JH Bioinformatics Note – 3

Hymenopellis Revisited, via a Circuitous Route

"It is not difficult to make a Phylogenetic Tree"

Jurrie Hubregtse

February 18, 2021

Hymenopellis gigaspora

Before proceeding — some important definitions

Homologous: two sequences are said to be homologous if they share a common evolutional ancestor. There are no degrees of homology; sequences are either homologous or not. While homology is a qualitative inference (sequences are homologous or not), identity and similarity are quantities that describe the relatedness of sequences. A high degree of similarity is often used to infer homology. Homology is an inference, not an observation (unless we have a time machine).

Homology: is a similarity in a sequence of a protein or nucleic acid or in the structure of an organ that reflects a common evolutionary origin. Molecules or sequences that exhibit homology are referred to as homologs.

Molecular phylogenetics: is the analysis of hereditary molecular differences, mainly in DNA sequences, to gain information on an organism's evolutionary relationships. The result of a molecular phylogenetic analysis is expressed in a phylogenetic tree.

Orthologs: are homologous sequences in different species that arose from a common ancestral gene during speciation.

Paralogs: are homologous genes within the same organism, that is, multiple genes that have evolved from a common ancestor.

Phylogenetic tree (Phylogram): is an evolutionary tree showing an estimate of phylogeny where the length of each branch is proportional to the amount of inferred evolutionary change.

Phylogeny: is the study of the evolutionary development of groups of organisms. The relationships are inferred based on the idea that all life is derived from a common ancestor.

Similarity and Identity are quantities that describe the relatedness between two sequences and are usually given as a percentage value.

Similarity represents the degree of resemblance between two sequences. This value can be obtained either as an alignment score between two sequences, or from a pair of aligned sequences by summing the characters that match exactly with the character mismatches, where various mismatched characters are weighted according to an evolutionary model. The total sum is then divided by the sequence length and a percentage similarity is calculated.

Identity between two aligned sequences is a measure of how well the sequences match up. The identity between two aligned sequences is calculated by summing characters that match exactly (gaps are ignored). This sum is then divided by the sequence length and a percentage is calculated. As the number of mismatches between the two sequences approaches zero, then both similarity and identity approach 100%.

Synapomorphy: is an inherited condition or character state that is directly attributed to the evolutionary process. (Some people consider homologies and synapomorphies to be equivalent.)

Taxon: (plural taxa) A formally named group of organisms/species (e.g. groups of organisms represented by the tips/leaves of a phylogenetic tree).

"In Nature's infinite book of
secrecy a little I can read."

William Shakespeare
Antony and Cleopatra (1623)

1 Introduction

This Note, *Hymenopellis* Revisited via a Circuitous Route, is the last in the series titled "It is not difficult to make a Phylogenetic Tree". In this Note I intend to summarise my heuristic journey, which I started about a year ago, into the making of phylogenetic trees. The reason for me to undertake this journey was driven by my interest in showing how organisms are related to each other. One such relationship that particularly interested me was that between two very different fungi, mushrooms and baker's yeast: one a gilled mushroom built from hundreds of thousands of filamentous cells, the other unicellular. Another group of fungi that I wanted to revisit belongs to the recently created genus *Hymenopellis* (Petersen and Hughes 2010). The results obtained from this phylogenetic experiment proved to be quite interesting.

I started my heuristic journey by reproducing some phylogenetic experiments I found in the taxonomic literature (see Bioinformatics Notes 1 & 2). In the literature I came across a number of phylogenetic reconstructions where the conclusions drawn from them caused me some conceptual difficulties. Most of these problematic phylogenetic reconstructions involved the use of the ITS nucleotide sequences (see JH Bioinformatics Note – 2 for a more detailed description of the ITS region). I soon realised that I had to go back to basics - I had to clearly understand what constituted an evolutionary based phylogenetic relationship and how to create a tree that represented such a relationship. To do this I needed to make a segue and include two extra sections "Basics of phylogenetic inference" and "Phylogenetic trees – gene based examples" before continuing.

2 Basics of phylogenetic inference

An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein; the second is the acceptance of the mutation by the species as the new predominant form. To be accepted, the new amino acid usually must function in a way similar to the old one: chemical and physical similarities are found between the amino acids that are observed to interchange frequently.

— Margaret Dayhoff (Taken from Pevsner 2015 p. 69)

Before proceeding I must emphasise that gene or nucleotide based phylogenetic trees are not the same as species trees. Gene/nucleotide evolutionary histories often do not coincide with each other and with patterns of evolutionary speciation. In this section some basic principles that underline the reconstruction of a molecular phylogenetic tree will be discussed. The methodology and vast array of software packages that can be used to construct phylogenetic trees will not be discussed here.

2.1 Evolution — genetic drift

The quote from Margaret Dayhoff gives us the basic evolutionary perspective that must be considered when constructing phylogenetic trees (evolutionary relationships). She describes the basic principles involved in the inheritance of new characteristics based on stochastic (random) mutations occurring in an organism's genetic material (DNA), which in this case is the code that describes a protein. I will continue using protein sequences in this study because they are frequently used for the analysis of evolutionary relationships, and they are often preferred for constructing phylogenetic trees.

As implied by Dayhoff's quote, DNA and RNA sequences (genes) are subject to gradual change over evolutionary time, resulting from the incorporation of new homologous characters caused by accepted mutations. The most common types of mutations are substitutions, insertions, and deletions (indels). Not all genetic mutations will be accepted by natural selection. Mutations that are harmful to the organism will be eliminated, whereas mutations that are neutral or beneficial to the organism will not be penalised by natural selection; they can accumulate over time and contribute to the process of genetic drift (this includes the whole genome).

Genetic drift is an ongoing process that never stops and is one of the mechanisms that results in speciation. The DNA that a child inherits from its parent is expressed as the child organism (phenotype). The book of life is recorded in every organism's genome; it is written by the evolutionary process because genetic mutations are recorded in the genome. By applying phylogenetic analysis, we try to read this book of life, and try to make sense of it. Producing a phylogenetic tree is just one way of doing this.

2.2 Constructing a molecular phylogenetic tree

A phylogenetic tree is a special type of tree, where molecular DNA/nucleotide sequences from selected taxa are used to infer their common evolutionary history. The procedure needed to produce a phylogenetic tree requires 3 basic steps: the selection of sequences, their alignment, and finally the phylogenetic relationships illustrated in the form of a tree. These 3 steps will be discussed in more detail below.

Step 1 is to select suitable sequences from which their evolutionary history is to be inferred. There are 2 main prerequisites for the selection of taxa and their gene/nucleotide sequences. (1) All chosen taxa must be homologous (they must all have the same common ancestor). For this purpose the common ancestor of the selected sequences must be recent enough for the sequence

information to have retained sufficient similarity for it to be used in the phylogenetic analysis. (2) The chosen genes must be orthologous. This is not a trivial exercise; violations of the orthology assumption could be caused by using a gene which is a paralog. This could easily occur for a multi-copy gene (a gene which has numerous copies in the genome i.e. the ITS gene region) where there is an instance of the gene which is not identical to the other copies.

Step 2, arguably the most important, is the multi sequence alignment (MSA) of the selected sequences. Generally a MSA by itself does not imply any evolutionary relationship between the sequences. An alignment is just a way to visualise the differences and similarities between sequences. A MSA intended for phylogenetic analysis differs from other forms of MSAs because the goal of its alignment procedure is to identify and align homologous features that have a common evolutionary origin. An evolutionary MSA is a set of inferences about historical events (mutations); it does not need to be optimally aligned with respect to any given scoring function (such as one calculated by an alignment program) but it does need to reflect the evolutionary relationship between the sequences. Once a MSA has been completed, homologous residues (new amino acids or nucleotides caused by historical mutations) are aligned in columns across the length of the sequences. Now the MSA has all the necessary information from which a phylogenetic tree can be inferred.

Unfortunately, to the best of my knowledge there is no alignment program or algorithm that can achieve an evolutionary alignment. But there are MSA approaches that will result in usable evolutionary MSAs. In Step 1, only homologous sequences were selected; this usually infers that they share significant identity. For homologous sequences it can also be inferred that similarities are representative of evolutionary events. Thus, it is possible, with due care, to use standard alignment software. The question then is which type of sequence is the best to align, amino acid or nucleotide?

Morrison (2006) answered this question in a very comprehensive 61 page paper, titled "Multiple sequence alignment for phylogenetic purposes". Morrison (2006) states that DNA nucleotide sequences are constructed using 4 characters A, G, C, and T (adenine, guanine, cytosine, and thymidine), while protein sequences have a 21 character set (i.e. 20 amino acids and 1 terminator). Because of the difference in size of these two character sets, it is significantly less likely that random alignments will be generated when using amino acid sequences, in contrast to nucleotide sequences. It has also been shown that, if an alignment accuracy (similarity) greater than 80% (a value normally preferred for the construction of phylogenetic trees) is required, then amino acid sequences require an identity of greater than 40%, while for nucleotide sequences an identity greater than 70% is required to achieve the same alignment accuracy. This clearly shows that it is much easier to accurately align amino acid sequences than it is to align nucleotide sequences. Therefore, when conducting any phylogenetic analysis, where possible it is preferable to use amino acid sequences.

Step 3 is the stage where the primary phylogenetic assessment in the form of a data set of aligned sequences is visualised in the form of an inferred phylogenetic tree (this process is sometimes referred to as the secondary phylogenetic

assessment). Currently there are a number of software packages that implement a variety of methods (i.e. Parsimony, Maximum Likelihood, or Bayesian Inference methods) to estimate trees from sequence data. The method used to eventually produce the tree will depend upon the researcher, who will choose the most appropriate method for the task.

What must be remembered is that all phylogenetic trees, irrespective of the methodology used to produce them, will be inaccurate. We will never know if a tree is correct because what actually happened in the past is unobservable and hence unknowable. The best that can be hoped for is that the estimated tree is the one that most closely approximates to the actual tree. Some biologists say that a phylogenetic tree is a hypothesis, but in my branch of science (engineering and physics) a hypothesis is only useful if it can be tested. We cannot test the validity of a phylogenetic tree. Therefore, the term hypothesis should not be used when referring to a phylogenetic tree: a phylogenetic tree should be referred to only as an inference (estimate).

Table 1: List of cytochrome c and myoglobin species sequence accession ID's used in Mammalian phylogenetic trees. All sequences were downloaded from GenBank. N.A. = Not Available.

Common Name	**Species**	**cytochrome c ID**	**myoglobin ID**
human	*Homo sapiens*	P99999	NP_005359
chimpanzee	*Pan troglodytes*	P99998	XP_001156591
rhesus monkey	*Macaca mulatta*	XP_014984096	XP_001082347
orangutan	*Pongo abelii*	XP_024104353	N.A.
orangutan	*Pongo pygmaeus*	N.A.	P02148
tree shrew	*Tupaia chinensis*	ELW66612	N.A.
tree shrew	*Tupaia glis*	N.A.	P02165
pig	*Sus scrofa*	P62895	NP_999401
horse	*Equus caballus*	NP_001157486	P68082
zebra	*Equus burchellii*	711086A	P68083
domestic cat	*Felis catus*	XP_006929319	XP_003989274
dog	*Canis familiaris*	NP_001183974	XP_850735
sperm whale	*Physeter catodon*	XP_007118006	P02185
goat	*Capra hircus*	ABA06540	XP_005680660
sheep	*Ovis aries*	P62896	P02190
cow	*Bos taurus*	NP_001039526	NP_776306
rabbit	*Oryctolagus cuniculus*	P00008	P02170
black rat	*Rattus rattus*	XP_032740407	XP_032773630
house mouse	*Mus musculus*	NP_031834	NP_038621
common wombat	*Vombatus ursinus*	XP_027728358	XP_027729286
kangaroo	*Macropus giganteus*	P00014	N.A.
kangaroo	*Osphranter rufus*	N.A.	P02194
chicken	*Gallus gallus*	P67881	XP_416292

3 Phylogenetic trees – gene based examples

The phylogenetic trees in Figures 1 and 2 are used to illustrate the above basic principles, using an example from Pevsner (2015).

3.1 Mammalian Relationships – Myoglobin Orthologs

Figure 1 shows a tree of myoglobin orthologs. The purpose for producing this tree was to illustrate the relationship between the primate species with the rest of mammalia.

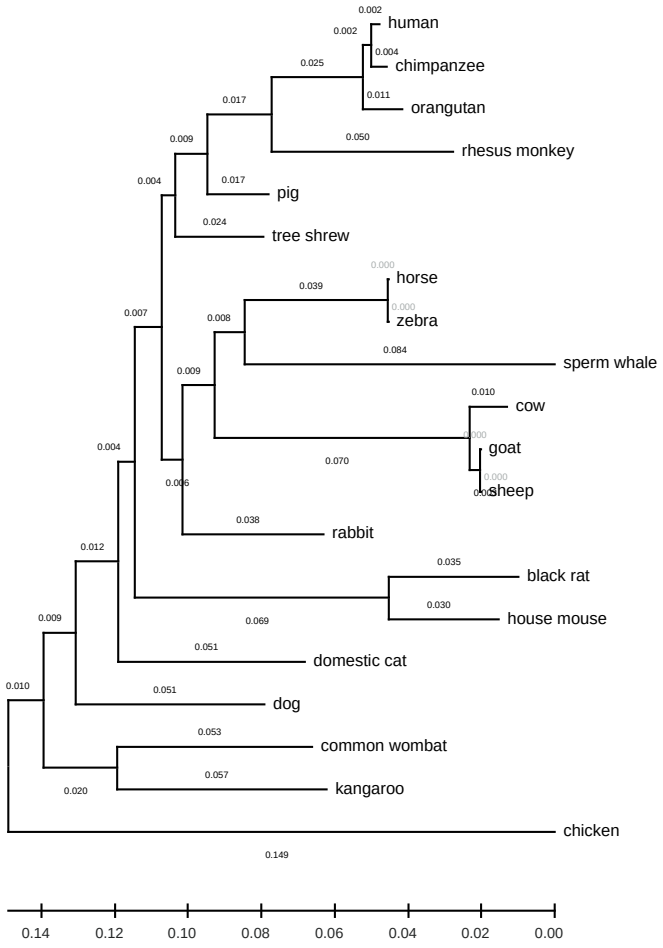

Figure 1: The Neighbor-Joining method was used to infer this phylogenetic tree based on aligned mammalian sequences of the myoglobin gene. The scale shows the evolutionary distances used to infer the phylogenetic tree.

Materials and Methods

To illustrate the relationship between the primates and the rest of the species in Mammalia a reasonable taxonomic spread of species is required. Table 1 contains the list of species (myoglobin sequences) to be used in this phylogentic experiment. The marsupials were from the Order Diprotodontia (kangaroo, common wombat), while the placental mammals were from the following Orders: Artiodactyla (pig, cow, sheep, goat); Carnivora (dog, domestic cat); Cetacea (sperm whale); Lagomorpha (rabbit); Perissodactyla (horse, zebra); Primates (human, chimpanzee, rhesus monkey, orangutan); Rodentia (black rat, house mouse); Scandentia (tree shrew). The outgroup chicken is from the Class Aves, Order Galliformes.

The myoglobin sequences were aligned using Mafft v7.471 (Katoh and Standley 2013) with the following command line "`mafft -ep 0 -genafpair -maxiterate 1000 [in-file.fasta] > [out-file.fasta]`". The MegaX software package was used to infer and visualise the evolutionary history of the taxa, using the Neighbor-Joining (NJ) method, and the evolutionary distances were computed using the p-distance method (Pevsner 2015).

Results

The myoglobin gene was used to illustrate the evolutionary history of a representative group of mammalian species. The phylogenetic gene tree that was produced seemed to be a reasonable representation of the evolutionary history of mammalian species. The tree showed that the marsupials are basal to the placental mammals and the remaining mammals were arranged in their characteristic groups.

It is also interesting to note that the horse (P68082) and zebra (P68083), although clearly separate and distinct species, have identical myoglobin gene sequence identities. This indicates that these species have, in evolutionary terms, speciated recently, hence they still have significant gene overlap.

3.2 Mammalian Relationships – Cytochrome C Orthologs

This phylogenetic experiment is a repeat of the one above but with a different gene, cytochrome c. The aim is to see if this gives a different result and if so, why. Cytochrome c is a member of the cytochrome complex of proteins. It is present in the cellular respiratory pathways in all eukaryotic organisms.
Normally, cytochrome c has a very low accepted mutation rate, and is highly conserved across a wide spectrum of species. This makes cytochrome c a useful gene candidate to explore evolutionary relationships for a broad range of species, and is why cytochrome c has been chosen for this experiment.

Materials and Methods

The cytochrome c sequences shown in Table 1 were aligned using Mafft v7.471 using the following command line "`mafft -ep 0 -genafpair -maxiterate 1000 [in-file.fasta] > [out-file.fasta]`". The MegaX software package was used to infer and visualise the evolutionary history of the taxa, using the Minimum Evolution (ME) method (Rzhetsky and Nei 1992). (Basically the ME method is to search for the NJ tree which has the minimum value of the total sum of branch lengths.)

Results

It is evident that the tree topologies between the myoglobin tree in Figure 1 and the cytochrome c tree in Figure 2 are significantly different, even though they were constructed using predominantly the same species. The difference that is most striking is the early divergence of the primate clade from the rest of the mammals, while most of the other groups (i.e. marsupials, rodents, cattle, etc.) are in their expected positions.

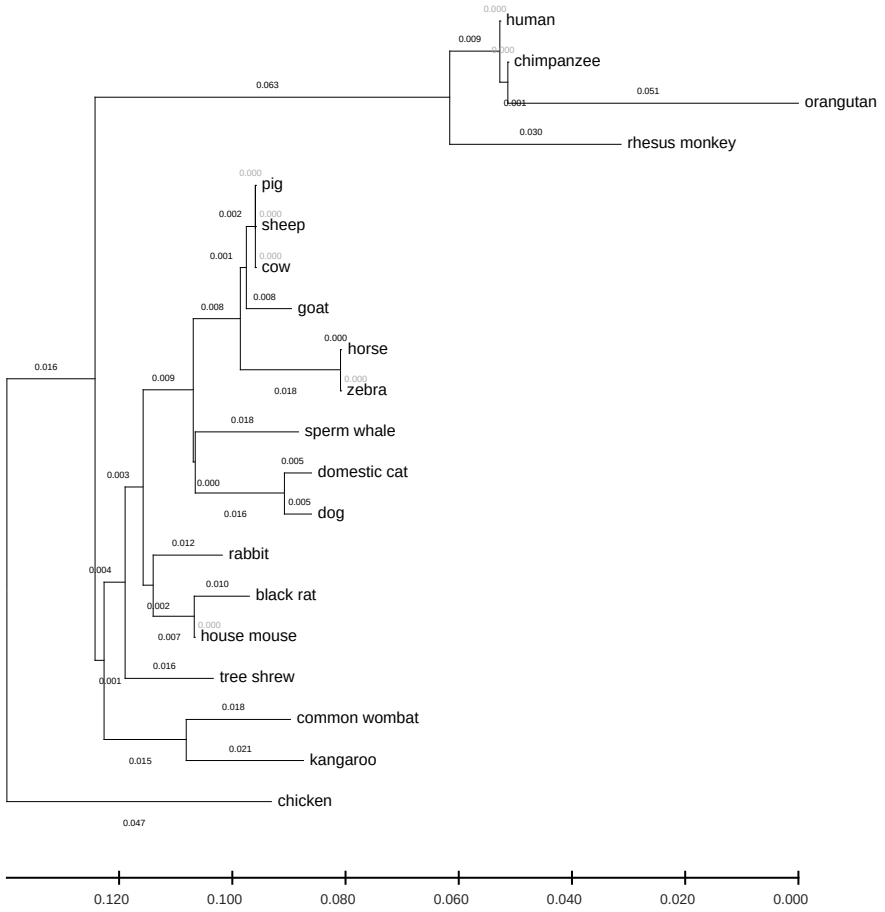

Figure 2: The Minimum-Evolution method was used to infer this phylogenetic tree based on aligned mammalian sequences of the cytochrome c gene. The scale shows the evolutionary distances used to infer the phylogenetic tree.

3.3 Discussion

The differences in tree topology between the myoglobin and the cytochrome c trees can be directly attributed to the choice of gene used. Whereas the myoglobin gene has a constant mutation rate throughout the mammalia, the same cannot be said for the cytochrome c gene. Normally the cytochrome c gene is highly conserved (having a very low amino acid substitution rate) but during different periods in its evolutionary history it has been shown to have periods of rapid substitution rates. Also, the relatively short amino acid sequence (104 amino acids) cannot be expected to provide precise divergence times and phylogenetic relationships in all cases. This is the reason why sequences for human (P99999) and chimpanzee (P99998) are identical.

In Figures 1 and 2, the significant difference in the position of the primate clade can be attributed to a period of relatively high mutation rate of the cytochrome c gene during the early formation of the clade, thus giving a false impression of how the primate clade evolved. Each gene has its own evolutionary history so care must be taken when selecting a gene to illustrate a particular phylogenetic relationship.

4 Are Mushrooms and Baker's Yeast related?

To state the question more precisely, are cultivated mushrooms (*Agaricus bisporus*) and baker's yeast (*Saccharomyces cerevisiae*) related? This is not a trivial question to answer: baker's yeast (*S. cerevisiae*) is in the subphylum Saccharomycotina, which is in the phylum Ascomycota, whereas the cultivated mushroom (*A. bisporus*) is in the phylum Basidiomycota. The evolutionary distance between these two species is huge, so it can be inferred that the common ancestor is a long way back in the past. To be able to produce a phylogenetic tree that spans such a huge evolutionary distance (history) requires the use of a gene that has a sequence mutation rate which is slow enough to be used to identify distantly separated species, but also fast enough to enable more closely related species to be separated.

Fortunately a gene from the cytochrome complex, called cytochrome c oxidase subunit 1 (COX1), has these attributes, making it a suitable candidate (Molitor *et al.* 2010). The COX1 gene is also used as the barcode for animals, and is the default marker adopted by the Consortium for the Barcode of Life (CBOL) for all groups of organisms, including fungi. Before 2010, this gene was relatively difficult to sequence, but after that date, and with the use of modern sequencing techniques, it has become more popular, resulting in an increase in the number of available fungi sequences.

4.1 Materials and Methods

The sequences for the 50 species chosen for this experiment were all downloaded from GenBank. To illustrate the large evolutionary distance that exists between baker's yeast and mushrooms, a spread of species from both Ascomycota and Basidiomycota is required. Three species of arbuscular mycorrhizal fungi from Glomeromycota were chosen as the outgroup, these being the most recent ancestors of all species in both the Ascomycota and Basidiomycota.

The sequences were aligned using Muscle version 3.8.31 and the inferred maximum-likelihood (ML) gene tree was produced using IQ-Tree version 2.0.5. IQ-Tree's inbuilt ModelFinder tested up to 546 protein models and the best-fit model chosen according to Bayesian Information Criterion (BIC) was the mtZOA+R4 model. The quick bootstrap option was set for 100,000 repeats.

4.2 Results

The ML gene tree in Figure 3 is well supported, demonstrating that COX1 gene can be aligned across distant taxonomic lineages. The inferred phylogenetic tree illustrates the evolutionary relationship between the cultivated mushroom (*A. bisporus*) and baker's yeast (*S. cerevisiae*). The sequence representing baker's yeast (*S. cerevisiae*) can be seen in the clade highlighted in , representing the subphylum Saccharomycotina. This subphylum is in the phylum Ascomycota, highlighted in , which connects to the node marked by Ⓐ. Saccharomycotina is also the sister group to the subphylum Pezizomycotina – this clade is highlighted in and contains most of the filamentous ascomycetes. The node marked by Ⓐ indicates when the fungal subkingdom Dikarya was formed; it is also when the most recent common

Figure 3: This consensus tree was obtained using the maximum-likelihood method (ML) based on aligned sequences of the COX1 gene. Ascomycota clade, Saccharomycotina clade, Pezizomycotina clade. Basidiomycota clade, Agaricomycotina clade. Outgroup Glomeromycota clade. Ⓐ indicates when the most recent common ancestor of baker's yeast and mushrooms existed.

ancestor of baker's yeast and mushrooms existed. The link highlighted in moving down from node Ⓐ represents the phylum Basidiomycota, which contains the subphylum Agaricomycotina, highlighted in . At the base of this clade the cultivated mushroom (*A. bisporus*) can be found.

4.3 Conclusion

By choosing an appropriate gene, the COX1 gene, it was possible to demonstrate, with a limited range of species, a well supported phylogenetic tree that spanned distant taxonomic lineages across the Kingdom Fungi. This phylogenetic tree clearly illustrated that baker's yeast (*S. cerevisiae*) and the cultivated mushroom (*A. bisporus*) had a common ancestor at node Ⓐ. This happened during the time when the subkingdom Dikarya with the two branches, Ascomycota and Basidiomycota, was formed.

5 Hymenopellis – revisited

All of the DNA analysis performed so far has used protein sequences to produce phylogenetic trees. The main reason for using sequences of protein encoded DNA is because they are expressed in the organism as an inherited characteristic that is subjected to the evolutionary processes. This makes protein based sequences ideal candidates for phylogenetic analysis. There is another type of DNA that is referred to as non-coding DNA, which is not expressed in the organism and most likely does not participate in the evolutionary process: mutations in this DNA will not result in any character changes in the organism. Because non-coding DNA does not participate in the evolutionary process, it is able to mutate rapidly without impacting on the organism. This rapid mutation makes sequences of non-coding DNA useful candidates for doing DNA analysis of closely related species.

One such a candidate is the nuclear ribosomal internal transcribed spacer (ITS) region, which includes two highly variable spacers (ITS1 and ITS2) consisting of non-coding DNA. Because of the variability of the ITS1 and ITS2 spacers it is possible to use this property to separate populations of various species by comparing the similarity between their sequences. By clustering groups of similar sequences it can be inferred that these groups may be the same species or species that belong in a particular genus. If two sequences have a similarity of 97% or more, then they can be considered as the same species. This is based on the valid assumption that nucleotide sequences with a high degree of similarity are most likely from the same or from a closely related population (species).

If populations of fungi have been separated for some time, this should show up as differences in their ITS sequences. In this experiment I chose to use ITS sequences to see if there are distinct population differences between northern and southern hemisphere species in the genus *Hymenopellis*. The reason for choosing this genus is because it has already undergone DNA analysis by Petersen and Hughes (2010), thus making the ITS sequences needed for this experiment readily available.

5.1 Materials and Methods

All of the sequences were downloaded from GenBank. Four *Oudemansiella* species were chosen for the outgroup, to act as the most recent ancestor to all the other species. Before proceeding with the multiple sequence alignment (MSA), a slight complication with the ITS region needs to be overcome. This complication is caused by the variable ITS1 and ITS2 spacers. On the one hand, there is the highly conserved 5.8s gene, and on the other hand we have the highly variable ITS1 and ITS2 spacers. To subject sequences featuring both highly variable and highly conserved parts to a MSA does not always produce an alignment suitable for species identification in the form of a tree. The problem is caused by the highly conserved 5.8s gene region: as this region is common in all the sequences, it will readily align and in the process skew the alignment of the variable information-bearing ITS1 and ITS2 spacers. To overcome this problem it is best to extract ITS1 and ITS2 spacers from the ITS data set and align them separately.

To identify and extract ITS1 and ITS2 spacers from the ITS data set, a Perl-based software tool called ITSx (Bengtsson-Palme *et al.* 2013) was used. Both the ITS1 and ITS2 data set were aligned separately using Muscle version 3.8.31. and then combined. The combined ITS1 and ITS2 data set was visualised as a maximum-likelihood (ML) tree, produced using IQ-Tree version 2.0.5. IQ-Tree's inbuilt ModelFinder tested up to 286 DNA models and the best-fit model chosen according to BIC was the HKY+F+G4 model. The quick bootstrap option was set for 100,000 repeats.

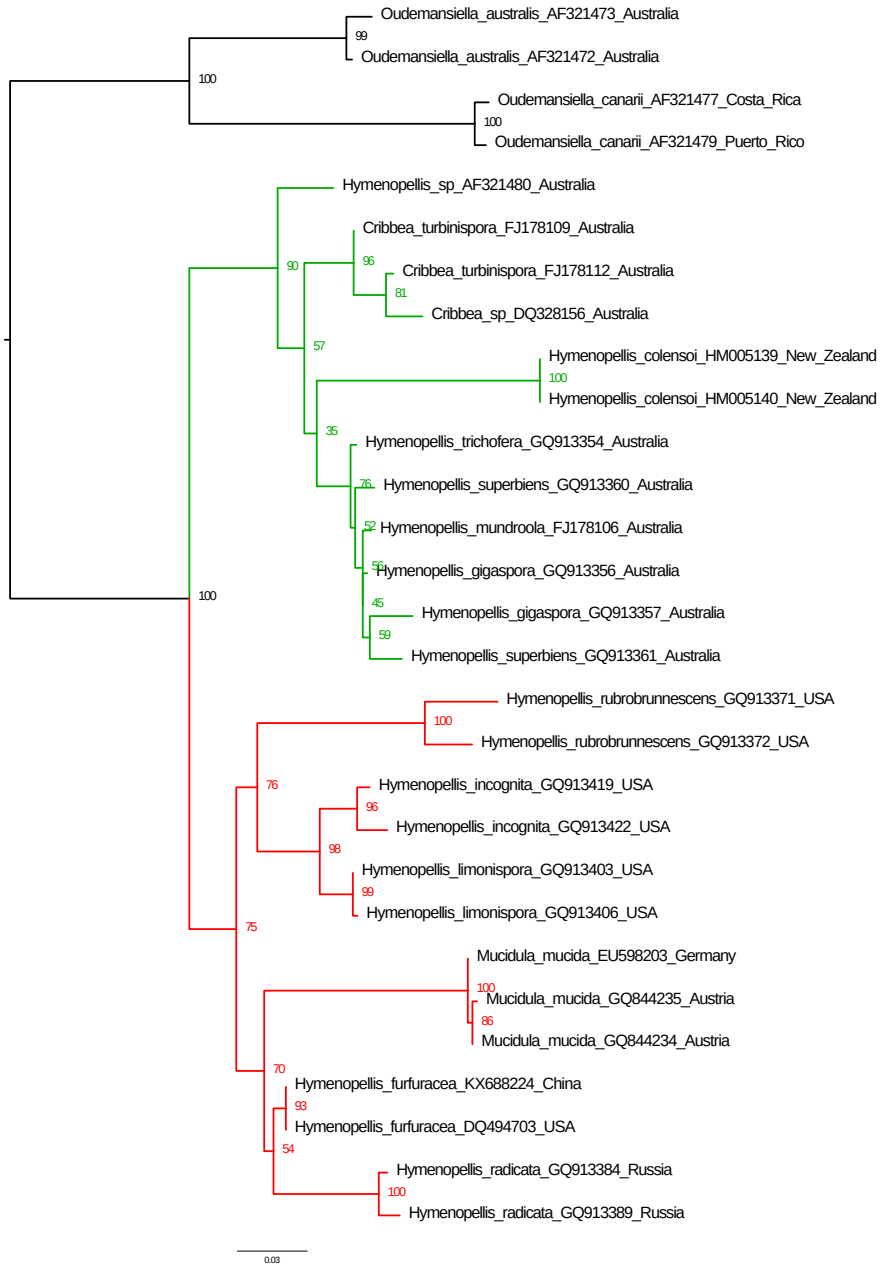

Figure 4: Maximum likelihood ITS1 and ITS2 consensus tree. Southern hemisphere (Australia, New Zealand) *Hymenopellis* species. Northern hemisphere clades of *Hymenopellis* species.

5.2 Results

The well supported maximum likelihood ITS1 and ITS2 consensus tree in Figure 4 clearly demonstrates that *Hymenopellis* and related species can be grouped into northern and southern hemisphere clades. This result was anticipated, but now there is clear evidence to show it. When two groups of species have been separated, even if the separation has not been all that long on the evolutionary time scale, differences will show up in the sequences of the highly variable ITS spacer region.

The result of this experiment does not necessarily mean that the genus *Hymenopellis* needs to be broken up into to separate genera, one for the northern and another for the southern hemisphere. All that can be said is that the two clades have been separated for a significant period of time. In the same manner the highly variable ITS spacer region should not be used as the sole determiner to identify closely related cryptic species. The two cryptic species may be in slightly different populations. This will show up in the ITS spacer region, but it cannot be assumed that the much slower process of speciation has occurred. Other evidence for this would be required.

6 Conclusion

This Note is largely the result of my isolation during the Covid 19 pandemic, which gave me time to delve into the process of producing phylogenetic trees. My haphazard heuristic approach eventually forced me to go back and look at the basics that define a phylogenetic tree and its evolutionary underpinning. I demonstrated the use of these basic principles by constructing a myoglobin and then a cytochrome c gene tree to illustrate the inferred relationships between the primates and the rest of the mammalian species. I then used the COX1 gene to illustrate that baker's yeast and the cultivated mushroom are related. In the process I also demonstrated that the COX1 gene would be an excellent gene to use in fungal phylogenetic analysis. I then proceeded to use the ITS spacer regions to show that the northern and southern hemisphere species of the genus *Hymenopellis* form two distinct clades.

Using the software to do all of this work was relatively easy and this supports my initial claim "It is not difficult to make a Phylogenetic Tree". The problem is that although making a phylogenetic tree with today's software is easy, this tends to hide the fact that the prerequisite understanding of phylogenetic analysis, how the DNA code, evolution, and organisms are all related, is hellishly complex, and at the same time an extremely exciting area for ongoing research. I have only managed to brush the surface of the new fields of "Bioinformatics" and "Functional Genomics" and unfortunately at my age I feel that is all I am capable of.

7 References

Bengtsson-Palme J, Ryberg M, Hartmann M, Branco s, Wang Z, Godhe A, De Wit P, Sánchez-García M, Ebersberger I, de Sousa F, Amend AS, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson KM, Vik U, Veldre V, Nilsson RH (2013) "Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data". *Methods in Ecology and Evolution* **Vol. 4 pp. 914–919.** https://doi.org/10.1111/2041-210X.12073

D'Andreano S, Cuscó A, Francino O (2020) "Rapid and real-time identification of fungi up to the species level with long amplicon Nanopore sequencing from clinical samples". bioRxiv preprint first posted online Feb. 6, 2020; doi: http://dx.doi.org/10.1101/2020.02.06.936708

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) "TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses". *Bioinfomatics* **Vol. 25(15), pp. 1972–1973.** doi:10.1093/bioinformatics/btp348

Katoh K and Standley DM (2013) "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". *Molecular biology and evolution* **Vol. 30(4) pp. 772–780.**

Molitor C, Inthavong B, Sage L, Geremia RA, Mouhamadou B (2010) "Potentiality of the cox1 gene in the taxonomic resolution of soil fungi". *FEMS Microbiological Letters* **Vol. 302(1), pp. 76–84.** https://doi.org/10.1111/j.1574-6968.2009.01839.x

Morrison DA (2006) "Multiple sequence alignment for phylogenetic purposes". *Australian Systematic Botany* **Vol. 19, pp. 479–539.**

Petersen RH and Hughes KW (2010) "The Xerula/Oudemansiella Complex (Agaricales)". *Nova Hedwigia* **Beiheft 137, pp. 1–625.**

Pevsner J. (2015) *Bioinformatics and Functional Genomics*, Third Edition, published by John Wiley & Sons Inc, UK.

Rzhetsky A and Nei M (1992) "A Simple Method for Estimating and Testing Minimum-Evolution Trees". *Molecular Biology and Evolution* **Vol. 9(5) pp. 945–967.**