Synopsis of the Thesis

Title: Protein structure and mutant phenotype prediction from analysis of mutant libraries by deep sequencing

by: Shruti Khare

Thesis Supervisor: Prof. Raghavan Varadarajan

Molecular Biophysics Unit,

Indian Institute of Science,

Bangalore-560012, India.

Proteins play a central role in all the biological processes. The immense diversity in protein structures and functions despite similar underlying composition is intriguing. The work presented in this thesis aims to provide a deeper understanding of protein structure-function relationships. It describes some techniques that were developed in order to probe these relationships. Chapter 1 provides a general introduction to the topics discussed in the thesis. Chapter 2 focuses on an important aspect of protein structures, the cavities. Although proteins are composed of regular arrangements of secondary structures, namely, α helices and β sheets, there are some irregularities. The packing density is not uniform throughout the protein resulting into formation of cavities. The role of cavities has been previously probed using mutagenesis studies. The mutations designed to fill the cavities were observed to improve stability and cavity creating mutations adversely affected the stability. Cavities are thus reported to be important contributors to stability. In chapter 2, we refine and benchmark a method for prediction of protein cavities based on molecular dynamics simulations.

The insights derived from the mutagenesis studies provide some basic understanding of substitution preferences in proteins. The exposed non-active site positions are more tolerant to mutations whereas, the buried positions are not. Introducing cavity filling and disulfide mutations in proteins have been demonstrated to improve stability. Engineering protein variants with improved stability has immense applications in biology. In chapter 3, we discuss an important application, i.e., immunogen designing. Surface glycoproteins of several viruses exhibit two conformations, namely the metastable prefusion conformation and the highly stable postfusion conformation adopted during the fusion of the virus with the host cell membrane. Stable immunogens exhibiting the prefusion conformation are promising

candidates for subunit vaccines. In chapter 3, we discuss immunogen designing for the respiratory syncytial virus (RSV).

In addition to stabilized mutants, temperature sensitive (Ts) mutants are another class of engineered proteins. The Ts mutants exhibit reduced activity levels above the permissive temperature. They have been extensively used in developmental biology. Ts mutants are excellent tools to modulate protein expression levels in cells. A model for prediction of Ts positions was developed previously. This model exploits residue hydrophobicity to infer residue burial in the structure solely based on the protein sequence. In the current work, we improved the accuracy of the model by incorporating structural information in the model. Chapter 4 describes the development and benchmarking of a server for the prediction of Ts mutants (TSpred). The TSpred server suggests a stereochemically diverse set of mutations at the putative buried positions which are would produce destabilization to different extents and at least one of them is likely to be Ts. The TSpred predictions were employed for designing live attenuated vaccine candidates for RSV.

The work thus far elaborates on factors contributing to protein stability and application of that information for rationally designing mutants to modulate protein structure and function. In order to gain a deeper understanding of the role of each protein residue in its function, simultaneous analysis of multiple mutants is essential. Site saturation mutagenesis techniques generate all nineteen mutations at each residue position of the protein and the mutant function is linked to a phenotypic readout like cell viability or binding to a ligand. The mutant libraries are deep sequenced using one of the available platforms like Illumina, SOLiD, 454 and Ion torrent, and analysed to estimate the relative proportions of each of the mutants in the library. Automated programs are necessary to analyse the large amount of data generated after deep sequencing. A pipeline for analysis of data generated from the Illumina sequencing platform is discussed in chapter 5.

A mutational sensitivity measure denoted as MSseq was previously derived for the Controller of Cell division or Death B protein (CcdB). The values of the MSseq parameter reflect mutant activity. The active or inactive phenotypes of various mutants were analysed as a function of residue burial. Additional insights about substitution preferences at buried positions were gained from this analysis. In addition to residue burial, the substitution preferences varied with the physico-chemical nature and the size of the wild type (WT) and the mutant side chains. The active site of the CcdB protein could be inferred based on the

trends in the mutational sensitivity values. We quantitated these effects and developed a model, detailed in chapter 6, for prediction of the mutant phenotypes using a fraction of the CcdB mutational data. The model was observed to perform better than two other machine learning based predictors, SNAP2 and SuSPect.

Chapter 7 describes an additional application of the mutational sensitivity data. By analysing the population distribution of the MSseq values, an empirical parameter, RankScore, was previously derived for each residue in CcdB. RankScore can be interpreted as a weighted average of the MSseq values. RankScore was found to correlate well with residue depth which measures the extent of burial of a residue. As the residue depths in the native structure correlated well ($r = 0.6$) with the RankScores, the residue depths in native-like models would also correlate well with RankScores. Based on this principle, native-like models could be distinguished from low quality decoys. In the analysis reported in chapter 7, we examine this methodology using decoy datasets for ~200 proteins. We also consider additional information like predicted secondary structure and model quality score to achieve better model discrimination.

Studies thus far describe analyses performed using single protein mutants. Furthermore, information about residue interactions is also important. During the course of evolution, as the maintenance of specific interactions is essential for protein function, residues participating in such interactions are either conserved or varied in a correlated manner. Several computational models analysing such correlations among mutations are available. Experimental techniques are also available for identification of the spatially proximal residues. In chapter 8, we analyse various computational programs using CcdB and Diacylglycerol kinase A (DgkA) proteins and the results are then compared with the available experimental data. Overall little overlap was observed between the predictions based on the natural sequence co-variation and the experimental data. Both the computational and experimental approaches can be applied in conjugation as they provide complementary information.

The analyses described in the current work would provide useful guidelines for rational design of mutations to modulate protein stability. This has important implications in immunogen designing. The tools developed as part of the current work can be applied for (i) rational designing of Ts mutants, (ii) the analysis of site saturation libraries, (iii) calculation

or prediction of substitution preferences, (iv) structure prediction using correlated mutations as constraints, or (v) protein model discrimination.

A small appendix section is also included in the thesis. Synonymous mutations with differential phenotypes were observed in our deep sequenced library. In order to analyse them further, we performed a multiple sequence alignment and analysed codon frequencies at different positions. However, only preliminary results are available and those are included in Appendix I section of this thesis.