國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

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

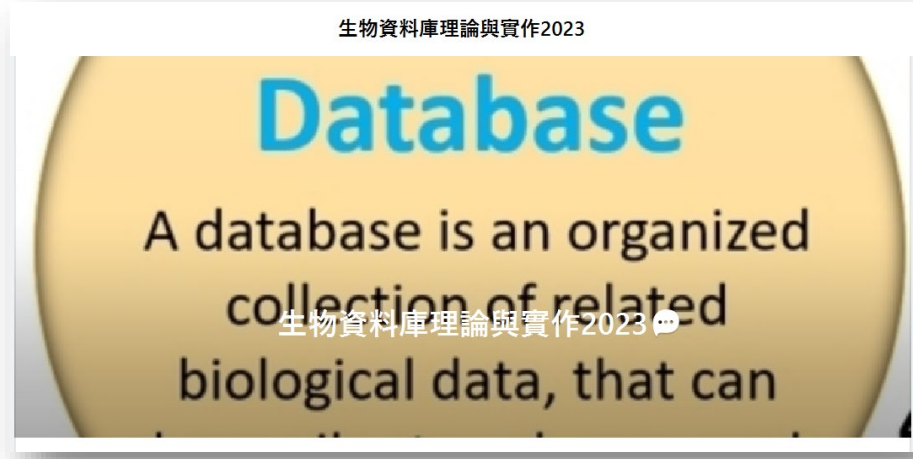
Lab：生物大數據與深度計算實驗室

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

Office Hour：basically Wed. afternoon (**actually anytime**)

Line ID: **francislee0215**

The Line Group of this course:



Education Background and Academic Experiences

Prof. Tzong-Yi Lee

李宗夷 (1980.02.15)

[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

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

Department of Computer Science and Information Engineering, National Central University

博士: 2004 ~ 2008

國立交通大學
生物資訊所

Published 10 SCI papers

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

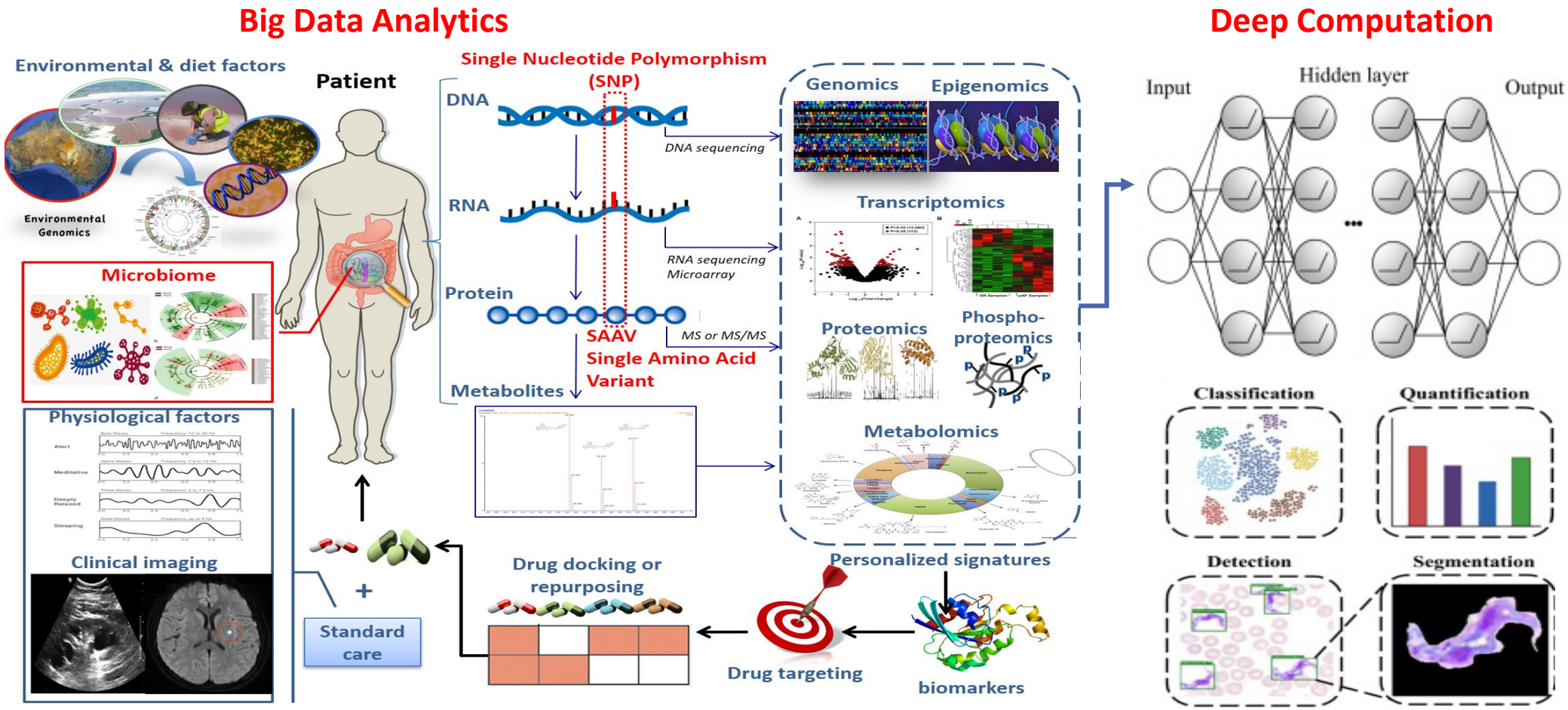
學術表現簡介

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

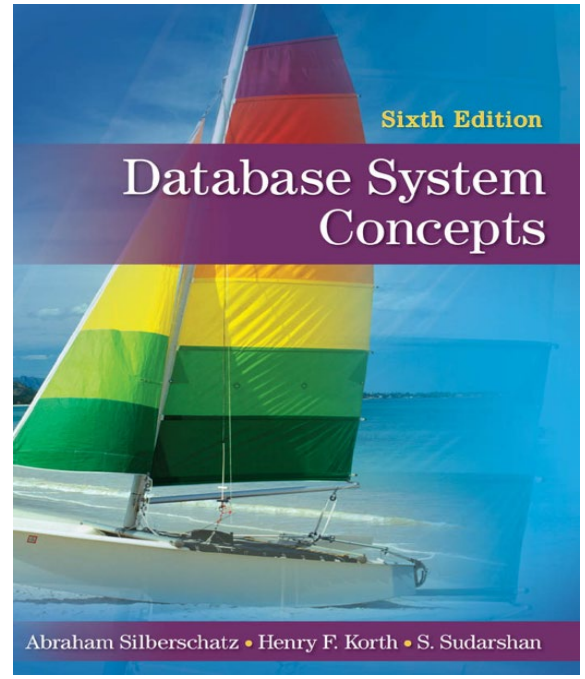


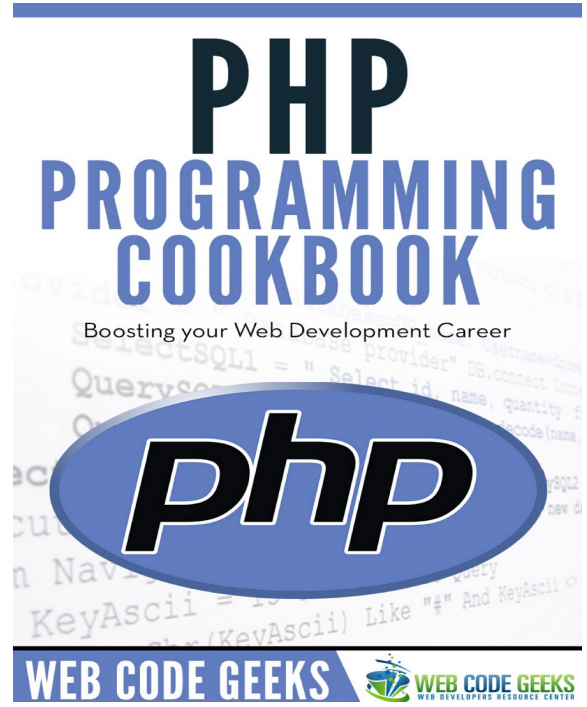
Course Objectives

- 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

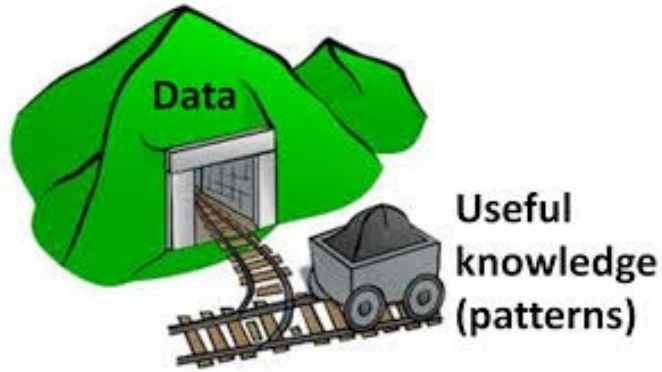
Promises and Expectations

- **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**

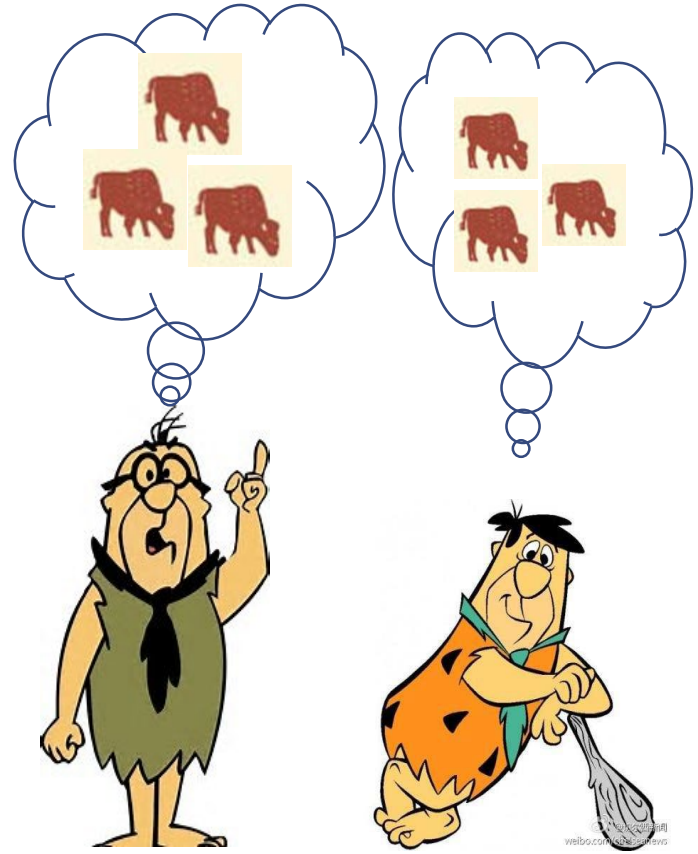
Why do we need Data?

- 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



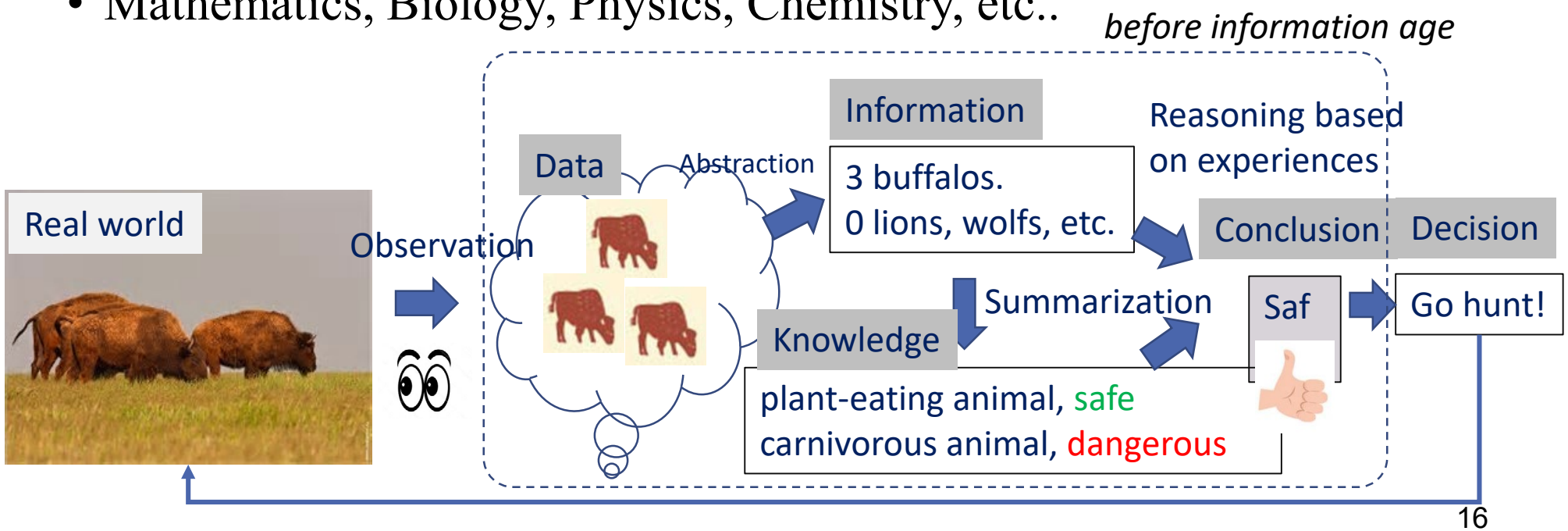
before information age



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

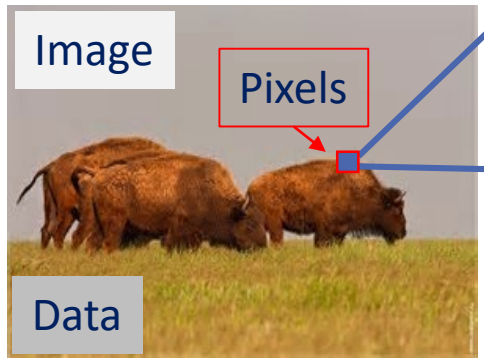
Data, Information and Knowledge

- 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..



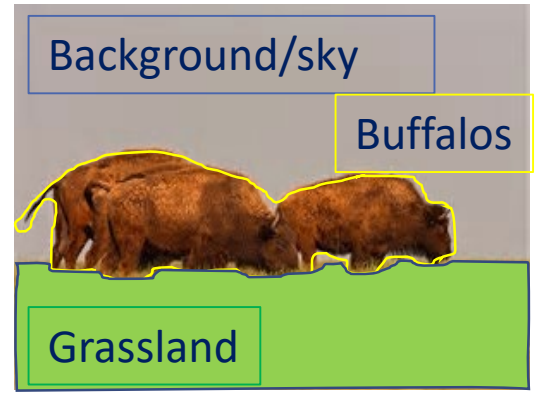
Data, Information and Knowledge

information age



R: 59 G: 73 B:120	R: 78 G: 95 B:141	R:130 G:147 B:191
R: 62 G: 76 B:123	R: 98 G:115 B:161	R:144 G:161 B:205

Image pre-processing
Contour detection
Feature extraction
Pattern recognition
Object classification
Overlapping estimation
etc.



Processed Data

Sampling

Camera

Information

Reasoning by using Expert system

Real world

Data processing in computer system

2 or 3 buffalos.
0 lions, wolfs, etc.

Conclusion

Decision

Knowledge of Human

Summarization

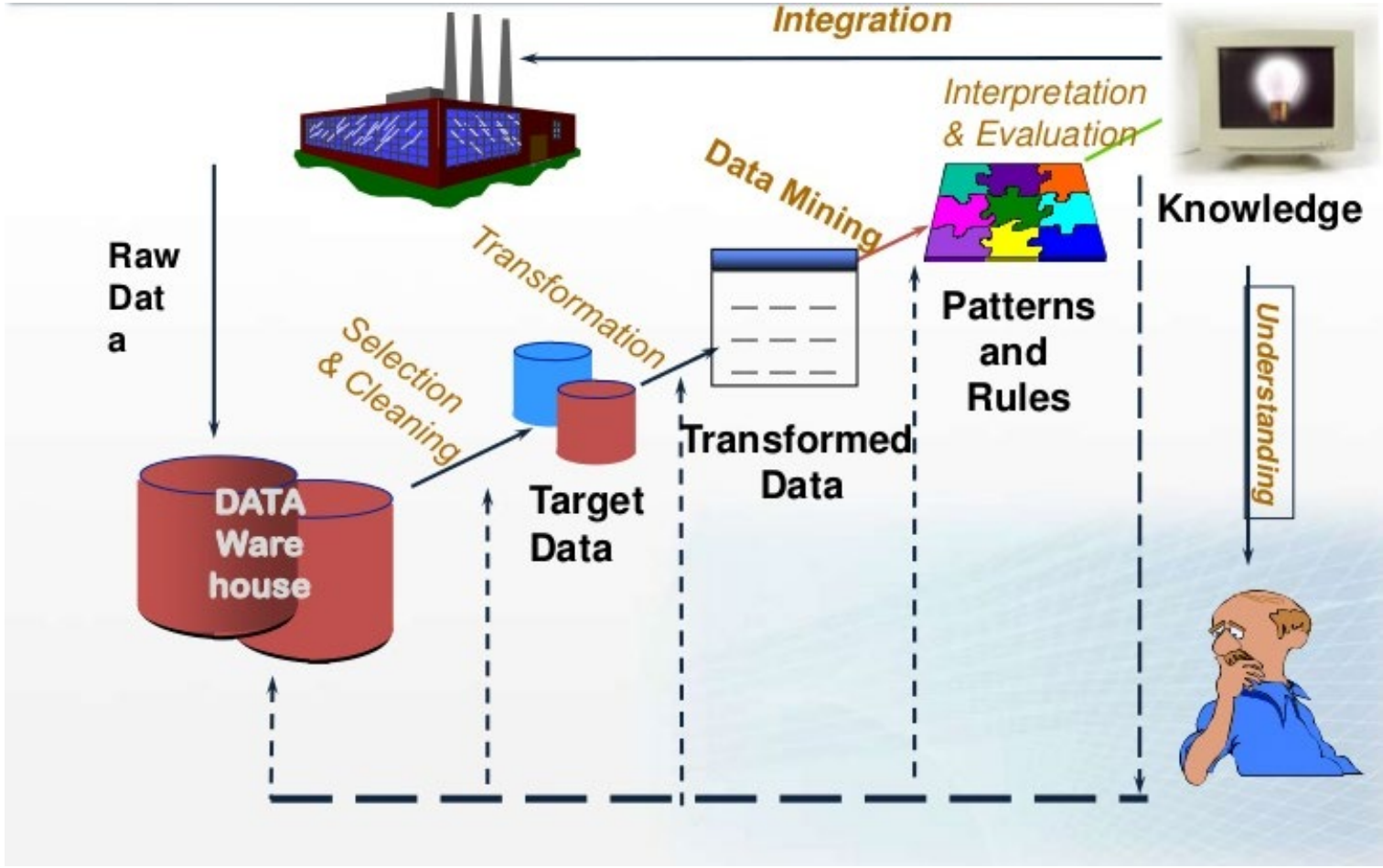
plant-eating animal, safe
carnivorous animal,
dangerous

Saf

Go hunt!

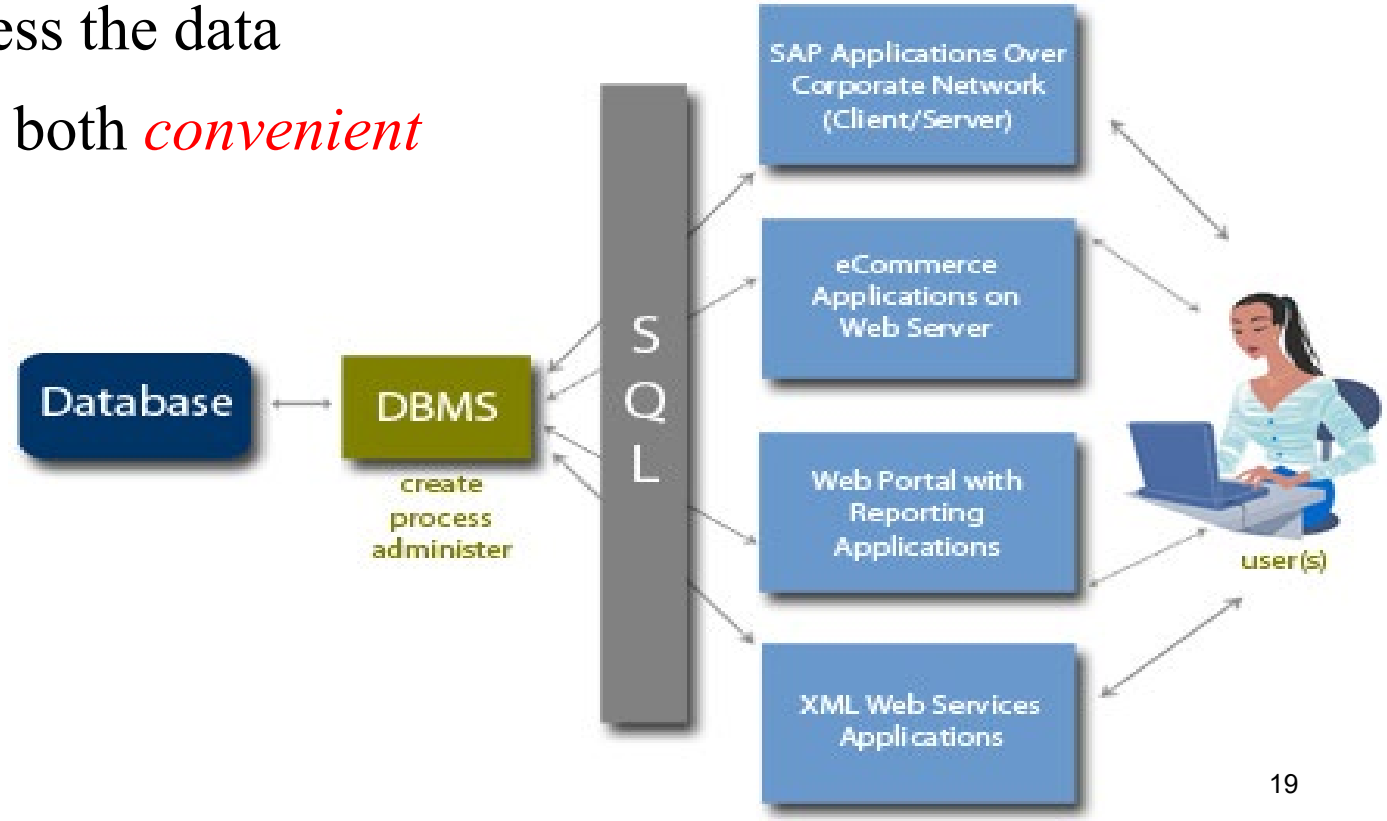
Knowledge Mining: from data to knowledge

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



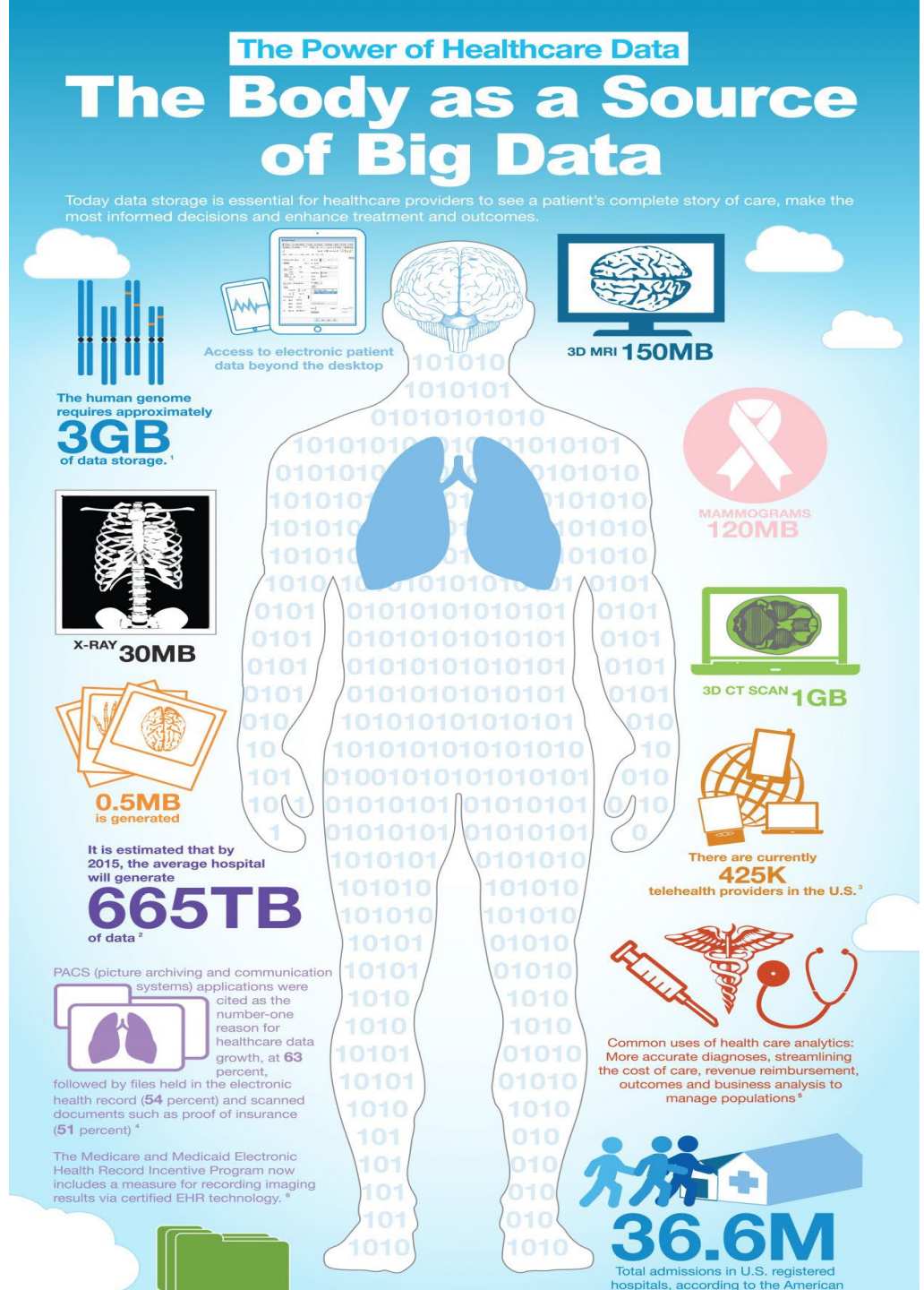
Database Management System (DBMS)

- **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



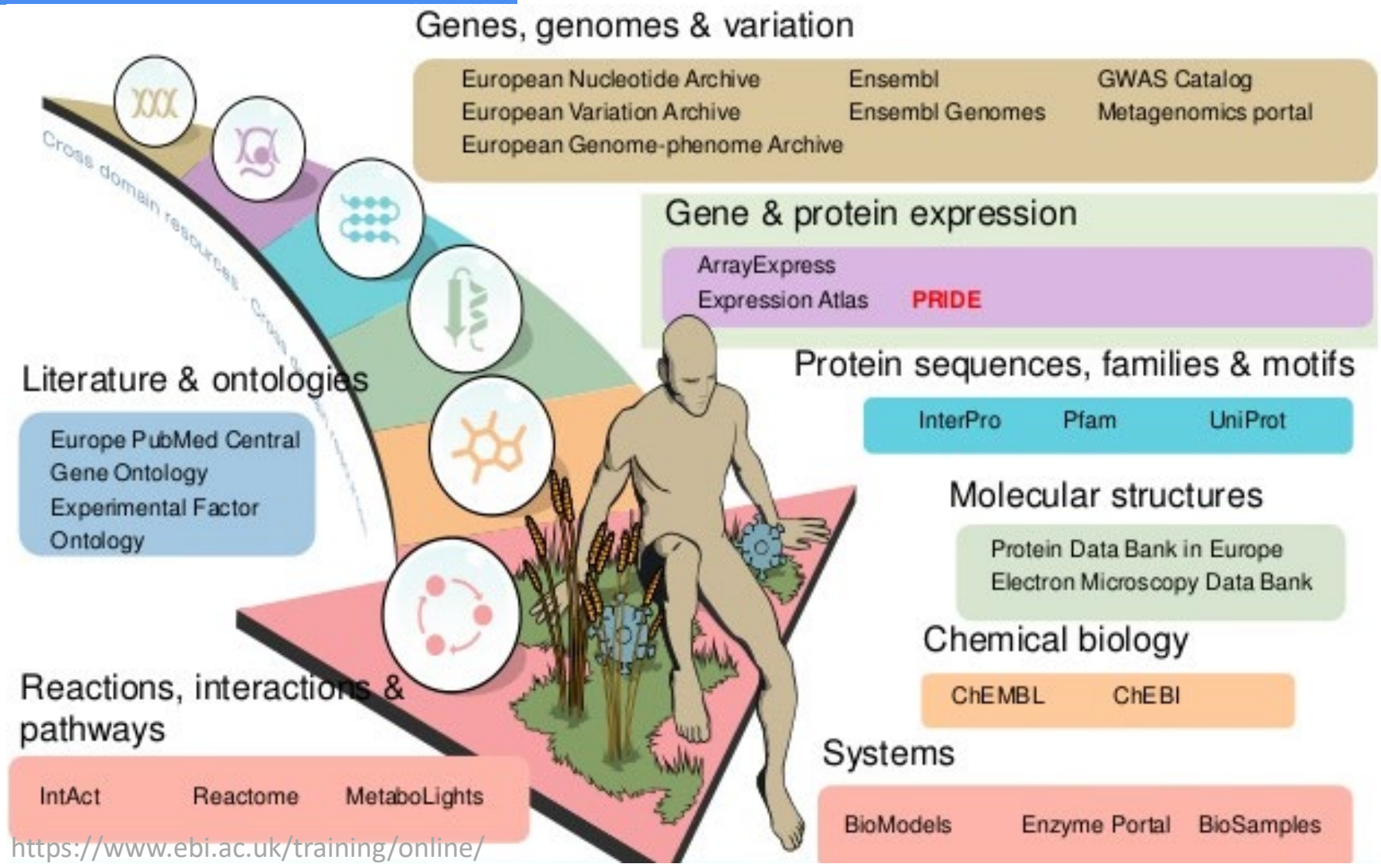
Why Need Database System

- **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



https://datafloq.com/read/body-source-big-data-infographic/

We are in the Data-driven World



We need database, even for evaluating a professor …

Tzong-Yi Lee (李宗夷)

FOLLOW

Professor, National Yang Ming Chiao Tung University
Verified email at nycu.edu.tw - Homepage

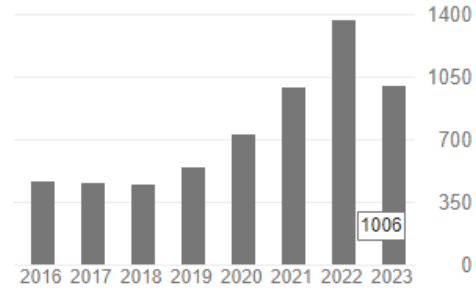
Bioinformatics Multi-Omics Data Analysis Systems Biology Machine Learning and Dee…
Precision Medicine

GET MY OWN PROFILE

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

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

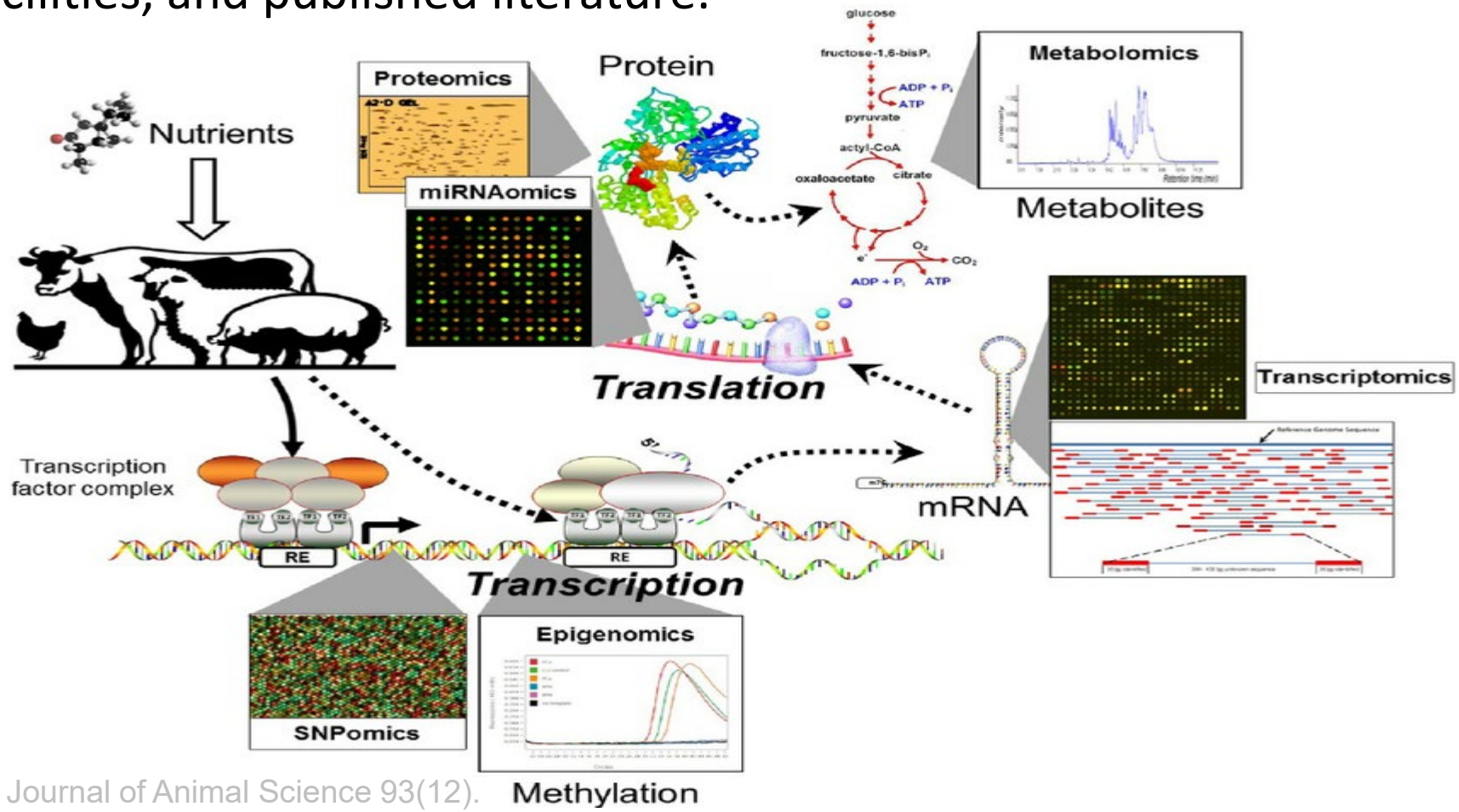
not available available

Based on funding mandates

Co-authors VIEW ALL

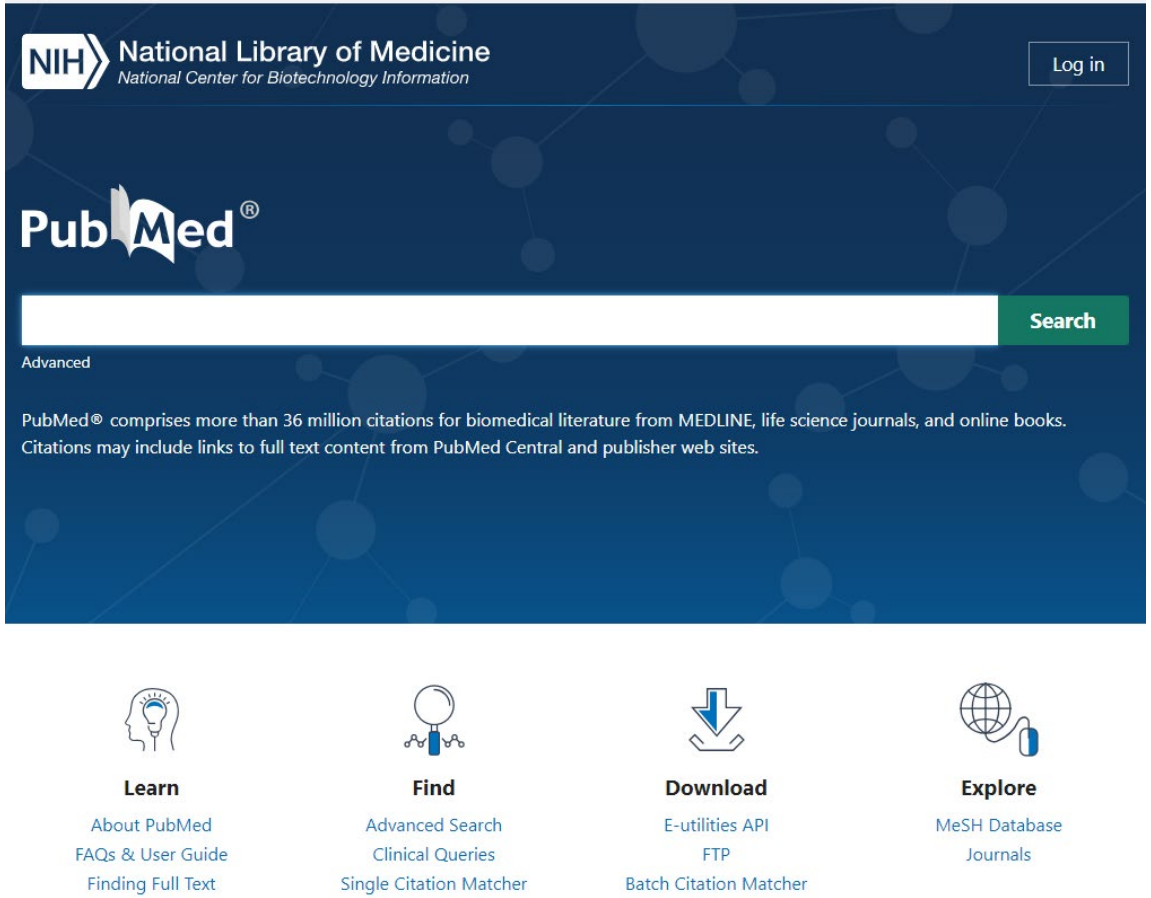
What is Biological Data?

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



Journal of Animal Science 93(12).

Biological text data



OXFORD

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 BC5CDR, KD-DTI 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 advantage 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 therapy, 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 30*M* 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-based 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 donwstream 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).

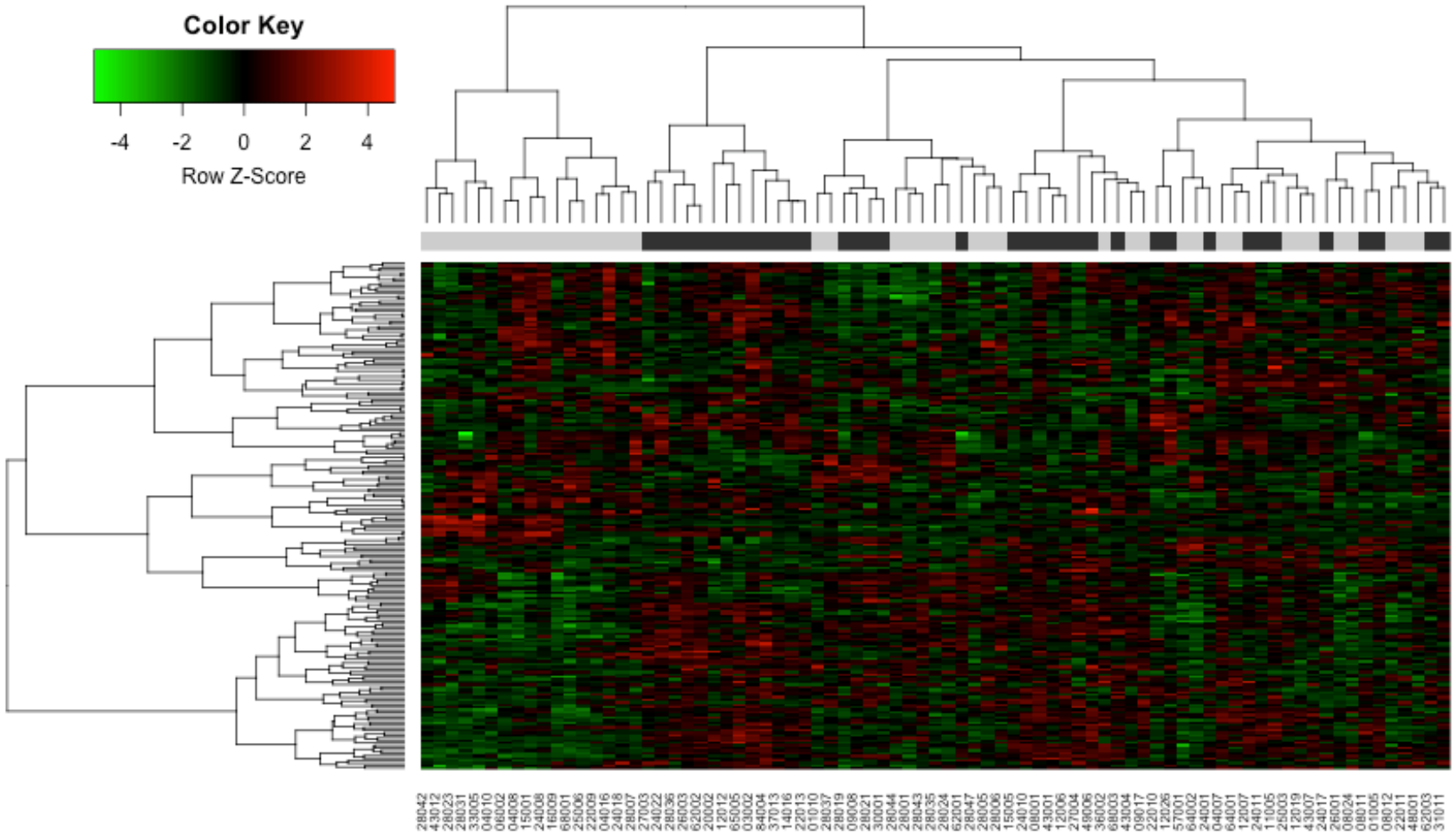
By witnessing the success of pre-training in general NLP, people explore adapting these techniques into biomedical domain. However, directly applying these models to the biomedical domain leads to unsatisfactory performance due to domain shift [8, 9]. A natural solution is to develop pre-training models 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

Sequence data

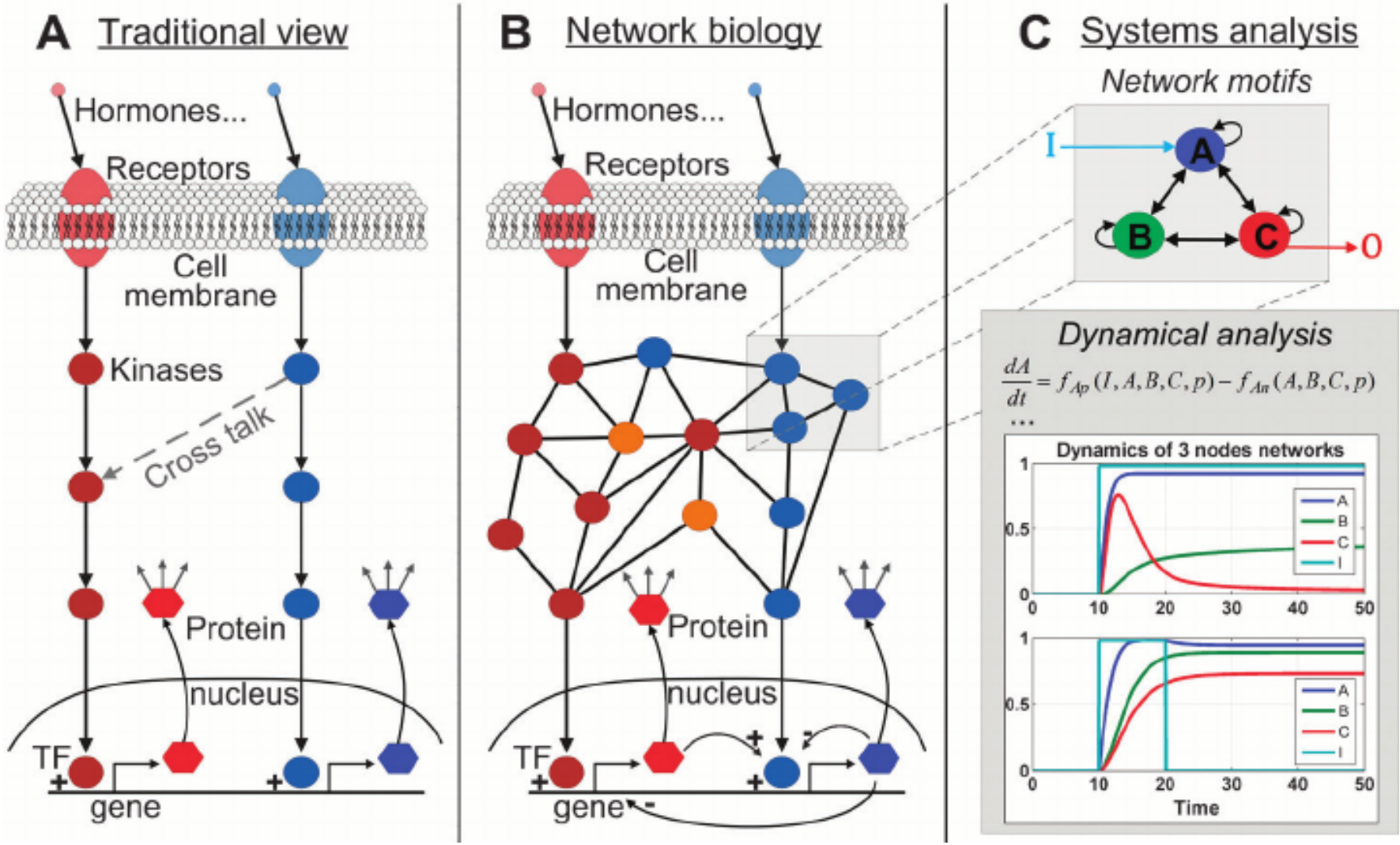
CTCACCGTCGTAAATCTATGATCTGGCTTGGCCTGCAGT
AGCTCTTTCATTTCGGGCTTATCTAATGCTGACTGGTCG
GTCCTGGCTACGCTCCAAAACGTACGTATTCGGGCCATC
GAGGCTAGCGGCACTTCGAGCGATCTATCGGGAGCTTTG
GCTATCGATCGGGCGATCGATGCTGACGTACGTAGCGCG
CGATCGAGCGCGGCTAGCTAGCGGCATCGTAGCTACGTA
GCTACGGCGCTATTTCGATCGAGTCGTGTCTAGTCGGAT
ATAGCTATGCATCTAGCTGAGGCATCTGAGCGGATCGAT
GCTAGGGCGATCGGAGCTAGCTGAGCTAGCTAGCTGAGC
GCTAGCGAGCGTACGAGCGATCGAGCGAGTCTAGCGAGC
GATTCTAGCGATATACATTAGCCCGATCGTATGCTAGCT
AGGGCTAGCATGCGGATCTATCGAGCGGCTATCTGAGCG
ATTCGATCGAGCGATCTAGCGAGCTATCGATCGAGCCGG

Gene expression data

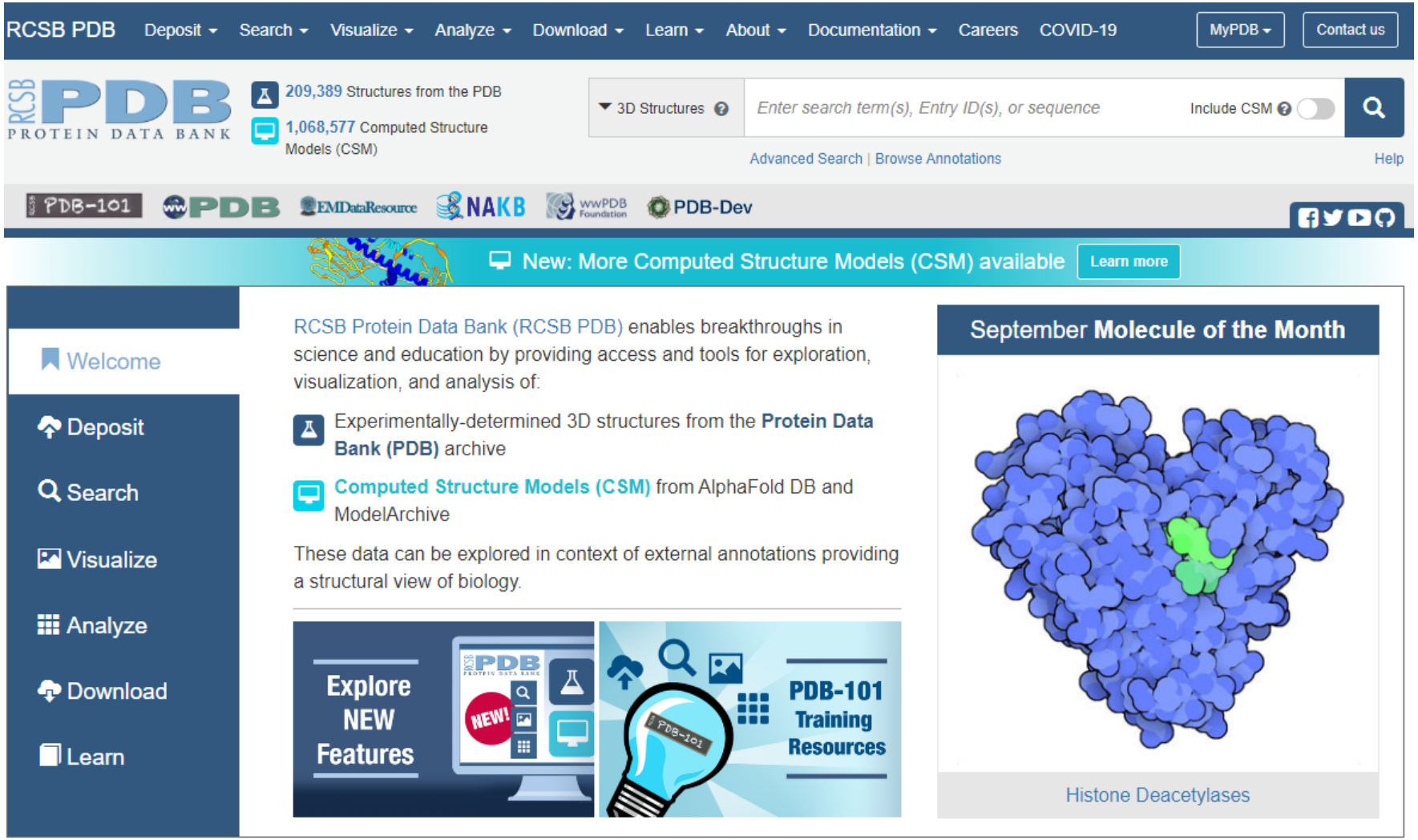


Biological network data

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

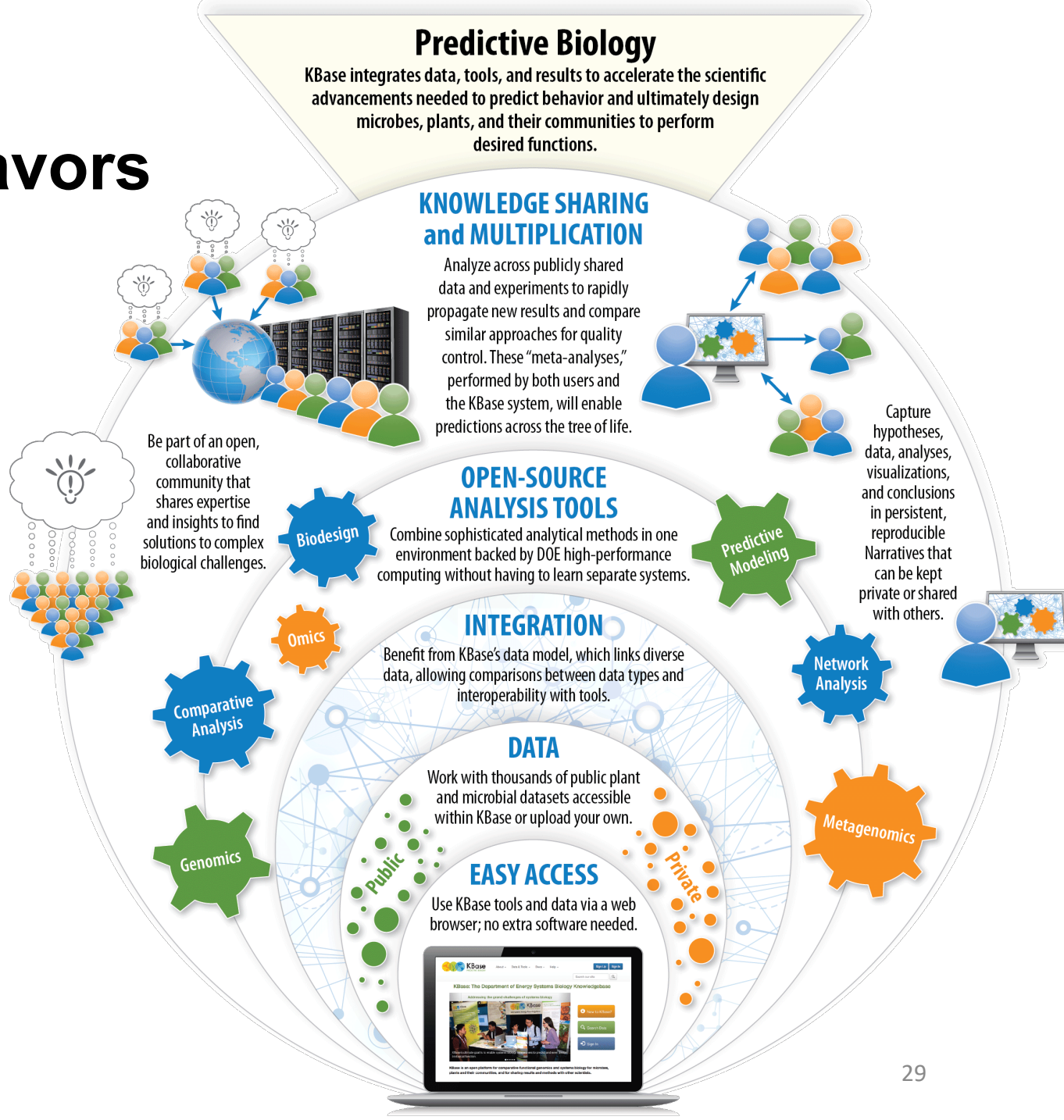


Protein structure data



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
…	…



Biological Database

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

1 Bibliographic Database

Contains article and research papers of different journals **Pubmed**

2 Sequence Database

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

3 Structure Database

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

4 Metabolic Database

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

5 Model organism Database

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

6 Enzyme Database

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

7 Disease Database

Disease related information

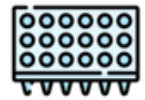
OMIM

8 Chemical Database

Data on several small organic molecules **PubChem**

9 Microarray Database

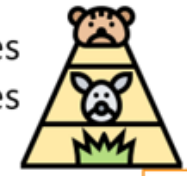
Gene expression data from microarray experiments **GEO**



10 Taxonomic Database

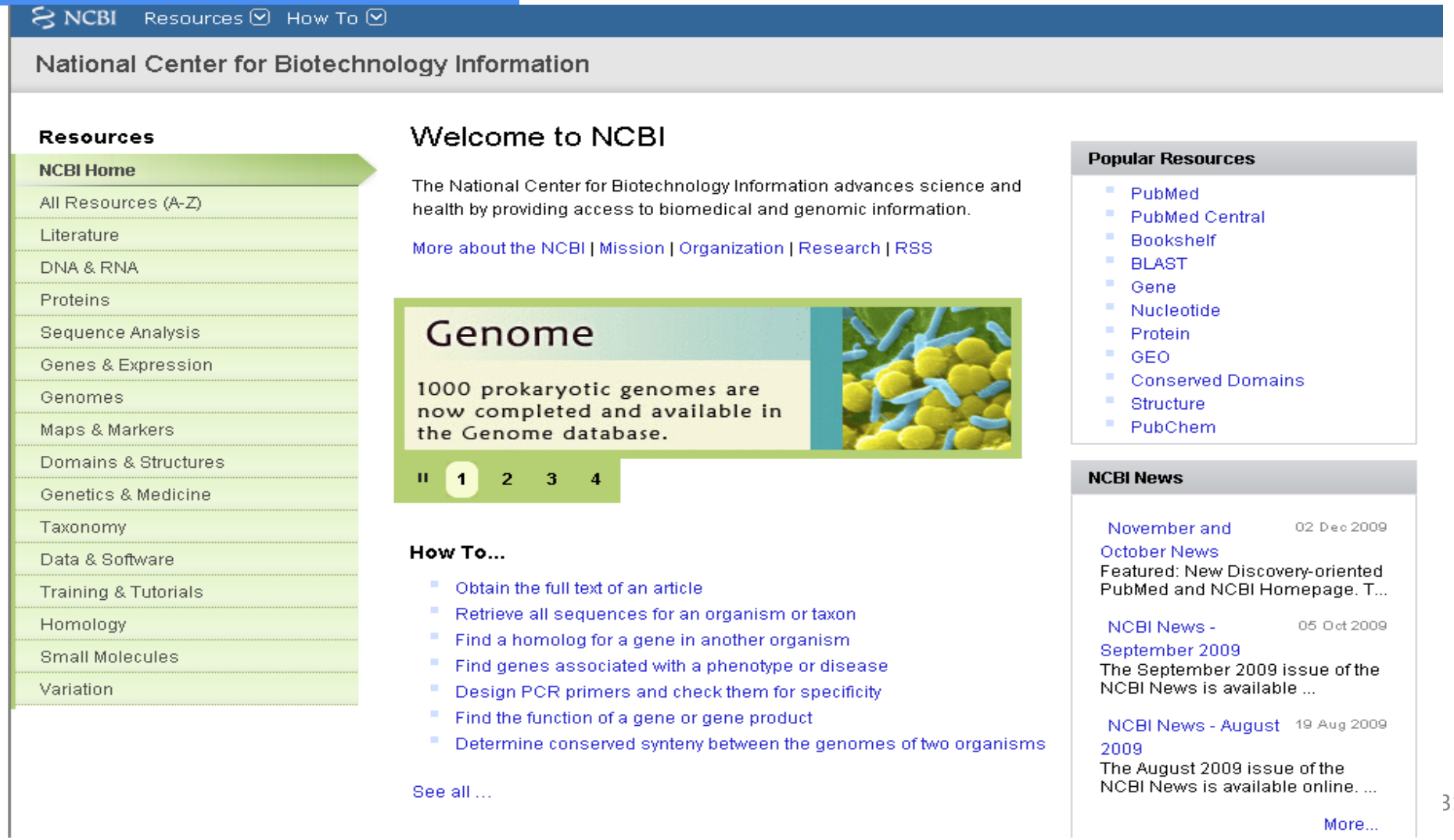
Database that provides information on earths species of animals, plants

Catalogue of life

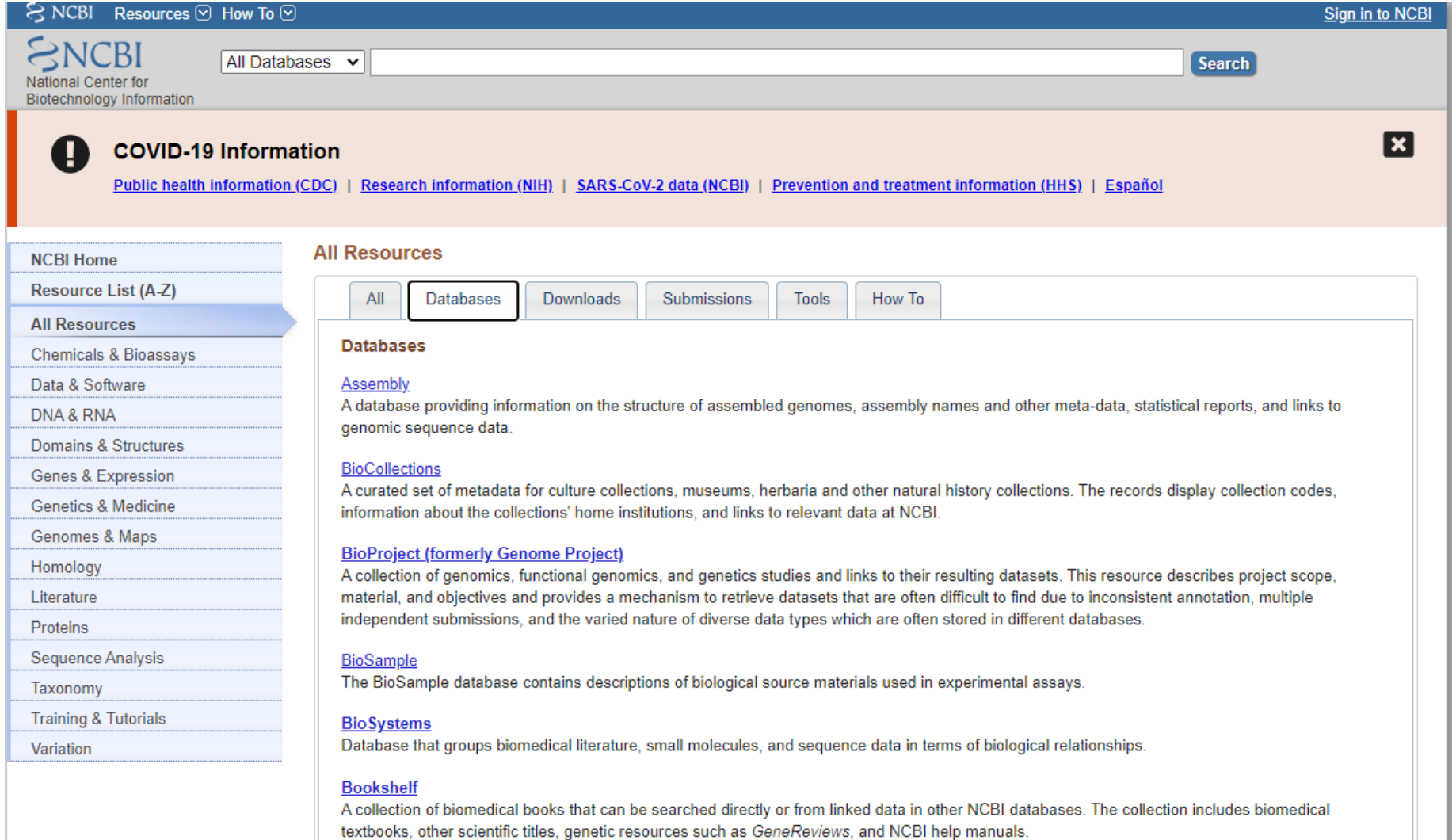


www.biologyexams4u.com

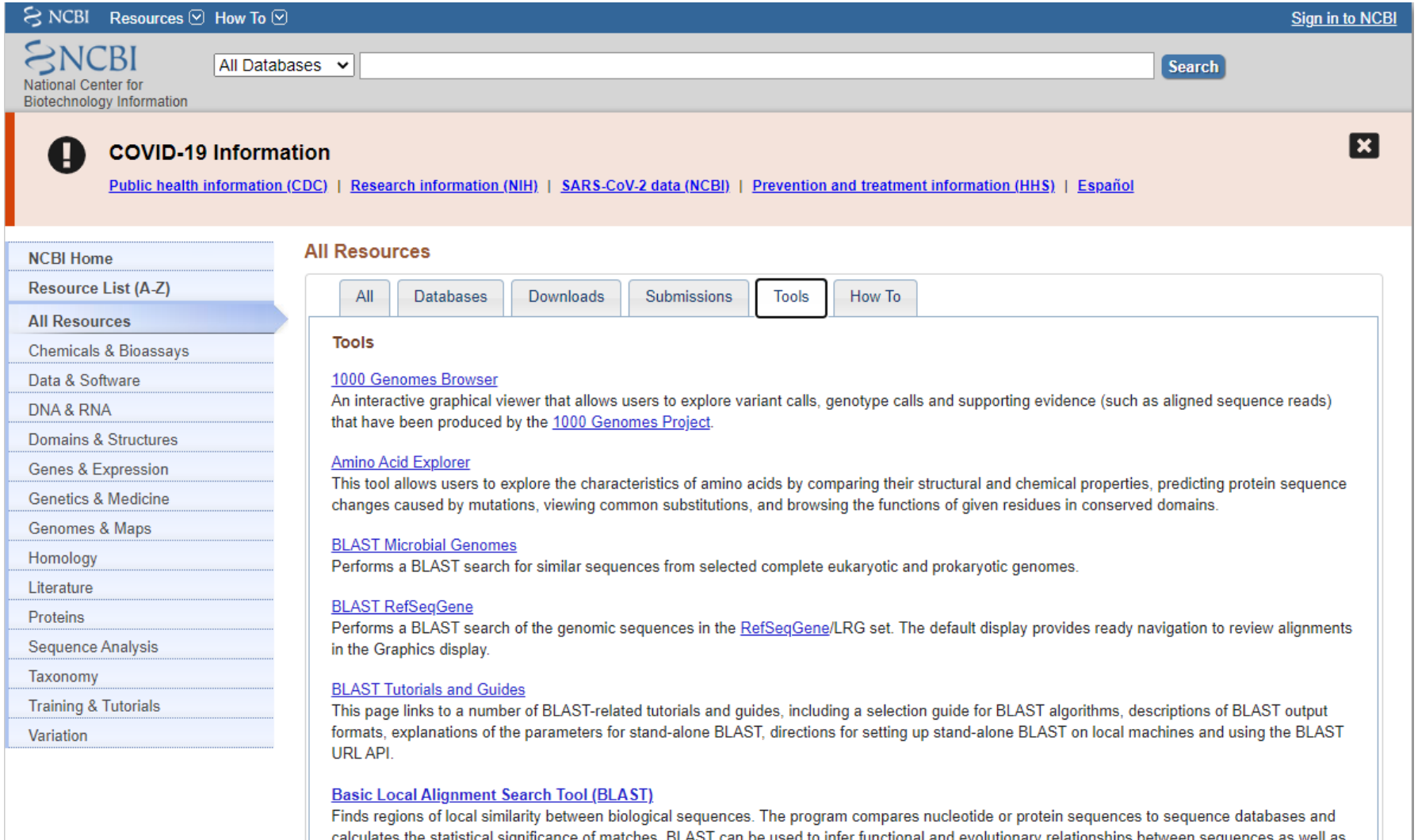
NCBI - http://www.ncbi.nlm.nih.gov/



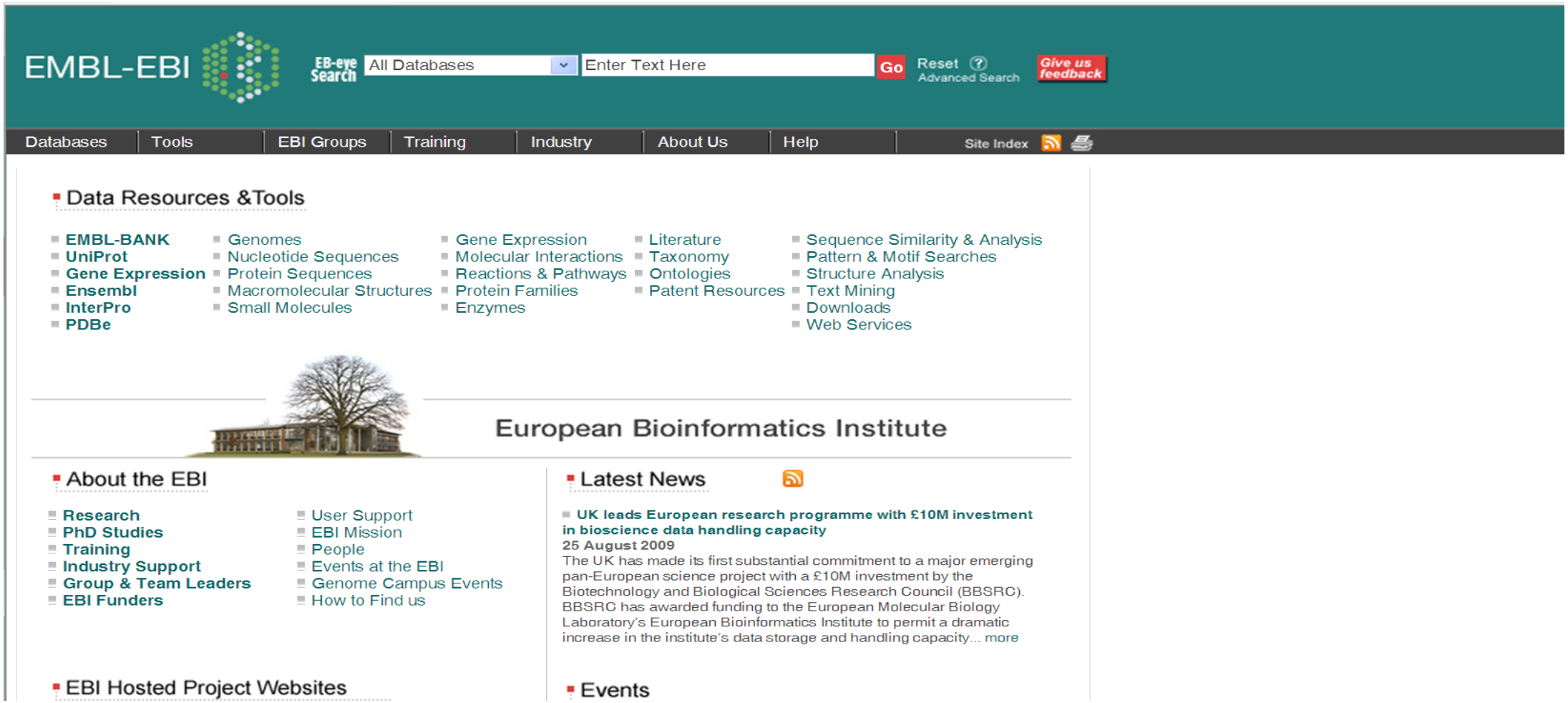
NCBI – all databases



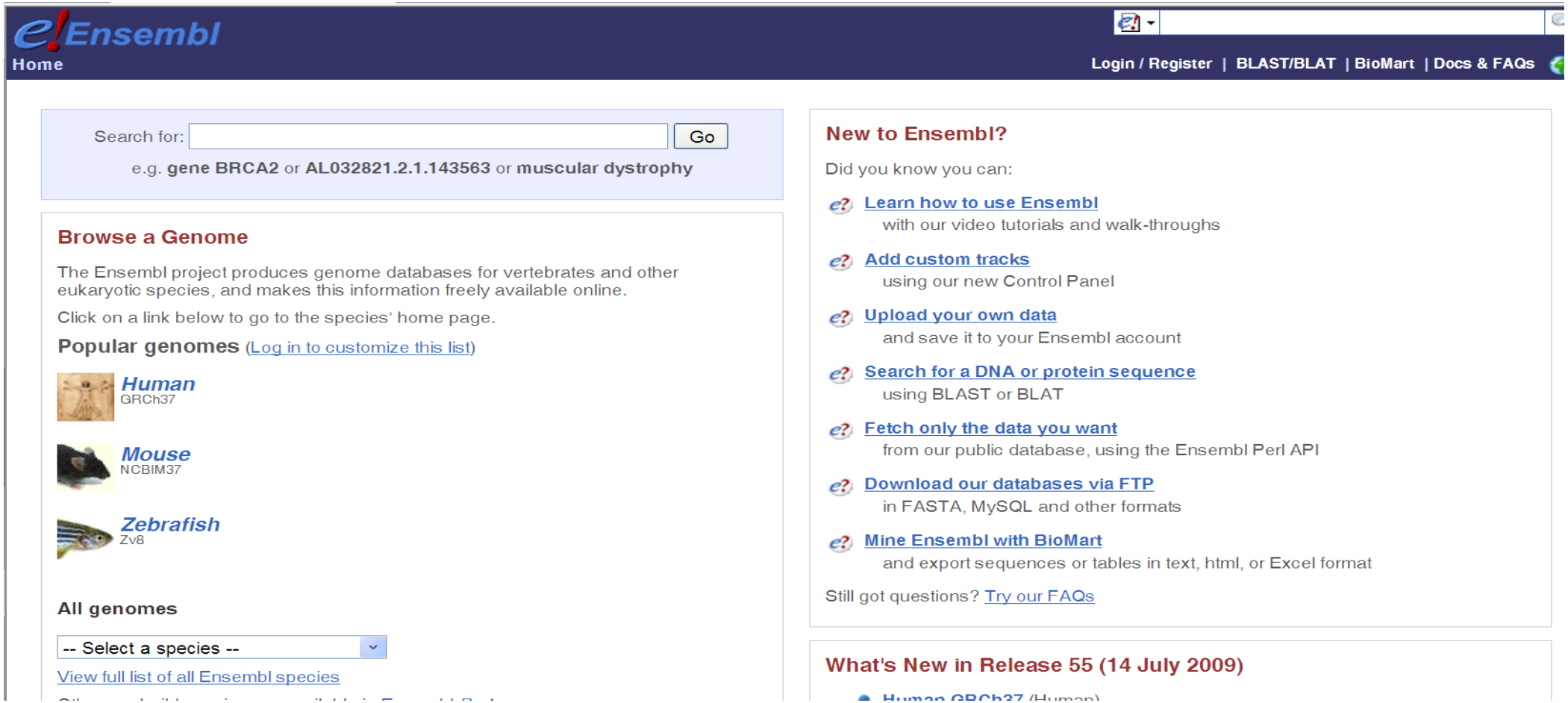
NCBI – all tools



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

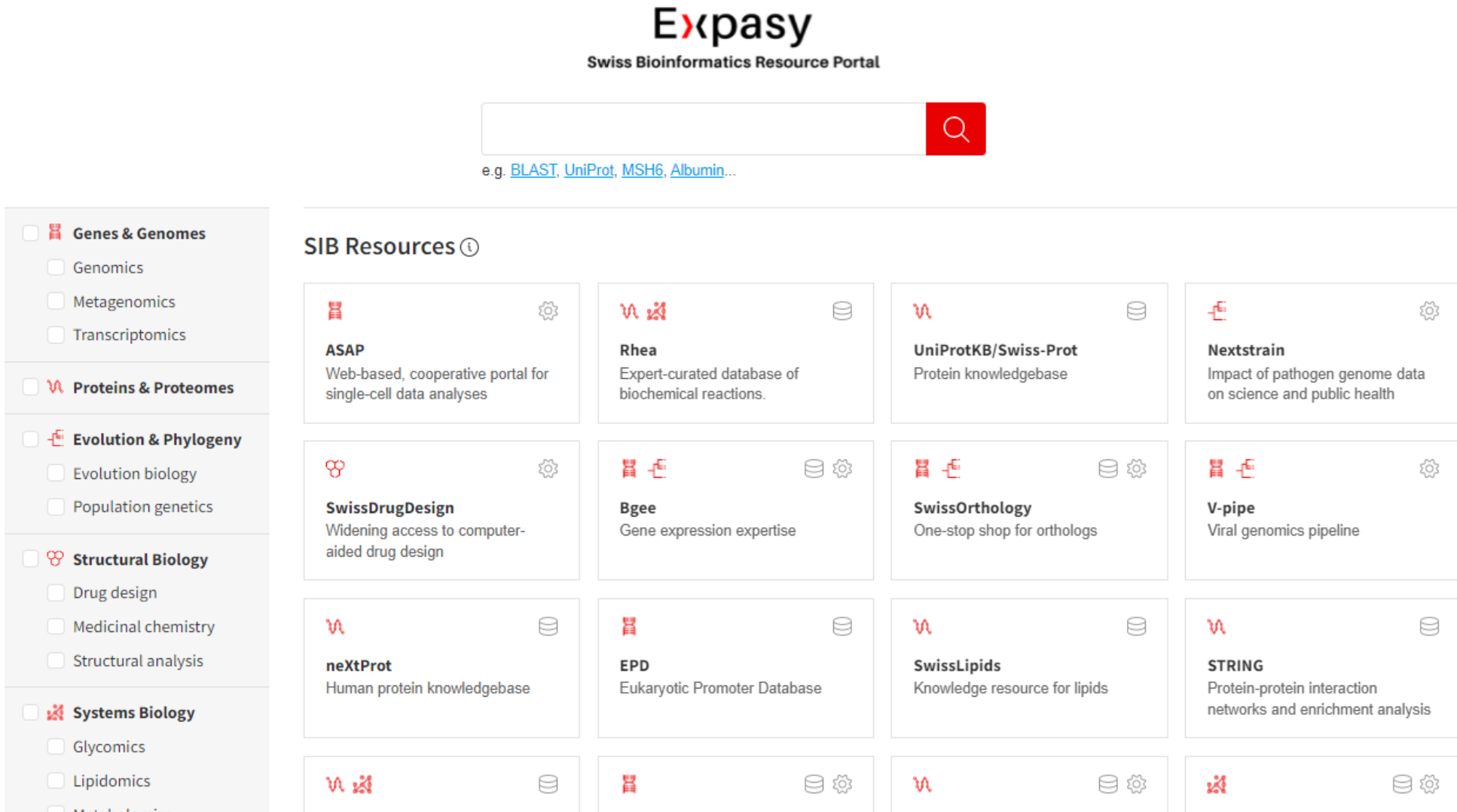


Ensembl - http://www.ensembl.org/index.html

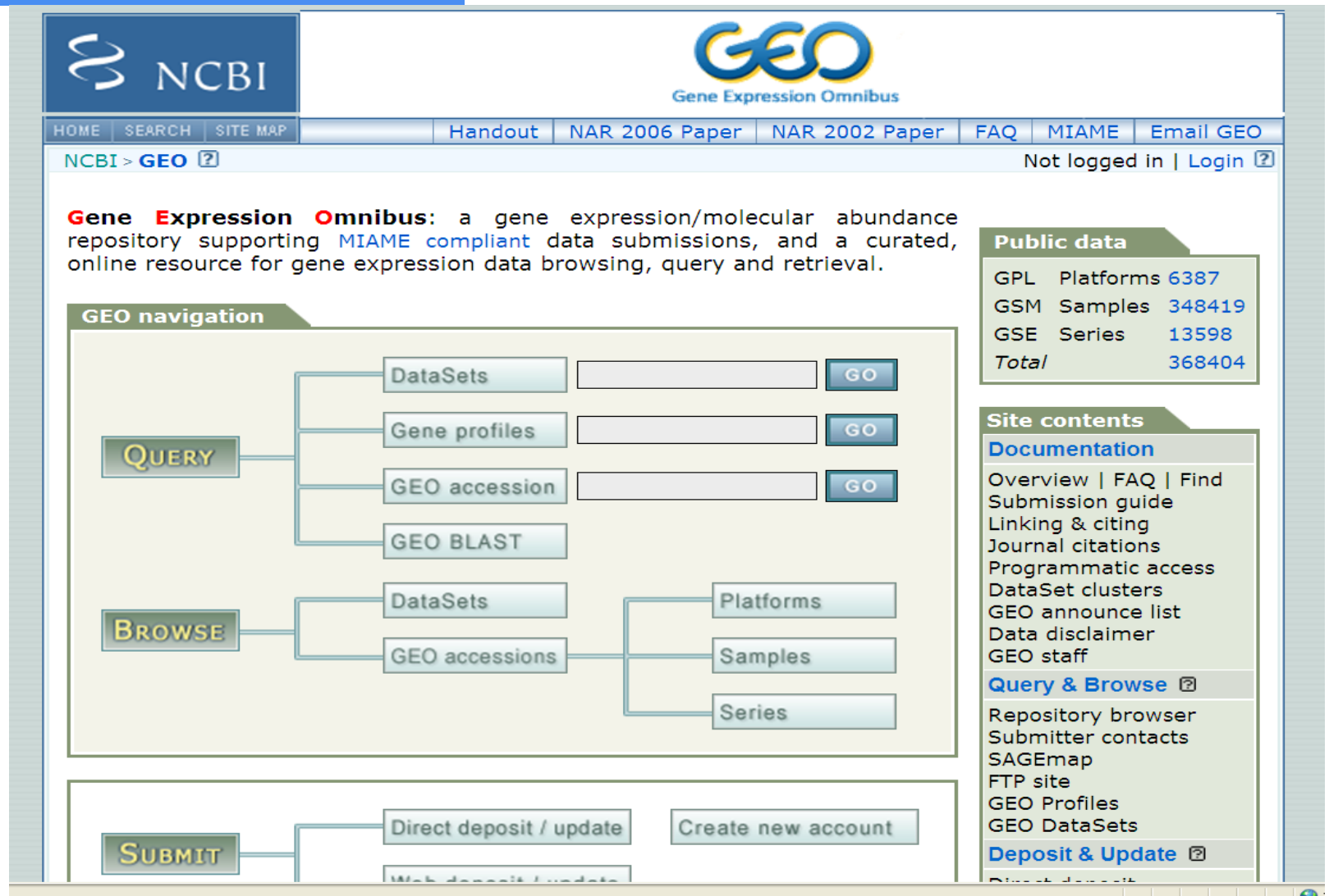


Expasy – Swiss Bioinformatics Resource Portal

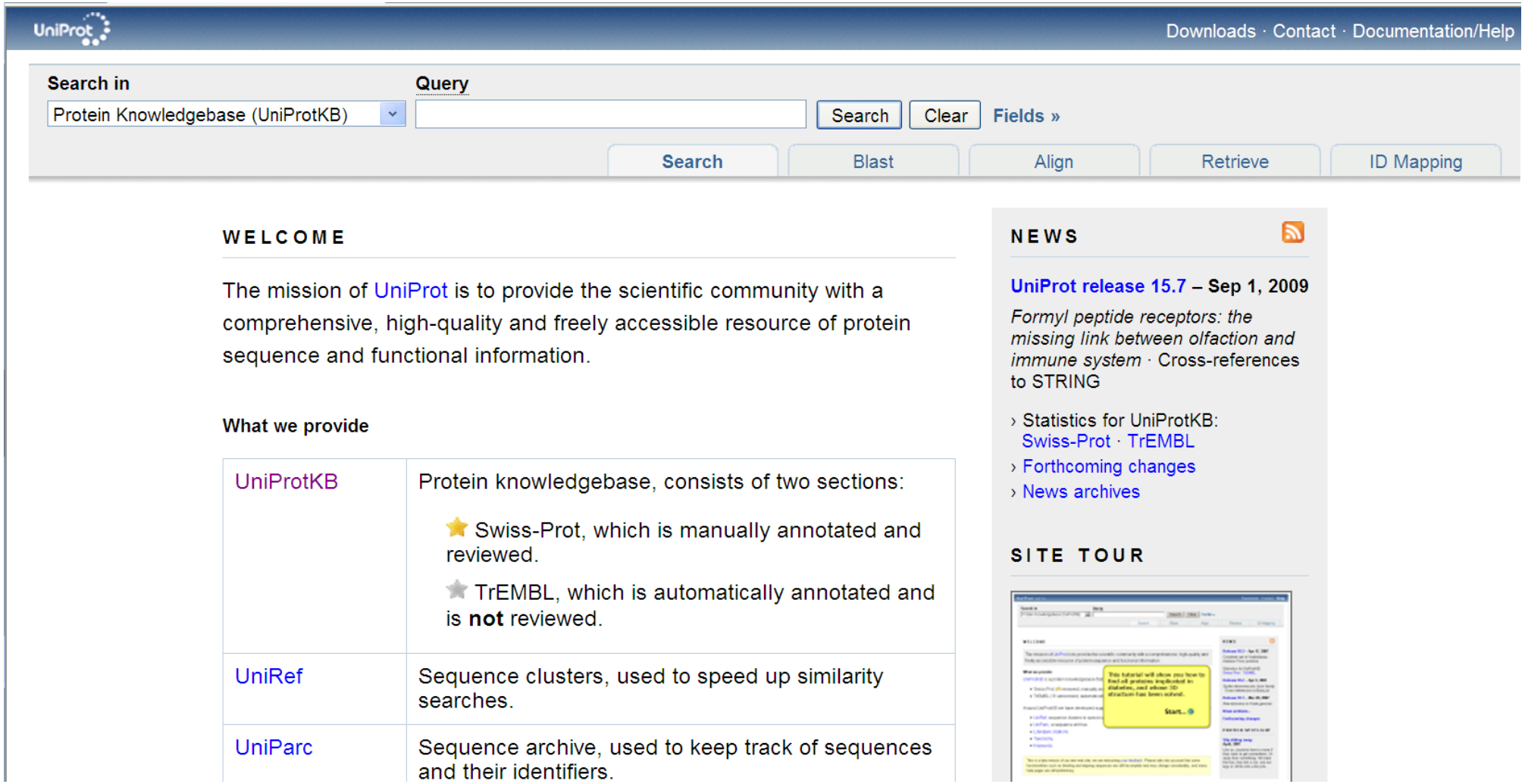
https://www.expasy.org/



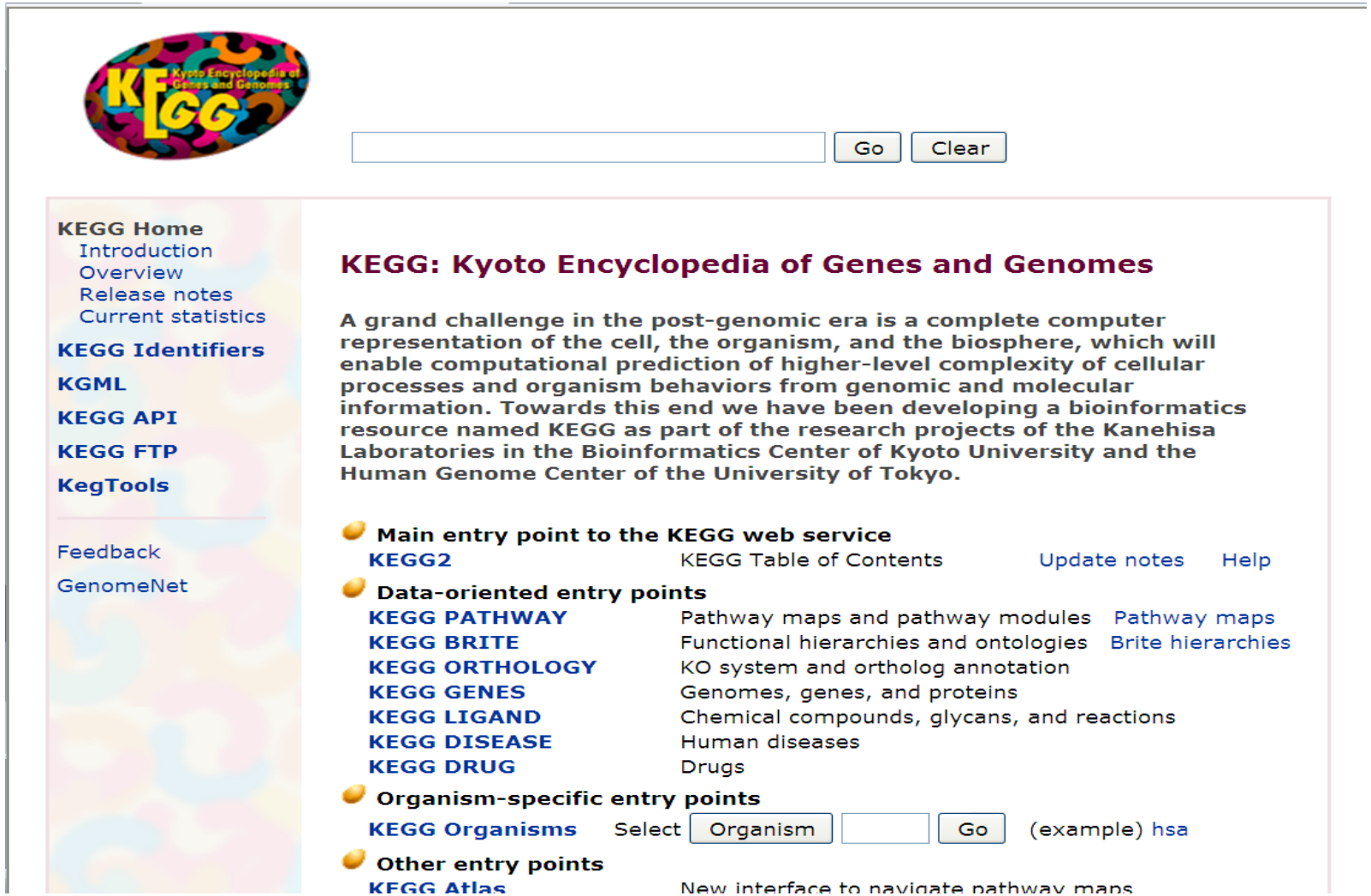
Gene Expression Resource: GEO - http://www.ncbi.nlm.nih.gov/geo/



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



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



Protein Structure Database: PDB - https://www.rcsb.org/

MyPDB Login

A MEMBER OF THE wwPDB

RCSB PDB PROTEIN DATA BANK

An Information Portal to Biological Macromolecular Structures

As of **Tuesday Sep 08, 2009** there are 60046 Structures | **PDB Statistics**

WHAT'S NEW | HELP | PRINT

PDB ID or keyword | Search | Advanced Search

- 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

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

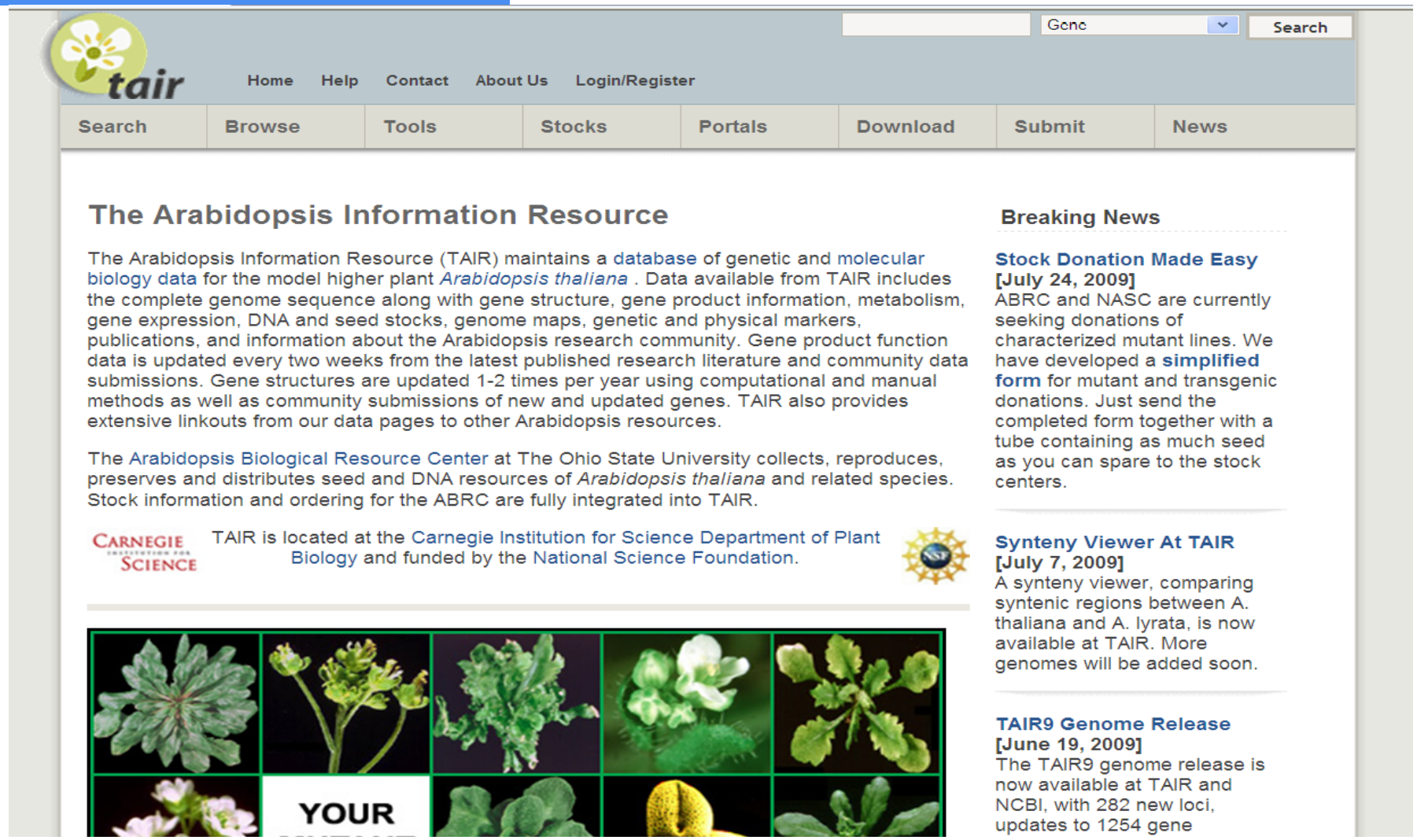
Improved Navigation of the RCSB PDB Website

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**

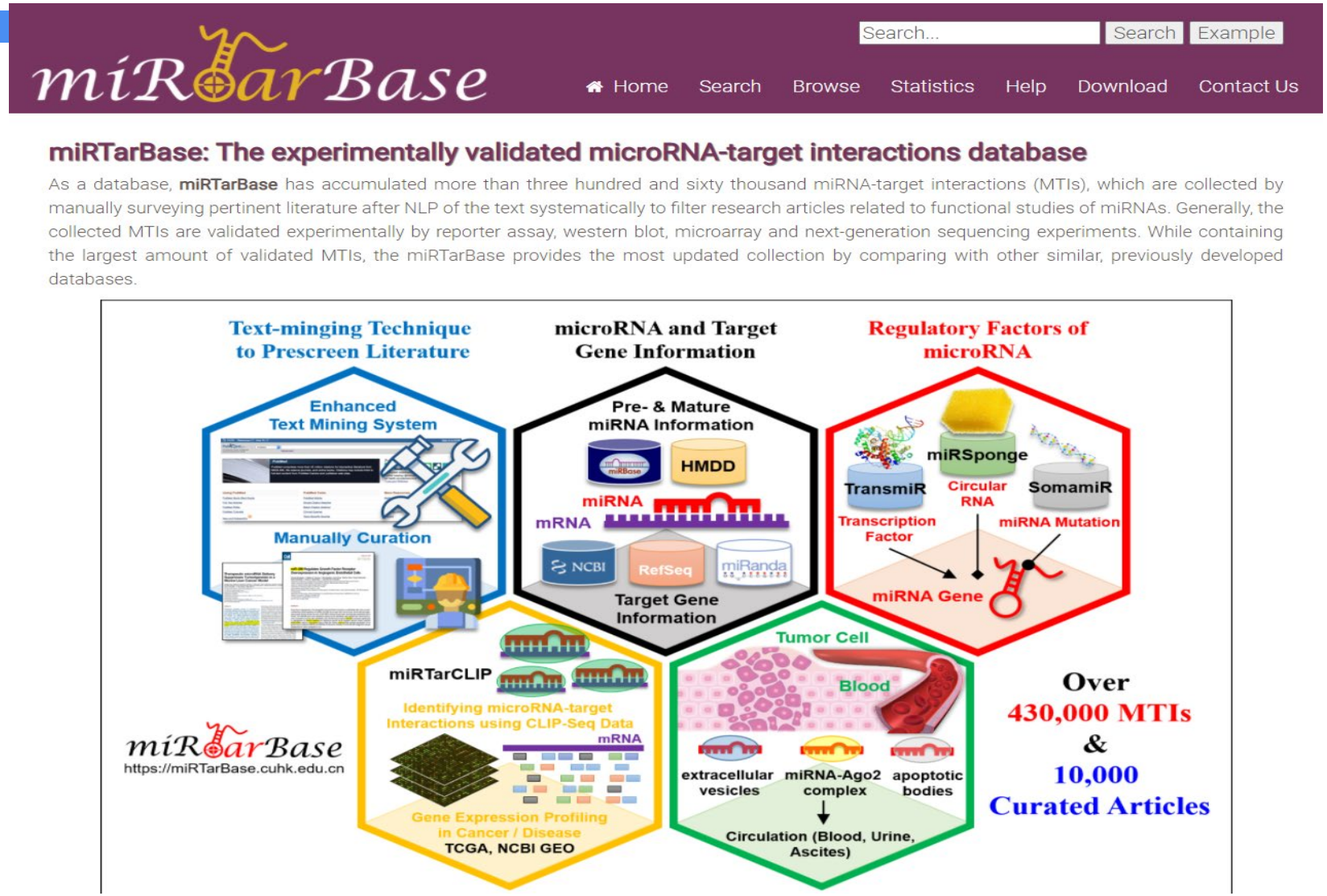
Plant-specific database: TAIR

http://www.arabidopsis.org/



miRTarBase – miRNA-target database

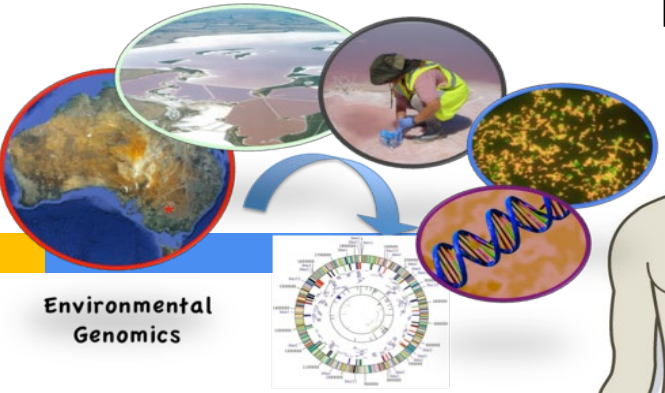
http://mirtarbase.cuhk.edu.cn/



dbPTM – Post-Translational Modifications

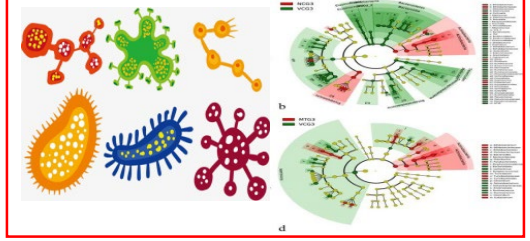
https://awi.cuhk.edu.cn/dbPTM/

Environmental & diet factors

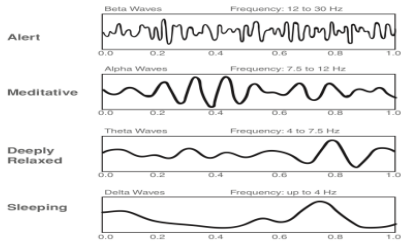


Environmental Genomics

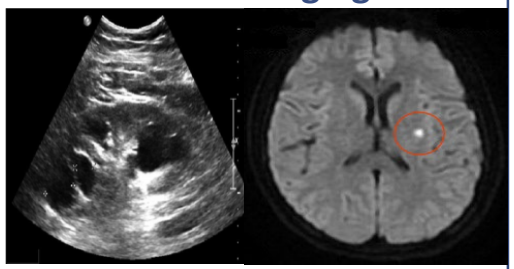
Microbiome



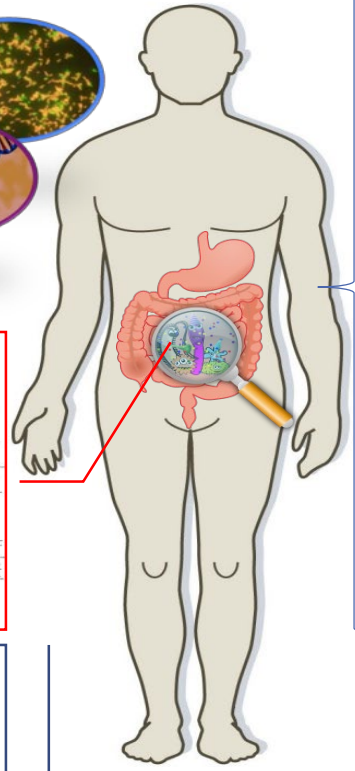
Physiological factors



Clinical imaging



Patient



Single Nucleotide Polymorphism (SNP)

DNA

DNA sequencing

RNA

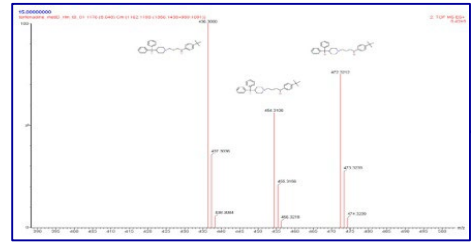
RNA sequencing
Microarray

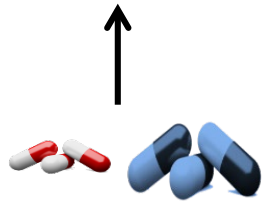
Protein

MS or MS/MS

SAAV
Single Amino Acid Variant

Metabolites

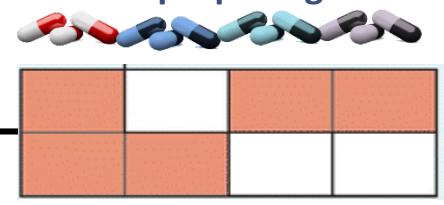




\+

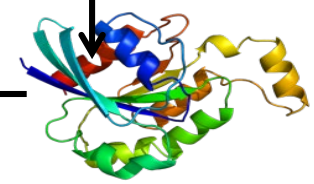
Standard care

Drug docking or repurposing



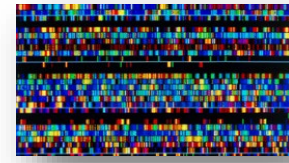
Drug targeting

Personalized signatures

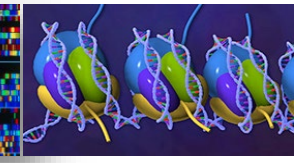


biomarkers

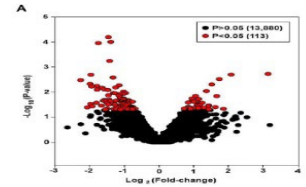
Genomics

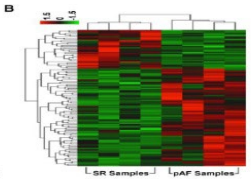


Epigenomics

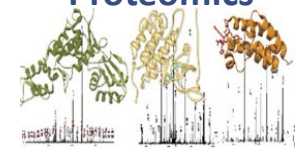


Transcriptomics

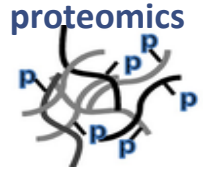




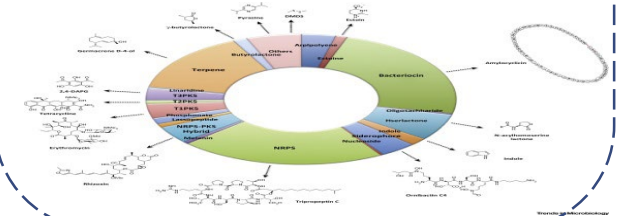
Proteomics



Phospho-proteomics



Metabolomics



What’s the aim of this course?