BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(6)

Comparative analysis to identify determinants of changing life style in *Thermosynechococcus elongatus* BP-1, a thermophilic cyanobacterium

Ratna Prabha[1, 5], Dhananjaya P Singh[1]*, Shailendra K Gupta[2], Sávio Torres de Farias[3] & Anil Rai[4]

[1]National Bureau of Agriculturally Important Microorganisms, Indian Council of Agricultural Research, Kushmaur, Maunath Bhanjan 275101, India; [2]CSIR-Indian Institute of Toxicology Research, Mahatma Gandhi Marg, Kaisarbagh, Lucknow 226001, India; [3]Departamento de Biologia Molecular, Universidade Federal da Paraiba, Brazil; [4]Indian Agricultural Statistical Research Institute, Indian Council of Agricultural Research, Pusa, New Delhi 110 012, India; [5]Department of Biotechnology, Mewar University, Gangrar, Chittorgarh, Rajasthan, India; Dhananjaya P Singh - Email: dpsfarm@rediffmail.com; Phone: +91-547-2530080, Fax: +91-547-2530358; *Corresponding author

Received December 04, 2012; Accepted December 22, 2012; Published March 19, 2013

Abstract:
A comparative genomics analysis among all forty whole genome sequences available for cyanobacteria (3 thermophiles–*Thermosynechococcus elongatus* BP-1, *Synechococcus* sp. JA-2-3B'a (2-13), *Synechococcus* sp. JA-3-3Ab and 37 mesophiles) was performed to identify genomic and proteomic factors responsible for the behaviour of *T. elongatus* BP-1, a thermophilic unicellular cyanobacterium with optimum growth temperature [OGT] of 55°C. Majority of genomic and proteomic characteristics for this cyanobacterium indicated contrasting features indicating its mesophilic behaviour while the role of mutational biasness and selection pressure is thought to be responsible for high OGT. Contradictory results were obtained for *T. elongatus* for synonymous codon usage, CvP-bias and amino acid composition with respect to thermophilic behaviour. Calculated J_2 index is lowest among all cyanobacterial genomes. Except for proline and termination codons, *T. elongatus* showed synonymous codon usage pattern which is expected for mesophiles. Results indicated that among cyanobacterial genomes, majority of genomic and proteomic determinants put *T. elongatus* very close to mesophiles and the whole genome of this organism represents continuous gain of mesophilic rather than thermophilic behavior.

Keywords: *Thermosynechococcus elongatus*, Thermophily, Genomics, Codon usage, CvP bias, J_2 index.

Background:
Thermophiles provide comprehensive physiological, biochemical and molecular insights into the biology of microbial life at high temperature **[1]**. Hyperthermophiles grow optimally above 65°C, thermophiles have optimal growth temperature (OGT) between 45 to 65°C while mesophliles grow well below 55°C (OGT 37°C) **[2]**. Possible relation between OGT of the organisms and nucleotide content of their genomes **[3, 4, 5, 6]** and specific trends in amino acid composition **[7, 8, 9]** has widely been worked out as signatures for thermophilicity. Genomic evidences for thermal adaptations suggest a positive correlation between G+C content and OGT because G:C pair is more stable than A:T but, there are contradictions too **[10]**. The composition of purine/pyrimidine dinucleotides is shown to correlate linearly with the OGT in Archaea **[11]**. Increase in the purine (A+G) load index in the genomes of thermophilic bacteria **[12]** represents possible primary factor of adaptation mechanism. Several other identified factors for the thermal

adaptations in the organisms include high core hydrophobicity **[13]**, high secondary structure propensity and packing density **[14, 15]**, increase in van der Waals interactions in thermophilic proteins **[16]**, ionic interactions **[17]**, high content of disulphide bonding **[18, 19]**, hydrogen bonding **[20]**, increase in Glu (*E*) and Lys (*K*) and decrease in Gln (*Q*) and His (*H*) residues **[21]**. Although genomic information facilitates the study of thermophily in the prokaryotes and determines thermostability at proteome level **[22]**, the central issue lies in finding out the adaptive strategy of nucleic acid molecules towards different OGTs **[10]**.

Despite significant efforts, compositional biases representing most definitive signatures of thermophilic adaptations in genomes and proteomes still remain elusive **[9]**. Thermophilic adaptations based on nucleotide biases are largely governed at the level of amino acid composition but additional adaptation in DNA sequence at the level of high-order correlation in nucleotides is inferred in prokaryotes using codon usage analysis **[3]** has also been suggested **[9]**. While thoroughly analysing *Thermosynechococcus elongatus* BP-1, a thermophilic unicellular rod-shaped cyanobacterium (OGT 55°C) inhabiting hot springs and comparing the same with 39 other cyanobacteria for whom whole genome sequences are available (NCBI, Genome database, May 2012), we found different levels of correlations among physical OGT, nucleotide and amino acid composition and codon biases. We discussed the relation of the observed pattern at nucleotide, proteome and amino acid level with the physical growth of *T. elongatus* and trends that appeared after comparative genomic analysis with other cyanobacterial organisms to find out determinants for the thermophilic or mesophilic behavior.

Methodology:

Sequences

Out of 42 complete genome sequences available for cyanobacteria (NCBI Genome, Apr. 2012), 40 sequences were taken for the study. For *Arthospira platensis* NIES-39 sequencing is incomplete and only one out of two sequences available for *Synechocystis* sp. PCC6803 was taken. Complete genome, protein and rRNA sequences were downloaded from NCBI for each of the 40 genomes **Table 1 (see supplementary material).** OGT of the organisms was obtained from the available literature.

DAMBE version 5.2.73 **[23]** was used for counting number of individual nucleotides (A, T, G, C) in each of the 40 genomes. PERL scripts were used to calculate the composition of different combination of dinucleotides [YR (TA, TG, CA, CG), RY (AT, AC, GT, GC), YY (TT, CC, TC, CT) and RR (AA, GG, AG, GA)]. J_2 Index is the subtraction of the frequency of all combinations of YR (TA,TG, CA,CG) and RY (AT, AC, GT, GC) from that of all YY (TT, CC, TC, CT) and RR (AA, GG, AG, GA) combinations **[11]**.

The index was calculated by the following formula: J_2 index = $\sum (F_{YY} + F_{RR} - F_{YR} - F_{RY})$. Total number of particular codons in the genome and relative synonymous codon usage (RSCU) was calculated through CUSP from EMBOSS package (http://www.ebi.ac.uk/Tools/emboss/).

The relationship between G+C content of RNA and OGT (OGT-RNA) was expressed by the equation - OGT-RNA= 2.91× (G+C) -103; Where, OGT-RNA is the OGT estimated in degree Celsius (°C) and G+C is the percentage of guanine and cytosine in 16S rRNA **[2]**.

Percent composition and total number of each amino acid in the proteome was calculated with the help of PERL script. The difference between charged (Lys, Arg, Asp, Glu) and polar-non-charged amino acid (Asn, Gln, Ser, Thr) i.e. CvP-bias and E+K / Q+H ratio in the proteome was calculated.

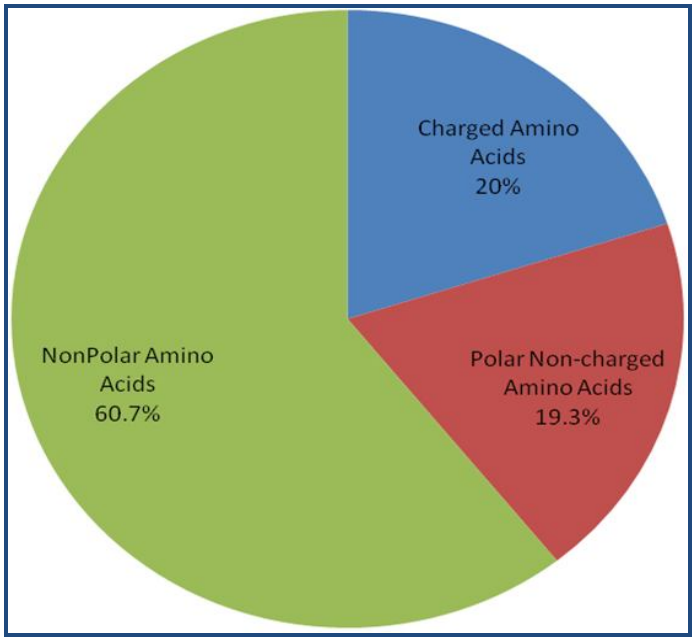

Figure 1: Distribution of charged amino acids (Asp, Glu, Arg, Lys), polar non-charged amino acids (Asn, Gln, Ser, Thr) and non-polar amino acids (Gly, Ala, Val. Leu, Ile, Phe, Trp, Tyr, Pro, Met, Cys, His) in *T. elongatus* proteome

Results & Discussion:

Analysis of genomes

All 40 genomes varied in size from 1.44 Mb (*Cyanobacterium* UCYN-A) to 9.04 Mb (*Nostoc punctiforme* PCC 73102). Genomic GC content varied from 30.8% (*Prochlorococcus marinus* subsp. pastoris CCMP1986 and *P. marinus* MIT 9515) to 62 % (*Gloeobacter violaceous* PCC 7421). In cyanobacterial genomes, number of CDS (coding sequences) vary from 1199 (Cyanobacterium UCYN-A) to 6312 (*Microcystis aeruginosa* NIES-843) **(Table 1).**

Out of these 40 cyanobacteria, except *T. elongatus,* two other cyanobacteria *Synechococcus* sp. JA-2-3B'a(2-13) (OGT 50 to 55°C) [24] and *Synechococcus* sp. JA-3-3Ab (OGT 50 to 60°C) **[24]** are also reported as thermophiles while rest of 37 are mesophiles. Thermophiles have shown higher proportion of G+C that varied from 30 to 60% or more, irrespective of their behaviour **[2]**. The genomic GC content in *T. elongatus* is 53.9% while over-all GC content of rRNA operon is 55.15% **Table 2 (see supplementary material).** In comparison to other two thermophilic cyanobacteria that showed higher GC content i.e. 58.5% (*Synechococcus* sp. JA-2-3B'a(2-13) and 60.2%

(*Synechococcus* sp. JA-3-3Ab) GC content of *T. elongatus* seems to favour mesophilic behaviour.

Purine load is a preferred index for thermophiles with low GC content like *T. elongatus* but this is uncommon in non-thermophilic organisms **[25]**. Purine load index i.e. the concentration of A+G is known to exhibit highest correlation with OGT and represents a primary adaptation mechanism to thermophily **[9]**. In contrast to other thermophilic cyanobacteria that showed nearly similar purine and pyrimidine nucleotides, *T. elongatus* showed no biasness towards purine but had higher pyrimidine (51%) than purine (49%) content **Table 3 (see supplementary material).** This organism, therefore, neither strongly favours GC bias nor nucleotide bias for its thermophilic character.

Combination of purine (R)/pyrimidine (Y) dinucleotide composition is shown to correlate linearly with the OGT among thermophilic Archaea **[11]**. A higher J_2 index is considered as important criteria for hyperthermophiles **[26]** and a positive J_2 value is reported for the sequences of all the thermophiles, while negative value represented mesophiles **[11]**. A positive J_2 index ranging from 0.003599 (*Gloeobacter violaceus* PCC 7421) to 0.148216 (*P. marinus* MIT 9301) was calculated for all the cyanobacteria studied except *Synechococcus* sp. RCC307 (J_2 index -0.00142) **Table 4 (see supplementary material).** Among thermophilic cyanobacteria under study, J_2 index calculated to be the least for *T. elongatus* (0.046) while *Synechococcus* sp. JA-3-3Ab and *Synechococcus* sp. JA-2-3B'a(2-13) showed J_2 index value as 0.079 and 0.083, respectively. It is suggested that for the thermophiles, J_2 value should be positive whereas for mesophiles, a negative value is recommended **[11]**. Although, *T. elongatus* showed a positive value (0.046) for J_2 index, this criterion is insufficient to establish this organism as mesophile or thermophile because, a high J_2 index value is important but not a sufficient criterion for thermophilc behaviour **[26]**.

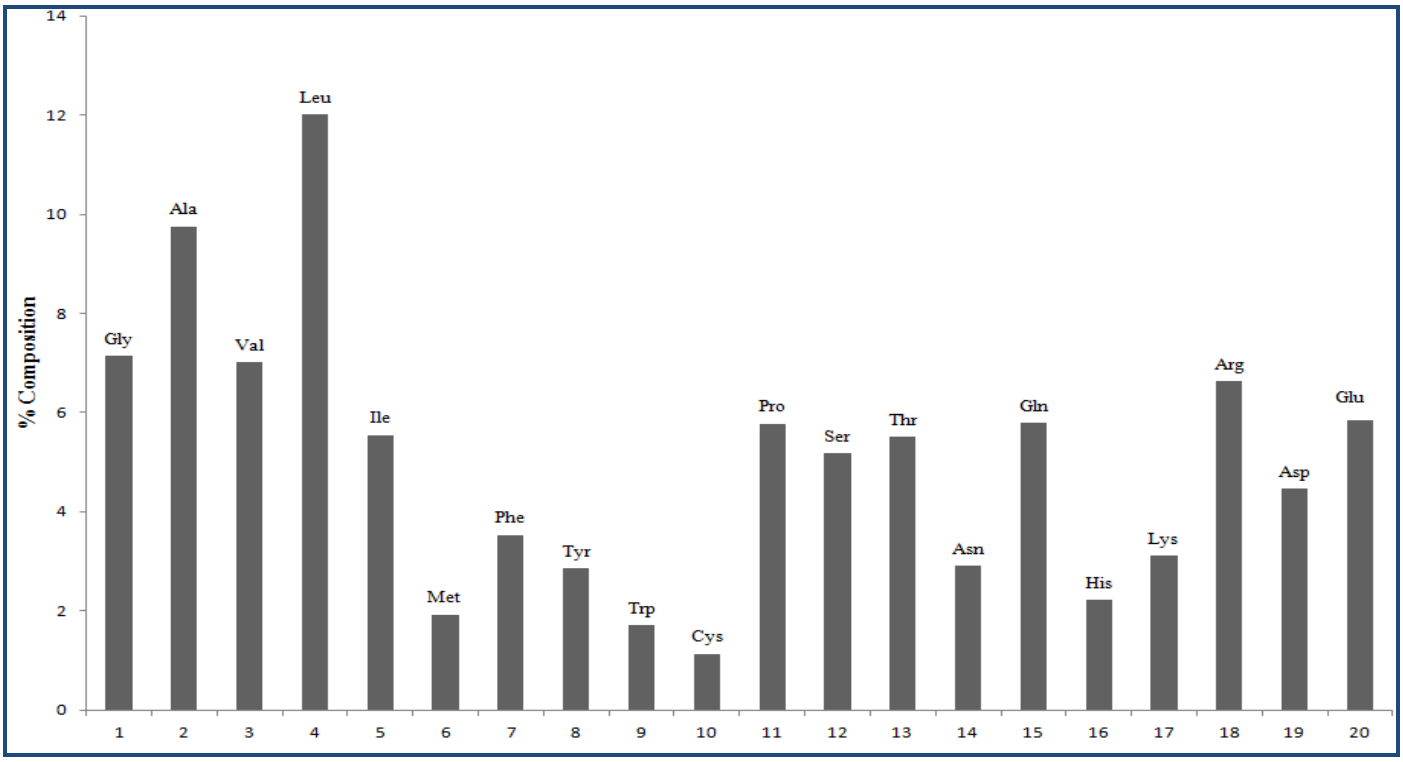

Figure 2: Distribution of different amino acids in *T. elongatus*

rRNA analysis

Significant correlation between G+C content of structural RNAs (rather than entire genome) and OGT **[27]** along with an additional preference for purine rich RNA (particularly adenine as compared to guanine in t-RNA and r-RNA but not in m-RNA) is reported in thermophiles **[28]**. Showing thermophilic tendency, *T. elongatus* uses purines more frequently than pyrimidine for both tRNA and rRNA but prefer guanine over adenine that further reflects mesophilic nature. While considering distribution of individual nucleotides i.e. A, T, G, C in 16S rRNA across all cyanobacteria, we identified that purine content is higher than pyrimidine in all of them where G occupied the maximum percentage and T the minimum. However, for A and C, there is different trend for thermophiles and mesophiles. Among thermophilic cyanobacteria (including *T. elongatus*) maximum percent content in rRNA is of C after G, whereas for mesophilic cyanobacteria it is for A. Comparative analysis of *T. elongatus* rRNA GC content (55.15%) with *Bacillus*-related mesophilic species (in which it varied from 52.7 to 54.4% in rRNA GC content and 42 to 58°C in growth temperatures) **[22]** showed behavioural similarity. But, significant difference is observed when we compare the genomic G+C content of *T. elongatus* (53%) and *Bacillus*-related mesophilic species (35.3-43.7%) in relation to the OGT. On the basis of 16S rRNA GC content,calculated OGT for *T. elongatus* was 57.25 °C and this correlated with the physical growth temperature of this organism. More or less, the same is reflected by the GC content of rRNA (52.7-54.4%) in *Bacillus*-related mesophilic species, OGT of which lies in the range of 42-58°C and the organism is a mesophile.

Analysis of protein content

In organisms, enhanced thermostability reflected specific trend of their amino acid composition **[7, 8, 29]** and that too, especially in the increased fraction of charged residues **[6, 30]** and/or enhanced content of hydrophobic residues **[9]**. We examined difference between charged (Lys, Arg, Asp, Glu) and polar non-charged amino acids (Asn, Gln, Ser, Thr) which were considered as a characteristic signature of thermophilic microorganisms **[26]**. *T. elongatus,* along-with other 39 cyanobacteria under study, did not show any preference for charged amino acids and there is no significant difference between charged and polar amino acids across all the cyanobacteria. Charged amino acids occupy 20%, polar-non-charged 19.3% and others 60.7% of entire proteome **(Figure 1)** and therefore, significant difference between charged and non-charged amino acids does not exist in *T. elongatus* **(Figure 2).**

CvP bias, the ratio of charged (Lys, Arg, Asp, Glu) and polar non-charged amino acids (Asn, Gln, Ser, Thr) is an important signature and a global characteristic of all thermophiles (OGT>55°C) in which it remains markedly higher than the mesophilic organisms **[26]**. Among all cyanobacteria, *Gloeobacter violaceus* PCC 7421 showed highest value for CvP bias i.e. 3.83. *Synechococcus* sp. JA-3-3Ab (2.713) and *Synechococcus* sp. JA-2-3B'a(2-13) (2.06) also have a higher value for CvP along with some other strains of *P. marinus* but *T. elongatus* showed a small CvP bias value of 0.68 that establishes a mesophilic life style for this cyanobacteria. The ratio of Glu (E) +Lys (K)/Gln (Q) +His (H) for *T. elongatus* proteome is calculated to be 1.12 that again indicates mesophilic behaviour. Other two thermophiles also show quite similar but comparatively higher value. Some strains of *P. marinus* showed higher value. In *T. elongatus* genome, 30 genes are predicted to be responsible for thermophily. On considering each gene (and their protein products) individually, similar results for CvP bias as of entire proteome except for three proteins namely chaperonin2, chaperonin GroEL and chaperonin GroES (4.2, 5.27 and 9.5 respectively) were found. Chaperonins are potentially thermostable proteins that are expected to favour thermophilic behaviour in organisms **[21]**.

Total fraction of the universal set of amino acids (Ile, Val. Tyr, Trp, Arg, Glu, Leu [IVYWREL]) in the proteome is considered to correlate with OGT **[9]**. In *T. elongatus* proteome, the fractional composition of the universal set of amino acids (IVYWREL) is 41.56%. Similar distribution of IVYWREL is observed across all the cyanobacteria. In comparison to mesophiles, thermophiles have significantly higher content of Val and Glu than Gln and Thr **[29]** along with the reduced frequency of His and Gln. No significant change is observed in the distribution of Val and Glu across meso- and thermophilic cyanobacteria. Leu is in high proportion (11.97%) in the proteome in terms of amino acid composition but the composition of His is only 2.2%. Across all the cyanobacteria under study, Leu occupies the largest proportion whereas Cys occupies the least and there remains no exception. These results indicated a thermophilic pattern for His only but a good proportion of Gln (5.76%) and Thr (5.48%) along with the low fraction of IVYWREL indicated mesophilic behaviour for *T. elongatus* **Table 5 (see supplementary material).** Similarly, among the amino acids Glu, Arg, Tyr, Asp and Lys that are reported to be abundant in thermophiles **[8, 31]**, *T. elongatus* proteome contains high proportion of Glu (5.84%) and Arg (6.61%) but other amino acids are poorly contained **(Table 5).**

Codon usage pattern

Synonymous codon usage pattern in thermophilic prokaryotes is different from mesophiles and the difference is a result of natural selection linked to thermophily. Moreover, this phenomenon is not restricted to specific set of genes and affects all of the genes within the genome **[32]**. Synonymous codon usage pattern in *T. elongatus* genome does not show any distinguishable thermophilic pattern. Except for proline and termination codons, *T. elongatus* shows synonymous codon usage pattern which is expected for mesophiles and not for thermophiles.

Thermophilic prokaryotes characteristically include increased usage of AGR codons for arginine, ATA for isoleucine and decreased usage of CGN for arginine **[6, 33, 34]**. For Arg, *T. elongatus* preferentially uses CGC out of the 6 codons available along with other two thermophilic cyanobacteria. For isoleucine, ATC is used by both *Synechococcus* sp. JA-3-3Ab and *Synechococcus sp.* JA-2-3B'a(2-13) whereas *T. elongatus* uses ATT which resembles mesophilic cyanobacteria. Biasness for arginine codon family containing the largest number of codons **[6]** in organisms is a signature of thermophiles and hyperthermophiles that represent important implications to thermostability **[21]**. *T. elongatus* preferentially uses CGC not only in arginine codon family but also in the entire genome followed by CGG. This is supported by the relative synonymous codon usage (RSCU) value for CGC which is highest followed by CGG and CGU in the entire genome. These results contradict with the fact that thermophiles and hyperthermophiles tend to employ AGR to encode arginine. Preferential usage of AGR for arginine implies positive error minimization and contributes to avoid mutations that harm protein thermostability and represent alternative mechanism of adaptation to proteins **[1]**. In *T. elongatus,* AGA and AGG are least used for Arg in the entire genome reflecting a mesophilic character of the genome because encoding for arginine by CGT and CGC are positive indicators for mesophiles **[35]**. Similar codon usage pattern was observed when all the 30 genes predicted for thermophily were considered individually (data not shown).

Entire proteome composition of *T. elongatus* showed that out of the total number of Leu, maximum is encoded by CUG codons. Thermophiles prefer GGR over GGY and *vice-versa* and this pattern is characteristically established for glycine (Gly) and arginine (Arg) **[24]**. It is observed that maximum cyanobacteria, including thermophiles preferentially use GGY over GGR. Also among GGY, GGC is more commonly used than GGT by thermophilic cyanobacteria. Like other mesophilic cyanobacteria and previous studies, preferences of *T. elongatus* genome for GGY and CGY codons for Gly and Arg again indicated mesophilic behaviour of the organism (data not shown).

A preference for specific nucleotides at 3rd position of codons of some amino acids has been shown by some thermophilic organisms **[36]**. Given a synonymous choice between T and C

(Asn, Asp, Cys, His, Phe, Tyr) at 3rd position, thermophiles systematically favour C-ending codons. Similar trend was also followed by the two thermophilic cyanobacteria (with an exception for phenylalanine). However, *T. elongatus* prefers T-ending codons instead of C-ending codons. When the choice at 3rd position is between A and G (Gln, Glu, Lys), G-ending codons are preferentially favoured **[36]**. *Synechococcus* sp. JA-3-3Ab and *Synechococcus* sp. JA-2-3B'a(2-13), like other thermophiles, show expected pattern but *T. elongatus* preferentially uses A-ending codons for these amino acids as is evident for mesophilic cyanobacteria. Among codons ending with pyrimidine, T-ending codons are most widely used across all mesophilic cyanobacteria including *T. elongatus* whereas *Synechococcus* sp. JA-3-3Ab and *Synechococcus* sp. JA-2-3B'a prefer C-ending codons as expected for thermophiles. G-ending codons also share a significant proportion among two thermophilic cyanobacteria whereas for *T. elongatus* and other mesophiles studied, they are the least used codons (data not shown). For termination codons, *T. elongatus* favours G-ending codons like thermophiles but mesophiles prefer A-ending termination codons. This characteristic reflects thermophilic behaviour of *T. elongatus*.

Thermal adaptations in T. elongatus

Potential adaptations to high temperatures in thermophilic organisms lie in their highly efficient and specialized DNA repair systems **[37]**. A specific DNA repair system largely confined to thermophiles is known to be regulated by a pathway consisting of more than 20 genes **[38]**. This pathway is missing in *T. elongatus* genome. Even no gene of *T. elongatus* showed significant similarity to the genes conferring the DNA repair system. The genes encoding fatty acid desaturases (*desA, desB, desC*) are missing reflecting the absence of highly unsaturated fatty acids in lipids composition which is characteristic of thermophiles **[39]**. Genes for replication, recombination and repair (as per COG classification) occupy a major composition (6.7049%) in *T. elongatus* genome along with a significant composition of genes involved in posttranslational modification, protein turnover, chaperones (4.0158%) showing thermal adaptation mechanisms.

Conclusion:

T. elongatus requires high temperature (55°C) for physical growth **[39]**. To a certain extent, genomic determinants but majority of proteomic determinants indicated mesophilic behavior of this organism with an OGT of (38.41°C) and that too, was supported by various reports **[22, 40, 41]**. In correspondence analysis performed for the proteomes of 279 prokaryotes, *T. elongatus* occupies the position under a large group of mesophiles **[34]**. Phylogenetically, *T. elongatus* is again found to be closely related to a large group of mesophiles suggesting a recent gain of thermophily by this organism **[34]**. Thermophilic adaptation mechanisms such as the presence of additional genes for heat-shock proteins **[39]**, 28 copies of group II introns comprising of almost 1.3% of the whole genome **[39, 42]**, and heat- induced groEL2 gene **[43]** gained by *T. elongatus* over the time has made it comfortable in mesothermophilic conditions. Additional features like presence of widely temperature compensated endogenous circadian rhythm is known in mesophilic organisms **[44]** and presumptive counterparts of all the genes involved in circadian clock system have also been identified in *T. elongatus* **[39]**. Adaptation to a new environment often necessitates a coordinated change in the genomic organization, rather than independent modifications of individual components **[11]**. Our results indicated that among the cyanobacteria, majority of genomic and proteomic determinants put *T. elongatus* very close to the mesophiles and the whole genome of this organism represents continuous gain of mesophilic rather than thermophilic behavior.

Acknowledgement:

Financial support from Indian Council of Agricultural Research, India in the form of "National Agricultural Innovation Project" entitled "Establishment of National Agricultural Bioinformatics Grid" (NABG) is gratefully acknowledged.

Reference:

[1] Farias ST & Linden MG, *Extremophiles.* 2006 **10**: 479 [PMID: 16830074]
[2] Nakashima H *et al. J Biochem.* 2003 **133**: 507 [PMID: 12761299]
[3] Hickey A & Singer GAC, *Genome Biol.* 2004 **5**: 117 [PMID: 15461805]
[4] Lambros RJ *et al. Extremophiles.* 2003 **7**: 443 [PMID: 14666404]
[5] Musto H *et al. FEBS Lett.* 2004 **573**: 73 [PMID: 15327978]
[6] Singer GA & Hickey DA, *Gene.* 2003 **317**: 39 [PMID: 14604790]
[7] Pe'er I *et al. Proteins.* 2004 **54**: 20 [PMID: 14705021]
[8] Szilagyi A & Zavodszky P, *Structure*. 2000 **8**: 493 [PMID: 10801491]
[9] Zeldovich KB *et al. PLoS Comput Biol.* 2007 **3**: e5 [PMID: 17222055]
[10] Basak S *et al. Bioinformation.* 2010 **4**: 352 [PMID: 20975899]
[11] Kawashima T *et al*. *Proc Natl Acad Sci USA.* 2000 **97**: 14257 [PMID: 11121031]
[12] Lin FH & Forsdyke DR, *Extremophiles.* 2007 **11**: 9 [PMID: 16957882]
[13] Schumann J *et al. Protein Sci.* 1993 **2**: 1612 [PMID: 8251936]
[14] Hurley JH *et al. J Mol Biol.* 1992 **224**: 1143 [PMID: 1569571]
[15] Querol E *et al. Protein Eng.* 1996 **9**: 265 [PMID: 8736493]
[16] Berezovsky IN *et al. FEBS Lett.* 1997 **418**: 43 [PMID: 9414092]
[17] Vetriani C *et al*. *Proc Natl Acad Sci.* 1998 **95**: 12300 [PMID: 9770481]
[18] Beeby M *et al. PLOS Biol.* 2005 **3**: 1549 [PMID: 16111437]
[19] Mallick P *et al. Proc Natl Acad Sci.* 2002 **99**: 9679 [PMID: 12107280]
[20] Jaenicke R, *Proc Natl Acad Sci.* 2000 **97**: 2962 [PMID: 10737776]
[21] Farias ST & Bonato MC, *Genet Mol Res.* 2003 **2**: 383 [PMID: 15011142]
[22] Takami H *et al*. *Nucleic Acids Res.* 2004 **32**: 6292 [PMID: 15576355]
[23] Xia X & Xie Z, *J Hered.* 2001 **92**: 371 [PMID: 11535656]
[24] Allewalt JA *et al. Appl Environ Microbiol.* 2006 **72**: 544 [PMID: 16391090]
[25] Lao PJ & Forsdyke DR, *Genome Res.* 2000 **10**: 228 [PMID: 10673280]
[26] Suhre K & Claverie JM, *J Biol Chem.* 2003 **278**: 17198 [PMID: 12600994]

[27] Galtier N & Lobry JR, *J Mol Evol.* 1997 **44**: 632 [PMID: 9169555]
[28] Trivedi S *et al. J Cell Mol Biol.* 2005 **4**: 61
[29] Kreil DP & Ouzounis CA, *Nucleic Acids Res.* 2001 **29**: 1608 [PMID: 11266564]
[30] Cambillau C & Claverie JM, *J Biol Chem.* 2000 **275**: 32383 [PMID: 10940293].
[31] Paz A *et al. Proc Natl Acad Sci.* 2004 **101**: 2951 [PMID: 14973185]
[32] Lynn DJ *et al. Nucleic Acids Res.* 2002 **30**: 4272 [PMID: 12364606]
[33] Farias ST & Bonato MC, *Genome Biol.* 2002 **3:** PREPRINT0006 [PMID: 12186639]
[34] Puigbo P *et al. Trends Genet.* 2007 **24**: 10 [PMID: 18054113]
[35] Carbone A *et al. Mol Biol Evol.* 2005 **22**: 547 [PMID: 15537809]
[36] Lobry JR & Chessel D, *J Appl Genet*. 2003 **44**: 235 [PMID: 12817570]
[37] Grogan DW, *Trends Microbiol.* 2000 **8**: 180 [PMID: 10754577]
[38] Makarova KS *et al. Nucleic Acids Res.* 2002 **30**: 482 [PMID: 11788711]
[39] Nakamura Y *et al*. *DNA Res.* 2002 **9**: 123 [PMID: 12240834]
[40] Dyer BD *et al. Archaea.* 2008 **2**: 159 [PMID: 19054742]
[41] Fujita M & Kanehisa M, *Genome Inform.* 2005 **16**: 174 [PMID: 16362920]
[42] Mohr G *et al*. *PLoS Biol*. 2010 **8**: e1000391 [PMID: 20543989]
[43] Sato S *et al. FEBS Lett.* 2008 **582**: 3389 [PMID: 18786533]
[44] Onai K *et al. J Bacteriol.* 2004 **186**: 4972 [PMID: 15262934]

Edited by P Kangueane
Citation: Prabha *et al.* Bioinformation 9(6): 299-308 (2013)

Supplementary material:

Table 1: General features of cyanobacterial whole genomes

S. No.	Organisms	Accession Number	Size (Mb)	GC%	CDS
1.	*Acaryochloris marina* MBIC11017	NC_009925.1	8.36	46.99	8383
2.	*Anabaena variabilis* ATCC 29413	NC_007413.1	7.11	41.39	5710
3.	*cyanobacterium* UCYN-A	NC_013771.1	1.44	31.1	1199
4.	*Cyanothece* sp. ATCC 51142	NC_010546.1	5.46	37.97	5304
5.	*Cyanothece* sp. PCC 7424	NC_011729.1	6.55	38.5	5710
6.	*Cyanothece* sp. PCC 7425	NC_011884.1	5.79	50.66	5327
7.	*Cyanothece* sp. PCC 7822	NC_014501.1	7.84	39.87	6642
8.	*Cyanothece* sp. PCC 8801	NC_011726.1	4.79	39.8	4367
9.	*Cyanothece* sp. PCC 8802	NC_013161.1	4.8	39.8	4444
10.	*Gloeobacter violaceus* PCC 7421	NC_005125.1	4.66	62	4430
11.	*Microcystis aeruginosa* NIES-843	NC_010296.1	5.84	42.3	6312
12.	*Nostoc azollae* 0708	NC_014248.1	5.49	38.33	3651
13.	*Nostoc punctiforme* PCC 73102	NC_010628.1	9.06	41.34	6689
14.	*Nostoc* sp. PCC 7120	NC_003272.1	7.21	41.22	6129
15.	*Prochlorococcus marinus* str. AS9601	NC_008816.1	1.67	31.3	1920
16.	*Prochlorococcus marinus* str. MIT 9211	NC_009976.1	1.69	38	1854
17.	*Prochlorococcus marinus* str. MIT 9215	NC_009840.1	1.74	31.1	1982
18.	*Prochlorococcus marinus* str. MIT 9301	NC_009091.1	1.64	31.3	1906
19.	*Prochlorococcus marinus* str. MIT 9303	NC_008820.1	2.68	50	2997
20.	*Prochlorococcus marinus* str. MIT 9312	NC_007577.1	1.71	31.2	1810
21.	*Prochlorococcus marinus* str. MIT 9313	NC_005071.1	2.41	50.7	2269
22.	*Prochlorococcus marinus* str. MIT 9515	NC_008817.1	1.7	30.8	1905
23.	*Prochlorococcus marinus* str. NATL1A	NC_008819.1	1.86	35	2193
24.	*Prochlorococcus marinus* str. NATL2A	NC_007335.2	1.84	35.1	2162
25.	*Prochlorococcus marinus* subsp. marinus str. CCMP1375	NC_005042.1	1.75	36.4	1883
26.	*Prochlorococcus marinus* subsp. pastoris str. CCMP1986	NC_005072.1	1.66	30.8	1717
27.	*Synechococcus elongatus* PCC 6301	NC_006576.1	2.7	55.5	2523
28.	*Synechococcus elongatus* PCC 7942	NC_007604.1	2.74	55.46	2662
29.	*Synechococcus* sp. CC9311	NC_008319.1	2.61	52.4	2892
30.	*Synechococcus* sp. CC9605	NC_007516.1	2.51	59.2	2645
31.	*Synechococcus* sp. CC9902	NC_007513.1	2.23	54.2	2306
32.	*Synechococcus* sp. PCC 7002	NC_010475.1	3.41	49.16	3187
33.	*Synechococcus* sp. RCC307	NC_009482.1	2.22	60.8	2534
34.	*Synechococcus* sp. WH 7803	NC_009481.1	2.37	60.2	2533
35.	*Synechococcus* sp. WH 8102	NC_005070.1	2.43	59.4	2519
36.	*Synechocystis* sp. PCC 6803	NC_000911.1	3.95	47.35	3575
37.	*Trichodesmium erythraeum* IMS101	NC_008312.1	7.75	34.1	4451
38.	*Thermosynechococcus elongatus* BP-1	NC_004113.1	2.59	53.9	2476
39.	*Synechococcus* sp. JA-2-3B'a(2-13)	NC_007776.1	3.05	58.5	2862
40.	*Synechococcus* sp. JA-3-3Ab	NC_007775.1	2.93	60.2	2760

Table 2: General features of *Thermosynechococcus elongatus* BP-1 genome*

Optimum growth temperature (OGT)	**55°C**
Circular chromosome size (bp)	2,593,857
Plasmid	Nil
G+C content (%)	53.9
Coding nucleotide sequences (%)	89
Total genes	2525
Protein coding genes	2475
Pseudogenes	None
Number of group II entrons	28
Number of rRNA operon	1
Mean G+C content of rRNA operon (%)	55.15
Number of tRNA genes	42
Mean G+C content	57.29

*Nakamura *et al.* 2002. NCBI database

Table 3: Distribution of nucleotides across all the cyanobacterial genomes

S. No.	Organisms	A	G	C	T	Sum(ACGT)
1.	*Acaryochloris marina* MBIC11017	1405907	1331248	1313820	1416315	5467290
2.	*Anabaena variabilis* ATCC 29413	1538573	1143064	1067624	1446787	5196048
3.	*Cyanobacterium* UCYN-A	403484	210506	174459	375761	1164210
4.	*Cyanothece* sp. ATCC 51142	1346257	867109	791788	1282536	4287690
5.	*Cyanothece* sp. PCC 7424	1502278	1008937	917141	1433873	4862229
6.	*Cyanothece* sp. PCC 7425	1091994	1201893	1180603	1135874	4610364
7.	*Cyanothece* sp. PCC 7822	12251	1091019	1015263	1471555	5126825
8.	*Cyanothece* sp. PCC 8801	1209535	825555	785204	1152366	3972660
9.	*Cyanothece* sp. PCC 8802	1208114	826839	789540	1152673	3977166
10.	*Gloeobacter violaceus* PCC 7421	764343	1295200	1323902	782892	4166337
11.	*Microcystis aeruginosa* NIES-843	1402912	1048271	989867	1309777	4750827
12.	*Nostoc azollae* 0708	855969	600367	533566	817030	2806932
13.	*Nostoc punctiforme* PCC 73102	1878175	1416960	1297693	1764802	6357630
14.	*Nostoc* sp. PCC 7120	1564017	1159633	1079169	1472216	5275035
15.	*Prochlorococcus marinus* str. AS9601	557263	277795	208400	474662	1518120
16.	*Prochlorococcus marinus* str. MIT 9211	487084	323507	262948	447035	1520574
17.	*Prochlorococcus marinus* str. MIT 9215	571499	283632	212711	487739	1555581
18.	*Prochlorococcus marinus* str. MIT 9301	547866	273738	205376	467443	1494423
19.	*Prochlorococcus marinus* str. MIT 9303	547472	602621	564453	562706	2277252
20.	*Prochlorococcus marinus* str. MIT 9312	558919	278836	208654	478515	1524924
21.	*Prochlorococcus marinus* str. MIT 9313	460852	531513	499306	484180	1975851
22.	*Prochlorococcus marinus* str. MIT 9515	555935	273819	204907	476382	1511043
23.	*Prochlorococcus marinus* str. NATL1A	552543	325654	255917	491082	1625196
24.	*Prochlorococcus marinus* str. NATL2A	547288	324689	255105	487227	1614309
25.	*Prochlorococcus marinus subsp. marinus* str. CCMP1375	513110	319649	256730	466683	1556172
26.	*Prochlorococcus marinus subsp. pastoris* str. CCMP1986	537071	265189	196914	461223	1460397
27.	*Synechococcus elongatus* PCC 6301	499941	662622	664207	537152	2363922
28.	*Synechococcus elongatus* PCC 7942	508549	671407	673866	546262	2400084
29.	*Synechococcus* sp. CC9311	511489	624858	589187	546429	2271963
30.	*Synechococcus* sp. CC9605	423222	662841	647511	442437	2176011
31.	*Synechococcus* sp. CC9902	437979	565417	534753	466682	2004831
32.	*Synechococcus* sp. PCC 7002	655485	658728	682849	648815	2645877
33.	*Synechococcus* sp. RCC307	392324	651394	636540	424740	2104998
34.	*Synechococcus* sp. WH 7803	415412	673631	663039	452606	2204688
35.	*Synechococcus* sp. WH 8102	424652	673104	648200	448346	2194302
36.	*Synechocystis* sp. PCC 6803	762017	775670	737064	803685	3113436
37.	*Trichodesmium erythraeum* IMS101	1539757	950692	757088	1397018	4644555
38.	*Thermosynechococcus elongatus* BP-1	510330	631560	639547	549104	2330541
39.	*Synechococcus* sp. JA-2-3B'a(2-13)	509082	784694	758421	548221	2600418
40.	*Synechococcus* sp. JA-3-3Ab	467375	766115	754534	499210	2487234

Table 4: Frequency of different dinucleotides and J_2 index for all cyanobacteria

S. No.	Organisms	YR bases	RY bases	YY bases	RR bases	J_2 INDEX
1.	*Acaryochloris marina* MBIC11017	0.232020111	0.232020111	0.26733798	0.268621798	0.071919556
2.	*Anabaena variabilis* ATCC 29413	0.239091371	0.239091371	0.24481707	0.277000189	0.043634517
3.	*Cyanobacterium* UCYN-A	0.224160782	0.224160782	0.248451953	0.303226483	0.103356871
4.	*Cyanothece* sp. ATCC 51142	0.222043157	0.222043157	0.261742864	0.294170823	0.111827374
5.	*Cyanothece* sp. PCC 7424	0.217948644	0.217948644	0.265577427	0.298525285	0.128205424
6.	*Cyanothece* sp. PCC 7425	0.229949789	0.229949789	0.272500235	0.267600187	0.080200843
7.	*Cyanothece* sp. PCC 7822	0.222068053	0.222069269	0.259804774	0.296057904	0.111725356
8.	*Cyanothece* sp. PCC 8801	0.22121355	0.22121355	0.266512681	0.291060219	0.1151458
9.	*Cyanothece* sp. PCC 8802	0.221253581	0.221253581	0.267087486	0.290405352	0.114985674
10.	*Gloeobacter violaceus* PCC 7421	0.249100169	0.249100169	0.256570521	0.245229141	0.003599326
11.	*Microcystis aeruginosa* NIES-843	0.221254367	0.221254367	0.262797038	0.294694228	0.114982531
12.	*Nostoc azollae* 0708	0.236206376	0.236206376	0.244958284	0.282628964	0.055174495
13.	*Nostoc punctiforme* PCC 73102	0.237477368	0.237477368	0.244226582	0.280818683	0.050090529
14.	*Nostoc* sp. PCC 7120	0.238641305	0.238641305	0.245030459	0.277686931	0.045434778

15.	*Prochlorococcus marinus* str. AS9601	0.213187504	0.213187504	0.236752191	0.336872801	0.147249985
16.	*Prochlorococcus marinus* str. MIT 9211	0.219667191	0.219667191	0.247250872	0.313414746	0.121331235
17.	*Prochlorococcus marinus* str. MIT 9215	0.212970725	0.212970725	0.237311485	0.336747065	0.148117101
18.	*Prochlorococcus marinus* str. MIT 9301	0.212945875	0.212945875	0.237274344	0.336833906	0.148216501
19.	*Prochlorococcus marinus* str. MIT 9303	0.238071253	0.238071253	0.256893509	0.266963984	0.047714986
20.	*Prochlorococcus marinus* str. MIT 9312	0.213360281	0.213360281	0.237265095	0.336014343	0.146558875
21.	*Prochlorococcus marinus* str. MIT 9313	0.238541893	0.238541893	0.259211479	0.263704735	0.045832427
22.	*Prochlorococcus marinus* str. MIT 9515	0.213065553	0.213065553	0.237808082	0.336060811	0.147737786
23.	*Prochlorococcus marinus* str. NATL1A	0.216780756	0.216780756	0.24285578	0.323582709	0.132876978
24.	*Prochlorococcus marinus* str. NATL2A	0.216785768	0.216785768	0.243059565	0.323368899	0.132856927
25.	*Prochlorococcus marinus subsp. marinus* str. CCMP1375	0.219951406	0.219951406	0.244915886	0.315181301	0.120194375
26.	*Prochlorococcus marinus subsp. pastoris* str. CCMP1986	0.21328804	0.21328804	0.237368495	0.336055426	0.146847841
27.	*Synechococcus elongatus* PCC 6301	0.244484058	0.244484058	0.263722011	0.247309872	0.022063766
28.	*Synechococcus elongatus* PCC 7942	0.244430297	0.244430297	0.263938789	0.247200618	0.022278813
29.	*Synechococcus* sp. CC9311	0.238244302	0.238244302	0.261595044	0.261916352	0.047022794
30.	*Synechococcus* sp. CC9605	0.244118363	0.244118363	0.256774555	0.254988718	0.023526546
31.	*Synechococcus* sp. CC9902	0.24137059	0.24137059	0.25814059	0.259118229	0.03451764
32.	*Synechococcus* sp. PCC 7002	0.229308932	0.229308932	0.27398903	0.267393105	0.082764272
33.	*Synechococcus* sp. RCC307	0.250356651	0.250356651	0.253815089	0.245471609	-0.001426605
34.	*Synechococcus* sp. WH 7803	0.245107809	0.245107809	0.260925474	0.248858908	0.019568764
35.	*Synechococcus* sp. WH 8102	0.246242425	0.246242425	0.253482088	0.254033061	0.015030299
36.	*Synechocystis* sp. PCC 6803	0.226714224	0.226714224	0.268156875	0.278414677	0.093143104
37.	*Trichodesmium erythraeum* IMS101	0.224743861	0.224743861	0.239047926	0.311464352	0.101024555
38.	*Thermosynechococcus elongatus* BP-1	0.238353772	0.238353772	0.271678667	0.251613789	0.046584912
39.	*Synechococcus* sp. JA-2-3B'a(2-13)	0.229097455	0.229097455	0.273376191	0.2684289	0.08361018
40.	*Synechococcus* sp. JA-3-3Ab	0.230138873	0.230138873	0.273932921	0.265789333	0.079444507

Table 5: CvP bias, E+K / Q+H ration and percentage of charged and polar amino acids

S. No.	Organisms	CvP bias	E+K/Q+H	Percentage of charged amino acids (Lys, Arg, Asp, Glu)	Percentage of polar amino acids (Ser, Thr, Asn, Gln)
1.	*Acaryochloris marina* MBIC11017	-1.97685	1.170276	20.1894	22.16624
2.	*Anabaena variabilis* ATCC 29413	0.598299	1.769986	23.2869	22.6886
3.	*Cyanobacterium* UCYN-A	-0.05583	2.206228	22.04482	22.10065
4.	*Cyanothece* sp. ATCC 51142	-0.50012	1.709463	21.64598	22.14609
5.	*Cyanothece* sp. PCC 7424	-0.43311	1.649176	21.66689	22.1
6.	*Cyanothece* sp. PCC 7425	-1.23924	1.137463	19.84591	21.08515
7.	*Cyanothece* sp. PCC 7822	-0.40962	1.661456	21.51902	21.92864
8.	*Cyanothece* sp. PCC 8801	-0.7354	1.590185	21.40951	22.14491
9.	*Cyanothece* sp. PCC 8802	-0.82057	1.580827	21.35785	22.17841
10.	*Gloeobacter violaceus* PCC 7421	3.832849	1.491982	21.39258	17.55973
11.	*Microcystis aeruginosa* NIES-843	0.958792	1.793338	22.37093	21.41214
12.	*'Nostoc azollae'* 0708	-0.94061	1.564043	20.80435	21.74496
13.	*Nostoc punctiforme* PCC 73102	-1.20163	1.542101	20.99593	22.19756
14.	*Nostoc* sp. PCC 7120	-1.72489	1.456483	20.58734	22.31223
15.	*Prochlorococcus marinus* str. AS9601	2.644807	3.38893	24.40054	21.75573
16.	*Prochlorococcus marinus* str. MIT 9211	2.017014	2.287032	22.78358	20.76657

17.	*Prochlorococcus marinus* str. MIT 9215	2.865965	3.408779	24.48015	21.61419
18.	*Prochlorococcus marinus* str. MIT 9301	2.684817	3.38293	24.38381	21.69899
19.	*Prochlorococcus marinus* str. MIT 9303	0.950023	1.370177	20.97774	20.02772
20.	*Prochlorococcus marinus* str. MIT 9312	2.577503	3.34481	24.30513	21.72763
21.	*Prochlorococcus marinus* str. MIT 9313	1.32384	1.325165	20.92975	19.60591
22.	*Prochlorococcus marinus* str. MIT 9515	2.53101	3.361183	24.2393	21.70829
23.	*Prochlorococcus marinus* str. NATL1A	2.144609	2.698487	23.49543	21.35082
24.	*Prochlorococcus marinus* str. NATL2A	2.13419	2.672473	23.46639	21.3322
25.	*Prochlorococcus marinus* subsp. marinus str. CCMP1375	1.850859	2.378521	22.9055	21.05464
26.	*Prochlorococcus marinus* subsp. pastoris str. CCMP1986	2.440824	3.371427	24.18849	21.74766
27.	*Synechococcus elongatus* PCC 6301	0.123241	1.050652	20.21106	20.08782
28.	*Synechococcus elongatus* PCC 7942	0.092173	1.051858	20.20777	20.1156
29.	*Synechococcus* sp. CC9311	1.223972	1.327008	20.94273	19.71875
30.	*Synechococcus* sp. CC9605	2.739618	1.296325	21.45741	18.71779
31.	*Synechococcus* sp. CC9902	1.967803	1.274706	21.26534	19.29754
32.	*Synechococcus* sp. PCC 7002	-0.02616	1.32542	20.61765	20.64381
33.	*Synechococcus* sp. RCC307	1.549061	1.133006	20.30257	18.75351
34.	*Synechococcus* sp. WH 7803	2.110156	1.186601	20.83803	18.72787
35.	*Synechococcus* sp. WH 8102	2.458723	1.23065	21.42925	18.97052
36.	*Synechocystis* sp. PCC 6803	-0.60852	1.3778	20.316	20.92452
37.	*Trichodesmium erythraeum* IMS101	0.298044	2.06676	22.67819	22.38015
38.	*Thermosynechococcus elongatus* BP-1	0.68236	1.12199	20.05073	19.36836
39.	*Synechococcus* sp. JA-2-3B'a(2-13)	2.06356	1.180282	20.74718	18.68362
40.	*Synechococcus* sp. JA-3-3Ab	2.713967	1.2213	20.8663	18.15233