**1算法和模型的相通性**

* **算法是为了解决一个特定问题而需要执行的一系列步骤或指令。**
* **模型是为了研究一个特定系统，通过抽象和简化而建立的代表这个系统的一种模式。**

**～～都需要输入数据，且都能够输出数据。**

**算法是根据输入数据得出输出数据——问题的解。**

**模型是根据输入数据得出输出数据，从而反映其所代表系统的特性。**

**2 Why use Unix？**

* **稳定性好: Over 25 years in industry and academia.**
* **开放性好：Supporting possible tasks in future.**
* **Internet上的操作系统：The software that powers the Web was invented in Unix, and many if not most web servers runs on Unix servers.**
* **科学软件的载体：Many good-quality, interesting and important scientific software are written for Unix.**
* **共享的乐园：Many programs can be downloaded and installed on Unix systems for free.**

**Linux**

* **Linux is a free, open source version of Unix.**
* **Linux can turn an ordinary PC into a powerful workstation.**

**Under Linux, inexpensive PCs regarded as “obsolete” by Windows user become startlingly flexible and useful workstations.**

* **Linux is an excellent platform for developing software.**

**3 计算机语言**

**编程灵活度上： （越灵活越容易编，但越易出错）**

**Perl > Java**

**C > C++**

**执行效率上：**

**C > C++ > Java > Perl**

**4 DNA序列比对与蛋白质序列比对**

**由于遗传密码的简并性，蛋白质序列比DNA序列更加具有同一性。**

* **一方面，用蛋白质序列进行序列比对的灵敏度高于用DNA序列进行序列比对的灵敏度，从而有利于寻找和联系亲缘关系较远的序列；**
* **另一方面，仅仅进行蛋白质序列比对可能丢失与进化过程直接有关的一些信息。**

**5 cDNA 文库**

* **提取出组织细胞的全部mRNA，在体外反转录成cDNA，与适当的载体（如噬菌体或质粒载体）连接后转化受体菌，则每个细菌含有一段cDNA，并能繁殖扩增，这样包含着细胞全部mRNA信息的cDNA克隆集合称为该组织细胞的cDNA文库。**
* **一个cDNA文库中的某两个克隆，可能来源于同一种mRNA，也可能不是；可能是全长，但一般不是全长。**

**6 获得cDNA全长的步骤**

**（1）获得全长cDNA序列及其信息（1）**

* **获得序列：实验操作往往很复杂，比如先通过特殊的限制性方法构建全长cDNA文库，再（根据已知的区段信息）筛选对应此种cDNA的特定克隆；**
* **获得序列信息：由于测序技术的限制（一次几百bp），一般很难直接测出全长序列，所以可能需要打断，测序，再拼接片段信息。**

**（2）获得全长cDNA序列及其信息（2）**

* **获得序列信息：通过构建普通的cDNA文库，进行高通量自动测序，我们可以得到大量的序列片断信息。运用序列自动拼接软件工具，拼接这些片断信息，可以同时推测多种全长cDNA序列信息。**
* **获得序列：之后可以根据全长信息在实验室中获得实际序列（如设计“全长”引物，PCR 筛选普通cDNA文库） 。**

**（3）cDNA克隆单次测序产生一个EST**

**通过自动测序仪对一个cDNA克隆单次测序很难产生整个克隆序列信息，往往只能产生一个片断序列信息，称为一个EST。也可以说，应用自动测序仪对一个cDNA克隆的一种“读法”产生一个EST。**

**这种读法可以从5‘端进行，也可以从3’端进行。**

**（4）从cDNA文库测序产生大量EST**

**（5）用EST拼接得到全长cDNA序列信息**

**运用序列自动拼接工具，进行EST序列信息拼接，可以同时获得多种全长cDNA序列信息**

**7 EST（ Expressed Sequence Tag）的意义**

* **它们是从cDNA测序产生的短序列信息，根据两端有重叠序列的EST可以组装获得全长的cDNA序列信息；**
* **它们代表在特定组织或发育阶段表达的基因，其相关分析是一种发现新基因和定位基因的有效方法。**
* **EST分析**

**尽管EST本身是不完整的甚至可能是不精确的DNA序列，但EST分析将为确定全长cDNA、寻找新基因和定位基因提供有价值的线索**

**8 EST的特性**

* **EST序列中除了A、G、T、C外，可能出现模糊碱基（如未知碱基N）；**
* **EST序列可能出现错误,其中插入或缺失将导致翻译时读码框移位（frame-shifts）；**

**错误率：genome 1/10kb，EST 1/100.**

* **在数据库中，EST数据可能是高度冗余的——交叉覆盖，甚至一个EST序列可能是另一个EST序列的一个片断。**

**9 基因组信息的三个层次**

* **染色体基因组，或简称基因组，即生物体内所有细胞中的遗传信息。－>DNA。**
* **表达基因组，或称转录基因组，即细胞某个特定生长阶段中基因组的表达部分。－>mRNA。**
* **蛋白质组，反映细胞特性和功能的所有蛋白质分子。 －>蛋白质。**

**10 序列分歧度---（sequence divergence）K是一种相异性指数。**

* **DNA序列分歧度：**

**设两个DNA序列的碱基差异值为N，序列长度为L，则差异率P＝N/L， 分歧度 —— K＝-¾ ln(1- 4P/3)**

* **蛋白质序列分歧度：**

**要考虑其密码子基础，区分同义变化（KS）和非同义变化（KA）。**

**11 最大似然法（maximum likelihood method）**

**（1）评估所选定的进化树能够产生实际观察到的数据的可能性：（基于置换）**

**（2）针对一个位点的进化，先把某种组合的核苷酸置于进化树的内部结点，根据取代函数计算每一段进化的可能性，将所有段的这种可能性相加，得到此进化树以此组合为进化途径产生此位点数据的可能性；**

**（3）换组合，再算。将所有组合的这种可能性相乘，得到此进化树产生此位点数据的可能性；**

**（4）换位点，再算。将所有位点的这种可能性相加，得到此进化树产生此（序列组）数据的可能性（似然值）。**

**（5）换进化树，再评估。具有最大似然值（产生此数据可能性最大）的进化树被认为是最可能的实际进化树。**

**12构建系统树的各种方法之比较**

**1）假设**

* **UPGMA：各分支置换速率一致；序列较短时易造成较大错误。**
* **邻接法：依赖于距离系数的准确性；序列短时，易有较大误差。**
* **最大简约法：无明显假设；当序列间的分歧度较大时，效果好，反之效果差。**
* **最大似然法：对进化速率和核苷酸置换型式的假设十分明确，但对违背假设的情形不敏感（Robust）。**

**2）计算时间**

**3）估计一致性（Consistency）**

* **距离矩阵法distance matrix method：进化速率恒定时一致；进化速率变化时不一致或难一致。**
* **最大简约法maximum parsimony method：不一致。**
* **进化简约法：转换/颠换＝1时一致；反之不然。**
* **最大似然法：一致性取决于建立似然函数的进化模型。**

**4）符合程度评价**

**(1)计算机模拟：模拟进化，用以评价构树方法的符合程度。**

**A.进化速率恒定时：最大简约法≤邻接法≈最小进化法；最大似然法依赖于进化模型。**

**B.进化速率可变时：最大简约法≤邻接法＜最大似然法； 但当 转换/颠换 远大于1时，邻接法＞最大似然法。**

**(2)实际进化：预先得到了实际的进化树（如实验室控制进化），再来检验分子构树的各种方法。克服了计算机模拟中参数选定的主观性。**

**5）一般构树方法选用策略**

**序列间有极高相似性：最大简约法。**

**序列间有较明显的相似性：距离矩阵法。**

**序列间没有较明显的相似性：最大似然法。**

**13 基因组大小**

**生物基因组大小和基因数目并不绝对成比例。**

**生物基因组大小与进化位置并不绝对相关。**

**14结构基因组分析——基因组作图**

* **Cytogenetic maps**
* **Genetic linkage maps**
* **STS maps**
* **RH maps**
* **Clone-based maps**

**15蛋白质组分析**

**1）分离--双向凝胶电泳（2D－gels）**

**第一向：等电聚焦（IEF），蛋白质沿pH梯度分离，至各自的等电点；**

**第二向：SDS电泳，按分子量分离。**

**2）2D－gels分析\***

* **图像分析：斑点位置和密度分析（包括斑点分离、背景消减等工作）。**
* **斑点配比：同种斑点的识别（蛋白差异表达分析的前提） 。**
* **聚类分析：蛋白质表达矩阵。**

**3)鉴定方法：**

**片段离子搜索 fragment ion searching**

**从头测序 de novo sequencing**

**肽质指纹分析（Peptide-mass fingerprinting）**

**肽质指纹分析原理**

**酶解拟鉴定蛋白质，测定实际肽段质谱图；**

**从蛋白质数据库中随机挑选一段蛋白质序列，理论酶解之，并计算出各个肽段的分子量，从而构建出理论质谱图；**

**将实际肽段质谱图与此蛋白序列的理论质谱图进行比较，进行相似性计分；**

**从数据库中重新挑选一段序列，重复以上两步；**

**最后，根据相似性计分从数据库中寻找最可能的蛋白序列。**

**4）蛋白质组分的自动化鉴定**

**高流通量筛选（High throughout screening, HTS）：**

**机器人自动处理——**

**转移双向凝胶图至PVDF膜、切割分离蛋白质组分；**

**1.控制酶解、传输至液相色谱分离、传输至质谱测肽段质量；**

**2.控制氨基酸组分分析；**

**自动搜索数据库进行蛋白质组分的鉴定。**

* **每天最小流量级达1000个蛋白质。**

**ExPASy--综合性蛋白质信息服务器**

**16 生物信息学软件**

* **学术途径：用于学术研究的软件一般可以免费或低费用获得。——自由软件**
* **商业途径：商业开发的软件需支付相当的费用，以获得产品和相应服务。——商业软件**

**自由软件的特点**

**获得：免费或低费用，作者愿意共享。英特网是获得自由软件的有效途径：有些可以直接通过浏览器访问运行（如Blast，Clustal，CAP3等） ；有些要下载（如Phylip） ，甚至在本地机器上编译。**

**使用：一般是作为专门用途的单独程序或一组程序。需要相当的熟练过程：用户必须熟悉输入/输出数据的格式，学会有效地运行这些程序。尤其是进行多个程序的分析较为麻烦。**

**商业软件的特点**

**获得：一般价格比较昂贵，特别是作为多用户用途的那些种类。以光碟或网上付费的形式提供产品以及相应的使用指南和升级服务。**

**使用：通常以集成功能的软件包形式开发。一般都提供运行程序的友好环境，利于不同功能的程序之间的相互调用或顺序运作。**

**17 生物信息学策略**

**核酸序列**

**新基因预测 蛋白质序列分析**

**数据库搜索 基序和模式识别**

**二重序列比对 二级结构预测**

**多重序列比对 三级结构预测**

**分子进化分析 评估与实验**

**信息是指在多种可能状态下的一种选择。  
当一种选择引起另一种选择时，我们理解为信息传递了。**

**生物信息学就是利用信息技术对生物信息进行搜集，加工，储存，分配，分析以解释这些信息数据所蕴涵的生物学意义的学科。狭义的生物信息学：专指应用信息技术储存和分析分子生物学数据，也被称为分子生物信息学。  
Perl 最初只是 Unix 系统管理员的一个工具，在工作日里被用在无数的小任务中。从那以后，它逐步发展成为一种全功能的程序设计语言，特别是在各种计算平台上，它被用作 Web 编程、数据库处理、XML 处理以及系统管理 ……  
UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.   
BLAST is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA.   
BLAST seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity.   
Unigene is an ongoing effort at NCBI to cluster EST sequences with traditional gene sequences.  
For each cluster, there is a lot of additional information included.  
Unigene is regularly rebuilt. Therefore, cluster identifiers are not stable gene indices.  
  
A phylogenetic analysis of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution.  
Two sequences that are very much alike will be located as neighboring outside branches and will be joined to a common branch beneath them.  
The object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.**

**Sequence Tagged Site (STS) maps——辅助性定位标记图**

**An STS is defined as a segment of genomic DNA that can be uniquely amplified by PCR using its primer sequences.   
A short DNA segment that occurs only once in the genome, the exact location and order of bases of which are known. STSs serve as landmarks on the physical map of the genome. STSs are helpful for chromosome placement of mapping and sequencing data from many different laboratories.**

**生物信息学发展阶段：  
前基因组时代：生物数据库的建立、检索工具的开发以及DNA和蛋白质序列分析。  
基因组时代：基因组测序，基因寻找和识别、网络数据库的建立和交互界面的开发等。  
后基因组时代：大规模基因组分析、蛋白质组分析及各种数据的比较和整合。**

**生物信息学研究内容：**

**上游——数据库管理技术；数据仓库、数据挖掘与数据库中的知识发现技术；分布式计算；图像处理和可视化技术**

**中游——数据库的构建；算法建立；统计模型建立；工具软件开发**

**下游——建立特定方向或自己的专用数据库；数据库检索的技术；数据分析：序列分析、进化分析等**

**数据库（Database, DB ）：统一管理的相关数据的集合。  
数据库管理系统（DB management system）\*：对DB进行管理的软件，提供DB的建立、查询、更新以及各种数据控制功能。  
数据模型（Data model，DM）：数据库结构和语义的一种抽象描述。由数据结构、数据操作和完整性约束三部分组成。**

**三代数据库系统：**

**根据其所采用的数据模型的特性，数据库分为：**

**第一代：层次、网状**

**第二代：关系(管理系统软件如：Oracle、DB2、Sybase、SQL Server、Informix等)**

**第三代：分布式、面向对象  
局域网：一般不需租用电话线路而直接建立专用通信线路，因此数据传输速率高于广域网。典型的局域网由一台或多台服务器和若干个工作站组成。  
广域网：利用电话交换网互联分布在世界各地的计算机和（局域）网络。Internet——现今世界上最大的广域计算机网络。  
网页是用户通过客户端浏览器观察到的超文本信息内容的页面。  
一个网站的主页是用户浏览某个网站时的入口。  
门户网页是服务于特定主题的、包含各种相关链接的网页，一般本身不含原始信息。  
  
Conceptional translation: 根据遗传密码表，理论上可以对任意一个DNA序列进行翻译而得到氨基酸序列，称为概念性翻译；这种通过计算机翻译而不是实验手段测定得到的蛋白质序列称为概念性序列。**

**六框翻译six-frame translation：先以所给DNA为模板，分别从（5’－3‘）第1、2、3个碱基开始翻译，得到3种翻译结果；再以其互补链为模板，依次从（5’－3‘）第1、2、3个碱基开始翻译，得到另外3种翻译结果。  
起始于起始密码子、终止于终止密码子的“中间没有终止密码子”的读码框，称作开放读码框（Open Reading Frame，ORF），又称可读框。**

**分子系统发育分析phylogenetic analysis——分子进化分析。**

**通过生物大分子序列比对，构建进化树，研究生物在分子水平的进化式样、方向、速率。   
同源Homology：最基本的意义就是具有共同祖先。在分子进化研究中，同源性一般是指两种核酸分子的核苷酸序列之间或两种蛋白质分子的氨基酸序列之间的相似程度。  
直系同源（orthology）：由“物种分化”而产生；可以反映物种血统上的同源性，即物种进化的历史。  
并系同源（paralogy）：由基因“多重化（duplicating）＋功能分化” 而产生。**

**类群group: 进化分析中分类单位的集合  
祖先类群（ancestral group）：原始类群；  
子裔类群（descendant group)：后代类群；  
单系类群（monophyletic group）：一个祖先类群的所有子裔类群的集合；  
内类群（ingroup）：一项研究所涉及的某一特定类群可称为内类群；  
外类群（outgroup）：不包含在内类群中又与之有一定关系的类群可称为外类群；  
姐妹群（sister group）：与某一类群在谱系关系上最为密切的类群称为姐妹群。**

**系统发育树（Phylogenetic tree）：表达类群间系统发育关系（进化关系）的一种树状图。  
有根树（rooted tree）—— 以外类群为树根的树。  
无根树（unrooted tree）—— 没有外类群为树根的树。  
标度树枝（scaled branch）系统树：树枝长度代表性状状态变异的数量，称为标度树枝系统树。  
非标度树枝（unscaled branch）系统树：树枝长度并不表示性状状态变异的数量，但所有节点的位置仍与分化时间相对应，称为非标度树枝系统树。  
基因树（gene tree）：根据同源基因所构建的系统树。  
物种树（species tree）：表达了物种的进化路径。**

**蛋白质组：(一个基因组表达产生的所有蛋白质的总体。)在某种外在和内在条件下，一个基因组表达产生的所有蛋白质的总体  
多态性：对于多细胞生物，组织特性不同，蛋白质组不同。  
动态性：蛋白质组随外在和内在条件而变化。  
蛋白质组学：对基因表达的蛋白质水平进行整体性研究，包括种类和表达量，阐释生命过程机制。**

**浏览（Browse）—— 信息的全面获取。  
查询（Query）  —— 特定信息的获取。  
关键词检索：以名称、记录号、分类学等级、文献题目等作为关键词提交，对数据库中的注释信息进行查询。   
序列相似性搜索：以序列本身作为提交信息对数据库中的序列信息进行查询。**

**序列对位排列（sequence alignment）：源序列与目标序列之间按碱基(或氨基酸）位置相对排列。其目标是使序列之间的相似程度最大。  
相似性记分：以记分矩阵作为序列相似性测度。  
全局（global）排列：对序列全长进行最优对位排列。  
局部（local）排列：通过对位排列使序列间的局部区域达到高度相似。**

**基因组是指一个由细胞组成的生物体的单倍细胞、细胞器或病毒所包含的所有基因与基因间序列的总体。  
单核苷酸多态性（SNP）：是指基因组内特定核苷酸位点上存在不同的碱基，其中每种在群体中的频率不小于1％。**

**人类基因组中SNP约占0.5-10‰。**

**简单序列长度多态性（SSLP）：卫星DNA (微－STR，小－VNTR） ，重复次数可变。**

**限制性片断长度多态性（RFLP）**

**单链构像多态性（SSCP） 同样长度但有单核苷酸差异的两条DNA序列，被单链化后电泳，由于构象差异而显示移动差异。**

**URF（unidentified reading frame, URF）非确定性读码框是指在DNA序列中识别出的一个可读框，但没有相关的生物学功能信息（也没有已知的同源物），它应该编码了一种蛋白质，但人们从未发现或确定过该蛋白质的功能.**

**Expressed Sequence Tags (ESTs) are short (usually about 300-500 bp), single-pass sequence reads from mRNA (cDNA). Typically they are produced in large batches. They represent a snapshot of genes expressed in a given tissue and/or at a given developmental stage. They are tags (some coding, others not) of expression for a given cDNA library.**

**利用计算机来协助**[**克隆**](http://www.bioon.com/Search.asp?Field=Title&ClassID=&keyword=克隆)**基因，称为“电子”基因**[**克隆**](http://www.bioon.com/Search.asp?Field=Title&ClassID=&keyword=克隆)**（sillcon cloning or virtual cloning），是与定位**[**克隆**](http://www.bioon.com/Search.asp?Field=Title&ClassID=&keyword=克隆)**、定位候选**[**克隆**](http://www.bioon.com/Search.asp?Field=Title&ClassID=&keyword=克隆)**策略并列的方法之一，即采用**[**生物信息学**](http://www.bioon.com/Search.asp?Field=Title&ClassID=&keyword=生物信息)**的方法延伸EST序列，以获得基因部分乃至全长的cDNA序列。EST数据库的迅速扩张，已经并将继续导致识别与**[**克隆**](http://www.bioon.com/Search.asp?Field=Title&ClassID=&keyword=克隆)**新基因策略发生革命性变化。**

**Cluster of orthologous groups（COG）——直系同源体簇**

**一个生物物种的基因组中，两个基因或可读框在各自全长的60％以上范围内，同一性不少于30％时，称为同源体。**

**不同物种中同源的同源体的集合，称为直系同源体簇（COG）**

**用EST搜索DNA数据库：**

* **数据库中如果找到匹配序列：**

**如果匹配的是CDS，则该EST可能属于此CDS，也可能是此CDS的同源序列；**

**如果匹配的是UTR，则该EST很可能就属于此UTR，因为UTR对于物种和基因通常都是特异的(不保守）**

* **数据库中如果找不到匹配序列则：**

**可能属“未知基因”，也可能只是“已知基因”的未被收录的UTR。**

**电子克隆不同于普通EST聚类－组装：**

**一般EST聚类－拼接通过聚类同一测序来源的EST来进行，可能不能得到全长信息。电子克隆可能借助同源序列拼接得到（不尽精确的）全长信息。**

**电子克隆可以以实际片断序列吊全长信息，借此设计引物，从而得到实际全长序列。**

**以上两点（在幻灯片中有下划线）代表了电子克隆不同于普通EST聚类－组装的意义。**

**基因组非编码区的意义：**

**非编码区的意义还不很明确，至少它在进化中的作用是巨大的。由于其非编码的特性，它为沉默突变的积累提供了一个平台。这应该是真核生物能够进化到如此复杂的一个重要原因。**