

# 基于散列辞典的蛋白质二级结构预测方法

南雨宏,陈 绮

(海南大学 信息科学技术学院,海南 海口 570228)

**摘 要:**提出一种易于修改的蛋白质二级结构预测算法。以蛋白质数据银行中 PDB 文本数据作为数据源,提取所有蛋白质氨基酸序列并以此建立样本数据库,然后针对  $\alpha$ -螺旋、 $\beta$ -折叠分别利用基于散列辞典的不同改进方法编程实现蛋白质二级结构序列片段预测,在预测过程中,随机抽取 68 421 个蛋白质中部分样本作为测试集,对未知序列根据建立的散列辞典中的片段使用正向最大匹配分词法进行切分对比。从实验结果来看,对未知序列片段预测的准确度达到了 83.9%,而且能够较好地体现片段之间的连接顺序。

**关键词:**蛋白质二级结构;序列片段;散列辞典; $\alpha$ -螺旋; $\beta$ -折叠

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2011)10-0168-03

## A Protein Secondary Structure Prediction Method Based on Hash-Dictionary

NAN Yu-hong, CHEN Qi

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

**Abstract:** This paper proposes a kind of easy to modify protein secondary structure prediction algorithm. Using PDB files from Protein Data Bank as a data source, extract all the protein amino acid sequences and build up a database, then for  $\alpha$ -helix,  $\beta$ -sheet, use different improved methods based on hash dictionary to implements the fragments prediction of protein's secondary structure. During the forecasting process, taking 68 421 samples as part of the protein in the test set. For unknown sequence according to the established fragments of hash dictionary use positive maximal matching points for segmentation lexical contrast. The results shows the prediction of segment reached 83.9% accuracy, but also to better reflect the sequence of amino acids connection.

**Key words:** protein secondary structure; sequence fragments; hash dictionaries;  $\alpha$ -helix;  $\beta$ -sheet

## 0 引言

蛋白质二级结构的预测是后基因组学的重要内容<sup>[1]</sup>,蛋白质的二级结构是指多肽链中主链原子的局部空间排布,是不涉及侧链部分的构象。蛋白质一级结构主要描述其化学组成结构,而二级结构则是在描述肽链中的一个局部区域的空间结构(Local Partial Structure)。它们是完整肽链构象(三级结构)的结构单元,是蛋白质复杂的空间构象的基础<sup>[2]</sup>。

目前,从直接测得、并从核酸序列翻译出的蛋白质序列的数据急剧增加,蛋白质数据库中已知的一级结构数目已经大大超过已测定空间结构的蛋白质数目<sup>[3]</sup>,所有蛋白质的空间结构都由实验来测得,如此巨大的工作量是不现实的,而蛋白质的二级结构又是联

系其一级结构和三维空间结构的纽带<sup>[4]</sup>,因此必须依靠计算机对海量数据的分析处理能力来对蛋白质的二级结构进行预测<sup>[5]</sup>,这是解决蛋白质空间结构预测问题关键的一步。

## 1 预测方法概述

蛋白质的二级结构中最有特点的是  $\alpha$ -螺旋和  $\beta$ -折叠<sup>[6]</sup>。利用这些具有特定结构的特殊片段,将待预测的未知氨基酸序列与蛋白质数据库中已知的氨基酸序列进行匹配,收集与其相匹配的蛋白质家族序列,通过设定相应的相似度阈值,能够找出最佳匹配。

在匹配检索的过程中,通过散列函数对已知蛋白质序列建立特定的哈希散列辞典,散列函数是一种将关键词转换成一个整数的函数,这个整数适合作为存储该关键词的数组的索引,如果散列函数值与关键词是一对一对的,我们就可以利用散列函数的值来访问相应的关键词。提前将哈希散列辞典加载到内存中,在检索过程中能够大大缩短匹配速度,最终达到高效

收稿日期:2011-03-26;修回日期:2011-06-10

基金项目:海南省自然科学基金资助项目(609003)

作者简介:南雨宏(1990-),男,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。

匹配、高准确度预测的目的<sup>[7]</sup>。

## 2 预测方法的实现

### 2.1 样本数据库的建立

为了保证原始数据的准确性,原始样本集文件选用蛋白质数据银行(Protein Data Bank)中的蛋白质分子文件作为原始数据抽取对象。截止 2010 年 11 月,该数据库共计收录了 68 421 个蛋白质分子,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<sup>[8]</sup>。

由于在二级结构预测研究中通常只做 3 种分类: H( $\alpha$ -螺旋)、E( $\beta$ -折叠)、C(无规则卷曲)<sup>[9]</sup>。针对二级结构手动建立数据库,能够省去自己编写程序读取文件内容来进行研究的繁琐步骤,减小出错概率。为存储、检索和处理数据带来许多方便。

在数据库中建立以下表项。

#### (1) 氨基酸注释表(ACID\_INFO)。

用于存放氨基酸 ID、单字母符号、三字母符号、氨基酸英文全称、氨基酸中文全称、氨基酸疏水值等。该表的字段还可以根据其它研究需要进行扩展(添加新字段),该表对于方便分析工作十分重要。

#### (2) 蛋白质一级结构表(SEQRES)。

包含蛋白质 ID、所在链 ID、氨基酸序列、氨基酸数目等字段。蛋白质一级结构即为实验测得的每个蛋白质的氨基酸序列,抽取每个蛋白质分子文件的 SEQRES 字段,并且参照氨基酸名称对照表,将所有三字母表示的氨基酸转换为单字母表示,存储转换后的所得序列。序列存储以蛋白质内的每条链为单位,这样便于从 PDB 文件中对特定的二级结构已知片段提取特定的序列。

#### (3) $\alpha$ -螺旋表和 $\beta$ -折叠表。

为了预测  $\alpha$ -螺旋和  $\beta$ -折叠,选取每个蛋白质分子文件的 HELIX 和 SHEET 字段,将所有的序列片段提取到数据库中,并且把每条记录都与一级结构表建立主外键关系。 $\alpha$ -螺旋表包含蛋白质 ID、 $\alpha$ -螺旋 ID 和  $\alpha$ -螺旋序列。 $\beta$ -折叠表包含蛋白质 ID、 $\beta$ -折叠 ID、连接点信息(氨基酸名称、连接位置、原子名称),每条肽链的氨基酸序列等信息。

### 2.2 已知片段哈希散列辞典的建立

蛋白质序列片段的查找,在很大程度上与文本语言的分词功能的实现非常相似。从 20 世纪末不断就有新的分词方法提出来<sup>[9]</sup>,而辞典的查询速度是决定分词算法效率的决定性因素,由于哈希结构在几种辞典查询结构中查询速度是最快的,文中辞典采用的哈希存储结构,以实现蛋白质序列片段的快速查找。

散列辞典的建立主要分为两步,第一步使用除余

法计算序列片段在散列表中的哈希地址;第二步使用二次探测法处理序列片段的地址冲突问题,使用该方法能够避免昂贵的乘法和除法运算。参考资文献[10]提出,当散列表大小为质数且至少一半为空时,新元素插入散列表产生碰撞的平均次数小于 2。

算法实现:

创建散列辞典,大小为序列片段总数 2 倍以上的质数。

```
int m = M;
```

```
String dict = new String[ M ];
```

(1) 使用除余法计算序列片段的哈希地址。

```
Address = Math. abs( word. hashCode() % m
```

(2) 使用二次探测法解决序列散列过程中的碰撞问题。

```
newAddress = address + 2 * colliCounter - 1;
```

```
if ( newAddress > m)
```

```
newAddress = newAddress - m;
```

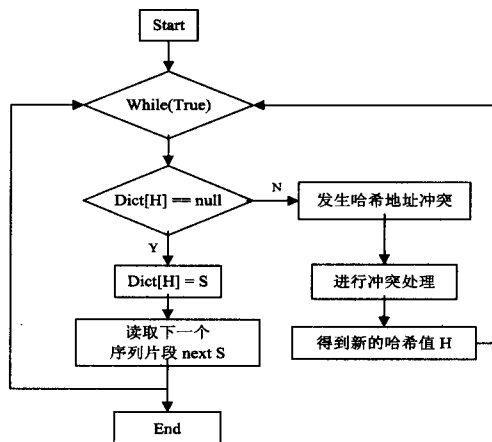


图 1 散列辞典建立流程

在进行二级结构预测时,需要将所有已测得序列加载到内存中,便于预测过程中的匹配、比较。使用 SELECT 语句选取数据库中所有不同的 HELIX 片段序列作为数据源,按照上述方法是对所有已知  $\alpha$ -螺旋提取到的序列片段建立散列辞典。

在  $\beta$ -折叠中,所有肽链序列片段的散列辞典建立过程基本相似,只需从数据库中抽取相应的字段即可。

### 2.3 使用 FMM 分词法对待测序列进行分词预测

FMM 分词法即为正向最大匹配分词法(Forward Maximum Matching Method)<sup>[11]</sup>。算法描述如下。

(1) 设自动分词辞典中最长氨基酸序列所含氨基酸数目为 I。

(2) 从待预测蛋白质序列中取第 0 到 I 的序列片段作为待匹配的片段(如果待切分片段长度小于 I,则 I 调整为待切分片段的总长),查找散列辞典,如果辞典中有这样的一个序列片段,则匹配成功,待匹配序列作为一个可能已知的序列片段被切分出来,转 6。

(3) 如果辞典中找不到这样一个序列片段, 则匹配失败。

(4) 待匹配序列中去掉最后一个氨基酸, I--。

(5) 重复步骤 2-4, 直到切分成功为止。

(6) I 重新赋值, 转 2, 直到切分出所有的片段为止 (通过对比计算, 得出当截取的序列片段长度大于等于 5 时才判定为有效序列片段。有效片段长度阈值高低与预测准确率成正比关系。所有长度小于规定阈值的片段视为无效片段, 有助于提高预测的准确度)。

分别建立  $\alpha$ -螺旋、 $\beta$ -折叠的散列辞典之后, 根据给定的未知氨基酸序列, 即可通过 FMM 分词法进行进一步预测。

$\alpha$ -螺旋的预测可以直接使用上述算法, 但是对于  $\beta$ -折叠, 由于每个片层的肽链之间存在关联, 要想完整预测出整个  $\beta$ -折叠部分, FMM 法需要进行以下改进。

当找到某个已经存在的肽链序列片段 (STRAND) 时, 回到数据库中检索包含该肽链的全部  $\beta$ -折叠, 由此可以得到多组不同的  $\beta$ -折叠结构。依次抽取每组  $\beta$ -折叠中的下一个肽链片段, 查看该片段是否在未知序列中出现, 若未知序列中存在该片段, 则继续查找下一个该组内的肽链片段, 直到该组的肽链片段在未知氨基酸序列中达到一定的阈值为止。若未知氨基酸序列中不存在这样的肽链片段, 则跳过该组  $\beta$ -折叠, 转向下一组继续搜索。

### 3 预测结果分析

目前评价预测结果的方法有很多种, 国际上通用以下几个评价指标。

(1) 整体准确率  $Q_3$ :

$$Q_3 = (P_a + P_b + P_c) / T \quad (1)$$

其中  $P_i (i \in \{a, b, c\})$  分别表示被正确预测出的三态 (H、E、C) 残基个数,  $T$  代表残基总数。

(2) 三态准确率  $Q_i$ :

$$Q_i = P_i / (P_i + U_i) \quad i \in \{H, E, C\} \quad (2)$$

其中  $P_i$  表示被正确预测为  $i$  态的残基的个数,  $U_i$  表示  $i$  态被预测为非  $i$  态的残基个数。  $Q_i$  可以用来衡量预测不足的情况。

(3) Motthew 系数:

$$C_i =$$

$$\frac{p_i n_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i) + (n_i + u_i) + (n_i + o_i)}} \quad i \in \{H, E, C\} \quad (3)$$

其中  $p_i$  表示  $i$  态被预测为  $i$  态,  $n_i$  表示非  $i$  态被预测为非  $i$  态,  $u_i$  表示  $i$  态被预测为非  $i$  态,  $o_i$  表示非  $i$  态被预测为  $i$  态。  $C_i = 0$  代表随机预测;  $C_i = 1$  代表完全预

测。

文中由于并未对无规则卷曲 (C) 的所有数据进行提取, 因此只采用方法 (2) 来评价  $\alpha$ -螺旋、 $\beta$ -折叠预测精度。

使用上述算法随机选取蛋白质数据中的部分蛋白质作为实验的测试数据集 (同时在数据库中删除这些蛋白质的相关数据)。该算法对测试集进行测试, 并且与测试集内蛋白质实际存在的  $\alpha$ -螺旋、 $\beta$ -折叠进行对比。

实验采用机器配置情况为: CPU Intel core 2 duo T5500 1.66GHz, 1GB 内存, 数据库中 Helix 片段记录数为 1 479 405, Beta 螺旋片层记录数为 1 574 666, 具体实验结果如表 1、表 2 所示。

表 1 建立散列辞典

	$\alpha$ -螺旋	$\beta$ -折叠
建立哈希表时间	1935 ms	2746ms
哈希表项数	371620	452101
最大冲突次数	1	0

表 2 部分预测结果统计

蛋白质名称	1o4l	1s73	1nod	2cpu	2eq5	平均值 (%)
序列长度 (单链)	166	294	496	423	228	
实际 $\alpha$ -螺旋个数	11	19	16	20	11	
检索所得 $\alpha$ -螺旋个数	13	17	12	22	10	
检索准确率 (%)	84.6	89.5	75	90.9	90.9	86.3
实际 $\beta$ -折叠个数	3	7	13	23	8	
检索所得 $\beta$ -折叠个数	6	7	15	12	11	
检索准确率 (%)	50	100	86.7	52.1	72.7	72.3
检索耗时 (ms)	31	100	181	132	61	
综合准确率 (%)	61.3	94.6	80.9	71.5	76.0	78.6

从总体实验结果来看, 预测平均精确度达到了 83.9%。其中  $\alpha$ -螺旋预测精度略高于  $\beta$ -折叠。这主要是由于  $\alpha$ -螺旋全部由单链组成, 且平均单链长度大于  $\beta$ -折叠, 因此字符串匹配精度更高。

### 4 结束语

张春霆院士提出利用氨基酸序列顺序来预测蛋白质结构类的算法, 将预测精度提到 90% 以上<sup>[12]</sup>。这说明基于序列的蛋白质结构预测方法非常值得进行进一步的讨论研究。

文中所实现的编程方法非常简单易懂, 研究者可以根据不同的研究需求对算法进行稍加修改, 即可针对自己所研究的特定内容建立数据库并进行预测。结果显示, 运用基于散列辞典的检索技术来匹配蛋白质二级结构序列的方法是高效合理的。

在得到检索结果之后, 有 2 条主要途径可以进行

(下转第 175 页)

不具有生物意义。

评价中应用 PKG 算法对训练数据集和测试数据集进行了实验,结合同样基于 EM 算法的 MEME 算法工具进行了对比实验,应用我们的评测方案在实验中搜集了大量实验参数,无论是在相同长度的 Motifs,还是在时间复杂度上都较 MEME 有所优化。通过运用信息量 IC 等指标对两种算法工具进行了定量检测,最后结合医学检测中常用的 ROC 曲线进行了特异度和敏感度的比较,进一步验证 PKG 算法的改进性。

结合生物实验中总结的经验方法对照通过算法获取的 Motifs 能够有效的提高预测算法的精确性,筛选后的数据集还需要生物实验来验证才能最终被确定。

文中的评价策略还存在一定的不足,例如本身在一个蛋白质家族的成员序列之间就存在一定的差异,因而通过随机选取的训练集和测试集序列就会有一定的偏差。此外受到实验中软硬件环境的制约,评价策略的效果也会受到一定的影响。因而,在后面的研究中还需要对训练集和测试集序列建立更好的数据模型,以提高通过文中评价策略筛选 Motifs 的精确性。

#### 参考文献:

- [1] 杜春鹏,朱云平,贺福初. 蛋白质家族模体 (motif) 的评价策略[J]. 北京生物医学工程,2005,24(2):97-102.
- [2] 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.
- [3] Timothy L B, Charles E. The value of prior knowledge in discovering Motifs with MEME[C]//Proceeding of the Third In-

ternational Conference on intelligent Systems for Molecular Biology. Menlo Park, California: [s. n.], 1995:21-29.

- [4] 王维彬,钟润添. 一种基于贪心 EM 算法学习 GMM 的聚类算法[J]. 计算机仿真,2007,24(2):65-68.
- [5] 张斐,徐利. 一种基于贪心 EM 的改进预测算法[J]. 计算机工程,2010,22(1):35-37.
- [6] Attwood T K, Croning M D R, Flower D R, et al. PRINT-S: the database formerly known as PRINTS[J]. Nucleic Acids Res, 2000,28:225-227.
- [7] 杜耀华,倪青山,王正志. 利用序列保守模体和局部构象信息预测转录因子结合位点[J]. 生命科学研究,2006,10(3):215-223.
- [8] Pavesi G, Mereghetti P, Zambelli F, et al. MoD Tools: regulatory Motif discovery in nucleotide sequences from co-regulated or homologous genes[J]. Nucleic Acids Res, 2006,34:566-570.
- [9] Attwood T K, Croning R D M, Flower D R, et al. PRINT-S: the database formerly known as PRINTS[J]. Nucleic Acids Res, 2000,28:225-227.
- [10] Grundy William N, 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.
- [11] Bailey T L, Elkan C. Unsupervised learning of multiple Motifs in biopolymers using expectation maximization[J]. Machine Learning, 1995,21:51-83.
- [12] Durbin R, Eddy S, Igogh A, et al. 生物序列分析,第三章:蛋白质和核酸的概率论模型[M]. 北京:清华大学出版社,2010.

(上接第170页)

比较分析:一是,比较未知蛋白序列与已知蛋白质序列的相似性;二是,查找未知蛋白质中是否包含于特定蛋白质家族或功能域有关的亚序列或保守区段。

#### 参考文献:

- [1] Eisenberg D, Marcotte E M, Xenarios I. Protein function in the post-genomic era[J]. Nature, 2000,405(6788):823-826.
- [2] Horton H R, Moran L A, Scrimgeour G. Principles of biochemistry[M]. 3rd ed. New Jersey: Pearson Education Inc, 2002.
- [3] 梁毅. 结构生物学[M]. 北京:科学出版社,2005.
- [4] Sundar H, Silver D, Gagvani N, et al. Skeleton based shape matching and retrieval[C]//Proceedings of International Conference on Shape Modeling and Applications. Seoul, Korea: [s. n.], 2003:130-139.
- [5] Szustakowski J D, Weng Z P. Protein structure alignment u-

sing a genetic algorithm[J]. Proteins, 2000,38:428-440.

- [6] 李晓琴,罗辽复. 由氨基酸序列预测蛋白质二级结构的进一步研究[J]. 内蒙古大学学报(自然科学版),1992(4):534-540.
- [7] 袁崇义. 离散数学及其应用[M]. 北京:清华大学出版社,2002.
- [8] 廖志华,谌容,陈敏,等. 生物学信息数据库简介[J]. 生物学教学,2006(1):61-62.
- [9] 张宁,张涛. 蛋白质二级结构预测样本集数据库的设计与实现[J]. 生物信息学,2006,4(4):163-166.
- [10] 李克清. 数据结构——C语言描述[M]. 武汉:华中科技大学出版社,2005.
- [11] 金澎,刘毅,王树梅. 汉语分词对中文搜索引擎检索性能的影响[J]. 情报学报,2006(1):21-24.
- [12] 张春霆. 蛋白质结构分类与结构类预测研究[J]. 中国科学基金,2000(5):44-45.