基于复合参数的蛋白质网络关键节点识别技术

黄海滨 [1,2] 杨路明 [1] 王建新 [1] 李绍华 [1,3]

摘 要 蛋白质的关键性与它在生物网络中对应节点的拓扑特性紧密相关. 把关键蛋白质识别看作是一类特殊的模式识别, 以分子之间量化的关系 — 拓扑参数作为识别依据: 从相关分析出发对蛋白质网络节点的关键性与其主要拓扑参数的相互关系进行研究, 发现参数对节点关键性识别能力的大小与两者之间的相关性有关; 研究复合参数识别度与独立参数识别度、独立参数相关性之间的关系, 提出复合参数的构造方法及异步识别算法. 仿真结果证实, 获得的识别度明显高于其他识别技术.

关键词 关键节点, 模式识别, 复合参数, 拓扑结构, 蛋白质网络
中图分类号 TP391

Identification Technique of Essential Nodes in Protein Networks Based on Combined Parameters

HUANG Hai-Bin[1,2] YANG Lu-Ming[1] WANG Jian-Xin[1] LI Shao-Hua[1,3]

Abstract The essentiality of a protein is correlated with its topological properties in a bionetwork. Viewing the identification of essential proteins as a special kind of pattern recognition by setting up the quantification of the relationship of molecules — topological parameter, this paper analyzes the correlation between a protein′s essentiality and its main topological parameters, studies the nature of the essential-node-judgement of the parameters, puts forwards theoretically the relation between the identification degree of combined parameters and that of the single parameters involved and the correlation between single parameters, and gives effective construction methods of combined parameters and their asynchronous recognition algorithm. The results show that the identification ability of the technique is obviously greater than that of the others.

Key words Essential node, pattern recognition, combined parameter, topological structure, protein network

生命活动是细胞内大量相互作用的产物, 本质上是蛋白质、DNA、RNA、小分子之间相互作用形成的生物网络的外在表现[1−2]. 不同蛋白质对生命活动的重要性是不一样的. Winzeler 等[3] 将关键蛋白质定义为通过基因剔除式突变 (Null mutation) 将其移除后造成有关蛋白质复合体功能丧失, 并导致生物体无法生存的蛋白质. 关键蛋白质的识别能够从系统水平上为生物学、医学等提供有价值的信息[4−5], 识别方法由最初和最直接的生物实验方法发展到了生物信息学方法[6−7]. 其中, 基于蛋白质网络拓扑参数的方法主要通过节点拓扑参数值的大小进行识别, 它是一种特殊的模式识别: 1) 它根据参数值的大小对节点的关键性进行判别, 识别过程实际上是一种统计分类, 所以属于模式识别的范畴; 2) 它不以节点或蛋白质分子本身的特性, 而以节点或分子之间已经量化的拓扑关系作为识别依据, 所以是一种特殊的模式识别. 目前基于拓扑参数的关键节点识别方法主要有几个特点: 1) 在参数的选择上, 主要凭经验而理论依据不足; 2) 在参数的使用上, 以单个参数值的大小进行识别, 参数在识别过程中是孤立的, 这种识别也称为独立识别; 3) 以寻求高识别度参数为主要目的, 只要参数的识别度不是最高, 都可能被抛弃; 4) 忽视参数之间的关系, 因而未能利用这些关系进一步挖掘节点关键性的信息; 5) 由于蛋白质及其相互关系的复杂性, 单个 (独立) 参数一般只能从某个方面反映节点的部分信息, 找到一个能非常有效地反映这种复杂性的独立参数非常困难, 所以依据独立参数获得的识别度不会很高.

Wasserman 和 Faust 在*Social Network Analysis* 一书中指出: “你不能只用一种指标, 每个指标都有它的优点和用处.” 文献 [8−10] 分别在研究基于生物特征融合的身份验证、融合全局与局部特征进行人脸识别、使用子模式典型相关分析方法进行人脸识别时, 通过融合不同的指标特征而获得更优的识别性能. 受到这些研究的启发, 本文将从以下几个方面对基于拓扑参数的关键节点识别作进一步研

收稿日期 2007-10-19 收修改稿日期 2008-04-24
Received October 19, 2007; in revised form April 24, 2008
国家自然科学基金 (60433020) 资助
Supported by National Natural Science Foundation of China (60433020)
1. 中南大学信息科学与工程学院 长沙 410083 2. 玉林师范学院数学与计算机科学系 玉林 537000 3. 广东商学院计算机科学与技术系 广州 510320
1. School of Information Science and Engineering, Central South University, Changsha 410083 2. Department of Mathematics and Computer Science, Yulin Normal College, Yulin 537000 3. Department of Computer Science and Technology, Guangdong Commercial College, Guangzhou 510320
DOI: 10.3724/SP.J.1004.2008.01388

究: 通过相关分析研究各主要参数与节点关键性的关系, 为参数的选择提供理论依据; 分析参数之间的相关性, 探讨利用多参数信息提高识别度的可能, 阐述复合参数构造的理论和方法; 提出基于复合参数的关键节点异步识别算法, 并通过仿真对算法的识别性能进行验证.

1 相关研究

1.1 节点的中心性测度

节点的重要性等价于该节点与其他节点的连接而使其具有的显著性, 我们称这种显著性为节点的中心性测度 (简称测度, 具体的测度也称为参数). 它主要用来衡量节点在网络中影响力的大小, 评估节点所代表的对象获得、控制信息及资源的能力. 节点度 (Degree, DE) 是最常用的测度, 其他常用的测度还有接近度 (Closeness, CO)、中介度 (Betweenness, BE)、信息度 (Information, IN)、特征向量 (Eigenvector, EI)、子图度 (Subgraph, SU) 和聚集系数 (Clustering coefficient, CU) 等[1, 11]. 这些测度从网络中寻找有效的特征信息来表达节点间的差异, 反映节点某些方面的显著性.

1.2 蛋白质网络的关键节点

文献 [1] 等研究显示, 酵母基因突变引起的致死性 (Lethality) 与因突变而缺失表达的蛋白质的节点度有关: 度越大, 它的缺失引起致死的可能性越大, 它就越有可能是关键的. 生物学上通过对 S.cerevisiae 和 E.Coli 的移除分析已经证实, 关键蛋白通常比其他蛋白具有更多的交互数量[1, 11]. 文献 [12] 指出, 不仅关键蛋白倾向于具有较高的节点度, 而且类似的趋势也反映在毒性调制 (Toxicity-modulating) 蛋白中, 认为显型 (Phenotype) 的非关键蛋白比非显型 (No-phenotype) 的非关键蛋白具有更高节点度, 这个结果与文献 [13] 中的相似, 后者认为关键蛋白比疾病蛋白 (Disease proteins) 更强相关于中枢节点 (Hubs). 文献 [14] 从由 4 743 个酵母蛋白质及其 23 294 种交互所构成的网络中, 按节点度从高到低取 1 061 个节点作为 hubs, 发现其中的 43 % 属于关键节点, 显著高于随机选择 20 % 的期望值, 关键节点的度大约是非关键节点的 2 倍. 同时还发现关键节点倾向于具有更高的聚集度, 相互之间也显现出更密切的关联. 文献 [11] 认为子图参数比度参数更能从结构上提供关键蛋白的重要信息, 按它的值从高到低取 1 % 的节点时, 其中 60 % 是关键蛋白. Yu[15] 的研究指出, 中介度高的节点是关键节点的可能性也比较大. 文献 [16] 以中介度和节点度为中心测度分析蛋白质在网络中的位置, 发现关键蛋白因倾向位于网络的中心而不大可能受到正向选择 (Positive selection), 对此文献 [17] 也指出蛋白质的位置越处于网络的中心, 它的进化越慢并且越有可能是关键的. 文献 [18] 把蛋白质网络和对应的遗传网络综合起来识别关键节点: 从 BioGRID 数据库收集了由 1 869 个基因的 12 850 种交互所组成的 GI 网络, 从多种来源中收集了涵盖 6 184 种蛋白质的 68 172 种交互构成的 PI 网络, 通过 BPM 分析获得 124 种枢纽 (pivot) 蛋白质, 其中的 72 个是关键的, 大大高于预期的 22.6 个. 文献 [19] 从关键交互作用的角度对由 4 126 个节点、7 356 条边构成的酵母蛋白质网络进行了研究, 与 2.92 % 的关键交互作用有关的节点中, 关键节点大约占 43 %.

2 复合参数关键节点识别研究

2.1 基本概念

定义 1. 设 $G = G(V, E)$ 是节点集为 V、边集为 E 的蛋白质网络, $V = \{v_1, v_2, \cdots, v_M\}$, $M = |V|$ 是节点数; $\delta = \{\delta_1, \delta_2, \cdots, \delta_T\}$ 是节点的拓扑参数集, δ_{T+1} 是节点的关键域; σ 是含有 $t\,(t \geq 2)$ 个参数的 δ 的子集.

定义 2. 称矩阵 $A = (a_{ij})_{M\times(T+1)}$ 为 G 的关键节点识别矩阵, 其中 a_{ij} 是节点 v_i 参数 δ_j 的值; $a_{i(T+1)} \in (0, 1)$ 表示 v_i 是否为关键节点, 取值为 1 说明 v_i 是关键节点. 如果 $A' \subseteq A$, 那么 A' 是 A 的行子矩阵.

定义 3. 对 $C \subseteq A$, 令 $\Gamma_S : (C, \delta_j) \rightarrow \acute{C}_j$ 表示根据参数 δ_j 对 C 进行排序后得到 $\acute{C}_j$, 并称之为 1-排序或独立排序, 其中 Γ_S^+ 和 Γ_S^- 分别表示升、降排序, 相应的排序结果分别为 $\acute{C}_j^+$ 和 $\acute{C}_j^-$.

定义 4. 令 $\Gamma_F : (C, m) \rightarrow \hat{C}$ 为筛选函数, 即选择 C 的前 m 个节点组成矩阵 $\hat{C}$.

定义 5. 对节点 v_i 的关键域 δ_{T+1} 赋值 1, 即 $a_{i(T+1)} = 1$, 就是将 v_i 识别为关键节点. 通过单个参数 δ_j 对 C 排序后取前 m 个节点并识别为关键节点, 称为 1-识别或独立识别; 通过参数集 σ 对 C 进行排序、筛选, 在此基础上识别出关键节点, 称为 t-识别或复合 t 参数识别. 如果将筛选下来 $\hat{C}$ 的全部节点识别为关键节点, 而 $\hat{C}$ 实际含有 m_E 个关键节点, 那么 $I = m_E/|\hat{C}|$ 是矩阵 $\hat{C}$ 的关键节点识别度.

我们以参数集 σ 包含两个参数 δ_j、δ_k 的情形 $(t = 2)$ 为例说明 t-识别: 1) 如果 $\sigma' = \varphi(\sigma)$, 即通过函数 φ 分别对每个节点的参数 δ_j、δ_k 进行计算, 使各个节点得到一个新的参数 σ', 然后通过 σ' 用 1-识别的方法进行识别, 这样的 2-识别称为 2-同步识别或复合 2 参数同步识别, 因为两参数是同时参与识别的; 2) 如果先通过其中一个参数 δ_j 进行 1-排

序并筛选得若干节点, 然后用参数 δ_k 对这些节点进行 1-识别, 这样的 2-识别称为 2-异步识别或复合 2 参数异步识别, 因为两参数参与识别有先有后.

2.2 复合参数异步识别算法

以下设 $C = A$, R_{jk} 为 A 中 δ_j 与 δ_k 的相关系数; $1 \le m_j$、$m_k \le M$. $0 \le R_{jk} \le 1$ 时 δ_k 与 δ_j 正相关, 取 $\hat{A}_j^- = \Gamma_F(\Gamma_S^-(A, \delta_j), m_j)$ 及 $\hat{A}_k^- = \Gamma_F(\Gamma_S^-(A, \delta_k), m_k)$; $-1 \le R_{jk} < 0$ 时 δ_k 与 δ_j 负相关, 取 $\hat{A}_j^- = \Gamma_F(\Gamma_S^-(A, \delta_j), m_j)$ 且 $\hat{A}_k^+ = \Gamma_F(\Gamma_S^+(A, \delta_k), m_k)$, 然后对 R_{jk} 取其绝对值参加计算; 这两种情形的其他分析步骤一致, 所以仅考虑 $0 \le R_{jk} \le 1$ 的情形.

引理 1. 设 $\acute{A}_j^- = \Gamma_S^-(A, \delta_j)$, $\hat{A}_k^-$ 如前述. $\acute{A}_j^-$ 中第 i 个节点 v_i 属于 $\hat{A}_k^-$ 的概率随 i 的分布为

$$p_{jki} = \frac{W_{jk} U_i D_i (M - m_k) + m_k}{M} \tag{1}$$

其中, $W_{jk} = \sqrt{R_{jk}(2 - \sqrt{R_{jk}})}$, U_i、D_i 分别为相容系数、偏离系数: $R_{jk} = 1$ 时, 若 $i \le m_k$, 令 $U_i = 1$ 且 $D_i = 1$, 否则 $U_i = m_k/i$、$D_i = i/(m_k - M)$; $0 \le R_{jk} < 1$ 时, $U_i = m_k/(m_k + i^\gamma - 1)$、$D_i = 1 - \alpha(i-1)/(M-1)$, $1 \le \alpha \le M$, $\gamma \ge 1$.

证明. 1) $R_{jk} = 1$ 时, δ_k 与 δ_j 完全正相关, 那么有 $\hat{A}_j^- = \Gamma_F(\Gamma_S^-(A, \delta_j), m_k) = \hat{A}_k^- = \Gamma_F(\Gamma_S^-(A, \delta_k), m_k)$, 即按 δ_j, δ_k 分别进行降序排序后取出的前 $m_k \le M$ 个节点完全一致. $i \le m_k$ 时, $\acute{A}_j^-$ 的节点 v_i (即 $\hat{A}_j^-$ 的节点) 同时也是 $\hat{A}_k^-$ 的节点, 即 $p_{jki} = 1$; $i > m_k$ 时, $\acute{A}_j^-$ 的节点 v_i 不同时属于 $\hat{A}_k^-$, 即 $p_{jki} = 0$; 换句话说, $\acute{A}_j^-$ 中 $1 \sim m_k$ 范围内的节点都同时属于 $\hat{A}_k^-$, 而在这个范围之外 $((m_k+1) \sim M)$ 都不存在属于 $\hat{A}_k^-$ 的节点.

2) $0 \le R_{jk} < 1$ 时, 随着 $\acute{A}_j^-$ 中 i 的增加, p_{jki} 也呈某种程度的下降, 假设 $i' = \lceil (M+1)/\alpha \rceil$ 处有 $p'_{jki} = m_k/M$. 由于 δ_k 与 δ_j 线性相关, 那么 $\acute{A}_j^-$ 的节点 v_i 出现在 $\hat{A}_k^-$ 的概率 p_{jki} 与 R_{jk} 和 m_k 呈正相关, 与 i 和 M 呈负相关. U_i 是 $\hat{A}_k^-$ 对 $\acute{A}_j^-$ 的节点 v_i 的相容系数, $0 \le U_i \le 1$; γ 用来调整 U_i 的下降速率, 它函数负相关于 m_k/M (γ 及前面的参数 α 通过回归分析确定); D_i 表示 i 偏离 i' 的程度; p_{jki} 与 U_i、D_i 呈正比关系.

p_{jki} 偏离 p'_{jki} 的最大幅度为 $1 - p'_{jki} = (M - m_k)/M$, 对 $\acute{A}_j^-$ 的节点 v_i 来说, 它出现在 $\hat{A}_k^-$ 的概率偏离 p'_{jki} 的幅度是在 $(M - m_k)/M$ 基础上由系数 U_i、D_i、W_{jk} 共同决定. 另外, 相关系数为 0 时 δ_k 与 δ_j 无线性相关, $\acute{A}_j^-$ 的节点 v_i 出现在 $\hat{A}_k^-$ 的概率 p_{jki} 与其位置 i 无关且等于 m_k 在 M 上的概率即 $p_{jki} = m_k/M$.

综上所述即得式 (1). □

引理 2. 设 $\hat{A}_j^-$ 及 $\hat{A}_k^-$ 如前述, p_{jk} 是 $\hat{A}_j^-$ 的所有节点出现在 $\hat{A}_k^-$ 的概率, 有

$$p_{jk} = \frac{\sum_{i=1}^{m_j} p_{jki}}{m_j} \tag{2}$$

证明. 由于 $\hat{A}_j^- = \Gamma_F(\Gamma_S^-(A, \delta_j), m_j) = \Gamma_F(\acute{A}_j^-, m_j)$, 即 $\hat{A}_j^-$ 是 $\acute{A}_j^-$ 的前 m_j 个节点. 根据引理 1, $\acute{A}_j^-$ 中第 i 个节点属于 $\hat{A}_k^-$ 的概率随其位置 i 的分布为 p_{jki}, 那么 $\hat{A}_j^-$ 的所有节点出现在 $\hat{A}_k^-$ 的概率是这 m_j 个节点出现在 $\hat{A}_k^-$ 的平均概率, 上式即为平均概率的计算. □

引理 3. 设 $\hat{A}_j^-$、$\hat{A}_k^-$ 如前述, $\hat{A}_{(T+1)}^- = \Gamma_F(\Gamma_S^-(A, \delta_{(T+1)}), N)$, $\tilde{A}_k = A - \hat{A}_k^-$. $p_{j(T+1)}$、p_{jk} 分别是 $\hat{A}_j^-$ 所有节点属于 $\hat{A}_{(T+1)}^-$ 和 $\hat{A}_k^-$ 的概率, $\tilde{p}_{k(T+1)}$ 是 $\tilde{A}_k$ 所有节点属于 $\hat{A}_{(T+1)}^-$ 的概率. 复合参数异步识别过程中, 从 $\hat{A}_j^-$ 中排除属于 $\tilde{A}_k$ 的节点后, 其余节点中关键节点的概率为

$$p_{\sigma\_2} = (2 - p_{jk}) p_{j(T+1)} - (1 - p_{jk}) \tilde{p}_{k(T+1)} \tag{3}$$

并且 $p_{j(T+1)} \le p_{\sigma\_2} \le 1$. 其中,

$$|\tilde{A}_k| = M - m_k$$

$$\frac{p_{j(T+1)}}{\beta} \le \tilde{p}_{k(T+1)} \le p_{j(T+1)} \le \frac{\beta}{\beta + (\beta - 1)(1 - p_{jk})},\ \beta > 1$$

证明. 根据引理 1 和 2, $\hat{A}_j^-$ 节点出现在 $\hat{A}_k^-$ 的概率为 p_{jk}. 由于 $\tilde{A}_k = A - \hat{A}_k^-$, $\tilde{A}_k$ 是 $\acute{A}_k^-$ 的最后 (δ_k 值最小) $M - m_k$ 个节点, 那么 $\hat{A}_j^-$ 节点出现在 $\tilde{A}_k$ 的概率 $\tilde{p}_{jk} = 1 - p_{jk}$, 因此从 $\hat{A}_j^-$ 中排除的、所有同时属于 $\tilde{A}_k$ 的节点的期望数为 $n = m_j \tilde{p}_{jk} = m_j(1 - p_{jk})$. 由于这 n 个节点是在 $\hat{A}_j^-$ 之中选择的, 设其中关键节点的概率为 p, 那么在其他条件相同的情形下, n 越大 (越接近 m_j) 则 p 越接近 $p_{j(T+1)}$, n 越小则 p 越接近 $\tilde{p}_{k(T+1)}$ (因为总是先将其中 δ_k 值最小的节点排除), 而任何同时位于 $\hat{A}_j^-$ 和 $\tilde{A}_k$ 的节点属于关键节点的概率是它在这两个集合中关键节点概率的函数, 即

$$p = \frac{n}{m_j} p_{j(T+1)} + \left(1 - \frac{n}{m_j}\right) \tilde{p}_{k(T+1)} = (1 - p_{jk}) p_{j(T+1)} + p_{jk} \tilde{p}_{k(T+1)}$$

从 $\hat{A}_j^-$ 排除属于 $\tilde{A}_k$ 的节点时也排除了期望值为 $n'=np$ 的关键节点, 因此剩余的 $m'=(m_j-n)=m_jp_{jk}$ 节点中关键节点出现的概率为

$$p_{\sigma\_2}=\frac{m_jp_{j(T+1)}-n'}{m'}=\frac{m_jp_{j(T+1)}-np}{m_jp_{jk}}$$

将 $p=(1-p_{jk})p_{j(T+1)}+p_{jk}\tilde{p}_{k(T+1)}$ 代入并整理得

$$p_{\sigma\_2}=(2-p_{jk})p_{j(T+1)}-(1-p_{jk})\tilde{p}_{k(T+1)}$$

当 $p_{j(T+1)}\geq\tilde{p}_{k(T+1)}$ 时,

$$p_{\sigma\_2}=(2-p_{jk})p_{j(T+1)}-(1-p_{jk})\tilde{p}_{k(T+1)}\geq (2-p_{jk})p_{j(T+1)}-(1-p_{jk})p_{j(T+1)}=p_{j(T+1)}$$

当 $\beta>1$ 且 $\dfrac{p_{j(T+1)}}{\beta}\leq\tilde{p}_{k(T+1)}\leq p_{j(T+1)}\leq\dfrac{\beta}{\beta+(\beta-1)(1-p_{jk})}$ 时,

$$p_{\sigma\_2}=(2-p_{jk})p_{j(T+1)}-(1-p_{jk})\tilde{p}_{k(T+1)}\leq (2-p_{jk})p_{j(T+1)}-\frac{(1-p_{jk})p_{j(T+1)}}{\beta}\leq \frac{\beta}{\beta+(\beta-1)(1-p_{jk})}\left(2-p_{jk}-\frac{1-p_{jk}}{\beta}\right)=1$$

□

推论 1. 根据引理 3: 1) p_{σ_2} 与 p_{jk} 呈负相关, 即两个独立参数之间的相关性越弱, 它们复合参数的异步识别度越高; 2) p_{σ_2} 与 $(p_{j(T+1)}-\tilde{p}_{k(T+1)})$ 呈正相关, 即在 1) 的基础上, 两个参数的独立识别能力越强, 它们复合参数的异步识别度越高.

3　仿真结果及分析

3.1　数据来源

我们主要引用了 4 个数据集, 预处理时将其中蛋白质的自作用去掉: 1) 文献 [11] 使用 Saccharomyces cerevisiae 的 PIN 数据集, 含有 2 224 个节点, 6 608 个相互作用; 2) 文献 [20] 引用的 DIP data 数据集, 包括 4 783 个节点, 14 455 个相互作用; 3) 文献 [14] 使用 Saccharomy ces cerevisiae 的 High-quality interaction network 数据集 (HQI), 包括 4 683 个节点, 22 665 个相互作用; 4) 文献 [20] 引用的 MIPS physical interaction data 数据集, 包括 1 875 个节点, 2 439 个相互作用. 另外综合文献 [11, 14] 所使用的关键节点集进行相关分析, 并对实验结果进行检验.

根据上述参数, 首先从总的角度对这些数据集关键节点的识别进行比较以观察方法的可靠性; 然后重点对 PIN 数据集进行详细分析并归纳复合参数识别算法的要点. 根据定义 1, 在以下分析中令参数集 $\delta=\{\mathrm{DE},\mathrm{CU},\cdots,\mathrm{IN}\}$, 关键域 $\delta_{(T+1)}=\mathrm{ES}$.

3.2　独立参数与复合参数的总体有效性分析

前述的许多研究得出了节点在网络拓扑结构上的显著性反映了蛋白质在功能上的关键性的结论, 而文献 [19] 则对此持怀疑态度, 但这两方面意见主要基于经验而理论依据不足, 为此我们首先对节点的拓扑参数与关键性进行相关分析. Pearson 相关系数计算方法为

$$R_{xy}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}\tag{4}$$

结果见表 1.

表 1　独立参数与节点关键性的相关性

Table 1　Correlation between parameter and essentiality

R	DE	CU	BE	EI	SU	CO	IN
ES_(PIN)	0.220	0.211	0.113	0.203	0.118	0.173	0.112
ES_(DIP)	0.205	0.203	0.092	0.222	0.016	0.052	0.027
ES_(MIPS)	0.206	0.172	0.115	0.023	−0.014	0.017	0.013
ES_(HQI)	0.241	0.233	0.086	0.013	0.016	0.011	0.022

表 1 反映出 DE、CU 与节点关键性 ES 的相关系数 R 明显地比其他参数大, 2-Sig. 显著水平 (表中未列出) 都非常高, 而且在各个数据集中均比较稳定. EI 在前两数据集中的 R 也比在后两数据集大. 仿真实验证实相关系数 R 越大的参数的关键节点识别度也越高, 部分结果见图 1.

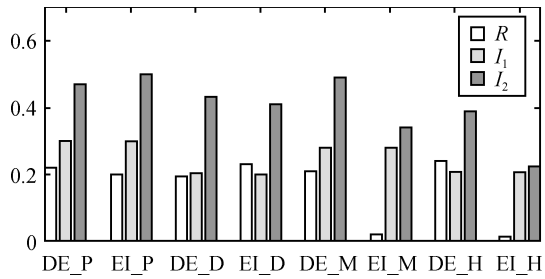

图 1　DE、EI 的独立识别性能与 R 的关系

Fig. 1　Relation of identifying abilities of DE, EI, and R

图 1 只示意了 DE、EI 两参数各自与 ES 的相关系数 R 及独立识别度 I 之间的关系, 图中后缀 _P、_D、_M、_H 分别代表上述 4 个数据集. 其中 I_1 是各个数据集中关键节点比率, 同一个数据集的 I_1 不变, 而 I_2 是这样获取的: 通过单个参数分别求数据集 1 %～50 % 范围内若干采样区间的识别度, 并以它们的均值作为 I_2 值. DE 在各个数据集中与 ES 的相关系数都比较大, 在图中的识别度 I_2 都明显高于 I_1. EI 的 R 值在前两个数据集较高而在后两者

较低, 相应地它在前两者的 I_2 也明显高于 I_1, 而在后两者的差别不大. 这些现象说明 R 越大的参数其识别度也越高, 支持了节点在网络拓扑结构上的显著性反映其功能关键性的观点.

考虑到 DE、CU 在 4 个数据集中与 ES 的相关程度比较高且稳定、可靠, BE 也相对稳定和可靠, 因此对它们重点观察: 1) DE、CU 之间的 R 值在不同数据集中有明显差异; 2) DE、BE 之间的 R 值大且比较稳定; 3) CU、BE 之间的 R 值 (R 小于 0 时以其绝对值进行比较) 都比较小. 这些相关系数 R 及有关的仿真结果分别在表 2、图 2 列出, 然后结合推论 1 对复合参数进行分析. 其中各个独立参数、复合参数的识别度 I 分别由独立识别法、复合 2 参数异步识别法 (参见定义 5) 获取, 它们是 1 %～50 % 范围内数据集若干采样区间的识别度的均值.

表 2 DE, CU, BE 等独立参数之间的相关分析

Table 2 Correlation between several single parameters

	DE ∼ CU	DE ∼ BE	BE ∼ CU
R_P	0.006	0.881	−0.104
R_D	0.059	0.850	−0.034
R_M	0.307	0.705	−0.043
R_H	0.549	0.486	−0.010

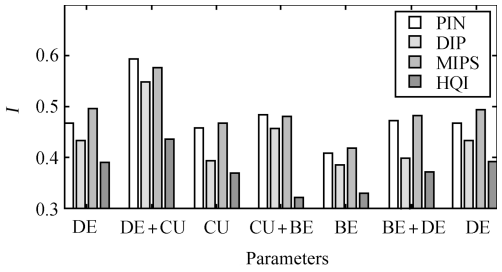

图 2 DE, CU, BE 及其复合参数总体识别性能

Fig. 2 Total identifying abilities of DE, CU, BE and their combined parameters

图 2 中按顺序将参数分成: a) (DE, DE + CU, CU), b) (CU, CU + BE, BE), c) (BE, BE + DE, DE) 等三组, 分别比较两独立参数与由它们构成的复合参数的识别性能: a) 组复合参数 DE + CU 的识别度在 4 个数据集中都明显高于它的两个独立参数 DE、CU; b) 组复合参数的识别度与它的两个独立参数无明显差异; c) 组的结果介于上述两组之间. 结合表 2 我们对上述现象进一步观察: 1) DE、CU 的相关系数在 PIN、DIP 数据中小 ($R < 0.06$) 而显著, 在 HQI 中较大 ($R = 0.549$) 而显著, 因此复合参数 DE + CU 在 PIN、DIP 比在 HQI 更能提高识别度; 2) DE 与 BE 的相关系数都比较大 ($R > 0.48$) 而显著, 因此复合参数 DE + BE 很难提高识别度; 3) 虽然 BE 与 CU 的相关系数比较小, 但 BE 与 ES 的相关系数也较小, 它的独立识别能力也比较低, 所以 CU + BE 的识别度没有显著地提高. 这些现象与推论 1 的论述是基本一致的.

3.3 基于相关分析的复合参数构造

根据上述分析, 独立参数的识别性能正比于它与关键节点的相关性, 由于这些相关程度比较有限, 因此独立参数的识别度都不高, 但是存在着通过多参数复合来提高识别度的可能, 这种可能性的大小又取决于独立参数之间的相关性, 为此根据推论 1, 以数据集 PIN 为例进一步探讨复合参数的构造及识别算法. 如无特别说明, 后面涉及的具体情形均基于数据集 PIN 和文献 [14] 引用的关键节点集. 根据式 (4) 得到该数据集各独立参数之间的相关系数, 表 3 列出部分有代表性的结果.

表 3 PIN 数据集的 DE, CU 与其他独立参数的相关性

Table 3 Correlation between DE (CU) and other parameters in PIN dataset

R	DE	CU	BE	EI	SU	CO	IN
DE	1.000	0.006	0.881	0.654	0.493	0.678	0.194
2-Sig.	–	0.760	0.000	0.000	0.000	0.000	0.000
CU	0.006	1.000	−0.104	0.050	0.026	0.005	0.134
2-Sig.	0.760	–	0.000	0.018	0.220	0.806	0.000

表 3 中 DE 和 CU 有如下主要特点: 1) 它们之间的相关系数很小, 显著水平很低, 可以认为它们之间不存在相关, 但两者与 ES 均有较显著的相关 (见表 1); 2) DE 与除 CU 之外的其他参数呈较明显的相关; 3) CU 与其他参数的相关性有明显差异. 图 3 以直观的形式表示了 DE 与 CU、BE 的关系: DE 上升, BE 也明显上升, CU 的升降则不明显, 在表 3 相应地发现 DE 与 BE 相关系数大 ($R = 0.881$), 与 CU 相关系数小 ($R = 0.006$).

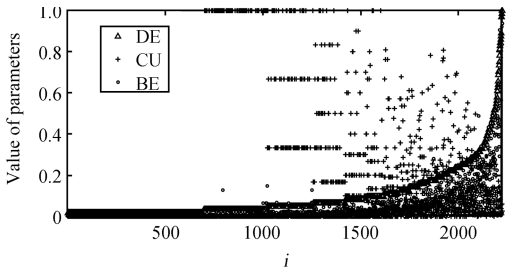

图 3 DE, CU, BE 等参数之间的相关性 (节点在横轴上按 DE 升序排列)

Fig. 3 Correlation between DE, CU and BE. Nodes sort ascending in X-coordinate according to DE

根据上述观察结合推论 1 初步推断: DE 和 CU

复合的识别性能明显好于它们各自的性能, 而且也好于 DE 与其他参数的复合; CU 的复合参数易于取得较好的识别性能. 下一节将在理论与实际识别度分析的基础上对此进一步加以验证, 同时对复合参数的异步、同步 (以下将同步特指为: 取独立参数的均值作为复合参数值) 识别算法进行比较和分析.

3.4　复合参数的异步识别性能及其比较

对表 3 构造如下复合 2 参数: 以 CU 或 DE 作为第一参数 (即引理 3 的 δ_j), 其他参数作为第二参数 (即 δ_k); 或者反过来. 通过第一参数 δ_j 取 $m_j = N = 670$, 通过第二参数 δ_k 在 $(10, 670)$ 范围内对 $m_k = 10, 25, 50, 100, 150, \cdots, 600, 650, 670$ 等进行识别并得出实际识别度. 在同样的条件上, 根据引理 $1 \sim 3$ 对不同的 R 值进行仿真计算得出理论识别度. 根据这些结果, 从以下几方面进一步研究:

1) 理论识别度与实际识别度的比较

图 4 主要以 DE + CU 为例分析复合参数的理论及实际识别度, 同时也与构成它的独立参数 DE 进行比较. 图中复合参数 DE + CU 理论值、实际值两条曲线的走向比较一致, 在整个采样空间上也没有出现太大的差异, 而独立参数 DE 的两条曲线的走向也基本一致, 从图 5 和图 6 也观察到类似的结果.

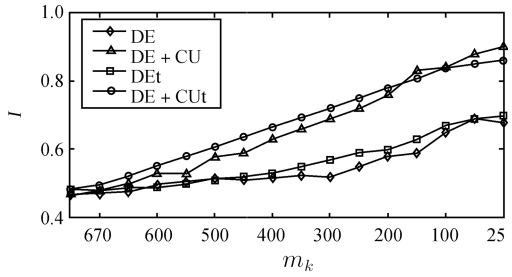

图 4　DE 及复合参数 DE + CU 识别度的理论值与实际值比较 (DEt, DE + CUt 为理论值)

Fig. 4　Comparing of theoretical and real identifying rates between DE and DE + CU (combined parameter), with DEt and DE + CUt to be theoretical

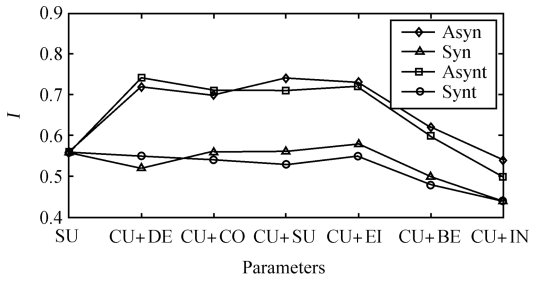

图 5　复合参数识别 (异步) 的有效性分析

Fig. 5　Analysis of identifying ability of combined parameter

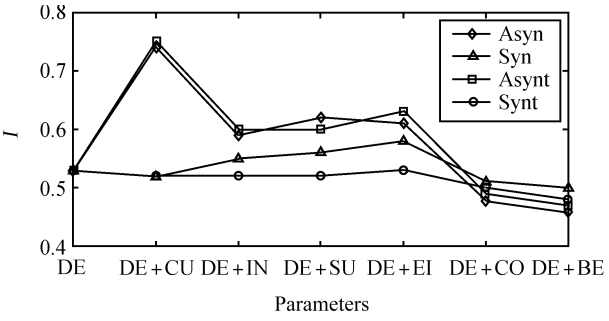

图 6　DE 及其复合参数异步/同步识别的比较

Fig. 6　Synchro / asynchronous identifying abilities of DE and its combined parameters

2) 复合参数与独立参数识别性能比较

根据 3.3 节的分析, 由 CU 参与构成的复合参数在总体上容易取得较好的识别度, 因此图 5 将它们与参数 SU (为了扩大比较范围, 且它在数据集 1) 局部小范围的识别度也比较好) 进行比较. 考虑到复合参数识别算法在 $m_k \in (400, 670)$ 区间内的识别主要受第一参数的影响 (见图 7), 复合参数的作用没有充分体现, 为了便于观察, 主要对 $m_k \in (10, 400)$ 的 10 个采样区间上的平均识别度进行分析. 图中 Asyn、Syn 分别是 CU 参与构成的复合参数异步、同步实际识别结果, Asynt、Synt 分别是对应的仿真计算结果. 从图 5 观察到: 1) 参数 SU 的平均识别度约为 56 %; 2) Asyn (Asynt) 曲线表明, 除了 CU + IN 外, 其他复合参数的异步识别能力明显高于 SU, 其中 CU + DE、CU + CO、CU + SU 及 CU + EI 的平均识别度都在 70 % 以上; 3) Syn (Synt) 曲线表明, 复合参数的同步识别能力对 SU 没有优势, 平均同步识别度大多数与 SU 基本上一样, CU + BE、CU + IN 的同步识别能力甚至低于 SU. 类似图 5 的几种情形还可以从图 6 观察到, 而图 7 主要是详细地比较独立参数 CU、SU 与复合参数 CU+SU 的异步识别结果: 从左至右, 随着第二参数作用的增强, 按两种次序构成的复合参数的性能均显著高于独立参数.

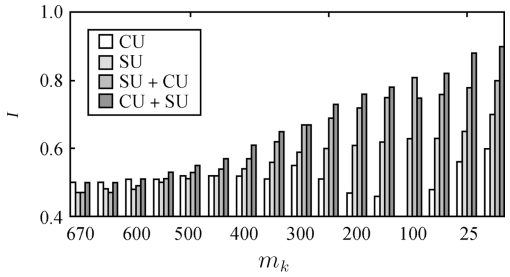

图 7　SU 与 CU 及其复合参数识别 (异步) 性能的比较

Fig. 7　Comparing the identifying abilities of SU, CU, and their combinations

根据上述观察, 在本文相关条件下有这样的关系: 复合参数的异步识别算法有助于提高识别度, 而同步识别算法很难提高识别度. 下面对这些现象进行更深入的观察和分析.

3) 复合参数的异步与同步识别分析

结合图 5 识别度的变化和表 3 发现: CU 与 DE、CO、SU、EI 的相关系数很小且不显著, 它们复合参数的异步识别度都比较高; CU 与 BE、IN 的相关系数相对较大且显著, 它们复合参数的异步识别度都比较低; 这些现象与推论 1 的推断一致: 独立参数的相关性越弱, 它们复合参数的异步识别能力越强.

为了进一步证实这个问题, 图 6 给出了 DE 及其复合参数在 $m_k \in (10, 400)$ 区间的平均识别度: DE 与 CU 几乎不存在相关性 ($R = 0.006$, 2-Sig. $= 0.76 > 0.05$), 它们复合的异步识别效果最好; DE 与其他参数相关性上升, 它们复合的异步识别能力降低; 相关性上升到一定程度后, 随着关键节点误删除比例的增加, 识别度甚至低于 DE 的独立识别能力, 异步识别失效. 上述结果与推论 1 及 3.3 节的设想是一致的, 即复合参数异步识别能力的大小除了与构成它的独立参数的识别能力有关之外, 还与独立参数之间相关性有关: 如果完全不相关, 互补性最强, 识别效果最好; 相关性上升, 互补性下降, 识别能力降低; 相关性上升到一定程度后, 在随机因素的干扰下识别度的稳定性下降且可靠性差, 甚至出现低于独立参数识别能力的现象; 完全相关时独立参数的复合没有意义. 在实际中这些相关性的变化是互动的, 某一方面的变化都将波及到其他方面.

异步与同步识别算法的不同在于, 前者通过第一个独立参数把识别度基本确定在一个比较高的水平, 第二个独立参数主要是从中将尽可能多的非关键节点去掉, 同时尽量减少关键节点的误删除. 后者在取双参数的平均值之后凸现出来的信息非常有限, 还可能因此而模糊了关键、非关键节点之间的差异, 所以很难有所提高.

3.5 关于复合参数异步识别算法的一些讨论

复合参数异步识别算法建立在独立参数及其相互关系的基础上, 理论研究及仿真表明这种算法是有效的, 这种有效性的生物学意义值得进一步探讨. 由于篇幅有限, 在此主要以 DE + CU 为例进行分析.

虽然节点度 DE 越高的蛋白质越有可能是关键的, 但是由于节点度只从一个侧面反映关键性, DE 高的蛋白质中总有相当一部分是非关键的, 单纯凭 DE 提供的信息获得的识别度不会很高; 聚集系数 CU 也存在同样的问题. 一般情形下 CU 依赖于 DE 且成反比关系[7], 即低度节点容易产生较高的聚集系数, 属于联系紧密的小模块, 但因与其他节点的关联程度低, 影响范围有限, 重要性可能不大. 另外, 由于小模块的数量比较多, 存在类似结构和功能模块的几率较大, 可替代性强, 它的缺失引起致死的可能性就低. 高度节点的聚集系数一般较低, 如果低到一定程度, 节点不大可能成为大的功能模块的核心, 地位就不象它的度所显现的那么重要, 所以成为关键的可能性也不会很高. 复合参数的异步识别算法正是考虑到这些关系的复杂性, 在相关分析的基础上力图找出参数之间相对合理的平衡, 尽可能多地发现参数之间的互补性并加以整合利用. 上述对几个数据集的仿真结果表明这种方法能够在一定程度上理出 DE 和 CU 之间的这种关系并筛选出反映这种关系的节点 (HH 节点). 较多的 HH 节点抵消掉了其他节点上的负相关, 使得 DE、CU 在整个数据集上几乎不显示出相关性. HH 节点的存在还意味着围绕着它可能形成规模较大而紧密的模块, 该蛋白质因处于中心位置而可能对模块结构和功能的稳定性起重要作用. 另外, 模块的规模越大, 复杂性也越高, 网络中结构和功能类似的模块数量必然少, 可替代性比较差. 一但中心蛋白质缺失, 模块发生重大改变的可能性较高, 并因可替代性差而容易致死.

4 结论

由于蛋白质结构、功能及相互关系的复杂性, 已有研究表明单个参数只能从某些方面提供有限的信息, 要找到具有很高识别能力的参数就目前来看是不现实的, 如何对现有单个参数的有限信息进行有效地整合以提高识别度是非常现实的问题. 上述工作正是基于多参数模式识别进行的一种尝试, 主要贡献在于: 1) 在识别上, 不是以对象的固有属性, 而是以对象之间的相互关系作为识别的依据; 2) 在算法上, 为模式识别提出一种新的、基于多特征 (复合参数) 的异步识别算法; 3) 在具体应用上, 为关键蛋白质的识别提供新的途径, 所提出的复合参数的构造方法及其异步识别算法不但有充分的理论依据, 在实践上也表现出明显高于原有方法的识别性能. 由于本文提供的是一种灵活开放的平台, 基于它可以集成更多有效信息来进一步优化识别性能, 即使以后发现新的效果更好的特征或参数, 通过引入其他相关信息与之整合仍有可能在它之上获得更高的识别效果, 因此在模式识别上具有广泛意义.

References

1 Jeong H, Mason S P, Barabási A L, Oltvai Z N. Lethality and centrality in protein networks. *Nature*, 2001, **411**(6833): 41−42

2 Lu Hong-Chao. Research on Genic Function by Clustering on Protein Network [Ph. D. dissertation], Institute of Computing Technology, Chinese Academy of Sciences, 2006
(卢宏超. 基于蛋白网络聚类的基因功能研究 [博士学位论文]. 中国科学院计算技术研究所, 2006)

3 Winzeler E A, Shoemaker D D, Astromoff A, Liang H, Anderson K, Andre B. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, 1999, **285**(5429): 901−906

4 SiegalM L, Promislow D E, Bergman A. Functional and evolutionary inference in gene networks: does topology matter? *Genetica*, 2007, **129**(1): 83−103

5 Parrish J R, Yu J, Liu G, Hines J A, Chan J E, Mangiola B A. A proteome-wide protein interaction map for campylobacter jejuni. *Genome Biology*, 2007, **8**(7): 1−19

6 Luo F, Yang Y F, Chen C F, Chang R, Zhou J Z, Scheuermann R H. Modular organization of protein interaction networks. *Bioinformatics*, 2007, **23**(2): 207−214

7 Yook S H, Oltvai Z N, Barabási A L. Functional and topological characterization of protein interaction networks. *Proteomics*, 2004, **4**(4): 928−942

8 Liu Hong-Yi, Wang Yun-Hong, Tan Tie-Niu. Multimodal data fusion for person authentication based on improved ENN. *Acta Automatica Sinica*, 2004, **30**(1): 78−85
(刘红毅, 王蕴红, 谭铁牛. 基于改进 ENN 算法的多生物特征融合的身份验证. 自动化学报, 2004, **30**(1): 78−85)

9 Wang Yun-Hong, Fan Wei, Tan Tie-Niu. Face recognition based on information fusion. *Chinese Journal of Computers*, 2005, **28**(10): 1657−1663
(王蕴红, 范伟, 谭铁牛. 融合全局与局部特征的子空间人脸识别算法. 计算机学报, 2005, **28**(10): 1657−1663)

10 Hong Quan, Chen Song-Can, Ni Xue-Lei. Sub-pattern canonical correlation analysis with application in face recognition. *Acta Automatica Sinica*, 2008, **34**(1): 21−30
(洪泉, 陈松灿, 倪雪蕾. 子模式典型相关分析及其在人脸识别中的应用. 自动化学报, 2008, **34**(1): 21−30)

11 Estrada E. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 2005, **6**(1): 35−40

12 Said M R, Begley T J, Oppenheim A V, Lauffenburger D A, Samson L D. Global network analysis of phenotypic effects: protein networks and toxicity modulation in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**(52): 18006−18011

13 Goh K I, Cusick M E, Valle D, Childs B, Vidal M, Barabási A L. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**(21): 8685−8690

14 Yu H, Greenbaum D, Xin L H, Zhu X, Gerstein M B. Genomic analysis of essentiality within protein networks. *Trends in Genetics*, 2004, **20**(6): 227−231

15 Yu H, Kim P M, Sprecher E, Trifonov V, Gerstein M B. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 2007, **3**(4): 713−720

16 Kim P M, Korbel J O, Gerstein M B. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular contex. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**(51): 20274−20279

17 Hahn M W, Kern A D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 2005, **22**(4): 803−806

18 Ulitsky I, Shamir R. Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. *Molecular Systems Biology*, 2007, **3**: 104

19 He X L, Zhang J Z. Why do hubs tend to be essential in protein networks? *PLoS Genetics*, 2006, **2**(6): 826−834

20 Deng M H, Zhang K, Mehta S, Chen T, Sun F Z. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 2003, **10**(6): 947−960

黄海滨 博士研究生, 副教授. 主要研究方向为模式识别, 人工智能, 生物计算. 本文通信作者.
E-mail: ylhhpin@163.com
(**HUANG Hai-Bin** Ph. D. candidate, associate professor. His research interest covers pattern recognition, AI, and biological computation. Corresponding author of this paper.)

杨路明 教授. 主要研究方向为人工智能, 模式识别, 计算机网络优化及安全.
E-mail: yang@mail.csu.edu.cn
(**YANG Lu-Ming** Professor. His research interest covers AI, pattern recognition, and computer network optimization and security.)

王建新 博士, 教授. 主要研究方向为计算机网络优化, 算法理论, 生物信息学.
E-mail: jxwang@mail.csu.edu.cn
(**WANG Jian-Xin** Ph. D., professor. His research interest covers computer network optimization, algorithm theory, and bioinformatics.)

李绍华 博士研究生, 副教授. 主要研究方向为计算机网络, 生物计算.
E-mail: sohually@163.com
(**LI Shao-Hua** Ph. D. candidate, associate professor. His research interest covers computer networks and biological computation.)