International Journal of Engineering & Technology, 7 (1.9) (2018) 221-224

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Plant and Animal sub cellular component localization prediction using multiple combination of various machine learning approaches

Bipin Nair B J [1] *, Ashik P V [1]

[1] *Department of Computer Science Amrita School of Arts & Science, Mysuru Amrita Vishwa Vidyapeetham, India*
**Corresponding author E-mail: bipinns@gmsil.com*

Abstract

Membrane proteins are encoded in the genome and functionally important in the living organisms. Information on subcellular localization of cellular proteins has a significant role in the function of cell organelles. Discovery of drug target and system biology between localization and biological function are highly correlated. Therefore, we are predicting the localization of protein using various machine learning approaches. The prediction system based on the integration of the outcome of five sequence based sub-classifiers. The subcellular localization prediction of the final result is based on protein profile vector, which is a result of the sub-classifiers.

Keywords: *Amino Acid Composition (AAC); Sub Cellular Localization; Gene Ontology (GO).*

1. Introduction

Bioinformatics has turns into a vital piece of numerous territories of biology, chemistry, computer science. In experimental molecular biology such as signal and sequence processing allow mining of useful outcome from a lot of crude information. Bioinformatics plays a role in the analysis of gene, protein appearance and regulation. The localization of protein helps us to evaluate in the role of protein.

Information on subcellular localization of protein is crucial to study of protein, discovery of drug target and system biology between localization and biological function are highly correlated. Recently various computational forecast strategies have been produced. There is still a need for higher accuracy in subcellular localization prediction.

In our research, we are taking the sample sequence of plant and animal cell protein then implementing various sub- prediction classifiers for feature selection and subcellular localization prediction. We are using sequence similarity, phylogeny profile and advanced gene ontology it improves the prediction of cellular localization of proteins.

2. Literature survey

Researches are reported in the area of sub cellular localization prediction in bioinformatics for predicted drug targets as follows.

Hisham [1] proposed a technique for Predicting Protein Subcellular Localization utilizing gram-negative, gram-positive and non-plant proteins strategies. It is essential for some applications particularly for the disclosure of novel protein and quality infection. For computational reason they utilized microscopic organism's proteins and plant proteins as dataset. The outcomes with six protein limitation datasets demonstrated that the strategy is promising and focused for foreseeing protein confinements. Kuo-Chen [2] proposed the strategy Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites utilizing GO, FunD, SeqEvo and Prediction Engine CE techniques. Webserver marker was made for eukaryotic systems that contain both single region and distinctive zone proteins. From the led trials, ProtLock and HSLPred are two indicators created for distinguishing protein subcellular confinement. W.Y. Yanga [3] proposed a strategy incorporating cell arranging structure the better expectation of protein subcellular areas utilizing SVM and Interdependent subcellular areas strategies. BaCelLo, RH and PLOC datasets are accustomed to deciding the elements of another protein. Paul proposed a strategy. Paul [4] proposed a method prediction with wolf psort using site definition, wolf psort system and classification algorithm. At that point he got the outcome SVM, which picks the leaf hub with the most astounding multiplicative likelihood. Kuo-Chen [5] proposed a method Plant- mPLoc top-down strategy to prediction which used subsampling test and jackknife test. Plant-mPLoc is a tool used for dealing with multiplex plant proteins that can exist at two, or more different location sites. Jackknife test yield a unique result for a given benchmark dataset. Hsiao [6] proposed a method for prediction using different classifiers such as k-nearest neighbor, Naive Bayesian method. Three techniques are using here cell fractionation, electron and fluorescence. The results indicate three compositions as protein feature for subcellular location prediction reaches accuracy. Chaohong [7] this method using K-Nearest Neighbor classifier for prediction. It has a higher predictive success of 88.3%. It uses SWISSPROT database. It gives a better result even though when large sequences are taken. Shibiao Wan [8] proposed a method prediction based on profile arrangement and GO which used Gene Ontology method-searching. Here explain different approaches for constructing GO vectors. The results shown that the fusion of PairPro SVM shows a higher prediction percentage. Kuo-Chen [9] Proposed a prediction algorithm to predict protein subcellular location. Predictive quality is achieved on both methods of self-consistency and overlay testing. This method got increase percentage of 22-30% which

is far better than the protLock algorithm. Qian Xu [10] proposed a method Multitask Learning for Subcellular Location Prediction using Baseline Multitask Learning and Kernel. Here he used two datasets Cell-Ploc and DBSubLoc. The result shows that it has a better prediction accuracy of 25% in best case. Dongjun Yu [11] proposed a method by Parallel Fusion Localization to predict Membrane Protein Sub cell. This method used Support vector machine (SVM) and Simple serial combination strategy. It demonstrate the efficacy of the parallel strategy. Fengmin Li [12] proposed a technique expectation in view of enhanced quadratic discriminant utilizing Increment of assorted variety joined with enhanced quadratic discriminant examination (IDQD).Here higher prescient achievement rates are acquired by the self-consistency test and the folding blade test and enhanced quadratic discriminant investigation is capable forecast. Dan Xie [13] proposed a technique utilizing themes in the expectation of eukaryotic protein for this help vector machine and two informational indexes PDB and SWISSPORT were utilized. This examination shows the consolidated theme technique is extremely compelling in eukaryotic protein subcellular restriction expectation. Hasan Ogul [14] proposed a method with new protein encoding schemes it used n-peptide compositions and Pairwise similarity encoding. This technique had the general precision of 91.3 percent, which is better than the accomplishments of a large number of the current strategies. Minghui Wang [15] A method for predicting evolutionary information and sequence information is proposed. Two datasets RH-2427 and SWN-Unique are used here. The overall accuracy of our method is 72.9%, more than 8% higher than the best available method, LOCnet. Eric Y.T [16] proposed a methods for protein subcellular localization using Support Vector Machine algorithm, k nearest neighbor method and C4.5 decision tree. The exploratory outcomes give additionally reference information to the generally utilized classifier and protein description strategies. BIPIN NAIR B J [17] proposed a method Comparative sequence analysis and 3D structure prediction using Needleman-Wunsch and Pair wise sequence Alignment Algorithm. In this work datasets from PDB as well as kannur4 Amrita hospital. In this method by getting which possess less time complexity and space complexity to get final optimal alignment. The 3D structural changes which is used to identify the stages of lymphoma and the algorithm to determine the various stages based on the maximum unique match property. David T. Okou [18] The Genome Analyzer Platform was proposed to sequence diploid target regions. Here, microarray-based genomic assay (MGS) was used to sequence on Illumine Genome An Alyzer microarray-based genomic selection (MGS) and sequenced at the IGA stage. The information suggests that MGS / IGA sequencing is a fully reproducible and accurate strategy that will surely increase the identification and elucidation of human genome variants that will be revealed by the focus of individual genome sequencing. Xumi Qu [19] proposed a method prediction with Multiple Sites Based on Multiple Features Fusion using k nearest neighbor, Jack Knife test and flexible neural tree (FNT) algorithms. In this technique five component extraction strategies: physical and synthetic properties, AAC, Stereo-substance property, N-terminal flag and amino acid list distribution. Xiaoqing Peng [20] proposed a method identify essential proteins for dissimilar species by participating protein subcellular localization information. Three methods are used here first one Evaluating the Importance of Each Compartment, The Importance of Interactions and Compartment Importance Centrality. The results show that CIC method has better performance to predict essential protein on four species.

In present literature, they are predicting the sub cellular localization for animal cell with some of the specific cell organelles protein using some machine learning algorithms.in our work we are doing the sub cellular localization prediction of both plant and animal cell using feature selection, and sub prediction methods. Finally, we are predicting the sub cellular localization of protein using k-nearest neighbor classifier. Through sub cellular localization, we can predict the target drug for various diseases.

3. Problem formulation

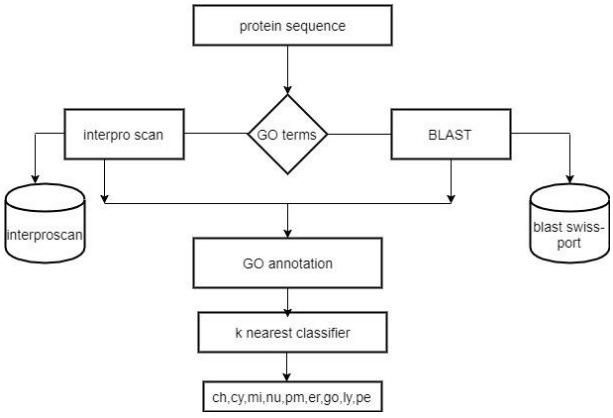
Despite the fact that there as of now exist a great deal of computational prediction techniques, there is still opportunity to get better. This is because of the way that the protein arranging process is exceptionally unpredictable and not yet surely knew. Just a little bit of proteins has obviously identifiable arranging signals in their essential succession. As an outcome, accessible expectation techniques are frequently either concentrated for the forecast of not very many localization with higher exactness or for the forecast of an extensive variety of limitations with diminished precision.

4. Problem definition

The aim of our work is to build a tool that can accurately predict the outcome of protein targeting in cells as well as cell organelles. Our approach for foreseeing protein subcellular confinement proposed framework incorporates a five sub-indicators in view of the amino corrosive structure, the recognition of arranging signals, phylogeny term profiles and advanced gene ontology terms. The predicted result of the proposed work improved than the existing work. In our work we deals with membrane protein sequence and predicts all of the main eukaryotic localizations based on a sequence that consists of a combination of animal and plant cell protein localizations. The prediction system based on the integration of the outcome of five sequence based sub classifiers N-Target, arranging signal, amino acid composition, phylogeny profile and Advanced gene ontology into a protein profile vector. The sub classifiers utilize the overall amino acid composition.

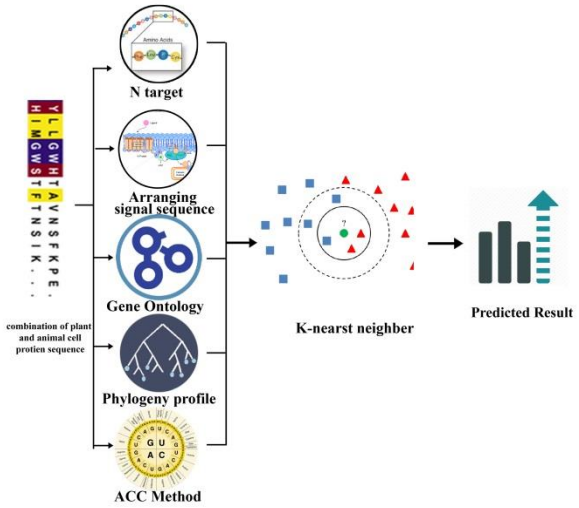
Sub prediction methods

Here we are taking protein sequence and n terminal targeting peptide used to predict plant and animal protein. Then the windowing technique used to scan the portion of n terminal and amino acid composition in the window used to get the outcome is the input for the final layer prediction. Arranging prediction is another sub-prediction here scanning protein sequence is the first step then select the secretory pathway in the membrane protein. Another sub-prediction technique is amino acid composition here converting protein sequence into twenty amino acid structure-based composition then the outcome is the subcellular localization on cell organelles. Advanced gene ontology here discovery of gene ontology information. GO is a technique for defining gene products in terms of the biological process in the cellular organelles. Interpro scan and blast are used to find gene annotation. Phylogeny profile sub-prediction is for finding the subcellular localization of the homologous protein. Here blast is used to create phylogeny profile.



The diagram represents the flow of the advanced gene ontology

ARCHITECTURE



5. Algorithm

Phylogenetic profile
//find the vector similarity between protein sequence and best sequence match
Input protein sequence
Input genome file
Initialize q is query sequence
Initialize g is genomes
Initialize Aqg is bit score
Initialize Aqq is self-bit score
Initialize Bqg is similarity rate
Bqg=Aqg+Aqq
Always Aqg less than Aqq
Bqg => 0<Bqg<1
Bqg close to 1 then sequence present in the genome
Bqg close to 0 then sequence absent in the genome
Gene ontology
Input protein sequence
GO terms set 1 vector
Otherwise
Go terms set 0 vector
InterProScan scan the GO terms
Output protein sequence signature profile

6. Data set

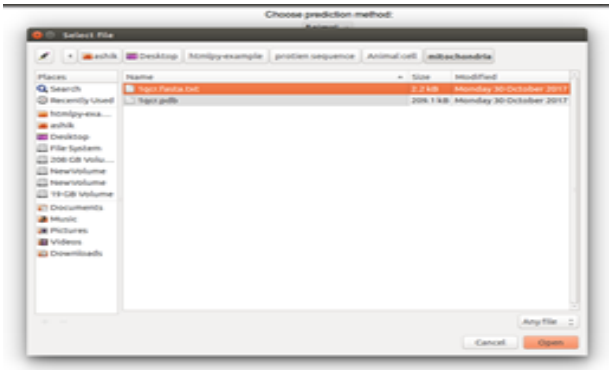
In this prediction method we are using the data sets are protein sequence fasta format files from animal and plant cell.

```
>5XJG:A|PDBID|CHAIN|SEQUENCE
DSSDEASVSPIADNEREAVTLLLGYLEDKDQLDFYSGGPLKALTTLVYSDNLNLQRSAALAFAE
ITEKYVRQVSREVLEP
ILILLQSQDPQIQVAACAALGNLAVNNENKLLIVEMGGLEPLINQMMGDNVEVQCNAVGCITNL
ATRDDNKHKIATSGAL
IPLTKLAKSKHIRVQRNATGALLNMTHSEENRKELVNAGAVPVLVSLLSSTDPDVQYYCTTALS
NIAVDEANRKKLAQTE
PRLVSKLVSLMDSPSSRVKCQATLALRNLASDTSYQLEIVRAGGLPHLVKLIQSDSIPLVLASV
ACIRNISIHPLNEGLI
VDAGFLKPLVRLLDYKDSEEIQCHAVSTLRNLAASSEKNRKEFFESGAVEKCKELALDSPVSVQ
SEISACFAILALADVS
KLDLLEANILDALIPMTFSQNQEVSGNAAAALANLCSRVNNYTKIIEAWDRPNEGIRGFLIRFL
KSDYATFEHIALWTIL
QLLESHNDKVEDLVKNDDDIINGVRK
>5XJG:B|PDBID|CHAIN|SEQUENCE
NREKDCSSSSEVESQSKCRKESTAEPDSLSRDTRTTSSLKSSTSFPISFKGSIDLKSLNQPSSL
LHIQVSPTKSSNLDAQ
VNTEQAYSQPFRY
>5XJG:C|PDBID|CHAIN|SEQUENCE
DSSDEASVSPIADNEREAVTLLLGYLEDKDQLDFYSGGPLKALTTLVYSDNLNLQRSAALAFAE
ITEKYVRQVSREVLEP
ILILLQSQDPQIQVAACAALGNLAVNNENKLLIVEMGGLEPLINQMMGDNVEVQCNAVGCITNL
ATRDDNKHKIATSGAL
IPLTKLAKSKHIRVQRNATGALLNMTHSEENRKELVNAGAVPVLVSLLSSTDPDVQYYCTTALS
NIAVDEANRKKLAQTE
PRLVSKLVSLMDSPSSRVKCQATLALRNLASDTSYQLEIVRAGGLPHLVKLIQSDSIPLVLASV
ACIRNISIHPLNEGLI
VDAGFLKPLVRLLDYKDSEEIQCHAVSTLRNLAASSEKNRKEFFESGAVEKCKELALDSPVSVQ
SEISACFAILALADVS
KLDLLEANILDALIPMTFSQNQEVSGNAAAALANLCSRVNNYTKIIEAWDRPNEGIRGFLIRFL
KSDYATFEHIALWTIL
QLLESHNDKVEDLVKNDDDIINGVRK
>5XJG:D|PDBID|CHAIN|SEQUENCE
```

Here is the data set in the form of Fasta format we can use the format of both animal and plant cell protein.

7. Experimental result

This is the user interface part of the prediction system here we can access the data set. Here we can choose animal or plant category.



There is an option for selecting Fasta format file from the file directory but the file format should be Fasta format otherwise cannot use this as input.

Here we have a text area, this is for accepting proper Fasta format protein sequence. This sequence used to predict the subcellular localization.

this above fig shows the outcome of the prediction system, here we get the different components of plant and animal cell organnel proteins like nuclear, Golgi apparatus,plasma membrane, cytoplasmic, mitochondrial, chloroplast, extracellular, peroxisomal, endoplasmic reticulum, lysosomal.

8. Conclusion

In our proposed prediction system predicting for subcellular localization, the prediction system integrates five sub-predictors based on the amino acid alignment in the discovery of arranging signals, n-target, phylogeny profile and gene ontology terms. The current proposed prediction performance is moderate.finally we are using k-nearest method for effective predicion for cellular localization

References

[1] Al-Mubaid, H., & Nguyen, D. B. (2014, November). New Feature Weighting Technique for Predicting Protein Subcellular Localization. In Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on (pp. 163-167). IEEE. https://doi.org/10.1109/BIBE.2014.35.

[2] Chou, K. C., & Shen, H. B. (2010). A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS One, 5(4), e9931. https://doi.org/10.1371/journal.pone.0009931.

[3] Yang, W. Y., Lu, B. L., & Kwok, J. T. (2011). Incorporating cellular sorting structure for better prediction of protein subcellular locations. Journal of Experimental & Theoretical Artificial Intelligence, 23(1), 79-95. https://doi.org/10.1080/0952813X.2010.506303.

[4] Horton, P., Park, K. J., Obayashi, T., & Nakai, K. (2006, February). Protein Subcellular Localisation Prediction with WoLF PSORT. In APBC (Vol. 39).

[5] Chou, K. C., & Shen, H. B. (2010). Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PloS one, 5(6), e11335. E. P. Wigner, "Theory of traveling-wave optical laser,"Phys .Rev. vol. 134, pp. A635–A646, Dec. 1965. https://doi.org/10.1371/journal.pone.0011335.

[6] Hsiao, H. C., Chen, S. H., Chang, J. P. C., & Tsai, J. J. (2008). Predicting Subcellular Locations of Eukaryotic Proteins Using Bayesian and k-Nearest Neighbor Classifiers. Journal of Information Science & Engineering, 24(5).

[7] Song, C., & Shi, F. (2010). Prediction of Subcellular Localization of Apoptosis Proteins by Dipeptide Composition. JDCTA, 4(1), 32-36. https://doi.org/10.4156/jdcta.vol4.issue1.4.

[8] Wan, S., Mak, M. W., & Kung, S. Y. (2011, September). Protein subcellular localization prediction based on profile alignment and Gene Ontology. In Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on (pp. 1-6). IEEE.

[9] Chou, K. C., & Elrod, D. W. (1999). Protein subcellular location prediction. Protein engineering, 12(2), 107-118. https://doi.org/10.1093/protein/12.2.107.

[10] Xu, Q., Pan, S. J., Xue, H. H., & Yang, Q. (2011). Multitask learning for protein subcellular location prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 8(3), 748-759. https://doi.org/10.1109/TCBB.2010.22.

[11] Yu, D., Wu, X., Shen, H., Yang, J., Tang, Z., Qi, Y., & Yang, J. (2012). Enhancing membrane protein subcellular localization prediction by parallel fusion of multi-view features. IEEE transactions on nan bioscience, 11(4), 375-385 https://doi.org/10.1109/TNB.2012.2208473.

[12] Li, F., & Zhou, H. (2011, October). Predicting protein subcellular location based on improved quadratic discriminant. In Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on (Vol. 4, pp. 1989-1992). IEEE. https://doi.org/10.1109/BMEI.2011.6098687.

[13] Xie, D., Li, A., Lin, X., Wang, M., Jiang, Z., & Feng, H. (2006, January). Using motifs in the prediction of eukaryotic protein subcellular localization. In Engineering in Medicine and Biology Society, 2005. IEEEEMBS 2005. 27th Annual International Conference of the (pp. 2802-2804). IEEE.

[14] Ogul, H., & Mumcuoglu, E. U. (2007). Subcellular localization prediction with new protein encoding schemes. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(2). https://doi.org/10.1109/TCBB.2007.070209.

[15] Ogul, H., & Mumcuoglu, E. U. (2007). Subcellular localization prediction with new protein encoding schemes. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(2). https://doi.org/10.1109/TCBB.2007.070209.

[16] Juan, E. Y., Chang, J. H., Li, C. H., & Chen, B. Y. (2011, June). Methods for Protein Subcellular Localization Prediction. In Complex, Intelligent and Software Intensive Systems (CISIS), 2011 International https://doi.org/10.1109/CISIS.2011.91.

[17] Bipin Nair B J, Pranav V, Athulya Viswan.(2017) Comparative sequence analysis and 3D structure prediction of various stages of lymphoma using a combined approach of pairwise and Needleman Wunsch (ICIIECS)

[18] Okou, D. T., Locke, A. E., Steinberg, K. M., Hagen, K., Athri, P., Shetty, A. C., & Zwick, M. E. (2009). Combining Microarray-based Genomic Selection (MGS) with the Illumina Genome Analyzer Platform to Sequence Diploid Target Regions. Annals of human genetics, 73(5), 502-513. https://doi.org/10.1111/j.1469-1809.2009.00530.x.

[19] Qu, X., Chen, Y., Qiao, S., Wang, D., & Zhao, Q. (2014, August). Predicting the subcellular localization of proteins with multiple sites based on multiple features fusion. In International Conference on Intelligent Computing (pp. 456-465). Springer, Cham. https://doi.org/10.1007/978-3-319-09330-7_53.

[20] Peng, X., Wang, J., Zhong, J., Luo, J., & Pan, Y. (2015, November). An efficient method to identify essential proteins for different species by integrating protein subcellular localization information. In Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on (pp. 277-280). IEEE. https://doi.org/10.1109/BIBM.2015.7359693.