第36卷 第14期 计 算 机 工 程 2010年7月
Vol.36 No.14 Computer Engineering July 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)14—0233—02 文献标识码：A 中图分类号：TP393

基于CGR的蛋白质相似性比较

徐 占，董洪伟
(江南大学信息工程学院，无锡 214122)

摘 要：从蛋白质结构特性出发，利用结构字母表和CGR游走技术将蛋白质三维结构信息转换到二维坐标空间中。通过分析所得图像找出蛋白质分子的主体结构，获得各结构点在CGR图中的坐标，利用Hausdorff距离判定要比较的蛋白质对象相似性。该方法实现了蛋白质相似性比较的结构-序列模式转变，利用Hausdorff距离比较两点集间相似性的优势，为蛋白质相似性比较提供了一种简便有效的方法。
关键词：结构字母表；主体结构；Hausdorff距离

Protein Similarity Comparison Based on CGR

XU Zhan, DONG Hong-wei
(School of Information Engineering, Jiangnan University, Wuxi 214122)

【Abstract】Starting from the characteristics of protein structures, a new method based on structure alphabet and Chaos Game Representation(CGR) is provided. Its can find the main structure of the protein easily by CGR, and make use of CGR and Hausdorff distance to complete protein similarity comparison. Compared with other methods based on amino acid sequence, this method can resolve the problem that different proteins are classified into one family just for the similar sequence of amino acid. The proteins with similar sequence of amino acid may be belong to different protein families, because they have different structures. The functions of protein depend on their structures, so this method can get better result of similarity compared with that of amino acid sequence.
【Key words】structure alphabet; main structure; Hausdorff distance

1 概述

现有的绝大多数有关蛋白质相似性比较的算法都围绕序列比对和结构比对。比较蛋白质的空间结构，可以发现蛋白质的结构共性、属于同一家族蛋白质的保守结构、与功能密切相关的结构域以及特定的空间结构模式，而这种模式在进行序列分析时却无法发现。

对序列-结构模式的发现方法主要分为2类：(1)从结构模式出发，寻找可能形成特定结构的序列；(2)从序列模式出发，搜索其可能形成的结构。

文献[1]用K-均值聚类方法研究蛋白质结构片段库并得到了一个结论，即为了准确地模拟蛋白质结构和构建近似结构，建立一个大的片段库是没有必要的。结构字母表的出现将蛋白质的三维结构转换为编码空间中的字符串序列，从而使三维空间中的结构模式转换为编码空间中的序列模式，这为运用Chaos Game Representation(CGR)游走技术研究蛋白质三维空间结构的相似性提供了可能。

2 结构字母表

DSSP是最有名且应用最广泛的结构字母表。表1列出了DSSP对二级结构所做的分类情况。

表1 DSSP八字母的二级结构字母

Letter	Name
E	beta strand
H	alpha helix
T	turn
S	bend
G	3-10 helix
B	short beta bridge
I	pi helix
C	random coil

文献[2]根据氢键的类型将蛋白质的二级结构分为8类。该方法定义了一个基于n-turns的层次结构特性，其中的氢键是第i个残基的CO与第$i+n$个(n=3, 4, 5)残基的NH形成的。该特性包括α螺旋、β折叠、无规则的螺旋扭曲和β凸起以及无规则卷曲。

2.1 STRIDE

最初的STRIDE字母表由7个字母组成，仅比表1中的结构字母少一个S。由于I结构数量过少不易被预测而且对预测结果影响很小，因此Rachel将字母I并入H中，得到一个6字母的字母表。

2.2 DSSP-EHL和STRIDE-EHL

DSSP-EHL和STRIDE-EHL字母表是在DSSP和STRIDE的基础上导出的， 2个字母表都含有3个字母，即{E, H, L}。3个字母分别代表螺旋、折叠和卷曲3种结构状态。在实际映射中，该字母表将DSSP中的G划归为螺旋类，将B划归为折叠类，将S和T划归为卷曲类。

3 本文方法

3.1 结构字母表的确定

在蛋白质结构预测方面，状态数目在7~10之间的字母表比状态数目为2或3的字母表得到的效果更好[3]。Rachel等将8字母DSSP中的I并入H中得到7字母的DSSP。为了对蛋白质结构进行更简便有效的描述，本文将I复原为单独的一个字母，即DSSP的原始8字母状态。

3.2 CGR游走

CGR是一种迭代映射技术，即把序列中的每一个单元映射到一个连续的坐标空间中。

作者简介：徐 占(1982－)，男，硕士研究生，主研方向：计算机图形学，分子图形学；董洪伟，副教授、博士
收稿日期：2009-12-27 **E-mail：**xuzhan03521@163.com

CGR 在本文中的应用如下：

首先，在笛卡尔坐标空间中建立一个边长为 1 的正八边形，在其 8 个顶点位置分别放置 8 个结构字母，由于得到的 CGR 图形受结构字母排列顺序的影响很大[4]，因此在排列时将结构相近的字母分开排列，即将 E, H, T, S, G, B, I, C 等 8 个结构字母按照逆时针方向给出。其次，将得到的结构字母序列按下式安放在正八边形内。

$$CGR_i=(Q_i+CGR_{i-1})/2 \quad (1)$$

其中，Q_i 为当前字母的坐标，CGR_{i-1} 为前一游走点的坐标。比如要安放的结构字母的序列段为 HTSTB，先将 H 放在中心点与对应点 H 的中间点，将 T 放在 H 与对应点 T 的中间点，按照相同方法依次完成所有字母的安放，得到图 1。然后分析得到的图形，找出蛋白质的主体结构，对比较对象进行粗分类，利用 Hausdorff 距离完成蛋白质相似性的比较。

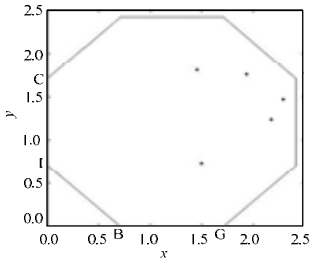

图 1　HTSTB 在图中的安放位置

3.3　Hausdorff 距离

作为研究分形和拓扑的重要概念，Hausdorff 距离[5]是描述 2 组点集之间相似程度的一种度量，即集合之间距离的一种定义形式：假设 2 组有限点集合 $A=\{a_1,a_2,\cdots,a_p\}$ 和 $B=\{b_1,b_2,\cdots,b_q\}$，则这 2 组点集间的 Hausdorff 距离定义为

$$H(A,B)=\max(h(A,B),h(B,A)) \quad (2)$$

其中，

$$h(A,B)=\max_{a_i\in A}\min_{b_j\in B}\| a_i-b_j \| \quad (3)$$

$$h(B,A)=\max_{a_j\in A}\min_{b_j\in B}\| b_j-a_i \| \quad (4)$$

其中，式(1)为 Hausdorff 距离的最基本形式，式(2)和式(3)分别称为从 A 集合到 B 集合和从 B 集合到 A 集合的单向 Hausdorff 距离，即 $h(A,B)$首先求出点集 A 中的每个点 a_i 到 B 集合中所有点 b_j 之间的距离最小值，然后从中选出最大值作为 $h(A,B)$的值。同理可得 $h(B,A)$的值。由式(1)可知，Hausdorff 距离是单向距离 $h(A,B)$和 $h(B,A)$两者中的较大者，它度量了 2 个点集间的最大不匹配程度。在本文中应用过程如下：

(1)在分别得到蛋白质对象的 CGR 之后，以各自结构点的坐标构成 2 个点集 A 和 B。

(2)利用上述公式计算 Hausdorff 距离。本文计算了 2 个点集中所有点之间的距离，因此，$h(A,B)$和 $h(B,A)$在这里是相等的。

(3)为了对蛋白质比较对象的相似性进行定量的分析，本文定义了一个新的相似度表达式，即

$$S=(D_{\max}-H(A,B))/D_{\max} \quad (5)$$

其中，D_{max} 是一个常数，即 2.613 1。它表示在边长为 1 的正八边形中任意 2 个结构点间可能的最大距离偏差。

4　实验与分析

4.1　粗分阶段

本文以肌红蛋白 101M、102M 为例进行实验。3 个蛋白质对象的结构字母序列可以在 PDB 数据库中找到。按照 3.2 节的方法，可以得到 101M 和 102M 的 CGR 图，见图 2、图 3。通过图像并结合式(5)各结构的贡献率，可以直观地找到 101M 和 102M 的主体结构，其结构贡献率见表 2。下一步的工作是对 101M 和 102M 的相似性进行比较。

$$P_i=n_i/N \ (i=H,T,S,G,B,I,C,E) \quad (6)$$

其中，n_i 为各个结构字母的数目；N 为总的结构数目。

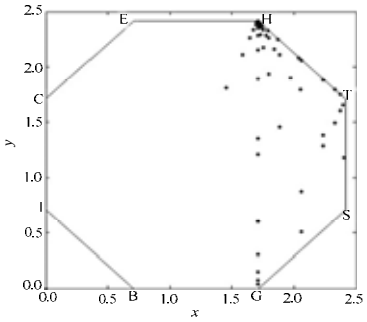

图 2　101M 在正八边形中的 CGR 图

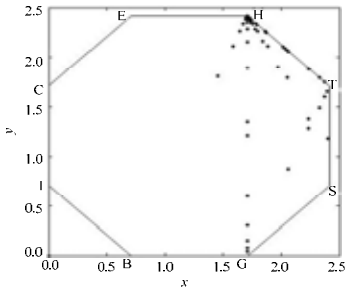

图 3　102 M 在正八边形中的 CGR 图

表 2　3 种蛋白质分子中各结构的贡献率和累积贡献率

结构字母	贡献率 P_i/(%)		累积贡献率/(%)	
	101M	102M	101M	102M
E	0.00	0.00	0.00	0.00
T	8.89	9.78	8.89	9.78
H	80.00	81.95	88.89	91.73
S	2.22	1.50	91.11	93.23
G	8.89	6.77	100.00	100.00
B	0.00	0.00	100.00	100.00
I	0.00	0.00	100.00	100.00
C	0.00	0.00	100.00	100.00

4.2　细化阶段

在细化阶段，由于在蛋白质结构的构成中，C, I, B 三结构字母所表示的结构在整个蛋白质结构中数量很少，因此为了更方便地对得到的 CGR 图进行曲线拟合等其他操作，本文将所得 CGR 图逆时针旋转了 90°，得到新的 101M 和 102M 的 CGR，见图 4、图 5。

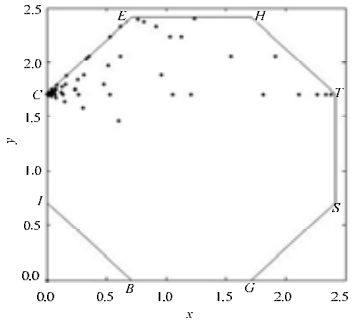

图 4　旋转后 101M 的 CGR 图

(下转第 237 页)