第 1-5 章　基因预测与功能注释

第一节　基因组序列构成与基因预测

一、基因组序列的基本构成

一个生物体的基因组是指一套染色体中完整的 DNA 序列。例如,生物个体二倍体体细胞由两套染色体组成,其中一套 DNA 序列就是一个基因组。也就是说,对于单倍体细胞,基因组是指编码序列和非编码序列在内的全部 DNA 分子。对于有性生殖物种的基因组,通常是指一套常染色体和两种性染色体的序列。基因组包括核基因组、线粒体基因组和叶绿体基因组等。

基因组 DNA 序列看似简单,其实其构成很复杂。真核生物核基因组一般包括 35-80% 比例的重复序列和约 5% 的蛋白编码序列,这些编码序列分布于整个基因组区域。相对而言,染色体中心粒附近重复序列多而编码序列分布少。一个蛋白质编码基因往往包含多个外显子或蛋白质编码序列,外显子被非编码的内含子隔开。如何从基因组序列中确定这些编码基因是生物信息学一个重要任务。基因组上除了重复序列和少量蛋白质编码序列,其余大量为非编码序列。非编码序列构成异常复杂,包括结构 RNA,如 tRNA、rRNA、snRNA(small nuclear RNA)以及调节 RNA。调节 RNA 会转录非编码 RNA 序列(如非编码小 RNA 和长 RNA),转录出来的 RNA 序列以多种形式参与编码基因表达,发挥重要的调控功能(第 2-4 章将重点介绍非编码 RNA 的预测);许多非编码序列包含假基因(特别是人类基因组),它们原来是编码序列,但由于进化过程中碱基变异等,丧失了编码蛋白质的功能。

以人类和水稻基因组为例。人类核基因组由 24 条不同染色体(1～22 号常染色体和 X、Y 两条性染色体)所对应的 24 个不同 DNA 分子所构成,30 多亿个碱基对(3.2×10^9bp),其中内含 2.0 万～2.5 万个蛋白质编码基因;线粒体基因组约长 16.6kb 长度,含有 13 个编码基因和 24 个非编码基因(tRNA 和 rRNA)。人类基因组约 1.5% 序列为编码蛋白质的基因序列,约 5% 序列为非编码的调控基因序列。重复序列占人类基因组至少 50%;重复序列可根据其来源和分布特点分为串联重复序列和分散重复序列,后者约占基因组的 45%。水稻基因组有 12 条染色体,核基因组序列总长约 400Mb,蛋白质编码基因总数达 3.9 万个,平均基因长度 2.85kb,每个基因 4.9 个外显子;重复序列(TE)相关基因 1.69 万个,平均长度 3.22kb,每个 TE 基因平均有 4.2 个外显子。重复序列占整个基因组约 40% 左右,主要是逆转座子和 DNA 转座子类重复序列。

以上述两个基因组一段 DNA 序列为例(图 1-5.1),说明它们的序列构成。图中利用生物信息学工具——基因组浏览器显示约 50kb 长度基因组序列及其对这一区段的生物信息学注释结果(分别截自 http://rice.plantbiology.msu.edu 和 http://genome.ucsc.edu)。水稻基因组 50KB 的区段中包含有 9 个蛋白质编码基因,其中最后一个基因存在交替剪切情况;后

面两行标出了重复序列分布和基因表达情况(基于一个 20 天幼苗叶片的 RNA-SEQ 转录组数据)。人类基因组的 50KB 区段仅包含一个蛋白质编码基因,提供的信息包括基因结构(包括交替剪切)、基因转录本(mRNA)、甲基化程度、与其他物种的同源基因序列保守性、SNP 分布、各类重复序列分别情况等。

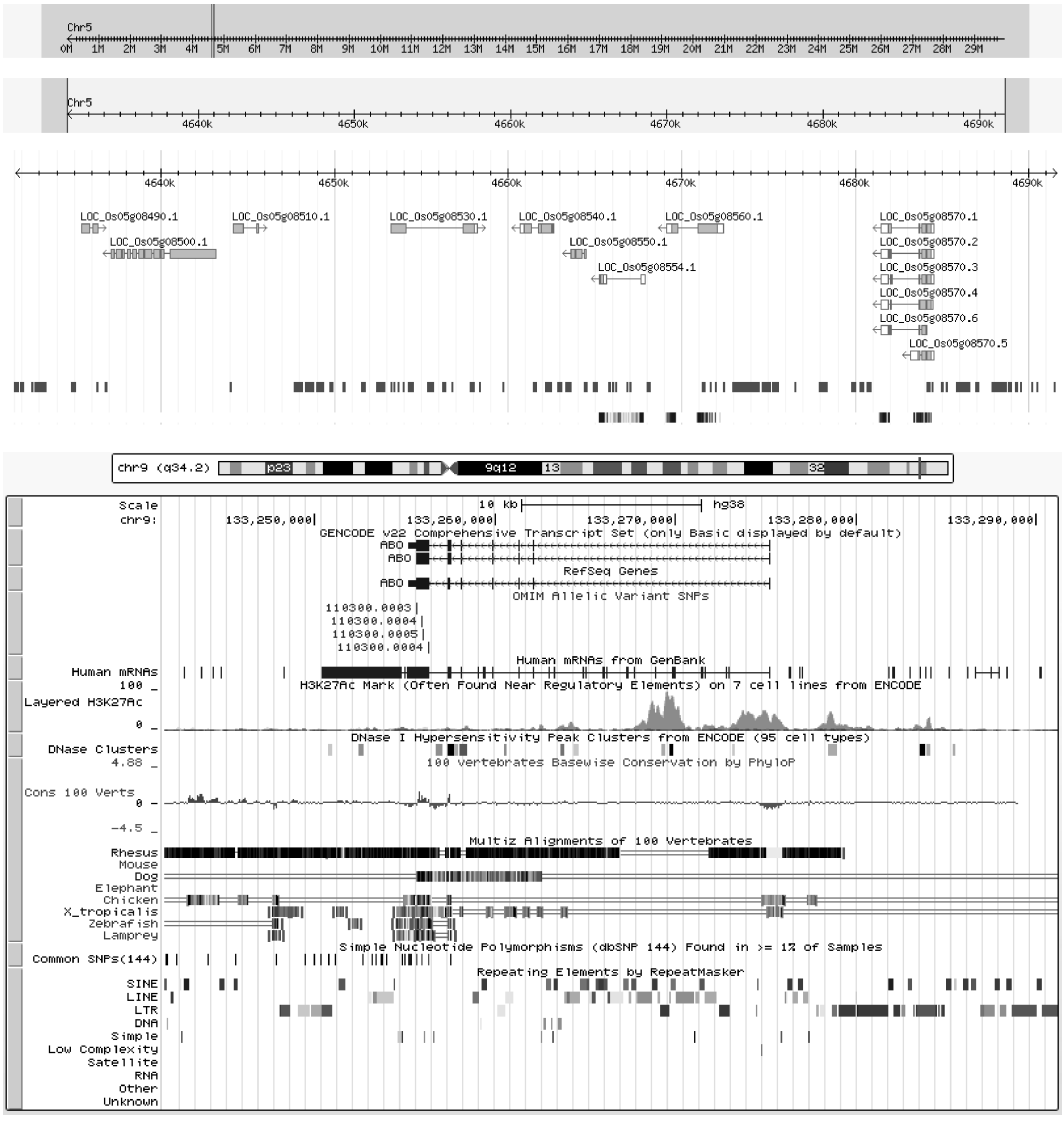


图 1-5.1 水稻(上)和人类(下)基因组序列构成列举(具体说明见文中)

不同物种基因组构成存在明显差异。上述植物(水稻)和人类基因组构成就存在明显差异,例如基因个数、基因密度、重复序列种类构成和假基因数量等,均存在明显不同。如果我们再看微生物(如酵母),其基因组构成明显不同于植物和人类,它们的基因组往往要小些,其重复序列比例明显不高(图 1-5.2)。即使是同一种类型生物,不同物种之间基因组构成也会千差万别,例如植物中的水稻和玉米基因组构成。玉米基因组由于物种分化后,转座子类重复序列大量增值,其基因组膨胀(约 2.5Gb),导致其基因组重复序列比例达到 85%以

上,远远高于其近缘同科物种水稻和高粱等基因组。

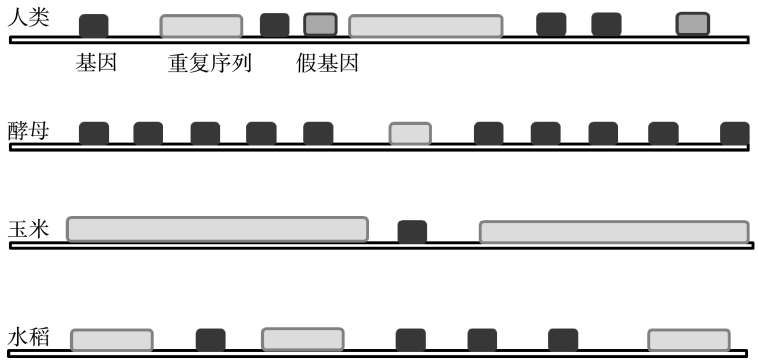


图 1-5.2　不同生物基因组构成比较模式图

图中仅标注出蛋白质编码基因、假基因和重复序列三种构成元素。

二、基因预测及其基本方法

在完成基因组序列拼接后,可以获得基因组的主要 DNA 序列,甚至可能是整个基因组各条染色体的序列。这些序列中包含有许多未知基因,将基因从这些基因组序列中找出来是生物信息学的一个重要任务。

基因组 DNA 序列上,一个蛋白质编码基因典型结构如图 1-5.3 所示。它包含编码和非编码序列,其编码序列(外显子)被非编码区(内含子)隔断,蛋白质编码区(CDS)包括大部分外显子序列(除了两端非翻译区域,即 UTR 序列)。从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)。基因表达后被转录成前体 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,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。

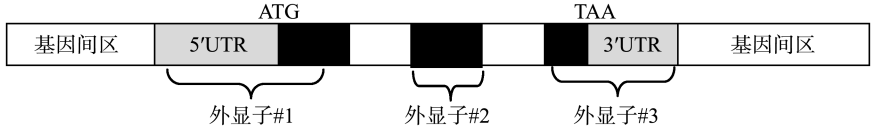


图 1-5.3　一种典型蛋白质编码基因的结构示意图

蛋白质编码区(CDS,黑色区域)包括大部分外显子序列(除了两端 UTR 序列),自起始密码子(ATG)开始,到终止密码子(TAA 等)结束。

所谓基因预测(genefinding)或注释(annotation)是指基因结构预测,主要预测 DNA 序列中编码蛋白质的区域(CDS)(图 1-5.3)。不过目前基因区域的预测,已从单纯编码区预测发展到整个基因结构的预测,如启动子、交替剪切等预测。基因预测并非易事,有许多因素会影响预测的准确性。例如(1)基因组 DNA 序列仅由 4 种碱基构成,其基因信号并不明显,背景噪音很大;(2)有些基因的外显子长度很短(如 3 个 bp 长度);(3)第一和最后一个外显子(包含 UTR 区域)预测尤其困难,无剪切信号可供判断;(4)基因存在大量交替剪切情况;(5)测序误差。不同类型生物基因构成特征存在差异,预测难点不同,如真核生物基因往往基因结构复杂,基因组上基因密度很低,存在大量交替剪切和假基因等;原核生物基因结构

简单,基因密度大,但其基因短,存在重叠基因等情况。

目前基因注释方法主要包括 2 大类:一个是同源比对方法,另一个是从头预测方法。这两种方法在实际应用中往往配合使用,即综合两种方法的预测结果,给出最终的预测结果。

同源比对方法(homology method)是利用近缘种已知基因进行序列比对,发现同源序列,并结合基因信号(外显子内含子剪切信号、基因起始和终止密码子等)进行基因结构预测(模式图见图 1-5.4a)。另外,通过测定目标物种转录组(RNA-seq)或其他基因表达序列(如早期的 EST 序列),可以获得大量目标物种转录本序列,将这些表达序列定位到基因组上,并结合基因信号,同样可以辅助基因编码区预测(图 1-5.4b)。

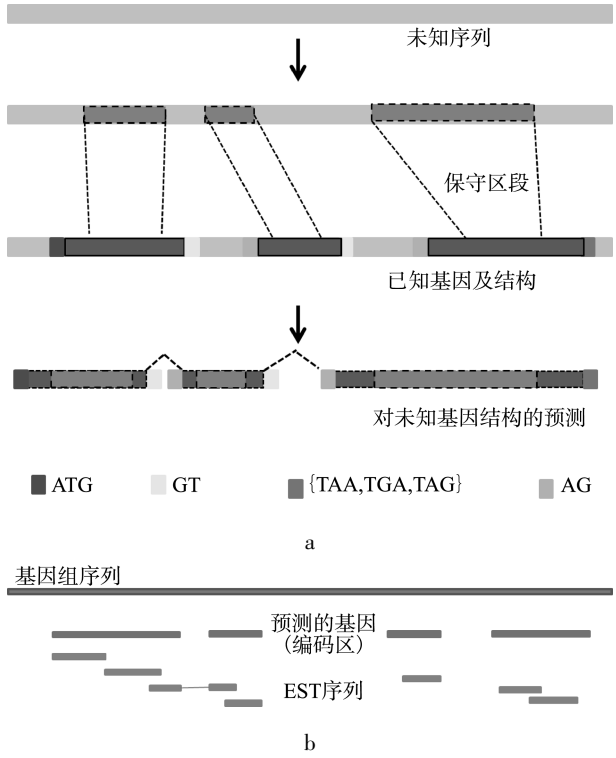


图 1-5.4 同源比对基因预测模式图

a:基于近缘物种已知基因结构;b:基于基因表达序列(如 EST)

从头预测方法(*ab initio* method)是生物信息学一个重要的研究领域,先后有一大批预测算法和相应程序被提出和应用。与同源比对方法不同,从头预测方法是根据编码区统计特征和基因信号进行基因结构的预测(图 1-5.5)。编码区特征的统计测验需要基于一定的基因模型。从头预测方法中,最早是通过序列核苷酸频率、密码子等特性进行预测(如 CpG 岛、最长 ORF 法等)。CpG 岛 (CpG island)一词是用来描述基因组中的一部分 DNA 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%,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,几乎所有的看家基因(housekeeping gene)及约 40%的组织特异性基因的 5′末端含有 CpG 岛,其序列可能落在基因转录的启动子及第一个外显子中。因此,在大规模基因测序中,如发现一个 CpG 岛,则预示可能在此存在基因。后来,一些其他方法陆续被提出,如隐马尔可夫模型(HMM)、神经网络(NN)、动态规划法(dynamic programming)等。

大约在上世纪 80-90 年代,HMM 模型用于基因预测应用研究开始出现,后续大量研究表明,HMM 模型用于基因预测表现良好。目前从头预测的主流方法均基于 HMM 概率模型(本章第 2 节将重点介绍)。基于概率模型算法,往往需要依赖于已知基因序列作为训练数据,如 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,对模型参数进行估计。由于训练所用序列的限制,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。

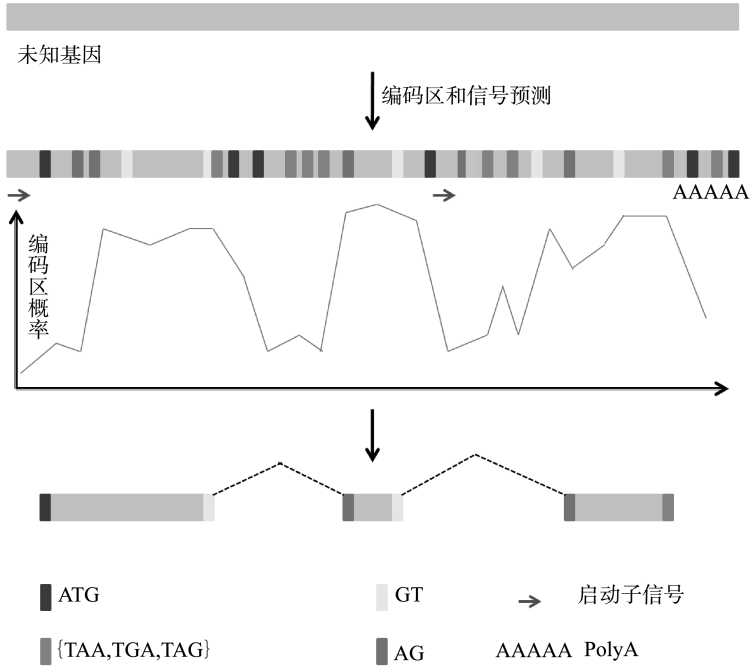


图 1-5.5　从头预测方法模式图

其中编码区概率估计往往根据一定概率模型(如 HMM)

近 20 年来,先后有一大批基因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。表 1-5.1 列出了部分目前基因从头预测主要工具及其相应算法。某一算法的优劣可以通过一定的标准,如敏感性(sensitive)和特异性(specifity)来衡量。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)对该序列进行预测,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(即预测准确)。则敏感性定义为 N_1/M,它表示程序预测的能力大小;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。

表 1-5.1 部分从头(*ab initio*)基因预测软件能力比较结果(Goel 等,2013)

程序名称	所用算法#	核苷酸层次&			外显子层次*				文献	软件网址
		敏感性	特异性	相关系数	敏感性	特异性	丢失的外显子	错误的外显子		
FGENESH	HMM	0.93	0.93	0.92	0.81	0.80	0.09	0.11	Salamov *et al*. 2000	http://www. softberry.com/berry.phtml
AUGUSTUS	HMM	0.88	0.93	0.89	0.72	0.84	0.20	0.08	Stanke *et al*. 2006	http://bioinf.uni-greifswald.de/webaugustus/
GENSCAN	HMM	0.94	0.89	0.90	0.78	0.74	0.08	0.14	Burge *et al*. 1997	http://genes. mit. edu/GENSCAN.html
GeneParser	DP	0.71	0.72	0.68	0.69	0.63	0.31	0.37	Synder *et al*. 1995	http://stormo.wustl.edu/src/GenParser/
Grail-I	NN	0.56	0.85	0.65	0.59	0.91	0.40	0.09	Xu *et al*. 1996	http://compbio. ornl.gov/grail-1.3

#:HMM:隐马尔可夫模型;DP:动态规划法; NN: 神经网络;

&:敏感性:真实编码序列被成功预测为编码序列的比例;特异性:预测为编码序列中确为编码序列的比例;相关系数:真实值和预测结果之间的相关性

*:敏感性:真实外显子被准确预测(包括拼接位点)的比例;特异性:预测为外显子的序列确为外显子的比例;丢失的外显子:未能预测出的真实外显子;错误的外显子:预测为外显子的序列实际不是外显子序列。

三、基因注释流程

在进行基因组序列注释过程中,一般会遇到两种情况:一是仅针对少量目标序列(如若干 BAC 克隆序列)进行基因注释,目的是为了解这些序列上可能的功能基因,二是针对一个新测序基因组进行全基因组水平的基因注释。对于第一种情况,可以利用在线开放基因预测平台和数据库搜索平台等,对目标序列逐条进行基因注释,这里就不再说明。下面仅对全基因组水平的基因注释过程进行描述。

2001 年 2 月,*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)的综述文章。文章对人类基因组序列如何解读进行了深入分析,比较全面地展示了当时对序列的理解能力和基因注释技术水平。经过 10 多年的发展,基因组测序和拼接技术已有很多变化,基因注释能力和工具已有很多改进。

基因组水平的基因注释往往需要本地化进行。以下对基因组基因注释过程进行简要说明:

在基因预测之前,一般首先会对全基因组进行重复序列鉴定和屏蔽。真核生物基因组中存在较高比例的重复序列。例如人类基因组上至少有 50%的重复区域。重复序列的存在对基因组注释的准确性会产生较大的影响,因此通常重复序列的鉴定是基因组注释的第一步。重复序列保守性很差,因而对不同物种都需构建该物种的重复序列库。由于有些基因在该物种中本身拷贝数很高(如组蛋白、维管蛋白等),容易误将这些基因上的部分片断当作

重复序列,导致最终无法预测出这些基因或基因结构预测不完整。因此,在构建的目标物种重复序列库中应排除掉这部分序列,即去除与已知物种基因相似性高的序列。在获得重复序列库后,可利用这部分序列将基因组中存在重复序列相似片段或区域“屏蔽”(mask)。所谓屏蔽就是将原序列中的“A、T、C、G”用“N”(hard mask)或小写的“a、t、c、g”(soft mask)表示,这样后续的基因预测软件将这部分序列按重复序列处理。对基因组中重复序列处理的好坏将直接影响后续基因注释的质量。

目前全基因组水平基因注释主要综合利用三种方法的预测结果:

(1) 从头预测

该方法的最大优势在于,其不需利用外部的证据来鉴定基因及判断该基因的外显子-内含子结构,而是利用各种概率模型和已知基因统计特征预测基因模型。然而这种方法的主要问题(a) 很多从头预测软件预测新物种基因时,是利用已有模式物种的基因进行统计参数估计。即使是非常相近的物种,它们之间的内含子长度、密码子频率、GC 含量等重要参数均会存在一定的差异。为了解决该问题,需要通过该物种的特定基因训练数据集获得统计参数。(b)足够的训练数据集可以在基因数量层次上保证准确,但内含子-外显子剪切位点的准确率仍然较低(60-70%)。

(2) 利用近缘物种已知基因蛋白序列进行同源比对获得间接证据

由于基因蛋白序列在相近物种间存在较高的保守性,因而这部分序列经常被作为基因注释过程中的主要证据,即将相近物种的已知蛋白序列联配到目标基因组上,获得这些蛋白序列在基因组上的对应位置,从而确定外显子边界。在这一过程中,选择高质量的物种注释结果作为辅助证据尤为关键,很多研究者由于引用了低质量的注释结果作为辅助证据,导致将注释错误从一个物种延续到另一物种。在软件工具选用方面,一般使用剪切位点识别度比较高的联配方法如 Spaln, Spidey 和 sim4 等软件,从而获得较为准确的外显子边界和剪切位点。

(3) 基于目标物种基因表达数据获得基因信息

在各种基因预测的证据中,转录组数据(如 RNA-seq)对基因注释的准确性提升有很大帮助。目前利用 RNA-seq 辅助注释的策略主要分为两种:①将 RNA-seq 数据独立拼接成转录本,然后将转录本定位到基因组上来确定基因的位置和结构;②直接将 RNA-seq 的读序数据联配到基因组上,然后再通过联配结果进行组装。目前对于这两种策略哪种更为准确看法不一,前者的主要问题在于 RNA-seq 本身的拼接质量,本身拼接的序列较短从而不能保证获得完整的转录本序列。目前三代测序技术已逐步可以解决该问题;对于后者,如果基因组中基因间隔很短,有时候会错误融合不同的基因。该策略的优势在于能够较为准确的确定剪切位点和外显子的边界。

当利用以上三种策略或工具完成注释后,会获得很多重叠或者有出入的基因结构。这时,可以通过基因注释整合工具,获得一个完整且较为准确的注释结果。目前使用较主流的整合工具为 EVidenceModeler(EVM)和 GLEAN。这类软件可以从各种来源的结构注释结果中选取最为可能的外显子,然后将它们合并整合成完整的基因结构。此外,Maker2 是一种将重复序列注释屏蔽、基因注释、注释结果整合等步骤综合一体的软件,目前也越来越被广泛运用于各种基因组注释项目。基因组注释,特别是复杂基因组注释一直是一个困难任务,大量研究人员还在不断开发新的工具,例如最新的利用云计算注释工具 xGBDvm,它可以进行

真核生物基因组的注释(Duvick 等,2016)。

经过上述步骤注释出来的基因集,通常还存在一定数量低质量的基因预测结果(假基因、ORF 太短等),需要再进行人工筛选。一般会过滤掉编码蛋白长度小于 50 个氨基酸、编码不完整、基因长度过长、基因中间存在大量'N'等情况的基因。

第二节 从头预测——隐马尔可夫模型(HMM)方法

如上所述,基因组序列基因结构预测大致分为两类方法。不同于同源比对方法,从头预测方法除了依据基因信号和蛋白质编码序列的统计特征外,一个有效概率模型往往非常重要,是保障真核和原核生物基因预测准确度的基础。目前在基因预测领域,主要应用的概率统计模型为隐马尔可夫模型(Hidden Markov Model, HMM)和神经网络等。下面重点介绍 HMM 方法。

一、马尔可夫和隐马尔可夫模型

马尔可夫模型,也叫马尔可夫过程或马尔可夫链(Markov chain),是俄罗斯数学家 Markov 在研究俄罗斯文学家普希金《奥涅金》作品不同音的出现规律时,于 1907 年提出的一个数学模型。它是研究随机过程中统计特征的一种概论模型。

假设存在这样一个随机变量序列(通常与时间有关),满足这样的条件:每个随机变量之间并非相互独立,并且每个随机变量只依赖序列中前面的随机变量。在很多类似的系统中,我们可以做出这样的假设:我们可以基于现在的状态预测将来的状态而不需要考虑过去的状态。也就是说,序列中将来的随机变量与过去的随机变量无关,它条件地依赖于当前的随机变量,这样的随机变量序列,通常称为一个 Markov 链,或者说这个序列具有 Markov 性质。其中所谓与过去状态无关,指的是先要由"过去"推导出"现状",由"现在"才能直接推导出"将来"。

马尔可夫过程是由一个个状态(所谓"态")构成,态之间的转换是以一定概率发生的。也就是说,"将来"与"现在"是通过一个概率去联系,同样"现在"与"过去"也是通过一个概率去联系,这样的概率称为转移概率。

对于一条 DNA 序列,我们可以构建一个简单的马尔可夫模型:

该模型中只有 4 个"态":A/T/G/C。对于一条 DNA 序列,它们之间以一定的概率转换。例如以下 DNA 序列:

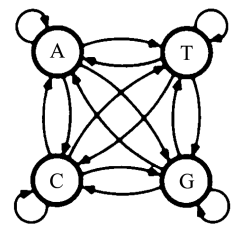


图 1-5.6

CTTCATGTGAAAGCAGACGTAAGTCA

从碱基 A 态向其他态(碱基)转移的次数如下:

A→T:1 次

A→G:3 次

A→C:1 次

在原状态转移(即 A→A):3 次

同样可以统计出其他碱基(态)之间的转换次数及其频率。

一个略为复杂的例子:为了建立识别一个基因内 5'端外显子和内含子间剪切位点方法,我们可以构建这样一个马尔可夫模型:

图 1-5.7

该模型除了起始和终止点,只有 3 个态:外显子、内含子和它们之间的 5' 端剪切位点。三个态之间的转换概率如图中标出。

基于马尔可夫概率模型,我们可以把任何一条序列概率描述为

$$P(x)=P(x_L,x_{L-1},\cdots,x_1)$$
$$=P(x_L|x_{L-1},\cdots,x_1)P(x_{L-1}|x_{L-2},\cdots,x_1)\cdots P(x_1)$$

如果我们限定任何位点 i 碱基的出现仅与其前一个碱基有关,即

$P(x_i|x_{i-1},\cdots,x_1)=P(x_i|x_{i-1})$,则上式

$$P(x)=P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})\cdots P(x_2|x_1)P(x_1)$$

作为马尔可夫模型的拓展并应用最为广泛的模型,隐马尔可夫序列模型是上世纪 60 年代 Leonard E. Baum 等人发展起来的。HMM 被广泛应用到多个领域,如上世纪 70 年代应用于语音识别;80 年代它首先被应用到序列分析中,用于序列联配(Bishop 和 Thompson 1986),而后其在生物信息学领域被广泛应用,在基因预测、功能域分析等方面得到很好应用,特别是基因组序列预测编码基因方面,取得了巨大成功,成为目前主流方法。

HMM 是一种用参数表示的用于描述随机过程统计特性的概率模型,是一个双重随机过程,由两个部分组成:马尔科夫链和一般随机过程。其中马尔科夫链用来描述状态的转移,用转移概率描述;一般随机过程用来描述状态与观察序列间的关系,用观察值概率描述。

继续以上述 5' 端外显子和内含子间剪切位点识别问题为例。其隐马尔可夫模型如下:

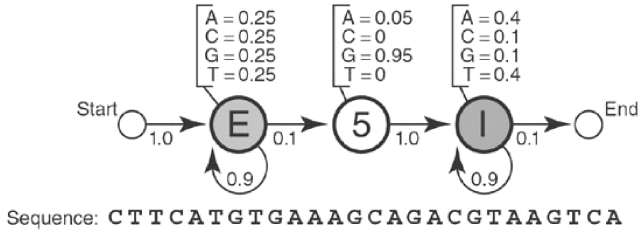


图 1-5.8

图中每个"态"的实际观察值有可能是 A/T/G/C 中任何一个,其观察值(碱基种类)有一个概率分布(根据训练数据集可获得),同时,3 个态的碱基概率分布并不一样。明显的,5' 端剪切位点上的碱基分布最多的是 G,也会出现 A,但 C/T 不会出现。

因此,对于 DNA 序列,基于 HMM 的解释是 DNA 序列任何一个位点的碱基是由一个由 A/T/G/C 四面体骰子随机产生的,每个位点都有一个自己的骰子,每个骰子产生的 A/T/G/C 概率不同。同时,DNA 序列特定位点出现什么样骰子符合马尔可夫模型特征,即它仅与序列中上一个位点的骰子有关。由于基因组序列虽然有其内在规律(如包含基因等),但总体上,序列的组成和分布具有很多随机性,也就是说基因的信号是很弱的,这种随机特征就使得 HMM 能够很好地解释 DNA 序列。

为何我们叫它隐马尔可夫模型,其"隐"何物? 如上所述,对于 HMM 模型,其状态转换

过程是未知的,不可观察的,我们能看到的就是以一定概率发生产生的特定观察值(对于DNA 序列就是具体碱基),因此,称之为"隐"马尔可夫模型。例如,对于上述 5'端外显子和内含子间剪切位点识别问题,对于某一条序列(如图例序列),其是否有真实剪切位点我们是看不到的,我们能看到的是其序列中 14 个位点包含 G 或 A,这 14 个位点都有可能是剪切位点,其中一个可能是真实的(如图 1-5.9)。我们的问题是如何判断这 14 个位点中哪个位点是真实的?或某一位点是真实剪切位点的概率是多少?

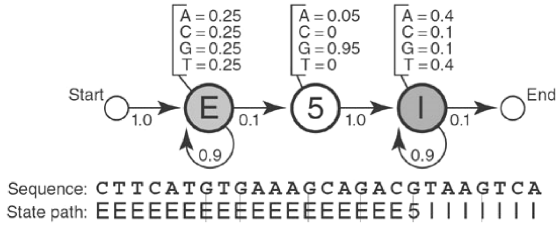


图 1-5.9

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)。如果假设当前状态与前面若干状态有关,这样就构成高阶马尔可夫模型。在不援引任何生物学机制的情况下,第 k 阶马尔科夫链假定在序列中某一位置上碱基的存在,只取决于前面 k 个位置上的碱基。1 阶链假定一个特定碱基存在于位置 i 的概率只取取决于在位置 $i-1$ 的 4 种碱基概率。相互独立的碱基所组成的序列即为 0 阶马尔科夫链。阶可以通过似然法估计。实际基因预测应用中,会使用高阶 HMM 模型,如 5 阶 HMM。

二、隐马尔可夫模型问题及其算法

隐马尔可夫模型在实际应用中会涉及 3 个基本问题,即评估问题(evaluation)、解码问题(decoding)和学习问题(learning)。评估问题是已知观察序列 O 和模型 λ,如何计算由此模型产生此观察序列的概率 $p(O|\lambda)$?解码问题是已知观察序列 O 和模型 λ,如何确定一个合理的状态序列,使之能最佳地产生 O,即如何选择最佳的状态序列。它是对观察值的最佳解释,揭示的是隐藏的马尔可夫模型的态序列。学习问题是如何根据观察序列不断修正模型参数,使 $p(O|\lambda)$ 最大。

针对上述 HMM 三个主要问题,已提出了相应的算法解决这三个问题:评估问题——向前和向后(Forward-backword)算法;解码问题——Viterbi 动态规划算法;学习问题——Baum-Welch 算法 (最大期望算法)。针对生物序列,我们往往会碰到大量评估问题和解码问题,例如找基因和功能域分析等。结合找基因问题,下节将具体介绍其中一个算法(Viterbi 算法)

三、HMM 基因预测模型及其应用

1. HMM 基因预测模型

HMM 是上世纪 90 年代最早在原核生物上用于基因预测。当时被用于大肠杆菌 *E.coli* 的基因预测(Krogh 等,1994),而在这之前,马尔科夫模型已在原核生物上被利用于基因预测(Borodovsky 和 McIninch,1993)。而后,HMM 被用于人类等真核生物基因组的基因预测(如 Burge 和 Karlin, 1997)。

如何构建一个基因组序列中蛋白质编码基因 HMM 模型? 一个简单 HMM 如下:

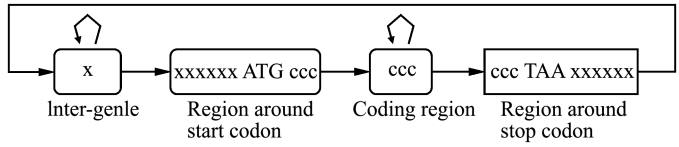


图 1-5.10

该模型把基因组序列看成一个包含 4 个"态"的随机过程,除了基因间区一个态(x),基因包含 3 个"态"(包含启始密码子区、编码区和包含终止子区)。但是,真实的情况要复杂得多,需要构建一个更加完备的模型才能涵盖基因组上的基因状况。为此可以构建如下 HMM 模型:

真核生物基因预测程序 GENSCAN 使用的 HMM 模型(Burge 和 Karlin, 1997)为最早也是当时最成功的基因预测算法及其程序。该 HMM 模型考虑了正负链、启动子区域、非编码 UTR 区、ploy-A 信号和单外显子基因等情况,把它们也分别作为"态"纳入模型中。实际证明,这样的模型取得了很好的预测效果。

在给定一条基因组序列,根据基因信号(如编码起始和终止密码子;外显子和内含子剪切信号等),有许多编码基因的可能性,我们如何确定最有可能的基因结构呢? 我们还是用上述 5' 端外显子和内含子间剪切位点查找的例子,来简单说明最后确定基因的基本过程:

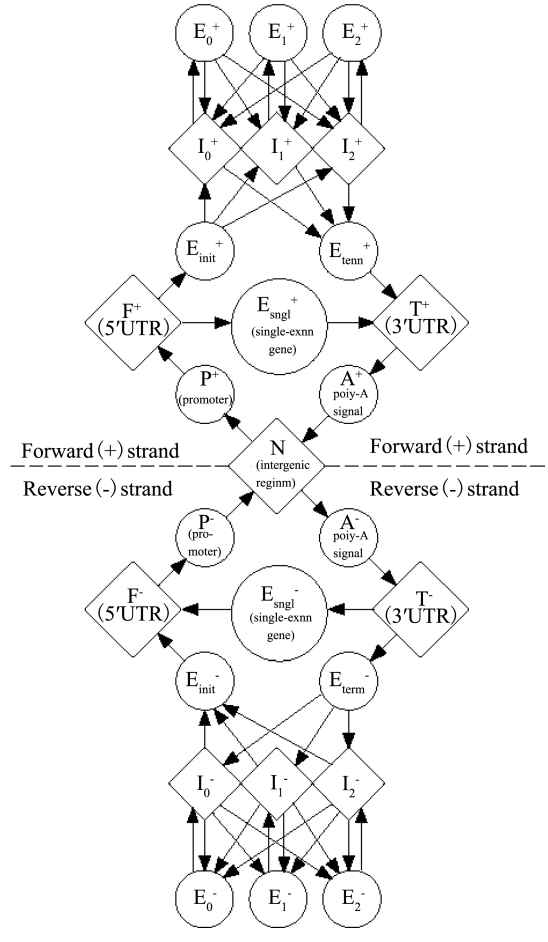


图 1-5.11　基因预测工具 GENSCAN 的 HMM 模型(Burge 和 Karlin, 1997)

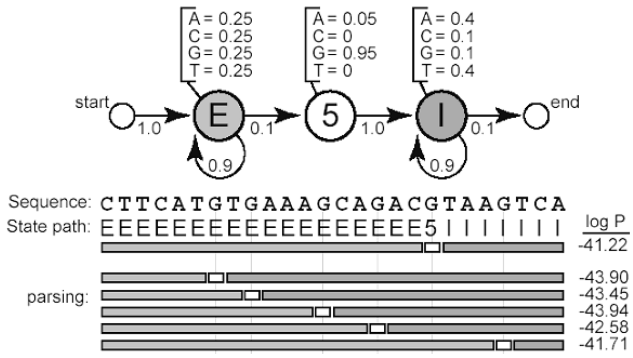


图 1-5.12

对于该例举序列,其序列内部有 14 个位点中包含 G 或 A,表明 14 个位点都有可能是剪切位点。对于每种可能位点,即每种马尔可夫链状态路径(parsing),我们基于它们转移概率和各个态的观察值概率,可以分别计算每条路径的发生概率:

$$P(S,\pi \mid HMM,\theta)$$

即具有参数 θ 的 HMM 模型,依据 π 路径产生的观察值即序列 S 的概率

$$\begin{array}{rl} x & \text{CTTCATGTGAAAGCAGACGTAAGTCA} \\ \pi & \text{EEEEEEEEEEEEEEEEEEEE5IIIIIII} \\ P(x_i \mid \pi_i) & \frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{95}{100}\cdot\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{1}{10}\cdot\frac{2}{5}\cdot\frac{1}{10}\cdot\frac{2}{5} \\ P(\pi_{i-1}\to\pi_i) & 1\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{1}{10}\cdot 1\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10} \end{array}$$

$$P(x \mid \pi) = P(\pi_0 \to \pi_1) \cdot \prod_{i=1}^{n} P(x_i \mid \pi_i) P(\pi_i \to \pi_{i+1})$$

对于第一个路径(最上面):

$$= \left(1 \times \frac{1}{4}\right)\left(\frac{1}{4} \times \frac{9}{10}\right)^{17}\left(\frac{1}{10} \times \frac{95}{100}\right)\left(\frac{2}{5} \times \frac{9}{10}\right)^{4}\left(\frac{1}{10} \times \frac{9}{10}\right)^{2}\left(1 \times \frac{2}{5}\right) \times \frac{1}{10}$$

$\text{Log}P = -41.22$(以 e 为底)

最下面的一条路径:

$$P(x \mid \pi) = P(\pi_0 \to \pi_1) \cdot \prod_{i=1}^{n} P(x_i \mid \pi_i) P(\pi_i \to \pi_{i+1})$$

$$= \left(1 \times \frac{1}{4}\right)\left(\frac{1}{4} \times \frac{9}{10}\right)^{21}\left(\frac{1}{10} \times \frac{95}{100}\right)\left(\frac{2}{5} \times \frac{9}{10}\right)\left(\frac{1}{10} \times \frac{9}{10}\right)\left(1 \times \frac{2}{5}\right) \times \frac{1}{10}$$

$\text{log}P = -41.71$

如此计算可以获得所有 14 个可能路径的发生概率(表 1-5.2).

实际应用中,有非常多的可能路径可以产生观察序列,这样往往需要一个动态规划算法——维特比算法(Viterbi algorithm)来获得最有可能的路径,即在给定的序列和 HMM 模型下给出 P 值最高的路径。

1967 年安德鲁·维特比(Andrew Viterbi)提出了维特比算法以解决解码问题。该算法假设给定 HMM 状态空间 S,初始状态 i 的概率为 π_i,从状态 i 到状态 j 的转移概率为 $a_{i,j}$。令观察到的输出为 $y_1,\ldots,y_T$。产生观察结果的最有可能的状态序列 $x_1,\ldots,x_T$。由递推关系

给出：

$$v_{1,k} = P(y_1 \mid k) \cdot \pi_k$$

$$v_{t,k} = P(y_t \mid k) \cdot \max_{x \in s}(a_{x,k} \cdot v_{t-1,x})$$

此处 $v_{t,k}$是前 t 个最终状态为 k 的观测结果最有可能对应的状态序列的概率。通过保存向后指针记住在第二个等式中用到的状态 x 可以获得维特比路径。声明一个函数 $Ptr(k, t)$，它返回若 $t>1$ 时，计算 $v_{t,k}$用到的 x 值或若 $t=1$ 时的 k。这样：

$$x_{t-1} = Ptr(x_t, t)$$

$$x_T = \text{argmax}_{x \in s}(V_{T,x})$$

2. HMM 基因预测模型的应用

原核生物基因的各种信号位点（如启动子和终止子信号位点）特异性较强且容易识别，因此相应的基因预测方法已经基本成熟。例如，Glimmer 是应用最为广泛的原核生物基因结构预测软件，准确度高，其应用的模型为内插值置换马尔科夫模型（Interpolated Markov Model，IMM）。然而，真核生物的基因预测难度则大的多。首先，真核生物中的启动子和终止子等信号位点更为复杂，难以识别；其次，真核生物中广泛存在可变剪切现象，使外显子和内含子的定位更为困难。因此，预测真核生物的基因结构需要运用更为复杂的算法，常用的有隐马尔科夫模型等，常用的软件有 Fgenesh、Augustus、GeneMark、SNAP、Genscan 等。

Fgenesh 是由英国 Sanger 中心的 Asaf 和 Victor 于 2000 年开发的，基于广义隐马尔可夫模型的真核生物基因预测软件。Fgenesh 软件对基因注释的准确性已经得到国际上认可，尤其是在植物基因预测方面应用非常广泛。该软件系列的成员还有 Fgenesh+、Fgenes、Fgenes-M、Fgenesh-M 和 Fgenesh_GC。其中 Fgenesh+是 Fgenesh 集成了蛋白比对和 cDNA 定位功能；Fgenes 是 Fgenesh 的前身，它主要采用线性判别式分析的方法来预测基因结构；Fgenes-M 和 Fgenesh-M 分别在 Fgenes 和 Fgenesh 的基础上集成了预测可变剪接的功能；Fgenesh_GC 则能够兼容非经典的 GC 剪接供体（在人类约占全部的 0.6%）。Fgenesh 为商业软件，由 Softberry 公司负责维护和发布。其在线版本（http://www.softberry.com/）仅可让用户输入单条序列进行基因注释。近些年来，Softberry 公司还开发了一套集成该公司各种软件的工具箱 MolQuest（http://www.molquest.com/molquest.phtml? topic=main），该工具箱可支持各种系统（有使用期限）。用户可以导入多条序列批量运行基因组注释工作，并且运行速度十分迅速。

Augustus 是由德国格赖夫斯瓦尔德大学数学与计算机科学学院的研究人员于 2006 年开发的。该软件目前自带了 75 个物种的基因模型参数，用户可以选择较近缘物种的模型参数来进行预测，其在线版本网址为 http://bioinf.uni-greifswald.de/augustus/submission.php。与 Fgenesh 不同，该软件是开源的，用户可以免费下载获取在本地环境下运行。

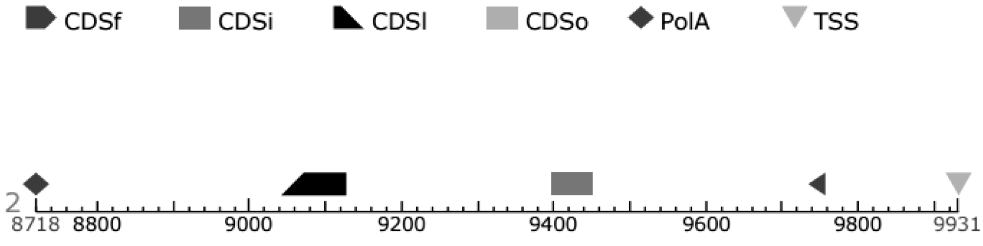
GeneMark（http://topaz.gatech.edu/GeneMark/）是由美国乔治亚理工大学研究人员于 1998 年开发。随着该团队对 GeneMark 工具包的不断更新，GeneMark 已包括一系列的软件，可用于不同类型物种的基因预测。例如 GeneMarkS 一般预测原核生物基因，GeneMark-ES 适合真核生物，MetaGeneMark 可用于宏基因组基因预测。该软件包与其他基于从头预测软件最大的不同在于，它可以利用目标物种基因组进行自我训练 HMM 参数，并用于后续注释。

以下举例说明如何进行一个基因组序列片段的基因预测：

一段来自番茄基因组约 120kb 基因组序列（GenBank 记录号 EU124734.1）需要进行基因

预测。比较简单的方法是将这段序列提交到 Fgenesh 的在线基因预测平台(http://www.softberry.com/)。该网站提供了近 300 个物种的预测基因的参数。如果物种来源已知,可以在网站上直接选择一致或相近的物种参数。如果序列来源未知,可将该序列先在 NCBI 上先进行搜索获知与这段序列来源最相近物种的信息,然后在 Fgenesh 网址上选择该物种进行基因预测。在本例子中,我们可以直接选择番茄的基因参数模型。Fgenesh 的结果报告很全面,提供了网页和 PDF 两种形式供用户查看。图 1-5.13 为 PDF 版本的部分结果截图,从图中我们可以获知这段 126,477bp 的番茄序列上,Fgenesh 共预测到包含 143 个外显子的 15 个基因。对于预测的每个基因,都会展示其在给定序列中位置及其结构。该软件默认的基因结构包括转录起始位点(TSS)、外显子/外显子区域(CDSf、CDSi、CDSl、CDSo)以及 PolyA 尾巴。在获得的预测结果中,还包括各个基因对应的 mRNA 序列和翻译后的蛋白序列。此外,基因结构注释结果文件的格式通常为 GFF3,而 Fgenesh、Augustus 等注释出来的格式均不统一,这时可以利用基因注释软件(如 EVM)中的脚本进行 GFF3 格式转换。

FGENESH 2.6 Prediction of potential genes in Tomato genomic DNA
Seq name: gi|157649042|gb|EU124734.1| Solanum lycopersicum chromosome 3 clone C03HBa0034
Length of sequence: 126477
Number of predicted genes 15: in +chain 5, in -chain 10.
Number of predicted exons 143: in +chain 50, in -chain 93.
Positions of predicted genes and exons: Variant 1 from 1, Score:2930.671875



2 -		PolA	8718		1.61			
2 -	1	CDSl	9040 -	9126	1.38	9040 -	9126	87
2 -	2	CDSi	9396 -	9449	7.82	9396 -	9449	54
2 -	3	CDSf	9733 -	9756	-3.28	9733 -	9756	24
2 -		TSS	9931		-4.08			

图 1-5.13 利用 Fgenesh 进行基因组序列编码基因预测例举

一段来自番茄~120kb 基因组序列(EU124734)的预测结果。共 15 个基因被预测出,图中仅列出其中一个基因的具体预测结果。图中 CDSf 代表基因模型中的第一个外显子;CDSi 代表中间的外显子;CDSl 表示最后一个外显子;如果仅有个外显子则用 CDSo 表示;TSS 和 PolA 分别代表转录起始位点和 PolyA 尾巴结构。

第三节 贝叶斯统计及其基因预测应用

一、贝叶斯统计与生物信息学

1. 贝叶斯统计简介

统计学中有两个主要学派,频率学派和贝叶斯学派。我们进行统计推测时,一般会涉及

三种信息:总体信息、样本信息和先验信息,上述两个学派对这三种信息的使用上有共同点也有不同点。频率学派或经典统计使用前两种信息,而贝叶斯统计基于三种信息进行统计推断。这里总体信息是指总体分布或总体所属分布族给我们的信息。例如,“总体是正态分布”,它提供我们许多信息。总体信息很重要,是我们统计推测的基础。样本信息是从总体抽取的样本给我们提供的信息。样本信息是最“新鲜”的信息,且越多越好,人们希望通过样本的加工和处理对总体的某些特征做出较为精确的统计推测。先验信息是抽样之前有关统计问题的一些信息,一般来说,先验信息主要来源于经验和历史资料。所以,贝叶斯统计与经典统计的主要差别在于是否利用先验信息。贝叶斯统计重视先验信息的收集、挖掘和加工,使它数量化,形成先验分布,加入到统计推测中来,以提高统计推断的质量。

贝叶斯统计源于英国数学家贝叶斯(T. R. Bayes, 1702-1761)发表的一篇论文“论有关机遇问题的求解”。在该论文中,贝叶斯提出著名的贝叶斯公式和一种归纳推理方法。贝叶斯方法长期未被普遍接受,直到二次大战后,在优化决策等领域开始不断被研究和完善,并陆续在工业、经济和管理等领域成功应用。如今,贝叶斯统计已趋成熟,已发展成一个有影响的统计学派,打破了经典统计学一统天下的局面。

贝叶斯学派的最基本观点是“任一个未知量 θ 都可看作一个随机变量,应该用一个概率分布去描述 θ 的未知状态”。这个概率分布是在抽样前就有的,是有关 θ 先验信息的概率陈述。这个概率分布被称为先验分布。因为任一未知量都有不确定性,而在表述不确定性程度时,概率和概率分布是最好的语言。例如工厂产品的不合格率 θ 是未知量,且每天都会有一些变化,把它看成一个随机变量是合适的,用一个概率分布去描述它也是恰当的。

一个先验分布的例子:学生估计一位新教师的年龄。依据学生们的生活经历,在看了新教师的照片后立即会有反应:“新教师的年龄在 30 岁到 50 岁之间,极有可能在 40 岁左右。”统计学家通过与学生们交流,明确这句话中“左右”可理解为±3 岁,“极有可能”可理解为 90%的把握。于是学生们对这位新教师年龄(未知量)的认识(先验信息)可综合为图 1-5.14 所示的概率分布,这也是学生们对未知量(新教师年龄)的概率表述。

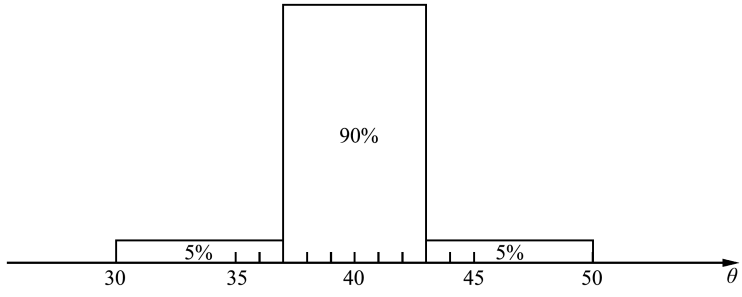


图 1-5.14　新教师年龄的先验分布

这里贝叶斯统计有两个问题与经典统计不一样。第一,未知量看作随机变量问题,该例所示的概率分布为未知量 θ 位于某个区间的概率。譬如,θ 位于 37 到 43 岁间的概率为 0.90,即

$$P(37 \leqslant \theta \leqslant 43) = 0.90$$

这种概率陈述在经典统计学中是不允许的,因为经典统计认为 θ 是常量,它要么在 37 岁到 43 岁之间(概率为 1),要么在这个区间之外(上述事件概率为零),不应有 0.9 的概率。

可在实际中类似的说法经常可以听到并使用。譬如:“明日降水概率为0.85”、“这场足球赛甲队获胜的概率只有0.6左右”,这种合理陈述的基础就是把未知量看作随机变量。

第二,主观概率问题。概率0.90不是在大量重复试验中获得的,而是学生们根据自己生活经历的积累,对该事件发生可能性所给出的估计,这样给出的概率在贝叶斯统计中是允许的,并成为主观概率。它与经典概率用频率确定的概率有相同的含义,只要它符合概率的三条公理即可。这一点经典概率学派是难以接受的,他们认为用大量重复试验的频率来确定概率才是“客观的”和符合科学的要求,而认为贝叶斯统计是“主观的”,因而(至多)只对个人做决策有用。这是当前对贝叶斯统计的主要批评。贝叶斯学派认为,引入主观概率及由此确定的先验分布,至少把统计的研究与应用范围扩大到不能大量重复的随机现象。其次,主观概率的确定不是随意的,而是要求当事人对所考察的事件有较透彻的了解和丰富的经验,甚至是这一行的专家,在此基础上确定的主观概率就能符合实际,把这样一些有用的先验信息引入统计推断中来只会有好处。当然误用主观概率与先验分布的可能性是存在的。贝叶斯学派也认为经典学派有关总体分布的选择也是经常主观的,其对答案的产生的影响要比先验分布选择所产生的影响来得大。

由此可见,贝叶斯统计方法是以坚实的概率论为基础,为统计推断提供了一套原则和灵活的方法。该体系明确告诉我们,该方法要求明确的先验知识、已有数据和假设;任何模型包括序列模型必须有概率意义,还可以用定量方法描述数据的变异和噪音,否则无法对模型进行严格的科学描述,无法确定模型是否与数据相吻合,最终也无法对模型和假设进行比较,无法对问题给出一个明确和唯一的解。

贝叶斯公式:

对于两个独立事件,它们的联合概率为

$$P(A,B)=P(A|B)\mathrm{P}(B)=P(B|A)P(A)$$

其中$P(A|B)$为条件概率,即事件B发生的情况下A事件发生的概率,反之其概率为$P(B|A)$。条件概率$P(B|A)$可以进一步写成

$$P(B|A)=\frac{P(A|B)P(B)}{P(A)}$$

事件A发生的概率$P(A)=P(A|B)P(B)+P(A|\overline{B})P(\overline{B})$。$P(A|\overline{B})P(\overline{B})$是事件B不发生情况下A事件发生的概率。

如果上式A=D(data),B=M(model),则贝叶斯公式:

$$P(M|D)=\frac{P(D|M)P(M)}{P(D)}$$

其中$P(M|D)$为后验概率(posteriori)、$P(D|M)$为似然概率(data likelihood)、$P(M)$为先验概率(priori)和$P(D)$为事实概率(evidence probability)。

在进行贝叶斯统计时,需要利用一个真实数据集作为训练数据集估计概率模型的参数,然后用于统计推断。

以一个古老赌博游戏为例,说明贝叶斯统计推断方法:

Alice和Bob在赌场玩一种古老赌博游戏。赌场在一个标有8个分隔的桌子上掷骰子,每轮比赛,第一次随机掷出骰子落入的分隔作为靶标,第二次再随机掷骰子,如果骰子落在靶标分隔内Bob得1分,落到其他位置Alice得1分。谁先得到6分谁获胜。一次游戏中,

Alice 已经以 5 比 3 领先于 Bob,问 Alice 最后取胜 Bob 的比率或概率有多大？这是一个科学推断的有趣问题,它在 13 世纪首先被提出,但答案千差万别,例如 2∶1和 3∶1。16 世纪中叶,法国数学家帕斯卡(Blaise Pascal)给出了 7∶1的答案,这被认为是概率论的起源。帕斯卡计算的依据是在 8 个分隔中,骰子随机落在其中一个分隔的概率是 1/8,即 Bob 获胜的概率是 1/8,所以他输给 Alice 的比率是 7∶1。当然你还可以用最大似然估计(maximum likelihood estimation)进行估计(即估计 $P(D|M)$):Bob 在前面 8 轮游戏中获胜 3 次(得到 3 分),获胜概率为 3/8,那么接下去他连续再赢 3 次得到累计 6 分的概率为 $(3/8)^3 = 27/512$,也就是 Alice 获胜的概率为 485/512,即两人获胜比率为 18∶1。那么贝叶斯会给出怎样的推测？贝叶斯以后验概率即 $P(M|D)$作为推断的依据。那么如何获得 $P(M|D)$？

假设 p 为 Bob 最后取胜 Alice 的概率(M),目前我们看到的结果是 Alice 已经以 5 比 3 领先于 Bob(D),根据根据贝叶斯公式

$$P(p \mid A=5, B=3) = \frac{P(A=5, B=3 \mid p)P(p)}{\int_0^1 P(A=5, B=3 \mid p)P(p)\,dp} = 1/11$$

即 Alice 最后取胜的概率为 10/11,或两者获胜的比率为 10∶1。

2. 贝叶斯统计在生物信息学领域应用

贝叶斯统计在生物信息学领域应用非常广泛,为生物信息学分析中重要方法之一。几个原因促成贝叶斯统计在生物信息学领域的重要地位:①生物信息学面对的是大量生物学序列数据,但对于这些数据产生相应机制或理论很不完善,具有高度不确定性,且大量冗余,而生物信息学家需要对这些数据进行归纳和推断,即在存在不确定性的情况下进行推理。度量不确定性正是贝叶斯统计的优势,它是进行这类推理的有效方法。贝叶斯统计使用概率论的语言来描述不确定性,并进行不确定性推理;②生物序列数据建模大多基于概率模型。已有信息或知识(先验知识或约束条件)对建模和基于模型进行统计推断具有重要作用,可以明显提高推断的准确性。贝叶斯统计推断首先利用所有背景信息和数据构建模型,然后使用概率论的语言赋予模型一个先验概率,通过概率计算,基于已有数据估计模型的后验概率或置信度,得到唯一的解,然后进行推断。贝叶斯统计方法的上述特征符合大规模生物序列数据要求;③贝叶斯统计经过上个世纪无数研究者的努力,其理论方法体系已日臻完善,而计算机技术的发展,使我们处理复杂模型的计算能力极大提高,通过机器学习方法可以对复杂模型参数等进行有效求解,包括含有几千个参数的模型和大量噪音的序列数据。由此可见,生物学领域存在大量基于观察数据进行推断的问题,而这些推断往往需要一个概率统计模型和不确定的参数或缺失的数据情况下进行("There is no shortage of problems in biology where we want to infer something from observed data, but the inference depends on uncertain parameters or missing data in a probability model")。这使贝叶斯统计在生物信息学领域的应用日益普遍。

贝叶斯统计最早在生物信息学领域的应用集中在序列联配、进化和模式识别(如基因和剪切位点预测),该方面具体应用已有大量论述(参见 Durbin 等,1998;皮埃尔-巴尔迪和索恩-布鲁纳克,2003;Mount,2004),本书仅对基因预测应用进行说明。

二、利用贝叶斯统计进行基因预测

如上所述,HMM 模型是目前基因组进行蛋白质编码基因预测的主要方法。HMM 模型

往往与贝叶斯统计关系密切,在实际预测中,HMM 模型经常利用贝叶斯统计进行统计推断,即利用后验概率进行统计推断。

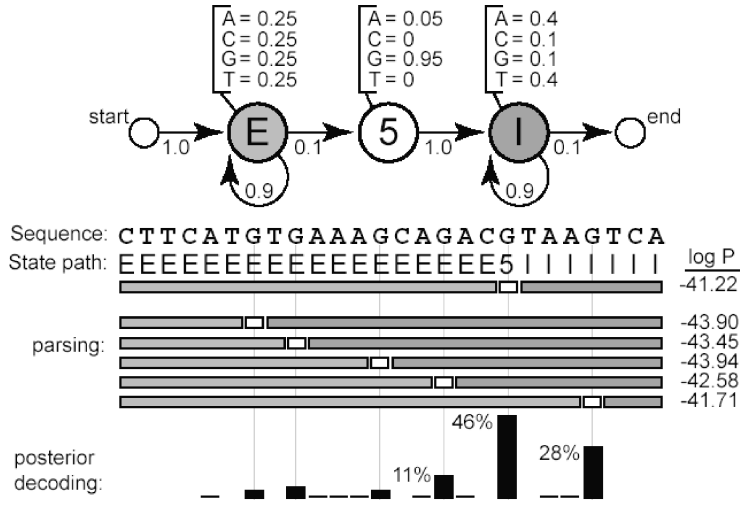


图 1-5.15 5'端外显子(E)与内含子(I)剪切位点(5)识别的隐马尔科夫模型(HMM)案例(引自 Eddy,2004)

还是利用上述 5'端外显子与内含子剪切位点识别例子。由于一些路径的发生概率很相近,如图中两个分别 log $P=-41.22$ 和 -41.71 相差不大,如何确定剪切位点到底发生在哪里呢?

此时我们需要结合后验概率来解决此问题。在本模型中,已知只有 A 或 G 位点可能发生可变剪切,我们知道一共只有 14 种可能性,根据前面的方法可以计算获得 14 种发生剪切路径的概率(表 1-5.2)。然后,用每一种剪切发生概率除以所有路径可能性概率之和,就是每一种剪切发生的可信度(在此处即后验解码概率),例如,此处所有可能性概率之和为:

$$P=\sum_{i=1}^{14}P_i=2.72E-18$$

P_i 为第 i 种方案发生的概率

所以第 i 种方案的后验解码概率

$$PDP_i=\frac{P_i}{\sum_{i=1}^{14}P_i}=\frac{1.25E-18}{2.72E-18}=46.2\%$$

由此我们得到所有 14 方案的后验概率(表 1-5.2)。由此我们可以推断,在所有 14 个可能剪切位点中,最大可能(46.2%)发生在该序列第 19 位 G 上(即第 5 个 G 上,也即第 1 个路径方案)。

表 1-5.2 外显子与内含子剪切位点识别隐马尔科夫模型(图 1-5.9)中各个可能路径(剪切方式)的联合概率和后验概率。

可能路径方案/剪切方式	剪切位点(位置/碱基)	发生概率 P	(log P)	后验概率(%)
1	19G	1.25E-18	-41.22	46.20
2	23G	7.66E-19	-41.71	28.20

续表

可能路径方案/剪切方式	剪切位点（位置/碱基）	发生概率 P	（log P）	后验概率（%）
3	16G	3.21E-19	-42.58	11.83
4	9G	1.35E-19	-43.45	4.96
5	7G	8.62E-20	-43.90	3.17
6	13G	8.22E-20	-43.94	3.03
7	5A	2.9E-21	-47.29	0.11
8	10A	4.43E-21	-46.87	0.16
9	11A	2.77E-21	-47.34	0.10
10	12A	1.73E-21	-47.81	0.06
11	15A	6.76E-21	-46.44	0.25
12	17A	1.06E-20	-46.00	0.39
13	21A	2.58E-20	-45.10	0.95
14	22A	1.61E-20	-45.57	0.59
合计		2.72E-18		100

第四节　基因功能注释

在获得基因结构注释信息后，我们希望能够进一步获得基因的功能信息。基因功能注释主要包括预测基因中的结构域、蛋白质功能和所在的生物学通路等。目前普遍采用序列相似性比对的方法对基因功能进行注释。

一、利用序列和结构域数据库进行注释

以下为基因功能注释中常用的几个数据库：

1. 利用 NR、Uniprot/SwissProt 数据库进行注释

当需要功能注释的序列数目不是很多时，可直接在 NCBI 网页上（http://blast.ncbi.nlm.nih.gov/Blast.cgi）选择需要进行比对的数据库（图 1-5.16），直接进行 BLAST 搜索，获与 NR 数据库（non-redundant protein sequence database）记录的最佳匹配，根据匹配上的已知功能基因，推断未知基因的功能。虽然 NCBI 在线注释可以一次提交多条序列，但是每次速度还是相对较慢，获得的注释结果需要手动整理。在线方式的优势在于结果中还会出现多个功能数据库的链接（Pfam、Interpro 等），适合对于少数几个特别感兴趣的序列，进行详细的功能了解。

若有大量的基因需进行功能注释时，如需注释某一物种上万条基因序列时，通常会采用本地化注释的方法，即利用 NCBI 提供的本地版本的 BLAST 程序（ftp://ftp.ncbi.nih.gov/blast/executables/blast+/LATEST/）和从 NCBI 上下载的 NR、SWISSPROT 数据库（ftp://ftp.ncbi.nih.gov/blast/db/）做比对。具体的本地 BLAST 操作方法可参考 NCBI 官网提供的使用

文档(http://www.ncbi.nlm.nih.gov/books/NBK279690/pdf/Bookshelf_NBK279690.pdf)。在利用 BLAST 进行功能注释时一般设定的 E-value 标准为 1e-7 或 1e-5,若有很多条记录满足该条件时,通常会选取匹配最好的记录(Best hit)作为该序列的功能注释结果。

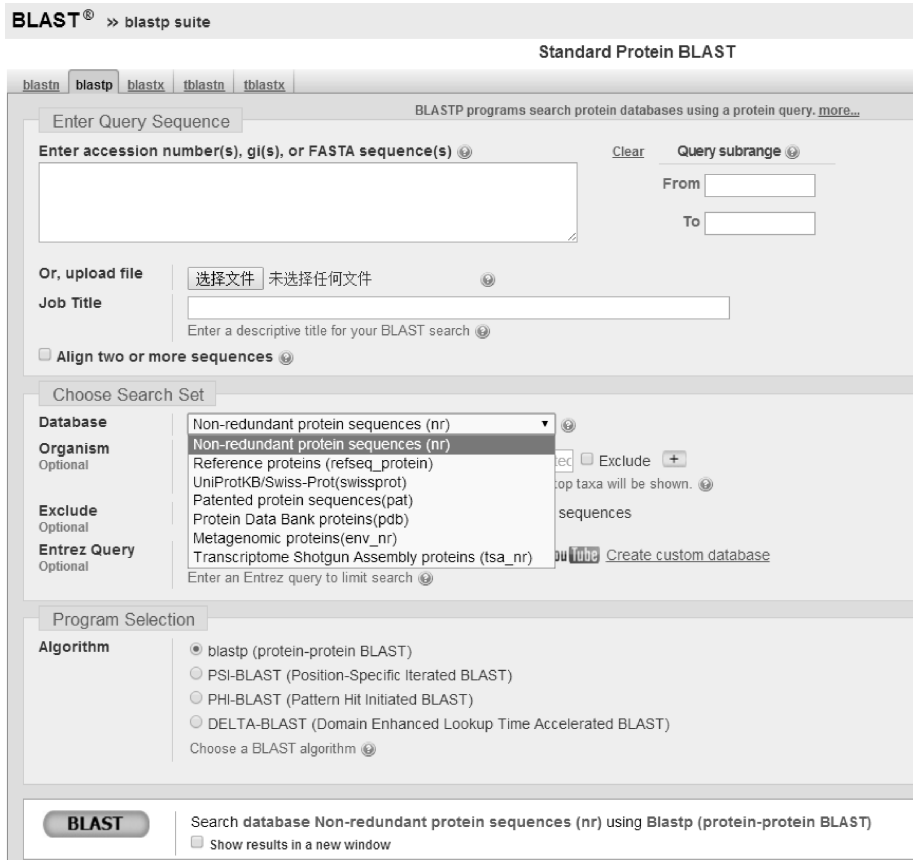


图 1-5.16 在线 NCBI 主页 BLAST 界面,可用于少量基因功能注释
可在下拉框中选择用于功能注释的数据库

2. 利用 Interpro 功能域数据库进行注释

使用 Interpro 数据库,可预测蛋白质功能域或重要位点。该数据库整合了 PROSITE、PFAM、PRINTS、ProDom、SMART、TIGRFAMs 等功能域数据库和 PIRSF、SUPERFAMILY、CATH-Genes3D 等其他不同类型数据库。根据需要可以选择注释数据库,获得相应的功能注释结果。在线 Interproscan(http://www.ebi.ac.uk/interpro/)目前一次仅支持单条蛋白序列的查询,结果的输出格式为 HTML 或者 GFF3。Interproscan 有本地化的版本(http://www.ebi.ac.uk/interpro/interproscan.html),在计算机资源充足的情况下,可利用多线程运行加快注释速度。此外,本地化版本还可输入 DNA 序列,获得 DNA 水平上的序列位点注释信息。输出格式可选择 GFF3、tsv 等便于用户查看操作。除了在线和本地化的 Interproscan 版本,也可使用 EBI 提供的 Perl、Puby 或 Python(http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan5_rest)程序进行远程比对,程序将序列递交到远程官网进行注释。这种方法较为方便,可在单机 WINDOWS 系统的 DOS 下运行“perl interproscan_lwp.pl --email <your@

email> [options] seqfile", 结果会返回到本地当前路径。

二、利用功能分类和代谢途径信息进行注释

1. 利用 GO 定义基因功能

GO 将功能分为三大类别,即细胞组分(cellular component)、分子功能(molecular function)、生物学过程(biological process)。获得 GO 注释最简单的方法是利用已做好 interproscan 的注释,直接从该结果中提取相关基因的 GO 注释信息。GO 注释信息统计和展示可用在线工具 WEGO(http://wego.genomics.org.cn/cgi-bin/wego/index.pl),后续的 GO 富集等分析可利用 AgriGO(http://bioinfo.cau.edu.cn/agriGO/)、GOEAST(http://omicslab.genetics.ac.cn/GOEAST/tutorial.php)等在线分析平台获得。

2. 利用 KEGG 等数据库生物学代谢通路信息

通常使用 KAAS(http://www.genome.jp/tools/kaas/)完成 KEGG 注释。通过该网站注释获得的结果包括对应 KO(KEGG Orthology)代号、KEGG 的代谢通路以及各个代谢通路对应的图谱等。KAAS 主要分为两种形式,即双向最好匹配(BBH)以及单向最好匹配(SBH),前者适用于全基因组基因序列的注释,后者适用于对个别基因进行注释的情况。

随着生物信息软件的发展与优化,出现了很多集成多种功能的基因功能注释方法。Blast2go(https://www.blast2go.com/)就是一个目前较为流行的,可在多操作系统下运行且具有综合用途的基因功能注释软件。其主要功能如下:可将序列比对到 NCBI 的 NR 数据库获得 NR 注释;通过 Blast2go 的数据库,将 NR 注释的结果转换为 GO 注释;进行 GO 分类、富集分析,以及整合 GO 概念关系图的制作;可获得 Enzyme Code 注释和 KEGG 通路图的制作等。

目前基因功能注释面临的问题很明显,注释工作是建立在相似性比对的基础上,因而非常依赖于外部数据,对某些研究较少的物种,其基因注释限制明显,无法得到功能信息。另外,序列相似并不表示生物学功能相似,需要考虑引入序列比对之外的方法,进一步完善基因功能注释工作。

第五节　基因序列构成分析

一、碱基构成与分布

1. 碱基构成

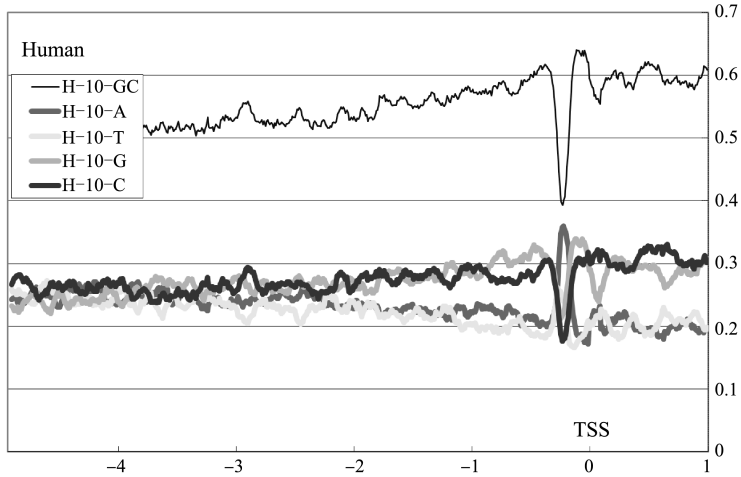
DNA 序列一个显而易见的特征是四种类型碱基的分布。几乎所有的研究都证明,DNA 序列碱基是以不同频率分布的。首先在基因组水平上,一个基因组的四种碱基构成会不一样,例如 9 个物种基因组 DNA 序列的碱基构成存在差异(表 1-5.3)。

表 1-5.3 九个物种基因组完整 DNA 序列的碱基组成*

基因组	名 称	碱基频率				总计(nt)
		A	C	G	T	
噬菌体						
λ	LAMCG	0.25	0.24	0.25	0.26	48 502
T7	PT7	0.27	0.23	0.24	0.26	39 936
ØX174	PX1CG	0.24	0.22	0.31	0.23	5 386
病毒						
花椰菜镶病毒	MCACGDH	0.37	0.21	0.23	0.19	8 016
人类乳头多瘤空泡病毒 BK	PVBMM	0.30	0.20	0.30	0.20	4 936
肝炎 B	HPBAYW	0.28	0.22	0.23	0.27	3 182
线粒体						
人类	HUMMT	0.31	0.31	0.25	0.13	16 569
牛	BOVMT	0.33	0.26	0.27	0.14	16 338
鼠	MUSMT	0.35	0.24	0.29	0.12	16 295

* 取自 GenBank 数据库

在基因水平上,碱基的构成表现出明显碱基分布特征。我们收集了 GenBank 数据库中几百条已知基因 DNA 序列, 将它们按照转录起始位点(TSS)对齐,然后以 10 个碱基长度窗口从左到右逐碱基滑动,计算每个窗口中四种碱基的频率,然后画成频率分布图(图 1-5.17)。从图中可见,基因碱基分布的一个总体趋势,即基因区域的 G/C 碱基比例会上升,A/T 比例下降,G+C 碱基比例超过 A+T;基因间碱基构成则正好相反。因此,在 TSS 区域附近,我们可以看到这四种碱基比例构成此消彼长的有趣现象。



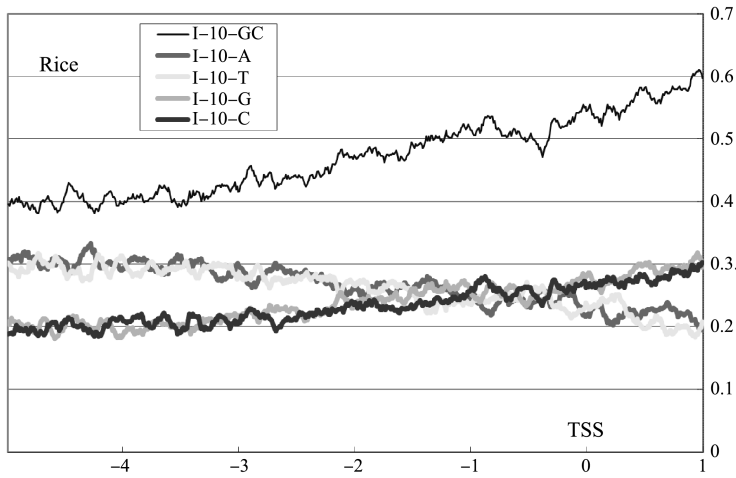


图 1-5.17　人类和水稻基因转录起始位点(TSS)附近碱基分布变化

随机挑选 100 条基因序列以转录起始位点对齐,按照 10nt 窗口长度逐碱基计算四种碱基频率。

2. 碱基相邻频率

分析 DNA 序列的主要困难之一是其碱基相邻的频率不是独立的。一个清晰的证据是碱基相邻的频率一般不等于单个碱基频率的乘积。如果 Pu 是序列中碱基 u 的频率,且 Puv 为两个相邻碱基 u 和 v 的频率,则

$$p_{uv} \neq p_u p_v$$

我们研究了水稻和人类基因组 DNA 序列两碱基相邻的频率(表 1-5.4)。数据来自这两个物种目前注释出来的所有基因的 DNA 序列,总长各为 168,717,208 和 1,506,657,427 个碱基。表中的比值为 16 种二个碱基相邻的频率除以相应的单个碱基频率的乘积。

表 1-5.4　人类和水稻中两碱基的相邻频率

相邻碱基对	观测频率/期望频率*	
	人类	水稻
CC	1.27	1.05
GG	1.22	1.03
CA	1.20	1.11
TG	1.19	1.11
AG	1.18	0.99
CT	1.15	0.99
TT	1.13	1.13
AA	1.13	1.11
GC	1.02	1.11
GA	0.99	1.05
TC	0.96	1.00
AT	0.88	1.02
GT	0.84	0.84
AC	0.83	0.86
TA	0.75	0.77
CG	0.26	0.83

* 期望频率为相应两个单个碱基频率的乘积

作为一个单个基因的例子,我们以鸡血红蛋白 β 链 mRNA 编码区的 438 个碱基为例(GenBank 记录号 J00860)说明相邻两个和三个碱基情况。表 1-5.5 列出了 4 种碱基和 16 种两个相邻碱基的数目。将该表看作 4×4 的表,计算行列独立性的卡方统计量,得到 $\chi^2 = 59.3$($\chi^2_{0.05,9} = 16.92$),表明行(第一碱基)列(第二碱基)之间存在明显的关联。

表 1-5.5 鸡 β 球蛋白基因序列(记录号 J00860)的相邻碱基分布

		第二碱基			
		A	C	G	T
第一碱基	A	23	26	23	15
	C	37	51	14	41
	G	25	38	36	19
	T	2	29	41	14
总计	/	87	144	117	89

我们进一步看其三碱基相邻情况。在编码区,存在某种约束来限制 DNA 序列编码氨基酸。在密码子水平上,这一约束与碱基相邻频率有关。表 1-5.6 列出了该序列中各遗传密码子的数量。尽管数目很小,难以做出有力的统计结论,但编码同一氨基酸的不同密码子(同义密码子)不是等同存在的,例如偏向于使用 GCC\CUG 等密码子。这种密码子偏倚必定与两碱基相邻频率水平有关。表 1-5.6 还清楚地表明,由于密码子第 3 位置上碱基的改变常常不会改变氨基酸的类型,因而对第 3 位置上碱基的约束要比第 2 位碱基小得多。从一个物种水平上看,密码子使用的偏好性非常明显,动物与植物以及微生物之间存在明显不同;同一类型,如植物的不同物种之间也有所不同。例如两个模式植物单子叶植物水稻与双子叶植物拟南芥,如果我们统计一下它们编码基因的碱基构成,可见明显差异(图 1-5.18)。与拟南芥相比,可以明显观察到水稻基因密码子对 G 和 C 碱基的偏好性。

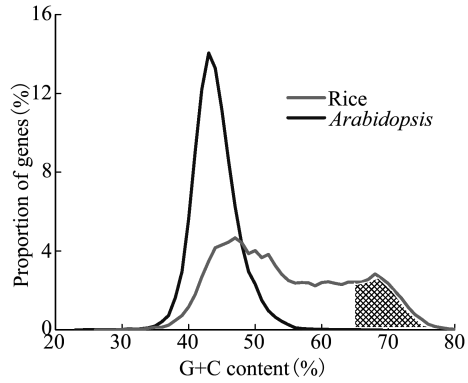


图 1-5.18 水稻和拟南芥不同 GC 含量蛋白质编码基因分布图

图中可以明显观察到水稻基因编码密码子对 G 和 C 碱基的偏好性(引自 Guo 等, 2007)

表 1-5.6 鸡 β 球蛋白基因(记录号 J00860)64 种可能的碱基三联体密码子及相应的氨基酸数

UUU Phe 3	UCU Ser 0	UAU Tyr 0	UGU Cys 2
UUC Phe 5	UCC Ser 5	UAC Tyr 2	UGC Cys 1
UUA Leu 0	UCA Ser 0	UAA Stop 0	UGA Stop 0
UUG Leu 0	UCG Ser 0	UAG Stop 0	UGG Trp 4
CUU Leu 1	CCU Pro 1	CAU His 3	CGU Arg 0
CUC Leu 6	CCC Pro 4	CAC His 4	CGC Arg 3
CUA Leu 0	CCA Pro 0	CAA Gln 1	CGA Arg 0

续表

CUG Leu 11	CCG Pro 0	CAG Gln 0	CGG Arg 0
AUU Ile 1	ACU Thr 3	AAU Asn 1	AGU Sre 0
AUC Ile 6	ACC Thr 4	AAC Asn 6	AGC Ser 2
AUA Ile 0	ACA Thr 0	AAA Lys 1	AGA Arg 0
AUG Met 1	ACG Thr 0	AAG Lys 9	AGG Arg 3
GUU Val 0	GCU Ala 4	GAU Asp 1	GGU Gly 1
GUC Val 5	GCC Ala 11	GAC Asp 5	GGC Gly 4
GUA Val 0	GCA Ala 0	GAA Glu 4	GGA Gly o
GUG Val 7	GCG Ala 1	GAG Glu 3	GGG Gly 3

相邻碱基之间的关联将导致更远碱基之间的关联,这些关联延伸距离的估计可以通过马尔科夫链方法得到(Javare 和 Giddings,1989)。如上所述,在不援引任何生物学机制的情况下,第 k 阶马尔科夫链假定在序列中某一位置上碱基的存在,只取决于前面 k 个位置上的碱基。也可以通过似然法进行类似估计。相关内容可参见第本章第 2 节。

二、DNA 行走与 Z 曲线

如果我们把 DNA 序列看作两个符号组成的符号序列,从一条 DNA 序列的首字母开始,每看到一个嘌呤(字母 A 或 G)就从横坐标的原点向左走一步,每看到一个嘧啶(字母 C 或 T)向右走一步。这就是所谓的“DNA 行走”。针对所选的序列,“DNA 行走”不是无规则行走,会反复经过原点。研究表明,编码序列比非编码序列更为“随机”。上述“DNA 行走”是在一维上实现,同时也可以在二和三维上实现,形成所谓行走曲线或路径(例举见图 1-5.19)。

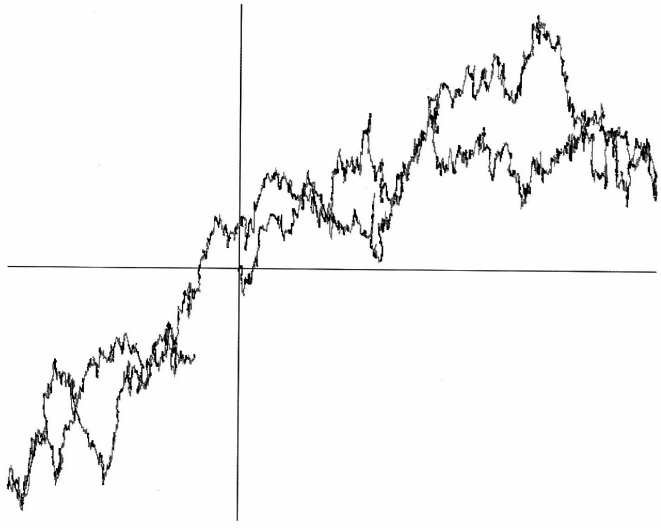


图 1-5.19　大肠杆菌 K12 菌株基因组序列的二维 DNA 行走曲线(郝柏林,2015)

生物信息学研究的一个重要原则,是所提出的方法及其结果要有生物学意义和具体用途。DNA 序列 Z 曲线是一个成功的例子。DNA 序列实际上是一种用 4 种字母表达的“语言”,只是其“词法”和“语法”规则目前还没有完全搞清楚。人类的语言有文字、声音两种基本表现形式,此外还有手语、旗语甚至图画语等特殊表达形式。同样,DNA 序列作为一种语言,其表达形式也不是唯一的。传统上,DNA 序列是用 4 种字母符号表达的一维序列。这是一种抽象形式,适合于存储、印刷和代数算法的处理,包括比较、排列和查找特殊序列等。我国学者张春霆院士开展了 DNA 序列三维空间曲线表示形式,即 DNA 序列几何表示形式的研究。几何形式虽然与符号形式完全等价,但显示了 DNA 序列的新特征。两种形式各有其特点,相互补充。这一新方法,为解读 DNA 序列信息提供了崭新的手段。

他们的研究始于对 4 种碱基对称性的观察,提出了用正面体表示碱基对称性。他们利用这种形式来表示任意长度的 DNA 序列。现将这种序列表示方法简述如下:

考察一个长为 L 的单 DNA 序列,方向(5'→3'或 3'→5')不限。从第一个碱基开始,依次考察此序列,每次只考察一个碱基。当考察到第 n 个碱基时($n=1,2,\cdots,L$),数一下从 1 到 n 这个子序列中四种碱基各自出现的次数。设 4 种碱基 A、C、G、T 出现的次数分别以 A_n、C_n、G_n、T_n 表示之,这里下标“n”是表明这些整数是从 1 到 n 这个子序列中数出来的。显然,它们都是正整数。根据正四面体的对称性可以证明,在正面体内存在唯一的一个点 P_n 与这四个正整数对应。点 P_n 构成了四个正整数的一一对应映射。点 P_n 坐标可用四正整数表达:

$$x_n=2(A_n+G_n)-n,$$
$$y_n=2(A_n+G_n)-n,$$
$$z_n=2(A_n+T_n)-n,$$
$$x_n,y_n,z_n\in[-n,n],n=1,2,\cdots,L,$$

其中 x_n,y_n 和 z_n 为点 P_n 的三个坐标分量。当 n 从 1 跑到 L 时,我们依次得到 P_1,P_2,P_3,$\cdots$,P_L 共 L 个点。将相邻两点用适当的曲线连接所得到的整条曲线,就称为表示 DNA 序列的 Z 曲线。可以证明,Z 曲线与所表示的 DNA 序列是一一对应的,即给定一 DNA 序列,存在唯一的一条 Z 曲线与之对应;反之,给定一条 Z 曲线,可找到唯一的一个 DNA 序列与之对应。换言之,Z 曲线包含了 DNA 序列的全部信息。Z 曲线是与符号 DNA 序列等价的另一种表示形式,一种几何形式。可以通过 Z 曲线对 DNA 序列进行研究。

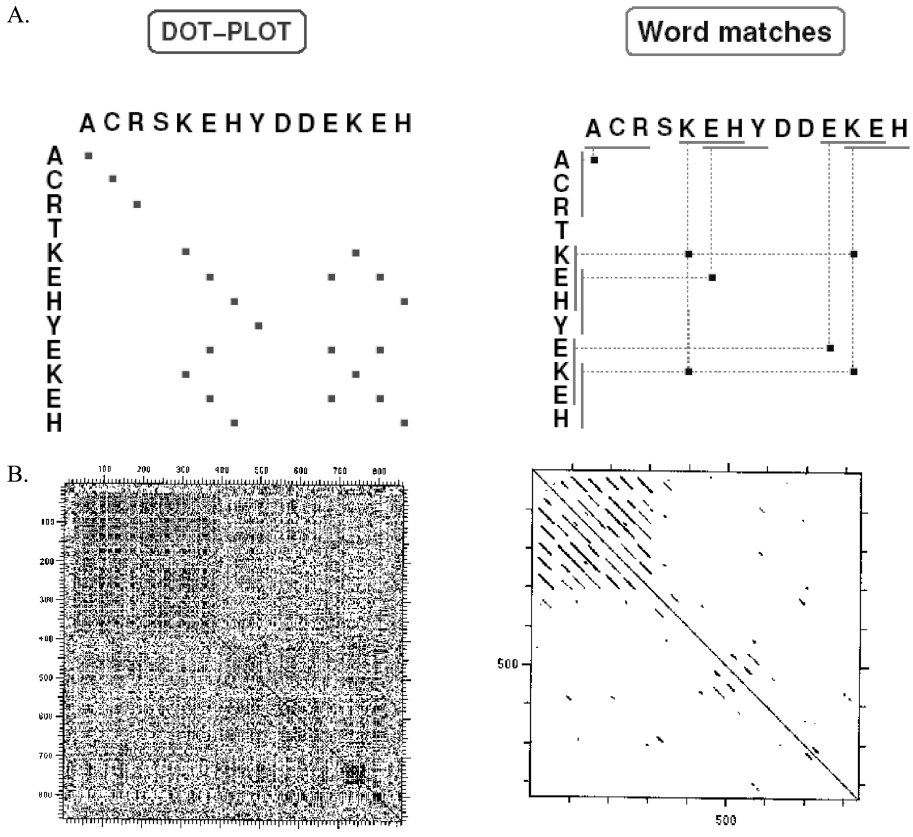
Z 曲线的三个分量具有明确的生物学意义:x_n 表示嘌呤/嘧啶碱基沿序列的分布。当从 1 到 n 的这个子序列中嘌呤碱基多于嘧啶碱基时,$x_n>0$,否则,$x_n<0$,当两者相等时 $x_n=0$。同样,y_n 表示氨基/酮基碱基沿序列的分布。当在子序列中氨基碱基多于酮基碱基时,$y_n>0$,否则,$y_n<0$,当两者相等时 $y_n=0$。z_n 表示强/弱氢键碱基沿序列的分布。当弱氢键碱基多于强氢键碱基时,$z_n>0$,否则 $z_n<0$,当两者相等时,$z_n=0$。这三种分布是相互独立的,表现在以下事实上:任何一种分布不能由其它两种分布的线性叠加表示出来。给定的 DNA 序列唯一地决定了这三种分布;三种分布唯一地描述了 DNA 序列。对 DNA 序列的研究就是通过对这三种分布的研究来进行。从方法学的角度来看,这是 DNA 序列的一种几何学研究途径。

三、同向重复序列分析

除了序列碱基关联特征外,我们常对重复序列,如同向重复序列(direct repeats)之类的

问题感兴趣。一个简单同向重复分析方式是点阵(dot matrix)方法,即把感兴趣的一条或两条序列分别放在两侧(如图 1-5.20A),比较每个位点的碱基/氨基酸,相同的碱基/氨基酸在图中打一个点,这样重复的区域就会出现连续的点。为了减少随机匹配导致的背景噪音,可以用一定长度的碱基/氨基酸字符串(所谓字)在图中打点。以人 LDL 受体基因序列为例,利用 DNA Strider 软件可以对其自身进行点阵画图(Mount, 2004)(图 1-5.20B)。如果逐个碱基比较,相同的打一个点,这样点阵图背景噪音很大(即随机造成的匹配很多);如果以字(如 23nt 长度)为单位,明显去除了背景噪音。可见该基因在 23nt 字长情况下,基因前部有 6 个重复模序(motif)。

基于序列相似性,点阵方法可以在更大尺度上(如染色体水平)进行同一物种或不同物种间基因组共线性区块的可视化分析。以不同染色体上注释基因之间相似性比较为基础,达到临界值的同源基因对打一个点,这样就可以比较不同染色体之间的相似性和进化关系了。一个具体例子(图 1-5.20C)可参加我们在水稻基因组上的分析结果(Zhang 等,2005),这里不再累述。



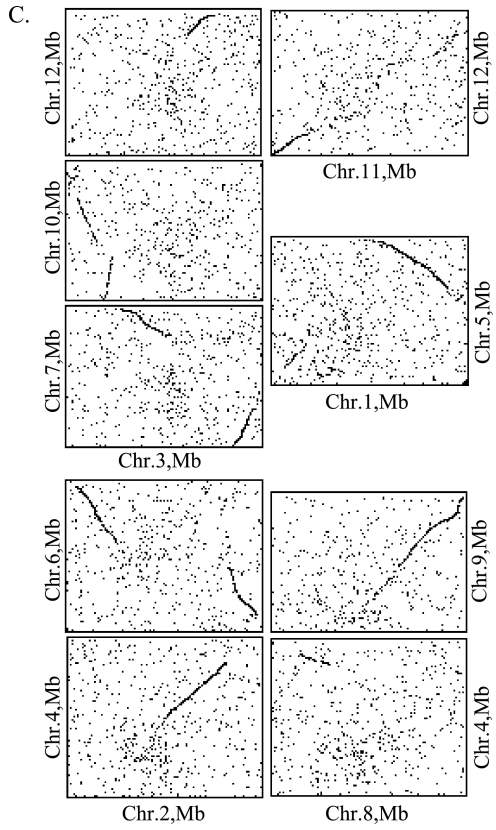


图 1-5.20 利用点阵(dot matrix)方法进行序列保守性分析列举

A.两条蛋白质序列不同字长(单个和三个氨基酸)的分析效果;B.人类 LDL 受体基因不同字长(单碱基和 23nt)的自身比较图;C.水稻 12 条染色体间比较图,图中可见基因组共线性区域(引自 Zhang 等,2005)。

点阵方法是一个很好的可视化分析方法,但在实际分析中,往往需要一个有效算法进行计算识别。Karlin 等(1983)给出了一个有效算法。该法采用特定的几组碱基字母组成的不同字符串或称为字码(word),只需要对整个序列搜索一次。给一碱基赋以一定值 α,例如 A、C、G、T 的值为 0、1、2、3。由 k 个字母组成的每一种不同字码具有如下字码值:

$$1+\sum_{i=1}^{k}\alpha_i 4^{k-i}$$

上述字码值的取值范围为 1 到 4^k。例如,5 字码 TGACC 的值为 $1+3\times4^4+2\times4^3+0\times4^2+1\times4^1+1\times4^0=459$。可先从低 k 值的字码开始搜索,记录序列中每一个位置 k 字码的字码值。只有在发现 k 字码长度重复的那些位置,考虑进行长度大于 k 的字码搜索。

以“TGGAAATAAAACGTAAGTAG”为例,我们可以计算该序列中所有两个碱基字码($k=2$)的初始位置和字码值(表 1-5.7)。对于长度大于 2 的同向重复或字符串的搜索可只限于两字码重复的初始位置。在本例中只有 4 个字码值存在重复出现。例如,在位置 4,5,8,9,10 和 15 均出现相同字码值(1),即至少均存在两个碱基的重复序列。进一步看三字码值及位置(表 1-5.7),发现字码值 1,45 和 49 在多个位点出现,说明在这些位置上有三碱基序列重复,例如字码值 1 在 4,8,9 位置上分别出现,其为“AAA”三碱基重复。继续以四字码值进一步搜索,未能发现重复出现的字码值,说明没有更长的重复序列。

表 1-5.7　序列“TGGAAATAAAACGTAAGTAAGTAG”的 2 和 3 字码值和位置(Karlin,1983)

2 字码值	碱基位置	3 字码值	碱基位置
1	4,5,8,9,10,15	1	4,8,9
2	11	2	10
3	16,19	3	15
4	6	4	5
5	–	45	13,17
6	–	49	7,14
7	12	51	18
8	–		
9	3		
10	–		
11	2		
12	13,17		
13	7,14,18		
14	–		
15	1		
16	1		

同样以鸡血红蛋白 β 链的 mRNA 编码区的 438 个碱基为例(J00860),鸡 β 球蛋白 DNA 序列进行同向重复序列搜索,发现不少位点出现 8-9 个碱基同向重复,最长的重复序列为 10 个碱基(表 1-5.8)。

表 1-5.8　鸡 β 球蛋白 DNA 序列中重复序列检测结果

字码长度(nt)	重复序列	起始位置
8	GCCCTGGC	79 418
	GCCAGGCT	85 377
	CCAGGCTG	86 378
	CAGGCTGC	87 379
	TCCTTTGG	130 208
	CCTTTGGG	131 209
	TGGTCCGC	176 398
	GGTCCGCG	177 399
9	GCCAGGCTG	85 377
	CCAGGCTGC	86 378
	TCCTTTGGG	130 208
	TGGTCCGCG	176 398
10	GCCAGGCTGC	85 377

Karlin 等(1983)还提出了特定序列内存在的最长同向重复序列的期望值及其统计显著性评价方法。对于长度为 n 的序列中,其核苷酸出现的位置为独立的假定下(相当于 0 阶马尔科夫链),其最长同向重复序列的期望长度和方差为:

$$\mu_L = \frac{0.635\ 9 + 2\ln n + \ln(1-p)}{\ln(1/p)} - 1$$

$$\sigma_L^2 = \frac{1.645}{(\ln P)^2}$$

其中,P 为序列中碱基频率的平方和:

$$P = \sum_{i=1}^{4} P_i^2$$

用尽可能接近最大长度期望均值的字码长度(μ_L)来开始同向重复序列的搜索,可以节省计算量。对于鸡 β 蛋白基因序列,A、C、G、T 四个碱基的次数分别为 87、144、118 和 89,因此 P =0.261 4,最长重复序列的期望长度为 8.13 且具有期望方差 0.9138。假定同向重复序列的长度呈正态分布。根据 95%的正态分布概率,理论上可以预期最长同向重复序列不超过 10。

四、蛋白质序列跨膜等特征分析

本节主要介绍蛋白质序列跨膜、抗体和疏水特性(所谓一级结构)分析内容,它不同于蛋白质序列二级结构(如 alpha, beta 等)和三级结构分析(详见第 1-7 章)。

1. 跨膜结构域预测

生物膜主要由脂双层和膜蛋白组成。膜蛋白决定了生物膜的功能特性。因此,不同类型的生物膜,其蛋白构成比例有很大的差别。另一个角度来说,生物体内的蛋白质中,20%~30%都属于膜蛋白,因此膜蛋白在细胞的功能中占据重要地位是显而易见的。膜蛋白主要包括外在膜蛋白或称外周膜蛋白(peripheral membrane protein)和内在膜蛋白或称整合膜蛋白(integral membrane protein)。在整合膜蛋白中,与细胞膜的结合也有几种不同的方式,其中主要方式是肽段直接跨过细胞膜,即跨膜蛋白(transmembrane protein),另外,还有通过脂肪基与细胞膜发生共价结合,蛋白本身可以在细胞内侧,也可以在细胞膜外侧。跨膜蛋白一般以疏水的 alpha- helix 与膜脂肪发生非共价结合,在细胞中常执行信号传导或转运通道功能,是研究最为集中的一类蛋白质。跨膜(TM)蛋白跨过整个脂膜,通常被分为两类 α-helical TM (AHTM) 和 TM β-barrel (TMB)蛋白。AHTM 定位在细菌细胞膜的内膜和真核生物的细胞膜上。它们的跨膜区域有极性的环链接而成的 α 螺旋。对 TMB 蛋白的了解还不多,它们的跨膜域为反向平行的桶装 β 链通道。

通过实验方法(X-ray 和 NMR 等)解析的 TM 蛋白三维结构非常有限。因此,人们开发了很多的方法用来预测蛋白质的跨膜结构域。这些方法中的大部分都只根据序列来识别跨膜结构。已有多种预测跨膜螺旋的方法,最简单的是直接观察以氨基酸为单位的疏水性氨基酸残基的分布区域,同时还有多种更加复杂的、精确的算法能够预测跨膜螺旋的具体位置和它们的膜向性。这些技术主要是基于对已知跨膜螺旋的研究,例如 TMHMM(http://www.cbs.dtu.dk/services/TMHMM/)。TMHMM 是一个基于隐马尔科夫模型(HMM)预测跨膜螺旋的程序,它综合了跨膜区疏水性、电荷偏倚、螺旋长度和膜蛋白拓扑学限制等性质,可对跨膜

区及膜内外区进行整体预测。

2. 亲疏水性

疏水性预测的方法依赖于疏水性的衡量尺度,每个氨基酸根据其一系列的物理特性(例如溶解性、跨越水-汽相时产生的自由能等),被赋予一个数值以代表其疏水性。具有更高正值的氨基酸具有更大的疏水性;反正则更加亲水。基于此,沿蛋白质序列的疏水性移动平均值或者称为亲/疏水性索引可以被计算出来。20 种氨基酸带有不同的侧链基团,亲疏水性明显不同(表 1-5.9),其中疏水性最强为异亮氨酸(亲水指数为 4.5),其次为缬氨酸(亲水指数为 4.2),亲水性最强的两种氨基酸分别是精氨酸(亲水指数为-4.5)和赖氨酸(亲水指数为-3.9)。使蛋白质整体或者不同的区域表现出不同的亲疏水性,从而影响蛋白质对应的结构功能。因此,根据基因序列翻译而成的氨基酸序列可以推测出其蛋白质的亲疏水性质。

表 1-5.9　二十种氨基酸 R 基亲水性(—)或疏水性(+)趋势

氨基酸	缩写符号	趋势值
甘氨酸	Gly	-0.44
丙氨酸	Ala	1.88
脯氨酸	Pro	1.66
缬氨酸	Val	4.22
亮氨酸	Leu	3.88
异亮氨酸	Ile	4.55
甲硫氨酸	Met	1.99
苯丙氨酸	Phe	2.88
色氨酸	Trp	-0.99
丝氨酸	Ser	-0.88
苏氨酸	Thr	-0.77
半胱氨酸	Cys	2.55
天冬氨酸	Asn	-3.55
谷氨酰胺	Gln	-3.55
酪氨酸	Tyr	-1.33
赖氨酸	Lys	-3.99
组氨酸	His	-3.22
精氨酸	Arg	-4.55
天冬氨酸	Asp	-3.55
谷氨酸	Glu	-3.55

3. 抗原

免疫细胞通常难以借助其表面受体识别整个蛋白质抗原分子,而是识别抗原肽分子上的一个特定部分,即表位(epitope),又称为抗原决定簇(antigenic determinant)。因为表位代表了抗原分子上的一个免疫活性区,负责与抗体分子或免疫细胞表面的抗原受体结合。一个蛋白质抗原,它不但含有 B 细胞、Th 细胞、CTL 细胞、NK 细胞等与免疫识别密切相关的表位结构,同时还含有一些对于保护性免疫不利的结构,如毒性或抑制性表位、优势非中和性

表位、病理及自身抗原交叉反应性表位等。人们还认识到同一分子的众多表位中,各表位对抗原性的贡献不同,有优势表位和非优势表位之分。此外,表位间的相对位置、构象特征、表位侧翼顺序等也与表位功能的表达密切相关。

蛋白质的二级结构是蛋白表位预测的重要参数之一,β转角为凸出结构,多出现在蛋白质抗原表面,有利于与抗体结合,较可能成为抗原表位。而α螺旋和β折叠结构规则不易变形,较难结合抗体,一般不作为抗原表位。蛋白质的二级结构与其他参数结合,进行综合分析得到可能的抗原表位。蛋白质的亲水性、蛋白表面可及性、柔韧性和抗原指数是蛋白质表位预测的重要参数,表面可及性指数需大于1,可及性高的区域会暴露于分子表面,反之可及性较低的区域则埋藏于分子内部的区域;柔韧性高的区域则具有一定的可塑性,形成表位的可能性较大,容易与抗体进行空间结合。亲水性和抗原指数都>0时,可能为蛋白质抗原表位区域。

以下我们以人CRP蛋白(C-reactive protein, UniProt记录号:P02741)为例进行其抗原表位预测。使用DNAstar(www.dnastar.com)和SOPMA(http://metadatabase.org/wiki/SOPMA)软件工具,利用相应算法(括号内)进行β-转角(Chou-Fasman)、抗原指数(Jameson-Wolf)、亲水性(Kyte-Doolittle)、表面可及性(Emini)、柔韧性(Karplus- Schulz)等预测。获得各个指标如图1-5.21结果。

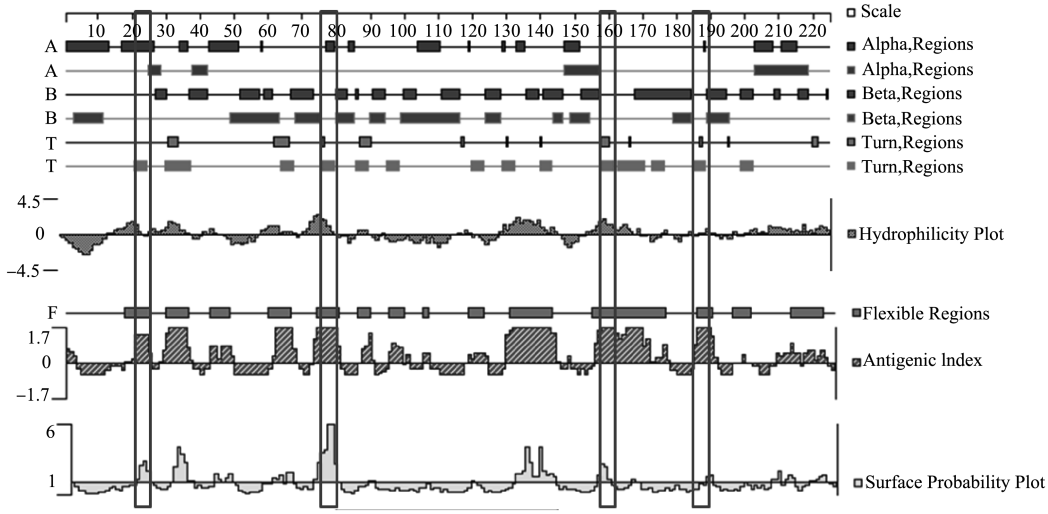


图1-5.21 人类CRP蛋白(UniProt记录号:P02741)抗原表位相关特征预测结果

根据上述各个指标结果,我们选定4个位点为候选抗原表位位点(图1-5.21方框标出)。4个位点具体位点信息如表1-5.10所示。

起始位点	终止位点	表位蛋白质序列
20	24	TDMSR
75	79	KRQDN
157	161	QDSFG
184	188	LSPDE

习　题

1. 生物序列碱基构成是否随机？请举例说明

2. 简述基因预测(注释)的一般方法和步骤

3. 简述马尔科夫模型和 HMM 模型

4. 请构建 DNA 序列的马尔科夫模型和 HMM 模型

5. 请构建多序列联配的马尔科夫模型和 HMM 模型

6. 贝叶斯统计与经典统计的不同点有哪些？为什么该统计方法在生物信息学领域应用广泛？