博士生课

生物信息学/系统生物学

章祥荪, 吴凌云, 王勇,张世华

http://zhangroup.aporc.org

中国科学院 数学与系统科学研究院

2011. 3. 23

第一讲　绪论

（一）什么是生物信息学/系统生物学？

（二）生物信息学的一些基本概念

（三）生物信息学/系统生物学的一些基本问题

（四）复杂网络与系统生物学

（五）结论

（一）

什么是生物信息学/系统生物学？

1.1 What Is Bioinformatics?

什么是生物信息学

- 美国NIH (美国国立卫生研究院)的定义 (2000)
  - Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

What Is Bioinformatics?

什么是生物信息学

- 美国NIH 的定义 (2000)

为深度利用生物的、医学的和行为学的数据而对计算工具和方法的的研究、开发或应用，包括对这些数据的获取、储存、组织、存档、分析或可视化。

解释一下这个定义，为什么在定义中没有出现信息这个词？

- Original data (数据/图形/声音) ----- Information (信息)
- 为特定的目的从数据中获得有效的部分称为信息
- 能减低事物不确定性的有效数据称为信息
- 从带有噪声的数据中滤出有效部分称为信息
- 从不完整数据中推得总体属性信息

我们这门课中主要针对的是“生物的”数据—OMICS （生物组学）数据

- Genomics 基因组学数据

  the term “genomics” encompasses a broader scope of biological study. A genome is the total sum of all an individual organism‘s genes. Thus, genomics is the study of all the genes of a cell, or tissue, at the DNA (genotype 基因型，生物体遗传上的组成，生物体基因的总体), mRNA (transcriptome 转录组，DNA的遗传信息被拷贝成RNA的过程), or protein (proteome 蛋白质组) levels.

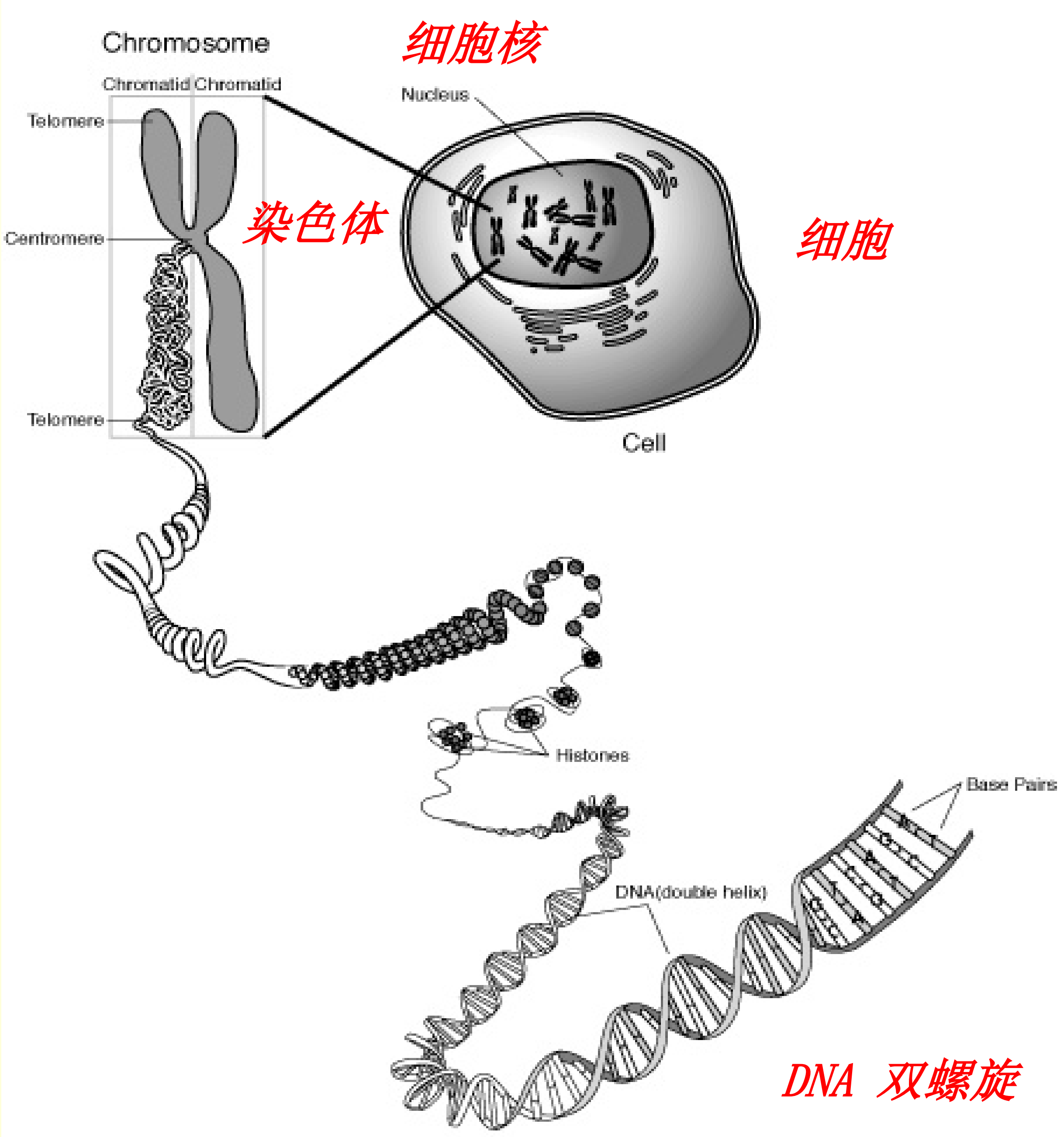
- Proteomics 蛋白组学数据

  Proteomics is the large-scale study of proteins, particularly their structures and functions

- Metabolomics 代谢组学数据

  Metabolomics is the scientific study of chemical processes involving metabolites （代谢物）. The metabolome （代谢组学）represents the collection of all metabolites in a biological cell, tissue, organ or organism.

- One of the challenges in the bioinformatics research is to integrate proteomic, transcriptomic , and metabolomic information to give a more complete picture of living organisms.

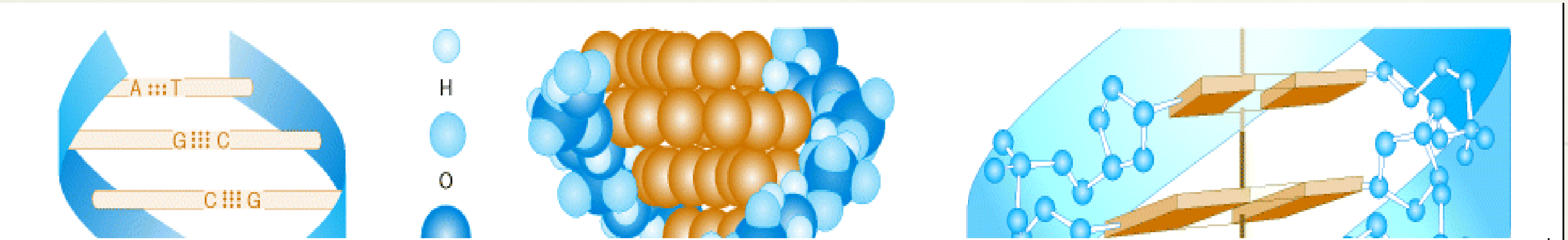


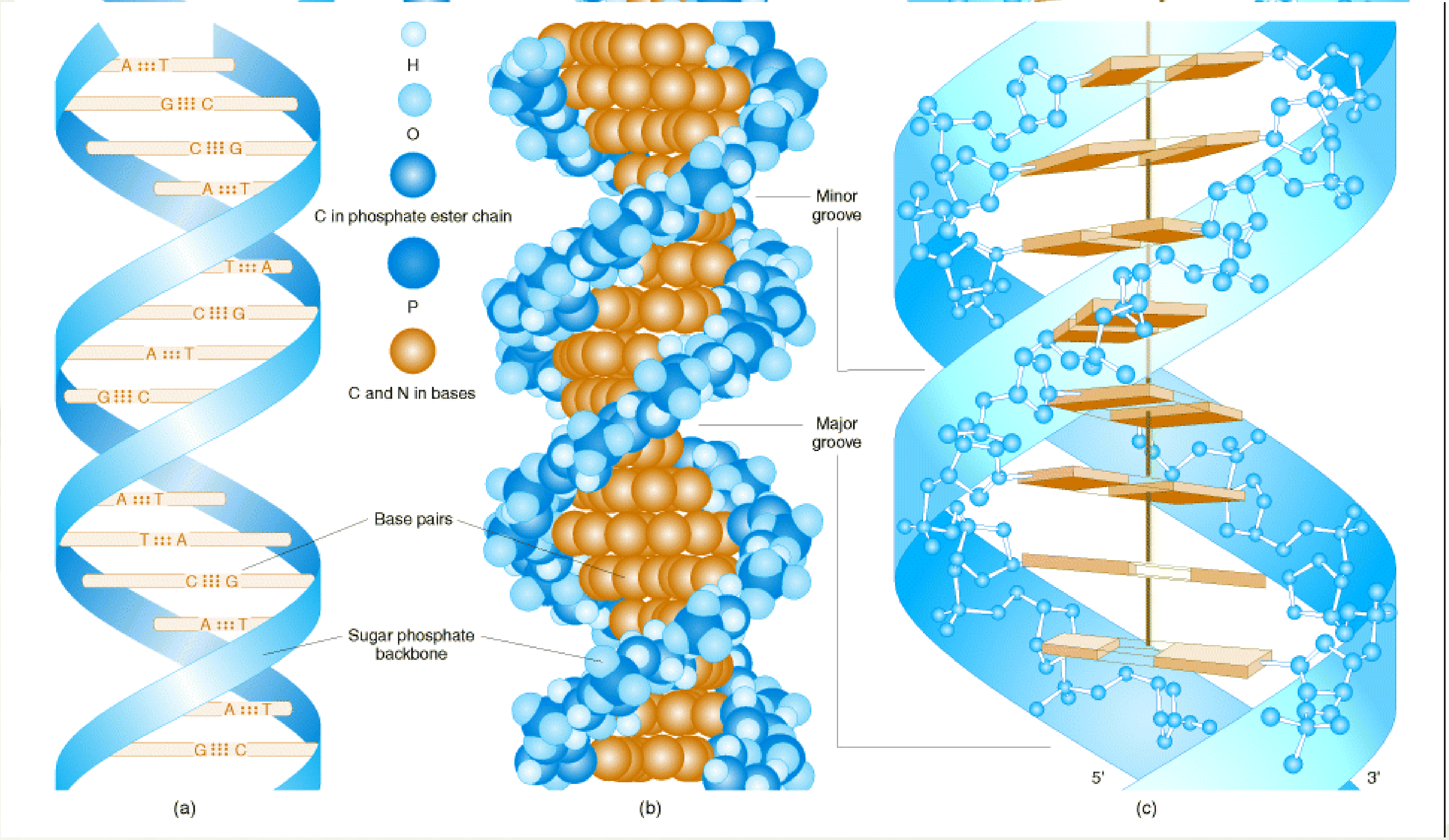
细胞核

染色体

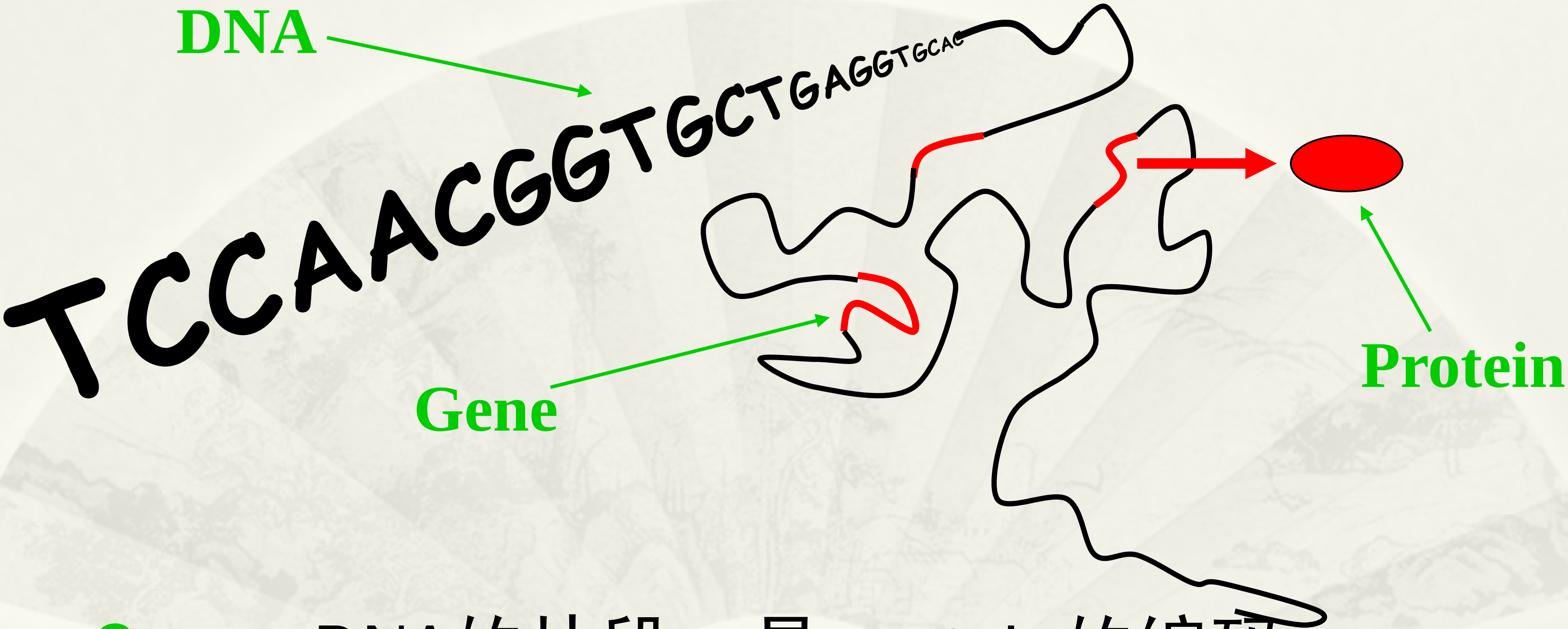
细胞

DNA 双螺旋


A ::: T
G ::: C
C ::: G
H
O


A ::: T
G ::: C
C ::: G
A ::: T
T ::: A
C ::: G
A ::: T
G ::: C
A ::: T
T ::: A
C ::: G
A ::: T
A ::: T
G ::: C
H
O
C in phosphate ester chain
P
C and N in bases
Minor groove
Major groove
Base pairs
Sugar phosphate backbone
5'
3'
(a)
(b)
(c)

DNA, Genes, and Proteins



Gene: DNA的片段，是 protein的编码

Proteins: 实现细胞的生物功能

回到生物信息学的定义

- 美国NIH 的定义 (2000)

  为深度利用生物的、医学的和行为学的数据而对计算工具和方法的的研究、开发或应用，包括对这些数据的获取、储存、**组织**、存档、**分析**或可视化。

组织 ----生物的、医学的有效信息的提取

1.2　What is Systems Biology

什么是系统生物学

一个例子:

- 传统的生物学认识：物种或生命的复杂程度应该与基因数目有直接的关系（*Reductionism，简化论，还原论*）。
- 线虫是一种低等动物，其基因组的基因数为1.9万多个。而人类基因组的基因总数仅仅是线虫的两倍。水稻基因组的基因总数在4.6万到5.5万之间，约是人的基因数的1.5倍。
- 在生命从简单到复杂，从低级到高级的进化过程中，起决定作用的不是基因个体数目，而是生命系统中简单元件的相互作用或网络的复杂性。因此有必要由整体的、合成的角度诠释生命系统。

简单的数学

- 对于一个100个不同元素的简单系统，每5个确定的元素分别完成一个功能，则该系统有20个功能；
- 对于一个有20个不同元素的较复杂的系统，任意5个元素组合会有不同的功能，就有15504种功能

1.2　什么是系统生物学

系统生物学的创始人 **Leroy Hood**，美国西雅图系统生物学研究所（全球第一个系统生物学研究所）所长：

- ***Many biological problems, particularly human diseases, fall into the category of "systems problems"***

  -- Leroy Hood

- *许多的生物问题，尤其是涉及到人类疾病方面的问题，都属于系统性的问题，不是研究清楚单个基因或者蛋白质就可以解决的。*

1.2 What is Systems Biology
什么是系统生物学(续)

Systems Biology: A Brief Overview

Hiroaki Kitano, *Science* 2002

- To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism.
- Properties of systems, such as robustness （鲁棒性）, emerge as central issues, and understanding these properties may have an impact on the future of medicine.
- However, many breakthroughs in experimental devices, advanced software, and analytical methods are required before the achievements of systems biology can live up to their much-touted (**大肆鼓吹的**) potential.

1. 2　什么是系统生物学(续)

这意味着要开发出能从整个系统中不同元素中提取出系统层面上的信息的工具；

还要有综合这些从不同生物学水平上（**DNA information, RNA information, protein information, protein interaction information, pathways and so forth**）所得到的信息的能力；

最终的目的是利用这些信息建立数学模型，能预测生物系统的结构和在受到某种刺激或扰动时生物系统的性质。

What is Systems Biology
什么是系统生物学(continue)

Systems Biology: A Brief Overview

Hiroaki Kitano, *Science* 2002, vol. 295

1) ***System structures***. These include the network of gene interactions and biochemical pathways, as well as the mechanisms by which such interactions modulate the physical properties of intracellular and multi-cellular structures.

2) ***System dynamics***. How a system behaves over time under various conditions can be understood through metabolic analysis, sensitivity analysis, dynamic analysis methods such as phase portrait and bifurcation analysis, and by identifying essential mechanisms underlying specific behaviors.

What is Systems Biology
什么是系统生物学(continue)

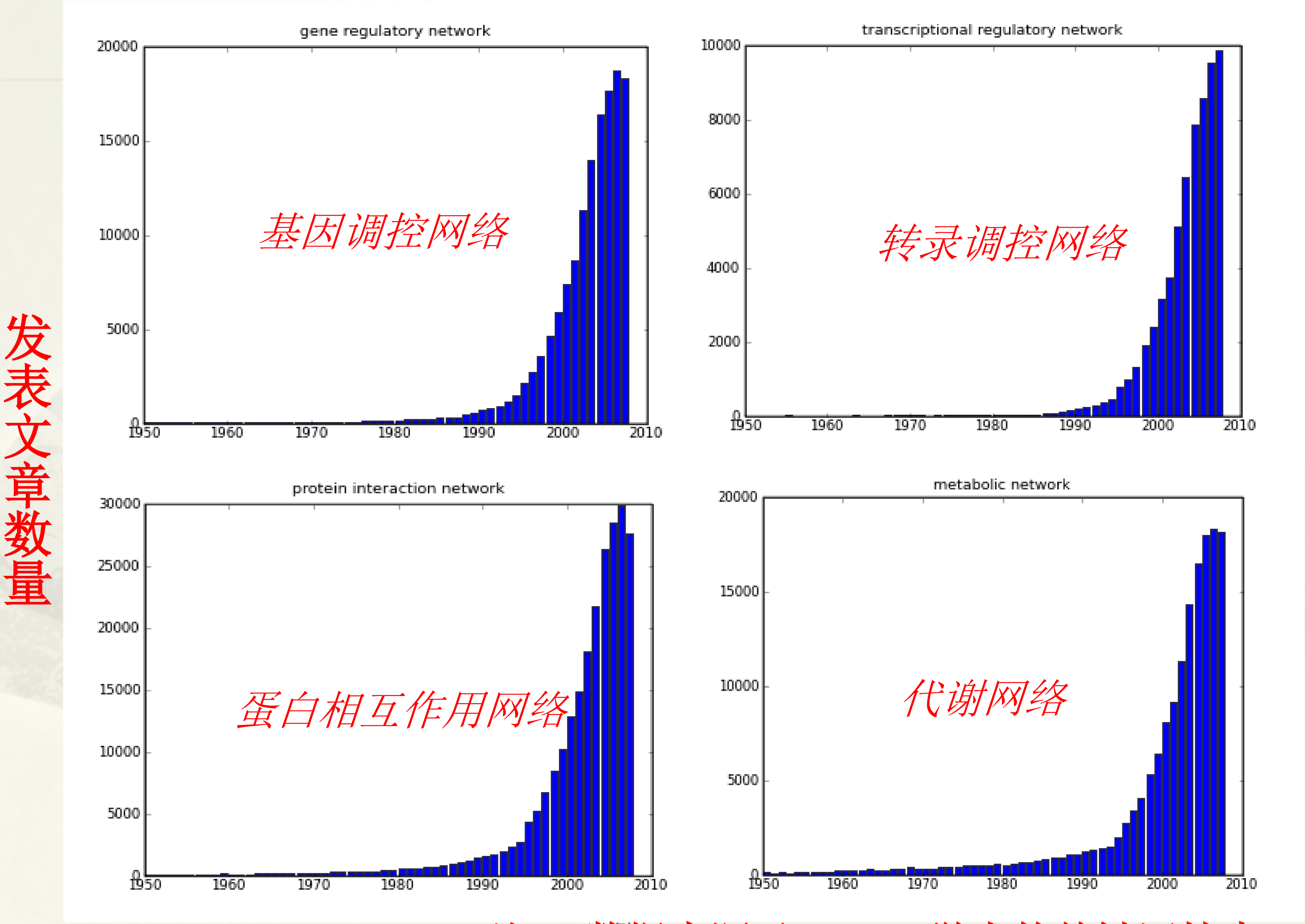
Systems Biology: A Brief Overview

Hiroaki Kitano, *Science* 2002, vol. 295

3) ***The control method***. Mechanisms that systematically control the state of the cell can be modulated to minimize malfunctions and provide potential therapeutic targets for treatment of disease.

4) ***The design method***. Strategies to modify and construct biological systems having desired properties can be devised based on definite design principles and simulations, instead of blind trial-and-error (**反复试验**).

生物分子网络是当前研究热点



注： 数据来源于Google 学术的关键词搜索

系统生物学（Systems Biology）

成为近年重要研究方向

Trey Ideker, et al, "Integrated genomic and proteomic analyses of a systemtically perturbed metabolic network", 4 May 2001 Vol 292 ***Science***

Michael T Laub, et al, "Global analysis of the genetic network controlling a bacterial cell cycle", 15 December, 2000 Vol 290, ***Science***

H. Jeong, et al."Lethality and centrality in protein networks", ***Nature*** ,Vol 411, 3 MAY 2001

George von Dassow, Eli Meir, "The segment polarity network is a robust developmental module", ***Nature*, **Vol 406,13 JULY 2000

H. Jeong, et al, "The large-scale organization of metabolics networks", ***Nature ,*** v407, 2000

Thomas Simon Shimizu, et al, "Molecular model of a lattice of signalling proteins inVolved in bacterial chemotaxis", ***Nature Cell Biology***, Vol 2, 2000

Michael B. Elowitz, et al, "A synthetic oscillatory network of transcriptional regulators``, ***Nature*** , v403, 2000

S. Kalir, et al, “Ordering genes in a flagella pathway by analysis of expression Kinetics from Living Bacteria", ***Science***, v292, 2001

Matthew Freeman, "Feedback control of intercellular signalling in development",
Nature, v408

Chunyan Xu, et al, "Overlapping activators and repressors delimit transcriptional response to receptor tyrosine kinase signals in the drosophila eye", ***Cell***, Vol.103, 2000

Thomas Surrey, Francois Nedelec, "Physical properties determining self-organization of motors and microtubules ", ***Science*** Vol 292 11 May 2001

Norbert Frey, et al, "Decoding calcium signals inVolved in cardiac growth and function ", ***Nature Medicine*** * Volume 6 * Number 11 * November 2000

Reka Albert, et al, "Error and attack tolerance of complex networks", ***Nature*** , v406, 2000

1.3 学习这门课的指导思想

- STATE-OF-THE-ART
  掌握“过程”而不是“结果”

- 各种数学分支的融会贯通，发现自己要进行补充的学问， 例如， “计算复杂性”

- 开放式的教与学

（二）
生物信息学的一些基本概念

- DNA(Deoxyribonucleic acid)，脱氧核糖核酸，是Cromosome , 染色体 (Human cells have 23 pairs of large linear nuclear chromosomes, 22 pairs of autosomes (常染色体） and one pair of sex chromosomes)的主要化学成分

- DNA的基本组成成分是核苷酸 (Nucleotide)。单个核苷酸由一个5碳糖连接一个或多个磷酸基团和一个含氮碱基组成, 硷基 有四种:

  腺嘌呤(adenine) A; 胞核嘧啶(cytosine) C;

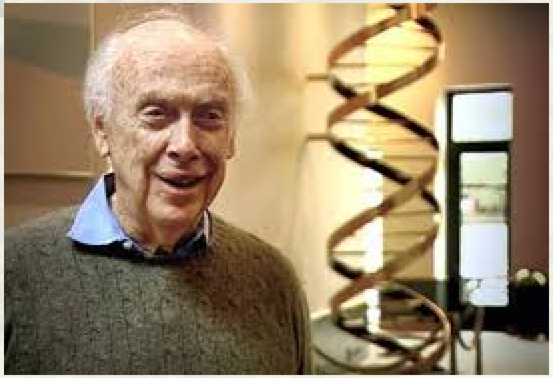
  鸟嘌呤(guanine) G; 胸腺嘧啶(thymine) T.

- DNA分子是由两条核苷酸(Nucleotide)链以互补配对(A-T, C-G)原则所构成的双螺旋结构(Watson-Click, 1953, Nature) 的分子化合物。

Rosalind Franklin, Francis Crick and James Watson

- Rosalind Franklin（1920～1958）拍摄到的DNA晶体照片，为双螺旋结构的建立起到了决定性作用。但“科学玫瑰”没等到分享荣耀，在研究成果被承认之前就已凋谢。

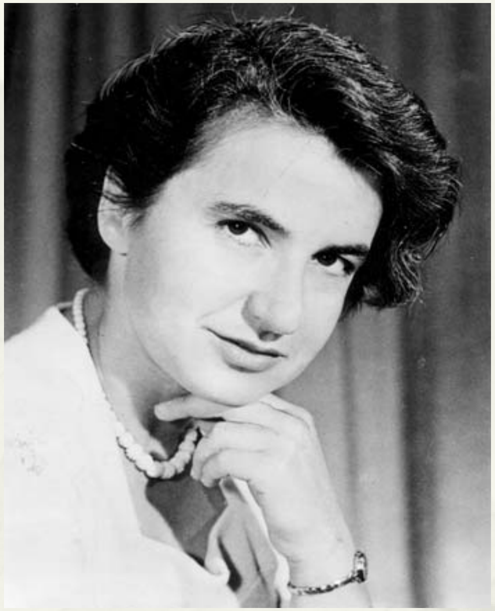
1953年的Watson和Crick都是名不见经传的小人物，37岁的Crick连博士学位还没有得到。受前人影响，他们按照3股螺旋的思路进行了很长时间的工作，结果陷于僵局。在发现正确的双股螺旋结构前2个月，他们看到蛋白质结构权威Pauling即将发表的关于DNA结构的论文，Pauling错误地确定为3股螺旋。Watson在认真考虑并向同事们请教后，决然地否定了权威的结论，在不到两个月内终于取得了后来震惊世界的成果。



James Watson was born on April 6, 1928 in Chicago, and at the age of 15 enrolled in Chicago University and majored in zoology. He received a Ph.D. in genetics from Indiana University, when he was 22 years old. In 1951, he joined Francis Crick at Cavendish Laboratory in Cambridge, England. He was only 23 years old when his greatest discovery was made

- Francis Crick was born on June 8, 1916 in Northampton, England, into a middle-class family. He began doing science experiments in his home when he was ten years old. He graduated with a degree in Physics from University College in London. He died of colon cancer in 2004 (aged 86).

- 在1953年2月底，33岁的Franklin已经在日记中写道，DNA具有两条链的结构。1953年3月17日，当Franklin将研究结果整理成文打算发表时，发现Watson和Crick破解DNA结构的消息已经出现在新闻简报中。
- 4月2日，Watson、Crick和Wilkins的文章送交《Nature》杂志，4月25日发表，接着他们在5月30日的《Nature》杂志上又发表了“DNA的遗传学意义”一文，更加详细地阐述了DNA双螺旋模型在功能上的意义。
- 她更不知道的是，Watson和Crick曾看过她拍摄的能验证DNA双螺旋结构的X射线晶体衍射照片，并由此获得了重要启发。

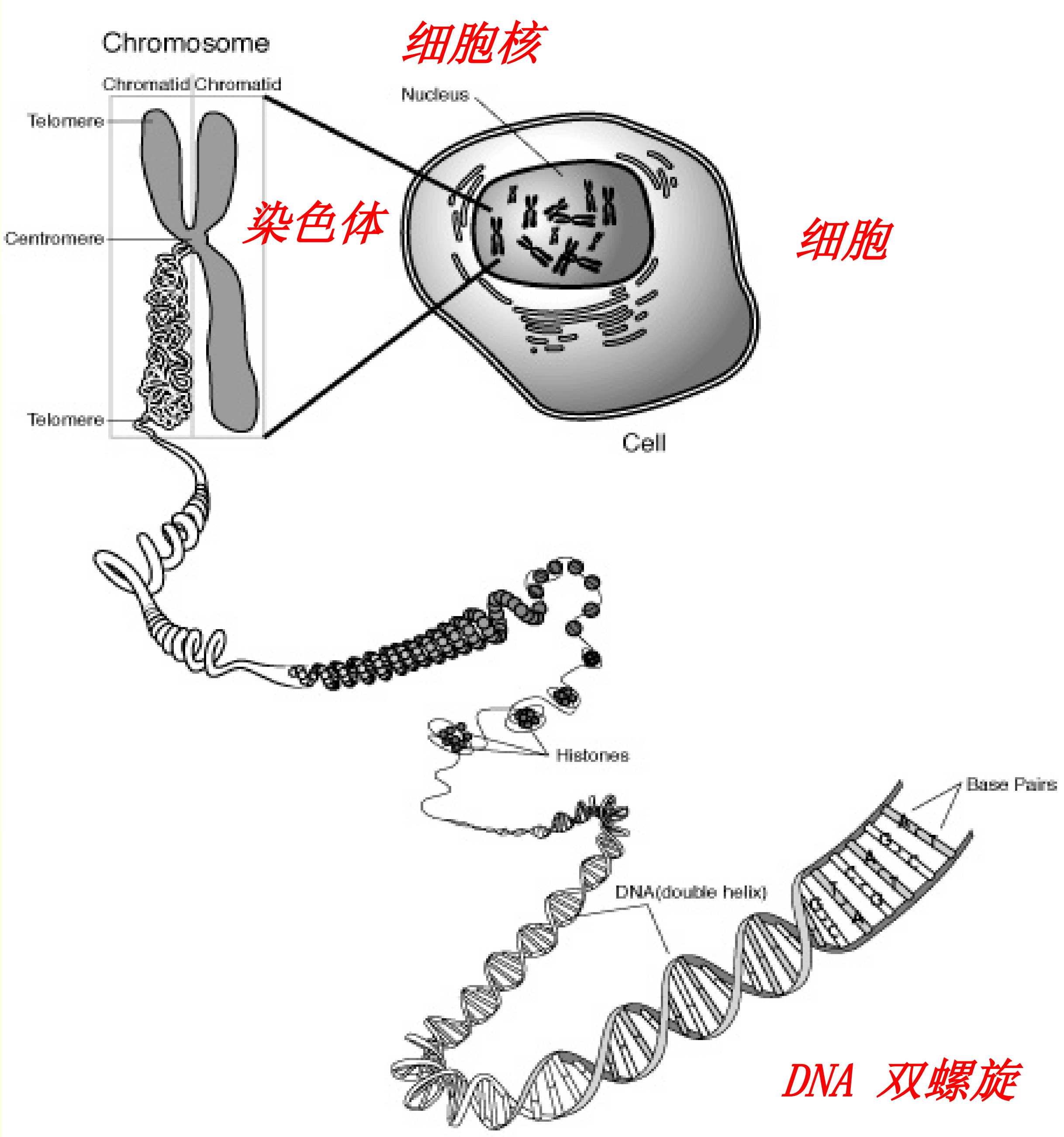


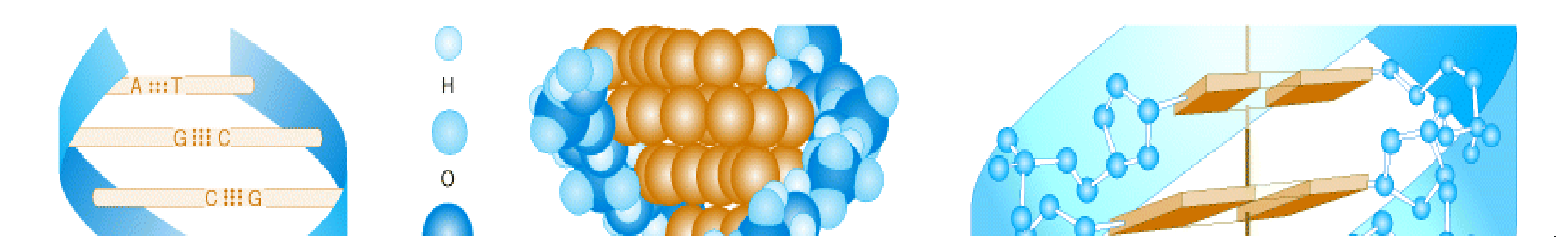
Rosalind Franklin, **Francis Crick and James Watson** (continue)

- 目前，科技界对Franklin的工作给予较高评价，对Wilkins是否有资格分享发现DNA双螺旋结构的殊荣存在很大争论。

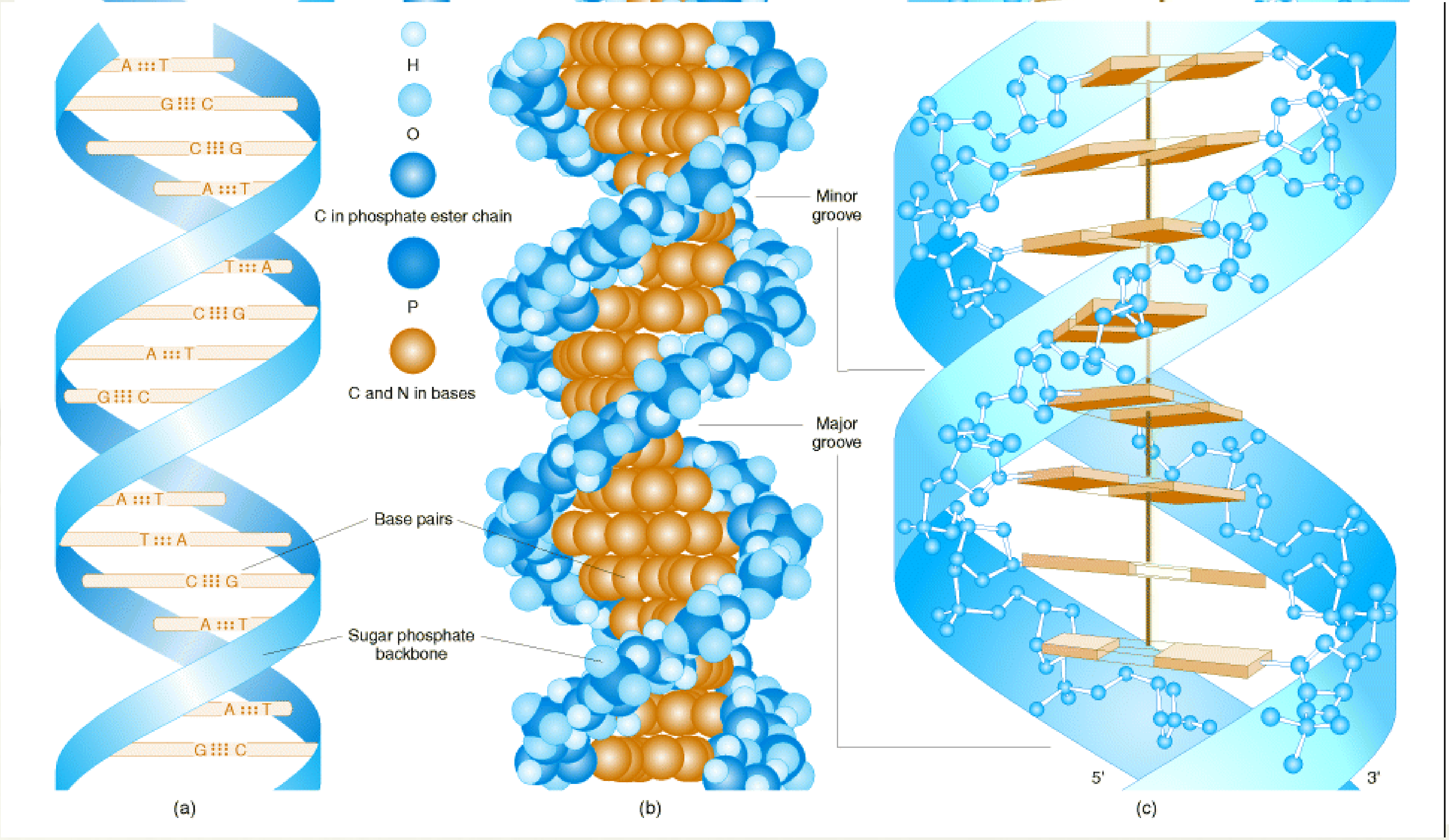
- 1962年，当Watson、Crick和Wilkins共同分享诺贝尔奖时，Franklin已经因长期接触放射性物质而患乳腺癌英年早逝。

- 这个故事的结局有些伤感。按照惯例，诺贝尔奖不授予已经去世的人。此外，同一奖项至多只能由3个人分享，假如Franklin活着，她会得奖吗？性别差异是否会成为公平竞争的障碍？后人为了这个永远不能有答案的问题进行过许多猜测与争论。那么我们应该从中吸取什么教训呢？


Chromosome
细胞核
Chromatid Chromatid
Nucleus
Telomere
染色体
细胞
Centromere
Telomere
Cell
Histones
Base Pairs
DNA(double helix)
DNA 双螺旋

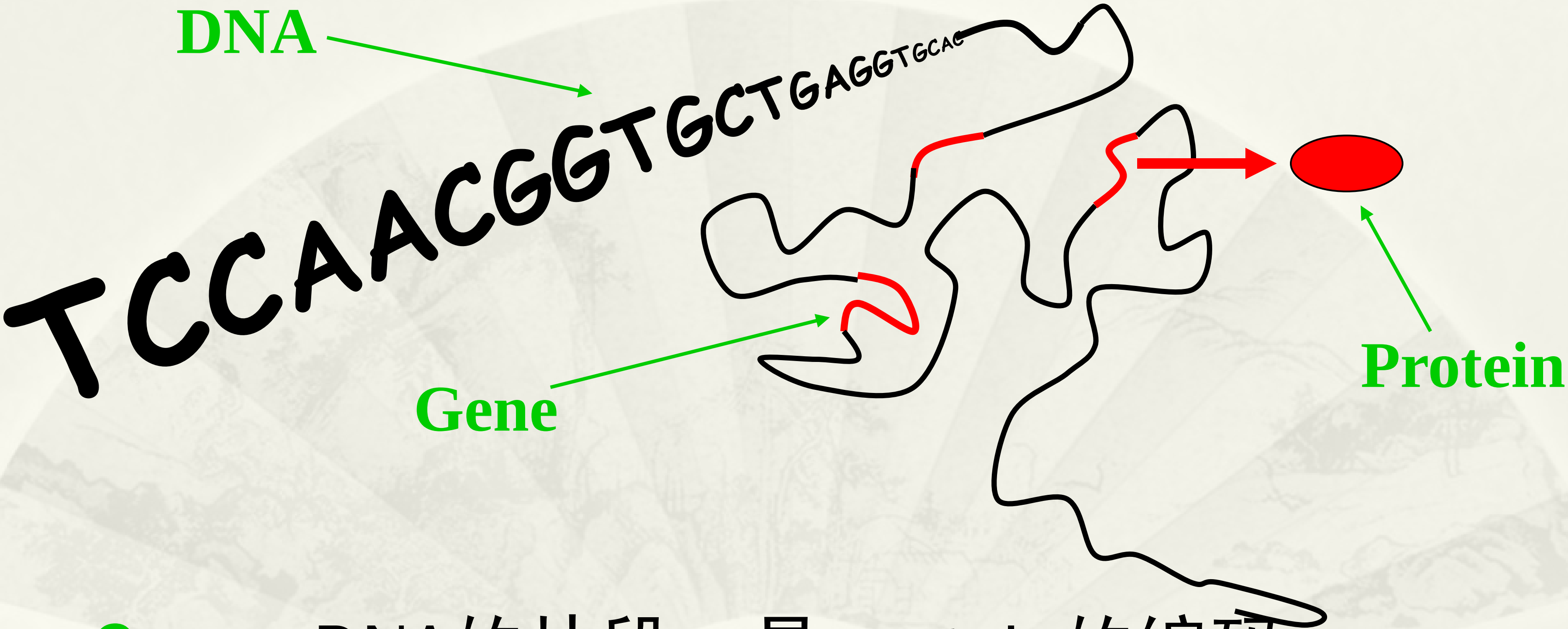


A ::: T
G ⁞⁞⁞ C
C ⁞⁞⁞ G
H
O



A ::: T
G ⁞⁞⁞ C
C ⁞⁞⁞ G
A ::: T
T ::: A
C ⁞⁞⁞ G
A ::: T
G ⁞⁞⁞ C
A ::: T
T ::: A
C ⁞⁞⁞ G
A ::: T
A ::: T
G ⁞⁞⁞ C
H
O
C in phosphate ester chain
P
C and N in bases
Minor groove
Major groove
Base pairs
Sugar phosphate backbone
5'
3'
(a)
(b)
(c)

DNA, Genes, and Proteins



Gene: DNA的片段，是 protein的编码

Proteins: 实现细胞的生物功能

How does the cell convert DNA into working proteins? — Gene expression 基因表达： Transcription （转录）and translation（翻译）

- In the first step, the information in DNA is transferred to a messenger RNA (mRNA，信使核糖核酸) molecule by way of a process called transcription (转录）. During *transcription*, the DNA of a gene serves as a template for complementary base-pairing, and an enzyme called RNA polymerase III （聚合酶）catalyzes the formation of a pre-mRNA molecule, which is then processed to form mature mRNA . The resulting mRNA is a single-stranded copy of the gene, which next must be translated into a protein molecule.

.

- During *translation*, which is the second major step in gene expression （**基因表达**）, the mRNA is “read” according to the genetic code, which relates the DNA sequence to the *amino acid* **(氨基酸)** sequence in proteins. Each group of three base pairs in mRNA constitutes a codon (密码子）, and each codon specifies a particular amino acid (hence, it is a triplet code). The mRNA sequence is thus used as a template to assemble—in order—the chain of amino acids that form a protein

所以，一切要从这里开始

$3*10^9$ （3G)

CCGGTCTCCCCGCCCGCGCGAAGTAAAGGCCCAGCGCAGCCCGCGCTCCTGCCCTGGGGCCTCGTCTT
TCTCCAGGAAAACGTGGACCGCTCTCCGCCGACAGTCTCTTCCACAGACCCCTGTCGCCTTCGCCCCCCCGG
TCTCTTCCGGTTCTGTCTTTTCGCTGGCTCGATACGAACAAGGAAGTCGCCCCCAGCGAGCCCCGGCTCCCC
CAGGCAGAGGCGGCCCCGGGGGCGGAGTCAACGGCGGAGGCACGCCCTCTGTGAAAGGGCGGGGCATGC
AAATTCGAAATGAAAGCCCGGGAACGCCGAAGAAGCACGGGTGTAAGATTTCCCTTTTCAAAGGCGGGAGAA
TAAGAAATCAGCCCGAGAGTGTAAGGGCGTCAATAGCGCTGTGGACGAGACAGAGGGAATGGGGCAAGGA
GCGAGGCTGGGGCTCTCACCGCGACTTGAATGTGGATGAGAGTGGGACGGTGACGGCGGGCGCGAAGGC
GAGCGCATCGCTTCTCGGCCTTTTGGCTAAGATCAAGTGTAGTATCTGTTCTTATCAGTTTAATATCTGATACG
TCCTCTATCCGAGGACAATATATTAAATGGATTGATCAATCCGCTTCAGCCTCCCGAGTAGCTGGGACTACAG
ACGGTGCCATCACGCCCAGCTCATTGTTGATTCCCGCCCCCTTGGTAGAGACGGGATTCCGCTATATTGCCT
GGGCTGGTGTCGAACTCATAGAACAAAGGATCCTCCCTCCTGGGCCTGGGCGTGGGCTCGCAAAACGCTGG
GATTCCCGGATTACAGGCGGGCGCACCACACCAGGAGCAAACACTTCCGGTTTTAAAAATTCAGTTTGTGAT
TGGCTGTCATTCAGTATTATGCTAATTAAGCATGCCCGGTTTTAAACCTCTTAAAACAACTTTTAAAATTACCTT
TCCACCTAAAACGTTAAAATTTGTCAAGTGATAATATTCGACAAGCTGTTATTGCCAAACTATTTTCCTATTTGT
TTCCTAATGGCATCGGAACTAGCGAAAGTTTCTCGCCATCAGTTAAAAGTTTGCGGCAGATGTAGACCTAGCA
GAGGTGTGCGAGGAGGCCGTTAAGACTATACTTTCAGGGATCATTTCTATAGTGTGTTACTAGAGAAGTTTCT
CTGAACGTGTAGAGCACCGAAAACCACGAGGAAGAGAGGTAGCGTTTTCATCGGGTTACCTAAGTGCAGTGT
CCCCCCTGGCGCGCAATTGGGAACCCCACACGCGGTGTAGAAATATATTTTAAGGGCGCG

(1250 characters)

在人类基因这本“天书”中有多少个字母? 你个人的基因密码存放在家中要占多少地方？要用什么车拉回家?

$3*10^{9}$ （3G)　1 页　3,000 字母

1 本书　250 页

4,000 本书

1本书1公斤，4000公斤

五隔的一米长书架16个

贴墙放要一个16平方米的房间

一家3口要一个套间?

当然我们不会那么笨！！！！！

DNA测序在当时不是一件容易的事

- 1985年首次提出要完整测出人类基因组的核苷酸序列.1990年人类基因组计划(HGP)正式由美国NIH和DOE启动,计划用15年、30亿美元。

- HGP得到的序列长约为29.1亿个硷基对， 是由在9个月中产生的2727万条DNA片段（每条平均长543个硷基对）拼装而成的，总长度是全序列的5.11倍。

- 发表在Science (vol. 291, 16, Feb. 2001)的文章 “The sequence of the human genome” 由274人署名， 除了Celera公司以外， 由13所大学和实验室共同完成。

《Science》2001

DNA sequencing（测序） refers to sequencing methods for determining the order of A, G, C and T—in a DNA.

- History:
- RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage (噬菌体) MS2, identified and published by Walter Fiers at the University of Ghent (Ghent, Belgium), between 1972[2] and 1976.[3]
- Prior to the development of rapid DNA sequencing methods in the early 1970s by Frederick Sanger at the University of Cambridge, in England and Walter Gilbert and Allan Maxam at Harvard,[4][5] a number of laborious methods were used. For instance, in 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.[6]
- The chain-termination method developed by Sanger and coworkers in 1975 soon became the method of choice, owing to its relative ease and reliability.[7][8]

Large-scale sequencing strategies

- Current methods can directly sequence only relatively short (300–1000 nucleotides long) DNA fragments in a single reaction.

- Large-scale sequencing aims at sequencing very long DNA pieces, such as whole chromosomes. Common approaches consist of cutting large DNA fragments into shorter DNA fragments. The fragmented DNA is cloned into a DNA vector, and amplified in *Escherichia coli* （大肠肝菌）. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence. This method does not require any pre-existing information about the sequence of the DNA and is referred to as ***de novo*** **(从无到有，从头开始）** sequencing. The different strategies have different tradeoffs in speed and accuracy; ***shotgun methods*** are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

New sequencing methods

- High-throughput sequencing, 2007-2008
- In vitro （试管内）clonal （无性(繁殖)系的）amplification, 2003-2005
- Parallelized sequencing, 2005
- Sequencing by ligation（连配性），2005
- Sequencing by hybridization（杂交）， 2005
- Mass spectrometry, 2005
- ....

In October 2006, the X Prize Foundation established an initiative to promote the development of ***full genome sequencing technologies***, called the Archon X Prize, intending to award $10 million to “the first Team” that can build a device and use it

- to sequence 100 human genomes within 10 days or less;
- with an accuracy of no more than one error in every 100,000 bases sequenced;
- with sequences accurately covering at least 98% of the genome;
- and at a recurring cost (续生成本) of no more than $10,000 (US) per genome.

这再次说明生物信息学是一门”State-of-the-Art”科学, 就连DNA测序(Sequencing) 这一常用的技术都不例外.

在我们这次课中不讲述测序方法，有很多教科书讲述这一技术.

基因功能组研究-大规模基因功能表达谱的分析

- 在长为29.1亿个硷基对的中，识别出‘3万’个基因

- 即使我们已经获得了人的完整基因图谱，那我们对人的生命活动能说明到什么程度呢？

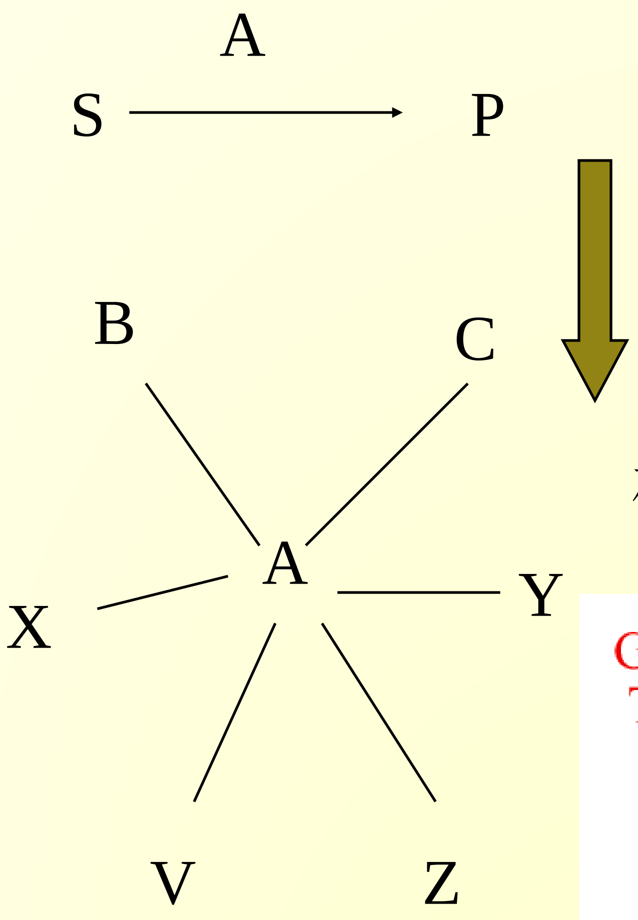
  例如：

  一个基因表达的产物是否同某一功能有关(*reductionism*)；

  多个基因之间的相关性说明什么问题(*systematist*)；

- 概括这些问题，其实质应该是：知道了DNA序列和基因，我们依然不知道它们是如何发挥功能的，或者说它们是如何按照特定的时间、空间进行基因表达的。

后基因组时代对生物信息学研究方向的影响 ----转向‘生物信息学 + 系统生物学’



序列 → 结构 → 功能

相互作用 → 网络 → 功能

Genome-centric Theory of Life ⇨ Network-control Theory of Life

Genome is the headquarters.

Genome is just the warehouse of parts used by the network.

Bio-molecular networks (生物分子网络)

By bio-molecular networks, we mean ：

Gene regulatory networks **(GRN, 基因调控网、转录调控网)**

Protein-protein interaction networks **(PPI, 蛋白交互网)**

metabolic networks **(新陈代谢网)**

Pathway network **(信号通路网)**

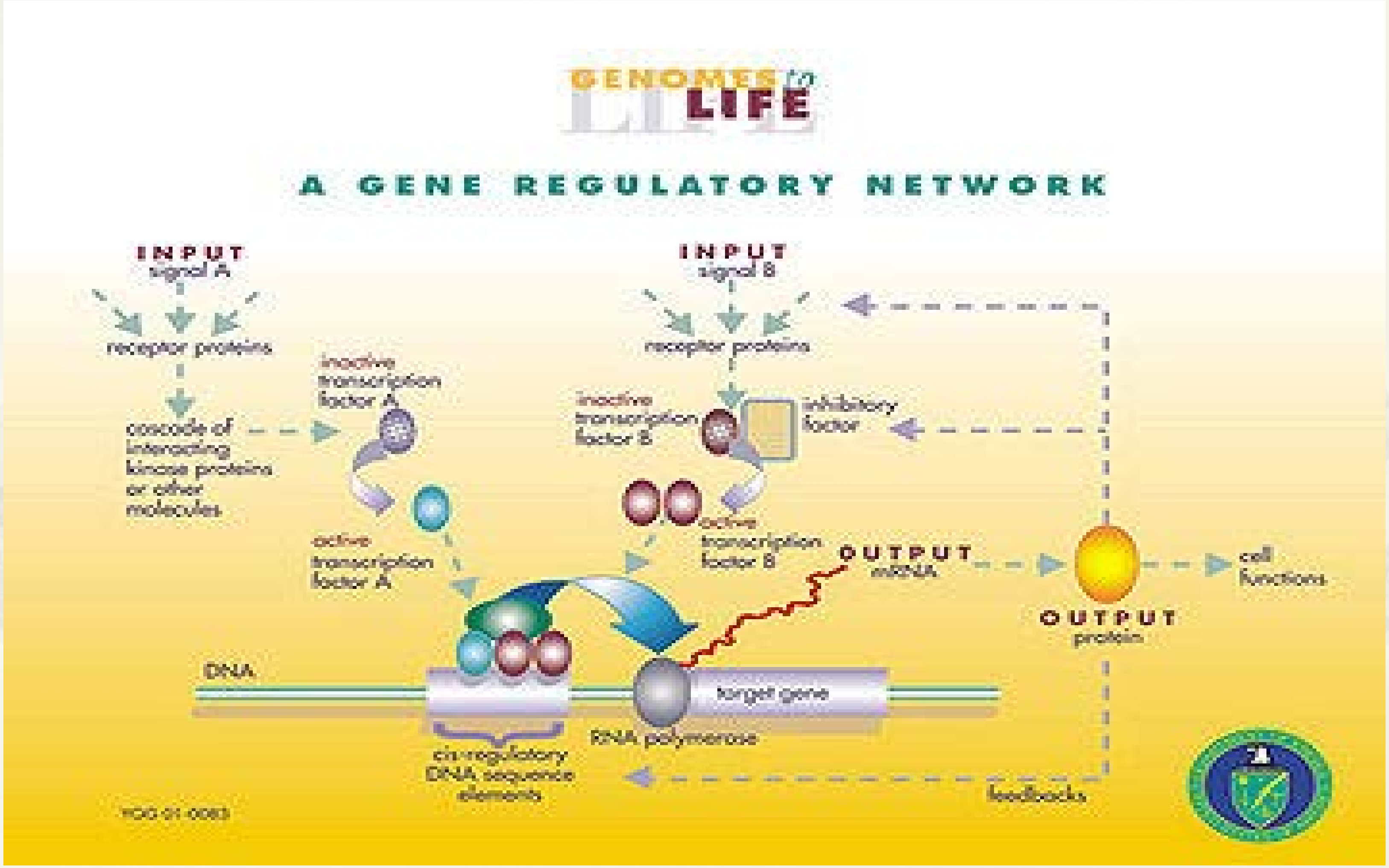
GRN, 基因调控网、转录调控网

- A gene **regulatory network** or **genetic regulatory network (GRN)** is a collection of DNA segments in a cell which interact with each other indirectly (through their RNA and protein expression products) and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA.
- In general, each mRNA molecule goes on to make a specific protein (or set of proteins). In some cases this protein will be structural, and will accumulate at the cell membrane or within the cell to give it particular structural properties. In other cases the protein will be an enzyme, i.e., a micro-machine that catalyses a certain reaction, such as the breakdown of a food source or toxin. Some proteins though serve only to activate other genes, and these are the transcription factors that are the main players in regulatory networks or cascades. By binding to the promoter region at the start of other genes they turn them on, initiating the production of another protein, and so on. Some transcription factors are inhibitory.

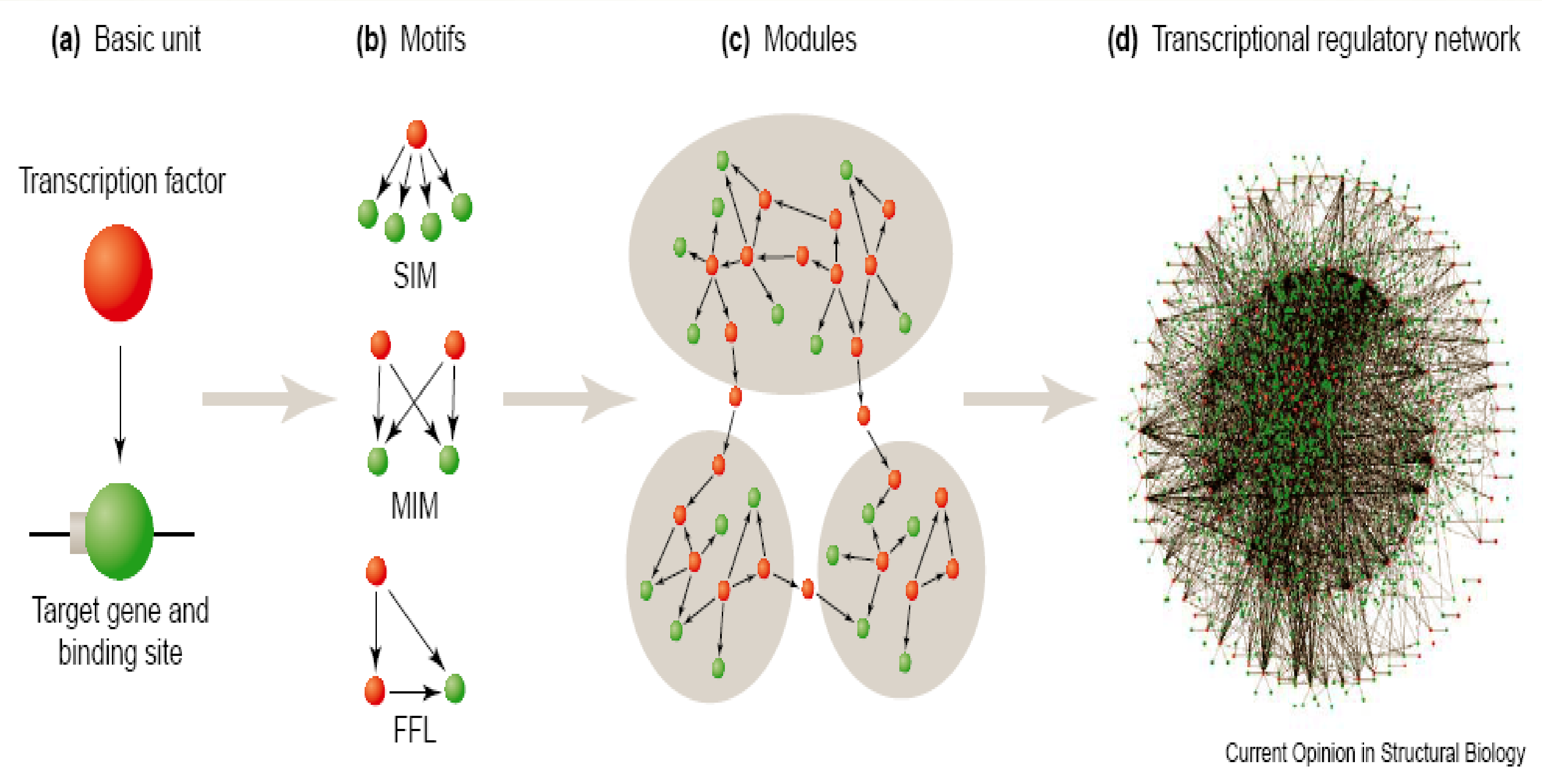
GRN, 基因调控网、转录调控网

- 基因调控网是细胞中一组基因，它们通过产生的RNA和蛋白表达产品以及细胞中的其他成分相互作用，以控制各基因转录为mRNA的速率。
- 一般来讲，一个mRNA产生一个蛋白。有的蛋白是细胞的构成材料，产生特定的生物功能；有的蛋白是酶，催化生化反应；更有一类蛋白专为激发一些基因的转录过程，通过绑定在基因的启动子(promoter)区域上，让该基因生产一个新的蛋白。
- 这些蛋白或RNA聚合酶称为Transcription factor (TF, 转录因子)，是基因调控网的主要节点。有些TF起抑制作用。

GRN


GENOMES to LIFE
A GENE REGULATORY NETWORK
INPUT signal A
receptor proteins
cascade of interacting kinase proteins or other molecules
inactive transcription factor A
active transcription factor A
INPUT signal B
receptor proteins
inactive transcription factor B
inhibitory factor
active transcription factor B
OUTPUT mRNA
OUTPUT protein
cell functions
DNA
target gene
RNA polymerase
cis-regulatory DNA sequence elements
feedbacks
YGG-01-0083

GRN 基因调控网络


(a) Basic unit
(b) Motifs
(c) Modules
(d) Transcriptional regulatory network
Transcription factor
Target gene and binding site
SIM
MIM
FFL

Current Opinion in Structural Biology

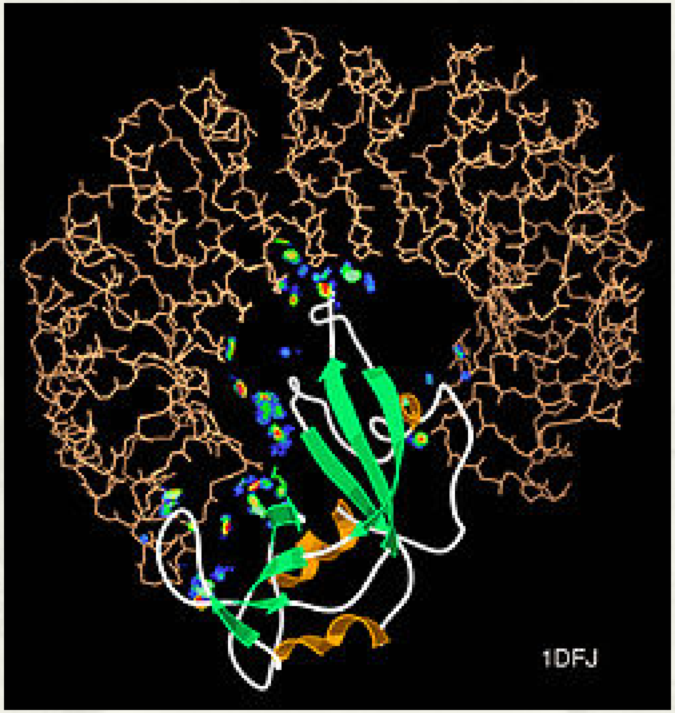
Protein-protein interaction (PPI) network 蛋白交互网

- **Protein–protein interactions (PPI)** occur when two or more proteins bind together, often to carry out their biological function.
- Many of the most important molecular processes in the cell such as DNA replication (复制) are carried out by large molecular machines that are built from a large number of protein components organised by their protein–protein interactions.
- Protein interactions have been studied from the perspectives of biochemistry, quantum chemistry, molecular dynamics, chemical biology, signal transduction and other metabolic or genetic/epigenetic networks.
- Indeed, protein–protein interactions are at the core of the entire interactomics (交互组学) system of any living cell.

蛋白交互网

- 两个蛋白绑定在一起实现某种生物功能。
- 许多重要功能，例如DNA的复制，是由一个由许多蛋白绑定在一起的复合体来完成的。
- 总多的蛋白绑定形成PPI网，是交互组学的核心。

The horseshoe shaped ribonuclease (核糖核酸酶) inhibitor (shown as wireframe) forms a protein–protein interaction with the ribonuclease protein. The contacts between the two proteins are shown as coloured patches



- PPI are important for the majority of biological functions.
- For example, signals from the exterior of a cell are mediated to the inside of that cell by protein–protein interactions of the signaling molecules. This process, called signal transduction (**信号传导**), plays a fundamental role in many biological processes and in many diseases (e.g. cancers).
- Proteins might interact for a long time to form part of a protein complex (**蛋白质复合体**), a protein may be carrying another protein (for example, from cytoplasm (**细胞质**) to nucleus (**细胞核**)or vice versa in the case of the nuclear pore importins), or a protein may interact briefly with another protein just to modify it (for example, a protein kinase (蛋白激酶) will add a phosphate (磷酸盐) to a target protein).
- In conclusion, protein–protein interactions are of central importance for virtually every process in a living cell. Information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches

Metabolic network 新陈代谢网

- A **metabolic network** is the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell. As such, these networks comprise the chemical reactions of metabolism as well as the regulatory interactions that guide these reactions.

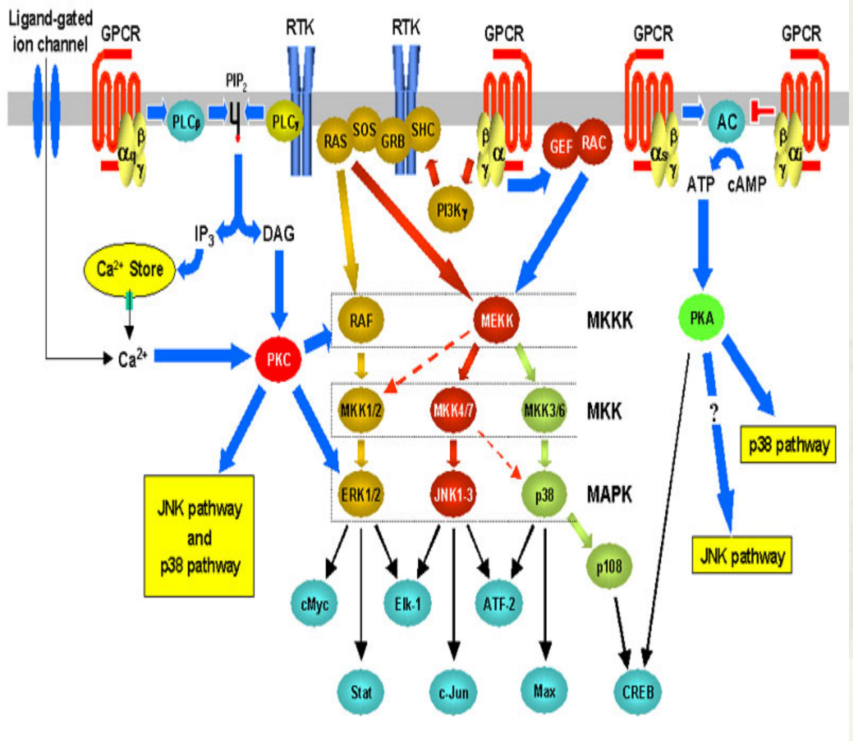
- With the sequencing of complete genomes, it is now possible to reconstruct the network of biochemical reactions in many organisms, from bacteria to human. Several of these networks are available online: Kyoto Encyclopedia of Genes and Genomes (KEGG)[1], EcoCyc [2], BioCyc [3] and metaTIGER [4]. Metabolic networks are powerful tools, for studying and modelling metabolism.

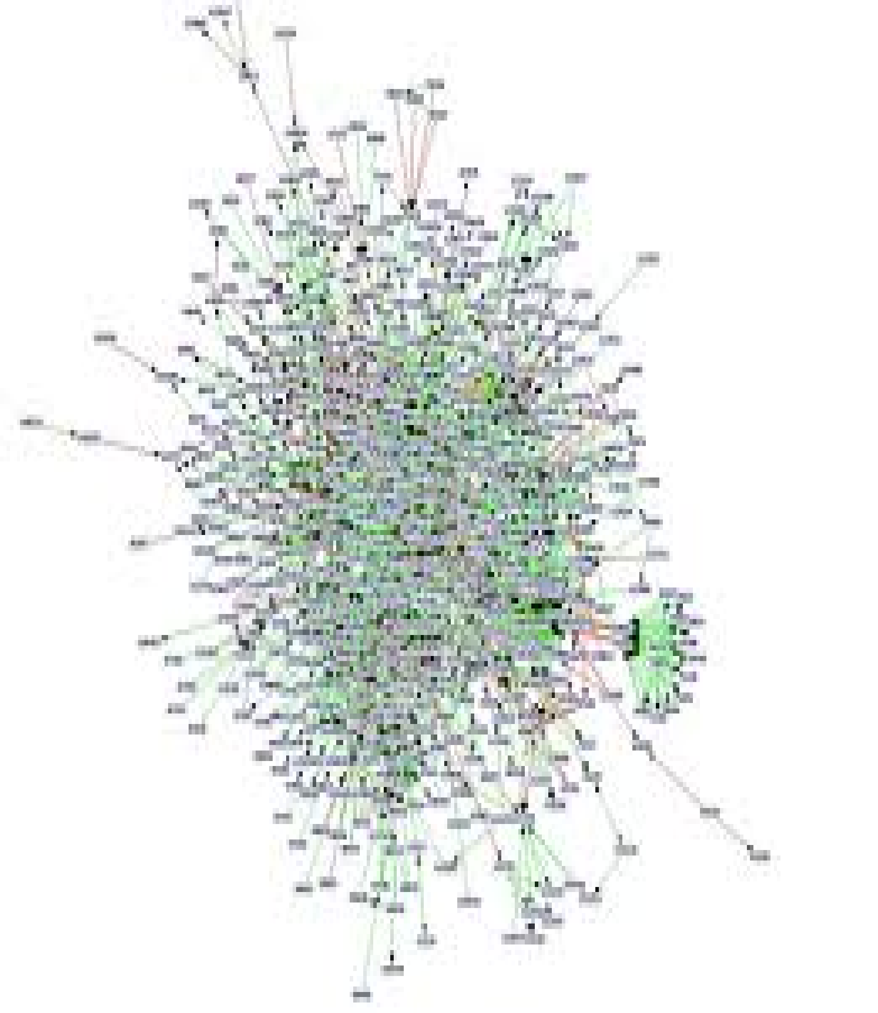
Signal transduction network 信号传导网络

- **Cell signaling** governs basic cellular activities and coordinates cell actions.
- The ability of cells to perceive and correctly respond to their microenvironment is the basis of development, tissue repair, and immunity as well as normal tissue homeostasis.
- Errors in cellular information processing are responsible for diseases such as cancer, autoimmunity (自身免疫性), and diabetes (糖尿病). By understanding cell signaling, diseases may be treated effectively and, theoretically, artificial tissues may be created.[citation needed]

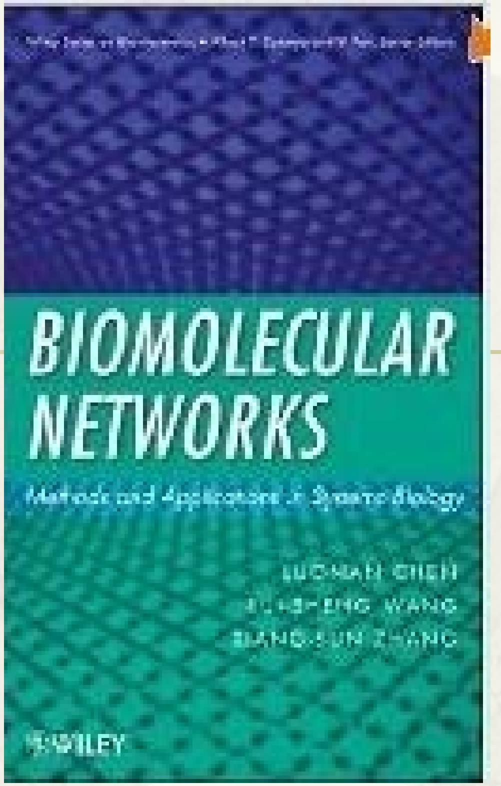
- Traditional work in biology has focused on studying individual parts of cell signaling pathways (信号通路). Systems biology research helps us to understand the underlying structure of cell **signaling networks** and how changes in these networks may affect the transmission and flow of information.
- Such networks are complex systems in their organization and may exhibit a number of emergent properties including bistability and ultrasensitivity.

信号传导网的例子


Ligand-gated ion channel
GPCR
RTK
RTK
GPCR
GPCR
GPCR
PIP₂
PLCβ
PLCγ
RAS
SOS
GRB
SHC
GEF
RAC
AC
αq
β
γ
α
αs
αi
PI3Kγ
IP₃
DAG
Ca²⁺ Store
ATP
cAMP
RAF
MEKK
MKKK
PKA
Ca²⁺
PKC
MKK1/2
MKK4/7
MKK3/6
MKK
?
p38 pathway
JNK pathway and p38 pathway
ERK1/2
JNK1-3
p38
MAPK
p108
JNK pathway
cMyc
Elk-1
ATF-2
Stat
c-Jun
Max
CREB



Book about Biomolecular networks



***Luonan Chen, Rui-Sheng Wang, Xiang-Sun Zhang*.**
***Biomolecular Networks: Methods and Applications in Systems Biology*.**
John Wiley & Sons, Hoboken, New Jersey. July, 2009.

（三）
一些基本问题

生物信息学研究的一些基本问题

- 分子生物学范畴内的研究的例子:

1) 单核苷酸(A-T-G-C)多态性的研究

2) 蛋白质结构预测

- 系统生物学范畴内的研究的例子

3） 基因调控网络的推断

4） 蛋白质相互作用网的预测

5） 利用分子生物网络研究疾病的机理和药物设计

3.1 单核苷酸 (A-T-G-C) 多态性 （**single-nucleotide polymorphisms ---SNP)** 的研究

3.1　单核苷酸多态性的研究

- DNA 序列的这种差异反映了人类的遗传性疾病 (genetics diseases) 和表象差异(phenotype differences)

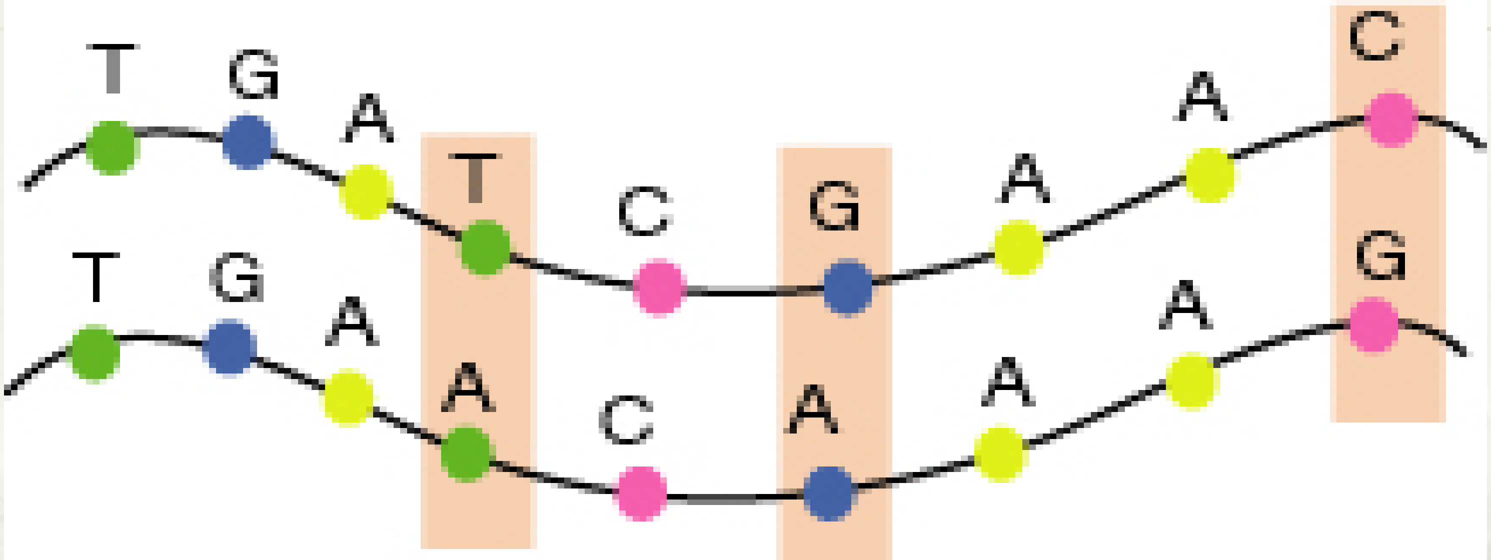
- 所有人的脱氧核糖核酸(DNA，由称为核苷酸 ‘nucleotide’的单体组成) 有 99% 是相同的

- 在一个群体中的 DNA序列的差异称为多态性，在单个核苷酸上的差异称为单核苷酸多态性(SNP：Single Nucleotide Polymorphism)

- 在人类DNA序列中将近每隔1250个核苷酸就有一个SNP(人类的DNA序列约30亿长, 故约有210万个SNP, 是一个很长的序列)，仅有1%的SNP导致蛋白的变异

- 把每个人的210万个SNP连在一起称为 Haplotype/Genotype

基本概念（续）



- Haplotype, 单体型：T， G， C; A， A， G
- Genotype, 基因型， 多倍体：

  （TA）/（AT)， （GA）/（AG)， （CG)/（GC)

单核苷酸多态性的研究（续）

于是, 我们有以下两类问题:

- 某个人的单体型组装(Haplotype Assembly)
  - 由一组SNP片段组装出某人的一对 haplotypes
- 为一个群体推断出一组haplotypes
  - 基于一组测试得到的 genotype, 推断一组haplotypes, 使其能表达测试得到的 genotype组(有不同的推断准则)

单体型组装问题

--- Conflict Graph (CG) and its Bipartization

- **Problem:**

由一组从某人染色体复制、打断后测得的SNP片段组装出该人的一对 haplotypes。

- 如何建立数学模型?

这可以看成一个组合优化问题

单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

		s_1				s_2						s_3	
f_1	G	C	A	T	G								
f_2	G	G	A	T	G	A	A	T					
f_3	G	C	A	T	G	C	A	T	G	C	A	A	
f_4			A	T	G	A	A	T	G	C	A	T	G
f_5										C	A	T	G

DNA sequence fragments

	s_1	s_2	s_3
f_1	0	—	—
f_2	1	0	—
f_3	0	1	0
f_4	—	0	1
f_5	—	—	1

SNP matrix

单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

有两个原因造成冲突:

- 两个片段(two fragments)取自染色体的两条不同的DNA序列
- 两个取自同一条染色体的 片段但带有测定误差（若这一误差不存在，则问题就简化成将上页的矩阵二分成相容的两部分的问题）

单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

定义一个平面图: $G = (V, E)$,

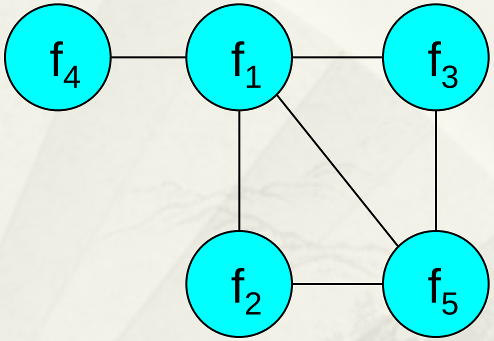
- 顶点集 V 由所有SNP片段组成
- 两个片段若相矛盾（来源不同或测试误差），则有一条边相连接。

单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

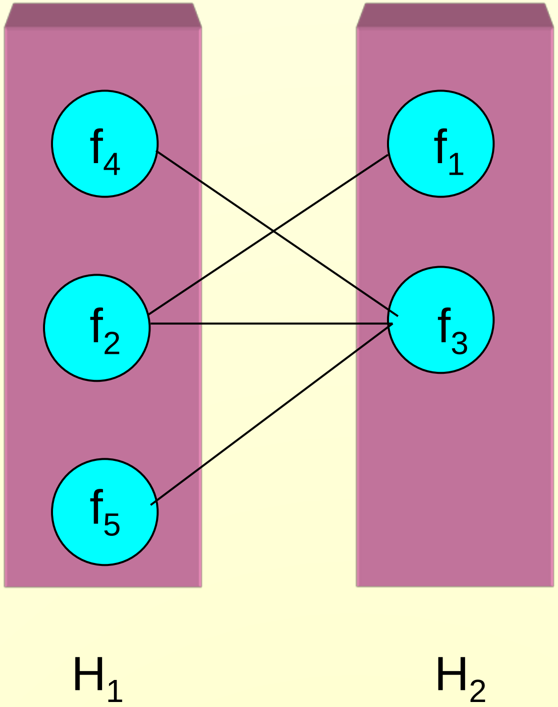
$$
\begin{array}{c|cccccc}
 & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\
\hline
f_1 & 0 & 1 & - & 0 & 0 & 1 \\
f_2 & 1 & 0 & - & - & 1 & - \\
f_3 & - & 0 & 1 & 0 & 1 & 0 \\
f_4 & - & 0 & 1 & - & 1 & 0 \\
f_5 & 1 & - & 0 & 1 & 0 & -
\end{array}
$$

SNP matrix



The conflict graph

当数据没有误差时, 冲突图是一个二分图 (指一个图可以分成两个不相交的子图, 每个子图中的顶点均无边相连)



A case when data is error-free.

单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

- 可以用Mathematica 5.1中的子程序 *BipartiteQ* 来计算判断一个图是否是一个二分图
- 一个图是否是二分图当且仅当它没有奇圈 (*a cycle with odd number of edges*) (S.Skiena,1990)

  （运筹学的研究到此为止）

- 如何从带有误差的数据中恢复出一对 haplotypes

  ⇔

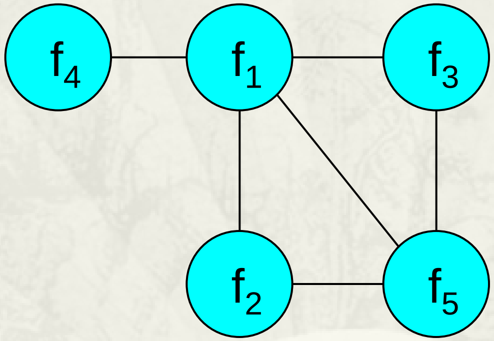
  如何将一个非二分图合理地变成一个近似二分图

  （运筹学的新问题）

单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

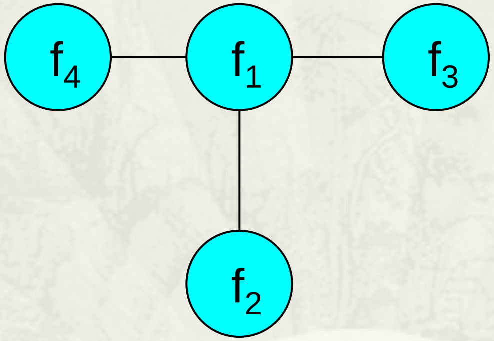
去掉一些顶点来得到二分图 (相当于去除一些受污染的SNP片段)



单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

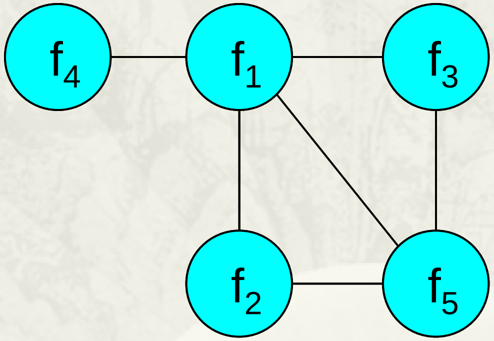
去掉一些顶点来得到二分图 (相当于去除一些受污染的SNP片段)



单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

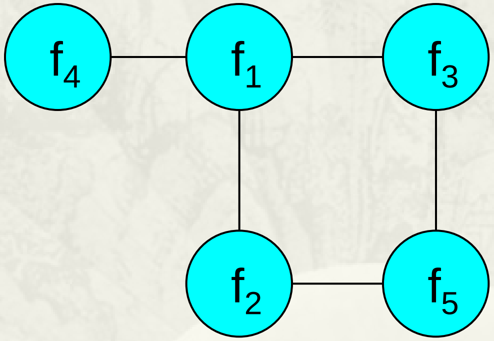
去掉一些边来得到二分图 (相当于去除一些SNP位置或改变某一片段上某一SNP的值)



单体型组装问题（续）

--- Conflict Graph (CG) and its Bipartization

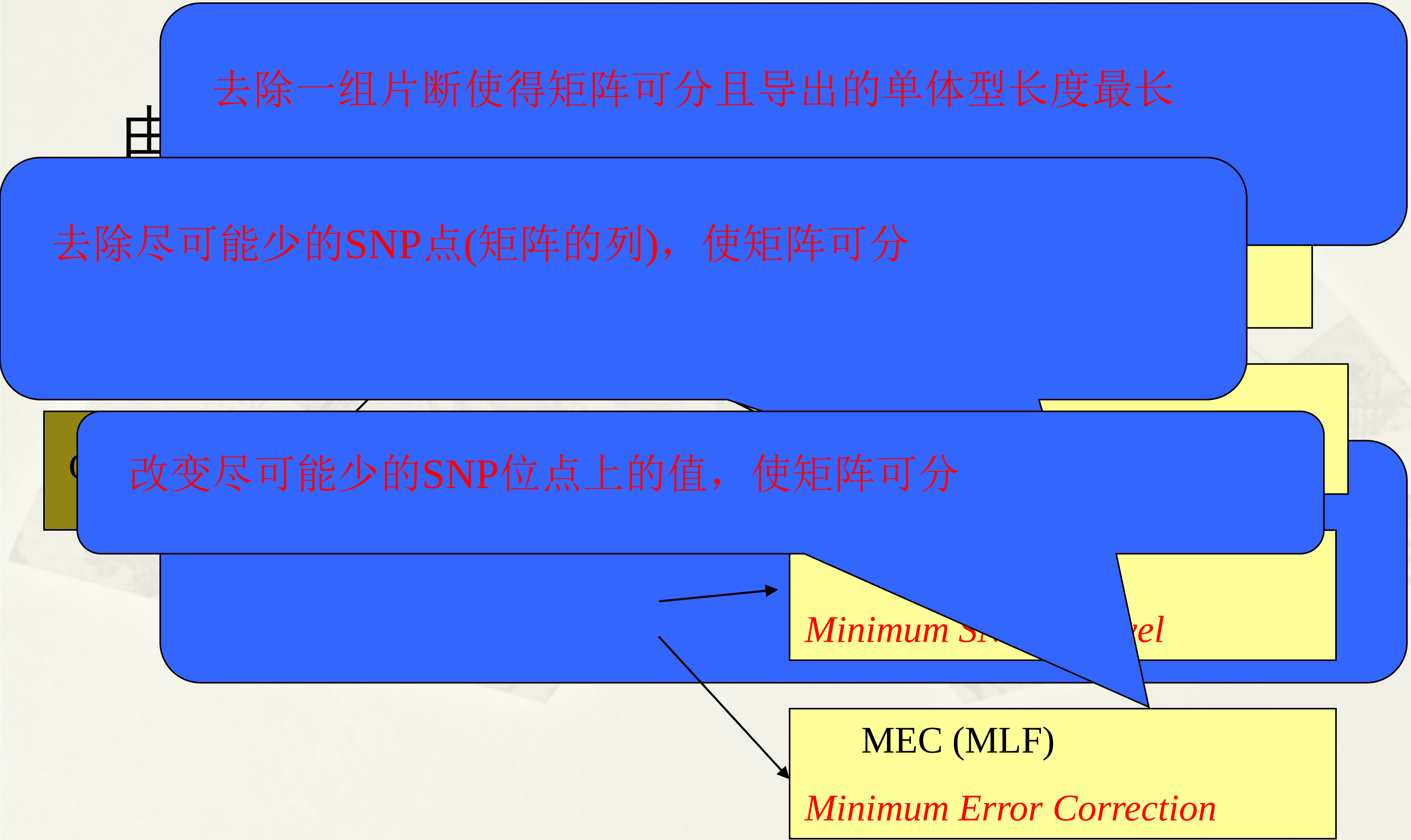
去掉一些边来得到二分图 (相当于去除一些SNP位置或改变某一片段上某一SNP的值)



单核苷酸多态性的研究（续）

- **组装问题(The haplotype assembly problem)**
  - Conflict Graph (CG, 冲突图) and its Bipartization (二分化)
  - 由冲突图和二分化得到的模型
  - 已有算法和复杂性分析
  - 直接依赖冲突图(Conflict Graph)的算法

--- 由冲突图及其二分化导出的模型 ---



去除一组片断使得矩阵可分且导出的单体型长度最长

去除尽可能少的SNP点(矩阵的列)，使矩阵可分

改变尽可能少的SNP位点上的值，使矩阵可分

问题计算复杂性的评估

- 设一个问题的大小，即它的二进制表达的长度，为 L，存在一个求解的算法A，所用的基本运算的总数为 P(L)，P 为一多项式函数，则称这个问题是多项式时间可计算的。

- 一个问题称为是 NP-hard，最简单的理解是对它到目前为止还找不到多项式的算法 （不要误会这是确切的定义）

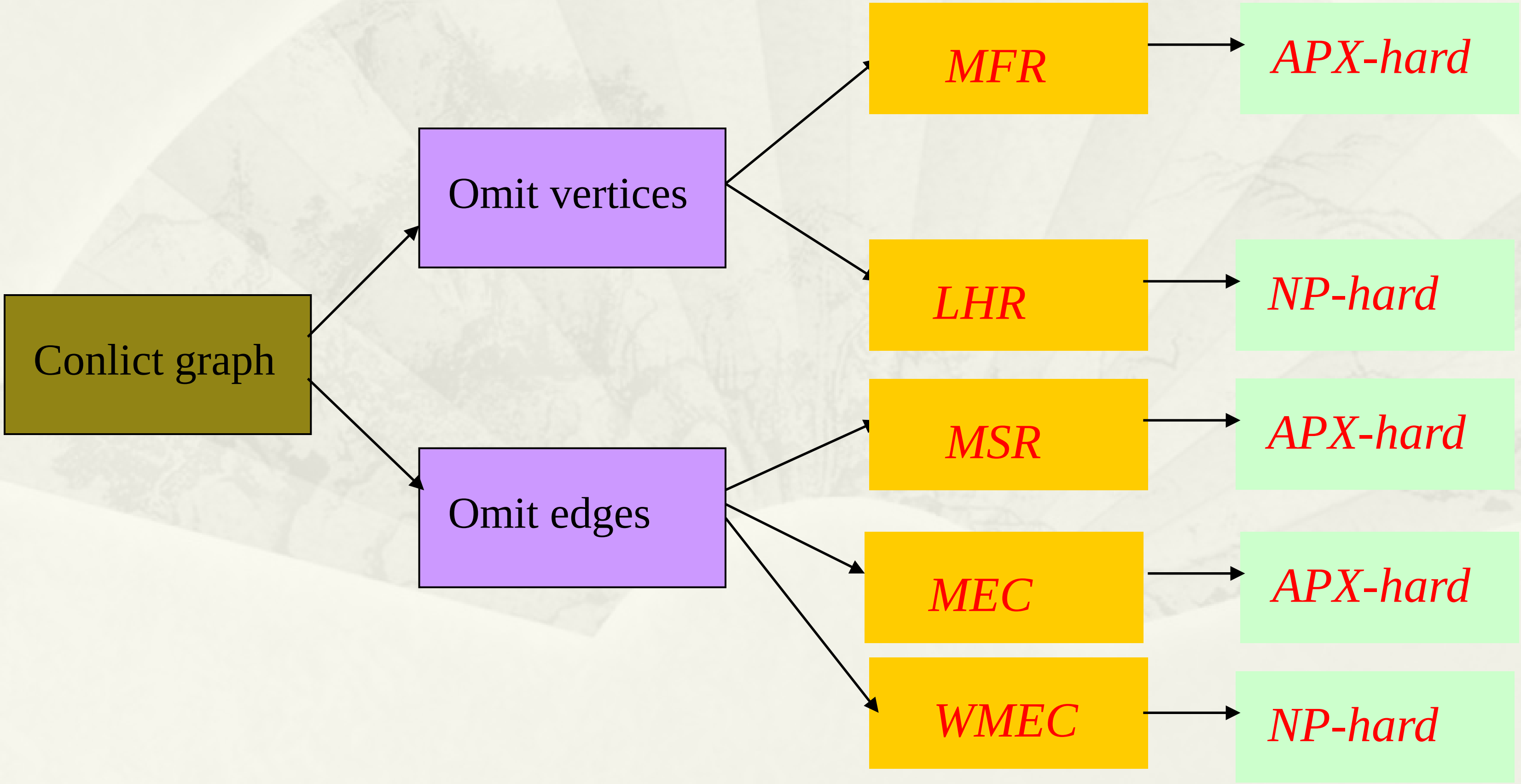
- 一个问题属于 APX (short for approximation algorithm)，是指存在一个多项式时间的近似算法，得到具有预定精度的最优解。

- 一类问题称为 PTAS (polynomial-time algorithm scheme)，是指有多项式算法，对任意精度能得到近似解。

- 一个问题是 APX-hard， 一个不严格的理解是，这是一个 NP-hard 问题，但不属于 PTAS。 因而是个坏消息。

---由冲突图及其二分化导出的模型（续）

计算模型的复杂性:



单体型组装问题--- 已有的算法（续）

- Algorithms for MLF (MEC) (Minimum Letter Flips, Minimum Error Correction)
  - An exact algorithm based on branch-and-bound method (分支定界法)and a heuristic method (启发式方法) based on genetic algorithm (遗传算法) are proposed to solve MEC in Wang R.-S., et al, 2005.
- Algorithms for WMLF (Weighted MLF)
  - A heuristic algorithm based on dynamic clustering (动态聚类)method is presented in Zhao Y.-Y et al, 2005 for WMLF.

直接基于冲突图的算法

以上从冲突图推导出来的模型并没有充分利用冲突图中带有的信息。而且，现有的图的理论和方法经改进后可以直接用来解haplotype assembly problem, 也就是, 图的二分化的问题。

The haplotype assembly problem

--- Algorithms directly based on conflict graph

- Vertex bipartization
  - Reed *et al* and Heuffner provided exact algorithms for the graph bipartization. Their algorithms seek for a minimal set of vertices by deletion to make the conflict graph bipartite. The algorithm given by Heuffner has time complexity $O(3^k \cdot |V||E|)$, where k is the number of vertices to delete.
  - The algorithm is so called fixed-parameter tractable or parameterized complexity

The haplotype assembly problem

--- Algorithms directly based on conflict graph

- Edge bipartization
  - Motivated by the works of Reed et al. and Heuffner, Guo et al. recently considered deleting edges instead of vertices to make the graph bipartite. Their algorithm has time complexity $O(2^k \cdot k|E|^2)$, where k is the number of edges to delete.

一点重要的评论

我们注意到，在 Reed *et al*. 的工作之前，在图论研究中极少有对图的最优二分化的研究，原因是大家知道这是一个NP-难问题。正是由于生物信息学研究的需要，推动图论学家回来研究这一问题并得到好的研究结果。

这一部分的内容依据我们的一篇有关haplotyping problem的综合性文章

Models and algorithms for the haplotyping problem,

Current Bioinformatics,

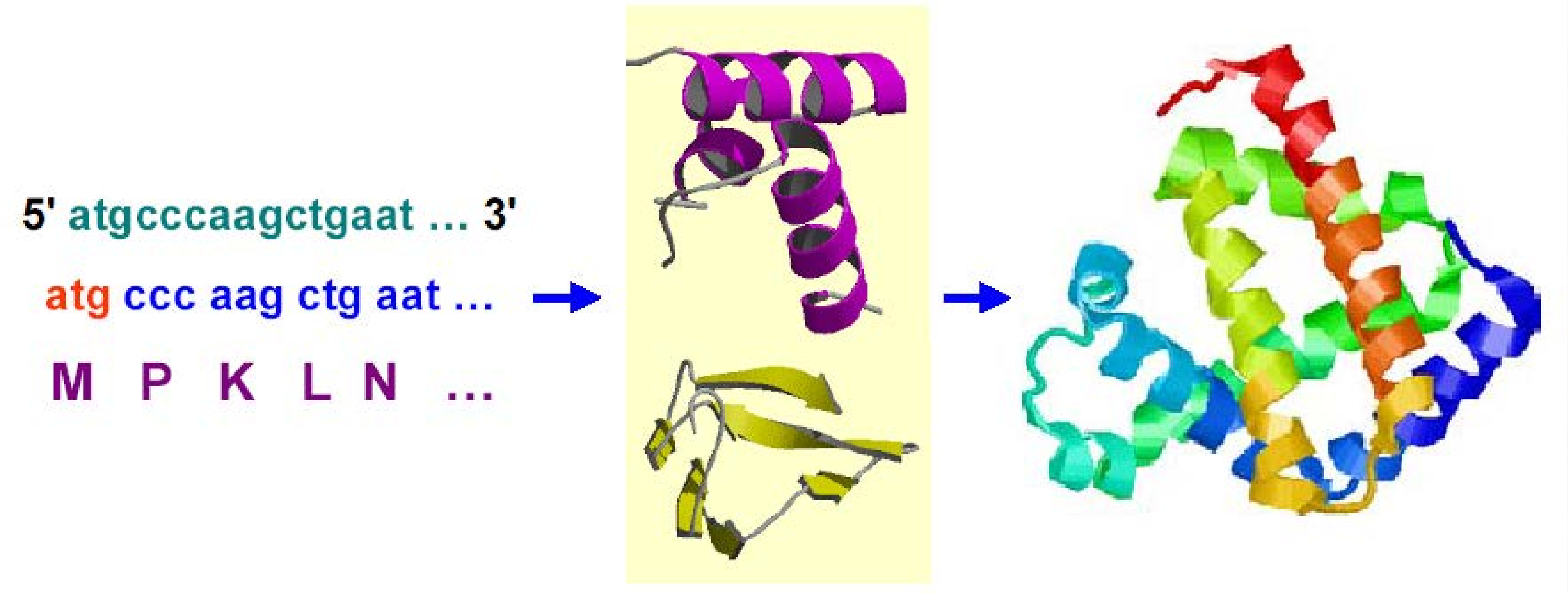
vol.1, no.1, pp.105-114, 2006

3.2　蛋白质结构预测问题

Protein Structure Prediction

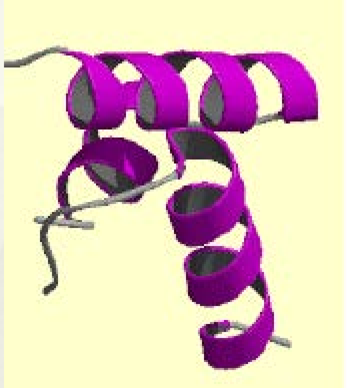
蛋白质结构预测问题Protein Structure Prediction

- Predict protein 3D structure from (amino acid, 20种氨基酸) sequence
- Sequence → secondary structure → 3D structure → function

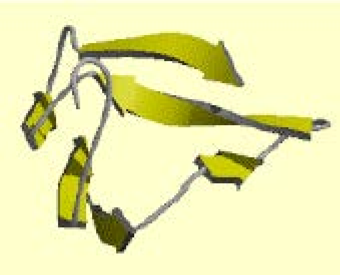


Proteins Secondary Structure （ continued ）

- α-helix (30-35%)
  α-螺旋

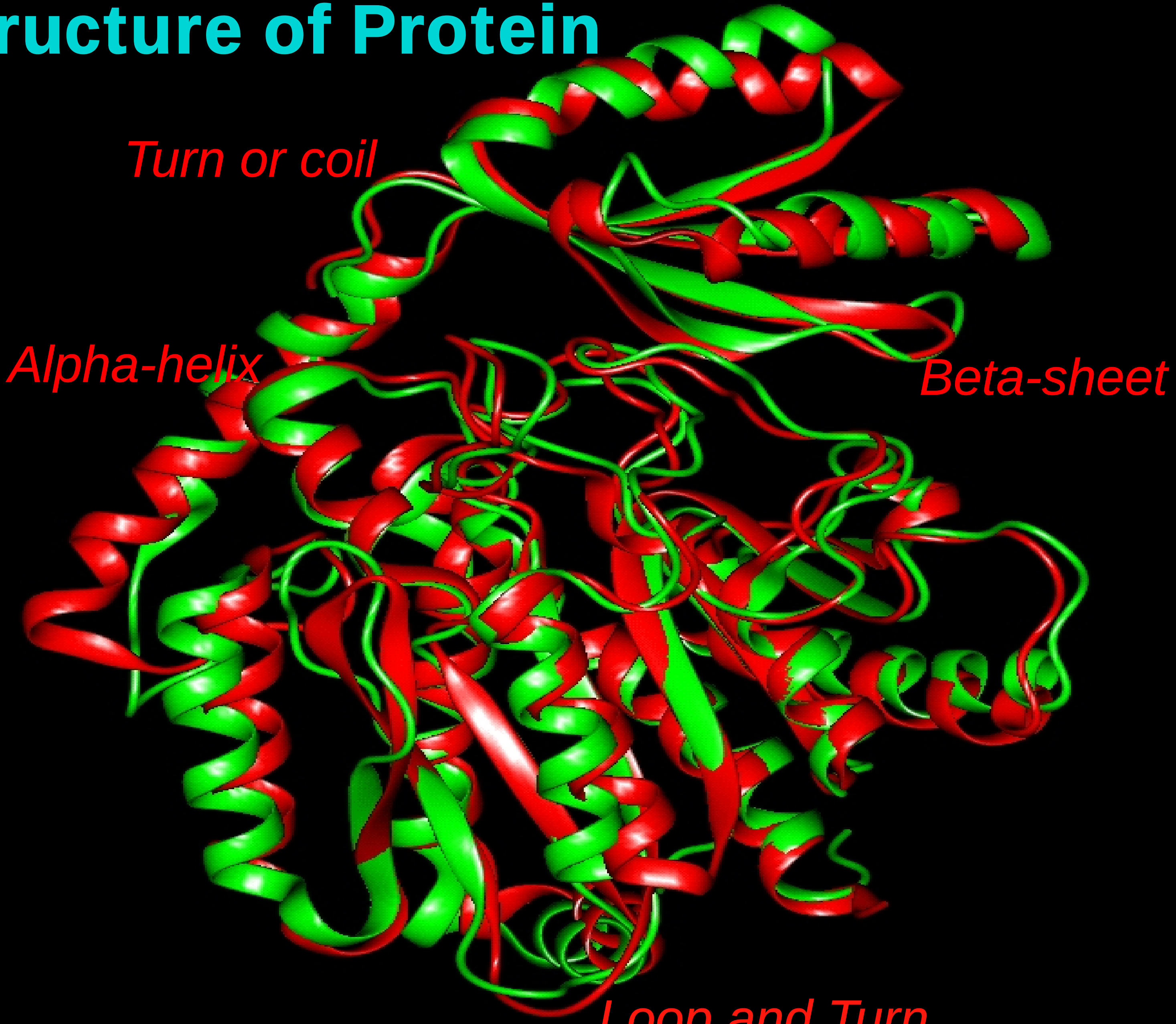


- β-sheet / β-strand (20-25%)
  β-折叠



- Coil (40-50%) 无规则卷曲
- Loop 环
- β-turn， β-转角

3D Structure of Protein

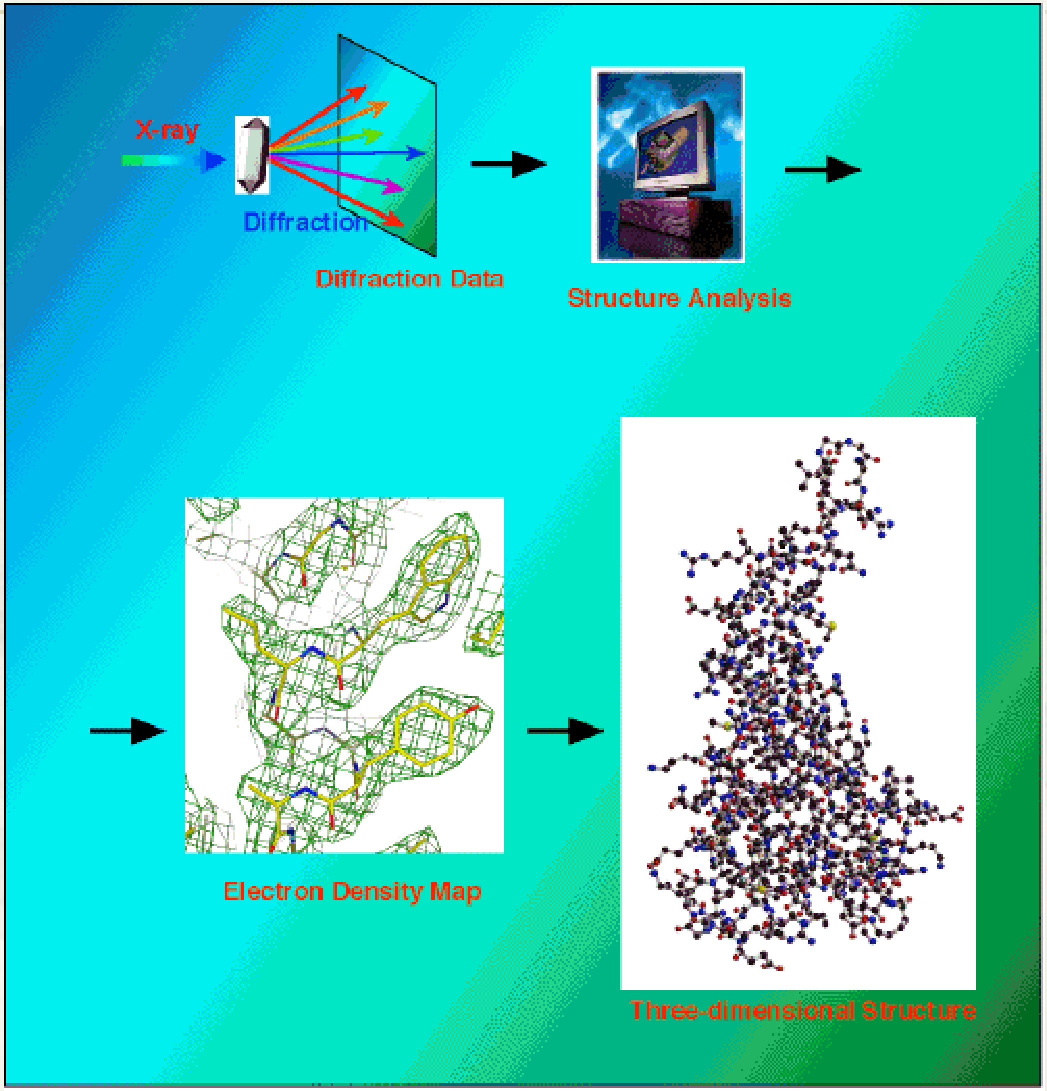


Protein 3D Structure Detection

X-ray diffraction

X-射线衍射法

- Expensive
- Slow



Protein Structure Prediction（ continued ）

- 结构预测的可能性在于：
  - 一级序列的信息唯一地确定了三级结构
  - 一级结构相似性大于50%时，几乎可以推断三级结构的相似性
- 结构预测的必要性
  - DNA sequence data » protein sequence data » structure data

	1994	1997	2002.10
Sequence (Swiss-Port)	40,000	68,000	114,033
Protein Structure (PDB)	4,045	7,000	18,838

Proteins Secondary Structure （ continued ）

——Three Methods of Protein Structure Prediction

- 同源建模法（Homology Modeling)

  Construct 3D model from alignment to protein sequences with known structure (基于一维结构的比对)

- 折叠识别法 Threading (fold recognition)

  Pick best fit to sequences of known 2D / 3D structures (folds) (基于空间结构的比对)

- 从头预测法 (*de novo,* 非比对方法)
  - 分子动力学（ Molecular dynamics ），蒙特卡洛方法
  - 格点方法（Lattice models）
  - 能量函数方法

蛋白质结构预测问题是生物信息学的最难的问题之一

The Critical Assessment of Techniques for Protein Structure Prediction --- **CASP**

蛋白结构预测技术评比

是每年举行一次的蛋白质结构预测的国际性竞赛

CASP: Critical Assessment of Techniques for Protein Structure Prediction

- CASP is a worldwide experiment for protein structure prediction taking place every two years since 1994. CASP provides research groups with an opportunity to objectively test their structure prediction methods and delivers an independent assessment of the *state of the art* in protein structure modeling to the research community and software users. Even though the primary goal of CASP is to help advance the methods of identifying protein three-dimensional structure from its amino acid sequence, many view the experiment more as a **"world championship"** in this field.

- More than 100 research groups from all over the world participate in CASP on the regular basis and it is not uncommon for the entire groups to suspend their other research for months while they focus on getting their servers ready for the experiment and on performing the detailed predictions.

Selection of target proteins

- In order to ensure that no predictor can have prior information about a protein's structure, it is important that the experiment is conducted in a double-blind fashion: Neither predictors nor the organizers and assessors know the structures of the target proteins at the time when predictions are made. Targets for structure prediction are either structures soon-to-be solved by X-ray crystallography or NMR spectroscopy, or structures that have just been solved (mainly by one of the structural genomics centers) and are kept on hold by the Protein Data Bank.

- If the given sequence is found to be related by common descent to a protein sequence of known structure (called a template), comparative protein modeling may be used to predict the tertiary structure (**三级结构**). Templates can be found using sequence alignment methods such as BLAST or FASTA or protein threading methods. Otherwise, *de novo* protein structure prediction must be applied, which is much less reliable but can sometimes yield models with the correct fold.

在以后各讲中会讨论比对方法, 我们在 这里仅讲一讲‘*de novo*’的概念, 让大家对于这一问题的复杂性有一个感性的认识

Protein Structure Prediction（continued）
——能量函数方法（*de novo*之一）

- 有很多文章作这方面的研究
- 利用20个氨基酸相互之间的引力/排斥力、同水分子之间的引力/排斥力，建立能量函数

$$E_{amber} = \sum_{bonds} K_{r_i}(r_i - r_{i,eq})^2 + \sum_{angles} K_{\vartheta_l}(\vartheta_l - \vartheta_{l,eq})^2$$

$$+ \sum_{dihehedrals} \frac{V_k}{2}[1+\cos(n_k\phi_k - \gamma_k)] + \sum_{i<j}\{\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - 2(\frac{\sigma_{ij}}{r_{ij}})^{6}] + c\frac{q_i q_j}{r_{ij}}\}$$

- 这是一个高阶非线性目标函数，一个由242氨基酸、4102个原子的蛋白质有12000个坐标变量，因而是目前没有办法解的问题

Protein Structure Prediction （ continued ）
——能量函数方法的简化模型

为了看一下问题的复杂性

- 问题的简化：

  1）简化吸引力/排斥力：将20个氨基酸分成疏水/亲水两类

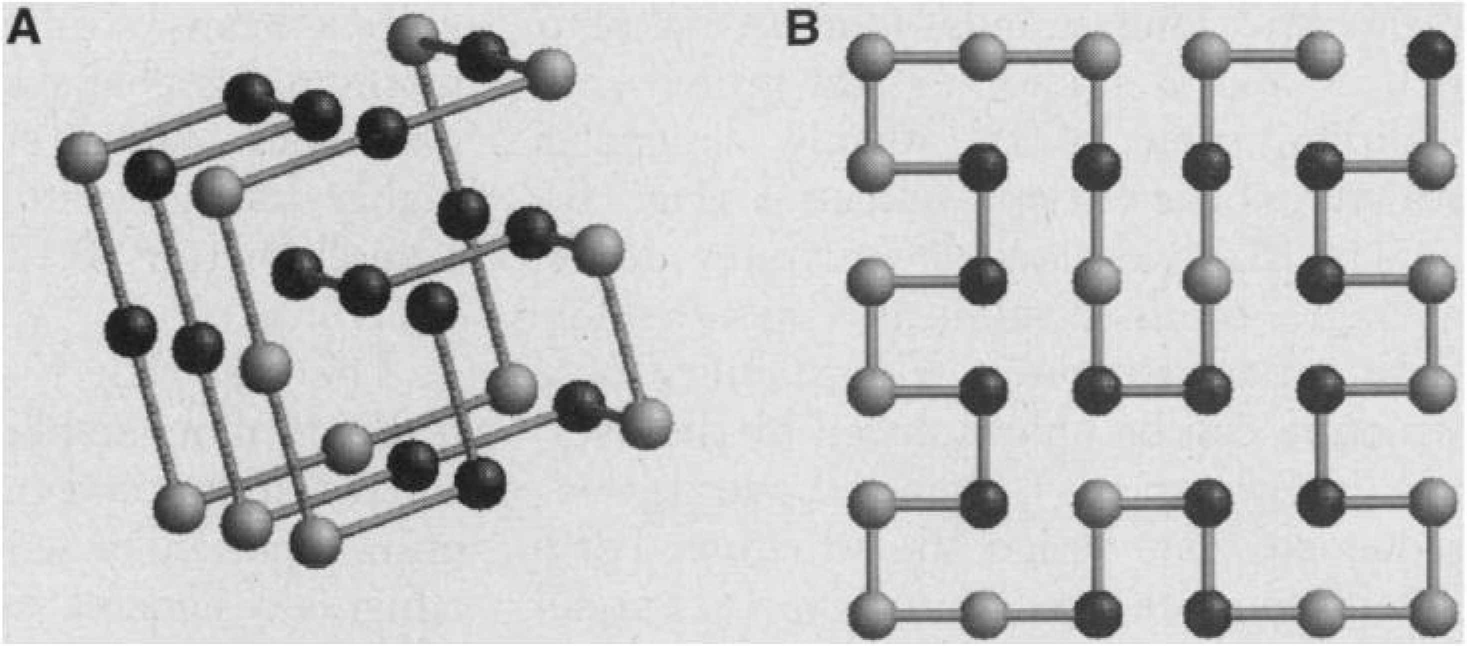
  2）采用格点坐标

  3）在二维空间上进行使能量函数值下降的自动折叠

Lattice Models:

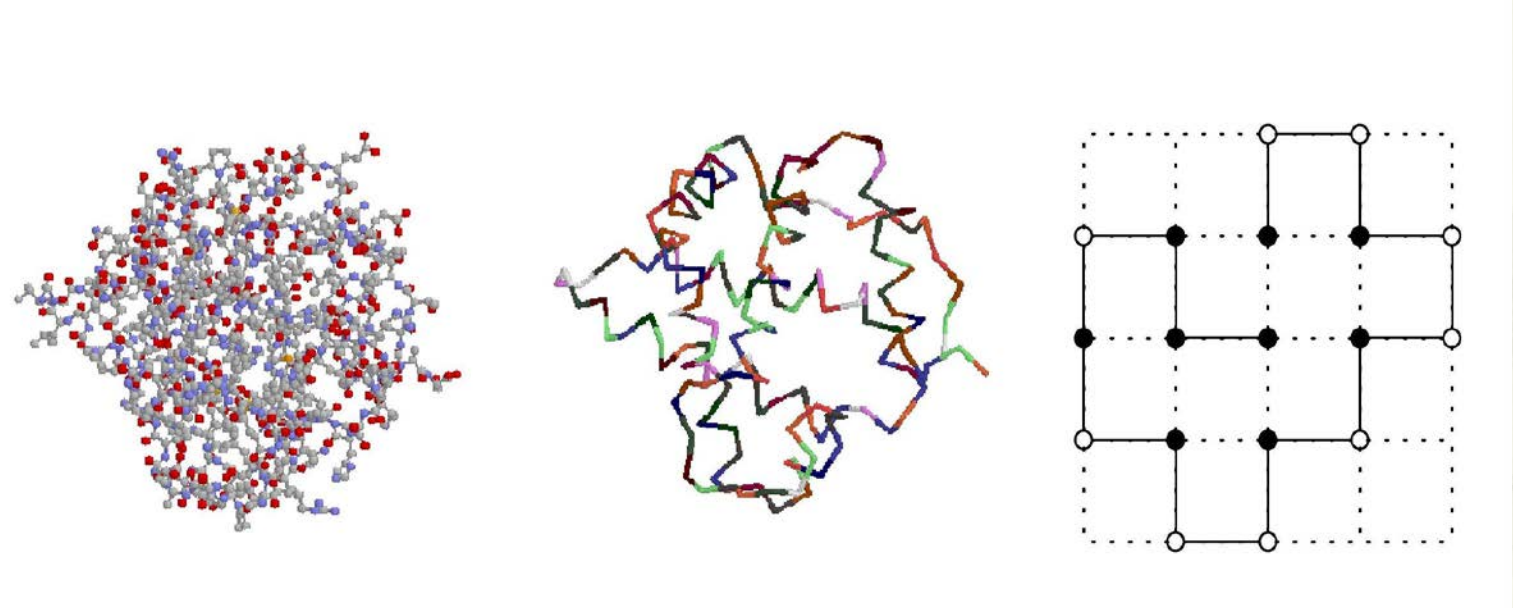
Emergence of Preferred Structures in a Simple Model of Protein Folding

Hao Li, R Helling, C Tang… - *Science, 1996*



Lattice Models

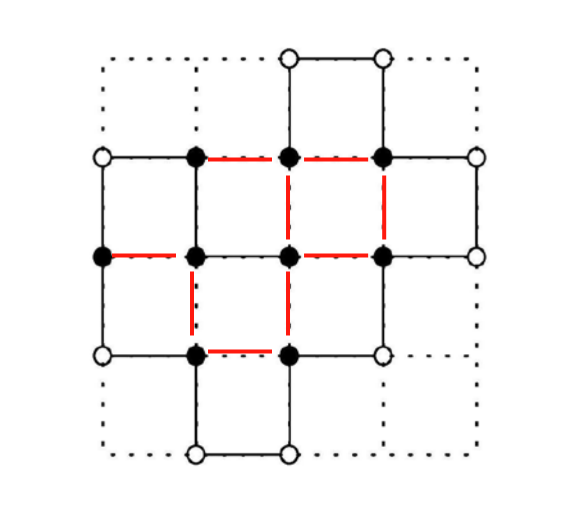
- Suppose that each amino acid occupies one point in a space lattice



- It is called an Exact Model

HP Model (Simple Model)

- Twenty amino acids can be divided into two classes:
  Hydrophobic/Non-polar (H) (疏水)
  Hydrophilic/Polar (P) (亲水)
- The contacts between H points are favorable



● 疏水氨基酸

○ 亲水氨基酸

— 共价键（ovalent bond）

— H-H contact

- Goal: 求得最大的 **H-H contacts**

Mathematical Model (in square lattice)

Let the both of sequence and lattice size be n,let $x_{ij} = 1/0$ for the i-th acid taking the j-th lattice point or not. Let $N(j)$ be the neighboring set of point j.

Let $|N(j)| = 1/2/3$ and the coordinates of point j be Y_j

$$i = H/P \Rightarrow f(i) = 1/0$$

$$\max \quad \sum_{j=1}^{n}[\sum_{i=1}^{n} f(i)x_{ij} \sum_{s\in N(j)} \sum_{i=1}^{n} f(i)x_{is}]$$

$$subject \quad to \quad \sum_{i=1}^{n} x_{ij} = 1, \quad j = 1,\dots,n$$

$$\sum_{j=1}^{n} x_{ij} = 1, \quad i = 1,\dots,n$$

$$\|\sum_{j=1}^{n} x_{ij}Y_j - \sum_{j=1}^{n} x_{(i-1)j}Y_j\| = 1, \quad i = 2,\dots,n$$

Complexity

- NP-hard problem even in the case of two dimensional HP model

  P.Crescenzi, et al.

  On the complexity of protein folding, Journal of Computational Biology, 5(3): 423-, 1998

  - Many local solutions

- GA(遗传算法) MC(蒙特卡洛) SA ()----- time consuming

Main Observation– 转化成组合问题

利用旅行推销员问题的模型和相关算法来解这个二维HP模型

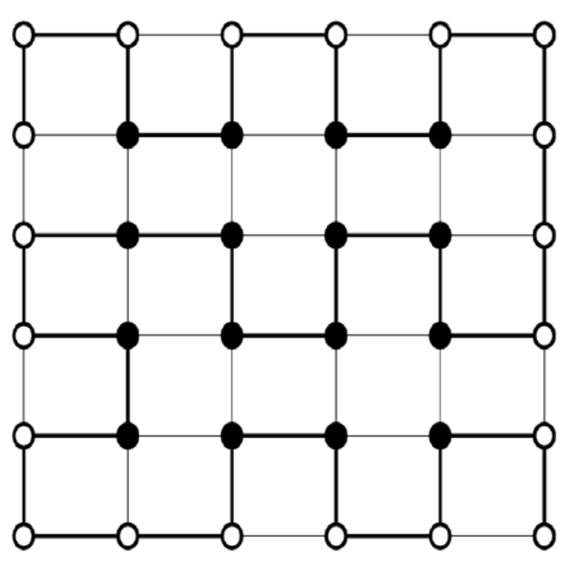
A Traveling Salesman Problem with an energy function concerning the H-H contacts that would be maximized.

自组织映射（ Self-Organization Map — A neural network method ）

- Existing algorithm
  - 利用解 TSP的自组织映射方法
  - 在目标函数中用上 HP Information
  - 在矩形格点上按排一个长度恰好的氨基酸序列

A 36-long sequence

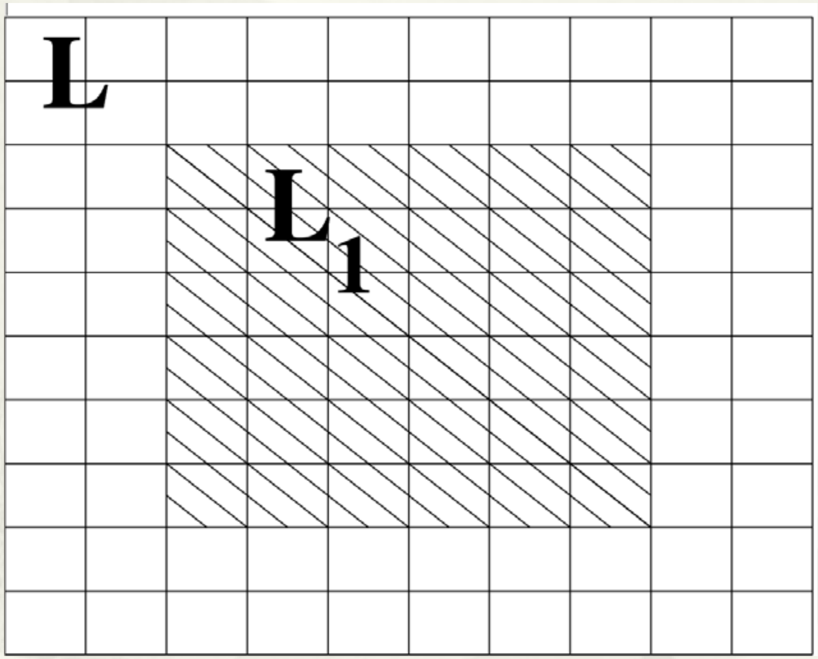
In a 6x6 lattice



An extended SOM Approach

Motivation

- Consider a bigger lattice than the sequence to have more flexible shapes than the only rectangular shape
- Equivalent to a PCTSP (Price Collecting Traveling Salesman Problem): a man travels only a part of the city set with some expectation.



Difficulties caused:

Number of cities > number of neurons

PCTSP

A traveling salesman who gets a prize f_k in every city *k* that he visits and pays a penalty p_l for every city l that he fails to visit, and who travels between cities i and j at cost c_{ij}, wants to minimize the sum of his travel cost and net penalties, while including in his tour enough cities to collect a prescribed amount f_0 of prize money.

The New SOM model is corresponding to the integer programming:

$$\max \quad \sum_{j=1}^{m}[\quad \sum_{i=1}^{n} f(i) x_{ij} \sum_{s\in N(j)} \sum_{i=1}^{n} f(i) x_{is}]$$

$$subject \quad to \qquad \sum_{i=1}^{n} x_{ij} + y_j = 1, \qquad j = 1,\ldots,m$$

$$\sum_{j=1}^{m} x_{ij} = 1, \qquad i = 1,\ldots,n$$

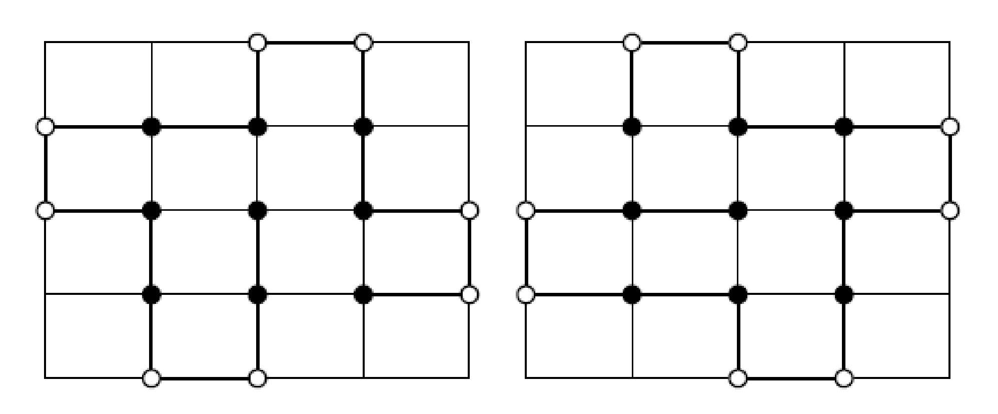
$$\| \sum_{j=1}^{n} x_{ij} Y_j - \sum_{j=1}^{n} x_{(i-1)j} Y_j \| = 1, \qquad i = 2,\ldots,n$$

where m>n and the total variables are (n+1)m.

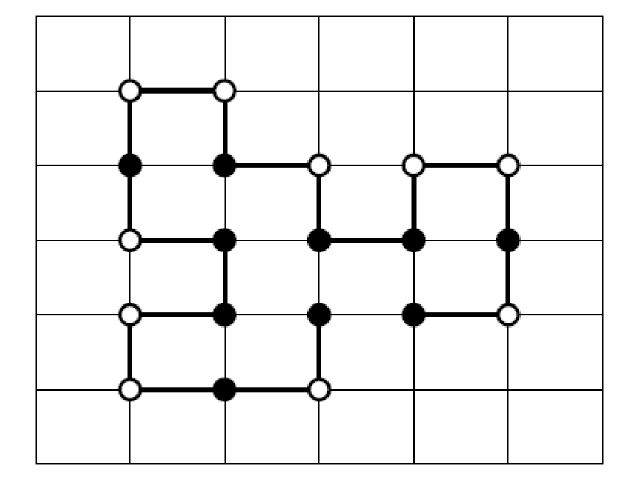
Numerical Results

1. Constructed HP sequences

   (Length of 17)



2. **HP benchmark (up to 36 amino acids)**



Protein Structure Prediction (continued）

就是连这样简单的模型，我们也只能得到36个氨基酸的最优折叠 ！

最小的蛋白约有150个氨基酸 ！

三维上的最优折叠 !?

这一部分的小结（思考问题）

- 听着有意思吗？
- 有

  ………（原因）
- 没有

  ………（理由）
- 不知道有没有

  （那就不要想理由了）

- 本课程的知识结构适合我吗？
- 寻找同我学科背景相近的Bioinformatics课本，对照阅读。为期末在班上作读书报告预作准备

- **试着回忆今天讲的课，能想到几点？‘一只耳朵进，一只耳朵出’，说明消化良好。**
- **可以在 http://zhangroup.aporc.org 上下载本段讲义。**

Thank you!

若要得到我们的研究工作的更多的信息，请使用以下网址

http://zhangroup.aporc.org