文章编号:2095-476X(2015)5/6-0085-03

基于伪氨基酸组成和多标记最近邻算法的抗菌肽功能类型预测

王晓, 杨鹏鹏, 王榕, 李辉

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:针对多数已有的计算方法无法同时预测抗菌肽的多种功能类型的问题,提出一种基于伪氨基酸组成和多标记最近邻算法的抗菌肽功能类型预测的系统方法:采用伪氨基酸组成抽取抗菌肽序列的特征向量,并且引入多标记最近邻算法作为预测引擎,同时预测抗菌肽的多种功能类型. 实验结果表明,本方法显著地提高了预测性能,为该领域的进一步研究提供了一个有用的工具.

关键词:抗菌肽;伪氨基酸组成;多标记分类;多标记最近邻算法

中图分类号:TP273;O811.4 **文献标志码**:A **DOI**:10.3969/j.issn.2095-476X.2015.5/6.017

Predicting functional types of antimicrobial peptides with pseudo amino acid composition and multi-label k-nearest neighbor algorithm

WANG Xiao, YANG Peng-peng, WANG Rong, LI Hui

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: In order to solve the problem that most of the existing computational methods can only predict one functional type of antibacterial peptides, a computational prediction method was developed for prediction of multiple functional types of antibacterial peptides based on the pseudo amino acid composition (PseAAC) and multi-label k-nearest neighbor(MLkNN) algorithm. It used the PseAAC to extract feature vector of antimicrobial peptide sequence, introduced the MLkNN algorithm as the prediction engine, and predicted a variety function type of antibacterial peptides simultaneously. Experimental results showed that the proposed method significantly improved the prediction performance, and it provided a useful tool for the further research in this field.

Key words: antimicrobial peptide; pseudo amino acid composition (PseAAC); multi-label classification; multi-label k-nearest neighbor(MLkNN) algorithm

0 引言

抗菌肽具有天然免疫特性,是传统抗生素药物的绝佳替代品,可以解决抗生素的耐药性问题[1]. 随着后基因组时代大量蛋白质序列的产生,已知是抗菌肽的序列与未知的蛋白质序列之间的差距越来越大. 用实验来确认哪些蛋白质序列是抗菌肽及搞清楚它们的功能类型,变得越来越不可行,迫切

收稿日期:2015-09-28

基金项目:国家自然科学基金项目(61402422);河南省教育厅科学技术研究重点项目(14A520063);郑州轻工业学院博士科研基金资助项目(2013BSJJ082)

作者简介:王晓(1982—),男,河南省卫辉市人,郑州轻工业学院讲师,博士,主要研究方向为机器学习与生物信息学.

需要开发基于序列的计算预测工具,以便快速而准确地识别抗菌肽和它们的功能类型.

从 APD 数据库可以看出,有大量的抗菌肽不止有 1 种功能,而是执行多种生物功能. 因此,不仅需要识别它们的功能类型,而且需要同时识别出它们的多种功能类型. 计算预测抗菌肽的多种功能类型对于基础研究和药物开发具有十分重要的意义. 目前,已经有一些计算预测工具出现[2-6],从而推动了该领域的快速发展. 但是,它们无法识别出抗菌肽的具体功能类型.

本文主要关注于识别抗菌肽的多种功能类型. 鉴于伪氨基酸组成(PseAAC)在预测蛋白质的各种属性中取得了良好的性能,本文拟采用 PseAAC 来提取蛋白质的特征,并且采用多标记最近邻算法(MLkNN)作为预测引擎,构建一个抗菌肽的多功能预测器,以期准确地预测抗菌肽的多种功能类型并显著地提高预测性能.

1 抗菌肽多功能预测器的设计

1.1 数据集

本文采用文献[7]所构建的数据集作为基准数据集,它包含抗菌肽和非抗菌肽子集,但由于本文只关注抗菌肽的多功能类型识别,因而只使用抗菌肽子集,用符号 Data_AMP 表示. Data_AMP 数据集中的肽序列从 ADP 数据库中获取. ADP 数据库中的抗菌肽序列原本共有 10 种功能类型,由于 Antiparasital, Anti-protist, AMPs with chemotactic activity, Insecticidal 和 Spermicidal 包含非常少的抗菌肽序列(不足 50 个),不具有统计显著性,故从 Data_AMP 数据集中删除,暂且不考虑它们,只保留 Antibacterial, Anticancer/tumor, Antifungal, Anti-HIV 和 Antiviral 这 5 种抗菌肽序列. 为了减少同源偏置和序列冗余的影响,采用 CD-HIT 程序过滤掉那些序列相似度 ≥40% 的肽序列[7]. 同时,为了考虑去除冗余和数据集大小之间的平衡,少于 150 个肽序列的功能类型子集不进行过滤操作,保留该功能类型的全部肽序列. 最后,Data_AMP 数据集共包含 878 个抗菌肽序列,其中,454 个属于 1 个功能类型,296 个属于 2 个功能类型,85 个属于 3 个功能类型,30 个属于 4 个功能类型,13 个属于 5 个功能类型. 表 1 给出每种功能类型拥有的抗菌肽数量.

表 1 数据集 Data_AMP 的统计信息

编号	功能类型	肽序列数量
1	Antibacterial	770
2	Anticancer/tumor	140
3	Antifungal	366
4	Anti-HIV	86
5	Antiviral	124

1.2 特征提取

由于蛋白质序列中的氨基酸残基顺序包含重要的信息,因此 K. C. Chou[8] 于 2001 年提出 PseAAC 的概念来代替传统的氨基酸组成(AAC). 至今,它已经广泛地渗透到蛋白质属性预测的多个领域,例如,蛋白质的超二级结构(supersecondary structure)的预测,细菌毒性蛋白质(bacterial virulent proteins)的识别,蛋白质亚细胞定位预测,蛋白酶家族和子家族类别(enzyme family and sub-family classes)的预测,等等. PseAAC 向量化蛋白质为 $(20+\xi\cdot\lambda)$ 维的特征向量,其中,前 20 维是传统的 AAC,而后 $\xi\cdot\lambda$ 维表示蛋白质氨基酸序列间的序列顺序信息. PseAAC 向量中的特征维数由两个重要的参数控制:选出的氨基酸指数数量 ξ 和蛋白质序列中的最大相关层数 λ. 需要注意的是 λ 必须小于训练集中最短蛋白质序列的长度,在 $\lambda=0$ 的极端情况下,PseAAC 退化为传统的 AAC.

1.3 预测引擎

抗菌肽多功能识别问题可以看作是一个机器学习领域中的多标记分类任务. 本文引入 MLkNN 算法作为抗菌肽多功能识别的预测引擎. MLkNN 是一个基于 kNN 算法的高效的多标记分类算法. 基于测试样本多个近邻的标记集合的统计信息,MLkNN 利用最大化后验规则确定测试样本的标记集合.

给定一抗菌肽数据集 X,其中包含的所有功能类型由集合 $Y=\{t_1,t_2,\cdots,t_5\}$ 表示,继而构成一训练集 $\{(p_1,Y_1),(p_2,Y_2),\cdots,(p_N,Y_N)\}$,其中 $Y_i(i=1,2,\cdots,N)\subseteq Y$ 是肽序列 $p_i(i=1,2,\cdots,N)\in X$ 的功能类型集合. 对一未知功能的肽序列 p,要想知道它的功能类型,首先要从数据集中获取它的 k 个最近邻,由 $N(p)$ 表示. 基于 $N(p)$ 中肽序列的功能集合,定义如下的成员计数向量:

$$C_p(t)=\sum_{n\in N(p)}y_n(t)\qquad t\in Y$$

其中,$C_p(t)$ 表示未知肽序列 p 的所有近邻中属于功能类型 t 的近邻个数;y_n 表示近邻肽序列 n 对应的功

能类型向量,当 $t \in Y_n$ 时 $y_n(t)$ 取值为 1,否则 $y_n(t)$ 取值为 0. 进而设 H_1^t 表示未知肽序列 p 具有功能类型 t 这一事件,而 H_0^t 代表未知肽序列 p 不具有功能类型 t 这一事件. 再设 $E_j^t(j \in \{0,1,\cdots,k\})$ 表示未知肽序列 p 的 k 个近邻中刚好有 j 个邻居肽具有功能类型 t 这一事件. 基于上面的设定,根据成员计数向量 $C_p(t)$ 提供的信息,可以通过最大化后验概率的准则确定未知肽序列 p 的功能类型向量:

$$y_p(t) = \arg\max_{b \in \{0,1\}} P(H_b^t \mid E_{C_p(t)}^t) \qquad t \in Y \quad ①$$

基于贝叶斯规则,式 ① 可重写为

$$y_p(t) = \arg\max_{b \in \{0,1\}} \frac{P(H_b^t)P(E_{C_p(t)}^t \mid H_b^t)}{P(E_{C_p(t)}^t)} = \arg\max_{b \in \{0,1\}} P(H_b^t)P(E_{C_p(t)}^t \mid H_b^t)$$

其中,先验概率 $P(H_b^t)(t \in Y, b \in \{0,1\})$ 和后验概率 $P(E_j^t \mid H_b^t)(j \in \{0,1,\cdots,k\})$ 均可以通过频率计数直接估计得到.

2　实验结果与讨论

本文采用 jackknife 测试评估所提方法的性能,并且采用 mlACC, mlPRE, mlREC, mlF_1 和 ACC 这 5 种性能评价指标. 以上 5 种指标是多标记生物数据属性识别中常用的性能评价指标,详细计算方法可以参照文献[9].

通过遍历所有的 PseAAC 的参数组合,选取 hydrophobicity 和 hydrophilicity 这两种氨基酸指数($\xi = 2$)用来计算蛋白质氨基酸序列间的相关因子,并且设置 $\lambda = 3$,由此可得,PseAAC 特征的维数为 $20 + 2 \times 3 = 26$. 选取该特征参数组合,再设置 MLkNN 算法的近邻数 $k = 5$,预测结果表现出最好性能.

本文所提方法与 iAMP-2L 预测器[7]在抗菌肽数据集 Data_AMP 上的预测性能比较结果见表 2. iAMP-2L 是目前唯一能够预测抗菌肽的多功能类型的预测器,因此本文所提方法仅与 iAMP-2L 进行比较是合理且充分的. 从表 2 可以看出,本文所提方法在几乎所有性能评价指标上都超过了 iAMP-2L 预测器,特别是绝对精度 ACC 达到了 46% 以上,显著超过了 iAMP-2L 方法. 由于 ACC 要求非常严格,必须完全正确地预测出测试肽序列的所有功能类型才算是预测正确,任何过预测或欠预测都被认为预测错误,因而可知,本文所提方法显著地改进了多功能抗菌肽的识别率.

表 2　本文方法与 iAMP-2L 的识别率比较 %

评价指标	预测方法	
	本文方法	iAMP-2L
mlACC	68.85	66.87
mlPRE	82.98	83.31
mlREC	77.81	75.70
mlF_1	80.31	79.32
ACC	46.76	43.05

3　结语

本文采用伪氨基酸组成抽取抗菌肽序列的特征向量,并且引入多标记最近邻算法作为预测引擎,开发了一个计算预测系统来预测抗菌肽的多种功能类型,实验结果表明,本文所提方法较 iAMP-2L 预测器显著地提高了预测性能,对抗生素替代药物的研制具有极其重要的意义. 为了更好地服务实验生物学,进一步的工作计划是把本文所提方法开发成在线预测服务网站.

参考文献:

[1] Riadh H, Ismail F. Current trends in antimicrobial agent research: chemo-and bioinformatics approaches[J]. Drug Discovery Today, 2010, 15(13/14): 540.

[2] Fjell C D, Hancock R E, Cherkasov A. AMPer: a database and an automated discovery tool for antimicrobial peptides[J]. Bioinformatics, 2007, 23(9): 1148.

[3] Lata S, Sharma B K, Raghava G. Analysis and prediction of antibacterial peptides[J]. BMC Bioinformatics, 2007, 8: 263.

[4] Lata S, Mishra N, Raghava G. AntiBP2: improved version of antibacterial peptide prediction[J]. BMC Bioinformatics, 2010, 11(S1): S19.

[5] Wang P, Hu L L, Liu G Y, et al. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods[J]. PLoS ONE, 2011, 6(4): e18476.

[6] Khosravian M, Faramarzi F K, Beigi M M, et al. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods[J]. Protein and Peptide Letters, 2013, 20(2): 180.

[7] Xiao X, Wang P, Lin W Z, et al. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types[J]. Analytical Biochemistry, 2013, 436: 168.

[8] Chou K C. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Proteins: Structure, Function, and Bioinformatics, 2001, 43(3): 246.

[9] Li G Z, Wang X, Hu X, et al. Multilabel learning for protein subcellular location prediction[J]. IEEE Transactions on NanoBioscience, 2012, 11(3): 237.