Research Article

Prediction of homologous genes by extracting *Glycine max* transcriptome using Hidden Markov Model

Rakesh Sharma[1,3], Monika[1], Vandana Nunia[4], Shailesh Kumar[2], S. L. Kothari[2], Sumita Kachhwaha[1,3*]

[1]*Bioinformatics Infrastructure Facility (DBT-BIF), University of Rajasthan, Jaipur- 302004, Rajasthan, India*

[2]*Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur- 303002, Rajasthan, India*

[3]*Departemnt of Botany, University of Rajasthan, Jaipur- 302004, Rajasthan, India*

[4]*Department of Zoology, University of Rajasthan, Jaipur- 302004, Rajasthan, India*

Received: 11 April 2019 Revised: 27 May 2019 Accepted: 7 July 2019

Abstract

Objective: The objective of the work is to develop a Hidden Markov Model (HMM) based approach for finding gene family from RNAseq data in *Glycine max.* **Material and Methods:** The publicly available RNAseq data for *Glycine max* was taken from **S**equence **R**etrieval **A**rchive (SRA) accession number SRR3090710, SRR3090711, SRR3090712 and SRR3090713. This quality of transcriptomics data was observed from FASTQC tool which was further filtered through Trimmomatic tool for filtering adapter and vector noises. Sequences of phred quality score ≥20 taken for further analysis where the sequences below this were removed to produce filtered data. The quality sequences were processed through Tuxedo protocol for alignment and assembly to produce transcript. The transcripts were processed by TransDecoder which identifies putative open reading frames and translates it to protein sequence. ClustalW was used for multiple sequence alignment generation. The alignment file for DREB (dehydration responsive element binding) candidate protein was used to build HMM profiles of translated transcripts. The HMM profile was used for finding the homologous sequences from RNAseq data. **Results:** The model developed through this method was tested by HMM search on E-value < 0.001 and we found 21 gene which have DREB like activity based on their HMM profile analysis. **Conclusion:** The method applied in this work is a novel method for identification of homologous gens for RNAseq datasets.

***Keywords*:** Transcriptomics, Hidden Markov Model, alignment, bigdata, gene family, homologous

Introduction

In the recent years the data driven science is booming in every filed of sciences, where they are applied to find the solutions for unsolved questions in a logical manner. This data driven science is also referred as bigdata analysis. The bigdata analysis is possible because of advancement in the computing capacity and development of advance software. In biological science the size of genomics data is growing rapidly and has a vast variety of big data applications. Analyzing DNA, RNA and proteins give molecular insight into the mechanism underlying the biological phenomena. DNA is dynamic in nature and evolving in every generation by means of mutation, homologous recombination and transposition. These changes results to variation in a population. These molecular variations offer a higher reproductive fitness to organisms, which also increases in number by means of reproduction. Duplication events in DNA create multiple copies of a gene, and these duplicated genes changed independently generation after generation. Duplicated genes can affect the quantity of gene product or it can acquire a novel function. Duplication events have contributed to vast diversity of plant and animal species. Production of floral structures, induction of disease resistance, adaptation to stress, grain quality, fruit shape, and flowering time are result of duplication events. Therefore, understanding the mechanisms and impacts of gene duplication will be important for future studies of plants in

Address for Corresponding Author:
Dr. Sumita Kachhwaha
Bioinformatics Infrastructure Facility (DBT-BIF),
University of Rajasthan, Jaipur- 302004, Rajasthan, India.
Email: kachhwahasumita@rediffmail.com

DOI: https://doi.org/10.31024/ajpp.2019.5.6.6

general and has potential for improving the agronomically important traits (Panchy et al., 2016). Because of advance sequencing platforms the genomes sequencing capacity is increasing exponentially, so the data of sequencing reads also. It is comparably easy to search for any gene human and their homologue, but plant genomes are many folds larger with numerous variety and species, plant biologist often encounter problems while detecting homologous genes. To decode the plant genome, there is a need to develop novel procedure, which can efficiently search species specific homologous genes. Orthologous or paralogous homologous genes have conserved pattern of similarity which can be identified using pattern searching algorithms. Now a days next generation sequencing technology called RNASeq, is used for transcriptome analysis of an organism. Using RNASeq platform both gene expression analysis and sequencing can be done simultaneously. It is easy to retrieve sequences of mRNA transcripts from RNASeq data, which can be converted into putative proteins. Protein sequences would provide better understanding of evolutionary signature than DNA sequences because of degeneracy in the genetic code. Human arm, dolphin's flipper, a bird's wing, and dog's leg are considered homologous structures. The underlying evolutionary signature is anatomical commonalities demonstrating descent from a common ancestor. Likewise, in proteins evolutionary signature are found in the functional domain of the proteins. So, if we have sequences of candidate homologous proteins for related varieties and species, we can detect their counterpart in newly sequenced species. This type of modelling can be employed using computational methods like Hidden Markov Models (HMM) (Yoon, 2009). Multiple sequence alignment provides all the parameters to run an HMM model. Present study is an application of HMMs for homologous gene finding by pattern recognition in a newly sequenced plant genome. We have used this pipeline to find homologous genes in soybean.

Material and methods

Data retrieval

Sequence Read Archive (SRA) is an open source repository of high throughput sequencing data (Leinonen et al., 2011). For comparative genomic studies, it provides raw sequencing reads of DNA and RNA and their alignments. For this study paired end raw sequenced reads of RNAseq is examining infection of different Soybean mosaic virus (SMV) isolates, L (G2 strain), LRB (G2 strain) and G7 (G7 strain) in *Glycine max* (Cultivar Williams 82) were retrieved from the SRA accession no. PRJNA308211 (Chen et al.,2016).

Data cleaning

Total 4 paired end samples containing raw sequence read were cleaned by Trim Galore and Trimmomatic (Bolger et al., 2014). FastQC (Andrews, 2010) was used to proofread and quality control. Phred score 20 was taken as a threshold for base quality. Each of the SRA files was converted into a FASTA file using the fast-dump tool (version 2.5.7). There are chances of abnormalities produced either during sequencing reaction or library preparation. Hence the quality of sequence reads was evaluated with FastQC tool. Low-quality sequences, PCR primers, base adapters and overrepresented sequences (commonly referred to as Illumina-specific sequences) were removed from RNA sequencing reads by using Trimmomatic toolkit v0.32.3. This tool starts scanning at the 5'end with a 4-base sliding window, and trims those bases whose average quality falls below a minimum threshold i.e. phred score of < 20. This resulted in reads with lengths ranging from as low as 20 bp to as high as 185 bp.

Transcriptome alignment and assembly

Tuxedo protocol (Trapnell et al., 2012) was used to align and assemble cDNA transcripts. This protocol uses Tophat and cufflinks for alignment and transcripts assembly respectively.

ORF finding and protein translation

TransDecoder (https://github.com/TransDecoder) tool identifies candidate coding regions within transcript sequences, such as those constructed based on RNA-Seq alignments to the genome using Tophat (Trapnell et al., 2009) and Cufflinks. TransDecoder take cDNA sequences as input and output putative protein sequences as fast file. The following command is used to run TransDecoder with minimum protein length threshold was set up in 100.

```
perl TransDecoder.LongOrfs -t cuffmerged_trascripts.fasta
```

Multiple sequence alignment and HMM modelling

Multiple sequence alignment provides all the three types of probabilities start, transition and emission. Homologous sequences of drab protein were retrieved from the NCBI. ClustalW used for multiple sequence alignment. The HMMER (Robert et al. 2011) software suite provides all other functionality to create an HMM model of drive protein sequences, given aligned known homologous genes. HMMER creates a profile of the aligned sequences that assigns a position-specific probabilistic models called "profile hidden Markov models" (profile HMMs) (Krogh et al., 1994). They store position-specific information about how conserved each column of the alignment is, and which residues are likely. Hmmsearch also search the model against a protein database which we have previously made using TransDecoder.

The following command is used to build an HMM profile

by using aligned sequences.

```
hmmbuild dreb_gmax.hmm  dreb_gmax.phylip
```

The following command is used to search an HMM against protein fasta database.

```
hmmsearch  dreb_gmax_cut_3.hmm
longest_orfs.pep  > Result.out
```

Results

The Tuxedo protocol over RNASeq data generates 12,720 transcripts. Total Cleaned reads and their alignment percentage is listed on table 1. These transcripts are taken as input for TransDecoder, which finds longest ORF (Open Reading Frame) and convert it into a FASTA file of protein sequences. Almost, 99% of all generated transcripts were converted to putative proteins based in ORF identified. For generating HMM pattern searching model, a multiple aligned sequence of DREB proteins obtained from ClustalW (Figure 1) were supplied to HMMBUILD tool.

HMMBUILD creates a mathematical model on the basis of alignment and stores it in a file automatically which can be searched over protein sequences to find a similar conservation pattern depicted in figure 1. HMM LOGO is a graphical image help to illustrate the HMM model (Figure 2). Hmmsearch finds the Pattern stored in HMM model and retrieve all the significant hits (E value < 0.001). Table 2 represents the output of Hmmsearch. All the parameters of table can be found at HMMER manual (http://hmmer.org/documentation.html). We have found 21 genes which is transcribed in control and infected condition of *Glycine Max*. Coordinates on genome can be found on a ninth column of table 2.

Table 1. Total number of sequences obtained for *Glycine max* in different accession number which were filtered and aligned. The alignment % is obtained from Tuxedo protocol

Accession No.	No. of sequences before cleaning	No. of sequences after quality trimming	Alignment %
SRR3090710	2780593	2588977	91.02
SRR3090711	3258771	3065956	92.00
SRR3090712	3573224	3377960	90.06
SRR3090713	3379495	3192642	90.12

Table 2. The search result for DREP homologous genes from HMM profile, last column of the table represents the coordinates of gene found homologous in *Glycine max*.

---- Full sequence ----			---- best 1 domain ---			--#dom--		
E-Value	Score	Bias	E-Value	Score	Bias	exp	N	Sequence
--------	------	-----	--------	------	----	----	--	---------
8e-77	259.1	0.0	9.6e-77	258.9	0.0	1.1	1	Gene.10368::4_9651973_9653652
7.3e-71	239.6	0.0	9.5e-71	239.2	0.0	1.1	1	Gene.11986::6_8383458_8384781
2.3e-66	224.8	5.2	1.3e-65	222.3	4.1	1.8	2	Gene.3571::13_38952352_38953872
2.3e-65	221.5	1.3	4.1e-65	220.6	1.1	1.4	2	Gene.2810::12_35894367_35895052
1.1e-42	147.0	0.0	1.2e-42	146.9	0.0	1.0	1	Gene.2079::11_4018075_4018578
3.3e-38	132.2	9.1	4.9e-38	131.7	9.1	1.3	1	Gene.1329::10_29589162_29590494
3.1e-37	129.1	0.4	3.8e-37	128.8	0.4	1.1	1	Gene.5571::16_2278691_2279335
1.1e-36	127.2	0.7	1.4e-36	126.9	0.7	1.1	1	Gene.12737::1_4794652_4795334
4.1e-36	125.3	0.8	5.9e-36	124.8	0.8	1.3	1	Gene.9025::20_23746135_23747170
7.5e-33	114.6	10.5	1.5e-32	113.7	10.5	1.6	1	Gene.12451::7_3661890_3665359
2.4e-32	113.0	10.0	4.4e-32	112.1	10.0	1.5	1	Gene.5321::16_1083430_1086860
4.4e-32	112.1	6.0	4.4e-32	112.1	6.0	1.6	1	Gene.9336::3_47664676_47666802
8.3e-31	107.9	2.1	1.2e-30	107.4	2.1	1.2	1	Gene.6979::18_9133067_9133951
5e-30	105.4	2.8	8.7e-30	104.6	2.8	1.3	1	Gene.8464::2_50328296_50329040
6.6e-28	98.4	6.8	8e-28	98.1	5.8	1.5	1	Gene.8458::2_48256678_48257607
2.2e-15	57.2	0.1	2.9e-15	56.8	0.1	1.2	1	Gene.8371::2_9219054_9220695
1.6e-14	54.4	0.1	1.9e-14	54.2	0.1	1.1	1	Gene.1363::10_42935024_42935923
2.5e-14	53.8	0.1	3.3e-14	53.4	0.1	1.2	1	Gene.513::1_27963731_27965260
4.3e-12	46.4	2.7	5.6e-12	46.1	2.7	1.2	1	Gene.10533:5_28389450_28390654
3.1e-05	23.9	0.2	3.4e-05	23.8	0.2	1.2	1	Gene.6549::18_39508163_39511857
0.00071	19.4	5.7	0.0014	18.5	5.7	1.5	1	Gene.10324::4_3819430_3820136

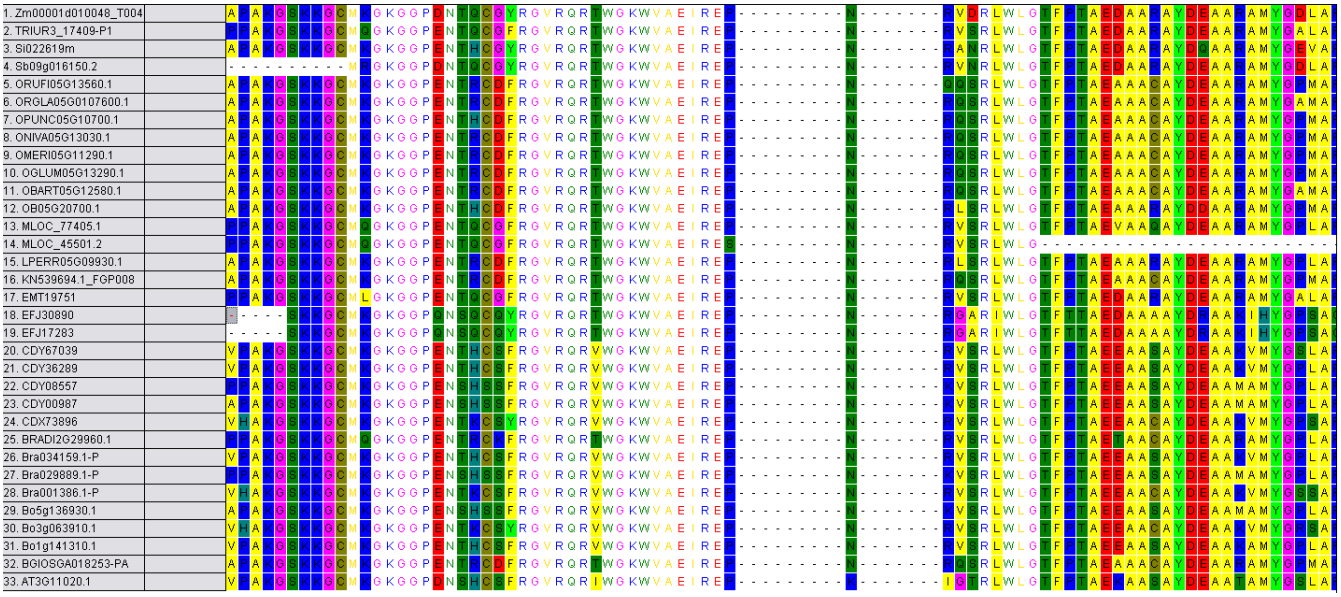

Figure 1. Blocks of conservation in multiple sequence alignment obtained for DREB protein after multiple sequence alignment by ClustalW

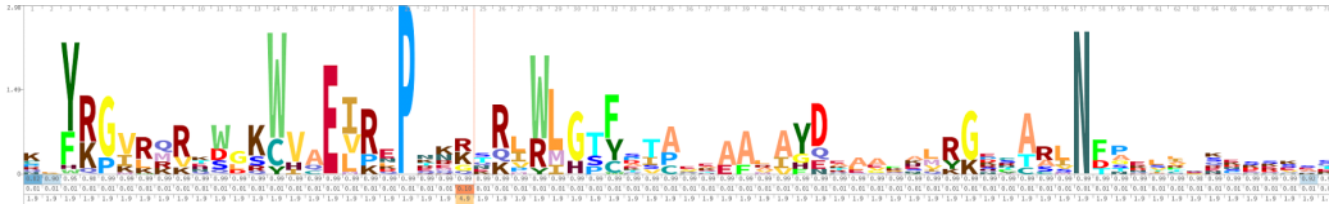

Figure 2. Size of symbols depicts the probability of occurrence of the amino acid at specific position

Discussion

According to evolutionary theory conserved pattern in DNA and proteins represent common ancestry. To find out these patterns we have employed a pipeline of open source tools. This pipeline is easy to implement by the biologist having minimal computational experience. RNASeq data provide detailed information about transcript sequences, which enable us to find all members of the gene family present in the transcriptome. HMM model generated by protein sequence alignment provides better result than DNA based analysis because of degeneracy in the genetic code. The extra advantage of using Tuxedo protocol is gene coordinates in output, which is automatically fetched by HMMSEARCH. The method is useful for researchers who are working on the evolutionary significance of gene duplication. It Provides functionality to find homologous genes in the newly sequenced plant genome. The method can be used as a single step in predicting protein-coding gene sequences with high accuracy. For more detailed annotation efforts, it offers an appropriate starting point for further refinement of annotations with additional supporting evidence.

Conclusion

High throughput sequencing technologies produce massive amount of genomic raw data. These datasets used in the wide range of applications including personalized medicine and plant breeding. However, to use these data effectively there is a need to develop algorithms, which can help us to decipher biologically relevant patterns in the data. In the present work we have developed a novel method for extracting of RNAseq data through hidden Markov models (HMMs) and introduce an HMM-based solution for finding gene family. Here we have taken DREB (dehydration responsive element binding) gene as a candidate for model building but this protocol can be applied for any gene with open source software application.

Acknowledgement

All authors are thankful to Bioinformatics Infrastructure Facility (BIF), University of Rajasthan, India, for providing infrastructure facility for data analysis and funding this research.

Conflicts of interest: Not declared.

References

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at:

http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15):2114–2120.

Chen H, Arsovski AA, Yu K, Wang A. 2016. Genome-Wide Investigation Using sRNA-Seq, Degradome-Seq and Transcriptome-Seq Reveals Regulatory Networks of microRNAs and Their Target Genes in Soybean during Soybean mosaic virus Infection. Plos One, 11(3): e0150582.

Krogh A, Brown M, Mian IS, Sjölander K and Haussler D. 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. Journal of Molecular Biology, 235:1501–1531.

Leinonen R, Sugawara H, Shumway M. 2011. International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Research, 39(Database issue):D19–D21.

Panchy N, Lehti-Shiu M, Shiu SH.2016. Evolution of Gene Duplication in Plants. Plant Physiology, 171(4): 2294-2316.

Robert D Finn, Jody Clements, Sean R Eddy. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research, 39 (suppl_2):W29–W37.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9):1105–1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L, 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3):562.

Yoon BJ. 2009. Hidden Markov Models and their Applications in Biological Sequence Analysis. Current Genomics 10(6):402–415.