文章编号:1004-9037(2013)05-0572-08

一种改进的 Affymetrix 外显子芯片原始数据分析方法

刘学军 张武军 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:选择性剪切与许多人类疾病有关,基因以及基因异构体水平的表达分析是揭示选择性剪切变化情况的常用研究方法,Affymetrix 外显子芯片为测量基因以及基因异构体表达水平提供了一种重要方法。由于外显子芯片基于杂交技术进行设计,实验数据中存在大量噪声,并且选择性剪切导致一个探针往往对应多个剪切异构体,这些给剪切异构体表达水平的计算带来了挑战。为此在先前提出的基于伽玛分布的概率模型(Gamma model for exon array data, GME)基础上,提出了 iGME 模型,进行基因以及异构体表达水平的计算。该模型利用已知的基因剪切异构体与探针的对应关系,模拟了条件独立的探针特性。通过采用真实实验数据进行验证,并与传统方法进行比较,结果表明 iGME 模型获得了较高的计算精度和更快的计算速度。

关键词:基因表达;选择性剪切;伽玛分布;概率模型;外显子芯片

中图分类号:TP399 **文献标志码**:A

Improved Method for Probe-level Affymetrix Exon Array Data Analysis

Liu Xuejun, *Zhang Wujun*, *Zhang Li*

(1. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China)

Abstract: Alternative splicing is associated with many human diseases. Analysis of gene and isoform expression is an approach to study the variation of alternative splicing. Affymetrix exon arrays provide an important tool to measure the expression of gene and isoform. The computation of gene and isoform expression level measured from Affymetrix exon arrays is challenging due to the noises caused by cross-hybridization and the multi-mappings between probes and isoforms. Based on previously devised Gamma model for exon array data(GME) method, this paper proposes an improved method, iGME, to compute gene and isoform expression. The iGME method uses the known mappings between probes and isoforms, and models the condition-independent probe effects. The new method is verified using three real data sets and compared with several traditional methods. The results show that iGME is more accurate and computationally efficient.

Key words: gene expression; alternative splicing; gamma distribution; probabilistic model; exon array

引 言

高等真核生物中普遍存在选择性剪切(Alternative splicing, AS)现象[1],即一个基因的多个外显子以多种选择方式进行连接,从而导致产生多种蛋白质异构体(Isoform),如图 1 所示。研究表明超过 94%的人类基因发生选择性剪切[2],同时选择性剪切与许多人类疾病有关[3]。因此,研究选择性剪切对于联系蛋白质和转录体的意义重大,同时它也是致病机制研究的重要内容之一。基因以及基因异构体水平的表达分析为揭示 AS 的变化情况提供了一种可行的研究方式。基因芯片由于其在参考基因序列上的高覆盖率、低成本、使用简单、

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:2013-06-12;**修订日期**:2013-08-10

数据收集方便等优点，在过去十多年来被广泛应用于高通量基因表达水平的测量，至今仍然是基因表达水平测量的可靠工具。特别对于低表达水平的基因，基因芯片技术仍具有明显的优势[4]。近年来，随着选择性剪切成为生物医学领域的研究热点，外显子芯片技术提供了一种测量基因异构体表达水平的方法，比如 Affymetrix 公司的外显子芯片(GeneChip exon arrays)。

Affymetrix 基因芯片采用 25 个碱基长度的探针来测量样本中转录本的丰度，Affymetrix 传统的 3'基因芯片使用完全匹配(Perfect match，PM)和错位(Mismatch，MM)这两种类型的探针，其中 PM 探针碱基序列和目标序列完全匹配。为了测量基因芯片实验中存在的交叉杂交以及其他类型的背景噪声，基因芯片在设计中引入 MM 探针，MM 探针仅将中间一位碱基换成互补碱基，其余各位与目标序列完全匹配[5]。而 Affymetrix 外显子芯片没有像传统芯片那样使用两种类型探针，而仅采用了 PM 探针，约 90% 的外显子被 4 个探针覆盖(如图 1 所示)，这样使外显子芯片获得了更高的集成度和覆盖率，从而可以在外显子、基因异构体以及基因等不同层次上测量转录本的表达水平[6]。一些传统 3'芯片原始数据分析方法，如鲁棒多芯片平均算法(Robust multi-array average，RMA)[7]以及探针对数灰度误差算法(Probe logarithmic intensity error，PLIER)[8]，由于仅采用 PM 探针信号进行计算，可以直接应用于外显子芯片基因以及外显子表达水平计算。根据已知的探针和外显子以及探针和基因的映射关系，可以进行外显子以及基因表达水平的计算，通过获得的外显子/基因表达比率，可以进行选择性剪切事件检测[9-11]。除可计算外显子/基因表达比率外，还可计算基因以及基因异构体表达水平。所计算得到的表达水平传递到后续分析中，可以进行更为精细的寻找差异表达、聚类、基因调控网络分析等研究。所以准确的基因以及异构体表达水平计算方法对后续分析以及生物学发现具有重要作用。

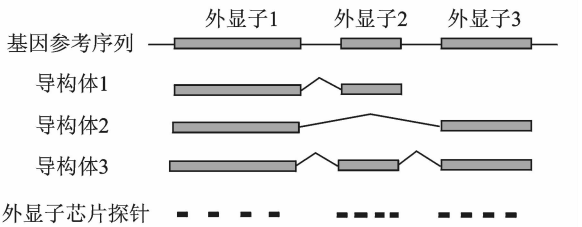

图 1　选择性剪切以及外显子芯片探针设计

基因表达水平的计算可以采用一些传统基因芯片数据处理方法，如 RMA 以及 PLIER，但是基因异构体表达水平的计算面临新的挑战，因为一个探针可能被多个基因异构体共享，该探针所测量到的信号可能来自所有可能映射到的异构体，故分离这些共享的探针信号是解决异构体水平计算，获得了准确基因表达水平的一个难点。此外，基于杂交技术的基因芯片原始数据一个重要特征是探针特性(Probe effects)，即探针信号分布与探针碱基序列内容有较强相关性。一些 3'芯片数据分析方法考虑到这种探针特性，获得了较好的分析结果[12-13]。最后，人们希望获得表达值的同时能够得到该表达值的方差，这样将计算结果的不确定程度一起传递到后续分析中，可以获得更具有意义的分析结果[14-15]。随着基因组注释信息的完善，人们可以获得探针和已知 Ensembl 转录本的映射关系，近年来出现了计算已知异构体表达水平的方法，如多映射贝叶斯基因表达计算方法(Multi-mapping Bayesian gene expression，MMBGX)[16]和多外显子芯片预处理方法(Multiple exon array preprocessing，MEAP)[17]。MMBGX 通过一个多层贝叶斯模型计算转录本表达水平，获得异构体表达水平的后验分布。该模型采用 MCMC(Markov chain Monte Carlo)求解，故计算效率较低，在实际中应用很困难。MEAP 采用非负矩阵分解的方法计算异构体表达水平的点估计值，而不能得到该估计值的分布情况。另外，这两种方法均没有考虑探针有效信号中的探针特性问题。

考虑到现有方法存在的这些问题，在先前工作中设计了基于伽玛分布的概率模型(Gamma model for exon array data，GME)[18]，利用通过 GATExplorer[19]获得的外显子芯片探针、异构体以及基因三者之间的对应关系，计算基因以及基因异构体的表达水平及其分布。该模型采用 R 语言实现，并包含国际生物信息学组件 Bioconductor 中的 puma[20]软件包中。该方法通过引入服从伽玛分布的隐含变量，有效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，将多异构体共享探针信号进行分离。模型采用最大似然法求解，计算较为简单。在文献[18]和[20]中通过实验验证了该方法能够获得较为准确的基因以及异构体表达水平。但是，该方法存在两个问题。首先该方法是一个多条件模型，即假设探针特性被多个实验条件共享，但目前并没有证据表明探针特性能够作为一种芯片特征被所有实验条件共享。当

实验条件较少的时候，探针特性表现较为分散，并不能用一个概率分布较好地表示一个芯片的探针特性。另外，由于多个实验条件同时处理，一个基因所包含的异构体个数也较多的时候，会导致模型参数个数较多，给模型优化带来困难，使计算速度较慢。为此，本文提出了改进的 GME(Improved GME, iGME)模型，取消对探针特性模拟的多条件共享，为每个实验条件分别模拟探针特性，以降低模型复杂程度，简化模型参数的优化计算，并在 3 个真实的实验数据集上进行验证，以证明该方法的有效性。

1 本文方法

1.1 GME 模型

GME 是一个概率模型，它利用通过 GATExplorer[19] 获得的外显子芯片探针、异构体以及基因三者之间的对应关系，计算基因以及基因异构体的表达水平及其分布。GME 的概率图模型表示如图 2 所示。对于每个基因，y_{aj} 是芯片 a 上观测到的第 j 个 PM 探针信号(用实心圆表示)，该探针信号来自该基因上的多个剪切异构体。$\boldsymbol{M}$ 矩阵表示探针与该基因异构体之间的对应关系，矩阵元素 M_{jk} 取值 0 或 1，0 表示探针 j 与异构体 k 没有映射关系，1 表示有映射关系。A 为实验中涉及的芯片总数，K 为该基因包含的异构体个数，J 为探针个数。假设 $y_{aj}=\sum_k M_{jk}s_{ajk}$，其中 s_{ajk} 为异构体 k 对探针 j 贡献的信号强度，且服从参数为 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，$y_{aj}$ 由以下伽玛分布产生

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

假设 β_j 为隐含变量(在图 2 中用空心圆表示)，且来自参数为 c 和 d 的伽玛分布，即

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

图 2 中 c,d,α_{ak} 以及 $\boldsymbol{M}$ 为模型超参数(用小实心点表示)，其中 $\boldsymbol{M}$ 已知。GME 模型产生探针信号的过程为：首先，根据式(2)中的伽玛分布为每个探针产生 β_j；然后，根据式(1)为每个探针产生在每个芯片上的信号 y_{aj}。由此可见，β_j 表示了探针特性，与芯片和实验条件无关。在该模型假设条件下，观测到的探针信号的对数似然为

$$L(\{y_{aj}\}\mid\{\alpha_{ak}\},c,d)=\sum_j\log\int \mathrm{d}\beta_j P(\beta_j\mid c,d)\prod_a P\left(y_{aj}\,\middle|\,\sum_k M_{jk}\alpha_{ak},\beta_j\right) \tag{3}$$

通过最大似然估计法计算参数 $\{\alpha_{ak}\},c,d$ 的估计值 $\{\hat{\alpha}_{ak}\},\hat{c},\hat{d}$。利用该估计值来计算剪切异构体表达水平 s_{ajk} 的概率密度分布函数

$$P(s_{ajk})=\int \mathrm{d}\beta_j P(s_{ajk}\mid\hat{\alpha}_{ak},\beta_j)P(\beta_j\mid\hat{c},\hat{d}) \tag{4}$$

假设基因的表达值是所含异构体表达值之和，即 $\sum_k s_{ajk}$，则根据伽玛分布的性质，基因表达值也服从伽玛分布，$\sum_k s_{ajk}\sim Ga\left(\sum_k\alpha_{ak},\beta_j\right)$。类似地，基因表达值的分布可以表示为

$$P\left(\sum_k s_{ajk}\right)=\int \mathrm{d}\beta_j P\left(\sum_k s_{ajk}\,\middle|\,\sum_k\hat{\alpha}_{ak},\beta_j\right)P(\beta_j\mid\hat{c},\hat{d}) \tag{5}$$

基因以及异构体的对数表达值的后验分布可以由式(4,5)进行计算，进而可得到其对数表达值的期望，并可用一个高斯分布来近似表示。高斯分布表示的基因以及异构体表达水平，可以较为方便地将原始数据中的不确定度传递到后续表达值分析中，以获得更具生物学意义的分析结果。在文献[18,20]中通过实验验证了该方法在基因以及异构体表达水平计算上的有效性。

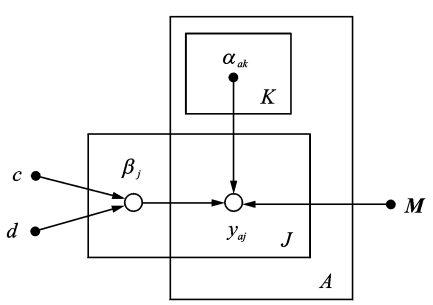

图 2 GME 的图模型表示

1.2 iGME 模型

GME 模型方法是一个多芯片模型，多个芯片不区分所属的不同实验条件，探针特性 β_j 被多个实验条件共享。当实验条件较少，探针特性表现较为分散的时候，用一个概率分布不能较好地表示一种类型芯片探针特性的统计特性。另外，由于实验中所有芯片同时处理，一次模型优化涉及的参数个数为 $A\times K+2$。如果实验涉及 100 个芯片，每个人类基因平均 4 个异构体，则每个基因平均需要计算的模型参数总数为 402 个，而人类外显子芯片设计为 4 万多基因，这样在较大数据集上模型优化较为困难，容易陷入局部极值点，且计算速度较慢。基于上述考虑，本文提出了 iGME 模型，取消对探针特性模拟的多条件共享，为每个实验条件分别模拟探针特性，这样可以更好地获得探针特性的统计

特征，同时降低模型复杂程度，简化模型参数的优化计算。

iGME 的概率图模型表示如图 3 所示，C 表示实验条件的个数，R 表示每个实验条件中涉及的芯片个数。由图 3 可以看出，探针特性 β_{cj} 与实验条件 c 有关，针对每个实验条件，产生该条件下的探针信号，具体数据产生过程如下：

(1)在实验条件 c 下，根据伽玛分布为每个探针产生 β_{cj}，$\beta_{cj}\sim Ga(c_c,d_c)$；

(2)对于该条件下的第 r 个芯片，根据伽玛分布产生探针信号 y_{crj}，$y_{rcj}\sim Ga\left(\sum_k M_{jk}\alpha_{crk},\beta_{cj}\right)$。

每个条件下探针信号的对数似然函数为

$$L(\{y_{crj}\}\mid\{\alpha_{crk}\},c_c,d_c)=\sum_j\log\int \mathrm{d}\beta_{cj}P(\beta_{cj}\mid c_c,d_c)\prod_r P\left(y_{crj}\mid\sum_k M_{jk}\alpha_{crk},\beta_{cj}\right) \tag{6}$$

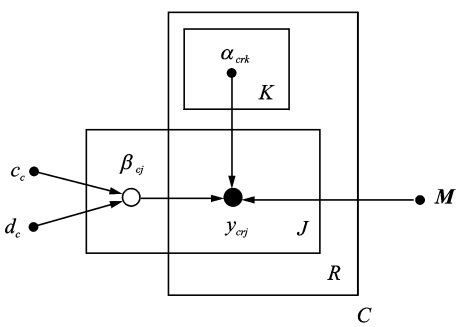

图 3　iGME 的图模型表示

通过最大似然法求解模型参数，在条件 c 下模型参数个数为 $R\times K+2$。由于 GME 模型中 $A=C\times R$，故 iGME 在每个条件下需要优化的参数个数大大少于 GME 模型参数，简化了参数计算。得到模型参数的最大似然估计值后，类似地根据式(4,5)可以得到条件 c 下异构体以及基因的表达水平后验分布分别为

$$P(s_{crjk})=\int \mathrm{d}\beta_{cj}P(s_{crjk}\mid\hat{\alpha}_{crk},\beta_{cj})P(\beta_{cj}\mid\hat{c}_c,\hat{d}_c) \tag{7}$$

$$P\left(\sum\nolimits_k s_{crjk}\right)=\int \mathrm{d}\beta_{cj}P\left(\sum\nolimits_k s_{crjk}\mid\sum\nolimits_k\hat{\alpha}_{crk},\beta_{cj}\right)\cdot P(\beta_{cj}\mid\hat{c}_c,\hat{d}_c) \tag{8}$$

最后，iGME 模型采用高斯分布来近似表示基因及异构体在每个条件下每个芯片上的对数表达值的期望。

1.3　iGME 模型处理流程

图 4 显示了采用 iGME 方法进行外显子芯片数据处理的流程图。对于包含 C 个实验条件的数据集，首先输入全部原始数据，然后按照条件依次处理每个条件下的数据，进行模型优化计算并计算基因及异构体表达水平。全部数据处理完之后进行基因表达值标准化。标准化是基因芯片数据处理中必要的步骤，以消除实验中不同芯片上的技术误差带来的整体信号偏移[21]。由于标准化不是本文主要工作，故本文采用了较为简单的基因表达水平对数中值中心化的方法进行标准化，即每个芯片最终获得的基因表达水平对数中值一致。最后，输出所获得的基因及异构体表达水平最终结果。

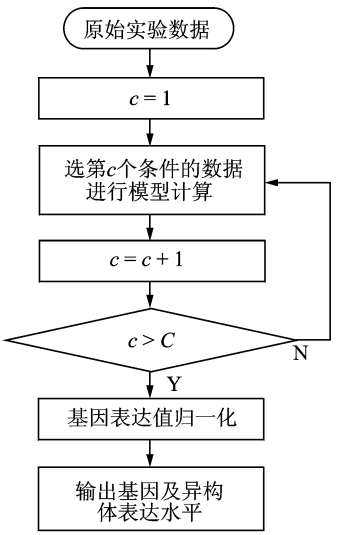

图 4　iGME 模型处理流程

2　实验数据集

本文采用基因芯片质量控制(Micro array quality control，MAQC)数据集[22]、人类头颈鳞状细胞癌(Head and neck squamous cell carcinoma，HNSCC)数据集[17]以及人类先天免疫反应数据集(Innate immune responses to vaccines，IIRV)[23]来验证 iGME 模型在计算准确度以及计算速度方面的性能，这 3 个数据集均采用了 Affymetrix 人类外显子芯片 Human Exon 1.0 ST。

2.1　MAQC 数据集

MAQC 项目通过测量高质量 RNA 样本中基因表达水平来比较不同测量平台的性能[22]。本文选取 Affymetrix 外显子芯片测量的通用人类参考 RNA(Universal human reference RNA，UHRR)以及人类大脑参考 RNA(Human brain reference RNA，HBRR)两组实验结果，每组实验包括 5 次

重复式样。该实验分别在 McGill University (MU)和 Virginia Tech (VT)两个独立的实验室进行了两次,本文随机选取 MU 获得的数据验证本文的方法。除了基因芯片实验外,MAQC 项目另外对大约 1 000 个基因进行了 qRT-PCR 实验,这些结果可以作为其他测量平台实验结果的参照,用以评价其他平台的性能。

在这些 qRT-PCR 数据中,本文参照文献[24]中的方法将两个实验条件下确定为差异(DE)的基因和没有差异(non-DE)的基因过滤出来,具体数据过滤方法请参考文献[24]。过滤出来的基因在外显子芯片上共有 305 个,其中 87 个 non-DE 基因,218 个 DE 基因。在 DE 基因中,有 116 个基因在 UHRR 样本中是表达值上升的,记为"DE+",而另外 102 个在 UHRR 样本中是表达值下降的,记为"DE-"。这些数据作为参考标准,可以绘制出从基因芯片实验中计算获得的相应基因表达水平结果的接收者操作特征(Receiver operating characteristic, ROC)曲线,并同时考虑基因在不同条件下的调控方向。

2.2 HNSCC 数据集

HNSCC 数据集也包含 qRT-PCR 验证数据,本文采用该数据集验证 iGME 在基因异构体表达水平上的计算性能。该数据集包含从口腔和喉部获得的 15 个细胞样本,采用 Affymetrix 外显子芯片进行表达值测量。在人类头颈鳞状细胞癌中,染色体 11q13 区域扩增是一种常见的基因变异现象。这 15 个样本分为两组,一组带有 11q13 扩增区域(记为 11q13+),包含 7 个样本,另外一组没有 11q13 扩增区域(记为 11q13-),包含 8 个样本。针对该数据集位于染色体 11q13 扩增区域并且与 HNSCC 相关的两个多异构体基因(ORAOV1 和 NEO1)进行了 qRT-PCR 实验。本文采用 iGME 计算与这两个基因相关的 4 个异构体的表达值,并采用基于概率模型的寻找差异表达方法 IPPLR[25] 计算差异异构体,将计算结果与 qRT-PCR 的结果进行对比,以验证 iGME 在基因异构体表达水平上的计算性能。根据文献[17]记录的实验结果,获得了 qRT-PCR 实验中 4 个异构体在两个不同条件中,以及一个条件同一基因的不同异构体共 8 种表达调控比较结果。

2.3 IIRV 数据集

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,注射疫苗的时候以及 4~6,24,72,168 h 的时候,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 10 个参与者的数据,共包括 $10\times5=50$ 个芯片,用来验证本文提出的 iGME 的计算效率,并与先前提出的 GME 方法进行对比。

3 结果和讨论

本文分别采用 MAQC 和 HNSCC 数据集对所提出的 iGME 方法在基因以及基因异构体表达水平计算两个方面进行计算准确度验证,并与目前流行的方法 RMA,PLIER 以及 MEAP 进行对比。另外,采用这两个数据集以及较大的 IIRV 数据集比较 iGME 和 GME 的计算速度。

3.1 基因表达水平计算结果验证

本文采用 MAQC 数据集验证 iGME 在计算基因表达水平方面的性能,并与之前的 GME 以及传统的 RMA 和 PLIER 进行对比。采用这 4 种方法计算 2.1 节过滤出来的 305 个具有 qRT-PCR 结果验证的基因的表达水平,根据所计算的基因表达水平,采用寻找差异基因的方法来确定差异基因,并与 qRT-PCR 的结果进行对比,由对比结果的一致性来评价基因表达水平计算方法的性能。iGME 采用 R 语言实现,GME 采用 Bioconductor 中的 puma 包实现,RMA 和 PLIER 采用 affy 包实现。

由于 MAQC 数据集涉及两个实验条件,每个条件有 5 次重复实验,故寻找差异基因的时候需要将每个条件下的重复实验结果进行结合,并计算差异重要性检验。由于 iGME,GME 以及 RMA 能够得到基因表达水平以及相应方差,所以本文采用基于概率模型的寻找差异方法(Improved probability of positive log-ratio, IPPLR)[25]来确定差异基因。PLIER 方法仅能得到基因表达水平的点估计值,不能获得该估计值的方差,故采用传统的 *t*-test 来寻找差异基因。

图 5 显示了 GME 和 iGME 获得的 MAQC 数据集 ROC 曲线,该曲线表示了基因芯片结算结果与 qRT-PCR 结果一致性的程度。ROC 曲线下面积越接近于 1,表示所获得的基因芯片计算结果越接近 qRT-PCR 结果。由于在真实的实验中人们往往更关注假阳率较小的差异基因结果,从图 5 可以看出,在假阳率小于 0.1 的范围中,iGME 相比 GME 得到 ROC 曲线下面积更大。表 1 第 1 行数

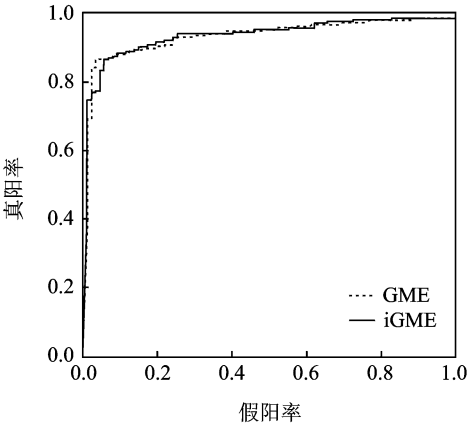

图 5　GME 和 iGME 方法获得的 MAQC 数据集 ROC 曲线

据显示了 iGME，GME，PLIER 以及 RMA 四种方法获得的 ROC 曲线下面积。可以看到在这四种方法中，RMA 获得了最为准确的基因表达水平计算结果，iGME 准确度比 GME 略微提高，PLIER 准确度最低。表 1 第 2 行显示了四种方法获得的差异基因在两个条件下表达值变化倍数与 qRT-PCR 结果相比的相关系数，该结果越接近于 1 表示相应计算方法越准确。由表 1 可以看到 RMA 获得了最高相关系数，iGME 相比 GME 结果略优，PLIER 结果最低；iGME 的基因表达水平计算精度相比 GME 得到略微提升。虽然 RMA 能够得到相对更为准确的基因表达水平计算结果，但是该方法不能计算异构体表达水平，故无法用来进行选择性剪切的研究。

表 1　MAQC 数据集计算结果

方法	iGME	GME	PLIER	RMA
ROC 曲线下面积	0.935 6	0.934 8	0.929 1	0.949 3
与 qRT-PCR 结果相关系数	0.867 8	0.862 5	0.794 1	0.885 7

3.2　异构体表达水平计算结果验证

采用具有 qRT-PCR 验证数据的 HNSCC 数据集来验证 iGME 模型计算异构体表达水平的准确度。在 HNSCC 数据集中，基因 ORAOV1 的两个异构体（ORAOV1-201 和 ORAOV1-202）以及基因 NEO1 的两个异构体（NEO1-201 和 NEO1-202）通过 qRT-PCR 实验进行了表达水平的验证。本文采用 iGME 对该数据集进行了处理，得到这 4 个异构体在 15 个芯片上的表达值以及方差。根据 qRT-PCR 结果，这 4 个异构体在 11q13＋样本中均为显著超表达。在 11q13＋和 11q13－两组样本中，异构体 ORAOV1-201 表达水平要高于 ORAOV1-202，而异构体 NEO1-202 表达水平要高于 NEO1-201。

表 2 显示了 qRT-PCR，iGME 和 GME 得到的 8 种表达调控变化方向，其中“＋”表示上升调控（Up-regulation），“－”表示下降调控（Down-regulation）。对于 iGME 和 GME，本文采用 IPPLR 计算了差异表达的重要性，用 max（PPLR，1－PPLR）表示，其中 PPLR（Probability of positive log ratio）表示正对数倍数的概率。PPLR 值大于 0.5 表示上升调控，小于 0.5 表示下降调控，max（PPLR，1－PPLR）越接近于 1，表示差异表达越显著。由表 2 可见 iGME 在全部 8 组对比结果中有 7 组（除了第 3 组对比）获得了较为显著的差异表达，而 GME 有 6 组（除了第 3 组和第 6 组）显著差异表达，故 iGME 获得了与 qRT-PCR 更为一致的结果，体现了更高的异构体表达水平计算精度。表 2 最后两列显示了 MEAP 方法的计算结果（数据来自文献[17]），其中第 2 行异构体 ORAOV1-202 表达水平 qRT-PCR 结果在条件 11q13＋中相比 11q13－要显著上升，而 MEAP 得到的变化倍数仅为 1.040，差异不明显。另外 MEAP 方法仅能得到每个异构体表达水平的点估计值，不能得到该值的方差，限制了在后续处理步骤中采用概率方法进行差异重要性检测，这在一定

表 2　HNSCC 数据集计算结果

比较分组			qRT-PCR（p-value＜0.1）	iGME		GME		MEAP	
				调控方向	max（PPLR，1－PPLR）	调控方向	max（PPLR，1－PPLR）	调控方向	变化倍数
ORAOV1	ORAOV1-201	11q13＋ vs. 11q13－	＋	＋	0.999 4	＋	1.000 0	＋	3.845
ORAOV1	ORAOV1-202	11q13＋ vs. 11q13－	＋	＋	1.000 0	＋	0.997 4	＋	1.040
NEO1	NEO1-201	11q13＋ vs. 11q13－	＋	＋	0.534 8	＋	0.574 0	＋	3.283
NEO1	NEO1-202	11q13＋ vs. 11q13－	＋	＋	0.966 5	＋	0.969 3	＋	2.459
ORAOV1	11Q13＋	ORAOV1-201 vs. 202	＋	＋	0.862 1	＋	0.915 5	＋	5.598
ORAOV1	11Q13－	ORAOV1-201 vs. 202	＋	＋	0.999 9	＋	0.572 3	＋	1.514
NEO1	11Q13＋	NEO1-201 vs. 202	－	－	0.999 9	－	0.999 9	－	9.761
NEO1	11Q13－	NEO1-201 vs. 202	－	－	0.942 2	－	1.000 0	－	13.032

程度上限制了该方法的实际应用。

3.3 计算速度比较

在大多真实的外显子芯片实验中，涉及的芯片数目往往几十到上百个[23,26-27]，先前提出的 GME 模型在计算大数据集的时候面临参数过多、优化困难的问题，而本文提出的 iGME 由于在每个实验条件中单独模拟探针特性的分布，在一次优化中大大减少了模型参数的个数，从而显著提高了模型计算速度。为了验证 iGME 模型相比之前提出的 GME 方法在计算效率上的提高，本文在 MAQC，HNSCC 和 IIRV 三个数据集上对比这两种方法的计算时间，结果如表 3 所示(计算环境为戴尔 Precision T3500 工作站，Intel Xeon CPU w3550 3.07 GHz，24 GB 内存)。由表 3 可以看出，随着数据集芯片个数增加，iGME 和 GME 所需的计算时间均增加。但是在每个数据集上 iGME 所需计算时间比 GME 少，并且随数据集增大，计算效率比 GME 显著提高。在涉及 50 个芯片的 IIRV 数据集上，iGME 在一天之内可以处理完，而 GME 则估计需要 40 多天，这使得 GME 在这种大规模的数据集上无法使用。另外，作者试图采用另外一个外显子芯片数据处理方法 MMBGX[16] 来处理本文中的 3 个数据集，但是在采用并行计算技术的前提下每个数据集估算计算时间仍均需 2 周以上，故同样难以使用。通过表 3 的比较可以看到，在较大数据集上 iGME 具有较高计算效率，使其具有较好的应用前景。

表 3 iGME 和 GME 方法在不同数据集上计算时间比较

数据集	iGME	GME
MAQC(10 个芯片，2 个实验条件)	3.95 h	5.4 h
HNSCC(15 个芯片，2 个实验条件)	6.33 h	10.42 h
IIRV(50 个芯片，5 个实验条件)	18.67 h	估计 42.5 d

4 结束语

本文针对 Affymetrix 外显子芯片数据分析中基因以及基因异构体计算中的探针特性模拟和模型计算效率的问题，在先前工作的基础上提出了改进的处理方法 iGME。该方法利用已知的探针、异构体以及基因之间的映射关系，解决了探针对基因异构体的多源映射问题；取消对探针特性模拟的多条件共享，为每个实验条件分别模拟探针特性，更好地获得探针特性的统计特征，降低了模型复杂程度。本文将 iGME 应用到具有 qRT-PCR 实验验证的 MAQC 和 HNSCC 数据集，以及较大的 IIRV 数据集进行性能验证。在 MAQC 数据集上验证了 iGME 基因表达水平计算的准确度，并和 GME 以及目前流行的方法 RMA 和 PLIER 进行了对比。在 HNSCC 数据集上验证了 iGME 基因异构体表达水平的准确度，并和 GME 以及 MEAP 方法进行了对比。此外，在这 3 个数据集上验证了 iGME 相对 GME 在计算效率方面的提高。对比结果表明，iGME 相比其他方法获得了较为准确的基因以及异构体表达水平计算结果，并且相比 GME 显著提高了计算效率，在选择性剪切研究中具有较好的应用前景。最后，虽然 iGME 获得了相对较高的计算精度和计算效率，但在现实中特别大规模的数据集上其计算效率仍然可能成为瓶颈。由于 iGME 逐个处理每个基因的数据，因此在后续工作中可以对现有计算程序进行并行化改进，充分利用多核处理器以及集群等大规模计算环境，使该模型更好地应用于大规模数据的处理。

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] 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.

[5] Lockhart D J, Dong H, Byrne M C, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays [J]. Nature Biotechnology, 1996, 14:1675-1680.

[6] Kapur K, Xing Y, Ouyang Z, et al. Exon arrays provide accurate assessments of gene expression [J], Genome Biology, 2007, 8(5): R82.

[7] 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.

[8] Affymetrix Whitepaper: Alternative transcript analysis methods for exon arrays [EB/OL]. http://

www. affymetrix. com/support/technical/whitepapers/exon_alt_transcript_analysis_whitep aper. pdf.

[9] Affymetrix. Alternative transcript analysis methods for Exon arrays [EB/OL]. (2005-10-11). http://media. affymetrix. com/support/technical/whitepapers/exon alt transcript analysis whitepaper. pdf.

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

[11] 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, 14:1470-1479.

[12] Liu X, Milo M, Lawrence N D, et al. A tractable probabilisticmodel for Affymetrix probe-level analysis across multiple chips[J]. Bioinformatics, 2005, 21: 3637-3644.

[13] Wu Z, Irizarry R A, Gentleman R, et al. A model-based background adjustment for oligonucleotide expression arrays[J]. Journal of the American Statistical Association, 2004, 99:909-917.

[14] Sun J, Kabán A, Raychaudhury S. Robust mixtures in the presence of measurement errors[C]// Proceedings of the 24th International Conference on Machine Learning. Corvalis, Oregon:[s. n.], 2007: 847-854.

[15] 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, 9:42-53.

[16] 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.

[17] 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.

[18] 赵志兰，刘学军，张礼. 一种基于概率模型的 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.
Zhao Zhilan, Liu Xuejun, Zhang Li. A probabilistic model for the analysis of RNA-Seq data [C]// The proceedings of CBME'2011 (CD). Wuhan: Chinese Society of Biomedical Engineering, 2011.

[19] Risueño A, Fontanillo C, Dinger M E, et al. GATExplorer: genomic and transcriptomic explorer; mapping expression probe to gene loci, transcripts, Exons and ncRNAs [J]. BMC Bioinformatics, 2010, 11:221.

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

[21] 严德春，王加俊. 改进的稳健 Lowess 标准化算法在基因芯片中的应用[J]. 数据采集与处理，2013，28(1)：82-86.
Yan Dechun, Wang Jiajun. Improved robust Lowess normalization method in analysis of gene chip [J]. Journal of Data Acquisition and Processing, 2013, 28(1): 82-86.

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

[23] Zak D E, Andersen-Nissen 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.

[24] 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.

[25] Zhang L, Liu X. An improved probabilisticmodel for finding differential gene expression[C]// Proceedings of the 2nd International Conference on BioMedical Engineering and Informatics, BMEI 2009. Tianjin, China:[s. n.], 2009.

[26] Agesen T H, Sveen A, Merok M A, et al. ColoGuideEx: A robust gene classifier specific for stage II colorectal cancer prognosis [J]. Gut, 2012, 61(11): 1560-1567.

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

作者简介：刘学军(1976-)，女，副教授，研究方向：生物信息学与机器学习，E-mail: xuejun. liu@nuaa. edu. cn；张武军(1989-)，男，硕士研究生，研究方向：生物信息学；张礼(1985-)，男，博士研究生，研究方向：生物信息学。