Spring 2018

Genetic Barcode Identification With Profile Hidden Markov Models

Vishrut Sharma
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Part of the [Computer Sciences Commons](#)

Recommended Citation

Sharma, Vishrut, "Genetic Barcode Identification With Profile Hidden Markov Models" (2018). *Master's Projects*. 603.

DOI: <https://doi.org/10.31979/etd.9fn2-bg55>

https://scholarworks.sjsu.edu/etd_projects/603

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Genetic Barcode Identification With Profile Hidden Markov Models

A project report presented to
The Department of Computer Science
San Jose State University

In partial fulfillment of the requirements for the class

CS 298

By Vishrut Sharma

May 2018

The Designated Project Committee Approves the Project Titled

Genetic Barcode Identification with Profile Hidden Markov Models

by

Vishrut Sharma

APPROVED FOR THE DEPARTMENTS OF COMPUTER SCIENCE

SAN JOSE STATE UNIVERSITY

May 2018

Dr. Philip Heller	Department of Computer Science
Dr. Laura Miller Conrad	Department of Chemistry
Dr. Robert Chun	Department of Computer Science

1. ABSTRACT

Genetic Barcode Identification with Profile Hidden Markov Models

by Vishrut Sharma

DNA barcoding is a method that uses an organism's DNA to identify its species. The gene cytochrome c oxidase I (COI) has been used effectively as a DNA barcode to identify organisms and elucidate relationships among species [1]. There also exists a database BOLD (Barcode Of Life Database) that contains COI sequences used for DNA barcoding for more than 1 million different species. Using BOLD to identify samples that have a match in the database is an uncomplicated process. However, this method fails to determine samples that are absent from the database. Given a sample that is not represented in BOLD but is similar to a represented sequence, it would be valuable to describe the sample at a higher taxonomic classification. Since COI is represented as long character sequences of amino acids, Hidden Markov Models (HMMs) can be used to associate an unknown DNA sequence with a taxonomic rank. In this work, I show that dynamically created Profile HMMs are an effective tool for such identification.

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to my advisor Dr. Philip Heller for his guidance, feedback, and patience throughout the course of this project. His mentorship was invaluable and without it, this project would not be possible. I want to thank my committee members, Dr. Laura Miller Conrad and Dr. Robert Chun for investing their time in this project and providing their valuable feedback. I am grateful to my friends who provided suggestions for my work. Finally, I would like to thank my parents. They have been my source of strength and motivation throughout this project and my life in general.

Table of Contents

1. BACKGROUND	4
1.1 History of Biological Classification.....	5
1.2 Drawbacks of Linnaean System.....	8
1.3 DNA Barcoding	9
1.4 Cytochrome C Oxidase (COI).....	11
2. METHOD	16
2.1 Goal.....	16
2.2 Preparation	16
2.3 Profile Hidden Markov Model Pipeline.....	16
2.4 Approach.....	17
3. RESULTS	19
4. DISCUSSION	28
4.1 Future Work	28
4.2 Conclusion	28
5. REFERENCES	30
6. APPENDIX.....	33
6.1 Hidden Markov Models	33
6.2 Multiple Sequence Alignment	37
6.3 Profile Hidden Markov Model.....	41

List of Figures

Figure 1– A sequence match found in BOLD [8]	13
Figure 2 - pHMM pipeline	17
Figure 3 – Simulation for Unknown Sequence	18
Figure 4 - Output After Building Tree (Children of Phylum Chordata)	20
Figure 5 - Output After Building Tree(Children of Family Bactrocera).....	20
Figure 6 - pHMM of Bactrocera (code of Dr. Philip Heller)	21
Figure 7 - Scores for Species of Carollia	21
Figure 8 - Scores for Species of Bactrocera.....	22
Figure 9 - Scores for Species of Mytilus.....	23
Figure 10 - Scores for Species of Lingula.....	23
Figure 11 - Scores for species of Watersipora	24
Figure 12 - Scores for species of Clathria.....	24
Figure 13 - Scores for Species of Zoanthus	25
Figure 14 - Scores for Species of Steinernema	25
Figure 15 - Scores for Species of Synchaeta.....	26
Figure 16 - Scores for Species of Paranoplocephala.....	26
Figure 17 - Scores for Species of Henricia	27
Figure 18 - Scores for Species of Prionospio.....	27
Figure 19 - Markov Process	34
Figure 20 - Hidden Markov Model	35
Figure 21 - Multiple sequence alignment using Clustal X.....	38
Figure 22 - Process of MSA Inside Clustal.....	40
Figure 23 - Match States in a PHMM	41
Figure 24 - Profile Hidden Markov Model	42
Figure 25 - Example Multiple Sequence Alignment.....	42
Figure 26 - PHMM Example	43

1. BACKGROUND

To classify means to assign anything to a group. In terms of biological classification, living organisms are categorized with other organisms that they are most closely related to. These small groups are then classified together to form larger groups and so on. This strategy of grouping allows scientists to observe relationships between closely related organisms and make sense of a hugely diverse array of life. In order to represent a maximum number of living things, biological classification follows a hierarchical structure. Every level in the hierarchy is called taxon and can range from very broad (include several different living organisms) to very specific (identify individual life forms). The science for naming and classifying living things is called Taxonomy.

It is estimated that there are at least 8.7 million different forms of life on earth [1]. Identifying, organising and naming these organisms is a daunting task. Taxonomic classification establishes a standard system that allows scientists to coordinate naming and grouping of life forms as without it organisms will be classified differently in different parts of the world. Following a common method for identification makes studying organisms easier and improves the speed of identification. This classification also shows the interconnections between different groups of organisms thereby demonstrating intimate evolutionary relationships. A standard system facilitates in identifying the common ancestors of different groups and study diversity within these groups.

1.1 History of Biological Classification

Greek philosopher Aristotle (384-322 B.C) was one of the greatest thinkers of his time and is well known for his contribution in the fields of politics, ethics, and psychology. He is also regarded as a taxonomist (people who study taxonomy) and devised a system of classification that divides living things into plants and animals. His work called “History of Animals” (*Historia Animalium* in Latin) grouped animals based on their habitat (land, air, and water) and divided plants into subgroups based on size (small, medium and large). He viewed all life to fit into a hierarchy that travelled from lowest to highest, with human beings at the top of the hierarchy. Aristotle’s view of life forms did not consider evolution as the basis of hierarchy and hence it had several organisms that did not fit well into his system of classification. Amphibians like frogs can live on both land and water and hence were misplaced in Aristotle’s hierarchy. He also grouped diverse specimens like birds, bats, and insects into the same group simply because they can fly. His system gave importance to the behaviour of living things instead of their innate similarities and differences. Intellectual capacity was also a factor in his system, hence plants were put at the bottom of the rank and animals with some form of reasoning were put above. Despite these serious conceptual flaws and misclassifications, Aristotle’s classification system existed for nearly 2000 years.

The English botanist John Ray (1627- 1705) is considered the father of natural history for his wide-ranging contributions in the field of botany. He also made significant contributions in taxonomy and fixed “species” as the basic unit of taxonomy. The basis of his classification system was first published in 1682 called “*Methodus Plantarum Nova*” and talked about the differences observed in monocotyledons (plants

that germinate with a single leaf) and dicotyledons (plants that germinate with two leaves). Building on this work Ray published 3 volumes that aimed to cover the evolution of plants titled *Historia Plantarum*. Instead of using behaviour and mental complexity as the basis of classification John Ray stressed on visible structural differences as the means of groupings. He studied anatomy in great detail and insisted on using lungs and cardiac organs too for performing classification. He classified about twenty thousand organisms based only on descriptive characters like nails, teeth, claws, and toes. His work showed the importance of morphology as the foundation for classification and paved way for better taxonomy.

Carl Linnaeus (1707 – 1778) a Swedish scientist is regarded as the father of modern taxonomy and his thoughts on taxonomic classification have molded much of our understanding about living organisms. In 1735 he published *Systema Naturae* that gave a concrete framework for the hierarchy of all living beings. He published 10 editions to this work, describing in detail the differentiating aspects studied for every level in his hierarchical system that is being used to date. Linnaeus introduced the concept of nested grouping that divided an organism into 7 groups. He concluded that only one form of an organism exists in the last group – species. Linnaeus studied structural similarities between organisms and created hierarchies based on morphology. In his hierarchy, the group at the top is the broadest and disintegrates into narrower sets as we move down the hierarchy. The position of an organism inside the hierarchical rank defines its relationships to other organisms and identifies the unifying features of the specific rank. Table 1 shows the breakdown for classification of human beings in the Linnaean system.

Classification Level	Name	Characterized by
Domain	Eukarya	Nucleus
Kingdom	Animalia	Multicellular, no- cell wall, ingest food
Phylum	Chordata	Spinal Chord
Class	Mammalia	Nurse offspring
Order	Primates	High intelligence
Family	Hominidae	Walk upright
Genus	Homo	Human
Species	<i>Homo sapiens</i>	Modern Human

Table 1: Classification of Humans [12]

Another big contribution of Carl Linnaeus to taxonomy is the formalization of a naming scheme for organisms known as Binomial Nomenclature. Before this scheme organisms were identified using long Latin names that were difficult to remember. Most of the names were not universal and would change for the same organism found at a different location. Binomial nomenclature solved many issues by providing a

universal naming scheme that aided in classification. This system names species using a 2 part scientific name. The first word is the genus of the organism and the second word describes a characteristic of the organism and is called specific epithet. This identifies a particular species as separate from others belonging to the same genus. Hence, a species name is a combination of the genus name and a specific epithet. Conventionally the first letter of genus is capitalized and the first letter of epithet is lowercase. For example the scientific name for human beings is *Homo sapiens*. This tells us that humans are part of genus *Homo*, and *sapiens* separates us from other species inside the genus like *Homo neanderthalensis*.

1.2 Drawbacks of Linnaean System

The Linnaean system focussed on morphological characteristics for identifying species. Only about 2 million species have been identified by humans till date. Specific morphological traits like size, shape, colour of organisms are used for identifying species, however, these traits are not reliable. Traditionally, taxonomists generate diagnostic keys by observing structural features and share these keys with other biologists who use them to identify unfamiliar species. Morphological features are often difficult to observe and undergo a change depending on the life stage and gender of the organism. For example, *Diptera* is identified largely based on male genitalia. The female will often not be identified because of the absence of a universal morphological marker. Markers are often not uniform across multiple environments and have a high degree of plasticity. Eg. Birds change the color of their feathers depending on seasons. The Dominance of a particular marker varies on individual specimens making them difficult to see. They cannot be used as keys for sibling and

cryptic species which have same morphological markers but belong to separate species. In the early 1900s, wrong classifications based on morphological markers prevented authorities from controlling the spread of malaria in Europe. Many times distinct species also exhibit similar appearance when adapting to a particular environment (convergent evolution). The collected specimens require a great deal of scrutiny and diligence to preserve the distinguishing features. Highly trained professional taxonomists are required to differentiate subtle differences between closely related species.

1.3 DNA Barcoding

The last decade has seen a shift from traditional morphology-based identification procedures to a robust technique called DNA barcoding. DNA is a molecule that encodes genetic information of an organism and it is always unique to a species. DNA barcoding uses a short standardized region of DNA as a marker to accurately identify species [1]. It mitigates the problems posed by identification using morphological characters (taxonomic impediment) [2]. The genetic marker behaves like a barcode similar to that found on products in a supermarket and makes species identification as simple as scanning the barcode of products inside a supermarket.

The idea of genetic barcodes gained attention in 2003 after Paul D Hebert a Professor at the University of Guelph published a paper titled "Biological identifications through DNA barcodes"[1]. He identified specific genes that can serve as a global bio-identification marker. This improves the accessibility of the Linnaean taxonomic system by allowing non-taxonomists to study and identify species [2]. His aim was to build a system that would accelerate the rate of species discovery allowing faster

sorting of specimens and find divergent taxa that can indicate new species [3]. The efforts of Herbert in tandem with other researchers resulted in the creation of an international initiative called the Consortium for the Barcode of Life (CBOL). Its main aim is to guide developments in DNA barcoding. In 2005, CBOL setup the Barcode of Life Data System (BOLD) to create a DNA barcode library, an open source platform to analyse DNA sequences. More than 200 organizations from 50 countries are now part of CBOL and share their data in the public database [4]. The database currently contains over 5.9 million DNA sequences from 542000 species.

Even if morphological markers are easily available DNA barcoding improves biodiversity data sources by being faster and cheaper. It is estimated that the cost of performing DNA barcoding is between \$2 and \$8 per sample [2]. The technology offloads a taxonomist's workload and allows him/her to study more important characteristics of organisms instead of searching for obscure physical markers. A single technician can replace many taxonomists for routine identification and misidentification due to human error is also eliminated. Barcoding technology can also correct the misclassifications that have happened in the past due to too much reliance on morphological characteristics [5]. A good example of this is the skipper butterfly (*Astraptes fulgerator*). This butterfly was discovered in 1775 using traditional morphological markers. Barcoding showed that the butterfly is not one but ten distinct species [5]. The classification of orchids has been revolutionized since the advent of DNA barcoding which was able to recognize 20,000 members of the family. It documented known species of orchids in the process uncovering small variations leading to the discovery of unknown species [6]. Since DNA barcode can be extracted

even from minuscule bits of a specimen including faeces and gut contents, researchers can explore unknown interactions within the food chain [7]. This is practically not possible to do using morphological markings.

DNA barcoding has broad applications in areas of health care, environment, and scientific discovery. It is used to measure food safety by verifying the ingredients inside packaged products and matching them with contents mentioned on their labels [7]. Even after food is cooked, DNA barcodes can be extracted and studied to verify ingredients [9]. The technology aids law enforcement agencies to crack down on illegal trade in endangered species [9]. There also exists a barcode of wildlife project whose aim is to build legal standards for DNA barcodes of endangered and threatened species. DNA comparisons can also be used to measure time periods between stages of evolution [10]. Mutations are a natural consequence of evolution causing minor changes to DNA. The degree of dissimilarity between DNA sequences indicates how long ago specimens shared a common ancestor [3].

1.4 Cytochrome C Oxidase (COI)

Cytochrome C Oxidase is an enzyme present in all Eukaryotes (cellular organisms whose genetic material resides in a nucleus). It is important for ATP synthesis carried out by mitochondria. COI plays a vital role in cellular respiration by breaking down food and releasing energy [4]. COI is called the “ideal DNA barcoding gene” because it can be used to uniquely identify any organism. In 2002, Hebert et al. proposed adoption of the cytochrome *c* oxidase I (COI) gene as a standard for molecular barcoding of animals.

It is used as a DNA barcode for the following reasons.

- Present in all species of organisms (universal)
- Short enough to be quickly sequenced.
- Easy to align and amplify.
- Shows variation across species, but is conserved within a species.
- Easy to obtain even from biological material under hard conditions.

While COI is also present in plants its rate of change is not much and hence it can't be used for DNA barcoding in plants [4]. COI successfully distinguishes 95% of animals and with easy to obtain DNA barcoding technology, it can completely replace morphological identification. BOLD is the go-to database to analyse and research samples of COI sequences. Its identification framework can search the database containing sequences contributed by researchers from across the globe and establish a match [2]. It accepts input as 5' region of COI gene and returns the matching species. The database also maintains all historical copies of COI databases, thereby allowing users to replicate past results. The Figure 1 shows a screenshot of a page indicating a match found for a queried COI sequence. The COI sequence matches with "Aethia cristatella" and BOLD displays the taxonomic hierarchy for the queried sequence.

Figure 1– A sequence match found in BOLD [8]

BOLD primarily uses BLAST (Basic Local Alignment Search Tool) algorithm for comparing COI sequences. Given a query to search for and a database (BOLD) to search against, BLAST finds all subsequences inside BOLD that resemble subsequences of the query. It uses statistical variation in sequences and heuristics to provide fast results [13]. Apart from this BLAST also provides a “expect value” for a query which is a confidence score for the sequence alignment.

In spite of huge support from the biology community around the world, the overall coverage of BOLD remains limited (over 1 million species) due to the sheer diversity of life forms that exist within nature. It aims to have an eventual target size of 100

million records, but efforts to grow are constrained by strict wetlab requirements like DNA isolation, cleaning, and amplification using Polymerase chain reaction (PCR). There exists another large database – GenBank containing collections of nucleotide sequences along with their protein translations[14]. This database is maintained by National Center for Biotechnology Information (NCBI) and accepts sequences from labs throughout the world. It has a fast rate of growth, doubling in size every 18 months and has over 162 million sequences. Hence there are many COI sequences in GenBank database that have not been submitted to BOLD. An adaptation of the ARBtrator algorithm [15], called CO-ARBtrator [16] has recovered over 1 million sequences from GenBank with very low error rates. About half of these sequences were not present in BOLD. By leveraging the data present inside BOLD and GenBank superior data sources can be developed. The Release of the CO-ARBtrator database, coupled with the ongoing growth of BOLD, presents an opportunity to develop identification algorithms that take full advantage of these databases. This work is an initial step in this direction

Here we report successful predictions of unknown COI sequences into the correct genus. To carry out experiments relating to unknown COI sequences we require a large reference database containing labelled sequences. BOLD is freely available for this purpose. High performing algorithms are needed to carry out identification. We explore the use of Profile Hidden Markov Models. Appendix A contains the detailed explanation of Profile Hidden Markov Models for those who are unfamiliar with them.

Building Hidden Markov Models for entire taxonomic hierarchy is computationally intractable. Instead, we can build dynamic HMMs for identifying unknown COI sequences when needed. Here we report promising results of this approach.

2. METHOD

2.1 Goal

Given a COI sequence of an unknown species whose match does not exist inside BOLD identify its genus.

2.2 Preparation

Dataset is a FASTA file containing 103837 COI sequences from BOLD. Every sequence contains taxonomic hierarchy as a comment from the level of phylum up to the level of species. Parse the FASTA file to construct a tree structure resembling taxonomic hierarchy of all sequences.

2.3 Profile Hidden Markov Model Pipeline

- 1) The Input to the pHMM pipeline is a file containing COI sequences of the training set. We deliberately remove one sequence from the genus, which will be later used as our test sequence.
- 2) We perform multiple sequence alignment over the training sequences using tool Clustal Omega. Clustal outputs the multiple sequence alignment to a FASTA file.
- 3) Build pHMM from multiple sequence alignment of the training sequence in java using the code of Dr. Philip Heller. Figure 2 shows the architecture of pHMM pipeline.

Figure 2 - pHMM pipeline

2.4 Approach

Follow below steps in order to measure the effectiveness of a pHMM to identify a new COI sequence in a known genus. Repeat steps to simulate unknown sequences belonging to 12 major Phyla (Annelida, Arthropoda, Bryozoa, Chordata, Cnidaria,

Echinodermata, Mollusca, Nematoda, Platyhelminthes, Porifera, Rotifera, and Brachiopoda).

1. Genus is represented inside tree constructed from the COI dataset.
2. We simulate an unknown sequence by deleting a sequence belonging to the genus.
3. Construct a pHMM trained on remaining sequences of the genus.
4. Construct more pHMMs for each genus in the deleted sequence's family.
5. Evaluate (score) deleted sequence using all pHMMs. Figure 3 shows the approach followed.

Figure 3 – Simulation for Unknown Sequence

3. RESULTS

Dynamic pHMMs correctly identified 92 COI sequences. At least one sequence was correctly identified for 12 phyla.

3.1 Error Rate

error rate = number of wrongly classified sequences / total number of sequences

Out of 106 sequences tested 10 were wrongly classified.

The total error rate for this experiment is 9.43%. Table 2 shows the count of species simulated as unknown COI sequences.

Phyla	Genus	No. of species tested	No. of correct identifications	No. of wrong identifications
Arthropoda	Bactrocera	42	37	5
Brachiopoda	Lingula	3	3	0
Bryozoa	Watersipora	3	3	0
Chordata	Carollia	7	7	0
Cnidaria	Zoanthus	8	8	0
Echinodermata	Henricia	3	3	0
Mollusca	Mytilus	10	10	0
Nematoda	Steinernema	5	5	0
Platyhelminthes	Paranoplocephala	5	4	1
Annelida	Prionospio	10	10	0
Porifera	Clathria	6	6	0
Rotifera	Synchaeta	4	0	4

Table 2 – Count of simulated sequences

Figures 4 – 6 are the outputs obtained from java pipeline before building pHMM.

Figures 7 – 18 are graphs showing scores obtained during testing of individual species against different genera.

```
Tree is built and contains 103837 number of nodes.
time taken to construct the tree is 1 seconds
Enter the node of whose children you want to extract..
Chordata
Finding children of Chordata

Actinopteri
Mammalia
Amphibia
Chondrichthyes
Ascidiacea
Thaliacea
Aves
Cladistia
```

Figure 4 - Output After Building Tree (Children of Phylum Chordata)

```
Tree is built and contains 103837 number of nodes.
time taken to construct the tree is 1 seconds
Enter the node of whose children you want to extract..
Bactrocera
Finding children of Bactrocera

invadens
ruiliensis
dorsalisD
kraussi
opilae
pyrifoliae
arecae
papayae
raiensis
paraverbascifoliae
umbrosa
occipitalis
kandiensis
dorsalisL
- - -
```

Figure 5 - Output After Building Tree (Children of Family Bactrocera)

Figure 6 - pHMM of Bactrocera (code of Dr. Philip Heller)

Figure 7 - Scores for Species of Carollia

Figure 8 - Scores for Species of Bactrocera

Figure 9 - Scores for Species of Mytilus

Figure 10 - Scores for Species of Lingula

Figure 11 - Scores for species of Watersipora

Figure 12 - Scores for species of Clathria

Figure 13 - Scores for Species of Zoanthus

Figure 14 - Scores for Species of Steinernema

Figure 15 - Scores for Species of Synchaeta

Figure 16 - Scores for Species of Paranoplocephala

Figure 17 - Scores for Species of Henricia

Figure 18 - Scores for Species of Prionospio

4. DISCUSSION

Species identification by DNA barcoding requires a reliable dataset of COI sequences. The BOLD database provides a large number of sequences and their classification hierarchy. Our objective was to use Profile Hidden Markov Models to identify the genus of an unknown COI sequence. We simulated unknown sequences spanning across 12 different phyla and evaluated them against dynamic pHMMs to identify their respective genera.

The error rate (<10%) suggests that this approach can be refined into a robust and useful classifier. Most errors appear in the phyla Arthropoda and Rotifera. Outside of those phyla, the error rate is 0.9%.

4.1 Future Work

Investigation into the Arthropoda and Rotifera sequences can explain why most errors appear in those phyla and can suggest refinements to the algorithm that can drive down the error rate.

The experiments can be extended both breadth-first and depth-first into the taxonomy tree in order to identify higher level ranks. The machine learning pipeline can be repurposed to allow more algorithms like support vector machines to examine effects on the error rate.

4.2 Conclusion

Profile Hidden Markov Models can be used to identify unknown COI sequences. The low rates of misclassification achieved establishes the strength of Hidden Markov

Models as an algorithmic technique to combine sequence data spread across multiple databases like BOLD and GenBank.

5. REFERENCES

- [1] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard, "Biological identifications through dna barcodes," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. 1512, pp. 313–321, 2003.
- [2] A. Valentini, F. Pompanon, and P. Taberlet, "Dna barcoding for ecologists," *Trends in Ecology & Evolution*, vol. 24, no. 2, pp. 110–117, 2009.
- [3] P. D. N. Hebert and T. R. Gregory, "The promise of dna barcoding for taxonomy," *Systematic Biology*, vol. 54, no. 5, pp. 852–859, 2005. [Online]. Available: <http://www.jstor.org/stable/20061289>.
- [4] P. D. Hebert, S. Ratnasingham, and J. R. de Waard, "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. Suppl 1, pp. S96–S99, 2003.
- [5] K. J. Gaston and M. A. O'Neill, "Automated species identification: why not?" *Philosophical transactions of the Royal society B: Biological sciences*, vol. 359, no. 1444, pp. 655–667, 2004.
- [6] P. D. Hebert, E. H. Penton, J. M. Burns, D. H. Janzen, and W. Hallwachs, "Ten species in one: Dna barcoding reveals cryptic species in the neotropical skipper butterfly *astraptes fulgerator*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 41, pp. 14 812–14 817, 2004.
- [7] A. Galimberti, F. De Mattia, A. Losa, I. Bruni, S. Federici, M. Casiraghi, S. Martellos, and M. Labra, "Dna barcoding as a new tool for food traceability," *Food Research International*, vol. 50, no. 1, pp. 55–63, 2013.

- [8]"Databases", V3.boldsystems.org, 2018. [Online]. Available:
http://v3.boldsystems.org/index.php/resources/handbook?chapter=2_databases.html.
 [Accessed: 10- May- 2018]
- [9] M. Y. Stoeckle, C. C. Gamble, R. Kirpekar, G. Young, S. Ahmed, and D. P. Little, "Commercial teas highlight plant dna barcode identification successes and obstacles," *Scientific reports*, vol. 1, p. 42, 2011.
- [10] Orlando, Ludovic, et al. "Revising the recent evolutionary history of equids using ancient DNA." *Proceedings of the National Academy of Sciences* 106.51 (2009): 21754-21759.
- [12] Life Science | Session 2", Learner.org, 2018. [Online]. Available:
<https://www.learner.org/courses/essential/life/session2/closer4.html>. [Accessed: 10- May- 2018].
- [13] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*. 1990 Oct 5;215(3):403-10.
- [14] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic acids research*. 2008 Jan;36(Database issue):D25.
- [15] Heller, P., Tripp, H.J., Turk-Kubo, K., and Zehr, J.P. "ARBitrator: a software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank". *Bioinformatics* 30(20): 2883-2890 (2014).
- [16] Heller, P., Casaletto, J., Ruiz, G., and Geller, J. "A COI sequence database, mined from GenBank". *Nature Scientific Data*, 2018 (in prep).
- [17] M. Stamp, *Introduction to Machine Learning with Applications in Information Security*. CRC Press, 2017.
- [18] Meyer, A. "Hmm and part of speech tagging." *Lecture Note 2012* (2011).

- [19] Huang, Xuedong D., Yasuo Ariki, and Mervyn A. Jack. "Hidden Markov models for speech recognition." (1990): 60-80.
- [20] Gilks, Walter R., Sylvia Richardson, and David Spiegelhalter, eds. Markov chain Monte Carlo in practice. CRC press, 1995.
- [21] Multiple sequence alignment. (2018, May 06). In Wikipedia. Retrieved March 15, 2015, from https://en.wikipedia.org/wiki/Multiple_sequence_alignment.

6. APPENDIX

6.1 Hidden Markov Models

A Markov Model uses probability theory for performing predictive analysis.

The Markov process is named after a Russian mathematician Andrey Markov and is described as a “memoryless” random process [17] . It is called memoryless because the probability distribution of events happening in the future does not depend on the sequence of events that happened before the present state. Some of the most popular Markov processes are found in physics demonstrated by the Poisson process and Wiener process, more commonly called Brownian motion. Generally speaking a process is called a Markov process if predictions are possible about the future states of the process based only on its present state [17]. Natural Markov processes contain a set of distinct states with traceable transitions between every state. We can describe processes observed in the environment around us as Markov processes. For example weather prediction and board games that use a dice. The Google page rank algorithm is also considered a Markov process with pages being the different states of the model and links representing the transitions between the processes

Figure 19 shows a simple Markov chain with 3 states A, B,C. There are traceable transitions between the states and the numbers suggest the probability of moving from one state to the next.

Figure 19 - Markov Process

The Markov chain is of excellent utility when we want to find the probability for events that we can observe around us. However, for events that are not directly observable, Markov chains cannot be applied. Example part of speech tagging, where we observe words but want to tag the correct part of speech to the next word. We do not explicitly see any part of speech tags like noun/verb, however we know that they are present [18]. Describing such a model in probabilistic terms is the function of a Hidden Markov Model (HMM).

It is obvious that a Hidden Markov Model contains a Markov chain, but Markov process taking place is not directly observable. What we can observe are symbols that have a direct correlation with the Markov process. Hence an HMM combines two distinct processes.

1. An invisible process of hidden states.

2. A visible process of symbols.

The fig 20 explains the two steps described above.

Figure 20 - Hidden Markov Model

The figure 20 is a generic representation of an HMM with some key notations. We know probabilities of moving from one state to another in a hidden process. This probability is called the state transition probability and is represented by A. We also know how the hidden process relates to the observations (symbols). This probability is represented as B and is known as the observation probability. We also know the initial state distribution i.e. the probability that a Markov process starts in a particular hidden state. This is usually written using the symbol π . Since a HMM is defined using these 3 important properties, it is expressed by the equation

$$\lambda = (A, B, \pi)$$

Ideally using the information we have from a Hidden Markov Model we can solve the following problems [17].

1. Scoring an observation sequence against a Hidden Markov Model.
2. Finding the best state sequence that represents the hidden process in a Hidden Markov Model.
3. Find a model that maximises the probability of occurrence of a given sequence. Also described as training a model to support a sequence of observations.

The three types of problems discussed above that an HMM can solve have a wide variety of applications in the real world. Today Hidden Markov Models are used in speech recognition systems as speech can be encoded as sequence of symbols within an audio range [19]. They are also used in the field of cyber security for identifying malware and benign files [17]. In this case assembly language instructions form the observable symbols and malicious/benign are the hidden states. Another important application of HMMs is gene finding. Here DNA sequences can be regarded as observable symbols in the set of four letters representing nucleotides. The disparity in probabilities for coding and non-coding parts of the DNA are exploited for finding the gene. HMMs were also crucial in the development of another set of models called branching processes that model chain reactions of nuclear physics and chemistry [20].

Hidden Markov Models while being extremely efficient for training and scoring do have some inherent deficiencies. The main deficiency is that HMMs assume the current state relies only on the previous state. This limits us from utilizing any

positional information contained inside the symbols. For many applications especially in the field of bioinformatics positional information is critical.

6.2 Multiple Sequence Alignment

Often in bioinformatics we require simultaneous arrangement and comparison of sequences of protein, DNA, RNA. This is performed to determine evolutionary relationships between biological sequences which point to a common ancestor. By performing a multiple sequence, alignment researchers can study molecular differences and determine sequence homology. The similarities between sequences obtained after performing multiple sequence alignment refer to functional equivalence and evolutionary connections between the organisms of the sequences. Performing phylogenetic reconstruction would be an extremely difficult task without having sequenced alignment.

Simply put a multiple sequence alignment is an alignment of more than 2 sequences.

When similarity is found between only two sequences, the alignment is called pairwise alignment. Broadly there are two types of multiple sequence alignments.

1. Local alignment – identify small regions showing high resemblance among sequences.
2. Global alignment – align sequences end to end showing high level sequence variation.

We can visualize fully aligned sequences as the rows of a matrix. Wherever the alignment between two rows does not match we insert a gap till all successive

columns contain the same characters. Fig 21 is an example of a multiple sequence alignment.

Figure 21 - Multiple sequence alignment using Clustal X [21]

There are 3 methods to perform multiple sequence alignment as mentioned below

1. Dynamic Programming

Dynamic programming divides a problem into smaller subsets and then combines the results from all smaller problems as the final result. The Needleman-Wunch algorithm is a dynamic programming algorithm commonly used to perform global pairwise alignment. One big disadvantage of using dynamic programming for performing multiple sequence alignment is the computation cost which rises exponentially as the number of sequences to align increases.

2. Progressive alignment –

In this method we first perform standard pairwise alignment for two sequences. This alignment is fixed. Following this another sequence is picked and aligned with the existing alignment. This step is iterated over until all sequences are fully aligned. It is important to note here that in progressive alignment the first pairwise alignment is never changed.

3. Iterative refinement.

This method is similar to progressive alignment in the way that at every iteration a new sequence is added onto the existing alignment. The difference here is that the initial alignment is reordered after adding a new sequence to achieve the best alignment.

Multiple sequence alignment is an important step in genetic barcode identification and a third party software – Clustal Omega (www.clustal.org/omega) is used for performing the multiple sequence alignment. Clustal is a command line tool and generates a single file with globally aligned sequences as output. The manner in which Clustal performs the alignment is explained below. Figure 5 shows the steps performed by Clustal to achieve multiple sequence alignment.

1. Clustal performs pairwise alignment between all sequence pairs.
2. Using the pairwise alignments Clustal generates a similarity matrix. This matrix contains a score indicating level of similarity between all the pairs of sequences.

The logic is that a low similarity score translates to higher degree of similarity between the two sequences.

3. Using the scores obtained Clustal develops a guide tree. It starts by putting two highly similar sequences into a group. Then Clustal progresses towards more dissimilar sequences and based on the similarity scores includes them into different groups.
4. It then aligns two most similar sequences and produces a pairwise alignment. Following the tree structure Clustal adds dissimilar sequences into the existing alignment and iteratively alters the initial alignment to achieve optimal alignment of all sequences.

Figure 22 - Process of MSA Inside Clustal

In order to classify unknown specimens into their correct species, we know that their COI gene sequence can be of great utility. Clustal provides an easy and systematic

way of aligning sets of gene sequences whose multiple sequence alignment can be compared to gene sequence of unknown specimens and check the resemblance between known and unknown sequences.

6.3 Profile Hidden Markov Model

The similarities between a gene of an unknown specimen and that of a known specimen maybe very subtle and hard to detect. Hence pairwise alignments may fail to detect what species a new found specimen may belong to. However an unknown sequence may have weak similarities with many sequences belonging to a species and this fact can help us assign an unknown specimen into the correct species. For this purpose we need to create a profile for a multiple sequence alignment that will help us use positional information of symbols into a Hidden Markov Model. The extra information from profiles helps an HMM to capture knowledge about the degree of conservation of a state in a multiple sequence alignment and the model generated is called a Profile Hidden Markov Model.

If the multiple sequence alignment does not contain any gaps then modelling a PHMM is simple. Every state is then called a match state. This is illustrated in fig 23. There are distinct observation probabilities for every match state.

Figure 23 - Match States in a PHMM [17]

Most multiple sequence alignments however do have gaps and hence we need to accommodate for insertions and deletions into the PHMM. The obvious way to model

insertions is to add “insertion” states. In the case of multiple insertions the insert states connect to themselves. Similarly we also add delete states into the PHMM to account for missing characters. The full structure of a PHMM with match and delete states is shown in figure 24.

Figure 24 - Profile Hidden Markov Model [17]

Consider the multiple sequence alignment in figure 25

A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C

Figure 25 - Example Multiple Sequence Alignment

From this matrix we can find the probability of occurrence of individual symbols at every position of the alignment.

Probability of A in first position = $4/5 = 0.8$

Probability of T in first position = $1/5 = 0.2$

Probability of C in second position = $4/5 = 0.8$

Probability of G in second position = $1/5 = 0.2$

After the 3rd column in the multiple sequence alignment 3 sequences have insertions, hence the probability of making an insertion is $3/5$ and not making insertion is $2/5$.

This profile of the multiple sequence alignment can be represented using the diagram as shown in fig 26. The insertion state is represented above the match states.

Figure 26 - PHMM Example

Based on the model described, we can now score any sequence against the model and find how closely it resembles the sequences of the multiple sequence alignment.

The score for the sequence ACACATC is calculated as below:

$$0.8 * 1 * 0.8 * 1 * 0.8 * 0.6 * 0.4 * 0.6 * 1 * 1 * 0.8 * 1 * 0.8 = 4.7 * 10^{-2}$$

If we perform similar calculation for a different sequence like TGCTAGG we get the result as $0.0023 * 10^{-2}$. The score for ACACATC is about 2000 times larger than score for sequence TGCTAGG. Therefore it is fair to say, the 1st sequence better fits

with the sequences of the multiple sequence alignment on which the Profile Hidden Markov Model was trained.

This straightforward training and scoring strategy is at the core of the concept used for performing genetic barcode identification. By scoring unknown gene sequences against dynamic profile Hidden Markov Models of several genera, this work tries to achieve correct placement into the taxonomy tree.