高技术通讯 2014 年 第24 卷 第3 期:228 ~235

doi:10.3772/j.issn.1002-0470.2014.03.002

基于相关分散搜索的基因表达数据双聚类①

孙俊玲②

(武汉大学计算机学院 武汉 430072)

(河南财政税务高等专科学校信息工程系 郑州 451464)

摘 要 为了发现来自基因表达数据的双聚类,提出了一种相关分散搜索方法,分散搜索是一种进化技术,它以按质量和多样性指标选择的一个小解集进化为基础。该算法所用适应度函数是基于基因之间的线性相关性来检测基因的移动和缩放模式,同时为了选择正相关基因,采用了一个改进方法。该算法已用酵母细胞周期数据集、人类 B 细胞淋巴瘤数据集和酵母应力数据集三个真实数据集进行了测试,同时与使用基因本体数据库的 CC,OPSM,ISA,BiMax 和 Samba 等方法进行了比较。实验结果表明该算法能发现大量移动和缩放模式双聚类。

关键词 双聚类,分散搜索,微阵列技术,移动模式,缩放模式

0 引 言

随着基因芯片和 DNA 微阵列检测技术的发展,对基因表达数据的分析成为当前研究的重点。双聚类是分析微阵列数据的重要途径,它可同时在基因和条件两个维度上分析基因表达数据,找出在部分条件下具有相似表达趋势的基因。本文提出了基于相关分散搜索的基因表达数据的双聚类方法,并通过实验研究了其性能。

1 相关研究

在微阵列分析背景下的双聚类研究首先是由 Cheng 和 Church 在 2000 年开始的,他们提出了 CC(Cheng and Church)算法[1]。这是一个贪婪迭代搜索方法,包括通过循环添加或删除行或列来建立一个双聚类,以提高由均方偏差(mean squared residue,MSR)[2]测度的质量。之后,出现了多种算法,如选代签名算法(iterative signature algorithm,ISA)[3],它使用了基因表达的简单线性模型,且假设在一个特定的方式下,每一个基因或条件的表达水平为一个正态分布,用于双聚类分析的统计算法(statistical algorithmic method for bicluster analysis,SAMBA)[4],它借助一个基于二分图的模型对双聚类进行详尽计算和评估,随后为找到最大权重子图,使用一个贪婪方法添加或删除节点;文献[5]利用保阶子矩阵(order preserving sub matrix,OPSM)算法实现的双聚类计算算法;谱双聚类[6]算法,它使用线性代数技术,具体说就是特征向量微积分,从输入数据中识别双聚类结构;格子模型[7]——一个统计建模方法,它把输入矩阵描述为叠加层,每一层对应一个双聚类;包含最全的二进制双聚类算法[8](binary inclusion-maximal biclustering algorithm,BiMax),它采用二进制值使数据集合离散化,需递归应用该方法直到检测到只有唯一值的子矩阵。

双聚类的几何特性可用来发现模式[9]。该类技术使用图像处理方法来寻找代表双聚类的超平面,如一些元启发式双聚类算法——进化方法[10]、多目标进化方法[11]、粒子群优化[12]、贪婪随机自适应搜索[13]和分布估计算法[14]。所有这些算法使用均方偏差(MSR)作为其适应函数的一部分。虽然 MSR 通常被用作质量标准,从生物学的角度看一些有趣的图案不能如此测量和检测。MSR 用来识别移动模式的双聚类是有效的,但不一定适合缩放趋势模式。当在所有基因中除了固定值之外的其他表

① 国家自然科学基金(60970063)资助项目。

② 女,1972 年生,副教授;研究方向:决策支持系统,数据挖掘和优化技术;联系人,E-mail:sun20130219@163.com

(收稿日期:2013-05-14)

达式值变化时,该组基因具有移动模式。当在所有基因中表达式值随一个固定值乘积变化时,该组基因具有缩放模式。为了发现数据中的这些模式,当基因值方差值高,即当基因表现为缩放模式时,立鲁伊斯证明 MSR 不是一个好的质量评估方法[15,16]。

其他算法被用来处理时间序列基因表达数据。在这类数据中,双聚类局限于那些相邻列。此约束已成为一个容易处理的双聚类问题。CCC-Biclustering[17]在线性时间内发现最大的连续列相关双聚类。首先,该算法对矩阵离散化处理,然后执行基于后缀树的字符串处理。由于在微阵列实验和离散化过程中,会发生有关基因值的错误,该算法鲁棒性较差。e-CCC-Biclustering[18] 是 CCC-Biclustering 的一个鲁棒性扩展,可以发现相似表达模式,如缩放模式,并提出了计算这些模式中出现错误的几点测量措施。

一个条件集合下的基因表达水平可以被看作是一个离散随机变量的值。因此,两个基因之间的线性关系,可以通过使用两个随机变量之间的相关系数进行研究。在本文中,这一事实促使使用基于基因之间的相关性测度。几个基于相关的测量方法在文献[19]中已被提出。文献[19]使用相关性定义,双聚类被看作是高维空间中的超平面,因此,问题被转化为嵌入在超平面的一组点的搜索。在文献[20]中,相关系数被用于贪心算法来产生一个双聚类。文献[21]提出了一种基于树结构的双聚类枚举算法,它使用基于 Spearman 秩相关性评价函数。本文提出了双聚类的分散搜索算法。分散搜索[22]是一个以种群为基础的方法,强调对随机过程的系统化处理。因此,初始种群的产生不是随机的,而是使用多样化[23]生成方法创建一组不同的初始解。此外,分散搜索包括一个改进方法,其目的是利用所提供的生成和组合方法的多样性。基因之间的线性相关性被包括在适应度函数中,以便评估分散搜索中双聚类的质量,从而改进移动和缩放模式的局限性。

2 算法描述

基于种群的优化算法是一个搜索程序,其代表试验解的一组个体进化以便找到问题的最优解。分散搜索采用多元化和加强搜索的战略,目的是避免局部极小并且找到高质量解。与其他进化算法不同,它强调随机过程的系统化处理。

基本上,优化过程由称为参考集的集合进化构成。根据其适应函数值,以及参考以前最好解种群中个体的分散性,最初该参考集由种群中最优解创建,通过使用组合方法和改进方法,该集合被更新,直到该集合稳定不变为止。当参考集稳定时,即当应用组合和改进方法后,它还包括相同的解,在下一次迭代前,初始化并重建参考集。即,参考集的初始建立是以质量和多样性为基础,但其更新只是根据质量标准进行。因此,在进化过程中的每一步,主要是当在参考集被重建时,当初始种群产生时,多样性标准被引入。而增强搜索是基于改进的方法,利用问题知识改进其解。

所提出的双聚类分散搜索伪代码在算法 1 中描述。分散搜索过程重复 numBi 次,numBi 是被发现的双聚类的数量,每次迭代参考集的最佳解是存储

算法 1:双聚类分散搜索算法

输入微阵列 $\boldsymbol{M}$,被发现的双聚类个数 numBi,最大迭代数 numIter,初始种群大小和参考集合大小 S。

输出:numBi 个双聚类集合 *Results*。

```
begin
  num ← 0, Results ← φ
while(num < numBi) do
    初始化种群 P
    P ← Improvement Method( P )
    //创建参考集合
    R1← S/2   来自 P 中双聚类  (根据适应函数)
    R2← S/2   最分散双聚类,参考 R1,来自 P \ R1(根据距离).
    RefSet ←(R1 ∪ R2)
  P ← P\ RefSet
    //初始化
    stable ← FALSE, i ← 0
    while(i < numIter) do
      while(NOT stable) do
        A ← RefSet
        B ← Combination Method(RefSet)
        B ← Improvement Method(B)
       RefSet ← S 最好双聚类,来自 RefSet ∪ B
        if(A = RefSet) then
          stable ← TRUE
        end if
      end while        //重建 参考集合
      R1← S/2 最好双聚类,  来自 RefSet
      R2←S/2 最分散双聚类,来自 P\ R1
      RefSet ←(R1 ∪ R2)
      P ← P \ RefSet
      i ← i + 1
    end while     //存储在 Results
    Results ←来自 RefSet 中最好的
    num ← num + 1
```

end while
end

在称为 *Results* 的集合中。因此,由 numBi 个双聚类形成的 *Results* 集合,作为算法 1 的输出。分散搜索主要包括一个用来生成初始种群的多样化生成方法,创建新的后代组合方法(Combination Method( ))以及一种强化搜索的改进方法(Improvement Method( ))。分散搜索的所有具体步骤详细介绍如下。

2.1 初始化阶段

形式上,微阵列是由 N 个基因和 L 个条件构成的实数矩阵 $\boldsymbol{M}$。矩阵元素 (i,j) 意味在条件 j 中的基因 i 的表达水平。双聚类 B 是由 $n \leqslant N$ 基因和 $l \leqslant L$ 条件组成的矩阵 $\boldsymbol{M}$ 的一个子矩阵。双聚类由长度为 $N + L$ 的二进制字符串编码。每个二进制字符串前 N 位与基因相关,剩余 L 位与条件相关。例如,字符串 0010110000 | 01100 用 10 个基因 $\{g_i\}_{1\leqslant i\leqslant 10}$ 和 5 个条件 $\{c_j\}_{1\leqslant j\leqslant 5}$ 表示微阵列一个双聚类。这个字符串对双聚类编码,它由基因 g_3,g_5 和 g_6 以及条件 c_2 和 c_3 组成。

产生的初始种群解尽可能多样化。因此,多样化的生成方法[23]取一个二进制字符串作为种子解,x_i,$i = 1,\cdots,n$,其中 n 是比特数,按以下规则产生解 x'_i,

$$x'_{1+kh} = 1 - x_{1+kh}$$
$$k = 0,1,2,3,\cdots,[n/h] \qquad (1)$$

其中 $[n/h]$ 是小于或等于 n/h 的最大整数,h 小于 n/5 的整数。x' 中所有剩余的位与 x 的对应位相同。

所有可能的种子解生成后,如果需要更多的解,通过使用最后解作为新的种子。

2.2 双聚类评价:适应值函数

最难找到的双聚类是指共同表现为移动和缩放模式的双聚类。该算法目的是发现这种类型的双聚类。如果它们由式

$$g_\gamma = \alpha g_x + \beta \qquad \alpha,\beta \in \mathbf{R} \qquad (2)$$

描述,则两个基因呈现移动和缩放模式。

因此,具有移动和缩放模式的两个基因是线性相关,因此具有相关性测量评价的适应度函数是一个能找到这类模式双聚类的评价函数。

利用两个变量 X 和 Y 之间的相关系数来测量它们线性依赖等级。它的定义是:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_X\sigma_Y} \qquad (3)$$

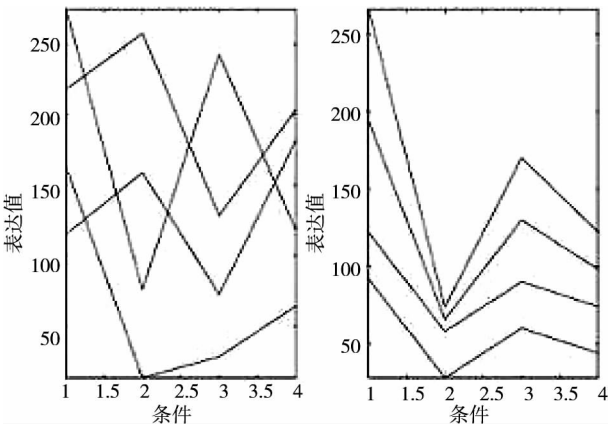
其中,$\mathrm{cov}(X,Y)$ 是变量 X 和 Y 的方差,$\bar{x}$ 和 $\bar{y}$ 分别是变量 X 和 Y 的平均值,σ_X 和 σ_Y 分别是 X 和 Y 的标准偏差。

给定一个由 N 个基因构成的双聚类 $B = [g_1,\cdots,g_N]$,B 的平均相关性 $\rho(B)$ 定义如下:

$$\rho(B) = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \rho(g_i,g_j) \qquad (4)$$

其中 $\rho(g_i,g_j)$ 是基因 i 和基因 j 的相关系数。由于 $\rho(g_i,g_j) = \rho(g_j,g_i)$,因此,只有 $\binom{N}{2}$ 个元素被考虑。

图 1 给出一个低相关基因双聚类和一个高度相关基因的双聚类。可以观察到完美的移动和缩放模式双聚类具有平均相关系数 1,没有模式的双聚类有接近 0 的平均相关性(具体 0.003)。



(a)低相关基因双聚类 (b)高度相关基因双聚类

图 1 基因之间的相关性

在这项工作中,具有高关联基因和高容量的双聚类优先。因此,用来评价双聚类质量的适应度函数定义如下:

$$f(B) = (1 - \rho(B)) + \sigma_\rho + M_1\left(\frac{1}{nG}\right) + M_2\left(\frac{1}{nC}\right) \qquad (5)$$

其中 nG 和 nC 分别是双聚类 B 的基因数和条件数,M_1 和 M_2 是控制双聚类 B 体积的惩罚因子,σ 是公式(4)值 $\rho(g_i,g_j)$ 的标准偏差。包含标准偏差是为了避免一个双聚类平均相关值高。双聚类可以包含剩下双聚类的几个非相关基因。最好的双聚类具有最小适应度函数值。因此,$(1-\rho(B))$ 可被用来识别具有高度相关基因双聚类。

此外,由于带噪声但具有移动模式的基因也可以呈现高相关,所以这种评价测量对噪声鲁棒性好。

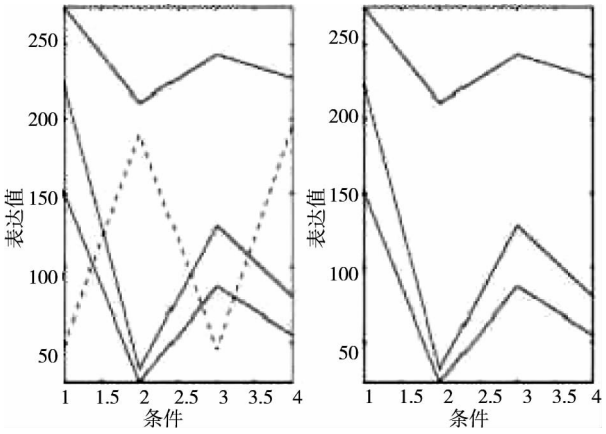
2.3 改进方法

当解必须满足一些约束或者为了加强搜索过程

时,本文采用了一种改进的分散搜索方法。这种方法取决于所研究的问题,通常是由经典的连续优化问题局部搜索组成。

本文目标是找到移动、缩放模式的双聚类。因此,只搜索有正相关基因的双聚类,所提出的改进方法目的是从初始种群双聚类或由组合方法获得双聚类中提取正相关基因。该改进方法伪代码在算法 2 中描述。

图 2 给出一个双聚类,其由 4 个基因组成:3 个高度相关基因和 1 个剩余的负相关基因。该双聚类的平均相关系数等于 0.0083,在应用改进方法后,双聚类的平均相关系数值等于 1。因此,通过消除负相关基因,双聚类体积减少,当改进方法被应用时,新双聚类平均相关系数将大于原来的双聚类。



(a)改进方法前双聚类　　(b)改进方法后双聚类

图 2　应用改进方法之前和之后的双聚类

2.4　参考集的建立

考察前最佳解集,根据其适应函数值以及在初始种群中分散度,创建参考集。

算法 2　改进方法

输入双聚类 $B=[g_1,\cdots,g_N]$

输出双聚类 $B'\subseteq B$ 使 $\rho(g_i,g_j)\geqslant 0\ \forall g_i,g_j\in B'$

```
begin
i ← 1, B' ← {g_i}, R ← { }
while(i < N) do
    j ← i + 1
    while(j ≤ N) do
        if(ρ(g_i, g_j) > 0) then
            if(g_j ∉ R) then
                B' ← B' ∪ {g_j}
            end if
        else
            R ← R ∪ {g_j}
        end if
        j ← j + 1
    end while
    i ← i + 1
end while
end
```

海明距离被用来衡量双聚类之间的距离。在更新过程中获得的参考集稳定后,参考集将被重建以便在搜索过程中引入多样性。因此,参考先前选择的最好解,根据适应度函数和来自最初种群中最分散的解,从已更新的参考集中用最好的双聚类,来重建参考集。

在进化过程中通过移除在参考集创建或重建中已经被考虑的解,更新初始种群。当初始种群是空的,通过采用多样化生成方法创建一个新种群。

2.5　组合方法和参考集更新

在搜索过程中,由组合方法引入新的解。采用均匀交叉算子,两个解被组合,一个新的解被生成。在参考集中所有双聚类对被组合,因此产生 $S*(S-1)/2$ 个新的双聚类,其中 S 是参考集的大小。交叉算子随机产生一个掩模码,当掩模码中有一个 1 时,其后代组成的值来自第一个双亲,当掩模中有一个 0 时,其对应后代的值来自第二个双亲。

在组合所有双聚类对后,从先前的参考集与新解集中选择最好的解加入。因此,根据它们的适应值函数,最好的解保持在参考集中。

3　实验与分析

为研究本文提出的算法的性能,将其应用于 3 个真实数据集的测试。第一个数据集(简称酵母集)是由 Cho 提出的具有 2884 个基因和 17 个实验条件酿酒酵母细胞周期表达[23]。第二个(简称淋巴瘤集)是具有 4026 个基因和 96 个条件的人类 B 细胞淋巴瘤表达数据[24]。这两个数据集可在文献[1]中得到,在那里原始数据被处理。第三个数据集(简称加施酵母集)是由 Gasch 提供的具有 2993 个基因和 173 个条件的酿酒酵母应力条件表达[25]。该数据集在文献[8]被使用,可以作为辅助数据下载。

算法 1 的内部参数如下:分散搜索迭代的最大数量为 20,参考集的大小为 10,初始种群的解个数为 200,每次运行发现的双聚类的数量为 100。M_1 和 M_2 参数是适应函数的权重,以便根据双聚类的所需尺寸来驱动搜索。当具有很多基因和条件的双聚类被需要时,可以使用高的 M_1 和 M_2 值。对于酵母细胞周期表达和淋巴瘤数据集的 *Results* 选择参

数值 $M_1=1$ 和 $M_2=1$。为了表示这些参数对双聚类体积的影响,对于加施酵母数据集的 *Results* 选择参数值 $M_1=1$ 和 $M_2=1$ 以及 $M_1=10$ 和 $M_2=10$。

3.1 *Results* 集合

表 1 显示了由本文提出的分散搜索应用所获得的 100 个双聚类中选择的 4 个双聚类的信息以及 100 个双聚类的平均值信息(粗体显示)。对于每个双聚类,表中给出了双聚类标识符、基因的数量、条件的数量、平均相关系数 $\rho(B)$ 和标准偏差 $\sigma(B)$。为了建立双聚类质量与其他算法的比较,MSR 和基因值方差值也在表中列出。基因方差值测量基因表达水平值如何不同。

表 1 由算法 1 找到的双聚类信息

双聚类标识符	基因数	条件数	体积	$\rho(B)$	$\sigma(B)$	MSR	基因方差值
biYeastN15	7	10	70	0.95	0.56	59.2	882.8
biYeastN21	11	9	99	0.92	0.47	205.2	1190.5
biYeastN24	9	9	81	0.92	0.45	142.9	1344.8
biYeastN40	13	8	104	0.89	0.45	368.2	2185.4
biYeast	22.27	6.46	133.1	0.90	0.48	321.0	1508.7
biLymphomaN1	14	14	196	0.92	0.43	3719.2	29180.0
biLymphomaN11	17	7	119	0.92	0.50	1607.9	10317.6
biLymphomaN15	21	10	210	0.86	0.43	1818.4	8351.2
biLymphomaN54	9	14	126	0.82	0.45	1292.6	6108.0
biLymphoma	10.81	11.53	123.7	0.85	0.45	2593.3	11643.07
bi1-GaschYeastN1	13	25	325	0.96	0.42	0.08	1.51
bi1-GaschYeastN10	12	22	264	0.95	0.48	0.06	1.19
bi1-GaschYeastN11	41	17	697	0.93	0.34	0.15	1.67
bi1-GaschYeastN25	19	10	190	0.93	0.43	0.19	0.89
bi1-GaschYeast	16.36	14.08	237.6	0.89	0.43	0.32	1.50
bi2-GaschYeastN1	54	39	2106	0.82	0.32	0.22	1.00
bi2-GaschYeastN4	43	32	1376	0.84	0.45	0.18	1.02
bi2-GaschYeastN9	48	24	1152	0.87	0.41	0.17	1.18
bi2-GaschYeastN27	33	28	924	0.84	0.39	0.13	0.72
bi2-GaschYeast	46.69	27.69	1269.4	0.72	0.34	0.38	1.02

图 3 和图 4 分别给出 4 个酵母和淋巴瘤数据集的双聚类,其数据在表 1 中列出。图 5 和图 6 描述了加施酵母数据集的双聚类。在图 5 的双聚类 bi1-gaschyeastn1, bi1-gasch yeastn10, bi1-gasch yeastn11 和 bi1-gasch yeastn25 获得权重值 $M_1=1$ 和 $M_2=1$, 在图 6 的双聚类 bi2-gaschy eastn1, bi2-gaschyeastn4, bi2-gaschyeastn9 和 bi2-gaschyeastn27 获得权重值 $M_1=10$ 和 $M_2=10$。值得注意的是,惩罚值越大,得到的双聚类体积也越大。取得参数 $M_1=M_2=1$ 的值的动机是要找到具有少量基因的双聚类,以便直

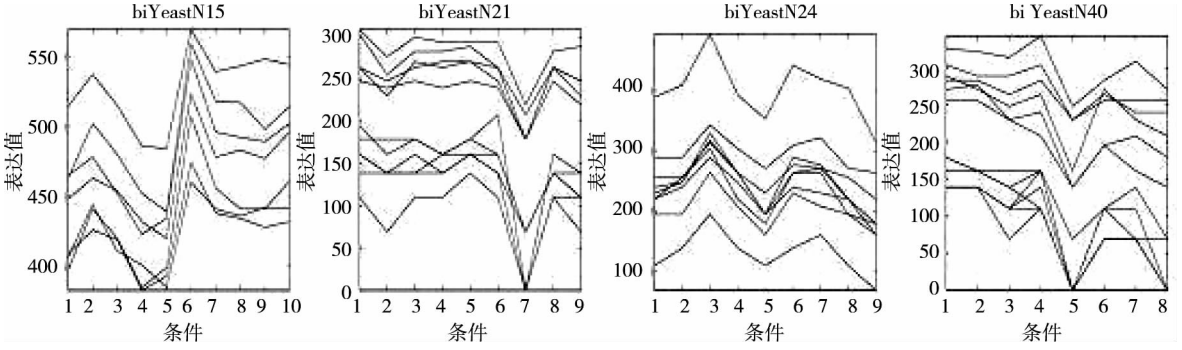


图 3 酵母数据集 *Results*(通过算法 1 从酵母数据集中发现的几个双聚类)

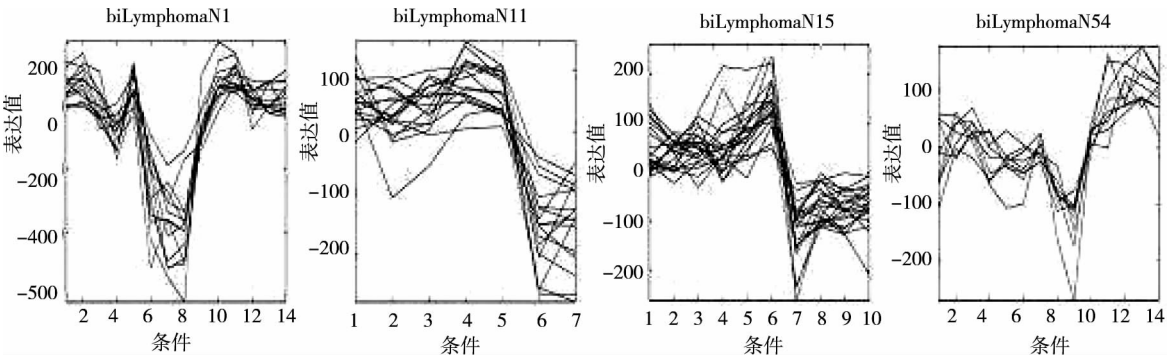


图 4　淋巴瘤数据集的 *Results*(通过算法 1 从淋巴瘤数据集发现的几个双聚类)

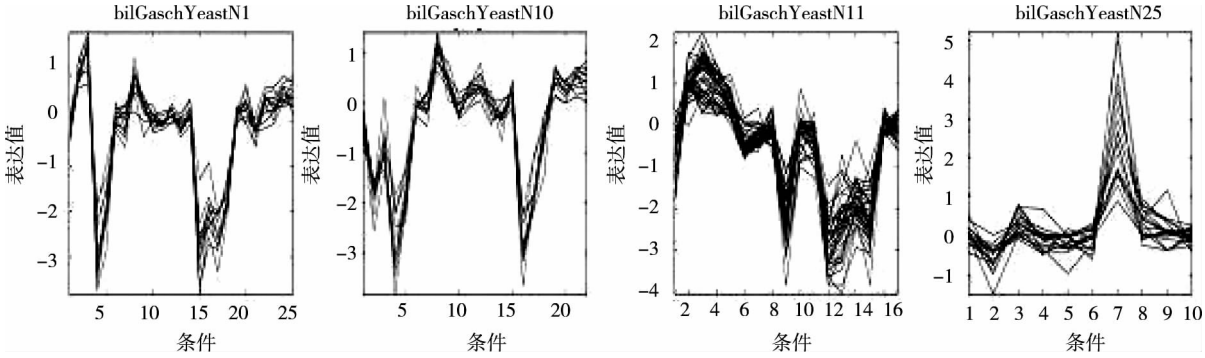


图 5　加施酵母数据集的 *Results*($M_1=1$,$M_2=1$)(通过算法 1 从加施酵母数据集发现的几个双聚类,其获得权重值 $M_1=1$ 和 $M_2=1$)

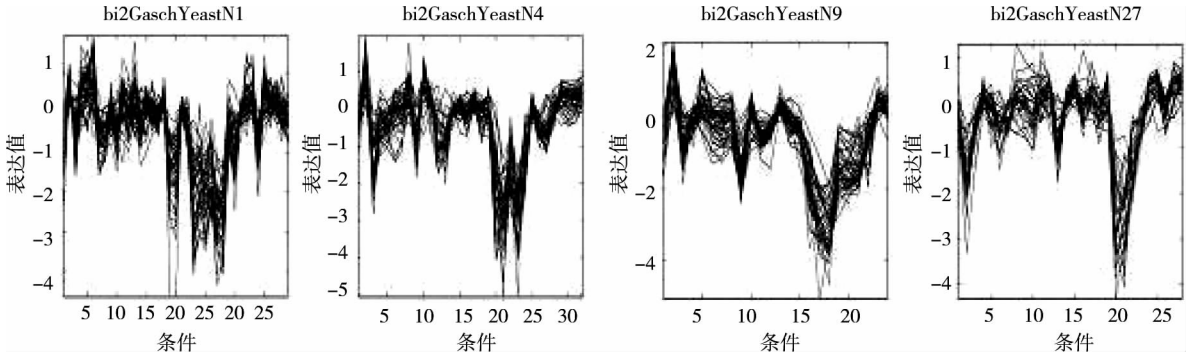


图 6　加施酵母数据集的 *Results*($M_1=10$,$M_2=10$)(通过算法 1 从加施酵母数据集发现的几个双聚类,其获得权重值 $M_1=10$ 和 $M_2=10$)

观显示移动和缩放模式。然而,主要的目标是找到共享相同 GO 的基因组,因此,它更适合寻找大量基因的双聚类。因此,$M_1=M_2=10$ 被认为是获取具有较高体积双聚类的参数。

3.2　讨论

在图 3 中表示的 4 个双聚类有一个高的平均相关值(见表 1)。在其中可以清楚地认出移动和缩放模式双聚类。使用 MSR 作为适应度函数,在大多数论文中都认为如果它的 MSR 小于 300,对应酵母数据集双聚类的质量好[2,21]。这个值依赖数据集是因为它依赖于表达式矩阵值标准偏差和平均值。从这个角度来看,biYeastN15,biYeastN21 和 biYeastN24 是好双聚类,由于平均相关值高,biYeastN40 不是好的双聚类。值得注意的是,相比其他的双聚类,biYeastN40 基因变异的值要高。

图 4 中描述的 4 个淋巴瘤数据集双聚类显示一组具有相似的行为和平均相关系数值高的基因(见表 1)。然而,一些作者因为 MSR 高于 1200,认为这些双聚类不好[2,21]。根据以前对酵母数据集的论述,该值太依赖数据集。注意,具有最低 MSR 值的

双聚类是 biLymphomaN54(1292.6),但该双聚类具有最低平均相关系数值(0.82)。在文献[16]中证明 MSR 不够精确到足以发现移动和缩放模式。使用 MSR 作为适应度函数的算法检测不到具有高基因方差值模式的双聚类。这些实例中报告的 *Results* 集证实了这个情况。

图 5 和图 6 显示了加施酵母数据集双聚类。可以观察到罚参数对基因和条件个数的影响。M_1 和 M_2 值越高,体积越大。从一个几何点观察,所有结果表明了相似行为的基因。例如,在 bi1-GaschYeastN25 中可以很清楚地观察到缩放模式,因为虽然所有基因增加具有不同强度的表达水平,但条件 6 和 8 之间的基因形状是相同的。此外,该平均相关系数值显示具有高度相关基因的加施酵母双聚类。由于先前的预处理,MSR 和基因方差值在其他两个不同数据集范围内变化。

应当指出的是,所有双聚类表现的移动和缩放模式同时具有高的平均相关系数。此外,标准偏差低,即,每对基因的相关系数有相似值且接近双聚类平均关联值。因此,由提出的分散搜索发现的具有高平均相关的所有双聚类不含非相关基因。

4 结 论

本文提出了从基因表达数据中发现双聚类的分散搜索方法。提出的分散搜索已作为价值函数来评价以基因之间相关性为基础的双聚类质量,目的是获得具有移动和缩放模式的双聚类。另外,一种包括从双聚类消除负相关基因的改进方法已被纳入用来增强搜索。该方法已经用三个真实数据集来测试:一个是与酵母细胞周期相关的数据集(酵母),一个是与酵母中不同应力条件集合相关的数据集(加施酵母),另一个是与 B 细胞淋巴瘤相关的数据集(淋巴瘤)。正如在文献[16]中被证明的那样,使用 MSR 不能探测到一组由移动和缩放模式基因组成的双聚类。使用基因本体论与其他几种方法进行了比较,结果表明本文提出的方法具有很好的性能。

未来的工作重点将参考基因间重叠度和适应度函数对本文提出的方法作一些改进。

参考文献

[1] Cheng Y, Church G. Biclustering of expression data. In: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, California, USA, 2000. 93-103

[2] Getz G, Levine E, Domany E. Couple two-way clustering analysis of gene microarray data. In: Proceedings of the National Academy of Sciences (PNAS), USA, 2000. 12079-12084

[3] Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 2003, 67: 1964-1981

[4] Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 2002, 18(90001): 136-144

[5] Ben-Dor A, Chor B, Karp R, et al. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of Computational Biology*, 2003, 10(3-4): 373-384

[6] Kluger Y, Basri R, Chang J, et al. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 2003, 13(4): 703-715

[7] Lazzeroni L, Owen A. Plaid models for gene expression data. *Statistica Sinica*, 2002, 12: 61-86

[8] Prelic A, Bleuler S, Zimmermann P, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 2006, 22(9): 1122-1129

[9] Gan X, Liew A, Yan H. Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinformatics*, 2008, 9(209): 1-15

[10] Divina F, Aguilar-Ruiz J. Biclustering of expression data with evolutionary computation. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18(5): 590-602

[11] Banka H, Mitra S. Evolutionary biclustering of gene expressions. *Ubiquity*, 2006, 7(42): 1-12

[12] Liu J, Li Z, Hu X, et al. Biclustering of microarray data with MOSPO based on crowding distance. *BMC bioinformatics*, 2009, 10(Suppl 4): S9

[13] Dharan S, Nair A. Biclustering of gene expression data using reactive greedy randomized adaptive search procedure. *BMC bioinformatics*, 2009, 10(Suppl 1): S27

[14] Joshua T Burdick, John I M. Deconvolution of gene expression from cell populations across the C. elegans lineage. *BMC Bioinformatics*, 2013, 14(204): 513-526

[15] Sjöstrand J, Arvestad L, Lagergren J. GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics*, 2013, 14(209): 216-228

[16] Aguilar-Ruiz J. Shifting and scaling patterns from gene expression data. *Bioinformatics*, 2005, 21(20): 3840-3845

[17] Madeira S C, Teixeira M C, Sá-Correia I, et al. Identification of regulatory modules in time series gene expression

data using a linear time biclustering algorithm. *IEEE/ACM Trans Comput Biology Bioinform*, 2010, 7: 153-165

[18] Madeira S C, Oliveira A L. A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithms for Molecular Biology*, 2009, 4: 215-226

[19] Madeira S, Oliveira A. Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on Computational Biology and Bioinformatics*, 2004, 1: 24-45

[20] Bhattacharya A, De R K. Bi-correlation clustering algorithm for determining a set of co-regulated genes. *Bioinformatics*, 2009, 25(21): 2795-2801

[21] Ayadi W, Elloumi M, Hao J K. A biclustering algorithm based on a Bicluster Enumeration Tree: application to DNA microarray data. *BioData Mining*, 2009, 2(9): 1203-1216

[22] Marti R, Laguna M. Scatter Search Methodology and Implementation in C. Boston: Kluwer Academic Publishers, 2003. 95-102

[23] Huang Y T, Lin X H. Gene set analysis using variance component tests. *BMC Bioinformatics*, 2013, 14(210): 276-289

[24] Alizadeh A. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 2000, 403: 503-511

[25] Gasch A P, Spellman P T, Kao C M, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 2000, 11(12): 4241-4257

Biclustering of gene gxpression data based on correlation scatter search

Sun Junling

(School of Computer, Wuhan University, Wuhan 430072)

(Department of Information Engineering, Henan Finance and Taxation College, Zhengzhou 451464)

Abstract

A correlation scatter search algorithm is presented with the aim of finding biclusters from gene expression data. Scatter search is an evolutionary technique that is based on the evolution of a small set of solutions chosen according to the quality and diversity criteria. This algorithm uses a fitness function to detect shifting and scaling patterns from genes based on the linear correlation among genes, and includes an improvement method in order to select just positively correlated genes. The proposed algorithm was tested with the three real data sets of the Yeast Cell Cycle dataset, human B-cells lymphoma dataset and Yeast Stress dataset, and the performance of the proposed method and fitness function were compared with that of the other algorithms of CC, OPSM, ISA, BiMax and Samba using the Gene the Ontology Database. The experimental results show that the proposed method can find a remarkable number of biclusters with shifting and scaling patterns.

Key words: bicluster, scatter search, microarray technology, shifting pattern, scaling pattern