ISSN 1007-7626
CN 11-3870/Q
中国生物化学与分子生物学报
Chinese Journal of Biochemistry and Molecular Biology
2007 年 3 月
23(3) :236～244

技术与方法·

GEO(Gene Expression Omnibus) :高通量基因表达数据库

刘 华*, 马文丽, 郑文岭
(南方医科大学基因工程研究所,广州 510515)

摘要 GEO(Gene Expression Omnibus)数据库包括高通量实验数据的广泛分类,有单通道和双通道以微阵列为基础的对 mRNA 丰度的测定;基因组 DNA 和蛋白质分子的实验数据;其中包括来自以非阵列为基础的高通量功能基因组学和蛋白质组学技术的数据也被存档,例如基因表达系列分析(serial analysis of gene expression,SAGE)和蛋白质鉴定技术.迄今为止,GEO 数据库包含的数据含概 10 000 个杂交实验和来自 30 种不同生物体的 SAGE 库.本文概述了 GEO 数据库的查询和浏览,数据下载和格式,数据分析,贮存与更新,并着重分析 GEO 数据浏览器中控制词汇的使用,阐述了 GEO 数据库的数据挖掘以及 GEO 在分子生物学领域中的应用前景. GEO 可由此公众网址直接登陆 http ://www. ncbi. nlm. nih. gov/projects/geo/.

关键词 基因表达; 数据库; 控制词汇; 数据挖掘
中图分类号 Q254 ;Q26

GEO (Gene Expression Omnibus) : High-throughput Gene Expression Database

LIU Hua*, MA Wen-Li, ZHENG Wen-Ling
(*Institute of Genetic Engineering, Southern Medical University, Guangzhou* 510515, *China*)

Abstract The Gene Expression Omnibus (GEO) database, the first public repository for gene expression data, premiered at National Center for Biotechnology Information (NCBI) in July 2000. The GEO database contains a wide assortment of high-throughput experimental data, including single and dual channel microarray-based experiments measuring the abundance of mRNA, genomic DNA and protein molecules. Data are also archived which origin from non-array-based high-throughput functional genomics and proteomics technologies, including serial analysis of gene expression (SAGE) and protein identification technology. To date, the GEO database contains data representing almost 10 000 hybridization experiments and SAGE libraries from 30 different organisms. This paper outlines the query and browse in GEO database, data download, format, data analysis, and deposit and update. Also, it focuses on the managing terminology used in the GEO-data browser, while describing the course of data mining and GEO's future applications in the field of molecular biology. GEO is publicly accessible at http :// www. ncbi. nlm. nih. gov/projects/geo/.

Key words gene expression; database; managing terminology; data mining

GEO(Gene Expression Omnibus)数据库,作为第一个基因表达数据的公共贮存库,2000 年 7 月在 NCBI 上首次公布于众.创建 GEO 的最初目的是适应高通量实验方法在将来的普遍发展,具有强大的数据收录功能.因此,它有着极大灵活性和与时俱进的设计风格,不需设立严格的登陆要求和标准. GEO 是支持符合 MIAME(minimum information about a microarray experiment)数据提交的基因表达/分子丰度库[1],关于基因表达数据浏览,查询和检索的在线资源. GEO 可作为广泛的高通量试验数据的公共贮存库.这些数据包括单通道和双通道的微阵列实

收稿日期: 2006-09-26, 接收日期: 2006-12-25
广东省重点实验室基金资助
*联系人 Tel: Tel :13560475855, E-mail :annylh2008 @126. com
Received: September 26, 2006; Accepted: December 25, 2006
Supported by Key Laboratory Program of Guangdong Province
*Corresponding author Tel: 13560475855
E-mail: annylh2008 @126. com

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

验[2,3],用于测量 mRNA[4,5],基因组 DNA 和蛋白质丰度,以及非阵列技术,如基因表达系列分析(SAGE)[6]和质谱分析蛋白组学数据. 至今公共数据中已有2 807个 GPL (platforms) 平台, 105 243个 GSM (GEO samples) 样品,4 476个 GSE (series) 系列.

1　数据库的构成

1.1　提交到 GEO 的数据分类

平台　平台记录描述阵列上的成分(例如, cDNAs, 寡聚核苷酸探针, ORFs, 抗体)或在实验中可检测和定量的成分(例如 SAGE 标签, 肽).

样品　样品记录描述个体样品信息,经过的处理,和每个元素的丰度测定,即关于被检测的 mRNA 样本,实验条件,和实验产生的基因表达测量数据信息.

系列　系列记录定义一组相关样品,样品间如何相关,以及是否有序和怎样排序. 就整体而言,系列提供试验的焦点描述. 系列记录也包含描述提取数据,概括结论,或分析的表格. 每一个系列记录指定唯一固定 GEO 登陆号(GPL xxx)[7].

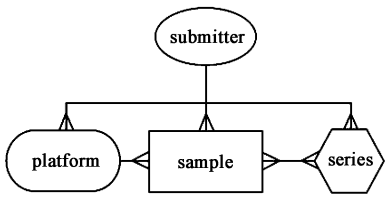


Fig. 1　The entity-relationship diagram for GEO

GEO 为了能够有效地检索,显示和分析数据,几个新的工具和特性已经开发. 为了创建这些工具, GEO 数据首先组合为可比较的集合,或 GEO 数据集(GDS). GDS 代表在生物学和统计学上有可比性的 GEO 样品. 两个新的数据库已经被创建用于查询这些数据-Entrez GEO 表达谱和 Entrez GEO 数据集. Entrez GDS 查询数据定义和原始的实验注解以确认感兴趣的实验. Entrez GEO 表达谱显示每一组数据的个体基因表达/分子丰度图谱.

1.2　GEO 数据集

GEO 数据集(GDSxxx)是当前的 GEO 样品数据库. GDS 记录代表生物学上和统计学上可比较的 GEO 样品集合及数据显示和分析工具的基础形式[8]. 在 GDS 内的样品是指相同的平台,也就是它们分享一组共同的探针. GDS 内的每一个样品值的测定假定通过等值的方法计算,即背景处理和标准化等条件在数据组中是连贯的. 它是存贮治疗基因表达和由基因表达库组成的分子丰度数据库.

1.3　GEO 表达谱

这个数据库贮存个体基因表达和由基因表达库组成的分子丰度图,可以通过基因注解或预处理的图谱特征寻找感兴趣的特殊图谱[9,10]. GEO 表达谱有利于强大的搜索和附注资源的链接.

2　查询和浏览

2.1　检索 GEO 数据的方式

使用显示条查看有登陆号的详细 GEO 记录,此工具对格式和查看的数据量有多个选项.

使用 Entrez GEO 数据集或 Entrez GEO 表达谱界面查看某个特殊领域或整个领域的所有 GEO 投稿, Entrez GEO 表达谱可查询预处理的基因表达/分子丰度图谱,而 Entrez GEO 数据集则查询所有的实验注解. 正如其它的 NCBI Entrez 数据库,可以使用布尔短语,并限定在支持的特征字段,能够进行有效的查询和挖掘.

使用 GDS 浏览器或查看当前的 GEO 存贮目录,可以浏览 GEO 数据和实验列表.

2.2　GEO 数据浏览器

2.2.1　控制词汇与非控制词汇

控制词汇是标准化术语的选择,具有固定的意义,用来避免不同标题下相关主题意义的弥散. 控制词汇处理有许多方式说明同一事物的情况,多年前已经用于信息学和其它领域. 例如,在航空控制领域,管理员和飞行员一样使用相同的术语描述航空器的不同类型,它们的高度,方向和速度,所有这些对航空工业的安全运行是至关重要的. 如果不使用这些词汇,将导致飞行员错误地转向,使别的飞行员转向或以错误的高度飞行,所有这些可能导致事故发生.

虽然因特网中非控制词汇的使用不会导致事故,但控制词汇是非常有用的,因为它们有助于元数据创建者和查找者使用相同的意义.

搜索引擎通过改变不同术语的搜索方式,使用先进的方法加快非控制词汇的搜索:

自由词搜索　自由词搜索包括以正文字符串的形式来自专业和日常用语的单词,短语或句子. 自由词搜索最普遍地用于使用者寻找地理学名称,其它固有名称,或当使用者寻找的概念在他们搜索的来源物中的陈述很少的时候. 查找引擎会匹配所有的单词,按照从最相关(如大多数单词匹配)到最不相关(如

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

一或两个单词匹配)的顺序排列. 通常自由词搜索比控制词汇搜索检索到更多的采样数,但是很可能返回的结果和正搜索的主题不相关.

关键词检索 关键词检索是至今为止因特网搜索最普遍的形式,被大多数搜索引擎所使用. 当使用关键词时,用于搜索的引擎将忽略所有的介系词,如 the, an, at, with, for.

2.2.2 控制词汇与非控制词汇的使用

控制词汇的使用比非控制词汇有更多的优点. 但是,可以用非控制词汇查找某些时段. 这两种方法最好一起使用,控制词汇的查找用于限定主题范围,非控制词汇的查找用于精炼主题范围内的查找. 对于它们本身而言,每一种方法都是独立的,而不是综合的搜索工具. 两种工具结合起来可以更为精确的,广泛的搜索工具.

因此,推荐 GEO 数据浏览器使用两种方法结合的方式,包括针对高级使用者的嵌套法,布尔运算,截词和近似查找. 扩展的,限制性和相关的术语的结构控制词汇的引入加强了 GEO 数据浏览器的查找能力.

GEO 数据浏览器仅仅搜索包含几千个记录的数据库,然而,搜索引擎如 Excite 在包含数百万记录的巨大资源库中查找. 尺寸的不同可能影响我们利用不同类型的查找技术. 很明显,使用者会采用简单的搜索策略,从查询结果中发现它们搜索的内容. GEO 数据浏览器必须包含供这种用户使用的界面[11]. 然而,这并不意味着高级的查找技术没有效用. 我们宁可主检索界面是单一的,与高级检索界面分离,这样不熟悉高级技术的用户不会被复杂的多功能界面所迷惑.

3 数据提交与下载

3.1 数据的提交与更新

一旦提交者确定了自己的 GEO 帐号,便可通过以下 3 种方法来存储数据:

交互的网络形式. 每个提交的平台和样本,都可以上传和验证文本格式的数据文件,通过一系列的网络格式来交互上传原始数据,这个过程十分简单,尤其是提交的条目较少时. 单个记录的更新也可用类似的网络格式来执行.

用简单的文本格式(SOFT text)和 SOFT 格式(SOFTformat)直接提交. SOFT 格式是为快速批量提交数据而设计的,从普通的电子表格和数据库应用软件,可以很容易得到这些文件. 单个 SOFT 文件可以容纳 2 份数据和多个平台,样品和序列原始数据,可以直接上传到数据库. 批量更新也可用 SOFT 格式来快速,有效完成.

以有效的 MAGE-ML 格式传输文件到 GEO.

3.1.1 网络存贮简述

提交者分 4 部分提供数据:平台,样品,序列,和补充数据[12]

步骤 1 输入联系信息创立一个 GEO 帐号. 这是必要的公共信息,给数据提供适当的可信度.

步骤 2 检查 GEO 中是否存在平台. 如果实验使用商业阵列(如,Affymetrix)完成,可以不必提交平台记录,在上述情况下可以直接进入步骤 4. 注解为限定平台的不需要 SAGE 数据;也直接进入步骤 4.

步骤 3 提交对平台的限定,平台记录包括出现在阵列中的元素(如,cDNAs,寡核苷酸,ORFs,抗体)的数据列表和描述性信息(但是没有杂交测量结果). 使用者首先要从下拉菜单中规定平台类型,然后必须以文本,表格限定的格式提供平台数据表(Table 1). 数据表的第一行必须包括专栏标题;此后每一行表示仅仅一个元素,或者阵列上的"点". 平台数据表需要一个有标志符的列被命名为"ID",每一个 ID 在平台中是唯一的,并包括序列标记;对用于索引的 Entrez GEO 文档应引出此信息,并提供附加描述栏.

Table 1 The data for a non-commercial nucleotide platform

ID	GB ACC	Gene symbol	Gene name
1	U83857	API5	apoptosis inhibitor 5
2	M61764	TUBG1	tubulin, gamma 1
3	NM_012094	PRDX5	peroxiredoxin 5

在数据表通过批准后,使用者要提供平台标题,组织,描述性信息,和撰稿人.

步骤 4 提交杂交数据(或 SAGE 标签计数资料)作为样品记录. 一个样品记录参照一个平台,描述单一杂交/实验状态的丰度测量. 首先要说明样品类型并参照原平台 GEO 登陆号. 然后使用者必须以文本,表格分隔的形式提供样品数据表. 数据表的第一行必须包括专栏标题. 样品数据表需要一个命名为"ID_REF"的列来匹配参考平台的"ID"栏和"VALUE"栏(或 SAGE 数据的"TAG"和"COUNT". 对于双通道实验,VALUE 反应标准化(量化)的对数比测量值. 对于单通道实验,VALUE 是标准化的信号运算数据(非对数转换).

GEO 数据显示和分析工具仅仅在使用标准化值时才有效. 如果用户数据库中的中位数变异很大,

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

那么此数据库是非标准化的,不能合并入 GEO 的查询和分析工具.

步骤 5　用户提交了所有的样品数据后,提交一个系列记录.一个系列把样品的相关组集合在一起,提供一个集中和整体的研究描述.反应亚型的信息也可以详细说明.

步骤 6　在样品和序列号经过处理和批准后,GEO 服务人员会给用户发送电子邮件确认 GEO 序列号,要求用户提供相应的补充数据文件.补充数据类型的样品包括 cDNA 阵列,tiff 扫描图像,Affymetrix,CEL,EXP 文件,GenePix,gpr 文件.补充数据文件将传送到 GEO 的私人 FTP 站点-FTP 详细地址在确认的电子邮件中提供.

3.1.2　更新

对个人记录的编辑和更新,提交者可以通过选择在 Web deposit/update 页面上的"UPDATE"部分进行操作[13].

3.2　数据下载和格式

3.2.1　GEO 记录

在 Accession Display 栏(位于 GEO 主页底部和每一条 GEO 记录顶部)有几种选择可供原始 GEO 记录的检索和显示."Scope"可以是一个登陆号或关于登陆的任何记录(平台,样品或序列)或所有(家簇)记录."Amount"指显示数据的数量,选项包括元数据,元数据和数据表的前 20 行,数据表,或整个元数据/数据表记录."Format"是记录是否以 HTML 或以 SOFT 格式显示.SOFT 是设计为数据检索或提交给 GEO 的可机读的 ASC11 文本格式.

3.2.2　GDS 记录

每 1 个 GDS 记录对数据组的下载有 3 种选择.完整的 SOFT 文档包含整个数据组的所有信息,包括对数据组的描述,类型,组织,亚群的定位等.另外,数据表包括标记物和数值.

4　数据检索与分析

4.1　数据挖掘的作用

1)证实感兴趣的基因的表达动向,这在个体实验中可以忽略[14,15];

2)确认实验室结果或感兴趣基因功能的学术汇报[16,17];

3)提供值得在实验室中进一步研究的可能的候选基因[18];

4)关于异常通道或基因相互作用形成假说[19];

5)发现特征基因的新作用[20].

4.2　数据检索

GEO 数据可以使用 Entrez GEO 数据集和 Entrez GEO 表达谱进行查询.Entrez GEO 表达谱查询预处理的基因表达/分子丰度图谱,即样品和系列记录,浏览器网址为 http://www.ncbi.nlm.nih.gov/geo/query/browse.cgi,而 Entrez GEO 数据集查询所有的实验注解(http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds-browse.cgi),正如其它的 NCBI Entrez 数据库,可以使用布尔短语,并限定在支持的特征字段,进行有效的查询和挖掘[21].

对于感兴趣的实验,使用 Entrez GEO 数据集进行属性限定,如基因名,GEO 登陆号,关键词,变异性,组织,创建日期和平台等.例如,使用检索词'dual channel [Experiment Type] AND metastasis AND human [Organism]'寻找人类新陈代谢的所有的双通道核苷酸微点阵实验数据组.检索信息显示了数据组标题,简短实验说明,分类法,实验变量类型和原始平台的链接,相关系列记录和完整 GDS 记录.一旦确定相关数据集,可进一步研究感兴趣基因的表达图谱.

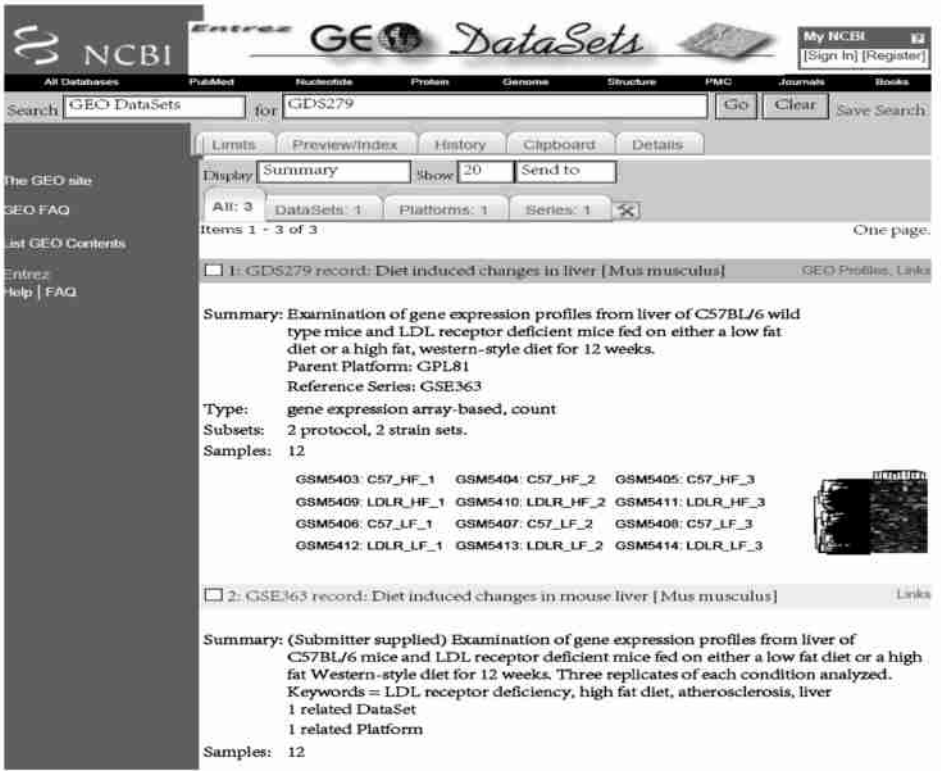
Entrez GEO 表达谱进行属性限定,如关键词,平台和样品类型,提交者,组织,发表日期和补充文档类型等.例如,利用检索词'Type 1 diabetes[GDS Text] AND apolipoprotein [Gene Description] NOT Homo sapiens [Organism]',检索到所有在非人类的物种中Ⅰ型糖尿病相关数据集中的载脂蛋白相关的基因资料.检索结果显示报告人的注解,简短实验信息,分类法和这个图谱的条形索引图.这个索引图对于快速、大量文档扫描、比较非常有用,单击索引图像可显示图谱的详细内容[22].

因为样品通常组合为系列内有意义的数据组,所以对一个系列及其相关样品和平台的检索更具有说明性.在 GEO 中检索某一感兴趣的数据系列的例子如下:

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

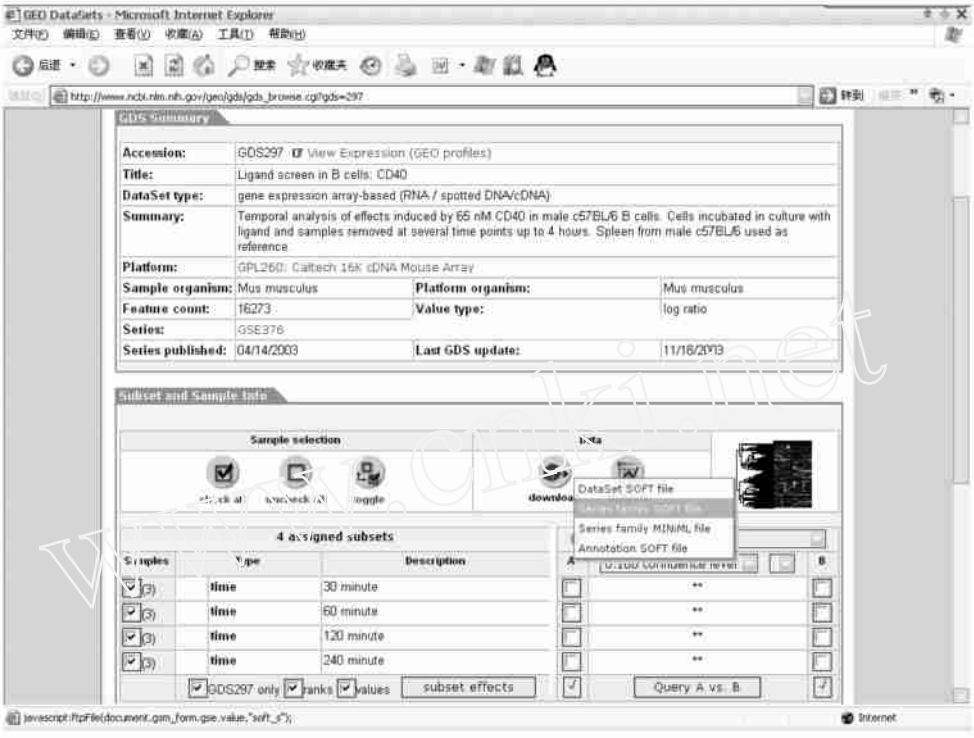
在 GEO 主页面的 GDS 检索框输入系列号 GDS279 ,单击 GO :

搜索结果显示为检索的 GDS 序列号及相关的平台与样品 ,单击 GDS279 record :



© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在“GDS Summary”上关于 GDS279 的描述是对数据的简要评价,在 download 中可以选择 SOFT 格式下载:



4.3　数据分析

通过 Entrez GEO 表达谱中的“Profile neighbors”、“Sequence neighbors”、“Links”等工具,可以找到感兴趣的相关数据. Profile neighbors 检索显示为相似类型数据组的其它基因/分子,由此可以推断某些普通功能元件或调控元件. Sequence neighbors 基于核苷酸序列相似性在所有 GEO 数据库寻找相关基因,因此可以用于鉴别同源序列如相关基因家族,或用于物种间对照[23~26]. Links 可以通过 GEO 数据库链接到其他 Entrez 数据库的相关纪录,包括 GenBank, PubMed, Gene, UniGene, OMIM, HomoloGene, Taxonomy, SAGEMap, and MapViewer[27,29].

辅助分析工具

除了 Entrez 查询系统以外,GEO 还提供了几个辅助工具来协助增强数据的挖掘和可视化[30].

4.3.1　Cluster heat maps ——聚类图

大多数据集提供了样本和基因等级聚类图[31,32].用户可以选择浏览这些聚类图,并选择感兴趣的多聚类部分,然后进行放大,下载,做成线性图表或直接链接到 Entrez GEO Profile 记录. GEO 提供 9 种预处理的分层聚类类型和用户确定的 K 均数和 K 中位数聚类(见 Fig. 3). 以 GDS279 为例的数据聚类分析如 Fig. 2 所示:

4.3.2　Query subset A versus B —比较 A 子集和 B 子集的查询

这个特性是通过计算一个数据集内、不同实验子集间的平均秩次或值的差别,来鉴别感兴趣的基因表达谱. 例如:对于数据集 GDS279,如果研究者想定位那些在高脂饮食与低脂饮食的老鼠之间表达高出 3 倍的基因,可以使用此复选框(Fig. 3 所示). 点击“Query A vs B”可以检索到符合条件的图谱:

4.3.3　Subset effects —子集效应

如果不同子集间的基因表达值或秩次存在显著性差异,那么这些表达谱就会被自动标记. 通过这个特性可以检索到所有相关的表达谱. 换句话说,对于某个特别的实验变量,如‘年龄’或‘血型’,一旦出现有意义的表达谱,该表达谱即会被标记.

4.3.4　Value distribution —数值分布

一个数据集中的每个样本均会有对应的箱丝图,可以大概了解一个数据集的数值分布状态.

4.3.5　GEO BLAST —GEO 的序列比对

该界面是通过 BLAST 来搜索感兴趣的,核苷酸序列相似的 GEO 基因表达谱. GEO BLAST 数据库包含了所有 GenBank 中的序列. 而且,是用 NCBI's BLAST 界面输出标准的 BLAST 比对结果,并且在适当的位置显示‘E’图标链接,点击‘E’图标即可直接链接到 GEO Profiles 数据[33].

5　前景与展望

为了支持公共使用和散布基因表达数据,NCBI

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Fig. 2 DataSet cluster analysis

Section of DataSet GDS279 uncentered correlation UPGMA hierarchical cluster analysis. Each column represents an individual sample, or hybridization; each row represents a gene, identified by a GenBank accession number. The light color indicates high expression and the darker color low expression. The dashed box can be moved and resized to select regions of interest, the data for which may be downloaded, or exported to Entrez GEO Profiles

开始了基因表达汇编(GEO)计划. GEO 努力建立一个基因表达数据仓库和在线资源,用于从任何物种或人造的来源检索基因表达数据. 来自于微阵列,高密度寡核苷酸阵列(HAD),杂交膜(filter)和 SAGE 的许多类型的基因表达数据都被接受,登记,和存档,作为一个公共数据集合 GEO 即将增加一系列预处理的数据的定义和描述,以及用于交互检索和分析这些表达数据的在线工具[34].

在提高索引,链接,搜索和显示功能的方面,GEO 资源正处于不断地发展中,以便进行更有效的数据挖掘. 由于 GEO 中存贮的数据来自不同的技术和原始资料,因此它们不一定具有可比性[36]. 基于这点,GEO 把 ProbeSet 限定为包含可比较资料的 GEO 样品的集合. 在 GEO 贮存库中的数据集合到其它 NCBI 资源以前,有必要对进入 ProbeSet 的 GEO 样品进行选择,以及针对这些数据开发有用的显示工具(Fig. 4). 另外,由于 GEO 贮存库的扩充,GEO 现在正发展一个完整的丰度测量数据库,它将支持个体丰度测量结果的查询与检索. 但是,在由复杂性和当前高通量基因表达和染色体组杂交实验的快速发展所带来的限制下,丰度测量仅仅在相似来源的数据组中才具有可比性. GEO 计划开发可比较的数据子集,以便于能够尽可能自由地查询丰度测量结果以及提供关于这些数据的有用概略图. GEO 凭着操作简单、数据全面、免费共享等特点,将在基因表达、数据挖掘、信息推广等中发挥重要作用,为后续研究提供了更好的平台.

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

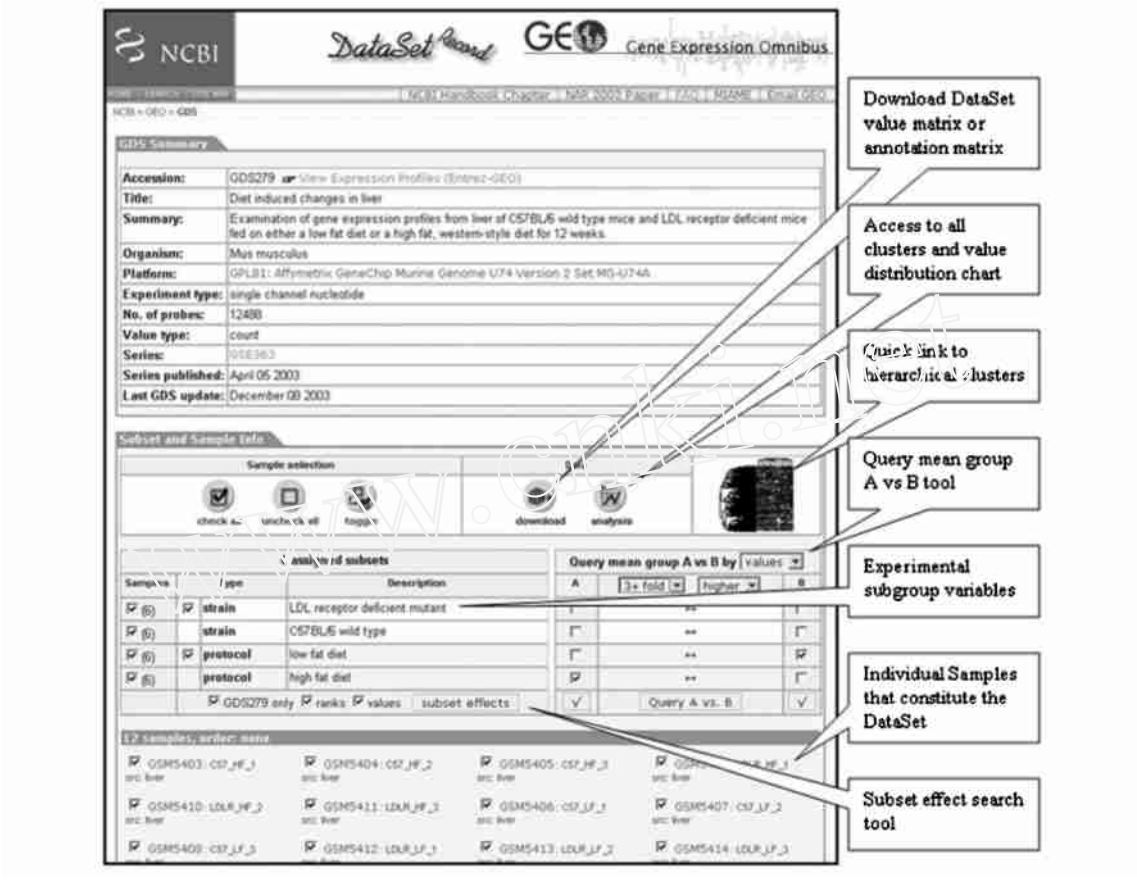


Fig. 3　GEO DataSet record

A screen shot of a typical DataSet record, GDS279, which investigates the effect of a high-fat diet on liver tissue in wild-type and LDL receptor-deficient mice. The locations of the main DataSet features and tools are indicated

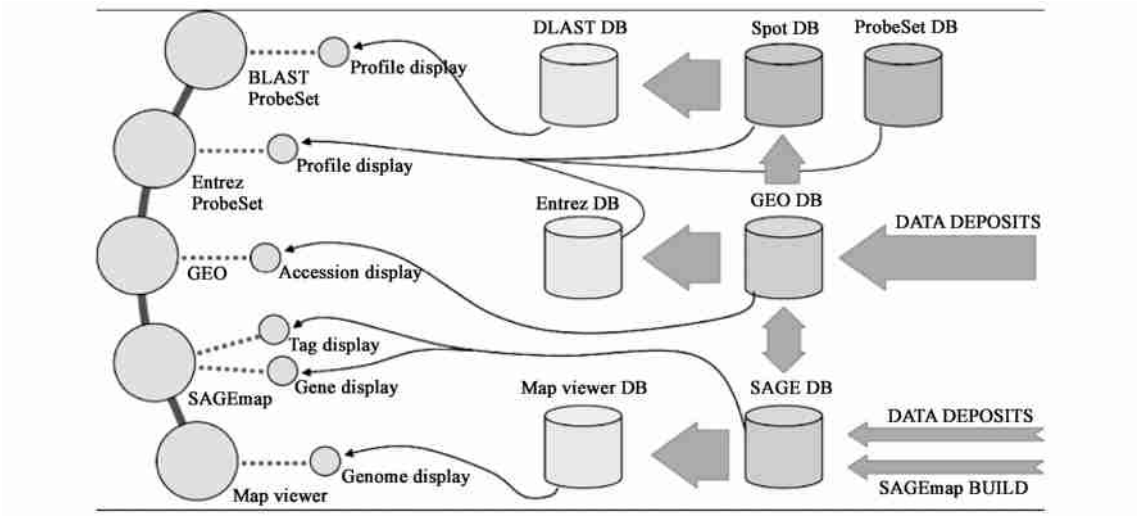


Fig. 4　Constellation of NCBI gene expression resources

Anticipated development of gene expression resources at NCBI is shown. Blue spheres represent Web sites, orange cylinders represent primary NCBI databases, green cylinders represent secondary databases, and yellow cylinders represent tertiary NCBI interface databases. Arrow represent data flow, and lines represent Web site links

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

参考文献(References)

[1] Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data[J]. Nat Genet, 2001, **29**(4): 365-371

[2] Lipshutz R J, Morris D, Chee M, *et al*. Using oligonucleotide probe arrays to access genetic diversity[J]. Biotechniques, 1995, **19**(3): 442-447

[3] Kononen J, Torhorst J, Kallioniemi O P, *et al*. Tissue microarrays for high-throughput molecular profiling of tumor specimens[J]. Nat Med, 1998, **4**(7): 844-847

[4] Chen Q, Qian K, Yan C. Cloning of cDNAs with PDCD2(C) domain and their expression during apoptosis of HEK293T cells[J]. Mol Cell Biochem, 2005, **280**(1-2): 185-191

[5] Lelandais G, Vincens P, Badel-Chagnon A, *et al*. Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms[J]. Bioinformatics, 2006, **22**(11): 1359-1366

[6] Griffith O L, Pleasance E D, Fulton D L, *et al*. Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses[J]. Genomics, 2005, **86**(4): 476-488

[7] Velculescu V E, Zhang L, Vogelstein B, *et al*. Serial analysis of gene expression[J]. Science, 1995, **270**(5235): 484-487

[8] Barrett T, Suzek T O, Troup D B, *et al*. NCBI GEO: mining millions of expression profiles--database and tools[J]. Nucleic Acids Res, 2005, **33**(Database issue): D562-566

[9] Jordan I K, Marino-Ramirez L, Wolf Y I, *et al*. Conservation and coevolution in the scale free-human gene coexpression network[J]. Mol Biol Evol, 2004, **21**(11): 2058-2070

[10] Lee H K, Hsu A, Sajdak J, *et al*. Coexpression analysis of human genes across many microarray data sets[J]. Genome Res, 2004, **14**(6): 1085-1094

[11] Schuler G D, Epstein J A, Ohkawa H, *et al*. Entrez: molecular biology database and retrieval system[J]. Methods Enzymol, 1996, **266**: 141-162

[12] Edgar R, Domrachev M, Lash A E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository[J]. Nucleic Acids Res, 2002, **30**(1): 207-210

[13] Wheeler D L, Church D M, Edgar R, *et al*. Database resources of the National Center for Biotechnology Information: update[J]. Nucleic Acids Res, 2004, **32**(Database issue): D35-40

[14] Shen D, He J, Chang H R. In silico identification of breast cancer genes by combined mutiple high throughput analyses[J]. Int J Mol Med, 2005, **15**(2): 205-212

[15] Gomez-Merino FC, Brearley CA, Ornatowska M, *et al*. a novel diacylglycerol kinase from Arabidopsis thaliana, phosphorylates 1-stearoyl-2-arachidonoyl-sn-glycerol and 1,2-dioleoyl-sn-glycerol and exhibits cold-inducible gene expression[J]. J Biol Chem, 2004, **279**(9): 8230-8241

[16] Puffenberger E G, Hu-Lince D, Parod J M, *et al*. Mapping of sudden infant death with dysgenesis of the testes syndrome (SIDDT) by a SNP genome scan and identification of TSPYL loss of function[J]. Proc Natl Acad Sci USA, 2004, **101**(32): 11689-11694

[17] Zhou X J, Kao M C, Huang H, *et al*. Functional annotation and network reconstruction through cross-platform of integration microarray data[J]. Nat Biotechnol, 2005, **23**(2): 238-243

[18] Calcagno A M, Ludwig J A, Fostel J M, *et al*. Comparison of drug transporter levels in normal colon, colon cancer, and Caco-2 cells: impact on drug disposition and discovery[J]. Mol Pharm, 2006, **3**(1): 87-93

[19] Calvo S, Jain M, Xie X, *et al*. Systematic identification of human mitochondrial disease genes through integrative genomics[J] Nat Genet, 2006, **38**(5): 576-582

[20] Chen Q, Qian K, Yan C. Cloning of cDNAs with PDCD2(C) domain and their expression during apoptosis of HEK293T cells[J]. Mol Cell Biochem, 2005, **280**(1-2): 185-191

[21] Altschul S F, Gish W, Miller W, *et al*. Basic local alignment search tool[J]. J Mol Biol, 1990, **215**(3): 403-410

[22] Jansen B J, Spink A, Saracevic T. Real life, real users, and real needs: A study and analysis of user queries on the web[J]. Information Processing Management, 2000, **36**(2): 207-227

[23] Barrett T, Edgar R. Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*[J]. Methods Mol Biol, 2006, **338**: 175-190

[24] Oue N, Hamai Y, Mitani, *et al*. Gene expression profile of gastric carcinoma: identification of genes and tags potentially involved in invasion, metastasis, and carcinogenesis by serial analysis of gene expression[J]. Cancer Res, 2004, **64**(7): 2397-2405

[25] Cheadle C, Cho-Chung Y S, Becker K G, *et al*. Application of z-score transformation to Affymetrix data[J]. Appl Bioinformatics, 2003, **2**(4): 209-217

[26] Koide T, Vencio R Z, Gomes S L. Global gene expression analysis of the heat shock response in the phytopathogen Xylella fastidiosa[J]. J Bacteriol, 2006, **188**(16): 5821-5830

[27] Wheeler D L, Barrett T, Benson D A, *et al*. Database resources of the National Center for Biotechnology Information[J]. Nucleic Acids Res, 2005, **33**(Database issue): D39-45

[28] Schena M, Shalon D, Davis R W, *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray[J]. Science, 1995, **270**(5235): 467-470

[29] Boyle J. SeqExpress: desktop analysis and visualisation tool for gene expression experiments[J]. Bioinformatics, 2004, **20**(10): 1649-1650

[30] Peters T A. The history and development of Log transaction analysis[J]. Library Hi Tech, 1993, **42**(11): 41-66

[31] Schuler G D, Epstein J A, Ohkawa H, *et al*. Entrez: molecular biology database and retrieval system[J] Methods Enzymol, 1996, **266**: 141-162

[32] Ott S, Hansen A, Kim S Y, *et al*. Superiority of network motifs over optimal networks and an application to the revelation of gene network evolution[J]. Bioinformatics, 2005, **21**(2): 227-238

[33] Zhou X J, Kao M C, Huang H, *et al*. Functional annotation and network reconstruction through cross-platform integration of microarray data[J]. Nat Biotechnol, 2005, **23**(2): 238-243

[34] Bassett D E Jr, Eisen M B, Boguski M S, *et al*. Gene expression informatics—it's all in your mine[J]. Nat Genet, 1999, **21**(1 Suppl): 51-55

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net