PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2010, 37(2): 208~217
www.pibb.ac.cn

一种融合表达谱相关性信息的激活子网辨识算法 *

李 非[1)**] 伯晓晨[2)**] 李 鹏[2)] 虞朝辉[3)] 彭宇行[1)***] 王升启[2)***]

([1)] 国防科学技术大学计算机学院，并行与分布处理国防科技重点实验室，长沙 410073；
[2)] 军事医学科学院放射与辐射医学研究所，北京 100850；[3)] 浙江大学医学院附属第一医院，杭州 310003)

摘要 传统表达谱数据分析方法集中于寻找差异表达基因和共表达基因集合，没有考虑基因表达产物之间已知的相互作用. 近年来在系统生物学的研究中发展了将基因表达谱与蛋白质相互作用网络进行整合分析的方法. 现有方法未能综合考虑基因表达差异性和相关性信息，容易导致辨识结果中重要功能分子缺失且生物学功能相关度不高. 提出一种融合表达谱差异性和相关性信息的激活子网辨识算法，能够在蛋白质相互作用网络中辨识高功能相关度的激活子网. 应用到人免疫缺陷病毒 HIV-1 感染过程的研究，结果表明，该算法可以有效避免仅考虑基因表达差异性所引入的偏差，揭示了高相关性低表达差异基因在相关通路中的关键性作用.

关键词 激活子网，表达谱，模拟退火算法，最大生成子树
学科分类号 Q71，TP39
DOI: 10.3724/SP.J.1206.2009.00519

随着微阵列技术及相关产业的高速发展，基因表达谱已经成为分子生物学中重要的高通量研究手段，高通量基因表达谱数据的积累随之迅速增长. 这些基因表达谱数据反映了基因在不同实验条件下的转录水平，通过后续的数据分析可以辨识参与关键生物过程的基因和相关通路，也可预测未知基因的功能和基因之间的转录调控关系. 传统的表达谱数据所使用的分析方法主要集中于两个方面：a. 基于统计假设检验寻找显著差异表达基因；b. 基于相关性聚类算法寻找共表达基因集合. 前者用于辨识潜在功能相关基因，后者用于挖掘功能相关的基因模块. 这些分析方法强调了基因表达谱的数值关系分析，没有考虑基因表达产物之间已知的相互作用，在一定程度上脱离了通过实验方法获得的大量基因调控知识基础，虽然有助于发现新的调控关系，但同时也不可避免地带来了大量的假阳性结果.

由于蛋白质、功能 RNA 等基因表达产物之间存在着复杂的关系，而这种关系本身又受到时间和空间约束，事实上，在一定条件下往往只有少数基因表达产物的相互关系被激活. 基因表达谱的本质就是特定的基因网络在一定时空约束下的表现. 基于这种观点，近年来，在系统生物学的研究中发展了将基因表达谱与蛋白质相互作用网络进行整合分析的方法[1–12]，以期在已知的蛋白质相互作用网络中辨识在特定条件下发生的相互作用关系子集，即激活子网[13–17] (active subnetwork). 在不同的文献中，该相互作用子集也被称为激活通路[18–19](active pathway)、应答功能模块[20] (responsive functional module)、条件应答子网[21] (condition responsive subnetwork)等.

Ideker 等[15]利用模拟退火搜索算法寻找包含显著差异表达基因的激活子网，并应用于酵母应激反应和 HIV-1 感染过程的分析，揭示了一些关键的通路. 类似的，Sohler 等[22]从初始蛋白集合扩展搜索包含显著差异表达基因的激活子网. 上述方法由于仅使用基因表达差异性对蛋白质进行筛选，并不对蛋白质相互作用进行筛选，可归为点赋权的最大子

* 国家重点基础研究发展计划(973)(2005CB321801)和国家自然科学基金(30600281)资助项目.
** 共同第一作者.
*** 通讯联系人. Tel: 0731-4574888
彭宇行. E-mail: pengyuxing1963@yahoo.com.cn
王升启. E-mail: sqwang@nic.bmi.ac.cn
收稿日期：2009-08-31，接受日期：2009-12-10

图搜索问题，因此也称为基于点的激活子网搜索方法(vertex based methods).

由于基于点的激活子网搜索方法仅采用基因的表达差异性作为评价子网激活度的准则，忽略了某些相对表达差异不显著，但在相关过程中发挥关键性作用的基因，如转录因子或信号蛋白等．针对上述问题，Ideker 等[15]提出一种通过修正搜索过程中子网扩展方法以改进搜索结果的方法．若子网中某基因在网络中具有较多邻居基因，属于 Hub 节点，则子网自动扩展以包含该基因的邻居基因．该方法避免了如下情形的出现，即网络中连接多个显著表达差异基因的 Hub 节点，仅由于自身表达差异不显著而无法被包含在激活子网中，从而间接导致与其相连的显著表达差异基因无法出现在最终的激活子网中．这种修正方法本质上是通过该基因与其他表达差异显著基因的连接关系修正其差异性评价，但由于该方法是针对某些搜索结果作局部偏差修正，缺乏普遍适用性，可能在一些情况下引入假阳性结果．此外，在基于点的激活子网定义中，已知的相互作用网络中的基因被认为都参与了实验相关的功能通路活动，但事实上，数据库中已知的相互作用网络是基因编码产物在各种条件下可能发生相互作用的综合图谱，而在某个特定过程中，基因之间的相互作用并非都被激活，原则上激活子网的相关性评价应只包含激活的相互作用子集合.

另一种方法是使用基因表达相关性对蛋白质相互作用进行筛选，可归为边赋权值的最大子图搜索问题，称为基于边的激活子网搜索方法(edge based methods)．研究表明，基因表达相关性与其相互作用关系参与的生物功能有着密切关联[23]．Han 等利用基因表达水平的相关性对癌症与衰老问题进行了讨论[1, 13]，结果表明，基因表达相关性可以用于辨识参与细胞分裂与分化的相关功能模块．可以认为基因之间表达水平相关性越强，其相互作用激活度越高，参与相关反应的可能性越大．Guo 等[21]提出以基因表达协方差相关系数作为网络中蛋白质相互作用的权值，通过模拟退火搜索算法寻找对应权值较大的连通子网作为激活子网.

基于边的激活子网搜索方法有效包含表达差异不显著但作用关键的基因，其结果反映了参与特定生物过程的蛋白质相互作用子集．但由于基因表达谱较之一般的时间序列要短得多，表达水平之间的相关性可能包含较多的随机成分．此外，维持细胞正常运转的必要过程涉及的蛋白质相互作用也呈现较高的相关性，但这类蛋白质相互作用广泛参与各种生物过程，并不与某生物过程或实验条件存在特异性的相关.

因此，有效的激活子网搜索方法应综合考虑基因表达差异性和基因表达相关性，搜索结果应与相关实验条件具有一定的特异相关性，同时包含表达差异不显著但在相关过程中发挥关键性作用的基因．本文提出一种新的激活子网辨识算法，结合各基因的表达差异性和基因之间的表达水平相关性，建立激活子网的综合辨识准则，有效减少辨识过程中的假阳性结果，并且能够获得由完整的关联信息所揭示的通路信息.

为验证算法有效性，我们将建立的激活子网辨识算法应用于人免疫缺陷病毒(HIV-1)感染过程的研究，基于人类蛋白质相互作用网络，利用 HIV-1 感染相关表达谱，辨识得到 HIV-1 感染相关的宿主蛋白及其相互作用子网．结果表明，由于评价指标融合了基因差异性和相关性所揭示的表达水平信息，有效避免仅考虑基因表达差异性所引入的偏差，揭示了高相关性低表达差异基因在相关通路中的关键性作用.

1 材料与方法

1.1 数据集

人类蛋白质相互作用网络来自 HPRD 数据库(Human Protein Reference Database)[24]，版本为 20070901，共包含 19 418 个基因，37 107 个相互作用．基因表达谱数据来自 GEO 数据库(Gene Expression Omnibus)[25]，登记号为 GSE9927，实验采用 Affymetrix Human Genome U133 Plus 2.0 Array 芯片，共有 54 675 个探针数据点．实验包含 20 个实验样本，其中 9 个样本采自正常 HIV-1 阴性人群，11 个样本采自 HIV-1 阳性人群，数据详细描述参见 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9927．表达谱数据经过标准化处理，并与人类蛋白质相互作用网络整合，去除在相互作用网络中无对应产物的基因以及在表达谱中无对应相互作用数据的基因后，得到 HIV-1 感染相关的宿主蛋白相互作用网络，共包含 8 545 个基因，32 913 个相互作用．整理后的相互作用网络数据和相关表达谱数据见 http://biotech.bmi.ac.cn/ActiveNetwork/data.csv.

结果分析中使用到的 HIV-1 与人类蛋白质之间的相互作用数据来自 HIV-1，Human Protein

Interaction Database[26]，HIV-1 相关通路来自 PID数据库(Pathway Interaction Database)[27]，蛋白质功能注释来自 GO 数据库 (Gene Ontology)[28]，使用 BiNGO[29]工具进行功能富集分析．Ideker 激活子网的计算采用 Cytoscape[30]插件 jActiveModule[15]完成．

1.2 方法

本文提出的激活子网辨识方法的目的在于寻找包含具有显著差异表达性且显著表达相关性的基因子网，方法分为两个步骤：a．建立激活子网的定量评价方法；b．在基因相互作用网络中搜索评价最优的激活子网．设 $G=(V, E)$表示基因相互作用网络，V 为网络 G 中包含所有基因 v_i 的集合，E 为网络 G 中包含的所有相互作用 e_{ij} 的集合，其中 e_{ij} 表示基因 v_i, v_j 之间的相互作用．$M=\{m_{ij}\}$为基因表达谱数据矩阵，其中 m_{ij} 表示基因 v_i 在第 j 组实验条件下的转录表达水平．为建立子网激活度的评价指标，首先计算基因表达差异度和相关度，然后计算子网的差异度和相关度，最后寻找具有显著差异性和相关性的子网作为激活子网．

1.2.1 子网表达差异度的计算方法．

基因表达差异度 $R_{\text{diff}}(v_i)$评价基因 v_i 在不同实验条件下表达水平的差异性．基因表达差异度的计算通常依赖于假设检验的统计显著性，如基于 t 检验[16]或更为复杂的统计模型，如 VERA[31]，计算基因在表达水平上的差异度．以 t 检验为例，基因表达差异度 $R_{\text{diff}}(v_i)$定义为：

$$R_{\text{diff}}(v_i)=1-P(v_i) \tag{1}$$

其中，$P(v_i)$为 t 检验统计显著性．

基于基因表达差异度，采用文献[15]的评价方法，基因网络的表达差异度定义为：

$$R^{\text{net}}_{\text{diff}}(G)=\frac{1}{\sqrt{|V|}}\sum_{\text{for each } v_i\in V}\Phi^{-1}(1-P(v_i)) \tag{2}$$

其中 Φ 为标准正态累积分布函数．由于该指标和网络的节点规模有关，因此需要对 $R^{\text{net}}_{\text{diff}}(G)$进行校正[15]，校正后的指标 $\overline{R^{\text{net}}_{\text{diff}}}(G)$为：

$$\overline{R^{\text{net}}_{\text{diff}}}(G)=\frac{R^{\text{net}}_{\text{diff}}(G)-\mu_k}{\sigma_k} \tag{3}$$

其中 $k=|V|$，μ_k，σ_k 为节点规模为 k 的随机网络差异度分布的均值和方差，其取值需要从背景网络中进行抽样估计．校正后的网络表达差异度指标基本与网络规模无关．

需要注意的是公式(3)给出的差异度 $\overline{R^{\text{net}}_{\text{diff}}}(G)$评价指标为相对值，其计算依赖于网络随机样本的分布趋势，即如果将同一子网内嵌到不同的网络结构中，其取值可能不同．使用校正后的相对值能够为不同网络规模建立一致的评价指标，并且通过对具有显著差异性或相关性的子网赋予较高的评价值，能够有效引导搜索过程，提高搜索效率．

1.2.2 子网表达相关度的计算方法．

基因 v_i, v_j 之间，基因产物之间相互作用的表达相关度 $R_{\text{corr}}(e_{ij})$的评价，基于基因 v_i, v_j 在不同实验条件下表达水平变化相关性．基因 v_i, v_j 表达水平变化相关性强度的定义可基于皮尔逊相关(Pearson′s correlation)指标或其他非参数相关性指标来计算．基因相互作用网络中包含许多可能发生的相互作用，通常仅有一部分相互作用直接参与相关实验过程．由于我们希望通过表达谱数据挖掘特定实验过程相关的激活子网，因此子网表达相关度的评价应基于基因产物间相互作用子集进行相关度计算，而该子集应由最有可能参与相关实验过程的相互作用组成，称为关键性通路 E_{key}．

上述关键性通路可以形式化表述如下：给定网络 $G=(V, E)$，需要确定子网 $G_{\text{key}}=(V, E_{\text{key}})$，其中 $E_{\text{key}}\subseteq E$，使得关键性通路 E' 与相关实验更加特异性相关．在确定关键性通路的过程中，我们考虑如下因素：由于基因之间表达水平相关性越强，参与相关反应的可能性越大[1]，E_{key} 应尽可能保留表达相关度较高的相互作用．同时我们无法从表达谱数据直接推断哪些相互作用直接参与相关实验，在无背景知识情况下，合理假设 E_{key} 为满足 G_{key} 连通性约束条件最小集合．从此定义出发，G_{key} 为由高表达相关度的相互作用构成的连通树，E_{key} 满足最大生成子树(maximum weight spanning tree，MST)定义，可通过确定性算法求解，即 $E_{\text{key}}=E_{\text{MST}}$．

计算子网表达相关度的具体方法是，首先计算相互作用的表达相关度，然后在此边赋权图上计算最大生成子树，并基于最大生成子树所包含的相互作用集合 E_{MST} 计算子网基因表达相关度，其定义为:

$$R^{\text{net}}_{\text{corr}}(G)=\frac{1}{|E_{\text{MST}}|}\sum_{\text{for each } e_{ij}\in E_{\text{MST}}}R_{\text{corr}}(e_{ij}) \tag{4}$$

与子网表达差异度相似，由此计算得到的网络相关度需要进一步校正过程．校正后的基因网络表达相关度定义为：

$$\overline{R^{\text{net}}_{\text{corr}}}(G)=\frac{R^{\text{net}}_{\text{corr}}(G)-\mu_k}{\sigma_k} \tag{5}$$

其中 $k=|V|$，μ_k, σ_k 为节点规模为 k 的随机网络相关度分布的均值和方差，与式(3)类似，其取值

需要从背景网络中进行抽样估计．类似地，公式给出的差异度 $\overline{R^{net}_{corr}}(G)$评价指标也为相对值．

1.2.3　激活子网的搜索方法．

在寻找激活子网的过程中，其子网表达差异度和相关度均为优化目标，因此问题本质是求解多目标优化问题．多目标优化问题的一种经典求解方法是使用目标函数线性聚合，将多个子目标聚合成向量函数，转换成易于求解的单目标优化问题．由于该求解方法属于前决策方法，即对可能解集合的选择决策在前，因此其求解结果为单一解而非 Pareto 最优集合，并且在某些情况下该求解方法无法获得非凸解[32]．但实际应用问题的求解域多为凸集[32]，并且在无有效后决策方法或高效求解方法的情况下，采用前决策方法预先对结果进行筛选，不失为高效易行的方法．

此外，即使转换为单目标优化问题，激活子网的搜索也属于最优连通子网求解问题(maximum weight connected subgraph problem)，即给定评价函数，在网络中寻找最优连通子网，使得其评价函数取得最大值．该问题属于 *NP* 问题[15]，在多项式时间内尚无有效的解决方法，需要采用随机优化算法，如模拟退火算法，最终获得的满意解能够有效逼近该问题的最优解．理论上还可以进一步证明，在恰当的参数设置下，模拟退火算法能够得到全局最优解，或者相当理想的次优解[33]．

还需要考虑的问题是：如果多目标之间的取值范围存在较大的数量级差异，在线性聚合目标函数中，某些目标的变化可能由于数值精度原因而被忽略，无法有效引导搜索过程．本文中子网差异度和相关度均为无量纲统计显著值，具有相似的统计分布趋势，从而避免了上述问题．

基于上述讨论，本文采用线性聚合函数方法将子网表达差异度和相关度转换为网络激活度，并使用模拟退火搜索算法寻找激活度较高的子网作为激活子网．基因网络激活度 $R^{net}_{active}(G)$的评价指标定义为网络基因表达差异度和网络基因表达相关度的线性加权之和：

$$R^{net}_{active}(G) = \overline{R^{net}_{diff}}(G) + \lambda \overline{R^{net}_{corr}}(G) \qquad (6)$$

其中 $\overline{R^{net}_{diff}}(G)$为基因网络表达差异度，$\overline{R^{net}_{corr}}(G)$为基因网络表达相关度，$\lambda$ 为权重因子．在多目标优化问题中，λ 体现了对不同目标重要性的决策．就本文而言，在无背景决策的情况下，以下计算中设 λ=1，也可调整 λ 寻找其他可行解．

激活子网搜索算法描述如下，设 G=(V, E)为相互作用网络，退火温度 $T = T_{start}$：

1．生成随机化网络 G_w=(V_w, E_w)，其中 $V_w \subset V$, $E_w \subset E$；

2．随机选取节点 $v \in V$，若 $v \in V_w$，则 V_w' = V_w− {v}，否则 V_w'=$V_w \cup$ {v}；

3．生成新的网络 G_w'=(V_w', E_w')，其中，E_w' = {$e_{ij}|v_i, v_j \in V_w, e_{ij} \in E_w$}；

4．构建 G_w'的最大生成子树，根据公式(6)，计算 G_w'的激活度 $R^{net}_{active}(G_w')$

5．以概率 p 决定是否以 G_w'取代 G_w，其中，p=Min{1, $e^{\Delta R/T_i}$}，ΔR=$R^{net}_{active}(G_w')$–$R^{net}_{active}(G_w)$．即若 $R^{net}_{active}(G_w') > R^{net}_{active}(G_w)$，则直接以 G_w'取代 G_w，若 $R^{net}_{active}(G_w') \leqslant R^{net}_{active}(G_w)$，则 G_w'取代 G_w 的概率为 p；

6．重复步骤 2～5，迭代 N 次，同时逐渐降低退火温度 T，以收敛至最终解 G_{opt}．

对于中等规模网络，以上搜索算法能够直接搜索得到较优结果，如果网络规模较大，搜索空间过大，可能使迭代时间难以接受．一种可行的方法是分步迭代优化，即在前一次搜索结果得到的激活子网中再次搜索最优子网，通过每步设置适当的迭代次数，算法能够在合理时间内完成，并给出较优的结果．本文采用了这种分步迭代优化方法．

2　结果与讨论

计算使用 HPRD 中已报道的人类蛋白质相互作用数据和 HIV-1 的表达谱数据，搜索参数设置为：N=10^6，T_{start}=1，T_{end}=0.01，最终结果如图 1 所示．

首次迭代得到规模为 2 440 节点的子网，激活度 $R^{net}_{active}(G_{opt})$=126.71, $\overline{R^{net}_{diff}}(G)$=46.16, $\overline{R^{net}_{corr}}(G)$=80.55. 在此子网中继续进行新的一轮迭代，搜索评价更高的子网，从而进一步缩小激活子网的规模．经过 4 轮分步迭代，最终得到的激活子网包含 94 个基因，激活度 $R^{net}_{active}(G_{opt})$=15.12, $\overline{R^{net}_{diff}}(G)$=5.24, $\overline{R^{net}_{corr}}(G)$=9.88, 详细结果见表 1，其中 $\overline{R^{net}_{diff}}(G)$和 $\overline{R^{net}_{corr}}(G)$定义参见公式(3)、(5)．$N_{before}$，$N_{after}$ 分别为搜索初始网络和结果网络包含节点数．

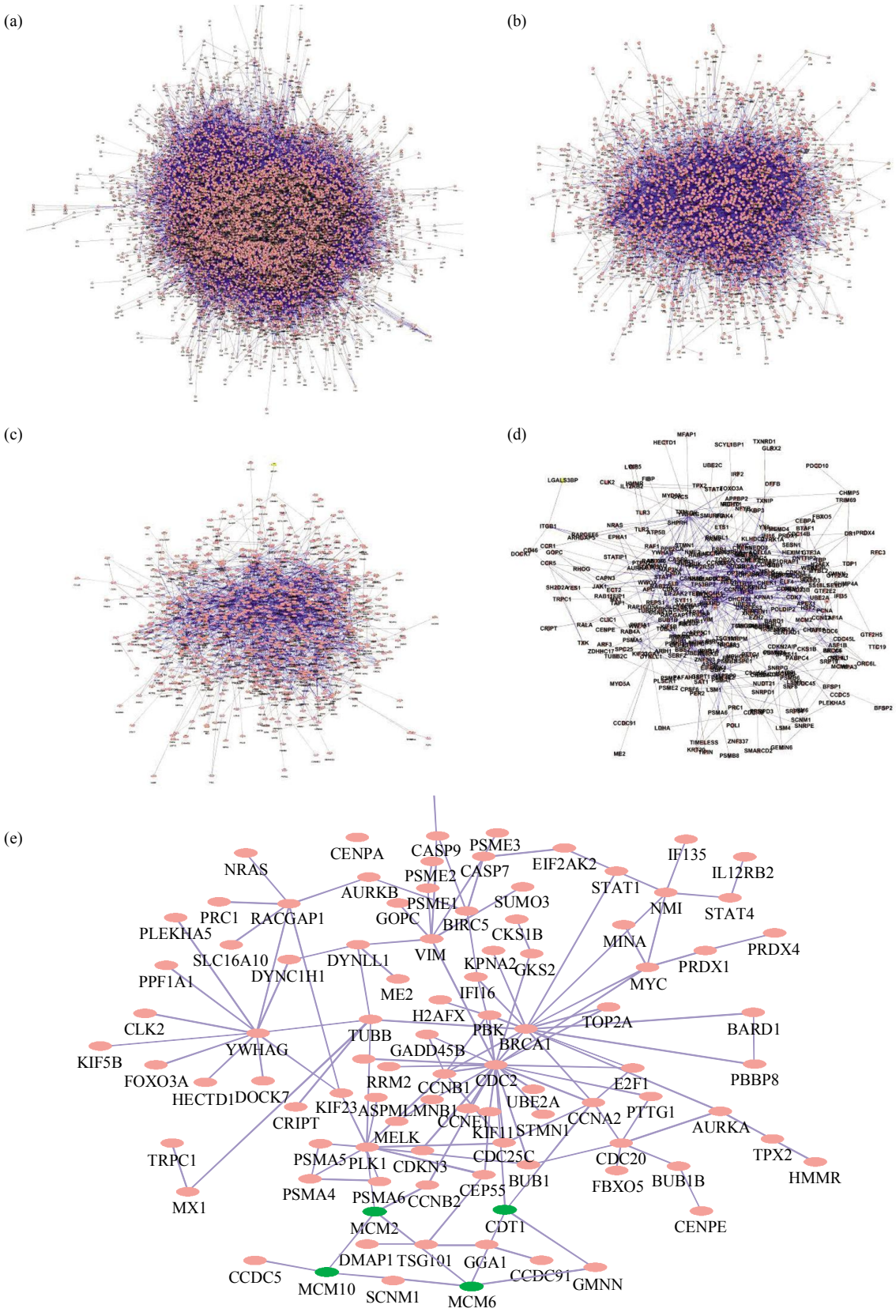

Fig. 1　Active subnetwork in human protein interaction network with HIV-1 expression profile

(a) Corresponding to initial network. (b), (c), (d), (e) Corresponding to active subnetworks after step 1, step 2, step 3, step 4. As shown in the figure, four active subnetworks are derived in sequences. (a) Whole human protein interaction network, with 8545 proteins. (b～e) Active subnetwork search in previous network, i.e., network (b) is searched in network (a), (c) is searched in (b). In (e), the non-differentially expressed but functional important genes CDT1, MCM6, MCM2, MCM10 are green colored.

Table 1　Scores of active subnetwork in each iteration step

Iteration step	N_{before}	N_{after}	$R^{net}_{active}(G_{opt})$	$R^{net}_{diff}(G)$	$\overline{R^{net}_{diff}}(G)$	$R^{net}_{corr}(G)$	$\overline{R^{net}_{corr}}(G)$
1	8 545	2 440	126.71	85.90	46.16	107.13	80.55
2	2 440	961	66.14	78.10	25.54	88.96	40.59
3	961	350	33.13	57.40	13.47	67.69	19.66
4	350	94	15.12	33.40	5.24	45.97	9.88

如图 2 所示，子网的激活度 $R^{net}_{active}(G)$趋向收敛，同时差异度 $\overline{R^{net}_{diff}}(G)$和相关度 $\overline{R^{net}_{corr}}(G)$在搜索过程中一致增大，并随退火温度的减低而逐步趋向收敛，经过 10^6 步迭代算法收敛，当模拟退火温度降至约 0.1 时，算法搜索过程基本平稳，趋近于最终结果.

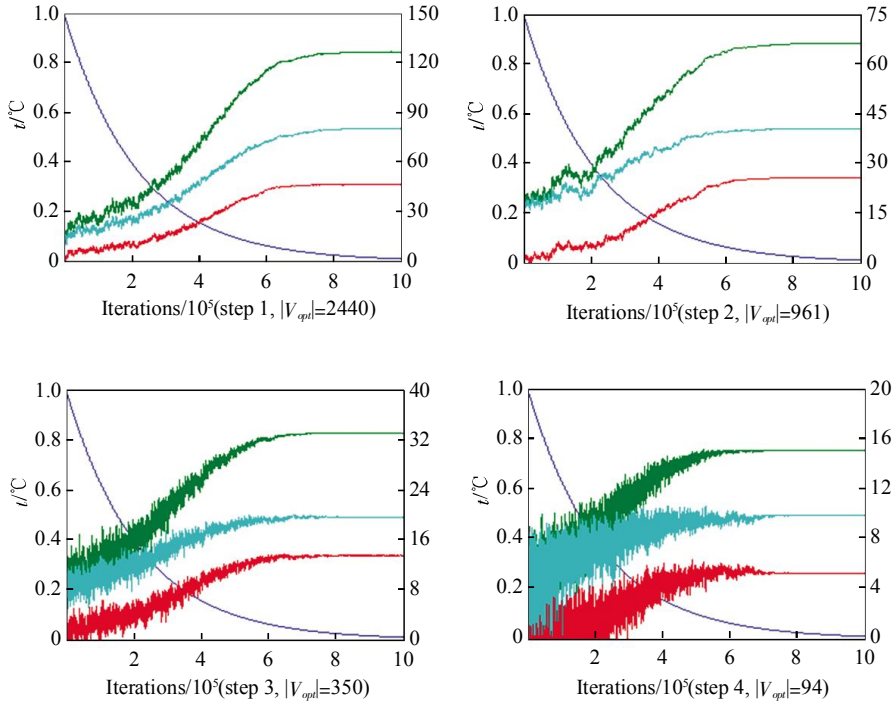

Fig. 2　Score and temperature versus number of iterations

Simulated annealing was performed for the network and expression profile shown in **Figure 2**, with parameters N=10^6, N_{start}=1, N_{end}=0.01. Annealing temperature (left vertical axis; solid blue trace) decreases geometrically over consecutive iterations. By the end of the run，scores for each of the five top scoring subnetworks have increased to a local maximum (right vertical axis; green trace = total network score(R^{net}_{active}), lightblue = network correlation score(R^{net}_{corr}), red = network difference score(R^{net}_{diff})).

根据所采用的表达谱数据，我们从三个方面评价辨识得到的激活子网与 HIV-1 病毒感染过程的相关程度：a．激活子网中与 HIV-1 存在相互作用的蛋白质比例；b．激活子网涉及的蛋白质 GO 功能注释分析；c．激活子网涉及的功能通路分析.

基于 HIV-1 病毒宿主相互作用数据集[26]，统计激活子网中与 HIV-1 病毒蛋白存在直接或间接相互作用的宿主蛋白比例，该比例随着逐步迭代搜索而逐渐增大，见表 2．在最终得到的包含 94 个蛋白质的激活子网中，已报道的直接或间接受到 HIV-1 病毒蛋白的调控蛋白质共有 30 个，占总蛋白质数的 32%．主要涉及 HIV-1 病毒的 Tat、Env 等关键蛋白的作用通路.

Table 2　The number of human proteins interacting with HIV-1 proteins in active subnetwork in each iteration step

All	Interacting with HIV-1	Percentage
8 545	1 029	12%
2 440	406	17%
961	199	21%
350	86	25%
94	30	32%

使用 BiNGO[29]，基于 GOA Human annotations 注释子集[34]，使用超几何分布统计和 Benjamini & Hochberg False Discovery Rate[35]校正方法对激活子网蛋白质进行 GO 功能富集分析．结果表明，激活子网的 94 个蛋白质涉及细胞增殖、细胞分化、细胞周期以及细胞凋亡等相关生物学过程、重点分布在 DNA 代谢过程和细胞骨架组装等过程，如图 3 所示．

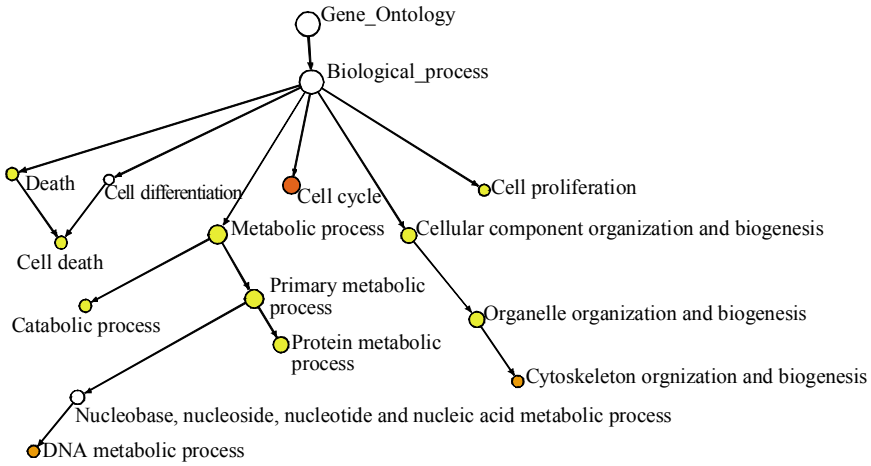

Fig. 3　Term enrichment of GO biological process in active subnetwork versus whole interaction network using BiNGO[29]

使用 PID[27]的通路分析表明，在激活子网包含的 94 个蛋白质中，涉及白介素介导通路、泛素 - 蛋白酶体介导的蛋白质降解、细胞凋亡及细胞周期调控等 HIV-1 显著相关的生物过程．

a．白介素介导的通路．白介素(IL)作为体内免疫调控的重要细胞因子，在调节机体免疫应答及免疫平衡上起着重要的作用，激活子网中的 NRAS、E2F1、IFI35、STAT1、MYC 参与了 IL2、IL6 等介导的通路，其中 STAT 涉及到 IL2、IL6、IL7、IL12、IL23、IL27 等介导的众多的通路．这显示了 HIV-1 病毒对免疫系统广泛的调节．

b．泛素 - 蛋白酶体介导的蛋白质降解．以往的研究表明，泛素 - 蛋白酶体通过破坏 HIV-1 病毒复合物而干扰病毒的入侵，最近的实验表明该通路对 HIV-1 同样发挥着积极的作用，即该通路对病毒的逆转录也十分重要[36]．存在于激活子网中的 PSMA4、PSMA5、PSMA6、PSME1、PSME2、PSME3 等泛素 - 蛋白酶体通路的重要组成分子参与到 CDC25A、CDC20、Cyclin D1、CD4 等众多细胞因子的降解过程．病毒蛋白与泛素 - 蛋白酶体相关通路之间的相互作用表明：HIV-1 可能通过窃取泛素 - 蛋白酶体的通路来完成其脱衣壳及逆转录过程．

c．细胞凋亡．HIV-1 可攻击人体免疫系统的中枢细胞，并诱导淋巴细胞的大量凋亡，从而破坏人体免疫系统，导致机体免疫能力下降．激活子网中的 CASP7、CASP9、CYCS、LMNB1、VIM 组成的 caspase 介导的细胞凋亡通路与之密切相关．

d．细胞周期调控．中心体的组装、解离、调控作为细胞周期过程的重要事件，是维持细胞周期运转所必需的过程．激活子网中 PLK1、CDC2、DYNC1H1、DYNLL1、TUBB、YWHAG 参与到细胞分裂过程中的中心体的组装、成熟及调控的通路．另外，激活子网还涉及 M/G1、G1/S、G2/M 及细胞周期阻滞等众多的细胞周期的调节通路，由于细胞周期检测点的激活和细胞周期在功能上是相互联系的[37]，因此 HIV-1 对人细胞周期的调控对于诱导感染细胞的凋亡至关重要．

实验结果还表明，与 Ideker 方法[15]相比，我们的方法能有效辨识功能重要的非表达差异显著基因，如激活子网中，标记为绿色的蛋白质 CDT1、MCM6、MCM2、MCM10 共同参与了细胞 DNA 合成和复制过程．虽然按照表达差异度从高到低排列，在 94 个激活子网蛋白质中分别位于第 78, 84, 85, 90，与激活子网中其他蛋白质相比其差异表达并不显著，但其由于这些蛋白质之间存在较强的表达水平相关性($P < 10^{-9}$)，最终出现在辨识结果中．此外，由于激活子网评价准则综合考虑了基因表达

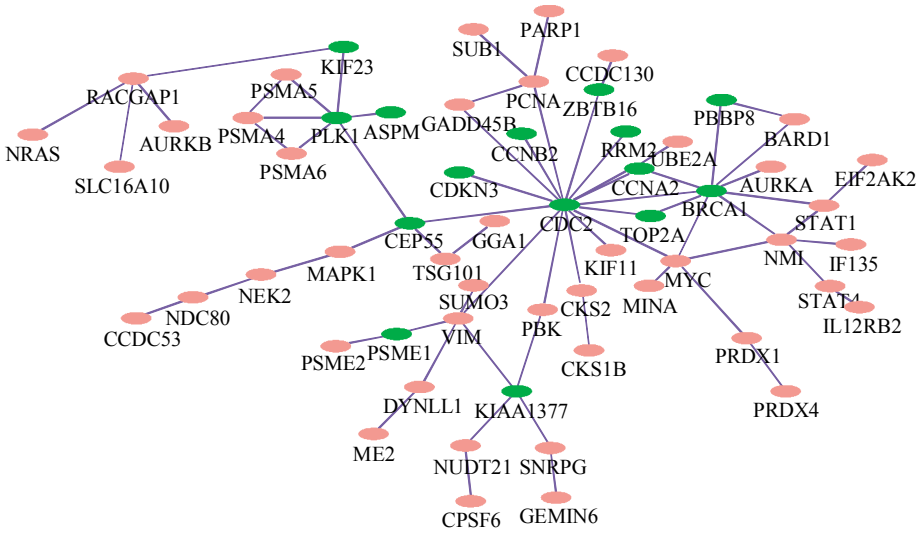

Fig. 4　Active subnetwork in human protein interaction network with HIV-1 expression profile using Ideker method

There are 60 genes in the active subnetwork with 47 genes are also included in the subnetwork by our method (**Figure 1e**). In these genes, 15 genes are non-differentially expressed with $P > 0.001$ (green).

相关性指标，最终结果倾向于包含更多的非表达差异显著基因，见表 3．如图 4 所示，Ideker 方法通过邻居基因的差异显著度修正方法也可以支持辨识部分功能重要的非表达差异显著基因，而由于其辨识过程并未利用表达水平相关性信息，具有一定的局限性．

Table 3　Comparison between different methods for identifying active subnetwork

	Number of genes	Percentage of non-differentially expressed genes($P > 0.001$)
Ideker 方法	60	15(25%)
本文方法	94	36(40%)

综上所述，本文提出的算法能够有效地在蛋白质相互作用网络中，利用表达谱数据辨识和定位相关激活子网，其结果不仅包含显著功能相关的差异基因，还包含高表达相关性但表达差异水平较弱的共表达基因，如图 1e 的 CDT1、MCM6、MCM2、MCM10．此外，由于算法融合了基因差异性和相关性所揭示的表达水平信息，提高了激活子网与已知调控通路的特异相关度，为辨识特定生物过程相关的通路提供了有效工具．

最优连通子网求解问题的本质复杂性使得激活子网搜索不能化简为确定性优化问题，搜索算法包含一定的随机因素．在计算过程中，可以通过多次运行以提高结果的稳定性，此外还可以进一步发展以先验知识为约束的启发式搜索算法，以有效缩小解空间．在某些情况下，用户还可以在搜索前通过人为指定激活子网中必须包含的特定节点，将问题转化为带约束的最优子网搜索问题，使得问题在一定程度得到简化．Lee[38]给出包含特定节点的带约束最优子网问题的启发式搜索算法，能够快速搜索中等规模网络的最优子网．此外，由于激活子网算法得到的激活子网在相当程度上与生物功能或者调控通路关联，因此本方法也可以经过推广后应用于预测未知网络模块的生物功能或者潜在药物靶点，这将是我们后续工作的重点．

参　考　文　献

[1] Xia K, Xue H, Dong D, *et al*. Identification of the proliferation/differentiation switch in the cellular network of multicellular organisms. PLoS Computational Biology, 2006, **2**(11): e145

[2] Zhu M, Gao L, Guo Z, *et al*. Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities. Gene, 2007, **391**(1−2): 113−119

[3] Auffray, C. Protein subnetwork markers improve prediction of cancer outcome. Mol Syst Biol, 2007, 3: 141

[4] Babur O, Colak R, Demir E, *et al*. PATIKAmad: putting microarray data into pathway context. Proteomics, 2008, **8**(11): 2196−2198

[5] Batada N N, Reguly T, Breitkreutz A, *et al*. Still stratus not altocumulus: further evidence against the date/party hub distinction. PLoS Biol, 2007, **5**(6): e154

[6] Camargo A, Azuaje F. Linking gene expression and functional network data in human heart failure. PLoS ONE, 2007, **2** (12):

e1347

[7] Cline M S, Smoot M, Cerami E, *et al*. Integration of biological networks and gene expression data using Cytoscape. Nat Protocols, 2007, **2**(10): 2366-2382

[8] Han J J, Bertin N, Hao T, *et al*. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 2004, **430**(6995): 88-93

[9] Jin G, Zhang S, Zhang X, *et al*. Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. PLoS ONE, 2007, **2**(11): e1207

[10] Kim P M, Sboner A, Xia Y, *et al*. The role of disorder in interaction networks: a structural analysis. Mol Syst Biol, 2008, **4**: 179

[11] Lu X, Jain V V, Finn P W, *et al*. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. Mol Syst Biol, 2007, **3**(1): 98

[12] Ramani A K, Li Z, Hart G T, *et al*. A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. Mol Syst Biol, 2008, **4**: 180

[13] Xue H, Xian B, Dong D, *et al*. A modular network model of aging. Mol Syst Biol, 2007, **3**: 147

[14] Ideker T, Thorsson V, Ranish J A, *et al*. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science, 2001, **292**(5518): 929-934

[15] Ideker T, Ozier O, Schwikowski B, *et al*. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics, 2002, **18**(1): S233-240

[16] Bandyopadhyay S, Kelley R, Ideker T. Discovering regulated networks during HIV-1 latency and reactivation. Pac Symp Biocomput, 2006, **354**: 66

[17] Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res, 2002, **12**(1): 37-46

[18] Tamada Y, Imoto S, Tashiro K, *et al*. Identifying drug active pathways from gene networks estimated by gene expression data. Genome Inform Ser Workshop Genome Inform, 2005, **16** (1): 182-191

[19] Vert J P, Kanehisa M. Extracting active pathways from gene expression data. Bioinformatics, 2003, **19**(2): 238-244

[20] Wu Z, Zhao X, Chen L. Identifying responsive functional modules from protein-protein interaction network. Mol Cell, 2009, **27** (3): 271-277

[21] Guo Z, Li Y, Gong X, *et al*. Edge-based scoring and searching method for identifying condition-responsive protein protein interaction sub-network. Bioinformatics, 2007, **23**(16): 2121-2128

[22] Sohler F, Hanisch D, Zimmer R. New methods for joint analysis of biological networks and expression data. Bioinformatics, 2004, **20**(10): 1517-1521

[23] Wang H, Azuaje F, Bodenreider O, *et al*. Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships//Comput Intell Bioinform Comput Biol, 2004. CIBCB′04. Proceedings of the 2004 IEEE Symposium. California: Institute of Electrical & Electronics Enginee, 2004: 25-31

[24] Mathivanan S, Ahmed M, Ahn N G, *et al*. Human proteinpedia enables sharing of human protein data. Nat Biotech, 2008, **26** (2): 164-167

[25] Barrett T, Suzek T O, Troup D B, *et al*. NCBI GEO: mining millions of expression profiles—database and tools. Nucl Acids Res, 2005, **33**(suppl 1): D562-566

[26] Pruitt K D, Dieffenbach C W, Ptak R G, *et al*. Cataloguing the HIV type 1 human protein interaction network. AIDS Research and Human Retroviruses, 2008, **24**(12): 1497-1502

[27] Schaefer C F, Anthony K, Krupa S, *et al*. PID: the pathway interaction database. Nucl Acids Res, 2009, **37** (suppl_1): D674-679

[28] Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: tool for the unification of biology. Nat Genet, 2000, **25**(1): 25-29

[29] Maere S, Heymans K, Kuiper M. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 2005, **21**(16): 3448-3449

[30] Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 2003, **13**(11): 2498-2504

[31] Ideker T, Thorsson V, Siegel A F, *et al*. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J Comput Biol, 2000, **7**(6): 805-817

[32] Deb K. Multi-objective optimization using evolutionary algorithms. Berkley: Wiley, 2001: 49-52

[33] Lundy M, Mees A. Convergence of an annealing algorithm. Math Program, 1986, **34**(1): 111-124

[34] Camon E, Magrane M, Barrell D, *et al*. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Research, 2004, **32**(Database issue): D262-266

[35] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statistical Society. Series B (Methodological), 1995, **57** (1): 289-300

[36] König R, Zhou Y, Elleder D. *et al*. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. Cell, 2008, **135**(1): 49-60

[37] Andersen J L, DeHart J L, Zimmerman E S, *et al*. HIV-1 Vpr-induced apoptosis is cell cycle dependent and requires bax but not ANT. PLoS Pathog, 2006, **2**(12): e127

[38] Lee H F, Daniel R. Dooly algorithms for the constrained maximum-weight connected graph problem. Naval Research Logistics, 1996, **43**(7): 985-1008

Discovering Active Subnetwork in Protein Interaction Network*

LI Fei[1)**], BO Xiao-Chen[2)**], LI Peng[2)], YU Zhao-Hui[3)], PENG Yu-Xing[1)***], WANG Sheng-Qi[2)***]

([1)] *National Laboratory for Parallel & Distributed Processing, College of Computer, National University of Defense Technology, Changsha* 410073, China;

[2)] *Beijing Institute of Radiation Medicine, Beijing* 100850, China;

[3)] *First Hospital of Zhejiang University Medical College, Hangzhou* 310009, China)

Abstract Traditional analysis of gene expression data focused on identifying differentially expressed and co-expressed genes, which didn′t take known interactions into consideration. In recent years, many methods have been developed to identify active subnetwork by integrating protein interaction networks with gene expression profiles. Current approaches failed to take full account of both difference and correlation in gene expression that may lead to false positive results. A new algorithm is proposed for identifying active subnetwork by considering both difference and correlation of gene expression profile. The algorithm is employed in the process of gene expression profiles of human immunodeficiency virus infection. The results showed that the algorithm can identify the active subnetwork that has extremely high biologically functional connectivity with human immunodeficiency virus, and effectively avoid the bias introduced by considering differences of gene expression profiles only, i.e., genes less differentially expressed are also included due to high correlations in gene expression.

Key words active subnetwork, gene expression data, simulated annealing algorithm, maximum weight spanning tree

DOI: 10.3724/SP.J.1206.2009.00519

*This work was supported by grants from National Basic Research Program of China (2005CB321801) and The National Natural Science Foundation of China (30600281).

**These authors contributed equally to this work.

***Corresponding author. Tel: 86-731-4574888

PENG Yu-Xing. E-mail: pengyuxing1963@yahoo.com.cn

WANG Sheng-Qi. E-mail: sqwang@nic.bmi.ac.cn

Received: August 31, 2009 Accepted: December 10, 2009