

基于蚁群聚类的蛋白质二级结构特征研究

高冶,陈绮

(海南大学信息科学技术学院,海南 海口 570228)

摘要:通过氨基酸序列来预测蛋白质功能与空间结构一直是生物信息学研究的重点之一。蛋白质二级结构是在一定的氨基酸残基的组成和排列顺序(即蛋白质一级结构)的基础上形成的,不同的氨基酸残基由于具有不同的理化特性,从而形成不同的蛋白质二级结构。文中以蛋白质数据库(PDB)为数据源建立了二级结构数据库,并选取疏水值、等电点等特征,利用蚁群聚类对二级结构进行聚类,其结果所表现出的特征符合既有规律,并为后期的预测工作提供了依据。

关键词:蛋白质二级结构;蚁群聚类;特征对比

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2013)06-0191-04

doi:10.3969/j.issn.1673-629X.2013.06.049

Research on Features of Protein Secondary Structure Based on Ant Colony Clustering

GAO Ye, CHEN Qi

(College of Information Science and Technology, Hainan University, Haikou 570228, China)

Abstract: It has been one focus of the bioinformatics study that predicts protein function and spatial structure through amino acid sequence. Protein secondary structure is formed on the basis of a certain amino acid residues in the composition and order (protein primary structure), has the formation of different protein secondary structure due to different physical and chemical properties. In this paper, established a protein secondary database based on the PDB database, then conducted a cluster of secondary structure by the hydrophobic value, isoelectric point and other features. With ant colony clustering cluster for secondary structure, the results demonstrate the characteristics are in line with the existing regular patterns and provide a basis for forecasting in the later.

Key words: protein secondary structure; ant colony clustering; feature contrast

1 背景介绍

蛋白质的结构预测对研究蛋白质空间结构与功能的关系,以及在此基础进行的蛋白质突变体设计、蛋白质药物设计等具有十分重大的意义。蛋白质的结构预测是指直接从氨基酸序列推断这一蛋白质的功能位点或者预测其三维结构,包括二级结构预测和三维结构预测,是目前生物信息学迫切需要解决的重要问题^[1]。其中,蛋白质的二级结构预测不但是联系蛋白质一级结构和三级结构的纽带,而且是由一级结构预测蛋白质三维结构的关键步骤。蛋白质二级结构预测是生物信息学中的分类问题也可以说是数学中的多维空间非线性映射问题。针对蛋白质二级结构预测的方法主要有神经网络、支持向量机、决策树等方法^[2]。现有的研究已经表明,根据上述的方法配合蛋白质部分残

基序列的特定属性,可以取得良好的预测效果。但是这些算法计算量大,只能使用较小的测试数据集(几百到几千个特定种类的蛋白质),实验结果适用范围有限,特征概括也较为繁琐。此外实现这些方法预测的前提是要有良好的多序列对差异性度量函数,这方面的研究到现在为止虽然涌现出很多不同的方法(概述不同的对比方法),但是依然没有一个良好的通用度量规范^[3,4]。上述研究方法大都注重对蛋白质序列结构信息局部特征进行提取,而对蛋白质序列结构信息整体特征提取的研究工作少之又少,因此工作的重点在对于生物数据库中所有蛋白质二级结构的三类特征进行宏观上的差异提取、发现特征。

在对蛋白质的结构频谱研究中,发现蛋白质序列与蛋白质结构存在一些潜在的关联。蛋白质二级结构

收稿日期:2012-09-22

修回日期:2012-12-15

网络出版时间:2013-03-05

基金项目:海南省自然科学基金资助项目(609003)

作者简介:高冶(1989-),男,硕士研究生,研究方向为数据挖掘;陈绮,教授,博士,硕士生导师,研究方向为数据挖掘。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130305.0819.049.html>

是在一定的氨基酸残基的组成和排列顺序(即蛋白质一级结构)的基础上形成的,不同的氨基酸残基由于具有不同的理化特性,从而形成不同的蛋白质二级结构,进而表现出不同的结构特征^[5]。因此,利用以蛋白质数据库 PDB 为数据源建立了二级结构数据库,并选取疏水值、等电点等特征,利用蚁群聚类对二级结构进行聚类,希望能探索蛋白质序列与结构间的潜在的关联。

2 前期工作

蛋白质的二级结构主要有以下四类: α -螺旋, β -折叠和无规卷曲。为了保证数据的准确性,数据源选用 PDB(Protein Data Bank)中的蛋白质分子文件作为抽取对象。截止 2010 年 11 月,该数据库共计收录了 68421 个蛋白质结构数据,这些原始数据是无冗余蛋白序列数据,它们全部为人工测量分析所得,具有最好的准确度。表 1 为目前所建立的数据库中收录情况,其中 Helix 记录是用来标识蛋白质分子中螺旋结构的部分,即 α -螺旋;Sheet 记录项标识了折叠片在蛋白质分子中 β -折叠;而 Coil 字段则记录了无规则卷曲部分。由于 PDB 提供的数据文件中,没有给出无规则卷曲部分的位置标识,故在数据中除去已提取出的 α -螺旋和 β -折叠,剩余部分即视为无规则卷曲部分。由于相当一部分无规则卷曲长度很小,大概只有 5 个或者更短的残基序列,如此短的序列片段不足以进行规律特征的提取,在这里将其视为噪声数据,通过数据清洗对其进行过滤。

表 1 PDB 数据库收录情况

Name	Number
Helix	1398430
Sheet	274202
Coil	389112

3 生成特征序列

蛋白质二级结构是在一定的氨基酸残基的组成和排列顺序(即蛋白质一级结构)的基础上形成的,不同的氨基酸残基由于具有不同的理化特性,从而形成不同的蛋白质二级结构,进而表现出不同的结构特征。选用了疏水值、等电点以及解离常数。解离常数(pK)是水溶液中具有一定解离度的溶质的极性参数。表 2 中 pK1 指 $\text{pK}-\text{COOH}$, pK2 指 $\text{pK}-\text{NH}_3^+$ 。氨基酸在一定 pH 条件下,某种氨基酸接受或给出质子的程度相等,分子所带的净电荷为零,此时溶液的 pH 值便是该氨基酸的等电点(pI)。这些性质都是由于其在水溶液中的电荷分布不均匀而产生的,这些性质都可以作为

蛋白质序列定量分析的重要数据源。表 2 为这些性质对照表。

表 2 氨基酸物化性质对照表

amino acids	pI	hydro	pK1	pK2
Glycine	5.97	-3.2	2.34	9.60
Alanine	6.02	1.8	2.34	9.60
Valine	5.97	4.2	2.32	9.62
Leucine	5.98	3.8	2.36	9.60
Lsoleucine	6.02	4.5	2.36	9.68
Serine	5.68	-0.8	2.21	9.15
Threonine	6.53	-0.7	2.63	10.43
Aspartic acid	2.97	-3.5	2.09	9.82
Asparagine	5.41	-4.5	2.02	8.8
Glutamic acid	3.22	-3.5	2.19	9.67
Glutamine	5.65	-3.5	2.17	9.13
Arpinine	10.76	1.8	2.17	9.04
Lysine	9.74	-3.9	2.18	8.95
Histidine	7.59	-3.2	1.82	9.17
Cysteine	5.02	2.5	1.71	8.33
Methionine	5.75	1.9	2.28	9.21
Phenylalanine	5.48	2.8	1.83	9.13
Tyrosine	5.66	-1.3	2.38	9.11
Tryptophan	5.89	-0.9	5.89	9.39
Proline	6.30	-1.6	1.99	10.60

依据上表内容,根据氨基酸的不同理化性质,每条特定的残基序列都能够对应生成一组包含相应属性的序列,已有研究表明,蛋白质二级结构的形成,与其各项理化指标有着必然的联系^[6]。因此,通过编程转换,将这样的序列作为聚类的输入参数。

4 基于蚁群算法的模型建立

蚁群算法是一种新型的模拟进化算法,此算法模拟一群真实蚂蚁的协作过程,每一只蚂蚁在其候选解空间中独立地搜索解并保存解的信息^[7]。一般来说,信息量越大的解被选中的可能也越大。蚁群算法根据聚类中心的信息把周围的数据归并到一起,从而得到聚类结果,是一种较新的全局化启发式算法,在此用它来对输入数据进行划分。蚁群算法将需要聚类的数据视为具有不同属性的蚂蚁,聚类中心则看成蚂蚁需要寻找的“食物源”。确定聚类中心的过程就是蚁群从蚁穴出发去寻找“食物源”的过程,在搜索时,不同的蚂蚁选择某个数据元素是相互独立的^[8]。

假如输入 $X = \{X_i \mid i = 1, 2, \dots, n\}$, $X_i = (x_{i1}, x_{i2}, \dots, x_{im})$ 有 n 个输入样本数据。

令 $d_{ij} = ||P(X_i - X_j)||_2$, d_{ij} 表示 X_i 到 X_j 之间的加权欧式距离, P 为加权因子,可以根据各分量在聚类中的贡献设定。

设 r 表示聚类半径, ε 表示统计误差, $\tau_{ij}(t)$ 是 t 时刻数据 X_i 到数据 X_j 路径上残留的信息量, 在初始时刻各条路径上的信息量相等且为 0。在路径上的信息量由式(1)给出。

$$\tau_{ij}(t) = \begin{cases} 1 & d_{ij} \leq r \\ 0 & d_{ij} > r \end{cases} \quad (1)$$

X_i 是否归并到 X_j 由下式给出:

$$P_{ij}(t) = \frac{\tau_{ij}^\alpha(t) \eta_{ij}^\beta(t)}{\sum_{s \in S} \tau_{is}^\alpha(t) \eta_{is}^\beta(t)}, S = \{X_s \mid d_{is} \leq r, s = 1, 2, \dots, i, j+1, \dots, N\} \quad (2)$$

这里 η_{ij} 表示 X_i 归并到 X_j 的期望程度, S 是蚂蚁 X_i 下一步可以选择的路径集合。如 $P_{ij} \geq P_0$, 则 X_i 归并到 X_j 领域。令 $C_j = \{X_k \mid d_{kj} \leq r, k = 1, 2, \dots\}$, C_j 表示所有归并到 X_j 领域的数据集合。

求出理想的聚类中心: $\bar{C}_j = \frac{1}{J} \sum_{k=1}^J X_k, X_k \in C_j$

基于蚁群聚类学习方法的具体算法步骤如下:

1) 首先根据经验或随机选择 M 个代表点。一般而言, 初始代表点的选择往往会影响到迭代的结果, 即可能得到的是局部最优解而不是全局最优解^[9]。针对这个问题, 可以尝试不同的初始代表点, 以免陷入局部最小点;

2) 初始化设定 $N, m, r, \varepsilon_0, \alpha, \beta, \tau_i(0) = 0, P_0, M$;

3) 计算 $d_{ij} = ||P(X_i - X_j)||_2 =$

$$\left(\sum_{k=1}^m P_k (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}};$$

4) 计算各路径上的信息量 $\tau_{ij}(t) = \begin{cases} 1 & d_{ij} \leq r \\ 0 & d_{ij} > r \end{cases}$

5) 计算 X_i 归并到 X_j 的概率 $P_{ij}(t) =$

$$\frac{\tau_{ij}^\alpha(t) \eta_{ij}^\beta(t)}{\sum_{s \in S} \tau_{is}^\alpha(t) \eta_{is}^\beta(t)}, S = \{X_s \mid d_{is} \leq r, s = 1, 2, \dots, i, j+1, \dots, N\},$$

判断 $P_{ij} \geq P_0$ 是否成立, 成立继续执行, 否则 $i+1$ 转(3);

6) 按 $\bar{v}_j = \frac{1}{J} \sum_{k=1}^J X_k, X_k \in v_j$ 计算该聚类中心;

7) 计算 $D_j = \sum_{k=1}^J \left(\sum_{i=1}^m (x_{ki} - v_{ji})^2 \right)^{\frac{1}{2}}$, 计算第 j 个聚类的偏离误差及总体误差, 其中 v_{ji} 表示 j 个聚类中心的第 i 个分量, 计算总体误差 $\varepsilon = \sum_{j=1}^k D_j$;

8) 判断 $\varepsilon \leq \varepsilon_0$ 是否成立, 成立输出聚类个数 c ; 若不成立, 转(3)继续迭代。

5 聚类结果及分析

每次从蛋白质数据库中随机抽取 1500 条三种二级结构的序列片段, 并通过程序比对得到相应的特征

序列, 如疏水值、解离常数。在序列长度选取上, 选取了长度为 30 的序列片段, 这是由于如果序列长度过短, 其序列排列组合概率出现的情况相对均匀, 无法反映特征, 而太长的序列则不具备普遍性, 不利于后面得到合理的聚类结果。将得到的 1500 条特征序列作为聚类算法的输入。将每次聚类得到的结果与上次聚类结果进行交叉比对分析, 最终得到稳定的类别。图 1 为聚类结果在空间中的表示, 可以看到几个类别在空间中区分明显, 取得了较好的聚类效果。为了更好地呈现聚类结果, 把各个类别的三种属性序列的特征集中在表 3 中体现:

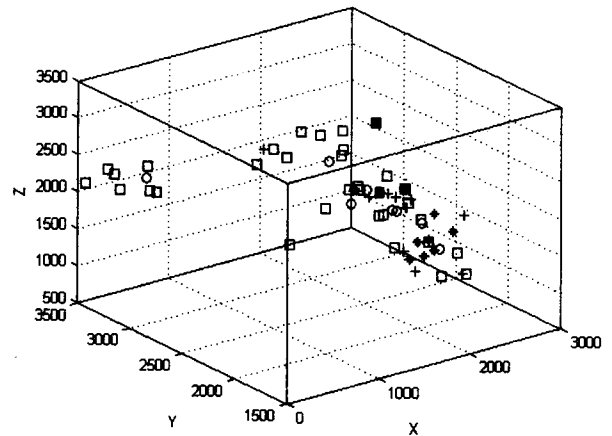


图 1 蚁群聚类结果 ($R = 100, t = 1000$)

从三种属性的聚类结果来看, 在 α 片段中, 根据 α 聚类结果 1~7 体现出了很好的亲疏水残基间隔模式, 其中类别 3 中的聚类结果, 亲疏水性特征序列基本呈对称模式, 特征尤为明显。在等电点性质序列中, 一定长度的 α 序列片段会有一个个明显的波峰出现。

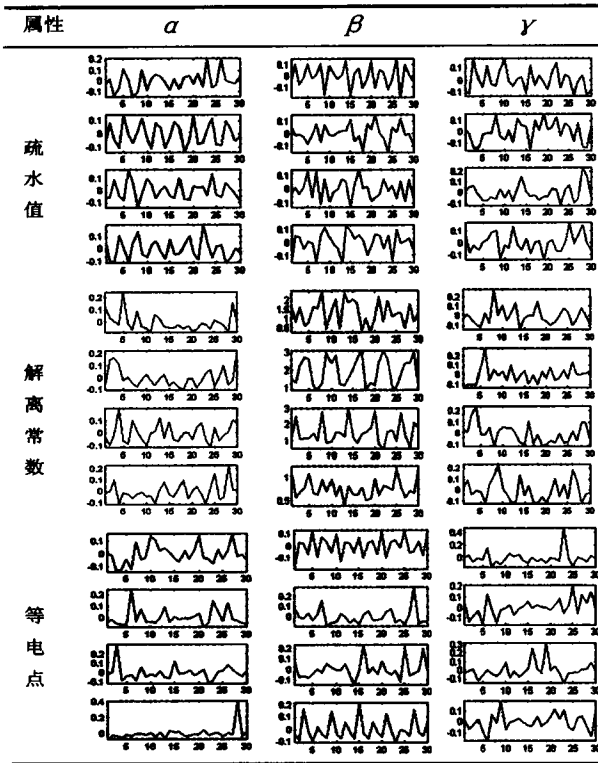
从 β 片段针对各个属性的聚类结果中均可以看出其表现出的特征非常明显, 通常一个长度为 30 的序列最起码会出现 2 到 3 个明显的折叠区间, 波峰、波谷的值以及出现间隔均相对固定, 反映在实际的氨基酸序列当中应该即为锯齿状的多重折叠结构。根据这些特征, 甚至还能够对其进行进一步细分, 例如针对解离常数属性聚类中, 类别 5, 6 的波峰出现之前均有一个相对缓和的坡度, 另一类则表现为波峰波谷交替出现且差值距中间点相对稳定^[10,11]。

针对无规则卷曲的聚类结果则呈现出明显的不均匀性, 特征较为模糊, 即说明, γ 结构松散, 没有统一的确定标准。因为从已知的实际结构中也可以看出无规则卷曲所代表的结构的类别特征本身就是很不明显的、无规则的。

上述结果表明基于密度聚类的方法能够高效并且合理地蛋白质三态序列进行聚类。总体来看, 聚类结果虽然相对比较分散, 但是每种都呈现出了不同的特征:

α 片段表现出很好的折叠性(疏水值正负交替);
 β 片层表现出周期规律性(与自身的折叠结构相关);
 γ 不规则片段表现比较琐碎,依据前两类可以排除得到这一类的分类。

表 3 主要物化性质的聚类结果图



通过对每次实验中,不同序列聚类得到的种类计数可以得出,83% 的 α 螺旋以及 87% 的 β 折叠体现出自身应有的特征。

6 结束语

通过上述实验的对比分析,发现了蛋白质三种二级结构各自的明显特征,但是这样的结果还是停留在定性分析之中,并没有特别详细的数据结果来充分进行说明,但是已经足够为下一步的研究奠定良好的基础。在后续研究中,将进一步合理优化序列相关度计算函数,同时增加长度不同的多序列对比分析,以提高

样本输入质量。另外,该聚类算法对单次运算的样本数目有着较为严格的要求,将会继续改进聚类算法以求满足更高的吞吐量,最终达到对蛋白质数据库中的全局序列进行一次性聚类分析,进一步量化提取上述特征,作为三维结构特征分析的依据,同时与其相应的三维空间结构建立关联规则。

参考文献:

- [1] 焉为家,郭雨珍.改进的粒子群算法求解蛋白质结果预测问题[J].计算机技术与发展,2011,21(12):109-112.
- [2] Zhang H X. The research of protein secondary structure prediction methods[D]. Dalian: Dalian University of Technology, 2004.
- [3] Chou P Y, Fasman G. Prediction of Protein Conformation [J]. Biochemistry, 1974, 13(2): 222-245.
- [4] Pollastri G. Structure, Function and Genetics [J]. Proteins, 2002, 47: 228-235.
- [5] Bo J, Guo T, Peng L W. Folding type-specific secondary structure propensities of amino acids, derived from α , β , α/β and $\alpha+\beta$ proteins of known structure [J]. BioPolymer, 1998, 45(1): 35-49.
- [6] Attwood T K, Croning R D M, Flower D R, et al. PRINTS: the database formerly known as PRINTS [J]. Nucleic Acids Res, 2000, 28: 225-227.
- [7] William N G, Bailey T L, Elkan C P, et al. Meta-MEME: Motif-based hidden markov models of biological sequences [J]. Computer Applications in the Biosciences, 1997, 13(4): 397-406.
- [8] Bailey T L, Elkan C. Unsupervised learning of multiple Motifs in biopolymers using expectation maximization [J]. Machine Learning, 1995, 21(1-2): 51-83.
- [9] 张斐,谭军,谢竞博.基于不同算法的 motif 预测比较分析与优化[J].计算机工程,2009,35(22): 94-96.
- [10] William N G, Bailey T L, Elkan C P, et al. Principles of protein folding - a perspective from simple exact models [J]. Protein Science, 1995(4): 561-602.
- [11] Moulton J, Fidelis K, Kryshchovych A, et al. Critical assessment of methods of protein structure prediction - Round VIII [J]. Proteins, 2009, 77(S): 1-4.

(上接第 190 页)

- [3] 王功明,郭新宇,赵春江,等.粒子系统的优化技术研究[J].计算机应用研究,2008,25(2): 495-497.
- [4] 孙少斌.分布式作战仿真技术[D].蚌埠:蚌埠坦克学院, 2010.
- [5] 孙少斌,李大鹏,陈璐.虚拟战场环境特殊效果仿真[J].火力指挥与控制,2010,35(12): 177-180.
- [6] Microsoft Corporation. Windows DirectX Graphics Documentation [M]. [s. l.]: [s. n.], 2009.
- [7] 汪继文,郑峰.基于 OpenGL 与粒子系统的喷泉模拟实

现[J].计算机技术与发展,2011,21(8): 161-164.

- [8] 张芹,吴慧中,张健.基于粒子系统的建模方法研究[J].计算机科学,2003,30(8): 144-146.
- [9] 李建明,吴云龙,何荣盛,等.基于粒子系统和 GPU 加速的喷泉实时仿真[J].系统仿真学报,2009,21(10): 3139-3141.
- [10] 贾丽.基于粒子系统的自然现象仿真[D].武汉:武汉理工大学,2008.
- [11] 冷冕冕.坦克分队分布式虚拟战场环境构建研究[D].蚌埠:蚌埠坦克学院,2009.

基于蚁群聚类的蛋白质二级结构特征研究

作者: 高冶, 陈琦, GAO Ye, CHEN Qi
作者单位: 海南大学信息科学技术学院, 海南海口, 570228
刊名: 计算机技术与发展 
英文刊名: Computer Technology and Development
年, 卷(期): 2013, 23(6)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjfz201306049.aspx