J Mol Evol (2004) 59:725–737
DOI: 10.1007/s00239-004-2663-2

JOURNAL OF MOLECULAR EVOLUTION

© Springer Science+Business Media, Inc. 2004

# Evolutionary Analysis of the Two-Component Systems in *Pseudomonas aeruginosa* PAO1

**Ying-Tsong Chen,[1] Hwan You Chang,[2] Chin Lung Lu,[3] Hwei-Ling Peng[1]**

[1] Department of Biological Science and Technology, National Chiao-Tung University, Hsinchu, Taiwan
[2] Institute of Molecular Medicine, National Tsing Hua University, Hsinchu, Taiwan
[3] Institute of Bioinformatics, National Chiao-Tung University, Hsinchu, Taiwan

Received: 19 May 2004 / Accepted: 6 July 2004 [*Reviewing Editor:* Dr. Martin Kreitman]

**Abstract.** Gene organization and functional motif analyses of the 123 two-component system (2CS) genes in *Pseudomonas aeruginosa* PAO1 were carried out. In addition, NJ and ML trees for the sensor kinases and the response regulators were constructed, and the distances measured and comparatively analyzed. It was apparent that more than half of the sensor-regulator gene pairs, especially the 2CSs with OmpR-like regulators, are derivatives of a common ancestor and have most likely co-evolved through gene pair duplication. Several of the 2CS pairs, especially those with NarL-like regulators, however, appeared to be relatively divergent. This is supportive of the recruitment model, in which a sensor gene and regulator gene with different phylogenetic history are assembled to form a 2CS. Correlation of the classification of sensor kinases and response regulators provides further support for these models. Upon comparison of the phylogenetic trees comprised of sensors and regulators, we have identified six congruent clades, which represent the group of the most recently duplicated 2CS gene pairs. Analyses of the congruent 2CS pairs of each of the clades revealed that certain paralogous 2CS pairs may carry a redundant function even after a gene duplication event. Nevertheless, comparative analysis of the putative promoter regions of the paralogs suggested that functional redundancy could be prevented by a differential control. Both codon usage and G+C content of these 2CS genes were found to be comparable with those of the *P. aeruginosa* genome, suggesting that they are not newly acquired genes.

**Key words:** Two-component system — *Pseudomonas aeruginosa* — PAO1 — Co-evolved — Recruitment model

## Introduction

The two-component system (2CS) is the means by which bacteria commonly regulate an adaptive response to versatile environments. A 2CS often comprises a sensor histidine kinase and a response regulator (Stock et al. 1990). The sensor kinase consists of at least one signal recognition (input) domain coupled to an autokinase (transmitter) domain. Signals binding to the input domain cause activation of the autokinase and, thereby, hydrolysis of an ATP molecule to phosphorylate a conserved histidine residue (Stock et al. 1989). The phosphate group is subsequently transferred to the conserved aspartate residue at the receiver domain of a response regulator.

Most sensor kinases contain one *i*nput domain and one *t*ransmitter domain and are hence called classical (IT-type) sensors. Some sensors contain both the sensor kinase signature and a *r*eceiver domain of the response regulator and are thus referred to as hybrid

*Correspondence to:* Hwei-Ling Peng; *email:* hlpeng@cc.nctu.edu.tw

or ITR-type sensory kinases (Ishige et al. 1994). A smaller fraction of the hybrid sensors possesses an additional *o*utput domain at the carboxyl terminus and are referred to as ITRO-type or unorthodox sensor kinases. The response regulator, in most cases, is a transcription factor for genes whose expressions correspond to the input signal. Phosphorylation of the aspartate residue activates the output domain to modulate an appropriate expression of the target gene (Parkinson 1993). Response regulators other than transcription factors have also been reported. For example, the CheY response regulator, after being phosphorylated by the kinase CheA, binds to a flagella motor to promote a clockwise rotation of the flagella (Macnab 1996).

The tight connection between the functionally coupled bacterial genes and their chromosomal vicinity is a common feature of bacterial genomes (Overbeek et al. 1999; Dandekar et al. 1998). Most of the 2CS genes encoding functionally coupled sensors and regulators are also physically linked as an operon in the genomes. Two models, the co-evolution and recruitment models, have been proposed to explain the evolution of 2CS genes. The co-evolution model proposes that the majority of the 2CS genes in a genome have been aroused by gene duplication and subsequent differentiation of the ancestral 2CSs (Koretke et al. 2000). This is supported by the fact that many of the coupled 2CS genes are concurrent in a genome. On the other hand, the recruitment model suggests that some of the 2CS operons have evolved as the result of an assembly of a sensor gene and a regulator gene from heterologous 2CSs. Signal transductions between 2CSs encoded by distantly located genes have been reported in the sporulation system of *B. subtilis* (Kobayashi et al. 1995) and *E. coli* hybrid sensor kinases, respectively (Hoch and Sihavy 1995). It is conceivable that, in the recruitment model, further assembly of such distantly located 2CS genes into an operon would be beneficial for a coordinate control of the system.

*Pseudomonas aeruginosa* is a flexible Gram-negative bacterium that grows in a variety of environmental habitats. Patients with cystic fibrosis, burn victims, and patients requiring extensive hospitalization are particularly at risk of *P. aeruginosa* infections (Goldberg et al. 2000). The complete genome sequence of *P. aeruginosa* PAO1 has been determined and published (Stover et al. 2000). The 6.3-Mb genome contains 5570 predicted genes, of which 123 2CSs were annotated according to the most recently updated database of the Pseudomonas Genome Project. The number of 2CS genes in the *P. aeruginosa* genome is relatively high in comparison with that in the *E. coli* and *Bacillus* genomes, which is likely to be advantageous for the bacteria to adapt to different environments. Nevertheless, the function of approximately two-thirds of the 2CS genes has not been characterized. In this study, we have performed analyses of the phylogenetic relationship of the 2CS genes in *P. aeruginosa* PAO1 in the hope that it might reveal some implication of their functions.

## Materials and Methods

### *Nucleotide Sequence Source and Sequence Analysis*

The known and putative 2CS genes annotated by the *Pseudomonas aeruginosa* Community Annotation Project (PseudoCAP) were obtained from the Web site http://www.pseudomonas.com. The sequences of sensor kinase genes, hybrid sensor genes, and response regulator genes were collected and processed into FASTA format. Analysis of the 2CS was performed by homology search using the BLAST programs provided by the National Center of Biotechnology Information through the Internet.

### *Multiple Sequence Alignment and Phylogenetic Estimation*

Neighbor-joining (NJ) trees built with the deduced amino acid sequences for sensor kinases and response regulators were done using CLUSTAL W 1.81 (Thompson et al. 1994). Default substitution matrix (Gonnet) was used for alignments, and the positions with gaps were excluded in the tree construction. The resultant trees were visualized by TreeView 1.6 (Page 1996) and MEGA2 (Kumar et al. 2001).

For the maximum likelihood (ML) analysis, multiple sequence alignments of the amino acid sequences of sensors and regulators from homologous gene clusters were performed, also using CLUSTAL W. The positions containing alignment gaps were subsequently excluded manually using BioEdit 4.8.6 (Hall 1999). Pairwise distances were analyzed by the PROML algorithm (with JTT amino acid change model) in PHYLIP 3.6 (Felsenstein 1993), and 1000 replications of bootstrap sampling were performed for each analysis. Graphical representations of the multiple amino acid sequence alignments, the sequence logos, are presented using WebLogo (Crooks et al. 2004).

### *GC%, and G + C Content in the Third Positions of Synonymous Codons (GC3s) and the Effective Number of Codons Used in a Gene ($N_c$)*

The GC% and G + C content in the GC3s were calculated using CodonW (Peden 1999). $N_c$, the measure of overall codon bias in a gene, was calculated using the CHIPS program with Wright's (1990) $N_c$ statistic for the effective number of codons used.

## Results and Discussion

### *Organization of 2CSs Encoding Gene Clusters*

The 123 annotated 2CSs in the *P. aeruginosa* PAO1 genome, including 64 sensor and 59 regulator genes, were chosen for this study. The discrepancy in the numbers of sensor kinases and regulator genes, compared to the earlier reports of 64 and 63 sensor kinases and regulator genes, respectively (Rodrigue

Group Ia
(R↱S↱)
0179(CY)-0178(CA)
0463(Omp)-0464(IT)
0756(Omp)-0757(IT)
1157(Omp)-1158(IT)
1179(Omp)-1180(IT)
1437(Omp)-1438(IT)
1799(Omp)-1798(IT)
2479(Omp)-2480(IT)
2523(Omp)-2524(IT)
2657(Omp)-2656(IT)
2686(Omp)-2687(IT)
2809(Omp)-2810(IT)
3045(Nar)-3044(ITRO)
3077(Omp)-3078(IT)
3192(Omp)-3191(IT)
4101(Omp)-4102(IT)
4381(Omp)-4380(IT)
4776(Omp)-4777(IT)
4885(Omp)-4886(IT)
4983(Omp)-4982(ITRO)
5200(Omp)-5199(IT)
5360(Omp)-5361(IT)
5483(Ntr)-5484(IT)

Group Ib
(S↰↱R↱S)
0928(ITRO)-0929(Omp)-0930(IT)

Group Ic
(R↰↱R↱S)
3948(Nar)-3947(UC)-3946(ITRO)

Group Id
(R↱X↱S)
0408(CY)-0409(CY)-X-X-X-0413(CA)
1456(CY)-X -1458(CA)

Group IIa
(S↱R↱)
0600(IT)-0601(Nar)
1098(IT)-1099(Ntr)
1336(IT)-1335(Ntr)
1636(IT)-1637(Omp)
1979(IT)-1980(Nar)
2882(IT)-2881(UC)
3878(IT)-3879(Nar)
4197(IT)-4196(Nar)
4494(IT)-4493(UC)
4546(IT)-4547(Ntr)
4725(IT)-4726(Ntr)
5124(IT)-5125(Ntr)
5165(IT)-5166(Ntr)
5262(IT)-5261(UC)
5512(IT)-5511(Ntr)

Group IIb
(S↱X↱R)
3704(CA)-X-3702(CY)

Group IIIa
(S↰↱R)
1396(ITR)-1397(Nar)
2571(IT)-2572(UC)

Group IIIb
(S↰X↱R)
4036(IT)-X-X-X-4032(Omp)
4293(IT)-X-X-4296(Nar)

Group IV
(Orphan S)
0471(UC)
1243(ITR)
1611(ITR)
1976(ITR)
1992(ITR)
2177(ITR)
2583(ITR)
2824(ITR)
3271(ITR)
3462(ITR)
3974(ITR)
4117(IT)
4112(ITRO)
4856(ITR)

Group V
(Orphan R)
0034(Nar)
2798(UC)
3346(UC)
3604(Nar)
3714(Nar)
4781(UC)
4843(UC)
5364(UC)

**Fig. 1.** Classification of the 123 annotated 2CS genes in the *P. aeruginosa* PAO1 genome. The arrows indicate the transcription direction of each gene. The gene numbers are given in accordance with that of PseudoCAP. R, regulator; S, sensor; X, any open reading frame; CA, CheA-like; CY, CheY-like; Nar, NarL-like; Ntr, NtrC-like; Omp, OmpR-like; UC, unclassified. The genes without an adjacent 2CS gene are listed as orphan S and orphan R.

et al. 2000), is most likely due to the recent refinement of the annotation by the Pseudomonas Genome Project.

All these 2CS genes were first classified according to their relative location, gene organization, and transcription orientation. As shown in Fig. 1, each sensor gene was found to be either located adjacent to a regulator gene by direct linkage or separated by less than three open reading frames (ORFs) except for 14 sensor genes which were assigned as orphan sensors in Group IV. In the most common type of gene organization as represented by the 29 2CS gene pairs in Group I, the regulator gene is located upstream to the sensor gene. Two 2CS gene clusters within this group contain an additional 2CS gene (a sensor gene in Group Ib and a regulator gene in Group Ic), which is transcribed in the direction opposite to that of the paired regulator and sensor genes. Group Id contains four gene clusters with one or three non-2CS ORFs located in between the regulator and the sensor genes.

Group II contains 16 pairs of 2CS genes, with the gene order sensor followed by regulator. There are four 2CS gene clusters in Group III, where the regulator and the sensor genes are transcribed divergently. The rest of the 2CS genes, including 14 sensors and 8 regulator genes, are not physically linked to any 2CS gene and are hence referred to as orphan sensors and regulators, respectively.

### *Analysis of the 2CS Genes Based on Functional Motifs*

These 2CS genes were further analyzed on the basis of the functional motifs of their gene products. The average length of the response regulator genes is approximately 850 bp. Twenty-four of the regulators are members of the OmpR transcription factor family, which forms the largest group of the 2CS response regulators in *P. aeruginosa* PAO1. Apart from the 11 NarL-, 8 NtrC-, and 5 CheY-type regulators, the rest of the 11 regulators with signal-receiving motifs were found not to contain the conserved C-domain for classification and therefore were listed as unclassified (Fig. 1).

In contrast to that of the regulator genes, the size of the sensor genes varies greatly, ranging from 650 to 7418 bp. Classification of the 64 sensor genes was as follows: 42 IT (classic), 12 ITR (hybrid), 5 ITRO (unorthodox), and 4 CheA type based on the functional motifs of their gene products (Rodrigue et al. 2000). The 650-bp PA0471, which encodes the iron sensor Fur, is the only unclassified sensor gene (Fig. 1).

Combining the analysis of gene organization and the structural motifs of these gene products, several interesting features were noted, as follows.

(1) Almost all (20 of 23) of the Group Ia gene clusters carry an OmpR-like regulator, and most of the OmpR-like regulators (22 of 24) have accompanying classical (IT) sensor genes located adjacently downstream. This observation indicates that the gene order of a regulator-to-classical sensor is preferred by the family of OmpR-like regulators. It is likely that most of the regulator–sensor pairs in Group Ia were co-evolved by duplication from an ancestral OmpR–IT pair so that the gene organization remained unchanged. Moreover, 15 of the 20 OmpR–IT 2CS gene clusters consist of the regulator gene overlapped with the downstream sensor gene, which also supports the co-evolution model. In the *E. coli* K12 MG1655 genome, 11 of the 14 2CSs of the OmpR-like family exert the gene order regulator-to-sensor according to the KEGG database (Kanehisa et al. 2002). A similar phenomenon is also observed in the genome of *Bacillus subtilis*, where all of the 14 2CS genes of the OmpR-family are in a regulator-to-sensor organization (Fabret et al. 1999). It is likely that most of the 2CS gene clusters of the OmpR family have originated from a common progenitor before the

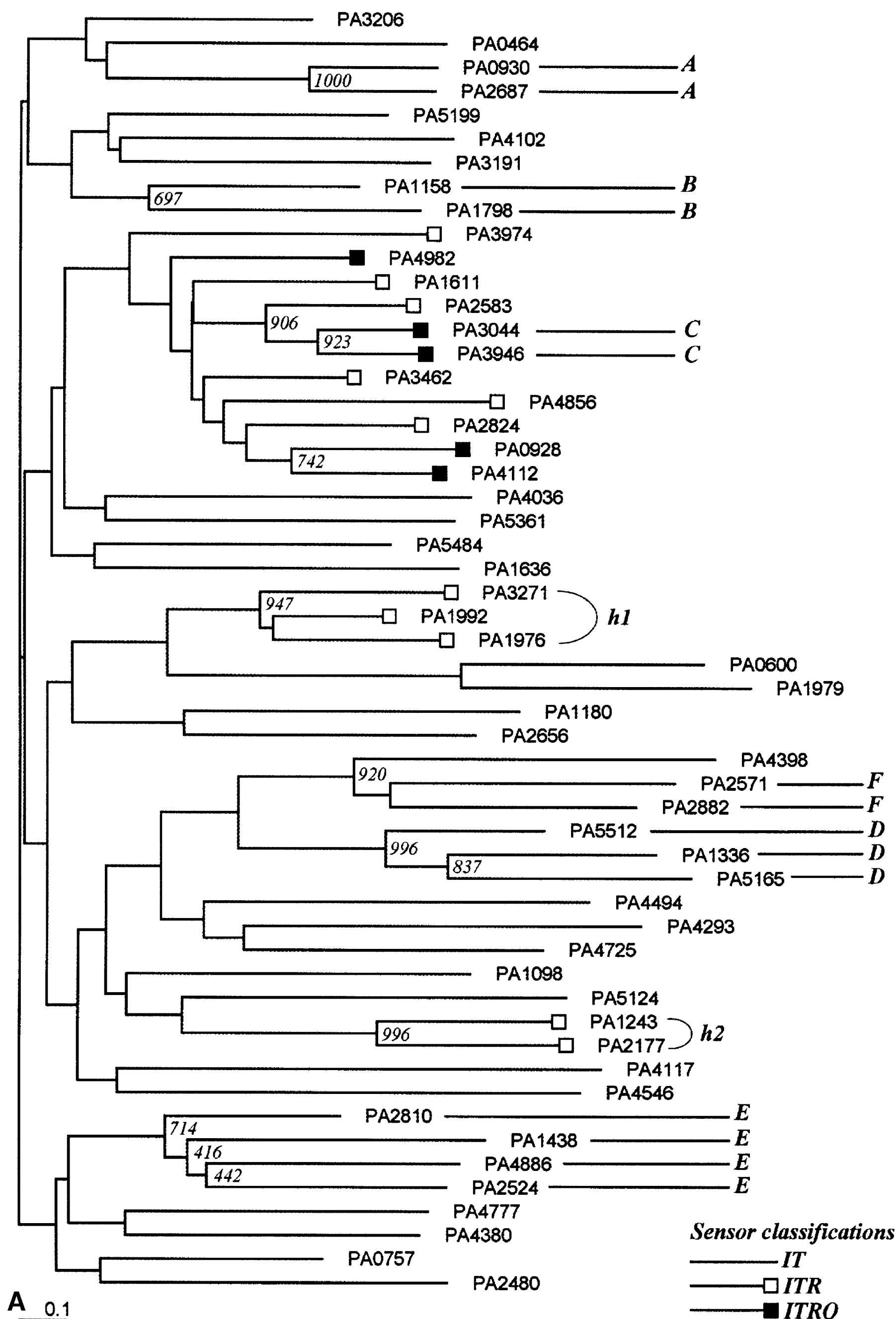



**Fig. 2.** Phylogenetic trees of the 54 sensor kinases (**A**) and 58 response regulators (**B**) constructed by CLUSTALW using the deduced amino acid sequences. The bootstrap values generated by 1000 sets of replications of tree construction are labeled at the internal nodes of the unrooted trees. The congruent clades in the two trees are labeled from A to F. The sensors IT, ITR, and ITRO and the regulators marked with NarL, NtrC, OmpR, and UC (unclassified) are shown.

speciation of the proteobacteria and Gram-positive bacteria.

(2) The NarL-like regulator genes classified in either Group I, Group II, Group III, or Group IV appeared to link to the corresponding genes of either IT-, ITR-, or ITRO-type sensors suggesting a different strategy from co-evolution. Instead, they are probably recruited components during evolution.

(3) Ten of eleven ITR-type hybrid sensors are orphans. An exception is PA1396, which is located next to a NarL-like regulator PA1397 in a divergently transcription orientation. The ITR-type sensor, also referred to as the hybrid sensor kinase, contains a regulator-like receiver domain following the input and transmitter domains. Interestingly, most of the ITRO-type sensors, which carried an additional Hpt (histidine-containing phosphotransfer) domain in comparison with that of the ITR-type sensors, are adjacent to a response regulator. It has been demonstrated that the phosphorelay specificity of the ITRO-type sensors, such as the BvgS of *Bordetella pertussis* and EvgS of *E. coli*, was determined by the Hpt domain (Perraud et al. 1998). The phosphorelay between the ITR-type kinases and the corresponding response regulators in *Vibrio harveyi* and *Saccharomyces cerevisiae* also occurred through Hpt modules, which are, however, encoded by genes located distantly from the 2CS genes (Freeman et al. 1999;

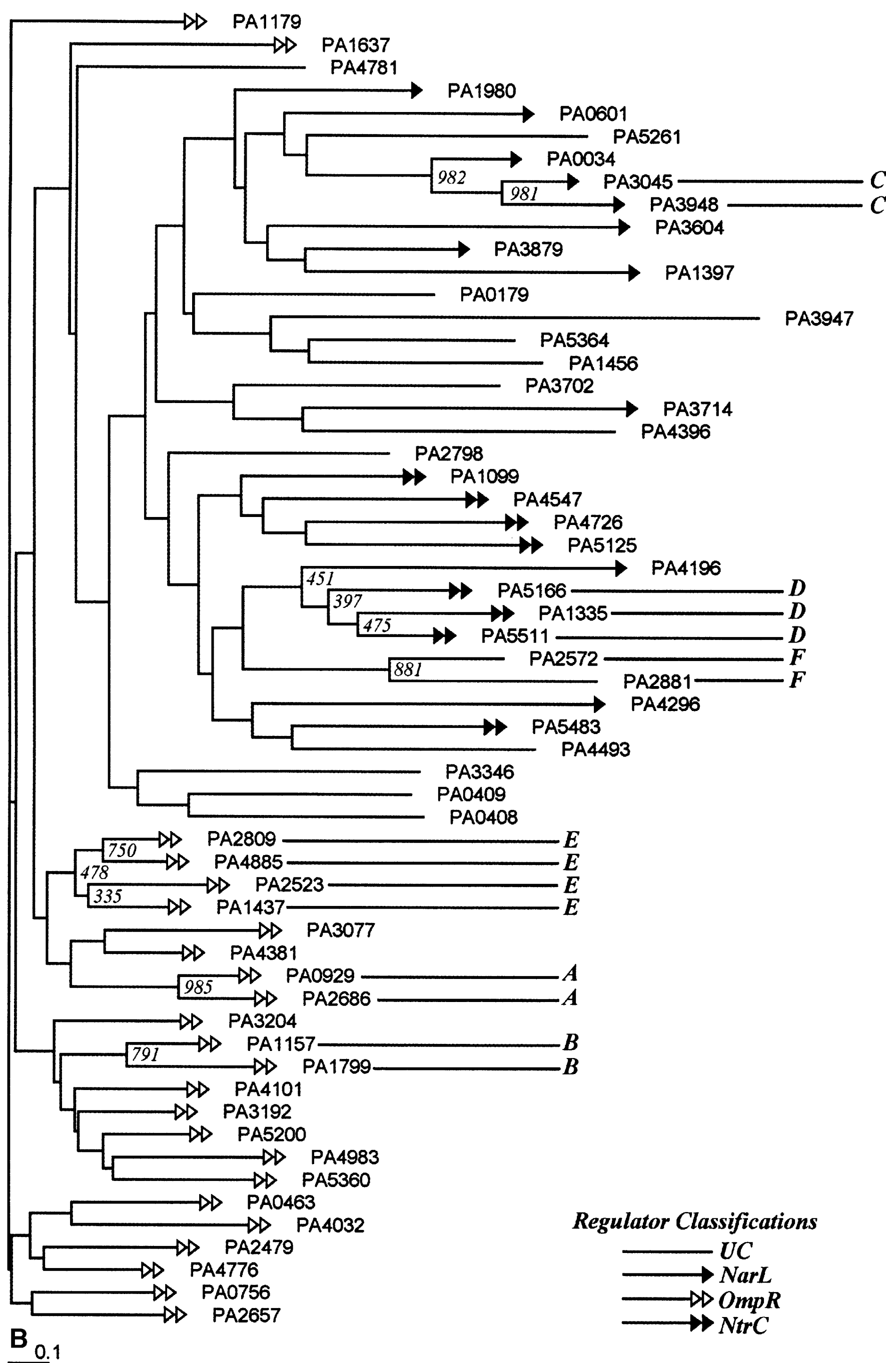



**Fig. 2.** Continued.

Posas et al. 1996). In the *P. aeruginosa* PAO1 genome, three such Hpt module-encoding genes have been identified (Rodrigue et al. 2000). It is likely that without a combined Hpt domain, the ITR sensors act in concert with the Hpt modules to perform the multiple-step phosphorelay in a manner similar to that of the ITRO systems. It has been proposed that a receiver or receiver-Hpt domain may be fused to an IT-type kinase to yield hybrid or unorthodox sensor kinases (Grebe and Stock 1999). Recruitment of these domains may confer on these systems an additional flexibility compared to the classical two-component signal transduction.

*Phylogenetic Analysis of the 2CS Genes*

The evolutionary relationships among the sensor and regulator genes were estimated by multiple sequence alignment of their deduced amino acid sequences using CLUSTAL W followed by the NJ method of tree construction. The 2CSs including four CheA-type sensors (PA0178, PA0413, PA1458, PA3704), five sensors (PA1396, PA3078, PA3878, PA4197, PA5262, PA0471) with extraordinary length, and one unclassified regulator (PA4843) were poorly aligned with the rest of sensors and thus were excluded from the NJ tree construction.

**Table 1.** The collective distance estimated for some of the 2CS pairs

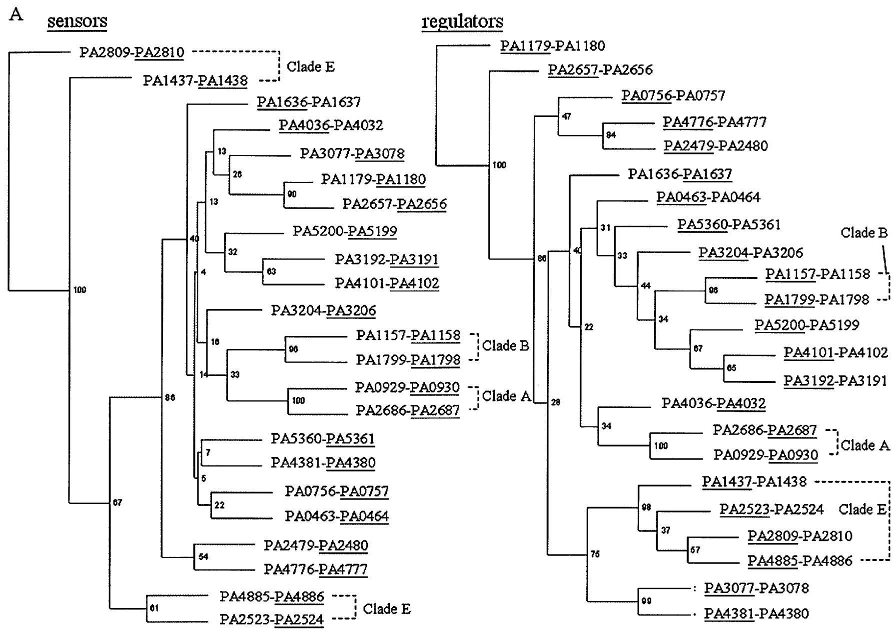
| $2CS_i$ | | $2CS_j$ | | Distance[a] | | |
|---|---|---|---|---|---|---|
| Sensor | Regulator | Sensor | Regulator | $S_{ij}$ | $R_{ij}$ | Clade/family |
| 0930 | 0929 | 2687 | 2686 | 0.81209 | 0.86742 | A |
| 1158 | 1157 | 1798 | 1799 | 1.78970 | 1.68556 | B |
| 3044 | 3045 | 3946 | 3948 | 1.10422 | 0.59288 | C |
| 1336 | 1335 | 5165 | 5166 | 1.38551 | 0.75003 | D |
| 5165 | 5166 | 5512 | 5511 | 1.54328 | 0.77573 | D |
| 1336 | 1335 | 5512 | 5511 | 1.50176 | 0.73861 | D |
| 1438 | 1437 | 2810 | 2809 | 1.37453 | 0.63785 | E |
| 2810 | 2809 | 4886 | 4885 | 1.47017 | 0.52720 | E |
| 2524 | 2523 | 2810 | 2809 | 1.64221 | 0.67896 | E |
| 2524 | 2523 | 4886 | 4885 | 1.65854 | 0.66942 | E |
| 1438 | 1437 | 4886 | 4885 | 1.77727 | 0.70514 | E |
| 1438 | 1437 | 2524 | 2523 | 1.85489 | 0.69369 | E |
| 2571 | 2572 | 2882 | 2881 | 1.32648 | 2.75433 | F |
| 0600 | 0601 | 1396 | 1397 | ND[b] | 2.57472 | NarL |
| 0600 | 0601 | 3878 | 3879 | ND* | 2.40411 | NarL |
| 0600 | 0601 | 3946 | 3948 | 19.73464 | 2.54159 | NarL |
| 0600 | 0601 | 4197 | 4196 | ND | 2.70931 | NarL |
| 0600 | 0601 | 4293 | 4296 | ND | 3.44983 | NarL |
| 1396 | 1397 | 1979 | 1980 | ND | 1.62793 | NarL |
| 1396 | 1397 | 4197 | 4196 | 10.11732 | 2.74051 | NarL |
| 1979 | 1980 | 3878 | 3879 | ND | 1.46125 | NarL |
| 1979 | 1980 | 3946 | 3948 | 18.40756 | 1.84689 | NarL |
| 1979 | 1980 | 4197 | 4196 | ND | 2.44646 | NarL |
| 1979 | 1980 | 4293 | 4296 | ND | 3.64490 | NarL |
| 3878 | 3879 | 4197 | 4196 | 20.78054 | 2.55514 | NarL |
| 3878 | 3879 | 4293 | 4296 | ND | 2.79356 | NarL |

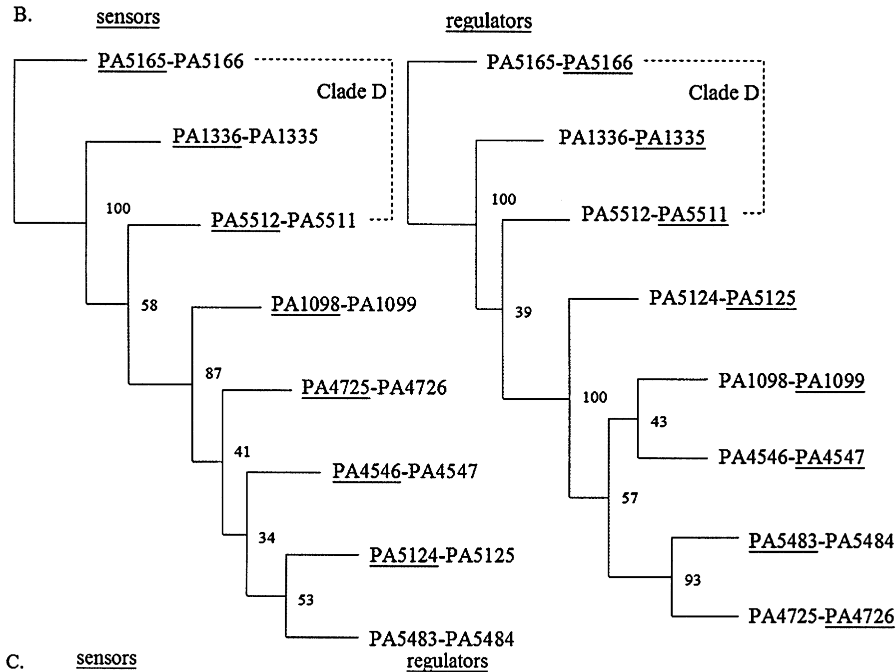
[a]The distance was calculated for each two 2CS pairs, in which $S_{ij}$ represents the distance between two sensors $S_i$ and $S_j$, and $R_{ij}$ is the distance between two regulators $R_i$ and $R_j$. $S_{ij}$ and $R_{ij}$ were calculated by PRODIST in Phylip using deduced amino acid sequences.
[b]Not determined. For the distances $S_{ij}$ or $R_{ij}$ exceeding the limitation in PRODIST, the distances are designated ND.

As shown in Fig. 2A, the ITR- and ITRO-type sensors apparently formed a group of subtrees except for branches h1 and h2. Their close association in the tree suggests that both ITR- and ITRO-type sensors share a common ancestor. However, only 4 ITRO- and 1 ITR-type sensors were observed in the 49 sensor–regulator pairs. The question of why the multistep phosphorelaying system is not favored in the bacteria remains to be answered.

Most OmpR-like and NtrC-like regulator encoding genes were found to form a cluster in the tree (Fig. 2B), whereas genes encoding the members of the NarL family and the unclassified regulators were scattered in the tree, indicating a lower sequence similarity. In order to analyze the historical associations between sensors and regulators, each node in the sensor tree was first assigned an association with the cognate regulator. Each terminal node in the sensor tree was therefore represented by its cognate regulator, while each of the internal nodes was represented by the union of the cognate regulators of all the descendants of that node. For each node in the sensor tree, the corresponding node in the regulator tree with the same descendent regulators was searched for. Subsequent to this, six clades (congruent monophyletic groups) composed of 11 sensor–regulator pairs, designated clades A to F, respectively, were identified (Fig. 2A and B). As shown in Table 1, the distances of the 2CS pairs in each clade calculated based on PRODIST were apparently shorter in contrast to those of the 2CS pairs containing NarL-like regulators. The 2CS pairs in each of the clades appeared to be the most closely related and most likely to be derived from a recent gene cluster-duplication event. This is coincident with the report of an extensive co-evolution of the 2CSs in 20 different genomes (Koretke et al. 2000).

In order to further assess the co-evolution relationships, ML estimation of the phylogeny was subsequently carried out for the specific groups of OmpR-, NtrC-, and NarL-like regulator-containing 2CSs (Fig. 3). As shown in Fig. 3A and B, the 2CS pairs of OmpR and NtrC families also form congruent clades, whereas the ML trees for the 2CS pairs of NarL-like regulators appear to show different topologies. In Fig. 3C, the PA1397 in the regulator tree is only one branch away from PA3879 (*narL*), however, their corresponding sensors PA1396 and PA3878 (*narX*) are distantly located from each other. This is supportive of the recruitment model that the



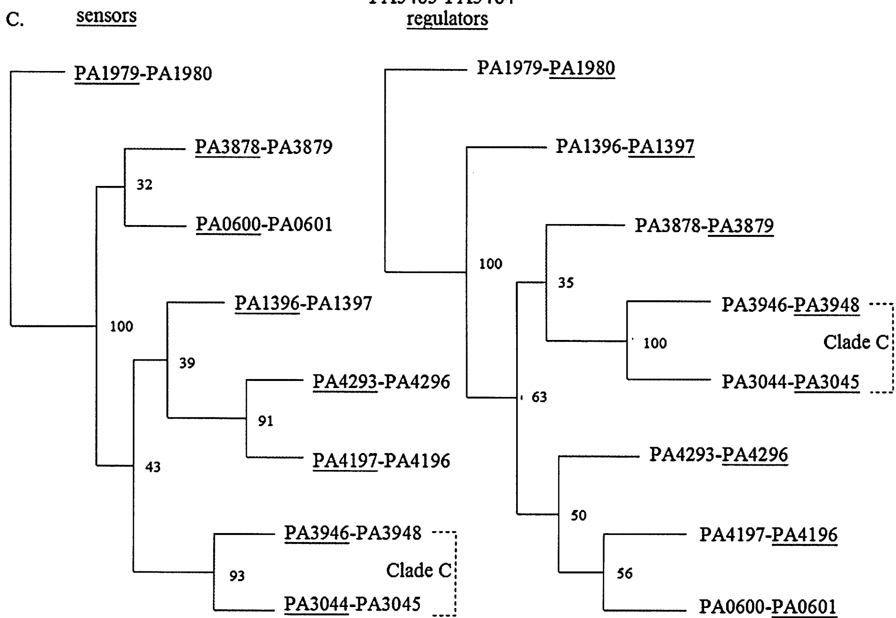
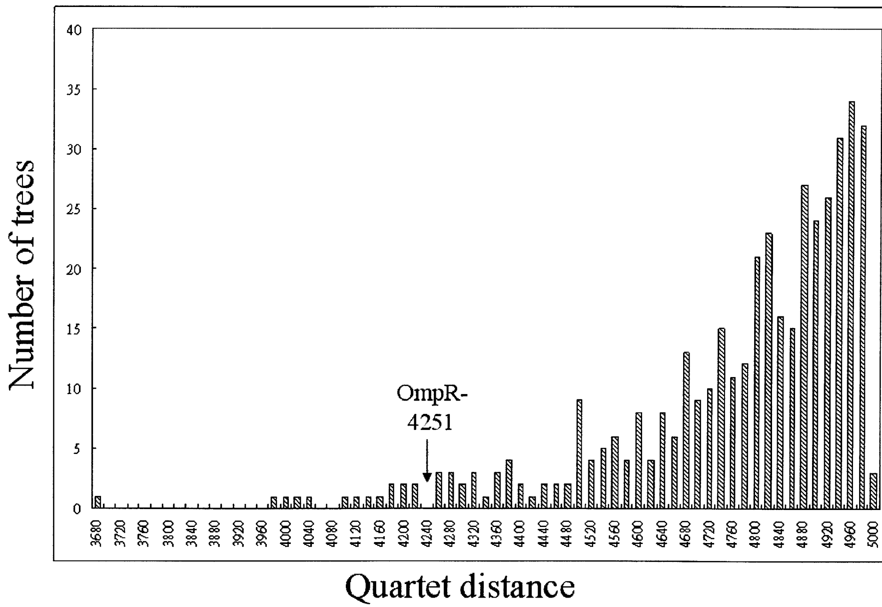




**Fig. 3.** The ML trees of the 2CSs carrying response regulators of the OmpR (**A**), NtrC (**B**), and NarL (**C**) types. The number shown at each node of the unrooted trees represents the bootstrap value of the operational taxonomic unit (OTU). The congruent clades in Fig. 2 are also indicated. The PA number of the OTU is shown with an underline and the PA number of the linked component is also marked.

A




B

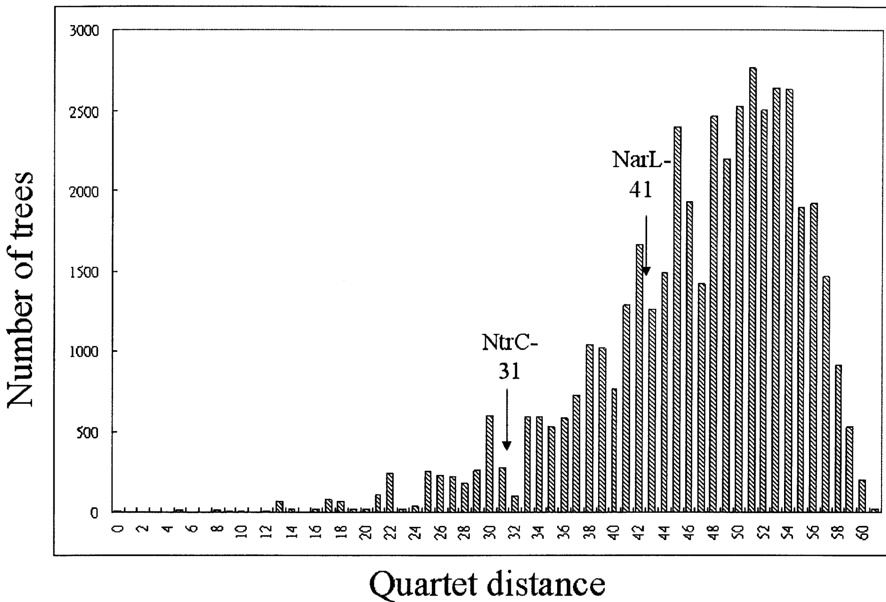



**Fig. 4.** The distribution of the dissimilarities of ML trees built with OmpR (**A**) and NtrC and NarL (**B**) groups measured by quartet metrics. The resolved and different quartets are scored in 300 random trees generated, respectively, with 23 and 8 OTUs using COMPONENT (Page 1993). The dissimilarities of the sensor and regulator trees are indicated for each of the groups.

2CS pairs are assembly products of a sensor and regulator. It is consistent with the distance-based analysis that the distances between these NarL-group 2CS pairs are relatively long or beyond determination (Table 1).

To measure the dissimilarities between the sensor and the regulator trees, the resolved and different quartets were determined (Estabrook et al. 1985). For trees of the 23 sensor and regulator pairs in the OmpR group, the quartet dissimilarity is 4251. For the eight sensor–regulator pairs in the NarL and NtrC groups, the values are 41 and 31, respectively. To compare the tree dissimilarities of different groups, random trees with the same number of OTUs are generated and, hence, the dissimilarities measured. Figure 3 shows the measurement of tree dissimilarities from the set of random trees. Fewer than 3.57 and 5.65% of the random trees have quartet dissimilarities smaller than the data for the OmpR and NtrC groups, respectively. In contrast, fewer than 19.49% of the random trees have quartet dissimilarities smaller than that calculated for the NarL group (Fig. 4). The results conclusively show that the tree congruencies of the OmpR and NtrC group are relatively higher than that of the NarL group.

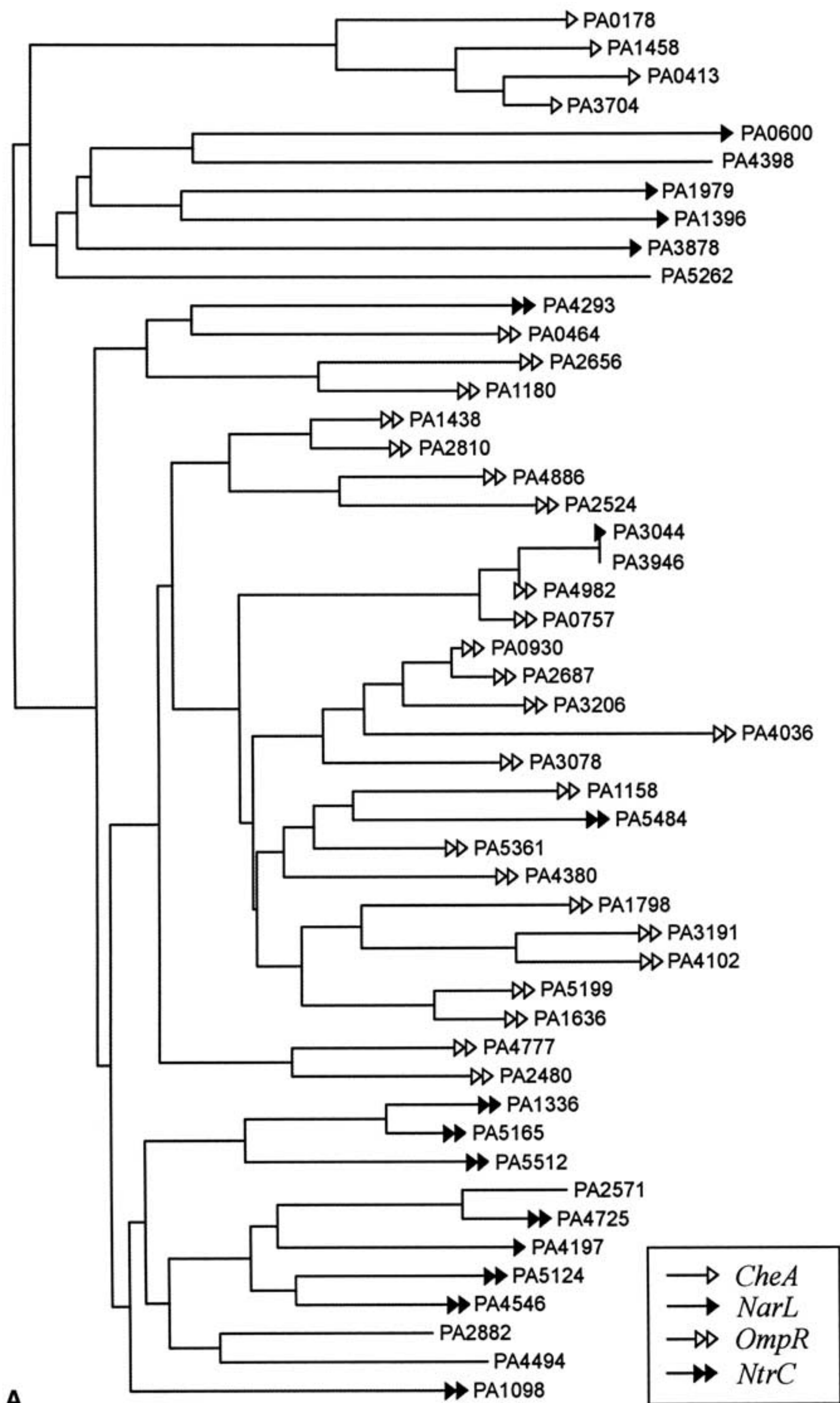



**Fig. 5.** Classification of sensor kinases by the sequence around the histidine. **A** Neighbor-joining tree built by the multiple sequence alignment of the sequences around the phosphorylated histidine residue using CLUSTAL W. The classifications of their cognate regulators, CheA-like, NarL-like, OmpR-like, and NtrC-like, are shown in the tree. **B** Graphical representations of the three major classes of kinases by multiple sequence alignments around the phosphorylated histidine residue. The amino acid symbols are shown at each position. The sequence conservation at each position is indicated by the overall height of the stack, while the relative frequency of each amino acid is indicated by the height of symbols within the stack.

### *Analysis of the Sequences Around the Phosphorelated Histidine of the Sensor Kinases*

It has been shown in the *B. subtilis* 2CSs that classification of the sequences flanking the histidine in the kinase could be correlated with their cognate regulators (OmpR, NarL, etc.) (Fabret et al. 1999). The sequences around the phosphorylated histidine residue were used to further classify the sensor kinases. We have found that the histidine-containing motifs could be classified into three homologous groups and their sequence logos are as shown in Fig. 5. The four CheY-type regulators were paired with the Class I kinases. Seven of nine NtrC-type regulators were paired with the Class II kinases. The two exceptions were PA5484, with a regulator-to-sensor transcrip-

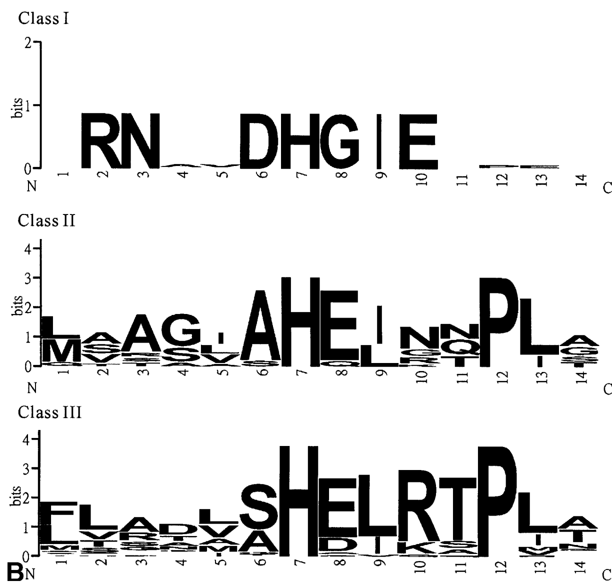



**Fig. 5.** Continued.

tion order, and PA4293, with two ORFs located in between the regulator and its cognate sensor. Most interstingly, all the OmpR-type regulators were paired with the kinases classified in Class III. Except for a few of the NarL-type regulators which were paired with the Class II and III sensors, most of the others were paired with the sensors showing low sequence similarity around the histidine residue. This is supporting evidence for the hypothesis that sensors paired with their cognate regulators of the NtrC or OmpR types have co-evolved as a unit from a common ancestor.

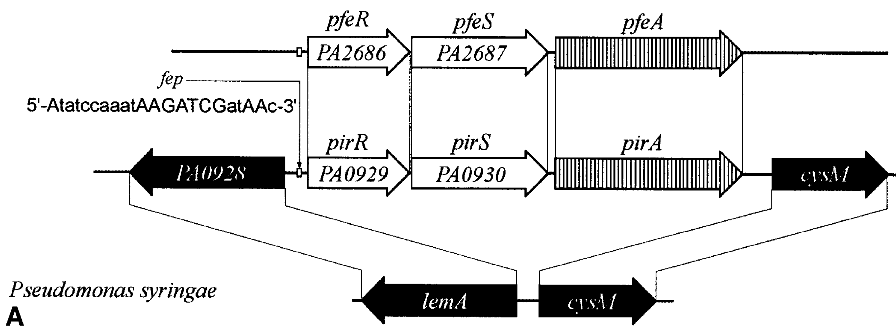
## *Functional Analysis of the Most Recently Duplicated 2CS Sensor–Regulator Pairs*

In order to assess functions of the 2CSs identified in each of the congruent clades, we compared these 2CSs with those of the known functions identified in other species and also their adjacent genes with the characterized properties. Several interesting findings were noted.

(1) The 2CS genes in one clade may contain a similar function. For instance, in clade A, the two 2CS gene pairs *pirS* (PA0930)/*pirR* (PA0929) and *pfeS* (PA2687)/*pfeR* (PA2686) are parts of the operons *pirRSA* and *pfeRSA*, respectively (Fig. 6A). Both operons encode siderophore-mediated iron uptake systems and are under the control of the Fur protein (Ochsner and Vasil 1996). This indicates that the paralogous groups continue to carry out a similar function after gene duplication. Functional redundancy is also seen in the members of clade C, PA3044/PA3045 and PA3946/PA3948, which are likely virulence-related 2CS paralogs (Fig. 6B). Both the 2CS gene pairs exhibit significant sequence homology with those of *Bordetella parapertussis bvgAS*, which has been demonstrated to participate in regulating the synthesis of many virulence factors (Bock and Gross 2001). Moreover, the regulator gene PA3947 has been reported to encode a homolog with a 45% sequence similarity to *Vibrio cholerae* virulence-related protein VieA (Lee et al. 1998) and to the regulator PvrR, which controls antibiotic susceptibility and biofilm formation in *P. aeruginosa* PA14 (Drenkard and Ausubel 2002). The clustering structure suggests a related function since gene clusters in a bacterial genome may possess the same function (Overbeek et al. 1999).

(2) Gene rearrangement may have occurred after duplication of the co-evolved 2CS gene cluster. The virulence-associated 2CS gene PA0928 (*lemA*) is adjacent to the *pirRSA* operon (Fig. 6A). In the *Pseudomonas syringae* genome, the *lemA* gene is clustered with the cysteine synthase encoding gene *cysM* (PA0932) in a divergently transcriptional direction, suggesting that the region has been subject to gene rearrangement during speciation of *P. syringae* and *P. aeruginosa*. PA0928, which resides relatively distant from the sensor PA0930 in the tree, appears to be recruited later by the 2CS pair PA0929/0930. As shown in Fig. 6B, the members of the 2CS pair PA3045 and PA3044 in clade C are cotranscribed. However, the PA3946 and PA3948 in this clade are transcribed divergently, which is probably indicative that the co-evolved 2CS gene clusters have undergone a rearrangement event after the gene duplication.

(3) A group of the functionally related 2CSs is probably required in controlling the translocation of metabolites and ions. As shown in Fig. 7A, all three gene pairs (PA1335/PA1336, PA5165/PA5166, and PA5511/PA5512) in clade D appear to be homologs of *Rhizobium meliloti* DctBD, which controls the transportation of C4-dicarboxylic acids in *R. meliloti* (Wang et al. 1989). The downstream genes clustered with PA5165/PA5166 are homologs of DctPQM (67, 47, and 72% sequence similarities) that are essential for transportation of the C4-dicarboxylates in *Rhodobacter capsulatus* (Shaw et al. 1991). Several genes encoding homologs of glutaminase–asparaginase (88% sequence similarity) of *Pseudomonas* 7A (Holcenberg et al. 1997), *E. coli* glutamyltranspeptidase (62% sequence similarity) (Suzuki et al. 1988), and *E. coli* glutamate–aspartate ABC transporters (>68% sequence similarities) (Oshima et al. 1996) were found upstream of the PA1335/PA1336. Moreover, a putative *S. typhimurium* amino acid permease-encoding gene was found adjacent to PA5511/PA5512. The regulator of the four 2CS gene pairs in clade E showed significant



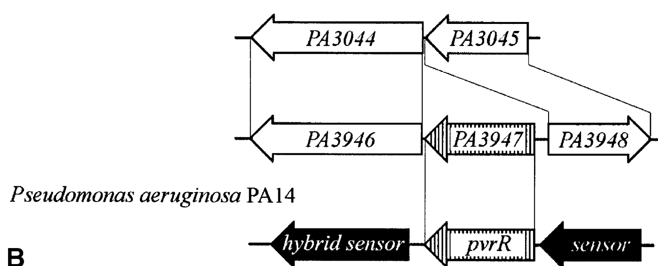




**Fig. 6.** **A** Schematic diagram of the gene organization of the *pfeRSA* and *pirRSA* gene clusters in the *P. aeruginosa* PAO1 genome. The thin likes linking genes show the regions with amino acid similarity, while arrows represent the transcriptional direction of genes. The PA numbers are labeled above the arrow for the 2CS genes. A *lemA–cysM* carrying segment of *P. syringae* is also shown below. The sequence and position of a putative *fep* binding site (Wang and Church 1992) identified in the upstream region of *pirRSA* are shown. **B** The two *bvgAS*-like gene clusters PA3044/PA3045 and PA3946/PA3947/PA3948 in the *P. aeruginosa* PAO1 genome (NA, NarL-like regulators; ITRO, unorthodox sensor kinase). A region of the *P. aeruginosa* PA14 genome containing the homolog of PA3947, *pvrR*, is shown below.

putative glutaminase-asparaginase
putative glutamyl-transpeptidase
putative glutamate-aspartate ABC transporters
PA1335
PA1336
ansB
ggt
homologs of DctPQM
PA5165
PA5166
PA5511
PA5512
putative amino-acid permease
A
PA1435
PA1436
PA1437
PA1438
PA2523
PA2524
PA2525
PA2526
PA2527
PA2528
czcA
czcB
czcC
czcR
czcS
PcopH
5'-cCAGCttAcgGAaATGtAATTACcTcgt-3'
PA2809
PA2810
PA2811
PA2812
PA4885
PA4886
PA4887
B

**Fig. 7.** Gene organization of the 2CSs in clade D and E. **A** The regions containing the 2CS genes in clade D in the *P. aeruginosa* PAO1 genome are shown. Arrows correspond to the transcription orientation of the genes. The PA numbers are labeled on the 2CS genes (open arrows). PA5165/PA5166 is adjacent to *dctPQM*, which encodes *R. capsulatus* C4-dicarboxylates transportation system homologs (Shaw et al. 1991). PA1336/PA1335 with the direct downstream genes of the *Pseudomonas 7A ansB* glutaminase–asparaginase homolog, *E. coli* γ-glutamyltranspeptidase, and four *E. coli* glutamate–aspartate ABC transporter homologs are shown. PA5512/PA5511 is shown adjacent to the gene homolog of the *S. typhimurium* putative amino acid permease. **B** Clusters of the known or putative small molecular transporter genes together with the four 2CSs in clade E are shown. The five genes in the *czc* gene cluster showed compelling similarities to the *R. eutropha czc* cation efflux system. A putative P*copH* transcription factor-binding site (Mills et al. 1994) identified at the upstream region of PA2523 is indicated. The hatched arrows represents nearby genes with known or predicted function involved in small molecular transport.

similarities (>74% sequence similarities) to the transcriptional regulator IrlR of *Burkholderia pseudomallei* (Jones et al. 1997) and CopR of *P. syringae* (Mills et al. 1994), which are related to Zn/Cd and Cu resistance, respectively. As shown in Fig. 7B, the genes PA2523/PA2524 are part of the *czcSRCBA* gene cluster, which have been reported as a Zn- and Cd-resistant determinant in *P. aeruginosa* and *R. eutropha* (Nies et al. 1995). Consolidation of all the above findings, together with the fact that the 2CS genes are part of the gene cluster with the functional similarity of being membrane-bound or small molecule-transporting proteins according to the Peudomonas Genome Project, suggests that even though divergent evolution is imminent, a conservation of function persists.

### *Comparative Promoter Analysis*

To elucidate whether the 2CSs which appeared to have resulted from gene pair duplication also shared the same transcriptional control, the upstream nontranslated sequence for each of the 2CS gene clusters of the six clades was collected and the sequences were searched using BLASTN against the DPInteract database in which *E. coli* transcription factor binding sites were collected (Robinson and Church 1994). In the iron transport operons, *pfeRS* and *pirRS*, of clade A, a *fep* signature (Wang and Church 1992) was identified in the upstream nontranslated region of the *pirR* (Fig. 6A). However, the signature was not found in that of the homolog *pfeR*, suggesting a different regulation for the two gene clusters. This finding is supported by the report that different levels of Fur proteins were required for regulatory control (Dean et al. 1996; Ochsner and Vasil 1996). Furthermore, a *P. syringae* CopR regulator binding site, P*copH* (Mills et al. 1994), identified upstream of the CopR gene homolog *czcR* is not found in the other members of clade E, which also suggested differential control of the gene expression for copper resistance (Fig. 7B).

We compared further the sequences with those of their homologous 2CSs in other *Pseudomonadaceae* including *P. aeruginosa*, *P. fluorescens*, *P putida*, *P. syringae*, and *Azotobacter vinelandii*. Surprisingly, sequence similarity was observed only when *P. aeruginosa* PA14 sequences were compared. The conserved sequences upstream of these 2CSs suggest a preserved response to regulatory signals in *P. aeruginosa*. Nevertheless, the degeneracy of the transcriptional factor binding sequences and the faster rate of accumulation of sequence variations in noncoding regions should be taken into consideration. It is conceivable that retaining functionally redundant genes is not economically favorable for the bacteria. The comparative promoter analysis indicated that the duplicated gene clusters have most likely evolved in order to perform similar functions under a different regulatory control.

### *GC%, GC3s, and $N_c$ of the 2CS Genes*

After analysis and comparison of the GC% and $N_c$ for the 2CS genes, two findings emerged. (i) None of the 123 2CS genes were found in regions with the typical low GC content of the PAO1 genome. (ii) No signs of recent horizontal gene transfer events were evident in comparison with the average GC% and codon usage. In view of these findings, and the fact that *P. aeruginosa* has a higher number of 2CS genes relative to its genome size, it is reasonable to speculate that this is the result of expansion of the bacterial genome and hence its ubiquity in nature.

In summary, we have analyzed the *P. aeruginosa* 2CS genes by the criteria of gene organization, functional motifs and sequence similarity determined by phylogenetic tree construction, and similarity measurements. Using these approaches, we have identified evidence that support both the co-evolution and the recruitment models for the evolution of 2CSs.

*Acknowledgments.* We thank Professor Fang-Rong Hsu, Taichung Healthcare and Management University, and Professors Yaw-Ling Lin and Yin-Te Tsai, Providence University, for their critical discussions during preparation of the manuscript. We are grateful also to Dr. Nandini Venkatesan for her critical reading of the manuscript. This work was supported in part by grant NSC91-2311-B-009-001 to Hwei-Ling Peng.

## References

Bock A, Gross R (2001) The BvgAS two-component system of *Bordetella* spp.: A versatile modulator of virulence gene expression. Int J Med Microbiol 291:119–130

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. URL http://weblogo.berkeley.edu

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci 23:324–328

Dean CR, Neshat S, Poole K (1996) PfeR, and enterobactin-responsive activator of ferric enterobactin receptor gene expression in *Pseudomonas aeruginosa*. 178:5361–5369

Drenkard E, Ausubel FM (2002) *Pseudomonas* biofilm formation and antibiotic resistance are linked to phenotypic variation. Nature 416:740–743

Estabrook GF, McMorris FR, Meacham CA (1985) Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst Zool 34:193–200

Fabret C, Feher VA, Hich JA (1999) Two-component signal transduction in *Bacillus subtilis*: How one organism sees its world. J Bacteriol 181:1975–1983

Felsenstein J (1993) PHYLIP 3.5 (phylogeny inference package) Department of Genetics, University of Washington, Seattle,

Freeman JA, Bassler B (1999) Sequence and function of LuxU: A two-component phosphorelay protein that regulates quorum sensing in *Vibrio harveyi*. J Bacteriol 181:899–906

Goldberg JB, Pier GB (2000) The role of the CFTR in susceptibility to *Pseudomonas aeruginosa* infections in cystic fibrosis. Trends Microbiol 8:514–520

Grebe TW, Stock JB (1999) The histidine protein kinase superfamily. Adv Microb Phys 41:139–227

Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98

Hoch J, Sihavy T (1995) Two-component signal transduction. ASM Press, Washington, DC

Holcenberg JS, Ericsson L, Roberts J (1997) Amino acid sequence of the diazooxonorleucine binding site of *Acinetobacter* and *Pseudomonas* 7A glutaminase–asparaginase enzymes. Biochemistry 17:411–417

Ishige K, Nagasawa S, Tokishita S, Mizuno T (1994) A novel device of bacterial signal transducers. EMBO J 13:5195–5202

Jones AL, Deshazer D, Woods DE (1997) Identification and characterization of a two-component regulatory system involved in invasion of eukaryotic cells and heavy-metal resistance in *Burkholderia pseudomallei*. Infect Infect 65:4972–4977

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acid Res 30:42–46

Kobayashi K, Shoji J, Shimizu T, Nakano K, Sato T, Kobayashi Y (1995) Analysis of a suppressor mutation *ssb* (*kinC*) of *sur0B20* (*spo0A*) mutation in *Bacillus subtilis* reveals that *kinC* encodes a histidine protein kinase. J Bacteriol 177:176–182

Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR (2000) Evolution of two-component signal transduction. Mol Biol Evol 17:1956–1970

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 17:1244–1245

Lee SH, Angelichio MJ, Mekalanos JJ, Camilli A (1998) Nucleotide sequence and spatiotemporal expression of the *Vibrio cholerae vieSAB* genes during infection. J Bacteriol 180:2298–2305

Macnab RM (1996) Flagella and motility. In: Neidhardt FC, et al. (eds) *Escherichia coli* and *Salmonella*, cellular and molecular biology. ASM Press, Washington DC, pp 123–145

Mills SD, Lim CK, Cooksey DA (1994) Purification and characterization of CopR, a transcriptional activator protein that binds to a conserved domain (cop box) in copper-inducible promoters of *Pseudomonas syringae*. Mol Gen Genet 244:341–351

Nies DH, Silver S (1995) Ion efflux systems involved in bacterial metal resistances. J Ind Microbial 14:186–199

Ochsner UA, Vasil ML (1996) Gene repression by the ferric uptake regulator in *Pseudomonas aeruginosa*: Cycle selection of iron-regulated genes. Proc Natl Acad Sci USA 93:4409–4414

Oshima T et al. (1996) A 718-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 12.7-28.0 min region on the linkage map. DNA Res 3:137–155

Overbeek R, et al. (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96:2896–2901

Page RD (1993) COMPONENT: Tree comparison software for Microsoft Windows, version 2.0 Natural History Museum, London

Page RD (1996) TreeView: An application to display phylogenetic trees on personal computers. Comput Appl Biosci 12:357–358

Parkinson JS (1993) Signal transduction schemes of bacteria. Cell 73:857–871

Peden JF (1999) Analysis of codon usage. PhD thesis. University of Nottingham. http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html

Perraud AL, Kimmel B, Weiss V, Gross R (1998) Specificity of the BvgAS and EvgAS phosphorelay is mediated by the C-terminal Hpt domains of the sensor proteins. Mol Micorbiol 27:875–887

Posas F, et al. (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

Robinson K, Church G (1994) A database on DNA-protein interactions (unpublished). http://arep.med.harvard.edu/dpinteract/

Rodrigue A, et al. (2000) Two-component systems in *Pseudomonas aeruginosa*: Why so many? Trends Microbiol 8:498–504

Shaw JG, Hamblin MJ, Kelly DJ (1991) Purification, characterization and nucleotide sequence of the periplasmic C4-dicarboxylate-binding protein (DctP) from *Rhodobacter capsulatus*. Mol Microbiol 5:3055–3062

Stock JB, Ninfa AJ, Stock AM (1989) Protein phosphorylation and regulation of adaptive responses in bacteria. Microbiol Rev 53:450–490

Stock JB, Stock AM, Mottonen JM (1990) Signal transduction in bacteria. Nature 344:395–400

Stover CK, et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 406:959–964

Suzuki H, Kumagai H, Echigo T, Tochikura T (1988) Molecular cloning of *Escherichia coli* K-12 *ggt* and rapid isolation of gamma-glutamyltranspeptidase. Biochem Biophys Res Commun 150:33–38

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Wang MX, Church GM (1992) A whole genome approach to in vivo DNA-protein interactions in *E. coli*. Nature 360:606–610

Wang YP, Birkenhead K, Boesten B, Manian S, O’Gara F (1989) Genetic analysis and regulation of the *Rhizobium meliloti* genes controlling C4-dicarboxylic acid transport. Gene 85:135–144

Wright F (1990) The ‘effective number of codons’ used in a gene. Gene 87:23–29