International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 09 Issue 10 ǁ 2021 ǁ PP. 75-81

Inferring Evolutionary Distance Using Kimura's Two Parameter Method

Jagdeep Kaur
Research Scholar
Desh Bhagat Foundation Group of Institution Ferozepur road Moga

Deepak Sharma
HOD &Assistant Professor
Desh Bhagat Foundation Group of Institutions Ferozepur road Moga

Abstract*:*
Bioinformatics is the application of computer science to the field of biology and medicine. With the help of computer tools, biological information is gathered and analysed. It is the science of managing, mining and interpreting information from biological sequences and structures. It deals with algorithms, databases and information systems, data mining, image processing and improving & discovering new models of computation. The substitution model is used to calculate the distances between the sequences of same family of an organism called evolutionary distance. If two sequences are similar, then they evolved from the same origin. In this researchkimura model isused to infer the evolutionary distance between two families.After evaluating the results it is observed that Kimura's model perform better as compare to other evolutionary distance measure methods.
Keywords: *Evolutionary method, Kimura's method, Phylogenetic tree, bioinformatics.*

Date of Submission: 20-09-2021 Date of acceptance: 05-10-2021

I. Introduction:

Today computer is used in almost every field including artificial Intelligence, natural language processing [32][33][34][35][36][37] and bioinformatics. Bioinformatics is the field of computer science used for managing, mining and interpreting information from biological sequences and structures. It deals with algorithms, databases and information systems, data mining, image processing and improving & discovering new models of computation. This scientific field deals with the computational management of all kinds of biological information. This information can be on genes and their products, whole organisms or even ecological systems. Mainly bioinformatics involves merger of different applications of mathematical, statistical, computational or molecular biological tools to gather different types of information and by analyzing them, researches can be carried out. Over the past few decades, major advancements in this field have led to an explosive growth in the biological information. The computerized databases are used to organize, store and index the data. Java, XML, Perl, C, C++, SQL and MATLAB are the programming languages popularly used in this field. The tools of bioinformatics include computer programs that help to reveal fundamental mechanisms. The biological problems related to the structure and function of macromolecules, disease processes, and evolution are contained in the tools. The applications of the tools is being categorised into sequence analysis, structure analysis, and function analysis. These three aspects of bioinformatics often interact to produce integrated and good results. Bioinformatics includes the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins. Extract DNA and Protein Sequences from Database. Whole of the database is being searched to compare the DNA's in a pair-wise fashion. DNA is transcribed to RNA which is further translated to proteins. This make possible to analyze the behaviour of the cell. After the alignments, a structure occurs in form of a tree.Two subfields consists of Bioinformatics are the development of computational tools and databases and the application of these tools and databases in generating biological knowledge to better understand living systems. These two subfields are complementary to each other. The application of the tools is being used in construction and crating of biological databases. The analyses of biological data often generate new problems that use to develop new and better computational tools.

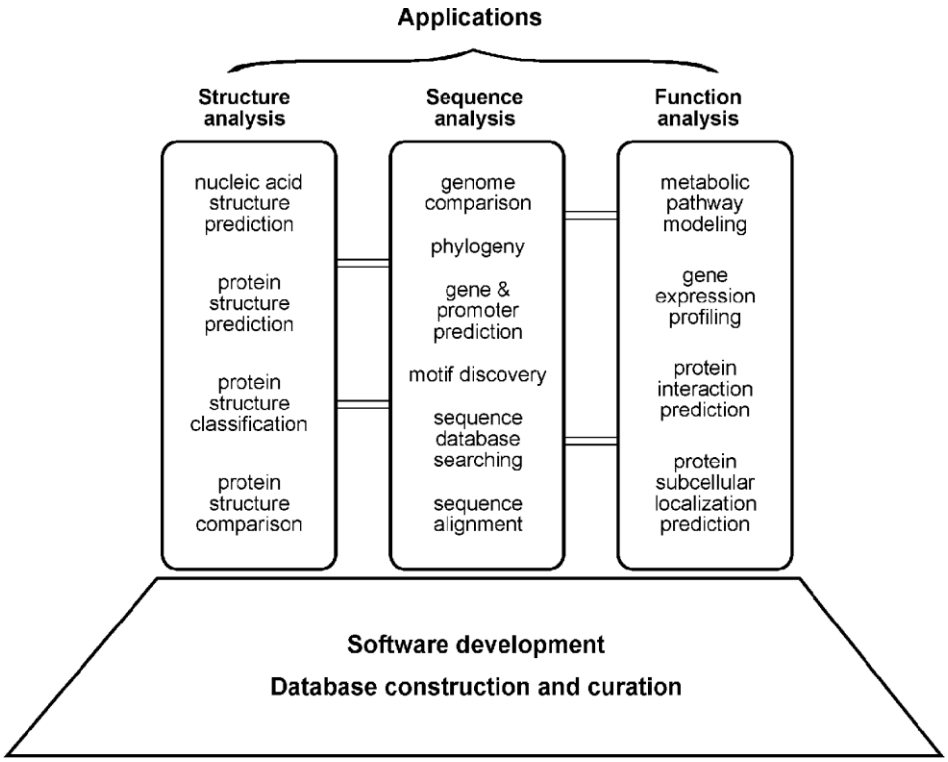

Figure 1.2: Overview of various subfields of bioinformatics [26]

The areas shown in Figure 1.2 consists of sequence analysis that include sequence alignment, sequence database searching, motif and pattern discovery, gene and promoter finding, reconstruction of evolutionary relationships, and genome assembly and comparison. Structural analyses include protein and nucleic acid structure analysis, comparison, classification, and prediction. The functional analyses include gene expression profiling, protein–protein interaction prediction, protein sub-cellular localization prediction, metabolic pathway reconstruction, and simulation.

Alignment of Sequences

In bio-informatics, sequence alignment is a way of arranging the primary sequences of DNA, RNA and proteins to identify regions of similarity. These regions of similarity may be a consequence of functional, structural or evolutionary relationships between the sequences. This is used to find the best-matching sequences. A sequence alignment is a scheme of writing one sequence on top of another where the residues in one position are deemed to have a common evolutionary origin. Aligned sequences of nucleotides or amino acid residues are represented as rows within a matrix. A letter or a stretch of letters may be paired up with dashes in the other sequence to signify such an insertion or deletion. Gaps are inserted between the residues so that residues with identical or similar characters are aligned in successive columns. If the same letter occurs in both sequences then the position is conserved in evolution. If the letters differ then take one residue or neither from two derives from an ancestral letter. Homologous sequences may have different lengths. Homologous sequences mean two or more sequences have a common ancestor.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        -------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Alignment Methods

Very short or very similar sequences can be aligned by hand. However, problem occur in alignment of lengthy, highly, variable or extremely numerous sequences that cannot be aligned by human effort and required computational approaches to align the sequences. Two categories are: global alignments and local alignments.Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set and of roughly equal size. Global alignment get the maximum match between the sequences as it assume that the two sequences are similar. This alignment attempts to match the two sequences from the end to the end even though if they are different in some parts. Calculating a global alignment is a form of global optimization that "forces" the alignment to span the entire length of all query sequences. A general alignment technique is called the Needleman-Wunsch algorithm and is based on dynamic programming. By contrast, Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. Local alignment searches for the part of the two sequences that match well. There is no attempt to "force" entire sequences into an alignment, just those parts that appear to have good similarity, according to some criterion are considered. The Smith-Waterman algorithm is a general local alignment method also based on dynamic programming. With sufficiently similar sequences, there is no difference between local and global alignments.

Existing literature

Zhou et al. (2019) discussed a method to found the nearest neighbors in biological databases using less distance computations that involved the similarity comparison between two objects. A substantial speedup technique for the well-studied k-nearest neighbor (k-nn) search is used, which is based on novel concepts of virtual pivots and partial pivots, such that a significant number of the expensive distance computations can be avoided. Some methods are included for k-nn searching and that are M-tree, OMNI, SA-tree, LAESA.

Zimek et al. (2019) studied the hierarchical and flat classification of proteins. The problem in classification of proteins was received significant attention. One feature of this problem is that expert-defined hierarchies of protein classes exist and can potentially be exploited to improve classification performance. They compared multiclass classification techniques that exploited the information in those class hierarchies and those that do not, using logistic regression, decision trees, bagged decision trees, and support vector machines as the underlying base learners.

Chen et al. (2018) discussed the classification trees that included nonparametric statistical learning methods having incorporated feature such as selection and interactions, possess intuitive interpretability, efficiently, and have high prediction accuracy when used in ensembles. However, it provided a brief introduction to the classification tree-based methods, a review of the recent developments, and a survey of the applications in bioinformatics and statistical genetics.

Kakiuchi et al. (2017) gave a characteristics of a new neighborhood that determined by three parameters. The neighborhood is generated from a special capacity that determined by three parameters. Neighborhood has a certain combination of contamination and gap from the model. Various new neighborhoods are obtained from changing the values of the three parameters. One of the parameters expressed the size of contamination and the others determined the size of gap from the model. It turns out that the introduced neighborhood is intuitively understandable and useful for developing minimax theory in robust inference.

Vijan et al. (2011) defined a biological sequence alignment for bioinformatics applications by using MATLAB. The biological sequence alignment is widely used operation in the field of bioinformatics and computational biology as it is used to determine the similarity between the biological sequences. The proposed method described the two basic alignment algorithms i.e. Smith Waterman for local alignment and Needleman Wunsch for global alignment. The algorithms have been developed and simulated using MATLAB for genome analysis and sequence alignment. The local and global alignment has been presented and the results are shown in the form of dot plots and local and global scores for the sequences. The goal is to develop a tool that can aided in the exploration, interpretation and visualization of data in the field of molecular biology.

Liu et al. (2012) developed an algorithm for bioinformatics by using pair-wise sequence alignment. Bioinformatics is the core of biotechnology. Sequence alignment is the most basic and important operation of bioinformatics. The information of functions, structure and evolution in biological sequence can be found by sequence comparison. The basic operation of sequence alignment is comparison. The proposed an algorithm that describes pairwise sequence alignment and explained the combination with instances.

Shehab et al. (2012) described an algorithm for sequence alignment by using fast dynamic method that based on bioinformatics. Their goal is to construct an algorithm for sequence alignment that based on the concepts of bioinformatics. Fast dynamic algorithm for sequence alignment (FDASA) is an implemented algorithm. This implemented algorithm based on making a matrix of M×N (M is the length of the first sequence, N is the length of the second sequence). The objectives of above mentioned algorithm are to filling the three main diagonal without filling the unused data and at the same time get an optimal solution; so that the execution time is decreased, the performance is high and the memory location decreased. Then compared the implemented

algortihm and between the dynamic algorithms Needleman-Wunsch algorithm, Smith-Waterman to test the execution time.

Mukunthan et al. (2011) proposed a technique that made an identification of unique repeated patterns, location of mutation in DNA finger printing by using artificial intelligence. Here the proposed method was described as Neural-Fuzzy pattern recognition (NFPR) system to reduces the complications in precisely analyzing and interpreting human deoxyribonucleic acid (DNA) samples. In above mentioned approach, the blend of bioinformatics and methods of neural network provided the advantages over conventional computational technique. It solved the problem that do not have an algorithmic solution or the available solutions that are too complex to be found, results in efficient DNA pattern analysis algorithm that identifies repeated patterns in the given human DNA sample by generation of unique identification number of an individual, location of occurrence of mutation in the mutated DNA sample with higher accuracy.

Naznin et al. (2010) proposed the decomposition with genetic algorithm (DGA) for multiple sequence alignment. With a certain research, they divide given sequences into two or more subsequences and then combine them together in order to find better multiple sequence alignments by applying a new GA based approach to the combined sequences. They introduced new ways of generating an initial population and of applying the genetic operators. To evaluate the proposed approach, they compared with well known methods such as T-Coffee, MUSCLE, MAFFT and ProbCons and their computational results have shown that the overall performance of the proposed decomposition with GA (DGA) method is better than the existing methods and the GA method (without decompositions).

Kimura's model :

The Kimura model adds one parameter to the Jukes-Cantor model in order to allow the rate of change between purines and pyr-midines (transversions) to be different from changes within purines or within pyr-midines (transitions). This is a rough-and-ready distance formula for approximating PAM distance by simply measuring the fraction of amino acids, that differs between two sequences and computing the distance. The rates of substitution in transition were assumed at a uniform rate α and transversions at a different, uniform rate of β. Furthermore, when the comparisons are being made, then the results produced are optimal as desired.

$R = \alpha + 2\beta$

The rate matrix of Kimura 2-parameter model and Proposed Kimura Model is different from Jukes Cantor model.

$$R = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \beta \\ \beta & \alpha & \beta & -2\beta-\alpha \end{bmatrix} \end{array}$$

Kimura's 2-parameter is more efficient as compared to Juke's cantor method because it calculates the distance based on transitions and transversions. In the present work, the Jukes Cantor distance model and Kimura's 2 parameter model and proposed Kimura distance models will be used for phylogenetic tree construction.

Phylogenetic Analysis

The development of a biological form from other pre-existing forms or current existing form through some modifications is known as evolution. The study of evolutionary history of some organisms using tree-like diagrams is known as phylogenetic tree construction or phylogenetic analysis. Each time a branch divides into a smaller branch, it shows the emergence of a new group of organisms. The most popular distance-based methods which are being used for the comparison are the Un-weighted pair group method with arithmetic mean (UPGMA) and Neighbor joining (NJ).

Database

The NCBI taxonomy website includes phylogenetic and taxonomic information from many sources available. These sources include the published literature, web databases, and taxonomy experts as well. While the NCBI taxonomy database is not a phylogenetic or taxonomic authority, it can be useful as a gateway to the NCBI biological sequence databases.The NCBI website can be searched through http://www.ncbi.nlm.nih.gov on internet. The MATLAB help browser is used to search the Web for information on NCBI through typing:Web ('http://www.ncbi.nlm.nih.gov/') in the command window. A separate browser window opens with the home page for the NCBI web site. The MATLAB used in statistical techniques for detecting peaks, selecting

features, and read genomic and proteomic data. With the help of bioinformatics toolbox all the proposed functions will be implemented.

II. Methodology

Very short or very similar sequences can be aligned by hand easily; however, most interesting problems require the alignment of lengthy, highly variable or extremely numerous sequences that cannot be aligned by human effort. The problem is to access a large amount of data and get the useful information. Due to the growing size and complexity of the biological data, it is necessary to get a correct tree. Only the correct alignment produces correct substitutions and this correct substitution produces correct phylogenetic tree. Incorrect alignment leads to incorrect substitution which shows systematic errors in the final tree and produces a wrong tree. Evolution is basically the study of changes in genes and proteins throughout different branches of the tree of life. Traditionally, phylogeny was assessed by comparing morphological features between the organisms from a variety of species taken. Molecular sequence data is being used for phylogenetic analysis through substitutional methods. When acquire the sequences for constructing the molecular phylogenetic trees, one can use either nucleotides or protein sequences. There are number of choices to represent the sequences. These formats are:

- Plain Text Format
- FASTA Format
- Genbank

The alignment of long and comparatively different sequences is very complex. This substitutional model is helpful for aligning the complex and extremely numerous sequences and find out the distances between the different species and make a phylogenetic tree which is correct in form using the distance based methods. The methodology for the present work involves the use of clustering analysis technique to compute the distances between the sequences. The present work will help the researchers to select the substitutional methods for finding the distances and different distance methods for constructing the phylogenetic tree.

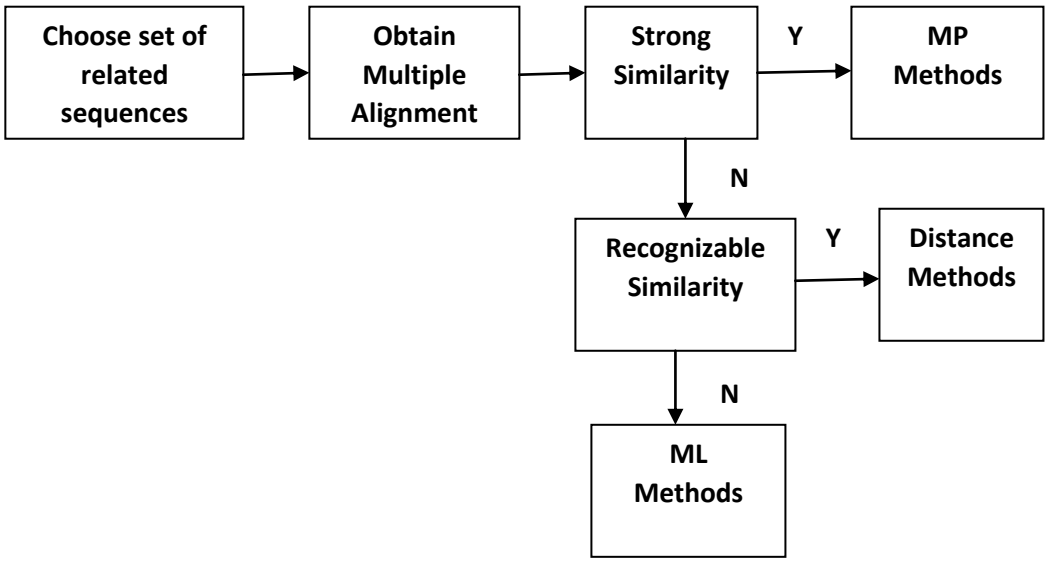

Figure: Phylogeny Flowchart

Implementation of Substitution Models

To measure the distance between the biological sequences numbers of substitutional models are used. These models are represented given below.

Juke's Cantor model

It is a nucleotide substitution model. For DNA sequences, the model assumes that all nucleotides are substituted with an equal rate. It is also called the *one-parameter model.* The rate of transitions (α) equals the rate of transversions (β).

The overall rate of substitution for any nucleotides was 3α. The initial probability (P) of site C and estimated substitution (K) was defined by the following equation:

$$P_{c(t)} = 1/4 + (3/4)e^{-4\alpha t} \quad (3.3)$$

$$K = -3/4 \ln[1-(4/3)(p)] \quad (3.4)$$

The rate of transitions (α) equals the rate of transversions (β).

Where p, is fraction of nucleotides that includes the rate of transition (P) and transversions (Q) i.e. (P+Q).

Kimura's 2-parameter model

It is a more sophisticated model. For DNA sequences, the model assumes that there are two different substitution rates, one for transition (α) and the other for transversion (β). It is also called the *two-parameter model*. According to this model, transitions occur more frequently than transversions. The rates of substitution in transition were assumed at a uniform rate α and transversions at a different, uniform rate of β.

The initial probability (P) of site C and estimated substitution (K) of Kimura was defined by the following equation:

$$P_{cc(t)} = 1/4 + (1/4)e^{-4\beta t} + (1/2)e^{-2(\alpha+\beta)t}$$

$$K = (1/2)\ln[1/(1 - 2P - Q)] + (1/4)\ln[1/(1 - 2Q)]$$

And the estimated substitution (K) of proposed Kimura was defined by

$$K = (1/2)\ln[1/(1 - 1.35P - Q)] + (1/4)\ln[1/(1 - 1.35Q)]$$

Where P, is fraction of nucleotides that are transitions and

Q, is fraction of nucleotides that are transversions

III. Conclusion

The substitutional model is used to calculate the distances between the sequences of same family of an organism. If two sequences are similar, then they evolved from the same origin. Analysis of phylogenetic tree can be done by nucleic acid and protein sequences. There are different sequence formats available from which FASTA format is used. The jukes cantor and kimura model are defined along with the distance matrix. The distance models are constructed for deciding that which method among the distance-based tree building methods yield an optimal tree at the end. These methods are computationally very fast and large numbers of sequences are easily handled. The overall advantage of these methods is the ability to make use of a large number of substitution models to correct distances in a distance metric.

UPGMA method had a critical assumption that the rate of nucleotide or amino acid substitution is constant for all the branches in the tree. The branch lengths can be used to estimate the dates of divergence, and the sequence-based tree mimics a species tree. The assumptions when violated with unequal substitution rates along different branches of the tree, produced an incorrect tree. The un-weighted distances also made an impact that a better tree must be produced for the species given. There are many applications in which UPGMA method is being applied for the result to come.

The neighbour-joining method gave an optimal tree for the input data and information given because it did not use the un-weighted distances. Also, this algorithm is especially useful when studying large numbers of taxa. With this algorithm, more extensive tree search can be carried upon as it has a better chance of finding the correct tree. A generalized NJ tree is also being developed in some cases in which optimal NJ tree is being chosen from a pool of NJ trees.

The evolutionary distances are displayed in jukes cantor model, kimura 2-parameter model and proposed kimura model. Each method has its own accuracy level. Above mentioned distance methods, will be used for phylogenetic tree construction.The overall computational results are being shown with respect to the evolutionary distances along both the distance-based methods i.eNeighbor-Joining Method and UPGMA tree building method. Thus, the user can analyze the different computation steps involved for the final assessment. This model is user friendly which provides different options to the user.

Future Scope

Following improvements and further enhancements regarding the developed analysis of phylogenetic trees can be made

- Database can be maintained for the known sequence of an organism and the search option can be provided to the user.
- The model can be extended to analyze large numbers of nodes that requires fast butaccurate algorithms to reconstruct and visualize evolutionaryhistories.
- The model further used to take a comparison with one tree building method with others.
- A new tool can be proposed for the phylogenetic tree to give one and only one optimal result without any error.
- Phylogenetic networks can be constructed from the new tool designed.
- Neighbor joining can be extended to enhance further to give an optimal result.

REFERENCES

[1]. Agrawal, A. and Xiaoqiu, H. (2008), "*Pairwise DNA Alignment with Sequence Specific Transition-Transversion Ratio Using Multiple Parameter Sets*", ICIT '08, pp.89-93.

[2]. Bogdanowicz, D. and Giaro, K. (2012), "*Matching Split Distance for Unrooted Binary Phylogenetic Trees*", IEEE/ACM Transactions on computational biology and bioinformatics, vol. 9, 1, pp. 150-160.

[3]. Chen, X., Wang, M. and Zhang, H. (2011), "*The use for classification trees for bioinformatics*", John Wiley & Sons, Inc. WIREs Data Mining KnowlDiscov, pp 55–63.
[4]. Guo, P., Chen, G. and Wang, Y. (2011), "*Constructing phylogenetic tree based on three-parameter model*", Key Engineering Materials, vol. 474-476, pp. 2193-2197.
[5]. Gronau, I., Moran, S. and Yavneh, I. (2010), "*Adaptive Distance Measures for Resolving K2P Quartets: Metric Separation versus Stochastic Noise*", Journal of Computational Biology, vol. 17, 11, pp. 1509-1518.
[6]. Huson, D.H. (2009), "*Drawing Rooted Phylogenetic Networks*", IEEE Transactions on computational biology and bioinformatics, vol. 6, 1, pp. 103-109.
[7]. Huson, D.H., Moulton, V. and Steel, M. (2009), "*Special Section: Phylogenetics*", IEEE/ACM transactions on computational biology and bioinformatics, vol. 6, 1, pp. 4-6.
[8]. Kakiuchi, I. and Kimura, M. (2011), "*Characterization of a new neighborhood determined by three parameters*", Technical Report of the NAS, pp. 1-11.
[9]. Krane, D. and Raymer, M. (2006), "*Fundamental concepts of bioinformatics*", Pearson Education Publishers.
[10]. Lesk, A.M. (2002), "*Introduction to Bioinformatics*", Oxford University Press, pp 180-200.
[11]. Lin, X., Meng, Z., He, X., Liu, Q., Liu, Y., Li, J. and Zhou, Y. (2010), "*A solution to integrate the phylogenetic tree's generation based on web*", ICBBE2010, pp. 1-3.
[12]. Liu, C. and Wang, F. (2012), "*Pair-wise sequence alignment algorithm in bioinformatics*", EEESym.2012, pp. 36-38.
[13]. Miyazawa, S. (2011), "*Advantages of a Mechanistic Codon Substitution Model for Evolutionary Analysis of Protein-Coding Sequences*", PLoS ONE, vol. 6, 12, pp. 54-69.
[14]. Mount, D.W. (2004), "*Bioinformatics: Sequence and Genome Analysis*", 2nd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.. ISBN 0-87969-608-7.
[15]. Mukunthan, B., Nagaveni, N. and Pushpalatha, A. (2011), "*Identification of unique repeated patterns, location of mutation in DNA finger printing using AI technique*", Journal of Bioinformatics and Sequence Analysis, vol. 3, 6, pp. 100-115.
[16]. Nakhleh, L. (2010), "*A Metric on the Space of Reduced Phylogenetic Networks"*, IEEE/ACM Transactions on computational biology and bioinformatics, vol. 7, 2, pp 1-5.
[17]. Naveed, T., Siddiqui, S.I. and Ahmed, S. (2009), "*Parellel Needleman-Wunsch Algorithm for Grid*", CLOUDS lab., pp. 1-6.
[18]. Naznin, F., Sarker, R. and Essam, D. (2010), "*DGA: Decomposition with genetic algorithm for multiple sequence alignment*", CIBCB 2010, pp. 1-8.
[19]. Pevsner, J. (2009), "*Bioinformatics and Functional Genomics*", A john wiley& sons, Inc. Publication, pp 215-221.
[20]. Rastogi, S.C., Mendiratta, N., Rastogi, P. (2007), "*Allignment of Multiple Sequences and Phylogenetic Analysi-Bioinformatics Methods and Applications*", 3rd edition, PHI publication, pp. 5-120.
[21]. Shehab, S.A., Keshk, A. and Mahgoub, H. (2012), "*Fast Dynamic Algorithm for Sequence Alignment based on Bioinformatics*", IJCA, vol. 37, 7, pp. 54-61.
[22]. Sohpal, V.K., Dey, A. and Singh, A. (2010), "*Sequence alignment and phylogenetic analysis of Human Herpes Simplex Virus (HHV) using bioinformatics tool*", Inderscience Enterprises Ltd., vol 3, pp. 68-88.
[23]. Torres, M., Dias, G., Gonçalves, G. and Vieira, C. (2011), "*Tool that Integrates Distance Based Programs for Reconstructing Phylogenetic Trees*", IEEE latinamerica transactions, vol. 9, 5, pp. 895-901.
[24]. Vijan, S. and Mehra, R. (2011), "*Biological Sequence Alignment for Bioinformatics Applications Using MATLAB*", IJCSET, vol. 2, 5, pp. 310-315.
[25]. Wang, L.S., Leebens-Mack, J., Wall, P.K., Beckmann, K., DePamphilis, C.W. and Warnow T. (2011), "*The Impact of Multiple Protein Sequence Alignment on Phylogenetic Estimation*", IEEE/ACM Transactions on computational biology and bioinformatics, vol. 8, 4, pp. 1108-1119.
[26]. Xiong J., (2006), "*Essential Bioinformatics*", United States Of America, Cambridge University Press, New York.
[27]. Zhou, J., Sander, J., Cai, Z., Wang, L. and Lin, G. (2010), "*Finding the Nearest Neighbors in Biological Databases Using Less Distance Computations*", IEEE/ACM Transactions on computational biology and bioinformatics, vol. 7, 4, pp. 669-680.
[28]. Zimek, A., Buckwald, F., Frank, E. and Kramer, S. (2010), "*A Study of Hierarchical and Flat Classification of Proteins*", IEEE/ACM Transactions on computational biology and bioinformatics, vol. 7, 3, pp. 563-571.
[29]. http://www.abpischools.org.uk/res/coresourceimport/resources04/cancer/images/dna1.gif
[30]. http://www.mysciencebox.org/files/images/RNA-codon.png
[31]. http://www.rothamsted.ac.uk/notebook/images/pept.gif
[32]. Mittal, Misha, Dinesh Kumar, and Sanjeev Kumar Sharma. "Grammar checker for asian languages: A survey." International Journal of Computer Applications & Information Technology 9.1 (2016): 163.
[33]. Sharma, Sanjeev Kumar, and G. S. Lehal. "Improving existing punjabi grammar checker." 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT). IEEE, 2016.
[34]. Sharma, Sanjeev Kumar, and Gurpreet Singh Lehal. "Using Hidden Markov Model to improve the accuracy of Punjabi POS tagger." 2011 IEEE International Conference on Computer Science and Automation Engineering. Vol. 2. IEEE, 2011.
[35]. Sharma, Sanjeev Kumar. "Sentence Reduction for Syntactic Analysis of Compound Sentences in Punjabi Language." EAI Endorsed Transactions on Scalable Information Systems 6.20 (2019): e4.
[36]. Sharma, Sanjeev Kumar. "Effect of Statistical POS Tagger on Syntactic Analysis of Punjabi Sentences." Indian Journal of Science and Technology 9.32 (2016).
[37]. Jindal, L., H. Singh, and S. K. Sharma. "A Framework for Grammatical Error Detection and Correction System for Punjabi Language Using Stochastic Approach." (2021).