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

Parnika P. Achrekar
San Jose State University

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

Part of the Computer Sciences Commons

Recommended Citation
DOI: https://doi.org/10.31979/etd.4ysq-wbmt
https://scholarworks.sjsu.edu/etd_projects/334

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.
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

1. Introduction..........................................................................................................................9
2. DNA Sequencing...................................................................................................................12
3. Understanding Automata.......................................................................................................17
4. Alergia Algorithm ..................................................................................................................19
5. Creating SFA using alergia algorithm..................................................................................22
6. DNA Samples of Living organisms.......................................................................................25
7. Test Results ............................................................................................................................33
8. Future Work ............................................................................................................................42
9. Conclusion .............................................................................................................................42
10. References.............................................................................................................................45
List of Tables

Table 1: List of Amino Acids...........................................................................................................14
Table 2: Comparison of Human, Chimpanzee and Banana DNA ........................................33
Table 3: Comparison of Human, Chimpanzee and Mouse DNA ...........................................34
Table 4: Comparison of Human, Monkey and Fruit Fly DNA .............................................35
Table 5: Comparison of Human, Dog and E. coli DNA .........................................................36
Table 6: Comparison of Human, Mouse and Yeast DNA .....................................................37
Table 7: Comparison of Human, Mouse and Yeast DNA .....................................................38
Table 8: Comparison of Human, Cat and Cow DNA ............................................................39
Table 9: Comparison of Human, Dog and Mushroom DNA ................................................40
Table 10: Comparison of Human, Dog and Rice DNA ..........................................................41
Table 11: Comparison of Human, Cow and E. Coli DNA ......................................................42
Table 12: Homologous Gene Summary Chart .........................................................................43
List of Figures

Figure 1: Process of Transcription ................................................................. 12
Figure 2: Process of Translation ................................................................. 13
Figure 3: Amino Acids Chart ................................................................. 15
Figure 4: Total Number of Species on Earth .................................................. 25
Figure 5: Relative Number of Named Species ................................................ 25
Figure 6: DNA sequence of Human ............................................................ 26
Figure 7: DNA sequence of Chimpanzee ..................................................... 27
Figure 8: DNA sequence of Monkey .......................................................... 28
Figure 9: DNA sequence of Mus Musculus (House Mouse) .......................... 28
Figure 10: DNA sequence of Banana .......................................................... 29
Figure 11: DNA sequence of Weed ............................................................ 29
Figure 12: DNA sequence of Drosophila Melanogaster ............................... 30
Figure 13: DNA sequence of Oryza sativa (Rice) ......................................... 31
Figure 14: DNA sequence of Agaricus bisporus (Mushroom) ....................... 32
Figure 15: DNA sequence of Felis Catus (Cat) ............................................. 32
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].

![Diagram of the translation process]

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:

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

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].

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:

\[ A = (S, P, i, \delta, T) \]

- \( S \): 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
- \( \delta \): 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]

<table>
<thead>
<tr>
<th>Deterministic Finite Automata</th>
<th>Non Deterministic Finite Automata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characterized as a 5 tuple state:</td>
<td>Characterized as a 5 tuple state:</td>
</tr>
<tr>
<td>( &lt;S, A, T, s_0, F&gt; )</td>
<td>( &lt;S, A, T, s_0, F&gt; )</td>
</tr>
<tr>
<td>( S ) is the set of states</td>
<td>( S ) is the set of states</td>
</tr>
<tr>
<td>( A ) is the alphabet</td>
<td>( A ) is the alphabet</td>
</tr>
<tr>
<td>( T ) is the transition function:</td>
<td>( T ) is the transition function:</td>
</tr>
<tr>
<td>( S \times A \rightarrow S )</td>
<td>( S \times (A \cup {\varepsilon}) \rightarrow PS )</td>
</tr>
<tr>
<td>( s_0 ) is the initial state</td>
<td>( s_0 ) is the initial state</td>
</tr>
<tr>
<td>( F ) is the set of accepting states.</td>
<td>( F ) is the set of accepting states.</td>
</tr>
</tbody>
</table>
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
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, \( S = \{ s_0, s_1, s_2, s_3, \ldots s_n \} \) and \( s_i = a_1 a_2 a_3 \ldots a_i \).

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:

\[
| p - \frac{f}{n} | < \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \text{ with probability larger than } (1 - \alpha).
\]

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.

\[
| \frac{f}{\tau_b} - \frac{f^*}{\tau_{b^*}} | > \sqrt{\frac{1}{2} \log \frac{2}{\alpha} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n^*}} \right)}.
\]

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].

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 4: Total Number of Species on Earth [8]

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  attccagct ttctatgcat tctggcataa gctagtctca caagccagag gacagccctt
61  gegeaagaag tttggcact ggcttgggga atagaagcca ctttaagcgc tggggagaag
121  gaacaccegg aaaaaccaaa aaaatggatcc a tgtgagcctt ggcaatagca caaatgcaat
181  gacccacttt gagaacactt taccacaatt taccacagtt gaccaagtct tacacacacac
241  acctgacacc ttcctcaagg accataacgc aatctttcgg tctgttgtt gccctcaccac
301  ococacacac cttcagagcc cccactgctt gactggagcc atggccacat cccaccaacc
361  aagttgcttt cctgtaaacc gctttttggg gctttgggca cctttttgtt gtttttgttgg
421  aacatcagtt cagaacccaa aacatgggaa cttatctctt tggctacacc tggagagacg
481  tgaagcccat tttctttagt taacttacca gaaacacaa taccgtttgt tctcctttat
541  aagtacagcc taacaccctgg ttactctag tcatctttg ctatatagac aggaaataa gacactggtt
601  aatcttggag ggggagaagg gggaaaaaggg gcagcgtcga aacactttct gttttgtact
661  agttctaggg ctaagcttgac agatctttcc ataccacaag cttaatctat atgtaagttg
721  ctacagcttc aaccttcgcc tcctttcacc tgaacctttt tagaacttac atctcaacaa
781  caaaaaatca aataaaaaac aagtatctcc tctactgtg cgtctaaagca tggctttttt
841  tcctttgggt gcaagatnct cccacanagc gatcttttct ctcttttttct gttttttctc
901  agagcagcag tccaccccttt taatattatt gtgggataac cccaaacacg gggagatagt
961  gatgatgctt ttgctaaaaa aaaaaaaaaa gtaagaccca ttcctcctac aacaaatcct
1021 atatagattaggg ccgggcggcg cggctgtcgc ctgtatcttc cagagctaaggg
1081  ggcgcagcag cggccagctt ataaggtgtg aagctgctgg taccacagct
1141  aacccccagct ttctcttttt taactttaaa atacgctttg gtttttgctgg gcgtgctctg
1201  tccagccatt ttaagggttg gaagcactgg ggtgcgtcat aacccgaggg cggagttgc
1261  agttgagagg gatcttctgc tggctttcgt cggattgga cggagcttgc ccccttaacc
1321  aaaaaataagtc aataactttt tccacacact aattacctct gctttggtgg
1381  aaccccgatt tgggtttttg aggctggttg aagaattttc tctctttttt tttttgttatt
1441  tcgggtgctt taacagtact tcaacagcag atagtttttt tctttttgta tcaagaaact
1501  gttttgtcct tttaagcttg ttcctttttc tttactgtt tttatagag tggctttgtc
1561  ttatattttt atagttctcc atctgggctg agggaagggc gcaatctcag ttgggtggtg
1621  ttgggtacct tttatatttttt tcttttcta ttaagatttt atgtgtgta tcaagacgcc
1681  aagagagctt aagacagagc gctgggttcct caagggagga ttataagttta tggtaagaac
1741  tgtctttgaa tttggttgatt tttctttatta attggtgaac ctttttaccttc
1801  aagagagctt aagacagagc gctgggttcct caagggagga ttataagttta tggtaagaac
1861  aagagagctt aagacagagc gctgggttcct caagggagga ttataagttta tggtaagaac
1921  tggggggttc tggggggttc tggggggttc tggggggttc tggggggttc tggggggttc
1981  agtttttacctttttgtttt tttttttttttt tttttttttttt tttttttttttt tttttttttttt
2041  ttatatctgt gttcagacat acctgtttgg gttgggggg gaaagacttatttggacac
2101  cagatgctct ttttattttg atagtttaat aagactttta aagatcttaa tttttaacctcc
2161  ttggtttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
DNA nucleotide sequence for Chimpanzee [19]:

```
1 cccgctgtctc ggyttgggtgc aatatttggg gcttaaggcc ttcctctcgc ttaagttttt ggtttttttt ctcctctcgc ttcttttta ttagccytac gggctgtgg ccagggaggg
ggcgggag ccagggaggg ccagggaggg ccagggaggg ggcgggag ccagggaggg ggcgggag ccagggaggg ggcgggag ccagggaggg ggcgggag ccagggaggg ggcgggag ccagggaggg
```

Figure 7: DNA sequence of Chimpanzee [19]
DNA nucleotide sequence for Monkey [19]:

<table>
<thead>
<tr>
<th></th>
<th>aagcttcttc gcggcgaacta ttctcataat cgccacacgga ctcctctct ccctgctatt</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>ctgcttaccc aatctcaact atgaacggac ccacagctgt ttataatgc ctcctcgaag</td>
</tr>
<tr>
<td>121</td>
<td>aacctcagcc ttacctccct taaacctgct tcagctttcc gcggacaaat ccctcaactt</td>
</tr>
<tr>
<td>181</td>
<td>agctctccct cccatactga atcctttgtc aaagatcctt ccctaataagt gtccttcc</td>
</tr>
<tr>
<td>241</td>
<td>ttccatccagt atcaataagt gtcgaagaata ttcataaatc atccctggtcc</td>
</tr>
<tr>
<td>301</td>
<td>ctctcatttt caaatcataa ccctacccat aatactcttg tcctctctct cttacaatct</td>
</tr>
<tr>
<td>361</td>
<td>gcctgcggaa ccgacgatcc aacgcataaat atcttttct cttgcccttt cttaaggttt</td>
</tr>
<tr>
<td>421</td>
<td>accttattct cccctatttt ttctctcttt gcgttagatt tacctttagt aagctatgtt ttaataaat</td>
</tr>
<tr>
<td>481</td>
<td>aagctttgctt cctagggctc gcctaggttt atctctctat atttctttcg aagctgattca</td>
</tr>
<tr>
<td>541</td>
<td>aggattgcte acctcgttc ccctaaaataa aactattttc ttcattcctt ttaataaat</td>
</tr>
<tr>
<td>601</td>
<td>atagccatcc attgctcca ccagcaaaaa atatctgagg aactcataat aaactgataa</td>
</tr>
<tr>
<td>661</td>
<td>atattgaca actccctatc atactcagcct cccctcttcc tcctccctcc gccctctctc</td>
</tr>
<tr>
<td>721</td>
<td>acccttatta atccttatca aatacactta atcataatac atgttaaac aacatcataa</td>
</tr>
<tr>
<td>781</td>
<td>tataaaatat gctttttatt tttttttttt tttttttttt ttttttttttt ttttttttttt</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>cttccggatc acagtggttac aactactctg gaaactaatt tgggaggttc tcacaaaaatt</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>tatacttttg cttccgggta ccctctgccc acctctctcc ggcctctttcc aagactgtct</td>
</tr>
<tr>
<td>121</td>
<td>ccaactcttg atactccagc agctctctctat gctcttctct gcctctctct cttactctc</td>
</tr>
<tr>
<td>181</td>
<td>agaagctttc gacaacctcc actgctctcc gcggagagat gttagctttt gttgcttttc</td>
</tr>
<tr>
<td>241</td>
<td>tttaaagagaa aagctgctct cctctaggg gttgctctcc cttactcatt aagctgcttt</td>
</tr>
<tr>
<td>301</td>
<td>aaagggcgtg cttctttgg gcgcctcttt cctctctct ggaagagctc gcctctctcc</td>
</tr>
<tr>
<td>361</td>
<td>cctctttttc cttctctct cctctctct cttactctct cctctctct cttactctct</td>
</tr>
<tr>
<td>421</td>
<td>ttccagagaa aagctgctct cctctctct cttactctct cctctctct cttactctct</td>
</tr>
<tr>
<td>481</td>
<td>ctccagagaa aagctgctct cctctctct cttactctct cctctctct cttactctct</td>
</tr>
<tr>
<td>541</td>
<td>cttccagagaa aagctgctct cctctctct cttactctct cctctctct cttactctct</td>
</tr>
<tr>
<td>601</td>
<td>aacccgagcc agtctttgag gcgcaaaaaa ccacttacct tttctctctt cttctctct</td>
</tr>
<tr>
<td>661</td>
<td>ttctcttttt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>721</td>
<td>tcaacccaaac ccctcttttt cttctctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>781</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>841</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>901</td>
<td>gctctctctt ccctcttttt cttctctct cttactctct cttactctct egaactgagat</td>
</tr>
<tr>
<td>961</td>
<td>tataagcgaa aagcgagagc ggctcttttt cttctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1021</td>
<td>gctctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1081</td>
<td>aacccagagc cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1141</td>
<td>aacccagagc cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1201</td>
<td>tataagcgaa aagcgagagc ggctcttttt cttctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1261</td>
<td>ttttcttttt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1321</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1381</td>
<td>ttttcttttt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1441</td>
<td>ttttcttttt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1501</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1561</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1621</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1681</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1741</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1801</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
<tr>
<td>1861</td>
<td>cttctctctt cttctctct cttactctct cctctctct cttactctct cttactctct</td>
</tr>
</tbody>
</table>
DNA nucleotide sequence for Banana [19]:

```
 1  tggattttaa  gctgggttta  aagattcaca  attcaattat  tatactcttg  actacgaagt
 61  caaagatact  gatacttgg  cagcttttcg  agtaacttcg  caacotgga  ttgcccgcga
121  aagaacaggg  ggtgctggag  ttgctgaact  tttaactggt  atacgagaca  sttgtgggac
181  tgaatgacct  acagctttgg  ctggagctgc  tatactcttg  aggcccgtgc  tcctttcctt
241  tggagctggt  aacaactata  tggctttagt  agttaacttc  tcgagccattt  ttaaagaagg
301  ttctgttaact  acactttta  cttctcttgg  ggttaatgta  tttggtttaa  ttaaagcttg
361  cgcttacgct  ttgaggtcct  tcgagactgc  cactttcatt  tccaaacttt  tccaaagccc
421  gctctacgagc  attcactggg  aagagcataa  ttgtctcaac  tagtctcctg  ccccttttgg
481  atgactctac  ttgtaaacttt  tgcgaacattc  tgggttttaa  tgcgctttaa  ccccttttgg
541  atgactctac  ttgtaaacttt  ccccttttgg
601  tttggtttaa  agagctggtg
```

Figure 10: DNA sequence of Banana [19]

DNA nucleotide sequence for Weed [19]:

```
 1  atgcattgca  tggtgtccct  ccatttctcg  cctctcctat  atttctgagt
 61  acacactatta  ataaaccttt  ttctgggata  gaatttttaa  cacaatttaa  atcaaatatt
121  aacaacaaaa  aatcaagatc  aagacctaatt  tgcaacccaa  tccaaagtta  ttgggcaaa
181  ctgccataag  tgcttactac  tatactcttt  aagagatcag  ccaaaactaa  ccccttttgg
241  taccagagt  ctaactttct  gttcctcctaa  tgaacataaa  tccaaagcct  ctaccatgct
301  cagacagtaa  acgtggaatg  aagagctgag  cggacagtta  ttggaatgga  aacatcttta
361  ctccacactt  tgcggagctg  tagagttcct  cgccttttca  gcaaaattca  tttgccccaa
421  gaatatcttt  ctatgttttt  agaattttcct  aacaatatttt  gaaatgatttt  atcaaatatttt
481  tagagtttata  atgcctttgc  ttcacaacttt  agGGGTTag  cagcataatgg  atacgcttttaa
541  ttacctcaaag  ttctgtttg  tgttttttgt  gcgtaaagc  agcttctttc  ccccttttgg
601  aagagatcag  ccaaaactaa  ccccttttgg
661  aaaaatatatt  ttctgggata  gaatttttaa  cacaatttaa  atcaaatatttt
721  atgcattaacc  aaacactaat  attaactatt  atatactggct  aacatcttgct  acctgccttt
781  gaaggaacat  cctttttctg  tcatataacttt  gtcgcctttc  gcaaatcttt  ccccttttgg
841  gagaacagac  aagacatcaca  tggactttggt  cctggagctgt  ggcctttttc  ccccttttgg
901  ttctgctcata  attcctgca  acgttctttc  cgggtcgact  gctgagtttt  acatctcttt
961  ctgcttctctt  ataatgctttt  ttaagtttctt  gatagtttggt  tagagtttctt  tctctgctgg
1021  gatgtagttc  atanctgcct  aacatctttc  ctcatacttt  gacatatttt  aagaaatatttt
1081  gctcataata  cacttataact  aggctattag  aagatatttt  ggttctaaa  gatattttgtt
1141  cctctcacta  aggctttttt  gagtagtggt  ccgagtcttta  tcaacacacat  ccccttttgg
1201  atcaagttcc  cttctccttc  ttttatttatt  agccttttgc  aagctttttt  ccccttttgg
1261  gacatatttt  gatattttgtt  cggctctttt  aatatttttt  ggttctaaa  gatattttgtt
1321  tcaacacacat  aacatctttc  ctcatacttt  gacatatttt  ggttctaaa  gatattttgtt
1381  atcatctttc  aatatttttt  ggttctaaa  gatattttgtt  ccaatatttt  ccccttttgg
1441  cctctcacta  aggctttttt  gagtagtggt  ccgagtcttta  tcaacacacat  ccccttttgg
1501  gatattttgtt  cggctctttt  aatatttttt  ggttctaaa  gatattttgtt
```

Figure 11: DNA sequence of Weed [19]
DNA nucleotide sequence for *Drosophila Melanogaster* (Fruit Fly) [19]:

```
1  gaattctttg aatatcoca gtcctagttac gcacccctctt caaccaggccg catttgacaa
 61  cattgtcgtt gagggtgtt gtcgctactt cgaagattag aaaaattttgc ttcttcgctg
121  tggaccaccc ctctcaccct agatctttttg ccagccgttt gctgataatt tctagtggtg
181  agcccaatcc caacgcattt cgggtttttcc ctgcaaccac aatgttgtg aaaaaataaa
241  ctgttgctgg gagggtcttt tccaggggg ctaataaca cacataagat aatcatcacc gtcctgaccc
301  gtgtctgccc gtcggctttg ttgcttactt cagatccata cttatatatg gatataatgt
361  ccagctctgc aagctgtcctt cggaggtttg ggtctgtgctt tctgtattat gttttttagg ctcctcctag
421  tgaagacctt ctctgtgcct ctagcttatt cgttattataa ttacccgtcct atagacacag
481  tgcgtcgtgatt ccagcttatt cgcattctcc gccgacctcg agtccaaatc ctatgcatat
541  gcctccgcgc acagccgccg atatcgagtc gttccagaaaa gatgttgctgc ctctcttgtc
601  ctgtgactcc agcagcaaga cttctctttg gacgccgatt gttggcctttt tggagtcgcc
661  atcgctgata tggctctcct gctgtatata cgcaccaaca acacagccag agccagatgt
721  ccagagaaaa cttgaaatgt ggcgccgcttc tcagctgcac ggcgagaccc ctggtggtgg
781  cgttgtctca ctgttggagc accgtgctcc gcggctggcc tgggtcttgg tgtgcttgga
841  cgagccgacc gcggaaaaac cggctctctg caggcggcag acatattcct agtygatag
901  gaccctgtgc acrgttgtgc gaccccaacga gtcgctcttg tggaggtggt ctctctgctc
961  cagccccagc cgccctcccg tccaccccgcc atcgccctcc ccccgccacc ccacagccga
1021  gcggctgtgcg acgtctctcgcc tctgtcgctc ttcagctgcg caatctctgcg tggttctccg
1081  caccctcgct ggacccgcgt gcctcgatct gctcgatagtc agctgccggg ccgggccccag
1141  gaaatttcgg gcggcggaaa acagccagga taagggctcc ccagacactt cattcgagca
1201  agcgccgaca gcacacaggc ggtggcctgg tgcgctctct caccctcttc cccaggggtt
1261  tctctggtgg agacgggctt cctggaatt ttataagggc agttcaagcc tccacgacc
1321  tgcgctgatt gcctcctcct cctctcttcc ctgccgcttc ccacccgacc ccacagccga
1381  gctgaggcag ccgaaaaacg gcggccagct gcaggatgct tggctctgg ccagccgaga
1441  gctgggaaaat cgcaagcgcg cgggctggcc ccagctccag gcgcagcaac aatggccggc cgggccccag
1501  gatgaaatt cgttctgac ggcgtccttt gcggctcatt gttttcgggc aaaaaattcc ggcgttatgg
1561  ttcagctctcg gacgacctcc cggccagcct gcagcagatg cgggcacacc ccacagccga
1621  ggagcaagag ctgcgctcag tggccagtct gaggctagct gtcctcattt ttcacaggg
1681  caagggaaag gcgggaggg gaacgcgacc gcgtcagagc cggccagctg cgggctaaaa
1741  gcgtattgga gaggctgggg tagactctgga ggtgccctcg cggagctgcc aaaaaaacgg
1801  caaatcgcgc acggcaaat ttcgctggcc ggggcccccg ccgggccccag ctcctattga
1861  cttccnncgc gcctggaacc gcacagtcgg ttgtggaacc ctgcttgccg gcctgaggga
1921  tctctctctg cccacacacc gcacacatca aatggtgagg aagggacagc tctctctctc gcctgagggg
1981  cttatctgtt cggcgagccc gcacacaca ggtjtgagg acatggttac gcctgagcggg
2041  ggaggccctac gtcgctgact tgcaggaaca agtgctagtct gcctctcttg cgggctgaaat
2101  gtgggctctg gcgtgctgaa gaaatggagc atggatgagc aagtttacac ggctctgagg
2161  gcgtaatgcgc caatgtgagc ctcgggagg aagtgcaagc cagctgcggg aagttgacca
2221  cgaactcggag ctgctcctgc gcgaggagat gctgctgagg aatgggcccc aaagaggagt
2281  gctgctggag cggcgagctcc ccatgctgatg gcgtgctgcc cttgccctcc gcgggccccag
2341  gtttctgagg gcgctgcgag cgtctgcgaa gccaccaact gattctggg gctcgtcctcc
2401  tagagggcag cttgcgagg gcgtgcgagct gcgctgatgc cttcgtgagg gctcgtcctcc
2461  ccccctctgg ctcggccgcc gcgctgtcgc ccgctgctac gcgtgctgcc ctcctctcag
2521  cccagcggct ctgctgctgc gcctgcgagc ccctgcgagc gcgcctgctg gactgcctgc
2581  gtcacctcggt gcggcgtcct gcggcgtcct gcgctgtgct gcggcgtcct gcggcgtcct
2641  ctcggtcgctt gactgtctggt cggagtcagc gacgctgcgtt ccaccagate
```

Figure 12: DNA sequence of *Drosophila Melanogaster* [19]
DNA nucleotide sequence for *Oryza sativa* (Rice) [19]:

```
  1 ttccaaacaa atgtgctcatt ggtottcggaa attcctedto ttgcacattgt gcaccacccgaa
  61 gctagtggacctctacgagactgtcttgtatatg atacattctttgtgtaacctgtctg
 121 taaggttggggcgctttattt gctgctttttg tttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
DNA nucleotide sequence for *Agaricus bisporus* (Mushrooms) [19]:

<table>
<thead>
<tr>
<th>1</th>
<th>aaccaggtgct</th>
<th>cattttcttt</th>
<th>tccttttttc</th>
<th>caacccctct</th>
<th>ctttcaccct</th>
<th>cttcgcctgt</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>tggctgcgcc</td>
<td>ggttggcatac</td>
<td>ccctatcgga</td>
<td>tccttttttc</td>
<td>caacccctct</td>
<td>ctttcaccct</td>
</tr>
<tr>
<td>121</td>
<td>ccctcctcatg</td>
<td>ctaagggcag</td>
<td>ttatcagaa</td>
<td>ggtgaagctc</td>
<td>tccttttttc</td>
<td>caacccctct</td>
</tr>
<tr>
<td>181</td>
<td>ggttcgcttg</td>
<td>ttctattctg</td>
<td>cttgcctcag</td>
<td>ggttcgcttg</td>
<td>ttctattctg</td>
<td>cttgcctcag</td>
</tr>
<tr>
<td>241</td>
<td>caggttttgt</td>
<td>tcctgttgtt</td>
<td>gggttgttct</td>
<td>caggttttgt</td>
<td>tcctgttgtt</td>
<td>gggttgttct</td>
</tr>
<tr>
<td>301</td>
<td>acgtgcttcg</td>
<td>ggtgtgtgtg</td>
<td>caggttttgt</td>
<td>tcctgttgtt</td>
<td>gggttgttct</td>
<td>caggttttgt</td>
</tr>
<tr>
<td>361</td>
<td>gacgacacga</td>
<td>ccggtgctgt</td>
<td>cgcctgtgtg</td>
<td>tgcctgtgtg</td>
<td>ccggtgctgt</td>
<td>cgcctgtgtg</td>
</tr>
<tr>
<td>421</td>
<td>ggtgctgtgt</td>
<td>cgcctgtgtg</td>
<td>cgcctgtgtg</td>
<td>tgcctgtgtg</td>
<td>ccggtgctgt</td>
<td>cgcctgtgtg</td>
</tr>
<tr>
<td>481</td>
<td>ttcttttttt</td>
<td>tccttttttt</td>
<td>tccttttttt</td>
<td>tccttttttt</td>
<td>tccttttttt</td>
<td>tccttttttt</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>1</th>
<th>ggcgggggga</th>
<th>gagggtcttca</th>
<th>agagacagca</th>
<th>aggaagtttt</th>
<th>ccatgggaca</th>
<th>gcocctggcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>tcacccgaaa</td>
<td>gcagctctgct</td>
<td>cgcctctcct</td>
<td>acaggggca</td>
<td>gcagctctgct</td>
<td>cgcctctcct</td>
</tr>
<tr>
<td>121</td>
<td>gcctggcaac</td>
<td>tcgctctgct</td>
<td>cgcctctcct</td>
<td>acaggggca</td>
<td>gcagctctgct</td>
<td>cgcctctcct</td>
</tr>
<tr>
<td>181</td>
<td>cggtcagcct</td>
<td>tgggctctgta</td>
<td>cgcctctcct</td>
<td>acaggggca</td>
<td>gcagctctgct</td>
<td>cgcctctcct</td>
</tr>
<tr>
<td>241</td>
<td>tcacggggcc</td>
<td>gggtggggttg</td>
<td>cgcctctcct</td>
<td>acaggggca</td>
<td>gcagctctgct</td>
<td>cgcctctcct</td>
</tr>
<tr>
<td>301</td>
<td>acgtgcttcg</td>
<td>ggtgtgtgtg</td>
<td>cgcctctcct</td>
<td>acaggggca</td>
<td>gcagctctgct</td>
<td>cgcctctcct</td>
</tr>
<tr>
<td>361</td>
<td>gacgacacga</td>
<td>ccggtgctgt</td>
<td>cgcctctcct</td>
<td>acaggggca</td>
<td>gcagctctgct</td>
<td>cgcctctcct</td>
</tr>
<tr>
<td>421</td>
<td>ggtgctgtgt</td>
<td>cgcctctcct</td>
<td>acaggggca</td>
<td>gcagctctgct</td>
<td>cgcctctcct</td>
<td>acaggggca</td>
</tr>
<tr>
<td>481</td>
<td>ttcttttttt</td>
<td>tccttttttt</td>
<td>tccttttttt</td>
<td>tccttttttt</td>
<td>tccttttttt</td>
<td>tccttttttt</td>
</tr>
</tbody>
</table>

Figure 15: DNA sequence of *Felis Catus* (Cat) [19]
7. **Test Results**

Comparison of Human, Chimpanzee and Banana

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

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.
### 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.
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.
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

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

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.
### 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

<table>
<thead>
<tr>
<th>Alpha</th>
<th>Human</th>
<th>Cat</th>
<th>Cow</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>98.717</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>98.219</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>95.420</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>93.685</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>91.133</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>89.993</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>88.913</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>86.215</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>85.931</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.951</td>
<td>84.231</td>
</tr>
</tbody>
</table>

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.
### 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

<table>
<thead>
<tr>
<th>Alpha</th>
<th>Human</th>
<th>Dog</th>
<th>Rice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>99.981</td>
<td>97.923</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>99.979</td>
<td>94.701</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>99.978</td>
<td>92.456</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>99.975</td>
<td>89.980</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>99.972</td>
<td>86.976</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>99.971</td>
<td>82.049</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>99.965</td>
<td>78.728</td>
</tr>
<tr>
<td>8</td>
<td>0.80</td>
<td>99.962</td>
<td>76.983</td>
</tr>
<tr>
<td>9</td>
<td>0.90</td>
<td>99.955</td>
<td>75.624</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>99.951</td>
<td>74.828</td>
</tr>
</tbody>
</table>

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.
### 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.

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

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


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


