Biological Data Management in a Dataspace Framework

Victor M. Markowitz

Lawrence Berkeley National Laboratory
Biological data management

- **Characteristics**
  - Data integration is **inherent** to biological data management
  - Systems need to be devised within
    - Application specific context
    - Time, cost, usability constraints

- **Problems**
  - Integration technologies have not lived up to initial hype *
    - No effect on need for & cost of database and application domain expertise
  - Poor data management practices hinder integration **
    - Lack of comprehensive documentation on data, methods, systems
  - Continuous growth of data types & resources ***

* **SIGMOD 2005** Session on Enterprise Information Integration: Successes, Challenges, Controversies
** **WDMBIO 2003** Workshop on Data Management for Molecular & Cell Biology Working Group on Data Integration
*** **Molecular Biology Collection: 2006 Update** (*Nucleic Acid Research*) lists over 850 bio databases, and increase of 130 compared to 2005.
Biological data management revisited

- **Knuckles & Nodes**
  - Bottom-up correlation of bio data resources ("nodes")
    - No incentive /funding outside application specific context
    - Correlation may differ depending on application

- **Bio Dataspace**
  - Bio data resources seen as ("co-existing") components in application specific context
    - Data management systems developed gradually: allow coping with evolution of application needs
  - Key elements
    - Dataspace catalog
    - Search & query
    - Data lineage

---

* "Integrating Biological Databases", Lincoln Stein, Nature Reviews 2003

** "From Databases to Dataspaces: A New Abstraction for Information Management”, M. Franklin, A. Halevy, D. Maier, SIGMOD Record, Dec 2005
An Example:

**Microbial Isolate Genome Data Management & Analysis (IMG)**

**Microbial Community Metagenome Data Management & Analysis (IMG/M)**

- IMG Genomes
  - 296 Bacteria
  - 263 Bacteria
  - 24 Archaea
  - 9 Eukarya
  - 347 Plasmids
  - 1661 Viruses

- IMG 1.0
- IMG 1.1
- IMG 1.2
- IMG 1.3
- IMG 1.4
- IMG 1.5
- IMG 1.6
- IMG 2.0

- Mar 05
- Jun 05
- Sep 05
- Dec 05
- Mar 06
- Jun 06
- Sep 06
- Dec 06
Microbial genome data views

Integration: via genes using (1) existing annotations  (2) gene relationships

1. Pfams (Pathways) in Genome 1
2. Gene interactions in Genome k

Database View
1. Gene - Pathway - Function - Genome - Scaffold
2. Orthologs - Paralogs - Homologs

Application View
- Pfams (Pathways)
- Genes
- Functions
- Pathways
- Genomes

Phylogenetic Profiler
- Find Genes
- With Homologs
- Without Homologs
- Ignoring Taxon

Example:
- Bacteria
- Proteobacteria
- Burkholderia
- Burkholderia mallei ATCC 2334H
- Burkholderia pseudomallei K96A
Microbial genome data management-- different resources implement data integration that gradually increases the coherence and completeness of microbial genome data (sequences & associated functional annotations)

evolving microbial genome data space
**Microbial genome data space**

### Data Generation/Processing

- **Data Processing Pipeline**
  - Sequence Data Files → Assembly → ORF Calling → Preliminary Functional Annotation

- **Data Review & Curation**
  - Final Review
  - Data Curation
  - Data Review
  - Revised Annotation Data Files

### Data Collection/Integration

- **Genomic**
  - Genbank
  - EMBL
  - RefSeq
  - EBI GR

- **Primary Resources**
  - KEGG
  - COG

- **Secondary Resources**
  - UniProt
  - CDD

- **Functional Resources**
  - Resource Consolidation

**Data generation/processing**— Sequence data & annotations generated by individual sequencing centers/providers using various

- Sequence assembly/finishing, gene prediction methods
- Data quality control procedures
Data collection-- Sequence data & annotations from individual sequence centers/providers are collected into an *archival* primary genomic data repository

chemist Common data format provides supports for (but does not ensure) syntactic consistency
Data collation-- Basic annotations (e.g., gene product names) are revised in the context of other (all) sequence data & annotations for semantic consistency and provided by secondary resources.

Sequence clusters provide context for consistency validation/ curation.
Microbial genome data space

Data integration (light)-- Sequence data & annotation integration relies on genes co-related via functional roles determined using existing functional classifications

- Functional categories provided by resources, such as COG, KEGG, UniProt
Data integration (tight) -- Sequence data & annotation integration relies on gene sequence similarity, gene clustering

 PROVIDES CONTEXT FOR EXPANDING FUNCTIONAL CHARACTERIZATION OF GENES (METABOLIC RECONSTRUCTION)
Microbial genome data space

Data space catalog -- Information on components and their relationships in the context of a specific application

- Components can be added, removed, or change roles
Summary

- **Dataspace**
  - Concept helps reasoning about biological data management:
    - Different systems seen as related components in an evolving context
    - Support for data provenance
  - Framework helps coping with data content growth, changing application needs

- **Desired effect**
  - Documentation of individual resources in the context of application specific bio dataspaces improves their usability
  - Support reuse of components across bio dataspaces