

综述

基因表达谱芯片的数据挖掘*

尤元海 张建中**

(中国疾病预防控制中心传染病预防控制所传染病诊断室)

摘要 随着基因芯片技术的迅速发展,表达谱芯片分析及 aCGH 等方法已被广泛应用于生命科学各个研究领域,由此产生的数据也呈指数级增长。如何从海量数据中获取有生物学意义的结果成为摆在生物学工作者面前的难题。对表达谱芯片数据挖掘方法进行了综述。介绍了基本分析思路,当前重点分析方向,如 GO 分析、pathway 与调控网络分析、聚类分析等计算法则和相关几款易用的分析软件。并介绍了几种科学自由计算软件在表达谱生物信息学分析中的应用。藉此为从事芯片分析的研究人员提供参考。

关键词 基因芯片 表达谱分析 数据挖掘

中图分类号 Q819

基因芯片是近二十年分子生物学领域发展起来的革命性技术之一,以其高通量快速并行的特点加快了生命科学研究的步伐。近几年随着表达谱芯片技术的日趋成熟,大量研究结果产生了海量的生物学数据,怎样从这些纷繁的基因表达数据中读懂其中蕴含的生物学意义成为摆在生物学工作者面前的一个新的难题。近几年芯片数据挖掘的一些新方法新思路在一定程度上降低了这项工作的复杂性,要很好地理解数据,不仅要明确研究目的和背景,掌握一些具有多种功能的生物统计分析软件也是必不可少的。目前针对基因表达谱芯片数据分析开发的在线和离线程序有几十种之多,对数据分析方法的报道也有很多,主要集中在数据前处理和聚类分析、判别分析方法方面,侧重于原理的介绍,而大多数从事芯片研究人员面对的一系列问题是:数据分析从何入手,哪些分析是必要的,有哪些易用的开源的程序可供使用,怎样能够快速选择并掌握合适的分析手段。本文拟针对以上问题就目前常用的表达谱芯片数据分析方法及各自特点作一综述,并结合相关研究心得介绍几款易用的软件,希望能够为芯片分析工作提供参考。

1 表达谱数据分析思路

尽管不同类型商业芯片的设计和检测方法有所不同,但所得结果的形式都大体相同,即为一定数量的差异基因列表,这些差异基因即为与实验处理因素相关的基因,基因数量往往有成百上千个。合理的芯片分析策略一般在实验前需要明确,一种思路是从总体上宏观地概括抽取信息,如不同样本间、不同时间点间全部差异基因的 GO 分析,从 GO 分类结果找到实验涉及的显著功能类别;将差异基因映射到通路,根据基因在通路中的位置及表达水平的变化算出受影响显著的通路;聚类分析找出共表达模块,从而预测未知的基因功能等。对于时间序列数据还可以构建新的调控网络。另一种思路是根据研究背景及文献挖掘结果选取感兴趣的部分深入分析在本次实验中的表达及功能变化,这样会更容易得到比宏观分析具有更重要意义的结果。对于表达谱数据分析通常没有一种通用的方法或软件适合于所有数据,就所得结果的可读性来讲一些商业软件展示界面更美观、功能更齐全,但往往需要较高的版权费用。如 Genespring, Pathway Architect, Pathway Assist。每种软件都有其各自的优势特点,所以分析之前要对各类软件加以了解,根据需要选择合适的软件,或者选择几种软件交互使用。无

收稿日期:2009-05-14 修回日期:2009-08-17

* 科技部社会公益项目(2004DIB2J065)

**通讯作者,电子信箱:zhangjianzhong@icdc.cn

论用哪种方法,都应该紧密结合研究背景分析数据而不是仅仅基于纯数学的分析。

2 表达谱数据的 GO (Gene Ontology) 分析

早期基因芯片差异基因分析面临的主要问题是如何与已知的生物学知识相结合以及从哪里获取这些生物学知识。人们首先想到的是结合生物学通路信息,但通路数据库如 KEGG, SWISS-PROT 所包含信息非常分散,缺乏系统的组织结构化。本体学概念的引入为基因功能数据挖掘提供了新的思路,一套本体实际上是一套词汇表,一套基因本体 (Gene Ontology, GO) 也就是一套与基因有关的树状的词汇表^[1]。GO 数据库目前主要由 GO 研究所维护,是一个用于生物功能注释术语分类的开放资源。由于近年来分子生物学的快速发展,有关基因功能注释的信息也在飞速增长,为了便于管理查询和进行基因功能的分析,GO 数据库综合了包括生物学进程、分子功能、细胞组分三个类别的基因本体术语分类,对不同信息源的信息进行整合、统一和标准化,以 DAG (定向非循环表) 结构组织起来,每个父节点下包含若干子节点,子节点可以作为下一层级的父节点进一步展开,分支延伸越远,展开越详细,匹配的 GO 条目就越具体。在这个层级结构中,一个生物学注释可以由一个基因集合表示,层级结构中不同的等级水平的条目具有不同数量的基因集合。在 GO 数据库及其系列分析程序问世之前,差异基因的功能分析是非常繁复费时的的工作,研究者需要花费数月时间检索大量以往的相关文献来分析与基因相关的功能,GO 分析不仅可以使这一工作在数分钟内完成,而且结果也更加准确可信,大大降低了假阳性的发生。目前已有多种免费 GO 分析工具可用,如 AmiGO (<http://amigo.geneontology.org>),既可以搜索某个基因相关的 GO 术语,也可以检索某个术语相关的全部基因。其它的还包括 Gostat, Gominer, Onto express, DAVID, Fatigo 等^[2-5]。差异基因 GO 分析的关键是用统计学方法进行基因富集,分析这些基因参与了何种生物学功能、生物进程以及亚细胞定位,目前常用的基因富集法是基于超几何分布,用 Fisher 精确检验或卡方检验完成的。Fisher 精确概率检验适于小样本量的计算 (小于 5),对于大样本计算 (大于 5) 卡方检验更为快速准确。算法详见参考文献^[6]。下面以 Onto Express 为例说明基因富集算法。Onto Express^[6] 是一个图形界面操作方便结果美观易读的分析软件,以差异基因列表及表达值

作为 input file,以所用芯片上的全部基因作为 reference file,找出差异基因相关的 GO 分类,用卡方检验计算 P 值,检验差异基因中与某功能条目 F 有关的基因是否显著,也就是判断 GO 分布数据是否符合随机分布的标准。GO 分布结果可以分别以饼图和条图的形式展示,还可以 p 值大小排序,便于分析。一般取基因数大于 3,校准 p 值 (corrected p value) < 0.05 的条目作为显著性结果。P 值的生物学意义决定于所提交的基因列表,例如,如果列表中均为上调基因而某功能条目显著,则认为此实验因素作用可能使这个功能激活;相反如果为下调基因并且某条目显著,则认为实验因素作用可能使这个功能抑制。

3 Pathway 分析与调控网络的推导

3.1 pathway 分析

传统的分子生物学研究侧重于生物体单一成分的研究,并未考虑到生物体内部成分间的相互作用和层次性关联,近些年系统生物学在探讨生物系统整体性质功能方面发挥出越来越重要的作用。系统生物学的宏观思路为分析生物体内部多个基因蛋白多层次的非线性相互作用及其复杂的动态网络的发展变化提供了强有力的支持^[7]。系统生物学实际上基于生物通路方面的研究成果。目前较为全面的通路数据库包括 KEGG, BIOCARTEA 等。Kegg (Kyoto encyclopedia of genes and genomes)^[8]是由日本京都大学生物信息中心维护的开放的生物通路数据库。以新陈代谢通路为主。biocarta 主要是信号转导通路,它的一个主要特点是研究者可以任意提交自行绘制的所涉及的通路,biocarta 没有对其准确性作分析验证。GenMAPP 提供了一定数量的生物学通路,并提供了便于分析的图形用户界面,研究者可根据需要绘制通路图。

芯片数据通路分析的第一步是差异基因的通路定位,一些商业软件如 Genespring 可以做到,基于 EASE 算法的开放在线程序 DAVID 也可以实现定位^[9]。目前的通路分析方法还存在很多局限性,例如只注意到基因集合定位到了哪个通路而忽略了其在通路中的位置,如果一个通路由某个基因产物触发或被单个受体激活,并且特定的蛋白没有表达,这个通路就会受到严重影响甚至关闭;相反,如果多个基因与某个通路相关但都只出现在通路的下游,那么其表达水平的变化就可能不会对通路造成很大影响。另外,一些基因往往有多个功能分布于不同的通路发挥不同的作用,要得

到相对准确的结果还必须考虑通路的拓扑结构。目前很少有能将基因差异表达值变化应用于通路分析的方法, Pathway express^[10] 提出了一种基于 IF (impact factor) 的通路分析方法, IF 值的计算基于两个模型③和④, 综合了差异基因的标化的差异表达值、通路中基因的统计学显著性以及信号通路的拓扑学三方面内容。模型具体算法详见参考文献[10]。Pathway express 主要基于 KEGG 库, 结果输出中自动把差异基因以不同颜色定位于通路中, 红色为上调, 蓝色为下调, 这些定位着上调和下调基因的通路图可以在 java 控制台中找到绝对路径, 在浏览器中打开或保存, 也可以 GML 格式导出, 然后直接导入 cytoscape^[11], 用 merge 节点功能把多个相关 pathway 连接起来, 显示互作网络, 并分别以红蓝色显示显著性通路中上调下调的基因(节点), 以及这些基因与其他基因间的相互作用(边), 可以从不同视角观察其位置, 不断放大就可以看到节点的基因名称。

$$PF(g) = \Delta E(g) + \sum_{u \in U_{S_g}} + \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} \quad (3)$$

$$IF(P_i) = \log\left(\frac{1}{P_i}\right) + \frac{\sum_{g \in P_i} |PF(g)|}{|\Delta E| \cdot N_{de}(P_i)} \quad (4)$$

其他的可视化工具还有 pathway studio, genmapp, arraypath, osprey 等。Biolayout 也是一款分子作用网络展示工具, 所不同的是结果为三维图形界面。^[12]

3.2 基因调控网络的推导

调控网络以单个基因为节点, 以其相互作用为边, 如果两节点间存在调控关系则边为有向的。无标度性是大多数生物网络的拓扑特性, 其特点是多数节点拥有少量连接, 而少数节点拥有大量连接, 这些高连接节点即为决定网络整体性质的关键性的节点中枢。

非时间序列数据主要采用高斯图形模型和贝叶斯网络模型进行推导。高斯图形模型算法主要包括总体相关系数的计算、基因与基因间偏相关系数的计算和显著性假设检验。Schafer 等^[13] 针对高斯模型中的假设的可靠性不足的问题改进了求解偏相关系数矩阵的方法, 并引入了贝叶斯模型, 从而提高了模型的可信度。这些模型的缺点是推导出的基因互作网络不具方向性, 而调控网络一般都具有很强方向性, 所以这种非方向性的网络很难给予准确的生物学解释。

时间序列数据主要应用动态贝叶斯概率模型, 由 Friedman^[14] 在 2004 年提出, 用于分析调控网络结构已知、基因表达动力学参数未知的情况, 结合了调控蛋白

浓度变化与靶基因表达速率变化的关联关系来描述, 但基因网络拓扑结构的确定也是一项繁复的工作。DREM (Dynamic Regulatory Events Miner) 是一款用于基因转录调控动力学建模分析的软件。输入时间序列表达谱数据和相关转录因子调控的基因, 计算基于存在交互作用的已注释的数据, 输出一个动态调控图, 图中高亮显示出时间序列中的分叉, 即为时间序列中存在调控作用关系的节点。由于目前对转录因子及其调控基因了解还比较少, 所以这个软件的应用仍受制于 TF reference file 数据的有限性, 尽管如此, DREM 仍为此类自动化分析方法的发展提供了一个很好的模式^[15]。

Pujana 等^[16] 建立了一个用于鉴定乳腺癌相关基因的网络模型, 以 4 个已知的编码乳腺癌肿瘤抑制因子的基因 BRCA1、BRCA2、ATM、CHEK2 为参照, 结合不同物种的功能基因组和蛋白质组的共表达谱数据构建了一个包含 118 个基因、涉及 866 种功能的网络, 这个集成的网络模型提出了一个把网络元件的可能性由低到高分类的分级系统, HMMR 是网络元件之一, 编码中心体亚基, 先前的研究对于这个基因与乳腺癌基因 BRCA1 之间的功能关联仍是未知的, 该研究通过双病例对照研究证明 HMMR 基因增加了乳腺癌发生的风险。

4 聚类分析

聚类分析是表达谱数据分析最常用的方法, 对于预测基因新功能及调控网络的构建具有重要意义。聚类分析用于探索完全未知的数据特征, 属于是无监督的聚类, 也称无监督模式识别, 这类训练样本没有类别标签, 主要用于确定两个特征向量间的相似度及合适的测度, 并选择一个算法方案, 基于选定的相似性测度对向量进行聚类。常用的相似性测度包括欧氏距离 (Euclidean distance)、马氏距离 (Manhattan distance)、明考斯基距离 (Minkowski distance)、皮尔逊相关距离 (Pearson correlation distance) 等, 关于距离的相关计算公式已有许多专业书籍做了详细说明, 这里不再赘述。聚类算法可分为层级聚类 (hierarchical clustering)、k-means 聚类、自组织图、主成分分析等。根据处理对象和目标的不同, 主要分为基因间聚类、样本间聚类和双向聚类。基因间聚类即比较表达矩阵中行与行之间的差异, 表达模式相似的行的集合即为共表达模块, 一般认为相似的表达模式往往代表着相似的基因功能, 从而根据共表达类别中已知基因的功能可以推测未知基因的功能。基因聚类分析中容易受到噪点基因的干

扰,而且要求算法的准确性和有效性较高,目前较常用的方法包括人工神经网络和模糊聚类。人工神经网络以自组织映射(Self-Organizing Map, SOM)为主,它采用的是结构简单的单层竞争性神经网络模式在输入端引入并与输出结点关联,其间的权重通过学习反复变更,直到达到终止标准,结果是相似的模式被分入同组,并为同一个单位神经元所代表。聚类分析软件种类很多,而且大都是开源的,常用的有 Mev, Acuity 等。

STEM(Short Time-serious Expression Miner)是一款专门用于时间序列芯片数据聚类分析开发的程序,最长可分析8个时间点。STEM采用一种新方法可以对短的时间序列表达数据聚类,并对聚类型别的可靠性进行统计推断。这种方法是首先定义了一套独立的有代表性的短时表达模型。这些模型与基因的可能的时序表达变化相匹配,基因表达时间序列被转换为开始于0时间点,每个基因基于其时序变化与模型的相关系数被排列到模型中去,然后计算出被排列到每个模型的基因数量。基因数量排列计算基于以下四步评价方法:起始时间点值的随机变换,基因表达值的重正化,把基因排列到最匹配的模型中,大量重复排序变换。所有排序变换中被分配到模型中的基因平均数用以评价应被分配到模式中的基因数量。接着计算出实际分配到每个模式中的基因数与假设的数目之间的统计学显著性。相似的具有显著统计学意义的模型被集合在一起形成聚类。STEM算法的优点是减少了聚类过程中的噪点干扰,提高了结果的特异性^[17]。应用STEM可以设定不同的模式数量优化聚类的准确度,数量越多聚类效果(均一性)越好,但基因数量也越少。

5 科学自由计算软件在分析中的应用

目前在生物芯片数据分析领域常用的科学计算软件主要包括 Matlab(<http://www.mathworks.cn>), Scilab(<http://www.scilab.org>),以及运行于R语言环境下的 bioconductor。

MATLAB 是 MATrix LABoratory 的缩写,早期主要用于现代控制中复杂的矩阵、向量的各种运算。MATLAB 提供了强大的矩阵处理和绘图功能,2006 年发布了主要针对蛋白质组和基因组分析的 bioinformatics tool box,利用该软件包可以方便地调用 clusters 函数完成 microarray 的聚类分析,用 clustergram 函数可以从层级聚类结果中生成热图和树状图。也可以根据个人需要编写新的程序。MATLAB 几乎囊括了目前所有科学计算分析方法。

SCILAB 一词来源于英文“Scientific Laboratory”词头的合并,是由法国国家信息、自动化研究院(INRIA)的科学家们开发的开源软件。与 MATLAB 类似,SCILAB 也是一种科学工程计算软件,其数据类型丰富,可以很方便地实现各种矩阵运算与图形显示,能应用于科学计算、数学建模、信号处理、决策优化、线性/非线性控制等各个方面。它还提供可以满足不同工程与科学需要的工具箱,例如 SCICOS,信号处理工具箱,图与网络工具箱等。可以说,就基本的功能如科学计算、矩阵处理及图形显示而言, MATLAB 能完成的工作 SCILAB 都可以实现。SCILAB 是只有几十兆大小,对系统配置要求不高,而 MATLAB 为 1G 以上,所以运行简单计算时用 matlab 就显得比较笨拙,这时用 scilab 更为方便。另外,SCILAB 提供的语言转换函数可以自动将用 MATLAB 语言编写的程序翻译为 SCILAB 语言,实现了与 MATLAB 的交互使用。

R 语言的语法形式与 S 语言基本相同(<http://www.r-project.org/>),是一个自由开源的软件,目前版本是 2.7.1, R 语言的实现主要是通过其自带的三百个扩展包, bioconductor 即是 R 语言环境开发的一个包含许多生物信息工具包的集成开放软件,主要针对基因芯片数据的管理和分析,可以方便地实现数据的可视化,绘出直方图(hist 命令)、盒图(boxplot 命令)、散点图(plot 命令)等^[18]。Stat、sma、cluster、class 包提供了用于聚类分析的函数。Annotate 包实现基因功能注释, goTools 包实现基因本体的图形化分析。

6 存在的问题与展望

系统生物学的理念正逐渐使生命科学的研究由分子水平向系统水平转变,但这方面的研究还处于初级发展阶段,已开发的软件和数据库都是基于已知的对于生物网络的了解,相对于整个生物网络的复杂性,还差得很远,很多未知的机制和功能有待深入全面的研究。对系统生物学知识的了解越深入,生物信息学的分析和预测结果准确性也就越高。

芯片数据分析面临的问题是如何从有限的数据结果挖掘出尽可能多的有意义的信息,这样就对研究人员知识结构提出了较高要求,要能够把数学、统计学、计算机科学与生物学、医学有机结合起来进行综合分析,而这种分析如果只是把计算统计学者和生物医学学者的各自专业的机械结合,往往得不到理想的结果,生物医学工作者注重提高统计计算方面的技术对于深入理解数据是非常重要的。虽然目前已有多种芯

片分析相关软件问世,而且 matlab 和 R 语言包这样的多功能集成软件在芯片数据分析中已得到广泛应用,但对于时间序列芯片数据,还没有一个实用程序能够进行动态网络和 pathway 显著性分析,且现有模型的应用具有很多局限性,一些多功能的更适用于非计算生物专业的集成应用软件将是未来智能系统生物学研究的主要方向之一。

在未来一段时间内,基因芯片高通量分析手段仍是组学研究中的利器,随着基因芯片方法学上的进一步完善,芯片数据将更加准确,下一步的主要任务将是完善下游对数据的系统科学的分析以及生物学意义的提取,这也是基因芯片分析的根本目的。

参考文献

- [1] Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology. *Nature Genet*,2000,25:25 ~ 29
- [2] Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*,2004, 20(4):578 ~ 580
- [3] Beissbarth T, Speed T P. Gostat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 2004,20(9):1464 ~ 1465
- [4] Draghici S, Khatri P, Bhavsar P, et al. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*, 2003, 31(13):3775 ~ 3781
- [5] Zeeberg B R, Feng W, Wang G, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 2003,4(4):R28
- [6] Khatri P, Bhavsar P, Bawa G. Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res*, 2004, 1(32):W449 ~ W456
- [7] 谭璐,姜璐. 系统生物学与生物网络研究. 复杂系统与复杂性科学, 2005,2(4):1 ~ 9
Tan L, Jiang L. *Complex Systems and Complexity Science*,2005,2(4):1 ~ 9
- [8] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*,2000,28:27 ~ 30
- [9] Dennis, G Jr, Sherman B T, Hosack D A, et al. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol*,2003,4(5):P3
- [10] Draghici S, Khatri P, Tarca A L, et al. A systems biology approach for pathway level analysis. *Genome Res*, 2007, 17(10): 1537 ~ 1545
- [11] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 2003, 13(11):2498 ~ 2504
- [12] Freeman T C, Goldovsky L, Brosch M, et al. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol*, 2007, 3(10): 2032 ~ 2042
- [13] Schafer J, Strimmer K. An empirical approach to inferring large graphical Gaussian models from microarray data. *Bioinformatics*, 2004, 21(6):754 ~ 764
- [14] Friedman N. Inferring cellular networks using probabilistic graphical models. *Science*,2004,303(5659):799 ~ 805
- [15] Ernst J, Vainas O, Harbison C T, et al. Reconstructing dynamic regulatory maps. *Mol Syst Biol*, 2007,3:74
- [16] Pujana M A, Han J D, Starita L M, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet*,2007,39(11):1338 ~ 1349
- [17] Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*,2006,7:191
- [18] 孙啸,谢建明,周庆. R 语言及 bioconductor 在基因组分析中的应用 北京 科学出版社 2006. 1 ~ 7
Sun X, Xie J M, Zhou Q. *The Application of R language and Bioconductor in Genome Analyzation*. Beijing: Science Press, 2006. 1 ~ 7

Data Mining from Microarray Gene Expression Profile

YOU Yuan-hai ZHANG Jian-zhong

(Department of Diagnosis, Institute of Communicable Disease Control and Prevention, Chinese Centre for Disease Control and Prevention, P. O. Box. 5, Changping, Beijing 102206, China)

Abstract Microarray technology are being performed more widely than ever before on many areas in lifescience, although the technology is still evolving, the challenge of performing a microarray experiment is no longer in the data generation, but in extracting useful information and utilizing it to get the results with biological meanings. Some methods and tools used for expressional microarray data mining based on previous work were summarized. These methods include gene clustering, GO analysis, regulating pathway analysis, and related algorithm. We hope this can be helpful for those researchers who are implementing expressional microarray for biological analysis.

Key words Microarray Expressional analysis Data mining