

# *In silico* subtractive hybridization of *Nostoc azollae* 0708 reveals that majority of proteins with hypothetical functions are unique in its proteome

PV Parvati Sai Arun

CR Rao Advanced Institute of Mathematics Statistics and Computer Sciences, Hyderabad, India

\*Corresponding Author: arun.uoh@gmail.com, Tel.: +91-70751-11851

Available online at: [www.isroset.org](http://www.isroset.org)

Received 24<sup>th</sup> Aug 2017, Revised 09<sup>th</sup> Sep 2017, Accepted 25<sup>th</sup> Sep 2017, Online 30<sup>th</sup> Oct 2017

**Abstract**— *Nostoc azollae* 0708 is a cyanobacterium and live as an endo symbiont. The other *Nostoc* species such as *Nostoc* PCC 7107, *Nostoc* PCC 7120, *Nostoc* PCC 7524, *Nostoc punctiforme* PCC 73102 has different adaptation styles in their life style. Their adaptation in the environment range from hot fresh water pools to soil. Availability of the sequenced proteomes in public databases has led to the application of *in silico* subtractive hybridization method to predict the unique proteins of *Nostoc azollae*, which may be responsible for its endophytic behavior. Upon careful analysis, it is found that comparative proteome analysis of *Nostoc azolle* with other considered *Nostoc* species shows that the percentages of unique protein content ranges from 4.2% to 5.2% of which majority of proteins have hypothetical functions.

**Keywords**— *Nostoc*, *In silico* subtractive hybridization, endophytic, hypothetical proteins, unique proteins.

## I. INTRODUCTION

Cyanobacteria are oxygenic photosynthetic organisms possessing a variety of metabolic pathways [1]. Several species of cyanobacteria were sequenced and deposited in public databases. These sequenced genomes serve as value resources for the researchers to address several question related to the biochemistry, ecology, adaptation of the cyanobacteria [1]. Of many sequenced cyanobacterial strains, there are few species of *Nostoc* genus. The bacterial species and strains belonging to the genus *Nostoc* show diversity in their adaptation. For example, *Nostoc* sp. PCC 7120 lives in soil, where *Nostoc* PCC 7107 lives in shallow ponds. Similarly *Nostoc* PCC 7524 has fresh water hot spring adaption whereas the *Nostoc azollae* 0708 lives as endo symbiont[2]. Here arises the fundamental question about the diversified adaptation of these *Nostoc* strains even though they belong to the same genus. *Nostoc azollae* is a endo symbiont found to be in symbiotic association with water fern *Azolla filiculoides* [3]. It is filamentous diazotrophic cyanobacterium capable of nitrogen fixation [4]. The genome sequence of *N.azollae* consists of two plasmids and one chromosome[5].

In this report, *Nostoc azollae* 0708 strain was selected as the target of interest and other *Nostoc* species as references, unique proteins for *Nostoc azollae* 0708 were predicted using *in silico* subtractive hybridization method. The study of

identification of unique genes, which are probable cause of endophytic behavior of *Nostoc azollae* 0708, by using *in silico* subtractive hybridization method, is first of its kind on *Nostoc* species.

## II. RELATED WORK

*In silico* subtraction hybridization method was first employed during the comparative study of *E.coli* and *Shigella* strains [6]. Later a web server was developed to perform parallel computing of *in silico* subtractive hybridization[7]. *In silico* subtractive hybridization method was applied in the comparative genome studies of bacteria such as *Erwinia amylovora*, *Pseudomonas savastanoi pv. glycinea* and many others[8, 9].

## III. METHODOLOGY

The \*.faa files of *Nostoc azollae* 0708, *Nostoc* sp PCC 7120, *Nostoc* PCC 7524, *Nostoc punctiforme* PCC 73102, and *Nostoc* PCC 7107 were downloaded from NCBI ftp ([ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old\\_refseq/Bacteria/](ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/)). Local database of all these proteomes were constructed. BLASTP was performed between *Nostoc azollae* as target organism and other *Nostoc* species as reference organisms considered in our analysis. The top hits containing the highest bit score were filtered from the BLAST output and homology value is calculated between each of the target protein of *Nostoc azollae* and its top hit containing highest bit score from other *Nostoc* species as

described earlier [6, 7]. The homology value is calculated as follows

$$H = I * Lh/Lq$$

Where H= Homolog value

I= percentage identity of the top hit with highest bit score from reference organisms.

Lh= protein sequence length of the top hit with highest bit score from reference organisms.

Lq= protein sequence length of the target organism.

In-house Perl scripts were developed perform the BLASTP and to calculate the homolog value.

#### IV. RESULTS AND DISCUSSION

The genome size of *Nostoc* species consider in this analysis ranges from 5.5 to 9.06 Mbp (Table 1). Out these species, *Nostoc azollae* 0708 is the smallest and *Nostoc punctiforme* PCC 73102 is the largest. The target organism *Nostoc azollae* 0708 has a total of 3,413 protein coding genes on its chromosome along with two plasmids[5].

**Table 1:** Overview of genome sequencing of *Nostoc* species.

Srain	Size (bp)	Total proteins	Total protein coding genes	Adaptation
<i>Nostoc azollae</i> 0708	5,486,145	3468	3413	Endo Symbiont
<i>Nostoc</i> sp. PCC 7120	7,211,789	5430	5366	Soil
<i>Nostoc</i> sp. PCC 7524	6,718,869	5420	5356	Fresh water hot spring
<i>Nostoc punctiforme</i> PCC 73102	9,059,191	6190	6087	Symbiotic heterotroph
<i>Nostoc</i> sp. PCC 7107	6,329,823	5329	5237	Shallow Pond

As the main goal of this report is to predict the unique proteins between the endo symbiont *Nostoc azollae* and other *Nostoc* species, the homology value which is less than or equal to 0.43 was considered for filtering the proteins. From the above prescribed cut off of homology score, It is found that, there are 229 of unique proteins between *Nostoc azollae* and *Nostoc* sp. PCC 7107 (3413 proteins of *Nostoc azollae* vs. 5329 proteins of PCC 7107) (Supplementary data). These 229 proteins may represent the core proteins of the *Nostoc azollae* species when compared to PCC 7107 (4.2%). Further analysis also indicates that when *N.azollae* is compared to *N. sp* PCC 7120(soil habitant), only 4.7% proteins fall under unique proteins with *H*-value greater than or equal to 0.43. Similarly, when *N.azollae* is compared with *N. PCC* 7524(fresh water hot spring), *N.punctiforme* PCC 73102 (Symbiotic heterotroph), 5.1%, 4.9% proteins are

unique with *H*-value greater than or equal to 0.43 respectively(Supplementary data).

**Table 2:** Total unique proteins present in *Nostoc azollae* when compared with other *Nostoc* species considered in our analysis. The proteins of *Nostoc azollae* which have homolog value less than or equal to 0.43 were considered as unique proteins.

Target strain	Reference Strains	Total proteins	Unique proteins
<i>Nostoc azollae</i> 0708 (Total 3413 proteins)	<i>Nostoc</i> sp. PCC 7120	5430	257
	<i>Nostoc</i> sp. PCC 7524	5420	278
	<i>Nostoc punctiforme</i> PCC 73102	6190	304
	<i>Nostoc</i> sp. PCC 7107	5329	229

This analysis shows that even though the genus of *Nostoc* has different species and adaptation, comparison of proteomes reveals that very few numbers of proteins are responsible for the endophytic behavior of the *Nostoc azollae*. Moreover after observing the unique proteins obtained from comparison of *Nostoc azollae* with other *Nostoc* species, it is found that majority of proteins found to be unique are hypothetical in nature, which means there is no function determined till today.

#### VI. V. CONCLUSION AND FUTURE SCOPE

In this report, application of *in silico* subtractive hybridization on *Nostoc* proteomes is first of its kind. The analysis done in this report opens a new gate way for the researchers to study about the adaptation of the *Nostoc* species and also to characterize the hypothetical proteins in relation its adaptation and survival.

None

#### VII. REFERENCES

- [1] P.V. Arun, R.K. Bakku, M. Subhashini, P. Singh, N.P. Prabhu, I. Suzuki, J.S. Prakash, "CyanoPhyChe: a database for physico-chemical properties, structure and biochemical pathway information of cyanobacterial proteins", PLoS One, Vol. 7, e49425, 2012.
- [2] R. Kottmann, I. Kostadinov, M.B. Duhaime, P.L. Buttigieg, P. Yilmaz, W. Hankeln, J. Waldmann, F.O. Glockner, "Megx.net: integrated database resource for marine ecological genomics", Nucleic Acids Res, Vol. 38, D391-395, 2010.
- [3] D.G. Adams, B. Bergman, S.A. Nierzwicki-Bauer, P.S. Duggan, A.N. Rai, A. Schüßler, Cyanobacterial-Plant Symbioses, in: E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, F. Thompson (Eds.) "The Prokaryotes: Prokaryotic Biology and Symbiotic Associations", Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 359-400, 2013.
- [4] R. Rippka, J. Deruelles, J.B. Waterbury, M. Herdman, R.Y. Stanier, "Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria", Microbiology, Vol. 111, pp.1-61, 1979.
- [5] L. Ran, J. Larsson, T. Vigil-Stenman, J.A. Nylander, K. Ininbergs, W.W. Zheng, A. Lapidus, S. Lowry, R. Haselkorn, B. Bergman, "Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium", PLoS One, Vol. 5 e11486, 2010.

- [6] S. Fukiya, H. Mizoguchi, T. Tobe, H. Mori, “*Extensive genomic diversity in pathogenic Escherichia coli and Shigella Strains revealed by comparative genomic hybridization microarray*”, J Bacteriol, Vol. 186 pp. 3911-3921, 2004.
- [7] Y. Shao, X. He, E.M. Harrison, C. Tai, H.Y. Ou, K. Rajakumar, Z. Deng, “*mGenomeSubtractor: a web-based tool for parallel in silico subtractive hybridization analysis of multiple bacterial genomes*”, Nucleic Acids Res, Vol. 38, W194-200, 2010.
- [8] Y. Zhao, M. Qi, “*Comparative Genomics of Erwinia amylovora and Related Erwinia Species-What do We Learn?*”, Genes (Basel), Vol. 2, pp.627-639, 2011.
- [9] M. Qi, D. Wang, C.A. Bradley, Y. Zhao, “*Genome sequence analyses of Pseudomonas savastanoi pv. glycinea and subtractive hybridization-based comparative genomics with nine pseudomonads*”, PLoS One, Vol. 6, e16451, 2011.

#### **AUTHORS PROFILE**

Dr. PV Parvati Sai Arun pursued B. Tech in Biotechnology., M.Tech in Bioinformatics., and Ph.D. Computational Biology from JNTU-Hyderabad and University of Hyderabad respectively in 2007, 2009 & 2015. He is currently working as Post-Doctoral Fellow from CR Rao Advanced Institute of Mathematics Statistics and Computer Sciences, Hyderabad Since 2015 He has published more than 5 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Computational genomics of cyanobacteria, Host pathogen interaction of Mycobacterium tuberculosis, Statistical genomics of NGS data and Metagenomics.