文章编号: 1673-1689(2010)01-0123-05

基于网络模块性的蛋白质序列聚类

梅娟[1,2], 何胜[1,2], 王正祥[1,2], 石贵阳[1,2], 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: 蛋白质的远同源性探测是结构基因组学和功能基因组学的主要研究任务之一。一些具有一定相似结构和功能、但序列相似性却较低的蛋白质组成蛋白质超家族, 则远同源性探测问题等价于对蛋白质超家族的识别问题。作者提出了一种基于模块性的聚类算法 ModuleFind, 该方法通过最大化蛋白质网络的模块性来寻找具有较强集团结构的划分。在蛋白质结构分类数据库(SCOP)超家族层次上进行的实验表明, 该方法得到的聚类结果更接近分类基准, 且具有较高的 F-测度值。

关键词: 蛋白质网络; 序列相似性; 远同源性; 模块性; 聚类; 蛋白质结构分类数据库
中图分类号: Q 7 文献标识码: A

Clustering Protein Sequences through Modularity Optimization

MEI Juan[1,2], HE Sheng[1,2], WANG Zheng-xiang[1,2], SHI Gui-yang[1,2], LI Wei-jiang*[1,2]
(1. Key Laboratory of Industrial Biotechnology, Ministry of Education, Jiangnan University, Wuxi 214122, China; 2. School of Biotechnology, Jiangnan University, Wuxi 214122, China)

Abstract: Remote homology detection between protein sequences is one of the principal research objectives in structural and functional genomics. Proteins with similar structure and function but low sequence similarity consist of protein superfamily. Therefore, the detection of remote homologues is the task of identifying protein superfamily. In this manuscript, a clustering algorithm, called ModuleFind, based on network modularity was presented. The method maximizes the modularity of protein network to find the partitioning with strong community structure. The resulting algorithm gives high quality of clusters quantified by F-measure that combines precise and recall, in the experiments of the detection of the remote homologues based on the superfamily level of SCOP database.

Key words: protein network, sequence similarity, remote homology, modularity, clustering, SCOP

从蛋白质序列推断蛋白质结构和功能是生物信息学的一个重要研究课题[1]。一类在结构和功

收稿日期: 2009-03-01
基金项目: 国家 863 计划项目(2006AA020204)。
作者简介: 梅娟(1980-), 女, 江苏盐城人, 发酵工程博士研究生。
* 通讯作者: 李炜疆(1964-), 男, 陕西榆林人, 理学博士, 教授, 博士生导师, 主要从事生物信息学、计算分子生物学方面的研究。Email: wjlee01@gmail.com

能上相似的蛋白质序列集合可定义为蛋白质家族,同一家族内的蛋白质具有同源关系,且有显著的序列相似性。通过寻找同源蛋白,可以识别蛋白质家族,从而推断结构与功能特征。

当蛋白质同源性较高时,借助序列比对算法,如 BLAST[2], FASTA[3], Needleman-Wunsch 算法[4]或 Smith-Waterman 算法[5],可以准确方便地发现蛋白质之间进化的关系。然而,许多序列相关性不显著的蛋白质之间也具有结构或功能相似性,因而可能具有更为遥远的共同祖先,即远同源性。这类蛋白质序列的集合可定义为蛋白质超家族。与同源性探测相比,远同源性探测则复杂得多,因为远同源蛋白间的序列相似性很低,处于随机涨落区域边缘(twilight zone),很难区分通过比对获得的序列特征是进化过程中功能约束还是随机突变导致的结果。换言之,远同源蛋白质的序列相似性指标是几乎淹没在噪音中的微弱信号。提取这些模糊信号的一个有效方法是利用一组蛋白质间的序列相似性,通过聚类处理获得可靠结果。这种方法称为图聚类算法[6-7]。

图聚类算法解决蛋白质远同源性探测问题主要分两步来实现,首先是蛋白质网络(节点是蛋白质,边的权重表示它所连蛋白质的序列相似程度)的构建,其中关键的是蛋白质间序列相似性的度量,通常采用的指标的有:BLAST E-value 的某些函数变换值[7-8],序列比对算法得到的相似性分数[10]等;然后采用聚类算法将构建的网络划分为"自然的"集团,使得每个集团中的蛋白质都尽可能地具有同源关系。文献中绝大部分的算法都是通过设定阈值对蛋白质间的序列相似性进行处理。这些方法大体上分为两类,第一类方法处理的对象是完全联通图,将权重小于设定的阈值的边舍弃,仍然连在一起的蛋白质就属于同一个集团。这类方法称为 CCA(Connected Component Analysis),GeneRAGE[11]是基于这种思想的一个方法。显然,阈值的选取直接影响聚类结果,保守的阈值只能够将相似性很大的蛋白质聚成一类,因而不利于寻找远同源蛋白质;放松的阈值易将本不属于同一类的蛋白质聚在一起。

第二类方法是基于单联聚类(Single Linkage Clustering)的思想,根据蛋白质间的序列相似性将蛋白质排列成树形结构,选择适当的阈值将整个树划分成许多子树,即对应着聚成的集团。这类方法便于我们研究蛋白质间的层次结构。SYSTERS[12], ProtoMap[13]和 ProClust[14]等都属于此类方法。

上述的方法都是局部算法,在确定蛋白质所属的类别号时,只考虑它与库中蛋白质间的序列相似性,忽略了该蛋白质以外的其他蛋白质间的序列相似性对结果的影响。作者提出了一种基于网络模块性(modularity)[15]的全局算法(ModuleFind),通过启发式策略搜索具有较强集团结构的蛋白质网络的划分,这样的划分对应着最后的聚类结果。与 TribeMCL[9]和 Spectral clustering[8]这两个全局算法相比,ModuleFind 对蛋白质超家族分类取得了较优的结果。

1 数据与方法

1.1 数据

文中的实验数据来源于 SCOP[16]数据库 1.73 版。该数据库依据蛋白质的 3D 结构将其归类,以树状的方式主要分为 4 个层次:结构类(class),折叠(fold),超家族(superfamily)及家族(family)。一般认为,同一超家族中的蛋白质存在共同的进化起源,但是由于它们的一级序列差别很大,仅从序列本身来看它们的同源关系不是非常明显。因此,SCOP 数据库中超家族的分类能够作为测试算法在聚类序列相似性很低,但实际上是相关的蛋白质时性能表现的基准。文献[8]中作者手工挑选了较难聚类的 6 个超家族。这里,我们类似地选取了 778 条序列,这些序列属于 6 个超家族:Globin-like, EF-hand, Cupredoxins, (Trans) glycosidases, Thioredoxin-like 和 Glucocorticoid receptor-like. 这些超家族选自 SCOP 数据库的子库 Astral-95,它是非冗余的,且库中任两个蛋白质的一致性(identity)不超过 95%。

表 1 列出了构建的库中各个蛋白质超家族包含的蛋白质序列和蛋白质家族的数目。每个超家族中平均有 12 个家族。

表 1 各个超家族包含的序列和家族的数目

Tab. 1 The numbers of sequences and families in each superfamiliy

超家族	序列数	家族数
Globin-like	110	4
EF-hand	123	11
Cupredoxins	95	7
(*Trans*) glycosidases	157	14
Thioredoxin-like	176	22
Glucocorticoid receptor-like	117	15

1.2 蛋白质相似性网络的构建

蛋白质相似性网络用无向带权图来表示, 图的节点为蛋白质, 连接两个节点的边的权重则表示相应蛋白质间序列相似的程度。本研究的蛋白质相似性网络采用了如下的构建过程: 首先用 NCBI 网站提供的 blastall 程序计算数据集中每两条序列间的 BLAST E-value 值, 作为蛋白质间距离的度量; 然后, 将这些值用 sigmoidal 函数做如下变换:

$$S_{ij} = \frac{1}{1 + \exp[20(\lg E_{ij} - 0.8)]} \quad (1)$$

其中, E_{ij} 表示序列 i 和 j 的 BLAST E-value 值。通常, 矩阵 E 不是对称的, 因为将序列 a 与 b 比对, 和序列 b 与 a 比对得到的 E-value 值一般是不等的。因此, 我们取 $A_{ij} = A_{ji} = \min(S_{ij}, S_{ji})$ 作为序列 i 和 j 的相似性, 这样得到的对称矩阵 A 称为相似性矩阵, 用作序列关联图的邻接矩阵。

1.3 算法

图聚类的目标是试图将构建的网络划分为"自然的"集团, 使得集团内部节点相似度较高, 而不同集团的节点之间相似度较低。为了量化这种集团结构, Newman 和 Girvan[15] 提出了模块性(modularity)这个评价指标。在这个框架下, 较大的模块性值对应着网络一个较好的划分, 于是把蛋白质序列的聚类问题转化为寻找蛋白质相似性网络最大模块性的优化问题。

1.3.1 模块性的定义　对于一个给定了划分, 由 n 个节点和 m 条边组成的网络, 其模块性的定义为:

$$Q = \sum_{i=1}^{k}\left(\frac{e_i}{m} - \left(\frac{a_i}{2m}\right)^2\right) \quad (2)$$

其中, k 是集团的个数, e_i 是集团 i 内部所有边的数目; a_i 是集团 i 中所有节点的度之和。研究中的蛋白质网络是带权网络, 相应的 e_i 表示集团 i 内部所有边的权重之和, a_i 表示所有与集团 i 中节点相连的边的度之和。

1.3.2 最大化模块性的启发式策略(ModuleFind)

最初, 将所有节点随机分配到任意多的集团中, 为了改进这个初始划分, 依次不断地将单个节点移动到使 Q 值增长最快的集团中, 直至 Q 值不能再增大为止。在这个过程中, 节点逐渐归并到部分集团中, 出现了不断增加的空集团, 导致非空集团的个数不断减少, 我们称之为收缩过程; 经历收缩过程后, 所有的节点处于较合适的集团中, 但这个状态只是相对于单节点移动的一个局部最优解。为了寻求更好的集团结构, 我们使所有节点以扰乱概率 p 离开原来所在的集团, 即由 p 来控制对收缩过程输出的集团结构的破坏程度。这个过程中非空集团的个数迅速变多, 因此我们称之为膨胀过程。整个算法流程就是将收缩和膨胀过程迭代多次求最优值的过程, 其形式化定义如下:

ModuleFind Algorithm:

(初始化)

1. 随机分配所有节点到 c_{max} 个集团中, 分别记录当前的集团结构和 Q 值为 c_{best} 和 Q_{best}
2. **repeat**

(收缩过程)

3. **repeat**
4. **for each** node i
5. 移动 i 到集团 $\beta = \arg\max_{\alpha} \Delta_{i\alpha}$
6. **endfor**
7. **until** Q 不能再增大
8. If $Q > Q_{best}$, 更新 c_{best} 和 Q_{best} (膨胀过程)
9. **for each** node i
10. 以扰动概率 p 来决定 i 是否移动到其他集团中
11. **endfor**
12. **until** NITER times

在上述算法中, NITER 表示整个过程中收缩和膨胀迭代的次数; $\arg\max_{\alpha} \Delta_{i\alpha}$ 表示使 $\Delta_{i\alpha}$ 最大的集团 α。

2 结果与讨论

为了验证和比较 ModuleFind 算法对蛋白质超家族分类的有效性, 作者与 Spectral clustering[8] 和 Tribe-MCL[9] 这两个比较著名的全局算法进行了比较。通常评价聚类结果常用的指标有 F-测度值(F-measure)、准确率等。作者以 F-测度值来评价聚类的效果。假设数据集中有 n 条序列, 由 SCOP 数据库提供的超家族的分类为 K, 算法的聚类结果为 L, n_g 和 n^h 分别表示表示 L 中第 L_g 类和 K 中第 K_h 类包含的蛋白质的数目, n_h^g 表示既在 L_g 中又在 K_h 中的蛋白质的数目, 则有:

精确率(Precise):

$$P(L_g, K_h) = \frac{n_g^h}{n_g} \quad (3)$$

召回率(Recall):

$$R(L_g, K_h) = \frac{n_g^h}{n^h} \quad (4)$$

F-测度值是精确率和召回率的组合, 这里取它们的等权重组合。于是, 对于聚类结果 L, 总的 F-测度值为:

$$F(L,K)=\frac{1}{n}\sum_{h}n^{h}\max_{g}\frac{2a_g^h}{n_g+n^h} \quad (5)$$

表 2 列出了 3 种算法对 1.1 节构建的蛋白质超家族数据集进行聚类得到结果的 *F*- 测度值。

表 2 3 种算法聚类结果的 *F*- 测度值

Tab. 2 Values of the *F*-measure given by three algorithms

算法	*F*- 测度值
Tribe-MCL	0.5193
Spectral clustering	0.7082
ModuleFind	0.7802

图 1~ 3 分别是 Tribe-MCL, Spectral clustering 和 ModuleFind 这 3 种算法的聚类结果。图中每一行表示一个集团, 较长的竖线表示超家族的边界, 较短的竖线表示一个蛋白质, 每个蛋白质在图中必有一个惟一的位置。假设某个蛋白质位于第 *a* 行第 *b* 列, 则表示属于第 *b* 个超家族的蛋白质处于算法输出的聚类结果中的第 *a* 个集团。图中从左到右的 6 个超家族依次是 Globin-like, EF-hand, Cupredoxins, (Trans)glycosidases, Thioredoxin-like 和 Glucocorticoid receptor-like。每个超家族中同一家族的蛋白质处于紧邻的位置。

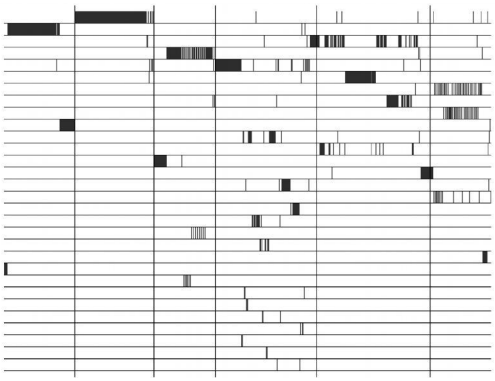

图 1 Tribe-MCL 的聚类结果

Fig. 1 Clustering result of Tribe-MCL

TribeMCL, Spectral clustering 和 ModuleFind 的聚类结果中分别包含 48, 10 和 11 个集团。为了使图比较清晰, 图 1 中只显示了前 30 个含蛋白质数目大于 1 的集团。从图中可以看出, Tribe-MCL 得到很多没有统计显著性的集团(只含有几个蛋白质), 但是它对 EF-hand 这个超家族聚类效果很好。Spectral clustering 对 Thioredoxin-like 和 Glucocorticoid receptor-like 这两个含家族数目较多的超家族聚类结果比较混乱。ModuleFind 将 Thioredoxin-like, Glucocorticoid receptor-like 和 (Trans)glycosidases 这 3 个超家族中的一些家族聚成独立的集团, 同时在 Globin-like, EF-hand 和 Cupredoxins 这 3 个超家族上的聚类效果近乎完美。

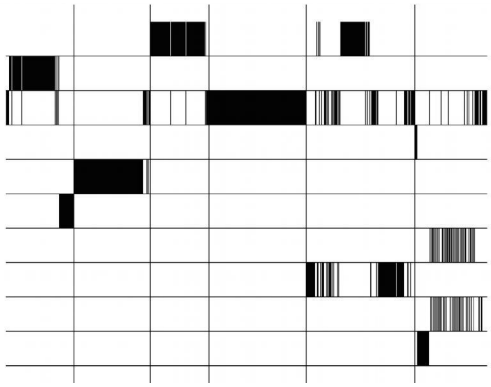

图 2 Spectral clustering 的聚类结果

Fig. 2 Clustering result of Spectral clustering

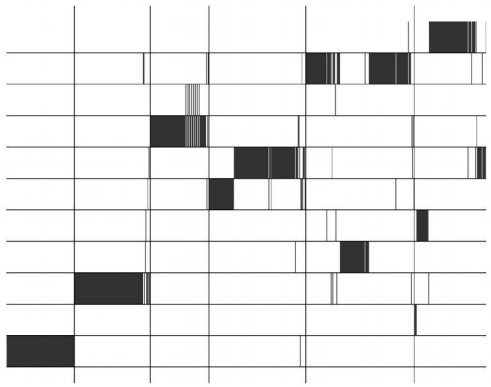

图 3 ModuleFind 的聚类结果

Fig. 3 Clustering result of ModuleFind

3 结 语

作者提出了一种基于网络模块性的蛋白质序列聚类算法, 并且构造了节点为蛋白质、边的权重为 BLAST E-value 值的蛋白质网络, 通过寻找具有最大模块性的蛋白质网络的划分来识别蛋白质超家族。在蛋白质结构分类数据库(SCOP)中的序列数据集上的实验结果表明, 该算法有较高的准确率。

参考文献(References):

[1] Nikolski M, Sherman D J. Family relationships: should consensus reign——consensus clustering for protein families[J]. **Bioinformatics,** 2007, 23(2): 71- 76.

[2] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool[J]. **J Mol Biol,** 1990, 215(3): 403- 410.

[3] Pearson W R, Lipman D J. Improved tools for biological sequence comparison[J]. **Proc Natl Acad Sci USA,** 1988, 85(8): 2444- 2448.

[4] Durbin R, Eddy S R, Krogh A, et al. Biological Sequence Analysis[M]. Cambridge: Cambridge University Press, 1998.

[5] Smith T F, Waterman M S. Identification of common molecular subsequences[J]. **J Mol Biol,** 1981, 147(1): 195- 197.

[6] Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology[J]. **Brief Bioinform,** 2006, 7(3): 243- 255.

[7] 梅娟, 王正祥, 石贵阳, 等. 复杂生物网络分析的图聚类方法研究进展[J]. 食品与生物技术学报, 2008, 27(5): 15- 20.
MEI Juan, HE Sheng, WANG Zheng-xiang, et al. Clustering protein sequences through modularity optimization[J]. **Journal of Food Science and Biotechnology,** 2008, 27(5): 15- 20. (in Chinese)

[8] Paccanaro A, Casbon J A, Saqi M A. Spectral clustering of protein sequences[J]. **Nucleic Acids Res,** 2006, 34(5): 1571- 1580.

[9] Enright A J, Van Dongen S, Ouzounis C A. An efficient algorithm for large-scale detection of protein families[J]. **Nucleic Acids Res,** 2002, 30(7): 1575- 1584.

[10] Liao L, Noble W S. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships[J]. **J Comput Biol,** 2003, 10(6): 857- 868.

[11] Enright A J, Ouzounis C A. GeneRAGE: a robust algorithm for sequence clustering and domain detection[J]. **Bioinformatics,** 2000, 16(5): 451- 457.

[12] Krause A, Stoye J, Vingron M. The SYSTERS protein sequence cluster set[J]. **Nucleic Acids Res,** 2000, 28(1): 270- 272.

[13] Yona G, Linial N, Linial M. ProtoMap: automatic classification of protein sequences and hierarchy of protein families[J]. **Nucleic Acids Res,** 2000, 28: 49- 55.

[14] Pipenbacher P, Schliep A, Schneckener S, et al. ProClust: improved clustering of protein sequences with an extended graph-based approach[J]. **Bioinformatics,** 2002, 18(2): 182- 191.

[15] Newman M E, Girvan M. Finding and evaluating community structure in networks[J]. **Phys Rev E,** 2004, 69: 026- 113.

[16] Murzin A G, Brenner, S E, Hubbard T, et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures[J]. **J Mol Biol,** 1995, 247(4): 536-540.

(责任编辑: 李春丽)