SRS - A Backbone for Genome Information and Data Grid Systems

Don Gilbert

Indiana University

gilbertd@bio.indiana.edu

Overview

- **Search/Retrieval in Genome Information systems**
- **Efficiency and complexity: RDBMS, SRS†, others**
- **Genome data federation: local and distributed**
- **Directories of data: automated S/R and the Grid**
- **SRS, LDAP and future biodata grids**

† Sequence Retrieval System, Lion Bioscience

Bioinformatics @ Indiana U. using SRS

- **Bio-info archiving and distribution**
  - **IUBio Archive, http://iubio.bio.indiana.edu/** -- public molecular biology data / 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 of Drosophila fruitfly
  - **euGenes, http://eugenes.org/** -- infosystem for 8 important eukaryotes with 180,000 genes
- **Bio-Data Grids**
  - **http://iubio.bio.indiana.edu/grid/** -- experimental distributed computing

Genome Information Systems

- **FlyBase, euGenes (SRS,Perl/Java)**
- **Wormbase (AceDB > RDBMS, BioPerl)**
- **Mouse GD, Sacc. GD (RDBMS)**
- **GeneCards (Glimpse > XMLquery)**
- **Ensembl (RDBMS,BioPerls)**
- **Nascent: many newly developing organism genome systems**

euGenes

- 8 eukaryote genomes in common summary data format
- Describes 180,000 known, predicted and orphan genes
- Gene Homologies with comparative summaries
- Genome map views and feature annotations
- Gene Ontology function, process and cell location integration
- Efficient information search and retrieval methods
- Extends FlyBase information system technology
- Updated (semi) automatically from several sources

Genome attributes in euGenes

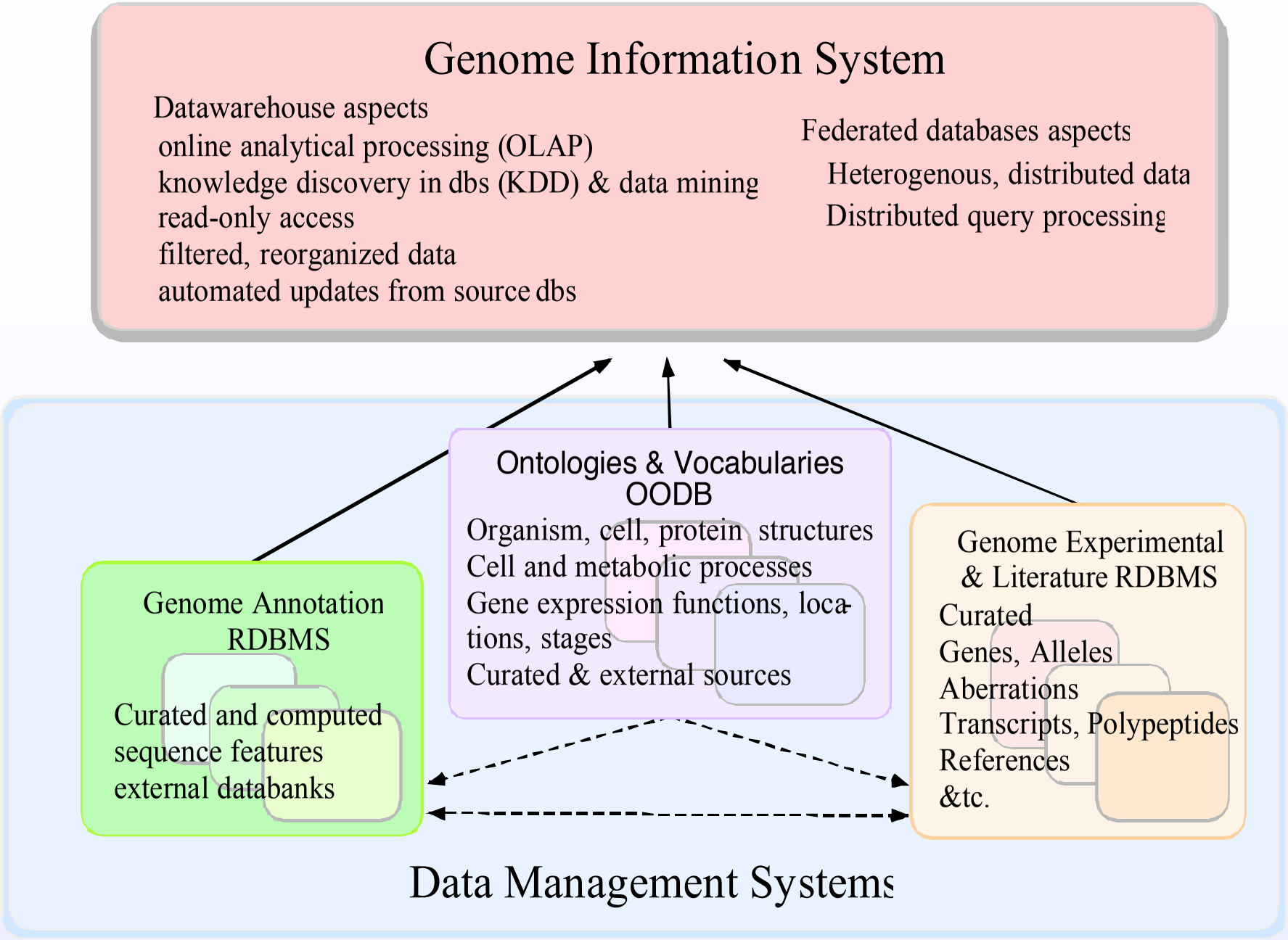
July 2002

	Genes	Located	Pre-dicted	Homo-logy	GO data	Genome kilobases	Genome features
Fruitfly	25,728	51%	35%	61%	27%	116,109	100,309
Human	53,210	84%	61%	39%	38%	3,118,900	258,835
Mouse	36,433	--	--	94%	22%	--	--
Mosquito	12,687	100%	91%	66%	--	231,408	161,565
Weed	28,129	100%	--	43%	35%	117,429	84,320
Worm	22,705	90%	78%	30%	29%	100,270	244,675
Yeast	7,222	92%	32%	34%	89%	12,156	13,876
Zebrafish	1,583	--	--	89%	--	--	--

Genes as extracted from genome project sources. These differ from true gene numbers by orphan gene records, prediction artifacts, unmerged predicted/expt. records, and unfinished sequencing gaps.

Search/Retrieval for Genome DBs

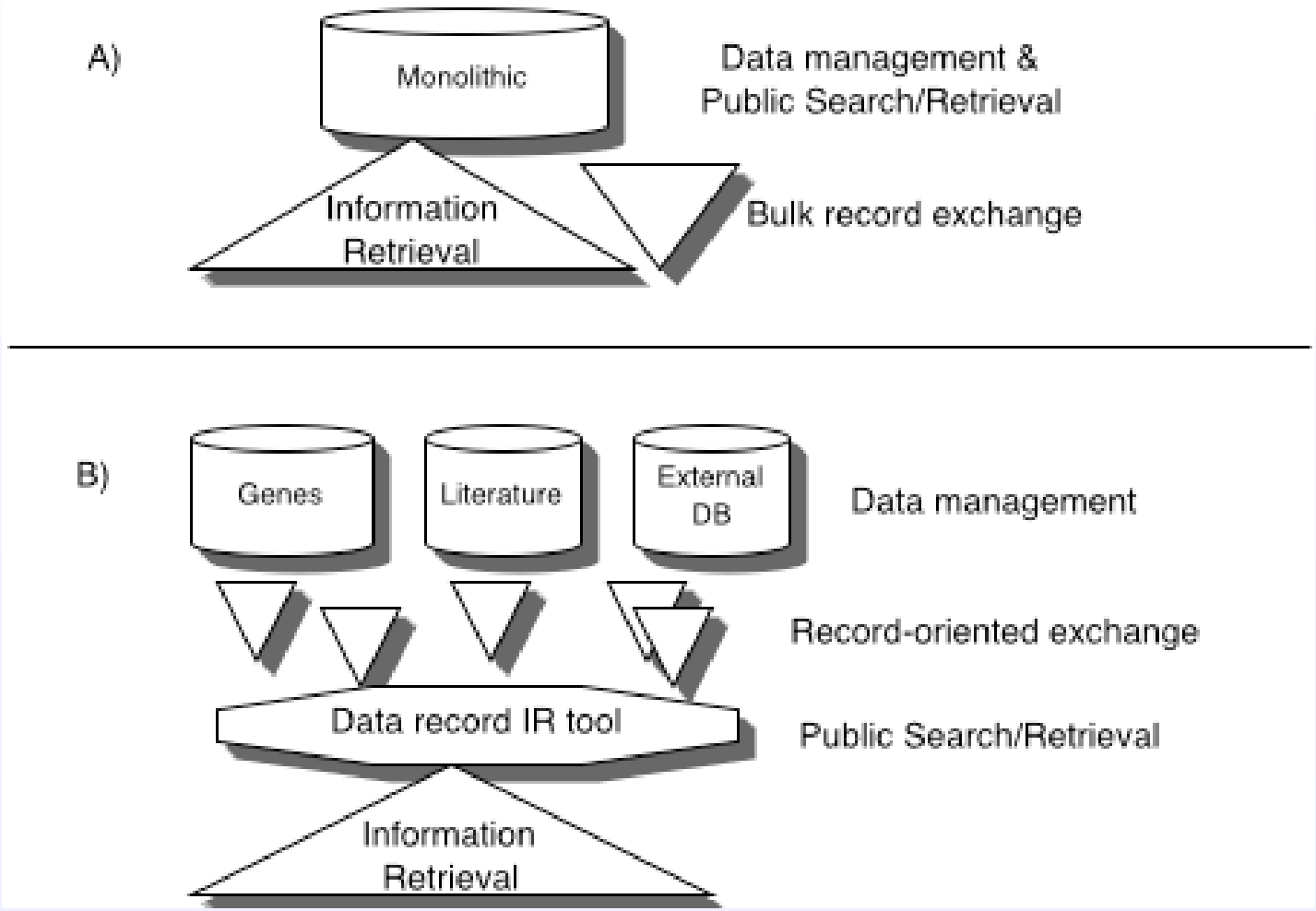
- **Separate management and public search/ retrieval has advantages in flexibility, speed**
- **Indexing methods for text databases (or rdbms exports) are accurate, efficient for high volume data, easy to implement for complexly structured biology data**
- **Sequence Retrieval System (SRS) is used in FlyBase and euGenes; GeneCards uses Glimmer and similar methods; Google and Digital library methods are related**


Genome Information System
Datawarehouse aspects
online analytical processing (OLAP)
knowledge discovery in dbs (KDD) & data mining
read-only access
filtered, reorganized data
automated updates from source dbs
Federated databases aspects
Heterogenous, distributed data
Distributed query processing
Ontologies & Vocabularies
OODB
Organism, cell, protein structures
Cell and metabolic processes
Gene expression functions, loca-
tions, stages
Curated & external sources
Genome Annotation
RDBMS
Curated and computed
sequence features
external databanks
Genome Experimental
& Literature RDBMS
Curated
Genes, Alleles
Aberrations
Transcripts, Polypeptides
References
&tc.
Data Management Systems

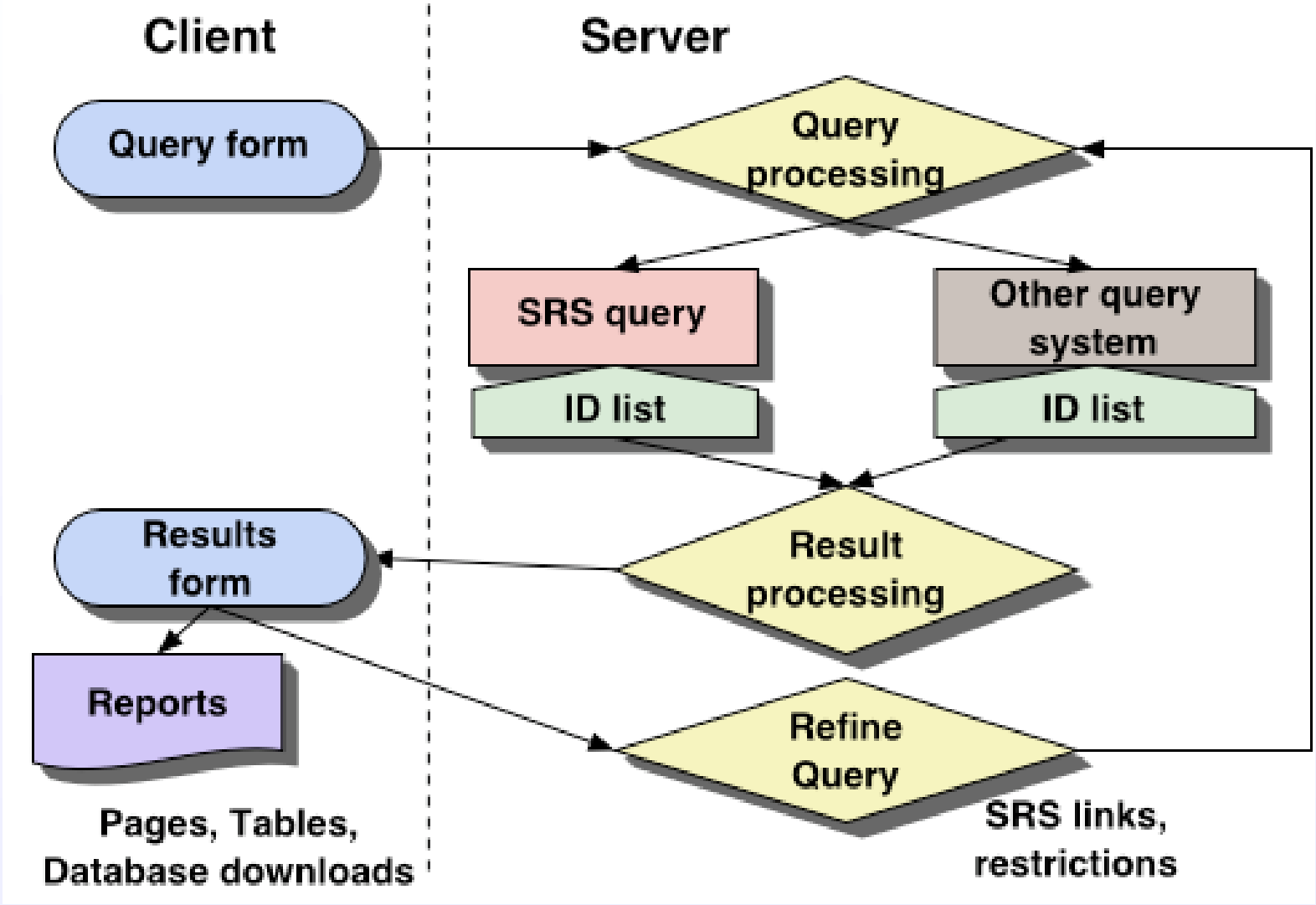
Anatomy of a Genome Info. System

- **Information structure**
  - Complex document structure; tabular data; etc.
  - Organize: Table of contents, Reports, Indexing
  - Browse contents; Search / retrieve from biological questions
  - Bulk data search / retrieve for bioinformatics
- **Information content**
  - Literature (abstracted and curated), Sequence and feature analyses, maps, controlled vocabulary/ontologies, people, biologics, contacts, etc.
  - Metadata describing primary data, along with protocols, notes, sources
- **Informatics / software**
  - Backend database, data collection, management, analyses
  - Front-end services (hypertext web, document search/retrieval); ease of understanding and usage (HCI)
  - Middleware glue code, interfaces, software, etc.
  - Specialized for genome data: maps, blast searches, ontologies

Single DB vs. Federated Info. S/R



FlyBase/euGenes Query System



FlyBase Query Results

FlyBase Genes query results

Query: ([libs={FBgn PFgn}-all:wing] *or* [libs-syn:wing]) *and* [libs-org:Dmel], No. matches= 1437
Bookmark FBquery: ([libs={FBgn PFgn}-all:wing] | [libs-syn:wing])& [libs-org:Dmel]

#	Symbol	Name	Map	Alleles	Stocks	Refs	DNA	Date
1	18w	18 wheeler	56F11	16	2	56	13	31 May 02
2	2R-F	-	-	2	1	3	-	31 May 02
...								
19	Act42A	Actin 42A	42A2	2	-	73	23	31 May 02
20	Act5C	Actin 5C	5C7	14	1	129	43	31 May 02

---------- Page and Sort results ----------

Batch Download

Fetch items: x All Items […] **Format:** [Spreadsheet] **Report content:** [Summary]
Report only Select fields: [Field list]

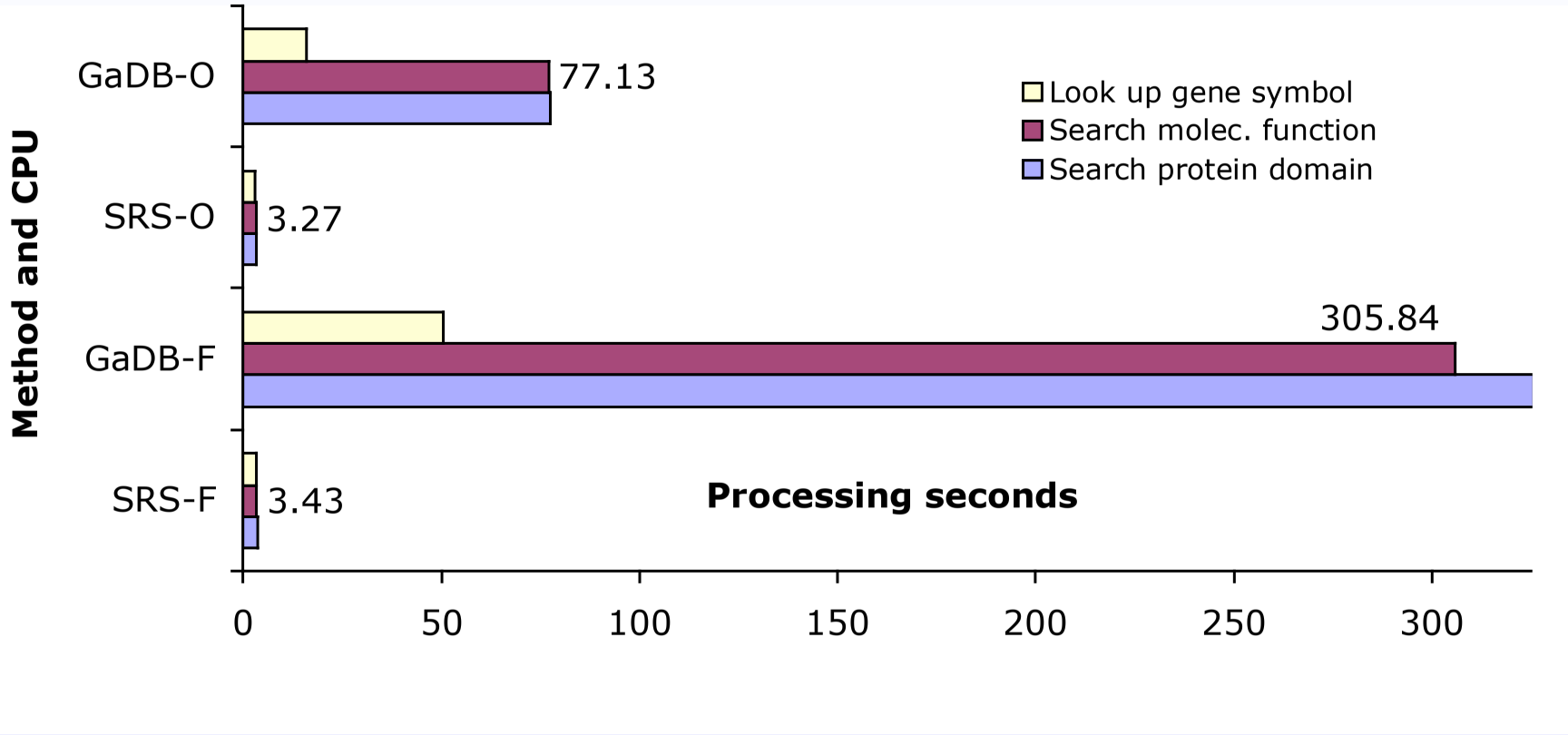
Refine query or find items in related data

Refine query ([libs={FBgn PFgn}-all:wing] *or* [libs-syn:wing]) *and* [libs-org:Dmel]
[and] [other fields] matches […..]

Search **Genes** , retrieve **Related Data Classes** (alleles, aberrations, transcripts, insertions, sequences …)

Efficiency of SRS versus RDB

Drosophila Genome Annotations
SRS or Gadfly DB relational database
Web search time (*shorter is better*; two computers - O,F)

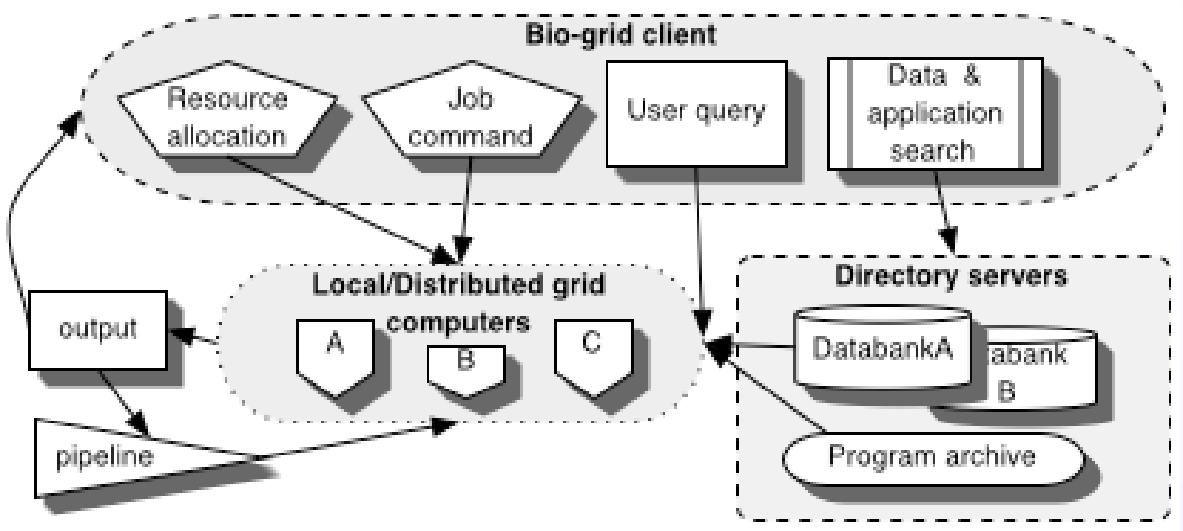


[-- Genomes to Grids --]

Science Data Grids

- **Infrastructure for distributed analyses**
  - **analyses distributed among 1000s of commodity computers**
  - **high-volume data distribution**
  - **data resource directories (catalogs)**
  - **security, authenticated use**
  - **peer-to-peer sharing and collaborations**
  - ***Data grid infrastructure still needs work***
- **Links**
  - **globus.org; eu-datagrid.org; ivdgl.org**

BioGrid Client-Server Aspects



- **Grid-aware client software**
- **Data and software resource directories**
- **Grid of processing computers**

Moving Data on the Grid

1. **@virtualdata= biodirectory "find protein coding sequences for species X,Y,Z"**
2. **@realdata= biodirectory "get locators for @virtualdata split 100 ways"**
3. **for i (1.. 100) { copydata(realdata[i],gridcpu[i]); runapp(gridcpu[i]) }**

Design of bio-data directories

1. Develop schema describing directory objects and attributes. Essential fields include ID/accession, data class / category, update time. Start with minimal directory descriptions.
2. Create directories of data records, with existing backend software such as SRS, RDMBS, Entrez, others.
3. Replicate directories among data centers; use for determining primary data to be fetched or mirrored.
4. Common exchange formats, schema, directory query syntax are necessary, implementation details are at the choice of a data center.

Directories of Genome data

- **For genome data, "broad and shallow" directories can federate the "narrow and deep" data-bases**
- **Science Grid computing**
  - **Needs efficient, authenticated discovery and distribution of high volume data**
- **LDAP directories**
  - **mature, efficient for high volumes, allows federated queries over distributed directories, and works well for SRS databanks and genome annotations;**
  - **As functional as BioDAS (distributed annotation); broader in scope, with generic client/server software**

LDAP? Why not xxx?

- **Why LDAP for bio-data directories?**
  - **Available now with many features needed**
- **Web/XML ?**
  - **Web/SOAP/WSDL/UDDI: SOAP for communication of directory requests, WSDL for an interface to the directory repository, UDDI to locate the service *(some assembly required…)***
  - **DSML: a direct conversion of LDAP to XML, for Web/XML interoperability to LDAP (e.g., http://www.dsmltools.org/); supported by industry (Msoft, Sun, others)**
- **CORBA? SQL? Wgetz? FTP?**

Light-weight Directory Features (LDAP)

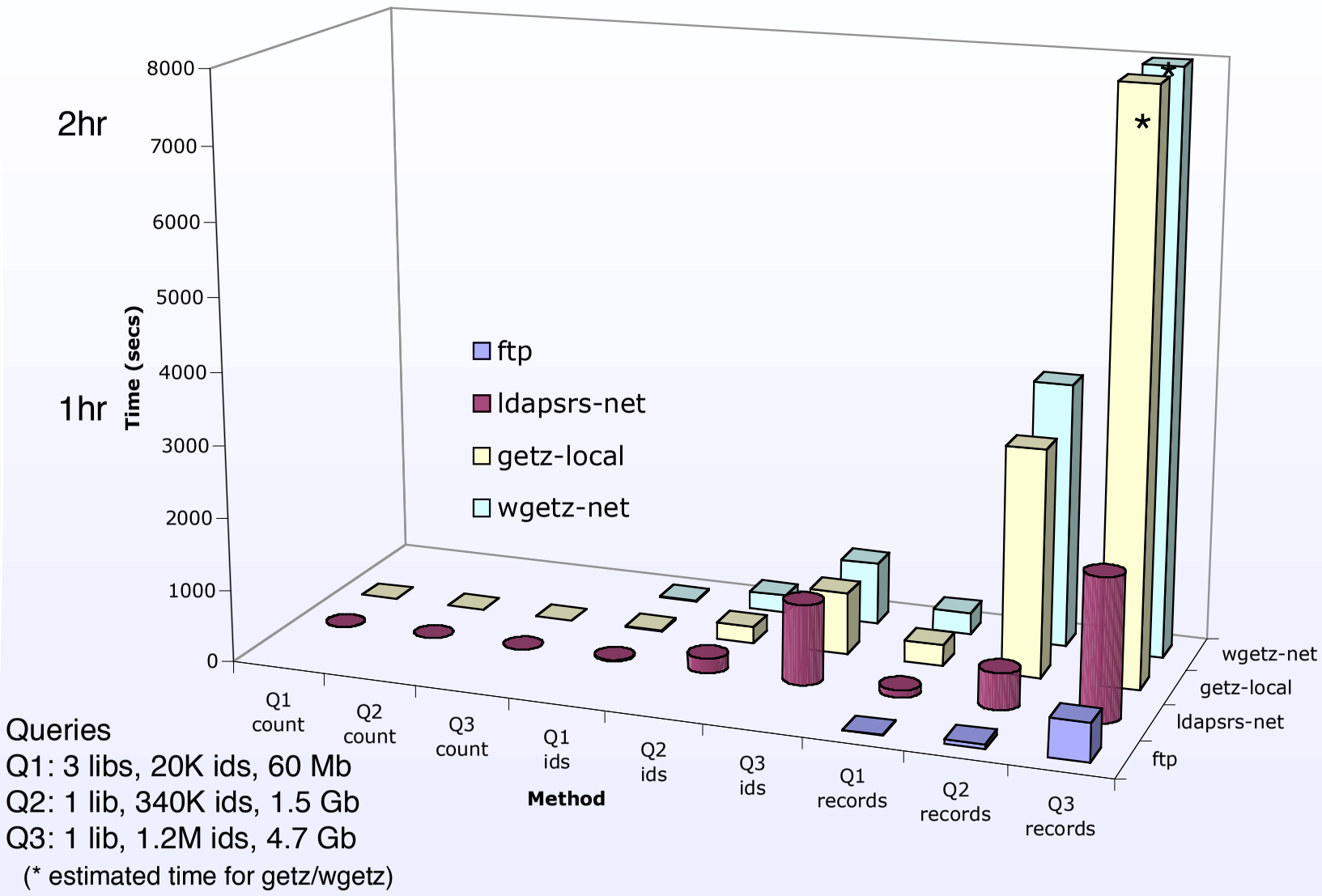
- Flexible, hierarchical directory of objects with identifiers to community definitions. Objects are simple or complex. Each has attributes (fields) composed of strings, numbers, binaries and complex structures
- Use many backend systems (including SRS); can be added to search/retrieval systems relatively easily
- Globally distributed searches of many directories
- Schema are documented: objects and attributes have unique identifiers and definitions (e.g. IETF RFC documents)
- Schema search/retrieval for directory 'discovery'
- Computable search, browse, retrieval; referrals to other servers and remote objects; extension mechanisms for new object types
- Replication of directories; mechanisms for peer group updates
- Security mechanisms for data transport, access and updates

SRS6 - LDAP gateway

- **Experimental SRS6 backend search compiled with OpenLDAP server**
  - **http://iubio.bio.indiana.edu/grid/directories/**
  - **ldap://iubio.bio.indiana.edu:3895/srv=srs**
- **Act like “getz” or “wgetz”, with LDAP query input and output**
- **Efficient, functional as network getz**

  surpasses wgetz for programmability, efficiency
- **Issue 1: convert ldap to srs query**
- **Issue 2: [.. must be something ..]**

SRS-LDAP efficiency

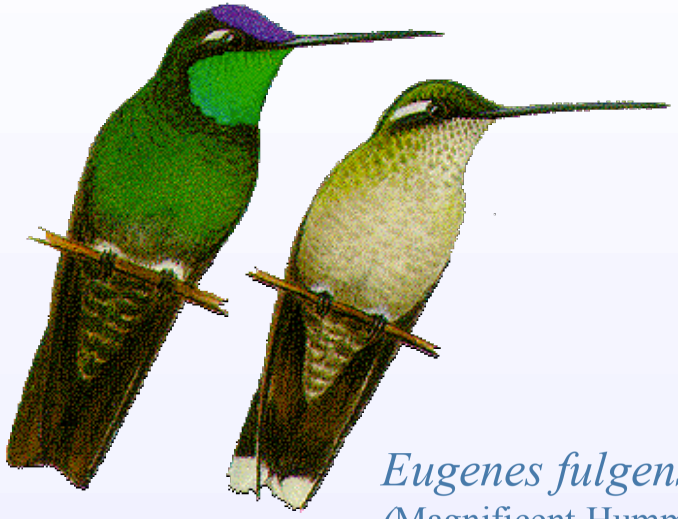


Queries
Q1: 3 libs, 20K ids, 60 Mb
Q2: 1 lib, 340K ids, 1.5 Gb
Q3: 1 lib, 1.2M ids, 4.7 Gb
(* estimated time for getz/wgetz)

Wrap-up

- **Beyond sequence retrieval with SRS to genome and biological information systems**
- **Federation of disparate data is “easy”**
- **Efficiency is high, an important factor in information systems**
- **Grid, future distributed computing needs flexible, efficient technology such as SRS.**

End of SRS - Genomes and Grids



Eugenes fulgens
(Magnificent Hummingbird, Costa Rica)

Don Gilbert
gilbertd@bio.indiana.edu

Indiana University
http://iubio.bio.indiana.edu/