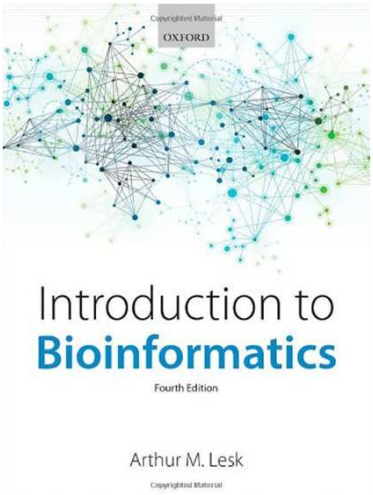
生物基因資訊科技簡介
Introduction to Bioinformatics



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
  - Gopal, Haake, Jones, and Tymann, ***Bioinformatics: A Computing Perspective***, McGraw-Hill 2009
  - **請購買正版書**
- Reference book:
  - Arthur M. Lesk, ***Introduction to Protein Science Architecture, Function, and Genomics*** , Oxford Press, 2004

2

- Grading:
  - Midterm – 40%
  - Homework & paper presentation – 30%
  - Project: protein sequence matching – 30%:
  - Office: Electrical Hall – Room EN308
- Phone: 05-5342601 ext. 4263
- Email: htchang@yuntech.edu.tw
- Homepage: http://teacher.yuntech.edu.tw/htchang/

3

Journal list

- Bioinformatics
- BMC bioinformatics
- Journal of Bioinformatics and Computational Biology
- Genomics, Proteomics & Bioinformatics

4

The Author

Arthur Lesk

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

5

Contents

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 (2014) the nucleotide sequences databanks contain >150 Gbp, >16M sequences
  - A very very large amount
  - 50 human genome equivalents
- The database of *macromolecular* structures contains ~100,000 entries, full 3-D coordinates of proteins, of average length ~400 residues.

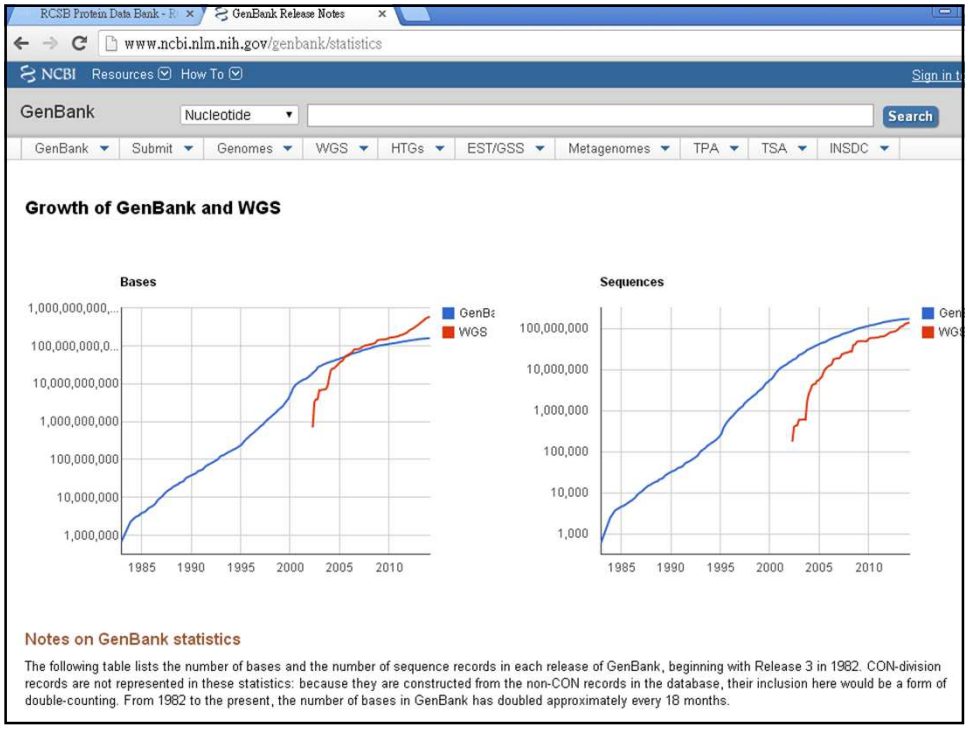
8

Growth of GeneBank and Protein Data Bank

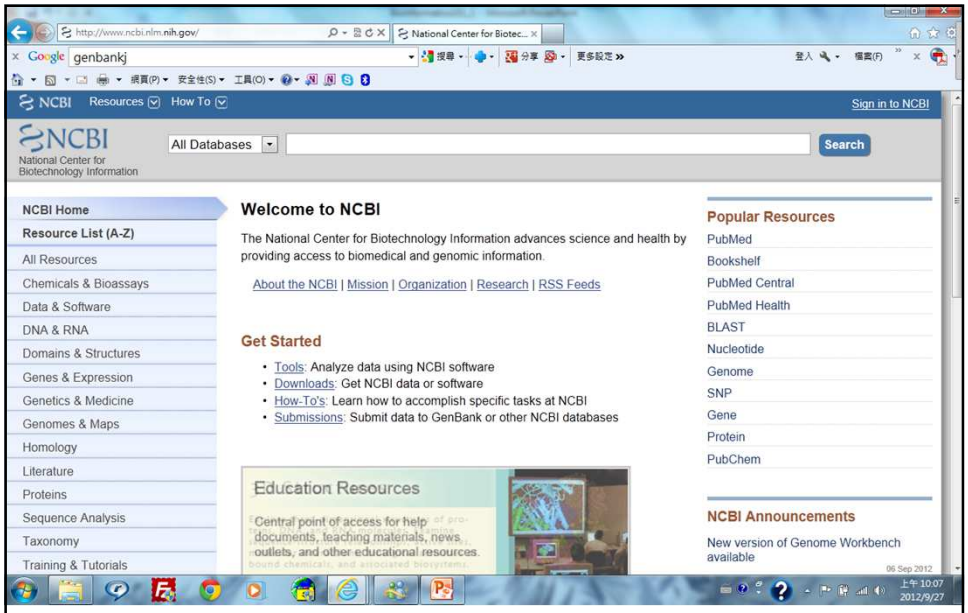
Current Holdings
38729 Structures
Last Update: 12-Sep-2006
PDB Statistics

111956 Structures
Last Update: 14-Sep-2015
PDB Statistics

Hsuan T. Chang, EE
Yuntech

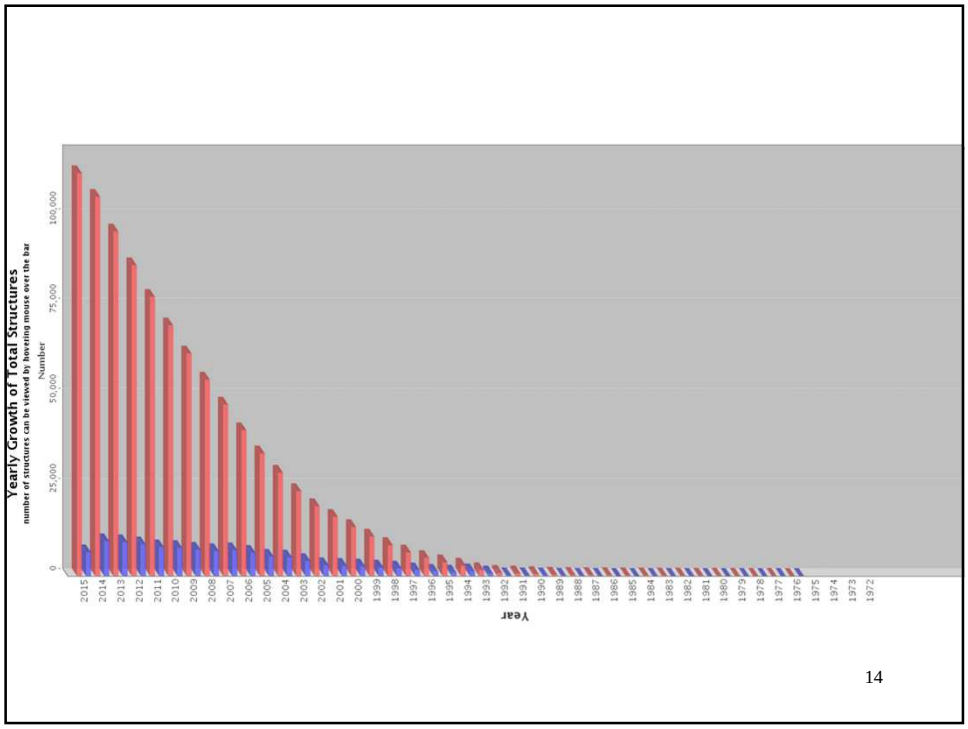


Year	nucleotide #	sequence #
2005	56,037,734,462	52,016,762
2006	69,019,290,705	64,893,747
2007	83,874,179,730	80,388,382
2008	99,116,431,942	98,868,465
2009	106,533,156,756	108,431,692
2011	126,551,501,141	135,440,924
2012	143,081,765,233	156,424,033
2014	157,943,793,171	171 123 749
2015	199,823,644,287	187,066,846



12


RCSB PDB PROTEIN DATA BANK
PDB-101
An Information Portal to Biological Macromolecular Structures
As of Tuesday Sep 25, 2012 at 5 PM PDT there are 84846 Structures | PDB Statistics
All Categories | Author | Macromolecule | Sequence | Ligand
Search | All Categories:
e.g., PDB ID, molecule name, author
Browse | Advanced
Customize This Page
Available on the App Store
PDB-101 Hide
Structural View of Biology
Understanding PDB Data
Molecule of the Month
Educational Resources
Author Profiles
MyPDB Hide
Login to your Account
Register a New Account
Home Hide
News & Publications
Usage/Reference Policies
Deposition Policies
Website FAQ
Deposition FAQ
Contact Us
Biological Macromolecular Resource
Full Description
Featured Molecules Hide
Structural View of Biology
List View of Archive By: Title | Date | Category
Biological Energy
Molecule of the Month
Pyruvate Dehydrogenase Complex
A combination of crystallography, NMR spectroscopy and electron microscopy is revealing the secrets of pyruvate dehydrogenase complex. The complex performs a central step in energy production, catalyzing the reaction that links glycolysis with the tricarboxylic acid cycle. The reaction is performed in three separate steps by three separate enzymes, but all three enzymes are linked efficiently together into one large multienzyme complex.
Full Article
Protein Structure Initiative Featured System
Bacterial Phosphotransferase System
Bacteria are thrifty cells. They often live in inhospitable
New Structures Hide
Latest Release
New Structure Papers
Search Unreleased Entries
New Features Hide
Advanced Search: All/Experimental
Latest features released:
Website Release Archive:
RCSB PDB News Hide
Weekly | Quarterly | Yearly
2012-09-25
Explore PDB Data Overviews
上午 10:40
2012/9/27


Yearly Growth of Total Structures
number of structures can be viewed by hovering mouse over the bar
Number
100,000
75,000
50,000
25,000
0
Year
2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 2005 2004 2003 2002 2001 2000 1999 1998 1997 1996 1995 1994 1993 1992 1991 1990 1989 1988 1987 1986 1985 1984 1983 1982 1981 1980 1979 1978 1977 1976 1975 1974 1973 1972

14

PDB Current Holdings Breakdown

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	**93446**	**1662**	**4647**	**4**	**99759**
NMR	**9723**	**1127**	**227**	**8**	**11085**
ELECTRON MICROSCOPY	**612**	**29**	**203**	**0**	**844**
HYBRID	**71**	**3**	**2**	**1**	**77**
other	**168**	**4**	**6**	**13**	**191**
Total	**104020**	**2825**	**5058**	**26**	**111956**

15

- 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.

16

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.

17

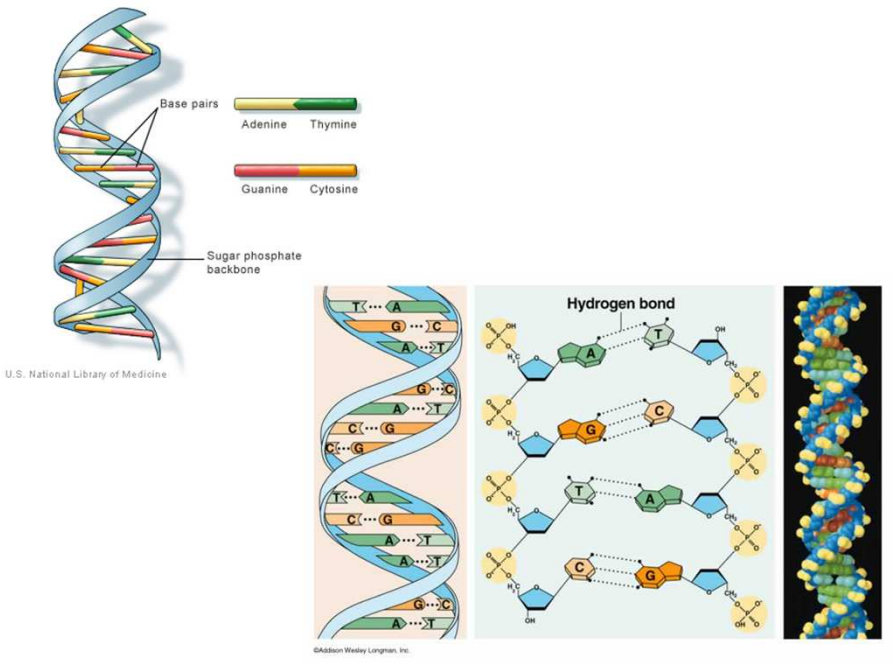
- Organisms are composed of cells. Every cell is an intimate localized ecosystem, not isolated from its environment but interacting within specific and controlled ways.
- Life is extended not only in space but in time
- We must try to read the past in contemporary (現代的) genomes.

18

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).

19



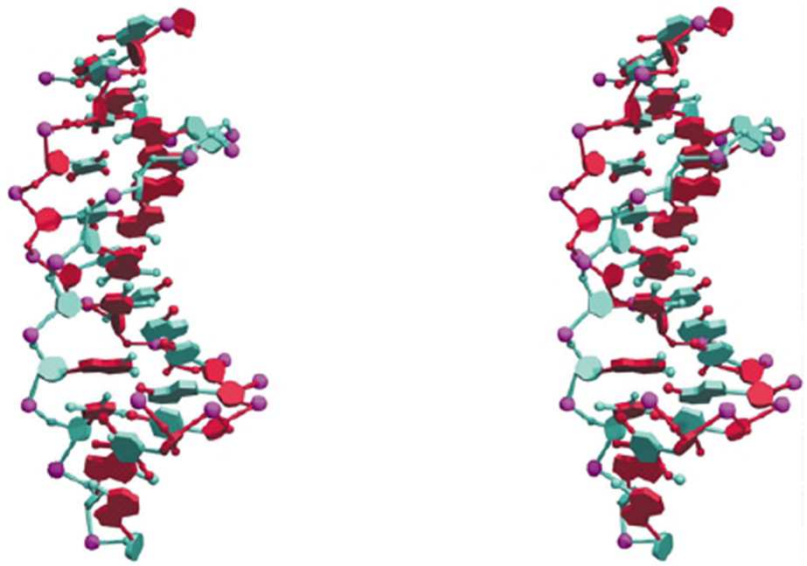


Plate I Double-helix of DNA. (See page 5.)

- 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.

22

- 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.

23

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

24

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

25

- 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.

26

- The translation of DNA sequences to amino acid sequences is very simple to describe logically; it 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.

27

- 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.

28

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.

29

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

30

Archival databanks of biological information

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

31

~~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?”

32

- 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 sequences in the database that adopt similar 3-D structures.
  - Given a protein structure, find sequences in the databank that correspond to similar structures.

33

- 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.

34

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.

35

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.

36

The World Wide Web

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

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

37

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

38

- 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.

39

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

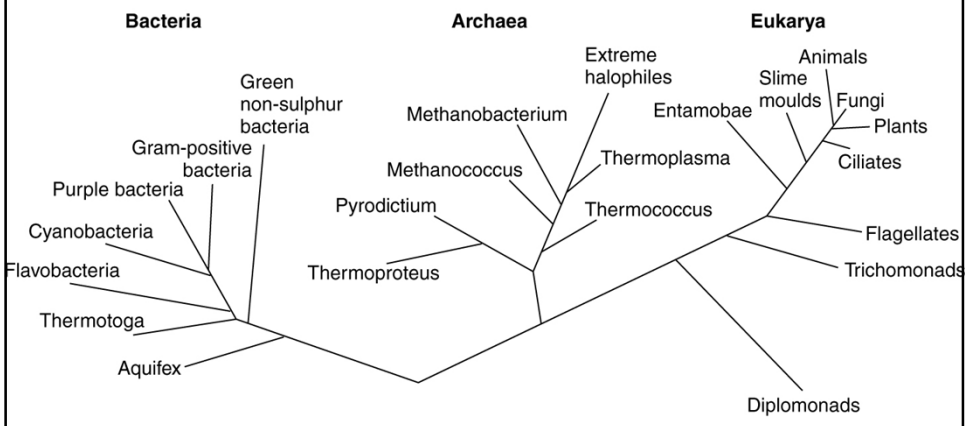
40

- 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.

41

- 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.

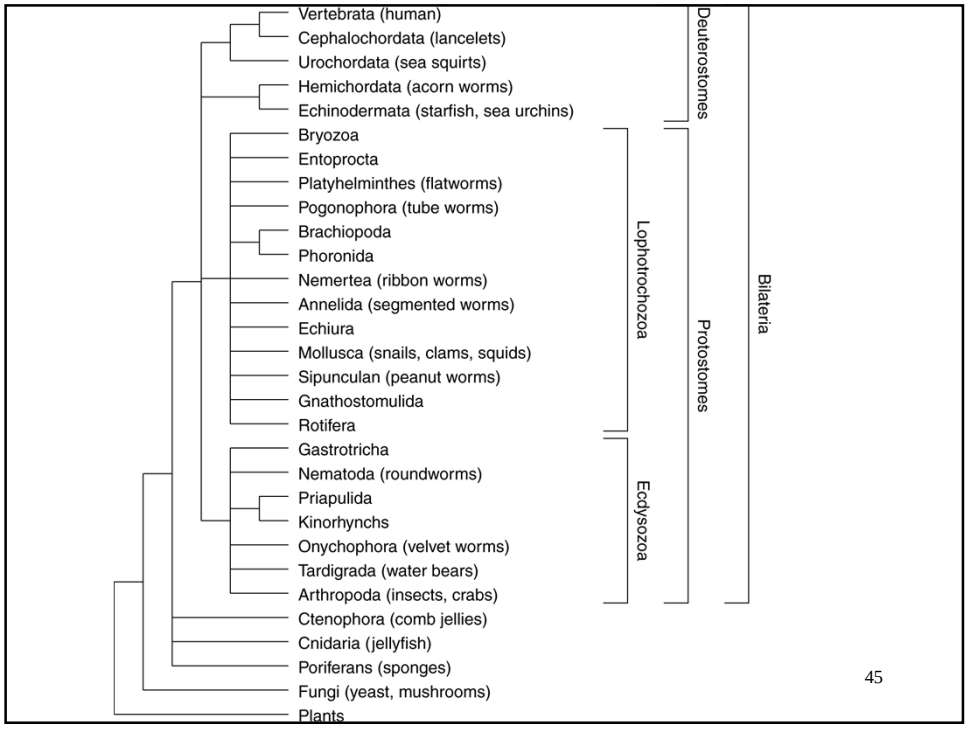
42

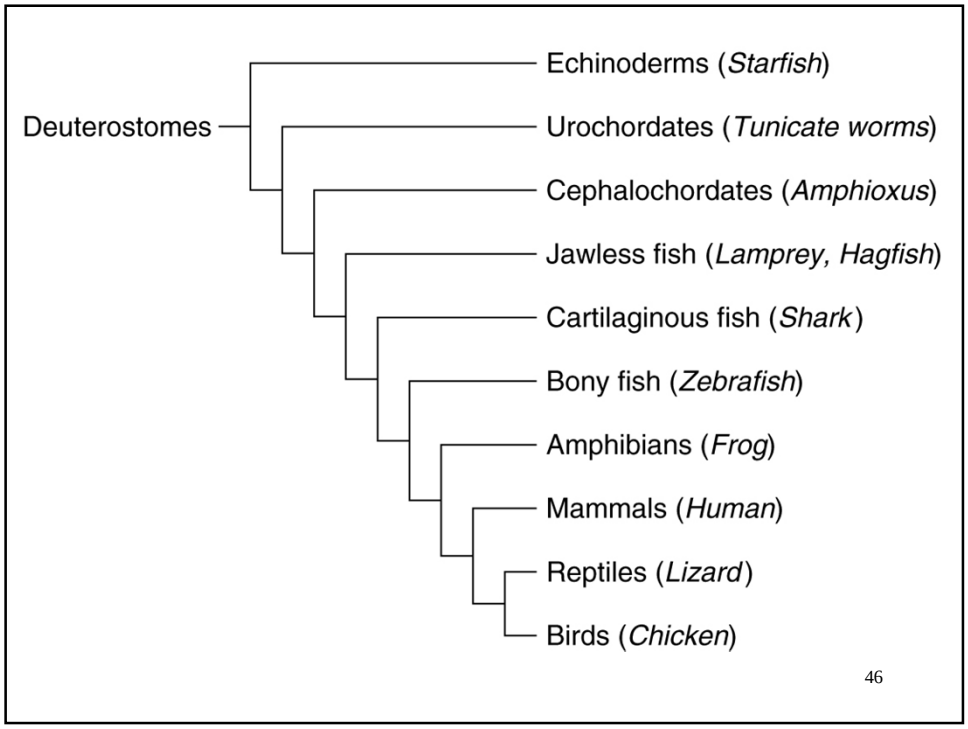


43

- 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)

44

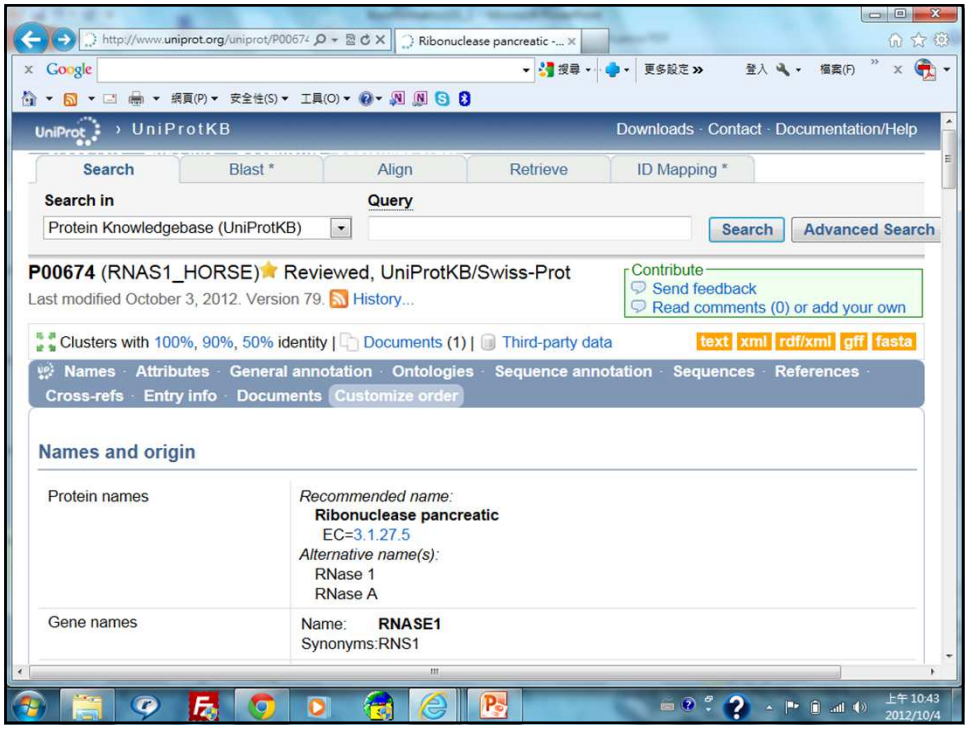

Vertebrata (human)
Cephalochordata (lancelets)
Urochordata (sea squirts)
Hemichordata (acorn worms)
Echinodermata (starfish, sea urchins)
Bryozoa
Entoprocta
Platyhelminthes (flatworms)
Pogonophora (tube worms)
Brachiopoda
Phoronida
Nemertea (ribbon worms)
Annelida (segmented worms)
Echiura
Mollusca (snails, clams, squids)
Sipunculan (peanut worms)
Gnathostomulida
Rotifera
Gastrotricha
Nematoda (roundworms)
Priapulida
Kinorhynchs
Onychophora (velvet worms)
Tardigrada (water bears)
Arthropoda (insects, crabs)
Ctenophora (comb jellies)
Cnidaria (jellyfish)
Poriferans (sponges)
Fungi (yeast, mushrooms)
Plants
Deuterostomes
Lophotrochozoa
Ecdysozoa
Protostomes
Bilateria
45


Deuterostomes
Echinoderms (*Starfish*)
Urochordates (*Tunicate worms*)
Cephalochordates (*Amphioxus*)
Jawless fish (*Lamprey, Hagfish*)
Cartilaginous fish (*Shark*)
Bony fish (*Zebrafish*)
Amphibians (*Frog*)
Mammals (*Human*)
Reptiles (*Lizard*)
Birds (*Chicken*)
46

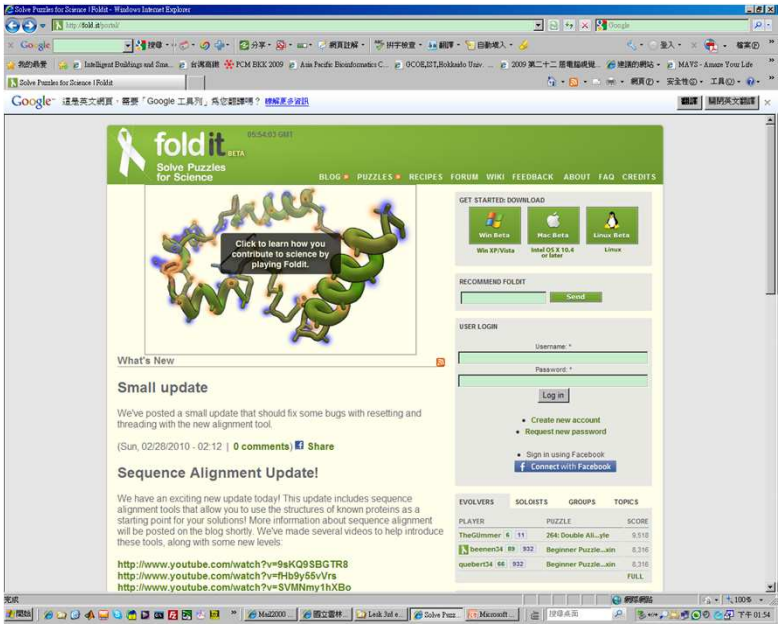
Case study

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

47



Term Project http://fold.it/portal/



50

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

Webpage description about this project – http://fold.it/portal/info/science

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

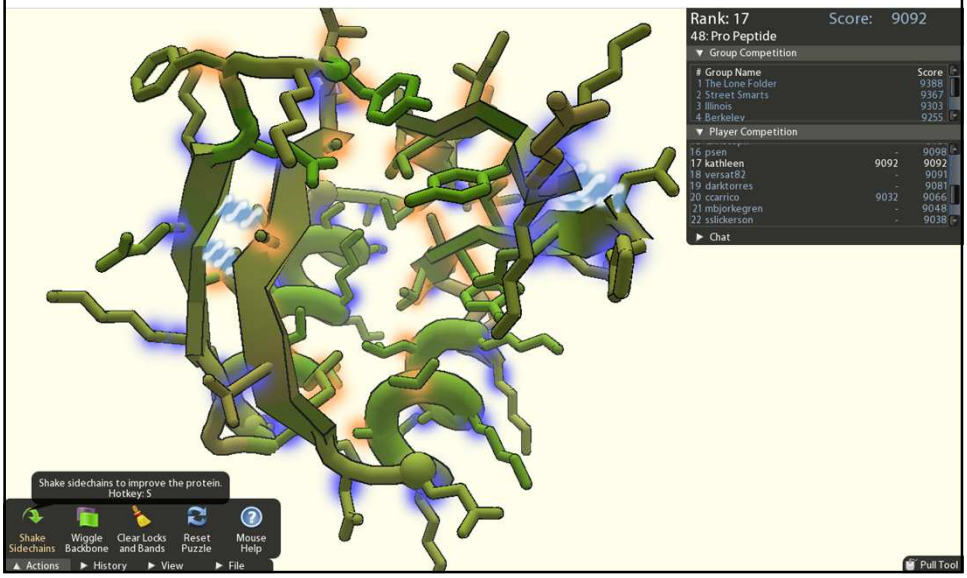
52

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



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. 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).

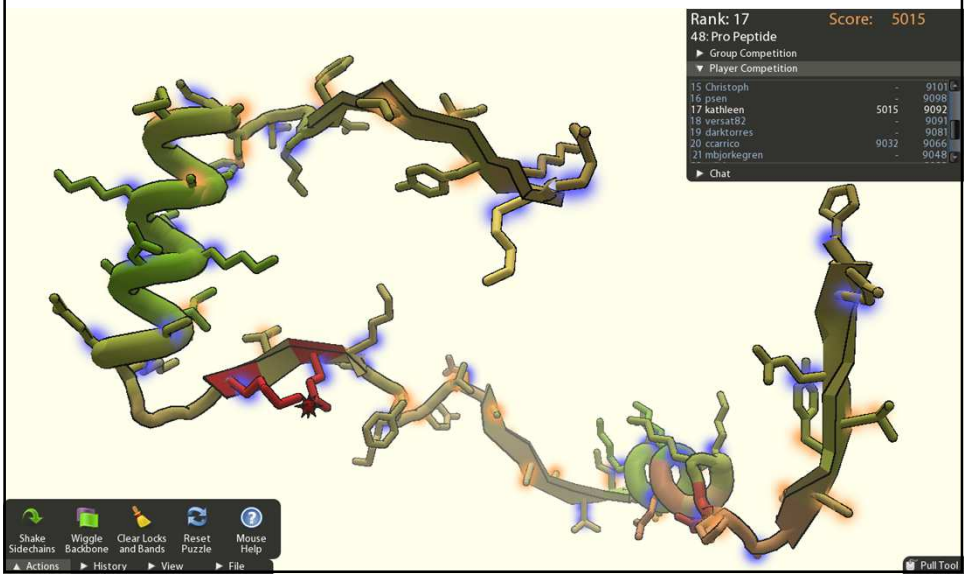
55

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. 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.

56

Unfolded (and unstable) Puzzle 48



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 will have a shape that recognizes the glucose and binds to it (like a lock and key) and chemically reactive amino acids that will react with the glucose and break it down to release the energy.

58

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.

59

Here's a small sample of what proteins do:

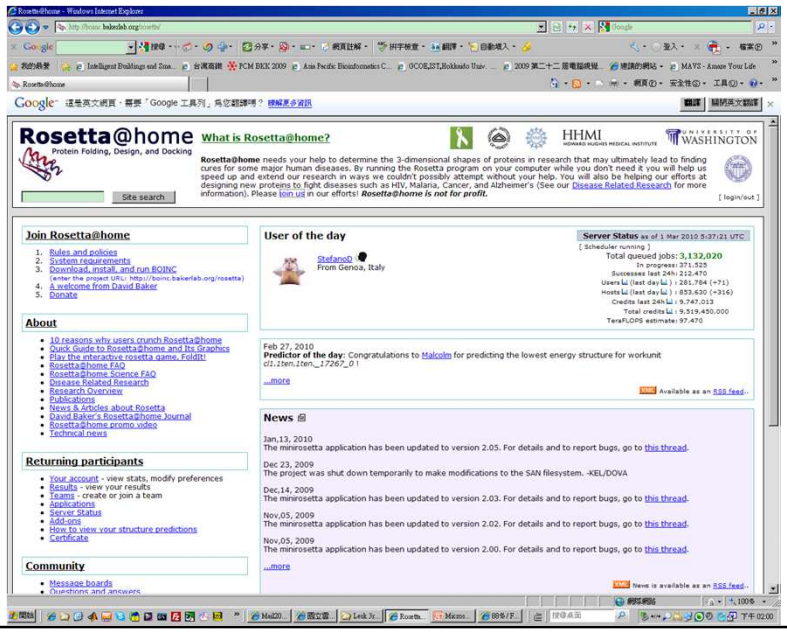
- Amylase starts the process of breaking down starch from food into forms the body can use.
- Alcohol dehydrogenase transforms alcohol from beer/wine/liquor into a non-toxic form that the body uses for food.
- Hemoglobin carries oxygen in our blood.
- Fibrin forms a scab to protect cuts as they heal.
- Collagen gives structure and support to our skin, tendons, and even bones.

60

- Actin is one of the major proteins in our muscles.
- Growth hormone helps regulate the growth of children into adults.
- Potassium channels help send signals through the brain and other nerve cells.
- Insulin regulates the amount of sugar in the blood and is used to treat diabetes.

61

Rosetta@Home Screensaver



62

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 the Rosetta program on your computer while you don't need it you will help us speed up and extend our research in ways we couldn't possibly attempt without your help. You will also be helping our efforts at designing new proteins to fight diseases such as HIV, Malaria, Cancer, and Alzheimer's.

63