

Biology Information Systems

Don Gilbert, Bioinformatics
Center for Genomics, Proteomics and Bioinformatics
and Biology Department, Indiana University

History

- IUBio Archive for biology software and data started 1989
- GenBank biosequence search using WAIS, Wide Area Information System; 1992
- FlyBase genome information system started 1993/1994.
- euGenes multi-genome information system started 1999.
- biology information retrieval system methods similar to SRS, Entrez (NCBI), Medline

Bio-information warehousing and distribution.

- IUBio Archive, <http://iubio.bio.indiana.edu/> and Bio-Mirrors, <http://www.bio-mirror.net/>
- public molecular biology software archive
- Sequence and related biology databanks
- **Future:** extend and improve methods for bioinfo. data-warehousing, re-distribution and user search services

Core bioinformatic data sets in Bio-Mirrors

Databank	Description	Home site
BlastDB	Biosequence databases for BLAST searches	NCBI
Blocks	Highly conserved regions of proteins	NCBI
DDBJ	DNA Data Bank of Japan	NIG
EMBL	The EMBL Nucleotide Sequence Database	EBI
Enzyme	Enzyme nomenclature database	ExPASy
GenBank	GenBank Sequence Database	NCBI
Genomes	Whole genome sequence section of GenBank	NCBI
InterPro	InterPro Protein databank	EBI
PIR	Protein Information Resource	NBRF
Pfam	The Pfam database of protein domains and HMMs	WUSTL
Prosite	Database of protein families and domains	ExPASy
Rebase	The Restriction Enzyme Database	NEB
RefSeq	NCBI Reference Sequences	NCBI
SRS Databanks	List of active SRS databases around world	EBI
SWISS-PROT	Annotated protein sequence database	ExPASy
Taxonomy	Species names	NCBI
TrEMBL	A supplement to SWISS-PROT	EBI
UniGene	Unique gene sequence collection	NCBI
euGenes	Eukaryote Genes Summary Databank	IUBio

Bioinformatics tools and services.

- **Future:** improve public molecular biology software archive and web-based services
- **Future:** enhance data integration with SRS – Seq. Retrieval System <http://iubio.bio.indiana.edu/srs/>
- **Future:** applications for Bio-Mirror and other bioinformatics centers.

Bioinformatics knowledge integration / discovery (BIKD)

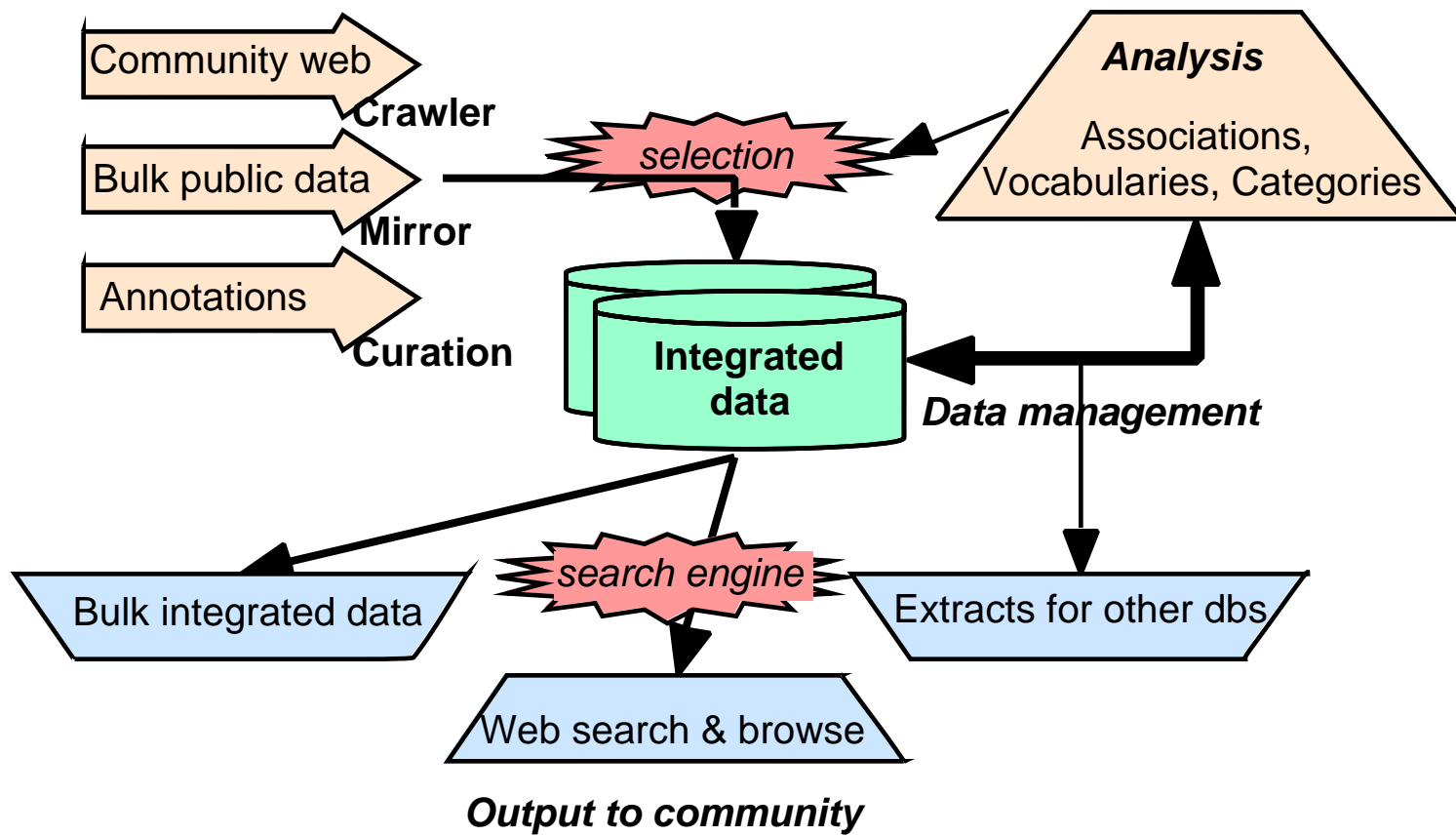
- **Future:** Research areas of knowledge integration for a broad spectrum of this molecular bioscience information, including automated Internet collecting and categorization.

BIKD methods include

- Standard text analysis methods .
- Concept indexing and categorization; to categorize documents in concept hierarchies.
- Factor analysis of the term-document matrix for term relevance, concept relevance, Internet linkage.

BIKD components include

- **web crawlers:** httdig, libwww robot, a focused crawler <IBM, 1999; Chakrabarti *et al.* 1999>, Harvest and others.
- **indexers:** glimpse, essence, lsearch, FreeWais, SRS
- **analyzers:** R statistics / SAS, text analysis tools, data mining and management tools.



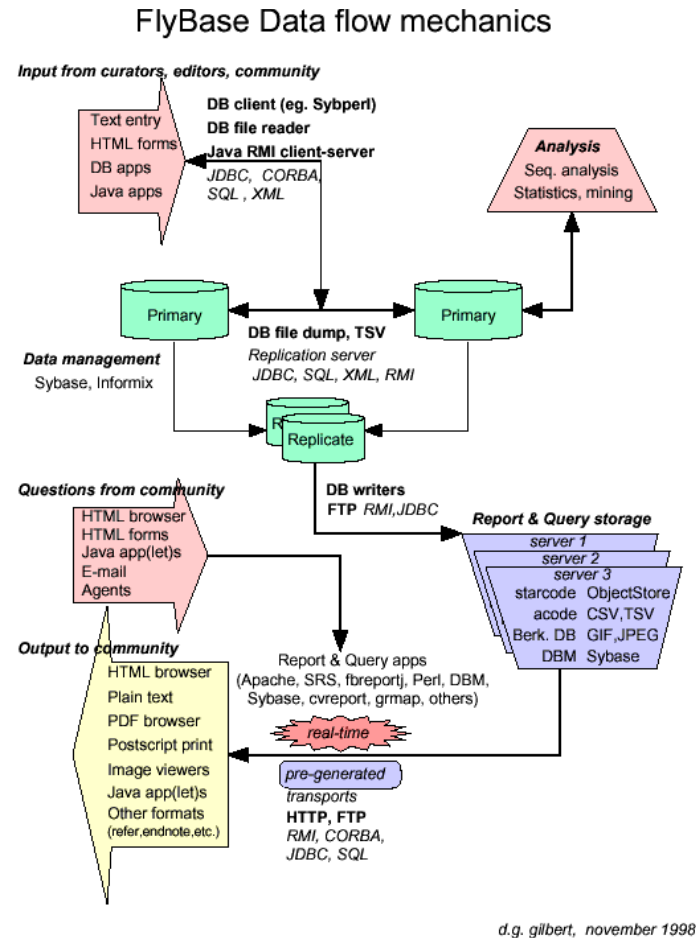
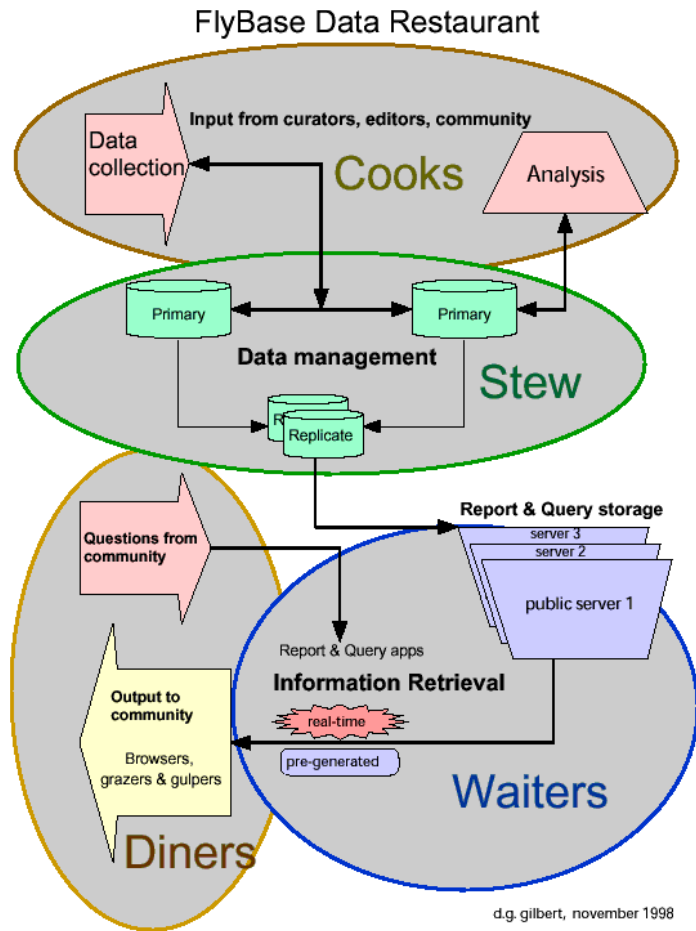
Bioinformatics knowledge integration overview

Genome information systems

FlyBase

- <http://flybase.bio.indiana.edu/>
- the primary genome information system for the fruitfly *Drosophila melanogaster*
- uses efficient information system methods for complex, document-object structured data; works well for multi-gigabyte data in many formats
- integrates hierarchical controlled vocabulary (ontology) structures for integrating knowledge of development, function and expression of genes

FlyBase Data flow overview



euGenes

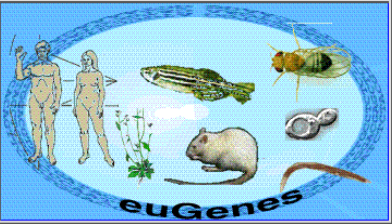
This project develops an information system for the collection, storage, integration, retrieval and distribution of genome data, including

- development of a flexible, usable whole-genome map display;
- development of analytical tools that can be used in the assembly, analysis, and interpretation of genomic data;
- creation of a database and software tools that provide easy access to up-to-date physical and genetic mapping and DNA sequencing information and allow linkage or integration of these data to related datasets (phenotypic, expression, structural data)

Current euGenes status

- <http://iubio.bio.indiana.edu/eugenest/>
- extends FlyBase technology to multiple eukaryote organism genomes
 - 7 popular genomes: human, fruitfly, worm, mouse, yeast, weed, zebrafish
 - a common summary of primary information on all known protein coding genes
- integrates diverse genome data into common format
 - gene symbol matching among data sets
 - gene homology (BLAST) calculations
 - genome feature annotation, chromosome location and molecular map integration
 - gene function, process and cell location vocabulary (Gene Ontology)
- common genome map views with access to genes and other features
- comparative summaries of genome homologies, features
- efficient information search and retrieval methods
- constantly updated in automatic way from many public sources
- compares favorably to other genome information systems for content, comprehensiveness and usability. See GeneCards, NCBI LocusLink, Proteome.com yeast and human services, single organism systems, others (Celera Discovery System?)

euGenes	
Organism	Available genes
Fruitfly Genes (Drosophila melanogaster)	23,321
Human Genes (Homo sapiens)	33,609
Mouse Genes (Mus musculus)	28,179
Weed Genes (Arabidopsis thaliana)	15,909
Worm Genes (Caenorhabditis elegans)	23,305
Yeast Genes (Saccharomyces cerevisiae)	7,524
Zebrafish Genes (Danio rerio)	1,039
All Genes summaries	
Help & Documents	Tools
Last updated: 22 February 2001	



euGenes
Genomic Information for Eukaryotic Organisms

euGenes provides a common summary of gene and genomic information from eukaryotic organism databases. This includes

- Gene symbol and full name,
- Chromosome, genetic and molecular map information,
- Gene product information (function, structure, and homologies).
- Links to extended gene information.

This summary is automatically maintained from the primary databases.

New features:


- [Human Genome Maps](#) (GoldenPath public draft) - 21 Feb 01
- [Gene Function/Location/Process](#) summaries (Gene Ontology) - Jan 01

[Search Help?](#)

for these words (appending wildcard '*' to words)

in these organisms:
excluding
predicted genes.

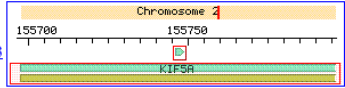
[euGenes](#) · [Fish](#) · [Fly](#) · [Human](#) · [Mouse](#) · [Weed](#) · [Worm](#) · [Yeast](#)



[Help](#) · [Preferences](#)

euGenes Report

Human Gene KIF5A

Symbol KIF5A	Full name kinesin family member 5A
Function	adenosinetriphosphatase , motor , Motor protein, Hydrolase
Process	kinesin family member 5A, non-selective vesicle transport , microtubule-based movement , synaptic transmission , Neuronal transmission
Cellular location	membrane fraction , kinesin , Cytoskeletal, Unspecified membrane, Cell body (soma), Tubulin-cytoskeleton associated
Protein domains	--
Chromosome Map location	
Ref. sequence	REFSEQ:NM_004984
Ref. protein	REFPROT:NP_004975
Similar genes	<i>Fruitfly</i> Khc FBgn0001308 (55%) <i>Mouse</i> Kif5a MGn0006756 (86%) <i>Weed</i> T8O18.9 ATgn0013133 (41%) <i>Worm</i> unc-116 CEgn0002994 (59%)
Synonyms	N-KHC, D12S1889, NKHC
Database ID	HUgn0003798
Ref. Database	LocusLink:3798
Database accessions	UNIGENE:Hs.192760 , OMIM:602821 , GDB:9864370 , CARDS:KIF5A
Database URL	http://www.ncbi.nlm.nih.gov/HomoloGene/homolquery.cgi?TEXT=3798[loc] http://www.proteome.com/databases/HumanPD/reports/3798.html

euGenes future:

- extend weed gene data set to full genome, add rice genome data;
- integrate literature (Medline) and additional bioinfo data (InterPro and protein family data)
- additional whole genome analyses (summarized and user-directed)

FlyBase & euGenes informatics details

Text database

- flexible, simple data format: a hierarchical object document structure but less complex than XML to suit existing software tools.
- data record for a gene or other bio-object contains most data for human-readable report.
- record contains any kind, number of fields and subrecords, along with ID and summary.
- file of all records for a class, and associated index for record retrieval by ID. Retrieval from gigabyte files is efficient for wide variety of software.

Browsing

- Sorted lists of data and query result subsets
- Lists show primary information of records (name, location, features) with hyperlinks to details.
- Lists are paged for viewing unlimited numbers
- Lists can be stepped to view entire range easily, and easily locate subranges of interest
- Lists can be re-sorted by key features

Searching

- SRS search engine - fast, field aware, efficient and easily tuned for huge and semi-structured text databases
- Boolean and regular expression matching

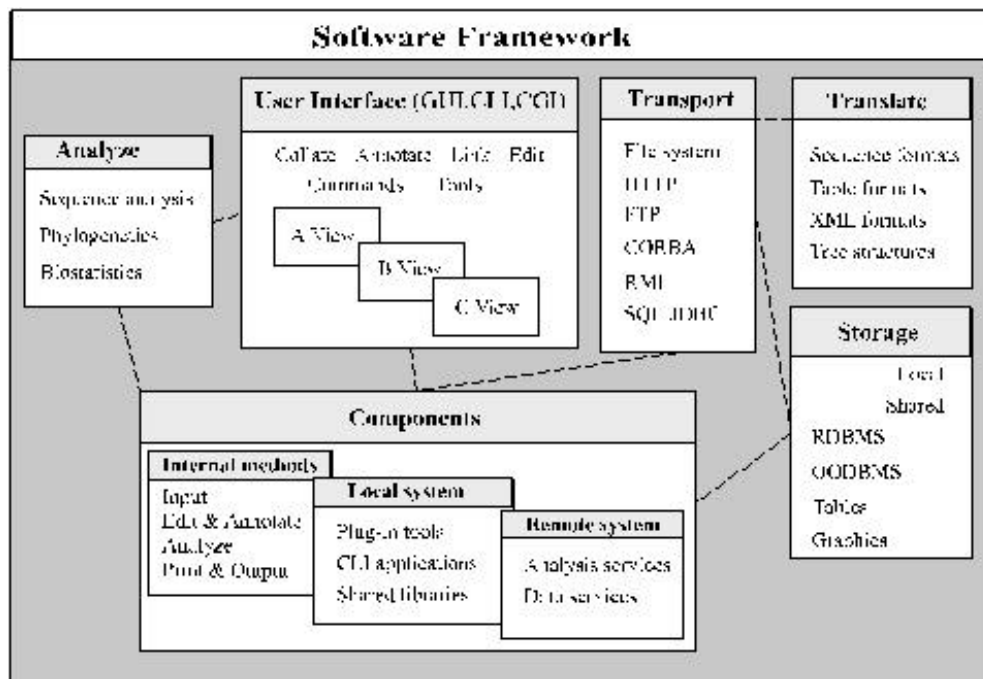
- multi-data class linking (database "joins")
- Search indices are regenerated when data is updated.
- Refinement of searches allow one to fine-tune query from first search

Reporting

- Convert data records to human readable reports at request time, with user configurations
- Present data fields selected and arranged as user desires, summarizing long lists
- Automatically generate data summaries
- Extensive hypertext links among reports
- Include pictures and maps generated from data
- Include external data directly with run-time internet lookups (e.g. PubMed abstracts)
- Software base is object-oriented Java with classes specific to each data field

Framework structure

- emphasis on Human-Computer Interaction factors for usability
- Java-based, strong user-interface, Internet client-server transport
- add and replace component tools and data sources; customize to needs
- integrates with external bioinformatics tools
- use standard formats, protocols, and component tools



build from existing components

- common code library of framework components (5 Megabytes of Java source)
- **SeqPup** is a biosequence editor and analysis framework;
<http://iubio.bio.indiana.edu/soft/molbio/java/apps/seqpup/>
- **gnomap** a whole-genome map display and discovery tool (in euGenes);
<http://iubio.bio.indiana.edu/soft/molbio/java/apps/gnomap/>
- **Phylodendron** a phylogenetic tree drawing program;
<http://iubio.bio.indiana.edu/soft/molbio/java/apps/trees/>
- **Readseq** a widely-used biosequence conversion tool;
<http://iubio.bio.indiana.edu/soft/molbio/readseq/java/>