山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

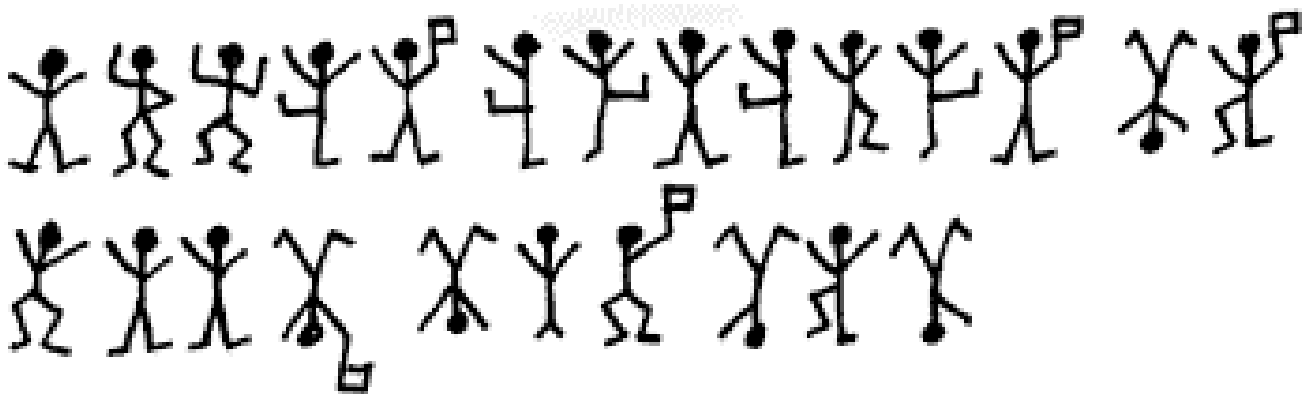
生物信息学

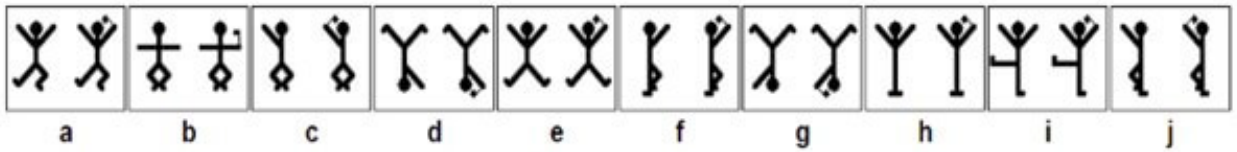
第一章: 导论

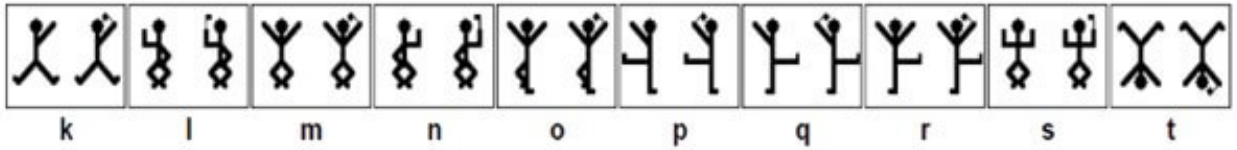
1. 生物信息学的概念

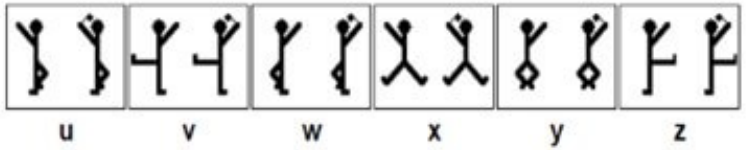
桂松涛
songtaogui@163.com

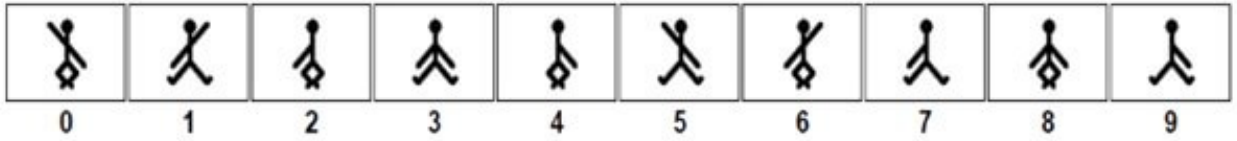
WHAT and WHY



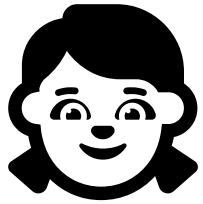








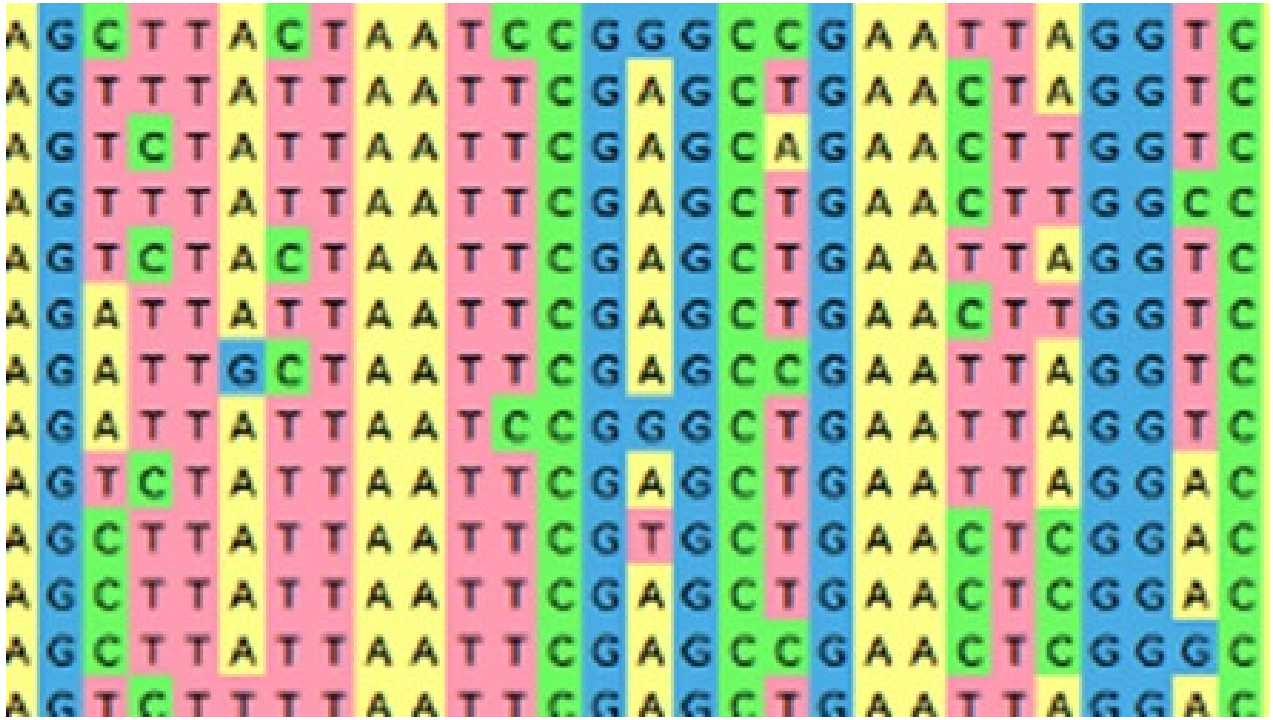
WHAT and WHY



你猜我上午去药店的路上，遇到谁了？

药店!药店!药店! 别问我怎么知道的

WHAT and WHY

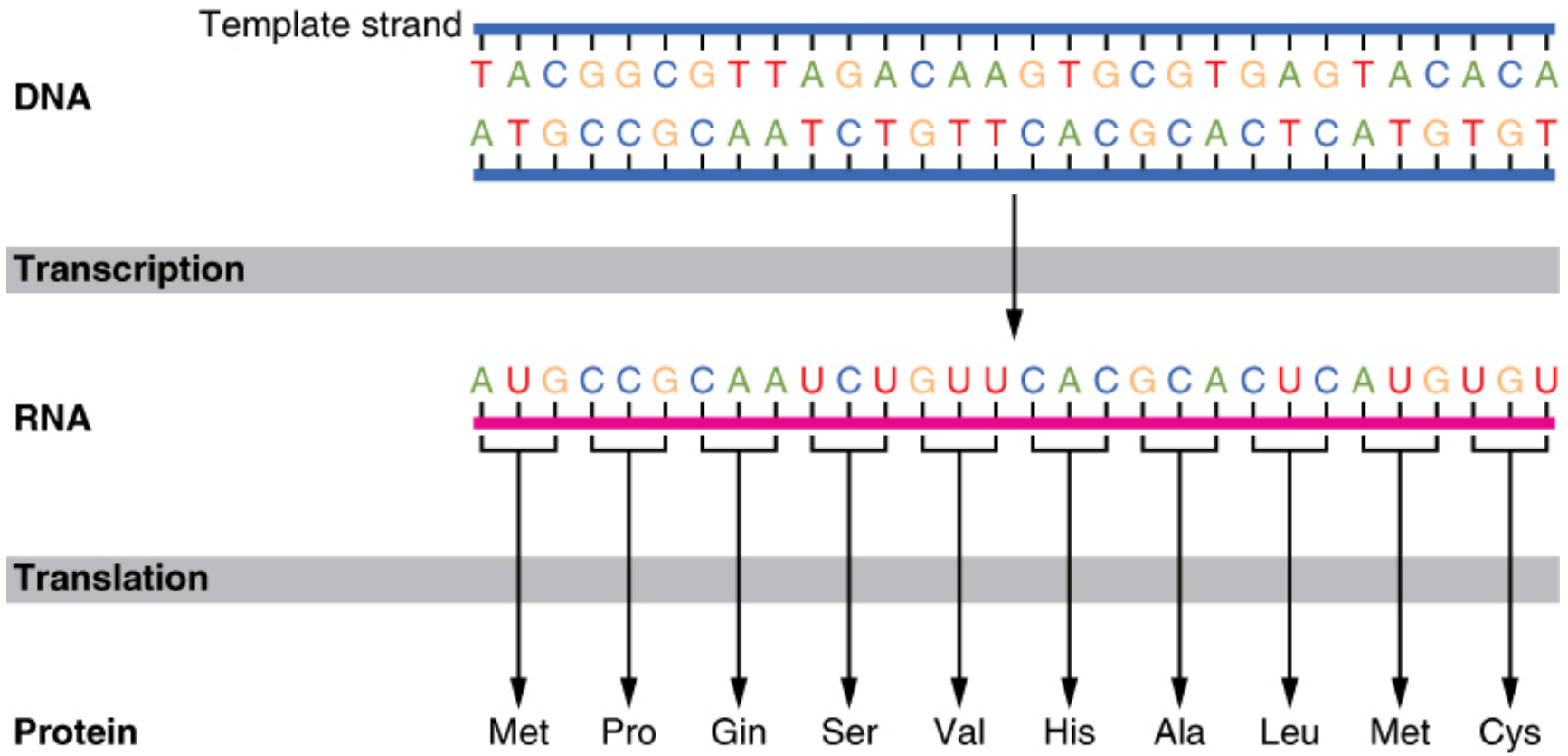


这段DNA序列是啥意思?

WHAT and WHY

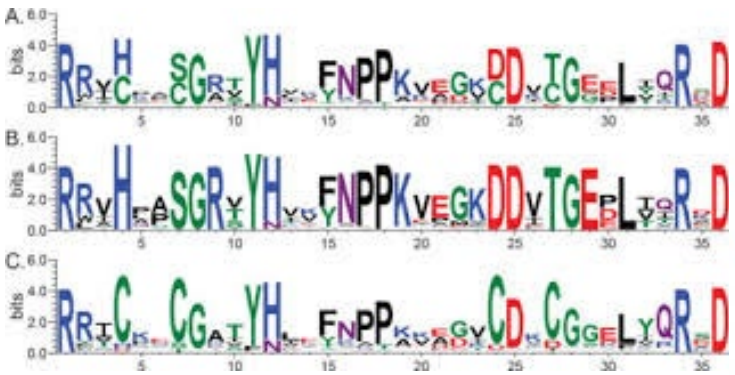
- 人类基因组包含3G碱基对
- 两个人之间的基因组相似度为99.9%以上 (只有约3M碱基的差别)
- 平均每600 bp 就可能出现一个SNP
- 平均基因长度(含调控区) 27 Kb
- 基因编码区只占总长度的1.5%

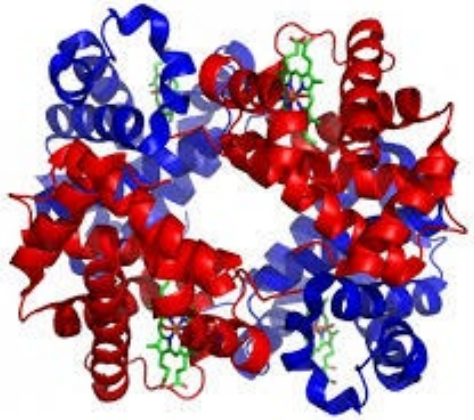
WHAT and WHY

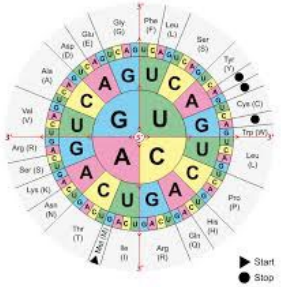

Template strand
DNA
T A C G G C G T T A G A C A A G T G C G T G A G T A C A C A
A T G C C G C A A T C T G T T C A C G C A C T C A T G T G T
Transcription
RNA
A U G C C G C A A U C U G U U C A C G C A C U C A U G U G U
Translation
Protein
Met
Pro
Gin
Ser
Val
His
Ala
Leu
Met
Cys

山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

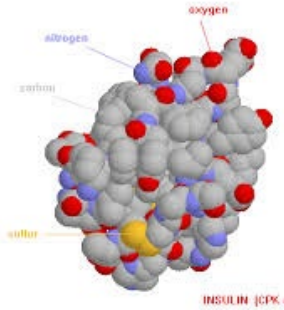
WHAT and WHY

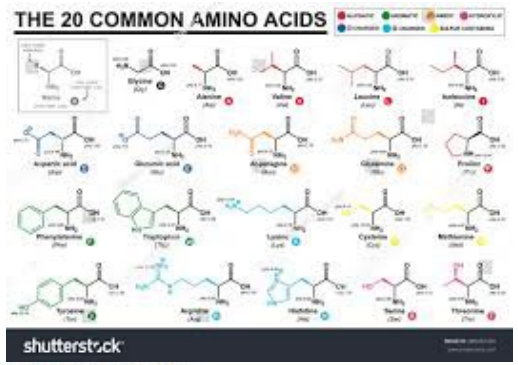


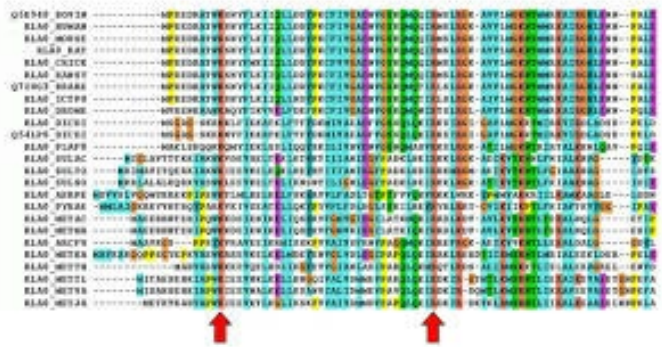


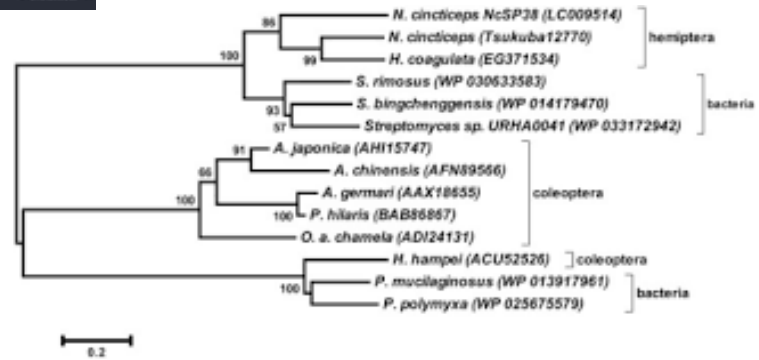


CCTTGATGCTCCTGGCACAAATGAGCAGAATCTCT
GACATGACTTTGGATTTCCCCAGGAGGAGTTTGAT
CAGCCATCTCTGTCCTCCATGAGCTGATCCAGCAG
ATCTTCAACCTCTTTACCACAAAAGATTCATCTGCTGCTTGGGATGAGGACCTCCTA
GACAAATTCTGCACCGAACTCTACCAGCAGCTGAATGACTTGGAAGCCTGTGTGATG
CAGGAGGAGAGGGTGGGAGAAACTCCCCTGATGAATGCGGACTCCATCTTGGCTGTG
AAGAAATACTTCCGAAGAATCACTCTCTATCTGACAGAGAAGAAATACAGCCCTTGT
GCCTGGGAGGTTGTCAGAGCAGAAATCATGAGATCCTCTCTTTATCAACAAACTTGC
AAGAAAGATTAAGGAGGAAGGAATAA, TGTGATCTCCCTGAGACCCACAGCCTGGA
TAACAGGAGGACCTTGATGCTCCTGGCACAAATGAGCAGAATCTCTCCTTCCTCCTG
TCTGATGGACAGACATGACTTTGGATTTCCCCAGGAGGAGTTTGATGGCAACCAGTT
CCAGAAGGCTCCAGCCATCTCTGTCCTCCATGAGCTGATCCAGCAGATCTTCAACCT

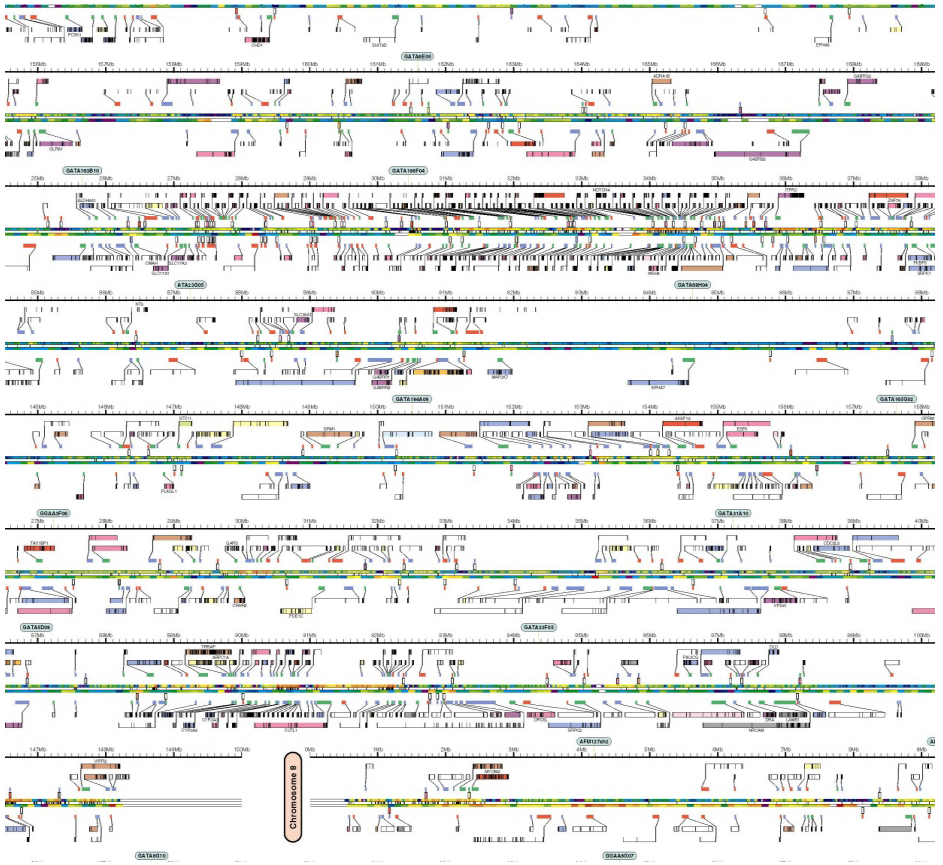








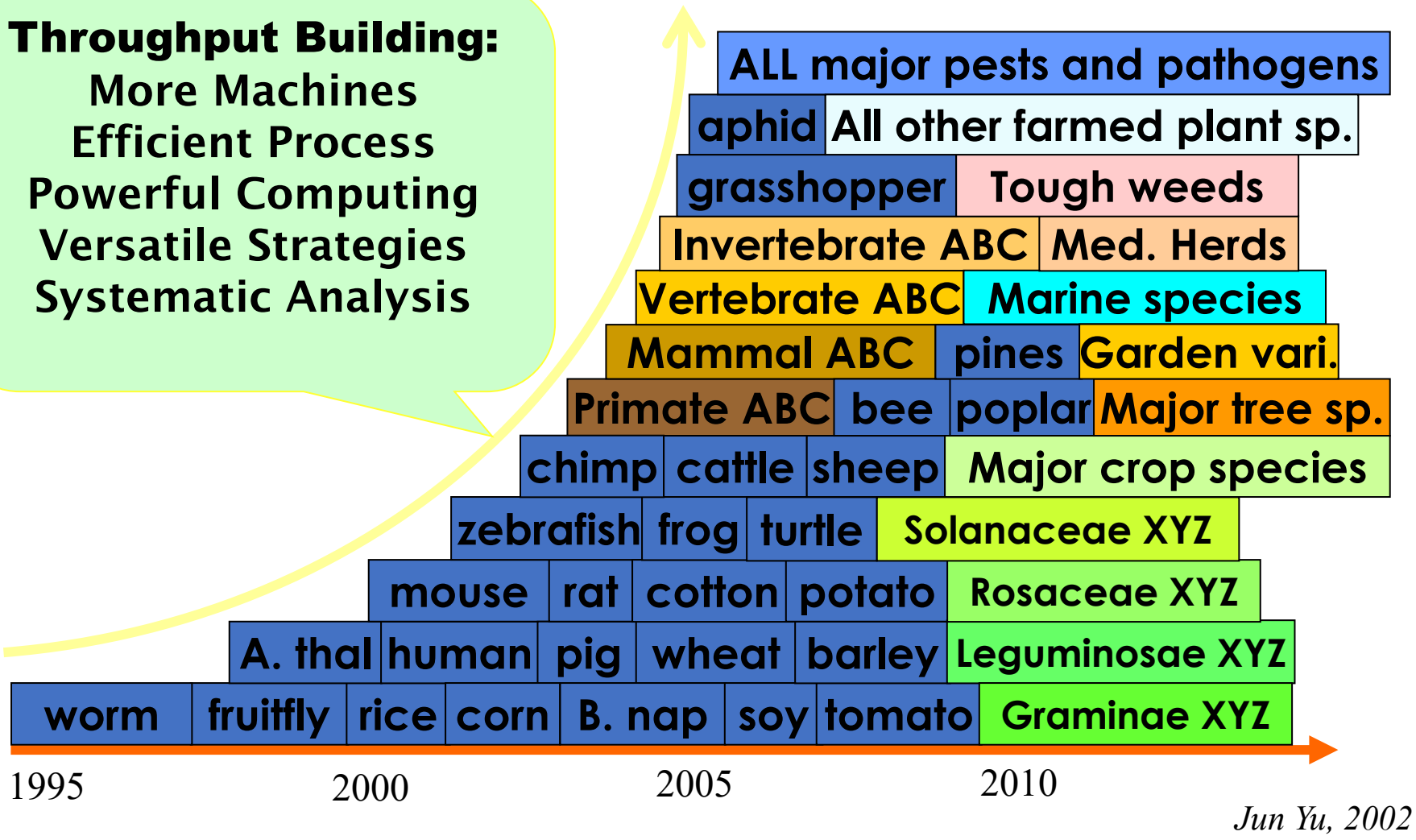
迅速增长的生物信息



"The Sequence of the Human Genome"
Science, 2001

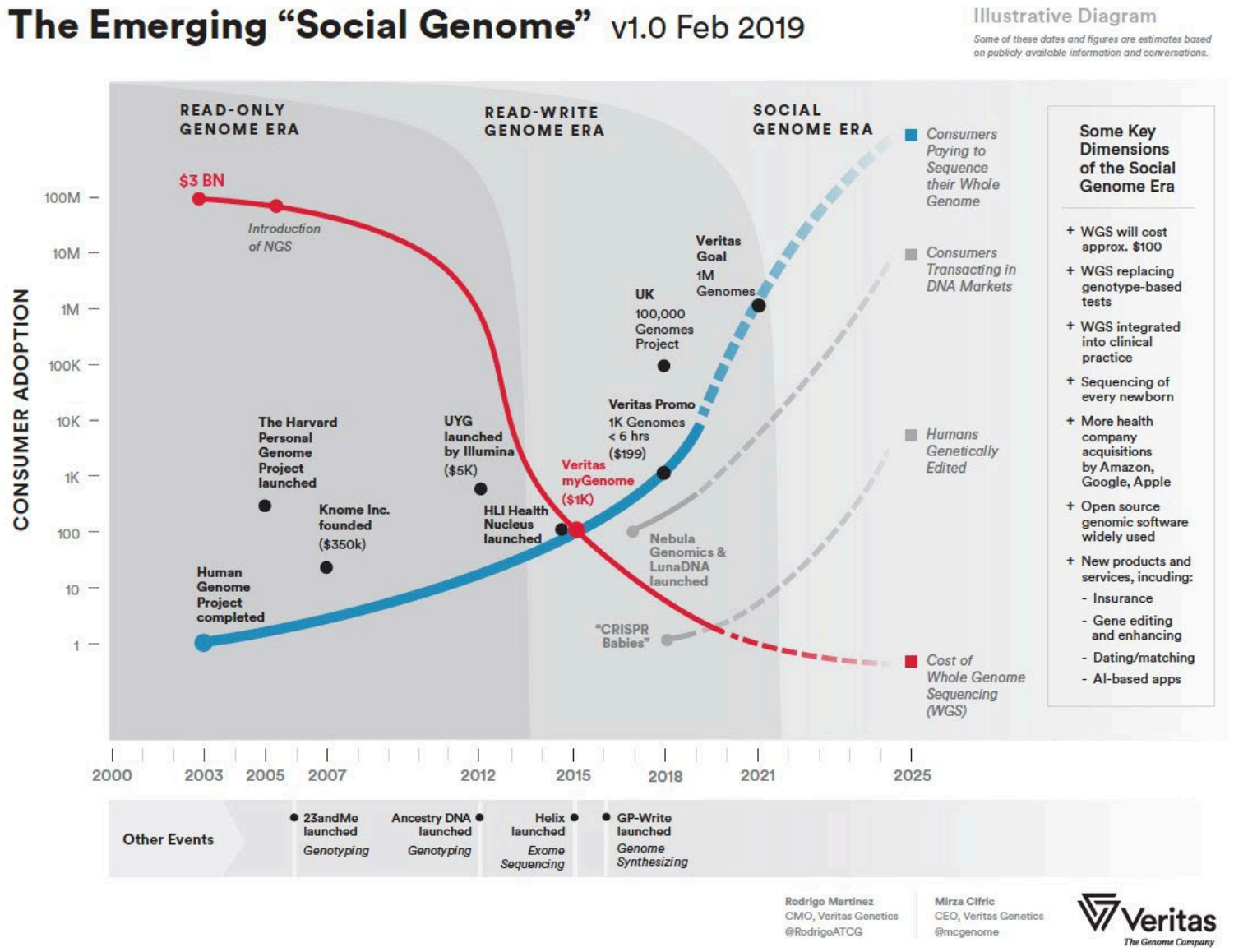
迅速增长的生物信息

Throughput Building:
More Machines
Efficient Process
Powerful Computing
Versatile Strategies
Systematic Analysis



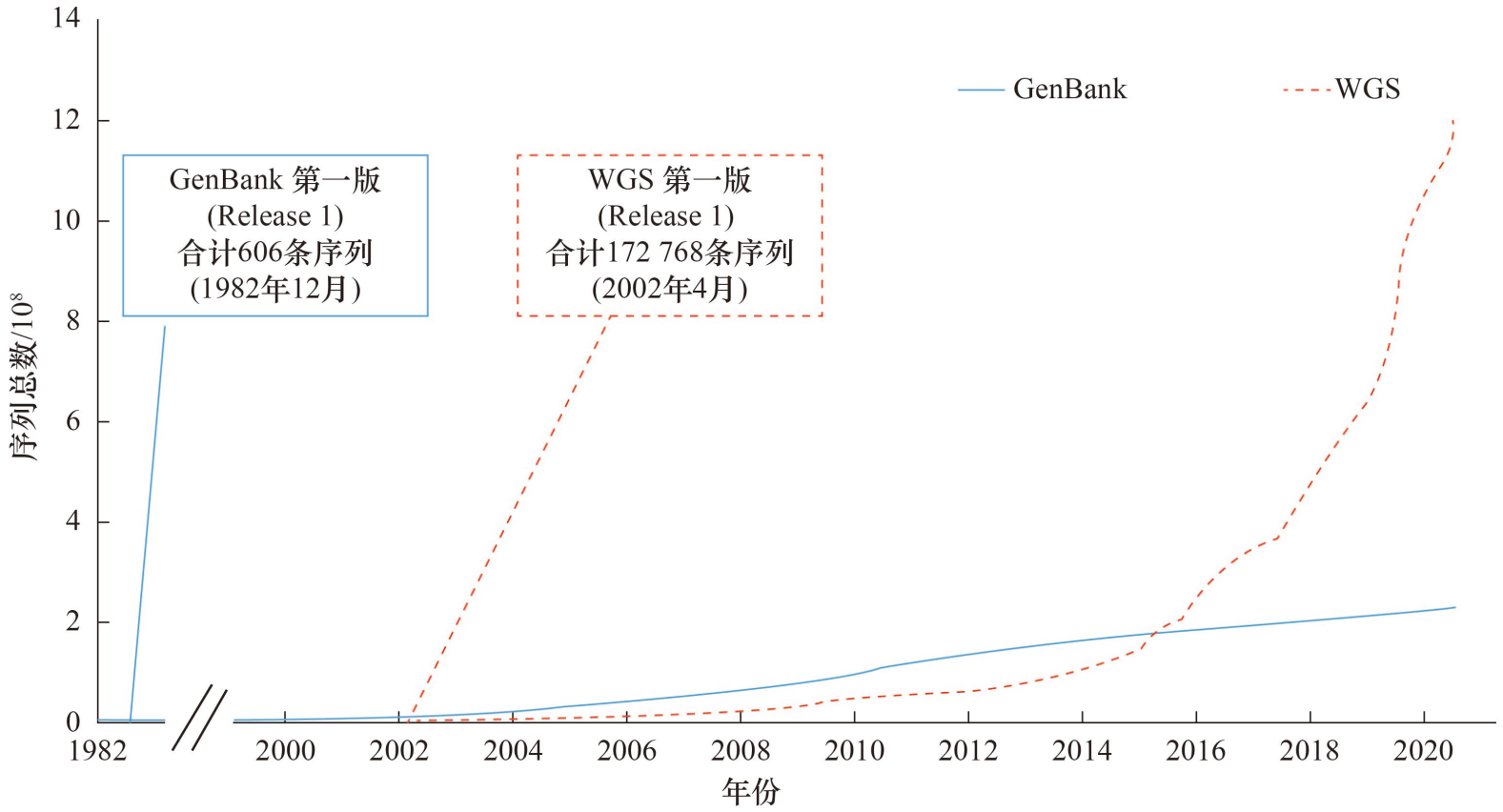
Jun Yu, 2002

迅速增长的生物信息


The Emerging "Social Genome" v1.0 Feb 2019
Illustrative Diagram
Some of these dates and figures are estimates based on publicly available information and conversations.
READ-ONLY GENOME ERA
READ-WRITE GENOME ERA
SOCIAL GENOME ERA
CONSUMER ADOPTION
100M
10M
1M
100K
10K
1K
100
10
1
$3 BN
Introduction of NGS
Human Genome Project completed
The Harvard Personal Genome Project launched
Knome Inc. founded ($350k)
UYG launched by Illumina ($5K)
HLI Health Nucleus launched
Veritas myGenome ($1K)
Veritas Promo 1K Genomes < 6 hrs ($199)
UK 100,000 Genomes Project
Veritas Goal 1M Genomes
Nebula Genomics & LunaDNA launched
"CRISPR Babies"
Consumers Paying to Sequence their Whole Genome
Consumers Transacting in DNA Markets
Humans Genetically Edited
Cost of Whole Genome Sequencing (WGS)
Some Key Dimensions of the Social Genome Era
+ WGS will cost approx. $100
+ WGS replacing genotype-based tests
+ WGS integrated into clinical practice
+ Sequencing of every newborn
+ More health company acquisitions by Amazon, Google, Apple
+ Open source genomic software widely used
+ New products and services, incuding:
- Insurance
- Gene editing and enhancing
- Dating/matching
- AI-based apps
2000
2003
2005
2007
2012
2015
2018
2021
2025
Other Events
23andMe launched Genotyping
Ancestry DNA launched Genotyping
Helix launched Exome Sequencing
GP-Write launched Genome Synthesizing
Rodrigo Martinez
CMO, Veritas Genetics
@RodrigoATCG
Mirza Cifric
CEO, Veritas Genetics
@mcgenome
Veritas
The Genome Company

迅速增长的生物信息

- 核酸数据库每14个月左右要翻一番
- 2024年, 公共核酸数据库: 8.2万亿个碱基对



迅速增长的生物信息

信息爆炸时代如何保持洞察力?

- 高效的信息获取能力
- 批判性思维
- 自我学习和持续学习能力
- 保持好奇心和探索精神

迅速增长的生物信息

- 有效管理
- 准确解读
- 充分使用

生物信息学的概念

太长不看系列

“生物信息学是研究生物信息的采集、处理、存储、传播、分析和解释等的一门学科，通过综合利用分子生物学、遗传学、计算机科学与技术，来揭示大量且复杂的生物数据中蕴含的生物学奥秘。”

“生物信息学是利用信息来理解生物学的一门科学。”

山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

生物信息学

第一章: 导论

2. 生物信息学发展史

桂松涛
songtaogui@163.com

生物信息学名词的提出

宝琳•霍格维
（Paulien Hogeweg）
（1943-）
荷兰乌得勒支大学教授

- 1978年提出Bioinformatics一词(官方认证的最早)

林华安 博士
（Hwa A. Lim）
（1957-）
马来西亚华裔, 佛州超算中心遗传学与生物物理学部门主任

- 1990年组织了世界第一个国际生物学信息学学术会议, 催生了`生物信息学`一词的出现。

生物信息学名词的提出

Bioinformatics

- **Biology**
- **Information**
- **Mathematics**

山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

生物信息发展史：Bioinformatics

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik.* (*Stockholm*), **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of

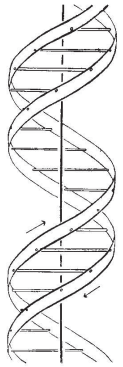
is a residue on each chain every 3·4 A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

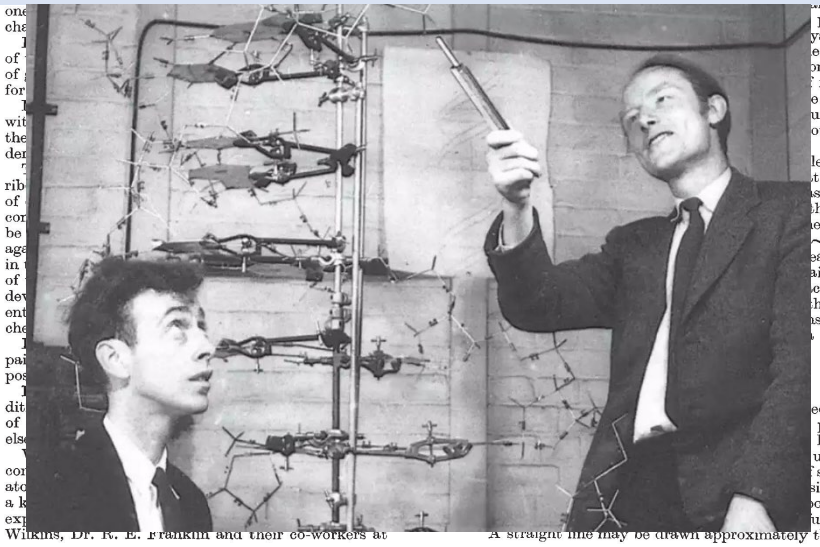
If it is assumed that the bases only occur in the

phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There



Wilkins, Dr. R. E. Franklin and their co-workers at

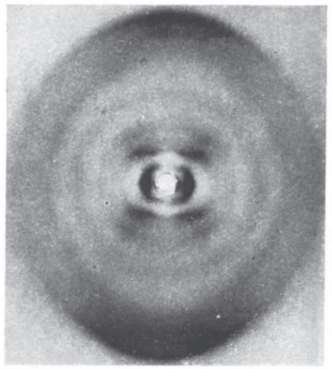
King's College, London. One of us (J. D. W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J. D. Watson
F. H. C. Crick
Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems,
Cavendish Laboratory, Cambridge.
April 2.

[1] Pauling, L., and Corey, R. B., *Nature*, **171**, 346 (1953); *Proc. U.S. Nat. Acad. Sci.*, **39**, 84 (1953).
[2] Furberg, S., *Acta Chem. Scand.*, **6**, 634 (1952).
[3] Chargaff, E., for references see Zamenhof, S., Brawerman, G., and Chargaff, E., *Biochim. et Biophys. Acta*, **9**, 402 (1952).
[4] Wyatt, G. R., *J. Gen. Physiol.*, **36**, 201 (1952).
[5] Astbury, W. T., Symp. Soc. Exp. Biol. 1, Nucleic Acid, 66 (Camb. Univ. Press, 1947).
[6] Wilkins, M. H. F., and Randall, J. T., *Biochim. et Biophys. Acta*, **10**, 192 (1953).

Molecular Structure of Deoxypentose Nucleic Acids

WHILE the biological properties of deoxypentose



parallel
ystalline
e X-ray
ons, one
nucleo-
e longer
uence of
ot made
leic acid
tion by
s shown
hat the
e inter-
~ 34 A.
eat of a
ain con-
ction as
han the
s on or
helical
ed) that
pattern
helix is
uniform
spacing
sity dis-
portional
unction.
A straight line may be drawn approximately through

We will now briefly analyse in physical terms some of the effects of the shape and size of the repeat unit or nucleotide on the diffraction pattern. First, if the nucleotide consists of a unit having circular symmetry about an axis parallel to the helix axis, the whole diffraction pattern is modified by the form factor of the nucleotide. Second, if the nucleotide consists of a series of points on a radius at right-angles to the helix axis, the phases of radiation scattered by the helices of different diameter passing through each point are the same. Summation of the corresponding Bessel functions gives reinforcement for the inner-

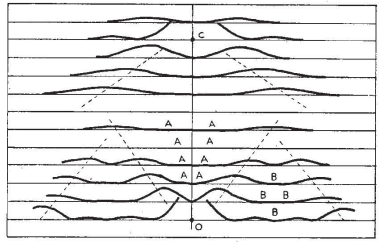


Fig. 2. Diffraction pattern of system of helices corresponding to structure of deoxypentose nucleic acid. The squares of Bessel functions are plotted about 0 on the equator and on the first, second, third and fifth layer lines for half of the nucleotide mass at 20 A. diameter and remainder distributed along a radius, the mass at a given radius being proportional to the radius. About *C* on the tenth layer line similar functions are plotted for an outer diameter of 12 A.

1953年，由沃森和克里克提出DNA双螺旋结构模型，开启了分子生物学时代

生物信息发展史：Bioinformatics

The Amide Groups of Insulin

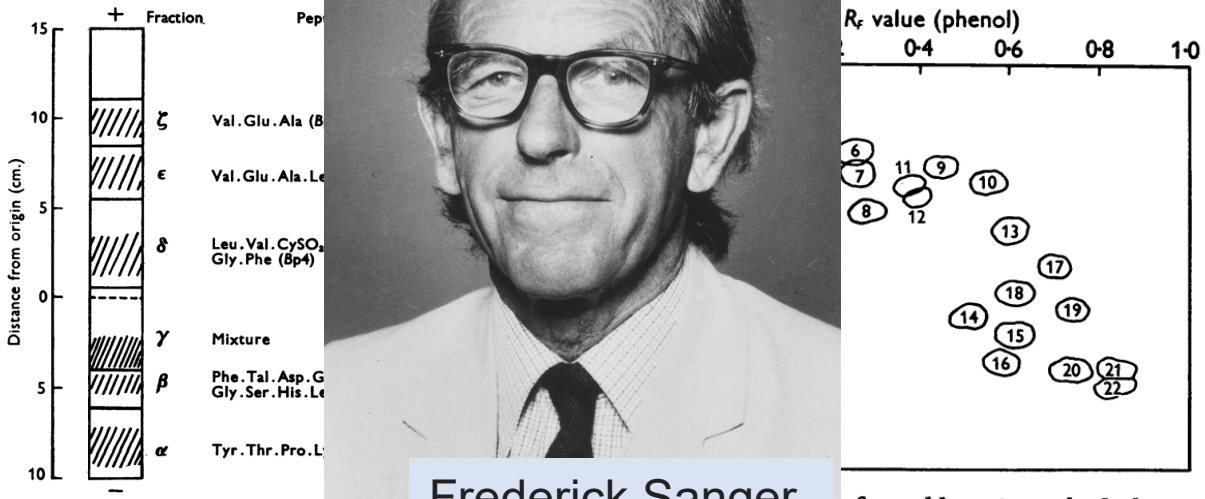
By F. SANGER,* E. O. P. THOMPSON† and RUTH KITAI
Department of Biochemistry, University of Cambridge

(*Received* 6 *September* 1954)

1955年, Sanger用二硝基氟苯（FDNB）法，首次成功地完成了第一个蛋白质-牛胰岛素的序列分析

† Present address: Wool Tex
343 Royal Parade Parkville, Vi

borohydride Chibnall & Rees (1952)
e three aspartic acids, one of which is



Frederick Sanger

Fig. 1. Ionophoresis of peptic hydroly
0·05M ammonium acetate.

of mould protease hydrolysate of
Expt. *A*m) (see Table 1).

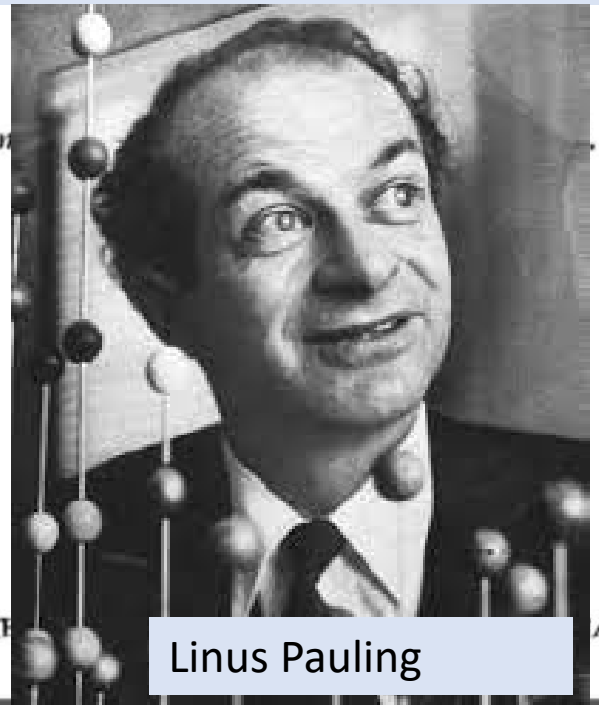
生物信息发展史：Bioinformatics

Evolving genes and proteins, 1965

Evolutionary Divergence and

Convergence in Proteins

1965年，祖卡坎德尔和鲍林提出的“分子钟”理论



Linus Pauling

生物信息发展史：Bioinformatics

1966年，我国第一次人工合成了胰岛素

人民日報

毛泽东思想武装的中国人民有志气有能力攀登前人没有攀登过的高峰

我国在世界上第一次人工合成結晶胰島素

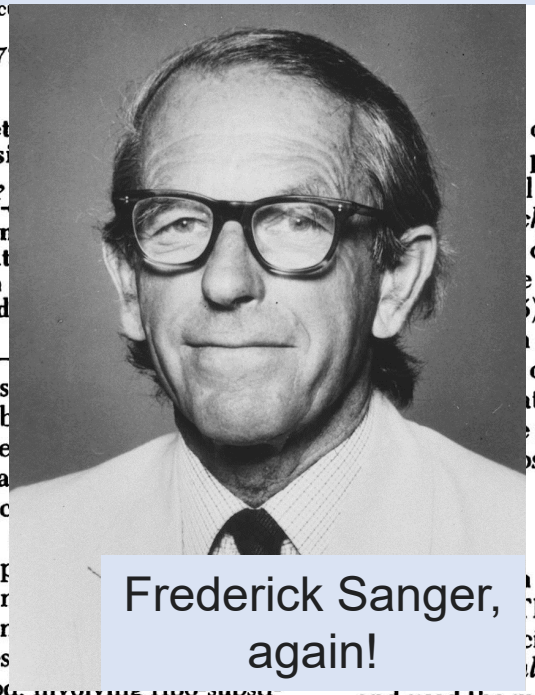
生物信息发展史：Bioinformatics

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

1977年，桑格等发表双脱氧链末端终止法，测定ϕX174序列。

Medical Research Council Laboratory of Molec

Contributed by F. Sanger, October 3, 197



Frederick Sanger, again!

ABSTRACT **A new method for det**
quences in DNA is described. It is si
minus" method [Sanger, F. & Coulson,
94, 441–448] but makes use of the 2′,3′-
cleoside analogues of the normal deoxyn
which act as specific chain-terminat
polymerase. The technique has been
bacteriophage ϕX174 and is more rapid
either the plus or the minus method.

The "plus and minus" method (1) is
simple technique that has made possib
the sequence of the genome of bacte
depends on the use of DNA polymera
regions of the DNA under controlled c
method is considerably more rapid
available techniques, neither the "p
method is completely accurate, and in
quence both must be used together, ar
tory data are necessary. W. M. Barnes
has recently developed a third method, involving ribo substi
tution, which has certain advantages over the plus and minus

of ribose in which the 3′-hydroxyl group is ori-
position with respect to the 2′-hydroxyl group.
l (ara) nucleotides act as chain terminating in-
cherichia coli DNA polymerase I in a manner
ddT (4), although synthesized chains ending in
e further extended by some mammalian DNA
). In order to obtain a suitable pattern of bands
extensive sequence can be read it is necessary
of terminating triphosphate to normal triphos-
t only partial incorporation of the terminator
dideoxy derivatives this ratio is about 100, and
syl derivatives about 5000.

METHODS

of the Triphosphate Analogues. The prepa-
TP has been described (6, 7), and the material is
cially available. ddA has been prepared by
l. (8). We essentially followed their procedure
and used the methods of Tener (9) and of Hoard and Ott (10)
to convert it to the triphosphate, which was then purified on

生物信息发展史：Bioinformatics

1988年，人类基因组计划提出

Perspective

A Turning Point in Cancer Research: Sequencing the Human Genome

RENATO DULBECCO

ONE OF THE GOALS OF CANCER RESEARCH IS TO ASCERTAIN the mechanisms of cancer. Efforts in this direction have been made by using model systems of limited complexity, such as cancer cells in vitro and oncogenic viruses. The use of cell cultures avoided the complexity of the whole animal but not the complexity of the animal genome. The use of oncogenic viruses seemed to circumvent this complexity by replacing it with the extraordinary simplicity of the viral genome. This simplicity made the study of viruses very productive. The persistence of the transformed state in a cell clone could be explained by the persistence of the viral genome in cells (*1*); genetic and molecular results showed that transformation is the consequence of the expression of one or a few viral genes. Finally, the viral transforming genes, or "oncogenes," and the proteins they specify were identified. The crowning development was the demonstration that in retroviruses the oncogenes are picked up from the cellular genome during the viruses' most recent history (*2*). As a result of these studies, cancer seemed to be locked to the expression of some viral gene; the possibility of a "hit-and-run" mechanism, in which the virus alters the cell and then vanishes, seemed excluded. Two types of oncogenes were identified: some which immortalize cells, and others which make them tumorigenic (*3*). In most cases oncogenes of both types are needed to cause a continuously growing tumor.

Subsequent work, however, blurred the distinction between immortalizing and transforming oncogenes by showing that their effects differ in primary cultures or permanent lines and in cells of different species (*4*). These findings suggested that the state of the cellular genes is important for the effect of oncogenes, in agreement with the great differences in cancer incidence and in the effects of chemical or viral carcinogens in different species.

These studies dealt with the initial cancer events. But natural cancers evolve slowly toward malignancy through many definable stages in a process called "progression" (*5*), which is the least understood but probably the most crucial phase in the generation of malignancy. Progression generates the marked heterogeneity of cancers (*6*) and their many chromosomal abnormalities (*7*); it must be differentiated from the initial action of oncogenes (*8*). Progression is observed in cells transformed by viruses. This is the case, for instance, of bursal lymphomas induced by avian leukosis viruses (*9*), of viral T-cell lymphomas in mice (*10*), and of leukemogenesis by Friend leukemia virus in cultures of mouse bone marrow cells (*11*). Stepwise transformation is observed also with DNA viruses (*12*). Fibroblastic cells from a variety of organs of a transgenic mouse containing *myc* and simian virus 40 (SV40) sequences, although expressing SV40 T antigen, were normal but became gradually transformed upon cultivation (*13*). In all these cases cellular changes occurring during culture growth determined full transformation. The "hit-and-run" hypothesis of viral transformation must be reconsidered.

A clue as to what these changes are is obtained by examining the heterogeneity of chemically induced rat mammary carcinomas with respect to several well-characterized markers. The expression of the markers is altered in different ways in different parts of the same cancer; the alterations seem to be clonal, being uniform in small parts of a tumor but different in adjacent parts (*14*). The closeness of the parts makes it unlikely that the differences are due to the environment; it is more likely that they are caused by structural changes of the genes, as is also suggested by the chromosomal rearrangements observed in cancers (*15*) and by the finding that each chemically or radiation-induced mouse sarcoma expresses a different class I major histocompatibility antigen, probably produced through gene rearrangement (*16*).

A major gap in our understanding of cancer is how the activity of an oncogene is related to the events of progression. But the first task is to ascertain whether the DNA of an advanced cancer is as heterogeneous as the phenotype of its cells. If it is so, a new field of cancer research opens up, possibly leading to the discovery of the genes whose activity or inactivity is responsible for infiltration and metastasis.

We are at a turning point in the study of tumor virology and cancer in general. If we wish to learn more about cancer, we must now concentrate on the cellular genome. We are back to where cancer research started, but the situation is drastically different because we have new knowledge and crucial tools, such as DNA cloning. We have two options: either to try to discover the genes important in malignancy by a piecemeal approach, or to sequence the whole genome of a selected animal species. The former approach seems less formidable, but it will still require a vast investment of research, especially if the important genes differ in cancers of different organs and if they encode regulatory proteins. A major difficulty for conventional approaches is the heterogeneity of tumors and the lack of cultures representative of the various cell types present in a cancer. I think that it will be far more useful to begin by sequencing the cellular genome. The sequence will make it possible to prepare probes for all the genes and to classify them for their expression in various cell types at the level of individual cells by means of cytological hybridization. The classification of the genes will facilitate the identification of those involved in progression.

In which species should this effort be made? If we wish to understand human cancer, it should be made in humans because the genetic control of cancer seems to be different in different species. Research on human cancer would receive a major boost from the detailed knowledge of DNA. Humans would become the preferred experimental species for cancer research with cells in culture or in immunodeficient mice. Because cancer could be defined in molecular terms, the agents capable of inducing cancer in humans could be identified by the combination of in vitro and epidemiological studies. Knowledge of the genes involved in progression would open new therapeutic approaches, which might lead to a general cancer cure if progression has common features in all cancers.

Knowledge of the genome and availability of probes for any gene would also be crucial for progress in human physiology and pathology outside cancer; for instance, for learning about the regulation of individual genes in various cell types. Many fields of

The author is in the Monoclonal Antibody Laboratory of the Armand Hammer Cancer Center, the Salk Institute, La Jolla, CA 92037.

人类基因组计划

开始于1988年

由美国能源部和国家医学研究院发起

美国、英国、法国、德国、日本和中国参加

经费三十亿美元

主要目标：

1、人类基因组测序

2、各种模式生物基因组测序

3、推动全基因组水平的高通量技术的发展

生物信息发展史：Bioinformatics

RESEARCH ARTICLE

Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, [illegible]
Hamilton O. Smith, J. Craig Venter†

1995年，H. influenza (流感嗜血杆菌)基因组: 第一个测序成功的基因组。

An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Haemophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The *H. influenzae* Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.

A prerequisite to understanding the complete biology of an organism is the determination of its entire genome sequence. Several viral and organellar genomes have

Homo sapiens (*11*). These projects, as well as viral genome sequencing, have been based primarily on the sequencing of clones usually derived from extensively mapped restriction

natural host is human. Six *H. influenzae* serotype strains (a through f) have been identified on the basis of immunologically distinct capsular polysaccharide antigens. Non-typeable strains also exist and are distinguished by their lack of detectable capsular polysaccharide. They are commensal residents of the upper respiratory mucosa of children and adults and cause otitis media and respiratory tract infections, mostly in

ly reduced the incidence of the disease in Europe and North America.

Genome sequencing. The strategy for a shotgun approach to whole genome sequencing is outlined in Table 1. The theory follows from the Lander and Waterman (*14*) application of the equation for the Poisson distribution. The probability that a base is not sequenced is $P_o = e^{-m}$, where m is the sequence coverage. Thus after 1.83 Mb of sequence has been randomly generated for the *H. influenzae* genome ($m = 1$, 1 × coverage), $P_o = e^{-1} = 0.37$ and approximately 37 percent of the genome is unsequenced. Fivefold coverage (approximately 9500 clones sequenced from both insert ends and an average sequence read length

山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

生物信息发展史：Bioinformatics

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

*A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the oper
the 20th century[1–3] sparked a scientific quest to un
nature and content of genetic information that h
biology for the last hundred years. The scientific pr
falls naturally into four main phases, corresponding r
four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

The sequence of the human genome is of interest in several respects. It is the largest genome to be extensively sequenced so far, being 25 times as large as any previously sequenced genome and eight times as large as the sum of all such genomes. It is the first vertebrate genome to be extensively sequenced. And, uniquely, it is the genome of our own species.

Much work remains to be done to produce a complete finished sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change as the sequence is finished, many points are already clear.

- The genomic landscape shows marked variation in the distribution of a number of features, including genes, transposable elements, GC content, CpG islands and recombination rate. This gives us important clues about function. For example, the developmentally important HOX gene clusters are the most repeat-poor regions of the human genome, probably reflecting the very complex
- The full set of proteins (the proteome) encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark G-bands' in karyotypes.
- Recombination rates tend to be much higher in distal regions (around 20 megabases (Mb)) of chromosomes and on shorter chromosome arms in general, in a pattern that promotes the occurrence of at least one crossover per chromosome arm in each meiosis.
- More than 1.4 million single nucleotide polymorphisms (SNPs) in the human genome have been identified. This collection should allow the initiation of genome-wide linkage disequilibrium mapping of the genes in the human population.

In this paper, we start by presenting background information on the project and describing the generation, assembly and evaluation of the draft genome sequence. We then focus on an initial analysis of the sequence itself: the broad chromosomal landscape; the repeat elements and the rich palaeontological record of evolutionary and biological processes that they provide; the human genes and proteins and their differences and similarities with those of other

2001年，人类基因组草图公布。

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1]
Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1]
George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5]
Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9]
...os,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1]
...har Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1]
...Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1]
...war Chandramouliswaran,[1] Rosane Charlab,[1]
...rancesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1]
...Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1]
Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1]
Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1]
Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1]
Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12]
Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1]
Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1]
Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1]
Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1]
Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1]
Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1]
Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1]
Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1]
Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1]
Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1]
Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1]
Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1]
Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1]
Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1]
Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1]
Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1]
Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1]
Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1]
Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1]
Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1]
Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1]
Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1]
Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1]
Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1]
Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1]
Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1]
Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1]
Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1]
Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1]
Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1]
Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

生物信息发展史：Bioinformatics

nature Vol 437|15 September 2005|doi:10.1038/nature03959

ARTICLES

Genome sequencing in microfabricated high-density picolitre reactors

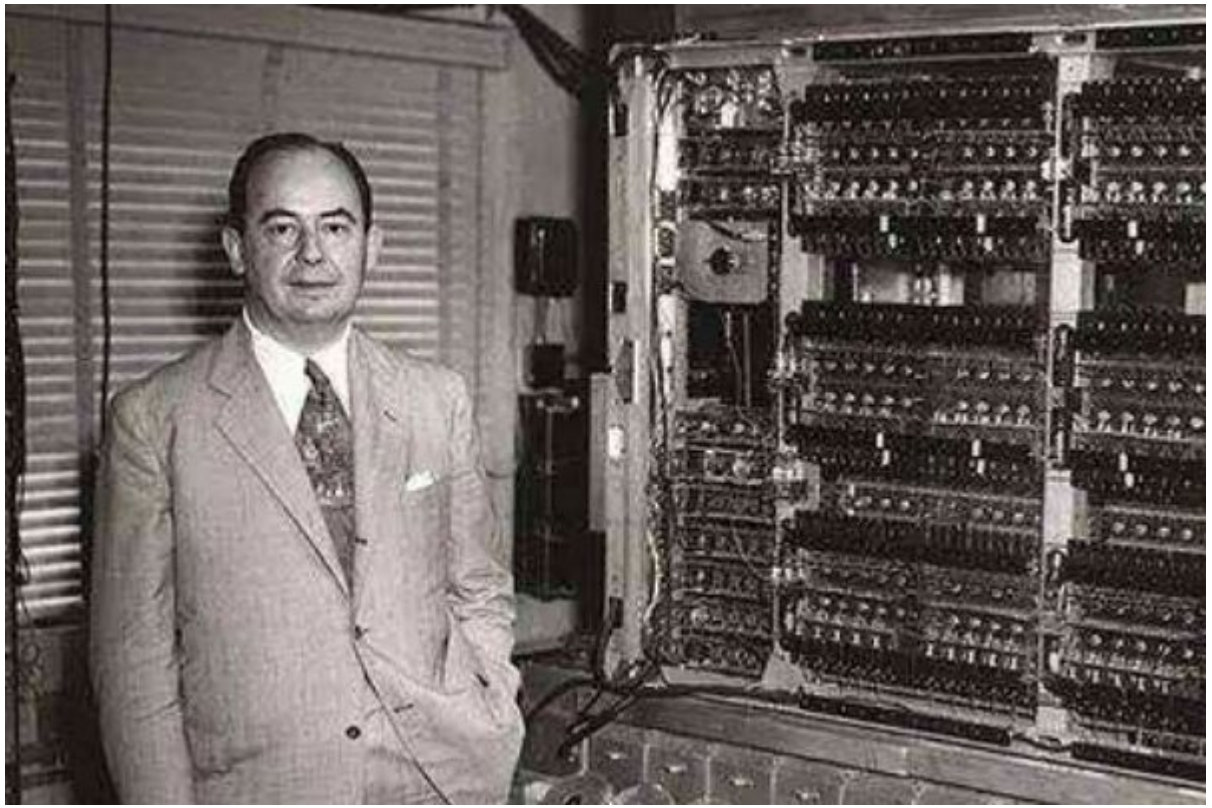
Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gom[illegible]yk[1], Szilveszter C. J[illegible] Kim[1], James R. Knigh[illegible] Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

2005年，新一代测序技术出现。

The proliferation of large-scale DNA-sequencing projects in recent years has driven a search for alternative methods to reduce time and cost. Here we describe a scalable, highly parallel sequencing system with raw throughput significantly greater than that of state-of-the-art capillary electrophoresis instruments. The apparatus uses a novel fibre-optic slide of individual wells and is able to sequence 25 million bases, at 99% or better accuracy, in one four-hour run. To achieve an approximately 100-fold increase in throughput over current Sanger sequencing technology, we have developed an emulsion method for DNA amplification and an instrument for sequencing by synthesis using a pyrosequencing protocol

生物信息发展史：Bioinformatics

约翰·冯诺依曼和计算机

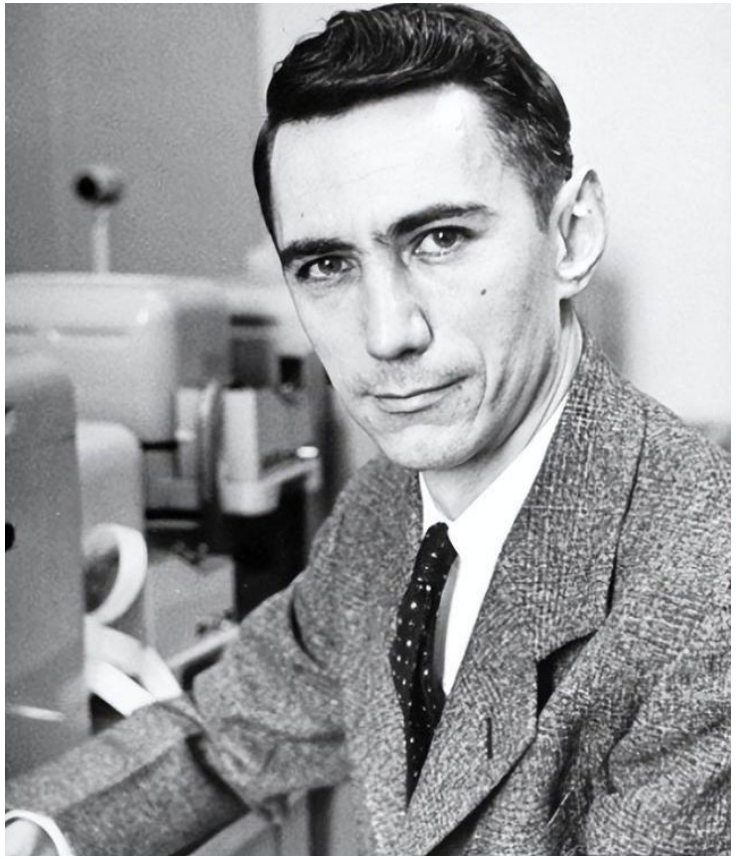


- 算符环理论
- 博弈论
- 蒙特卡洛方法
- 冯诺依曼体系

计算机制造的三个基本原则，即采用二进制逻辑、程序存储执行以及计算机由五个部分组成(运算器、控制器、存储器、输入设备、输出设备)

生物信息发展史：Bioinformatics

信息论之父: 克劳德·艾尔伍德·香农



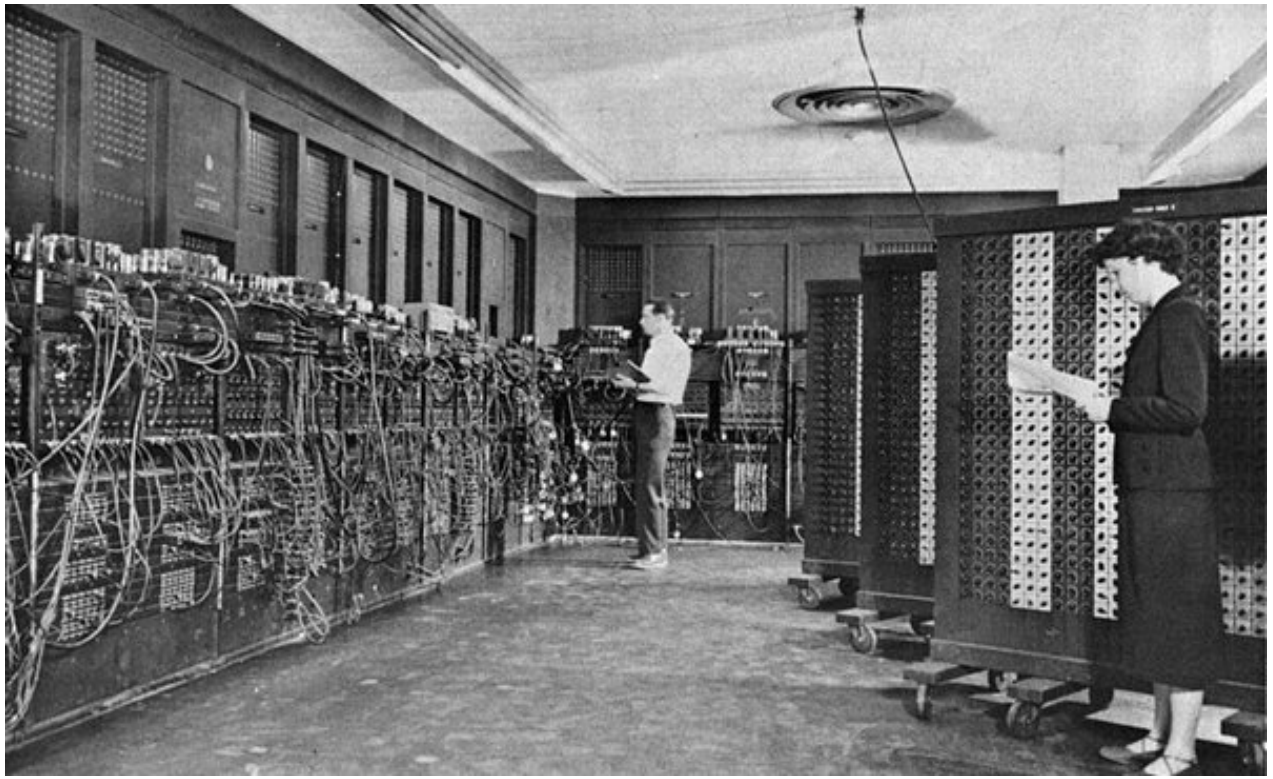
$$H(X) = -\sum_{x \in \aleph} p(x) \log p(x)$$

生物信息发展史：Bioinformatics

计算机科学与人工智能之父:艾伦·麦席森·图灵

生物信息发展史：Bioinformatics

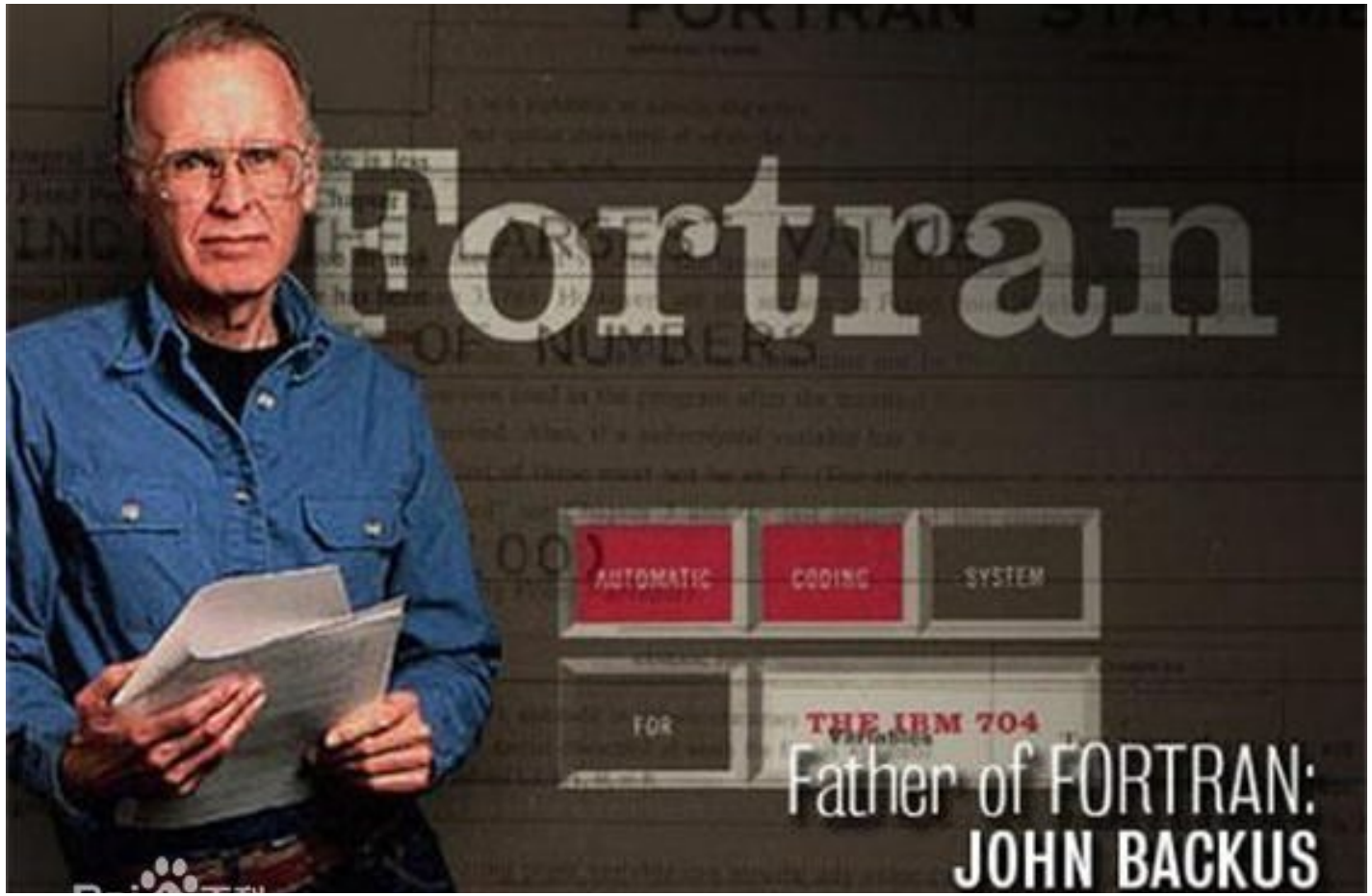
1946年，第一台计算机，ENIAC（埃尼亚克）



ENIAC：长30.48米，宽1米，占地面积约63平方米，30个操作台，约相当于10间普通房间的大小，重达30吨，耗电量150千瓦，造价48万美元。它包含了17,468 真空管7,200水晶 二极管, 1,500 中转, 70,000 电阻器, 10,000 电容器，1500继电器，6000多个开关，每秒执行5000次加法或400次乘法，是继电器计算机的1000倍、手工计算的20万倍。

生物信息发展史：Bioinformatics

1951年，IBM公司的约翰·贝克斯在纽约正式对外发布Fortran语言，称为FORTRAN I。



生物信息发展史：Bioinformatics

1969-1991: WWW的诞生

- 1969年，ARPANet建立，对计算机网络技术发展做出重要贡献。
- 1970年后，出现了E-mail、Ethernet、TCP协议。
- 1980年后，以IBM为代表的个人计算机开始普及。
- 1991年，World Wide Web协议被建立。

生物信息发展史：Bioinformatics

1991: Linux操作系统

Linus Torvalds

生物信息学大事记

1952年，肯德鲁用计算机程序来解析蛋白结构。（acta Cryst，1952）

1962年，**Dayhoff**开发序列分析软件COMPROTEIN。

1965年，**Dayhoff**出版一个蛋白质数据库Atlas（第一年65条序列），发展为1983年的PIR。

1966年，**Dayhoff**对蛋白质家族进化深入研究。（ Science，1966 ）

1967年，Fitch发表系统发育树（ Science，1967 ）

1970年，Hesper提出“Bioinformatics”单词，生物信息学概念被定义。

1970年，Needleman和Wunsch提出全局比对算法。（J. Mol. Biol.，1970）

1977年，Protein Data Bank（PDB）数据库建立。

1978年，**Dayhoff**提出氨基酸序列比对的PAM矩阵。

生物信息学大事记

Dr. Margaret Oakley Dayhoff
The Mother of Bioinformatics

生物信息学大事记

1981年，Smith和Waterman发表**局部比对算法**。（J. Mol. Biol., 1981）

1982年，建立核酸序列数据库**Genbank**（最初606条序列）。

1983年，建立Protein information Resource（PIR）蛋白数据库。2003年（4200万）

1988年，Lipman和Pearson发表FastA算法。（PNAS, 1988）

1990年，Altschul发表**Blast算法**。（J. Mol. Biol., 1990）

1997年，Altschul发表Gapped BLAST和PSI-BLAST算法。（Nucleic Acids Research）

1997年，Chris Burge等发明了GENSCAN算法。（J. Mol. Biol., 1997）

2002年，Kent建立BLAT算法。（Genome Research, 2002）

2003年，Ouzounis对前期的生物信息学发展进行了总结。（BIOINFORMATICS, 2003）

生物信息学发展的四个阶段

(1) 萌芽期(60-70年代)：以Dayhoff的替换矩阵和Neelleman-Wunsch算法为代表，它们实际组成了生物信息学的一个最基本的内容和思路：序列比较。它们的出现，代表了生物信息学的诞生(虽然“生物信息学”一词很晚才出现)；

(2) 形成期(80年代)：以分子数据库和FASTA等相似性搜索程序为代表。在这一阶段，生物信息学作为一个新兴学科已经形成，并确立了自身学科的特征和地位；

(3) 基因组测序时代(90年代-至今)：以模式基因组测序与BLAST为代表；

(4) 高通量测序时代（2005 -至今）：以第二和三代测序技术和基因组重测序为代表。

我国生物信息大事记

1994年，中国终于获准加入互联网，并在同年5月完成全部中国联网工作。

1998年，中国人类基因组学南方中心和北方中心分别在上海和北京成立。

1999年，华大基因在北京成立。

2003年，中科院北京基因组所成立。

2008年，魏丽萍老师发表了中国生物信息学发展情况。（PLoS Computational Biology, 2008）

我国生物信息学的开拓者

郝柏林 院士
理论物理所 进化发育分析

陈润生 院士
生物物理所 ncRNA

张春霆 院士
天津大学 Z曲线DNA分析

李衍达 院士
清华大学 基因表达调控

孙之荣 教授
清华大学 分子网络分析

罗辽复 教授
内蒙古大学 基因组进化

山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

生物信息学

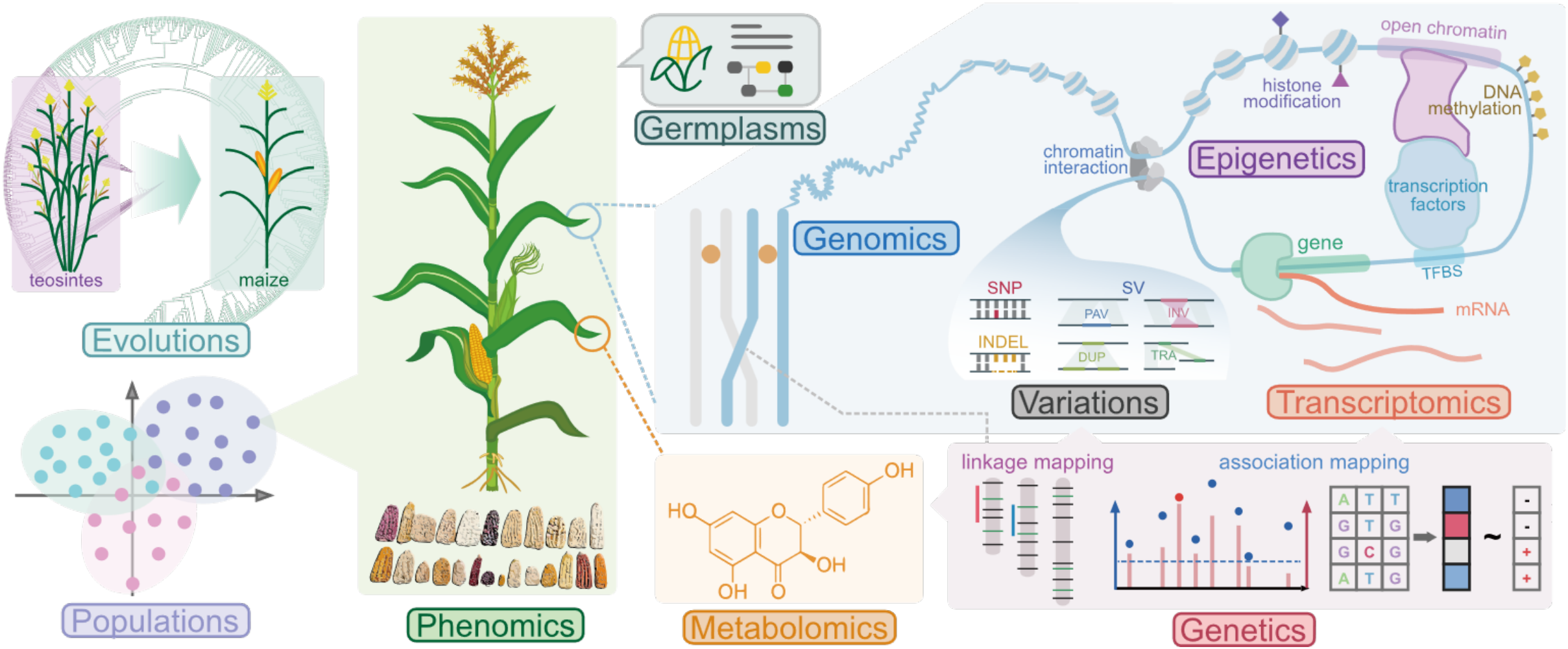
第一章: 导论

3. 生物信息学研究内容

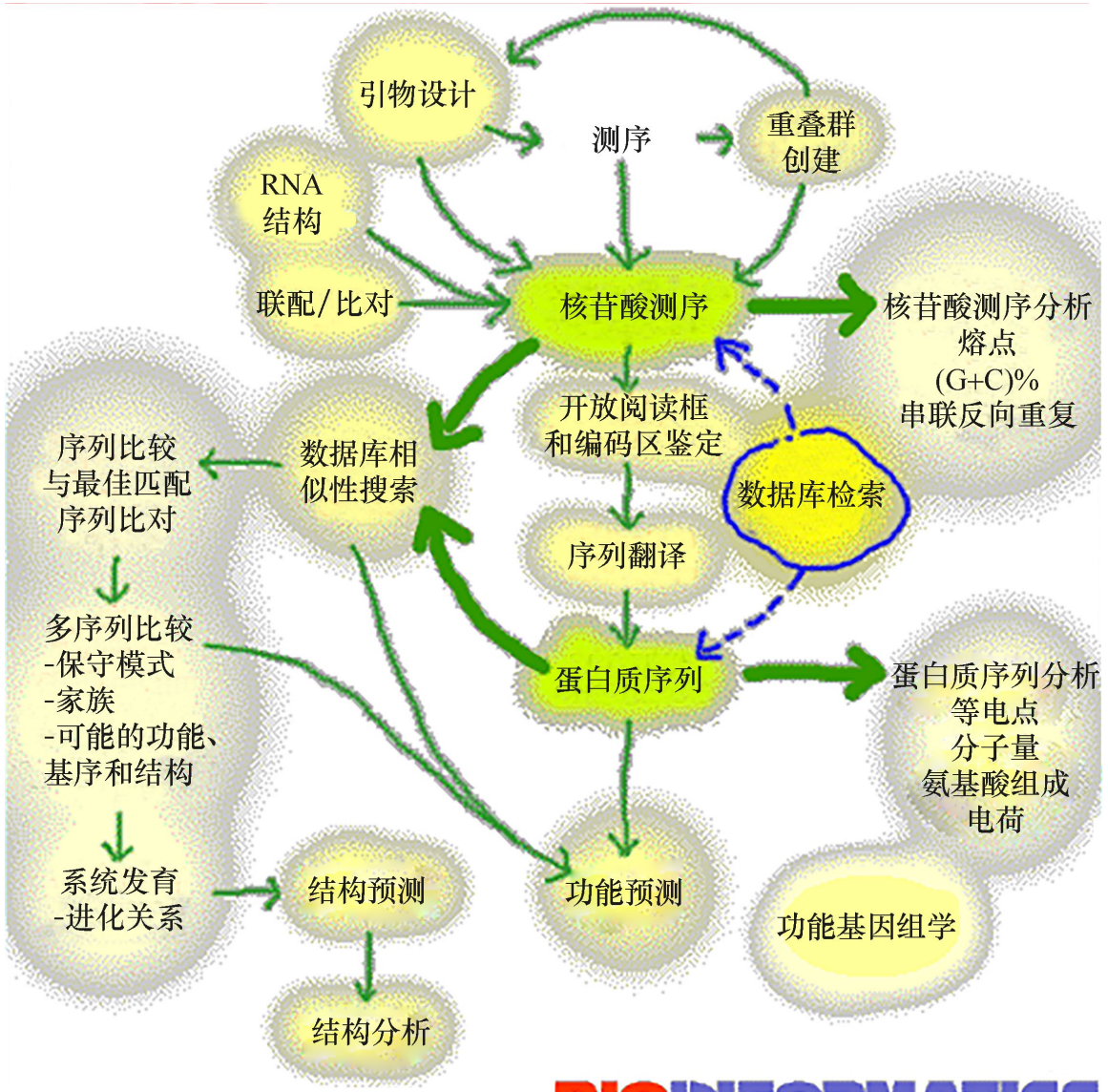
桂松涛
songtaogui@163.com

生物信息研究内容

Overview



生物信息学早期研究路线图



山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

生物信息加点技能树

> Bioinformatics ≈ Biology + Programming + Mathematics

新手村技能加点

1. 基本的 *生物学和组学* 知识: 测序(二代/三代), RNA-Seq, 甲基化, ChIP-seq, ATAC-seq, 单细胞 ...
2. 了解常用的 *生物数据库* 资源: NCBI, Ensemble, UCSC, Uniprot, CoGe, Panther ...
3. *Linux系统* 相关: Shell(Bash)编程, 作业调度, 超算使用 ...
4. 掌握至少一门 *通用编程语言* : python, perl, R, julia, rust, java, C++ ...
5. 基本的 *数据分析* 技巧: 表格操作, 数据可视化 ...
6. 基本的 *统计分析* 知识: 数据分布、假设检验、线性回归、聚类 ....
7. 常见的 *生物信息学算法* : 动态规划, 遗传算法, BWT转换, 序列比对 ...

> 延伸阅读: 生物信息等级划分

山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

生物信息进阶转职路线

Biology

Programming

Mathematics

群体遗传学
发育生物学
进化生物学
系统生物学

算法
数据结构
高性能计算

概率统计
数据挖掘
优化建模

生信工具使用者的加点建议

我以湿实验为主, 偶尔使用常规分析流程、现成的生信软件做一些基础分析, 我要学啥?

- 基本的linux操作;
- 对组学技术和测序建库原理有本的认识;
- 了解常用的生信数据库和软件;
- 善用搜索工具找到自己需要的软件, 并能看懂说明书;
- 学会向生信专业人员描述问题, 快速获得帮助;

"切, 电脑里又种不出粮食"

- 编程可以快速训练*抽象*的能力;
- 生信从业者情绪不稳定(久坐+熬夜+头发少), 对我们友好一些 🥹;

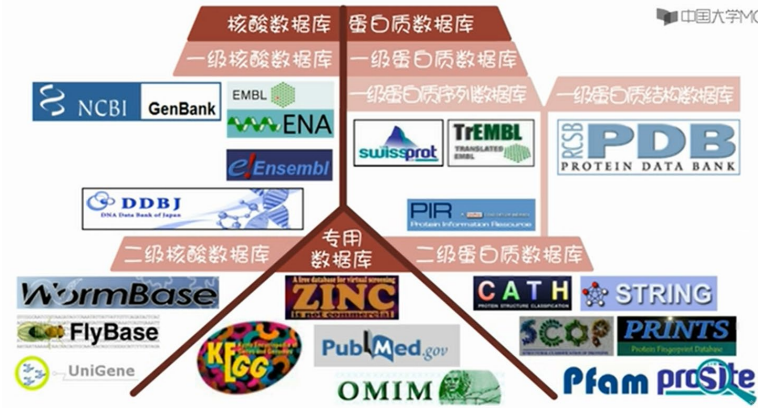
生物信息学之我见

- 快速正反馈
- 终身学习
- 头脑风暴
- 学好数理化
- 科学问题驱动

程序员有三种美德: 懒惰,急躁和傲慢

---- Larry Wall

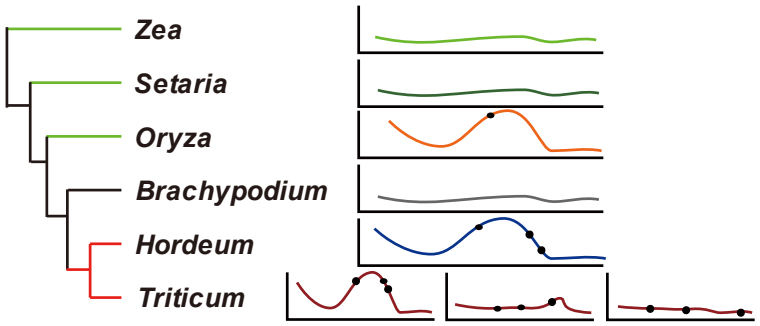
课程研究内容



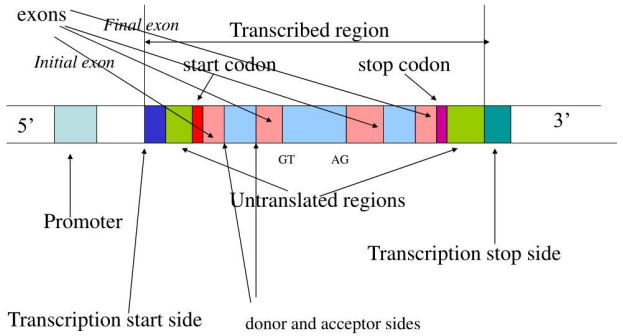
生物信息数据库及其应用

		A	T	G	C	T
	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
G	-4	-1	0	0	-2	-4
C	-6	-3	-2	-1	1	-1
T	-8	-5	-2	-3	-1	2

序列比对算法



进化分析方法



核酸序列分析

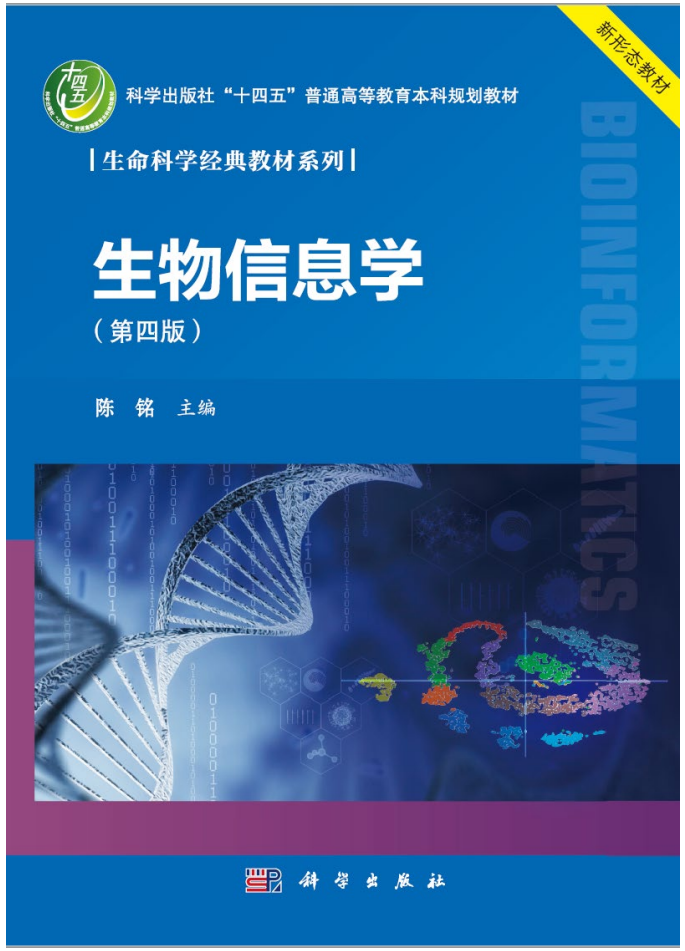
课程概览

理论课部分（16课时）

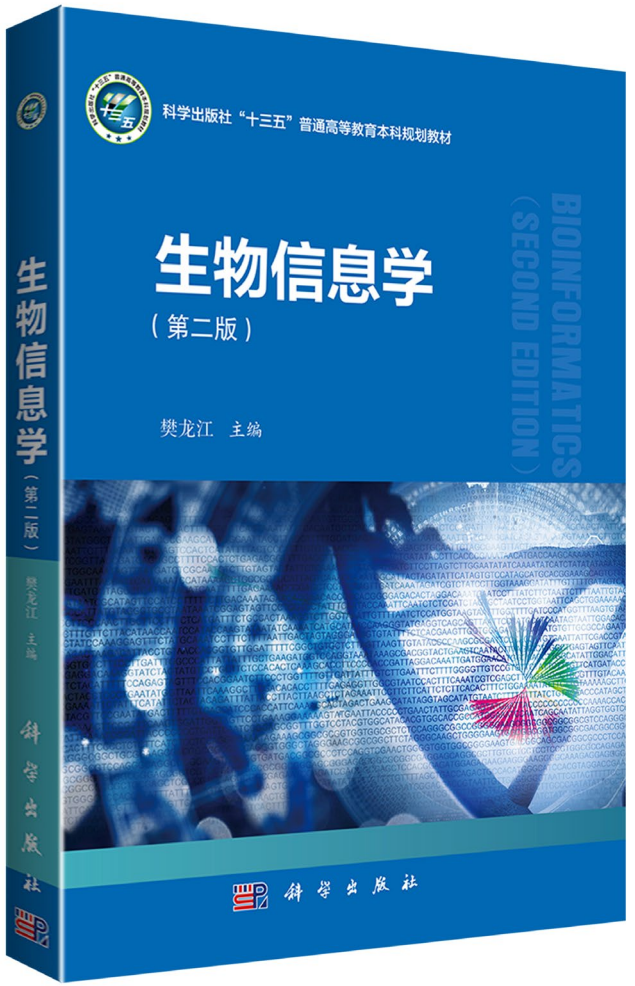
1. 生物信息学导论 （2课时）
2. 生物信息数据库（2课时）
3. 序列比对（4课时）
4. 进化分析（4课时）
5. 核酸序列分析（4课时）

实验课部分（16课时）

参考书推荐



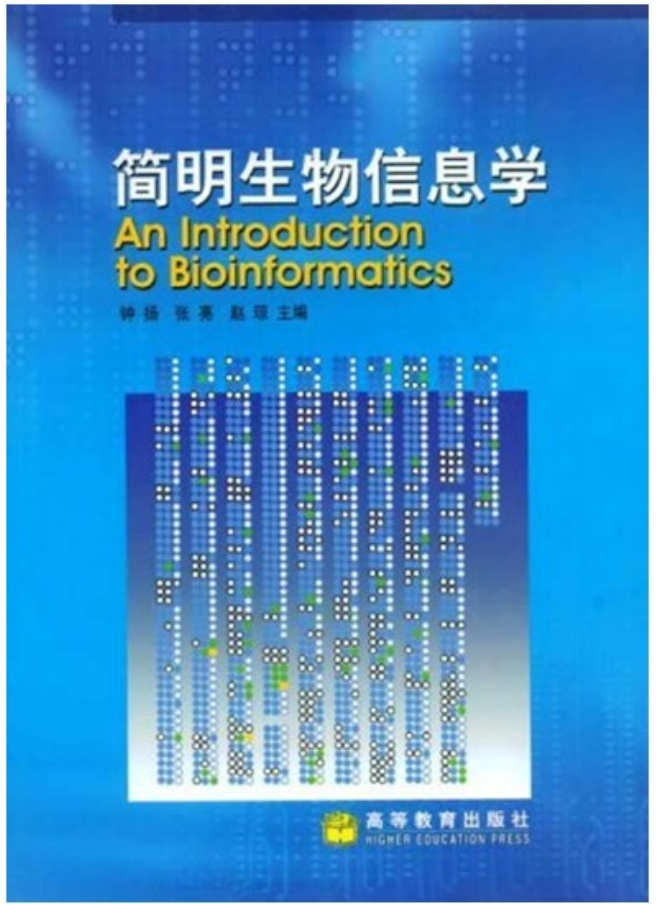
陈铭 主编
科学出版社, 2022



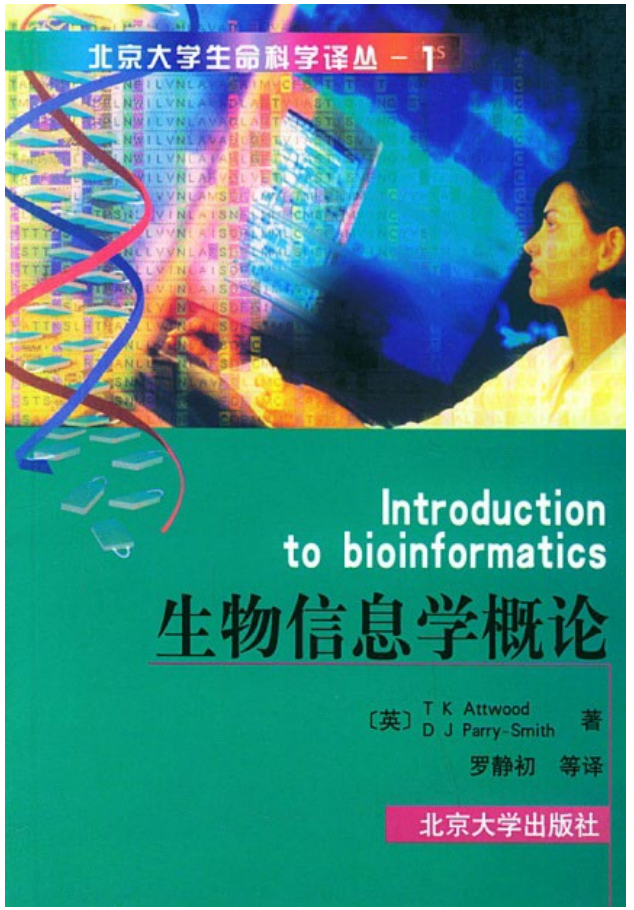
樊龙江 主编
科学出版社, 2021

吕巍 主编
高等教育出版社

参考书推荐



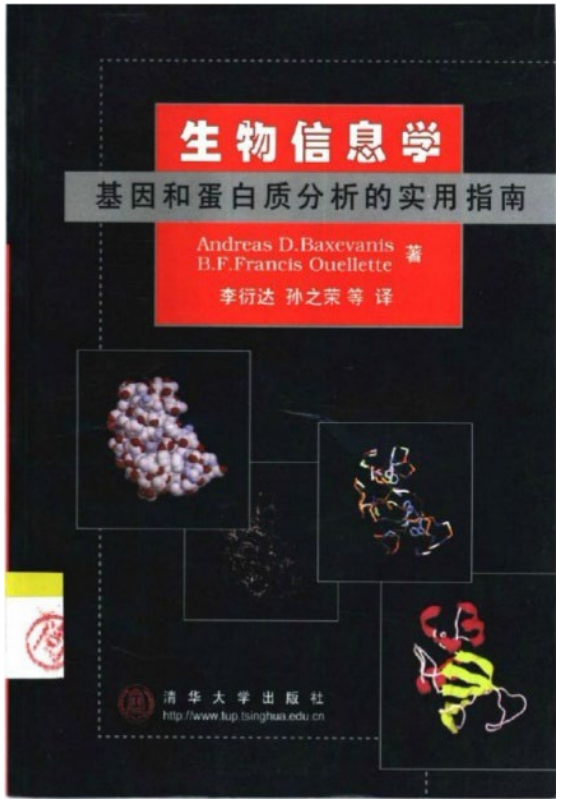
《简明生物信息学》
钟扬，张亮，赵琼 主编
高等教育出版社



《生物信息学概论》
罗静初 等译
北京大学出版社

参考书推荐


科学出版社
生命科学前沿丛书
合作策划
生物信息学
方法与实践


生物信息学
基因和蛋白质分析的实用指南
Andreas D.Baxevanis
B.F.Francis Ouellette 著
李衍达 孙之荣 等 译
清华大学出版社
http://www.tup.tsinghua.edu.cn


现代生物技术前沿
赵国屏 等 编著
生物信息学
科学出版社
www.sciencep.com

考核方式

平时表现(出勤+测验) : 20%

期末测试: 80%

绝知须躬行

- 课上内容以框架、大纲为主;
- 提供自学资源、资料;
- 90%以上的内容需要课下**自学**掌握;

山東農業大學
SHANDONG AGRICULTURAL UNIVERSITY

生物信息学

Thanks!

桂松涛

山东农业大学 生命科学学院

songtaogui@163.com