Summary and future prospects

ıcer, an important health concern for the both developing and developed countries ls to intense social and monetary consequences. 14.1 million new cases of cancer nearly 8 million deaths due to cancer have been reported in 2012 across the world .UK, 2014). In India, these numbers are alarming with around 1 million new cases around seven lakhs deaths due to cancer in 2012 (Mallath et al., 2014). There is a need to review and improve the therapeutic techniques to offer appropriate and rdable treatment of cancer in both developing and developed nations.

personalized therapy has come with the hope to make more successful cancer tment. Personalized treatment is the targeted therapy that aims at the right delivery lrug to the right patient at the right time (Verma, 2012). The dawn of next :ration sequencing gives insights for the development of personalized medicine. A hora of information is being generated to study the molecular mechanisms of :er. Besides conventional chemotherapy, other therapeutic mechanisms are also :loped to combat the limitations of conventional approach.

A interference (RNAi), which is used to suppress gene expression as well as ated genes, has exciting potential for cancer therapeutics (Pecot et al., 2011), gh problems remain for the efficient of the small interfering RNA (siRNA) :cules. Also, genome-wide RNAi screens have been analyzed with NGS-driven (Sims et al., 2011), so there is a lucrative scope of investigating these data to get ghtful information.

ımmary, this research work emphasizes on mining relevant biomarkers for cancer nosis and therapeutics from genomic information resources like TCGA, MIC, and CCLE. A large amount of information present in these databases, rated from various experimental sequencing techniques, can act as the goldmine lucidating molecular mechanisms of cancer, detecting cancer diagnostic markers leveloping therapeutic mechanisms for tailor-made treatment.

first section of this thesis comprises deriving critical biomarker genes, playing ortant role in cancer mechanisms, in different groups of patients based on various cal features like the age, sex, tissue type and germ layers from which cancer e has been originated. Analyses of set of highly expressed genes and rexpressed genes covering most of the patients in different categories indicate the

preference of differentially expressed genes in particular stage or cancer type. We have attempted to select a panel of six genes from each category that covers maximum number of patients from that category.

In the next part, an attempt has been made to identify important biomarker genes that can discriminate early and late stages of various cancers on the basis of RSEM values derived from high-throughput RNA-seq data. We applied three machine-learning classifiers on six selected cancers available from the TCGA including bladder urothelial carcinoma (BLCA), colorectal adenocarcinoma (COAD), breast invasive carcinoma (BRCA), clear cell kidney carcinoma (KIRC), skin cutaneous melanoma (SKCM), and thyroid carcinoma (THCA). The key strength of this study is that machine-learning techniques have been applied to a large number of cancer samples with the panel of selected features (less than hundred genes) to differentiate early and late stages of cancer on the basis of gene expression profile. We found that Random Forest classifier is consistent in classifying the two stages in most of the cancers. We were able to distinguish early and late staging with satisfactory performance in six types of malignancies. We have implemented the models in a web application 'CancerSP', available freely at http://crdd.osdd.net/raghava/cancersp. Future prospects of above two sections may include statistical validation of the gene biomarkers in different subgroups. Further, these candidate genes can be validated experimentally to confirm the biomarker status of these genes in large population size.

The next part involves the development of the resource (CSiRdb) of cancer-specific potential siRNAs. CSiRdb catalogs the cancer-specific siRNA sequences that have been found potentially efficient against the mutated fragments (neonucleotide regions) of 144 cancer gene targets absent in normal condition. In addition to the detailed analyses of each target, their siRNA and different types of mutations, gene essentiality data from COLT is also incorporated. Since this work is based on the cell line data, hence to make it more pragmatic, we have implemented partially and fully personalized modules to handle patient level data. In future, the siRNA stored in the repository can be taken to experimental level to check their efficacy *in vitro* or *in vivo*.

The subsequent part of the thesis explains the development of the advanced tools for the improvisation of the siRNA-based therapy including prediction of siRNA efficacy based on heterogeneous data and studying immunotoxic effects of siRNAs. Here, we

have developed models for predicting siRNA efficacy trained on siRNAs tested in diverse experimental conditions and systems. In past, different tools for siRNA efficacy have been designed. The unique feature about this tool is that it contains the heterogeneous data, and we have been able to achieve a decent correlation of 0.58 with a significant reduction in feature space up to 64. These models are provided to the scientific community in the form of a web server, GMPesi, where the user can predict the efficacy of siRNA before testing in wet lab experiments. Future work with respect to this work involves improving the efficacy prediction using more parameters in feature space including thermodynamic or structural information. Further, robustness of the models can be validated in more diverse and heterogeneous dataset.

Next section discusses the work related to the immunological effects of siRNAs. We have developed a platform to predict immunomodulatory potentials of siRNA using sequence information. The RNA sequences collected from literature and RNAimmuno database (Olejniczak et al., 2012) were taken as immunomodulatory oligoribonucleotides (IMORNs) while the miRNAs found to be circulating in the human fluids were taken from the miRandola database (Russo et al., 2012) and considered as Non-immunomodulatory sequences (Non-IMORNs). We have found that pentanucleotide composition can be used to differentiate two classes of siRNAs. The section was concluded with the development of a web server, imRNA, where user can design siRNA with the desired immunotoxicity.

siRNA therapy is impeded by the blockades for siRNA to reach their proposed targets inside the cell that in turn affects the gene silencing activity (Zorde Khvalevsky et al., 2013). So there is need to develop efficient delivery systems with minimum side effects. Diverse macromolecules have been delivered by cationic cell penetrating peptides inside the cell (*viz.* antisense oligonucleotides, antibodies, peptides, and plasmid DNA). Unlike utilization of the traditional endocytotic pathways, CPP mediated siRNA have been shown to directly enter the cell (Meade and Dowdy, 2007). So, next section is focused on the development of the manually curated storehouse of different CPPs reported in the literature. The resource 'CPPsite' is supported by various analysis tools apart from primary and secondary information about each of the 843 entries (Gautam et al., 2012). It is the first repository that structured the information of CPP and dedicated to scientific community. We utilized the CPPsite and found that CPP are different from other peptides in their sequence,

structure and motif information (Gautam et al., 2012) . Using the above information, we developed CellPPD web server, which can predict highly efficient CPPs with >90% of accuracy (Gautam et al., 2013). Composition of arginine was observed to play a significant role in discriminating the peptides. CellPPD also has a facility for designing of CPPs in such a manner that single mismatch mutation can be optimized on the web server to achieve CPP sequence of desirable properties. Protein scan facility is useful for identifying CPP-like regions in a query protein. In future, further attempts can be made to check the efficiency of siRNA delivery by doing docking studies involving siRNA and peptides with molecular dynamics simulations.

In the thesis, we have touched upon the different angles relevant to design better siRNA therapeutics for personalized therapy. We performed analyses of biomarkers in different cancers to discriminate the early and late stages of the disease, designed cancer-specific siRNAs, developed generalized prediction algorithm for siRNA designing using most heterogeneous data, explored immunomodulatory effects of siRNA and designed the better CPPs for efficient delivery of the molecules. In the course, we have developed different resources and prediction servers that are in public domain for the scientific community. We anticipate that sincere efforts put in this study would bring fruits with the validation of suggested/predicted siRNA in real life.