GENOME ANNOTATION-STATE OF THE ART

RAHUL BANIK* AND SAYAK GANGULI

DBT Centre for Bioinformatics, Presidency University, Kolkata

Genome annotation is of enormous significance in interpretation of the necessary information obtained from raw sequence data produced by genome sequencing projects. The process aids in identifying the biological significance of raw sequence data and thus putting our understanding of biological processes in proper context. There are two interrelated types of genomic annotation: structural and functional. Structural annotation deals with identified the genome elements (such as genes, promoters, and regulatory elements) whereas functional annotation assigns functions to these structural elements. Structural annotation is defined as finding genes in genomic DNA. Structural annotation is of two types: one, prediction based and another, sequence similarity based. Prediction based algorithms designed to find structures of gene(s) based on nucleotide sequence and composition whereas similarity based prediction is alignment with mRNA sequences (ESTs) from the same or related species to identification of motifs. In case of functional annotation, Gene Ontology (GO) plays an integral part. GO describes three attributes of gene product(s): molecular function, biological process and cellular component. The basic steps of functional annotation include BLAST, Mapping of GO terms, Annotation using an annotation rule, and finally statistical analysis of GO term distribution differences between groups of sequences. In this study, two unannotated sequences were used as samples for the analyses for functional annotation. After mapping the sequences, GO terms that were generated, describe the molecular function (F), biological process (P), and cellular component (C) e.g. DNA Binding (F), biosynthetic process (P), protein complex (C). Enzyme codes and KEGG pathway maps were generated for each sequence, which describe the different pathways like purine and pyrimidine metabolism, pentose phosphate pathway etc. Another important aspect is evidence code distribution i.e. the quality of annotation. Here, IEA (Inferred from electronic assay) is dominant.

Genome sequence of an organism is an important resource of information for all aspects of biological research. But only the sequence of genome is nothing without its proper annotation. The proper annotation pathway or pipeline is required for identify the key features of genes in the genome and their biological functions. (Conesa *et al* 2008)

An annotation is any comment or note (collectively known as metadata) that is attached to the data and describe how, when or where the data were collected. The recent advancement of next generation sequencing technologies helps us to sequence DNA and RNA much more rapidly, thus the amount of raw sequence data deposited in the sequence databases very large volume within a short period of time. Hence, the newly generated large volume of raw sequence data needs to annotate in proper way and speed i.e. not only manually but also automated. (Conesa *et al* 2008; Conesa *et al* 2005).

In general, genome annotation can be classified in two ways. One is structural annotation and another one is functional annotation. Structural annotation deals with identified the genome elements (such as genes, promoters, and regulatory elements) whereas functional annotation assigns function to these structural elements. Structural annotation can be done

*Corresponding author: ***E-mail:*** *rbanik2009@gmail.com* by two ways. One is homology based method and another is ab-initio method. Similarity based prediction is alignment with mRNA sequences (ESTs) from the same or related species to identification of motifs whereas ab-initio prediction algorithm is designed to find structure of gene(s) use compositional features of the DNA sequence to define coding segments (exons). (Conesa *et al*, 2005; Nagraj *et al* 2007)

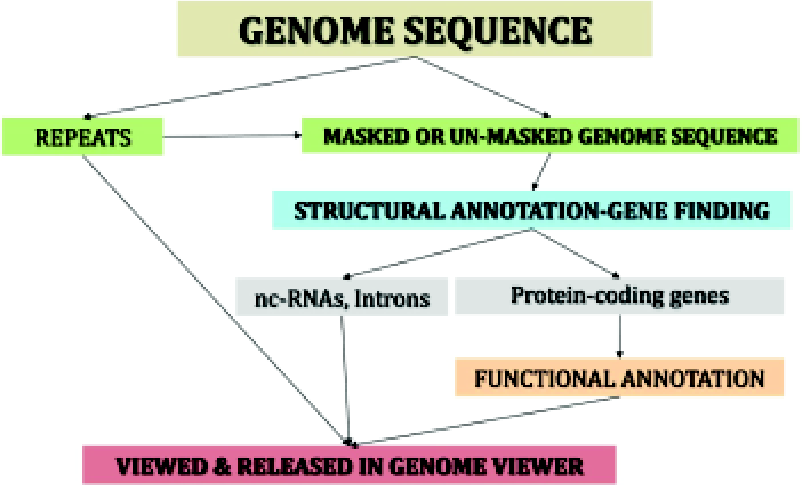
GENOME SEQUENCE
REPEATS
MASKED OR UN-MASKED GENOME SEQUENCE
STRUCTURAL ANNOTATION-GENE FINDING
nc-RNAs, Introns
Protein-coding genes
FUNCTIONAL ANNOTATION
VIEWED & RELEASED IN GENOME VIEWER

Fig 1: ***Schematic Representation of the Genome Annotation Pipeline.***

Functional annotation allows categorization of genes in functional classes; molecular function, biological process and cellular component. Gene ontology (GO) developed by GO consortium play an important role for functional annotation of genes.

MATERIALS AND METHODS

The basic steps of functional annotation is BLAST the unannotated sequences, mapping of GO terms, annotation of GO terms and quality evaluation of annotation data. In this study, an unannotated sequence was used as sample for the analysis of function annotation.

Step 1: The first step is to find the sequence similarity by BLAST searching. Homology search was performed by using NCBI nr database. BLAST expectation value (E-value) and hit number thresholds were provide to get significant results. Depending upon the best BLAST hits the next step i.e. GO mapping is performed.

Step 2: In GO term mapping step using the best BLAST hit gene identifiers (gi) and gene accessions retrieves all GO annotation for the hit sequences, together with evidence code (EC) distribution. Evidence code is actually used as quality level of annotation.

Step 3: Annotation of GO terms was performed by applying an annotation rule (AR) to the obtained ontologies. The rule finds the most specific GO terms with a certain level of reliability.

Step 4: Finally, another level of more specific and accuracy of the annotation is done by applying Annex and GO Slim for further refinements.

RESULTS AND DISCUSSION

In this study, when we perform the BLAST on the unannotated sequence against NCBI nr database, the BLAST result information is useful for choosing the annotation cutoff parameters at the annotation step. Furthermore, the species distribution shows a great majority of Anaplasma sequence within the BLAST hits followed by Wolbachia species.

Depending upon the best BLAST hits the GO mapping was performed. Total seven GO terms were generated. Five terms describes the molecular function and two others describe the biological process. Evidence code distribution refers to IEA (Inferred from Electronic Assay) is dominant rather than any other evidence code. Mapping database using here UNIPROTKB.

Then the annotation result gives us more specific GO terms. After that the further level of annotation refinement using Annex and GO Slim more specific and summarized form of GO annotation terms is obtained.

CONCLUSION

Automated genome annotation, has progressed a long way within a short period of time, as the number of sequenced genomes increased exponentially. So genome annotation processes are immensely important for identification of novel genes and biological function of those gene products.

REFERENCES

Conesa A. and Gotz S. (2008): “Blast2GO: A comprehensive suite for functional analysis in plant genomics”, International Journal of plant genomics, vol. **2008**, p.1-12.

Conesa A. and Gotz S., *et al.*, (2005): “Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research”, Bioinformatics, vol. **21**(18), p.3674.

Nagraj SH, Gasser RB, Ranganathan S., (2007): “A hitchhiker's guide to expressed sequence tag (EST) analysis”, Briefings in Bioinformatics. **8**(1):p.6–21.

FURTHER BIBLIOGRAPHIES

Andrade M A, *et al*. (1999): “Automated genome sequence analysis and annotation”, Bioinformatics., **15**, p.391–412.

Ashburner M., Ball CA, *et al*.(2000): “Gene Ontology: tool for the unification of biology”, Nature Genetics. vol. **25** (1), p.25-29.

Curwen V., and Eyras E, *et al.*, (1999): “The Ensembl automatic gene annotation system”, Genome Res. **14**, p.942–950.

Dolan M. E., *et al*.(2005): “A procedure for assessing GO annotation consistency”, Bioinformatics. **21**, p.136-143.

King O. D., *et al*.(2003): “Predicting gene function from patterns of annotation”, Genome Res.**13** p.896-904.

Lewis S. E. (2004): “Gene Ontology: looking backwards and forwards”, Genome Biol, **6**, p.103.

S. Myhre, and H. Tveit. (2006): “Additional gene ontology structure for improved biological reasoning”, Bioinformatics. vol. **22**(16), p.2020-2027.