第四章 基因组测序及分析

人类基因组和其它一些生物基因组的大规模测序将成为科学史上的一个里程碑。基因组测序带动了一大批相关学科和技术的发展，一批新兴学科脱颖而出，生物信息学、基因组学、蛋白质组学等便是一批最前沿的新兴学科。可以说，基因组测序及其序列分析使整个生命科学界的真正认识了生物信息学，生物信息学也真正成为了一门受到广泛重视的独立学科。

基因组测序及其分析实际是人类的又一场“淘金”和“探险”运动。哥伦布等一大批探险家在几百年前发现了美洲、澳洲等一大批新大陆，最终使人类认识了地球上的每一块处女地。于是有人形象地把人类目前的基因组研究形象地比喻为“地球探险”，并把基因组研究称为基因组地理（genomic geography）。我们不妨想象一下，人类基因组的各条染色体就如同人类基因“地球”上的 7 大洲，寻找新基因和搞清楚基因组结构与功能的过程恰如开垦地球上的每一块处女地，而这些处女地上可能蕴藏着无穷的宝藏。目前人类全基因组序列已基本测定完成，另有一大批生物也已完成基因组测定或正在进行。世界上无数大型测序仪（最好的测序仪一次可以阅读 1000 多个碱基）日夜不停地运转，每日获得的序列数据以百万和千万计。同时，来自政府和企业的大量投资，使整个世界的测序能力与日俱增。面对基因组的天文数据，分析方法举足轻重，大量新的分析方法被提出和改进，大量重要基因被发现；大量来自基因组水平上的分析比较结果被公布，这些结果正在改变人类已有的一些观念。

第一节 DNA 测序及序列片段的拼接

一．DNA测序的一般方法[1]

1．DNA 测序的基本原理

DNA 序列测定的工作基础是在变性聚丙烯酰胺凝胶(测序胶)上进行的高分离度的电泳过程。这些所谓的测序胶能在长达 500bp 的单链寡核苷酸中分辨出一个脱氧核苷酸的差异。操作时，在相应的待测 DNA 区段产生一套标记的寡核苷酸单链，它们有固定的起点，但另一端是按模板序列连续终止于各不相同的核苷酸。确定每个脱氧核糖核苷酸的序列的关键，是在 4 个独立的酶学或化学反应中产生终止于所有不同的 A、T、G、C 位点的寡核苷酸链，而这 4 个反应的寡核苷酸产物在测序胶的相邻泳道中都能被一一分辨出来。由于在 4 个泳道中再现了所有的可能寡核苷酸链，DNA 的序列能从图 4.1 所示的 4 个寡核苷酸“阶梯”中依次直接读出。

实际上，从一套测序反应中所能获得的信息量受限于测序胶的分离度。虽然最新的测序技术经常可从一套测序反应中测到高达 500 核苷酸的信息，但获得的可靠序列信息大约在 300 个核苷酸。因此，如果待测 DNA 的区段在 300 核苷酸以

[1]本部分内容主要取自F．奥斯伯，R．E．金斯顿等．精编分子生物学实验指南，北京：科学出版社，1998

内，所需的工作只是简单地将此片段克隆于合适的载体，以产生一个能方便地进行测序的重组 DNA 分子。

对于大片段 DNA 的序列测定，往往需要将其切割成能单独进行测定的小片段，这可通过随机的或有序的方式进行。下一节将讨论测定大片段 DNA 的策略。

目前广泛应用于 DNA 序列测定的方法有酶学的双脱氧法和化学裂解法，在产生寡核苷酸“阶梯”的技术上，两者截然不同。酶学双脱氧法是利用 DNA 聚合酶合成与模板互补的标记拷贝，化学裂解法是一套碱基专一的化学试剂作用于标记好的 DNA 链。这两种方法下面将进一步描述。

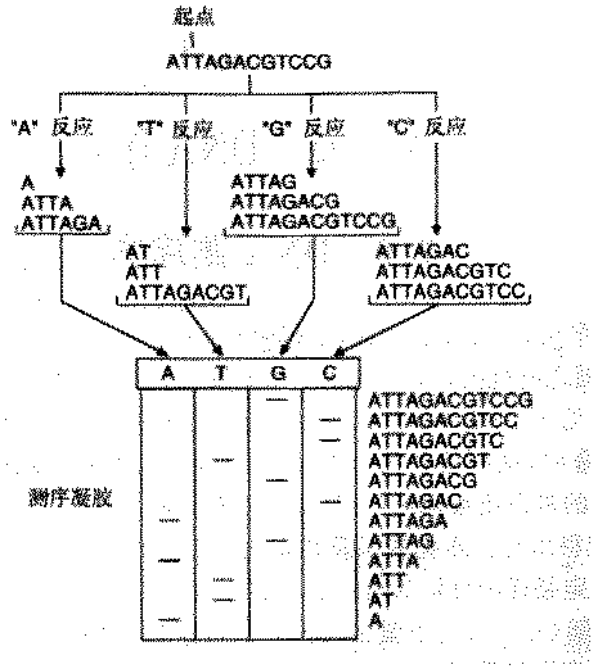


图 4.1 DNA 测序的一般策略。进行 DNA 序列测定时，在 4 个独立的反应中，各产生一套放射性标记的单链寡核苷酸，它们有固定的起点，另一端终止于不同的 A、T、G 或 C 位点。每个反应的产物在高分离度的聚丙烯酰胺凝胶上电泳分级。经放射自显影，DNA 序列可从凝胶上直接读出（奥斯伯等，1998）。

2．双脱氧测定法(Sanger 法)

双脱氧法或酶法利用DNA聚合酶合成单链DNA模板的互补拷贝，这一方法最先（1977）由F．Sanger 及其合作者提出。DNA聚合酶不能起始DNA链的合成，而能在退火于“模板”DNA的引物 3′端上进行链的延伸（如图 4.2）。通过与模板碱基的特异性配对，脱氧核糖核苷酸(dNTP)被掺入到引物的生长链上。链的延伸是通过引物生长端的 3′羟基与被掺入脱氧核糖核苷酸的 5′磷酸基的反应形成磷酸二酯键，在总体上看，链是从 5′→3′方向延伸的。

双脱氧测序法利用了DNA聚合酶能从双脱氧核糖核苷酸(ddNTP)为底物的特性。当ddNTP被掺入到延伸着的引物的 3′端时，由于链上 3′羟基的缺如，链的延伸就终止于G、A、T或C。在 4 个测序反应中，每个反应只需各加入 4 种可能的ddNTP中的一种，就将产生如图 4.1 所示的 4 个序列阶梯。调整每个测反应中的ddNTP与dNTP的比例，使引物的延伸在对应于模板DNA上的每个可能掺入ddNTP的位置都

有可能发生终止。以这种测序方式，每个延伸反应的产物是一系列长短不一的引物延伸链，它们都具有由退火引物决定的固定的 5′端以及终止于某一ddNTP的不定的 3′端。

图 4.2 中介绍了两种双脱氧测序的工作方案。最早期的双脱氧法，本章称之为 Sanger 法，是利用大肠杆菌 DNA 聚合酶 I 大片段(或称 Klenow 片段，Klenow 酶)发展起来的。“ 标记／终止法 ”则利用了一种修饰的 T7DNA 聚合酶，在两个独立的反应中分别进行引物的标记和双脱氧核苷酸的掺入终止。引物与模板退火后，标记反应发生在 4 种低浓度 dNTP(其中 1 种是放射性标记)中，DNA 的合成持续到一种或多种 dNTP 被耗竭为止，这样可保证掺入全部的标记的脱氧核糖核苷酸。链终止反应在 4 个独立的反应中进行，每个反应除了含有 4 种 dNTP 外，还各含 4 种 ddNTP 中的一种,而高浓度的 dNTP 保证 DNA 逐次合成至生长链因 ddNTP 的掺入而终止。

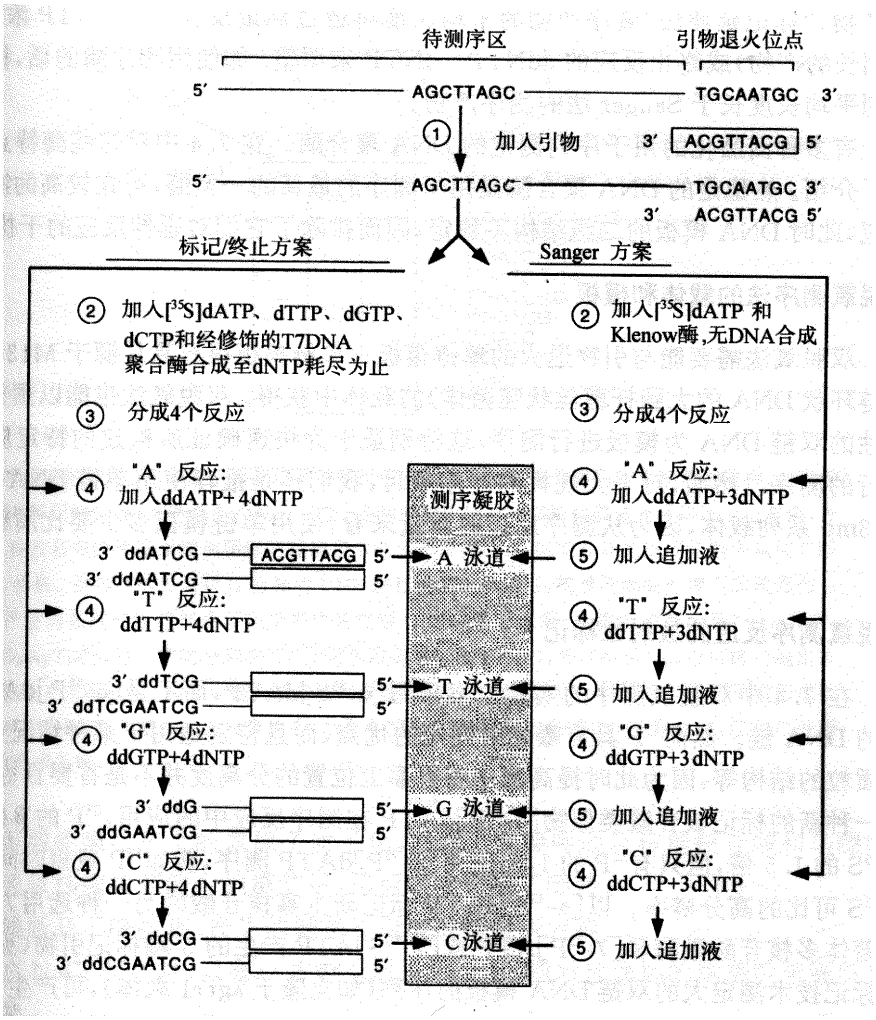


图 4.2　双脱氧测序法。在图示的每种方法中，单链 DNA 片段与引物退火后进行聚合反应(步骤 1),在 Sanger 法中((右图),加入 Klenow 酶和放射标记的 dATP(步骤 2),然后,分成 4 份进行反应(步骤 3),分别加入其余的 3 种 dNTP 和加入 ddATP、ddTTP、ddGTP 和 ddCTP 其中的一种(步骤 4)。DNA 合的成进行至摄入 ddNTP 后被终止。追加 dNTP(步骤 5)使未被终止的链再延伸以产生更高分子量的 DNA。“ 标记／终止法 ”(左图)说明略。在每种方法中，反应终止后，样品加样于测序胶的相邻泳道上，进行电泳分离（奥斯伯等，1998)。

Sanger 法测序产物的平均链长取决于 ddNTP：dNTP 的比例，比例高时，得到较短的产物；“ 标记／终止法 ” 测序产物的平均长度可通过标记反应中 dNTP 浓度(高浓度能得到长的产物)或终止反应的 ddNTP:dNTP 来调整。

有多种商品化的用于序列测定的 DNA 聚合酶。热稳定的 DNA 聚合酶是用于测序的最新的一类酶，可在高的温度进行测序反应。此时 DNA 模板的二级结构不稳定，因而排除了它们对延伸反应的干扰。

3．化学测序法(Maxam-Gilbert 法)

在A．Maxam和W．Gilbert(1977)发展的DNA化学测序法中，与碱基发生专一性反应的化学试剂在一种或两种特定核苷酸位置上随机断裂已纯化的 3′端或 5′端标记DNA链，产生 4 套寡聚脱氧核糖核苷酸。在随后的测序胶放射自显影中，仅末端标记的片段显迹，故可得到如图 4.3 所示的 4 种DNA阶梯。

肼、硫酸二甲酯(DMS)或甲酸可以专一性地修饰 DNA 分子中的碱基，这构成了化学测序法的基础，加入吡啶可催化 DNA 链在这些被修饰核苷酸处断裂。化学法的特异性基于第 1 步反应中肼、硫酸二甲酯，或甲酸仅与 DNA 链上小部分特定碱基的作用，而第 2 步的哌啶断裂必须定量反应。第 1 步反应的化学机制如下：

G 反应： DMS 使鸟嘌呤的 7 位氮原子甲基化，其后断开第 8 位碳原子和第 9 位氮原子间的化学键，哌啶置换了被修饰鸟嘌呤与核糖的结合。

G+A 反应： 甲酸使嘌呤环上的氮原子质子化，削弱了腺嘌呤脱氧核糖核苷酸和鸟嘌呤脱氧核糖核苷酸中的糖苷键，然后哌啶置换了嘌呤。

T+C 反应： 肼断开了嘧啶环，产生的碱基片段能被哌啶所置换。

C 反应： 在 NaCl 存在时，只有 C 才能与肼发生反应，随后被修饰的胞嘧啶被哌啶置换。

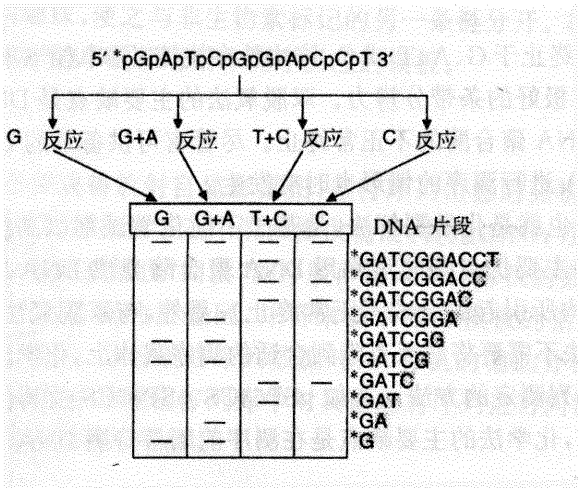
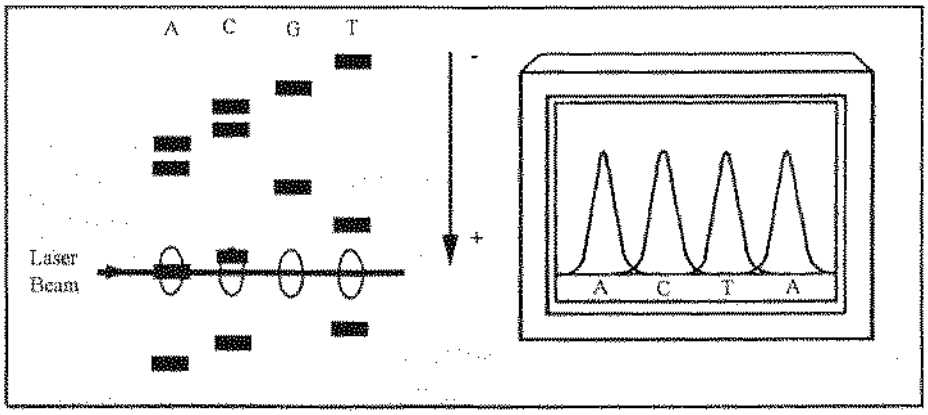


图 4.3 化学测序的策略。图中表示四个化学裂解反应产物经凝胶电泳分离后的寡核苷酸阶梯。“ * ”表示DNA片段上^{32}P标记的位置。本例是在片段的 5′端。凝胶右侧的片段 3′端加阴影的碱基表示经化学修饰后，在哌啶介导的链间切割中从核苷酸链上被取代的碱基（奥斯伯等，1998)。

4．荧光自动测序仪

自动化测序仪使凝胶电泳、DNA 条带检测和分析过程全部自动化。目前，所

有的商品化 DAN 自动化测序仪的设计都是以酶法(即 Sanger 法)测序反应产生荧光标记或放射性标记的测序产物为基础，它们都具有数据收集的能力，并含有进一步分析处理的程序。荧光标记物通过引物或 ddNTP 掺入到测序产物中。4 种碱基产生 4 种颜色的荧光反应，所以以单泳道或毛细管电泳就可以分辨出相应的寡核苷酸产物。



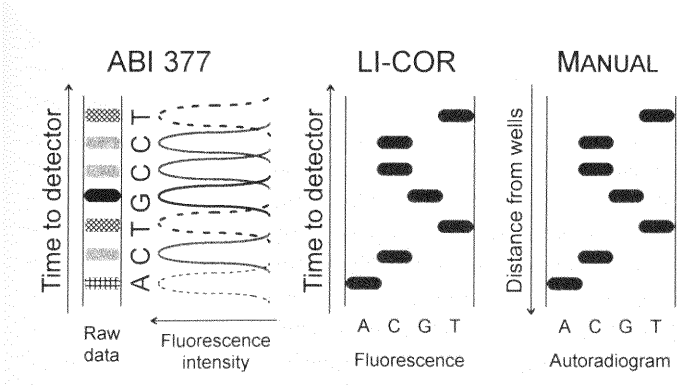


图 4.4 全自动测序仪基本操作原理

下面结合两种型号的 DNA 自动测序仪介绍自动测序原理。

ALF全自动激光荧光DNA测序系统（automated laser fluorescent DNA sequencer）是由德国海德堡（Heidelberg）欧洲分子生物学试验室（EMBL）W. Ansorge和B. Sproat提出和设计的。与同位素测序系统相比，ALF不但在仪器硬件设计上，而且在驱控仪器的软件功能上都作了很大改进。操作中能直接分析原始数据，也可以及时处理收集过程中获取的数据。最近推出的ALF express™全自动激光荧光核酸测序仪，则是利用电泳原理把荧光标记的DNA片段通过测序胶电泳分离。该仪器本身设计独特，提供快速可靠的核酸测序、片段分析、HLA序列定型及突变检测等。在人类基因组大规模序列测定中，该设备起到了重要的初筛作用。ALF express™系统采用非放射性的单一Gy5 荧光素标记引物或dNTPs进行核酸测序和片段分析，沿用Sanger双脱氧核酸末端终止测序法，使用Cy5 荧光标记的引物与模板进行退火。测试时，把A、C、G、T四种反应物分别加到凝胶板上的样品槽内，上样程序与手工测序相同。另外，在仪器电泳单元的下方是由激光枪(laser source)和探测器排列组成的探测系统：每个样品道后面都有一个探测

器，激光能透过凝胶的每一条泳道，当DNA条带迁移到探测区域并遇上激光时，DNA上的荧光标记立刻被激活，放出光信号；此荧光信号由泳道前的光探测器接收，并将信息输送给电脑进行分析和保存(图 4.4)。电泳结束后，电脑将收集到的信号(原始数据)进行处理，从而获得最终序列。

早在 1987 年 Perkin Elmer(PE)Applied Biosystems 公司就推出 DNA 自动测序仪，其专利是分别采用 4 种荧光染料进行标记且在同一个泳道测序，具有极大的优越性。377 型全自动 DNA 测序仪是 PE 公司近年推出的新型测序仪，它采用专利的四种荧光染料标记，并采用激光检测方法，具有测序精确度高、每个样品判读序列长(700bp)、一次电泳可测定样品数量多(64 个)、不需要同位素测序，方法灵活多样等特点，在人类基因组测序和 cDNA 文库测序研究中应用极其广泛。此外，该仪器在各种应用软件的辅助下还可以进行 DNA 片段大小分析和定量分析，应用于基因突变分析 SSCP、DNA 指纹图谱分析、基因连锁图谱表达水平的研究，有着极其广泛的应用前景。其原理是采用四种荧光染料标记终止物 ddNTP 或引物，经 Sanger 测序反应后，产物 3′端(标记终止物 ddNTP 法)或 5′端(标记引物法)带有不同荧光标记，一个样品的 4 个测序可以在一个泳道内电泳，从而降低了测序泳道间迁移率差异对精确性的影响。由于增加了一个电泳样品的数目，可一次测定 64 个或更多样品。经电泳后各个荧光谱带分开，同时激光检测器同步扫描，激发出的荧光经光栅分光后打到 CCD 摄像机上同步成像。也就是代表不同碱基信息的不同颜色荧光经光栅分光，经 CCD 成像，因而一次扫描可检测出多种荧光，传入电脑。其测序速度高达 200bp/h，比 373 型 DNA 测序仪速度大大提高。最后经过软件分析后输出结果。

自动化测序仪的发明促进了人类基因组的大规模测序行动。自动化测序效率高，而且测序的质量也比手工操作好。由于 DNA 多聚酶和荧光底物的不断更新，在很长一段时间内，荧光自动化测序将会处于主导地位。

二．DNA片段测序策略[2]

1．鸟枪测序法(shotgun sequencing)

大分子 DNA 被随机地“敲碎”成许多小片段，收集这些随机小片段并将它们全部连接到合适的测序载体；小片段测序完成后，根据重叠区计算机将小片段整合出大分子 DNA 序列。这就是所谓的鸟枪测序法(见图 4.6)。鸟枪测序法可以迅速获得 90%左右的片段序列结果，但随后测序效率明显下降，这是因为随后测定的随机片段越来越多地是重复已测序完成的片段。因此，一般通过合成特定的寡核苷酸引物来测定剩余少量未知片段。

有三种方法可用来将DNA大片段切割成小片段：限制性内切酶、超声波处理和DNA酶I降解(加Mn^{2+})。在这三种方法处理前， DNA的纯化非常重要，要去除载体DNA或仅由载体DNA产生的片段。

[2]本部分内容译自Alphey L. DNA Sequencing—From Experimental Methods to Bioinformatics, BIOS Scientific Publishes Limited , 1997

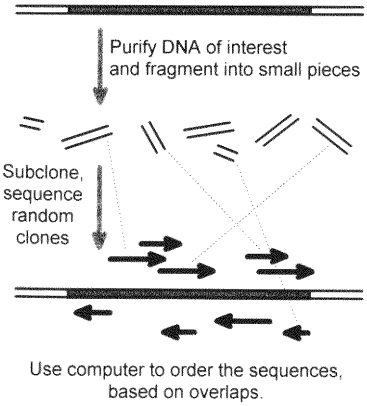


图 4.6 鸟枪法测序过程

鸟枪测序法的优点是成本低、快速、易于自动化操作，它的缺点是在测序后期，大量重复测序使测序效率变低。

1995 年第一个细胞有机体——流感嗜血(*Haemophilus influenzae*)全基因组序列被完成，这是完全用鸟枪法策略直接完成的，说明鸟枪法用于微生物基因组测序是有效的。研究者直接将全基因组 DNA 打成 1.6～2.0kb 大小的片段分别克隆，共使用了 19687 个模板，进行了 28443 个测序反应，组建了 140 个片段重叠群，测序用时 3～4 个月，耗费 100 万美金左右。

2．引物步查法(Primer walking)

引物步查法是一种渐进式测序策略，也是最简单的一种测序策略。该方法适合于双脱氧测序，并绕开了亚克隆小片段DNA的要求。最初的序列数据是通过利用载体上的引物获得的，一旦新的序列被确认，与新获得序列的 3′端杂交的寡核苷酸就能合成，并能以之为引物进行下一轮的双脱氧测序反应。这样，从两头向中间，序列被一步步测序(见图 4.7)

引物步查法相对较慢，因为序列仅从两头测得。每一步均需要一个测序反应(凝胶电泳)、数据分析、新引物设计和合成。这些过程将至少需要几天时间，如果引物供应不畅，可能时间还要更长。该方法适合于短 cDNA 片段，不适合于长 cDNA 片段，同时不宜自动化处理，因为每一反应需要一个不同的引物，这些引物将依据上一次反应结果而定。引物步查法成本相对较高，每一步都需要合成一个新引物，这制约了该技术的广泛应用。但是，最近寡核苷酸合成的成本已显著下降，所以成本问题有望解决。该技术的优点在于它的简单，不需要亚克隆或其它一些操作，实际操作时间不多，在其测序过程中，分析者有大量时间可以干其它一些事情。

引物步查法将合成一套覆盖整条序列的测序引物，如果序列需要重复测序，如测定序列突变等位位点，这套引物则成为很有用的资源。

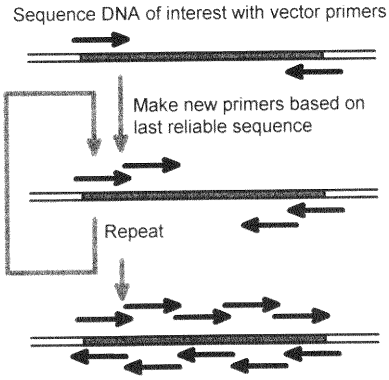


图 4.7 引物步查法测序过程

3．限制性酶切—亚克隆法(Restriction endonuclease digestion and subcloning)

原理上讲，序列的信息可以从其已知的限制性内切酶位点中获得。用限制性内切酶酶切并亚克隆一个适当大小的片段，使酶切位点附近的未知片段与载体已知序列相邻，这样就可以用载体的引物去测定未知序列；可以很方便地利用 2 个或更多位点切除一个未知克隆片段并用 DNA 聚合酶再将酶切下来的克隆产物再接合上去。由于所选用的内切酶不可能产生粘性末端，所以正常情况下，有必要用 Klenow 或 T4DNA 聚合酶把它们转变为平端。该方法示意图见图 4.8.

该方法的关键一步是需要一张准确的限制性内切酶谱，而且这些酶切位点间最好都相隔几百个碱基。对于一个熟练的研究者来说，制作一张酶切图并不难，但是酶切位点的分布则是一个随机问题，所以，不可能位点距离总是符合该方法的测序。利用该方法可以得到整条片段的大部分序列。由于该方法是基于酶切图，所以对于尚有哪些缺口(gap)，缺口有多大都很清楚，这有助于进一步的分析。

该方法难以自动化分析，因为它依赖于一套特定的亚克隆过程，而这些过程在每次的测序计划均是不同的。可能最常用的方法是用未知片段中的少量酶切位点，每个位点作为未知片段的一个新起点，然后用引物步查法在每个方向进行测序。这种混合方法较单用引物步查法可以显著减少整个片段的测序时间。

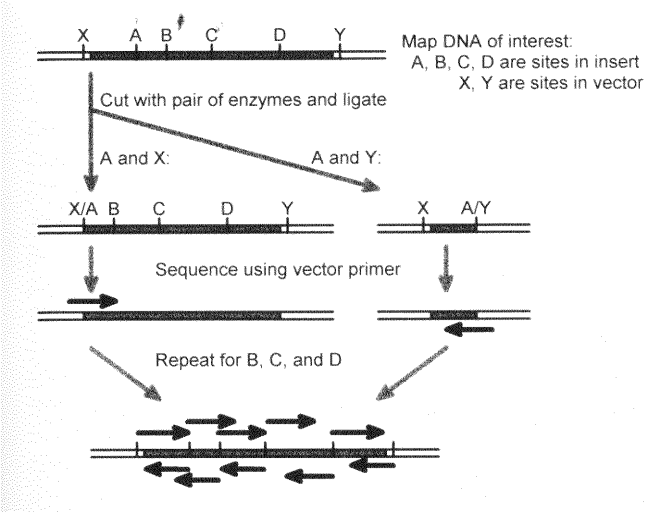


图 4.8 限制性酶切—亚克隆法测序过程

三、基因组测序策略

1．逐步克隆（clone by clone）：从遗传图谱、物理图谱到基因组图谱[3]

基因组测序涉及 DNA 的大规模测序，它是一项如同“曼哈顿登月计划”一样的庞大工程，是人类在现有技术水平的重重障碍中科学技术的又一次进步。根据现有的技术水平，人类还无法对基因组这样的复杂 DNA 大分子直接进行测序，而只能采取分而治之的测序基本策略，即将基因组 DNA 分割成一定大小的片段，然后分别对这些片段进行测序。这样便产生了这样一个问题：如何将这些片段准确地拼接起来？目前的测序方法(上节)每次反应只能测定 500bp 左右长度的 DNA 片段，而一般一条染色体的长度对于 400-500bp 长度如同天文数字。所以，要进行诸如人类基因组测序，则必须在 2 个方面取得突破：一是将基因组 DNA 大分子分割并构建适合于测序的 DNA 片段库，而且库中的片段要覆盖整条序列；二是在整条线性序列上建立一定数量的“路标”，使切割下来的 DAN 片段能准确拼装回去。遗传图谱和物理图谱便是这样的“路标”图。人类遗传和物理图谱于 1998 年的建成使最终人类基因组测序成为可能。

基因组上的 DNA 相当稳定，因此可以构建含有这些 DNA 片段的新生物体。克隆技术是把基因组上的片段插入不同生物载体，并转染到一些生物体中使其生存和稳定复制，由此可以分析由小片段 DNA 组成的基因组拷贝(克隆群)。目前选用插入的载体包括酵母、细菌、粘粒、噬菌体等。

遗传图谱(genetic map)又称连锁图谱(linkage map)或遗传连锁图谱(genetic linkage map)，是指基因组内基因和专一的多态性DNA标记(marker)相对位置的图谱，其研究经历了从经典的基因连锁图谱到现代的DNA标记连锁图谱的过程。构建遗传图谱的基本原理是真核生物遗传过程中会发生碱数分裂，此过程中染色体要进行重组和交换，这种重组和交换的概率会随着染色体上任意两点间相对距离的远近而发生相应的变化。根据概率大小，人们就可以推断出同一条染色体上两点间的相对距离和位置关系。正因为如此，我们得到的这张图谱也就只能显示标记之间的相对距离。我们称这一距离(概率)为遗传距离(cM)，由此构建的图谱也称为遗传图谱。遗传图谱的“路标”(遗传标记)已经历了几次从“粗”到“细”的大的演变，或者说，从第 1 代标记向第 2 代、第 3 代标记的过渡。经典的遗传标记(第 1 代标记)最初主要是利用蛋白质或免疫学等的标记，70 年代中后期建立起来的限制性片段长度多态性(RFLP)方法成为第 1 代的DNA标记，这类标记在整个基因组中确定的位点数目可达 10^5以上。第 2 代标记为可变数量串联重复序列(Variable number tandem repeat,VNTR)，包括微、小卫星(microsatellite/minisatellite)或短串联重复(short tandem repeat,STR或short sequent length polymorphysm,SSLP)标记等。第 3 代标记是一类称作SNP(single nuleotide polymorphysm)的遗传标记系统，即单核苷酸多态性标记。

遗传图谱上的各种DNA标记正如地图上标明的河流、山川，基因组中的这些标记种类繁多，随着人类基因组等计划的进行，人们不断发现一些新的标记，而且这些标记在地图上的密度也越来越高，迄今已经有好几个版本的图谱发表出来。在Internet网上的GDB(geneome database)网页上可以方便地查找到迄今已

[3]本部分内容取自陈竺、杨焕明等人的文章，见：贺林．解码生命—人类基因组计划和后基因组计划，北京：科学出版社，2000

发表的各种遗传标记(http://gdbwww.gdb.org)。

遗传图谱的构建是人类基因组研究必不可少的一步，它对搞清基因的功能、定位及分离克隆新基因、排列 DNA 片段、研究染色体上基因的排列顺序等起到不可估量的作用。遗传图谱在过去几年的人类基因组研究中发挥了巨大的作用，以致同样的策略也被应用于其它模式生物。

物理图谱是描述位于染色体上的基因和生物学界标独特并有确定位置及实际距离的染色体结构。任何图谱都是一系列路标及客观物(objects)按其固有的顺序和可能的距离构建出来的。客观物的顺序应不随构图方法的不同而不同，但它们之间的距离则可能不一致。在遗传图谱中按重组率来估计实际距离会有很大的偏差。物理图谱可以理解为用物理学方法而不是遗传学方法定位的由客观物组成的任何图谱，而通常物理图谱是指高分辨率(high-resolution)的物理图谱，即基因组长片段限制性酶切图谱和重叠克隆图谱等，但整合物理图谱还应包括只能粗略分辨路标位置但不能准确排位的染色体图谱(chromosome map)和遗传连锁图谱。

人类基因组测序的开展还得益于另一项突破：随着脉冲场电泳技术(pulsed-field gel electrophoresis,PFGE)、YAC 克隆、BAC 和 PAC 克隆的出现，可以把切割基因组后产生的大片段 DNA 准确地分离和纯化，并插入能转入 DNA 大片段的载体，转染酵母细胞形成 YAC 克隆库或转染大肠杆菌形成 BAC 克隆库。这些载体可载入 10Mb 长度(相当于人类全基因组碱基长度的 1/300)的 DNA 片段。全基因组的 YAC 克隆库及 BAC 克隆库保证了基因组分析的完整性和准确性。可以用杂交技术等来发现重叠克隆，以此进行克隆片段的排序。对于大片段 DNA 克隆进行再切割，并载入粘粒、细菌或噬菌体，即可构建相应于特定 YAC 或 BAC 克隆的亚克隆(subcloning)，供测序使用。这一系统过程的建立为大规模测序打下了坚实的基础。

构建物理图谱最终是要统一到基于 STS 的物理图谱。STS(sequence-tagged site,序列标签位点)的概念首先由 Olson 于 1989 年提出，目的是建立一套人类基因组统一的生物学界标。STS 本身是随机地从人类基因组上选择出来的长度在 200～300bp 左右的特异性短序列。STS 路标的建立一般是从噬菌体 M13 上构建特定染色体克隆开始，STS 概念的提出是物理构图的一次革命，由于特定 STS 在一套基因组结构中只出现一次，统一地把相应的克隆库中的克隆进行排序变得更准确和更科学。如果两个或两个以上的克隆包含有相同的 STS，则它们之间存在重叠。基于 STS 的物理图谱的重要性在于(1)它们可用来特异地定义 YAC、粘粒或噬菌体克隆；(2)STS 可鉴定出与特定克隆存在重叠的克隆；(3)在计算机数据库中的各种物理图谱可以用 STS 这种通用语言统一起来。基于 STS 的物理图谱不但可对染色体图谱、限制性酶切位点为路标的限制性酶切图、重叠探针杂交的 YAC 克隆片段重叠群(contig)图谱及其亚克隆重叠排序，以及新近发展的其它新方法构建的物理图谱进行整合，也可对遗传图谱、基因图谱等各类图谱进行整合，最终完成系统、统一的基因组终极图谱。最终完成的人类基因组核苷酸序列相当于 STS 密度最高的基因组物理图谱。

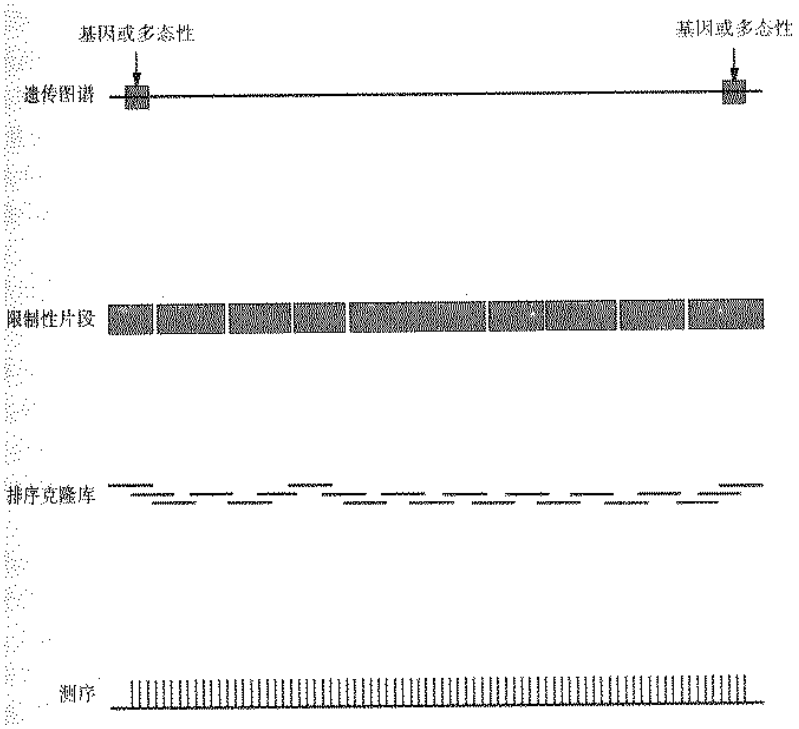


图 4.5 人类基因组的各种图谱。最粗糙的图谱是遗传图谱，它根据相邻标记（如基因和多态片段）间的重组率来测量相互间的距离；具有 1-2Mb 长度的限制性酶切片段可被分离和构建物理图谱；YAC 等长度在 40-400kb 的插入片段排列构建高分辨率物理图谱；碱基序列为最高分辨率物理图谱。

综上所述，广义上各种基于路标位点构建的物理图谱方法从低分辨率到高分辨率可主要分为以下几种：

(1)对路标进行粗略定位的染色体图谱即细胞遗传图谱(cytogenetic map)，通常使用原位杂交(ISH)或荧光原位杂交(FISH)技术确定含有路标 DNA 片段在染色体上的区带位置和分布。DNA 片段可被定在 2～10Mb 的范围内。

(2)cDNA 图谱是在细胞遗传图谱上显示 cDNA 或 ESTs(expressed sequence tags)，即表达 DNA(外因子)的区带位置。部分 cDNA 序列可作为路标。

(3)利用家系分离分析法(pedigree segregate analysis)可确定具有多态性的遗传标记位点在遗传连锁图谱上的位置，最新的人类基因组遗传连锁图谱已把标记间的平均距离缩小到 1cM 以下，即粗略地对应于物理图谱中的 1Mb 范围内。

(4)辐射杂种图谱是利用体细胞遗传技术(somatic cell genetic approach)构建高分辨率、长范围连续的人类基因组图谱。基本原理为，人为地用放射线打断染色体，制备出含有特定人类染色体或片段的杂交细胞系，并利用类似于传统的减数分裂构图原理确定路标间的距离和位置，最高的分辨率可达到 50kp。

(5)脉冲场电泳的长片段限制性位点(macrorestrietional site)图谱，即限制性酶切位点指纹(restriction enzyme fingerpriting)图谱是描述以稀有酶切位点为生物学界标的顺序和距离，以及形成基因组或染色体区域上的酶切图谱。由于些法是从大片段入手，常常又称为“从上到下”(Top-down)构图法；此外，区域性 DNA 大片段有利于较精细制图，如 YAC 克隆插入片段分析便于重叠图谱的分析，此方法可把 DNA 片段定位在 100kb 到 1Mb 范围内。

(6)由 DNA 片段重叠群(contig)形成的小组合，即相连组合图谱，或称重叠克隆群(overlapping sets of cloning)图谱描述存在于重叠的 DNA 片段克隆的顺序和距离。通常通过粘粒重叠克隆把 DNA 片段定位在小于 2Mb 的范围内，相对于长片段限制性酶切位点图谱，这种构图法也被称为“从下到上”(Bottom-up)法。

(7)序列标签位点(sequence-tagged site,STS)构成了 STS 基础上整合图。它是基因组上筛选特异序列，其最终密度至少达到平均每 100kb 左右一个，最终将把各种方法构建的图谱整合起来，完成准确完整的系统物理图谱。

(8)部分及全基因组测序是分辨率最高的物理图谱，而目前要构建的高分辨率(<100kb)物理图谱上路标序列本身也是基因组序列信息的一部分。

此外，一些构建物理图谱的方法还包括基因组序列抽样(genomic sequence sampling,GSS)和可见图谱(optical map)等。GSS 是结合片段限制性酶切和 STS 的一种作图法，分辨率可达到 1～5kb；可见图谱则是结合限制性酶切、电泳和 FISH 技术通过观察单个 DNA 大分子在限制性酶切作用下的图象来作图。

低分辨率物理图谱在人类基因组计划中本身是独立的部分，但从染色体区带－表达基因区域－遗传学距离－物理学实际距离－碱基序列这一过程来看，低分辨率染色体分带可看作粗略的物理图谱，碱基序列则是最精密的物理图谱。低分辨率图谱上的一些路标常常被用在高分辨率图谱的构建中，结合其它路标形成高密度路标分布的图谱，同时这些高密度路标可以重新在低分辩率图谱进行验证，形成高分辩率与低分辩率相结合的整合物理图谱。每种图谱都有各自的优缺点，所以即使对同一基因组研究，不同的实验室会采用不同的作图方法，但最终各种图谱的结果应能统一起来，相互补充和完善。

表 4.1 部分物种基因大小和遗传/物理距离关系

物种	拉丁名（英文名）	基因组大小（kb）	物理距离（kb/cm）
水稻	Oryza sativa (rice)	4.30×10^5	300
玉米	Zea mays (maize)	2.5×10^6	2140
小麦	Triticum aestivum (wheat)	1.6×10^7	
大麦	Hordeum vulgare (barley)	5.0×10^6	
燕麦	Avena sativa (oat)	1.1×10^7	
大豆	Glycine max (soybean)	1.2×10^6	
高粱	Sorghum bicolor (sorghum)	7.50×10^5	
马铃薯	Solanum tuberosum (potato)	8.4×10^5	
油菜	Brassica napus (rape)	1.1×10^6	
陆地棉	Gossypium hirsutum (upland cotton)	2.1×10^6	
黑麦	Secale cereale (rye)	9.1×10^6	
甜菜	Beta vulgaris subsp. Vulgaris (beet)	7.58×10^5	1100
西红柿	Lycopersicon esculentum (tomato)	9.5×10^5	510
拟南芥	Arabidopsis thaliana (thale cress)	1.20×10^5	139
洋葱	Allium cepa (onion)	1.5×10^7	
向日葵	Helianthus annuus (sunflower)	3.0×10^6	
菜豆	Phaseolus vulgaris (kidney bean)	6.3×10^5	

人	Homo sapiens	3.3×10^6	1000
小鼠	Mus musculus (mouse)	2.5×10^6	1800
大鼠	Rattus norvegicus(rat)	2.75×10^6	
线虫	Caenorhabditis elegans	9.7×10^4	250
果蝇	Drosophila melanogaster	1.37×10^5	500
大肠杆菌	Escherichia coli	4.6×10^3	
酵母	Saccharomyces cerevisiae	1.21×10^4	4.8
流感嗜血杆菌	Haemophilus influenzae	1.8×10^3	

表 4.2 基因组物理图谱数据库的部分相关网站。最新情况见附件。

数据库	网址
图谱相关资料	
STS 数据库(dbSTS)	http://www.ncbi.nlm.nih.gov/dbSTS/index.html
EST 数据库(dbEST)	http://www.ncbi.nlm.nih.gov/dbEST/index.html
CpG 数据库 (CpG island database)	http://biomaster.uio.no/CpGdb.html
细胞遗传图谱	
GDG	http://gdbwww.gdb.org
MGD	http://www.informatics.jax.org/mgol.html
辐射杂种图谱	
RHdb	http://www.ebi.ac.uk/RHdb
克隆重叠图谱	
YAC 克隆	
CEPH-GENETHON 整合图谱	http://www.cephb.fr/ceph-genethon-map.html
STS/YAC MAP	http://www_genome.wi.mit.edu/
BAC 和 PAC 克隆	http://www.tree.caltech.edu/
粘粒克隆	http://gea.lif.icnet.uk/
整合图谱	http://cedar.genetics.soton.ac.uk/public_html

表 4.2 中列举的物理图谱数据库的数据主要来自人类基因组，但同时也包含了其它的一些生物体。

构成物理图谱的 4 个基本要素之一可复制 DNA 片段(clonable fragment)(另 3 个要素是路标、单位、顺序)主要包括辐射杂种细胞(RH)、YAC、BAC、PAC 等。对于这些 DNA 大片段的测序一般需要将其再细分为能单独进行序列分析的小片段，目前有三种常用方法：鸟枪测序法、引物步查法和限制性酶切—亚克隆法。

2. 全基因组鸟枪法 (whole-genome shortgun)

在基因组水平上，全基因组鸟枪法和逐步克隆测定法是目前广泛应用的两个测序策略。小的单分子基因组，如细菌和小基因组(<10Mb)可直接用鸟枪法测序。虽然有人提出用鸟枪法直接测序人类基因组(Weber 和 Mayers，1997)，但由于人类基因组中存在高比例的重复序列(尤其是 LINE，2-7kb)、克隆文库不可避免的间隙和基因的多态性等原因，鸟枪法的片段组装几乎是不可能的。受读序长度的限制，一个反应无法跨过 LINE。鸟枪法在小组因组(1-5Mb)测序方面已取得了非常好的效果，例如流感嗜血杆菌(*H. influenzae*, 1.9Mb)、枝原体(*M. genitalium*, 0.58Mb)和甲烷球菌(*M. jannaschii*)基因组均用此法完成测序。逐步克隆测定法则通过建立克隆文库(YAC、BAC、PAC、Cosmid、Fosmicl、噬菌体、质粒)，然后用鸟枪法进行克隆片段的测序。所以，大规模测序的两个前沿基本都是采用鸟枪法(图 4.9)。

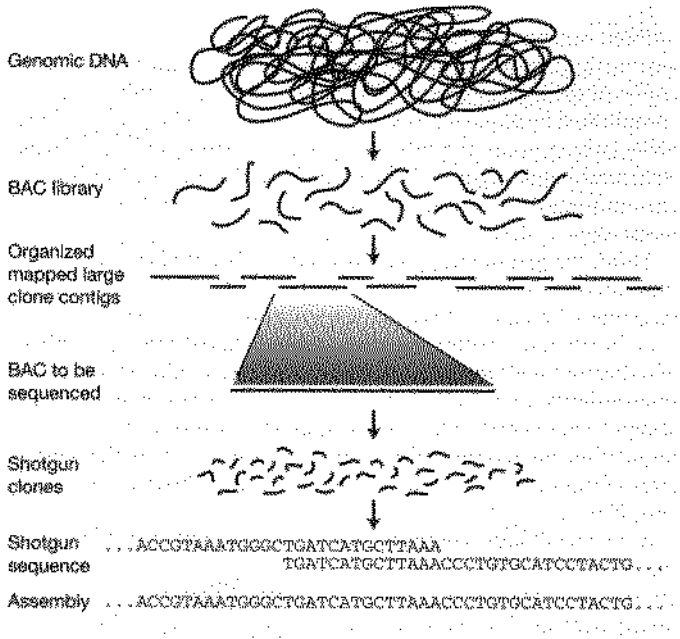


图 4.9 鸟枪法测序策略。基因组的逐步克隆测序包括图中的所有步骤：DNA 单链→构建 BAC 文库→鸟枪法克隆测序→组装；全基因组鸟枪法测序则省去中间的构建 BAC 文库步骤。

四、序列片段的拼接方法

无论是逐步克隆测序还是全基因组鸟枪法测序，都存在片段拼接组装的难题。目前 DNA 自动测序仪每个反应只能测序 500bp 左右，如何将这些片段拼接成完整的 DNA 序列呢？Lander 和 Waterman(1988)提出利用“指纹”(fingerprinting)随机克隆进行基因组作图的算法，它为大量鸟枪法随机测序的片段用计算机进行自动拼接提供了可能。这种技术不仅避免了传统的亚克隆策略的大量繁琐工作，还使测序具有一定的冗余性(即一定数量的重复)，保证了测序中每个碱基的准确性。

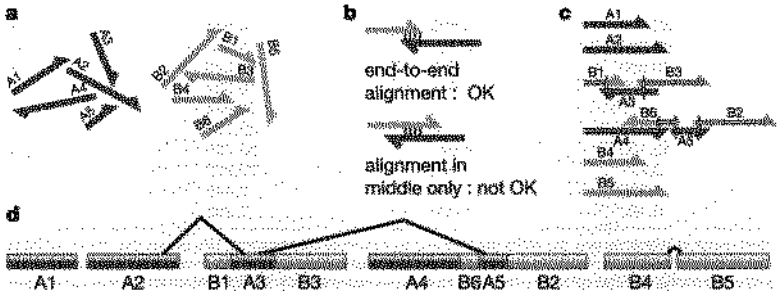


图 4.10 序列拼接示意图。a-d 步骤表示从单一克隆片段拼装成基因组草图的过程；A1-A5 和 B1-B6 分别表示由鸟枪法克隆 A 和 B 获得的序列重叠群。

目前 DNA 序列拼接应用的主要软件是由美国华盛顿大学 Phil Green 实验室

开发的 Phred-Phrap-Consed 系统。Green 也因研制该系统而在人类基因组研究历史上占有一席之地(见 *Science* 2001 年 2 月 16 日人类基因组专刊“A history of Human Genome Project”一文)。Phred(测序器)是一种碱基识别系统(base-caller)，它根据自动测序仪信号按顺序识别碱基，估计测序错误率等。Phrap(组装器)是根据 Phred 的结果从头组装由鸟枪法产生的不同的短序列。Consed(校对器)与 Phrep 组成一个有机整体，利用 Phrap 组装的序列由 Consed 编辑、整合人工校对结果等。目前 36 个国家 900 多个实验室都在使用上述系统。非赢利研究机构或个人可申请免费利用该系统。

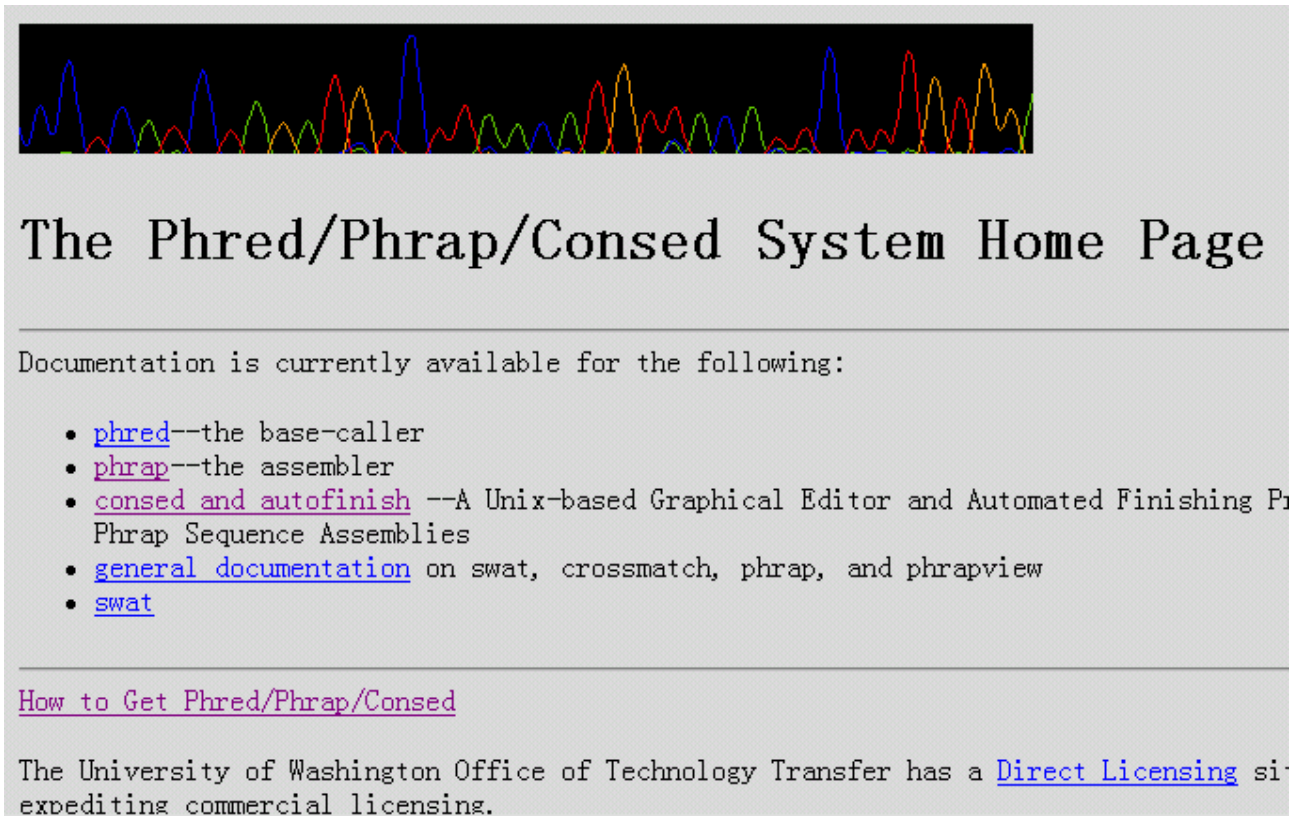


图 4.11 自动测序组装系统 Phred-Phrap-Consed 主页

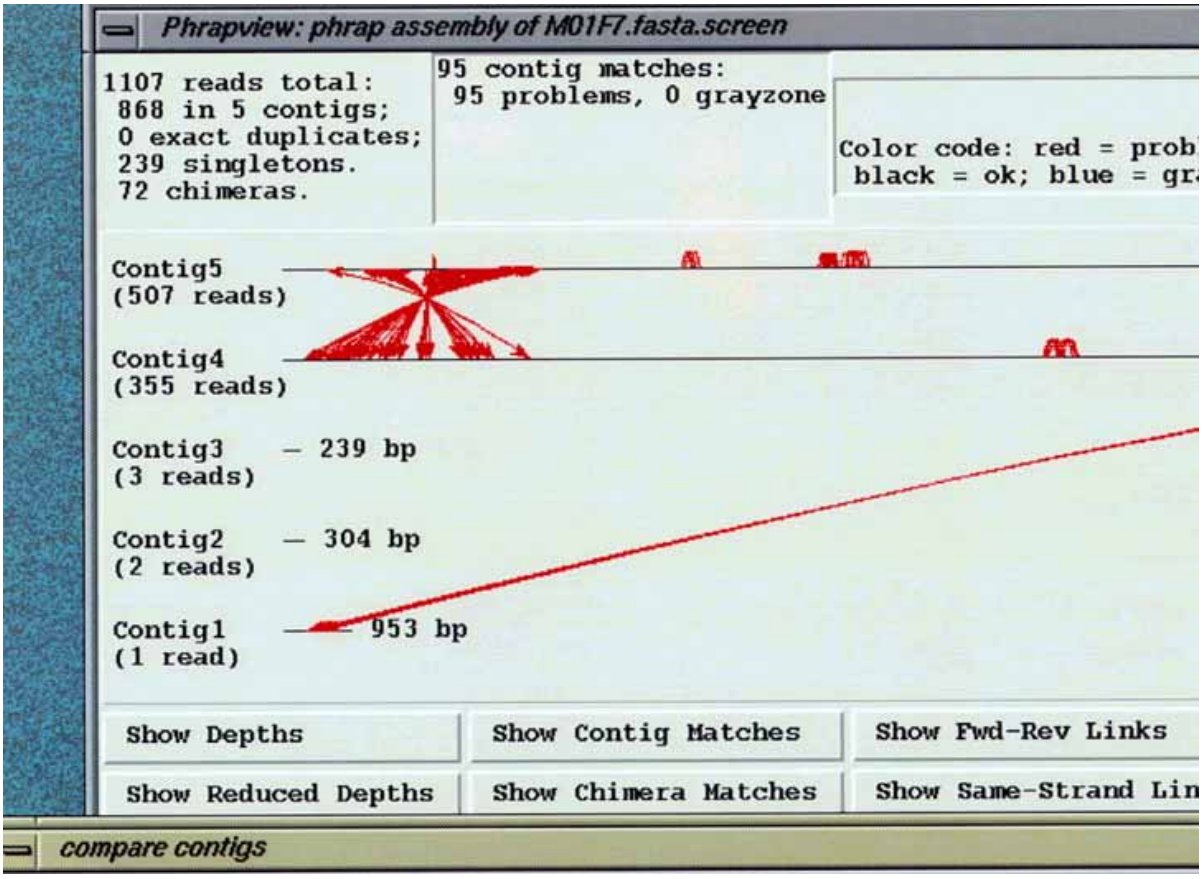


图 4.12 Phrap 程序中序列重叠群比对结果显示窗口

Phrap 拼接鸟枪法序列的的方法也是通过列线(aligment)查找匹配序列。其列线算法采用的是 Smith-Waterman 算法和 Needleman-Wunsch 算法（可选择），替换矩阵(缺省为 BLOSUM50)、空位设置罚值和空位扩展罚值(缺省值分别为-12 和-2)、E 值(缺省值 1.0)等都在列线比对中被应用。Phrap 的算法中使用了一个新参数——Z 值(Z-score)。当数据库序列长度变化很大时(实际情况往往如此)，理论分析和经验研究都表明列线值敏感性下降，即判别由随机性产生匹配的能力下降。Z 值的引入便是为了解决这一问题。Z 值定义如下：

$$Z = [s - f(n)] / \sqrt{g(n)}$$

其中 s 和 n 为原始列线值和数据库序列长度，f(n)和 g(n)分别是序列长度为 n 的序列列线值平均数和变异度。由此，Z 值的平均数为零，标准差为 1,与序列长度 n 无关。相对而言，Z 值与数据库大小无关，这一特性与原始列线值 s 相似，但与 E 值不同，所以，Z 值是比 s 值更合理的一个指标尺度。

五、EST 测序及其分析

（待补充）

第二节 基因组注释：基因区域的预测

一．从序列中寻找基因

1．基因及基因区域预测

在完成序列的拼接后，我们得到的是很长的 DNA 序列，甚至可能是整个基因组的序列。这些序列中包含有许多未知的基因，将基因从这些序列中找出来是生物信息学的一个研究热点。

基因一词最早是由丹麦遗传学家约翰逊(Johannsen W.)于 1909 年提出，而在这之前，遗传学创始人孟德尔用“遗传因子”表达了对基因的朦胧认识。基因的概念随着遗传学、分子生物学等的发展而不断完善。从分子生物学角度看，基因是负载特定生物遗传信息的 DNA 分子片段，在一定条件下能够表达这种遗传信息，产生特定的生理功能。基因按其功能可分为结构基因和调控基因：结构基因可被转录形成 mRNA，并进而转译成多肽链；调控基因是指某些可调节控制结构基因表达的基因。在 DNA 链上，由蛋白质合成的起始密码开始，到终止密码子为止的一个连续编码序列称为一个开放阅读框(Open Reading Frame,ORF)。结构基因多含有插入序列，除了细菌和病毒的 DNA 中 ORF 是连续的，包括人类在内的真核生物的大部分结构基因为断裂基因，即其编码序列在 DNA 分子上是不连续的，或被插入序列隔开。断裂基因被转录成前体 mRNA，经过剪切过程，切除其中非编码序列(即内含子)，再将编码序列(即外显子)连接形成成熟 mRNA，并翻译成蛋白质。假基因是与功能性基因密切相关的 DNA 序列，但由于缺失、插入和无义突变失去阅读框而不能编码蛋白质产物。

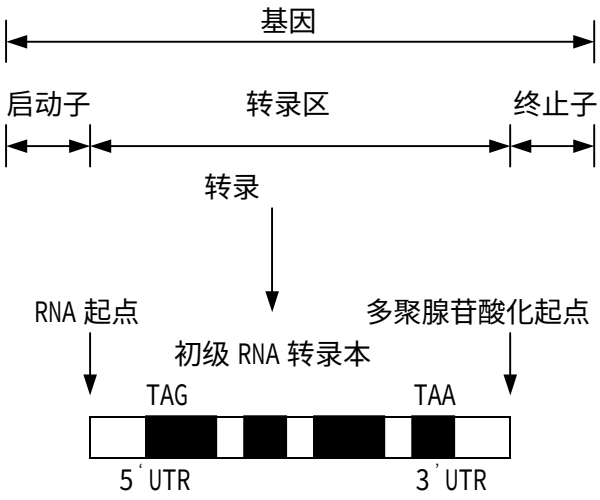


图 4.13 一种典型的真核蛋白质编码基因的结构示意图。其编码序列（外显子）是不连续的，被非编码区（内含子）隔断。

所谓基因区域预测，一般是指预测 DNA 序列中编码蛋白质的部分，即外显子部分。不过目前基因区域的预测已从单纯外显子预测发展到整个基因结构的预测。这些预测综合各种外显子预测的算法和人们对基因结构信号(如 TATA 盒等)的认识，预测出可能的完整基因。

某一算法的优劣可以通过一定的标准衡量：敏感性(sensitive)和特异性

(specifity)。假设待测序列中有M条序列是基因序列，而剩余的为非基因序列。我们用某一程序(算法)对待测序列进行预测，共预测出N条基因序列，而这N条序列中有N_1条确实为基因。则敏感性定义为N_1/M，它表示程序预测的功能；特异性定义为N_1/N，它表示程序预测结果的可靠程度。敏感性和特异性往往是一对矛盾。

基因区域的预测是一个活跃的研究领域，先后有一大批预测算法和相应程序被提出和应用，其中有的方法对编码序列的预测准确率高达 90%以上，而且在敏感性和特异性之间取得了很好的平衡。预测方法中，最早是通过序列核苷酸频率、密码子等特性进行预测(如最长 ORF 法等)，随着各类数据库的建立和完善，通过相似性列线比对也可以预测可能的基因。同时，一批新方法也被提了出来，如隐马尔可夫模型(Hidden Markov Model，HMM)、动态规划法(dynamic programming)、法则系统(ruled-based system)、语言学(linguistic)方法、线性判别分析(Linear Discriminant Analysis，LDA)、决策树(decision tree)、拼接列线(spliced alingment)、博利叶分析(Fourier analysis)等。

表 4.3 列出了 claverie(1997)对部分程序预测基因区域能力的比较结果，表中同时列出了相应算法和程序的网址。

目前基因区域预测的各种算法均基于已知基因序列。如相似性列线比较算法是完全依赖于已知的序列，而象 HMM 之类的算法都需要对已知的基因结构信号进行学习或训练，由于训练所用的序列毕竟是有限的，所以对那些与学习过的基因结构不太相似的基因，这些算法的预测效果就要大打折扣了。要解决以上问题，需要对基因结构进行更深入的研究，寻找隐藏在基因不同结构中的内在统计规律。

表 4.2 部分程序预测基因区域能力的比较结果（claverie，1997）

程序名称	所用算法	作者	预测对象	敏感性(%nucl)	物异性(%nucl)	敏感性(%exact exon)	特异(%exact exon)	丢失性的外显子(%)	错误的外显子(%)	网址
FGENEH	LDA	solovyev et al 1995	基因结构	83	93	73	78	15	11	http://dol.imgen.bcm.tmc.edu:9331/gene-inder/gf.html
GeneID	RB	Guigo et al 1992	基因结构	69	77	42	46	28	24	http://geneid@darwin.bu.edu www.imim.es/GeneIdentification/Geneid/geneid_input.html
GeneParser	DP	Snyder&Stormo 1993	基因结构	66	79	35	40	29	17	http://Beagle.colorado.edu/~eesnyder/GeneParser.html
Genie	HMM,DP	Henderson et al 1997	基因结构	87	88	69	70	10	15	http://www-hgc.lbl.gov/inf/genie.html
GenLang	LM	Dong&Searls 1994	基因结构	72	79	51	52	21	21	http://www.chil.upenn.edu/~sdong/genlang_home.html
GENSCAN	HMM,DP	Burge&Karlin 1997	基因结构	93	93	78	81	9	5	http://genomic.stanford.edu/GENSC-ANW.html
HEXON	LDA,DP	Solovyer et al 1994	基因外显子	88	80	71	65	10	27	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
MORGAN	DT	—	基因结构	83	79	58	51	14	-	http://www.cs.jhu.edu/labs/compbio/morgan.html
MZEF	—	Zhang 1997	基因外显子	87	95	78	86	14	7	http://Clio.cshl.org/genefinder
VEIL	HMM,DP	Krogh et al 1994	基因结构	83	72	53	49	19	-	http://www.cs.jhu.edu/labs/compbio/veil.html

注释：

①LDA：线性判别分析；RB：法则系统；DP：动态规划法；HMM：隐马尔可夫模型；DT：决策树；②敏感性(%nucl)：实际编码序列被成功预测为编码序列；特异性(%nucl)：预测为编码的序列实际确定为编码序列；敏感性(%exact exon)：实际的外显子被准确预测(包括拼接位点)；特异性(%exact exon)：预测为外显子的序列与实际外显子准确符合；丢失的外显子(%)：未能预测出的实际外显子；错误的外显子(%)：预测为外显子的序列实际不是任何外显子的片段。

2．发现基因的一般过程

从序列中发现基因可以理解为基因区域预测和基因功能预测 2 个层次。生物信息学在这 2 个层次上均形成具有自身学科特色的算法和手段，以下便简单描述通过生物信息学手段发现基因的一般过程。有关基因功能的预测将在以后的章节中进一步论述，同时本小节描述的发现过程只是生物信息学手段的一种可选策略。

以下主要根据 Gene Discovey(http://bioinformatics .weizmann.ac.il)：

第一步：获取 DNA 目标序列

① 如果你已有目标序列，可直接进入第 2 步；

② 可通过 PubMed 查找你感兴趣的资料；通过 GenBank 或 EMBL 等数据库查找目标序列。

第二步：查找 ORF 并将目标序列翻译成蛋白质序列

利用相应工具，如 ORF Finder、Gene feature(Baylor College of Medicine)、GenLang(University of Pennsylvania)等，查找 ORF 并将 DNA 序列翻译成蛋白质序列。

第三步：在数据库中进行序列搜索

可以利用 BLAST 进行 ORF 核苷酸序列和 ORF 翻译的蛋白质序列搜索。

第四步：进行目标序列与搜索得到的相似序列的整体列线(global alignment)

虽然第三步已进行局部列线(local alignment)分析，但整体列线有助于进一步加深目标序列的认识。

第五步：查找基因家族

进行多序列列线(multiple sequence alignment)和获得列线区段的可视信息。可分别在 AMAS(Oxford University) 和 BOXSHADE(ISREC,Switzerland)等服务器上进行。

第六步：查找目标序列中的特定模序

① 分别在 Procite、BLOCK、Motif 数据库进行 profile、模块(block)、模序(motif)检索；

② 对蛋白质序列进行统计分析和有关预测

第七步：预测目标序列结构

可以利用 PredictProtein(EMBL) 、 NNPREDICT(University of California)等预测目标序列的蛋白质二级结构。

第八步：获取相关蛋白质的功能信息

为了了解目标序列的功能，收集与目标序列和结构相似蛋白质的功能信息非常必要。可利用 PubMed 进行搜索。

第九步：把目标序列输入“提醒”服务器

如果有与目标序列相似的新序列数据输入数据库，提醒(alert)服务会向你发出通知。可选用 Sequence Alerting(EMBL)、Swiss-Shop(Switzerland)等服务器。

3．解读序列(making sense of the sequence)

在 2001 年二月份的第二星期里(12 日-18 日)，*Science* 和 *Nature* 同时刊发了具有划时代意义的人类基因组研究专刊。在 *Science* 的专刊中，有一篇题为“解读序列”(making sense of the sequence)(Galas D.J.)的综述文章。文章对序列，特别是人类基因组序列如何解读进行了深入分析，比较全面地展示了人类目前对序列的理解能力和技术现状。以下内容摘译自该篇文章。

利用基因组序列解决生物学问题已经具备了其自身(学科)特色，它被冠以“功能基因组学”。自从 1996 年酶母(*Sacharomyces cerevisiae*)基因组序列被公布，我们已熟悉用全基因组序列来研究基因表达模式等等生物学问题。虽然我们还不知道约 1/3 酶母基因的功能，但是我们知道所有与细胞功能有关的可能的蛋白质和 RNA 均由我们已知的序列编码。

根据目前对基因的分析结果，哺乳动物一个基因的转录产物平均有 2～3 种或者更多。从现有序列数据估计，人类的基因数约为 3 万，这意味着人类基因组编码了约有 9 万或更多种蛋白质。但是，以上由现有序列数据推测的结论有很多不确定因素。重叠序列群(contig)是由单个测序反应测得的序列(通常 400～800 碱基长度)拼装而成的一条连续片段，重叠序列群的数量和长度分布是基因分析的两个重要参数。正如美国 NCBI2000 年 12 月 12 日的报告所说，目前公共数据库中最大的重叠序列群为 28.5Mb，其中 43 个超过 1Mb，566 个在 250Kb～1Mb 之间，而 1628 个在 100～250Kb。这意味着长度大于 100Kb 的重叠序列群总长度约 600Mb——不足人类基因组全部序列的 20%；而基因组的一半序列是由 22Kb 或更小的重叠序列群所涵盖。由于基因的长度(一般估计为 30000 碱基对)大于或等于重叠序列群，这说明一定比例的人类基因不可能只在一个重叠群中；在一个重叠群中发现一个最长的基因，如肌联蛋白(Titin)基因(约 250Kb,内含 200 多个外显子)，比发现一个短的简单基因，如嗅感受蛋白基因(平均小于 2Kb)的概率小得多。但要将序列缺口和重叠群扩大还要籍以时日。因此，在不久的将来，基因的合成将通过组配重叠群“镶嵌物”(mosaic)，或称为“支架”(scaffold)来完成，这意味着重叠群间的拼接又将增加序列数据的不确定性。

要想将所有的基因都落入拼装而成的无缺口的支架片段中似乎还不可能，但是组装成的基因的大致轮廓将变得很清楚。这就象一个被重新复原的古希腊花瓶，虽然花瓶的残缺部分被用陶土填补，而整个花瓶的轮廓已很清晰。文特尔(Venter)等人进行基因拼装和分析的方法中，一人重要的参数是支架的大小和分布。据报道，支架的平均长度超过 1Mb，而 10Mb 以上的支架占整个基因组的 25%，支架间的缺口平均只有 2Kb。这些为基因分析者提供了高档次的序列数据。

从一给定序列片段中，通过相似性比较发现基因的效果决定于简单的统计量和重叠群在基因组中的覆盖率。当该覆盖率达到 90%以上，那就意味着几乎所有的基因(或至少是基因片段)均可在序列数据中找到。因此，利用本周公布的数据(指 *Science* 和 *Nature* 的人类基因组专刊)，通过相似性搜索来发现任何一个基因几乎都是可能的。

但是必须注意的是，这样确定的基因可能还具有随意性。这是因为某一生物，例如果蝇(*Drosophila*)的一条具有高度相似的受体基因序列可能来自几个不同的同源基因，而这些基因可能具有相同或完全不同的功能，甚至可能是一些没有功能的假基因(pseudoge)。也就是说，共同的功能域(domain)或模序(motif)可能在几个基因同时存在。使用 BLAST 搜索工具可能还是目前发现相似序列的最佳途径。NCBI 网站简明的介绍内容有助理解不断增多的 BLAST 系列工具的特性，有些小册子介绍了 BLAST 近似算法的统计特色和局限。BLAST 算法并不适合于所有目的的近似估计，但使用者应有这样的认识，即任何一种算法都有可能错过一些特殊相似性。例如，由于对一些相隔相似性(interrupted similarity)的忽略，使间隔越大，获得相似性统计显著的可能性越小。新的一些方法试图利用编码区的结构因素来进行相似性比对，这突破了相似显著性方法的局限。

虽然在基因组序列基因的自动化识别方面已取得巨大进步，但根据序列构建

准确的基因模型(model of genes)还需要大量的人力，即“ 手工操作 ”(“ hand-on ” effort)。基因的最佳模型是其全长 mRNA 序列。RNA 序列(以 cDNA 形式)可以将基因组序列基因的外显子结构串联起来，而不必考虑这些片段身处何方——片段的连续性、顺序和方向并不影响串联过程。但是，假基因和高度相同的重复序列可能使这一策略失灵，这引起了对收集更多全长 cDNA 序列数据的争论。

大致有 2 条途径可以发现基因：(1)基于同源性的方法，包括已知 mRNA 序列的应用；(2)基因家族和特殊序列间的比较。最初的方法包括利用各种计算机手段分析外显子和其它序列信号，如酶切位点等。

在每一个基因模型中，与调控相关的序列位置和结构往往是最难完成的注释(annotation)之一。在一些情况下，可以通过诸如模序(motif)(检索)来寻找和鉴定这些重要序列区段，但是我们目前对调控区段的鉴定和预测能力还很有限和不可靠。特定基因组间的比较是获得这些区段的一条途径，它建立在可以通过比较找出保守区的假设基础上。新的一些实验方法，例如列阵技术可以定位基因组水平的转录位点，同样可以有效地检测出基因组顺式调节(cis-regulatory)信号。

目前已有很多工具可以用于自动注释工作，对于这些工具的特点本文不做进一步论述。将统计学和启发式机器学习方法(heuristic methods) 相结合来分析基因和基因特征是目前流行的趋势(例如隐马尔可夫模型、神经网络和贝叶斯网络)。它们在发现基因方面最有效的方法并不是在准确建模方面，而是常与同源性方法配合使用。影响这些算法有效性的因素包括测序误差和统计偏差，例如碱基组成。数据的噪音(noise)会极大降低这些方法的效果，所以以上基于误差率较高的序列草图的预测结果将明显劣于基于完成序列的预测。

GENSCAN(http://genes.mit.edu/GENSCAN.html)是被广泛用于基因查寻和预测的软件之一，但是一些新软件，如 Genie 也不逊色。Genie (http://www_hgc.lbl.gov/inf/genie.html)是一种隐马尔可夫模型(HMM)系统，它可以整合不同来源的信息，如信号传感器(酶切位点、起始密码等)、内含子和外显子、mRNA EST 的列线和肽序列等。其它软件工具，如 GENEBUILDER、GLIMMERM、FGENES、GRAIL 等，最近也都被评价过。有一个简单的办法可以比较这些软件的优劣：利用果蝇基因组数据为例，GASPI 项目(Genome Annotation Assessment Project)(www.fruitfly.org/GASPI)对真核生物基因组注释的进展和存在的问题进行很好的比较分析。另外利用拟南芥(Arabidopsis)基因组也进行了相同的比较分析。

*Nature*和 *Sciece* 上的两篇人类基因组分析论文分别使用了各自的基因分析系统。由公共资金资助的人类基因组计划(IHGSC)(论文发表在 *Nature* 上)使用的是一个称为“ Ensembl ”的系统，它使用 GENSCAN 进行初步预测，GENSCAN 利用 mRNA、EST 和蛋白质模序信息进行比对；然后使用 GeneWise(www.Sanger.ac.uk/software/Wise2/)进行蛋白质匹配分析，GeneWise 曾被用于果蝇基因组分析。以文达尔(Venter)为代表的私人公司(论文发表在 *Science* 上)使用的是一种称为“ otto ”的专家注释系统(rule-based expert system for annotation)，该系统力图将人的一些智能纳入程序中。

二、最长 ORF 法等：基于编码区特性

基因区域或蛋白质编码区的识别，特别是对高等真核生物基因组 DNA 序列中编码区的识别仍未能实现完全自动化。将每条链按 6 个读框全部翻译出来，然

后找出所有可能的不间断开放阅读框(ORF)往往有助于基因的发现。预测基因组的全部编码区或称为开放阅读框的方法概括来说也可以分为三类：一类是基于编码区所具有的独特信号，如始起密码子、终止密码子等；二是基于编码区的碱基组成不同于非编码区，这是由于蛋白质中 20 种氨基酸出现的概率、每种氨基酸的密码子兼并度和同一种氨基酸的兼并密码子使用频率不同等原因造成的；三是通过同源性比较搜寻蛋白质库或 dbEST 库寻找编码区。前二类方法主要是利用编码区的特性来寻找，本小节对这二类方法做简单描述。

最长 ORF 法：

在细菌基因组中，蛋白质编码基因从起始密码 ATG 到终止密码平均有 100bp，而 300bp 长度以上的 ORF 平均每 36Kb 才出现一次，所以只要找出序列中最长的 ORF(>300bp)就能相当准确地预测出基因。

在真核生物中，全长 cDNA 的编码区一般也可以用最长 ORF 法，如水稻的 3 万多条的全长 cDNA 的编码区预测（见 KOME DATABASE)。但是，要十分小心的是，这一预测有时也会出错。例如：以下全长 cDNA 的编码蛋白序列应为 4-029B，而非最长的 4-029A。

>4_029

ATCGGCCATTACGGCCGGGGACACAACAAACCAACAAACATCATAATTAACCTCTTCCTCCCAAGTAGT
CATCTGCCAACATGAAAGCCCTCGCACTCTTCTTCGTACTTTCCCTCTATCTCCTCGCCAACCCAGCTC
ATTCCAAGTTCAATCCCATCCGCCTCCGCCCCGCCCACGAAACGGCGTCGTCCGAAACTCCGGTGCTCG
ACATCAACGGCGACGAAGTCCGGGCCGGCGAAAATTACTACATTGTCTCCGCCATATGGGGCGCCGGCG
GAGGAGGCCTGAGACTCGTCCGATTGGATTCCTCCTCGAACGAATGCGCCAGCGACGTGATCGTATCCC
GGAGCGACTTCGACAACGGCGACCCGATTACCATCACGCCGGCGGACCCGGAATCCACCGTCGTCATGC
CGTCGACGTTCCAGACCTTCAGATTCAACATTGCGACCAACAAACTCTGCGTAAACAACGTAAACTGGG
GGATCAAGCACGACAGTGAATCCGGGCAATATTTCGTGAAAGCCGGCGAGTTCGTCTCCGACAATAGCA
ACCAGTTCAAGATTGAGGTGGTCAACGACAACCTTAACGCTTACAAAATCAGTTATTGTCAGTTCGGCA
CCGAGAAATGCTTCAACGTTGGCAGATACTACGACCCGTTGACCAGGGCTACGCGTTTGGCTCTCAGTA
ATACTCCCTTCGTGTTTGTGATCAAACCTACTGATATGTAATGAGCACCGGTGTTGAGGTTGCATGCAT
GTTATGGAGCTATGCTAAATAAGTAACGTTGCAACTTTGACAACGTTGTACGTGTAATAATAAGAATAA
ACATGCAATAAATCCGAGCTTGTTGTGTTGTGTAAATTTAACTATCTTAAATGAATAAGCATAATATTA
TCTATGCGAAAAAGAAAAAATAATAAAAAAAATTCATGTTCCGCCGCCTCGGCCCAGTCAACTCTGAAT
CCAAGCAAGCTTATGCATGCGGCCCAAATTCAAGCTCAATTGGCCAATTCGCCTATAGGGAGTCGTATT
ACATTCATGGCCGTCGTTTTACACGTCGGGACTGGGAAAACCCTGGGGTTACCCAACTTATCCCCTTGG
GCCCATTCCTCC

>4_029A ORF:69..755 Frame -2 Most length 687

MQPQHRCSLHISRFDHKHEGSITESQTRSPGQRVVVSANVEAFLGAELTITDFVSVKVVVDHLNLELVA
IVGDELAGFHEILPGFTVVLDPPVYVVVYAEFVGRNVESEGLERRRHDDGGFRVRRRDGNRVAVVEVAPG
YDHVAGAFVRGGIQSDESQASSAGAPYGGDNVVIFAGPDFVAVDVEHRSFGRRRFVGGAEADGIELGMS
WVGEEIEGKYEEECEGFHVGR

>4_029B ORF:81..731 Frame +3 second length 651

MKALALFFVLSLYLLANPAHSKFNPIRLRPAHETASSETPVLDINGDEVRAGENYYIVSAIWGAGGGGL
RLVRLDSSSNECASDVIVSRSDFDNGDPITITPADPESTVVMPSTFQTFRFNIATNKLCVNNVNWGIKH
DSESGQYFVKAGEFVSDNSNQFKIEVVNDNLNAYKISYCQFGTEKCFNVGRYYDPLTRATRLALSNTPF

VFVIKPTDM

利用编码区与非编码区密码子选用频率的差异进行编码区的统计学鉴别方法：由于内含子的进化不受约束，而外显子则受到选择压力，因此内含子的序列要比外显子更随机。这是目前各种预测程序中被广泛应用的一种方法，如GCG(Genetic Computer Group 研制，一种通用核酸、蛋白质分析软件包)的TestCode、美波士顿大学 GeneID 和 Baylor Medcine College 的 BCM Gene Finder等程序均利用了这一方法。具体方法描述可参阅相关程序说明。

CpG岛：CpG岛(CpG island)一词是用来描述哺乳动物基因组DNA中的一部分序列，其特点是胞嘧啶(C)与鸟嘌呤(G)的总和超过 4 种碱基总和的 50%，即每 10 个核苷酸约出现一次双核苷酸序列CG。具有这种特点的序列仅占基因组DNA总量的 10%左右。从已知的DNA序列统计发现，几乎所有的管家基因(House-Keeping gene)及约占 40%的组织特异性基因的 5′末端含有CpG岛，其序列可能包括基因转录的启动子及第一个外显子。因此，在大规模DNA测序计划中，每发现一个CpG岛，则预示可能在此存在基因。另外，AT含量也可以作为编码区的批示指标之一。

三、序列相似性比较法

近年来相似比较算法也被应用于预测可能存在的基因。这一方法之所以可以预测新基因，主要有以下几个原因：

(1)大约已经有 50%的基因有了对应的 EST，已知的蛋白质序列也越来越多；

(2)不少原核生物和酶母的全序列已经测定。研究表明有将近一半的脊椎动物基因可以通过 BLAST 在酶母、细菌和线虫的序列数据库中找到相似性相当高的序列；

(3)大多数EST 都采用每个克隆分别从 5′和 3′测序，克服了早期EST只代表 3′外显子的缺点。

许多基因预测的程序都已经整合了同源比较算法。

下面举例说明如何通过人类 EST 数据库搜索和拼接与已知基因高度同源的人类新基因：

①以已知基因 cDNA 序列对 EST 数据库进行 BLAST 分析，找出与已知基因 cDNA 序列高度相似的 EST；

②用 Seqlab 的 Fragment Assembly 软件构建重叠群，并找出重叠群的一致(consensus)序列；

③比较各重叠群的一致序列与已知基因关系(图 4.14)。通常有两种情况，一是 EST 足够多，可形成一个覆盖全长的重叠群，以此拼接基因全长序列；另一情况则是，EST 形成几个重叠群，所以可以拼接基因的几段序列。

④对编码区蛋白质序列进行比较，并与已知基因蛋白质的功能域(domain)进行比较分析，推测新基因的功能。

⑤用新基因序列或 EST 序列对 STS 数据库进行 BLAST 分析，如果某一 EST(非重复序列)与某一 STS 有重叠，那么，STS 的位置即确定了新基因的定位。

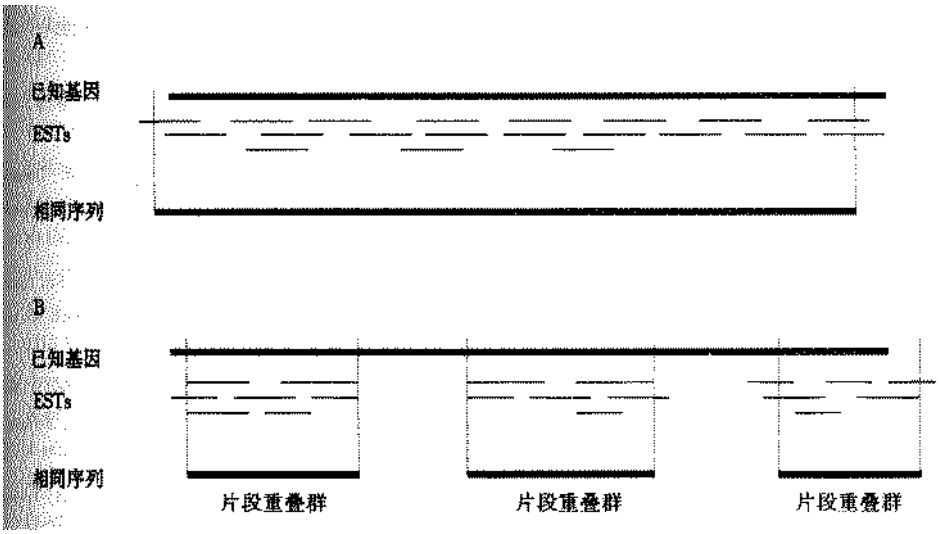


图 4.14 应用以知基因对 EST 数据库进行同源性比较构建的两种 EST 重叠群情况（贺林，2000）

四、隐马尔可夫模型(HMM)

改进目前数据库搜索技术的灵敏性和速度的一条可行办法，是通过蛋白质家族的多序列列线(mutiple alingment)建立一致序列(consensus sequence)。与两条序列的列线比对不同，一致序列可揭示更多的信息，如家族内保守程度不一的残基位置，残基插入和缺失的可能性等等。一致序列的所有表述形式，例 profile、模块(block)等都可视为隐马尔可夫链(Hidden Markov Model,HMM)的特例。

HMM 是最近几十年发展起来的时间序列模型，已在语音识别(speech recognition)、离子通道记录、最佳特征识别等方面被应用。HMM 也被较早地应用于生物信息学上的一些问题，如 DNA 编码区、蛋白质超级家族(supper family)的构模等。但是，直至上世纪 90 年代中叶，HMM 才与机器学习技术结合，被系统地应用于整个蛋白质家族和 DNA 区段的建模、列线和分析。HMM 与神经网络、随机模型(stochastic grammar)和贝叶斯网络(Bayesian networks)关系极其密切，或者可视为它们的一个特例。HMM 将 DNA 序列的形成看作一个随机过程，编码和非编码的 DNA 序列在核苷酸选用频率上有所不同而对应于不同的马尔可夫模型。由于这些马尔可夫模型的统计规律是未知的，而 HMM 能够自动寻找出其隐藏的统计规律，因而被称为隐马尔可夫模型。对于处理复杂的 DNA 序列，HMM 需要学习不同 DNA 序列结构的信息。

初阶(first order)或称为 0 阶离散HMM是一种时间序列随机通用模型，由有限的状态集S、离散字符表A、转换(transition)概率矩阵T=(t_{ji})和散发(emission)概率矩阵E=(e_{ix})定义。字符散发，系统由一种状态随机地向另一种状态进化。假设系统处于状态i，它存在t_{ji}概率转变为状态j，而字符x散发的概率为e_{ix}。因此，对于HMM来说，系统的每一个状态只与 2 个不同的骰子(dice)节点有关：散发节点和转换节点。0 阶马尔可夫链假设散发和转换仅由现状态决定，而与过去的状态无关。而字符的散发只有模型系统本身可以识别，即所谓“ 隐藏 ”

(hidden)。

图 4.15 给出了一个非常简单的HMM例子。例子中，最后观察到的序列为ATCCTTTTTTTCA。我们可以想象有 2 个“ DNA节点 ”(DNA dice)：第一个节点的散发概率向量为(e_{1A}=0.25,e_{1C}=0.25,e_{1G}=0.25,e_{1T}=0.25)，第二个节点的散发概率向量为(e_{2A}=0.1,e_{2C}=0.1,e_{2G}=0.1,e_{2T}=0.7)。而转换概率如图所示。

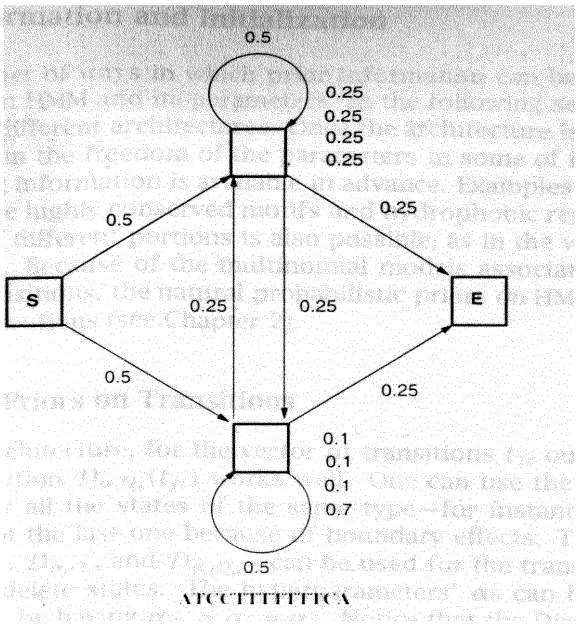


图 4.15 一个简单的 HMM 例子。例子中除了开始（S）和结束（E）两个状态外，还有两个中间状态（Baldi and brunak, 1998）。

对于生物序列而言，HMM 的字符当然是 20 个字母的氨基酸或 4 个字母的核苷酸。但依据不同的问题，其它的一些字符也可使用，如 64 个字母的三联体字母，3 个字母(,β,coil)的二级结构等。当然，HMM 模型并非如上所举的仅有 2 个节点例子那么简单。图 4.16 给出了一个最基本和被广泛应用的“ 左－右 ”(left－right)结构模型——标准线性结构模型。所谓“ 左－右 ”结构是指该结构中不存在从一种状况回复到已有状况的情况。对于 HMM 模型，一个蛋白质家族如同语音识别中一个词的不同语调。

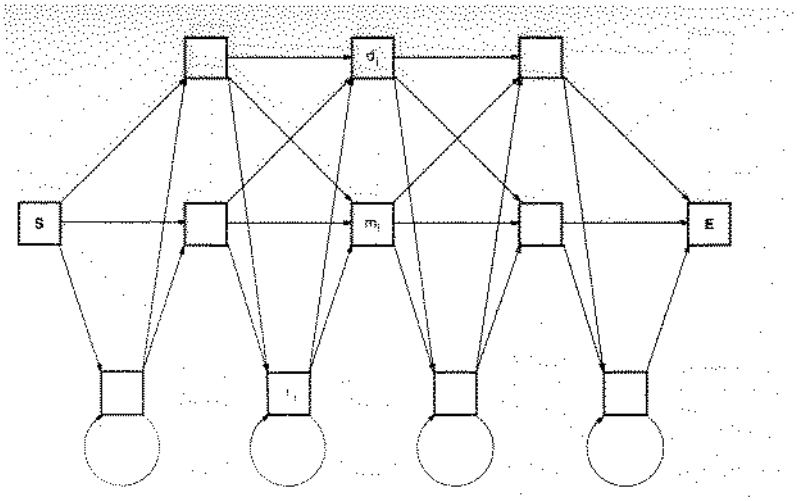


图 4.16 HMM的标致结构。S和E表示开始和结束状态，d_i、m_i和i_i分别表示缺失、

维持和插入状态（Baldi and brunak, 1998）。

一旦一个蛋白质家族成功地构建了 HMM 模型，则该模型可以用于多个领域：①多序列列线；②数据库序列数据的挖掘和分类；③结构分析和模式查找。

五、神经网络

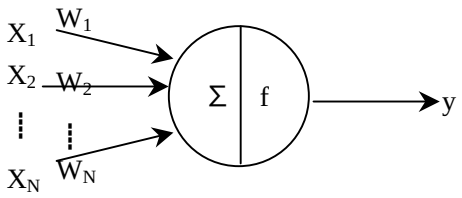
1、神经网络的基本原理

神经网络(NN)属于信息科学理论范畴，它是随着信息科学的开创而发展起来的，目前发起的“神经计算机”革命对计算机产业产生了空前的推动作用。

神经网络是由大量的简单处理单元(即神经元)构成的非线性动力学系统，它具有的学习算法能使其对事物和环境具有很强的自学习、自适应和自组织能力。它能解决常规信息处理方法难以解决或无法解决的问题，尤其是那些属于思维(形象思维)和推理方面的问题。

在人工神经网络中，神经元常被称为“处理单元”，有时从网络的观点出发又把它称为“节点”。人工神经元是生物神经元的一种近似，在功能上讲它只是一阶逼所。它仅仅近似地模拟了生物神经元的三个过程。

处理单元的结构：



（1）输入与输出

（2）加权系数

（3）神经元函数：如常用的活化函数 S 型(Sigmoid)函数

目前应用的一些神经模型包括：感知器(perceptron)模型、反向传播网络(backpropagation network, BP 或 BPN)模型、自组织特征映谢模型(self-organizing feature map, SOFM)、回归网络(recurrent network)模型(Hopfield 提出)、混合网络和混合系统模型。其中混合系统模型是指把神经网络与常规信号处理系统模型结合起来，以便分别取其所长，目前在语音信号处理中将神经网络与隐马尔柯夫模型(HMM)结合起来进行语音识别即是一例，同时在生物信息学上除了应用 NN(BPN)外，也将此混合系统模型加以应用。

神经网络的学习规则：神经网络中的神经元是一个具有相当自适应能力的处理单元，它所连接的权可以根据一定的规则来调整。比较流行的几种规则：

（1）Hebb 规则

（2）感知器学习规则

（3）δ学习规则

（4）Widrow-Hoff 学习规则

（5）相关学习规则

（6）胜者取全学习规则

2、BPN 神经网络

以下重点介绍 BPN 神经网络。

BPN网络由输入层、输出层以及若干隐层节点互连而成的一种多层网，它的输入和输出是在[0，1]或[-1，+1]区间连续取值，每个处理单元对输入的加权和y_i加以非线性处理，得到其活性输出。最常用的非线性函数为Sigmoid函数：

$$f_{(yi)}=\begin{cases}\dfrac{1}{1+e^{-y}} & (0,1)\quad (f_{(x)}=\dfrac{1}{1+e^{-x}})\\ \dfrac{1-e^{-y}}{1+e^{-y}} & (-1,+1)\end{cases}$$

BPN为前馈网络，对其训练所采用的算法是反向传播法，这是一种有导师学习方法。它利用了均方误差和梯度下降法来实现对网络连接权的修正。对网络权值修正的目标是使网络实际输出与规定输出之间的均方误差(mean squared error，MSE)最小。对于一个处理单元的情况下，如果网络有K个训练样本$\{E^k\}$，对应的正确输出为$\{C^k\}$，网络的权为W，则用 ε 表示MSE为：

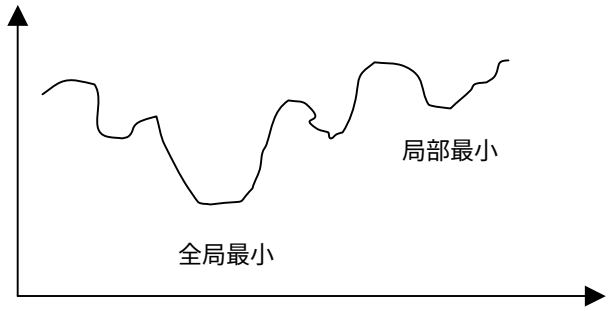
$$\varepsilon=\frac{1}{k}\sum_{k=1}^{k}(w\cdot E^k-C^k)^2$$

它可以看成是权的函数 ε (w)，则我们可以按下式来修正权值：

$$W^*=W-\rho\triangle\varepsilon(W)$$

其中，ρ是一个大于零的小数，它规定了修改的步幅。

梯度下降法的基本思想：首先设置权 W 的一组初值，然后，连接计算均方误差相对于权的梯度，并按上式一小步小步地修正权值，当满足一定的准则时（比如 MSE 进入到下限的某一范围时）即停止。这时称为算法收敛。对于梯度下降算法来说，最大的问题是不能保证收敛到全局最优。



更新权值公式可进一步分解为：

$$W_{ij}^{*} = Wij + \rho\delta_i X_j, \qquad \delta i = -\frac{\partial \varepsilon}{\partial y_i}$$

$$\downarrow$$

$$W_{ij}^{*} = W_{ij} + \alpha\Delta W_{ij} + \rho\delta_i X_j, \qquad \Delta W_{ij} = W_{ij}^{*} - W_{ij}$$

一般ρ选0.1以下，α选0.9，初始权值可在(-2/q，2/q)之间选择(q为一个神经元的输入数)。δ_i为求误差梯变过程的一个中介量。

新的梯度法：共轭梯度(conjugate gradient)法和准牛顿(quasi-Newton)法。

克服局部极小问题：可随机换一组初始权值重新训练一次。

3．神经网络的应用

神经网络技术应用于生物序列分析领域已有较长历史。1982年利用氨基酸序列，神经网络便被用于预测核糖体结合位点。但是神经网络在本领域的真正应用，是在1986年多层神经网反向传播(即BPN)学习算法被广泛应用后才开始的，特别是1988年该技术被应用于蛋白质二级结构预测后，该技术已在多方面取得了很好的应用效果。一些程序已采用了神经网络方法，例如比较有名的GRAIL程序(Gene Recognition and Analysis Internet Link)(由美Oak Ridge国家实验室研制，http://avalon.epm.ornl.gov/Grail-bin/EmptyGrailForm/)。GRAIL采用了神经网络技术，具有根据范例而“学习”的功能。向GRAIL程序提供了一组已知序列，其中序列的外显子和内含子的位置都已通过实验确定；神经网络根据设计运行，确定特定类型的简化特性，当一条待测序列输入后，网络可以利用已建立的序列与特性之间的关系，找出待测序列的外显子等。

神经网络主要应用于以下几个方面：

①序列编码分析；
②蛋白质二级结构预测；
③单肽及其切割位点预测；
④遗传密码的结构和起源分析；
⑤真核生物基因寻找和内含子剪接位点预测。

六、RNA二级结构预测

尽管现有一些RNA折叠程序可以预测RNA二级结构，但这类分析仍然是一

门艺术。RNA 折叠有助于找出 RNA 分子中可能的稳定茎区，但对给定的 RNA 分子来说，这一结果的生物学意义究竟有多大，还是一个未知数。即使有此局限性，二级结构的预测还是有助于找出 mRNA 控制区以及 RNA 分子中可能形成稳定折叠结构的区段。

预测二级结构的最大难题是对三级结构中既有的相互作用进行模型处理，然后将此处理结果回归成一级结构要素，以用于折叠结构的预测。诚然，现有的 RNA 折叠程序并未考虑核酸分子中可能的三级结构。这些程序只能定出有限数目的二维结构的能学参数，由此推测的二维最稳定结构，可能与三维最稳定结构相去甚远，因为三维亿个结构里的环区可以与环区相互作用，螺旋区可以堆积，还会出现各种的非 Watson-Crick 碱基对结构。

目前已有一些比较有名的预测程序，例如 MFOLD [M 代表多(multi)，从早期的 RNAFold 程序或 GCG 软件包的 FOLD 程序扩充而成]，由加拿大国家研究基金会的 Michael Zuker 设计。除对碱基配对的标准能学进行分析外，MFOLD 还考虑到了碱基堆积的能量及单碱基统计的熵。这一程序的 VMS、VNIX、DOS 和 Macintosh 版本可以从许多软件组合中找到。尽管 MFOLD 的输出是文本形式的(图 4.17A)，但有几个程序可以将预测结构转化为图示形成(例如由 Don Gillbert 设计的 Loop Viewer，见图 4.17B)。

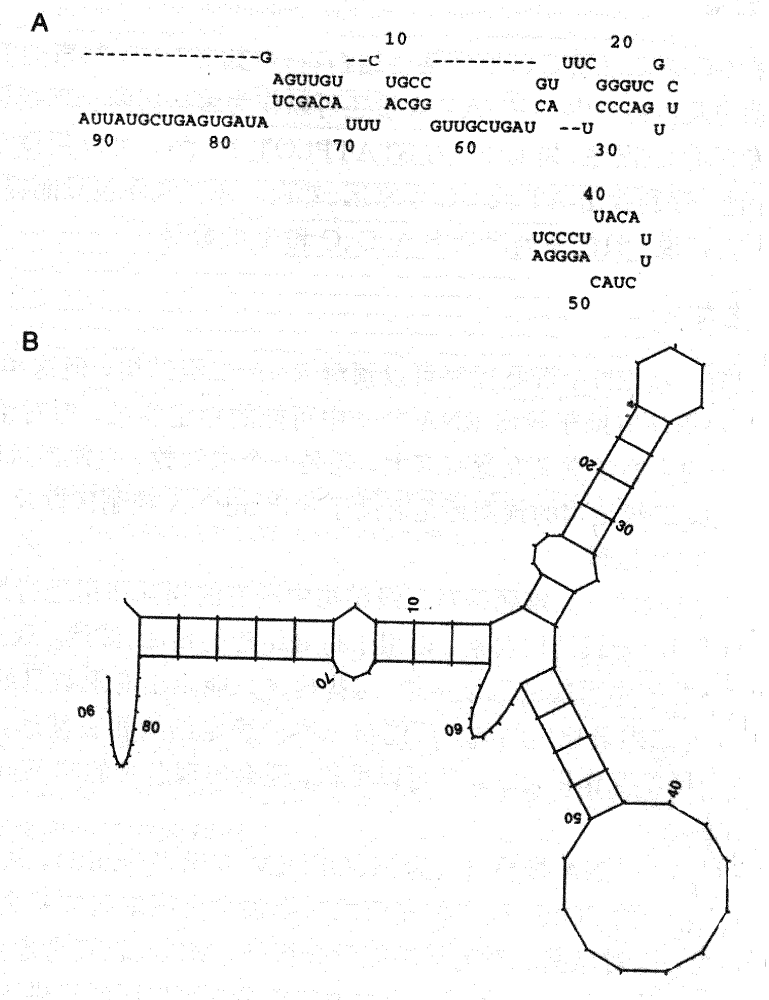


图 4.17 RNA 二级结构的文本输出结果（A）和图形显示(B)。分别由 GCG 的 FOLD 和 Squiggles 程序生成。

第三节 基因组分析

一、基因组分析：生物信息学发展的“史记”

自从1995年第一个可以独立生存的生物被基因组测序以来(Fleischmann et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science*. 1995，269:496-512)，每年在 *NATURE* 和 *SCIENCE* 杂志上都会发表一些重要生物基因组测序完成后的分析文章。这些大文章（Article）中对基因组的分析可谓登峰造极，往往包括了当时想得到的和可以做得到的序列分析手段，它们代表着当时生物信息学发展的最新高度。可以说，这些文章是生物信息学发展史的另类记录。

以下列出了一些重要基因组分析文章，感兴趣的读者不妨对他们的分析内容或方法做些比较：

1977 First biology: Phage ΦX174 (5.386kb)

Sanger F, Air G M, Barrell B G, *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 1977, 265:687-695

1982 Phage lambda genome

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol.* 1982, Dec 25;162(4):729-73

1983 Phage T7 genome (39.937kb)

Dunn,J.J. and Studier,F.W. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol*. 1983, 166 (4), 477-535

1995 First bacterial genomes (1.8 Mb)

Fleischmann et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995 Jul 28;269(5223):496-512

1996 Yeast genome

Genome sequence of the yeast *S. cerevisiae* Overview of the yeast genome. H. W. MEWES *et al*. *Nature* 387, suppl. 7-8 (29 May 1997)

1997 *E. coli* genome

The Complete Genome Sequence of *Escherichia coli* K-12. Frederick R. Blattner, et al. *Science*, Volume 277, Number 5331, Issue of 5 Sep 1997, pp. 1453-1462.

1998 Worm (multicellular) genome

Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. The *C. elegans* Sequencing Consortium. *Science* Dec 11 1998: 2012-2018.

1999 Fly genome

The Genome Sequence of *Drosophila melanogaster.* Mark D. Adams, et al. *Science* Mar 24 2000: 2185-2195.

2000 First plant genome: *Arabidopsis thaliana*

Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* THE ARABIDOPSIS GENOME INITIATIVE. *Nature* 408, 796-815 (14

December 2000)

2001 Human genome

The Sequence of the Human Genome. J. Craig Venter, et al. *Science* Feb 16 2001: 1304-1351.

Initial sequencing and analysis of the human genome. THE GENOME INTERNATIONAL SEQUENCING CONSORTIUM. *Nature* 409, 860-921 (15 February 2001)

2002 First crop genome: Rice (ssp. *indica* and *japonica*) genomes

A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *indica*). Jun Yu, et al. *Science* Apr 5 2002: 79-92.

A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). Stephen A. Goff, et al. *Science* Apr 5 2002: 92-100.

Sequence and analysis of rice chromosome 4. Qi Feng, et al. *Nature* 420, 316 - 320 (21 Nov 2002) Letters to Nature

The genome sequence and structure of rice chromosome 1. Takuji Sasaki, et al. *Nature* 420, 312 - 316 (21 Nov 2002) Letters to Nature

In-Depth View of Structure, Activity, and Evolution of Rice Chromosome 10. The Rice Chromosome 10 Sequencing Consortium. *Science* Jun 6 2003: 1566-1569.

2003 Dog genome

The Dog Genome: Survey Sequencing and Comparative Analysis. Kirkness *et al*. *Science*, Volume 301, Number 5641, Issue of 26 Sep 2003, pp. 1898-1903

2004 Rat genome

Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Rat Genome Sequencing Project Consortium. *Nature* 428, 493-521 (1 Apr 2004)

二、比较基因组学[4]

比较基因组学是基因组学的重要分支,它是随着人类和其它生物基因组的大规模测序发展起来的新科学,现已成为研究生物基因组最重要的策略与手段之一。

与比较解剖学、比较组织学等学科一样，比较基因组学使用是遗传学的重要方法——异同的比较，但该学科的特点是在整个基因组的层次上比较，如基因组的大小、基因数量的多少、特定基因的存在或缺失、基因(或标记序列片段)的位置及排列顺序、特定基因或片段的组织等等。而最重要，也是最体现比较基因组学科特点的是全基因组的核苷酸序列的整体比较。随着世界各国基因组计划的实施，除了人类基因组，许多模式生物基因组测序也已完成或正在进行中，如大肠杆菌、酶母、果蝇、线虫、小鼠、鱼、拟南芥等。同时，美国的“食物基因组计划”，几乎包括了所有重要作物：小麦、玉米、大豆、马铃薯、南瓜、棉花，

[4]本部分内容取自陈竺、杨焕明等人的文章，见：贺林．解码生命—人类基因组计划和后基因组计划，北京：科学出版社，2000

而我国的水稻、家蚕、微生物等基因组计划也在进行中或已完成。这些基因组全序列数据将成为比较基因组的最基本研究对象。

认同所有生物的基因组都有共同的进化史，即进化上的共性是比较基因组学的理论依据，可以说，没有进化上的关系，就没有比较基因组学。进化是基因组比较的最重要主题，所以目前基因组比较的生物信息学方法主要来自系统进化分析的一些方法，例如系统进化树的构建方法等。相关内容请参见第五章。基因组比较急需发展针对整个基因组的专用算法。基因组是一种具有大尺度、巨量特点的研究对象，它有其特有问题，必须有特定的算法才能充分挖掘和利用基因组信息。

以下对基因组学分析中经常涉及的四个最基本概念进行介绍：

1、相似性：

相似性(similarity,有时也用analogy)就是简单比较得出的两者之间的相同程度。相似性本身的含义，并不要求与进化起源是否同一，与亲缘关系的远近、甚至于结构与功能有什么联系。核苷酸与氨基酸序列的测定，使原先“模糊”的描述有了定量的指标——百分比。不同基因组之间、不同基因或不同物种的“同一”基因，都可以用%来表示异同程度。

2、同源性：

同源性(homology)是具有严格定义的进化学词汇：在进化上起源同一。同源性可以用来描述染色体—“同源染色体”、基因—“同源基因”和基因组的一个片断—“同源片断”。

在进化上起源同一的两段核苷酸序列，特别是功能较重要的保守区断或基因，一般表现为相似。迄今有证据表明，同源的基因的的确确在核苷酸(或氨基酸)序列上具有较高程度的相似，这就带来了这两个词的混用。如我们有时把“相似搜索(similarity searching)”说成是“同源搜索(homology searching)”。在比较两段序列时，正常的描述应该是：这两个片断可能同源(或这两个基因有可能为同源基因)，因为它们的核苷酸(或氨基酸)的相似程度为 80%。“80%的同源”的说法是不正确的(还有 20%的不同源?)，也是不符合事实与定义的。

相似性与同源性是两个不同的概念，相互之间并没有直接的等同关系。相似的不一定同源，因为在进化的过程中，来源不同的基因或序列由于不同的独立突变而“趋同”并不罕见；同源一般表现为相似，但同源并不一定比非同源的相似程度要高。我们只是在进化的过程的一个时间点上加以观察。功能相似或相同也不一定必然同源。非同源基因的代谢功能替换已有不少证据，其它表型相似也不一定反映了同源，不同基因的不同突变就有可能产生“表型模拟”。

而同源又有两种不同的情况即垂直方向的(orthology)与水平方向的(paralogy)。

3、直系同源：

直系同源(orthology)是比较基因组学中最重要的定义。直系同源的定义是：

(1)在进化上起源于一个始祖基因并垂直传递(vertical descent)的同源基因；

(2)分布于两种或两种以上物种的基因组；

(3)功能高度保守乃至于近乎相同，甚至于其在近缘物种可以相互替换；

(4)结构相似；

(5)组织特异性与亚细胞分布相似。

在这些条件中，垂直传递和功能相同是最重要的。如多种抗药性基因，在细菌、果蝇、河豚鱼、小鼠、人类的基因组中都存在，其结构相似，功能都与多种药物的抗性有关。直系同源基因的鉴定是比较基因组的研究线索和内容，直系同源的存在是基因组进化的重要证据，因此对直系同源的定义与条件的掌握甚为严格。鉴定直系同源的实际操作标准(practical criteria)为：

如基因组Ⅰ中的A基因与基因组Ⅱ中的A′基因被认为是直系同源，则要求：

(1)A′的产物比任何在基因组Ⅱ中所发现的其它基因产物都更相似于A产物；

(2)A′与A的相似程度比在任何一个亲缘关系较远的基因组中的任一基因都要高；

(3)A编码的蛋白与A′编码的蛋白要从头到尾都能并排比较，即含有相似以至于相同的模序(motif)。

3、旁系同源：

旁系同源(paralogy)基因是指同一基因组(或同系物种的基因组)中，由于始祖基因的加倍而横向(horizontal)产生的几个同源基因。

直系与旁系的共性是同源，都源于各自的始祖基因。其区别在于：在进化起源上，直系同源是强调在不同基因组中的垂直传递，旁系同源则是在同一基因组中的横向加倍；在功能上，直系同源要求功能高度相似，而旁系同源在定义上对功能上没有严格要求，可能相似，但也可能并不相似(尽管结构上具一定程度的相似)，甚至于没有功能(如基因家族中的假基因)。旁系同源的功能变异可能是横向加倍后的重排变异或进化上获得了另一功能，其功能相似也许只是机械式的相关(mechanistically related)，或非直系同源基因取代新产生的非亲缘或远缘蛋白在不同物种具有相似的功能。在真细菌与古细菌的基因组中，30%～50%的基因属旁系同源，在真核基因组的比例更高(Koonin EV and Galperin MY,1997)。

相似与同源，直系与旁系需要在定义上加以明确，但实际应用中很难截然分开。与别的常用术语也很难明确界定。但基因家族或多基因家族(gene family, multigene family)的原来的定义较侧重于结构，因而一个直系基因可以与几个旁系基因同属于一个基因家族。在这一定义上，旁系同源可以说是一个基因家族中的其他成员(Huynen et al, 1997)。

随着不同物种全基因组序列的阐明，上述概念愈见重要并更明确。从已知的 7 个物种的全基因组序列比较，如所有的保守基因都据同源关系而加以分类(Tatusov RL et al.,1997)，可归纳出 720 个直系同源簇(clusters of orthologous groups,COG)，每一 COG 由一个直系同源蛋白或存在于至少 3 个种系(lineage)的直系的旁系同源组(orthologous sets of paralogs)组成。而基因家族又因大批基因及产物序列而赋予新的内容，这对于扩大对生物过程的认识与操作基因的能力有很大的意义(Henikoff et al.,1997)。