doi:10.3969/j.issn.1672-5565.2014.02.10

生物信息学基因表达差异分析

卢 汀
(中国科学院水生生物研究所,湖北 武汉 430072)

摘 要:基因的差异化表达由多种因素共同导致,并且与许多疾病的发生和发展有密切联系,对差异化表达的基因进行生物信息学以及生物统计学的分析对于研究细胞调节机制和疾病机理有着重要意义。目前,对差异化表达的基因有以下几种主流的研究方法:DNA 微阵列 (DNA microarray),抑制性消减杂交 (SSH),基因表达连续性分析 (SAGE),代表性差异分析 (RDA),以及 mRNA 差异显示 PCR (mRNA DDRT-PCR)。目前许多基因差异化表达数据是建立在时段(time series)基础上,因此对基于时间变化的基因差异化表达分析变得尤为重要。本文将对差异化表达基因的几种主流方法进行详细阐述,并介绍一种基于傅里叶函数的时段基因差异化表达分析。

关键词:生物信息学;基因;差异化表达;时段
中图分类号:Q344+.13 **文献标志码**:A **文章编号**:1672-5565(2014)-02-140-05

Bioinformatics analysis for gene differential expression

LU Ting
(*Institute of Hydrobiology*, *Chinese Academy of Sciences*, *Wuhan* 430072, *China*)

Abstract: Gene differential expression can be caused by multiple factors, and related with genesis and development of many diseases. Bioinformatics and biostatistics analysis for gene differential expression are widely used for studying cellular regulation mechanism and disease mechanism. Currently, there are several main methods for studying gene differential expression, DNA microarray, suppression subtractive hybridization (SSH), serial analysis of gene expression (SAGE), representational difference analysis (RDA), and mRNA differential display PCR (mRNA DDRT-PCR). Nowadays, much gene differential expression data is time series based, therefore the analysis for time series based gene differential expression data is critical. This review will elucidate several main methods for gene differential expression, and introduce a Fourier algorithm-based bioinformatics analysis for time-series gene differential expression.

Keywords: Bioinformatics; Gene; Differential expression; Time-series

2006年,随着美英科学家宣布完成了人类1号染色体的基因测序图,历时16年的浩大工程——人类基因组计划终于划上了句号。这个与曼哈顿工程和阿波罗登月计划并称20世纪人类三大科学项目的计划对人类了解疾病以及治疗疾病带来了深远的影响,从此,基因诊断,基因治疗等全新的概念走上了历史的舞台。与此相伴的是其他一系列物种的基因组计划相继完成。但是紧接着我们发现基因组信息仅仅是遗传信息发挥作用的第一个层次,基因在不同生物体以及生物体不同状态下,比如发育,交配,衰老等,产生的不同程度的表达对生物体的生命活动则产生了更直接的影响。因此,对差异化表达的基因进行数据分析就显得尤为重要。基因差异化表达包括基因种类的变化和基因表达量的变化。传统的基因表达差异分析主要是对 mRNA 的丰度以及有无进行分析比较。现代生物信息学分析方法则包括 DNA 微阵列 (DNA microarray),抑制性消减杂交 (SSH),基因表达连续性分析 (SAGE),代表性差异分析 (RDA),以及 mRNA 差异显示 PCR (mRNA DDRT-PCR)。基于时间段的基因差异化表

收稿日期:2013-11-06;**修回日期**:2014-01-05.
作者简介:卢汀,男,生物信息学研究助理,研究方向:生物信息学和统计学;E-mail:luting_bioinfo@163.com.

达数据通常用动态生物系统和基因调控网络进行分析,微阵列技术使得对高通量基因组差异化表达分析成为可能[1]。

1　DNA 微阵列 (DNA microarray)

DNA 微阵列 (DNA microarray)技术又称为基因芯片技术 (Gene chips)。在 DNA 芯片上,一段由寡聚核苷酸或者 cDNA 构成的脱氧核苷酸序列被固定在片基上,待测样本中的 RNA 被提取出来并进行反转录染色并与片基上的基因序列进行分子杂交 (Molecular hybridization),待测样本中的由 RNA 反转录得来的 cDNA 会和片基上固定的互补 DNA 特异性结合。杂交后的基因芯片经洗脱出去未杂交的 cDNA 片段,并进行荧光检测,便通过荧光强度平判断待测样本中 RNA 的表达水平[2]。基因芯片技术由芯片的制备,杂交,和检测三方面组成。基因芯片技术是半导体产业与分子生物学的结合。由于在极小的基片 (玻片,硅片,尼龙膜)上集成了大量的寡核苷酸或者 cDNA,一次性对高通量 (High-throughput)的基因进行分析成为可能,克服了传统技术操作复杂,自动化程度低,检测序列少的问题。基因芯片技术被广泛应用于生物医学科研,疾病诊断等各个方面,对于治疗水平的提高有着重大而深远的意义[3]。许多研究者目前已经将 DNA 芯片技术用于检测基于时序 (Time series)的基因表达状况[4-6]。目前储存 Microarray 数据的公共数据库包括 NCBI 的 Gene Expression Omnibus (GEO)[7]. EBI 的 ArrayExpress[8] 和 Stanford Microarray Database (SMD)[9]。在这些数据库中储存了大量基于时序的 DNA microarray 数据。

2　抑制性消减杂交 (SSH)

抑制性消减杂交起源于代表性差异分析法 (Representational difference analysis, RDA)。RDA 是一种以杂交为基础的研究基因组之间差异的方法。1996 年,Diatchenko L 等[10]将“抑制性 PCR 理论”[11]与 RDA 相结合,建立了一种旨在分离差异表达基因的方法。该方法的原理为杂交二极动力学:高丰度单链 cDNA 在退火时产生同源杂交速度快于低丰度的单链 cDNA,这样就可以使原来在丰度上有差异的单链 cDNA 相对含量达到基本一致。而抑制 PCR 的原理则是利用链内退火优先于链间退火的规律是非目的的序列片段 3’端反向重复序列在退火时产生类似发卡的互补结构,从而导致无法与引物配对,这样就可以选择性的抑制了非目的基因片段的扩增。SSH 将需要检测的细胞称为“检测子”,将对照细胞 mRNA 称为“驱赶子”,当 mRNA 合成 cDNA 后,经限制性内切酶消化成为两份,然后连上不同的接头,然后进行两轮杂交反应。两轮杂交后的样品经末端补平后,用一对与接头外测序列对应的巢式引物进行第二次 PCR 扩增,产物直接用于差减 cDNA 文库构建。通过 SSH 可以得到某种细胞相对其他组织的差异基因的全面信息,较好的解决了低丰度基因难以得到的问题,甚至可以得到一些传统方法没有得到的新基因[12-13]。SSH 技术的优势主要表现在以下方面:不同丰度的 mRNA 分子能够趋于一致,从而提高了差减的灵敏度,克服了不同 mRNA 分子拷贝数目不同对杂交结果的影响。通过一轮差减杂交就可以对差异表达的 cDNA 分子实现 1 000 多倍的富集[14]。一次 SSH 就可以分离到多达几百个的差异表达的基因,阳性率高达 94%[11]。

3　基因代表连续性分析 (SAGE)

1995 年,Velculescu VE 等[15]建立了一种高效快速研究基因表达的方法,称为基因代表连续性分析 (SAGE),主要理论依据有两个:(1) 一个短寡核苷酸序列(9~11 bp) (来自转录物内特定位置)含有足够多的鉴定一个转录物特异性的信息,可以作为区别转录物的标签。(2) 这些短寡核苷酸序列式可以串联在一起的。形成大量多联体 (Concatemer),然后,这些克隆到载体的多联体被测序并使用 SAGE 专用软件进行分析,基因表达种类就可以被确定下来,同时,根据标签的数量还可以确定基因的表达丰度。尽管如此,SAGE 技术本身还是存在一个缺陷:SAGE 标签对于识别基因来源通常太短了(14 个碱基)[16]。但是这个问题可以通过延伸 SAGE 标签到 3’cDNA 来解决[17-18]。改进后的技术可以将标签的长度延展到 21 至 26 个碱基[16,19]。即便如此,标签的长度依然在基因识别的过程中产生了一些问题[20-21],LongSAGE 标签与已知的数据库中的任何表达序列不相吻合[22-23]。另一方面,同样的标签常常与两个或者更多的基因序列吻合,使得进一步的分析变得困难[24]。

在 SAGE 技术中,生物素标记的 bio-Oligo(dT)为引物合成双链 cDNA,然后用限制酶酶切,捕获 3’cDNA。在此,产物被分为两部分,分别与包含有 iIS 型内切酶位点的 a, b 连接子连接。iIS 型内切酶的作用位点处于识别位点以外。经过酶切得到 9~

10 bp 的标签序列。以每两个标签的钝端结合后产生的 PCR 模板为基础,以基于 a, b 连接子的引物进行 PCR 反应,得到大量包含两个不同来源标签的序列,然后再进行酶切和连接,就能把多个不同的标签连接在一起,克隆至质粒载体后集中测序[25]。SAGE 的结果最后通过统计学处理得到,根据标签出现频率的高低来判断所属基因表达的丰度。

4 代表性差异分析 (RDA)

代表性差异分析 (Representational difference analysis, RDA)是一种对微阵列方法进行补充的技术,通常用于识别在测试样本和对照样本中差异表达的基因[26-27]。这种技术后来被进一步发展成为 ROMA 技术 (representation oligonucleotide microarray analysis, ROMA)。ROMA 技术用阵列技术来进行类似的分析。这种技术可以用来探测 DNA 甲基化的差异[28]。这种技术依赖于 PCR 技术来对在两个几乎相同的 DNA 类别 (被称为"driver"和"tester" DNA)消化得来的非同源 DNA 片段进行扩增。通常来说"tester" DNA 包含了一段与"driver" DNA 非同源的序列。当两个 DNA 类别被混合时,driver 序列被过量加入 tester 序列。在 PCR 过程中,双链 DNA 片段在大约 95℃变性,然后在退火温度退火。因为 driver 和 tester 序列几乎是相同的,过量的 driver DNA 会和同源的 tester DNA 片段结合。这样就阻碍了 PCR 扩增,因此就没有同源片段的增加。然而,与 driver 和 tester 序列不同的片段就会扩增。

5 mRNA 差异显示 PCR (mRNA DDRT-PCR)

DDRT-PCR 技术是目前在筛选和克隆差异表达基因方面最有效的方法[29]。1992 年,位于美国波士顿的 Dena-Farber 癌症研究所的 Liang. P 和 Pardee.AD 创立了该技术。DDRT-PCR 技术的最大优势是简便,它把 poly-A RNA 逆转录技术,多聚酶链式反应 (PCR)和聚丙烯酰胺凝胶电泳 (PAGE)技术结合,这样我们就能看到并比较 2 个样本或者更多样本之间的基因表达谱[30]。在真核细胞的 mRNA 3' 端有一个长度为 30 ~ 300 的多聚苷酸 (PolyA)尾巴,与 3' 相连接的两个碱基有 12 种组合: TT, GG, CC, AT, CT, TC, TG, GC, CG, AC, GT, AG。总 mRNA 被从样品中提取出来,在逆转录酶作用下启动 mRNA 反转录,合成 cDNA 模板链,以该模板链为基础,加入 TaqDNA 聚合酶,dNTP, 5' 和 3' 引物进行 PCR 扩增,扩增后的 cDNA 用变性或非变性的聚丙烯酰胺凝胶电泳分离差异片段,如果在一种细胞型中的总 mRNA 扩增出某条别的细胞型没有扩增出来的带,就可以认为这条带代表的基因是特异性表达的,一旦找到这些特异性带,就可以从 cDNA 文库中筛选出相应的基因[31-32]。

虽然这项技术有着操作简单,灵敏度高,效率高等优点[32],但是也有一些缺陷:产生假阳性条带;对高拷贝的 mRNA 有很强的倾向性;片段较短,不含编码信息;安全系数低;易造成污染等[33]。针对这些缺陷, 研究者对该技术进行了以下一些改进:(1) cDNA 合成底物由 mRNA 改为总 RNA, 可避免 poly(dT)柱纯化 mRNA 时的污染[30];(2) 3' 引物由 OligoT11MN 简化为 OligodT12M,由 12 种减少到 3 种,进一步利用单碱基把 Oligo(dT)引物锚定于 poly-A 尾的起始端,同时在锚定引物和随机引物的 5' 引入限制酶切位点[34];(3)荧光标记电泳条带[35];(4) DNase 处理 RNA 样品已消除 DNA 的可能污染[36-37];(5)提高差异阳性带验证的可靠性[38-44]。

6 基于傅里叶函数的时序(Time series)数据分析

傅里叶分析是数学分析中的一个重要分支, 主要分析傅里叶函数变换及其性质。研究领域由最初的直线群,圆周群扩展到了一般的抽象群。最近一系列研究都集中在了基于傅里叶系数对基于时序的基因表达数据的分析上。这些数据都是通过收集在不同时段的 DNA 芯片实验数据得到的。研究人员通过使用基于傅里叶系数(Fourier coefficients)控制的 FDR 和基于模型的群扫描识别在时段上差异表达的基因,并且与 GP 扫描技术进行了比较。最近光谱混合内核引入高斯混合的傅里叶变换内核,这些内核能够发现模式和推断和模型的负协方差,说明 GP 和傅立叶方法之间的关系。基于傅里叶函数的方法可以识别基于时序的差异化表达基因,并且可以找到有相同生物进程特点的基因,对于将来的数据分析有重要意义。

7 结论和展望

通过对以上几种基因差异化表达的方法进行比较,我们可以发现他们各自具有各自的特点,各自有各自的优缺点。随着技术的发展,DDRT-PCR 技术已经逐渐成为分离,克隆和分析差异表达基因的主流技术,这项技术由最初的医学研究领域逐渐扩展

到其他领域,包括高等植物研究,随着 DDRT-PCR 技术的不断完善,在分子水平上研究园马铃薯,西红柿,豌豆,草莓,苹果和芒果生长发育,成熟衰老以及抗病抗虫等等问题都可以应用该技术[44]。

参考文献(References)

[1] JAEHEE K, ROBERT T O, HASEONG K. A method to identify differential expression profiles of time-course gene data with Fourier transformation [J]. BMC Bioinformatics, 2013, 14:310.

[2] STEARS R L, MARTINSKY T, SCHENA M. Trends in microarray analysis [J]. Nat Med, 2003, 9(1): 140-145.

[3] KAN T, SHIMADA Y, SATO F, et al. Gene expression profiling in human esophageal cancers using cDNA microarray [J]. Biochem Biophys Res Commun, 2001, 286(4): 792-801.

[4] HENDRIKSEN P J, DITS N F, KOKAME K, et al. Evolution of the androgen receptor pathway during progression of prostate cancer [J]. Cancer Res, 2006, 66(10):5012-5020.

[5] BAGHDOYAN S, LAMARTINE J, CASTEL D, et al. Id2 reverses cell cycle arrest induced by {gamma}-irritation in human HaCaT keratinocytes [J]. J Biol Chem, 2005, 280(16):15836-15841.

[6] CHO R J, HUANG M, CAMPBELL M J, et al. Biological methods for cell-cycle synchronization of mammalian cells [J].Biotechniques, 2001, 30(6):1322-1326,1328,1330-1321.

[7] EDGAR R, BARRETT T. NCBI GEO standards and services for microarray data [J]. Nat Biotechnol, 2006, 24(12):1471-1472.

[8] PARKINSON H, KAPUSHESKY M, SHOJATALAB M, et al. ArrayExpress-a public database of microarray experiments and gene expression profiles [J]. Nucleic Acids Res, 2007, 35(Database issue): D747-750.

[9] DEMETER J, BEAUHEIM C, GOLLUB J, et al. The stanford microarray database: Implementation of new analysis tools and open source release of software [J]. Nucleic Acids Res, 2007, 35(Database issue): D766-770.

[10] DIATCHENKO L, LAU Y F, CAMPBELL A P, et al. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries [J]. Proc Natl Acad Sci USA, 1996, 93: 6025-6030.

[11] SIEBERT P D, CHENCHIK A, KELLOGG D E, et al. An improved PCR method for walking in uncloned genomic DNA [J]. Nucleic Acids Res, 1995, 23: 1087-1088.

[12] VOSTEIN O D, THIES W G, HOFMANN M. A high throughput screening for rarely transcribed differentially expressed genes [J]. Nucleic Acids Res, 1997, 25:2598-2602.

[13] KUANG W W, THOMPSON D A, HOCH R V, et al. Differential screening and suppression subtractive hybridization identified genes differentially expressed in an estrogen receptor-positive breast carcinoma cell line [J]. Nucleic Acids Res, 1998, 26:1116-1123.

[14] TCHERNITSA O I, ZUBER J, SERS C, et al. Gene expression profiling of fibroblasts resistant toward oncogene-mediated transformation reveals preferential transcription of negative growth regulators [J]. Oncogene, 1999, 23, 18(39):5448-54.

[15] VELCULESCU V E, ZHANG L, VOGELSTEIN B, et al. Serial analysis of gene expression [J].Science, 1995, 20, 270(5235):484-7.

[16] MATSUMURA H, REUTER M, KRÜGER D H, et al. SuperSAGE [J]. Methods Mol Biol, 2008, 387:55-70.

[17] MATSUMURA H, REICH S, ITO A, et al. Gene expression analysis of plant host-pathogen interactions bySuperSAGE [J]. Proc Natl Acad Sci USA, 2003, 100: 15718-15723.

[18] SAHA S, SPARKS A B, RAGO C, et al. Using the transcriptome to annotate the genome [J]. Nat Biotechnol, 2002, 20:508-512.

[19] WAHL M B, HEINZMANN U, IMAI K. LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse [J]. Bioinformatics, 2005, 21:1393-1400.

[20] ZHANG L, ZHOU W, VELCULESCU V E, et al. Gene expression profiles in normal and cancer cells [J]. Science, 1997, 276:1268-1272.

[21] LEE S, CHEN Jianjun, ZHOU Guolin, et al. Generation of high quality and quantity of tag/ditag for SAGE analysis [J]. BioTechniques, 2001, 31:348-354.

[22] WAHL M, SHUKUNAMI C, HEINZMANN U, et al. Transcriptome analysis of early chondrogenesis in ATDC5 cells induced by bone morphogenetic protein 4 [J]. Genomics, 2004, 83:45-58.

[23] SIDDIQUI A S, KHATTRA J, DELANEY A D, et al. Large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells [J]. Proc Natl Acad Sci USA, 2005, 102:18485-18490.

[24] CHEN Jianjun, LEE S, ZHOU Guolin, et al. High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequence into 3' complementary DNAs [J]. Genes Chromosomes Cancer, 2002, 33(3):252-261.

[25] VELCULESCU V E, ZHANG L, ZHOU W, et al. Char-

acterization of the yeast transcription [J]. Cell, 1997, 88 (2): 243-251.

[26] HUBANK M, SCHARZ D G. Identifying dieferences in mRNA expression by representational analysis of cDNA [J]. Nucleic Acids Res, 1994, 22: 5640-5648.

[27] HUBANK M, SCHATZ D G. cDNA representational difference analysis: a sensitive and flexible method for identification of differentially expressed genes [J]. Methods Enzymol, 1999, 303: 325-349.

[28] LISITSYN N, WIGLER M. Cloning the differences between two complex genomes [J]. Science, 1993, 259: 946-951.

[29] MATZ M V, LUKYANOV S L. Different strategies of differential display: areas of application [J]. Nucleic Acids Research, 1998, 26(24): 5537-5543.

[30] STEIN J, LIANG P. Differential display technology: a general guide [J]. Cellular Molecular Life Sciences, 2002, 59: 1235-1240.

[31] XIA T, JIANG Y H, ZHOU H T, et al. Progress in mRNA differential display [J]. Development and Reproductive Biology, 1996, 5(1): 60-72.

[32] LIEVENS S, GOORMACHTIG S, HOLSTERS M A. A critical evaluation of differential display as a tool to identify genes involved in legume nodulation: looking back and looking forward [J]. Nucleic Acids Research, 2001, 29(17): 3459-3468.

[33] FROST M R, GUGGENHEIM J A. Mammalianpolyadenylation sites: implications for differential display [J]. Nucleic Acids Research, 1999, 27(5): 1368-1391.

[34] VON D K H, ALBRECHT C, MAYHAUS M, et al. Identification of genes regulated by muscarinic acetylcholine receptors: application of an improved and statistically comprehensive mRNA differential display technique [J]. Nucleic Acids Research, 1999, 27(10): 2211-2218.

[35] CHEN J J W, PECK K. Non-radioisotopic differential display method to directly visualize and amplify differential bands on nylon membrane [J]. Nucleic Acids Research, 1996, 24(4): 793-794.

[36] MATZ M, USMAN N, SHAGIN D, et al. Ordered differential display: a simple method for systematic comparison of gene expression profiles [J]. Nucleic Acids Research, 1997, 25(12): 2514-2542.

[37] MALHOTRA K, FOLTZ L, MAHONEY W C, et al. Interaction and effect of annealing temperature on primers used in differential display RT-PCR [J]. Nucleic Acids Research, 1998. 26(3): 854-856.

[38] ZHANG H, ZHANG R, LIANG P. Differential screening of gene expression difference enriched by differential display [J]. Nucleic Acids Research, 1996, 24(12): 2454-2455.

[39] BUESS M, MORONI C, HIRSCH H H. Direct identification of differentially expressed genes by cycle sequencing and cycle labeling using the differential display PCR primers [J]. Nucleic Acids Research, 1997, 25(11): 2233-2235.

[40] VOGELI L R, BURCKERT N, BOLLER T. Rapid selection and classification of positive clones generated by mRNA differential display [J]. Nucleic Acids Research, 1996, 24(7): 1358-1386.

[41] BONNET S, PREVOT G, BOURGOUIN C. Efficient re-amplification of differential display products by transient ligation and thermal asymmetric PCR [J]. Nucleic Acids Research, 1998, 26(4): 1130-1131.

[42] GUPTA R, THOMAS P, BEDDINGTON R S P, et al. Isolation of developmentally regulated genes by differential display screening of cDNA libraries [J]. Nucleic Acids Research, 1998, 26(19): 4538-4539.

[43] FROST M R, GUGGENHEIM J A. Prevention ofdepurination during elution facilitates the reamplification of DNA from differential display gels [J]. Nucleic Acids Research, 1999, 27(15): e6.

[44] CABONI E, LAURI P, WATILLON B, et al. Isolation of mRNA species related to the rooting induction in almond and apple through the differential display technique [J]. Biologia Plantarum, 1997, 39(1): 99-104.