Constructing Phylogenetic Trees using Multiple Sequence Alignment

Ryan M. Potter

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2008

Program Authorized to Offer Degree:
Institute of Technology – Tacoma

University of Washington
Graduate School

This is to certify that I have examined this copy of a master's thesis by

Ryan M. Potter

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____
Isabelle Bichindaritz

_____
Joseph Felsenstein

_____
Menaka Muppa

Date: _____

In presenting this thesis in partial fulfillment of the requirements for a master's degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Any other reproduction for any purpose or by any means shall not be allowed without my written permission.

Signature_____

Date_____

University of Washington

Abstract

Constructing Phylogenetic Trees using Multiple Sequence Alignment

Ryan M. Potter

Chair of the Supervisory Committee:
Professor Isabelle Bichindaritz
Computing and Software Systems

Phylogenetics is the study of evolutionary relatedness amongst organisms. The genetic relationships between species can be represented using phylogenetic trees. Advances in genomics have enriched the range of computational methods available for assisting experts in building these trees. Among other methods, these trees can be built by comparing genetic sequences of various species. The current implementations of multiple sequence alignment have limitations that prevent them from constructing accurate phylogenetic trees when sequences with low similarity are contained in the dataset. The purpose of this project is to modify the ClustalW sequence alignment algorithm so that it can be used to construct a more accurate tree when highly divergent sequences are present. The modifications to the existing algorithm consist of two parts. First, the highly divergent sequences are identified within the dataset by analyzing the pairwise alignment scores. Next the guide tree, which is used to determine the order that the sequences are aligned in, is modified so that the highly divergent sequences are aligned last. Mitochondrial genome sequences of species with known phylogenetic trees are used as a dataset for testing. ClustalW and PHYLIP provide a variety of methods for constructing

trees using the multiple sequence alignment as input. These trees are compared to the known tree to determine which version of the algorithm provides a more accurate tree. The results of this study show that the modified version of ClustalW produces a more accurate evolutionary tree in the majority of all the tests. In addition, the modified algorithm is more capable of correctly placing the highly divergent sequences in the phylogenetic tree.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank all the committee members for taking the time to proofread my thesis. A special thanks goes to Dr. Felsenstein for getting me headed in the right direction. Also a special thanks goes to Dr. Bichindaritz for all the help she has provided me throughout this process. Her guidance helped immensely in completing this thesis.

Chapter 1

INTRODUCTION

There are somewhere between 5 and 100 million living species of organisms alive on earth today. There is evidence that suggests that all of these organisms are genetically related. These genetic relationships can be represented by an evolutionary tree called the tree of life. The tree of life represents the phylogeny of all organisms, which is the history of the organism's lineage as they change through time. Large scale projects are taking place under the sponsorship of the National Science Foundation (NSF) Assembling the Tree of Life (ATOL) initiative.

Organisms have evolved over time from ancestral forms to more derived forms. These new forms keep many of their ancestral features. Some of these features gradually change to help organisms adjust to their environment. Studying the phylogeny of organisms can help explain the similarities and differences among species [15].

There are various techniques used to create phylogenetic trees and most of them rely on aligned genetic sequences to perform this task. Probably the most popular genetic sequence alignment algorithm is ClustalW [14]. Although successful in its domain, ClustalW is very sensitive to highly divergent sequences. Therefore the purpose of this project is to modify the ClustalW sequence alignment algorithm so that it can be used to construct a more accurate tree when highly divergent sequences are present.

In chapter 2, this relationship between phylogenetics and multiple sequence alignment is explained. In addition, a popular program used for multiple sequence alignment is presented. The specific problem this thesis attempts to solve is outlined in chapter 3. Chapter 4 describes the new method proposed to improve upon the existing tree construction methods. The dataset used for testing is described in chapter 5. The results of the tests are analyzed in chapter 6. Similar work is discussed in chapter 7 and future work is explored in chapter 8. Lastly, the paper finishes with some conclusions from this research project in chapter 10.

Chapter 2

BACKGROUND INFORMATION

2.1 Phylogenetics

Phylogenetics is an area of research concerned with finding the genetic relationships between various organisms. Originally, phylogenetics mainly used morphological features such as size, color, fur, or other physical characteristics to determine relationships. Modern phylogenetics relies on information extracted from genetic material such as DNA, RNA or protein sequences [12].

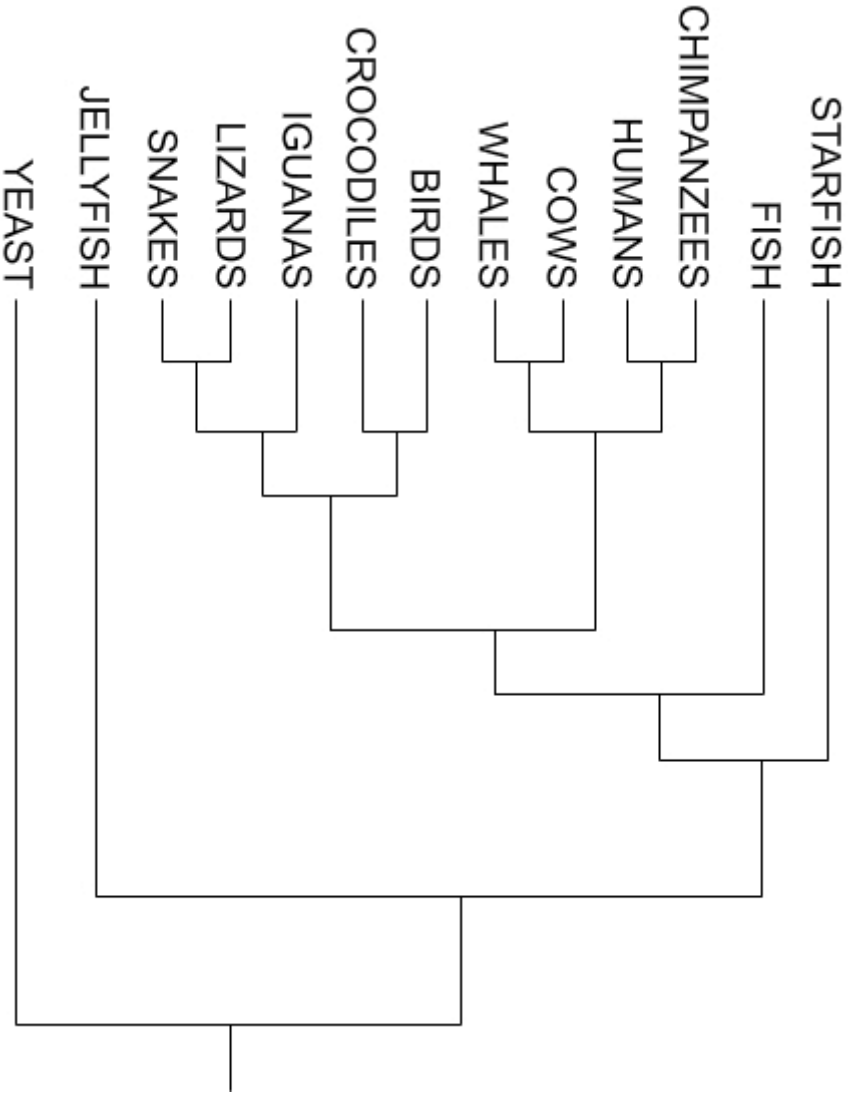

Figure 2.1: Known phylogenetic tree.

A way of visually representing these relationships is with a phylogenetic tree as seen in Figure 2.1. These trees show the evolutionary relationships amongst various species by way of common ancestors. Each node in the tree with descendants represents the most recent common ancestor of the descendents.

Using Figure 2.1, the tree would imply that iguanas share a common ancestor with the snakes and lizards [1, 13].

2.2 Multiple Sequence Alignment

Sequence alignment is a way of arranging sequences of DNA, RNA, or proteins in order to distinguish regions of similarity. A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences such as protein, DNA, or RNA. Typically it is implied that the set of sequences share an evolutionary relationship, which means they are all descendents from a common ancestor. These regions may correspond to functional, structural, or evolutionary relationships between the sequences. Alignments can reflect a degree of evolutionary change between sequences that are descendants from a common ancestor. There is a relationship between phylogenies and sequence alignments [4].

To find the globally optimum alignment, a dynamic programming technique can be used if one uses a parsimony approach and a particular scoring scheme. There is no universally agreed upon scoring scheme. This approach is computationally expensive and impractical since it has been shown to be a NP-complete problem [2]. Instead, heuristics are commonly used to perform a multiple sequence alignment. This research focused on studying one heuristic approach called progressive alignment. One popular program that employs a progressive alignment method is ClustalW.

2.3 ClustalW

ClustalW is a popular program used for multiple sequence alignment and for preparing phylogenetic trees. Its portability amongst various computing platforms is the main reason for its widespread use. Due to its popularity and the availability of source code, ClustalW was used for this project. The progressive alignment algorithm used by ClustalW to perform a multiple sequence alignment can be broken down into three major steps.

First, all pairs of sequences are aligned separately and then a distance matrix is calculated giving the divergence of each pair of sequences. A full dynamic programming alignment is calculated for each pair using two gap penalties, one for opening a gap and another for extending a gap. The score in the distance matrix is computed by taking the number of identities in the best alignment divided by the number of residues compared excluding gap positions. Then that number is multiplied by 100 and subtracted from 1.0 to give a value between 0 and 1.0.

Next, a guide tree is calculated which will be used to guide the final multiple alignment process. This tree is calculated by using the distance matrix from the first step and a Neighbor-Joining clustering algorithm. Weights are also assigned to each sequence depending on their distance from the root of the tree. By contrast, in the original Clustal progressive alignment algorithm, all sequences would be equally weighted.

Finally, the sequences are progressively aligned according to the branching order in the guide tree. To do this a series of pairwise alignments are used to align larger and larger groups of sequences. First, proceed from the tips of the rooted tree towards the root. At each alignment a full dynamic programming algorithm is used with penalties for opening and extending gaps. Each step aligns two existing alignments or sequences. Gaps that are present in the older alignments stay in place. When all the sequences have been considered a final alignment is produced. That final alignment can then be used to construct a phylogenetic tree for those species [14].

One disadvantage of a progressive alignment approach is that, once an alignment has been performed involving some of the species, this alignment is never reconsidered despite what other decisions are made for the remaining species. This can lead to inaccuracies in the final alignment [4].

Chapter 3

PROBLEM STATEMENT

In ClustalW the sequences that are hardest to align are the sequences with the lowest similarity to other sequences in the set. When one or more of these highly divergent sequences are contained in the dataset, the phylogenetic tree constructed based on the multiple sequence alignment tends to be inaccurate.

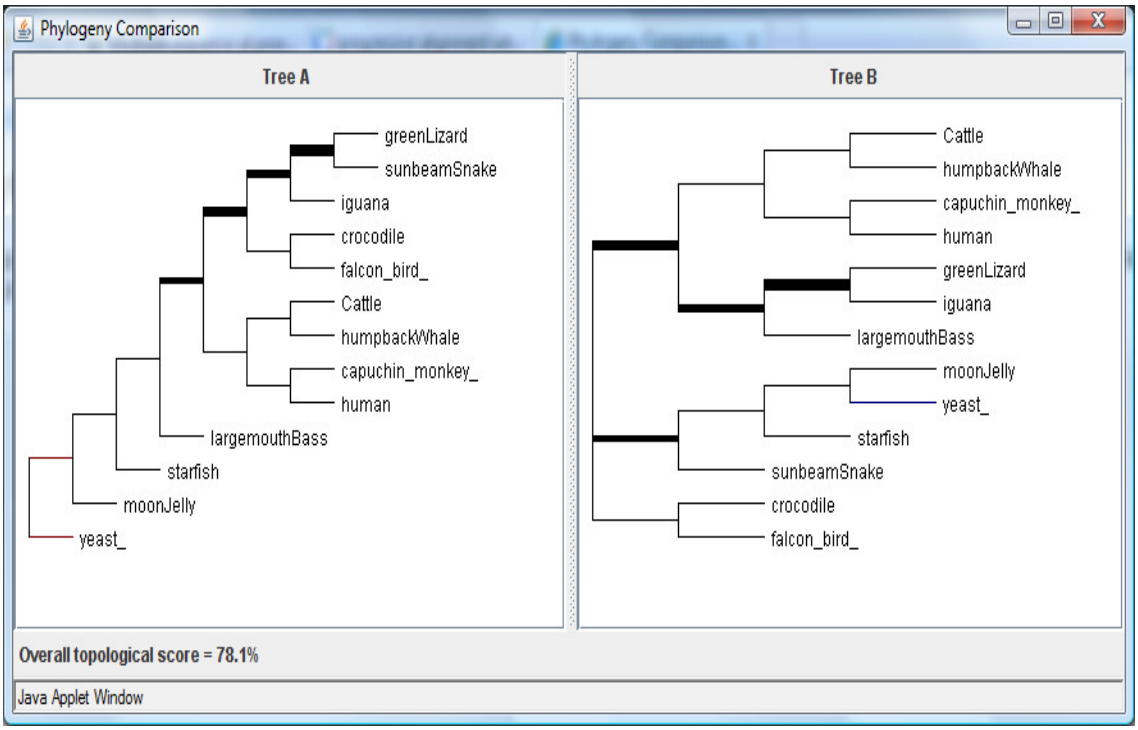

Figure 3.1: Comparison of trees.

Figure 3.1 allows us to see the errors generated by ClustalW when three highly divergent sequences are contained in the dataset. These highly divergent sequences are yeast, moon jellyfish, and starfish. Tree A represents the known tree,

which is the accurate phylogenetic tree for those species [13]. Tree B is the phylogenetic tree produced using ClustalW to generate both the multiple sequence alignment and phylogenetic tree. In Tree A we can see that the three highly divergent sequences are placed closely to the root of the tree. In contrast, Tree B places these same sequences further away from the root, which is incorrect.

It is the goal of this project to modify the ClustalW algorithm so that the phylogenetic trees produced are more accurate when highly divergent sequences are present in the dataset.

Chapter 4

NEW METHOD FOR GUIDE TREE CONSTRUCTION

To accomplish the goal of this project, the ClustalW algorithm was modified. The alteration to the existing algorithm consists of two parts. First, the highly divergent sequences are detected. Second, the guide tree is modified so that these identified sequences are aligned last.

During the initial pairwise alignment, ClustalW assigns a score to each pair of sequences. This score is a percentage value based on their similarity. A score closer to 0 would indicate that the two sequences share little in common. After all the pairwise alignment scores are computed, a detection algorithm is used to identify highly divergent sequences. In the modified algorithm, highly divergent sequences are defined as having a pairwise alignment score of less than 10.

After the pairwise alignment, ClustalW generates a guide tree which is used to order the sequences for the final multiple sequence alignment. Placing the divergent sequences closer to the root of the tree will ensure that they are aligned last. ClustalW does not always generate the guide tree in this manner so the guide tree is modified.

There are a few possible scenarios that need to be addressed. The tree inferred by the clustering algorithm in ClustalW is unrooted. So if there is a highly divergent sequence detected then the tree needs to be rooted using that highly divergent sequence as the root. If there are no divergent sequences detected the guide tree is not changed. If there is one divergent sequence, then it is placed

closest to the root of the tree. If there are 2 or more divergent sequences, then we first check to see if any of those divergent sequences have a high pairwise alignment score between each other. If they do, then they will be placed in the same cluster on the tree closest to the root. If they do not, then they will be placed in separate clusters and the one with the lowest pairwise alignment score will be placed closer to the root.

Chapter 5

DATASET

The dataset for this experiment consisted of various mitochondrial genome sequences. The mitochondrial genome is the genetic material inside the mitochondria which is found in eukaryotic organisms. All plants, animals, fungi and protists are eukaryotic organisms [6].

Mitochondrial genome sequences were used mainly because of their size. They generally only contain 16,000 to 20,000 base pairs whereas the full human genome contains approximately 3 billion base pairs [7]. Since aligning sequences is computationally expensive, aligning entire genome sequences would not be feasible.

The mitochondrial genome sequences were gathered from a database of sequences hosted by the University of Montreal [3]. All of these sequences were converted to the FASTA format, which is a format that can be used as input in ClustalW [11]. Table 5.1 shows the species that were used in this study.

Table 5.1: Species used for testing.

Species	Common name	Name used in test
Acanthaster brevispinus	Starfish	Starfish
Aurelia aurita	Moon Jelly	Jellyfish
Boa constrictor	Boa	Snake2
Bos taurus	Cattle	Cattle
Candida glabrata	Yeast	Yeast
Candida zemplinina	Yeast	Yeast2
Cebus albifrons	Capuchin Monkey	Monkey
Crocodylus niloticus	Crocodile	Crocodile
Falco peregrinus	Falcon	Bird
Homo sapiens	Human	Human
Iguana iguana	Iguana	Iguana
Lacerta viridis viridis	Green Lizard	Lizard
Megaptera novaeangliae	Humpback Whale	Whale
Micropterus salmoides	Largemouth Bass	Fish
Strigops habroptilus	Kakapo	Bird2
Tetraodon nigroviridis	Puffer Fish	Fish2
Xenopeltis unicolor	Sunbeam Snake	Snake

Having a known phylogenetic tree is also important for testing. This means that the tree is believed to accurately show the evolutionary relationships between the various species included in the tree. All of the species listed in Table 5.1 were chosen because they are in the known tree. Figure 2.1 illustrates the tree that is used as the known tree for this project [13].

Chapter 6

ANALYSIS

The first step in comparing the original ClustalW algorithm to the modified version was selecting sequences to align. Using a subset of the sequences in Table 5.1, ten test cases were devised. Each test consisted of performing a full alignment on the sequences using the original version of ClustalW and the modified version of ClustalW with all the default settings. Then ClustalW and three programs within the PHYLIP package were used to infer phylogenetic trees using the aligned sequence as input [5].

After the trees were constructed they needed to be compared to the known tree to see which was more accurate. In order to do this, a pairwise comparison of phylogenies was performed to determine the similarity between two trees. This algorithm works by pairing up each branch in one tree with a matching branch in the second tree. Then it finds the optimum 1-to-1 map between branches in the two trees in terms of a topological score [10, 17].

As input this method takes two phylogenetic trees. One tree would be the known tree for the selected species, which is based on the tree in Figure 5.1. The second tree would either be the tree produced using an unmodified version of ClustalW or the tree produced using the modified version of ClustalW. This similarity method produces a value between 0 and 100. A score of 100 would indicate that the two trees being compared are identical.

To see which version of the ClustalW algorithm performed better, the similarity scores were compared. Table 6.1 describes the species being used in each test and which species in the set are considered the highly divergent species.

Table 6.1: Species used for each test.

Test Number	Species Used	Highly Divergent Species
1	snake, lizard, iguana, crocodile, bird, whale, cow, human, monkey, fish, yeast	Yeast
2	snake, lizard, iguana, crocodile, bird, whale, cow, human, monkey, fish, jellyfish	Jellyfish
3	iguana, crocodile, whale, human, fish, yeast	Yeast
4	snake, lizard, iguana, whale, cow, human, monkey, fish, yeast	Yeast
5	snake, lizard, iguana, crocodile, bird, human, monkey, fish, yeast	Yeast
6	snake, lizard, iguana, crocodile, bird, whale, cow, human, monkey, fish, yeast, yeast2	Yeast, Yeast2
7	lizard, iguana, crocodile, bird, whale, cow, human, monkey, fish, yeast, yeast2, jellyfish	Yeast, Yeast2, Jellyfish
8	snake, lizard, iguana, crocodile, bird, whale, cow, human, monkey, fish, yeast, starfish, jellyfish	Yeast, Jellyfish
9	snake, snake2, lizard, iguana, crocodile, bird, bird2, whale, cow, human, monkey, fish, fish2, yeast, yeast2	Yeast, Yeast2
10	fish, snake, iguana, crocodile, whale, human, jellyfish	Jellyfish

Table 6.2 shows the test results when ClustalW was used to infer the phylogenetic tree. The original ClustalW column shows the similarity between the known phylogenetic tree and the phylogenetic tree produced using the original, unmodified version of ClustalW to perform the multiple sequence alignment. The modified ClustalW column shows the similarity between the known phylogenetic

tree and the phylogenetic tree produced using the version of ClustalW that contains the modifications proposed in this paper. A positive value in the difference column shows that the modified version provided an alignment that could be used to construct a more accurate phylogenetic tree. A value of zero indicates that both the original and modified version of ClustalW produced the same phylogenetic tree. A negative value means that the original ClustalW could be used to calculate a more accurate phylogenetic tree. In every single test, the original ClustalW could not be used to place any of the highly divergent sequences correctly in the phylogenetic tree. So the last column in the table is used to indicate that yes, the modified version of ClustalW produced a multiple sequence alignment that resulted in a phylogenetic tree that accurately placed the highly divergent sequence or no, it did not.

In 7 out of 10 tests, the modified version of ClustalW constructed a more accurate phylogenetic tree. In test 3 both methods produced the exact same tree. In tests 5 and 8, the original version of ClustalW provided a more accurate tree. Besides looking at just the similarity score it is interesting to see where the most divergent sequences were placed within the evolutionary tree.

In tests 2, 5, 6, 9 and 10, the highly divergent sequences were correctly placed in the tree using the modified version of ClustalW. In contrast, the original version of ClustalW placed none of them correctly. In many instances, the original version placed them far away from the root.

Table 6.2: ClustalW results (H.D.S. stands for Highly Divergent Sequence)

	ClustalW			
Test	Original ClustalW	Modified ClustalW	Difference	H.D.S placed correctly?
1	72.9%	81.9%	9.0%	No
2	70.8%	76.0%	5.2%	Yes
3	55.6%	55.6%	0.0%	No
4	70.8%	73.6%	2.8%	No
5	70.8%	68.1%	-2.7%	Yes
6	76.1%	80.4%	4.3%	Yes
7	82.2%	87.8%	5.6%	No
8	78.9%	75.1%	-3.8%	No
9	80.0%	83.2%	3.2%	Yes
10	58.3%	77.1%	18.8%	Yes

PHYLIP is a software package that amongst many programs contains seven applications to infer phylogenetic trees based on DNA sequences. Three of these programs were used to determine if the modified version of ClustalW could outperform ClustalW using a variety of tree building software. The applications used from the PHYLIP package are DNAPars, DNAComp, and DNAMl [5].

DNAPars, DNAComp, and DNAMl use different methods for inferring trees than ClustalW. DNAPars uses parsimony, which is a method that provides the tree with the least number of evolutionary changes. DNAComp uses a compatibility method which is a modification of the algorithm used in DNAPars. The compatibility method is similar to parsimony, but as species are added it can calculate the minimum number of base changes that could be required at a specific site. DNAMl uses a maximum likelihood method and can determine different rates of evolution at different sites [5]. Experts do not rely on one method for

constructing a phylogenetic tree. Therefore, it is useful to compare the results for more than one tool to see if the modified version of ClustalW can construct a more accurate tree using a variety of methods.

The test results for DNAPars can be seen in Table 6.3. Using the same methods as the previous test, the modified version of ClustalW provided nearly the same results as the original version. Each version had 2 instances where they performed better than the other. In 6 of the tests, the trees produced were of the same accuracy as one another. However, it can be seen that the modified version of ClustalW correctly placed the highly divergent sequences in 7 out of the 10 tests, whereas the original placed none of them correctly. This is interesting because even though the difference in accuracy between the different versions is 0, the modified version correctly placed those sequences in each of those trees. This shows that if the difference is zero that they are not necessarily the same tree.

Table 6.3: DNAPars results

	DNAPars			
Test	Original ClustalW	Modified ClustalW	Difference	H.D.S placed correctly?
1	82.3%	82.3%	0.0%	Yes
2	82.3%	89.6%	7.3%	No
3	66.7%	66.7%	0.0%	Yes
4	79.2%	79.2%	0.0%	Yes
5	75.0%	75.0%	0.0%	Yes
6	81.3%	81.3%	0.0%	Yes
7	82.2%	87.6%	5.4%	Yes
8	81.2%	77.1%	-4.1%	No
9	84.4%	84.4%	0.0%	Yes
10	68.8%	58.3%	-10.5%	No

Table 6.4 shows the test results when DNAComp was used to prepare the evolutionary trees. In this set of tests, the modified version of ClustalW always provided a tree of equal difference or a more accurate tree. In tests 2, 5, and 6 this difference was an impressive high double digit difference which indicates that there was a significant improvement. This method did not perform as well as the previous two methods in correctly placing the highly divergent sequences. The new method of ClustalW only helped correctly place the highly divergent sequences in 3 out of the 10 tests, which is better than the original version since it did not place any of them correctly.

Table 6.4: DNAComp results

	DNAComp			
Test	Original ClustalW	Modified ClustalW	Difference	H.D.S placed correctly?
1	57.9%	69.8%	11.9%	No
2	69.8%	87.5%	17.7%	No
3	66.7%	66.7%	0.0%	No
4	79.2%	79.2%	0.0%	No
5	59.2%	75.0%	15.8%	No
6	57.4%	81.3%	23.9%	Yes
7	71.1%	71.1%	0.0%	Yes
8	66.6%	78.0%	11.4%	No
9	76.1%	76.1%	0.0%	Yes
10	68.8%	77.1%	8.3%	No

Table 6.5 shows the test results when DNAMl was used to construct the phylogenetic trees. In tests 6 and 8, the original version of ClustalW produced a more accurate tree. In tests 1, 2, 4, 5, 9 and 10, the modified version of ClustalW did better. Like the previous test, the modified version correctly placed the highly

divergent sequences in only 3 of the tests and the original version placed none correctly.

Table 6.5: DNAMl results

	DNAMl			
Test	Original ClustalW	Modified ClustalW	Difference	H.D.S placed correctly?
1	77.1%	80.9%	3.8%	No
2	79.2%	80.9%	1.7%	No
3	66.7%	66.7%	0.0%	No
4	72.2%	79.8%	7.6%	No
5	68.1%	74.2%	6.1%	No
6	84.4%	72.6%	-11.8%	Yes
7	87.6%	87.6%	0.0%	Yes
8	93.3%	72.5%	-20.8%	No
9	82.6%	87.0%	4.4%	Yes
10	58.3%	81.2%	22.9%	No

By comparing the accuracy and the placement of highly divergent sequences, the modified version of ClustalW does show a significant improvement. Out of the combined 40 tests, the modified version correctly placed the highly divergent sequences in 18 tests compared to the original versions 0. In addition, the modified version led to a more accurate tree in 21 tests, a tree of the same similarity in 13 tests and a worse tree in 6 tests. Using a variety of programs to infer the tree shows that this new approach is not dependent on one phylogenetic method for positive results.

Although an increase of a few percent may not seem like a lot, it is important to consider the overall accuracy of the tree. If the accuracy is in the 70^{th} to 80^{th} percentile, then an increase of 5% or more is a fairly good improvement. This new method also provides good results for a variety of test cases. The highly

divergent sequences were varied, as was the number of other sequences. Since the new method did not outperform the original in every test, there is no guarantee that it will always lead to a better tree. To get the best results, the user should use a variety of methods and interpret the results to determine which alignment is the best for their situation.

Chapter 7

DISCUSSION

ClustalW already has a couple of features implemented to deal with divergent sequences. The first feature delays the alignment of divergent sequences until the more similar sequences are aligned first. This may give a better chance of correctly placing gaps within the alignment. This approach is similar to the modified version of ClustalW presented in this paper, but the implementation is different. The modified version guarantees that the highly divergent sequences are aligned last whereas the method provided by ClustalW does not. The test results show that the original ClustalW was not able to properly place any of the divergent sequences, but the modified version was able to in approximately half of the tests.

The second feature ClustalW offers is sequence weights, which are calculated directly from the guide tree. Closely related sequences will receive low weights and highly divergent sequences will receive high weights. These weights are then used for scoring during the final alignment step. The purpose is to try and eliminate scoring bias for sequences that are very similar [8]. One problem with this approach is that the weights are based on the guide tree. So if the clustering algorithm provides bad results then the guide tree could calculate incorrect weights.

Similar research was conducted by Vescovo, Aude, and Polaillon to show that improvements to guide tree construction influence alignment accuracy. Three different clustering methods outperformed the Neighbor-Joining, which is the algorithm implemented in ClustalW. These methods were considered to be better

because they produced guide trees that were different from ClustalW and those new guide trees increased the accuracy of the multiple sequence alignment [16]. Their results support the findings of this project because it shows that the guide tree impacts the accuracy of the final alignment and that there is room for improvement in the current implementation of ClustalW.

Of course the results presented in this study cannot be considered as definitive. They would require a much larger test set. However the improvement trend is undeniable and encourages pursuing this investigation further.

Chapter 8

FUTURE WORK

ClustalW is not the only progressive alignment program available. Work could be done to compare the results of ClustalW with other programs such as T-Coffee to see what types of differences exist [9]. This could be useful in potentially determining if one program is better suited for a specific type of dataset.

There are other approaches to solving the multiple sequence alignment problem besides using a progressive alignment method. Hidden Markov models, iterative methods and genetic algorithms are just a few different methods currently being used to try and find better alignments. Future work could include researching these methods to compare the advantages and disadvantages with programs like ClustalW.

It is also important to look at the software used to infer the phylogenetic trees. There are many different methods for constructing the tree based on the multiple sequence alignment. Modifications to these methods could yield better results as well. Since the trees are based on genetic data there are important limitations to consider since there is still a lot that remains to be known about genetic sequences. As more knowledge is gained about genetic sequences, this knowledge should be valuable to phylogenetics [4].

Chapter 9

EDUCATIONAL STATEMENT

9.1 Graduate Work Contribution

This research thesis helped build on my graduate coursework in TCSS 588 Bioinformatics by allowing me to study genomics in more depth. I was able to utilize the skills from this class in order to understand the problem domain. Furthermore, I was able to use the knowledge I gained in TCSS 543 Advanced Algorithms to analyze how the ClustalW algorithm worked and how to make improvement without sacrificing efficiency. Lastly, using the skills I gained in the TCSS 598 Master's Seminar class I was able to conduct research that assisted in achieving my project goal.

9.2 New Learning

This project allowed me to explore phylogenetics, which was an area of science that interested me, but I had no previous experience in. I was able to research the subject domain and see what kinds of problems exist. I gained experience in using some of the current tools available to biologists. Overall I was able to improve my research and writing skills. Being able to research a topic of my own interest was the reason I chose to attend graduate school. Now that this experience is over I am very grateful that I was able to find a topic that I cared about. It makes this type of work so much more fun and rewarding.

Chapter 10

CONCLUSION

This thesis has proposed an improvement to ClustalW sequence alignment algorithm that enables the construction of a more accurate tree when highly divergent sequences are present. In the majority of the tests performed, the modified version of ClustalW produced more accurate trees than the original version. It was also able to correctly place the highly divergent sequences in nearly half of the tests. This shows that the modified version of ClustalW is an improvement. The results are encouraging and mandate testing it on larger test sets. However, like all current methods for constructing evolutionary trees, this method does not ensure the correct phylogenetic tree will be produced. In order to get the best results it is important for the user to have some expert knowledge so that they can interpret the results and adjust parameters within the program to get the best phylogenetic tree.

BIBLIOGRAPHY

[1] Bichindaritz, I., Potter, S., and S.F.S. “Knowledge-Based Phylogenetic Classification Mining”, Industrial Data Mining Conference, Perner, P. (Edt.), Leipzig, SPRINGER-VERLAG Lectures Notes in Artificial Intelligence, 2004 163-172.

[2] Bonizzoni, Paola, and Gianluca Della Vedova. "The Complexity of Multiple Sequence Alignment with SP-Score That is a Metric." Theoretical Computer Science. 259 (2001): 63-79.

[3] University of Montreal. "Complete Mitochondrial Genome Sequences." 23 Oct. 2007. Evolutionary & Integrative Genomics at the Université De Montréal. 6 May 2008 <http://www.bch.umontreal.ca/ogmp/projects/other/mt_list.html>.

[4] Felsenstein, Joseph. Inferring Phylogenies. Sunderland, Massachusetts: Sinauer Associates, Inc., 2004. 496-520.

[5] Felsenstein, Joseph. “PHYLIP Phylogeny Inference Package.” Cladistics 5 (1989):164-166.

[6] Henze, K, and W Martin. "Evolutionary Biology: Essence of Mitochondria." Nature 426 (2003): 172-176.

[7] International Human Genome Sequencing Consortium. "Finishing the Euchromatic Sequence of the Human Genome." Nature 431 (2004): 931-945.

[8] Jeanmougin, Francois, Julie D. Thompson, Manolo Guoy, Desmond G. Higgins, and Toby J. Gibson. "Multiple Sequence Alignment with Clustal X." Trends in Biochemical Sciences (1998): 403-405.

[9] Notredame, Cedric, Desmond G. Higgins, and Jaap Heringa. “T-Coffee: A novel method for fast and accurate multiple sequence alignment.” J Mol Biol 302 (2000): 205-217.

[10] Nye, Tom M. W., Pietro Lio, and Walter R. Gilks. "A Novel Algorithm and Web-Based Tool for Comparing Two Alternative Phylogenetic Trees." Bioinformatics Advance Access (2005).

[11] Pearson, W.R. “Rapid and Sensitive Sequence Comparison with FASTP and FASTA.” Methods in Enzymology 183 (1990):63-98.

[12] Shamir, Ron. "Algorithms in Molecular Biology." Tel Aviv University School of Computer Science. Fall 2001. Tel Aviv University.

[13] Theobald, Douglas L. "29+ Evidences for Macroevolution." TalkOrigins. Department of Biochemistry, Brandeis University. 6 May 2008 <http://www.talkorigins.org/>.

[14] Thompson, Julie D., Desmond G. Higgins, and Toby J. Gibson. "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research 22 (1994): 4673-4680.

[15] Tree of Life Project. "What is Phylogeny?" Tree of Life Web Project. 6 May 2008 <http://tolweb.org>.

[16] Vescovo, Laure, Jean-Christophe Aude, and Geraldine Polaillon. "Guide structure calculation: a critical step for the accuracy of progressive multiple sequence alignment algorithms." Bioinformatics (2005): 1-2.

[17] Robinson, D.F., and Foulds, L.R. "Comparison of phylogenetic trees." Mathematical Biosciences 53 (1981): 131–147.