12 Summary and future prospects

This thesis is mainly focused on the AARS enzyme family, but we also developed several prediction methods for the ncRNAs and protein-RNA interactions. Recently, many noncoding transcripts proved to be involve in the cellular activities, those were earlier thought to be functionally inert or called as junk part of the genome. The growing knowledge about the role of ncRNAs in the cellular processes and availability of the huge amount of sequences gained tremendous attention recently but still the prediction and classification of ncRNA families is in its infancy stage. Although, several algorithms have been developed for predicting ncRNAs, but accuracy of the prediction is a major concern. Initially, we developed a method for discriminating coding and noncoding RNAs and thereafter developed a method for the further classification of ncRNAs into corresponding classes. By employing simple tri-nucleotide composition approach, we achieved 0.98 MCC for discriminating coding and noncoding RNAs. Our approach is very fast and suitable for the genome-wide predictions whereas previous methods used more complex features and are CPU-intensive. To classify ncRNAs into their respective classes, we used graph properties from the predicted structures because it is known fact that RNA structure is responsible for their biological function. Although, this graph properties based approach was already used by GraPPLE method, but they used only libSVM classifier for model development, whereas we used a variety of classifiers (BayeNet, NaiveBayes, MultilayerPerceptron, IBk, libSVM, SMO and RandomForest) and found that RandomForest is most suitable classifier along with graph properties-based approach for classifying ncRNAs. An online tool -- *RNAcon* (http://crdd.osdd.net/raghava/rnacon) has been developed for the service of RNA biologists. We hope that *RNAcon* will be useful in the annotate pipeline for transcripts in various NGS projects.

The protein-RNA interactions are involved in the variety of cellular processes and play an important role in the biological functions. The experimental determination of the both protein interacting nucleotides and protein interacting residues are tedious and time-consuming process. The increasing gap between the solved structures of the protein-RNA complexes and available sequences of RNA-binding

proteins demands sequence-based prediction methods. In this, *in-silico* prediction methods can be useful to detect these binding sites. In the past, several methods have been developed for predicting RIRs but no method available for PINs so far. We applied SVM with many approaches and found that tri-nucleotide composition based approaches performed well to predict PINs in the RNA sequences and developed *RNApin* (http://crdd.osdd.net/raghava/rnapin/) web-server. The existing RIRs prediction methods were focused on the individual or mono protein interacting residues. However, combination of different residues plays an important role in the protein-RNA interaction. Therefore, we applied PSSM-profile of patterns approach and developed separate SVM-based methods for the RNA-interacting mono residues (RIMRs), di-residues (RIDRs), tri-residues (RITRs), tetra residues (RITTRs) and penta residues (RIPRs). It is essential for the detection of RNA interacting sites in the sequences of RNA-binding proteins. All prediction models have been implemented in a webserver called *RNAint* (http://crdd.osdd.net/raghava/rnaint/). We anticipate that *RNApin* and *RNAint* methods are important for biologists to investigate and predict protein-RNA interactions in the future.

There are more than 2000 families of ncRNAs known yet (Burge et al., 2013) and tRNA is a major family of ncRNAs and play important role by specific binding with the AARS enzyme. Additionally, post-transcriptional base modification of the tRNA involves in the various cellular processes and affects codon-anticodon interactions (Agris et al., 2007). The post-translational modification of proteins is the well-explored area and several prediction methods have been developed whereas limited attempt has been made to understand and predict the post-transcriptional modifications. The tRNA modifications have direct role in the genome structure and codon usage (Novoa *et al* 2012) but only few rules have been established yet to discriminate modified and unmodified base in the raw tRNA sequences. We analyzed different kingdom and sub-cellular location wise modifications and developed tool for the prediction of uridines modifications (UMs) because this is most abundant in the tRNAs. There are different nucleotides preferred in the modified and unmodified uridines. We integrated binary and structural information and used in the SVM-based machine learning for prediction model development. We also developed separate models for the prediction of pseudouridine and

dihydrouridine. Based on these prediction models, a webserver called *tRNAmod* (http://crdd.osdd.net/raghava/trnamod/) has been developed. It has two different modules for the tRNA sequences and whole genome based predictions.

The AARSs play a central role in protein translation by covalently linking the correct amino acid to its cognate transfer RNA (Safro and Moor, 2009). Some AARSs perform editing activity to remove mischarged tRNAs, and this activity is important for the fidelity of the translation (Safro and Moor, 2009). There are twenty different AARSs found in all organisms, and each one is specific for single amino acid (Rajbhandary, 1997) and these AARSs further divided into two classes (class-1 & class-2), each containing ten enzymes (Eriani et al., 1990). We integrated PROSITE domain based information and compositional information to develop SVM-based method for the prediction and classification (class-1 & class-2) of AARSs. A webserver *icaars* (http://crdd.osdd.net/raghava/icaars/) has been developed to predict and classify AARSs. There are two different sets of cytosolic and mitochondrial AARSs present in the eukaryotic cells. In the process of horizontal gene transfer, mitochondrial AARSs genes were transferred into the nucleus, and now both cytosolic and mitochondrial AARSs encode from the nucleus, coexist in the cytosol and further mitochondrial AARSs transfer into the mitochondria (Duchêne et al., 2009). The experimental determination of the location of AARSs is labor-intensive and tedious. Therefore, developed a prediction method *MARSpred* (http://crdd.osdd.net/raghava/marspred/) using selected attributes of split amino acid composition (SA-SAAC) approach.

AARSs are interesting antibacterial drug targets and many natural compounds and antibiotics specifically target AARSs and inhibit the growth or survival of the target bacteria (Ochsner et al., 2007). While all essential twenty AARSs represent a potential drug targets, but they have not been exploited completely and only one marketed drug mupirocin is available against the IleRS of *Staphylococcus aureus* (salleRS). Therefore, it is indispensable to explore this protein family and design inhibitors against the AARSs of pathogens. In order to investigate the AARS family as a potential drug targets, we compared the complete human cytosolic AARSs (hcAARSs) and mitochondrial AARSs (hmAARSs) sets with the AARSs of four

different pathogens, *Haemophilus influenza* (hiAARSs), *Mycobacterium tuberculosis* (mtAARSs), *Staphylococcus aureus* (saAARSs) and *Streptococcus pneumoniae* (spAARSs). Ideally, potential drug candidate of pathogen's AARS should be different not only from the hcAARS but also from the hmAARS. The complete set of experimentally validated hcAARSs and hmAARSs were absent; therefore, we have predicted remaining AARS with the help of our previously developed algorithms of *icaars* and *MARSpred*. Initially we compared the full length protein sequences of human (both hcAARSs and hmAARSs) and pathogen's AARSs and then we extracted the catalytic domains from the complete protein sequences and again performed the domain-domain comparisons. Thereafter, we compared the structures of catalytic domains. The structural availability was low in the PDB; therefore, we modeled remaining structures using different homology modeling software. Finally, we found that mtArgRS, mtGlyRS, saArgRS, saHisRS, saIleRS, spIleRS, hiPheRS, mtPheRS, spPheRS, saPheRS, hiTyrRS, mtTyrRS, spTyrRS and saTyrRS are potential drug targets. It is important to explore the inhibitors against these potential drug targets. Therefore, we searched the experimentally validated inhibitors against these potential drug targets and found 45 inhibitors against hiPheRS and 42 inhibitors against spPheRS (Montgomery et al 2009). We have developed QSAR based prediction models, where first we calculated descriptors from these inhibitors using different software such as VLife, PaDEL etc. Further we selected the only those descriptors, which are highly correlated with the IC_{50} or pIC_{50} value of inhibitors using attribute selection of WEKA software. Finally, we achieved 0.83 and 0.94 correlation value using only 20 and 12 descriptors of PaDEL for hiPheRS and spPheRS respectively. These prediction models have been implemented into the form of a web-server called *iPheRS* (http://crdd.osdd.net/raghava/iphers/).

To conclude, in this thesis a systematic study of different predictions such as noncoding RNAs predictions, prediction of protein-RNA interaction, prediction of tRNA modifications, prediction of AARS and their sub-cellular location and prediction of inhibitors against the pathogen's AARSs, have been carried out. The success of GNU project motivated many biologists to contribute in the open-source bioinformatics. This is a philosophy to distribute biological software freely and make them available for all the users globally. World Wide Web (WWW) is the most

powerful way to provide these services and software into the public domain. Therefore, we implemented all the prediction methods into different web-servers, and all are freely available for the service of global scientific community. We hope that these prediction methods will be useful in the solving different biological problems.