DISCUSSION

Bioinformatics and computational molecular biology are emerging interdisciplinary field between computer science and molecular biology. This field encloses the study of computational problems arising from the huge amount of biological data. Molecular biology domain is very complex in nature due to their vast size and variations. The main pathway of molecular biology is central dogma, in which DNA convert into mRNA (transcription) and mRNA convert into protein (translation). In processing of DNA convert into mRNA, DNA divided into two parts coding (mRNA) and noncoding (junk DNA). This study suggest the prediction of noncoding RNAs (either it is long and short) in the molecular biology domain. For this investigation raw data was collected from the public domain like NCBI, data type ESTs, GSS and WGS (Johnson et al., 2008). From ESTs and GSS the total 20,296 redundant sequences were collected and removed the redundancy through EGassembler online server and developed non-redundent data (Masoudi-Nejad et al., 2006). In the processing of EGassembler 31 ESTs were trashed, 4 in dust, 19 by short, 5966 sequences were trimmed out, resulting into a total of 14,266 ESTs; after removing redundancy and complexity through EGassembler, 11,693 non-redundent sequences were found (Table-4.1).

5.1 Identification of IncRNAs through ORF based pipeline

In this study, in house pipeline was developed and generated novel long noncoding RNA in *C. roseus*. The ORF (Open Reading Frame) based approach used to identify long noncoding RNAs. The rich EST resource were used to identify novel noncoding transcripts as a good resource. In the first step, set of full length cDNA sequences (11,693 sequences) were run through NCBI ORF finder (Wheeler et al., 2003). Sequences whose ORF length was greater 100 amino acid were eliminted and total
number of such sequences were 10,455. Total remaining sequences 1238 were used for further processing in house pipeline. According to Boerner and McGinnis (2012), in case of long noncoding RNA, a transcript length should be at least 200 nt. For the identification of long noncoding RNAs transcript length is an important feature. In house pipeline, predicted those ORF sequences, which were greater than and equal to 200 nucleotide and less than and equal to 100 amino acids in length (Fig-4.1). In next step, protein homology was searched against 1238 sequences. BLASTx was performed for searching protein coding sequences against nrdatabase. This criteria divided these sequence into three parts (a) pure noncoding (b) noncoding weak (c) coding sequences. Through this pipeline 104 long noncoding RNAs were identified. Standalone BLAST program was used for finding short noncoding RNA in putative 104 IncRNA candidates. After the BLAST output, one potential micro RNA was found from putative IncRNAs candidates. The remaining 103 IncRNA candidates were classified as IncRNA with no potential precursor. Resulting 76 sequences were found pure noncoding (do not show significant result with translated BLAST) and 27 sequences were considered as noncoding weak (showing significant result with tBLAST less than 30% identity with e-value >0.001) (Table-4.2) and excluded 1135 sequences that were considered as coding sequences (showing significant result with translated BLAST against the protein database greater than 30% identity) (Jia et al., 2010). In the end, through this pipeline 103 potential long noncoding RNAs were identified.

5.2 Identification of IncRNAs through CPC based pipeline

CPC (Coding Potential Calculator) is the SVM based classifier that is used to assess the protein coding potential of the translated sequences and based on six biologically meaningful sequence feature (Kong et al., 2007). This tool was used in
parallel with the described computational pipeline to identify the most comprehensive set of noncoding RNAs (Boerner, 2012). The set of full length 11,693 cDNA sequences were computed through CPC (Fig-4.2) and analyzed results in the form of noncoding, noncoding weak, coding and coding weak. Coding and coding weak sequences (7725 sequences) were eliminated and rest 3968 noncoding sequences were divided into two forms (2121 noncoding and 1847 noncoding weak). Those sequences were considered whose transcript length was greater than and equal to 200 nt and retrieved 1190 sequences. In next step, those sequences were analysed whose transcript length was less than and equal to 100 amino acid. Finally 707 putative long noncoding RNAs were found.

5.3 Combination between ORF based pipeline and CPC pipeline

After analysis of the both pipeline (ORF and CPC) results, 103 potential long noncoding RNAs were found, which were common in CPC and ORF based pipeline. Using in silico approach a comprehensive set of potential long noncoding RNAs were developed for further analysis.

5.4 Annotation of long noncoding RNAs

The 103 lncRNAs candidate were further classified to their other members like short noncoding RNA. The mode of this noncoding data set is illustrated in (Table-4.3). According to McGinnis and Boerner (2012), 103 lncRNAs on the basis of ORF size in amino acids were classified and identified linc1257 (44-68 amino acid in length) -12 members, NRON (69-71 amino acid in length) -13 members, HOTTIP (72-86 amino acid in length) -51 members, NEAT1 (87-97 amino acid in length) -20 members and RepA (98-105 amino acid in length) -7 members. The function of HOTTIP is cromosomal looping and cromatin modification (Ma et al., 2012). NEAT1 is present in paraspeckles
NRON is an antisense LncRNA and their function has not been clear yet. Linc1257 is long intergenic noncoding RNA that function is not clear (Zhang et al., 2013), the function of RepA (only define in mouse) is not clear yet in plants (Wilusz et al., 2009). Analysis was done in LncRNAs sequences. Adenine was more dominant nucleotide among four. LncRNA sequences showed vast variability in their length from 201 to 303 nt with an average length of 243.3 ± 29.92 nt (Table 4.5). The composition of the four nucleotides (A, C, G and U) were important parameters. The percentage of each nucleotide of *C. roseus* was not equal. Adenine was dominant and comprised 28.36 ± 5.62% (Table 4.5) of the total nucleotide composition, while uracil 27.52 ± 6.00% constituted 27.52 ± 6.00% (Fig-4.4) followed by guanine at 20.06 ± 5.66% and cytosine at 24.03 ± 7.01% (Table 4.5). The identified potential LncRNA sequences of *C. roseus* AU% were ranged from 41.13 to 69.05 with an average of 55.88 ± 6.66% (Fig-4.5) and GC% ranged from 30.95 to 58.87 with an average of 44.09 ± 6.66% (Fig-4.6). The A/U% was ranged from 0.43 to 4.17 with an average of 1.10 ± 0.46 and C/G% ranged from 0.48 to 9.00 with an average of 1.41 ± 1.08 in *C. roseus* sequences (Table 4.5).

### 5.5 Identification of short noncoding RNAs (miRNAs)

Although miRNA associated research is one of the most modern research topic in biological or biomedical fields (Zhang et al., 2007). However, in silico based prediction of other types of target specifically involved in the phase of translation or post translation that may not be identified experimentally easily in lab. Using various computational approaches, it was possible to identify wide spread complementary regions of the miRNAs with the computational methods that can be useful for plants.

For the computational identification of conserved miRNAs in *C. roseus*, a
reference dataset of 324 non-redundant miRNAs of *Arabidopsis thaliana* (collected from miRbase) were searched against whole genome shotgun contig sequences of *C. roseus*. Homology based approach defined for the identification of novel miRNAs. In mature sequences, mismatches should be only less than and equal to three. A total number of 53 miRNAs were found from homology approach but 23 miRNAs were reported in earlier studies (Prakash *et al.*, 2015). A set of 30 potential conserved miRNAs were reported in this study (*Table-4.4*). Although all mature miRNA sequences are well conserved along with plants but only those sequences were selected whose showed one to three mismatches against the miRNAs sequences of *Arabidiopsis thaliana* (model plant). The 30 potential miRNAs in *C. roseus* 73.33% (22) were located in 5’ arm of the stem loop hair pin structure while 26.67% (8) were located in 3’ arm, suggesting that *C. roseus* miRNAs are located in both the arms of the structure void any preference (*Table-4.6*). This property of mature miRNAs are typically confined to the stem-loop hairpin structure, which is an union with miRNAs in plant species (Zhang *et al.*, 2008).

5.6 Annotation of short noncoding RNAs

The identified 30 miRNAs belonged to 22 families (Cro-miR156b, Cro-miRNA162, Cro-miR169, Cro-miR172, Cro-miR395, Cro-miR398, Cro-miR399e Cro-miR403, Cro-miR414, Cro-miR780.2, Cro-miR827, Cro-miR830, Cro-miR833, Cro-miR838, Cro-miR841a, Cro-miR845a, Cro-miR2111b, Cro-miR5632, Cro-miR5641, Cro-miR5650, Cro-miR5653 and Cro-miR5658). The number of miRNAs in each family range from one to five members whereas from present annotation the miRNAs were found as distributed randomly (*Fig-4.7*). It was observed that in *C. roseus* most of the miRNA sequences have uracil as their first nucleotide dominant in nature (*Table-4.6*). This type of pattern distribution is in consensus with restricted miRNAs in other plants.
(Wang et al., 2012). According to Huang et al. (2011), previously reported that U is the predominant nucleotide at the 5' end of mature miRNA sequences in plants and proposed that U may play an important role in miRNA biogenesis through recognition of targeted miRNA precursors by the RNA-induced silencing complex (RISC). *C. roseus* pre-miRNAs sequences showed great variability in their length from 79 to 221 nt (*Table-4.7*) with an average length of 171 ± 54 nt. The length distribution of pre-miRNAs in *C. roseus* is similar to those pre-miRNA reported in other plant species such as *A. thaliana*. CarmiR5641 exhibited the shortest precursor length of 79 nt, whereas CarmiR169b-5p exhibited the longest precursor length of 221 nt (*Table-4.7*). The percentage of the four nucleotides (A, C, G and U) in *C. roseus* pre-miRNAs were not equal. Uracil was dominant and comprised 31.66 ± 3.92% (*Fig-4.9*) of the total nucleotide composition while adenine constituted 30.82 ± 4.01% (*Table-4.7*) followed by guanine at 19.19 ± 4.43% (*Table-4.7*) and cytosine at 18.32 ± 4.25% (*Table-4.7*). The nucleotide composition of the identified potential *C. roseus* pre-miRNA precursor sequences had (A+U) content ranging from 53.75 to 74.72 with an average of 62.58 ± 5.27 (*Fig-4.10*) and (G+C) content ranging from 25.27 to 46.25 with an average of 37.41 ± 5.27 (*Fig-4.11*). The average A/U ratio of the potential *C. roseus* pre-miRNA precursor sequences was 0.98 ± 0.20 (*Table-4.6*).

5.7 Characterization of *C. roseus* miRNAs

*C. roseus* miRNA precursor tend to form a stem loop hairpin in their secondary structure. The precursor sequence of 221 nucleotides were extracted (100 upstream and 100 downstream from the BLAST hit) and used for hairpin structure prediction (Gupta et al., 2015). The annotation process of all identified miRNAs was initiated by secondary structure prediction through MFold (Zuker, 2003b). The Mfold used with following
parameters: (1) The RNA sequence is linear, (2) Folding temperature is fixed at 37ºC, (3) Ionic condition: 1M NaCl (no divalent ion), (4) Percent suboptimality number is 5, (5) Enter an Upper bound on the number of computed folding is 50, (6) Window parameter select according to sequence length, (7) Maximum interior/bulge loop size is 30, (8) Maximum asymmetry of an interior/bulge loop is 30 and (9) Maximum distance between paired bases is no limit. These are significant step for deciding the outcome of the pre-miRNAs. After finding novel structure, following different parameter were studied: (1) Minimum free energy (MFE), (2) Adjusted minimum fold energy (AMFE) and (3) Minimal fold energy index (MFEI) for all predicted structure. We developed hairpin secondary structure through MFold web server (Zuker, 2003a). In *C. roseus* the minimum free energy (MFE) of pre-miRNA ranged between -10 to -75 with an average of -54.50 ± 16.88 kcal/mol (Fig-4.12), however MFE are strongly and positively correlated with their sequence length (Zhang et al., 2006). To normalize the potential effect of sequence length on MFE, it is a better way to distinguish miRNAs from other RNAs (Zhou et al., 2009) and measured adjusted MFE, AMFE and MFEI for this study. Here Minimum free energy index (MFEI) of pre-miRNAs were ranged between -0.42 to -1.98 with an average of -0.85 ± 0.30 kcal/mol (Fig-4.13) and Adjusted minimal folding energy (AMFE) were found between -12.65 to -54.6 with an average of -32.64 ± 9.39 (Fig-4.14).

The structure analysis was performed with the following criteria: (1) Potential mature miRNA has not been more than three mismatches compared with a known mature miRNA, (2) The selected sequence could fold into an appropriate stem-loop hairpin secondary structure, (3) The mature miRNA localized in any one arm of the stem-loop structure, (4) No loop or break in the miRNA or miRNA* sequences and (5) Not more
than 6 mismatches between the predicted mature miRNA sequence and its opposite miRNA*. Predicted secondary structures have high negative MFE and Minimal Folding Free Energy Index (MFEI) values. All those pre-miRNAs structure that satisfied the above criteria were selected in order to identify their respective family. All outputs of Mfold predicted structures satisfied the aforesaid criteria were presented in Fig-4.16. However, MFEs are strongly and positively correlated with their pre-miRNA sequence length. Longer the pre-miRNA sequences show more degree of freedom (and lower the MFEs), with which the sequences have to formed stable secondary structures.

After structure prediction, miRNA targets were predicted through psRNATarget web server (Dai and Zhao, 2011). 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. The psRNATarget program was executed with the following parameters: (1) score or maximum expectation at 3, (2) length of complementarity 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 web server by scoring and the criteria was used: each G:U wobble pairing was assigned 0.5 points and all other non-canonical Watson-Crick pairing were assigned 1.0 point each. Total score for an alignment (query-subject) was calculated based on 20 nt subsequences and the minimum score was considered after the alignment among the sequences. Because targets complementarity to the miRNA 5’ end appear to be critical, mismatches other than G:U wobbles at position 2-7 at the 5’ end were further penalized by 0.5 points in the final score. Sequences were considered as miRNA targets if the total score less than 3.0 points.
All miRNAs that satisfied the above criteria were selected. Out of thirty, twenty three miRNAs were found in *C. roseus* with their corresponding target genes. In *C. roseus* at least one target mRNA was identified for most of the *Catharanthus* miRNA families (Table-4.7). Twenty three conserved miRNA targets were found through psRNATarget web server. Cro-miR156 targets SPL15 (squamosa promoter binding like protein-like 15), which is a trans acting factor and these transcription factors are known to play role in the control the expression of the genes involved in regulation of metabolic processes. Cro-miR162b targets MNS3 (alpha mannosidase 3), which is involved in endoplasmic reticulum-associated degradation, Cro-miR169a targets GGT1 (Gamma glutamyl transpeptidase 1), which cleaves the gamma-glutamyl bond of extra cellular glutathione, glutathione conjugates and other gamma-glutamyl compound. GGT is involved in control cells and shows an increased activity presenting malignant status of breast tissue. Cro-miR169b and Cro-miR169i target NF-A8 (nuclear factor Y, subunit A8), which stimulates the transcription of various genes by recognizing and binding to a CCAAT motif in promoters. Cro-miR172b-5p and Cro-miR 172e-5p target ADC1, ARGDC1, ARGDC, SPE1 (Arginine decarboxylase1), which specifically catalyzes the decarboxylation of L-arginine to agmatine. ADC1 is involved in polyamine biosynthesis. Cro-miR172b-3p and Cro-miR172e-3p targets TOE2 (target of early activation tagged 2), which is involved in DNA binding. Cro-miR172d-3p targets AP2, FLO2, FL1 (Integrasc-type DNA-binding super family protein), which is involved in sequence-specific DNA binding transcription factor activity. Cro-miR395b targets APS4 (pseudo uridine synthase/archaeosine transglycosylase-like family protein), which is involved in RNA binding. Cro-miR399f targets UBC24, ATUBC24, PHO2 (phosphate 2), which is involved in phosphate-signaling pathway in plants. Cro-miR403-3p targets AGO2
(Argonaute family protein), which is involved in gene-silencing pathways guided by small RNAs. Cro-miR414 targets DUF584, which regulates the expression of proteins associated with leaf senescence. Cro-miR833a-3p and Cro-miR5650 target protein kinase superfamily protein, the enzymes of this family are related to substrate specificities and modes of regulation. Cro-miR841a-5p targets pseudogene (similar to Putative histone H2A) and its function are not clear yet. Cro-miR845a targets transposable element gene, which is involved in DNA sequences that move from one location to another in the genome. Cro-miR2111b-5p targets Galactose oxidase/kelch repeat superfamily protein, which catalyses the oxidation of the hydroxyl group at the C6 position in D-galactose. Cro-miR5632-3p targets ZIP metal ion transporter family, which is involved in transporting a variety of cations including cadmium, iron, manganese and zinc. Cro-miR5658 targets ARM repeat superfamily protein, which is involved in RNA cap binding, DNA methylation, meristem structural organization, primary miRNA processing. Comprehensive characterization of all the identified Catharanthus miRNAs and their targets genes in different tissues would be helpful to understand the tissue specific expression of all the miRNAs. No targets were predicted in (Cro-miR169e, Cro-miR399e, Cro-miR827, Cro-miR830, Cro-miR838, Cro-mir5641 and Cro-miR5653). The functions of Cro-miR398c-3p and Cro-miR780.2 are not clear yet.

5.8 Identification of flower developing miRNAs

According to Luo et al. (2013), nine conserved miRNA families are involved in flower development. These are miR156, miR159, miR160, miR164, miR165, miR166, miR167, miR169, miR172 and miR319. MiRNAs play an essential role in plant flowering time and floral organ identity. In C. roseus two miRNA families were found that are helpful in flower development. MiRNA 156 and 172 have shown involvement in
flowering time control under special environmental conditions (Luo et al., 2013). The over expression of miR156 can delay flowering significantly (Wang et al., 2012). A relationship between miRNAs and flowering-related biological processes in miR156 family in C. roseus was found. Studies show that miR172 regulates the initiation of flowering. Homology based prediction was used for flower development in C. roseus. Those miRNA sequences in other plants were identified, which are helpful in flower development. In C. roseus Cro-miR156-5p targets SPL15 (squamosa promoter binding like protein-like 15) that belongs to transcription gene family to control the transition from the vegetative phase to the floral phase. In this study, it was found that Cro-miR172-3p and Cro-miR172e-3p target AP2-like gene, such as TARGET OF EAT2 (TOE2), which is involved in flowering time and Cro-miR172d-3p target to AP2 gene that is involved in floral organ identity.

Previous study suggested that the function of miR172 is highly conserved among plants (Prakash et al., 2015). MiR156, miR172 target unigenes included MADS-box transcription factors (Li et al., 2011). The MADS-box transcription factors are well known to determine the identity of flower organs (Michaels and Amasino, 2001) and hormone responding factors involved in floral development (Li et al., 2011; Michaels and Amasino, 2001). Through comparative genomics approach two novel families (miR156, miR172) were found that are helpful in flower development in C. roseus. In here, standalone BLAST program was used for identifying those miRNAs in other plants, which are involved in flower development (Li et al., 2011; Michaels ana Amasino, 2001). In here, standalone BLAST program was used for identifying those miRNAs in other plants, which are involved in flower development. The BLAST output file showed score, identities and e-value (Table-4.8). Higher identities and massive e-values were found in
miR156 family members in different plants. In C. roseus, miR156 family showed 100% similarity with miR156 family members of Zea mays, Populus trichocarpa, Medicago truncatula, Arabidopsis thaliana and Oryza sativa and 95% similarity with some members of Zea mays, Populus trichocarpa, Medicago truncatula and Oryza sativa (Table-4.8).

The same process was repeated in miRNA172 family members in different plants. In C. roseus, five members of miR172 family (Cro-miR172b-5p, Cro-miR172b-3p, Cro-miR172d-3p, Cro-miR172e-5p and Cro-miR172e-3p) were identified. Cro-miR172b-5p showed 100% similarity with miR172 family members of Oryza sativa and Zea mays. Some members of Oryza sativa, Glycine max, Zea mays, Populus trichocarpa and Oryza sativa showed 94% similarity. Cro-miR172b-3p showed 100% similarity with Glycine max, Populus trichocarpa, Oryza sativa, Arabidopsis thaliana, Zea mays, Populus trichocarpa and Oryza sativa, and 95% similarity with some members of Medicago truncatula and Populus trichocarpa. Some members of Oryza sativa and Zea mays showed 94% similarity and Populus trichocarpa showed 93% similarity.

Cro-miR172d-3p showed 100% similarity with Glycine max, Zea mays, Populus trichocarpa, Oryza sativa, Arabidopsis thaliana, Zea mays, and Medicago truncatula and Populus trichocarpa showed 95% similarity. Oryza sativa, Zea mays showed 94% similarity and Populus trichocarpa showed 93% similarity with C. roseus.

Cro-miR172e-5p showed 100% similarity with Oryza sativa, Zea mays, Populus trichocarpa and Oryza sativa and 94% similarity with Glycine max, Zea mays, Populus trichocarpa, Oryza sativa and Arabidopsis thaliana.

Cro-miR172e-3p showed 100% similarity with Glycine max, Zea mays, Populus trichocarpa, Oryza sativa, Arabidopsis thaliana, Oryza sativa and some members of
Medicago truncatula, Populus trichocarpa, Oryza sativa, Zea mays showed 94% similarity, and Populus trichocarpa showed 93% similarity (Table-4.9). Further CLUSTRALX2 software was used for multiple sequence alignment. Multiple sequence alignment was used for finding the similarity among the sequences that are helpful in flower development in different plants (Fig-4.19, Fig-4.21). Thereafter, phylogenetic tree was constructed through MEGA for showing close and distantly relationship between the miR156 (Fig-4.20) and miR172 (Fig-4.22) family members in different plants.

5.9 Prediction and analysis of motif regions in RNA binding protein

Present study was based on in silico approach for the prediction and analysis of motif regions in RNA binding protein. RNA binding protein consists of RRM domain and glycine residues. Twelve plant sequences were collected which were rich in RNA binding activity and has a constructive relationship between RRM domain and RNA binding proteins. First a domain architecture in retrieving sequences was developed through NCBI Conserve Domain Database (Marchler-Bauer et al., 2007) and Scan Prosite (de Castro et al., 2006). After analysis, out of twelve plant sequences, eleven sequences show at least one RRM binding domain, but in Nicotiana sylvestris two RRM domains were found (Fig-4.23).

Furthermore, conservation was defined among the sequences and developed a phylogram for showing the relationship among the sequences. Multiple sequence alignment was performed among the sequences and showed high conserved residues with their respective positions. These results concluded that two highly conserve regions were present (RRM region and glycine rich regions) (Fig-4.24A). For phylogenetic analysis, plant profile showed two major clusters (Fig-4.24B), cluster-I consists of seven species.
(Sinapis alba, Arabidopsis thaliana, Daucus corata, Brassica napus, Solanum tuberosum, Catharanthus roseus and Sesamum indicum) and showed close relationships with Catharanthus RNA binding protein. Cluster-II consists of five species (Oryza sativa, Zea mays, Sorghum bicolor, Hordeum vulgare and Nicotiana sylvestris) and showed distantly relationships with Catharanthus RNA binding protein. One motif was identified with the help of MEME tool in C. roseus.

The primary sequence analysis of protein was calculated through Expasy’s Protparam. The program computes the extinction coefficient of 276, 278, 279, 280 and 282 nm wavelengths and, in addition, 280 nm has been elected since proteins absorb light strongly. Extinction coefficient of protein at 280 nm was 17,085 M$^{-1}$ cm$^{-1}$. The computed extinction coefficient can help in the quantitative study of protein-protein and protein-ligand interaction in solution. The instability index provides to determine the stability of protein in a test tube. There are definite dipeptide, which is particularly divergent in the unstable protein compared with those in the stable once. This method assigned a weight value of instability, which is feasible to compute an instability index (II). A protein whose instability index is slighter than 40 is estimated as stable. The value above 40 estimates that protein may be unstable. The instability index value of the protein was found 55.80, which indicates that the protein is unstable. The aliphatic index elucidated as the relative volume of a protein occupied by aliphatic side chains (A, V, I and L). It is estimated as a positive factor for the increment of the thermal stability of globular protein. Aliphatic Index (AI) of the protein sequence was 34.74, the very low aliphatic index of the protein sequence indicates that these proteins may be unstable at low temperature. The Gravy average hydropathy (GRAVY) value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids. The number of
residues in the sequence were divided. A GRAVY index of protein was -0.796. This low value shows the probability of better interaction with water. The secondary structure analysis through EBI PDBsum tool, showed two helices and four beta sheets of the given sequence (Fig:4.25A, B). A three dimensional structure was constructed through comparative modeling approach. This approach provides reasonable results based on the assumption that the tertiary structure of the two proteins will be similar. 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. BLAST scanning results were more identical with crystallographic structure *E.coli* (PDB ID: 4C7Q) while the template was determined on the basis of higher sequence identity. Present investigation found 86% sequence identity with 72 conserved residues. 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 resulting 3D structure of protein was sorted according to the score calculated from discrete optimization and protein energy (DOPE) scoring function. Selected 10 models whose DOPE score were low. The sterio-chemical quality of these models were checked through Ramachandran Plot using Procheck tool. Ramachandran plot showed result in (Phi-Psi) pairs whose had 93.7% residue in most favoured regions, 4.2% residues in core region and additional allowed region, 0.0% residues in generously allowed regions and 2.1% residues in disallowed regions (Fig-4.29A). These values showed a good quality of model. The packing quality of each residue of the model was checked through Verify_3D program and protein structure analysis with ProSA web server (Wiederstein and Sippl, 2007). Both programs showed our model structure quality was good (Fig-4.29B).
5.10 Development of CRONRDB-\textit{Catharanthus roseus} noncoding RNA Database

Noncoding RNA genes information were organized in Microsoft Access database for both long and short noncoding RNAs with their specific fields like miRNAID, Description, WGSID, Mature length, Position, Mature sequence 5’ and 3’, MFE, MFEI, A content, C content, G content, U content, Target and pubmed id. Fourteen fields were assigned for micro RNA information and same process were applied in long noncoding RNA. The field “miRNAID” and “lncRNAID” were compositely choosen to assigned as a primary key in micro RNA table. A unique number was given to “Cro-miR” for micro RNA and to “Cro-IncR” for long noncoding RNA. The sequence information stored in MS Access treated as a back-end database. A friendly Graphical User Interface (GUI) to operate the database was developed through HTML (Hyper Text Markup Language) web pages. The external style sheets like CSS (Cascading Style Sheets) that allow changing the appearance and layout of all the pages on website. The front page of the browser was connected to the back-end MS Access using server-side scripting language ASP (Active Server Pages) and the ‘ASP.net’ executed on the server. 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 (Fig-4.30).