

近邻传播聚类算法在蛋白质结构预测中的应用

黄 旭¹, 吕 强^{1,2}, 杨凌云¹, 钱培德^{1,2}

(1 苏州大学 计算机科学与技术学院, 江苏 苏州 215006; 2 江苏省计算机信息处理技术重点实验室, 江苏 苏州 215006)

摘 要: 聚类是蛋白质结构预测中重要的后处理步骤, 许多结构预测中都采取了不同的聚类算法. 而 AP 聚类算法通过在数据点之间传递消息, 经过若干次迭代后达到一种稳定状态, 是构思巧妙的聚类算法. 文中把 AP 聚类算法应用于蛋白质结构预测中, 并在 7 个不同的数据集上进行了实验. 结果表明, 在采用 RMSD 进行结构相似性度量的情况下, AP 算法有 67% 的结果优于 Rosetta 聚类算法或相当, 是一种适合蛋白质结构聚类的算法.

关键词: 蛋白质结构预测; AP 聚类算法; RMSD

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2010)11-0154-04

Application of Affinity Propagation Clustering Algorithm in Protein Structure Prediction

HUANG Xu¹, LV Qiang^{1,2}, YANG Ling-yun¹, QIAN Pei-de^{1,2}

(1 School of Computer Science and Technology, Soochow University, Suzhou 215006, China;

2 Jiangsu Provincial Key Lab for Computer Information Processing Technology, Suzhou 215006, China)

Abstract: Clustering decoys is a necessary post-processing step for protein structures prediction. There are many clustering algorithms used in protein structures prediction. The Affinity Propagation (AP) clustering algorithm is ingenious as it propagate some information among all the data points, and reach a steady state after some iterations. This paper applies the affinity propagation clustering algorithm in protein structure prediction, and test it in seven different data sets. And then, we compare the exemplars with their native conformations with the RMSD metric. Results show that the Affinity Propagation clustering algorithm is better than Rosetta clustering algorithm under 67%, and it is fit for protein structure prediction.

Key words: protein structure prediction; Affinity Propagation clustering algorithm; RMSD

1 引言

蛋白质结构预测是用计算的方法^[1], 从蛋白质的氨基酸序列出发, 预测蛋白质三维结构的过程. 通常的做法是由预测过程产生出相当数量的具有较低能量的候选结构(decoys), 然后通过聚类的方法寻找最稳定的聚类中心, 普遍认为这一稳定的聚类中心最接近于最低自由能状态^[2].

聚类是一种无监督的模式识别方式^[3]. 聚类算法大致分为分类数目已知与分类数目未知两类问

题^[4], 蛋白质候选结构的聚类显然属于第二类. 然而对同一目标问题而言, 不同的聚类策略也将导致不同的结果. 目前一种常用的聚类算法是 Rosetta 所采用的聚类算法, 该算法逻辑简单, 但聚类效率高.

但在 Rosetta 聚类算法中, 聚类半径是一个相当关键的算法参数. 而根据事先设定的聚类半径对候选集合进行硬性划分, 很难保证这种划分是合理的.

2 近邻传播聚类算法

文中采用近邻传播(Affinity Propagation, AP)

聚类算法^[5]对蛋白质结构预测所产生的候选结构集合进行聚类,取得了较好的效果.近邻传播聚类算法以数据点对之间的相似性作为输入,在数据点之间交换消息,经过若干次迭代之后,消息的传播达到一种稳定状态.在这样一个过程中,有两种类型的消息:一种被称作 responsibility (记作 $r(i,k)$),另一种是 availability (记作 $a(i,k)$).

消息 $r(i,k)$ 由数据点 i 发出,并由潜在的聚类中心 k 接收,该消息反映了点 k 主动作为点 i 所在类中心的意愿.该值越大,说明作为类中心的意愿越强烈.而消息 $a(i,k)$ 由潜在的聚类中心 k 发出,并由数据点 i 接收,反映了点 i 选择点 k 作为自己所在类中心的可能性.该值越大,说明选择对方作为聚类中心的意愿越强烈.因此这两种消息实际上反映了在数据点集合中“自荐”与“选举”的交互.消息值也是随着迭代而不断变化的,最终的平衡状态应该呈现一种合理的聚类划分.通过这种方式,AP 算法避免了聚类半径所带来的硬性划分而使类别之间的界限更加自然.

消息 availability 初始化为 0: $a(i,k) = 0$, 而 $r(i,k)$ 和 $a(i,k)$ 由下列公式进行迭代计算:

$$r(i,k) \leftarrow s(i,k) - \max_{k's.t.k' \neq k} \{a(i,k') + s(i,k')\}$$
 (1)

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i's.t.i' \in \{i,k\}} \max\{0, r(i',k)\}\}$$
 (2)

$$a(k,k) \leftarrow \sum_{i's.t.i' \in \{i,k\}} \max\{0, r(i',k)\}$$
 (3)

其中 $s(i,k)$ 是数据点 i 与 k 之间的相似性.

在消息 availabilities 和 responsibilities 不断传播的过程中,二者可以共同确定聚类中心.对于点 i , 具有最大值的 $a(i,k) + r(i,k)$, ($i,k = 1,2,\cdots,n$) 指出了点 i 属于中心点 k 所对应的类(此时 $k \neq i$),反之,若 $k = i$,则表明点 i 将选择自己作为其所在类的中心.

AP 算法中有一个重要的参数是 preference,该参数表明了数据点选择自己作为聚类中心的意愿,该值正好补全了相似性矩阵中缺失的对角线. Preference 值越大,将导致最终聚类结果中类的数目越多,相应地,每一类中成员的数目会更少.

3 数据集与实验过程

文中选择了 7 个不同的数据集.包括 I-TASSER^[6]产生的:1abv_、1af7_、1cqkA、1csp_、1dcjA、256bA 等 6 个数据集,这些数据下载自 <http://>

zhang. ~ bioinformatics. ku. edu/ I - TASSER/decoys/. 同时还选择了数据集 2hq7_,这是采用 Rosetta 结构预测程序在本地产生的.这 7 个数据集相关数据如表 1 所示.

表 1 实验用到的数据集

序号	数据集	蛋白质	长度	数据集规模	来源
1	1abv_ - 10	1abv_	103	1250	I - TASSER
2	1af7_ - 10	1af7_	72	1250	I - TASSER
3	1cqkA - 10	1cqkA	101	2000	I - TASSER
4	1csp_ - 10	1csp_	67	1250	I - TASSER
5	1dcjA - 10	1dcjA	73	2000	I - TASSER
6	256bA - 10	256bA	106	2000	I - TASSER
7	2hq7_ - 10	2hq7_	141	920	ROSETTA

为提高实验效果,文中对每一数据集进行了三种不同标准的相似性计算,得出三个不同的相似性矩阵,然后分别应用于 AP 与 Rosetta 聚类算法.这样在 7 个数据集上可获得 21 组比较数据.采用的相似性度量标准包括 RMSD(Root Mean Square Deviation)^[7]、TM-score 以及 GDT_TS(Global Distance Test_Total Score).其中,RMSD 是反映两个结构之间相应原子间位置差异的量,在不同的精度下可以采用不同的计算方法,比如全原子 RMSD 与 C_α 原子的 RMSD 等.文中采用 C_α 原子的 RMSD.

$$RMSD = \sqrt{\frac{\sum_{i=1}^n d_i^2}{n}}$$
 (4)

其中, d_i 是不同结构中相对应的第 i 对 C_α 原子之间的距离,单位为 Å, n 是蛋白质氨基酸链上 C_α 原子的数目. RMSD 越小说明两个结构相似性越高.因此, RMSD 表明的是两个结构之间的差异性,对于 AP 算法而言,需转换为相似性才能作为输入.文中采用取负数的方法进行转换.

尽管 RMSD 可以指出模型的误差,但在有些情况下,即使大多数 C_α 原子都重合得很好,而局部的错误(如尾部定向错误)也会导致大的 RMSD 值^[8]. 文献^[6,8-10]提出的 TM-score 可以避免这种情况. TM-score 可以评估两个结构空间上的相似性,公式如下:

$$TM\text{-score} = \frac{1}{L} \sum_{i=1}^L \frac{1}{1 + d_i^2/d_0^2}$$
 (5)

其中 L 是蛋白质长度, d_i 是第 i 对残基之间的距离, $d_0 = 1.24 \sqrt[3]{L - 15} - 1.8$. TM-score 值域区间为 $[0,1]$,当两个结构完全重合时, TM-score 值为 1.

GDT_TS^[9] 定义为在阈值 $D_0 = 1, 2, 4, 8 \text{Å}$ 时 GDT 值的平均值, 而 GDT 反映了骨架上相对应的 C_α 原子距离在某个阈值 D_0 之内的残基比例. GDT_TS 计算公式如下:

$$\text{GDT_TS} = \frac{1}{4}(N_1 + N_2 + N_4 + N_8) \quad (6)$$

其中 N_i 为 $D_0 = i \text{Å}$ 时的 GDT 值. GDT_TS 值域区间也是 $[0, 1]$, GDT_TS 值越大说明两个结构的相似性越高.

文中在采用 1.6GHz CPU 的 IBM pServer 平台上计算数据点之间的相似性矩阵, 在采用 3GHz Pentium 4 CPU 的 PC 机上进行 AP 聚类. 为便于比较, 同时采用 Rosetta 聚类算法在相同场景下进行了聚类实验.

如前所述, 聚类参数的不同也会导致聚类结果有差异, 在 Rosetta 中聚类半径就是这样一个参数. 在 AP 算法中, preference 参数也会导致聚类结果的差异. 为弥合这种差异, 文中针对 21 种不同场景下的聚类都采用了相同的方式进行处理. 首先尝试选取边界参数 P_0 与 P_{N+1} , 使得在参数 P_0 下每个点自成一类, 而在参数 P_{N+1} 下所有点聚成一类. 然后将区间 $[P_0, P_{N+1}]$ 分为 $N+1$ 等份以获取 $P_1, P_2, \dots, P_N$ 共 N 个参数, 最后分别采用这 N 个参数进行聚类. 文中选择 $N=5$. 在每一次聚类尝试中, 根据类的大小选择最大的 R_{MAX} 个聚类中心, 文中选择 $R_{\text{MAX}} = 10$. 接下来分别统计这 $N \times R_{\text{MAX}}$ 个聚类中心点的频数、位次, 以及类的大小等信息, 了解聚类中心的稳定性、类的密度, 并据此选择所谓的最佳聚类中心. 由于在聚类过程中并没有参考天然结构的信息, 因此这种选择是合理的, 也是公平的. 通过这种方法选择出来的聚类中心作为本次聚类的最终结果.

4 结果分析

文中分别在不同的 7 个数据集上计算 3 组不同的相似性矩阵, 然后分别将这些相似性矩阵作为 AP 与 Rosetta 聚类算法的输入, 通过调整聚类参数进行多次尝试, 最后对聚类中心进行择优. 选出每组实验最终的聚类结果后, 计算该聚类中心与天然结构的差异程度. 文中采用的是 RMSD 评价指标. 结果如表 2 所示.

由于 RMSD 表明了两个结构之间的差异性, 因此 RMSD 较小的结果更好. 由表中可以看出, 在 21 组比较数据中, AP 算法有 10 组优于 Rosetta, 有 4 组比较结果相同, 合计占有所有比较实验的 67%; 只有 7 组

劣于 Rosetta, 仅占 33%. 因此, 实验结果表明采用 AP 算法的聚类结果要优于 Rosetta 的聚类结果.

表 2 AP 与 Rosetta 聚类算法结果比较

序号	数据集	相似性度量	AP	ROSETTA
1	1abv_-10	RMSD	12.510	12.500
2	1abv_-10	TM-score	12.441	12.500
3	1abv_-10	GDT_TS	12.199	12.500
4	1af7_-10	RMSD	8.878	7.897
5	1af7_-10	TM-score	8.878	8.884
6	1af7_-10	GDT_TS	8.878	9.088
7	1cqkA-10	RMSD	8.938	8.721
8	1cqkA-10	TM-score	8.938	8.721
9	1cqkA-10	GDT_TS	8.733	8.851
10	1csp_-10	RMSD	2.427	2.427
11	1csp_-10	TM-score	2.498	2.478
12	1csp_-10	GDT_TS	2.427	2.561
13	1dcjA-10	RMSD	9.785	10.015
14	1dcjA-10	TM-score	9.967	10.688
15	1dcjA-10	GDT_TS	10.015	10.015
16	256bA-10	RMSD	3.227	3.438
17	256bA-10	TM-score	3.227	3.057
18	256bA-10	GDT_TS	3.438	3.438
19	2hq7_-10	RMSD	12.441	16.621
20	2hq7_-10	TM-score	16.952	15.967
21	2hq7_-10	GDT_TS	16.952	16.952

5 结束语

AP 聚类算法作为一种构思巧妙的聚类算法可以用于蛋白质结构聚类. AP 算法的输入是数据点两两之间的相似性矩阵, 由于蛋白质空间结构相似性计算的复杂性, 致使相似性的计算耗费了大量时间, 如何提高相似性的计算速度是下一步需要重点解决的问题. 此外, 对于较大规模的数据集往往得到规模巨大的相似性矩阵, 受内存限制, AP 算法有时候不能正确处理. 下一步还将针对基于 AP 算法的增量聚类做进一步研究.

参考文献:

- [1] Christian B Anfinsen. Principles that govern the folding of protein chains[J]. Science, 1973(181):223-230.

- [2] Zhang Y, Skolnick J. Spicker: approach to clustering protein structures for near-native model selection[J]. Journal of Computational Chemistry, 2004(25):865-871.
- [3] 张昕, 彭宏, 郑启伦. 基于微粒群算法的聚类分析[J]. 微电子学与计算机, 2006, 23(9): 94-95.
- [4] 高尚, 汤可宗, 杨静宇. 一种新的基于混合蚁群算法的聚类方法[J]. 微电子学与计算机, 2006, 23(12): 38-40.
- [5] Brendan J Frey, Delbert Dueck. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.
- [6] Yang Zhang. I-tasser server for protein 3d structure prediction[J]. BMC Bioinformatics, 2008(9):40-47.
- [7] Andriy Kryshchak, Maciej Milosian, Lukasz Szajkowski, et al. Casp6 data processing and automatic evaluation at the protein structure prediction center[J]. Proteins: Structure, Function, and Bioinformatics, 2005(7):19-23.
- [8] Sitao Wu, Jeffrey Skolnick, Yang Zhang. Ab initio modeling of small proteins by iterative tasser simulations[J]. BMC Bioinformatics, 2007(5):17-26.
- [9] Machael Tress, Iakes Ezkurdia, Osvaldo Grana, et al. Assessment of predictions submitted for the casp6 comparative modeling category[J]. Proteins: Structure, Function, and Bioinformatics, 2005(7):27-45.
- [10] Yang Zhang, Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality[J]. Proteins: Structure, Function, and Bioinformatics, 2004, 57(4):702-710.

作者简介:

黄旭 男,博士研究生.研究方向为智能信息处理新技术.
吕强 男,教授,博士生导师.研究方向为生物信息计算、元启发搜索、并行分布计算.
杨凌云 男,硕士研究生.研究方向为智能信息处理新技术.
钱培德 男,教授,博士生导师.研究方向为中文信息处理理论.

(上接第 153 页)

4.4.3 结果评估

通过表 2 和图 2 可见,采用设计的综合预测模型和算法,能够实现预测,预测数据精度得到提高,满足预测目的和要求,具有实际的应用价值.

5 结束语

网络设备工作负载由于工作环境、影响因素等复杂性,没有具体标准,特别是定量预测、计算和分析,不容易准确预测.根据文中提出的综合预测模型,采用给出的具体预测方法及步骤,实际应用预测效果良好,为网络设备动态工作负载定量预测提出了参考方法,在网络建设、管理和应用工作中有一定的实用价值.

参考文献:

- [1] 元雪冬,李霞,梁鸿,等. 计算网格中动态负载平衡策略

研究[J]. 微电子学与计算机, 2009, 26(5): 21-24.

- [2] 邹心遥,姚若河. 基于 BP 神经网络的电子元器件寿命预测[J]. 微电子学与计算机, 2009, 26(1): 52-53.
- [3] 曹建华,刘渊,戴悦. 一种基于灰色神经网络的网络流量预测模型[J]. 计算机工程与应用, 2008, 44(5): 155-157.
- [4] 曹建华. 基于灰色神经网络模型的网络流量预测算法研究[D]. 无锡:江南大学, 2009.
- [5] 李福琴. 灰色模型的稳定性和建模精度研究[D]. 武汉: 武汉理工大学, 2007.
- [6] 胡春光,高燕,李颖. 一种扩展滑动窗口算法[J]. 微电子学与计算机, 2007, 24(8): 106-112.
- [7] 吴昌友. 神经网络的研究及应用[D]. 哈尔滨:东北农业大学, 2007.

作者简介:

张靖 男, (1972-), 硕士, 副教授. 研究方向为软件工程和计算机网络.