Argos
& Genome Directories
& Lucegene ('Lucy Jean')

A Replicable Genome infOrmation System of Common Components

GMOD Meeting, Sept. 2003

Don Gilbert, gilbertd@indiana.edu

Focus on Genome Data Access

- Bioscientists are data-mining to study 1000s of genes rather than 1.
- Web page scraping and bulk files not enough
- Need Internet search & retrieval of genome objects distributed among many sources
- Simple, flexible client program model
- Efficient for high volumes (10^5 objects; >1 GB sizes)

Three building blocks

- ***Argos*** is a framework for distributing common components with implemented genome data systems
- ***LuceGene***, SRS,… are backends to search & retrieve data objects efficiently from any flat-file
- ***Genome Directory System*** includes WebServices, GridServices, LDAP, OAI,… Internet standard interfaces to search backends

Argos

- Reduce install & replication effort
  - Replace common { fetch, compile, install, configure,…} loop for packages of software & data
  - Compatible with most GMOD efforts
  - Compare to EnsEMBL, WormBase, other distributable systems
- Reference servers
  - http://www.gmod.org/argos/
  - http://eugenes.org/argos   http://flybase.net/flybase-ng
- General contents

common/

    java/ ; perl/ -- program libraries and packages
    servers/ -- major programs (BLAST, PostgreSQL, others)
    systems/ -- OS executables of programs

daphnia/, eugenes/, flybase/ -- *implemented organism genome systems*

centaurbase/ -- *sample testing system*

docs/ & install/ -- Argos instructions and usage

ROOT/ -- common directory of projects, each as virtual host web service in ROOT

Argos common parts

- *Java* common library, Ant builds, XML Tools, Web Services (Axis), Lucene for “Google”-like searches
- *Perl* common library of BioPerl, GBrowse, others
- *Servers* include
  - Apache, Tomcat web servers
  - MySQL, PostgreSQL databases
  - BLAST (NCBI)
- *Systems* compiled for
  - **apple**-powerpc-darwin, **intel-linux**, **sun**-sparc-solaris

Argos features

- Common genome & IT tool set
  - Share benefits of “best of breed” genome tools
  - Common parts are tested & maintained by others
  - Minimal IT expertise (no compiles or system management)
- *To do* for Common set
  - Mod-perl for Apache web server (& Perl runtime)
  - More GMOD tools (Gbrowse; Cmap; …)
  - …

Argos features

- Flexible project packages
  - Project needs specify tool set (compare EnsEMBL all-in-one)
  - Own look’n’feel web pages, contents, functions
  - Security with protected and public sections (including collaborative editing, updates)
- *To do* for packages
  - Improve package configuring
  - More integration of common & project parts
  - …

Argos features

- Easy replication to any Unix computer
  - ‘Live’ copy with rsync keeps servers up-to-date
  - Local cluster/grid for high-volume traffic
  - Works on common workstations, laptops
- *To do* for replication
  - File sync useless for Postgres updates; transactions?
  - One-click install & documentation
  - Improve auto-update; need more post-update processing

Argos advanced features

- Data mining (Genome Directory component)
  - Fulfill need to search & retrieve 1000s of genes
  - Simple, computable, industry standards for distributed query & retrieval of big data (Web Services, Grid Services, LDAP)
  - Use to update personal, lab databases with genome links
- *To do* for Data mining
  - Much !

Argos comparisons

- EnsEMBL
  - See install instructions - not hard, but harder than auto-replication
- WormBase, Gramene
  - ??
- Redhat, MacOSX, other system package auto-updaters
  - no data replication; mature; focused on system-level updates
- Globus Grid package management, PacMan
  - Also offers binary program replication; install on remote systems; more configuring
  - Data replication is immature (less useful than rsync, wget, ftp mirror) but includes directory management
- Others?

Daphnia Example System

wFleaBase -- proto-Daphnia genome system

Cgi-bin -- Web programs(Perl)

Common -- Link to common, shared tools

Conf -- Site configurations for web, data

Data -- Bulk data & FTP site folder

Dbs -- Project databases: blast, lucene, mysql

Indices -- Database indices

Lib -- Program libraries

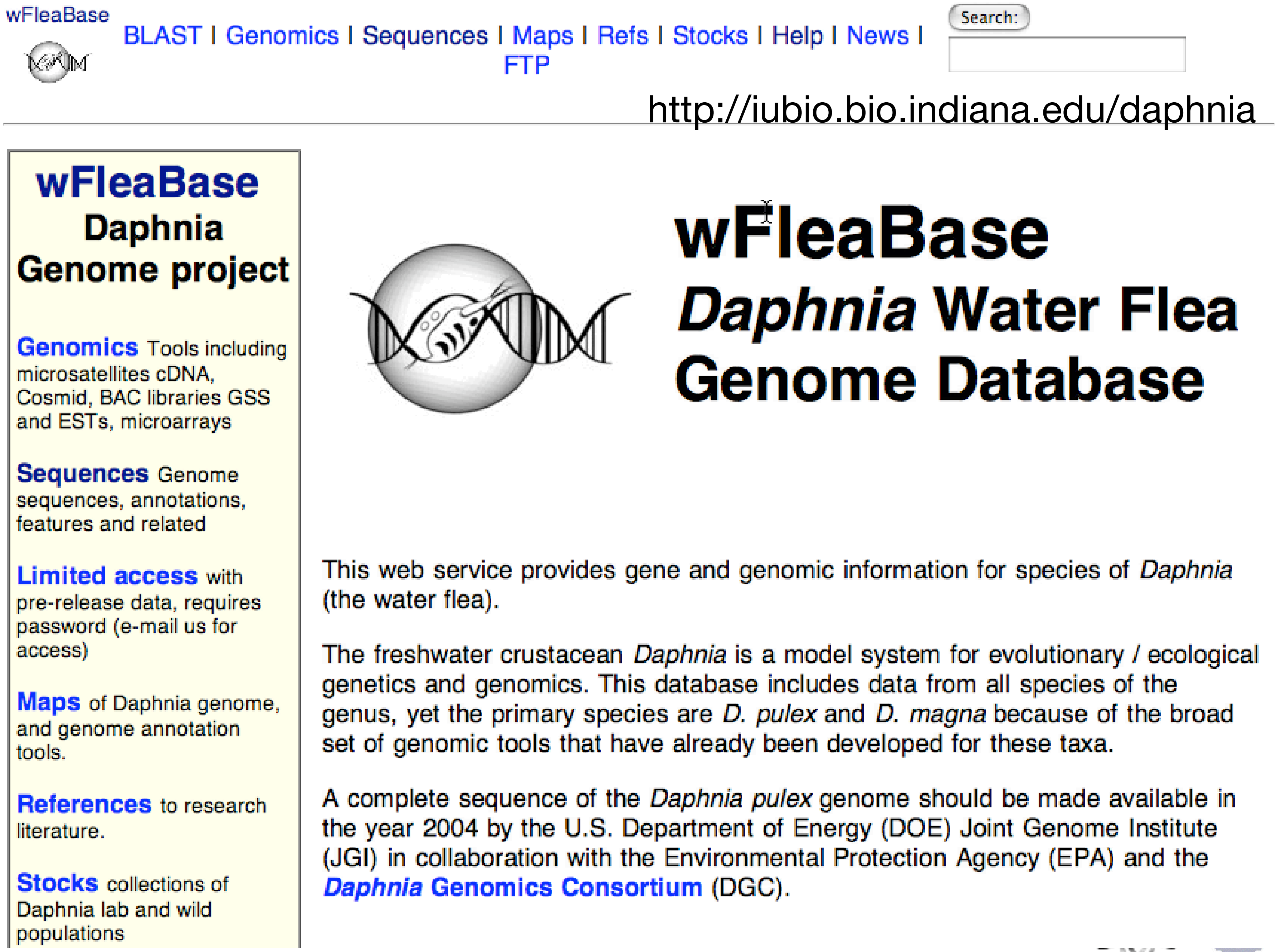
Web -- Web structure and documents

Genomics, Sequences, Maps, Literature, Stocks, Docs, other

includes Public and Protected (project member only) parts

Webapps -- Web programs (Java)

includes Search system, Secure web and editing



http://iubio.bio.indiana.edu/daphnia

BLAST | Genomics | Sequences | Maps | Refs | Stocks | Help | News | FTP

wFleaBase
Daphnia Genome project

Genomics Tools including microsatellites cDNA, Cosmid, BAC libraries GSS and ESTs, microarrays

Sequences Genome sequences, annotations, features and related

Limited access with pre-release data, requires password (e-mail us for access)

Maps of Daphnia genome, and genome annotation tools.

References to research literature.

Stocks collections of Daphnia lab and wild populations

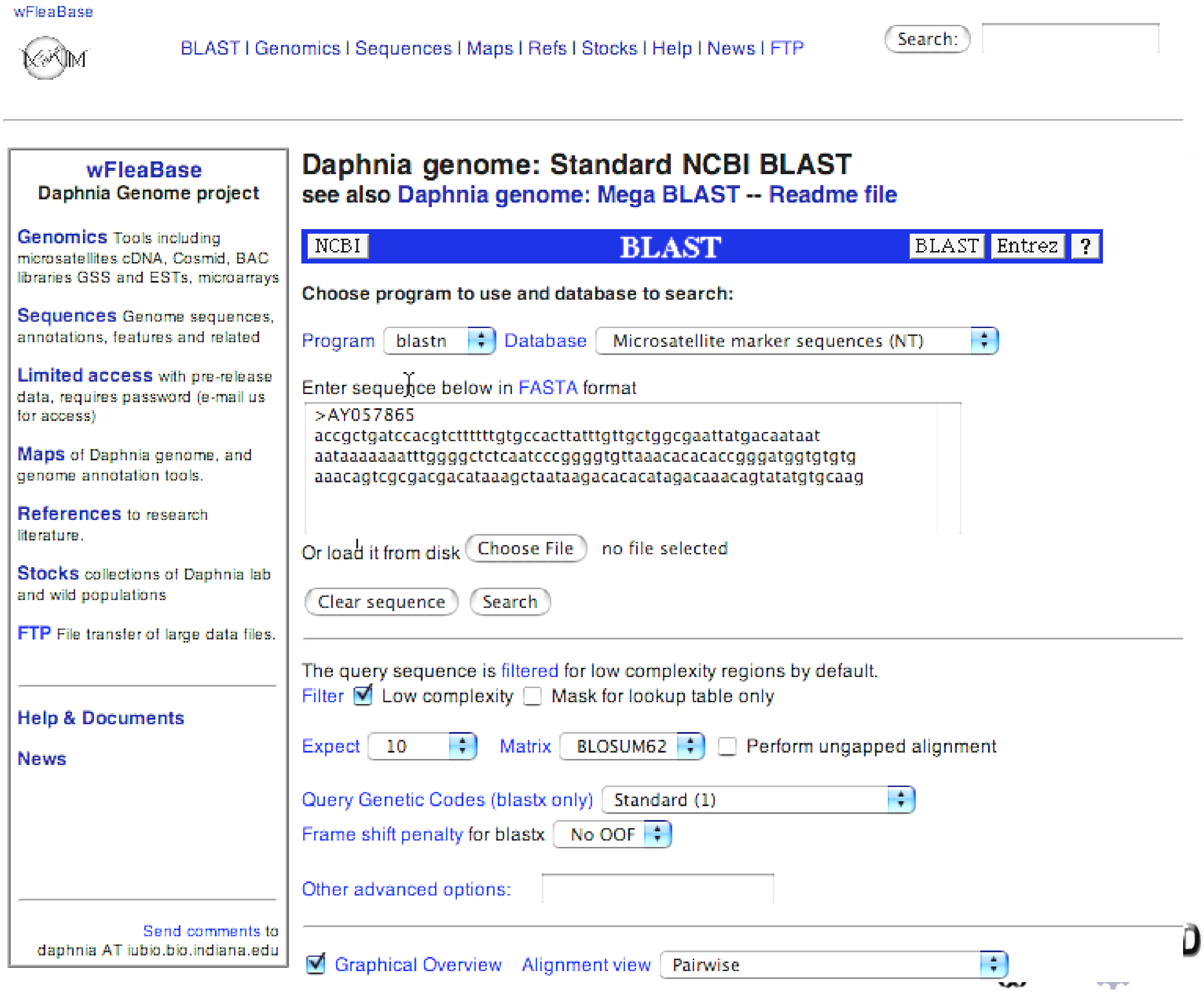
wFleaBase
Daphnia Water Flea Genome Database

This web service provides gene and genomic information for species of *Daphnia* (the water flea).

The freshwater crustacean *Daphnia* is a model system for evolutionary / ecological genetics and genomics. This database includes data from all species of the genus, yet the primary species are *D. pulex* and *D. magna* because of the broad set of genomic tools that have already been developed for these taxa.

A complete sequence of the *Daphnia pulex* genome should be made available in the year 2004 by the U.S. Department of Energy (DOE) Joint Genome Institute (JGI) in collaboration with the Environmental Protection Agency (EPA) and the ***Daphnia* Genomics Consortium** (DGC).

BLAST wFleaBase



wFleaBase

BLAST | Genomics | Sequences | Maps | Refs | Stocks | Help | News | FTP

Search:

wFleaBase
Daphnia Genome project

Genomics Tools including microsatellites cDNA, Cosmid, BAC libraries GSS and ESTs, microarrays

Sequences Genome sequences, annotations, features and related

Limited access with pre-release data, requires password (e-mail us for access)

Maps of Daphnia genome, and genome annotation tools.

References to research literature.

Stocks collections of Daphnia lab and wild populations

FTP File transfer of large data files.

Help & Documents

News

Send comments to
daphnia AT iubio.bio.indiana.edu

Daphnia genome: Standard NCBI BLAST

see also Daphnia genome: Mega BLAST -- Readme file

Choose program to use and database to search:

Program: blastn
Database: Microsatellite marker sequences (NT)

Enter sequence below in FASTA format

```
>AY057865
accgctgatccacgtcttttttgtgccacttatttgttgctggcgaattatgacaataat
aataaaaaaaatttggggctctcaatcccggggtgttaaacacacaccgggatggtgtgtg
aaacagtcgcgacgacataaagctaataagacacacatagacaaacagtatatgtgcaag
```

Or load it from disk: Choose File — no file selected

The query sequence is filtered for low complexity regions by default.

Filter: ☑ Low complexity □ Mask for lookup table only

Expect: 10
Matrix: BLOSUM62
□ Perform ungapped alignment

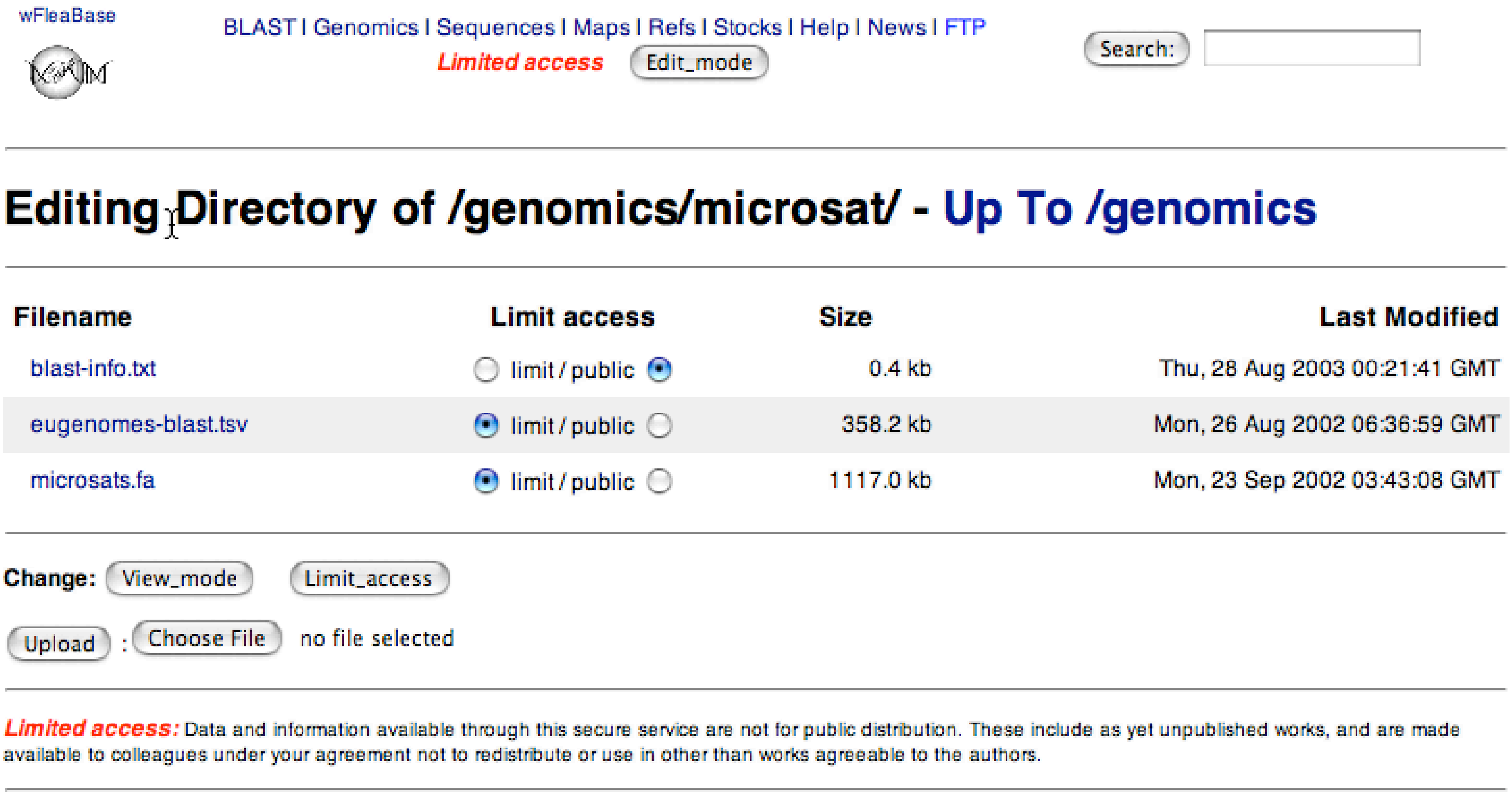
Query Genetic Codes (blastx only): Standard (1)

Frame shift penalty for blastx: No OOF

Other advanced options:

☑ Graphical Overview
Alignment view: Pairwise

Edit wFleaBase



wFleaBase

BLAST | Genomics | Sequences | Maps | Refs | Stocks | Help | News | FTP

Limited access Edit_mode

Search:

Editing Directory of /genomics/microsat/ - Up To /genomics

Filename	Limit access	Size	Last Modified
blast-info.txt	○ limit / public ●	0.4 kb	Thu, 28 Aug 2003 00:21:41 GMT
eugenomes-blast.tsv	● limit / public ○	358.2 kb	Mon, 26 Aug 2002 06:36:59 GMT
microsats.fa	● limit / public ○	1117.0 kb	Mon, 23 Sep 2002 03:43:08 GMT

Change: View_mode Limit_access

Upload : Choose File no file selected

Limited access: Data and information available through this secure service are not for public distribution. These include as yet unpublished works, and are made available to colleagues under your agreement not to redistribute or use in other than works agreeable to the authors.

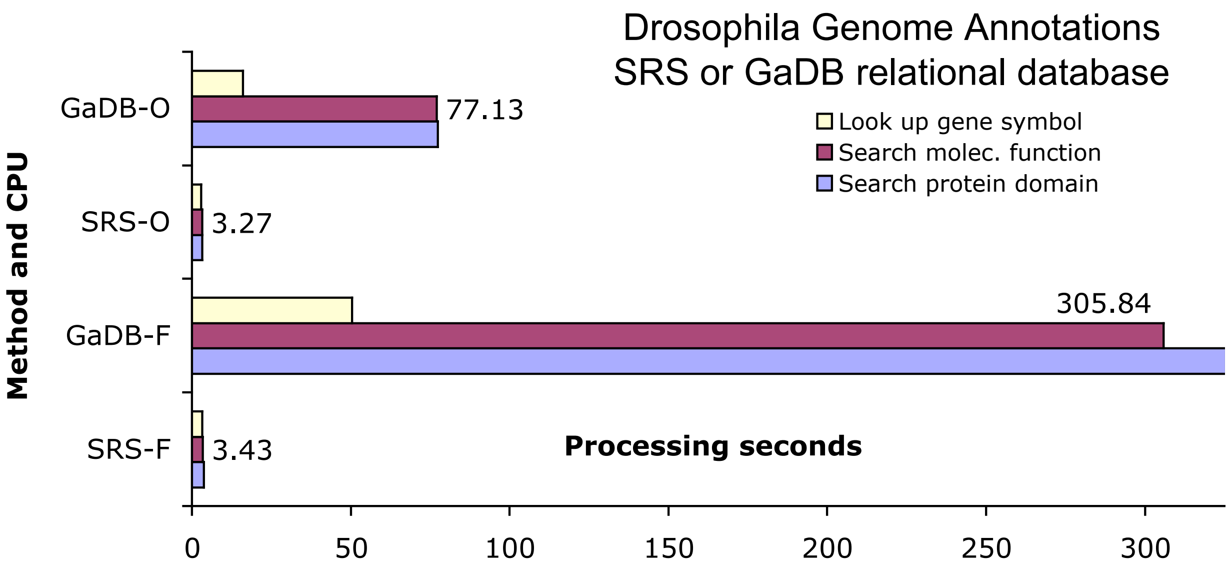
Lucegene ('Lucy Jean')

for Genome Information Search and Retrieval

Info. Retrieval for Genomes

- IR text search/retrieval tools tuned for data access, not management
- Good for a wide range of semi-structured and complex structured data
- Better functional match for textual data common in biology than numeric, table-oriented RDBMS
- Easier to add new data (e.g. SRS parses 100s of existing bio-databanks)

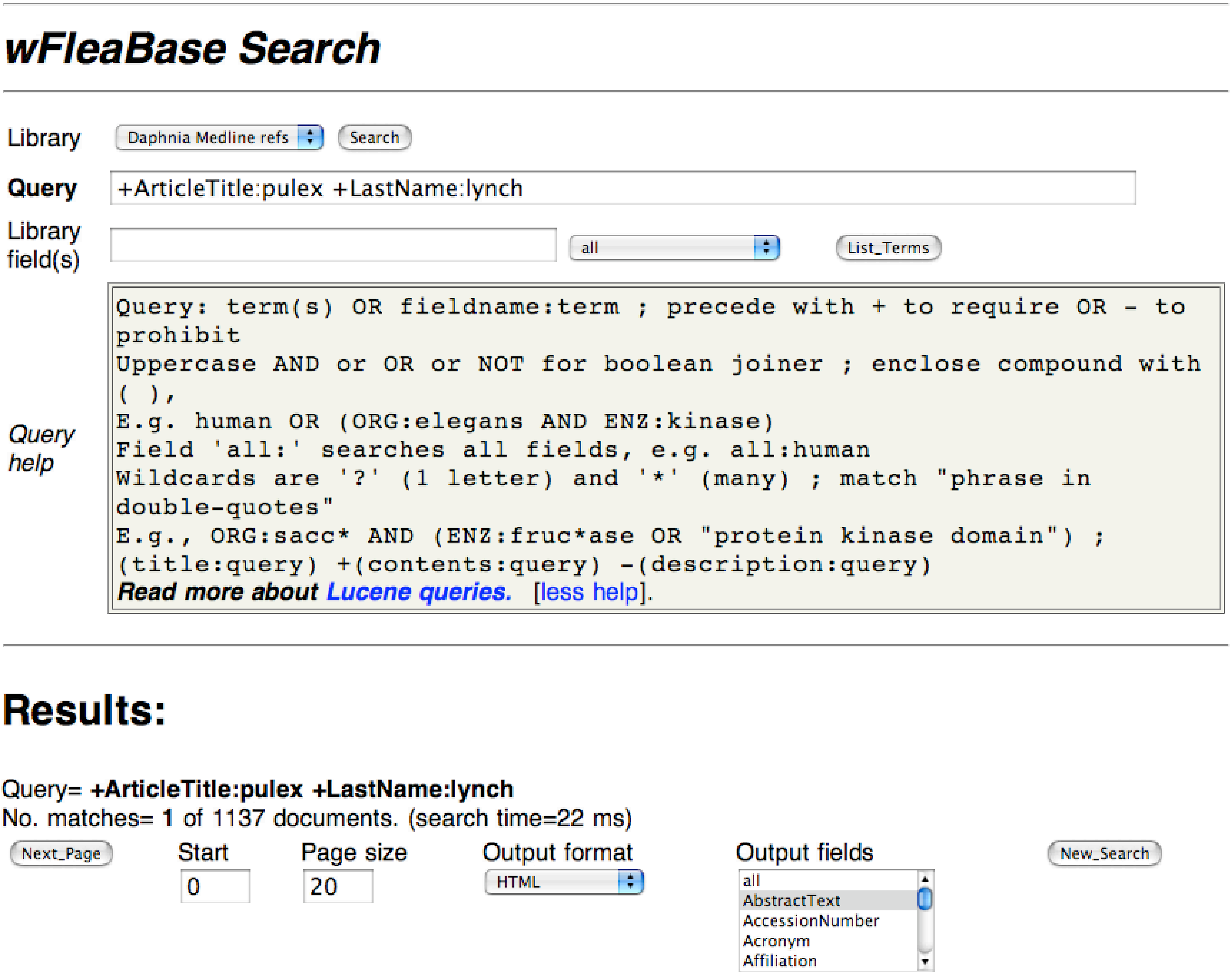
- Faster by orders of magnitude at search of complex data (no table joins; data is extremely *non-normal*)



Lucene and LuceGene

- Lucene open-source project at jakarta.apache.org/lucene
  - Common text search features: booleans, phrases, word stemming, fuzzy and field range searches, relevance ranking
  - Comparable to Glimpse, Exite, WAIS, Isearch, ht/dig, Alta-vista, Google backends
  - Author Doug Cutting has written text search engines for Apple and Excite
- LuceGene additions
  - Data input adaptors for HTML; XML (e.g. MedLine); FlyBase flatfile; Biosequences (GenBank, EMBL, etc.)
  - Basic output formats for XML, HTML via XSLT, Text, Spreadsheet
- Tested with
  - 100,000s of FlyBase Genes, References, Game and Chado XML annotations
  - euGenes gene summaries & Daphnia Medline, Sequences, HTML documents
- LuceGene/Lucene needs
  - Range search improvements (inefficient, dies w/ large range)
  - Links/joins among databases
  - Output adaptors and work? (or rely on data source formatting)

wFleaBase Search



wFleaBase Search

Library: Daphnia Medline refs | Search

Query +ArticleTitle:pulex +LastName:lynch

Library field(s): all | List_Terms

Query help

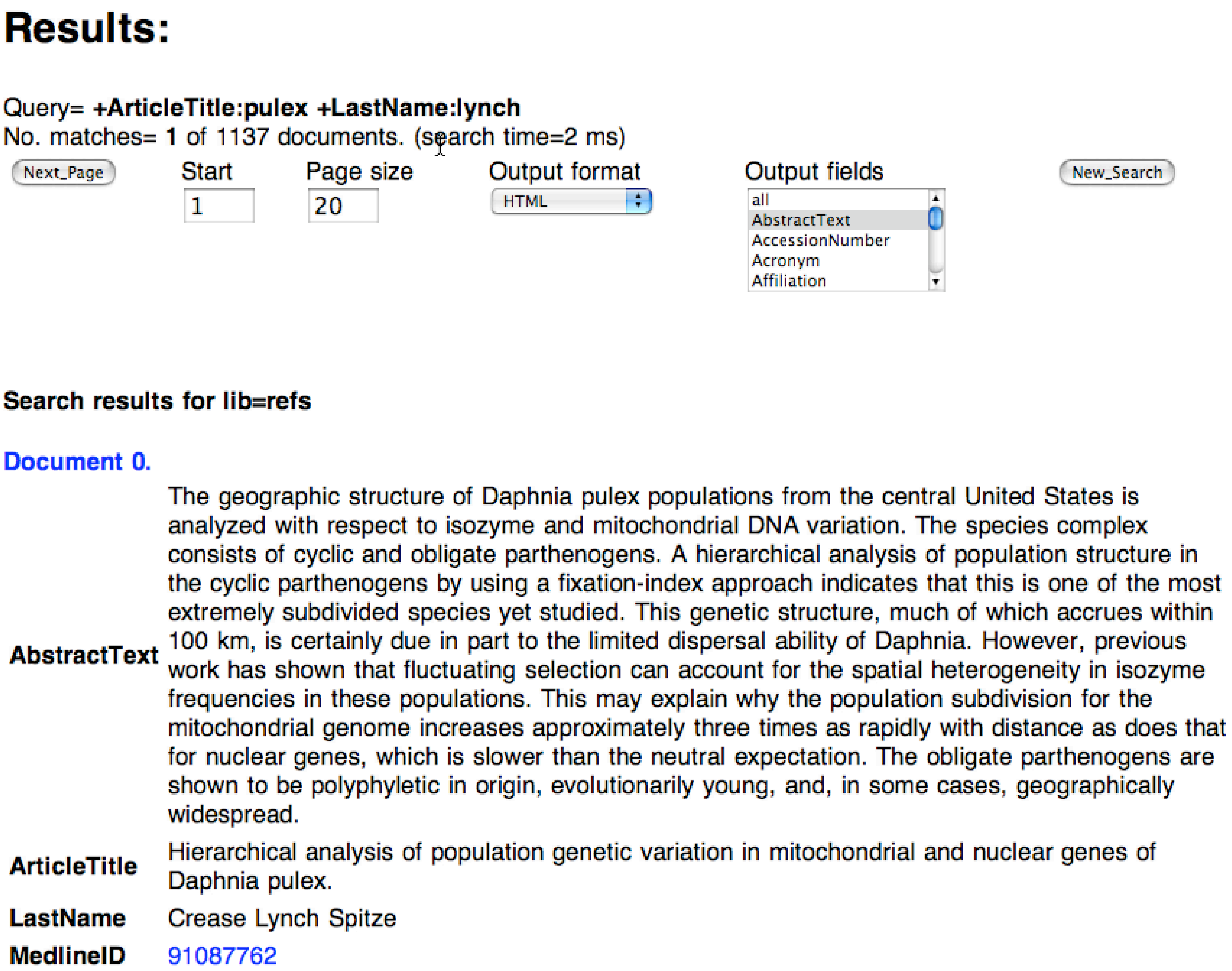
```
Query: term(s) OR fieldname:term ; precede with + to require OR - to
prohibit
Uppercase AND or OR or NOT for boolean joiner ; enclose compound with
( ),
E.g. human OR (ORG:elegans AND ENZ:kinase)
Field 'all:' searches all fields, e.g. all:human
Wildcards are '?' (1 letter) and '*' (many) ; match "phrase in
double-quotes"
E.g., ORG:sacc* AND (ENZ:fruc*ase OR "protein kinase domain") ;
(title:query) +(contents:query) -(description:query)
```

Read more about Lucene queries. [less help].

Results:

Query= **+ArticleTitle:pulex +LastName:lynch**
No. matches= **1** of 1137 documents. (search time=22 ms)

Search wFleaBase



Results:

Query= **+ArticleTitle:pulex +LastName:lynch**
No. matches= **1** of 1137 documents. (search time=2 ms)

Start: 1
Page size: 20
Output format: HTML
Output fields: all, AbstractText, AccessionNumber, Acronym, Affiliation

Search results for lib=refs

Document 0.

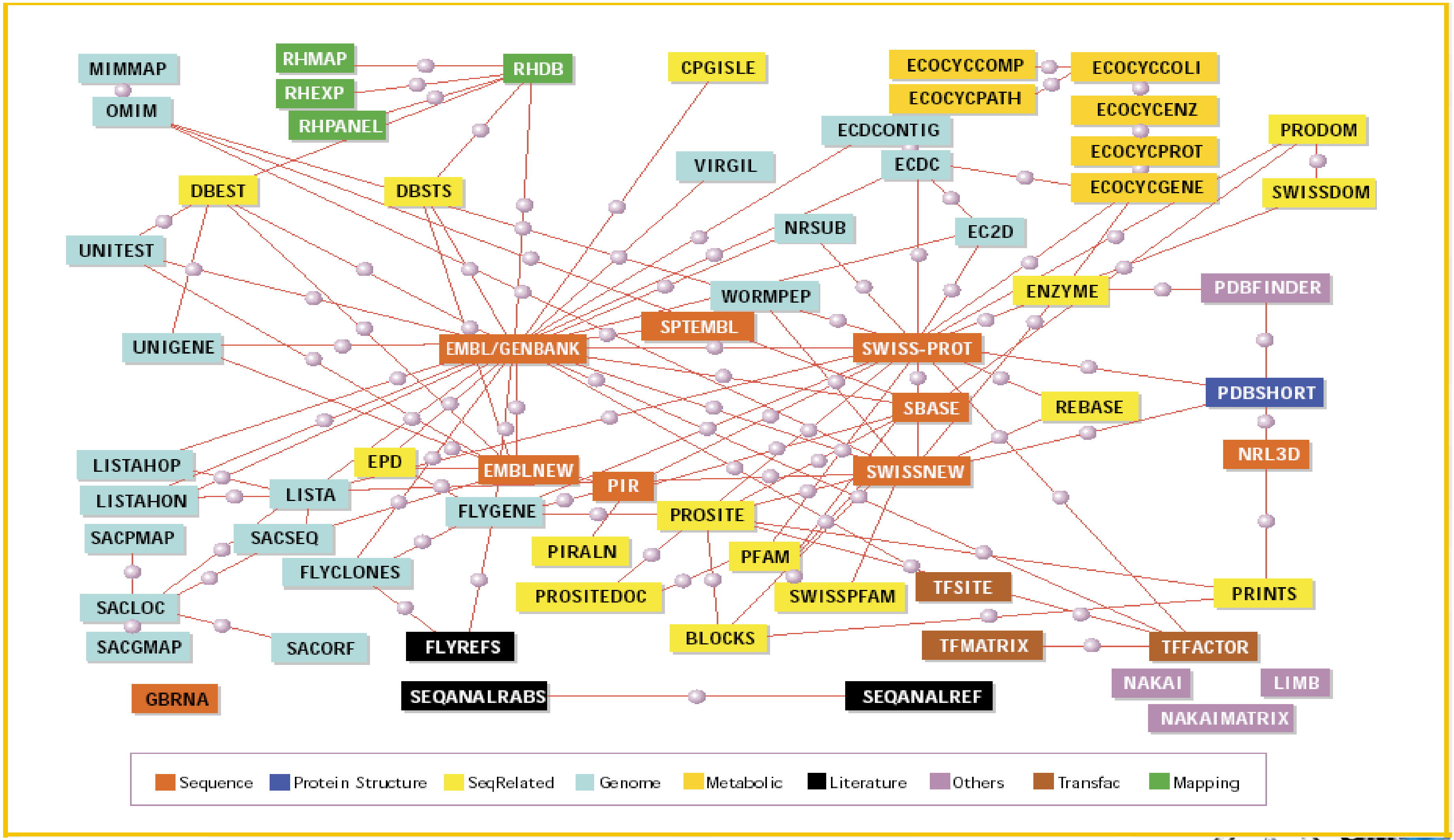
AbstractText	The geographic structure of Daphnia pulex populations from the central United States is analyzed with respect to isozyme and mitochondrial DNA variation. The species complex consists of cyclic and obligate parthenogens. A hierarchical analysis of population structure in the cyclic parthenogens by using a fixation-index approach indicates that this is one of the most extremely subdivided species yet studied. This genetic structure, much of which accrues within 100 km, is certainly due in part to the limited dispersal ability of Daphnia. However, previous work has shown that fluctuating selection can account for the spatial heterogeneity in isozyme frequencies in these populations. This may explain why the population subdivision for the mitochondrial genome increases approximately three times as rapidly with distance as does that for nuclear genes, which is slower than the neutral expectation. The obligate parthenogens are shown to be polyphyletic in origin, evolutionarily young, and, in some cases, geographically widespread.
ArticleTitle	Hierarchical analysis of population genetic variation in mitochondrial and nuclear genes of Daphnia pulex.
LastName	Crease Lynch Spitze
MedlineID	91087762

Genome Data Directories

for Data Grid and related Internet distributed search standards

Constellation of Bio-Data

(SRS - Lion Bioscience)



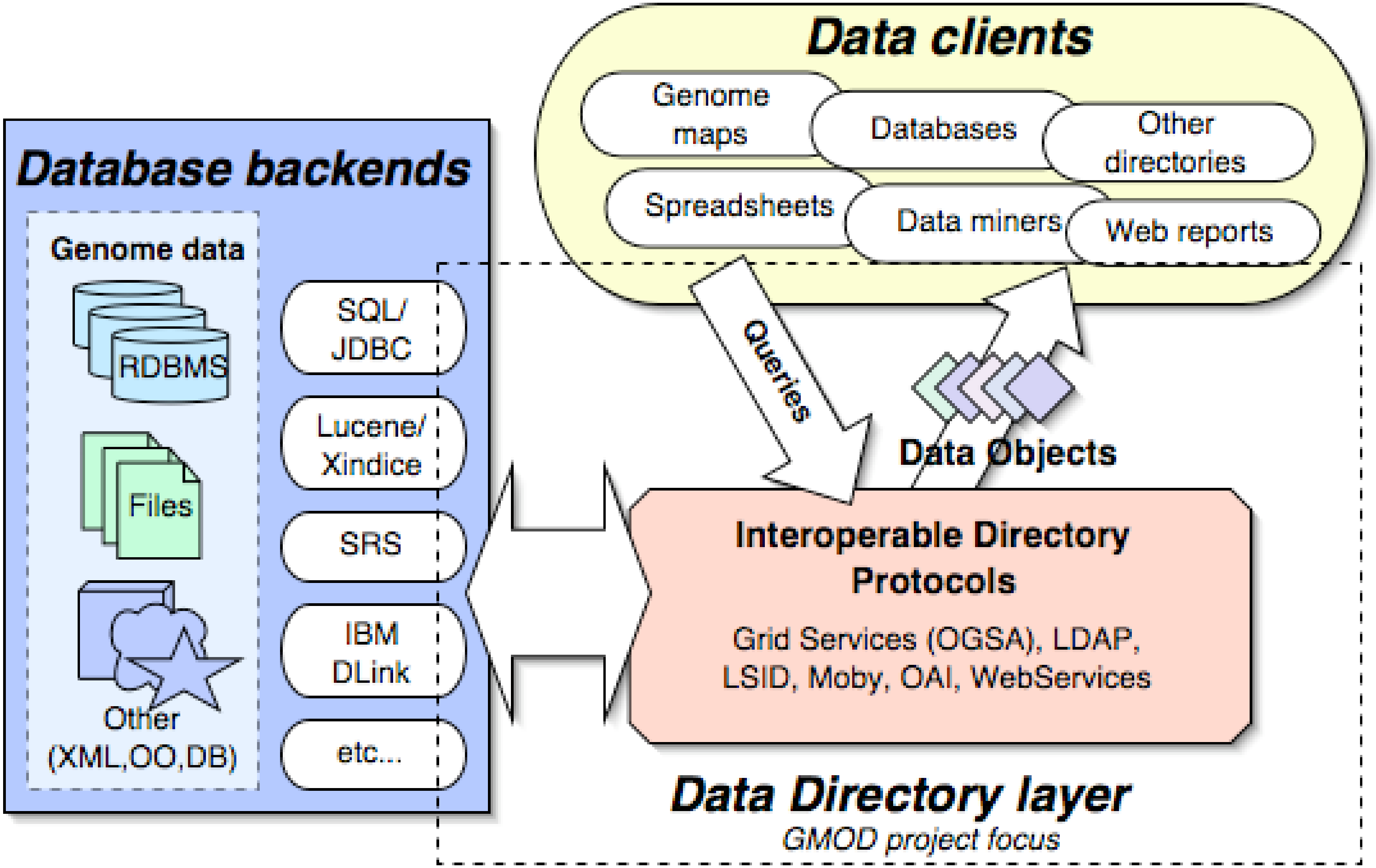
Directories of Genome Data

- **Directories are a necessary step for bio grids**
  - *"broad and shallow"* directories federate the *"narrow and deep"* databases
- **Bio-Data Access Tools**
  - SRS, Sequence Retrieval System; Entrez ; AceDB; Genome relational databases (Ensembl, FlyBase, WormBase) ; IBM DiscoveryLink; BioDAS ; BioMoby
- **Directory services for data access**
  - Layer onto access tools for common query/retrieval
  - **LDAP:** mature, efficient for high volumes, query distributed directories ; works well with bio-access tools
  - **Web Services:** XML messages over Web ; wide industry support , standards are in progress

Directory Aspects

- Build on existing technology
- Efficient for millions of objects
- Queries distributed across directories
- Support existing and new data access
- Simple client program methods
- Flexible, common schema for objects
- Replicate directories among bioinformatics centers
- Peer-to-peer directories for collaborations
- Strong authentication and security

Directory Components



Directory Standards

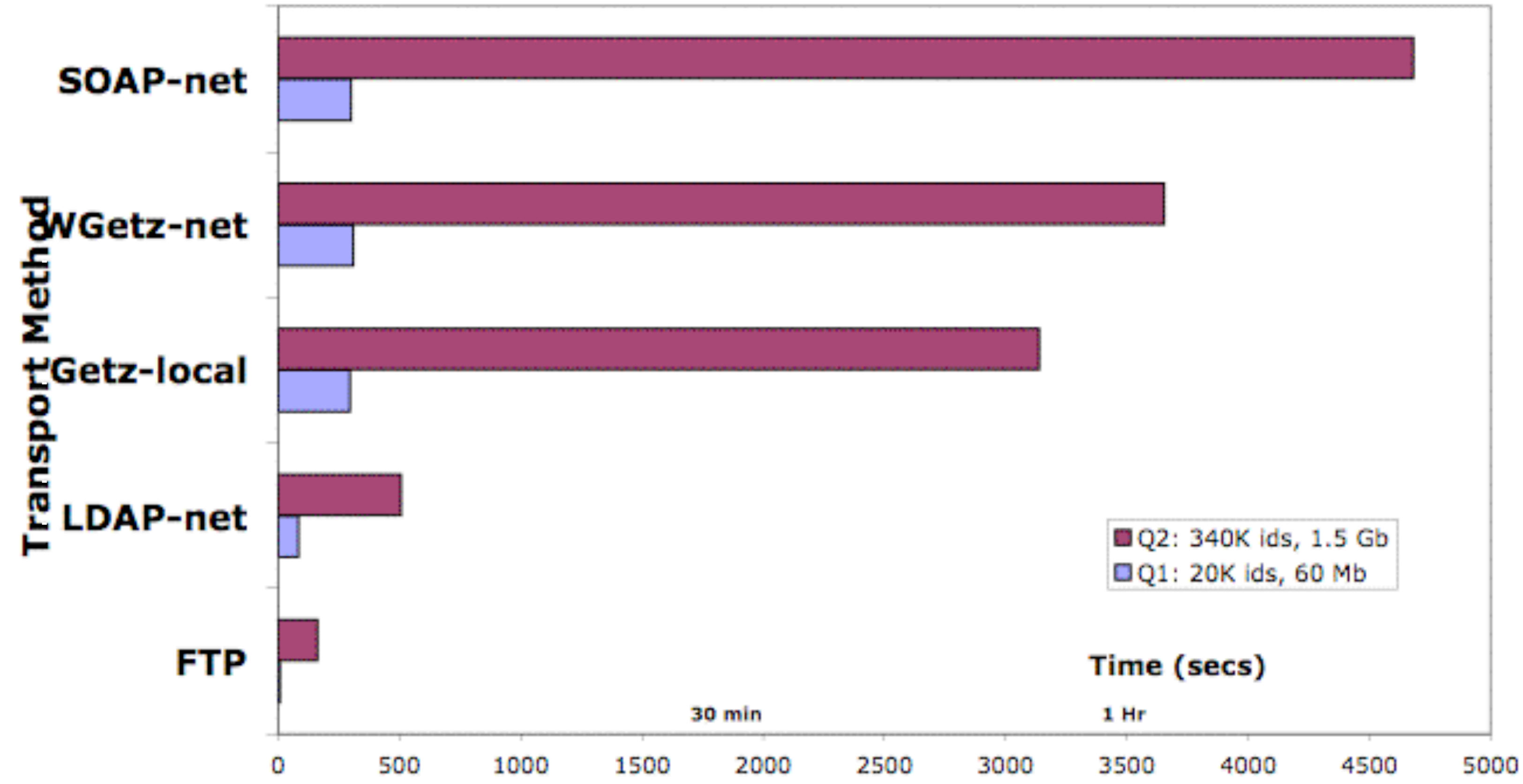
- Open Grid Services Architechture (OGSA)
  - SOAP based; query support for XML-SQL, Xpath, Xquery.
  - Data Access project: http://www.ogsa-dai.org.uk/
- Lightweight Directory Access (LDAP)
  - Robust system for distributed search and retrieval
  - Object-centric, optimized for efficient read operations
  - Hierarchical, distributed and replicated in nature
- Life Sciences ID (LSID)
  - new standard for bio-object naming, with LDAP and WebServices implementations
- Moby project web services repository system

Directory Tests

- Design and test distributed access with LDAP and Web Services
- SRS backend for efficient search/retrieval from GenBank, SwissProt/TrEMBL, LocusLink, Medline, many others
- Find & fetch 20,000 to 1.2 million objects
- LDAP is ~10x faster than WebServices
- Tests in progress for IUBio, FlyBase data

Directory Tests

BioDirectory Search/Retreive Efficiency



FTP > 5x LDAP > 5x Web	FTP	LDAP-net	Getz-local	WGetz-net	SOAP-net
■ Q2: 340K ids, 1.5 Gb	162	506	3138	3653	4680
□ Q1: 20K ids, 60 Mb	8	84	297	309	300

Q1/Q2 - Query biosequence directories ; **FTP** - no query selection
Q1 = {swissprot trembl refseq}-des:kinase , 20K records; **Q2** = genbank-org:drosophila , 340K records

gilbertd@bio.indiana.edu, Oct 2002

Directory Issues

- Basic Web-Services and LDAP access working in testing form; not stable nor finalized
- Bio-Data categorization, schema, and meta-data for directories need work
- Grid (OGSA), OAI, other interfaces to be developed

Directory tests at

http://iubio.bio.indiana.edu/biogrid/directories/