Database 1

Biological databases and ontology

Cheong Xin Chan (CX)

c.chan1@uq.edu.au

Australian Centre for Ecogenomics
School of Chemistry & Molecular Biosciences
The University of Queensland

SCIE2100 | BINF6000 | Bioinformatics 1 – Introduction

Outline

- **Biological data and data types**
- **Biological databases**
  - Main functions and applications in bioinformatics
  - Characteristics and types of biological databases
  - Issues and challenges
- **Organisation of biological data and Ontology**
  - Definition and rationales
  - Basic terminology and types of ontologies
  - Linking data in Semantic Web
- **Biological ontologies**
  - Development and resources of biological ontologies
  - Gene Ontology and other examples
  - Issues and challenges

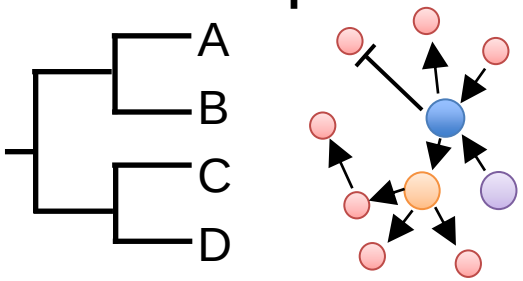
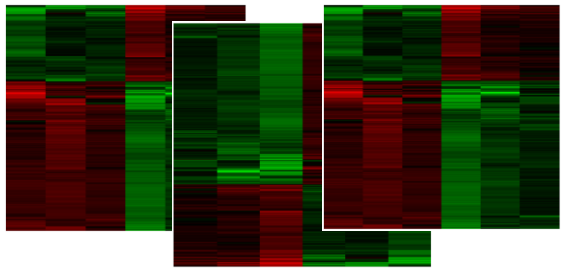
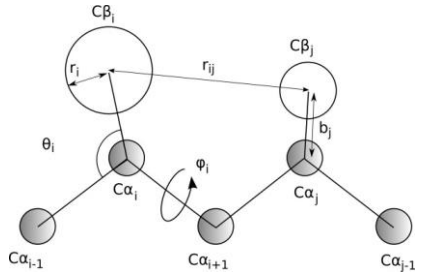
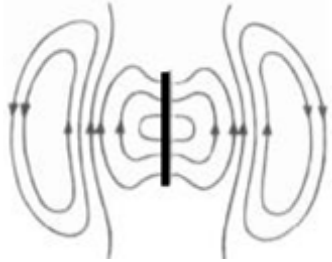
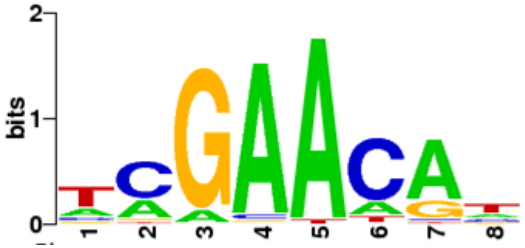
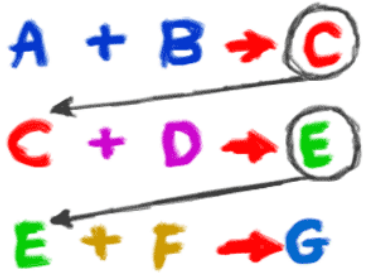
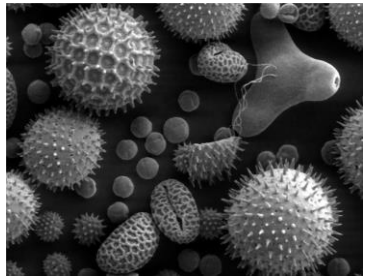
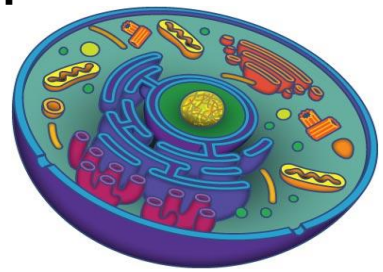
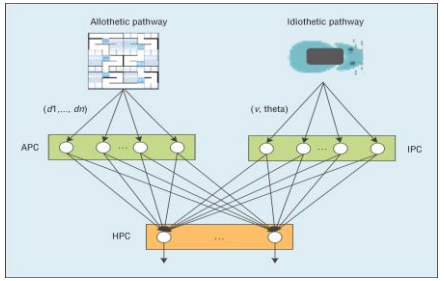
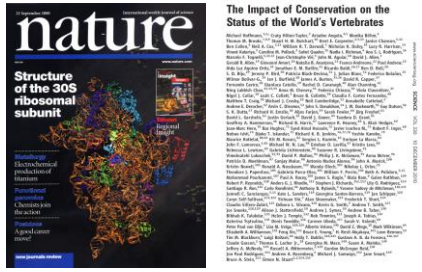
Biological data

include:

(a) information or measurements generated from biological resources and/or experiments (i.e. the **data**), and

(b) information that describes and/or related to characteristics/features of a biological entity/an experiment (i.e. the **metadata**, or the *data* about the data)

Biological data types

Sequences CGACTACGATCAGCTA CGACGACTACGATCAG CATGCATCACAGCTAG CATCGACTAGCATCGA TCACGACTAGCGCATG	**Graphs** 	**High-dimensional data** 	**Geometric information** 
Scalar & vector fields 	**Patterns** 	**Constraints** 	**Images** 
Spatial information 	**Models** 	**Prose/Literature** 	**Declarative knowledge (hypotheses, evidence)**

Data on a biological entity can be associated with one or more of these types: e.g., a protein might have associated with it two-dimensional images, three-dimensional structures, one-dimensional sequences, annotations of these data structures, etc.

Biological data in the digital world

- Biological data are **heterogeneous**
- Ideally these data should be **shareable** and **interoperable** among diverse laboratories and computer systems
- Most data are now in digital, machine-readable forms
- Common digital formats of biological data (non-exhaustive): en.wikipedia.org/wiki/Biological_data

Example 1: sequences

Sequence data

Multiple sequence alignment

```
>sequence1
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPY
IGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIP
FHPYYTIKDFLGLLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRS
VPNKLGGVLALFLSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPY
TIIGQMASILYFSIILAFLPIAGXIENY
```

FASTA format

FASTA format

```
>Seq5541
LNPSAPTVVVLGWLGASIKHLAKY
>Seq7032
QKPKRPTAVLLGWFAAKHKNLSKY
>Seq7152
MRPARGFVVVLGWFGAQDKHLKKY
>Seq2877
LVPWAPTAVLLGWVGCQMRYLRKY
>Seq0056
----RPLVLTLGWLGANERHLGKY
>Seq0781
GFLLNPLVIVMGWHGCKPRYLSKY
>Seq2239
QDPASVIVVLLGWYACHPKVLAKY
>Seq5612
KFPKVPIVMLLGWAGCQDRYLMKY
>Seq4904
VFSEEPVVILLGWAGSRDKHLAKY
>Seq5924
EIPDLPLVILLGWGGCSDKNLAKY
>Seq7619
EIPDQPVVILLGWGGCRDKNLAKY
```

PHYLIP format

```
 11 24
Seq5541 LNPSAPTVVVLGWLGASIKHLAKY
Seq7032 QKPKRPTAVLLGWFAAKHKNLSKY
Seq7152 MRPARGFVVVLGWFGAQDKHLKKY
Seq2877 LVPWAPTAVLLGWVGCQMRYLRKY
Seq0056 ----RPLVLTLGWLGANERHLGKY
Seq0781 GFLLNPLVIVMGWHGCKPRYLSKY
Seq2239 QDPASVIVVLLGWYACHPKVLAKY
Seq5612 KFPKVPIVMLLGWAGCQDRYLMKY
Seq4904 VFSEEPVVILLGWAGSRDKHLAKY
Seq5924 EIPDLPLVILLGWGGCSDKNLAKY
Seq7619 EIPDQPVVILLGWGGCRDKNLAKY
```

```
#NEXUS
[MySeqData.phy -- data title]

[Name: Seq5541          Len:    24  Check: 3190F9FF]
[Name: Seq7032          Len:    24  Check: 5463206F]
[Name: Seq7152          Len:    24  Check: 1CC258CA]
[Name: Seq2877          Len:    24  Check: D21B9C87]
[Name: Seq0056          Len:    24  Check: CF841852]
[Name: Seq0781          Len:    24  Check:  A625362]
[Name: Seq2239          Len:    24  Check: 9F481732]
[Name: Seq5612          Len:    24  Check: AC34C3CF]
[Name: Seq4904          Len:    24  Check: 817ABE64]
[Name: Seq5924          Len:    24  Check: C2CBAE73]
[Name: Seq7619          Len:    24  Check: C2CBAE73]


begin data;
 dimensions ntax=11 nchar=24;
 format datatype=protein interleave missing=-;
  matrix
  Seq5541  LNPSAPTVVVLGWLGASIKH LAKY
  Seq7032  QKPKRPTAVLLGWFAAKHKN LSKY
  Seq7152  MRPARGFVVVLGWFGAQDKH LKKY
  Seq2877  LVPWAPTAVLLGWVGCQMRY LRKY
  Seq0056  ----RPLVLTLGWLGANERH LGKY
  Seq0781  GFLLNPLVIVMGWHGCKPRY LSKY
  Seq2239  QDPASVIVVLLGWYACHPKV LAKY
  Seq5612  KFPKVPIVMLLGWAGCQDRY LMKY
  Seq4904  VFSEEPVVILLGWAGSRDKH LAKY
  Seq5924  EIPDLPLVILLGWGGCSDKN LAKY
  Seq7619  EIPDQPVVILLGWGGCRDKN LAKY

;
  end;
```

NEXUS format

Example 2: sequence records

```
LOCUS       HG941718             5109767 bp    DNA     circular BCT 03-APR-2015
DEFINITION  Escherichia coli ST131 strain EC958 chromosome, complete genome.
ACCESSION   HG941718
VERSION     HG941718.1  GI:641682562
DBLINK      BioProject: PRJEA61443
            BioSample: SAMEA2272019
KEYWORDS    complete genome.
SOURCE      Escherichia coli O25b:H4-ST131
  ORGANISM  Escherichia coli O25b:H4-ST131
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1
  AUTHORS   Forde,B.M., Ben Zakour,N.L., Stanton-Cook,M., Phan,M.D.,
            Totsika,M., Peters,K.M., Chan,K.G., Schembri,M.A., Upton,M. and
            Beatson,S.A.
  TITLE     The complete genome sequence of Escherichia coli EC958: a high
            quality reference sequence for the globally disseminated multidrug
            resistant E. coli O25b:H4-ST131 clone
  JOURNAL   PLoS ONE 9 (8), E104400 (2014)
   PUBMED   25126841
   REMARK   Publication Status: Online-Only
REFERENCE   2  (bases 1 to 5109767)
  AUTHORS   Beatson,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-OCT-2011) The University of Queensland, Centre for
            Infectious Disease Research, St. Lucia, Brisbane, QLD 4072,
            AUSTRALIA
FEATURES             Location/Qualifiers
     source          1..5109767
                     /organism="Escherichia coli O25b:H4-ST131"
                     /mol_type="genomic DNA"
                     /strain="EC958"
                     /serotype="O25b:H4"
                     /db_xref="taxon:941322"
...
```

GenBank format (NCBI USA)

```
ID   HG941718; SV 1; circular; genomic DNA; STD; PRO; 5109767 BP.
AC   HG941718;
PR   Project:PRJEA61443;
DT   11-MAR-2014 (Rel. 120, Created)
DT   03-APR-2015 (Rel. 124, Last updated, Version 6)
DE   Escherichia coli ST131 strain EC958 chromosome, complete genome
KW   complete genome.
OS   Escherichia coli ST131
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
OC   Enterobacteriaceae; Escherichia.
RN   [1]
RP   1-5109767
RA   Beatson S.;
RT   ;
RL   Submitted (05-OCT-2011) to the INSDC.
RL   The University of Queensland, Centre for Infectious Disease Research, St.
RL   Lucia, Brisbane, QLD 4072, AUSTRALIA.
RN   [2]
RX   DOI; 10.1371/journal.pone.0104400.
RX   PUBMED; 25126841.
RA   Forde B.M., Ben Zakour N.L., Stanton-Cook M., Phan M.D., Totsika M.,
RA   Peters K.M., Chan K.G., Schembri M.A., Upton M., Beatson S.A.;
RT   "The complete genome sequence of Escherichia coli EC958: a high quality
RT   reference sequence for the globally disseminated multidrug resistant E.
RT   coli O25b:H4-ST131 clone";
RL   PLoS One 9(8):e104400-e104400(2014).
...
FH   Key             Location/Qualifiers
FH
FT   source          1..5109767
FT                   /organism="Escherichia coli ST131"
FT                   /strain="EC958"
FT                   /serotype="O25b:H4"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:1359206"
...
```

EMBL format (EBI Europe)

- **Standardised definitions** within a format – these can be format-specific
- **Cross-references** allow for integration of information from different databases

Example 3: protein structures

```
HEADER    TRANSFERASE                                 29-JUL-07   2Z6C
TITLE     CRYSTAL STRUCTURE OF LOV1 DOMAIN OF PHOTOTROPIN1 FROM
TITLE    2 ARABIDOPSIS THALIANA
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: PHOTOTROPIN-1;
COMPND   3 CHAIN: A, B;
COMPND   4 FRAGMENT: UNP RESIDUES 180-308, LOV1 DOMAIN;
COMPND   5 SYNONYM: NON-PHOTOTROPIC HYPOCOTYL PROTEIN 1, ROOT
COMPND   6 PHOTOTROPISM PROTEIN 1;
COMPND   7 EC: 2.7.11.1;
COMPND   8 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: ARABIDOPSIS THALIANA;
SOURCE   3 ORGANISM_COMMON: MOUSE-EAR CRESS;
SOURCE   4 ORGANISM_TAXID: 3702;
SOURCE   5 GENE: PHOT1, JK224, NPH1, RPT1;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 562
...
ATOM      1  N   VAL A 184      63.131  56.497  -7.951  1.00 66.06           N
ATOM      2  CA  VAL A 184      63.000  57.402  -6.758  1.00 66.09           C
ATOM      3  C   VAL A 184      61.746  58.281  -6.840  1.00 65.59           C
ATOM      4  O   VAL A 184      60.942  58.285  -5.910  1.00 65.83           O
ATOM      5  CB  VAL A 184      64.308  58.238  -6.490  1.00 66.38           C
ATOM      6  CG1 VAL A 184      64.019  59.583  -5.767  1.00 67.05           C
ATOM      7  CG2 VAL A 184      65.335  57.397  -5.699  1.00 67.19           C
ATOM      8  N   SER A 185      61.574  59.014  -7.941  1.00 64.93           N
ATOM      9  CA  SER A 185      60.374  59.839  -8.108  1.00 64.24           C
ATOM     10  C   SER A 185      59.090  59.005  -8.233  1.00 63.86           C
ATOM     11  O   SER A 185      58.967  58.134  -9.107  1.00 63.28           O
ATOM     12  CB  SER A 185      60.513  60.821  -9.276  1.00 64.53           C
...
```

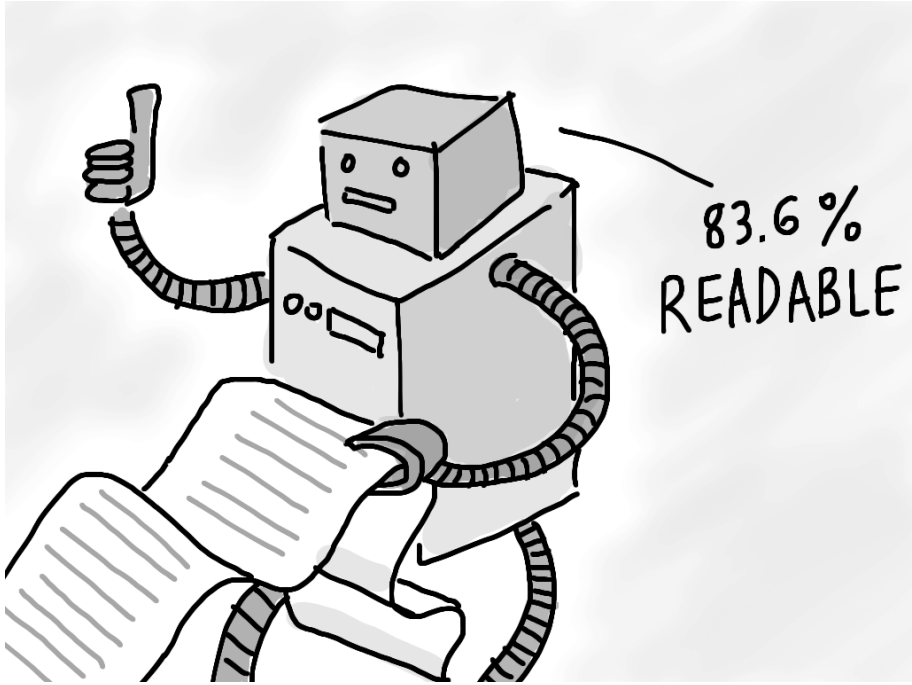
PDB format – the standard representation for macromolecular structure data

Biological databases

Two main functions:

Emilio Quintana (https://www.flickr.com/photos/eq/)

Make biological data available to scientists



UXMastery.com

Make biological data available in computer-readable form

Applications of databases in bioinformatics (and modern molecular biosciences)

Identification of biological entities

- genes, proteins, metabolites, reactions etc.
- homologs, orthologs, gene/protein families

Inference of function

- sequence/structural similarity, prediction or association

Hypothesis generation

- Source data for experiments/to train and test tools, models and methods

Characteristics of biological databases

Type of data	Data entry & quality control	Primary or derived data
• Nucleotide • Protein • Macromolecular structure • Gene expression • Metabolic pathways • Protein-protein interactions • …	• Scientists (teams) deposit data directly • Appointed curators add and update data • Are erroneous data removed or marked? • Type and degree of error checking • Consistency, redundancy, conflicts, updates	• Primary databases: experimental results directly into database • Secondary databases: results of analysis of primary databases • Aggregate of many databases • Links to other data items • Combination of data • Consolidation of data
Technical design	**Maintainer status**	**Availability**
• Flat-files • Relational database (MySQL) • Object-oriented database (e.g. XML, JSON formats, PostgreSQL)	• Large, public institution (e.g. EMBL, NCBI) • Quasi-academic institute (e.g. Swiss Institute of Bioinformatics, J. Craig Venter Institute) • Academic group or scientist • Commercial company	• Publicly available, no restrictions • Available, but with copyright • Accessible, but not downloadable • Academic, but not freely available • Proprietary, commercial; possibly free for academics

One or more of these factors affect which database resources does one use, and how.

Types of biological databases:
Sequence databases

What can you get?

- Sequences, from whole genomes to protein isoforms
- Biomolecular and chemical structures
- Functional annotation
- Analysis tools

What can you do with them?

- Homology inference
- Phylogenetic analysis
- Sequence-based feature prediction, e.g. conserved patterns (motifs)
- Functional analysis

International Nucleotide Sequence Database Collaboration (INSDC)

- GenBank www.ncbi.nlm.nih.gov
- EMBL-EBI www.ebi.ac.uk
- DNA DataBank of Japan www.ddbj.nig.ac.jp

Mandatory sequence submission to public repositories (making data freely available) prior to scientific publications

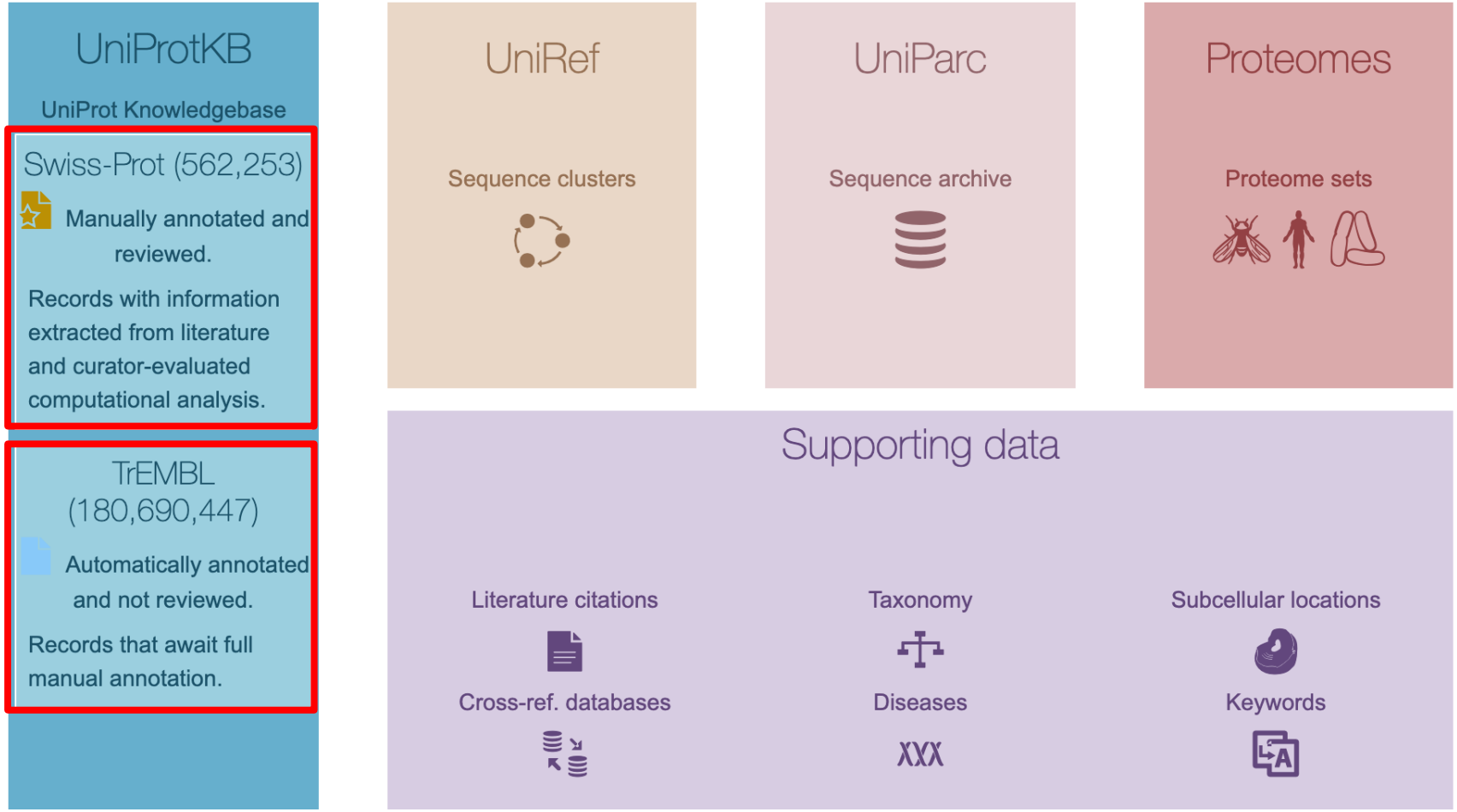
Growth of sequence databases

1960s	• First Atlas of Protein Sequence and Structure; Vol. 1 (1965) • Vol. 4 (1969): > 300 protein sequences; 16 DNA sequences
1970s	• Recognition of the role of databases in collecting & managing sequence data • Efforts to make sequence data available to biologists
1980s	• Sequence databases established by NIH and EMBL • DDBJ/EMBL/GenBank (1982); **606** nucleotide seqs (Dec 1982)
1990s	• Internet • Increase in sequence submission to public repositories
2000s	• Continued growth of sequence data including whole genomes
2010s+	• Challenges posed by high-throughput sequencing • **1.79** billion nucleotide sequences in GenBank including WGS (Feb 2021)

Types of biological databases:

Annotation databases

UniProt Knowledgebase (UniProtKB):
www.uniprot.org



Cross-referencing to other databases

- Identifiers, alternative names, accession numbers

Annotation

- Description of function
- Isoforms and sequences
- Localisation
- Post translational modifications
- Domains, motifs, signal sequences
- Interactions

Useful for

- Collecting sets of functionally related proteins (e.g. families)
- Identifying protein features

Annotation databases

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

Gene Ontology database (geneontology.org)

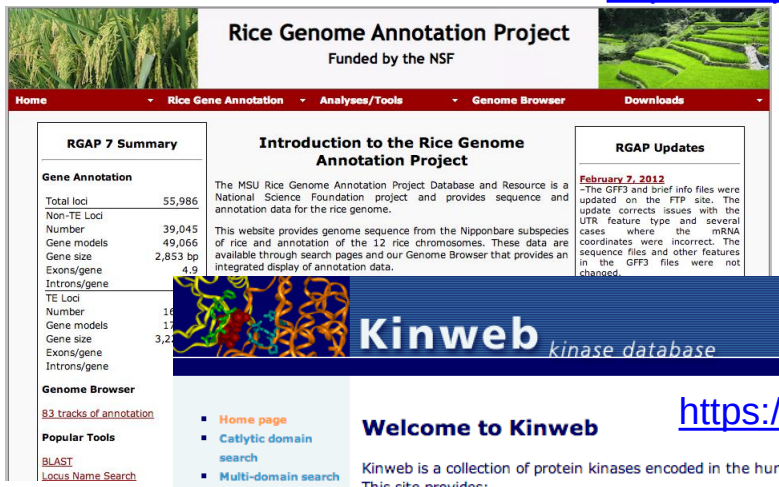
- Unified attributes of gene and gene product across all species
- Maintained by the GO Consortium

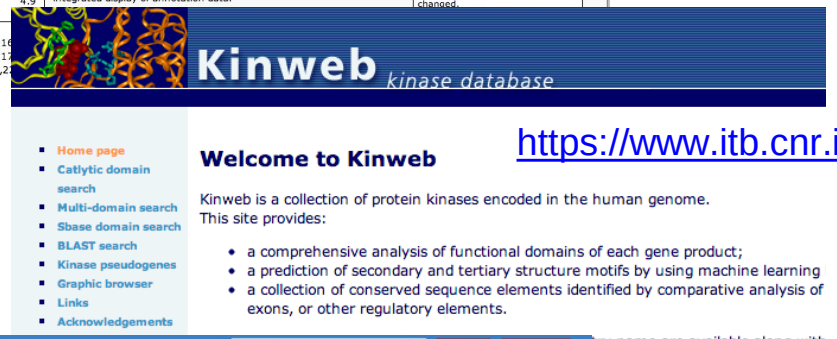
Types of biological databases:

Context-specific databases

- Established and maintained by the research communities
- Species-specific
- Disease-specific
- System-specific
  e.g. kinases (Kinweb), nuclear receptors (NURSA), allosteric molecules (ASD), membrane transporters (transportDB)

http://rice.plantbiology.msu.edu/





https://www.itb.cnr.it/kinweb/

http://mdl.shsmu.edu.cn/ASD/

Advantages

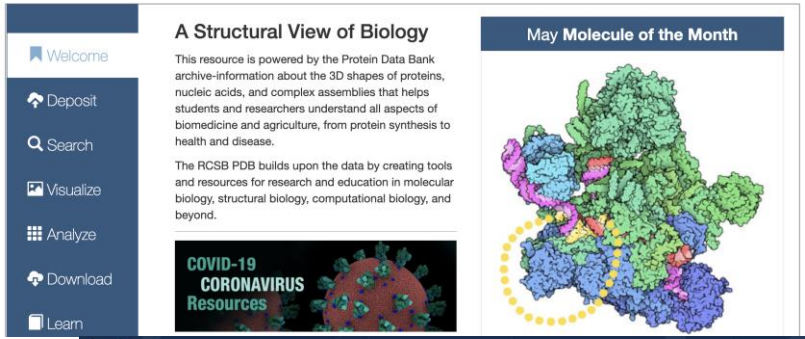
- Maintained by the research community
- Data curation by expert community
- Detailed, high-quality annotation

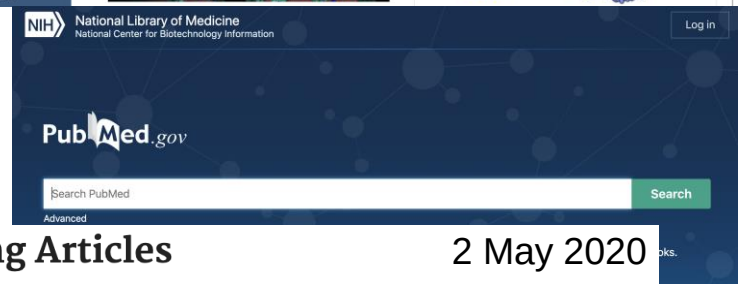
Disadvantages

- Maintained by the research community
- Idiosyncratic naming conventions
- Poor mapping to external DBs
- Sporadic updates

Other context-specific databases

- **InterPro:** www.ebi.ac.uk/interpro
  - domain models of proteins from ~20 primary resources
- **Protein Data Bank (PDB):** www.rcsb.org/pdb/
  - high-resolution 3D structural data of proteins
- **Gene Expression Omnibus (GEO):** www.ncbi.nlm.nih.gov/geo/
  ArrayExpress: www.ebi.ac.uk/arrayexpress/
  - high-throughput functional genomics (gene expression) data
- **Oncomine:** www.oncomine.org
  - cancer microarray data & web-based data-mining platform
- **SymbioGBR:** http://www.symbiogbr.org/
  - database of coral symbionts from the Great Barrier Reef
- **KEGG Pathway Database:** www.genome.jp/kegg/pathway.html
  - pathways of molecular interaction and reaction networks for metabolism, information processes, cellular processes, organismal systems & human diseases
- **PubMed (NCBI):** www.ncbi.nlm.nih.gov/pubmed/
  - scientific literature relevant to biomedical fields
- and many more …





Trending Articles 2 May 2020

PubMed records with recent increases in activity

COVID-19 diagnosis and management: a comprehensive review.
Pascarella G, et al. J Intern Med. 2020. PMID: 32348588

Hypercoagulation and Antithrombotic Treatment in Coronavirus 2019: A New Challenge.
Violi F, et al. Thromb Haemost. 2020. PMID: 32349133

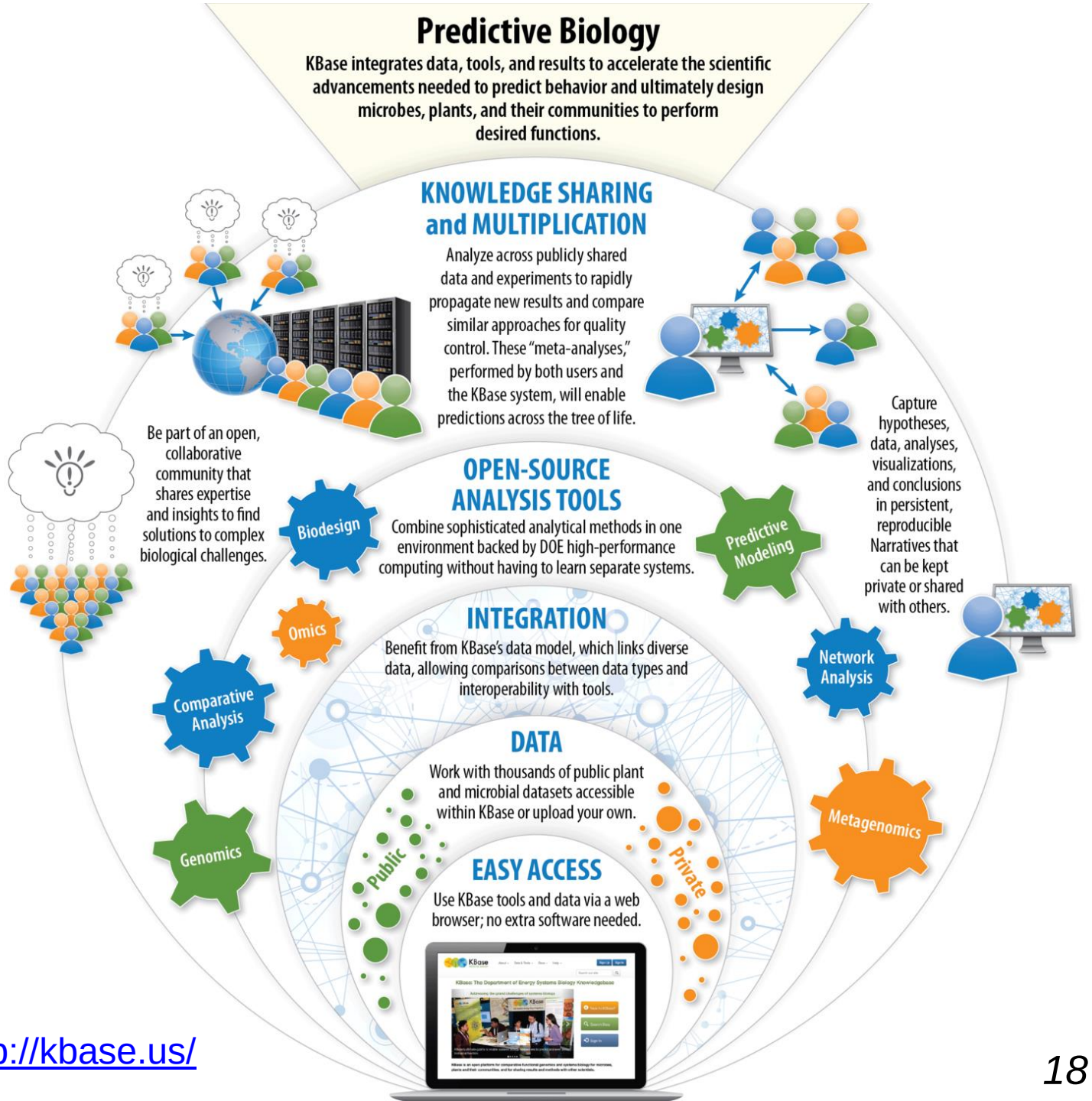
A SARS-CoV-2 protein interaction map reveals targets for drug repurposing.
Gordon DE, et al. Nature. 2020. PMID: 32353859

DIC in COVID-19: Implications for Prognosis and Treatment?
Seitz R, et al. J Thromb Haemost. 2020. PMID: 32344469

Finding the dose for hydroxychloroquine prophylaxis for COVID-19; the desperate search for effectiveness.
Al-Kofahi M, et al. Clin Pharmacol Ther. 2020. PMID: 32344449

Databases as a knowledgebase

- integrates and processes data (usually from multiple databases) and uses expert knowledge to give answers, recommendations and expert advice
- one-stop platform allows for multi-scale modelling, data integration, collaboration and sharing



A workflow example:
U.S. Department of Energy's Systems Biology Knowledgebase (KBase) http://kbase.us/

Biological databases: Issues and challenges

Data heterogeneity

- various data types & data quality impact the ease of curation, automation and integration

Proprietary issues

- private databases are not readily accessible to the research community

Disparate terminology

- standardisation of terms or formats through time

Shareability & interoperability

- interfaces for data exchange & data-format description, interfaces to recognise data-model intersections, to exchange metadata and to parse queries

Organisation of biological data & ontology

- Data are useless if they are inaccessible or incomprehensible to others
- Data organisation is central to extracting useful information from the data
- Aim: one large, distributed information resource with **common controlled vocabularies**, related **user interfaces**, and **practices**
- Emphasis: **shareability** and **interoperability** of large-scale **heterogeneous** data

Ontology

Greek origin: *ontos* – being or the nature of things; *logia* – science, study, theory

- defines (specifies) the **concepts**, **relationships**, and other distinctions that are relevant for modeling a domain
- takes the form of the definitions of **representational vocabulary** (classes, relations etc.), which provide meanings for the vocabulary and **formal constraints** on its coherent use
- commonly based on agreed-upon understanding of a domain, i.e. a **joint terminology** between members of a community of interest
- a vocabulary of concepts and relations rich enough to enable us to express knowledge and intention **without semantic ambiguity**

Why do we need ontology?

- to **share common understanding** of the structure of information among people or software agents
- to **enable reuse** of domain knowledge
- to make domain **assumptions explicit**

Natalya Noy & Deborah McGuinness, Stanford University:
http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

Basic terminology

Class

Person

Organisation

Relations

has employer,
is employee of
….

Person —has employer→ Organisation

Instance (Individual)

Triple

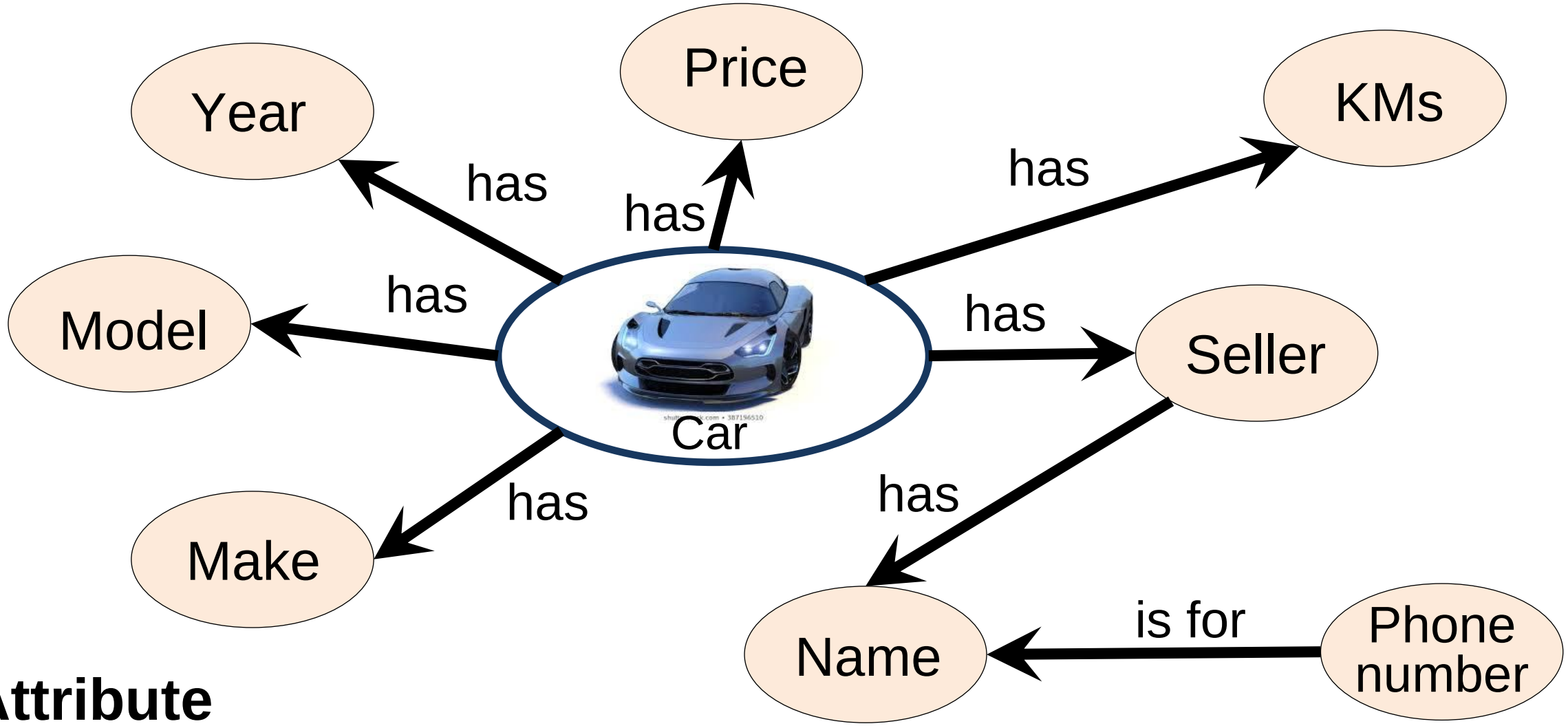
Subject
Andrew

—**Predicate** has employer→

Object
University of Queensland

https://www.youtube.com/watch?v=jfUPLuPL3Ho

Example: a car advertisement

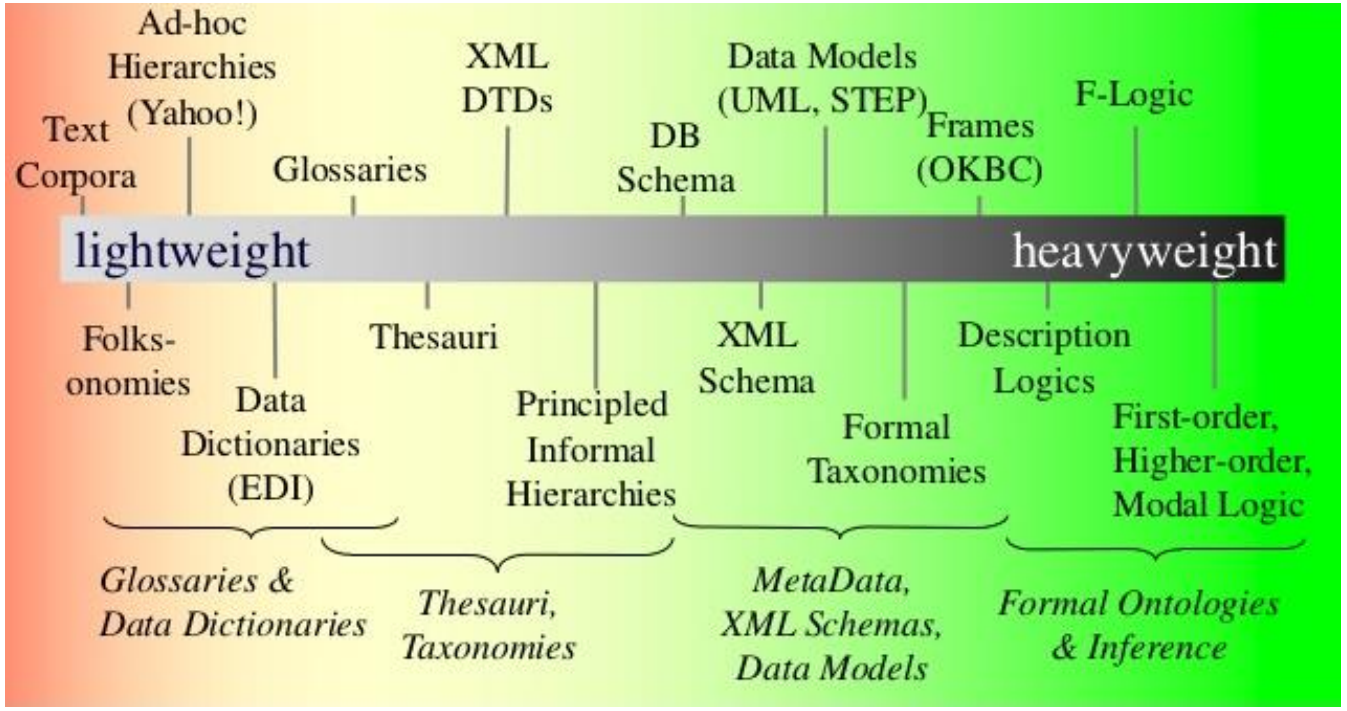


Attribute
(property, characteristics)

Types of ontology

Light weight ontologies

- glossaries, dictionaries
- thesauri
- taxonomies or conceptual hierarchies
- typically ***is-a*** relations



Fernando Silva Parreiras. *Model-Driven Software Development with Semantic Web Technologies.* http://www.slideshare.net/fparreiras/modeldriven-software-development-with-semantic-web-technologies

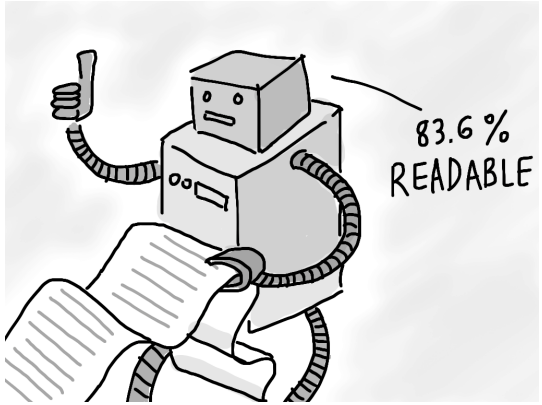
Heavy weight ontologies

- data models
- definition of concepts with axioms (i.e. established statements)
- logical formalisation and inference

Linking data in Semantic Web

- In Semantic Web, ontologies are **collections of statements** (written in e.g. Resource Description Framework *RDF*) that define the **relations between concepts** and specify **logical rules for reasoning** about them
- a common framework that allows data to be **shared and reused** across application, enterprise and community boundaries
- enables machines to “understand” and respond to complex human requests based on their meaning
- the relevant information sources need to be semantically structured

http://www.w3.org/standards/semanticweb/data

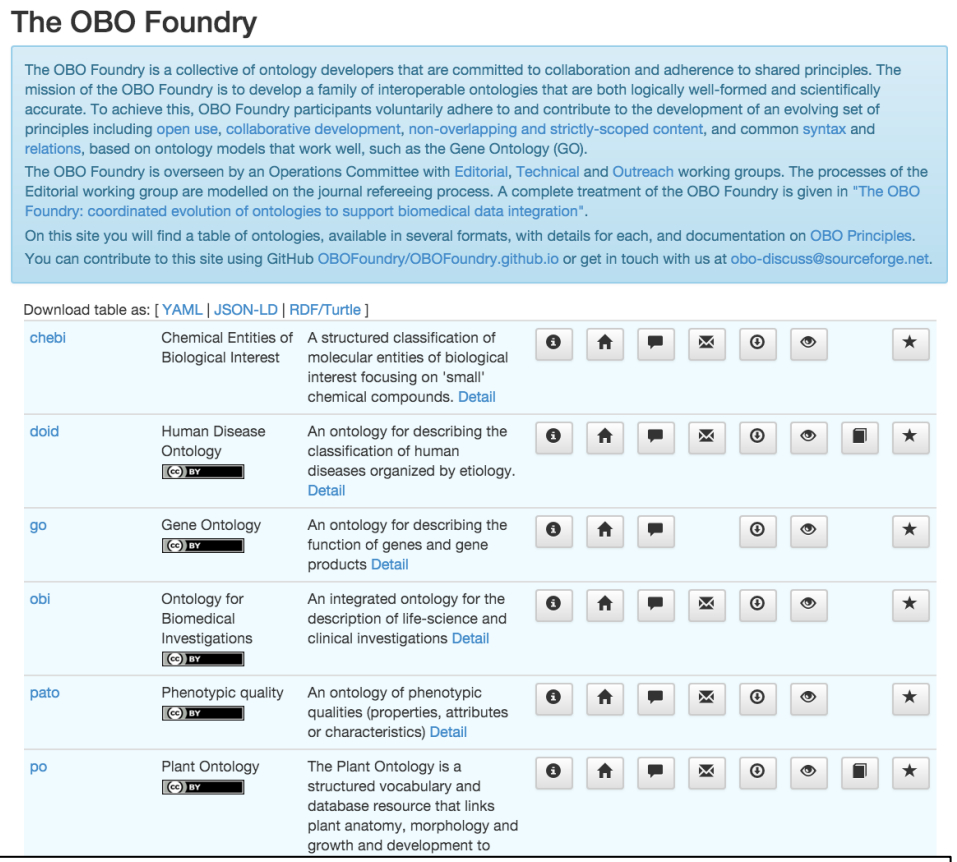


Biology is rapidly changing from a descriptive to a data-driven discipline in which the discovery of novel findings depends on the comparison and integration of massive data sets. As a consequence, **ontologies—systematic descriptions of specific biological attributes**—are becoming more and more important for describing the existing biological knowledge.

Jensen LJ & Bork P (2010) Ontologies in Quantitative Biology: A Basis for Comparison, Integration, and Discovery. *PLoS Biology* 8(5): e1000374.

Development of biological ontologies

- The OBO Foundry www.obofoundry.org
- Collective of ontology developers
- Adoption of a growing set of principles specifying best-practices in ontology development
- To foster **interoperability** of ontologies
- Projects include Human Disease Ontology, Gene Ontology etc.



The OBO Foundry

The OBO Foundry is a collective of ontology developers that are committed to collaboration and adherence to shared principles. The mission of the OBO Foundry is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate. To achieve this, OBO Foundry participants voluntarily adhere to and contribute to the development of an evolving set of principles including open use, collaborative development, non-overlapping and strictly-scoped content, and common syntax and relations, based on ontology models that work well, such as the Gene Ontology (GO).

The OBO Foundry is overseen by an Operations Committee with Editorial, Technical and Outreach working groups. The processes of the Editorial working group are modelled on the journal refereeing process. A complete treatment of the OBO Foundry is given in "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration".

On this site you will find a table of ontologies, available in several formats, with details for each, and documentation on OBO Principles.

You can contribute to this site using GitHub OBOFoundry/OBOFoundry.github.io or get in touch with us at obo-discuss@sourceforge.net.

Download table as: [YAML | JSON-LD | RDF/Turtle]

chebi	Chemical Entities of Biological Interest	A structured classification of molecular entities of biological interest focusing on 'small' chemical compounds. Detail
doid	Human Disease Ontology	An ontology for describing the classification of human diseases organized by etiology. Detail
go	Gene Ontology	An ontology for describing the function of genes and gene products Detail
obi	Ontology for Biomedical Investigations	An integrated ontology for the description of life-science and clinical investigations Detail
pato	Phenotypic quality	An ontology of phenotypic qualities (properties, attributes or characteristics) Detail
po	Plant Ontology	The Plant Ontology is a structured vocabulary and database resource that links plant anatomy, morphology and growth and development to

Anatomy Reference Ontology

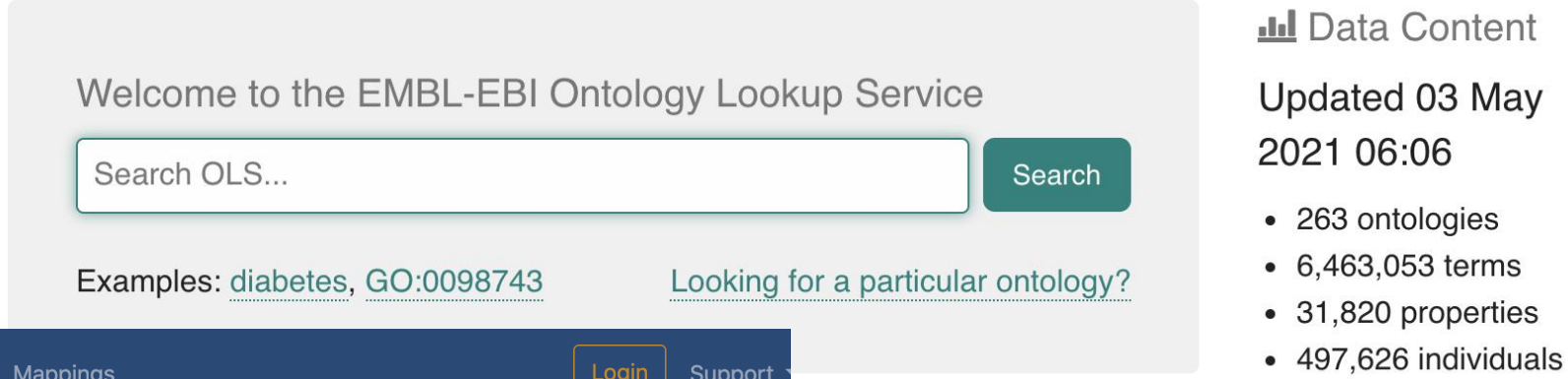
What does OBO stand for?
The Open Biological and Biomedical Ontologies Foundry (formerly known *as The Open Biomedical Foundry*, which was previously known as *The Open Biological Foundry*)

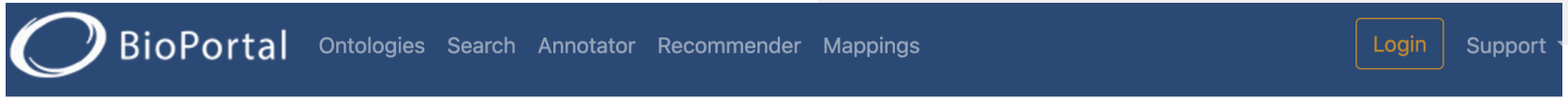
Smith B *et al.* (2007) *Nature Biotechnology* 25: 1251-1255.

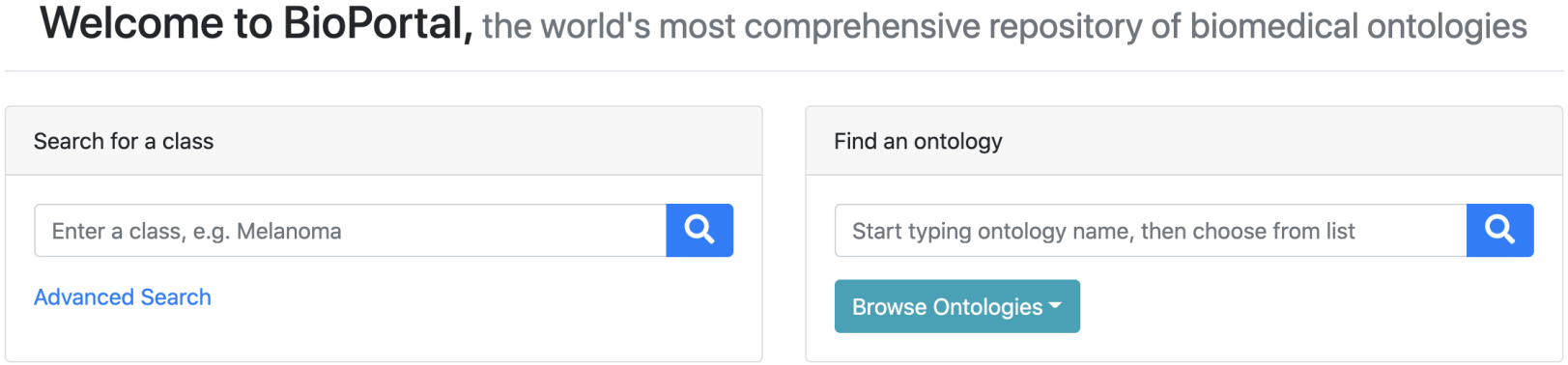
Resources of biological ontologies

EBI Ontology Lookup Service
https://www.ebi.ac.uk/ols
- Ontologies from OBO Foundry

NCBO BioPortal
bioportal.bioontology.org
- open repository; user can add notes, review & map





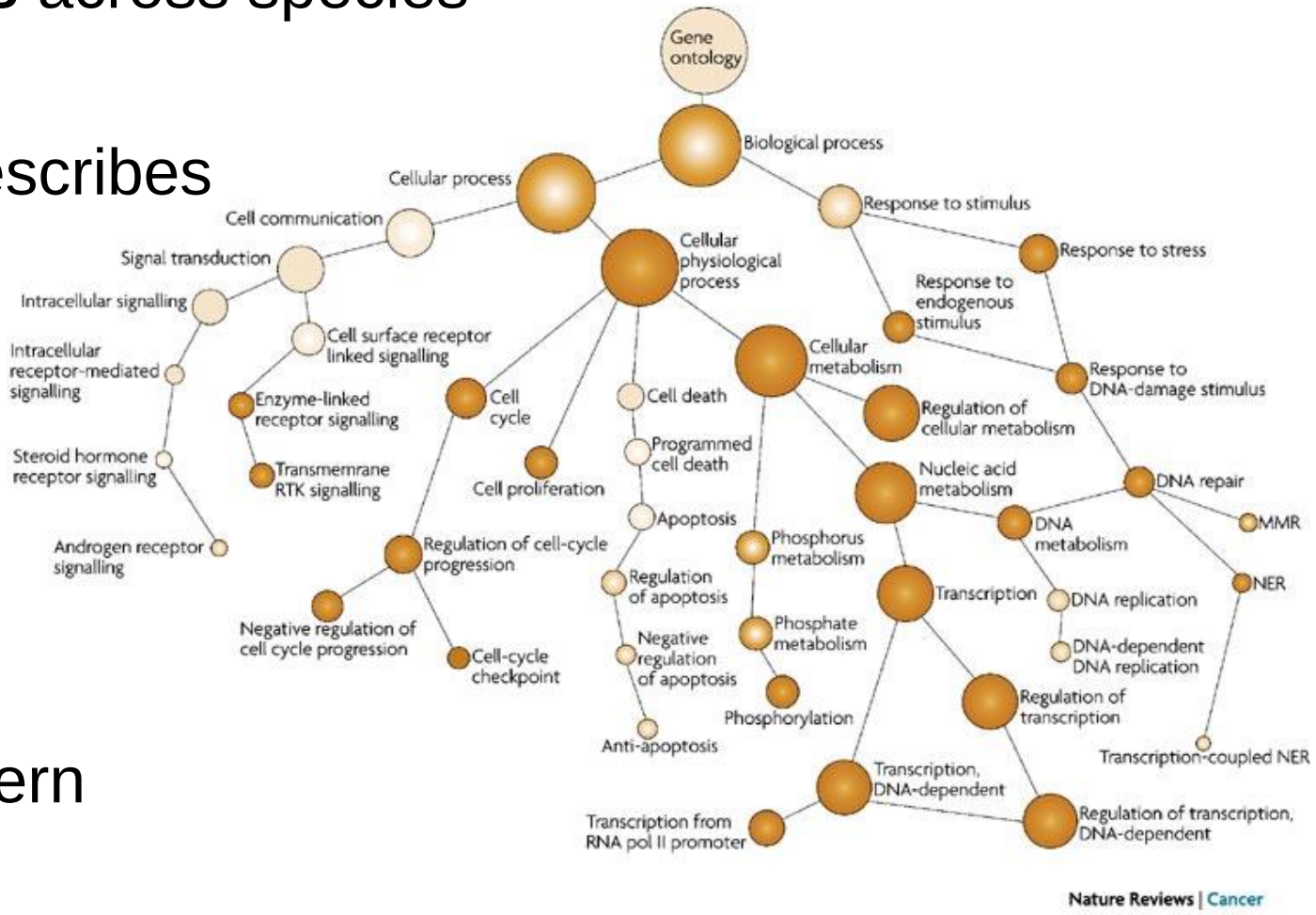


BioPortal Statistics	
Ontologies	868
Classes	9,914,067
Properties	36,286
Mappings	73,435,253

Gene Ontology

http://geneontology.org/

- Aims to standardise the representation of **gene and gene product attributes** across species and databases
- Controlled vocabulary that describes characteristics of gene products, and the associated annotation data
- Terms organised **resembling hierarchy**
- Widely used ontology in modern biological research



"The goal of the **Gene Ontology Consortium** is to produce **a dynamic, controlled vocabulary** that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is **accumulating and changing**."

Ashburner *et al.* (2000) *Nature Genetics*, 25: 25-29

Gene Ontology terms

http://geneontology.org/stats.html

Statistics for release 2021-02

Organised in **three** structured, **species-independent** ontologies that describe gene products based on their association to aspects of:

biological process	cellular component	molecular function

← Level 1 terms

Ontology

Property	Value
Valid terms	44085 (Δ = -6)
Obsoleted terms	3125 (Δ = 17)
Merged terms	2252 (Δ = 12)
Biological process terms	28748
Molecular function terms	11153
Cellular component terms	4184

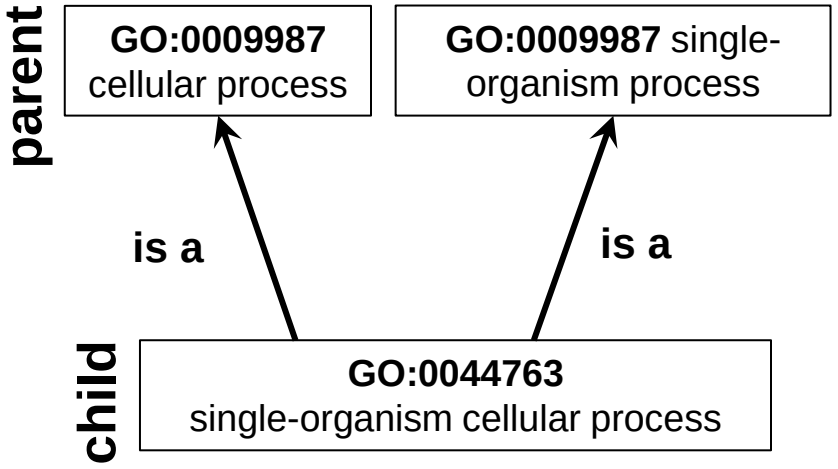
These ontologies resemble an **hierarchy**:

- **child terms** – more specialised; **parent terms** – less specialised
- a term may have **more than one** parent term (unlike an hierarchy)
- terms may be connected to parent terms via **different relations**

Example:

```
id: GO:0044763
name: single-organism cellular process
namespace: biological_process
def: "Any process that is carried out at the cellular
level, occurring within a single organism."
is_a: GO:0009987 ! cellular process
is_a: GO:0044699 ! single-organism process
```

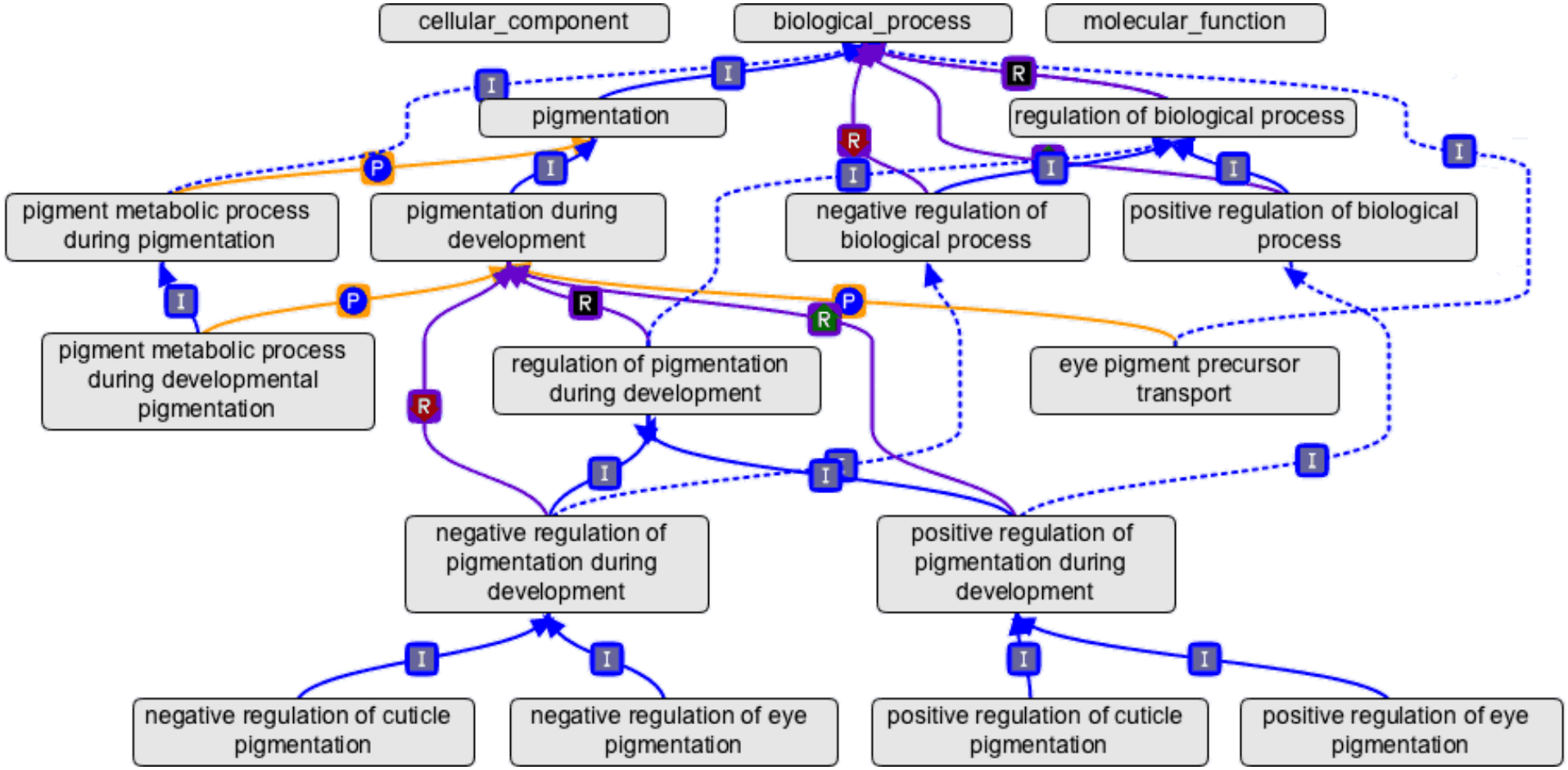
OBO format



Gene Ontology graph

Structure of GO can be described in terms of a graph:

- **node**: a GO term; **edge** (arc): the relationship between the two terms (nodes)

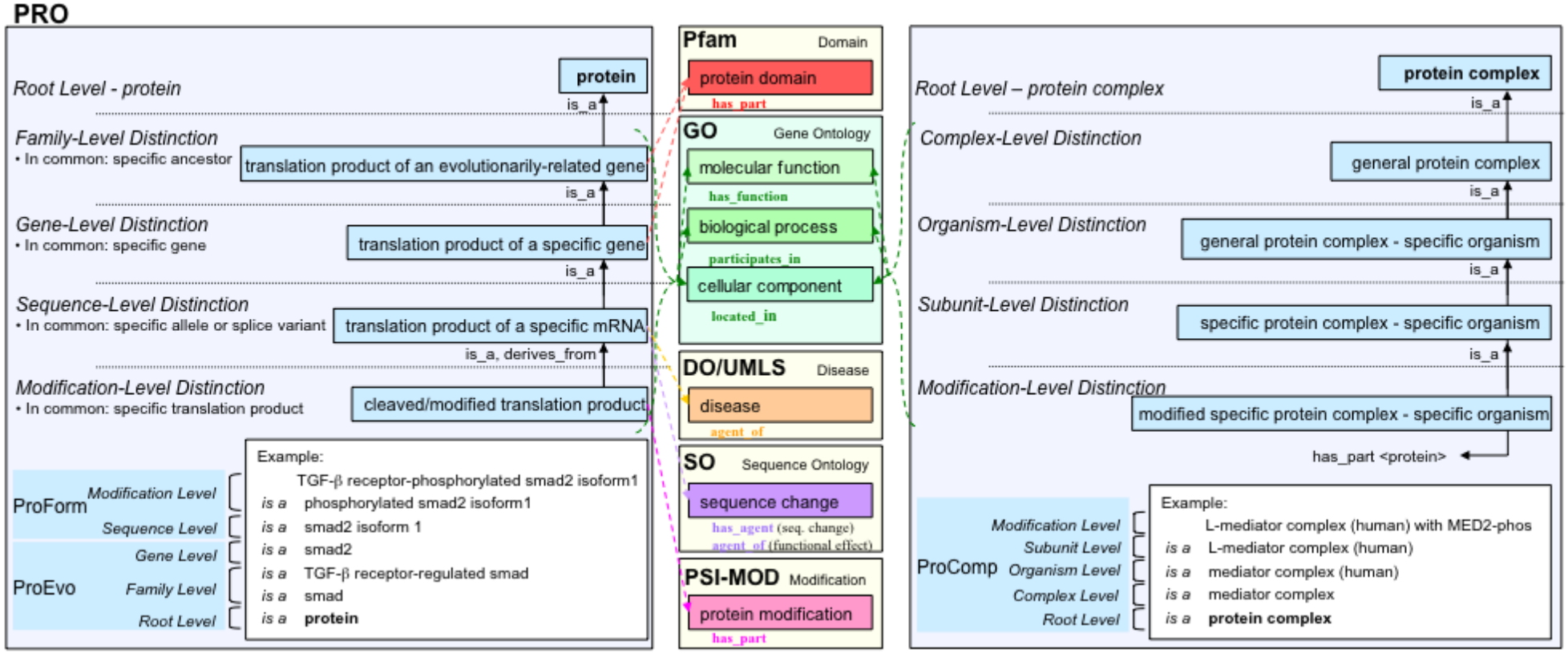


Relations **I: is a** **P: part of**
R: regulates (one process directly affects the manifestation of another process or quality)

Protein Ontology

PRO pir.georgetown.edu/pro/

- ontological representation of protein-related entities
- three sub-ontologies:
  - ProEvo (based on evolutionary relatedness)
  - ProForm (protein forms produced from a given gene locus); and
  - ProComp (protein-containing complexes)

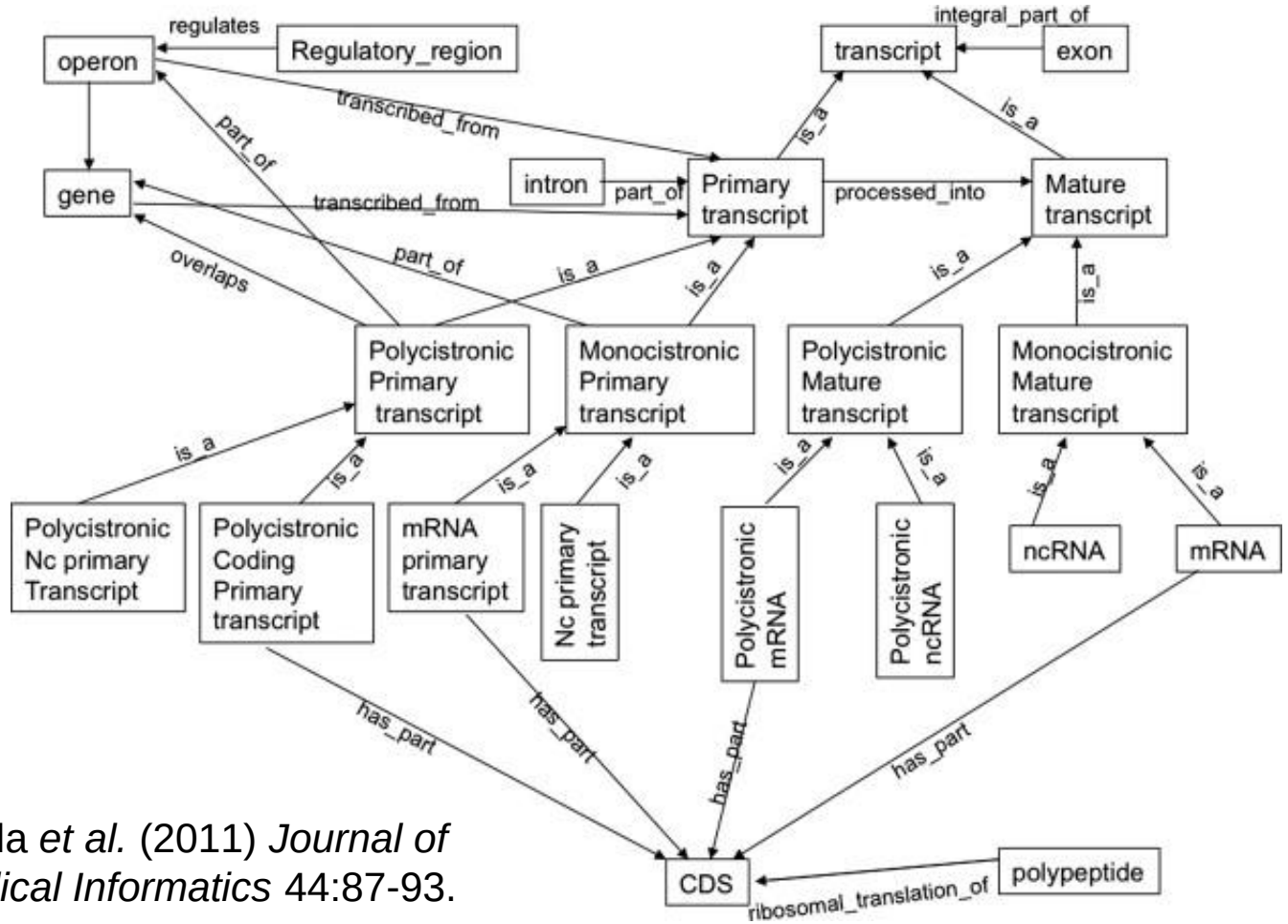


Sequence Ontology

SO: www.sequenceontology.org

- describes features and attributes of biological sequences, e.g. as defined by their disposition to be involved in a biological process, e.g. **binding_site** and **exon**

- describes primary annotations of nucleic acid sequence, and of mutations
- a structured SO within databases allows for query for e.g. *genes whose transcripts are edited, or trans-spliced, or are bound by a particular protein*.

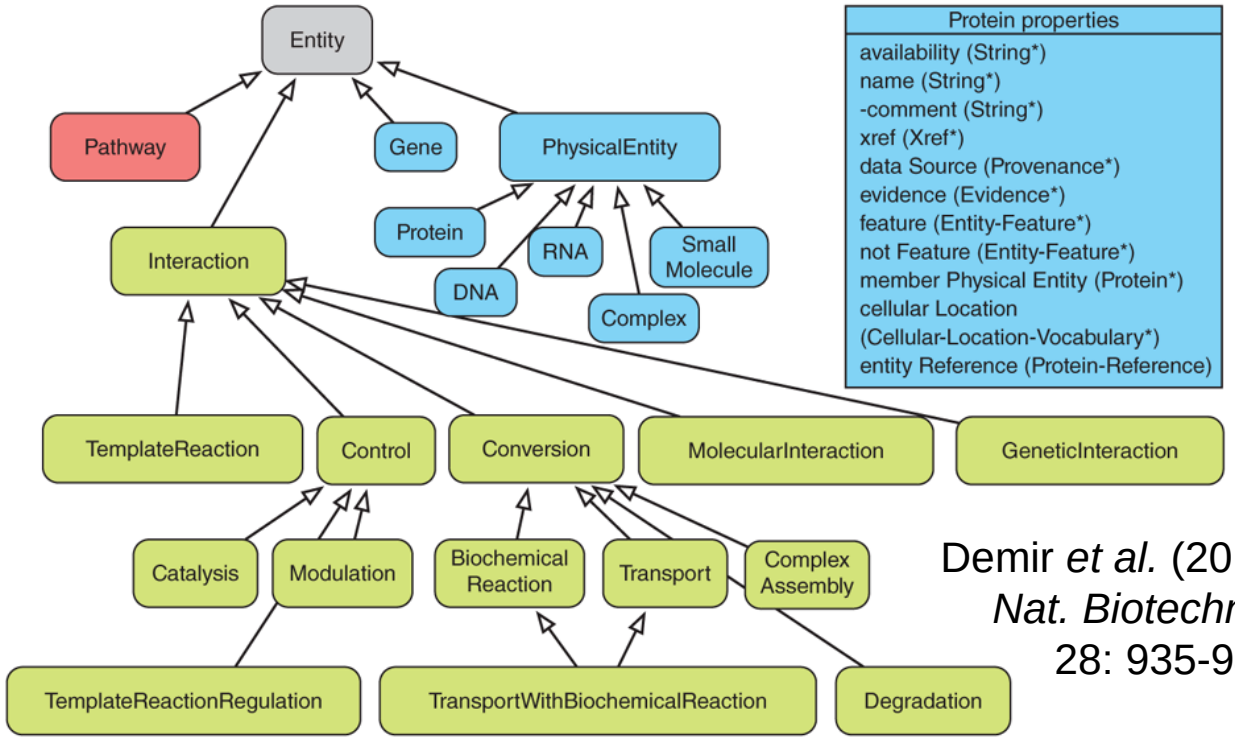


Mungalla *et al.* (2011) *Journal of Biomedical Informatics* 44:87-93.

Pathway Ontology

BioPax: www.biopax.org

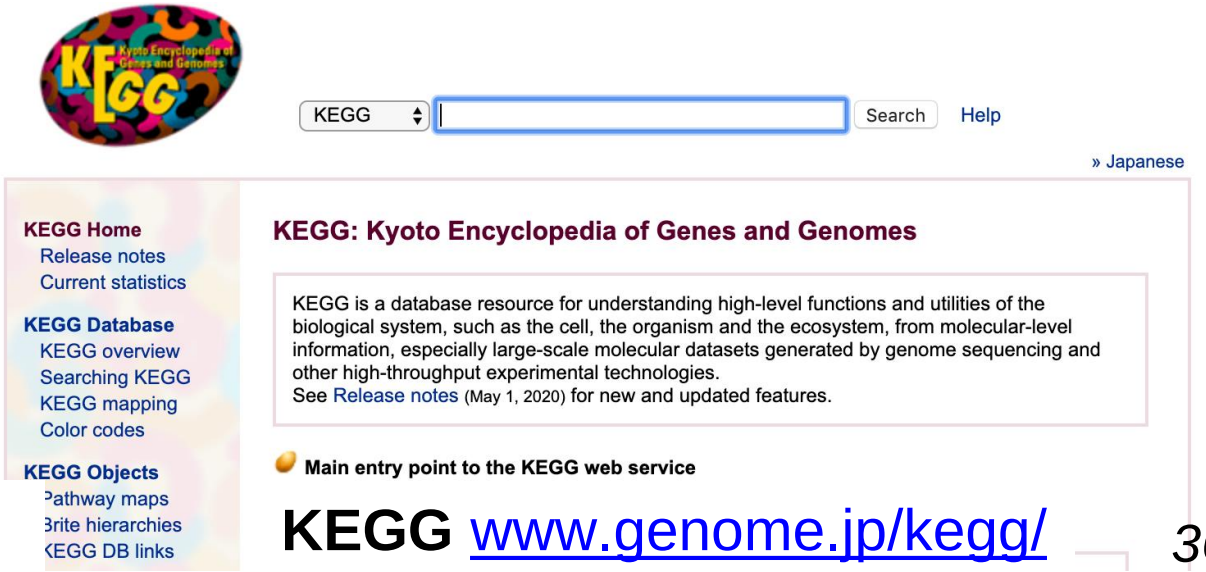
- formalisation of biochemical pathways; enables integration, exchange, visualisation and analysis of biological and signalling pathways, gene regulations, genetic interactions
- iterative development with increasing levels of biological knowledge modelled



Demir *et al.* (2010) *Nat. Biotechnol.* 28: 935-942.

Examples of databases:

Reactome www.reactome.org

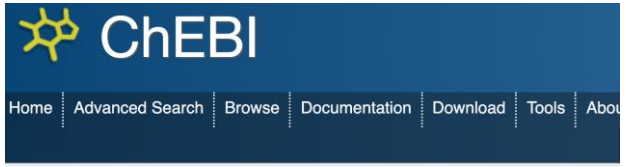


KEGG www.genome.jp/kegg/

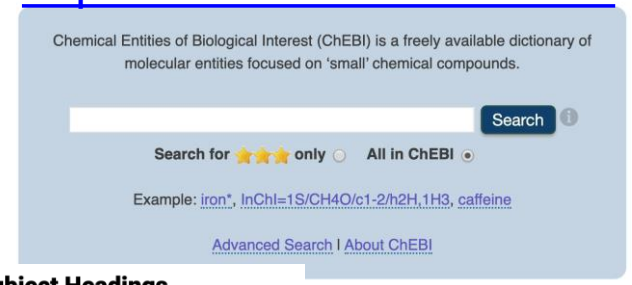
Other examples of biological ontologies

- Species-specific
  - *C. elegans* phenotype (wbphenotype)
- Chemical entities
  - Chemical Entities of Biological Interest (chebi)
- Molecular interactions
  - Protein modification (PSI-MOD), molecular interactions (PSI-MI)
- Investigations/experiments
  - Ontology for Biomedical Investigations (obi)
- Biomedical literature
  - Medical Subject Headings (MeSH)
- Many more …

Most are listed in www.obofoundry.org



https://www.ebi.ac.uk/chebi/



www.nlm.nih.gov/mesh/

Ontobee:
http://www.ontobee.org/

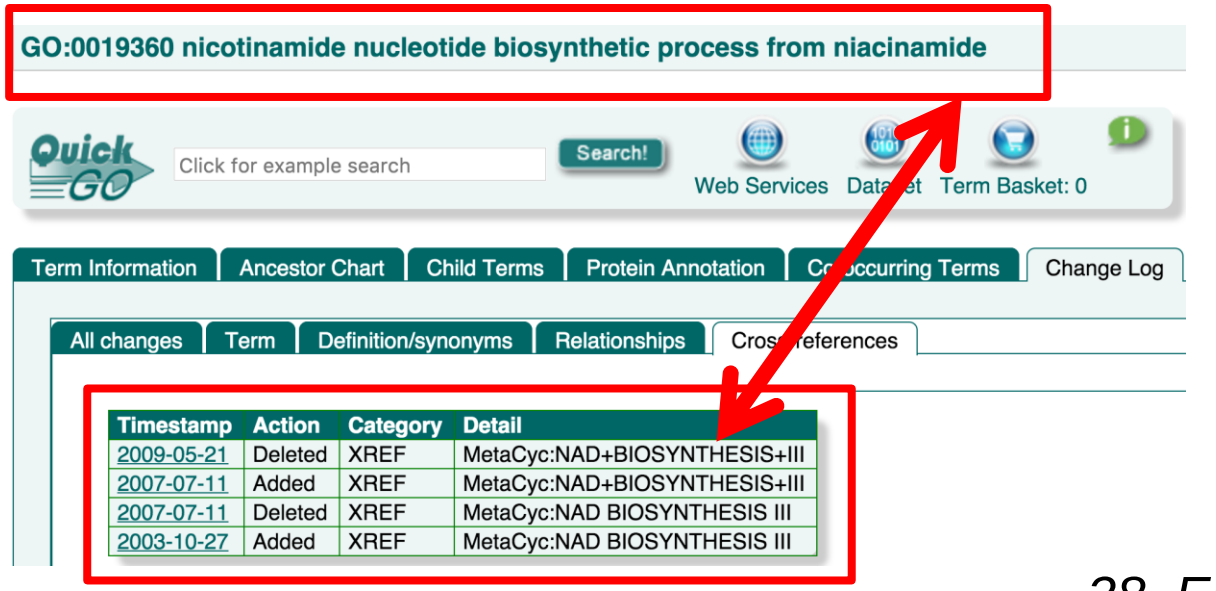
ontology data sharing, visualization, query, integration, and analysis

Biological ontologies: Issues and challenges

- Ontologies for complex entities e.g. **genotypes and phenotypes**; some on-going projects include:
  - Genotype ontology (GENO) to characterise genetic variation
  - Human Phenotype Ontology (HPO) Project: to phenotypic abnormalities encountered in human disease

https://hpo.jax.org/app/

- Terms represented in **several**, possibly **overlapping**, ontologies - may cause errors in data cross-linking



- **Funding** support for maintenance