3.1 MATERIALS

3.1.1 Datasets

This study is based on identification of long and short noncoding RNAs through comparative genomics approach. Three different datasets were used for the identification of noncoding RNA analysis.

(1) Expressed sequence tags (EST)

(2) Genome Survey Sequences (GSS)

(3) Whole Genome shotgun contigs (WGS).

3.1.1.1 Dataset I

For the identification of long and short noncoding RNAs, full length of cDNA sequences were retrieved from NCBI dbEST database. A total of 20,186 redundant sequences were retrieved from dbEST database and stored for the present study. A lot of redundancy was found among the sequences. Redundancy and complexity are removed through EGassembler. After removing contamination, a non-redundent data set was developed for the identification of novel noncoding RNAs (long and short both) in C. roseus. NCBI dbEST database is used as retrieval system for finding EST data (http://www.ncbi.nlm.nih.gov/dbEST_summary.html). Expressed sequence tags (ESTs) are the most commonly nucleotide-sequence product from the plant genomes (Rudd, 2003). ESTs submitted to all three international sequence databases (GenBank, EMBL and DDBJ), can be accessed. The same ESTs are also available from the NCBI’s dbEST, the database of Expressed Sequence Tags (Baxevanis et al., 2001). The ESTs database has the maximum set of contamination related with them 1.63%, which was removed through EGassembler before further processing (Falgueras et al., 2010). EGassembler software
has been used to trim out the redundant sequences and impurities (Johnson et al., 2008).

3.1.1.2 Dataset II

A total of 115 GSS redundant sequences were retrieved from Genbank nucleotide database and a lot of redundancy was found among the sequences. Redundancy and complexity were removed through EGassembler. After removing contamination, a non-redundent data set was developed for the identification of novel noncoding RNAs in C. roseus. The GSS division of GenBank is analogous to the EST division where most of the sequences are genomic in origin rather than cDNA (mRNA). The data type of GSS division is random "single pass read", genome survey sequences, cosmid/BAC/YAC end sequences, exon trapped genomic sequences, Alu PCR sequences and transposon-tagged sequences (Johnson et al., 2008).

3.1.1.3 Dataset III

A complete set of 79,302 whole genome shotgun contigs were retrieved from NCBI and developed a local library for finding noncoding RNAs. Whole Genome Shotgun projects are based on genome assemblies of partial genomes or incomplete chromosomes of prokaryotes and eukaryotes that are usually being sequenced by a whole genome shotgun strategy. It was submitted on 25 August 2014 to Department of Plant Biology, Michigan State University, 612 Wilson Road, Room 166, East Lansing, MI 48824, USA. The Assembly database also tracks the association between an assembly submitted to the International Nucleotide Sequence Database Consortium (INSDC) and the assembly represented in the NCBI RefSeq project. It allows users easy access to download sequences and annotations for current versions of genome assemblies from the NCBI genomes FTP site (Kitts et al., 2015).
3.1.2 Reference dataset of noncoding RNA

For the identification of short noncoding RNAs, 324 non-redundant miRNA sequences of *Arabidopsis thaliana* (model plant) were retrieved from miRBase (microRNA database) (Kozomara and Griffiths-Jones, 2011). Redundancy was manually removed among the miRNA sequences on the basis of sequence similarity and developed a non-redundant data set. There were no redundancy found among the sequences. Using these miRNAs as a reference dataset of miRNA sequences, a non-redundent data set was developed for the identification of novel miRNAs in *C. roseus*.

3.1.2.1 MiRBase (http://www.mirbase.org/)

MiRBase is the main online repository for all microRNA sequences. It provides whole evidence for microRNA annotations and it resumes previous annotations. Deep-sequencing technologies have been used for the discovery of novel microRNA. These data can be use as a substitute for relative expression levels of microRNA sequences. It provides detailed facts for microRNA observations and alternative isoforms of mature microRNAs and allow us to revisit previous annotations (Griffiths-Jones *et al*., 2008).

3.1.3 Bioinformatics tools and web servers for noncoding RNA identification

In the present study, the following tools and web servers were used.

3.1.3.1 EGassembler (http://www.genome.jp/tools/egassembler)

EGassembler is used for clustering & assembly of the sequences. The sequence redundancy were removed through EGassembler program. After removal of the redundancy coding (contigs) & noncoding (singletons) regions were identified. Using EGassembler, different processes like Sequence cleaning, Repeat Masking, VectorMasking, Organelle Masking and Sequence Assembly were performed. PolyA/PolyT tail filtering, Low complexity filtering, Low quality filtering, short sequence
were removed through sequence cleaning process. Microsatellites, LINEs, SINEs, LTR, RNA pseudogenes, other Interspersed repeats (Masking against Rep Base, TIGR, TREP, user’s database) were removed through repeat masking process. Vector adaptor & other contamination (Masking against, NCBI’s UniVec, EMBL emvec, User’s database) were removed through Vector masking process. Plastids, Mitrochontria, plasmids, Nucleomorph (Masking against NCBI’s organelle database) were removed through Organelle masking process In the end, sequence assembly process differentiated sequences in to Contigs & Singletones (Masoudi-Nejad et al., 2006).

3.1.3.2 ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/)

A DNA sequence has six potential translation frames, three in the forward (+) direction and three in the reverse (-) direction (Dwivedi and Mishra, 2012). The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The ORF Finder is helpful in preparing whole and accurate sequence submissions. It is also a package with the Sequin sequence submission software (Johnson et al., 2008).

3.1.3.3 Basic local Alignment Tool (BLAST) (http://blast.ncbi.nlm.nih.gov/Blast.cgi)

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between the sequences. BLAST compares protein or nucleotide sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as helped to categorize members of gene families. BLASTx (Translate BLAST) is used for removing the coding region among the sequences and searched protein database using a translated
nucleotide query against the swiss prot (Wheeler et al., 2003). In this research work, it was very important to remove coding regions among the sequences and through BLASTx coding region (with nucleotide sequences of *C. roseus* against the protein non-redundant database) were removed. BLASTn is used for finding the similarity among the reference miRNA sequences against the (*Catharanthus roseus* G.Don taxid: 4057) whole genome shotgun contigs and after comparative study homologous miRNA sequences were found. In nucleotide BLAST, searched a nucleotide database using a nucleotide query (Johnson et al., 2008).

3.1.3.4 Coding Potential Calculator (http://cpc.cbi.pku.edu.cn)

*De-novo* method is based on feature selection technique that used for the identification of ncRNAs for 99% results. Coding Potential Calculator (CPC) is the Support Vector Machine based classifier to assess the protein-coding potential of a transcript based on six biologically meaningful sequence features. Tenfold cross-validation on the training dataset and independent testing on three large standalone datasets showed that CPC can distinguish coding from noncoding transcripts with high accuracy. In addition to predicting the coding potential of the input transcripts, the CPC web server also graphically displays detailed sequence features and additional annotations of the transcript that may facilitate users further analysis (Kong et al., 2007). It shows the result in the form of noncoding, noncoding weak, coding and coding weak data. Excluded coding and coding weak regions and further worked on noncoding and noncoding weak data.
3.1.3.5 Mfold (http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form)

Michael Zuker professor of mathematical sciences developed this tool for predicting the secondary structure of RNA ad DNA mainly by using thermodynamic methods. This algorithms have been extensively used for drug design and worked on DNA folding. DNA folding has been very popular with the biotechnology society. Nick Markham, a computer scientist who worked with professor Zuker created the UNAFold software package that both replaces and extends mfold. Secondary structures for all probable miRNA precursor sequences were predicted by MFold. This software computed the minimum free energy (MFE) contribution for various possible secondary structures. Less the minimum free energy of the structure showed high thermodynamic stability (Zuker, 2003).

3.1.3.6 psRNATarget: A Plant Small RNA Target Analysis Server

(http://plantgrn.noble.org/psRNATarget/)

psRNATarget is a plant small RNA target analysis server that depends on two important analysis. First reverse complementary match between miRNA and target transcript using a proven scoring schema and second target site accessibility evaluation by calculating unpaired energy (UPE) required to open secondary structure around miRNA’s target site on mRNA. PsRNATarget incorporates recent discoveries in plant miRNA target recognition, e.g., it distinguishes translational and post-transcriptional inhibition and it reports the number of miRNA/target site pairs that may affect miRNA binding activity to target transcript. psRNATarget is designed for high-throughput sequence analysis of next generation data with an efficient distributed computing back-end pipeline that runs on a Linux cluster. The server front-end integrates three simplified user friendly interfaces to accept user submitted or preloaded miRNAs, transcript sequences and outputs a
comprehensive list of miRNA target pairs along with the online tools (Dai and Zhao, 2011).

3.1.3.7 CLUSTALX2

The CLUSTALX and CLUSTALW is used for multiple sequence alignment (MSA). In the present study CLUSTALX2 was used for multiple sequence alignment to find the conservation among the sequences. CLUSTALX is completely written in C++ programming language. ClustalX2 is the new version of the ClustalX graphical alignment tool. The graphical interface sections of ClustalX2.0 have been completely rewritten using the Qt GUI toolbox. Qt is an easy to use, multi-platform C++ GUI toolkit. The Qt toolbox provides a native look and feel on Windows, Linux and Mac platforms. ClustalX2.0 has the same functionality as ClustalX. It has included new code for UPGMA guide trees as an alternative to the usual Neighbor-Joining guide trees. This helps to speed up the alignment of extremely large data sets of tens to thousands of sequences. It has also included an iterative alignment facility to increase alignment accuracy (Bateman, 2007).

3.1.3.8 MEGA5.0

Molecular Evolutionary Genetics Analysis (MEGA5.0) is a user friendly software for mining online databases, building sequence alignments and phylogenetic trees and using methods of evolutionary bioinformatics in basic biology, biomedicine and evolution. The newest addition in MEGA5.0 is a collection of maximum likelihood (ML) analyse for inferring evolutionary trees, selecting best fit substitution models (nucleotide or amino acid), inferring ancestral state and sequences (along with probabilities) and estimating evolutionary rates. In computer simulation analyse ML tree inference algorithms in MEGA5.0 compared satisfactorily with other software packages in terms of computational efficiency and estimated the accuracy of the phylogenetic trees, substitution parameters.
and rate variation among sites. This version of MEGA is proposed for the Windows platform and it has been configured for effective use on Mac OS X and Linux desktops (Tamura et al., 2011).

3.1.4. Sequence analysis for RNA binding protein

In sequence analysis, different type of software and web servers were used for finding the RNA binding protein.

3.1.4.1 Conserved domain database (CDD)


The conserved domain database (CDD) is a part of NCBI’s Entrez database system and serves as a primary resource for the annotation of conserved domain footprints on protein sequences in Entrez. CDD is a protein annotation resource that consists of a collection of well annotated multiple sequence alignment models for ancient domains and full length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST CD-Search & Batch CD-Search. CD-Search is NCBI’s interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST that scan quickly a set of pre calculated position specific scoring matrices (PSSMs) with a protein query. CDTree is a helper application for web browser that allows to interactively view and examine conserved domain hierarchies curated at NCBI (Marchler-Bauer et al., 2011).

3.1.4.2 PROSITE (http://www.expasy.ch/prosite/)

PROSITE is a documented database using patterns and profiles as motifs descriptors. PROSITE is an annotated collection of motifs descriptors dedicated to the identification of protein families and domains. The motifs descriptors used in PROSITE
are either patterns or profiles, which are derived from multiple alignments of homologous sequence (de Castro et al., 2006)

3.1.4.3 MEME (Multiple EM for Motif Elicitation) (http://meme.nbcr.net)

MEME (Multiple EM for Motif Elicitation) is one of the most commonly used tools for searching for novel ‘signals’ sets of biological sequences. Applications include the discovery of new transcription factor binding sites and protein domains. MEME works by searching for repeated, ungapped sequence patterns that occur in the DNA or protein sequences provided by the user. Users can perform MEME searches via the web server hosted by the National Biomedical Computation Resource and several mirror sites. Through the same web server, users can also access the Motif Alignment and Search Tool to search sequence databases for matches to motifs encoded in several popular formats (Bailey et al., 2009).

3.1.5 Tools for Protein (primary, Secondary and tertiary structure) Analysis

Study of protein analysis was performed through different types of software. The following offline and online tools were used for the protein analysis and their structure prediction.

3.1.5.1 ProtParam (http://web.expasy.org/protparam/)

The program ProtParam, a component of the ExPASy server estimates many basic physicochemical properties of a polypeptide on the basis of its sequence. It is a home of UniProtKB/Swiss-Prot sequence database. It is a manually annotated protein knowledge base established in 1986 and is now maintained by Swiss Institute of Bioinformatics. The annotations include the functions of the protein, post-translational modifications, carbohydrates phosphorylation, domains and sites, calcium binding regions, ATP-binding
sites, secondary structure, quaternary structure, homodimer, heterotrimer, similarities to other proteins, diseases associated with deficiencies in the protein, sequence conflicts and variants, etc. PROTPARAM computes the physiochemical properties of protein through its amino acid sequence on the various parameters such as the molecular weight, theoretical isoelectric point, instability index, aliphatic index, GRAVY. The aliphatic index defines as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine and leucine) (Porto et al., 2014). It is the positive factor for thermostability which explored by Ikai (1980). The primary sequence analysis of protein was calculated through Expasy’s Protparam.

3.1.5.2 PROCHECK (http://nihserver.mbi.ucla.edu/SAVES/PROCHECK)

PROCHECK gives the details of stereochemistry of a protein structure. It reads the atom coordinates in protein data bank format. Its output comprises the number of plots in the postscript format and comprehensive residue. It gives an assessment of the overall quality of the structure as compared with well refined structures of the same resolution and also highlight regions. The PROCHECK programs are useful for assessing the quality of protein structures (Laskowski et al., 1993). The sterio chemical quality of the models were checked through Ramachandran plot using Procheck tool.

3.1.5.3 PDBsum (http://www.ebi.ac.uk/pdbsum)

PDBsum provides summary information about each experimentally determined structural model in the Protein Data Bank (PDB). It accepts users’ own PDB format files and generates a private set of analyses for each uploaded structure. Since its inception in 1971, the Protein Data Bank (PDB) has released over 55,000 experimentally determined structural models of proteins and nucleic acids. The archiving management and quality control of these models are performed by worldwide PDB (wwPDB) now a days, a
Materials and Methods

Chapter 3

consortium whose partners comprise: the Research Collaboratory for Structural Bioinformatics (RCSB), the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (EBI), the Protein Data Bank Japan (PDBJ) at Osaka University and more recently, the BioMagResBank at the University of Wisconsin-Madison (Laskowski, 2009). Secondary structure analysis has been done through EBI PDBsum tool.

3.1.5.4 PFAM (http://pfam.xfam.org/)

PFAM is used for multiple sequence alignment and HMM profiles of protein domains. Pfam is a comprehensive database of conserved protein and provide query sequences for these motifs. These tools are based on profile models (HMMER3). PFam which is a widely used database of protein families, currently containing more than 13,000 curated protein families. The last years generated, 1840 new families and 80% increased coverage of the UniProt Knowledge base. Proteins are generally composed of one or more functional regions called as domains. Domains define the nature and function of the protein. There are two machinery on Pfam, Pfam-A and Pfam-B. Pfam-A manually curated families and entries cover a sequences in the underlying sequence database in order to give a coverage of known proteins. Pfam-B families can be useful for functionally conserved regions when no Pfam-A entries are found. Pfam provides the various analysis tools for sequence search against the Pfam matches that identify the family and domain identification can be done through it (Finn et al., 2010).

3.1.5.5 ESPript SERVER (http://escript.ibcp.fr/ESPript/ESPript/)

ESPript is a FORTRAN based program and web server was made in 1999. ESPript display multiple sequence alignments in the format of PostScript figure, PDF images in PNG that can adorned with secondary structure elements. ESPript was linked to three major webtools: ProDom that identifies protein domains and predict protein, which
predicts secondary structure elements and NPS@, which runs sequence alignment programs. It enables the creation from a single Protein Data Bank identifier to a multiple sequence alignment figure adorned with secondary structure elements of each sequence of known 3D structure. Similar 3D structures are superimposed in turn with the program PROFIT and a final figure is drawn with BOBSCRIPT, which shows sequence and structure conservation along the Ca trace of the query (Gouet et al., 2000).

3.1.6 Computational tools for homology modeling and validation

The following standalone computational tools and web servers were used for tertiary structure prediction and for the validation of generated structure.

3.1.6.1 MODELLER 9v5

MODELLER is used to modeled the three dimensional structure of protein, their assemblies by satisfaction and spatial restraints through the homology modeling. For the development of 3D structure user provides the alignment of template and target sequence. MODELLER calculates the model with all non hydrogen atoms and generates the restraints on dihedral angles, distances, pairs of dihedrals and output 3D, which satisfies all these restraints. MODELLER can do the alignment between the structures, sequences of protein, looping in model and clustering of proteins (John and Sali, 2003).

3.1.6.2 Structural Analysis and Verification Server (SAVeS) (http://nihserver.mbi.ucla.edu/SAVES/)

SAVeS stands for Structural Analysis and Verification Server. It was used to check and validate the three dimensional structure of protein after the refinement of protein model. It includes various servers for the validation of the structure such as Ramachadran plot and Verify-3D (Bowie et al., 1991).
3.1.6.2.1 Ramachandran plot

The Ramachandran plot is a plot of the torsional angles phi (φ) and psi (ψ) of the residues (amino acids) contained in a peptide. In sequence order φ is the N(i-1),C(i),Ca(i),N(i) torsion angle and ψ is the C(i),Ca(i),N(i),C(i+1) torsion angle. The plot was developed in 1963 by G. N. Ramachandran by plotting the φ values on the x-axis and the ψ values on the y-axis. Plotting of the torsional angles in this way graphically shows, which combination of angles are possible. The torsional angles of each residue in a peptide define the geometry of its attachment to its two adjacent residues by positioning its planar peptide bond relative to the two adjacent planar peptide bonds, thus the torsional angles determine the conformation of the residues and the peptide. Many of the angle combinations and conformations of residues are not possible because of steric hindrance. Through Ramachandran plot, protein structural scientists can determine which torsional angles are permitted and can obtain insight into the structure of peptides (Hollingsworth and Karplus, 2010).

3.1.6.2.2 VERIFY-3D (http://services.mbi.ucla.edu/Verify_3D/)

VERIFY-3D is used to verify the three dimensional structure of protein. This server resolves the issue of mistraced or wrongly folded model through the testing of 3D structure against the amino acids of protein. VERIFY-3D program was used to determine the compatibility of the atomic model (3D) with its own amino acid sequence (1D). It also provides the visual analysis of quality of a putative crystal structures for a protein (Eisenberg et al., 1997).

3.1.6.2.3 ProSA (www.prosa.services.came.sbg.ac.at)

ProSA (Protein Structure Analysis) is largely user base server and employed in the refinement and validation of experimental protein structures as well as it performs the
structure prediction and modeling. The analysis of protein structures is generally difficult and cumbersome exercise. ProSA program is the web based tool and display the value of Z-scores and energy plots. In particular, the quality scores of a protein are displayed in the context of all known protein structures and highlighted in a 3D molecule viewer. This process encountered in the validation of protein structures obtained from X-ray analysis, NMR spectroscopy and theoretical calculations (Wiederstein and Sippl, 2007).

3.1.7 Visualization Tools

The visualization tools were used to display the alignment and tertiary structures in graphic formats. The following web server and standalone software were used to visualize the images.

3.1.7.1 pymol (www.pymol.org)

PYMOL is used to visualize and analysis of 3D protein structure. PYMOL was written in C or FORTRAN and it is the cross platform molecular graphics system. It has integrated python interpreter. It supports common representations of macromolecule structures such as wire bonds, cylinders, spheres, ball and stick, dot surfaces, wire mesh surfaces, back bone ribbons and cartoons ribbons, which was generated by Molscript. PYMOL can display the atoms and label them. Pymol can be used to evaluate the hydrogen bonding interactions and distances. Pymol can read the crystallographic structures, which are in format of X-PLOR and calculate the electron density of them (DeLano, 2002).

3.1.8 Development of CRONRDB-Catharanthus roseus noncoding RNA Database

Noncoding RNA genes information were organized in Microsoft Access database for both long and short ncRNA sequences. The sequence information stored in MS Access and it is treated as a back-end database. A friendly Graphical User Interface (GUI) operate
the database that was developed by HTML (Hyper Text Markup Language) webpages. The external style sheets (Cascading Style Sheets) were used for changing the appearances and layout of all the pages in web site. The front page of the browser were connected to the back-end. MS Access uses server-side scripting language ASP (Active Server Pages) and ASP.net is executed on the server and the plain HTML results are sent back to the browser. The database was web enabled and test version of the database was initially hosted at a local server.

3.2 METHODOLOGY

Identification and annotation of noncoding RNAs (long and short) was done using comparative genomics based approach. Those families were identified that were involved in flower development in *C. roseus*. Prediction and analysis of motif regions in RNA binding protein was done through *in silico* approach. Finally a database was developed in MS Access with the help of HTML (Hyper Text Markup Language), CSS (Cascading Style Sheets), Java Script, ASP (Active Server Pages), PERL (Practical Extraction Report Language) and CGI (Common Gateway Interphase).

3.2.1 Retrieval of dataset

The EST and GSS sequences of *C. roseus* were retrieved from dbEST (20,181 ESTs) and GSS (115 sequences) database. Two different datasets were used for the identification of noncoding RNA analysis. A lot of redundancy was found in these datasets. EGassembler web server was used for removing the redundancy and complexity among the sequences and developed a non-redundent data set for further processing. After removing contamination through EGassembler, 11,693 non-redundent sequences were obtained. Finally 2,663 contigs, 8,959 singletons were found through EST analysis and 13 contigs, 58 singletons were found through GSS data analysis. After EST and GSS data
analysis, total 11,693 sequences (2,676 contigs and 9,017 singletons) were found. These contigs and singletons were found in different tissue types of C. roseus like Substraction library, Substration leaf specific cDNA library, Root CR03 cDNA library, Root cDNA library, Phytoplasma infected library, Leaf specific cDNA library of one year old, Leaf CR02 cDNA library, leaf cDNA library, Flower bud CR01 cDNA library, Crunigene stem, Crunigene leaf, Crunigene root, different display cDNA clone, crolfing, CR04 cDNA library, CR05 cDNA library, Alpha tubulin gene library, Beta tubulin gene library and GSS (Table 3.1).

Table-3.1: EGassembler results

<table>
<thead>
<tr>
<th>S.No</th>
<th>Tissue Type</th>
<th>Total sequences (NCBI dbESTs)</th>
<th>EGassembler Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Contig</td>
</tr>
<tr>
<td>1</td>
<td>Subtraction Library</td>
<td>345</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>subtracted leaf specific cDNA library</td>
<td>55</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Root CR03 cDNA library</td>
<td>2236</td>
<td>278</td>
</tr>
<tr>
<td>4</td>
<td>root cDNA library</td>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>Phytoplasma infected library</td>
<td>258</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>Leaf specific cDNA library of one year old</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Leaf CR02 cDNA library</td>
<td>1997</td>
<td>186</td>
</tr>
<tr>
<td>8</td>
<td>leaf cDNA library</td>
<td>55</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Flower bud CR01 cDNA library</td>
<td>2178</td>
<td>253</td>
</tr>
<tr>
<td>10</td>
<td>differential display cDNA clone</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>CrUniGene stem</td>
<td>458</td>
<td>61</td>
</tr>
<tr>
<td>12</td>
<td>CrUniGene root</td>
<td>5039</td>
<td>757</td>
</tr>
<tr>
<td>13</td>
<td>CrUniGene leaf</td>
<td>4785</td>
<td>787</td>
</tr>
<tr>
<td>14</td>
<td>CROLF1NG</td>
<td>575</td>
<td>77</td>
</tr>
<tr>
<td>15</td>
<td>CR04 cDNA library</td>
<td>1164</td>
<td>121</td>
</tr>
<tr>
<td>16</td>
<td>CR05 cDNA library</td>
<td>943</td>
<td>99</td>
</tr>
<tr>
<td>17</td>
<td>alpha tubulin gene library</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
3.2.2 Identification of noncoding RNA (short and long noncoding RNA)

For the identification of the noncoding RNA (long and short noncoding RNA), a general pipeline was developed that was further divided in two (ORF and CPC based) pipelines (Fig-3.1).

![Flow chart for EST and GSS based identification of IncRNAs in C. roseus](image)

**Fig-3.1**: Flow chart for EST and GSS based identification of IncRNAs in *C. roseus*
3.2.2.1 The ORF based pipeline

In this pipeline those transcripts were selected, which were greater than and equal to 200 nt in length. In next step selected those transcript that were less than and equal to 100 amino acid in length and compared these transcript to protein non-redundant database. Those transcripts were selected, which have e-value above 0.001. Those sequences showed insignificant results with protein database these are considered as pure noncoding, which showed significant results against protein BLAST with less than 30% identity, these are considered as noncoding weak and significant result with greater than 30% identity, these are considered as coding sequences. For further analysis coding sequences were removed and worked on noncoding and noncoding weak data.

3.2.2.2 CPC based pipeline

The full length of 20,296 cDNA sequences were uploaded using CPC for finding noncoding RNAs. The output data was analyzed in the form of noncoding, noncoding weak and coding and coding weak. The coding regions were excluded and worked on noncoding and noncoding weak data. Those transcripts were selected in CPC based pipeline, which were greater than and equal to 200 nt and less than and equal to 100 aa in length.

3.2.3 Identification of short noncoding RNAs

Comparative genomics based approach was used for the identification of novel miRNAs (Fig-3.2).
3.2.3.1 Micro RNA database repository

A total of 324 miRNAs sequences were retrieved of *Arabidopsis thaliana* (model plant) from miRBase deposited in (release 19.0, July 19, 2012). MiRBase is the main online repository for all microRNA sequences. It provides whole evidence for microRNA annotations and it resumes previous annotations. A complete set of 79,302 whole genome shotgun assembly sequences of *C. roseus* were used in the present investigation.

3.2.3.2 Whole Genome Sequence (WGS) based comparative genomic resource

Homology based tools like nucleotide BLAST from NCBI were used for finding the similarity among the mature miRNA sequences against the *C. roseus* whole genome shotgun contigs and eliminated protein coding sequences respectively by BLASTx.
3.2.4 Prediction of potential miRNAs

For finding potential miRNAs of *C. roseus*, EST and WGS sequences were compared with reported plant miRNA sequences through nucleotide BLAST. Those miRNA sequences were selected, where ≤3 nt mismatch with the subject sequence (*Arabidopsis thaliana*). After using the comparative analysis mature sequences were found. Precursor miRNA sequence were extracted 200 nt from BLAST program (100 upstream and 100 downstream). These precursor sequences were aligned against swissprot database using BLASTx for removing the coding regions. The noncoding sequences were used for structure prediction.

3.2.5 Secondary structure prediction

Secondary structure prediction for all probable miRNA precursor sequences were predicted through MFold web server. This software computes the minimum free energy (MFE) contribution for different possible secondary structures. Less the minimum free energy showed high thermo dynamical stability of the structure. After that adjusted minimal folding energy (AMFE) and minimum free energy index (MFEI) were calculated. The AMFE and MFEI values were calculated using the following equation.

\[
\text{AMFE} = \left( \frac{\text{MFE}}{\text{length of the pre-miRNA sequence}} \right) \times 100
\]

\[
\text{MFEI} = \left( \frac{\text{MFE}}{\text{length of the pre-miRNA sequence}} \right) \times 100 \div (G+C)\% 
\]

The secondary structure of precursor sequences gathering the following five criteria were chosen as potential miRNA homologues:
(1) The sequence should fold into a complete stem loop hairpin.

(2) A mature 18-22 nt miRNA sequence must be located on one arm of the hairpin structure.

(3) Any predicted 18-22 mature miRNAs should not have more than 3 nt substitutions compared to existing known miRNAs.

(4) The predicted secondary structure should have ≥ 10 Kcal mole⁻¹ free energy.

(5) The A+U content should be in the range of 30-70%.

All pre-miRNAs those satisfied the above criteria were selected in order to identify *C. roseus* miRNA family (Lakshmana Reddy *et al.*, 2012).

### 3.2.6 Conservation and Phylogenetic Analysis of miRNA precursors in *C. roseus*

The predicted miRNAs were classified in to 22 different family based on their precursor miRNA (pre-miRNA) sequences. Conservation showed by multiple sequence alignment among the family members. The evolutionary relationships were analyzed among members of each miRNA family. The alignment file was used to compare areas of conserved sequences through Bioedit software (http://www.mbio.ncsu.edu/bioedit/bioedit.html). Phylogenetic analysis showed close and distantly relationships among the families of different plants and MEGA software was used for phylogenetic analysis (Tamura *et al.*, 2011).

### 3.2.7 miRNA Target prediction

In present study, psRNA Target tool was used for target prediction of miRNA.
3.2.7.1 Target gene prediction

A total of 23 targets were found through psRNATarget. Predicted mRNA targets are essential in order to understand the biological functions of miRNAs. After structure prediction, the potential targets of the newly identified miRNAs were predicted through psRNA Target, in which the score for an alignment depends on the length of the query and the query-subject alignment. Identification of miRNAs and their targets in *C. roseus* will help not only in understanding the mechanism of control the cellular processes but also help in controlling the traits. psRNATarget program was executed with following parameters: (1) score or maximum expectation at 3, (2) length of complementary scoring at 20, (3) target accessibility at 25, (4) flanking length around target 17 bp upstream and 13 bp downstream and (5) range of central mismatch leading to translational inhibition in between 9-11. All predicted target genes were evaluated through psRNATarget by scoring and total score for an alignment was calculated based on 20 nt subsequences, and the minimum score was considered the total score for the query-subject alignment. All miRNAs that were satisfied the above criteria were selected (Dai and Zhao, 2011).

3.2.8 Identification of noncoding RNA that are helpful in flower development

After previous studies those noncoding RNAs were identified that were helpful in flower development in *C. roseus*.

3.2.8.1 Conservation analysis

The predicted miRNAs were classified in to different families. Conservation showed by multiple sequence alignment among the family members. The evolutionary relationships were analyzed among members of each miRNA family that were involved in
flower development. The alignme-nt file was used to compare areas of conserved sequence through Bioedit software (http://www.mbio.ncsu.edu/bioedit/bioedit.html).

3.2.8.2 Phylogenetic analysis

Phylogenetic analysis was considered on those sequences that were involved in flower development through MEGA (Tamura et al., 2011). Evolutionary relationships (Close and distantly) were found among the members of the same family through phylogeny.

3.2.9 Motif identification in RNA binding protein through In silico approach

The present study was based on in silico approach for the prediction and analysis of motif regions in RNA binding protein. A domain architecture was developed for retrieving sequences through NCBI Conserve Domain Database and Scan Prosite. Furthermore, conservation was defined between the sequences and developed a phylogram for showing the relationships among sequences. Multiple sequence alignment was performed among the sequences and showed highly conserved residues with their respective positions. Phylogenetic analysis showed close and distantly relationships among the sequences. Motif was identified with the help of MEME tool in *C. roseus*.

The primary sequence analysis of protein was calculated through Expasy’s Protparam and secondary structure analysis through EBI PDBsum tool. A three dimensional structure was constructed using comparative modeling approach. The three dimensional structure of the RNA binding protein of *C. roseus* was absent in Protein data bank. A theoretical three dimensional structure was developed. Comparative modeling predicts the 3D structure of hypothetical model of the given protein sequence (target),
primarily based on the alignment of the template and generate 50 models with the help of software modeler9v5. The sterio-chemical quality of these models were checked through Ramachandran Plot using Procheck tool. Ramachandran plot (Phi-Psi) pairs was prescribed and after the analysis of these values confirmed a good quality of model, the packing quality of each residue of the model was checked through Verify_3D program and protein structure analysis was completed through ProSA web server. Both programs show the quality of the model structure is good or not.

3.2.10 Database development

Noncoding RNA genes information were organized in Microsoft Access database for both long and short noncoding RNAs with their specific fields. The sequence information stored in MS Access was treated as a back-end database. A friendly Graphical User Interface (GUI) operate the database and it was developed through HTML (Hyper Text Markup Language) web pages. The external style sheets like CSS (Cascading Style Sheets) that were used for changing the appearance and layout of all the pages on the website. The front page of the browser was connected to the back-end MS Access using server-side scripting language ASP (Active Server Pages). ASP.net was executed on the server and the plain HTML results were sent back to the browser. The database was web enabled and test version of the database was initially hosted at a local server.