

基于近邻传播的时间序列基因表达谱聚类算法

周 运^{a,b,c}, 徐久成^c, 徐存拴^{a,b}

(河南师范大学 a. 生命科学学院; b. 省部共建细胞分化调控国家重点实验室培育基地;

c. 计算机与信息工程学院, 河南 新乡 453007)

摘 要:聚类是识别基因表达数据蕴含的关键基因调控模块的一种有效方法, 基因表达谱的相似性度量是聚类的关键问题. 然而, 一般的相似性度量方法不能刻画时间序列基因表达谱数据所蕴含的时间延迟、反向相关和局部相关等复杂的基因调控关系. 针对时间序列基因表达谱数据, 提出一种基于近邻传播和动态规划的相似性度量方法和聚类算法. 在大鼠再生肝细胞基因表达谱数据集上的聚类结果与基因功能富集分析结果高度一致, 证明算法在时间序列基因表达谱数据聚类上的有效性.

关键词:近邻传播; 时间序列; 反向相关; 瞬时相关; 基因表达谱

中图分类号:TP391.9

文献标志码:A

随着生物高通量技术的广泛应用和微阵列芯片成本的不断降低, 产生了大量的时间序列基因表达谱数据, 为揭示基因表达和精确调控提供了可能. 时间序列基因表达谱数据分析的一个重要任务是识别基因功能模块, 如转录因子及其调节的靶基因等, 为后续的实验和定量分析提供重要的参考^[1].

聚类算法被广泛运用于基因模块的识别^[2-4]. 聚类就是把一组数据点分割成若干模块, 模块内部的数据点相似性尽量高, 而模块间的数据点相似性尽可能低. 度量数据点相似性常用的方法有欧式距离、马氏距离、皮尔森相关系数、信息熵等, 这些方法能有效识别具有相似表达模式的基因模块. 但基因之间的调控关系非常复杂, 有时在时间上并不同步, 如转录因子和下游靶基因之间的表达调控可能存在时间延迟, 瞬时调控等关系, 这是上述度量方法无法识别的, 造成聚类结果不能完全反映真实的基因调控过程. 一般相似性度量方法存在的问题主要有: 1) 适合识别同步调控和表达的基因模块, 不能识别在时间上有延迟调控关系的上下游基因^[5], 而高时间分辨率的基因表达谱数据中蕴含着丰富的这种基因调控关系. 2) 在生物体内, 上游的转录因子和 microRNA 等调控分子或激活或抑制下游靶基因的表达, 所以时间序列基因表达谱中广泛存在反向表达(受抑制)的情况^[6]. 3) 没有考虑基因表达的瞬时相关性, 即基因之间的调控可能只在一定的时间子序列相关, 而在其他时间点不相关, 比如, 一个突变基因的表达可能在某个时间点高表达, 在后面的时间点几乎恢复到突变前的正常表达水平.

近年来许多学者在研究基因表达谱数据的聚类问题, 试图从不同角度挖掘基因表达谱数据中蕴含的调控规律. 文献[7]提出了一种多聚类算法的框架模型用于肿瘤聚类, 在不同的聚类结果基础上用一个一致性函数得到最终的聚类结果. 文献[8]采用近邻传播算法来对药物靶标进行聚类, 预测出 123 个潜在的药物靶分子. 为了解决生物数据对传统聚类算法带来的时间和空间复杂性上的挑战, 文献[9]提出并行框架来加速相似性矩阵的构建, 并利用近邻传播算法进行聚类, 取得了很好的效果. 这些算法大多针对基因表达谱数据高维高噪声的特点, 从计算效率上进行改进, 对基因表达的时间属性没有进行更多的考虑. 最近文献[10]针对时间序列基因表达谱数据, 提出一种基于近邻传播的聚类算法, 该算法采用滑动窗口机制来探索基因之间的关系, 但该算法的空间复杂性和时间复杂性都较高.

收稿日期: 2015-04-03; 修回日期: 2015-09-17.

基金项目: 国家 973 前期研究专项基金(2012CB722304); 河南省基础与前沿计划研究项目(122300410355).

第 1 作者简介(通信作者): 周 运(1978-), 男, 河南信阳人, 河南师范大学副教授, 河南师范大学博士研究生, 研究方向为生物信息学, E-mail: zy@htu.cn.

针对时间序列基因表达谱蕴含的丰富的基因调控关系,本文首先基于动态规划算法,提出了一个刻画时间延迟相关、反向相关和瞬时相关属性的基因表达谱相似性度量方法,然后基于近邻传播算法进行基因聚类,在聚类精度和运行效率上取得了较大提高。

1 常用聚类方法

传统聚类大多是一种无监督机器学习形式,目标是把一组数据点分成几个类,使得同一类中的数据点相似性较高,类间数据点的相似性较低。

1.1 K-means

给定一个自定义聚类数 K , K -means 算法是一种基于数据点相似性度量的聚类方法,属于无监督学习方法。此算法以 K 为参数,把 n 个数据点分为 K 个聚类,以使同一类中的数据点具有较高的相似性,而且类间数据点有较低的相似性。相似性度量通常采用数据点到类中心的距离的平均值。 K -means 试图找到分类 $c(i), c: \mathcal{L} \mapsto \mathcal{H} = \{1, \dots, K\}$ 和相应的类中心 $\{\mu_k\}_{k \in \mathcal{H}}$, 使得目标函数(1)最小。

$$\mathcal{J}(\{c(i)\}, \{\mu_k\}) = \sum_{k \in \mathcal{H}} \sum_{i: c(i)=k} \|X_i - \mu_k\|^2, \quad (1)$$

这里, $\|\cdot\|^2$ 表示欧氏距离,数据点 X_i 即基因表达谱, $\{c(i)\}$ 和 $\{\mu_k\}$ 分别表示聚类和类中心。

由于基因表达谱数据的高噪声和不精确度量,这类无监督聚类算法应用到基因表达谱数据上往往具有较低的聚类精度,而且基因表达是个时序过程,该聚类算法也没有考虑基因表达谱的时间属性。

1.2 K-centres

当数据点之间的相似性矩阵 $s(X_i, X_j)$ 不能由类似的聚类均值给出时,一般用聚类中的一个数据点本身作为类中心(centre),由此产生了目标函数(2)。

$$\mathcal{J}(\{e(X_i)\}) = - \sum_{X_j \in \mathcal{J}_s, \text{ s.t. } X_j \neq e(i)} s(X_i, e(i)), \quad (2)$$

其中, $e(i)$ 表示基因表达谱 X_i 所属类的类中心,本身也是一个数据点。

K -centres 算法的目标函数其实是 K -means 目标函数的优化启发式方法。对于高维基因表达谱数据, K -centres 方法很难获得较好的聚类效果,算法的鲁棒性和执行效率在高维空间也大幅下降。

1.3 AP 算法

AP(affinity propagation, 近邻传播)^[11] 是一种半监督聚类算法。与传统的 K -means, K -centres 等算法相比,它有很多优势,如不需事先指定类的个数,对初值的选取不敏感,对相似性矩阵的对称性没有要求等。AP 算法将每个数据点看成图中的一个节点,迭代的过程即是在图中节点间不断传播信息来找到聚类集合。AP 算法的输入是数据点的相似性矩阵,计算两个数据点的相似性采用距离的负数,也就是说距离越近,相似性越大。主对角线上的数值 $s(k, k)$ 表示数据点和其自身的相似性,称为偏向参数,一般情况下 $s(k, k)$ 都相等,取非主对角线上的所有数的中位数。这个值很重要,它的大小与最后得到的类的数目有关,一般而言这个数越大,得到的类的数目就越多。AP 算法是通过不断迭代更新代表矩阵 $R = [r(i, k)]_{N \times N}$ 和适选矩阵来找到每个类的类代表点及其所属数据点。这两个矩阵初始化都为 0, N 是所有样本的数目。 $r(i, k)$ 表示第 k 个样本适合作为第 i 个样本的类代表点的代表程度, $a(i, k)$ 表示第 i 个样本选择第 k 个样本作为类代表点的适合程度。迭代更新公式(3)和(4)如下:

$$r(i, k) = s(i, k) - \max_{k' \neq k} (a(i, k') + s(i, k')), \quad (3)$$

$$a(i, k) = \begin{cases} \min\{0, r(k, k) + \sum_{i' \notin (i, k)} \max(0, r(i', k))\}, & i' = k, \\ \sum_{i' \neq k} \max(0, r(i', k)), & i = k. \end{cases} \quad (4)$$

每次更新后可以确定当前样本 i 的类代表点 k , 类代表点的确定是使 $\{a(i, k) + r(i, k)\}$ 最大的 k 值,如公式(5)所示,

$$e(i) = \operatorname{argmax}_{k \in \mathcal{H}} (a(i, k) + r(i, k)). \quad (5)$$

如果 $i = k$ 的话,那么说明样本 i 本身就是这个类的类代表点,否则 i 属于 k 为类代表点的聚类。迭代停止的条件是所有样本的归属不再变化,或者迭代了 n 次没有变化(n 的值自定义)。

AP 聚类算法的迭代过程容易产生震荡,所以一般每次迭代都加上一个阻尼系数 $\lambda^{[12]}$,如公式(6)和(7)所示.

$$r_n(i, k) = \lambda * r_o(i, k) + (1 - \lambda) * r(i, k), \quad (6)$$

$$a_n(i, k) = \lambda * a_o(i, k) + (1 - \lambda) * a(i, k). \quad (7)$$

为方便说明,公式(3)、(7)中将基因表达谱 i, k 描述为数据点, r_o 和 r_n 分别表示迭代前后第 k 个样本适合作为第 i 个类的代表点的代表程度, a_o 和 a_n 分别表示迭代前后第 i 个样本选择第 k 个样本为类代表点的适合程度.

2 基因表达谱的数据表示

X_i 表示基因 i 在时间点 t 的 mRNA 表达量, $i \in \mathcal{J} = \{1, 2, \dots, N\}$, $t \in \mathcal{T} = \{1, 2, \dots, T\}$, X 表示 N 个基因在 T 个时间点的表达谱矩阵,对于基因 i 的完整表达谱 $X_i = [X_{i1}, X_{i2}, \dots, X_{iT}]^T$.

3 相似性度量方法

AP 算法的输入是数据点的相似性矩阵,两个数据点之间的相似性度量通常采用欧氏距离等公式计算.但对于基因表达谱数据,这些度量方法无法刻画在时间序列中存在的时间延迟(转录调控)、反向相关(抑制表达)和瞬时相关(局部相关)特性.所以简单的距离度量方法不适合度量时间序列基因表达数据,需要更复杂的相似性度量方法.

本文选择文献[13]的相似性打分方法,然后用局部序列比对发现两个表达谱的局部相关性.虽然是个近似方法,但很容易用动态规划算法实现.假设所有的基因表达谱 X_i 已用 Z-Score 规范化,即每个基因表达谱的均值为 0,标准方差为 1.定义任意两个基因 X_i 在 t_1 ($t_1 \in \{1, 2, \dots, T\}$)时间点和 X_j 在 t_2 ($t_2 \in \{1, 2, \dots, T\}$)时间点的表达量打分记为 $M_{t_1, t_2} = X_{it_1} X_{jt_2}$, M 为两个基因表达谱的打分矩阵.然后共表达矩阵 E 和反向共表达矩阵 D 用公式(8)和(9)计算.

$$E_{t_1, t_2} = \max(E_{t_1-1, t_2-1} + M_{t_1, t_2}, 0), \quad (8)$$

$$D_{t_1, t_2} = \max(D_{t_1-1, t_2-1} - M_{t_1, t_2}, 0). \quad (9)$$

初始化条件是 $E_{0, t} = 0$, $E_{t, 0} = 0$ 和 $D_{0, t} = 0$, $D_{t, 0} = 0$.通过算法 1 可获得所有基因表达谱间的相似打分矩阵 $\psi(i, j)$.

算法 1

输入:规范化基因表达谱

输出:相似性打分矩阵 ψ

① 对任意 $t \in \mathcal{T}$ 初始化 E_{t0} , E_{0t} , D_{t0} , D_{0t} 为 0;

② 初始化 $t_1 = t_2 = 1$;

③ 按公式(8)和(9)计算 E_{t_1, t_2} 和 D_{t_1, t_2} ;

④ 如果 $t_1 < T$ 并且 $t_2 \leq T$,置 $t_1 = t_1 + 1$,返回到③;

⑤ 如果 $t_1 = T$ 并且 $t_2 < T$,置 $t_1 = 1$, $t_2 = t_2 + 1$ 返回到③;

⑥ 若 $\alpha = \max_{t_1, t_2} [E_{t_1, t_2}]$ 和 $\beta = \max_{t_1, t_2} \{D_{t_1, t_2}\}$,置 $\psi(i, j) = \max(\alpha, \beta)$.

如果 $\psi(i, j)$ 值来自矩阵 E 的主对角线,表示两个表达谱相似;如果 $\psi(i, j)$ 值来自矩阵 E 的下三角阵里,表示两个表达谱有时间延迟关系($t' < T$);如果 $\psi(i, j)$ 来自矩阵 D 的主对角线,表示两个表达谱反向相关;如果 $\psi(i, j)$ 对应 E 最大值位置上 D 矩阵相应值为 0,表示两个表达谱在该时间点瞬时相关.近邻传播聚类算法的输入相似性矩阵可由公式(10)变换得到:

$$s(i, j) = \psi(i, j) - T. \quad (10)$$

4 基于近邻传播的基因表达谱聚类算法

算法 2

输入:基因表达谱相似性矩阵 $s(i,j)$ 、最大迭代次数、偏向参数 $s(k,k)$ 和阻尼系数 λ

输出:基因聚类

- ① 初始化代表矩阵 $r(i,k)$ 和适选矩阵 $a(i,k)$ 为 0;
- ② 迭代计算公式(3) 和公式(4);
- ③ 如果所有基因所属类别不再变化或者迭代了 n 次没有变化,迭代过程停止.

5 实验结果分析

5.1 实验数据及参数设置

本文采用河南师范大学细胞分化调控国家重点实验室 Rat Genome 230 2.0 芯片检测大鼠 2/3 肝切除后再生肝恢复过程肝细胞在 2 h,6 h,12 h,24 h,30 h,36 h,72 h,120 h,168 h 9 个时间点的基因表达谱,数据集存储在 NCBI GEO 数据库 (accession number:GSE55434). 根据真假手术两组数据的 fold change>3 和 F 检验的 $p\text{-value}<0.05$,挑选出 1068 个肝再生相关基因,按照 Z-Score 进行规范化处理.

近邻传播算法的输入是相似性矩阵 s ,偏向参数 $s(k,k)$ 设为矩阵 s 除对角线外的数值的中位数,阻尼系数 λ 设为 0.9.

5.2 结果分析

把表达谱矩阵输入到上述聚类算法,得到 90 个聚类. 为了验证算法的有效性,挑选其中 5 个聚类的基因进行基因功能富集分析如表 1 所示.

表 1 5 个聚类的基因功能富集分析结果

聚类	分类	功能注释	p-Value 值	相关基因	相关基因数	基因总数	比例/%
1	Cancer	Cancer	4.83E-07	ARHGAP11A, ASPM, AURKA, AURKB, BUB1, BUB1B, CCNA2, CCNB1, CCNB2, CCNF, CDC20, CDC25B, CDK1, CENPF, CENPK, CKS2, DEPDC1, DLGAP5, ESPL1, FAM83D, HMGB2, HMMR, KIF11, KIF20A, KIF20B, KIF22, KIF2C, KNSTRN, MAD2L1, MKI67, NUSAP1, PARPBP, PLK1, PLK4, PRC1, PTTG1, RACGAP1,RRM2,SMC4,SPC25,TOP2A, TROAP,UBE2C,UBE2T	44	62	71.0
2	Cancer	epithelial neoplasia	1.41E-03	TNFRSF9, BCL3, C11orf82, CDCA2, FOXM1, IL34, KIAA0101, LCN2, NDUFC2,NUF2,RAD51,RAD54L,RFC3, STMN1,SUV39H1,TRIP13,TTK,TUBB	18	26	69.2
18	Cellular Growth and Proliferation	proliferation of cells	4.50E-04	MYC,CGREF1,CXCR7,FABP4,IMPDH2, LPIN1,MAP3K8	7	13	53.8
38	Cellular-Growth and Proliferation	proliferation of cells	1.42E-05	HGF, APOE, AVPR1A, CCND1, Ly6a, MBL2, MCM8, MEF2C, PUS7L, SERPINE1,TGFBI	11	14	78.6
43	Cancer	adenocarcinoma	5.31E-05	CXCR4, CKAP4, DIAPH3, ELOVL3, FAM171B, FGF13, KCNC1, LSM7, P2RY10,SLC24A2	10	12	83.3

从功能富集结果来看,在 cluster1 的 62 个基因有 44 个与癌症有关,cluster2 的 26 个基因有 18 个与上皮细胞瘤的发生有关,cluster18 的 13 个基因中有 7 个与肝细胞增殖有关,cluster38 的 14 个基因中有 11 个与肝细胞增殖有关,cluster43 的 12 个基因有 10 个恶性腺瘤有关.

为说明方便,挑选两个已知在肝再生和肿瘤发生中起关键作用的两个聚类讨论,分别是基因 MYC 和 HGF 所在的聚类 18(表 2)和聚类 38(表 3),聚类结果如图 1 所示.

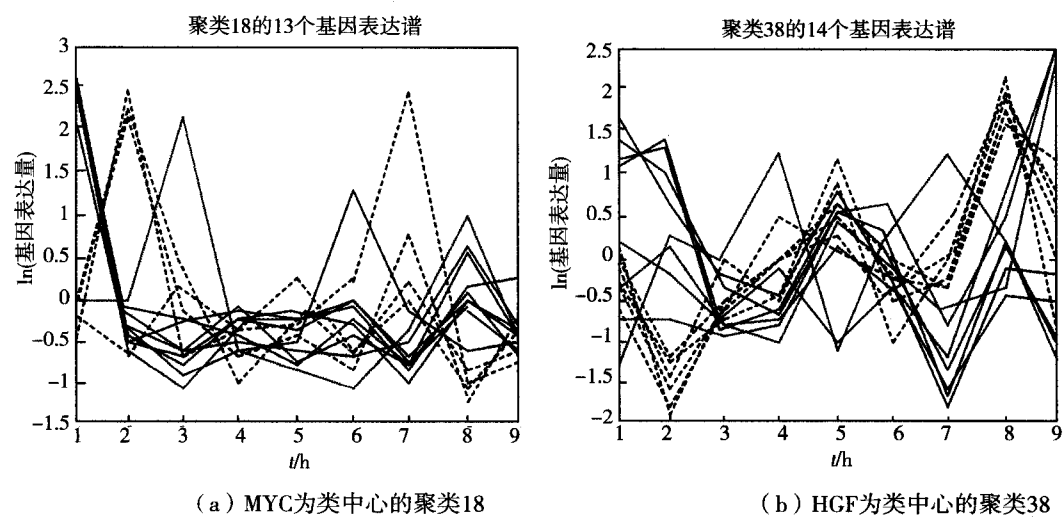


图1 肝细胞基因表达谱聚类结果

图 1 中,实线表示和类中心一致;虚线表示和类中心反向相关;圆点虚线表示延迟相关. 聚类(a)中,类中心 MYC 是在肝再生和癌症发生过程中起重要作用的一个原癌基因,是一个转录调控因子,起重要的调控作用,一些延迟表达的基因如 CGREF1,COBLL1,IMPDH2,LTV1 和 TGFB1 可能受它调控,文献[14]研究发现 MYC 基因在大鼠肝再生中延迟 TGFB1 的 mRNA 表达,文献[15]研究发现了 MYC 基因在黑素瘤发展过程中对 IMPDH2 基因的表达起了抑制作用. 此外,该类基因主要富集在“Myc Mediated Apoptosis Signaling”信号通路中,参与细胞凋亡过程,对肝再生的启动和终止起重要作用[16],这与聚类结果一致. 通过基因表达谱的延迟表达特性,可以预测其他可能没有发现的调控关系,为实验验证提供参考.

表 2 聚类 18 的基因相似性度量

探针号	基因名	家族	平均相似度	反向	延迟	瞬时
1370361_at	CGREF1	other	0.851 38	0	-2	0
1376868_at	Cobll1	other	0.851 38	0	-6	0
1397050_at	CPNE8	other	0.851 38	0	0	0
1367940_at	CXCR7	G-protein coupled receptor	0.851 38	0	0	0
1368271_a_at	FABP4	transporter	0.851 38	0	0	0
1370023_at	GJA4	transporter	0.851 38	0	0	0
1388629_at	IMPDH2	enzyme	0.851 38	0	-1	0
1383149_at	LOC684122	other	0.851 38	0	0	0
1377599_at	LPIN1	phosphatase	0.851 38	0	0	0
1372312_at	LTV1	other	0.851 38	0	-1	0
1369393_at	MAP3K8	kinase	0.851 38	0	0	0
1368308_at	MYC	transcription regulator	0.851 38	0	0	0
1393072_at	UBE2Q2	enzyme	0.851 38	0	-1	0

聚类(b)中,CCND1 细胞周期素是类中心,该类中 14 个基因中 11 个与参与细胞增殖活动,HGF(肝细胞生长因子)是一个在肝再生和肝癌发生过程中起重要作用的一个肝细胞生长因子,TGFB1 转化生长因子,AVPR1A 是 G 蛋白连接受体,都是在细胞增殖过程中非常重要的基因,这说明聚类结果与基因的实际功能高度一致. 此外,该类基因主要富集在“HGF Signaling”信号通路中,在肝再生过程中扮演刺激细胞增殖、分化、侵袭和迁移等多种角色[17].

6 结 论

生物学理论认为具有相似功能的基因可能具有相似的表达谱,但在发生异常的生理活动中一些基因的

表达受到其他基因的调控,可能存在延迟表达、反向相关和瞬时相关等特性. 目前大多数聚类算法只考虑基因表达谱向量间简单的相似性度量,本文综合考虑基因的延迟表达、反向表达(抑制)和瞬时表达特性对基因表达谱之间的相似性进行度量,然后运用近邻传播进行聚类. 实验结果显示聚类结果与基因功能富集分析结果具有高度的一致性,说明算法可靠和有效. 但由于生物生理活动的复杂性,现在对基因的功能还不可能完全清楚,所以很难有一个十分精确的标准来评价这类算法的准确度.

由于近邻传播算法的输入是数据点间的相似性矩阵,不同类型数据需要选择和定义合适的相似性度量,为聚类提供了一个灵活的框架. 未来计划在基因表达谱相似性度量上引入概率模型刻画基因之间的调控关系. 当基因规模达到约 10^4 时,基因之间相似性度量的计算开销也是需要考虑的问题.

表 3 聚类 38 的基因相似性度量

探针号	基因名	家族	平均相似度	反向	延迟	瞬时
1371288_at	APOE	transporter	0.818 76	1	0	0
1368316_at	AQP8	transporter	0.818 76	1	-1	0
1369664_at	AVPR1A	G-protein coupled receptor	0.818 76	1	-1	0
1383075_at	CCND1	other	0.818 76	0	0	0
1395778_at	CEP72	other	0.818 76	1	-1	0
1395689_at	FAM169A	other	0.818 76	1	-2	0
1387701_at	HGF	growth factor	0.818 76	1	-1	0
1370056_at	Ly6a (includes others)	other	0.818 76	1	-2	0
1379240_at	MBL2 *	other	0.818 76	1	-1	0
1381130_at	MCM8	enzyme	0.818 76	0	0	0
1373410_at	MEF2C	transcription regulator	0.818 76	0	0	0
1378541_at	PUS7L	other	0.818 76	0	0	0
1392264_s_at	SERPINE1	other	0.818 76	1	-2	0
1371913_at	TGFB1	other	0.818 76	1	-1	0

参 考 文 献

[1] 徐存拴,章静波. 大鼠肝再生的功能基因组学研究[M]. 北京:高等教育出版社,2009:349-674.

[2] He F, Zeng A P. In search of functional association from time-series microarray data based on the change trend and level of gene expression[J]. BMC Bioinformatics,2006,7(69):1-15.

[3] Weber C C, Hurst L D. Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation[J]. Genome biology,2011,12:R23.

[4] Wang G, Zhao Y H, Zhao X G, et al. Efficiently mining local conserved clusters from gene expression data[J]. Neurocomputing,2009,73(7/8/9):1425-1437.

[5] Leone M, Sumedha, Weigt M. Clustering by soft-constraint affinity propagation: applications to gene-expression data[J]. Bioinformatics,2007,23(20):2708-2715.

[6] Kiddle S J, Windram O P F, McHattie S, et al. Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana[J]. Bioinformatics,2010,26(3):355-362.

[7] Yu Z W, Chen H T, You J, et al. Hybrid Fuzzy Cluster Ensemble Framework for Tumor Clustering from Biomolecular Data[J]. IEEE-ACM transactions on computational biology and bioinformatics,2013,10(3):657-670.

[8] Feng Y H, Wang J T. Prediction of Drug Target Proteins with Affinity Propagation Algorithm[J]. Research journal of biotechnology, 2013,8(6):38-42.

[9] Wang M, Zhang W, Ding W, et al. Parallel Clustering Algorithm for Large-Scale Biological Data Sets[J]. PLOS ONE,2014,9(4):e91315.

[10] Chiu T Y, Hsu T C, Yen C C, et al. Interpolation based consensus clustering for gene expression time series[J]. BMC Bioinformatics, 2015,17:e117.

[11] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science,2007,315(5184):972-976.

[12] Song J, Nicolae D L. A sequential clustering algorithm with applications to gene expression data[J]. Journal of the korean statistical society,2009,38(2):175-184.

[13] Jiang Q, Marisa D F, Jimmy L, et al. Beyond synnexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new biologically relevant interactions[J]. J Mol Biol,2001,314:1053-1066.

[14] Loredana M, Elda P, Ersilia M, et al. Hepatocyte 'priming' and increase in transforming growth factor-beta1 mRNA expression are de-

- layed in hypothyroid versus euthyroid rats during liver regeneration[J]. *Int J Mol Med*, 2006, 17(6): 1063-8.
- [15] Mannava S, Grachtchouk V, Wheeler L J, et al. Direct role of nucleotide metabolism in C-MYC-dependent proliferation of melanoma cells[J]. *Cell Cycle*, 2008, 7(15): 2392-400.
- [16] Zhang D, Li C, Zhang A, et al. mTOR-Dependent Suppression of Remnant Liver Regeneration in Liver Failure After Massive Liver Resection in Rats[J]. *Digestive diseases and sciences*, 2015, 60(9): 2718-2729.
- [17] Gui Y, Yeqaneh M, Ramanathan S, et al. SOCS1 controls liver regeneration by regulating HGF signaling in hepatocytes[J]. *Journal of hepatology*, 2011, 10(55): 1300-1308.

A cluster Algorithm for Time-course Gene Expression Profile Based on Affinity Propagation

ZHOU Yun^{a,b,c}, XU Jiucheng^c, XU Cunshuan^{a,b}

(a. College of Life Science; b. Key Laboratory for Cell Differentiation Regulation;

c. College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)

Abstract: Clustering is an effective method used to identify key gene regulation modules from time-course gene expression data. An important problem in clustering is similarity measure between gene expression profiles. However, general similarity measurement methods are not suitable for gene expression data with time delay, invert correlation and transient correlation feature. Since temporal feature reflect complex regulation relationships between genes, a similarity measurement method for time-course gene expression data and a clustering algorithm based on affinity propagation and dynamic programming methods were proposed. Clustering results in hepatocyte gene expression dataset during rat liver regeneration showed that proposed algorithm is high degree of consensus with gene function enriched analysis. Results shows validity of proposed algorithm applied in time-course gene expression data.

Keywords: affinity propagation; time-course; invert correlation; transient correlation; gene expression prolifera-