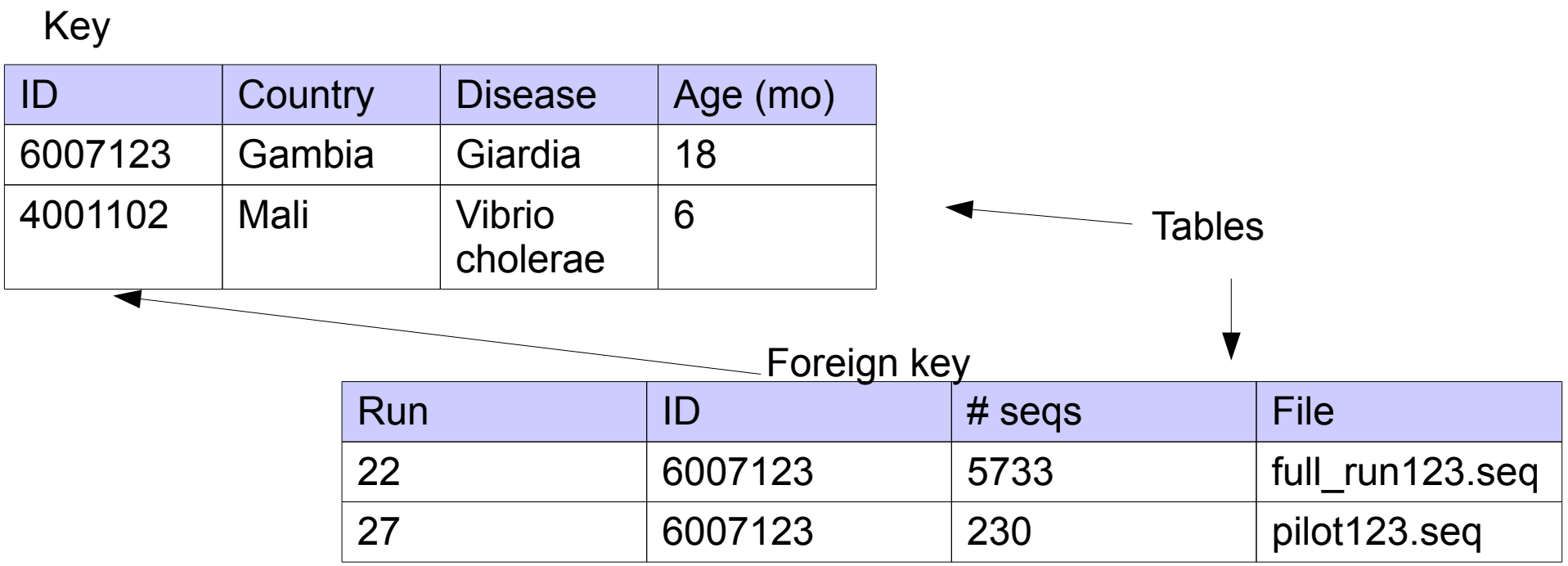
CMSC423: Bioinformatic Algorithms, Databases and Tools

Biological databases

What's a database?

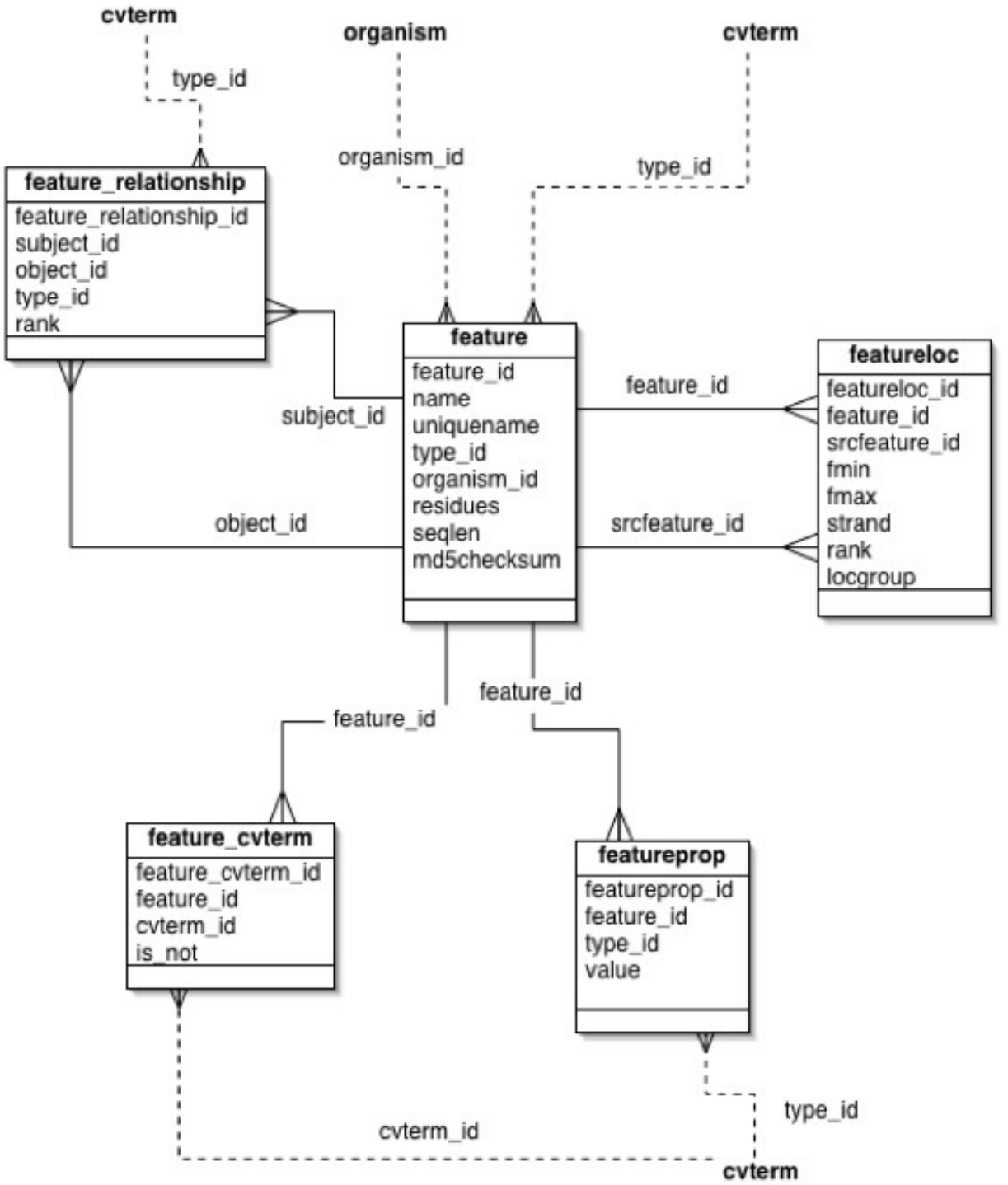
- Take CMSC424 for in-depth view
- Essentially a collection of Excel sheets or tables
  (note: only true for the “relational model” - most popular)



Chado

- http://www.gmod.org
- Relational schema for storing biological data types in a relational database (e.g. MySQL, Oracle, Sybase, ...)

```
SELECT o.organism_id,o.abbreviation,o.genus,o.species,
            o.common_name,  count(f.feature_id) as n_features,
            o.comment
  FROM organism o LEFT JOIN feature f USING (organism_id)
  GROUP by o.organism_id,o.abbreviation,o.genus,o.species,
                  o.common_name,o.comment
  ORDER BY o.genus,o.species
```


cvterm
type_id
organism
organism_id
cvterm
type_id
feature_relationship
feature_relationship_id
subject_id
object_id
type_id
rank
feature
feature_id
name
uniquename
type_id
organism_id
residues
seqlen
md5checksum
featureloc
featureloc_id
feature_id
srcfeature_id
fmin
fmax
strand
rank
locgroup
subject_id
feature_id
object_id
srcfeature_id
feature_id
feature_id
feature_cvterm
feature_cvterm_id
feature_id
cvterm_id
is_not
featureprop
featureprop_id
feature_id
type_id
value
type_id
cvterm_id
cvterm

Chado...more

- Bio... generally provide ability to interface with relational database.
- Understanding SQL and Chado is useful irrespective of language used.
- Relational DB particularly useful for web services
- Gbrowse example....if time

Biological databases

- General
  - GenBank - US
  - EMBL - Europe
- Specialized by data type
  - NCBI Trace Archive – raw sequencing data
  - SwissProt – curated protein information
  - KEGG – biological pathways
  - Gene Expression Omnibus – microarray data
- Specialized by organism
  - ZFIN – zebrafish
  - SGD – yeast
  - WormBase - worms

What data gets stored?

- DNA
  - string of letters
  - quality information, maybe chromatograms
  - location of genes (ranges along a chromosome)
- Proteins
  - string of letters
  - protein domains
  - 3D coordinates of each atom
- Pathways
  - graph of interactions between genes

For all – often store link to scientific articles related to data

How the data get accessed

- Gene by gene/object by object – targeted at manual inspection of data
  - usually lots of clicking involved
  - simple search capability
  - similarity searches in addition to text queries
- Bulk – targeted at computational analyses
  - often programmatic access through web server
  - most frequently – just bulk download (ftp)

NCBI - National Center for Biotech. Info.

- Virtually all biological data generated in the US gets stored here!
- One-stop-shop for biological data
- Primarily focused on gene-by-gene analyses
- Provides simple scripts for programmatic access
- Provides ftp access for bulk downloads

http://www.ncbi.nlm.nih.gov

EMBL European Molecular Biology Lab.

- European version of NCBI
- BioMart query builder

http://www.ebi.ac.uk/embl/

Expasy proteomics server

- Home of Swisprot and other useful information on proteins

http://www.expasy.org

Kyoto Encyclopedia of Genes & Genomes

- Central repository of pathway information

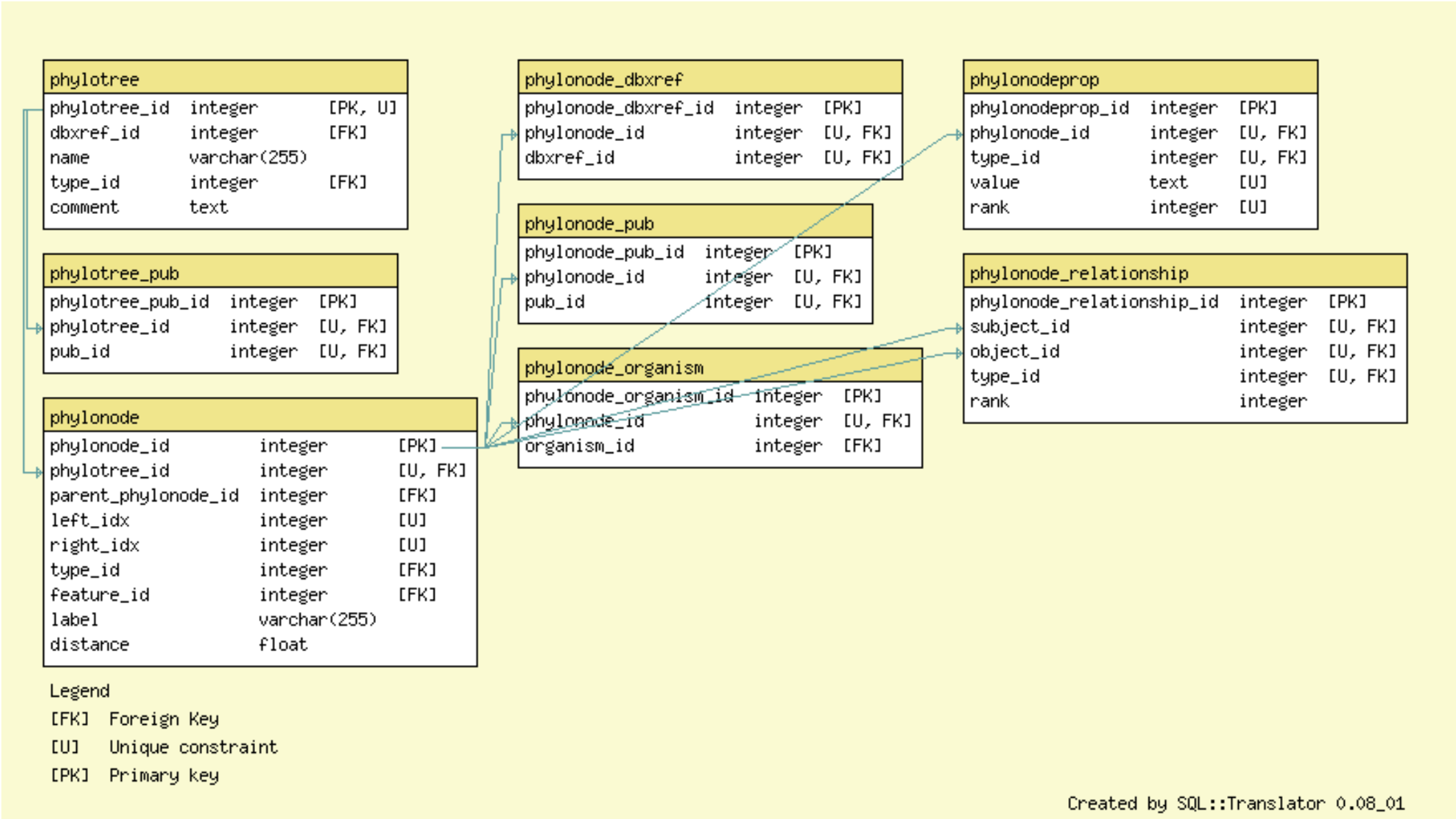
http://www.genome.jp/kegg/

Genome browsers

- UCSC Genome Browser – http://genome.ucsc.edu
- ENSEMBL Genome Browser – http://www.ensemble.org
- Gbrowse http://www.gmod.org

Direct database access - SQL

- CHADO schema – www.gmod.org



SQL

```
select pt.phylotree_id, pn.parent_phylonode_id, po.organism_id
from phylotree pt, phylonode pn, pylonode_organism po
where
  pt.name = "Archaea" and
  pt.phylotree_id = pn.phylotree_id and
  pn.phylonode_id = 1000 and
  po.phylonode_id = pn.parent_phylonode_id


# Selects parent node and organism IDs for archaeon with ID 1000
```

Programmatic database access

```
use DBI;

my $dbh = DBI->connect("dbi:Sybase:server=SERV;packetSize=8092",
                       "anonymous", "anonymous");
if (! defined $dbh) {
      die ("Cannot connect to server\n");
}

my $mysqlqry = <STDIN>;

$dbh->do("set textsize 65535");

my $qh = $dbh->prepare($mysqlqry) || die ("Cannot prepare\n");
$qh->execute() || die ("Cannot execute\n");

while (my @row = $qh->fetchrow()){
      processrow($row);
}
```

NCBI programmatic access

- http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
  - must write your own HTTP client (LWP Perl module helps)
  - queries go directly to web server
  - data returned in XML
- http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=doc&m=obtain&s=stips
  - stub script provided (query_tracedb)
  - queries still go through web server
  - data returned in a variety of user selected formats
- For both, limits are set on the amount of data retrieved, e.g. less than 40,000 records at a time
- Download procedure:
  - figure out # of records to be retrieved ("count" query)
  - read data in allowable chunks
  - combine the chunks

Biological Ontologies

- Gene Ontology. http://www.geneontology.org
  The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. (text from GO homepage)

- Note: similar to semantic web
- GO not the only one: http://www.obofoundry.org

Exercises

- Write a simple script that wraps around the query_tracedb script from NCBI and allows a user to download an entire data-set without worrying about “page limits”.

  Note: set your page limit to 4000 records (rather than the 40,000 allowed) so as not to overload the NCBI servers.
- Create a FASTA file containing all recA genes found in bacteria. Note: you can use a combination of manual queries and additional scripts (sometimes an NCBI query doesn't quite return what you want)