Archaea 2, 11–30

Distribution, structure and diversity of "bacterial" genes encoding two-component proteins in the Euryarchaeota

MARK K. ASHBY[1,2,3]

[1] *Department of Basic Medical Sciences, Biochemistry Section, University of the West Indies, Mona Campus, Kingston 7, Jamaica*

[2] *Current address: School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, U.K.*

[3] *Corresponding author (m.ashby@qmul.ac.uk)*

Received June 13, 2005; accepted January12, 2006; published online January 23, 2006

Summary The publicly available annotated archaeal genome sequences (23 complete and three partial annotations, October 2005) were searched for the presence of potential two-component open reading frames (ORFs) using gene category lists and BLASTP. A total of 489 potential two-component genes were identified from the gene category lists and BLASTP. Two-component genes were found in 14 of the 21 Euryarchaeal sequences (October 2005) and in neither the Crenarchaeota nor the Nanoarchaeota. A total of 20 predicted protein domains were identified in the putative two-component ORFs that, in addition to the histidine kinase and receiver domains, also includes sensor and signalling domains. The detailed structure of these putative proteins is shown, as is the distribution of each class of two-component genes in each species. Potential members of orthologous groups have been identified, as have any potential operons containing two or more two-component genes. The number of two-component genes in those Euryarchaeal species which have them seems to be linked more to lifestyle and habitat than to genome complexity, with most examples being found in *Methanospirillum hungatei, Haloarcula marismortui, Methanococcoides burtonii* and the mesophilic Methanosarcinales group. The large numbers of two-component genes in these species may reflect a greater requirement for internal regulation. Phylogenetic analysis of orthologous groups of five different protein classes, three probably involved in regulating taxis, suggests that most of these ORFs have been inherited vertically from an ancestral Euryarchaeal species and point to a limited number of key horizontal gene transfer events.

Keywords: histidine kinase, hybrid kinase, response regulator.

Introduction

Two-component systems are one of the key means by which bacteria respond to environmental changes (Hoch 2000, Stock et al. 2000, Alves and Savageau 2003, Hellingwerf 2005). They are assumed to be of bacterial origin, having radiated into archaea and some eukaryotes by horizontal gene transfer (HGT) (Koretke et al. 2000). Two-component systems consist of a sensor and a response protein. The sensor protein is characterized by a histidine kinase (HK) made up of two main domains, a phosphoacceptor (HisKA) and a histidine kinase ATPase (HATPase) and, in many cases, other sensory domains are present (Galperin et al. 2001, Zhulin et al. 2003). The response protein (response-regulator, RR) is characterized by a response regulator domain that has a conserved aspartate residue. The histidine kinase autophosphorylates a conserved histidine residue in response to a signal and the phosphate group is then transferred to the conserved aspartate residue of the response-regulator. The transfer of the phosphate group to the response-regulator elicits a response causing a change in taxis, development or gene expression. Histidine kinases and response-regulators are sometimes found together in a single polypeptide known as a hybrid kinase (Hoch 2000, Stock et al. 2000).

The recognition of the Archaea as a distinct division of life has been strengthened by the availability of a number of complete genome sequences, representing three phyla (Euryarchaota, Crenarchaeota and Nanoarchaeota). This has, in turn, enabled a more rigorous phylogenetic analysis based on the fusion of ribosomal protein sequences (Matte-Tailliez et al. 2002, Brochier et al. 2004, Bapteste et al. 2005, Makarova and Koonin 2005) and clusters of conserved orthologous genes (COGs) (Makarova and Koonin 2003). Analysis of genome sequences has revealed genes of bacterial origin in the genomes of archaea and vice versa (Nelson et al. 1999). The importance of HGT in the evolution of prokaryotes and the implications for phylogeny and definition of species is still being discussed (Ochman et al. 2000, Forterre et al. 2002, Boucher et al. 2003, Koonin 2003, Kurland et al. 2003, Lawrence and Hendrickson 2003).

For bacteria, it has been shown that the number of two-component genes possessed by an organism is related to the complexity of its genome, its physiology and the changeability of its habitat (Ashby 2004, Galperin 2005). The greater the value of any of those parameters, the greater the need for regulation of cellular activities.

The aim of this study was to analyze the complement of genes in archaeal genomes that could encode two-component proteins. The putative two-component proteins were classified by their domain structure, and the number of each class was

determined for each species. Potential orthologous groups and those that may be part of operons, with two or more two-component genes, are indicated. Phylogenetic analysis of possible orthologous groups representing five classes of protein, three associated with taxis, is shown.

Materials and methods

Genome sequence data

The list of publicly available Euryarchaeal genome sequences is shown in Table 1, along with brief details of the habitat, physiology, genome size, putative number of open reading frames (ORFs) and the abbreviation used with gene sequences (Makarova and Koonin 2003). The sequences and annotations for the annotated sequences (up to October 2005) were accessed at the Integrated Microbial Genomes (IMG) server (http://img.jgi.doe.gov/pub/main.cgi) and HaloLex (http://www.halolex.mpg.de/). The identity of potential two component genes was determined by reference to the published assignments located at http://www. tigr.org/tdb/ (Bult et al. 1996, Smith et al. 1997, Kawarabayasi et al. 1998, Klenk et al. 1997, Ng et al. 2000, Deppenmeier et al. 2002, Slesarev et al. 2002,

Table 1. Strain description of members of the Euryarchaea for which there are publicly available genome sequences as of October 2005. The gene name prefix is used with the gene designations to aid in identification of the species.

Strain	Genus	Gene name prefix	Temp opt. (°C)	Physiology	Genome size (MB)	No. protein-coding genes
Archaeoglobus fulgidus DSM 4304	Archaeoglobales	Aful	83	Anaerobic, sulphate-reducing chemolitho- or chemorgano-autotroph	2.18	2,456
Ferroplasma acidarmanus Incomplete	Thermoplasmales		40	Acidophile	1.97	1740
Haloarcula marismortui	Halobacteriales	Hma	37	Aerobic halophilic chemorganotroph	4.3	4,242
Halobacterium sp. NRC-1	Halobacteriales	Halo	37	Aerobic halophilic chemorganotroph	2.57	2,656
Methanocaldococcus jannaschii DSM 2661	Methanococcales		85	Chemolithoautotroph, anaerobe, methanogen	1.66	1,758
Methanococcus maripaludis S2	Methanococcales	Mmar	37	Mesophilic anaerobe, hydrogenotrophic methanogen	1.67	1,742
Methanopyrus kandleri AV19	Methanopyrales		110	Chemolithoautotroph. anaerobe, methanogen	1.69	1,691
Methanococcoides burtonii DSM 6242 Incomplete	Methanosarcinales	Mbur	~0	Anaerobe, psychrotolerant methanogen	~2.6	~2872
Methanosarcina acetivorans C2A	Methanosarcinales	Mace	37	Chemoorganoheterotroph, anaerobe, N_2-fixing, methanogen	5.75	4,540
Methanosarcina barkeri str. fusaro	Methanosarcinales	Mbar	37	Chemoorganoheterotroph, anaerobe, methanogen	4.87	3,380
Methanosarcina mazei Go1	Methanosarcinales	Mmaz	37	Chemoorganoheterotroph, anaerobe, N_2-fixing, methanogen	4.1	3,371
Methanospirillum hungatei JF-1	Methanomicrobiales	Mhun	37	Chemoorganoheterotroph anaerobe, methanogen	3.53	3,356
Methanothermobacter thermautotrophicus str. Delta H	Methanobacteriales	Mthe	65	Chemolithoautotroph, anaerobe, N_2-fixing methanogen	1.75	1,914
Natronomonas pharaonis	Halobacteriales	Npha	20	Haloalkaliphile	2.75	2,843
Picrophilus torridus DSM 9790	Thermoplasmales		50	Extreme acidophile	1.58	1,582
Pyrococcus abyssi GE5	Thermococcales	PAB	96	Anaerobic heterotroph	1.76	1,769
Pyrococcus furiosus DSM 3638	Thermococcales		96	Anaerobic heterotroph	1.91	2,115
Pyrococcus horikoshii OT3	Thermococcales	PH	98	Anaerobic heterotroph	1.74	2,110
Thermococcus kodakaraensis	Thermococcales	Thermoco	102	Anaerobic heterotroph	2.09	2,358
Thermoplasma acidophilum DSM 1728	Thermoplasmales		59	Fac. anaerobe, chemorganotroph, thermoacidophile	1.56	1,482
Thermoplasma volcanium GSS1	Thermoplasmales		60	Fac. anaerobe, chemorganotroph, thermoacidophile	1.55	1,499

Galagan et al. 2002, Cohen et al. 2003, Baliga et al. 2004, Falb et al. 2005). This was supplemented by BLASTP (Altschul et al. 1997) searches of each genome with a battery of two component domains (domains used include receivers from CheY and OmpR, HisKA/HATPase and Hpt; see Table A1) from *Methanosarcina acetivorans* and *E. coli* K12 at IMG (http://img.jgi.doe.gov/pub/main.cgi), The Integrated Genome Resource (TIGR, http://www.tigr.org/tdb/) or the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/).

Bioinformatic analysis

Putative two-component domains were initially assigned in ORFs by Pfam batch analysis (http://www.sanger.ac.uk/Software/Pfam/; Bateman et al. 2004). Domains were recorded for each two-component gene only if they were scored as "Pfam's trusted match thresholds." Domain assignments were checked and modified using the more extensive domain assignments at InterPro (http://www.ebi.ac.uk/interpro/). The results were used to classify the putative two-component proteins by domain organization using a nomenclature adapted from Ohmori et al. (2001), with each group subdivided by the organization of the identified signalling domains. The cartoon style diagrams that present the domain organization of the deduced sequences were constructed from these data, with the sizes of the domains roughly in proportion to each other. For clarity, each gene name begins with a four character acronym (except for *Haloarcula marismortui*, *Pyrococcus abyssi* GE5 and *Pyrococcus horikoshii* OT3, see Table 1) followed by either the locus tag that can be found at IMG or HaloLex (http://img.jgi.doe.gov/pub/main.cgi and http://www.halolex.mpg.de/) or the gene object identifier if the sequencing or annotation is incomplete.

To determine orthologous groups, orthology information, based on the bidirectional best hits from BLASTPs of each organism against each other organism polypeptide, is accessible at IMG (http://img.jgi.doe.gov/pub/main.cgi). This definition is not completely accurate, but it provides a useful approximation as it is not always possible to know whether the polypeptides arose from a single gene present in the last common ancestor (orthologues) or from a gene duplication within a genome (paralogues). Alignments for phylogenetic analysis were performed by TCoffee (Notredame et al. 2000) and accessed at the Centre Nationale de la Recherche Scientifique website (http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi) and ClustalW alignments (Thompson et al. 1994) were performed at the European Bioinformatics Institute (http://www.ebi.ac.uk/clustalw/). Representatives from three bacterial phyla were included in the alignments (chosen by having the best match to one of the archaeal ORFs, either as an orthologue or by BLASTP at IMG). Phylogenetic analysis by neighbor-joining (Bootstrap 250) was performed using MEGA version 3.0 (Kumar et al. 2004) and by maximum-likelihood (Felsenstein 1996) using Molphy, accessed at the Institut Pasteur, biological software website (http://bioweb.pasteur.fr/intro-uk.html#phylo).

Closely linked two-component genes and probable operons that contain two or more two-component genes were constructed from the chromosome map images at IMG (http://img.jgi.doe.gov/pub/main.cgi), TIGR (http://www.tigr.org/tdb/) and the biology of extremophiles website (http://www-archbac.u-psud.fr/homepage.html)

Results and discussion

The structural classification of potential two-component proteins is shown in Table 2. No two-component encoding gene could be found in the Crenarchaeota or Nanoarchaeota (data not shown). Sensor domains are drawn as ellipses and two-component (HisKA, HATPase_c and response regulator) and output domains are drawn as rectangles. Parentheses followed by figures indicate the number of similar domains that may be found in the proteins listed in each subclass.

Histidine kinases

Different types of histidine kinases (HK) are listed in Table 2A. Histidine kinases contain two domains; a dimerization and a phosphoacceptor domain (HisKA or HisKA_2) and a HATPase_c domain (Grebe and Stock 1999). HisKA and HisKA_2 are part of a His kinase A phosphoacceptor domain superfamily that also includes HWE_HK and HisKA_3 (Karniol and Vierstra 2004; Pfam accession CL0025).

HKI Histidine kinase Is are HKs containing HisKA and HATPase domains. There may be other domains in some of these examples which are not currently recognized. The HKI ORFs vary greatly in size, ranging from 175 to 592 amino acids in length.

HKII Histidine kinase IIs are HKs containing sensor GAF and PAS/PAC domains. The GAF domains (c<u>G</u>MP phosphodiesterase, <u>a</u>denylyl cyclases, bacterial transcription factors <u>F</u>hlA) are associated with small molecule binding, in particular cAMP and cGMP (Aravind and Ponting 1997, Ho et al. 2000, Anantharaman et al. 2001). The GAF domain is usually found in combination with PAS (*Drosophila* <u>p</u>eriod clock protein, vertebrate <u>a</u>ryl hydrocarbon receptor nuclear translocator and *Drosophila* <u>s</u>ingle-minded protein) or PAC (<u>P</u>AS-<u>a</u>ssociated <u>C</u>-terminal motif) domain, or both. One class of PAS domains is known to bind cofactors such as heme and FAD (Bibikov et al 2000, Sardiwal et al. 2005). Sensing of light, oxygen or redox potential by PAS domains requires cofactors, whereas sensing signals such as voltage, xenobiotics and nitrogen availability does not (Ponting and Aravind 1997, Gilles-Gonzalez and Gonzalez 2004). The PAC domains are proposed to contribute to the PAS domain fold. The shared feature of GAF and PAS/PAC domains is the binding of a diverse set of regulatory small molecules that often remain unidentified; all three domains are common signal transduction system components (Anantharaman and Aravind 2001, Zhulin et al. 2003). There is one example containing Cache and one containing SBP_bac_3 (bacterial extracellular solute-binding proteins,

family 3). The Cache domain is a signalling domain found in animal calcium channel subunits and it is thought to form an extracellular or periplasmic ligand sensor (Anantharaman and Aravind 2001). SBP_bac_3 is involved in active transport of solutes across the cytoplasic membrane and in the initiation of signal transduction pathways (Tam and Saier 1994). This is by far the largest subgroup of ORFs, containing 161 out of the total of 489 (33% of the total).

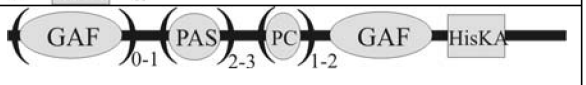
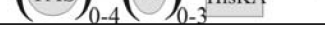
Table 2. Compilation and cartoon diagrams of all putative archaeal open reading frames encoding two-component proteins. Abbreviations; Cach = Cache; ConA = ConA-like glucanase; Gly = Glycos_transf_2; HKA = His_KA (and KA2); HATP = HATPase_c; Hkd = H-kinase_dim; H10 = HTH_10; PC = PAC; and SBPBac3 = SBP_bac_3. See Table 1 for gene name prefixes.

A. Histidine kinases

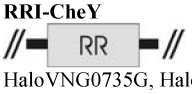
HKI.  HaloVNG2180.3, AfulAF1467, AfulAF0208, MtheMTH459, MtheMTH1124, MtheMTH123, MtheMTH444, MtheMTH356, Mbur_401648670, Mbur_401649920, Mbur_401657500, Mbur_401643290, MaceMA3962, MbarA_0679, MmazMM0948, HmarrnAC0477, HmarrnAC1693, HmarrnAC2945, HmarrnAC2997, HmarrnAC3370, HmarrnAC3440, Mhun_401789380, Mhun_401789570 (HisKA_2), Mhun_401795600
HKII
HKII + GAF 
HKII + GAF HmarrnAC0247, HmarrnAC0964, NphaNP0730A
HKII + GAF1PASPAC. GP - HaloVNG0736G, MaceMA1878, MaceMA0970, MmazMM2886, MbarA_2594, HmarrnAC0413, HmarrnB0295. PG - HaloVNG0716G (No PAC), HaloVNG2037.3, Mbur_401647970 (No PAS), Mbur_401650390 (No PAS), Mbur_401653070, Mbur_401638250.
HKII + GAF2PASPAC GP - Mbur_401656610 (No PAC), MaceMA2890, MaceMA2294, MaceMA2266 (PAC_1), MbarA_3538, MbarA_0815 (PAS_1) PG - HaloVNG1175G (No PAC), HmarrnAC0410 PGP - HaloVNG0916G $(PAC)_1$, Mbur_401642670 (no PAC), NphaNP1882A (PAC_1)
HKII + GAF3PASPAC GP - MaceMA1628, MmazMM0518, NphaNP0090A (PAC_1) PGP - AfulAF1483 $(PAC)_2$, GPGP - NphaNP1912A
HKII + GAF4PASPAC GP - MaceMA3370 PGP - MbarA_2935, MmazMM3295, NphaNP1622A (PAC_3)
HKII + mPASPAC - PGP MaceMA1270 $(PACPAS)_6$, MaceMA1630 (PAC_5PAC_2), MaceMA0551 (PAS_6PAC_5), MaceMA1646 (PAS_5PAC_4), MaceMA0203 $(PASPAC)_5$, MbarA_3036 (PAS_5PAC_4), MbarA_1944 $(PASPAC)_5$ GP - MbarA_3447 $(PASPAC)_5$
HKII + mGAFPASPAC $(PAS)_2$-G-PASPAC-G-$(PASPAC)_4$ - MmazMM0168 G-$(PASPAC)_4$-G-$(PAS)_7(PAC)_8$ - HmarrnB0180 PASPAC-G-$(PASPC)_4$-G - MbarA_3037 $(PAS)_2$PAC-G-$(PAS)_7(PAC)_6$-$(GAF)_2$ -HmarrnB0299 
HKII + PASPAC 
HKII + 1 PASPAC HaloVNG1234C, AfulAF1515 (No PAC), AfulAF0450, AfulAF1639, AfulAF0021 (No PAC), MtheMTH1260 (NoPAC), MtheMTH360, MtheMTH292 (NoPAC), Mbur_401644730, Mbur_401643030, Mbur_401657970, Mbur_401641290, Mbur_401650600 (NoPAS), MaceMA2757, MaceMA1957, MaceMA0619, MaceMA1704, MaceMA0620 (NoPAS), MaceMA3481, MbarA_1535, MbarA_2362, MbarA_1687 (No PAC), MbarA_1668 (No PAC), MmazMM1781, MmazMM1915, MmazMM1777 (No PAS), MmazMM2990, HmarrnAC0789 (no PAC), HmarrnAC1114 (No PAC), HmarrnAC1391, HmarrnAC1486 (No PAC), HmarrnAC2219 (No PAS), HmarrnAC2461, HmarrnAC2477, HmarrnAC2789 (No PAC), HmarrnAC3347, HmarrnAC3349 (No PAC), HmarrnAC3482, HmarrnB0133(No PAC), NphaNP1506A, NphaNP1640A (No PAC), NphaNP1804A, NphaNP1850A (No PAC), NphaNP3536A (No PAC), NphaNP3752A.
HKII + 2 PASPAC AfulAF0770, AfulAF1184, AfulAF1452, MtheMTH619 (No PAC), MtheMTH823, MtheMTH468 (No PAC), Mbur_401638920 $(PAS)_1$, Mbur_401653820, MaceMA4026, MaceMA2294, MaceMA2082, MaceMA2732, MbarA_2184, MbarA_2907, MbarA_1483 (1PAS), HmarrnAC0991, HmarrnAC1096, NphaNP0144A, NphaNP1154A, NphaNP1914A, Mhun_401786790 (1PAS)
HKII + 3 PASPAC HaloVNG1374G $(PAS)_2$, AF2109, MtheMTH174 (1PAC), Mbur_401643080 $(PAC)_2$, Mbur_401655300, MaceMA1627 (2PAC), MaceMA0758 (2PAS), MaceMA3346, MaceMA1149, MaceMA0490, MaceMA3543, MbarA_2691 (1PAC), MmazMM1671, MmazMM2748, MmazMM2178 (2PAC), MmazMM2646 (2PAS), HmarrnAC0086, HmarrnAC1501 (2PAS), NphaNP0862A, NphaNP6080A $(PAC)_2$.

continued on facing page

Table 2. Cont'd. Compilation of all putative archaeal open reading frames encoding two-component proteins.

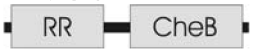
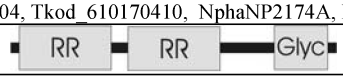
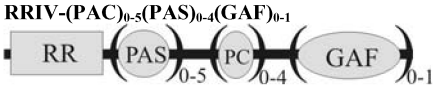
HKII + 4 PASPAC AfulAF0410 $(PAC)_3$, MaceMA1991 (3PAS), MaceMA1645, MbarA_1666, MbarA_2680, MbarA_2876 (PAS_3), MbarA_2552, MmazMM2276, MmazMM2773, MmazMM1093, MmazMM0169 (no PAS), MmazMM2435 (3PAC), MmazMM2275 (3PAS).	
HKII + 5 PASPAC MaceMA2784 $(PAS)_4$, MaceMA0759 $(PAS)_4$, MaceMA0552, MbarA_0404 $(PAS)_3$, MmazMM2515, MmazMM2277, MbarA_3250, NphaNP4606A $(PAC)_4$.	
HKII + mPASPAC MaceMA1844 (7PAC, 8PAS), MaceMA1274 (6PAS, 7PAC), MaceMA1322 (6PAS, 7PAC), MaceMA1470 (5PAC, 6PAS), MaceMA3368 $(PACPAS)_7$,	
HKII + Cache Mbur_401646300	
HKII + SBP Mbur_401644740	
HKIII	
HKIII MmarMMP1303, MmarMMP1120, HmarrnAC0988, NphaNP0142A	
HKIII + PASPAC ConA-like glucanase Mbur_401641830	
HKIII + GAFPASPAC HmarrnB0156	
HKIII + CHASE4 AfulAF1721, MaceMA2553, MmazMM3099.	
HKIII + CHASE4PASPAC MaceMA1739 (No PAC), MaceMA2555, MaceMA2348 (No PAC), MmazMM2629, MmazMM3101.	
HKIII + Cache Mbur_401655290, Mbur_401649130 , MmazMM2955 (PASPAC), Mhun_401789550	
HKVI	
HKVI  **No P2 -** MmazMM1325., MaceMA0014, HmarrnAC2205, MbarA_0984, Mhun_401776240, Mhun_401784470 **P2** HaloVNG0971G, AfulAF1040, MaceMA3066, MmazMM0328, Mhun_401793120, NphaNP2172A **$(P2)_2$ -** Mbur_401647520, MmarMMP0927, PAB1332, PH0484, Tkod_610170420/30	
HATPase_c	
HATPase_c Mbur_401643110, Mbur_401662170, Mbur_401641680, Mbur_401642700, Mbur_401657760, MbarA_1241, HmarrnAC0457, HmarrnAC2337.	
HATPase_c + PASPAC AfulAF0893, HmarrnAC0130, NphaNP2516A $(PASPAC)_2$., NphaNP6064A $(PAS)_3(PAC)_2$.	
HATPase_c + $(GAF)_2$,$(PASPAC)_n$ MaceMA4561 $(PAS)_2(PAC)_3$, MaceMA0863 $PAS(PAC)_2$	
HATPase_c + PAS MaceMA0777, MmazMM1931, HmarrnAC1346, HmarrnAC2616 $(PAS)_2$, HmarrnAC2694, NphaNP3356A, Mhun_401806820. **HATPase_c + HAMP** HaloVNG1375.3	
His_KA	
His_KA Mhun_401787170	
His_KA $(PAC)_{1-2}(PAS)_{2-3}(GAF)_{1-2}$ AfulAF0277 $(PAC)_2(PAS)_3$-GAF, AfulAF2420 332 $(PASPAC)_2$GAF. AfulAF0448 $GAF(PAS)_2PACGAF$.	
His_KA + $(PAS)_{0-4}(PAC)_{0-3}$. AfulAF1035 $(PAS)_2$. AfulAF2032 $(PAS)_4(PAC)_3$.	

B. Response regulators

RRI-CheY  HaloVNG0735G, HaloVNG2036G, HaloVNG0974G, HaloVNG0917G (213aa), AfulAF1063, AfulAF1898, AfulAF0449, AfulAF2249, AfulAF1256, AfulAF2419, AfulAF1384, AfulAF1473, AfulAF1042, MtheMTH549, MtheMTH445, MtheMTH447 (277aa), MmarMMP1304, MmarMMP0933, Mbur_401656780 (361aa), Mbur_401641730, Mbur_401640890, Mbur_401660760,

continued overleaf

Table 2. Cont'd. Compilation of all putative archaeal open reading frames encoding two-component proteins.

Mbur_401647540, Mbur_401657490, Mbur_401644620, Mbur_401658280 (287aa), Mbur_401641840, MaceMA4671, MaceMA1268, MaceMA1269, MaceMA2861, MaceMA3068, MaceMA2445 (292aa), MaceMA1468, MaceMA1469, MaceMA0016, MaceMA0018, MaceMA2012, MaceMA4376, MaceMA1366, MbarA_3448, MbarA_2388, MbarA_3321, MbarA_2896, MbarA_0988, MbarA_0986, MbarA_3109 (260aa), MbarA_1051, MbarA_3248, MbarA_2510, MmazMM0330, MmazMM0049, MmazMM3007 (292aa), MmazMM3206, MmazMM1068, MmazMM2351, MmazMM2516, MmazMM2880, MmazMM2953, MmazMM2954, MmazMM1327, MmazMM1328, PAB1330, PH0482, HmarpNG4019, HmarrnAC0361, HmarrnAC0411, HmarrnAC0536 (201aa), HmarrnAC1118, HmarrnAC1494, HmarrnAC2168, HmarrnAC2194, HmarrnAC3308, HmarrnB0297, Tkod_610170400, NphaNP0028A, NphaNP0140A, NphaNP0516A, NphaNP0682A, NphaNP2102A, NphaNP2906A, Mhun_401793100, Mhun_401778650, Mhun_401785170, Mhun_401776210 (284aa), Mhun_401789540, Mhun_401778900, Mhun_401780830, Mhun_401781980, Mhun_401783330, Mhun_401783500, Mhun_401783520, Mhun_401785150, Mhun_401786800, Mhun_401789100, Mhun_401789600, Mhun_401791380, Mhun_401792540, Mhun_401793270, Mhun_401793820 (220aa), Mhun_401794450, Mhun_401795290 (216aa), Mhun_401795560, Mhun_401796600, Mhun_401799230 (214aa), Mhun_401799550, Mhun_401799610 (211aa), Mhun_401799660 (208), Mhun_401799870, Mhun_401800570 (352aa), Mhun_401801850, Mhun_401801900, Mhun_401801910, Mhun_401803600, Mhun_401803780, Mhun_401804010, Mhun_401806750 (238aa).	
RRIII-(PACPAS)GAFHTH_10 HmapNG7159 – $(PASPAC)_2(GAF)_2$, HmapNG7223 – $(GAF)_3$	
RRIII-DUF24 HmarrnB0301	
RRIV-CheB  HaloVNG0973G, AfulAF1041, MmarMMP0926, Mbur_401647530, MaceMA3067, MaceMA0015, MbarA_0985, MmazMM0329, MmazMM1326, PAB1331, PH0483, HmarrnAC2204, Tkod_610170410, NphaNP2174A, Mhun_401793110, Mhun_401776230	
RRIV-Glycos_transf_2 MtheMTH548	
RRIV-$(PAC)_{0-5}(PAS)_{0-4}(GAF)_{0-1}$  MtheMTH1607, HmarrnAC3271, NphaNP1846A, NphaNP2716A –RR-GAF MtheMTH1764 – RR-PAC-GAF HmarrnAC2211 – RR-GAF-PAS Mhun_401786010 – RR-GAF-PAS_2 MtheMTH440, HmarrnAC0339, HmarrnAC0674, HmarrnAC2109, Mhun_401793830, Mhun_401773960, Mhun_401775740, Mhun_401792240 – RR-PAS Mhun_401789530, Mhun_401795570, Mhun_401803130 - RR-PASPAC MtheMTH457 – RR-PAS-PAC-$(PAS)_2$ HmarrnAC0356 – RR-$(PASPAC)_2$-GAF HmarrnAC1142 – RR-PAC-PAS-PAC Mhun_401776490 – RR-$(PAS)_3(PAC)_2$ Mhun_401796390 RR-$(PAS)_5(PAC)_4$.	

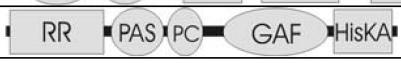
C. Hybrid kinases

HY1  MtheMTH901, HmarrnAC0301, HmarrnAC0412, HmarrnAC0475, HmarrnB0296, NphaNP3458A, Mhun_401776220, Mhun_401784480, Mhun_401792550, Mhun_401801890, Mhun_401799920 (No HisKA)	
HYI+PASPAC 	
HY1+PASPAC MtheMTH902 (No PAS), Mbur_401644630, Mbur_401651890, Mbur_401641180, HmarrnAC0487 (No PAC), NphaNP0138A (No PAS), Mhun_401789580 (No HATPase), Mhun_401796780, Mhun_401800550 (HisKA_2), Mhun_401806740 (HisKA_2)	
HYI+$(PASPAC)_2$ MtheMTH446 (No PAC), MaceMA1267, Mbur_401642690, HmarrnAC0574, HmarrnAC0848 $(PAC)_1$, HmarrnAC1806 $(PAS)_1$, NphaNP0188A, NphaNP0622A, Mhun_401803140 (PAC_1, No HATPase), Mhun_401795550 (No HATPase), Mhun_401785160, Mhun_401795880, Mhun_401803320 (No PAC, HATPase), Mhun_401803800 (No PAC), Mhun_401803980	
HYI + mPASPAC HaloVNG5037G – $(PAS)_5(PAC)_3$, MbarA_3247 - $(PAS)_2(PAC)_3$,. MmazMM2518, HmarrnAC2044, HmarrnAC3050, Mhun_401801140, Mhun_401805200 - $(PAS)_3(PAC)_2$, HmarrnAC0075, NphaNP6028A, NphaNP6202A, Mhun_401774520, Mhun_401777310, Mhun_401801130, Mhun_401801880 (HisKA_2) – $(PACPAS)_3$, Mhun_401777350, - $(PAS)_4(PAC)_2$, HmarrnAC1626, NphaNP2742A – $(PASPAC)_4$., Mhun_401774510 $(PASPAC)_5$, Mbur_401644900 $(PAS)_5(PAC)_4$ (No HATPase), Mhun_401795210 $(PASPAC)_6$, Mhun_401774210 (6PAS, 4PAC), NphaNP2656A (7PAC, 8PAS), Mhun_401790080 $(PAS)_4$,	
HYI + GAF HmarrnAC0794	

continued on facing page

Table 2. Cont'd. Compilation of all putative archaeal open reading frames encoding two-component proteins.

<table>
<tr><td colspan="2">HYI + PASPACGAF
 

HmarrnAC2372, HmarrnAC3379, NphaNP5274A - PAS-GAF
NphaNP5120A - PASPAC-GAF
HmapNG7155, HmarrnAC2533, HmarrnAC2692 - GAF-PAS-PAC
HmapNG7156 - $(PAS)_2$-PAC-GAF-$(PACPAS)_2$
NphaNP0626A - PASPAC-GAF-$(PACPAS)_2$-GAF
HmarrnAC2416 - PASPAC-GAF-$(PAS)_2(PAC)_3$.
HmarrnAC3361, Mhun_401786010 - PASPAC-GAF-$(PASPAC)_3$.
HmarrnAC1495 - $(PAS)_2$-PAC-GAF-PASPAC-GAF-$(PASPAC)_2$
HaloVNG2334 - $(PAS)_2$-PAC-GAF-$(PAC)_3(PAS)_2$-GAF
Mhun_401793840 - GAF-$(PASPAC)_3$
Mhun_401775750 - GAF-PASPAC-HisKA-$(PAS)_4(PAC)_2$
Mhun_401779810 - $(PAS)_3(PAC)_2$-HisKA-$(PAS)_2$PAC-GAF-$(PAS)_4(PAC)_3$
NphaNP4696A - RR-PASPAC-HisKA-$(PAS)_4(PAC)_3$</td></tr>
<tr><td>HYII+Hpt
MaceMA2013</td><td> 
</td></tr>
<tr><td>HYIII
$PAC(PAS)_2$ - MaceMA2256.
$(PACPAS)_2$ - MmazMM3205, MmazMM2881</td><td> 
</td></tr>
<tr><td>HYIII+CHASE4HAMP
MaceMA4377, MbarA_1052</td><td> 
</td></tr>
<tr><td>HYIIHisKA
AfulAF1472</td><td> 
</td></tr>
</table>

HKIII Histidine kinase IIIs are HKs that possess a HAMP "linker" (histidine kinase, adenylyl cyclase, methyl-accepting chemotaxis protein and phosphatase) domain. The HAMP domain is usually associated with the transmission of a signal across a membrane from periplasmic ligand-binding domains (Aravind and Ponting 1999, Appleman and Stewart 2003, Zhu and Inouye 2004). Eight examples of HKIIIs have an N-terminal putative periplasmic signalling CHASE4 domain and four have an N-terminal periplasmic signalling Cache domain (Anantharaman and Aravind 2001, Zhulin et al. 2003). These domains are positioned next to the HAMP domain, presumably for efficient transfer of the signal.

HKVI Histidine kinase IVs are the CheA-like chemotaxis signalling proteins that contain an N-terminal Hpt (histidine phosphotransfer) and Hkd (histidine kinase dimerization) domain and a C-terminal CheW domain. Some contain one or two P2 domains between the Hpt and Hkd. The Hpt domain is involved in mediating phosphotransfer from one receiver domain to another (Hoch 2000). Hkd (H-kinase-dim) is the dimerization domain of CheA and CheW that interacts with methyl-accepting chemotaxis proteins (MCPs), relaying signals to CheY, and thereby affecting flagellar rotation (West et al. 1995). The P2 domain is involved in enhancing the interaction of CheY with the HK (Jahreis et al. 2004, Stewart and van Bruggen 2004). *Thermococcus kodakaraensis* has two open reading frames with a frame shift mutation that probably encodes for a CheA-like protein. All of the HKVI genes discussed are located close to other genes that could be involved with signal transduction and are probably transcribed as single operons (see Table A2).

HATPase_c These contain no dimerization or phosphoacceptor domains currently recognized at INTERPRO.

His_KA There are five groupings that contain His_KA without a discernable HATPase_c domain.

Response regulators

Response regulators (RR) are listed in Table 2B. These contain a characteristic receiver (RR/T_reg) domain, which is about 120 amino acids long and contains a conserved aspartate residue about halfway along the molecule that accepts a phosphate group from an HK.

RR I Response regulator Is are simple orphan (no other domain detected) RRs, representing the second largest group of two-component ORFs (24% of the total).

RRIII Response regulator IIs contain an RR fused to a potential DNA binding domain. Such regulators are found only in *H. marismortui.* Of these, there are only three examples that contain either the HTH_10 or DUF24 domain (PF04967 and PF 01638). These are the only RRs that are possibly transcriptional regulators, but there may be other currently unidentified DNA-binding domains in other RRs or hybrid kinases.

RRIV Response regulator IVs contain an N-terminal RR fused to output or signal domains. There are 16 examples of CheB fused to the RR. The CheB domain is related to methylesterase and is likely to be concerned with chemotaxis (West et al. 1995). There is one example of two RRs fused to a glycosyl transferase domain in *M. thermoautotrophicus* (Pfam Accession number: PF00353). The glycosyl transferase domain is involved in transferring sugar moieties from a donor to recipient molecules. There are a lot of *Methanospirilum hungatei* ORFs fused with PAS/PAC or GAF domains, or both, however, as the

annotation is incomplete, some of these ORFs may turn out to be part of hybrid kinases.

Hybrid kinases

Hybrid kinases (HY) are shown in Table 2C. They are defined as containing both HK and RR domains. The nomenclature is based on the position and number of RR with respect to the HK. There is an incomplete HYI in *A. fulgidus* that has a PAC/PAS and GAF sensor domain, but no discernable HATPase domain.

HYI Hybrid kinase Is have a single RR N-terminal to the HK.

HYII Hybrid kinase IIs have a single RR C-terminal to the HK. There is only one example in *M. acetivorans.*

HYIII Hybrid kinase IIIs have two RRs either N or C-terminal to the HK.

Distribution of putative two-component ORFs

The total number of ORFs within each class of two-component proteins, for each species of Euryarchaeota, is shown in Table 3. No two-component ORFs were found in the four Crenarchaeota species or *N. equitans* (data not shown). No two-component ORF was found in *M. jannaschii*, *M. kandleri*, *P. furiosus* or in any of the members of the Thermoplasmales. The three other *Pyrococcus* species each have only three two-component ORFs (*Thermococcus kodakaraensis* HKVI that has a frame shift Tkod_61070420/30, that could be a sequencing error has been counted as one), representing 0.17% of the protein-coding capacity of the genome. *Methanococcus maripaludis* and *Halobacterium* also have a small number, six (0.34%) and 16 (0.64%), respectively. *Archaeoglobus fulgidus*, *M. thermoautotrophicus* and the four Methanosarcinales groups have a comparatively large number of two-component ORFs, from 23 to 67. This represents from 1.03 to 1.48% of the coding capacity of the four complete genome assignments. *Haloarcula marismortui* has the largest number of two-component encoding genes, of the complete annotations, which represents 1.93% of the total protein coding capacity of the genome. *Methanospirillum hungatei* appears to have the largest number of two-component genes at 87, though the annotation of the genome is incomplete (so no percentage is given in Table 3). The HKs form half to two-thirds of the two-component ORFs for each species (except *Pyrococcus* sp. and *Methanospirillum hungatei*). The DNA-binding domains (putative) were only detected as part of the RRs in *H. marismortui*.

The PAS/PAC and GAF sensory domains are found in 293 of the 489 putative proteins surveyed. These sensory domains are absent in the *Pyrococcus* sp. A total of 18 ORFs were found that contain the HAMP domain that would in most cases be involved in transferring signals from sensor domains detecting information outside the cell.

Orthologous groups

Potential orthologous groups are shown in Table 4. These results are based on the bidirectional best hits from BLASTPs at IMG. The identification of orthologous groups at IMG may not be correct in all cases as some groupings may include ORFs that are due to gene duplication, hence a paralogue (in a different organism) rather than an orthologue. It is, nevertheless, a useful tool for assigning putative orthologous groups when no functional information is available. The groups have been named with a three letter acronym for ease of reference (see Table 4). In *arr18/19*, RRIV-CheB, the grouping was modified from the information at IMG based on the phylogenetic analysis presented in Figure 1 (see below). There are many orthologous groups that contain two or three members, partic-

Table 3. Distribution of euryarchaeal two-component open reading frames. The total number of identified two-component genes are shown (Total 2-C) and are given as a percentage of the total protein coding capacity of each genome (% 2-C). Abbreviations: HK = histidine kinase; RR = response regulator; HY = hybrid kinase; and incomp = incomplete.

Species	HKI	HKII	HKIII	HKVI other	HK total	HK	RRI	RRIII	RRIV total	RR	HY	Total 2-C	% 2-C
Archaeoglobus fulgidus	2	10	1	1	6	20	9	0	1	10	1	31	1.26
Haloarcula marismortui	6	23	2	2	6	39	10	3	8	21	22	82	1.93
Halobacterium sp. NRC-1	1	7	0	1	1	10	4	0	1	5	2	17	0.64
M. thermoautotrophicus	5	7	0	0	0	12	3	0	5	8	3	23	1.2
Methanococcus maripaludis	0	0	2	1	0	3	2	0	1	3	0	6	0.34
Methanococcoides burtonii	4	17	3	1	5	30	9	0	1	10	5	45	incomp
Methanosarcina acetivorans	1	38	4	2	3	48	13	0	2	15	4	67	1.48
Methanosarcina barkeri	1	21	0	1	1	24	10	0	1	11	2	35	1.03
Methanosarcina mazei	1	20	4	2	1	28	12	0	2	14	3	45	1.33
Methanospirillum hungatei	3	1	1	3	1	9	36	0	12	48	30	87	incomp
Natronomonas pharaonis	0	17	1	1	3	22	6	0	3	9	11	42	incomp
Pyrococcus abyssi	0	0	0	1	0	1	1	0	1	2	0	3	0.17
Pyrococcus horikoshii	0	0	0	1	0	1	1	0	1	2	0	3	0.14
Thermococcus kodakarensis	0	0	0	1	0	0	1	0	1	2	0	3	0.13
												489	

Table 4. Potential orthologous two-component open reading frames.

Gene name	Classification	Orthologue number
MaceMA3962, MbarA_0679, MMAZMM0948	HKI	*ahk1*
AfulAF208	HKI	*ahk2*
Mbur_401643110	HATPase	
MaceMA1878, MbarA_3037	HKII	*ahk3*
MaceMA1628, MMAZMM0518	HKII	*ahk4*
MaceMA2890, MbarA_2935, MmazMM3295, MtheMTH174	HKII	*ahk5*
Mhun_401800550	HYI PASPAC	
MaceMA1646, MbarA_3036	HKII	*ahk6*
MaceMA0203, MbarA_1944	HKII	*ahk7*
MmazMM2515, MbarA_3447, MaceMA1470	HKII	*ahk8*
MaceMA0552, MmazMM0168	HKII	*ahk9*
MaceMA2294, Mbur_401657970	HKII	*ahk10*
MbarA_2362, MmazMM2990	HKII	*ahk11*
MaceMA0619, MMAZMM1777	HKII	*ahk12*
MaceMA1704, MmazMM2990	HKII	*ahk13*
MaceMA0620, MmazMM1781, MbarA_1535	HKII	*ahk14*
MaceMA1149, MmazMM2178	HKII	*ahk15*
MaceMA0490, MmazMM1671, MbarA_2876	HKII	*ahk16*
MaceMA1645, MbarA_2680, MmazMM2748, Mhun_401774840	HKII	*ahk17*
MbarA_1687, MmazMM1931, Mhun_401789380, HmarrnAC0789	HKII	*ahk18*
HaloVNG1234C, HmarrnAC3482	HKII	*ahk19*
AfulAF0770, MmazMM2518, MaceMA3405	HKII	*ahk20*
MaceMA4026, MmazMM0889	HKII	*ahk21*
MaceMA2294, Mbur_401657970, HaloVNG1374G, Mhun_401803800, AfulAF0450, NphaNP1154A	HKII	*ahk22*
HmarrnB0156	HKIII	
AfulAF2109, Mhun_401800710, HmarrnAC1626	HKII	*ahk23*
MaceMA3368, Mbur_401643080, MaceMM2773, MbarA_3250	HKII	*ahk24*
MaceMA1627, MmazMM0172, MbarA_0404	HKII	*ahk25*
MaceMA1991, MtheMTH468, MmazMM1915	HKII	*ahk26*
MbarA_1666, Mhun_401774840	HKII	*ahk27*
MmazMM2276, MaceMA1274	HKII	*ahk28*
MmazMM1093, MbarA_3538	HKII	*ahk29*
MaceMA0759,	HKII	*ahk30*
Mhun_401779810	HYI	
MaceMA1844, MtheMTH823	HKII	*ahk31*
HmarMMP1303,	HKII	*ahk32*
Mbur_401650390	HKIII	
(NphaNP1640A)	HKII	*ahk33*
HmarrnAC2616	HATPasePAS	
(NphaNP1804A)	HKII	*ahk34*
HmarrnAC0130	HATPasePAS	
HmarrnB0133, NphaNP3536A	HKII	*ahk35*
MaceMA2553, MmazMM3099	HKIII	*ahk36*
MaceMA2555, MMAZMM3101	HKIII	*ahk37*
AfulAF1721, Mhun_401794250	HKIII CHASE	*ahk38*
MaceMA1739, MmazMM2629	HKIII CHASE	*ahk39*
MaceMA3066, MmazMM0328, AfulAF1040, Mbur_401647520	HKVI	*ahk40*
MaceMA0014, MmazMM1325, MbarA_0984	HKVI	*ahk41*
PAB1332, PH0484, Tkod_610170420/30, MmarMMP0927	HKVI	*ahk42*
HaloVNG0971G, HmarrnAC2205, NphaNP2172A, Mhun_401793120	HKVI	*ahk43*
MaceMA0777, MmazMM1931	HATPase+	*ahk44*
Mhun_401789380	HKI	
HmarrnAC0789	HKII	
MaceMA0863,	HATPaseGAF	*ahk45*
Mbur_401650390	HKII	

continued overleaf

Table 4. Cont'd. Potential orthologous two-component open reading frames.

Gene name	Classification	Orthologue number
HaloVNG2036G, HmarrnAC1118 (NphaNP2906A 50%)	RRI	*arr1*
MmarMMP1304, Mbur_401640890, Mhun_401785170	RRI	*arr2*
MaceMA4671, MmazMM2880	RRI	*arr3*
MbarA_2510, MmazMM2953	RRI	*arr4*
MmazMM2954, Mbur_401641840	RRI	*arr5*
MtheMTH447, Mbur_401660760, Mhun_401806750	RRI	*arr6*
HaloVNG0974G, AfulAF1042, MmarMMP0933, Mbur_401647540, MaceMA3068, MmazMM0330, PAB1330, PH0482, HmarrnAC2194, Tkod_610170400 , Mhun_401793100, NphaNP2102A	RRI	*arr7*
Mbur_401660760, MaceMA1469, MbarA_3448, MMAZMM2516, MtheMTH1764 (PACGAF), Mhun_401793830 (PAS)	RRI	*arr8*
MaceMA1268, MbarA_3248, HmarrnAC0411, AfulAF1473, MtheMTH445, Mhun_401778650, HmarrnAC0536, (NphaNP0140A)	RRI	*arr9*
MaceMA0016, MbarA_0986, MmazMM1327	RRI	*arr10*
MaceMA2861, MbarA_2388, MmazMM0049, HmarrnAC0339 (PAS)	RRI	*arr11*
MaceMA4376, MbarA_1051, MmazMM1068, AfulAF2419, Mbur_401644620, Mhun_401789540	RRI	*arr12*
MaceMA1366, MbarA_2896, MMAZMM2351, Mbur_401657490, Mhun_401803780	RRI	*arr13*
MaceMA2445, MbarA_3109, MMAZMM3007, Mbur_401658280, Mhun_401776210	RRI > 200aa	*arr14*
HmarrnAC3308, Mhun_401796600	RRI > 200aa	*arr15*
MaceMA0018, MbarA_0988, MMAZMM1328	RRI	*arr16*
HmapNG7223,	RRIII-HTH	*arr17*
Mhun_401799920	HYIHATPase	
MaceMA0015, MbarA_0985, MmazMM1326	RR CheB	*arr18*
PAB1331, PH0483, Tkod_610170410, HaloVNG0973G, HmarrnAC2204, MaceMA3067, MmazMM0329, Mbur_401647530, AfulAF1041, MmarMMP0926, NphaNP2174A, Mhun_401793110	RR CheB	*arr19*
MtheMTH1607	RR IV	*arr20*
AfulAF0448	His_KAPACPASGAF	
MtheMTH440,	RRIVPAS	*arr21*
Mbur_401642690, Mhun_401789580	HYI PASPAC	
HmarrnAC1142	RRIVPAS	*arr22*
Mhun_401803140	HYI PASPAC	
Mhun_401784480, Mbur_401644630	HYI	*ahy1*
MaceMA1267, MbarA_3247, MtheMTH446, Mhun_401795550	HYI PASPAC	*ahy2*
AfulAF1472	HYIHisKA	
HmarrnAC3379,	HYI PASGAF	*ahy3*
HaloVNG1175G	HKIIPASGAF	
HmarrnAC2533	HYI PASGAF	*ahy4*
HaloVNG0916G, (NphaNP1912A)	HKIIPASGAF	
HmapNG7156, HaloVNG2334	HYI PASGAF	*ahy5*
MaceMA2256, MmazMM2881	HYIII	*ahy6*
MaceMA4377, MbarA_1052	HYIII	*ahy7*
HmarrnAC0475	HYI	*ahy8*
Mhun_401787170	HisKA	
Mhun_401801890,	HYI	*ahy9*
MtheMTH444	HKI	
HmarrnAC2044, Mhun_401801140	HYIPASPAC	*ahy10*
Mhun_401774520, HaloVNG5037G, Mbur_401644900	HYIPASPAC	*ahy11*
HmarrnAC3050, Mhun_401777310	HYIPASPAC	*ahy12*

ularly within the Methanosarcinales. The more interesting groups are those that have more members or that have members in different genera. These will be discussed below, particularly those that are part of taxis operons (*ahk40–43, arr7, arr10, arr18* and *arr19*).

Phylogenetic analysis

Figures 1 to 5 show neighbor-joining and the supplementary Figures A1 to A5 contain the maximum likelihood phylogenetic analyses of alignments from TCoffee analysis. Phylogenetic analysis of the same ORFs was also performed on align-

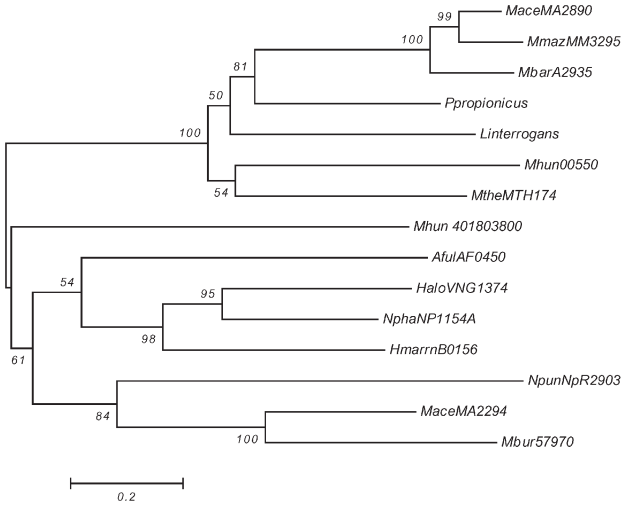


Figure 1. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from histidine kinase II groups *ahk5* and *ahk22*. Group *ahk5* = MaceMA2890, MbarA_2935, MmazMM3295, MtheMTH174 and Mhun_401800550; and group *ahk22* = MaceMA2294, Mbur_401657970, HaloVNG1374G, Mhun_401803800, AfulAF0450, NphaNP1154A and HmarrnB0156. The bacterial ORFs are Ppropionicus, NpunNpR2903 and Linterrogans. Abbreviations: Ppropionicus = *P. propionicus*_500413890; NpunNpR2903 = *N. punctiforme*; and Linterrogans = *L. interrogans*-LA2540

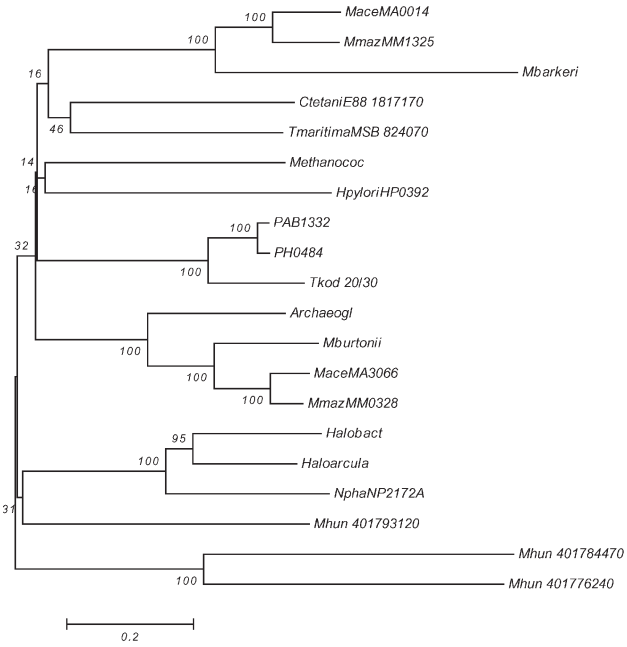


Figure 2. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from the histidine kinase IV groups *ahk40–43*. Group *ahk40* = MaceMA3066, MmazMM0328, AfulAF1040 and Mbur_401647520; group *ahk41* = MaceMA0014, MmazMM1325 and MbarA_0984; group *ahk42* = PAB1332, PH0484, Tkod_610170420/30 and MmarMMP0927; and group *ahk43* = HaloVNG0971G, HmarrnAC2205, NphaNP2172A and Mhun_401793120. The bacterial ORFs are CtetaniE17170, TmaritimaMSB_824070 and HpyloriHP0392. Abbreviations: CtetaniE17170 = *C. tetani*E88_1817170; TmaritimaMSB_824070 = *T. maritima*; and HpyloriHP0392 = *H. pylori.*

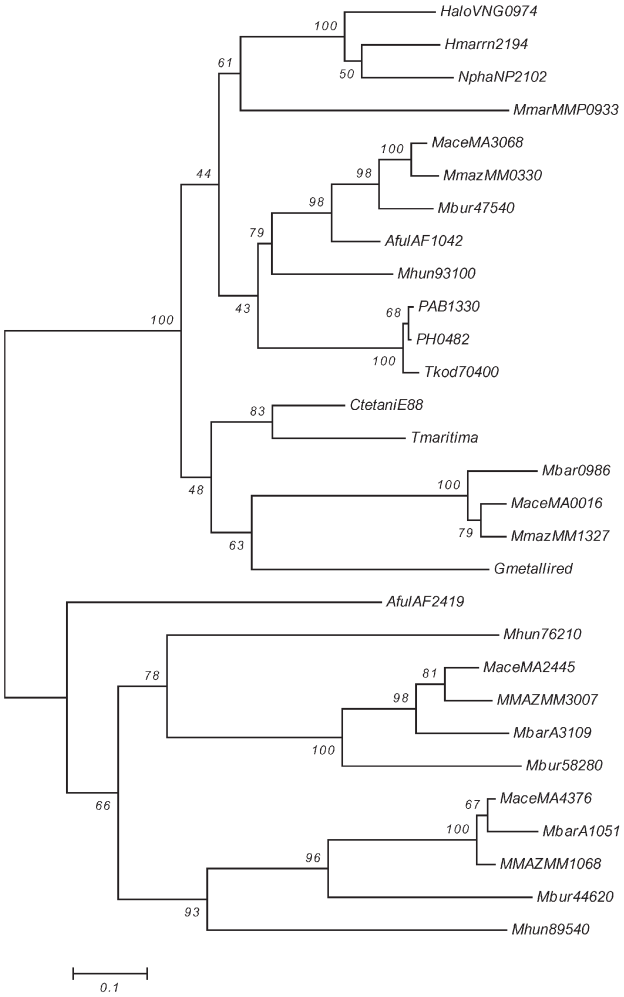


Figure 3. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from the response regulator I orphans *arr7, arr10, arr12* and *arr14*. Group *arr7* = HaloVNG-0974G, AfulAF1042, MmarMMP0933, Mbur_401647540, Mace-MA3068, MmazMM0330, PAB1330, PH0482, HmarrnAC2194, Tkod_610170400 , Mhun_401793100 and NphaNP2102A; group *arr10* = MaceMA0016, MbarA_0986 and MmazMM1327; group *arr12* = MaceMA4376, MbarA_1051, MmazMM1068, AfulAF2419, Mbur_401644620 and Mhun_401789540; and group *arr14* = Mace-MA2445, MbarA_3109, MMAZMM3007, Mbur_401658280 and Mhun_401776210. The bacterial ORFs are CtetaniE88, Tmaritima and Gmetallired. Abbreviations: CtetaniE88 = *C. tetani*E88_1817-150; Tmaritima = *T. maritima*MSB8_24050; and Gmetallired = *G. metallireducens*_401349760.

ments made by ClustalW (data not shown), but the results were not found to differ significantly. Figure 1 contains the ORFs from the two HKII groups, *ahk5* and *ahk22,* with the three closest bacterial ORFs (to MaceMA2890) from Cyanobacteria, Firmicutes and Proteobacteria. Figure 2 is composed of ORFs from the four HKVI 'CheA like' groups, *ahk40–43,* and the three closest bacterial ORFs (to MaceMA0014) from Thermatogae, Firmicutes and Proteobacteria. Figure 3 is the analysis of a number of RRI orphans, *arr7* and *arr10* (from putative taxis operons), *arr12* and *arr14* (> 200 amino acids)

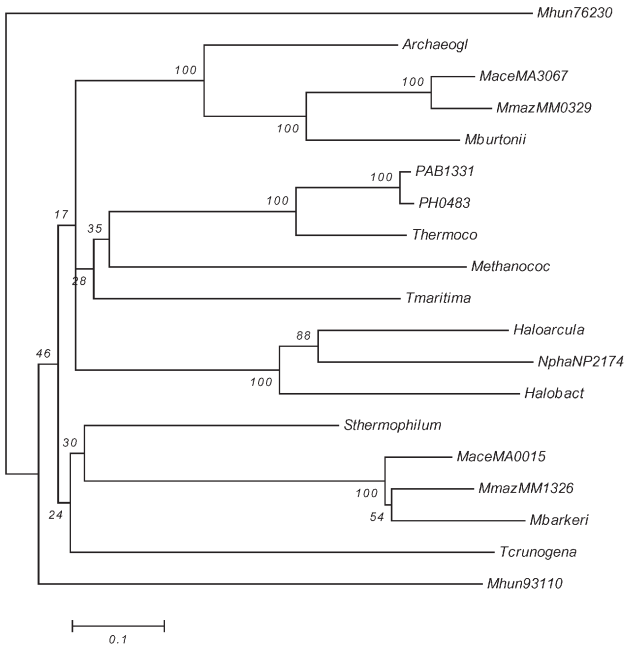


Figure 4. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from response regulator IV group *arr18* and *arr19*. Group *arr18* = MaceMA0015, MbarA_0985 and MmazMM1326; and group *arr19* = PAB1331, PH0483, Tkod_610170410, HaloVNG0973G, HmarrnAC2204, MaceMA3067, MmazMM0329, Mbur_401647530, AfulAF1041, MmarMMP0926, NphaNP2174A and Mhun_401793110. The bacterial ORFs are Sthermophilum, Tmaritima and Tcrunogena. Abbreviations: Thermoco = *T. kodakaraensis*; Sthermophilum = *S. thermophilum*_3768360; Tmaritima = *T.maritima*MSB8_21060; and Tcrunogena = *T. crunogena* Tcr0758

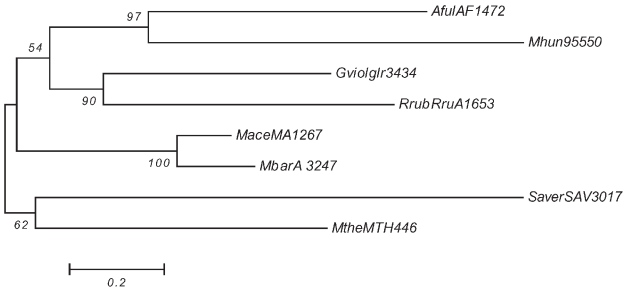


Figure 5. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from hybrid kinase 1 group *ahy2*. The bacterial ORFs are Gviolglr3434, RrubRruA1653 and SaverSAV3017. Abbreviations: Gviolglr3434 = *G. violaceus*; RrubRruA1653 = *R. rubrum*; and SaverSAV3017 = *S. avermitilis.*

with bacterial ORFs from Thermatogae, Firmicutes and Proteobacteria (closest to MaceMA3068). Figure 4 shows the results for the two RRIV CheB orthologous groups from taxis operons, *arr18* and *arr19* with bacterial representatives from Thermatogae, Proteobacteria and Actinobacteria (closest to MaceMA0015). Figure 5 shows results for *ahy2* and bacterial representatives from Cyanobacteria, Actinobacteria and Proteobacteria.

Linked genes

Genes that are located close to each other on the genome and transcribed in the same orientation are shown in Table A2. Most of these are likely to be part of operons. This provides clues to some cognate pairs of HKs, HYs and RRs. All putative HKVI encoding genes are located with other "chemotaxis" genes in "chemotaxis operons," including two such operons for *M. acetivorans* and *M. mazei*. Included in these "chemotaxis operons" are the orthologous groups, *ahk40–43, arr7, arr10* and *arr18* and *arr19* (see Table 4).

Conclusions

Distribution of two-component ORFs

Ten species, representing four genera have at least 17 putative two-component ORFs. Some of these two-component ORFs are quite sophisticated in structure, including the multiple sensor HKIIs, CheA-like HKVIs and the hybrid kinases. The results presented here show that a number of euryarchaeal species have an extensive array of two-component sensory ORFs. These proteins may sense a number of different internal signals by means of PAS/PAC domains and their associated cofactors (Ponting and Aravind 1997, Gilles-Gonzalez and Gonzalez 2004). In addition, the potential to sense other small molecules (particularl cNMPs) via the GAF domains (Aravind and Ponting 1997, Ho et al. 2000, Anantharaman et al. 2001) and extracellular signals, by the CHASE4 and Cache putative sensory domains, via HAMP domains (Anantharaman and Aravind 2001, Zhulin at al. 2003) shows that these organisms (in particular *H. marismortui, Natronomonas pharaonis, Methanospirillum hungatei* and the Methanosarcinales) possess sophisticated and complex sensory networks. As yet, none of these putative two-component genes have a functional name, so functions can be assigned only by similarity. The DNA-binding RRs are common in bacteria that regulate gene expression (Ashby 2004, Galperin 2005). However only three RRs have been identified with putative DNA binding domains, all in *H. marismortui*. If regular indiscriminate HGT were taking place, one would expect to see more DNA-binding RRs in archaeal sequences. Presumably the large number of orphan RRs are involved in regulation of cellular activity by interacting directly with other proteins. Transcriptional control is probably maintained by the many DNA-binding domains that have been identified as part of one-component systems in archaea (Ulrich et al. 2005). In these systems the DNA-binding output domain is linked directly to a sensor domain without any phosphotransfer.

Of the species that have the most two-component genes, *H. marismortui* and *Natronomonas pharaonis* are halophilic and the Methanosarcinales and *Methanospirillum hungatei* are mesophiles. The mesophiles coexist with a large and diverse population of bacteria, giving ample opportunity for HGT, whereas the opportunity for HGT in the halophilic organisms would be more restricted. This begs the question of how the distribution of two-component genes that can be seen in the

Euryarchaeota arose. Was it through HGT exclusively or by vertical transfer from a common ancestral euryarchaeal organism coupled with gene duplications?

Phylogeny and inheritance of two-component ORFs

The phylogenetic analysis of five different sets of orthologous ORFs, chosen because they are found in most of the species that contain two-component ORFs (Figures 1–5), were found to closely match the published phylogenies for these organisms (Matte-Tailliez et al. 2002, Brochier et al. 2004, Bapteste et al. 2005).

For *ahk5* and *ahk22*, shown in Figure 1, the phylogeny of each group agrees with the current phylogeny of these organisms and the position of the three bacterial examples indicates that the two groups may have arisen through a separate HGT event in an ancestral euryarchaeal species for *ahk5* and possibly, into an ancestral methanogen for *ahk22*.

The results for the CheA-like HKVI ORFs are shown in Figure 2. *Ahk40*, *ahk42* and *ahk43* (except Mhun_401793120) cluster together and probably represent vertical inheritance from a single HGT event into an ancestral Euryarchaeota species (one bacterial ORF from *T. maritima* giving the best match). *Ahk41* appears to be a separate group, found in the Methanosarcinales, that clusters on its own and seems to be more closely associated with the Firmicutes and Proteobacterial examples, presumably representing a separate HGT event. The three *Methanospirillum hungatei* ORFs seem to be due to separate (Mhun_401793120 probably should not be in *ahk43*) HGT events and Mhun_401784470 and Mhun_401776240 are probably true paralogues.

Figure 3 shows the results for four orphan RR groups. The two groups, associated with putative taxis operons *ahk7* and *ahk10*, group closely together, however, *ahk10*, which is found only in the Methanosarcinales is probably due to an HGT event into a direct ancestor of this group. The other two orthologous groups, *arr12* and *arr14* are quite separate from the first two mentioned groups (*arr7* and *arr10*) and probably arose from separate HGT events into the ancestors of methanogens (AfulAF2419 appears to be a distant member of *ahk12*).

Figure 4 shows the results for the two RRIV-CheB orthologous groups associated with taxis. The *arr18* orthologous group found in Methanosarcinales groups separately from *arr19*, being closer to two of the bacterial ORFs. Therefore *arr18* appears to be the result of a separate HGT event in an ancestor of the Methanosarcinales, whereas *arr19* appears to be the result of an HGT event into an ancestor of Euryarchaeota.

The phylogeny for *ahy2* (the biggest hybrid kinase orthologous group), shows that these members probably arose from more than one HGT event. The combined results for the orthologous groups found in potential taxis operons are shown in Table A2.

The operon that contains HKVI (*ahk40/42/43*), RRI (*arr7*) and RRIV-CheB (*arr19*) appears to have arisen as an HGT event that transferred the whole operon into an ancestor of the Euryarchaeota. In contrast, the taxis operon containing HKVI (*ahk41*), RRI (*arr10*) and RRIV-CheB (*arr18*) appears to have arisen from a separate HGT event of the whole operon into a direct ancestor of the Methanosarcinales.

The results presented here suggest that HGT has taken place from bacterial species both into ancestral Euryarchaeota and more recently into the methanogens. However the large numbers of two-component genes in the mesophilic methanogens and the Halobacteriales probably reflect their well known metabolic flexibility (Bapteste et al. 2005, Falb et al. 2005). This in turn, necessitates an increased requirement for regulation of cellular activity in a changing environment rather than the increased potential for HGT from bacteria. Most of the two-component ORFs that can be observed in these groups of organisms are probably derived from paralogous gene duplication events, the number of two-component ORFs observed would be driven by the requirement to control cellular activity as the organisms evolve. A limited number of HGT events could be sufficient to account for the diversity of phosphotransfer and sensory domains.

Any function of two-component ORFs is inferred by homology to known bacterial genes (e.g. HKVI and chemotaxis) and awaits in situ or in vitro studies, or both. This highlights the importance of interfacing between bioinformaticians and biochemists to plan experiments in an informed way, particularly where orthologues are identified and found in more than one genus and hence may play central roles in cellular regulation.

Acknowledgments

Mark Ashby was supported by New Initiative Funding from the University of the West Indies. The author wishes to thank John Allen, Elke Dittmann, Conrad Mullineaux and Ruth-Sarah Rose for critical reading of this manuscript.

References

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

Alves, R. and M.A. Savageau. 2003. Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: differences in molecular structure and physiological function. Mol. Microbiol. 48:25–51.

Anantharaman, V. and L. Aravind. 2001. The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors. TIBS 26:579–582.

Anantharaman, V., E.V. Koonin and L. Aravind. 2001. Regulatory potential, phyletic distribution and evolution of ancient, intracelluar small-molecule-binding domains. J. Mol. Biol. 307:1271–1292.

Appleman, J.A. and V. Stewart. 2003. Mutational analysis of a conserved signal-transducing element: the HAMP linker of the *Escherichia coli* nitrate sensor NarX. J. Bact. 185:89–97.

Aravind, L. and C.P. Ponting. 1997. The GAF domain: an evolutionary link between diverse phototransducing proteins. TIBS 22: 458–459.

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.

Ashby, M.K. 2004. Survey of the number of two-component response regulator genes in the complete and annotated genome sequences of prokaryotes. FEMS Microbiol. Lett. 231:277–281.

Baliga, N.S., R. Bonneau, M.T. Facciotti et al. 2004. Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. Genome Res. 14:2221–2234.

Bapteste, E., C. Brochier and Y. Boucher. 2005. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. Archaea 1:353–363.

Bateman, A., L. Coin, R. Durbin et al. 2004. The Pfam protein families database. Nucl. Acids Res. 32:D138–141.

Bibikov, S.I., L.A. Barnes, Y. Gitin and J.S. Parkinson. 2000. Domain organisation and flavin adenine dinucleotide-binding determinants in the aerotaxis signal transducer Aer of *Escherichia coli*. Proc. Natl. Acad. Sci. USA 97:5830–5835.

Boucher, Y., C.J. Douady, R.T. Papke, D.A. Walsh, M.E.R. Boudreau, C.L. Nesbø, R.J. Case and W.F. Doolittle. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu. Rev. Genet. 37: 283–328.

Brochier, C., P. Forterre and S. Gribaldo. 2004. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. Genome Biology 5:R17.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton and J.D. Gocayne. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus janaschii*. Science 273:1058–1073.

Cohen, G.N., V. Barbe, D. Flament et al. 2003. An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abysii*. Mol. Microbiol. 47:1495–1512.

Deppenmeier, U., A. Johann, T. Hartsch et al. 2002. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J. Mol. Microbiol. Biotechnol. 4:453–461.

Falb, M., F. Pfeiffer, P. Palm, K. Rodewald, V. Hickmann, J. Tittor and D. Oesterhelt. 2005. Living with two extremes: conclusions from the genome sequence of *Natromonas pharaonis*. Genome Res. 15:1336–1343.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.

Forterre, P., C. Brochier and H. Philippe. 2002. Evolution of the Archaea. Theoret. Popul. Biol. 61:409–422.

Galagan, J.E., C. Nusbaum, A. Roy et al. 2002. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. Genome Res. 12:532–542.

Galperin, M.Y. 2005. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. BMC Microbiol. 5:35.

Galperin, M.Y., A.N. Nikolskaya and E.V. Koonin. 2001. Novel domains of the prokaryotic two-component signal transduction systems. FEMS Microbiol Lett 203:11–21.

Gilles-Gonzalez, M.-A. and G. Gonzalez. 2004. Signal transduction by heme-containg PAS-domain proteins. J. Appl. Physiol. 96: 774–783.

Grebe, T.W. and J.B. Stock. 1999. The histidine protein kinase superfamily. Adv. Microbial. Physiol. 41:139–227.

Hellingwerf, K.J. 2005. Bacterial observations: a rudimentary form of intelligence? Trends Microbiol. 13:152–8.

Ho, Y.-S., L. Burden and J.H. Hurley. 2000. Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor. EMBO J. 19:5288–5299.

Hoch, J.A. 2000. Two-component and phosphorelay signal transduction. Curr. Opin. Microbiol. 3:165–170.

Jahreis, K., T.B. Morrison, A. Garzón and J.S. Parkinson. 2004. Chemotactic signaling by an *Escherichia coli* CheA mutant that lacks the binding domain for phosphoacceptor partners. J. Bact. 186:2662–2672.

Karniol, B. and R.D. Vierstra. 2004. The HWE histidine kinases, a new family of bacterial two-component sensor kinases with potentially diverse roles in environmental signaling. J. Bact. 186: 445–453.

Kawarabayasi, Y., M. Sawada, H. Horikawa et al. 1998. Complete sequence and gene organisation of the genome of a hyper-thermophilic archaebacterium, *Pysococcus horikoshii* OT3. DNA Res. 5:55–76.

Klenk, H-P., R.A. Clayton, J.-F. Tomb et al. 1997. The complete sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeglobus fulgidus*. Nature 390:364–370.

Koonin, E.V. 2003. Horizontal gene transfer: the path to maturity. Mol. Microbiol. 50:725–727.

Koretke, K.K., A.N. Lupas, P.V. Warren, M. Rosenberg and J.R. Brown. 2000. Evolution of two-component signal transduction. Mol. Biol. Evol. 17:1956–1970.

Kumar, S., K. Tamura and M. Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinform. 5:150–163.

Kurland, C.G., B. Canback and O.G. Berg. 2003. Horizontal gene transfer: A critical view. Proc. Natl. Acad. Sci. USA 100: 9658–9662.

Lawrence, J.G. and H. Hendrickson. 2003. Lateral gene transfer: when will adolescence end? Mol. Microbiol. 50:739–749.

Makarova, K.S. and E.V. Koonin. 2003. Comparative genomics of archaea: how much have we learned in six years, and what's next? Genome Biology 4:115.

Makarova, K.S. and E.V. Koonin. 2005. Evolutionary and functional genomics of the Archaea. Curr. Opin. Microbiol. 8:586–94.

Matte-Tailliez, O., C. Brochier, P. Forterre and H. Philippe. 2002. Archael phylogeny based on ribosomal proteins. Mol. Biol. Evol. 19:631–639.

Nelson, K.E., R.A. Clayton, S.R. Gill et al. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermatoga maritima*. Nature 399:323–329.

Ng, W.V., S.P. Kennedy, G.G. Mahairas et al. 2000. Genome sequence of *Halobacterium* species NRC-1. Proc. Natl. Acad. Sci. USA 97:12176–12181.

Notredame, C., D.G. Higgins and J. Heringa. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–17.

Ochman, H., J.G. Lawrence and E.A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405: 209–304.

Ohmori, M., M. Ikeuchi, N. Sato et al. 2001. Characterization of genes encoding multi-domain proteins in the genome of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. 8:271–284.

Ponting, C.P. and L. Aravind. 1997. PAS: a multifunctional domain family comes to light. Curr. Biol. 7:R674–R677.

Sardiwall, S., S.L. Kendall, F. Movahedzadeh, S.C. Rison, N.G. Stoker and S. Djordjevic. 2005. A GAF domain in the hypoxia/NO-inducible *Mycobacterium tuberculosis* DosS protein binds haem. J. Mol. Biol. 353:929–936.

Slesarev, A.I., K.V. Mezhevaya, K.S. Makarova et al. 2002. The complete genome of hyperthermophile *Methanopyrs kanleri AV19* and monophyly of archaeal methanogens. Proc. Natl. Acad. Sci. USA 99:4644–4649.

Smith, D.R., L.A. Doucette-Stamm, C. Deloughery et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional analysis and comparative genomics. J. Bact. 179:7135–7155.

Stewart, R.C. and R. van Bruggen. 2004. Association and dissociation kinetics for CheY interacting with the P2 domain of CheA. J. Mol. Biol. 336:287–301.

Stock, A.M., V.L. Robinson and P.N. Goudreau. 2000. Two-component signal transduction. Annu. Rev. Biochem. 69:183–215.

Tam, R. and M.H. Saier. 1994 Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. Microbiol. Rev. 57:320–46.

Thompson, J.D., D.G. Higgins and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. 22:4673–80.

Ulrich, L.E., E.V. Koonin and I.B. Zhulin. 2005. One-component systems dominate signal transduction in prokaryotes. Trends Microbiol. 13:52–56.

West, A.H., E. Martinez-Hackert and A.M. Stock. 1995. Crystal structure of the catalytic domain of the chemotaxis receptor methylesterase, CheB. J. Mol. Biol. 250:276–290.

Zhu, Y. and M. Inouye. 2004. The HAMP linker in histidine kinase dimeric receptors is critical for symmetric transmembrane signal transduction. J. Biol. Chem. 279:48152–48158.

Zhulin, I.B., A.N. Nikolskaya and M.Y. Galperin. 2003. Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in bacteria and archaea. J. Bact. 185:285–294.

Appendix

Table A1 shows the two-component protein domains used for BLASTP. Figures A1 to A5 show maximum likelihood phylogenetic analyses. Table A2 shows closely linked genes that may be part of operons.

Table A1. Two-component protein domains from *M. acetovrans* (*M. ace*) and *Escherichia coli* K12 (*E. coli*) used for BLASTP searches, showing the online accession numbers and the amino acid range that was used. Abbreviation: aa = amino acids.

Domain	Species	Gene	Amino acid region
CheY Receiver	*E. coli*	NP_416396.1	
	M. ace	MA0016	
	M. ace	MA3068	
OmpR Receiver	*E. coli*	NP_417864.1	aa 6–124
Histidine kinase	*E. coli*	NP_417863.1	aa 234–439
	M. ace	MA0490 (HisKA_2)	aa 639–847
Hpt	*E. coli*	NP_415513.1	aa 815–896
	M. ace	NP_614988	aa 5–106

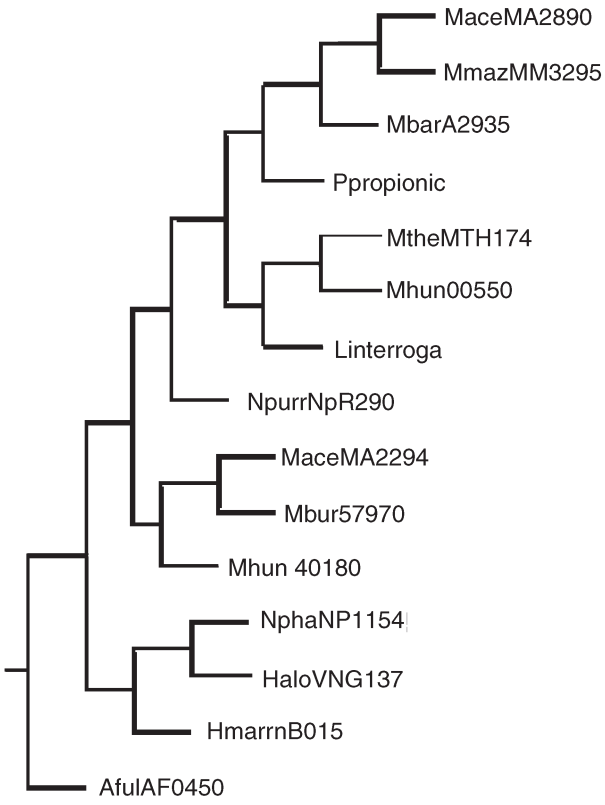


Figure A1. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from histidine kinase II groups *ahk5* and *ahk22*. Group *ahk5* = MaceMA2890, MbarA_2935, MmazMM3295, MtheMTH174 and Mhun_401800550; and group *ahk22* = MaceMA2294, Mbur_401657970, HaloVNG1374G, Mhun_401803800, AfulAF0450, NphaNP1154A and HmarrnB0156. The bacterial ORFs are Ppropionicus, NpunNpR2903 and Linterrogans. Abbreviations: Ppropionicus = *P. propionicus*_500413890; NpunNpR2903 = *N. punctiforme*; and Linterrogans = *L. interrogans*-LA2540.

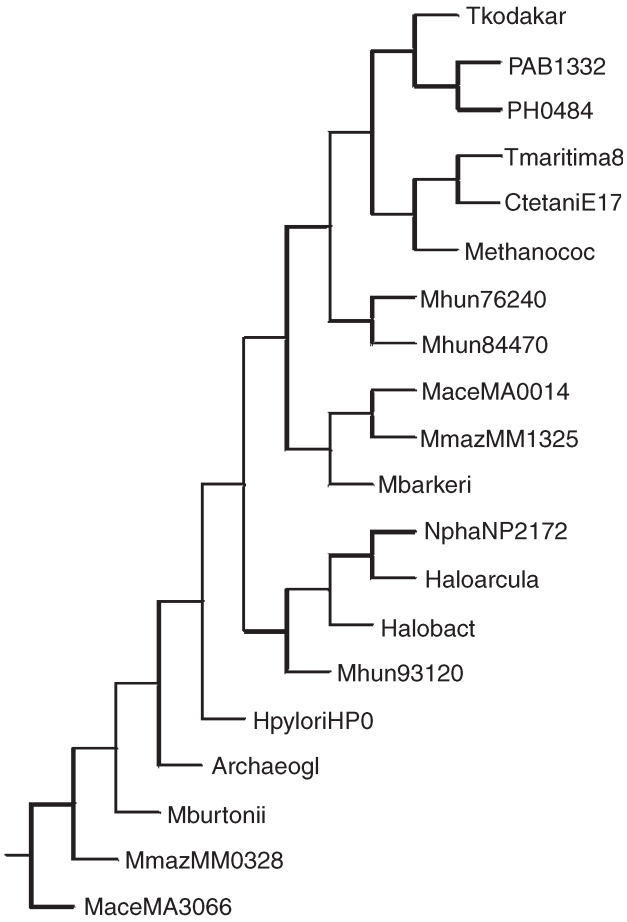


Figure A2. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from histidine kinase VI groups *ahk40–43*. Group *ahk40* = MaceMA3066, MmazMM0328, AfulAF1040 and Mbur_ 401647520; group *ahk41* = MaceMA0014, MmazMM1325 and MbarA_ 0984; group *ahk42* = PAB1332, PH-0484, Tkod_610170420/30 and MmarMMP0927; and group *ahk43* = HaloVNG0971G, HmarrnAC2205, NphaNP2172A and Mhun_401793120. The bacterial ORFs are CtetaniE17170, Tmaritima-MSB_824070 and HpyloriHP0392. Abbreviations: CtetaniE17170 = *C. tetani*E88_1817170; *T. maritima* = TmaritimaMSB_824070; and *H. pylori* = HpyloriHP0392.

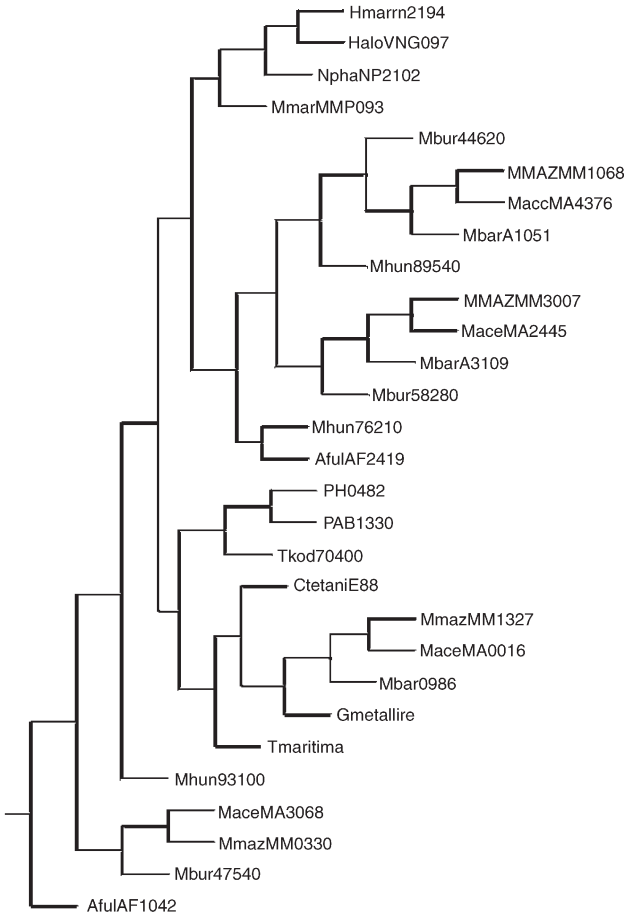


Figure A3. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from response regulator I groups *arr7*, *arr10*, *arr12* and *arr14*. Group *arr7* = HaloVNG0974G, AfulAF1042, MmarMMP0933, Mbur_401647540, MaceMA3068, MmazMM0330, PAB1330, PH0482, HmarrnAC2194, Tkod_610170400, Mhun_401793100 and NphaNP2102A; group *arr10* = MaceMA0016, MbarA_0986 and MmazMM1327; group *arr12* = MaceMA4376, MbarA_1051, MmazMM1068, AfulAF2419, Mbur_401644620 and Mhun_401789540; and group *arr14* = MaceMA2445, MbarA_3109, MMAZMM3007, Mbur_401658280 and Mhun_401776210. The bacterial ORFs are CtetaniE88, Tmaritima and Gmetallired. Abbreviations: CtetaniE88 = *C. tetani*E88_1817150; Tmaritima = *T. maritima*MSB8_24050; and Gmetallired = *G. metallireducens*_401349760.

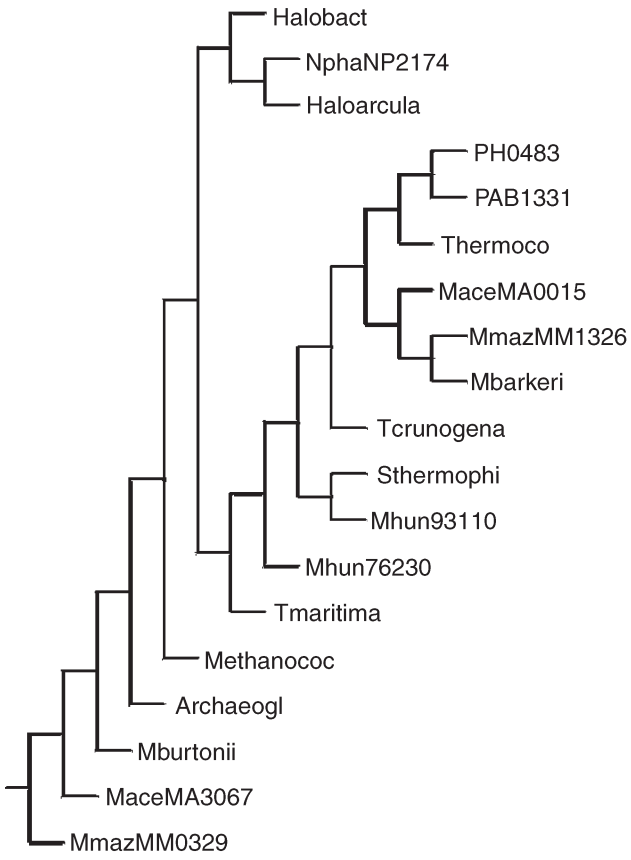


Figure A4. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from response regulator IV groups *arr18* and *arr19*. Group *arr18* = MaceMA0015, MbarA_0985 and MmazMM1326; and group *arr19* = PAB1331, PH0483, Tkod_610170410, HaloVNG0973G, HmarrnAC2204, MaceMA3067, MmazMM0329, Mbur_401647530, AfulAF1041, MmarMMP0926, NphaNP2174A and Mhun_401793110. The bacterial ORFs are Sthermophilum and Tmaritima, Tcrunogena. Abbreviations: Thermoco = *T. kodakaraensis*; Sthermophilum = *S. thermophilum*_3768360; Tmaritima = *T.maritima*MSB8_21060; and Tcrunogena = *T. crunogena* Tcr0758.

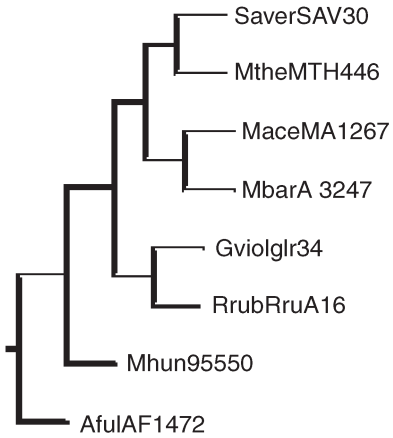


Figure A5. Phylogenetic analysis by neighbor-joining of putative two-component open reading frames (ORFs) from hybrid kinase 1 group *ahy2*. The bacterial ORFs are Gviolglr3434, RrubRruA1653 and SaverSAV3017. Abbreviations: Gviolglr3434 = *G. violaceus*; RrubRruA1653 = *R. rubrum*; and SaverSAV3017 = *S. avermitilis*.

Table A2. Closely linked genes that may be part of operons.

Assigned gene no.	Classification	Orthologue number	Additional information
Archaeoglobus fulgidus			
AF448	His_KA PACPASGAF		
AF449	RRI-CheY		
AF450	HKIII + 1 PASPAC		
AF1042	RRI-CheY	*arr7*	AF1055 mcp
AF1041	RRIV-CheB	*arr18*	AF1044CheW
AF1040	HKVI	*arr40*	AF1039 CheC
AF1472	HYI-HisKA	*ahy2*	
AF1473	RRI-CheY	*arr9*	
AF2419	RRI-CheY	*arr12*	
AF2420	His_KA PACPASGAF		
Halobacterium			
VNG0974G	RRI-CheY	*arr7*	VNG0976G CheW
VNG0973G	RRIV-CheB	*arr19*	VNG0970G CheC1
VNG0971G	HKVI	*ahk43*	VNG0967G CheD
VNG0966G CheR	HKIII	*ahk22*	
VNG1374G	HATPase HAMP		
VNG1375G			
VNG2036C	RRI-CheY	*arr1*	
VNG2037C	HKII		
Haloarcula marismortui			
rrnAC0410	HKII + 2PASPAC	*arr9*	
rrnAC0411	RRI-CheY		
rrnAC0412	HYI		
rrnAC0413	HKII + 1PASPAC		
rrnAC0456	HATPase_c+ gyra		Top6B
rrnAC0457	HATPase_c		Top6A
rrnAC2204	RRIV-CheB	*arr19*	*rrnAC2206 CheR*
rrnAC2205	HKVICheW-CheA	*ahk43*	
rrnAC2692	HYI + PASPACGAF		
rrnAC2694	HATPase_c + PAS		
Methanobacter thermoautotrophicus			
MTH444	HKI	*ahy9*	
MTH445	RRI-CheY	*arr9*	
MTH446	HYI + $(PASPAC)_2$	*ahy2*	
MTH447	RRI-CheY	*arr6*	
MTH457	RRIV-PACPAS		
MTH459	HKI		
MTH548	RRIV-G_transf		
MTH549	RRI-CheY		
MTH901	HYI		
MTH902	HYI-PASPAC		
Methanococcus maripaludis			
MMP0926	RRI-CheB	*arr19*	MMP0925 CheW
MMP0927	HKVI	*ahk42*	MMP0928 CheD
MMP0933	RRI-CheY	*arr7*	MMP0929 mcp
MMP1303	HKIV	*ahk32*	
MMP1304	RRI-CheY	*arr2*	
Methanosarcina acetivorans			
MA0014	HKVI	*ahk41*	
MA0015	RRI-CheB	*arr19*	
MA0016	RRI-CheY	*arr10*	
MA0018	RRI-CheY	*arr16*	MA0019 mcp MA0020 CheW

continued on facing page

Table A2. Cont'd. Closely linked genes that may be part of operons.

Assigned gene no.	Classification	Orthologue number	Additional information
Methanosarcina acetivorans cont'd			
MA0551	HKII + PASPAC		
MA0552	HKIII	*ahk9*	
MA0619	HKIII	*ahk12*	
MA0620	HKIII	*ahk14*	
MA0758	HKIII		
MA0759	HKIII	*ahk30*	
MA1267	HYI + PASPAC	*ahy2*	
MA1268	RRI-CheY	*arr9*	
MA1269	RRI-CheY		
MA1270	HKII + PASPAC		
MA1468	RRI-CheY		
MA1469	RRI-CheY	*arr8*	
MA1470	HKIII	*ahk8*	
MA1627	HKIII	*ahk25*	
MA1628	HKII + PASPAC	*ahk4*	
MA1645	HKIII	*ahk17*	
MA1646	HKII + PASPAC	*ahk6*	
MA2012	RRI-CheY		
MA2013	HYII + Hpt		
MA3066	HKVI	*ahk40*	MA3063 CheR
MA3068	RRI-CheY	*arr7*	MA3064 CheD
MA3067	RRI- CheB	*arr18*	MA3065 CheC
			MA3070 CheW
MA3368	HKIII	*ahk24*	
MA3370	HKII + PASPAC		
MA4376	RRI-CheY	*arr12*	
MA4377	HYIII CHSE4HP	*ahy7*	
Methanosracina barkeri			
MbarA_0984	HKVI	*ahk41*	MbarA_0983 CheR
MbarA_0985	RRIV-CheB	*arr19*	
MbarA_0986	RRI-CheY	*arr10*	
MbarA_0988	RRI-CheY		MbarA_0989 mcp
Opp orientation to above			MbarA_0990 CheW
MbarA_1051	RRI-CheY	*arr12*	
MbarA_1052	HYIII	*ahy7*	
MbarA_3036	HKII		
MbarA_3037	HKII		
MbarA_3247	HYIPASPAC		
MbarA_3248	RRI-CheY		
MbarA_3250	HKII		
MbarA_3447	HKII		
MbarA_3448			
Methanosracina mazei			
MM0168	HKII	*ahk9*	
MM0169	HKIII		
MM0328	HKVI	*ahk40*	MM3025 CheR
MM0329	RRI- CheB	*arr18*	MM0326 CheD
MM0330	RRI-CheY	*arr7*	MM0327 CheC
			MM0332 CheW
			MM0333 mcp
MM1325	HKVI	*ahk41*	MM1323 CheC
MM1326	RRI-CheY	*arr19*	MM1324
MM1327		*arr10*	CheB
MM1328 opp orientation to above	RRI-CheY	*arr16*	MM1329 mcp
			MM1330 CheW

continued overleaf

Table A2. Cont'd. Closely linked genes that may be part of operons.

Assigned gene no.	Classification	Orthologue number	Additional information
Methanosracina mazei cont'd.			
MM2275	HKIII	*ahk28*	
MM2276	HKIII		
MM2277	HKIII		
MM2515	HKIII	*ahk8*	
MM2516	RRI-CheY	*arr8*	
MM2880	RRI-CheY	*arr3*	
MM2881	HYIII	*ahy6*	
MM2953	RRI-CheY	*arr4*	
MM2954	RRI-CheY	*arr5*	
MM2955	HKVIPASPACCach		
MM3205	HYIII		
MM3206	RRI-CheY		
Pyrococcus abyssi			
PAB1330	RRI-CheY	*arr7*	PAB1329 CheR
PAB1331	RRVI-CheB	*arr19*	PAB1333 CheC
PAB1332	HKVI	*ahk42*	PAB1334 CheC
			PAB1335CheW
			PAB1336 mcp
Pyrococcus horikoshi			
PH0482	RRI-CheY	*arr7*	PH0481 CheB
PH0483	RRIV-CheB	*arr19*	
PH0484	HKVI	*ahk42*	