonin, E. V. & LA (2004). Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. *Gene*, **335**, 73–88.
109. Yamada, K., Miyata, T., Tsuchiya, D., Oyama, T., Fujiwara, Y., Ohnishi, T. *et al*. (2002). Crystal structure of the RuvA-RuvB complex: a structural basis for the Holliday junction migrating motor machinery. *Mol. Cell*, **10**, 671–681.
110. Botella, M. A., Parker, J. E., Frost, L. N., Bittner-Eddy, P. D., Beynon, J. L., Daniels, M. J. *et al*. (1998). Three genes of the *Arabidopsis* RPP1 complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. *Plant Cell*, **10**, 1847–1860.
111. Lupas, A. N. & Martin, J. (2002). AAA proteins. *Curr. Opin. Struct. Biol.* **12**, 746–753.
112. Putnam, C. D., Clancy, S. B., Tsuruta, H., Gonzalez, S., Wetmur, J. G. & Tainer, J. A. (2001). Structure and mechanism of the RuvB Holliday junction branch migration motor. *J. Mol. Biol.* **311**, 297–310.
113. Zhang, X., Chaney, M., Wigneshweraraj, S. R., Schumacher, J., Bordes, P., Cannon, W. & Buck, M. (2002). Mechanochemical ATPases and transcriptional activation. *Mol. Microbiol.* **45**, 895–903.
114. Inohara, N., Koseki, T., del Peso, L., Hu, Y., Yee, C., Chen, S. *et al*. (1999). Nod1, an Apaf-1-like activator of caspase-9 and nuclear factor-kappaB. *J. Biol. Chem.* **274**, 14560–14567.
115. Linhoff, M. W., Harton, J. A., Cressman, D. E., Martin, B. K. & Ting, J. P. (2001). Two distinct domains within CIITA mediate selfassociation: involvement of the GTP-binding and leucine-rich repeat domains. *Mol. Cell. Biol.* **21**, 3001–3011.
116. Acehan, D., Jiang, X., Morgan, D. G., Heuser, J. E., Wang, X. & Akey, C. W. (2002). Three-dimensional structure of the apoptosome: implications for assembly, procaspase-9 binding, and activation. *Mol. Cell*, **9**, 423–432.
117. 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.
118. 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.
119. 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.
120. Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
121. Barton, G. J. (1993). ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.* **6**, 37–40.
122. Neuwald, A. F., Liu, J. S., Lipman, D. J. & Lawrence, C. E. (1997). Extracting protein alignment models from the sequence database. *Nucl. Acids Res.* **25**, 1665–1677.
123. Cuff, J. A. & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Struct. Funct. Genet.* **40**, 502–511.
124. Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**, 525–539.
125. Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S. *et al*. (2003). CDD: a curated Entrez database of conserved domain alignments. *Nucl. Acids Res.* **31**, 383–387.
126. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S. *et al*. (2004). The Pfam protein families database. *Nucl. Acids Res.* **32**, D138–D141.
127. Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J. *et al*. (2004). SMART 4.0: towards genomic data integration. *Nucl. Acids Res.* **32**, D142–D144.
128. 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.
129. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1245.
130. Swofford, D. L. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.* Sinauer Associates, Sunderland, MA.
131. Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357–358.
132. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proc. Natl Acad. Sci. USA*, **87**, 4576–4579.
133. Yeats, C., Bentley, S. & Bateman, A. (2003). New knowledge from old: *in silico* discovery of novel protein domains in *Streptomyces coelicolor*. *BMC Microbiol.* **3**, 3.

Edited by J. Thornton

(Received 14 May 2004; received in revised form 27 July 2004; accepted 10 August 2004)