武汉大学生命科学学院
2004—2005学年第二学期期末考试
《生物信息学》试卷

专业生基 学号2002310600021 姓名[redacted] 得分_____

一、名词解释（用中文）：（4×4=16分）

1. 数据仓库（Data warehouse）：
2. 开放读码框（Open reading frame，ORF）：
3. 序列表达标签（Expressed sequence tag，EST）
4. 单系类群（Monophyletic group）：

二、填空：（17×1=17分）

1. 促成生物信息学产生的原因主要是：一方面，生物学数据 的极剧增长 另一方面，计算机与信息科学 技术的飞速发展 [illegible]
2. 数据库的出现是计算机应用的一个里程碑，它使计算机应用从过去单一的 [illegible] 转向了以 数据处理 为主，从而使计算机进入了各行各业乃至普通家庭；同时也为生物信息学的迅速发展提供了条件。
3. 与传统数据库中高度结构化的数据相比，网上数据最大的特点是 结构 特征较弱——每个站点独立并可能 异构，这是当今数据管理技术研究的热点，也是挑战。
4. 目前，最重要的公共核苷酸序列数据库是 Genbank、EMBL 和 DDBJ，它们每天进行数据交换，同时更新序列资料，所以无论是 [illegible] 都基本等价。
5. Genbank 数据库主要有两种查询方式，即 [illegible] 检索和 BLAST 搜索。
6. 蛋白质数据库主要分为三个层次，即 [illegible] 的序列 数据库、模式 数据库和 结构 数据库。
7. 蛋白质组分析有两个出发点，即在蛋白质组的背景下研究蛋白质的 [illegible] 和蛋白质的 [illegible]

三、选择（单项或多项，漏选得1分，错选得零分。5×3=15分）

1. WWW____。（ADE）
   A. 是一种信息网络　B. 是一种物理网络　C. 是对“Internet”的另一种称谓
   D. 基于超文本组织结构　E. 是Internet上的一种信息服务

2. 以下_____属于蛋白质一级复合数据库。（AB）
   A. NRDB　B. UniProt　C. InterPro　D. Swiss-Prot
   [illegible]

3. 用蛋白质序列搜索蛋白质序列数据库用_____。（B）
   A. BLASTn　B. BLASTp　C. BLASTx　D. tBLASTn　E. tBLASTx

4. 在基因组作图中，属物理结构图的有_____。（ABC）

A. 染色体核型图　B. 细胞遗传学图　C. STS 图　D. 遗传连锁图

5. ExPASy 是一个综合性的___服务器。（C）
A. 基因信息　B. 基因组信息　C. 蛋白质信息　D. RNA 信息

四、判断（正确--V；错误—X。 6×2=12 分）

1. 统一资源定位器（URL）实现了单一计算机在整个 Internet 中的定位。（X）
2. CLUSTAL 是一种多序列比对软件。（√）
3. 序列表达标签（EST）中可能会出现模糊碱基甚至错误碱基。（√）
4. 在构建进化树的几种方法中，最大简约法比最大似然法的运算速度更慢。（X）
5. UPGMA 是一种距离构树法。（√）
6. 直系同源体族（COG）中也包含并系同源的基因。（X）

五、英汉互译：(2×5=10 分)

1. A STS （Sequence Tagged Site）is a short DNA segment that occurs only once in the genome, the exact location and order of bases of which are known. STSs are helpful for chromosome placement of mapping and sequencing data from many different laboratories.
2. 在过去的十年中，世界范围内的分子生物学家“制造”的信息数量经历了真正的爆炸。急速成长的生物信息学领域（大致定义为分子生物学和计算生物学的交叉）已成为一个全职的工作，其中的研究已促成了无数的重要发现，而且很可能揭示更多的自然之谜。

六、问答：（30 分）

1. 作为一个跨学科领域，生物信息学研究可以形象地划分为“上、中、下游”三个亚领域。请按这种划分描述生物信息学的研究内容。（6 分）
2. (1) 请描述用最大似然法构建系统发育树的算法过程。(2) 请叙述肽质指纹图谱分析的基本原理。 两小题任选一题。（8 分）
3. 生物信息学的自由软件和商业软件的特点各是什么？（6 分）
4. 已知来自某物种的一段 DNA 的序列信息，如何运用生物信息学的方法，尽可能多地得到其它各种相关信息？ 请结合图示说明。(10 分)

武汉大学生命科学学院
2005－2006学年度第二学期期末考试
《生物信息学》（A）试卷

专业_____ 学号_____ 姓名_____ 得分_____

一、名词解释（用中文）：（5×4＝20分）

1. 六框翻译（Six-frame translation）
2. 序列对位排列（Sequence alignment）
3. 序列表达标签（Expressed sequence tag，EST）
4. 单系类群（Monophyletic group）：
5. 非确定性读码框（unidentified reading frame，URF）

二、填空：（18×1＝18分）

1. 生物信息学的资源主要以_____形式管理，以_____形式共享。
2. 数据_____的出现是计算机应用的一个里程碑，它使计算机应用从过去单一的_____转向了以_____为主，从而使计算机进入了各行各业乃至普通家庭，同时也为生物信息学的迅速发展提供，条件。
3. 目前，最重要的公共核苷酸序列数据库是 Genbank、_____和_____，它们每天进行数据交换，同时更新序列资料，所以无论是_____还是_____都基本等价
4. 序列相似性分析可以用来进行_____推测、结构推测和_____推测。
5. 一般来说，基因预测主要通过_____预测、_____预测和序列相似性搜索进行。
6. 目前，分子系统_____构建方法有三大类：_____、_____、_____。
7. 自由软件的特点是可以免费或低费用获得，英特网是获得自由软件的有效途径，有些可以直接通过_____运行（如 Blast，Clustal，CAP3 等）；有些需要_____（如 Phylip），甚至在本地机器上_____。

三、单项选择（6×2＝12分）

1. （　）早期的一些生物信息源采用的只是_____的管理方式，随着数据量的急剧增加，越来越多地采用_____主流技术。
   A. 手工……文件系统　B. 手工……数据库　C. 文件系统……数据库

2. （　）以下哪种对 WWW 的描述不正确。
   A. 是一种信息网络　B. 是对“Internet”的另一种称谓
   D. 基于超文本组织结构　E. 是 Internet 上的一种信息服务

3. （　）NRDB 是蛋白质一级复合数据库，由 NCBI 建立；另一个和它基本等价的、主要由 EBI 负责维护的数据库是：
   A. UniProt　B. Swiss-Prot　C. PDB　D. PIR

MegaBlast 搜索或高相似度序列

4. (B) 用蛋白质序列搜索蛋白质序列数据库用_____。
A. BLASTn B. BLASTp C. BLASTx D. tBLASTn E. tBLASTx

5. (C) 在基因组图中，____属功能标记图。利用STS进行的序列检索
A. 染色体核型图 B. 细胞遗传学图 C. STS 图 D. 遗传连锁图

6. (B) Ensembl 是著名的____服务器。
A. 基因信息 B. 基因组信息 C. 蛋白质信息 D. RNA 信息

四、判断（正确--V；错误--X。 10×1＝10 分）

1. (X) IP 地址实现了单一文档在整个 Internet 中的定位。
2. (V) FASTA 格式是描述核酸或蛋白质序列信息的著名格式，很多分析软件都能识别这种格式。
3. (V) PROSITE 是著名的蛋白质二级数据库，主要储存蛋白质基序的信息。
4. (X) 在使用 BLAST 过程中，设置 E 值越低，则搜索敏感性越高，命中的特异性越低。
5. (V) 概念性翻译时，如果选用序列很长，六框翻译中可能不止一框含有正确的（对应 CDS 的）ORF。
6. (V) 序列表达标签（EST）中可能会出现模糊碱基甚至错误碱基。
7. (X) CLUSTAL 是一种多序列比对软件。改为：BLAST
8. (X) UPGMA 是一种距离构树法。NJ FM ME
9. (X) 直系同源体族（COG）中也可能包含并系同源的基因。
10. (X) GCG 是著名的生物信息学集成环境，属自由软件。 Wisconsin Package

五、英汉互译：(2×5＝10 分)

1. A phyiogenetic 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.
2. 在过去的十年中，世界范围内的分子生物学家"制造"的信息数量经历了真正的爆炸。急速成长的生物信息学领域（大致定义为分子生物学和计算生物学的交叉）已成为一个全职的工作，其中的研究已促成了无数的重要发现，而且很可能揭示更多的自然之谜。

六、问答：(30 分)

1. 作为一个跨学科领域，生物信息学研究可以形象地划分为"上、中、下游"三个亚领域。请按这种划分描述生物信息学的研究内容。(6 分)
2. 在面向靶蛋白的药物开发中，计算机方法是如何起作用的？(5 分)
3. 生物大分子序列数据库的记录一般分为哪两个部分，相应的查询可以分为哪两种方式？(4 分)。
4. 请描述用最大似然法构建系统发育树的算法过程。(5 分)
5. 已知来自某物种的一段蛋白质的序列信息，如何运用生物信息学的方法，尽可能多地得到其它各种相关信息？请结合图示说明。(10 分)

protein

一、名解

1.

六框翻译：对任意给定的一段 DNA 序列，不知道其读码方向（即不知其是正义链还是反义链），也不能确定其编码区是否从第一个碱基开始，则必须将其所有的读码框全部都翻译出来，即六框翻译——

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。

2.

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。

3.

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。

- 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。

4.

单系类群（monophyletic group）：一个祖先类群的所有子裔类群的集合，或称为“进化枝”（clade）。

5.

非确定性读码框（unidentified reading frame ，URF）：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。

五、英汉互译(句)系统发育分析

1. 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.

对于一个有相关的核苷酸或者蛋白质序列的家族进行系统发育分析，是来判断这个家族在进化中是如何起源和发展的。两个十分相近的序列会被定位于相邻的外部分支上而且会合并到一个在其之下的共同的分支。系统发育分析的目的就是揭示系统树上所有的分支关系和分支长度。

2. 在过去的十年中，世界范围内的分子生物学家“制造”的信息数量经历了真正的爆炸。抓住这个信息的瀑布现在在大多数情况下已成为一个全职的工作，而在不久前，这些信息的分析主要是由那些宁愿拿吸管而不是敲键盘的心怀不满的学生来进行。急速成长的生物信息学领域（大致定义为分子和计算生物学的交叉），已经促成了无数的重要发现，而且很可能

一、名词解释

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。

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。

六框翻译：对任意给定的一段 DNA 序列，很难确定其编码区是否从第一个碱基开始，也不知道其读码方向（即不知其是正义链还是反义链），则必须将其所有的读码框全部都翻译出来，即六框翻译——

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。

序列表达标签（EST）：是从 cDNA 文库中生成的一些很短的序列（300-500bp），它们代表在特定组织或发育阶段表达的基因，有时可代表特定的 cDNA。EST 可能是编码的，也可能不是，而两端有重叠序列的 EST 可以组装成全长的 cDNA 序列。

开放阅读框：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。

系统发育分析：

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。比如，小鼠和大鼠的肌红蛋白。

二、填空

1、近 20 余年，以基因组计划实施划分，生物信息学经过哪三个阶段：前基因组时代、基因组时代、后基因组时代

2、与传统数据库相比，网上数据最大的特点是结构化特征较弱，既每个站点独立并可能异构，这是当今数据管理技术研究的热点，也是挑战

3、生物信息学资源主要以数据库形式管理，以网络访问形式共享

4、目前，最常用的公共核苷酸序列数据库是 Genbank、EMBL 和 DDBJ，它们每天进行数据交换，同时更新序列资料，所以无论是投送数据还是查询数据基本等价。

5、蛋白质数据库主要分为三个层次，即一级数据库、二级（蛋白质模式）数据库和三级（结构）数据库。

6、DNA 序列进化的一个基本过程就是核苷酸随时间而变化，我们可以用核苷酸置换模型来描述这个基本过程的机制。

7、目前，分子系统树构建的方法有三大类：距离矩阵法、简约法、最大似然法

8、基因组信息可以分三个层 次，即[illegible]基因组、[illegible]基因组和[illegible]基因组。

9、蛋白组分析有两个出发点，即在蛋白质水平下研究蛋白质的动态信息和蛋白质的相互关系。

三、选择

1、早期一些生物信息资源采用的只是_____的管理方式，随着数据量急剧增加，越来越多地采用_____主流技术。 （C）

A 手工 文件系统 B 手工 数据库 C 文件系统 数据库

2、用 DNA 序列搜索蛋白质序列数据库用_____ （C）

A、BLASTn B、BLASTp C、BLASTx D、tBLASTn E、tBLASTx

3、NRDB 是蛋白质一级复合数据库，由 NCBI 建立，另一个和它基本等价的，主要由 EBI

揭示更多的自然之谜。

In the past ten years, the quantity of the information "made" by molecular biologists all around the world had undergone a real explosion. It has been a full-time job to grab this waterfall of information in most cases. However, just not a long time ago, the analysis of the information was mainly made by the unsatisfied students who would rather take the tubes than knock the keyboards. The fast growing field of bioinformatics(approximately defined as the crossing of molecular biology and computational biology) has already led to numerous important findings and more discoveries of the mysteries of the nature.

六、问答题

1. 上游：有针对性的计算机技术开发；
中游：利用这些技术建立相关数据库、工具、算法、软件等；
下游：利用这些工具有效管理和处理生物学数据。

上游：
1.数据库管理技术。
2.数据仓库、数据挖掘与数据库中的知识发现技术。
3.分布式计算（网格计算等）
4.图像处理和可视化技术。

中游：
1.数据库的构建。
2.算法建立。
3.统计模型建立。
4.工具软件开发。

下游：
1.建立特定方向或自己的专用数据库。
2.数据库检索的技术。
3.数据分析：序列分析、进化分析等。

2. 计算机辅助药物设计（computer aided drug design）是以计算机化学为基础，通过计算机的模拟、计算和预算药物与受体生物大分子之间的关系，设计和优化先导化合物的方法。计算机辅助药物设计实际上就是通过模拟和计算受体与配体的这种相互作用，进行先导化合物的优化与设计。

计算机辅助药物设计根据受体的结构是否已知，分为直接药物设计和间接药物设计。

计算机辅助药物设计的一般原理是，首先通过X-单晶衍射技等技术获得受体大分子结合部位的结构，并且采用分子模拟软件分析结合部位的结构性质，如静电场、疏水场、氢键作用位点分布等信息。然后再运用数据库搜寻或者全新药物分子设计技术，识别得到分子形状和理化性质与受体作用位点相匹配的分子，合成并测试这些分子的生物活性，经过几轮循环，即可以发现新的先导化合物。因此，计算机辅助药物设计大致包括活性位点分析法、数据库搜寻、全新药物设计。

1. 活性位点分析法

该方法可以用来探测与生物大分子的活性位点较好地相互作用的原子或者基团。用于分析的探针可以是一些简单的分子或者碎片，例如水或者苯环，通过分析探针与活性位点的相互作用情况，最终可以找到这些分子或碎片在活性部位中的可能结合位置。由活性位点分析得到的有关受体结合的信息对于全新药物的设计具有指导性。

目前，活性位点分析软件有 DRID、GREEN、HSITE 等。另外还有一些基于蒙特卡罗、模拟退火技术的软件如 MCSS、HINT、BUCKETS 等。

2. 数据库搜寻

目前数据库搜寻方法分为两类。一类是基于配体的，即根据药效基团模型进行三维结构数据库搜寻。该类方法一般需先建立一系列活性分子的药效构象，抽提出共有的药效基团，进而在现有的数据库中寻找符合药效基团模型的化合物。该类方法中比较著名的软件有 Catalyst 和 Unity，而以前者应用更普遍。另一类方法是基于受体的，也称为分子对接法，即将小分子配体对接到受体的活性位点，并搜寻其合理的取向和构象，使得配体与受体的形状和相互作用的匹配最佳。在药物设计中，分子对接方法主要用来从化合物数据库中搜寻与受体生物大分子有较好亲和力的小分子，从而发现全新的先导化合物。分子对接由于从整体上考虑配体与受 体的结合效果，所以能较好地避免其他方法中容易出现的局部作用较好，整体结合欠佳的情况。目前具代表性的分子对接软件主要有 DOCK、F1exX 和 GOLD。

3. 全新药物设计

数据库搜寻技术在药物设计中广为应用，该方法发现的化合物大多可以直接购买得到，即使部分化合物不能直接购买得到，其合成路线也较为成熟，可以从专利或文献中 查得，这都大大加快了先导化合物的发现速度。但是，数据库搜寻得到的化合物通常都是已知化合物，而非新颖结构。近年来，全新药物设计越来越受到人们的重视，它根据受体活性部位的形状和性质要求，让计算机自动构建出形状、性质互补的新分子，该新分子能与受体活性部位很好地契合，从而有望成为新的先导化合物[1][19]；它通常能提出一些新的思想和结构类型，但对所设计的化合物需要进行合成，有时甚至是全合成。全新药物设计方法出现的时间虽然不长，但发展极为迅速，现 已开发出一批实用性较强的软件，其主要软件有 LUDI、Leapfrog、GROW、SPROU 等，其中 LUDI 最为常用。

LUDI 是由 Bhöm 开发的进行全新药物设计的有力工具，已广泛地被制药公司和科研机构使用[16]，其特点是以蛋白质三维 结构为基础，通过化合物片段自动生长的方法产生候选的药物先导化合物。它可根据用户确定好的蛋白质受体结合部位的几何形状和物理化学特征（氢键形成能力、 疏水作用位点），通过对已有数据库中化合物的筛选并在此基础上自动生长或连接其他化合物的形式，产生大量候选先导化合物并按评估的分值大小排列，供下一步 筛选；可以对已知的药物分子进行修改，如添加/去除基团、官能团之间的连接等。在受体蛋白质结构未知的情况下，此模块也可以根据多个已知的同系化合物结构 的叠合确定功能团，再根据功能团的空间排列和理化性质推测可能的蛋白质受体结合部位特征，根据此特征进行新型药物设计。目前研究人员利用 LUDI 设计出数 十个针对不同疾病的活性化合物。

3.

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。——关键词检索（Retrieval）

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。——序列相似性搜索（Search） （第三章 PART 2）

4. 最大似然法（maximum likelihood method）：

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）

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乘，得到此组合此位点此进化树为真的可能性；

换组合，再算。将所有组合的这种可能性相加，得到此位点此进化树为真的可能性；

换位点，再算。将所有位点的这种可能性相乘，得到此进化树为真的可能性（似然值）。

换进化树，再评估。具有最大似然值的进化树被认为是最可能的实际进化树。

5.

（1）

1、把蛋白序列对应的核酸序列找到。

2、根据核酸序列做 BLAST （对 dbSNP 数据库 http://www.ncbi.nlm.nih.gov/SNP/snpblastByChr.html）。

3、结果中，可以得到你的序列上所以已知的 SNP。

以编码 5-hydroxytryptamine (serotonin) receptor 的 HTR1A 基因为例，先进入 entrez，选择 gene 选项，在 NCBI 中搜 HTR1A 基因，到如下图界面，点基因缩略图，然后点 graphics。

（2）

使用 BioEdit 软件对人脂联素蛋白质序列进行分子质量、氨基酸组成和疏水性等基本性质分析：

蛋白质序列的蛋白质同源性分析：点击 BLAST：查看与之同源的蛋白质；

蛋白质序列的 motif 结构分析；

进行二级结构预测；

三维结构预测。

分子系统发育分析：分子进化分析：

就是通过生物大分子序列的对比，构建进化树来研究物种之间或分子种群的进化关系，与同源

一、名词解释

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学。

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。

六框翻译：对任意给定的一段DNA序列，很难确定其编码区是否从第一个碱基开始，也不知道其读码方向（即不知其是正义链还是反义链），则必须将其所有的读码框全部都翻译出来，即六框翻译——

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。

序列表达标签（EST）：是从cDNA文库中生成的一些很短的序列（300-500bp），它们代表在特定组织或发育阶段表达的基因，有时可代表特定的cDNA。EST可能是编码的，也可能不是，而两端有重叠序列的EST可以组装成全长的cDNA序列。

开放阅读框：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。

系统发育分析：进化分析 分析系统发育的树：指对于进化分析，通过生物大分子的序列建立一树状进化树

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，即物种进化史明在分子水平的历史。比如，小鼠和大鼠的肌红蛋白。

二、填空

1、近20余年，以基因组计划实施划分，生物信息学经过哪三个阶段：前基因组时代、基因组时代、后基因组时代

2、与传统数据库相比，网上数据最大的特点是结构化特征较弱，既每个站点独立并可能异构，这是当今数据管理技术研究的热点，也是挑战

3、生物信息学资源主要以数据库形式管理，以网络访问形式共享

4、目前，最常用的公共核苷酸序列数据库是Genbank、EMBL和DDBJ，它们每天进行数据交换，同时更新序列资料，所以无论是投送数据还是查询数据基本等价。

5、蛋白质数据库主要分为三个层次，即一级数据库、二级（蛋白质模式）数据库和三级（结构）数据库。

6、DNA序列进化的一个基本过程就是核苷酸随时间而变化，我们可以用核苷酸置换模型来描述这个基本过程的机制。

7、目前，分子系统树构建的方法有三大类：距离矩阵法、简约法、最大似然法

8、基因组信息可以分三个层次，即结构基因组、功能基因组和比较基因组。

9、蛋白组分析有两个出发点，即在蛋白质水平下研究蛋白质的动态信息和蛋白质的相互关系。

三、选择

1、早期一些生物信息资源采用的只是_____的管理方式，随着数据量急剧增加，越来越多地采用_____主流技术。 （C）

A 手工 文件系统 B 手工 数据库 C 文件系统 数据库

2、用DNA序列搜索蛋白质序列数据库用_____ （C）

A、BLASTn B、BLASTp C、BLASTx D、tBLASTn E、tBLASTx

3、NRDB是蛋白质一级复合数据库，由NCBI建立，另一个和它基本等价的，主要由EBI

负责维护的数据库是：　（A）

A、Uniprot　B、Swiss-prot　C、PDB　D、PIR

4、ExPASy 是一个综合的_____服务器　（C）

A 核苷酸信息　B 基因组信息　C 蛋白质信息　D 生物信息学工具

5、目前较通用的两种分子系统发育分析软件包是：　（BC）

A、DNAstar　B、PHYLIP　C、PAUP　D、GCG

6、请将一下构树方法的计算时间由少到多排列　（CDAB）

A 最大简约法　B、最大似然法　C、距离矩阵法　D、进化简约法

距离矩阵法<进化简约法<最大简约法<最大似然法.

四、简答

1、蛋白组是指什么？蛋白组分析主要有哪些步骤与关键技术？

答：蛋白组：一个基因组表达产生的所有蛋白质的总体。（抽象）

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。（具体）

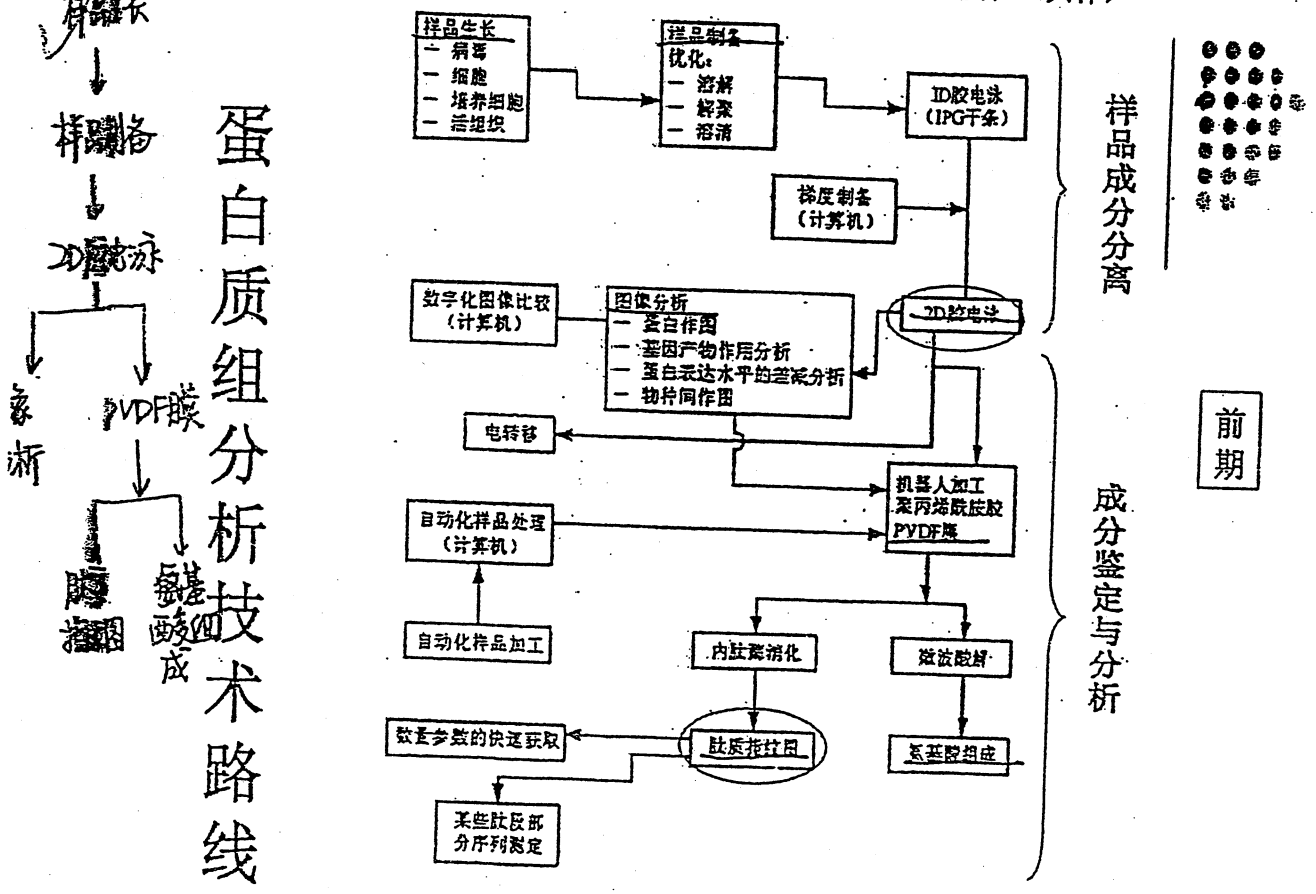


图 7-1　大规模蛋白质组分析技术路线

2、生物信息学的自由软件和商业软件的特点是什么？

答：生物信息学工具的种类：自由软件与商业软件

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。

五、问答

1、请描述最大似然法构建系统发育树的算法过程

答：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）

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乘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；

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

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

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。

2、请叙述肽指纹图谱分析的基本原理

答：（1）酶解拟鉴定蛋白质，测定实际肽段质谱图；

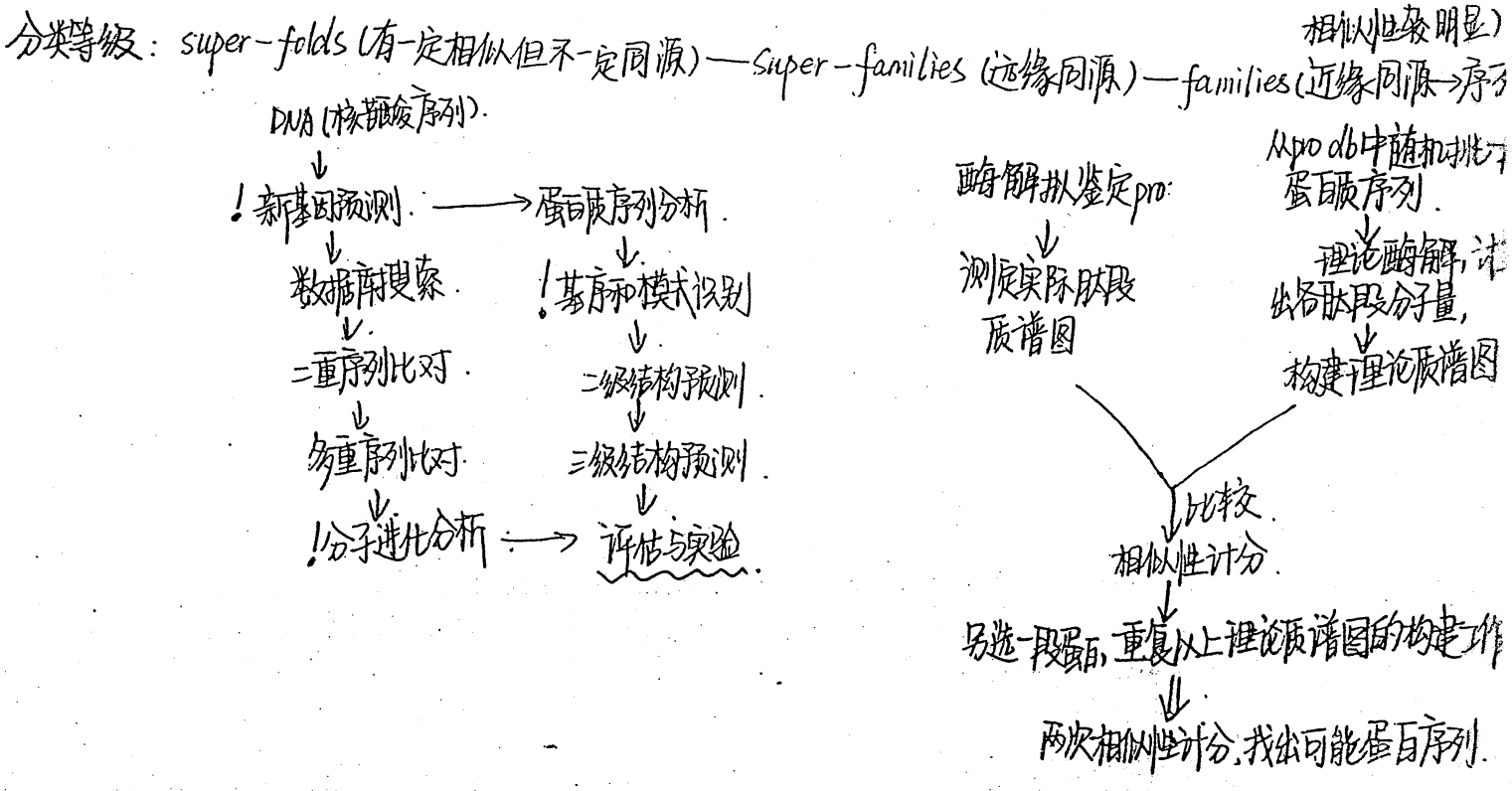
（2）从蛋白质数据库中随机挑选一段蛋白质序列，理论酶解之，并计算出各个肽段的分子量，从而构建出理论质谱图；

（3）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；

（4）从数据库中重新挑选一段序列，重复以上两步；

（5）最后，根据相似性计分从数据库中寻找最可能的蛋白序列。

3、已知来自某物种的一段 DNA 序列信息，如何运用生物信息学方法，尽可能多的得到其他各种相关信息？请结合图　示说明



武汉大学生命科学学院
2007－2008 学年度第二学期期末考试
《生物信息学》（A）试卷 —— 评分标准

任课教师：
2008.6.25

一、填空：（14×1＝14分）

1. 促成生物信息学产生的原因主要是：一方面，生物学数据 指数增长 ；另一方面，计算机与信息科学 迅猛发展 。

2. 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。

3. 目前，最重要的公共核苷酸序列数据库是 Genbank， EMBL 和 DDBJ ，它们每天进行数据交换，同时更新序列资料，所以无论是 投递数据 还是 查询 都基本等价。

4. DNA 序列进化的一个基本过程就是 核苷酸 随时间而变化。我们可以用 置换 模型来描述这个基本过程的机制；一般来说，我们用矩阵来表示这种模型。

5. 分子进化树的构建方法的原理或者基于核苷酸（氨基酸）在进化过程中的 置换 ，或者基于目标序列之间的 相异性 。

6. 自由软件的特点是可以免费或低费用获得。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）；有些需要 下载 （如 Phylip），甚至在本地机器上 编译 。

二、单项选择（6×2＝12分）

1. （ C ）早期的一些生物信息源采用的只是_____的管理方式，随着数据量的急剧增加，越来越多地采用_____主流技术。
A. 手工……文件系统　B. 手工……数据库　C. 文件系统……数据库

2. （ B ）以下哪种对 WWW 的描述不正确。
A. 是一种信息网络　B. 是对“Internet”的另一种称谓
D. 基于超文本组织结构　E. 是 Internet 上的一种信息服务

3. （ B ）以下哪一个数据库是数据源库，而不是次级数据库：
A. RefSeq　B. Swiss-Prot　C. PROSITE　D. UniProt

4. （ B ）用蛋白质序列搜索蛋白质序列数据库用_____。
A. BLASTn　B. BLASTp　C. BLASTx　D. tBLASTn　E. tBLASTx
DNA→DNA　P→P　DNA搜索蛋白质

5. （ C ）以下构树方法计算速度最快的是：
A. 最大简约法　B. 最大似然法　C. 距离矩阵法　D. 进化简约法

6. （ B ）Ensembl 是著名的_____服务器。
A. 基因信息　B. 基因组信息　C. 蛋白质信息　D. RNA 信息

三、判断（正确--V；错误--X。 10×1＝10分）

1. （ X ）IP 地址实现了单一文档在整个 Internet 中的定位。
2. （ V ）Perl 语言是一种解释性的脚本语言。
3. （ V ）FASTA 格式是描述核酸或蛋白质序列信息的著名格式，很多分析软件都能识别这种格式。
4. （ X ）目前最著名的蛋白质序列数据库是 Protein Data Bank （PDB）。
5. （ X ）在使用 BLAST 过程中，设置 E 值越低，则搜索敏感性越高，命中的特异性越低。
6. （ X ）EST 是 cDNA 文库自动测序得到的 CDS 序列片段信息。
7. （ V ）Unigene 是 NCBI 维护的一个 EST 聚类库。
8. （ V ）一个 STS 由一对 PCR 引物来定义，它在相应的基因组中只出现一次。
9. （ V ）直系同源体族（COG）中也可能包含并系同源的基因。
10. （ X ）Phylip 是一种著名的多序列比对软件，可以免费获得。

四、名词解释（用中文）：(6×4=24 分)

1. 算法（algorithm）：
   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。

2. 六框翻译（Six-frame translation）：
   对任意给定的一段 DNA 序列，很难确定其编码区是否从第一个碱基开始，也不知道其读码方向（即不知其是正义链还是反义链），则必须将其所有的读码框全部都翻译出来，即六框翻译。

3. 序列拼接（Sequence assembly）
   根据序列片段头尾的重复序列进行对位而得到更长序列信息的一种方法。

4. 分子系统发育分析（Phylogenetic analysis）：
   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。

5. 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。

6. 电子 PCR（Electronic-PCR）：
   是模拟 PCR 寻找特定 DNA 序列中的 STS 的一种计算机方法，要求已知该序列的核苷酸排列信息。

五、英汉互译：(2×5=10 分)

1. Cluster of orthologous groups (COG): The availability of multiple complete genome sequences spurred both the demand for the construction of an evolutionary classification of genes from these genomes. Such a classification system based on orthologous relationships between genes appears to be a natural framework for comparative genomics.
   评分点：基本语义正确 3 分，专业单词正确 1 分，翻译流畅 1 分。

2. 目前生物信息学取得的重大进展主要在于：从收集和整理原始数据到发展新的、精辟的方法去分析数据。所有这些都在一个信息和技术自由共享的环境中。生物信息

实验数据 → 数据库 → 分析 → 结果 → 数据库

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。

1)DNA序列分析的对象是基因未知性质的DNA序列——针对基因组
EST分析的对象是表达了的DNA——针对转录基因组

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 —— $K=-\frac{3}{4}\ln(1-\frac{4P}{3})$

- 蛋白质序列分歧度：

要考虑其密码子基础，区分同义变化（K_S）和非同义变化（K_A）。

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加，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

距离矩阵法 < 进化简约法 < 最大简约法 < 最大似然法

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 GCG

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 blast2 二级结构预测

多重序列比对 Clustalw 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。 产生源于自然选择

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 ..... bioperl是一组Perl模块

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. Basic Local Alignment Search Tool

Unigene is an ongoing effort at NCBI to cluster EST sequences with traditional gene sequences. EST聚类序

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 有针对性的计算机技术开发

中游——数据库的构建；算法建立；统计模型建立；工具软件开发 利用这些技术建立相关数据库、工具、软件

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 利用这些工具有效管理和处理生物学数据

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。和管理一类计算机管理一数据库

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。

数据控制功能、完整性、安全性、并发控制、数据库恢复

RefSeq (Reference Sequence) 非冗余2次序

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基因，称为“电子”基因克隆（silicon cloning or virtual cloning），是与定位克隆、定位候选克隆策略并列的方法之一，即采用生物信息学的方法延伸 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新基因策略发生革命性变化。根据大量EST具有相互重叠的性质，通过序列比对而延伸，最终得到cDNA全长序列信息。

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。

已知来自某物种的一段cDNA序列信息，如何运用生物信息学方法，尽可能地得到其它相关信息（请结合图示说明）？

protein (blastp, tblastn) nucleotide (blastn, tblastx)

序列相似性分析 { 意义：进化关系推测；功能推测
应用：比对—精确性 { blast2 proteins; 多重 alignment
多搜索—速度

基因预测

序列特征分析

DNA：基因预测（CDS预测、非CDS特征位点预测）
非基因序列特征预测

蛋白质：亲疏水性分析、磷酸化位点、结构域预测等

EST 拼接软件：phrap, Cap3, Staden, Cat, TIGR等

Genomic Database：NCBI Genome；UCSC；Ensembl

Protein Structure Alignment

Sequence Alignment 序列之间的相似性最大

Structure Alignment (RMSD) 序列同相应的球基碳原子空间位置最接近

蛋白质信息数据库

序列 源库 → 模式二次库(motif、模式、特征型式)
↓ 复合库 ↓ 模式复合库

结构 源库 → 分类二次库.

构象对预测: pMHC. pdbUP.
多序列比对: clustalW. ~
基因组数据库: NCBI. UCSC. Ensembl.
EST聚类: Unigen

核酸 Genbank EMBL DDBJ ……

蛋白质
- 序列源: NRDB, UniProt; Swiss-Prot, TrEMBL, PIR, NRL-3D ……
- 序列模式二次: Interpro; PROSITE, PRINTS, Profiles, Pfam ……
- 结构源: PDB / MSD
- 结构分类二次: SCOP, CATH, MMDB, FSSP ……

	Query Sequence	Database type
blastp	aa	protein
blastn	nucleotide	nucleotide
blastx	nucleotide (six-frame)	protein
tblastn	aa	nucleotide (six-frame)
tblastx	nucleotide (six-frame)	nucleotide (six-frame)

起始序列 → 检索 → 识别序列 → 调用 → 相关序列 → 对位排列 / 文献
相关序列 → 预测 → 结构 → 分析 → 功能分析