Abstract

Post-genomic molecular biology embodies high-throughput experimental techniques and hence is a data-rich field. The goal of this research work is to develop bioinformatics methods to utilize publicly available biological data of green algae in order to produce new metabolic pathway knowledge and to aid mining of newly generated data. As an example of knowledge or hypothesis generation from network creation to network analysis, consider function prediction of biological molecules. Assignment of identification of enzyme function is a non-trivial task owing to the fact that one or more potential enzymatic protein may lead to function identification in pathway metabolic network and at the same time involvement of enzymes in different biological processes; depending on their centrality value in particular pathway biological system. The availability of different databases and various approaches of genome annotation will lead to some unorganized architecture of databases, as user point of view organization of annotation information must be data comprehensive and organism specific along with ease in mining without Internet support. The multiple biological data of specific organism are available in scattered format and there is a lack of integrative approach among these biological data. Such genome annotation databases lack in providing the integrative view of different biological entities- especially enzymes and metabolites in a specific network. Therefore, we need to find out different ways to represent biological data in network architecture. Here we apply data integration approach to provide rich representation that enables pathway names based text mining of biological data in terms of integrated networks and conceptual spaces. A new tool called MetAlgNet has been developed during this research work, which follows integrative approach. The publicly available green algae genome annotated data can be used to aid mining of important biological enzymes in metabolic networks. We developed an integrative bioinformatics approach that utilizes publicly available knowledge of enzyme-metabolites interactions, network topological analysis like betweenness, closeness and degree for assigning node importance with quantitative values. Unidentified protein must be assigned to a particular biological protein with the help of Support Vector Machine (SVM) methods. We have reconstructed a phylogenetic tree from a set of specific metabolic enzyme protein sequences with the help of ete2 module in our tool. We applied this approach to create a random network from repetitive observations of the biological data about 1. *Chlamydomonas reinhardtii*, 2. *Ostreococcus lucimarinus*, 3. *Ostreococcus tauri* and 4. *Volvox carteri*, which are stored in a standalone
annotation database. The application of our software revealed importance of role of potential enzymes in biological functions in view of network centrality values, which were calculated by various algorithms. The results provided in this thesis indicate that integration of heterogeneous biological data facilitates advanced mining of data to create metabolic pathway networks. The methods can be applied for gaining insight into functions of enzymes, metabolites and other molecules, as well as for offering interpretation of functional evolution of metabolites with help of topological analysis and reconstruction of phylogenetic tree from sequence data.

Summary of Chapters

Chapter 1 presents a bird eye view of the research problem and the basic idea about system biology analysis of genomic data with a combination of omics technologies.

Chapter 2 on metabolic network reconstruction focuses on the workflow of genome annotation and elaborative representation of 1) gene regulatory (2) protein networks and (3) metabolic networks.

Chapter 3 covers various kinds of reconstruction resources used in making a model. Its broadly explains various genome annotation databases used in the genome annotation process. This chapter describes annotation workflow for identification of biological functions from raw genome sequence data and in-house database architecture for storing data.

Chapter 4 talks about metabolic network evolution study and interpretation of results with organic evolution. The six classes of enzymatic protein sequence data are analyzed using the node degrees of orthologous proteins, which are crucial for understanding the evolution of orthologous proteins. The results demonstrated the evolutionary importance of the fatty acid syntheses and the photosynthetic system in algae.

Chapter 5 presents the organization of various functions of the MetAlgNet tool. Also, it demonstrates the process flow for reconstruction and provides technical details related to organism complexity for integration with biological data. The chapter also highlights the features of the tool ENZPRED for determining the proteins class predication of un-annotate protein. It is very useful for unidentified protein classification with class. This chapter covers data annotation details and information about construction of integrated pathway metabolic network along with discussion of centrality analysis function.
Chapter 6 covers metabolic network analysis which consists of dynamics network statistics, a degree of the network, Betweenness and Closeness values. The analysis represents fifty five biological pathway network construction views and respective identification of potential enzymes from each network. It also shows a rooted phylogenetic tree for each metabolic pathway network.

**Goal of this Research Work**

The goal of this research work is to

1. develop biological data integrated tools and analysis methods for metabolic pathway network construction of green algae species.
2. design an algorithm to create a random network from multiple annotated biological data of green algae.
3. develop a tool for prediction of enzymatic class from protein sequence with help of SVM.
4. perform different centrality analysis of metabolic pathways network in different green algae species.
5. reconstruct phylogenetic tree based on circle type.