第 4 章　核酸序列分析

细胞中的核酸有两大类——DNA 和 RNA，前者携带着决定个体性状的遗传信息，后者参与遗传信息的表达与调控，它们在生命活动中起着重要的作用。生物体的遗传信息储存于 DNA 分子上，表现为特定的核苷酸排列顺序，通过复制将遗传信息传递给后代。DNA 分子中的遗传信息转录到 RNA 分子中，再由 RNA 翻译生成体内各种蛋白质，最终执行其特定的生物功能。ACGT(U)四种核苷酸是构成各种生物体核酸序列的基本组分，不同生物体具有不同的排列顺序，不同的排列顺序蕴含着不同的生物信息。核酸序列中包含着生物的遗传信息和进化信息，从海量的已经测序的核酸序列中获取和挖掘信息，是生物信息学的研究目的，序列分析是重要的途径之一，也是对核酸序列进行生物信息学分析的首要步骤。

4.1 DNA 序列信息分析

DNA 是主要的遗传物质，是携带遗传信息的载体之一。DNA 序列是指 DNA 的一级结构，由 ACGT 四种碱基组成，DNA 序列又可称为碱基序列，不同种属的 DNA 碱基组分存在差异，与遗传密码子的使用偏好和 DNA 甲基化程度具有相关性。

DNA 主要携带两类遗传信息，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 两种形式），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基因序列；另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分子特异地识别结合，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。遗传信息储存于具有特征信息的 DNA 序列中，根据这些特征信息设计不同的算法，能够在海量的序列数据中挖掘出具有生物学功能的特征信息。

本节将介绍 DNA 序列一级结构的基本信息和特征信息分析方法。DNA 基本信息分析主要包括序列组分分析、序列转换、限制性内切酶位点分析；序列的特

征信息分析主要包括开放阅读框（Open Reading Frame，ORF）分析、启动子及转录因子结合位点分析和 CpG 岛（CpG island）识别。

4.1.1 DNA 序列的基本信息

(1) DNA 序列组分分析

物理化学性质是 DNA 的基本性质，不同物种其 DNA 的物理化学性质具有差异性。DNA 物理化学性质主要由碱基组成决定，碱基组成有两种方法表示——碱基比例（base ratio）和 GC 百分比含量简称 GC 含量（GC content）。

奥地利犹太生物学家 Erwin Chargaff 用层析和电泳技术分析组成 DNA 的碱基，提出了 DNA 碱基组成的 Chargaff 规则：同一生物的 DNA 碱基含量是 A=T，G=C，A+G=C+T；且（A+T）/（G+C）的比值因生物种类不同而异。

原核生物中不同种属的 GC 含量从 25%到 75%不等，这种组分差异可用于识别细菌种类。真核生物物种间 GC 含量的差别不如原核生物明显，但真核基因组中不同区域 GC 含量存在差异。GC 含量与物种的密码子使用频率有关，而且与 DNA 双链的熔解温度有关，是进行核酸杂交的重要参数。

核酸碱基组成可通过一些常用软件直接获得，如：BioEdit 和 DNAMAN。BioEdit 是 Tom Hall 开发的一个生物序列编辑器，其基本功能是提供蛋白质核酸序列的编辑排列处理和分析，如：序列比对、序列检索、引物设计、系统发育分析等。DNAMAN 是美国 Lynnon Biosoft 公司开发的高度集成化的分子生物学应用软件，可完成核酸和蛋白质序列的综合分析工作，包括多重序列比对、引物设计、限制性酶切分析、蛋白质分析、质粒绘图等。

核酸碱基组成分析工具及其网址如下：

BioEdit： http://www.mbio.ncsu.edu/BioEdit/bioedit.html

DNAMAN： http://www.lynnon.com。

例 4.1 人类 CD9 基因序列组分分析

应用 BioEdit 软件，以人类 CD9 基因序列（序列号 AY422198）为例进行核酸组分分析。打开 BioEdit 输入 AY422198 序列，选中该序列，点击“sequence”

下拉菜单“Nucleic Acid”中的“Nucleotide Composition”项，即可得到序列组分分析结果，见图 4.1。其结果显示 CD9 基因序列中四种碱基 A、C、G、T 的含量分别是 25.58%、25.51%、25.46%、23.44%，GC 含量为 50.97%。

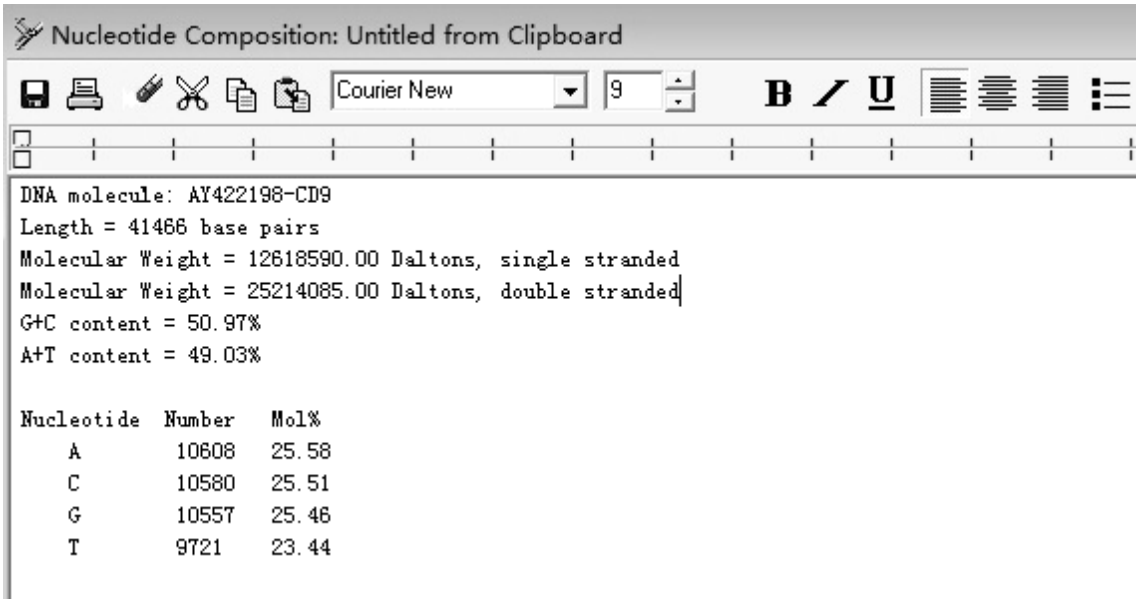


图 4.1 AY422198 序列组分分析结果界面

(2) 序列转换

DNA 序列具有双链性、双链互补性及开放阅读框在两条链上存在的特性，因此进行序列分析时，经常需要针对 DNA 序列进行各种转换，例如反向序列、互补序列、互补反向序列、显示 DNA 双链、转换为 RNA 序列等。

序列转换可使用的软件有 DNASTAR、BioEdit、DNAMAN 等。DNASTAR 软件是 DNASTAR 公司开发的 Lasergene 程序组，是核酸序列和蛋白质序列的综合分析工具，其中的 EditSeq 程序能够实现核酸 DNA 序列的各种转换。DNASTAR 网址：http://www.dnastar.com/。

例 4.2 人类 CD9 基因序列转换

应用 DNASTAR 软件，以人类 CD9 基因（序列号 AY422198）前 300bp 序列为例进行序列转换。打开 DNASTAR 软件“EditSeq”程序，输入 AY422198 序列，选中该序列，点击“Goodies”下拉菜单中的“Reverse Sequence”项，可得到该序列的反向序列；点击 “Reverse Complement”则可得到该序列的反向互补序列，如图 4.2 所示。

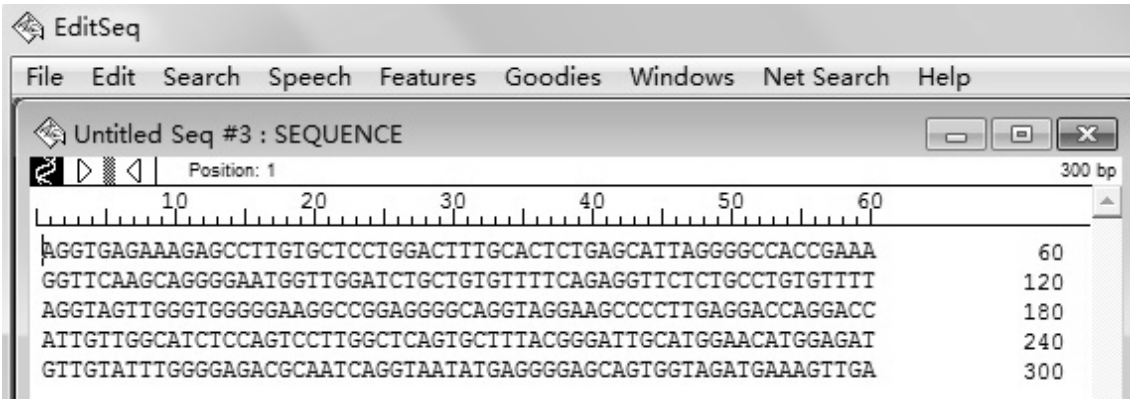


图 4.2　反向互补序列界面

(3) 限制性内切酶位点分析

在生物体内有一类酶，具有将异源性 DNA 切断的功能，可以限制异源 DNA 的侵入并使之失活，但对自身 DNA 无损伤作用，它们能够维持细胞原有遗传信息的完整性。这种具有切割作用的内切核酸酶被称为限制性核酸内切酶，简称内切酶。它们能够识别 DNA 的特异顺序序列即识别位点，并在识别位点内部或周围切割双链 DNA。内切酶是分子生物学和基因工程中重要的工具酶，限制性内切酶位点分析是分子克隆的基础。

内切酶分为Ⅰ、Ⅱ、Ⅲ型三大类。基因工程中的内切酶一般指的是Ⅱ型内切酶。这类内切酶具有专一的识别和切割位点，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。序列中被内切酶识别的位点多数为回文对称结构，长度一般为 4-8 个碱基，常见为 6 个碱基，切割位点在 DNA 两条链相对称的位置。例如，基因工程中常用的两个内切酶 EcoRⅠ和 HindⅢ的识别序列和切

割位置如下：

EcoRⅠ　G↓AATTC　　HindⅢ　A↓AGCTT

　　CTTAA↑G　　TTCGA↑A

核酸序列内切酶位点识别依据内切酶所识别的序列结构信息进行预测分析。常用内切酶的资源是限制酶数据库（Restriction Enzyme dataBase，REBASE），由新英格兰生物实验室建立，收录了内切酶的所有信息，包括内切酶识别序列和切割位点、甲基化酶、甲基化特异性、其他相关酶、酶类产品的商业来源及公开发表的和未发表的参考文献。REBASE 提供了内切酶的查询工具、识别位点序列信息及内切酶酶切双链 DNA 的三维结构等信息;分析工具具有提供理论酶切消化图谱、序列比对、酶切位点分析等功能。

限制性内切酶位点分析常用的工具是 NEBCutter2，可接收 DNA 序列并产生酶切位点分析结果。NEBCutter2 使用的内切酶来源于 REBASE 数据库，它的识别位点列表每天根据 REBASE 数据库数据同步更新。此外，很多 DNA 分析的商业软件都含有酶切位点分析功能，如集成化分析软件 BioEdit 、DNAMAN 和 DNASTAR 等。

限制性内切酶位点分析常用数据库和工具网址：

REBASE：　http://rebase.neb.com/rebase/rebase.html

NEBCutter2：http://tools.neb.com/NEBcutter2/。

例 4.3 人类 CD9 基因序列内切酶位点分析

应用 REBASE 数据库的分析工具箱链接 NEBCutter2 分析软件获得人类 CD9 基因（序列号 AY422198）序列内切酶图谱。首先登录 REBASE 数据库页面进入 REBASE Tools 工具箱，登入如图 4.3 所示的 NEBCutter2 分析页面。序列输入有三种方式：键入 GenBank 序列号（如：AY422198）、直接粘帖序列及从本地序列文件输入;输出方式有两种:直线形和环形,本例选择直线形;内切酶库选择“NEB enzymes”。提交序列后页面返回酶切位点分析结果，见图 4.4。选择序列某区域点击“ZOOM in”进行放大得到该区域详细的酶切位点，见图 4.5。图 4.4 中的识别

位点标示“blunt end cut”为平末端、 “5' extension”为粘性末端 5'端突出、“3' extension” 为粘性末端 3'端突出。NEBCutter2 还可提供单一酶切或多选酶切切位点识别和模拟消化图谱。

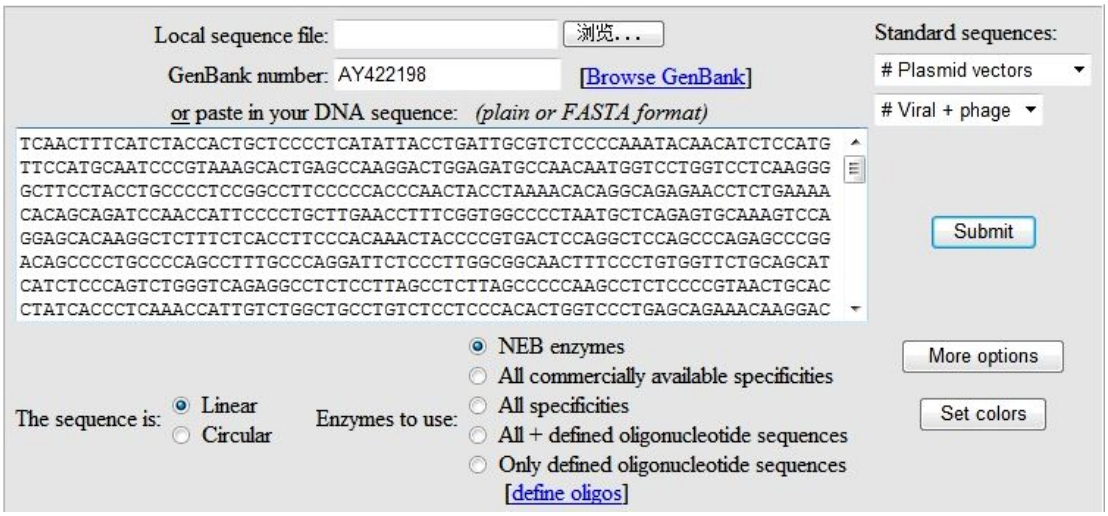


图 4.3 NEBCutter2 分析页面

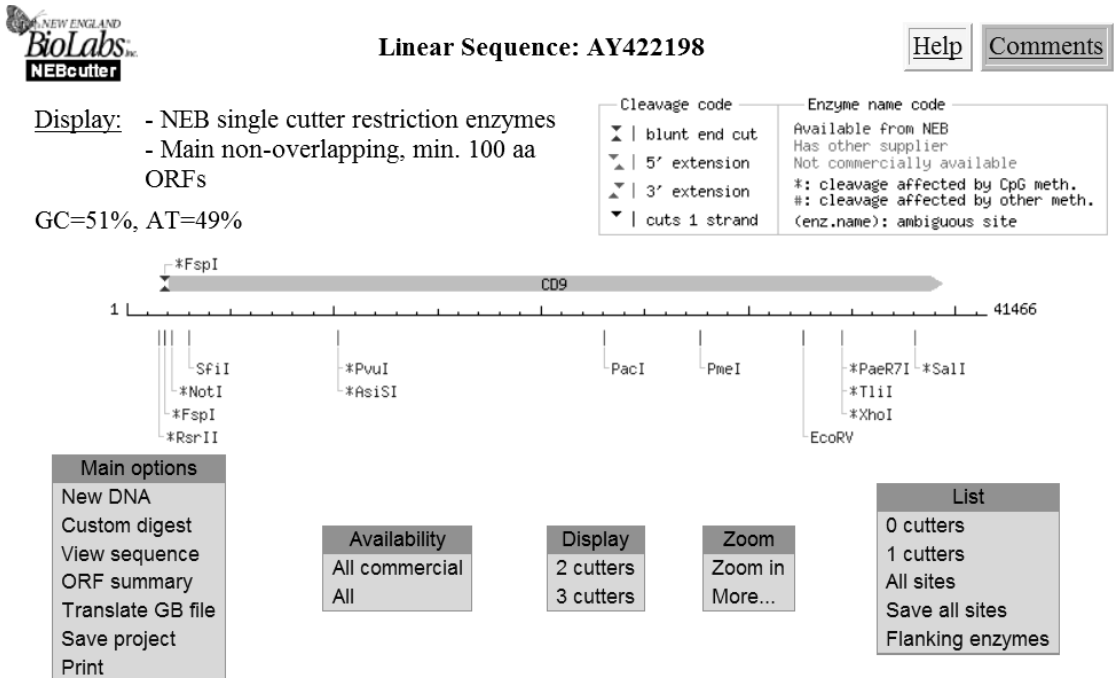


图 4.4 AY422198 序列酶切位点分析结果

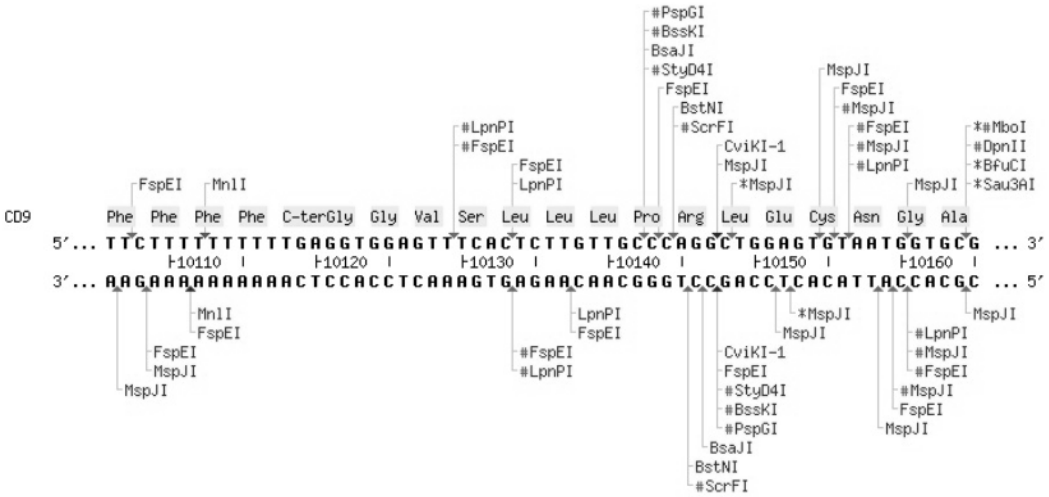


图 4.5　AY422198 序列 10106-10165bp 区域放大后的酶切位点图

4.1.2 DNA 序列的特征信息

(1) 开放阅读框分析

完整的开放阅读框是从 5'端起始密码子（ATG）到终止密码子（TAA、TAG、TGA）之间的一段编码蛋白质的碱基序列。一个 ORF 存在一个潜在的编码序列（Coding Sequence，CDS），不同的 ORF 翻译成氨基酸可得到不同的蛋白编码。DNA 序列中一个 ORF 对应一个候选的 CDS，ORF 分析是对 DNA 序列是否为编码序列的初步判别，是判断该序列是否为 CDS 的方法之一。一条 DNA 序列可能存在六种阅读框，如图 4.6 所示，对于任意一条给定的核酸序列，根据密码子的起始位置，可以按照三种方式进行解释，其反向互补序列又含有三种阅读框顺序。

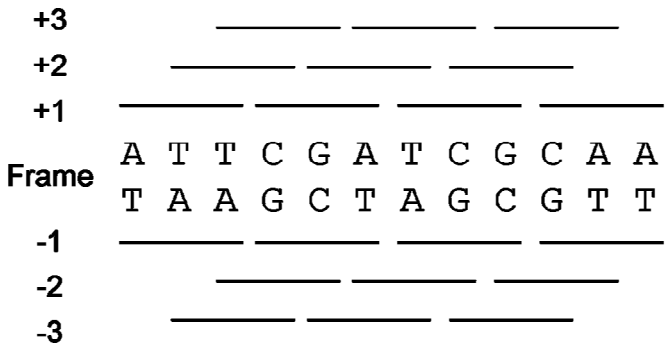


图 4.6　DNA 序列阅读框顺序

ORF 的预测程序主要是对编码区进行特征统计、相关模式的识别或利用同源比对的方法进行识别。原核生物编码区通常只含有一个单独的 ORF，识别方法相对简单，即最长 ORF 法。而真核生物的编码区被内含子分隔成数个不连续的外显子，其编码区序列分析更趋复杂。

ORF 分析常用的工具为 NCBI 在线分析工具 ORF Finder，网址为：http://www.ncbi.nlm.nih.gov/gorf/gorf.html 。其它集成化的软件有 BioEdit 、DNAMAN 和 DNASTAR 等。

例 4.4 大肠杆菌基因组序列 ORF 分析

以大肠杆菌基因组序列 U00096 前 2800bp 片段为例，应用 NCBI 的 ORF Finder 分析该片段可能存在的 ORF。登录 ORF Finder 主页，粘帖该序列片段。点选“Genetic Codes”下拉菜单中的“Bacterial Code”。点击“OrfFind”提交序列后得到分析结果，见图 4.7。图中显示六种相位中长度大于 100bp 的可能的阅读框共 17 个：左侧每一条绿色条形框代表一个预测的可能的 ORF，右侧阅读框依据长度从长到短排列，“Frame”栏“+1”表示 DNA 正链第一相位读码顺序。点击最长的条形框可获得该阅读框编码的氨基酸，该片段位于正链第一相位 337bp 至 2799bp，编码 820 个氨基酸，见图 4.8。为验证 ORF 预测的可靠性，ORF Finder 提供蛋白序列与蛋白序列比对的 blastp 工具，来查询氨基酸序列在蛋白数据库中的相似序列。如图 4.9 所示，该编码氨基酸序列与蛋白数据库中多条序列具有高度相似性，说明该 ORF 具有较高的可靠性。

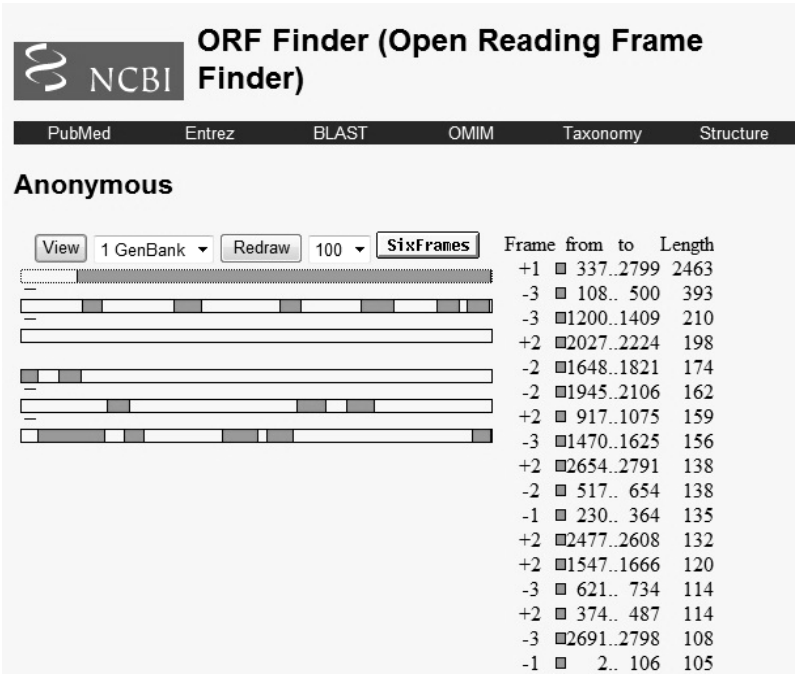


图 4.7　ORF 预测结果页面

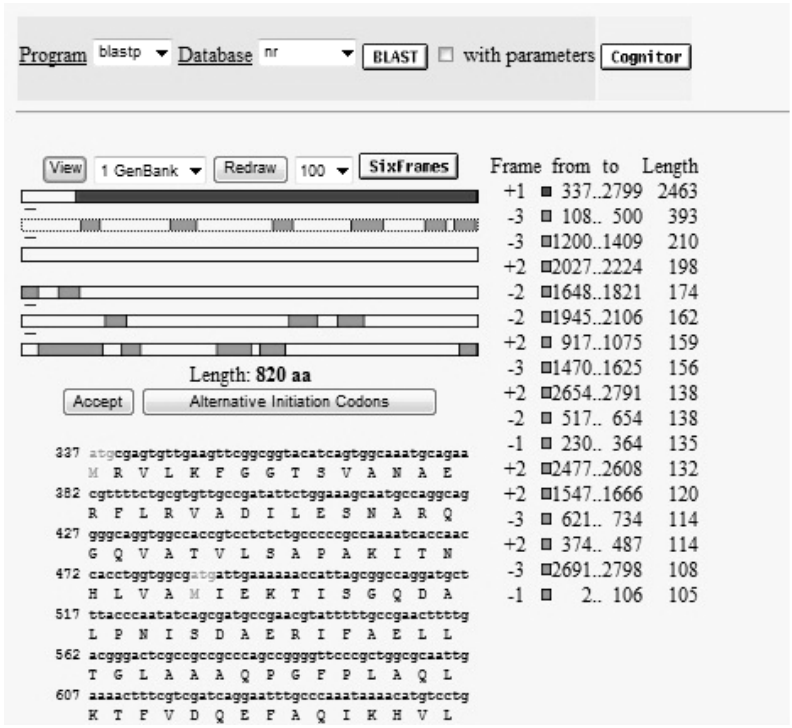


图 4.8　预测最长 ORF 的编码氨基酸

Accession	Description	score
ZP_07176840.1	homoserine dehydrogenase [Escherichia coli MS 200-1] >gb|EFJ61	1687
ABI99494.1	bifunctional aspartokinase I/homeserine dehydrogenase I [Escheric	1687
NP_414543.1	fused aspartokinase I and homoserine dehydrogenase I [Escherichia	1687
YP_001742118.1	bifunctional aspartokinase I/homoserine dehydrogenase I [Escheric	1686
EFU52714.1	homoserine dehydrogenase [Escherichia coli MS 153-1]	1686
EGV45125.1	bifunctional aspartokinase I/homoserine dehydrogenase I [Escheric	1685
EGB81574.1	homoserine dehydrogenase [Escherichia coli MS 60-1]	1685
ACX41207.1	aspartate kinase [Escherichia coli DH1] >dbj|BAJ41858.1| bifuncti	1685
ZP_03071390.1	aspartokinase/homoserine dehydrogenase I [Escherichia coli 101-1	1685
YP_002396091.1	bifunctional aspartokinase I/homoserine dehydrogenase I [Escheric	1685
AAL90885.1	aspartokinase I-homoserine dehydrogenase I [Escherichia coli]	1684
ZP_07222897.1	homoserine dehydrogenase [Escherichia coli MS 78-1] >gb|EFK71	1684
EFU57654.1	homoserine dehydrogenase [Escherichia coli MS 16-3]	1683

图 4.9 预测最长 ORF 编码氨基酸序列 blastp 结果页面

(2) 启动子及转录因子结合位点分析

DNA 序列中储存着调控信息，其中转录调控控制基因的转录活性。真核基因转录调节主要通过顺式作用元件和反式作用因子的相互作用而实现。启动子是一段 RNA 聚合酶识别、结合和起始转录的特定 DNA 序列，属于顺式作用元件。转录因子结合位点（Transcription Factor Binding Site，TFBS）位于启动子中，是与转录因子结合的 DNA 序列，长度约为 5~20 bp，它们与转录因子相互作用进行基因的转录调控。识别基因的调控区序列特征是研究基因功能、基因的转录调控规律、识别新基因及解析基因组结构的途径之一。挖掘调控区序列特征信息的方法主要有同源匹配法和模式识别法。

原核基因启动子区具有明显共同一致的序列，真核基因启动子区与多种转录因子相互作用共同完成转录调控，其调控机制更加复杂。真核生物启动子的-25~-35 区域含有 TATA 序列，是 RNA 酶的识别区，可使转录精确地起始，称为核心启动子元件；-70~-80 区域含有 CCAAT 序列，-80~-110 区域含有 GCCACACCC 或 GGGCGGG 序列，这两个区域控制着转录的起始频率。TATA 框上游的保守序

列称为上游启动子元件（Upstream Promoter Element，UPE）或上游激活序列（Upstream Activating Sequence，UAS）。真核转录因子数量大、种类多、作用机制复杂。同一转录因子能够同时调控多个基因，虽然与不同基因序列的结合位点具有一定的保守性，但又存在一定的可变性，结合位点是较短的 DNA 片段，在整个基因组中会存在大量的重复序列，这些特点给正确识别 TFBS 带来一定的难度，也使得预测方法普遍存在较高的假阳性率。

启动子位点和转录因子结合位点信息储存在相关数据库中。EPD (Eukaryotic Promoter Database)是一个有注释的非冗余的真核生物 RNA 聚合酶 II （Pol II）启动子数据库，其中的转录起始位点（Transcription Start Site，TSS）都通过实验获得。TRANSFAC 是真核生物转录调控信息的数据库，收录的数据都经过实验验证，包含转录因子、转录调控关系以及转录因子结合位点等相关信息，涵盖的物种主要有人、酵母、线虫、拟南芥、果蝇、大鼠、小鼠等，它通过文献挖掘收集数据，具有较高的质量。

启动子和转录因子结合位点常用数据库网址如下：

EPD： http://www.epd.isb-sib.ch

TRANSFAC：http://www.gene-regulation.com/pub/

DBTSS： http://dbtss.hgc.jp/index.html

TRRD： http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/

一些保守的功能区如启动子、增强子、转录因子结合位点等可通过序列分析获得相应的序列特征信息。其分析工具能直接搜索目的 DNA 序列中是否含有已知位点的序列模式。Promoter Scan 由美国明尼苏达大学维护，根据转录因子结合序列同源性分析预测 DNA 中的启动子区；Promoter2.0 基于遗传算法的人工神经网络技术预测脊椎动物启动子区 Pol II 和其它调控因子结合位点的信息。比较常用的启动子识别工具还有：TfBlast（TRANSFAC BLAST），可以通过比对算法找出目标 DNA 序列中可能存在的转录因子结合位点；TESS（Transcription Element Search System）是转录元件搜索系统。

常用启动子、转录因子结合位点分析工具网址如下：

Promoter Scan：http://www-bimas.cit.nih.gov/molbio/proscan/

Promoter2.0： http://www.cbs.dtu.dk/services/Promoter/，

TfBlast： http://www.gene-regulation.com/cgi-bin/pub/programs/tfblast/

TESS ： http://www.cbil.upenn.edu/tess/

例 4.5 人类 ALB 基因序列启动子分析

以人类 ALB 基因的序列（序列号为 NC_000004.11）为例，选取该序列中第一外显子之前长 2000bp 和其 3'方向顺延 100bp 共计 2100bp 长度的序列进行启动子识别。应用 Promoter Scan（1.7 版本），粘贴 FASTA 格式提交序列，预测结果见图 4.10。Promoter Scan 程序默认预测阈值为 53.00，分值越高代表预测准确性越大。程序预测该序列启动子区位于正链的 1710 到 1960 之间，预测结果分值为 53.70，图中显示了可能与该区域结合的转录因子的名称、编号、位置及权重。预测该序列正链有 4 个转录因子（PEA1、Albumin_US2、 CTF/NF-1、CTF）的结合位点：点击转录因子编号（TFD #）列可得到该转录因子的详细信息。由于目前启动子的预测方法存在较高的假阳性率，因此为提高识别率，在进行启动子预测时最好同时参考基因结构信息，如 CpG 岛、外显子/内含子等信息。

Proscan: Version 1.7
Processed Sequence: 2100 Base Pairs
Promoter region predicted on forward strand in 1710 to 1960
Promoter Score: 53.70 (Promoter Cutoff = 53.000000)
Significant Signals:

Name	TFD #	Strand	Location	Weight
PEA1	S01595	+	1892	1.539
AP-1	S00982	-	1898	1.613
Albumin_US2	S00627	+	1926	50.000
CTF/NF-1	S00696	+	1944	1.765
CTF	S00301	+	1945	2.993

图 4.10 Promoter Scan 预测 NC_000004.11 序列启动子结果

(3) CpG 岛识别

哺乳类动物基因组中 5% ~10%是 CpG(二核苷酸)，CpG 的聚集称 CpG 岛；其中 70% ~80%呈甲基化状态，称为甲基化的 CpG(mCpG)。人类和小鼠分别有 55.9%和 46.9%的基因与 CpG 岛有着密切的关联。CpG 岛经常在脊椎动物基因的 5'区域被发现，主要位于基因的启动子和第一外显子区域，这一特点有助于基因的识别。CpG 岛同时是表观遗传学中重要的作用区域，CpG 岛甲基化是基因转录活性的调控因素之一，CpG 岛甲基化异常常伴随着疾病发生。

CpG 岛的 GC 含量达到 55%、CpG 二核苷酸的出现率（观测值与期望值的比率）达到 0.65 且长度不少于 500 bp，符合这三个条件的 DNA 序列更趋向于分布在基因的 5'端区域。传统的 CpG 岛识别方法依据上面三个序列特征：GC 含量、序列长度、二核苷酸的出现率；另一主要方法是基于统计学特征的识别方法，如马尔柯夫链和隐马尔柯夫链识别 CpG 岛。

EMBL 提供的 CpG 岛的计算工具是在线分析软件 EMBOSS 的 CpGPlot/CpGReport/ Isochore，基于传统的窗口滑动法，一般默认的 CpG 岛跨度至少为 200bp，GC 含量>50%，CpG 出现频率>0.6，符合这些参数的区域都默认为 CpG 岛。其它分析工具还有 CpG Island Searcher、 CpGcluster 等。

CpG 岛识别的在线分析工具网址如下：

EMBOSS： http://www.ebi.ac.uk/embosscpgplot/

CpG Island Searcher：http://cpgislands.usc.edu/

CpGcluster： http://bioinfo2.ugr.es/CpGcluster/。

例 4.6 人类 TERT 基因序列 CpG 岛分析

以人类 TERT 基因序列（序列号 NG_009265.1）为例，选取序列中 4000-5300bp 区域共 1301 个碱基序列，包含第一个内含子序列。参数设置中 Program 选择 cpgplot 以直方图形式显示预测结果；Window 是计算 CG 含量和 CpG 岛的窗口大小，默认为 100；Step 是窗口移动的碱基数，默认为 1；Obs/Exp 是最小平均观察值与期

望值之比，默认为 0.6；MinPC 是每 10 个移动窗口中最小的平均 CG 含量，默认为 50；Length 设定 CpG 岛的最小长度，默认为 200；Reverse 用于选择预测目标序列的反向序列，默认为 no；Complement 用于选择预测目标序列的互补序列，默认为 no；参数设置见图 4.11。

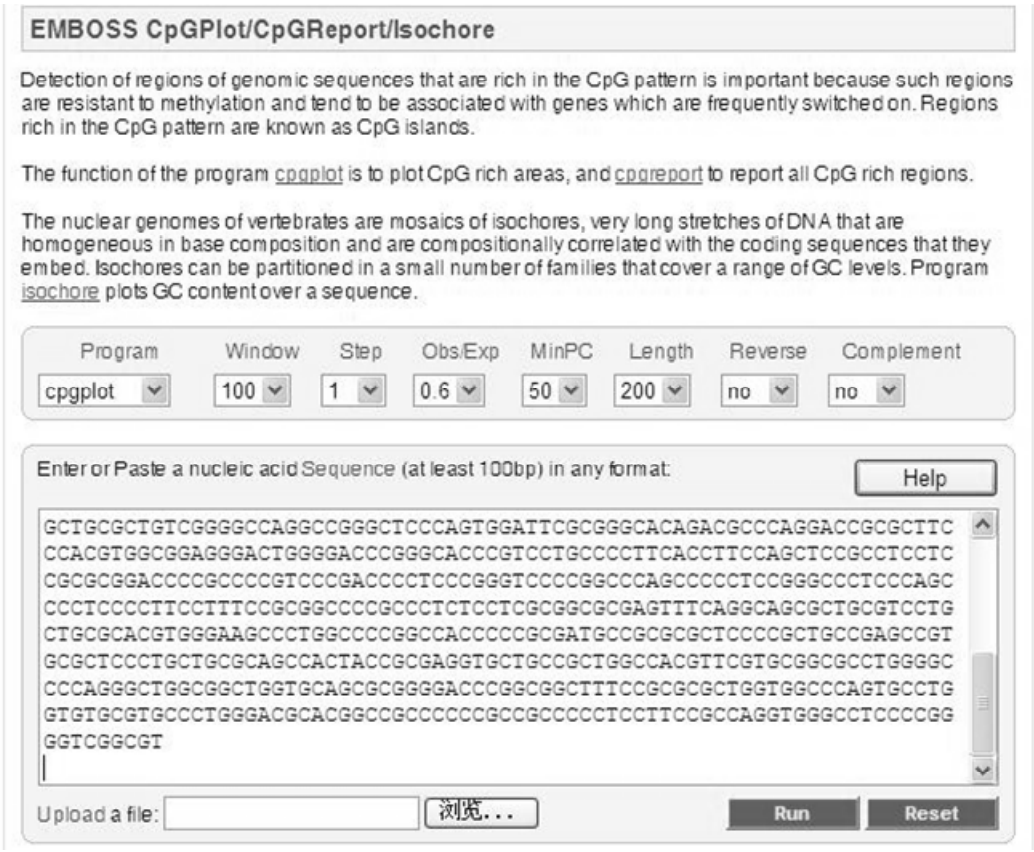
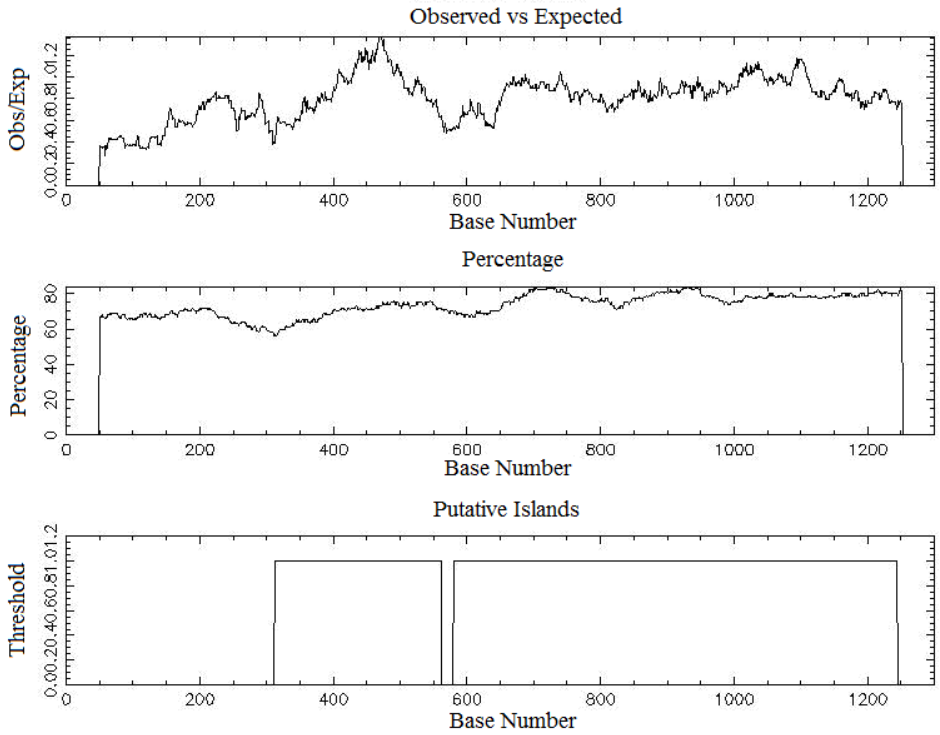


图 4.11 EMBOSS 在线分析软件页面

预测结果如图 4.12 所示。图中上部显示预测的参数；中部显示了三个直方图，包括 Obs/Exp 图、CG 含量百分比图及以上面两个图综合判断而得出的 CpG 岛预测结果图；最后给出预测结论，该序列可能存在两个 CpG 岛，分别为长 250bp 的 313-562 区段和长 665bp 的 581-1245 区段。

EMBOSS CpGPlot Results

Cpgplot Results	
Window	100
Step	1
Obs/Exp	0.6
MinPC	50
Length	200
Reverse	no
Complement	no
Cpgplot output	cpgplot-20101220-0318166257.output
Cpgplot PNG	cpgplot-20101220-0318166257.1.png
SUBMIT ANOTHER JOB	



CPGPLOT islands of unusual CG composition
EMBOSS_001 from 1 to 1301
Observed/Expected ratio > 0.60
Percent C + Percent G > 50.00
Length > 200
Length 250 (313..562)
Length 665 (581..1245)

图 4.12 NG_009265.1 序列 4000-5300bp CpG 岛预测结果

4.2 基因组结构注释分析

基因组序列主要构成成分是基因序列、重复序列和基因间序列。基因组时代最重要的工作之一是对海量的基因组数据进行注释，基因组注释包括基因组结构注释和基因组功能注释。基因组结构注释中的核心是基因识别，而为了提高基因识别效率需要首先寻找并屏蔽重复的和低复杂性的序列。本节将主要介绍基因组重复序列分析和基因识别方法。

4.2.1 重复序列分析

(1) 重复序列基本概念

重复序列（repetitive sequence）是指在基因组中不同位置出现的相同或对称性序列片段。重复序列在顺式调控元件如启动子、增强子、终止子处被大量发现，在真核生物中广泛分布。重复序列大致可以分成三类，即低度重复序列(1ow repetitive sequences)、中度重复序列(moderately repetitive sequences)和高度重复序列(highly repetitive sequences)。对于真核生物的核酸序列而言，在进行基因识别之前首先应该把大量的简单重复序列标记出来并去除，避免重复序列对预测程序产生干扰，尤其是涉及数据库搜索的程序。

(2) 重复序列分析方法

不同重复序列数据库储存了不同类型的重复序列信息。美国遗传信息研究所（GIRI）的 RepBase 是常用的真核生物 DNA 重复序列数据库；ALU 数据库是人类及其他灵长类代表性的 Alu 重复片段，可以通过 NCBI 的 BLAST 序列搜索程序检测序列中的 Alu 序列；其他重复序列数据库有 LINE-1 数据库、短的串联重复序列数据库 STR 等。RepeatMasker 是比较常用的重复序列分析工具，由美国华盛顿大学维护，通过与已知重复序列数据库比对搜索基因组序列中的相似序列，用于识别、分类和屏蔽重复元件，包括低复杂性序列和散在重复。

重复序列分析常用数据库和分析工具网址：

RepBase：http://www.girinst.org/repbase/

RepeatMasker：http://www.repeatmasker. org/

LINE-1：http://line1.bioapps.biozentrum.uni-wuerzburg. de/

STR：http://www.cstl.nist.gov/div831/strbase/

例 4.7 人类 15 号染色体 RP11-79C23 克隆序列重复序列分析

以人类 15 号染色体 RP11-79C23 克隆序列（序列号 AC138701.3）为例，使用 RepeatMasker 程序进行重复序列分析。登录 RepeatMasker 主页面，进入 RepeatMasker 分析页面。搜索引擎有 Cross_match、ABBlast 及 RMBlast 三种：Cross_match 在三者中速度慢但精度高；ABBlast 速度快精度略低；RMBlast 是 NCBI Blast 工具的兼容版。

RepeatMasker 分析结果如图 4.13 所示，分为两部分。第一部分是总体结果显示，包括序列长度、GC 含量及总屏蔽率。如图 4.13 所显示，AC138701.3 序列长度为 145239 bp，GC 含量为 36.56%，被屏蔽的碱基数是 95308bp，占全序列的 65.62%。第二部分是详细的被屏蔽重复序列的说明，包括重复序列的种类如 SINEs、LINEs、LTR elements、DNA elements、Total interspersed repeats、Small RNA、Satellites，Simple repeats、Low complexity 等，以及每种中含有的元件数量、长度和所占序列的百分比。如图 4.13 中 SINEs 的元件数为 50，长度为 12274bp，占总序列的 8.45%；LINEs 的元件数为 33，长度为 29189 bp，占总序列的 20.10%；LTR 元件数为 23，长度为 11823 bp，占总序列的 8.14 %等。综合结果显示 AC138701.3 序列中各种重复元件的覆盖率为 65.62 %。

Summary:

file name: RM2sequpload_1288248422
sequences: 1
total length: (145239 bp excl N/X-runs)
GC level: 36.56 %
bases masked: 95308 bp (65.62 %)

	number of elements	length occupied	percentage of sequence
SINEs:	50	12274 bp	8.45 %
ALUs	43	11392 bp	7.84 %
MIRs	7	882 bp	0.61 %
LINEs:	33	29189 bp	20.10 %
LINE1	26	27148 bp	18.69 %
LINE2	6	1906 bp	1.31 %
L3/CR1	1	135 bp	0.09 %
LTR elements:	23	11823 bp	8.14 %
ERVL	6	3960 bp	2.73 %
ERVL-MaLRs	7	2215 bp	1.53 %
ERV_classI	9	4636 bp	3.19 %
ERV_classII	1	1012 bp	0.70 %
DNA elements:	10	3348 bp	2.31 %
hAT-Charlie	4	560 bp	0.39 %
TcMar-Tigger	4	2325 bp	1.60 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats		56634 bp	38.99 %
Small RNA:	1	104 bp	0.07 %
Satellites:	5	34407 bp	23.69 %
Simple repeats:	25	2252 bp	1.55 %
Low complexity:	38	1911 bp	1.32 %

图 4.13 RepeatMasker 对 AC138701.3 序列分析结果

4.2.2 基因识别

(1) 基因识别基本概念

基因识别是识别基因组序列中的编码基因。随着测序方法的进步，更多的物

种基因组测序不断完成，基因识别是基因组研究的基础性工作。生物信息学中基因识别的方法主要包括同源性方法、隐马尔可夫模型、人工神经网络、动态规划法、基于规则的识别方法、语义学的方法和决策树方法等。

原核基因为连续基因，结构简单，其编码区是一个完整的 DNA 片段，见图 4.14；真核基因比原核基因复杂，具有复杂的调控机制，编码区是非连续的，被内含子分割为若干个小片段，见图 4.15。两者基因结构不同，故应用的基因识别方法也不同。

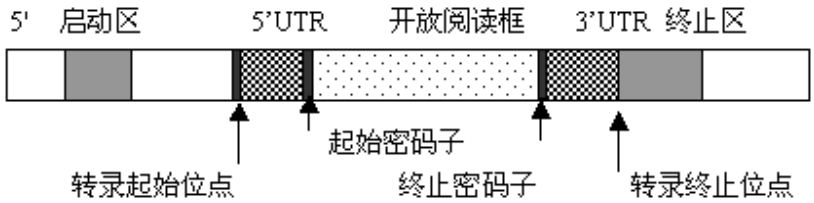


图 4. 14 原核基因结构（引自孙啸，陆祖宏，谢建明．生物信息学基础）

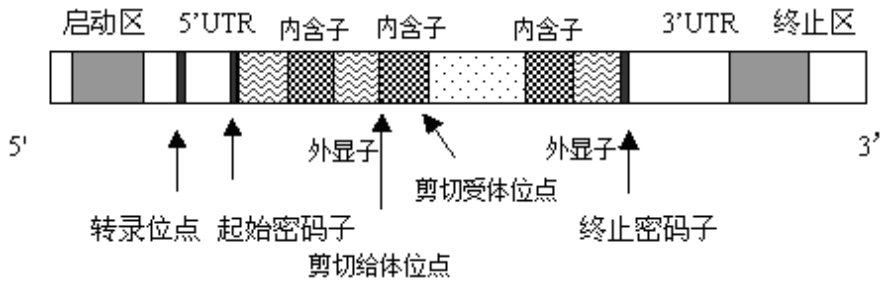


图 4.15 真核基因结构（引自孙啸，陆祖宏，谢建明．生物信息学基础）

(2) 基因识别方法

原核生物基因识别常用方法：一是基于序列同源性的算法，使用序列比对工具 BLAST 或 FASTA 来进行数据库搜索，利用已知序列的信息识别基因；一是基于序列组成统计学特征的算法，也称为从头预测（*ab initio*）方法，利用编码区组成特性和一些功能位点信息识别基因。

原核基因识别常用工具：GeneMarkS 是由美国乔治亚理工学院开发，采用迭代隐马尔可夫模型（iterative Hidden Markov model）方法的识别工具；Glimmer

是美国马里兰大学开发的基于插入式马尔可夫模型（Interpolated Markov Models，IMMs）的识别工具。

真核基因识别常用方法：一是基于特征信号的识别方法，利用真核编码区域一些具有特征的序列信息，例如：上游启动子区特征序列（TATA box、CAAT box、GC box）；5'端外显子位于核心启动子 TATA 盒的下游，含有起始密码子；内部的外显子两端的给体位点和受体位点；3'端外显子的下游包含终止密码子和 polyA 信号序列。综合多个序列特征信息确定外显子的边界，识别编码区域。一是基于统计学特征的方法，对已知编码区进行统计学分析找出编码规律和特性，通过统计值区分外显子、内含子和基因间区域。统计学特征主要包括密码子使用偏好性和双联密码子出现频率。此外，真核基因识别也可以采用同源序列比较的方法获得编码区信息。在实际应用中常常联合几种方法，以提高识别效率。

真核基因识别常用工具 GENSCAN 是美国麻省理工大学开发的脊椎动物基因预测软件，它使用广义隐马尔可夫模型（GHMM）根据基因的整体结构进行基因预测，包括外显子、内含子、基因间区域、转录信号、翻译信号、剪接信号等信息，能在基因组 DNA 序列识别完整的外显子-内含子结构，能识别多个基因，具有同时处理正、反两条链的功能。其它常用识别工具有美国橡树岭国家实验室支持的 GRAIL，是利用神经网络技术同时组合各种编码度量的识别方法。

常用基因识别工具的网址：

GENSCAN：http://genes.mit.edu/GENSCAN.html

GRAIL：　http://compbio.ornl.gov/Grail-1.3/

GeneMarkS：http://opal.biology.gatech.edu/GeneMark/

Glimmer：　http://cbcb.umd.edu/software/glimmer/

例 4.8 人类 CD9 序列基因识别分析

应用 GENSCAN 在线分析工具，分析类人 CD9 序列（序列号 AY422198）基因结构。登陆 GENSCAN 主页，物种选择 Vertebrate（脊椎动物），判别阈值为 1.00，序列名称中填写 cd9 AY422198，预测选项选择 Predicted peptides only，序列框中

粘贴序列，点击“Run GENSCAN” 运行，如图 4.16 所示。分析结果见图 4.17，显示该序列被预测出的 10 个外显子的信息。

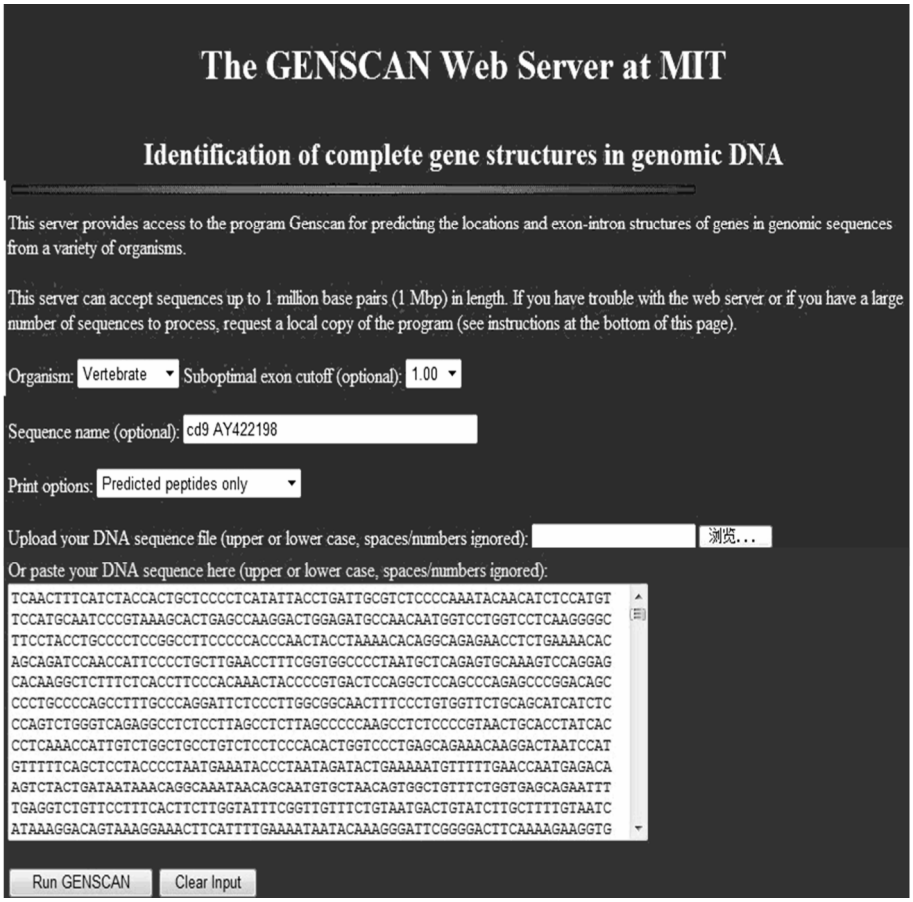


图 4.16 GENSCAN 在线分析界面

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	2030	2299	270	1	0	98	39	306	0.436	23.67
1.02	Intr	+	7489	7614	126	0	0	123	72	74	0.908	10.18
1.03	Intr	+	20012	20123	112	1	1	106	49	40	0.454	1.85
1.04	Intr	+	26834	27067	234	0	0	29	94	212	0.525	13.56
1.05	Intr	+	34168	34265	98	2	2	115	105	100	0.964	14.13
1.06	Intr	+	34948	35022	75	0	0	115	92	107	0.997	13.51
1.07	Intr	+	36765	36923	159	2	0	103	15	291	0.736	23.48
1.08	Intr	+	37012	37101	90	0	0	79	92	96	0.757	9.29
1.09	Intr	+	37728	37811	84	2	0	43	101	229	0.947	19.72
1.10	Term	+	39299	39364	66	1	0	110	40	67	0.937	2.04
1.11	PlyA	+	39776	39781	6							1.05

图 4.17 AY422198 序列分析结果界面

图中主要参数如下：

Gn.Ex：gene number, exon number (for reference)

Type: **Init** = Initial exon (ATG to 5' splice site)

Intr = Internal exon (3' splice site to 5' splice site)

Term = Terminal exon (3' splice site to stop codon)

Sngl = Single-exon gene (ATG to stop)

Prom = Promoter (TATA box / initation site)

PlyA = poly-A signal (consensus: AATAAA)

S：DNA strand (+ = input strand; - = opposite strand)

Begin：beginning of exon or signal (numbered on input strand)

End：end point of exon or signal (numbered on input strand)

Len：length of exon or signal (bp)

Fr：reading frame (a forward strand codon ending at x has frame x mod 3)

Ph：net phase of exon (exon length modulo 3)

I/Ac：initiation signal or 3' splice site score (tenth bit units)

Do/T：5' splice site or termination signal score (tenth bit units)

CodRg：coding region score (tenth bit units)

P：probability of exon (sum over all parses containing exon)

Tscr：exon score (depends on length, I/Ac, Do/T and CodRg scores)

GenBank 给出 AY422198 序列编码区信息如下，包含 8 个外显子：

CDS join(2030..2095, 26959..27067, 34168..34265, 34948..35022, 36765..36863, 37012..37101, 37728..37811, 39299..39364)

预测结果和 GenBank CDS 信息的对比见表 4.1，有 6 个外显子完全匹配，GENSCAN 多识别出两个外显子，另有两个外显子的 3'或 5'端位置预测出现偏差，这与 GENSCAN 特性有关。

表 4.1　AY422198 序列基因 GENSCAN 预测结果与 GenBank 对比

外显子编号	预测结果（碱基位置）	GenBank CDS	对　比
1.01 Init +	2030——2299	2030..2095	5'端匹配
1.02 Intr +	7489——7614	—	不匹配
1.03 Intr +	20012——20123	—	不匹配
1.04 Intr +	26834——27067	26959..27067	3'端匹配
1.05 Intr +	34168——34265	34168..34265	匹配
1.06 Intr +	34948——35022	34948..35022	匹配
1.07 Intr +	36765——36923	36765..36863	匹配
1.08 Intr +	37012——37101	37012..37101	匹配
1.09 Intr +	37728——37811	37728..37811	匹配
1.10 Term +	39299——39364	39299..39364	匹配

4.3 RNA 序列分析

RNA 既是携带遗传信息的主要生物大分子，也是重要的功能单位。RNA 包括 mRNA、tRNA、rRNA 三种主要形式，参与蛋白质的生物合成；还包括微小 RNA（miRNA）、小干扰 RNA（siRNA）等，参与生物调控。mRNA 属于编码 RNA，miRNA 、siRNA、tRNA、rRNA 属于非编码 RNA。本节将重点介绍 mRNA 和 miRNA 部分生物学特征的分析方法。

4.3.1 mRNA 可变剪接分析

(1) 基本概念

可变剪接也称选择性剪接，是指生物体中的单一基因能够生成多个转录本，从而产生远多于基因数目的蛋白质，完成机体的复杂功能及精细调节。大多数高等真核生物的基因都存在可变剪接的现象。

可变剪接受时间和空间的限制，在不同的组织中，在相同组织的不同细胞中，在同一组织的不同发育阶段，对病理过程的不同反应等过程中均会产生不同的剪接变体。有研究表明 94％以上的人类基因存在可变剪接，其中多达 50％的致病突变会影响剪接。可变剪接的异常改变使得基因在转录后期产生异常的剪接变体，编码出异常的蛋白质，导致人类遗传疾病甚至癌变。可变剪接种类形式多样，主要通过以下几种方式产生 mRNA 前体的产物，见图 4.18。

①外显子跳跃(exon skipping) ，也称为外显子加入或转移，是可变剪接最常见的一种方式。

②5'端或 3'端的可变剪接(alternative 5' or 3' splicing)，基因的 5'端或 3'端外显子被有选择地延长或缩短。

③外显子互斥(mutually exclusive exons)，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。

④多个外显子选一(one-of-N)，从一个大的外显子集合中选出其中一个外显子，作为可变剪接的产物。

⑤选择性起始(alternative initiation)，同一个基因使用两个或两个以上不同的启动子，在 5'端就会产生不同的转录产物。

⑥选择性终止(alternative termination)，同一个基因中出现两个或两个以上的 poly A 位点，使转录产物有不同的 3'端。

⑦内含子保留（intron retention），在剪接的过程中有的内含子被保留下来，起到了外显子的作用。

在这几种可变剪接方式中外显子跳跃最为常见，选择性终止的方式比较常见，外显子互斥方式相对较少见，内含子保留方式是最少见的。

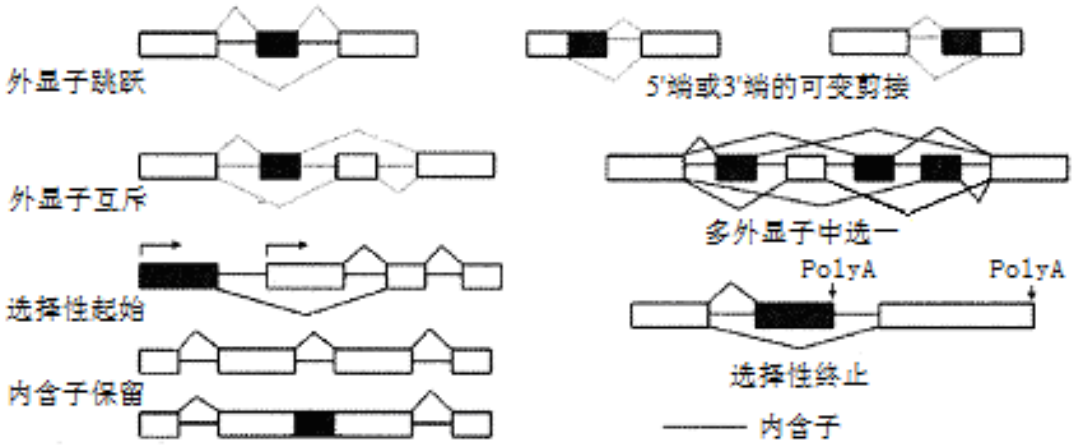


图 4.18 可变剪接种类（林鲁萍等. 基因选择性剪接的生物信息学研究概况.）

(2) 分析方法

可变剪接数据资源根据数据来源的不同分成两大类：一类基于文献报道的数据库，通过收集、整理已有的实验数据和文献报道建立的数据库；另一类基于 EST 数据的可变剪接数据库，主要是采用 EST 序列数据与基因组或 DNA、mRNA 序列进行比对的方法，发现新的或已经存在的可变剪接形式后建立的数据库或数据集。以下是常用的可变剪接数据库：

ASTD（Alternative Splicing and Transcript Diversity Databases）可变剪接和转录多样性数据库，由 ASD 数据库发展形成，2012 年整合至 Ensembl 基因组数据库。提供人、小鼠、大鼠、斑马鱼、线虫、果蝇等多物种可变剪接数据，是目前常用的可变剪接数据库。

ASD (Alternative Splicing Database)包括了多种模式生物的可变剪接数据，

ASD 由以下三个子数据库组成：AEDB (Alternative Exon Database)，经实验验证的可变外显子；AltExtron，由 EST 与全基因组序列比对得到的可变剪接数据；AltSplice，收集了通过计算的方法得到的可变剪接事件及其模式。

ASAP（the Alternative Splicing Annotation Project）数据库是通过全基因组范围内比对 EST 数据得到的人和小鼠的可变剪接数据库。ASAP 提供基因的外显子、内含子结构、可变剪接、组织特异性可变剪接、可变剪接产生的蛋白质异构体等信息。

可变剪接常用数据库网址：

ASTD：http://www.ebi.ac.uk/asd/index.html

ASD： http://www.ebi.ac.uk/asd/

ASAP：http//www.bioinformatics.ucla.edu/ASAP

剪接位点的精确定位是确定真核生物基因结构的关键，生物实验主要采用外显子连接芯片和外显子芯片等高通量技术，这些方法具有探针设计和数据分析相对复杂、无法识别未知剪接位点的局限性。生物信息学已开发出包括从头预测法、基于 EST/cDNA 序列比对法和基于 RNA-seq 数据识别等多种方法。从头预测算法主要采用支持向量机、概率模型、隐马尔可夫模型、神经网络和二次判别分析法等技术预测剪接位点。可变剪接过程的调控机制具有多样性，主要由剪接因子与调节蛋白相互作用来进行调节，剪接因子主要有外显子增强子(ESE)、外显子抑制子(ESS)、内含子增强子(ISE)和内含子抑制子(ISS)。

目前常用应用工具都结合可变剪接调控因子预测进行可变剪接的分析，如 ASPicDB 就是一个由可变剪切产生的带注释的转录本和蛋白变异体数据库，能够在基因、转录本、外显子、蛋白质或剪切位点水平进行分析，提供两类蛋白质类型（球状蛋白和跨膜蛋白）及有关定位、PFAM 结构域、信号肽（signal peptides）、跨膜片段和卷曲螺旋片段的信息。ESEfinder、RESCUE-ESE 等用于外显子增强子的预测。

可变剪接预测工具网址：

ASPicDB： http://t.caspur.it/ASPicDB/index.php

ESEfinder： http://rulai.cshl.edu/cgi-bin/tools/ESE3/

RESCUE-ESE：http://genes.mit.edu/burgelab/rescue-ese/

例 4.9 人类 TP53BP1 基因可变剪接分析

登陆 ASPicDB 主页，见图 4.19。进入“Search”页面，有三种检索方式：按基因、关键词和基因本体（Gene Ontology）进行检索，见图 4.20。本例应用关键词方式，在查询框中键入 TP53BP1，查询结果得到 1 个 TP53BP1 结果列表，见图 4.21，表中显示该基因有 17 个可变剪接转录本，14 个蛋白变异体。点击“ASPic Results”中的 show 链接可得到人类 TP53BP1 基因 17 个预测转录本和 14 个预测蛋白变异体，结果如图 4.22 和 4.23 所示。在转录本列表信息中包括外显子个数、序列长度、编码区域、编码长度、异构体类型等信息。在预测的 17 个转录本中有三个转录本（TP53BP1.tr14、TP53BP1.tr15、TP53BP1.tr17）没有蛋白序列信息。

结果中还包括基因结构图（Gene Structure View）、预测的剪接位点（Predicted Splice Sites）及内含子列表（Intron Table）等信息。

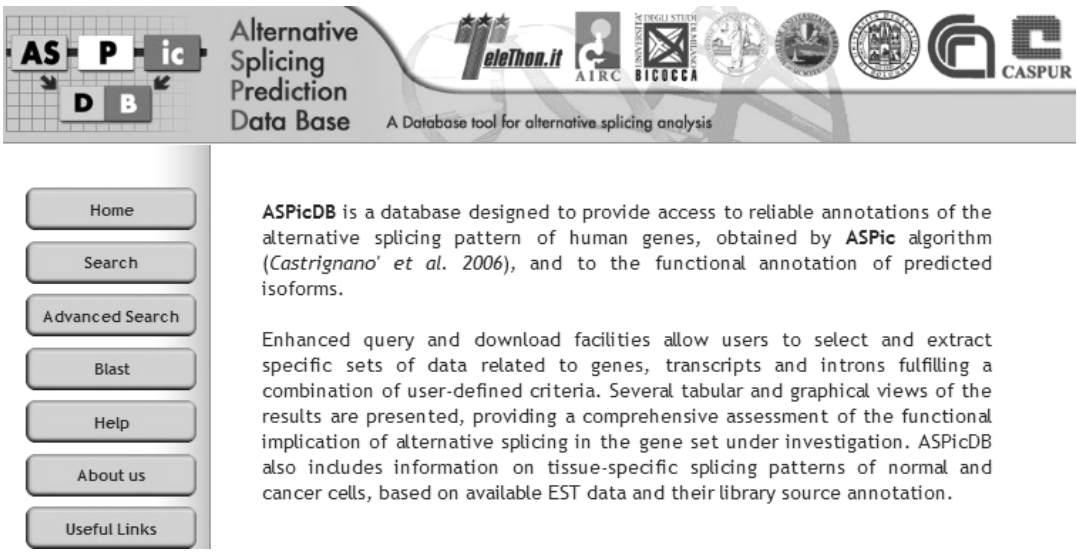


图 4.19 ASPicDB 主页面

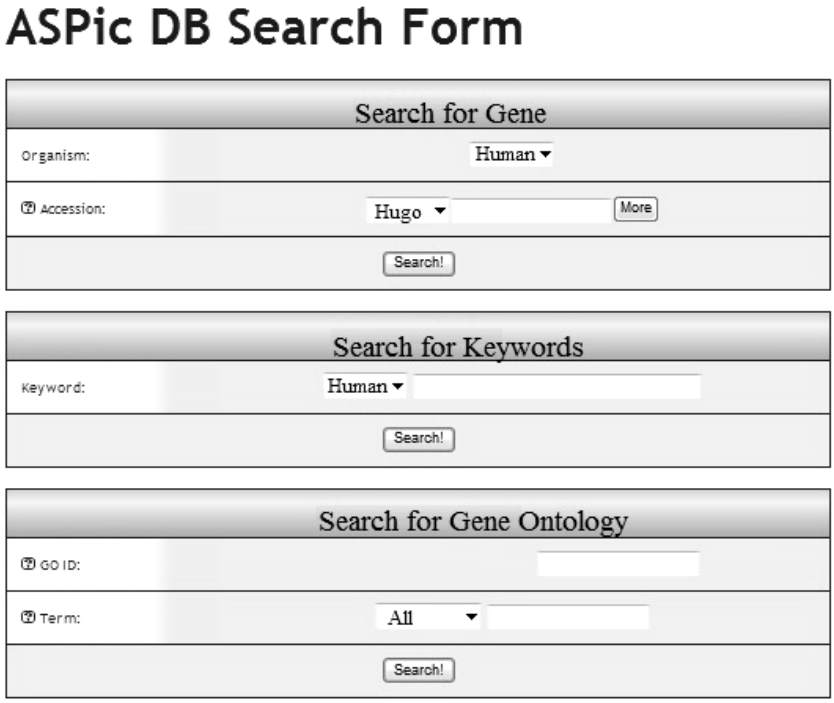


图 4.20 ASPicDB 分析页面

Gene Results

Results found: 1

Accession	Description	Organism	Coordinates	Aspic Results	Alt. Trans	Alt. Proteins
TP53BP1 Hs.440968 NM_001141979	tumor protein p53 binding protein 1	human	chr: 15 start:41449550 end:41590218 strand: R	show	17	14

图 4.21 基于 ASPicDB 的人类 TP53BP1 基因查询结果

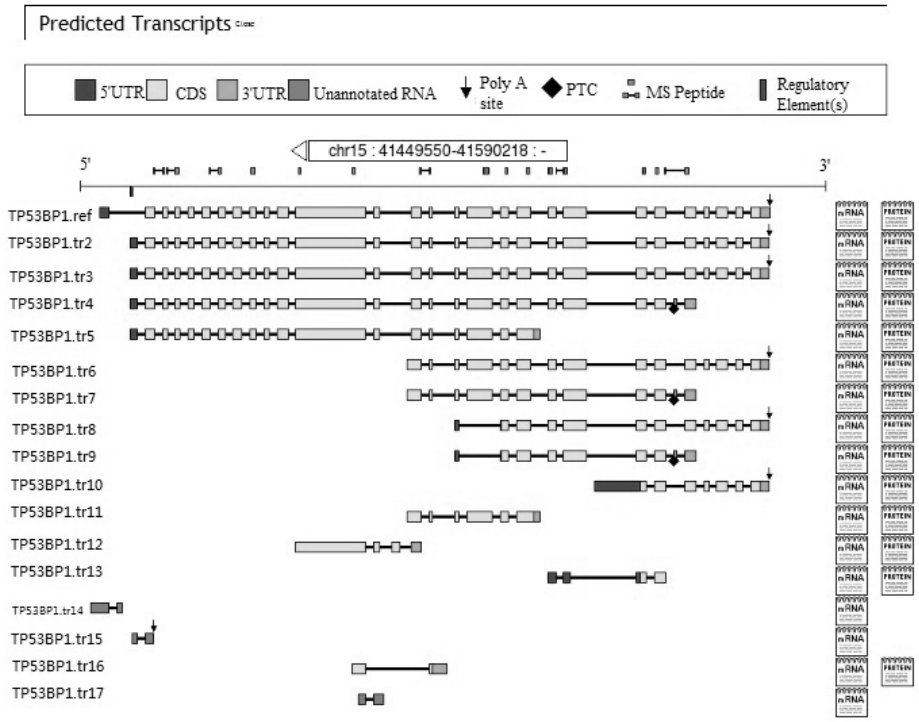


图 4.22 基于 ASPicDB 的人类 TP53BP1 基因 17 个转录本结果

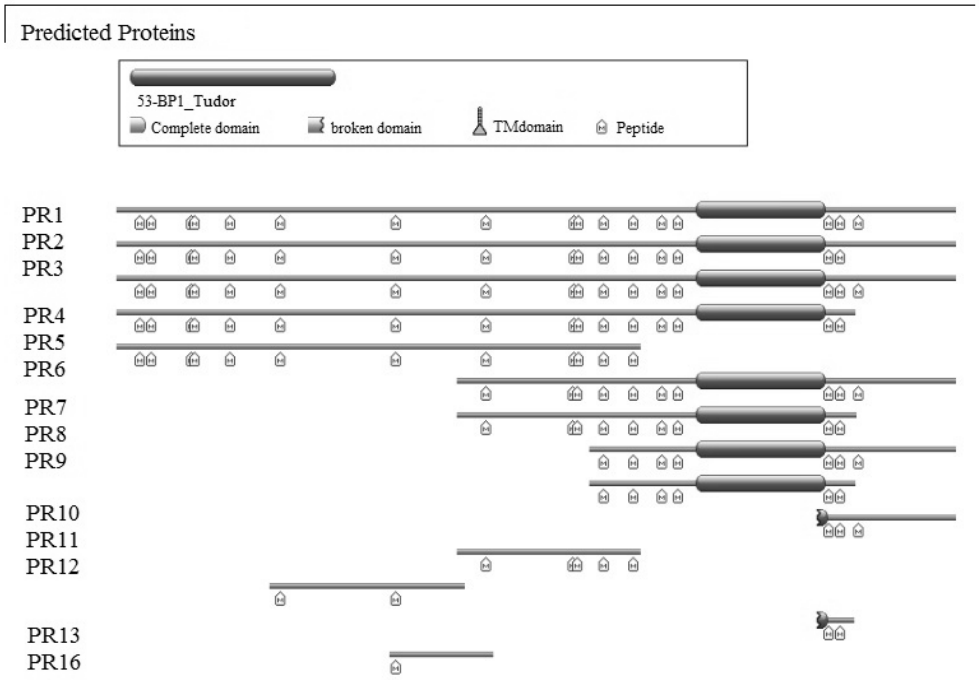


图 4.23 基于 ASPicDB 的人类 TP53BP1 基因预测的 14 个蛋白变异体结果

4.3.2 miRNA 与靶基因预测分析

(1) 基本概念

miRNA 归属小 RNA 范畴，广泛存在于真核生物中，是长度约为 19～24nt 的内源性非编码单链 RNA，不具有开放阅读框，不编码蛋白质，进化上具有高度的保守性。miRNA 通过 RNA 诱导的沉默复合物（RISC）与靶基因 3'非翻译区（3'UTR）结合，导致靶基因 mRNA 降解或者抑制其翻译，从而调控基因转录后的表达。

编码基因在核内产生长度为几百到几万 nt 的初始 miRNA（pri-miRNA），被一种多蛋白复合物剪切生成长度为 60-70nt、具有茎环二级结构的单链前体 miRNA（pre-miRNA），pre-miRNA 在胞质中经 Dicer 酶及其辅因子加工形成 19-24nt 的 miRNA 及其互补体。miRNA 前体在各个物种间具有高度的进化保守性，茎部保守性最强，环部可以容许更多的突变位点存在。

miRNA 与其靶基因间是多对多的关系：一个 miRNA 可调控多个靶基因，一个基因也可受多个 miRNA 调控。miRNA 的作用机制取决于 miRNA 与靶 mRNA 的互补程度，包括完全互补型和不完全互补型。miRNA 与靶 mRNA 完全互补致靶基因 mRNA 降解，不完全互补致靶基因 mRNA 翻译抑制，而不影响其稳定性。

miRNA 通过作用于相应靶基因 mRNA 完成生物学功能，如个体发育的调控、参与细胞分化和组织发育、调控基因表达等，miRNA 异常与疾病发生具有相关性。

(2) 分析方法

miRNA 分析主要包括 miRNA 预测和 miRNA 靶基因预测两方面。

① miRNA 预测

miRNA 主要通过 cDNA 克隆测序和计算预测两种方法获得。早期克隆测序直接、可靠，但很难克隆出在不同时期表达或只在特定组织或细胞系中表达的 miRNA，由于它的固有局限性，也很难捕获表达丰度较低的 miRNA。近年来随着该研究的发展，生物信息学预测 miRNA 方法成为一条重要辅助途径，优势是不受 miRNA 表达的时间和组织特异性以及表达水平的影响。

常用 miRNA 预测方法主要有 5 种：a）同源片段搜索方法，将已知 miRNA 或 pre-miRNA 序列在自身或其它相近基因组中用比对算法搜索同源序列，结合序列二级结构特征进行筛选；b）基于比较基因组学的预测方法，依据进化过程中的保守性在多物种中搜索潜在的 miRNA；c）基于序列和结构特征打分的预测方法，根据已知 miRNA 序列和结构的特征对全基因组范围中能形成茎环结构的片段进行筛选，是发现非同源、物种特异 miRNA 的方法；d）结合作用靶标的预测方法，依据 miRNA 与其靶基因序列间的碱基互补配对的保守性的特点预测 miRNA；e）基于机器学习的预测方法，通过对阳性 miRNA 和阴性 miRNA 数据集的训练来构建区分两者的分类器，根据所得分类器对未知序列进行预测，其中支持向量机（SVM）方法是目前 miRNA 分类和预测最常用的机器学习方法。常用 miRNA 预测软件有 MIRSCAN、MiPred、miRFinder 等。

② miRNA 靶基因预测

miRNA 通过与靶基因 mRNA3' UTR 不精确互补配对使靶 mRNA 裂解或抑制其翻译，二者相互作用以 miRNA:mRNA 二聚体结构形式存在。miRNA 序列 5'端的 2-8nt 称为种子区域，在 miRNA 靶基因预测中起主导作用。种子区域具有保守性，与靶 mRNA 序列能较好地互补配对结合，且在不同物种中靶序列也是保守的，这些特征是靶基因预测方法的重要依据。miRNA 靶基因预测方法主要有 2 类：基于种子区域互补和保守性的规则预测，常用软件有 miRanda、Targetscan 等；基于机器学习方法训练参数进行靶基因预测，常用软件有 PicTar、miTarget 等。

③ miRNA 数据库资源

miRBase 是集 miRNA 序列、注释信息以及预测的靶基因数据为一体的数据库，是目前存储 miRNA 信息最主要的公共数据库之一；TarBase 数据库是存储真实 miRNA 与靶基因间关系的数据库；miRGen 是整合了 miRNA 靶基因数据、基因组注释信息以及位置关系的综合数据库。miRNA 分析常用数据库和预测软件网址如下：

miRBase：http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5

TarBase： http://diana.cslab.ece.ntua.gr/tarbase/

miRGen： http://www.diana.pcbi.upenn.edu/miRGen/v3/miRGen.html

MIRSCAN：http://genes.mit.edu/mirscan/

MiPred： http://www.bioinf.seu.edu.cn/miRNA/

miRFinder：http://www.bioinformatics.org/mirfinder/

miRanda： http://cbio.mskcc.org/mirnaviewer

Targetscan： http://www.targetscan.org/

PicTar： http://pictar.mdc-berlin.de/。

参考文献

[4.1] R.M.特怀曼著，陈淳，徐沁等译．高级分子生物学要义．北京：科学出版社，2000.
[4.2] Roberts RJ, Vincze T, Posfai J, et al. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Research, 2010, 38: 234-236.
[4.3] 侯琳,钱敏平,朱云平等．转录因子结合位点生物信息学研究进展．遗传, 2009, 31(4): 365-373
[4.4]Prestridge, D.S.. Predicting Pol II Promoter Sequences Using Transcription Factor Binding Sites. J. Mol. Biol. 1995, 249:923-32.
[4.5] Zhao Z, Han L. CpG islands: algorithms and applications in methylation studies. Biochem Biophys Res Commun. 2009, 382(4): 643-5.
[4.6] 李霞．生物信息学．北京：人民卫生出版社, 2010.
[4.7] 孙啸，陆祖宏，谢建明．生物信息学基础．北京：清华大学出版社, 2005.
[4.8] Burge, C. and Karlin, S. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 1997, 268: 78-94.
[4.9] 林鲁萍，马飞，王义权．基因选择性剪接的生物信息学研究概况．遗传, 2005, 27(6): 1001-1006.
[4.10] 王科俊，吕俊杰，冯伟兴等．可变剪接与疾病的生物信息学研究概况．生命科学研究, 20l1, l15 (1): 86-94.
[4.11] Martelli PL, D'Antonio M, Bonizzoni P, et al. ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing.Nucleic Acids Res. 2011, 39: 80-5.
[4.12] 侯妍妍，应晓敏，李伍举. microRNA 计算发现方法的研究进展．遗传, 2008, 30(6): 687-696.
[4.13] Li Li, Jianzhen Xu, Deyin Yang, et al. Computational approaches for microRNA studies: a review. Mamm Genome, 2010, 21: 1-12.