ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 5,Sep. 2015,pp. 1078－1084
DOI:10.16337/j.1004-9037.2015.05.019

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于并行计算的大规模外显子芯片数据分析

张武军　刘学军　张　礼

(南京航空航天大学计算机科学与技术学院,南京,210016)

摘　要:快速准确地计算出转录组表达水平对转录组研究具有重要的作用。本文针对伽玛分布的概率模型(Gamma model for exon array data, GME)在处理大规模外显子芯片数据集上效率低下的特点,提出一种充分利用多核处理机或者集群环境来提高效率的并行计算方法。首先分析GME模型的原理,其次分析模型并行算法的选择,最后在不同规模的数据集上分析并行计算的效率。通过实验验证了并行计算极大地提高了模型的计算效率。实验结果表明,与先前的串行计算相比,并行计算使得GME模型更适用于大规模的外显子芯片分析。

关键词:基因表达;并行计算;外显子芯片;概率模型

中图分类号:TP399　　**文献标志码**:A

Large-Scale Exon Array Data Analysis Based on Parallel Computing

Zhang Wujun, Liu Xuejun, Zhang Li

(College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing, 210016, China)

Abstract: The accurate and fast calculation of transcriptome expression level plays an important role in transcriptome research. Based on the previously devised Gamma model for exon array data (GME), a parallel computing method is proposed to improve the computational efficiency of GME on large-scale Affymetrix exon chip datasets by taking full advantage of multi-core or cluster computation environment. The principles of the GME model and the parallel computing strategy are introduced. The proposed method is verified using real datasets with various scales. The experimental results show that the proposed parallel computing approach greatly improves the efficiency of GME model. Thus the GME model is applicable for the analysis on large-scale exon array datasets.

Key words: gene expression; parallel computing; exon chips; probabilistic model

引　　言

高等真核生物中普遍存在选择性剪切[1]现象,即一个基因在转录过程中选择性地剪切基因序列中的外显子进行连接,从而形成蛋白质异构体,这是生物体内蛋白质多样性的原因之一。研究表明,超过94%的人类基因发生了选择性剪切[2],同时这些选择性剪切还与人类的许多疾病相关[3]。因此,选择性剪切的研究是深入了解病理机制的重要内容之一,尤其是针对大规模数据[4-5]的选择性剪切研究。

基金项目:国家自然科学基金(61170152)资助项目;中央高校基本科研业务费专项(CXZZ11-0217)资助项目。
收稿日期:2014-05-30;**修订日期**:2014-06-23

近年来,随着生物信息学的发展,目前针对大规模数据的选择性剪切研究技术手段主要有两种:基于高通量测序技术(RNA sequencing, RNA-Seq)[6]与基因芯片技术(Mircoarray)[7]。RNA-Seq 是基于高通量测序技术对转录组进行研究的一种新方法,其核心思想是通过将 RNA 序列数据映射到参考基因组或转录组上得到量化的基因表达值与剪切异构体表达值,具有信噪比高、分辨率高、所需样本少等优势。基因芯片技术又称 DNA 微阵列技术,是把大量已知序列探针集成在同一基片上,将标记过的干靶核苷酸序列与芯片特定位点上的探针杂交,通过检测杂交信号,对生物细胞中的基因信息进行分析的一种技术。基因芯片技术具有在基因参考序列上覆盖率高、成本低、使用简单和数据易收集等特点。近年来,随着高通量测序技术的发展,RNA-Seq 已成为转录组可变剪切及表达谱的主流方法,但在针对大规模的数据(生物样本数>30)分析时,基因芯片技术具有明显费用低、对低表达水平的基因稳定等优势[8],仍是大规模数据的选择性剪切研究的主要实验方法[4-5]。

随着大规模选择性剪切研究成为生物医学领域的研究热点,Affymetrix 公司提供了一种外显子芯片用来测量基因剪切异构体表达水平。测量外显子芯片数据的基因剪切异构体表达水平是指从外显子芯片上获得 PM 探针的灰度值,通过分析计算获得基因或者异构体的表达水平,为后续分析提供依据。比如可以根据已知的探针和外显子以及探针和基因的映射关系,通过获得的外显子/基因的表达比率来进行选择性剪切事件的检测[9-10],或者所计算的表达水平传递到后续分析中可以进行更为精细地寻找差异表达、聚类、基因调控网络分析等研究。所以如何快速有效地从基因芯片中的原始数据中分析出表达结果成为研究者们的目标。目前针对外显子芯片也出现了很多数据分析方法,有各自的优点和缺点。一些传统方法如鲁棒多芯片平均算法模型(Robust multi-array average,RMA)[11]和探针对数灰度误差算法(Probes logarithmic intensity error,PLIER)[12],因其两者都仅仅采用完全匹配(Perfect match,PM)探针的灰度值计算表达值,所以不仅可以用于传统的基因芯片数据分析,也可以用于外显子芯片的数据分析,但它们均无法计算出剪切异构体的表达水平。因为基因芯片上的一个探针可能被多个剪切异构体所共享,所以如何合理分离这些探针信号是计算剪切表达水平的难点。除此之外,人们还希望可以获得剪切异构体表达值的方差,这样就可以将结果的不确定程度一起传递到后续分析中,以此获得更有意义的分析结果[11,13]。由于基因芯片是个多步骤的过程,不确定性可能发生在任何一个实验步骤中,这导致了最后的实验结果很难准确表示基因的真实值,而概率方法能很好地模拟这种不确定性。目前已有一些方法可以用来计算剪切异构体的表达值和方差,如外显子芯片预处理方法(Multiple exon array preprocessing,MEAP)[14]和多源映射贝叶斯基因表达计算方法(Multi-mapping Bayesian gene expression,MMBGX)[15]。MEAP 采用非负矩阵分解的方法计算剪切异构体的点估计值,但无法得到该估计值的分布情况。MMBGX 通过一个多层贝叶斯模型来计算转录本的表达水平,以此来获得剪切异构体表达水平的后验分布。该模型采用 MCMC(Markow chain Monte Carlo)求解,计算效率很低。同时,这两个方法均没有考虑有效信号中的探针特性问题。

针对上述问题,本文在先前的工作中设计出了基于伽玛分布的概率模型(Gamma model for exon array data,GME)[16],该模型通过 GATExplorer[17,19]获得的外显子芯片探针、剪切异构体以及基因三者的映射关系来计算基因和剪切异构体的表达水平、方差以及置信区间。该方法通过引入服从伽玛分布的隐含变量,有效地模拟了探针信号的探针特性,并利用伽玛分布随机变量的叠加性质,将被多个剪切异构体共享的探针信号进行分离,该模型采用最大似然估计法求解,计算较为简单。该模型采用 R 语言实现,已包含在生物信息学组件 Bioconductor 中的 Puma[18]软件包中。文献[16,18]通过基因芯片质量控制(Microarray quality control, MAQC)数据集[17]和头颈部鳞状细胞癌(Head and neck squamous cell carcinoma, HNSCC)数据集[12]验证了该模型能够获得较为准确的基因和异构体表达水平。但该模型在大规模外显子芯片实验中仍然存在计算效率问题,GME 算法每次优化涉及的参数随着芯片个数的增加线性增加,如果芯片个数达到 30 以上,GME 的计算速度会变得非常缓慢,严重阻碍了该方法在实

际中的应用。针对这一现象,本文利用多核处理机和集群网络资源,在先前提出的 GME 模型基础上引入并行计算技术,并在不同规模的数据集上验证引入并行计算后模型计算效率得到显著提高。

1 GME 模型分析概述

1.1 GME 模型

GME 是一个通过外显子芯片探针、剪切异构体以及基因三者之间的映射关系而建立的一个双层伽玛模型,用来计算基因以及剪切异构体的表达水平及其分布。GME 的概率模型图如图 1 所示。对于芯片中的每个基因,y_{aj} 表示芯片 a 上观察到的第 j 个 PM 探针信号灰度值,该探针被这个基因的多个剪切异构体所共享;$\boldsymbol{M}$ 矩阵表示探针与剪切异构体之间的对应关系;M_{jk} 为 0 或者 1,0 表示第 j 探针与第 k 个剪切异构体没有映射关系,1 表示有映射关系 ;A 为实验条件的芯片总数;K 为该基因对应的剪切异构体个数;J 为探针的个数。假设 $y_{aj}=\sum_k M_{jk}s_{ajk}$,其中 s_{ajk} 服从参数为 α_{ak} 和 β_j 的伽玛分布 $s_{ajk}\sim Ga(\alpha_{ak},\beta_j)$。根据服从伽玛分布随机变量的性质,则有

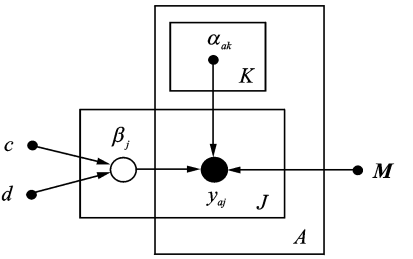

图 1 GME 图模型

Fig. 1 GME modules

$$y_{aj}\sim Ga\left(\sum_k M_{jk}\alpha_{ak},\beta_j\right) \tag{1}$$

进一步假设 β_j 服从参数为 c 和 d 的伽玛分布,即

$$\beta\sim Ga(c,d) \tag{2}$$

则该基因对应的剪切异构体服从以下分布

$$p(s_{ajk})=\int \mathrm{d}\beta_j p(s_{ajk}\mid\alpha_{akc},\beta_{aj})p(\beta_{aj}\mid c,d) \tag{3}$$

观察到的探针信号的对数似然函数为

$$L_a(\alpha_{ak},c,d)=\sum_j \log\int \mathrm{d}\beta_j p(\beta_j\mid c,d)\prod_a p\left(y_{aj}\mid\sum_k M_{jk}\alpha_{ak},\beta_j\right) \tag{4}$$

通过最大似然估计计算出参数$\{\alpha_{ak}\},c,d$ 的估计值$\{\hat{\alpha}_{ak}\},\hat{c},\hat{d}$,利用该估计值就计算出基因和剪切异构体的表达水平。可以看出,GME 模型是一个多芯片模型,模型通过每个基因所对应的剪切异构体所对应的探针,用最大似然估计优化出参数$\{\alpha_{ak}\},c,d$,最后利用这些参数计算最终的表达值水平和方差。模型每次优化涉及的参数个数为 $A*K+2$,所以当芯片个数 A 增大时,模型每次需要优化的参数线性增长,采用最大似然估计优化参数的效率急剧下降,这导致了串行优化大规模外显子芯片中的 4 万多个基因的计算效率很低,特别是剪切异构体多的基因,计算效率更低。针对模型在优化过程中基因之间的计算过程相对独立、互不影响这一特点,希望可以对外显子芯片中的 4 万多个基因进行合理的分割,引入并行计算,让不同的处理器或者集群网络中的各个结点去处理分割后的子任务,达到成倍提高数据处理效率的目的。

1.2 GME 模型的并行算法

对于并行计算,可以用不同的并行算法来分配任务,不同的并行算法可能会对程序的执行效率产生很大影响。一般情况下任务分配方法有两种:静态任务分配和动态任务分配。静态任务分配在程序运行前就已经决定好任务的划分。它把任务平均到每个结点上,对于集群中各个结点的计算能力差别不大时,计算效率最好。动态分配任务在集群中各个结点的计算性能差别较大时采用,这时候需要衡量机器的计算能力,给计算能力强的多分配任务,计算能力弱的少分配。

因为外显子芯片上的各个基因复杂程度不同,估计每个基因计算的时间不可取,同时随着芯片个数

的增加,每个基因优化计算所需的时间非线性,所以本文无法采用静态方法或者动态方法来分配任务使得各个结点在同一时间段完成任务。本文采用两者相结合的方法,算法流程图如图 2 所示。

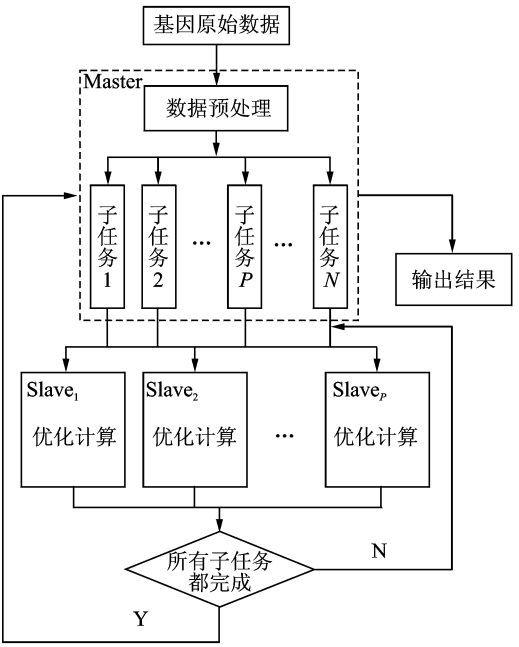

图 2 GME 模型并行算法流程图

Fig. 2 Parallel algorithm flow chart of GME modules

具体的并行算法如下。

(1) 对外显子芯片的原始数据进行预处理获得所有优化计算所要到的数据。

(2) 将外显子芯片上的 4 万多个基因数据等分成 N 份,N 值要远大于集群中的 Slave 结点个数 P,同时也不可以过大,避免把总任务分的太多,造成过多的上下文切换,将大部分时间损耗在数据通信上,同时也不可以太小,避免出现长时间等待某一个 Slave 结点任务结束的现象。本文在程序中默认将 N 值设置为 Slave 结点个数 P 的 10 倍。用户也可以根据芯片规模的大小手动的调节 N 的大小。一般情况下规模越大,N 设置的值越大。

(3) 将 N 份任务的前 P 份传递给 P 个 Slave 结点去优化计算。

(4) 若 Slave 结点中的某个结点完成任务,则此结点自动去 Master 结点取下一任务,以此类推,直至 N 份子任务全部完成,将结果返回给集群中的 Master 结点,输出结果。

本文采用这种方法将 GME 模型进行并行化,实现函数为 gmoExon,包含在生物信息学组件 Bioconductor 软件 Puma 中,可以从 http://www.bioconductor.org/packages/release/bioc/html/puma.html 下载使用。

1.3 GME 模型的并行算法实现

集群的并行环境是通过构建一个高速网络系统,以充分利用高速局域网上的计算机资源来达到快速处理大规模问题的目的。目前,MPI(Message passing interface)[20] 与 PVM(Parallel virtual machine)[21] 都是给用户提供基于消息传递的并行环境。MPI 是基于消息传递的并行编程标准。在标准的串行程序设计语言(C, Fortran 和 C++)基础上,再加入实现进程间通信的 MPI 消息传递库函数,就构成了 MPI 并行程序所依赖的库函数。在并行编程模式中,每个进程享有独立的地址空间,同时它们也

只能访问各自的地址空间,相互之间的访问必须通过显示的消息传递来实现。PVM 是一种通用的网络并行程序开发环境,一般进行网络间的进程消息传递,它可以把多个同构或异构的计算机组成一个易于管理的、可扩展的和易编程使用的并行计算资源。采用 PVM 构造一个全互连结点的虚拟机,此后在此虚拟机上可以动态地创建和管理线程。

本文实现 GME 模型并行计算的基础是一个支持 MPI 和 PVM 消息传递标准的 R 语言软件包 Snow,Snow 包是 R 语言能实现高性能并行计算的基础之一,它的使用思想是在集群中 Master 结点创建多个线程,这多个线程可以来自 Master 结点自身或者集群网络中个各个 Slave 结点。创建完成后通过 Snow 软件包的 VlusterApply 函数向各个线程传递数据,这些数据可共享,也可独立。各个线程获得各自需要的数据后分别完成各自的任务,最终通过 Master 结点收集各个线程的结果直至结束。

2 模型并行前后效率对比

本文采用人类先天免疫反应数据集(Innate immune responses to vaccines, IIRV)[22] 验证不同芯片规模采用并行计算后,GME 模型的效率变化。此数据集采用 Affymetrix 人类外显子芯片 Human Exon 1.0 ST。IIRV 数据集用来研究人类对疫苗的先天免疫反应。在该实验中,对实验对象注射 MRKAd5/HIV 疫苗后一周内测量 HIV 相关细胞的反应。样本采集自 5 个时间点,注射疫苗时以及 4-6,24,72,168 h 时,在每个时间点对样本进行外显子基因芯片实验,测量基因表达水平。本文选取其中 6 个参与者的数据,共包括 30 个芯片,在这 30 个芯片中分别选取 10,15,25,30 个芯片测量 GME 模型在不同芯片规模的效率变化。

GME 在不同规模的 IIRV 数据集和不同并行度下的效率变化如图 3 所示。并行计算环境结点为 IBM 刀片服务器,Intel(R) Xeon(R) CPU X5560 2.80 GHz,内存 32 GB。从图 3 中可以看出模型在没有引入并行计算时,在小规模芯片数据集(<10)所需的时间很少,但随着芯片规模的增加,GME 模型所需要的时间越来越多,尤其当芯片规模增加到 30 个时,GME 模型处理时间需要 5 天左右(136 h)。而且在大多真实的大规模外显子芯片实验中,涉及的芯片个数往往达到几十到上百个[4-5],按照如图 3 所示 GME 模型的效率变化,原先串行的 GME 模型很难用于大规模的芯片数据处理。

引入并行计算后,从表 1 中可以清楚的看到在 2,4,8 不同的并行度下 GME 模型的效率变化:随着并行度的增加,GME 模型所需要的时间越来越少。从图 3 可看到,在芯片规模比较少时,效率提高得还不太明显,但随着芯片规模的变大,效率变化得越来越明显,特别是在芯片规模达到 30 个时,并行度为 8 时计算只需要 19.2 h 左右,相对于原先的 136.0 h 有了极大的提高,同时这一效率值还可以通过增加并行度继续提高。

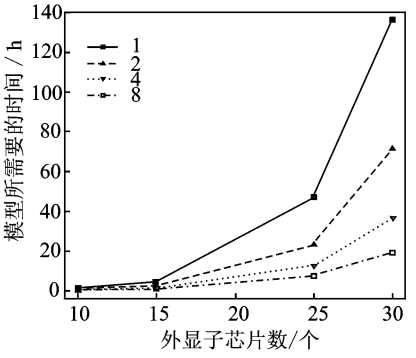

图 3 模型在不同规模 IIRV 外显子芯片和不同并行度下的效率

Fig. 3 Efficiency of modules under different numbers of IIRV exons chips and thread

表 1 不同规模 IIRV 芯片在不同并行度的效率比较

Table 1 Comparison of efficiency using different numbers of IIRV chips and thread h

数据集	GME(1)	GME(2)	GME(4)	GME(8)
‖RV(2 个参与者,5 个实验条件,10 个芯片)	1.5	1.2	0.7	0.4
‖RV(3 个参与者,5 个实验条件,15 个芯片)	5.0	2.6	1.3	1.0
‖RV(5 个参与者,5 个实验条件,25 个芯片)	47.6	23.2	12.5	7.5
‖RV(6 个参与者,5 个实验条件,30 个芯片)	136.0	71.6	36.5	19.2

3 结束语

本文针对 GME 模型处理大规模外显子芯片数据效率极低的特点,引入利用多核处理机或集群环境资源的并行计算,使得模型更好地适用于大规模的数据处理。并行计算的 GME 模型实现在生物信息学组件 Bioconductor 的最新 Puma 软件包中。虽然此并行计算是针对外显子芯片设计,但是其并行化思想对其他类型的基因芯片的大规模数据处理也具有较好的适用性。

参考文献:

[1] Valenzuela A, Talavera D, Orozco M, et al. Alternative splicing mechanisms for the modulation of protein function: Conservation between human and other species [J]. Journal of Molecular Biology, 2004, 335(2):495-502.

[2] Wang E T, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes [J]. Nature, 2008, 456(7221):470-476.

[3] Cáceres J F, Kornblihtt A R. Alternative splicing: Multiple control mechanisms and involvement in human disease [J]. Trends in Genetics, 2002, 18:186-193.

[4] Taylor B S, Schultz N, Hieronymus H, et al. Integrative genmonic profiling of human prostate cancer [J]. Cancer Cell, 2010,18(1): 11-22.

[5] Bullard J H, Purdom E, Hansen K D, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments[J]. BMC Bioinformatics, 2010, 11:94-101.

[6] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics [J]. Nature Reviews Genetics, 2009, 10 (1): 57-63.

[7] Service R F. Microchip arrays put DNA on the spot[J]. Science, 1998, 282(5388):396-399.

[8] Łabaj P P, Leparc G G, Linggi B E, et al. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling [J]. Bioinformatics, 2011, 27(13): i383-i391.

[9] Purdom E, Simpson K M, Robinson M D. FIRMA: A method for detection of alternative splicing from exon array data [J]. Bioinformatics, 2008, 24:1707-1714.

[10] Xing Y, Stoilov P, Kapur K, et al. MADS: A new and improved method for analysis of differential alternative splicing by exon-tiling microarrays [J]. RNA, 2008, 14(8): 1470-1479.

[11] Irizarry R A, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data [J]. Biostatistics, 2003, 4:249-264.

[12] Affymetrix Whitepaper. Alternative transcript analysis methods for exon arrays[EB/OL]. http://www.affymetrix-.com/support/tenchnical/whitepapers/exon_alt_transcript_analysis_whitepaper.pdf, 2005-10-11.

[13] Liu X, Rattray M. Including probe-level measurement error in robust mixture clustering of replicated microarray gene expression[J]. Statistical Applications in Genetics and Molecular Biology, 2010, 9:42.

[14] Chen P, Lepikhova T, Hu Y, et al. Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants [J]. Nucleic Acids Research, 2011, 39:e123.

[15] Turro E, Lewin A, Rose A, et al. MMBGX: A method for estimating expression at the isoform level and detecting differential splicing using whole-transcript Affymetrix arrays [J]. Nucleic Acids Research, 2010, 38:e4.

[16] 高珍珠, 刘学军, 张礼. 一种基于概率模型 Affymetrix 外显子芯片原始数据分析方法[C]//2011 中国生物医学工程联合学术年会论文集(光盘版). 武汉: 中国生物医学工程学会, 2011.
Gao Zhenzhu, Liu Xuejun, Zhang Li. A probabilistic model for the analysis of Affymetrix exon arrays data[C] //Proceeding of CBME2011(CD). Wuhan: Chinese Society of Biomedical Engineering, 2011.

[17] Risueño A, Fontanillo C, Dinger M E, et al. GAT Explorer: Genomic and transcriptomic explorer, mapping expression probe to gene loci, transcripts, exons and ncRNAs [J]. BMC Bioinformatics, 2010, 11:221.

[18] Liu X, Gao Z, Zhang L, et al. Puma 3.0: Improved uncertainty propagation methods for gene and transcript expression analysis [J]. BMC Bioinformatics, 2013, 14:39.

[19] Consortium M. The micro array quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements [J]. Nature Biotechnology, 2006, 24:1151-1161.

[20] Jihching C, Liang C, Tzeng H Y. A multi-streaming SIMD architecture for multimedia applications[C]// Conference on Computing Frontiers 2009. Ischia, Italy: [s. n.],2009:51-60.

[21] Konuru R B, Otto S W, Walpole J. A migratable user-level process packages for PVM [J]. Journal of Parallel and Distributed Computing,1977, 15(1): 3-40.

[22] Zak D E, Andersen N E, Peterson E R, et al. Merck Ad5/HIV induces broad innate immune activation that predicts CD8$^+$ T-cell responses but is attenuated by preexisting Ad5 immunity [J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(50):E3503-12.

作者简介:

张武军(1989-),男,硕士,研究方向:生物信息学,E-mail: 350121197@qq.com。

刘学军(1976-),女,教授,研究方向:生物信息学。

张礼(1985-),男,博士,研究方向:生物信息学。