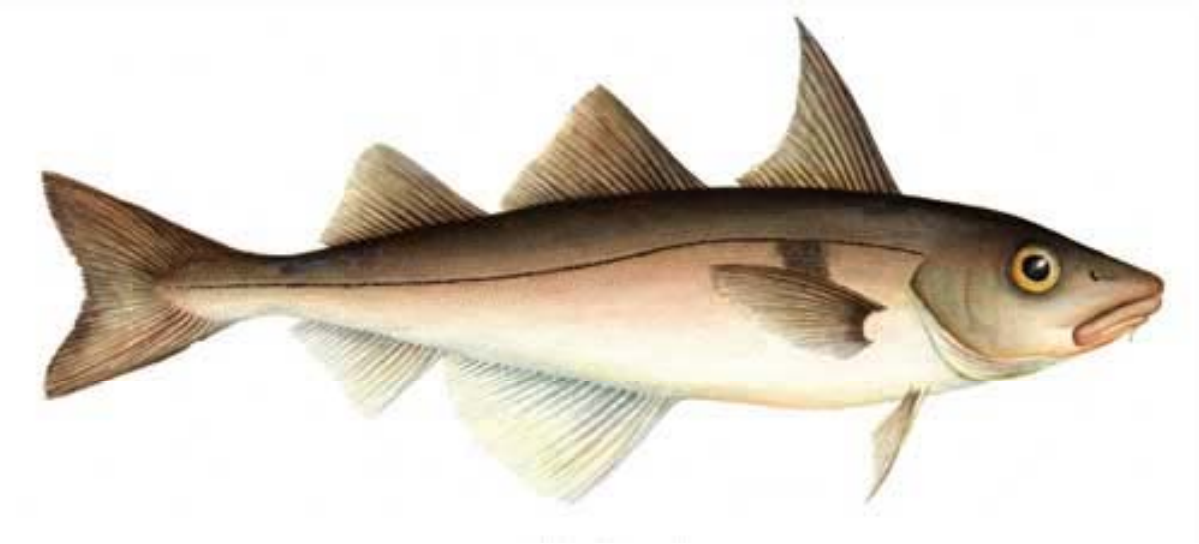
Project in Bioinformatics
Bio 2304 Introduction to Bioinformatics
Institute of Medical Biology
Department of Molecular Biotechnology
University of Tromsø
Autumn 2005

The mysterious Haddock-sequence



By

Gr. 4

ABBREVIATIONS

A	
A	Adenine
B	
BLAST	Basic Local Alignment Search Tool
C	
C	Cytosine
D	
DNA	Deoxy Ribonucleic Acid
G	
G	Guanine
J	
JC	Jukes- Cantor
M	
ME	Minimum Evolution
MEGA	Molecular Evolutionary Genetics Analysis
MP	Maximum Parsimony
N	
NCBI	National Centre for Biotechnology Information
ND2	NADH- dehydrogenase subunit 2
NJ	Neighbour- joining
O	
ORF	Open Reading Frame
OTU	Operation Taxonomic Units
P	
PAM	Point Accepted Mutation
PD	Proportion- Distance (p- distance)
PDB	Protein Data Bank
T	
T	Thymine
U	
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
V	
VS	Versus

CONTENTS

Introduction …….5

Aim …….6

Materials and Methods …….7
- BLAST …….7
- CLUSTAL algorithm …….8
- MEGA algorithm …….9
- Maximum Parsimony …….9
- Bootstrapping ……10
- Minimum Evolution ……10
- Neighbor-Joining ……10
- Jukes-Cantor ……11
- PubMed ……11
- Swiss-PDB Viewer ……13

Results ……13
- Trees constructed from sequence-set 1 ……15
- Trees constructed from sequence-set 2 ……18

Discussion ……20

Conclusion ……22

References ……23

SUMMARY

The aim of this project was to find out as much as possible about a DNA-sequence that we, in advance, knew nothing about, by the use of different available data-resources. The assignment instructed us to put emphasis on the relationship between this sequence and other more, or less, related sequences. By inserting the sequence in question as a query in the BLASTn[1] database, we obtained a number of so-called "hits". These hits represented other isolated DNA-sequences with a certain degree of similarity to the sequence under study.

After selecting two assortments of sequences of varying similarity we constructed multiple alignments by the use of the alignment-program Clustal X[2]. In order to be able to construct a fairly reliable phylogenetic tree on the basis of these alignments, we were forced to manually manipulate them so that the gaps introduced by the program correlated with the reading-frame. With this being done, we used the program Mega2[3] to estimate distances and construct phylogenetic trees using different tree-building methods.

The results showed that some of our expectations were correct. For instance, for one of the alignments, four of the sequences were grouped together in all the trees constructed, and the branches were assigned very high significance-values when confidence-tests were applied. These sequences originated from species we knew to be closely related in advance. The rest of the sequences did not show a uniform relatedness, and also did not score high on the significance-tests. Therefore it was difficult to infer anything about the evolutionary paths taken by these sequences.

INTRODUCTION

When an entire genome, or a smaller DNA- sequence has been sequenced, half of the job is yet to be done. Even when the exact content of bases in the sequence is known, little is still known about its function. To get this knowledge it's necessary to spend time in the laboratory.

By using bioinformatic tools before starting the work in the laboratory, you can get a clue about the function of the sequence by comparing it to related sequences, and thereby save a lot of time on experiments. There are many sequence-databases where you can search for similar sequences, and find out about their function and evolutionary relationships.

To be able to compare the unknown sequence with other known sequences, you need to use different search tools. These search tools usually consist of one or more algorithms. Some algorithms use global alignments and others use local alignments. Some take into consideration reversions, while others do not. The result of the search depends on the algorithm being used.

When sequences related to the one under study are found, their entries often contain information about what kind of protein they code for, and they thereby give the researcher a clue about the protein encoded by the unknown sequence.

If one wishes to investigate the relationships between some of these sequences,
the next step is to choose sequences of interest to make a multiple alignment, which can further be used to construct phylogenetic trees.

Phylogeny is the evolutionary history of a group of species. By making a phylogenetic tree, the species are set in connection with each other. There are many methods that can be used to make phylogenetic trees.

AIM

The aim of this study was to find out as much as possible about the sequence[4] below, including an investigation of the relationship between this and other, closely or distantly related sequences that we selected among several BLASTn-results. The only information given us about the sequence in advance was that it was isolated from the fish *Melanogrammus aeglefinus*, or Haddock, which is its common name.

```
ATGAACCCCTTTATTCTCTCTATCCTCTTATTAAGCCTGGGCTTAGGCACAACGCTAACCTTTGCAAGCTCTCATTGGCT
ATTAGCCTGAATAGGCCTAGAGATTAGTACCCTAGCTATTATTCCTCTAATATCACAACACCACCATCCGCGAGCAGTAG
AGGCTACAACAAAATATTTTATTACACAAGCTGCAGCAGCTGCCTTAATTCTATTTGCTAGCACCACAAATGCATGATTT
ACAGGACAATGAGACATTAACTTTAACCTTCATTTTTTCCCAGCTTCTATACTCACAATAGCTTTAGCCCTAAAAATAGG
TCTTGCCCCAGTACACTTCTGATTACCAGAAGTGCTTCAGGGACTTGACCTAACAACAGGACTAATTTTATCTACTTGAC
AGAAGCTGGCACCTTTTATCTTAATATGTCAAATTATGCCAGTGAATTCTAACTTAATTACTTTTCTGGGTGTCACATCA
ACGCTCGTAGGAGGTTGAGGAGGATTAAACCAAACCCAATTACGTAAAATTTTAGCCTACTCATCTATTGCCCACCTTGG
TTGAATAATTCTTGTTATACAATTTAATCAACAATTAGCCCTTCTAGCTTTAATTATTTACATCCCTATAACTTTCTCAA
CTTTCATAATTTTTAAAACTAATTCTTCTACCACAGTGAACACATTAGCTGCCTCCTGATCTAAAACCCCTGCCCTAACA
GCAATCACCCCTTTAATCCTTCTCTCTTTAGGAGGCCTGCCCCCTCTTTCTGGGTTCATACCCAAATGAATAATTCTCCA
AGAATTAACGAAGCAGGATATCCCTCTAACTGCCTCAATTGCTGCATTAAGCGCACTATTAAGTCTTTATTTTTATCTAC
GCGTGTCCTACGCAATAACTTTAACCATTTCACCTAATAATCTTAATGCAACAACCCCTTGACGACTGCAAACAACAGCA
TCCACTTTACCTCTCGCTATTTCAGCAACAATTTCTGCTATACTTTTGCCCCTGGCCCCCGCAACCTTGGCTTTATTGTC
TCTTT
```

Questions that we wanted to find answers to were:

- Does the sequence contain any open reading-frames? I.e. is the sequence protein-coding?
- What protein-product, if any, does the sequence code for?
- Assuming that the sequence indeed is protein coding, which protein family does it belong to?
- Are there any conserved domains in this and similar sequences?
- Can we say anything about the reliability of the phylogenetic trees we construct on the basis of our sequence-alignments?
- Do the selected sequences show the same relationship when aligned both on the basis of nucleotide- and amino-acid-composition?

Our hypothesis was that sequences originating from closely related species would show a greater degree of relatedness than sequences from more distantly related species.

MATERIALS AND METHODS

Starting with the assigned sequence, we used several known computer programs and databases to find out information from the DNA- sequence, then aligned it with other sequences from several species and finally constructed several phylogeny trees using different tree-building programs.

We started with a BLASTn search, for finding sequences, which showed similarity to our given sequence.

To find sequences of similarity to a sequence under study, a very useful tool is **BLAST.**
The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
To know which BLAST program to use, there is a constructed table, which is called the "BLAST Program Selection Guide". It is divided into several categories according to the nature and size of the query and the primary goal of the search. Starting from the query sequence on the left and cross-referencing to the right, a user will arrive at the specific BLAST program best suited for that search.
Several variations of the BLAST algorithm are available for searching protein or nucleotide sequence databases using protein or nucleotide query sequences.
The working process for BLAST is like this:
To search a large database efficiently, i.e. BLASTn first breaks down the query sequence into words, or subsequences of a fixed length (4 is the default word length). All possible words in the query sequence are calculated by sliding a window equal in size to the word length over the query sequence. And once all of the words in the query sequence have been determined, words composed mostly of common amino acids will be discarded. The sequences in the database are then searched for occurrences of the search words. Each time a word match is found in the database, the match is extended in both directions from the matching word until the alignment score falls below a given threshold. Since the alignment is ungapped, the extension only involves adding additional residues to the matching region and recalculating the score according to the scoring matrix.
Numerous sequence alignment and database search algorithms have been developed for various specific types of sequence searches. As described above, BLASTn searches for protein sequence databases for polypeptide sequences. Other variation of BLAST, included BLASTN and BLASTX, allow searching of nucleotide sequence databases and translating from nucleotide sequences to protein sequences prior to searching, respectively.

We got several hits, the best hit was the one with the E-value of 0, or the one with the lowest E-value.

The part of the sequences that showed similarity to our query sequence turned out to be coding regions (ORFs), coding for the protein NADH-dehydrogenase subunit 2 (ND2). Next we identified this part of each sequence, and then copied it. We chose different sequences of varying resemblance to the query-sequence

The sequences in the set came from these species:

Melanogrammus aeglefinus = Haddock
Merlangius merlangus = Whiting
Gadus morhua= Atlantic cod
Theragra chalcogramma = Alaska Pollock
Lota lota = Burbot
Polymixia japonica = Silver eye
Zenion japonicum = Japanese zenion
Ditrema temmincki
Anguilla megastoma = Polynesian longfinned eel
Sargocentron rubrum = Redcoat
Sarotherodon linnellii = Blackfin tilapia
Coreoleuciscus splendidus = Korean splendid dace

All these sequences were then collected in one document, which was next used to conduct a multiple alignment. This is a necessary step before constructing phylogenetic trees. Using CLUSTALX or BioEdit5 could do this, but since the BioEdit program only align two and two sequences at a time (e.g. no multiple alignments), we decided on using CLUSTALX.

D.G. Higgins and P.M. Sharp in1988 first described the Clustal algorithm. It begins by aligning closely related sequences and then adds increasingly divergent sequences to produce a complete multiple sequence alignment. First the algorithm constructs a phylogenetic tree to determine the degrees of similarity among the sequences being aligned.
The tool provides the basic information for identification of the conserved sequence regions. Selection of a scoring matrix for alignments can present a significant problem in multiple sequence alignments. Some matrices, such as PAM-1 and BLOSSUM 90 are appropriate for closely related sequences, while other are appropriate for very divergent sequences. They calculate the best match for the selected sequences and line them up so that the similarities and the differences can be seen. Sequences can be aligned in their entire length (global alignment) or only in contained regions (local alignment). The program is due for pairwise and multiple alignments. The Global alignment needs to insert gaps (representing insertion/ deletion) in order to obtain the highest possible similarity scores. The ClustalX is automatic for multiple alignments of DNA and protein
sequence. The tool can provide a tree after running alignment and has adjustable parameters. The goal of a multiple alignment is to align regions of similar structural or functional importance among sequences.

In CLUSTALX, we had to adjust several alignment parameters, such as the pairwise parameters; gap opening [0-100]: 10.00, gap extension [0-100]: 0.10
On multiple parameters; gap opening [0-100]: 10.00, gap extension [0-100]: 0.20
And the output format options should be in CLUSTAL format and FASTA format. After adjusting the values mentioned above, we did the complete alignment.
We aligned both the nucleotide and the amino acid sequences for all the species that we searched for. And when we drew the phylogeny trees, we did that for both the nucleotide and the amino acid sequences.

Programs such as CLUSTALX however, introduce gaps into the alignment in order to obtain the highest possible similarity-score. The problem is that the program does not discriminate between gaps that correlate with the codon reading-frame and those that do not, with the

result being that gaps can span one and a half codon or simply one single nucleotide. This would in reality lead to nonsense mutations, due to a shift of the reading-frame.

By using the editing resources in the MEGA2 program, we could manipulate the alignments by moving gaps so that they correlated with the reading-frame, using the amino acid-sequence of the original alignments as guidelines. Not until this was completed could we estimate the average distances between the sequences, and thereby choose which tree-building methods to use. A JK-estimate of less than 0.05 would imply the use of P-dist or JC substitution-estimation models. If the value exceeded this, there were several considerations to be taken when choosing which model to use:

- If the number of nucleotides was high we should use P-dist or JC.
- If the number of transitions/ transversions and the number of nucleotides was high we should use Kimura or Gamma distance.
- If the number of nucleotides was relatively small, but number of sequences large, we should use P-dist.

When the distances had been calculated, and models choosed, we could move on to constructing trees by using the tree-building facilities of MEGA.

MEGA(Molecular Evolutionary Genetics Analysis) is an integrated tool for automatic and manual sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, and testing evolutionary hypotheses.
MEGA includes the following tree-building and substitution-estimating models: MP[6], ME[6], NJ[6], UPGMA[6]; nucleotides and proteins.

Maximum Parsimony
Parsimony implies that simpler hypotheses are preferable to more complicated ones.
Maximum parsimony is a character-based method that infers a phylogenetic tree by minimizing the total number of evolutionary steps required to explain a given set of data, or in other words by minimizing the total tree length. The steps may be base or amino-acid substitutions for sequence data, or gain and loss events for restriction site data.
The maximum parsimony method searches all possible tree topologies for the optimal (minimal) tree.
Before considering how it is that parsimony is used to infer phylogenetic relationships, it is necessary to consider what sites within a multiple sequence alignment might have useful information content for a parsimony approach. The multiple sequence alignment contains positions that fall into two categories in terms of their information content for a parsimony analysis: those that have information, **informative**, and those that don't, **uninformative**. The definition of an informative site is as follows. A site is informative only when there are at least two different kinds of nucleotides at the site, each of which is represented in at least two of the sequences under study.
Since the Maximum Parsimony method may result in more than one equally parsimonious tree, a consensus tree should be created. Consensus trees are a convenient way to summarise the agreement between two or more trees. A consensus tree is not an optimal tree. It is a summary of how well the original trees agree (i.e., how congruent). A consensus tree is NOT a phylogeny. Any consensus tree may be used as a phylogeny only if it is identical *in topology* to one of the original equally parsimonious trees. Typically a consensus tree for a set of equally parsimonious, minimal length trees, is longer than any of the original set, because a consensus tree is usually less resolved (contains more polytomies). So interpreting a

consensus tree as a phylogeny implies that polytomies contained in it indicate multiple divergences rather than the proper interpretation that they indicate lack of resolution.

Bootstrapping[6]

Bootstrapping is a way of testing the reliability of the dataset. It is the creation of pseudoreplicate datasets by resampling. Bootstrapping allows you to assess whether or not stochastic effects have influenced the distribution of characters. In phylogenetic analyses nonparametric bootstrapping is the most commonly used confidence-testing method. The pseudoreplicate datasets are generated by randomly sampling the original character matrix to create new matrices of the same size as the original. The frequency with which a given branch is found is recorded as the bootstrap proportion. These proportions can be used as a measure of the reliability (within limitations) of individual branches in the optimal tree.
Thus bootstrap analysis:

- Is a statistical method for obtaining an estimate of error
- Is used to examine how often a particular cluster in a tree appears when nucleotides or amino acids are resampled
- Is used to evaluate the reliability of a tree

Minimum evolution

Given an unrooted metric tree for n sequences there are (2n-3) branches, each with length d_i. The sum of these branch lengths is the length L of the tree:

$$L = \sum_{i=1}^{2n-3} e_i$$

Figure 1. The formula of the length L of the tree.

The minimum evolution tree is the tree which minimizes L. This method is similar in spirit to parsimony, discussed above. However the length in this case is computed from the pairwise distances between the sequences rather than from the fit of individual nucleotide sites to a tree.

Neighbor-Joining

Neighbor-joining is a simplified version of the minimum evolution (ME) method, but generates only one final tree.
The method starts with a star-like tree with all species coming off a single central node regardless of their number. Neighbours are then sequentially found that minimize the total length of the branches on the tree. The principal difference between the neighbour-joining methods is the way in which they determine the sum of all branch lengths with each reiteration of the process.
The tree's total branch length can be determined with this formula:
$S_{12} = (1/2(N-2))(\Sigma(d_{1k}+d_{2k})+(1/2)d_{12} + (1/N-2)(\Sigma d_{ij})$
Keppler's formula: $Q_{12}=(N-2)d_{12}- \Sigma d_{1i} - \Sigma d_{2i}$.
Where any pair of species can take positions 1 and 2 in the tree, N is the number of species represented in the distance matrix, k, is an accepted out-group, and d_{ij} is the distance between species *i* and *j*.

All possible pairs of species are considered in each round of the process, and the pairing that makes a tree with the overall smallest branch lengths (the smallest values for S or Q) is grouped so that a new distance matrix can be generated.

Jukes Cantor

Where substitutions are common, there are no guarantees that a particular site had not undergone multiple changes during the course of evolution, as depicted in this figure:

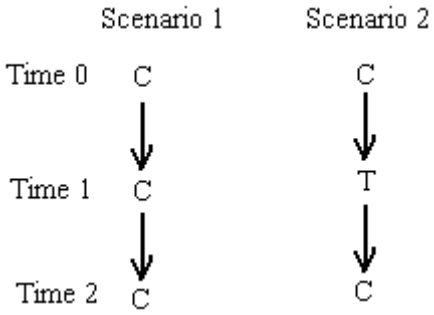


Figure 2. The scenarios of possible changes during the course of evolution.

To address that possibility, Jukes and Cantor assumed that each nucleotide was just as likely to change into any other nucleotide. Using that assumption, they created a mathematical model, in which the rate of change to any one of the three alternative nucleotides was assumed to be α, and the overall rate of substitution for any given nucleotide was 3α.

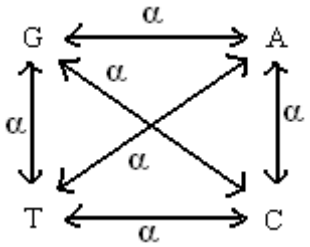


Figure 3. Jukes and Cantor formula of the possibility of changes.

In the model, if a site within a gene was occupied by a C at time 0, then the probability (p) that that site would still be the same nucleotide at time 1 would be $P_{C(1)} = 1-3\alpha$. Because a reversion (back mutation) to C could occur if the original C changed to another nucleotide in that first time span, at time 2 the probability, $P_{C(2)}$, would be equal to $(1-3\alpha)P_{C(1)} + \alpha [1-P_{A(1)}]$.

By using different tree-building methods we then constructed several phylogenetic trees on both the basis of the nucleotide and amino acid sequences.

PubMed

PubMed was developed by the National Centre for Biotechnology Information (NCBI). With this tool you can search for nucleotide and protein sequences, protein structures, complete genomes, taxonomy and many others things. PubMed was designed to provide access to citations from biomedical literature.

Aligning the animal tree.

By searching in PubMed, we also found the NADH dehydrogenase subunit 2 (ND2) sequence of different animals:

Zenion japonicum: Japanese Zenion (fish)
Anguilla Megastoma: Polynesian longfinned eel (fish)
Mustelus manazo: shark
Lyciasalamandra atifi: salamander
Gorilla gorilla: gorilla…
Strongylocentrotus purpuratus: sea urchin
Pupa strigosa: opisthobranch gastropod

We started out doing an alignment based on the nucleic acid composition.
After doing the alignment in ClustalX, it was soon obvious that this would not be possible, because it would mean days of work to align these sequences that were so different.

Next step was to try to use the amino acid sequences to do the alignment, and to leave out two plants first included, because they were much longer then the rest of the sequences. After aligning these amino acid sequences with high parameters in ClustalX, we got a result we could use to make a phylogenetic tree.
For pairwise parameter we used the values: Gap opening: 35,00; Gap extension: 0,75.
For multiple parameter we used the values: Gap opening: 15,00; Gap extension: 0,3.
Pupa strigosa turned out to be quite different from the rest of the organisms. When we tried to remove this from the alignment, we got no difference in the tree, so we decided to keep it in the alignment.

We also tried finding ORFs by using the program BioEdit early in our investigation. BioEdit is a user-friendly, graphically oriented biological sequence editor written for Windows by a graduate student at North Carolina State University. BioEdit is intended to supply a single program that can handle most simple sequence and alignment editing and manipulation functions that researchers are likely to do on a daily basis, as well as a few basic sequences analyses. This way, researchers will save time, not having to use several different programs to execute relatively small tasks. BioEdit is however not the most accurate or powerful alignment-program, and compensates for this by including automatic online links to other Web-based alignment-programs such as BLAST and Clustal W.
Among other features included in the program are:

- In-colour alignment and editing with separate nucleic acid and amino acid colour tables.
- Plasmid drawing interface for automated creation of plasmid vector graphic from a DNA sequence.
- Sort sequences by name, locus, definition, accession, pid/nid, references, comments or by residue frequency in a selected column.
- ORF searching with user-defined preferences.
- Amino acid and nucleotide composition analyses and plots.
- Protein hydrophobicity/ hydrophilicity plots.

To summarise, one might say that BioEdit is a broad-spectered sequence- and alignment editor that offers an easy way to edit and manipulate sequences and alignments for further use in other, more specialized programs.

As it turned out that the sequence originated from the mitochondrial genome, we were forced to cancel this procedure. There are certain differences between the mitochondrial amino acid code and the code for amino acids in the nuclear genome that BioEdit does not allow the user to distinguish between. What we could use BioEdit to find out however was the GC-content of the sequence.

Swiss-PDB Viewer

To find out that the protein had an available structure, we used the Swiss- PDB Viewer to search for the protein. We couldn't find any related structure for the protein, but instead we download the PDB id from the Protein Data Bank, which was also the human NADH-dehydrogenase subunit 2.
Swiss-PdbViewer is an application that provides a user-friendly interface allowing to analyze several proteins at the same time. The proteins can be superimposed in order to deduce structural alignments and compare their active sites or any other relevant parts. Amino acid mutations, H-bonds, angles and distances between atoms are easy to obtain thanks to the intuitive graphic and menu interface. Swiss-PdbViewer is tightly linked to Swiss-Model, an automated homology-modelling server developed within the Swiss Institute of Bioinformatics (SIB) in collaboration between GlaxoSmithKline R&D and the Structural Bioinformatics Group at the Biozentrum in Basel.

RESULTS

The sequence that was investigated turned out to be the coding region of the gene encoding the NADH dehydrogenase subunit 2 (ND2, figure 5 below). This polypeptide is a part of the large NADH-ubiquinone oxidoreductase-complex situated in the inner mitochondrial membrane of vertebrates (also called complex 1, see fig 4 below). These complex catalyses the transfer of two electrons from NADH to ubiquinone, a very important step in the process of conserving energy obtained from oxidation of energy-rich molecules in the form of a proton motive force that can ultimately be used to generate ATP. All hits in the databases searched, showed that the sequence in question is indeed a part of the mitochondrial rather than the nuclear genome.

The GC-content of the sequence is 39,43 %

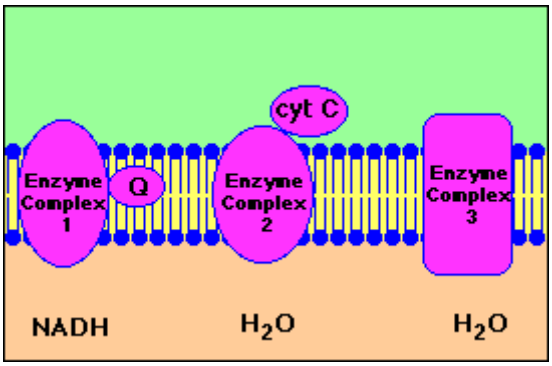
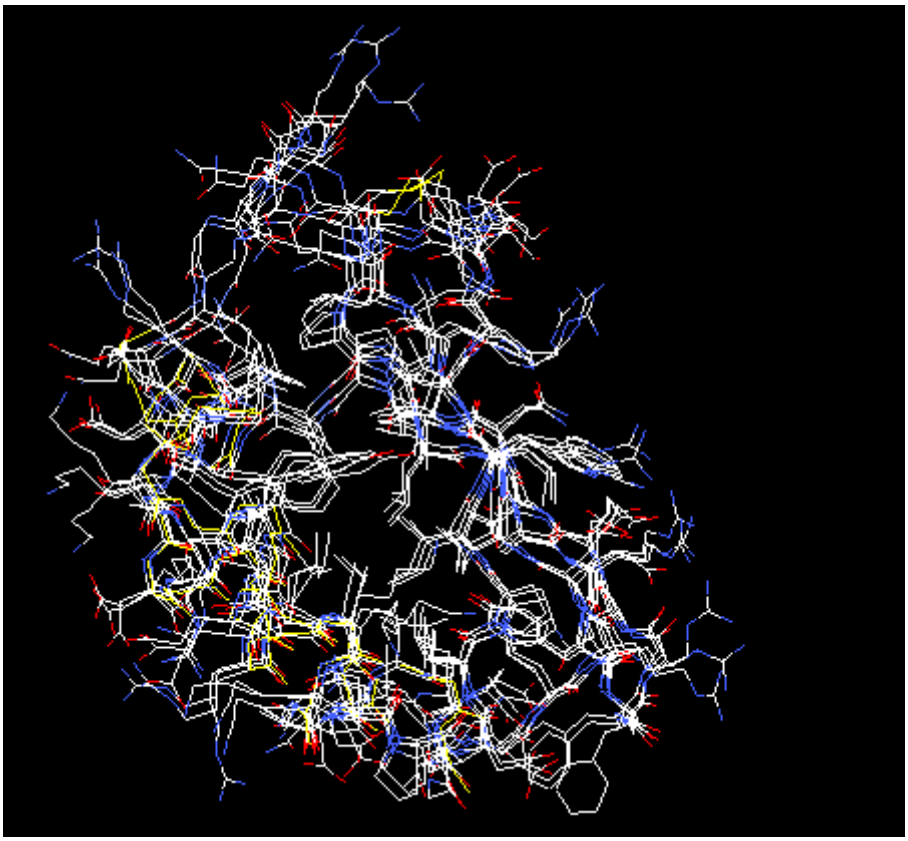


Figure 4: Schematic picture of part of the electron transport chain in vertebrate mitochondria. NADH-ubiquinone oxidoreductase complex is the one at the far left.


Figure 5. Structure of NADH dehydrogenase subunit2 isolated from Homo sapiens.

Trees constructed from sequence-set 1

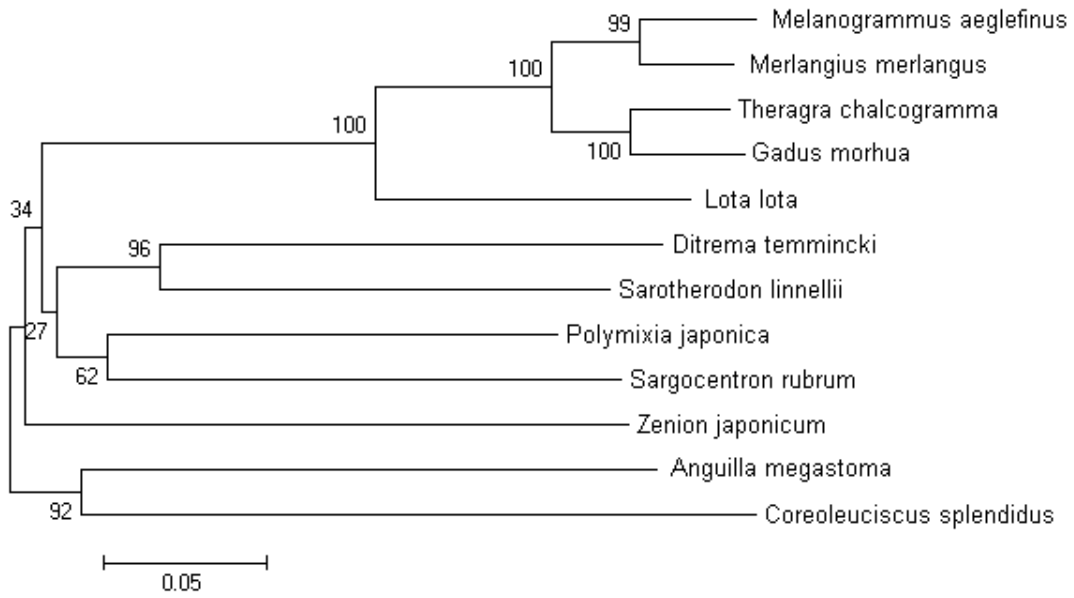


Figure 6: Tree constructed by the NJ-distance method using the JC-substitution estimation model. All three codon-positions are included. A bootstrap test of 1000 resamples is applied.

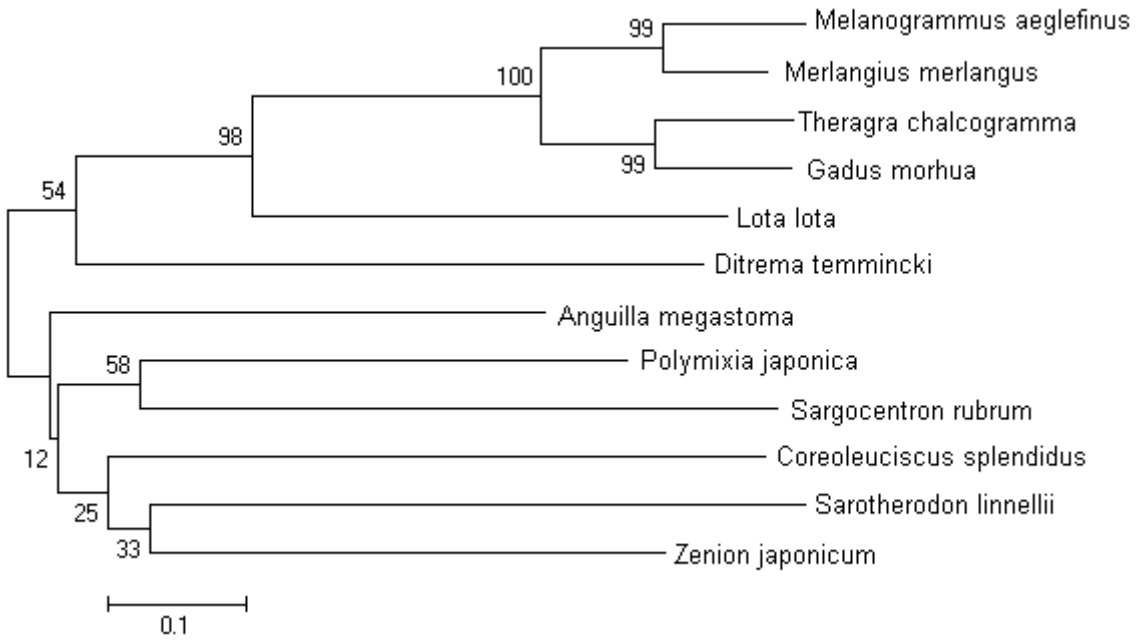


Figure 7: Tree constructed by the NJ-distance method using the JC-substitution estimation model. Only the third codon-position is included. A bootstrap test of 1000 resamples is applied.

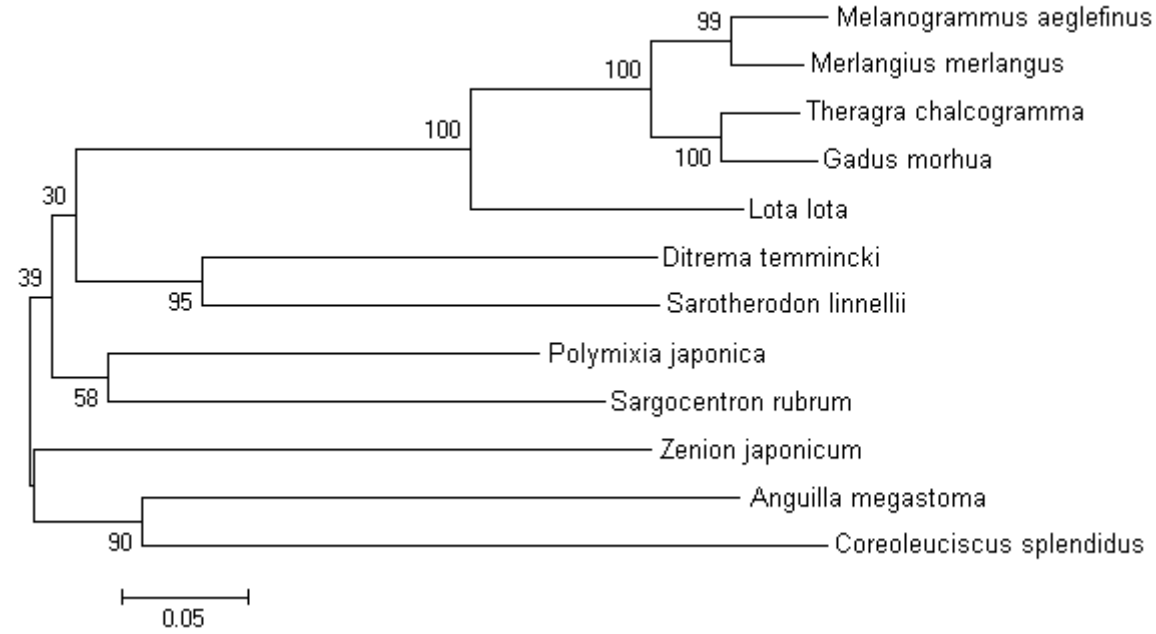


Figure 8: Tree constructed by the ME-distance method, using the JC-substitution estimation model. All three codon-positions are included. A bootstrap test of 1000 resamples is applied.

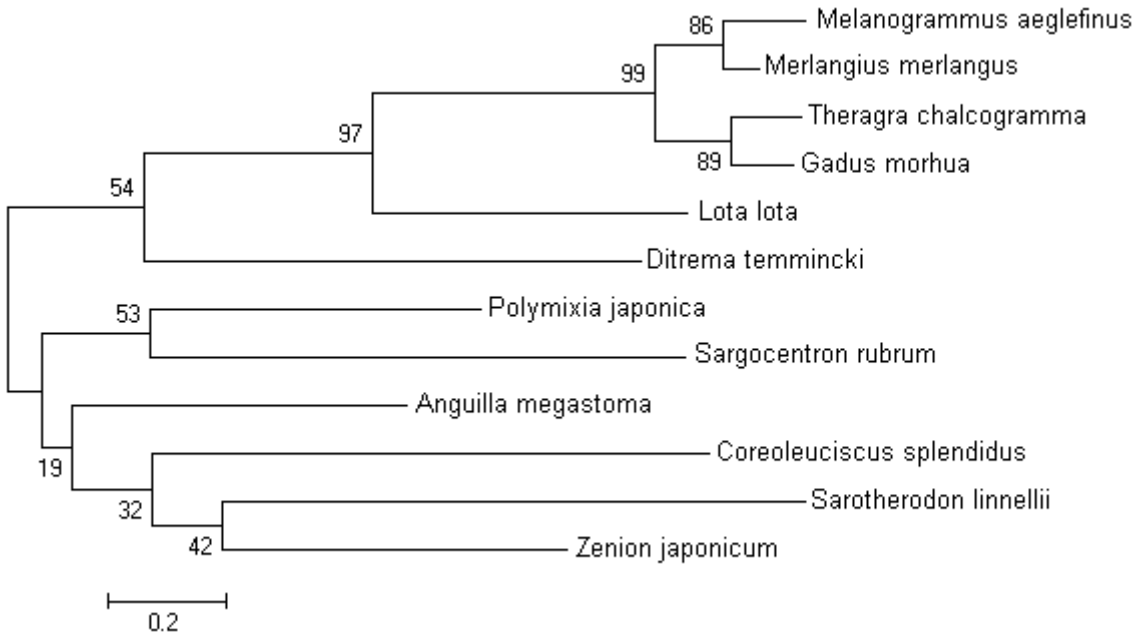


Figure 9: Tree constructed by the ME-distance method using the JC-substitution estimation model. Only the third codon-position is included. A bootstrap test of 1000 resamples is applied.

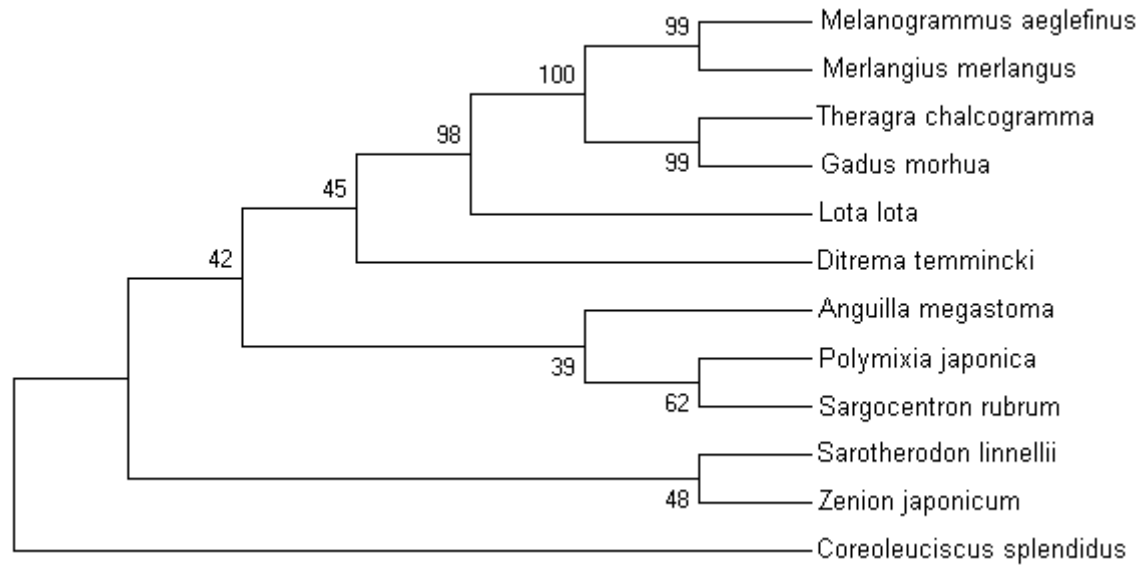


Figure 10: MP-Consensus tree. Bootstrap: 1000 resamples

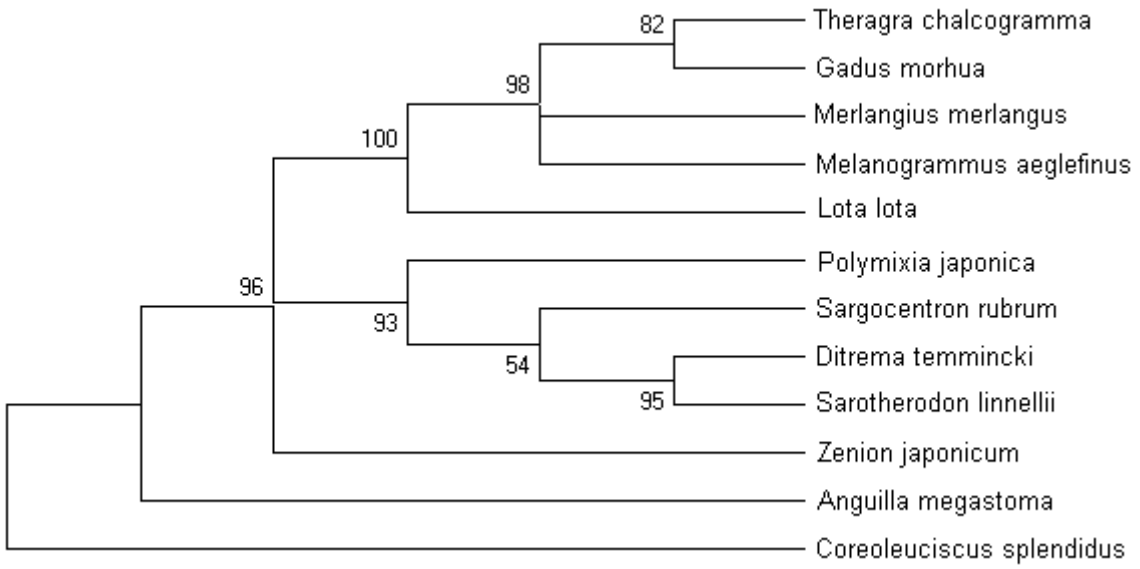


Figure 11: Bootstrap Consensus tree based on the amino acid sequence. Constructed by NJ-distance-method using P-dist. Bootstrap: 1000 resamples.

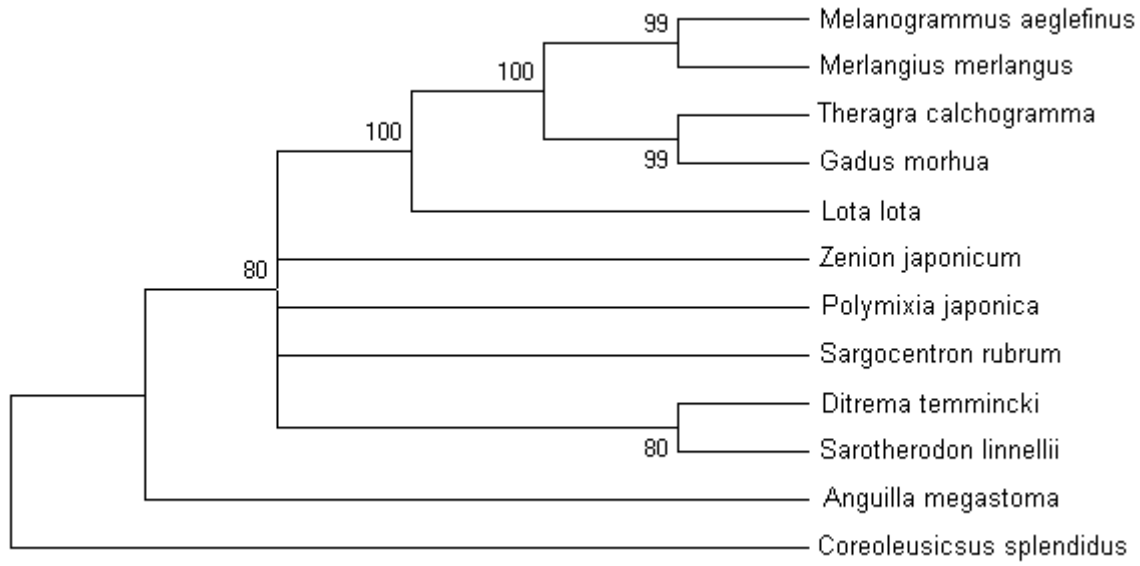


Figure 12: Bootstrap Consensus tree based on the amino acid sequence. Constructed by ME-distance-method using P-dist. Bootstrap: 1000 resamples.

Trees constructed from sequence-set 2

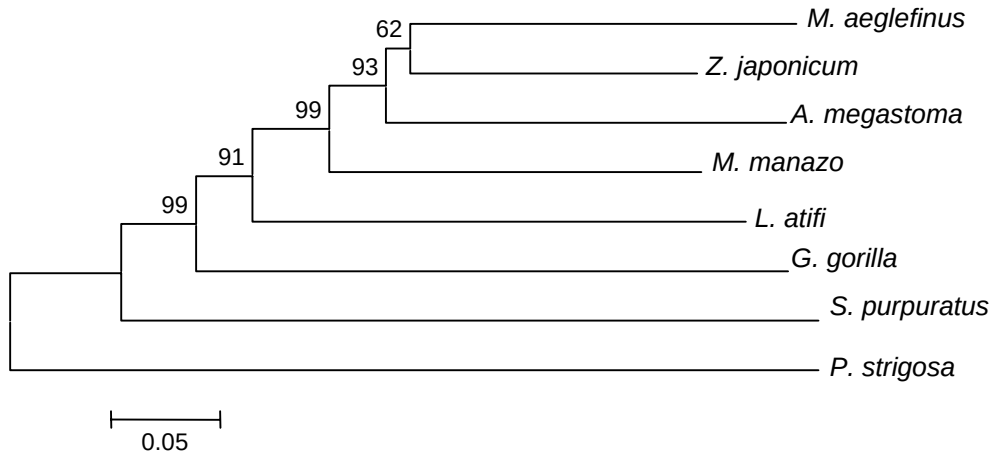


Figure 13: Tree constructed by the NJ- distance method, based on the amino acid sequence. A bootstrap test of 1000 resamples is applied

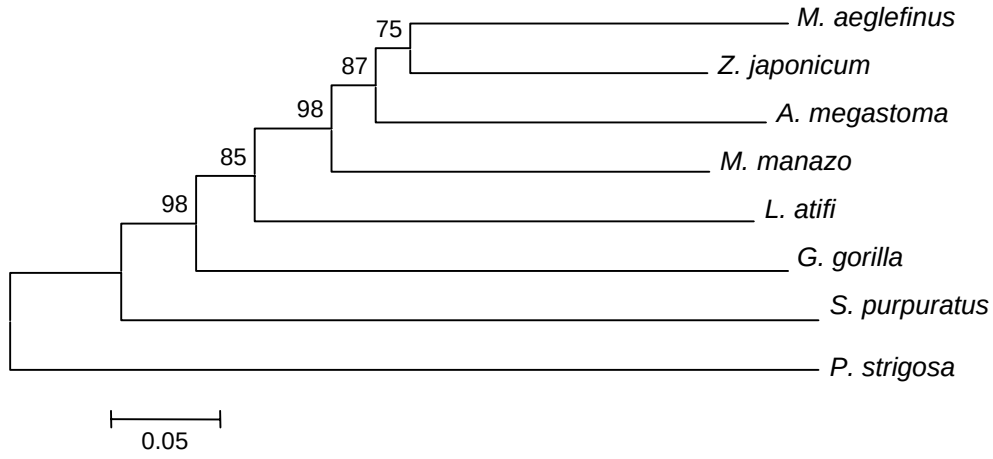


Figure 14: Tree constructed by the ME-distance-method, based on the amino acid sequence. A bootstrap test is applied with 1000 resamples.

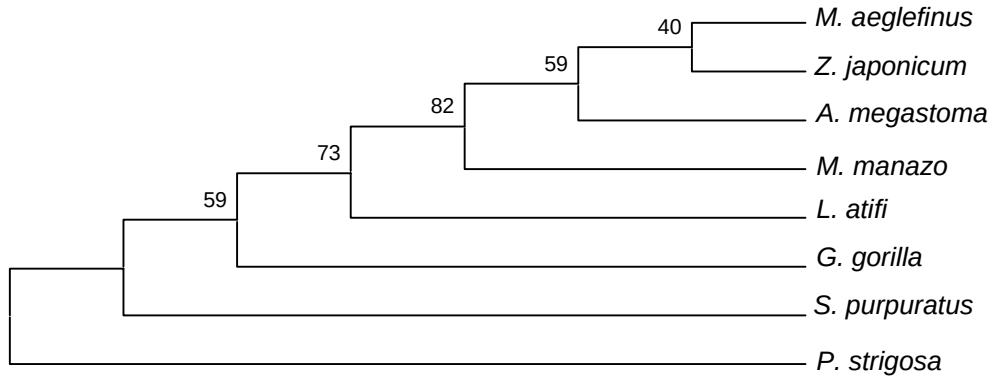


Figure 15: Tree constructed by the MP-distance method, based on the amino acid sequence. A bootstrap test of 1000 resamples is applied.

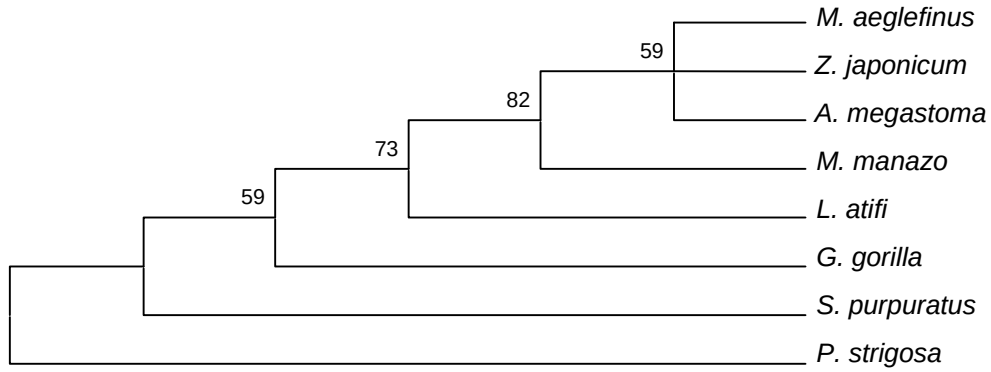


Figure 16: Bootstrap Consensus tree based on the amino acid sequence. Constructed by MP-distance-method. A bootstrap test of 1000 resamples is applied.

DISCUSSION

A phylogenetic tree is never the "true answer" to how the evolution has worked over the years. It can, at the best, give a reasonably good picture of how the course of evolution *most likely* has been. All phylogenetic trees constructed by computers (with the aid of the researcher), are thus only *inferred trees*.

Sequence- set 1.

When crating a distance-matrice by using JC in MEGA, we obtained an average distance-value of 0,375. This value is above the threshold-value for using P-dist as substitution-estimation model when building trees (e.g. our sequences did not collectively show a high enough degree of relatedness to use P-dist). Due to the fact that the sequences in our alignment did not contain a large number of nucleotides, *and* the fact that our alignment did not contain many sequences, we could not automatically follow the guidelines given for tree construction. By discussing the matter with our supervisor, we decided on using JC and ME.

By comparing figures 6 and 7 of the trees constructed on the basis of sequence-set 1, we observe that by including only the last codon-position we obtain lower bootstrap-values for the most related sequences than when all positions are included. This is consistent with the theory that position 1 and 2 are more conserved than position 3, because mutations in this position most often lead to synonymous substitutions (exchange of the nucleotide but no exchange of the amino acid). The closest related sequences will be highly conserved when considering the amino acid content.

For both trees, the bootstrap values among most of the sequences were low. The upper part of the trees was as expected, with high bootstrap values of approximately 100% significance. Which meant that this was the most probable branching to be obtained for these sequences, and that the sequences shared a common ancestor.

The tree-constructions in figures 6, 7, 8, and 9 were based on the Distance method. This might give a true relationship between homologous sequences, for example that the species *M. aeglefinus* was more closely related to *M. merlangus* with a bootstrap value of 99 than compared to *S. linnellii*. The trees were unrooted and had twelve external nodes.

The tree constructed by MP with bootstrap 1000 (figure 10), brought out a bright true graphical of the evolutionary relationship among these species. From this tree we might suggest that the specie *C. splendidus* could be an outgroup. The species *M. aeglefinus*, *M. merlangus*, *T. chalcogramma* and *G. morhua* all belong to one family. The tree had very good bootstrap values, and this supported the observed results.

When using character-based methods, the most parsimonious tree will be the one with the fewest evolutionary changes for all sequences to derive from a common ancestor. This method is a more time-consuming method than the distance methods. Maximum parsimony is perhaps the most popular method for reconstructing ancestral relationships. MP can also lead to the incorrect tree if the amount of the changes is sufficiently divergent in different branches.

Character methods produce more than one tree that is " optimal". For simplicity the program can create a single consensus-tree that summarizes a set of the most parsimonious trees. It is this consensus tree that is depicted in figure 10.

The two trees in figure 11 and 12 constructed from amino acid sequence were very similar, even though one of them was based on NJ, P-dist and the other based on ME, P-dist methods. If we get multiple trees, we can look for the regions that are similar. Those are the regions we can be more confident are correct. The reason we decided to align the sequences based on the amino acid composition was to see if this would lead to significant changes in the relationships between the sequences. Very closely related sequences show a very low degree of differences at the nucleotide-level, and close to none on the amino acid-level. More distantly related sequences may however contain little difference when compared on the amino acid-level, but still be relatively different from each other when compared on the nucleotide-level. This is due to the occurrence of synonymous substitutions. As for the trees described above, these trees did show the same kind of relationship between the topmost sequences, and varying relationships among the rest.

The obtained results for five of the sequences (from *M. aeglefinus, M. merlangius, T. calchogramma, G. morhua* and *L. lota*) showed the same relationship when constructing the

trees based on both nucleotide and amino acid sequences. Trees constructed by distance methods (ME and NJ) and character based methods (MP) also showed the same relationship This observation tells us that this part of the tree can be trusted. Whenever Distance methods and character-based methods provide similar trees, one can assume the relationships to be a fairly reliable picture of the true course of evolution.

There are two other fishes that show clear relation with a bootstrap over 50. They are *P. japonica* and *S. rubrum*. Out from the trees we can assume that these are closer related to each other than to any other fishes they are compared to.

Sequence-set 2.

When making this tree, it did not matter what kind of method that was used, all the trees showed the same branching.
P. strigosa was sufficiently different from the other sequences to be defined as an outgroup. This is a seashell, and the simplest organism in the sequence- set.

All the trees have very high bootstrap- values, which indicates that this line of evolution for this sequence was very likely. The only thing that was not for sure, was weather the haddock was more related to *Z. japonicum* or *A. megastoma*. In figure 15, you can see the tree made by the MP method as a consensus tree. In this tree the three fishes are equally branched.
The assumption is that it is more related to *Z. japonicum*, because this has a bootstrap- value higher than 50 in all but one methods. *A. megastoma* is an eel, and at least from a morphologic point of view, this fish is more different to the haddock then the Japanese zenion. This also means it is highly possible that the sequence of the eel has had a slightly different evolution than the other two fishes. Looking at the phylogenetic trees from sequence- set 1 neither weaken, nor confirms this assumption.

The reason why the Gorilla leaves the root of the tree so early, may be because it separated from the other organisms early in the evolution, and that its sequence is very different from the sequence of the fish.
P. strigosa and *S. purpuratus* were the most primitive organisms here, and they separated from the tree first. The shark seems to be the closest relative to the fishes, as we would expect.

CONCLUSION

With all the tree-building methods used to provide trees from the alignment based on sequence set 1, we got a very high bootstrap value for four of the species in the same family, and a fifth fish that belongs to the same order. This is consistent with our assumptions.

The sequences from the rest of the species were not close relatives, and in reality these species live in different oceans like the Pacific Ocean and the Atlantic Ocean. This is why the branching of the trees differed a lot after what kind of tree- building method we used. It is clearly not possible to give a reliable assumption about relation here.
All trees constructed from the second alignment obtained high bootstrap-values. With trees having these high bootstrap-values, and all the trees showing the same line of evolution, it is reliable.

REFERENCES

1. http://www.ncbi.nlm.nih.gov/BLAST/
2. http://clustalw.genome.ad.jp/ CLUSTALX, school database program.
3. http://www.megasoftware.net/ MEGA3, a newer version also available for download from this web page. The group used MEGA2 computer program, provided from school.
4. Department of Molecular Biotechnology at University of Tromsø.
5. BioEdit, computer program available from school.
6. Fundamental Concepts of Bioinformatics (2004), Dan E. Krane, Michael L. Raymer.
7. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
8. Swiss PDB-Viewer, computer program from school.