SciForum
Mol2Net

MetAlgNet :Metabolic Pathway Network Reconstruction from Algae Genome Annotation Data

Kirtan Dave [1]*, DarshanChoksi [1], Hetalkumar Panchal [1]

[1] G.H. Patel P.G. Dept. of Computer Science and Technology, Sardar Patel University, Vallabh Vidyanagar, Gujarat, India; E-Mail: kirtandave11@gmail.com

* E-Mail: kirtandave11@gmail.com; Tel: +91-02692-236829; Fax: +91-02692-236829;

Published: 4 December 2015

Abstract: Post-genomic molecular biology embodies high-throughput experimental techniques and hence it is a data-rich field. The goal of development of this tool is to utilize free available biological data of green algae in order to produce new metabolic pathway knowledge and to aid mining of newly generated data. The variety of biological sequence and functional information are stored in different online database, so getting annotation information of genome from different database is challenging task for reconstruction of pathways. 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. 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. The application of our software is 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 work 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.

Keywords: metabolic networks, green algae, centrality, Python

1. Introduction

The advancement of new technology leads to production of large amount of biological data such as high-throughput sequencing data, metabolomics data, transcriptomics data and

many more.The metabolic networks are complex due to their size and the presence of bimolecular reactions; so combined knowledge of biology, computer science and graph theory will help understand molecular network complexity[1]. Within the biological sciences, one of the primary challenges is to investigate how the collective behavior of cells, tissues, or organisms can be understood in terms of the properties of their molecular constituents from a metabolic network [2]. There is an essential role of metabolic networks in all biological processes of a living cell. Some are like biochemical pathways to protein interactions and gene regulation to cellular communication. Traditionally, genes and proteins involved in different functionalities have been studied in isolation or in small clusters. However, the complex nature of a cell cannot be fully understood by studying individual components in isolation. To investigate this intricate connectivity of cellular systems, the analysis of complex networks has become an important part of molecular biology [3]. Cellular system can be viewed as a combination of omics technologies, data integration, analysis, mining, and visualization often involving use of these techniques iteratively over hypothesis driven systematic experimental design to gain increased understanding of the structure and dynamics of the biological systems [4]. In Fig-1 an integrative bioinformatics starts with the integration of multiple datasets from one or more omics and also possibly from multiple organisms, and forms the basis for systems biology analysis.

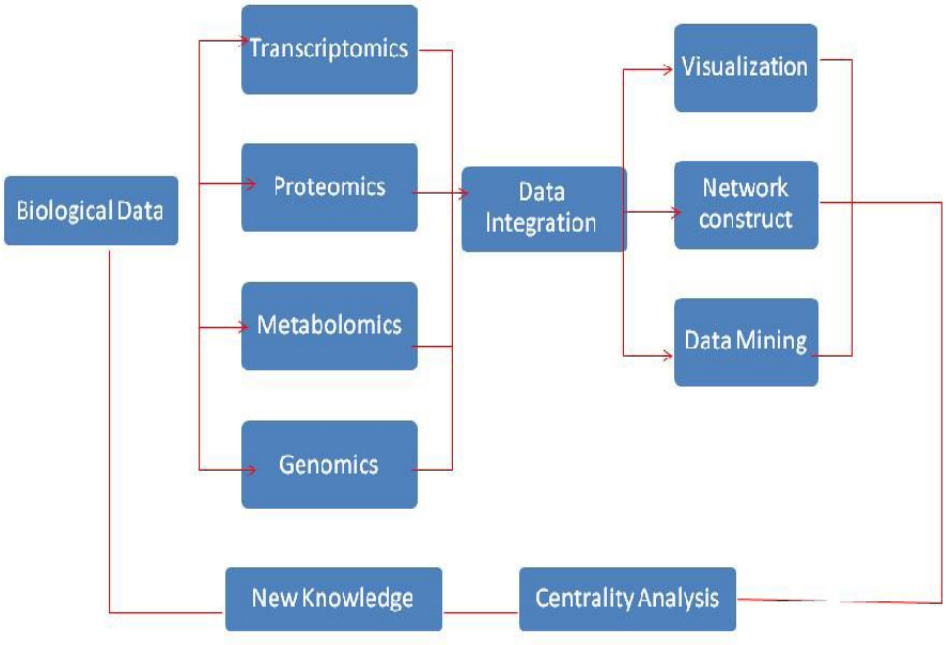

Figure 1. Basic schema for **MetAlgNet**

2. Results and Discussion

MegAlgNet data integration system facilitates mining of biological data and hence exploration ofsome useful patterns, novel relationships between different biological entities from the data, andmay provide novel insights into metabolic functions and context- specific biological functions.

The MetAlgNet creates a network from annotated data and the purpose for developing

this network is to get knowledge of potential element from topological analysis with generated network. The software created random network from GMT file (see Fig. 3).We created 55 different pathway networks from the standard biological pathway as per KEGG database; the main purpose of creating this network is getting inference from it. However, the tool has ability to generate more number of networks with respective search term.

Along with centrality, we also reconstruct phylogenetic tree from respective annotated data of particular pathway. Here, we summarize result of pathways which generated from MetAlgNet data mining that were four main result in consideration 1) Network generated from particular pathway from specific search term 2) identification of potential node of respective network with help of node ranking algorithm 3) degree, closeness and betweenness centrality calculation and bar chart generation and 4) phylogenetic tree generation from sequence data.The interpretation lead to identification of major role of particular enzyme in network and chemical compound. The networks given below are generated using 1.*Chlamydomonas reinhardtii*, 2.*Ostreococcus lucimarinus*, 3.*Ostreococcus tauri*and 4.*Volvox carteri*. So, collecting all data from each of the organism database tool, creates a comprehensive GMT file.

The GMT file is further utilized for creating a network of enzymes and metabolites. The resulting networks showed surprisingly high level of connectivity across different stages of linear metabolic pathways via enzyme and metabolite interactions. The centrality analysis plays major role to identify a potential node in the network. If the network has a very high average closeness value, it leads to more organized functional units or modules. The degree could indicate a central role in a biological network. It may indicate relevance of a node as functionally capable of holding together other nodes in the network. Betweenness of a node effectively indicates the capability of a node to bring in distant nodes to perform communication in network (see Fig 4) .

3. Materials and Methods

Primary requirement for annotation is collecting genome data of desired organisms. However, if complete annotation data is not available, so we can annotate data with available genome annotationpipeline. The raw data collected from NCBI sequence read archive database or DDBJ or EBI-SRA database[5-7] for 1.Chlamydomonas reinhardtii, 2.Ostreococcus lucimarinus, 3.Ostreococcus tauri and 4.Volvox carteri. In context of four different algae, there is list of data available, which we have downloaded and used in annotation. The major data used in making database are listed below in table 1. However, there are lots of incomplete data, so we try to avoid use in study[8-10].

We took permanent draft and complete sequenced data for study. The study also considers other source annotated data of respected algae. There are many community based genome annotation projects going on like OrcAE (Online Resource for Community Annotation of Eukaryotes) and Phytozome (JGI annotation resource). Both online databases provide very much useful annotation data which contain Gene locus name, transcript name, protein name, PFAM, Panther ID, KOG, EC NO, KEGG Orthology and Gene ontology[11].

Table 1. The various data numbers representing each database, 1. CRE (Chlamydomonasreinhardtii) 2. OLU (Ostreococcuslucimarinus) 3. OTA (Ostreococcustauri) and 4.VCN (Volvox carteri).

	Org_db	**Org_anno_db**	**Org_cds_db**	**Org_gene_db**	**Org_protein_db**	**Org_rxn_db**
CRE	113	19526	19526	3433	19526	4635
OLU	109	7796	7796	3146	1196	5624
OTA	108	6912	6912	3309	6912	4067
VCU	114	15285	14971	3484	14971	3915

Data from various public data sources were collected into our local database systems The curation of a free available data from database involves several steps required in the curation process.

Multiple molecular biology databases provide descriptions of biological systems at different levels of abstraction. Some common biological information, along with names of primary databases providing information is indicated in figure-2.

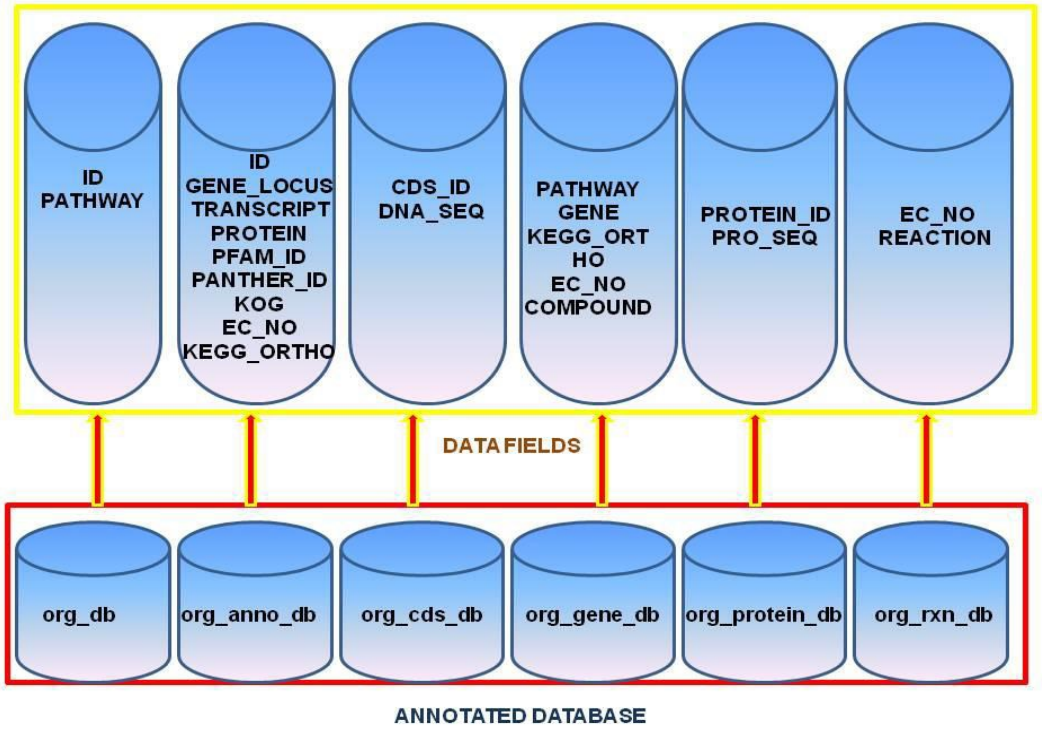

Figure 2. Overview of our in-house database architecture

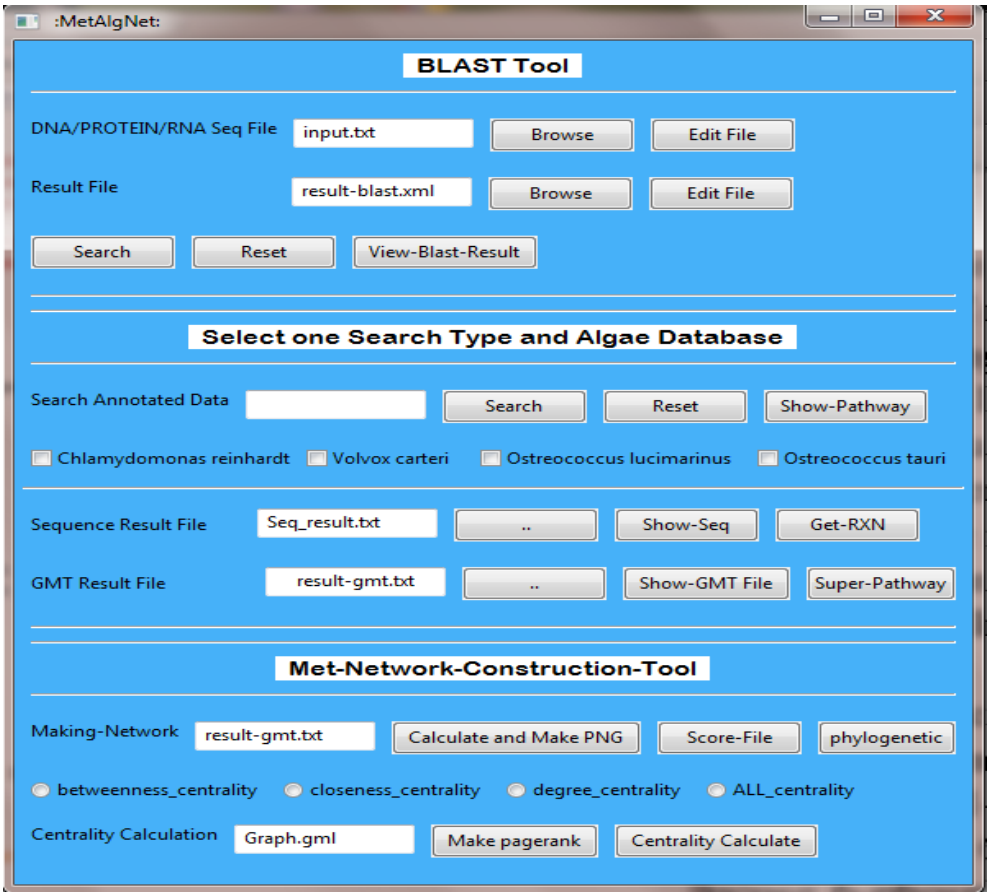

Figure 3. The MetAlgNet tool

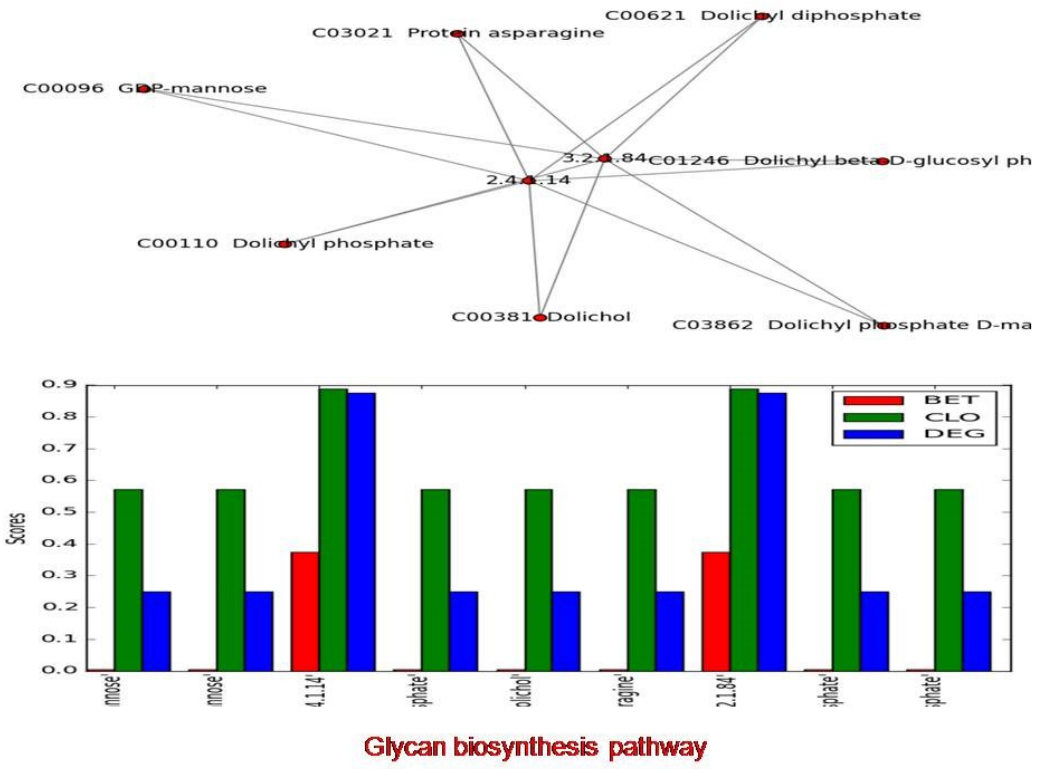

Figure 4. Sample result output

4. Conclusions

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. 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 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.

Acknowledgments

We special thanks to Jignesh V. Smart for debug the **MetAlgNet** Program. This work is supported by the Councilof Scientific & Industrial Research (CSIR) - Senior Research Fellowship (SRF) grant no. 112511/2k11/1, India at G. H.Patel Post Graduate Department of Computer Science and Technology (GDCST), Sardar Patel University.

Conflicts of Interest

"The authors declare no conflict of interest."

References

1. Chain, P.S.G.; Grafham, D.V.; Fulton, R.S.; FitzGerald, M.G.; Hostetler, J.; Muzny, D.; Ali, J.; Birren, B.; Bruce, D.C.; Buhay, C.*, et al.* Genome project standards in a new era of sequencing. *Science (New York, N.Y.)* **2009**, *326*, 10.1126/science.1180614.

2. Hwang, D.; Rust, A.G.; Ramsey, S.; Smith, J.J.; Leslie, D.M.; Weston, A.D.; de Atauri, P.; Aitchison, J.D.; Hood, L.; Siegel, A.F.*, et al.* A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 17296-17301.

3. Reeves, G.; Thornton, J.; Excellence, t.B.N.o. Integrating biological data through the genome. *Hum Mol Genet* **2006**, *15*, R81 - 87.

4. Christensen, C.; Thakar, J.; Albert, R. Systems-level insights into cellular regulation: Inferring, analysing, and modelling intracellular networks. *Systems Biology, IET* **2007**, *1*, 61-77.

5. NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Research* **2015**, *43*, D6-D17.

6. Kodama, Y.; Shumway, M.; Leinonen, R. The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research* **2012**, *40*, D54-D56.

7. Pruitt, K.D.; Tatusova, T.; Maglott, D.R. Ncbi reference sequences (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **2007**, *35*, D61-D65.

8. Palenik, B.; Grimwood, J.; Aerts, A.; Rouzé, P.; Salamov, A.; Putnam, N.; Dupont, C.; Jorgensen, R.; Derelle, E.; Rombauts, S.*, et al.* The tiny eukaryote ostreococcus provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104*, 7705-7710.

9. Merchant, S.S.; Prochnik, S.E.; Vallon, O.; Harris, E.H.; Karpowicz, S.J.; Witman, G.B.; Terry, A.; Salamov, A.; Fritz-Laylin, L.K.; Maréchal-Drouard, L.*, et al.* The chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **2007**, *318*, 245-250.

10. Prochnik, S.E.; Umen, J.; Nedelcu, A.M.; Hallmann, A.; Miller, S.M.; Nishii, I.; Ferris, P.; Kuo, A.; Mitros, T.; Fritz-Laylin, L.K.*, et al.* Genomic analysis of organismal complexity in the multicellular green alga volvox carteri. *Science* **2010**, *329*, 223-226.

11. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.*, et al.* Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* **2012**, *40*, D1178-D1186.

© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions defined by MDPI AG, the publisher of the Sciforum.net platform. Sciforum papers authors the copyright to their scholarly works. Hence, by submitting a paper to this conference, you retain the copyright, but you grant MDPI AG the non-exclusive and un-revocable license right to publish this paper online on the Sciforum.net platform. This means you can easily submit your paper to any scientific journal at a later stage and transfer the copyright to its publisher (if required by that publisher). (http://sciforum.net/about).