Automated fungal DNA sequence comparison

Dylan Glotzer '11[1], Anders Ohman '10[1], Darcy Young '11[1] Faculty Sponsor: David Hibbett[1]

[1]Department of Biology, Clark University, Worcester, MA

Introduction

Tansley is an ongoing research project that seeks to examine fungal nucleotide sequences of the highly variable internal transcribed spacer (ITS) region of ribosomal DNA. Ribosomal DNA codes for ribosomal RNA, which translates messenger RNA into proteins. However, the ITS region has no effect on an organism's fitness, so theITS region is under weak selection pressures and can therefore vary widely between groups of organisms. ITS sequences provide information about genetic variation both between and among species of fungi. A major goal of Tansley is to automate the process of gathering and analyzing sequences can be done with a single all-encompassing script. During the past year, much work has been done on the four independent processes of the Tansley suite:

(1) downloading sequences from online databases and parsing them for submission into the sequence data set;
(2) clustering sequences atan 80% similarity value;
(3) clustering sequences at a 93% similarity value, or subclustering;
(4) adding new sequences to the clusters, which can facilitate the identification of new environmental sequences.

These diverse tasks can now be covered by one large-scale script written in the Perl programming language.

A related project focuses on community ecology of wood-decay fungi. Sequences generated in this project have been used to demonstrate the automated process of updating the cluster dataset with newly-generated environmental sequences.

Methods

Informatics

First, all available fungal ITS sequences were downloaded from the NCBI's GenBank, which is a database of publicly available DNA sequences. These sequences were categorized as "environmental" or "specimen-based" submissions based on a number of keywords found in their database entry. Additionally, the variable ITS1 and ITS2 spacer regions were extracted from less variable gene regions that code for structural RNA so that they could be examined independently.

Second, sequences were clustered at 80% similarity. In other words, all sequences in a cluster were within 80% similarity of a single arbitrarily chosen reference sequence. To generate each cluster, one sequence was compared to all other non-clustered sequences using the program ClustalW. If the comparison gave a result greater than 80% similarity, it would be joined to the cluster.

This process was repeated, forming subclusters of 93% similarity within each more inclusive 80% group.

Data generation

Fungal DNA was extracted directly from wood and from fungal subcultures that grew from decaying wood onto agar media. The highly variable ITS region of fungal ribosomal DNA with primers ITS1 and ITS4 was amplified with PCR and sequenced. Sequences were edited in the software program Sequencher (Figure 1).

Clustering wood decay sequences

The ITS sequences generated from wood were compared to the reference sequence of each cluster in the dataset. The sequence was placed into the most similar cluster, provided the similarity score was above 80%. This was repeated for the subclusters at 93% similarity. In both cases, ClustalW was used again (Figure 1).

Contact
aohman@clarku.edu
dglotzer@clarku.edu
dayoung@clarku.edu

Poster created for Fall Fest 2010 at Clark University, Worcester, MA

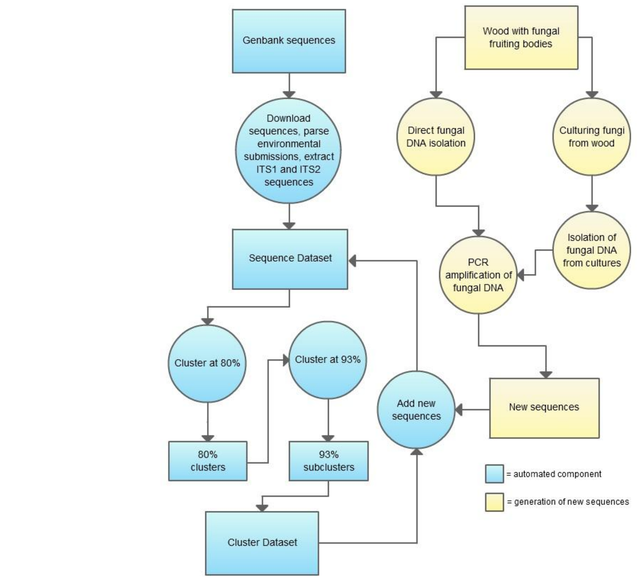


Figure 1: Flow chart describing the informatics (blue) and data collection (yellow) components of the project.

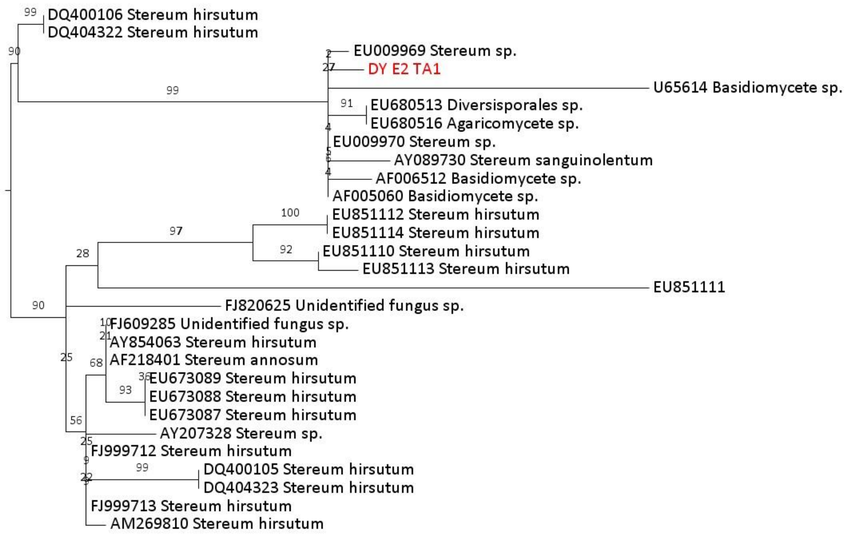


Figure 3: Phylogenetic tree of cluster 17231 at 93% similarity (highlighted in yellow in Table 1). Red sequence was generated in this study.

Figure 4: *Stereum sanguinolentum* (commanster.eu)

Figure 5: *Stereum hirsutum* (mykoweb.com)

Acknowledgements

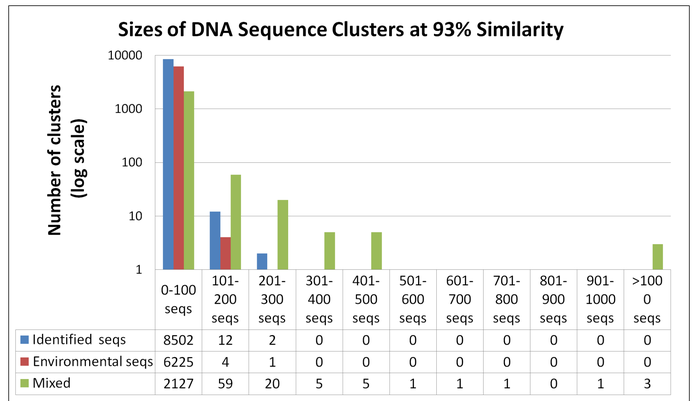
We thank R. Henrik Nilsson and the NSF Assembling the Fungal Tree of Life grant (DEB-0732968 to David Hibbett).

References

Nilsson RH, Kristiansson E, Ryberg M, Larsson K. 2005. Approaching the taxonomic affiliation of unidentified sequences in public databases – an example from the mycorrhizal fungi. *BMC Bioinformatics* **6:**178.

Nilsson RH, Bok G, Ryberg M, Kristiansson E, Hallenberg N. 2009. A software pipeline for processing and identification of fungal *ITS* sequences. *Source Code for Biology and Medicine* **4:**1.

Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH. 2009. An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. *New Phytologist* **181:** 471-477.



	0-100 seqs	101-200 seqs	201-300 seqs	301-400 seqs	401-500 seqs	501-600 seqs	601-700 seqs	701-800 seqs	801-900 seqs	901-1000 seqs	>1000 seqs
Identified seqs	8502	12	2	0	0	0	0	0	0	0	0
Environmental seqs	6225	4	1	0	0	0	0	0	0	0	0
Mixed	2127	59	20	5	5	1	1	1	0	1	3

Figure 2: Histogram detailing the distribution of cluster sizes containing Identified and/or environmental sequences on a logarithmic scale.

Results and Conclusions

Informatics

Out of 91,225 total clustered sequences 30,217 sequences were listed as environmental or unidentified in GenBank, meaning that about one-third of the fungal sequences in GenBank are not identified to species level. This highlights an advantage of the clustering portion of Tansley because it provides more information about sequences by matching them with related sequences based only on similarity and independent of name, which facilitates inferring evolutionary relationships.

For example, at the 93% similarity level, there were 16,969 total clusters. Of these, 6,230 contained exclusively environmental or unidentified sequences, while 2,223 clusters contained a mix of identified and non-identified sequences. Therefore, the Tansley script can identify potential relationships among many of the previously unidentified environmental sequences in GenBank.

Data generation and clustering of wood decay sequences

56 newly generated fungal sequences isolated from wood were grouped into 25 clusters at 93% similarity (Table 1). Several of these clusters did not include sequences from GenBank. This suggests that some of the newly isolated sequences come from fungal species that are not yet described in GenBank. Figure 3 is a phylogenetic tree depicting the relationships among sequences in cluster 17231 (highlighted in yellow in Table 1). The newly-generated sequence in this cluster, shown in red, is most closely related to species like *Stereum sanguinolentum* (Figure 4) and *Stereum hirsutum* (Figure 5). These results demonstrate how the clustering pipeline streamlines the process of gathering related sequences for identification and analysis of new sequences.

cluster ID	cluster ID at 80%	cluster ID at 93%	total # of sequences in cluster	# new wood decay sequences in cluster	example species
16975	136	6	222	2	*Fusarium mangiferae*
17120	264	2	27	1	*Umbelopsis isabellina*
17129	589	5	19	1	*fungus sp.*
17187	598	5	23	2	*Mortierella hyalina*
17198	601	2	69	6	*Mortierella gamsii*
17203	601	7	1	1	
17227	628	24	1	1	
17231	694	4	30	1	*Stereum hirsutum*
17236	694	9	1	1	
17243	1053	7	6	2	*Mucor hiemalis*
17248	1213	5	4	1	*Podospora glutinans*
17265	1603	5	28	5	*Coniophora olivacea*
17269	1803	3	7	7	
17270	1803	4	4	4	
17271	1803	5	1	1	
17274	1866	3	10	1	*Mucor hiemalis*
17280	2037	2	4	2	*Trichaptum abietinum*
17286	2405	6	1	1	
17287	3778	1	10	4	*Hyphoderma puberum*
17290	4776	2	1	1	
17292	5268	2	2	2	
17293	5268	3	6	6	
17294	5268	4	1	1	
17296	5514	2	1	1	
17298	6114	2	1	1	

Table 1: Description of clusters containing newly generated fungal sequences isolated from wood. Sequences in cluster 17231, highlighted in yellow, were made into a phylogenetic tree (Figure 3).