Title	Circadian input kinases and their homologs in cyanobacteria: Evolutionary constraints versus architectural diversification
Author(s)	Baca, I; Sprockett, D; Dvornyk, V
Citation	Journal Of Molecular Evolution, 2010, v. 70 n. 5, p. 453-465
Issued Date	2010
URL	http://hdl.handle.net/10722/90482
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Circadian input kinases and their homologs in cyanobacteria: Evolutionary constraints vs architectural diversification

Ivan Baca[1], Daniel Sprockett[2], and Volodymyr Dvornyk[3 §]

[1] Institute of Genetics, National Academy of Sciences of Moldova, Chisinau, Moldova,

[2]Department of Biological Sciences, Kent State University, Kent, OH 44242-0001, USA,

[3]School of Biological Sciences, University of Hong Kong, Pokfulam Rd., Hong Kong SAR, P.R. China

[§]Corresponding author

Email addresses:

IB: bidmd@yahoo.com

DS: dsprocke@kent.edu

VD: dvornyk@hku.hk

Abstract

The *cikA* gene encodes a protein relaying environmental signals to the central circadian oscillator in cyanobacteria. The CikA protein has a variable architecture and usually consists of four tandemly arrayed domains: GAF, histidine kinase (HisKA), histidine kinase-like ATPase (HATPase_c), and a pseudo-receiver (REC). Among them, HisKA and HATPase_c are the least polymorphic, and REC is not present in heterocystic filamentous cyanobacteria. CikA contains several conserved motifs that are likely important for circadian function. There are at least three types of circadian systems, each of which possess a different set of circadian genes. The originally described circadian system (*kaiABC* system) possesses both *cikA* and *kaiA*, while the others lack either only *cikA* (*kaiABC*$^{\Delta}$) or both (*kaiBC*). The results we obtained allowed us to approximate the time of the *cikA* origin to be about 2600-2200 MYA and the time of its loss in the species with the *kaiABC*$^{\Delta}$ or *kaiBC* system between 1100 and 600 MYA. Circadian specialization of CikA, as opposed to its non-circadian homologs, is a result of several factors, including the unique conserved domain architecture and high evolutionary constraints of some domains and regions, which were previously identified as critical for the circadian function of the gene.

Running title: Circadian input kinases of cyanobacteria

Key words: CikA, prokaryotes, GAF, divergence, constraints

Introduction

Cyanobacteria are the simplest organisms known to have an endogenous circadian clock (Kondo and Ishiura, 1999). The model species *Synechococcus elongatus* PCC 7942 contains a cluster of three tandemly arrayed genes, *kaiA*, *kaiB*, and *kaiC*, which has been identified as a key element of the circadian system (Ishiura et al., 1998). Possession of a circadian clock has been shown to enhance the adaptive fitness of cyanobacteria in a wide variety of environmental conditions (Woelfle et al., 2004).

In addition to the three *kai* genes, which were thought to be indispensable for circadian oscillation (Ishiura et al., 1998; Kitayama et al., 2003; Xu et al., 2003), several other genes have been identified to control the input and output of the cyanobacterial clock (see Golden and Canales, 2003, for review). An evolutionary analysis of various components of the circadian system (Dvornyk et al., 2003; Dvornyk et al., 2004; Dvornyk and Knudsen, 2005; Dvornyk, 2006a) suggested that cyanobacteria have at least two types of the system, those with and those without *kaiA* (hereafter referred to as *kaiABC* and *kaiBC* system, respectively). The species lacking *kaiA* possess a timing mechanism, although it is less robust than the original *kaiABC* system (Holtzendorff et al., 2008; Axmann et al., 2009).

The *cikA* (*c*ircadian *i*nput *k*inase) gene encodes a bacteriophytochrome-like histidine kinase involved in the input signaling of the clock (Schmitz et al., 2000). CikA was reported to have three distinct domains: GAF, histidine protein kinase, and a receiver domain (Mutsuda et al., 2003). This structure is typical for bacteriophytochromes (Fankhauser, 2001). However, CikA is missing a conserved cysteine residue, which serves as a bilin ligand in the sensor domain of typical phytochromes. Because of this deletion, it is categorized as an unusual

bacteriophytochrome (Schmitz et al., 2000). A recent study showed that CikA senses light not by a chromophore binding to the GAF domain, but through detecting quinones (Ivleva et al., 2006). The concentration and redox state of quinones in a cell is light dependent. Mutants that are *cikA*-deficient have a shorter circadian period of gene expression and altered phasing of rhythmicity (Schmitz et al., 2000).

In this study, we have analyzed the occurrence, domain architecture, level of variation and phylogeny of the *cikA* gene in order to reconstruct the evolutionary history and to determine the evolutionary factors that have been operating on this component of the cyanobacterial circadian system. We have also attempted to estimate a timeline for key events in the evolution of both *cikA* and the entire circadian system. The present work provides new data about the probable functional importance of various structural motifs of the CikA protein, and significantly updates our knowledge about the evolution of the cyanobacterial circadian system as a whole.

Methods

DNA and protein sequences

The homologous sequences of the CikA proteins, 16S rRNA and 23S rRNA genes were retrieved from the GenBank non-redundant database using gapped PSI-BLAST (with 3 iterations) and BLASTN tools (Altschul et al., 1990; Altschul et al., 1997). The following GenBank accession numbers of the sequences from *S. elongatus* PCC 7942 were used as queries: AAF82192.1, AF132930.1 and CP000100.1, respectively. Only the sequences from completely sequenced cyanobacterial genomes were utilized for the phylogenetic analysis. Since CikA is a member of a large family of bacteriophytochromes, two criteria were used to filter the sequences for the subsequent analyses. First, the proteins should have at least three (GAF-HisKA-HATPase_c) of the four domains arrayed in the same order as in the originally

described CikA. Second, all these domains should display sufficiently high homology to *bona fide* CikA (bit score of 177 was used as a lower limit of homology). With such an approach, some proteins having formally higher similarity score (usually limited to the HisKA-HATPase_c domains) but lacking the above domain architecture were excluded from the analyses.

The sequences were aligned using MUSCLE (Edgar, 2004). The aligned 16S rRNA and 23S rRNA sequences were trimmed and concatenated. The CikA protein sequences were manually adjusted based on the available data about the protein's structure (Mutsuda et al., 2003) to match the putative domains. The *cikA* nucleotide sequences were aligned against the aligned protein homologs using RevTrans v. 1.4 (Wernersson and Pedersen, 2003) available online at http://www.cbs.dtu.dk/services/RevTrans/. The list of the used sequences is given in Supplementary data online Tables S1-S3.

Analysis of variation and phylogenetic reconstruction

The DNA substitution model that fitted the data best was determined for the concatenated 16S rRNA and 23S rRNA genes using the hierarchical test as implemented in the ModelTest 3.0 software (Posada and Crandall, 1998) and HYPHY (Kosakovsky-Pond et al., 2005). Based on the results of this test, the Tamura-Nei model of substitutions with gamma distribution (Tamura and Nei, 1993) and $\alpha = 0.40$ was used for further phylogenetic analysis of the concatenated rRNA genes. For the CikA protein homologs' phylogenetic reconstruction, two empirical amino acid replacement matrices were tested: WAG (Whelan and Goldman, 2001) and LG (Le and Gascuel, 2008). The latter yielded a tree with significantly better likelihood scores. In the reconstruction of the species tree, the 16S rRNA and 23S rRNA genes of the proteobacterium *Rhodobacter sphaeroides* ACC 17029 were used as an outgroup.

The rate of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) was calculated using the Pamilo-Bianchi-Li method (Pamilo and Bianchi, 1993; Li, 1993). The rate of synonymous nucleotide substitutions could not be estimated due to saturation. The MEGA 4 software package (Tamura et al., 2007) was used for the computations of d_N.

The phylogenetic tree of the CikA-like proteins (in total 1113 sites) was constructed using the maximum-likelihood (ML) algorithm implemented in the PHYML 3.0 software (Guindon and Gascuel, 2003). The phylogeny of the concatenated rRNA (4529 sites) genes was reconstructed using two approaches: the ML method as described above and the Bayesian relaxed clock phylogeny as implemented in the BEAST software (Drummond and Rambaut, 2007) with MCMC run for 10 million generations and trees sampled every 1000 steps. The reliability of tree topologies inferred with the ML approach was statistically evaluated using nonparametric bootstrap (100 replications) and the approximate likelihood-ratio test (aLRT) (Anisimova and Gascuel, 2006). Branch lengths in the species tree were then estimated using the ML with local clock and the above specified parameters of the Tamura-Nei model of substitutions with gamma distribution (Tamura and Nei, 1993) implemented in PAML v. 4.1 (Yang, 2007). In a case of the Bayesian phylogeny reconstruction, the maximum clade credibility tree was inferred using TreeAnnotator v.1.5.3, which is included in the BEAST software package.

Reconstruction of the evolutionary time scale

The inferred 16S-23S rRNA tree was tested for the presence of global and local molecular clock using the respective algorithms implemented in HYPHY (Kosakovsky-Pond et al., 2005). Based on the results of the test, the model with local clock was used for the further analysis.

Two internal calibration points were used for the time estimates. One point (CP1) is based on the fossil record about the origin of cyanobacteria and is constrained either by ~3500 MYA (Schopf and Packer, 1987; Walsh, 1992; Kazmierczak and Altermann, 2002; Altermann and Kazmierczak, 2003; Schopf et al., 2007), or by ~2600 MYA (Summons et al., 1999). Another point (CP2), which refers to the appearance of heterocystic cyanobacteria, was calibrated using both molecular and geological data and was estimated between 2450 and 2100 MYA (Tomitani et al., 2006). We used an average value of 2200 MYA for the analysis. These computations were conducted using PAML v. 4.1 (Yang, 2007). We also used the Bayesian relaxed clock phylogeny estimation (Drummond and Rambaut, 2007) as described above. Uncertainty in the estimates was indicated by 95% highest posterior density (95% HPD) intervals.

Modeling of the 3D structure of the GAF domain in the *bona fide* CikA protein

The GAF domain and the adjacent N-terminal region both are critical for the circadian function of CikA, as they control phosphorylation of the kinase domains (Mutsuda et al., 2003). Therefore, we modeled a 3D structure of the GAF domain and mapped on it the conserved motifs identified in this study. The 3D structure was modeled using a majority consensus sequence (in total 180 amino acids) from the alignment of the region between 20 *bona fide* CikA proteins. The initial model was constructed using $(PS)^2$ Protein Structure Prediction Server (http://ps2.life.nctu.edu.tw/) with following options selected: both PSI-BLAST (Altschul et al., 1997) and IMPALA (Schaffer et al., 1999) for template search, and RAMP (http://software.compbio.washington.edu/ramp/ramp.html) for the model building. The obtained initial model was then optimized using the MolProbity server at http://molprobity.biochem.duke.edu/index.php (Lovell et al., 2003). The quality of the

model's versions was assessed with PROCHECK (Laskowski et al., 1993), ERRAT (Colovos and Yeates, 1993), and VERIFY_3D (Luthy et al., 1992), as implemented in SAVES (http://nihserver.mbi.ucla.edu/SAVES_3/), and ProQ (Wallner and Elofsson, 2003) (http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi) The model built upon a consensus of the following three templates (PDB ID codes): 2K2N (Cornilescu et al., 2008), 2OOL (Yang et al., 2007) and 2O9C (Wagner et al., 2007) yielded the best scores.

Identification of amino acid residues of potential functional importance in the CikA proteins

The level of conservation is usually correlated to the functional importance of a particular amino acid site or a sequence motif (Kimura, 1983; Graur and Li, 2000). If particular sites in one protein subfamily are more conserved (or fixed) as compared to other subfamilies, they are assumed to be more functionally important for that subfamily. We applied ConSeq (Berezin et al., 2004) available at http://conseq.tau.ac.il/index.html) to identify which of the conserved residues are of potential functional significance.

Results

Architecture, occurrence, and phylogeny of the *cikA* genes and their homologs

Comparison of the *S. elongatus* CikA against the NCBI Conserved Domain Database (Marchler-Bauer et al., 2003) suggested that the protein consists of four tandemly arrayed functional domains: a GAF domain, a histidine kinase phosphoacceptor domain (HisKA), a histidine kinase-like ATPase domain (HATPase_c), and a signal receiver domain (REC) (Fig. 1). HisKA and HATPase_c are usually considered to be components of a single histidine-protein kinase (HPK)

unit and are referred to as dimerization and catalytic (ATP/ADP-binding phosphotransfer) domains, respectively (Stock, 1999). A BLAST search of available completed bacterial genomes revealed several hundred *cikA* homologs that may be classified as two-component histidine kinases. However, the majority of the matches were limited to either two (HisKA and HATPase_c) or three (HisKA, HATPase_c, and REC) domains. A relatively small number of the homologs possessed the GAF domain. Both GAF and its adjacent N-terminal region are critical for autophosphorylation of the HisKA domain and their deletion negatively affects CikA expression (Mutsuda et al., 2003). Therefore, only genes indicating homology to GAF as well as the HisKA and HATPase_c domains were selected for the further analyses. A number of the homologous genes containing other domains in addition to the originally described *cikA* four-domain architecture (GAF – HisKA – HATPase_c – REC) were found in cyanobacteria. The other domains included, but were not limited to, sensor domains (e.g., PAS, GAF, and the others), CBS, REC, HPT and the others. For example, the genes from the closely related thermophilic Yellowstone isolates *Synechococcus* sp. JA-2-3B'a(2-13) (JA-2 YP_476763) and *Synechococcus* sp. JA-3-3Ab (JA-3 YP_476201) (Table S1, Supplementary material online) have the following domain architecture: GAF – CHASE – PAS – PAS – GAF – HisKA – HATPase_c – REC – REC – HPT (Fig. S1, Supplementary material online).

Both ML and Bayesian phylogenetic analyses of the CikA-like proteins yielded an identical tree topology featuring five major distinct clades with high statistical support. They are designated hereafter as A1-A5 (Fig. 2). Clade A1 includes the originally described CikA from *S. elongatus* PCC 7942 (Schmitz et al., 2000) and its closest homologs, which are thus presumed to have a circadian function. These proteins usually have a fairly stable four-domain architecture GAF – HisKA –

HATPase_c – REC (as in CikA of *S. elongatus* PCC 7942), except for those of filamentous heterocystic Nostocales (genera *Anabaena* and *Nostoc*) lacking the REC domain (Fig. S1, Supplementary material online). All other CikA-like proteins are from cyanobacteria possessing the original *kaiABC* system, except for *Gloeobacter violaceus* PCC 7421, which does not have any *kai* genes (Nakamura et al., 2003). However, these homologs are more variable in their architecture, featuring various additional domains upstream in the N-terminal region or downstream in the C-terminal region, as described above (Fig. S1, Supplementary material online). The different domain architectures of the *cikA* homologs are likely indicative of their different functional assignments. Importantly, only the proteins from clade A1 (presumably the *bona fide* CikA proteins) occur in the genomes of all studied cyanobacterial species with the original *kaiABC* system. The proteins from clades A2-A5 were found only in a subset of the species presented here. These phylogenetic patterns are similar to those of other known circadian genes and their non-circadian homologs (Dvornyk, 2006b).

A notable feature of the *cikA* homologs is their high diversification in some species. The number of the homologs can vary considerably, from a single gene copy in *G. violaceus* PCC 7421, *Trichodesmium erythraeum* IMS101, and the others, up to the eight copies present in *Acaryochloris marina* MBIC11017 (Table S1, Supplementary material online).

Divergence of the *cikA* homologs and *cikA*-like genes

The *cikA* homologs from cyanobacteria exhibit domain-specific patterns of nucleotide variation. HisKA and HTPase_c are usually the most conserved domains (Table 1). The genes from A1, which presumably have a circadian function, appear to be the second least polymorphic after those of A2 (Table 1). Despite this, the REC domain of the genes from clade A1 displays the lowest level of conservation.

The CikA proteins of clade A1 have several regions that are much more conserved than in the other clades. These regions may be of particular importance for circadian functionality (Fig. 1 and Fig. S2, Supplementary material online). Specifically, one of these conserved motifs corresponds to the part of the N-terminal region immediately preceding the GAF domain (Fig. 3). The N-terminal region upstream of the GAF domain was previously shown to enhance phosphorylation of the HisKA domain (Mutsuda et al., 2003), however, no specific fragment of this region was identified as a major contributor to this function. The analysis of variation suggests that this fragment (motif 1) corresponds to amino acid residues 168-183 (numbering refers to positions in the CikA protein of *S. elongatus* PCC 7942). Notably, the whole N-terminal region preceding the GAF domain in CikA of *Lyngbya* sp. PCC 8106 and *Arthrospira maxima* CS-328 consists only of this motif (Fig. 3). This is further evidence for its functional significance. Another conserved region, motif 2, is located near motif 1 and includes residues 199-210 (Fig. 3).

The C-terminal region of the GAF domain exhibits a much lower level of variation than its other regions. Within its C-terminal section, a highly conserved segment of 10 residues (pos. 309-318, motif 3) was identified in the CikA proteins (Fig. 3). A search against the functional motif databases returned no apparent homologs of this motif, and its function remains undetermined.

In the *cikA* genes of clade A1, the highly conserved segments in the HisKA and HATPase_c domains are located near previously identified functionally important motifs (Schmitz et al., 2000; Mutsuda et al., 2003). For example, motif 4 (pos. 578-587) is immediately adjacent to the G-X-G motif (pos. 574-576) in the G box (Fig. S2, Supplementary material online). The G-X-G motifs are critical for ATP binding and are located in the loops that shape the top and bottom of the binding pocket (Obermann

et al., 1998). The segment between these motifs is more conserved in clade A1 than in the others, suggesting that the existing tertiary structure of the pocket is extremely important for the circadian function of the HATPase_c domain and CikA. Notably, the patterns of variation within the pocket highlight its different tertiary structure among the different clades.

The REC domain does not occur in some members of clade A1: it is missing in heterocystic filamentous Nostocaceae. Unlike the other three domains, REC is most polymorphic in clade A1, specifically at the functionally important fixed positions 678Asp, 708Thr, and 727Lys in clades A2 and A3. The conserved Asp residue functions as phosphoryl acceptor, while the conserved Lys is essential for the β5α5 loop and formation of the dimer upon phosphorylation (Solá et al., 1999). Likewise, the intermolecular recognition site immediately following the phosphorylated residue (Müller-Dieckmann et al., 1999) is conserved in the other clades, while being polymorphic in A1.

The model of the CikA GAF domain

The quality assessment parameters of the constructed 3D model are presented in Table 2. They show that the predicted model is of good quality. The model basically follows the experimentally determined structure of the bacterial PAS-related domains (Wagner et al., 2007; Yang et al., 2007; Cornilescu et al., 2008): it consists of five antiparallel β-sheets located between two groups of α-helices (see Fig. S3, Supplementary material online). The β-sheets shape the bottom of a bilin binding pocket that is characteristic of GAF domains. The conserved motifs 1, 2, and 3 correspond to helix α1 and sheets β1 and β5, respectively.

Evolutionary constraints associated with functional specialization of the CikA proteins

The results of the Conseq analysis (Fig. S3, Supplementary material online) suggest that the conserved motifs identified in the CikA proteins may be functionally and/or structurally important. Motif 1 seems to be primarily functional: most of its residues are exposed and seven of them are determined as functionally important (Fig. S4). Motif 3 is mainly of structural significance, as all of its residues but one are buried. Motifs 2 and 4 are important both functionally and structurally, as they possess buried and exposed amino acid residues in about equal proportions. Furthermore, motif 4 is located near the ATP-binding pocket (Fig. 1, see above), which suggests the histidine in this position is of critical importance to CikA phosphorylation.

Evolutionary time estimates

We based our reconstruction of the time scale for the key events in evolution of CikA and the circadian system on the following facts (see Figs. 2 and 4 for reference): (1) *bona fide* CikA is present only in group S1; (2) the REC domain is missing in filamentous heterocystic Nostocaceae; (3) KaiA is missing in clade S3 (Dvornyk et al., 2003). These data constrain the time of the CikA origin to the period between nodes 1 and 2; the CikA loss in S2 and S3 – between nodes 3 and 5; the REC domain loss – between CP2 and node 4; the KaiA loss – between nodes 6 and 7. Table 3 presents the resulting time estimates for the respective nodes based on the results of the ML and Bayesian analyses. However, it should be taken into account that they are likely biased toward the higher values, because CP1 was placed in the node, which is apparently not basic for cyanobacteria, as other yet unknown cyanobacterial species evolutionarily older than Gloeobacter may potentially exist.

Discussion

CikA was identified as an important input component in the *kaiABC* system of *S. elongatus* PCC 7942 (Schmitz et al., 2000) and was hypothesized to transfer a signal to the central oscillator (presumably KaiA) through the yet unknown associated response regulator (Mutsuda et al., 2003; Zhang et al., 2006). However, as the results of the present study show, neither the *cikA* gene nor its apparent GAF-containing homologs occur in the cyanobacteria of clades S2 and S3 (Fig. 4). The species of clade S2 contain the *kaiA* gene, while those of clade S3 lack it. This leads to two conclusions. First, *cikA* was lost before *kaiA*. Second, if the species from either S2 or S3 possess circadian rhythmicity, they should utilize a different molecular mechanism for signal input. It was hypothesized earlier that information about the light signal may be delivered to the central oscillator not through a photoreceptor, but indirectly through sensing the redox state of the cell (Ivleva et al., 2006) by using another input element, such as LdpA (Ivleva et al., 2005). In contrast to CikA, LdpA occurs in all known cyanobacteria (Dvornyk, 2005); Dvornyk, unpublished]. This suggests that the indirect transfer of the light signal might have become a primary mechanism of the circadian input after the loss of *cikA*. The recent results of biochemical studies provide support for this hypothesis (Holtzendorff et al., 2008; Axmann et al., 2009). On the other hand, such an indirect signal transfer mechanism might have existed before the origin of *cikA*.

The circadian system of the species from clade S2 (Fig. 4) is of particular interest, because it shares structural features of the two systems: it has *kaiA* that makes it more similar to the *kaiABC* system, but lacks *cikA* as the *kaiBC* system does. In addition, other features of this circadian system (e.g., the evolutionary history of the *cpmA* gene regulating the circadian output) position it closer to the *kaiBC* type

(Dvornyk et al., 2003; Dvornyk et al., 2004; Dvornyk and Knudsen, 2005; Dvornyk, 2006b). If the *kaiABC*$^{\Delta}$ system of clade S2 is as functionally transitional between the *kaiABC* and *kaiBC* as it is structurally intermediate, then studying it may provide important information about the functional evolution of the original *kaiABC* system into the *kaiBC* system.

In the original study of CikA in *S. elongatus* PCC 7942, this protein was classified as a non-typical bacteriophytochrome, due to the missing conserved Cys residue at position 285 (Schmitz et al., 2000) that normally serves as a bilin ligand for phytochromes (Li and Lagarias, 1992). This suggests that this particular CikA may interact with another protein (possibly possessing the GAF domain) to replace the bilin acceptor (Mutsuda et al., 2003). However, given that most of the CikA proteins in clade A1 do possess this critical cysteine residue, the proposed mechanism may represent only one of the possible variants. Recent findings by Narikawa et al. (Narikawa et al., 2008) support this view, and show that GAF domains retaining the conserved C285 residue may function as a violet light sensor. This, in turn, explains the functional assignment of motifs 2 and 3 (Fig. 3), which ensure proper 3D configuration of the bilin-binding pocket (Fig. S3, Supplementary material online).

A recent study of CikA in *S. elongatus* PCC 7942 proposed positive regulation of CikA phosphorylation by the GAF domain and negative regulation by the REC domain (Mutsuda et al., 2003; Zhang et al., 2006). Autophosphorylation is essential for the circadian function of CikA (Mutsuda et al., 2003). This process involves all three principal domains of the protein (GAF as a positive regulator, HisKA as a phosphoacceptor, and HATPase_c as an ATP binder) and therefore assumes a close inter-domain interaction. For such an interaction, a corresponding tertiary structure of the CikA protein is critical for autophosphorylation. On the other hand, since CikA

belongs to the superfamily of sensor kinases of bacterial two-component signal transduction systems, it is expected to supply a phosphoryl group to a specific, yet unidentified, response regulator(s) (Schmitz et al., 2000; Mutsuda et al., 2003). This putative function of CikA assumes a close correspondence of its structure to that of the response regulator. The highly conserved motifs identified in the present study (Fig. 3) are probably critical for maintaining the tertiary structure of CikA, ensuring said physical interactions both between the domains and between each domain and the response regulator.

Recently, the pseudo-receiver REC domain was confirmed as necessary for the CikA circadian function in *S. elongatus* PCC 7942 by entraining the clock through sensing the redox state of cellular quinones (Ivleva et al., 2006) and repressing the kinase activity of the protein (Mutsuda et al., 2003; Zhang et al., 2006). However, REC is most polymorphic in clade A1 as compared to the other clades (Table 1) and is even missing in some cyanobacteria (e.g., Nostocales). According to the principle of functional constraint in molecular evolution, functional importance of a protein or domain directly correlates with its level of conservation (Kimura, 1983; Graur and Li, 2000). Thus the results of our study raise several questions about the functional significance of REC. Specifically, how is the protein autokinase activity of the REC-deficient CikA controlled? How is quinone sensing conducted? One possibility is that these REC-associated functions in some cyanobacterial species are performed by an unidentified response regulator protein.

Interestingly, two components of the cyanobacterial circadian system, CikA and SasA, which respectively control the input and output pathways in the *kaiABC* system, are autophosphorylated *in vitro* and have a similar domain structure (Sensor Domain – HisKA – HATPase_c) to the one that is common in two-component sensory

transduction histidine kinases (Dutta et al., 1999; Iwasaki et al., 2000; Mutsuda et al., 2003). The similar domain organization of CikA and SasA suggests that they both may have originated through the common evolutionary mechanism: the fusion of the sensor domain with a double-domain two-component sensory transduction histidine kinase. Two-component sensory transduction histidine kinases are a large superfamily widely distributed in prokaryotes (Nagaya et al., 1993). Proteins of this superfamily display a diversity of domain organizations (Dutta et al., 1999) that are the apparent result of multiple gene fusion events. Recent findings suggest gene fusions have played a major role in the evolution of protein-domain architectures (Yanai et al., 2002). These fusions were probably quite common in evolution of the CikA-like proteins. While the core domains, GAF, HiSKA and HTPase_c, exhibit relatively high similarity between the different members of clades A2-A5, the domain organization of the proteins varies greatly (Figs. 2 and S1, Supplementary material online).

The aggregated data of the current study and previous functional studies of *cikA* (Schmitz et al., 2000; Mutsuda et al., 2003; Zhang et al., 2006; Ivleva et al., 2006) provide evidence that various evolutionary mechanisms have resulted in functional specialization of *cikA* as a circadian gene. After the duplication of the ancestral two-component histidine kinase, *cikA* experienced neofunctionalization through accretion of specific domains (GAF and REC) while maintaining conservation at functionally important sites and domain architecture. While the CikA-like proteins from clades A2-A5 experienced significant diversification of the domain organization, the original CikA protein maintained its high level of conservation (Fig. S1, Supplementary material online). The acquired circadian function was then maintained by strong purifying selection in the regions conferring this function. This is a common pattern

for circadian genes, which are usually more conserved than their non-circadian paralogs (Dvornyk et al., 2004; Dvornyk, 2006b).

The results of the present study make it possible to roughly estimate the probable time of origin and the main events in the evolution of the *cikA* gene. The following time estimates are approximate, especially with regard to the very early stages of this evolution, and will be updated and corrected as additional genomic data are accumulating. According to the ML estimates, depending on the accepted date of the appearance of cyanobacteria, the lower time limit of *cikA* origin is placed about 2900 MYA, while the upper limit is placed about 2200 MYA (Table 3). The gene was lost in groups S2 and S3 in the period between 1050 and 600 MYA. The broad range of the estimate is due to the absence of data about the circadian system in cyanobacterial species that are phylogenetically positioned between *S. elongatus* PCC 7942 and *Synechococcus* sp. RCC307. The loss of the REC domain in filamentous Nostocales seems to have occurred before the loss of *cikA*, falling within the period of approximately 2200-900 MYA (Fig. 4, Table 3). The Bayesian analysis yielded similar to the above estimates for all nodes, except for nodes 3 and 5 (Table 3). The latter pull the date of the *cikA* loss back by almost two-fold, between 1900 and 1000 MYA. This large difference between the ML and Bayesian estimates is likely not due to the phylogenetic uncertainty, because both methods produced the same tree topology (Fig. 4). However, the most recent studies suggest that the ML analysis tends to give more accurate estimates of branch length and, respectively, divergence times (Schwartz and Mueller, 2010).

The *kaiABC* system was initially thought to have evolved from the *kaiBC* system through the addition of *kaiA*, which presumably originated about 1000 MYA (Dvornyk et al., 2003). The recent analysis of newly available, more extensive

genomic data suggests a different scenario: the current *kaiBC* system evolved stepwise during 1050-400 MYA from the *kaiABC* through the loss of *kaiA* and other components, including *cikA* (Dvornyk, 2006a; Dvornyk, 2009). Apparently, these losses were associated with the origin and rapid radiation of cyanobacteria in clades S2 and S3 (Fig. 4). Interestingly, the time limits of this radiation (between nodes 5 and 6, i.e., about 600-500 MYA, Fig.4 and Table 3) correspond to the period around the well-known Cambrian explosion. Furthermore, according to the results of the current study, *kaiA* was lost between 500 and 400 MYA (Table 3), which corresponds to the hypothesized upper time limit of the last of the three periods proposed to describe the role of UV radiation in the evolution of cyanobacteria (Garcia-Pichel, 1998). This period is thought to last between 1000 and 400 MYA and was associated with the increase of atmospheric oxygen content and the formation of the earth's ozone shield (Canfield and Teske, 1996; Garcia-Pichel, 1998). Recently, it was suggested that the loss of *kaiA* and the associated decrease in circadian oscillator robustness occurred due to the adaptation of *Prochlorococcus* to a temperature-stable ecological niche that does not require a robust oscillator (Holtzendorff et al., 2008; Axmann et al., 2009). This scenario may not fully explain the observed phylogenetic patterns, however, since *S. elongatus* PCC7942 possesses *kaiA* yet often occurs with *Prochlorococcus* in the same ecological niches (Partensky et al., 1999). Further studies may help to determine the factors that triggered the large-scale reduction of the *Prochlorococcus* genome (Dufresne et al., 2003) including the loss of several circadian genes.

An intriguing finding of our study is that none of the species from clades S2 and S3 have any GAF-containing genes. Obviously, these genes were present in the genome of the common ancestor of *S. elongatus* PCC 7942 and *Synechococcus* sp. RCC307 at the time point corresponding to node 3 (Fig. 4). However, for reasons yet

unknown, all GAF-containing genes were lost in the lineage leading to *Synechococcus* sp. RCC307. Of course, this loss might occur in any time point between nodes 3 and 5, but a more accurate estimate will only be possible when more genomic data in this lineage become available. The loss of *cikA* in some cyanobacterial lineages suggests that the evolution of this gene and its homologs follows the birth-and-death scenario (Nei and Rooney, 2005).

The results of the 16S-23S rRNA phylogenetic analysis provide some insights into molecular systematics of cyanobacteria. Specifically, assigning the species from clades S2 and S4 to the same genus *Synechococcus* seems unjustified. In fact, they are phylogenetically quite distant from each other as well as from *S. elongatus* PCC 7942. The polyphyly of the genus *Synechococcus* has been comprehensively discussed elsewhere (Urbach et al., 1998; Honda et al., 1999; Robertson et al., 2001). In total, up to eight *Synechococcus* lineages are recognized on the basis of 16S rRNA and other genes analyses (Honda et al., 1999; Robertson et al., 2001). However, this topic warrants a separate, more comprehensive investigation involving a much larger number of available strains.

All previous evolutionary studies have shown that the elements of the circadian system usually have lower variation than their apparent non-circadian homologs; this variation is specific to each type of the system and is maintained by strong purifying selection (Dvornyk et al., 2004; Dvornyk, 2005). While the core genes, *kaiB* and *kaiC*, are common among the different circadian system types, the input and output signal pathways differ significantly, thereby conferring functional and selective constraints to each type. The *kaiABC* system originally described in *S. elongatus* PCC 7942 is apparently evolutionarily oldest among the three; however it is still unclear what happened about 1050-600 MYA and 400 MYA that resulted in loss of *cikA* and *kaiA*,

respectively. Further comparative evolutionary and functional studies of all types of the cyanobacterial circadian system are needed to reveal the specific molecular mechanisms that have been developed and utilized during this system's evolution towards functional specialization.

Acknowledgements

This work was supported by the University of Hong Kong start-up fund for VD.

Abbreviations

cikA, circadian input kinase A; MYA, million years ago; NCBI, National Center for Biotechnology Information

Figures

Fig. 1. Domain architecture of the CikA protein with the mapped motifs of putative functional and/or structural importance. Dashed boxes represent regions not present in CikA of all species.

Fig. 2. Unrooted maximum-likelihood tree of the CikA homologs in cyanobacteria**.** Bar, 0.1 substitutions per site. Maximum-likelihood probabilities of the node support <0.5 and bootstrap <50 are not shown. A1-A5 refer to the subfamilies of the CikA homologs. Clade A1 comprises *bona fide* CikA proteins. For the designations of the proteins see Supplementary data file 1.

Fig. 3. Alignment of the GAF domains of the cyanobacterial CikA proteins. Motifs 1, 2, and 3 are underlined. Block arrow indicates putative cysteine ligand. Black and grey-shaded backgrounds indicate different degree of conservation (black is the most conserved). The upper numbers indicate positions in the alignment of the full sequences; the numbers on the right indicate positions in the respective sequences. Visualized using Genedoc (Nicholas and Nicholas, Jr., 1997).

Fig. 4. Maximum-likelihood tree with local clock of the concatenated 16S rRNA and 23S rRNA genes from cyanobacteria**.** Bar, 1 substitution per site. ML and posterior probabilities of the node support are shown in numerator, bootstrap proportion values are shown in denominator. ML values<0.5 and bootstrap <50 are not shown. ML and posterior probabilities values equal 1.00 are shown without decimals. S1-S4 refer to the groups of the species with different composition of the circadian system. Species with the *bona fide* CikA are boxed, those belonging to Nostocales and lacking the REC domain are shown in bold. Black boxes indicate the calibration points. Black circles and numbers mark nodes that correspond to the key events in evolution of CikA and the circadian system: the CikA origin – between nodes 1 and 2;

the CikA loss – between nodes 3 and 5; the REC domain loss – between CP2 and node 4; the KaiA loss – between nodes 6 and 7. See text for the details.

Tables

Table 1. Patterns of nonsynonymous nucleotide substitutions (d_N) in the different regions of the *cikA* genes and *cikA*-like homologs.

Clade	Domain					Average over gene[2]
	GAF	HisKA	HTPase_c	REC[1]	Rest of gene	
A1	0.478 ± 0.039	0.272 ± 0.036	0.268 ± 0.032	0.495 ± 0.050	0.486 ± 0.039	0.417 ± 0.017
A2	0.229 ± 0.061	0.137 ± 0.063	0.274 ± 0.042	0.250 ± 0.040	0.614 ± 0.111	0.343 ± 0.051
A3	0.582 ± 0.115	0.453 ± 0.145	0.550 ± 0.105	–	0.879 ± 0.147	0.578 ± 0.088
A4	0.491 ± 0.039	0.260 ± 0.038	0.444 ± 0.046	0.446 ± 0.037	0.643 ± 0.055	0.484 ± 0.018
A5	0.615 ± 0.044	0.310 ± 0.047	0.451 ± 0.042	0.460 ± 0.068	0.685 ± 0.042	0.554 ± 0.022
Average over domain	0.767 ± 0.038	0.510 ± 0.043	0.566 ± 0.039	0.799 ± 0.055	0.920 ± 0.043	0.723 ± 0.025

[1] the number of the analyzed sequences is smaller due to the exclusion those lacking the REC domain;

[2] the REC domain is excluded as most genes from clade A2 lack it.

Table 2. Quality assessment parameters for the constructed 3D structure of the GAF domain.

Parameter	Value
Ramachandran plot	91.0% core, 7.2% allowed, 1.8% generously allowed, 0.0% disallowed
All Ramachandrans	8 labelled residues (out of 178)
Chi1-chi2 plots	1 labelled residues (out of 124)
Main-chain	6 better, 0 inside, 0 worse
Side-chain	5 better, 0 inside, 0 worse
Residue properties	Max.deviation: 2.3, bad contacts: 14, bond length/angle: 7.5, Morris et al class: 1 1 2
G-factors	Dihedrals: 0.06, covalent: -0.22, overall: -0.04
M/c bond lengths	98.7% within limits
M/c bond angles	92.8% within limits
Planar groups	100.0% within limits
Verify 3D score >0.2	65.19% of the residues
Errat overall quality factor	71.512
Predicted LGscore	4.477
Predicted MaxSub	0.267

Table 3. Maximum-likelihood and Bayesian time estimates, MYA, for the nodes (Fig. 4) corresponding to the major events in evolution of CikA and the circadian system. Based on the 16S-23S rRNA tree.

	Maximum-likelihood		Bayesian
Node	2600*	3500	3259 (2749-3793)**
1	2598 ± 15	2875 ± 219	3026 (2534-3515)
2	2223 ± 145	2217 ± 195	2583 (2186-2973)
3	1031 ± 105	1079 ± 111	1943 (1372-2524)
4	910 ± 63	952 ± 66	912 (400-1523)
5	588 ± 38	614 ± 40	1035 (576-1531)
6	499 ± 45	521 ± 48	670 (403-971)
7	407 ± 26	426 ± 28	565 (336-836)

* Time for calibration point 1, MYA,

** Posterior mean (95% HPD).

Supplementary material online

Supplementary data file 1 – List of sequences used in the study.

File format: pdf

Description: Tables S1, S2 and S3 provide a list of *cikA*, 16S rRNA and 23S rRNA sequences used in the study.

Supplementary Figure S1.

File format: pdf

Description: Phylogeny of the domain architectures of the *bona fide* CikA proteins and CikA-like proteins. Shown according to the clades in Fig. 1. Clades with the same domain architectures are collapsed.

Supplementary Figure S2.

File format: pdf

Description: Full multiple alignment of the *bona fide* CikA proteins. Black and grey-shaded backgrounds indicate different degree of conservation (black is the most conserved). The upper numbers indicate positions in the alignment of the full sequences; the numbers on the right indicate positions in the respective sequences. Visualized using Genedoc (Nicholas and Nicholas, Jr., 1997).

Supplementary Figure S3.

File format: pdf

Description: A theoretical 3D model of the GAF domain of the CikA protein. α-helices are shown in red, β-sheets are shown in yellow. A putative bilin-binding

residue, C285, which is conserved in most CikA proteins, is shown. Visualized with Sirius 1.2 (San Diego Supercomputer Center).

Supplementary Figure S4.

File format: pdf

Description: Predicted amino acid sites of functional and structural significance in the CikA protein as determined by ConSeq (Berezin et al., 2004). The predicted residues are mapped on the CikA protein of *S. elongatus* PCC 7942.

References

Altermann W, Kazmierczak J (2003) Archean microfossils: a reappraisal of early life on Earth. Res Microbiol 154:611-617

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402

Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol 55:539-552

Axmann IM, Duhring U, Seeliger L, Arnold A, Vanselow JT, Kramer A, Wilde A (2009) Biochemical evidence for a timing mechanism in *Prochlorococcus*. J Bacteriol 191:5342-5347

Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 20:1322-1324

Canfield DE, Teske A (1996) Late Proterozoic rise in atmospheric oxygen concentration inferred from phylogenetic and sulphur-isotope studies. Nature 382:127-132

Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2:1511-1519

Cornilescu G, Ulijasz AT, Cornilescu CC, Markley JL, Vierstra RD (2008) Solution structure of a cyanobacterial phytochrome GAF domain in the red-light-absorbing ground state. J Mol Biol 383:403-413

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214

Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S, Galperin MY, Koonin EV, Le Gall F, Makarova KS, Ostrowski M, Oztas S, Robert C, Rogozin IB, Scanlan DJ, Tandeau de Marsac N, Weissenbach J, Wincker P, Wolf YI, Hess WR (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci USA 100:10020-10025

Dutta R, Qin L, Inouye M (1999) Histidine kinases: diversity of domain organization. Mol Microbiol 34:633-640

Dvornyk V (2005) Molecular evolution of *ldpA*, a gene mediating circadian input signal in cyanobacteria. J Mol Evol 60:105-112

Dvornyk V (2006a) Evolution of the circadian clock mechanism in prokaryotes. Isr J Ecol Evol 52:343-357

Dvornyk V (2006b) Subfamilies of *cpmA*, a gene involved in circadian output, have different evolutionary histories in cyanobacteria. Microbiology 152:75-84

Dvornyk V (2009) The circadian clock gear in cyanobacteria: Assembled by evolution. In: Ditty, J. L., Mackey, S., and Johnson, C. H. (eds) *Bacterial Circadian Programs*. Springer-Verlag, Berlin-Heidelberg, pp 241-258

Dvornyk V, Deng HW, 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, Knudsen B (2005) Functional divergence of circadian clock proteins in prokaryotes. Genetica 124:247-254

Dvornyk V, Vinogradova ON, Nevo E (2003) Origin and evolution of circadian clock genes in prokaryotes. Proc Natl Acad Sci USA 100:2495-2500

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797

Fankhauser C (2001) The phytochromes, a family of red/far-red absorbing photoreceptors. J Biol Chem 276:11453-11456

Garcia-Pichel F (1998) Solar ultraviolet and the evolutionary history of cyanobacteria. Orig Life Evol Biosph 28:321-347

Golden SS, Canales SR (2003) Cyanobacterial circadian clocks - timing is everything. Nature Rev Microbiol 1:191-199

Graur D, Li WH (2000) Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, MA

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704

Holtzendorff J, Partensky F, Mella D, Lennon JF, Hess WR, Garczarek L (2008) Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium *Prochlorococcus marinus* PCC 9511. J Biol Rhythms 23:187-199

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

Ishiura M, Kutsuna S, Aoki S, Iwasaki H, Andersson CR, Tanabe A, Golden SS, Johnson CH, Kondo T (1998) Expression of a gene cluster *kaiABC* as a circadian feedback process in cyanobacteria. Science 281:1519-1523

Ivleva NB, Bramlett MR, Lindahl PA, Golden SS (2005) LdpA: a component of the circadian clock senses redox state of the cell. EMBO J 24:1202-1210

Ivleva NB, Gao T, LiWang AC, Golden SS (2006) Quinone sensing by the circadian input kinase of the cyanobacterial circadian clock. Proc Natl Acad Sci USA 103:17468-17473

Iwasaki H, Williams SB, Kitayama Y, Ishiura M, Golden SS, Kondo T (2000) A *kaiC*-interacting sensory histidine kinase, SasA, necessary to sustain robust circadian oscillation in cyanobacteria. Cell 101:223-233

Kazmierczak J, Altermann W (2002) Neoarchean biomineralization by benthic cyanobacteria. Science 298:2351

Kimura M (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge

Kitayama Y, Iwasaki H, Nishiwaki T, Kondo T (2003) KaiB functions as an attenuator of KaiC phosphorylation in the cyanobacterial circadian clock system. EMBO J 22:2127-2134

Kondo T, Ishiura M (1999) The circadian clocks of plants and cyanobacteria. Trends Plant Sci 4:171-176

Kosakovsky-Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283-291

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307-1320

Li L, Lagarias JC (1992) Phytochrome assembly. Defining chromophore structural requirements for covalent attachment and photoreversibility. J Biol Chem 267:19204-19210

Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96-99

Lovell SC, Davis IW, Arendall WB, III, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins 50:437-450

Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356:83-85

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH (2003) CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res 31:383-387

Müller-Dieckmann HJ, Grantz AA, Kim SH (1999) The structure of the signal receiver domain of the *Arabidopsis thaliana* ethylene receptor ETR1. Structure 7:1547-1556

Mutsuda M, Michel KP, Zhang X, Montgomery BL, Golden SS (2003) Biochemical properties of CikA, an unusual phytochrome-like histidine protein kinase that resets the circadian clock in *Synechococcus elongatus* PCC 7942. J Biol Chem 278:19102-19110

Nagaya M, Aiba H, Mizuno T (1993) Cloning of a sensory-kinase-encoding gene that belongs to the two-component regulatory family from the cyanobacterium *Synechococcus* sp. PCC7942. Gene 131:119-124

Nakamura Y, Kaneko T, Sato S, Mimuro M, Miyashita H, Tsuchiya T, Sasamoto S, Watanabe A, Kawashima K, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A, Nakazaki N, Shimpo S, Takeuchi C, Yamada M, Tabata S (2003) Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. DNA Res 10:137-145

Narikawa R, Kohchi T, Ikeuchi M (2008) Characterization of the photoactive GAF domain of the CikA homolog (SyCikA, Slr1969) of the cyanobacterium *Synechocystis* sp. PCC 6803. Photochem Photobiol Sci 7:1253-1259

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121-152

Nicholas, K. B. and Nicholas, H. B., Jr. GeneDoc: a tool for editing and annotating multiple sequence alignments. 1997

Obermann WM, Sondermann H, Russo AA, Pavletich NP, Hartl FU (1998) In vivo function of Hsp90 is dependent on ATP binding and ATP hydrolysis. J Cell Biol 143:901-910

Pamilo P, Bianchi NO (1993) Evolution of the *Zfx* and *Zfy* genes: Rates and interdependence between the genes. Mol Biol Evol 10:271-281

Partensky F, Hess WR, Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev 63:106-127

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818

Robertson BR, Tezuka N, Watanabe MM (2001) Phylogenetic analyses of *Synechococcus* strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. Int J Syst Evol Microbiol 51:861-871

Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15:1000-1011

Schmitz O, Katayama M, Williams SB, Kondo T, Golden SS (2000) CikA, a bacteriophytochrome that resets the cyanobacterial circadian clock. Science 289:765-768

Schopf JW, Kudryavtsev AB, Czaja AD, Tripathi AB (2007) Evidence of archean life: Stromatolites and microfossils. Precambrian Res 158:141-155

Schopf JW, Packer BM (1987) Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. Science 237:70-73

Schwartz RS, Mueller RL (2010) Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. BMC Evol Biol 10:5

Solá M, Gomis-Rüth FX, Serrano L, González A, Coll M (1999) Three-dimensional crystal structure of the transcription factor PhoB receiver domain. J Mol Biol 285:675-687

Stock J (1999) Signal transduction: Gyrating protein kinases. Curr Biol 9:R364-R367

Summons RE, Jahnke LL, Hope JM, Logan GA (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. Nature 400:554-557

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-1599

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

Tomitani A, Knoll AH, Cavanaugh CM, Ohno T (2006) The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. Proc Natl Acad Sci USA 103:5442-5447

Urbach E, Scanlan DJ, Distel DL, Waterbury JB, Chisholm SW (1998) Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (Cyanobacteria). J Mol Evol 46:188-201

Wagner JR, Zhang J, Brunzelle JS, Vierstra RD, Forest KT (2007) High resolution structure of *Deinococcus* bacteriophytochrome yields new insights into phytochrome architecture and evolution. J Biol Chem 282:12298-12309

Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073-1086

Walsh MM (1992) Microfossils and possible microfossils from the Early Archean Onverwacht Group, Barberton Mountain Land, South Africa. Precambrian Res 54:271-293

Wernersson R, Pedersen AG (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 31:3537-3539

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

Woelfle MA, Ouyang Y, Phanvijhitsiri K, Johnson CH (2004) The adaptive value of circadian clocks: an experimental assessment in cyanobacteria. Curr Biol 14:1481-1486

Xu Y, Mori T, Johnson CH (2003) Cyanobacterial circadian clockwork: roles of KaiA, KaiB and the *kaiBC* promoter in regulating KaiC. EMBO J 22:2117-2126

Yanai I, Wolf YI, Koonin EV (2002) Evolution of gene fusions: horizontal transfer versus independent events. Genome Biol 3:RESEARCH0024

Yang X, Stojkovic EA, Kuk J, Moffat K (2007) Crystal structure of the chromophore binding domain of an unusual bacteriophytochrome, RpBphP3, reveals residues that modulate photoconversion. Proc Natl Acad Sci USA 104:12571-12576

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591

Zhang X, Dong G, Golden SS (2006) The pseudo-receiver domain of CikA regulates the cyanobacterial circadian input pathway. Mol Microbiol 60:658-668

Figure 1


GAF
HisKA
HATPase_c
REC
motif 1
motif 2
motif 3
G-X-G motifs
motif 4

Figure 2

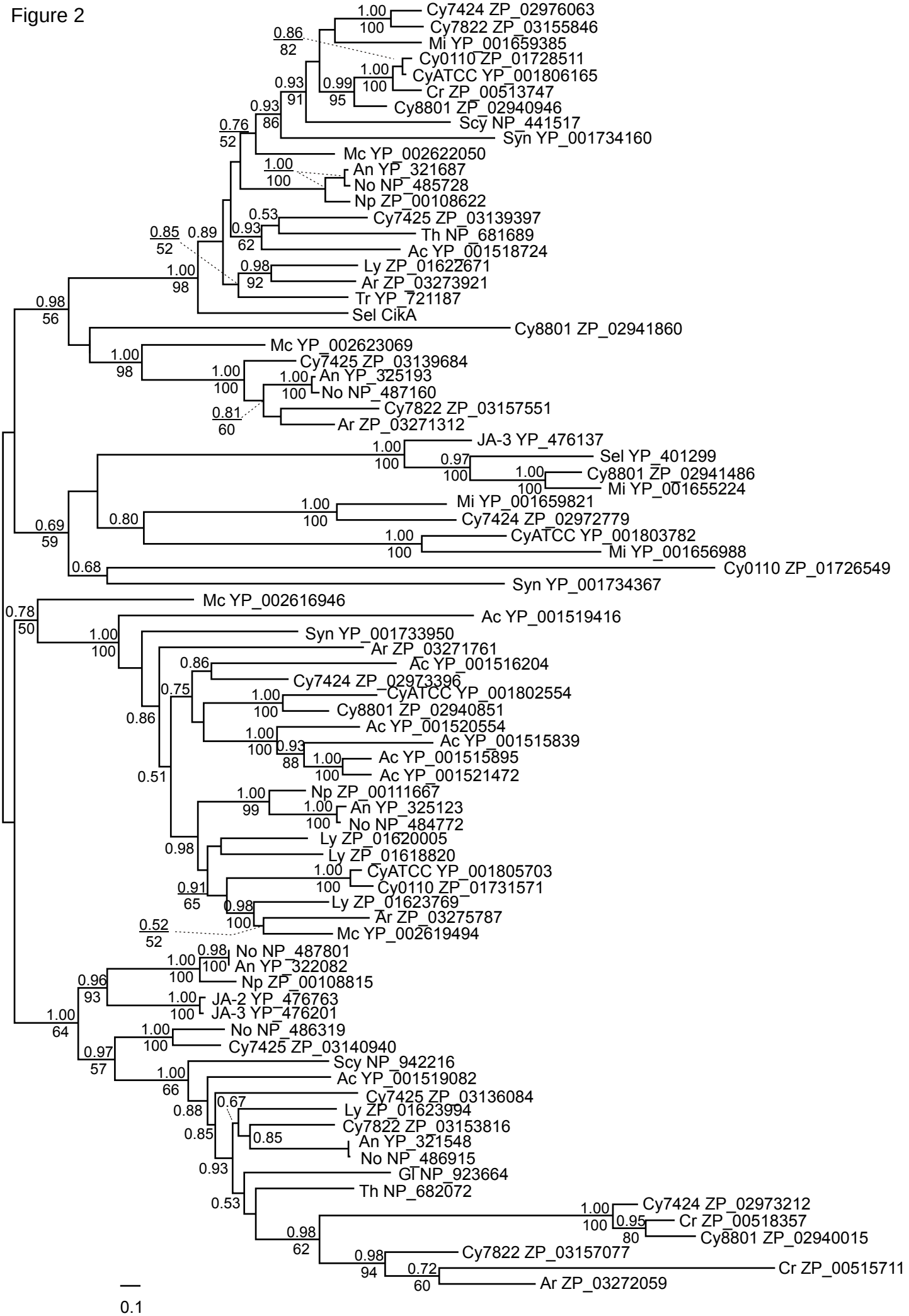

Cy7424 ZP_02976063
Cy7822 ZP_03155846
Mi YP_001659385
Cy0110 ZP_01728511
CyATCC YP_001806165
Cr ZP_00513747
Cy8801 ZP_02940946
Scy NP_441517
Syn YP_001734160
Mc YP_002622050
An YP_321687
No NP_485728
Np ZP_00108622
Cy7425 ZP_03139397
Th NP_681689
Ac YP_001518724
Ly ZP_01622671
Ar ZP_03273921
Tr YP_721187
Sel CikA
Cy8801 ZP_02941860
Mc YP_002623069
Cy7425 ZP_03139684
An YP_325193
No NP_487160
Cy7822 ZP_03157551
Ar ZP_03271312
JA-3 YP_476137
Sel YP_401299
Cy8801 ZP_02941486
Mi YP_001655224
Mi YP_001659821
Cy7424 ZP_02972779
CyATCC YP_001803782
Mi YP_001656988
Cy0110 ZP_01726549
Syn YP_001734367
Mc YP_002616946
Ac YP_001519416
Syn YP_001733950
Ar ZP_03271761
Ac YP_001516204
Cy7424 ZP_02973396
CyATCC YP_001802554
Cy8801 ZP_02940851
Ac YP_001520554
Ac YP_001515839
Ac YP_001515895
Ac YP_001521472
Np ZP_00111667
An YP_325123
No NP_484772
Ly ZP_01620005
Ly ZP_01618820
CyATCC YP_001805703
Cy0110 ZP_01731571
Ly ZP_01623769
Ar ZP_03275787
Mc YP_002619494
No NP_487801
An YP_322082
Np ZP_00108815
JA-2 YP_476763
JA-3 YP_476201
No NP_486319
Cy7425 ZP_03140940
Scy NP_942216
Ac YP_001519082
Cy7425 ZP_03136084
Ly ZP_01623994
Cy7822 ZP_03153816
An YP_321548
No NP_486915
Gl NP_923664
Th NP_682072
Cy7424 ZP_02973212
Cr ZP_00518357
Cy8801 ZP_02940015
Cy7822 ZP_03157077
Cr ZP_00515711
Ar ZP_03272059
A1
A2
A3
A4
A5
0.1

Figure 3

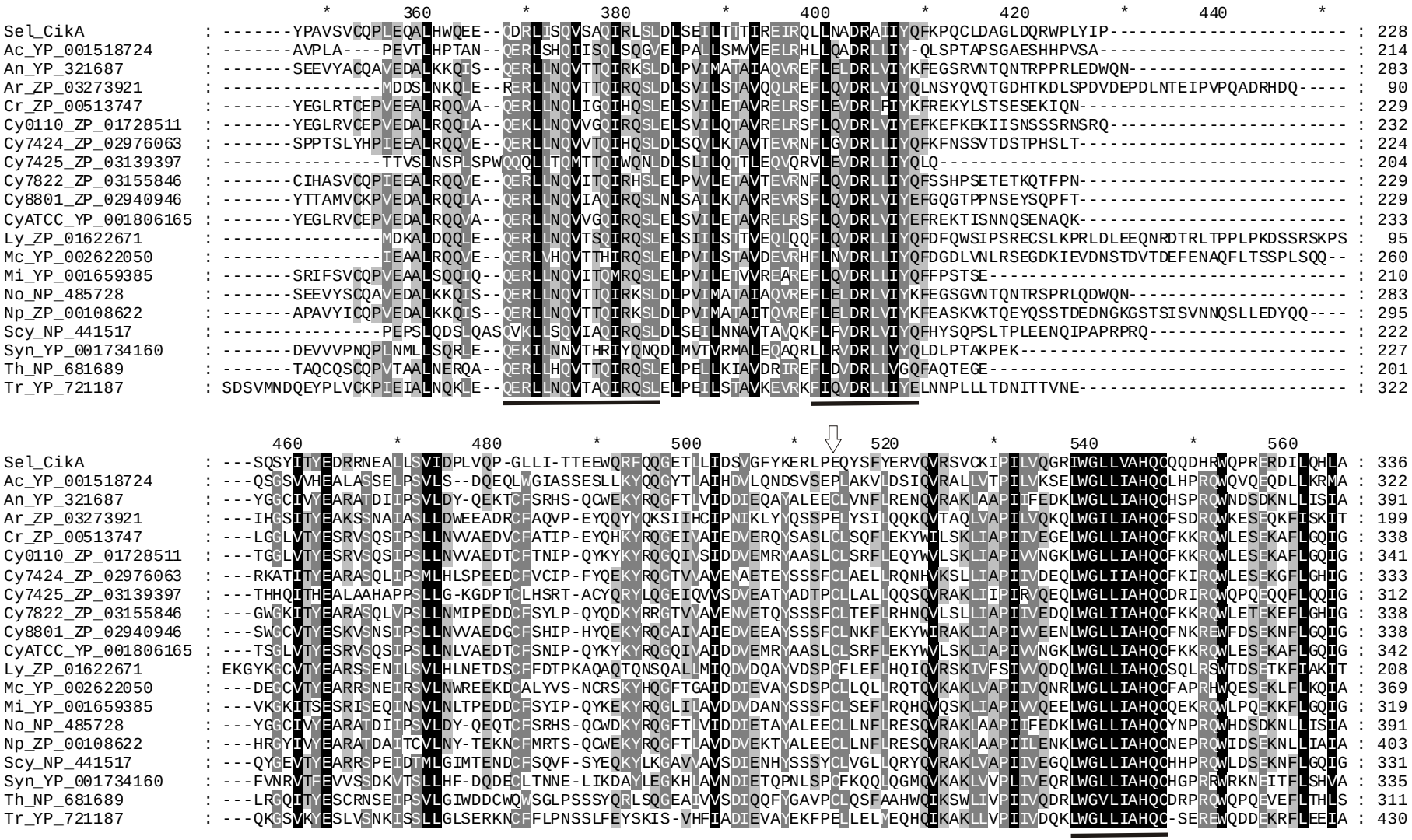

Sel_CikA
Ac_YP_001518724
An_YP_321687
Ar_ZP_03273921
Cr_ZP_00513747
Cy0110_ZP_01728511
Cy7424_ZP_02976063
Cy7425_ZP_03139397
Cy7822_ZP_03155846
Cy8801_ZP_02940946
CyATCC_YP_001806165
Ly_ZP_01622671
Mc_YP_002622050
Mi_YP_001659385
No_NP_485728
Np_ZP_00108622
Scy_NP_441517
Syn_YP_001734160
Th_NP_681689
Tr_YP_721187

Figure 4

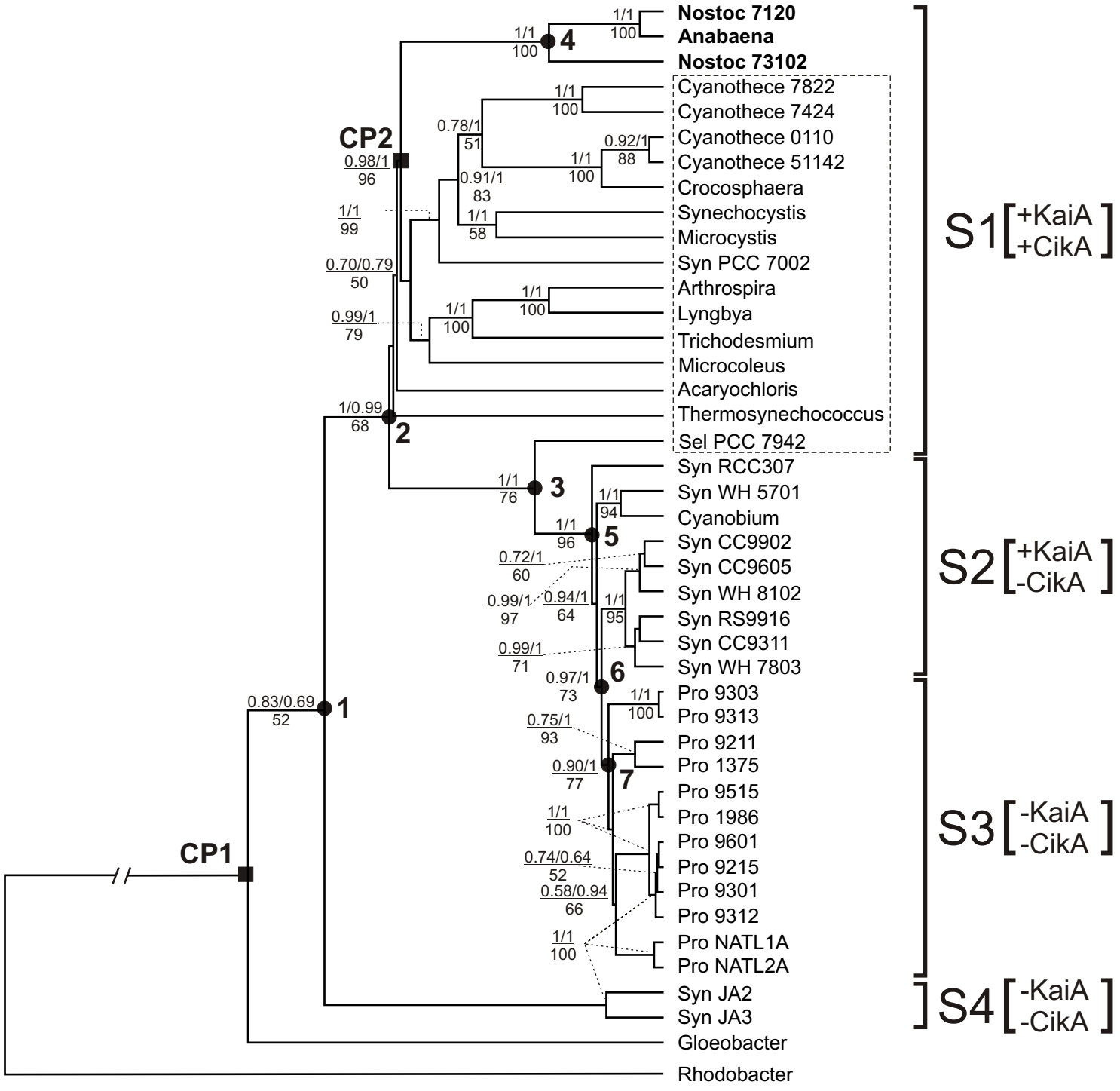

Nostoc 7120
Anabaena
Nostoc 73102
Cyanothece 7822
Cyanothece 7424
Cyanothece 0110
Cyanothece 51142
Crocosphaera
Synechocystis
Microcystis
Syn PCC 7002
Arthrospira
Lyngbya
Trichodesmium
Microcoleus
Acaryochloris
Thermosynechococcus
Sel PCC 7942
Syn RCC307
Syn WH 5701
Cyanobium
Syn CC9902
Syn CC9605
Syn WH 8102
Syn RS9916
Syn CC9311
Syn WH 7803
Pro 9303
Pro 9313
Pro 9211
Pro 1375
Pro 9515
Pro 1986
Pro 9601
Pro 9215
Pro 9301
Pro 9312
Pro NATL1A
Pro NATL2A
Syn JA2
Syn JA3
Gloeobacter
Rhodobacter
CP1
CP2
1
2
3
4
5
6
7
S1 +KaiA +CikA
S2 +KaiA -CikA
S3 -KaiA -CikA
S4 -KaiA -CikA
1

Table S1 – List of the *cikA* sequences used in the study.

Species and strain	Length, bp	GenBank nucleotide sequence accession number	Protein accession number	Designation of sequence in figure
Acaryochloris marina MBIC11017	2208	NC_009925	YP_001518724	Ac YP_001518724
	2790	NC_009925	YP_001515839	Ac YP_001515839
	4545	NC_009925	YP_001516204	Ac YP_001516204
	2667	NC_009925	YP_001519082	Ac YP_001519082
	2652	NC_009925	YP_001515895	Ac YP_001515895
	2715	NC_009925	YP_001520554	Ac YP_001520554
	2790	NC_009928	YP_001521472	Ac YP_001521472
	6123	NC_009925	YP_001519416	Ac YP_001519416
Anabaena variabilis ATCC 29413	2031	NC_007413	YP_321687	An YP_321687
	3825	NC_007413	YP_325193	An YP_325193
	4653	NC_007413	YP_322082	An YP_322082
	5463	NC_007413	YP_321548	An YP_321548
	5019	NC_007413	YP_325123	An YP_325123
Arthrospira maxima CS-328	1878	NZ_ABYK01000018	ZP_03273921	Ar ZP_03273921
	5310	NZ_ABYK01000001	ZP_03271312	Ar ZP_03271312
	3156	NZ_ABYK01000049	ZP_03275787	Ar ZP_03275787
	4176	NZ_ABYK01000005	ZP_03272059	Ar ZP_03272059
	4101	NZ_ABYK01000003	ZP_03271761	Ar ZP_03271761
Crocosphaera watsonii WH 8501	2277	NZ_AADV02000001	ZP_00513747	Cr ZP_00513747

	2391	NZ_AADV02000007	ZP_00515711	Cr ZP_00515711
	1707	NZ_AADV02000134	ZP_00518357	Cr ZP_00518357
Cyanothece sp. ATCC 51142	2298	NC_010546	YP_001806165	CyATCC YP_001806165
	3255	NC_010546	YP_001802554	CyATCC YP_001802554
	2595	NC_010546	YP_001805703	CyATCC YP_001805703
	2235	NC_010546	YP_001803782	CyATCC YP_001803782
Cyanothece sp. CCY0110	2295	NZ_AAXW01000009	ZP_01728511	Cy0110 ZP_01728511
	2529	NZ_AAXW01000057	ZP_01731571	Cy0110 ZP_01731571
	2697	NZ_AAXW01000002	ZP_01726549	Cy0110 ZP_01726549
Cyanothece sp. PCC 7424	2274	NZ_ABOY01000027	ZP_02976063	Cy7424 ZP_02976063
	3858	NZ_ABOY01000007	ZP_02973396	Cy7424 ZP_02973396
	1734	NZ_ABOY01000006	ZP_02973212	Cy7424 ZP_02973212
	1839	NZ_ABOY01000004	ZP_02972779	Cy7424 ZP_02972779
Cyanothece sp. PCC 7425	2202	NZ_ABVJ01000010	ZP_03139397	Cy7425 ZP_03139397
	5934	NZ_ABVJ01000012	ZP_03139684	Cy7425 ZP_03139684
	3918	NZ_ABVJ01000022	ZP_03140940	Cy7425 ZP_03140940
	2664	NZ_ABVJ01000001	ZP_03136084	Cy7425 ZP_03136084
Cyanothece sp. PCC 7822	2289	NZ_ABVE01000004	ZP_03155846	Cy7822 ZP_03155846
	4461	NZ_ABVE01000008	ZP_03157551	Cy7822 ZP_03157551
	4281	NZ_ABVE01000001	ZP_03153816	Cy7822 ZP_03153816
	4842	NZ_ABVE01000006	ZP_03157077	Cy7822 ZP_03157077
Cyanothece sp. PCC 8801	2292	NZ_ABLR01000005	ZP_02940946	Cy8801 ZP_02940946
	2535	NZ_ABLR01000004	ZP_02940851	Cy8801 ZP_02940851

	1329	NZ_ABLR01000008	ZP_02941860	Cy8801 ZP_02941860
	1731	NZ_ABLR01000002	ZP_02940015	Cy8801 ZP_02940015
	2076	NZ_ABLR01000006	ZP_02941486	Cy8801 ZP_02941486
Gloeobacter violaceus PCC 7421	2571	NC_005125	NP_923664	Gl NP_923664
Lyngbya sp. PCC 8106	1926	NZ_AAVU01000031	ZP_01622671	Ly ZP_01622671
	3678	NZ_AAVU01000006	ZP_01620005	Ly ZP_01620005
	3084	NZ_AAVU01000048	ZP_01623769	Ly ZP_01623769
	4884	NZ_AAVU01000001	ZP_01618820	Ly ZP_01618820
	5184	NZ_AAVU01000054	ZP_01623994	Ly ZP_01623994
Microcoleus chthonoplastes PCC 7420	2370	NW_002435256.1	YP_002622050	Mc YP_002622050
	5568	NW_002435259.1	YP_002623069	Mc YP_002623069
	4425	NW_002435233.1	YP_002616946	Mc YP_002616946
	3069	NW_002435248.1	YP_002619494	Mc YP_002619494
Microcystis aeruginosa NIES-843	2196	NC_010296	YP_001659385	Mi YP_001659385
	2238	NC_010296	YP_001656988	Mi YP_001656988
	1992	NC_010296	YP_001655224	Mi YP_001655224
	2229	NC_010296	YP_001659821	Mi YP_001659821
Nostoc punctiforme PCC 73102	2052	NZ_AAAY02000045	ZP_00108622	Np ZP_00108622
	5667	NZ_AAAY02000005	ZP_00111667	Np ZP_00111667
	4680	NZ_AAAY02000034	ZP_00108815	Np ZP_00108815
Nostoc sp. PCC 7120	2031	NC_003272	NP_485728	No NP_485728
	3861	NC_003272	NP_487160	No NP_487160
	3900	NC_003272	NP_486319	No NP_486319

	4938	NC_003272	NP_484772	No NP_484772
	4653	NC_003272	NP_487801	No NP_487801
	5454	NC_003272	NP_486915	No NP_486915
Synechococcus elongatus PCC 7942	2265	NC_007604	YP_399663	Sel CikA
	1914	NC_007604	YP_401299	Sel YP_401299
Synechococcus sp. JA-2-3B'a(2-13)	4581	NC_007776	YP_476763	JA-2 YP_476763
Synechococcus sp. JA-3-3Ab	4572	NC_007775	YP_476201	JA-3 YP_476201
	1965	NC_007775	YP_476137	JA-3 YP_476137
Synechococcus sp. PCC 7002	2283	NC_010475	YP_001734160	Syn YP_001734160
	2325	NC_010475	YP_001733950	Syn YP_001733950
	2826	NC_010475	YP_001734367	Syn YP_001734367
Synechocystis sp. PCC 6803	2253	NC_000911	NP_441517	Scy NP_441517
	3630	NC_005229	NP_942216	Scy NP_942216
Thermosynechococcus elongatus BP-1	2190	NC_004113	NP_681689	Th NP_681689
	4062	NC_004113	NP_682072	Th NP_682072
Trichodesmium erythraeum IMS101	2562	NC_008312	YP_721187	Tr YP_721187

* identical to the sequence of *S. elongatus* PCC 7942

Table S2 – List of 16S rRNA sequences used in the study.

Species and strain	Length, bp	GenBank nucleotide sequence accession number	GI/EMB List	Designation of sequence in the figure*
Acaryochloris marina MBIC11017	1501	NC_009925	gi|158333233:1409149-1410649	Acaryochloris
Anabaena variabilis ATCC 29413	1488	NC_007413	gi|75906225:1002918-1004405	Anabaena
Arthrospira maxima CS-328	1482	NZ_ABYK01000016	gi|209525145:29326-30807	Arthrospira
Crocosphaera watsonii WH 8501	1408	NZ_AADV02000003	gi|67921358:105686-107093	Crocosphaera
Cyanobium sp. PCC 7001	1481	NZ_ABSE01000018	gi|194271858:58330-59810	Cyanobium
Cyanothece sp. ATCC 51142	1489	NC_010546	gi|172034917:3950932-3952420	Cyanothece 51142
Cyanothece sp. CCY0110	1479	NZ_AAXW01000002	gi|126655026:c195826-194348	Cyanothece 0110
Cyanothece sp. PCC 7424	1483	NZ_ABOY01000032	gi|186903529:26610-28092	Cyanothece 7424
Cyanothece sp. PCC 7425†	N/A	N/A	N/A	N/A
Cyanothece sp. PCC 7822	1483	NZ_ABVE01000002	gi|196255427:394510-395992	Cyanothece 7822
Cyanothece sp. PCC 8801†	N/A	N/A	N/A	N/A
Gloeobacter violaceus PCC 7421	1485	NC_005125	gi|37519569:1571231-1572715	Gloeobacter
Lyngbya sp. PCC 8106	1493	NZ_AAVU01000018	gi|119488018:81300-82792	Lyngbya
Microcoleus chthonoplastes PCC 7420	1488	NZ_ABRS01000062	gi|194018314:60739-62226	Microcoleus

Microcystis aeruginosa NIES-843	1477	NC_010296	gi|166085114:1885814-1887290	Microcystis
Nostoc punctiforme PCC 73102	1410	AF027655	gi|186463002:2021489-2022977	Nostoc 73102
Nostoc sp. PCC 7120	1489	NC_003272	gi|17227497:2375734-2377222	Nostoc 7120
Prochlorococcus marinus str. AS 9601	1465	NC_008816	gi|123967536:323018-324482	Pro 9601
Prochlorococcus marinus str. MIT 9211	1465	NC_009976	gi|159902540:342283-343747	Pro 9211
Prochlorococcus marinus str. MIT 9215	1465	NC_009840	gi|157412338:319913-321377	Pro 9215
Prochlorococcus marinus str. MIT 9301	1465	NC_009091	gi|126695337:322587-324084	Pro 9301
Prochlorococcus marinus str. MIT 9303	1401	NC_008820	gi|124021714:243682-245082	Pro 9303
Prochlorococcus marinus str. MIT 9312	1465	NC_007577	gi|78778385:317580-319044	Pro 9312
Prochlorococcus marinus str. MIT 9313	1465	NC_005071	gi|33862273:208864-210328	Pro 9313
Prochlorococcus marinus str. MIT 9515	1465	NC_008817	gi|123965234:332422-333886	Pro 9515
Prochlorococcus marinus str. NATL1A	1465	NC_008819	gi|124024712:385837-387301	Pro NATL1A
Prochlorococcus marinus str. NATL2A	1394	NC_007335	gi|162958048:370879-372272	Pro NATL2A
Prochlorococcus marinus subsp. *marinus* str. CCMP1375	1465	NC_005042	gi|33239452:353331-354795	Pro 1375
Prochlorococcus marinus subsp. *pastoris* str. CCMP1986	1465	NC_005072	gi|33860560:313061-314525	Pro 1986

Synechococcus elongatus PCC 7942	1490	NC_007604	gi|81298811:568589-570078	Sel PCC 7942
Synechococcus sp. CC9311	1477	NC_008319	gi|113952711:529961-531437	Syn CC9311
Synechococcus sp. CC9605	1440	NC_007516	gi|78211558:473828-475267	Syn CC9605
Synechococcus sp. CC9902	1479	NC_007513	gi|78183584:373319-374797	Syn CC9902
Synechococcus sp. JA-2-3B'a(2-13)	1323	NC_007776	gi|86607503:1447574-1448896	Syn JA2
Synechococcus sp. JA-3-3Ab	1324	NC_007775	gi|86604733:2310397-2311720	Syn JA3
Synechococcus sp. PCC 7002	1452	NC_010475	gi|170076636:1461361-1462812	Syn PCC 7002
Synechococcus sp. RCC307	1498	NC_009482	gi|148241099:348737-350234	Syn RCC307
Synechococcus sp. RS9916	1439	NZ_AAUA01000001	gi|116072916:c817919-816481	Syn RS9916
Synechococcus sp. WH 5701	1440	AY172832	gi|87303064:1-2054	Syn WH 5701
Synechococcus sp. WH 7803	1497	NC_009481	gi|148238336:534540-536036	Syn WH 7803
Synechococcus sp. WH 8102	1464	NC_005070	gi|33864539:2083060-2081597	Syn WH 8102
Synechocystis sp. PCC 6803	1489	NC_000911	gi|16329170:3325053-3326541	Synechocystis
Thermosynechococcus elongatus BP-1	1491	NC_004113	gi|22297544:2335243-2336733	Thermosynechococcus
Trichodesmium erythraeum IMS101	1482	NC_008312	gi|113473942:3137164-3138645	Trichodesmium
Rhodobacter sphaeroides ATCC 17029	1460	NC_009049	gi|126460778:87772-89231	Rhodobacter

[*] refers to the concatenated 16S rRNA and 23S rRNA sequences

[†] either absent in the GenBank or not used because of absence of the respective 23S rRNA

Table S3 – List of the 23S rRNA sequences used in the study.

Species and strain	Length, bp	GenBank nucleotide sequence accession number	GI/EMB List
Acaryochloris marina MBIC11017	2880	NC_009925	gi|158333233:5638205-5641084
Anabaena variabilis ATCC 29413	2830	NC_007413	gi|75906225:3896066-3898895
Arthrospira maxima CS-328	2882	NZ_ABYK01000016.1	gi|209494330:31287-34168
Crocosphaera watsonii WH 8501	2881	NZ_AADV02000003.1	gi|67856470:107505-110384
Cyanobium sp. PCC 7001	2886	NZ_ABSE01000005.1	gi|194271845:94-2979
Cyanothece sp. ATCC 51142	2875	NC_010546	gi|172034917:3952751-3955625
Cyanothece sp. CCY 0110	2862	NZ_AAXW01000002	gi|126655026:c193919-191058
Cyanothece sp. PCC 7424	2886	NZ_ABOY01000032.1	gi|186689375:28427-31312
Cyanothece sp. PCC 7425*	N/A	N/A	N/A
Cyanothece sp. PCC 7822	2882	NZ_ABVE01000004	gi|196255427:c391299-394180
Cyanothece sp. PCC 8801*	N/A	N/A	N/A
Gloeobacter violaceus PCC 7421	2781	NC_005125	gi|37519569:c1570772-1567992
Lyngbya sp. PCC 8106	2878	NZ_AAVU01000018.1	gi|119488018:83267-86144
Microcoleus chthonoplastes PCC 7420	2884	NZ_ABRS01000062.1	gi|194018314:62741-65624

Microcystis aeruginosa NIES-843	2878	NC_010296	gi|166362741:1887656-1890533
Nostoc punctiforme PCC 73102	2887	NC_010628	gi|186680550:2023598-2026484
Nostoc sp. PCC 7120	2828	NC_003272	gi|17227497:2377736-2382034
Prochlorococcus marinus str. AS 9601	2874	NC_008816	gi|123967536:325023-327896
Prochlorococcus marinus str. MIT 9211	2876	NC_009976	gi|159902540:344441-347316
Prochlorococcus marinus str. MIT 9215	2874	NC_009840	gi|157412338:321917-324790
Prochlorococcus marinus str. MIT 9301	2876	NC_009091	gi|126695337:324618-327493
Prochlorococcus marinus str. MIT 9303	2873	NC_008820	gi|124021714:245970-248842
Prochlorococcus marinus str. MIT 9312	2874	NC_007577	gi|78778385:319585-322458
Prochlorococcus marinus str. MIT 9313	2876	NC_005071	gi|33862273:211158-214033
Prochlorococcus marinus str. MIT 9515	2874	NC_008817	gi|123965234:334434-337307
Prochlorococcus marinus str. NATL1A	2874	NC_008819	gi|124024712:387935-390808
Prochlorococcus marinus str. NATL2A	2875	NC_007335	gi|162958048:372962-375836
Prochlorococcus marinus subsp. *marinus* str. CCMP1375	2873	NC_005042	gi|33239452:355462-358334
Prochlorococcus marinus subsp. *pastoris*	2875	NC_005072	gi|33860560:315074-317948

str. CCMP1986			
Synechococcus elongatus PCC 7942	2878	CP000100.1	gi|81298811:570624-573501
Synechococcus sp. CC9311	2864	NC_008319	gi|113952711:532200-535063
Synechococcus sp. CC9605	2866	NC_007516	gi|78211558:476063-478928
Synechococcus sp. CC9902	2866	NC_007513	gi|78183584:375574-378439
Synechococcus sp. JA-2-3B'a(2-13)	2808	NC_007776	gi|86607503:1449625-1452432
Synechococcus sp. JA-3-3Ab	2807	NC_007775	gi|86604733:c1109571-1106765
Synechococcus sp. PCC 7002	2811	NC_010475	gi|170076636:1463345-1466155
Synechococcus sp. RCC307	2866	NC_009482	gi|148241099:350832-353697
Synechococcus sp. RS9916	2865	NZ_AAUA01000004.1	gi|116072916:c812870-815734
Synechococcus sp. WH 5701	2884	NZ_AANO01000001.1	gi|87282231:2913-5796
Synechococcus sp. WH 7803	2866	NC_009481	gi|148238336:536808-539673
Synechococcus sp. WH 8102	2865	NC_005070	gi|33864539:c1870985-1873849
Synechocystis sp. PCC 6803	2883	NC_000911	gi|16329170:c2451721-2448839
Thermosynechococcus elongatus BP-1	2871	NC_004113	gi|22297544:c2334822-2331207
Trichodesmium erythraeum IMS101	2881	NC_008312	gi|113473942:3139115-3141995
Rhodobacter sphaeroides ATCC 17029	2918	NC_009049	gi|126460778:89899-92816

* either absent in the GenBank or not used because of absence of the respective 16S rRNA

Figure S1

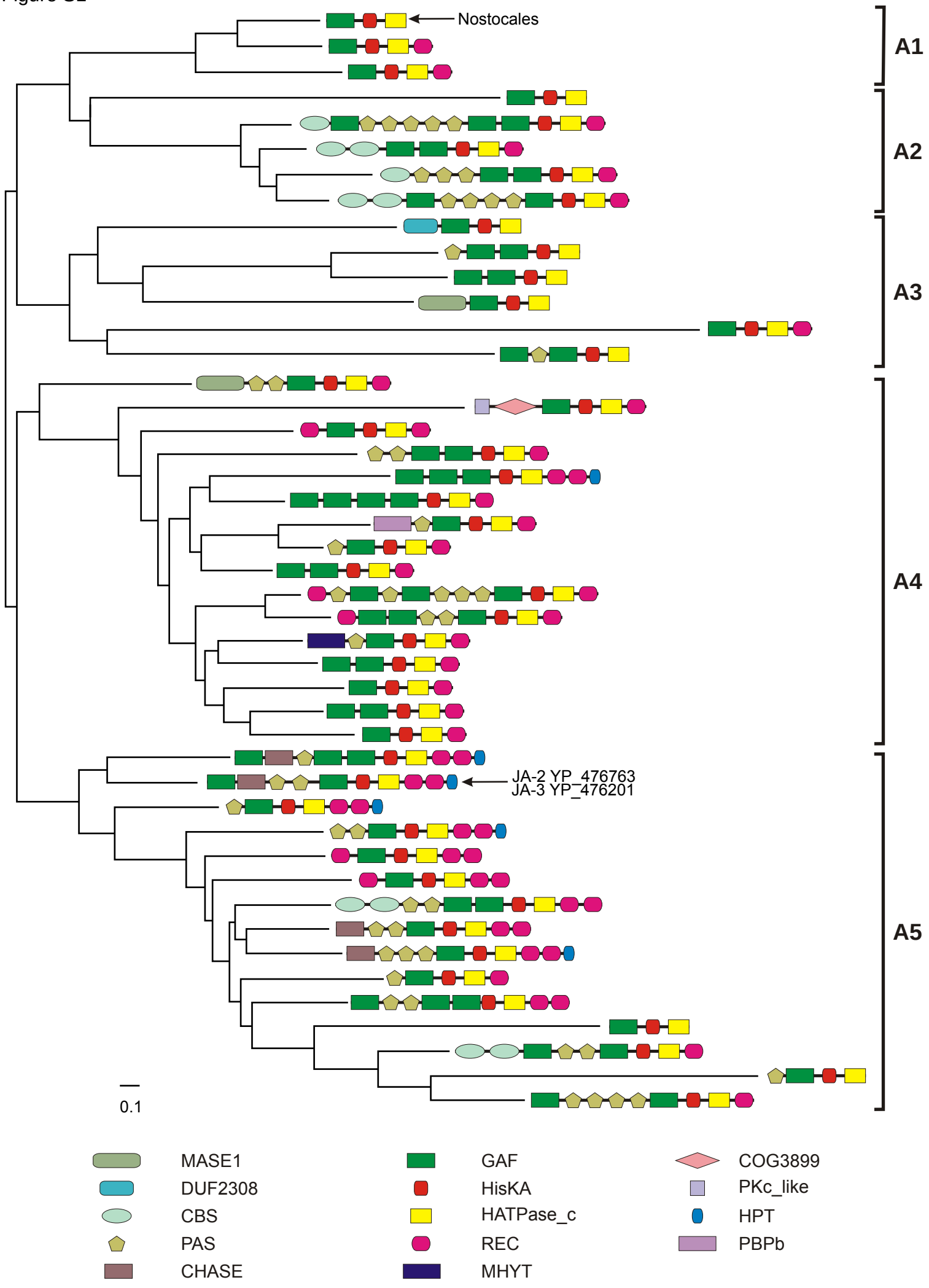
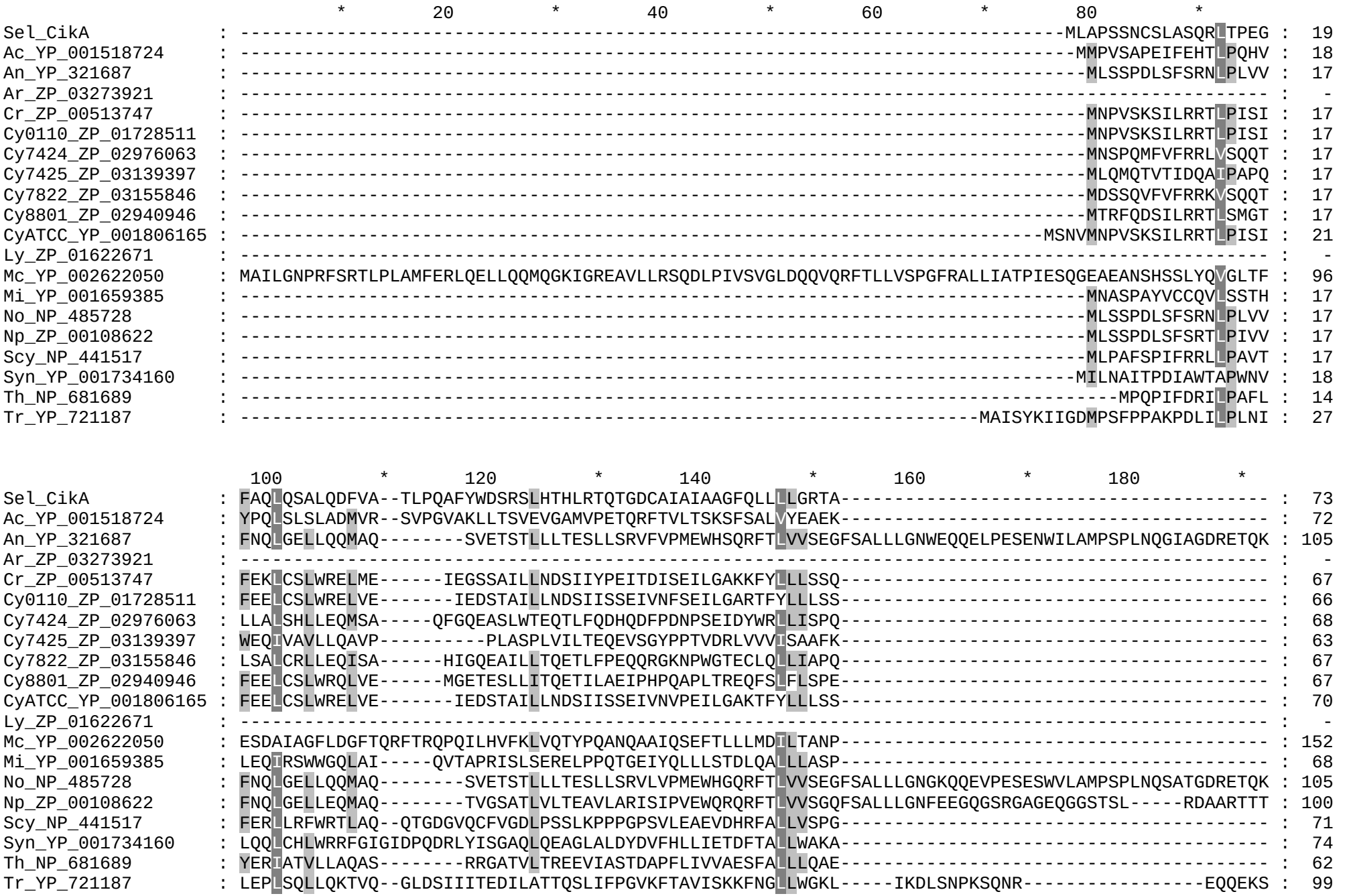
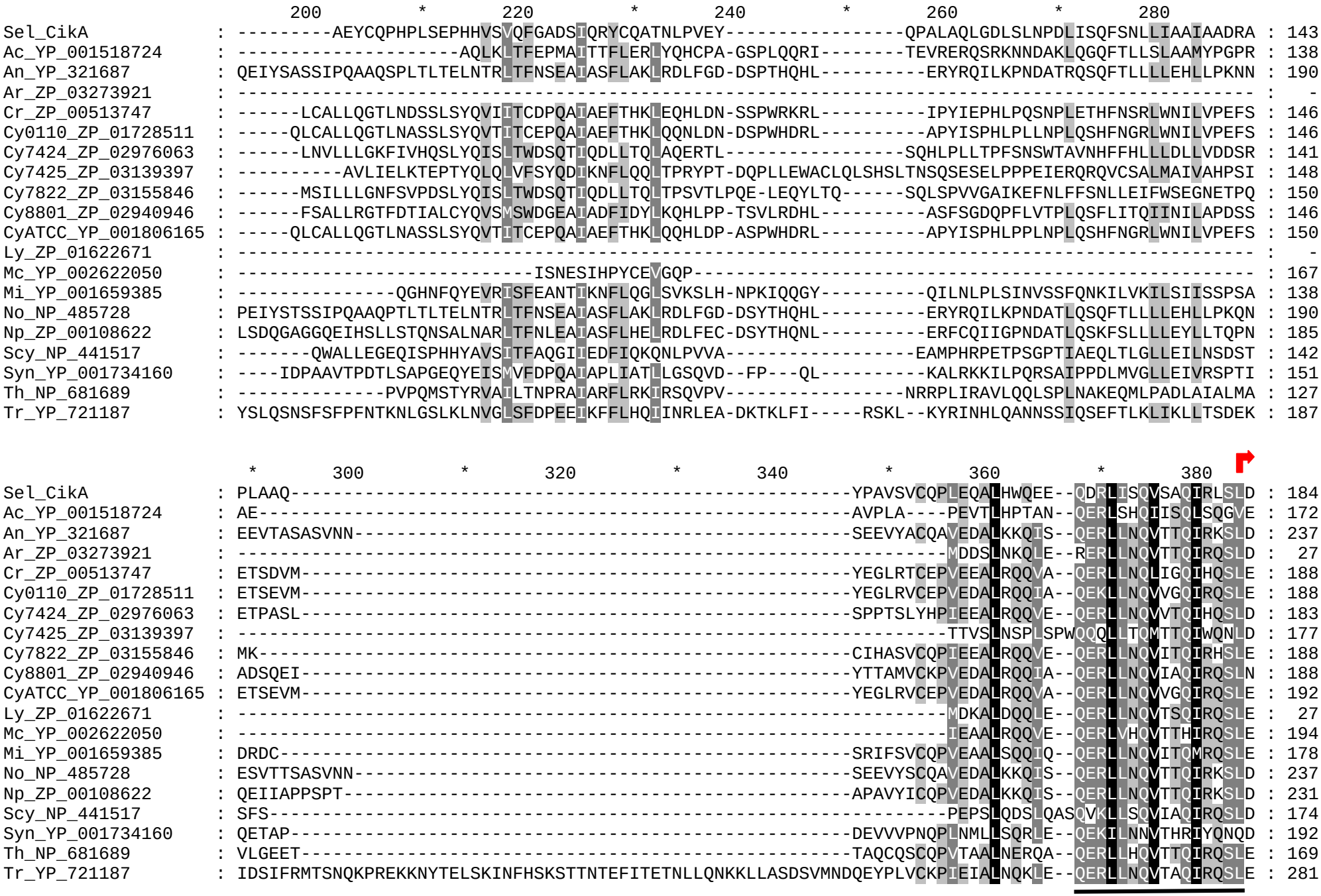
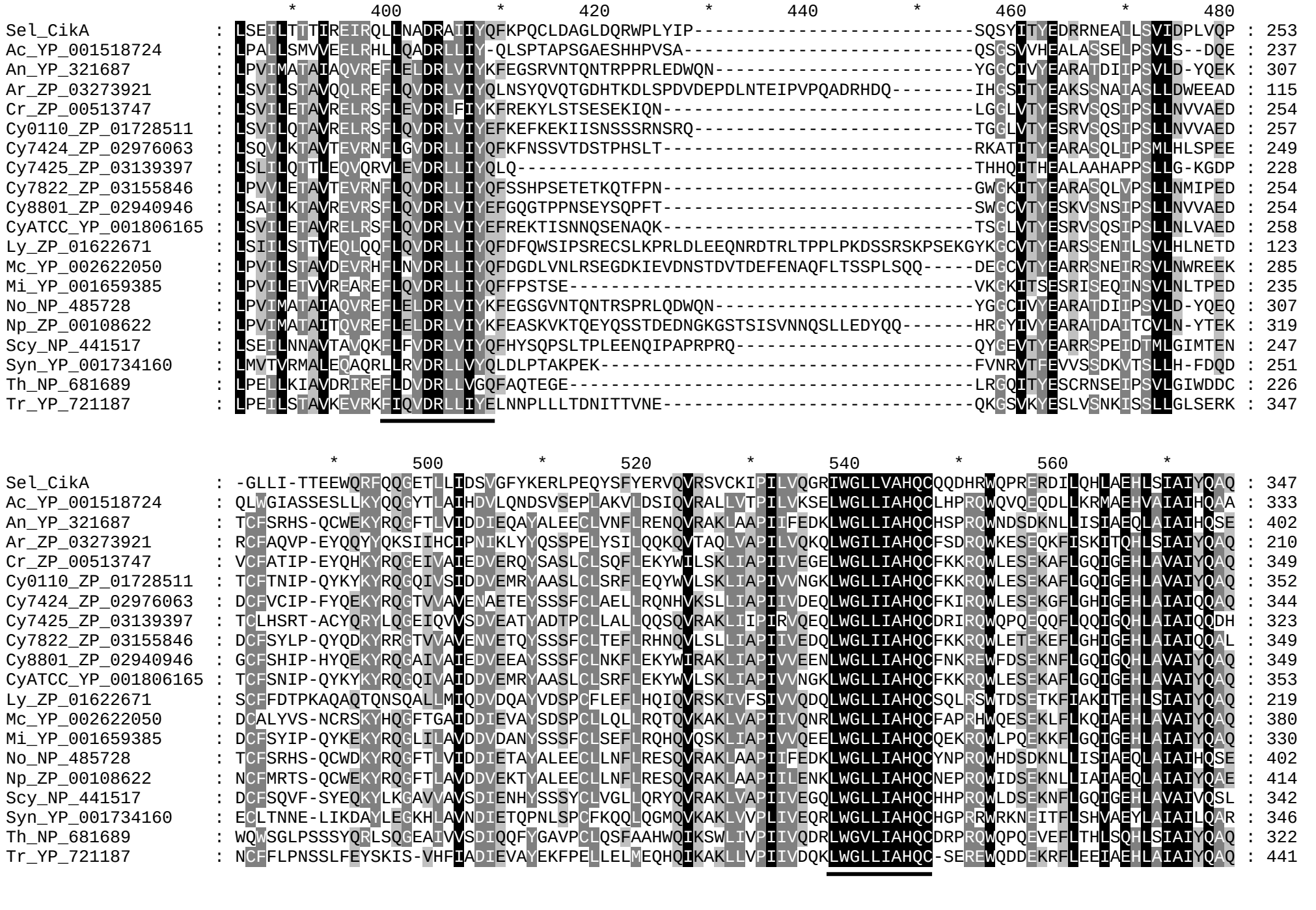
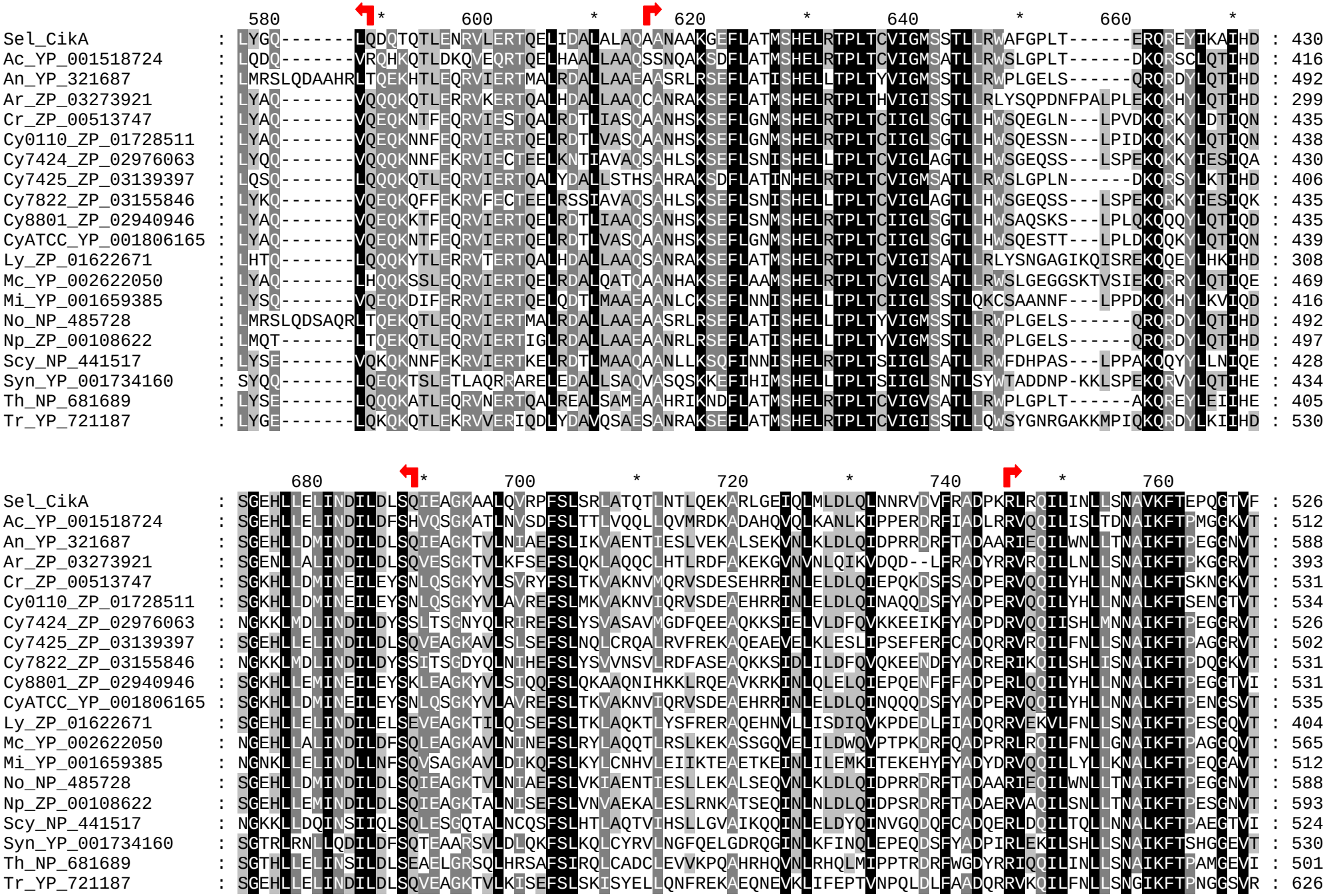

Nostocales
A1
A2
A3
A4
A5
JA-2 YP_476763
JA-3 YP_476201
0.1
MASE1
DUF2308
CBS
PAS
CHASE
GAF
HisKA
HATPase_c
REC
MHYT
COG3899
PKc_like
HPT
PBPb

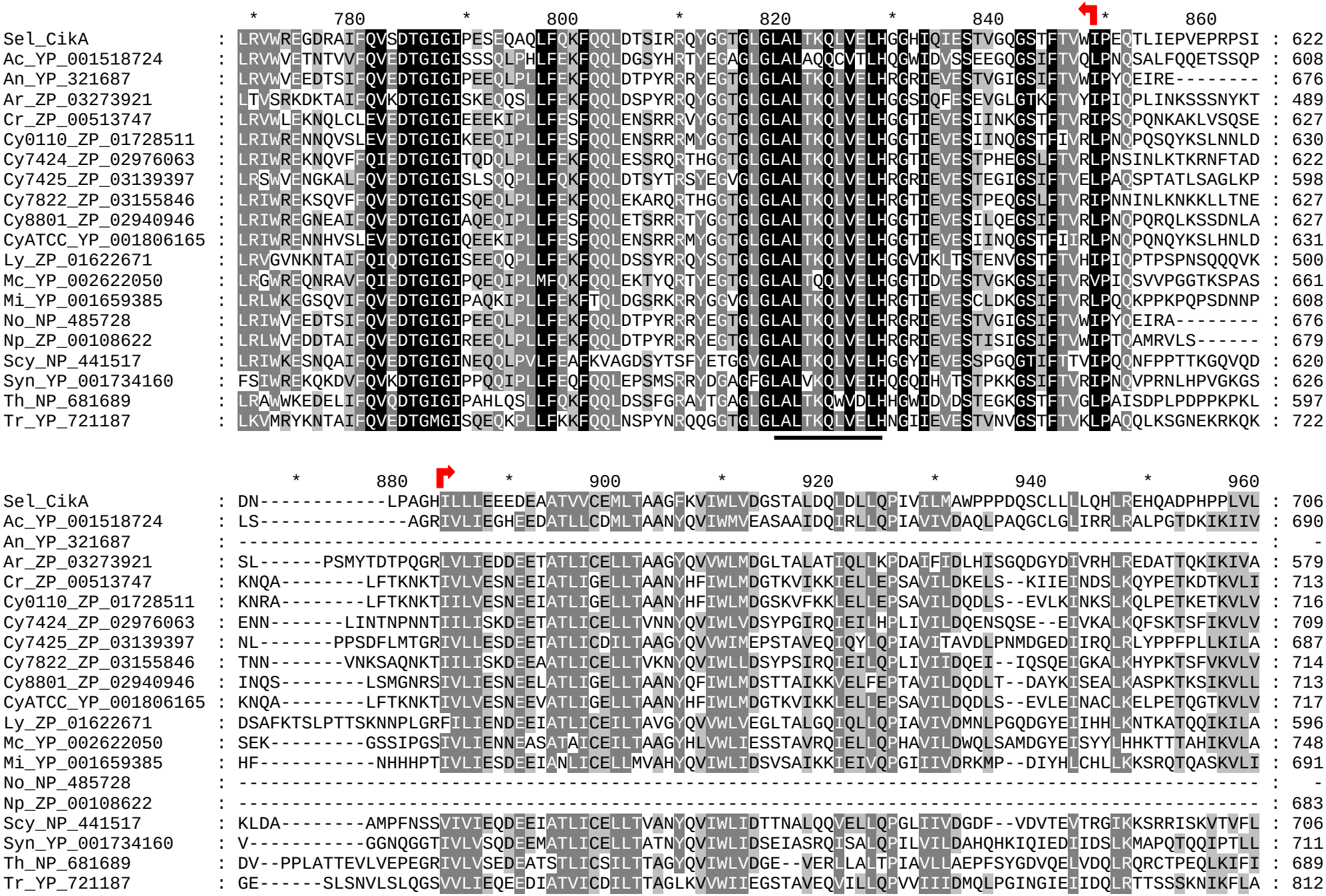
Figure S2


Sel_CikA
Ac_YP_001518724
An_YP_321687
Ar_ZP_03273921
Cr_ZP_00513747
Cy0110_ZP_01728511
Cy7424_ZP_02976063
Cy7425_ZP_03139397
Cy7822_ZP_03155846
Cy8801_ZP_02940946
CyATCC_YP_001806165
Ly_ZP_01622671
Mc_YP_002622050
Mi_YP_001659385
No_NP_485728
Np_ZP_00108622
Scy_NP_441517
Syn_YP_001734160
Th_NP_681689
Tr_YP_721187


200 * 220 * 240 * 260 * 280
Sel_CikA : ---------AEYCQPHPLSEPHHVSVQFGADSIQRYCQATNLPVEY----------QPALAQLGDLSLNPDLISQFSNLLIAAIAADRA : 143
Ac_YP_001518724 : --------------------AQLKLTFEPMAITTFLERLYQHCPA-GSPLQQRI--------TEVRERQSRKNNDAKLQGQFTLLSLAAMYPGPR : 138
An_YP_321687 : QEIYSASSIPQAAQSPLTLTELNTRLTFNSEAIASFLAKLRDLFGD-DSPTHQHL----------ERYRQILKPNDATRQSQFTLLLLEHLLPKNN : 190
Ar_ZP_03273921 : ---------- : -
Cr_ZP_00513747 : ------LCALLQGTLNDSSLSYQVIITCDPQAIAEFTHKLEQHLDN-SSPWRKRL----------IPYIEPHLPQSNPLETHFNSRLWNILVPEFS : 146
Cy0110_ZP_01728511 : -----QLCALLQGTLNASSLSYQVTITCEPQAIAEFTHKLQQNLDN-DSPWHDRL----------APYISPHLPLLNPLQSHFNGRLWNILVPEFS : 146
Cy7424_ZP_02976063 : ------LNVLLLGKFIVHQSLYQISLTWDSQTIQDLLTQLAQERTL----------SQHLPLLTPFSNSWTAVNHFFHLLLDLLVDDSR : 141
Cy7425_ZP_03139397 : ----------AVLIELKTEPTYQLQLVFSYQDIKNFLQQLTPRYPT-DQPLLEWACLQLSHSLTNSQSESELPPPEIERQRQVCSALMAIVAHPSI : 148
Cy7822_ZP_03155846 : ------MSILLLGNFSVPDSLYQISLTWDSQTIQDLLTQLTPSVTLPQE-LEQYLTQ------SQLSPVVGAIKEFNLFFSNLLEIFWSEGNETPQ : 150
Cy8801_ZP_02940946 : ------FSALLRGTFDTIALCYQVSMSWDGEAIADFIDYLKQHLPP-TSVLRDHL----------ASFSGDQPFLVTPLQSFLITQIINILAPDSS : 146
CyATCC_YP_001806165 : -----QLCALLQGTLNASSLSYQVTITCEPQAIAEFTHKLQQHLDP-ASPWHDRL----------APYISPHLPPLNPLQSHFNGRLWNILVPEFS : 150
Ly_ZP_01622671 : ---------- : -
Mc_YP_002622050 : ----------ISNESIHPYCEVGQP---------- : 167
Mi_YP_001659385 : --------------QGHNFQYEVRISFEANTIKNFLQGLSVKSLH-NPKIQQGY----------QILNLPLSINVSSFQNKILVKILSIISSPSA : 138
No_NP_485728 : PEIYSTSSIPQAAQPTLTLTELNTRLTFNSEAIASFLAKLRDLFGD-DSYTHQHL----------ERYRQILKPNDATLQSQFTLLLLEHLLPKQN : 190
Np_ZP_00108622 : LSDQGAGGQEIHSLLSTQNSALNARLTFNLEAIASFLHELRDLFEC-DSYTHQNL----------ERFCQIIGPNDATLQSKFSLLLLEYLLTQPN : 185
Scy_NP_441517 : -------QWALLEGEQISPHHYAVSITFAQGIIEDFIQKQNLPVVA----------EAMPHRPETPSGPTIAEQLTLGLLEILNSDST : 142
Syn_YP_001734160 : ----IDPAAVTPDTLSAPGEQYEISMVFDPQAIAPLIATLLGSQVD--FP---QL----------KALRKKILPQRSAIPPDLMVGLLEIVRSPTI : 151
Th_NP_681689 : -------------PVPQMSTYRVAILTNPRAIARFLRKIRSQVPV----------NRRPLIRAVLQQLSPLNAKEQMLPADLAIALMA : 127
Tr_YP_721187 : YSLQSNSFSFPFNTKNLGSLKLNVGLSFDPEEIKFFLHQIINRLEA-DKTKLFI-----RSKL--KYRINHLQANNSSIQSEFTLKLIKLLTSDEK : 187
* 300 * 320 * 340 * 360 * 380
Sel_CikA : PLAAQ----------YPAVSVCQPLEQALHWQEE--QDRLISQVSAQIRLSLD : 184
Ac_YP_001518724 : AE----------AVPLA----PEVTLHPTAN--QERLSHQIISQLSQGVE : 172
An_YP_321687 : EEVTASASVNN----------SEEVYACQAVEDALKKQIS--QERLLNQVTTQIRKSLD : 237
Ar_ZP_03273921 : ----------MDDSLNKQLE--RERLLNQVTTQIRQSLD : 27
Cr_ZP_00513747 : ETSDVM----------YEGLRTCEPVEEALRQQVA--QERLLNQLIGQIHQSLE : 188
Cy0110_ZP_01728511 : ETSEVM----------YEGLRVCEPVEDALRQQIA--QEKLLNQVVGQIRQSLE : 188
Cy7424_ZP_02976063 : ETPASL----------SPPTSLYHPIEEALRQQVE--QERLLNQVVTQIHQSLD : 183
Cy7425_ZP_03139397 : ----------TTVSLNSPLSPWQQQLLTQMTTQIWQNLD : 177
Cy7822_ZP_03155846 : MK----------CIHASVCQPIEEALRQQVE--QERLLNQVITQIRHSLE : 188
Cy8801_ZP_02940946 : ADSQEI----------YTTAMVCKPVEDALRQQIA--QERLLNQVIAQIRQSLN : 188
CyATCC_YP_001806165 : ETSEVM----------YEGLRVCEPVEDALRQQVA--QERLLNQVVGQIRQSLE : 192
Ly_ZP_01622671 : ----------MDKALDQQLE--QERLLNQVTSQIRQSLE : 27
Mc_YP_002622050 : ----------IEAALRQQVE--QERLVHQVTTHIRQSLE : 194
Mi_YP_001659385 : DRDC----------SRIFSVCQPVEAALSQQIQ--QERLLNQVITQMRQSLE : 178
No_NP_485728 : ESVTTSASVNN----------SEEVYSCQAVEDALKKQIS--QERLLNQVTTQIRKSLD : 237
Np_ZP_00108622 : QEIIAPPSPT----------APAVYICQPVEDALKKQIS--QERLLNQVTTQIRKSLD : 231
Scy_NP_441517 : SFS----------PEPSLQDSLQASQVKLLSQVIAQIRQSLD : 174
Syn_YP_001734160 : QETAP----------DEVVVPNQPLNMLLSQRLE--QEKILNNVTHRIYQNQD : 192
Th_NP_681689 : VLGEET----------TAQCQSCQPVTAALNERQA--QERLLHQVTTQIRQSLE : 169
Tr_YP_721187 : IDSIFRMTSNQKPREKKNYTELSKINFHSKSTTNTEFITETNLLQNKKLLASDSVMNDQEYPLVCKPIEIALNQKLE--QERLLNQVTAQIRQSLE : 281


Sel_CikA
Ac_YP_001518724
An_YP_321687
Ar_ZP_03273921
Cr_ZP_00513747
Cy0110_ZP_01728511
Cy7424_ZP_02976063
Cy7425_ZP_03139397
Cy7822_ZP_03155846
Cy8801_ZP_02940946
CyATCC_YP_001806165
Ly_ZP_01622671
Mc_YP_002622050
Mi_YP_001659385
No_NP_485728
Np_ZP_00108622
Scy_NP_441517
Syn_YP_001734160
Th_NP_681689
Tr_YP_721187


580 * 600 * 620 * 640 * 660 *
Sel_CikA : LYGQ-------LQDQTQTLENRVLERTQELIDALALAQAANAAKGEFLATMSHELRTPLTCVIGMSSTLLRWAFGPLT------ERQREYIKAIHD : 430
Ac_YP_001518724 : LQDQ-------VRQHKQTLDKQVEQRTQELHAALLAAQSSNQAKSDFLATMSHELRTPLTCVIGMSATLLRWSLGPLT------DKQRSCLQTIHD : 416
An_YP_321687 : LMRSLQDAAHRLTQEKHTLEQRVIERTMALRDALLAAEAASRLRSEFLATISHELLTPLTYVIGMSSTLLRWPLGELS------QRQRDYLQTIHD : 492
Ar_ZP_03273921 : LYAQ-------VQQQKQTLERRVKERTQALHDALLAAQCANRAKSEFLATMSHELRTPLTHVIGISSTLLRLYSQPDNFPALPLEKQKHYLQTIHD : 299
Cr_ZP_00513747 : LYAQ-------VQEQKNTFEQRVIESTQALRDTLIASQAANHSKSEFLGNMSHELRTPLTCIIGLSGTLLHWSQEGLN---LPVDKQRKYLDTIQN : 435
Cy0110_ZP_01728511 : LYAQ-------VQEQKNNFEQRVIERTQELRDTLVASQAANHSKSEFLGNMSHELRTPLTCIIGLSGTLLHWSQESSN---LPIDKQKKYLQTIQN : 438
Cy7424_ZP_02976063 : LYQQ-------VQQQKNNFEKRVIECTEELKNTIAVAQSAHLSKSEFLSNISHELLTPLTCVIGLAGTLLHWSGEQSS---LSPEKQKKYIESIQA : 430
Cy7425_ZP_03139397 : LQSQ-------LQQQKQTLEQRVIERTQALYDALLSTHSAHRAKSDFLATINHELRTPLTCVIGMSATLLRWSLGPLN------DKQRSYLKTIHD : 406
Cy7822_ZP_03155846 : LYKQ-------VQEQKQFFEKRVFECTEELRSSIAVAQSAHLSKSEFLSNISHELLTPLTCVIGLAGTLLHWSGEQSS---LSPEKQRKYIESIQK : 435
Cy8801_ZP_02940946 : LYAQ-------VQEQKKTFEQRVIERTQELRDTLIAAQSANHSKSEFLSNMSHELRTPLTCIIGLSGTLLHWSAQSKS---LPLQKQQQYLQTIQD : 435
CyATCC_YP_001806165 : LYAQ-------VQEQKNTFEQRVIERTQELRDTLVASQAANHSKSEFLGNMSHELRTPLTCIIGLSGTLLHWSQESTT---LPLDKQQKYLQTIQN : 439
Ly_ZP_01622671 : LHTQ-------LQQQKYTLERRVTERTQALHDALLAAQSANRAKSEFLATMSHELRTPLTCVIGISATLLRLYSNGAGIKQISREKQQEYLHKIHD : 308
Mc_YP_002622050 : LYAQ-------LHQQKSSLEQRVIERTQELRDALQATQAANHAKSEFLAAMSHELRTPLTCVIGLSATLLRWSLGEGGSKTVSIEKQRRYLQTIQE : 469
Mi_YP_001659385 : LYSQ-------VQEQKDIFERRVIERTQELQDTLMAAEAANLCKSEFLNNISHELLTPLTCIIGLSSTLQKCSAANNF---LPPDKQKHYLKVIQD : 416
No_NP_485728 : LMRSLQDSAQRLTQEKQTLEQRVIERTMALRDALLAAEAASRLRSEFLATISHELLTPLTYVIGMSSTLLRWPLGELS------QRQRDYLQTIHD : 492
Np_ZP_00108622 : LMQT-------LTQEKQTLEQRVIERTIGLRDALLAAEAANRLRSEFLATISHELLTPLTYVIGMSSTLLRWPLGELS------QRQRDYLQTIHD : 497
Scy_NP_441517 : LYSE-------VQKQKNNFEKRVIERTKELRDTLMAAQAANLLKSQFINNISHELRTPLTSIIGLSATLLRWFDHPAS---LPPAKQQYYLLNIQE : 428
Syn_YP_001734160 : SYQQ-------LQEQKTSLETLAQRRARELEDALLSAQVASQSKKEFIHIMSHELLTPLTSIIGLSNTLSYWTADDNP-KKLSPEKQRVYLQTIHE : 434
Th_NP_681689 : LYSE-------LQQQKATLEQRVNERTQALREALSAMEAAHRIKNDFLATMSHELRTPLTCVIGVSATLLRWPLGPLT------AKQREYLEIIHE : 405
Tr_YP_721187 : LYGE-------LQKQKQTLEKRVVERIQDLYDAVQSAESANRAKSEFLATMSHELRTPLTCVIGISSTLLQWSYGNRGAKKMPIQKQRDYLKIIHD : 530
680 * 700 * 720 * 740 * 760
Sel_CikA : SGEHLLELINDILDLSQIEAGKAALQVRPFSLSRLATQTLNTLQEKARLGEIQLMLDLQLNNRVDVFRADPKRLRQILINLLSNAVKFTEPQGTVF : 526
Ac_YP_001518724 : SGEHLLELINDILDFSHVQSGKATLNVSDFSLTTLVQQLLQVMRDKADAHQVQLKANLKIPPERDRFIADLRRVQQILISLTDNAIKFTPMGGKVT : 512
An_YP_321687 : SGEHLLDMINDILDLSQIEAGKTVLNIAEFSLIKVAENTIESLVEKALSEKVNLKLDLQIDPRRDRFTADAARIEQILWNLLTNAIKFTPEGGNVT : 588
Ar_ZP_03273921 : SGENLLALINDILDLSQVESGKTVLKFSEFSLQKLAQQCLHTLRDFAKEKGVNVNLQIKVDQD--LFRADYRRVRQILLNLLSNAIKFTPKGGRVT : 393
Cr_ZP_00513747 : SGKHLLDMINEILEYSNLQSGKYVLSVRYFSLTKVAKNVMQRVSDESEHRRINLELDLQIEPQKDSFSADPERVQQILYHLLNNALKFTSKNGKVT : 531
Cy0110_ZP_01728511 : SGKHLLDMINEILEYSNLQSGKYVLAVREFSLMKVAKNVIQRVSDEAEHRRINLELDLQINAQQDSFYADPERVQQILYHLLNNALKFTSENGTVT : 534
Cy7424_ZP_02976063 : NGKKLMDLINDILDYSSLTSGNYQLRIREFSLYSVASAVMGDFQEEAQKKSIELVLDFQVKKEEIKFYADPDRVQQIISHLMNNAIKFTPEGGRVT : 526
Cy7425_ZP_03139397 : SGEHLLELINDILDLSQVEAGKAVLSLSEFSLNQLCRQALRVFREKAQEAEVELKLESLIPSEFERFCADQRRVRQILFNLLSNAIKFTPAGGRVT : 502
Cy7822_ZP_03155846 : NGKKLMDLINDILDYSSITSGDYQLNIHEFSLYSVVNSVLRDFASEAQKKSIDLILDFQVQKEENDFYADRERIKQILSHLISNAIKFTPDQGKVT : 531
Cy8801_ZP_02940946 : SGKHLLEMINEILEYSKLEAGKYVLSIQQFSLQKAAQNIHKKLRQEAVKRKINLQLELQIEPQENFFFADPERLQQILYHLLNNALKFTPEGGTVI : 531
CyATCC_YP_001806165 : SGKHLLDMINEILEYSNLQSGKYVLAVREFSLTKVAKNVIQRVSDEAEHRRINLELDLQINQQQDSFYADPERVQQILYHLLNNALKFTPENGSVT : 535
Ly_ZP_01622671 : SGEHLLELINDILELSEVEAGKTILQISEFSLTKLAQKTLYSFRERAQEHNVLLISDIQVKPDEDLFIADQRRVEKVLFNLLSNAIKFTPESGQVT : 404
Mc_YP_002622050 : NGEHLLALINDILDFSQLEAGKAVLNINEFSLRYLAQQTLRSLKEKASSGQVELILDWQVPTPKDRFQADPRRLRQILFNLLGNAIKFTPAGGQVT : 565
Mi_YP_001659385 : NGNKLLELINDLLNFSQVSAGKAVLDIKQFSLKYLCNHVLEIIKTEAETKEINLILEMKITEKEHYFYADYDRVQQILLYLLKNALKFTPEQGAVT : 512
No_NP_485728 : SGEHLLDMINDILDLSQIEAGKTVLNIAEFSLVKIAENTIESLLEKALSEQVNLKLDLQIDPRRDRFTADAARIEQILWNLLTNAIKFTPEGGNVT : 588
Np_ZP_00108622 : SGEHLLEMINDILDLSQIEAGKTALNISEFSLVNVAEKALESLRNKATSEQINLNLDLQIDPSRDRFTADAERVAQILSNLLTNAIKFTPESGNVT : 593
Scy_NP_441517 : NGKKLLDQINSIIQLSQLESGQTALNCQSFSLHTLAQTVIHSLLGVAIKQQINLELDYQINVGQDQFCADQERLDQILTQLLNNALKFTPAEGTVI : 524
Syn_YP_001734160 : SGTRLRNLLQDILDFSQTEAARSVLDLQKFSLKQLCYRVLNGFQELGDRQGINLKFINQLEPEQDSFYADPIRLEKILSHLLSNAIKFTSHGGEVT : 530
Th_NP_681689 : SGTHLLELINSILDLSEAELGRSQLHRSAFSIRQLCADCLEVVKPQAHRHQVNLRHQLMIPPTRDRFWGDYRRIQQILINLLSNAIKFTPAMGEVI : 501
Tr_YP_721187 : SGEHLLELINDILDLSQVEAGKTVLKISEFSLSKISYELLQNFREKAEQNEVKLIFEPTVNPQLDLFAADQRRVKQILFNLLSNGIKFTPNGGSVR : 626


780
800
820
840
860
Sel_CikA : LRVWREGDRAIFQVSDTGIGIPESEQAQLFQKFQQLDTSIRRQYGGTGLGLALTKQLVELHGGHIQIESTVGQGSTFTVWIPEQTLIEPVEPRPSI : 622
Ac_YP_001518724 : LRVWVETNTVVFQVEDTGIGISSSQLPHLFEKFQQLDGSYHRTYEGAGLGLALAQQCVTLHQGWIDVSSEEGQGSIFTVQLPNQSALFQQETSSQP : 608
An_YP_321687 : LRVWVEEDTSIFQVEDTGIGIPEEQLPLLFEKFQQLDTPYRRRYEGTGLGLALTKQLVELHRGRIEVESTVGIGSIFTVWIPYQEIRE-------- : 676
Ar_ZP_03273921 : LTVSRKDKTAIFQVKDTGIGISKEQQSLLFEKFQQLDSPYRRQYGGTGLGLALTKQLVELHGGSIQFESEVGLGTKFTVYIPIQPLINKSSSNYKT : 489
Cr_ZP_00513747 : LRVWLEKNQLCLEVEDTGIGIEEEKIPLLFESFQQLENSRRRVYGGTGLGLALTKQLVELHGGTIEVESIINKGSTFTVRIPSQPQNKAKLVSQSE : 627
Cy0110_ZP_01728511 : LRIWRENNQVSLEVEDTGIGIKEEQIPLLFESFQQLENSRRRMYGGTGLGLALTKQLVELHGGTIEVESIINQGSTFIVRLPNQPQSQYKSLNNLD : 630
Cy7424_ZP_02976063 : LRIWREKNQVFFQIEDTGIGITQDQLPLLFEKFQQLESSRQRTHGGTGLGLALTKQLVELHRGTIEVESTPHEGSLFTVRLPNSINLKTKRNFTAD : 622
Cy7425_ZP_03139397 : LRSWVENGKALFQVEDTGIGISLSQQPLLFQKFQQLDTSYTRSYEGVGLGLALTKQLVELHRGRIEVESTEGIGSIFTVELPAQSPTATLSAGLKP : 598
Cy7822_ZP_03155846 : LRIWREKSQVFFQVEDTGIGISQEQLPLLFEKFQQLEKARQRTHGGTGLGLALTKQLVELHRGTIEVESTPEQGSLFTVRIPNNINLKNKKLLTNE : 627
Cy8801_ZP_02940946 : LRIWREGNEAIFQVEDTGIGIAQEQIPLLFESFQQLETSRRRTYGGTGLGLALTKQLVELHGGTIEVESILQEGSIFTVRLPNQPQRQLKSSDNLA : 627
CyATCC_YP_001806165 : LRIWRENNHVSLEVEDTGIGIQEEKIPLLFESFQQLENSRRRMYGGTGLGLALTKQLVELHGGTIEVESIINQGSTFIIRLPNQPQNQYKSLHNLD : 631
Ly_ZP_01622671 : LRVGVNKNTAIFQIQDTGIGISEEQQPLLFEKFQQLDSSYRRQYSGTGLGLALTKQLVELHGGVIKLTSTENVGSTFTVHIPIQPTPSPNSQQQVK : 500
Mc_YP_002622050 : LRGWREQNRAVFQIEDTGIGIPQEQIPLMFQKFQQLEKTYQRTYEGTGLGLALTQQLVELHGGTIDVESTVGKGSIFTVRVPIQSVVPGGTKSPAS : 661
Mi_YP_001659385 : LRLWKEGSQVIFQVEDTGIGIPAQKIPLLFEKFTQLDGSRKRRYGGVGLGLALTKQLVELHRGTIEVESCLDKGSIFTVRLPQQKPPKPQPSDNNP : 608
No_NP_485728 : LRIWVEEDTSIFQVEDTGIGIPEEQLPLLFEKFQQLDTPYRRRYEGTGLGLALTKQLVELHRGRIEVESTVGIGSIFTVWIPYQEIRA-------- : 676
Np_ZP_00108622 : LRLWVEDDTAIFQVEDTGIGIREEQLPLLFEKFQQLDTPYRRRYEGTGLGLALTKQLVELHRGRIEVESTISIGSIFTVWIPTQAMRVLS------ : 679
Scy_NP_441517 : LRIWKESNQAIFQVEDTGIGINEQQLPVLFEAFKVAGDSYTSFYETGGVGLALTKQLVELHGGYIEVESSPGQGTIFTTVIPQQNFPPTTKGQVQD : 620
Syn_YP_001734160 : FSIWREKQKDVFQVKDTGIGIPPQQIPLLFEQFQQLEPSMSRRYDGAGFGLALVKQLVEIHQGQIHVTSTPKKGSIFTVRIPNQVPRNLHPVGKGS : 626
Th_NP_681689 : LRAWWKEDELIFQVQDTGIGIPAHLQSLLFQKFQQLDSSFGRAYTGAGLGLALTKQWVDLHHGWIDVDSTEGKGSTFTVGLPAISDPLPDPPKPKL : 597
Tr_YP_721187 : LKVMRYKNTAIFQVEDTGMGISQEQKPLLFKKFQQLNSPYNRQQGGTGLGLALTKQLVELHNGIIEVESTVNVGSTFTVKLPAQQLKSGNEKRKQK : 722
880
900
920
940
960
Sel_CikA : DN-----------LPAGHILLLEEEDEAATVVCEMLTAAGFKVIWLVDGSTALDQLDLLQPIVILMAWPPPDQSCLLLLQHLREHQADPHPPLVL : 706
Ac_YP_001518724 : LS-------------AGRIVLIEGHEEDATLLCDMLTAANYQVIWMVEASAAIDQIRLLQPIAVIVDAQLPAQGCLGLIRRLRALPGTDKIKIIV : 690
An_YP_321687 : ---------- : -
Ar_ZP_03273921 : SL------PSMYTDTPQGRLVLIEDDEETATLICELLTAAGYQVVWLMDGLTALATIQLLKPDAIFIDLHISGQDGYDIVRHLREDATTQKIKIVA : 579
Cr_ZP_00513747 : KNQA--------LFTKNKTIVLVESNEEIATLIGELLTAANYHFIWLMDGTKVIKKIELLEPSAVILDKELS--KIIEINDSLKQYPETKDTKVLI : 713
Cy0110_ZP_01728511 : KNRA--------LFTKNKTIILVESNEEIATLIGELLTAANYHFIWLMDGSKVFKKLELLEPSAVILDQDLS--EVLKINKSLKQLPETKETKVLV : 716
Cy7424_ZP_02976063 : ENN-------LINTNPNNTIILISKDEETATLICELLTVNNYQVIWLVDSYPGIRQIEILHPLIVILDQENSQSE--EIVKALKQFSKTSFIKVLV : 709
Cy7425_ZP_03139397 : NL-------PPSDFLMTGRIVLLESDEETATLICDILTAAGYQVVWIMEPSTAVEQIQYLQPIAVITAVDLPNMDGEDIIRQLRLYPPFPLLKILA : 687
Cy7822_ZP_03155846 : TNN-------VNKSAQNKTIILISKDEEAATLICELLTVKNYQVIWLLDSYPSIRQIEILQPLIVIIDQEI--IQSQEIGKALKHYPKTSFVKVLV : 714
Cy8801_ZP_02940946 : INQS--------LSMGNRSIVLIESNEELATLIGELLTAANYQFIWLMDSTTAIKKVELFEPTAVILDQDLT--DAYKISEALKASPKTKSIKVLL : 713
CyATCC_YP_001806165 : KNQA--------LFTKNKTIVLVESNEEVATLIGELLTAANYHFIWLMDGTKVIKKLELLEPSAVILDQDLS--EVLEINACLKELPETQGTKVLV : 717
Ly_ZP_01622671 : DSAFKTSLPTTSKNNPLGRFILIENDEEIATLICEILTAVGYQVVWLVEGLTALGQIQLLQPIAVIVDMNLPGQDGYEIIHHLKNTKATQQIKILA : 596
Mc_YP_002622050 : SEK---------GSSIPGSIVLIENNEASATAICEILTAAGYHLVWLIESSTAVRQIELLQPHAVILDWQLSAMDGYEISYYLHHKTTTAHIKVLA : 748
Mi_YP_001659385 : HF-----------NHHHPTIVLIESDEEIANLICELLMVAHYQVIWLIDSVSAIKKIEIVQPGIIVDRKMP--DIYHLCHLLKKSRQTQASKVLI : 691
No_NP_485728 : ---------- : -
Np_ZP_00108622 : ---------- : 683
Scy_NP_441517 : KLDA--------AMPFNSSVIVIEQDEEIATLICELLTVANYQVIWLIDTTNALQQVELLQPGLIIVDGDF--VDVTEVTRGIKKSRRISKVTVFL : 706
Syn_YP_001734160 : V----------GGNQGGTIVLVSQDEEMATLICELLTATNYQVIWLIDSEIASRQISALQPILVILDAHQHKIQIEDIIDSLKMAPQTQQIPTLL : 711
Th_NP_681689 : DV--PPLATTEVLVEPEGRIVLVSEDEATSTLICSILTTAGYQVIWLVDGE--VERLLALTPIAVLLAEPFSYGDVQELVDQLRQRCTPEQLKIFI : 689
Tr_YP_721187 : GE------SLSNVLSLQGSVVLIEQEEDIATVICDILTTAGLKVVWIIEGSTAVEQVILLQPVVIIIDMQLPGINGIEIIDQLRTTSSSKNIKFLA : 812

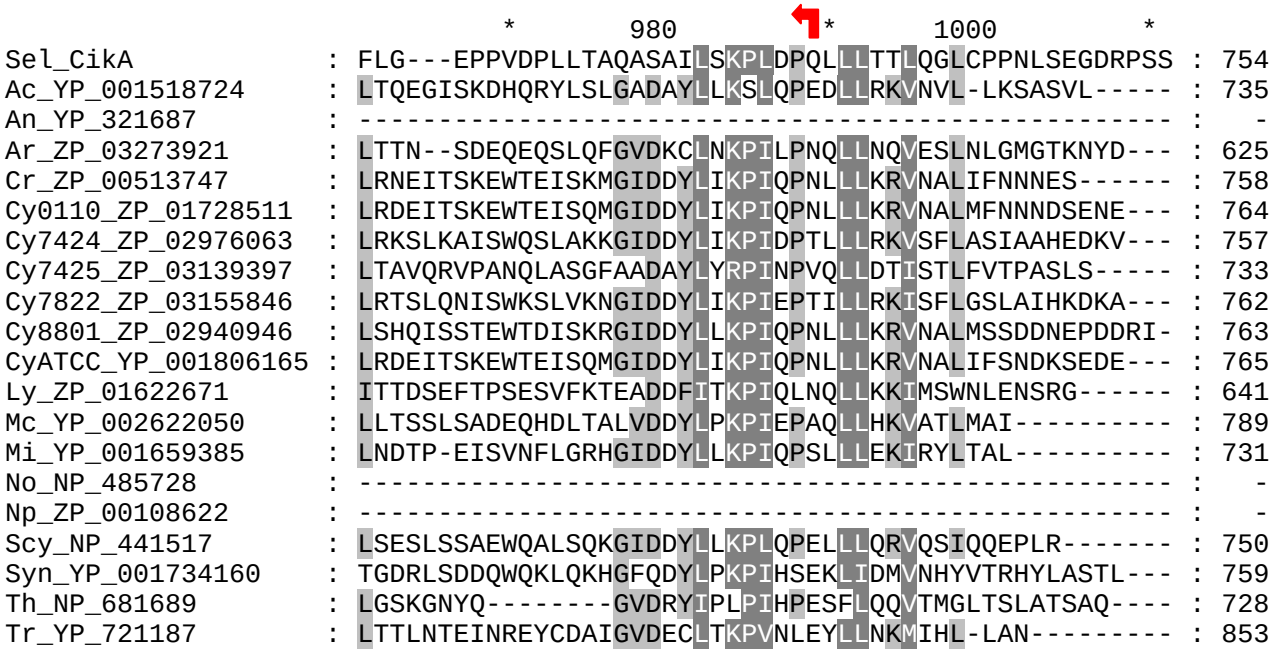
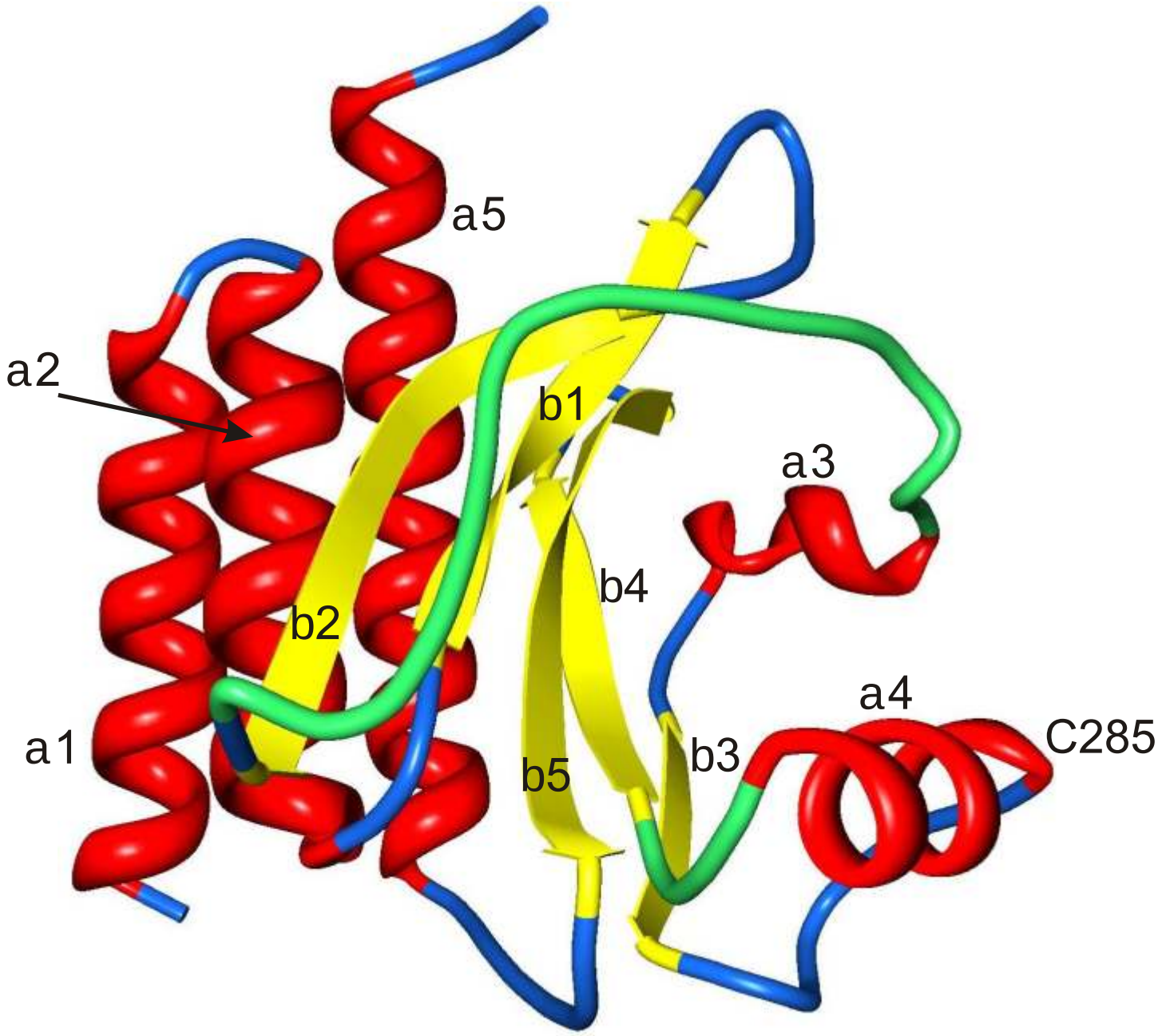

980
1000
Sel_CikA : FLG---EPPVDPLLTAQASAILSKPLDPQLLLTTLQGLCPPNLSEGDRPSS : 754
Ac_YP_001518724 : LTQEGISKDHQRYLSLGADAYLLKSLQPEDLLRKVNVL-LKSASVL----- : 735
An_YP_321687 : ---------- : -
Ar_ZP_03273921 : LTTN--SDEQEQSLQFGVDKCLNKPILPNQLLNQVESLNLGMGTKNYD--- : 625
Cr_ZP_00513747 : LRNEITSKEWTEISKMGIDDYLIKPIQPNLLLKRVNALIFNNNES------ : 758
Cy0110_ZP_01728511 : LRDEITSKEWTEISQMGIDDYLIKPIQPNLLLKRVNALMFNNNDSENE--- : 764
Cy7424_ZP_02976063 : LRKSLKAISWQSLAKKGIDDYLIKPIDPTLLLRKVSFLASIAAHEDKV--- : 757
Cy7425_ZP_03139397 : LTAVQRVPANQLASGFAADAYLYRPINPVQLLDTISTLFVTPASLS----- : 733
Cy7822_ZP_03155846 : LRTSLQNISWKSLVKNGIDDYLIKPIEPTILLRKISFLGSLAIHKDKA--- : 762
Cy8801_ZP_02940946 : LSHQISSTEWTDISKRGIDDYLLKPIQPNLLLKRVNALMSSDDNEPDDRI- : 763
CyATCC_YP_001806165 : LRDEITSKEWTEISQMGIDDYLIKPIQPNLLLKRVNALIFSNDKSEDE--- : 765
Ly_ZP_01622671 : ITTDSEFTPSESVFKTEADDFITKPIQLNQLLKKIMSWNLENSRG------ : 641
Mc_YP_002622050 : LLTSSLSADEQHDLTALVDDYLPKPIEPAQLLHKVATLMAI---------- : 789
Mi_YP_001659385 : LNDTP-EISVNFLGRHGIDDYLLKPIQPSLLLEKIRYLTAL---------- : 731
No_NP_485728 : ---------- : -
Np_ZP_00108622 : ---------- : -
Scy_NP_441517 : LSESLSSAEWQALSQKGIDDYLLKPLQPELLLQRVQSIQQEPLR------- : 750
Syn_YP_001734160 : TGDRLSDDQWQKLQKHGFQDYLPKPIHSEKLIDMVNHYVTRHYLASTL--- : 759
Th_NP_681689 : LGSKGNYQ--------GVDRYIPLPIHPESFLQQVTMGLTSLATSAQ---- : 728
Tr_YP_721187 : LTTLNTEINREYCDAIGVDECLTKPVNLEYLLNKMIHL-LAN--------- : 853

Figure S3



Legend:

The conservation scale:

1	2	3	4	5	6	7	8	9

Variable **Average** **Conserved**

e - An exposed residue according to the neural-network algorithm.

b - A buried residue according to the neural-network algorithm.

f - A predicted functional residue (highly conserved and exposed).

s - A predicted structural residue (highly conserved and buried).

X - Insufficient data - the calculation for this site was performed on less than 10% of the sequences.

Figure S4