iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39417

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Reducing Dimensionality in Remote Homology Detection

S. Dinesh[1], R. Bindhu[2], T. Anitha[3], R. Gunasundhari[4]
[1, 2, 4] *Assistant Professor* [3]*Ph.D Research Scholar*
[1,2]*Department of Computer Science*
[4] *Department of Information Technology*
[3]*Department of computer science and Applications,*
[1, 2, 4] *G.T.N Arts College*
[3]*The Gandhigram Rural Institute (Deemed to be University)*

Abstract: Homology detection plays a major role in bioinformatics. Different type of methods is used for Homology detection. Here we extract the information from protein sequences and then uses the various algorithm to predict the similarity between protein families. SVM most commonly used the algorithm in homology detection. Classification techniques are not suitable for homology detection because theyare not suitable for high dimensional datasets. Soreducing the higher dimensionality is very important than easily can predict the similarity of protein families.
Keywords: Homology detection, Protein, Sequence, Reducing dimensionality, BLAST, SCOP.

I. INTRODUCTION

Homology detection is very important in the biological field. It is one of the research problems. so many classification methods are used by classification of the protein sequence. Classify the protein family and superfamily is called as homologies. One protein family to another protein family sequence may not high similarity. But sometimes they are may structurally and functionally relate. Dimensionality reducing is to reduce the random variable counting then we have to analyze the algorithm performance. So many powerful methods and algorithms are used to solving this problem. the most commonly used method SVM (support vector machine) is a supervised learning based algorithm[3-6,11]. The smith-waterman algorithm is also usedfor homology detection problem [9,11,14]. Traditional biological studies have been only focused on model systems, but remote homolog studies provide high-level resources to investigate other species family. Remote homology detection can be used to compare a query sequence to similar sequences in large data sets. the biggestproblem in remote homology identification is all the methods are based on the sequence of the protein. Most of the available methods use the information about closely related sequences of protein, protein structure prediction, and the structure-structure comparison is used for identification in remote homology. Whatever method we used in this problem all the method have some error prediction. Here uses the SCOPhierarchy, it has a different level of hierarchy (i.e, class, family, superfamily and fold) [1-14]. Hidden Markov Model (HMMs) is the most powerful approaches in remote homology detection. HMMs is a structure based one. Whatever method and algorithms used in this problem it gives an output fully depends on protein sequence.

II. METHOD

A. Sequence-Sequence Comparison

William R. Pearson et al [15].uses this method. The remote homology detection is using most common method is a sequence-sequence comparison, protein sequence containing all the information about structure functions. Sequence comparison using pairwise alignments that provide correct relationship information between proteins. Some other methods alsoused for remote homology detection, but that type of methods used by query sequences against a sequence database. Sequence similarity searching is the first and important step in analysis of newly determined sequences .the most commonly used similarity searching programs, like BLAST, PSI-BLAST, FASTA, SSEARCH , HMMMER3.that programs produce accurate statistical estimation of protein sequence similarities.

1) *Structural Alignment:* Structural alignment is used to represent the protein sequences, it also represents the RNA sequences also. It gives a secondary and tertiary information about the structure of proteins .that is used foraligning sequences purposes. Marianne M. Lee, Ralf Bundschuh, andMichael K. Chan1 et al [16] uses the LESTAT algorithm. LESTAT is used in structural Alignment method. It is a structure-based sequence alignment algorithm., this methodcomprised of the following steps:

a) We construct our initial model by taking samples of amino acid blocks and block separation distances from three structural homolog with low-sequence identity.
b) Next we generate a block containing position which specifies the matrix score.
c) Align the query sequentially.
d) Sequenced Result with Optimal Alignment are used in this algorithm.

Sequences with reasonable statistical significance are used to generate a new BPSSM, repeating steps (2)–(4), until the refinement reaches convergence. Structural alignment uses the two or more sequences and produces the local alignments is based on structural information. These structures are more evolutionarily conserved than sequence. But the sequence comparison not only decides the similarity.

B. Sequence-Structure Comparison BLAST

Bin Liu, Lei Lin, Xiaolong Wang and Xuan Wang et al [11][7-9,13]. used to Protein sequence structure, which is generated by aligning the closely related protein families .each amino acid is located at each position of multiple sequence alignments. it better reflect on protein families than a single sequence. thus the sequence-structure alignment gives a sensitive remote homology detection than pair wise sequence alignment.

1) *Hidden Markov Model:* Noah M. Daniels, Andrew Gallant, Norman Ramsey, Lenore J. Cowen et al [8]. Uses this method. Hidden Markov Model (HMMs) is the most powerful approaches in remote homology detection. HMMs is a structure based on HMM. But, HMM is slower than PSI-BLAST. HMM, performance is higher than PSI-BLAST [14] approaches like HMMER, SAM, and META-MEME .these are the most common frequently used models. Other models are,JA(Jumping Alignment), DIALIGN, Family Pairwise Search (FPS)
2) *Structure-Structure Comparison:* It is also called as profile-profile comparison. correct and long alignment has been obtained from structure-structure comparison. it is a more sensitive approach for remote homology detection compare to sequence-profile comparison approaches. such as PSI-BLAST and HMMs.it may give the same order output of the BLAST approach [1]or give an improved output of the BLAST approach. The profile-profile comparison tools contain: COMPASS, PROF_SIM, COACH, HH search, FORTE, HMAP, and SP3.
3) *Phylogenetic Analysis:* Phylogenetic analysis and sequence alignment are closely related with each other. Phylogenetic analysis user for construction and interpretation of phylogenetic trees. it used to classify the evolutionary relationship between homologous genes and it represents the genomes of divergent species.

III. TOOLS

The tools are mainly used for predict the structural similarities of protein. Bin Liu, Lei Lin, Xiaolong Wang and Xuan Wang et al [11] [7-9,11,13].uses the BLAST tool used to align the query sequences from the selected target database. Jian-huaYeh and Chun-Hsing Chen et al [10]. Use the SVM and PSI-BLAST tools used for classification. Noah M. Daniels, Andrew Gallant, Norman Ramsey and Lenore J. Cowen et al [8] proposed the HMMER tools for alignment. Yuchen Yanga, Erwin Tantosob, and Kuo-Bin Lic [5] proposed the RQA method for users to find the relationship between structures. Jian-huaYeh, Chun-hsingChen[10] proposed the MAST tool used for alignment and search purposes.

IV. DATABASES

A. ASTRAL

The ASTRAL provides the databases and tools for analyzing the protein structures and their sequences [3]. It derived from, and augments the SCOP: Structural Classification of Proteins database. The dataset contains 54 families and 4352 distinct sequences. It provides 1.53 families from a given super family then Remote homology is simulation will be used.

B. SCOP- Structural Classification of Proteins

The Structural Classification of Proteins (SCOP) database is manually classification of protein structural domains based on their similar sequence of amino acid [1-14]. Here to classify the relationship between proteins. Sequenced similar proteins are placed on different super families. The common ancestor is used for classification of family protein sequences. The SCOPs database is freely accessible on the internet.

V. PERFORMANCE MEASURE

Different type of algorithms and methods are used for remote homology detection but the accuracy of the result is also different from one to one. if we use the smith-waterman algorithm it gives a less than 95% percent identity on SCOP 1.59 [12]database latent semantic methods give a ROC score 0.9435 such as SVM-Ngram, SVM-Motif-LSA, SVM-Pattern-LSA [10]. remote C3D method using The profile-based methods such as SVM-DT, SVM-PDT-Profile and composition-based methods such as SMV-LA, SVM-RQA, SVM-PDT used for remote homology detection It will give a ROC score is 0.948 with SCOP 1.53 and 0.936 with SCOP 1.55 [1].

$$TPR(T)= \int f1(x)dx$$
$$FPR(T)= \int f0(x)dx$$

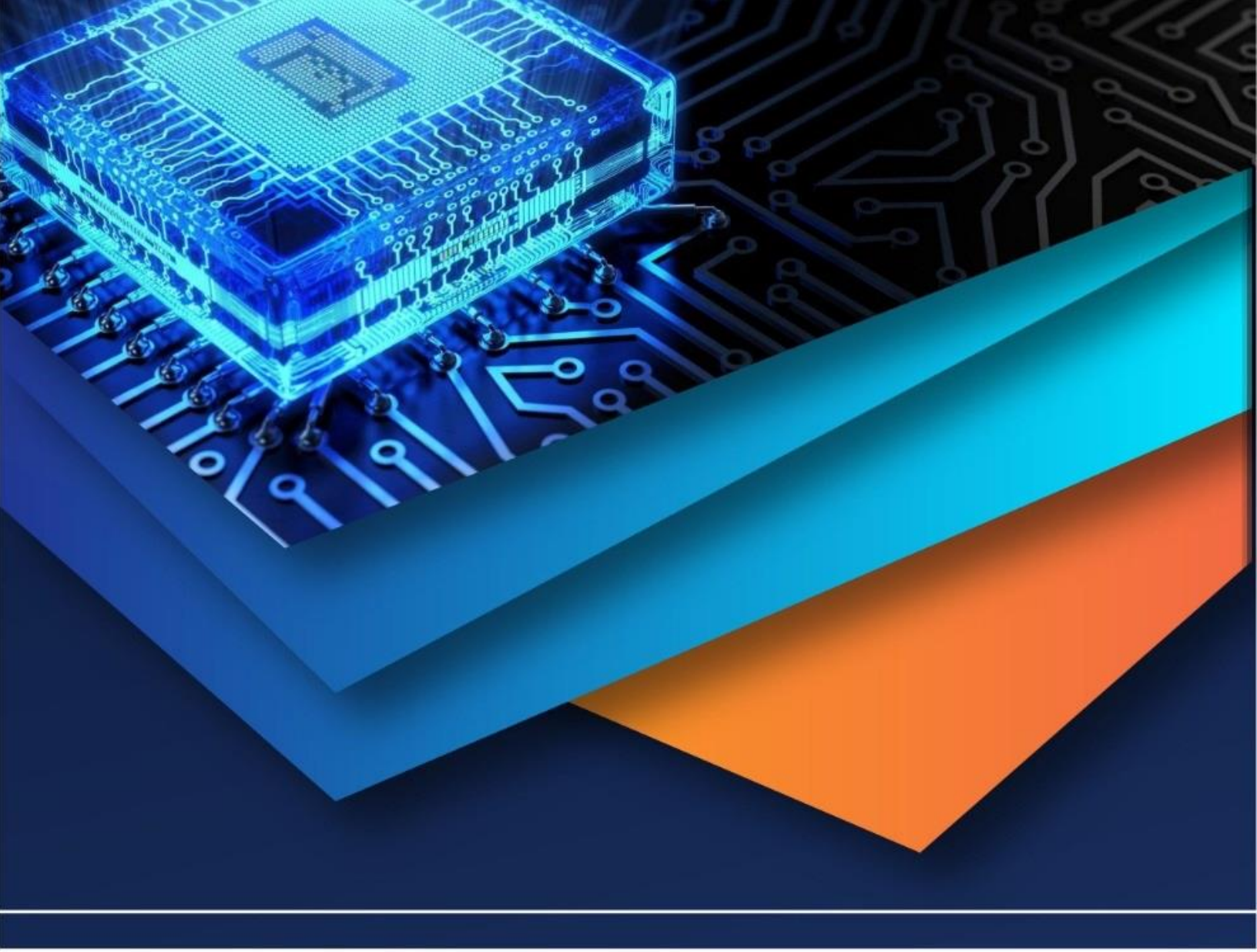
$$z = \frac{| < AUC >_1 - < AUC >_2 |}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}}$$

VI. CONCLUSION

Remote homology detection contains so many methods, tools and algorithms. Here we discussed commonly used methods, tools, and algorithms. Remote homology detection depends on the sequence of proteins. Reducing dimensionality is used to give an accurate result of remote homology detection. The problem lies on classification of sequence of protein families from super families remote homology detection but here some error prediction is also possible. The result fully depends on the sequence of proteins.

REFERENCES

[1] Oscar Bedoya n, IreneTischer ,Reducing dimensionality in remote homology detection using predicted contact maps, Computers in Biology and Medicine 59(2015)64–72.
[2] Tommijaakkola , mark diekhans andDavid Haussler, A Discriminative Framework for Detecting Remote Protein Homologies, Journal Of Computational Biology,Volume 7, Numbers 2,2000,Mary Ann Liebert, Inc.,Pp. 95–114.
[3] Qi-wen Dong, Xiao-long Wang and Lei Lin, Application of latent semantic analysis to protein remote homology detection,School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, Vol. 22 no. 3 2006, Pp. 285– 290.
[4] Hilmi M. Muda , PutehSaad,Razib M.Othman , Remote protein homology detection and fold recognition using two-layer support vector machine classifiers, Computers in Biology and Medicine 41(2011)687–699.
[5] YuchenYanga, Erwin Tantosob and Kuo-Bin Lic, Remote protein homology detection using recurrence quantification analysis and amino acid physicochemical properties, Journal of Theoretical Biology 252 (2008) 145–154.
[6] Isaac Arnold Emerson ,ArumugamAmala, Protein contact maps: A binary depiction of protein 3D structures, PhysicaA 465 (2017) 782–791.
[7] PietroLovato, Alejandro Giorgetti, and ManueleBicego, A Multimodal Approach for Protein Remote Homology Detection , IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 12, No. 5, September/October 2015.
[8] Noah M. Daniels, Andrew Gallant, Norman Ramsey, Lenore J. Cowen, MRFy: Remote Homology Detection for Beta-Structural Proteins Using Markov Random Fields and Stochastic Search , Mathematics Department and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139,Department of Computer Science, Tufts University, Medford, MA 02451.
[9] Bin Liu, Junjie Chen, MingyueGuo, and Xiaolong Wang, Protein remote homology detection and fold recognition based on Sequence-Order Frequency Matrix, IEEE Transactions On Computational Biology And Bioinformatics, Tcbb-2017-04-0157.
[10] Jian-huaYeh, Chun-hsing Chen, Protein Remote Homology Detection Based on Latent Topic Vector Model, 201O,International Conference on Networking and Information Technology.
[11] Bin Liu, Lei Lin, Xiaolong Wang andXuan Wang, Protein remote homology detection using order profiles,2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing.
[12] AsaBen-Hur and Douglas Brutlag, Remote homology detection: a motif based approach, Department of Biochemistry, B400 Beckman Center, Stanford University, CA 94305-5307, USAVol. 19 Suppl. 1 2003, pages i26–i33.
[13] Soft Ngram representation and modeling for protein remote homology detection PietroLovato, Marco Cristani, Member, IEEE and ManueleBicego, Member, IEEE,2016 Transactions on Computational Biology and Bioinformatics.
[14] Bin Liu, Junjie Chen and Xiaolong Wang , Application of learning to rank to protein remote homology detection, School of Computer Science and Technology, Key Laboratory of Network Oriented IntelligentComputation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China and Gordon Life Science Institute, Belmont, MA 02478, USA.
[15] William R. Pearson, An Introduction toSequence Similarity (“Homology”) Searching, Bioinform. 42:3.1.1-3.1.8. C @ 2013 by John Wiley & Sons, Inc.Marianne M. Lee, Ralf Bundschuh, and Michael K. Chan, Distant homology detection using a LEngth and STructure-based sequence Alignment Tool (LESTAT), The Ohio State Biophysics Program, The Ohio State University, Columbus, Ohio 43210.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY