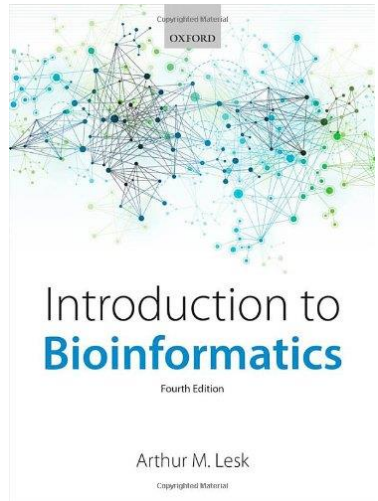
生物基因資訊科技簡介
Introduction to Bioinformatics



Instructor: Dr. Hsuan-Ting Chang
Department of Electrical Engineering
National Yunlin University of Science & Technology

1

Course description

- Text book:
  - Arthur M. Lesk, ***Introduction to Bioinformatics***, 4th Edition, Oxford Press, 2013 ，藝軒圖書出版社代理
  - **請購買正版書**(Legal copy only in classroom)

- Reference books:
  - Baxevanis, Ouellette, ***Bioinformatics***, 4th Edition, Wiley-Interscience, 2019，藝軒圖書出版社代理
  - Arthur M. Lesk, ***Introduction to Protein Science Architecture, Function, and Genomics*** , Oxford Press, 2004
  - Arthur M. Lesk, ***Introduction to Bioinformatics***, 5th Edition, Oxford Press, 2019

2

- Grading:
  - Midterm – 40%
  - Homework – 20%
  - Paper presentation – 40%
- Office: Electrical Hall – Room EN307
- Phone: 05-5342601 ext. 4263
- Email: htchang@yuntech.edu.tw
- Homepage: https://teacher.yuntech.edu.tw/htchang/

3

The Author

Arthur Lesk

Professor of Biochemistry and Molecular Biology, The Pennsylvania State University, USA

4

Contents (4th Ed.)

1: Introduction
2: Genome organization and evolution
3: Scientific publications and archives: media, content and access
4: Archives and information retrieval
5: Alignments and phylogenetic trees
6: Structural bioinformatics and drug discovery
7: Introduction to systems biology
8: Metabolic pathways
9: Gene expression and regulation

5

Contents (5th Ed.)

1: Introduction
2: From genetics to genomes
3: The panorama of life
4: Alignments and phylogenetic trees
5: Structural bioinformatics and drug discovery
6: Scientific publications and archives: media, content, access, and presentation
7: Artificial intelligence and machine learning
8: Introduction to systems biology
9: Metabolic pathways
10: Control of organization and organization of control

6

Chapter 1

Introduction

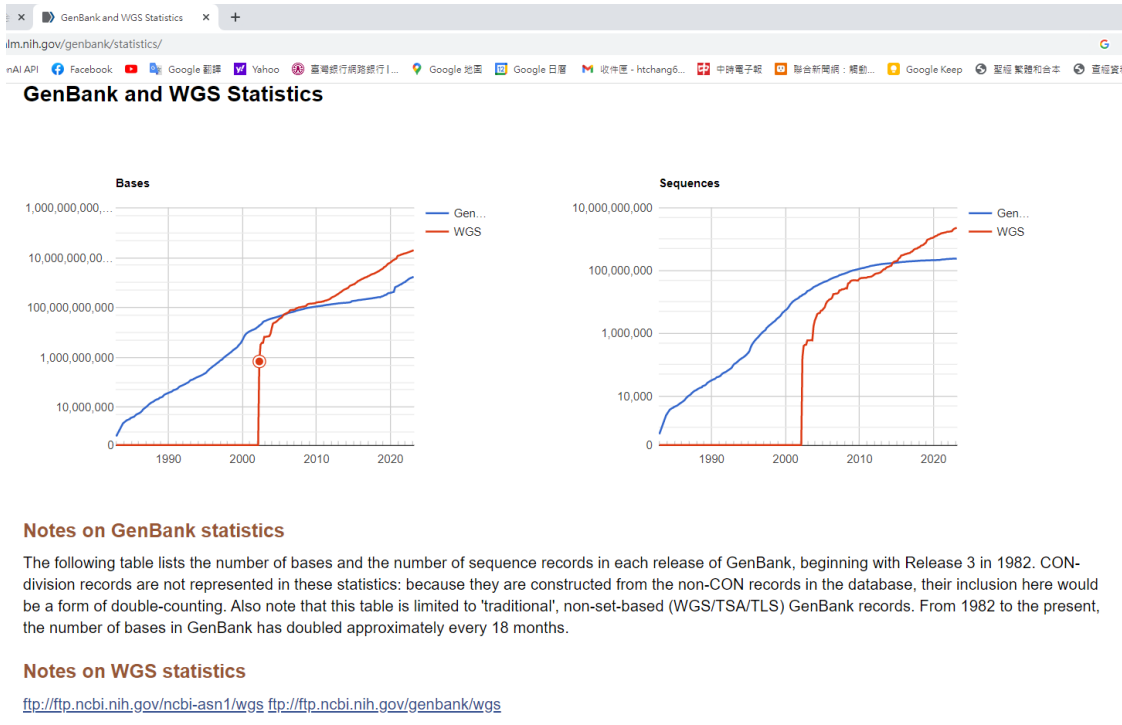
7

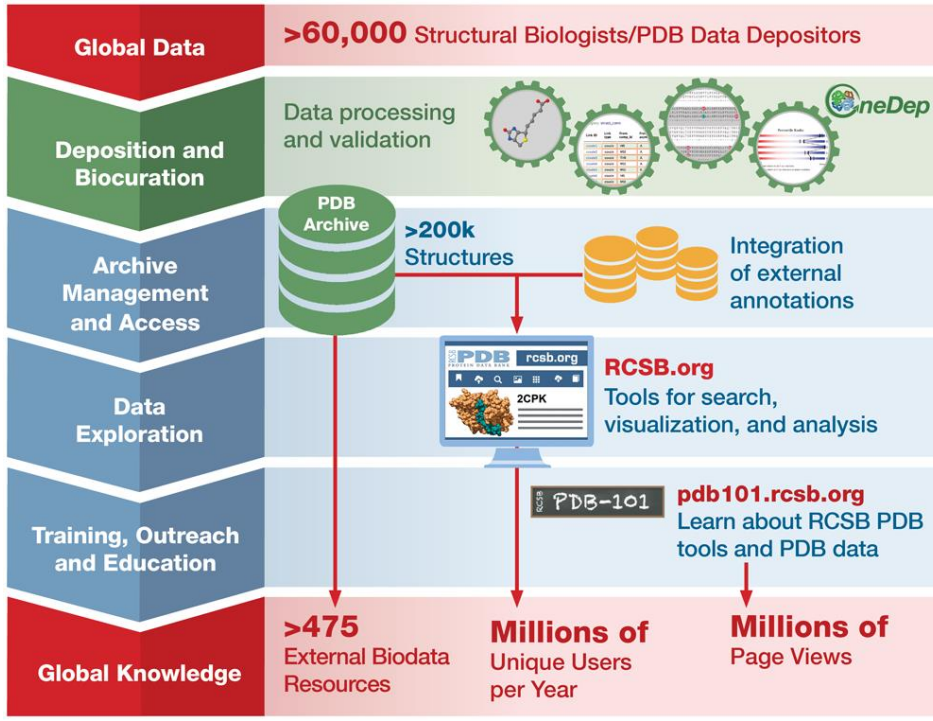
- Biology has traditionally been an observational rather than a deductive (演繹性) science.
- Modern genomic sequencing has converted biology into a deductive science
- Life does obey principles of physics and chemistry, but for now life is too complex
- Currently (2023) the nucleotide sequences databanks contain >2 Tbp, >246M sequences
- The database of *macromolecular* structures contains >20K entries, full 3-D coordinates of proteins

8

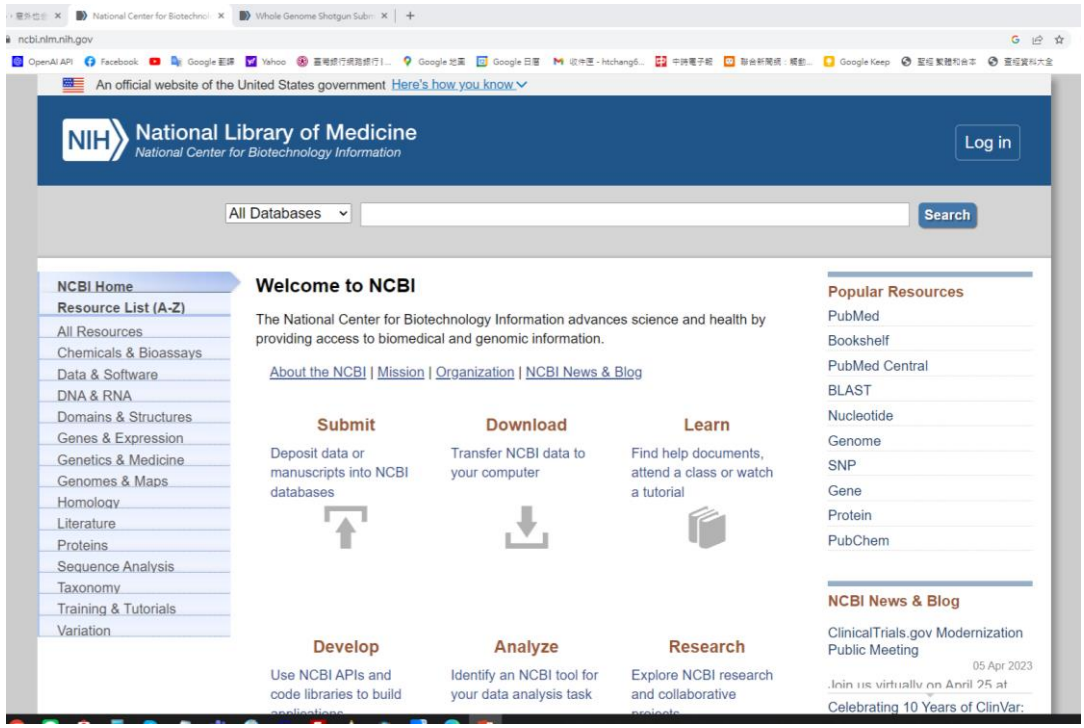
GeneBank Overview

- GenBank ® is the NIH genetic sequence database, an annotated collection of all ***publicly available*** DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42).
- GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI.
  - These three organizations exchange data on a ***daily*** basis.
- A GenBank release occurs ***every two months*** and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank.
  - Release notes for previous GenBank releases are also available.
  - GenBank growth statistics for both the traditional GenBank divisions and the Whole-genome sequencing (WGS) division are available from each release.

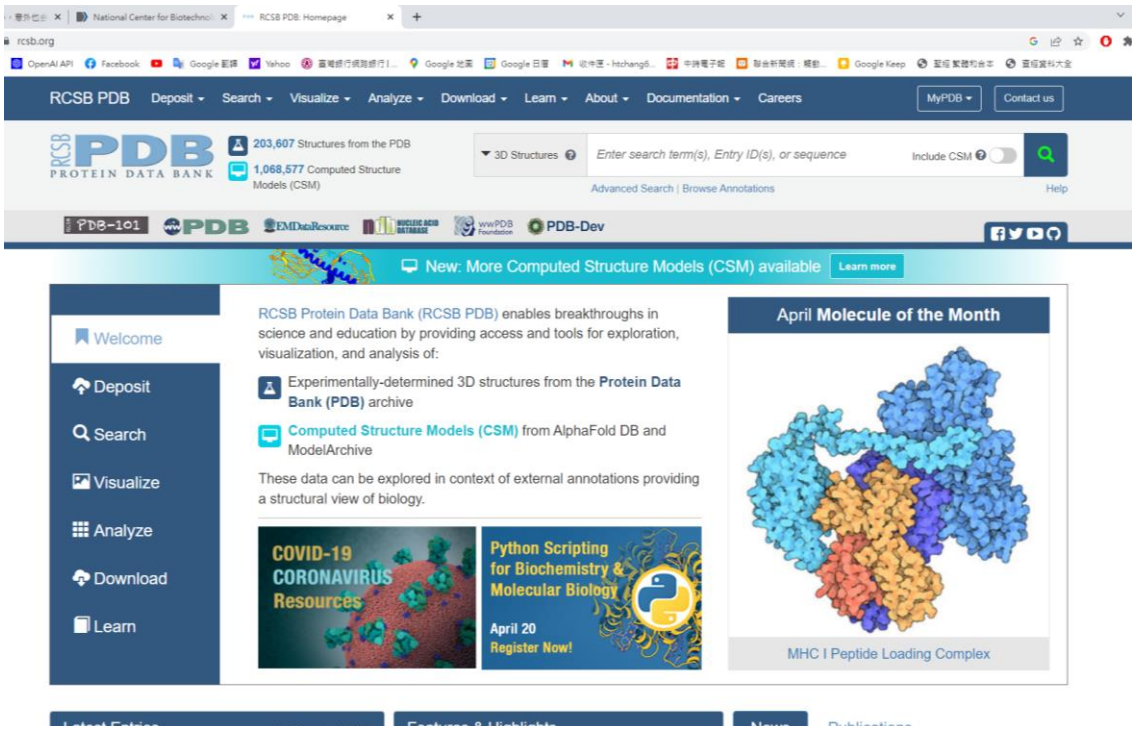



Global Data
>60,000 Structural Biologists/PDB Data Depositors
Deposition and Biocuration
Data processing and validation
OneDep
Archive Management and Access
PDB Archive
>200k Structures
Integration of external annotations
Data Exploration
rcsb.org
2CPK
RCSB.org
Tools for search, visualization, and analysis
Training, Outreach and Education
PDB-101
pdb101.rcsb.org
Learn about RCSB PDB tools and PDB data
Global Knowledge
>475 External Biodata Resources
Millions of Unique Users per Year
Millions of Page Views

11

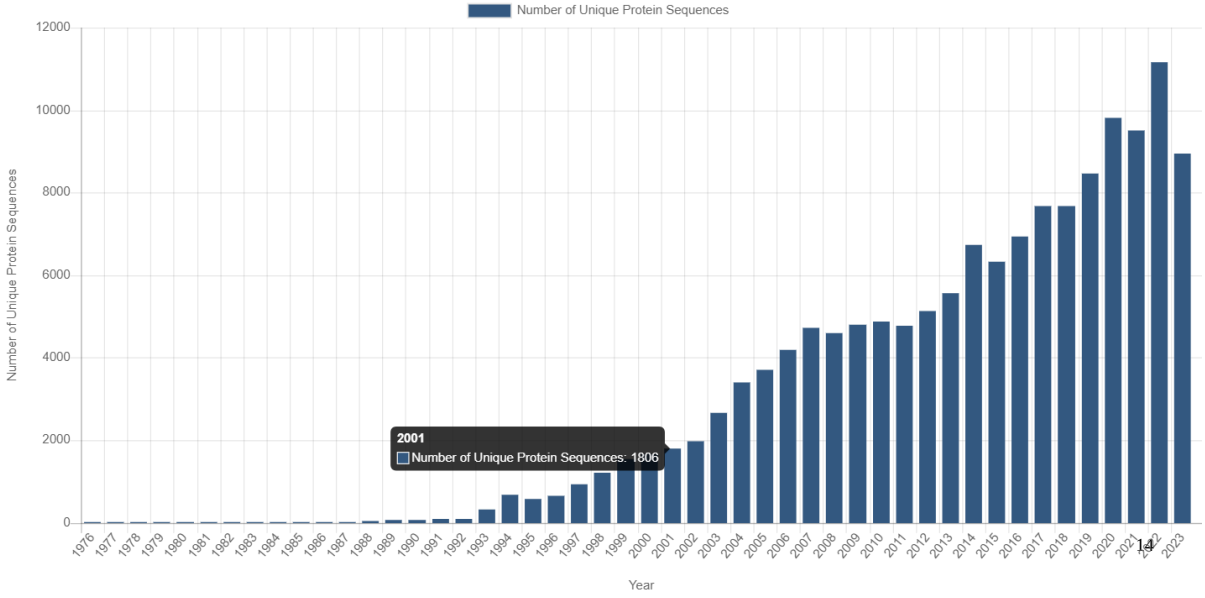

ncbi.nlm.nih.gov
An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
All Databases
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About the NCBI | Mission | Organization | NCBI News & Blog
Submit
Deposit data or manuscripts into NCBI databases
Download
Transfer NCBI data to your computer
Learn
Find help documents, attend a class or watch a tutorial
Develop
Use NCBI APIs and code libraries to build
Analyze
Identify an NCBI tool for your data analysis task
Research
Explore NCBI research and collaborative
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI News & Blog
ClinicalTrials.gov Modernization Public Meeting
05 Apr 2023
Join us virtually on April 25 at
Celebrating 10 Years of ClinVar:

12



13

PDB Statistics: Number of Unique Protein Sequences within Released PDB Structures (Annual) at Identity 95%



14

- Commensurately (同量的) ambitious goals scientist aim:
  1. Saw life clearly & saw it ***whole***. That is, to understand ***integrative aspects*** of the biology of organisms.
  2. To ***interrelate*** sequence, 3-D structure, interactions, and function of individual proteins, nucleic acids and protein-nucleic acid complex.
  3. To use data on contemporary (當代的) organisms as a basis for travel backward and forward in time – back to deduce events in evolutionary history, forward to greater deliberate scientific modification of biological systems.
  4. To support ***applications*** to medicine, agriculture and other scientific fields.

15

Life in Space and Time

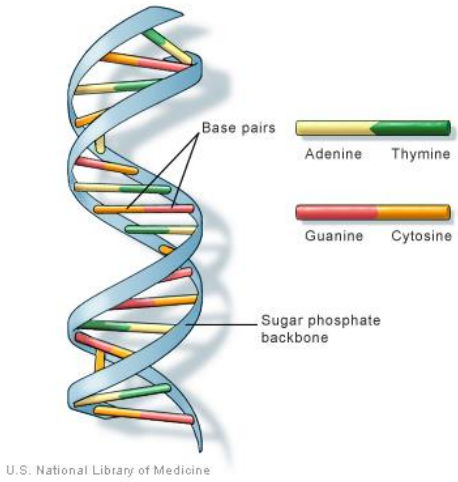
- Biological organism: a natural-occurring, self-reproducing device that effects controlled manipulations of matter, energy and information.
- Local ecosystems are stable until their environmental conditions change or they are invalid.
- Occupying each ecosystem are sets of species, which evolve by Darwinian selection or genetic drift.
- The generation of variants may arise from natural ***mutation***, or the recombination of genes in sexual reproduction, or direct gene transfer.

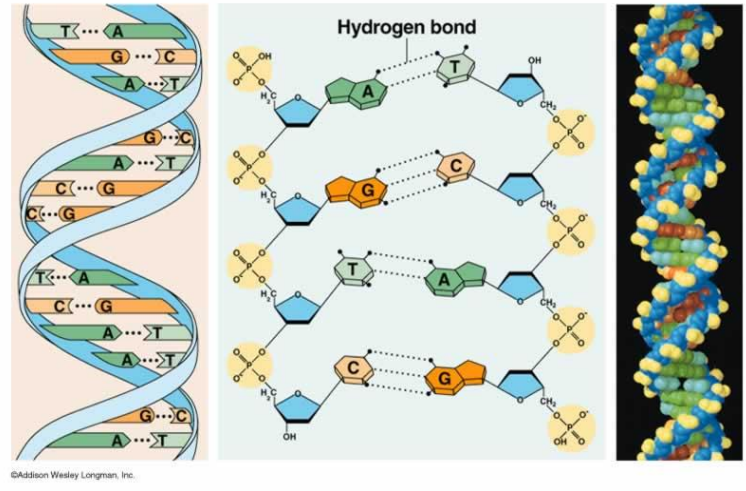
16

Dogmas: Central and Peripheral

- The blueprint for potential development and activity of any individual is the genetic material, DNA, or, in some viruses, RNA.
- DNA molecules are long, linear, chain molecules containing a message in a four-letter alphabet.
- Implicit in the structure of DNA are mechanisms for self-replication and for translation of genes into proteins.
- The double-helix, and its internal self-complementarity providing for accurate replication, are well known (Plate 1).

17





18

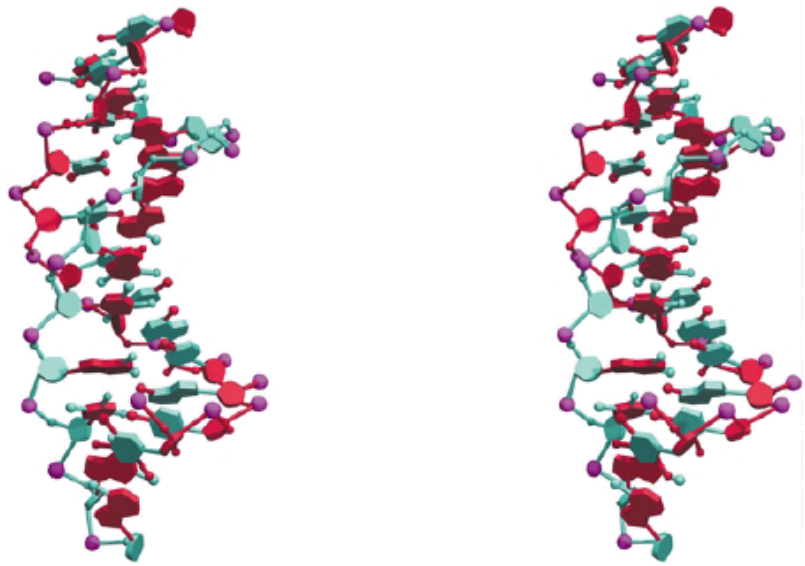


Plate I Double-helix of DNA. (See page 5.)

19

- Near-perfect replication is essential for stability of inheritance; but some imperfect replication, or mechanism for import of foreign genetic material, is also essential, else evolution could not take place in a sexual organisms.
- The strands in the double-helix are anti-parallel; directions along each strand are named 3' and 5'.
- In translation to protein, the DNA sequence is always read in the 5'→3' direction.

20

- Implementation of genetic information occurs, initially, through the synthesis of RNA and proteins.
- Proteins are the molecules responsible for much of the structure and activities of organisms.
- Both nucleic acid and proteins are long, linear chain molecules.
- The genetic ‘code’ is in fact a cipher: successive triplets of letters from the DNA sequence specify successive amino acids; stretches of DNA sequences encipher amino acid sequences of proteins.

21

TABLE 1.2

The genetic code mapping codons to amino acids.

First position	*Second position* G	A	C	U	*Third position*
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	STOP	Ser	Leu	G
	STOP	STOP	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	U

AUG
Start codon

22

TABLE 1.1

The twenty amino acids commonly found in proteins.

	One-letter code	*Three-letter code*	*Name*
1	A	Ala	Alanine
2	C	Cys	Cysteine
3	D	Asp	Aspartic Acid
4	E	Glu	Glutamic Acid
5	F	Phe	Phenylalanine
6	G	Gly	Glycine
7	H	His	Histidine
8	I	Ile	Isoleucine
9	K	Lys	Lysine
10	L	Leu	Leucine
11	M	Met	Methionine
12	N	Asn	Asparagine
13	P	Pro	Proline
14	Q	Gln	Glutamine
15	R	Arg	Arginine
16	S	Ser	Serine
17	T	Thr	Threonine
18	V	Val	Valine
19	W	Trp	Tryptophan
20	Y	Tyr	Tyrosine

23

- In most organisms not all of the DNA expressed as proteins or RNAs. Some regions of the DNA sequence are devoted to control mechanisms, and a substantial amount of the genome of higher organisms appears to be ‘junk’.
  - We do not yet understand its function.
  - Garbage we throw away, but junk we keep
- The amino acid sequence of a protein dictates its 3-D structure.
- For each natural amino acid sequence, there is a unique stable native state that under proper conditions are adopted spontaneously.

24

- The translation of DNA sequences to amino acid sequences is very simple to describe logically; its is specified by the ***genetic code***.
- The folding of the polypeptide chain into a precise 3-D structure is very difficult to describe logically.
- The functions of proteins depend on their adopting the native 3-D structure.

25

- Paradigm
  - DNA sequence determines protein sequence
  - Protein sequence determines protein structure
  - Protein structure determines protein function
  - Regulatory mechanisms deliver the right amount of the right function to the right place at the right time

- This paradigm does not include levels higher than the molecular level of structure and organization.

26

Observables & Data Archives

- Databanks includes:
  - (1) an archive of information
  - (2) a logical organization or `structure' of that information (schema)
  - (3) tools to gain access to it
- Contain nucleic acid and protein sequences, macromolecular structures and functions, expression patterns and networks of metabolic pathways and control cascades.

27

- They include:
  - Archival databanks of biological information
  - Derived databanks
  - Bibliographic databanks
  - Databanks of web sites

28

Archival databanks of biological information

- DNA & protein sequences, including annotation
- Variations, such as compilation (編輯) of haplotypes (連鎖不平衡)
- Nucleic acid and protein structures, including annotation
- Databanks focused on organisms, including genome databases
- Databanks of protein expression patterns
- Databanks of metabolic pathways
- Databanks of interaction patterns and regulatory networks

29

Derived databanks

- The mechanism of access to a databank is the set of tools for answering questions such as
  - Does the databank contain the information I require?
  - How can I assemble selected information from the databank in a useful form?
  - Indices of databanks are useful in asking “What can I find some specific piece of information?”

30

- A databank without effective modes of access is merely a data graveyard.
- Possible kinds of database queries:
  - Given a sequence, or fragment of a sequence, find sequences in the database that are similar to it.
  - Given a protein structure, or fragment, find protein structures in the database that are similar to it.
  - Given a sequence of a protein of unknown structure, find structures in the database that adopt similar 3-D structures.
  - Given a protein structure, find sequences in the databank that correspond to similar structures.

31

- One wishes to study relationships between information contained in separate databank.
  - This requires links that facilitate simultaneous access to several databanks.
- Research in databank interactivity – how can databanks ‘talk to one another’, without too great sacrifice of the freedom of each one to structure its own data.

32

Information flow in bioinformatics

- Reorganization of data may involve:
  - Simply integrating the new entry into a general or specialized search engine
  - Extracting useful subsets of the data
  - Deriving new types of information from the original data
  - Recombining data in different ways
  - Re-annotating the data, including provision of different constellations (群集) of links.

33

- Organisms are composed of ***cells***. Every cell is an intimate localized ecosystem, not isolated from its environment but interacting within specific and controlled ways.
- Life is extended not only in ***space*** but in ***time***
- We must try to read the past in contemporary (現代的) genomes.

34

Curation (保存), Annotation, & Quality Control

- Databank entries comprise raw experimental results, and supplementary information, or annotations. Each of these has its own sources of error.
  - The quality of the data depends on the art of experiments.
- Annotations include information about the source of the data and the methods used to determine them.
  - Identify the investigators responsible
  - Cite relevant publications
  - Provide links to related information in other databanks
  - In databanks, annotations include feature tables: list of segments of the sequences that have biological significance.

35

The World Wide Web

- Browser
- Links
- Search engine
- Bookmarks or my favorite

- Enter information, & launch a program that returns within your session

36

Computers & Computer Science

- Bioinformatics would not be possible without advances in computers:
  - Fast & high-capacity storage media
  - Information retrieval & analysis programs
  - Facilities of computer networks & WWW for distributing information

37

- Computer science:
  - Analysis of algorithms: an algorithm is a complete and precise specification of a method for solving a problem
  - Data structures, & information retrieval: data organization & user interface
  - Software engineering: high level languages such as C, C++, PERL, Python.

38

Biological Classification & Nomenclature (專門術語)

- Living things are divided into units called ***species*** – groups of similar organisms with a common gene pool.
- Linnaeus classified living things according to a hierarchy: Kingdom (界), Phylum (門), Class (綱), Order (目), Family (科), Genus (屬) and Species (種).
- For identification it generally suffices to specify the binomial: Genus and Species
  - *Homo sapiens* for human
  - *Drosophila melanogaster* for fruit fly

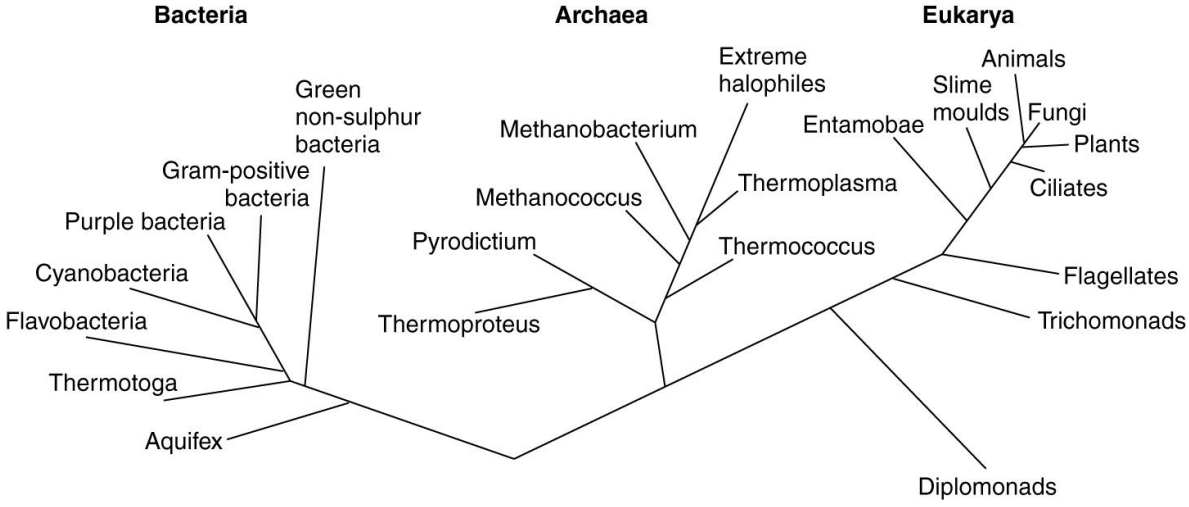
39

- Originally the Linnaean system was only a classification based on observed similarities.
  - Characteristics derived from a common ancestor are called *homologous*
  - Other apparently similar characteristics may have arisen independently by *convergent evolution*.
  - Conversely, truly homologous characters may have *diverged* to become very *dissimilar* in structure and function.

40

- On the basis of 155 rRNAs, Woese divided living things more fundamentally into three Domains (a level *above* kingdom in the hierarchy):
  - Bacteria(細菌), Archaea (古生菌), and Eukarya (Fig. 1.2)
- Bacteria and archaea are prokaryotes (原核生物); their cells do not contain nuclei.
- We ourselves are Eukarya (真核生物) – organisms containing cells with nuclei, including yeast and all multicellular organisms.

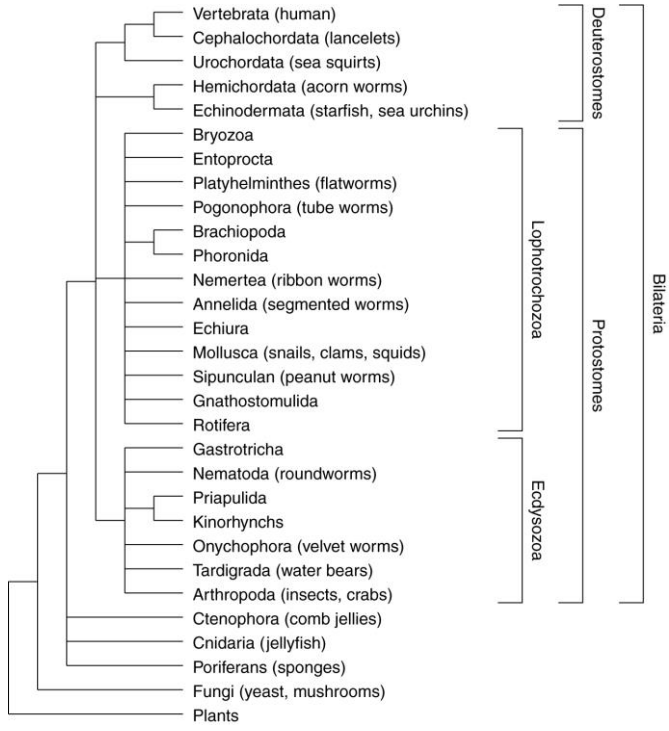
41



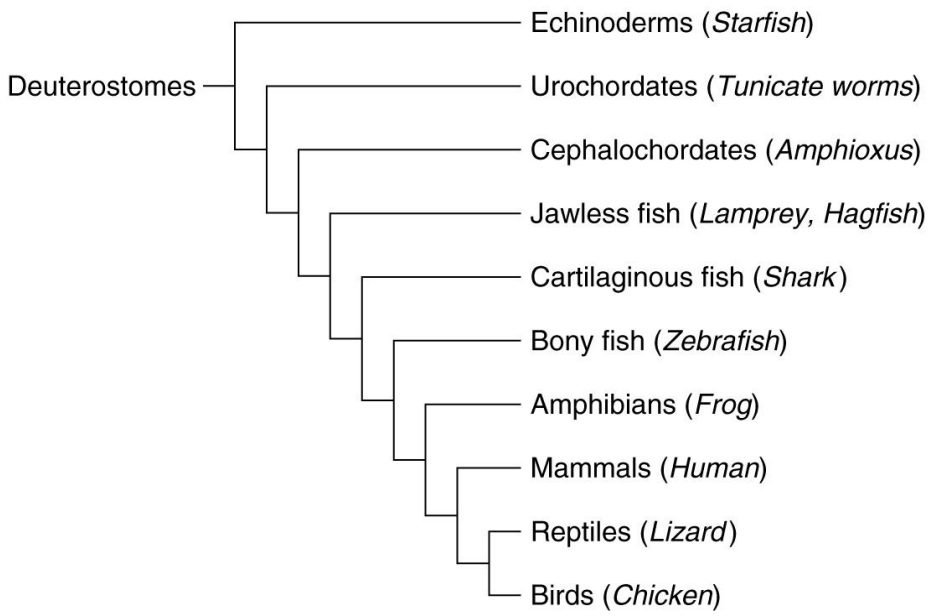
42

- Archaea:
  - The obvious differences in lifestyle
  - The absence of a nucleus
  - In some ways more related on a molecular level to eukarya than to bacteria
- It is likely that the archaea are the *closest* living organisms to the root of the tree of life.
- Figures 1.3 and 1.4. Deuterostomes (後口動物)

43



44

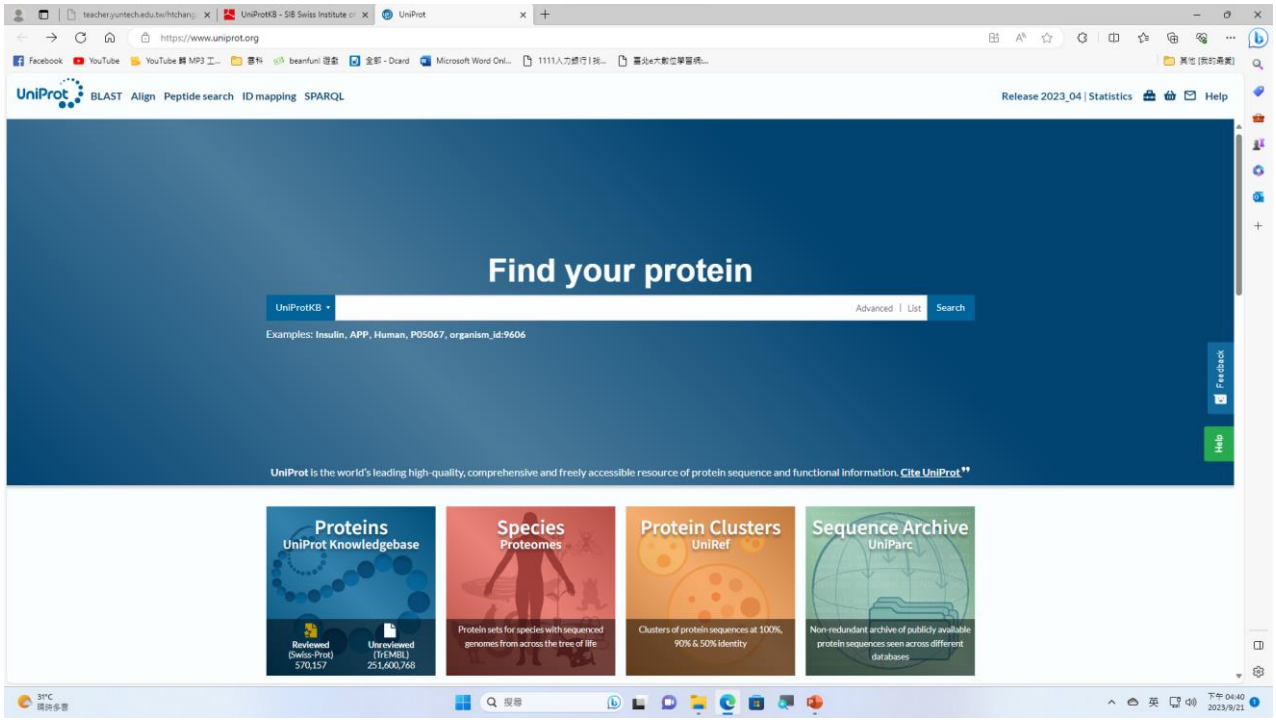


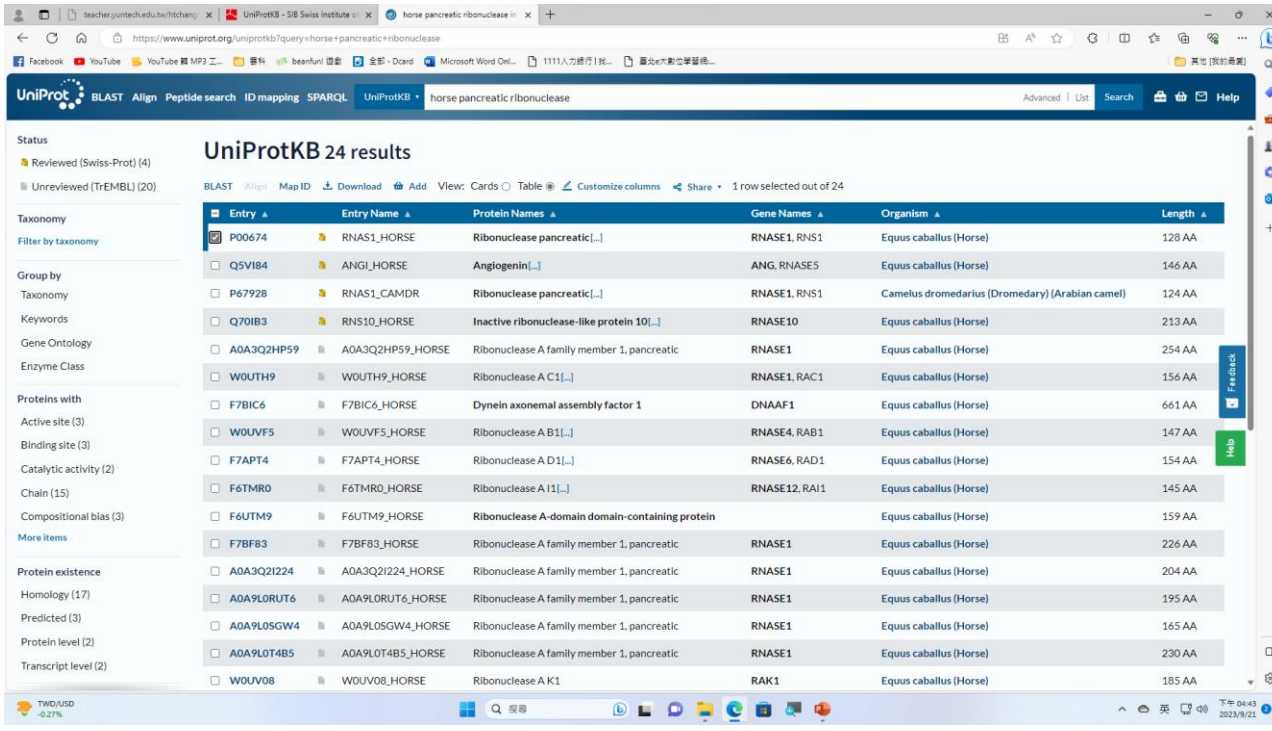
45

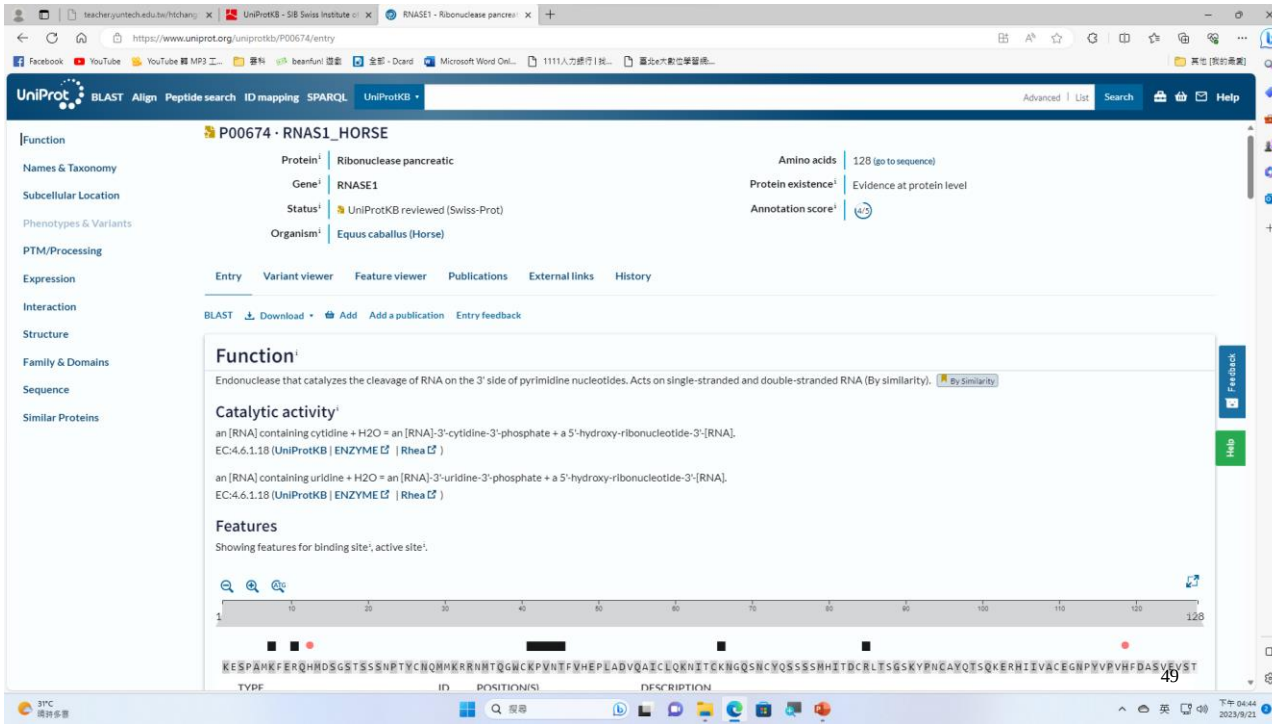
Case study

- Retrieve the amino acid sequence of horse pancreatic ribonuclease
- Use the ExPASy server at the Swiss Institute for Bioinformatics: https://www.expasy.org
- *Databases → UniProt → Type "horse pancreatic ribonuclease" in Query → Select P00674*

46

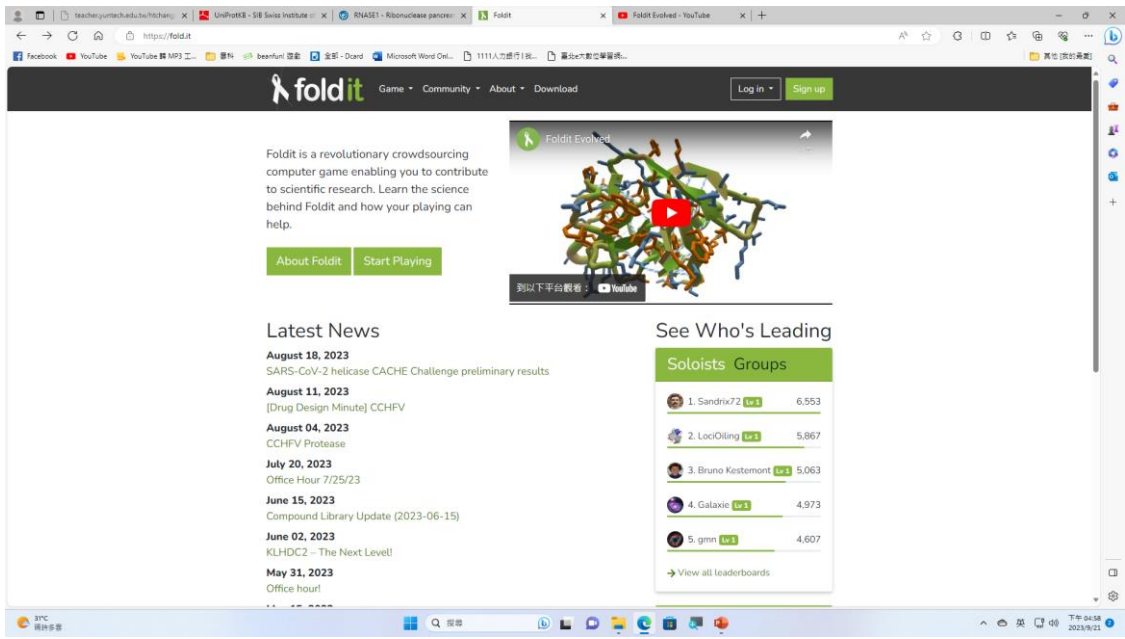

UniProt
BLAST Align Peptide search ID mapping SPARQL
Release 2023_04 | Statistics Help
Find your protein
UniProtKB
Advanced | List
Search
Examples: Insulin, APP, Human, P05067, organism_id:9606
UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt
Proteins
UniProt Knowledgebase
Reviewed (Swiss-Prot) 570,157
Unreviewed (TrEMBL) 251,600,768
Species
Proteomes
Protein sets for species with sequenced genomes from across the tree of life
Protein Clusters
UniRef
Clusters of protein sequences at 100%, 90% & 50% identity
Sequence Archive
UniParc
Non-redundant archive of publicly available protein sequences seen across different databases
Feedback
Help


UniProt
BLAST Align Peptide search ID mapping SPARQL
UniProtKB
horse pancreatic ribonuclease
Advanced | List
Search
Help
Status
Reviewed (Swiss-Prot) (4)
Unreviewed (TrEMBL) (20)
Taxonomy
Filter by taxonomy
Group by
Taxonomy
Keywords
Gene Ontology
Enzyme Class
Proteins with
Active site (3)
Binding site (3)
Catalytic activity (2)
Chain (15)
Compositional bias (3)
More items
Protein existence
Homology (17)
Predicted (3)
Protein level (2)
Transcript level (2)
UniProtKB 24 results
BLAST Align Map ID Download Add View: Cards Table Customize columns Share 1 row selected out of 24
Entry | Entry Name | Protein Names | Gene Names | Organism | Length
P00674 RNAS1_HORSE Ribonuclease pancreatic[...] RNASE1, RNS1 Equus caballus (Horse) 128 AA
Q5VI84 ANGI_HORSE Angiogenin[...] ANG, RNASE5 Equus caballus (Horse) 146 AA
P67928 RNAS1_CAMDR Ribonuclease pancreatic[...] RNASE1, RNS1 Camelus dromedarius (Dromedary) (Arabian camel) 124 AA
Q70IB3 RNS10_HORSE Inactive ribonuclease-like protein 10[...] RNASE10 Equus caballus (Horse) 213 AA
A0A3Q2HP59 A0A3Q2HP59_HORSE Ribonuclease A family member 1, pancreatic RNASE1 Equus caballus (Horse) 254 AA
W0UTH9 W0UTH9_HORSE Ribonuclease A C1[...] RNASE1, RAC1 Equus caballus (Horse) 156 AA
F7BIC6 F7BIC6_HORSE Dynein axonemal assembly factor 1 DNAAF1 Equus caballus (Horse) 661 AA
W0UVF5 W0UVF5_HORSE Ribonuclease A B1[...] RNASE4, RAB1 Equus caballus (Horse) 147 AA
F7APT4 F7APT4_HORSE Ribonuclease A D1[...] RNASE6, RAD1 Equus caballus (Horse) 154 AA
F6TMR0 F6TMR0_HORSE Ribonuclease A I1[...] RNASE12, RAI1 Equus caballus (Horse) 145 AA
F6UTM9 F6UTM9_HORSE Ribonuclease A-domain domain-containing protein Equus caballus (Horse) 159 AA
F7BF83 F7BF83_HORSE Ribonuclease A family member 1, pancreatic RNASE1 Equus caballus (Horse) 226 AA
A0A3Q2I224 A0A3Q2I224_HORSE Ribonuclease A family member 1, pancreatic RNASE1 Equus caballus (Horse) 204 AA
A0A9L0RUT6 A0A9L0RUT6_HORSE Ribonuclease A family member 1, pancreatic RNASE1 Equus caballus (Horse) 195 AA
A0A9L0SGW4 A0A9L0SGW4_HORSE Ribonuclease A family member 1, pancreatic RNASE1 Equus caballus (Horse) 165 AA
A0A9L0T4B5 A0A9L0T4B5_HORSE Ribonuclease A family member 1, pancreatic RNASE1 Equus caballus (Horse) 230 AA
W0UV08 W0UV08_HORSE Ribonuclease A K1 RAK1 Equus caballus (Horse) 185 AA
Feedback
Help



49

An Interesting Project: http://fold.it/portal/



Foldit is a revolutionary new computer game enabling *you* to contribute to important scientific research.

Webpage description about this project – https://fold.it/

51

Page Contents:

- What is protein folding?
- Why is this game important?
- Foldit Scientific Publications
- News Articles about Foldit
- News Articles about Rosetta
- Rosetta@Home Screensaver
- Community Rules
- Privacy Policy

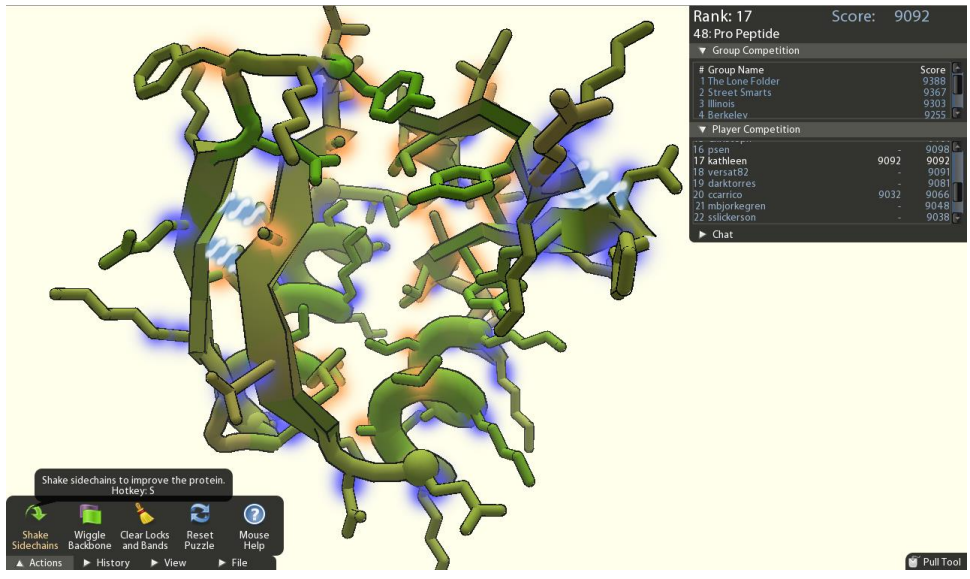
52

What is a protein?

- Proteins are the workhorses in every cell of every living thing.
- Your body is made up of trillions of cells, of all different kinds: muscle cells, brain cells, blood cells, and more.
- Inside those cells, proteins are allowing your body to do what it does: break down food to power your muscles, send signals through your brain that control the body, and transport nutrients through your blood.
- Proteins come in thousands of different varieties, but they all have a lot in common. For instance, they're made of the same stuff: every protein consists of a long chain of joined-together amino acids.

53

Folded up Puzzle 48



54

What are amino acids?

- Amino acids are small molecules made up of ***atoms of carbon, oxygen, nitrogen, sulfur, and hydrogen***.
- To make a protein, the amino acids are joined in an ***unbranched*** chain, like a line of people holding hands.
- Just as the line of people has their legs and feet "hanging" off the chain, each amino acid has a small group of atoms (called a ***sidechain***) sticking off the main chain (***backbone***) that connects them all together.

55

- There are ***20 different kinds*** of amino acids, which differ from one another based on what atoms are in their sidechains.
- These 20 amino acids fall into different groups based on their ***chemical properties***: acidic or alkaline (鹼), hydrophilic (water-loving) or hydrophobic (greasy).

56

What shape will a protein fold into?

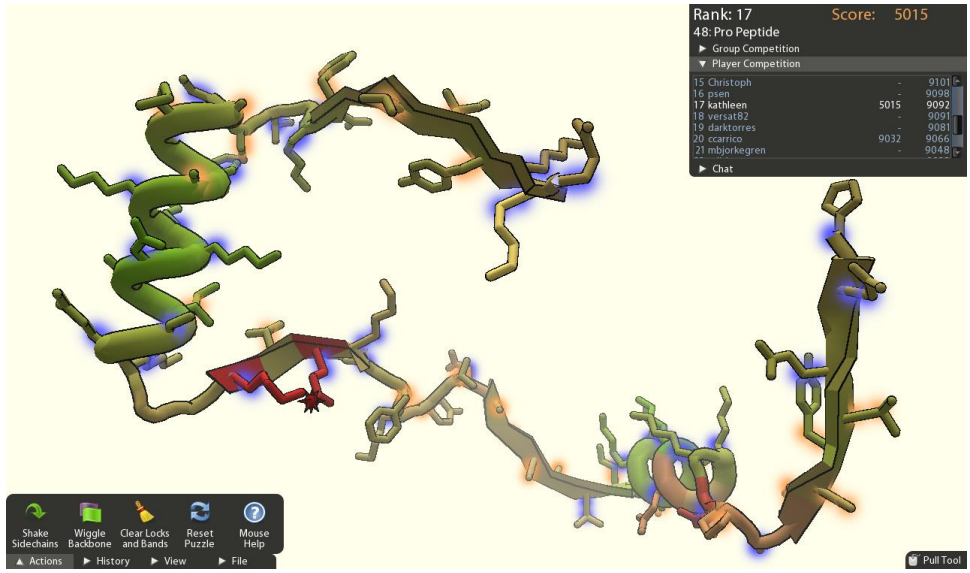
- Even though proteins are just a long chain of amino acids, they ***don't like to stay stretched out in a straight line***.
- The protein ***folds up to make a compact blob***, but as it does, it *keeps some amino acids near the center of the blob, and others outside*; and it keeps some pairs of amino acids close together and others far apart.
- Every kind of protein folds up into a ***very specific shape*** -- the same shape every time.

57

- Most proteins do this all ***by themselves***, although some need ***extra help*** to fold into the right shape.
- The unique shape of a particular protein is the ***most stable state*** it can adopt. Picture a ball at the top of a hill -- the ball will always roll down to the bottom.
- If you try to put the ball back on top it will still roll down to the bottom of the hill because that is where it is most stable.

58

Unfolded (and unstable) Puzzle 48



59

Why is shape important?

- This structure ***specifies the function*** of the protein.
- For example, a protein that breaks down glucose (葡萄糖) so the cell can use the energy stored in the sugar
- The protein will have a shape that ***recognizes*** the glucose and ***binds*** to it (like ***a lock and key***)
- Chemically reactive amino acids will react with the glucose and ***break it down to release the energy***.

60

What do proteins do?

- Proteins are involved in ***almost all of the processes*** going on inside your body: they ***break down food*** to power your muscles, ***send signals through your brain*** that control the body, and ***transport nutrients through your blood***.
- Many proteins act as ***enzymes***, meaning they catalyze (speed up) chemical reactions that wouldn't take place otherwise.
- But other proteins power muscle contractions, or act as chemical messages inside the body, or hundreds of other things.

61

Here's a small sample of what proteins do:

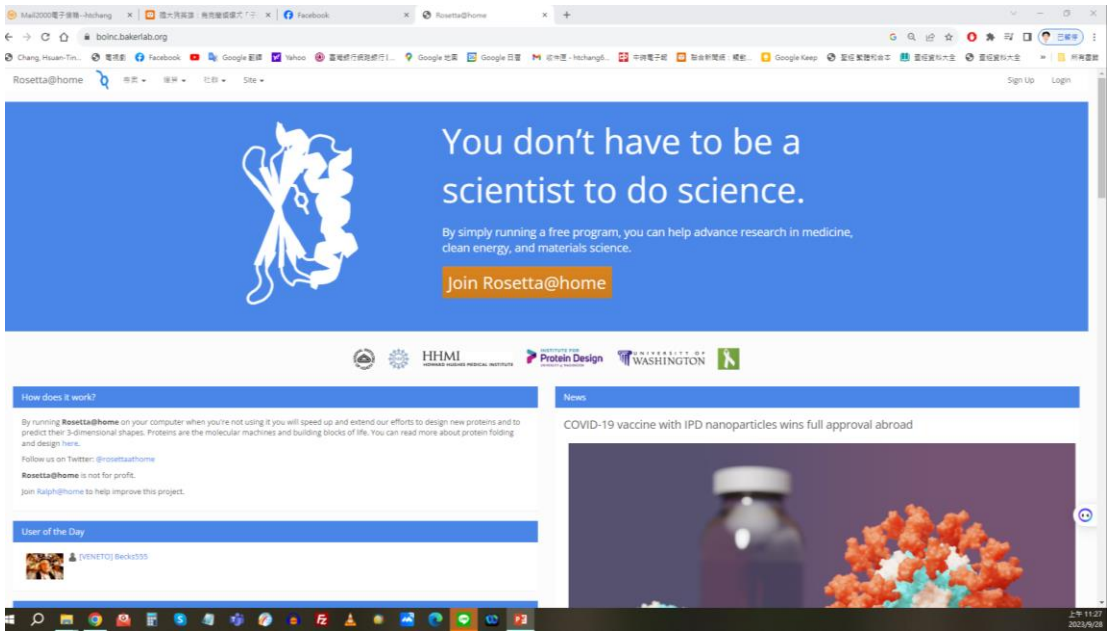
- ***Amylase*** (澱粉酶) starts the process of breaking down starch from food into forms the body can use.
- ***Alcohol dehydrogenase*** (乙醇脫氫酶) transforms alcohol from beer/wine/liquor into a non-toxic form that the body uses for food.
- ***Hemoglobin*** (血紅素) carries oxygen in our blood.
- ***Fibrin*** (纖維蛋白) forms a scab to protect cuts as they heal.
- ***Collagen*** (膠原) gives structure and support to our skin, tendons, and even bones.

62

- ***Actin*** (肌動蛋白) is one of the major proteins in our ***muscles***.
- ***Growth hormone*** (生長激素) helps ***regulate the growth*** of children into adults.
- ***Potassium channels*** (鉀通道) help send signals through the brain and other nerve cells.
- ***Insulin*** (胰島素) regulates the amount of sugar in the blood and is used to treat diabetes.

63

Rosetta@Home



64

- **Rosetta@home** is a volunteer computing project researching protein structure prediction on the Berkeley Open Infrastructure for Network Computing (BOINC) platform, run by the Baker lab.
- Rosetta@home aims to predict protein–protein docking and design new proteins with the help of about fifty-five thousand active volunteered computers processing at over 487,946 GigaFLOPS on average as of September 19, 2020.[4] Foldit, a Rosetta@home videogame, aims to reach these goals with a crowdsourcing approach.

65

- Rosetta@home needs your help to determine the 3-dimensional shapes of proteins in research that may ultimately lead to finding cures for some major human diseases.
- By running Rosetta@home on your computer when you're not using it you will speed up and extend our efforts to design new proteins and to predict their 3-dimensional shapes. Proteins are the molecular machines and building blocks of life.
- Though much of the project is oriented toward basic research to improve the accuracy and robustness of proteomics methods, Rosetta@home also does applied research on malaria, Alzheimer's disease, and other pathologies.[5]

66