Evolution of Two-Component Signal Transduction

Kristin K. Koretke, Andrei N. Lupas, Patrick V. Warren, Martin Rosenberg, and James R. Brown

SmithKline Beecham Pharmaceuticals, Collegeville, Pennsylvania

Two-component signal transduction (TCST) systems are the principal means for coordinating responses to environmental changes in bacteria as well as some plants, fungi, protozoa, and archaea. These systems typically consist of a receptor histidine kinase, which reacts to an extracellular signal by phosphorylating a cytoplasmic response regulator, causing a change in cellular behavior. Although several model systems, including sporulation and chemotaxis, have been extensively studied, the evolutionary relationships between specific TCST systems are not well understood, and the ancestry of the signal transduction components is unclear. Phylogenetic trees of TCST components from 14 complete and 6 partial genomes, containing 183 histidine kinases and 220 response regulators, were constructed using distance methods. The trees showed extensive congruence in the positions of 11 recognizable phylogenetic clusters. Eukaryotic sequences were found almost exclusively in one cluster, which also showed the greatest extent of domain variability in its component proteins, and archaeal sequences mainly formed species-specific clusters. Three clusters in different parts of the kinase tree contained proteins with serine-phosphorylating activity. All kinases were found to be monophyletic with respect to other members of their superfamily, such as type II topoisomerases and Hsp90. Structural analysis further revealed significant similarity to the ATP-binding domain of eukaryotic protein kinases. TCST systems are of bacterial origin and radiated into archaea and eukaryotes by lateral gene transfer. Their components show extensive coevolution, suggesting that recombination has not been a major factor in their differentiation. Although histidine kinase activity is prevalent, serine kinases have evolved multiple times independently within this family, accompanied by a loss of the cognate response regulator(s). The structural and functional similarity between TCST kinases and eukaryotic protein kinases raises the possibility of a distant evolutionary relationship.

Introduction

Two-component signal transduction (TCST) pathways form the central signaling machinery in bacteria (for reviews, see Stock, Ninfa, and Stock 1989; Parkinson and Kofoid 1992; Hoch and Silhavy 1995). In response to a stimulus, typically extracellular, the kinase component autophosphorylates at an internal histidine (the H-box). The high-energy phosphate group is then transferred to an aspartyl residue on the response regulator component (hence, ''two-component'' signal transduction), which modifies cellular behavior via an effector domain (fig. 1). Although most systems use a linear phosphorelay from one kinase to one response regulator, some use more complicated paths, involving a branching of the signal (chemotaxis) or multiple phosphorylated components (sporulation, adaptation to anaerobic conditions).

In bacteria, TCST systems mediate adaptive responses to a broad range of environmental stimuli. These include citrate uptake and catabolism (Cit), aerobic respiration (Arc), osmoregulation (EnvZ/OmpR), stress-induced sporulation (Kin/Spo), *N*-acetylmuramoyl-L-alanine amidase biosynthesis (Lyt), nitrate and nitrite metabolism (Nar), nitrogen regulation (Ntr), phosphate regulation (Pho), host recognition for pathogen invasion (Vir), and chemotaxis (Che) (Stock, Ninfa, and Stock 1989; Parkinson and Kofoid 1992; Hoch and Silhavy 1995). TCST systems also exist in certain non-animal eukaryotes and in some Archaea (Alex and Simon 1994; Loomis, Shaulsky, and Wang 1997). In plants, they mediate photosensitivity (Schneider-Poetsch et al. 1991; Yeh and Lagarias 1998) and ethylene response (Chang et al. 1993); in fungi, they mediate osmoregulation (Maeda, Wurgler-Murphy, and Saito 1994; Krems, Charizanis, and Entian 1996; Posas et al. 1996) and hyphal development (Alex, Borkovich, and Simon 1996; Alex et al. 1998); and in the slime mold, they mediate *Dictyostelium discoideum* osmoregulation (Schuster et al. 1996) and fruiting body formation (Singleton et al. 1998). A thorough cataloguing of protein kinases found in the genome of the nematode *Caenorhabditis elegans* failed to find any true orthologs to prokaryotic histidine kinases, which suggests that TCST systems do not occur in metazoans (Plowman et al. 1999).

Histidine kinases and response regulators are modular proteins, containing multiple homologous and heterologous domains (Stock, Ninfa, and Stock 1989; Parkinson and Kofoid 1992; Hoch and Silhavy 1995). The three domains required for phosphotransfer, corresponding to the kinase, the H-box, and the response regulator, are homologous in all TCST systems and represent their defining element. In addition, histidine kinases generally contain an N-terminal transmembrane sensory domain and response regulators generally contain a C-terminal effector domain; these are specific to individual TCST systems and determine their specificity. Some kinases and response regulators also contain PAS domains, which enable them to sense the redox potential (Zhulin, Taylor, and Dixon 1997); SH3-like domains (Bilwes et al. 1999), which appear to mediate protein complex formation; or a second type of His-acceptor domain called

Key words: two-component signal transduction, coevolution, evolution of histidine kinase domain.

Address for correspondence and reprints: James R. Brown, SmithKline Beecham Pharmaceuticals, 1250 South Collegeville Road, UP1345, Collegeville, Pennsylvania 19426-0989. E-mail: james_r_brown@sbphrd.com.

Mol. Biol. Evol. 17(12):1956–1970. 2000

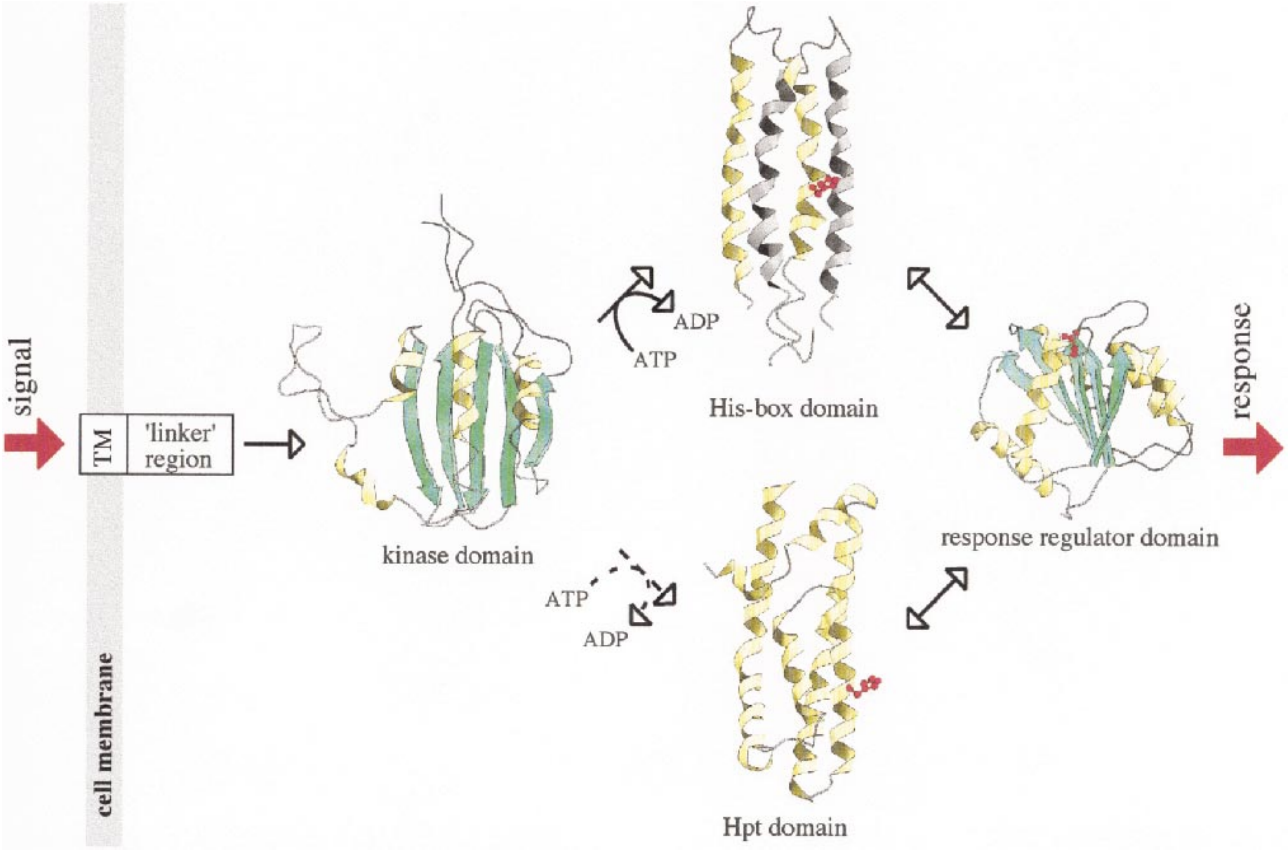


FIG. 1.—Schematic diagram illustrating phosphorelay in two-component signal transduction systems. The signal (usually coming from the extracellular environment) is transduced to the ''linker'' region, which is located adjacent to the last transmembrane helix of the membrane-bound protein. The linker interacts with the kinase domain to induce autophosphorylation at a histidine (shown in red) located in the His-box domain or—in CheA—in the Hpt domain. The phosphate is then transferred to an aspartate residue (shown in red) in a response regulator domain. The phosphorylated response regulator elicits the appropriate response within the cell (typically via the DNA-binding activity of an effector domain). In some systems, an Hpt domain serves as a regulatory phosphate sink for the phosphorylated response regulator.

Hpt, which serves as a regulatory phosphate sink (Kato et al. 1997; Xu and West 1999). Most domains found in TCST systems can occur either as stand-alone proteins or within larger polypeptides, and in one class of kinases, all domains required for phosphotransfer are found on the same polypeptide chain (hybrid kinases).

An exception to the phosphorelay described here is found in the chemotaxis kinase CheA, where the H-box has lost the catalytic histidine and is used exclusively for dimerization, while an Hpt domain that may have originally served a regulatory function is now used for phosphorelay to CheY and CheB (Bilwes et al. 1999; Dutta, Qin, and Inouye 1999). Further variability in His-acceptor domains is seen in the Spo0B protein, whose H-box domain forms a dimeric four-helix bundle similar to canonical H-boxes but of opposite handedness (Varughese 1998).

In addition to the three phosphorelay domains described above, a fourth domain conserved broadly in TCST systems has recently been recognized (Park and Inouye 1997; Aravind and Ponting 1999). Termed the ''linker'' region (or HAMP domain), it is typically found at the C-terminal end of the last transmembrane segment in many histidine kinases, chemoreceptors, bacterial nucleotidyl cyclases, and phosphatases, and mutations show that it plays a critical role in signal transduction. Some proteins contain multiple copies in tandem, suggesting that it represents an autonomously folding unit. Its ability to regulate kinase activity in *trans* (e.g., between chemoreceptors and CheA) and the variable nature of the segments connecting it to the H-box suggest that it acts through direct interaction with the kinase domain rather than through propagation of conformational change along the polypeptide chain. Thus, four protein domains appear to be typically involved in the signal transduction pathway from the extracellular sensor domain to the cytoplasmic effector (fig. 1).

TCST systems represent one of the most studied and best understood areas of bacterial physiology. Recently, they have also emerged as attractive targets for anti-microbial drug development (Barrett et al. 1998; Lange et al. 1999; Throup et al. 2000). Here we report the results of a detailed phylogenetic and structural study of genomic TCST sequences, undertaken to explore the origin and evolution of TCST systems.

Materials and Methods

Database Searches

Complete genomic sequences of the Bacteria (12 species), Archaea (4 species), and eukaryotes (2 species

and other eukaryotic GenBank entries) were searched for proteins homologous to histidine kinases and response regulators from *Escherichia coli* (species listed in table 1). *Escherichia coli* was selected as the source of query sequences, since the biochemical and genetic characteristics of TCST pathways in this species are well known (reviewed in Stock, Ninfa, and Stock 1989; Hoch and Silhavy 1995). In order to collect all possible TCST proteins, two separate searches of the complete genome sequence of *E. coli* were performed with the program PSI-BLAST (Altschul et al. 1997) applying an EXPECT threshold of 0.01. The H-box and kinase domains of the histidine kinase phoR and the associated response regulator receiver domain phoB were used as query sequences for these searches. The histidine kinase search converged after five iterations, revealing a total of 32 open reading frames (ORFs) with significant sequence similarity to phoR (P[N] $\leq$ 0.01), while the response regulator search converged after three iterations, revealing 37 ORFs with significant similarity to phoB.

The identified *E. coli* histidine kinases and response regulators were then used as query sequences to search the other genomic databases for homologs using PSI-BLAST. A subject protein was considered a putative histidine kinase if the conserved H, N, D, and G boxes were present in order. A particular ORF was considered to be homologous to an *E. coli* response regulator if all three conserved boxes, D1, D2, and K, were found in order.

The kinase domain of *E. coli* CheA was the query sequence for the SENSER run used to identify homologous kinase/nucleotide-binding domains within the nonredundant database. An identified sequence was validated as a kinase/nucleotide-binding domain by the presence of the N, D, G, and, possibly, F boxes.

Multiple-Sequence Alignment and Phylogeny

Similar alignment and phylogenetic methodologies were applied to histidine kinase/PDK and response regulator data sets. Full-length proteins were initially aligned using the program CLUSTAL W, version 1.8 (Thompson, Higgins, and Gibson 1994), with the BLOSUM62 (Henikoff and Henikoff 1992) similarity matrix and gap opening and extension penalties of 10.0 and 0.05, respectively. The alignment of conserved sequence blocks was later refined using the program MACAW (Schuler, Altschul, and Lipman 1991). Since MACAW cannot analyze more than 32 sequences at once, the sequences were subdivided according to the CLUSTAL W clustering into groups of 32 or fewer sequences. Each subgroup was aligned in MACAW using the options of pairwise segment overlap and Gibbs sampler based on the BLOSUM62 similarity matrix. The aligned subgroups of sequences were then concatenated into a single large multiple-sequence alignment, which was further refined manually using the program SEQLAB of the GCG, version 9.0, software package (Womble 2000). The final alignment for either protein family included all accepted members of the family, with positions truncated down to the most conserved regions (133 and 107 amino acid residues for histidine kinase and response regulator alignments, respectively).

To align the kinase/nucleotide-binding domains, a database of all sequences identified from SwissProt was generated. With the *E. coli* CheA kinase domain as a query sequence, we used PSI-BLAST to search against the database. The individual alignments from the converged PSI-BLAST run were extracted and converted into a multiple-sequence alignment. Thirty histidine kinases, four anti-sigma factors, four phosphate dehydrogenase kinases, six DNA mismatch repair proteins (mutL), four topoisomerase VI proteins, six heat shock 90 proteins, three gyraseB proteins, and three topoisomerase II proteins were randomly selected from the complete alignment. This smaller multiple-sequence alignment was then further refined, taking into consideration hydrophobicity patterns and secondary-structure units between the different protein families, and then truncated down to the most conserved residues/secondary-structure units (82 residues). Accession numbers for all sequences and the multiple-sequence alignments are available from one of the authors (kristin_k_koretke@sbphrd.com) on request.

Phylogenetic trees were constructed by maximum parsimony and distance methods for each set of alignments. A distance matrix of pairwise comparisons of the proportion of different amino acids per site was constructed using the program PROTDIST of the PHYLIP, version 3.572c, package (Felsenstein 1993). In our analysis, we invoked the ''Dayhoff'' program option, which estimates the expected amino acid replacements per position (EAARP) using a replacement model based on the Dayhoff 120 matrix (Dayhoff, Eck, and Park 1972). The programs SEQBOOT, NEIGHBOR, and CONSENSE were used to derive a neighbor-joining (NJ) tree that was replicated in 500 bootstraps. Maximum-parsimony (MP) analysis was done using the program PAUP*, version 4.0 (Swofford 1999). The number and length of minimal trees were estimated by 100 replicate random heuristic searches. Confidence limits for the branch points were estimated by 1,000 bootstrap replicate random heuristic searches. Quartet maximum-likelihood (ML) analysis was attempted on these data sets as well, using the program PUZZLE version 4.0.2 (Strimmer and von Haeseler 1996). However, the low ratio of aligned residues to operational taxonomic units (OTUs) in both histidine kinase and response regulator data sets tended to make resolution of major branch points in ML trees difficult; therefore, those trees were not reported.

Results

Phyletic Distribution of TCST Proteins

Genes encoding histidine kinases and response regulators are found in all three domains, or urkingdoms, of life: Archaea, Bacteria, and Eucarya (Woese, Kandler, and Wheelis 1990). Across a set of representative genomes, we identified a total of 183 histidine kinase and 220 response regulator homologs (see table 1). In the following, we will refer to these proteins as histidine kinases and response regulators, even though in most

Table 1
Global Distribution of Histidine Kinase and Response Regulator Classes Across Different Species

	Afu		Mth		Hybrid		Che		Cit		Lyt		Nar		Ntr		Pho		Syn		Rest		Total	
Species	Hk	Rr	Hk	Rr	Hk	Rr	A	Y, V, B	Hk	Rr	Hk	Rr	Q	Rr	B	C	R	B	Hk	Rr	Hk	Rr	Hk	Rr
Bacteria																								
Ec	0	0	0	0	5	5	1	2	2	2	2	2	3	5	3	4	12	14	1	0	0	3	29	37
Hi	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	2	4	0	0	0	0	4	5
Hp	0	0	0	0	1	0	1	5	0	0	0	0	0	0	1	1	1	1	0	0	0	3	4	10
Rp	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	3	4	4
Bs	0	0	0	0	0	1	1	4	3	3	3	3	9	9	5	0	12	13	0	0	2	2	35	35
Mtu	0	0	1	1	0	0	0	0	0	0	0	0	3	2	0	0	8	9	1	0	0	0	13	12
Mg/Mp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bb	0	0	0	0	1	1	2	4	0	0	0	0	0	0	1	1	0	0	0	0	0	1	4	7
Tp	0	0	0	0	0	0	1	2	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2	3
Aa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	1	1	0	0	0	0	3	4
Syn	0	0	1	1	12	11	3	4	0	0	0	0	0	5	0	0	8	13	12	21	2	6	38	61
Archaea																								
Af	12	7	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	2	14	11
Mj	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mth	0	0	15	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	16	10
Ph	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
Ta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eucarya																								
Ca	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
Cc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Dd	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	4
Nc	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Sc	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
Sp	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
At	0	0	0	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3
Pf	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ce/Dm/Hs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note.—Histidine kinases (Hk or gene name) and response regulator (Rr or gene name) classes correspond to the functional and species-specific clusters shown in figure 2. These include proteins specific to *Archaeoglobus fugidulus* (Afu), *Methanobacterium thermoautotrophicum* (Mth), and Synechocystis (Syn), histidine kinase and response regulator fusions (Hybrid), and functional characterized TCST families Che, Cit, Lyt, Nar, Ntr, and Pho. Uncharacterized histidine kinases and response regulators (Rest) are also given. The total numbers of putative histidine kinases and response regulators in each species are also given. For the Bacteria and the Archaea, only complete genome sequences are reported. Among the Bacteria, these species include the Proteobacteria, *Escherichia coli* (Ec), *Haemophilus influenzae* (Hi), *Helicobacter pylori 26695* (Hp), and *Rickettsia prowazekii* (Rp); the Gram-positive bacteria, *Bacillus subtilis* (Bs), *Mycobacterium tuberculosis* (Mtu), *Mycoplasma genitialium* (Mg), and *Mycoplasma pneumoniae* (Mp); the spirochaetes, *Borrelia burgdorferi* (Bb) and *Treponema pallidum* (Tp); an extreme thermophilic bacterium, *Aquifex aoelicus* (Aa); and a cyanobacterium, *Synechocystis* PCC 6803. The Archaea include *A. fuligidus* (Af), *Methanococcus jannaschii* (Mj), *M. thermoautotrophicum* (Mth), *Pyrococcus horikishii* (Ph), and *Thermoplasma acidophilum* (Ta). Among the Eucarya, the genomes searched include the fungi *Candida albicans* (Ca), *Cyanidium caldarium* (Cc), *Neurospora crassa* (Nc), *Saccharmyces cerevisiae* (Sc), and *Schizosaccharomyces pombe* (Sp); the slime-mold *Dictostylium doscoideum* (Dd); the protist *Plasmodium falciparum* (Pf); the plant *Arabodopisis thaliana* (At); and the animals *Caenorhabiditis elegans* (Ce), *Drosophila melanogaster* (Dm), and *Homo sapiens* (Hs).

cases no experimental evidence for such activity has been obtained, and occasionally, residues known to be critical for phosphotransfer are missing (see *Materials and Methods* for criteria of selection). The largest number of TCST proteins in a single genome (39 histidine kinases and 55 response regulators) was found in the cyanobacterium Synechocystis PCC 6803. It is unclear, however, whether other cyanobacteria will also be found to contain a large TCST complement, since large variations were observed between related bacterial species. Thus, among the proteobacteria, the number of TCST proteins ranged from 66 in *E. coli* to 9 in *Haemophilus influenzae* and 8 in *Rickettsia prowazekii*; among Gram-positive bacteria, the number ranged from 70 in *Bacillus subtilis* to none in *Mycoplasma genitalium* and *Mycoplasma pneumoniae*; and among spirochetes, the number ranged from 10 in *Borrelia burgdorferi* to 5 in *Treponema pallidum* (table 1). Large variations were also seen between organisms from the same environmental niche, for example, between the intracellular pathogens Mycoplasma and Rickettsia. However, in general, more TCST proteins were found in free-living species (Escherichia, Bacillus, Synechocystis) than in pathogenic ones (Haemophilus, Rickettsia, Mycoplasma).

Among the Archaea, *Methanobacterium thermoautotrophicum* and *Archaeoglobus fulgidus* contained the largest numbers of histidine kinases (16 and 14, respectively) and response regulators (10 and 11, respectively), although they still contained fewer than free-living bacteria. *Pyrococcus horikoshii* had only a single histidine kinase and two response regulators (corresponding to the chemotaxis proteins CheA, CheY, and CheB), while *Methanococcus jannaschii, Aeropyrum pernix,* and *Thermoplasma acidophilum* had none. Among eukaryotes, only fungi, slime molds, and plants appeared to contain TCST proteins, and these typically occurred in small numbers. The complete genome of the yeast *Saccharomyces cerevisiae* yielded only a single histidine kinase and three response regulators, while none were found in the partial genome of the apicomplexan protist *Plasmodium falciparum.* No histidine kinases were found in the genomes of human, *Drosophila melanogaster,* or *C. elegans.* The latter search confirmed findings of an earlier study of nematode signal transduction pathways (Plowman et al. 1999) and provided independent verification of our search strategies. In total, five complete genomes and one partial genome of those surveyed in this study lacked TCST proteins.

Phylogenetic Analyses of TCST Proteins

Phylogenetic trees based on the NJ method showed extensive similarity in the clustering of cognate histidine kinases and response regulators (fig. 2). Despite the low ratio between the number of aligned residues and the number of OTUs , which restricted the resolution of the phylogenetic analyses, provisional bootstrapping support was obtained for 7 of 11 major phylogenetic clusters in the histidine kinase tree and for 6 of 12 clusters in the response regulator tree (i.e., these occurred in >50% of 500 random bootstrap replicates; indicated by black dots in fig. 2). Further support for the histidine kinase NJ tree topology was obtained in MP analyses, where only 10 minimal-length trees, each of 9,910 steps, were found after 100 random replicate heuristic searches. A strict consensus tree revealed that the 10 MP trees differed only in the rearrangement of some terminal taxa and confirmed the main clusters derived in the NJ tree. However, MP analysis of the response regulator alignment failed to converge on a small number (<100) of minimal-length trees, which was likely due to a lower ratio of aligned residues to OTUs than in the histidine kinase alignment.

The TCST proteins of the archaea *A. fulgidus* and *M. thermoautotrophicum* primarily formed species-specific clusters, labeled Arf and Mth, respectively, in figure 2. Strong bootstrap support was obtained both for histidine kinase clusters and for the Arf response regulator cluster. Whereas the Arf clusters contained only proteins from Archaeoglobus, the Mth clusters also contained one protein each from *Mycobacterium tuberculosis* and Synechocystis. Two histidine kinases of Archaeoglobus and one of Methanobacterium, as well as four response regulators of Archaeoglobus and three of Methanobacterium, occurred elsewhere in the tree. Of these ''errant'' proteins, one histidine kinase and one response regulator from each organism formed bootstrap-supported clusters with proteins of Synechocystis (SAM), while one response regulator of Archaeoglobus and two of Methanobacterium clustered together outside the major phylogenetic groups. The remaining errant proteins of Archaeoglobus corresponded to CheA, CheY, and CheB and clustered accordingly.

Three other pairs of clusters (Lyt, Cit, and Che) received significant bootstrap support. Lyt was named for the LytS and LytT proteins, involved in *N*-acetylmuramoyl-L-alanine amidase biosynthesis (Lazarevic et al. 1992), and Cit was named for CitA and CitB, involved in the expression of citrate-specific fermentation genes (Bott, Meyer, and Dimroth 1995). Che was named for the chemotaxis proteins (Stock, Ninfa, and Stock 1989; Parkinson and Kofoid 1992; Stock et al. 1993; Hoch and Silhavy 1995; Bilwes et al. 1999), of which it is exclusively composed. The kinases (orthologous CheA proteins from different organisms) clustered together with strong bootstrap support, but the response regulators (CheB, CheY, and CheV) formed three separate clusters. Two of these, one containing CheB proteins and the other CheY proteins from Gram-positive bacteria, spirochetes, and archaea, received significant bootstrap support, whereas the third one, formed by CheV and by CheY proteins from proteobacteria, did not.

The four remaining pairs of clusters showing strong correlation (Pho, Ntr, Nar, and Hybrid) were by far the largest ones and, correspondingly, had the lowest resolution. Pho contained TCST systems involved in phosphate regulation (PhoR/PhoB; Stock, Ninfa, and Stock 1989; Hoch and Silhavy 1995), virulence (PhoQ/PhoP; Hoch and Silhavy 1995), osmoregulation (EnvZ/OmpR; Hoch and Silhavy 1995), and anaerobic nitrite reduction (ResD/ResE; Hoch and Silhavy 1995; Nakano et al.

1998); Ntr contained systems regulating nitrogen assimilation (NtrB/NtrC and NtrY/NtrX; Stock, Ninfa, and Stock 1989; Hoch and Silhavy 1995), acetoacetate metabolism (AtoS/AtoC; Jenkins and Nunn 1987), and hydrogenase activity (HydH/HydG; Stoker et al. 1989); and Nar contained regulators of anaerobic respiration (NarQ/NarP and NarX/NarL; Darwin et al. 1998), sugar phosphate uptake (UhpB/UhpA; Hoch and Silhavy 1995), and degradative enzyme expression (DegS/DegU; Hoch and Silhavy 1995). The Hybrid cluster pair contained all eukaryotic kinases and all but two response regulators, in agreement with a previous analysis (Pao and Saier 1997), as well as many bacterial TCST systems, particularly from *E. coli* and Synechocystis. This cluster was named for the fact that approximately two thirds of its members, including all eukaryotic kinases except phytochrome, were hybrid kinases (i.e., they contained kinase and response regulator domains within the same polypeptide). All bacterial hybrid kinases except for five proteins of Synechocystis (three of them CheA homologs) fell into this cluster.

One major cluster, which appeared in similar locations in both trees, was labeled Syn, as it was formed almost exclusively by proteins from Synechocystis. It contained a single ''errant'' response regulator, from *B. burgdorferi,* and three ''errant'' kinases, one each from *E. coli, R. prowazekii,* and *M. tuberculosis.* No correlation between the two trees could be established for Syn, as only one kinase in this cluster has been studied experimentally (*E. coli* KdpD; Jung, Tjaden, and Altendorf 1997), and it signals through a response regulator of the Pho cluster (KdpE).

Despite the absence of bootstrap support for some of the largest clusters in our phylogenetic trees, most kinases and response regulators from experimentally studied TCST systems were found in cognate clusters. However, several TCST kinases, particularly among the hybrid class, appear to have recruited additional response regulators from noncognate phylogenetic clusters, such as ArcB (Georgellis et al. 1998) and TorS (Simon et al. 1994), which signal via the Pho cluster regulators ArcA and TorR, respectively, or RcsC (Hoch and Silhavy 1995), which signals via the Nar cluster regulator RcsB.

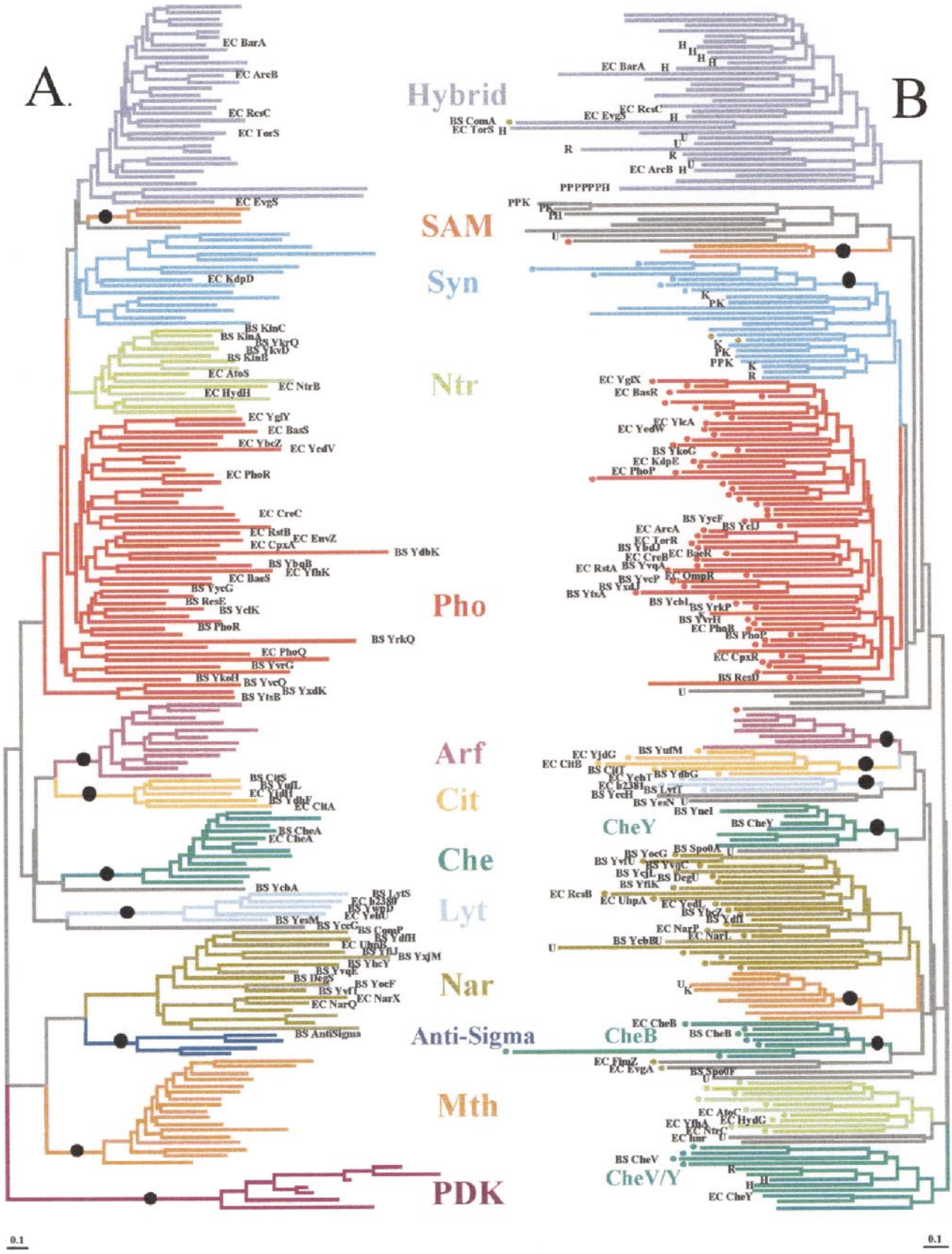
Several interesting observations follow from the phylogenetic clusters presented here: No cluster contained proteins from both Archaea and eukaryotes, although a specific evolutionary relationship has long been postulated between these two groups (reviewed in Brown and Doolittle 1997). Bacterial phylogeny similarly did not correlate well with the observed clustering. Despite the considerable sizes of some clusters, none contained representatives from each bacterial species, and no bacterium had a representative in all of the clusters. Among bacteria, no cluster predominated across all species: Pho contained nearly a third of all TCST proteins detected in *E. coli, B. subtilis,* and Synechocystis and two thirds of those in *M. tuberculosis,* but none from spirochetes. In the latter, Che which is missing from *A. aoelicus* and *M. jannaschii,* was predominant (even though both organisms encode flagellar proteins and are motile).

Three phylogenetically distinct groups within the kinase tree have serine rather than histidine kinase activity and do not act in conjunction with a response regulator. These are the plant phytochromes (Yeh and Lagarias 1998), found in the Hybrid cluster, the bacterial anti-sigma factors (Schurr et al. 1996), and the mitochondrial pyruvate dehydrogenase kinases (PDKs; Popov et al. 1993; Thelen et al. 1998). The latter two formed separate clusters with high bootstrap support (fig. 2). Anti-sigma factors were suggested to be evolutionarily linked to the Nar cluster, whereas PDKs formed a distinct outgroup to all histidine kinases (no functional outgroup to response regulators was identified). As the outgroup, PDK sequences were not counted among the 183 kinases in this study.

Concordance of Linkage Groups and Carboxyl-Terminal Domain Structure with Phylogeny

Recent studies have highlighted the connection between the functional coupling of genes and their chromosomal vicinity, in the form of either gene fusions (Enright et al. 1999; Marcotte et al. 1999) or gene clusters (Dandekar et al. 1998; Overbeek et al. 1999). Of the 183 kinases in this study, 28 were hybrid kinases (''gene fusions'') and 84 were concurrent on the chromosome with response regulator genes, either as part of an operon or within 20 bases of one another (''gene clusters''). Among these, 25 hybrid kinases and 75 chromosomally clustered TCST systems had their kinase and response regulator modules in cognate phylogenetic groups (fig. 3). Eight of the 28 hybrid kinases were also concurrent on the chromosome with a separate response regulator; however, none of these gene pairs clustered in cognate groups (see also the previous section). The extensive concordance between chromosomal linkage and phylogenetic classification supports the validity of the analyses presented here.

Further support was obtained from an analysis of the carboxyl-terminal domains of response regulators, which elicit the appropriate adaptive response, typically through DNA binding. With the exception of seven proteins, all response regulator domains with homologous carboxyl-terminal domains fell within the same cluster (indicated by dots in fig. 2). The phylogenetic position of the errant proteins may be due to the limited resolution of the response regulator tree, rather than to domain shuffling. In agreement with this view, some of the errant proteins were found in their cognate clusters in preliminary phylogenetic analyses, and none of the response regulator clusters with bootstrap support had members among the errant proteins or contained errant proteins themselves. Because of the lower resolution of the response regulator phylogeny, the support obtained from this analysis was particularly useful for the major clusters that had not obtained bootstrap support (Pho, Nar, and Ntr). In the Pho cluster, only three proteins did not have a recognizable ''winged-helix'' DNA-binding domain (Martinez-Hackert and Stock 1997), and only



two proteins with a winged-helix domain were found outside the cluster. The corresponding numbers for Nar, whose carboxyl-terminal domains contain a ''helix-turn-helix'' DNA-binding motif (Baikalov et al. 1996), were three and five, respectively. Finally, in the Ntr cluster, whose carboxyl-terminal domains also contain a helix-turn-helix motif (Volkman et al. 1995) but are preceded by a σ54-interacting ATPase domain (Weiss et al. 1991), only proteins with an Ntr-like DNA-binding domain were found, and no Ntr-like proteins occurred outside the cluster.

Origin of the Histidine Kinase Fold

Response regulators are not recognizably related to other known protein families beyond a general structural similarity to P-loop NTPases (such as Ras) (Lukat et al. 1991; Stock et al. 1993), which has hitherto not been considered sufficient to infer homology. Histidine kinases, however, are clearly related to Hsp90, MutL, and type II topoisomerases in the ATP-binding domain (Tanaka et al. 1998; Bilwes et al. 1999; Dutta, Qin, and Inouye 1999). The conserved structural core consists of an antiparallel, four-stranded β-sheet flanked on one side by three α-helices, which surround the ATP-binding site. In addition, an αβ element, which is in an equivalent structural position, is circularly permuted in the sequence of histidine kinases relative to other proteins with this fold. The structural similarity is mirrored in a set of conserved sequence motifs, primarily associated with nucleotide binding, which strongly imply descent from a common ancestor (fig. 4*A*). Phylogenetic analysis of the sequence data by distance methods indicates that all kinases in this superfamily arose from a single ancestral protokinase (fig. 4*B*). Because of the low branch point of PDKs, it is unclear whether this ancestor had Ser- or His-phosphorylating activity.

We used a sensitive sequence comparison routine, SENSER (Golbik et al. 1999; Koretke et al. 1999), to search for protein families distantly related to histidine kinases and response regulators. SENSER uses exclusively sequence information, yet has a sensitivity comparable to that of advanced fold recognition methods (Koretke et al. 1999). Searches with response regulator sequences did not uncover protein families that could either serve as phylogenetic outgroups or be used for further understanding of the evolutionary origin of response regulators. Searches with histidine kinase ATPase domains, however, suggested a distant similarity to the small lobe of protein kinases. The structures of both protein folds are known and have not hitherto been considered similar, but on investigation, we found that the suggested similarity was centered on a topologically equivalent αββ element (yellow in fig. 5*A*) and that an additional β-hairpin has a similar structural position but is circularly permuted in protein kinases (green in fig. 5*A*). In both folds, the nucleotide is bound in a similar location and with a similar orientation relative to the protein, and important nucleotide interactions are produced by two glycine-containing loops which originate from the same parts of the structure (blue in fig. 5*A*). There are also substantial differences between the two folds. There are no highly conserved residues in equivalent structural positions, except for a glutamate residue (black in fig. 5*B*). However, this residue activates a water molecule during nucleotide hydrolysis in topoisomerases and Hsp90, whereas it positions an invariant lysine residue for interaction with the α and β phosphates of the nucleotide in protein kinases. Also, the angles and distances of the nucleotide relative to the β-sheet are different in the two folds (fig. 5*B*).

Discussion

Histidine Kinases and Response Regulators Have Coevolved

In this study, we analyzed the phylogenetic relationships between the TCST systems from 14 complete and 6 partial genomes. Their components represent highly evolved multigene families, and the pattern and process of proliferation of such interacting yet structurally unrelated proteins is an important problem in evolutionary biology. A priori, two competing models for the evolution of novel TCST systems appear plausible: the recruitment model and the coevolution model. The recruitment model suggests that novel TCST systems evolve through gene duplication of one component, which then co-opts components from heterologous systems to yield a new specificity. This model is supported by the structural similarity of response regulators, in which only a few residues are sufficient to determine specificity, and by the observed crosstalk of TCST systems within an organism. From the phylogenetic perspective, this mechanism would result in an incongruent clustering of cognate histidine kinases and response regulators. The coevolution model suggests that novel

←

FIG. 2.—Phylogenetic trees of (*A*) histidine kinase and (*B*) response regulators generated using the neighbor-joining (NJ) distance method. Different colors and labels indicate histidine kinases and response regulators which either belong to the same functional signal transduction cascade based on the membership of previously characterized two-component signal transduction systems or are from the same species. Coding of C-terminal domain comparisons are as follows: (1) colored dots represent domains that are homologous within a class; (2) ''U'' indicates domains that are unique among the response regulator proteins; (3) ''K'' indicates a complete histidine kinase domain C-terminal to the response regulator; (4) ''H'' indicates an Hpt domain C-terminal to the response regulator; and (5) ''P'' indicates a PAS/PAC domain C-terminal to the response regulator. Response regulator proteins lacking a C-terminal region are blank. Histidine kinase trees are based on multiple-sequence alignments for the histidine kinases/pyruvate dehydrogenase kinases and response regulators comprised of 190 and 220 protein sequences, respectively. Multiple-sequence alignments were edited to include only the most readily aligned amino acid positions, which were 133 and 107 residues for the histidine kinases and response regulators, respectively. The NJ phylogeny was constructed using the program NEIGHBOR of the PHYLIP, version 3.57c, package (Felsenstein 1993). The scale bar represents 0.1 expected amino acid replacements per site, as estimated by the program PROTDIST using the Dayhoff PAM substitution matrix. Nodes occurring in >50% of 500 random bootstrap replicates in the NJ distance trees are indicated with black dots.

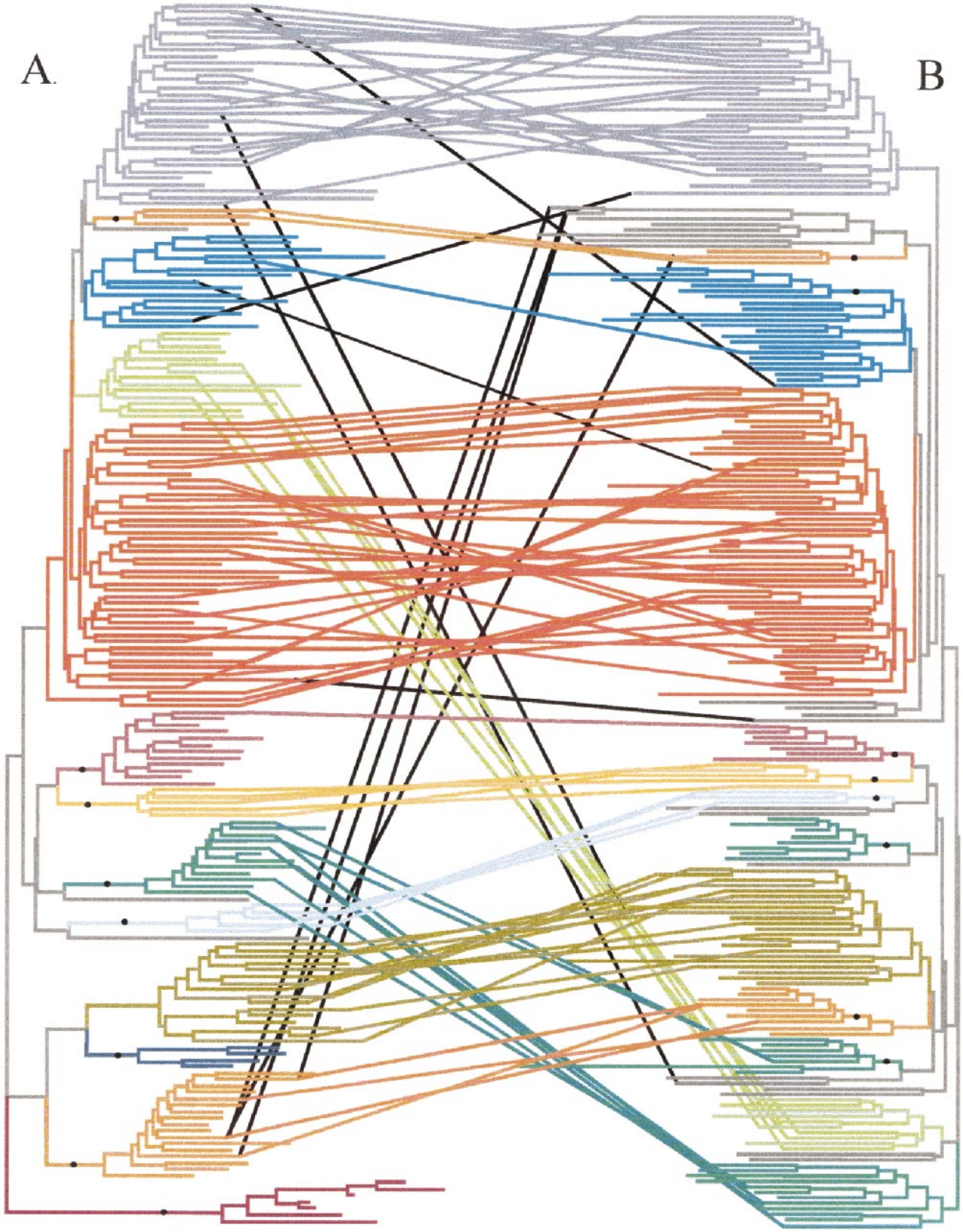


FIG. 3.—The concordance of histidine kinase (*A*) and response regulatory (*B*) protein phylogeny with chromosome location of coding genes. Neighbor-joining trees are identical to those in figure 2. Lines link histidine kinase and response regulator proteins which either are encoded by concurrent genes on the genome or are fused polypeptides. Colored lines connect histidine kinase and response regulator proteins belonging to the same signaling pathway or species phyletic cluster. Black lines indicate those proteins which are not in the same phyletic cluster although they are physically linked on the chromosome.

A

```
                hhhhhhhhhhhh                    sssssss                 sssss                              hhhhh
chea_ecoli  riIDPLTHLVRNSLDHgielpekrlaagknsv..GNLILSAEhqgg.............NICIEVTDDGAGLnrerilakaasqgltvsenmsddEVAML
envz_ecoli  .iKRAVANMVVNAARYgn...............GWIKVSSGtepn.............RAWFQVEDDGPGIap.....................EQRKH
ntrb_ecoli  qiEQVLLNIVRNALQAlgpeg............GEIILRTRtafqltlhgeryrl...AARIDVEDNGPGIpp.....................HLQDT
phor_ecoli  qlRSAISNLVYNAVNHtpeg.............THITVRWQrvph.............GAEFSVEDNGPGIap.....................EHIPR
arcb_ecoli  rlRQILWNLISNAVKFtqg..............GQVTVRVRydegd............MLHFEVEDSGIGIpq.....................DELDK
etr1_arath  rlMQIILNIVGNAVKFskqgsi............SVTALVTKsdtraadffvvptgshfYLRVKVKDSGAGInp.....................QDIPK
phy1_tobac  rlQQVLANFLLVCVNStpsg..............GQLSISGTltkdrigesvqla....LLEVRISHTGGGVpe.....................ELLSQ
narx_ecoli  .lLQIAREALSNALKHsqa...............SEVVVTVAqndn.............QVKLTVQDNGCGVpenairsn.....................
sp22_bacsu  eiKTVVSEAVTNAIIHgyeence...........GKVYISVTledh.............VVYMTIRDEGLGItdle.....................EARQP
pdk1_human  hlYHMVFELFKNAMRAtmehhanrgvypp.....IQVHVTLGne...............DLTVKMSDRGGGVpl.....................RKIDR
gyrb_ecoli  glHHMVFEVVDNAIDEalaghc............KEIIVTIHa...............DNSVSVQDDGRGIptgihpeegvsaa.....EVIMT
hs82_yeast  nkEIFLRELISNASDAldkirykslsdpkqletePDLFIRITpkpe............QKVLEIRDSGIGMtkaelinnlgtia.....KSGTK
mutl_ecoli  rpASVVKELVENSLDAga...............TRIDIDIErgg..............AKLIRIRDNGCGIkkdel.....................ALALA
tp2b_human  glYKIFDEILVNAADNkqrdknm...........TCIKVSIDpe...............SNIISIWNNGKGIpvvehkvekvyvpalifgq....LLTSS
tp6b_sulsh  alYQTVRELIENSLDAtdvhgil...........PNIKITIDliddarq..........IYKVNVVDNGIGIpp.....................QEVPN
                 #  ## * #                       # # #                    # # * * *#

                hh                          hhhhhhhhhh    ssssss                                          sssssssss
chea_ecoli  IFAPGFStaeqvtdvsgr.......GVGMDVVKRNIQkmgg.HVEIQSKqgt.....................................GTTIRILLP
envz_ecoli  LFQPFVRgdsartisgt........GLGLAIVQRIVDnhng.MLELGTSerg.....................................GLSIRAWLP
ntrb_ecoli  LFYPMVSgreggt............GLGLSIARNLIDqhsg.KIEFTSWpg.....................................HTEFSVYLP
phor_ecoli  LTERFYRvdkarsrqtggs......GLGLAIVKHAVNhhes.RLNIESTvgk.....................................GTRFSFVIP
arcb_ecoli  IFAMYYQvkdshggkpat.......GIGLAVSRRLAKnmgg.DITVTSEqgk.....................................GSTFTLTI.
etr1_arath  IFTKFAQtqslatrssggs......GLGLAISKRFVNlmeg.NIWIESDglgk.....................................GCTAIFD..
phy1_tobac  MFGTEAEasee..............GISLLISRKLVKlmng.EVQYLREagr.....................................STFII....
narx_ecoli  ........................HYGMIIMRDRAQslrg.DCRVRRResg.....................................GTEVVVTF.
sp22_bacsu  LFTTKPElers..............GMGFTIMENFMD.....DVSIDSSpem.....................................GTTIRLTKH
pdk1_human  LFNYMYStaprprvetsravplagfGYGLPISRLYAQyfqg.DLKLYSLegy.....................................GTDAVIYIK
gyrb_ecoli  VLHAGGKfddnsykvsgglh.....GVGVSVVNALSQ.....KLELVIQ....REGKIHRQIYehgvpqa.PLAVTGETekt.......GTMVRFW.P
hs82_yeast  AFMEALSagadvsmigqf.......GVGFYSLFLVAD.....RVQVISKs...NDDEQYIWESnagg....SFTVTLDEvnerigr...GTILRLFLK
mutl_ecoli  RHATSKIasldd1eaii.........SLGFRGEALASIssvs.RLTLTSRta..EQQEAWQAYAegr.....DMNVTVKPaahpvgttleVLDLFYNTP
tp2b_human  NYDDDEKkvtggrn...........GYGAKLCNIFST.....KFTVETAck..EYKHSFKQTWmnnmmktsEAKIKHFDged.......YTCITFQ.P
tp6b_sulsh  AFGRVLYsskyvnrqtrgmy.....GLGVKAAVLYSQmhqdkPIEIETSpvnsKRIYTFKLKIdinkne..PIIVERGSventrgfh..GTSVAISIP
             *                        *#*#              # #                                                 # #
```

B

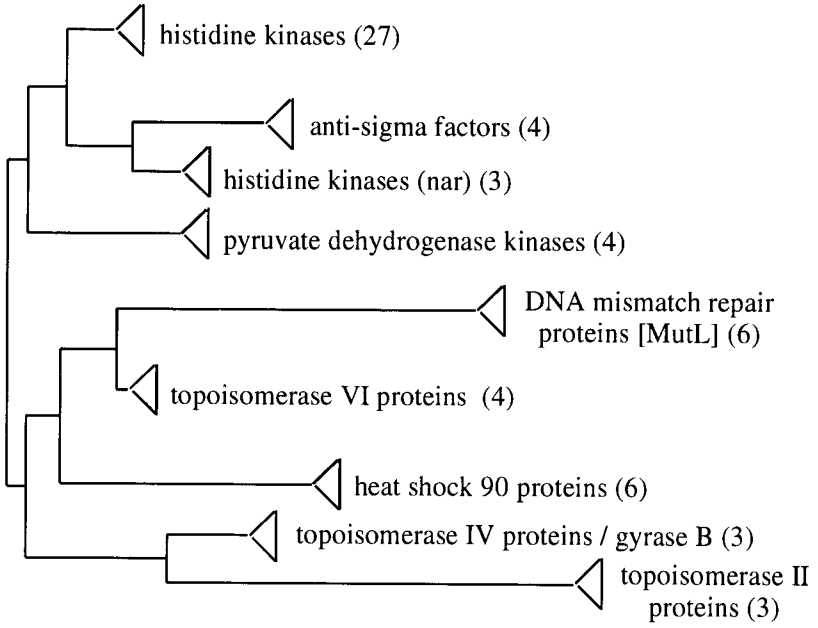


FIG. 4.—*A*, A representative set of sequences showing the multiple-sequence alignment of the kinase/nucleotide-binding domain. The highly conserved residues that define this domain are indicated below by "*" while conserved hydrophobicity patterns are indicated by "#." The conserved secondary-structure units of CheA, gyraseB, and heat shock 90 protein are labeled "h" for helix and "s" for beta strand. The residues in capital letters were the ones used for the phylogenetic analysis. *B*, Neighbor-joining phylogenetic tree of the kinase/nucleotide-binding domains. Methodology and labeling follow those of figure 2. The support for the division between kinase- and nucleotide-binding domains remains below 50%. Number of sequences used for each protein family are in parentheses.

TCST systems evolve by global duplication of all their components and subsequent differentiation. This model is supported by the fact that many TCST systems are concurrent on the chromosome. Phylogenetically, this mechanism would produce congruent gene trees for histidine kinases and response regulators.

Our results support the coevolution model. Despite the large number of proteins considered, which limited

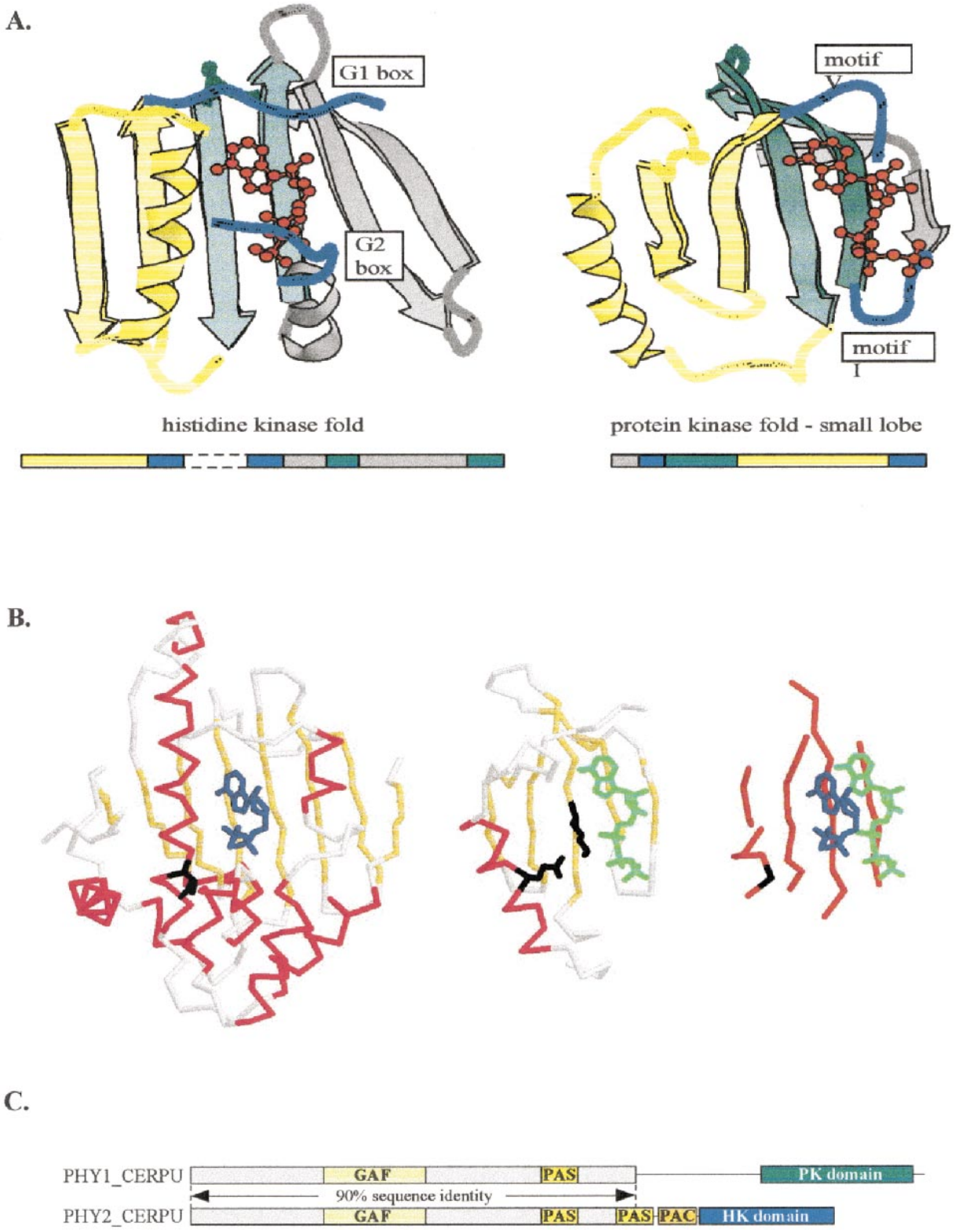


FIG. 5.—*A,* Comparison of the conserved structural cores in the folds of the histidine kinase superfamily (left) and the small lobe of protein kinases (right). The regions shown in yellow and green are structurally equivalent but circularly permuted. The bound nucleotide is shown in red, and two glycine-containing loops involved in nucleotide binding are shown in blue. Secondary-structure elements shown in gray are peculiar to the individual folds, although they may occupy topologically equivalent positions. *B,* Differences in nucleotide binding between the Hsp90 ATPase domain (left) and the small lobe of protein kinase C (center). Structurally equivalent Cα positions, as computed by STAMP (Russell and Barton 1992), are shown on the right. A glutamate residue that is highly conserved in both folds is shown in black. In protein kinases, this residue (in motif III; Hanks, Quinn, and Hunter 1988) interacts with an equally conserved lysine residue (in motif II). *C,* Comparison of phytochromes 1 and 2 of the moss *Ceratodon purpureus.* Domains are labeled as identified by sequence searches against the SMART database (Schultz et al. 2000). The two proteins are 90% identical in their amino-terminal 750 residues, but they signal through histidine and protein kinase domains, respectively.

the resolution of the analysis by lowering the ratio of aligned residues to OTUs, the trees obtained for the histidine kinase and response regulator domains showed congruent clustering (fig. 2). Although the precise evolutionary relationships between clusters were not supported by strong bootstrap values, the overall topology of the NJ histidine kinase tree was verified by heuristic search results for the minimal-length MP tree. No such support was obtained for the response regulator tree, which had a lower resolution, but its validity was verified by the occurrence of two superclusters (Arf/Cit/CheY/Lyt and Nar/Mth) that were also found in the histidine kinase tree. The clusters themselves were statistically much more robust, with over half receiving bootstrap support of >50% in the distance-based analysis. As required by the coevolution model, pairs of histidine kinases and response regulators that are known to interact were overwhelmingly found in cognate clusters, as were 89% of histidine kinase and response regulator pairs that are linked on the chromosome (fig. 3). Coevolution has previously been proposed for eukaryotic TCST proteins, as well as for two hybrid kinases of *E. coli* (BarA and RcsC) (Pao and Saier 1997). Hybrid kinases, however, also provide evidence for a recruitment mechanism at work. For example, four of the five hybrid kinases of *E. coli* (ArcB, TorS, RcsC, and EvgA) are known to signal through response regulators found in noncognate clusters, and the fifth one, BarA, is thought to do so as well (via OmpR, found in the Pho cluster) (Hoch and Silhavy 1995). Recruitment is also observed in the chemotaxis and sporulation systems. Nevertheless, coevolution appears to be the strongly predominant mechanism by which novel TCST systems arise. Similar patterns of molecular coevolution have been observed in other interacting proteins, such as neuropeptides and their receptors (Darlinson and Richter 1999) and chaperonin subunits (Archibald, Logsdon, and Doolittle 1999).

Coevolution is not limited to TCST proteins, but also extends to the domains forming them. Both domain shuffling and domain swapping are comparatively rare events, and, with few exceptions, response regulators having homologous carboxyl-terminal domains were found within one cluster. These results agree well with a previous study performed on 49 bacterial response regulators by Pao and Saier (1995), who found that classes of response regulators, defined by homology of their carboxyl-termini, generally formed distinct phylogenetic clusters. Pao and Saier's (1995) classes 1–5 correspond—in order—to our phylogenetic clusters Ntr, Pho, Nar, CheB, and Hybrid; classes 6 and 7 were phylogenetically heterogeneous in both studies.

Archaeal and Eukaryotic TCST Systems Evolved Through Separate Horizontal Transfers of Bacterial Genes

The conventional view of the universal tree is that Archaea and eukaryotes are sister groups rooted in the bacteria, and that all three urkingdoms of life are separate, monophyletic groups (Woese, Kandler, and Wheelis 1990; Brown and Doolittle 1997). However, the phylogenies depicted here clearly deviate from this canonical view in several fundamental respects. First, Archaea and eukaryotes are not sister groups. Eukaryotic histidine kinases and response regulators cluster with bacterial TCST proteins in a monophyletic group (Hybrid) that is not closely related to any of the archaeal clusters. Second, the Archaea do not form a monophyletic group. Most of their TCST proteins form species-specific (either *A. fulgidus* [Arf] or *M. thermoautotrophicum* [Mth]) clusters that are separate and most closely related to bacterial clusters (Arf to Cit and Mth to Nar). Third, although TCST proteins do not occur universally in any of the urkingdoms, their representation is much more limited in Archaea and eukaryotes. Despite the large number of sequenced bacterial genomes, only Mycoplasmas have so far been found to lack TCST systems, and these are obligate pathogens known to have greatly reduced their gene complement. Among the Archaea, however, the eight sequenced genomes have already uncovered four that lack TCST proteins entirely (table 1) and two that contain only a single system (Che). Among eukaryotes, TCST proteins are limited to fungi, slime molds, and plants. Thus far, no representatives have been found in animals or protists, which contain only two histidine kinase homologs, pyruvate dehydrogenase kinase and branched-chain alpha-ketoacid dehydrogenase kinase, which are not TCST proteins and form an outgroup to the histidine kinase tree (figs. 2 and 4*B*).

These observations suggest that TCST systems originated in bacteria after their separation from the last common ancestor and radiated into Archaea and eukaryotes through multiple horizontal gene transfer events. The basic forms of two-component signaling, as defined by the effector domain structure, presumably arose early in bacterial evolution, hence the widespread representation of diverse bacterial species in most of the major phylogenetic clusters. However, significant diversification of TCST systems occurred throughout the subsequent bacterial speciation, resulting in multiple species-specific subclusters within the major clusters, as well as in the striking Synechocystis-specific cluster Syn, which contains 12 histidine kinases and 21 response regulators from this organism.

As suggested above, TCST proteins radiated into the Archaea and Eucarya (eukaryotes) after these groups emerged as separate urkingdoms and were well into their speciation phases. This scenario of multiple horizontal gene transfers is the most parsimonious explanation for contemporary TCST gene distributions. The alternative view, that the last common ancestor already contained the basic TCST forms, which were selectively lost in most archaeal and eukaryotic branches, would require a large number of independent gene loss events occurring nearly concurrently with rapid gene evolution in the lineages retaining TCST genes.

The results presented here contribute to a growing body of evolutionary studies which suggest that horizontal gene transfer, rather than being a rare and isolated event, is a major motive force in organismal evolution (Golding and Gupta 1995; Doolittle 1999). They also

show that novel functional requirements may arise during the evolution of a species, which remain unfilled by endogenous genes, allowing acquired genes to establish themselves and rapidly diversify.

Histidine Kinases and Eukaryotic Protein Kinases—Homology or Analogy?

A distance-based phylogenetic analysis of proteins containing a histidine kinase–like ATP-binding domain, which include type II topoisomerases, Hsp90, and MutL, indicated that all kinase domains with this fold are monophyletic and confirmed that the PDKs form an outgroup to the histidine kinase clade (fig. 4*B*). Searches for more distantly related proteins surprisingly suggested a similarity to the small lobe of protein kinases, which is also involved in ATP binding. The structurally similar region covers virtually the entire conserved core of both folds but is circularly permuted in the protein kinase small lobe (fig. 5*A*). The amino- and carboxyl-termini of the histidine kinase fold are in close proximity, however (a prerequisite of circular permutation), and circular permutation events have been documented in the evolution of many protein folds (see the SCOP database at http://scop.mrc-lmb.cam.ac.uk/scop/; Lo Conte et al. 2000), including the histidine kinase fold (Bilwes et al. 1999). Although the protein kinase small lobe is part of the so-called ATP-grasp fold (jointly with the peptide-binding large lobe), a recent structural analysis by Grishin (1999) has concluded that only the large lobe is homologous among the members of this fold, with the small lobe having been recruited among structurally similar but unrelated proteins. Our analysis supports an independent evolutionary origin of the small and large lobes of eukaryotic protein kinases.

Despite the fact that the structural similarity between histidine kinases and protein kinases is remote, the signaling pathways in which the two types of kinase operate present intriguing similarities. Both form homodimers, which phosphorylate in *trans* and generally contain an extracellular sensory domain, which binds signaling molecules asymmetrically at the subunit interface. They have a common mode of signal transduction, as shown by the phytochromes Phy1 and Phy2 of the moss *Ceratodon purpureus*, which are 90% identical in the chromophore-binding domain, yet signal through protein kinase and histidine kinase domains, respectively (fig. 5*C*). Functional chimeras have also been constructed between the sensory domain of the Tar chemoreceptor and both protein and histidine kinases (Moe, Bollag, and Koshland 1989; Utsumi et al. 1989). Finally, as discovered recently, both kinase types use adaptors with an SH3-fold for interaction with other proteins (Bilwes et al. 1999). Similar parallels can be drawn between response regulators and the Ras family of G-proteins (Lukat et al. 1991). Not only are both encoded by large multigene families, linking different sensory inputs to specific effector outputs, but they are both activated by a high-energy phoshoanhydride bond. Their striking structural similarity, particularly in the active site, has previously been interpreted as evidence for homology rather than analogy (Artymiuk et al. 1990). Although each of these similarities may have arisen by convergent evolution, the combination of structural and functional parallels that can be drawn throughout signaling pathways in bacteria and eukaryotes suggest that a prototypical signal transduction pathway may already have existed in the last common ancestor and that this pathway utilized protein phosphorylation. If so, yet a third group of kinases (possibly showing similarly profound structural changes) remains to be discovered in Archaea, where bacterial- and eukaryotic-type kinases are rare and clearly acquired by horizontal transfer.

Acknowledgments

We thank Drs. John Throup and Martin Burhnam for useful discussions. This work was supported by Defense Advanced Research Projects Agency Grant N65236-97-1-5810 to Dr. M. Rosenberg. The content of the information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

LITERATURE CITED

ALEX, L. A., K. A. BORKOVICH, and M. I. SIMON. 1996. Hyphal development in *Neurospora crassa*: involvement of a two-component histidine kinase. Proc. Natl. Acad. Sci. USA **93**:3416–3421.

ALEX, L. A., C. KORCH, C. P. SELITRENNIKOFF, and M. I. SIMON. 1998. COS1, a two-component histidine kinase that is involved in hyphal development in the opportunistic pathogen *Candida albicans*. Proc. Natl. Acad. Sci. USA **12**: 7069–7073.

ALEX, L. A., and M. I. SIMON. 1994. Protein histidine kinases and signal transduction in prokaryotes and eukaryotes. Trends Genet. **10**:133–138.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHÄFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

ARAVIND, L., and C. P. PONTING. 1999. The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. FEMS Microbiol. Lett. **176**:111–116.

ARCHIBALD, J. M., J. M. LOGSDON, and W. F. DOOLITTLE. 1999. Recurrent paralogy in the evolution of archaeal chaperonins. Curr. Biol. **9**:1053–1056.

ARTYMIUK, P. J., D. W. RICE, E. M. MITCHELL, and P. WILLETT. 1990. Structural resemblance between the families of bacterial signal-transduction proteins and of G proteins revealed by graph-theoretical techniques. Protein Eng. **4**:39–43.

BAIKALOV, I., I. SCHRODER, M. KACZOR-GRZESKOWIAK, K. GRZESKOWIAK, R. P. GUNSALUS, and R. E. DICKERSON. 1996. Structure of the *Escherichia coli* response regulator NarL. Biochemistry **35**:11053–11061.

BARRETT, J. F., R. M. GOLDSCHMIDT, L. E. LAWRENCE et al. (22 co-authors). 1998. Antibacterial agents that inhibit two-component signal transduction systems. Proc. Natl. Acad. Sci. USA **95**:5317–5322.

BILWES, A. M., L. A. ALEX, B. R. CRANE, and M. I. SIMON. 1999. Structure of CheA, a signal-transducing histidine kinase. Cell **96**:131–141.

BOTT, M., M. MEYER, and P. DIMROTH. 1995. Regulation of anaerobic citrate metabolism in *Klebsiella pneumoniae*. Mol. Microbiol. **18**:533–546.

BROWN, J. R., and W. F. DOOLITTLE. 1997. Archaea and the prokaryote-to eukaryotes transition. Microbiol. Mol. Biol. Rev. **61**:456–502.

CHANG, C., S. F. KWOK, A. B. BLEECKER, and E. M. MEYEROWITZ. 1993. *Arabidopsis* ethylene-response gene ETR1: similarity of product to two-component regulators. Science **262**:539–544.

DANDEKAR, T., B. SNEL, M. HUYNEN, and P. BORK. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. **23**:324–328.

DARLINSON, M. G., and D. RICHTER. 1999. The 'chicken and egg' problem of co-evolution of peptides and their cognate receptors: which came first? Results Probl. Cell Differ. **26**: 1–11.

DARWIN, A. J., E. C. ZIEGELHOFFER, P. J. KILEY, and V. STEWART. 1998. Fnr, NarP, and NarL regulation of *Escherichia coli* K-12 napF (periplasmic nitrate reductase) operon transcription in vitro. J. Bacteriol. **180**:4192–4198.

DAYHOFF, M. O., R. V. ECK, and C. M. PARK. 1972. A model of evolutionary change in proteins. Pp. 89–99 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Vol. 5. National Biomedical Research Foundation, Washington, D.C.

DOOLITTLE, W. F. 1999. Lateral genomics. Trends Cell Biol. **9**:M5–M8.

DUTTA, R., L. QIN, and M. INOUYE. 1999. Histidine kinases: diversity of domain organization. Mol. Microbiol. **34**:633–640.

ENRIGHT, A. J., I. ILIOPOULOS, N. C. KYRPIDES, and C. A. OUZOUNIS. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature **402**:86–90.

FELSENSTEIN, J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author (http://evolution.genetics.washington.edu/phylip.html), Department of Genetics, University of Washington, Seattle.

GEORGELLIS, D., O. KWON, P. DE WULF, and E. C. LIN. 1998. Signal decay through a reverse phosphorelay in the Arc two-component signal transduction system. J. Biol. Chem. **273**:32864–32869.

GOLBIK, R., A. N. LUPAS, K. K. KORETKE, W. BAUMEISTER, and J. PETERS. 1999. The Janus face of the archaeal Cdc48/p97 homolog VAT: protein folding versus unfolding. Biol. Chem. **380**:1049–1062.

GOLDING, G. B., and R. S. GUPTA. 1995. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. Evol. **12**:1–6.

GRISHIN, N. V. 1999. Phosphatidylinositol phosphate kinase: a link between protein kinase and glutathione synthase folds. J. Mol. Biol. **291**:239–247.

HANKS, S. K., A. M. QUINN, and T. HUNTER. 1988. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science **241**:42–52.

HENIKOFF, S., and J. G. HENIKOFF. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**:10915–10919.

HOCH, J., and T. SILHAVY. 1995. Two-component signal transduction. ASM Press, Washington, D.C.

JENKINS, L. S., and W. D. NUNN. 1987. Regulation of the ato operon by the atoC gene in *Escherichia coli.* J. Bacteriol. **169**:2096–2102.

JUNG, K., B. TJADEN, and K. ALTENDORF. 1997. Purification, reconstitution, and characterization of KdpD, the turgor sensor of *Escherichia coli.* J. Biol. Chem. **272**:10847–10852.

KATO, M., T. MIZUNO, T. SHIMIZU, and T. HAKOSHIMA. 1997. Insights into multistep phosphorelay from the crystal structure of the C-terminal HPt domain of ArcB. Cell **88**:717–723.

KORETKE, K. K., R. RUSSELL, R. COPLEY, and A. N. LUPAS. 1999. Fold recognition using sequence and secondary structure information. Proteins **37**(S3):141–148.

KREMS, B, C. CHARIZANIS, and K. D. ENTIAN. 1996. The response regulator-like protein Pos9/Skn7 of *Saccharomyces cerevisiae* is involved in oxidative stress resistance. Curr. Genet. **29**:327–334.

LANGE, R., C. WAGNER, A. DE SAIZIEU, N. FLINT, J. MOLNOS, M. STIEGER, P. CASPERS, M. KAMBER, W. KECK, and K. E. AMREIN. 1999. Domain organization and molecular characterization of 13 two-component systems identified by genome sequencing of *Streptococcus pneumoniae*. Gene **237**: 223–234.

LAZAREVIC, V., P. MARGOT, B. SOLDO, and D. KARAMATA. 1992. Sequencing and analysis of the *Bacillus subtilis lytRABC* divergon: a regulatory unit encompassing the structural genes of the N-acetylmuramoyl-L-alanine amidase and its modifier. J. Gen. Microbiol. **138**:1949–1961.

LO CONTE, L., B. AILEY, T. J. HUBBARD, S. E. BRENNER, A. G. MURZIN, and C. CHOTHIA. 2000. SCOP: a structural classification of proteins database. Nucleic Acids Res. **28**: 257–259.

LOOMIS, W. F., G. SHAULSKY, and N. WANG. 1997. Histidine kinases in signal transduction pathways of eukaryotes. J. Cell Sci. **110**:1141–1145.

LUKAT, G. S., B. H. LEE, J. M. MOTTONEN, A. M. STOCK, and J. B. STOCK. 1991. Roles of the highly conserved aspartate and lysine residues in the response regulator of bacterial chemotaxis. J. Biol. Chem. **266**:8348–8354.

MAEDA, T., S. M. WURGLER-MURPHY, and H. SAITO. 1994. A two-component system that regulates an osmosensing MAP kinase cascade in yeast. Nature **369**:242–245.

MARCOTTE, E. M., M. PELLEGRINI, H. NG, D. W. RICE, T. O. YEATES, and D. EISENBERG. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science **285**:751–753.

MARTINEZ-HACKERT, E., and A. M. STOCK. 1997. The DNA-binding domain of OmpR: crystal structures of a winged helix transcription factor. Structure **5**:109–124.

MOE, G. R., G. E. BOLLAG, and D. E. KOSHLAND JR. 1989. Transmembrane signalling by a chimera of the *Escherichia coli* aspartate receptor and the human insulin receptor. Proc. Natl. Acad. Sci. USA **86**:5683–5687.

NAKANO, M. M., T. HOFFMANN, Y. ZHU, and D. JAHN. 1998. Nitrogen and oxygen regulation of *Bacillus subtilis* nasDEF encoding NADH-dependent nitrite reductase by TnrA and ResDE. J. Bacteriol. **180**:5344–5350.

OVERBEEK, R., M. FONSTEIN, M. D'SOUZA, G. D. PUSCH, and N. MALTSEV. 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA **96**:2896–2901.

PAO, G. M., and M. H. SAIER JR. 1995. Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. J. Mol. Evol. **40**:136–154.

———. 1997. Nonplastid eukaryotic response regulators have a monophyletic origin and evolved from their bacterial precursors in parallel with their cognate sensor kinases. J. Mol. Evol. **44**:605–613.

PARK, H., and M. INOUYE. 1997. Mutational analysis of the linker region of EnvZ, an osmosensor in *Escherichia coli.* J. Bacteriol. **179**:4382–4390.

PARKINSON, J. S., and E. C. KOFOID. 1992. Communication modules in bacterial signaling proteins. Annu. Rev. Genet. **26**:71–112.

PLOWMAN, G. D., S. SUDARSANAM, J. BINGHAM, D. WHYTE, and T. HUNTER. 1999. The protein kinases of *Caenorhabditis elegans.* A model for signal transduction in multicellular organisms. Proc. Natl. Acad. Sci. USA **96**:13603–13610.

POPOV, K. M., N. Y. KEDISHVILI, Y. ZHAO, Y. SHIMOMURA, D. W. CRABB, and R. A. HARRIS. 1993. Molecular cloning of the p45 subunit of pyruvate dehydrogenase kinase. J. Biol. Chem. **268**:26602–26606.

POSAS, F., S. M. WURGLER-MURPHY, T. MAEDA, E. A. WITTEN, T. C. THAI, and H. SAITO. 1996. Yeast HOG1 MAP kinase cascade is regulated by a multistep phosphorelay mechanism in the SLN1-YPD1-SSK1 ''two-component'' osmosensor. Cell **86**:865–875.

RUSSELL, R. B., and G. J. BARTON. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins **14**:309–323.

SCHNEIDER-POETSCH, H. A., B. BRAUN, S. MARX, and A. SCHAUMBURG. 1991. Phytochromes and bacterial sensor proteins are related by structural and functional homologies. Hypotheses on phytochrome-mediated signal-transduction. FEBS Lett. **281**:245–249.

SCHULER, G. D., S. F. ALTSCHUL, and D. J. LIPMAN. 1991. A workbench for multiple alignment construction and analysis. Proteins **9**:180–190.

SCHULTZ, J., R. R. COPLEY, T. DOERKS, C. P. PONTING, and P. BORK. 2000. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. **28**:231–234.

SCHURR, M. J., H. YU, J. M. MARTINEZ-SALAZAR, J. C. BOUCHER, and V. DERETIC. 1996. Control of AlgU, a member of the sigma E-like family of stress sigma factors, by the negative regulators MucA and MucB and *Pseudomonas aeruginosa* conversion to mucoidy in cystic fibrosis. J. Bacteriol. **178**:4997–5004.

SCHUSTER, S. S., A. A. NOEGEL, F. OEHME, G. GERISCH, and M. I. SIMON. 1996. The hybrid histidine kinase DokA is part of the osmotic response system of *Dictyostelium.* EMBO J. **15**:3880–3889.

SIMON, G., V. MEJEAN, C. JOURLIN, M. CHIPPAUX, and M. C. PASCAL. 1994. The torR gene of *Escherichia coli* encodes a response regulator protein involved in the expression of the trimethylamine N-oxide reductase genes. J. Bacteriol. **176**:5601–5606.

SINGLETON, C. K., M. J. ZINDA, B. MYKYTKA, and P. YANG. 1998. The histidine kinase dhkC regulates the choice between migrating slugs and terminal differentiation in *Dictyostelium discoideum.* Dev. Biol. **203**:345–357.

STOCK, A. M., E. MARTINEZ-HACKERT, B. F. RASMUSSEN, A. H. WEST, J. B. STOCK, D. RINGE, and G. A. PETSKO. 1993. Structure of the Mg(2+)-bound form of CheY and mechanism of phosphoryl transfer in bacterial chemotaxis. Biochemistry **32**:13375–13380.

STOCK, J. B., A. J. NINFA, and A. M. STOCK. 1989. Protein phosphorylation and regulation of adaptive responses in bacteria. Microbiol. Rev. **53**:450–490.

STOKER, K., W. N. REIJNDERS, L. F. OLTMANN, and A. H. STOUTHAMER. 1989. Initial cloning and sequencing of hydHG, an operon homologous to ntrBC and regulating the labile hydrogenase activity in *Escherichia coli* K-12. J. Bacteriol. **171**:4448–4456.

STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

SWOFFORD, D. L. 1999. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer, Sunderland, Mass.

TANAKA, T., S. K. SAHA, C. TOMOMORI et al. (13 co-authors). 1998. NMR structure of the histidine kinase domain of the *E. coli* osmosensor EnvZ. Nature **396**:88–92.

THELEN, J. J., M. G. MUSZYNSKI, J. A. MIERNYK, and D. D. RANDALL. 1998. Molecular analysis of two pyruvate dehydrogenase kinases from maize. J. Biol. Chem. **273**:26618–26623.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

THROUP, J. P., K. K. KORETKE, A. P. BRYANT et al. (12 co-authors). 2000. A genomic analysis of two-component signal transduction in *Streptococcus pneumoniae.* Mol. Microbiol. **35**:566–576.

UTSUMI, R., R. E. BRISSETTE, A. RAMPERSAUD, S. A. FORST, K. OOSAWA, and M. INOUYE. 1989. Activation of bacterial porin expression by a chimeric signal transducer in response to aspartate. Science **245**:1246–1249.

VARUGHESE, K. I., MADHUSUDAN, X. Z. ZHOU, J. M. WHITELEY, and J. A. HOCH. 1998. Formation of a novel four-helix bundle and molecular recognition sites by dimerization of a response regulator phosphotransferase. Mol. Cell **2**:485–493.

VOLKMAN, B. F., M. J. NOHAILE, N. K. AMY, S. KUSTU, and D. E. WEMMER. 1995. Three-dimensional solution structure of the N-terminal receiver domain of NTRC. Biochemistry **34**:1413–1424.

WEISS, D. S., J. BATUT, K. E. KLOSE, J. KEENER, and S. KUSTU. 1991. The phosphorylated form of the enhancer-binding protein NtrC has an ATPase activity that is essential for activation of transcription. Cell **67**:155–167.

WOESE, C. R., O. KANDLER, and M. L. WHEELIS. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. Proc. Natl. Acad. Sci. USA **87**:4576–4579.

WOMBLE, D. D. 2000. GCG: the Wisconsin package of sequence analysis programs. Methods Mol. Biol. **132**:3–22.

XU, Q., and A. WEST. 1999. Conservation of structure and function among histidine-containing phosphotransfer (Hpt) domains as revealed by crystal structure YPD1. J. Mol. Biol. **292**:1039–1050.

YEH, K. C., and J. C. LAGARIAS. 1998. Eukaryotic phytochromes: light-regulated serine/threonine protein kinases with histidine kinase ancestry. Proc. Natl. Acad. Sci. USA **95**:13976–13981.

ZHULIN, I. B., B. L. TAYLOR, and R. DIXON. 1997. PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox. Trends Biochem. Sci. **9**:331–333.

ANTONY DEAN, reviewing editor

Accepted September 6, 2000