在这里仅以一成功的例子阐明这一系统的工作原理，以便举一反三。Tate P et al.（1998）在哺乳细胞中设置基因陷阱扫描新基因，结果他们使一小批新的染色体蛋白和蛋白基因掉入了“陷阱”。在这一工作中，主要依靠由报告基因（β-galactosidase-neomycin phosphotransferase，*geo*）组成的基因陷阱构建体（construct）。这一构建体具有在3种不同的阅读框架上，缺失自身的启动子和起始密码 ATG，并且在5′端带有切拼接收器（splice acceptor）的特点（图 6.14A）。通过药物来筛选整合有载体的细胞，其原理是只有以正确的方向整合到内含子（intron）的表达基因才有可能在框架上被切拼成基因转录本（transcript）的一部分。这一转录本具备了编码 neomycin phosphotransferase 和 β-galactosidase（*lacZ* / β-gal）融合蛋白（fusion protein）能力（Freidrich G and Soriano P, 1991），所以可以很容易用简单的组化法对后者染色以达到鉴定的目的。对 β-gal 加入蛋白质功能域可以展示所产生的嵌合蛋白在亚细胞区域的变化情况（Gossler A *et al.*, 1989; Skarnes WC *et al.*, 1992; Burns N *et al.*, 1994）。这种融合蛋白的稳定表达对细胞功能没有明显的影响。

载体设计使快速的通往由 pGT1-3 基因陷阱插入破坏的基因的分子通路变为容易。掉入“陷阱”的基因序列可通过用 5′RACE 法（5′rapid amplification of cDNA ends）获得（图 6.14），而基因的特征可由 EST 和基因序列数据库帮助分析。

实验用的细胞可是维持在 LIF（leukemia-inhibitory factor）中的雄性 ES 细胞系（Tate P *et al.*，1996）或一般的体细胞杂交体。分别为 50g 最高浓度的 pGT1，2 和 3 由 HindIII 切割后呈线性、再由电穿孔法转入细胞，用含 Geneticin/G418 的培养基进行筛选。用 ES 细胞进行研究的优点是，一旦需要即可把它们整合入胚胎中进行分化，形成小鼠体的一部分，以进行基因功能的研究。

研究结果显示，基因陷阱法是一种有效的鉴别新蛋白质的方法，特别对鉴定那些处于非常特定的亚细胞区域的蛋白质更为有效。

（贺　林　薛　红）

6.10　“计算机杂交”

6.10.1　概论

核酸分子杂交是以碱基之间的 A 对 T 和 G 对 C 配对为基础，核酸分子之间能否杂交取决于两个因素：①杂交条件；②核酸分子之间的同源性（或相似程度）。当在严格条件下，只有同源性很高或完全同源的分子才能相互杂交；当在宽松条件下，有一定同源性即能相互杂交。现已开发出许多软件，可对你已获得的序列与计算机核酸序列及蛋白质序列数据库进行同源性比较，或对数据库中不同物种间的序列进行比较，我们可通俗地称之为“计算机杂交”。

自从人类基因组计划开展以来，有关研究成果和数据迅速增长。在所有这些数据中，增长最快的是 DNA 序列的数据以及由之而来的蛋白质氨基酸序列的数据。在基因组测

序方面，根据基因组监视表（genome monitoring table；http://www.ebi.ac.uk/~sterk/genome-MOT）的统计数据，在人类基因组大规模测序方面，到1998年11月16日已测定序列有233 221 491bp，占全基因组6.9%。模式生物基因组中，酵母（*S. Cereviae*）全基因组（12Mb）已完成测序，线虫（*C. Elengans*）全基因组（100Mb）也与1998年底完成测序。由于基因的发现属于知识产权保护范围，基因本身是药物开发和临床诊断的基础，因此人类致病基因和有功能基因的克隆成为国际竞争焦点之一。GenAtlas数据库（http://www.citi2.fr/GENATLAS）1998年11月11日收录的人类遗传病表型1 800种，基因7 800个，遗传标记18 100个。由于许多大的制药公司竞相投资对表达序列测序，EST数据急剧增加，1998年11月18日GenBank中收录的EST达1 967 546条，其中人类EST达1 170 219条（http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html）。

在怎样收集、整理、分析序列数据方面，已有许多软件可供利用，并有一些专门书籍详细介绍操作方法（Griffin AM and Griffin HG, 1994；Bishop MJ, 1998）。在此我们介绍序列的同源性比较在EST数据库整理及新的人类基因克隆等方面的应用。

6.10.2　序列比较在EST数据库整理中的应用

自从Venter（1991）提出EST概念以来，EST测序倍受重视。其原因主要有三，①人类基因组大约有3%为编码序列，这些编码序列分散在人的基因组序列中。通过大规模的EST测序，可拼接出大约一半人类基因。②如果一个EST在基因组中只出现一次，那么它还可做为STS。由EST构建的物理图叫表达图或转录图（expression or transcript maps）（Boguski MS and Schuler GD, 1995）。③EST测序过程是从cDNA文库中随机挑取克隆，用载体引物从5′和3′端进行一轮测序。从挑取克隆到测序和数据输入全过程可自动化，省时省资。

EST测序也存在一些缺点：①EST很短，没有给出完整的表达序列；②由于只是一轮测序结果，出错率达2%～5%，并且有时有载体序列和核外mRNA来源的cDNA序列污染；③有时出现相嵌克隆；④高度冗余序列等。因此，有必要对原始数据库进行整理加工。

公共数据库中，EST数据主要来自3个数据库：EMBL（European Molecular Biology Laboratory）（Stoesser G *et al.*，1997），GenBank（Benseon DA *et al.*，1997），DDBJ（DNA Databank of Japan）（Tateno Y and Gojobori T, 1997）。这3个数据库每天进行数据交换。这些EST数据为原始数据（raw data），经过质量监控（quality control）去掉序列中含N高于3%的序列和长度短于100bp的序列，以及通过序列比较加工而成为"增值"的数据。

以序列的同源比较为基础的EST数据加工主要包括：①对3′端EST进行序列比较，将EST进行分组（或簇），每一组（或簇）代表一个基因；②对大规模随机测得的EST进行整合形成可能的人类一致序列（Tentative Human Consensus，THC）。EST分析软件包括对cDNA序列进行分组的ICAtools软件包（Parson J *et al.*，1992；Parson J, 1995），对大规模随机测序的EST进行整合组装的TIGR软件（Sutton GG *et al.*，1995；Adams MD *et al.*，1995）。

经过加工整理的EST数据库有：①UniGene和UniEST：UniGene计划的目的是构建人类基因转录图谱（Boguski MS，1995；Schuler GD *et al.*，1996）。UniGene是来自于GenBank中的非冗余性基因，而UniEST是与UniGene不匹配的非冗余性EST。UniEST

96.0 版本包括来自人类 325 488 个 EST 中的 53 356 组 EST (Bishop MJ, 1998)。②Mer基因检索目录:Merck 基因检索目录计划(Aaronson JS *et al.*, 1995)是大规 EST 测序的产物,其目的是建立一个相似基因检索目录。该目录还提供插入 cDNA 片段的长度,这样有于将 EST 组装成一致序列时,估计 EST 的相对位置和间隙的大小(Bishop MJ1998)。③TIGR 的人类基因检索目录(HGI):TIGR (The Institute for Genomic Research)的 HGI 的目的是综合 TIGR 自己的数据和公共数据库的数据,建立一个非冗余的,代表人类全部基因的检索目录,同时还提供相应基因序列,蛋白质序列,定位信息和相关文献(Bishop MJ 1998)。④STACK 和 SaniGene:STACK (Sequence Tag Alignment Consensus Knowledge-base)数据库是具有一定相同程度的表达序列的数据库(Bishop MJ, 1998)。

6.10.3 序列同源性比较在新基因克隆中的应用

人类基因组计划的开展使得基因的定位、测序和功能研究等方面积累了大量数据,充分利用这些已有数据资料,可加速人类基因克隆研究。因此,有必要在此讨论序列的同源比较在新基因克隆方面的应用。

序列比较(sequence alignment)可分为整体比较(global aligmment)和局部比较(local alignment)两大类。由于有重要功能的序列往往是保守序列,所以对数据库进行序列比较检索时,局部比较更为常用。目前最流行的可用于序列比较的算法有 FASTA 和 BLAST。而对数据库进行检索常用的为 BLAST (Basic Local Alignment Search Tool)。这种搜寻常用于检索新的 DNA 序列是否已经发表,它与什么序列有同源性等。许多网址提供 BLAST 搜寻界面,最常用的是 NCBI 提供的界面。NCBI BLAST 网页有"basic"和"advanced"搜寻两类。一般情况下应用"basic" BLAST,只有在有详细相关资料时才用"advanced" BLAST。BLAST 有包括多种测序:

- Blastn:核酸序列与核酸数据库比较;
- Blastp:氨基酸序列与蛋白质数据库比较;
- Blastx:核酸序列的 6 种翻译序列与蛋白质数据库比较;
- tBlastn:蛋白质序列与核酸数据库所有 6 种翻译序列的比较;
- tBlastx:核酸序列的 6 种翻译序列与核酸数据库所有 6 种翻译序列比较。

当选定好程序和需要搜寻的数据库范围,输入序列(常用 FASTA 格式),然后就可开始搜寻。

现在国际上相当数量的人类新基因克隆都是从同源 EST 分析开始的。Koonin EV (1997)通过计算机分析,鉴定出一个大的核糖体相关蛋白超家族(superfamily of ribosome associated proteins),进而推测 Gadd45 是细胞周期检测处(checkpoint)信号通道成员。Hardiman 等通过同源分析方法,克隆、定位了人类 WNT 基因家族的新成员 WNT10B (Hardiman G *et al.*, 1997)。Rossi DI *et al.* (1997)通过同源序列分析发现了人的细胞趋化因子(chemokines) MIP-3alpha 和 MIP-3beta。夏家辉等通过同源序列分析克隆并定位了一个新的蛋白激酶基因 DyRk3,并确定了该蛋白激酶在蛋白激酶超家族中的分类位置(夏家辉等,1998)。刘春宇等通过对 EST 数据库同源分析识别和定位了人的 auxilin 基因(刘春宇和夏家辉,1998)。

下面介绍应用同源比较在人类 EST 数据库中识别和拼接与已知基因高度同源的人类新基因(图 6.15)。

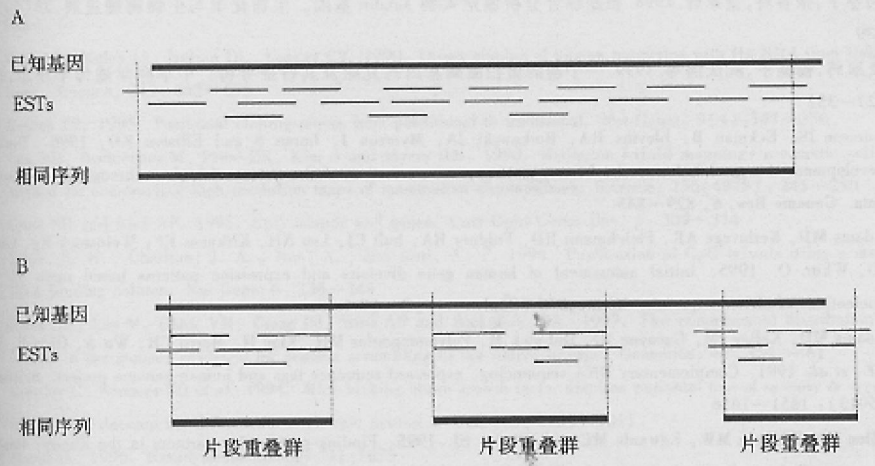

图 6.15　应用已知基因对 EST 数据库进行同源性比较构建 EST 重叠群

• 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 序列高度同源的 EST。

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

• 比较各重叠群的一致序列与已知基因关系(图 6.15)。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(图 6.15A);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部分序列(图 6.15B)。例如用牛的 auxilin cDNA 序列对 EST 数据库进行 Blast 分析得到的 EST 可形成 3 个与牛的 auxilin cDNA 对应的重叠群(图 6.16)。可用 PCR 在 cDNA 文库中将重叠群之间的区域扩增并测序。

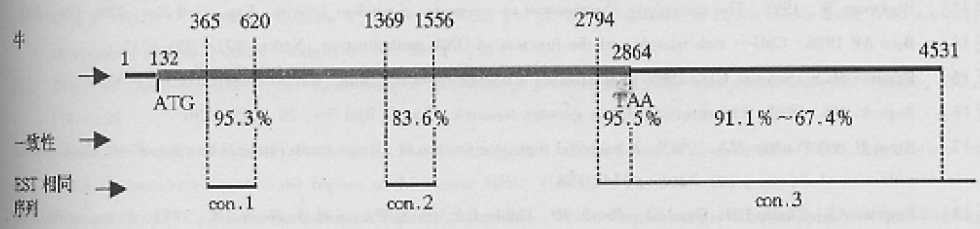

图 6.16　牛 auxilin 基因与人 auxilin EST 重叠群对比图(刘春宇和夏家辉,1998)

• 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。

• 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 的定位即确定了新基因的定位。

(于常海　杨新平)