Computational analysis of transcriptome of Indian major carp, Labeo rohita (Hamilton-Buchanan, 1822) for functional annotation

Naresh Sahebrao Nagpure, Iliyas Rashid, Ajey Kumar Pathak*, Mahender Singh, Shri Prakash Singh & Uttam Kumar Sarkar

National Bureau of Fish Genetic Resources, Canal Ring Road, P.O.- Dilkusha, Lucknow-226002, India; Ajey Kumar Pathak - Email: pathakajey@rediffmail.com; Phone: +91-522-2442440, 2442441, Fax: +91-522-2442403; *Corresponding author

Received September 21, 2012; Accepted October 01, 2012; Published October 31, 2012

Abstract: A total of 1671 ESTs of Labeo rohita were retrieved from dbEST database and analysed for functional annotation using various computational approaches. The result indicated 1387 non-redundant (184 contigs and 1203 singletons) putative transcripts with an average length of 542 bp. These 1387 transcript sequences were matched with Refseq_RNA, UniGene and SwissProt on high threshold cut-off for functional annotation along with help of gene ontology and SSRs markers. We developed extensive Perl programming based modules for processing all alignment files, comparing and extracting common hits from all files on a threshold, evaluating statistics for alignment results and assigning gene ontology terms. In this study, 92 putative transcripts predicted as orthologous genes and among those, 44 putative transcripts were annotated with gene ontology terms. The annotated orthologous gene of our result associated with some very important proteins of L. rohita involved in biotic and abiotic stresses and glucose metabolism of spermatogenic cells etc. The unidentified transcripts, if found important in expression profiling can be vital resource after re-sequencing. The predicted genes can further be used for enhancing productivity and controlling disease of L. rohita.

Keywords: Expressed sequences tag, Functional annotation, Gene, Labeo rohita, Putative transcripts, Transcriptome

Background: Indian major carp, Labeo rohita (Hamilton-Buchanan, 1822), commonly known as ‘rohu’ belongs to the family Cyprinidae is available in lakes, ponds, rivers and other water bodies of India and adjacent countries [1]. It is a highly preferred carp and fetches high market price. India is by far the largest producer of rohu, followed by Bangladesh and Myanmar. Although a considerable work has been carried out on various aspects including aquaculture, biology, seed production etc., however the species has not been adequately studied with respect to its genome, proteome and transcriptome that may provide critical information for enhancing its aquaculture productivity, growth, disease resistance and other traits. Identification and characterization of transcriptome of L. rohita can provide valuable information to undertake future research in functional genomics. Transcriptome explores the part of the genome that is functionally active in a particular tissue [2]. An expressed sequence tag (EST) is a fraction of a transcribed cDNA sequence [3] and important for identification of transcripts of a gene [4]. Large-scale EST data represents a snapshot of the transcriptome of an organism while cloning and sequencing of ESTs is also an effective approach for the recovery of full length cDNA, discovery of novel genes and molecular markers development [5]. ESTs being one of the important genomic resources, their numbers in public databases are rapidly increasing. Globally, 32,300 fish species have been...
described [6] that offer a unique opportunity for studies of transcriptomics, genomics and evolutionary biology in fishes. Despite, the global importance of fishes, information on transcriptome is available only for few species [7, 8]. Therefore, the utilization of available biological data reported across these species should be a new starting point for expanding studies in functional genomics.

In the light of above, we performed transcriptome analysis and identified differentially expressed genes in different tissues of L. rohita by utilizing available EST information in other fish species. Homologous transcripts, similarity with mRNA and translated product of these ESTs were predicted by the sequential application of various methodologies. Initially, we downloaded ESTs of L. rohita from dbEST database with complete information. We further developed an EST database using MySQL database management system on Linux platform. The ESTs sequences were refined by removing vector contamination segments and assembled into contigs and singletons. A contig was referred to tentative consensus sequence [9], while both contig and singleton were collectively considered as putative transcripts (PTs) [10], which ultimately represented subsequence of individual gene. We also analyzed PTs for repeat motifs and stored them into database along with ESTs sequences. This unique set of sequences were used as query sequence for analysis such as repetitive motif (microsatellites), RNA search, and homolog transcript search with zebrafish dataset and prediction of hypothetical protein and functional annotation. A total of 92 transcripts of L. rohita were putatively and functionally annotated using similarity searches, SSRs mapping and establishment of gene ontology [11] relationship.

Methodology:
Database design and development
A database 'lrest' was designed in MySQL on Linux to manage the downloaded EST data in GB and FASTA files. The data was then parsed through Perl parser according to schema and managed into database comprised of three tables: 'fishinfo' for physical information of fish, 'estinfo' holding the records related to ESTs information. The database also holds the results of the analysis in the third table 'analysis'. Relationships among the tables were created based on primary and foreign keys. The diagrammatic representation of data organization and other details have been explained in (Figure 1).

Primary analysis
ESTs collected from dbEST, NCBI in GB (for annotation) and FASTA (for sequences) format (February 2012) [12] were used for the analysis. For the downloaded EST sequences were cleaned by using SeqClean program [13] and UniVec database [14]. The resultant quality ESTs were assembled through CAP3 [15] assembly program which generated unique gene sequences. These unique sequences were used as query sequences for alignment analysis. Further, the singletons and contigs were analyzed to identify simple sequence repeat (SSR) motifs with the help of MicroSatellite identification tool (MISA) [16], a publicly available Perl script for SSR search within nucleotide sequences. This program produces types (simple perfect, compound perfect and imperfect), repeat unit length (mono- to hexa-nucleotide) and length and repeat sequence class of microsatellites. The result of Cap3 program related to contig and singleton and results of the MISA program were managed into the ‘analysis’ table of the database. We established relationship of the ‘analysis’ table with other tables through primary and foreign keys of database (Figure 1).

![Figure 1: The dataflow diagram of 'lrest' database comprising with three tables 'fishinfo', 'estinfo' and 'analysis'.](image)

Similarity search
The Refseq_RNA [17] and Swiss-Prot [18] databases provide more comprehensive, integrated, non-redundant and well-annotated set of sequences for similarity search. These two databases along with UniGene dataset of zebrafish were downloaded from ftp server of NCBI (ftp.ncbi.nih.gov). The annotation of 1387 PTs was performed on the basis of significant hits using local BLAST [19] against Refseq_RNA and UniGene database using BLASTn and Swiss-Prot protein databases using BLASTx. The Nucleotide search against Refseq_RNA and UniGene databases were obtained on high threshold score > 250 and identities > 85%, while putative transcripts were compared against the protein database using BLASTx with significant matches at a set threshold of score > 100 and identities > 70%.

Analysis programs
For the detection of number of clustered sequences present in different contigs the CAP3 assembly files (.cap.ace) were analyzed using a Perl script (CAP3ResultParser.pl). We wrote another Perl script (TissueStatsClust.pl) program for manipulation of total number of PTs and ESTs for individual tissues. A third program ‘AlignmentAnalysis’ was also written in Perl to generate the information about targeted databases and detail statistics about alignments by using ‘lrest’ database in the backend and alignment files. Initially the program took alignment files generated by BLASTn and BLASTx program for PTs query against UniGene set of zebrafish and Refseq_RNA database and Swiss-Prot database respectively. In its first step, the program separated the hits and non hits PTs sequences against corresponding target database and subsequently the program extracted only significant hits on the threshold Score >
250 and Identities > 85% for zebrafish and Refseq RNA and alignment Score > 100 and Identities > 70% for Swiss-Prot. These significant results for all the databases were further stored in separate files. The program retrieved pre-computed SSRs from 'Irest' database and mapped same motif patterns into aligned sequences of zebrafish with corresponding PTs for providing strength to our putative annotation results. The program 'AlignmentAnalysis' used all the three alignments files and subsequently merged for PTs through corresponding ids, serving as primary key into table 'analysis' of 'Irest' database. Average length of PTs and length range of PTs were also processed by this program. We also downloaded two files 'gene2refseq' and 'gene2go' from Entrez Gene[20] for assigning the gene ontology terms like biological process, cellular component and molecular function to the putative transcript in functional annotation process. A program 'GOAssign.pl' written in Perl was used for assigning GeneID and GO terms to the annotation results with computational mapping of both files 'gene2refseq' and 'gene2go'(Figure 2).

Figure 2: Flowchart for our method explaining functional annotation of transcriptome.
Results and Discussion:
Clustering and SSRs
The EST data were processed and managed into a database using MySQL on Linux platform. All cleaned EST sequences were assembled with CAP3 software that resulted in 184 contigs and 1203 singletons i.e. 1387 unique sequences putatively corresponding to as many genes with an average length of 542 bp. These 1387 PTs were used to identify SSRs motifs using MISA program and derived SSRs ranging in length from 2 to 6 bp were considered as repeat motifs. The minimum numbers of repeats were 6 for dinucleotides, 5 for trinucleotides, and 10 for tetranucleotides. In SSR analysis, observation was made on their types (di to tetra nucleotides), number of repeats, and percentage frequency of occurrences and their distribution in the sequence. The results of unique sequences (PTs) along with all SSR information was parsed and manipulated using Perl parser method according to database schema and data were managed into a separate table 'analysis' of 'lrest' database.

The data of table 'analysis' were used for further application in transcriptome analysis study. After analyzing the ESTs, we found majority of them were from brain (1317) followed by liver (236), spleen (1) and 40 ESTs from unknown tissue. The organ wise distribution of contigs, singletons and clustered ESTs were determined through Perl script 'TissueStasClust.pl' that uses computational mapping of accession id and GenBank annotation stored in 'lrest'. The contigs derived from more than one organ were not considered in this statistics obtained by the above program Table 1 (see supplementary material). Only 1625 ESTs were found distributed into 168 contigs derived from single organ and 1203 singletons.

Putative annotation
The annotation of the unique 1387 PTs was done on the basis of the best match data with Refseq_RNA database using BLASTn.

Only 439 (31.66%) sequences of PTs showed alignment with 2406 sequence of Refseq_RNA database and remaining PTs showed no hits. Out of 2406 alignments with Refseq_RNA, 664 alignments were obtained for 256 PTs on threshold with identities > 85% and score > 250 that could be considered as highly significant homologs. The score > 250 was used so that the alignment coverage reached 50% of query length to the target, while identities > 85% provided highly conserved and homologous sequences. Figure 3 describes the result of above alignment generated by 'AlignmentAnalysis.pl' which used blast alignment file for parsing and calculating the statistics of alignment between query (putative transcript) and target sequences (selected database). Similarly, 1387 PTs sequences of L. rohita were aligned with D. rerio dataset of UniGene collections using a BLASTn with score > 250 and sequence identities > 85%. A total of 455 (32.81%) PTs overlapped, while remaining PTs could not align with any of the transcripts of D. rerio. Out of 701 alignments with D. rerio sequences, 332 alignments were obtained for 258 PTs sequences on above threshold values. The aligned sequences on above threshold parameters were considered as highly significant homologs. Figure 4 describes the result of above alignment generated by our own script 'AlignmentAnalysis.pl'. The pre-computed SSRs motifs stored in the database were used to map these motifs as a marker of D. rerio homologous sequences corresponding to the PTs. In this map we found a strong support to annotate the gene of L. rohita. Identification of repetitive elements and nucleotide polymorphisms in query as well as target sequences provided strong support for annotation.

Hypothetical protein analysis
Swiss-Prot databases were used for hypothetical protein prediction for 1387 transcripts query with the help of BLASTx and program aligned 1102 PTs hits with 36517 protein
sequences with alignment percentage 79.46%. Further, the set threshold values i.e. score >100 and identities >70% for SwissProt resulted in alignment of 145 PTs with 1900 protein sequences (Figure 5).

Figure 5: The output of ‘AlignmentAnalysis.pl’ showing details on transcripts alignment statistics with Swiss-Prot database along with results of functional annotation.

Comparative analysis and functional annotation

Alignment of ESTs with Refseq RNA, D. rerio transcripts and Swiss-Prot resulted in 256 PTs, 258 PTs and 145 PTs respectively as significant hits in the similarity search analyses. The pre computed and stored SSRs of related PTs sequences were mapped with the aligned transcripts of zebrafish and the presence of SSRs motifs in corresponding target transcripts provided strength to this annotation. All significant alignments were merged with their PT ids and found that only 92 PTs were common hit for all the three databases and referred to as annotated on the basis of gene and protein prediction and mentioned into the supplementary ‘S1.xls’available with us. Functional categorization of predicted protein sequences for 92 PTs was done through gene ontology terms like biological process, cellular component and molecular function. The Program ‘GOAssign.pl’ efficiently mapped gene ontology terms for 44 PTs sequences and result about gene ontology has been mentioned into the supplementary file ‘S2.xls’ available with us.

Our final result based on putative and functional annotation of transcriptome of L. rohita for 92 PTs was executed and the analysis was supported by several approaches for transcriptome and proteome based similarity searches as well as computational mapping of SSRs and gene ontology. The annotated results were explained in the supplementary file S1. We identified several genes and proteins along with function, process and location based on high identities (85-94%) and matches (70-95%) in the similarity search, e.g. Contig93 (H0758335; H0758369; H0758347) as protein L40-1, Contig120 (GR977092; H0758682) as L29, Single1101 (H0762402) as L5a, Single1116 (JK492853) as L28-like and L32, Single290 (GR958016) as L17, Single478 (GR977096) as L36a, Single595 (H0758410) as L38, Single965 (H0762207) as L35, Single626 (H0758770) as S6 ribosomal protein or transcript variant 1, Single79 (GR426887) as ribosomal protein S16, Single814 (H0758752) as ribosomal protein S33. Additionally few important orthologous genes viz. Single288 (GR958033), Single1001 (H0762255), Single886 (H0758554) on high identities >92% were also annotated as heat shock protein (HSPs) associated with cell stress and cell growth [21]. Similarly, we identified protein specific for biotic stress viz. Single311 (GR958037) as thioredoxin-like 1, Single777 (H0758701) as F-box protein 11a (fbx11a), Single315 (GR958041) as CoA synthetase family member 2 (acs2), Single1046 (H0762321) as 3-monoxygenase/tryptophan 5-monoxygenase activation protein, gamma polypeptide 2 (ywhag2) and Contig92 (H0762389; H0762458) as high-mobility group box 30.

Likewise annotation of some PTs lead to identification of important proteins such as Contig184 (H0758515; H0758524 H0758557; H0758384; H0758383) as Epденmyin, Single985 (H0762234) as coiled-coil domain-containing protein, Single786 (H0758714) as fibroblast growth factor 8/ androgen-induced growth factor, Single1169 (K492845) as ba2 globin and ba1 globin, Contig105 (H0762218; H0762362; JK492857; JK492849; H0758519; GR958118) as hemoglobin alpha adult-1, Contig163 (H0758295 H0762456 GR463879) as fatty acid-binding protein, Single780 (H0758705) as PKR-interacting protein 1, Contig92 (H0762389 H0762458) as high-mobility group. The Contig68 (H0758673; H0762288) showed 529 matches on identities 91% to D. rerio NADH dehydrogenase (ubiquinone) Fe-S protein 4. The NADH dehydrogenase is associated with the electron transport chain reaction and involved in the glucose metabolism of spermatogenic cells.

Conclusion:

Out of 1387 putative transcripts, 256 PTs showed alignment with important proteins such as Contig184 (H0758515; H0758524 H0758557; H0758384; H0758383) as Epденmyin, Single985 (H0762234) as coiled-coil domain-containing protein, Single786 (H0758714) as fibroblast growth factor 8/ androgen-induced growth factor, Single1169 (K492845) as ba2 globin and ba1 globin, Contig105 (H0762218; H0762362; JK492857; JK492849; H0758519; GR958118) as hemoglobin alpha adult-1, Contig163 (H0758295 H0762456 GR463879) as fatty acid-binding protein, Single780 (H0758705) as PKR-interacting protein 1, Contig92 (H0762389 H0762458) as high-mobility group. The Contig68 (H0758673; H0762288) showed 529 matches on identities 91% to D. rerio NADH dehydrogenase (ubiquinone) Fe-S protein 4. The NADH dehydrogenase is associated with the electron transport chain reaction and involved in the glucose metabolism of spermatogenic cells.

Acknowledgement:

Authors are grateful to National Agricultural Bioinformatics Grid (NABG), under National Agricultural Innovation Project (NAIP), ICAR, New Delhi for providing financial support and the Director, NBFGR, Lucknow for providing necessary facilities and guidance.

References:

[1] Talwar PK & Jhingran AG, Inland Fishes of India and Adjacent Countries. Oxford & IBH Publishing Co. Pvt. Ltd. 1991; 219

[2] Veclucescu VE et al. Cel. 1997 88: 243 [PMID: 9008165]
[3] Adams MD et al. Science. 1991 252: 1651 [PMID: 2047873]
[4] Kim JM et al. DNA Res. 2006 13: 275 [PMID: 17213182]
[5] Manickavelu A et al. DNA Res. 2010 17: 211 [PMID: 20360266]
[6] http://www.fishbase.org/search.php
[7] Li P et al. BMC Genomics. 2007 8: 177 [PMID: 17577415]
[8] Christoffels A et al. BMC Bioinformatics. 2006 7: 52 [PMID: 17254304]
[9] Fei Z et al. Plant Journal. 2004 40: 47 [PMID: 15361140]
[10] Xia JH et al. DNA Res. 2001 8: 513 [PMID: 12086997]
[11] Harris MA et al. Nucleic Acids Res. 2004 32: D258 [PMID: 14681407]
[12] Benson DA et al. Nucleic Acids Res. 2012 40: D48 [PMID: 22144687]
[13] http://compbio.dfci.harvard.edu/cgi/software/
[14] ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/
[15] Huang X & Madan A. Genome Res. 1999 9: 868 [PMID: 10508846]
[16] http://pgrc.ipk-gatersleben.de/misa/misa.html
[17] Sayers EW et al. Nucleic Acids Res. 2012 40: D13 [PMID: 22140104]
[18] Bairoch A & Apweiler R. Nucleic Acids Res. 2000 28: 45 [PMID: 10592178]
[19] Altschul SF et al. J Mol Biol. 1990 215: 403 [PMID: 2231712]
[20] Maglott D et al. Nucleic Acids Res. 2011 39: D52 [PMID: 21115458]
[21] Knowlton AA & Salfity MJ, Biosci. 1996 21: 123

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Edited by P Kangeane

Citation: Nagpure et al. Bioinformation 8(21): 1005-1011 (2012)
### Supplementary material:

**Table 1: Organ wise distribution of contigs, singletons and total number of ESTs**

| Tissues | Contigs | EST Clusters | Singletons | Total Accessions |
|---------|---------|--------------|------------|------------------|
| Brain   | 150     | 366          | 926        | 1292             |
| Testes  | 18      | 56           | 20         | 76               |
| Liver   | -       | -            | 221        | 221              |
| Spleen  | -       | -            | 1          | 1                |
| Unknown | -       | -            | 35         | 35               |
| Sum     | 168     | 1203         | 1625       |                  |

© 2012 Biomedical Informatics