Protocol of Rice Genome Annotation through Comparative Functional Genomics Approach

S Kushwaha, M Shakya

Citation

S Kushwaha, M Shakya. *Protocol of Rice Genome Annotation through Comparative Functional Genomics Approach*. The Internet Journal of Genomics and Proteomics. 2008 Volume 4 Number 1.

Abstract

Identification & characterization of genes and proteins are very important task, but these are slow processes as compared to the genome sequencing due to lack of annotation protocol. In this paper, efforts have been made to characterize Oryza sativa var. Japonica genome on the basis of comparative functional genomics. Various online tools are used in order to annotate the hypothetical sequences of rice for characterization and depiction of sub-cellular location. The results obtained from the bioinformatics tools are compared statistically through confidence score for high specificity and sensitivity. In the present study chromosome-1 is studied for rice genome interpretation, which consists of 500 hypothetical proteins. Results showed that chromosome-1 is characterized by Nuclear (50%) Mitochondrial (21%), Plastid (12%), Secretary Protein (2%) Plasma Membrane (6%), Cytoplasmic (5%), Ext. Cellular Proteins (4%) with eleven ESTs, which are concerned to root and their homologs are present in the chromosome 2, 4 and 9.

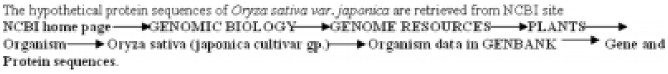
INTRODUCTION

Genome sequencing of animals, plants and microbes is going very fast and genomic data increasing at very rapid rate., So storage of data and transformation of these data into information are critically needed. Genomic analysis of cereal crops like rice, wheat and maize, will contribute greatly to improvement to their productivity. Rice genome is very important among the cereal crops because of its small genome size (430 Mb) and high degree of chromosomal co-linearity with other cereal crops [22] like maize, wheat, barley and sorghum. It is a major food supply source for more than half of the world's population. In the countries like Asia, Africa, and Latin America where the demand for rice is at the top priority, the population is continuously increasing [9]. There is need to develop novel techniques to breed new varieties of rice. Following the successful completion of human genome project, a new era of whole genome science has emerged ranging from humans to plants and yeast [4]. Comparisons between distantly related genomes provide insight into the universality of biological mechanisms and identify experimental models for studying complex processes. The IRGSP, a public consortium of publicly funded laboratories has generated finished quality sequence of the entire genome using the clone-by clone sequencing strategy [13,21] and made it available to public domain. With the completion of the sequencing process, annotation is a dynamic process essential to add the value to the genome [2,4,5].

The major task in genome annotation is to identify the genes termed as structural annotation, which relies on the computational methods. Considering the importance of comparative genetics in the forefront of new knowledge on plant genomes and genes, comparative bioinformatics remains an essential strategy to gain new insights on the needs and expectations on rice genomics. The information regarding genes, their proteins and their specificity is obtained from cellular and subcellular locations of proteins [1,8,3]. Bioinformatics approaches are helping in expedite the determination of protein cellular and subcellular locations [10,11,12,14]. To explore this problem, proteins were classified [6], according to their specific characterization and subcellular locations [16,17,19], into the following 12 groups: (1) Chloroplast, (2) Cytoplasm, (3) Cytoskeleton, (4) Endoplasmic Reticulum, (5) Extracellular, (6) Golgi Apparatus, (7) Lysosome, (8) Mitochondria, (9) Nucleus, (10) Peroxisome, (11) Plasma Membrane (12) Vacuole. ESTs are c-DNA clone that has been arbitrarily chosen and subjected to single-pass sequencing in both directions, which gives us a rough canvassing of a tissue or organisms transcriptional content [7]. They provide a highly cost & time effective method of accessing the desired feature. The cellular location (tissue) identification by the ESTs (japonica variety11) from root is reported in NCBI EST database [23]. EST similarity search is explored with the help of BLAST. It is anticipated that the classification scheme, concept and prediction protocol can expedite the property determination of new genes and their protein. It may also use in the prioritization of genes for potential molecular targets identification [20,21]. Here we, have made a comparative functional genomics analysis of results obtained through various tools. Then statistical approach is used i.e. we have assigned confidence level to the functionally annotated sequences.

MATERIAL & METHODS

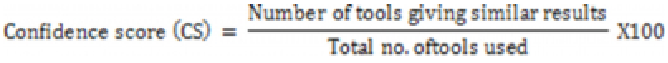
Figure 1



The hypothetical protein sequences of *Oryza sativa var. japonica* are retrieved from NCBI site
NCBI home page ⟶ GENOMIC BIOLOGY ⟶ GENOME RESOURCES ⟶ PLANTS ⟶ Organism ⟶ Oryza sativa (japonica cultivar gp.) ⟶ Organism data in GENBANK ⟶ Gene and Protein sequences.

When work was started chromosome1 was containing of 500 hypothetical sequences. The relevance of work is increased as the genome sequencing is complete but genes detail is still to be explored in the gene bank. Each individual sequence is now run with the help of functional annotating tools. The tools used here are Interpro, SVMProt (Support vector machine based), Pfam (conserved domain based), GFSelector, MIPS BLAST, TAIR Tools, and PROTFun. For subcellular localization prediction, the tools used here are SubLoc, ESL Predicts, TargetP, Cello, Psort, Predator, Mito-II, Chloro-I, LOC-tree [15,17].These tools are used for identification and characterization of genes products i.e. Proteins. For calculation of confidence score (CS) the following formula is used

Figure 2



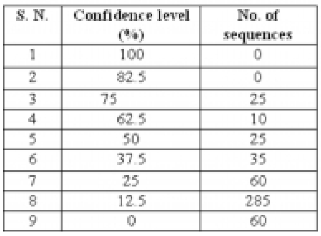
$$\text{Confidence score (CS)} = \frac{\text{Number of tools giving similar results}}{\text{Total no. oftools used}} \text{X100}$$

RESULTS & DISCUSSIONS

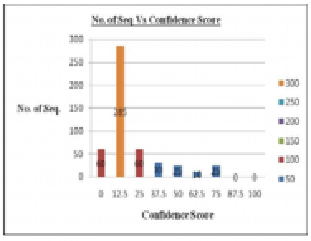
The results from all the tools (Interpro, Pfam, GFSelector, MIPS BLAST, TAIR Tools, PROTFun and SVMProt.) are stored in the form of excel sheet. Results from these tools are then compared by calculating confidence level for the functions i.e. it gives a statistical analysis of tools giving common function for one sequence. This can lead to a conclusion that how much confidence is there for any function assigned to a sequence. From the 500 hypothetical protein sequences, 25 sequences have been assigned a protein function with 75 % confidence level, 10 sequences with 62.5 % confidence level, 25 sequences with 50 % confidence level, 35 sequences with 37.5 % confidence level, 60 sequences with 25 % confidence level, 285 sequences with 12.5 % confidence level and remaining 60 sequences are with very low percentage confidence [Table-1]. The graphical representation of Table-1 is shown in Graph-1.

Figure 3

Table 1: Confidence level for the identification & characterization of sequences of chromosome-1 and Graph-1 is Graphical representation of Table-1



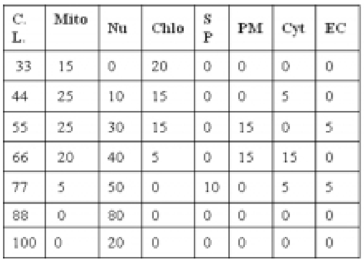
S. N.	Confidence level (%)	No. of sequences
1	100	0
2	82.5	0
3	75	25
4	62.5	10
5	50	25
6	37.5	35
7	25	60
8	12.5	285
9	0	60



The characterization of protein which showed 12.5% cutoff score were selected for subcellular localization prediction and then the confidence level of these proteins was calculated statistically. For subcellular localization prediction, the tools used here are SubLoc, ESL Predicts, TargetP, Cello, Psort, Predator, Mito-II, Chloro-I, LOC-tree [14,15,16]. The results from all subcellular localization prediction tools are stored in the form of excel sheet. Results from these tools are then compared by calculating confidence level for the functions i.e. it gives a statistical analysis of tools giving common function for one sequence. This can lead to a conclusion that chromosome-1 is characterized by Nuclear (Nu.) (50%) Mitochondrial (Mito)(21%), plastid (Chlo.) (12%), Secretary Protein (SP)(2%), Plasma Membrane (PM) (6%), Cytoplasmic proteins (Cyt) (5%), Ext. Cellular Proteins (EC) (4%) [Table-2].

Figure 4

Table 2: Confidence level at subcellular locations of oryza sequences of chromosome-1 and Graph-2 is Graphical representation of Table-2.



C. L.	Mito	Nu	Chlo	S P	PM	Cyt	EC
33	15	0	20	0	0	0	0
44	25	10	15	0	0	5	0
55	25	30	15	0	15	0	5
66	20	40	5	0	15	15	0
77	5	50	0	10	0	5	5
88	0	80	0	0	0	0	0
100	0	20	0	0	0	0	0

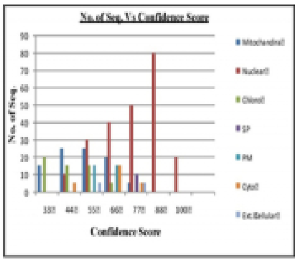
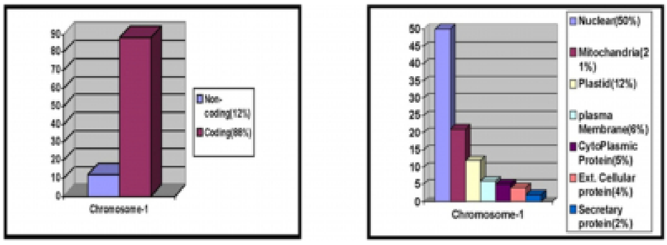


Figure 5

Figure 1: Non-Coding and coding proportion of chromosome-1 of Fig-2 shows Characterization of coding region of chromosome-1 of .



Here cellular location (tissue) identification is done by the ESTs [23]. In NCBI EST database has 44 entries for the Oryza sativa (japonica -11, Indica-33). EST similarity search is investigated with the help of BLAST. BLAST-N and BLAST-X similarity analysis results show that beside chromosomes-1, chromosome-2, 4and 9 also concerned with root. Genomic location of chromosomes is considered for >90% of the EST sequences. This information will be valuable in selecting possible candidate genes from regions of the oryza genome, which can be mapped genetically.

CONCLUSION

The different tools with different logic and algorithms were used to analyse these sequences and their results indicate common functions for these sequences, but with different levels of confidence. The sequences with higher confidence levels can be given more priority for the purpose of research and development to improve rice cereal. The sequences with low confidence levels imply that they do not have significant homology with data sets already present in databases. In characterization process 60 sequences are not showing the results i.e. 12% of chromosome1, have no functional significance (non-coding sequences) [fig-1]and 88% of chromosome 1 has biological activity [Molecular Regulatory- Nuclear (50%), mitochondrial (21%), plastid (12%), Secretary Protein (2%), PlasmaMembrane (6%), Cytoplasmic (5%), Ext. CellularProteins (4%)] of Oryza [fig-2]. All these predictions are made on the basis of bioinformatics tools &techniques, by statistical analysis.

FUTURE DIRECTIONS

The present model can be further extended with same modifications, if necessary for analysis of other varieties of Oryza sativa and other plants as well as animal which are in sequencing phases. The information generated from the present analysis may also prove to be useful in developing databases, tools and softwares for analysis of data. The databases used here are in great demand, which can integrate results from other automated systems and can give updated results. Tools and resources are being developed to maximally interpret and annotate the rice genome sequence. These include improvement of gene prediction programs and identification of molecular resources for mapping. The annotations will be more refined as the sequences are updated; i.e. genome sequences and c-DNA sequences are revised with time and improvement of technologies. Moreover continuous efforts are needed to interlink the data for which comparative bioinformatics is needed. Results from various tools need to be compared and integrated so that annotations can be generated from different bioinformatics strategies. After a decade or so, we will be able to decode information available in nucleotide sequences and by comparison of information from related cereal species. This knowledge will contribute positively to breeding of new varieties with more favourable traits as well as leading varieties by knowing the proteins of tissue part and their concerned genes.

ACKNOWLEDGEMENTS

We are grateful to Department of Bioinformatics, MANIT, Bhopal, India for support and cooperation.

References

1. Hagit S., Annette H. , Scott B., Torsten B., Pierre D., and Oliver K., SherLoc: "High-Accuracy Prediction of Protein Subcellular Localization by integrating Text and Proteins Sequence Data", Bioinformatics, 23, 1410-1417, 2007.
2. Mewes H. W., Amid C., Arnold R., Frishman D., Guldener U., Mannhaupt G., Munsterkotter M., Pagel P., Strack N., Stumpen V., Warfsmann J., and Ruepp A., MIPS: "Analysis and annotation of proteins from whole genomes", Nucleic Acids Res., 32 :D41, 44, 2004.
3. Yamanishi Y., Vert J.P., and Kanehisa M., "Protein network inference from multiple genomic data: A supervised approach", Bioinformatics, 20:i363- i370, 2004.
4. Karen R. C., Shuai W., Rama B., Maria C. C., Kara D., Selina S. D., Stacia R. E., Becket F., Dianna G. F., Jodi E. H., Eurie L. H., Laurie I. T., Robert N., Anand S., Barry S., Chandra L. T., Rey A., Gail B., Qing D., Christopher L., Mark S., David B., and Cherry J. M. "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms", Nucleic Acids Res., 32:D311-314, 2004.
5. Huh W.K., Falvo J. V., Gerke L.C., Carroll A.S., Howson R.W., Weissman J.S., and O'Shea E. K., "Global analysis of protein localization in budding yeast", Nature, 425:686- 691, 2003.
6. Wu C. H., Huang H., Yeh L. S. and Barker W. C. "Protein family classification and function annotation", Comput. Biol. Chem. 27, 37-47, 2003.
7. Mootha V. K., Bunkenborg J., Olsen J. V., Hjerrild M., Wisniewski J. R., Stahl E., Bolouri M. S., Ray H. N., Sihag S., Kamal M., Patterson N., Lander E. S., and Mann M. (2003) " Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria". Cell 115, 629-64, 2003.

8. Fen, Z. P., "An overview on predicting the subcellular location of a protein". In Silico Biol. 2, 291-303, 2002.
9. Takuji Sasaki, "Rice Genomics to understand rice plant as an assembly of genetic codes", Current Science, 83, 834-839, 2002.
10. Emanuelsson O., and Heijne G. V. "Prediction of organellar targeting signals". Biochim. Biophys. Acta. 1541,114-119, 2001.
11. Emanuelsson O., Nielsen H., Brunak S., and Von H. G. " Predicting subcellular localization of proteins based on their N-terminal amino acid sequence", J. Mol. Biol. 300, 10051016, 2000.
12. Nakai K. "Protein sorting signals and prediction of subcellular localization", Adv. Protein Chem. 54, 277-344, 2000.
13. Corthals G. L., Wasinger V. C., Hochstrasser D. F., and Sanchez J. C. "The dynamic range of protein expression:a challenge for proteomic research", Electrophoresis 21, 1104-1115, 2000.
14. Cokol M., Nair R., and Rost B., "Finding nuclear localization signals", EMBO Rep, 1:411-415, 2000.
15. Nakai K., and Horton P. "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization", Trends Biochem. Sci. 24, 34-35, 1999.
16. Chou K. C. and Elrod D. W. "Prediction of membrane protein types and subcellular locations", Protein ,34, 137-153, 1999a.
17. Eisenhaber F. and P. Bork.Wanted: "subcellular localization of proteins based on sequence", Trends in Cell Biology, 8, 169-170, 1998.
18. Andrade M. A., O'Donoghue S. I. and B. Rost, "Adaptation of protein surfaces to subcellular location", J. Mol. Biol. 276, 517-525,1998.
19. Cedano J., Aloy P., Perez-Pons J. A. and Querol E. "Relation between amino acid composition and cellular location of proteins", J. Mol. Biol. 266, 594-600, 1997.
20. Klose J., and Kobalz U. "Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome", Electrophoresis, 16, 1034-1059, 1995.
21. Sasaki T. and Burr B., "The International Rice Genome Sequencing Project: the effort to completely sequence the rice genome, The map based sequence of the Rice genome", Nature ,436,793-800, 2000.
22. Liang F., Quackenbush J. (2000) "An optimized protocol for analysis of EST sequences". Nucleic Acids Res., 28, 3657-3665, 2000.

Author Information

Sandeep Kushwaha, M.Sc., M.Phil.
Department of Bioinformatics, MANIT

Madhvi Shakya, Ph.D.
Department of Bioinformatics, MANIT