

Biological Databases: Theories and Practice

430032

Course Introduction

Instructor: Prof. LEE, Tzong-Yi (李宗夷)
Email: leetzongyi@nycu.edu.ctw

*Professor
Institute of Bioinformatics and Systems Biology,
National Yang Ming Chiao Tung University*

Lecture outline



1. Course information
2. Assessment
 - Assignments + Midterm Exam + Final Project
3. Lecture schedule
 - Topics to be covered in this course
4. Introduction to...
 - a. What is Data, Information, and Knowledge?
 - b. How to obtain Knowledge from Data?
 - c. Why do we need biological databases?
5. Brief Introduction of Biological Databases

Course Instructor and TA

Instructor



Prof. Lee, Tzong-Yi (李宗夷)

Email: leetzongyi@nycu.edu.tw

Office: 賢齊館317室, 博愛校區

Phone: 03-5712121 #56947

Teaching Assistant



Mr. Chiu, Yen-Peng (邱彥鵬)

Email: SilverGojo4@gmail.com

Office: 賢齊館421室, 博愛校區

Phone: 03-5712121 #56948

Contact Information



李宗夷 教授

(Prof. Tzong-Yi Lee)

Email: leetzongyi@nycu.edu.tw

Office: 賢齊館 BIO317

Lab: 生物大數據與深度計算實驗室

<https://sites.google.com/view/biomics>

Office Hour: basically Wed. afternoon
(actually anytime)

Line ID: **francislee0215**

The **Line Group** of this course:

生物資料庫理論與實作2023

Database

A database is an organized
collection of related
biological data, that can

生物資料庫理論與實作2023

Education Background and Academic Experiences

Prof. Tzong-Yi Lee



李宗夷 (1980. 02. 15)

B.S., 1998 ~ 2002
M.S., 2002 ~ 2004



Department of Computer
Science and Information
Engineering, National
Central University



國立交通大學
生物資訊所
Published 10 SCI papers

生物大數據與深度計算實驗室
生物大數據分析
多體學深度學習計算
空間轉錄體
病原菌抗藥性
抗菌肽設計
智慧醫療

有庠傑出教授: 2015 ~
教授: 2015 ~ 2018
副教授: 2012 ~ 2015
助理教授: 2009 ~ 2012



元智大學
資訊工程學系

2016: YZU研究傑出獎
2014: 科技部百人拓荒計畫
2012: YZU研究傑出獎
2011: YZU青年學者研究獎

終身聘副教授: 2021 ~ 2022
生物資訊主任: 2019 ~ 2022
珠江人才: 2019 ~ 2022
校長青年學者: 2019 ~ 2022
副教授: 2018 ~ 2022



香港中文大學(深圳)
生命與健康科學學院

玉山青年學者: 2023.08 ~
教授: 2023.02 ~



國立陽明交通大學
生物資訊及系統生物研究所

學術表現簡介

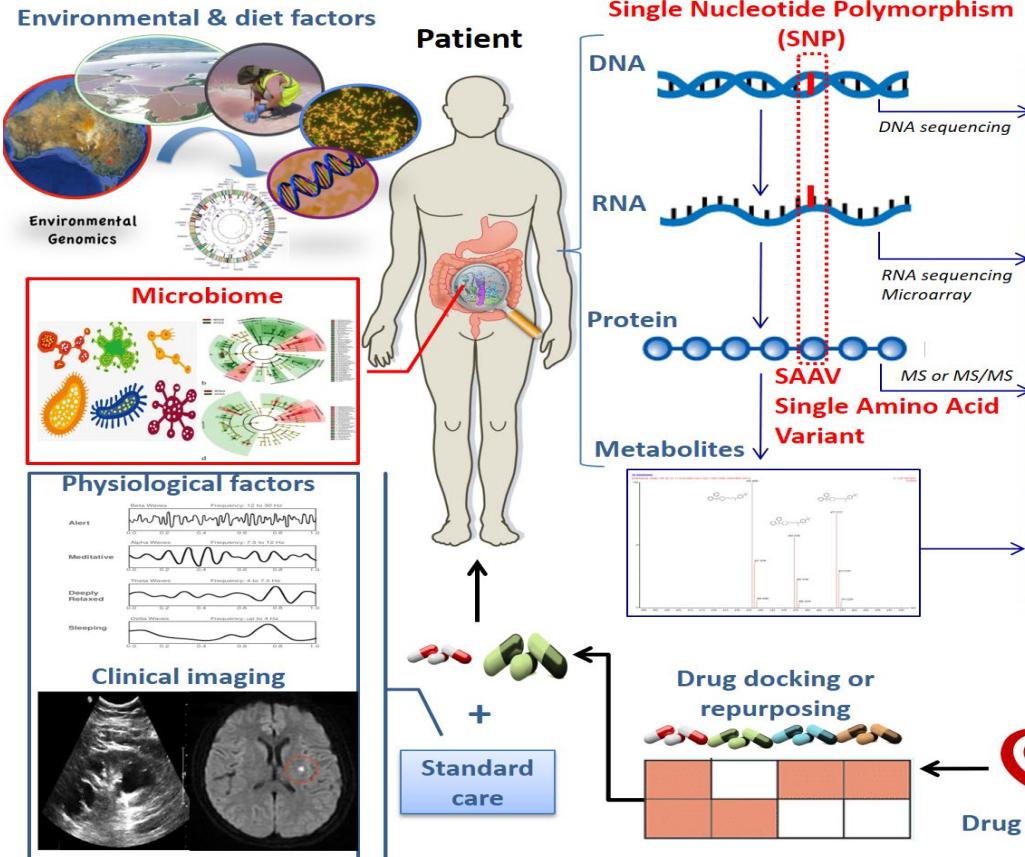
已發表超過**130**篇 SCI期刊論文(**其中第一或通訊作者文章共有98篇**)，包含**Nucleic Acids Research** (SCI IF:19.160) 19篇、**Briefings in Bioinformatics**(SCI IF:13.994) 9篇、**PNAS** (SCI IF:12.777) 1篇、**Cell Reports**(SCI IF:9.995) 1篇、**Microbiology Spectrum** (SCI IF:9.043) 2篇等。論文總引用次數超過**7600**次，H指數為**44**。於**2022年獲選為全球前2%學者 (World's Top 2% Scientists in Bioinformatics)**。目前**有4篇高被引文章**。

生物大數據與深度計算實驗室 (BiOmics Lab)

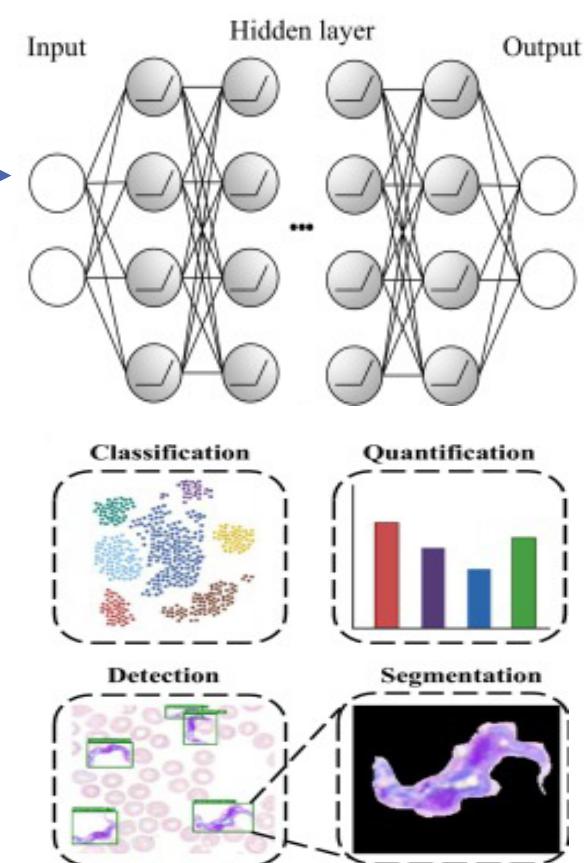
<https://sites.google.com/view/biomics>

• Big Data Analytics and Deep Computation of Systems Biology

Big Data Analytics



Deep Computation



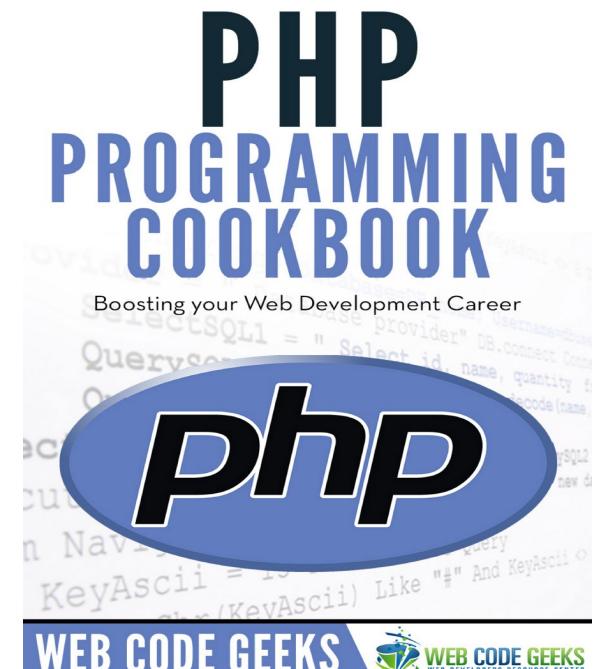
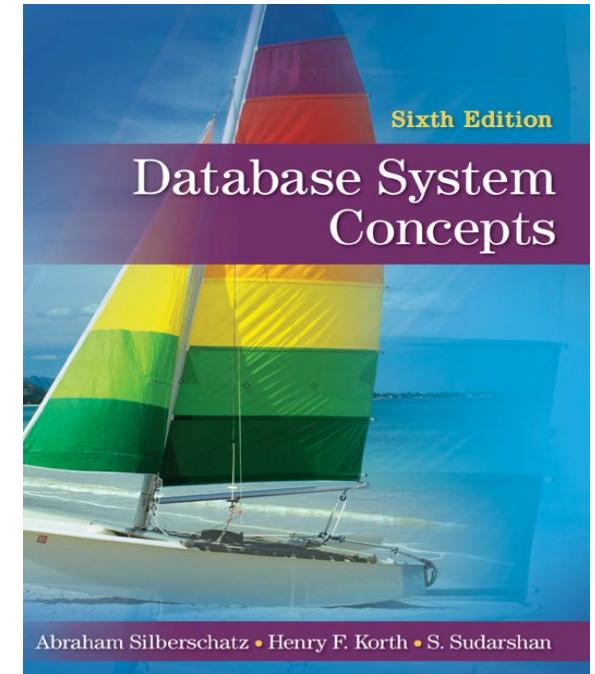
Course Objectives



- This course focuses on the design and implementation of biological **database systems**, it also discusses contemporary concepts and approaches about **web interface design** for database access and data management.
- Topics such as database management system (DBMS), structured query language (SQL), database design, PHP web programming for accessing database, web interface design, and big data analytics for biomedical applications are covered.
- On the practical side, students are given exercises to accumulate **hands-on experiences** on applying the learned concepts and well-developed tools to develop database system on biological data.

Textbooks

- The three books are available from E3 system
 - Abraham Silberschatz, Henry F. Korth, S. Sudarshan, “*Database System Concepts*”, Tata McGraw - Hill, 6th edition, 2011
 - WCGs (Web Code Geeks), “*PHP Programming Cookbook*”
 - Kai Hwang and Min Chen, “*Big-Data Analytics for Cloud, IoT and Cognitive Computing*”, WILEY, First edition.



Assessment

- **Homework Assignments (20%)**
 - 2 assignments (each for 10% of total credit)
- **Midterm exam - written examination (30%)**
 - Tentatively on 8th Week (Oct. 31st)
- **Final Project (50%)**
 - Oral presentation (15th and 16th weeks)
 - Project report (upload in 17th week)

週次	上課日期	課程進度、內容、主題
1	2023-09-12(二)	Course Introduction and Biological Databases
2	2023-09-19(二)	Database System Architectures and Relational Model
3	2023-09-26(二)	Introduction to SQL I
4	2023-10-03(二)	SQL Practice (Self-study)
5	2023-10-10(二)	Natioanl Day Holiday
6	2023-10-17(二)	Introduction to SQL II
7	2023-10-24(二)	Database Schema and Entity-Relationship Model
8	2023-10-31(二)	Midterm Exam
9	2023-11-07(二)	Application Design and Database Connection
10	2023-11-14(二)	PHP Programming and Web Interface Design I
11	2023-11-21(二)	PHP Programming and Web Interface Design II
12	2023-11-28(二)	Introduction of Biological Databases
13	2023-12-05(二)	Database Design and Normalizations
14	2023-12-12(二)	Data Warehousing and Knowledge Mining
15	2023-12-19(二)	Final Project Presentation
16	2023-12-26(二)	Final Project Presentation

Final Project

- Each group has 2 – 3 students to work together as a team to work on a project with a topic of interest
 - Must be a topic of general interest and biological importance
- Each project should have:
 - Construction of a database system (e.g. MySQL, DB2,)
 - Relational schema of database (e.g. E-R model) primary key
 - Web interface (e.g. PHP) of accessing and managing the content of databases
 - The involved data mining and statistical analyses (**bonus grading**)API, search (functional)

combine with the function of plotting
feature analysis

Project report (Within 10 A4 pages)



- Introduction to the project topic
- Key problems involved
- Motivation and goals
- Materials and data source
- Design of database schema
- Methods (involved data mining and statistical approaches)
- Results (web interface, developed functions, case study, etc.)
- Prospects

Learning outcomes



- Before Midterm Exam
 - Understand the fundamentals of database system architectures
 - Apply Structured Query Language (SQL) on database establishment
- After Midterm Exam
 - Learn PHP web programming and database connection
 - Learn database schema and database normalization

Promises and Expectations



- **Promises**

- Lecture content, teaching pace and projects tailored according to your interests
- Putting up lecture slides in time
- Quick response to emails
- Step-by-step guidance for MySQL practice and Web programming
- Prompt and fair grading of project deliverables, associated with the final project report

- **Expectations**

- Attending lectures, punctuality
- *Active class participation*
- **Upload assignments / report in time**

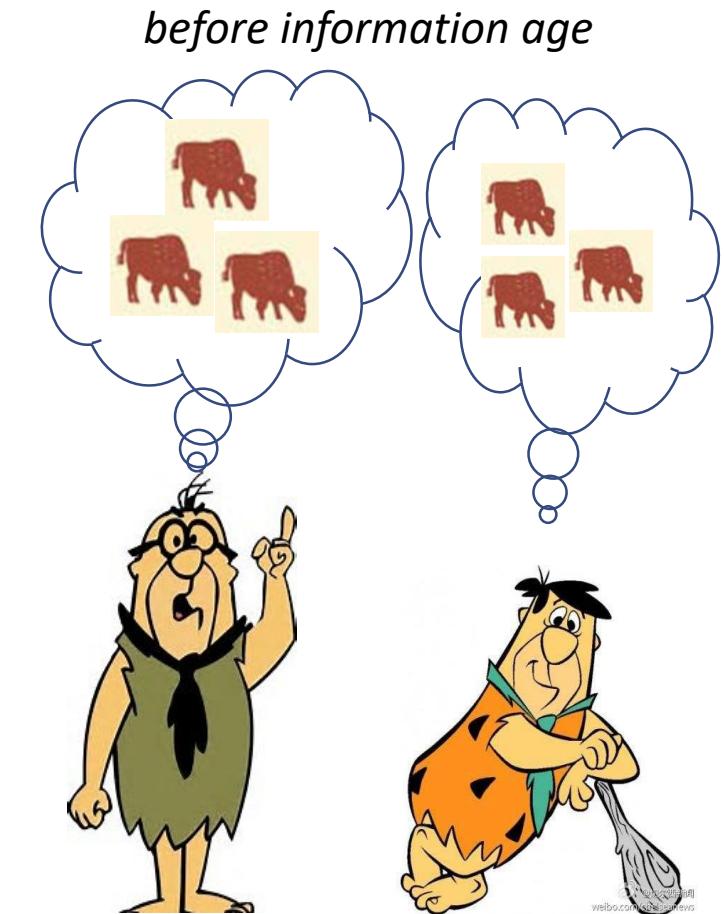
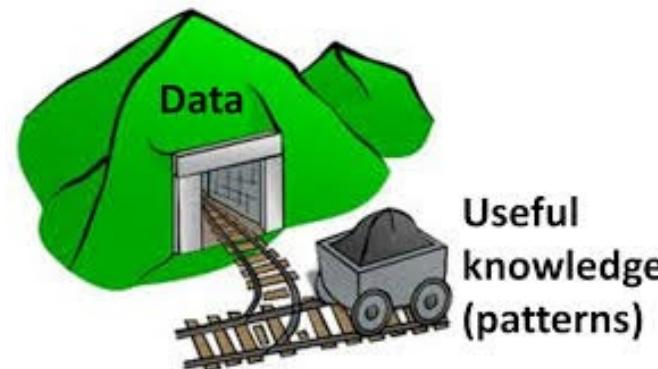
Why do we need Data?

- Data are conceptual raw facts that can be generated, stored and further **refined as information** within human brain.
- Data and information can be delivered with voice and gestures, and it also can be painted, written, recorded and duplicated on paper or other lightweight carrier.



American Buffalo (Bison).

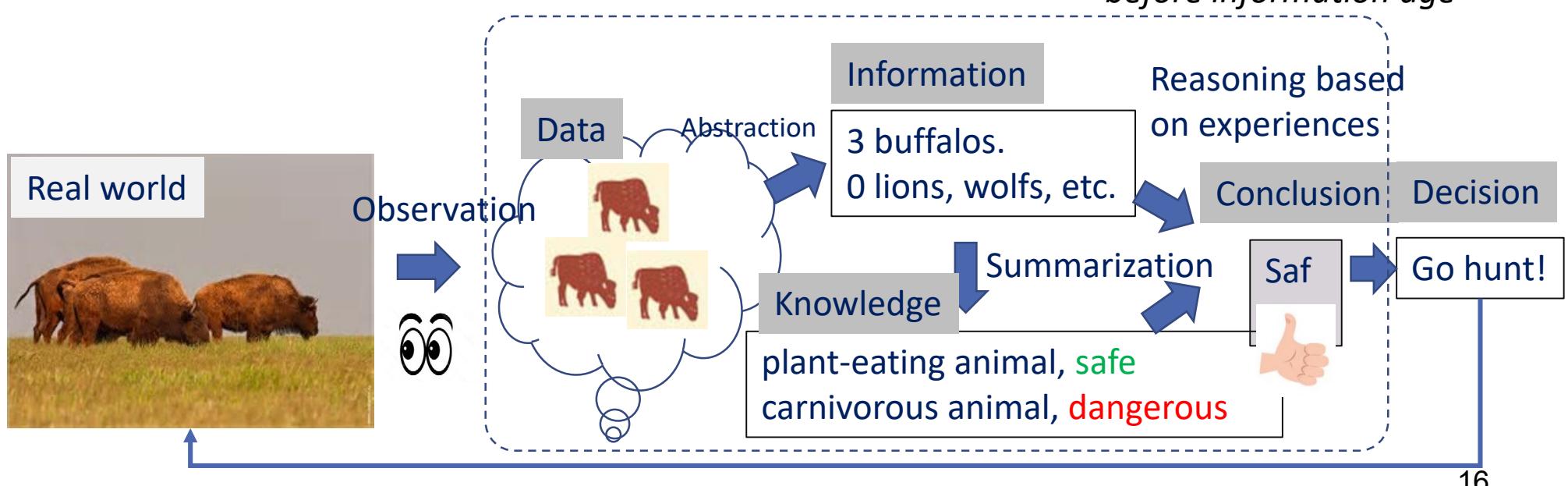
http://xihuashe.lofter.com/post/103b29_13f6566



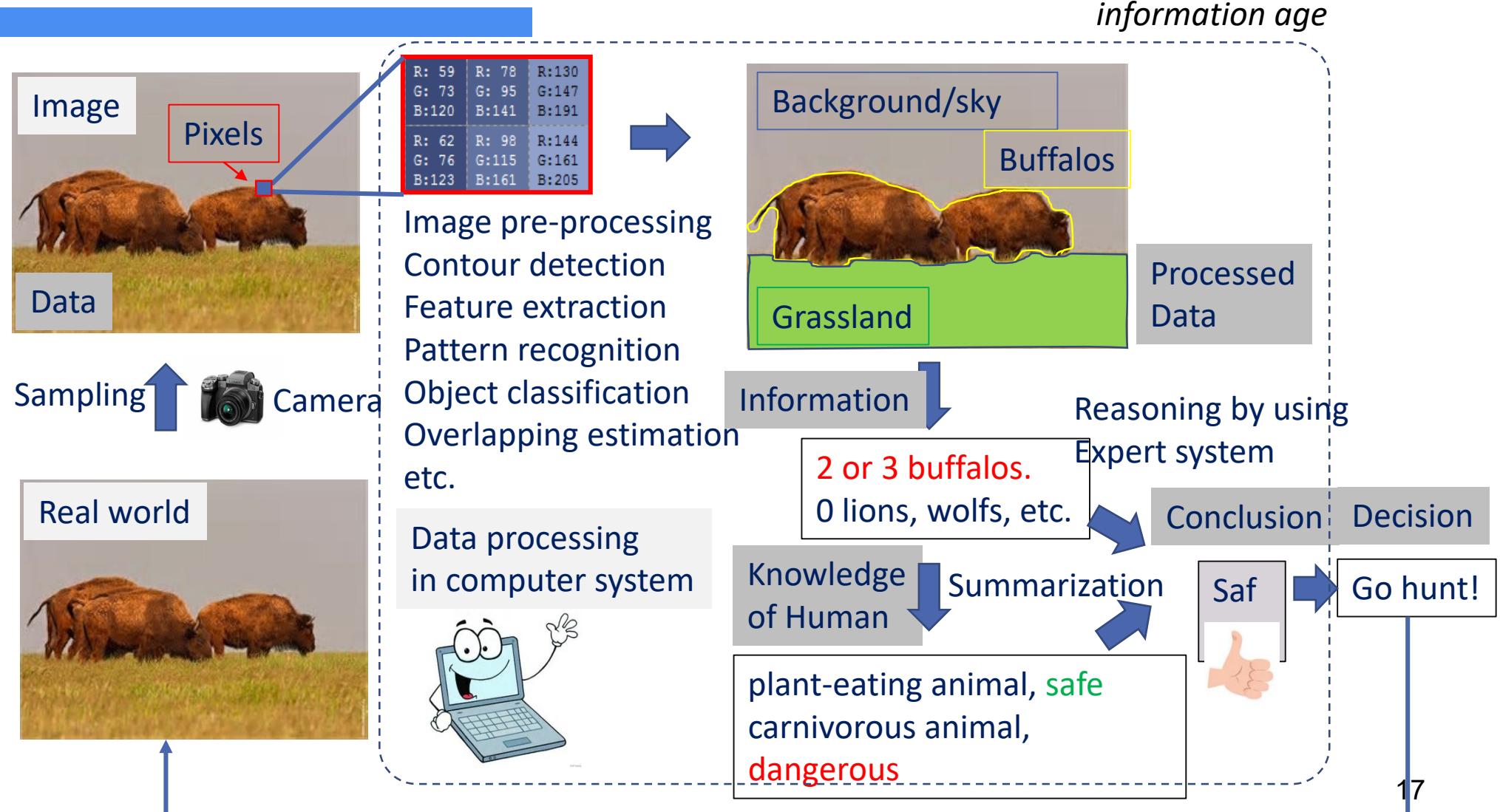
Cartoon characters of The Flintstones.
<http://www.nipic.com>

Data, Information and Knowledge

- Data: observation records of facts, based on knowledge.
- Information: accurate, relevant, and timely description of facts.
- Knowledge: a theoretical or practical understanding of a subject.
 - Mathematics, Biology, Physics, Chemistry, etc..

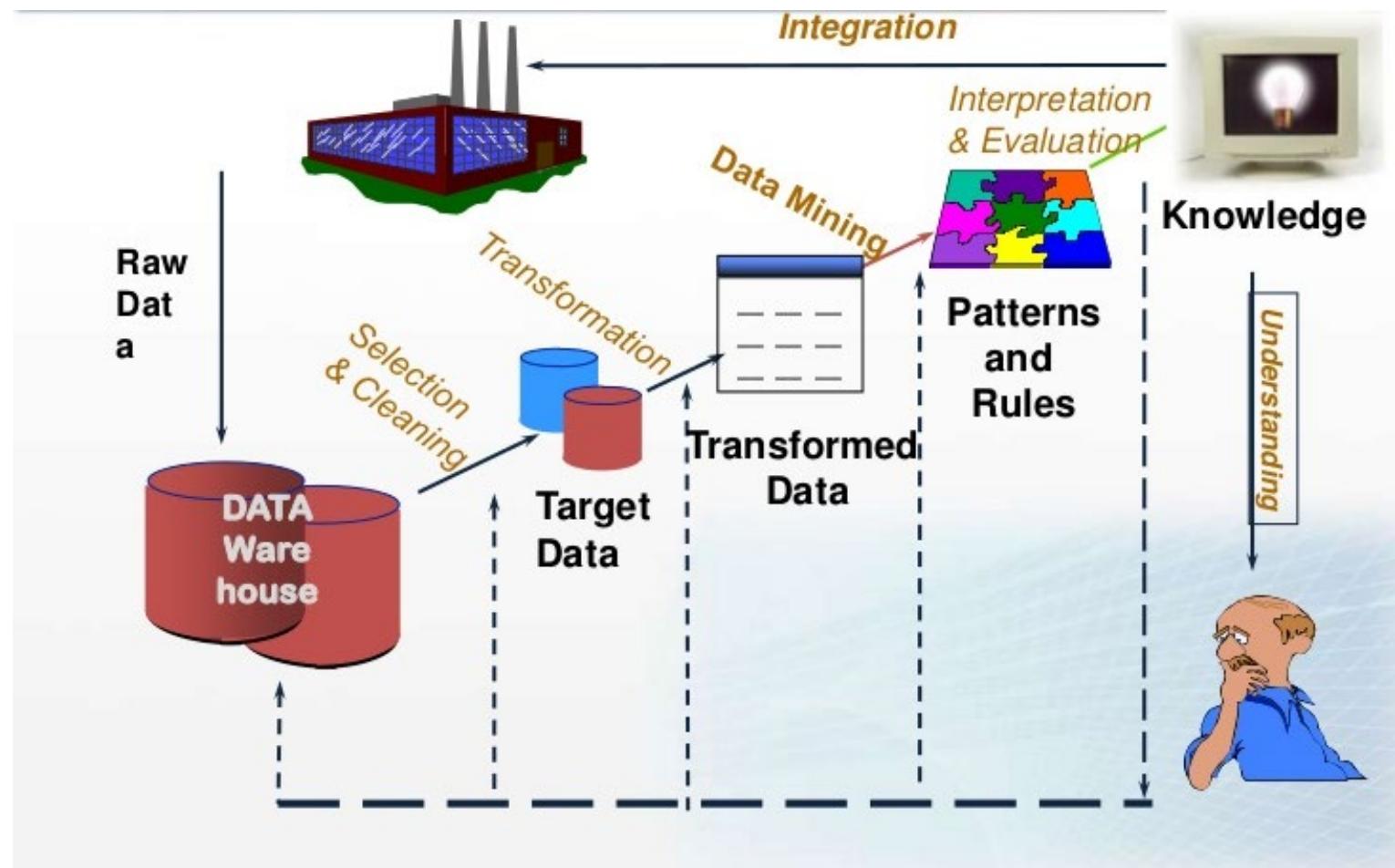


Data, Information and Knowledge



Knowledge Mining: from data to knowledge

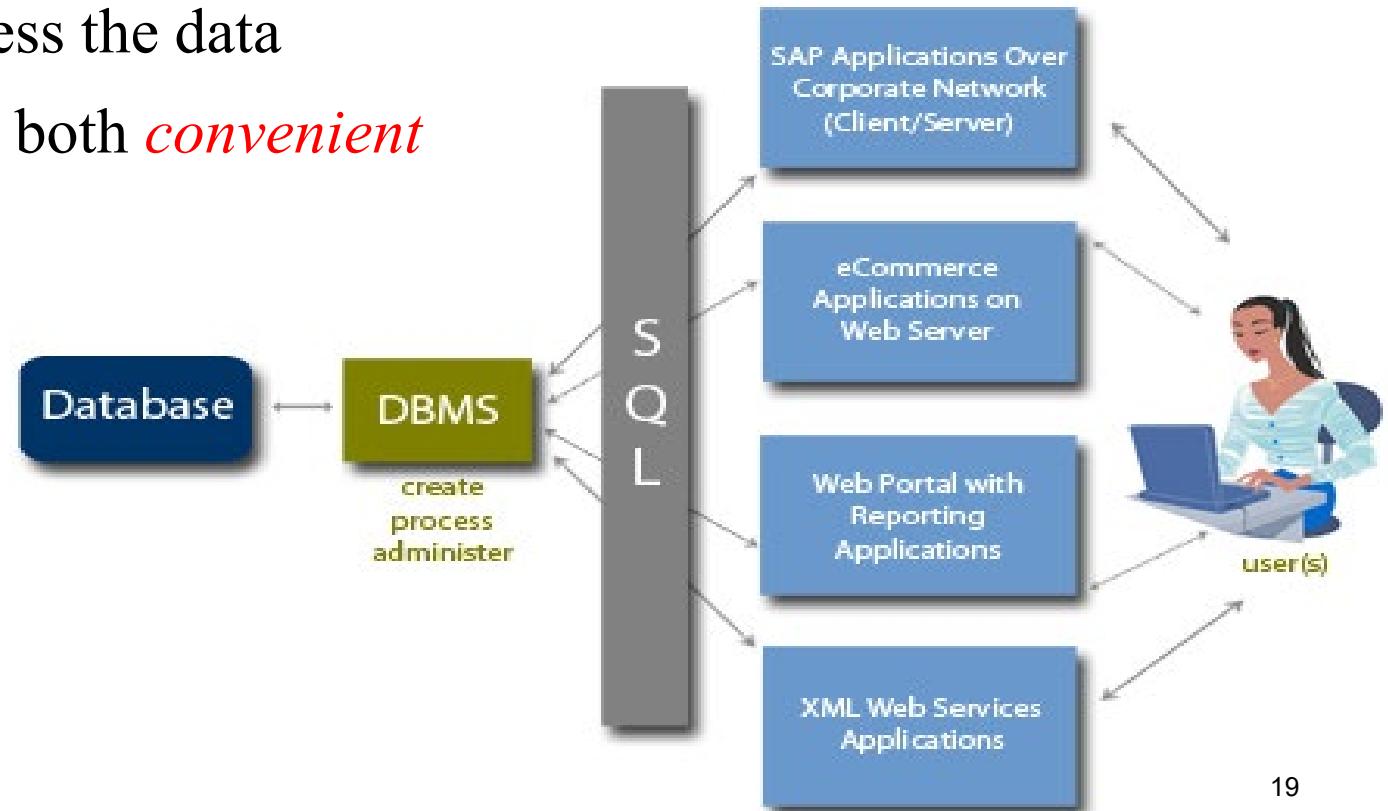
- Knowledge mining is the automatic extraction of non-obvious, hidden knowledge from large volumes of data.
 - Selection
 - Transformation
 - Pattern recognition



Database Management System (DBMS)

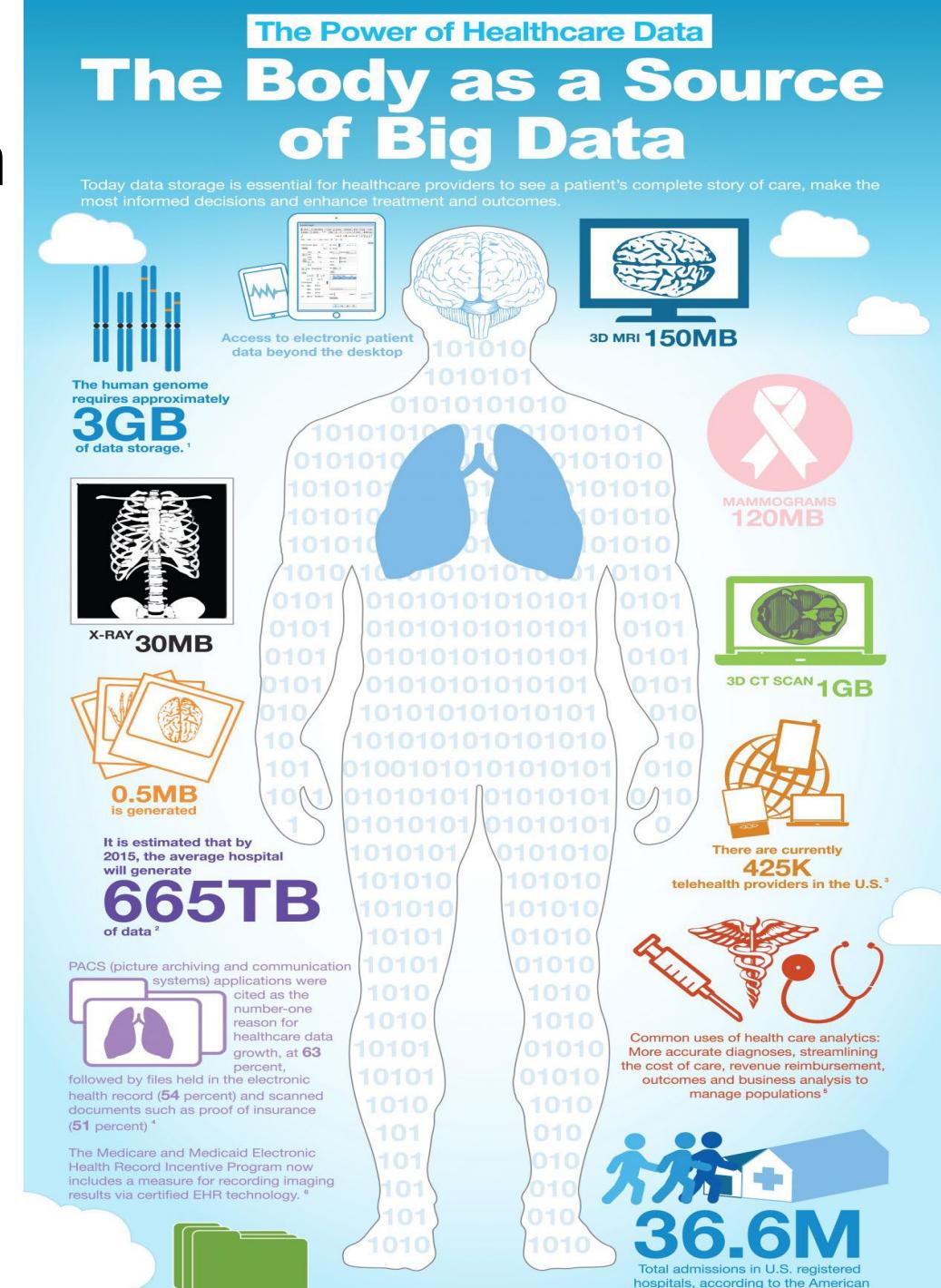
- ▶ DBMS contains information about a particular enterprise

- Collection of interrelated data
- Set of programs to access the data
- An environment that is both *convenient* and *efficient* to use

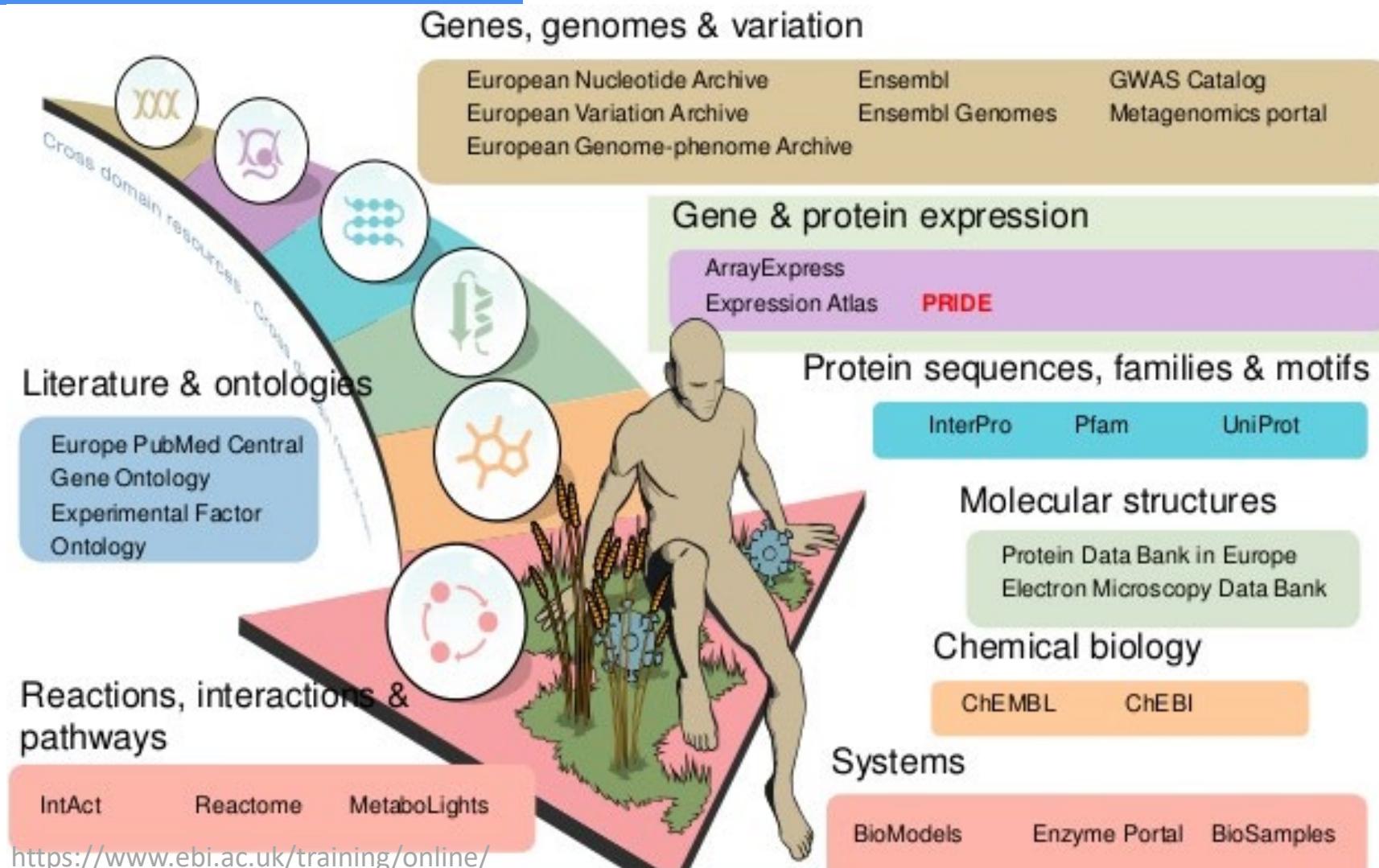


Why Need Database System

- **Large data size!!**
- Each adult human has 10^{13} - 10^{14} cells
- Most of them contain two copies of DNA with 3×10^9 nucleotides (each is called a “haploid genome”)
- Humans have 20,000-25,000 genes that produce proteins
- High-throughput sequencing technology



We are in the Data-driven World



We need database, even for evaluating a professor ...



Tzong-Yi Lee (李宗夷)

FOLLOW

[GET MY OWN PROFILE](#)

Professor, National Yang Ming Chiao Tung University
Verified email at nycu.edu.tw - [Homepage](#)

Bioinformatics Multi-Omics Data Analysis Systems Biology Machine Learning and Deep...
Precision Medicine

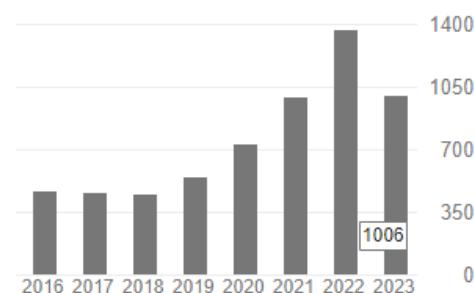
TITLE	CITED BY	YEAR
miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database HY Huang, YCD Lin, J Li, KY Huang, S Shrestha, HC Hong, Y Tang, ... Nucleic acids research 48 (D1), D148-D154	1017	2020
KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns YH Wong, TY Lee, HK Liang, CM Huang, TY Wang, YH Yang, CH Chu, ... Nucleic acids research 35 (suppl_2), W588-W594	402	2007
KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites HD Huang, TY Lee, SW Tzeng, JT Horng Nucleic acids research 33 (suppl_2), W226-W229	397	2005
PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants CN Chow, HQ Zheng, NY Wu, CH Chien, HD Huang, TY Lee, ... Nucleic acids research 44 (D1), D1154-D1160	350	2016
PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups WC Chang, TY Lee, HD Huang, HY Huang, RL Pan BMC genomics 9 (1), 1-14	332	2008
PlantPAN3. 0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants CN Chow, TY Lee, YC Hung, GZ Li, KC Tseng, YH Liu, PL Kuo, HQ Zheng, ... Nucleic acids research 47 (D1), D1155-D1163	290	2019

[GET MY OWN PROFILE](#)

Cited by [VIEW ALL](#)

All Since 2018

Citations	7622	5098
h-index	44	36
i10-index	96	90



Public access [VIEW ALL](#)

1 article	19 articles
-----------	-------------

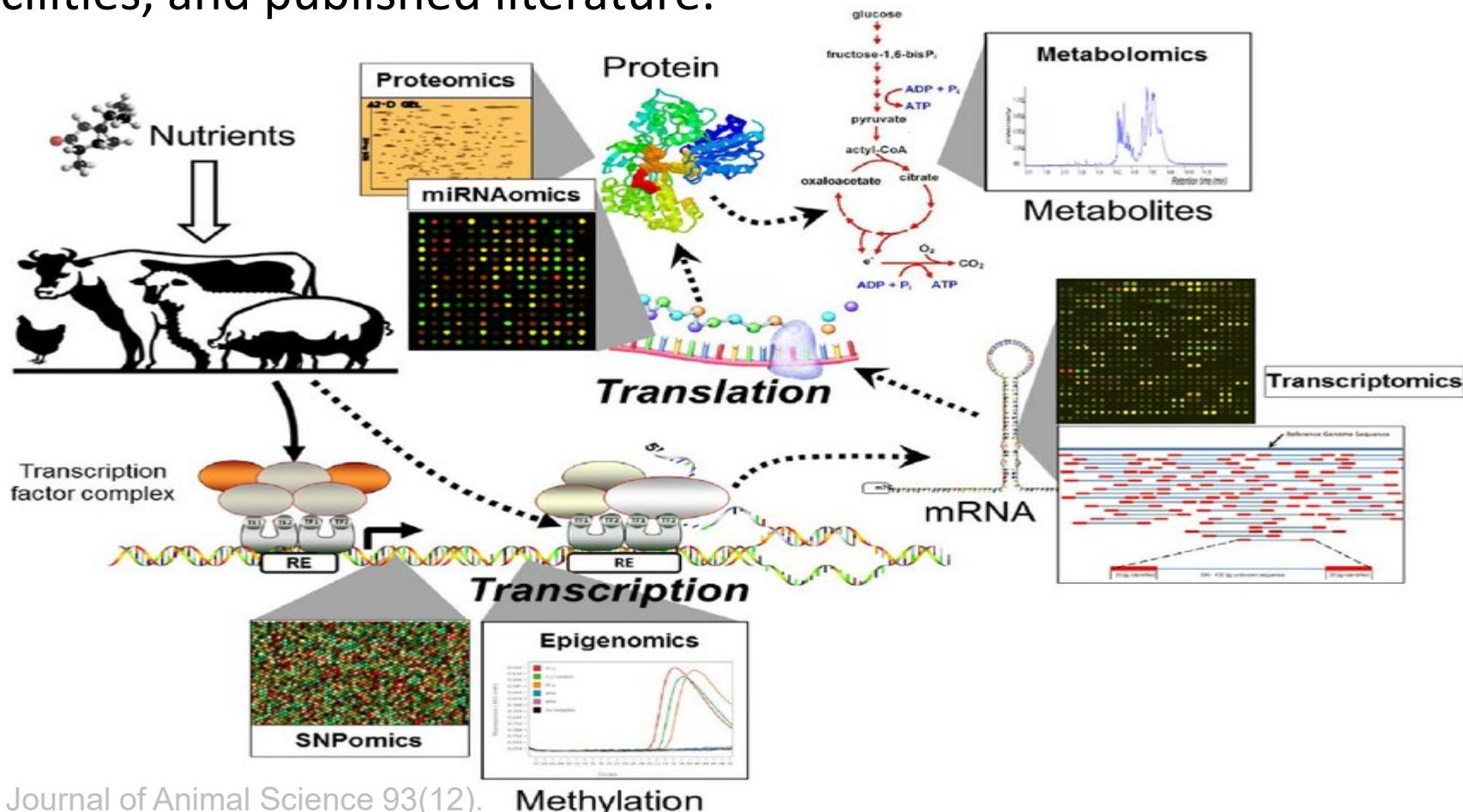
not available	available
---------------	-----------

Based on funding mandates

Co-authors [VIEW ALL](#)

What is Biological Data?

- The raw data collected from scientific experiments, high-throughput facilities, and published literature.



Biological text data



Briefings in Bioinformatics, 2022, 23(6), 1–11

<https://doi.org/10.1093/bib/bbac409>

Advance access publication date 24 September 2022

Problem Solving Protocol

BioGPT: generative pre-trained transformer for biomedical text generation and mining

Renqian Luo , Liai Sun, Yingce Xia , Tao Qin , Sheng Zhang , Hoifung Poon and Tie-Yan Liu

Corresponding authors: Tao Qin, Microsoft Research AI4Science, Beijing, China, E-mail: taoqin@microsoft.com; Renqian Luo, Microsoft Research AI4Science, Beijing, China, E-mail: renqianluo@microsoft.com; Yingce Xia, Microsoft Research AI4Science, Beijing, China, E-mail: yinxia@microsoft.com

Abstract

Pre-trained language models have attracted increasing attention in the biomedical domain, inspired by their great success in the general natural language domain. Among the two main branches of pre-trained language models in the general language domain, i.e. BERT (and its variants) and GPT (and its variants), the first one has been extensively studied in the biomedical domain, such as BioBERT and PubMedBERT. While they have achieved great success on a variety of discriminative downstream biomedical tasks, the lack of generation ability constrains their application scope. In this paper, we propose BioGPT, a domain-specific generative Transformer language model pre-trained on large-scale biomedical literature. We evaluate BioGPT on six biomedical natural language processing tasks and demonstrate that our model outperforms previous models on most tasks. Especially, we get 44.98%, 38.42% and 40.76% F1 score on BSCDR, KD-DT1 and DDI end-to-end relation extraction tasks, respectively, and 78.2% accuracy on PubMedQA, creating a new record. Our case study on text generation further demonstrates the advantages of BioGPT on biomedical literature to generate fluent descriptions for biomedical terms.

Keywords: biomedical literature, generative pre-trained language model, text generation, text mining

Introduction

Text mining and knowledge discovery from biomedical literature play important roles in drug discovery, clinical pathology, pathology research, etc. Typical tasks include recognizing named entities in the articles, mining the interaction between drugs and proteins/diseases/other drugs, answering questions given reference text, generating abstracts for given phrases/words, etc. People have accumulated large amounts of literature in the previous studies. For example, PubMed (<https://pubmed.ncbi.nlm.nih.gov/>), one of the most popular biomedical search engines, covers more than 30M articles and the number still rapidly increases every day as new discoveries are continuously coming out. Therefore, automatically mining the knowledge from literature becomes an urgent demand.

Pre-training models have demonstrated their powerful capability in natural language processing (NLP). On the GLUE benchmark, a widely used benchmark for natural language understanding, pre-training methods outperform non-pre-training methods by a large margin [1] (<https://gluebenchmark.com/leaderboard>). There are two main kinds of pre-training models: (1) the BERT-like models [2–4], mainly for language

understanding tasks; (2) the GPT-like models [5–7], mainly for language generation tasks.

These models are first pre-trained on large-scale corpora collected from the Web via self-supervised learning task (e.g. masked language modeling for BERT, auto-regressive language modeling for GPT), and then fine-tuned on specific downstream tasks. The BERT-like models are widely used in sequence classification and sequence labeling, where we need to encode the complete document. In comparison, the GPT-like models are often used in generation tasks (e.g. abstract generation, knowledge triplet generation).

By witnessing the success of pre-training in general NLP, people explore adapting these techniques into biomedical domain. However, directly applying these models to the biomedical domain leads to unsatisfactory performance due to domain shift [8, 9]. A natural solution is to develop pre-training models on biomedical texts (e.g. PubMed). BioBERT [10] and PubMedBERT [9]) are two representative BERT-like models pre-trained on biomedical domain, and they obtain superior performances than general pre-trained models on biomedical benchmarks. However, previous works mainly focus on BERT models which are more appropriate

Renqian Luo is a Researcher at Microsoft Research AI4Science. His research interests include machine learning and deep learning with applications to natural language processing and science.

Liai Sun is a graduate student at Peking University. Her research interests are natural language processing, deep learning and machine learning.

Yingce Xia is a Senior Researcher at Microsoft Research AI4Science. His research interests include deep learning, machine learning, natural language processing and drug discovery.

Tao Qin is a Senior Principal Researcher/Manager at Microsoft Research AI4Science. His research interests include deep learning, machine learning, reinforcement learning, and their applications to natural language processing, speech, computer vision, game and science.

Sheng Zhang is a Senior Researcher at Microsoft Research. His research focuses on natural language processing, semantic parsing and information extraction.

Hoifung Poon is a Senior Director at Microsoft Research. His research interests lie in advancing machine learning and NLP to overcome the knowledge and reasoning bottlenecks in precision medicine.

Tie-Yan Liu is a Distinguished Scientist of Microsoft, an Assistant Managing Director of Microsoft Research Asia and Microsoft Research AI4Science. He is a fellow of the IEEE, the ACM and the AAIA.

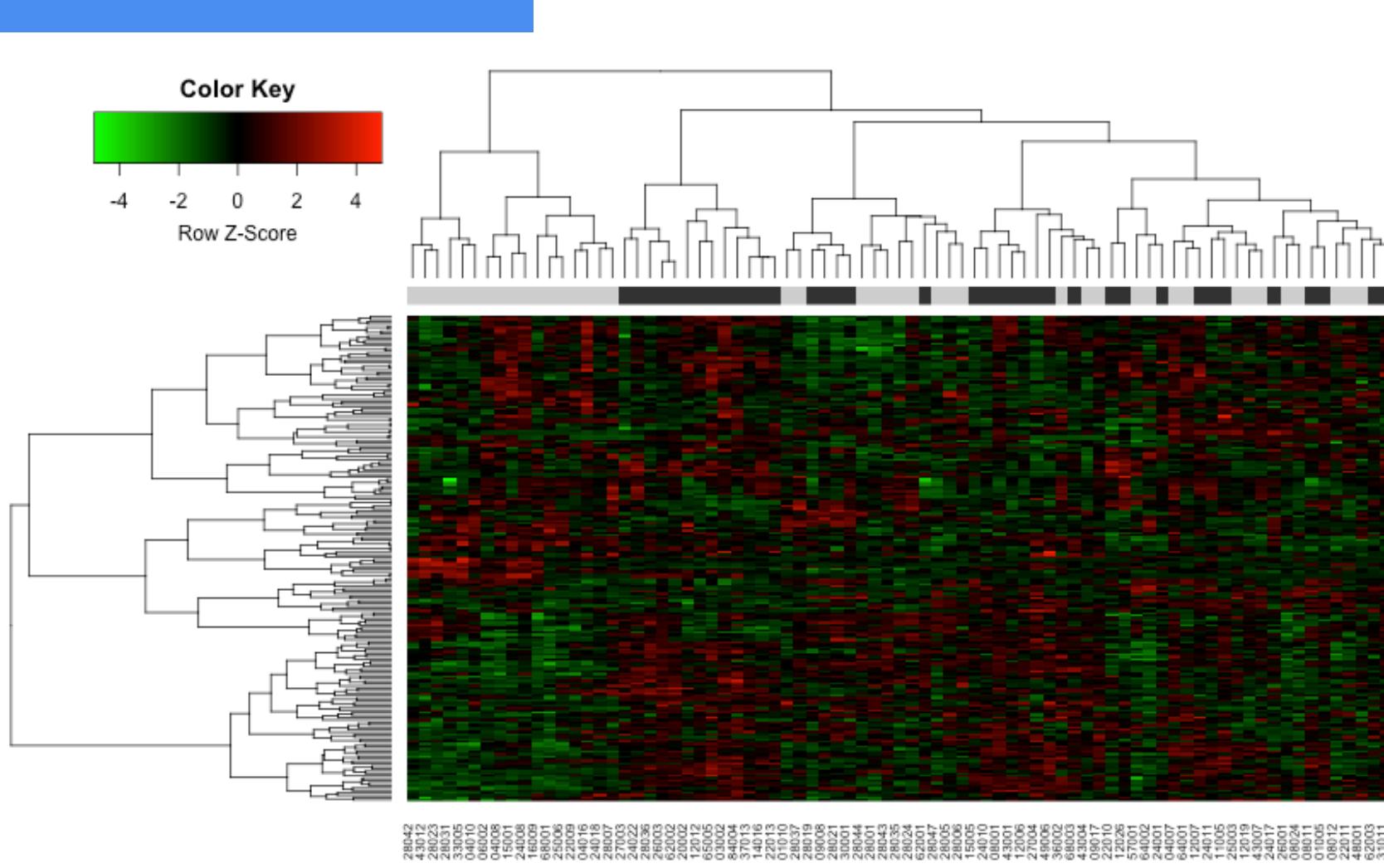
Received: June 16, 2022. Revised: August 5, 2022. Accepted: August 23, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Sequence data

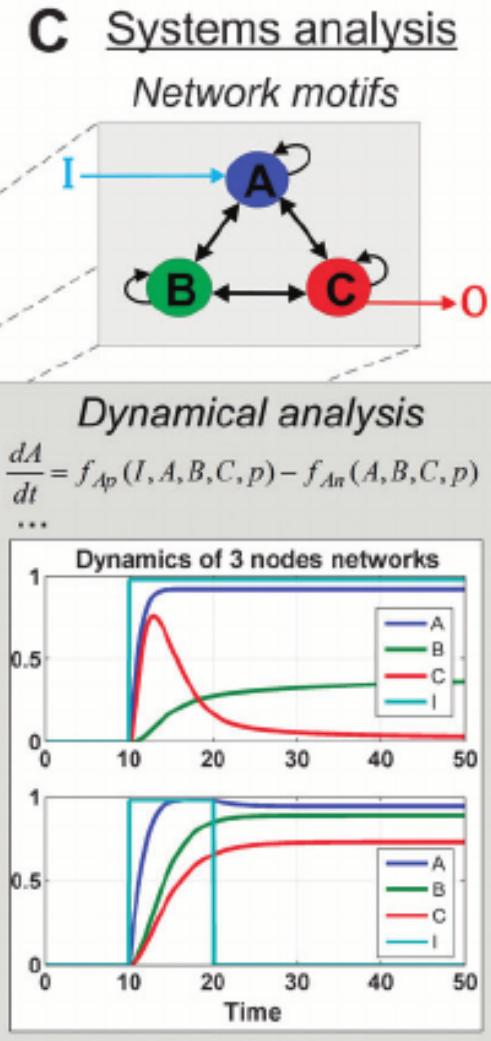
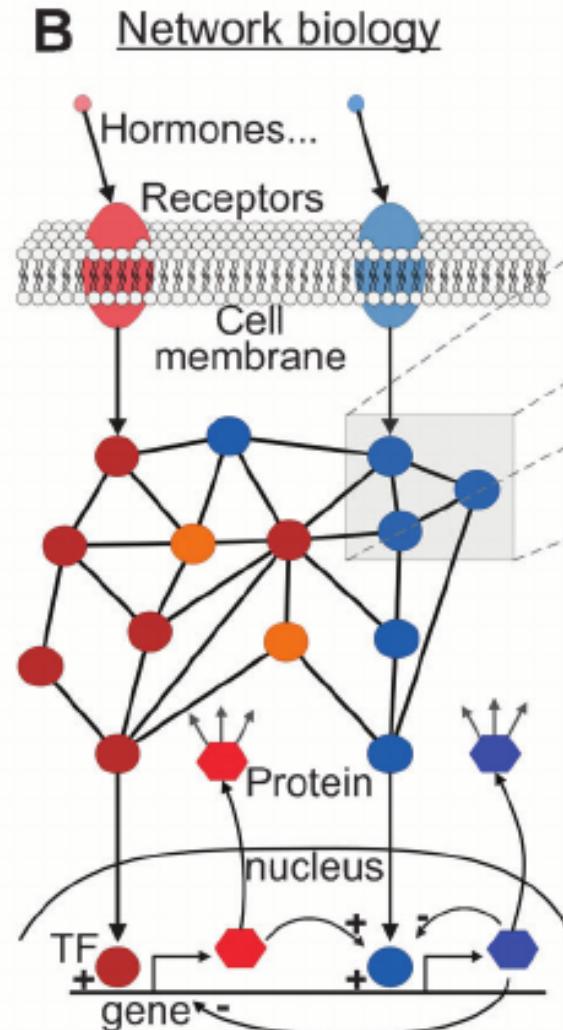
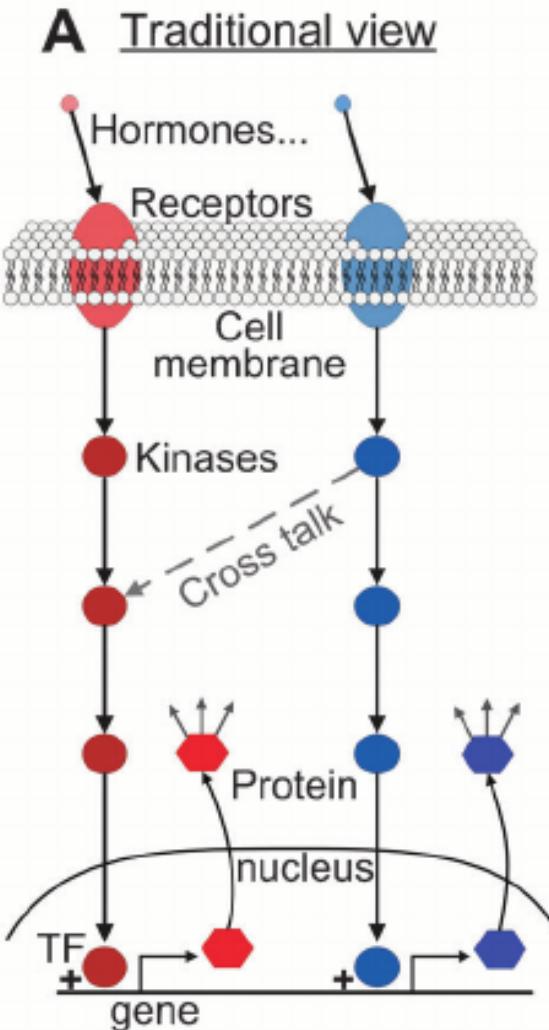
CTCACCGTCGAAATCTATGATCTGGCTTGGCCTGCAGT
AGCTCTTCATTTCGGGCTTATCTAATGCTGACTGGTCG
GTCCTGGCTACGCTCCAAAACGTACGTATTGGGCCATC
GAGGCTAGCGGCACTCGAGCGATCTATCGGGAGCTTG
GCTATCGATCGGGCGATCGATGCTGACGTACGTAGCGCG
CGATCGAGCGCGGCTAGCTAGCGGCATCGTAGCTACGTA
GCTACGGCGCTATTTCGATCGAGTCGTGTCTAGTCGGAT
ATAGCTATGCATCTAGCTGAGGCATCTGAGCGGATCGAT
GCTAGGGCGATCGGAGCTAGCTGAGCTAGCTAGCTGAGC
GCTAGCGAGCGTACGAGCGATCGAGCGAGTCTAGCGAGC
GATTCTAGCGATATACTAGCCCCGATCGTATGCTAGCT
AGGGCTAGCATGCGGATCTATCGAGCGGCTATCTGAGCG
ATTCGATCGAGCGATCTAGCGAGCTATCGATCGAGCCGG

Gene expression data



Biological network data

https://en.wikipedia.org/wiki/Biological_network



Protein structure data

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB 209,389 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

▼ 3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Advanced Search | Browse Annotations Help

PDB-101 wwPDB EMDDataResource NAKB wwPDB Foundation PDB-Dev

New: More Computed Structure Models (CSM) available Learn more

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

Explore NEW Features

PDB-101 Training Resources

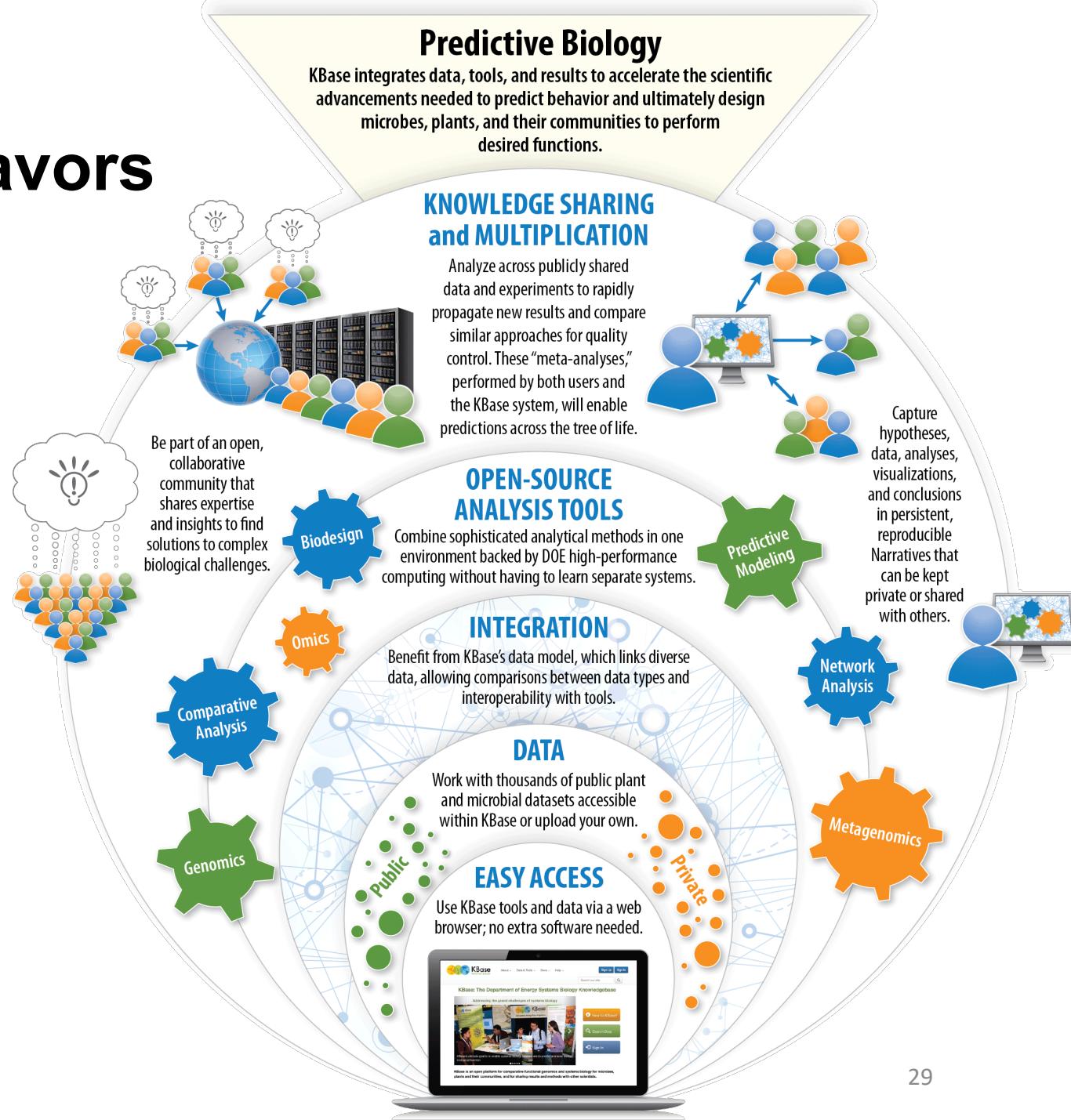
September Molecule of the Month

Histone Deacetylases



International endeavors

Consortium	Purpose
The Human Genome Project (HGP)	Sequence the human genome
The International HapMap Project	Develop haplotype map of the human genome
Encyclopedia of DNA Elements (ENCODE)	Catalog and characterize human DNA elements
Model Organism Encyclopedia of DNA Elements (modENCODE)	Catalog and characterize model organism DNA elements
1000 Genomes Project	Identify most genetic variants with at least 1% frequencies
The Cancer Genome Atlas (TCGA)	Build an atlas of genomic changes in cancer genomes
...	...



Enzyme Database

Contains data about structure and function of various enzymes **BRENDA**



6

Disease Database

Disease related information



7

OMIM

Chemical Database

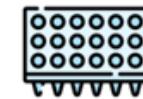
Data on several small organic molecules **PubChem**



8

Microarray Database

Gene expression data from microarray experiments



9

GEO

Taxonomic Database

Database that provides information on earth's species of animals, plants

Catalogue of life



10

Biological Database

A database is an organized collection of related biological data, that can be easily stored, accessed and managed

1



Contains article and research papers of different journals
Pubmed

2



Contains protein and nucleotide sequence **GenBank, DDBJ, PIR**

3



Contains 3D structure of proteins and nucleic acids **PDB**

4



Contains data about various biological pathways **KEGG, MetaCyc**

5



Contains indepth biological data of studied model organism. **Flybase, RGD**

NCBI - <http://www.ncbi.nlm.nih.gov/>

NCBI Resources How To

National Center for Biotechnology Information

Resources

- NCBI Home
- All Resources (A-Z)
- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Genome

1000 prokaryotic genomes are now completed and available in the Genome database.



1 2 3 4

How To...

- Obtain the full text of an article
- Retrieve all sequences for an organism or taxon
- Find a homolog for a gene in another organism
- Find genes associated with a phenotype or disease
- Design PCR primers and check them for specificity
- Find the function of a gene or gene product
- Determine conserved synteny between the genomes of two organisms

[See all ...](#)

Popular Resources

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

NCBI News

November and October News	02 Dec 2009
Featured: New Discovery-oriented PubMed and NCBI Homepage. T...	
NCBI News - September 2009	05 Oct 2009
The September 2009 issue of the NCBI News is available ...	
NCBI News - August 2009	19 Aug 2009
The August 2009 issue of the NCBI News is available online. ...	

31

[More...](#)

NCBI – all databases

The screenshot shows the NCBI homepage with a blue header bar at the top. Below the header, there's a banner for COVID-19 Information. On the left, a sidebar lists various resources like NCBI Home, Resource List (A-Z), and All Resources. The main content area is titled 'All Resources' and has tabs for All, Databases, Downloads, Submissions, Tools, and How To. The 'Databases' tab is selected. Under 'Databases', several sections are listed: Assembly, BioCollections, BioProject (formerly Genome Project), BioSample, BioSystems, and Bookshelf. Each section has a brief description.

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

COVID-19 Information X

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

All Resources

All Databases Downloads Submissions Tools How To

Databases

Assembly
A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.

BioCollections
A curated set of metadata for culture collections, museums, herbaria and other natural history collections. The records display collection codes, information about the collections' home institutions, and links to relevant data at NCBI.

BioProject (formerly Genome Project)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.

BioSample
The BioSample database contains descriptions of biological source materials used in experimental assays.

BioSystems
Database that groups biomedical literature, small molecules, and sequence data in terms of biological relationships.

Bookshelf
A collection of biomedical books that can be searched directly or from linked data in other NCBI databases. The collection includes biomedical textbooks, other scientific titles, genetic resources such as *GeneReviews*, and NCBI help manuals.

NCBI – all tools

The screenshot shows the NCBI homepage with a blue header bar at the top. The main content area has a light gray background. On the left, there's a vertical sidebar with a light blue background containing a list of links. The 'Tools' link in this sidebar is highlighted with a blue arrow pointing towards the main content area. The main content area features a dark blue header with the NCBI logo, 'Resources', 'How To', and a 'Sign in to NCBI' button. Below this is a search bar with dropdown menus for 'All Databases' and 'Search'. A red banner at the top of the main content area contains a white exclamation mark icon, the text 'COVID-19 Information', and links to various COVID-19 resources. The main content area is titled 'All Resources' and includes tabs for 'All', 'Databases', 'Downloads', 'Submissions', 'Tools' (which is selected and highlighted with a black border), and 'How To'. The 'Tools' section lists several tools with descriptions:

- 1000 Genomes Browser**: An interactive graphical viewer that allows users to explore variant calls, genotype calls and supporting evidence (such as aligned sequence reads) that have been produced by the [1000 Genomes Project](#).
- Amino Acid Explorer**: This tool allows users to explore the characteristics of amino acids by comparing their structural and chemical properties, predicting protein sequence changes caused by mutations, viewing common substitutions, and browsing the functions of given residues in conserved domains.
- BLAST Microbial Genomes**: Performs a BLAST search for similar sequences from selected complete eukaryotic and prokaryotic genomes.
- BLAST RefSeqGene**: Performs a BLAST search of the genomic sequences in the [RefSeqGene](#)/LRG set. The default display provides ready navigation to review alignments in the Graphics display.
- BLAST Tutorials and Guides**: This page links to a number of BLAST-related tutorials and guides, including a selection guide for BLAST algorithms, descriptions of BLAST output formats, explanations of the parameters for stand-alone BLAST, directions for setting up stand-alone BLAST on local machines and using the BLAST URL API.
- Basic Local Alignment Search Tool (BLAST)**: Finds regions of local similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as

EBI - <http://www.ebi.ac.uk/>

The screenshot shows the homepage of the European Bioinformatics Institute (EBI). The top navigation bar includes links for 'Databases', 'Tools', 'EBI Groups', 'Training', 'Industry', 'About Us', 'Help', 'Site Index', and links for 'Give us feedback' and 'Advanced Search'. The main content area features a section titled 'Data Resources & Tools' with links to various databases like EMBL-BANK, UniProt, Gene Expression, Ensembl, InterPro, and PDBe, as well as tools for Genomes, Nucleotide Sequences, Protein Sequences, Macromolecular Structures, Small Molecules, Gene Expression, Molecular Interactions, Reactions & Pathways, Protein Families, Enzymes, Literature, Taxonomy, Ontologies, Patent Resources, Text Mining, Downloads, and Web Services. Below this is a photograph of the EBI building and a tree. A horizontal line separates this from the 'European Bioinformatics Institute' logo. The 'About the EBI' section lists Research, PhD Studies, Training, Industry Support, Group & Team Leaders, and EBI Funders. The 'Latest News' section features a recent article about a £10M investment in bioscience data handling capacity. The 'Events' section is also present.

Data Resources & Tools

- [EMBL-BANK](#)
- [UniProt](#)
- [Gene Expression](#)
- [Ensembl](#)
- [InterPro](#)
- [PDBe](#)
- [Genomes](#)
- [Nucleotide Sequences](#)
- [Protein Sequences](#)
- [Macromolecular Structures](#)
- [Small Molecules](#)
- [Gene Expression](#)
- [Molecular Interactions](#)
- [Reactions & Pathways](#)
- [Protein Families](#)
- [Enzymes](#)
- [Literature](#)
- [Taxonomy](#)
- [Ontologies](#)
- [Patent Resources](#)
- [Text Mining](#)
- [Downloads](#)
- [Web Services](#)

European Bioinformatics Institute

About the EBI

- [Research](#)
- [PhD Studies](#)
- [Training](#)
- [Industry Support](#)
- [Group & Team Leaders](#)
- [EBI Funders](#)
- [User Support](#)
- [EBI Mission](#)
- [People](#)
- [Events at the EBI](#)
- [Genome Campus Events](#)
- [How to Find us](#)

Latest News

■ [UK leads European research programme with £10M investment in bioscience data handling capacity](#)
25 August 2009
The UK has made its first substantial commitment to a major emerging pan-European science project with a £10M investment by the Biotechnology and Biological Sciences Research Council (BBSRC). BBSRC has awarded funding to the European Molecular Biology Laboratory's European Bioinformatics Institute to permit a dramatic increase in the institute's data storage and handling capacity... [more](#)

Events

EBI Hosted Project Websites

Ensembl - <http://www.ensembl.org/index.html>

The screenshot shows the Ensembl homepage with a blue header bar at the top. Below the header, there's a search bar with placeholder text "Search for: e.g. gene BRCA2 or AL032821.2.1.143563 or muscular dystrophy". To the right of the search bar is a "Go" button. The main content area is divided into several sections:

- Browse a Genome**: A section about the Ensembl project producing genome databases for vertebrates and other eukaryotic species. It includes links to "Popular genomes" (Human, Mouse, Zebrafish) and a dropdown menu for "All genomes".
- New to Ensembl?**: A section listing various ways to interact with Ensembl:
 - [Learn how to use Ensembl](#) (with video tutorials)
 - [Add custom tracks](#) (using the Control Panel)
 - [Upload your own data](#) (saving to your account)
 - [Search for a DNA or protein sequence](#) (using BLAST or BLAT)
 - [Fetch only the data you want](#) (using the Ensembl Perl API)
 - [Download our databases via FTP](#) (in FASTA, MySQL formats)
 - [Mine Ensembl with BioMart](#) (exporting sequences or tables)
- What's New in Release 55 (14 July 2009)**: A section highlighting changes in the latest release.

ExPasy – Swiss Bioinformatics Resource Portal

<https://www.expasy.org/>

The screenshot shows the ExPasy homepage with a search bar and a sidebar containing category filters.

SIB Resources

- ASAP**: Web-based, cooperative portal for single-cell data analyses.
- Rhea**: Expert-curated database of biochemical reactions.
- UniProtKB/Swiss-Prot**: Protein knowledgebase.
- Nextstrain**: Impact of pathogen genome data on science and public health.
- SwissDrugDesign**: Widening access to computer-aided drug design.
- Bgee**: Gene expression expertise.
- SwissOrthology**: One-stop shop for orthologs.
- V-pipe**: Viral genomics pipeline.
- neXtProt**: Human protein knowledgebase.
- EPD**: Eukaryotic Promoter Database.
- SwissLipids**: Knowledge resource for lipids.
- STRING**: Protein-protein interaction networks and enrichment analysis.

Category Filters (Sidebar)

- Genes & Genomes**
 - Genomics
 - Metagenomics
 - Transcriptomics
- Proteins & Proteomes**
- Evolution & Phylogeny**
 - Evolution biology
 - Population genetics
- Structural Biology**
 - Drug design
 - Medicinal chemistry
 - Structural analysis
- Systems Biology**
 - Glycomics
 - Lipidomics
 - Metabolomics

Gene Expression Resource: GEO -

<http://www.ncbi.nlm.nih.gov/geo/>

The screenshot shows the NCBI Gene Expression Omnibus (GEO) homepage. At the top left is the NCBI logo, and at the top right is the GEO logo with the text "Gene Expression Omnibus". The top navigation bar includes links for HOME, SEARCH, SITE MAP, Handout, NAR 2006 Paper, NAR 2002 Paper, FAQ, MIAME, Email GEO, and a login link. A sidebar on the right displays "Public data" statistics: GPL Platforms 6387, GSM Samples 348419, GSE Series 13598, and a total of 368404. Another sidebar lists "Site contents" such as Documentation, Programmatic access, and various submission and browse links. The main content area features a "GEO navigation" section with three main categories: QUERY, BROWSE, and SUBMIT. The QUERY section contains links for DataSets, Gene profiles, GEO accession, and GEO BLAST, each with a "GO" button. The BROWSE section contains links for DataSets, GEO accessions, Platforms, Samples, and Series. The SUBMIT section contains links for Direct deposit / update, Create new account, and Web deposit / update.

Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

GEO navigation

QUERY

- DataSets
- Gene profiles
- GEO accession
- GEO BLAST

BROWSE

- DataSets
 - Platforms
 - Samples
 - Series
- GEO accessions

SUBMIT

- Direct deposit / update
- Create new account
- Web deposit / update

Public data

GPL Platforms	6387
GSM Samples	348419
GSE Series	13598
Total	368404

Site contents

Documentation

- Overview
- FAQ
- Find
- Submission guide
- Linking & citing
- Journal citations
- Programmatic access
- DataSet clusters
- GEO announce list
- Data disclaimer
- GEO staff

Query & Browse

- Repository browser
- Submitter contacts
- SAGEmap
- FTP site
- GEO Profiles
- GEO DataSets

Deposit & Update

UniProt - <http://www.uniprot.org/>

The UniProt homepage features a prominent search bar at the top. The search interface includes a dropdown menu for "Search in" (set to "Protein Knowledgebase (UniProtKB)"), a query input field, and buttons for "Search", "Clear", and "Fields »". Below the search bar are several navigation links: "Search", "Blast", "Align", "Retrieve", and "ID Mapping".

WELCOME

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.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none">★ Swiss-Prot, which is manually annotated and reviewed.★ TrEMBL, which is automatically annotated and is not reviewed.
UniRef	Sequence clusters, used to speed up similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.

NEWS

UniProt release 15.7 – Sep 1, 2009

Formyl peptide receptors: the missing link between olfaction and immune system · Cross-references to STRING

› Statistics for UniProtKB:
Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives

SITE TOUR



Metabolic Pathway: KEGG - <http://www.genome.jp/kegg/>



KEGG Home
Introduction
Overview
Release notes
Current statistics

KEGG Identifiers

KGML

KEGG API

KEGG FTP

KegTools

Feedback

GenomeNet

KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

Main entry point to the KEGG web service
KEGG2 [KEGG Table of Contents](#) [Update notes](#) [Help](#)

Data-oriented entry points

KEGG PATHWAY	Pathway maps and pathway modules	Pathway maps
KEGG BRITE	Functional hierarchies and ontologies	Brite hierarchies
KEGG ORTHOLOGY	KO system and ortholog annotation	
KEGG GENES	Genomes, genes, and proteins	
KEGG LIGAND	Chemical compounds, glycans, and reactions	
KEGG DISEASE	Human diseases	
KEGG DRUG	Drugs	

Organism-specific entry points
KEGG Organisms (example) [hsa](#)

Other entry points
KEGG Atlas

New interface to navigate pathway maps

Protein Structure Database: PDB -

<https://www.rcsb.org/>

PDB
PROTEIN DATA BANK

WHAT'S NEW | HELP | PRINT

- Home
- News & Publications
- Policies
- FAQ
- Contact
- Feedback
- About Us

- Deposition
- All Deposit Services
- Electron Microscopy
- NMR
- Validation Server
- BioSync Beamline
- Related Tools

- Search
- Advanced Search
- Latest Release
- Latest Publications
- Sequence Search
- Ligand Search
- Unreleased Entries
- Browse Database
- Histograms

- Tools

An Information Portal to Biological Macromolecular Structures
As of **Tuesday Sep 08, 2009** there are 60046 Structures [?](#) | [PDB Statistics](#) [?](#)

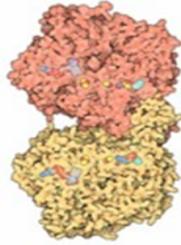
PDB ID or keyword [Advanced Search](#)

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the [wwPDB](#), the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Molecule of the Month: Xanthine Oxidoreductase



Our diet includes a wide variety of different molecules. Many of these molecules are broken down completely and used to generate the metabolic energy that powers our cells. Others are disassembled piece-by-piece and recycled to build our own proteins and nucleic acids. The ones that are left over are broken down and discarded. Xanthine oxidoreductase, shown here from PDB entry [1fo4](#), is the last stop for extra purine nucleotides (ATP and GTP) in our cells. Purines are broken down in several steps, ultimately yielding uric acid, which is excreted from the body.

[■ Read more ...](#) [■ Previous Features](#)

PSI Featured Molecule: Toxin-antitoxin VapBC-5

News

- Complete News
- Newsletter
- Discussion Forum
- Job Listings

08-September-2009
Improved Navigation of the RCSB PDB Website

RCSB PDB web pages been reorganized to make navigating the website and search results easier and more intuitive.

The **left-hand menu** now groups frequently-used webpages into sections that can be moved up and down to create a left-hand menu ordered by user interest. This customized menu will then appear on every web page. Several enhancements have also been added to the **query result pages**.

Plant-specific database: TAIR

<http://www.arabidopsis.org/>

The screenshot shows the homepage of The Arabidopsis Information Resource (TAIR). At the top, there's a navigation bar with links for Home, Help, Contact, About Us, and Login/Register. Below that is a secondary navigation bar with links for Search, Browse, Tools, Stocks, Portals, Download, Submit, and News. The main content area features a logo with a flower and the word "tair". A large heading says "The Arabidopsis Information Resource". Below it, a paragraph describes TAIR as maintaining a database of genetic and molecular biology data for *Arabidopsis thaliana*. It mentions the complete genome sequence, gene structures, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, and publications. The text also notes that gene product function data is updated every two weeks from research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods, along with community submissions of new and updated genes. TAIR provides extensive linkouts to other Arabidopsis resources. A note from the Carnegie Institution for Science states that TAIR is located at their Department of Plant Biology and funded by the National Science Foundation. The NSF logo is present. On the right side, there's a "Breaking News" section with three entries: "Stock Donation Made Easy [July 24, 2009]", "Synteny Viewer At TAIR [July 7, 2009]", and "TAIR9 Genome Release [June 19, 2009]". Each news item includes a brief description. At the bottom, there's a grid of images showing various parts of an Arabidopsis plant, with the word "YOUR" overlaid on one of the images.

Breaking News

Stock Donation Made Easy
[July 24, 2009]
ABRC and NASC are currently seeking donations of characterized mutant lines. We have developed a **simplified form** for mutant and transgenic donations. Just send the completed form together with a tube containing as much seed as you can spare to the stock centers.

Synteny Viewer At TAIR
[July 7, 2009]
A synteny viewer, comparing syntenic regions between *A. thaliana* and *A. lyrata*, is now available at TAIR. More genomes will be added soon.

TAIR9 Genome Release
[June 19, 2009]
The TAIR9 genome release is now available at TAIR and NCBI, with 282 new loci, updates to 1254 gene

41

miRTarBase – miRNA-target database

<http://mirtarbase.cuhk.edu.cn/>

The screenshot shows the miRTarBase homepage with a purple header containing the logo, search bar, and navigation links (Home, Search, Browse, Statistics, Help, Download, Contact Us). Below the header, a main section highlights the database's validation process, target gene information, regulatory factors, and experimental techniques like CLIP-Seq and gene expression profiling. A large callout box on the right emphasizes the database's size: Over 430,000 MTIs and 10,000 curated articles.

miRTarBase: The experimentally validated microRNA-target interactions database

As a database, miRTarBase has accumulated more than three hundred and sixty thousand miRNA-target interactions (MTIs), which are collected by manually surveying pertinent literature after NLP of the text systematically to filter research articles related to functional studies of miRNAs. Generally, the collected MTIs are validated experimentally by reporter assay, western blot, microarray and next-generation sequencing experiments. While containing the largest amount of validated MTIs, the miRTarBase provides the most updated collection by comparing with other similar, previously developed databases.

Text-mining Technique to Prescreen Literature

Enhanced Text Mining System
Manually Curation

microRNA and Target Gene Information

Pre- & Mature miRNA Information
mRNA
Target Gene Information

Regulatory Factors of microRNA

miRSponge
Circular RNA
SomamiR
TransmiR
Transcription Factor
miRNA Mutation
miRNA Gene

miRTarCLIP

Identifying microRNA-target Interactions using CLIP-Seq Data
mRNA
Gene Expression Profiling in Cancer / Disease TCGA, NCBI GEO

Tumor Cell

Blood
extracellular vesicles
miRNA-Ago2 complex
apoptotic bodies
Circulation (Blood, Urine, Ascites)

Over 430,000 MTIs & 10,000 Curated Articles

miRTarBase
<https://mirtarbase.cuhk.edu.cn>

dbPTM – Post-Translational Modifications

<https://awi.cuhk.edu.cn/dbPTM/>

dbPTM

HOME STATISTICS SEARCH BROWSE ▾ ANALYSIS ▾ RESOURCE DOWNLOAD

Quick Search (Search by UniProt ID, AC and keywords of gene/protein names) - eg.: CHK2_HUMAN / Histone
CHK2_HUMAN Search

Recent Update History

The updated dbPTM 2019 is now accessible.
Administrator Time 10:00 am at 29th june

PTM Data Updated.
Administrator Time 2:00 pm at 1st june

908,917 Sites Experimental PTM Sites

130+ PTM Types Collecting PTM Types

30+ Databases Integrated Databases

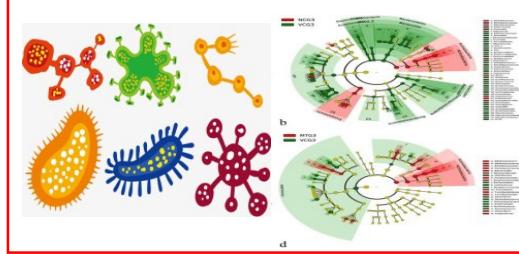
30+ Datasets Benchmark Datasets

Environmental & diet factors

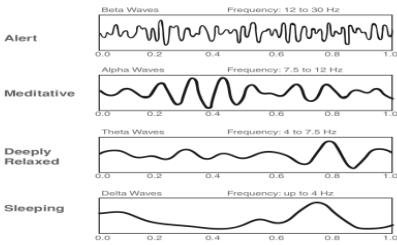


Environmental Genomics

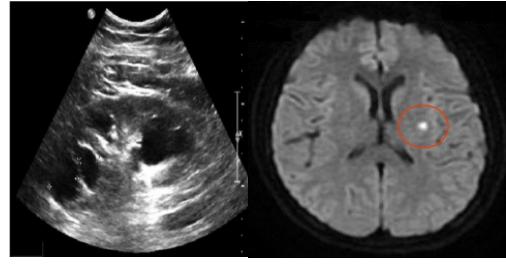
Microbiome



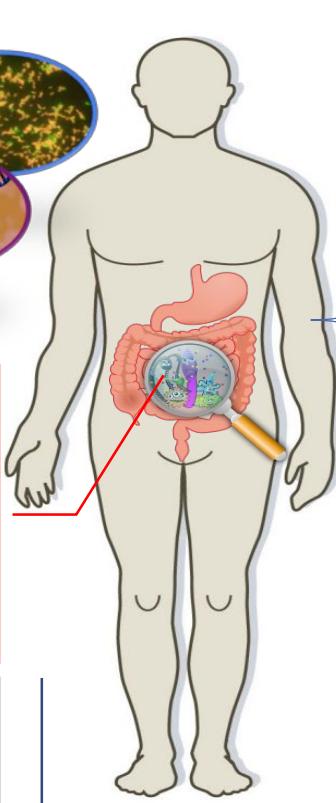
Physiological factors



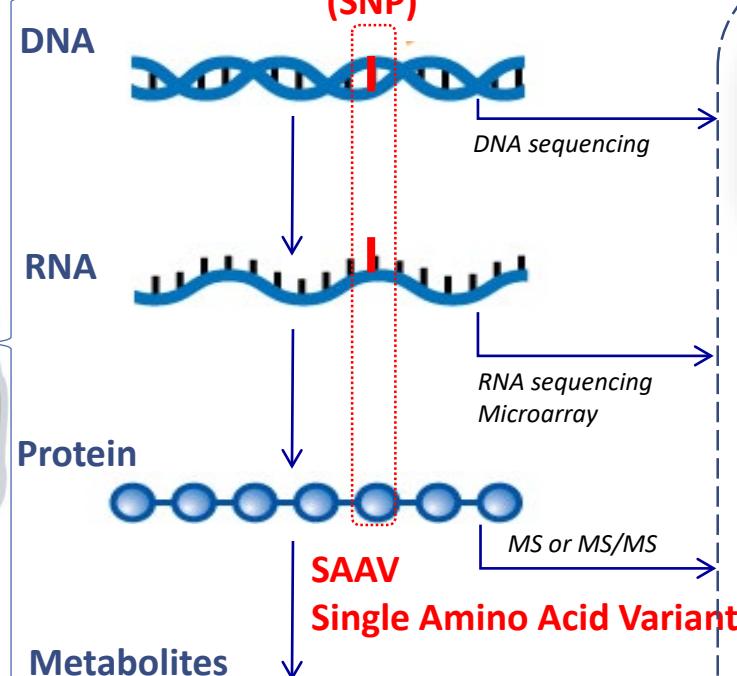
Clinical imaging



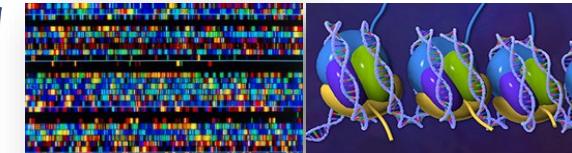
Patient



Single Nucleotide Polymorphism (SNP)

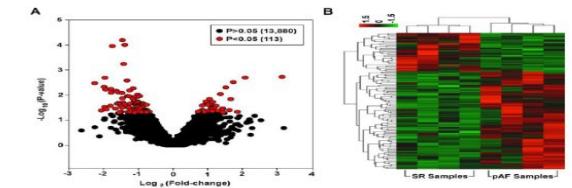


Genomics

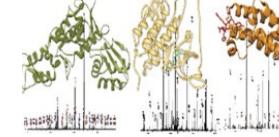


Epigenomics

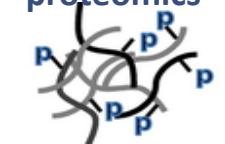
Transcriptomics



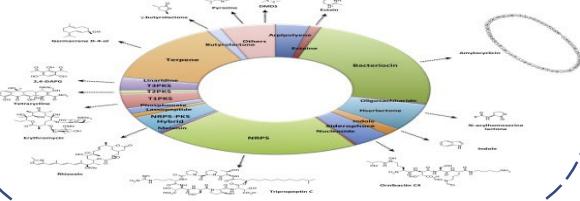
Proteomics



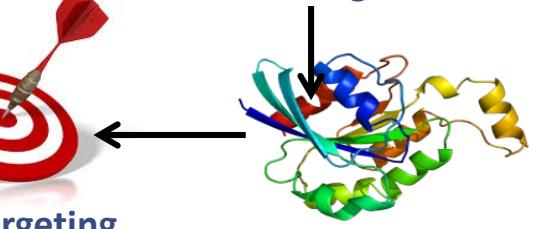
Phospho-proteomics



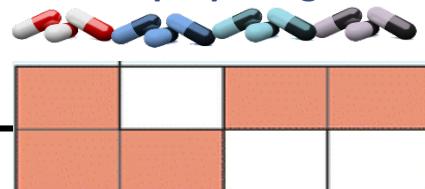
Metabolomics



Personalized signatures



Drug docking or repurposing



+

Standard care

What's the aim of this course?

