HITWA08 - Health Information Technology Forum
26 November 2008

Current State of Bioinformatics in Australia

Dr. Amandeep S. Sidhu
Centre for Comparative Genomics, Murdoch University
http://ccg.murdoch.edu.au/

Overview

• What is Bioinformatics?
• Key Bioinformatics Groups and Activities
• Key Issue: Biomedical Data Integration
• Integration Bottleneck: Ontologies
• Open Issues
What is Bioinformatics?

Bioinformatics is the science for storing, managing, analysing, visualising and integrating various digitised biological information – to aid in decision making.

Data stockpiles
- Genomics, gene expression, proteomics, metabolomics, gene networks, hospital data, population data, etc.

Data Analysis and Mining
- New discoveries through data integration
- Sharing/communicating data via the internet
- New computational methods for data analysis

Impact on society
- Hospital Management
- Medical breakthroughs
- Improved decision making
A Biomedical Informatics Perspective of Life Science Processes

Key Groups and Activities
3rd Health Information Technology Forum (HITWA08) 'E-Health Landscapes'
26 November 2008, Perth

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian Genome Research Facility</td>
<td><a href="http://www.agrf.org.au/">http://www.agrf.org.au/</a></td>
</tr>
<tr>
<td>Australian Proteome Analysis Facility</td>
<td><a href="http://www.proteome.org.au/">http://www.proteome.org.au/</a></td>
</tr>
<tr>
<td>ARC Centre of Excellence for Bioinformatics</td>
<td><a href="http://www.bioinformatics.org.au/">http://www.bioinformatics.org.au/</a></td>
</tr>
<tr>
<td>Australian e-health Research Centre</td>
<td><a href="http://www.e-hrc.net/">http://www.e-hrc.net/</a></td>
</tr>
<tr>
<td>Queensland Facility for Advanced Bioinformatics</td>
<td><a href="http://www.qfab.org/">http://www.qfab.org/</a></td>
</tr>
</tbody>
</table>

**CCG’s Spectrum of Activities**

**Comparative Genomics (Crop plants, Production Animals, Human, Pathogenic Bacteria)**

**Outputs**
- Identify important biological features/genome characterisation/evolution
- New Markers for mapping agronomic/disease traits
- Vaccine candidates
- Gene silencing

**Outcomes**
- Generation of new knowledge
- Improved speed of identification
- Improved speed to incorporate into breeding programs/therapeutics/diagnostics

**Integrated Systems development**

**Outputs**
- Web-based (secure) systems
- Integration of diverse data sources
- Novel visualisation and analysis tools
- Grid computing

**Outcomes**
- Generation of new knowledge
- Ability to conduct novel R&D
- Access/communication/curation
CCG Team consists of
• Research Scientists
• Bioinformaticians
• Bioinformatics Officers
• Software Developers
• System Administrators

Key Issue: Biomedical Data Integration
Example: Integrating publicly available Asthma genomic data

- Internet resource 1:
  - NCBI 58 asthma associated (genome regions) QTLs
- Internet resource 2:
  - NCBI 144 asthma annotated genes
- Internet resource 3:
  - NCBI many GEO datasets asthma gene expression studies
- Internet resource 4:
  - NIH Genetic Association DB 46 positively associated asthma genes
- Internet resource 5:
  - PubMed 50 microarray and asthma related papers
Initial questions

- Do the ‘NCBI’ annotated genes co-locate with the Asthma QTLs?
  - What about the NIH Genetic Association DB 46 positively associated asthma genes?
- Do any of the 58 microarray gene expression studies identify common genes that are upregulated/downregulated?
  - Are the non-randomly clustered?
- Are there any available bioinformatics resources that integrate all this information or allow a researcher to do so? How to go about this?
Asthma Genomic data mapped to the Human Genome

Human Genome Mapped Asthma QTL
Nature of Biomedical Data

- The biological data sets are organized in loose hierarchies that reflect our understanding of complex living systems.
- They range from plain text of laboratory records, nucleic acids and proteins, three-dimensional structures of molecules, biomedical images, to various experimental outputs from technologies in laboratories and hospitals.
- Data from these systems is incomplete and often inconsistent which presents challenges in informatics, modelling, and simulation.
- Furthermore, biomedical research is an expanding phase, and many fields of biology are still in the developing stages.
Inconsistency in Biomedical Data

- Creators of biological data sources use data descriptors with specific intended meanings in their own contexts.
- Heterogeneity arises because different biological data sources have different vocabularies and semantics.
- The creator of a data source structures the information in the source such that the interests of the creator and the immediate community are best satisfied.
- The creator is frequently less concerned about the use of the data source by a secondary system or community.

Problem of Biomedical Data Integration

- Typically, researchers require integrated access to data from multiple domains, which requires resolving terms that have slightly different meanings across the communities.
- This is further complicated by the observations that terminology is being used either is not explicitly identified or the terminology evolves over time.
- Data access methods change regularly. These changes are an attempt to keep up with the scientific evolution occurring in the community at large.
Data Integration Solutions

Integrated Data Retrieval

- Data Integration Solutions can be divided into three major categories:
  - Link-driven approaches using hyperlinks
  - View integration approaches using standardized schema
  - Semantic integration using common vocabulary
Need for Common Languages

• Mechanisms of intercommunication between databases of different structure and format is the need for common semantic standards.

• As noted by Hafner (Hafner and Fridman, 1996), general biological data resources are databases rather than knowledge bases: they describe miscellaneous objects according to the database schema, but no representation of general concepts and their relationships is given.

• (Schulze-Kremer, 1998) addressed this problem by developing ontologies for knowledge sharing in molecular biology.

Ontologies in Biomedicine & E-health
Ontologies

- The term ontology is originally a philosophical term referred to as “the object of existence”. The computer science community borrowed the term ontology to refer to a “specification of conceptualisation” for knowledge sharing in artificial intelligence (Gruber, 1993).
- Ontologies provide a conceptual framework for a structured representation of the meaning, through a common vocabulary.
- Vocabulary usually includes concepts, relationships between concepts, and possibility of defining ontology rules and axioms.
- Ontologies in biomedicine have emerged because of the need for a common language for effective communication across diverse sources of biological data and knowledge.

The Gene Ontology

The Gene Ontology (GO) is a collaborative effort to create a controlled vocabulary of gene and protein roles in cells, addressing the need for consistent descriptions of gene products in different databases.

http://www.geneontology.org/
The Protein Ontology

The Protein Ontology (PO) provides a unified vocabulary for capturing declarative knowledge about protein domain and to classify that knowledge to allow reasoning.


The SNOMED Ontology

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) Ontology includes “a Core terminology of over 364,000 health care concepts with unique meanings and formal logic—based definitions organized into hierarchies.

http://www.ihtsdo.org/snomed-ct/
The Health Records Ontology

Health Records Ontology is an effort to utilize SNOMED terminology as a basis & extend it to define a minimal set of concepts to integrate diversity in various formats to achieve a standardization in communication for healthcare delivery.

Open Issues
CCG Projects

- Plant Breeding Web (PB Web)
- Wheat Genome Sequencing and Annotation
- Protein Ontology (PO) Project
- Truth in Labelling of Fish
- Beef CRC Project
- Health Records Ontology (HRO) Project
- Yet Another Bioinformatics Interface (YABI)
- ... and more
Wheat Genome Sequencing and Annotation

Truth in Labelling of Fish
Data Integration Website for Program 3: Adaptation and Cattle Welfare

An interactive resource where Beef CRC researchers can upload and view integrated data. Click on a Project Box in the image below to upload data for integration to a Project Page. Click on the Integrated Data Genome Overview box to view the integrated data. Click for help.

---

**Beef CRC Project**

**Data Integration Website for Program 3: Adaptation and Cattle Welfare**

- An interactive resource where Beef CRC researchers can upload and view integrated data.
- Click on a Project Box to upload data for integration to a Project Page.
- Click on the Integrated Data Genome Overview box to view the integrated data.

---

**Yet Another Bioinformatics Interface**

- A Web 2.0 Internet application
- Simple design: drag & drop concept
- Simplifying
  - omics workflow creation
  - Customisation
  - Data analysis
- Design your own workflow