第 37 卷 第 2 期
Vol.37 No.2
计 算 机 工 程
Computer Engineering
2011 年 1 月
January 2011

• 博士论文 • 文章编号：1000—3428(2011)02—0020—02 文献标识码：A 中图分类号：R318.04

基因芯片数据多尺度分析的 *y~n* 曲线模型

罗万春[1a]，易 东[1b]，李辉智[2]，龚利红[3]

(1. 第三军医大学 a. 数学教研室；b. 卫生统计学教研室，重庆 400038；2. 西南政法大学刑侦学院，重庆 400031；
3. 重庆通信学院机电教研室，重庆 400035)

摘 要：针对离散的基因芯片表达数据难以进行系统分析的问题，提出基因芯片数据的 *y~n* 曲线模型，将基因芯片数据转化为唯一的一列信号，从而可以对其进行多尺度分析。对人类胎儿脑发育过程中大脑皮质基因表达基于 *y~n* 曲线模型进行小波多尺度降噪，结果证明，*y~n* 曲线的降噪效果比用基因芯片原始数据更好。

关键词：基因芯片；*y~n* 曲线；多尺度分析

y~n Curve Model of Multi-scale Analysis to DNA Microarray Data

LUO Wan-chun[1a], YI Dong[1b], LI Hui-zhi[2], GONG Li-hong[3]

(1a. Department of Mathematics; 1b. Department of Health Statistics, Third Military Medical University, Chongqing 400038, China;
2. Criminal College, Southwest University of Political Science and Law, Chongqing 400031, China;
3. Department of Machine and Electron, Chongqing Communication College, Chongqing 400035, China)

【Abstract】 The model of *y~n* curve is set up to solve the problem that discrete gene chip data are difficult to be systematically analyzed, which makes the data be changed into a row of sole signal, and can be studied by multi-scale analysis. Results of denoising the cerebral cortex gene expression of developing human brain based on wavelet multi-scale analysis show that the denoised effect of *y~n* curve is better than that of raw data.

【Key words】 DNA microarray; *y~n* curve; multi-scale analysis

DOI: 10.3969/j.issn.1000-3428.2011.02.007

1 概述

基因芯片技术对基因的大规模研究有重要意义。每张芯片上的基因数量成千上万，实验研究往往选择几个时间点，从而通过实验可以获得基因在不同时间点的表达强度。这些基因表达数据往往含有噪声[1]，因此，有必要研究一种可以对基因表达数据进行降噪的方法。多尺度分析是降噪的有效工具，但基因芯片数据的时间点往往很少，难以应用多尺度分析进行时频信号处理，每一个时间点(一张基因芯片)的基因数量很大，基因之间并不存在明确的函数关系。因此，本文尝试建立基因表达的数学模型，使得基因表达强度可以较好地进行多尺度降噪。

2 基因芯片数据分析的 *y~n* 曲线模型

微阵列技术可以同时检测芯片上所有基因的表达强度，表达强度的数据采集一般按照基因在芯片上的位置顺序进行。基因芯片实验一般选择几个时间点以观察研究对象随时间变化的特点，不妨记试验次数为 T，数据预处理后的有效表达基因个数为 N，则获得的有效基因芯片实验数据矩阵 $\boldsymbol{A}=(x_{ij})_{N\times T}$。为使排列方法唯一，将基因的表达均值向量 $\boldsymbol{A}_{\text{mean}}=(\frac{1}{T}\sum_{j=1}^{T}x_{ij})_{N\times 1}$ 按照从小到大的顺序排序，并按此顺序将矩阵 $\boldsymbol{A}$ 重新排序，得到新的基因表达矩阵：

$$\boldsymbol{A}_{\text{new}}=(S_{ij})_{N\times T} \tag{1}$$

其中，S_{ij} 为重新排序后第 i 个基因在时间点 j 的表达强度。

在矩阵 $\boldsymbol{A}_{\text{new}}$ 中，每一列为一张基因芯片的基因表达数据。为描述方便，不失一般性，取出 $\boldsymbol{A}_{\text{new}}$ 中的任意一列，并记第 n 个基因表达强度为 S_n，f_n 为前面 n 个基因表达强度的叠加，建立如下模型：

$$\begin{cases} f_n = f_{n-1}+s_n \\ f_0 = 0 \end{cases} \quad n=1,2,\cdots,N \tag{2}$$

对 f_n 作多项式拟合，拟合函数为：

$$\hat{y}=\sum_{k=0}^{m}a_k n^k \tag{3}$$

其中，多项式的阶数 m 可以根据数据特点适当调整，一般当 m=3 或 m=4 时可达到较好的拟合效果。

定义 称如下函数刻划的曲线为基因芯片数据分析的 *y~n* 曲线，f_n、$\sum_{k=0}^{m}a_k n^k$ 分别满足式(2)和式(3)，则：

$$y(n)=f_n-\sum_{k=0}^{m}a_k n^k,\quad n=1,2,\cdots,N \tag{4}$$

综上所述，*y~n* 曲线的构造步骤为：

(1)排序，按式(1)得到排序后的基因表达数据；

(2)求 f_n，按式(2)求出 f_n；

(3)得到 *y~n* 曲线，按式(4)建立 *y~n* 曲线。

3 *y~n* 曲线的性质

性质 1(存在性) 对于任一基因芯片任一时刻测出的基因表达数据序列 $\{S_n\}$，总存在一条 *y~n* 曲线。

证明：对于任意的一张基因芯片，无论何物种，只要存在基因 k 的表达强度 $S_n\neq 0$，则 $f_n\neq c$，(c 为常数)。只要 f_n(n=1,

基金项目：国家自然科学基金资助项目“基因时空表达信息多尺度分析方法研究”(30872184)

作者简介：罗万春(1971－)，男，副教授、博士，主研方向：生物医学预测，优化建模；易 东，教授、博士生导师；李辉智，副教授、硕士；龚利红，讲师、硕士

收稿日期：2010-05-24 **E-mail：**aluo@tmmu.edu.cn

2,…, N)中存在 3 点不共线，则 $y(n)\neq 0$。按式(1)获得的基因表达数据，总可以取得 3 个点，使得 $f_1=S_1$、$f_2=S_1+S_2$、$f_3=S_1+S_2+S_3$ 3 点不共线，即：

$$\begin{vmatrix}1 & f_1 & 1\\ 2 & f_2 & 1\\ 3 & f_3 & 1\end{vmatrix}=\begin{vmatrix}1 & s_1 & 1\\ 2 & s_1+s_2 & 1\\ 3 & s_1+s_2+s_3 & 1\end{vmatrix}\neq 0$$

展开行列式，得到 $S_3-S_2\neq 0$，亦即 $S_3\neq S_2$。即只要有 2 个基因表达强度不相等即可，这对于需要研究的基因芯片数据来说是可以做到的，否则没有研究的必要。

性质 2(唯一性) 对于任一基因芯片任一时刻测出的基因表达数据 S_n，若存在相应的 y~n 曲线，则该曲线必定唯一。

性质 3(有界性) 任一基因芯片的 y~n 曲线，y 必定是关于 n 的有界函数。

性质 4(离散性) 任一基因芯片的 y~n 曲线模型必定具有离散性。

性质 5(零均值性) 若 $y(n)$满足式(4)，则：$\frac{1}{N}\sum_{n=1}^{N} y(n)=0$。

性质 6(等价性) 由 S_n(n=1, 2, …, N)表示的一列基因表达数据和 y~n 曲线等价。

4　应用举例

4.1　胎儿大脑皮质的基因表达数据集

数据来源于香港城市大学生物化学系曾志雄博士的基因芯片实验数据[2]。为了研究人类胎儿脑发育过程中大脑皮质的基因表达和基因之间的调控关系，从医院收集非疾病原因终止妊娠的胎儿，按照不同的发育阶段以 4 周为周期分为 7 个时期(孕期分别为 9 周、12 周、16 周、……、36 周)。从胎儿大脑皮质提取 mRNA，并标记荧光染料 Cy5，对照物标记荧光染料 Cy3。将样品与对照等量混合，与 10 080 个点的高密度基因芯片杂交实验，得到按照空间位置排列关系的 10 080×7 个 *ratio* 值矩阵：

$$ratio_{ij}=\frac{R_f-R_b}{G_f-G_b}$$

其中，R_f、R_b 分别表示红色前景荧光强度和背景荧光强度；G_f、G_b 分别表示绿色前景荧光强度和背景荧光强度，i=1, 2,…, 10 080, j=1,2,…,7。

第 i 个基因的第 j 个时间点的表达强度[3]为 $x_{ij}=\mathrm{lb}ratio_{ij}$，因此，对第 i 个基因，必须 7 个时间点的 *ratio* 值均大于 0，以此剔除无关基因，获得新的基因表达矩阵$(x_{ij})_{956\times 7}$，即基因表达有意义的 956 个基因在 7 个时间点的表达强度。

4.2　大脑皮质基因的 y~n 曲线

在多项式拟合过程中，本文选取 m=4，并用可决系数 R^2 刻划拟合效果，R^2 越接近于 1，表明标准误差越小。7 张芯片表达数据曲线拟合的可决系数 R^2 分别为：0.996 8, 0.981 0, 0.990 8, 0.974 3, 0.996 9, 0.981 9, 0.950 6，拟合效果较好。y~n 曲线如图 1~图 3 所示(限于篇幅，只列出前 3 张图)。

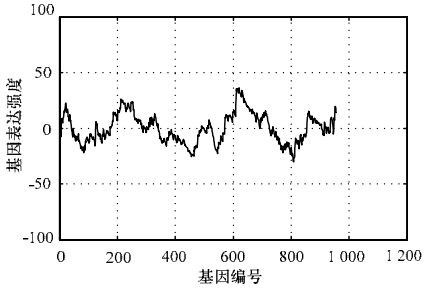


图 1　大脑皮质第 1 张基因芯片的 y~n 曲线

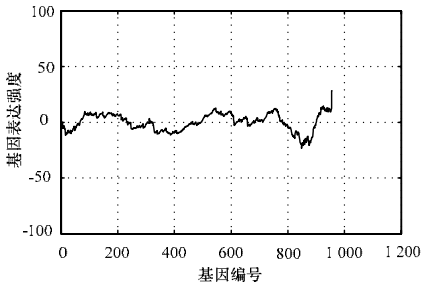


图 2　大脑皮质第 2 张基因芯片的 y~n 曲线

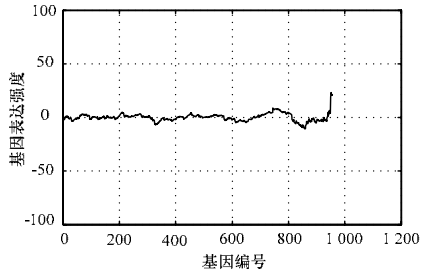


图 3　大脑皮质第 3 张基因芯片的 y~n 曲线

4.3　多尺度降噪的结果

对 y~n 曲线进行多尺度降噪并还原成基因芯片数据，最后通过信噪比[4]来刻画分析降噪效果，信噪比定义为：

$$S_{\mathrm{nr}}=10\ln\left(\frac{\|X\|^2}{\|X-S\|^2}\right)$$

其中，$\|X\|$是长度为 N 的序列 $X=\{x_n\}$的范数，此处定义为 $\|X\|=(\sum_{n=1}^{N}x_n^{\,2})^{\frac{1}{2}}$；$S$ 为重构信号。信噪比越大，说明降噪效果越好。

笔者曾对胎儿大脑皮质的基因芯片原始数据做了降噪效果分析，结果比较理想[5]，并进一步在 y~n 曲线模型的基础上作多尺度降噪。依次选择了常见的 16 种小波函数 db1(sym1, Haar)、db2(sym2)、db3(sym3)、db4、db5、db6、db7、db8、db9、db10、sym4、sym5、bior1.3、bior2.4、coif1、coif2 做多尺度分析，按照频率分解为 10 层，将原始芯片数据与 y~n 曲线模型获得的 7 张芯片数据的信噪比均值作比较，结果如图 4 所示。所有数据处理均用 Matlab R2007a 编程实现。

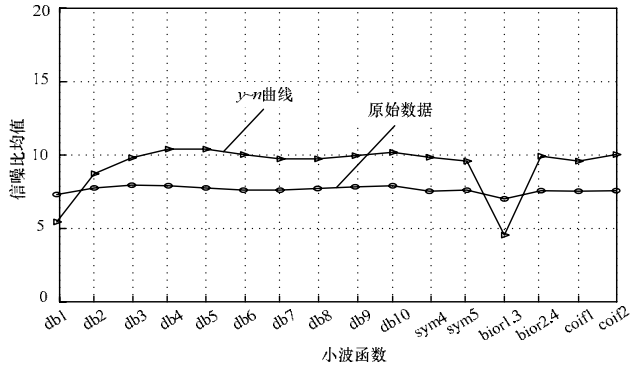


图 4　16 种小波函数对 y~n 曲线和原始数据的信噪比比较

可以看出，除 db1(sym1, Haar)和 bior1.3 外，其余小波用 y~n 曲线的信噪比均明显好于原始数据的信噪比，这可能是因为这 2 种小波函数的分段函数每个阶段的常数函数特点与 y~n 曲线的非线性特点不吻合。而 y~n 曲线与小波函数都具有零均值性，从而使得其降噪效果更好。用原始数据处理时，db3 小波函数的多尺度降噪效果最好，在每个点均有较高的

(下转第 24 页)