计算生物学

王扬帆

Email: yfwang@ouc.edu.cn

中国海洋大学遗传育种实验室

2014 年6 月

内容提要

- 背景
  - 课程要求
  - 课程主页
  - 参考资料
  - 考核方式
- 简介
  - HGP
  - 计算生物学

课程要求

本课程的层次构成

核心内容：

定量分析和数学建模 → 生命科学

两个层次：

—— 定量分析方法在现代生物学领域的初步应用

—— 现代生物学中的定量分析方法和数学建模实践

目的：

—— 启发思想，培养自主学习的能力

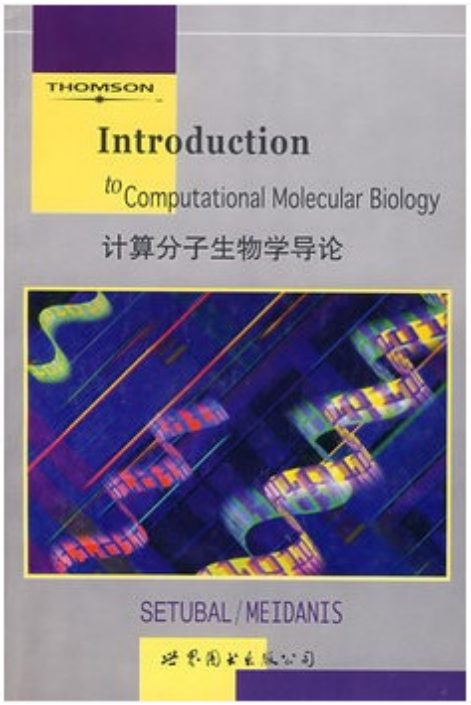
课程主页

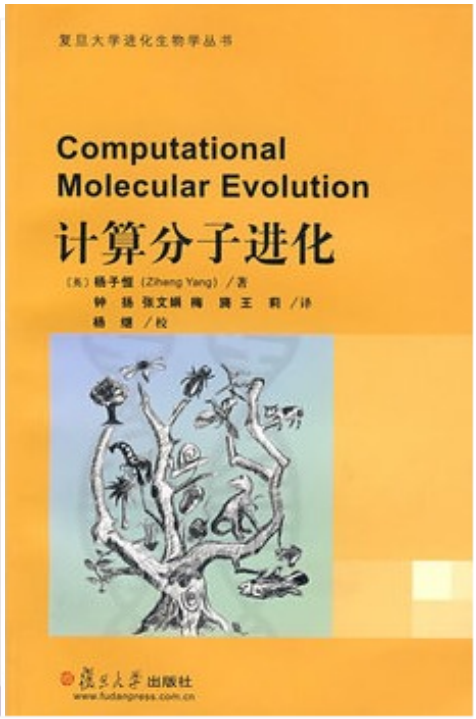
- 课程主页 http://vision.ouc.edu.cn/zhenghaiyong/courses/cb/2014fall/index.html

课次	周次	时间	内容	主讲人
1	二	2014年10月13日周一	[简介]课堂事务；课程定位和主要内容。	王扬帆
2	三	2014年10月20日周一	[CS]UNIX/Linux操作系统。	郑海永
3	四	2014年10月27日周一	[CS]开发环境（GCC、Makefile等）。	郑海永
4	五	2014年11月03日周一	[CS]O|B|F；LAMP；项目。	郑海永
5	六	2014年11月10日周一	[讨论]UNIX/Linux，开发环境，O|B|F，LAMP。	武斌
6	七	2014年11月17日周一	[MA]最大似然法；贝叶斯。	王扬帆
7	八	2014年11月24日周一	[BI&MA]系统发育重建。	王扬帆
8	九	2014年12月01日周一	[BI&MA]系统发育重建。	王扬帆
9	十	2014年12月08日周一	[BI&MA]系统发育重建。	王扬帆
10	十一	2014年12月15日周一	[讨论]系统发育重建。	王扬帆
11	十二	2014年12月22日周一	[BI&MA]分子遗传育种。	王扬帆
12	十三	2014年12月29日周一	[BI&MA]分子遗传育种。	王扬帆
13	十四	2015年01月05日周一	[BI&MA]分子遗传育种。	王扬帆
14	十五	2015年01月12日周一	[讨论]分子遗传育种。	王扬帆

参考资料


生物信息学数据分析丛书
理解生物信息学
Understanding Bioinformatics
〔英〕M. 泽瓦勒贝 J.O.鲍姆 著
李亦学 郝 涛 主译


THOMSON
Introduction
to Computational Molecular Biology
计算分子生物学导论
SETUBAL / MEIDANIS


复旦大学进化生物学丛书
Computational
Molecular Evolution
计算分子进化

考核方式

- 教学与考核方式

讲义授课
读书报告（1次）
课堂讨论（~ 6学时）

讨论（分组讨论发言、读书报告，占总成绩40%）

实践作业报告（占总成绩60%）

HGP

- 从人类基因组计划（HGP）说起

HGP

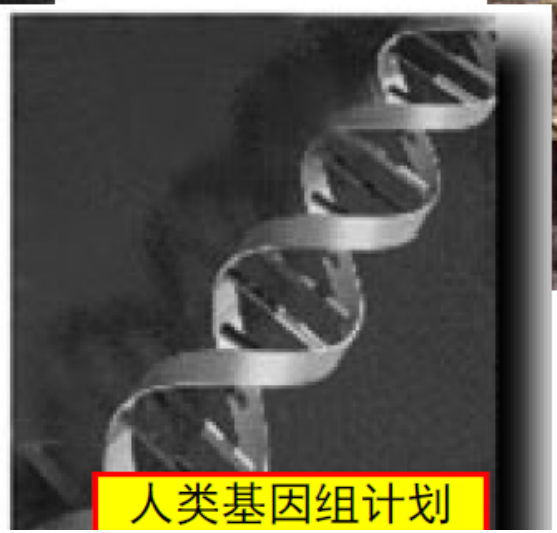

GOD... THE HUMAN GENOME CODE'S BEEN UNRAVELLED
DAMN HACKERS!!! NOW, I HAVE TO CHANGE THE PASSWORD

HGP


IT'S A BOY!
...
... SMOKER, HOMOSEXUAL, AND INSURANCE AGENT, WITH A 75% ALZHEIMER-PROBABILITY!
GENOME

HGP

- 20世纪
三大科学计划



HGP

Why HGP?

1961年，美国总统Kennedy提出两个科学计划：

☑登月计划

“我们选择登月”
(1962年Kennedy在Rice大学演讲)

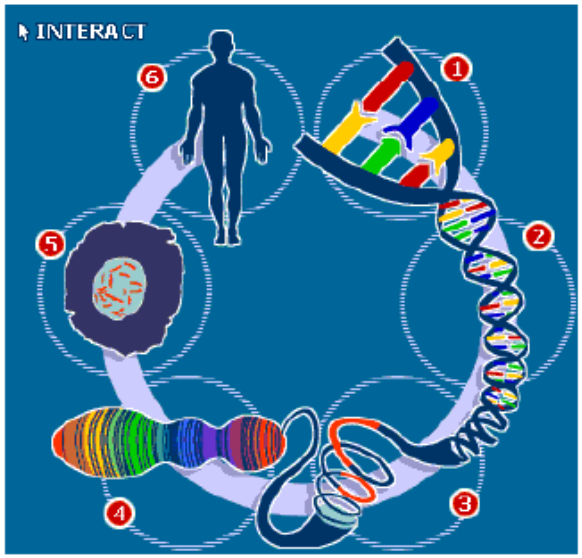
攻克肿瘤计划 ← 人类遗传信息的复杂性

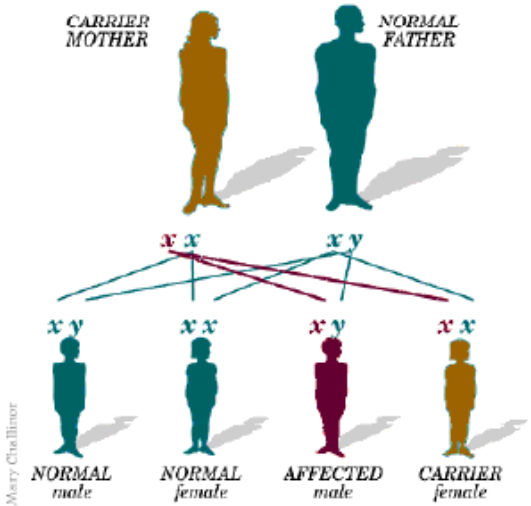
人类基因组计划
(HGP，Human Genome Project)
目标：整体上破解人类遗传信息的奥秘
1、“曼哈顿原子弹计划”历史遗留问题之产物
2、 对生命科学和医学的科学影响

HGP

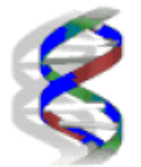
DNA、基因、基因组

生命活动三要素：物质、能量、信息





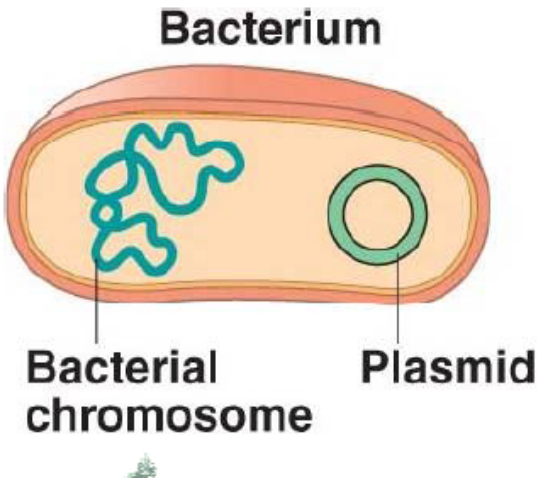
- **DNA**: 遗传物质(遗传信息的载体)→ 双螺旋结构
  A, C, G, T四种基本字符的复杂文本
- 基因（Gene）：具有遗传效应的DNA分子片段

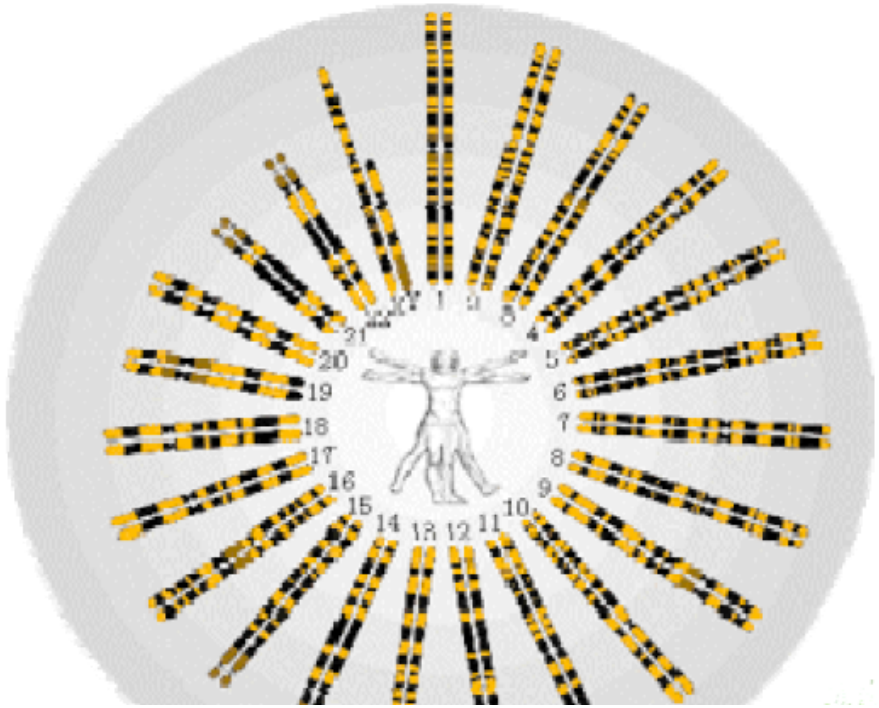


HGP

- 基因组(Genome)：包含细胞或生物体的全套遗传信息的全部遗传物质
  Winkles（1920），GENes+chromosOEs
  原核生物(细菌、古细菌、病毒等)
  真核生物(真菌、植物、动物等)

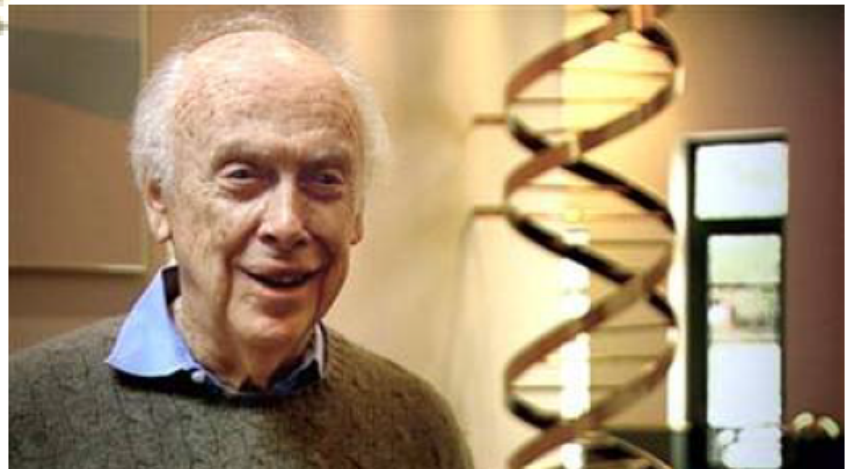
人类基因组：
3.2×10^9 bp，含有2万多个基因





HGP

- 尽管比之于人类登月，**HGP**的投入资金要少得多，但**HGP**对人类生活的影响要更为深远。因为随着这个计划的完成，**DNA**分子中编码的遗传信息将对人类存在的化学基础作出最终的回答。这将不仅帮助我们理解我们是如何作为健康的人发挥正常功能的，而且也将在化学水平上解释遗传因子在各种疾病，如癌症、早老痴呆症、精神分裂症等一些严重危害人类健康的疾病中的作用。毕竟对人类自身更深入的了解是人类活动中最重要的一个部分。



——Watson ,1990,《Science》

HGP

基因组研究的新革命

NATURE|Vol 445|25 January 2007

ESSAY

Putting the pieces together

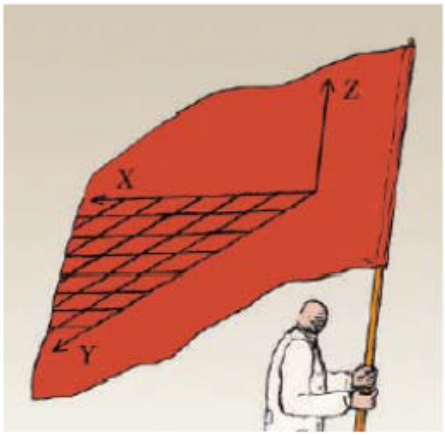
Biology's next revolution

The emerging picture of microbes as gene-swapping collectives demands a revision of such concepts as organism, species and evolution itself.

Nigel Goldenfeld and Carl Woese

One of the most fundamental patterns of scientific discovery is the revolution in thought that accompanies a new body of data. Satellite-based astronomy has, during the past decade, overthrown our most cherished ideas of cosmology, especially those relating to the size, dynamics and composition of the Universe.

Similarly, the convergence of fresh theoretical ideas in evolution and the coming avalanche of genomic data will profoundly alter our understanding of the biosphere — more powerful early forms of HGT.



KAPUSTA

Refinement through the horizontal sharing of genetic innovations would have triggered an explosion of genetic novelty, until the level of complexity required a transition to the current era of vertical evolution. Thus, we regard as regrettable the conventional concatenation of Darwin's name with evolution, because other modalities must also be considered.

This is an extraordinary time for biology, because the perspective we have indicated places biology within a context that must necessarily engage other disciplines more

HGP

- 已完成测序的10,599个基因组

种类	数目	备注
古细菌(Archaea)	132	
古细菌病毒(Archaeal virus)	51	
真细菌(Bacteria)	1,983	其中有的测定了2个以上的菌株
真核生物(Eukaryo)	161	包括酵母、线虫、果蝇、蚊子、拟南芥、人等
病毒(Virus)	2,842	包括不同亚类或不同株系
类病毒(Viroid)	56	包括不同亚类或不同株系
噬菌体(Phage)	1,029	包括不同亚类或不同株系
细胞器(Organelle)	3,510	包括线粒体和叶绿体
质粒(Plasmid)	835	

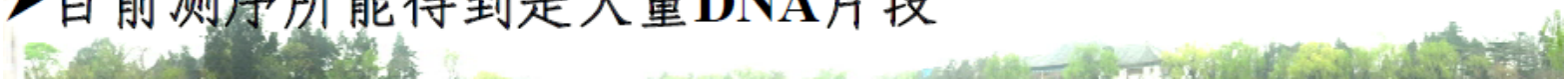
（http://www.ebi.ac.uk/genomes/，2012年7月）

HGP

- **Gene-swapping collectives**
- **Microbiota**（微生物群落）
- **Microbiome**（微生物群落基因组，“微生物组”）
- **Metagenomics**（宏基因组学、元基因组学、环境基因组学、多源基因体学）

Microbiome的主要性质：

➢不依赖培养的微生物群落（数量极多的基因组的集合，而且来自不同的物种）
➢与群落的生态环境或宿主有十分重要的相互作用
➢基因组之间存在活跃的、复杂的遗传物质交换关系，基因组内存在快速的、复杂的基因组演化
➢目前测序所能得到是大量DNA片段



HGP

最新研究：消化道就像细菌乐园

【美联社华盛顿2010年3月3日电】最新研究发现，人的消化道就像是一个动物园，充满了各种各样的细菌。科学家说，这其实是件好事。

一项旨在给人体内胆非人类基因分类的国际研究的初步结果显示，约170种细菌在普通人的消化道内都可以找到。研究还发现，有肠道炎症的人肠道内的细菌种类较少。

我们体内超过99%的不同种类基因其实都不属于我们自身，而是来自于微生物。研究者之一、中国遗传学研究人员王俊说，分析我们体内的细菌遗传性将大大改进人类基因组图谱的绘制。

“细菌统治着地球，其中也包括我们的身体。”研究者之一、欧洲分子生物学实验室的研究人员拉埃说，“我认为人们应该认识到，我们其实不是人类——而是能够行走的细菌的宿主，它们对我们的幸福和健康至关重要。”

通过对124名成年人的分析，研究人员发现，大多数人的消化系统都有很多相似之处。至少有57种细菌存在于所有人都肠道内。研究人员一共找到了约1000种不同的细菌，估计还有150种左右尚未找到。

HGP

HGP的研究特色

1、大协作研究：
以学科为中心➔以问题为中心，多学科合作

2、研究的计划性和有序性：
正反双方共同参与，制定更科学、更全面的研究计划

3、商业竞争促进基础研究：
1998年Celera公司的加入

4、政府与国家的作用：
美：领导与推动
英：始于1989年2月，贡献为1/3左右
法：始于1990年6月，贡献为3%左右
日：始于1990年，贡献为7%左右

HGP

5、可持续性：

太空观测和基因组计划都是科学上出色的计划，每一个都是科学上迈出的一大步。但是两者之间存在着一个刺眼的差别：开支方面有四十倍的差别。开支的差别是至关重要的，因为这意味着可持续性。当一个计划足够便宜到成为一条可以无限向未来延伸的系列的第一个时，它是可持续的。而当一个计划太昂贵，以至不经过重大的政治斗争就无法重复时，它就是不可持续的。可持续计划带来新计划的开始，不可持续计划则标志着老时代的结束。

《The Sun, the Genome, and the Internet
——Tools of Scientific Revolution》

HGP

HGP带来的科学挑战

●各学科参与、协作：生命科学、数学、物理学、化学、计算机科学、材料科学以及伦理、法律等社会科学……

HGP ——Pandora's Box

“It’s never the same!”

计算生物学

- **首要科学问题**

  如何找到记载在基因组DNA一维结构上控制生命时间、空间的调控信息的编码方式和调节规律。

  应用数学、复杂系统理论、信息论、非线性科学……

  催生➔**生物信息学、计算生物学、系统生物学**

- **DNA芯片技术**

  交叉性技术领域：物理学、微电子信息技术、生化技术、信息技术……

- **结构生物学**

  前沿领域之一：生物物理学、生物化学、晶体学、波谱学、光谱学以及X射线晶体衍射技术、核磁共振技术……

计算生物学

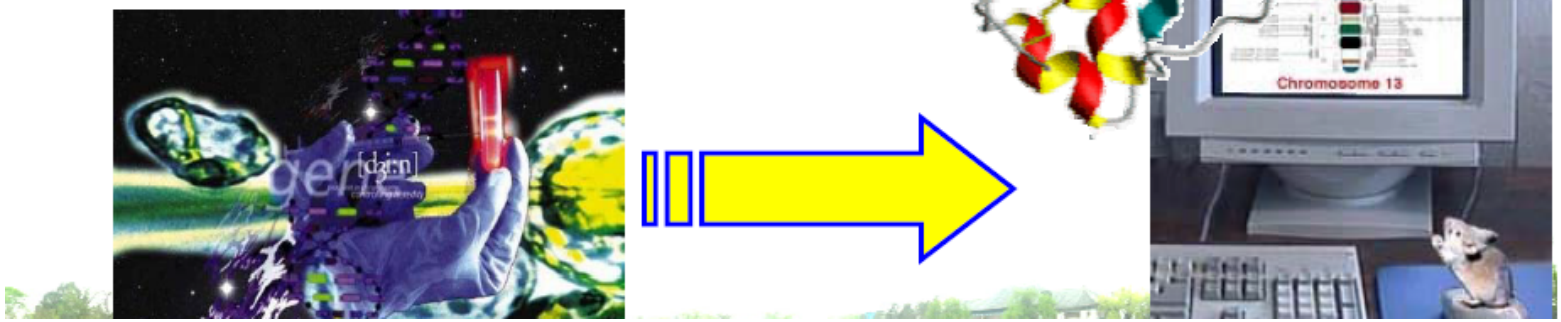
面对堆积如山的生物学数据

计算生物学

Walter Gilbert
(1932-)
Harvard University,
Biological Laboratories

……新的生物学研究模式的出发点应该是理论的。科学家将从理论推测出发，然后再返回到实验中去，追踪或验证这些理论假设。……生物学家不仅必须成为计算机学者，而且也要改变他们研究生命现象的途径。

——W. Gilbert, Towards A Paradigm Shift in Biology, *Nature*, 1991



计算生物学

- **传统生物学：**实验科学

现代生物学的发展：

1、高通量数据获取日益实现自动化、半工业化

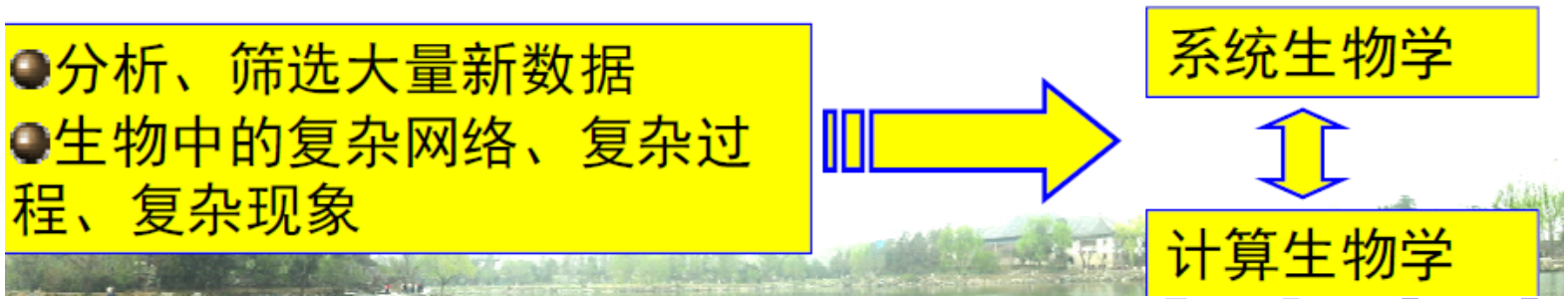
从数据库中实现数据挖掘、知识发现

2、海量数据

难以完全依赖实验手段对新数据进行分析，必须借助计算机实现分析和筛选

3、更复杂层次的生物学问题

复杂的基因调控网络、代谢网络；细胞间信号转导过程；生物个体全部基因表达变化……



计算生物学

计算生物学、理论生物学与实验研究

- 实验永远起着决定作用
- 计算/理论生物学的发展离不开实验生物学的贡献
- 实验生物学日益依赖计算/理论生物学的指导

计算生物学

- 美国国家卫生研究院（NIH）的定义：

Computational Biology (The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.)

开发和应用数据分析、理论方法、数学模型和计算机仿真技术，用于生物学、行为学和社会群体系统的研究。

Bioinformatics (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.)

为拓展生物学、医学、行为学和卫生学数据的用途，而进行有关计算机方法手段的研究、开发与应用，包括此类数据的采集、存贮、整理、归档、分析与可视化。

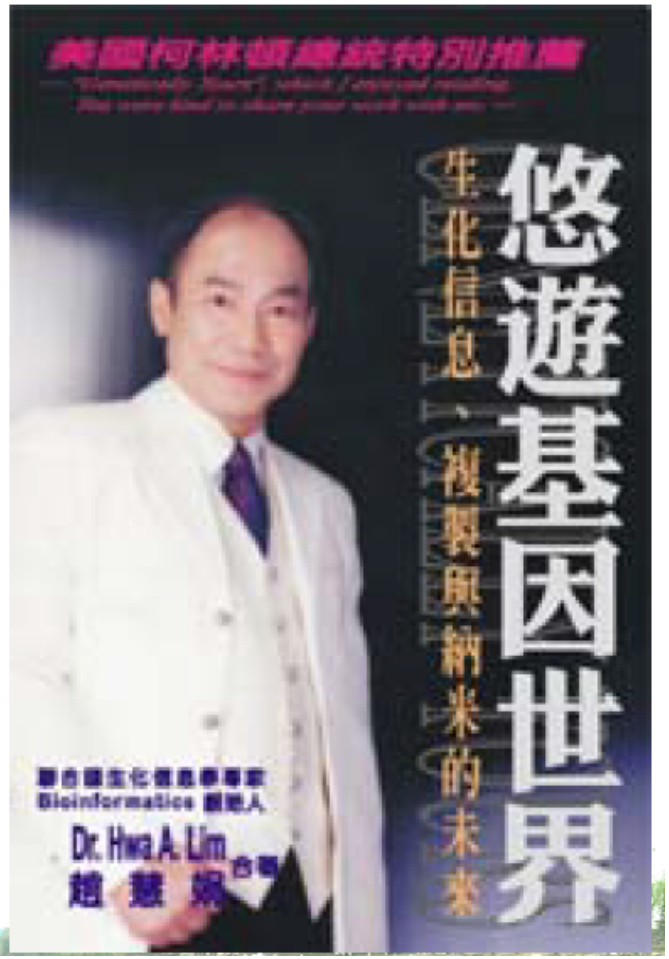
计算生物学

- **生物信息学（Bioinformatics）的来源**

●Dr. Hwa A. Lim （林华安）1987年提出
“Bio-informatique” → “Bioinformatics”

1955年出生于马来西亚。University of Texas at Dallas分子与细胞生物学Adjunct Professor、中国科学院基因遗传研究所客座教授。1981年Imperial College, London University毕业，1986年获得美国Rochester University生化物理学博士学位，30岁取得佛罗里达州立大学终生教授。1992年受聘担任美国国家癌症中心及美国国家科学基金会审核委员。1995年后，历任多家生物科技公司生化信息执行长、副总裁等高层管理职位。1997年，创立结合软件与数据分析的专业顾问公司D’ Trends，服务生物技术、制药及卫生保健等机构。

林华安



计算生物学

- **多学科交叉、互相推动发展**

  生物学、物理学、化学、数学、计算机科学、信息科学、系统科学.....

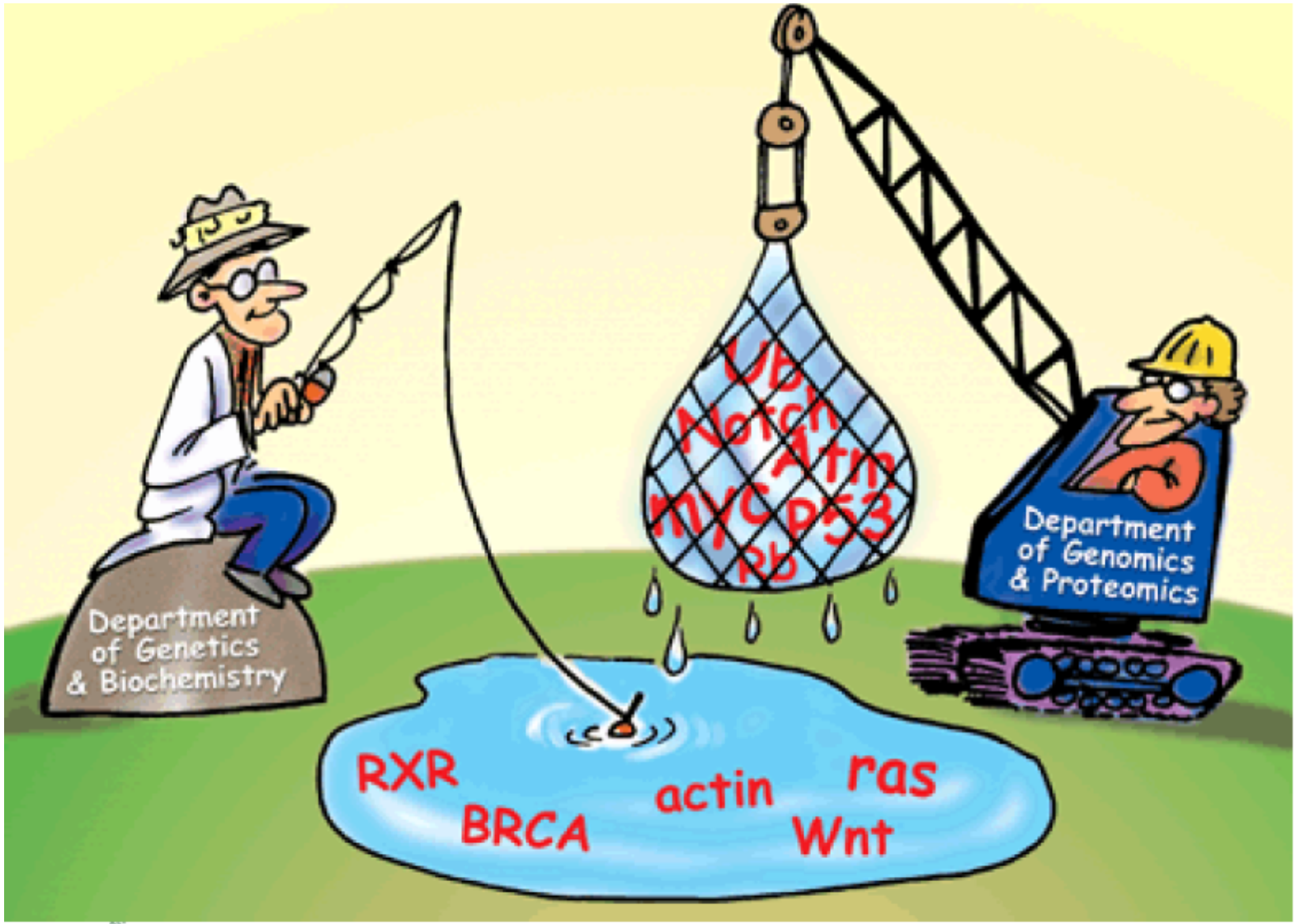
生物信息学/计算生物学

揭示基因组蛋白质组信息结构的复杂性、遗传语言的根本规律

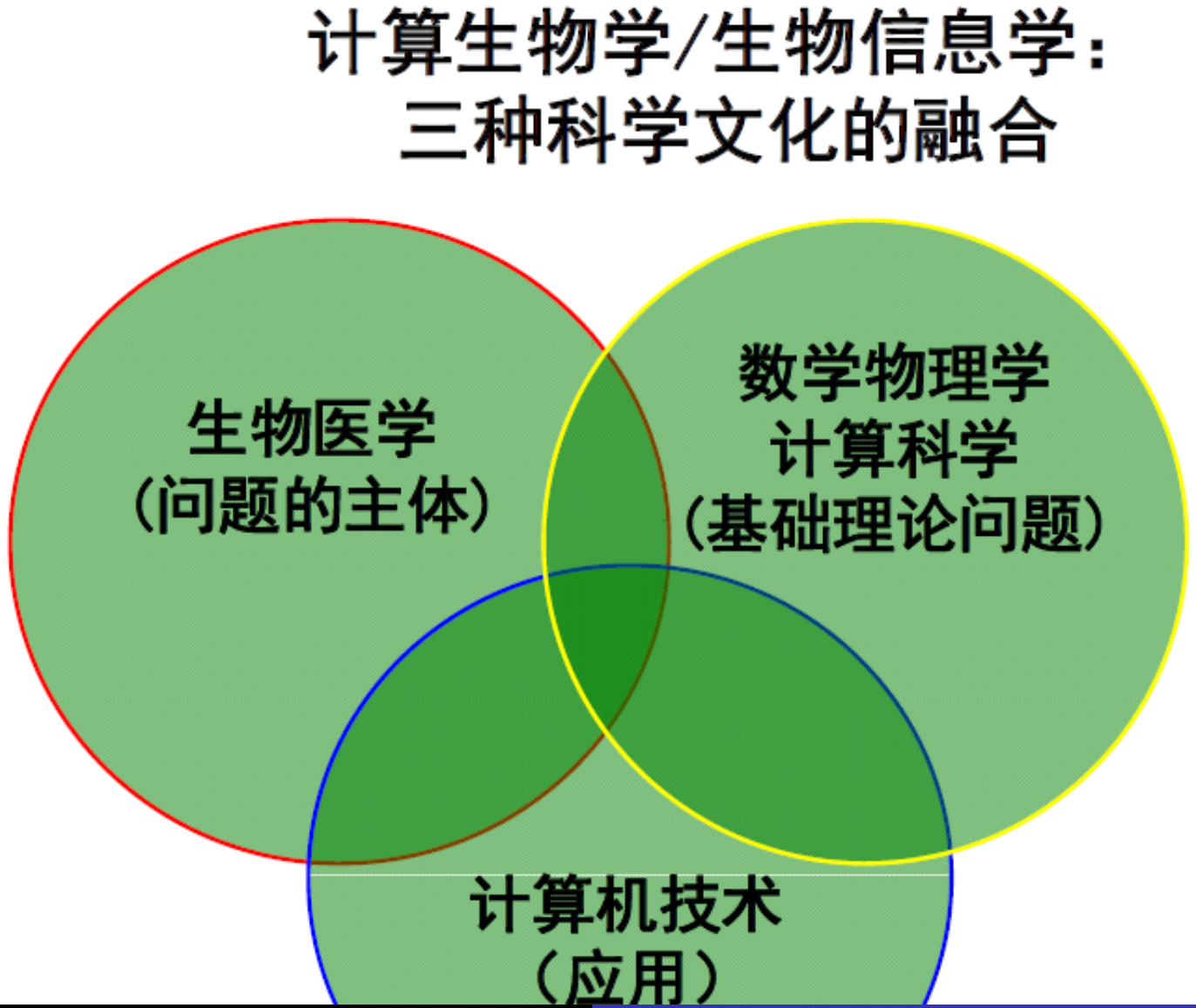
- 生物学信息的计算机处理
- 计算生物学 / 理论生物学

以基因组DNA序列信息为源头，识别基因组序列中代表蛋白质和RNA基因的编码区，阐明非编码区的信息特征，破译隐藏在DNA序列中的遗传语言规律；
同时，归纳、整理与基因组遗传语言信息释放及其调控相关的转录谱和蛋白质谱的数据，从而认识代谢、发育、分化、进化的规律。

计算生物学


Department of Genetics & Biochemistry
Department of Genomics & Proteomics
RXR
BRCA
actin
ras
Wnt

计算生物学


计算生物学/生物信息学：
三种科学文化的融合
生物医学
（问题的主体）
数学物理学
计算科学
（基础理论问题）
计算机技术
（应用）

计算生物学

计算生物学的主要研究内容

- 生物信息数据的收集、存储、管理与提供
- 基因组学：基因组序列特征的分析、比较
- 功能基因组学：基因功能发现、表达分析及突变检测
- 生物大分子结构模拟和药物设计
- 生物信息分析的技术与方法研究
- 应用与发展研究

计算生物学

1 核酸和蛋白质序列分析研究

DNA序列 RNA序列 蛋白质

由重复的核苷酸或氨基酸单元组成的线性高分子，具有高度有序并能完成特定生物学功能的三维结构

目的

揭示序列蕴含的更高级的结构和功能信息

主要思想

具有相似序列的分子，可能具有相似的三维结构和生物学功能。
首要任务：提取反映结构、功能性质的序列特征

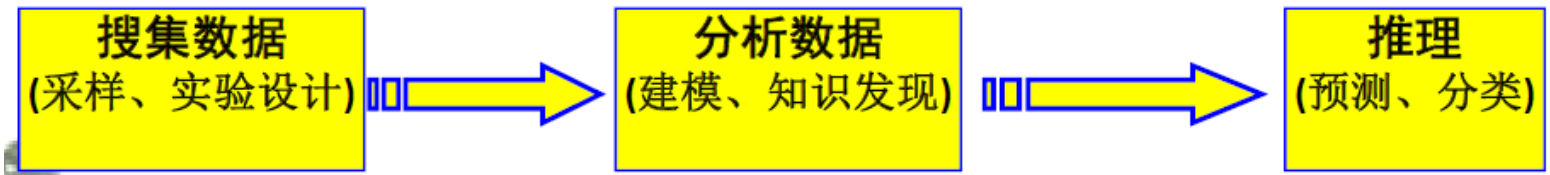
主要方法

基于数据挖掘或知识发现（**data-mining, knowledge discovery**）的方法：
统计方法、机器学习、神经网络等

计算生物学

统计方法

从事物的外在数量上的表现去推断事物可能的规律性

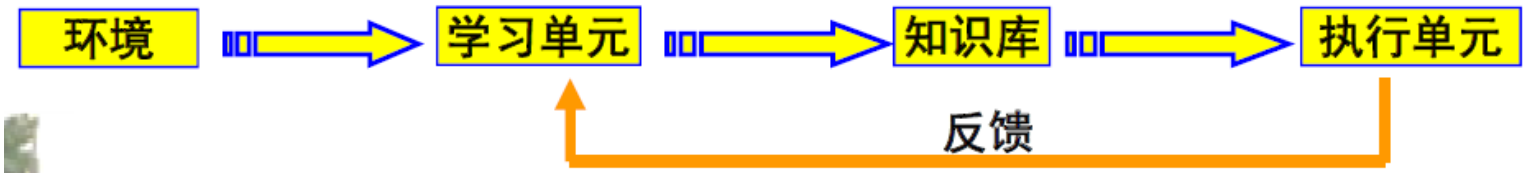


- 回归分析：多元回归、自回归...
- 判别分析：线性判别分析、非线性判别分析...
- 聚类分析：系统聚类分析、动态聚类分析...

计算生物学

机器学习方法

学习：系统通过执行某种过程而改进它的性能



规则归纳	AQ算法…
决策树	基于可读规则与决策树的数据分类方法
范例推理	直接使用过去的经验或解法来求解给定的问题
遗传算法	通过模拟自然进化过程搜索最优解（自然选择、突变进化）

计算生物学

人工神经网络方法

模仿人脑神经网络的结构和某些工作机制，利用大量的神经元连成网络来实现大规模并行计算。通过学习，改变神经元之间的连接强度。

McCulloch-Pitts模型

多层感知器模型

反传网络模型

计算生物学

序列分析中的主要算法

生命科学中的问题		数理问题/算法
相似性搜寻	两两序列比对	寻优算法
	相似序列的数据库搜寻	——动态规划算法
	多序列比对	——模拟退火算法
	系统发育树重建	——遗传算法
	蛋白质三维结构比对	——人工神经网络方法

计算生物学

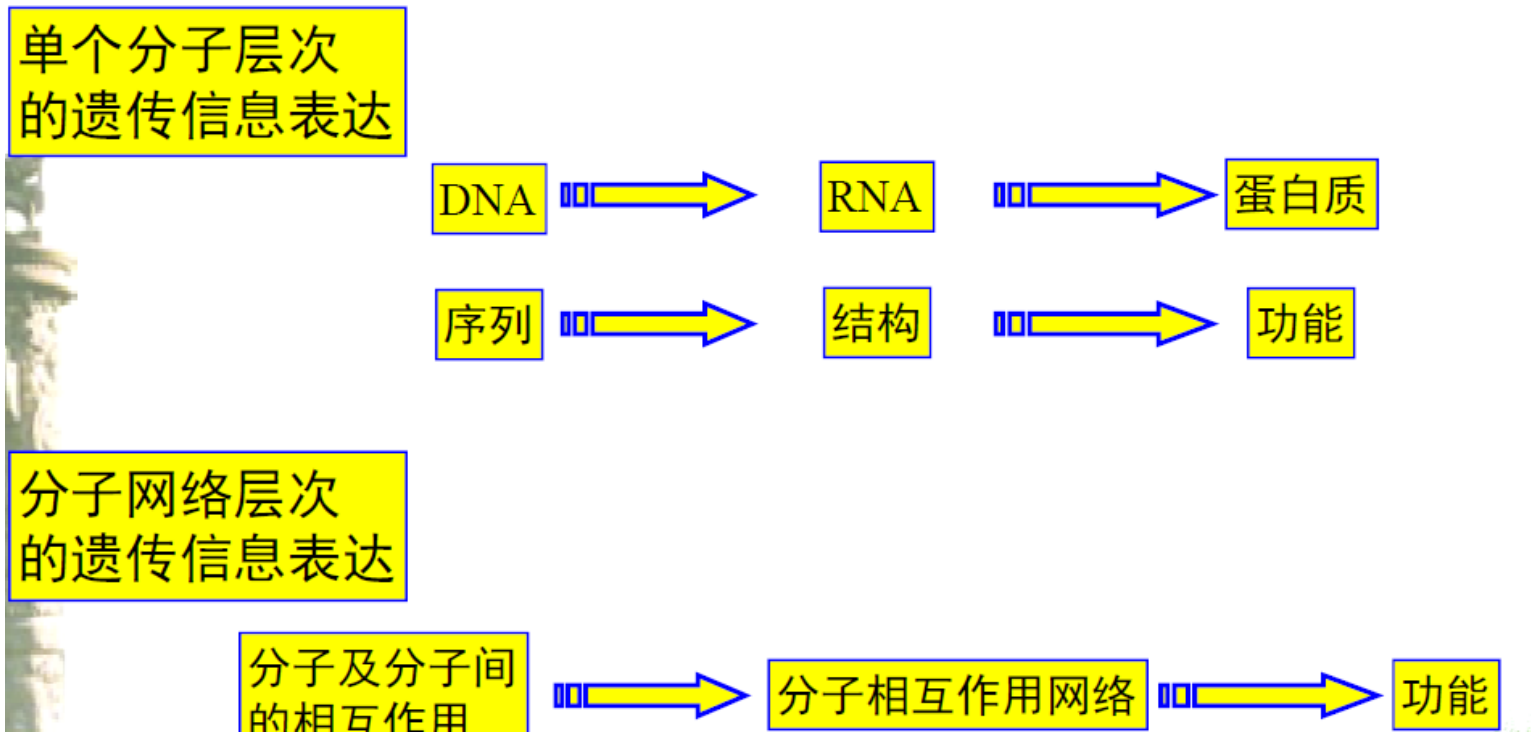
生命科学中的问题		数理问题/算法
结构/功能的从头预测	RNA二级结构预测 RNA三级结构预测 蛋白质三级结构预测	寻优算法 ——动态规划算法 ——模拟退火算法 ——遗传算法 ——人工神经网络方法

计算生物学

生命科学中的问题		数理问题/算法
结构/功能的基于知识预测	Motif提取 功能部位预测 细胞定位预测 编码区预测（基因结构预测） 跨膜片段预测 蛋白质二级结构预测 蛋白质三维结构预测	模式识别、学习算法 ——判别分析方法 ——人工神经网络算法 ——随机过程方法(HMM) ——语言学(形式语法) ——演化、遗传算法

计算生物学

2 生物分子相互作用的复杂系统模拟



计算生物学

系统水平的生物信息学方法

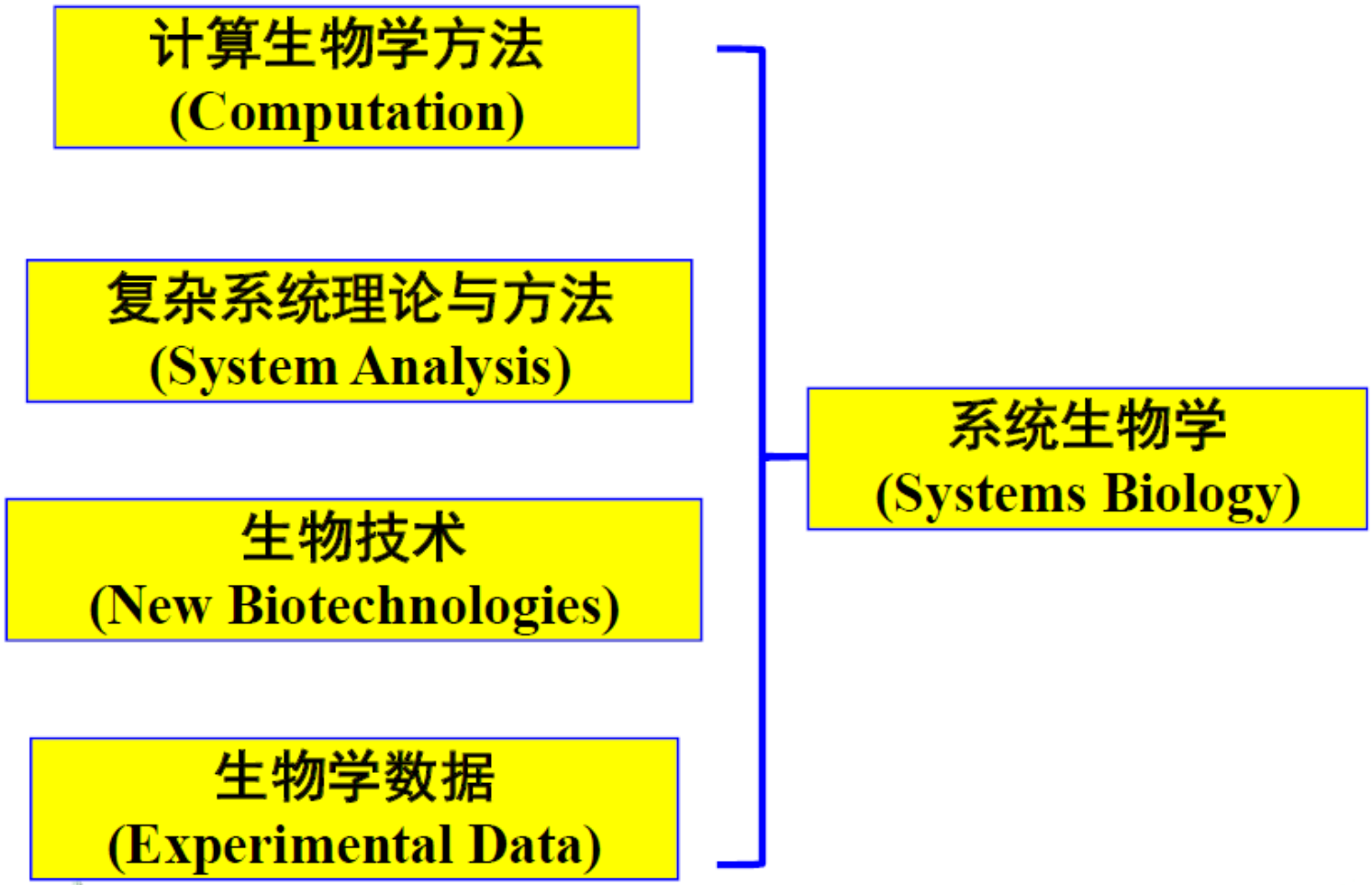
主要目标 **Predict the dynamics of systems so that the validity of the underlying assumptions can be tested.**

主要思想 复杂网络系统分析

主要途径 1、静态研究：路径计算、二元关系和演绎…（离散数学方法）

2、动态研究：微分方程组、随机过程…（网络的时间依赖演化）

计算生物学


计算生物学方法
(Computation)
复杂系统理论与方法
(System Analysis)
生物技术
(New Biotechnologies)
生物学数据
(Experimental Data)
系统生物学
(Systems Biology)

计算生物学

初级层面

基于现有的生物信息数据库和资源，利用成熟的计算生物学和生物信息学工具（专业网站、软件）解决生物学问题

——生物信息数据库（NCBI、EBI等）

——基因组序列分析、序列比对软件（GCG、BLAST、CLUSTAL等）

——转录组数据分析软件（Bowite、BLAT）、代谢组数据分析软件（SIMCA-P）

——系统发育树构造软件（PHYLIP、PALM等）

——分子动力学模拟软件（GROMACS、NAMD等）

计算生物学

- **中级层面**

  利用数值计算方法、数理统计方法和相关的工具，研究计算生物学和生物信息学问题

  ——概率、数理统计基础

  ——科学计算基础

  ——现有的数理统计和科学计算工具（**EXCEL**、**SPSS**、**SAS**、**MATLAB**等）

  ——建立有特色的生物学数据库

计算生物学

- **高级层面**

  提出有重要意义的计算生物学和生物信息学问题；自主创新，发展新型方法，开发新型工具，引领计算生物学和生物信息学领域研究方向。

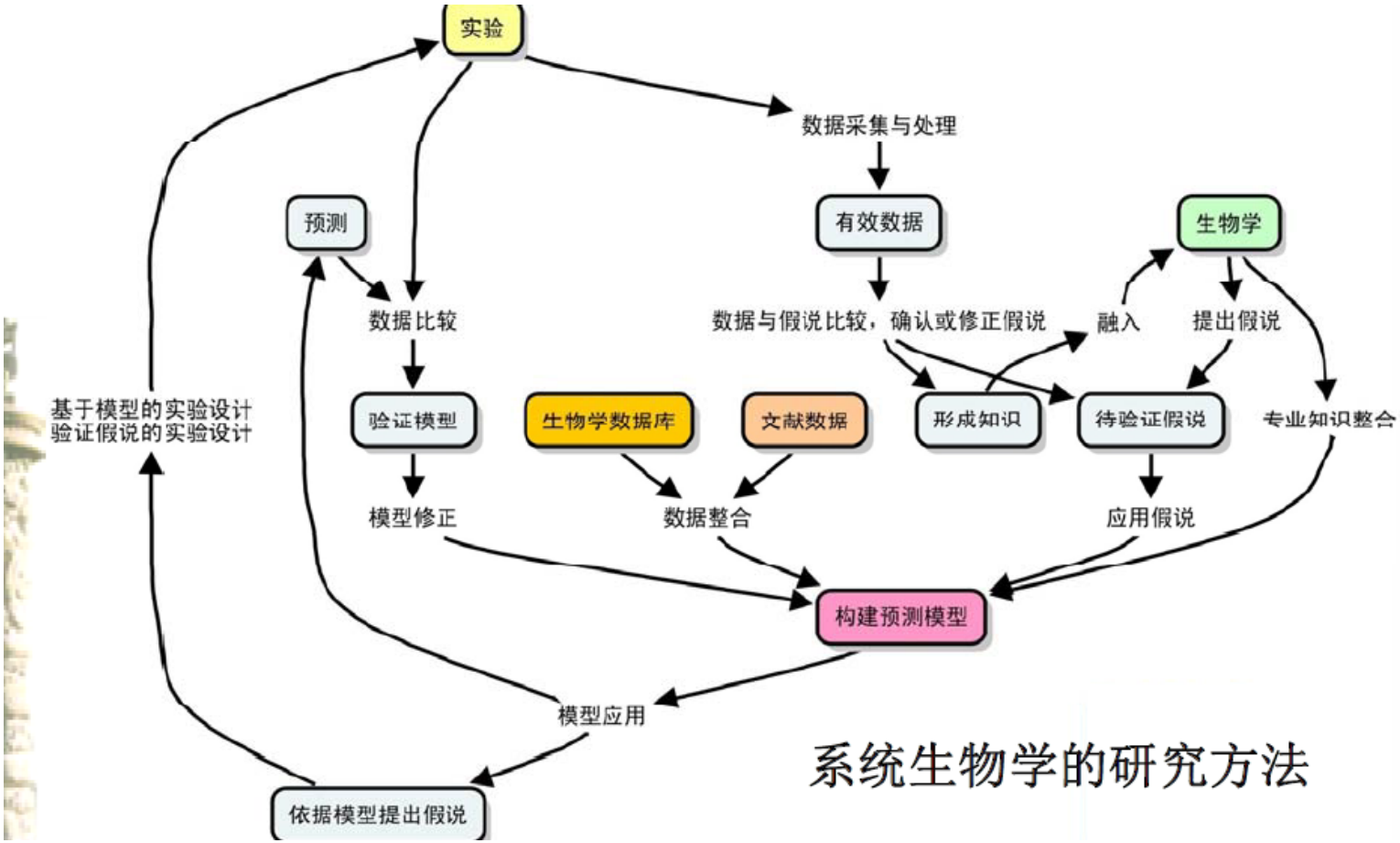
  ——面向生物学领域，解决生物学问题，**wet & dry lab**

  ——数学、物理、化学、计算科学等思想和方法

  ——建立模型，发展算法

  ——自行编程，开发软件，建立网页（**Linux**系统、**C/C++**、**PERL**、数据库技术）

计算生物学


实验
数据采集与处理
预测
有效数据
生物学
数据比较
数据与假说比较，确认或修正假说
融入
提出假说
基于模型的实验设计
验证假说的实验设计
验证模型
生物学数据库
文献数据
形成知识
待验证假说
专业知识整合
模型修正
数据整合
应用假说
构建预测模型
模型应用
依据模型提出假说
系统生物学的研究方法

计算生物学

计算生物学/生物信息学的“降龙十八掌”

计算生物学

（1）

生物信息数据库及其查询搜索方法

(Database & searching)

——对分子生物信息数据库的种类以及某些具体数据库的掌握和了解

——从现有数据库中熟练获得需要的数据信息（尤其是二级数据库）

——能熟练地进行数据库查询和数据库搜索（数据库查询系统Entrez、SRS；搜索工具BLAST等）

计算生物学

- （2）

计算生物学软件和工具的应用

(Software & application)

利用成熟的计算生物学工具（专业网站、软件）解决生物学问题

——基因组序列分析、序列比对软件（GCG、BLAST、CLUSTAL等）

——系统发育树构造软件（PHYLIP、PALM等……）

——基因芯片检测分析软件（商业软件ScanArray、Array-Pro等 ……）

——分子动力学模拟软件（GROMACS、NAMD等……）

计算生物学

- （2）

计算生物学软件和工具的应用

(Software & application)

利用成熟的计算生物学工具（专业网站、软件）解决生物学问题

——基因组序列分析、序列比对软件（GCG、BLAST、CLUSTAL等）

——系统发育树构造软件（PHYLIP、PALM等……）

——基因芯片检测分析软件（商业软件ScanArray、Array-Pro等 ……）

——分子动力学模拟软件（GROMACS、NAMD等……）

计算生物学

（3）

概率论基础

(Probability theory)

——随机事件、概率

——随机变量、概率分布

——大数定律、中心极限定理

"Most of the problems in computational sequence analysis are essentially statistical."

——"Biological sequence analysis"

计算生物学

- （4）

统计学基础

(Statistical methods)

——样本和统计量（方差、均值……）

——参数估计、假设检验

——基本的统计分析（方差分析、协方差分析、回归分析）

——常用统计软件的运用（SPSS、SAS）

计算生物学

（5）

基于频率的组分分析方法和权重矩阵方法

(Composition analysis & weight matrix method)

——词汇频率反映具有生物学意义的序列特征

——核酸组分、氨基酸组分、密码子使用频率

——*k*-tuples/*k*-mers频率分析

计算生物学

（6）

信息论方法

(Information method)

——信息符号、状态空间（ACGT四种符号，及其所有可能的排列）
——信息的度量是信息符号出现何种状态的一种不确定性程度，信息的获得要对不确定性进行否定。

——信息熵（Shannon，1948）

$$H = -\sum_i p_i \log p_i$$

——信息熵H刻画了由$\{p_i\}$表示的随机试验结果的先验不确定性，或观察到输出时所获得的信息量。

计算生物学

- （7）

期望最大化（EM）方法

(Expectation Maximization)

——EM算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。在每一迭代循环过程中交替执行两个步骤：E步（Expectation step），在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望；M步（Maximization step），用极大化对数似然函数以确定参数的值，并用于下步的迭代。算法在E步和M步之间不断迭代直至收敛。该方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。

——适用于具有隐变量的模型和问题，如神经网络模型中的隐节点、HMM模型中的隐状态

计算生物学

（8）

动态规划方法

(Dynamic Programming)

——一种常用的多阶段决策的寻优算法

——基本思想：在状态空间中，根据目标函数，通过递推，求出一条从状态起点到状态终点的最优路径（代价最小的路径）。其策略是将一个问题递归分解为两个规模更小的相似子问题。

——动态规划在生物信息学研究中用得最多的方面是DNA序列或者蛋白质序列比对，或应用于隐Markov模型中寻找最优的隐状态序列。

计算生物学

- （9）

迭代方法

(Iteration)

——不断用变量的旧值递推新值的过程

——迭代的目的通常是在状态空间找到目标函数收敛的稳定解

——在运用模式识别方法时，对系统参数的学习通常要经过迭代来实现

——迭代必须能够不断逼近稳定解

计算生物学

（10）

回归、拟合、相关性分析、关联分析

(Regression, fitting, correlation & association)

——经典的统计分析方法

——Regression: the relation between selected values of x and observed values of y (from which the most probable value of y can be predicted for any value of x)

——主要目的：描述和预测自变量与因变量间的关系

计算生物学

（11）

判别分析方法

(Discriminant analysis)

——用于判别样品所属类型的统计分析方法

条件：已知研究对象总体的类别数目及其特征（如：分布规律，或各类的训练样本）

目的：判断未知类别的样本的归属类别

——多元判别分析、线性判别分析、非线性判别分析

——基因识别、医学诊断、人类考古学 .....

计算生物学

（12）

聚类分析方法

(Clustering method)

——聚类分析（群分析）是实用多元统计分析的一个新分支，正处于发展阶段。理论上尚未完善，但应用十分广泛。实质上是一种分类问题，目的是建立一种分类方法，将一批数据按照特征的亲疏、相似程度进行分类。

——条件：研究对象总体的类别数目未知，也不知总体样本的具体分类情况

——目的：通过分析，选定描述个体相似程度的统计量、确定总体分类数目、建立分类方法；对研究对象给出合理的分类。(“物以类聚”是聚类

计算生物学

（13）

Markov模型的应用

(Markov model)

——Markov过程：从一种状态转移到另一种状态时，过程仅取决于前面n种状态，是一种有序n模型。 n是影响下一个状态选择的状态数。

——最简单的Markov过程是一阶过程，状态的选择完全取决于前一状态，这种选择是依照概率来选择的。

——状态的选择是概率的，而非确定的。故Markov过程本质上是一种随机过程。

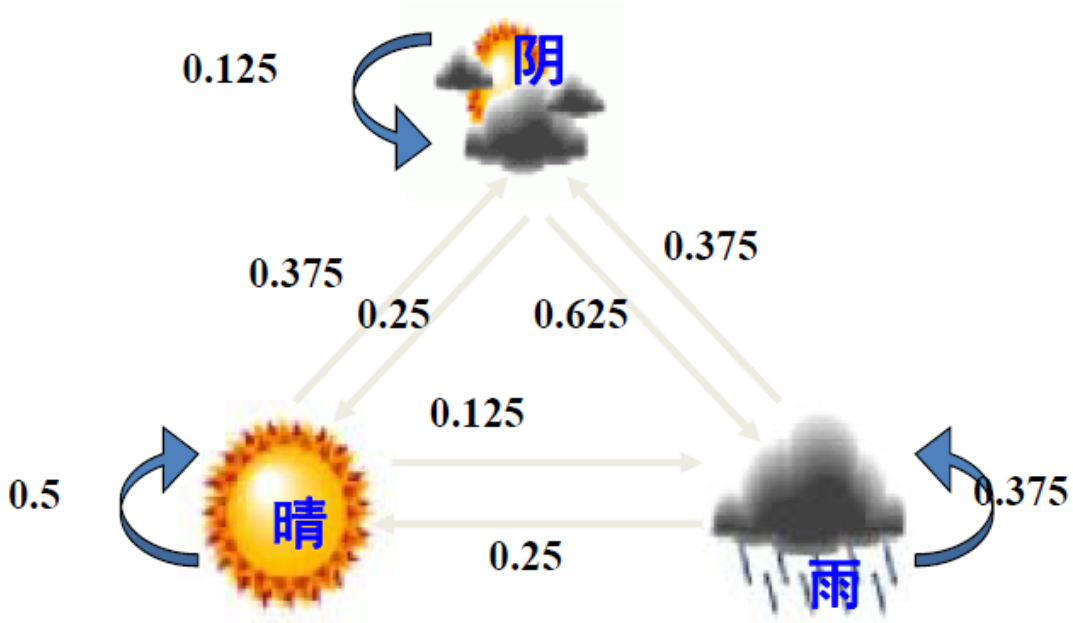
计算生物学

（1）天气状态：

晴

阴

雨



（2）状态转移矩阵：

某地天气状态的一阶转移概率矩阵		昨天的天气		
		晴	阴	雨
今天的天气	晴	0.5	0.25	0.25
	阴	0.375	0.125	0.375
	雨	0.125	0.625	0.375

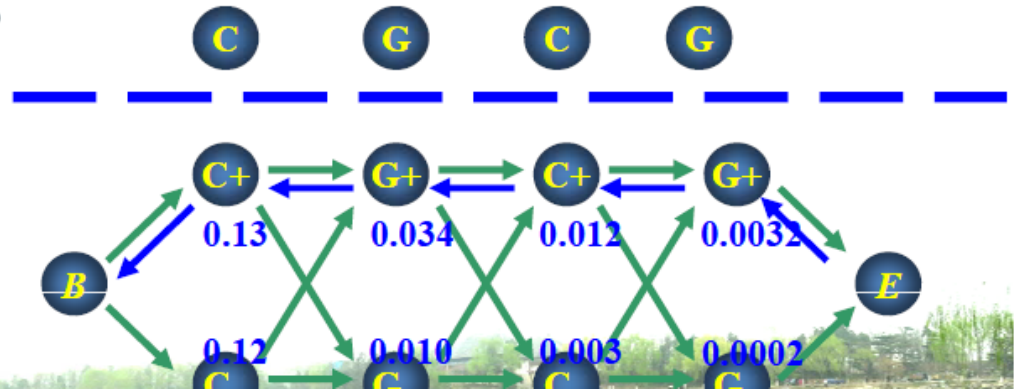
计算生物学

- （14）

隐Markov模型方法

(HMM method)

——将核苷酸序列看成一个随机序列，DNA序列的不同功能部分在核苷酸的选用频率上对应着不同的Markov模型。由于这些Markov模型的统计规律是未知的，而HMM能够自动寻找出它们隐藏的统计规律。对于复杂的DNA序列，HMM必须学习不同的基因结构的信号。



计算生物学

（15）

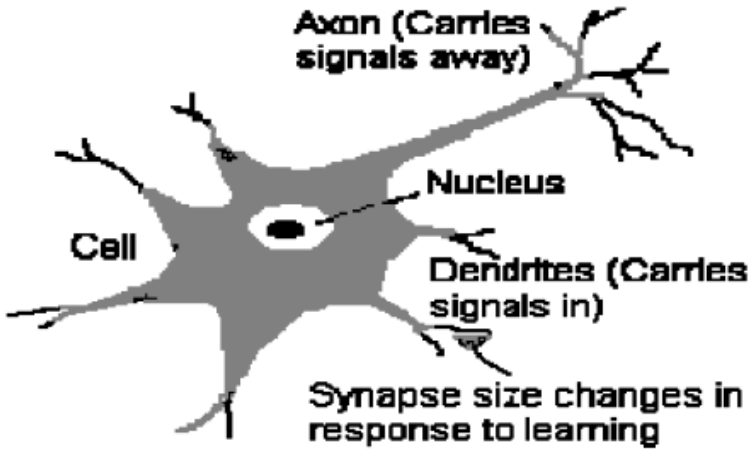
感知器与人工神经网络方法

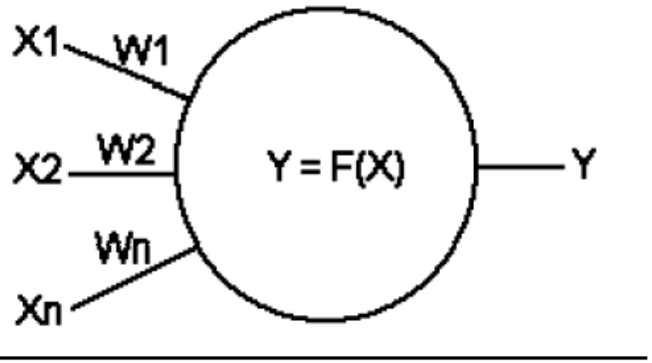
(Perceptron & ANN method)

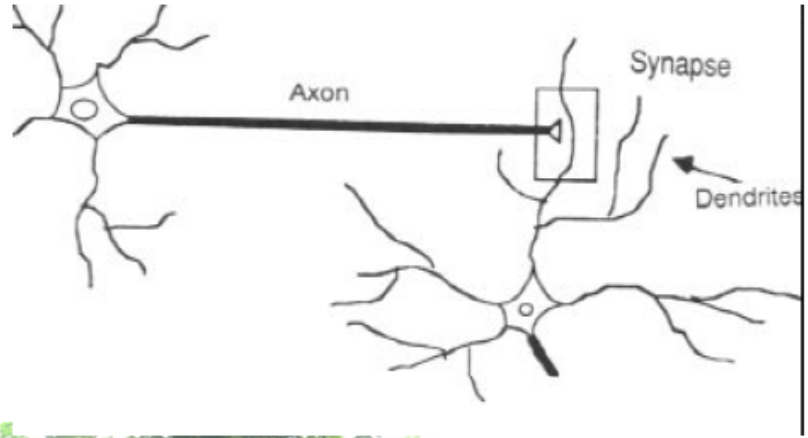
——A collection of mathematical models that emulate some of the observed properties of biological nervous systems and draw on the analogies of adaptive biological learning.

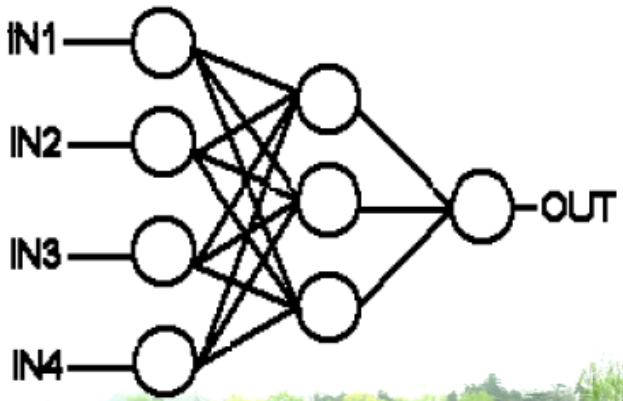
——The key element of the artificial neural network (ANN) model is the structure of the information processing system. It is composed of many highly interconnected processing

计算生物学


Axon (Carries signals away)
Nucleus
Cell
Dendrites (Carries signals in)
Synapse size changes in response to learning


X1
W1
X2
W2
Y = F(X)
Y
Wn
Xn


Axon
Synapse
Dendrites


IN1
IN2
IN3
IN4
OUT

计算生物学

（16）

决策树、支持向量机及其它模式识别方法

(Decision tree & SVM method)

——模式识别是在输入样本中寻找特征并识别对象的一种方法。

——模式识别主要有两种方法，一种是根据统计特征进行识别，另一种是根据对象的结构特征进行识别，而后者常用的方法为句法识别。

——在基因识别中，对于DNA序列上的功能位点和特征信号的识别常用到模式识别。

Thank you!

Email: yfwang@ouc.edu.cn