Chapter 7

Concluding Remarks
Chapter 7

Concluding Remarks

Post-genomic molecular biology is a data-rich field of research, and a variety of such data are publicly available. This research work is aimed at developing biological data integrated tools and analysis methods for metabolic pathway network construction of green algae species. The research has been carried out for integration of heterogenous types of biological data by applying data mining techniques for systems biology. An attempt is made to contribute to the areas of data integration, visualization, mining, and study of network topology, which are important elements of systems biology.

The thesis presents an integrated database which is a consolidation of a number of heterogeneous biological databases of four different algae 1. *Chlamydomonas reinhardtii*, 2. *Ostreococcus lucimarinus*, 3. *Ostreococcus tauri* and 4. *Volvox carteri*, and a software system called MetAlgNet that enables retrieval and visualization of biological relationships across the annotated data sources. The work uses repetitive dataset observation, an approach to indentify interactions of metabolic enzymes, which is the basis of the construction of integrated networks. The creation of interactions network is through retrieval of data from multiple annotated databases, and the MetAlgNet software system allows visualization of the networks. Integrative text-based mining of the data from 24 various databases is facilitated by representing the annotated data as raw material for network construction, and visualizing the similarities using different python library.

The MetAlgNet-based data mining approach may facilitate discovery of novel or unexpected relationships among enzymes and metabolites, formulation of new hypotheses, data annotation, interpretation of new experimental data, and construction and validation of new network-based models of biological systems. Results presented in chapter 6. High-throughput experimental techniques of postgenomic era are poor at quality and reproducibility. Therefore, integration of multiple types of data is desirable for bringing more confidence into analysis. More than fifty five metabolic pathway analysis shows that on an average one to three potential enzymes in each metabolic pathway plays crucial role in reactions.
Our approach takes advantage of connectivity of different annotated metabolic data of respective green algae in heterogeneous interactome network constructed by MetAlgNet, and shows that connectivity-based approach is superior to traditional pathway analysis. The findings from this study establish the applicability of our network analysis strategy, and support the hypothesis that modeling of local network topology dynamics can be used as an effective tool to study the activity of biological modules. Also, omics data are ever expanding and this poses challenges to updating and mining of data. The data warehousing approaches for data integration are really useful and effective from user point of view. It is not possible to completely avoid these problems, but by taking standards-based approach to data integration, we can minimize the problem of data integration. The integration approach is still found missing in online biological data available with different databases. It is better to develop databases which are interconnected with specific groups of organisms. The diversity of the data and the fact that not all data sources adapt the standards forces us to create our own schemas. We adapted a combination of multiple approaches in data integration. Although we imported all the databases to the local warehouse, the individual schemas were kept intact. We created an additional semantic mapping with the help of Python cursor and SQLite database to facilitate resolution of entities across databases, which often doesn’t need to change even when a new data source is added. The integration of data across databases and sophisticated queries are handled using Python programs.

The results of the data integration techniques such as network creation to network centrality analysis presented in this thesis are all from four green algae. But the technique of data integration is applicable more broadly to any organism for which we have large scale genome annotation data availability. As enzyme identifiers are the central entities to data integration in our method, data mining shows different interaction databases that use consistent identifiers. We used basic biological pathway names as the standard identifiers for search and afterward created networks from mined datasets, mainly because search term directly works as primary identifier among different heterogeneous databases of the four different algae. It is well known that publicly available molecular biology data include many false positives and thus quality filtering of these data is essential. There is a need to make exhaustive automatic curation efforts as well as sophisticated mining approaches to address these problems. In this thesis, we tried to identify potential enzymes along with centrality score from various pathway networks. Usage of high quality source data would be enhancing the value of our approaches and predication of unknown enzymes.