6. SUMMARY

The present dissertation was directed towards exploration of Vaccine Adjuvants, a relatively untapped aspect of vaccine development that might play a major role in the success of subunit vaccines, if significant advances in the understanding and design of adjuvants are made. Currently, the search for novel adjuvants is mostly hypothesis-driven with a foresight that the immunomodulatory properties of molecules, specifically the capabilities of activating the innate immune responses could be modified to make them applicable as vaccine adjuvants. This is in contrast to the earlier methods of searching adjuvant candidate molecules that were more empirical in nature emanating trial-and-error efforts for discovering novel adjuvants. The end of the twentieth century saw a surge in the mechanistic understanding of the mammalian innate immune system especially the role of Pattern Recognition Receptors (PRRs) and their ligands in shaping the overall immunogenic response to the pathogenic attack. Lata and Raghava (Lata and Raghava, 2008) compiled the PRR knowledge in the form of a database while a comprehensive database, InnateDB, harboring all the genes and proteins involved in the innate immune reactions came much later (Lynn et al., 2008). A more focused informatics work on vaccine adjuvants began only in 2012 with the arrival of Vaxjo (Sayers et al., 2012), the first database dedicated to the molecules in clinical use or under experimentation.

Lacunae in the rational approaches for development of vaccine adjuvants as well as the absence of computational aide tools engendered the current study where we made attempts to develop *in silico* methods of rational adjuvant design. The motive for taking up such a task of tool development was to impart an impetus to a scientific field that has been slogging for over a century in its strife to mine candidate molecules from the nature for advancing the therapeutic capabilities of shaping desired immune reactions against infections, allergies and cancer. In the first half of the dissertation, the primary focus was to train and develop

machine-learning models for predicting immunomodulatory molecules that may be adjuvant candidates and package these methods into web-based informatics tools available freely to the scientific community for use. The latter half of the study was aimed at demonstrating the how these tools are capable of supplementing the rationality of the epitope-based vaccine design.

Numerous tools have been developed for predicting antigenic/immunogenic regions in antigens/proteins such that a choice of desired immune responses is available with different kinds of epitopes. For example, tools have been developed for predicting B cell epitopes (CBTOPE, Bcepred, ABCpred, LBtope, IgPred)(Haste Andersen et al., 2006; Kringelum et al., 2012; Liang et al., 2010; Rubinstein et al., 2009; Wee et al., 2010) and T cell epitopes (ProPred I, MHC2Pred, IL4pred, IFNepitope, CTLPred) (Nielsen and Lund, 2009; Peters et al., 2005; Reche et al., 2004; Zhang et al., 2010). Majority of such tools are focused on the prediction of immunological outcomes based on the adaptive immune system. Very few immunological *in silico* tools are concerned with the innate immune system such as the InnateDB (Lynn et al., 2008) and the PRRdb (Lata and Raghava, 2008).

In the present work, a primary focus was on the ligands of the PRRs, which are specialized receptors of the innate immune system. We selected two classes of biomolecules, namely the oligodeoxynucleotides (DNA) and the peptides, in our study due to the massive scale of diversity these molecules offer in terms of the nucleotide and amino acid sequences for criteria-based exploration of potential therapeutic molecules. The prediction tools capable of identifying novel therapeutic molecules in these classes could be employed over the sequencing data that has amassed in colossal proportions owing to the 'sequencing rush' among the investigators in their pursuit of understanding the genomic basis of biological phenomena.

Accordingly, our study began with the collection of experimentally validated immunomodulatory sequences of oligodeoxynucleotides (IMODNs) from the existing literature. This formed the positive set of sequences to be used for development of prediction models aimed at identification of novel IMODNs. Imparting the ability to distinguish the IMODNs from non-IMODNs would require training of the models with negative sequences too. Hence, random fragments of the same sequence length range as the positive sequences were extracted from the human genome CpG islands. Preliminary observation revealed sequence-based significant differences between the positive and the negative sequences such as the preference for thymidine and thymidine-rich motifs in the IMODN sequences as compared to the non-IMODNs. These features were then used for developing the Support Vector Machine-based prediction models capable of identifying new sequences that may be IMODNs based on compositional features. The pentanucleotide composition-based model performed the best among the various models tested for their performance based on the sensitivity, specificity, accuracy and the Matthews Correlation Coefficient (MCC). Addition of the motif information to the pentanucleotide models improved the performance marginally. Finally, the best performing prediction models were framed into an *in silico* web-based platform freely available to the scientific community for searching novel IMODN-based adjuvant candidates using various tools incorporated into the webserver 'VaccineDA'.

The next step in our pursuit for contributing to the rational design of vaccine adjuvant was the exploration of peptide-based immunomodulators for their promising adjuvant potential. In this case too, only the experimentally verified immunomodulatory peptides (IMODPs) were considered for assembling the set of the positive sequences. For compiling the set for the negative sequences, sequences of peptides experimentally found to be circulating in the human serum were considered as non-immunomodulatory peptides (non-IMODPs). Sequence analysis indicated trends in the difference of amino acid compositions between IMODPs and

the non-IMODPs such as the abundance of arginine and paucity of leucine in IMODPs as compared to the non-IMODPs and human proteins in general. Taking clues from such sequence-based differences, these were used as features for developing SVM-based prediction models with capabilities of identifying novel sequences that may act as peptide adjuvants in vaccines. The prediction model taking the dipeptide composition as well as the motif information performed the best among various combinations of features checked for their prediction performance. For equipping the scientific community with a handy tool proficient in searching and identifying new peptides that may act as adjuvants, the prediction models were packaged into a web-based server called 'VaxinPAD'. The server provides a couple of tools for adjuvant discovery to the user.

The concluding work of this dissertation was directed at the demonstration of how adjuvant prediction tools could supplement the rational design of epitope-based vaccines. For this purpose, common pathogenic bacterial species were taken up and a computational pipeline was executed for searching novel peptide epitope sequences that may advance into epitope-based vaccines by experimental methods. Two kinds of proteins, namely the virulence factors and the proteins encoded from the essential genes, were taken as the vaccine targets. Virulence factors have been the primary targets for vaccine designing but this strategy suffers a setback when the virulence factors resemble the intermediates of the host cellular signaling pathways, a survival tactic used by the pathogen, which the investigators named as the 'Molecular Mimicry'. In such cases, targeting the proteins encoded by genes essential for the survival of the pathogen might offer better vaccine targets, although some essential gene proteins might also be the virulence factors.

The virulence factor proteins were taken from the 'Virulence Factor Database' or the VFDB (Chen et al., 2016) while the essential gene proteins were picked up from the DEG (Zhang et

al., 2004) or the 'Database of Essential Genes'. For the sake of comparison, proteins of only those bacterial species were taken for which the proteins were present in both the VFDB as well as the DEG. For 14 such bacterial species, VFDB had 1459 proteins while majority of the proteins for these organisms in DEG could be categorized into 4 groups, those present in the membrane (membrane proteins), those present in the envelope (envelope proteins), protein that are secreted (secretory proteins) and those which are involved in the molecular repair processes within the cell (repair proteins). Hence, the DEG proteins considered for the present analysis were 420 membrane proteins, 36 envelope proteins, 52 secretion proteins and 38 repair proteins. The pipeline for epitope identification used for searching vaccine candidates in these proteins consisted of a tool for prediction of MHC Class II binding epitopes (Propred), a cytotoxic T-lymphocyte epitope prediction tool (CTLpred), a linear B-lymphocyte epitope prediction tool (LBtope) and a tool for predicting peptide based immunomodulators (VaxinPAD) that may act as vaccine adjuvants. The pipeline was used for fishing two types of epitopes, the first category being that of predicted T-cell epitopes those are predicted MHC II binders as well as predicted adjuvants. The second category of epitopes searched from bacterial proteins was the predicted B-cell epitopes that were also the predicted MHC II binders as well as the predicted adjuvants. Using the default parameters, the predicted T-cell epitopes, B-cell epitopes, MHC II binders and the adjuvants were mostly in thousands individually but the motive of the present study itself was to have a reasonable number of the 'best' epitopes that could be experimentally tested. Hence, the aforementioned categories of epitope search seemed logical for rationally arriving at the epitopes that may be best in terms of their suitability of recognition by the immune system for evoking immunogenic responses as well as their reasonable number feasible for testing experimentally. Thus, the 1459 VFDB proteins yielded 3622 nonamer peptides predicted to be T-cell epitopes, MHC II binders and self-adjuvants as well as 606 peptides as predicted B-

cell epitopes, MHC II binding and self-adjuvants. On the other hand, 420 membrane proteins from the DEG provide 1307 predicted T-cell epitopes and 155 predicted B-cell epitopes that are also predicted MHC II binders and adjuvants through the pipeline used here suggesting these peptides as vaccine candidates. The 36 DEG envelope proteins when subjected to the epitope prediction pipeline, yield 104 predicted T-cell epitopes and 20 predicted B-cell epitopes with probability that these will bind MHC Class II molecules and would also be adjuvants. In cases of 52 secretory proteins and 38 repair proteins taken from the DEG, the predicted T-cell epitopes and B-cell epitopes were respectively 173 and 23 for secretory proteins and 138 and 44 for the repair proteins, all of them being also the positively predicted MHC II binders as well as adjuvants. In its final recommendation for vaccine candidates, this study concluded that 252 epitopes (9-mers) sourced from proteins belonging to 13 bacterial species might offer the best set of choice for the immunogenic components of vaccines, given the experimentally verified bacterial essential gene proteins as well as the virulence factors apart from the predictive power of the tools used in the prediction pipeline used.

The present study is focused on development of sequence-based tools for identification of novel vaccine adjuvants and designing of epitope-based vaccines. Currently, peptide mimics of therapeutic molecules are being evaluated as alternatives to minimize the adverse effects of the corresponding therapies. This strategy is being tested for immunomodulators too (Shanmugam et al., 2012). Future advancements in this area would rely on the structural investigation of the PRR ligands for designing their peptide mimics. Nevertheless, the known limitations of peptide molecules such as their stability and toxicity would accompany such explorations. Thus, the rational design of adjuvants, in future, would require deployment of diverse addressing the multiple aspects like biomolecular sequence, structure, stability, toxicity, metabolism and formulation.