文章编号: 1673-1689(2010)05-0782-06

16S rDNA 序列关键位点的信息论分析

周大为[1,2], 于雪飞[2], 周少刚[2], 李炜疆*[1,2]
(1. 江南大学 工业生物技术教育部重点实验室, 江苏 无锡 214122; 2. 江南大学 生物工程学院, 江苏 无锡 214122)

摘 要: 16S rDNA 中含有较多的菌种特异性信息, 但是如何使用这些信息进行菌种鉴定却因不同菌种间序列的特征变化缺少规律而变得困难。作者使用信息论方法对那些种内较保守, 种间变异明显的关键位点进行研究。结果表明, 16S rDNA 序列中仅有少数位点对菌种的分类是重要的。基于少数关键位点的相似度比传统的全序列鉴定有着更好的统计学特性。

关键词: 16S rDNA; 分类; 信息论; 熵
中图分类号: Q 93.331 文献标识码: A

Information-Theory Analysis for Key Sites of 16S rDNA Sequence

ZHOU Da-wei[1,2], YU Xue-fei[2], ZHOU Shao-gang[2], LI Wei-jiang*[1,2]
(1. Key Laboratory of Industrial Biotechnology, Ministry of Education, Jiangan University, Wuxi 214122, China; 2 School of Biotechnology, Jiangnan University, Wuxi 214122, China)

Abstract: 16S rDNAs carry significant species-specific genetic information, but species identification is not straight forward because the sequence identity varies greatly for different species. In this manuscript, we analyze the key sites that are well conserved within species but significantly varied between species using an information theoretic approach. The method proposed by us was just using the sites (key sites) which contained more genetic information than that of the normal sites of the sequences, to classify the bacterium rather than using the overall sequence. Only a few sites of the whole sequences are responsible for the classification of species. The results presented here demonstrated that key-sites based similarity measures have much better statistical characteristics than that of the traditional overall sequence identity scores

Key words: 16S rDNA, taxonomy; information theory; entropy

随着生物技术的飞速发展, 传统的微生物鉴定方法常常难以鉴定众多的生长习性复杂的微生物, 因而基于基因组序列的分子鉴定受到广泛关注。在细菌基因组中, 编码 16S rRNA 的 rDNA 基因具有良好的进化保守性, 适宜分析的长度(约为 1.5 kb)[1-4], 以及与进化距离相匹配的良好变异性[5], 所以成为细菌分子鉴定的标准标识序列。目前 16S rDNA 的序列信息已经广泛应用于菌种鉴定和系统发生学研究[4, 6-7]

16S rDNA 分子测序是目前使用的主流菌种鉴

收稿日期: 2009-09-08

* 通信作者: 李炜疆(1964-), 男, 陕西榆林人, 理学博士, 教授, 博导, 主要从事生物信息学与计算分子生物学研究。Email: wjlee01@gmail.com

定方法[8], 测得待测菌种的 rDNA 序列, 将序列送至基因库中进行同源性比对, 同源性相似性最高, 且至少达到 97% 以上可认为是同一种。由于 16S rDNA 反映的是进化上的距离, 对于同一属内的微生物鉴定获得的同源相似度都十分高, 此时对于种的鉴定及分类是极为不利的, 同一种间的差异往往只有 2%, 即 30 个碱基左右[9], 此时大量相同的、保守的序列信息被计入了比较过程, 从而掩盖了菌种间的序列信息的不同。

Shannon[10] 在 1948 年提出信息熵的概念用以衡量不确定性, 同时也指出信息量[11] 的多少要依靠信息消除未知, 也就是得到信息后信息熵的减少来衡量。对于 16S rDNA, 使用全序列比对的效率较低, 这时对序列信息的深入研究就显得十分重要, 使用统计学方法找出加入分类信息后, 信息熵变化最为剧烈的、信息量最大的位点, 对这些位点研究, 找出这些位点的变化规律[12], 用以指导菌种分类鉴定工作。

1　实验材料

分类信息来自 NCBI 的菌种分类目录, 所使用的对齐后的16S rDNA 序列取自 Greengenes[13], 不同菌种在进化过程中 16S rDNA 也发生了碱基和长度的改变, 以及在实际的测序工作中具体方法不同, 得到的序列起点、长度也不完全相同, 为方便对不同序列进行对比分析, Greengenes 使用 NAST 算法[14] 将 1.5 kb 左右的 16S rDNA 序列依照保守区定位、对齐扩展为 7 682 个字符长度的格式, 用在序列的头尾添加数量不同的". "来将长度和起点不统一的序列凑齐, 在序列中用字符"-"表示对齐时产生的插入缺失, 并且将这样处理后的序列称为 Greengenes 扩展序列。(Greengenes 计划由 Lawrence Berkeley National Laboratory 等机构开展, 截至 2009 年 2 月, 共收集了 298 922 条 16S rDNA 序列[15]。)

在 Greengenes 开放下载的数据库中, 各菌种的序列数量分布极不均匀, 若种属包含的样本过少, 则无法得到有统计学意义的研究结果。作者从数据库中以属内至少有 9 个菌种, 每个菌种至少有 8 条以上序列为挑选条件扫描该数据库, 得到两个符合条件的属: *Salmonella* 属中选取了符合要求的 9 个亚种, 每个亚种均有 9 条以上序列, 见表 1。*Streptomyces* 属中含有 8 条以上序列的共有 49 个种, 属内序列总计 950 条, 见表 2。

表 1　沙门氏菌属数据集中的菌种

Tab. 1　Species included in the Salmonella dataset

菌种名	样本量	菌种名	样本量	菌种名	样本量
typhi	9	*heidelberg*	17	*agona*	20
virchow	14	*kentucky*	14	*dublin*	14
saintpaul	12	*newport*	14	*hadar*	14

表 2　链霉菌属数据集中的菌种

Tab. 2　Species included in the Streptomyces dataset

菌种名	样本量	菌种名	样本量	菌种名	样本量
acidiscabies	14	*olivochromogenes*	22	*humidus*	10
anulatus	28	*phaeochromogenes*	42	*hygroscopicus*	21
atratus	12	*phaeopurpureus*	11	*lavendulocolor*	11
aurantiacus	12	*platensis*	21	*lividans*	16

2　实验方法

2.1　种间差异度

当种内序列变化较小, 混乱度较低; 同时种间混乱度较高, 序列变化剧烈时, 分类工作易于进行。

信息熵[16] 为度量系统混乱程度的概念, 信息熵越大则表示描述的系统越混乱。给定一组 Greengenes 扩展序列, 为了定量描述特定位点上碱基变异的程度, 定义位点信息熵, 见式(1)。

$$H(x) = -\sum_{\alpha} p_{\alpha}(x)\ln p_{\alpha}(x), \qquad (1)$$

式(1) 中, x 为该位点在扩展对齐的16S rDNA 序列中的位置; $\alpha \in \{A, C, G, T, -\}$ 为该位点上可能出现的所有碱基以及因对齐而产生的插入缺失(indel, 用"-"表示); $p_{\alpha}(x)$ 为在位点 x 处碱基出现的频率; 而 $H(x)$ 则表示在 x 处的信息熵, 当各字母出现频率相同时, 信息熵最大, 此时 x 处最为混乱, 相反, 如果只出现一种字母, 此时信息熵最小, x 是数据集中的保守位点。$H(x)$ 不同则反映了该菌种的 16S rDNA 共同特性出现的概率不同, 每一位点对应的混乱程度不同, 即信息熵不同。对于给定属内的序列样本, 不同菌种的信息熵(即种内信息熵)也不相同, 种内信息熵可以用表达式(2) 表示:

$$H^{(s)}(x) = -\sum_{\alpha} p_{\alpha}^{(s)}(x)\ln p_{\alpha}^{(s)}(x), \qquad (2)$$

式(2) k, $p_{\alpha}^{(s)}(x)$ 为 S 菌种在位点 x 处碱基 α 出现的频率; $H^{(s)}(x)$ 则表示在菌种 S 范围内 x 处的种内信息熵。种内信息熵描述了在特定菌种内的位点变异情况。对于给定菌种内的16S rDNA 序列, 如果某个位点上的碱基变化很小, 信息熵较小, 则此位点对

于该菌种是保守的[17]，反映了该菌种的共同特性，因而可能对菌种鉴定有贡献；反之，如果某个位点上的碱基变化很大，则该位点对菌种鉴定的贡献就很小，见式(3)

$$H_M(x) = \sum_s \frac{N^{(s)}}{N} H^{(s)}(x) \qquad (3)$$

式(3)中，$N^{(s)}$ 为 S 菌种的序列条数；N 代表在研究的数据集中所有属内的序列数量；$\frac{N^{(s)}}{N}$ 为 S 菌种占数据集中总样本的比例；$H_M(x)$ 为种内信息熵的加权求和，也就是加入分类信息条件后的条件信息熵。一般来说，由于加入了更多的信息，混乱程度得以减小，属内的条件信息熵比属内的信息熵小。可以证明，属内信息熵 $H(x)$ 减去条件信息熵 $H_M(x)$ 为该分类信息所有的信息量，即互信息 $I(x)$，见式(4)。

$$I(x) = H(x) - H_M(x), \qquad (4)$$

可以看出，只有当属内信息熵较大同时条件信息熵较小时互信息才会较大，也就是说互信息较大的位点是在属内较为不保守、同时在种内较为保守的，即作者将挑选位点进行分类的要求，定义互信息为种间差异度，表示位点对分类的贡献。

2.2 统计涨落处理

由于数据集中大多数物种的序列样本数目都不大，所以必须估量统计涨落对种间差异度的影响。为此，作者将数据集中各序列的物种分类属性用洗牌的方式随机打乱，由此模拟位点变异与分类无关时由于统计涨落导致的种间差异度，记作 $I_{\text{shuffle}}(x)$。重复模拟 100 次，计算平均值 $< I_{\text{shuffle}}(x) >$ 标准差 $\sigma_{\text{shuffle}}(x)$，定义 Z-value，见式(5)。

$$Z(x) = \frac{I(x) - < I_{\text{shuffle}}(x) >}{\sigma_{\text{shuffle}}(x)} \qquad (5)$$

$Z(x)$ 体现了实际观测到的种间差异度与随机涨落之间差异的显著性，当 $Z(x) >> 1$ 时，位点 x 处碱基的出现频率与物种即存在着显著关联，亦即该位点可用于物种分类。显著性是位点用于分类的必要条件，而种间差异则是充分条件，只有当一个位点有良好的显著性时，该位点才有用于分类的可能，而能不能使用还要看它的种间差异度，位点对分类的贡献的大小。

2.3 数值实验

统计涨落是统计平均值在其值附近有微小变动的现象，增大统计样本的数量可以部分抵消该现象。本实验所采用的数据集由于样本量较小，易于出现统计涨落。

模拟实验可以估计由于统计涨落而产生的种间差异度(模拟所得到的 $< I_{shuffle}(x) >$ 约为 0.14)，而使用正确分类信息得到的种间差异度，见图 1。则体现了在 16S rDNA 上不同位点对分类的贡献程度。

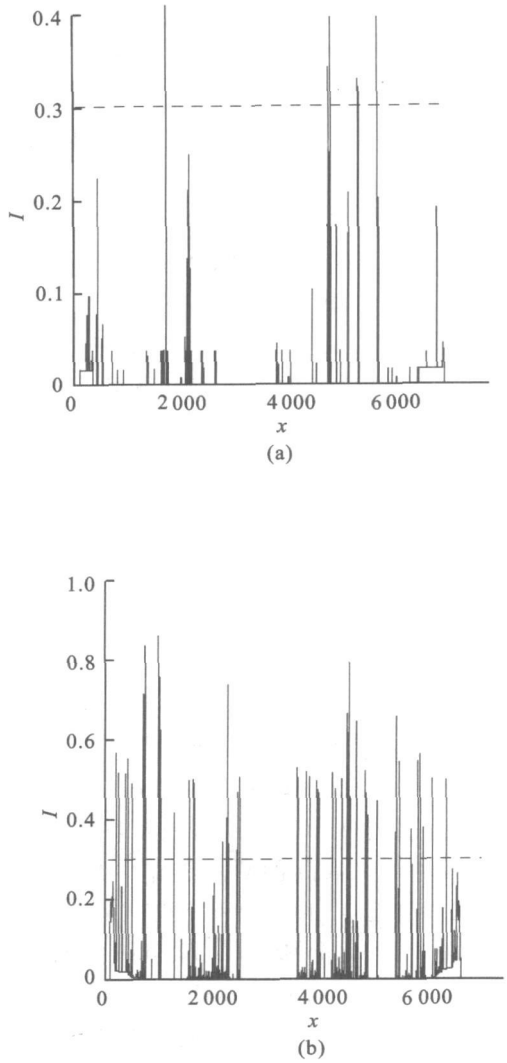


图 1 种间差异度(I)随着所研究的沙门氏和链霉菌属中对齐后的 16S rDNA 序列而变化

Fig. 1 **Interspecies discrepancy (I) varied along the aligned 16S rDNA sequences of the two genera investigated in this study, *Salmonella* (a) and *Streptomyces* (b)**

种间差异度与显著性的关系

只有当选定位点的种间差异度和 Z-value 都足够大时，该位点才有可能应用于分类，对这两者进行研究，找出其中规律，可以指导下一步的位点选择工作。

以种间差异度为纵坐标，Z-value 为横坐标研究种间差异度和显著性之间的关系，发现两者间呈明显正相关，见图 2。当 Z-value 增大时，种间差异度也有一定程度的增加。

$I(x) > 0.3$ 时, 相应位点的 Z-value 值全部大于 10, 此时这些位点既有统计学上的显著性, 同时也对分类有足够多的贡献, 称为关键位点。*Salmonella* 属共得到 8 个关键位点, 而在 *Streptomyces* 属得到 92 个关键位点。

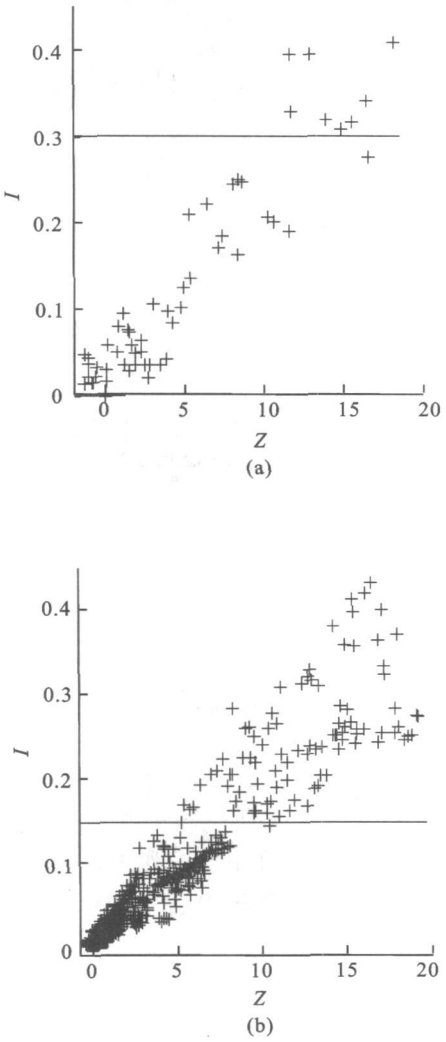


图 2　从沙门氏(A)和链霉菌(B)属数据集中计算所得种间差异度和 Z 值之间的关联

Fig. 2　Dependence between the interspecies discrepancy and the Z values calculated from the two data-sets, *Salmonella* (a) and *Streptomyces* (b)

式(6) 表示打分方法, 当被比较的两条序列的位点出现'.'或者两位点的字母相同时记零分, 而当两位点字母不同时记一分。

$$\theta(a, b) = \begin{cases} 0, & \text{if } a \text{ or } b = \text{'.'} \text{ or } a = b; \\ 1, & \text{otherwise} \end{cases} \qquad (6)$$

在得到关键位点后, 对于给定的两条序列 i 和 j, 定义关键位点分数, 见式 7。

$$D_{ij}^{\text{key}} = \sum_{x \in \text{key sites}} \theta(\alpha_i(x), \alpha_j(x))。 \qquad (7)$$

通常的序列比较所得的分数相当于不区分关键位点(或将所有位点都视为关键位点) 的序列分数, 即整体分数, 见式(8)。

$$D_{ij}^{\text{overall}} = \sum_{x \in \text{all sites}} \theta(\alpha_i(x), \alpha_j(x)), \qquad (8)$$

其中, $\alpha_i(x)$ 为序列 i 在位点 x 处的字母; D_{ij}^{key} 为第 i 和第 j 条序列使用关键位点打分后得到的分数; D_{ij}^{overall} 为使用全序列位点打分得到的分数。

定义序列相似度, 见式(9)。

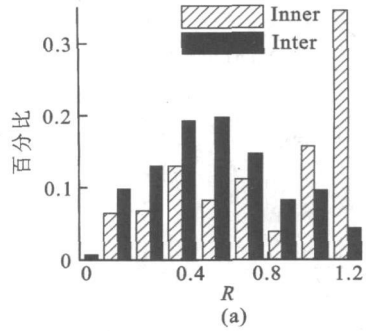
$$R = 1 - \frac{D}{D_{\max}} \qquad (9)$$

其中, R 为相似度; D 为使用特定方法得到的分数; $D_{\max}$ 为使用该种打分方法理论上可以达到的最大分数。

由于不同序列的原始长度不一致, 而在比较的过程中还会忽略掉两条序列都有的插入缺失, 为了在同一体系中比较不同长度序列的打分情况, 使用相似度可以忽略序列长度的差异[18], 打分情况以百分数来描述。

2.3.1　对 *Salmonella* 属的数据集打分　使用关键位点, 同时也使用全序列 *Salmonella* 属打分, 从图 3 可以看出, 使用少数关键位点得到的结果比利用全部信息的结果要好, 这主要是因为使用全部信息时不仅使用的对分类有帮助的位点的信息, 同时也使用了大量保守的位点的信息, 这就使得有价值的信息被稀释从而不能表现在结果里, 而仅仅使用关键位点则避免了这一问题。

2.3.2　对 *Streptomyces* 属的数据集打分　使用同样的方法对 *Streptomyces* 属的样本研究见图 4, 发现仅使用关键位点(92 个) 得到的分类效果比使用全序列好得多。从图 4 可以看出, 使用种间差异度较大的位点可以很好地对大部分同属内不同种间的菌株分类, 仅对小部分菌株发生误判, 而全序列比对的分类效果十分不理想, 对大量的菌株误判[19]。



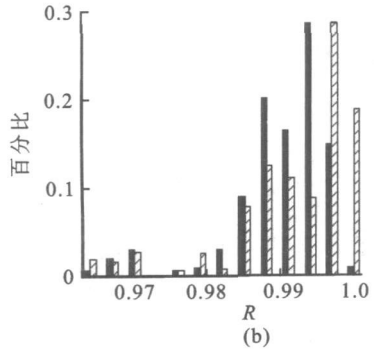


图 3 沙门氏菌属数据集中所有序列按照两种打分策略所得到的分布:以关键位点(a)和全序列(b)表示的相似度

Fig. 3 Distributions of all sequences in the Salmonella dataset according to two scoring schema: similarity upon key sites (a) and the overall sequence (b)

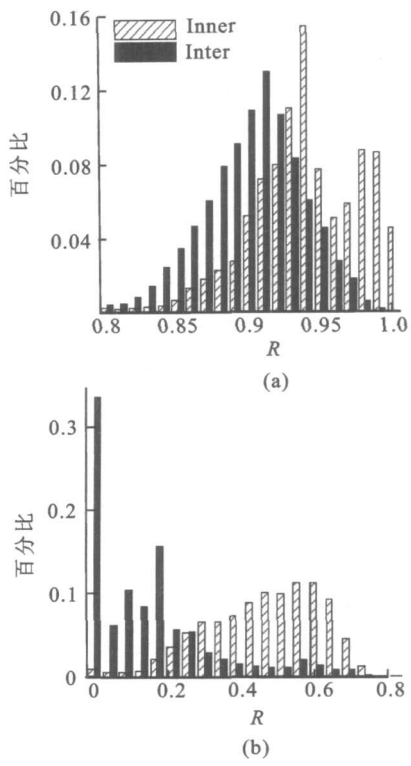


图 4 在 16S rDNA 序列间使用两种不同打分方法比较所得到的相似性分数分布:(a)全序列和(b)关键位点

Fig. 4 Distributions of the similarity scores between 16S rDNA sequence pairs with the scores defined by (a) overall sequence and (b) the key sites

3 结 语

因为不同位点在分类中的种间差异度不同,同时,比对全序列信息的分类效果不显著,作者认为应该使用种间差异度较大的位点开展菌种的分类鉴定工作。全序列包含有序列中所有的信息,也包含关键位点中的信息,但其分类效果不如只使用关键位点的效果明显,这是因为在采用全序列比对时使用了太多的冗余信息,其中的噪音掩盖了对分类有贡献的位点的信息。而使用种间差异度比较大的位点分类则部分避免了该问题,分类效果优于全序列比对。本方法在计算中出现的一个问题是,随着序列条数的增加,种间差异度比较大,同时也比较显著的位点也在增加,这就为分类带来了新的难题,必须针对不同的样本分别训练才能得到较好的分类效果,同时,如果样本中序列条数较少,则不能得到统计学上有意义的结果,导致本方法在这种情况下不能被应用。如何应用该方法对未知新物种的发现和分类,将是我们下一步工作的重点。

参考文献(References):

[1] Gutell R R, Larsen N, Woese C R. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective[J]. **Microbiol Rev,** 1994, 58(1): 10-26.

[2] Fox G W, Woese C R. 5S RNA secondary structure[J]. **Nature,** 1975, 256(5517): 505-507.

[3] Noller H F, Woese C R. Secondary structure of 16S ribosomal RNA[J]. **Science,** 1981, 212(4493): 403-411.

[4] Woese C R, Fox G E, Zablen L, et al. Conservation of primary structure in 16S ribosomal RNA[J]. **Nature,** 1975, 254(5495): 83-86.

[5] Pace N R, Olsen G J, Woese C R. Ribosomal RNA phylogeny and the primary lines of evolutionary descent[J]. **Cell,** 1986, 45(3): 325-326.

[6] Fox G E, Stackebrandt E, Hespell R B, et al. The phylogeny of prokaryotes[J]. **Science,** 1980, 209(4455): 457-463.

[7] Yunhong Kong S L O, Wun Jun Ng, Wen-Tso Liu. Diversity and distribution of a deeply branched novel proteobacterial group found in anaerobic-aerobic activated sludge processes[J]. **Environmental Microbiology,** 2002, 4(11): 753-757.

[8] Wagner J, Short K, Catto-Smith A G, et al. Identification and characterisation of pseudomonas 16S ribosomal DNA from ileal biopsies of children with crohn's disease[J]. **PLoS ONE,** 2008, 3(10): 3578.

[9] Woo P C Y, Lau S K P, Teng J L L, et al. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories[J]. **Clinical Microbiology & Infection,** 2008, (14): 908–934.

[10] Shannon C E. A mathematical theory of communication[J]. **Bell System Technical Journal,** 1948, (27): 379–423.

[11] Swanson R, Vannucci M, Tsai J W. Information theory provides a comprehensive framework for the evaluation of protein structure predictions[J]. **Proteins,** 2009, 74(3): 701–711.

[12] Van de Peer Y, Chapelle S, De Wachter R. A quantitative map of nucleotide substitution rates in bacterial rRNA[J]. **Nucleic Acids Res,** 1996, 24(17): 3381–3391.

[13] DeSantis T Z, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB[J]. **Appl Environ Microbiol,** 2006, 72(7): 5069–5072.

[14] DeSantis T Z, J Hugenholtz P, Keller K, et al. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes[J]. **Nucleic Acids Res,** 2006, 34: 394–399.

[15] Cole J R, Wang Q, Cardenas E, et al. The ribosomal database project: improved alignments and new tools for rRNA analysis[J]. **Nucl Acids Res,** 2009, 37: 141–145.

[16] Chen H D, Chang C H, Hsieh L C, et al. Divergence and Shannon information in genomes[J]. **Phys Rev Lett,** 2005, 94(17): 102–103.

[17] Fernandes F, Freitas A T, Almeida J S, et al. Entropic profiler-detection of conservation in genomes using information theory[J]. **BMC Res Notes,** 2009, 2(1): 72–74.

[18] 梅娟,何胜,王正祥,等. 基于网络模块性的蛋白质序列聚类[J], 食品与生物技术学报, 2010, 29(1): 123–127.
MEI Juan, HE Sheng, WANG Zheng-xiang, et al. Clustering protein sequences through modularity optimization[J]. **Journal of Food Science and Biotechnology,** 2010, 29(1): 123–127. (in Chinese)

[19] Mei J, S, He. et al. Revealing network communities through modularity maximization by a contraction-dilation method [J]. **New Journal of Physics,** 2009, 11: 221–226.

(责任编辑: 李春丽)