

# **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

# 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	Graphs	High-dimensional data	Geometric information
<pre>CGACTACGATCAGCTA CGACGACTACGATCAG CATGCATCACAGCTAG CATCGACTAGCATCGA TCACGACTAGCGCATG</pre>			
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](https://en.wikipedia.org/wiki/Biological_data)

# Example 1: sequences

```
>sequence1
MDSKGSSQKGSRLLLLLVSNLLCQGVSTPVCPNGPGNCQVSLRDLFDRAMVSHYIHDLS
EMFNEFDKRYAQKGFITMALNSCHTSSLPTPEDKEQAQQTHEVLMSSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSILGQNPTAEALQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIKY
IGTNLVEWIWGGFSVDKATLNRFFAFHFLPFTMVALAGVHLTFLHETGSNNPLGLTSSDKIP
FHPYYTIKDFLGLLILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRS
VPNKLGGVLALFLSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDDLTLTWIGSQPVEYPY
IIIGQMASILYFSIILAFLPIAGXIENY
```

FASTA format

PHYLIP format

```
11 24
Seq5541 LNPSAPTVVVLGWL GASIKHLAKY
Seq7032 QKPKRPTAVLLGWFAAKHKNLSKY
Seq7152 MRPARGFVVVLGWF GAQDKHLKKY
Seq2877 LVPWAPTA VLLGWVG CQMRYLRKY
Seq0056 ---RPLVLT LGWLGANERH LGKY
Seq0781 GFLLNPLVIVMGWHGCKPRYLSKY
Seq2239 QDPASVIVVLLG WYACHPKVLAKY
Seq5612 KFPK VPIVMLLGWAGCQDRYLMKY
Seq4904 VFSEEPV VILLGWAGSRDKHLAKY
Seq5924 EIPDLPV ILLGWGGCS DKNLAKY
Seq7619 EIPDQPV VILLGWGGCRDKNLAKY
```

Sequence  
data

FASTA format

```
>Seq5541
LNPSAPTVVVLGWL GASIKHLAKY
>Seq7032
QKPKRPTAVLLGWFAAKHKNLSKY
>Seq7152
MRPARGFVVVLGWF GAQDKHLKKY
>Seq2877
LVPWAPTA VLLGWVG CQMRYLRKY
>Seq0056
----RPLVLT LGWLGANERH LGKY
>Seq0781
GFLLNPLVIVMGWHGCKPRYLSKY
>Seq2239
QDPASVIVVLLG WYACHPKVLAKY
>Seq5612
KFPK VPIVMLLGWAGCQDRYLMKY
>Seq4904
VFSEEPV VILLGWAGSRDKHLAKY
>Seq5924
EIPDLPV ILLGWGGCS DKNLAKY
>Seq7619
EIPDQPV VILLGWGGCRDKNLAKY
```

Multiple sequence  
alignment

```
#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 LNPSAPTVVVLGWL GASIKHLAKY
Seq7032 QKPKRPTAVLLGWFAAKHKN LSKY
Seq7152 MRPARGFVVVLGWF GAQDKHLKKY
Seq2877 LVPWAPTA VLLGWVG CQMRY LRKY
Seq0056 ----RPLVLT LGWLGANERH LGKY
Seq0781 GFLLNPLVIVMGWHGCKPRYLSKY
Seq2239 QDPASVIVVLLG WYACHPKVLAKY
Seq5612 KFPK VPIVMLLGWAGCQDRYLMKY
Seq4904 VFSEEPV VILLGWAGSRDKHLAKY
Seq5924 EIPDLPV ILLGWGGCS DKNLAKY
Seq7619 EIPDQPV VILLGWGGCRDKNLAKY

;
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 025b:H4-ST131  
**ORGANISM** Escherichia coli 025b: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 025b:H4-ST131 clone  
**JOURNAL** PLoS ONE 9 (8), e104400 (2014)  
 25126841  
**PUBMED**  
**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**  
 source Location/Qualifiers  
 1..5109767 /organism="Escherichia coli 025b: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 025b:H4-ST131 clone";  
**RL** PLoS One 9(8):e104400-e104400(2014).  
**...**  
**FH** Key Location/Qualifiers  
**FH** 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

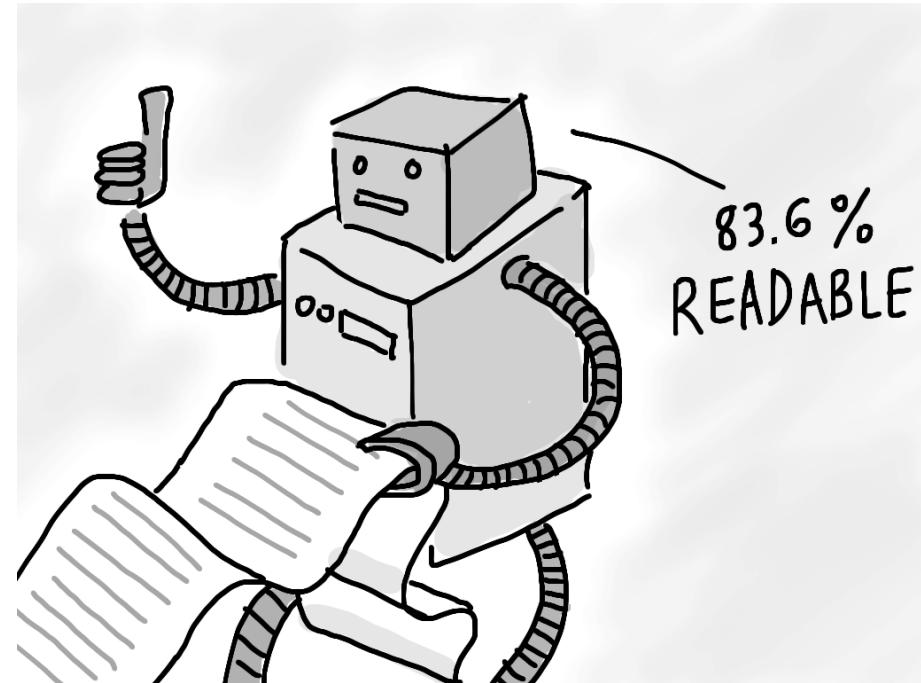
# Biological databases

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

<b>Type of data</b> <ul style="list-style-type: none"><li>• Nucleotide</li><li>• Protein</li><li>• Macromolecular structure</li><li>• Gene expression</li><li>• Metabolic pathways</li><li>• Protein-protein interactions</li><li>• ...</li></ul>	<b>Data entry &amp; quality control</b> <ul style="list-style-type: none"><li>• Scientists (teams) deposit data directly</li><li>• Appointed curators add and update data</li><li>• Are erroneous data removed or marked?</li><li>• Type and degree of error checking</li><li>• Consistency, redundancy, conflicts, updates</li></ul>	<b>Primary or derived data</b> <ul style="list-style-type: none"><li>• Primary databases: experimental results directly into database</li><li>• Secondary databases: results of analysis of primary databases</li><li>• Aggregate of many databases</li><li>• Links to other data items</li><li>• Combination of data</li><li>• Consolidation of data</li></ul>
<b>Technical design</b> <ul style="list-style-type: none"><li>• Flat-files</li><li>• Relational database (MySQL)</li><li>• Object-oriented database (e.g. XML, JSON formats, PostgreSQL)</li></ul>	<b>Maintainer status</b> <ul style="list-style-type: none"><li>• Large, public institution (e.g. EMBL, NCBI)</li><li>• Quasi-academic institute (e.g. Swiss Institute of Bioinformatics, J. Craig Venter Institute)</li><li>• Academic group or scientist</li><li>• Commercial company</li></ul>	<b>Availability</b> <ul style="list-style-type: none"><li>• Publicly available, no restrictions</li><li>• Available, but with copyright</li><li>• Accessible, but not downloadable</li><li>• Academic, but not freely available</li><li>• Proprietary, commercial; possibly free for academics</li></ul>

*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](http://www.ncbi.nlm.nih.gov)
- EMBL-EBI  
[www.ebi.ac.uk](http://www.ebi.ac.uk)
- DNA DataBank of Japan  
[www.ddbj.nig.ac.jp](http://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](http://www.uniprot.org)

The screenshot shows the UniProtKB homepage. It features two main sections highlighted with red boxes:

- Swiss-Prot (562,253)**: Manually annotated and reviewed. Records with information extracted from literature and curator-evaluated computational analysis.
- TrEMBL (180,690,447)**: Automatically annotated and not reviewed. Records that await full manual annotation.



### 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](http://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)

## 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

The screenshot shows the RGAP 7 Summary page. It features a header with the project name and funding information, followed by a main content area with a table of genome statistics and a sidebar with links to various tools and resources.

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



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

## Welcome to Kinweb

Kinweb is a collection of protein kinases encoded in the human genome. This site provides:

- a comprehensive analysis of functional domains of each gene product;
- a prediction of secondary and tertiary structure motifs by using machine learning
- a collection of conserved sequence elements identified by comparative analysis of exons, or other regulatory elements.



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

# Other context-specific databases

- **InterPro:** [www.ebi.ac.uk/interpro](http://www.ebi.ac.uk/interpro)
  - domain models of proteins from ~20 primary resources
- **Protein Data Bank (PDB):** [www.rcsb.org/pdb/](http://www.rcsb.org/pdb/)
  - high-resolution 3D structural data of proteins
- **Gene Expression Omnibus (GEO):** [www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/)
- **ArrayExpress:** [www.ebi.ac.uk/arrayexpress/](http://www.ebi.ac.uk/arrayexpress/)
  - high-throughput functional genomics (gene expression) data
- **Oncomine:** [www.oncomine.org](http://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](http://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/](http://www.ncbi.nlm.nih.gov/pubmed/)
  - scientific literature relevant to biomedical fields
- and many more ...

## Trending Articles

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.](#)

Violli 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.](#)

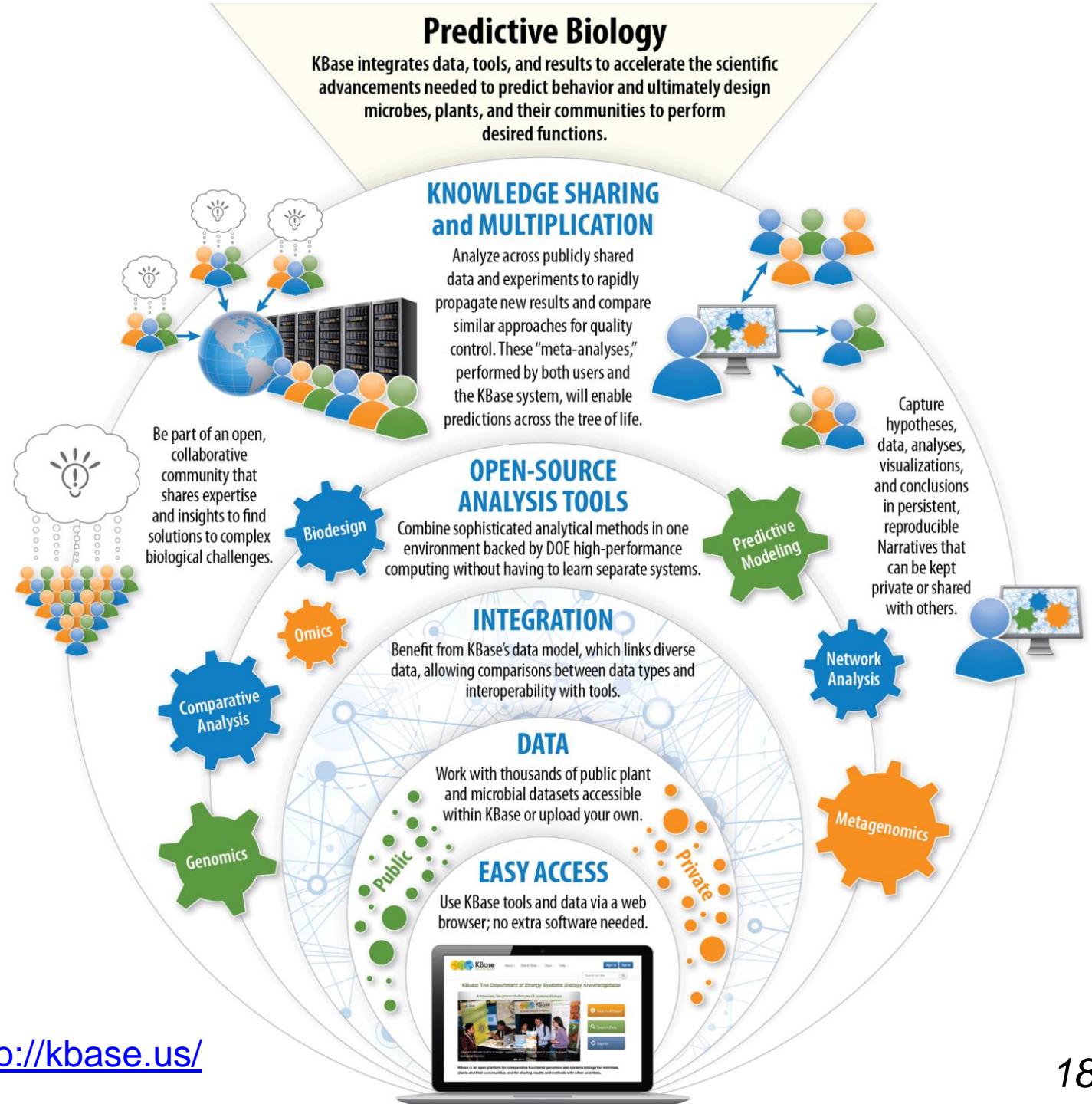
Al-Kofahi M, et al. Clin Pharmacol Ther. 2020. PMID: 32344449

# 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**

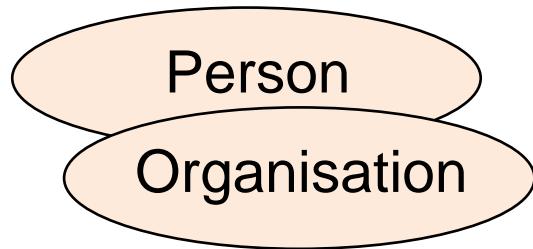
# *Why do we need ontology?*

- 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](http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html)

## Class



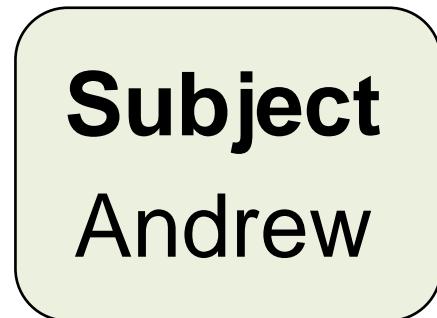
# Basic terminology

**Relations**  
has employer,  
is employee of  
....



## Instance (Individual)

### Triple

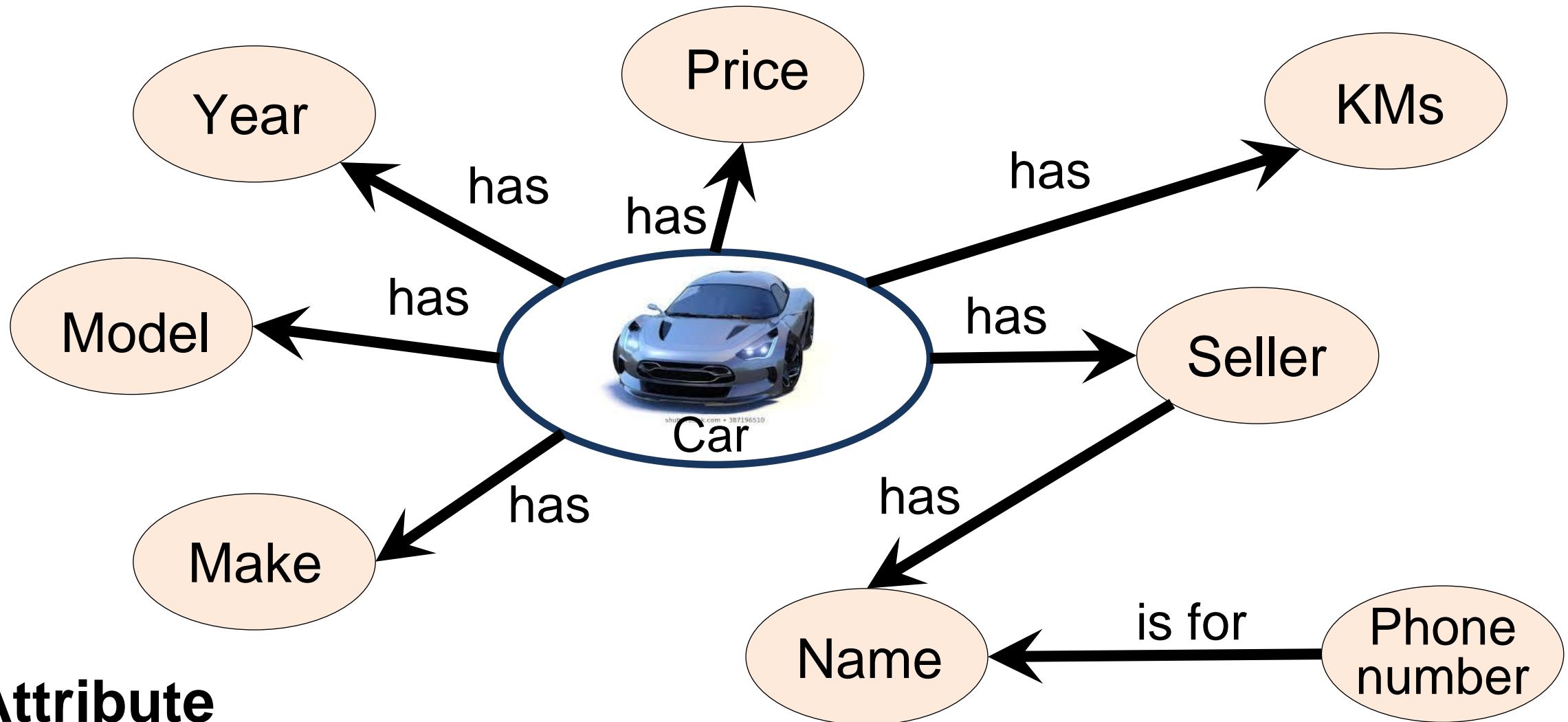


### Predicate

has employer



## Example: a car advertisement

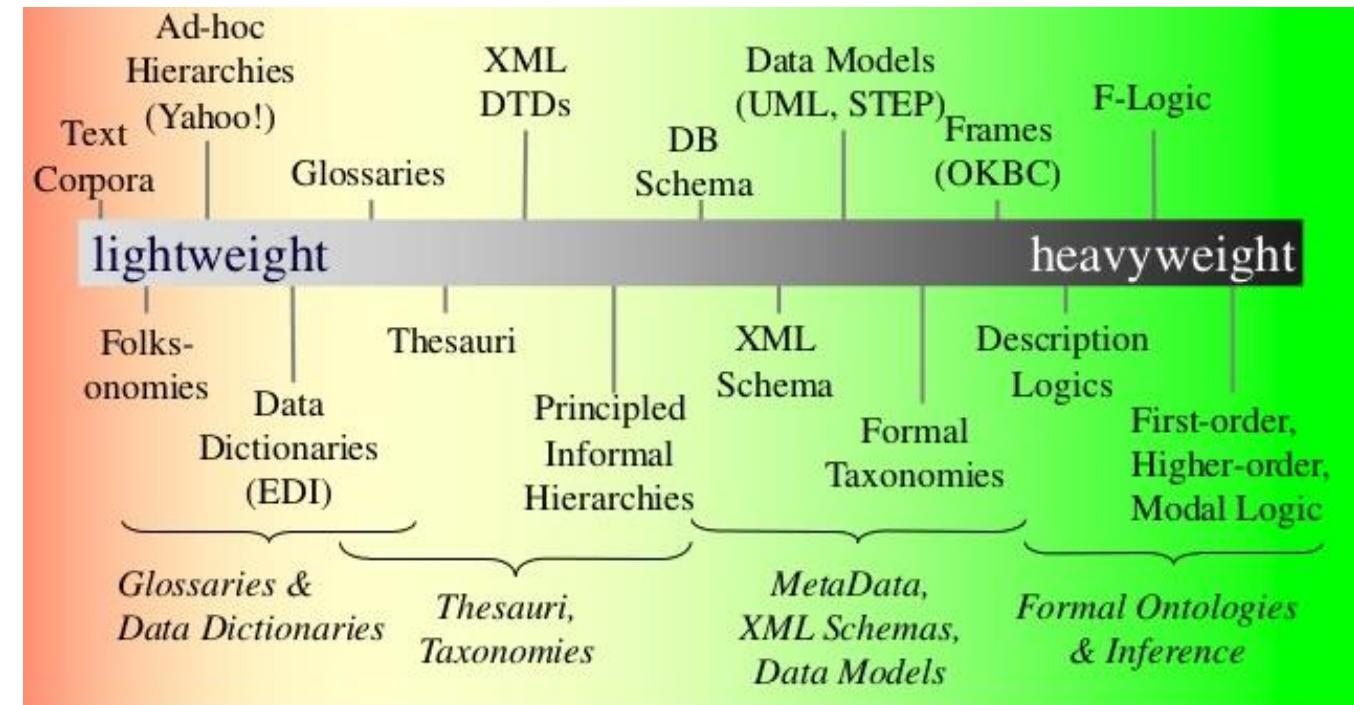


**Attribute**  
(property, characteristics)

# Types of ontology

## Light weight ontologies

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



## Heavy weight ontologies

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

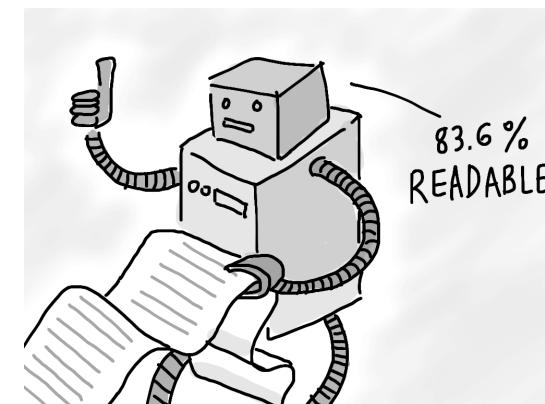
Fernando Silva Parreiras. *Model-Driven Software Development with Semantic Web Technologies.*

<http://www.slideshare.net/fparreiras/modeldriven-software-development-with-semantic-web-technologies>

# 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](http://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.github.io](https://github.com/OBOFoundry/OBOFoundry.github.io) or get in touch with us at [obo-discuss@sourceforge.net](mailto:obo-discuss@sourceforge.net).

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

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

*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)*

# Resources of biological ontologies

## EBI Ontology Lookup Service

<https://www.ebi.ac.uk/ols>

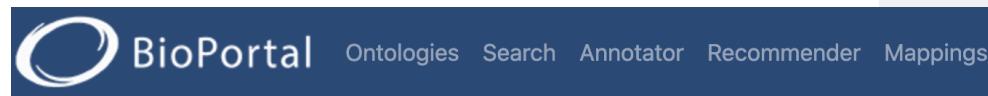
- Ontologies from OBO Foundry



## NCBO BioPortal

[bioportal.bioontology.org](http://bioportal.bioontology.org)

- open repository; user can add notes, review & map



Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class

Enter a class, e.g. Melanoma



Advanced Search

Find an ontology

Start typing ontology name, then choose from list



Browse Ontologies ▾

### Data Content

Updated 03 May  
2021 06:06

- 263 ontologies
- 6,463,053 terms
- 31,820 properties
- 497,626 individuals

### BioPortal Statistics

Ontologies 868

Classes 9,914,067

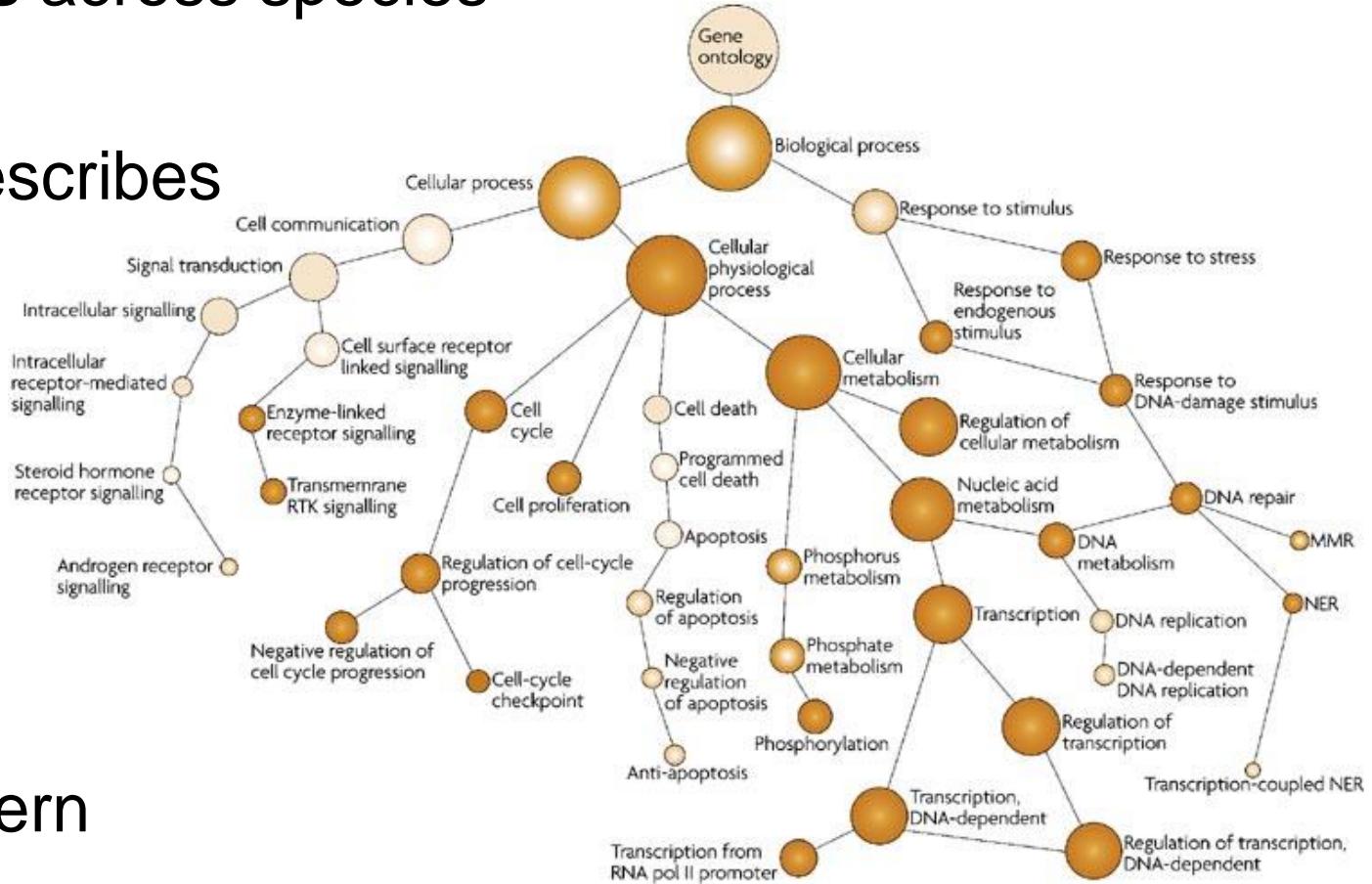
Properties 36,286

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



Nature Reviews | Cancer

“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:

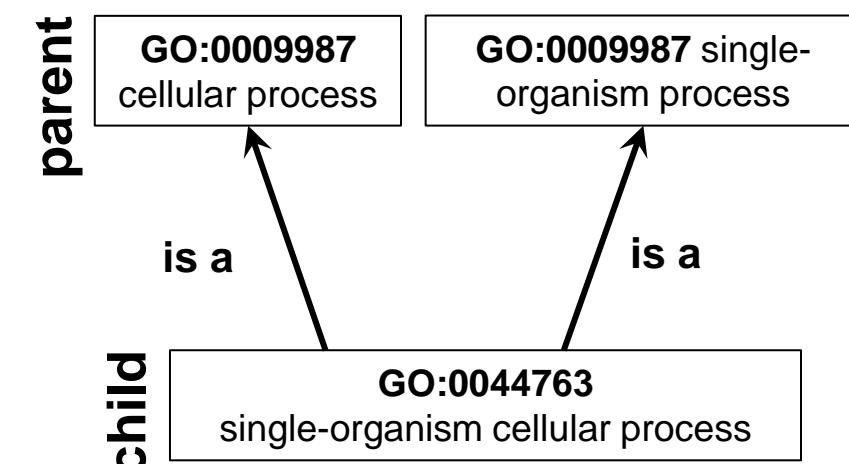


These ontologies resemble an **hierarchy**:

- **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
```

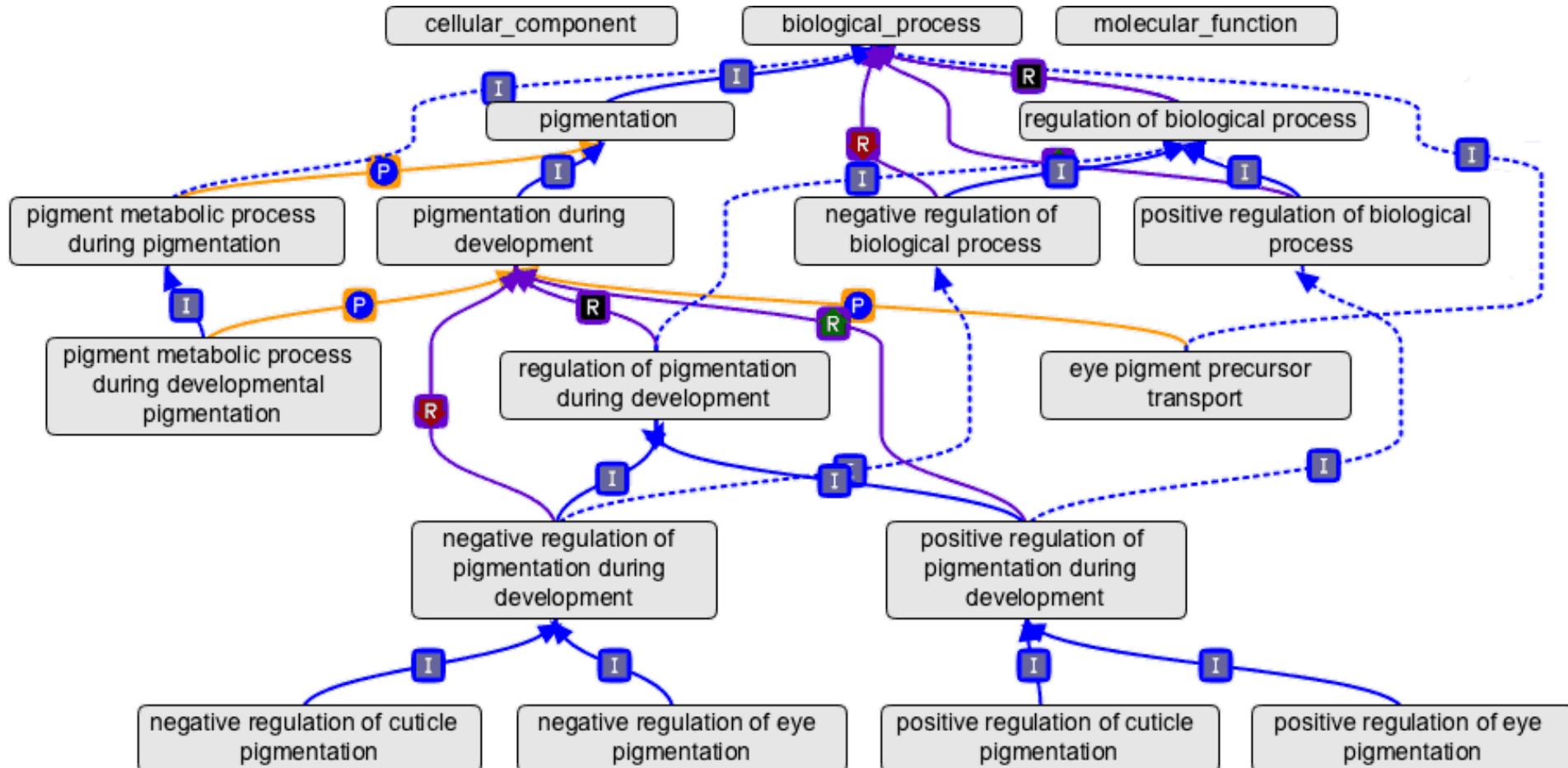


OBO format

# 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

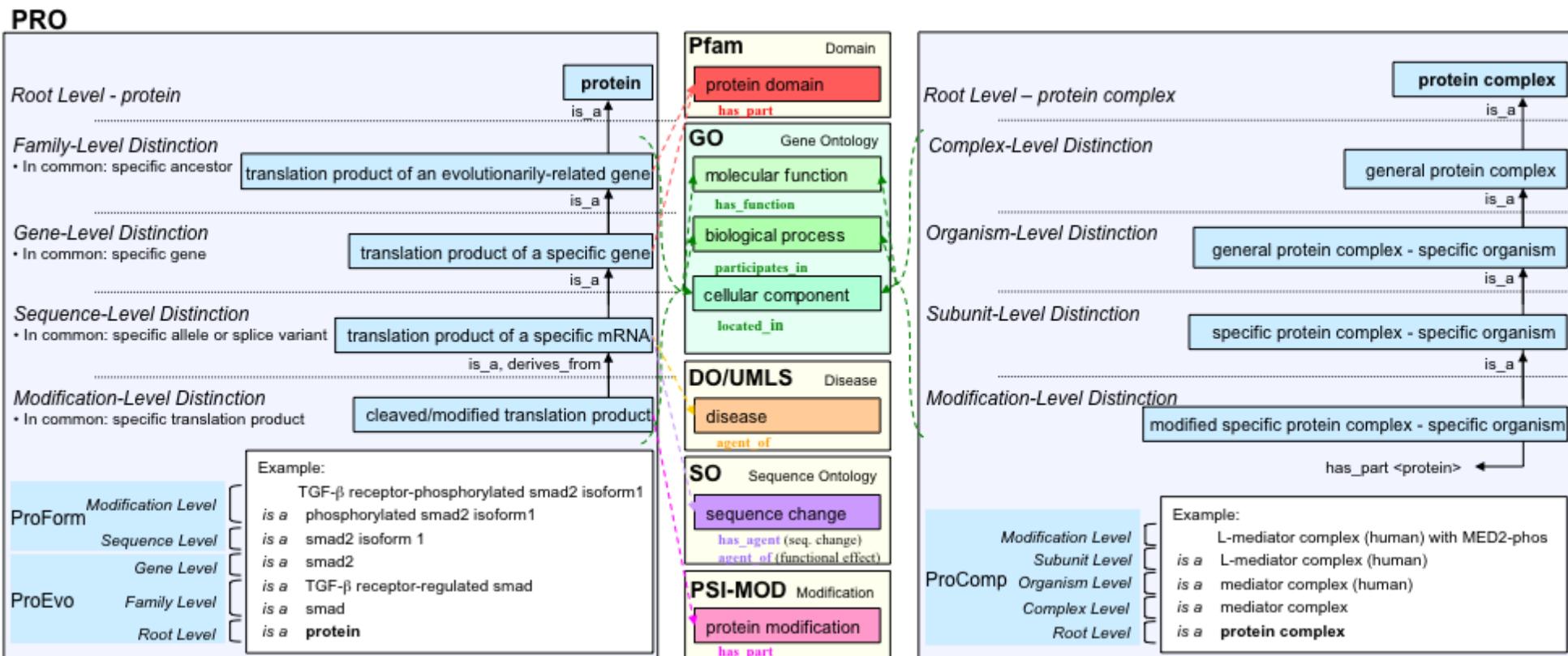
P: part of

R: regulates (one process directly affects the manifestation of another process or quality)

# Protein Ontology

PRO [pir.georgetown.edu/pro/](http://pir.georgetown.edu/pro/)

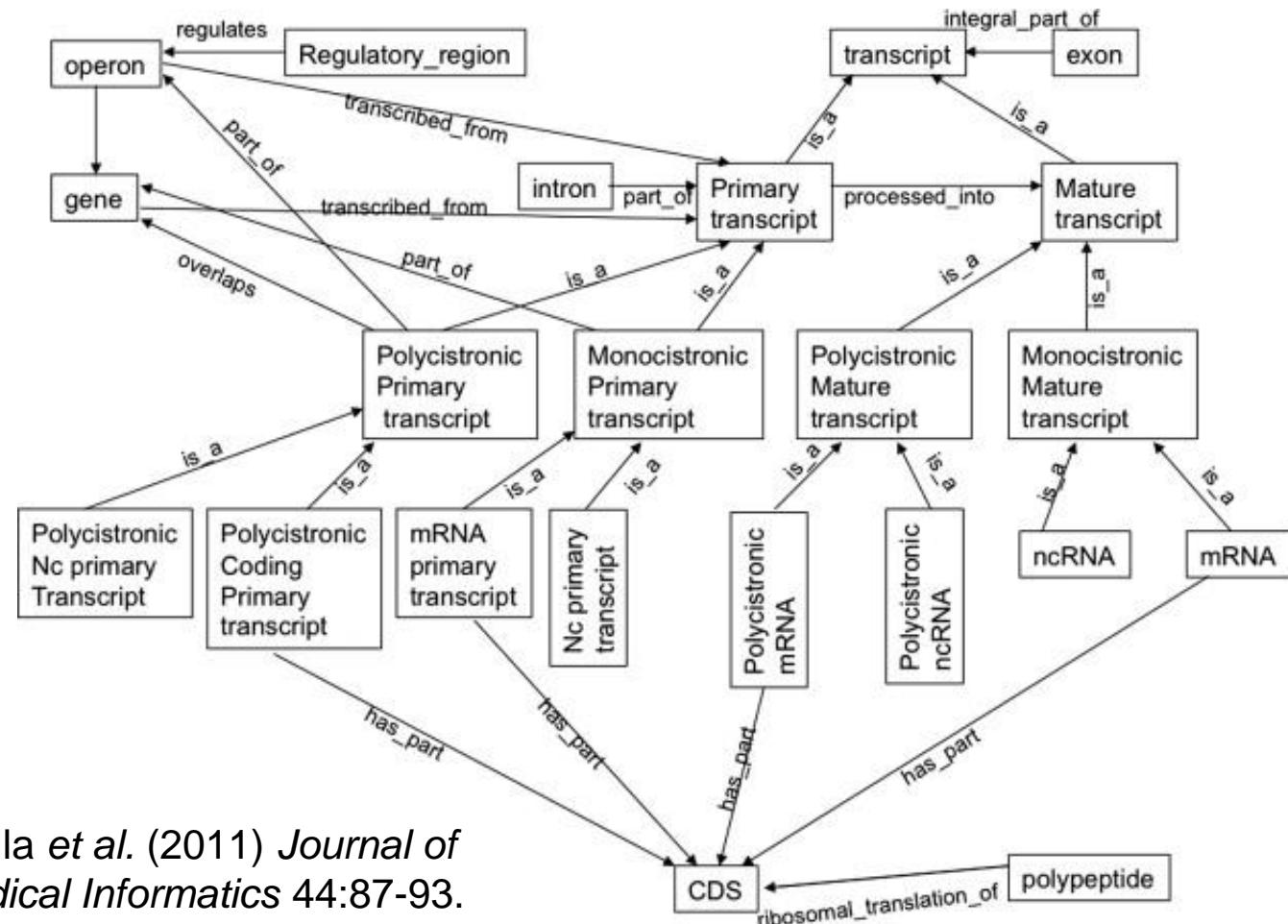
- ontological representation of protein-related entities
- three sub-ontologies:
  - ProEvo (based on evolutionary relatedness)
  - ProForm (protein forms produced from a given gene locus); and
  - ProComp (protein-containing complexes)



# Sequence Ontology

SO: [www.sequenceontology.org](http://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.*

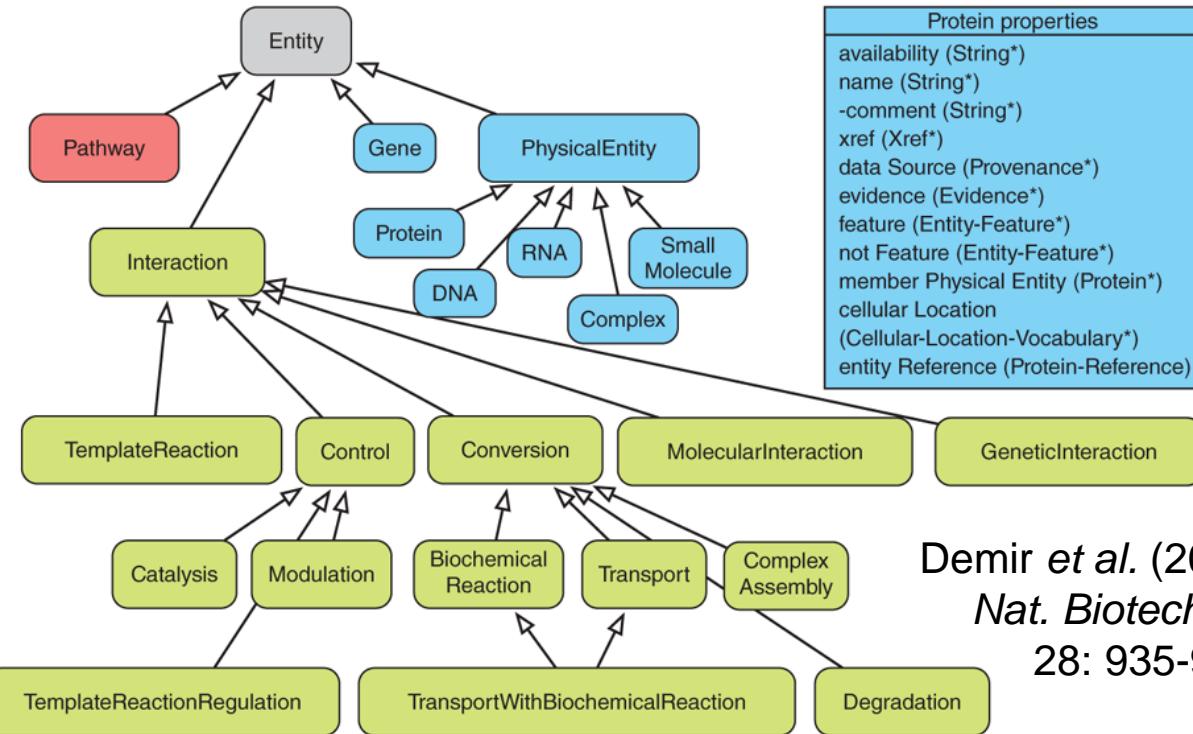


Mungalla et al. (2011) *Journal of Biomedical Informatics* 44:87-93.

# Pathway Ontology

BioPax: [www.biopax.org](http://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.

The Reactome website features a search bar at the top with the placeholder "Find Reactions, Proteins and Pathways" and a "Go!" button. Below the search bar is a text input field containing "e.g. 095631, NTN1, signaling by EGFR, glucose". The main content area displays four icons: "Pathway Browser" (a stack of three circles), "Analyze Data" (a bar chart), "ReactomeFIViz" (two overlapping circles), and "Documentation" (a document icon). A navigation bar at the top includes links for "About", "Content", "Docs", "Tools", "Community", and "Download".

Examples of databases:



Pathway  
Browser

Visualize and interact with Reactome biological pathways



Analyze Data  
Merges pathway identifier



ReactomeFIViz  
Designed to find pathways and



Documentation  
Information to browse the

Reactome [www.reactome.org](http://www.reactome.org)



KEGG

Search Help

» Japanese

**KEGG Home**  
Release notes  
Current statistics  
**KEGG Database**  
KEGG overview  
Searching KEGG  
KEGG mapping  
Color codes  
**KEGG Objects**  
Pathway maps  
Brite hierarchies  
KEGG DB links

## KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes](#) (May 1, 2020) for new and updated features.

Main entry point to the KEGG web service

**KEGG** [www.genome.jp/kegg/](http://www.genome.jp/kegg/)

# Other examples of biological ontologies

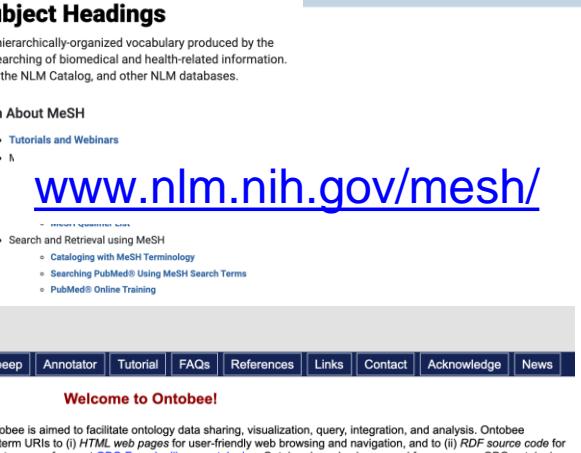
- Species-specific
  - *C. elegans* phenotype (wbphenotype)
- Chemical entities
  - Chemical Entities of Biological Interest (chebi)
- Molecular interactions
  - Protein modification (PSI-MOD), molecular interactions (PSI-MI)
- Investigations/experiments
  - Ontology for Biomedical Investigations (obi)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

Most are listed in  
[www.obofoundry.org](http://www.obofoundry.org)

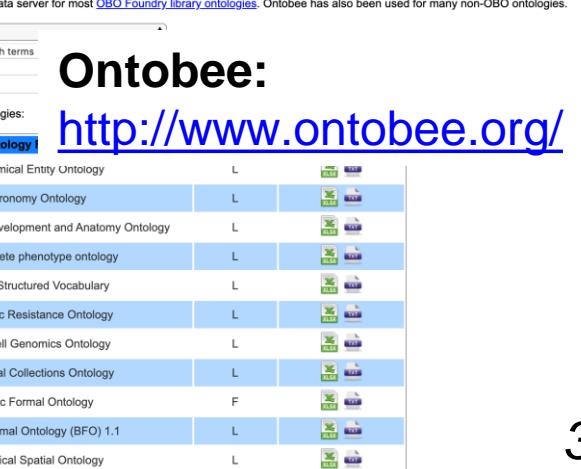
ontology data sharing,  
visualization, query,  
integration, and  
analysis



The screenshot shows the ChEBI homepage. At the top, there's a navigation bar with links for Home, Advanced Search, Browse, Documentation, Download, Tools, and About. Below the navigation is a search bar with placeholder text "Search for iron\*, InChI=1S/CH4O/c1-2/h2H,1H3, caffeine". There are also buttons for "Search", "Search for ★★★ only", and "All in ChEBI". Below the search bar, there's a brief description: "Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds." At the bottom of the page are links for "Advanced Search" and "About ChEBI".



The screenshot shows the MeSH homepage. It features a "Welcome to Medical Subject Headings" header. Below it, a paragraph describes MeSH as a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine, used for indexing, cataloging, and searching of biomedical and health-related information. It includes subject headings from MEDLINE/PubMed, the NLM Catalog, and other NLM databases. The page has sections for "What's New" (with a link to the "What's New" page) and "Learn About MeSH" (with links to "Tutorials and Webinars" and "Search and Retrieval using MeSH"). A search bar at the top right contains the URL "www.nlm.nih.gov/mesh/".



The screenshot shows the Ontobee homepage. It features a "Welcome to Ontobee!" header. Below it, a paragraph describes Ontobee as a linked data server for ontologies, designed to facilitate ontology data sharing, visualization, query, integration, and analysis. It uses dynamic dereferencing to present individual ontology term URIs as user-friendly web pages and provides source code for Semantic Web applications. Ontobee is the default linked data server for most OBO Foundry library ontologies. The page includes a search bar, a table of ontologies with columns for No., Ontology Prefix, and Ontology, and a footer with links to various ontologies like AEO, AGRO, AMPHX, APO, APOLLO\_SV, ARO, BCGO, BCO, BFO, BFO11, and BSPO.

**Ontobee:**  
<http://www.ontobee.org/>

# 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
- Terms represented in **several**, possibly **overlapping**, ontologies
  - may cause errors in data cross-linking
- **Funding** support for maintenance



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

A screenshot of the QuickGO website interface. At the top, a search bar contains the term "GO:0019360 nicotinamide nucleotide biosynthetic process from niacinamide". Below the search bar are various navigation links: "Quick GO", "Click for example search", "Search!", "Web Services", "Dataset", "Term Basket: 0", and an "i" button. The main content area shows tabs for "Term Information", "Ancestor Chart", "Child Terms", "Protein Annotation", "Co-occurring Terms", and "Change Log". The "Term Information" tab is active. Below these tabs is a navigation bar with buttons for "All changes", "Term", "Definition/synonyms", "Relationships", and "Cross-references". A red box highlights the "Term" button. At the bottom of the page is a table titled "Change Log" with columns: "Timestamp", "Action", "Category", and "Detail". The table lists four entries with red boxes highlighting the "Detail" column.

Timestamp	Action	Category	Detail
2009-05-21	Deleted	XREF	MetaCyc:NAD+BIOSYNTHESIS+III
2007-07-11	Added	XREF	MetaCyc:NAD+BIOSYNTHESIS+III
2007-07-11	Deleted	XREF	MetaCyc:NAD BIOSYNTHESIS III
2003-10-27	Added	XREF	MetaCyc:NAD BIOSYNTHESIS III

## Database 2

# Searching for sequences in databases

---

**Cheong Xin Chan (CX)**

c.chan1@uq.edu.au

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

# Outline

- **Basic concepts of sequence searching**
  - The general approach and concept of sub-sequences ( $k$ -tuples or  $k$ -mers)
- **FastA**
  - Basic principles
  - Hashing-and-chaining algorithm
- **Basic Local Alignment Search Tool (BLAST)**
  - Differences between FastA and BLAST
  - Basic parameters of BLAST
  - BLAST algorithm using Finite State Machine

## The basic concepts

- to search for **similar** sequences (in a database) to a query, **alignment** is necessary
- Needleman-Wunsch or Smith-Waterman algorithm is too slow for this purpose; faster approach was developed: **FastA** and **BLAST**
- assumption: a good local alignment should have some identical subsequences (i.e. **exact matches** of sub-sequences)
- these sub-sequences at a fixed length  $k$ , are referred to as  **$k$ -tuples** or  **$k$ -mers**
- usually, smaller  $k$  (e.g. 2–3) is used for protein sequences; larger  $k$  (e.g. 3–6) for DNA sequences

Protein 2-tuples ( $k = 2$ ) → **AN**, **AR**, ...

*Example*

DNA 4-tuples ( $k = 4$ ) → **TAAA**, **TAAC**, ...

# General approach for sequence searching

1. Pre-process **query** string, e.g. generate short sub-sequences to search for
2. Quickly **align sub-sequences** with sequences in the database, keeping only high-scoring sub-alignments
3. Attempt to **join sub-alignments**, creating tentative scores
4. Perform a **thorough alignment** on high-scoring sequences

Sub-sequences at defined length  $k$  are known variously as  **$k$ -tuples**,  **$k$ -mers**, **words**, or  **$n$ -grams** (of length  $n$ )

# FastA: searching similar sequences

- developed by Lipman & Pearson (1985)
- first described as FastP (P for protein); FastA (A for all) work on both nucleotide and protein sequences
- **FASTA** format – the most common text-based representation of biological sequences



David L. Lipman

FastA Server: [http://fasta.bioch.virginia.edu/fasta\\_www2/](http://fasta.bioch.virginia.edu/fasta_www2/)

## Key steps:

1. observe the pattern of word hits
2. identify word-to-word matches of a given length
3. mark potential matches
4. perform an optimised search using a Smith-Waterman type of algorithm



William R. Pearson

# FastA: searching using hashing and chaining

## Hashing

- an efficient approach for data storage and retrieval, common in computing
- in FastA, each distinct data entry (e.g.  $k$ -tuples) is assigned a **unique** integer (this saves storage space), i.e. the **index** (or **key**)
- relevant data (the *values*), i.e. position(s) at which the corresponding  $k$ -tuple is found on the query sequence, are stored in relation to each **index**
- data are structured in a *key-value* relationship

# FastA: searching using hashing and chaining

## Hashing: assigning indices

At  $k = 3$  for nucleotide sequences:

- there are  $4^3 = 64$  possible 3-tuples
- let number  $e(N)$  be a distinct value for each nucleotide  $N$ :  $e(A) = 0$ ,  $e(C) = 1$ ,  $e(G) = 2$  and  $e(T) = 3$
- a 3-tuple, represented as  $x_i x_{i+1} x_{i+2}$ , is assigned a value  $C_i$ :

$$C_i = e(x_i)4^2 + e(x_{i+1})4^1 + e(x_{i+2})4^0$$

$$\begin{aligned} C_i(\text{AAA}) &= e(\text{A})4^2 + e(\text{A})4^1 + e(\text{A})4^0 \\ &= 0 \times 4^2 + 0 \times 4^1 + 0 \times 4^0 \\ &= 0 \end{aligned}$$

$$\begin{aligned} C_i(\text{CAA}) &= e(\text{C})4^2 + e(\text{A})4^1 + e(\text{A})4^0 \\ &= 1 \times 4^2 + 0 \times 4^1 + 0 \times 4^0 \\ &= 16 \end{aligned}$$

Example

- the  $C_i$  is the **index** representation of the 3-tuple

# FastA: searching using hashing and chaining

## Chaining: creating a look-up table

record the position(s) on a sequence at which the  $k$ -tuples occurred (the *values*), and assign to the corresponding **index** (the *key*)

consider **TAAAACTCTAAC** (at  $k = 3$ ):

*Example*

	<b>Index (key)</b>	<b>Position(s) (values)</b>
<b>AAA</b> →	0	2, 3
<b>AAC</b> →	1	4, 10
<b>AAG</b> →	2	–
<b>AAT</b> →	3	–
...	...	...
<b>TTT</b> →	63	–

# FastA: searching using hashing and chaining

## Chaining:

consider **TAAAACTCTAAC** (at  $k = 3$ ):

*Example*

Position number	1	2	3	4	5	6	7	8	...
Nucleotide	T	A	A	A	A	C	T	C	...

Index of 3-tuple ( $C_i$ )

AAA AAC AAG AAT ACA ACC ACG ACT TAA

0	1	2	3	4	5	6	7	...	48	...
2	4	0	0	0	0	0	5	...	1	...

First position where the 3-tuple is found on the sequence

Position number

1	2	3	4	5	6	7	...
9	3	0	10	0	0	0	...

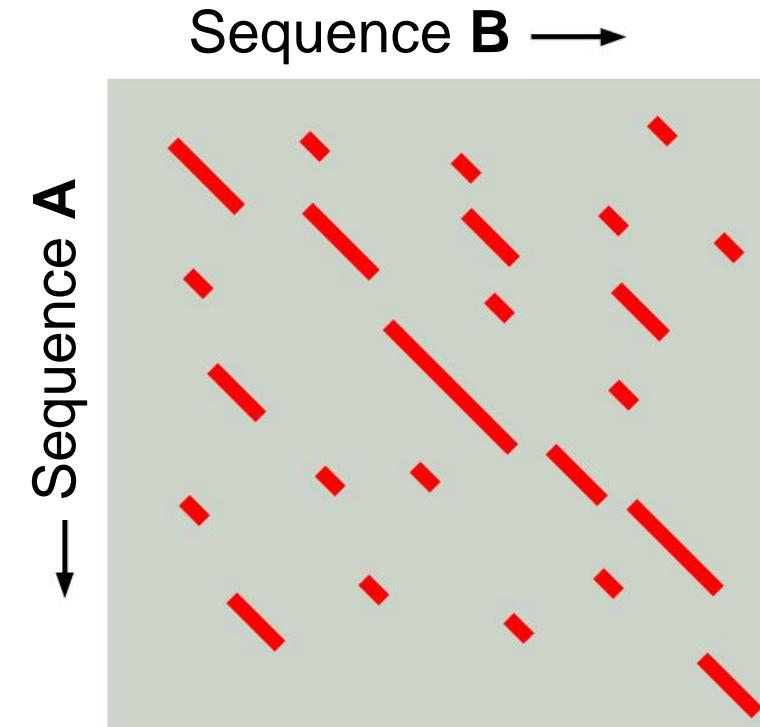
Next position the same 3-tuple is found

**AAA** found at position 2 then at 3 .... ; **AAC** found at position 4 then at 10 ... etc.

# FastA: searching using hashing and chaining

## Step 1: observe the pattern of word hits

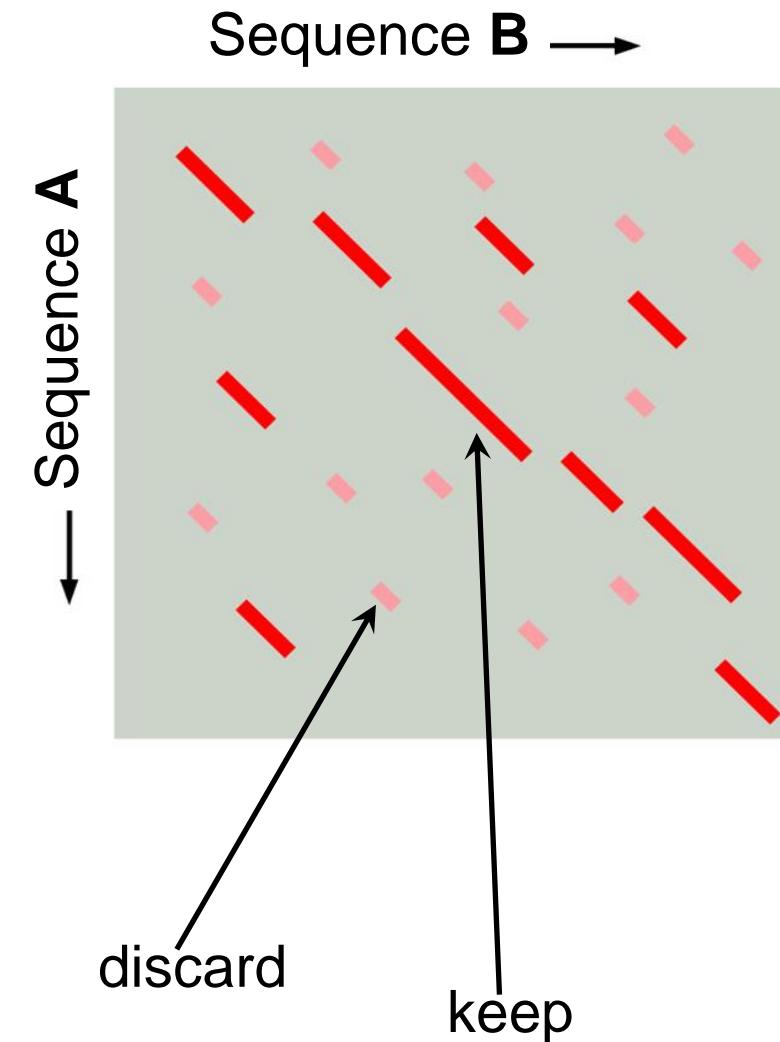
- identify perfect matches between sequence A and sequence B, using *k*-tuples via **hashing and chaining**
- look for high-density local regions between the two sequences; these are locally aligned regions



# FastA: searching using hashing and chaining

**Step 2:** identify word-to-word matches of a given length

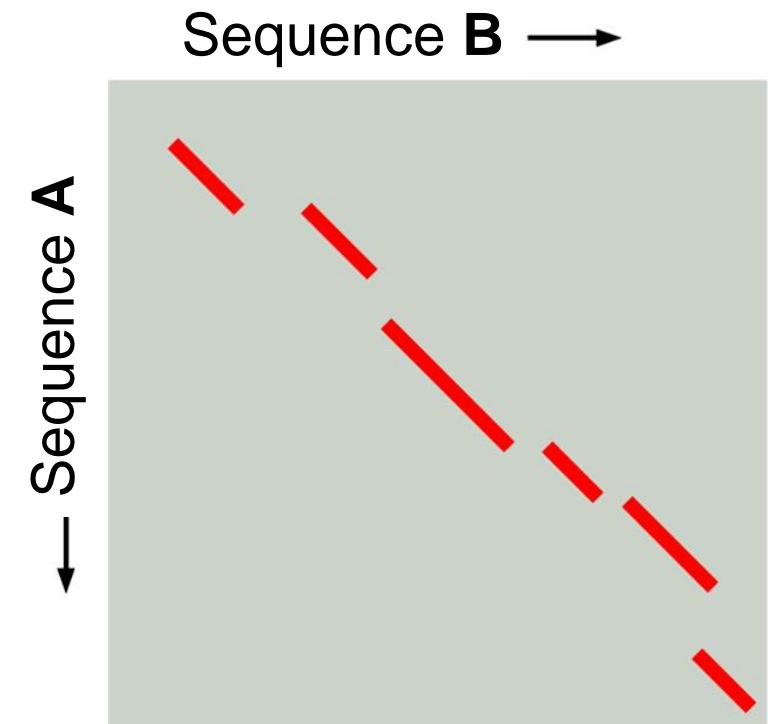
- re-score the aligned regions using a substitution scoring matrix
- keep only those contributing to the highest score, e.g. the top 10 aligned regions; discard the others



# FastA: searching using hashing and chaining

## Step 3: *mark potential matches*

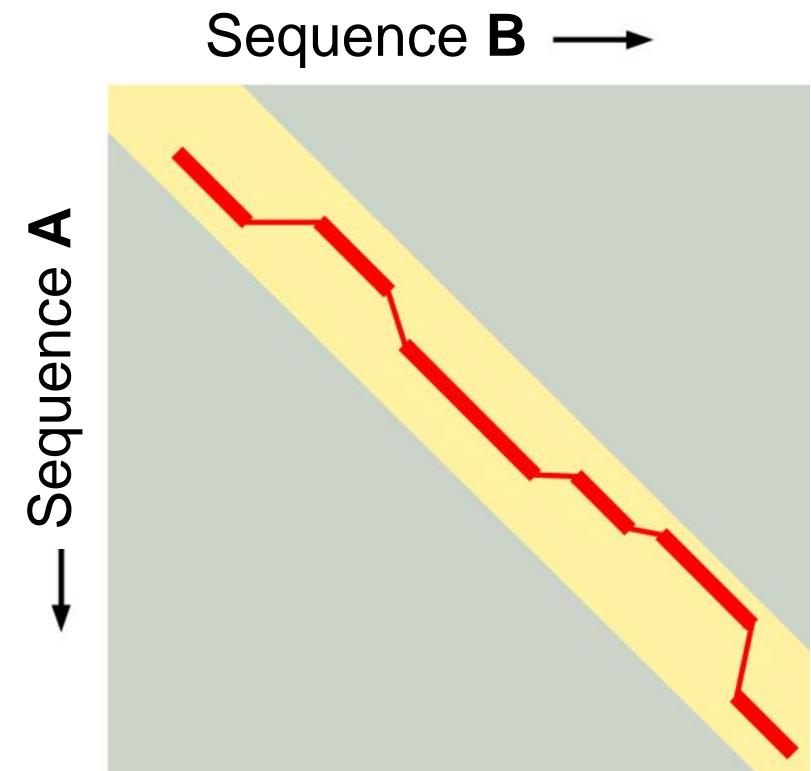
- initial regions with scores greater than a cut-off value are *joined* into an approximate alignment with gaps
- **initial similarity score** (“*initn*”) is calculated from scores of individual alignments with joining penalties
- this score is used for **preliminary ranking** of database sequences
- **init1** is the score of the single best initial region (i.e top of the rank)



# FastA: searching using hashing and chaining

## Step 4: *local alignment using SW algorithm*

- for suitably high-scoring database sequences, local alignment is performed between the two sequences using a Smith-Waterman algorithm
- The final score, “opt” is used for final sequence ranking; significance is calculated



# BLAST: Basic Local Alignment Search Tool

- developed by Altschul et al. (1990)
- search for **High-scoring Segment Pairs** (HSPs) contained in a **statistically significant** alignment
- most commonly used tool for sequence searching in major databases:  
<http://blast.ncbi.nlm.nih.gov/Blast.cgi>
- similar approach to FastA but with **two** key differences:

---

FastA	BLAST
uses identical $k$ -tuples <b>(exact matches)</b>	uses $k$ -mers in the target that score above a <b>threshold <math>T</math></b> (allowing for <b>non-exact matches</b> for protein sequences)
algorithm based on <b>hashing and chaining</b>	algorithm based on <b>finite state machine</b> (FSM), also known as <b>finite state automata</b> (FSA)

---

# BLAST: the key parameters

## Word size, i.e. $k$ in $k$ -mer

- typically **3** (2-4) for protein sequences, **11** for nucleotide sequences

## Threshold $T$ (typically an integer between 11 and 19):

- only  $k$ -mer (exact or inexact) matches that score **equal to or greater than  $T$**  (a neighbourhood score threshold) are used for seeding

## Scoring (substitution) matrix

- typically BLOSUM62; the optimal scoring matrix depends on expected sequence similarity

## Drop-off $X$

- the amount of drop in score that will stop the sequence extension, typically 20

# BLAST algorithm

**Phase 1:** remove low-complexity region or repeats in the query sequence

## *Example*

**Low-complexity** region in a sequence composes of very little variation (i.e. fewer distinct residues)

- they might yield high alignment scores but not due to biologically significance, e.g. between two non-homologous sequences (simply by chance)
  - these regions in the sequence database and in the query are commonly masked before a search

# BLAST algorithm

**Phase 2:** *identify  $k$ -mers in query with scores above threshold  $T$  for seeding*

## Finite state machine (FSM)

a computational algorithm used to model a variety of phenomena that are described by a succession of states that control the behavior of a system

- given the current state and an input, the next state of the FSM can be **deterministically determined** (not probabilistic)
- given a list of  $k$ -mers in the query (input), the FSM can quickly produce all possible matches with scores exceeding a **threshold  $T$** ; these  $k$ -mers are used for seeding (to start the alignment process)

# BLAST algorithm

**Phase 2: identify  $k$ -mers in query with scores above threshold  $T$  for seeding**

Consider a 3-mer **CHH**, and  $T = 19$

# query

Based on BLOSUM62 matrix, each of the following transitions has a score  $\geq 19$ :

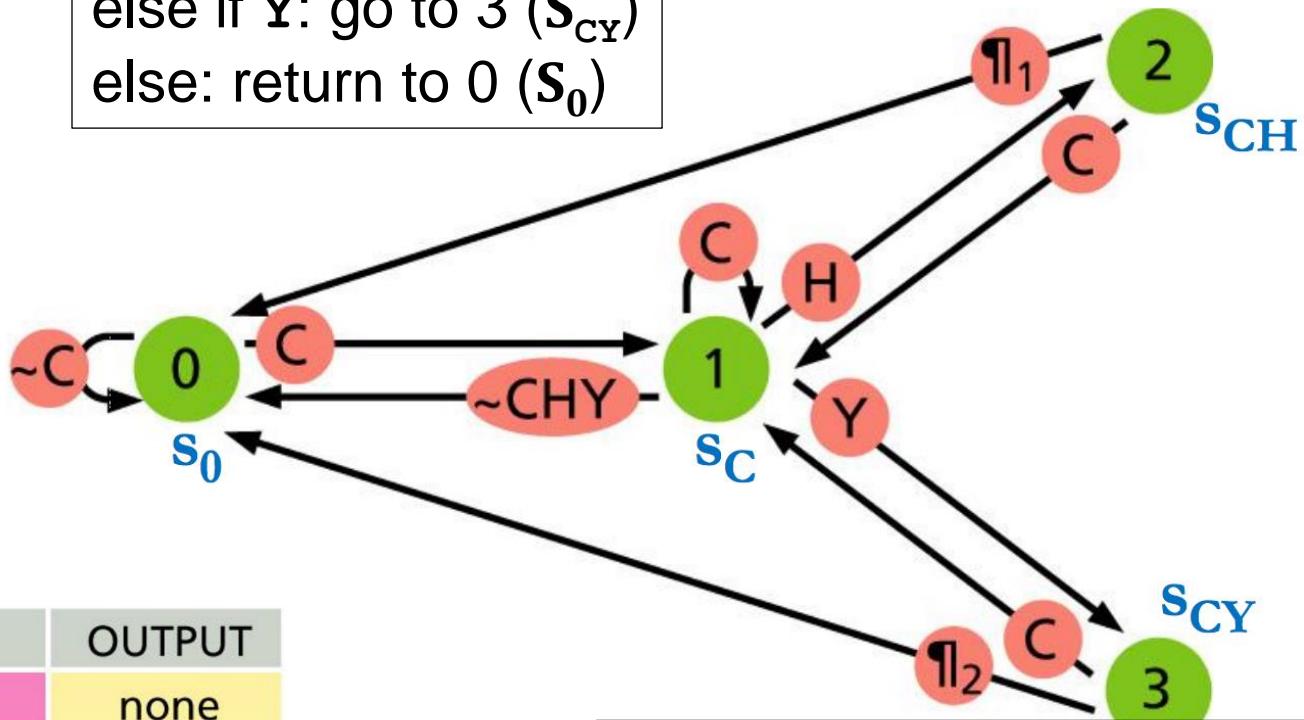
- CHH → CHH:  $9 + 8 + 8 = 25$
  - CHH → CHY:  $9 + 8 + 2 = 19$
  - CHH → CYH:  $9 + 2 + 8 = 19$

## *Example*

# FSM

## Example

**State 0 ( $S_0$ ):**  
if C: go to 1 ( $S_c$ )  
else: stay here



anything but CHY

	INPUT	OUTPUT
$\#_1:$	~CHY	none
	H	CHH
	Y	CHY
$\#_2:$	~CH	none
	H	CYH

anything but CH

**State 1 ( $S_c$ ):**  
if C: stay here  
else if H: go to 2 ( $S_{CH}$ )  
else if Y: go to 3 ( $S_{CY}$ )  
else: return to 0 ( $S_0$ )

**State 2 ( $S_{CH}$ ):**

if H: output CHH & return to 0 ( $S_0$ )  
else if Y: output CHY, return to 0 ( $S_0$ )  
else if C: return to 1 ( $S_c$ )  
else: return to 0 ( $S_0$ )

**State 3 ( $S_{CY}$ ):**

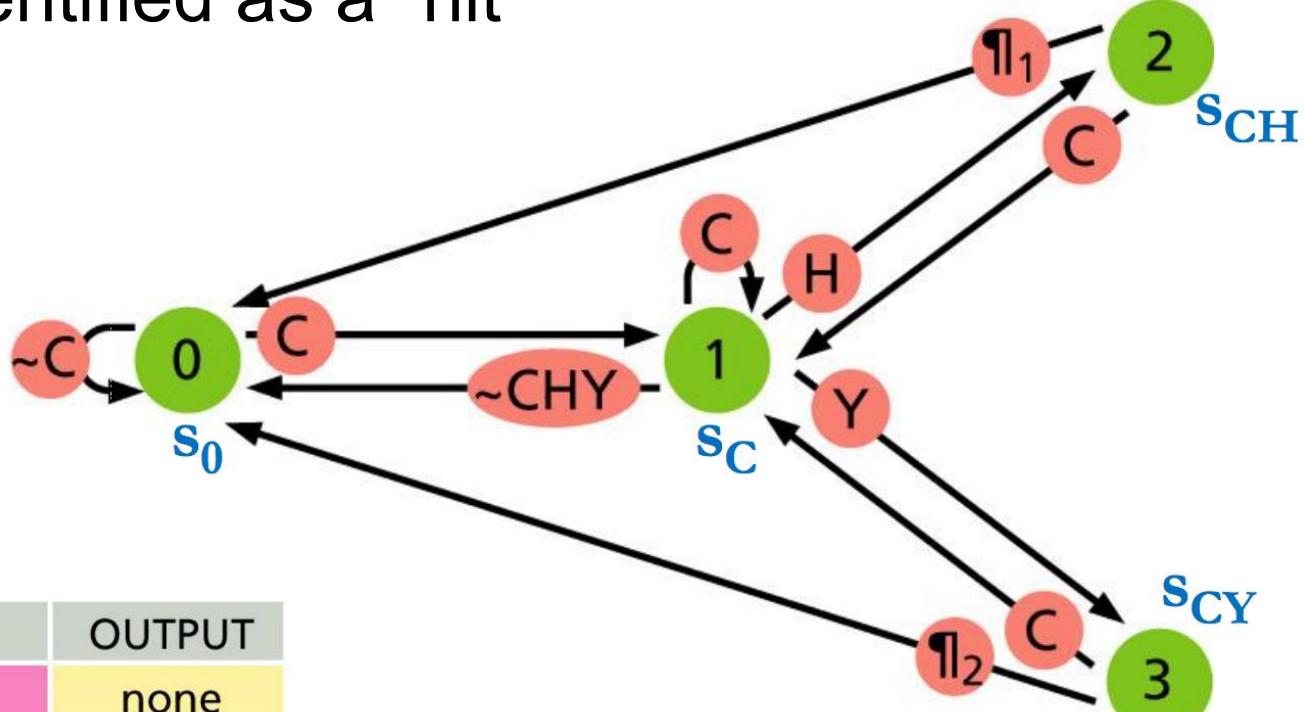
if H: output CYH & return to 0 ( $S_0$ )  
else if C: return to 1 ( $S_c$ )  
else: return to 0 ( $S_0$ )

# FSM

## Example

e.g. a (subject) sequence from the database

If the input is **CHCYHC**, the states visited are 0-1-2-1-3-0-1, with CYH identified as a “hit”



anything but CHY

INPUT	OUTPUT
$\text{I}_1:$ ~CHY	none
H	CHH
Y	CHY
$\text{I}_2:$ ~CH	none
H	CYH

anything but CH

**State 3 ( $S_{CY}$ ):**  
if **H**: output **CYH** & return to 0 ( $S_0$ )  
else if **C**: return to 1 ( $S_c$ )  
else: return to 0 ( $S_0$ )

# BLAST algorithm

**Phase 2:** identify  $k$ -mers in query with scores above threshold  $T$  for seeding

Query	Example							
LSNETDKRPPFIETAERL <b>RDQ</b> HKKDY.....	L	S	N	E	T	D	K	R
LSN	LSN	SNE	NET	ETD				
					<b>FSM</b>			
<b>RDQ</b> 16	QDQ 12	EDQ 11	RDN 11	<b>RDB</b> 11	BDQ 10	RDP 10		
RBQ 14	REQ 12	HDD 11	RDD 11	ADQ 10	XDQ 10	RDT 10		
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10		
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10		
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9		
...								

This list represents  $k$ -mers (in any target/database sequence) that would give a score of  $T$  or higher when aligned with the query sequence; they are used for **seeding**

# BLAST algorithm

**Phase 3:** *identify matches of these k-mers in each database sequence*

For each target sequence in the database:

- scan for hits with the compiled list of *k*-mers in Phase 2
- extend the hits to form **high-scoring segment pairs (HSPs)**
- Extension proceeds in **both** directions from the *k*-mer until the score drops by more than *X* relative to the current best score
- identify the highest scoring pair, i.e. maximal segment pair (MSP); if needed, combine two or more HSPs into a longer alignment
- final alignment score, **S**

# BLAST algorithm

Query: GSVEDTTGSQSLAALLNKCKT**PQG**QRLVNQWIKQPLMDKNRIEERLNLVAFVDAELRQTLQEDL

**PQG 18**

PEG 15

PRG 14

PKG 14

PNG 13

PDG 13

PHG 13

**T = 13**

**PMG 13**

PSG 13

---

PQA 12

PQN 12

etc.

extension proceeds in both directions

Query: 325 SLAALLNKCKT**PQG**QRLVNQWIKQPLMDKNRIEERLNVEA 365  
+LA++L+ TP G R++ +W+ P+ D + ER + A  
Sbjct: 290 TLASVLDCTVT**PMG**SRMLKRWLHMPVRDTRVLLERQQTIGA 330

**High-scoring segment pair (HSP)**

Example

# BLAST algorithm

**Phase 4:** evaluate the statistical significance of the alignments

## Expect Value (*E*-value)

- the number of alignments with a score **equal to or greater than** the observed score, that would be expected by chance alone in searching a database of n sequences
- *E*-value = 1: we would expect to see 1 match with the observed score or higher simply by chance
- the smaller an *E*-value, the more “significant” the match is
- dependent on the size of the database