基因表达序列标签(EST)数据分析手册

胡松年 主编

浙江大学出版社

基因表达序列标签(EST)数据分析手册

胡松年　主编

浙江大學出版社

内容提要

本手册以基因表达序列标签(Expressed Sequence Tag,EST)的处理分析流程为主线，结合大量实例，以"跟我学"的方式较为系统地介绍了EST测序，EST数据分析平台的构建，EST的文献检索和数据库查询，EST序列的聚类组装，EST的功能分类、代谢途径及基因产物的诠释等方面所涉及的方法和常用软件。此外，还介绍了如何利用大量的网上和专业软件来从基因和蛋白质水平分析基因及其蛋白产物的基本特性和功能。从核苷酸水平介绍了开放阅读框寻找、基因功能预测、基因结构分析、选择性剪切分析、基因多态性位点分析、基因表达调控区域分析、序列GC含量和密码子使用偏性统计，以及酶切位点和引物设计等内容。从氨基酸水平介绍了理化性质、结构域、二级结构、三级结构和系统进化等方面的分析。

本手册充分发挥网络优势，借助图文讲解，内容前沿新颖，为从事生物学、医学、农学、计算机科学等领域的科研及教学人员提供了一本十分有用的工具书，也是帮助读者掌握基因和EST数据分析方法的实验指导教材。

图书在版编目(CIP)数据

基因表达序列标签(EST)数据分析手册 / 胡松年主编. —杭州：浙江大学出版社，2005.5

ISBN 7-308-04186-7

Ⅰ.基… Ⅱ.胡… Ⅲ.基因表达—数据—分析—手册 Ⅳ.Q753

中国版本图书馆CIP数据核字(2005)第028631号

责任编辑 应伯根 张 利
封面设计 赵 胜
出版发行 浙江大学出版社
(杭州浙大路38号 邮政编码310027)
(网址:http://www.zjupress.com)
(E-mail:zupress@mail.hz.zj.cn)
排 版 浙江大学出版社电脑排版中心
印 刷 浙江大学印刷厂
开 本 730mm×980mm 1/16
印 张 15.5
字 数 275千字
印 数 0001-2000
版 印 次 2005年5月第1版 2005年5月第1次印刷
书 号 ISBN 7-308-04186-7/Q·049
定 价 45.00元

浙江大学沃森基因组科学研究院

www.wigs.zju.edu.cn

基因组科学研习班

实习指导用书

序

基因组学或称基因组生物学是一个发展迅速、数据与信息量浩大的新兴生命科学领域。在这个领域的发展过程中，新技术、新方法和新概念层出不穷，有时也会令人眼花缭乱。这样一来，有个“向导”，有个“路标”，有个“地图”，都会是雪中送炭，使人豁然开朗，甚至感觉到海阔天空。这就是我们浙江大学沃森基因组科学研究院和杭州华大基因研发中心的老师、学生和员工们努力工作，开办学习班，编写教材的基本目的。由胡松年和薛庆中老师主编，我们的学生们参与写作的《基因组数据分析手册》已于2003出版发行。现在二位老师又和新一批学生们将它的续篇——《基因表达序列标签(EST)数据分析手册》呈现给读者。我们确实应该感谢这些探路和铺路的勇者，向他们数谢和致敬。

基因组生物学是一门涉及面很广的学科，不仅仅是因为有生命，就有基因和基因组，而且数据本身的复杂性也不是一群人、几个部门、乃至一个学科能够理解和懂得的，需要交流、交叉和合作，建立一个和谐、有力的系统。基因组学的一个非常突出的特点是它的基本研究手段是所谓“数据导向”，至少其研究分量要比“假说导向”的研究多，数据从而成为主要的决定因素。对于从事基因组学研究的人来讲，没有原始数据，创新的可能性就小了，科学也就难做得多了，因为科学的本质是创新，而不是将别人的来西拿过来。这不仅仅是“嚼别人嚼过”的事情，而是能不能生存的问题。对于其他仅仅使用数据的生物相关学科而言，从“数据共享”的殿堂里拿数据来用，是可以的。可是同样的数据别人也有，

你的优势在哪里呢？毫无疑问，分析数据、理解数据、运用数据的能力变得“性命攸关”了。其实，从数据到信息仅仅是第一个“瓶颈”，另一个“瓶颈”是从信息到知识的转化。数据到信息的转化，造就了生物信息学，不仅吸引了诸多的计算机科学家和数学家来“淘金”，也引来物理和工程学家来据“乌龙”，于是又有了所谓的多系统生物学，使本来学科纷纭的生物学领域更加“车水马龙”，“山头林立”。这其实是好事情，数据里的信息可以从不同的角度被解读。从信息到知识的转化，则需要生物学家们也来共同努力。生物学家既熟悉数据的真伪，也急需信息来指导新的实验设计，同时也验证信息的准确性。但是信息毕竟不是很容易理解的知识，只有专家或具有基本知识的人才可以解读。所以，想利用信息的人也要具备基本知识。这样一来，从数据到信息，再从信息到知识，一个新的生命科学“流水线”就产生了。这个“流水线”的工作数率在生命科学的发展和应用上成为关键和必由之路。

我的同事们和我本人最高兴做的事，也是我们的科研领域和本职工作，那就是成为这条流水线上的一个“零件”和这条“流水线”的“润滑剂”。

2005 年 3 月 18 日于杭州

编者的话

基因表达序列标签(Expressed Sequence Tag,EST)是通过随机挑选cDNA文库并进行单向测序所获得的序列(约为60－500bp),其上携带着表达基因的部分遗传信息。和基因组序列相比,EST序列具有快速、简便、易得的优势,其价值得到了越来越多的科学家的认可,已成为基因组学研究领域中不可或缺的一项内容。由于EST在研究上的重要性,1993年NCBI专门建立了EST数据库dbEST(database of EST),系统地收集和保存所有的EST数据,记录了每个EST序列的登记号、测序引物、碱基序列、cDNA文库构建方法、组织来源等详细资料。到2004年10月1日,dbEST已收录来自700余个物种2400万个EST序列,这些海量的数据中蕴涵了大量有价值的生物信息。

总体来说,EST可用于发现和预测新基因、新的SNP位点和分子标记,构建基因表达谱和基因组物理图谱等研究,这些应用将在本书第一章中有较为详细的介绍。对任何一个物种来说,系统地对其进行EST测序和分析,往往是获得该物种基因的总体组成及其表达特征的最直接而快捷的方法。

随着基因组学领域的飞速发展,EST数据呈爆炸性增长,研究人员面对的不再是一条或几条EST序列,而是成千上万条EST序列。那么,如何根据研究目的采用适当的EST测序手段,如何选择和使用EST分析工具,如何系统而高效地从中挖掘其所蕴含的生物学信息,已成为研究人员尤其是初学者和研究生所十分关注的问题。为推动基因组学这一新的学科在国内的普及和发展,浙江大学沃森基因组科学研究院和杭州华大基因研发中心

自 2002 年以来已成功举办了 13 期基因组生物信息学培训班，被《科学美国人》评为 2002 年度世界科学领袖的杨焕明和于军每次都前来培训班讲课，受到来自全国各地 29 个省市的科研单位和高等院校学员们的好评。2003 年 5 月，胡松年和薛庆中主编了《基因组数据分析手册》，为初学者系统介绍了基因组数据的处理分析过程及所涉及的相关软件工具，但每期培训班中仍有不少学员强烈建议增加 EST 分析的内容。为了使培训班能更加贴近学员的实际需要，我们从 2004 年 7 月起，根据近年来从家猪、水稻等基因组计划中对海量 EST 数据分析所积累的经验，详细地制定了 EST 的培训计划和内容，并组织具体从事 EST 分析的研究生和工作人员编写相应章节的讲义。为确保培训质量，每个培训老师都经过多次试讲，并邀请浙江大学薛庆中、傅衍、张传溪等教授参加旁听并提出批评和建议。2005 年 2 月在第 13 期基因组生物信息学培训班上，我们正式推出了这一全新的内容，并得到了学员们的一致好评。结合学员们所提出的建议，我们在原讲义的基础上编写了这本《基因表这序列标签(EST)数据分析手册》。希望这本书能对 EST 致据分析的学习和应用起到积极的推动作用。

本手册沿续了《基因组数据分析手册》的写作风格，采取“跟我学”的方式，重点向读者介绍 EST 测序分析过程中所常用的工具和方法，并对其中的常见问题进行探讨。下面对全书的主要内容作一简要说明。

本书第 1 章概要描述了 EST 测序分析的整个流程，包括 cDNA 文库构建、测序、拼接及分析等内容，简介了 EST 在基因组学、分子生物学等方面的应用。

第 2 章主要介绍了搭建 EST 分析平台所需的软硬件设施。极据数据处理能力的大小，我们推荐了不同配量的服务处理器，并简介了 EST 分析所常用的 Unix 和其他多用户操作系统及其安

装方法，并列举一些常用的Unix命令。同时，以4个常用软件为例，分别介绍了在Windows操作系统以及Unix系统下安装各种软件的方法和步骤。

第3章重点介绍了EST分析处理过程中常用的两个综合性数据库dbEST和UniGene的内容和使用方法，同时介绍了一些常用的文献检索数据库，对批量文献的查询和管理也作了概述。

第4章简要介绍了几种常用的聚类组装软件的用法。考虑到目前研究人员多采用Windows操作系统，本章重点对Windows操作系统下的EST聚类组装进行了讲解。

第5章根据大规模EST数据的分析流程，分别从网上运行和本地运行两个方面系统地介绍了BLAST比对方法，EST数据的注释、多序列比对工具Clustal W的使用以及亲缘树的绘制。

第6章围绕着如何对EST进行功能注释、功能分类以及代谢途径分析等重要内容，重点介绍了GO、KEGG、COG等常用工具的使用。

最后两章较为详尽地介绍了如何利用大量的网上和专业软件来从基因和蛋白质水平分析基因及其蛋白产物的基本特性和功能。第7章主要以核苷酸序列为研究对象，介绍了开放阅读框寻找、基因功能预测、基因结构分析、选择性剪切分析、基因多态性位点分析、基因表达调控区域分析、序列GC含量和密码子使用偏性统计，以及针对实验设计酶切位点和引物等多方面的内容。第8章则针对所有蛋白质序列，推荐了一些蛋白质数据库，如SWISS－PROT、PDB等，对这些数据库的熟练使用将使我们更全面地了解和认识细胞内各种蛋白特定的结构功能、表达时空、加工修饰、代谢凋亡等特性。同时，本章还从蛋白质的理化性质、二级结构、结构域、三维结构和系统进化分析5个方面介绍了相关的分析方法和常用工具。

这本手册由于是多位编者参与撰写，各章内容侧重不尽一致。我们在汇总成书时，对全部文字和图表作了认真的修整，力求文笔流畅一致。同时在书的最后，建立了常用术语的索引，以方便读者查询。然而，基因组学这门前沿学科的发展是如此迅猛和广博，即便是我们也会对一些最新出现的术语和方法难以作出非常准确的说明和注释。因此，对本书中所出现的错误和缺点，恳请误者批评指正。

在本书的编写过程中，得到了浙江大学沃森基因组科学研究院程家安院长、杨焕明副院长热请的鼓励，于军教授特索为本书作了序，他们为本书的顺利完成提供了有力的支持。最后，还要对陈爱华者师在书稿策划、组织工作中所作的贡献表示衷心的感谢。

胡松年

2005 年 3 月于杭州

目　录

第1章　基因表达序列标签(EST)及其应用

胡松年

1.1　简　介

基因表达序列标签(Expressed Sequence Tag,EST)是从cDNA文库中随机挑取单克隆进行测序,所获得的序列片段,序列长度一般约为60－500bp。

为能较快速地发现新基因,早在20世纪80年代,有人曾提出对cDNA序列进行大规模测序,但对此想法一直存在争论,反对者认为cDNA序列只含有基因编码区的序列,缺乏基因调控区域的信息。90年代初Craig Venter首次从3个人脑组织的cDNA文库中随机抽取609个克隆测序分析得到了一组表达序列标签数据。这一研究成果的发表,开创了cDNA大规模测序(EST测序)时代。

和基因组序列相比,EST序列具有快速、简便、易得的优势,其价值得到了越来越多科学家的认可。1993年NCBI专门建立了EST数据库dbEST(database of EST),系统地收集和保存所有的EST数据。在dbEST中记录了每个EST序列的登记号、测序引物、碱基序列、cDNA文库构建方法、组织来源等详细资料。到2004年10月1日,dbEST已收录23 970 155个EST序列,分别来自736个物种,其中人和小鼠的占大多数。

本章简要介绍了EST在基因组学、分子生物学等方面的应用,描述了EST测序分析的整个流程,包括cDNA文库构建、测序、拼接及基因分析等内容,说明EST的研究和分析是连接结构基因组学和功能基因组学不可或缺的桥梁。

1.2 EST 的应用

EST 数据主要有以下几个方面的应用:绘制物理图谱和转录图谱、识别基因、电子 PCR 克隆、基因预测、发现 SNP 和研究基因表达水平等。

下面将分别对这些应用进行介绍。

1.2.1 基因图谱的绘制

转录图谱是指不同基因的转录产物在基因组上的分布位置。由于 EST 来源于生物体不同组织或同一组织不同发育时期的 cDNA 文库,因此,通过对 EST 数据进行分析整理,可以绘制不同的转录图谱。通过对这些转录图谱的比较,有可能发现组织间或同一组织不同发育时期特异转录表达的基因。

物理图谱是以特异的 DNA 序列为标记的染色体图谱,标记之间的距离以物理距离如碱基对(bp,Kb,Mb)表示。其中最为普遍的序列标志为序列标签位点(Sequence-Tagged Sites,STS),因此,物理图谱又常被称为 STS 图谱。STS 在基因组中是惟一存在的一段特异性短序列,长度一般在 200－300bp 间。它们来源于随机的基因组序列、遗传标记序列(如微卫星标记)及表达基因序列等。基因组中大多数基因为单拷贝序列,以表达基因的 EST 序列作为 STS,应用于物理图谱的绘制。有以下优点:

EST 序列中没有内含子的存在,因此,用非翻译区或同一外显子的特异性序列作为 STS,其 PCR 产物大小在 cDNA 及基因组模板中是相同的;基因家族成员间虽然其编码区具有很强的保守性,但各自的 3′非翻译区序列的保守性通常较差,若以 3′非翻译区序列为 STS,在基因组中较容易地分辨各基因家族成员。

1.2.2 基因识别

由于各数据库中 EST 的数目远比其他的核苷酸序列多,因而利用 EST 数据库搜寻新基因已成为基因识别的重要手段。通过在数据库中对 EST 序列进行比对(详见第 5 章),可以识别同一物种中基因家族的新成员(paralog genes),在不同物种间功能相同的基因(ortholog genes),以及同一基因的不同剪切模式。

1.2.3　电子 PCR 克隆

获得全长 cDNA 克隆(full-length cDNA)是进行基因表达和功能研究的前提条件。在 cDNA 文库构建过程中许多基因(尤其是大基因)cDNA 都缺少 5′端的序列信息。因此,通过随机测序获得全长 cDNA 十分费时耗力。

EST 数据库中存在着大量同一基因的 EST 序列冗余,通过聚类拼接(详见第 4 章),可获得较长的一致性序列(consensus sequence)。这一拼接过程通常借助计算机完成,所以又称为电子 PCR 克隆(e-PCR clone)。

目前,NCBI 的 UniGene (http://www.ncbi.nlm.nih.gov/UniGene)(见图 1.1)、TIGR 的 Gene Index (http://www.tigr.org/tdb/tgi/)(见图 1.2)和南非的 STACK (http://www.sanbi.ac.za/Dbases.html)数据库(见图1.3)都可提供同一基因转录本的 EST 序列,因而有利于电子 PCR 克隆。

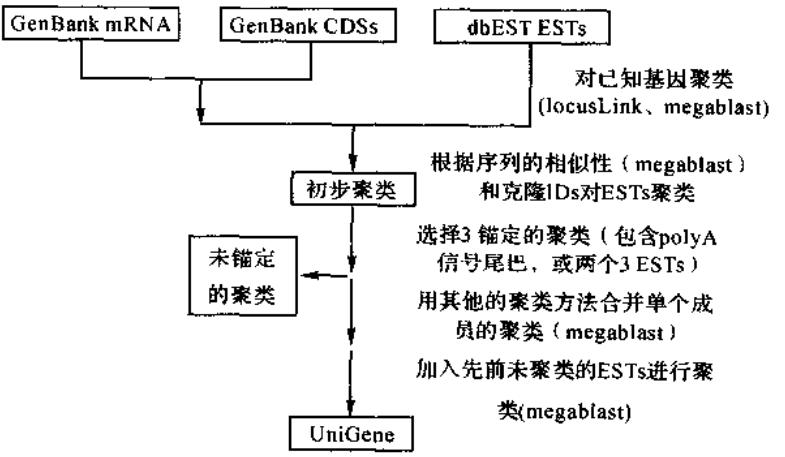

图 1.1　UniGene 数据库

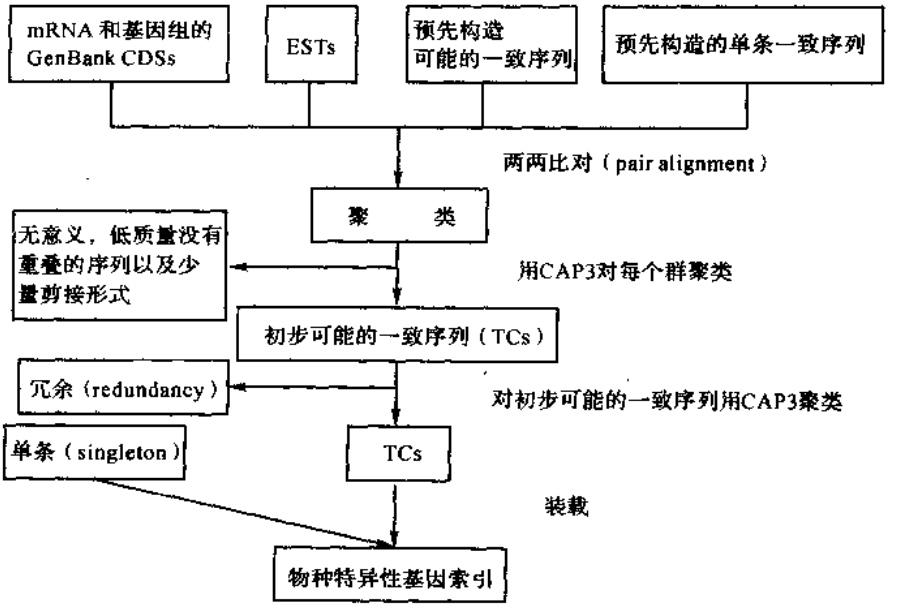

图 1.2　TIGR 基因指数

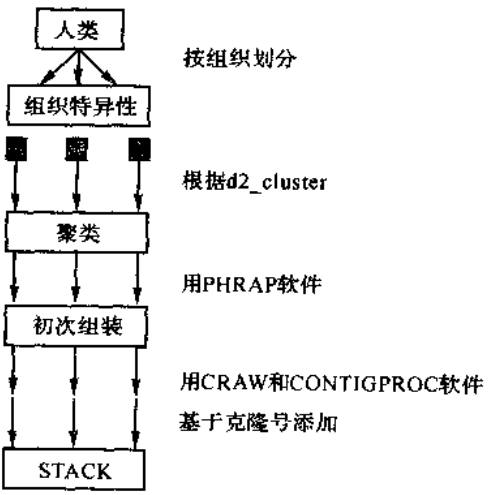

图 1.3　STACK 数据库

利用电子 PCR 克隆所得到的一致性序列，结合 PCR 技术和其他实验技术手段，可以较快捷地获得全长 cDNA 序列。大多数 cDNA 文库都是通过 oligo(dT)逆转录而成，这里仅对 cDNA 5′末端克隆所常用的实验方法作简单介绍：

● cDNA 文库 5′末端扩增法。根据电子 PCR 克隆的一致性序列，在 cDNA 文库序列近 5′端设计下游特异引物，利用 cDNA 文库的混合 DNA 为模板，进行 PCR 扩增。选取最长的 PCR 产物克隆测序，获得更多的 cDNA 5′端序列信息。若所行 PCR 产物仍未包含完整 5′末端序列，可依上述策略重新设计下游引物，直到获得完整的 5′末端序列为止。

● cDNA 末端快速扩增(Rapid Amplification of cDNA Ends，RACE)法。这是目前获得全长 cDNA 序列最常用的方法。传统的 5′RACE 有一条根据 cDNA 序列设计的特异引物，而另一条为简并的多聚同聚体锚定引物，这样，其扩增产物中往往含有全长和断裂的 cDNA 产物。它们的特异性较低，其扩增效果易受引物、一链合成的特异性、目标 mRNA 的丰度和复杂度，及扩增序列的长度等多因素的影响。通常尚需经过多次设计和实验优化才能成功。近来 Invitrogen 公司开发的 GeneRacer(tm)试剂盒针对全长 5′加帽的 mRNA RACE 技术，大大提高了特异性和成功率。

● cDNA 文库直接筛选法。根据电子 PCR 克隆的一致性序列，在其近 5′端设计一对特异引物，利用原位杂交或 PCR 扩增 cDNA 混合矩阵文库(pooled library)的方法直接筛选阳性克隆，通过随机对若干阳性克隆进行 5′测序，以获得全长 cDNA 序列。

1.2.4　基因识别

当一个物种的全基因组测序完成后，首要的工作就是对其基因组中所包含的全部基因进行预测。由于不同物种在碱基组成、重复序列、基因结构等方面存在较大的差异，迄今的基因预测软件不可能对所有的物种都达到很高的准确度。因此，对预测基因的验证就显得至关重要。EST 来源于基因组中转录出的 mRNA，每一条 EST 均代表了特定发育时期和生理状态表达基因的部分序列。因此，使用合适的比对参数，将预测基因与同物种的所有 EST 进行比对，有助于对基因识别。

此外，EST 对预测基因的交替剪切和 3′非翻译区很有效。以 EST 为训练集，可提高基因预测算法的准确度和灵敏度。