Subfamilies of *cpmA*, a gene involved in circadian output, have different evolutionary histories in cyanobacteria

Volodymyr Dvornyk^{1,2}

Correspondence

Volodymyr Dvornyk
vdrvornyk@kent.edu

¹Department of Biological Sciences, Kent State University, Kent, OH 44242-0001, USA

²Laboratory of Molecular Population Genetics and Evolution, M. G. Kholodny Institute of Botany, National Academy of Sciences of Ukraine, vul. Tereshchenkivska 2, Kiev, Ukraine

The *cpmA* gene mediates an output signal in the cyanobacterial circadian system. This gene and its homologues are evolutionarily old, and occur in some non-photosynthetic bacteria and archaea as well as in cyanobacteria. The gene has two functional domains that differ drastically in their level of polymorphism: the N-terminal domain is much more variable than the PurE homologous C-terminal domain. The phylogenetic tree of the *cpmA* homologues features four main clades (C1–C4), two of which (C1 and C3) belong to cyanobacteria. These cyanobacterial clades match respective ones in the previously reported phylogenetic trees of the other genes involved in the circadian system. The phylogenetic analysis suggested that the C3 subfamily, which comprises the genes from the cyanobacteria with the *kaiBC*-based circadian system, experienced a lateral transfer, probably from evolutionarily old proteobacteria about 1000 million years ago. The genes of this subfamily have a significantly higher nonsynonymous substitution rate than those of C1 (2.13×10^{-10} and 1.53×10^{-10} substitutions per nonsynonymous site per year, respectively). It appears that the functional and selective constraints of the *kaiABC*-based system have slowed down the rate of sequence evolution compared to the *cpmA* homologues of the *kaiBC*-based system. On the other hand, the differences in the mutation rates between the two cyanobacterial clades point to the different functional constraints of the systems with or without *kaiA*.

Received 1 August 2005

Revised 30 September 2005

Accepted 4 October 2005

INTRODUCTION

The endogenous circadian clock with a period of about 24 h is a key mechanism, which regulates a variety of physiological processes in accordance with daily changes in environmental conditions and thus helps organisms to adapt to the changing environment (Johnson & Golden, 1999).

Cyanobacteria are the only prokaryotes known so far to exhibit circadian rhythmicity (Ishiura *et al.*, 1998; Kondo & Ishiura, 2000). Their circadian system has been comprehensively studied in the unicellular *Synechococcus elongatus* PCC 7942 (Ditty *et al.*, 2003; Johnson & Golden, 1999). A core element of the system in this species is a cluster of three genes: *kaiA*, *kaiB* and *kaiC* (Ishiura *et al.*, 1998). In addition to the *kai* genes, several other genes have been shown to participate in the regulation of circadian rhythmicity, mainly by mediating input/output signals (Iwasaki *et al.*, 2000; Kutsuna *et al.*, 1998; Schmitz *et al.*, 2000; Tsinoiremas *et al.*, 1996).

Unlike *S. elongatus* PCC 7942, which possesses a single *kaiABC* cluster, other cyanobacterial species may have multiple copies of the *kaiB* and *kaiC* genes (Dvornyk *et al.*, 2002, 2003; Kaneko *et al.*, 2001). The *kaiA* gene may either occur in a single copy or be absent in some cyanobacteria (Dvornyk *et al.*, 2003; Nakamura *et al.*, 2003). Along with the data on the phylogenetic analysis of the *kaiB* and *kaiC* genes, this suggests that cyanobacteria probably have two types of the circadian system, *kaiABC*- and *kaiBC*-based, respectively (Dvornyk *et al.*, 2003). The *kaiBC*-based system perhaps also has components involved in input/output pathways, similar to those of the *kaiABC*-based system of *S. elongatus* PCC 7942. Based on the results of evolutionary analyses of the key components (*kaiB*, *kaiC* and *sasA* genes), it has been hypothesized that each type of cyanobacterial circadian system has specific functional and selective constraints, which shape its evolution and preclude lateral transfers of the components between the systems (Dvornyk *et al.*, 2003, 2004). A large proportion of these constraints are thought to be due to the emergence of the *kaiA* gene (Dvornyk *et al.*, 2004, 2005).

The *cpmA* gene (circadian *phase modifier*) was reported to be involved in an output pathway of the circadian clock

Abbreviations: AIC, Akaike information criterion; AIR, 5'-phosphoribosyl-5-amino-4-imidazole; LRT, likelihood ratio test; MYA, million years ago.

(Katayama *et al.*, 1999). Inactivation of *cpmA* dramatically affects activity of the *kaiA* promoter, but has no effect on the *kaiBC* promoter. These data suggest that, like *sasA*, *cpmA* may also have specific evolutionary constraints within each of the two types of the cyanobacterial circadian system, and that at least some of the constraints resulted from the appearance of *kaiA* in the course of the system's evolution. To test this hypothesis, I performed a comprehensive evolutionary and structural analysis of the *cpmA* genes from cyanobacteria.

METHODS

DNA and protein sequences. The annotated and homologous sequences of the *cpmA* genes were retrieved from the GenBank non-redundant database by using gapped BLASTP and PSI-BLAST (Altschul *et al.*, 1997) with the respective amino acid sequences of *S. elongatus* PCC 7942 (GenBank accession AAD29318.1) as a query. All sequences with bit score cutoff above 120 were considered homologues. The retrieved protein and nucleotide sequences were aligned using CLUSTALW (Thompson *et al.*, 1994) and manually adjusted based on structural considerations (Katayama *et al.*, 1999). Alignments of the nucleotide sequences were modified manually according to the respective amino acid alignments.

For the comparative phylogenetic analysis, I used sequences of the 16S rRNA gene and the RecA protein from the respective or closely related strains. The 16S rRNA gene is a common marker for evolutionary studies of cyanobacteria (Garcia-Pichel *et al.*, 2001; Honda *et al.*, 1999). RecA is encoded by a housekeeping gene, which was previously shown to be useful for phylogenetic reconstructions of bacteria (Lloyd & Sharp, 1993). The respective sequences were obtained from the public databases, aligned with CLUSTALW (Thompson *et al.*, 1994) and adjusted by visual inspection. A list of the sequences used is given in Table 1.

Phylogenetic analysis. Due to saturation, the rate of synonymous nucleotide substitutions was not estimated. The rate of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) was calculated using the modified Nei–Gojobori method (Nei & Gojobori, 1986) with the Jukes–Cantor correction for multiple substitutions at the same site and a transitions/transversions ratio of 1.2. The MEGA 3.1 software (Kumar *et al.*, 2004) was used for the computations of d_N .

To determine the DNA substitution model best fitting the data, the ModelTest 3.0 software was used (Posada & Crandall, 1998). Based on the results of this test, the Tamura–Nei model of substitutions (Tamura & Nei, 1993) with gamma distribution was employed for further phylogenetic analysis of the 16S rRNA genes.

The alternative topologies of the phylogenetic trees of the CpmA protein sequences and 16S rRNA nucleotide sequences were initially constructed using the neighbour-joining (NJ) algorithm (Saitou & Nei, 1987) and then compared using the Shimodaira–Hasegawa test (Shimodaira & Hasegawa, 1999) to find the best topology. The RecA tree was obtained using the NJ method and the Poisson correction distance assuming equal substitution rate among amino acid sites. The statistical support for the nodes of the trees was obtained by the interior branch test (Sitnikova *et al.*, 1995). The phylogenetic trees were built using the respective options of the MEGA 3.1 software (Kumar *et al.*, 2004). Branch lengths in the CpmA tree were estimated using maximum-likelihood (ML) approach and the WAG matrix of amino acid substitutions (Whelan & Goldman, 2001) with a gamma parameter. Because the ML test rejected the global clock hypothesis for

the CpmA proteins, the local clock model (Yoder & Yang, 2000) was applied to date their evolution. The analyses were performed using the PAML software (Yang, 1997).

To detect lateral transfers, the congruency of the obtained phylogenetic trees of species (16S rRNA tree) and the CpmA proteins was estimated using the likelihood-ratio test (LRT, Kishino & Hasegawa, 1989).

Analysis of selective constraints. The nonsynonymous/synonymous rate ratio, $\omega = d_N/d_S$, is commonly used as a measure of selective pressure at the protein level. Depending on its value ($\omega = 1$, < 1 and > 1) it may indicate neutral evolution, purifying selection and positive selection, respectively. The d_N/d_S ratio may either vary among amino acid sites but be averaged over all sequences (site-specific model) or differ between branches of a phylogenetic tree (branch-specific model). Comprehensive description of these models has been provided elsewhere (Nielsen & Yang, 1998; Yang *et al.*, 2000; Yang & Nielsen, 2002). All computations were performed with the PAML software (Yang, 1997). The nested models were compared pairwise by the LRT. According to the LRT, twice the log-likelihood difference, $2\Delta\ell = 2(\ell_1 - \ell_0)$, follows a χ^2 distribution with $df = p_1 - p_0$, where p is the number of free parameters in the model. To compare non-nested models, the Akaike information criterion was used: $AIC = -2\ell + 2p$, where p is the number of independent parameters in the model (Akaike, 1974). The model with the lowest AIC is considered the most appropriate.

RESULTS

Phylogeny of the *cpmA* genes

As revealed by the BLAST search, *cpmA* homologues, besides occurring in cyanobacteria, are also found in some other bacteria and archaea. This confirms the previously reported data about homologues of the cyanobacterial CpmA in some archaea (Katayama *et al.*, 1999). However, no CpmA homologues have been found either in the photosynthetic α -proteobacteria (i.e. *Rhodobacter*, *Rhodospirillum* and *Rhodopseudomonas*) or in *Chloroflexus*, which all are reported to have *kai* genes and/or a *kaiBC* operon laterally transferred from cyanobacteria (Dvornyk *et al.*, 2003). In the screened genomes, the *cpmA* gene usually occurred in a single copy. The only exception was *Trichodesmium erythraeum*, which contained two copies. Those were identical by their nucleotide sequence, but different in length: one gene is shorter by 90 nucleotides from the 3'-end (Table 1).

The 16S rRNA tree obtained featured three major clades, which corresponded to Cyanobacteria, Proteobacteria and Archaea (Fig. 1). Importantly, the cyanobacterial clade had 100 % statistical support. The topology of this clade was essentially in accordance with the conventional taxonomy of cyanobacteria (Castenholz, 1989), as well as with previously reported molecular data (Lyra *et al.*, 2001; Turner, 1997). It indicated that apparently no horizontal transfers of 16S rRNA genes have occurred between the cyanobacterial species studied.

The phylogenetic tree of the *cpmA* genes showed their separation into four well-defined clades (Fig. 1). Clade C1 contained the genes only from cyanobacteria. Clade C2 comprised the genes from proteobacteria, while clade C4

Table 1. Sequences used in the study

Species and strain	cpmA			16S rRNA			RecA		Designation of sequence in the figures
	Length (nt)	GenBank nucleotide sequence accession no.	Protein accession no.	Length (nt)	GenBank nucleotide sequence accession no.	Length (aa)	Protein accession no.		
Archaea									
Archaeoglobus fulgidus DSM 4304	690	NC_000917	NP_070103.1	1492	11497621	221	NP_070921.1	Archaeoglobus	
Methanothermobacter thermautotrophicus ΔH	753	NC_000916	NP_276407.1	1479	15678031	234	AAB86165.1	Methanothermobacter	
Methanococcoides burtonii DSM 6242	771	NZ_AADH01000021	ZP_00147513.1	1476	MB16SRNAG	225	ZP_00563456.1	Methanococcoides	
Methanocaldococcus jannaschii DSM 2661	771	NC_000909	NP_247133.1	1475	15668172	221	AAB98241.1	Methanocaldococcus	
Methanopyrus kandleri AV19	708	NC_003551	NP_613960.1	1512	NC_003551	316	NP_614315.1	Methanopyrus	
Methanosarcina acetivorans C2A	834	NC_003552	NP_615395.1	1430	NC_003552	267	NP_615048.1	M. acetivorans	
Methanosarcina barkeri Fusaro	807	NZ_AAAR01001960	ZP_00079120.1	1478	AF028692	243	ZP_00542441.1	M. barkeri	
Methanosarcina mazei Goel	813	NC_003901	NP_633640.1	1474	NC_003901	248	NP_633393.1	M. mazei	
Bacteria									
Thermotogae									
Thermotoga maritima MSB8	756	NC_000853.1	NP_228773.1	1559	15642775	356	NP_229655.1	Thermotoga	
Firmicutes									
Clostridium acetobutylicum ATCC 824	747	NC_003030	NP_347412.1	1510	NC_003030	349	NP_348440.1	Cl. acetobutylicum	
Clostridium thermocellum ATCC 27405	762	NZ_AABG02000056	ZP_00061531.1	1239	NZ_AABG030 00051	349	ZP_00504247.1	Cl. thermocellum	
Lactobacillus plantarum WCFS1	741	NC_004567	NP_783950.1	1571	NC_004567	380	NP_785779.1	Lactobacillus	
Moorella thermoacetica ATCC 39073	771	NZ_AADT02000009	ZP_00330044.1	1268	NZ_AADT020 00007	345	ZP_00576110.1	Moorella	
Staphylococcus epidermidis ATCC 12228	777	NC_004461	NP_763927.1	1554	NC_004461	349	NP_764518.1	Staphylococcus	
Thermoanaerobacter tengcongensis MB4T	762	NC_003869	NP_622425.1	1527	NC_003869	307	NP_622992.1	Thermoanaerobacter	
Proteobacteria									
Desulfotalea psychrophila LSv54	765	CR522870	CAG35536.1	1492	AF099062	355	YP_066694.1	Desulfotalea	
Desulfovibrio desulfuricans G20	1029	NZ_AABN02000022	ZP_00131082.1	1476	AF098671	351	Dde2373*	Desulfovibrio	
Magnetococcus sp. MC-1	771	NZ_AAAN01000179	ZP_00043977.1	1353	NZ_AAAN020 00028	390	EAN28147.1	Magnetococcus	
Myxococcus xanthus	753	AF449411	AAL56587.1	1515	MXA233929	342	P48291	Myxococcus	
Wolinella succinogenes DSM 1740	747	NC_005090-1	NP_906400.1	1481	WOLRR16SA	350	NP_907645.1	Wolinella	
Cyanobacteria									
Anabaena sp. PCC 7120	789	NC_003272	NP_487925.1	1489	17131110	357	BAB74971.1	Anabaena	
Anabaena variabilis ATCC 29413	789	NZ_AAEA01000013	ZP_00159800.1	1439	AB016520	358	YP_325417.1	A. variabilis	
Crocospaera watsonii WH 8501	795	NZ_AADV01000074	ZP_00175911	1402	NZ_AADV010 00003	358	ZP_00519282.1	Crocospaera	
Gloeobacter violaceus PCC 7421	819	NC_005125	NP_924408.1	1407	AF132790	369	NP_926362.1	Gloeobacter	
Nostoc punctiforme PCC 73102	789	NZ_AAAY02000058	ZP_00107377.1	1410	AF027655	359	ZP_00105721.1	Nostoc	
Prochlorococcus marinus CCMP 1375	672	NC_005042	NP_875743.1	1465	NC_005042	379	NP_876107.1	Prochlorococcus CCMP	
Prochlorococcus marinus MED4	657	NC_005072	NP_893395.1	1147	AF001466	370	NP_893679.1	Prochlorococcus MED	
Prochlorococcus marinus MIT 9313P	684	NC_005071	NP_894191.1	1400	AF053399	379	CAE21902.1	Prochlorococcus MIT	
Synechococcus elongatus PCC 7942	783	AF117208	AAD29318.1	1411	AF132930	356	ZP_00163563.2	Synechococcus PCC	
Synechococcus sp. WH 8102	651	NC_005070	NP_897701.1	1464	NC_005070	375	CAE08577.1	Synechococcus WH	
Synechocystis sp. PCC 6803	840	NC_000911	NP_442897.1	1489	1653715	354	NP_443045.	Synechocystis	
Thermosynechococcus elongatus BP-1	762	NC_004113	NP_681979.1	1491	AP005376	355	BAC09641.1	Thermosynechococcus	
Trichodesmium erythraeum IMS101	786	NZ_AABK02000013	ZP_00072207.1	1438	AF013030	360	ZP_00672580.1	Trichodesmium	
	696	NZ_AABK02000151	ZP_00075135.1						

*Joint Genome Institute draft genome sequence (version as of 6 June 2005).

Fig. 1. Congruent ML tree of the CpmA homologues and 16S rRNA genes. Interior-branch test values <50% are not shown. Branches *a* and *b*, for which positive selection was tested, are in bold. For the designations of the genes, see Table 1.

included those mainly from methanogenic archaea. The phylogenetic position of clade C3 deserves particular consideration for two important reasons. First, it definitely does not match the taxonomic position of the respective cyanobacterial species (Turner, 1997) and their positions in the 16S rRNA tree (Fig. 1). Nor does it correspond to the position of these species in the RecA phylogeny (Fig. 2). The most obvious explanation of the incongruence is lateral transfer from some early branching non-photosynthetic bacteria, as the tree topology suggests. This assumption is supported by the results of the Kishino–Hasegawa test (Kishino & Hasegawa, 1989), which strongly rejected the null hypothesis of 16S rRNA and CpmA tree compatibility ($P < 0.001$) using either the 16S rRNA or the CpmA sequence data and with all alternative trees.

Second, clades C1 and C3 have essentially the same topology as do the subtrees of the *kaiBC* operons and the *sasA* genes from the respective species (Dvornyk *et al.*, 2004). Clade C1 includes the *cpmA* genes from the species with all three *kai* genes. The only exception is *Gloeobacter violaceus*, which has no *kai* genes (Nakamura *et al.*, 2003). This incongruence, when analysed against the topologies of the CpmA tree and the 16S rRNA and RecA subtrees of cyanobacteria (Figs 1 and 2), suggests that the *cpmA* gene of *G. violaceus* was laterally transferred from one of the evolutionarily younger species. Clade C3 contains *cpmA* from cyanobacteria, which possess either all three clock genes (*Synechococcus* sp. WH 8102) or only two of them, *kaiB* and *kaiC* (three *Prochlorococcus* strains). The topology of this clade is also

characteristic of the other components of the circadian system (i.e. *kaiBC*, *sasA* and *ldpA*; Dvornyk *et al.*, 2004). Along with data from the phylogenetic analyses of 16S rRNA (Fig. 1) and RecA (Fig. 2), this suggests that the *kaiA* gene was laterally transferred to *Synechococcus* sp. WH 8102 from some other cyanobacteria. However, the taxa from which the transfer occurred can not be determined due to the limited data available.

Structure, polymorphism and mutation rates of the *cpmA* genes

The length of the CpmA protein homologues varied from 216 (*Synechococcus* sp. WH 8102) to 333 (*Geobacter metallireducens*) amino acid residues. According to the NCBI Conserved Domain Database (Marchler-Bauer *et al.*, 2003), CpmA belongs to COG1691, the superfamily of NCAIR mutase (PurE)-related proteins, which catalyse the carboxylation of 5'-phosphoribosyl-5-amino-4-imidazole (AIR) to 5'-phosphoribosyl-5-aminoimidazole-4-carboxylic acid in purine biosynthesis (Meyer *et al.*, 1992; Tiedeman *et al.*, 1989; Watanabe *et al.*, 1989). However, this classification is based on the general function prediction only. An analysis of the *cpmA* nucleotide polymorphism suggested that the gene probably consists of two functional domains. Each of them occupies about half of the gene. The N-terminal domain of the encoded protein is 85–203 amino acid residues long (corresponding to residues 1–120 of the *S. elongatus* PCC 7942 CpmA sequence). The C-terminal domain is less variable in length and spans 131–158 residues.

Fig. 2. Phylogeny of the RecA proteins from the studied species as inferred by the NJ method. Interior-branch test values <50% are not shown.

It contains two hydrophobic cores (Katayama *et al.*, 1999). This domain is homologous to the N-terminal segments of the other related proteins, AIR carboxylase (pfam00731) and NCAIR mutase (COG0041), though this homology is weak.

The two domains of the *cpmA* gene showed striking differences in the level of nucleotide polymorphism (Table 2). The C-terminal domain was much more highly conserved than the N-terminal one in all four major clades. The genes of C3 had a lower mean d_N than those of C2. To test whether this difference is due to the more stringent evolutionary constraints in C3, the age of the respective nodes and the nonsynonymous substitution rate were estimated. The time of *S. elongatus* PCC 7942 speciation, about 1000 MYA (Dvornyk *et al.*, 2003), was used as a calibration point (node 1, Fig. 1). Based on this, the time of nodes 2, 3 and 4

was estimated to be 1207.75 ± 1.16 , 1040.35 ± 2.54 and 2018.04 ± 7.74 MYA, respectively. Importantly, the time estimates for the nodes 1 and 3 were similar; this was in general agreement with the 16S rRNA tree, which suggested a close time of origin for the genera *Synechococcus* and *Prochlorococcus*. The respective mutation rates for non-synonymous sites are given in Table 2. As the data suggest, the *cpmA* homologues from subfamilies C2 and C3 have the highest mutation rate, while the genes from two other subfamilies have a much lower one. The mutation rates in C2 and C3 are similar, so the differences in their d_N estimates are likely to be due to the different evolutionary ages of the subfamilies.

On the other hand, the mutation rate in C1 is 35–46% lower than that in either C2 or C3. Given that the *cpmA* genes of this subfamily have at least one more important function,

Table 2. Patterns of nonsynonymous nucleotide substitutions (d_N) in the different regions of the *cpmA* genes

Clade	Domain				Mean over gene	μ
	N-terminal	μ^*	C-terminal	μ		
C1	0.422 ± 0.037	2.11	0.151 ± 0.026	0.76	0.305 ± 0.024	1.53
C2	0.686 ± 0.054	2.84	0.274 ± 0.038	1.13	0.494 ± 0.032	2.04
C3	0.640 ± 0.037	3.08	0.256 ± 0.040	1.23	0.442 ± 0.037	2.13
C4	0.852 ± 0.066	2.11	0.391 ± 0.040	0.97	0.623 ± 0.037	1.54
Total	0.869 ± 0.061		0.396 ± 0.045		0.622 ± 0.038	

* $\times 10^{-10}$ substitutions per nonsynonymous site per year.

regulation of the *kaiA* promoter (Katayama *et al.*, 1999), as compared to C2 and C3, the lower mutation rate may be due to the evolutionary constraints related to this function.

Selective constraints in evolution of *cpmA* genes

The site-specific likelihood models assume variable selective pressures among sites but no variation among branches of a phylogenetic tree (Nielsen & Yang, 1998; Yang *et al.*, 2000). The parameter estimates and log-likelihood values for these models are given in Table 3. Among the discrete models, M3 fits the data significantly better than M0, M1 or M2, according to the LRT (for nested models) or AIC criteria (for nonnested comparisons: M1/M3, M2/M3). This suggests heterogeneous selective pressures along whole *cpmA* sequences. However, ω values smaller than 1 indicate no positive selection on the whole set of the *cpmA* genes (Table 3). Among the continuous distribution models, model M8 was significantly better than M7 ($P < 0.0001$). However, the mean ω value estimated under the best model (M8) is 0.066, and the proportion of sites under positive selection $p_1 = 0$, again implying strong purifying selection. These data favour the notion that, for the whole dataset, the main force acting is negative selection, with evidence for heterogeneous selection rates on different codons.

To test variation of the ω ratio among lineages, I proceeded from the hypothesis that selective constraints of the *cpmA* gene in the *kaiABC*-based circadian system might be influenced by the appearance of the *kaiA* gene and subsequent acquisition by *cpmA* of a new function, namely regulation of the *kaiA* promoter activity (Katayama *et al.*, 1999). This might result in positive selection along branch *a* (Fig. 1). On the other hand, the lateral transfer of the *cpmA* gene from non-photosynthetic proteobacteria to photosynthetic

cyanobacteria, if resulting in an acquisition of a new function by the gene, might also be followed by the period of positive selection (along branch *b*). I compared the log likelihood values under two models: the one-ratio model (M0 in the notation of Goldman & Yang, 1994), which assumes the same ω parameter for the entire tree, and the branch-specific-ratio model, which assumes the same d_N/d_S ratio (ω_0) for the background lineages, while specifying either the same ($\omega_1 = \omega_a = \omega_b$, model 1) or different ($\omega_a \neq \omega_b$, model 2) ratios for lineages *a* and *b* (Fig. 1). The likelihoods of both models ($\ell_1 = -22132.94$ and $\ell_2 = -22132.56$) were significantly better ($P < 0.0001$) than that of M0 ($\ell_0 = -22167.19$), suggesting rejection of the one-ratio model and, consequently, different d_N/d_S ratios along the lineages *a* and *b* as compared to the background ratio. The LRT statistic for the comparison of models 1 and 2 is $2\Delta\ell = 2 \times 0.38 = 0.76$, with $P = 0.38$ and $df = 1$. It indicates that model 2 is not significantly better than model 1 and thus the ω ratios for lineages *a* and *b* are similar.

To test whether ω_1 is significantly higher than 1, the log likelihood value was calculated under the two-ratios model but with $\omega_1 = 1$ fixed, giving the log likelihood value -22133.73 . The two-ratios model that does not place the constraint on ω_1 is not significantly better ($\ell_1 = -22132.94$); the test statistic is $2\Delta\ell = 2 \times 0.79 = 1.58$, with $P = 0.21$ and $df = 1$. So ω_1 is not significantly greater than 1 at the 5 % significance level (see Yang, 1998). However, this model averages the ω value along the sequence. Therefore, I further applied Yang and Nielsen's branch-site model B (Yang & Nielsen, 2002), which is an extension of the M3 model and potentially allows for positive selection at some sites in both background and foreground lineages. This model suggested no sites under positive selection in the background branches ($\hat{\omega} = 0.085$) and about 8 % positively selected sites in the

Table 3. Parameter estimates and log-likelihood values under different models of variable ω ratios among sites

p is the number of free parameters for the ω ratios, d_N/d_S = mean over sites. Parameters indicating positive selection are given in bold. For the detailed description of the models, see Nielsen & Yang (1998) and Yang *et al.* (2000).

Model	<i>p</i>	ℓ	AIC	ω	Parameters
M0: one ratio	1	-22167.19	-44336.38	0.046	$\hat{\omega} = 0.046$
Site-specific models					
M1: neutral	1	-23390.03	-46782.06	0.821	$p_0 = 0.179$, $\hat{\omega}_0 = 0$, $p_1 = 0.821$, $\hat{\omega}_1 = 1$
M2: selection	3	-23205.91	-46405.82	0.948	$p_0 = 0.174$, $\hat{\omega}_0 = 0$, $p_1 = 0.696$, $\hat{\omega}_1 = 1$, $p_2 = 0.130$, $\hat{\omega}_2 = 1.948$
M3: discrete ($K = 3$)	5	-21251.96	-42493.92	0.107	$p_0 = 0.274$, $\hat{\omega}_0 = 0.005$, $p_1 = 0.340$, $\hat{\omega}_1 = 0.044$, $p_2 = 0.386$, $\hat{\omega}_2 = 0.121$
M7: β	2	-28151.18	-56298.36	0.093	$p = 0.109$, $q = 1.021$
M8: β and ω	4	-21221.93	-42435.86	0.066	$p_0 = 1.000$, $p = 0.703$, $q = 9.566$, $p_1 = 0.000$, $\hat{\omega} = 7.725$
Branch-site model					
Model B	5	-21333.31	-42656.62		$p_0 = 0.323$, $\hat{\omega}_0 = 0.009$, $p_1 = 0.596$, $\hat{\omega} = 0.085$, ($p_2 + p_3 = 0.081$), $\hat{\omega}_2 = 35.182$

foreground branches ($\hat{\omega}=35.182$). However, model B yielded the log likelihood value -21333.31 , which does not fit the data better than the site-specific model M3 (Table 3). This suggests that, although d_N/d_S ratios vary along the different branches as well as among codons, they are probably not above 1 and thus do not indicate any positive selection.

DISCUSSION

The results of the phylogenetic analyses suggest that the *cpmA* homologues are evolutionarily old. The observed homology of CpmA to the PurE and AIR carboxylase proteins implies their probable common evolutionary ancestry. However, this homology is weak, and one could assume that, due to the large divergence, CpmA does not necessarily perform functions similar to those of PurE and AIR carboxylase. Moreover, as compared to the *purE* genes, which are ubiquitous in prokaryotes, the *cpmA* homologues occur in far fewer species. This may also suggest some alteration of their functions. However, lack of comprehensive molecular data about CpmA makes it impossible to infer these functions with reasonable accuracy.

Evolution of the *cpmA* genes of cyanobacteria has involved at least one influential horizontal transfer from proteobacteria, most likely from the early branching ones (Fig. 1). The fact that clade C3 is monophyletic suggests that transfer of this subfamily probably occurred before the speciation of *Prochlorococcus*. Based on the results of the molecular clock analysis, the time of the transfer may be estimated as about 1000 MYA. This is close to the time of origin of two other circadian-related genes, *kaiA* and *ldpA* (Dvornyk *et al.*, 2003; Dvornyk, 2005). This correspondence is probably not accidental and may be in favour of the assumption that the most important events resulting in the 'evolutionary upgrade' of the *kaiBC*-based system to the *kaiABC*-based one (e.g. appearance of *kaiA*, *ldpA* and maybe some other as yet unknown genes) occurred about 1000 MYA (Dvornyk *et al.*, 2003).

Horizontal (lateral) gene transfer is important in the evolution and adaptation of prokaryotes (Gogarten *et al.*, 2002; Koonin *et al.*, 2001). Recent data showed that, in the evolution of the cyanobacterial circadian system, lateral transfers of its various elements have been a common event (Dvornyk *et al.*, 2003, 2004; Dvornyk & Nevo, 2004). However, except for a few documented transfers of the *kaiB* and *kaiC* genes from cyanobacteria to proteobacteria (Dvornyk *et al.*, 2003), the other transfers have occurred only between cyanobacteria with the same type of system, either the *kaiABC*- or the *kaiBC*-based one (Dvornyk *et al.*, 2004; Dvornyk, 2005).

The phylogenetic trees of all the components of the cyanobacterial circadian system studied so far (e.g. the *kaiBC* operon, *sasA*, *ldpA* and *cpmA*) have featured two clades (corresponding to C1 and C3 in Fig. 1). While the topology of each clade may vary, a set of the species within the clade is always the same. That is, none of the species of clade C3

ever appears in C1, and vice versa. The observed major congruence between the phylogenies of the various components of the circadian system (e.g. the *kaiBC* operon, *sasA*, *ldpA* and *cpmA*) has several important implications. First, it indicates that evolution of the genes involved in circadian control is probably system-specific, i.e. the genes within each of the two types of the cyanobacterial circadian system evolve concordantly. Second, each type of circadian system has specific functional and selective constraints, which preclude horizontal transfers of the components between the systems. The only exception to this rule is *Synechococcus* sp. WH 8102, which has all three *kai* genes. This might suggest that its circadian system is the *kaiABC*-based one. However, phylogenetic analyses of all currently known components of the system unambiguously place this species together with *Prochlorococcus*, which has the *kaiBC*-based system. Based on the analyses of the 16S rRNA (Fig. 1) and RecA (Fig. 2) sequences, *Synechococcus* sp. WH 8102 is phylogenetically positioned between *S. elongatus* PCC 7942 and *Prochlorococcus*. The tree topologies of the different circadian components may correspond merely because the 16S rRNA and RecA trees are similar and the rate of lateral transfer is low. However, this is unlikely in the case being considered. As mentioned above, the phylogenetic trees of the elements of the cyanobacterial circadian system studied so far are not fully congruent within the two main clades (Dvornyk *et al.*, 2003, 2004). The incongruence may be due to a number of analytical and biological factors (Rokas *et al.*, 2003), e.g. the choice of optimality criterion (Huelsenbeck, 1995), limited data availability (Cummings *et al.*, 1995), taxon sampling (Graybeal, 1998), specific assumptions in the modelling of sequence evolution (Yang *et al.*, 1994), natural selection or genetic drift (Maddison, 1997; Martin & Burg, 2002), and others. Among the possible reasons, horizontal transfers of the *cpmA* homologues seem to be the most feasible explanation of the observed incongruence. Apparently, horizontal transfers have been rather common between the species of the same clade. For example, multiple transfers of the *kaiBC* operon between closely related filamentous cyanobacteria in the Nostocaceae were recently reported (Dvornyk & Nevo, 2004). As follows from the results of the molecular phylogenetic analyses using small ribosomal subunit RNA (Turner, 1997), 16S rRNA (Fig. 1) and RecA (Fig. 2), the unicellular cyanobacteria *S. elongatus* PCC 7942 and *Prochlorococcus* are close relatives and thus clustered together. This might suggest that they should either have the same type of circadian system or experience horizontal transfers of some its components. However, in the phylogenies of all known circadian-related genes, *S. elongatus* PCC 7942 and *Prochlorococcus* sp. always appear in clearly separated clades (Dvornyk *et al.*, 2004; Dvornyk, 2005; Dvornyk & Knudsen, 2005). This means that the cyanobacterial system of *S. elongatus* PCC 7942 is different from that of *Prochlorococcus* sp. and is closer to the system of more phylogenetically distant filamentous cyanobacteria (*Nostoc*, *Anabaena*). So far, no apparent horizontal transfers between the different systems have been detected, except for the *kaiA* gene. Furthermore, the significant differences in

substitution rates between the components of the two systems (Dvornyk *et al.*, 2004) are in favour of the assumption that the functional constraints are indeed system-specific.

In *S. elongatus* PCC 7942, *cpmA* was shown to modify the phasing and amplitude of class 1 rhythms (Katayama *et al.*, 1999), which have their circadian peak near subjective dusk and their trough at subjective dawn (Liu *et al.*, 1995). In particular, the disruption of *cpmA* significantly affects expression of *psbAI* and *psbAII* (Katayama *et al.*, 1999), which encode two forms of the photosystem II reaction centre D1 protein (Golden *et al.*, 1986; Schaefer & Golden, 1989). Given that among the prokaryotes with the *cpmA* homologues (Table 1) only cyanobacteria are photosynthetic, the genes of subfamilies C1 and C3 (Fig. 1) might be expected to perform similar functions related to photosynthesis. On the other hand, an important role of *cpmA* in the *kaiABC*-based circadian system is regulation of the *kaiA* promoter activity (Katayama *et al.*, 1999). In this system, *kaiA* is critical for circadian oscillation, because it functions as an activator of the *kaiBC* promoter (Ishiura *et al.*, 1998). The exact mechanism of the effect of *cpmA* on circadian-like expression of the photosynthesis-related genes is still unknown. It may involve *kaiA* as an essential element. In such a case, in the *kaiBC*-based system lacking *kaiA*, the relevant function of *cpmA* is abolished and thus the genes of C3 seem not to be associated with the circadian system. However, if control of circadian-dependent expression of the photosynthesis-related genes by *cpmA* employs a different mechanism, without *kaiA*, and this mechanism is available in all cyanobacteria, another scenario is possible. It may imply that the lower mutation rate in the C1 subfamily is due to the constraints conferred by the interaction with the *kaiA* promoter, but in the rest, the genes of C1 and C3 perform essentially the same circadian-associated functions. While the former is probable, the latter seems more unlikely than the previous scenario. Indeed, as mentioned above, the lateral transfer of the C3 subfamily to cyanobacteria occurred about 1000 MYA, i.e. after the *kaiB* and *kaiC* genes formed an operon and joined into the circadian protosystem that happened between 3500 and 2320 MYA (Dvornyk *et al.*, 2003), and after photosynthesis originated (Xiong *et al.*, 2000). Moreover, the *cpmA* genes were probably transferred from evolutionarily old non-photosynthetic proteobacteria, i.e. *a priori* were not involved in regulation of photosynthesis. If the transferred *cpmA* genes were to acquire a new photosynthesis-related function, this process should hypothetically involve either positive Darwinian selection for functional divergence (Ohno, 1970) or relaxation of selective constraints (Zhang *et al.*, 1998). However, as the results of the present study suggest, the *cpmA* genes of the C3 subfamily are under strong purifying selection. Although the genes of subfamily C1 are under strong purifying selection too, they, in contrast to C3, have evolved together with photosynthesis-related genes and the other components of the circadian system. Hence their evolution should be concordant to ensure proper interactions between the genes.

The results of the present study suggest that the functional and selective constraints of the *kaiABC*-based system have slowed down the rate of sequence evolution of the *cpmA* genes as compared to the homologues of the *kaiBC*-based system. This supports the hypothesis that the emergence of *kaiA* was one of the most influential events in the evolution of the system and resulted in changes of selective and functional constraints of the other system components (Dvornyk *et al.*, 2004; Dvornyk, 2005). A vector of these changes probably depends on the role of a particular element in the system and may be in favour of either relaxing the constraints (e.g. for the *sasA* and *ldpA* genes: Dvornyk *et al.*, 2004; Dvornyk, 2005) or making them more stringent (as for the *cpmA* gene). However, purifying selection probably remains one of the primary selective forces of their evolution.

The present study of the *cpmA* genes provides further evidence for the complex nature of evolution of the prokaryotic circadian system and the different time of origin of its components. It also raises new questions to be answered. For example, the system with *kaiA* remains functional even after disruption of some other elements, e.g. *sasA* (Iwasaki *et al.*, 2000), *ldpA* (Katayama *et al.*, 2003) or *cpmA* (Katayama *et al.*, 1999), despite the lower fitness of the mutants. Does the *kaiBC*-based system do similarly? How does the N-terminal domain of *cpmA* contribute to the circadian function of the gene? The evolutionary data obtained provide a basis for further studies of molecular mechanisms underlying the control of circadian rhythmicity in cyanobacteria.

ACKNOWLEDGEMENTS

I wish to thank Professor Outi Savolainen (University of Oulu, Finland) and Dr Claus Vogl (University of Veterinary Medicine, Vienna, Austria) for their critical reading and helpful comments on the early draft of the paper.

REFERENCES

- Akaike, H. (1974). New look at statistical model identification tree. *IEEE T Automat Contr* **AC19**, 716–723.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
- Castenholz, R. W. (1989). Subsection IV, order *Nostocales*. In *Bergey's Manual of Systematic Bacteriology*, pp. 1780–1793. Edited by J. T. Staley, M. P. Bryant, N. Pfennig & J. G. Holt. Baltimore: Williams & Wilkins.
- Cummings, M. P., Otto, S. P. & Wakeley, J. (1995). Sampling properties of DNA sequence data in phylogenetic analysis. *Mol Biol Evol* **12**, 814–822.
- Ditty, J. L., Williams, S. B. & Golden, S. S. (2003). A cyanobacterial circadian timing mechanism. *Annu Rev Genet* **37**, 513–543.
- Dvornyk, V. (2005). Molecular evolution of *ldpA*, a gene mediating circadian input signal in cyanobacteria. *J Mol Evol* **60**, 105–112.

- Dvornyk, V. & Knudsen, B. (2005). Functional divergence of circadian clock proteins in prokaryotes. *Genetica* **124**, 247–254.
- Dvornyk, V. & Nevo, E. (2004). Evidence for multiple lateral transfers of the circadian clock cluster in filamentous heterocystic cyanobacteria *Nostocaceae*. *J Mol Evol* **58**, 341–347.
- Dvornyk, V., Vinogradova, O. N. & Nevo, E. (2002). Long-term microclimatic stress causes rapid adaptive radiation of *kaiABC* clock gene family in a cyanobacterium, *Nostoc linckia*, from the "Evolution Canyons" I and II, Israel. *Proc Natl Acad Sci U S A* **99**, 2082–2087.
- Dvornyk, V., Deng, H. W. & Nevo, E. (2004). Structure and molecular phylogeny of *sasA* genes in cyanobacteria: insights into evolution of the prokaryotic circadian system. *Mol Biol Evol* **21**, 1468–1476.
- Dvornyk, V., Vinogradova, O. N. & Nevo, E. (2003). Origin and evolution of circadian clock genes in prokaryotes. *Proc Natl Acad Sci U S A* **100**, 2495–2500.
- Garcia-Pichel, F., Lopez-Cortes, A. & Nubel, U. (2001). Phylogenetic and morphological diversity of cyanobacteria in soil desert crusts from the Colorado plateau. *Appl Environ Microbiol* **67**, 1902–1910.
- Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226–2238.
- Golden, S. S., Brusslan, J. & Haselkorn, R. (1986). Expression of a family of *psbA* genes encoding a photosystem II polypeptide in the cyanobacterium *Anacystis nidulans* R2. *EMBO J* **5**, 2789–2798.
- Goldman, N. & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725–736.
- Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol* **47**, 9–17.
- Honda, D., Yokota, A. & Sugiyama, J. (1999). Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains. *J Mol Evol* **48**, 723–739.
- Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst Biol* **44**, 17–48.
- Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C. R., Tanabe, A., Golden, S. S., Johnson, C. H. & Kondo, T. (1998). Expression of a gene cluster *kaiABC* as a circadian feedback process in cyanobacteria. *Science* **281**, 1519–1523.
- Iwasaki, H., Williams, S. B., Kitayama, Y., Ishiura, M., Golden, S. S. & Kondo, T. (2000). A *kaiC*-interacting sensory histidine kinase, SasA, necessary to sustain robust circadian oscillation in cyanobacteria. *Cell* **101**, 223–233.
- Johnson, C. H. & Golden, S. S. (1999). Circadian programs in cyanobacteria: adaptiveness and mechanism. *Annu Rev Microbiol* **53**, 389–409.
- Kaneko, T., Nakamura, Y., Wolk, C. P. & 18 other authors (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8**, 205–213.
- Katayama, M., Tsinoremas, N. F., Kondo, T. & Golden, S. S. (1999). *cpmA*, a gene involved in an output pathway of the cyanobacterial circadian system. *J Bacteriol* **181**, 3516–3524.
- Katayama, M., Kondo, T., Xiong, J. & Golden, S. S. (2003). *ldpA* encodes an iron-sulfur protein involved in light-dependent modulation of the circadian period in the cyanobacterium *Synechococcus elongatus* PCC 7942. *J Bacteriol* **185**, 1415–1422.
- Kishino, H. & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in *Hominioidea*. *J Mol Evol* **29**, 170–179.
- Kondo, T. & Ishiura, M. (2000). The circadian clock of cyanobacteria. *Bioessays* **22**, 10–15.
- Koonin, E. V., Makarova, K. S. & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* **55**, 709–742.
- Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.
- Kutsuna, S., Kondo, T., Aoki, S. & Ishiura, M. (1998). A period-extender gene, *pex*, that extends the period of the circadian clock in the cyanobacterium *Synechococcus* sp. strain PCC 7942. *J Bacteriol* **180**, 2167–2174.
- Liu, Y., Tsinoremas, N. F., Johnson, C. H., Lebedeva, N. V., Golden, S. S., Ishiura, M. & Kondo, T. (1995). Circadian orchestration of gene expression in cyanobacteria. *Genes Dev* **9**, 1469–1478.
- Lloyd, A. T. & Sharp, P. M. (1993). Evolution of the *recA* gene and the molecular phylogeny of bacteria. *J Mol Evol* **37**, 399–407.
- Lyra, C., Suomalainen, S., Gugger, M., Vezie, C., Sundman, P., Paulin, L. & Sivonen, K. (2001). Molecular characterization of planktic cyanobacteria of *Anabaena*, *Aphanizomenon*, *Microcystis* and *Planktothrix* genera. *Int J Syst Evol Microbiol* **51**, 513–526.
- Maddison, W. P. (1997). Gene trees in species trees. *Syst Biol* **46**, 523–536.
- Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C. & 24 other authors (2003). CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res* **31**, 383–387.
- Martin, A. P. & Burg, T. M. (2002). Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. *Syst Biol* **51**, 570–587.
- Meyer, E., Leonard, N. J., Bhat, B., Stubbe, J. & Smith, J. M. (1992). Purification and characterization of the *purE*, *purK*, and *purC* gene products: identification of a previously unrecognized energy requirement in the purine biosynthetic pathway. *Biochemistry* **31**, 5022–5032.
- Nakamura, Y., Kaneko, T., Sato, S. & 16 other authors (2003). Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* **10**, 137–145.
- Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418–426.
- Nielsen, R. & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.
- Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin: Springer.
- Posada, D. & Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.
- Rokas, A., King, N., Finnerty, J. & Carroll, S. B. (2003). Conflicting phylogenetic signals at the base of the metazoan tree. *Evol Dev* **5**, 346–359.
- Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.
- Schaefer, M. R. & Golden, S. S. (1989). Light availability influences the ratio of two forms of D1 in cyanobacterial thylakoids. *J Biol Chem* **264**, 7412–7417.
- Schmitz, O., Katayama, M., Williams, S. B., Kondo, T. & Golden, S. S. (2000). CikA, a bacteriophytochrome that resets the cyanobacterial circadian clock. *Science* **289**, 765–768.
- Shimodaira, H. & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**, 1114–1116.
- Sitnikova, T., Rzhetsky, A. & Nei, M. (1995). Interior-branch and bootstrap tests of phylogenetic trees. *Mol Biol Evol* **12**, 319–333.
- Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**, 512–526.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (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.

Tiedeman, A. A., Keyhani, J., Kamholz, J., Daum, H. A., III, Gots, J. S. & Smith, J. M. (1989). Nucleotide sequence analysis of the *purEK* operon encoding 5'-phosphoribosyl-5-aminoimidazole carboxylase of *Escherichia coli* K-12. *J Bacteriol* **171**, 205–212.

Tsinoremas, N. F., Ishiura, M., Kondo, T., Tanaka, K., Takahashi, H., Johnson, C. H. & Golden, S. S. (1996). A sigma factor that modifies the circadian expression of a subset of genes in cyanobacteria. *EMBO J* **15**, 2488–2495.

Turner, S. (1997). Molecular systematics of oxygenic photosynthetic bacteria. *Plant Syst Evol [Suppl]* **11**, 13–52.

Watanabe, W., Sampei, G., Aiba, A. & Mizobuchi, K. (1989). Identification and sequence analysis of *Escherichia coli* *purE* and *purK* genes encoding 5'-phosphoribosyl-5-amino-4-imidazole carboxylase for *de novo* purine biosynthesis. *J Bacteriol* **171**, 198–204.

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691–699.

Xiong, J., Fischer, W. M., Inoue, K., Nakahara, M. & Bauer, C. E. (2000). Molecular evidence for the early evolution of photosynthesis. *Science* **289**, 1724–1730.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **15**, 555–556.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**, 568–573.

Yang, Z. & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**, 908–917.

Yang, Z., Goldman, N. & Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol* **11**, 316–324.

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.-M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449.

Yoder, A. D. & Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol* **17**, 1081–1090.

Zhang, J., Rosenberg, H. F. & Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A* **95**, 3708–3713.