Amelia Antrim

November 25, 2013

**A Program for Estimating the Taxonomic Category of an Organism**

**Given a Small Stretch of its DNA or Amino Acid Sequence**

For centuries, a central interest of naturalists, biologists, and zoologists has been the categorization of organisms. Early naturalists were interested primarily in discovering and identifying new types, or species, of organisms. A hierarchical system of classification was forwarded by Swedish medical doctor and botanist Carolus Linnaeus in his 1735 book *Systema Naturae*.[[1]](#footnote--1) Linnaeus gave each species a Latin name in the form of binomial nomenclature, consisting of the genus and species categories. He further grouped genera into broader categories, called families and orders. Linnaeus was a creationist and believed that these natural orders of species demonstrated the perfection of Biblical design and creation. While this idea was overturned by the concepts of natural selection and branching evolution in Charles Darwin’s *On the Origin of Species* in 1859,[[2]](#footnote-0) the Linnean system of classification has remained relatively unchanged. [[3]](#footnote-1)

An extension of the theory of evolution via natural selection is the common descent of related organisms; species rise from other species. Following from this observation, Darwin hypothesized that all organisms may have risen from a single common ancestor: “I should infer from analogy that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed."[[4]](#footnote-2) Viewing the diversity of life as such, one can visualize the diversification of species as a branching tree of evolution, with each clade, or group of organisms consisting of a common ancestor and its descendants, being a branch on the tree. The practice of classifying organisms and their common descent based on common traits is known as taxonomy. However, the ability to sequence DNA and RNA has led to a revolution in classification. Due to convergent evolution, in which organisms under similar selection pressures evolve similar morphology, and homology, in which similar structures in distantly related organism does not reveal common descent, the underlying genetic information of organisms has proven a more reliable method of estimating common descent than traditional morphological taxonomy. Thus, the science of phylogenetics has been developed to categorize organisms based on molecular sequencing data and common descent.

With the rise of genetics and bioinformatics, many tools have been developed to compare sequences with reference sequences from organisms whose full or partial genomes have been sequenced. One such tool is the National Center for Biotechnology Information’s (NCBI) BLAST. BLAST, or Basic Local Alignment Search Tool, which searches a genome database to find “regions of local similarity between sequences” by comparing nucleic acid or amino acid sequences to predetermined sequence databases and calculating statistical significance of these matches.[[5]](#footnote-3) However, BLAST is primarily designed for aligning sequences derived from a known organism with a section of the genome of that organism; that is, BLAST is useful for determining changes and origins of gene sequences in an already-identified organism, rather than determining the organism’s phylogeny based on genetic information. Additionally, BLAST can only search for matches within its defined set of sequenced genomes, which consists of 226 organisms (52 vertebrates, 17 invertebrate animals, 19 protozoans, 121 plants, and 17 fungi). [[6]](#footnote-4) While this is certainly a sizeable sampling, with new studies estimating 8.7 million eukaryotic and 105 to 107 prokaryotic[[7]](#footnote-5) species currently living on Earth[[8]](#footnote-6), and many more species that have gone extinct, the vast majority of testable species are not represented in this database. Furthermore, NCBI currently does not provide a tool for estimating phylogeny based on phylogenetic information.

Recently, a procedure called DNA barcoding has been developed to identify the species of an animal or plant using a roughly 600-base pair region of its mitochondrial DNA or chloroplast DNA, respectively. However, the “barcode” of an organism is a specific region in the mitochondrial CO1 gene (cytochrome c oxidase 1), or, in the case of plants, two chloroplast gene regions called matK and rbcL.[[9]](#footnote-7) Because of this, you would need to have access to a complete mitochondrial genome of the organism or enough of the genome to be able to access and identify that gene and isolate it from the rest of the mitochondrial genome. Furthermore, this method is specific to plants and animals, excluding fungi, single-celled protozoans, and all prokaryotes.[[10]](#footnote-8)

Therefore, I have decided to implement a tool that estimates the phylogeny of an organism based on a short segment of its genetic information for my final research project. Such a program has many potential utilities distinct from the capabilities of both BLAST and DNA barcoding. For example, running this program on a small amount of fossil DNA may help elucidate from which type of organism this DNA came. Additionally, a refined version of this program could help identify the origin of genes for organism for which good annotated reference genomes have not yet been developed.[[11]](#footnote-9) Unlike barcoding, while this program is capable of categorizing existing organisms into species groups, the intention of this program is to examine sequence origins within a genome and to categorize an organism that may not be an identified species or have any sort of annotated sequence data on record. This program has widespread other applications; for example, due to retroviral incorporation of viral DNA into human germ cells, the human genome is composed of approximately 8% inserted viral genetic material.[[12]](#footnote-10) Given a short segment of human DNA, this program might be able to determine whether the sequence is of viral or human descent.[[13]](#footnote-11)

**My Program**

For this project, I have developed a program, which I refer to as a “taxonomic classifier”. My program is adaptable, allowing users to use a default training set which I have hand-gathered, or to input their own training files. I have allowed the user to build their training dataset from scratch, but I have also included getter and setter methods, as well as other methods I thought maybe useful for the user to write their own main method or driver class for more personalized and flexible functionality.

I represented each taxonomic group as an object that include the group’s name, the larger superclass that contains this class (which I have dubbed the “parent class”), a list of subclasses that constitute this class (“child classes”), the training dataset, the files containing genomic data about this taxonomic group, and, after training, a bigram model of that taxon’s data.

Once a training dataset is built, the program allows the user to input the name of FASTA-format “unknown” sequences. The file is parsed, and the resulting dataset will be compared to a taxonomic hierarchy represented by Taxonomic Group objects. Each taxonomic group object contains information about parent and child classes, allowing traversal of the taxonomy. Therefore, once the most probable option has been chosen from a set of same-level groups, the unknown sequence will then be compared to the child groups of the most likely group, thus traversing the taxonomy from the root to the most specific group represented in the training dataset.

To gather genomic information for my project, I chose to use the NCBI database, as NCBI provides one of the most complete set of bioinformatic databases freely available to the public. In particular, the RefSeq database contains an extensive collection of non-redundant genomic reference sequences, which can be downloaded in multiple formats.[[14]](#footnote-12) I haphazardly selected organisms to represent the phylogenetic groups from which I drew my training dataset.

I have used this hand-selected dataset to test multiple “unknown” test sequences. However, I chose to write into my program a class representing phylogenetic categories and methods allowing the user to create their own training dataset specific the problem they choose. I had originally considered implementing a system of random training sequence retrieval from the NCBI RefSeq database; however, lack of taxonomic information on the RefSeq website and strong bias in organism type made this unfeasible. Perhaps as more organisms’ genomes are sequenced, RefSeq will present a more representative database of genomic sequences, but at present, hand-picking a representative training set remains more feasible within the scope of my project.

For my training dataset, I used short subsets of the genomes of the following organisms, within the hierarchy depicted below:

Cellular Organisms

* Bacteria (Bacteria)
  + *Bacillus thuringiensis* (bacterium commonly used as pesticide)
* Eukaryota (Eukaryotes)
  + Metazoa (Animals)
    - Mammalia (Mammals)
      * *Pan troglodytes* (chimpanzee)
      * *Mus musculus* (mouse)
    - Arthropoda (Arthropods)
      * *Drosophila melanogaster* (fruit fly)
      * *Anopheles gambiae* (mosquito)
      * *Apis mellifera* (honey bee)
    - Amoeba (Amoebas)
      * *Dictyostelium discoideum* (slime mold)
      * *Chaos carolinense* (giant amoeba)
      * *Amoeba proteus* (no common name - amoeba often used in science demonstrations)
    - Actinopterygii (ray-finned fish)
      * *Oreochromis niloticus* (Nile tilapia)
      * *Danio rerio* (zebrafish)
  + Embryophyta (Plants)
    - *Arabidopsis thaliana* (common research plant closely related to Chinese cabbage)
    - *Populus trichocarpa* (California poplar)
  + Fungi (Fungi)
    - *Saccharomyces pastorianus* (yeast species)

Because phylogenetic categories are currently a hotly-debated topic given the rise of cheap gene sequencing, I have chosen to use the classifications presented on the NCBI’s Taxonomy database. As the NCBI refers to these classifications as “taxonomy,” I have chosen as a convention torefer to the categorizations as “taxonomy” rather than “phylogeny” both in my program and henceforth in this paper.

My genomic file format of choice is FASTA, as these files are easily obtained from the NCBI databases and are one of the most prevalent formats for genomic data storage. I have written methods solely to parse FASTA files in this code. However, the code could be easily adapted to utilize different genetic file types using the BioPython package ([www.biopython.org](http://www.biopython.org)).

**Calculations**

To determine the most likely phylogeny, I have used a sample tree of taxonomic categories in which to place my organisms (Figure 1). In my test runs of my program, I chose to create a more detailed taxonomy and test organisms primarily within Eukarya, due to the prevalence of genomic research on eukaryotic organisms, and the relatively well-established evolution of these species. Due to the incomplete and dynamic nature of prokaryotic taxonomy,[[15]](#footnote-13) categorizing prokaryotes using this program is possible yet problematic. Additionally, while I have created a test sequence that differentiates between eukaryotic and bacterial DNA and amino acid sequences, the small, circular chromosomal genomes of prokaryotic organisms makes distinguishing between the two groups trivial in many situations. However, the ability to discern between eukaryotic segments and segments of non-eukaryotic origin remains important when dealing with questions such as the incorporation of viral information into the human genome, as viruses are neither prokaryotes nor eukaryotes (and are debatably not living at all).

Additionally, I chose to limit my testing to taxonomic groups with published non-mitochondrial reference sequences available for more than one species. I excluded mitochondrial and chloroplast DNA in this run of the test due to the potential prokaryotic origins of mitochondrial and chloroplast DNA. The endosymbiotic theory, forwarded by Dr. Lynn Sagan in 1966, states that eukaryotic cells may have originated when one prokaryotic cell was engulfed within another prokaryotic cell during endocytosis, and the two began living together in a symbiotic relationship.[[16]](#footnote-14) Under this view, these engulfed prokaryotes gave rise to mitochondria and chloroplasts, resulting in numerous similarities between prokaryotic genetic information and the sequences contained within the mitochondria and chloroplasts. As such, I excluded mitochondrial and chloroplast DNA from my training dataset to avoid obfuscating the distinction between prokaryotic and eukaryotic gene sequences. However, this program is designed so that the user may build and customize the tree to fit his or her own research needs.

**Challenges and Future Directions**

While this program effectively categorizes three of my four test organisms (the giant amoeba was categorized as a fungus, but still was correctly placed within Eukaryota), I encountered several challenges, and there are undoubtedly ways that the program could be improved over time.

The first and most daunting challenge I encountered was navigating the NCBI resources. These databases contain tremendous amounts of information, but many properties of the websites are not well documented. Determining out how to obtain the information necessary to build a training set and select small segments of DNA from whole-genome shotgun sequences proved a daunting task.

The first challenge facing the implementation of this program is the current incompleteness of the genomic information available. Likely only a small fraction of the species on earth have been discovered, and, of those known species, even fewer have available sequenced reference genomes. I had hoped, for example, to represent many of phyla within the Metazoa (animal) kingdom, but several groups I had hoped to include, such as Mollusca (mollusks) and Annelida (segmented worms), as these phyla include many diverse and interesting organisms; however, no chromosomal sequences were available on RefSeq’s genome database for these phyla. Additionally, all chromosomal data was of very large size, making training n-gram models on this data impossible on my computer within a reasonable amount of time. Even downloading the sequences and haphazardly cutting them down to a reasonable size took a significant amount of time. A helpful tool in implementing a project like this would be a premade set of n-gram models for given taxa, preventing the necessity of gathering the data manually, or at the least a convenient way to retrieve random short segments of a sequence. Finally, given a high-performance computer, I would like to run my analyses on much larger datasets. Once I have gained more skill in computer science and refined this method of categorization, I hope to implement an interactive interface for this program, making the application more user-friendly as well as more accurate.

Additionally, there are drawbacks to categorizing organsisms based on a randomly acquired segment of DNA. Some regions of DNA such as homeobox (HOX) genes are well-conserved and remain relatively unchanged among a wide variety of organisms. Depending on which genes are used for the training data and which genes are acquired as the sample to categorize, the similarity between genes could potentially outweigh the similarity among organisms, causing a miscategorization of the sample. Thus, given enough computing power, the data would ideally be trained on entire genomes, or at least entire chromosomes, and would encompass a broader variety of organisms.

While I concede caveats and flaws in my particular implementation of this program so far, I believe that my application is a step toward a method of categorizing an organism and that a program built upon these principles could, if finely-tuned enough, eventually accurately categorize an organism based on genetic information, and I hope to improve my project over time.

**Appendix: A Guide to Interpreting the Output of this Program**

The print taxonomy method of the unknown sequence class prints out the estimated taxonomy, beginning from the broadest categorization and progressing to the most specific. The print method also gives the perplexity of the data as compared to each subsequent taxonomic group. I have attached a test run of my program on several genomic samples. For clarity, I have written a small header on each test run giving the actual taxonomy of the sequence, since my test sequences were known to test the performance of this program. Humans, common carp, and Staphylococcus bacteria were correctly categorized, while the giant amoeba was misclassified as a fungus but was correctly placed within the more broad domain of Eukaryota. Output is shown below:

The taxonomy of a human (Mammalia) is estimated as follows:

The perplexity of this organism's genome when compared to Eukarya is 3.88734593957

The perplexity of this organism's genome when compared to Metazoa is 3.88779645956

The perplexity of this organism's genome when compared to Mammalia is 3.88250879765

The taxonomy of Staphylococcus aureus (Bacteria) is estimated as follows:

The perplexity of this organism's genome when compared to Bacteria is 3.80761733257

The taxonomy of the common carp (ray-finned fish) is estimated as follows:

The perplexity of this organism's genome when compared to Eukarya is 3.92766225573

The perplexity of this organism's genome when compared to Metazoa is 3.92960130126

The perplexity of this organism's genome when compared to Actinopterygii is 3.94319357506

The taxonomy of the giant amoeba is estimated as follows:

The perplexity of this organism's genome when compared to Eukarya is 3.90615895904

The perplexity of this organism's genome when compared to Fungi is 3.90998119253

1. O’Neil, Dennis. *Introduction to Linnean Classification*. 2012. <http://anthro.palomar.edu/animal/animal_1.htm>. Nov 2013. [↑](#footnote-ref--1)
2. Darwin, Charles and William Bynum. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life / edited and with an indruction by William Bynum. 2009. [↑](#footnote-ref-0)
3. O’Neil, *Introduction to Linnean Classification*. [↑](#footnote-ref-1)
4. Darwin, On the origin of species. [↑](#footnote-ref-2)
5. NCBI Basic Local Alignment Search Tool (BLAST). <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. [↑](#footnote-ref-3)
6. *Blast Homepage and Selected Search Pages*. National Center for Biotechnology Information, 2013. <ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf>. Nov 2013. [↑](#footnote-ref-4)
7. Whitman, William B., David C. Coleman, and William J. Wiebe. *Prokaryotes: The unseen majority.* Proceedings of the National Academy of Sciences, 1998. <http://www.pnas.org/content/95/12/6578.full#ref-88>. Nov 2013. [↑](#footnote-ref-5)
8. Sweetlove, Lee. *Number of species on Earth tagged at 8.7 million.* Nature, 2011. <http://www.nature.com/news/2011/110823/full/news.2011.498.html>. Nov 2013. [↑](#footnote-ref-6)
9. Hebert, Paul. *What is DNA Barcoding?* Consortium for the Barcode of Life, 2013. <http://www.barcodeoflife.org/content/about/what-dna-barcoding>. Nov 2013. [↑](#footnote-ref-7)
10. Pearson, Helen. *DNA barcodes tag species.* Nature, 2004. <http://www.nature.com/news/2004/040927/full/news040927-2.html>. Nov 2013. [↑](#footnote-ref-8)
11. Andrew Kitchen, University of Iowa Anthropology Department, Personal Communication, October 2013. [↑](#footnote-ref-9)
12. Feschotte, Cedric. *Virology: Bornavirus enters the genome.* *Nature* **463**, 39-40 (7 January 2010). <http://www.nature.com/nature/journal/v463/n7277/full/463039a.html>. Nov 2013. [↑](#footnote-ref-10)
13. Gevorg Gregoryan, Dartmouth College Computer Science Department, Personal Communication, October 2013. [↑](#footnote-ref-11)
14. NCBI Reference Sequences (RefSeq). http://www.ncbi.nlm.nih.gov/refseq/. [↑](#footnote-ref-12)
15. http://onlinelibrary.wiley.com/store/10.1111/j.1574-6976.2001.tb00571.x/asset/j.1574-6976.2001.tb00571.x.pdf;jsessionid=7527D0A6870E02498086F53253E4884C.f04t03?v=1&t=hofya0sr&s=157b14f23743cb37345a82a6b7718b71559aa602 [↑](#footnote-ref-13)
16. Sagan, Lynn. On the origin of mitosing cells. Journal of National Institute of Health Research 1993 Mar;5(3):65-72 (orig. 1967). <http://www.ncbi.nlm.nih.gov/pubmed/11541392>. Nov 2013. [↑](#footnote-ref-14)