Genomes to Grids

Bio Data Distribution for Grid Computing

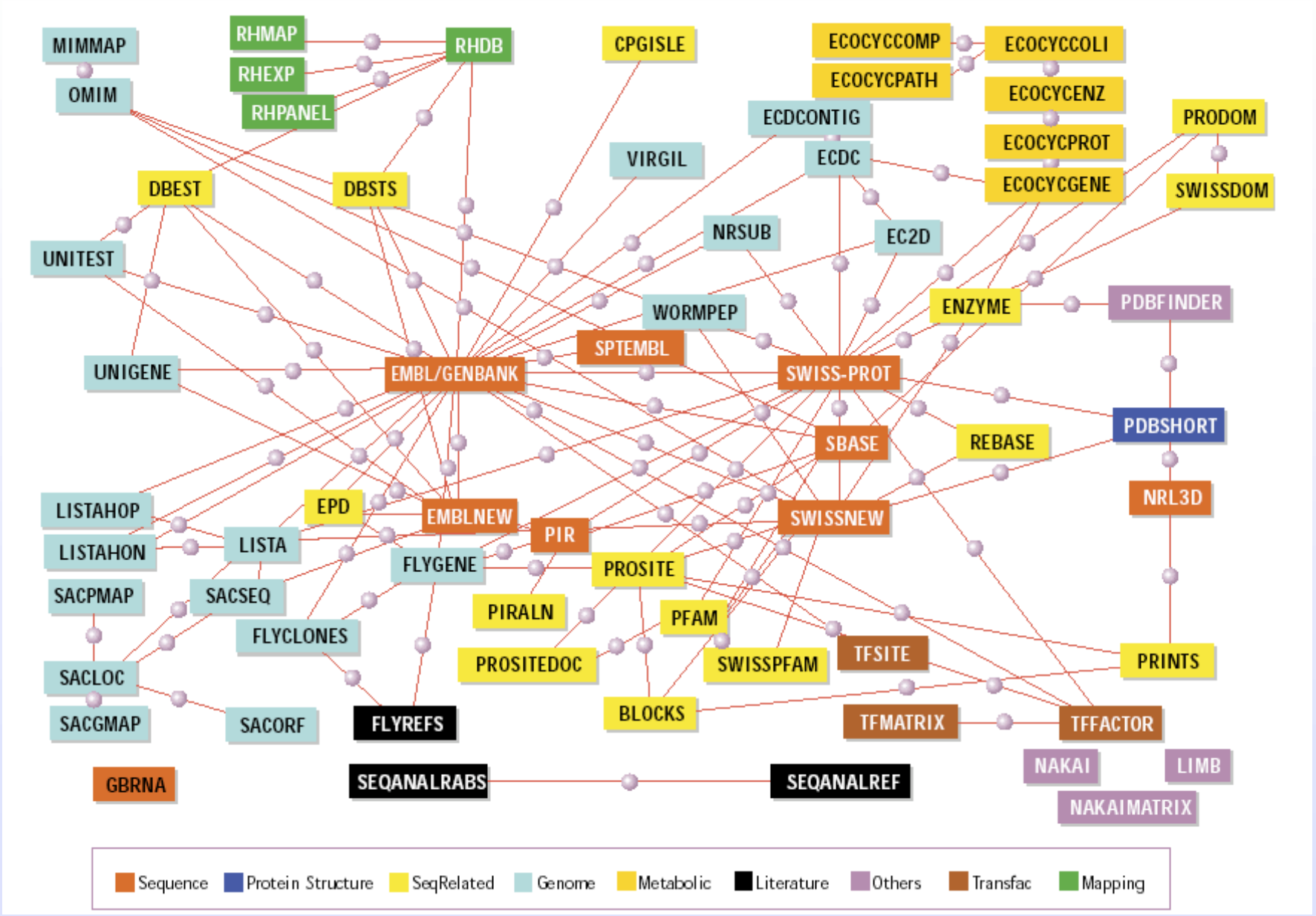
Biologists have discovered many millions of genes and genome features, now part of the bio-data "library" distributed on computers around the world. Grid computing methods for finding and using interesting genome knowledge from this mountain of data are discussed - their promise and practical concerns for building usable bioinformatics grids.

Don Gilbert, gilbertd@bio.indiana.edu, December 2002

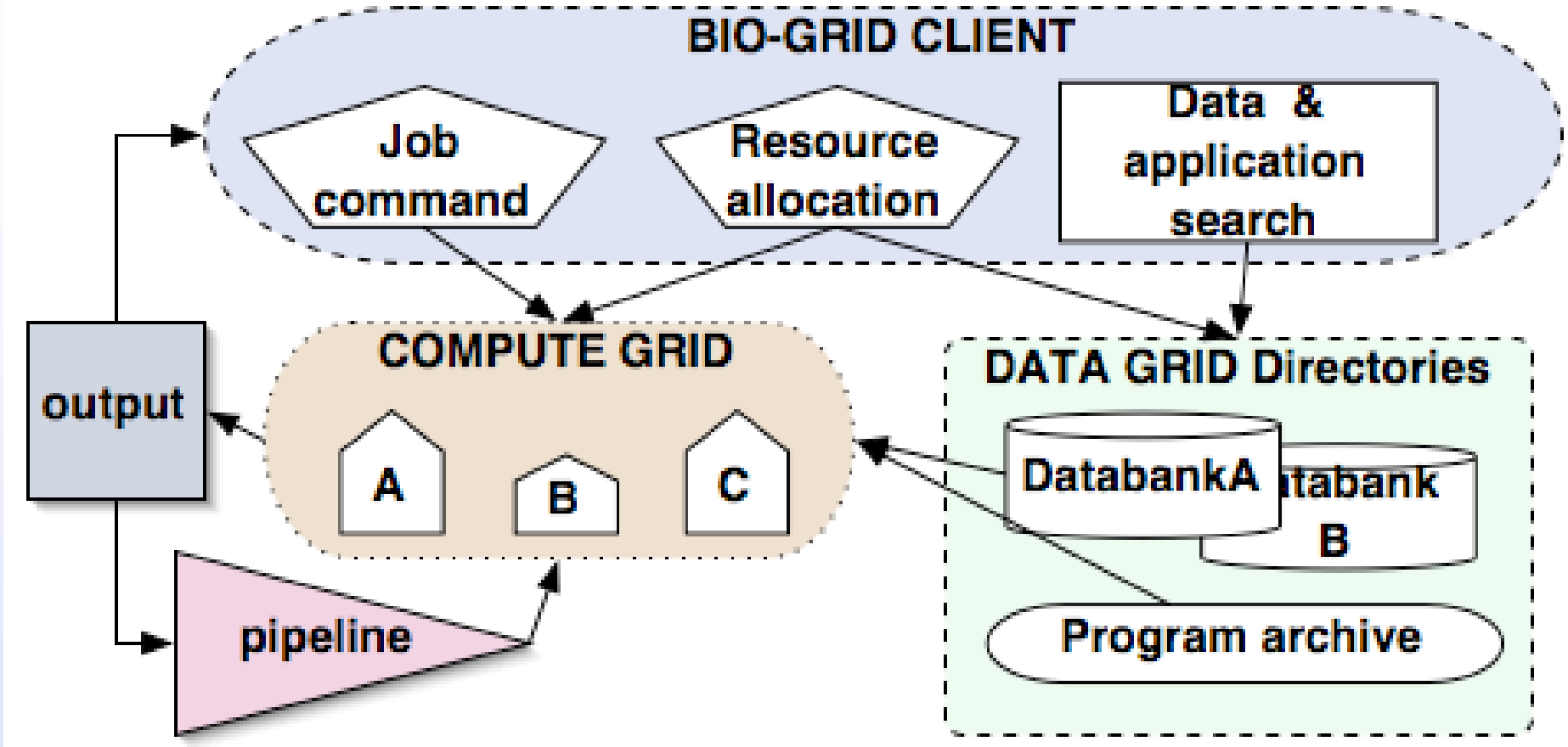
Bio Databanks, EBI, Sept. 2002

Databank	Contents	Entries
EMBL	DNA Sequences	18,800,000
SWALL	Protein sequences	900,000
InterPro+	Protein motifs	1,000,000
HGBASE	SNP database	1,500,000
	Metabolic Pathways	250,000
MEDLINE	Literature	11,350,000
Total		33,800,000

Constellation of Bio-Data (SRS - Lion Bioscience)



BioGrid Schematic



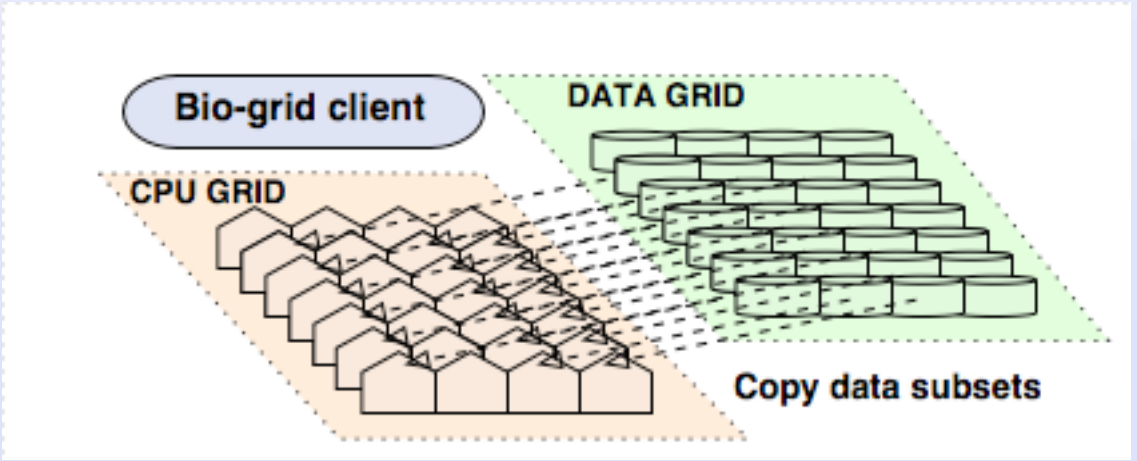
- **Grid-aware client software**
- **Data and software directories**
- **Grid of processing computers**

Directories of Genome Data

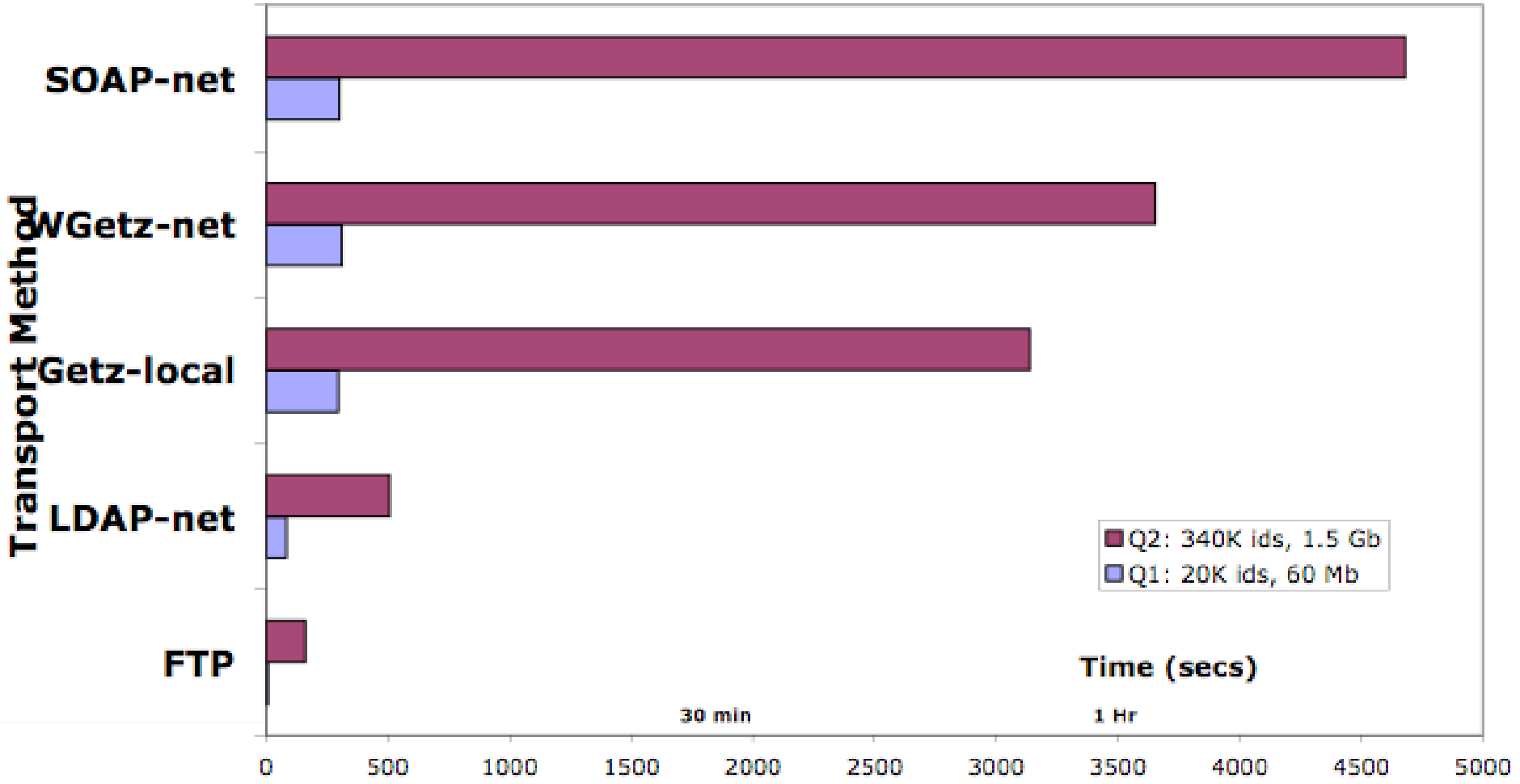
- **For genome data, *"broad and shallow"* directories federate the *"narrow and deep"* data-bases**
- ***BioData access tools***
  - ***SRS - Sequence Retrieval System; Entrez ; AceDB***
  - ***RDBMS ; Ensembl ; IBM DiscoveryLink; BioDAS ; BioMoby***
- ***Directory services - Data tools + LDAP , Web Services***
  - **LDAP:** mature, efficient for high volumes, allows federated queries over distributed directories, and works well for SRS databanks and genome annotations
  - **Web Services:** new, simple & complex for XML messages over Web ; has wide industry support , but its many standards are in flux

Moving Data on the Grid

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



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

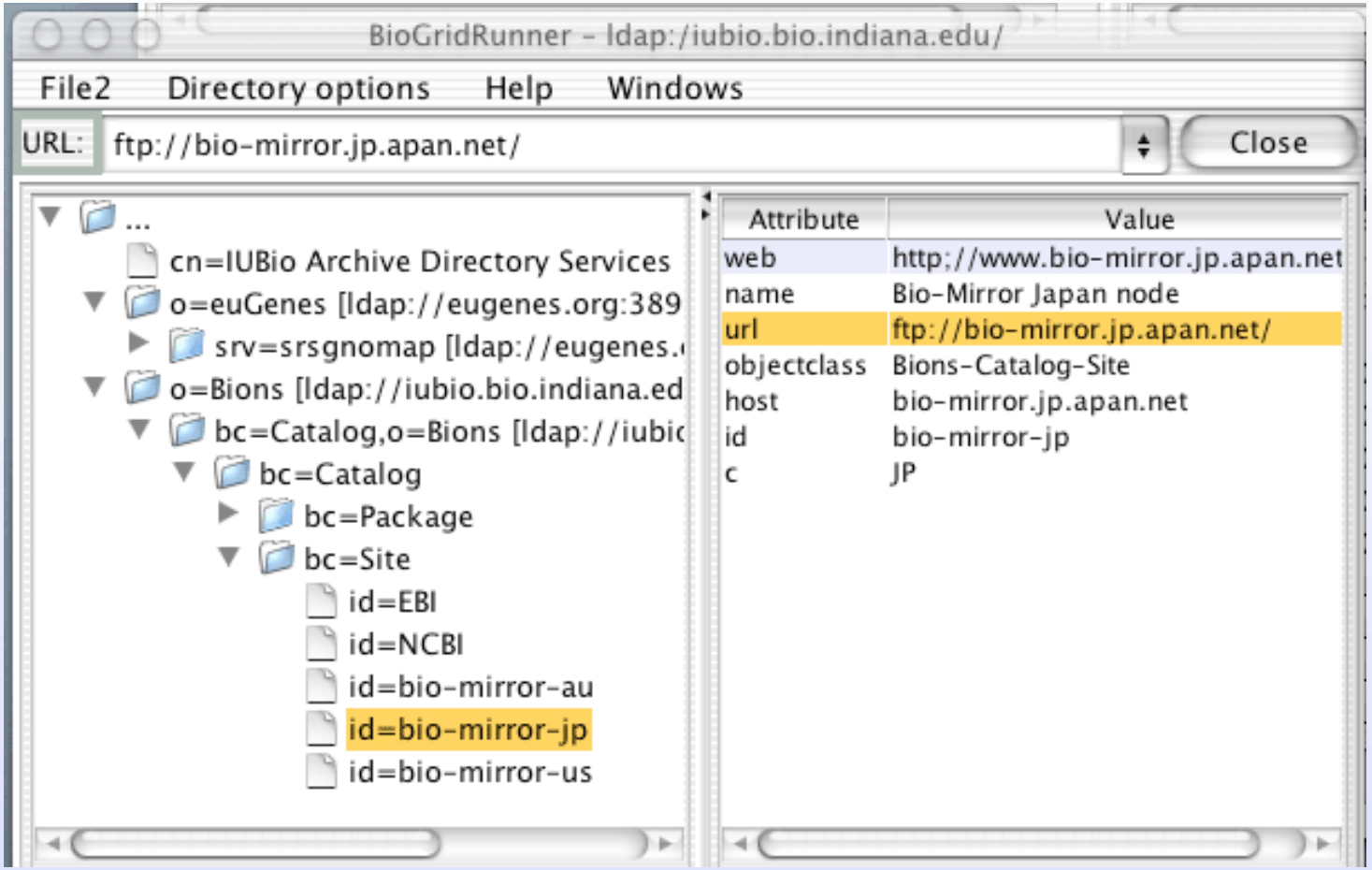
Q1/Q2 - Query biosequence directories ; **FTP** - no query selection

Q1 = {swissprot trembl refseq}-des:kinase , 20K records; **Q2** = genbank-org:drosophila , 340K records

gilbertd@bio.indiana.edu, Oct 2002

Using Bio Directories

- Simple client software
- Automated use
- People use
- Discovery
- Search by many criteria
- Retrieve bulk subsets



Wrap up

- **Future of Bio-data on Grids**
  - **High volume, complex, changing, distributed data**
  - **Computationally find and use this data**
- **Best methods for Bio-data to Grids**
  - **Efficient selection and transport to grid computers**
  - **LDAP works well ; Web-XML is usable ; Others?**
- **Community needs and uses**
  - **Common data descriptions, schema, ontologies**
  - **Simple, practical, flexible grid methods ; use existing dbs**

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