A bioinformatics approach towards sequence analysis for revealing dark matters

Radhe Shyam Thakur and Rajib Bandopadhyay*
Department of Biotechnology, Birla Institute of Technology, Mesra, Ranchi 835 215, India

The biotechnological approach of introducing C4 plant genes into the C3 plants showed unpredicted behaviour in transgenic C3 plants. It was revealed that some noncoding parts of genome had a very important role in the complete functional gene expression. It was also found that these parts were evolutionary conserved along with the gene. So it becomes a prerequisite to prevail functional role of these ‘dark matters’ for complete functional manifestation of the genome sequence. Comparative genomics approach is used to functionally annotate the noncoding parts of the genome, while wet laboratory protocols need much time, chemicals and instrument facilities. Therefore, as an alternative, using bioinformatics approach, a workflow is designed from the preexisting tools, which is much more reliable and can be used to guide the wet lab experiments to validate and functionally annotate the dark matters of the genome.

Keywords: Annotation, bioinformatics pipeline, comparative genomics, dark matters, genome analysis

Introduction

Totipotent cells very well converted the two dimensional genomic code into multidimensional tissues, organs and finally organism. This property is inherited from generation to generation through germ line cells. According to central dogma of molecular biology, sequences are transcribed and then translated, i.e., protein, controlling the biological processes. The gene concept and further studies showed that, instead of whole of the genomic sequences, only coding sequences control the complete gene expression. This introduced the concept of coding and noncoding sequences. In a biotechnological approach, some C4 (plant) genes along with its promoters and enhancers were introduced in C3 plant for enhanced crop production. The introduced genes over expressed in the transgenic C3 plant. The advancement in sequencing and annotation techniques like single molecule sequencing (SMS) technology, Ion Torrent, and Miseq and Hiseq2000, especially RNA sequencing, resulted in piling up of huge sequence data and added a new dimension to the genomic world, suggesting that noncoding part too played key role in functional gene expression

Some portion of genomes are only transcribed and never translated into protein. These transcripts are known as noncoding RNAs or dark matter. It is also a well established fact that the functional part, even noncoding part (dark matters), is evolutionary conserved even during divergent evolution and controls the biological processes by different mechanism like RNAi (RNA interference). There are 574 different families of noncoding RNAs reported in Rfam. Few important among them are summarized in Table 1.

A detailed survey over the different (whole) genome databases shows that although the genomes are sequenced, we are ignorant with respect to the functional annotation. According to the present status of the model plant Arabidopsis thaliana, as reported by TAIR, almost 45% of the whole genome still remains unannotated. The annotation percentage to biological function is further reduced at cellular and molecular level. As we are still far away from deciphering the DNA language, it is a great challenge to functionally

<table>
<thead>
<tr>
<th>Name</th>
<th>Length (nt)</th>
<th>Function(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro RNA</td>
<td>20-24</td>
<td>Translation repression, mRNA degradation</td>
</tr>
<tr>
<td>siRNA</td>
<td>20-24</td>
<td>mRNA degradation</td>
</tr>
<tr>
<td>rasiRNA</td>
<td>20-29</td>
<td>Translational silencing via chromatin remodeling</td>
</tr>
<tr>
<td>piRNA</td>
<td>26-31</td>
<td>Epigenetic regulation, prevent retrotransposon, transposon</td>
</tr>
<tr>
<td>tasiRNA</td>
<td>21-23</td>
<td>Lateral root growth and development</td>
</tr>
<tr>
<td>snoRNA</td>
<td>80-100</td>
<td>Chemical modification of other RNA</td>
</tr>
</tbody>
</table>

*Author for correspondence:
Tel: +91-651-2276223; Fax: +91-651-2275401
Email: rajib_bandopadhyay@bitmesra.ac.in

ΨFor ‘Supplementary Matter’ see the paper at www.nopr.niscair.res.in or www.niscair.res.in
annotate the whole genome. Wet laboratory approaches have already been developed but they are still time consuming and need blind perception to proceed on. The recently developed approaches like TriAnnot, involving comparative genomics approach, have drawbacks of ignoring repeats. So it cannot be applied on larger genomes. Also, with the discovery of small noncoding RNAs like piRNA from the repeated regions of the genome, and keeping their importance in mind, a new approach is needed. In the present study authors are proposing a workflow using different bioinformatics tools for the in silico analysis of the genome for the functional annotation of the dark matters.

**Materials and Methods**

In order to search for dark matters, the target and the reference species were selected from the phylogenetic tree (Fig. 1) considering that all should be from different families. The target is shown by a red spot and the reference species are shown in blue spots.

The whole genome sequence and the coding sequences for chloroplast were downloaded in the fasta format from the Chloroplast Genome Database (http://chloroplast.cbio.psu.edu/). The tools TBA, lav2maf, maf2fasta, songle_cov2 were taken from Multiz tools package and is available freely from Miller laboratory website: http://www.bx.psu.edu/miller_lab. RNAz is a stable and conserved RNA predictor and was downloaded freely as RNAz package from http://www.tbi.univie.ac.at/~wash/RNAz/.

**Tools Deployed**

- **blastzWrapper**
  It processes two sequence files each containing one or more contigs and runs Blastz. Blastz is a redesigned tool based on gapped BLAST algorithm for pairwise local alignment of long sequences at genomic level. It searches for the almost perfect match between two sequences and then extends that fragment in both the direction using high score strings based on dynamic programming. It is highly specific in aligning constrained sequences and highly sensitive for sequences from a particular family as reported in human-mouse alignment. The default output format of the aligned sequences is lav format.

  The sequence header must be in proper format as described below for input:

  >string1:string2:int1char:int3

  Where, string1 is species name, string2 is chromosome name, int1 is start position, char is +/- denoting direction of reads, int3 is source size.

  The above format of input is not required if there is a single fasta sequence for the two aligning files.

- **single_cov2**
  This tool is used to parse out the overlapped regions of the aligned sequences. It accepts the input in maf format and output is too in maf format.

- **degapseq**
  It is a tool from Emboss package and used to remove all special characters and gaps from the sequences in fasta format.

- **TBA (Threaded Blockset Aligner)**
  The beauty of this tool is that it is a global local aligner. It searches for the best location of the blocks aligned in pairwise alignment. It assumes that the order and orientation of the block remain the same. Inversions and duplications are not considered. TBA is well adapted for aligning much mega base size of multiple genomes and can be projected onto any of the reference genome. This results in the best alignment positions for considering the evolutionary conservation among the different diverged sequences.

**Fig. 1—Dendogram representing the phylogenetic relation between whole genome already sequenced and being sequenced. [Adopted from CoGe database (The place to Compare Genome, Berkeley, USA); Available at: http://genomevolution.org/wiki/index.php/Sequenced_plant_genomes#Phylogenetic_Tree; Recently, Tomato genome sequenced has been published in Nature, 31st May, 2012.] Different species from different families are selected (depicted in blue spot) for evolutionary analysis of a target (depicted in red spot) to reveal dark matters.**
RNAz

This tool combines comparative sequence analysis and secondary structure prediction for the detection of the functional RNA. It uses following two basic steps:

1. The consensus secondary structure is computed and sequence conservation is calculated. A consensus minimum free energy (MFE) is calculated for alignment by averaging the energy contribution from individual sequence and covariance term for every mutation.

2. The thermodynamic stability is checked using the z-score, which is normalized using both sequence length and base composition. It can be calculated without sampling from the shuffled sequences. The z-score is calculated by generating synthetic sequences of different length and base composition. Then, using the support vector machine (SVM), regression analysis is done.

Rav2maf

It converts lav formatted aligned files into maf format which could be further analyzed.

maf2fasta

It converts a maf format file in fasta format according to reference genome passes as argument.

Workflow

From the phylogenetic tree several diverged species were selected from the different families; the diverged species ensure the analysis of real dark matters. With target and selected diverged species, the coding sequence (CDS) were subtracted from the whole genome sequence to get the noncoding sequences (NCS). Then the NCS of the selected species were locally aligned pairwise with target to get the conserved sequences of the target with respect to different species. Further, these pairwise aligned sequences were multiple aligned to form a triplet aligned file, i.e., target, species-1 and species-2 block. Here a triplet block was considered to select the real dark matters. It is possible that too close species may contain no real dark matters, i.e., unannotated part, that really do not have any function and have descended to next generation, or too diverged species would have lost the dark matter during evolution. A multiple alignment which searches for the local block and fits it at the best score was used to maintain the functional integrity of dark matter with the related gene. Then, using transcriptome modeling, structurally stable transcripts were predicted, which could be validated using the wet laboratory.

In order to execute this lengthy and error prone workflow smoothly and to restrict the error, whole program was divided into 3 phases and codes were written to execute it in a phase wise single command. The 3 phases are as follows:

**Phase I**: This phase took two sequence files as arguments and it subtracted the second sequence file from the first file. First, it locally aligned the two files and the matched pairs were extracted out. In this case, files containing chloroplast whole genome and CDS were given as arguments to extract the NCS.

**Phase II**: This phase took three sequence files as arguments. Considering first argument as reference, it locally aligned with other two files and then multiply aligns of the aligned files to give the conserved parts. The NCS of target and selected species were given as arguments and the code was executed to give CNS.

**Phase III**: This phase took multiple aligned blocks as argument. It predicted thermodynamically stable and conserved RNA secondary structure from the multiple aligned blocks. In this case the triplet blocks were taken into account.

The complete workflow is summarized in Fig. 2 and the scripts for all the phases are given as the supplementary material S1.

The target (*Arabidopsis*) and all selected species (Grape, Poplar and Rice) were analyzed with the Phase I to get the NCSs. Phase II aligned the NCSs derived from Phase I in triplet block with all the possible valid combinations considering *Arabidopsis* as reference. The valid combinations were *Arabidopsis*, Poplar and Vitiferous (APV); *Arabidopsis*, Poplar and *Oryza* (APO); and *Arabidopsis*, Vitiferous and *Oryza* (AVO). Using the Phase III thermodynamics, stable and conserved RNA were predicted on all the three triplet blocks, i.e., APO, APV and AVO.

Results and Discussion

Out of 1121 loci predicted in the genomic sequences producing non-coding RNAs, 163 were found having Sequence Conservation Index (SCI) more than 1 with respect to the selected diverged species, showing they are fully conserved during the evolution and they might have some role in gene expression. Among 163 predicted loci, 45 were found having z-score less than −2, showing they have a
stable secondary structure and have thermodynamic existence. All (163) predictions have a Support Vector Machine (SVM) Probability more than 9. So these 45 loci are the most potential targets for the gene responsible for the production of noncoding RNA, which really have a biological meaning as they are proved to be evolutionary conserved as represented by SCI and have a stable secondary structure as represented by z score (Supplementary Table S1). The whole result is represented in the form of pie chart in Fig. 3.

The proposed approach gave the most potential target from the genomic sequences, which were conserved and produced stable RNA. This approach has major advancements over the previous approaches used for annotating the genome. The following described constraints were eliminated in this approach:

1. The selected and number of the species for generating phylogenetic footprints were fixed in previous approaches. For example, four species from the Triticale genome were annotated using TriAnnot. The present approach eliminated this constrain by dividing the whole program in phasewise scripts. Species for studies are passed as arguments and can be iterated as many times user prefers. Thus, results accuracy can be enhanced.

2. The repeats in the genomic sequences were ignored as they were thought they have no function. But with the discovery of piRNA from the repeats, it became a necessity to consider it. The present approach took all the repeats for the footprints and finally predicts or annotates the function.

3. The training datasets formed the base for functional prediction. In case, if no training dataset was properly matched, the tools overwhelmed and predicted inappropriate results. The present approach is a de novo technique eliminating the constraint.

4. The present approach is assembly of different phase wise scripts. This makes the program highly portable, reusable, debug and easy to modify according to the new developments in bioinformatic tools and sequence information.
Conclusion

Although present approach is quite slow in eliminating the negative points due to its de novo nature and it still needed validation from the wet laboratory approach, dark matters seems to be revealed by the combined computational and wet laboratory approaches. However, the question that does the location of the genomic sequences has any role in its expression still remains unanswered and needs to be further considered and analyzed in order to achieve a blueprint of the genome and complete understanding the biological mechanisms. Other important approach is also cloud computing, which has been used very recently to solve different biological problems and may be used to reveal dark matters from huge number of genome sequences.

Acknowledgement

Authors gratefully acknowledge BTISNet SubDIC (BT/BI/065/2004) for providing internet facilities. We are also thankful to Professors S Madhekar and Mr Mannmohan Gupta for providing computational facilities at the Department of Applied Physics, Birla Institute of Technology (BIT), Mesra. RB is thankful to BIT for getting Cumulative Professional Development Grant (CPDG; Ref No. GO/PD/2011-12/269/3523; dated August 04, 2011) during preparation of this manuscript.

References

16 http://www.arabidopsis.org/portals/genAnnotation-genome_snapshot.jsp