JH Bioinformatics Note – 2

Hericium Revisited

"It is not difficult to make a Phylogenetic Tree"

Jurrie Hubregtse

November 18, 2020

Hericium coralloides

Introduction

This Note is the next in the series "It is not difficult to make a Phylogenetic Tree". As I see it, inferring a phylogentic tree from a set of text strings (sequences) involves getting a computer to do a lot of text string manipulation, from which, eventually, various degrees of similarity between the text strings can be illustrated in the form of a phylogenetic tree. In this very basic description of phylogenetic analysis some very elegant software is used. However, people have been using elegant software for a long time now (think of word processors, spread sheets, etc.), but this is not the hard part. The hard and most important part of producing a phylogenetic tree is to determine what needs to be illustrated and how to select the appropriate sequences needed in order to construct the phylogenetic tree. Once this has been achieved, making the phylogenetic tree is easy because there is ample high quality software to do that. But, once your phylogenetic tree has been constructed, the question you need to ask is, "Can I be confident in this phylogenetic tree, considering the methodology used in its construction, and is it fit for purpose?" If the answer is yes, all is well, but if the answer is no, then you need to start again and go right back to the sequence selection stage.

Recently, it came to my attention that the Australian species *Hericium coralloides* may be phylogenetically similar to the New Zealand species *Hericium novae-zealandiae*. Since there are some *H. novae-zealandiae* sequences available in GenBank, I thought that it would be interesting to repeat the phylogenetic experiment presented in the article by Jumbam *et al.* (2019), which showed the reconstruction of a phylogenetic tree of the genus *Hericium*, but this time with the inclusion of *H. novae-zealandiae* sequences.

This phylogenetic experiment "*Hericium* Revisited" illustrates many of the issues that can be encountered when using sequence data from GenBank. Apart from numerous misidentifications of taxa, largely due to the morphological similarity between *Hericium* species, there is also a considerable variation in sequence lengths and quality – as well as the problem of inadequate species sampling – to overcome.

This phylogenetic experiment requires the use of genes found in the RNA gene cluster shown in Fig. 1. This gene cluster consists of three genes, the 18S gene (referred to as the small sub-unit or SSU region), the 5.8S gene plus its internal transcribed spacer regions, ITS1 and ITS2 (the ITS1-5.8S-ITS2 is usually referred to as the ITS region), plus the 28S gene (referred to as the large sub-unit or LSU region).

The ITS, SSU and LSU regions evolve at different rates; ITS is highly variable compared to the SSU and LSU regions. This high degree of variability of the ITS region can be explained by the relatively low evolutionary pressure on the ITS spacer regions, which are not expressed in the organism (phenotype). Because of this variability, the ITS region is used as the official barcoding marker for species-level identification in fungi, for which it is well suited. The ITS region can usually provide robust phylogenetic reconstructions within species and genus levels, but less so within family levels. The SSU and LSU markers are recommended for phylogenetic reconstructions at both family and generic levels.

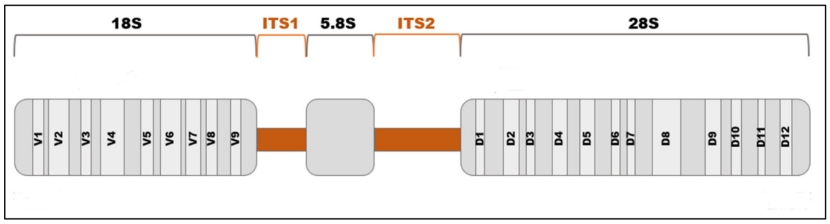

Figure 1: This RNA gene cluster consists of the small sub-unit or SSU region, the ITS region (the highly variable ITS1 and ITS2 regions are marked in orange), plus the large sub-unit or LSU region. The dark grey areas are sections which are highly conserved (See D'Andreano *et al.* 2020).

Experiment 1 – *Hericium* revisited

This experiment was based on the phylogenetic reconstruction of the genus *Hericium* as described in the article by Jumbam *et al.* (2019). The intention was to follow their method as closely as possible, by using many of the same sequences and phylogenetic tools. The experiment was not exactly the same, because of the inclusion of some extra sequences – those relating to *H. novae-zealandiae* plus a single sister clade (a closely related genus) for the outgroup. Using a single sister clade with multiple species as an outgroup is regarded as good practice, but is not essential (Lou *et al.* 2010; Lyons-Weiler *et al.* 1998). These minor changes should have no significant impact upon the phylogenetic tree topology. During reconstruction of the *Hericum* phylogenetic tree an extra tool was used, namely 'AliViewer', which is an alignment viewer and editor (Larsson 2014). This tool enables sequence alignments to be visually inspected and edited if necessary.

Materials and Methods

In this experiment the outgroup consisted of 4 species in the genus *Dentipellis* and the ingroup consisted of 46 *Hericium* species. All of the sequences were downloaded from GenBank using the NCBI Taxonomy Browser WEB page: (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root).

The downloaded ITS and LSU sequences are listed in Table 1 and species names that are not in bold have been taken directly from the article by Jumbam *et al.* (2019). There are 50 ITS sequences and 14 LSU sequences. The table also has 8 highlighted rows, which contain sequences that span both the ITS and LSU regions and have the same GenBank sequence ID's in both the ITS and LSU columns.

Table 1: List of ITS and LSU sequences used for analysis. All species not in bold are from the article by Jumbam *et al.* (2019). The highlighted rows contain sequences that span both the ITS and LSU regions and have the same GenBank sequence ID's in both the ITS and LSU columns. N.A. = Not Available

Species	**strain/culture**	**region/country**	**ITS GenBank ID**	**LSU GenBank ID**
Dentipellis dissita	NH6280	N.A.	AF506386	AF506386
Dentipellis fragilis	Dai 9009	N.A.	JQ349108	JQ349094
Dentipellis leptodon	CDS 125879	NZ	MH864083	MH875541
Dentipellis leptodon	FCUG 2983	NZ	JQ716932	N.A.
H. abietis	CBS 125851	Europe	MH863807	MH875271
H. abietis	CBS 243 48	Canada	AY534579	MH867882
H. abietis	HMJAU 23283	China	JF430070	N.A.
H. abietis	K 107270	UK	EU784259	N.A.
H. abietis	NH6990	Canada	AF506456	AF506456
H. alpestre	DSM 108284	Germany	MK491173	MK491173
H. alpestre	JQ716936	Romania	JQ716936	N.A.
H. alpestre	NH13240	Russia	AF506457	AF506457
H. americanum	AFTOL	469 USA	DQ206987	DQ411538
H. americanum	CBS 129 40	N.A.	MH856062	N.A.
H. americanum	CBS 493 63	USA	AY534581	N.A.
H. americanum	AF506458	Canada	AF506458	AF506458
H. bembedjaense	JB 46	Cameroon	MK683483	MK683482
H. bharengense	KD 14 005	India	MK335755	N.A.
H. bharengense	CBS kd10658E	India	JN185603	N.A.
H. cirrhatus	F794	Germany	AF506385	AF506385
H. cirrhatus	K 125827	Europe	EU784260	N.A.
H. cirrhatus	K 135658	Europe	EU784261	N.A.
H. coralloides	ATCC 52480	Canada	AY534584	N.A.
H. coralloides	FCUG 3209	Argentina	JQ716934	N.A.
H. coralloides	FCUG 426	France	JQ716935	N.A.
H. coralloides	Cui 14825	China	MH085948	MH085962
H. coralloides	HMJAU 23285	China	JF430072	N.A.
H. coralloides	HMJAU 23287	China	JF430074	N.A.
H. coralloides	HMJAU 4368	China	JF430062	N.A.
H. coralloides	HMJAU 4990	China	JF430064	N.A.
H. coralloides	IFO 7716	USA	AY534582	N.A.
H. coralloides	K 104978	Europe	EU784262	N.A.
H. coralloides	K 61470	UK	EU784263	N.A.
H. coralloides	NH282	Sweden	AF506459	AF506459
H. coralloides	NYBG 0011	USA	KY432832	N.A.
H. erinaceus	MK463636	USA	MK463636	N.A.
H. erinaceus	CBS 202 31	Netherlands	MH855186	MH866638
H. erinaceus	CBS 260 74	Netherlands	DQ185922	MH872587
H. erinaceus	CBS 448 85	CzechRepublic	DQ185926	N.A.
H. erinaceus	DAOM 19644	Canada	JN649345	JN649345
H. erinaceus	K 62494	Europe	EU784265	N.A.
H. erinaceus	KUMC 1035	SthKorea	DQ185919	N.A.

Continued on next page

Species	strain/culture	region/country	ITS GenBank ID	LSU GenBank ID
H. erinaceus	NH12163	Russia	AF506460	AF506460
H. erinaceus	Zh001	Tunisia	KY655903	KY655904
H. flagellum	MG649451	Poland	MG649451	N.A.
H. rajchenbergii	GR1997	Argentina	JX403945	N.A.
H. rajchenbergii	GR2041	Argentina	JQ716939	N.A.
H. yumthangense	JX855033	India	JX855033	JX855034
H. novae-zealandiae	MN044068	NZ	MN044068	MN044060
H. novae-zealandiae	MN044067	NZ	MN044067	MN044059

The first step in the experiment was to produce the aligned ITS dataset. To do this the method described in the article by Jumbam *et al.* (2019) was used. The ITS sequences shown in Table 1 were aligned using Muscle v3.8.31 (Edgar 2004) using its default options.

The alignment was trimmed using a command line software package called TrimAl (Capella-Gutiérrez *et al.* 2009). As per the article, the command line options used were -gt 0.60 and -cons 0.50. The -gt option sets the gap threshold; its value range is [0–1], which is the fraction of sequences in a column without gaps. The -gt 0.60 tells TrimAl to accept only columns that are 60% or more gap free. The -cons option sets the conservation percentage; its value range is [0–100], which is the minimum percentage of the columns in the original alignment to conserve. The -cons 0.50 tells TrimAl to conserve at least 0.5% or more of the alignment. However, if the number of columns that achieve or pass the -gt threshold is less than the number of columns fixed by the -cons parameter, then TrimAl relaxes the -gt threshold value in order to achieve the minimum conservation percentage. The often-used range of values given for the -cons option is 50–90%, not 0.50% as stated in the article.

An automated trimming package such as TrimAl needs to be exercised with care because the trimming process can unintentionally remove some of the phylogenetic signal (Tan *et al.* 2015).

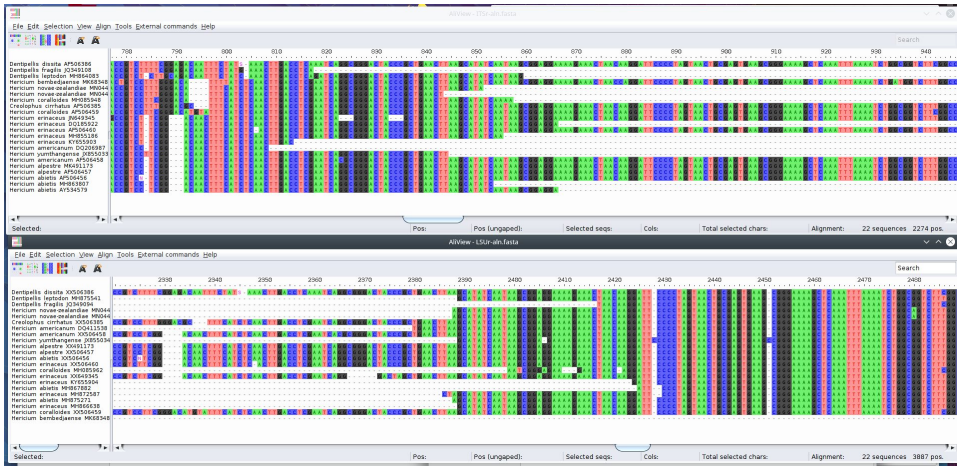

Figure 2: Aligned ITS and LSU sequences in AliView ready to be trimmed.

The second step was to produce the merged ITS+LSU dataset. The methodology used by Jumbam *et al.* (2019) was not described except that they used a software package called MEGA7 (Kumar *et al.* 2015), which is a general purpose phylogenetics package; the current version is MEGA X. Since I had written my own software tool to merge aligned sequences I used that instead of MEGA X.

Muscle was used to align both the ITS and LSU datasets. Because both these datasets contain 9 sequences that span both the ITS and LSU regions, it was necessary to trim the alignments in such a way that the ITS dataset contained only sequence data belonging to the ITS gene region, and the LSU dataset contained only data belonging to the LSU gene region.

The ITS and LSU alignments were arranged in AliView, as seen in Figure 2, where the top alignment belongs to the ITS dataset and the bottom alignment belongs to the LSU dataset. For the ITS alignment, the sequence region in view is at the end of the ITS region. For the LSU alignment, the sequence region in view is at the start of the LSU region. These sequences were then manually trimmed so that the ITS and LSU alignments contained only their relevant sequence data. After trimming, the ITS and LSU datasets were combined to form a single ITS+LSU dataset. There were now 2 aligned datasets, the ITS (produced earlier) and the ITS+LSU, ready for analysis.

Similar maximum likelihood (ML) analysis was performed on each of the 2 datasets using IQ-TREE. In both cases the best-fit evolutionary models used were those chosen by IQ-TREE's inbuilt ModelFinder based on Bayesian Information Criterion (BIC). For the ITS dataset the best-fit model was the HKY+F+G4 model and for the ITS+LSU dataset it was the TNe+R2 model. IQ-TREE's rapid bootstrapping option was set to 100,000 replicates for all phylogenetic reconstructions.

Results

It is evident that the phylogenetic tree topologies shown in Figures 3 and 4, produced using the method described by Jumbam *et al.* (2019), are not similar to those illustrated in their article. The reasons for these differences will be addressed below.

What can be noted is that the *H. novae-zealandiae* clade, highlighted in , is well embedded in the *H. coralloides* complex of species, highlighted in . The *H. novae-zealandiae* clade contains not only two *H. novae-zealandiae* specimens but also an Argentinian *H. coralloides* (JQ716934) that is very closely related to *H. novae-zealandiae*.

Some other clades of interest are *H. abietis* and *H. americanum* , both well formed clades, and *H. erinaceus* , which seems to be split into 2 clades, both of which are highlighted.

The phylogenetic tree reconstructed using the ITS+LSU dataset shown in Figure 4 has a topology consistent with the tree in Figure 3. The topology of this tree is not particularly robust or useful because many of its backbone branches do not have strong bootstrap support and the LSU data sample is too small to allow any conclusions to be drawn about the phylogenetic position of the the two *H. novae-zealandiae* specimens.

Figure 3: This ML phylogenetic reconstruction is based on the method described by Jumbam *et al.* (2019). The ITS dataset used is shown in Table 1.

The third step was to determine why there is such a difference between the phylogenetic reconstructions described above and the one presented in the article by Jumbam *et al.* (2019). The direction taken was to use the unorthodox conservation value chosen for the automated trimming package Trimal. The -cons 0.50 value that was used would have produced the same result as the default value of zero and would have been computationally more efficient. Assuming that there was a transcription error in the article and the more likely

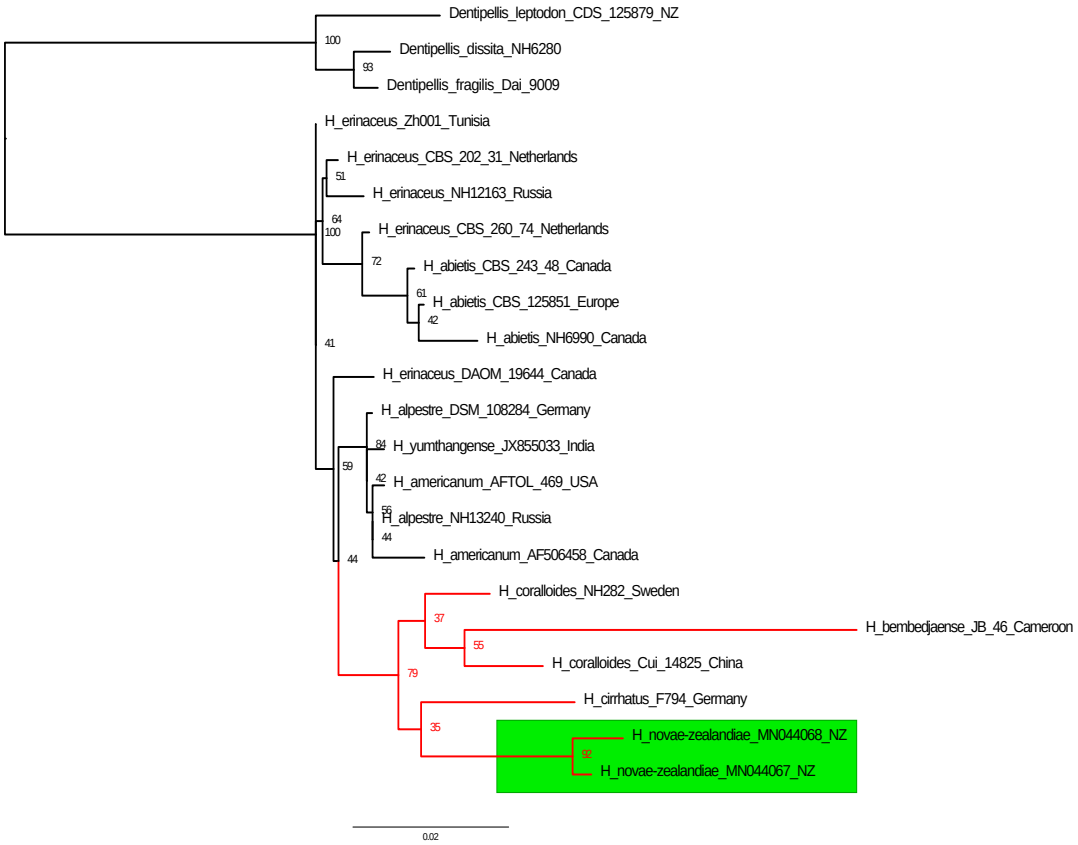

Figure 4: Phylogeny of *Hericium*, reconstructed from the combined ITS and LSU datasets using species in Table 1. Dataset was aligned using Muscle, then manually trimmed, and IQ-TREE was used to produce the maximum likelihood inference tree.

conservation value of 50% was intended, this would have a significant impact on the phylogenetic reconstruction.

This time the method was repeated but with the TrimAl command line parameters set to -gt 0.60 -cons 50; IQ-TREE was used to make the ML reconstruction. Figure 5 shows the resultant topology of the revised phylogenetic construction and it is very similar to the one that was published. However, a number of other issues presented themselves during the reconstruction.

During the ML reconstruction, IQ-TREE printed a warning that 41 sequences contained more than 50% gaps/ambiguity. Such a large level of gaps/ambiguity should be enough to stop the reconstruction and a thorough examination of the alignment should be conducted before proceeding. To determine what was causing this problem the pre- and post- trimmed alignments were viewed in AliView. It was surprising to see that length of the pre-trimmed alignment was 3885 bp (base pairs) because the usual length of an aligned ITS region is in the order of 700 bp. The excess length was due to one sequence MK491173 starting in the SSU region, passing through the ITS region, and well into the LSU region, a length well over 3000 bp. Also, there were 7 other sequences AF506XXX that spanned partially across the ITS region and all the way into the LSU region. After trimming the alignment with TrimAl, with the conservation value set at 50%, the resultant length was reduced to half, 1942 bp

Figure 5: Phylogeny of *Hericium*, reconstructed from the ITS dataset using species in Table 1. The dataset was aligned using Muscle, trimmed using revised TrimAl options, and ML inference tree was produced using IQ-TREE.

as expected. Figure 6 shows the ITS1 spacer region, with 7 sequences exhibiting horizontal gaps across this vital region; not shown are 9 sequences that traverse well into the LSU region, leaving gaps in the remaining 41 sequences. It is clear from visual inspection that the alignment produced would most likely not produce a reliable and representative phylogenetic tree.

IQ-TREE tries to make a ML reconstruction from the problematic alignment

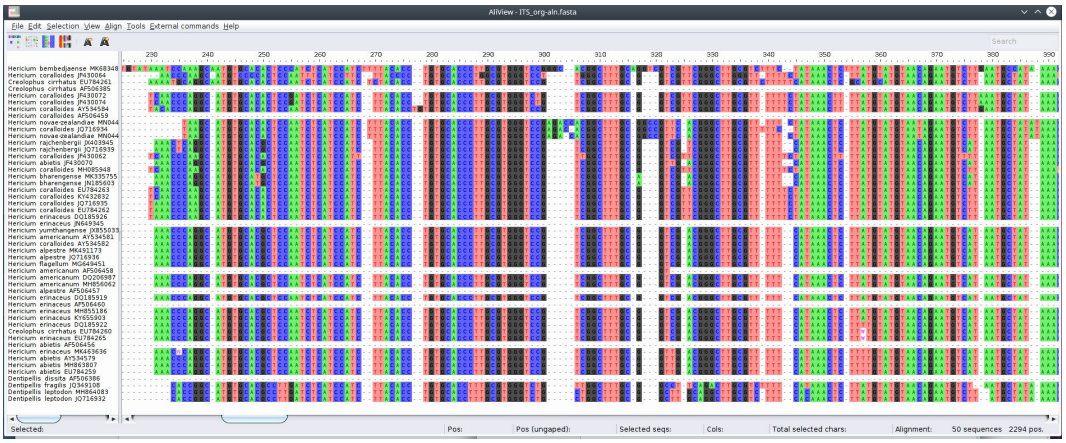

Figure 6: Aligned ITS dataset from Table 1.

because that is what the program is meant to do. But the topology of the phylogenetic reconstruction it produced, shown in Figure 5, is influenced more by the gaps in the alignment than by the phylogenetic relationships between the various *Hericium* species. At this stage it would be recommended that this alignment should not be used to draw any phylogenetic conclusions about the genus *Hericium*.

Experiment 2 – *Hericium* revisited again

The problems encountered while constructing the above phylogenetic trees were largely the result of poorly selected sequences, which produced problematic alignments resulting in unreliable phylogenetic trees. To overcome these problems the sequence selection was revised and problematic sequences were removed and replaced. Most of the replacement sequences came from the same species. No sequences from the LSU gene region were selected or used because an inadequate number of sequence samples were available from GenBank. The phylogenetic tree in Figure 8 was based on the method outlined by Jumbam *et al.* (2019) but this time there was no need to use TrimAl.

Materials and Methods

Table 2 shows the sequences downloaded from GenBank that were used in the revised phylogenetic reconstruction. The new sequences are highlighted.

Table 2: List of updated ITS sequences used for analysis. The highlighted rows contain the new sequences. N.A. = Not Available

Species	region/country	ITS GenBank ID
Dentipellis fragilis	N.A.	JQ349108
Dentipellis fragilis	Russia	MG734825
Dentipellis leptodon	NZ	MH864083

Continued on next page

Species	region/country	ITS GenBank ID
Dentipellis leptodon	NZ	JQ716932
H. abietis	Europe	MH863807
H. abietis	Canada	MH856328
H. abietis	Canada	AY534579
H. abietis	Canada	MF955013
H. abietis	UK	EU784259
H. alpestre	Italy	MN861076
H. alpestre	Romania	JQ716936
H. americanum	USA	MN906129
H. americanum	USA	MN906132
H. americanum	USA	MN906131
H. americanum	USA	DQ206987
H. americanum	N.A.	MH856062
H. americanum	USA	AY534581
H. bembedjaense	Cameroon	MK683483
H. bharengense	India	MK335755
H. bharengense	India	JN185603
H. coralloides	Canada	AY534584
H. coralloides	Argentina	JQ716934
H. coralloides	France	JQ716935
H. coralloides	Ukraine	MG549952
H. coralloides	Russia	MG735348
H. coralloides	China	MH085948
H. coralloides	China	MH085949
H. coralloides	China	JF430072
H. coralloides	China	JF430074
H. coralloides	USA	MT110631
H. coralloides	USA	MT551926
H. coralloides	USA	AY534582
H. coralloides	Europe	EU784262
H. coralloides	UK	EU784263
H. erinaceus	India	MT448853
H. erinaceus	USA	MT731944
H. erinaceus	Netherlands	MH860849
H. erinaceus	Netherlands	DQ185922
H. erinaceus	Slovenia	AM981220
H. erinaceus	CzechRepublic	DQ185926
H. erinaceus	China	MH085955
H. erinaceus	SthKorea	KP004969
H. erinaceus	SthKorea	KP004980
H. flagellum	Poland	MG649451
H. rajchenbergii	Argentina	JX403945
H. rajchenbergii	Argentina	JQ716939
H. yumthangense	India	NR155021
H. yumthangense	India	JX855033
H. novae-zealandiae	NZ	MN044068
H. novae-zealandiae	NZ	MN044067

The 50 ITS sequences listed in Table 2 were aligned using Muscle. The

ends of the alignment were manually trimmed in AliView, with the resultant alignment length being 695 bp. The ML phylogenetic reconstruction was done using IQ-TREE. As in the first method, IQ-TREE's inbuilt ModelFinder was used to determine the best-fit model, which was K2P+R2 according to BIC criteria. The rapid bootstrapping option was set to 100,000 replicates.

Results

Figure 7 shows the ITS1 spacer region of the revised sequences listed in Table 2. The horizontal gaps that were present in the original alignment are now gone. Some of the remaining vertical gaps are largely due to a few species, such as *H. bembedjaense* from the Cameroon (with the GenBank ID = MK683483), which are exhibiting a larger than usual number of mutations; these may be real or possibly a sequencing artifact. Although these sequences have a large number of mutations they can still be accommodated within the alignment. Species with sequences that exhibit a large number of mutations have long branch lengths in the final phylogenetic tree.

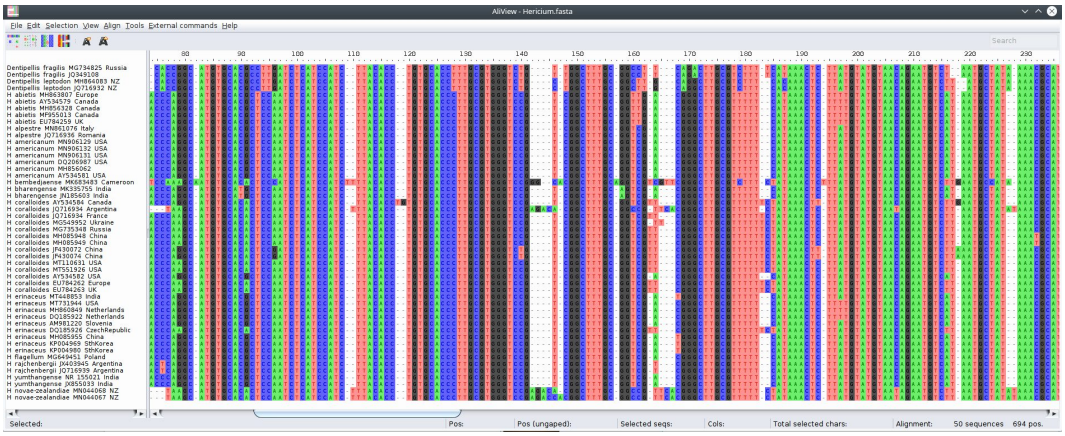

Figure 7: Aligned ITS dataset from revised sequences shown in Table 2. The area in view is the ITS1 spacer region.

It is also apparent that the vast majority of the columns of data in the ITS1 region have a single colour, indicating that the column consists of a single letter (nucleotide A, C, G, T) showing that the data for each sequence in that column is the same. This is not normally the case for the highly variable ITS1 region, and indicates that there is a low interspecific sequence variation, resulting in a weakened phylogenetic signal with which to construct a phylogenetic tree.

Figure 8 shows the phylogenetic reconstruction for the genus *Hericium* using the revised ITS dataset constructed from sequences in Table 2. The revised sequences have resulted in a topology consistent with the initial phylogenetic tree that was produced, but not with the topology of the tree presented in the article by Jumbam *et al.* (2019).

Figure 8: Phylogeny of *Hericium*, reconstructed from the revised ITS dataset. Muscle was used to align the dataset and IQ-TREE was used to produce the maximum likelihood inference tree.

Discussion

During this phylogenetic experiment, 3 phylogentic reconstructions of the genus *Hericium* have been attempted. In the first attempt the method described in the article by Jumbam *et al.* (2019) was followed, but resulted in a tree topology that was inconsistent with the one in the published article. The second attempt

was to determine the cause of the difference between the two topologies and to produce a phylogenetic tree that was similar to the published version. This was successfully done. In the process it became evident that there were major issues with the sequence selection for the original phylogenetic reconstruction. A revised set of sequences was chosen and a new phylogenetic reconstruction was attempted. This final attempt at reconstructing the *Hericium* genus went smoothly without any issues presenting themselves, apart from revealing that there is a low interspecific sequence variation in the ITS gene region between the various *Hericium* species. This low interspecific sequence variation resulted in a phylogenetic reconstruction with a tree that had just adequate bootstrap support from which phylogenetic inferences could be made.

In all of the phylogenetic reconstructions the position of *H. novae-zealandiae* remained in the *H. coralloides* complex. From this it can be said with a high degree of confidence that *H. novae-zealandiae* belongs in the *H. coralloides* clade. Another interesting discovery was that the Argentinian species of *H. coralloides* (GenBank ID JQ716934) is extremely closely related to *H. novae-zealandiae* and both can be considered to be the same species.

From the results obtained it is difficult to say whether *H. novae-zealandiae* belongs to a complex of species, or to a complex of species names in the *H. coralloides* clade. Further work by taxonomists will be required to resolve this problem.

The final reconstruction had well supported positions for the *H. abietis*, *H. americanum*, and *H. erinaceus* clades. It was observed that *H. erinaceus* had two separate and well defined clades, possibly indicating that *H. erinaceus* can be broken up into 2 separate species. More work with a different set of genes will most likely be needed to verify this observation.

These phylogenetic experiments have shown that the quality of a multiple sequence alignment is reflected in the final topology of the phylogenetic reconstruction. Therefore it is important that a high quality alignment be achieved, and this can best be done by paying special attention to the sequences and outgroup species that are selected. It is also important to assess the quality of sequence alignments before proceeding to the phylogenetic tree construction stage. Usually this can be achieved by visual inspection: problems with an alignment can often be detected visually because the human eye is very good at picking up any anomalous variations within an aligned pattern.

Conclusion

By repeating the phylogenetic experiment described by Jumbam *et al.* (2019), it was possible to show, with a significant amount of phylogenetic evidence, that *H. novae-zealandiae* clearly belongs in the *H. coralloides* clade and is closely related to an Argentinian *H. coralloides*.

Phylogenetic reconstructions are pivotal to our understanding of fungal taxonomy and evolution. As demonstrated, by repeating a phylogenetic experiment described in the literature we can gain a deeper understanding of the taxonomy and relationships between species of fungi.

References

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

Edgar, RC (2004) " MUSCLE: multiple sequence alignment with high accuracy and high throughput". *Nucleic Acids Research* **Vol. 32(5), pp. 1792–1797.**

Jumbam B, Haelewaters D, Kock RA, Dentinger BTM, Henkel TW, Aime MC (2019) "A new and unusual species of *Hericium* (Basidiomycota: Russulales, Hericiaceae) from the Dja Biosphere Reserve, Cameroon". *Mycological Progress* **Vol. 18, pp. 1253–1262.**

Available from:-
`https://www.researchgate.net/publication/336926773_A_new_and_unusual_species_of_Hericium_Basidiomycota_Russulales_Hericiaceae_from_the_Dja_Biosphere_Reserve_Cameroon`

Kumar S, Stecher G, Tamura K (2016) "MEGA7: Molecular Evolutionary Genetics Analyses version 7.0 for bigger datasets". *Molecular Biology and Evolution* **Vol. 33, pp. 1870–1874.**

Larsson A (2014) "AliView: A fast and lightweight alignment viewer and editor for large datasets". *Bioinformatics* **Vol. 30(22), pp. 3276–3278.**

Lou A-R, Zhang Y-Z, Qiao H-J, Shi W-F, Murphy RW, Shu C-D (2010) "Outgroup selection in tree reconstruction: a case study of the family Halictidae (Hymenoptera: Apoidea)". *Acta Entomologica Sinica* **Vol. 53(2), pp. 192–201.**

Lyons-Weiler J, Hoelzer GA, Tausch RJ (1998) "Optimal outgroup analysis" *Biological Journal of the Linnean Society* **Vol. 64, pp. 493–511.**

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies". *Molecular Biology and Evolution* **Vol. 32, pp. 268–274.** https://doi.org/10.1093/molbev/msu300

Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, Dessimoz C (2015) "Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference". *Systematic Biology* **Vol. 64, pp. 778–791.**

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) "Jalview Version 2 - a multiple sequence alignment editor and analysis workbench". *Bioinformatics* **Vol. 25(9), pp. 1189–1191.** doi: 10.1093/bioinformatics/btp033