**生物信息学重点整理**

**第二章 分子生物学数据库**

**1.分子生物信息数据库分类：**

基因组数据库

核酸和蛋白质一级结构序列数据库

生物大分子(主要是蛋白质)三维空间结构数据库

以上述三类数据库和文献资料为基础构建的 “二次数据库 ”

**2.一、二次数据库特点对比：**

一次数据库的数据量大，更新速度快，用户面广，通常需要高性能的计算机硬件、大容量的磁盘空间和专门的数据库管理系统支撑。

二次数据库的容量则要小得多，更新速度也不象一次数据库那样快，也可以不用大型商业数据库软件支撑。

**3.生物信息数据库的特征：**

（1）数据库的更新速度不断加,数据量呈指数增长趋势

（2）数据库使用频率增长更快

（3）数据库的复杂程度不断增加

（4）数据库网络化

（5）面向应用

（6）先进的软硬件配置

**4.各类型数据库代表：**

（1）核酸序列数据库 ：

GenBank 美国国家生物技术信息中心（数据格式：Genbank）

EMBL 欧洲分子生物学实验室

DDBJ 日本国立遗传学研究所

（2）蛋白质序列数据库：

PIR 美国生物医学基金会建立

SWISS-PROT 瑞士生物信息学研究所和欧洲分子生物学实验室共同维护

（3）蛋白质结构数据库：

PDB 美国Brookhaven实验室大分子数据库

**5.数据库的组成：**

**（1）表（建立数据库）**

表是数据库中最基本的对象 , 有表可以建立其他对象关系

表及其表之间的关系构成数据库的核心

**（2） 查询（查询调用数据库）**

查询是从一个或多个表( 或查询 )中选择一部分数据,将它们集中起来,形成一个全局性的集合,供用户查看

**（3）窗体**

窗体是用户与数据库交互的界面 , 是数据库维护的一种最灵活的方式

**窗体的数据源 可以是表 , 也可以是查询**

**（4）报表**

Access 中的报表与现实生活的报表是一样的 , 是一种按指定的样式格式化的数据形式

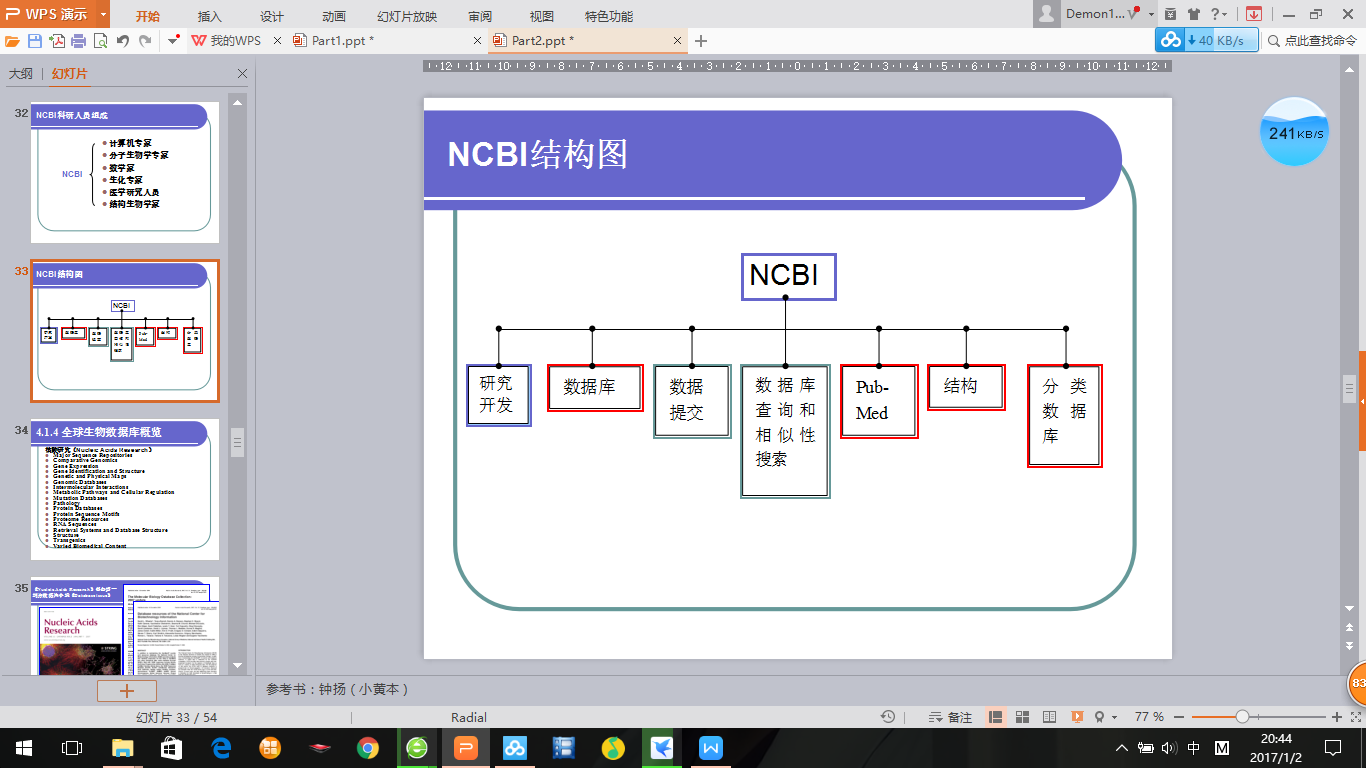
**（5）宏**

宏是若干个操作的组合 , 可用来简化一些经常性的操作

**（6）模块**

在模块中 , 用户可用 VBA（是VB自动化语言） 语言编写

**6.数据库结构：**



**第三章 序列比对**

**1.全局比对（global alignment）：**从全长出发，考察序列之间的整体相似性。（nw算法）

**2.局部比对（local alignment）**：针对于序列中的某些特殊片段，比较这些片段之间的相似性。（sw算法）

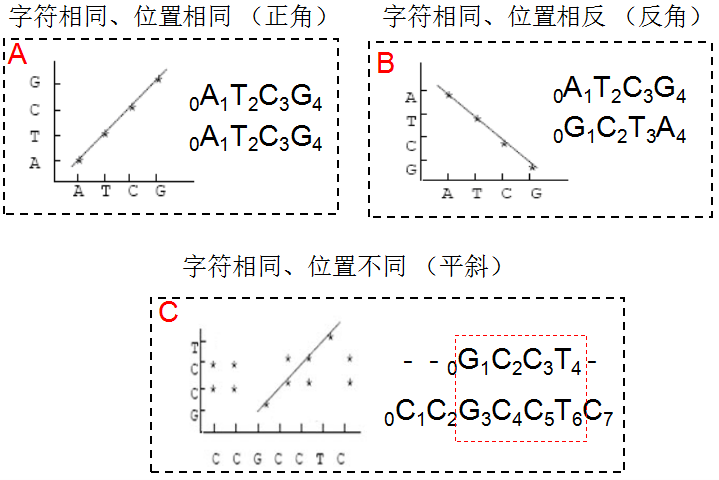
**3.序列比对定义:** 序列比对就是对序列进行基本编辑操作，通过字符匹配和替换，或者插入和删除字符，使比对序列中相同的字符尽可能地一一对应，并计算序列相似程度，搜索最优比对形式，该过程为序列比对。

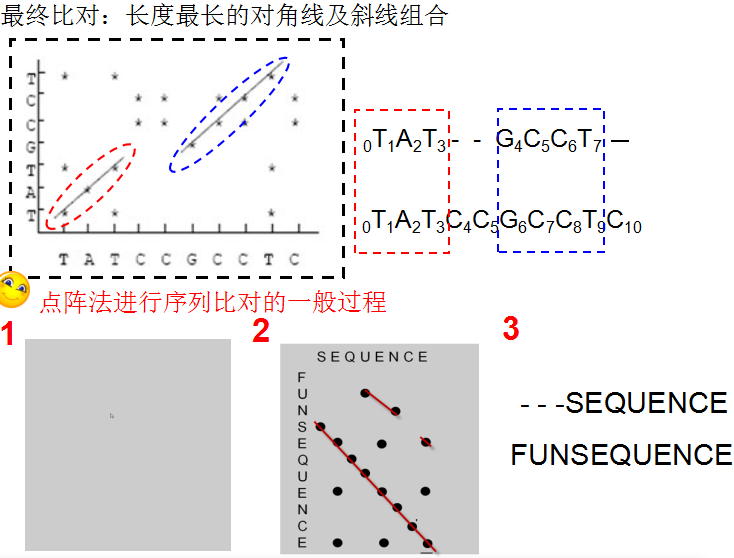
**4.直系同源(Orthologous)** :来自不同物种的同源序列，是随着物种形成而产生。

**5.旁系同源(Paralogous)** :来自同一物种的同源序列，是通过类似于基因复制机制而产生的同源序列。

**6.双序列比对方法：**

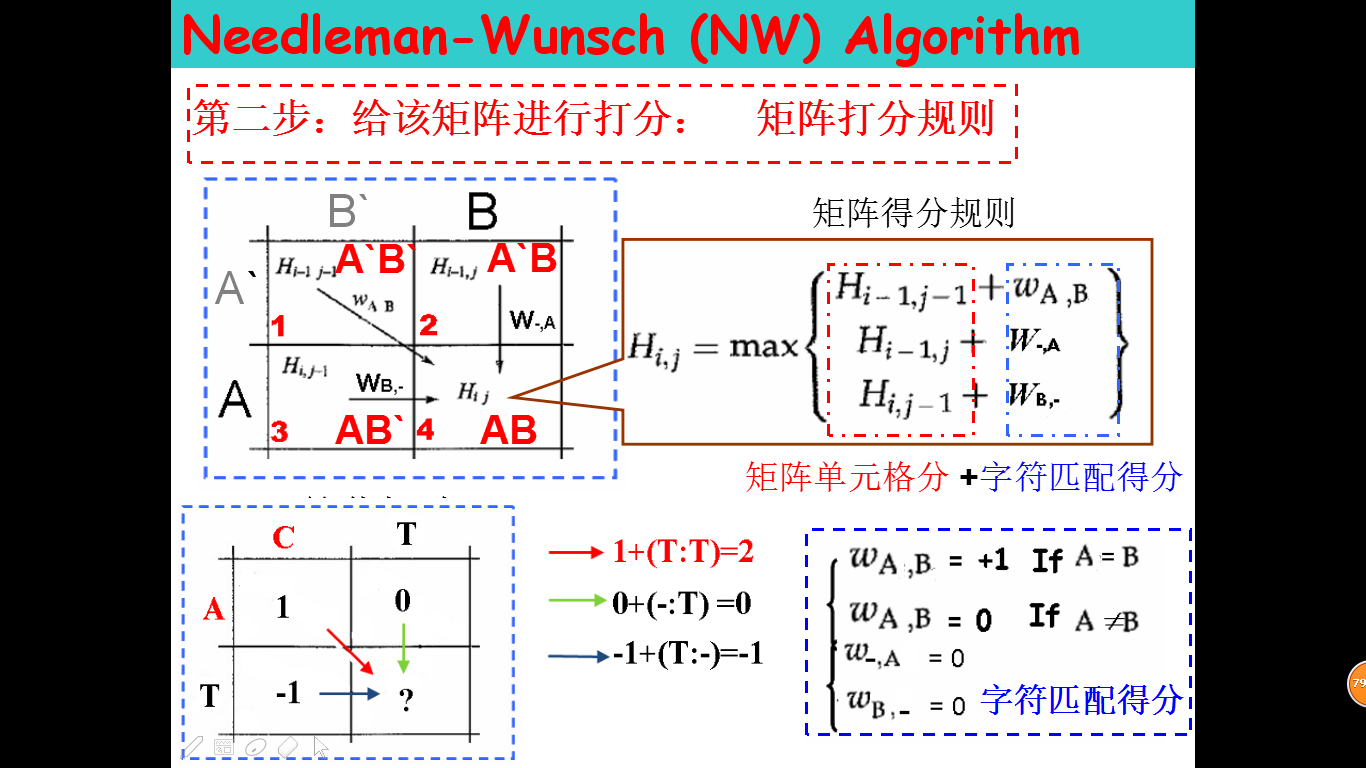
**第一种方法：点阵法 （对角线作图法）**

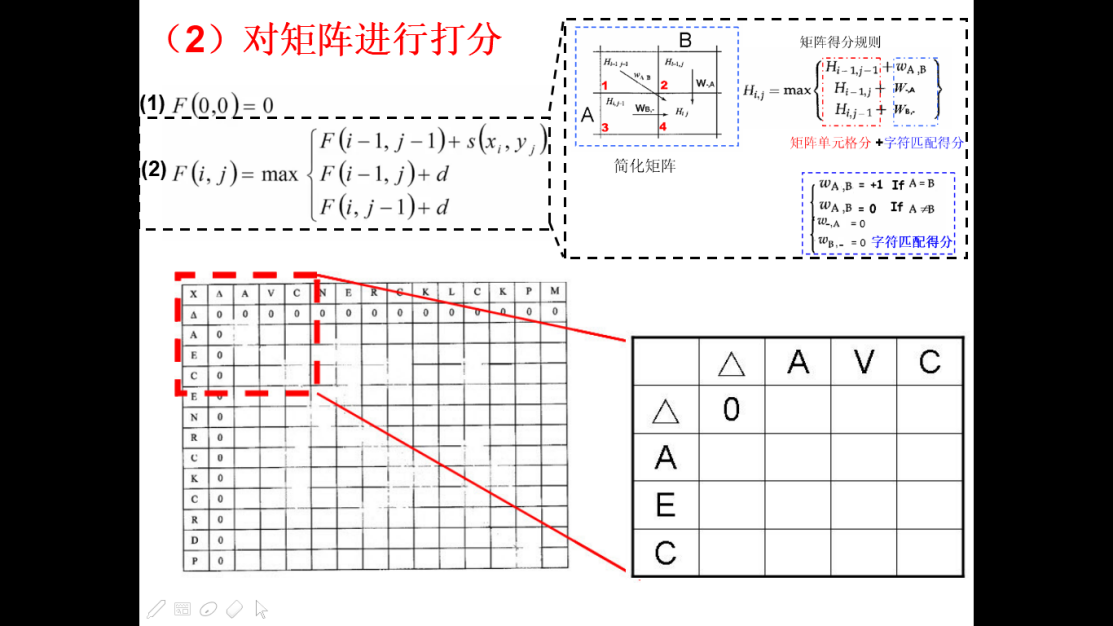


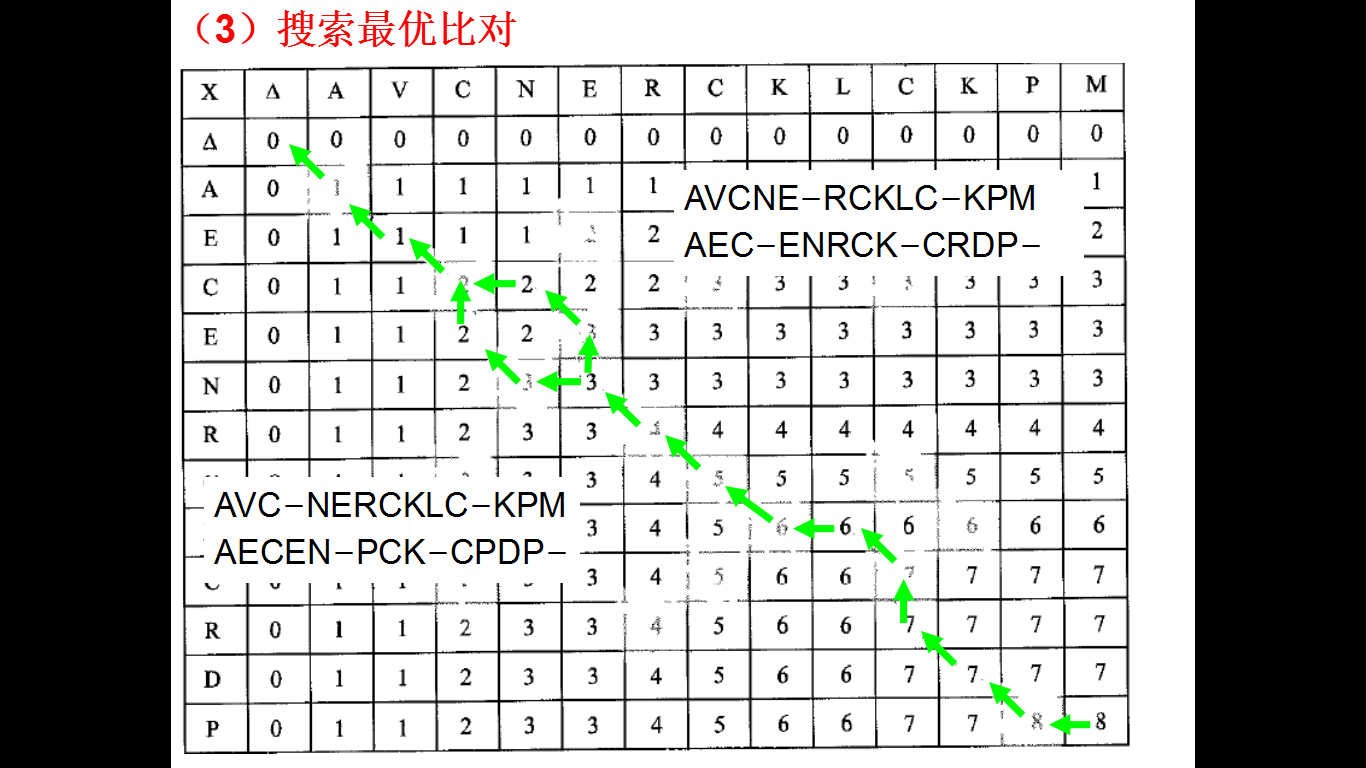


**第二种方法：动态规划算法**

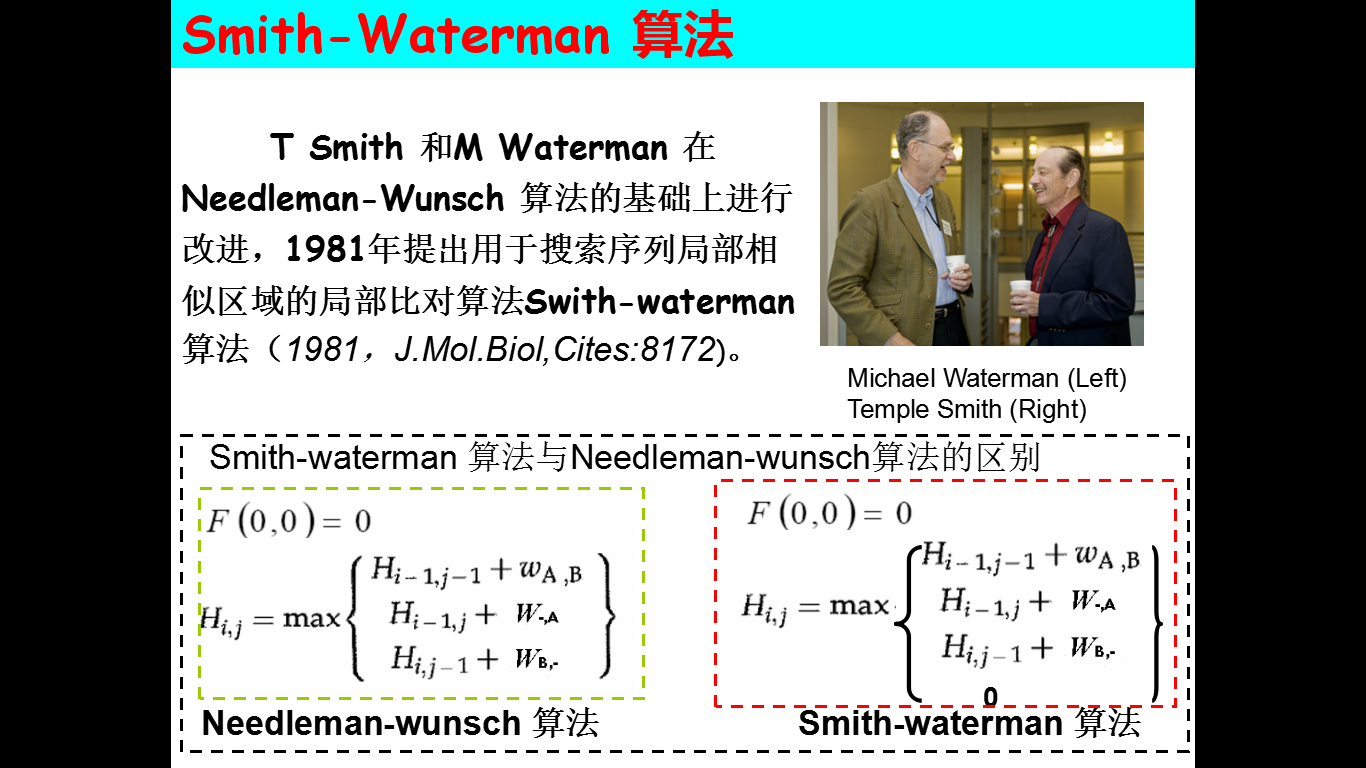
**Needleman-Wunsh 算法：**







**Swith-waterman算法：**



**第三种方法：启发式算法（K-元法/字串法)**

**7.打分矩阵：**

（1）DNA序列的打分矩阵：

①等价矩阵

②BLAST矩阵

③转换—颠换矩阵：转换（Transition):如果DNA碱基的变化（碱基替换）保持环

数不变,如A→G，C→T；

颠换 ( Transversion) :如果DNA碱基的变化后环数发生变化, 如A→C，A→T 等。

（2）蛋白质序列的打分矩阵：

①等价矩阵

②遗传密码矩阵GCM：通过计算一个氨基酸残基转变到另一个氨基酸残基所需的密码子变化数目而得到，矩阵元素的值对应于代价。

③疏水矩阵：根据氨基酸残基替换前后疏水性的变化而得到得分矩阵，若一次氨基酸替换疏水特性不发生太大的变化，则这种替换得分高，否则替换得分低。

**④PAM矩阵：**

基于氨基酸进化的可接受点突变模型，通过统计、计算相似序列中的两种不同氨基酸之间替换发生的频率评价两类氨基酸相似程度。

特点：对角矩阵

**同一个PAM矩阵，分值越大，两条序列的相似度越大**

**不同的PAM（n）矩阵，n值越大，两条序列的相似度越小**

PAM-1突变概率矩阵:

在一个PAM 时期内（蛋白质序列1%发生突变所需时间），任意两个氨基酸相互突变的频率

**⑤BLOSUM矩阵：**

通过统计相似蛋白质序列的替换率而得到的。

构建：从包含有500多个蛋白质家族的数据库中，通过多序列比对方法得到2000多个氨基酸保守模块，从中获得氨基酸替换频率

特点：对角矩阵

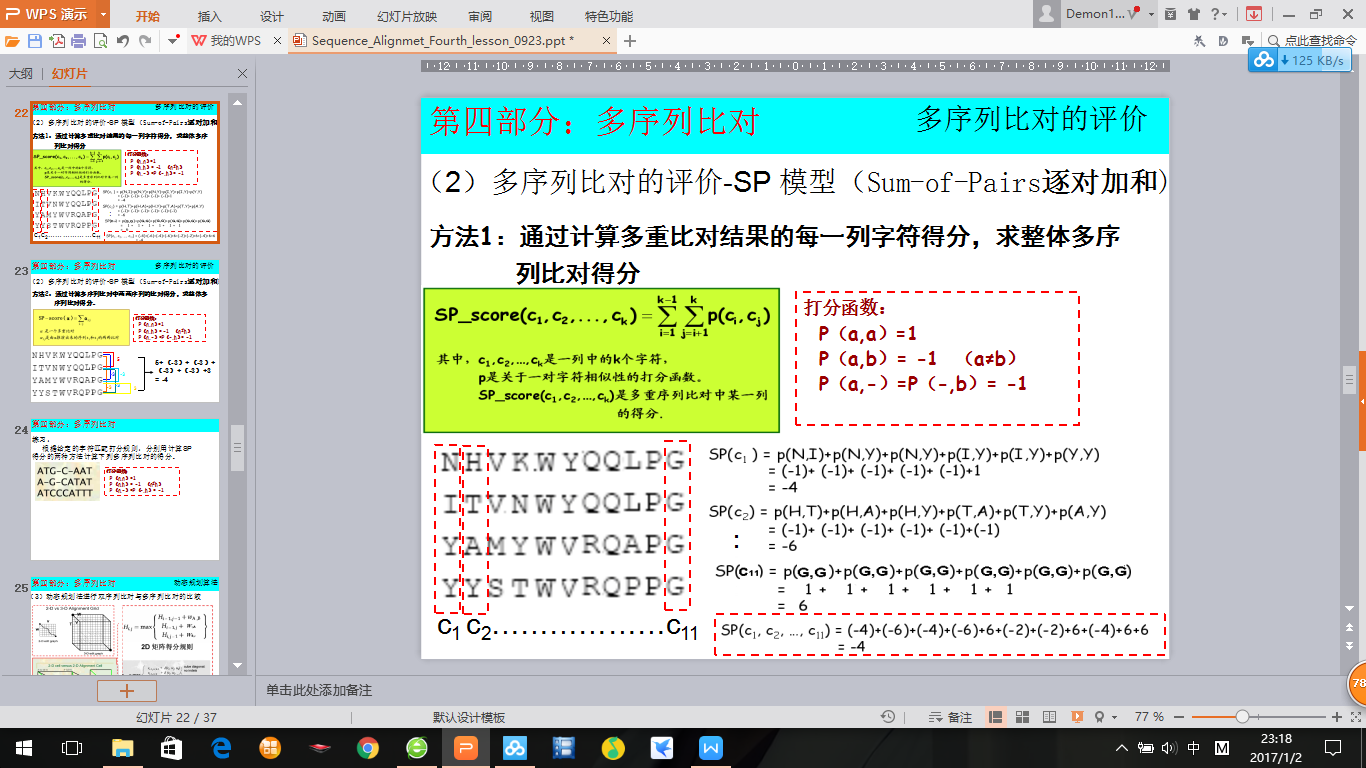
**同一个BLOSUM矩阵，分值越大，两条序列的相似度越大**

**不同的BLOSUM（n）矩阵，n值越大，两条序列的相似度越大**

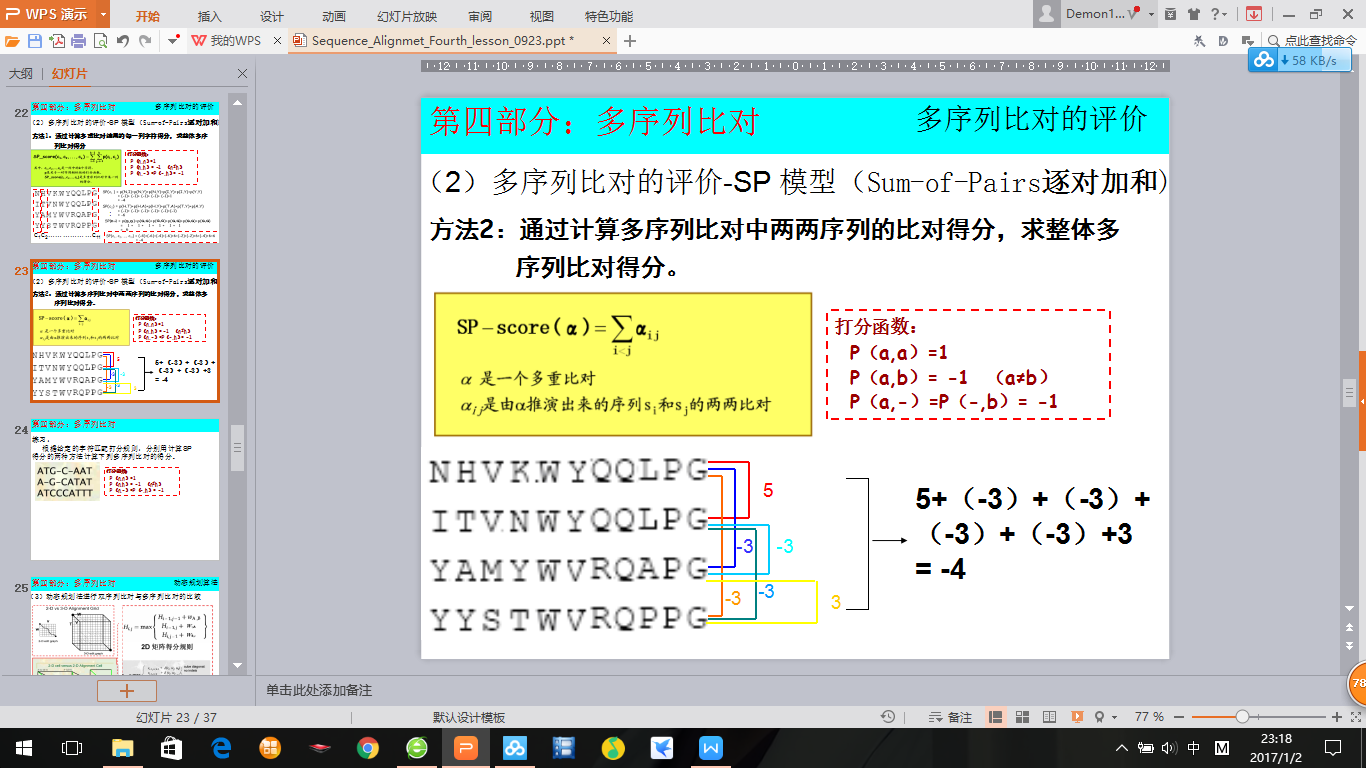
**8.多序列比对：**

SP 模型：

方法1：通过计算多重比对结果的每一列字符得分，求整体多序列比对得分。（先列后行）



方法2：通过计算多序列比对中两两序列的比对得分，求整体多序列比对得分（先行后列）



（1）动态规划法

两种策略：

（2）星形比对算法（渐进法）

基本思想：在给定的若干序列中，选择一个核心序列，通过该序列与其它序列的两两比对形成所有序列的多重比对，从而使得在核心序列和任何一个其它序列

方向的投影是最优的两两比对。

（3）CLUSTAL W算法：CLUSTAL 算法是一种渐进的比对方法。

**第四章核酸序列分析**

**1.DNA携带遗传信息：**

（1）一类信息储存于具有功能活性的DNA序列中，能够通过转录过程形成RNA（主要有编码RNA和非编码RNA），其中编码RNA含有编码蛋白质的氨基酸序列信息，这类DNA序列主要是指遗传的基本单位即功能序列。

（2）一类信息属于调控信息，主要存在于特定DNA的区域，能被各种功能性蛋白质分子特异性的识别结合，进而完成各种生物过程，例如启动子和增强子调控基因的表达。

**2.DNA序列的基本信息分析：**

序列长度

碱基比例

GC含量：计算公式：（G+C）/（A+T+C+G）×100%

序列转换：反向序列、互补序列、反向互补序列、DNA双链、转换为RNA序列

限制性核酸内切酶：是可以识别DNA 的特异序列、并在识别位点或其周围切割双链DNA 的一类内切酶。

特点：限制酶分为I、II、III型三大类。其中，II型限制酶能识别专一的、短的DNA序列，并在识别位点或附近切割双链DNA。具有专一的识别和切割位点，即回文结构。

切割末端：黏性末端、平滑末端

回文对称（palindrome）：特指DNA 的一种具有反向重复的结构。具有这种结构的DNA，其一条链从左向右读和另一条链从右向左读的序列是相同的。

**3.DNA序列的特征信息分析：**

**（1）开放阅读框（Open Reading Frame，ORF）：**在给定的阅读框架中，不包含终止密码子的一串序列，是生物个体的基因组中可能作为蛋白质编码序列的部分，包含从5’ 端翻译起始密码子(AUG) 到终止密码子(UAA、UAG、UGA) 之间的一段编码蛋白质的碱基序列。

**（2）编码序列（Coding DNA Sequence，CDS）：**经过实验证实的可以编码蛋白质的碱基序列。一条DNA序列可能存在六种相位（frame）。

**（3）CpG岛：**在基因组的某些区段，CpG 保持或高于正常概率，这些区段被称作CpG岛

识别依据：

①传统的CpG岛识别方法主要依据三个序列特征：GC含量、CpG岛长度、CpG二核苷酸出现的频率

CpG 岛长度：至少200bp

GC 含量：超过50%

CpG 的观察值与预测值的比率：高于60%

②升级版标准：CpG岛长度不少于500bp，GC含量达到55%，CpG二核苷酸出现率为65%

**4.重复序列（repetitive sequence, repeated sequence）：**真核生物基因组中重复出现的核苷酸序列，一般不编码多肽，在基因组内可成簇排布，也可散布于基因组。

**分类：**（按重复次数）

①低度重复序列（lowly repetitive sequence）：2 ～ 10 个拷贝

②中度重复序列（moderately repetitive sequence）：重复几十次到几千次，平均长300bp

③高度重复序列（highly repetitive sequence）：重复几百万次，少于10 个核苷酸残基组成的短片段（按组织形式）

①串联重复序列：卫星DNA：着丝粒

小卫星DNA：端粒

微卫星DNA：内含子

②散在重复序列：

短散在重复序列（SINE，500bp以下）：ALU数据库（人类及灵长类代表性Alu重复片段）

长散在重复序列（LINE，1000bp以上）：LINE—1数据库

长末端重复（LTR）转座子

**工具及特点：**

RepeatMasker：识别、分类和屏蔽重复序列

Cross\_match：速度慢、精度高

ABBlast：速度快、精度略低

RMBlast：NCBI Blast 的兼容版

HMMER：只适用于人类基因组序列

**5.基因（gene）：**产生一条多肽链或功能 RNA 所需的全部核苷酸序列。一段具有特定功能和结构的连续的DNA片段，携带着遗传信息，是编码蛋白质或RNA分子遗传信息、控制性状的基本遗传单位。

**6.基因识别：（gene prediction，gene finding）：**使用生物学实验或计算机等手段识别DNA 序列上的具有生物学特征的片段，是基因组研究的基础。对象主要是蛋白质编码基因（还有RNA 基因和调控因子等）。

**7.基因识别的方法：（简/问）**

①间接识别法（Extrinsic Approach）：利用已知的mRNA 或蛋白质序列为线索在DNA 序列中搜寻所对应的片段

②从头计算法（Ab Initio Approach）：基因预测（通常仍需实验证实）。

基于基因的两种类型的特征：

“信号”：由一些特殊的序列构成，通常预示着周围存在着一个基因。

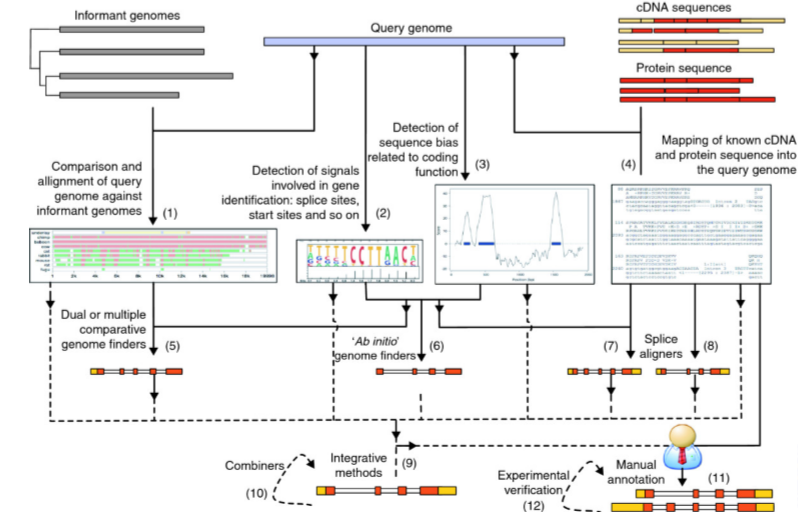
启动子，剪接供体和受体位点，起始和终止密码子，polyA 位点

“内容”：蛋白质编码基因所具有的某些统计学特征

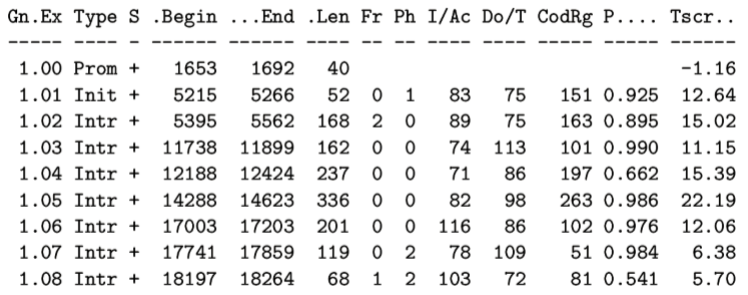
密码子使用偏好性，双联密码子出现频率，基因组等值区

③比较基因组学的方法：自然选择的力量使得基因和DNA 序列上具有生物学功能的片段较其他部分有较慢的变异速率，在前者的变异更有可能对生物体的生存产生负面影响，因而难以得到保存

1. **基因识别策略：**



**9.GENSCAN结果解读：**



Init: initial exon; Intr: internal exon; Term: terminal exon Sngl: single-exon gene;

Prom: promoter region; PlyA: polyA signal

**10.选择性剪接（alternative splicing）：**又称可变剪接，指用不同的剪接方式（选择不同的剪接位点组合）从一个mRNA 前体产生不同的mRNA 剪接异构体的过程。

**11.可变剪接种类：**

（1）外显子跳跃：外显子加入或转移，是可变剪接最常见的一种方式

（2）3‘端可变剪接：基因的3‘端外显子被有选择的延长或缩短

（3）5‘端可变剪接：基因的5‘端外显子被有选择的延长或缩短

1. 外显子互斥：两个相邻的外显子只能有其中一个外显子被包含在剪接产物中

（5）内含子保留：在剪接过程中有的内含子被保留下来，起到了外显子的作用

**12.microRNA：**是真核生物中广泛存在的一种长约20 到24 个核苷酸的内源性非编码单链RNA 分子。miRNA 通过RNA 诱导沉默复合体（RISC）与靶基因的3’非翻译区（3’UTR）相结合，导致靶基因mRNA降解或者抑制其翻译，从而调节基因转录后的表达。

数据库：miRBase、TarBase、miRGen

作用方式：

①miRNA与靶mRNA完全互补致靶基因mRNA降解

②不完全互补致靶基因mRNA翻译抑制，而不影响其稳定性

**miRNA预测方法：**

①同源片段搜索方法

②基于比较基因组学的预测方法

③基于序列和结构特征打分的预测方法

④结合作用靶标的预测方法

⑤基于机器学习的预测方法

MiRscan、MiPred、miRFinder

**miRNA靶基因预测方法：**

①基于种子区域互补和保守性的规则预测：miRanda、 TargetScan

②基于机器学习方法训练参数进行靶基因预测：PicTar、 miTarget

**第五章 基因组功能注释分析**

**1.基因组的组装版本：**

人类基因组的UCSC hg19——NCBI Genome Reference Consortium GRCh37

小鼠基因组的UCSC mm10——NCBI Genome Reference Consortium GRCm38

hg19：人类基因组的UCSC组装版本

**2.基因组的坐标系统：**

①0-based坐标系统：从0开始标记——【start,end】

②1-based坐标系统：从1开始标记——【start,end】

例如：



0-based，half-open：[3,5)

1-based，fully-closed：[4,5]

1. **基因组注释常用格式：FASTA、BED、GFF、VCF**

（1）FASTA格式：

起始标识符为“>”，后面紧跟序列的ID以及可有可无的描述信息。

下面一行或数行是具体的序列。

（2）BED格式：

BED6 ：chrom, start, end, name, score, and strand

第一列：chrom——染色体名

第二列：chromstart——特征的起始位置

第三列：chromend——特征的终止位置

第四列：name——特征名

第五列：score——分值（0~1000）

第六列：strand——链性（+代表正链，-代表负链）

**4.基因组坐标的逻辑运算模式：**

（1）交集（intersect）：提取出两组基因组特征坐标中完全重叠的坐标位置或有重叠的基因组特征。

（2）减法（subtract）：除去完全重叠的坐标位置，或去除有重叠的基因组特征而只保留完全没有重叠的特征（A—B与B—A不同）。

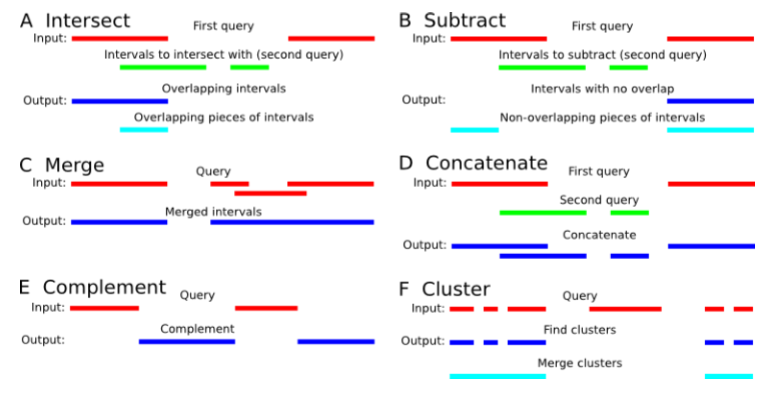
（3）合并（merge）：把多个有重叠的坐标位置或基因组特征合并成一个大的坐标或特征

（4）串联（concatenate）：简单地把两个坐标合并起来，不进行其他操作

（5）补集（complement）：依据基因组坐标全集，对当前的这组坐标或特征取补集

（6）聚类（cluster）：根据设定的最小坐标间隔以及聚类需要的最小记录数目，将所有符合要求的坐标聚合成一个坐标，或把所有符合要求的特征聚合成一个大的特征

（7）联合（join）：比较两组坐标或特征，根据坐标的重叠情况，把两组坐标或特征中相应的记录对应起来



**5.富集分析：**

（1）数据库：

①GO （Gene Ontology）：

biological process，BP，生物学过程

molecular function，MF，分子功能

cellular component，CC，细胞组份

两大关系：is\_a、part\_of

②KEGG （Kyoto Encyclopedia of Genes and Genomes）

（2）分析工具：DAVID （Database for Annotation, Visualization and Integrated Discovery）

其中Functional Annotation Chart：富集分析

**6.序列标识（sequence logo）：**是基于DNA、RNA和蛋白质的多序列比对信息，把多序列的保守性信息通过图形表示出来。

数据：多序列比对信息

横轴：序列坐标位置

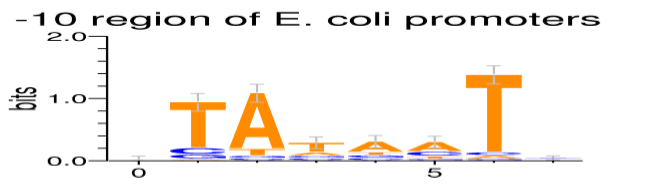
纵轴：比特，计量单位

总高度：信息量/保守性

相对高度：相对频率

位置自上而下：频率由大到小

例子：



1. **蛋白质序列信息分析**
2. **Uniprot数据库的组成：**Swiss-prot、TrEMBL、PIR-PSD
3. **蛋白质序列分析层次：**

**（1）基本信息分析**

①氨基酸组成分析

AACompIdent：通过将未知蛋白质的氨基酸组成百分比与Swiss—TrEMBL数据库中的蛋白质理论氨基酸组成百分比进行匹配

②理化性质分析

ProtParam：可以计算蛋白质的理化性质包括分子量、理论等电点、氨基酸组成、原子组成、消光系数、体内半衰期、不稳定系数、脂肪指数、总平均亲水性

③亲疏水性分析

ProtScale：分值大于0表示疏水，小于0表示亲水

**（2）特征信息分析**

①跨膜区分析

Tmpred：基于统计学，通过权重矩阵打分对蛋白质跨膜区进行预测、分析

TopPred：基于氨基酸疏水性进行预测

PHDhtm：利用机器学习方法进行预测

②信号肽分析

SignalP：基于蛋白质序列预测信号肽的有无及位置的工具

Predisi：基于序列比对

③卷曲螺旋分析

COILS：利用滑动窗口法基于氨基酸在卷曲螺旋不同位置上出现的频率进行预测

④糖基化位点分析：NetOGlyc 4.0

⑤磷酸化位点分析：NetPhos

基于序列对于蛋白质进行理化性质分析：

Compute pI/Mw

ProtParam

PeptideMass

PeptideCutter

ProtScale

**（3）功能信息分析**

1、基于蛋白质基序

PROSITE：基于Motif 进行功能分析的网站

2、蛋白质的结构域和功能位点

InterPro：基于Domain 进行功能分析的网站

3、蛋白质的同源性

blastp、FASTA

1. **蛋白质结构分析**
2. **测定蛋白质结构的实验方法：**

X-线晶体衍射法、核磁共振光谱法 (NMR)、冷冻电子显微镜法

1. **数据库：**

**（1）SCOP数据库：**对已知结构蛋白质的结构域进行分类的数据库，从不同层次对蛋白质进行分类，以反映不同蛋白质结构和进化关系之间的相关性。

**（2）CATH**：一个蛋白质结构域分类数据库，从不同层次上对蛋白质结构域进行分类，该数据库的分类基础是蛋白质结构域。

**（3）PDB：**存储运用实验方法测定的生物大分子的三维结构信息。

PDB数据库中每个分子结构具有的检索号，通常由大写字母和数字0-9组成的四个字符（eg：2HI4、2BEG）

**3.二级结构预测**

**（1）Chou—Fasman方法（第一代）：**

每种氨基酸残基出现在蛋白质二级结构中的倾向或频率是不同，具有一定规律。依据每种氨基酸残基形成不同二级结构的倾向性进行预测。

**①构象倾向因子计算：**



**②二级结构预测准则：**

<1> α-螺旋准则：

Ⅰ.寻找螺旋的起点（螺旋核）

规则:“4/6 氨基酸残基中螺旋构象倾向因子大于 1.0”

Ⅱ.确定螺旋的终点

规则:“连续4个残基的螺旋构象倾向因子小于1.0, 或出现Pro”

Ⅲ.确定螺旋片段

规则：序列的长度大于6，同时该序列的螺旋构象倾向因子的平均值大于1.03 并且内部不含有Pro。

<2> β-折叠准则：

Ⅰ.寻找折叠的起点（折叠核）

规则：“4/6 氨基酸残基中折叠构象倾向因子大于 1.0”

Ⅱ.确定折叠的终点

规则：“连续4个残基的折叠构象倾向因子小于1.0”

Ⅲ.确定折叠片段

规则：序列片段的螺旋构象倾向因子的平均值大于1.05

<3> β-转角准则：

Ⅰ.通过形成转角结构的概率寻找可能的四肽片段

Ⅱ.四肽片段中形成转角结构的构象倾向因子的平均值大于1.0, 同时也大于形成螺旋和

折叠的倾向因子的均值

<4> 重叠准则：

Ⅰ.给序列中每个氨基酸指定不同结构的构象倾向因子

Ⅱ.根据氨基酸α-螺旋的倾向因子的值和α-螺旋的判定规则预测α-螺旋

Ⅲ.根据氨基酸的β-折叠倾向因子的值和β-折叠的判定规则预测β-折叠

Ⅳ.根据氨基酸的β-转角倾向因子的值和β-转角的判定规则预测β-转角

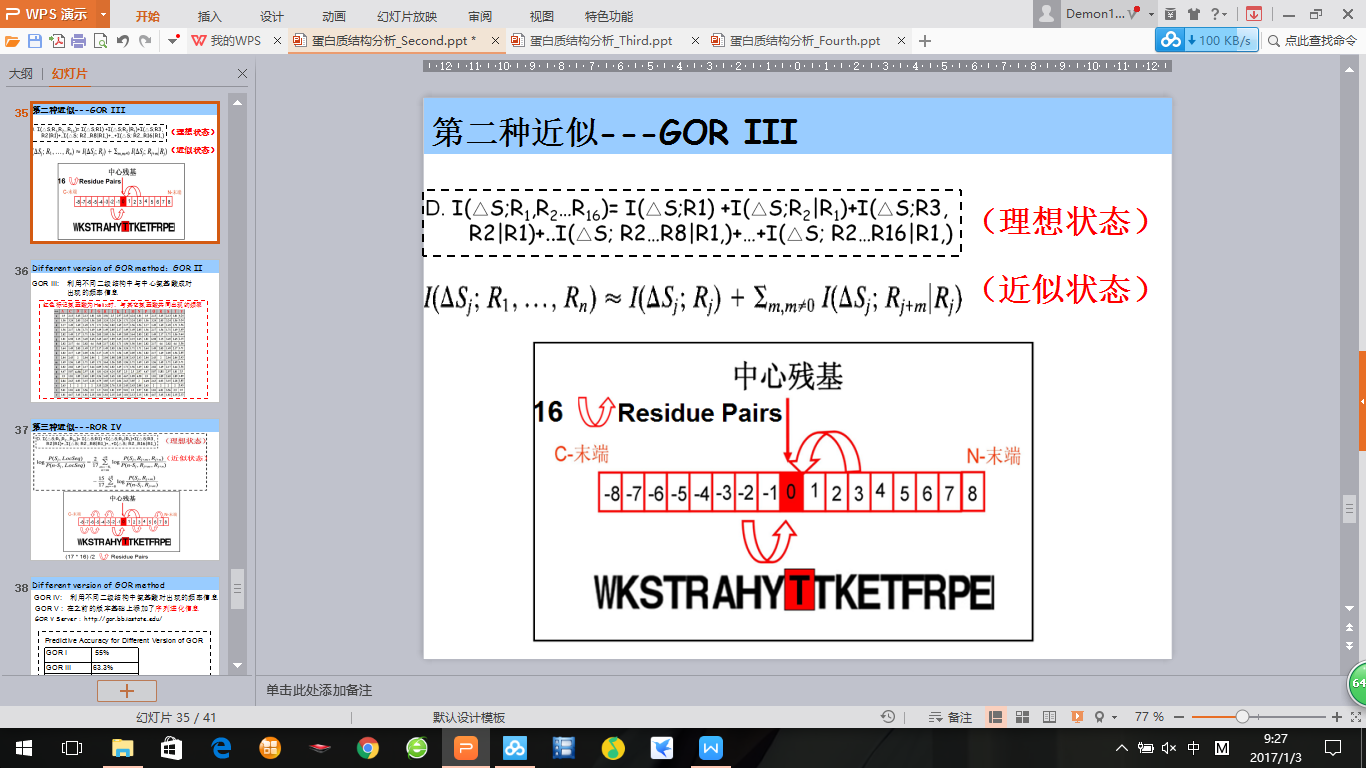
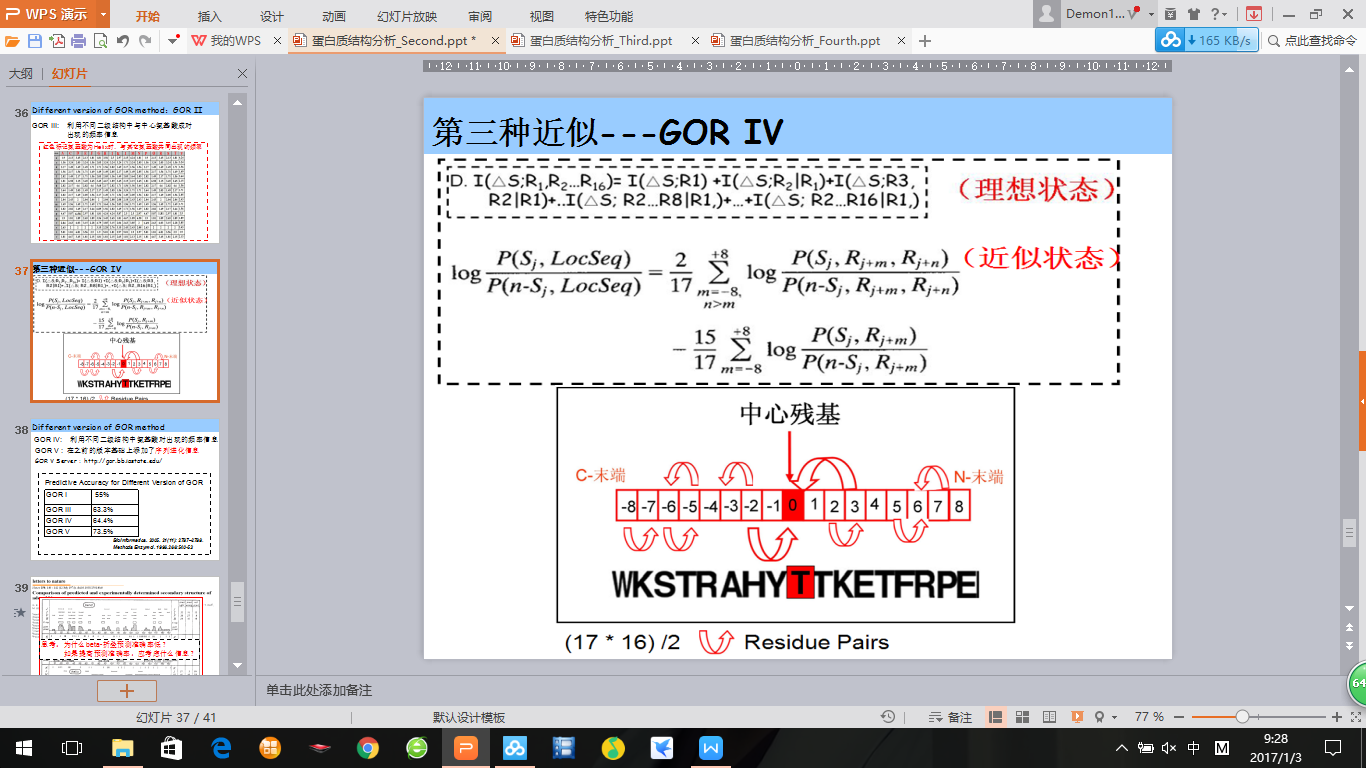
1. **GOR方法（第二代）：**

邻近氨基酸之间的相互作用对于形成二级结构会产生影响,因此要从某个氨基酸残基所处的局部环境对于可能形成的二级结构形式进行预测。

GOR = 信息熵 + 贝叶斯

**四代GOR方法区别：**

GOR II与GOR I相比，统计的数据增多267 Proteins

GOR IV:利用不同二级结构中氨基酸对出现的频率信息

GOR V:在之前的版本基础上添加了序列进化信息

**（3）PHD方法（第三代）：**

通过将三个简单的人工神经网络组合成一个复杂网络，并结合蛋白质进化信息对于蛋白质二级结构进行预测，其中每个简单网络模型的输入、输出信息都有所不同。

①第一个网络 Sequence-to-Structure：

Input（13 个氨基酸组成的窗口）

Output: 中间氨基酸形成三种二级结构的概率

②第二个网络 Structure-to-Structure：

Input（17 个氨基酸网络的输出）

Output（中间氨基酸形成三种二级结构形式的概率）

③第三个网络Jury Decision

Input（不同参数的第二层网络的输出）

Output（氨基酸形成二级结构的预测最终形式）

PHD方法的预测流程：（训练模型+预测）

第一步：BLAST在SWISS-PROT数据库中搜索同源序列

第二步：运用MaxHom程序进行同源序列比对

第三步：根据比对获得蛋白质局部及全局序列信息

第四步：运用已经训练好的神经网络进行结构预测

1. **NSSP预测方法（第四代）：**

运用最小邻近法根据蛋白质的序列相似性对蛋白质二级结构进行预测

利用最小邻近法预测蛋白质二级结构的流程：

①运用滑动窗口法，将待预测蛋白序列以19个氨基酸长度进行分解

②运用特异序列打分矩阵对窗口中的序列在数据库中进行序列相似搜索

③寻找、统计50条与窗口序列相似的序列片段中处于中心位置氨基酸形成不同二级结构形式的频率

（5）多元结构预测方法

**4.三级结构预测**

**（1）同源建模方法（homology modeling）**

如果两个蛋白质序列相似，则它们的结构也相似。

适用条件：如果能搜索到同源性较高的模板蛋白（同源性）30%），使用同源建模方法基于模板蛋白预测目标蛋白的结构

同源建模法流程（六大步）：

找模板、定保守、建骨架、连一起、补侧链、量分选

①应用序列比对方法(BLAST)在PDB数据库中寻找与目标蛋白序列相似的蛋白质作为模

板蛋白，并获得相应的蛋白质结构文件

选取模板蛋白的原则：

（A）模板蛋白与目标蛋白的序列相似度要高 (序列）

（B）选择高分辨率蛋白结构作为模板蛋白 （结构）

（C）尽量选取与目标蛋白功能相似的模板蛋白 （功能）

②构建出包含序列信息和二级结构信息的序列—结构比对，确定目标蛋白与模板蛋白之间的保守区域（序列相似度高且无空位，同时位于Helix和Sheet结构）

保守区域的特征：

Ⅰ.蛋白质序列相同或相近

Ⅱ.位于蛋白质二级结构上

Ⅲ.保守区域之间是不连续的

③构建目标蛋白中保守区域的蛋白质骨架结构

刚体组装法

片段匹配法

空间限制法

仿真进化法

④构建保守区域之外（非保守区）的其它区域的蛋白质骨架结构

数据库搜索法：选参考原子 — 测原子间距离 — 找相似片段

系统构象搜索法：随机产生构象 — 准确评价构象 — 择优选择构象

⑤构建目标蛋白中所有残基侧链原子的结构

⑥评价目标蛋白的预测结构，选择最优结构作为最终预测结果。

**（2）串线法（threading）**

通过确定蛋白质氨基酸序列和蛋白质结构之间的匹配方式来预测目标蛋白的空间结构

适用条件：如果能搜索到同源性较差的模板蛋白（同源性10%~30%），使用串线法预测目标蛋白的结构

**（3）从头预测法（ab initio）**

不依赖其它蛋白质结构而仅仅基于蛋白质的序列信息通过蛋白质折叠规律来预测蛋白质可能的三级结构。

适用条件：如果连低同源性的模板蛋白（同源性<10%）也搜索不到，使用从头预测法预测目标蛋白的结构

**第八章 生物信息学与人类复杂疾病**

**1.单基因疾病(Monogenic diseases)：**

也称孟德尔式遗传疾病，遗传因素在发病过程中占主导地位，而且主要与单个基因的功能或表达水平上的异常有关。

例子：血友病、白化病。

特点：

（1）种类多，目前已经确认的单基因疾病大约有1万种

（2）发病率低，通常为1/10000或更低，因而这类疾病被称为稀有疾病

**2.复杂疾病(Complex Disease)：**

绝大多数疾病的发生与遗传、环境、生活方式和年龄等多种因素有关, 因而被称为复杂疾病。

例子：肿瘤、心脑血管病、代谢系统疾病、神经系统疾病、精神和行为异常等

特点：

（1）每个基因或其他因素的影响通常比较小，单独存在时一般不足以致病，而当多种因素叠加在一起协同作用时，疾病发生的风险就显著增加。

（2）在普通人群中发病率较高(一般不少于1%)，所以也叫“常见疾病”

（3）是遗传易感性、生理状态和环境等多种因素交互作用的结果，患病风险与各因素之间存在不同程度和不同层次的关联，因此也导致相应的分子机理十分复杂。

**3.复杂疾病的遗传特性：**

（1）不符合孟德尔规律，而且疾病基因型与表型之间存在多因素致病、多基因多层次调控以及临床表型复杂等特征

（2）遗传易感性不一定是对疾病表型本身的直接影响，而可能是通过影响疾病的中间性状的间接后果。

**4.复杂疾病的影响因素：**

（1）遗传因素：遗传风险与多个基因的变异有关

（2）环境因素：

①直接或间接地影响基因表达或其他分子的功能及活性

②环境中的放射性元素、紫外线和重金属等可能使DNA分子发生变异，使得某些基因的表达异常

③直接影响某些组织或器官中特定基因或蛋白质的功能，进而调节和影响其他组织和器官中基因的表达。

④环境因素与基因之间的交互作用

⑤基因的个体差异使环境对人体的影响不同

⑥具有相同的基因型的个体比如同卵双胞胎在不同的环境中可能发展出不同的表型

（3）年龄和个体的生活方式：通过与基因之间的交互作用影响疾病的发生和发展

**5.研究复杂疾病发生的分子机理方法：**

（1）多基因和蛋白质参与的人类疾病

人类孟德尔遗传数据库OMIM

癌症基因数据库CGAP

（2）表观遗传与人类疾病

DNA修饰、蛋白修饰

（3）单核苷酸多态性与人类疾病

单核苷酸多态性数据库dbSNP

基因型和表型数据库dbGaP

（4）非编码RNA与基因调控

rRNA，tRNA，snRNA，snoRNA 和microRNA

**6.复杂疾病分子机理分析的技术：**

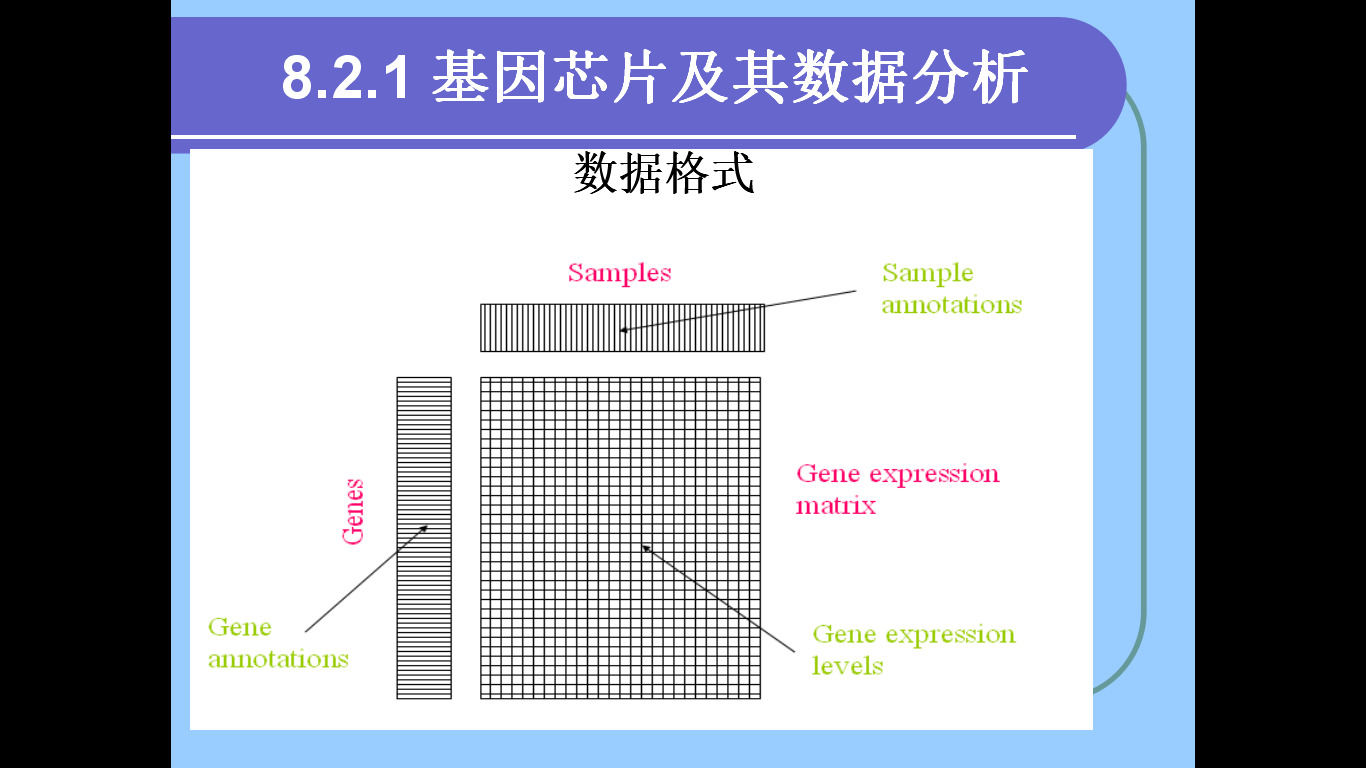
基因芯片、蛋白质组

1. **基因芯片**

**原理:基于Northern杂交原理，即以核酸分子的碱基互补配对原则，用序列已知的DNA片段特异性的测定样品中的mRNA分子的种类及丰度。**制备芯片时，作为探针的DNA片段被预先固定在固体基质；芯片使用时，待测样品中的mRNA分子被提取出来并被反转录为用放射性同位素或荧光染料标记的cDNA；随后探针与标记的靶基因进行杂交，并通过放射自显影或荧光检测等技术测定杂交的强度，进而获得样品中mRNA分子种类及浓度等信息。

**类型**：点样芯片、原位合成芯片、光纤微珠芯片

**数据特点**：通过不同平台获取的基因表达数据，其数据格式和测量精度都是有差异的，因此需要前期的数据预处理才能进行后续分析，预处理包括基因表达数据的提取、数据过滤、数据标准化等。



如何分析：

①对基因进行聚类

识别功能相关的基因

识别基因共表达模式

②对样本进行聚类

质量控制

检查样本是否按已知

类别分组发现亚型