



142  
689  
THS



2010



This is to certify that the  
thesis entitled

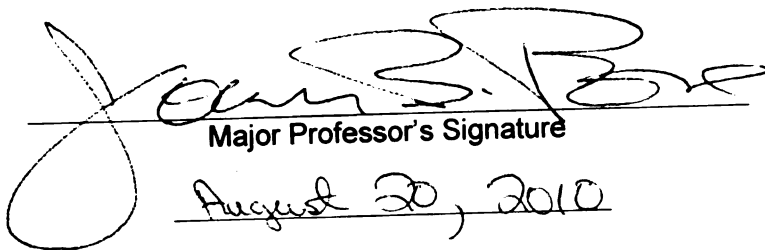
TWILIGHT ZONE:  
PROTEIN SEQUENCE SEARCH AND CLASSIFICATION

presented by

Jiarong Guo

has been accepted towards fulfillment  
of the requirements for the

M.S. degree in Fisheries and Wildlife

  
Major Professor's Signature

August 30, 2010

Date

*MSU is an Affirmative Action/Equal Opportunity Employer*

**PLACE IN RETURN BOX** to remove this checkout from your record.  
**TO AVOID FINES** return on or before date due.  
**MAY BE RECALLED** with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

**TWILIGHT ZONE:  
PROTEIN SEQUENCE SEARCH AND CLASSIFICATION**

**By**

**Jiarong Guo**

**A THESIS**

**Submitted to  
Michigan State University  
in partial fulfillment of requirements  
for degree of**

**MASTER OF SCIENCE**

**Fisheries and Wildlife**

**2010**



# ABSTRACT

## TWILIGHT ZONE: PROTEIN SEQUENCE SEARCH AND CLASSIFICATION

By

Jiarong Guo

Homology search is important for gene annotation and classification. The emergence of Next Generation sequencing techniques and metagenomics give homology search new challenges. In this study, I evaluate the commonly used homology search tools: BLAST and HMMER to find new genes (*nirK*) in our metagenomic data and classify closely related genes (*amoA* and *pmoA* in UniProtKB). BLAST false positive problem is found when comparing BLAST and HMMER searching results in metagenomic data.

Furthermore, I also describe methods which use phylogenetic trees to refine the results of common homology search methods. We evaluated these methods with the above genes, and find narrow phylogenetic sampling of genes (*nirK*) will limit sequence annotation and some genes (*amoA* and *pmoA*) are probably misclassified.

*Keywords:* homology search; BLAST false positive; phylogenetic tree; sequence misclassification

## **ACKNOWLEDGEMENT**

Through the past two years, I have got support from many people. Without their help, none of this work would have happened.

First, I want to thank my advisor C. Titus Brown. As a young professor, he is exceptionally good at advising students. His intellectual insights and scientific training helped me greatly throughout my research.

Second, I want to thank Adina Howe. She helped me a lot revising my thesis and also gave me great support as a friend.

My committee members, Joan B. Rose and Weiming Li, gave me many advices in the committee meeting. I am very grateful to them.

Everyone in C. Titus Brown's Lab helped me revising my thesis or presentation. Thank you all.

I also want to thank Jiaguo Qi and R. Jan Stevenson, two professors in ZHEJIANG UNIVERSITY and MSU program. They helped me a lot during my first year in MSU.

At last but by no means least, I want to thank my parents. They are the two always supporting me.

## TABLE OF CONTENT

LIST OF TABLES.....	vi
LIST OF FIGURES .....	vii
Chapter 1 .....	1
Background.....	1
1.2 Pairwise comparison: BLAST .....	2
1.3 Profile method: PSI-BLAST, HMMER, and Pfam database .....	5
1.4 Objective of research .....	8
Chapter 2.....	9
Comparing results of BLAST and HMMER nitrite reductase genes ( <i>nirK</i> ) search .....	9
2.1 Introduction: .....	9
2.2 Methods: .....	10
2.2.1 Samples/Metagenomic Datasets: .....	10
2.2.2 Data Analysis: BLAST and HMMER comparison .....	11
2.3 Results and conclusion: .....	14
Chapter 3.....	20
Using phylogenetic structure for refining genes identifications.....	20
3.1 Introduction: .....	20
3.2 Methods: .....	26
Using the tree squeezing method to find more <i>nirK</i> genes: .....	26
3.3 Results and Conclusion:.....	28
Chapter 4.....	31
Classification of <i>amoA</i> and <i>pmoA</i> .....	31
4.1 Introduction: .....	31
4.2 Data: .....	32
4.3 Methods: .....	33
4.3.1 BLAST and HMMER comparison (Figure 19) .....	33
4.3.2 Building the tree .....	34
4.4 Results: .....	35
4.4.1 The use of BLAST and HMMER to differentiate between <i>amoAs</i> and <i>pmoAs</i> in UniprotKB .....	35
4.4.2 Using the tree clustering method to differentiate <i>amoAs</i> and <i>pmoAs</i> .....	39
4.4.2 Using low nodes and high nodes for tree clustering .....	42
4.4.3 Investigating sequences identified by tree clustering that disagreed with UniProt Annotations.....	46
4.5 Conclusion:.....	50
Chapter 5.....	52
Conclusions and future work .....	52
5.1 Conclusions: .....	52
5.1.1 Biology is increasingly reliant on sequence similarity with the use of new sequencing technology, and sequence similarity may not give accurate gene annotations. ....	52



5.1.2 Annotations are sometimes incorrect. ....	52
5.1.3 Narrow phylogenetic sampling is a problem. ....	53
5.2 Novel approaches: .....	53
5.3 Future directions: .....	53
5.4 Parting Thoughts:.....	54
<b>Bibliography .....</b>	<b>55</b>

## **LIST OF TABLES**

Table 1. BLAST FP (False Positive Rate) at different E-value cutoffs.....	17
Table 2. Numbers of sequences classified by various methods.....	45

## LIST OF FIGURES

Figure 1. Simple examples of global and local alignment.....	3
Figure 2. The BLAST algorithm.....	4
Figure 3. Simple illustration of profile search. ....	6
Figure 4. Nitrogen cycling .....	10
Figure 5. Eight datasets sequenced by roche 454 pyrosequencing .....	11
Figure 6. BLAST and HMMER comparison processes in metagenomic dataset.....	12
Figure 7. BLAST hits and domain mapping .....	13
Figure 8. BLAST and HMMER search comparison.....	14
Figure 9. Distribution of identities in BLAST search result of six well known <i>nirKs</i> against dataset1 and dataset2. ....	15
Figure 10. All BLAST hits matched back to a query <i>nirK</i> . ....	16
Figure 11. E-value distribution of BLAST hits matched to domain region (blue) and a subset that were also found by HMMER domain search (green). ....	18
Figure 12. Simple illustration of neighbor-joining process. ....	22
Figure 13. A hypothetical sub-tree of tree of whole protein space with cluster F1, A and F2 in a specific arrangement. ....	24
Figure 14. A hypothetical sub-tree of tree of whole protein space with cluster F1, F2 and A in a specific arrangement. ....	25
Figure 15. E-value distribution of 3055 <i>nirKs</i> from FunGene (blue) and 32 sequences from tree squeezing method (green) .....	27
Figure 16. Definition of tree squeezing method. ....	28
Figure 17. A hypothesis explaining why only small number of sequences is squeezed. .	29
Figure 18. Phylogenetic tree of all 3055 FunGene <i>nirK</i> sequences.....	30
Figure 19. BLAST and HMMER comparison flow chart.....	34
Figure 20. Flow chart of constructing phylogenetic tree of <i>amoAs</i> and <i>pmoAs</i> . ....	35
Figure 21. E-value distribution of <i>amoA</i> BLAST hits. ....	36
Figure 22. E-value distribution of <i>amoA</i> HMMER hits.....	37
Figure 23. E-value distribution of <i>pmoA</i> BLAST hits. ....	38



Figure 24. E-value distribution of <i>pmoA</i> HMMER hits.....	39
Figure 25. Tree of known <i>amoAs</i> and <i>pmoAs</i> . .....	41
Figure 26. Trees of all <i>amoAs</i> and <i>pmoAs</i> . .....	42
Figure 27 Simple example trees for super node analysis.....	44
Figure 28. Alignments of two correctly classified <i>amoAs</i> , two wrongly classified <i>pmoAs</i> found by low node method, two well known <i>amoAs</i> and two well known <i>pmoAs</i> .....	47
Figure 29. Closeups of two possibly misclassified <i>pmoAs</i> in tree of all <i>amoA</i> and <i>pmoA</i> (A1 and B1) and two known <i>amoAs</i> (arrow) that help classify the two possibly misclassified <i>pmoAs</i> in tree of all known <i>amoAs</i> and <i>pmoAs</i> . .....	49

**NOTE: Images in the thesis are presented in color.**

# **Chapter 1**

## **Background**

### **1.1 Importance of homology search**

Homologous proteins are proteins that are derived from a common ancestor. They often have similar sequences, structures, and functions. Because homology is often inferred from sequence similarity, accurate sequence annotations based on homology search is an important area in biology. Most protein sequences are annotated by sequence similarity, and only a small number of sequences are biologically verified by molecular experiments. UniProtKB is a large database of protein sequences. Within the UniProtKB database, the biologically verified sequences and those sequences that can be annotated significantly by homology searches are stored in the SwissProt database, which is maintained manually. The other sequences, which are annotated by automatic sequence homology searches, are stored in the TrEMBL database (Wu, Apweiler et al. 2006). The size of the SwissProt database (518,415 protein sequences) is much smaller than the TrEMBL database (11,397,958 protein sequences), indicating that the majority of current proteins in the UniProtKB database are annotated by automatic homology searches. Problems with homology searching algorithms may result in incorrect protein annotations throughout the database.

The emergence of metagenomics presents a new challenge to using homology searches to annotate and/or classify sequences. Metagenomics is the study of DNA recovered directly from environmental samples, and it investigates the diversity and functional pathways represented by microbial communities (Riesenfeld, Schloss et al. 2004). Given the fact

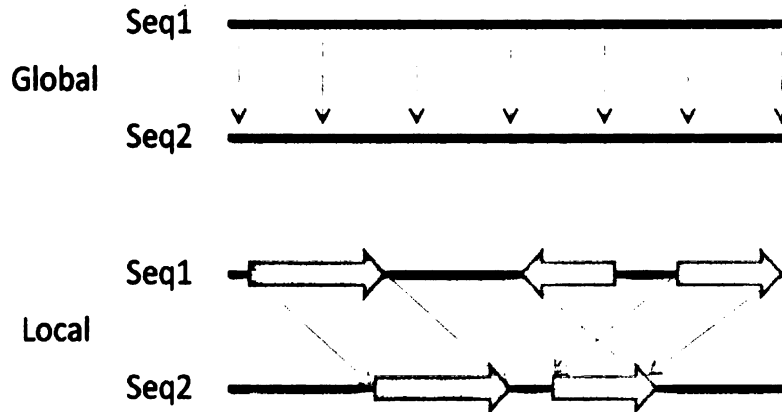
that most microbes cannot be cultured in standard lab conditions, metagenomics is an alternative approach to studying the microbial world. With the introduction of next generation sequencing (NGS) techniques, metagenomic data have rapidly accumulated in the last five years. The resulting large number of sequences and their characteristic short lengths present challenges to sequence annotation.

The most popular homology searching methods use pairwise comparisons or profile models (also known as position specific scoring matrices).

## **1.2 Pairwise comparison: BLAST**

Pairwise comparison is used to find the best matching local or global alignment of two sequences. Global alignment tries to align every position in both sequences and is more effective for similar sequences of roughly equal size. It also has the advantage of being able to align multiple sequences. Local alignment, on the other hand, compares all segments of all possible lengths and attempts to align the similar sequence regions together (Figure 1). These sequence regions could be sequence motifs, which are defined as amino-acid or nucleotide patterns that might have a biological significance. It is more useful to detect these similar sequence motifs within their larger dissimilar sequence context. The Needleman-Wunsch algorithm is a well-known method for global sequence alignment, while the Smith-Waterman algorithm is famous for performing local alignment. The Smith-Waterman algorithm is more sensitive at finding remote homologous sequences, but is not fast enough for large database homology search (Needleman and Wunsch 1970; Smith and Waterman 1981; Altschul, Gish et al. 1990).



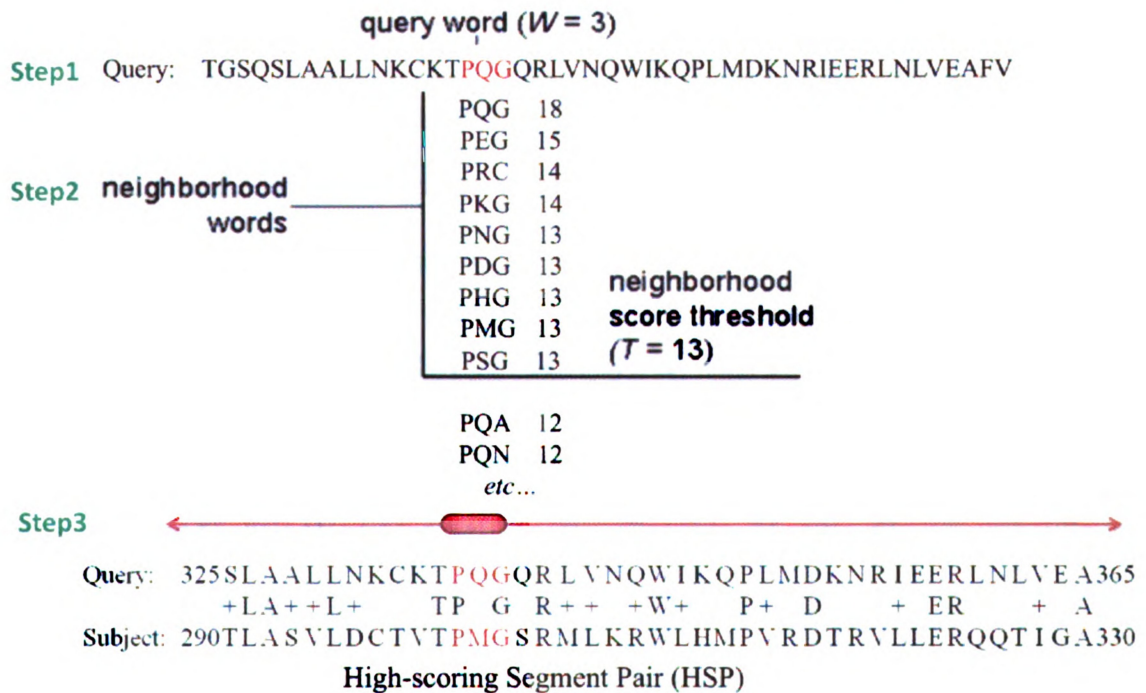


**Figure 1.** Simple examples of global and local alignment. Global alignment aligns sequences that have similar size and are generally similar over entire length. Local alignment is good at finding small similar regions (including inversions) within whole sequences that global alignment is not able to detect.

BLAST (Altschul, Madden et al. 1997) is the most popular bioinformatics tool for searching homology of DNA or protein sequence in big databases. It performs a local alignment of a query sequence against a sequence in database. BLAST, by using a heuristic method called the word method (Figure 2), is able to perform pairwise alignments with significantly improved speed and efficiency without losing much accuracy (Altschul, Gish et al. 1990). BLAST efficiently searches for matches between a query and a reference by finding the hot spots of high scoring words in database sequences. It avoids aligning the query sequence with a large portion of database sequences that have no significant matches. In the word method, non-overlapping  $k$ -letter subsequences (words) in the query sequences are listed. By default  $k$ ,  $k$  is 3 for protein and 11 for DNA. All  $k$ -letter words ( $20^k$  for protein and  $4^k$  for DNA in total) are aligned with words in the list and scored with a substitution matrix. The high scoring words

(neighborhood words) that score above a threshold  $T$  are collected. These neighborhood words are the ‘queries’ actually searched in the database. When an exact match of one word is found, the alignment is extended in both directions to find the ungapped alignment with the highest score, higher than the bit score threshold. These ungapped alignments, called HSP (high scoring pair), are reported by BLAST.

## The BLAST Search Algorithm



**Figure 2.** The BLAST algorithm. BLAST applies a heuristic search method which finds  $k$ -letter words (default = 3 in blastp) scoring at  $T$  (neighborhood score threshold) when aligned to a specific  $k$ -letter word in the query and scored with a substitution matrix. Words scoring above  $T$  (neighborhood words) are searched in database and then extended in both directions until the score starts to decrease. The resulting locally optimal scoring

alignment is called HSP (high scoring pair) and later reported by BLAST if it has a score higher than S (BLAST score cutoff) or E-value lower than a specified cutoff.

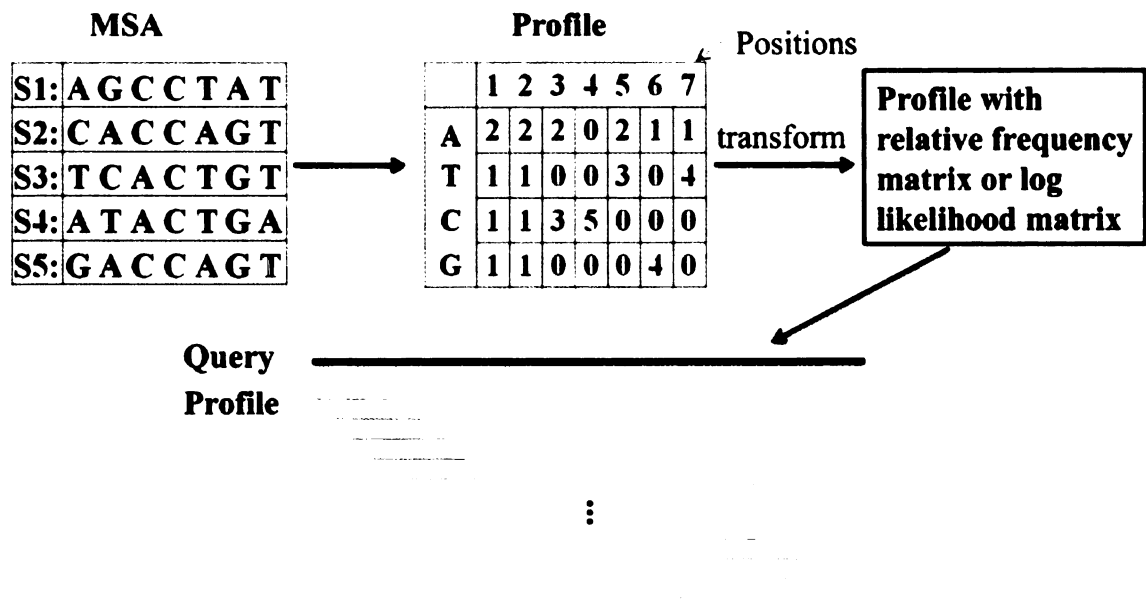
When pairwise sequence identity is high, putative homology and non-homology can be distinguished by pairwise comparison methods including BLAST. However, pairwise comparisons perform poorly on sequence alignments when the pairwise sequence identity goes down to 20-35% (for protein sequence), the twilight zone of protein sequence alignment (Rost 1999). The best scoring pairwise alignment becomes uncertain when aligning two remote sequences with low identity. Thus, pairwise comparison performs poorly at finding remote homologies in the twilight zone. It is not clear whether the difficulty of using alignment methods in this zone is caused by a technical problem (i.e., detection of statistical significance) or is a special feature of evolution (Rost 1999).

### **1.3 Profile method: PSI-BLAST, HMMER, and Pfam database**

The profile method, also called position specific scoring method, is another method popularly used for homology searches in sequence databases. A profile is a probability model, made from a multiple sequence alignment of several homologous sequences (usually from the same protein family). In the profile, a position-specific scoring system for insertions, deletions and substitutions is assigned, which shows that some columns of the multiple sequence alignment are more conserved for several amino acids (or nucleotides) while some are more open to gaps or indels (Figure 3). The profile represents conserved or non-conserved sequence information of the multiple sequence alignment. Compared to the Smith-Waterman algorithm, BLAST, or other traditional pairwise alignment methods that use position-independent scoring parameters (like



BLOSUM62), profile methods are more precise and more sensitive to detect remote homology (distantly related sequences). The use of profiles is more effective at capturing important domains or motifs that are conserved in all query sequences (Gribskov, McLachlan et al. 1987).



**Figure 3.** Simple illustration of profile search. Profile with absolute frequency at each position (column) is made from MSA (multiple sequence alignment). The profile is transformed into a new one with relative frequency matrix or log likelihood matrix. The new profile is aligned with every 7 base pair region of query sequence. The regions with scores above cutoff are reported.

Profile methods were introduced in the late 1980s. Similar methods like “flexible patterns” (Barton 1990), “templates” (Bashford, Chothia et al. 1987) were introduced at the same time. One such example is PSI-BLAST which implements the profile method in the BLAST2 package. This program combines pairwise local alignment and the profile method. The first round BLAST results are integrated into a profile sequence, then the

profile sequence is used as query in a second round search, and new hits are added to the profile for the next round search. PSI-BLAST is much more sensitive than standard protein-protein BLAST in picking up evolutionarily distant homology (Altschul, Madden et al. 1997). Recently, PSI-BLAST has also been reported to have a homologous over-extension problem (where it picks up many non-homologies) because the noise introduced by one non-homology can be amplified through iteration (Gonzalez and Pearson 2010).

In this study, I choose HMMER (Eddy 1998) as a representative tool implementing the profile method. The basic advantage of HMMER is that it uses a formal probabilistic basis called the “hidden Markov model” to guide how all the parameters in a position-specific scoring system should be set. The HMMER profiles constructed are called “Profile Hidden Markov Models” (profile HMMs), which are statistical models of multiple sequence alignment or even single sequences.

HMMER has a consistent theory for setting parameters in profile HMMs so it is applicable to a large database of profile HMMs and large scale sequence analysis. The Pfam database is a large collection of protein domains represented by profile HMMs. It is an important part of the Interpro annotation system and is one of the most popular databases for protein sequence annotation and analysis.

In the past, profile HMM methods were computationally expensive. HMMER2 is about 100x slower than comparable BLAST searches. A new version of HMMER, HMMER3, combines the power of using a probabilistic model with high computational speed, is now essentially the same speed as BLAST (Eddy 2009).

HMMER also has the function (hmmalign) to align multiple sequences to a profile HMM. The hmmalign gives highly reproducible and high quality alignments when profile seeds are well selected because all the sequences are aligned to the profile HMM. It is also faster than pair-wise alignment methods like MUSCLE and CLUSTALW because sequences are only aligned once each to the profile HMM (Wu and Eisen 2008). Thus, this is a good alignment option for phylogenetic analysis of sequences from the same family.

The shortcoming of the profile method is that the choice of seed sequence for profile is critical (garbage in garbage out) and the method is affected by the quality of underlying MSA (Madera and Gough 2002).

## **1.4 Objective of research**

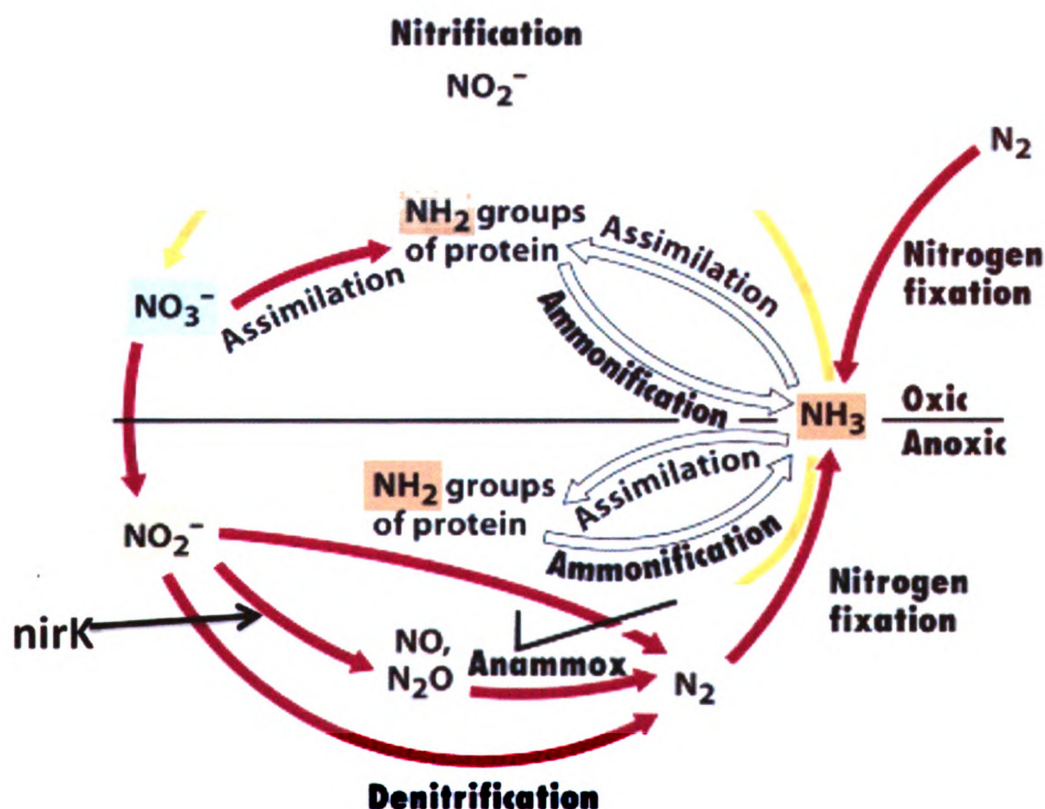
In this study, I evaluate the above described protein annotation methods (pairwise comparison implemented by BLAST and profile method implemented by HMMER) to find and classify proteins in metagenomic datasets. Our evaluations included: (1) comparing the usage of BLAST and HMMER to find new environmentally-relevant genes (*nirK*) in a soil metagenomic data and (2) the ability of BLAST and HMMER packages to classify closely related genes (*pmoA* and *amoA*) in the UniProt protein database. Furthermore, I also describe a method which uses phylogenetic trees to identify sequence homology. I evaluated this method with the above genes and compared our results to other protein annotation methods (HMMER).

## Chapter 2

### Comparing results of BLAST and HMMER nitrite reductase genes (*nirK*) search

#### 2.1 Introduction:

Nitrite reductase genes are a crucial part of global nitrogen cycle. *NirK* is such a gene and is involved in denitrification. Denitrification is a microbe-facilitated process that uses oxidized nitrogen as an alternative electron acceptor to produce energy in environments where oxygen is limited. The end product of denitrification is molecular nitrogen (N<sub>2</sub>), which may be released back to the atmosphere. Denitrification is also useful in wastewater treatment and bioremediation, though it also causes the emission of a greenhouse gas which may damage the ozone layer and/or cause nutrient loss in agriculture (Tiedje 1988). Nitrite reductase catalyzes the reduction of nitrite to nitric oxide (NO). *NirK* is a copper based metalloprotein.

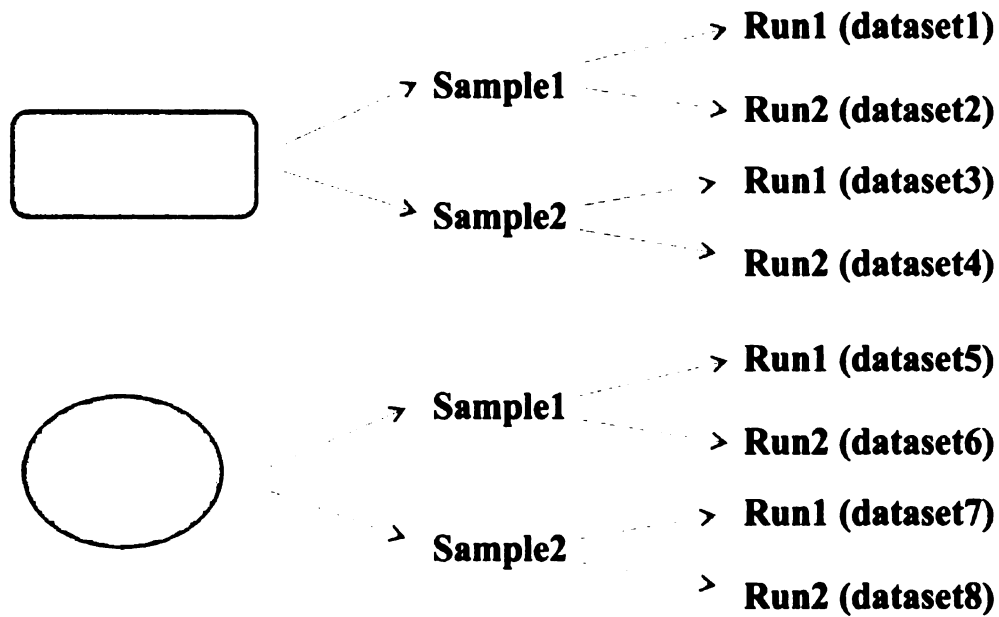


**Figure 4.** Nitrogen cycling (Madigan, Martinko et al. 2006).

## 2.2 Methods:

### 2.2.1 Samples/Metagenomic Datasets:

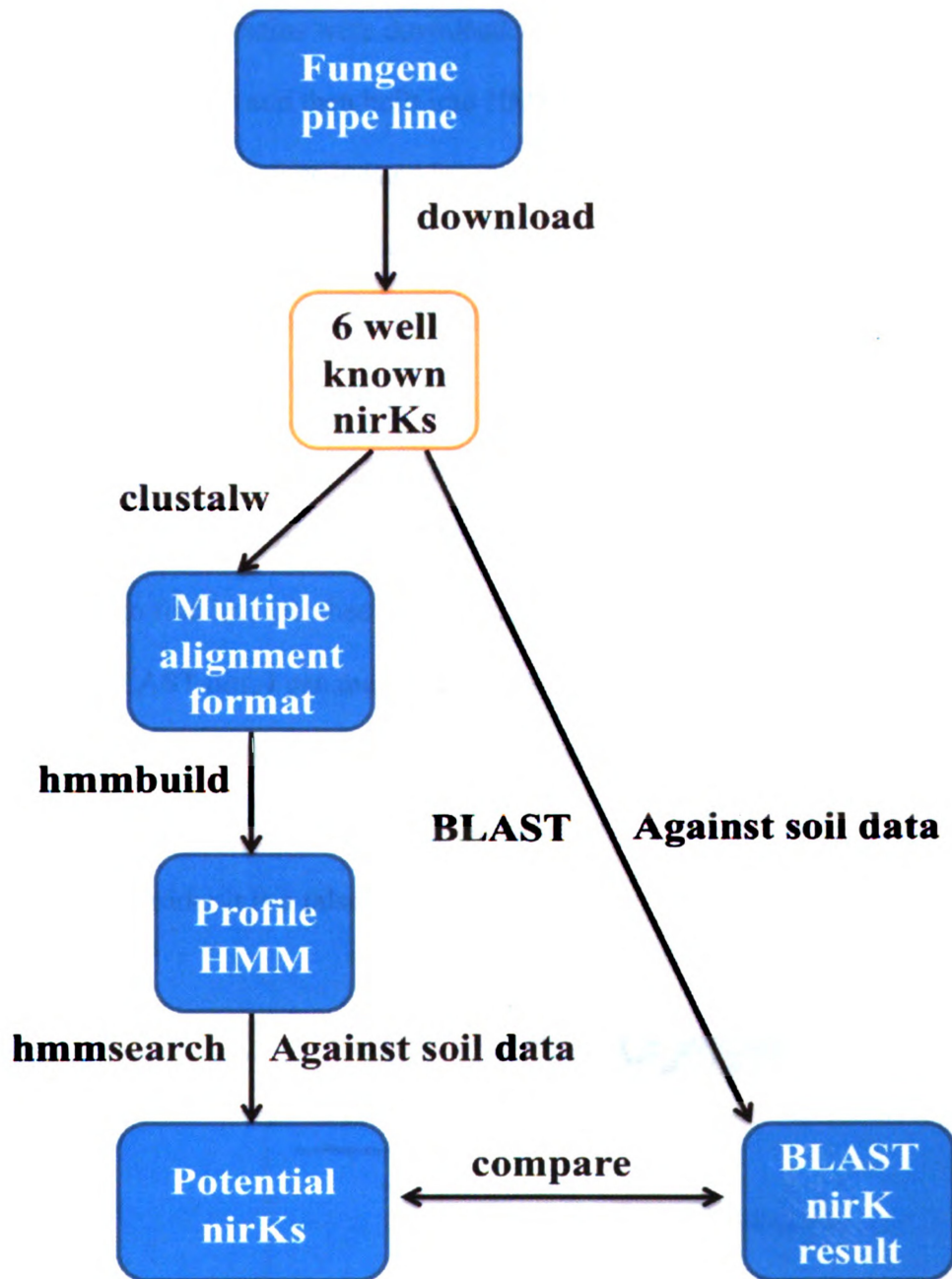
For our studies, I used 454 (Metzker 2010) sequenced metagenomic data of various soil samples. A total of four soil samples were studied: two agricultural soil samples (biological replicates) and two forest soil samples (biological replicates) from Kalamazoo, MI (KBS LTER – Kellogg Biological Station Long Term Ecological Research). For each sample, DNA was extracted directly from the soil sample and sequenced with two 454 pyrosequencing runs (Figure 5). The average resulting sequence read length was about 250 base pairs, much shorter than read lengths generated by traditional Sanger sequencing. I evaluated the ability to identify *nirK* genes using BLAST and HMMER.



**Figure 5.** Eight datasets sequenced by roche 454 pyrosequencing.

### 2.2.2 Data Analysis: BLAST and HMMER comparison

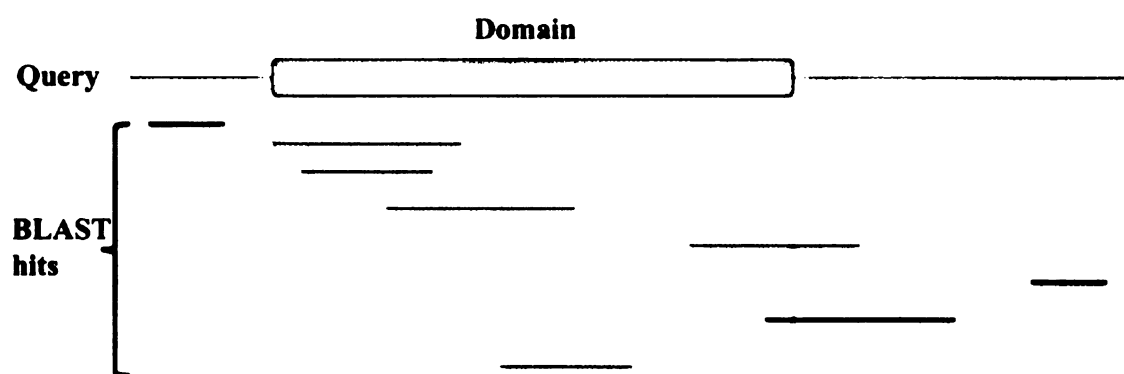
The DNA reads in the metagenomic data were translated into amino acid reads. Six well known *nirK* sequences from the FunGene (functional gene pipeline & repository) website were used to blast against the various soil metagenomic datasets using default parameters (E-value cutoff was 10). Additionally, the six *nirKs* were also used to generate a profile HMM using HMMER3 and searched against our metagenomic data using default parameters (E-value cutoff was 10) (Figure 6). I chose to use BLAST and HMMER default parameters with a relatively relaxed E-value cutoff to include all possible real *nirK* sequences. Subsequent filtering steps could be used to filter out more false positive hits (see Chapter3)



**Figure 6.** BLAST and HMMER comparison processes in metagenomic dataset.

To investigate the difference between BLAST and HMMER search results, I evaluated which part of the query *nirK* sequence, domain region or non-domain region, blast hits were matched to. Functional domains of *nirK* (cu-oxidase, cu-oxidase\_2, and cu-oxidase\_3) are shown on the FunGene website, and the multiple sequence alignment

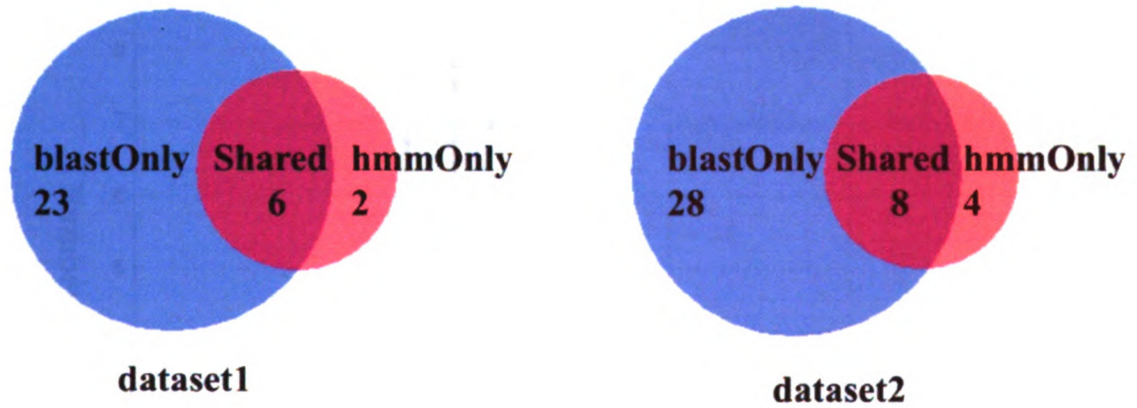
(MSA) of these domains were downloaded from the Pfam database (stockholm format for use with HMMER) and then built into HMMs of the domains. Domain regions of the query seed sequence were located by searching the domain HMMs against the query seed sequences at an E-value cutoff of 0.01. Then, BLAST hits were matched back to the query seeds based on the start and end position of matches that are shown in the BLAST output. If a BLAST hit had more than 20 amino acids matched to domain region of query, I say it is a match to the domain region. If a query was not matched to the domain region, I removed it from the BLAST and HMMER comparison (Figure 7). HMMER is supposed to find hits matched to the conserved region, and thus, by removing non-domain BLAST hits, I can make the results from BLAST and HMMER comparable. I assume that if both BLAST and HMMER identify a sequence as *nirK*, the annotation is likely to be correct. If a sequence is matched by BLAST but not by HMMER domain search, I consider it is a false positive.



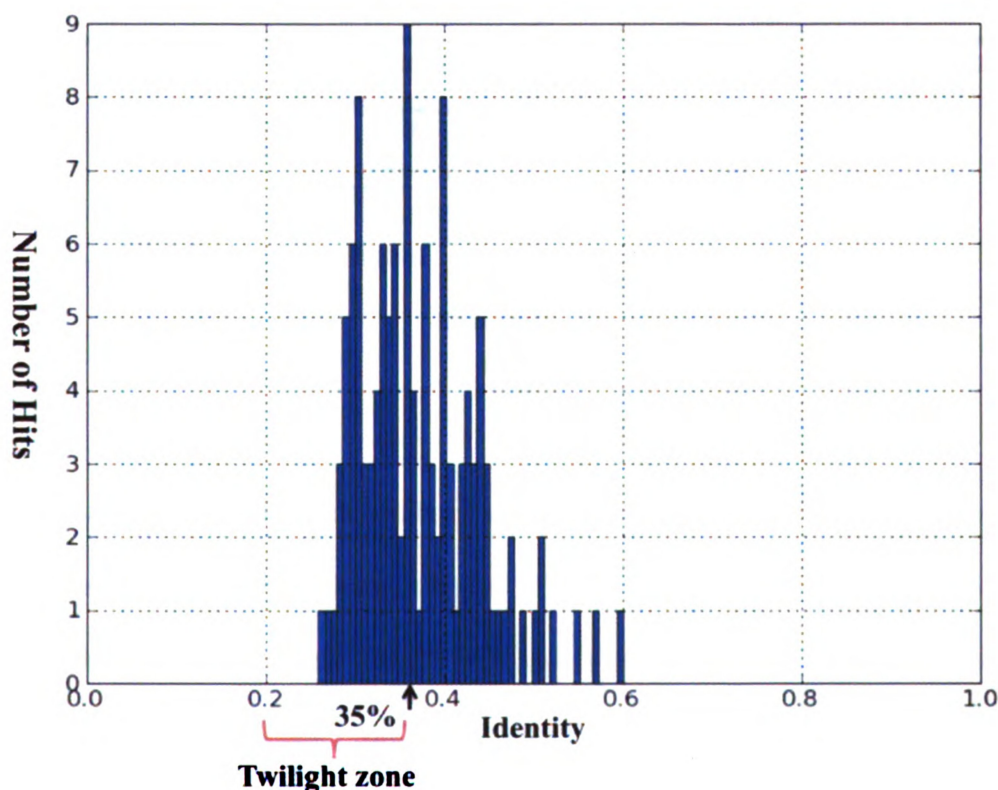
**Figure 7.** BLAST hits and domain mapping. HMMER is good at finding conserved regions (domain), so I are interested in BLAST hits matched to domain region (red lines) in order to compare results from two tools.



### 2.3 Results and conclusion:

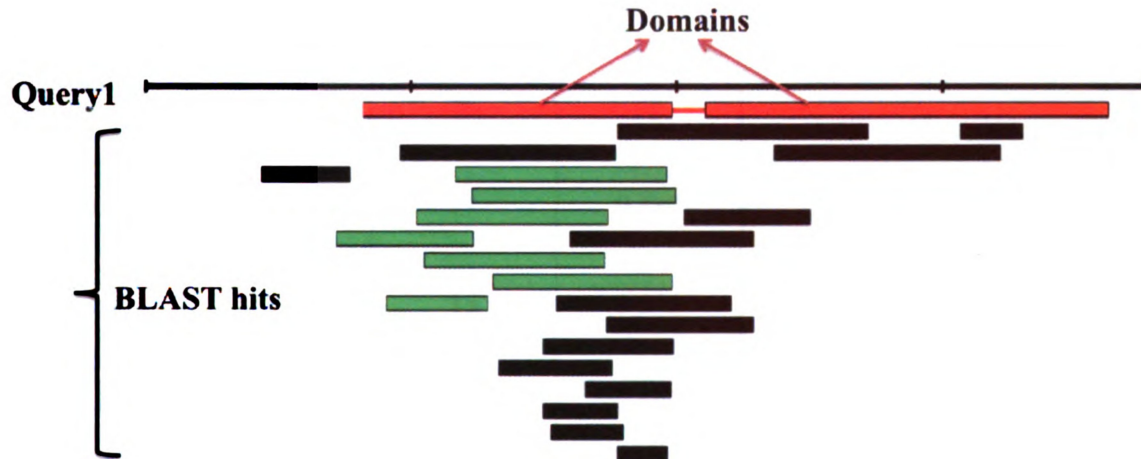


**Figure 8.** BLAST and HMMER search comparison. “BlastOnly” are the hits obtained only by BLAST; “hmmOnly” are the hits obtained by HMMER only; “shared” are overlap of BLAST and HMMER hits. Search results from both datasets show BLAST gets more hits than HMMER search.



**Figure 9.** Distribution of identities in BLAST search result of six well known *nirKs* against dataset1 and dataset2.

In BLAST and HMMER search comparison (Figure 9), BLAST has 29 hits while HMMER has 8 hits from dataset, and BLAST has 36 hits while HMMER has 12 hits from dataset2. I can see BLAST gets more hits than HMMER, and many BLAST hits are in the twilight zone (20%-35% identity) (Figure 9), suggesting that there may be false positive hits in the BLAST result. By matching the BLAST hits back to the query sequence, I were able to see whether the hits are matched to a domain (functional) region or non-domain (non-functional) region (Figure 10). The total number of false positive hits using this method was then calculated (Table 1).



**Figure 10.** All BLAST hits matched back to a query *nirK*. Query1 is a well known *nirK* seed. The red bands show the domain region of query1. BLAST hits are shown as black (not found by HMMER domain search) and green (also found by HMMER domain search).

For *nirK* searches in our metagenomic data, the FP rate does not change much when the E-value cutoff gets lower (more stringent) (Table 1, Figure 11).

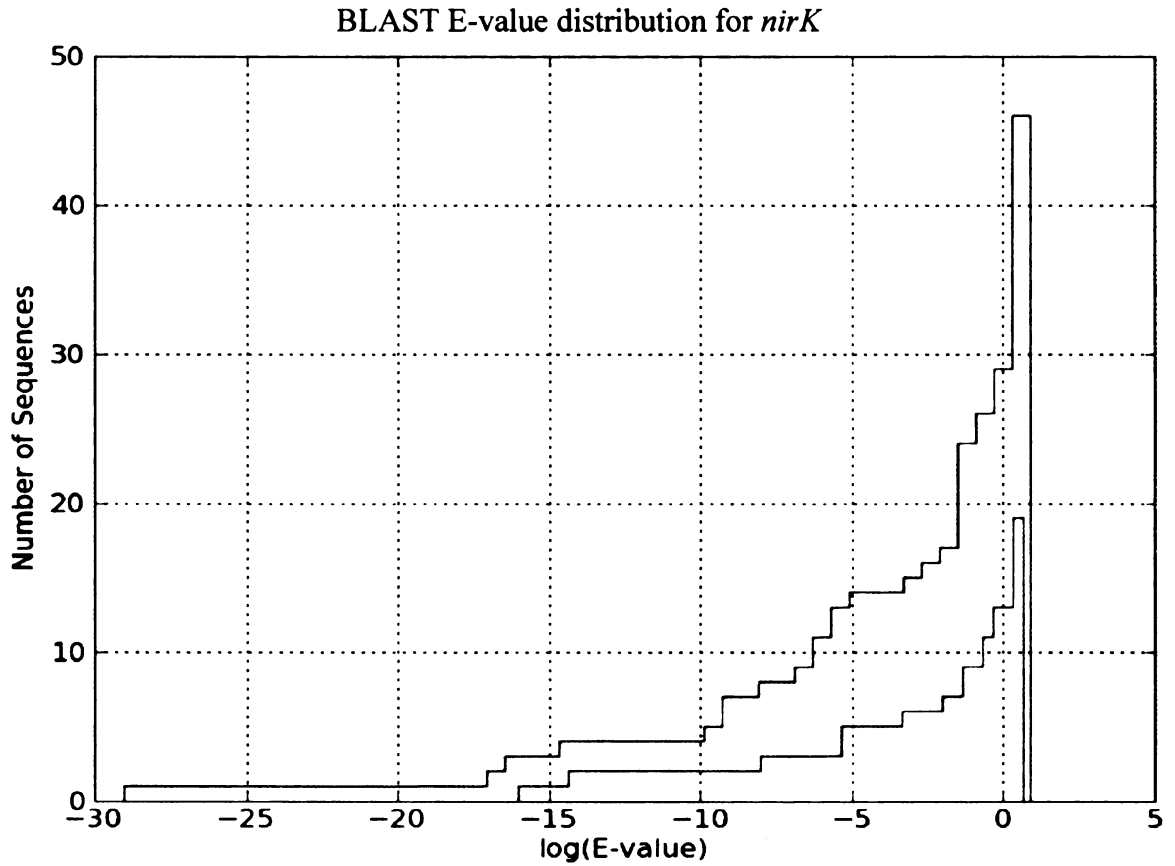
**Table 1.** BLAST FP (False Positive Rate) at different E-value cutoffs.

Cutoff	A	B	FP
10	19	27	0.587
1	13	15	0.536
0.1	9	14	0.609
0.01	6	10	0.625
0.001	6	9	0.600
0.0001	5	9	0.643

A is BLAST hits matched to domain region and also found by HMMER domain search;

B is BLAST hits matched to domain region but not found by HMMER domain search. FP

=  $B / (A+B)$ .



**Figure 11.** E-value distribution of BLAST hits matched to domain region (blue, A+B in Table 1) and a subset that were also found by HMMER domain search (green, A in Table 1). BLAST hits matched to domain region (blue) has similar E-value distribution shape as its subset that are also found by HMMER domain search, which indicates the BLAST false positive rate (FP) does not change much when E-value cutoff is changed.

For this part of study, I found BLAST might give false positives by comparing BLAST and HMMER *nirK* search result in metagenomic data. Many BLAST hits had their alignment identity in the twilight zone of sequence alignment (20%-35%). When I treated the overlap of BLAST hits and HMMER hits as true positive hits, the false positive rate did not change much when E-value cutoff was changed. This indicated that two different

tools, BLAST and HMMER, might share some same characteristics, which needs further study.

It is important to note that the significance of hits, E-value, is not considered here, and this should be studied more in the future. For example, if I have a BLAST hit matched to the query's domain region but the E-value is very high ( $>10$ ), the hit is probably a false positive. However, if a low (stringent) cutoff were chosen to reduce false positives, remote homologs may be overlooked. When searching for novel genes or remote homologs, picking a high E-value cutoff and then using other methods to filter for better ones may be a good strategy.

## **Chapter 3**

### **Using phylogenetic structure for refining genes identifications**

#### **3.1 Introduction:**

Pairwise comparison and profile methods are commonly used for homology evaluation, but they have their weaknesses, i.e., BLAST has poor performance for sequences within the twilight zone (20-35% sequence identity) (Rost 1999) and profile methods rely heavily on seed sequences for making a statistical profile (Loewenstein and Linial 2008). Moreover, homology searching annotation methods may give inaccurate functional annotations for paralogous sequences (homology through gene duplication) or orthologous sequences (homology through speciation) which may have different functions (homolog with different functions). Comparative genomics studies have given some useful methods to address these challenges, such as gene neighboring methods, protein domain architectures, and tree clustering methods (Singh, Doerks et al. 2009). Here, I focus on the tree clustering method to refine gene identification methods for more accuracy.

Phylogenetic trees show evolutionary relationships among different species or molecular sequences based on their phenotype or sequence similarity. Here I focus on molecular sequence phylogenetic trees. There are two types of trees: rooted and unrooted. In a rooted tree, a unique node represents the most common ancestor of all species or molecular sequences at the leaves under this node. In an unrooted tree, the relatedness of branch nodes and leaves is illustrated without assumptions about ancestry. The root can

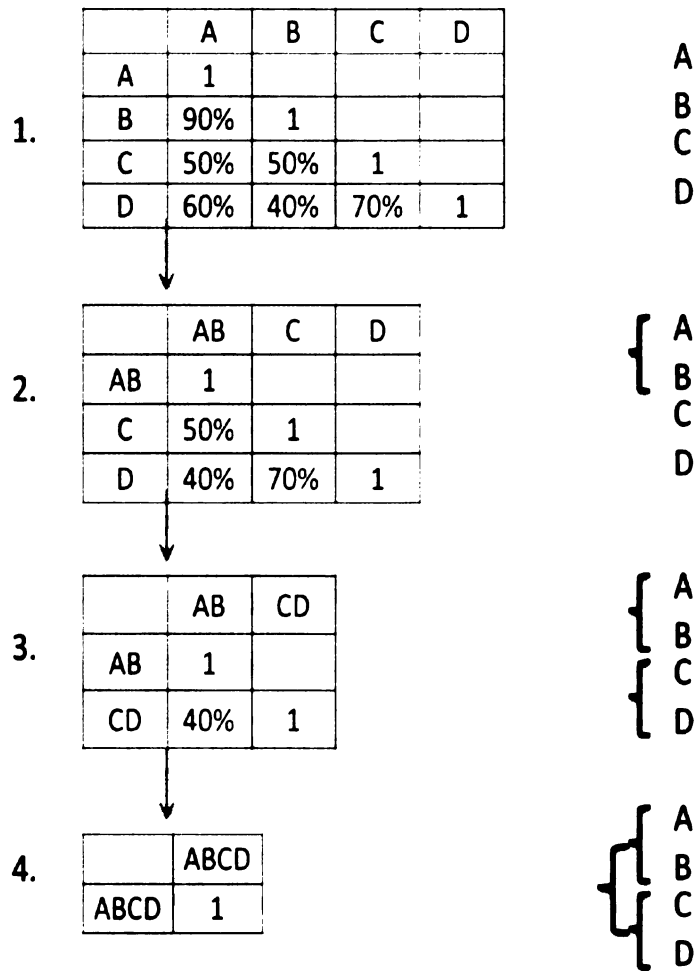
be set at any part of the unrooted tree if no other information is provided. Using an uncontroversial out-group is the most common approach to make a rooted tree. The out-group should be distantly related to the sequence of interest. If too distant, it will add noise to the phylogenetic analysis.

Three types of methods are commonly used to construct trees: distance matrix, maximum parsimony, and maximum likelihood.

The distance matrix method calculates all possible pairwise distances from a multiple sequence alignment. It includes neighbor-joining and UPGMA (Unweighted Pair Group Method with Arithmetic mean). The former produces unrooted trees and does not assume constant rate of evolution, while the latter produces rooted trees and does assume constant rate of evolution. The distance matrix methods are generally computationally fast and thus are good for large sequence data analysis (Felsenstein 2004; Mount 2004).

Neighbor-joining tree is the one used my thesis. Figure 12 shows the basic idea of neighbor-joining process.





**Figure 12.** Simple illustration of neighbor-joining process. A, B, C and D stand for 4 sequences. Their pair wise sequence similarities are shown in tables. In every step, two sequences with highest similarity are first joined.

The maximum likelihood method is a more advanced method, which applies an explicit model of evolution, such as Jukes-Cantor or generalized time-reversible (GTR) models of nucleotide evolution and the JTT (Jones-Taylor-Thornton) model of amino acid evolution, to tree estimation. It is statistically well founded, but computationally much more expensive than the distance matrix method, and so is not fast enough for phylogenetic analysis of a large number of sequences (Felsenstein 2004).

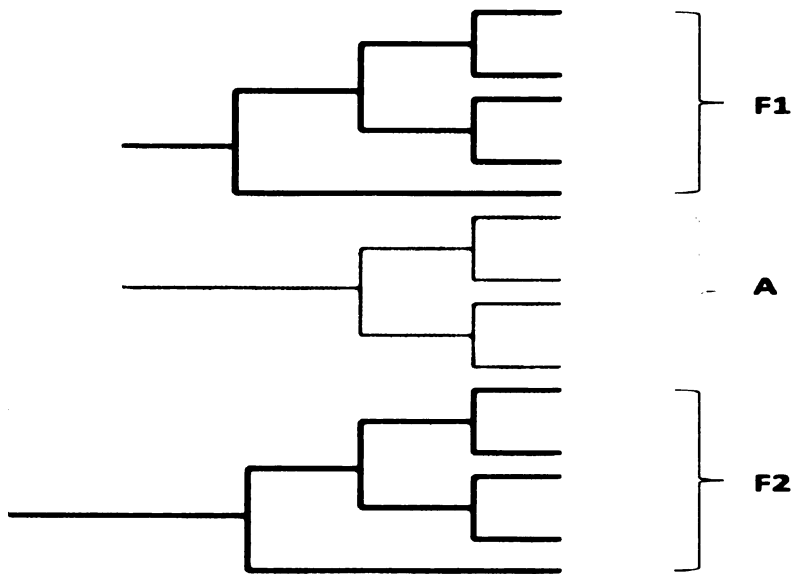
## **The Tree Clustering Method**

The tree clustering method is a systematic approach to find undetected relations among homologous sequence families or subfamilies based on sequence clusters in a phylogenetic tree (Loewenstein and Linial 2008; Singh, Doerks et al. 2009).

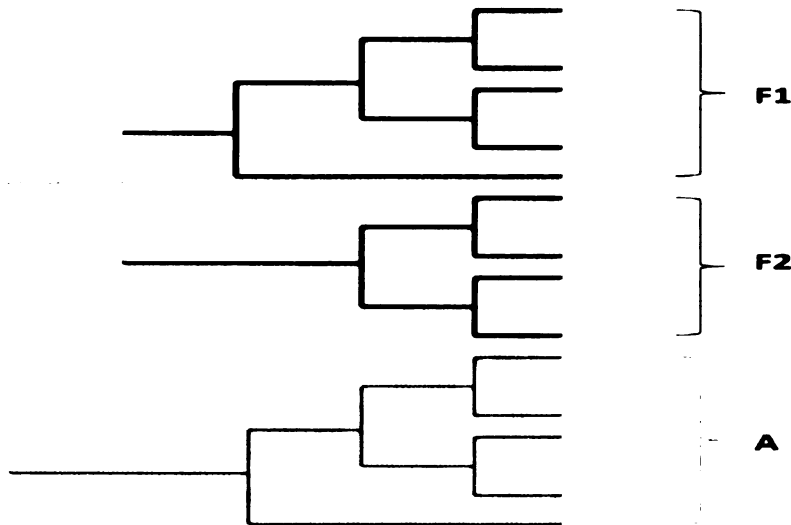
The ProtoNet database (Kaplan, Friedlich et al. 2004) is a successful application of the tree clustering method. It provides protein classification based on a phylogenetic tree that shows hierarchical organization of the whole protein sequence space and evolutionary relatedness among protein families. The tree is constructed by a modified UPGMA algorithm, which can handle all the UniProtKB sequences. High correspondence is shown between clusters in the tree and protein families classified in other databases, such as Pfam and SCOP. Further, overlooked and new functional connections between families and sequences can be discovered in the tree, because it provides a global view of the protein space hierarchy from close subfamily proteins to distantly related superfamily proteins (Loewenstein and Linial 2008).

There are several reasons why the tree clustering method reveals relations undetectable by profile methods. First, tree clustering is based on pairwise comparisons of all family members (in a distance matrix tree), while in the profile method, families are represented as a single statistical model (Gribskov, McLachlan et al. 1987). Similarities expressed by only some remote family members can be detected in tree but are overlooked by profile method if the remote members are not included in the profile seeds (Loewenstein and Linial 2008). Second, remote homologs picked up in the tree can help group family members. Third, the profile method relies too much on the quality of the underlying MSA

(multiple sequence alignment), which is decided by choice of seed sequences and MSA algorithm. The MSA tools also have the MSA uncertainty problem (different tools give different alignments) (Wong, Suchard et al. 2008) when aligned sequences are distantly related. Examples of hypothetical tree clustering methods are shown in Figures 13 and 14.



**Figure 13.** A hypothetical sub-tree of tree of whole protein space with cluster F1, A and F2 in a specific arrangement.



**Figure 14.** A hypothetical sub-tree of tree of whole protein space with cluster F1, F2 and A in a specific arrangement.

In Figure 13, hypothetical subfamilies F1 and F2 are known to belong to the protein family, F. Sequences in subfamily A are sequences from a metagenomic dataset with no significant matches to known sequences in a reference database. Because of its position in the tree between known subfamilies, I can infer that A is a subfamily of F. This example demonstrates the advantage of the tree clustering method (a global view of all sequence similarities). If using HMM profiles, and only some well known sequences are picked to make a profile HMM, the sequences in A may not be detected by that profile.

In Figure 14, F1, F2 are the same hypothetical subfamilies as described in Figure 13. The position of subfamily A is now outside the known families. I cannot infer confidently A is a new subfamily of F due to the relative position of A, F1 and F2 in the tree though there is still a possibility.

Although using tree clustering can identify functions that are undetectable by other methods, this method also has its shortcomings. Like profile comparison methods, the tree is constructed based on MSA and thus errors or uncertainties used in MSA construction may result in inaccurate tree building.

In this chapter, “tree squeezing method”, an application of the general tree clustering idea, was used to refine the possible 3055 *nirK*s down from FunGene website. I found the phylogenetically narrow known *nirK*s limited the application of tree squeezing method.

### **3.2 Methods:**

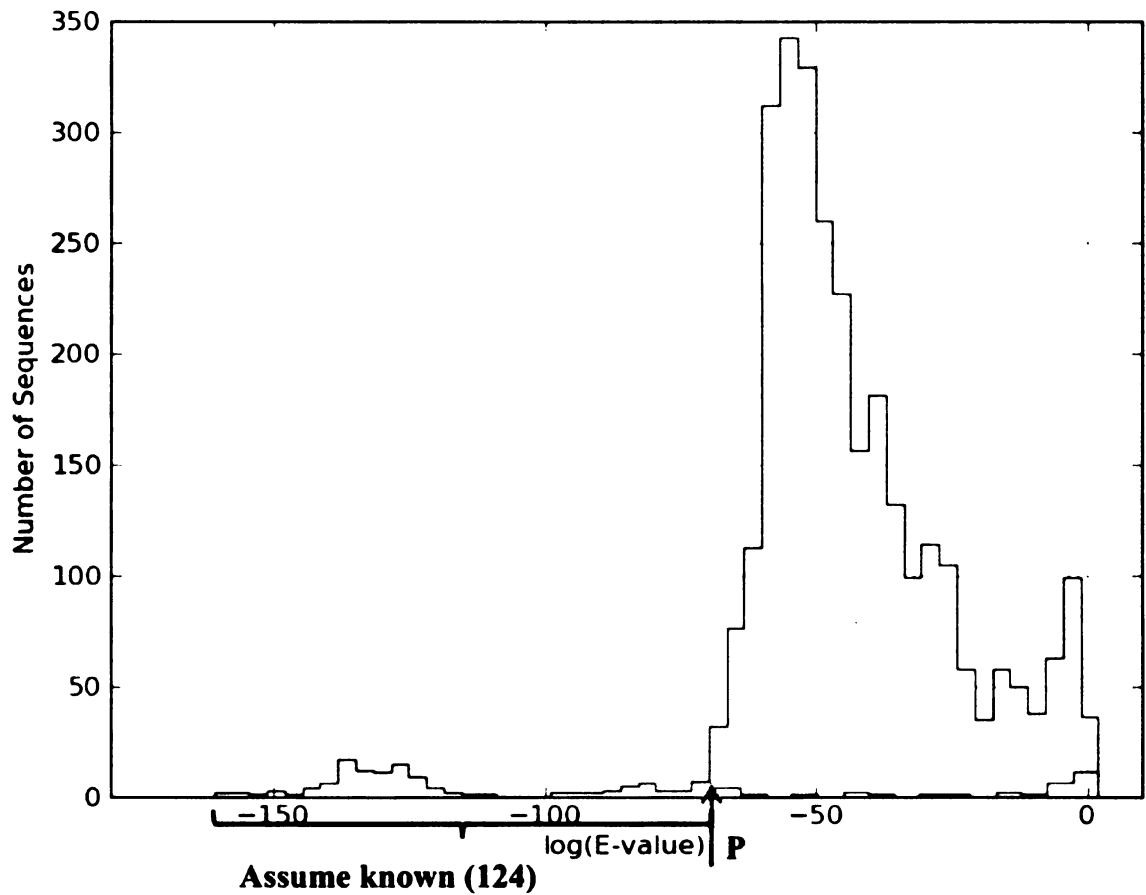
#### **Using the tree squeezing method to find more *nirK* genes:**

I evaluated the use of the tree clustering method, also called the tree squeezing method, to identify true *nirK* genes from those listed on the FunGene website. The *nirK* genes listed on FunGene are based on a profile matching of sequences in the NCBI non-redundant database which match a well-annotated seed model of 6 well-known *nirK* genes.

#### **Building the tree**

In order to build a tree for this method, I initially needed to identify additional “known” sequences (in addition to the six seed sequences). To do this, I searched the six seed *nirK* profile HMM (see Chapter 1) against all 3,055 *nirK*s listed on the FunGene website. From the distribution of E-values of all hits, I selected a cutoff (arrow in Figure 16) and treated hits with E-values lower than the cutoff as known *nirK*s and hits with E-values higher than the cutoff as potential *nirK*s. As a result, 124 *nirK*s were picked as known,

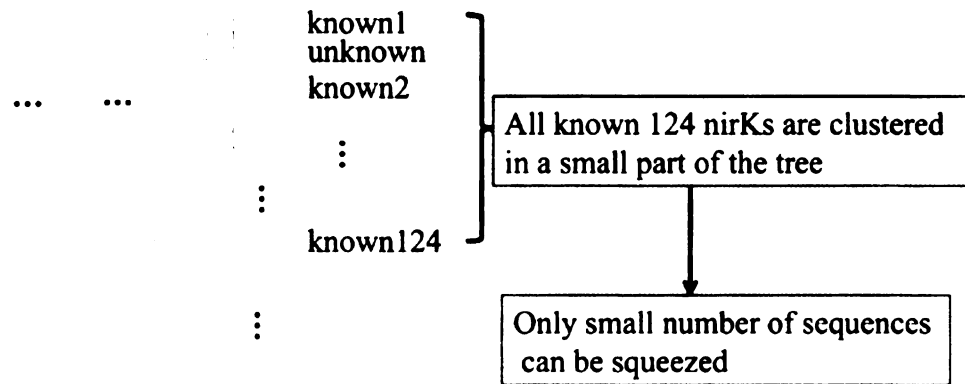
and the other were marked as unknown (Figure 15). All 3,055 sequences were aligned to the 6 seed *nirK* profile using the `hmmalign` function in the HMMER package. The MSA was then used to construct a neighbor-joining tree by QuickTree (Howe, Bateman et al. 2002). QuickTree, an implementation of the neighbor-joining algorithm, can quickly construct phylogenetic tree of thousands of sequences.



**Figure 15.** E-value distribution of 3055 *nirK*s from FunGene (blue) and 32 sequences from tree squeezing method (green). Sequences on the left of P (stringent cutoff) are treated as known *nirK*s.

In the tree shown in Figure 16, if an unknown sequence L or a branch node N is neighbored with a known *nirK* or a node of known *nirK*s, and there is another known

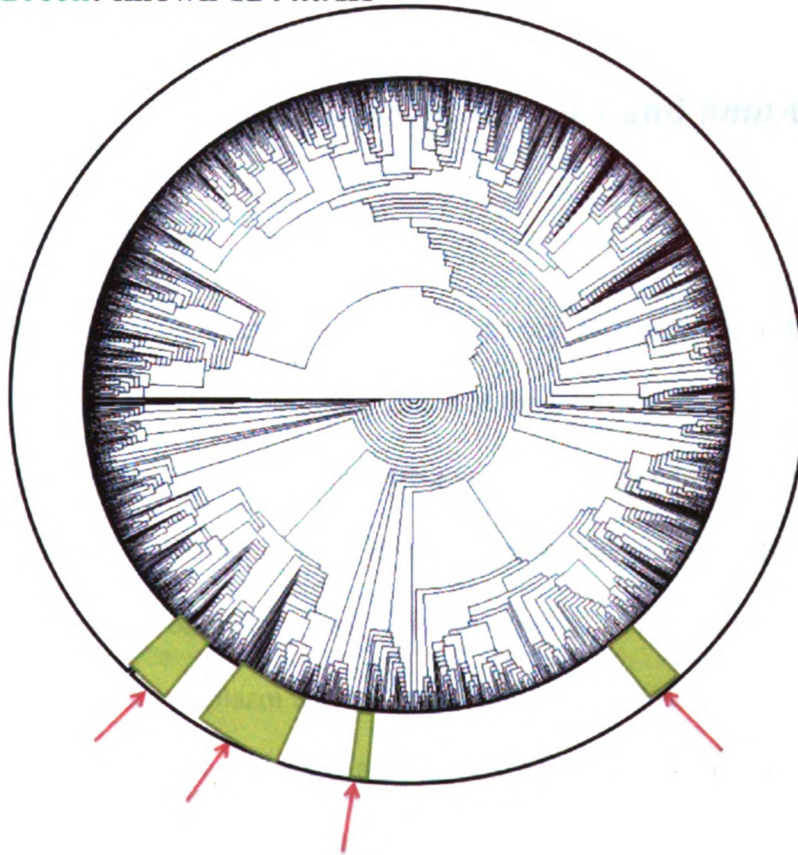




**Figure 17.** A hypothesis explaining why only small number of sequences is squeezed.



**Green: known 124 *nirKs***



**Figure 18.** Phylogenetic tree of all 3055 FunGene *nirK* sequences. The 124 known *nirKs* are marked green. The tree shows the 124 *nirKs* treated as known are not phylogenetically diverse.

An explanation for the small number of squeezed sequences (32) is that the 124 known *nirKs* are not phylogenetically diverse and clustered in a small part of the whole tree (Figure 17). The position of 124 known *nirKs* (green) in tree in Figure 18 supports this conclusion (Figure 18).

Though I can use the tree squeezing method to add confidence to the conclusion that the 32 squeezed potential *nirK* genes are valid, further validation with biological experiments is required.

## Chapter 4

### Classification of *amoA* and *pmoA*

#### 4.1 Introduction:

Methanotrophs are a special group of bacteria utilizing methane as the only carbon and energy source. They are commonly found at the interface of aerobic and anaerobic environments (i.e., hot spring), and are a very important part of the global methane cycle (Oremland and Culbertson 1992). Methane monooxygenase (MMO) is the enzyme that catalyzes methane oxidation. Two different types of MMOs have been found, the soluble form (sMMO) in cytoplasm and the particular form (pMMO) bound to the membrane. Despite the functional similarity, they do not show any sequence similarity. The switch between their expressions is regulated by a copper ion ( $\text{Cu}^{2+}$ ) (Stanley, Prior et al. 1983). Only pMMO is present in all methanotrophs.

Ammonia monooxygenase (AMO) is the enzyme that catalyzes oxidation of ammonia, which is crucial for the global nitrogen cycle. AMO is found only in ammonia oxidizing nitrifying microbes (Holmes, Costello et al. 1995).

pMMO and AMO are both integral membrane proteins, having similar sequences and structures and thought to be evolutionarily related, despite their different physiological functions. *PmoA* and *amoA* are genes coding the alpha subunit of pMMO and AMO respectively. They are around 450 bp long and highly conserved (McTavish, Fuchs et al. 1993; Semrau, Chistoserdov et al. 1995). Their protein products are grouped in the same Pfam domain family called AMO (PF02461) and the same family (IPR003393) in

InterPro database. *PmoAs* are found in  $\alpha$ - and  $\gamma$ -Proteobacteria while *amoAs* have been detected in  $\beta$ -,  $\gamma$ -Proteobacteria and archaea. Some *amoAs* in  $\gamma$ -Proteobacteria are more similar to *pmoAs* from  $\gamma$ -Proteobacteria than other *amoAs* (Holmes, Costello et al. 1995) suggesting that these two genes share a common ancestry.

Given their sequence similarity, BLAST and HMMER3 fail to separate these two very closely related genes. I evaluate the resolution of the tree clustering methods to classify *pmoAs* and *amoAs*.

## **4.2 Data:**

Initially a collection of well-known sequences for BLAST and HMMER searches was created:

Eight well-known *pmoA* seed sequences were downloaded from FunGene, and six *amoA* seed sequences from well studied ammonia oxidizing nitrifying bacteria were downloaded from UniProtKB database. These genes were from species including *Nitrosomonas europaea* ATCC 19718, *Nitrosomonas eutropha* C-71, *Nitrospira briensis* C-128, *Nitrosovibrio tenuis* Nv1, *Nitrosococcus oceani* ATCC 19707, and *Nitrospira multiformis* ATCC 25196. The whole UniProtKB database release 2010\_05 (including Swiss-Prot and TrEMBL) was also downloaded.

Sequences are downloaded from UniProtKB to build a tree for tree clustering:

Sequences annotated as *amoA* (n = 176) from cultured were downloaded from UniProtKB. The keywords for the search were: (*amoA* AND (name:ammonia OR

name:*amoA*) NOT gene:*amoB* NOT gene:*amoC* AND gene:*amoA* NOT  
taxonomy:environmental NOT organism:Sali).

The 10,470 sequences annotated as *amoAs* that were sequenced directly from environmental samples (not cultured) were downloaded from UniProtKB. The keywords for the search were: (*amoA* AND (name:ammonia OR name:*amoA*) NOT gene:*amoB* NOT gene:*amoC* AND gene:*amoA* AND taxonomy:environmental).

Another 233 sequences annotated as *pmoAs* from other research using the culturing methods were downloaded from UniProtKB. Keywords for the search were: (*pmoA* NOT taxonomy:environmental AND (name:methane OR name:*PmoA*) AND gene:*pmoA*).

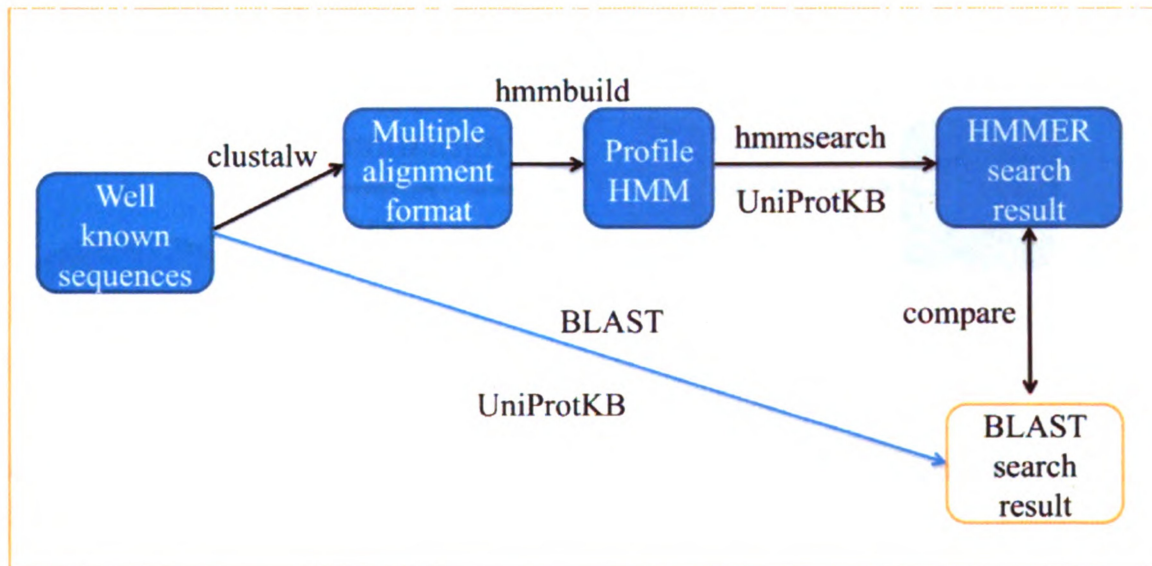
Finally 3062 sequences annotated as *pmoAs* that were sequenced directly from environmental samples (not cultured) were downloaded from UniProtKB using keywords: (*pmoA* AND taxonomy:environmental AND (name:methane OR name:*PmoA*))

## **4.3 Methods:**

### **4.3.1 BLAST and HMMER comparison (Figure 19)**

The 6 *amoA* seed sequences and 8 *pmoA* seed sequences were separately searched against UniProtKB database using NCBI-BLAST blastall with parameters: -p blastp -v 100000 -b 100000 -e 10.

Also, the 6 *amoA* seed sequences and 8 *pmoA* seed sequences were separately made into profile HMMs by hmmbuild in HMMER3. The two resulting HMMs were searched using HMMER against UniProtKB database using hmmsearch with default parameters.

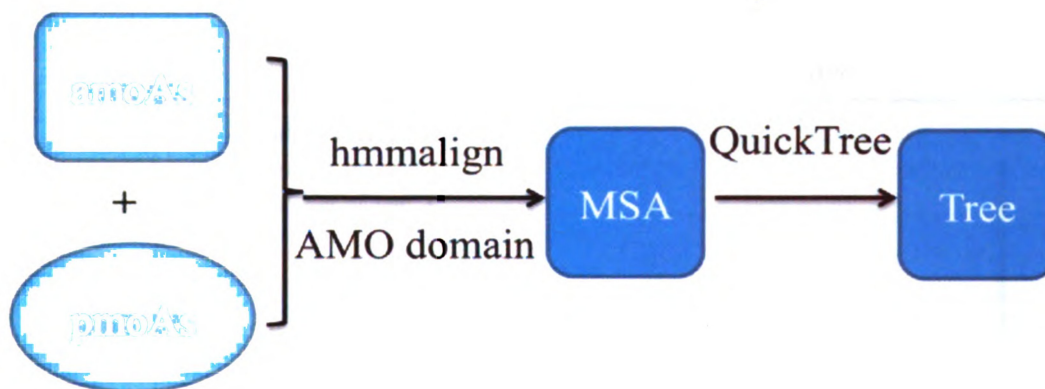


**Figure 19.** BLAST and HMMER comparison flow chart.

#### 4.3.2 Building the tree

First, the 176 sequences annotated by UniProtKB as *amoAs* and 233 sequences annotated as *pmoAs* from cultured bacteria were combined and aligned by *hmmalign* in HMMER3 using the AMO Pfam domain family (PF02461) profile HMM. The alignment was made into a neighbor-joining by QuickTree.

Second, all 10,646 sequences annotated by UniProtKB as *amoAs* and 3,295 sequences annotated by UniProtKB as *pmoAs* from both culturing method and environmental samples were combined and aligned by *hmmalign* in the HMMER3 package using AMO Pfam domain family profile HMM. The alignment was made into a neighbor-joining tree by QuickTree.



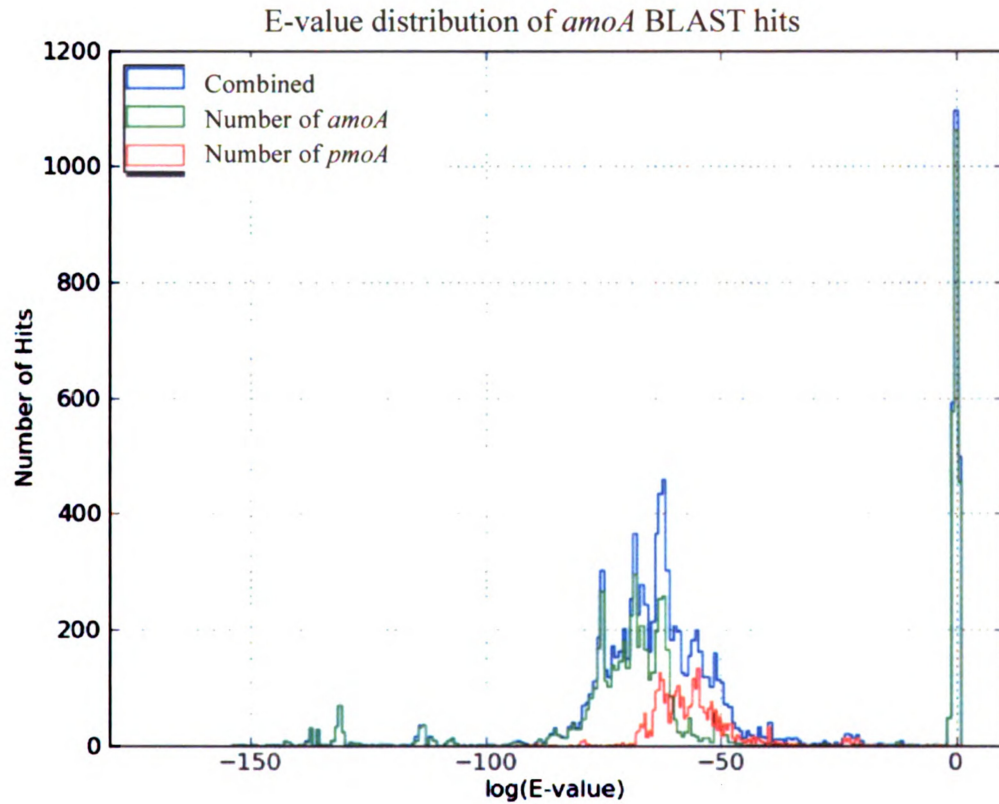
**Figure 20.** Flow chart of constructing phylogenetic tree of *amoAs* and *pmoAs*.

## 4.4 Results:

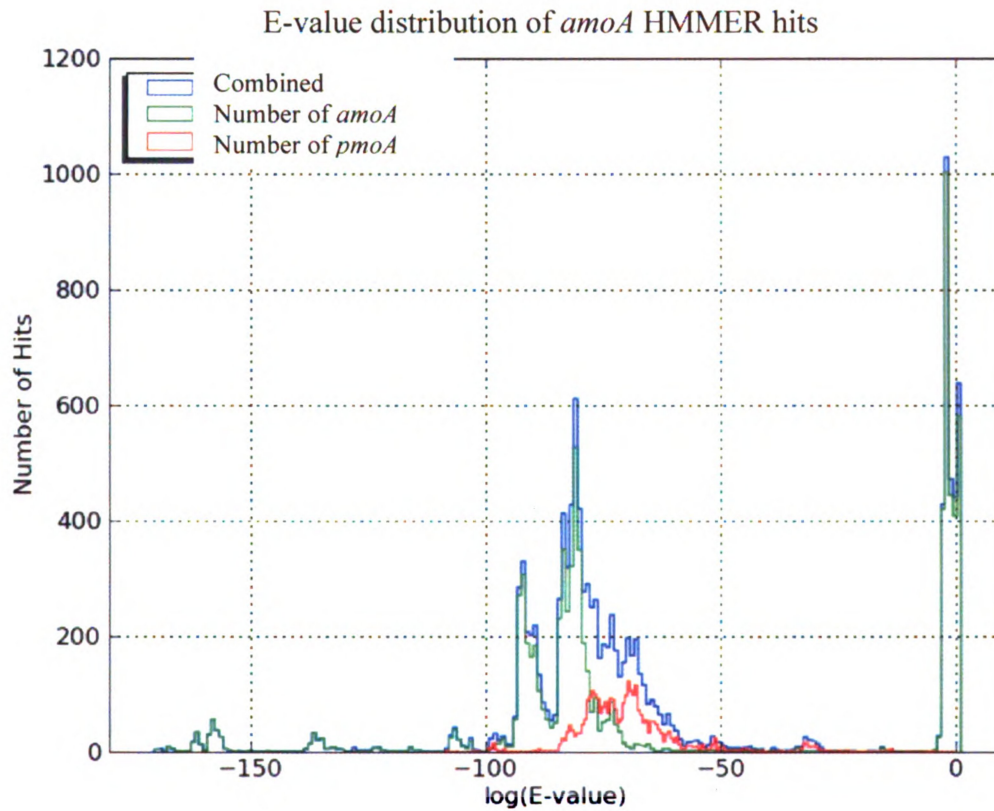
### 4.4.1 The use of BLAST and HMMER to differentiate between *amoAs* and *pmoAs* in UniprotKB

BLAST and HMMER search results against the UniProtKB database showed that sequences of *amoA* and *pmoA* could not be separated effectively (Figure 21-24). For example, in Figure 21, BLAST hits for *amoA* consisted of both sequences annotated as *amoA* and *pmoA* in the UniProtKB database. If a very low E-value was used, *amoAs* identified by BLAST were largely also annotated as *amoAs* in UniProtKB; however, using very stringent E-value cutoffs resulted in less sequences identified in general. Due to the high similarity of these genes, BLAST and HMMER did not have enough resolution to separate these *amoA* and *pmoA* sequences in UniProtKB.



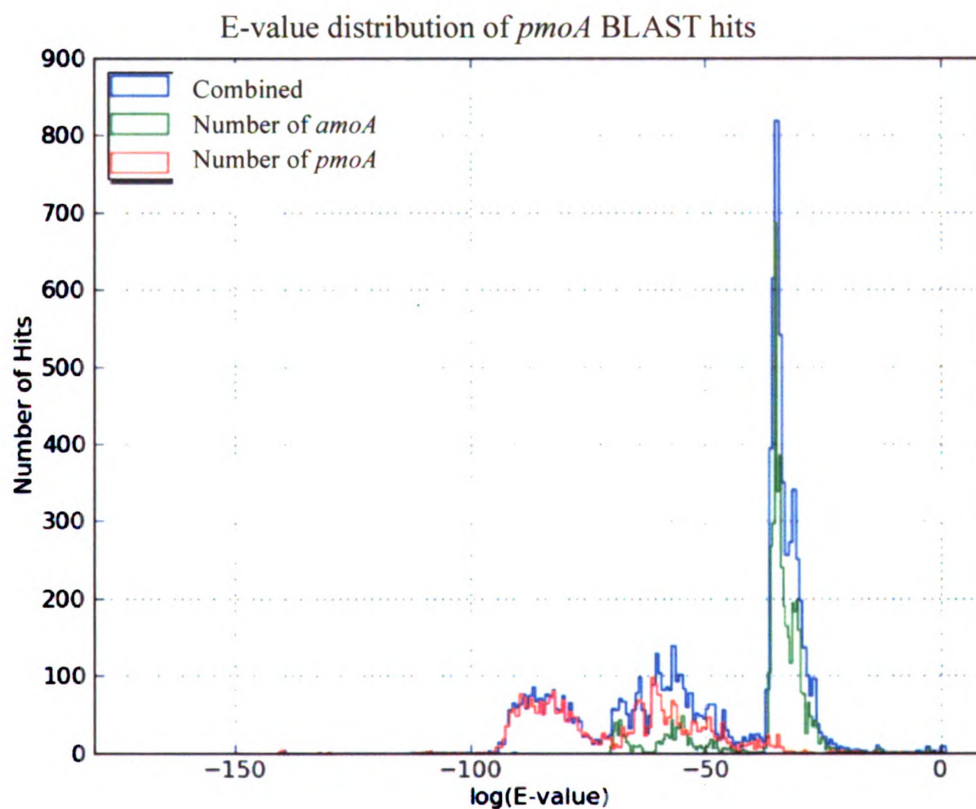


**Figure 21.** E-value distribution of *amoA* BLAST hits. Blue is overall E-value distribution of *amoA* BLAST hits. Green is a subset of overall BLAST hits that are annotated as *amoA* in UniProtKB. Red is a subset of overall BLAST hits that are annotated as *pmoA* in UniProtKB.

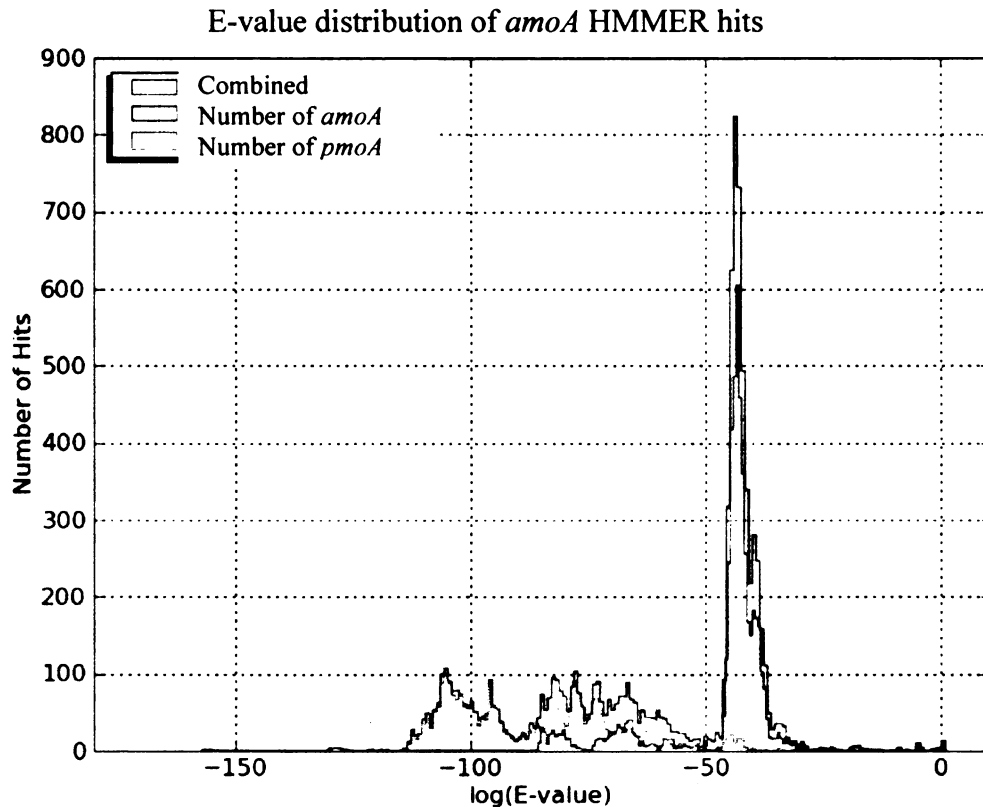


**Figure 22.** E-value distribution of *amoA* HMMER hits. Blue is overall E-value distribution of *amoA* HMMER hits. Green is a subset of overall HMMER hits that are annotated as *amoA* in UniProtKB. Red is a subset of overall BLAST hits that are annotated as *pmoA* in UniProtKB





**Figure 23.** E-value distribution of *pmoA* BLAST hits. Blue is overall E-value distribution of *pmoA* BLAST hits. Green is a subset of overall BLAST hits that are annotated as *amoA* in UniProtKB. Red is a subset of overall BLAST hits that are annotated as *pmoA* in UniProtKB.



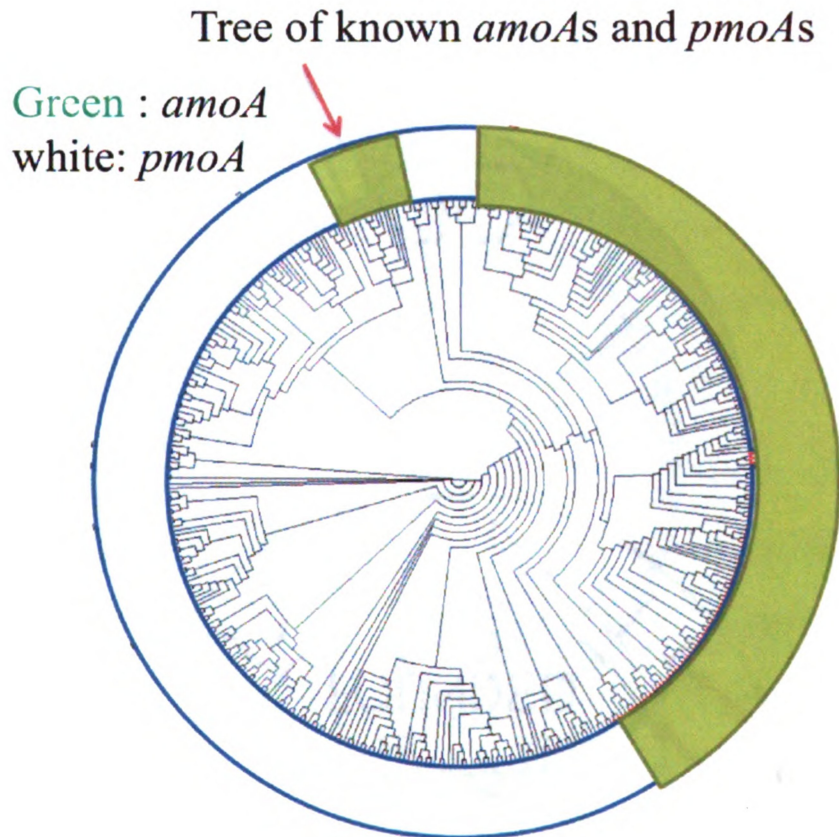
**Figure 24.** E-value distribution of *pmoA* HMMER hits. Blue is overall E-value distribution of *pmoA* HMMER hits. Green is a subset of overall HMMER hits that are annotated as *amoA* in UniProtKB. Red is a subset of overall HMMER hits that are annotated as *pmoA* in UniProtKB.

#### 4.4.2 Using the tree clustering method to differentiate *amoAs* and *pmoAs*

To effectively use the tree clustering method, I needed to identify a significant number of “known” *amoA* and *pmoA* sequences. Thus, I treated sequences from culturing methods as known sequences and those sequenced directly from environmental sample as unknown. The reasons supporting this assumption were: 1) most well known *amoAs* or *pmoAs* are obtained by culturing ammonia oxidizing species or methanotrophic species

and the culturing process is actually a screening step for ammonia oxidizing or methanotrophic species and thus increased the confidence that these annotations are real; and 2) environmental sequences, especially those sequenced by next generation sequencing techniques, are more likely to be incorrectly annotated because they are sequenced directly from environmental samples, without culturing first.

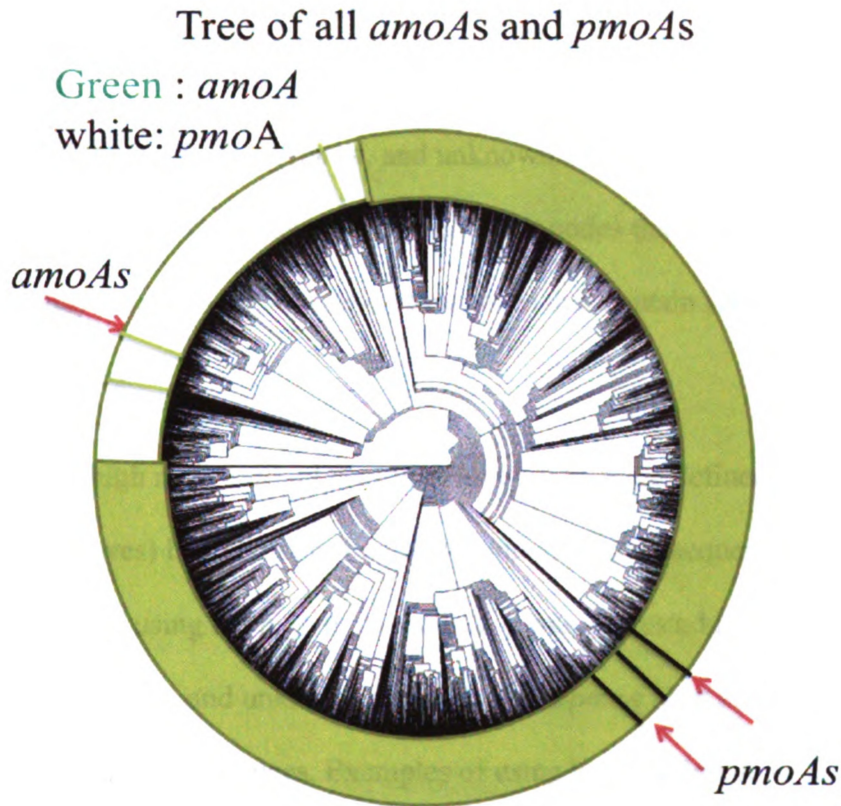
The tree built from “known” *amoAs* and *pmoAs* (sequences from culturing method) showed *amoAs* and *pmoAs* separated very well (Figure 25). There are also segregations for different taxonomy groups within *amoAs* or *pmoAs*. This separation is promising evidence that I can apply the tree clustering method for *amoA* and *pmoA* classification.



**Figure 25.** Tree of known *amoAs* and *pmoAs*. The green colored sequences are *amoAs*.

The others are *pmoAs*. Tree is made from known *amoAs* (from culture) and known *pmoAs* (from culture). *AmoAs* and *pmoAs* separate well on tree though a small cluster of *amoAs* from  $\gamma$ -proteobacteria are closer to *pmoAs* from  $\gamma$ -proteobacteria than other *amoAs* (referenced by an arrow in Tree).

Using all *amoAs* and *pmoAs* (sequences from both cultures and environmental samples) to build a tree, I observed that there was, in general, good separation of *amoAs* and *pmoAs* (Figure 26). However, some *pmoAs* were located on *amoA* tree clusters, and some *amoAs* were located on *pmoA* tree clusters suggesting possible incorrectly annotated sequences.



**Figure 26.** Trees of all *amoAs* and *pmoAs*. Tree is made from all *amoAs* and *pmoAs* (including those from culture and environmental samples). The majority of *amoAs* and *pmoAs* separate from each other though some *pmoAs* are scattered in *amoA* clusters and some *amoAs* are scattered in *pmoA* clusters.

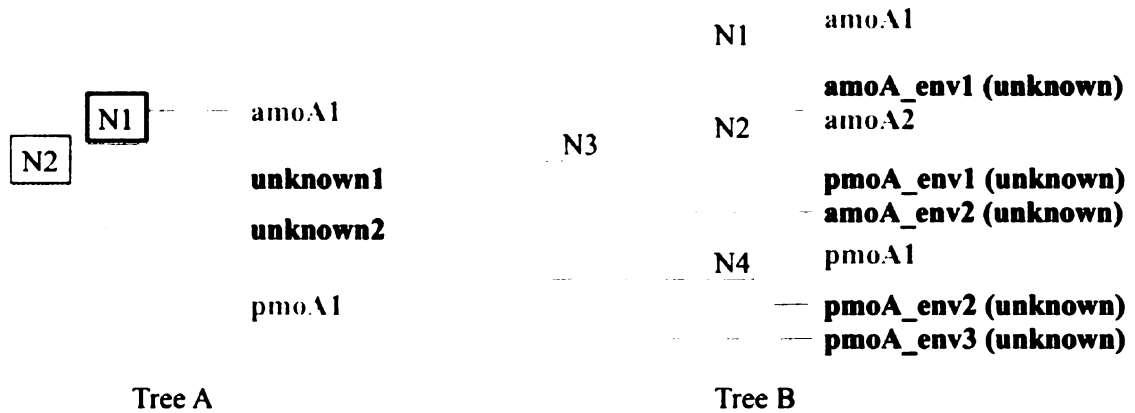
#### 4.4.2 Using low nodes and high nodes for tree clustering

To further classify unknown sequences (sequenced directly from environmental samples) based on the tree clusters, I implemented a low node method and a high nodes method to attempt to group together *amoA* and unknown sequences or *pmoA* and unknown sequences in neighboring tree branches.

1) Using the low node method, I identified the low node as the rightmost (closest to leaves) branch nodes that only contain a list of sequence types. The list can be any combination of *amoA*, *pmoA*, and unknown sequences. To identify *amoA* and *pmoA*, respectively, I was interested in *amoA* low nodes that contain only known *amoAs* and unknown sequences, and *pmoA* low nodes that contain known *pmoAs* and unknown sequences.

2) The high node method uses high nodes which are defined as the leftmost (furthest from leaves) branch nodes that only contain a list of sequence types. For identify *amoA* and *pmoA* using the high node method, I was interested in *amoA* nodes containing only known *amoAs* and unknown sequences, and *pmoA* high nodes containing known *pmoAs* and unknown sequences. Examples of using both the low node and high node methods to identify an unknown gene are shown in Figure 27.

The low node method gave more accurate classification, while the high node classified more unknown sequences but with less accuracy. Since unknown sequences outside the low node can also be included in high node (unknown2 of Tree A in Figure 27), the high node method is able to classify more unknown sequences and the sequences classified by low node method is a subset of those classified by high node method.



**Figure 27.** Simple example trees for super node analysis. Known *amoAs* are marked as *amoA1*, *amoA2*, ... Known *pmoAs* are marked as *pmoA1*, *pmoA2*, ... Environmental sequences are treated as unknown sequences for classification. Those environmental sequences annotated as *amoAs* in UniProtKB are marked as *amoA\_env1*, *amoA\_env2*, ... Those annotated as *pmoAs* in UniProtKB are marked as *pmoA\_env1*, *pmoA\_env2*, ... In Tree A, the low node of *amoA* and ‘env’ (environmental) sequence is N1; the high node of *amoA* and ‘env’ sequence is N2, which include one more ‘env’ sequence (*amoA\_env2*) than low node method. In Tree B, the low node of *amoA* and ‘env’ sequence is N1 and N2, which shows *pmoA\_env1* may be incorrectly annotated; the high node of *amoA* and ‘env’ sequence is N3, which includes one more ‘env’ sequence than low node method. The low node of *pmoA* and ‘env’ sequence is N4; the high node of *pmoA* and ‘env’ sequence is still N4.

As Table 2 shows both low node and high node methods have sequences classified differently from UniProtKB annotation. High node method was able to classify the most unknown sequences (12,980/13,532), while low node method classified only a small number of sequences (528/13,532).

Using the low node, a total of 291 and 237, respectively, could be classified as potential *amoAs* or *pmoAs* (Table 2). Using the high node, a total of 10,242 and 2,738, respectively, could be classified as potential *amoAs* or *pmoAs* (Table 2). Several sequences identified by the tree squeeze method disagreed with annotations in UniProtKB and require further investigation.

**Table 2.** Numbers of sequences classified by various methods.

	Low Node				High Node			
	<i>amoA</i>	<i>pmoA</i>	unclassified	sum	<i>amoA</i>	<i>pmoA</i>	unclassified	sum
Total	291	237	13004	13532	10242	2738	552	13532
UniProt <i>amoA</i>	289	7	10174		10159	31	280	
UniProt <i>pmoA</i>	2	230	2830		83	2707	272	



#### **4.4.3 Investigating sequences identified by tree clustering that disagreed with UniProt Annotations**

##### Comparing sequence alignments

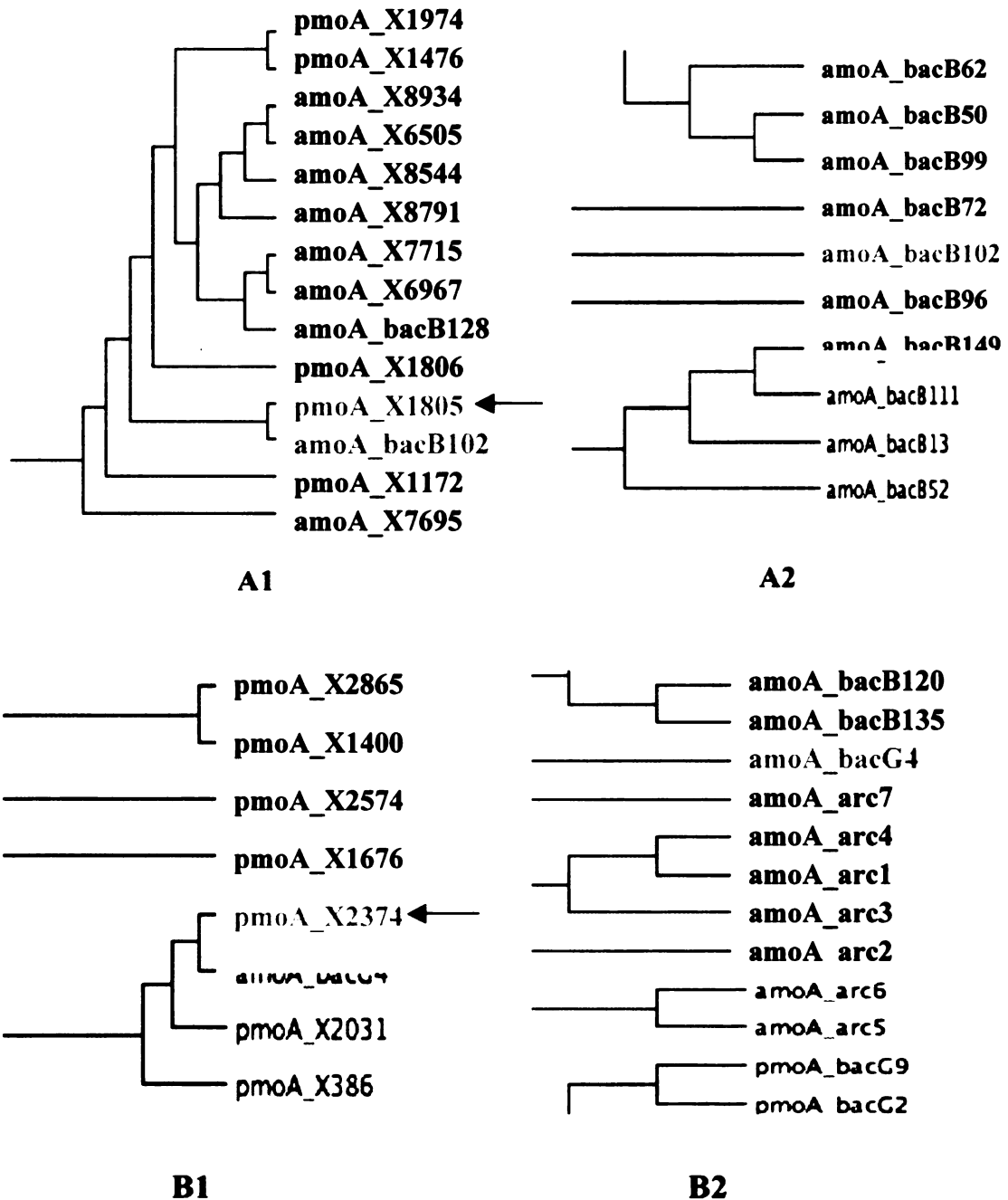
When checking the alignment of some UniProtKB *amoAs* that agree with low node method (correctly classified) and UniProtKB *pmoAs* (incorrectly classified), I found that they are very similar and I could not recognize *amoAs* and *pmoAs* from each other based on the alignment alone (Figure 28)



I evaluated the position of the possibly misclassified genes in the tree. In general, I expected that either (1) a possibly misclassified *pmoA* was in a cluster of known or unknown *amoAs* or (2) a possibly misclassified *amoA* was in a cluster of *pmoAs*. By investigating the nodes surrounding the sequence in question, I hoped to better understand the misclassification.

1) For example, in Figure 29.A1, the known neighboring *amoA*, *amoA\_bacB102*, helped the classification of *pmoA\_X1805* and the surrounding *amoAs* (known or unknown) strengthened the possibility that this may be a misclassification of *pmoA\_X1805* in the UniProt database. Note that the position of this known *amoA* was also evaluated in relation to other known *amoAs* in the tree of all known *amoAs* and *pmoAs* to further support this hypothesis (Figure 29.A2).

# Evaluating sequence positions in tree



**Figure 29.** Closeups of two possibly misclassified *pmoA*s in tree of all *amoA* and *pmoA* (A1 and B1) and two known *amoA*s (green and grey shaded) that help classify the two possibly misclassified *pmoA*s in tree of all known *amoA*s and *pmoA*s.

2) Another example of a possible misclassified gene was *pmoA\_X2374* which was in a cluster of *pmoAs* (known or unknown) (Figure 29.B1). Looking more closely at its neighboring *amoA*, *amoA\_bacG4*, I see that it was located in a big cluster of both *amoAs* and *pmoAs* of the tree of known *amoAs* and *pmoAs*. In this situation, I had to be skeptical about whether *amoA\_bacG4* was correctly annotated. This example highlighted the need for more reference sequences before more conclusions could be drawn (here the known *amoAs* or *pmoAs* still may be incorrectly annotated).

#### **4.5 Conclusion:**

In general, this study found that tree clustering method gave higher resolution for separating two genes with similar sequences but different functions, such as *amoA* and *pmoA*. The low node method gave better accuracy in classification, while the high node method classified more unknown sequences with lower but still reasonably good accuracy provided by the tree hierarchy. The high node gave more aggressive assertions that the unknown sequences outside lode nodes are also classified with lower confidence. Meanwhile, the difference between tree clustering method classification and UniProtKB annotation showed that some of the annotations of *amoAs* and *pmoAs* in UniProtKB were probably wrong, especially those of genes sequenced directly from environmental samples. These sequences need further computational or biological verification.

The stronghold of the tree clustering method is also its shortcoming—it is based on an underlying tree. Problems in multiple sequence alignments and tree constructing tools

will result in incorrect classification with the tree clustering method. An example of this would be sequences containing multiple domains which are difficult to cluster on a tree.

I used the tree clustering method to differentiate *amoA* and *pmoA* mainly for family classification. However, I could also study the evolutionary relationship of *amoA* and *pmoA*. Significant efforts have focused on quantification of these genes in samples due to their ecological importance (Webster, Embley et al. 2005; De Corte, Yokokawa et al. 2009). A close evolutionary relationship has been inferred based on their high sequence similarity, but their evolutionary mechanism still needs further study. The multiple *amoA* or *pmoA* copies within individual strains have resulted from gene duplication events (paralogous evolution) not horizontal gene transfer (Klotz and Norton 1998). But the reason that initiated the paralogous evolution and the time when paralogous evolution started with respect to speciation is still unknown. What is the evolutionary relationship between *amoA* and *pmoA*, convergent or divergent? I may use comparative genomics, phylogenetics, or other methods to get a better understanding of their evolution mechanism.

## Chapter 5

### Conclusions and future work

#### 5.1 Conclusions:

##### **5.1.1 Biology is increasingly reliant on sequence similarity with the use of new sequencing technology, and sequence similarity may not give accurate gene annotations.**

When using BLAST and HMMER to identify *nirK* genes in a metagenomics dataset, I found that BLAST hits could be unreliable, with several false positives within the BLAST results. Using these tools to classify *amoA* and *pmoA*, these genes could not be differentiated due to sequence similarity. The examples shown in this study exemplify that computational methods infer homology from similarity, but similarity does not necessarily mean homology. A major challenge to using these methods is a lack of standards or controls to assess the “truth”. Biological experiments can be used but are expensive and take long time. For most genes, there are only a few sequences which are biologically verified. I should be cautious when using annotation tools and consider the possibility of incorrect annotation.

##### **5.1.2 Annotations are sometimes incorrect.**

The classifications of *amoA* and *pmoA* genes in our tree which disagree with UniProt annotations highlight the possibility of incorrect annotations in reference databases. Homology searches based on incorrect annotations will result in future incorrect

annotations. Tools which could evaluate the accuracy of annotations would be very useful for making sound biological conclusions from homology searches.

### **5.1.3 Narrow phylogenetic sampling is a problem.**

Our ability to detect novel *nirK* and reliably classify *pmoA/amoA* is strongly affected by the lack of diverse *nirK/pmoA/amoA*. Tools which could highlight important “sequencing gaps” would be useful in providing more resolution to maximize gene discovery using homology searches to annotate sequences.

## **5.2 Novel approaches:**

Here I used the tree clustering method as a filtration step downstream of initial homology search. The general tree clustering idea, which assumes more similar sequences are closer in a tree, is already widely used (Loewenstein and Linial 2008; Singh, Doerks et al. 2009). In this study, I presented and evaluated some new “tricks” for discovering “sequence closeness” including “tree squeezing”, “low node and “high node” methods to complement other methods such as BLAST or HMMER. These methods helped refine the result of BLAST and HMMER search and found possible misclassification of current reference database.

## **5.3 Future directions:**

A phylogenetically diverse set of well known sequences of a gene is very important for annotation and classification of new sequences. Sequences that could contribute substantially to annotation and classification can be picked out from the phylogenetic tree for further biology experiment verification.



I present here tools evaluated with specific examples. A natural extension of this work would be to use these tools for other datasets, metagenomic and otherwise. For example, the low/high node methods used here could be applied to the classification of other protein families with closely related subfamilies such as those that can be found in the ProtoNet database.

The tree clustering method may also be a method to identify novel groups of genes within a family. For example, big nodes close to the root with all unknown or new sequences are possibly new subgroup of the family.

#### **5.4 Parting Thoughts:**

A repetitive theme throughout this study was that tools that are currently used, whether BLAST, HMMER, or even a reference database, can often give the incorrect results which could result in wrong conclusions. Understanding the limits of the tools and their underlying methodology is important in their effective use. Much like experimental methods, using computational tools requires preliminary thought and planning.

## Bibliography

- . Retrieved July 21, 2010, from [http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST\\_algorithm.html](http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html).
- . Retrieved August 15, 2009, from <http://fungene.cme.msu.edu/index.spr>.
- Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." *J Mol Biol* **215**(3): 403-410.
- Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* **25**(17): 3389-3402.
- Barton, G. J. (1990). "Protein multiple sequence alignment and flexible pattern matching." *Methods Enzymol* **183**: 403-428.
- Bashford, D., C. Chothia, et al. (1987). "Determinants of a protein fold. Unique features of the globin amino acid sequences." *J Mol Biol* **196**(1): 199-216.
- De Corte, D., T. Yokokawa, et al. (2009). "Spatial distribution of Bacteria and Archaea and amoA gene copy numbers throughout the water column of the Eastern Mediterranean Sea." *ISME J* **3**(2): 147-158.
- Eddy, S. R. (1998). "Profile hidden Markov models." *Bioinformatics* **14**(9): 755-763.
- Eddy, S. R. (2009). "A new generation of homology search tools based on probabilistic inference." *Genome Inform* **23**(1): 205-211.
- Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland, Mass., Sinauer Associates.
- Gonzalez, M. W. and W. R. Pearson (2010). "Homologous over-extension: a challenge for iterative homology searches." *Nucleic Acids Res* **38**(7): 2177-2189.
- Gribskov, M., A. D. McLachlan, et al. (1987). "Profile analysis: detection of distantly related proteins." *Proc Natl Acad Sci U S A* **84**(13): 4355-4358.
- Holmes, A. J., A. Costello, et al. (1995). "Evidence That Particulate Methane Monooxygenase and Ammonia Monooxygenase May Be Evolutionarily Related." *Fems Microbiology Letters* **132**(3): 203-208.
- Howe, K., A. Bateman, et al. (2002). "QuickTree: building huge Neighbour-Joining trees of protein sequences." *Bioinformatics* **18**(11): 1546-1547.
- Kaplan, N., M. Friedlich, et al. (2004). "A functional hierarchical organization of the protein sequence space." *BMC Bioinformatics* **5**: 196.
- Klotz, M. G. and J. M. Norton (1998). "Multiple copies of ammonia monooxygenase (amo) operons have evolved under biased AT/GC mutational pressure in ammonia-oxidizing autotrophic bacteria." *Fems Microbiology Letters* **168**(2): 303-311.
- Loewenstein, Y. and M. Linial (2008). "Connect the dots: exposing hidden protein family connections from the entire sequence tree." *Bioinformatics* **24**(16): i193-199.

- Madera, M. and J. Gough (2002). "A comparison of profile hidden Markov model procedures for remote homology detection." Nucleic Acids Res **30**(19): 4321-4328.
- Madigan, M. T., J. M. Martinko, et al. (2006). Brock biology of microorganisms. Upper Saddle River, NJ, Pearson Prentice Hall.
- McTavish, H., J. A. Fuchs, et al. (1993). "Sequence of the gene coding for ammonia monooxygenase in *Nitrosomonas europaea*." J Bacteriol **175**(8): 2436-2444.
- Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.
- Mount, D. W. (2004). Bioinformatics : sequence and genome analysis. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.
- Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-453.
- Oremland, R. S. and C. W. Culbertson (1992). "Importance of Methane-Oxidizing Bacteria in the Methane Budget as Revealed by the Use of a Specific Inhibitor." Nature **356**(6368): 421-423.
- Riesenfeld, C. S., P. D. Schloss, et al. (2004). "Metagenomics: genomic analysis of microbial communities." Annu Rev Genet **38**: 525-552.
- Rost, B. (1999). "Twilight zone of protein sequence alignments." Protein Eng **12**(2): 85-94.
- Semrau, J. D., A. Chistoserdov, et al. (1995). "Particulate methane monooxygenase genes in methanotrophs." J Bacteriol **177**(11): 3071-3079.
- Singh, A. H., T. Doerks, et al. (2009). "Discovering functional novelty in metagenomes: examples from light-mediated processes." J Bacteriol **191**(1): 32-41.
- Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-197.
- Stanley, S. H., S. D. Prior, et al. (1983). "Copper Stress Underlies the Fundamental Change in Intracellular Location of Methane Mono-Oxygenase in Methane-Oxidizing Organisms - Studies in Batch and Continuous Cultures." Biotechnology Letters **5**(7): 487-492.
- Tiedje, J. M. (1988). Ecology of denitrification and dissimilatory nitrate reduction to ammonium. New York, John Wiley & Sons.
- Webster, G., T. M. Embley, et al. (2005). "Links between ammonia oxidizer species composition, functional diversity and nitrification kinetics in grassland soils." Environ Microbiol **7**(5): 676-684.
- Wong, K. M., M. A. Suchard, et al. (2008). "Alignment uncertainty and genomic analysis." Science **319**(5862): 473-476.
- Wu, C. H., R. Apweiler, et al. (2006). "The Universal Protein Resource (UniProt): an expanding universe of protein information." Nucleic Acids Res **34**(Database issue): D187-191.

Wu, M. and J. A. Eisen (2008). "A simple, fast, and accurate method of phylogenomic inference."  
Genome Biol **9**(10): R151.



MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 03063 7940