

Genome Information Systems

euGenes and FlyBase

and other molecular biology databanks

Ethoinformatics Workshop
19-20 April 2002
CISAB, Indiana University, Bloomington, IN

Don Gilbert
gilbertd@bio.indiana.edu

Overview

- Bio-information warehousing and distribution
 - IUBio Archive, <http://iubio.bio.indiana.edu/> -- public molecular biology software archive
 - Bio-Mirrors, <http://www.bio-mirror.net/> -- Sequence and related biology databanks
- Genome information systems
 - FlyBase, <http://flybase.bio.indiana.edu>, genome infosystem for *Drosophila* fruitfly
 - euGenes, <http://iubio.bio.indiana.edu/eugenesis/> infosystem for human, fly, worm, and other complex genomes, genome maps, query and retrieval examples
- *New*: Bio-Data Grid
 - <http://iubio.bio.indiana.edu/grid/> distributed computing for bioinformatics

History at IUBio

- IUBio Archive for biology software and data 1989-
 - early Internet biosequence information search and retrieval; similar to SRS, NCBI's GenBank (Entrez), Medline
 - GenBank biosequence search using WAIS, Wide Area Information System; 1992-1995; switch to SRS in 1995
 - Bio-Mirror world-wide data distribution, 1998-
- FlyBase genome information system, 1993-
 - One of first genome DBs, along with ACeDB (*C. elegans*)
- euGenes multi-genome information system, 1999-
 - One of first multi-organism collections of complex eukaryote genome information

Table 1. BioMirror databanks (partial list)

Section	Mbytes	Updated	Databank source
blast	8053	06-Mar-2002	Biosequence databases for BLAST searches
embl/new	1316	15-Mar-2002	EMBL daily from EBI
embl/release	9824	12-Mar-2002	The EMBL Nucleotide Sequence Database
eugenes	184	16-Mar-2002	Eukaryote Genes Summary Databank
genbank	22541	17-Mar-2002	GenBank Sequence Database
geneontology	91	15-Mar-2002	Vocabularies of gene functions and roles
interpro	56	15-Nov-2001	InterPro Protein databank
ncbigenomes	4875	16-Mar-2002	Whole genome sequence section of GenBank
pdb	8252	15-Mar-2002	Protein Data Bank of 3-D macromolecules
swissprot	67	07-Mar-2002	Annotated protein sequence database
taxonomy/ebi	5	11-Mar-2002	Taxonomy data
taxonomy/ncbi	47	17-Mar-2002	Species names
unigene	1078	15-Mar-2002	Unique Gene Sequence Collection

These public bio-sequence databanks have been built up over decades from the contributed work of 1000s of scientists, partly in response to the goal for common collection of important information all would benefit from, and partly in response to organized community inducements to publish data along with research reports.

Large public bio-sequence databanks are updated daily or weekly. Timely distribution and updates to local databanks is becoming a task that stretches network and compute resources. Validating and correcting this community-owned data is also a problem for science community resources, where ownership/responsibility is vested in 1000s of scientists. New methods to manage and distribute these shared data are needed and in development.

Uses of genome infosystems

- Accumulate research knowledge of 100,000s of genes from many organisms
 - Find and learn of gene function, cell, phenotypic effects, expt. literature, DNA and protein coding, genome mapping, alleles and variants, names, and more
- Part of ‘digital library’ of biology
 - Link to and from other sources, sinks of knowledge
- Source of validated reference data to incorporate in other projects
 - Extract subsets for use in other research

Anatomy of a Genome InfoSystem

- Information structure
 - Records of hierarchical, complex documents; Tables of rows and columns of numbers, others
 - Table of contents, Reports, Indexing (as a reference book)
 - Browse thru available structure.
 - Search and retrieve according to biological questions
 - Bulk data selection & retrieval for other uses
- Information content
 - Primary: Literature (referenced, abstracted and curated), Sequence and feature analyses, maps, controlled vocabulary/ontologies relevant to biology, people and biologics contacts, etc.
 - Metadata describing primary data, along with protocols, notes, sources
- Informatics / software
 - “backend” database, data collection, management, with some analyses
 - “frontend” information services (hypertext web, document search/retrieval methods); ease of understanding and usage (HCI)
 - “middleware” glue code, software, etc.
 - Specialized for genome data: maps, blast searches, ontologies

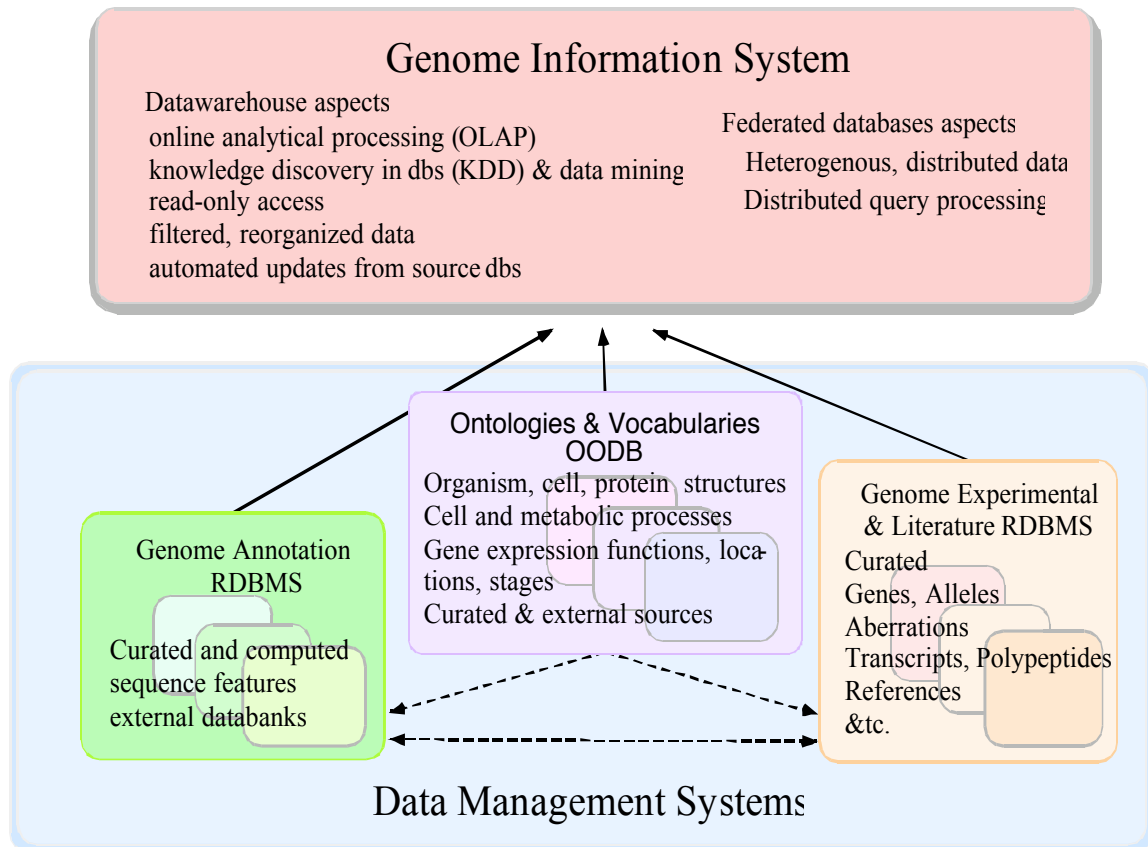


Figure 1. Genome Information system parts (FlyBase)

Related systems

- ACeDB (www.acedb.org) - C. elegans genome database, object-oriented database
- e-Prints Literature database (eprints.org) - MySQL, Perl
- GeneX Gene expression database (genex.ncgr.org) - PostgreSQL, Perl
- SRS sequence retrieval (srs.ebi.ac.uk) - flexible information retrieval system for complex, huge text bio-databanks
- Yeast genome database (genome-www.stanford.edu/Saccharomyces), Mouse (informatics.jax.org), Human (www.gdb.org and www.ncbi.nih.gov/LocusLink/)

FlyBase

<http://flybase.bio.indiana.edu/>

- primary genome information for *Drosophila melanogaster*
 - Genes & alleles, proteins/transcripts, stocks, aberrations, literature, sequences, constructed genes, anatomy & development,
- uses efficient information system methods to handle this complex, document-object structured data
- integrates hierarchical vocabularies (ontology) function and expression of genes

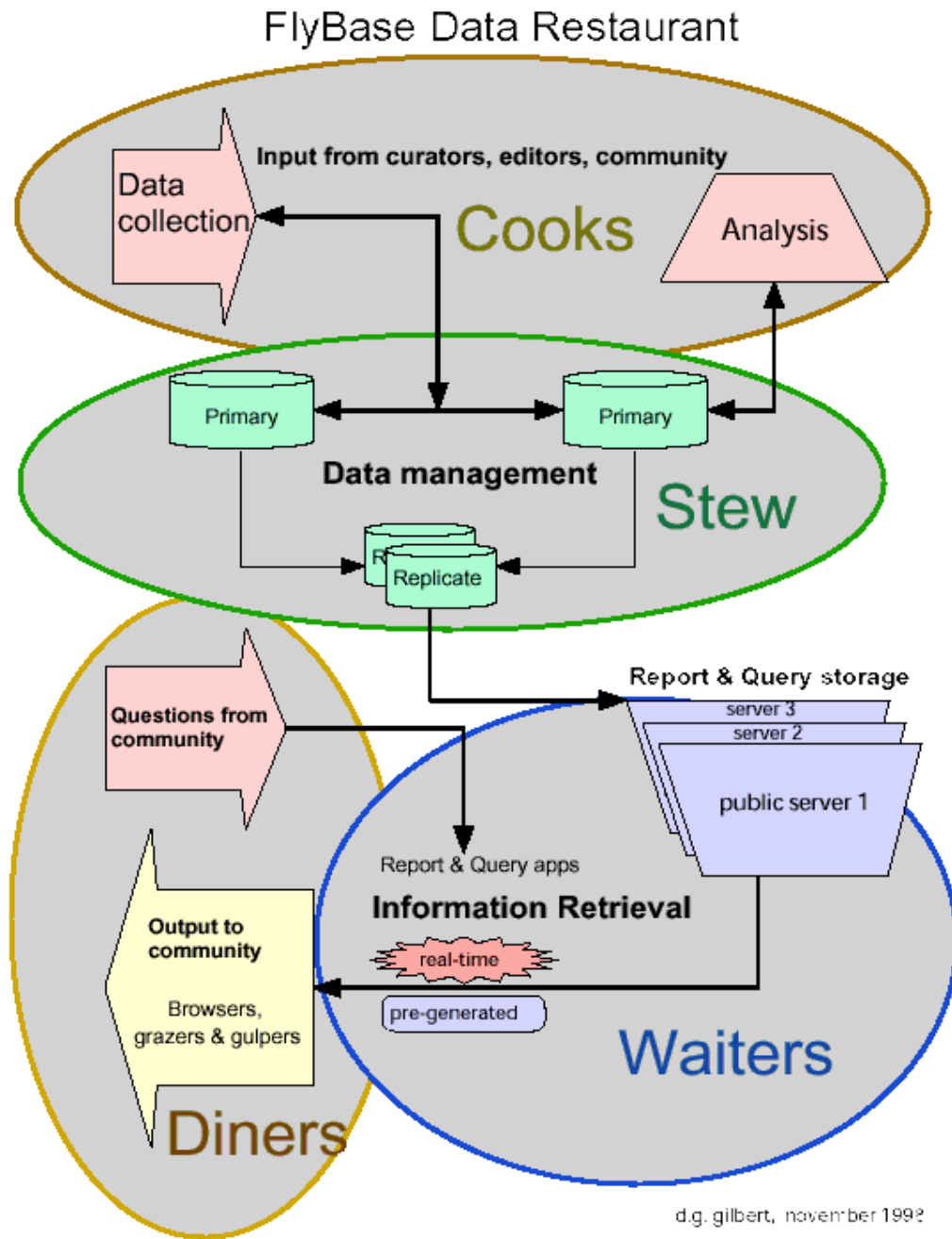


Figure 2. FlyBase Data Flow schematic

euGenes

<http://iubio.bio.indiana.edu/eugenesis/>

- Describes 150,000 known, predicted and orphan genes, using consistent gene symbols, identifiers, and synonyms
- extends FlyBase technology to human, fruitfly, worm, mouse, yeast, weed, zebrafish (rice is coming...)
- integrates diverse genome data into common format
- gene homologies (BLAST) with comparative summaries of genome homologies, features
- genome feature annotation, chromosome location and molecular maps
- gene function, process and cell location vocabulary (Gene Ontology) integration
- common genome map views with links to genes and other features
- efficient information search and retrieval methods
- constantly updated from many public sources
- compares favorably to other genome information systems for content, integration, comprehensiveness and usability. See GeneCards, LocusLink, Proteome.com, single organism systems, others

Table 2. Genome attributes in euGenes, July 2001

	Genes reported	Located	Homo logy	GO data	Genome kilobases	Genome features
Fruitfly	23,649	56%	44%	31%	116,094	41,570
Human	37,049	66%	76%	--	3,310,005	1,575,667
Mouse	28,210	--	88%	20%	--	--
Weed	26,819	100%	18%	14%	116,702	54,053
Worm	21,881	100%	27%	27%	100,090	207,478
Yeast	7,226	90%	30%	88%	12,155	13,594
Zebrafish	1,221	--	87%	--	--	--

euGenes data sources

- FlyBase, BDGP/Celera sequence (fruitfly)
- LocusLink, Golden Path (public) sequence (human)
- WormBase / ACeDB (C. elegans)
- TAIR, TIGR weed sequence (Arabidopsis)
- Mouse Genome Database
- Saccharomyces Genome Database
- Zebrafish ZFIN system
- Gene Ontology (GO) Consortium
- NCBI GenBank, SwissProt, PIR and related sequence data

FlyBase & euGenes informatics details

Text database

- flexible, simple data format for a heterogenous, hierarchical object document structure
- data record for a gene or other bio-object contains most data for human-readable report (*denormalized*).
- record contains any kind, number of fields and subrecords, along with ID and summary.
- file of all records for a class, and associated indices. Efficient search/retrieval from huge files for various software.

Browsing

- Sorted lists of data and query result subsets
- Lists show primary information of records with hyperlinks to details.
- Lists are paged for viewing unlimited numbers
- Any subset of lists can be retrieved for view or analysis, in multiple formats

Searching

- SRS search engine - fast, field aware, efficient and easily tuned for large and semi-structured text databases
- Boolean and regular expression matching
- Multiple data class linking ("joins")
- Refinement of searches: easily add new parameters to a query to focus results from first search

Reporting

- Convert data to readable reports at request time, with user options
- Present data fields selected and arranged as user desires, summarizing long lists
- Automatically generate data summaries
- Extensive hypertext links among reports, runtime configured
- Include pictures and maps generated from data
- Include external data with runtime Internet lookups (e.g. PubMed)
- Object-oriented Java software with classes specific to each data field

Properties for data exchange

- Metadata for
 - data types (audio, video, graphics, tables), experimental design, author (Dublin Core fields), links and Ids, literature, taxonomy
- Data exchange language
 - XML doc. definitions & schema
 - Minimal information for all users
 - Controlled vocabularies of science terms, ontologies
- Central & Distributed (lab maintained) databases and information services
- Repositories & Curated databases
 - Self/author archiving; staffed data collecting and curating
- Data distribution, sharing agreements and methods, authorizations
- *Examples*: Gene expression databases and repositories; Science literature archives

Summary of EthoInformatics community databases from Genome Informatics perspective

The EthoInformatics workshop participants described a set of needs and goals for sharing animal behavior data in a public way, for the future good of research in and related to this subject. Among needs are aspects of organizing existing databases, and encouraging researchers to build their own new ones, along with desire for a common, central repository or database. Data types range from large volume, rich media (video, sound, and other types), raw text and tabular data in various forms, from structured to ad-hoc observations. Methods to encourage sharing while allowing researchers to retain copyrights and other data controls are needed.

From a genome informatics perspective, look carefully at these relevant related informatics for examples and template methods

- Emerging *Gene Expression* databases and data exchange community process
 - Very similar in numerous aspects to defined needs of EthoInformatics; distributed lab use and central repositories of common data; community process for developing common exchange languages with many biological equivalents
 - Links: mged.org, genex.org
- Lessons from *GenBank*
 - Not as close a match to EthoInformatics as Gene Expression
 - Learn from success, failure of GenBank/EMBL extensive publicly shared bio-data
 - Success of carrot/stick approach requiring scientists to publish data when publishing articles or getting funding; animal behavior shared data will involve similar community forces - journal and grant agency help could be essential
 - Failure: significant public databank error due to data ownership by scientists; no inducements to update
 - Cleaned data possible with primary extensively shared databank
- Digital library's *Open Archive Initiative*
 - metadata database as a candidate prototype distributed database and data cataloging package for EthoInformatics
 - Existing open-source, well documented framework for metadata (about data) that is flexible enough to cover basic sharing of widely variable sources and kinds of data that are included in a common subject hierarchy in a distributed searchable fashion
 - Links: openarchives.org and esp. ePrints.org
- *Science GRID*
 - Has longer term (5 year) distribution methods to offer ethoinformatics
 - infrastructure for high-volume data distribution and analysis, including data resource directories (or catalogs), with standard methods for security, authenticated use, peer-to-peer sharing and efficient high-volume distributed use
 - Links: globus.org; eu-datagrid.org; www.communitygrids.iu.edu; ivdgl.org

- Common exchange language and ontologies
 - Critical component for a community of shared data with distributed model
 - Minimum information about a microarray experiment (MIAME at mged.org) and gene expression ontologies (www.mged.org/ontology), gene ontology (geneontology.org) and related examples offer detailed reasons and solutions for ethoinformatics to draw on
- Distributed &/or Central databases
 - Federation of distributed databases is harder; an important but longer term goal
 - Practical solutions in genomics field still are limited and in progress
 - Central prototype database / repository a good step on road
 - Lab/project databases are important step; building a template database in open-source sharable, cookbook way to encourage more projects using common structure is important
- Payoff in genomics:
 - one gene/one species studies now losing to importance of 1000s genes/100s species studies using central, shared public data.
 - Can animal behavior hope to answer important new questions with centrally organized shared data?
- Funding:
 - NSF is the most supportive of a range of informatics efforts, but funding is stretched thin among many sciences
 - NIH sees importance of bioinformatics
 - Applications to health for ethoinformatics, including comparative studies, behavior genetics, animal models of neurobiology & human behavior, make it suitable for funding from NIH
 - DOE and USDA include bioinformatics support, but more focused on their specific interests