Journal of Harbin Engineering University Vol 44 No. 5
ISSN: 1006-7043 May 2023

Protein Structure Prediction Using Needle Man Wunsch Algorithm

Rohit Mishra
Department of Computer Science and Engineering
United Institute Of Technology
Prayagraj,India
rohitmishra.academic@gmail.com
Prashant Pandey
Department of Computer Science and Engineering
United Institute Of Technology
Prayagraj,India
prashant02002@gmail.com
Amit Tiwari
Department of Computer Science and Engineering
United Institute Of Technology
Prayagraj,India
kumartiwariamit@gmail.com
Ravi Singh
Department of Computer Science and Engineering
United Institute Of Technology
Prayagraj,India
ravisingh98138@gmail.com
Sameer Devwanshi
Department of Computer Science and Engineering
United Institute Of Technology
Prayagraj,India
devwanshishubham@gmail.com
Sakshi Singh
Department of Computer Science and Engineering
United Institute Of Technology
Prayagraj,India
vsakshisingh2121@gmail.com

Abstract— Protein Sequence Analysis is a very important part of Bioinformatics which deals with the study of amino acids which is the building block of proteins. In Protein sequence analysis we subject a protein or the peptide sequence to various kinds of analytical methods which compares and contrasts their features in order to establish various relationships among important factors. In order to implement we are using the Needleman Wunsch algorithm for performing the pairwise alignment of the two protein sequence. As a result, this algorithm when paired with local alignment generates a score which is also known as the alignment score of the two protein sequences. Upon further study multiple inferences can be obtained and used for various medical and research purposes.

Keywords— Protein Sequence, Needle man Wunsch Algorithm, local alignment.

I. INTRODUCTION

A protein sequence is a unique series of amino acids that are linked together by peptide bonds to form a protein molecule.[1] Each protein has its own specific sequence of amino acids, which is determined by the genetic code stored in the DNA of

an organism.[2] The sequence of amino acids in a protein molecule determines its three-dimensional structure and its function. The study of protein sequences is important for understanding the structure and function of proteins, and for designing drugs and therapies that target specific proteins.[3] Protein sequence analysis is a critical field in bioinformatics that involves the study of the amino acid sequences that make up proteins.[4] Proteins are complex molecules that play essential roles in various biological processes, such as catalysis, signaling, and transport, among others.[5] Understanding the structure and function of proteins is crucial to elucidating their role in various biological processes and developing new drugs and therapies.[6] Protein sequence analysis involves the use of computational tools to predict and analyze protein structures, functions, and interactions. The analysis includes various methods such as sequence alignment, motif identification, and prediction of secondary and tertiary structures.[7] additionally, protein sequence analysis can also be used to identify potential drug targets, study protein evolution, and investigate protein-protein interactions [8].

The lengths of the protein sequences vary and so a method needs to advised which can predict and compare the protein structures in a rapid and efficient manner, moreover traditional methods have become outdated as they are not able to deal with the tremendous temporal complexity. So local alignment of the two protein sequences can play an instrumental role in comparing the structures of the proteins although this requires an extensive knowledge of the proteins and its constituents failing which proper conclusions cannot be drawn. Adoption of this methodology can prove to be a boon in the field of Bio-informatics as it can help in various fields such as drug identification, disease prediction and protein redesigning.

II. LITERATURE SURVEY

The main aim of protein sequence analysis is to explore the pattern of body constituents such as DNA, protein sequences and RNA. This process involves the sequence alignment of the given proteins thus extracting the patterns in sequences which will later aid in classification. This process of sequence analysis can be performed in various ways such as sequence comparison, mutation detection etc. It also helps in performing evolutionary analysis and prediction of gene/protein structures. [9]

The current analysis and some theses have proved that every protein sequence-structure pair shows distinctive properties and they can be classified based on that. moreover, a protein sequence can be used to find or predict the structure of a protein or vice versa. This will help in showing how a particular change in the structure of the protein will cause change in its sequence. [10]

Using only the protein sequence in order to find out the structure is a tough and hectic task let alone the prediction of a three-dimensional structure. While significant progress has been made in recent years, the accuracy of current methods is still limited, particularly for large and complex proteins.[11]

The algorithms and methods that are currently being used consume a large space and time complexity which is not much efficient. In addition to this validating and benchmarking the accuracy of new methods and tools for protein sequence analysis remains a challenge. While there are established methods for benchmarking protein structure prediction methods, developing similar standards for other types of analysis, such as motif discovery or functional annotation, is still an active area of research.[12]

The protein structure and its sequence show a striking resemblance i.e., through a protein structure, the sequence can be obtained and in addition to this sequence can be used to predict the structure of the Protein.[13]

Predicting the structure of a protein can be made much more efficient if a large dataset is present for the prediction of protein structures from their sequences, presently less sequences are compared and so the result derived also does not show all the possibilities, and so to cover every possibility it is important to cover a large dataset.[14].Protein redesigning has gained much importance as by applying it present proteins can be made more efficient and their properties can also be improved which will help in tackling diseases with ease as compared to the normal protein [15].Earlier the world of protein and its constituents was treated as model which was thought to be hard to deal and navigate with as the domain of protein , its related peptides and amino acids is pretty vast and it is also

not easy to classify them due to the large extent of similarities possessed by them.

III. PROPOSED WORK

In our project we want to implement protein sequence analysis by performing local alignment on the two protein sequence.

Local Alignment - Local alignment is a method for aligning protein sequences that focuses on identifying the regions of highest similarity or homology, rather than the entire sequence. This approach is useful for identifying conserved regions or domains within proteins that may have different functions or evolutionary histories than other regions of the protein.

A. Block diagram

Fig 1 is the diagrammatic representation of processes involved in our system, starting from giving the input to obtaining the result as the output. This system involves 6 major steps which are served on the backbone of Needleman Wunsch algorithm.

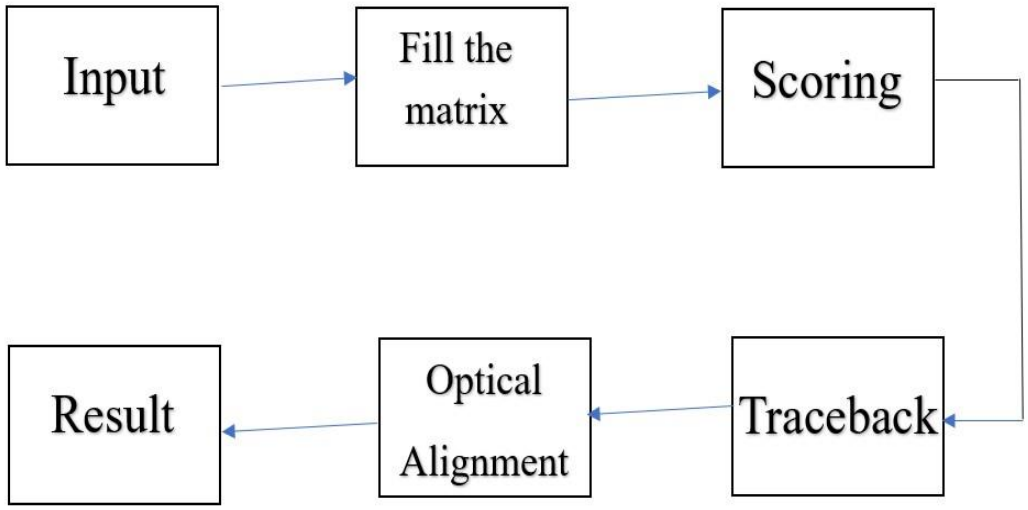

Fig 1: Block Diagram

B. Algorithm

The NM Wunsch algorithm or the Needleman Wunsch algorithm is a dynamic programming algorithm which is basically implemented in order to align two protein sequences.

The steps involved in the Needleman-Wunsch algorithm are:

Step 1. Initialization: The process starts by initializing or rather creating a two-dimensional score matrix (score matrix) whose dimensions are (m+1) x (n+1), where m and n are the lengths of the two sequences to be aligned. The matrix is filled with scores and initial values, which depend on the gap penalty and substitution matrix used.

Step 2. Fill the matrix: The matrix is then filled in row-by-row, column-by-column order. At each cell (i,j), the algorithm calculates three scores: the score obtained from alignment of the ith character of the sequence A with the jth character of the sequence B (diagonal), the score obtained from inserting a gap in the first sequence (left), and the score obtained from inserting a gap in the second sequence (up). The score at cell (i, j) is the maximum of these three scores.

Step 3. Traceback: Once the matrix is completely filled, the algorithm traces back from the bottom-right corner of the matrix to the top-left corner, following the path of highest scores. This path corresponds to the optimal alignment of the two sequences.

Step 4. Output: The algorithm outputs the aligned sequences, along with the alignment score.

It is worth noting that in some implementations of the Needleman-Wunsch algorithm, step 1 and step 2 are combined into a single step, where the matrix is initialized and filled at the same time. Identify

applicable funding agency here. If none, delete this text box.

C. Methodology

In order to achieve our objective, we are comparing two protein sequences in order to know the similarities and differences between them. We are using the local alignment method in order to reduce the time and space complexity of the problem, comparing a single protein sequence with the whole sequences in the protein database resulted in excessive time consumption which is not an ideal scenario. The algorithm that we are using in order to achieve our objective is the needle man Wunsch algorithm. Efforts have been made earlier to perform protein sequence alignment with the help of this algorithm but that was performed with global alignment but we have used local alignment and pairwise alignment method

IV. COGNITIVE ANALYSIS

The cognitive analysis of protein sequence-structure relationships involves examining how the primary sequence of a protein relates to its three-dimensional structure i.e., the way in which the protein folds and interacts with itself and other molecules. This analysis can be done using various computational methods, such as machine learning algorithms and bioinformatics tools.[11]

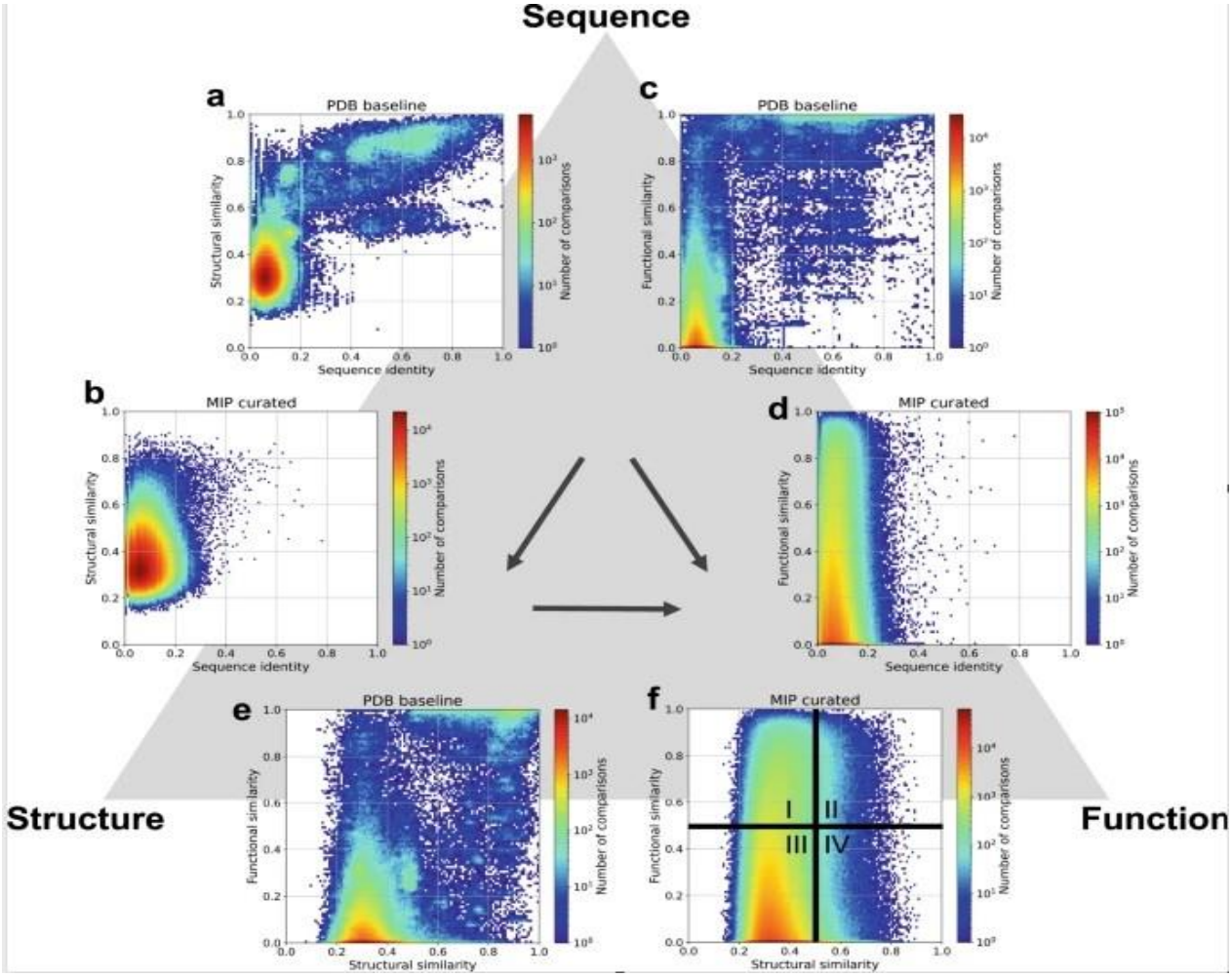

Fig 2: Structure Sequence Relationship

V. RESULT ANALYSIS

The results obtained from the protein sequence analysis are of vital importance as it can help in the prediction of three-dimensional structure of the proteins and related peptides.

The comparison of the two structures of protein can help in determining the extent of damage that a foreign body or virus can cause on the protein or the human body at large. The alignment score is used to resemble the quality or goodness of the alignment of the protein sequences. The algorithm used is developed by keeping in mind that regardless of the length or the complexity of the protein sequence, best alignment is obtained. The efficiency of the algorithm increases when we use protein sequences who are of similar length i.e., which have not undergone massive mutations over their lifetime. This also shows the importance for urgency required in the field of Protein sequencing and analysis as if considerable amount of time has passed between two mutations than it would be difficult to establish the relations between them.

Fig 3: Result(A)

```
The substitution matrix for the alphabet ACGT is:
   0  1  2  3
0  3 -1 -1 -1
1 -1  3 -1 -1
2 -1 -1  3 -1
3 -1 -1 -1  3
Scoring Matrix :

      0       0       0       0       0       0       0       0       0       0       0       0       0
      0       3       3       0       0       0       0       3       0       3       0       0       0
      0       3       6       3       0       0       0       3       2       3       2       0       0
      0       0       3       9       6       3       0       0       6       3       2       1       3
      0       0       0       6      12       9       6       3       3       5       2       5       2
      0       0       0       3       9      11      12       9       6       3       8       5       4
Identities = 9/11 (81.8%), Gaps = 2/11 (18.2%), Mismatches = 0/11 (0.0%)

Query  1     TTA-GCTGATC  11
             ||| ||| |||
Sbjct  1     TTAGGCT-ATC  11

Score of the alignment =
21
```

Fig 4: Result(B)

VI. FUTURE SCOPE

Despite of tremendous leaps made recently in the field of bioinformatics and protein sequencing and analysis we are still far off from what could be done in this field and this was quite evident during the covid-19 phase as majority of the time was invested in finding the how the virus is interacting with the body and how it mutates the constituents of the body. It can help in early treatment and prevention of diseases and in addition to this it can help in producing more efficient medicines and drugs in order to treat diseases. The analysis of a protein sequence or a peptide will help in learning biological properties which will serve as a building block for predictive models for solving future biological problems. As machine learning and artificial intelligence algorithms continue to improve, they may be able to analyze large-scale protein sequence data more efficiently and accurately. This could lead to new insights into protein structure and function, as well as the development of new drugs and therapeutics. Post-translational modifications (PTMs) are essential for regulating protein function and activity, but they can be difficult to identify and analyze. As technology improves, there may be new ways to identify and analyze PTMs, providing insights into their role in protein function and disease. The future of protein sequence analysis looks bright, with

many exciting developments on the horizon. As new techniques and technologies are developed, we can expect to gain a deeper understanding of the structure and function of proteins, as well as new insights into the complex interactions that occur within and between cells.

VII. CONCLUSION

In conclusion, this study provides valuable insights into the changing structure of proteins which can later be used to check for rise in any disease or prevent a critical situation such as the COVID-19 pandemic. Through a comprehensive review of the literature and the analysis of various data we found that the study of protein structure can be very beneficial in preventing diseases, predicting their extent, minimizing its effect and drug identification for a particular disease. These findings can have important implications for the bioinformatics and Healthcare sector.

Overall, this research highlights the urgent need for more research on the intersection of protein sequences and the diseases affecting human body. By continuing to investigate these issues we can better understand the complex relationship between these important factors and work towards solutions that promote a world with a cure for every disease.

REFERENCES

[1] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995 Apr 7;247(4):536-40. doi: 10.1006/jmbi.1995.0159. PMID: 7723011.

[2] Náray-Szabó, Gábor & Perczel, András. (2014). Protein structure and dynamics. INTERNATIONAL JOURNAL OF TERRASPACE SCIENCE AND ENGINEERING. 6. 7-16.

[3] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42. doi: 10.1093/nar/28.1.235. PMID: 10592235; PMCID: PMC102472.

[4] Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. (2014) ECOD: An Evolutionary Classification of Protein Domains. PLoS Comput Biol 10(12): e1003926. https://doi.org/10.1371/journal.pcbi.1003926

[5] Wan X, Tan X (2021) A protein structural study based on the centrality analysis of protein sequence feature networks. PLoS ONE 16(3): e0248861. https://doi.org/10.1371/journal.pone.0248861.

[6] Fan Pu, Scott A. Ugrin, Andrew J. Radosevich, David Chang-Yen, James W. Sawicki, Nari N. Talaty, Nathaniel L. Elsen, and Jon D. WilliamsAnalytical Chemistry 2022 94 (39), 13566-13574DOI: 10.1021/acs.analchem.2c03211 7. Gan HH, Perlow RA, Roy S, Ko J, Wu M, Huang J, Yan S, Nicoletta A, Vafai J, Sun D, Wang L, Noah JE, Pasquali S, Schlick T. Analysis of protein sequence/structure similarity relationships. Biophys J. 2002 Nov;83(5):2781-91. doi: 10.1016/s0006-3495(02)75287-9. PMID: 12414710; PMCID: PMC1302362.

[7] Wan X, Tan X. A protein structural study based on the centrality analysis of protein sequence feature networks. PLoS One. 2021 Mar 29;16(3):e0248861. doi: 10.1371/journal.pone.0248861. PMID: 33780482; PMCID: PMC8006989.

[8] Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC, Bradley DG. Genetic evidence for Near-Eastern origins of European cattle. Nature, 2001, vol. 410, p. 1091

[9] Gan HH, Perlow RA, Roy S, Ko J, Wu M, Huang J, Yan S, Nicoletta A, Vafai J, Sun D, Wang L, Noah JE, Pasquali S, Schlick T. Analysis of protein sequence/structure similarity relationships. Biophys J. 2002 Nov;83(5):2781-91. doi: 10.1016/s0006-3495(02)75287-9. PMID: 12414710; PMCID: PMC1302362.

[10] Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008 Jun;18(3):342-8. doi: 10.1016/j.sbi.2008.02.004. Epub 2008 Apr 22. PMID: 18436442; PMCID: PMC2680823.

[11] Aniba MR, Poch O, Thompson JD. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Res. 2010 Nov;38(21):7353-63. doi: 10.1093/nar/gkq625. Epub 2010 Jul 17. PMID: 20639539; PMCID: PMC2995051.

[12] Hin Hark Gan 1, Rebecca A Perlow, Sharmili Roy, Joy Ko, Min Wu, Jing Huang, Shixiang Yan, Angelo Nicoletta, Jonathan Vafai, Ding Sun, Lihua Wang, Joyce E Noah, Samuela Pasquali, Tamar

Schlick: Analysis of protein sequence/structure similarity relationships , vol 1,p.01

[13] John-Marc Chandonia, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K Fox, Steven E Brenner, SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D553–D559, https://doi.org/10.1093/nar/gkab1054

[14] Yuting Xu, Deeptak Verma, Robert P. Sheridan, Andy Liaw, Junshui Ma, Nicholas M. Marshall, John McIntosh, Edward C. Sherer, Vladimir Svetnik, and Jennifer M. JohnstonJournal of Chemical Information and Modeling 2020 60 (6), 2773-2790DOI: 10.1021/acs.jcim.0c00073.

[15] Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017, 18, 186, DOI: 10.1186/s13059-017-1319-7

[16] Bernard, G.; Chan, C. X.; Chan, Y. B.; Chua, X. Y.; Cong, Y.; Hogan, J. M.; Maetschke, M. A.; Ragan, M. A. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. Briefings Bioinf. 2019, 20, 426– 435, DOI: 10.1093/bib/bbx067

[17] Katoh, K.; Misawa, K.; Kuma, K. I.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002, 30, 3059– 3066, DOI: 10.1093/nar/gkf436

[18] Domazet-Lošo, M.; Haubold, B. Alignment-free detection of local similarity among viral and bacterial genomes. Bioinformatics 2011, 27, 1466–1472, DOI: 10.1093/bioinformatics/btr176

[19] Gupta, M. K.; Niyogi, R.; Misra, M. A 2D Graphical Representation of Protein Sequence and Their Similarity Analysis with Probabilistic Method. MATCH Commun. Math. Comput. Chem. 2014, 72, 519– 532

[20] El-Lakkani, A.; Lashin, M. An efficient method for measuring the similarity of protein sequences. SAR QSAR Environ. Res. 2016, 27, 363–370, DOI: 10.1080/1062936x.2016.1174735

[21] Ghosh, S.; Pal, J.; Maji, B.; Bhattacharya, D. K. A sequential development towards a unified approach to protein sequence comparison based on classified groups of amino acids. Int. J. Eng. Technol. 2018, 7, 678, DOI: 10.14419/ijet.v7i2.9546

[22] Kong, F.; Yao, Y. H.; Dai, Q.; He, P. A. A sequence-segmented method applied to the similarity analysis of long protein sequence. MATCH Commun. Math. Comput. Chem. 2013, 70, 431– 450, DOI: 10.1109/ISB.2012.6314157

[23] Yu, L.; Zhang, Y.; Gutman, I.; Shi, Y.; Dehmer, M. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. Sci. Rep. 2017, 7, 46237, DOI: 10.1038/srep46237

[24] Randić, M. 2-D graphical representation of proteins based on physicochemical properties of amino acids. Chem. Phys. Lett. 2007, 440, 291– 295, DOI: 10.1016/j.cplett.2007.04.037

[25] Yao, Y.-H.; Dai, Q.; Li, L.; Nan, X. Y.; He, P. A.; Zhang, Y. Z. Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation. J. Comput. Chem. 2010, 31, 1045–1052, DOI: 10.1002/jcc.21391

[26] Yu, C.; Cheng, S. Y.; He, R. L.; Yau, S. S. T. Protein map: an alignment-free sequence comparison method based on various properties of amino acids. Gene 2011, 486, 110, DOI: 10.1016/j.gene.2011.07.002