Challenges faced in the integration of biological information

Chapter 2

Challenges faced in the integration of biological information

MBI301-Dataming & Data Analytics

Mamta Sagar
Department of Bioinformatics
University Institute of Engineering & Technology, CSJM University, Kanpur

Life science discovery process

In the last half of the century, a highly focused, hypothesis driven approach known as reductionist molecular biology, gave scientist to identify and characterize molecules, the fundamental building blocks of living system

- Scientist not only have to continue reductionist strategies for further elucidation of the **structure & function of individual components**, but they also have to adopt a system- level approach in biology.

System analysis demand not just knowledge of the parts-genes, proteins and macromolecular entity-but also **knowledge of the connection of these molecular parts and how they work together**

- In other the pendulum of bioscience is now swing away from reductionist approach and towards synthetic approach
- Characteristic of system biology and of an integrated biology capable of quantitative and or detailed qualitative predictions

Information Driven Process

Genomics → Gene Expression profile →

→ Proteomics → System Biology

- In every step, database searches and computational analysis of data are an integral part of the discovery process
- As we choose complex systems for study, experimentally generated data must be combined with data derived from databases and computationally derived model or simulation for best interpretation

Modeling & simulation of protein- protein interaction, cellular process need more experimental observation to fill in missing quantitative details for mature efforts

An information integration environment for life science discovery

The basis needs are

1. On demand access & retrieval of most up to date data & the ability to perform complex query
2. Access of best of breed analytical tools and algorithms for extraction of useful information
3. A robust information integration infrastructure that connects various computational steps

The nature of Biological data

- Diversity: ranging from gene & protein-pathways-networks-cell tissues-organism population
- Variability: Different individuals & species vary tremendously. For ex. Structure & function of organs vary across age & gender

Data sources in life science

- Nucleic acid, protein
- Biological data base are autonomous
- Heterogeneous
- Dynamic
- Computational tools are required specific I/O formats and broad domain knowledge

Challenges in information integration

- Heterogeneity
- Biological Data Complexity

Data integration (fundamental function features &)

1.Accssing & retrieving relevant data from a broad range

2. Transforming the retrieved data into designated data model for integration

3. Providing a rich common data model for abstracting retrieved data & presenting integrated data

4. Providing high level expression language to compos complex query & to facilitate data manipulation, transformation and integration task

5. Managing query optimization & complex issues

Data warehouse

- Assemble data sources into a centalized sysyetm with a global schema & indexing system for integration & navigation

Federation approach

- Do not require centralized system Underlying data remains autonomous
- Maintain a common data model& rely on schema mapping to translate source db schema into a target schema
- A data dictionary is used CORBA & OMG: to encapsulate the heterogeneity & to facilitate interoperation of disparate components

Mediator approach

- Introduce & intermediate processing layer to decouple underlying heterogeneous distribute data sources and client layer of end user & applications
- Mediator layer is collection of software components performing the task of data integration
- Most dbase mediator use a wrappers layer t o handle the task of data access, retrieval, & translation

Advantages of mediator approach

- Support a high level query language for data transformation & manipulation, facilitate the composition of complex query
- Flexibility, scalability & modularity: handle data source schema changes
- OPM,IBM discovery link, TAMBIS, TSIMMIS

Meta data specification

Provide documentation on other data managed within an application

Data provenance & data accuracy

Move to the next stage of development, more & more secondary db with value added annotation will be developed.

Question ?

What are the differences b/w federation & mediator approach of data mining?

References

- Bioinformatics: Managing Scientific Data Z. Lacroix, Lacroix Zoe Critchlow Terence, T. Critchlow Published 2013 Computer Science