

Quantitative Exploration of the Occurrence of Lateral Gene Transfer Using Nitrogen Fixation Genes as a Case Study

by Jason Lin

Advisor: Professor Peter Bickel

Introduction

Under the concept of evolution, organisms constantly evolve in order to stay fit to the environment and predefined conditions. Such idea has existed and remained essential in biology and science for years, ever since Darwin's theory of evolution and natural selection. However, the tree-based view of evolutionary relationships in the Darwinian model cannot explain some similarities observed among otherwise distantly related bacterial organisms. Similarly to the discovery and formation of quantum mechanics, some biologists have recognized that the process of lateral gene transfer (LGT) may explain genetically the reason that some traits proliferate in distinct organisms. Through a series of basic statistical analyses on different genomes against established results of LGT, we developed a method for inferring the possibility of lateral gene transfer.

The Darwinian Evolution

Traditionally the idea of evolution assumes that genes are inherited in a top-down fashion, such that a species has acquired its genes from an ancestral species, which had presumably undergone a series of natural selection and only the fittest survived and reproduced offspring. Through this model, a species inherits its genes from an immediate ancestor, who inherits its genes from its immediate ancestor species. In essence, the Darwinian mechanism drives evolution in a "vertical" fashion. Genetically this mechanism is often referred as *vertical gene transfer*.

Lateral Gene Transfer

The basic concept of lateral gene transfer (LGT) relies on the possibility that through interspecific interactions some individuals may acquire genes from other organisms nearby. With the presence of a vector, such as a virus, a gene can be extracted from an individual and later injected inside the host. The new gene will be retained through several later generations when the new additions prove to aid the survival of the individual's offspring. Otherwise, if the gene is extraneous or offers no selective advantage it may be lost from the population. Sometimes a gene is transferred, but the gene is lost. Other times, it is accepted and it is incorporated in the genome. This phenomenon is referred to as lateral gene transfer. LGT occurs when the outcome favors evolutionary selection if the inserted gene appears advantageous to host, such as proteins that maximize nutrient uptake, develop antibiotic resistance, or maintains better and overall livelihood of the organism. In addition, transfer of genes between organisms as described above is a random process that happens frequently, resulting in a higher possibility of success of LGT.

Recently, genetic analyses, using the increasing amount of available genomic data, are providing new evidence that LGT may play more of a role than previously accepted "Vertical transfer" has been the more dominant theory, but now LGT has appeared to provide more evidence for

acceptance in general. The difficulty in identifying occurrences of LGT is that its nature of gene acquisition has also made it hard to differentiate from the traditional descent mechanisms that many have come to accept over the years. It is difficult to detect because the results of various tests for LGT may sometimes be explained by different biological phenomena. Gene loss, where genes have mutated so that they are no longer recognizable, is often used instead of LGT as an explanation. However, more research has shown that LGT does occur in various species, with prokaryotic organisms being the most prevalent due to their biological simplicity, such as the absence of nuclear boundaries between its genetic information and the cytoplasm, and the form of a rigid and chemically more complex cell wall.

Detection of LGT

There are several methods that have been proposed as ways to detect LGT. A typical method is to compare the evolutionary trees of different genes, as LGT should cause different genes to have different phylogenetic trees, therefore a comparison of those evolutionary trees based on gene distribution would reveal the effect of LGT. In addition, finding several close species matches and comparing their lineage distances may also reveal LGT, as if two species are evolutionary distant yet a genetic analysis show some similarities, then LGT may have taken place for a particular gene. Other mechanisms often include a study of gene distributions across different organisms, “unusual” nucleotide compositions or behaviors in some regions of species genome, or often some combinations of different methods to minimize the shortcomings of individual methods.

The Experiment

For this experiment, BLAST (Basic Local Alignment Search Tool) is used to find the best alignment and similarity score between two sequences, the *query* and the *subject*. After finding similarities, BLAST calculates the statistical significance of the matches from the raw scores and matches in relation to the likelihood of finding a particular sequence at random. For the experiment, the expected values of this probability, or E-value, are taken for comparison purposes in the analysis. By the statistical software R, the results are collected and used for plots and qualitative comparisons. Through various case studies, the results of BLAST comparisons are used to make plausible statistical conclusions.

Some evidence of lateral gene transfer is seen from the statistically significant matches and similarities in gene expressions in organisms. By repeating similar procedures on different case studies and comparing the results to both positive and negative controls, sufficient evidence for LGT may surface and suggest further analysis on a biological level. The goal is to develop a method that can be used to detect lateral gene transfer through statistical methods. Below, we discuss our positive and negative control sequence sets for LGT. We have used these data to evaluate the behavior of BLAST scores in the absence or presence of LGT. We compare these results with a similar analysis on a third data set where there is debate about the occurrence of LGT. Finally, we use another sequence set, 16S rRNA, as a basis for relative distance between organisms.

Case Studies

16S rRNA

16S is used traditionally to construct the tree of life. The tree of life assumes a vertical evolutionary model and displays the phylogenies of different organisms from a common ancestor. Quantitatively 16S rRNA is used to determine relative distances between organisms, and the results are used to model possible lineages. When the E-values of the 16S rRNA BLAST scores are compared, the closer the value is to the control organism, the closer the organism is in phylogeny. Due to its highly conserved nature and crucial role for life functions, 16S rRNA is acquired through the traditional vertical fashion from an ancestral species, and less likely to be laterally transferred. 16S is part of a very important complex and needed for basic functions and survival of life. Due to this inflexibility and conservation, LGT is not likely to have occurred since an organism likely has very protective mechanisms to prevent mutations and changes to such a critical element.

Ribosomal Proteins

Four ribosomal proteins *rpsA* (S1), *rpsB* (S2), *rpsC* (S3), and *rpsD* (S4) are studied in the experiment as a negative control for LGT. Because they are highly conserved like the 16S rRNA and critical for basic livelihood functions, the ribosomal proteins are typically not laterally transferred, thus making them suitable as negative controls.

CTX-M

The gene coding for the *CTX-M* enzyme is determined to be a highly probable candidate for LGT. *CTX-M* has a great function in antibiotic resistance, and with the relatively rapid evolution of different bacterial antibiotic resistance the probability of acquisition via LGT cannot be overlooked. A group of organisms are selected to make up the positive control group for LGT, in contrast to the ribosomal proteins which make up the negative control. The extremely low E-values among the collection of organisms with *CTX-M* variants imply that *CTX-M* is highly conserved across the spectrum of organisms. In a comparison with 16S rRNA, the results also demonstrate that the organisms with *CTX-M* may not be close in lineages.

Nitrogen Fixation

Nitrogen fixation is the process of fixing atmospheric nitrogen to nitrates, an usable form for plants. This process depends on a series of genes in various species of bacteria. Nitrogen fixation provides a recycling of materials in the nitrogen cycle, similar to the carbon cycle. There are many genes involved in this process, but there are four necessary genes, namely *nifD*, *nifK*, *nifE* and *nifN*. An organism cannot perform nitrogen fixation in the absence of those genes. Since most nitrogen fixers reside in the soil or similar environments, it is probable that LGT may have been the mechanism of gene inheritance. However, there is debate in the literature about the occurrence of LGT for this trait.

Additional Organisms

Some additional organisms are introduced to the case studies as a way of verification. Based on a couple of articles, a select few genomes are obtained for BLAST comparison tests. The purpose of acquiring additional organisms is to understand whether it is possible to assume that LGT has occurred in the presence of a particular gene, based on known habitat and biological background.

Results

The BLAST comparisons of *Nostoc sp. PCC 7120* with the TIGR genome library produces the following distributions (only the top twenty significant matches are listed) as figures 1a-d:

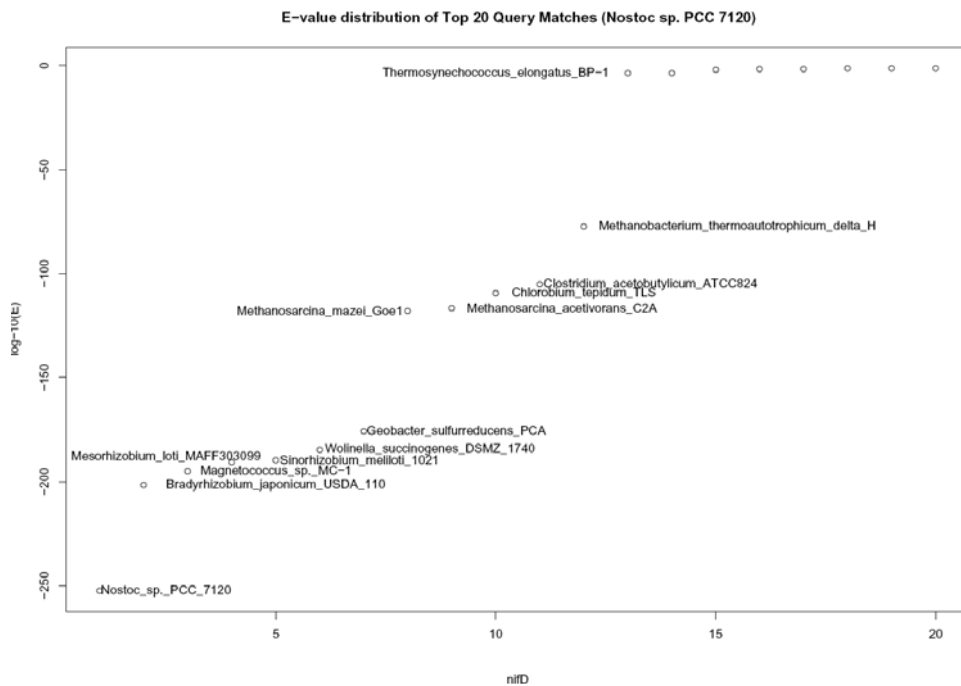


Fig 1a. *nifD* Comparison

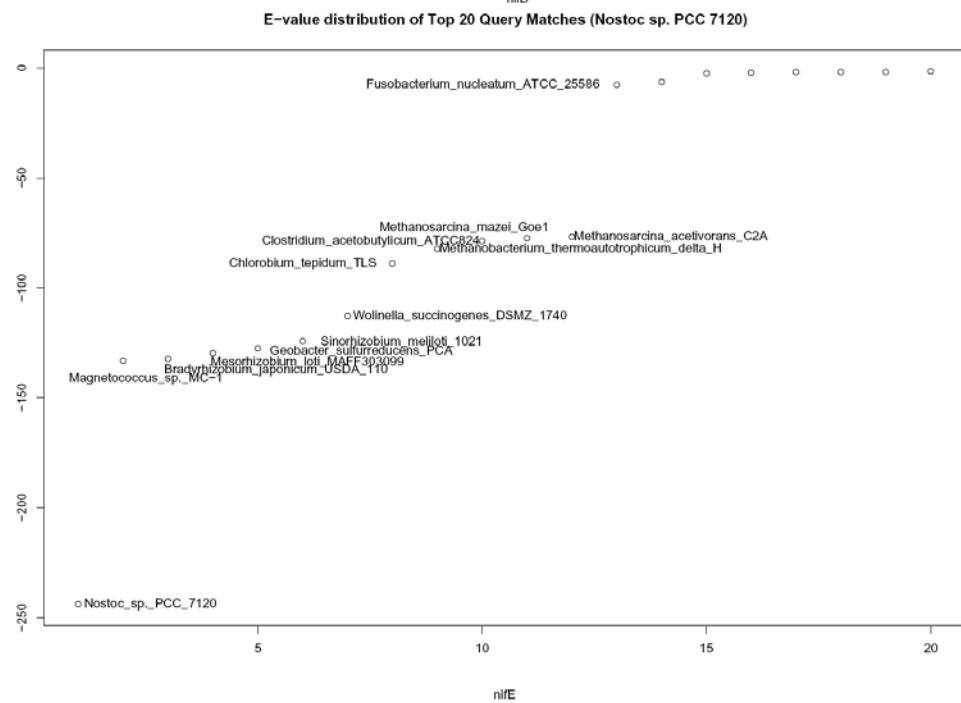


Fig 1b. *nifE* Comparison

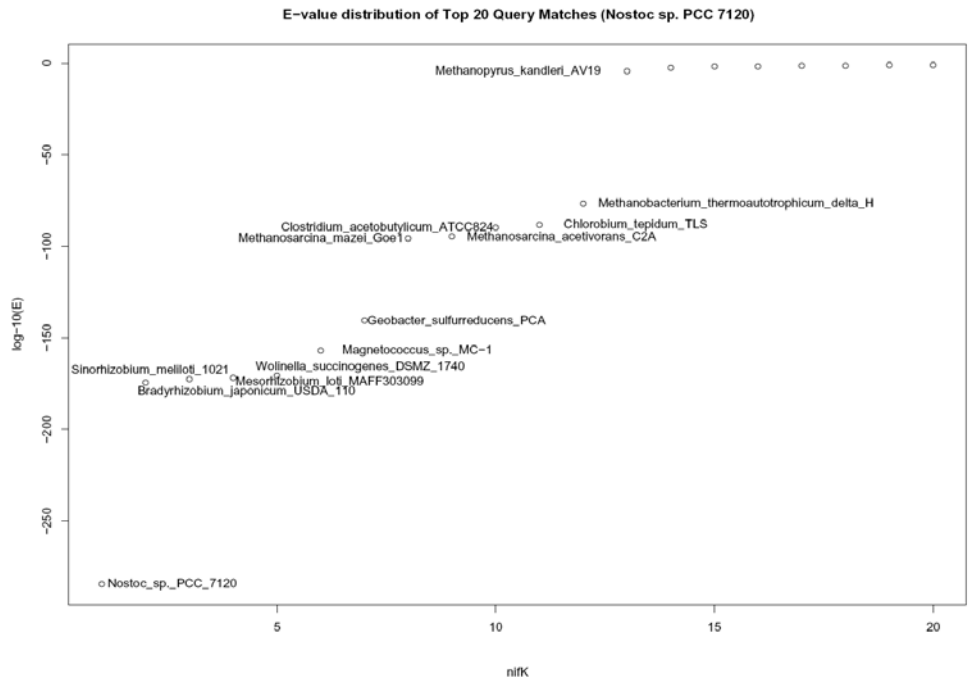


Fig 1c. *nifK* Comparison

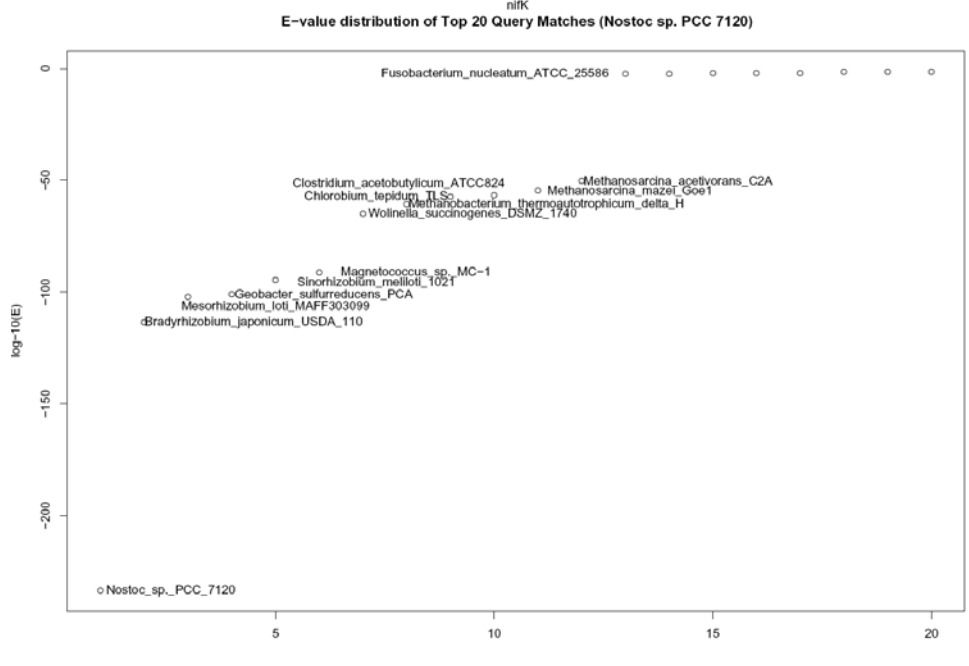


Fig 1d. *nifN* Comparison

Figure 1a-d. A distribution of the top 20 BLAST matches against select genes in *Nostoc sp. PCC7120*. Each figure contains a comparison of E-values (in logarithmic scale) to a reference BLAST result from *Nostoc sp. PCC7120* as labeled. A more negative value indicates that the gene sequence is closer to that of the reference by BLAST comparison criteria.

From the distributions above, we have observed that those that are previously identified as positive nitrogen fixers systematically appear in the top twenty significant matches for all four of the *nif* genes. These top matches are:

Nostoc sp. PCC7120 (control)
Bradyrhizobium japonicum USDA 110
Chlorobium tepidum TLS
Clostridium acetobutylicum ATCC824
Dehalococcoides ethenogenes 195(addition)
Desulfovibrio vulgaris Hildenborough (addition)
Fusobacterium nucleatum ATCC 25586
Geobacter sulfurreducens PCA
Magnetococcus sp. MC-1
Mesorhizobium loti MAFF303099
Methanobacterium thermoautotrophicum delta H
Methanosarcina mazel Goe1
Methanosarcina acetivorans C2A
Sinorhizobium meliloti 1021

From the distributions in Figure 1, we also observe that among these significant matches, there are two distinct clusters with no overlapping, when they are evaluated with *Nostoc sp. PCC 7120* as control. This is a hint that members of these two separate groups are closer to members within the group than those outside the group, perhaps in phylogenetic terms and also ecological reasons. However, within a group we also observe a rather diverse distribution of species and therefore it is more likely that due to their similar environmental preference and ecological roles that these clusters are formed, and that LGT may have occurred after the formation of these clusters. Their closeness in habitat may have led to the acquisition of nitrogen fixation genes through LGT out of necessity and amelioration. In spite of those with significant BLAST matches, most are distant relatives, as revealed in the *16S-nif* correlation plot (fig 2):

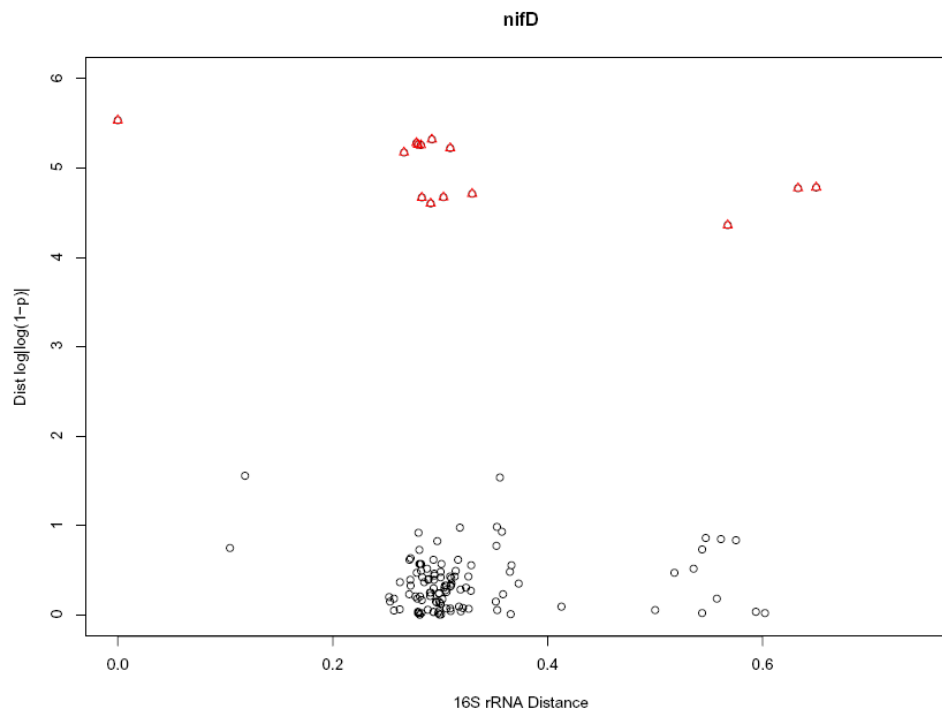
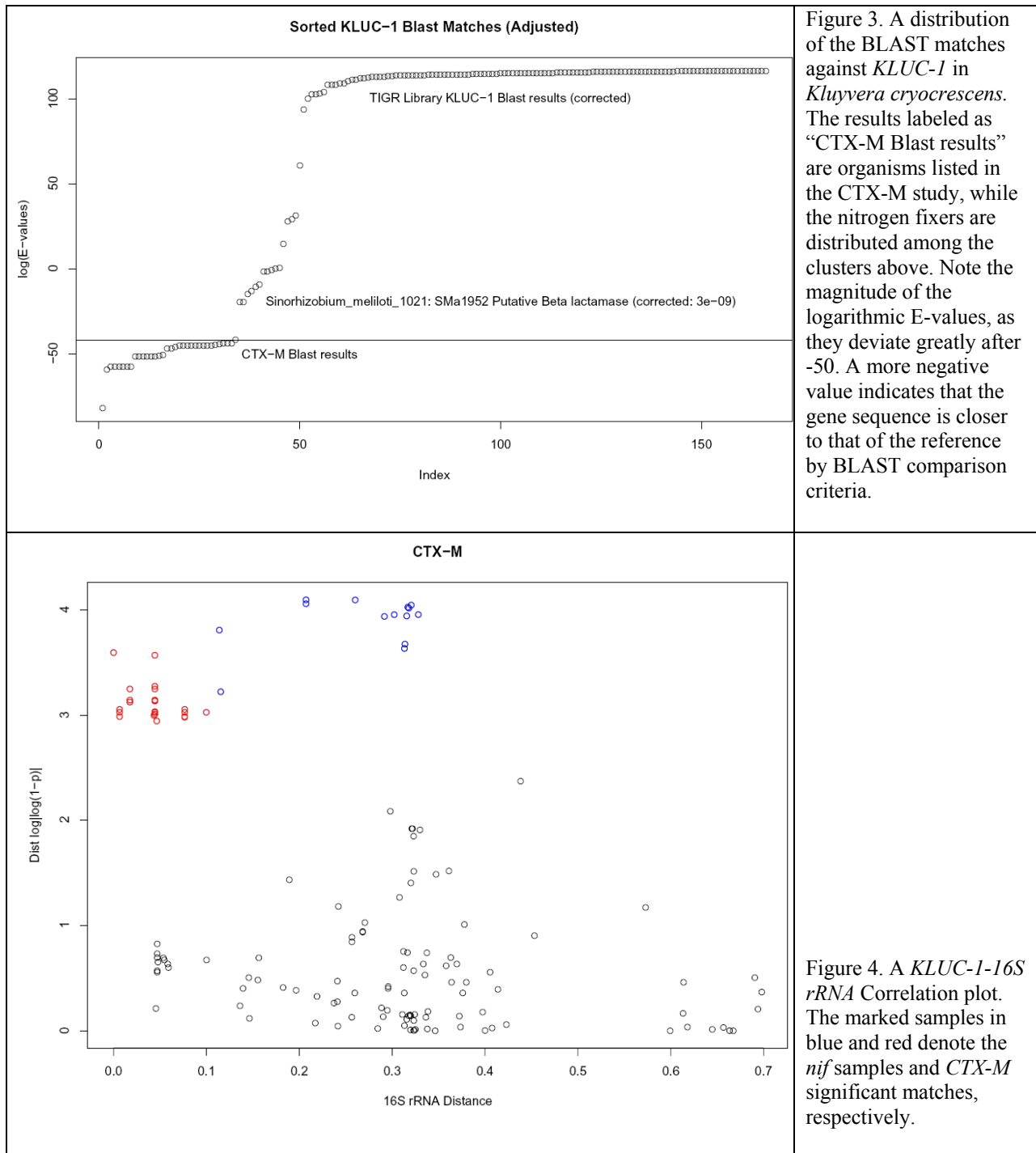


Figure 2. A *nifD*-*16S* rRNA Correlation plot. The marked samples in red denote the significant matches as observed in Figure 1.

From figure 2, we see that among the significant matches (marked in red), not all of them are tightly clustered as compared to the rest of genomes in the TIGR library (in black) below. A tight clustering in the bottom banding implies normality such that their *16S rRNA* distance is closely distributed about a mean (hence the large cluster).

We also observe similar trends from the *CTX-M* matches, as they are, previously established, highly probable candidates for lateral gene transfer.



Figures 3 and 4 both produce similar results compared to those obtained from the *nif* group. Since *CTX-M* is a highly probably LGT candidate, this qualitative comparison suggests that the *nif* genes are also likely LGT candidates. This suggestion is well in line with the fact that the organisms found in the significant *nif* results coexist in very similar habitats.

A comparison to ribosomal proteins (S1 – S4) yields an entirely different result:

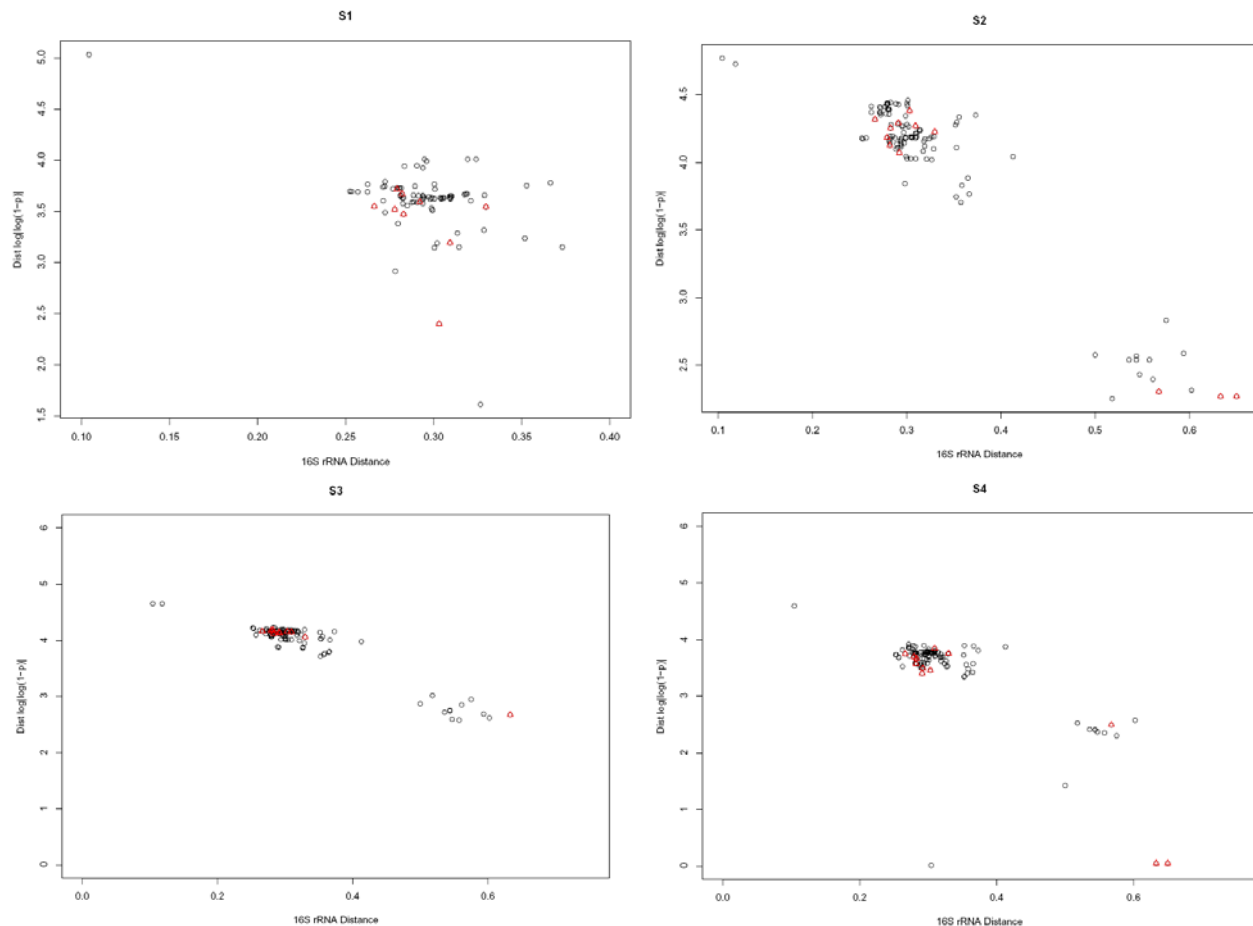


Figure 5. Ribosomal Proteins-*16S rRNA* Correlation Plots. The ribosomal proteins tested are proteins S1 ~ S4, obtained from TIGR genome library. The samples marked in red are those from the nitrogen fixation group in Figures 1 and 2.

From the correlation plots above, we no longer observe a distinct banding among the *nif* samples, but instead only one group. All the marked samples are intermingled with the rest of genomes in the library, and there is no longer any separation with each of the four ribosomal proteins. Because of their highly conserved nature, all genomes yield relatively close results with each other. LGT typically does not occur in genes highly specific and critical such as these ribosomal proteins. The correlation plots measure the relative distance among each organism (by their *16S rRNA* distance), compared to the relative closeness of a particular gene (measured by E-values or probabilities, as shown here). In comparison with the previous two results (*nif* and *CTX-M*), we can observe differences in the fundamental distributions of genetic differences between a laterally transferred gene and a non-lateral transferred gene.

We have three controls and intend to compare the differences between each group in order to infer possibilities of LGT and possibly extend to other case studies. From the three case study plots of BLAST E-values against *16S* relative distance, we observe internal clustering within the library, whereas for *nif* and *CTX-M*, we see a distinctive separation of matches far from the rest of the data. This pattern implies some qualitative similarities between the *nif* genes and *CTX-M* that is not exhibited in ribosomal proteins.

Discussions and Conclusion

With the results of nitrogen fixation in mind, there is supporting evidence about whether a particular gene is laterally transferred, by qualitatively comparing its overall distribution with those of known cases. However, this claim needs to be further verified through similar procedures. We have observed that the nitrogen fixers in our study are separated into two clusters, and that they share similar environment but have distinctive biological lineages. From this observation, we wish to extend and quantify the similarities to further observe whether later gene transfer has occurred in each case. Generally most LGT experiments to-date utilize qualitative analysis and have no rigorous testing mechanisms, therefore new quantitative techniques need to be developed in order to theorize and apply this interesting phenomenon. Further analyses, genetically or statistically, should confirm that nitrogen fixation behaves closer to *CTX-M* than ribosomal proteins in terms of gene acquisition. The eventual goal of the project is to quantify the differences between gene sequences, and once a quantitative method is established then this project can be extended to other case studies to develop a method of predicting the likelihood of a laterally transferred gene.

Methods and Materials

Complete Genomes

The library of organisms is established from the TIGR CMR database. The complete sequences available on TIGR via ftp.tigr.org the database is separated into individual genomes. Additions such as *D. vulgaris Hildenborough* and *D. ethenogenes 195* are added at a later date for the purpose of comparison.

BLAST

BLAST scores are obtained by a comparison between a query (e.g., *Nostoc sp. PCC 7120*) and a database of sequences with default parameters. Initially the filter option is on by default, and later in the experiment the filter option is turned off. The results presented in all sections, however, will be results obtained with the filter option off. For proteins, blastp is used, while blastn is used for nucleic sequences.

Nitrogen Fixation (NIF)

The *nif* protein sequences are obtained from Genbank. Comparisons of BLAST results between *Nostoc sp. PCC 7120* and *M. mazei Goel1* (chosen randomly from the group of nitrogen-fixing organisms) through the similar groupings showed similar results based on different controls, then it was decided that *Nostoc sp. PCC 7120* remained the control organism throughout the *nif*

experiment. Additional *nif* results are obtained after the acquisition of *D. vulgaris Hildenborough* and *D. ethenogenes 195* in the database.

After BLAST is run via Blast, the E-value of the best sequential match for each organism is recorded and tabulated. For subsequent experiments, values are recorded for statistical analysis in R .

CTX-M

The *CTX-M* protein sequences are obtained from NCBI using the accession numbers given in the original reference. The *CTX-M* protein sequences are compared via BLAST 2.0 among the group of selected organisms via WU Blast using *Kluyvera cryocrescens (KLUC-1)* as a query.

16S rRNA sequences obtained for the *CTX-M* species (see below) are compared similarly through BLAST 2.0. The BLAST results of *CTX-M* among the group contain unusable results for E-values because of their very strong and close identity. In order to obtain a set of E-values for the purpose of comparison, the E-values are calculated based on the published computational algorithm by NCBI (available www.ncbi.nih.gov).

Following the steps outlined in *NCBI*, using results of *16S Kluyvera cryocrescens* consensus sequence against other organism as the basis for λ and K , the bit scores and E-values of *KLUC-1* matches are read in R, then a linear regression model is fitted to natural log of E-values against the bit scores such that the unknown coefficient can be obtained. Since

$$E = mn \cdot (2^{-s'}) \quad (1)$$

where E as the E-values and s' the adjusted bit scores of a particular sequence, taking logs on both sides of the equality yields

$$\log E = \log mn + (-s') \cdot \log 2 \quad (2)$$

By a linear regression model with two coefficients as:

$$\log E = \hat{\beta}_0 + \hat{\beta}_1 x_1 \quad (3)$$

Assuming all other values remain, the unknown coefficient product (mn) can be computed.

After the coefficient product mn is calculated, the BLAST results of *CTX-M* among the control group can be extrapolated by computing the log of the E-values. Previously, BLAST has produced unusable E-values due to a hardware limitation in floating point calculations. The problem has been overcome by using the log of E-values, since most of the analyses are conducted with the log of E-values in the experiment, so comparisons can still be made.

Ribosomal Proteins

Another control for the BLAST comparison is based on BLAST results from the ribosomal proteins S1, S2, S3 and S4 of *Nostoc sp. PCC 7120* against those proteins from other organisms.

Because ribosomal proteins are well-protected within a cell, and vital for the life of organisms, lateral gene transfer is less likely to occur among the ribosomal proteins. The proteins are obtained from TIGR through a batch download for each organism within our library. From the batch download, some genome apparently contains two copies of each ribosomal protein, dubbed either sequence A or B. Only one sequence is taken for each genome by selecting the longer sequence or the A strand if both sequences are identical.

After the sequences are obtained, a comparison against ribosomal protein sequences of *Nostoc sp. PCC 7120* using BLAST 2.0 is conducted, and the results are recorded alongside with all other BLAST results. The results from ribosomal proteins will establish a basis of comparison as the negative control due to the nature of those genes. The sequences are extremely conserved such that lateral gene transfer is unlikely to transfer the entire sequences from one organism to another.

16S rRNA

The next step of the analysis is to establish relative distances between species. To do so, the *16S rRNA* of all organism in the library are extracted from *Ribosomal Database Project 9 (RDP 9*, available rdp.cme.msu.edu) and the sequences are compared to *Nostoc sp. PCC 7120* via BLAST 2.0 under the usual parameters. The results of BLAST are stored and used for later comparisons.

For the organisms with complete genomes, we were able to obtain each of their *16S rRNA* sequences. But, for the *CTX-M* data, RDP did not contain the sequence for the particular strains in the data set. Therefore, *16S rRNA* sequences for five to six pathogenic strands of each species in the *CTX-M* data set were obtained and a consensus sequence was constructed. For each of the alignments, the beginning extraneous segments are truncated near position 35, such that all alignments begin at approximately the same point. A similar truncated is also performed near position 1495 for all sequences, where all flanking regions of sequences are truncated after a previously selected marker. After truncations, sequences are screened such that gaps are removed, and the most frequent base is chosen for the consensus sequence. If only one base occupies a particular position in one organism and gaps appear for all other organisms, then that single base is chosen. In the event of a tie, the base from organisms that have more dominant bases before the position is chosen.

References

- Andersson, J., Doolittle, W. Ford, et al. *Are There Bugs in Our Genome? Science*, June 8, 2001 v292 i5523 p1848
- Blankenship, R.E., Raymond, J., Siefert, Janet L., Staples, C.R. *The Natural History of Nitrogen Fixation. Mol. Biol. Evol.* 21(3):541-554. 2004
- Bonnet, R. *Growing Group of Extended-Spectrum β -Lactamases: the CTX-M Enzymes. Antimicrobial Agents and Chemotherapy*, Jan. 2004. p. 1-14 Vol. 48, No.1
- Cannone, J.J., Gutell, R.R., Lee, J.C. *The Accuracy of Ribosomal RNA Comparative Structure Models. Current Opinion in Structural Biology* (2002) 12:301-310

Doolittle, W. Ford. *Phylogenetic Classification and the Universal Tree*. Science, June 25, 1999
v284 p2124-2128

Doolittle, W. Ford, Gogarten, J. Peter, Lawrence, J. *Prokaryotic Evolution in Light of Gene Transfer*. *Molecular. Biology and Evolution*. 19(12):2226–2238. 2002

Fox G.E., Willson, R.C., Zhang, Z. *Identification of Characteristic Oligonucleotides in the Bacterial 16S Ribosomal RNA Sequence Dataset*. *Bioinformatics* 18:244-250 (2002).

Postgate, John. *Nitrogen Fixation*, 3rd ed. Cambridge University Press, New York, NY, 1998.