Executive Summary of the Minor Research Project

APPLICATIONS OF DIRECTIONAL DATA IN BIOINFORMATICS

Proteins are workhorses of all living organisms. Understanding the functions of biomolecules such as proteins, DNA, RNA, etc. requires the knowledge of shape of these biomolecules. It is now well-established that 3-dimensional structure of protein determines its functioning. The prediction of Protein structure is the one of the important research problem in the field of structural bioinformatics. How one dimensional sequence of amino acids folds into 3-dimensional structure is still mystery to the life scientists. The $(\phi, \psi)$ angles are key parameters in determining the 3-D structures of the Proteins. Similarly, in Genome Analysis, organizations of genomes are also problem of interest to the life scientists. The modeling of conformational angles needs to be addressed through statistical point of view. In the present research, attempts have been made to propose bivariate circular distributions to model the conformational angles, which plays crucial role in the formation of 3-D structures of the proteins.

Two bivariate circular distributions on torus have been proposed in the present research. Their distributional properties are investigated and inference problems are studied. The proposed distributions are applied to protein data set and circular genomes data set as well. By looking into Ramachandran plots of various protein molecules, we found that bivariate skewed distributions could be most suitable candidate to model such data. Accordingly, we proposed a “Skewed Bivariate Circular Distribution” by applying the technique of perturbation. The technique of perturbation is popularly used in linear Statistical Data Analysis. The distributional properties of the proposed model are investigated and related inference problems are addressed. The parameters are estimated by the method of maximum likelihood estimation. The utility of the model is explored by applying it to the bivariate data related to the protein molecules.

In Genome Analysis, comparison of genomes is the most important aspect of studying homology between two species of organisms. The genome comparison can be made at various levels like Nucleotide content, compositional biases of leading and lagging strands in replication, dinucleotide frequencies, the occurrence of repeats, RNA structures, coding densities etc. In the present work, a distribution on the torus is developed. The proposed model is capable of
modeling skewness and bimodality of marginal distributions as well as joint distributions. The proposed distribution is a member of the exponential family of distributions. It can be easily extended to a distribution on hyper-torus for studying a group of homologous circular genomes simultaneously. Further, we have given an extension of this model that is capable of incorporating multimodal marginal distributions and used it for protein local structure modeling.

The proposed family of bivariate circular distributions retains several important properties such belonging to the exponential family, bivariate normal approximations for highly concentrating data. We study several important features like marginal and conditional distributions, expression for normalizing constant, conditions for modality, test for independence of this distribution. We finally show the effectiveness of this model by fitting it to the bivariate data of dihedral angles in a protein molecule, shared orthologs by a paired circular genomes and also derived test for independence of the organizations of circular genomes. The BVSM distribution can be efficiently used to model bivariate circular data obtained from other features of genome organization such as shared TFBSs, shared coding or non-coding genes and shared repeated elements between circular genomes.

We have applied proposed GBSM model to fit the bivariate circular data of shared orthologs by circular prokaryotic genomes. Orthologous gene means gene evolved from the same ancestor gene. Shieh et al (2011) have studied the distributions of shared orhologs by using BGVM distributions. They have analyzed data of nine prokaryotic bacterial circular genomes: Bacillus anthracis Ames (G-1), Bacillus anthracis sterne(G-2), Bacillus subtilis (G-3), Oceanobacillus (G-4), Staphylococcus(G-5), Lactobacillus(G-6), Clostridium(G-7), Thermotoga (G-8) and Sulfolobus (G-9).

The likelihood ratio test for bivariate symmetry is applied to the seven pairs of prokaryotic genomes. The genomes pairs (G-6, G-7) and (G-7, G-8) were found to be asymmetric. The marginal distributions of the proposed generalized bivariate sine model (GBSM) are flexible to model asymmetric and/or bimodal data. PP-plots are used to assess Goodness of fit test for the marginals. We applied likelihood Ratio (LR) test based on the distribution of the shared orthologs to test whether the organization of each pair of the prokaryote genomes are independent or not. If the hypothesis of independence is rejected then it indicates that their genome organizations are correlated. The values of LR-test statistic computed for each of the seven pairs of circular genomes. The hypothesis of independence is rejected
strongly for all the pairs of genomes except the pair (Thermotogo, Sulfolobus). These results are consistent with the results obtained by Shieh et al. (2011) by applying the Bivariate Generalized von Mises (BGVM) distributions. The proposed GBSM model was found better than BGVM model by Shieh et al (2011) for seven pairs of prokaryote genomes.

Mr. D. N. Kalange  
Principal Investigator

Mr. S. G. Phalake  
Co-investigator