THE UNIVERSITY of EDINBURGH
School of Biological Sciences

[Using phylogenetic profiles to study functional associations between cyanobacterial proteins and nitrogen fixation]

[Comparative and Evolutionary Genomics]

Student Exam Number: …B196466……

The file name of your uploaded document **must** include your exam number

1.Introduction

Nitrogen-fixing cyanobacteria reduce free molecular nitrogen in the atmosphere to nitrogen compounds that can be used by plants through the action of nitrogenase in their cells. The nitrogenase complex consists of molybdenum ferritin, encoded by the nifD and nifK genes, and ferritin, encoded by the nifH gene.[1] Nitrogenase and nifH genes exist only in nitrogen-fixing microorganisms, while nitrogenase is encoded by the nitrogen-fixing gene nif, which has a high degree of interspecies homology.[2] In the molybdenum-iron nitrogenase system, the nifH gene encodes ferritin and is also involved in the biosynthesis of iron-molybdenum cofactor (FeMoco) and directs the incorporation of FeMoco into molybdenum-ferritin. NifH is a relatively conserved functional gene among all biological nitrogen-fixing species so this gene is the main research object of this experiment.[3]

In this experiment, phylogenetic profiles were used to predict which proteins in cyanobacterial proteins were functionally related to nitrogen fixation, by finding orthonormal groups functionally related to nifH. Since the phylogenetic mapping method is based on homology search, the results largely depend on the protein sequences in the reference organisms. This time, ten cyanobacterial taxa were selected, of which four taxa can fix nitrogen, and six cannot fix nitrogen. Phylogenetic maps rely on protein-related evolution; when two proteins share a phylogenetic profile, they are evolutionarily related. Phylogenetic profiles can only be calculated accurately when several complete proteomes are compared, and two proteins with similar phylogenetic profiles are considered functionally related. Therefore, protein clustering based on phylogenetic profiles can provide functional information about the unknown protein when grouped with one or more proteins with known functions.[4] Sequence similarity relationships among protein family members contain information about their evolutionary history, such as the relative timing of horizontal transfer events, and the differential acceleration or deceleration of specific organisms in response to selection pressures.[5] This paper presents the representation and comparison of phylogenetic profiles and uncovers correlations between protein functional associations.

2.Methods

2.1 Preparation of protein sets

Select ten suitable cyanobacterial types from one literature, ensure that half of them are nitrogen-fixing cyanobacteria, and ensure that all ten taxa have genome-wide RefSeq proteomes.[6] These ten genome-wide RefgSeq proteomes were then downloaded from NCBI. The ten selected protein names were 'Nostoc azollae' 0708, Crocosphaera watsonii WH 8501, Lyngbya sp. PCC 8106, Synechococcus sp. JA-3-3Ab, Acaryochloris marina MBIC11017, Gloeobacter violaceus PCC 7421, Microcystis aeruginosa NIES-843, Synechococcus sp. RS9917, Arthrospira maxima CS-328 and Synechococcus thermophila BP-1.[6] The first four are nitrogen-fixing cyanobacterial proteins, and the latter six are non-nitrogen-fixing cyanobacterial proteins. Results are shown in Table 1.

species name	type	Rank	NCBI BLAST name	Taxonomy ID
'Nostoc azollae' 0708	Nitrogen fixation	strain	cyanobacteria	551115
Arthrospira maxima CS-328	Non-Nitrogen fixation	strain	cyanobacteria	513049
Crocosphaera watsonii WH 8501	Nitrogen fixation	strain	cyanobacteria	165597
Lyngbya sp. PCC 8106	Nitrogen fixation	species	cyanobacteria	313612
Synechococcus sp. JA-3-3Ab	Nitrogen fixation	species	cyanobacteria	321327
Acaryochloris marina MBIC11017	Non-nitrogen fixation	strain	cyanobacteria	329726
Gloeobacter violaceus PCC 7421	Non-nitrogen fixation	strain	cyanobacteria	251221
Microcystis aeruginosa NIES-843	Non-nitrogen fixation	strain	cyanobacteria	449447
Synechococcus sp. RS9917	Non-nitrogen fixation	species	cyanobacteria	221360
Synechococcus thermophila BP-1	Non-nitrogen fixation	species	cyanobacteria	2022662

Table 1 Details of the 10 taxa

2.2 Predicted orthogonal groups and Obtaining a phylogenetic map

Orthofinder(Emms and Kelly 2015) takes the genome-wide protein set as input, finds orthogroup and direct homologs, infers rooted gene trees for all Orthogroup, and identifies all gene duplication events in these gene trees.[7] It inferred a rooted species tree for the species being analyzed and mapped gene duplication events from gene trees to branches of the species tree. It also provided comprehensive statistics for comparative genomic analysis and output predicted orthogroup with a default expansion value of 1.5. Finally, the output of OrthoFinder is converted into a phylogenetic map utilizing a script.[8]

2.3 Discovery of the phylogenetic profile of the nifH gene

Search by keyword for matches of cyanobacterial nifH proteins in 10 proteome combinations, and then query the Pfam database for this protein.[9] Check the Pfam results to see if the domain content indicates that the protein is correctly annotated, and select an annotated nifH protein for a BLASTP query to find the top few best matches and find their corresponding orthogroups. Finally, this orthogonal group was asked if it contained other proteins that were also annotated as cyanobacterial nifH proteins.

2.4 Finding Orthogroup with phylogenetic profiles similar to the nifH phylogenetic profile

For two strings of equal length, the Hamming distance refers to the number of characters that are not the same. It characterizes the similarity between the two sequences to a certain extent. First, convert the sequences into binary numbers, and then The Hamming distance is then

obtained by calculating the number of different corresponding position values. Generally, the low Hamming distance in the phylogenetic spectrum is selected. In the best case, when the Hamming distance is 0, the series of orthogonal groups will have functional connections. The proteins in these functionally related orthogroups are then queried, their functions are studied, and these known functional connections are classified.[10]

3.Result

3.1 OrthoFinder predicts orthogroup

Using OrthoFinder to predict orthogroups within the ten cyanobacterial taxa previously found, the following results were obtained: A total of 42,552 genes, 34,556 in the orthologue and 7,996 unassigned. The number of orthologues was 4820, the number of species-specific orthologues was 32, and the number of genes in the species-specific orthologue was 276.

3.2 Discovery of the phylogenetic profile of the nifH gene

In a combined file of 10 cyanobacterial protein sequences, look up the cyanobacterial nifH protein by the keyword "nitrogenase ferritin". A total of 6 matches were found and the results are shown in Figure 1.

```
>WP_007305800.1 nitrogenase iron protein [Crocosphaera watsonii]
>WP_009784199.1 nitrogenase iron protein [Lyngbya sp. PCC 8106]
>WP_013192381.1 nitrogenase iron protein [Trichormus azollae]
>WP_013192322.1 nitrogenase iron protein [Trichormus azollae]
>WP_013190628.1 nitrogenase iron protein [Trichormus azollae]
>WP_011430651.1 MULTISPECIES: nitrogenase iron protein [unclassified Synechococcus]
```

Figure 1 Six sequences with the cyanobacterial nifH gene

By querying these six proteins in the Pfam database, it was found that the six protein sequences were correctly annotated, and the protein sequence "WP_007305800.1" was selected as an example, as shown in Table 2.

Family	Description	Entry type	Clan	HMM length	Bit score	E-value
Fer4 NifH	4Fe-4S iron sulfur cluster binding proteins, NifH/frxC family	Domain	CL0023	271	423.3	8.0e-130

Table 2 Pfam result of protein sequence "WP_007305800.1"

Select the protein sequence named "WP_007305800.1" in Figure 1 and perform a BLASTP search on it. The BLASTP results are shown in Figure 2.

```
Database: User specified sequence set (Input: protein_together.fasta).
           42,552 sequences; 12,776,417 total letters


Query= WP_007305800.1 nitrogenase iron protein [Crocosphaera watsonii]

Length=296
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

WP_007305800.1 nitrogenase iron protein [Crocosphaera watsonii]       603     0.0
WP_009784199.1 nitrogenase iron protein [Lyngbya sp. PCC 8106]        534     0.0
WP_013190628.1 nitrogenase iron protein [Trichormus azollae]          466     2e-167
WP_013192381.1 nitrogenase iron protein [Trichormus azollae]          462     2e-165
WP_013192322.1 nitrogenase iron protein [Trichormus azollae]          447     8e-160
WP_011430651.1 MULTISPECIES: nitrogenase iron protein [unclassifi...  436     6e-156
WP_038001202.1 MULTISPECIES: ferredoxin:protochlorophyllide reduc...  155     2e-45
WP_164921031.1 ferredoxin:protochlorophyllide reductase (ATP-depe...  143     4e-41
WP_002759152.1 ferredoxin:protochlorophyllide reductase (ATP-depe...  140     6e-40
WP_007305348.1 ferredoxin:protochlorophyllide reductase (ATP-depe...  140     8e-40
WP_041439195.1 MULTISPECIES: ferredoxin:protochlorophyllide reduc...  139     2e-39
WP_013191289.1 ferredoxin:protochlorophyllide reductase (ATP-depe...  138     7e-39
WP_010468480.1 MULTISPECIES: ferredoxin:protochlorophyllide reduc...  137     1e-38
WP_011142366.1 ferredoxin:protochlorophyllide reductase (ATP-depe...  136     2e-38
WP_009787956.1 ferredoxin:protochlorophyllide reductase (ATP-depe...  134     2e-37
WP_006624178.1 MULTISPECIES: ferredoxin:protochlorophyllide reduc...  134     2e-37
WP_012264281.1 AAA family ATPase [Microcystis aeruginosa]             43.5    9e-05
```

Figure 2 BLASTP result of sequence "WP_007305800.1"

As can be seen from Figure 2, Crocosphaera watsonii and Lyngbya sp. PCC 8106 were the two best matches with an E-value of 0 and the highest scores. The orthogroup was found by searching for the corresponding IDs of the two proteins in the Orthogroup, and the results are shown in Table 4. These two proteins are found in the same Orthogroup “OG0002254”. In this orthogroups, also contain “WP_013190628.1”, “WP_013192322.1”, “WP_013192381.1”, and “WP_011430651.1”. Similarly, finding the sequences of these four proteins and retrieving the four IDs in the Pfam database, the result shows that all proteins were annotated as nifH. Furthermore, Table 5 shows the phylogenetic profile of orthogroup “OG0002254”.

Another finding was that these four proteins, together with the two proteins generated by the previous BLASTP results, were the six genes containing the cyanobacterial nifH protein keyword generated by the initial search of all proteins and that these six genes were found to be in an orthogroup.

OG0002254	WP_013190628.1	WP_013192322.1	WP_013192381.1	WP_007305800.1	WP_009784199.1	WP_011430651.1

Table 4 orthogroup “OG0002254” which is an orthogroup of proteins “Crocosphaera watsonii” and “Lyngbya sp. PCC 8106”

orthogroup	'Nostoc_azollae'_0708	Acaryochloris_marina_MBIC11017	Arthrospira_maxima_CS-328	Crocosphaera_watsonii_WH_8501
OG0002254	0	1	0	0

Gloeobacter_violaceus_PCC_7421	Lyngbya_sp.PCC_8106	Microcystis_aeruginosa_NIES-843	Synechococcus_sp.JA-3-3Ab	Synechococcus_sp.RS9917	Thermosynechococcus_vestitus_BP-1
1	0	1	0	1	0

Table 5 the phylogenetic profile of orthogroup “OG0002254”

3.3 Finding orthogroup with phylogenetic profiles similar to the nifH phylogenetic profile

The results in Table 6 show all orthogroup with a Hamming distance of 0, including the previously

found orthogroup OG0002254 and 11 other orthogroups. Furthermore, it also shows several orthogroup with Hamming distance 1.

There are 12 orthologous clusters in Table 6 with a Hamming distance of 0. These orthologues were predicted to be functionally associated with nifH and functionally associated with nitrogen fixation. These 12 orthogroups were then searched in the Pfam database. The results show that orthogroup “OG0002254” is all the 4Fe-4S iron-sulfur cluster binding proteins and belongs to NifH/frxC family, which is encoded by the nifH gene, orthogroup “OG0002820” is all the NifT/FixU protein, orthogroup “OG0002856”,” OG0002857”,”OG0002872”,”OG0002874” is all the Nitrogenase component 1 type Oxidoreductase, orthogroup “OG0002873” is all the Nitrogen fixation protein NifW, orthogroup “OG0002902” is all the NifZ domain and the protein function of orthogroup “OG0002902” is unknown.

So orthogroup “OG0002856”,” OG0002857”,”OG0002872”, and ”OG0002874” has the function linkages, and their common function is they are the Nitrogenase component 1 type Oxidoreductase. This four orthogroups can be divided into one category, and the others mentioned above each form a category. In addition to the orthogroups “OG0002902”, this orthogroups may be divided into a separate class because the protein function is unknown, or it may be grouped into other orthogroups.

orthogroup	**Hamming Distance**
OG0002254	0
OG0002820	0
OG0002856	0
OG0002857	0
OG0002870	0
OG0002871	0
OG0002872	0
OG0002873	0
OG0002874	0
OG0002877	0
OG0002888	0
OG0002902	0
OG0000123	1
OG0001285	1
OG0001825	1

Table 6 12 orthogroups whcih Hamming Distance is 0 and 3 orthogroups whcih Hamming Distance is 1

4.Discussion

From the results, it can be inferred that 4Fe-4S iron-sulfur cluster binding proteins, NifT/FixU protein, Nitrogenase component 1 type Oxidoreductase, Nitrogen fixation protein NifW, and NifZ domain are all related to the nifH gene and associated with nitrogen fixation. The two components of nitrogenase, molybdenum-iron protein (MoFe protein) and ferritin (Fe protein), have been identified in the literature.[11] Neither of these proteins has nitrogenase activity alone,

but only when they are integrated together to form a complex does it have nitrogenase activity.[12] The Fe protein is encoded by the nifH gene and has three functional sites on the Fe protein, one of which is the 4Fe-4S iron-sulfur cluster binding protein. Several homologous domains with other prokaryotic symbiotic genes (nifH, fixA, fixU, and nifT) were also found in these loci, as well as the Nitrogenase component 1 type Oxidoreductase, which belongs to the second component of the nitrogen-fixing enzyme, also known as diazoxide reductase.[13] These are evidence of the functional linkage between these proteins. Finally, the literature also shows that "NafH, NifW and NifZ are co-purified with MoFe proteins produced by A. vinelandii strains," which also demonstrates the functional link between NifW and NifZ and nitrogenase.[14]It can be seen that the orthogroups calculated from the Hamming distance all carry proteins that are functionally linked to the NifH gene and act as nitrogen fixers, and the orthogroups are excellent predictors.

The limitations of this experiment are that the number of protein sets selected is insufficient and the ten cyanobacterial protein taxa are too few. Criteria could be set for the selection of protein sets, and in addition a weight-based phylogenetic profile could be constructed, thus indirectly reducing the over-reliance on the selection of reference genomes. The method used to predict the functional linkage of proteins is relatively homogeneous, and other approaches such as cluster analysis can be adopted, using hierarchical clustering algorithms and K-mean clustering algorithms, which can be used to achieve better clustering results by taking advantage of the strengths and weaknesses of these two methods. The future direction is that this experiment can be used for functional linkage studies in other species to predict the relationship between some disease-causing proteins and to determine their relationship with the disease itself, in order to promote the development of vaccines and drugs.

References

[1] Wang, Yue, et al. "Microalgal hydrogen production." Small Methods 4.3 (2020): 1900514.
[2] Hennecke, H., et al. "Concurrent evolution of nitrogenase genes and 16S rRNA in Rhizobium species and other nitrogen fixing bacteria." Archives of Microbiology 142.4 (1985): 342-348.
[3] Gaby, John Christian, and Daniel H. Buckley. "A comprehensive aligned nifH gene database: a multipurpose tool for studies of nitrogen-fixing bacteria." Database 2014 (2014).
[4] Eisenberg, David, et al. "Protein function in the post-genomic era." Nature 405.6788 (2000): 823-826.
[5] Ramarathnam, Rampriya, and Shankar Subramaniam. "Phylogenomics of Orthologous Protein Families in Prokaryotes: Comparison of Evolutionary Profiles." Current Bioinformatics 9.2 (2014): 113-131.
[6] Latysheva, Natasha, et al. "The evolution of nitrogen fixation in cyanobacteria." Bioinformatics 28.5 (2012): 603-606.
[7] Emms, David Mark, and Steven Kelly. "SHOOT: phylogenetic gene search and ortholog inference." Genome biology 23.1 (2022): 1-13.
[8] Barker D (2022) makephyprofiles.sh. Unpublished coursework material, Comparative and Evolutionary Genomics (PGBI11115), University of Edinburgh.
[9] Baliga, Nitin S., et al. "Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea." Genome research 14.11 (2004): 2221-2234.

[10] Pellegrini, Matteo, et al. "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proceedings of the National Academy of Sciences 96.8 (1999): 4285-4288.
[11] Hänsch, Robert, and Ralf R. Mendel. "Physiological functions of mineral micronutrients (cu, Zn, Mn, Fe, Ni, Mo, B, cl)." Current opinion in plant biology 12.3 (2009): 259-266.
[12] Jeong, Ho-Sang, and Yves Jouanneau. "Enhanced nitrogenase activity in strains of Rhodobacter capsulatus that overexpress the rnf genes." Journal of bacteriology 182.5 (2000): 1208-1214.
[13] Presta, Luana, et al. Molybdenum cofactors and their role in the evolution of metabolic pathways. Springer, 2015.
[14] Jimenez-Vicente, Emilio, et al. "Sequential and differential interaction of assembly factors during nitrogenase MoFe protein maturation." Journal of Biological Chemistry 293.25 (2018): 9812-9823.