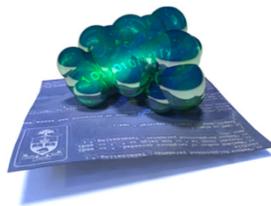


A
BIOINFORMATICS
COURSE

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 identifiers 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 (<i>S. cerevisiae</i> 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 (<i>Saccharomyces</i> 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”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	} “Hypothetical”, i.e. computationally derived from genome annotation
NM_123456	:	mRNA	
NP_123456	:	Protein	
XM_123456	:	mRNA	
XP_123456	:	Protein	

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

Similar proteins¹

		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)					
	AQA0LBVSP8	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. across 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.

DATA INTEGRATION

Simple integration:
cross-references

The screenshot shows the UniProt website interface for protein P39678 (Mbp1). The left sidebar contains a navigation menu with categories like 'Entry', 'Publications', 'Feature viewer', 'Feature table', 'Ontology & classification', 'PTM / Processing', 'Interaction', 'Structure', 'Family & Domains', 'Sequence', 'Similar proteins', 'Cross-references', 'Miscellaneous', and 'Top'. The main content area is titled 'Cross-references' and lists various databases with links to retrieve related information. The 'Sequence databases' section includes GenBank (X74158, Z74104, U75008, BK008538), RefSeq (A47325, NP_010227.5, NM_00180115.1), and 3D structure databases (PDB entry 1BM8, 1L3G, 1MB1). Other sections include Protein-protein interaction databases (BioGRID, DIP, Intact, MINT, STRING), PTM databases (PTMbase), Proteomic databases (MaxQB, PRIDE), Protocols and materials databases (Structural Biology, Knowledgebase), Genome annotation databases (EnsemblFung, GeneD, KEGG), Organism-specific databases (EUPanCP, SGD), Phylogenetic databases (GeneTree, HOGENOM, IPFtree, K0, OMA, OrthoDB), Enzyme and pathway databases (BioCyc), Miscellaneous databases (EvolutionaryTrace), and Family and domain databases (CD, Gene3D, InterPro, Pfam, PRINTS, ProDom).

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)

Search NCBI databases

mbp1 AND "Saccharomyces cerevisiae"[organism]

Results found in 17 databases for "mbp1 AND "Saccharomyces cerevisiae"[organism]"

Literature		Genes			
Books	1	books and reports	EST	0	expressed sequence tag sequences
MeSH	0	ontology used for PubMed indexing	Gene	148	collected information about gene loci
NLM Catalog	0	books, journals and more in the NLM Collections	GEO DataSets	17	functional genomics studies
PubMed	92	scientific & medical abstracts/citations	GEO Profiles	129	gene expression and molecular abundance profiles
PubMed Central	473	full-text journal articles	HomoloGene	2	homologous gene sets for selected organisms

Health		Proteins			
ClinVar	0	human variations of clinical significance	PopSet	1	sequence sets from phylogenetic and population studies
dbGaP	0	genotype/phenotype interaction studies	UniGene	0	clusters of expressed transcripts
GTR	0	genetic testing registry	Proteins		
MedGen	0	medical genetics literature and links	Conserved Domains	0	conserved protein domains
OMIM	1	online mendelian inheritance in man	Protein	131	protein sequences
PubMed Health	0	clinical effectiveness, disease and drug reports	Protein Clusters	0	sequence similarity-based protein clusters

Genomes		Chemicals			
Assembly	0	genome assembly information	Structure	3	experimentally-determined biomolecular structures
BioCollections	0	museum, herbaria, and other biorepository collections	Chemicals		
BioProject	1	biological projects providing data to NCBI	BioSystems	1	molecular pathways with links to genes, proteins and chemicals
BioSample	30	descriptions of biological source materials	PubChem BioAssay	0	bioactivity screening studies
Clone	0	genomic and cDNA clones	PubChem Compound	0	chemical information with structures, information and links
dbVar	0	genome structural variation studies	PubChem Substance	0	deposited substance and chemical information
Genome	1	genome sequencing projects by organism			
GSS	0	genome survey sequences			
Nucleotide	136	DNA and RNA sequences			
Probe	1	sequence-based probes and primers			
SNP	0	short genetic variations			
SRA	0	high-throughput DNA and RNA sequence read archive			
Taxonomy	0	taxonomic classification and nomenclature catalog			

DATA INTEGRATION

The EBI integrates data into **UniProtKB** – the UniProt Knowledge Base.

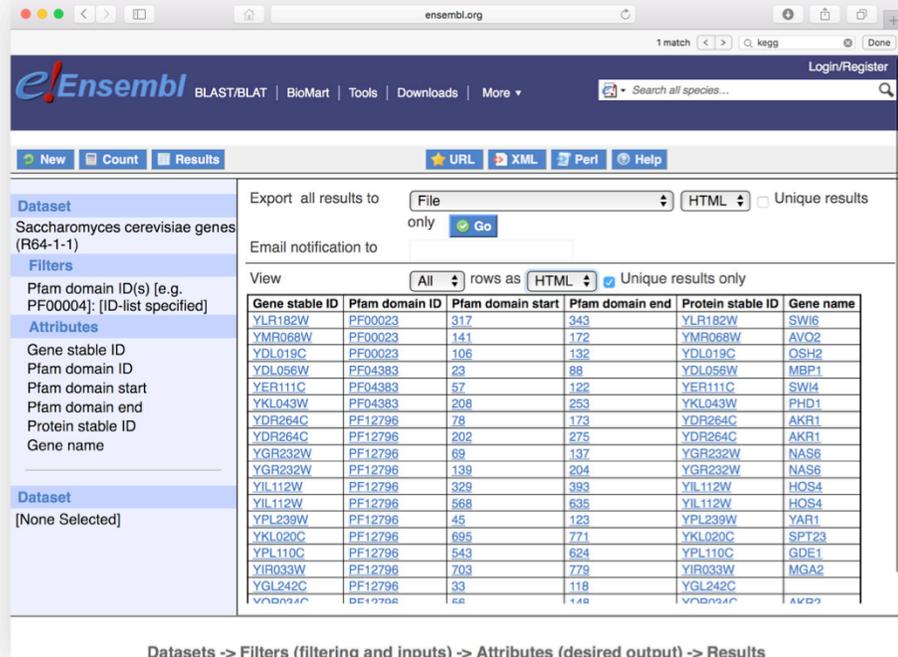
The screenshot displays the UniProt website homepage. At the top, there is a navigation bar with the UniProt logo, a search bar containing 'UniProtKB', and a search button. Below the navigation bar, a banner states the mission of UniProt: 'The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.'

The main content area is divided into several sections:

- UniProtKB Knowledgebase:** A vertical sidebar on the left lists 'Swiss-Prot (555,594) Manually annotated and reviewed.' and 'TrEMBL (90,050,711) Automatically annotated and not reviewed.'
- UniRef:** A box labeled 'Sequence clusters' with a circular icon.
- UniParc:** A box labeled 'Sequence archive' with a database icon.
- Proteomes:** A box with an icon of a person and a cell.
- Supporting data:** A central grid of icons for 'Literature citations', 'Cross-ref. databases', 'Taxonomy', 'Diseases', 'Subcellular locations', and 'Keywords'.
- News:** A section on the right with social media icons and news items like 'Forthcoming changes' and 'UniProt release 2017_09'.
- Getting started:** A section on the bottom left with links for 'Text search', 'BLAST', 'Sequence alignments', 'Retrieve/ID mapping', and 'Peptide search'.
- UniProt data:** A section on the bottom middle with links for 'Download latest release', 'Statistics', 'How to cite us', 'Submit your data', and 'SPARQL'.
- Protein spotlight:** A section on the bottom right featuring an image and the article 'A Touch Of Warmth'.

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.



The screenshot shows the Ensembl BioMart interface. The search criteria are: Dataset: Saccharomyces cerevisiae genes (R64-1-1); Filters: Pfam domain ID(s) [e.g. PF00004]; [ID-list specified]; Attributes: Gene stable ID, Pfam domain ID, Pfam domain start, Pfam domain end, Protein stable ID, Gene name. The results table is as follows:

Gene stable ID	Pfam domain ID	Pfam domain start	Pfam domain end	Protein stable ID	Gene name
YLR182W	PF00023	317	343	YLR182W	SWI6
YMR068W	PF00023	141	172	YMR068W	AVO2
YDL019C	PF00023	106	132	YDL019C	OSH2
YDL056W	PF04383	23	88	YDL056W	MBP1
YER111C	PF04383	57	122	YER111C	SWI4
YKL043W	PF04383	208	253	YKL043W	PHO1
YDR264C	PF12796	78	173	YDR264C	AKR1
YDR264C	PF12796	202	275	YDR264C	AKR1
YGR232W	PF12796	69	137	YGR232W	NAS6
YGR232W	PF12796	139	204	YGR232W	NAS6
YIL112W	PF12796	329	393	YIL112W	HOS4
YIL112W	PF12796	568	635	YIL112W	HOS4
YPL239W	PF12796	45	123	YPL239W	YAR1
YKL020C	PF12796	695	771	YKL020C	SPT23
YPL110C	PF12796	543	624	YPL110C	GDE1
YIR033W	PF12796	703	779	YIR033W	MGA2
YGL242C	PF12796	33	118	YGL242C	
YDR034C	PF12796	56	148	YDR034C	AKR2

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