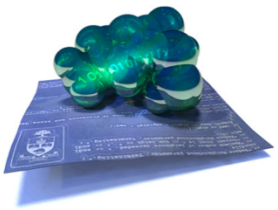
A
Bioinformatics
Course

Data Integration



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

Databases need **Identifiers** for their records ...

- ... identifiers that will **uniquely** refer to (retrieve) specific entities (sequence records)
- ... identifiers that can **cross-reference** secondary aspects of a record (features, annotations)
- ... identifiers that are **stable** through time
- ... identifiers that can **track** a history of sequence updates (versions)

Modified from Francis Ouellette, Canadian Bioinformatics Workshop 2006, Database lecture

Unique, stable, traceable identidiers are the key to unlock information resources, and to construct a network of relationships.

Selected Unique Identifiers in Biological Data

Data Type	Unique Identifier	Example	Databases
Biomolecule (DNA, Protein)	Accession Number.Version	DAA11800.1 P39678 NP_010227.1	GenBank UniProt RefSeq
Macromolecular Structure	PDB Code	1BM8 1L3G	PDB
Taxonomy Tree Node	TaxID	559292 (S. cerevisiae S288C)	NCBI/EMBL/DDBJ Taxonomy
Descriptive terms	GOID	GO:0071931	Gene Ontology

All of the above identifiers uniquely specify an information item in their respective database, they are all related to yeast Mbp1.

GO:0071931 is "positive regulation of transcription involved in G1/S transition of mitotic cell cycle" – one of the terms in the Biological Process Ontology of the GO project.

What's in a name ... ?

(All these IDs refer to essentially the same biological entity.)

`SWI4`	Standard name
`ART1`	Synonym
`YER111C`	Systematic name (*Saccharomyces* Genome Database)
`P25302`	Swiss-Prot / UniProt ID / GenPept accession number
`SWI4_YEAST`	Swiss-Prot name
`X51606`	Nucleotide accession number of the gene (Genbank)
`SCSWI4`	Locus name of the gene (Genbank)
`AAC03209`	Protein accession number (Genbank), from gene translation
`GI:603350`	GeneInfo: unique NCBI-internal identifier for `AAC03209.1`
`CAA35949`	Protein accession number (Genbank), imported from EMBL
`NP_011036`	RefSeq ID of the protein
`NP_011036.1`	RefSeq ID with version-number

It is often useful to recognize the database from the identifier. In particular you should be able to recognize SwissProt, RefSeq and PDB identifiers.

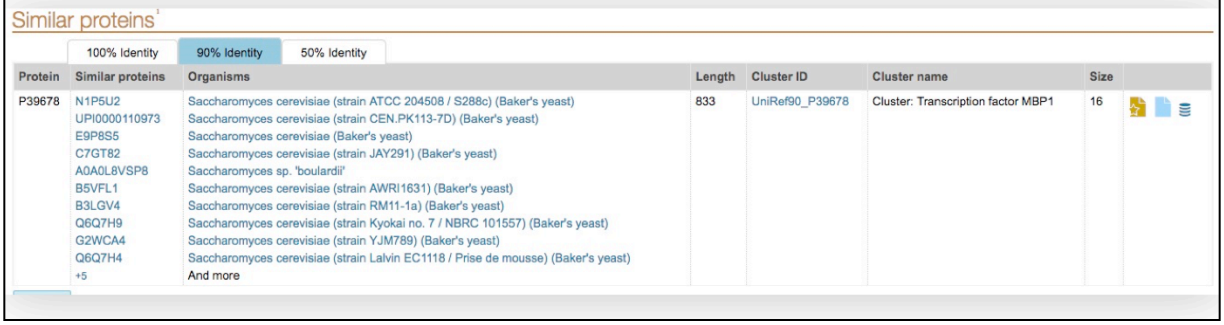
The “Systematic Name” YER111C also happens to be a “Locus” identifier, since it is constructed from the ID of the chromosome, and the index of the ORF among all ORFs, counting outward from the centromere. YER111C is a Yeast gene (“Y”); on chromosome V (five; “E” is the fifth letter in the alphabet); it is on the “R”ight arm of the chromosome; it is the “111”$^{\text{th}}$ ORF counting outwards; and it is encoded on the the “C”rick-strand: the (-)-strand, or bottom strand, i.e. the coding sequence is the reverse complement of the chromosome sequence that is deposited in the database.

Data Redundancy

The NCBI RefSeq project collects identical sets of sequences sequenced from the same organism into a single identifier.

NT_123456	:	Genomic contig	
NM_123456	:	mRNA	
NP_123456	:	Protein	
XM_123456	:	mRNA	"Hypothetical", i.e. computationally derived from genome annotation
XP_123456	:	Protein	

At the EBI, UniRef clusters group redundant sequences at various levels of similarity.



Similar proteins

100% Identity | 90% Identity | 50% Identity

Protein	Similar proteins	Organisms	Length	Cluster ID	Cluster name	Size
P39678	N1P5U2	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	833	UniRef90_P39678	Cluster: Transcription factor MBP1	16
	UPI0000110973	Saccharomyces cerevisiae (strain CEN.PK113-7D) (Baker's yeast)				
	E9P8S5	Saccharomyces cerevisiae (Baker's yeast)				
	C7GT82	Saccharomyces cerevisiae (strain JAY291) (Baker's yeast)				
	A0A0L8VSP8	Saccharomyces sp. 'boulardii'				
	B5VFL1	Saccharomyces cerevisiae (strain AWRI1631) (Baker's yeast)				
	B3LGV4	Saccharomyces cerevisiae (strain RM11-1a) (Baker's yeast)				
	Q6Q7H9	Saccharomyces cerevisiae (strain Kyokai no. 7 / NBRC 101557) (Baker's yeast)				
	G2WCA4	Saccharomyces cerevisiae (strain YJM789) (Baker's yeast)				
	Q6Q7H4	Saccharomyces cerevisiae (strain Lalvin EC1118 / Prise de mousse) (Baker's yeast)				
	+5	And more				

Large numbers of redundant sequences can obscure results without adding useful information. Since many sequencing projects are active for a variety of reasons, many sequences from model organisms have been deposited into the databases multiple times. Redundancy is currently a major problem in sequence database searches. RefSeq attempts to hold only one sequence for all identical sequences in the database, and provide a high standard of annotation. Therefore there are more GenBank sequences, but if a RefSeq sequence exists, it is the authoritative, most highly annotated one of the set.

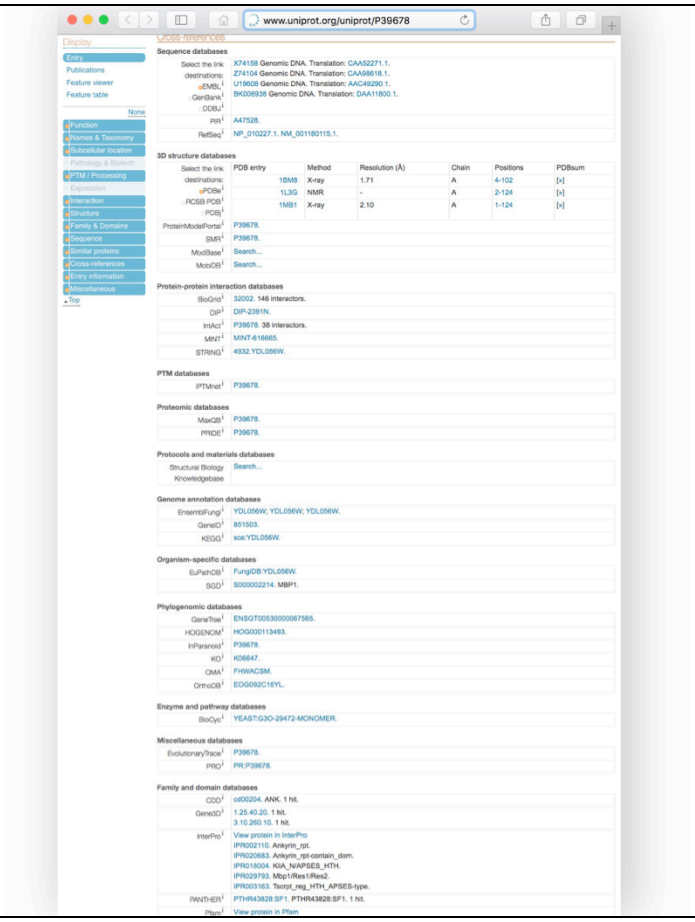
UniRef clusters sequences at three different levels: 100% identity for fully redundant sequences, 90% identity for sequences with trivial changes, e.g. accross strains; 50% similarity for sequence families.

The integration challenge is the single largest bottleneck in Bioinformatics !

Most bioinformatics data and procedures have been available through Web interfaces only. Interfaces have been poorly defined. Database architectures are not compatible. Data models suffer from legacy problems. Semantics differ ...

With all the stored and available sequence and annotation data, the challenges to cross-reference information have become very apparent.

Simple integration: cross-references



A simple approximation to the integration challenge is to provide cross-references. A cross-reference indicates that some information exists at the target database, but the schemas are not actually joined. For example the cross reference on the P39678 UniProt (yeast Mbp1) entry to RefSeq NP010227 retrieves the identical sequence, but the cross reference to PDB ID 1BM8 only concerns part of the Mbp1 sequence, the APSES DNA-binding domain. Thus a cross reference can't guarantee that it is or remains valid for the exact molecule that it annotates, and it is up to the user to take non-identical sequence numbers, sequence variants, post-translational modifications, partial coverage etc. etc. into account.

Data Integration

Data Integration across schemas:

- Federated databases – distinct databases, distributed query, merged result
- Data warehouses – all data in one database *(falling out of favour)*

Semantic integration:

- Ontologies

Data integration that is based on shared data model schemas is often done via federated databases. Only the cross-referencing tables are replicated across all databases, the tables that store the actual information are held in distinct databases. Queries are distributed across the different databases and the results are merged.

Integrating legacy databases in this way is often not possible, because the keys may describe mutually incompatible perspectives on the same entity.

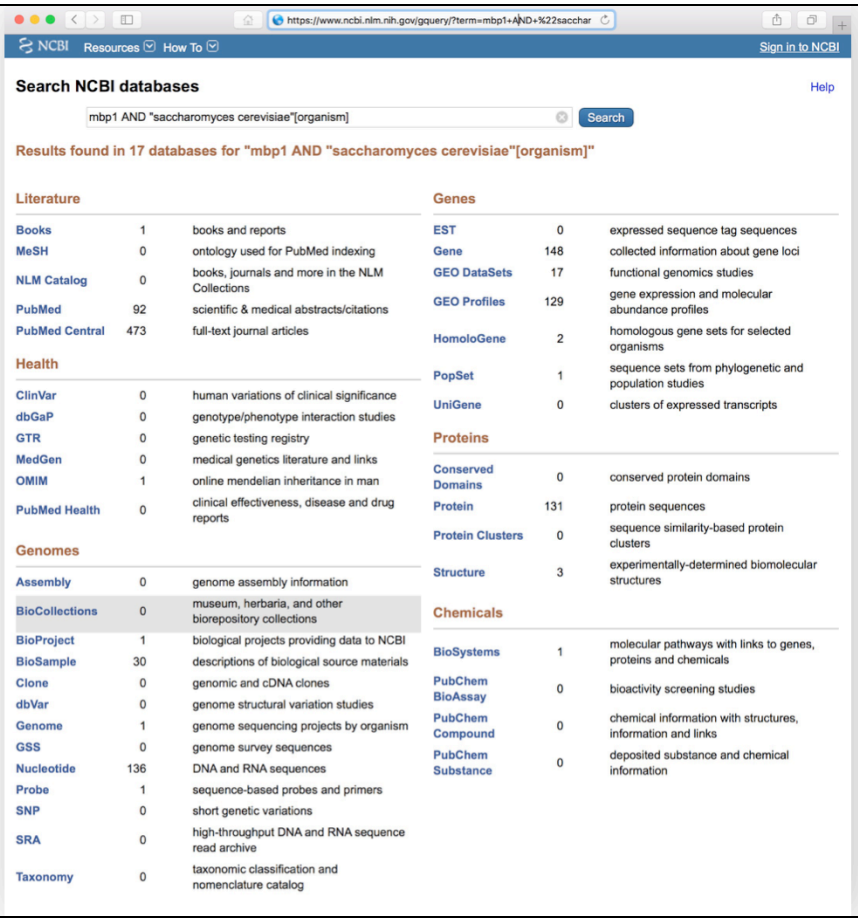
This problem can sometimes be overcome with semantic integration, i.e. focussing on the **meaning** of an entity rather than on an abstract identifier.

Data Integration

The Entrez system is NCBI's integration solution; Entrez Global Query is its search and retrieval system.

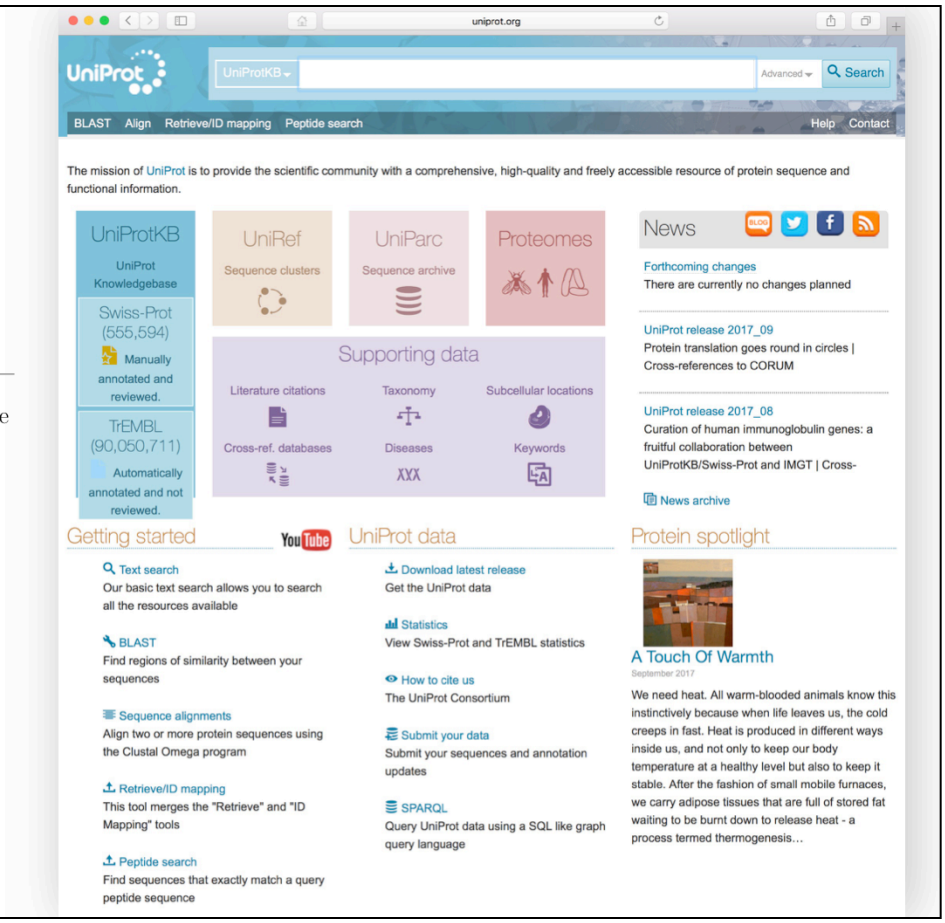
- Federated
- Programmable API

  (via E-utils)



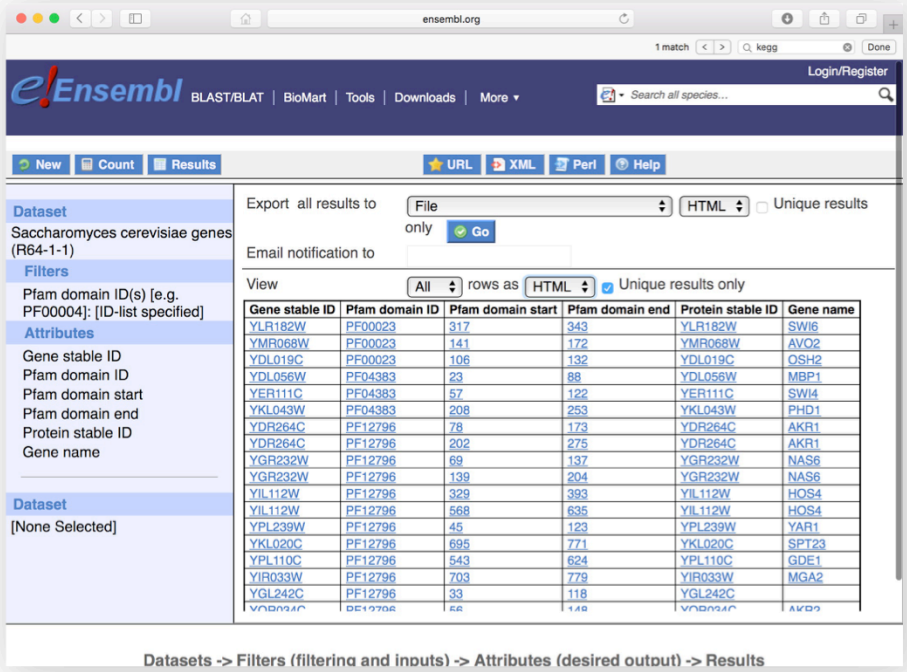
Data Integration

The EBI integrates data into UniProtKB – the UniProt Knowledege Base.



Data Integration

Ensembl is the EBIs **model organism genome data** integration project, and the BioMart data mining tool allows programmatic access and data download of its tables.



Sample search in the yeast genome returning GeneID, PfamDomain ID, start and end coordinates, and gene name of all genes with a KilA-N or ankyrin domain annotation.

A convenient interface to BioMart functions is provided by the Bioconductor biomaRt package.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA