大数据驱动的生命科学研究方法

-- 生物信息学概论

宁康
NING Kang

Professor,
Director, Department of Bioinformatics and Systems Biology
School of Life Science, Huazhong University of Science and Technology
ningkang@hust.edu.cn

2020/10, Wuhan

生物信息学研究的三个层面

初级层面
中级层面
高级层面

初级层面

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

——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）

——系统发育树构造软件（PHYLIP、PALM、MEGA等）

——分子动力学模拟软件（GROMACS、NAMD等）

——搜集、整理有特色的生物信息学数据集

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）

——建立有特色的生物信息学数据库

高级层面

提出有重要意义的生物信息学问题；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

——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）

从事生物信息学研究应具备多方面的科学基础：

（1）、一定的计算能力，包括相应的软、硬设备。要有各种数据库或者能与国际、国内的数据库系统进行有效的交流。要有发达、稳定的互联网络系统；
（2）、强有力的创新算法和软件。没有算法创新，生物信息学就无法获得持续的发展；
（3）、与实验科学，特别是与自动化的大规模高通量的生物学研究方法与平台技术建立广泛、紧密的联系。这些技术，既是产生生物信息数据的主要方法，又是验证生物信息学研究结果的关键手段。

从事生物信息学研究的人员必须具备多学科交叉的知识。

生物信息学的“降龙十八掌”

第一式　见龙在田

（1）

要掌握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）

——数据库技术、互联网技术

第二式　飞龙在天

（2）

要学会生物信息学软件和工具的应用

(Software & application)

利用成熟的生物信息学工具（专业网站、软件）解决生物信息学问题

——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……）

第三式　鸿渐于陆

（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

——几乎用于生物信息学的各个方面

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

——“Biological sequence analysis”

（4）

数理统计基础

(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）

——几乎用于生物信息学的各个方面

第五式　羚羊触藩

（5）

基于频率的组分分析方法和权重矩阵方法

(Composition analysis & weight matrix method)

——符号（如碱基）频率反映具有生物学意义的序列特征，如内含子剪接位点的发现，KOZAK规则的发现等

——核酸组分、氨基酸组分、密码子使用频率

——主要用于具有特定生物学意义的序列特征的分析

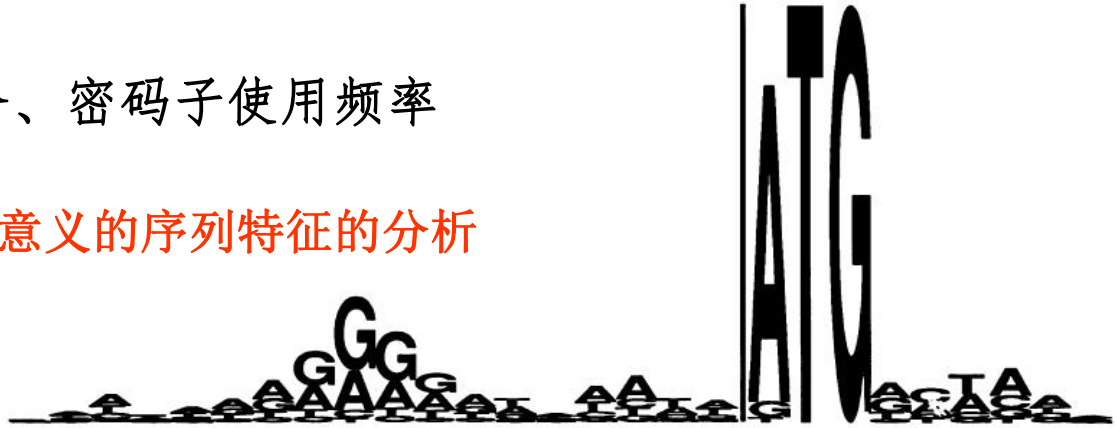


Figure 1. Logo for *E. coli* ribosome binding sites. Only −18 to +8 of the −20 to +13 site is shown. The first translated codon is just to the right of the 2 bit high vertical bar. 149 natural sites were used to create the logo (9).

权重矩阵分析方法举例

——针对序列信号（一段核酸、蛋白），计算每一位点所使用的词汇或叫符号（*碱基、氨基酸*）频率，频率的偏好性反映信号的序列特征（**sequence pattern**）。

例：人类基因内含子/外显子剪接位点的序列特征分析

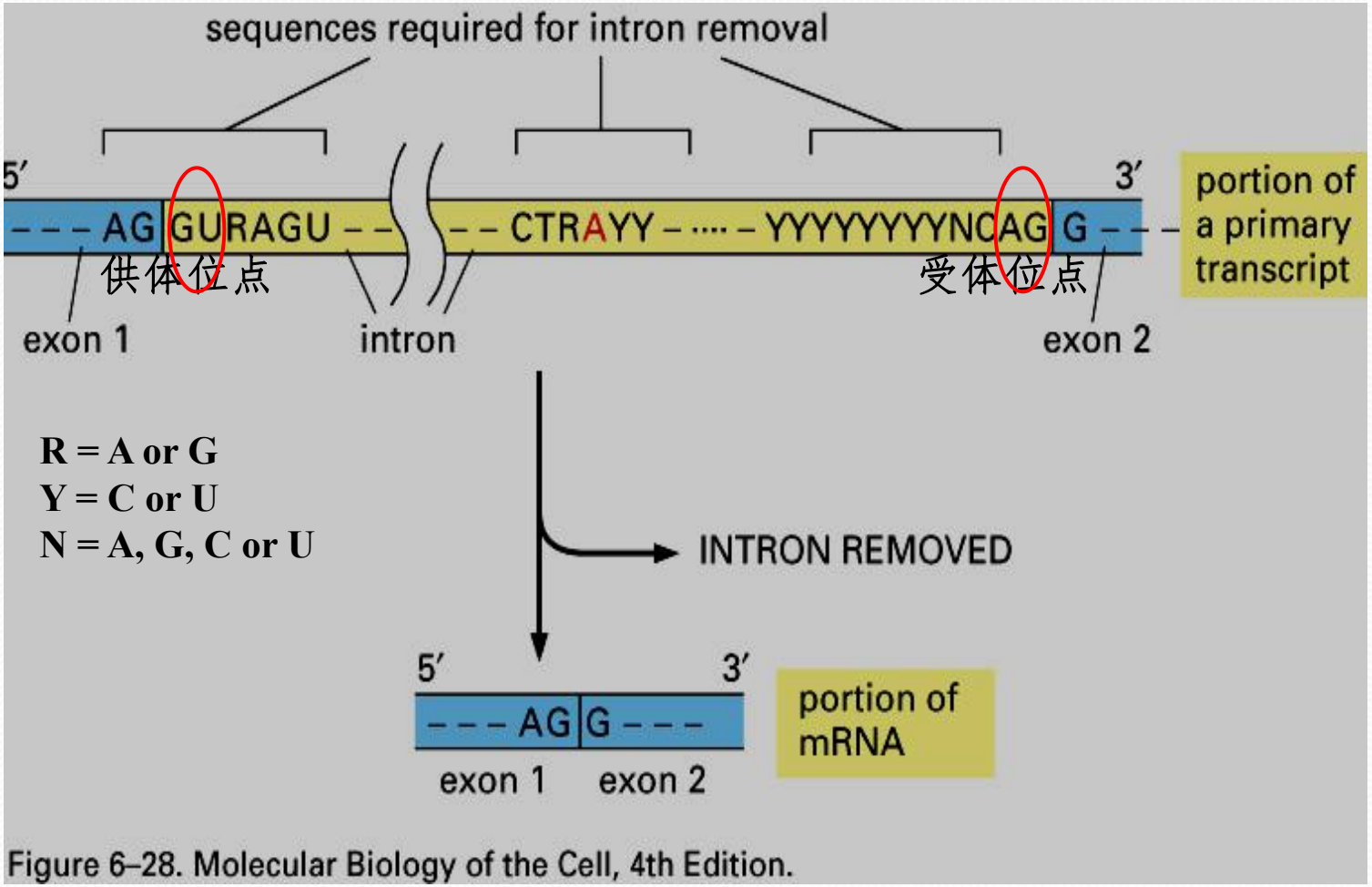


Figure 6–28. Molecular Biology of the Cell, 4th Edition.

Bayesian打分函数用于剪接位点预测的公式

The likelihood that a property value *v* (of a new structure) is drawn from the splicing site is:

$$P(site \mid \upsilon) = \frac{P(\upsilon \mid site)P(site)}{P(\upsilon \mid site)P(site) + P(\upsilon \mid nonsite\ )P(nonsite\ )'}$$

Score for the overall likelihood of the query sequence being a site is:

$$\sum_{\substack{\text{properties at} \\ \text{associated volumes}}} \log\left(\frac{P(site \mid \upsilon)}{P(site)}\right)$$

Say we have a sequence S= $S_1S_2\ldots S_n$. Then one need to calculate

$$\frac{P(S|\text{splice site})}{P(S|\text{background})}$$

So to look for a donor site in the sequence, we might calculate

（6）

信息论方法

(Information method)

——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。

——生物信息的符号如ACGT四种符号，状态空间即其所有可能的排列

——用于结构预测

——信息熵 $$H = -\sum_i p_i \log p_i$$

——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。

第七式　利涉大川

（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。

——适用于具有隐变量的模型和问题，

——用于结构的识别，如Motif识别的MEME方法、HMM中的Baum-Welch算法

第八式　神龙摆尾

（8）

动态规划方法

(Dynamic Programming)

——一种常用的多阶段决策的寻优算法
——动态规划用得最多的方面是DNA序列或者蛋白质序列比对

第九式　密云不雨

（9）

迭代方法

(Iteration)

——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

——用于上述某些方法的方法

第十式　突如其来

（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
——主要目的：描述和预测自变量与因变量间的关系
——用于上述某些方法的方法

第十一式　双龙取水

（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

——用于基因识别、医学诊断、人类考古学

第十二式　鱼跃于渊

（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。

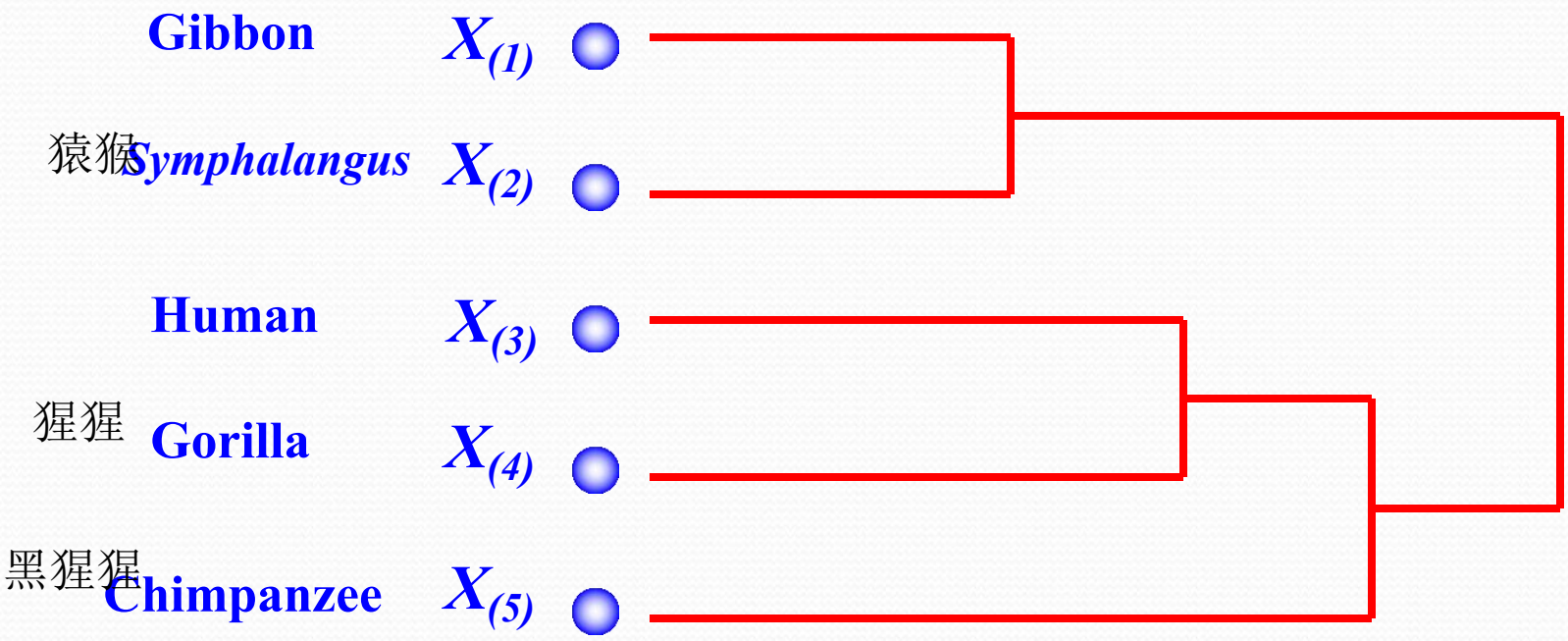
——条件：研究对象总体的类别数目未知，也不知总体样本的具体分类情况

——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”是聚类分析的基本出发点)

——定性、经验的分类的局限
分类较粗、数据量小、凭借经验

——谱系聚类法（系统聚类法）、动态聚类法、模糊聚类法

——生物信息学中的聚类分析问题：
根据DNA芯片获得的基因表达数据进行基因聚类（数据量庞大）
蛋白质相互作用网络的分类
根据不同物种的大分子序列进行相似性比较并构建系统发育树



（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。

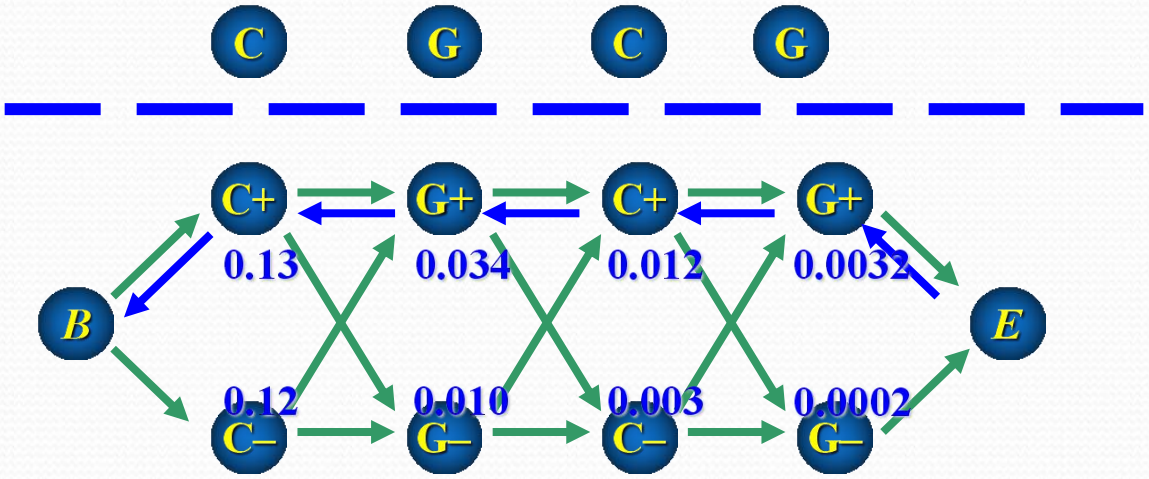
——状态的选择是概率的，而非确定的。故Markov过程本质上是一种随机过程。

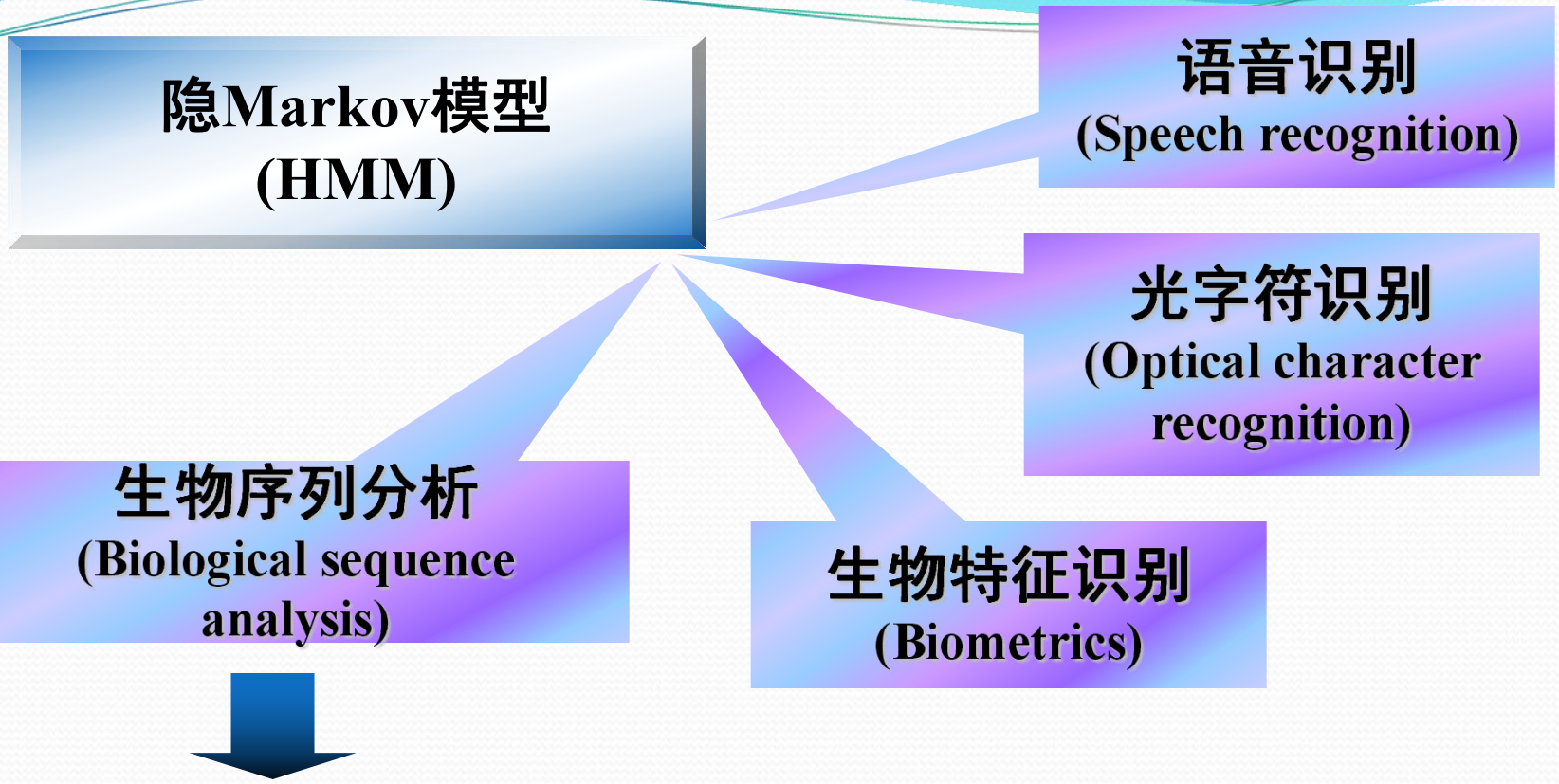
（14）

隐Markov模型方法

(HMM method)

——将核苷酸序列看成一个随机序列，DNA序列的编码部分与非编码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。对于高等生物这样复杂的DNA序列，HMM必须学习不同的基因结构的信号。





（1）序列比较与搜寻（尤其是多序列比对）

（2）基因及信号的识别、预测（包括DNA编码与非编码区的识别、真核基因剪接位点信号识别、非编码区的转录调控信号识别、信号肽识别……）

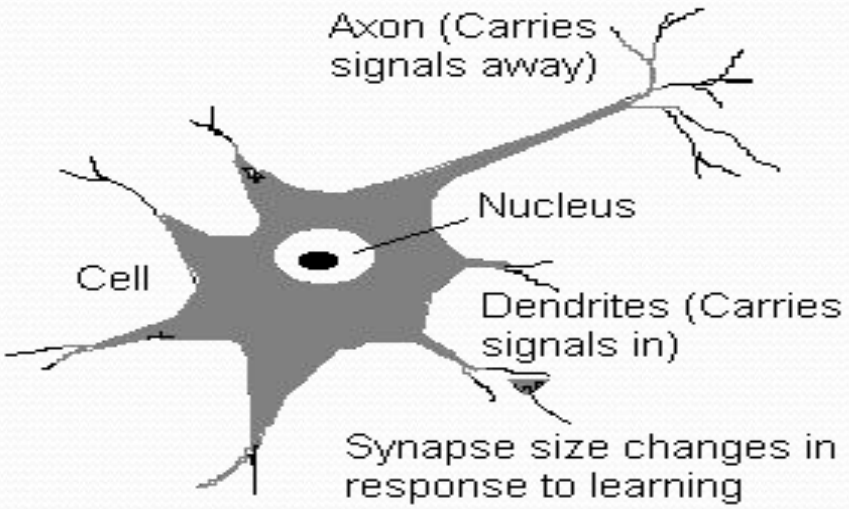
（3）蛋白质二级结构、家族、超家族预测、分类等……

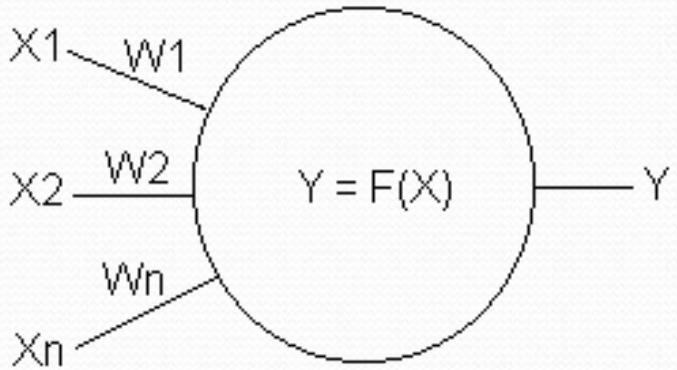
（15）

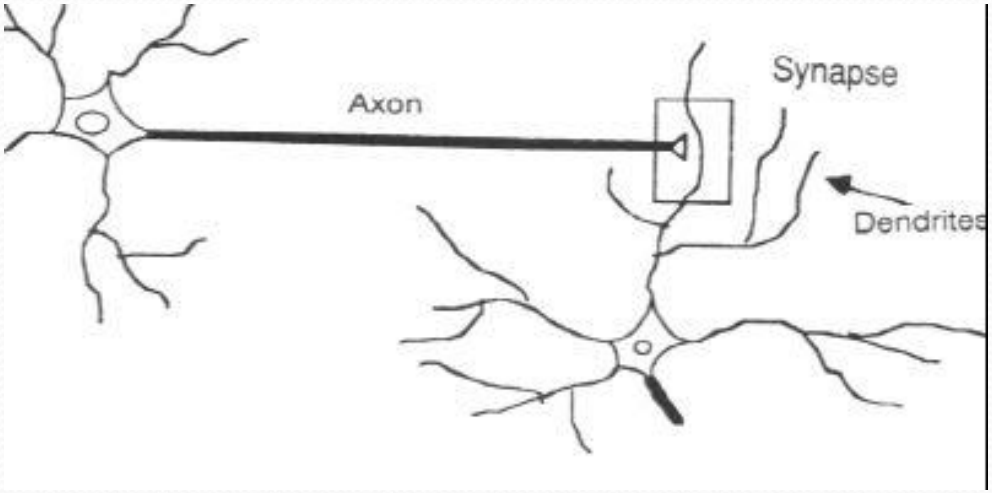
感知器与人工神经网络方法

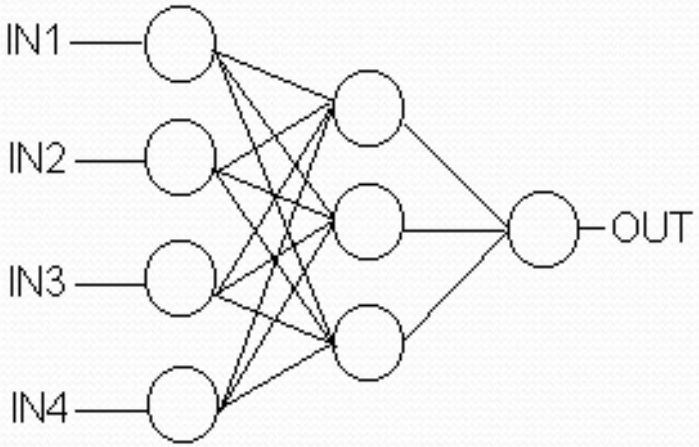
(Perceptron & ANN method)

——计算机人工神经网络是对大脑神经网络的模拟，在生物信息学研究中，无论是基因识别还是蛋白质结构预测，神经网络都取得了比其它方法更为准确的结果。


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


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


Axon
Synapse
Dendrites


IN1
IN2
IN3
IN4
OUT

第十六式　龙战于野

（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都需要用到模式识别。

第十七式　履霜冰至

（17）

微分方程的数值方法

(Numerical methods)

——分子动力学模拟：研究生物大分子的构象，主要还是用基于半经验势函数的分子动力学方法，而量子力学则在确定势函数的参数和研究局部性质时起作用。对蛋白质进行动力学研究是利用计算机进行模拟实验的基础。

——分子动力学得到一组动力学微分方程，要求得到初值问题的解。

——微分方程的数值求解：有限差分法、有限元法

（18）

最终要诀：各类方法综合运用

All in one!

——综合运用不同的研究方法

——始终面向生物学问题

——知识和技能的学习方法

——文献的查阅和阅读方法

——中、英文论文的写作方法

生物信息学的“东邪西毒南帝北丐中顽童”

——东邪：de Bruijn图算法（基因组拼接）

——西毒：近似算法和似然估计（进化树分析）

——南帝：核函数（基因和物种分类问题）

——北丐：统计建模（序列分析）

——中顽童：深度学习和生成对抗神经网络GAN（判断问题）

Slides credits

——生物信息学研究方法概述：北京大学生物信息中心

——生物统计学：中国科学院计算技术研所所

谢谢！