Data Mining Solution for Transcription as a part of Gene Expression

Sukhjit Singh Sehra
Assistant Professor
Guru Nanak Dev Engineering College, Ludhiana, Punjab
sukhjitsinha@gmail.com

Sumeet Kaur Sehra
Assistant Professor
Guru Nanak Dev Engineering College, Ludhiana, Punjab
sumeetksehra@gmail.com

ABSTRACT
The biological data is available in different formats and is comparatively more complex. Knowledge discovery from these large and complex databases is the key problem of this era. Data mining and machine learning techniques are needed which can scale to the size of the problems and can be customized to the application of biology. In the present research work, transcription is simulated as a part of gene expression. It involves conversion DNA to pre-mRNA which is based on the concept of replacing Uracil with Thymine. Association mining is utilized to retrieve the results. The process of splicing is also implemented based on GU-AG rule which removes the introns from the input pre-mRNA sequence. The final output sequence is mRNA. The output sequence is mRNA. Induction and Cluster Analysis technique of data mining are used to retrieve the results.

Keywords
Gene Expression, Induction, Splicing, Transcription.

1. INTRODUCTION

1.1 Transcription
DNA is the main building block of a living organism. The information stored in DNA is used to make a more transient, single standard polynucleotide called RNA (ribonucleic acid). The process of making a RNA copy of a Gene is called transcription and is accomplished through the enzymatic activity of an RNA polymerase. There is a one to one correspondence between the nucleotide used to make RNA (G, A, U, C, where “U” is an abbreviation for Uracil) and the nucleotide sequence in DNA (G, A, T, C respectively). The next process of converting that information from nucleotide sequence in RNA to the mRNA. The first eukaryotic genomic sequence is obtained, which genes” intervening sequence” (introns) the interrupt the coding regions. RNA splicing is a process that removes the non-coding part (introns) and joins exons in a primary transcript. The introns are removed from pre-mRNA based on GU-AG rule. The sequence after skipping the intron part is m-RNA which contain exons only. The next process of converting that information from nucleotide sequence in mRNA to the amino acid sequence that make protein is called translation. The process of conceptual translation of gene sequences into corresponding amino acid sequence of protein using the genetic code [3][7].

1.2 Format of DNA Sequence
There are four basic types of molecules involved in life: (1) small molecules, (2) proteins, (3) DNA and (4) RNA. Proteins, DNA and RNA are known collectively as biological macromolecules. DNA is the main information carrier molecule in a cell. DNA may be single or double stranded. A single stranded DNA molecule, also called a polynucleotide, is a chain of small molecules, called nucleotides. There are four different nucleotides grouped into two types, purines: adenosine and guanine and pyrimidines: cytosine and thymine. They are usually referred to as bases (in fact bases are the only distinguishing element between different nucleotides, and denoted by their initial letters, A, C, G and T. This sequence is stored in text files. There is very long sequence (Chain of different combinations of ATCG) stored in the file. There are different formats that are used to organize this sequence data e.g. Plain Text, FASTA Format, Genbank, Genetic Computer Group Format (GCG) [5]. Plain text format is chosen for the present work as no further detail is needed for the problem.

1.3 Cluster Analysis Using Data Mining
Cluster analyzes the data objects without consulting a class label. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to in other clusters. Each cluster that is formed can be viewed as a class of objects from which rules can be derived. Figure 1.3 below shows how several clusters might form a hierarchy. When a hierarchy of clusters like this is created the user can determine the right number of clusters that adequately summarizes the data while still providing useful information (at the other extreme, a single cluster containing all the records is a great summarization but does not contain enough specific information to be useful) [2].
1.4 Induction
A database is a store of information but more important is the information, which can be inferred from it. There are two main inference techniques available i.e. deduction and induction. Deduction is a technique to infer information that is a logical consequence of the information in the database e.g. the join operator applied to two relational tables where the first concerns employees and departments and the second departments and managers infer a relation between employee and managers. Induction can be described as the technique to infer information that is generalized from the database as in the example mentioned above to infer that each employee has a manager. This is higher level information or knowledge in that it is a general statement about objects in the database. The database is searched for patterns or regularities [2].

2. METHODOLOGY
2.1 DNA to pre-mRNA Conversion
DNA is the main building block of a living organism. A cell acts as biological catalysts called as enzymes. Information stored in DNA is used to make more transient, single-stranded polynucleotide called RNA (ribonucleic acid). The process of making an RNA copy of a gene is called transcription and is accomplished through the enzymatic activity of a RNA polymerase. DNA to pre-mRNA conversion is done as shown in flow chart of fig.3.1. DNA sequence is accepted in the form of Plain Text format and then the extra spaces or gaps are trimmed from it. There is a one to one correspondence between the nucleotide used to make the RNA (G, A, U and C where U is an abbreviation for Uracil). In the process of DNA to RNA, T is replaced by U which results in RNA sequence.

Fig. 1: Hierarchy of Clusters.

Fig. 2: DNA to pre-mRNA Conversion
2.2 RNA to mRNA Conversion

The messenger RNA (mRNA) copies of prokaryotic genes correspond perfectly to the DNA sequence present in the organism’s genome with the exception that the nucleotide uracil (U) is used in place of thymine (T). In case, translation by ribosomes almost always begins while RNA polymerases are still actively transcribing a prokaryotic gene. Eukaryotic RNA polymerases also use uracil (U) in place of thymine (T), but much more striking differences are commonly found between the mRNA molecules by ribosomes and the nucleotide sequence of the eukaryotic gene. In eukaryotic genome the two step of genes expression are physically separated by the nuclear membrane, with transcription occurring within the nucleus and translation occurring after mRNA has been exported to the cytoplasm. RNA to mRNA conversion is computed in accordance to flow chart of fig.3.2. The RNA sequence is accepted in the form of plain text format. The first eukaryotic genomic sequence is obtained, which contain genes “intervening sequence” (introns) that interrupt the coding regions. RNA splicing is a process that removes the non-coding part (introns) and joins exons in a primary transcript.

![Fig. 3: RNA to mRNA Conversion.](image)

The different kind of introns have been found in eukaryotic cells through only one of that type, the one that conform to a GU-AG rule is predominantly associated with eukaryotic protein coding genes. The GU-AG rule, the first two nucleotides at the 5’end of the RNA sequence of all of these introns are in variably 5’-GU-3’ and the last two at the 3’ end of the intron are always 5’-AG-3’. Additional nucleotides associated with these 5’ and 3’ splice junction as well as an internal branch site [4][6]. In the present work, introns are removed from pre-mRNA based on GU-AG rule. The final sequence after skipping the intron part is m-RNA which contains exons only.

3. RESULTS AND DISCUSSION

The user inputs the DNA sequence. To convert DNA to pre-mRNA, Thymine is replaced by Uracil in the existing sequence. The extra spaces present in the input sequence are also trimmed. Association mining is utilized to retrieve the results. The next process is to convert pre-mRNA to mRNA. In this case, splicing is implemented as a part of transcription. GU-AG rule is followed to remove the introns. The output sequence is mRNA. Induction and Cluster Analysis technique of data mining is used to retrieve the results.
4. CONCLUSION
The model constructed is useful for determining the pre-mRNA and mRNA for the input DNA sequence. The provision is also there to input the pre-mRNA or mRNA sequences directly to retrieve the results at the relevant stages. All the possible amino acid combinations are shown at the output. The present work will assist the researchers to predict the relevant outputs at the different stages of gene expression. This can further assist to predict the type or class of the gene under consideration.

5. FUTURE SCOPE OF WORK
Following improvements regarding the developed model of bioinformatics can be made:

- For the final amino acid sequence, different protein structures (primary, secondary and tertiary) can be predicted.
- The model can be extended to have a search retrieval tool for performing search in 3D structure 3D structure databases like Protein Database (PDB).
- Various types of RNA sequences (rRNA, tRNA etc) can also be generated for a given DNA sequence.

6. REFERENCES