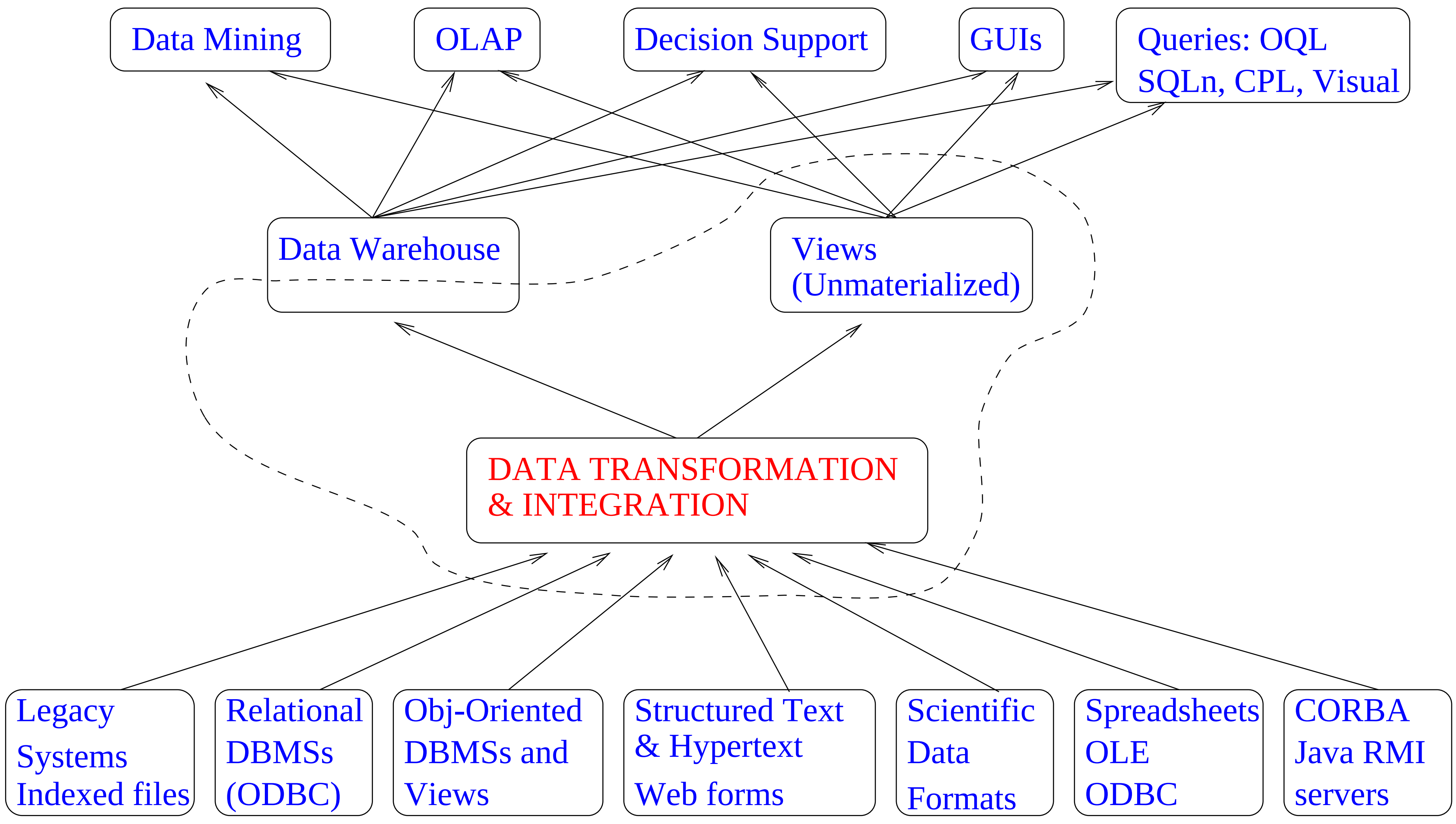
Integrating Biological (and other) Databases

Peter Buneman

University of Pennsylvania

Molecular biology databases are:

- Numerous, and growing in number.
- Decentralized, and becoming more so.
- Evolving in structure.
- Highly cross-referenced.
- Heavily "curated".


Data Mining
OLAP
Decision Support
GUIs
Queries: OQL
SQLn, CPL, Visual
Data Warehouse
Views
(Unmaterialized)
DATA TRANSFORMATION
& INTEGRATION
Legacy
Systems
Indexed files
Relational
DBMSs
(ODBC)
Obj-Oriented
DBMSs and
Views
Structured Text
& Hypertext
Web forms
Scientific
Data
Formats
Spreadsheets
OLE
ODBC
CORBA
Java RMI
servers

The variety of molecular biology databases.

From: `http://www.infobiogen.fr/services/dbcat/`

Domain	Number of DBs
DNA	59
RNA	20
Protein	72
Genomic	57
Mapping	28
Protein structure	18
Literature	37
Miscellaneous	113

Growing at 30-40 per year!

```
ID   11SB_CUCMA     STANDARD;      PRT;   480 AA.
AC   P13744;
DT   01-JAN-1990 (REL. 13, CREATED)
DT   01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE)
DT   01-NOV-1990 (REL. 16, LAST ANNOTATION UPDATE)
DE   11S GLOBULIN BETA SUBUNIT PRECURSOR.
OS   CUCURBITA MAXIMA (PUMPKIN) (WINTER SQUASH).
OC   EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE;
OC   VIOLALES; CUCURBITACEAE.
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. KUROKAWA AMAKURI NANKIN;
RX   MEDLINE; 88166744.
RA   HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
RL   EUR. J. BIOCHEM. 172:627-632(1988).
RN   [2]
RP   SEQUENCE OF 22-30 AND 297-302.
RA   OHMIYA M., HARA I., MASTUBARA H.;
RL   PLANT CELL PHYSIOL. 21:157-167(1980).
CC   -!- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC   -!- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC       BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC       DISULFIDE BOND.
CC   -!- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
DR   EMBL; M36407; G167492; -.
DR   PIR; S00366; FWPU1B.
DR   PROSITE; PS00305; 11S_SEED_STORAGE; 1.
KW   SEED STORAGE PROTEIN; SIGNAL.
FT   SIGNAL        1     21
FT   CHAIN        22    480       11S GLOBULIN BETA SUBUNIT.
FT   CHAIN        22    296       GAMMA CHAIN (ACIDIC).
FT   CHAIN       297    480       DELTA CHAIN (BASIC).
FT   MOD_RES      22     22       PYRROLIDONE CARBOXYLIC ACID.
FT   DISULFID    124    303       INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT   CONFLICT     27     27       S -> E (IN REF. 2).
FT   CONFLICT     30     30       E -> S (IN REF. 2).
SQ   SEQUENCE   480 AA;  54625 MW;  D515DD6E CRC32;
     MARSSLFTFL CLAVFINGCL SQIEQQSPWE FQGSEVWQQH RYQSPRACRL ENLRAQDPVR
     RAEAEAIFTE VWDQDNDEFQ CAGVNMIRHT IRPKGLLLPG FSNAPKLIFV AQGFGIRGIA
     IPGCAETYQT DLRRSQSAGS AFKDQHQKIR PFREGDLLVV PAGVSHWMYN RGQSDLVLIV
     FADTRNVANQ IDPYLRKFYL AGRPEQVERG VEEWERSSRK GSSGEKSGNI FSGFADEFLE
     EAFQIDGGLV RKLKGEDDER DRIVQVDEDF EVLLPEKDEE ERSRGRYIES ESESENGLEE
     TICTLRLKQN IGRSVRADVF NPRGGRISTA NYHTLPILRQ VRLSAERGVL YSNAMVAPHY
     TVNSHSVMYA TRGNARVQVV DNFGQSVFDG EVREGQVLMI PQNFVVIKRA SDRGFEWIAF
     KTNDNAITNL LAGRVSQMRM LPLGVLSNMY RISREEAQRL KYGQQEMRVL SPGRSQGRRE
//
```

Metadata

An Entry in Swissprot

Data

The Complexity of Metadata

```
...
AC   P13744;
OC   EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE;
OC   VIOLALES; CUCURBITACEAE.
RN   [1]
...
RA   HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
RL   EUR. J. BIOCHEM. 172:627-632(1988).
RN   [2]
...
RA   OHMIYA M., HARA I., MASTUBARA H.;
RL   PLANT CELL PHYSIOL. 21:157-167(1980).
...
```

Just the citation and classification data requires *four* tables and two additional ranking fields in a relational representation.

Simple queries require five-way joins!

Database Cross References

```
ID   11SB_CUCMA     STANDARD;      PRT;   480 AA.
AC   P13744;
...

RX   MEDLINE; 88166744.
...
DR   EMBL; M36407; G167492; -.
DR   PIR; S00366; FWPU1B.
DR   PROSITE; PS00305; 11S_SEED_STORAGE; 1.
...
```

Cross references to eight other databases

Schema Evolution

```
...
CC   -!- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC   -!- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC       BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC       DISULFIDE BOND.
CC   -!- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
...
```

Structure is added to the comment fields which now contain "controlled" keywords.

CPL/Kleisli

A database integration system that exploits a generic approach to complex types. It comprises

- A query language (CPL). Also OQL.
- A type inference system – robustness against schema modification.
- An optimization module – capable of "offloading" work.
- Drivers for numerous sources:
  - Relational: Sybase and Oracle
  - OO: ObjectStore, Shore, O_2
  - ASN.1 – generic interchange format
  - SRS – a file indexing system
  - Sequence matching packages: FASTA, BLAST
  - ACEDB and other "lightweight" object managers.
  - Various non-generic data formats
  - In the works: CORBA, XML

A Kleisli OQL Query for a fragment of Swissprot

The first author on any paper that studies the same plant protein as `"TELL, W."`:

```
SELECT first(r2.Auths)
FROM   Swissprot e,
       e.Refs r1,
       e.Refs r2,
WHERE  "PLANTA" in e.Taxons
       r1.Refnum <> r2.Refnum
       "TELL, W." in r1.Auths
```

This requires a 5-way join in a relational representation of the same data.

A Surprisingly Hard Query

Is my protein sequence already patented?

Databases involved:

Name	Type	Where
SCOP	Indexed Files	Cambridge, UK
NCBI	Indexed Data Exchange Format (ASN.1)	Baltimore, USA
ISS Patent-aa	Flat file	ISS, Singapore
ISS Scop-aa	Flat file	ISS, Singapore
NCBI Blast	Homolog Search Program	Baltimore, USA
WU Blast	Homolog Search Program from Wash. U.	Singapore
USPTO Isite	WWW patent text search	Washington, USA

None of these is relational!

What's next?

- Understanding inter-database constraints for
  - Further optimization
  - Integrity and updates
- Data Provenance: How did this piece of data get into my database?
- Data models and ontologies. Making complex models understandable to scientists.
- "Plug and play" data transformation.
- A unified data format/exchange protocol? Useful, but it's not the main problem.