Let 1:
Bioinformatics introduction

曹志伟

同济大学

Cause Overview

- 34 课时 （32+2）

- 考核方式：

  课堂 （10%） + 作业 （40%） + 期末 （50%）

曹志伟,张晓艳，唐凯临
zwcao@tongji.edu.cn
助教： 汪源 （小哥哥）
QQ：78934505
微信： czw1013
电话: 65980296
医学楼1015房间

Outline of lecture 1

1. Introduction

2. Bioinformatics History
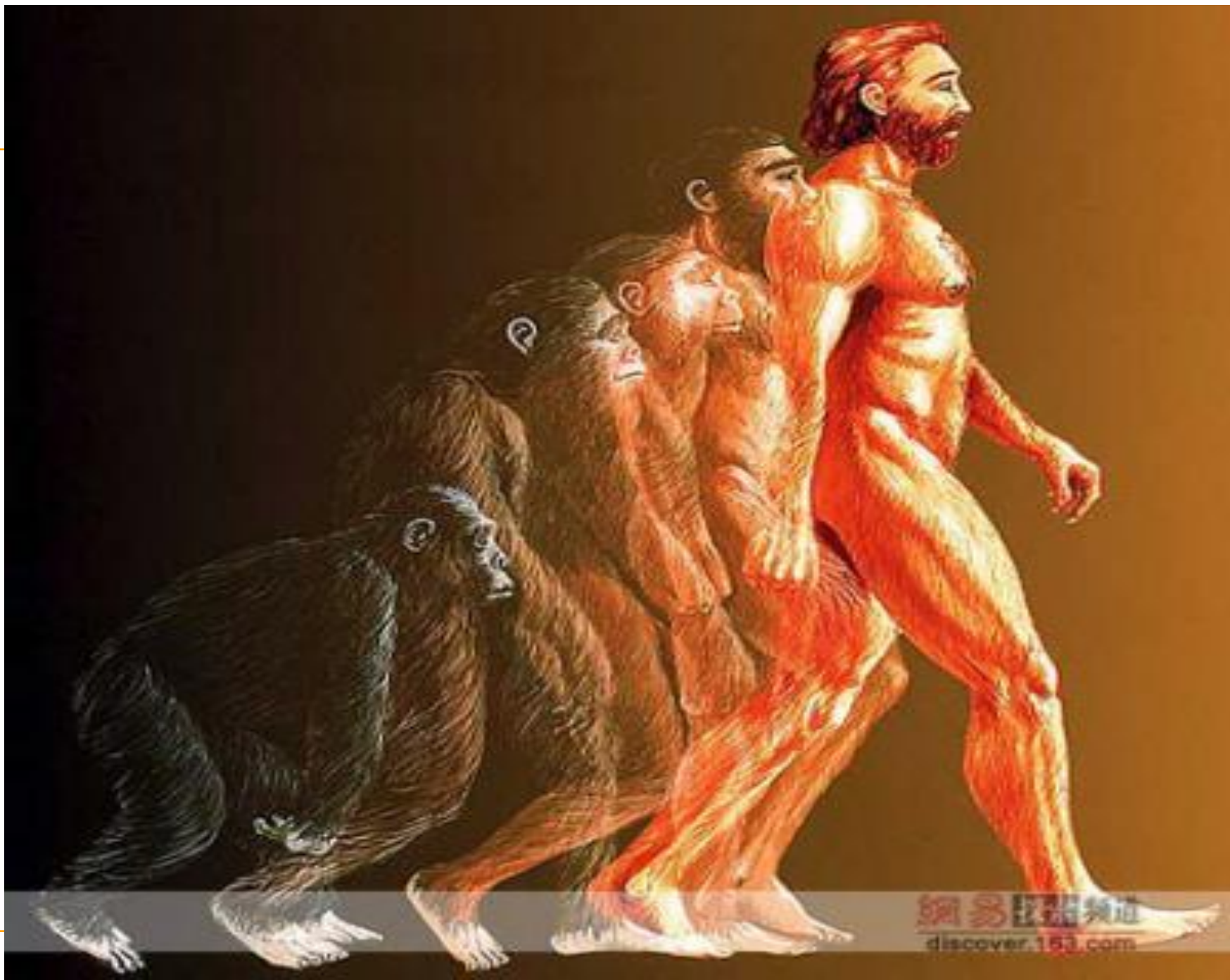
3. Future of bioinformatics

1. Introduction

- Your question?

1. Introduction

- 生命从哪里来…

- 生命是什么…

- 生命往哪里去…?

人类的由来

人，是什么？

─ We are the digital archives of the African Pliocene, even of Devonian seas; walking repositories of wisdom out of the old days. You could spend a lifetime reading in this ancient library and die unsated by the wonder of it.
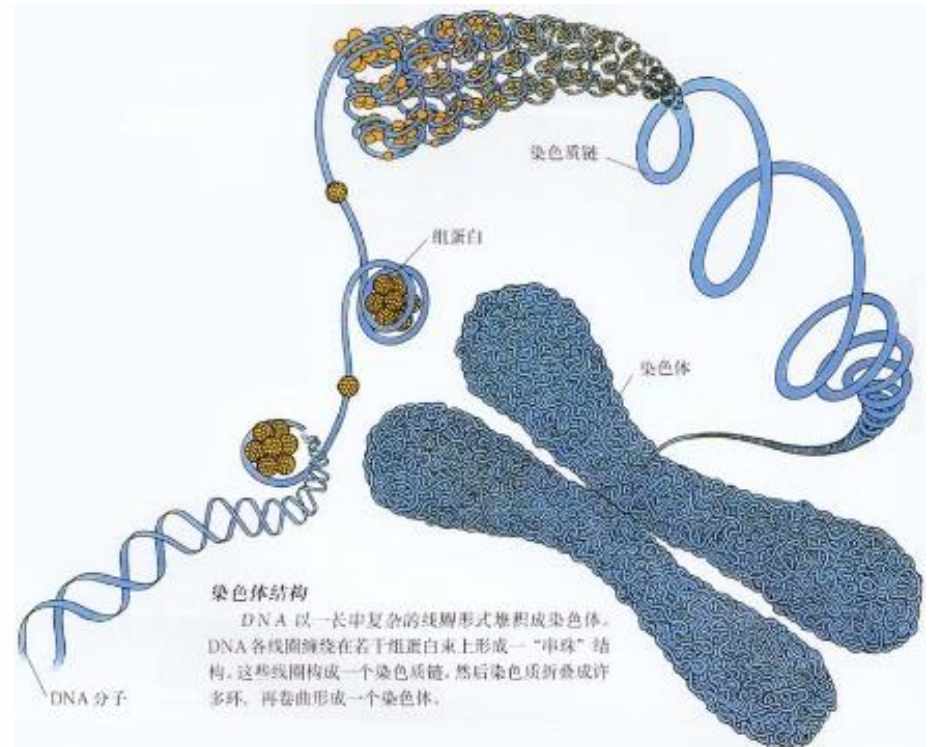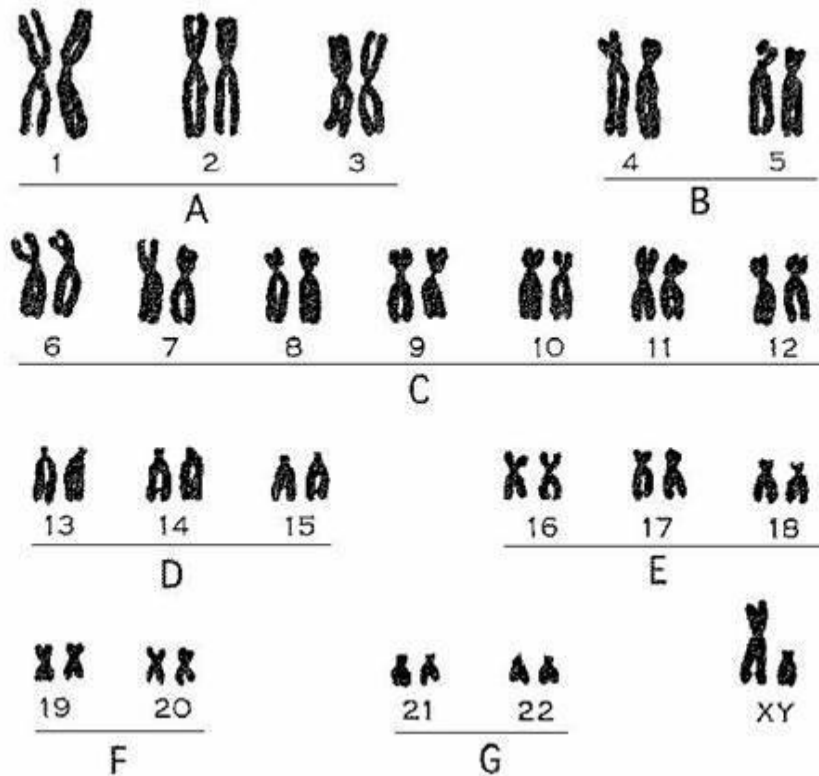

翻译：

人，是什么？

一 We are the digital archives of the African Pliocene, even of Devonian seas; walking repositories of wisdom out of the old days. You could spend a lifetime reading in this ancient library and die unsated by the wonder of it.
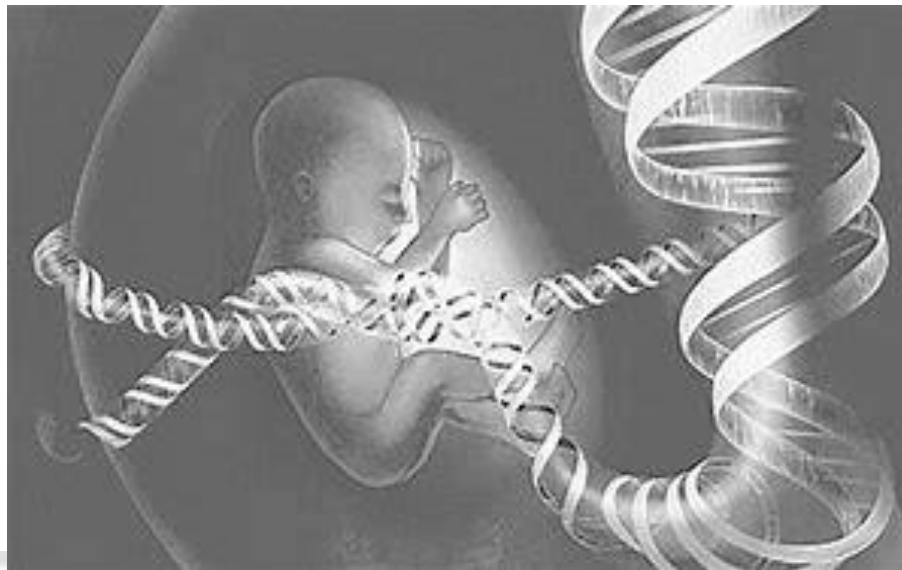
 我们
是非洲上新纪、甚至是泥盆纪的电子文档，
我们
是来自远古时代行走的灵物。

你可以花一辈子的时间
阅读这本古老的书，

却为它的奥妙
死
不
瞑
目。

------ 达尔文（1809-1882）

digital archives

This is Life !

5'CCTCCAGCTTCTCTCTTTGTGTCTAACAAACAAGATAAAAATGAATAAATAAAT
AAATAAGTAAAAGACAAGAGAAAGTAAAAACAAAAAATTGATTCATAGCTAGTC
AAAGAACCATATACATTGAAGACGGTCTCAAGAACTTCATAGCTGAAGGCTCCGT
AATATGATTCAGGTTTATTATTTCCAGCGGGGAAGAATAACTGCAGCAATTATAA
GTACAGGGTCAATAGACTAACCAAGACATCAAGGTTATGTAGAAACTTCTAATAA
ATAAATGTTAAAGTAGAAAACCTCATATTTGCCAGGAGAATTAACAGGCAGGGCG
AGCACAGCTATGGTTAGCTTCTCTTGGTTGTCTTGGCTAACCACGTAAACAGTGCT
TCCTGCAGGAACGCTGACTACTGTTCCACGCTGTACATTATAGGACTCTTTGTTTT

人类基因组测序
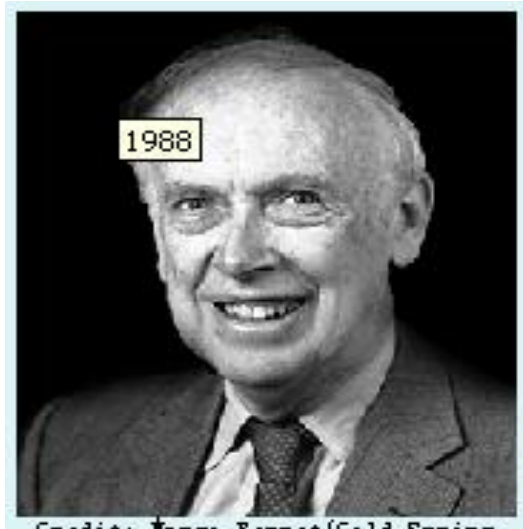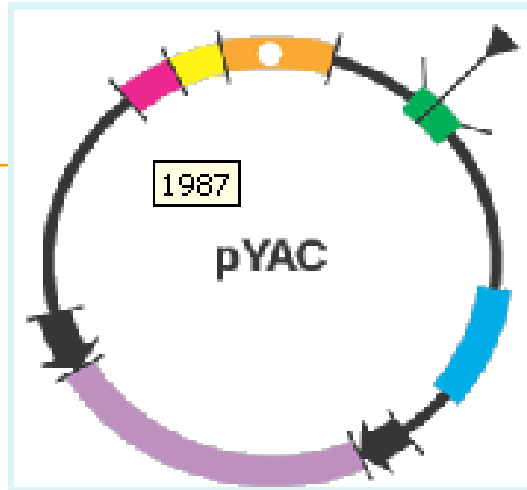
1988年，国际人类基因组织（HUGO）成立 。

1988年，Watson担任负责人，美国国立卫生研究院参与HGP。
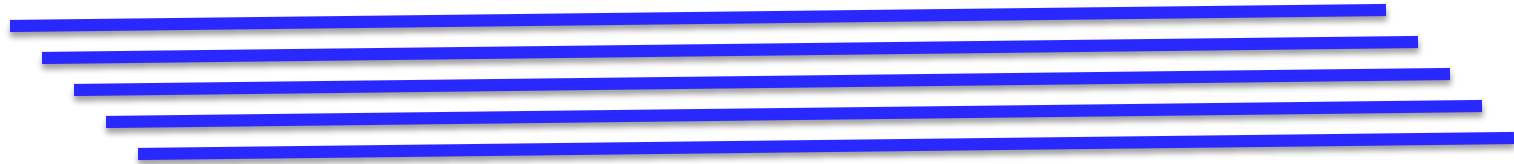
1990年，美国国会批准正式HGP，随后法国、英国、意大利、德国、日本 、中国等陆续启动HGP计划。
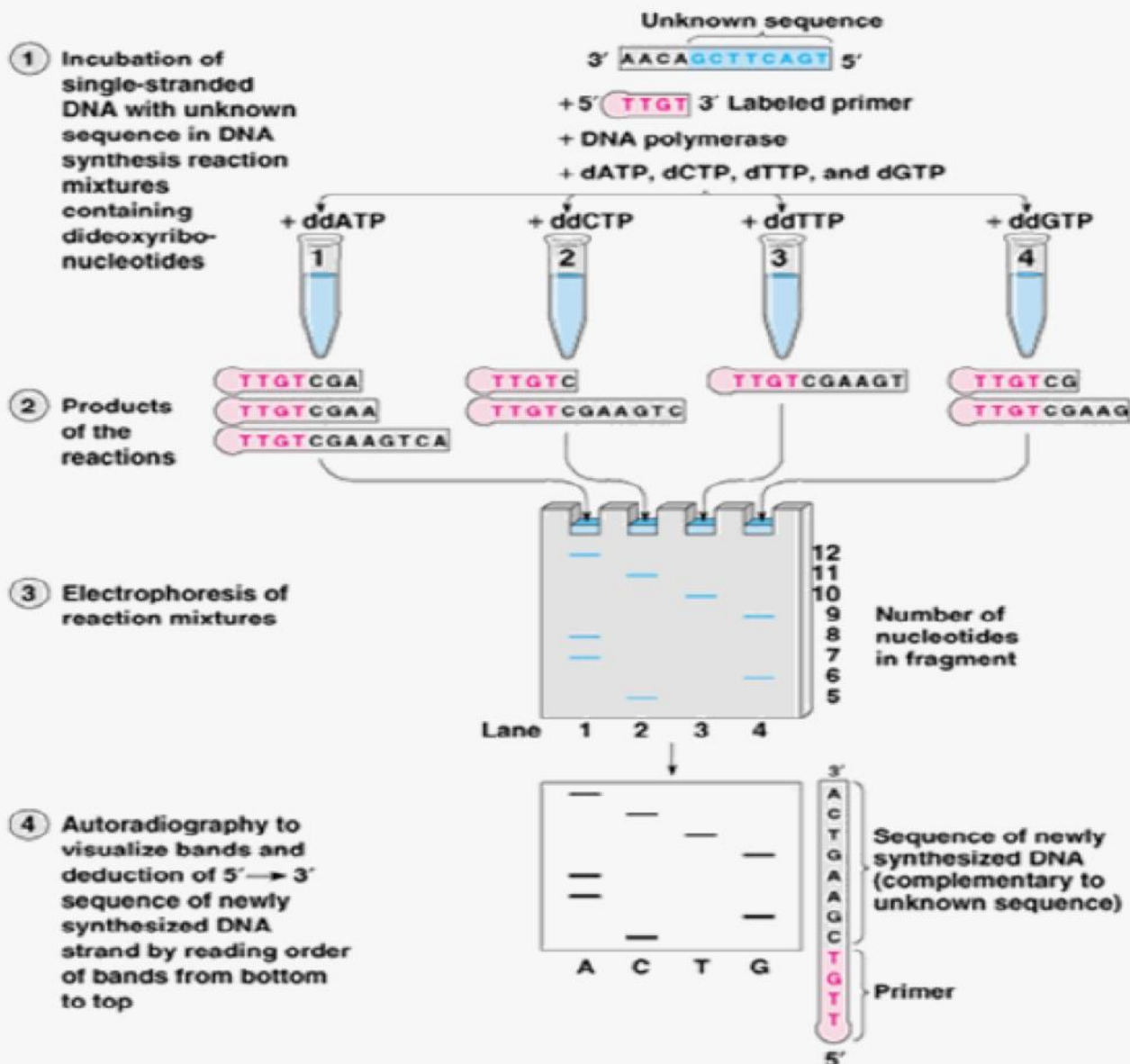
How to sequence??

讨论一下， 举手发言

Dr. Fred Sanger

Frederick Sanger was awarded the prize in both 1958 and 1980. He is the fourth person in the world to have been awarded two Nobel Prizes and the only person to receive both in chemistry.
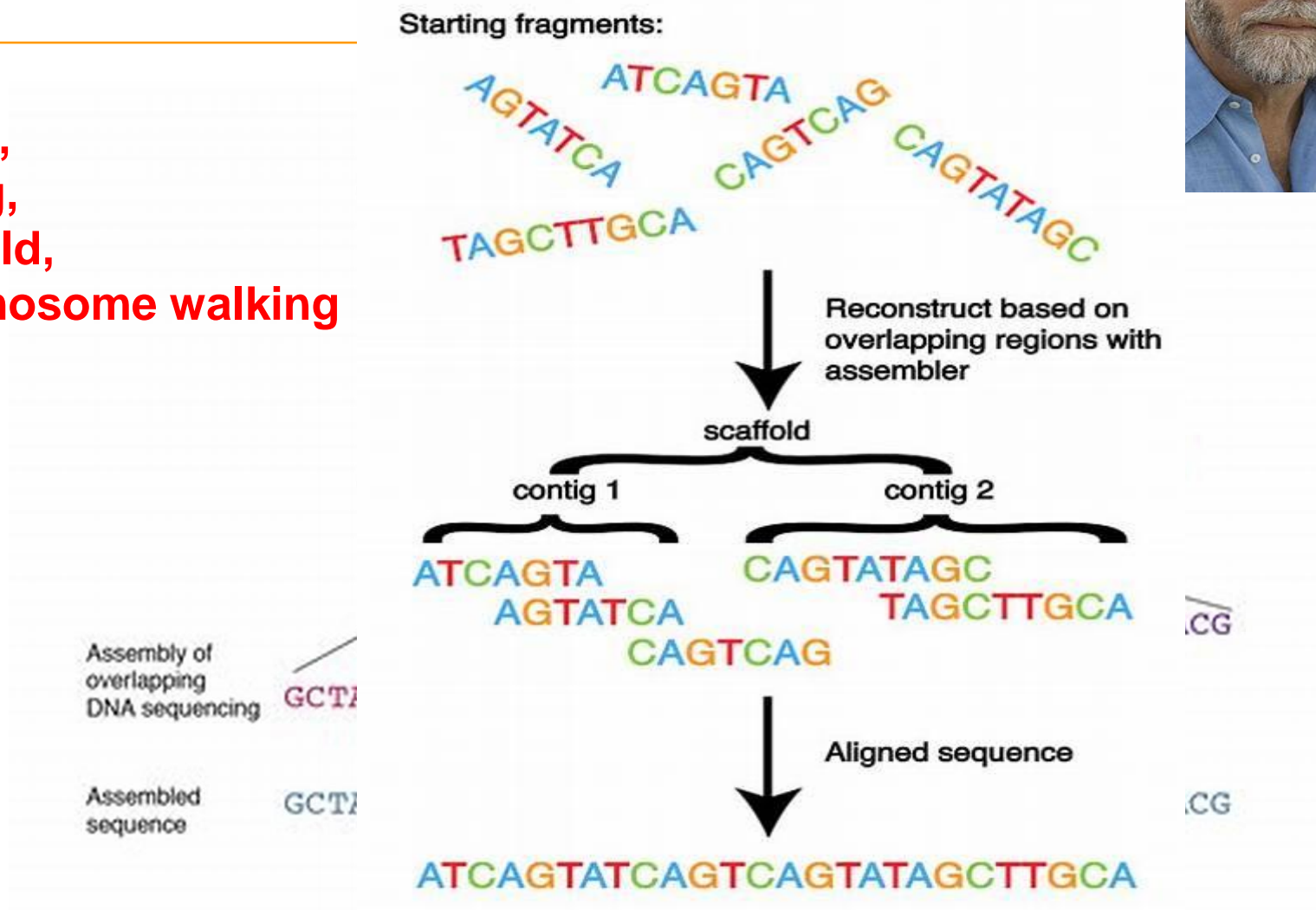
"dideoxy" sequencing technique (Sanger et al., 1977)

DNA双脱氧链终止法测序

1. Incubation of single-stranded DNA with unknown sequence in DNA synthesis reaction mixtures containing dideoxyribo-nucleotides

Unknown sequence
3' AACAGCTTCAGT 5'
+5' TTGT 3' Labeled primer
+ DNA polymerase
+ dATP, dCTP, dTTP, and dGTP

+ ddATP   + ddCTP   + ddTTP   + ddGTP
1   2   3   4

2. Products of the reactions

TTGTCGA   TTGTC   TTGTCGAAGT   TTGTCG
TTGTCGAA   TTGTCGAAGTC   TTGTCGAAG
TTGTCGAAGTCA

3. Electrophoresis of reaction mixtures

Lane 1 2 3 4

Number of nucleotides in fragment: 12 11 10 9 8 7 6 5

4. Autoradiography to visualize bands and deduction of 5' → 3' sequence of newly synthesized DNA strand by reading order of bands from bottom to top

A C T G

3'
A C T G A A G C T G T T
5'
Sequence of newly synthesized DNA (complementary to unknown sequence)
Primer

©Addison Wesley Longman, Inc.

http://blog.csdn.net/dyllove98

Shotgun sequencing

**Reads,
Contig,
Scaffold,
Chromosome walking**

Starting fragments:

ATCAGTA
AGTATCA
CAGTCAG
CAGTATAGC
TAGCTTGCA

Reconstruct based on overlapping regions with assembler

scaffold

contig 1

contig 2

ATCAGTA
AGTATCA
CAGTCAG

CAGTATAGC
TAGCTTGCA

Assembly of overlapping DNA sequencing   GCT

Assembled sequence   GCT

Aligned sequence

ATCAGTATCAGTCAGTATAGCTTGCA
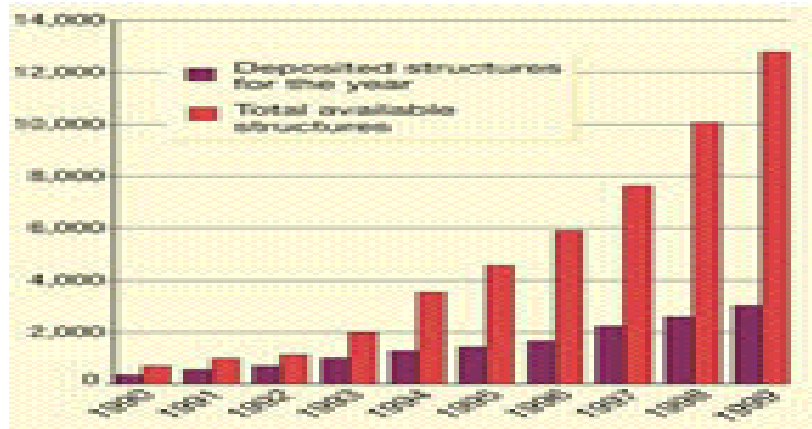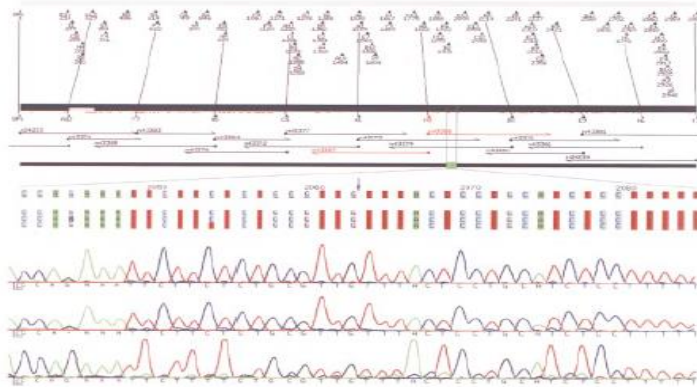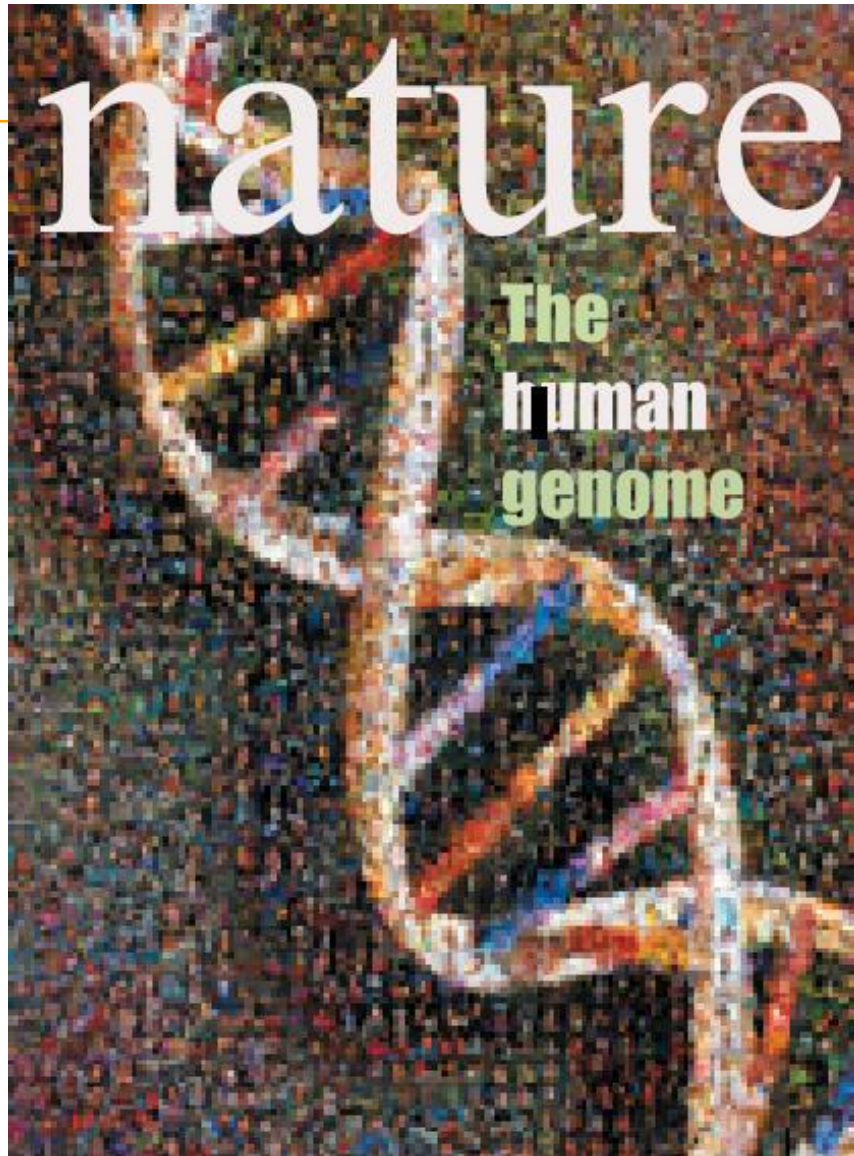
**1998年，科学家Ventor
等人组建Celera公司，同
时开展人类基因组测序
计划，与HUGO竞争**

**2000年6月，人类基因组
草图公布**

**2001年2月，人类基因组
图正式公布，宣告人类
基因组计划第一阶段任
务结束**

Everyone's genome

Nature 409, 813 (2001)

HGP

Science 291, 1304 (2001).

Cerela

人类基因组计划从此被写入历史

开工：1990年；

竣工：2003年；

参与国：美国、英国、德国、法国、日本和中国；

耗资：26亿美元；

成果：排出人类遗传物质中30亿个遗传密码的顺序。

人类基因组草图顺序

1）国际联合体的测序结果：

    总长： 2 692 Mb

    基因数： 26 383

2）Cerela Genomics 的测序结果：

    总长： 2 847 Mb

    基因数： 31 778

All life mainly depends on critical molecules

DNAs

- Hold information on how cell works

RNAs

- Act to transfer short pieces of information to different parts of cell
- Provide templates to synthesize into protein

Proteins

- Form enzymes that send signals to other cells and regulate gene activity
- Form body's major components (e.g. hair, skin, etc.)

The "New" Biology: *X*-omics

Traditional reductionistic approach:

- One gene/protein/reaction at a time.
- Test/validate isolated models at bench.
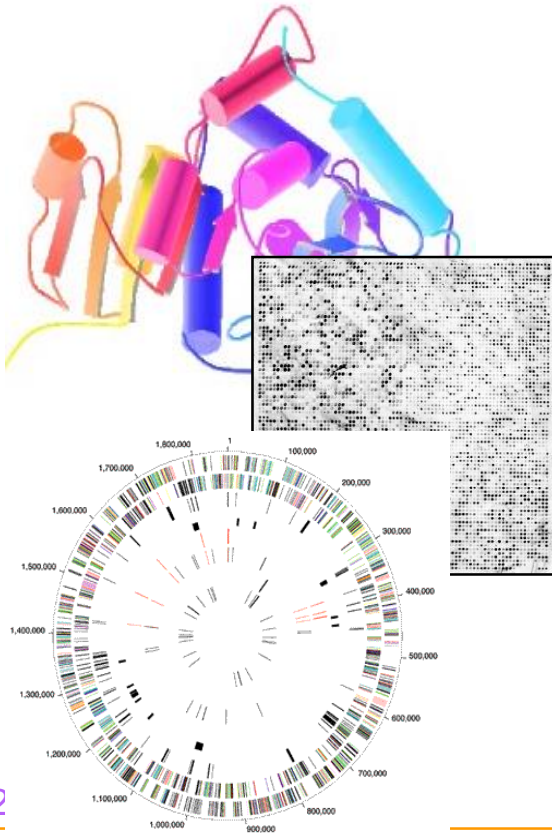
New "systems" approach:

- All DNA/RNA/proteins surveyed at once.
- Need to
    - Manage data globally (across labs, sites, ...)
    - Analyze large batches of intermediate results.
    - Provide links to minute details when required.

Bioinformatics

Bio-
information

+

Computer

Introduction to Bioinformatics

Bioinformatics is "the *in-silico* study of the information content and information flow in biological systems and processes".

- Michael Liebman in "Bioinformatics: An Editorial Perspective" (http://www.netsci.org/Science/Bioinform/feature01.html)

- Annotate -> store -> search/retrieve -> analyze -> visualize

Why Bioinformatics?

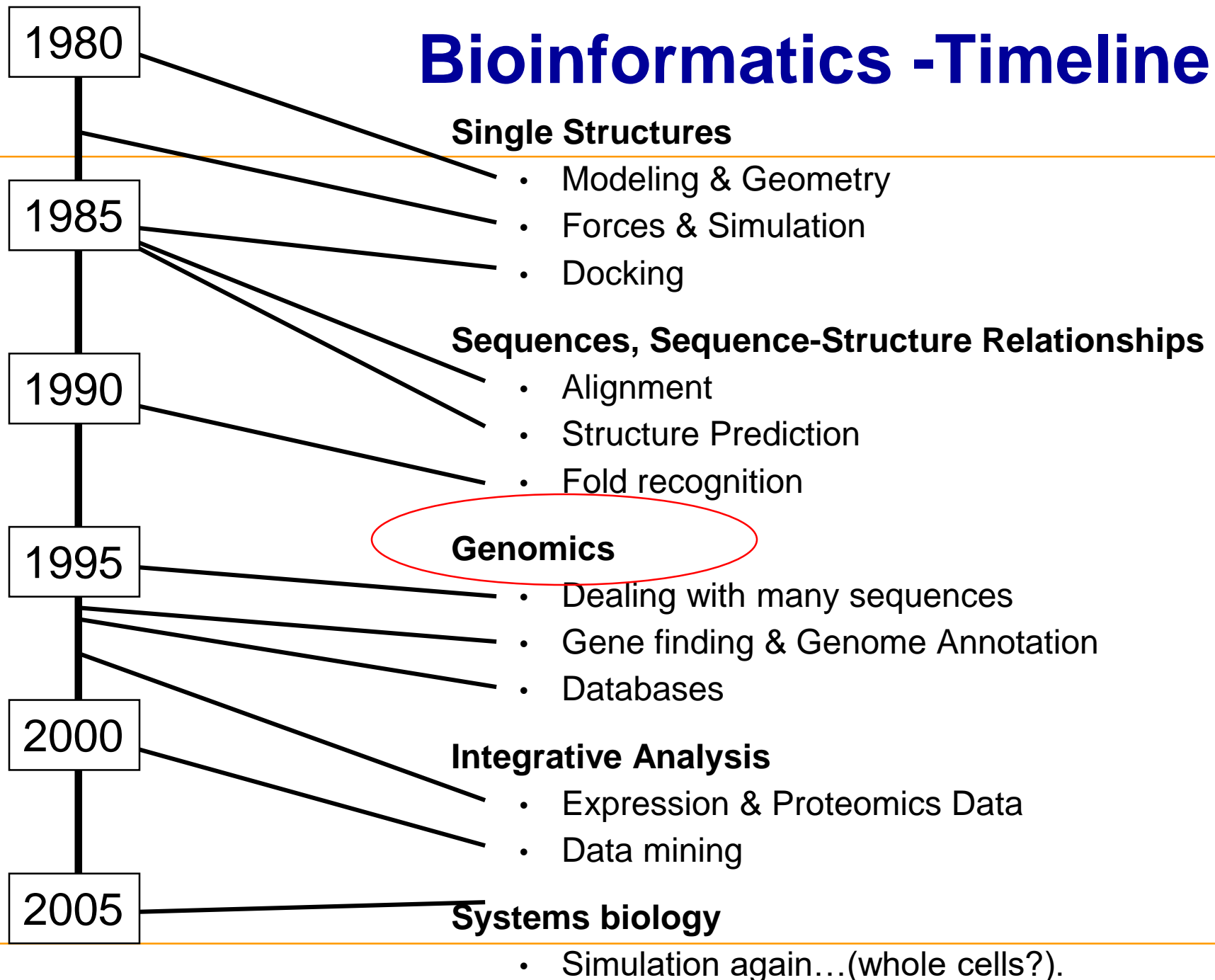Bioinformatics is the combination of biology and computing.

-Omics technologies have created massive amounts of information that can only be efficiently analyzed with computers.

As the information becomes ever larger and more complex, more computational tools are needed to sort through the data.

- Bioinformatics to the rescue!!!

Bioinformatics is the electronic infrastructure of molecular biology.

Bioinformatics -Timeline

1980

Single Structures

- Modeling & Geometry
- Forces & Simulation
- Docking

1985

Sequences, Sequence-Structure Relationships

- Alignment
- Structure Prediction
- Fold recognition

1990

Genomics

- Dealing with many sequences
- Gene finding & Genome Annotation
- Databases

1995

Integrative Analysis

- Expression & Proteomics Data
- Data mining

2000

Systems biology
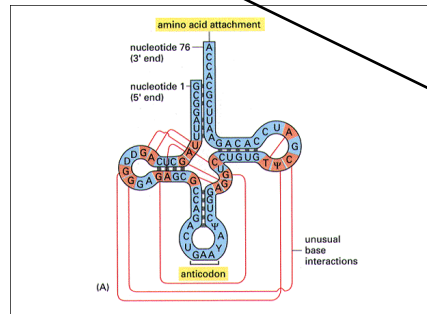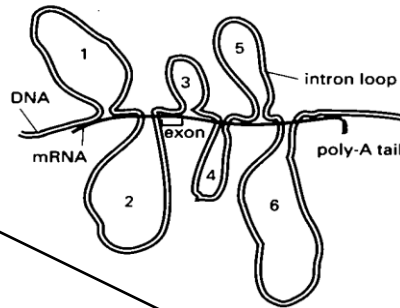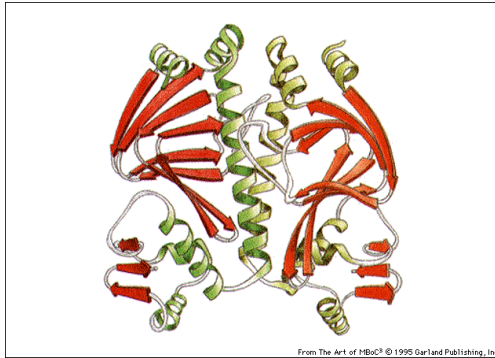
- Simulation again…(whole cells?).

2005

Bioinformatics Technologies

- Bioinformatics database development

- Sequence analysis

- Sequence comparison
  - Pair-wise
  - Mutiple sequence Alignment (MSA)

- Structure Modeling

- Evolution

- -omics analysis

Goal of Bioinformatics ?

Goals Leading Toward Predictive Biology

Gene Sequence Data

↓

Gene Identification

↓

Structure Prediction

↓

Protein Circuit & Regulatory Network Discovery

↓

Biosimulation