基于分块矩阵 MCL 识别时序动态蛋白质网络链功能模块的研究与实现

张锦雄　　李陶深

广西大学计算机与电子信息学院，南宁，530004

摘　要　蛋白质通过相互作用发挥其生物功能，功能模块是参与特定生物过程时在不同时间不同地点彼此相互作用的蛋白质集合，识别蛋白质功能模块已成为蛋白质组学研究主要任务之一。本文研究开发用于时序动态蛋白质相互作用网络链识别功能模块的马尔可夫算法 IFM-TDPINC，该算法针对蛋白质相互作用的时间顺序，构造一个逻辑上由单个邻接矩阵表示的时序动态蛋白质相互作用网络链，运用马尔可夫过程聚类跨多个时序动态蛋白质相互作用网络的蛋白质簇，并基于 GPU 加速马尔可夫聚类过程。实验结果表明，与对比算法相比，算法 IFM-TDPINC 在稠密 PPI 数据集 STRING 上识别出综合指标#PM×FAM 较好的功能模块，且在阳性预测值 PPV 优于其它算法，同时能准确识别一定数量的功能模块。

关键字　蛋白质功能模块，蛋白质-蛋白质相互作用，时序动态蛋白质相互作用网络链，马尔可夫聚类，GPU 加速

Research and Implementation of A sub-block-matrix-multiplicative Based MCL for Identifying Functional Modules from Temporal Dynamic Protein Interaction Network chain

Zhang Jin Xiong　Li Tao Shen

School of Computer, Electronics & Information Guangxi University
Nanning 530004 China
zhangjx@gxu.edu.cn

Abstract—Proteins carry out themselves biological function by interacting with each other. Functional modules consist of proteins that participate in a particular cellular process while binding each other at a different time and place. Identifying protein functional modules has become one of the key tasks of proteomics research. In this paper, a sub-block-matrix-multiplicative based MCL method, called IFM-TDPINC, is proposed to identify functional modules from temporal dynamic protein interaction network chain. According to the time sequences of protein interactions during a particular biological process, IFM-TDPINC constructs a temporal dynamic protein interaction network chain represented by a logically single adjacency matrix, exploits Markov Clustering to cluster protein clusters across several temporal dynamic protein interaction networks, and accelerates the process of Markov Clustering by using GPU. The experimental results show that IFM-TDPINC performs better than the competing methods on dense PPI dataset STRING in terms of #PM×FAM, is superior to the competing methods with regard to PPV, and is capable of accurately identifying a quite quantity of protein functional modules.

Key words—Protein functional module; protein-protein interaction; temporal dynamic protein interaction network chain; Markov Clustering; GPU accelerating

1　引　言

在细胞系统中，生命有机体的大多数生物过程离不开蛋白质的参与，蛋白质-蛋白质相互作用(Protein-Protein Interaction, PPI)是蛋白质参与生物过程的主要方式，生命有机体的生物过程是蛋白质一系列复杂时序相互作用的结果。蛋白质功能模块是在不同时间不同场所时序地参与特定生物过程的蛋白质集合组成[1]，准确认知功能模块参与生物过程的全流程对于理解生命有机体功能机理和结构组织具有重要意义。于是，以 PPI 数据为基础识别蛋白质功能模块的研究成为了蛋白质组学研究的主要任务之一。以蛋白质为结点，以 PPI 为边，PPI 数据可以建模为网络，从而形成蛋白质-蛋白质相互作用网络(PPI Network, PPIN)，于是基于 PPIN 系统地识别蛋白质功能模块的研究引起了人们的关注。

基因表达数据是一组基因在若干均匀间隔时间点上转录的 mRNA 的丰度采样值，它可以反映一组基因在整个采样过程的动态表达模式。显然，基因表达具有动态性，也就是说，基因编码的蛋白质具有时序动态性，自然地蛋白质相互作用也随之具有时序动态性。

随着高通量生物学实验技术的发展，细胞水平的蛋白质相互作用数据和基因表达数据日益增加，为建模蛋白质及其相互作用时序动态行为提供了可能[2]。

当前开放数据库中的蛋白质相互作用数据是在不同的时间地点条件下产生的，这些蛋白质相互作用数据仅说明蛋白质之间存在相互作用，但却没有说明这些相互作用在何时发生。蛋白质相互作用网络是由缺乏时间信息的相互作用数据构成的，因此它是静态的。生物过程中的蛋白质相互作用是时序地发生的，也就是说，生物过程中时序的蛋白质相互作用是跨多个时间点发生的。因此一个特定生物过程对应的功能模块必然存在于跨时间的动态蛋白质相互作用网络中。所以建模细胞系统时序动态蛋白质相互作用网络是准确识别功能模块的有效途径。鉴于功能模块中蛋白质相互作用的时序性，本文研究时序动态蛋白质相互作用网络的构建及从中识别蛋白质功能模块的算法。

2 相关数据集分析

2.1 *R* 可靠蛋白质相互作用网络

给定一个带可靠性得分的蛋白质相互作用数据集*PPIS*=(*PNS*, *PPS*, *Score*)，*PNS*={1, ..., *M*}为相互作用蛋白质结点集，|*PNS*|=*M*，*PPS*={(*i*, *j*)|*i*, *j*=1, ..., *M*}为蛋白质相互作用集，*Score*(*i*, *j*)={1, ..., 999}[3]为相互作用(*i*, *j*)的可靠性得分。另有一个蛋白质相互作用数据集*PPI*=(*PN*, *PP*)，*PN*={1, ..., *N*}为相互作用蛋白质结点集，|*PN*|=*N*，*PP*= {(*i*, *j*)|*i*, *j*=1, ..., *N*}为蛋白质相互作用集。于是用*PPIS*的可靠性得分按式(1)给*PPI*打分得到*PPIs*=(*PN*, *PP*, *s*)。

$$s(i,j)=\begin{cases}0 & , \ if \quad (i,j)\notin PP\\ 1 & , \ if \quad (i,j)\in PP,\ \ (i,j)\notin PPS\\ Score(i,j) & , \ if \quad (i,j)\in PP,\ \ (i,j)\in PPS\end{cases} \tag{1}$$

进一步，本文选取*R*为可靠性得分阈值，并构造*R*可靠相互作用网络*PINs*|$_R$=(*PN*, *PP*|$_R$, *s*)，其中*PP*|$_R$={(*i*,*j*)|*s*(*i*, *j*)≥*R*, *i*, *j*=1, ..., *N*}。设*AR*为*R*可靠相互作用网络*PINs*|$_R$的邻接矩阵，则邻接矩阵*AR*的元素$ar_{i,j}$按式(2)计算。

$$ar_{i,j}=\begin{cases}1 & , \ if \quad s(i,j)\ge R\\ 0 & , \ otherwise\end{cases} \tag{2}$$

其中，*i*, *j*=1, ..., *N*, *R*∈{1, ..., 999}[3]。

2.2 基因表达及蛋白质活跃

假设有*N*个基因在*T*个时刻经归一化的表达值，则基因表达数据可用矩阵*GEV*：$R^{N\times T}$表示，记*GEV*={$gev_{i,t}$}，*i*=1，...，*N*，*t*=1，...，*T*。

假设基因*i*及其对应的基因表达数据为$gev_{i,t}$，*t*=1，...，*T*，令$\overline{gev_i}=\frac{1}{T}\sum_{t=1}^{T}gev_{i,t}$表示基因*i*的基因表达平均值，于是用$ap_{i,t}$表示基因*i*编码的蛋白质在时刻*t*的活跃情况，如此形成矩阵*AP*：{0，1}$^{N\times T}$。当时刻*t*基因*i*的表达值大于等于平均值时，$ap_{i,t}$=1表示基因*i*编码的蛋白质在时刻*t*活跃，否则$ap_{i,t}$=0表示基因*i*编码的蛋白质在时刻*t*不活跃，其定义如式(3)所示。

$$ap_{i,t}=\begin{cases}1 & , \ if \quad gev_{i,t}\ge\overline{gev_i}\\ 0 & , \ otherwise\end{cases} \tag{3}$$

其中，*i*=1, ..., *N*，*t*=1, ..., *T*。

2.3 基因表达相关性

假定有基因*i*和*j*，它们对应的基因表达数据为$gev_{i,t}$和$gev_{j,t}$，*t*=1，...，*T*。基因*i*和*j*的表达相关性可用皮尔森相关系数(Pearson correlation coefficient，*pcc*)度量，于是基因*i*和*j*的皮尔森相关系数可按式(2-4)计算[4]。

$$pcc(i,j)=\frac{\sum_{t=1}^{T}(gev(i,t)-\overline{gev_i})(gev(j,t)-\overline{gev_j})}{\sqrt{\sum_{t=1}^{T}(gev(i,t)-\overline{gev_i})^2}\sqrt{\sum_{t=1}^{T}(gev(j,t)-\overline{gev_j})^2}} \tag{4}$$

其中，$\overline{gev_i}$和$\overline{gev_j}$分别表示基因*i*和*j*的基因表达平均值。

式(4)量化了两个基因在整个采样周期之间的表达相关性，而基因编码的蛋白质相互作用则必然是相应基因在相同时刻短期（前后共5个时间点）内共同表达，因此需要按式(5)量化以某时刻*t*为中心短期内两个基因的表达相关性。

$$pcc(i,j)^t=\frac{\sum_{k=t-2}^{t+2}(gev(i,k)-\overline{gev_i^t})(gev(j,k)-\overline{gev_j^t})}{\sqrt{\sum_{k=t-2}^{t+2}(gev(i,k)-\overline{gev_i^t})^2}\sqrt{\sum_{k=t-2}^{t+2}(gev(j,k)-\overline{gev_j^t})^2}} \tag{5}$$

其中，$\overline{gev_i^t}=\frac{1}{5}\sum_{k=t-2}^{t+2}gev_{i,k}$，$\overline{gev_j^t}=\frac{1}{5}\sum_{k=t-2}^{t+2}gev_{j,k}$。

对于连续时间点活跃的蛋白质，其编码基因*i*应该在连续时间点上均有显著表达，因此可按式(6)计算.

$$pcc(i,i)^t=\frac{\sum_{k=t-2}^{t+2}(gev(i,k)-\overline{gev_i^t})(gev(i,k+1)-\overline{gev_i^{t+1}})}{\sqrt{\sum_{k=t-2}^{t+2}(gev(i,k)-\overline{gev_i^t})^2}\sqrt{\sum_{k=t-1}^{t+3}(gev(i,k)-\overline{gev_i^{t+1}})^2}} \tag{6}$$

其中，$\overline{gev_i^{t+1}}=\frac{1}{5}\sum_{k=t-1}^{t+3}gev_{i,k}$。

该编码基因i在连续时间点上短期基因表达自相关性。

基于上述对数据集的处理，接下来首先提出时序动态蛋白质相互作用网络链(Temporal Dynamic Protein Interaction Network Chain, TDPINC)的构造及其数据结构，然后介绍基于分块矩阵的马尔可夫聚类MCL的优化方案。

3 核心概念

3.1 时序动态蛋白质相互作用网络链 TDPINC

给定时序动态蛋白质网络链$G=(V, E)=\{G^1, G^2,\ldots, G^t,\ldots,G^T\}$，其中子图$G^t=(V^t,E^t)$，$V^t=\{i|ap_{i,t}=1\}$表示时刻$t$活跃蛋白质结点集，$V=\bigcup_{t=1}^{T}V^t$表示任意时刻活跃蛋白质结点集，$E^t=\{(i,j)|s(i,j)\geq R, ap_{i,t}\times ap_{j,t}=1\}$表示时刻$t$活跃蛋白质$R$可靠相互作用集合，$E^t$中的元素是无向边，也称为网内边，用$(<v^t, v^{t+1}>|ap_{v,t}\times ap_{v,t+1}=1)$表示由时刻$t$活跃且时刻$t+1$也活跃的结点$v$构成的跨时间点的有向自边，也称为跨网边，于是有：

$$E=(\bigcup_{t=1}^{T}E^t)\cup(\bigcup_{t=1}^{T-1}\{<v^t,v^{t+1}>|ap_{v,t}\times ap_{v,t+1}=1\}),\ i, j, v=1, \ldots, N, t=1, \ldots, T$$

。实质上，E是由所有的网内边和跨网边构成的集合。由于存在跨网边，这样的一组有序图集构成了一个逻辑上连接的网络链（如图1所示），图中红色边即为跨网边。如果任意两个相邻子图间均有跨网边，则所有子图便连成一个逻辑整体。E中所有边的权值可按式7计算。

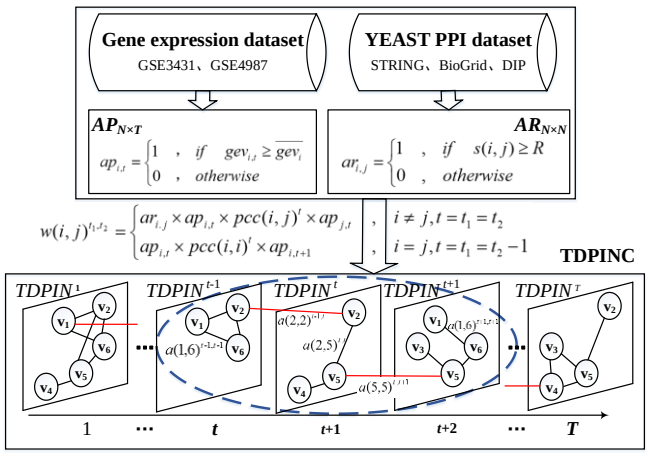

图 1 时序动态蛋白质相互作用网络链

$$w(i,j)^{t_1,t_2}=\begin{cases}ar_{i,j}\times ap_{i,t}\times pcc(i,j)^t\times ap_{j,t} & ,\ i\neq j, t=t_1=t_2\\ ap_{i,t}\times pcc(i,i)^t\times ap_{i,t+1} & ,\ i=j, t=t_1=t_2-1\end{cases} \tag{7}$$

其中，$i=1,...,N$，$j=1,...,N$，$t, t_1, t_2=1,...,T$。

假设$A^{t,t}$为G^t的邻接矩阵，其中$t=1,...,T$，于是邻接矩阵$A^{t,t}$的元素$a(i,j)^{t,t}$可按式(3-1)取$w(i,j)^{t,t}$的值。类似地，用跨网矩阵$A^{t,t+1}$反映跨网边，$t=1,...,T-1$，跨网矩阵$A^{t,t+1}$的元素$a(i,j)^{t,t+1}$可按式(3-1)取$w(i,j)^{t,t+1}$的值。

既然时序动态蛋白质网络链G是一个逻辑整体，因此可用单一矩阵A表示，于是有

$$A=\begin{pmatrix}A^{1,1} & A^{1,2} & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & A^{2,2} & A^{2,3} & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & \ddots & \ddots & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & A^{t,t} & A^{t,t+1} & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & A^{t+1,t+1} & A^{t+1,t+2} & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & \ddots & \ddots & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & A^{T-1,T-1} & A^{T-1,T}\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & A^{T,T}\end{pmatrix}$$

易见，A是一个上三角子块矩阵，而且非零子块分布在正对角线和次对角线子块位置，其中正对角线上的子块矩阵均是对称矩阵，而次对角线上的子块矩阵是非对称的，它反映了时序有向性。

涉及特定生物过程的蛋白质功能模块镶嵌在细胞周期的时序动态蛋白质网络链中，图1中椭圆圈内跨网连接的结点集有可能对应一个蛋白质功能模块。因此，构建单一矩阵A能保持蛋白质功能模块在时序动态蛋白质网络链G中的逻辑完整性。

3.2 马尔可夫聚类 MCL

马尔可夫聚类[5]以模拟网络流的随机游走方式，对网络转移概率矩阵交替地执行扩展和膨胀操作，以强化稠密连接区域的网络流，弱化稀疏连接区域的网络流，从而实现网络流随机游走概率的分配与分化，最终根据不同的概率完成图的划分并达到聚类的目的。本文将马尔可夫聚类过程用于时空动态蛋白质网络实现功能模块识别。

扩展操作的实质是执行矩阵幂乘运算。设有方阵P，则方阵P的n次幂运算为：$P^n=\overbrace{P\cdot P\cdot\ldots\cdot P}^{n}=P^{n-1}\cdot P$。本文通过循环调用矩阵乘实现扩展操作，并将整个扩展操作简记为：$\mathscr{E}:P\rightarrow P'$。

膨胀操作实质是对矩阵每个元素执行r次方后在列方向上进行归一化。设矩阵P'：$R^{N\times N}$和非负实数r，经膨胀操作后的矩阵为P''，P_{ij}'和P_{ij}''分别表示矩阵P'和P''的元素，于是矩阵元素P_{ij}''可按下式计算。

$$P_{ij}''=\frac{(P_{ij}')^r}{\sum_{k=1}^{N}(P_{kj}')^r} \tag{8}$$

本文将整个膨胀操作简记为：$\mathscr{F}:P'\rightarrow P''$。

对于网络链G的邻接矩阵A，MCL算法需要进行幂乘运算。由于A是由子块矩阵构成的，因此MCL算法中对A的幂乘运算就要采用分块矩阵乘法。另外，考虑A的子块仅出现在主对角线和上三角的次对角线位置，为此，基于分块矩阵MCL算法的递推分块矩阵乘可以按下述优化。

假定要计算矩阵A的幂乘A^n，按定义有：$A^n=\overbrace{A\cdot A\cdot\ldots\cdot A}^{n}=A^{n-1}\cdot A$。现将$A$的幂乘运算按累乘的方式表示成递推关系：$\begin{cases}C^1=A\\ C^{k+1}=C^k\cdot A,\quad k=1,\ldots,n-1\end{cases}$，此处$k$表示递推的轮次。

由于A是分块矩阵，因此C^{k+1}的计算是逐个子块进行，于是C^{k+1}的第I行第J列子块C_{IJ}^{k+1}可按式9计算。

$$C_{IJ}^{k+1}=\sum_{K=I}^{J}C_{IK}^{k}A^{K,J},\quad 1\le I\le J\le T,\quad k=1,\ldots,n-1 \tag{9}$$

为了加快收敛速度，可以将最新计算出来的C_{IJ}^{k+1}尽早用于后续子块矩阵的计算，于是将C_{IJ}^{k+1}的计算修改如下式。

$$C_{IJ}^{k}=\sum_{K=I}^{J}C_{IK}^{k}A^{K,J},\quad 1\le I\le J\le T,\quad k=1,\ldots,n-1 \tag{10}$$

在k一定时，C_{IJ}^{k}的计算过程可以按如下递推式计算。

$$C_{IJ}^{k}=\begin{cases}C_{II}^{k}A^{I,I}, & 1\le I=J\le T\\ \sum_{K=I}^{J}C_{IK}^{k}A^{K,J}, & 1\le I<J\le T\end{cases},k=1,\ldots,n-1 \tag{11}$$

可见，C^k_{IJ}的计算过程是可沿对角线方向，从主对角线向右上角逐个子块计算，计算过程如图 2 所示。红色箭头反映在同一斜线上子块计算的推进顺序，绿色箭头反映斜线子块计算的推进顺序。

3.3 基于 GPU 加速的 MCL

马尔可夫聚类算法的扩展操作是矩阵幂乘运算，它是MCL算法中耗时最多的操作。对于上述分块矩阵A的幂乘运算，上面的分析讨论提供了解决方案。显然这样的解决方案需要进行大量的矩阵乘法运算，利用GPU加速的矩阵乘法是高效实现解决方案的关键。考虑到GPU存储容量不足的情况，本文定义了支持自适应分块矩阵乘方式的模块SBBMM，模块SBBMM是以模块mmgpu为核心通过多次调用以实现矩阵乘法。

本文基于分块矩阵 MCL 的优化方案，设计实现识别 TDPINC 中的功能模块的算法 IFM-TDPINC(Identifying Functional Modules in Temporal Dynamic Protein Interaction Network Chain)。

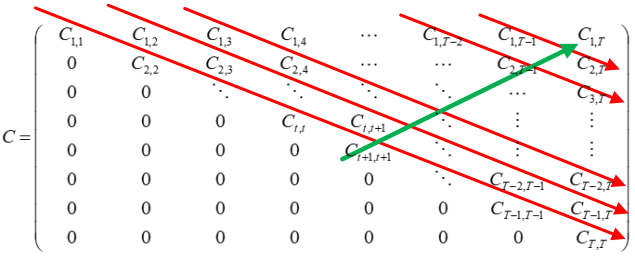

图 2 递推分块矩阵乘的子块矩阵计算顺序

4 算法 IFM-TDPINC 设计

4.1 算法思想

本文算法 IFM-TDPINC 首先将基因表达数据与蛋白质相互作用数据集成产生时序动态蛋白质相互作用网络链 TDPINC，然后将 TDPINC 看成一个整体并用分块矩阵 A 表示，随后利用经优化的基于分块矩阵的 MCL 算法划分 TDPINC 以实现蛋白质簇聚类，最后对所有蛋白质簇进行除重处理形成最终的蛋白质功能模块。由于基于分块矩阵的 MCL 算法需要进行大量的矩阵乘法运算，因此本文运用 GPU 加速其中的矩阵乘法运算。

4.2 算法描述

实现基于分块矩阵的 MCL 算法识别时序动态蛋白质相互作用网络链功能模块算法 IFM-TDPINC 的流程图如图 3 所示。

算法 1 中第 2 行是要确保后续构造时序动态蛋白质相互作用网络链是在 R 可靠连接边的基础上进行的。

算法IFM-TDPINC的伪代码见算法1。

算法1：算法IFM-TDPNC

输入：基因表达数据，已打分*PPI*数据。

输出：蛋白质功能模块集

Begin

生成蛋白质活跃矩阵*AP*和蛋白质*R*可靠相互作用矩阵*AR*

For *t*=1 to *T*

计算 $a(i,j)^{t,t}$ 以构造 $A^{t,t}$

Endfor

For *t*=1 to *T*-1

计算 $a(i,i)^{t,t+1}$ 以构造 $A^{t,t+1}$

Endfor

调用基于分块矩阵的MCL算法

蛋白质簇除重

输出蛋白质功能模块

End

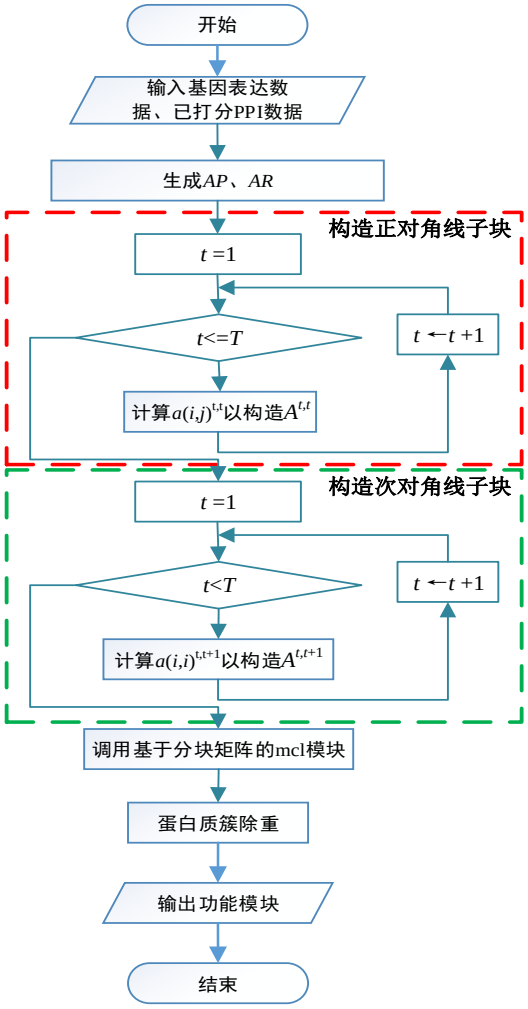

图 3 算法 IFM-TDPNC 的流程图

算法 2：基于分块矩阵 MCL 算法

输入：矩阵 A

输出：结点聚类

```
1   Begin
2.    将 A 转换为转移概率矩阵 P
3.   P′←P
4.   For k=1 to n-1
5.     For L=1 to T
6.       For I=1 to T-L+1
7.         J←I+L-1
8.         For K=I to J
9.           调用模块SBBMM和MulADD。
10.        Endfor
11.      Endfor
12.    Endfor
13.  Endfor
14.  逐元素计算r次方，并在列方向累加。
15.  逐元素归一化。
16.  若P″未收敛，则P←P″，然后转3。
17.  输出结点聚类
18.  End
```

算法2中的第3行相当于执行P的1次幂运算，4-13行循环n-1次执行矩阵累乘($P'\leftarrow P'\cdot P$)以完成n次幂运算，因此整个3-13行实现了扩展操作。第14-15行是实现膨胀操作的两个步骤，每个步骤都要对矩阵的元素扫描一次。基于分块矩阵的MCL算法的程序流程图如图4所示，其中基于GPU的矩阵乘模块SBBMM的程序流程图如图5所示，模块SBBMM中调用的基于GPU的矩阵乘法模块mmgpu的程序流程图如图6所示。

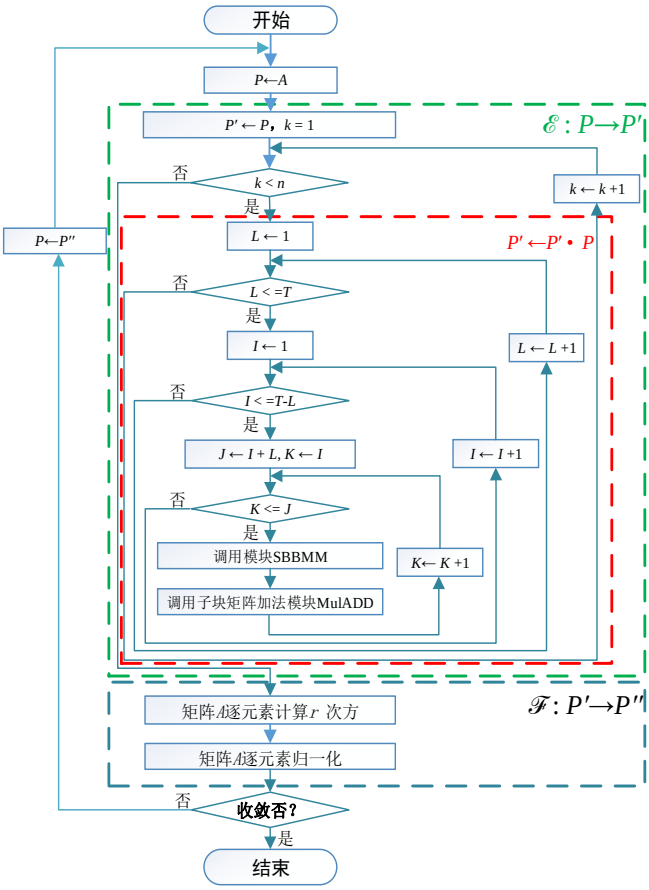

图 4 基于分块矩阵的 MCL 算法的程序流程图

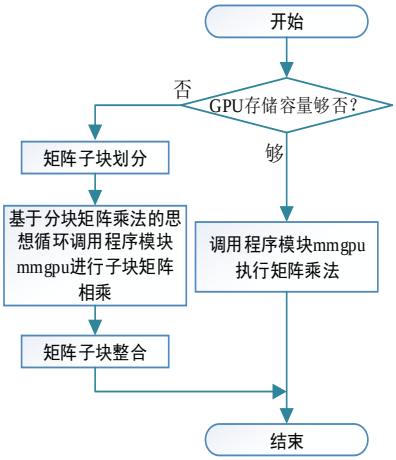

图 5 基于 GPU 的分块矩阵乘法模块 SBBMM 的程序流程图

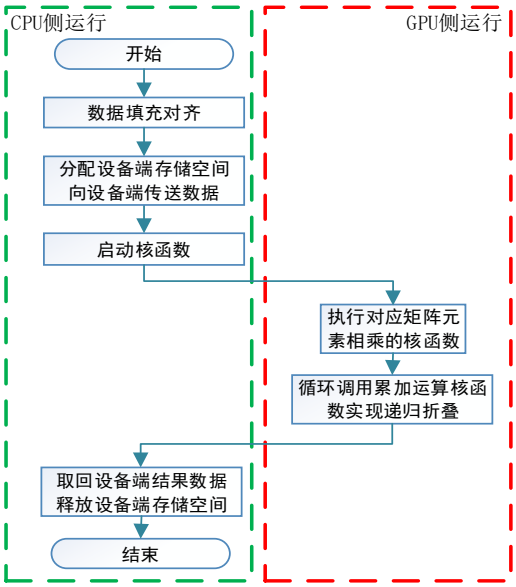

图 6 基于 GPU 的矩阵乘法模块 mmgpu 的程序流程图

5 实验

5.1 实验环境

本文算法实验测试的硬件和软件环境主要配置如下：

硬件环境：

处理器： Intel(R) Core(TM) i7-10510U CPU @1.80GHz 2.30GHz

内存容量：20 GB

GPU： NVIDIA GeForce MX230

软件环境：

操作系统类型：Windows 10，64位操作系统

编程环境：Microsoft Visual Studio 2017

算法采用C++/CUDA编程语言实现。

5.2 实验数据集

在实验中，本文选择模式生物酿酒酵母的3个蛋白质相互作用PPI数据集。第一个数据集来自于STRING数据库版本10[3]，其中包含6418个蛋白质和939998对交互，每对交互均带有可靠性得分数据。第二个数据集由5811个蛋白质和256516交互组成，其来源于BioGrid数据库3.4.128版本的酵母PPI数据[6]。第三个酵母PPI数据集源于DIP数据库，其发布日期为2015/07/01，其中包含5022个蛋白质和22381对交互[7]。

实验选择的基因表达数据集分别为GSE3431[8]和GSE4987[9]。数据集GSE3431中的每个基因包含36个以25分钟间隔收集的原始基因表达数据。每连续12个采样值构成一个周期的基因表达数据，共形成3个周期的基因表达谱。本文对3个周期中对应时间点的基因表达值进行平均，每个基因形成12个表达平均值。对于数据集GSE3431中每个基因，本文对这12个基因表达平均值进行归一化后去分析计算短期皮尔森相关系数，并构造时序动态蛋白质相互作用网络链TDPINC。数据集GSE4987是野生型W303a细胞的基因表达数据，它是以5分钟间隔、2小时一个周期采样2个周期获得的数据，因此数据集GSE4987包含每个周期25个、共两个周期的50个原始基因表达数据。类似地，对于数据集GSE4987中的每个基因，可计算得到25个基因表达平均值并进行归一化。同样本文使用GSE4987中每个基因的25个归一化后的表达值去计算短期皮尔森相关系数，并构造时序动态蛋白质相互作用网络链TDPINC。

此外，本文从http://current.geneontology.org/products/pages/downloads.html下载酿酒酵母的GO标注文件，然后按生物过程标注生成金标准功能模块集，其中包含2117个蛋白质功能模块，它被用作基准数据计算基于匹配的评价指标[10]。

5.3 实验结果比对

为评价算法IFM-TDPINC的性能，本文和已有同类算法ClusterONE[11]、MCODE[12]、SPICi[13]、MCL[9]、APcluster[14]、NCMine[15]和TICONE[16]进行了实验性能比较分析。实验在3个不同PPI数据集DIP、BioGrid和STRING与2个不同基因表达集GSE3431和GSE4987的各种组合中进行，各种算法准确识别出的不同规模功能模块的数量分布对比见表1，各种算法识别出的功能模块质量评价指标[10]对比见表2。

从表1结果可见，本算法IFM-TDPINC能准确地识别出的蛋白质功能模块的数量，仅在PPI数据集BioGrid和DIP上少于算法SPICi，其余情况均较多。从表1结果易见，各种算法准确识别出的蛋白质功能模块的规模均较小，且大多数为2-3。此外，算法MCODE、MCL、NCMine和TICONE未能在PPI数据集STRING上准确识别出任何蛋白质功能模块，算法NCMine和TICONE未能在PPI数据集DIP上准确识别出任何蛋白质功能模块，算法MCODE未能在PPI数据集BioGrid上准确识别出任何蛋白质功能模块。从整体看，本算法IFM-TDPINC准确地识别出的蛋白质功能模块的数量虽然不是最多，但仍具有一定程度准确识别蛋白质功能模块的能力。

从表2结果可见，算法SPICi在PPI数据集BioGrid上整体占优，在PPI数据集BioGrid和STRING上也获得*Acc*、*MMR*和*FAM*的最高值。本文算法IFM-TDPINC在敏感度*Sn*指标上表现较差，仅在*PPV*指标上一致占优。就综合指标*#PM*×*FAM*而言，本文算法IFM-TDPINC在稠密数据集STRING上有较好表现，在稀疏数据集DIP上表现不如算法SPICi和MCL，在稠密度居中的数据集

BioGrid上表现仅次于算法SPICi而好于其它算法。

综上所述，本文算法IFM-TDPINC具有一定程度准确识别蛋白质功能模块的能力，在稠密PPI数据集上识别蛋白质功能模块的表现相对较好，在敏感度*Sn*指标上表现较差，在阳性预测值*PPV*指标上占有优势。

.

表 1 准确识别出的不同规模功能模块的数量分布

PPI datasets	methods	Expression datasets	准确识别出的不同规模功能模块的数量							总数
			规模 2	3	4	5	6	7	8	
STRING	IFM-TDPINC	GSE34	9	1	0	0	0	0	0	10
		GSE49	16	0	0	0	0	0	0	16
	APcluster		0	0	0	0	0	1	0	1
	ClusterONE		0	0	0	0	0	0	0	0
	SPICi		4	2	1	0	0	1	0	8
	MCL		0	0	0	0	0	0	0	0
	NCMine		0	0	0	0	0	0	0	0
	TICONE		0	0	0	0	0	0	0	0
	MCODE		0	0	0	0	0	0	0	0
BioGrid	IFM-TDPINC	GSE3431	18	1	1	1	0	0	0	21
		GSE4987	13	0	0	0	0	0	0	13
	APcluster		4	2	0	3	0	0	0	9
	ClusterONE		1	1	0	1	0	0	0	3
	SPICi		18	10	3	1	2	0	0	34
	MCL		1	0	0	0	0	0	0	1
	NCMine		0	1	1	0	0	0	0	2
	TICONE		2	0	0	0	0	0	0	2
	MCODE		0	0	0	0	0	0	0	0
DIP	IFM-TDPINC	GSE3431	10	0	1	1	0	0	0	12
		GSE4987	7	0	0	0	0	0	0	7
	APcluster		8	0	0	0	0	0	0	8
	ClusterONE		6	0	0	1	0	0	0	7
	SPICi		8	0	2	2	0	0	0	12
	MCL		7	1	0	1	0	0	1	10
	NCMine		1	1	0	1	0	0	0	3
	TICONE		0	0	0	0	0	0	0	0
	MCODE		0	0	0	0	0	0	0	0

表 2 各种算法识别出功能模块的评价指标对比

PPI datasets	methods	Expression datasets	*prec*	*rec*	*fm*	*Sn*	*PPV*	*Acc*	*MMR*	*FAM*	*#PM×FAM*
STRING	IFM-TDPINC	GSE3	0.44	0.15	0.23	0.13	0.83	0.33	0.10	0.58	5.8
		GSE4	0.38	0.18	0.25	0.12	0.82	0.32	0.11	0.61	9.76
	APcluster		0.32	0.10	0.15	0.35	0.46	0.40	0.08	0.59	0.59
	ClusterONE		0.17	0.08	0.11	0.47	0.41	0.44	0.08	0.60	0
	SPICi		0.34	0.17	0.22	0.40	0.51	0.45	0.11	0.73	5.84
	MCL		0.05	0.003	0.005	0.96	0.17	0.4	0.005	0.41	0
	NCMine		0.04	0.05	0.05	0.53	0.18	0.31	0.08	0.45	0
	TICONE		0.15	0.05	0.07	0.14	0.35	0.22	0.07	0.34	0
	MCODE		0.00	0.00	NAN	0.24	0.16	0.20	0.01	0.21	0
BioGrid	IFM-TDPINC	GSE3431	0.50	0.21	0.30	0.19	0.78	0.38	0.13	0.72	15.12
		GSE4987	0.42	0.17	0.24	0.11	0.84	0.31	0.10	.0.59	7.67
	APcluster		0.34	0.21	0.26	0.30	0.53	0.40	0.14	0.74	6.66
	ClusterONE		0.41	0.22	0.29	0.30	0.55	0.41	0.13	0.76	2.28
	SPICi		0.48	0.31	0.38	0.32	0.66	0.46	0.18	0.95	32.3
	MCL		0.28	0.07	0.11	0.30	0.33	0.31	0.05	0.44	0.44
	NCMine		0.05	0.11	0.07	0.30	0.24	0.27	0.10	0.47	0.94
	TICONE		0.27	0.07	0.11	0.09	0.53	0.22	0.07	0.36	0.72
	MCODE		0.08	0.001	0.003	0.21	0.11	0.15	0.007	0.16	0
DIP	IFM-TDPINC	GSE343	0.51	0.15	0.23	0.12	0.85	0.32	0.09	0.56	6.72
		GSE498	0.48	0.12	0.19	0.09	0.89	0.28	0.07	0.47	3.29
	APcluster		0.31	0.23	0.26	0.20	0.51	0.32	0.14	0.69	5.52
	ClusterONE		0.36	0.17	0.23	0.16	0.68	0.33	0.11	0.61	4.27
	SPICi		0.53	0.23	0.32	0.18	0.78	0.38	0.14	0.74	8.88
	MCL		0.25	0.23	0.24	0.18	0.54	0.31	0.15	0.69	6.9
	NCMine		0.35	0.32	0.33	0.20	0.55	0.33	0.17	0.81	2.43
	TICONE		0.27	0.02	0.04	0.04	0.72	0.17	0.02	0.21	0
	MCODE		0.48	0.04	0.07	0.10	0.47	0.21	0.03	0.28	0

6 结束语

本文研究实现了一个基于 GPU 加速的分块矩阵 MCL 识别时序动态蛋白质相互作用网络功能模块的算法 IFM-TDPINC。该算法具有一定程度准确识别相当数量蛋白质功能模块的能力，且准确识别的蛋白质功能模块规模为 2-3 的较多；在识别稠密 PPI 数据集蛋白质功能模块时，综合指标#PM×FAM 的表现相对较好，阳性预测值 PPV 优于对比算法；但敏感度 Sn 较差，准确识别功能模块的总量有待提高。未来将进一步改进算法以提高精确识别蛋白质功能模块能力以及敏感度。

参 考 文 献

[1] SPIRIN V, MIRNY L. Protein complexes and functional modules in molecular networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(21):12123-12128.

[2] 李敏, 孟祥茂. 动态蛋白质网络的构建、分析及应用研究进展[J]. 计算机研究与发展, 2017, 54(6): 1281~1299.

[3] STRING http://string-db.org/ cgi/download_page.pl[DB].

[4] Goh KI, Cusick ME, Valle D, et al. The human disease network[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104 (21): 8685-8690.

[5] VAN DONGEN S M. Graph clustering by flow simulation[D]. Utrecht: University of Utrecht, 2000.

[6] Stark C, Breitkreutz BJ, Reguly T, et al. BioGRID: a general repository for interaction datasets[J]. Nucleic Acids Research, 2006, 34: 535-539.

[7] Salwinski L, Miller CS, Smith AJ, et al. The database of interacting proteins: 2004 update[J]. Nucleic Acids Research, 2004, 32:449-451.

[8] Tu BP, Kudlicki A, Rowicka M, et al. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes[J]. Science, 2005, 310(5751):1152-1158.

[9] Pramila T, Wu W, Miles S, et al. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle[J]. Genes Dev, 2006, 20(16):2266-2278.

[10] ZHANG J X, ZHONG C, HUANG Y R, LIN H X, et al. A method for identifying protein complexes with the features of joint co-localization and joint co-expression in static PPI networks [J]. Computers in Biology and Medicine, 2019, 111:103333.

[11] Nepusz T., Yu H., Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks[J]. Nature Methods, 2012, 9(5):471-472.

[12] Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks[J]. BMC Bioinformatics, 2003, 4(1):2.

[13] Jiang P., Singh M. SPICi: a fast clustering algorithm for large biological networks[J]. Bioinformatics, 2010, 26(8):1105-1111.

[14] Frey B.J., Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315 (5814):972-976.

[15] Shu T., Kengo K. NCMine:Core-peripheral based functional module detection using near-clique mining[J]. Bioinformatics, 2016, 32(22):3454–3460.

[16] Wiwie C., Kuznetsova I., Mostafa A., et al. Time-resolved systems medicine reveals viral infection-modulating host targets[J], Systems Medicine, 2019, 2(1): 1-9.