第60卷第1期　　大连理工大学学报　　Vol. 60, No. 1
2020年1月　　Journal of Dalian University of Technology　　Jan. 2020

文章编号：1000-8608(2020)01-0069-06

基于广义线性模型的基因表达水平预测

师豪杰，顾　宏，徐晓璐，秦　攀*

（大连理工大学 控制科学与工程学院，辽宁 大连　116024）

摘要： 组蛋白修饰是生物体中普遍存在的一种现象，能够以不同的调控方式影响基因表达，且随着高通量测序技术的飞速发展，大量的测序数据使得探究组蛋白修饰信号与基因表达水平之间的内在联系成为可能．由于基因表达数据存在零膨胀现象，提出了一种基于广义线性模型框架的主从模型，能够以较高精度从组蛋白修饰信号预测基因表达水平．首先通过人类全基因组注释文件中的基因位点信息，筛选出包含完整基因位点信息的表达数据；其次，根据基因位点信息，定位并提取出组蛋白修饰数据中基因特定位点的特征信息，构建设计矩阵；最后结合响应变量数据零膨胀的特点，构建主从模型，以GM12878细胞系为例，与现有的多种回归算法进行对比，验证了所提模型的有效性．

关键词： 广义线性模型；主从模型；组蛋白修饰；基因表达

中图分类号： Q-34；TP301.6　　**文献标识码：** A　　**doi：** 10.7511/dllgxb202001010

0　引　言

在真核生物中，核小体是生物遗传物质染色质的基本组成单位，它们由DNA和核心组蛋白结合构成[1]．这些核心组蛋白的结构容易发生改变，包括组蛋白甲基化、乙酰化、磷酸化、泛素化等，这些改变统称为组蛋白修饰[2-3]．组蛋白修饰会影响生物体中遗传物质的表达，其影响基因表达的机制一般有两种：一是改变DNA的空间结构，从而改变转录因子对DNA的可及性；二是为转录激活因子和抑制因子的募集提供特异性的结合表面，从而影响基因表达[4]．不同于DNA突变，组蛋白修饰是可逆转的，这种差异使得研究组蛋白修饰对基因表达的调控机制具有重要意义，研究组蛋白修饰可以进一步推动癌症等疾病的表观遗传药物研究的进展[5-7]．

目前针对组蛋白修饰调控基因表达的研究主要分为两类，一类是通过传统的机器学习方法探究组蛋白修饰与基因表达的内在联系．例如Cheng等在2011年首次提出了提取组蛋白修饰特征的方法，并通过双向层次聚类、支持向量机（support vector machine，SVM）以及支持向量回归（support vector regression，SVR）研究了组蛋白修饰的内在联系以及对基因表达水平高低的影响[8]；Dong等于2012年通过随机森林（random forest）对组蛋白修饰、组蛋白变体以及DNA超敏感位点对基因表达水平高低的影响进行了研究[9]．另一类是通过神经网络和深度学习对组蛋白作用机制进行探究．例如Singh等于2016年基于卷积神经网络研究了5种组蛋白修饰特征对基因表达的调控机制[10]；并于2017年通过基于注意力的深度学习方法研究了人体56种细胞类型染色质标记和基因表达之间的关系[11]．这些研究大都是以基因表达水平的中位数为衡量标准，将基因表达划分为高表达和低表达，对基因表达水平的高低进行分类建模，并达到了较高的精度．相对于前述的分类模型，对基因表达的实值数值建立回归预测模型的难度较大，研究成果相对较少．如2011年Cheng等通过SVR预测基因表达，其相关系数仅达到0.7．从上述论文的研究成果来看，基因表达的实值数值预测精度仍有提升的空间，通过改善回归预测模型能进一步明确不同基因位点对基因表达的影响程度．

收稿日期： 2019-08-09；　**修回日期：** 2019-11-08．
基金项目： 国家自然科学基金资助项目(81872247)．
作者简介： 师豪杰(1995-)，男，硕士生，E-mail：shihj0511@163.com；顾　宏(1961-)，男，教授，博士生导师，E-mail：guhong@dlut.edu.cn；秦　攀*(1978-)，男，副教授，硕士生导师，E-mail：qp112cn@dlut.edu.cn．

针对上述问题，本文以人体B淋巴(GM12878)细胞系作为研究对象，该细胞系主要执行机体的体液免疫功能，基因表达较为活跃，适合于基因表达水平的研究，且前述论文均对此细胞系进行了深入的研究. 以每百万读取比对的转录本数(transcripts per million，TPM)作为衡量基因表达水平高低的标准[12]，由于生物体中在同一时刻有很大一部分基因并未进行表达，TPM表达数据包含大量零值，此类零膨胀数据会降低回归模型的预测精度. 为此，本文结合基因表达数据中零膨胀的特点[13]，在广义线性模型(generalized linear model，GLM)[14]框架下构建一个主从模型：第一步先构建一个基于二项分布的广义线性模型，对组蛋白修饰特征是否对基因的表达产生影响进行分类；第二步对产生影响的基因构建一个基于正态分布的GLM模型，预测基因的表达值，从而对基因的表达水平以及组蛋白修饰作用的机制有一个更准确的认识.

1 数据来源及预处理

1.1 数据来源

本文中所采用的组蛋白修饰数据和基因表达数据(RNA-seq)均来自ENCODE(encyclopedia of DNA elements)数据库，即DNA元素百科全书[15]. ENCODE旨在构建一个全面的人类基因组功能元件清单，包括在蛋白质、DNA和RNA水平上作用的元件，以及控制细胞基因表达活跃或沉默的调控元件等. ENCODE数据库整合了大量的多组学的实验技术和方法，囊括了最完善的表观遗传学、转录组学等多种组学数据，为对各种数据的进一步研究提供了丰富的数据资源.

人类参考基因组(hg19)注释数据来自于UCSC(http://genome.ucsc.edu/)数据库，该注释文件包含了染色体编号(chromosome number)、基因名称(gene name)、基因ENSEMBL-ID、基因类型(gene type)、正负链(strand)、转录起始位点(transcription start sites，TSS)、转录终止位点(transcription terminal sites，TTS)等信息. 这些信息完整地表示了一个基因所在位置、长度、染色体编号等信息，对组蛋白修饰数据后续的基因定位、基因数据合并等具有重要作用.

本文选取GM12878细胞系作为研究对象，选取了该细胞系的10组组蛋白修饰数据、RNA-seq数据以及hg19注释文件进行分析研究. 表1详细列出了本文中所使用的10种组蛋白修饰，并对这10种组蛋白修饰所在区域以及该组蛋白的功能类别进行简要介绍[16-17].

表1 组蛋白修饰功能介绍

Tab. 1 Introduction of histone modification function

组蛋白修饰名称	关联区域	功能类别
H3k9me3	异染色质区域	抑制作用
H3k9ac	增强子区域、启动子区域	激活作用
H3k36me3	转录区域	结构标记
H3k4me1	增强子区域	远端作用标记
H3k4me2	转录区域	抑制作用
H3k4me3	启动子区域	激活作用
H3k79me2	增强子区域、启动子区域	抑制作用
H3k27me3	多梳蛋白抑制区域	抑制作用
H4k20me1	转录区域	抑制作用
H3k27ac	增强子区域、启动子区域	激活作用

1.2 数据预处理

hg19注释文件的处理，首先保留参考基因组中的编码蛋白基因(protein coding)，该部分基因可以指导蛋白质的合成，同时保留表达类型为外显子(exon)的基因，即编码相应RNA外显子的DNA中的区域. 其次，对相同ID的基因数据进行去重，只保留其中一条，并去除基因中后续分析不需要的部分信息，最终剩余22 642个基因的信息用于后续分析.

由于RNA-seq数据集中的基因通过ENSEMBL-ID来表示，该类型的基因标识符是由ENSEMBL基因组数据库项目进行标注命名的，通用R语言(https://www.r-project.org/)程序包org.Hs.eg.dg将基因的ENSEMBL-ID转换为常用的基因名称. 然后提取RNA-seq数据中每个基因对应的TPM值，记为T_{pm}，该值是当前较为标准的衡量基因表达水平的数值，是一种优化的RPKM计算方法，可以用于同一物种不同组织基因表达水平的比较，其计算公式如下：

$$T_{\mathrm{pm}}=\frac{r_g\times r_l\times 10^6}{l_g\times T} \tag{1}$$

$$T=\sum_{g\in G}\frac{r_g\times r_l}{l_g} \tag{2}$$

式中：r_l表示测序读长，即每次测序读取比对的平均核苷酸数目；l_g为基因长度；r_g为读取比对到相应基因的总数目.

通过对hg19注释文件的处理，得到了人体中22 642个基因的注释信息，包括基因转录起始位点、转录终止位点、染色体编号以及正负链等信

息. 由于少数基因的注释信息有部分缺失，在后续数据处理过程中无法对该基因在组蛋白修饰数据上进行定位，进而无法提取对应基因位置的组蛋白修饰特征，故而删除该部分基因数据. 最后，本文将基因表达数据与基因注释信息进行比对，筛选出 20 114 个注释信息完整的基因，生成一个 20 114 维的响应变量的实现值向量 $\boldsymbol{y}$.

组蛋白修饰的原始数据为 ENCODE 数据库中的 Bigwig(bw)文件，共有 10 组数据. bw 文件是由 Wiggle(wig)文件格式转换而来的二进制压缩文件，相较于 wig 文件更加节省空间，但无法直接使用其提供的信息. wig 文件是一种用于展示连续值数据的格式，如组蛋白修饰、转录组数据等，表示基因组一个区域内信号的高低，更便于提取组蛋白修饰在不同基因位点的特征值. 本文使用 UCSC 提供的格式转换工具 BigwigToWig 将该 bw 格式数据转换为步长为 25 bp 的 wig 文件，然后通过求和取平均值的方法将该文件转换为步长为 50 bp 的 wig 文件，为后续更方便地提取基因位点信息做好准备.

由于基因起始位点和终止位点周围的组蛋白修饰对基因表达的作用较为显著，且其作用方式为长程效应，即较长距离的组蛋白修饰都会对基因表达产生影响. 由于基因长度的限制，为了保证基因样本的数量，本研究只提取了基因起始位点前和基因终止位点后 4 000 bp 的组蛋白修饰特征. 组蛋白修饰特征的具体提取方法如图 1 所示，首先通过 hg19 注释文件对每个基因进行定位，根据每个基因的 TSS 和 TTS，将每个基因 TSS 前 4 000 bp 的基因序列划分为 40 个箱，TTS 后 4 000 bp序列也划分为 40 个箱，每个箱长度为 100 bp，每个基因共对应 80 个箱. 考虑到基因长度的限制，本次提取的特征不包含基因内部的组蛋白修饰特征提取. 通过组蛋白修饰文件提取对应位点的特征信息，每 100 bp 提取一个值，将该值作为该位点的特征. 每种组蛋白提取了转录起始位点前 40 个特征，终止位点后 40 个特征，共 80 个特征，合并 10 组组蛋白的特征数据，生成一个包含定常项的 20 114×801 的设计矩阵 $\boldsymbol{X}$.

通过对 RNA-seq 数据、hg19 注释文件以及 10 组组蛋白修饰数据进行数据预处理，最终得到了一个 20 114 × 801 的设计矩阵 $\boldsymbol{X}$ 和长度为 20 114维的响应变量 $\boldsymbol{y}$，用于后续的建模分析.

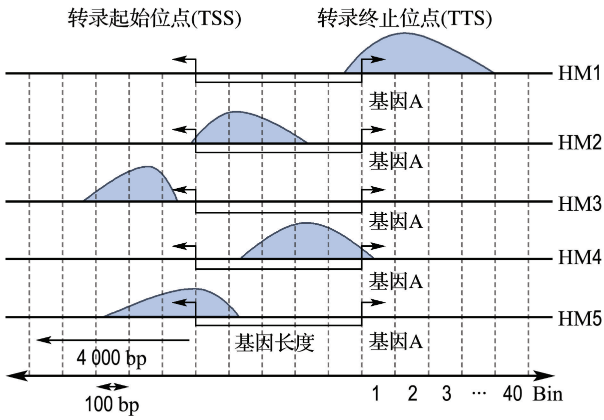

图 1　组蛋白修饰特征提取示意图

Fig. 1　Sketch map of histone modified feature extraction

2　模型构建

通过之前的数据预处理，已经整合出了一个维数为 20 114×801 的设计矩阵 $\boldsymbol{X}$，以及对应的响应变量 $\boldsymbol{y}$. 其中 $\boldsymbol{x}_i \in \boldsymbol{R}^{801}$，对应样本的解释变量，即 $\boldsymbol{x}_i$ 表示某个基因对应的特征值；y_i 表示对应样本的响应变量，即对应基因的总体表达水平值.

由于响应变量中零值所占比例较大(25.3%)，为典型零膨胀数据. 本文在 GLM 框架下提出了一种主从模型，将建模过程分为两个过程对该数据集进行建模分析. 第一个过程先对数据中的零值和非零值进行分类，用来确定“零”是否出现，根据响应变量是否为零，给数据添加 0-1 标签，响应变量值为零则添加 0 标签，否则添加 1 标签，通过构建一个基于二项分布的 GLM 来进行分类，对响应变量中是否出现“零”值进行分类. 当第一个过程分类结果为 1 时，则进入第二个过程，此时响应变量数据为连续型实值，选取适当的分布形式，建立一个基于正态分布的 GLM 对该部分数据进行回归分析.

在对数据进行建模过程中，首先对特征矩阵数据进行分析，图 2 为特征矩阵中 10 个特征在每个样本处的箱形图，每种组蛋白选取了一个特征. 由于其数值分布范围较广，需要对特征矩阵进行标准化处理，使得 $x_{ij} \in [0,1]$，以解决不同数据之间存在不可比性的问题，并消除了奇异样本数据导致的不良影响. 具体标准化方法如下：

$$x_{ij} = \frac{\log_2(x_{ij}+1)}{\log_2(\max(x_j)+1)} \tag{3}$$

第一过程：通过响应变量值给每个样本添加标签(0 或 1)，将响应变量值 y_i 不为零的样本标记为正类，用 1 表示，响应变量值 y_i 为零的样本

标记为负类,用 0 表示.通过 GLM 对该数据集构建一个分类模型.

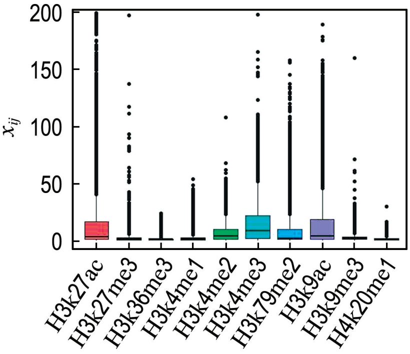

图 2 标准化前组蛋白修饰特征值分布箱形图

Fig. 2 Box-plot of histone modified feature before standardization

根据之前给定的标签值,响应变量值 y_i 服从二值分布,对于任意一个样本 $\boldsymbol{x}_i$,其对应的标签值为 y_i,则每个基因样本是否表达的分布函数如下:

$$f_y(y_i|\nu_i)=P(Y=y_i|\nu_i)=\nu_i^{y_i}(1-\nu_i)^{1-y_i} \tag{4}$$

其中 $y_i\in\{0,1\}$,$0<\nu_i<1$,并且 $E(Y)=\nu_i$,$Var(Y)=\nu_i(1-\nu_i)$.以逻辑函数为其链接函数,即:

$$g(\nu_i)=\ln\left(\frac{\nu_i}{1-\nu_i}\right)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 \tag{5}$$

其中 $\boldsymbol{\beta}_1$ 为预测二值分布参数 ν_i 的系数向量,将所有样本的分布函数进行连乘得到其似然函数:

$$L(\boldsymbol{\beta}_1)=\prod_{i=1}^{n}f_y(y_i|\nu_i)=\prod_{i=1}^{n}\nu_i^{y_i}(1-\nu_i)^{1-y_i} \tag{6}$$

对上式取负对数运算,得到其负的对数似然函数,如下式所示:

$$(\boldsymbol{\beta}_1)=-\ln\prod_{i=1}^{n}f_y(y_i|\nu_i)=$$
$$-\sum_{i=1}^{n}(y_i\ln\nu_i+(1-y_i)\ln(1-\nu_i)) \tag{7}$$

对上式求偏导并置零可得:

$$\frac{\partial}{\partial\nu_i}=-\frac{y_i}{\nu_i}+\frac{1-y_i}{1-\nu_i}=0 \tag{8}$$

其中 $\nu_i=\dfrac{1}{1+\exp(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1)}$,解出分类模型的参数向量 $\boldsymbol{\beta}_1$,即可构建好第一过程中的分类模型.

第二过程:通过构建好的分类模型,提取出正类数据,对该部分数据进行回归建模,由于响应变量中还存在小部分零值,且非零数据分布不均匀,对其进行变换,使得响应变量值 y_i 大致符合正态分布,利于对该部分数据进一步建模分析,变换函数为

$$y_i=\ln\left(\frac{y_i+1}{\mathrm{median}(y_i)+1}\right) \tag{9}$$

其中 median(y_i)表示 y_i 的样本中位数.

图 3 为对标签为 1 的响应变量值 y_i(即 TPM 值)进行变换后的分布图.

由图 3 可知,该部分数据大致呈现正态分布形式,所以通过基于正态分布的 GLM,对该部分数据进行建模,其分布函数如下所示:

$$f(y_i|\mu_i,\sigma_i^2)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma_i^2}\right) \tag{10}$$

其中 μ_i 为第 i 个样本所服从分布的均值,σ_i^2 为该分布的方差.通过链接函数将正态分布的参数 μ_i 与对应第 i 个样本的解释变量 $\boldsymbol{x}_i$ 相联系,本研究并未估计方差,其链接函数如下所示:

$$g(\mu_i)=\mu_i=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 \tag{11}$$

其中 $\boldsymbol{\beta}_2$ 为预测正态分布参数 μ_i 的系数向量,$\boldsymbol{x}_i$ 为单个样本的全部解释变量构成的向量.同样通过连乘可以得到基于正态分布的 GLM 关于期望 μ_i 的似然函数,并通过最大似然法求出 $\boldsymbol{\beta}_2$,其似然函数形式如下:

$$L=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(\frac{-(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2)^2}{2\sigma_i^2}\right) \tag{12}$$

求解结果为

$$\hat{\boldsymbol{\beta}}_2=\mathrm{argmax}\ L\ (\boldsymbol{\beta}_2) \tag{13}$$

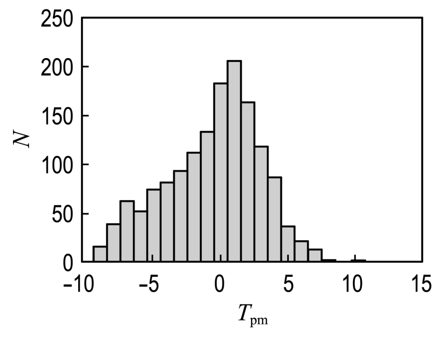

图 3 TPM 值分布直方图

Fig. 3 Distribution histogram of TPM

3 实验结果与分析

本文研究了人体 GM12878 细胞系的 10 组组蛋白修饰特征数据和基因表达数据,通过构建一个基于 GLM 框架的主从模型对该数据集进行回归分析.

第一部分主模型先对响应变量是否为零进行分类,计算分类精确度等信息量并画出 ROC 曲线,作为分类精度的性能评价指标.其中精确度计算公式如下:

$$P=T_{\mathrm{p}}/(F_{\mathrm{n}}+F_{\mathrm{p}}) \tag{14}$$

式中：T_p 表示将正类预测为正类数，F_n 表示将正类预测为负类数，F_p 表示将负类预测为正类数.

第二部分从模型对响应变量中分类不为零的数据集构建回归模型，计算模型的决定系数 R^2，用来衡量该模型的拟合精度，并对实验结果进行进一步的讨论. 其计算公式如下：

$$R^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \Big/ \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{15}$$

对于建立好的分类模型，本文通过 ROC 曲线来衡量模型的分类效果，通过十折交叉验证，得出其分类模型的精确度为 0.898，说明本文所构建的分类模型具有较高的精确度，能够较为准确地判断组蛋白修饰信号是否对基因表达有促进作用. 图 4 为本文所构建分类模型的 ROC 曲线，其中 R_t 为真阳性率，R_f 为假阳性率.

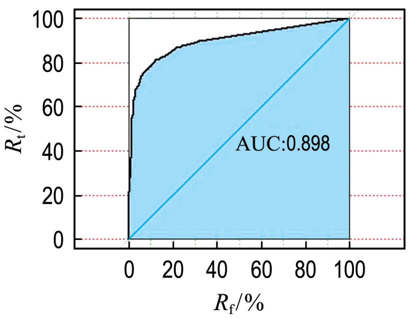

图 4　分类模型 ROC 曲线

Fig. 4　ROC curve of classification model

由于目前对组蛋白修饰数据影响基因表达水平的研究大都以分类为主，即通过基因表达水平的中位数来界定基因表达水平的高低，然后构建一个分类模型来对表达水平高低进行分类. 但是预测基因表达水平值的研究相对较少，所以本文通过对比多元线性回归（multiple linear regression，MLR）、SVR 和神经网络（neural_network）这些常用的建模方法对基因表达数据进行回归分析，并对比最终结果，证明本文所提出的 GLM 框架下的主从模型的有效性. 图 5 为多种回归算法拟合结果对比图.

由图 5 可知，本文所提出的 GLM 主从模型在相关系数 C 和决定系数 R^2 方面相比于其他回归算法有较大的提升，其相关系数为 0.873，决定系数达到 0.749，并且其均方误差值为 0.044，得到了较好的预测结果，证明该模型在该数据集具有较好的拟合效果. 其他 3 种回归算法中，MLR 结果较差，表明基因的表达水平和组蛋白修饰特征之间的线性关系不显著；神经网络的结果相对于 SVR 有较大的提升，证明将神经网络运用于该研究上有一定的可行性；在对 SVR 模型进行回归预测时，本文也比较了在回归预测之前加入 SVM 分类过程对最终预测结果的拟合程度是否具有显著差别，其分类前后进行回归预测的结果分别为 0.460 和 0.495，说明该算法并不适用于零膨胀型数据且提升空间较小. 本文所提出的主从模型充分考虑了基因表达数据中零膨胀的现象，在建模初期便对该现象进行分析研究，所以得到了更好的拟合效果，更好地明确了组蛋白修饰在基因表达过程中的作用.

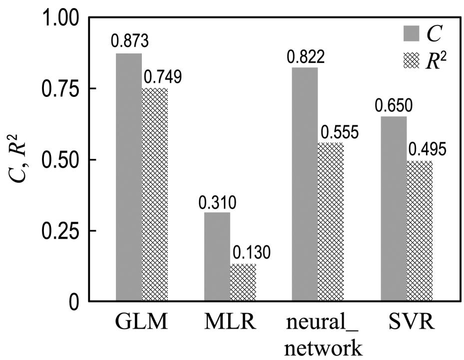

图 5　不同算法结果比较

Fig. 5　Results comparison of different algorithms

4　结　语

本文综合分析了人类全基因组注释数据、GM12878 细胞系的组蛋白修饰数据及该细胞系基因表达数据，不同于先前对基因表达水平高低构建分类模型的研究，构建了通过组蛋白修饰特征预测基因表达的主从模型，对基因表达水平进行回归分析. 通过提取每个基因特定位点的组蛋白修饰特征值，结合基因表达值的零膨胀特点，构建了一个基于 GLM 框架的主从模型，对组蛋白修饰特征对基因表达的影响进行建模分析，并得到了较好的拟合效果，这对研究者们进一步明确组蛋白修饰在调控基因表达过程中发挥的作用具有较大帮助.

参考文献：

[1]　DONG Xianjun，WENG Zhiping. The correlation between histone modifications and gene expression [J]. **Epigenomics**，2013，**5**(2)：113-116.

[2]　PETERSON C L，LANIEL M-A. Histones and histone modifications [J]. **Current Biology**，2004，**14**(14)：R546-R551.

[3] ROTH S Y, DENU J M, ALLIS C D. Histone acetyltransferases [J]. **Annual Review of Biochemistry**, 2001, **70**(1): 81-120.

[4] LI Bing, CAREY M, WORKMAN J L. The role of chromatin during transcription [J]. **Cell**, 2007, **128**(4): 707-719.

[5] BANNISTER A J, KOUZARIDES T. Regulation of chromatin by histone modifications [J]. **Cell Research**, 2011, **21**(3): 381-395.

[6] HERCEG Z. Epigenetics and cancer: towards an evaluation of the impact of environmental and dietary factors [J]. **Mutagenesis**, 2007, **22**(2): 91-103.

[7] SHANG Yongfeng. Epigenetics and cancer [C] // **Abstract of Academic Conference and 10th Genera Congress of the Chinese Society of Biochemistry and Molecular Biology**. Nanjing: The Chinese Society of Biochemistry and Molecular Biology, 2010: 34.

[8] CHENG Chao, YAN K K, YIP K Y, *et al*. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets [J]. **Genome Biology**, 2011, **12**(2): R15.

[9] DONG Xianjun, GREVEN M C, KUNDAJE A, *et al*. Modeling gene expression using chromatin features in various cellular contexts [J]. **Genome Biology**, 2012, **13**(9): R53.

[10] SINGH R, LANCHANTIN J, ROBINS G, *et al*. Deep chrome: deep-learning for predicting gene expression from histone modifications [J]. **Bioinformatics**, 2016, **32**(17): 639-648.

[11] SINGH R, LANCHANTIN J, SEKHON A, *et al*. Attend and predict: understanding gene regulation by selective attention on chromatin [J]. **Advances in Neural Information Processing Systems**, 2017, **30**: 6785-6795.

[12] WAGNER G P, KIN K, LYNCH V J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples [J]. **Theory in Biosciences**, 2012, **131**(4): 281-285.

[13] LAMBERT D. Zero-inflated Poisson regression, with an application to defects in manufacturing [J]. **Technometrics**, 1992, **34**(1): 1-14.

[14] NELDER J A, WEDDERBURN R W M. Generalized linear models [J]. **Journal of the Royal Statistical Society**, 1972, **135**(3): 370-384.

[15] The ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project [J]. **Science**, 2004, **306**(5696): 636-640.

[16] KUNDAJE A, MEULEMAN W, ERNST J, *et al*. Integrative analysis of 111 reference human epigenomes [J]. **Nature**, 2015, **518**(7539): 317-330.

[17] GRUNSTEIN M. Histone acetylation in chromatin structure and transcription [J]. **Nature**, 1997, **389**(6649): 349-352.

Prediction of gene expression level based on generalized linear model

SHI Haojie, GU Hong, XU Xiaolu, QIN Pan*

(School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Histone modification is a common phenomenon in organisms, which can affect gene expression in various ways. With the rapid development of high-throughput sequencing technology, adequate sequencing data make it possible to explore the relation between histone modification and gene expression level. A master-slave model based on the generalized linear model framework is proposed, which can predict gene expression levels from histone modification signals with high precision. Firstly, gene locus information from the human genome-wide annotation file is used to screen out the expression data which contain the complete gene locus information. Secondly, according to the gene locus information, the characteristics of the gene-specific locus in the histone modification data are located and extracted, and then the design matrix is constructed. Finally, combined with the zero-expansion characteristics of the response variable data, the master-slave model is constructed, then compared with the existing multiple regression algorithms by using the data of GM12878 cell line, the validity of the proposed model is proved.

Key words: generalized linear model; master-slave model; histone modification; gene expression