Pattern Recognition of DNA Sequences using Automata with application to Species Distinction

A Thesis
Presented to
The Faculty of the Department of Computer Science
San José State University

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By
Parnika P Achrekar
December 2013

SAN JOSE STATE UNIVERSITY

The Designated Thesis Committee Approves the Thesis Titled

Pattern Recognition of DNA Sequences using Automata with emphasis on Species Distinction

By
Parnika P Achrekar

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

December 2013

____

Dr. T. Y. Lin, Department of Computer Science Date

____

Dr. Chris Tseng, Department of Computer Science Date

____

Mr. Amit Sant, Software Engineer at Apple Inc Date

ABSTRACT

"Darwin wasn't just provocative in saying that we descend from the apes—he didn't go far enough, we are apes in every way, from our long arms and tailless bodies to our habits and temperament." said Frans de Waal, a primate scientist at Emory University in Atlanta, Georgia. 1.3 million Species have been named and analyzed by scientists. This project focuses on capturing various nucleotide sequences of various species and determining the similarity and differences between them. Finite state automata have been used to accomplish this. The automata for a DNA genome is created using Alergia algorithm and is used as the foundation for comparing it to the other species DNA sequences.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank each and every person who has contributed towards the completion of this project. Working on this project was an exciting experience. Knowledge and experience gained from this project will remain with me as an ingratiating memory.

I would like to express my special thanks of gratitude to my project advisor Dr. T. Y. Lin who gave me this golden opportunity to this wonderful project. His guidance and cooperation have helped me in completing this project successfully. Thanks for the benevolent support and kind attention. I would also like to thank my committee members Dr. Tseng and Mr. Amit Sant for their support and patience.

I would also like to thank our department for providing us with the necessary software required in our project. I'm also thankful to the library for providing necessary books and materials required to learn different concepts for our project.

Last but not the least, sincere thanks to my parents for inspiration and blessings, to my brother's constant moral support and encouragement without which project completion would have been next to impossible. I would also like to take this opportunity to thank my friends Mona, Mini, Krupali and Nikhil for being there for whenever I needed them.

Table of contents

List of Tables

List of Figures

1. Introduction

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Almost all the cells in a human body have the same DNA. Most DNA is found in the cell nucleus (where it is called nuclear DNA) however a small amount of DNA can also be discovered in the mitochondria (where it is called mitochondrial DNA or mtDNA). DNA molecules are double-stranded helices, consisting of two long biopolymers made of simpler units called nucleotides. DNA nucleobase contains 4 chemical bases: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) [15].

RNA or ribonucleic acid is an important molecule with long chains of nucleotides. A RNA nucleotide contains a nitrogenous base, a ribose sugar, and a phosphate [15]. RNA, just like DNA, is equally important for living beings. RNA is usually single stranded unlike DNA which is double stranded. RNA nucleobase is made up of 4 chemical bases: Adenine (A), Guanine (G), Cytosine (C) and Uracil (U) [2].

DNA chemical bases pair up with each other, A with T and C with G, forming units called base pairs. A sugar molecule and a phosphate molecule are attached to each base. DNA in humans contains around 3 billion bases and these are similar in two people for about 99% of the total bases. These bases are sequenced differently for different information that needs to be transmitted [15]. This is similar to the way that different sequences of letters form words and sequences of words form sentences.

The study of abstract machines and the computational difficulties that can be resolved using these abstract machines is called automata. Automata theory is closely related to formal language theory, as the automata are often classified by the class of formal languages they are able to recognize. A finite representation of a formal language that may be an infinite set can be automata [1].

Automata theory has been used to analyze the pattern of text data to find the writer and find the similarity and differences between him and others [5]. In biology, automata theory has been of vital importance. DNA nucleotide genomes have been symbolized using Cellular automata [13]. Hence, the study of DNA nucleobase pairs can be achieved using the automata theory.

A human DNA has approximately three billion base pairs. Searching a single gene from these vast base pairs that contribute to the human genome is known as DNA sequencing. In late 1970's, primary technique for DNA sequencing was established however scientist could sequence very few base pairs.

An enormous volume of information can be captured from one million bases or more. Matching the dissimilarity between the vast DNA sequences can help in understanding evolution, adaptation and immunity. The Human Genome Project (HGP) was dedicated to evolving innovative and improved tools to obtain gene economically, more rapidly and practical for scientists to achieve. Its popular sequencing of the human genome has provided scientists with a fundamental design of the human being [12].

In this project, we will create the automata of the DNA nucleotide sequence by appropriately representing the base pair sequences in the form of numerical symbols. We will further create a PTA (Prefix Tree Acceptor) to compare the sequence with various other species.

2. DNA Sequencing

A segment of DNA that is transferred from parents to children is known as gene. They are systematized and wrapped in components called chromosomes. Humans have 23 pairs of chromosomes which makes them different from other creatures. A gene also codes for a single protein molecule also known as polypeptide which is also used for protein synthesis. It comprises of two steps: Transcription and Translation [9].

Transcription: The sequence of one gene is replicated in an RNA molecule [15].

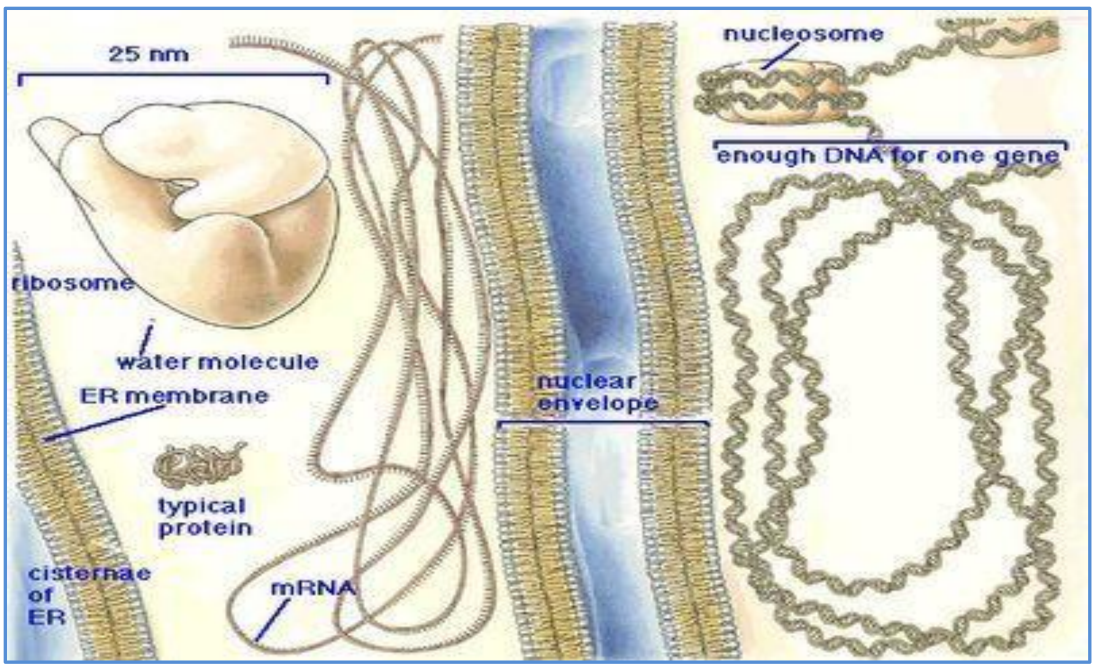

Figure 1: Process of Transcription [17]

Translation: The RNA molecule acts as a cypher for the formation of an amino-acid chain (a polypeptide) [15].

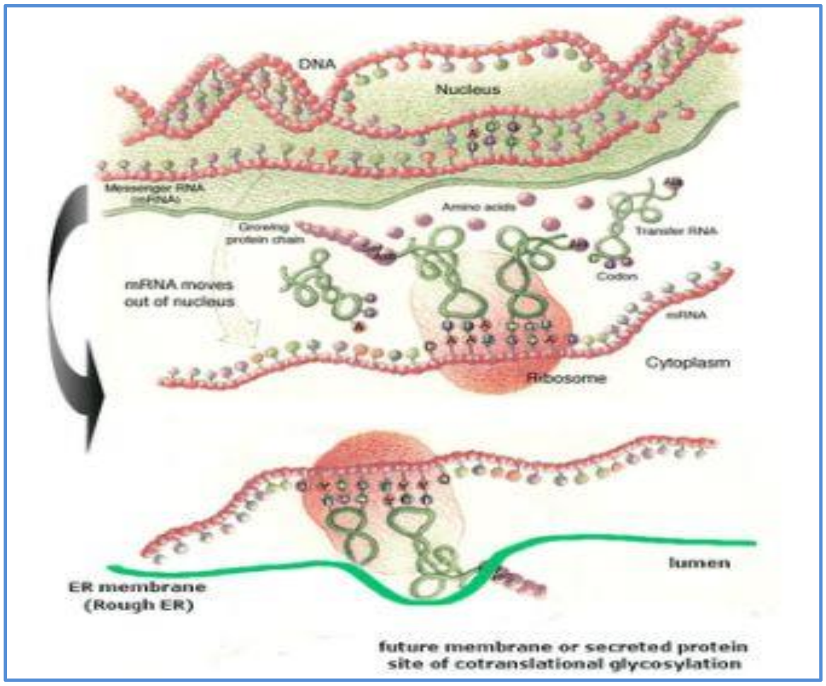

Figure 2: Process of Translation [17]

Translation of DNA to RNA into a sequence of amino acids marks the beginning of protein synthesis [9][15]. The main structure of protein is a thorough sequence of amino acids in a polypeptide string. A set of 20 naturally occurring amino acids exists today. Asparagine was discovered in 1806 followed by Cysteine, Leucine and Glucine [9].

Types of Amino Acids:

Amino Acid	one letter code	three letter code
L-alanine	A	Ala
L-arginine	R	Arg
L-asparagine	N	Asn
L-aspartic acid	D	Asp
L-cysteine	C	Cys
L-glutamine	Q	Gln
L-glutamic acid	E	Glu
glycine	G	Gly
L-histidine	H	His
L-isoleucine.	I	Ile
L-leucine	L	Leu
L-lysine	K	Lys
L-methionine	M	Met
L-phenylalanine	F	Phe
L-proline	P	Pro
L-serine	S	Ser
L-threonine	T	Thr
L-tryptophan	W	Trp
L-tyrosine	Y	Tyr
L-valine	V	Val

Table 1: List of Amino acids [2]

Amino acids are categorized into four major sets based on the properties of the "R" group in each amino acid. The types of amino acids are namely polar, nonpolar, positively charged, or negatively charged [9]. Polar amino acids have "R" groups that are hydrophilic, which hunt for contact with aqueous solutions. Nonpolar amino acids are the opposite of hydrophilic; they avoid contact with liquid [10].

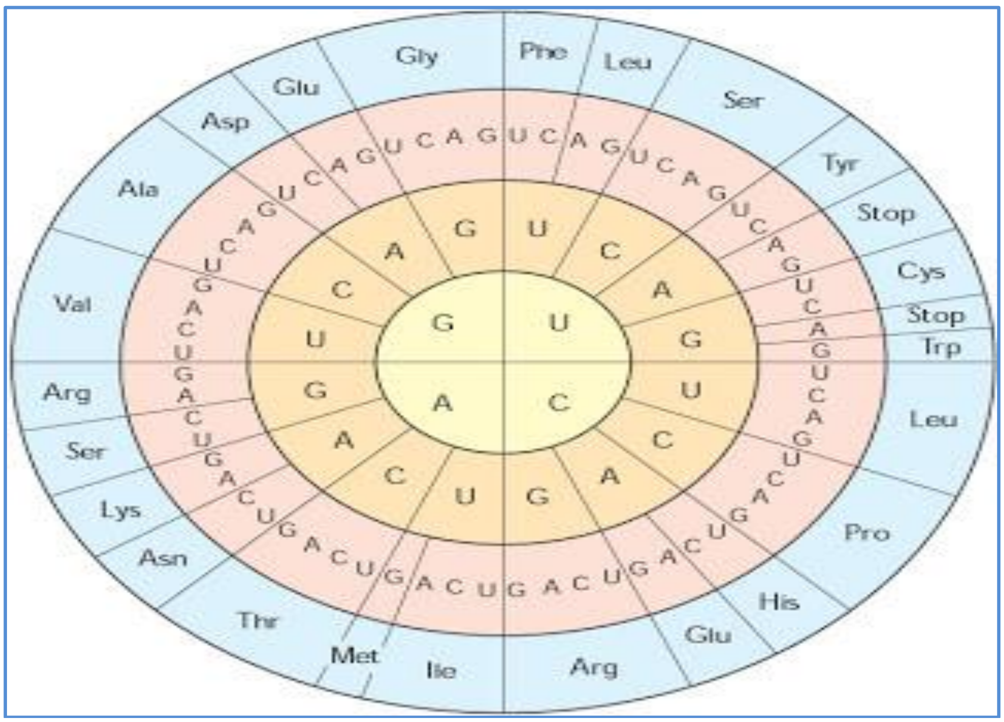

Figure 3: Amino Acids Chart [2]

There are 8 different types of essential amino acids: isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan and valine. The remaining 12 are non-essential amino acids [10]. Essential amino acids perform various functions in your body including supervising insulin and maintaining healthy hair, skin, and nails.

They act as the elementary building blocks of the human body. Deficiency in amino acids can lead to lower energy levels. It could also slower the rate of metabolism and cause skin and hair loss, indigestion, insomnia, stress etc. Obesity can be avoided by getting all the required amino acid, which in turn can help in throwing waste away from the bloodstream.

3. Understanding Automata

In this section, we will understand the use of Finite Automata for representing DNA genomes [1] [3].

3.1. Finite automaton 'A' is defined as follows:

$$A=(S, P, i, \delta, T), \text{ where}$$

- S: is a finite set known as set of states
- P: finite input alphabet

$$P = \{A, C, G, T\} \text{ or } \{A, C, G, U\}$$

- i: fixed element of A called as initial state
- δ: is a function:

$$\delta: S \times A \rightarrow S$$

It is known as the transition function.

- T: is a subset of S known as terminal state.

3.2. Non-Deterministic Finite Automata:

Non-deterministic finite automata can be in various states at a single instance of time [14]. Transition from one state on an input can be to any set of states.

DFA vs NFA [14]

Deterministic Finite Automata	**Non Deterministic Finite Automata**
Characterized as a 5 tuple state: <S, A, T, s_0, F>	Characterized as a 5 tuple state: <S, A, T, s_0, F>
S is the set of states	S is the set of states
A is the alphabet	A is the alphabet
T is the transition function: $S \times A \rightarrow S$	T is the transition function: $S \times (A \cup \{\varepsilon\}) \rightarrow PS$
s_0 is the initial state	s_0 is the initial state
F is the set of accepting states.	F is the set of accepting states.

4. Alergia Algorithm

Our main focus is on an algorithm that can encode the strategy for understanding the DNA sequences. This algorithm belongs to the family of functions that can be determined as Stochastic Finite State Transducer (SFST) [16][18]. Stochastic Moore machine is nothing but the probabilistic distribution of symbols.

We will use Alergia algorithm for our DNA recognition which is discussed as follows.

```
Algorithm Alergia
Input:
        S: sample set of strings
        α: 1 – confidence level
Output:
        SFA
Begin
        A = stochastic prefix tree acceptor from S
        Do (for j = successor(first node(A) to last
        node(A))
             Do (for i = firstnode(A) to j)
                  If compatible(i,j)
                       Merge (A,i,j)
                       Determinize(A)
                       Exit (i loop)
                  End if
             End for
        End for
        Return A
End algorithm
```

There are 4 major groups of amino acids: Polar, Non polar, positively charged and negatively charged. To build automata we have to convert these to numerical.

Hence, we will enumerate them in the following way:

NonPolar-0

Glycine (G) – GGU, GGC, GGA, GGG;

Alanine (A) – GCU, GCC, GCA, GCG;

Valine (V) – GUU, GUC, GUA, GUG;

Leucine (L) – CUU, CUC, CUA, CUG, UUA, UUG;

Isoleucine (I) – AUU, AUC, AUA;

Proline (P) – CCU, CCC, CCA, CCG;

Methionine (M) – AUG;

Phenylalanine (F) – UUU, UUC;

Tryptophan (W) – UGG

Polar-1

Serine (S) – UCU, UCC, UCA, UCG;

Threonine (T) – ACU, ACC, ACA, ACG;

Cysteine (C) – UGU, UGC;

Asparagine (N) – GAU, GAC;

Glutamine (Q) – CAA, CAG;

Tyrosine (Y) – UAU, UAC

Polar Acidic-2

Aspartic Acid (D) – GAU, GAC;

Glutamic Acid (E) – GAA, GAG

Polar Basic-3

Lysine (K) – AAA, AAG;

Arginine (R) – CGU, CHC, CGA, CGG, AGA, AGG;

Histidine (H) – CAU, CAC

Figure 3 shows that UAA, UAG and UGA are stop codons. We will group them in the final stage as 4.

Stop Codons-4

UAA,

UAG,

UGA

5. Creating SFA using Algorithm Alergia

Let us assume there are 'n' strings, $S=\{s_0, s_1, s_2, s_3, \dots s_n\}$ and $s_i = a_1a_2a_3\dots a_i$.

Once the SFA is build, we start merging the states [16]. Two states can be merged when they are compatible i.e. they have equal transition probabilities for every input $a \in A$ and the end nodes must be same as well.

$$q_i \equiv q_j \Rightarrow \forall a \in A, \text{ where } p_i(a) = p_j(a) \text{ and } \delta_i(a) \equiv \delta_j(a)$$

It's very difficult to find equal frequencies hence states are accepted to be same if they fall under a confidence range.

Given the probability p and frequency n for n values, a confidence range can be defined as:

$$\left|p - \frac{f}{n}\right| < \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}} \ \textit{with probability larger than } (1-a).$$

The probabilities are calculated and these values of vital importance for the process of merging. Algorithm Alergia will reject the states if these values are greater than the confidence range.

$$\left|\frac{f}{n} - \frac{f'}{n'}\right| > \sqrt{\frac{1}{2}\log\frac{2}{\alpha}}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n'}}\right).$$

The above equation helps in merging the compatible states. After merging all the compatible states, we get a SFA [16] which is an estimate of the initial one.

A DNA nucleotide sequence can be represented in the form of numerical depending on the 4 groups of amino acids discussed in Chapter 4 as follows:

Sequence 1: AUG AGA CCA GCG AGG ACA CCU GAU GAA UGA

Input 1: 0 3 0 0 3 1 0 2 2 4

Sequence 2: AUG CUC CAU CAA UGG GAC AAA UUU UUC UGG

Input 2: 0 0 3 1 0 2 3 0 0 0

Sequence 3: AUG AUC ACC UGU GAU AAG GUU AUU CCU CAU

Input 3: 0 1 1 1 2 3 0 0 0 3

Sequence 4: AUG UCU GAG GAC GAA CGU UCU UGG GAU AAA

Input 4: 0 1 2 2 2 3 1 1 2 3

Sequence 5: AUG CCU CAU GAU AAG AUC UGU CAU GUU ACC

Input 5: 0 0 3 1 3 1 1 3 0 1

Sequence 6: AUG AUU CCC UAU GAU GAG AAG GAC AAA UCU

Input 6: 0 0 0 1 2 2 3 2 3 1

Sequence 7: AUG CAU UAU GAU CAU GAC AAA CCU AUC GAU

Input 7: 0 3 1 1 3 2 3 0 1 2

Sequence 8: AUG CCU GAU AUU UGU CAU GUU GAG UAU ACC

Input 8: 0 0 1 0 1 3 0 2 1 1

Sequence 9: AUG GAU AAG GAA AAA UCA GAC CUU CCC CAU

Input 9: 0 1 3 2 3 1 1 0 0 3

Sequence 10: AUG AAA AAG GAU UGU CAA GAU AUC GAG CAC

Input 10: 0 3 3 2 1 1 2 0 2 3

Above are a few examples of DNA sequences being represented numerically. Once this is done we can now use Algorithm Alergia to build a prefix tree acceptor (PTA) [3][16]. The algorithm then merges all the compatible states in PTA and creates stochastic finite automata [16][17][18]. This automaton is an estimate of the initial one.

6. DNA samples of living organisms

There are approximately 8.7 million species of species on our planet out of which 6.5 million are from land and the remaining from the seas [8].

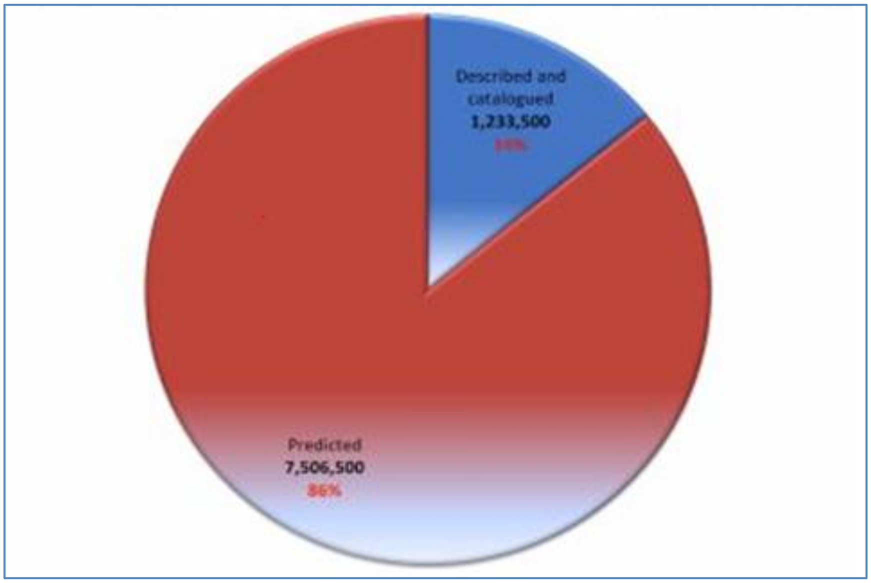

Figure 4: Total Number of Species on Earth [8]

As shown in the above figure, only 1.8 million species have been categorized and known to mankind. This clearly states that around 75-90% of them are yet to be discovered.

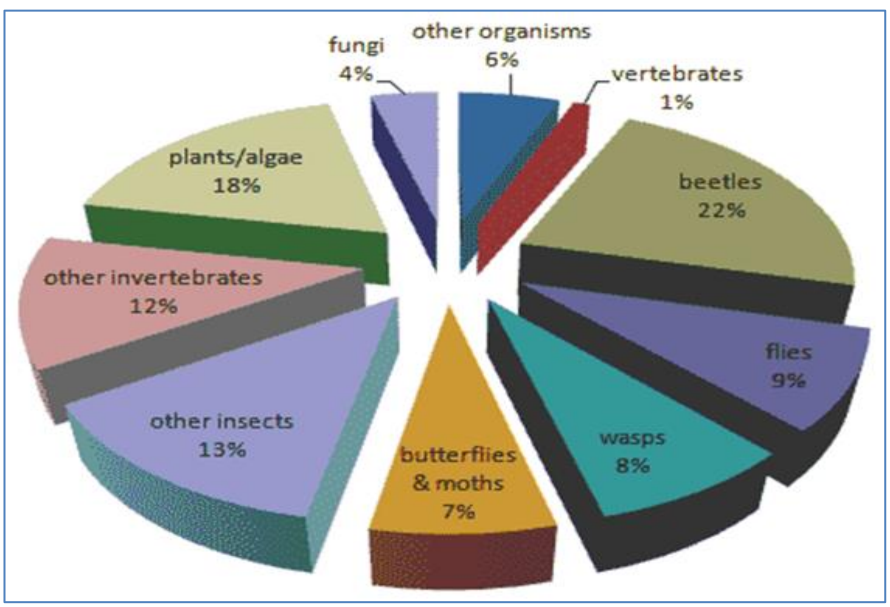

Figure 5: Relative Number of Named Species [7]

The above chart shows that there are approximately 12% of invertebrates such as arthropod, mollusk, annelid, coelenterate etc. Vertebrates, categorized by the existence of spinal cord, include mammals (human beings), birds, reptiles, amphibians etc. Our percentage is the lowest amongst all [7].

Our goal is to find the similarity between different species. Below are samples of DNA sequences [15][19] of some species:

DNA nucleotide sequence for Homo sapiens (Human) [19]:

```
   1 atttccagct ttctatgcat tctggcaaaa gctagctcca caagccagag gacagccctt
  61 gagagaaaga tttaggcact ggcttttgaa atagaaagca cctcaaatgc tggggagaag
 121 gaacacacag aaaatagcaa aaaaggatcc agtgagacct gggcaatgca caaatgcaat
 181 gcaccacttt gagacaatca gctttcaatt tacacaagca gtaacaatgc tccaaaccac
 241 accctgcagc tgtcccatgc accatcaggg aaatctctga tgctgctggt gccctgccag
 301 caccactacc cactgctgca tctaactgct gactgcagtc attgccccat cctcactccc
 361 atggattctg cctgtaacct gctcttggaa tctctgactt ctaaagtcta gcgtttatgg
 421 aatactacac agccacacaa aataatgaaa tcatatcttt tgtaccaaca tggatgcagg
 481 tgaaggccat tatccttagt gaaattaaca gaaaaccaaa taccgtatgt tctcacttat
 541 aagtgacagc taaacactgg ttactcatgg acataaaaat aggaacaata gacactgggg
 601 aatactggag gggggaagga gggaaaggaa caacagttga aaaactaact gttggttact
 661 atgctcagga catgggtgac agtatcattc ataccccgaa cttcaatatc atgtaatgta
 721 ctcatgtaac aaacctgcac atgtaccccc tgaatctaaa ataagttgaa attacaaaaa
 781 acaaaaaata aaataaaaca aagtttaggg tgctaagtga tggcagccag ggtgtgttta
 841 tacatcagct gcaagaaatg ccagaaaagg gaatatctgg catttttagc tgtcgtatca
 901 agaggcaaga tccacctcat taaatattag gtgggaattc ccaaaacacg gggagaagat
 961 gatgatgttg tgtagaaaaa aaaaaaaaaa gtaagagcca ttcactccac acacaaatgc
1021 ataaaacatt tagaattggg ccgggcgcag tggctcacgc ctgtaatccc agcacttggg
1081 gaggccgaga cgggcagatc atgaggtcag gagatcgagg tcatcctggc taacacagtg
1141 aaatcccgtc tctactaaaa atacaaaaaa atagccaggt gtggtggcgg gcgcctgtag
1201 tcccagctac ttaggaggct gaggcaggag aatggcatga acccaggagg cggagcttgc
1261 agtgagcaga gatcatgcca ctgcactcca gcctaggcga cagtgagact ccacctcaaa
1321 aaaaaaatcc atttagaatt aatatgaaat tgccatcaga aattacctct ggggagtgga
1381 accagagcta tagtttcagg agtgggtgag agaagattct tacttctcat tttatatgtt
1441 tcggtagtat ttaagaattt tataagcgac atatgtttct tttttgatt tcaaagaact
1501 ggtttacttt ttaagacctg tctctttctt tagaactgct tttaaaaaga ggctggaacg
1561 ttttaattaa attatgtacc ctctgctttc aggaagggag gccactcaga tttggtggcg
1621 gtggttacca ttcatttttt cattcattta tcaaagattt attgattgta tgcaaggccc
1681 aagaaagatg aaagacagag gctctgttct caaggaggga attaatgtta tgatgagaaa
1741 tgtctttgaa tgtcttgggt tttgtgttat tttcttacat attggtgaac cttttacttc
1801 agatagtaag taccctctac tatacagctt taactagatt tacttacgtt ttttcctatt
1861 aaatggaatt aggaaatata agttgtacat cttcacaatg atttccaagc taaatgatgt
1921 tggtggggtc tttgaaatga gttactgtgg aagtatttta tgctcttgaa cttctgtgga
1981 agtattttat gctcttgaat ttcattcaag aattcaattt aacttcattt aaagatttca
2041 ttagaattag gtgacatcac cttatgtttt gtgttggttt gcaaaagact tattgctagc
2101 cagatgtgct ccttttgctg atagtaatat aagcattcta aaagttctaa tttctaagcc
2161 ttggatttaa tacaaaacca taggtaataa agatgtataa aaatctagca cggagtccgg
2221 acgcggtggc tcatgcctgt aatcccagca ctttgggaag tcgaggtggg tggatcacct
```

Figure 6: DNA sequence of Human [19]

DNA nucleotide sequence for Chimpanzee [19]:

```
   1 ccacgcgtcc gggtggtgcc aaattctggg gcctaggcat ttccctcgct ttatgttttt
  61 ggtttttttt cttccttcaa tctctttgat taggccgtac gtggctgtgg caaggagttg
 121 gggaaaaaaa ttataaaaac aggaaagaga gaaagcacag ccagagcccc ggcttcgcga
 181 gccgccgggg agggggcgga ggaggctgag ccaggcagag tcgccagcgg agactcgcga
 241 gtggcgcgcg ggaggagcgg ctgccggcgc tgggcttgcc ttgctgctgc tgctgctgcc
 301 tccccaccgc cttttttttt ttttaatctg gagcggggtg gggagtggga accggagaga
 361 aagcaaaata ttaaaaagcc ccaaagacag ccagcaggag cgcggtgccc gatggcttcg
 421 ctgtaccaga ggttcactgg caagatcaac acctcgaggt ccttccccgc gcccccggag
 481 gcgagtcacc tcctgggcgg ccaggggccc gaggaggacg gcggcgcagg agccaagccc
 541 ctcggcccgc gggcgcaggc ggcggcgccc cgggagcgcg gcggcggcgg cggcggcgcg
 601 ggtggccggc cccggttcca gtaccaggcg cggagcgatg gtgacgagga ggacgagctg
 661 gtggggagta accctccgca gaggaattgg aaaggaatag caattgcact gcttgtcatt
 721 ctggtcatct gctccttgat cgtcacctcg gtcatacttc tgacaccagc ggaagataat
 781 agtctgtctc aaaagaagaa ggtcactgta gaagatctct tcagtgaaga cttcaaaatt
 841 catgaccccg aggctaagtg gataagtgat acagaattca tctacagaga acagaaagga
 901 acagtgagac tgtggaatgt tgaaacaaat acttctactg tcttaataga aggcaaaaaa
 961 attgaatcat taagagccat cagatatgaa atatctccag atagagagta tgcacttttt
1021 tcatacaatg tggaacccat gaagaaagtg aagtccagga agttgacatt gcctcattca
1081 aaatcatgtg actcattagc agtaagtcaa gcctgtagcc cagcttgtca ccagggctgt
1141 tttcttcatt acatcaccat gtctcttcct cttcactgcc tgcgtgacta tgtctcggca
1201 gtcaatggat acagcacagc attgccagct tgccatgtac aaggggacc tgtttcagat
1261 attccatgga gaccctggct ggaggattgc aggagagtcc caggaggcag gactgccaat
1321 ggcaccaggc ttcgcagcca tgcacctgca gccctcaggc agcactgtcc attgtcatac
1381 gagtgtggca ggtgtgaggc atcgcatctg ctcaccccgg ggataatgca cagcagctac
1441 aggcagattt cgggccagag agcaaccgag tgagccttgc agcctctgct gccagcacag
1501 gcttgttcct tcaacactgg tggagagaga cacgctgtca tcaggcccaa gaaatactgc
1561 cttccccatc ctatccctgg tcactgggtg cccgcagagt gtcccagagg agggagggag
1621 ggaccctcca ctggttcaaa tggcctgttc tcagagatgc agcaatagac cctcgtgaat
1681 actgaactga taatcatggg aaggagactg gctctcctgg attccctcat gattcctctg
1741 agtgacaatg tgatgttggc cgactgtgtc ttcttcagaa tatcatatac acttgaggtc
1801 tccaggagcc tccaattaca ttattttcct ggctcataca gtgacaagta attcttatcc
1861 tggattcctc gttactgaga cttttcttgc ctttttttgtt agcttatgat ttattctagg
1921 acttcctcca acaggttata cttaactgtc tacctcagtc tctggaagtt ttaaaaatgt
1981 tcagctaaat aaaagaagta gattctccct ggaaaccaaa aaaaaaaaaa aaaaaaaaaa
2041 aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaa
```

Figure 7: DNA sequence of Chimpanzee [19]

DNA nucleotide sequence for Monkey [19]:

```
  1 aagcttctcc ggcgcaacta ttctcataat cgcccacgga ctcacctcct ccatgctatt
 61 ctgcttagcc aattccaact atgaacgcac ccacagtcgt gttataatgc tctcccgagg
121 acttcaagcc ttacttccac taatggcctt ttgatgattc gcagcaaatc ttaccaatct
181 agccctaccc cccactatca atctaatagc agagctcctt gttattacag cttcattttc
241 ttgatctcat atcactatca tactaatagg gctcaacata ctaatcacag ccctctattc
301 cctctacata tttatcacaa cacaacgagg aaaacttaca caccacacaa ctaacataaa
361 gccctcattc acacgagaaa acacactaat attccttcac cttgccccaa ttattcttct
421 atcccttaac cctagcatca tcctaggatt tacctcttgt aagtatagtt taattaaaac
481 accagattgt gaatctgact atagaggcca gtcacttctt atttaccgag aaaactcgca
541 aggattgcta acccatgttc ccataattaa aactatggtt ttctcaactt ttaaaggata
601 atagctatcc attggcctta ggagccaaaa atattggtgc aactccaaat aaaagtaaca
661 attatgtaca cctccattat aataacagcc ctcgtctccc taattcttcc aatcattgcc
721 acccttatta accctaataa aaaacagtca tatccaaact atgtaaaaac aactacaata
781 tatgccttca tcaccagcct tatcccaata actctctacc tctttctaaa tcaagaggca
841 actatttgaa gttggcattg aacaacaacc caaacactaa atctaacatt aagctt
```

Figure 8: DNA sequence of Monkey [19]

DNA nucleotide sequence for Mus Musculus (House Mouse)[19]:

```
   1 ctgggattac aagctggtac aactactctg gaaatcagtt tggcaggtcc tcagaaaatt
  61 taatgtaatg ctacccgagg acccactaat accactcctg gcatctaccc agaagatgct
 121 ccaatatgta ataatgacac atgctctact atgttcatag caccctattt tataataacc
 181 agaagctgga aagaacccag atgtccctca gcagaggaat ggatacagaa aatgtggtat
 241 atttatacaa tggaatacta ctcagctatt aaaaggagtg aattcatgaa attcttaggc
 301 aaatggatgg aattagaaaa tatccttagt gaggttaccc aatcacaaaa gaacacacat
 361 ggtatacact cactgataag tggatactag cccagaagtt cagaataccc aagatacaat
 421 ttaaagacaa aatgaagctc aagaagaagg aagaccaaag tatggatact ttggtccttc
 481 ttagaagggg aaacaaaata cccatgggag gagttacaga gacagcgtgt ggagcagaga
 541 ctgaaggaaa ggccattgag agacttcccc acctagagat ccatcccata tacagtcacc
 601 aaacccagac agtattgtgg atgccaacca gtgctttctg acaggagcct gatatagctg
 661 tctccttgag aggctcttgc cagtgcctga caaatacaga gatggatgct ctctctgagc
 721 acaaggcctc cagtggagaa gctagagaaa ggaccaaagg agctgaagga gcttgcagcc
 781 ccataggagg aacaacaata tgaaccaacc agtacctcca gagctcccag ggagtaaacc
 841 accaaccaga gagtatgcat ggtgggactc atgactccag ctgcacatgt agcagaggat
 901 ggtcttattg gacatcaatg ggaggagagg cgcttggtcc tgagaagact taatgtccca
 961 gtataggaaa atgccaggac agggaagcgt gggtgggtgg gtggtgagca gggggttggg
1021 ggagagaata gggggttttc agaggggaaa ccaggaaagg ggattacatt tgaaatgtaa
1081 ataaagagaa aaatctaata aaaaaacatt atgttacaat aaaaaaaatg agagaaatta
1141 gtaaagagcc aatgttttaa gtggaggtat tgaataccaa taatactatg ttcagatttc
1201 tgaagaactg ccagaatgat ttctagaggg gttataccag cttgcaatcc catgaaggag
1261 tttttctctt tcaccacatc cttgccagca cctgctgtca cctgagtttt tgatcttagc
1321 cattctgatt ggtgagaggt ggaatctcag ggccgttttg atttgccttt ctctgatgac
1381 tgaggatatt ttctattcca tgatcttatt caggattcta cattgtagtt agtccttata
1441 tctccttata tattctttta ttaattgatc atttttaatt tacatttcaa atgttattcc
1501 cccttcccca tccccctctg caaatccccc ctatctcact cctgcacttc tatgagggtg
1561 ctcccctacc cactcattca ctcctgcctc actgccctag cattccctta cgctggggca
1621 tcgagccttc acagggccaa gggcctcctc tcccattaat gccagataag gccatcctct
1681 gctacatatg gagctagacc catggatcct tccatgtgta ttctttggtt ggtgtttttt
1741 tttttttttt tttttagtcc ctgggagcac tgggcaatct ggctggttga tactgttttt
1801 cttcctatgg ggttgaaaac ttcttcaact ccttagtcct tcccctaact cttccattgg
```

DNA nucleotide sequence for Banana [19]:

```
  1 tggatttaaa gctggtgtta aagattacaa attgacttat tatactcctg actacgaagt
 61 caaagatact gatatcttgg cagcattccg agtaactcct caacctggag ttccgcccga
121 agaagcaggg gctgcggtag ctgccgaatc ttctactggt acatggacaa ctgtgtggac
181 tgatggactt accagtcttg atcgttacaa agggcgatgc taccacatcg aggccgttgt
241 tggggaggaa aatcaatata ttgcttatgt agcttatcct ttagaccttt ttgaagaagg
301 ttctgttact aacatgttta cttccattgt gggtaatgta tttggtttca aagccttacg
361 agctctacgt ctggaggatc tgcgaattcc cacttcttat tccaaaactt tccaaggccc
421 gcctcacggc attcaggttg aaagagataa gttgaacaag tatggtcgtc ccctattggg
481 atgtactatt aaaccaaaat tgggattatc tgcaaaaaac tacggtagag cggtttatga
541 atgtctacgt ggtggacttg attttaccaa agatgatgaa aacgtgaact cacagccatt
601 tatgcgttgg agag
```

Figure 10: DNA sequence of Banana [19]

DNA nucleotide sequence for Weed [19]:

```
   1 atgcattgca tggctgttcg ccatttcgct ccatcgtcat cgctctccat attttcgagt
  61 actaatatta ataatcattt ttttggtaga gaaattttta caccaaaaac atctaatatt
 121 acaacaaaaa aatcaagatc aagacctaat tgcaatccaa tccaatgtag tttggccaaa
 181 agccctagta gtgatactag tacaattgtt agaagatcag ccaactatga tcctcccatt
 241 tggtcttttg atttcattca gtctcttcca tgcaaatata agggagaacc ctatacaagt
 301 cgatcgaata agctaaaaga agaagtgaaa aagatgttag ttggaatgga aaactcttta
 361 gtccaacttg agttgattga tacattacaa agacttggaa tatcttatca ttttgagaat
 421 gaaatcattt ctattttgaa agaatatttc actaatatta gtactaataa aaaccctaaa
 481 tatgatttat atgccactgc tctcgaattt aggcttttac gcgaatatgg atatgcaata
 541 cctcaagaaa tatttaatga ttttaaggac gagacgggaa agttcaaagc gagtattaaa
 601 aatgatgata ttaagggagt attggcttta tatgaagctt cattctatgt gaaaaatggt
 661 gaaaatattt tggaggaagc tagggttttc acaacagaat atctcaaaag atatgtaatg
 721 atgattgatc aaaacataat attaaatgat aatatggcaa tattagtgag acatgccttg
 781 gagatgccac ttcattggag gactataaga gcagaagcta agtggttcat tgaagaatat
 841 gagaagacac aagacaagaa tggcactttg cttgaatttg cgaaattgga tttcaacatg
 901 cttcaatcaa tatttcaaga agatctaaaa catgtctcga ggtggtggga acattctgag
 961 cttggaaaga ataaaatggt ttatgctaga gatagattgg tagaggcttt tctatggcag
1021 gttggagtaa gatttgagcc acaattcagc cactttagga gaatatctgc aagaatatat
1081 gctctaatta caatcataga tgacatatat gatgtgtatg gaacattgga agagttagag
1141 cttttcacca aggctgttga gagatgggat gcgaagacca tacacgagtt accagattat
1201 atgaagttgc ctttctttac tttatttaac accgtaaatg aaatggcgta tgatgtatta
1261 gaagagcata attttgtcac cgttgaatac ctcaagaact cgtgggcaga gttatgtagg
1321 tgctatttgg aagaggcaaa atggttctat agcggataca aaccaacctt gaaaaaatat
1381 attgagaacg cctcgctttc aataggagga caaattattt ttgtatatgc tttttctct
1441 cttacaaagt ccataacaaa cgaggcctta gagtccttgc aagagggtca tcacgctgca
1501 tgtcgccaag gatccttaat gttacgactt gcagatgatc taggaacatt gtcggatgaa
```

Figure 11: DNA sequence of Weed [19]

DNA nucleotide sequence for Drosophila Melanogaster (Fruit Fly) [19]:

```
   1 gaattcttga atatatccaa gtctagttac gcaccttctt caccaggcga catttgacaa
  61 cattgtcgtt gagcggatgt gtcgtcatat cgaagagtag aaaattttgc ttttccgtcg
 121 tgagcacacc cttctccacc agatttttgg ccagacgttc gcgtacattt ttcagttggt
 181 agcgcaattt caacggattc caggtttcac ctgccacaac aataggttat acaaaacata
 241 cttggcgaaa tggcaggcgc taaatacaca ccactaagat attcaatcca gctctgcacc
 301 gtctccgggg gatctgtttc cttaatgtgt ttaagtccct ccaccactaa gatattcaat
 361 ccagctctgc accgtctccg gggatctgt ttccttaatg tgtttaagtg cctcatcgag
 421 tagaacgtct cccgtctgct gatccgattt cagtattaat ttccttgtac atagaccacg
 481 tcgccgcatt ccagatttct cgatcatcac gcgacctcgc agtccaagct ctatgagaat
 541 gcatccgcgc aagccgcttg atatgcagtc gttccagaaa gatgtgtagc cctccttgtc
 601 cttgagtccc agcagcagaa cctcctccat gagcgttagt cgtgtttcct tggagtcgcc
 661 atcgtcgata ttgtcctcct ggtctcatac acgcacacaa acacagcgag agcgagatgt
 721 ccgagaaaaa cctgaaagtg ggcgcccggg tcgagctgac cggcaaggat ctgcttggca
 781 cggttgccta cgtggggatg accagcttcg cgtcggcaag tgggtgggcg tcgtgctgga
 841 cgagccgaag ggcaaaaaca gcggctccat caagggccag cagtacttcc agtgcgatga
 901 gaactgtggc atgtttgtgc gacccacgca gctgcgtctg ctggaggctg ctcctggcag
 961 caggcgcagc atcgaggatg tcagcggggc tacgcccacg gctgcccaac ccacaaaggc
1021 gcggctgagc agctctcgca cctcgctctc ctccagtcgc caatcgctgc tgggttcccg
1081 cacccagttg accacttctc tgagtgaacg cactgcctcc agcagcagta ttggcccgag
1141 gaaatctttg gcgccgcaaa acagcaagga taaggagtcc cccagcactt cattggcaga
1201 aggagcccca gcagcaagcg gtggcaacgg tgccgttcgc atgcctcctc caaacgggct
1261 tccttcgtgg agacgggctt ccttgaaatt cttaagccgc agttcacgcc ttcccagcca
1321 ctgcgatcgc cctctttcac catgccctcc aactccggtg ctgaagacaa ggttcgccct
1381 gctggaggca cagaaaacga gcgccgagct gcaggctcag ctggctgatc tcaccgagaa
1441 gctggaaact ttaaagcagc gcaggaacga ggataaagaa aggttgcggg agttcgacaa
1501 gatgaagatt cagtttgagc agcttcaaga gtttcgaacg aaaatcatgg gtgctcaggc
1561 ttcgcttcag aaggagttac tgcgcgccaa acaggaggcc aaggatgcaa tcgaggccaa
1621 ggagcagcat gctcaggaaa tggcagatct ggcagacaat gtggagatga tcacgctgga
1681 caaggaaatg gccgaggaga aggccgacac gctgcagctg gagctagagt cctccaagga
1741 gcgtattgaa gagttggagg tagatctgga gctcttacgc tcggagatgc aaaacaaggc
1801 cgaatctgcc atcggaaata tttctggcgg cggcgattcg ccgggcctct ctacttatga
1861 attcaaacag ctggagcaac agaacattcg tttgaaggaa acactagtgc gtctgaggga
1921 tctatctgct cacgacaagc acgacatcca aaagttgagc aaggaactgg agatgaagcg
1981 ctctgaagtc accgaactgg agcgcaccaa ggagaagctt agtgccaaga ttgatgaact
2041 ggaggccata gtcgccgact tgcaggaaca agtcgatgct gcacttggtg ccgaggaaat
2101 ggtggagcag ctggctgaaa agaaaatgga attggaagac aaagtaaaac tgctcgagga
2161 ggaaattgcc caattggagg ccttggagga agtgcacgaa cagctggtgg agagtaacca
2221 cgaactggag cttgatctgc gcgaggaatt ggatctcgcc aatggggcca aaaaggaggt
2281 gctgcgagag cgggatgctg ccattgaaac catctatgat cgcgaccaaa ctatcgttaa
2341 gtttagggaa ctggtacaga agctaaacga ccaactaact gagttaaggg atcgcaattc
2401 tagcaacgaa aaggagtcgt tgcaggatcc cagtttgaaa atggtcaccg aaaccatcga
2461 ctacaaacaa atgttcgccg aatccaaggc ttacactcgc gccatcgacg ttcaactgcg
2521 ccagattgag ctgagccagg ccaatgagca tgtccagatg cttaccgcct tcatgcctga
2581 gtcattcatg agtcgcggtg gcgatcacga ctcaatcctt gtgattctgc tcatttcacg
2641 cattgtcttt aagtgcgcac attgtcgttt cgcaaacgag agagcgtttc ccaccagtgg
```

Figure 12: DNA sequence of Drosophila Melanogaster [19]

DNA nucleotide sequence for Oryza sativa (Rice) [19]:

```
   1 tcccaaaaca atgtgtctat ggtcttccga attcctagtc tcagcattgt gcaccaccga
  61 gctaggttgc agactatcac gatctgcttg atatatagtg tcaatttggt gtgtaccaac
 121 taaaggttgg tttgcattta ccgtctttct ttgtttatta gcaattgttt ctcgctgagt
 181 ggccatactt cttcctctct ttttagtgag tggaagttga gtggttttat ttggtacctc
 241 cactctttct ggcgcattct gagcgggaat gaaagattta gtcacacctt tataattggt
 301 aaatgcatct ggcagattat ttgcaagtct ttgcaaatgt ataattttct gaacttgaag
 361 ttcagtttca gtagtacgtg ggtctgaggc tggaacacct tgggcatccc aatcaatttc
 421 ctggcattct ttctggtact tgaagtctcc ccctaatgcc gggaaatgtt actcatcaaa
 481 gatagagtca gcgaaccagg cagtaaatag atcacatgtt aagggttcta aatactttat
 541 gatcgacgga gatttgaatc ccacatagat ccccactttc ctgtgtgggc ccatagcagt
 601 acgctgtggt ggtgagatcg gtatgtatac aacacaaccg aacttacgca aatgggaaat
 661 atttggaaga tttccacgta ctaactgcat tggggaagtt tcatgatatg cagttggtcg
 721 tagttggaca aggtcagcag cgtgcagaac tgcatgaccc caacacgacg aaggtaattt
 781 gcaattcatc aataatggtc gagtaataag cttaattctt tttatcaatg attcagccaa
 841 accattttgt gtgtggacat atggaacaaa gtgttgaacc tgaattccca atgccataca
 901 ataatcatcg aaagcatggg atgtaaattc ggcagcattg tccatacgga ttgattgaat
 961 cctatgttca gggtaatttg ccttcagcct tataatttga gacattaatt tggcaaaggc
1021 atggtttcgt gtcgatagaa gacacacatg agaccatcta gtagatgcat caatcagaac
1081 cataaagtac ctaaacggtc cagatcttgg cacaataggg ccatagatat ctccttgaat
1141 gcgttcaagg aatttaagtg gttcggctct aattttgaga taagatggtc tcaaaatcag
1201 tttcccagta gcacatgcag tgcatacgaa atcggaggat ttgggaaatt tgtcagtgat
1261 caaatgatga ccaatagagt tgccaataat ttttctcatc atcccgatac tagggtgccc
1321 aagtcgatca tgccaagtgt ggaatgcatc aacattttga aaaattactt tgtacgtaac
1381 atgtgcaatg ggcttaatgt atgtatagta caatcccgat gtgagagatg gaattttctc
1441 gcaaatgcat ttgccatatc tgttttgttt ggttaagaga agaaattctt ctcgattatc
1501 catatgggtt tcaatgtgaa acccattttg acggatatct ctataactta gtagggtacg
1561 ggttgaatca agatacaata aagcatcctt gattgtaatt tgtgtaccca ttgggagtgt
1621 aataattgct catcctgagc caactatcac agtatcgcgc ccagtgatag tcaaaacttt
1681 gccttctctc tttttgagag tttgaaagta tttgatctcc ctaagtatag agtttgtggt
1741 accactgtcc acaagacata attcctctcc aatcggagtg atatccttag acatctataa
1801 tgaaagaaga attgcttgat taagaattct ttatccaata tatatacata cataaaataa
1861 ttaaaacatc agatacatag tatgacgttt acaaatgtta atagtacata ctctaatgac
1921 tagcaagtct tataacctta taatataagg gagtttgtac tcatcgactt attacaacca
1981 ttattgtttt aacaaactat aggatatcaa tatactgtct caaacacact gagattaaag
2041 cagctttatc tctaagtggg acgcactgag attacagtaa atctccaagt gggtccgttg
2101 agcagtattc gatgagcatg tcatccattg cagaaaatgc agcggtatcc tctgggagaa
2161 gagcgaggtt gttctctggt tcaataggag cctagtgaga actttcaaca tccggtcttt
2221 cttttgtaag atgaagtgag cttcaaatct tagttcctca gaagactttt tcgcctttag
2281 ggatttctga tacaggagaa caagatgttt ttgggatgtg gcaatcttta gtgacatgat
2341 agtcagatcc acacctgttg caatgcctgt tgctattgca acgaggttgt ggtgccttac
2401 ccttctcttt tcctttctat cgaccattgg atttgcacct tgttgttatg ttgcgttttc
2461 cagtcagatt cttagggtta ttcgaggaat ttcccttgaa tcctttaagt gcgatactgt
2521 tggtttagta tcttgtcctc cggaaacata gttgatagag ttttctctat ctttctgctt
2581 tcgttggctc ttatcgcaaa atatcaactt ggagcaaatg ttgtgaacag catgattgta
2641 ttctgccaca gtttaaaatc ctgtaggcgt aaatgaatcc agccataatt agcctcatgc
```

Figure 13: DNA sequence of Oryza sativa (Rice) [19]

DNA nucleotide sequence for Agaricus bisporus(Mushrooms) [19]:

```
   1 accgacgatg catttctctt tgtcttttgc cacccttgct ctcttagtcg cttcggctgt
  61 tggtgcgccc gctgcgatcc actctatcga gactttcgat ggcgagacta ctggaaagca
 121 catcatcatg ctcaaggaag gagtcaagaa ggaggatctc ttcgccaact tcaaggccaa
 181 ggtcgctgta tcccatcagt gggaactgat caatggcttt gccggtgaat tcgacgagga
 241 gacactgaac gagcttcgcg caaaccccaa cgttgagagc atttccgagg acggcctgat
 301 gcacaccatg actactcaaa ccaatgcgcc atggggcctc gcccgattga gctccactac
 361 aaggctcagt aaccagaacg ccgcagctct gaccttcagc tacaccttcg atgcttccgc
 421 cggaagtggc gttgatattt tcattgttga taccggcatt ctcacaacgc acagtcaatt
 481 cggtggtcgt gcagcttggg gagagacctt cggtccctac gcagaccgtg atggcaacgg
 541 tcatggtact catgtcgccg gtactgctgc tggaagccaa ttcggtgttg ctaaatctgc
 601 caacgtcttc gccgttaagg tactcagcga tgaaggttcc ggttcgatca ccgatatcgt
 661 ttccggcttg aacttcgtcg gccaaagagc tgcgtccagt ggccgaccca cgattgcatc
 721 catgtctcta ggtggtggtg cctccagcag tctggacagt gcagtagctt ctctcacgaa
 781 cagtggtgtt cacgttaccg tcgctgccgg aaatgataat gccaacgccg cgaatacatc
 841 tcccgctcgt gctccttccg ccattactgt cggcgcatct actaccggcg acgctcgtgc
 901 ttcattctcc aactttggaa gcgttgtcga catcttcgct cccggccaga gcgtcatcag
 961 ttcttggatc ggtagcaaca ctgataccaa ctgcatctca ggaacttcca tggcaactcc
1021 ccatattgca ggactcgtcg cttacttgat cagtcttcaa ggaaacgtga gccccgctgc
1081 catgagcacc aagatcaagt ccctcagttt gaagggtgtc atcagtggaa ttccttaagg
1141 aagcccttga gagttgctga accgggtgtt acgaatttcg aagccgcata ttgaaatttg
1201 gaatgtatca tcatcattat tcctttgttt tttaaaaatc aagtcaagga atatacactt
1261 tgcaaaaaaa aaaaaaaaaa
```

Figure 14: DNA sequence of Agaricus bisporus(Mushroom) [19]

DNA nucleotide sequence for Felis Catus (Cat) [19]:

```
   1 ggcgggggga ggagggtcta agagagcaga aggaaggttt ccatgggaca ggccctcgcc
  61 tcaacccggg gatcctggtg cgcctcctcc aaggcggcca cgaggggggcg ccgcggccgc
 121 gcctgcgaac tcacctgtgc agaagcaggc acgcggctgt tctcagccgg cgggatccag
 181 cgggcaggtg tgggttcgag cgcgcagagc ttcctgattt tcggtccccc agcgcgggtg
 241 tccaggcccg ggggtggggt gactggcttg gggctgagc ccctcaggtg gagccatcgc
 301 actgtgtctc cttgaaacca ggctctgagc agagagagaa acagagatgt gtgggcgctt
 361 ctccggctgg gggacgtcct cctgcgtgtc actctcaggc gggcgcagcc ggcccggtgt
 421 tgaccgccgc gtgggcgccc cgacgggcgg agggagaggg aagacgagcg gtaagcaaat
 481 cagtgtggag gggagaagac ggaggagacc tccggcaagg agaggaagga agcggagggg
 541 ggaggcggga agaggaggag aagcatcaga cctgaaatcc gaggtgggag gggagctggg
 601 ggcagggaga ccgggtgtgt ggggcgggtg gcggggcggg ggtgagtgag aggaggccgt
 661 ttgcggcctg aaccggggag gccttatgaa atgaggcagc ggtgggcgcg gttctcggcg
 721 gtagaattcc acgggctgtg gaaattccag ggctgttgct tggattgcct gaagaagacg
 781 tgtgtgtcgg gttagggtgg ttgagacagg agtgggtgca gagggttctg ggtgcggggg
 841 aggcaagtga ccgtgtgtgt acagtgtgag gctgcattgg ggcggcgtga aagcaagtca
 901 cgctaatctg gcgagagaga tcatggtcgg gaacgtactt ttttccagag tgaggcatgt
 961 gtgttccgcc gaggacctac tgaccctctg tgattttcct caagtatgcg cagttcggct
1021 gcgcttgtgc tctctcgagg taactggtgt ttaaagcatc aaacgcgttt tggtgttttg
1081 ctgtatcttt gttttgcttg tccttttagt ttaagagttt tgccccagca tctcagagat
1141 acttgtgaat aatcaccaaa atggccctta ttttgtatat ttcgtttact tgttcctttc
1201 ttatttgtag tttgtggttc attcttagtt tttcttgtgg tttatgtgca agataactta
1261 gagtaacgtt cctgatggag tttggagtgt atttaaatga ttcgagttag tttttccctg
```

Figure 15: DNA sequence of Felis Catus (Cat) [19]

7. Test Results

Comparison of Human, Chimpanzee and Banana

	Alpha	Human	Chimp	Banana
1	0.10	99.981	99.949	89.933
2	0.20	99.979	97.816	87.154
3	0.30	99.978	95.342	81.706
4	0.40	99.975	94.721	74.585
5	0.50	99.972	92.808	70.633
6	0.60	99.971	90.368	63.707
7	0.70	99.965	89.886	59.961
8	0.80	99.962	88.386	54.822
9	0.90	99.955	86.371	52.666
10	1.00	99.951	84.731	49.595

Table 2: Comparison of Human, Chimpanzee and Banana DNA

The above table shows that the DNA of chimpanzee has 84% similarity with Human DNA and DNA of banana is 49% similar to human DNA.

Comparing Human, Chimpanzee and Mouse

	Alpha	Human	Chimp	Mouse
1	0.10	99.981	99.949	97.933
2	0.20	99.979	97.816	97.154
3	0.30	99.978	95.342	96.706
4	0.40	99.975	94.721	94.585
5	0.50	99.972	92.808	93.633
6	0.60	99.971	90.368	92.707
7	0.70	99.965	89.886	91.961
8	0.80	99.962	88.386	89.822
9	0.90	99.955	86.371	82.666
10	1.00	99.951	84.731	81.595

Table 3: Comparison of Human, Chimpanzee and Mouse DNA

The above table shows that the DNA of chimpanzee has 84% similarity with Human DNA and DNA of banana is 81% similar to Mouse DNA.

Comparing Human, Monkey and Fruit Fly

	Alpha	Human	Monkey	Fruit Fly
1	0.10	99.981	99.941	79.103
2	0.20	99.979	97.814	72.974
3	0.30	99.978	95.360	66.286
4	0.40	99.975	94.722	64.605
5	0.50	99.972	92.800	61.993
6	0.60	99.971	90.363	57.127
7	0.70	99.965	89.898	53.581
8	0.80	99.962	88.545	49.232
9	0.90	99.955	86.371	46.116
10	1.00	99.951	84.931	44.685

Table 4: Comparison of Human, Monkey and Fruit Fly DNA

The above table shows that the DNA of monkey has 84% similarity with Human DNA and DNA of Fruit Fly is 44% similar to human DNA.

Comparing Human, Dog and E. Coli (bacteria)

	Alpha	Human	Dog	E. Coli
1	0.10	99.981	97.923	39.202
2	0.20	99.979	94.701	32.346
3	0.30	99.978	92.456	29.282
4	0.40	99.975	89.980	22.167
5	0.50	99.972	86.976	17.593
6	0.60	99.971	85.049	12.152
7	0.70	99.965	83.728	09.361
8	0.80	99.962	82.983	07.991
9	0.90	99.955	80.624	05.668
10	1.00	99.951	77.828	03.120

Table 5: Comparison of Human, Dog and E. Coli DNA

The above table shows that the DNA of Dog has 77% similarity with Human DNA and DNA of E. Coli is 3% similar to human DNA.

Comparing Human, Mouse and Yeast

	Alpha	Human	Mouse	Yeast
1	0.10	99.981	99.191	58.111
2	0.20	99.979	97.664	52.912
3	0.30	99.978	95.850	49.282
4	0.40	99.975	94.102	46.629
5	0.50	99.972	92.810	41.908
6	0.60	99.971	90.303	37.133
7	0.70	99.965	89.678	34.592
8	0.80	99.962	88.685	31.225
9	0.90	99.955	87.371	29.193
10	1.00	99.951	86.931	27.662

Table 6: Comparison of Human, Mouse and Yeast DNA

The above table shows that the DNA of Mouse has 86% similarity with Human DNA and DNA of Yeast is 27% similar to human DNA.

Comparing Human, Fruit fly and Weed

	Alpha	Human	Fruit Fly	Weed
1	0.10	99.981	78.717	58.125
2	0.20	99.979	71.285	52.936
3	0.30	99.978	67.453	49.222
4	0.40	99.975	64.636	46.695
5	0.50	99.972	62.125	42.901
6	0.60	99.971	59.984	33.198
7	0.70	99.965	55.920	29.598
8	0.80	99.962	52.615	25.233
9	0.90	99.955	48.331	22.180
10	1.00	99.951	44.231	18.690

Table 7: Comparison of Human, Fruit Fly and Weed DNA

The above table shows that the DNA of Fruit Fly has 44% similarity with Human DNA and DNA of Weed is 18% similar to human DNA.

Comparing Human, Cat and Cow

	Alpha	Human	Cat	Cow
1	0.10	99.981	98.717	97.989
2	0.20	99.979	98.219	96.026
3	0.30	99.978	95.420	94.894
4	0.40	99.975	93.685	92.695
5	0.50	99.972	91.133	89.430
6	0.60	99.971	89.993	88.925
7	0.70	99.965	88.913	86.686
8	0.80	99.962	86.215	82.135
9	0.90	99.955	85.931	79.248
10	1.00	99.951	84.231	76.666

Table 8: Comparison of Human, Cat and Cow DNA

The above table shows that the DNA of Cat has 84% similarity with Human DNA and DNA of Cow is 76% similar to human DNA.

Comparing Human, Dog and Mushroom

	Alpha	Human	Dog	Mushroom
1	0.10	99.981	97.923	89.471
2	0.20	99.979	94.701	82.895
3	0.30	99.978	92.456	79.346
4	0.40	99.975	89.980	77.908
5	0.50	99.972	86.976	69.786
6	0.60	99.971	82.049	66.012
7	0.70	99.965	78.728	61.623
8	0.80	99.962	76.983	54.979
9	0.90	99.955	75.624	49.801
10	1.00	99.951	77.828	42.213

Table 9: Comparison of Human, Dog and Mushroom DNA

The above table shows that the DNA of Dog has 77% similarity with Human DNA and DNA of Mushroom is 42% similar to human DNA.

Comparing Human, Dog and Rice

	Alpha	Human	Dog	Rice
1	0.10	99.981	97.923	58.309
2	0.20	99.979	94.701	46.786
3	0.30	99.978	92.456	41.523
4	0.40	99.975	89.980	37.960
5	0.50	99.972	86.976	33.986
6	0.60	99.971	82.049	29.112
7	0.70	99.965	78.728	25.011
8	0.80	99.962	76.983	22.951
9	0.90	99.955	75.624	18.208
10	1.00	99.951	74.828	15.420

Table 10: Comparison of Human, Dog and Rice DNA

The above table shows that the DNA of Dog has 74% similarity with Human DNA and DNA of Rice is 15% similar to human DNA.

Comparing Human, Cow and E. Coli(bacteria)

	Alpha	Human	Cow	E. Coli
1	0.10	99.981	97.130	39.202
2	0.20	99.979	94.195	32.346
3	0.30	99.978	92.222	29.282
4	0.40	99.975	89.900	22.167
5	0.50	99.972	86.928	17.593
6	0.60	99.971	82.022	12.152
7	0.70	99.965	81.123	09.361
8	0.80	99.962	79.646	07.991
9	0.90	99.955	77.186	05.668
10	1.00	99.951	76.925	03.120

Table 11: Comparison of Human, Cow and E. Coli DNA

The above table shows that the DNA of Cow has 76% similarity with Human DNA and DNA of E. Coli is 3% similar to human DNA.

Following is a table which shows the similarity between different species. For example, the Human and Chimps are 87% similar (84% according to our test result), Dog and Mouse are 82% similar (87% according to our test result). The results below are almost in accordance with the tests we have conducted.

Homologs	Human	Chimp	Dog	Mouse	Rat	Fruit Fly
Human	--	29529 87% 84%	27761 81% 77%	26830 79% 81%	23860 70% 73%	13276 39% 44%
Chimp	18898 87% 84%	--	16865 78% 71%	16194 75% 79%	14283 66% 68%	7673 35% 38%
Dog	28144 82% 77%	27139 89% 82%	--	26740 88% 91%	23816 78% 74%	22771 75% 69%
Mouse	16384 83% 81%	15674 82% 78%	16066 84% 87%	--	14067 74% 76%	7887 41% 45%
Rat	12409 70% 73%	11907 90% 92%	12184 92% 89%	12420 94% 91%	--	6592 50% 49%

Table: Homologous gene Summary Chart [21]

8. Future Work

Although 1.8 million species are discovered today, all their DNA nucleotides are not easily accessible to study the differences and the similarities between these organisms. Also, DNA can be represented in 3D structures [12][20] depending on the behavioral patterns of proteins in the amino acids. This can be achieved in future research.

9. Conclusion

Pattern recognition of sequential symbolic data using automata theory was proposed in 2005 by Dr. Lin [1] and is being researched since then by him and his students. His student, Nikhil Kalantri has proposed an approach for author identification using the Alergia algorithm for pattern recognition.

In this project, two or more species can be compared on the basis of their DNA genome. The nucleotide sequences help us understand and learn the theory of life and the evolution of living organisms by comparing two species or by comparing the two organisms of the same species. For mathematical results, theory of automata proves to be vital importance. A PTA formed by the use of Alergia helps us understand the DNA genome in a better way.

10. References

1. P.Baliga and T.Y.Lin: Kolmogorov Complexity Based Automata Modeling for Intrusion Detection. Proceeding of the 2005 IEEE International Conference on Granular Computing, " July 25-27, Beijing, China (2005)

2. Pevsner, J. & Wiley, J.: Bioinformatics and Functional Genomics. (2003)

3. M.Young-Lai and F.Tompa: Stochastic Grammatical Inference of Text Database Structure. Machine Learning (2000)

4. Bosnacki D., Eikelder H.M.M., Steijaert M., Vink E.: Stochastic Analysis of Amino Acid Substitution in Protein Synthesis. In: CMSB 2008, LNBI 5307, 367–386, Springer-Verlag Berlin Heidelberg (2008)

5. Cotter, N., Gesteland, R., & Murdock, M.: Neural network based pattern recognition for sequenced DNA autoradiograms. In: International Joint Conference on Neural Networks, 2, 909 (1991)

6. Burks C., Farmer D.: Towards Modeling DNA Sequences as Automata. In: Physica D: Nonlinear Phenomena, Volume 10, 157-167 (1984)

7. Information about existing species: http://www.backyardnature.net, [Online – May 2013]

8. Total number of estimated species on Earth: http://www.plosbiology.org, [Online – June 2013]

9. Martin, J. C. & Hawk, J. F.: DNA sequence analysis by optical pattern recognition. In: The International Society for Optical Engineering, 938, 238-45(1988)

10. Anderson C., Brunak S.: Representation of Protein Sequence Information by Amino Acid Subalphabets. In: American Association for Artificial Intelligence, Volume 1, 97-104 (2004)

11. Crick F.: The Great Ideas of Today 1980, Encyclopedia Britannica, 644-683(1980)

12. Paul Barry, Michael Moorhouse.: Bioinformatics, Biocomputing and Perl: An Introduction to Bioinformatics Computing Skills and Practice (2004)

13. Pierre Baldi, Soren Brunak.: Bioinformatics: The Machine Learning Approach (Adaptive Computation and Machine Learning)

14. Rajeev Motwani, Jeffrey D. Ullman, John E. Hopcroft.: Introduction to Automata Theory, Languages, And Computation (2003)

15. Jir Poner , Filip Lanka.: Computational Studies of RNA and DNA. Challenges and Advances in Computational Chemestry and Physics (2006)

16. Ferdinand Wagner, Ruedi Schmuki, Thomas Wagner, Peter Wolstenholme.: Modeling Software with Finite State Machines: A Practical Approach (2006)

17. Justin Davis.: Finite State Machine Datapath Design, Optimization, and Implementation (Synthesis Lectures on Digital Circuits and Systems) (2008)

18. David J Corner.: Digital Logic and State Machine Design (The Oxford Series in Electrical and Computer Engineering) (1994)

19. DNA Sequences- http://www.ncbi.nlm.nih.gov/gene, [Online – June 2013]

20. Peter H. Raven, George B. Johnson, Jonathan B. Losos, and Susan R. Singer, Biology (7th edition).

21. Comparison of multiple species- http://eugenes.org/, [Online – May 2013]