学校编码： 10384　　分类号＿＿＿＿＿密级＿＿＿＿＿
学　　号： 24320130154085　　UDC＿＿＿＿＿

厦门大学

博 士 学 位 论 文

基于多信息融合的生物大分子序列预测方法研究

Research of Biological Macromolecule Sequence Prediction Method Based on Multi-Information Fusion

魏乐义

指导教师姓名： 廖 明 宏 教 授

专 业 名 称： 计算机科学与技术

论文提交日期： 2016 年 4 月

论文答辩日期： 2016 年 5 月

学位授予日期： 2016 年 月

指 导 教 师：＿＿＿＿＿

答辩委员会主席：＿＿＿＿＿

2016 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1. 经厦门大学保密委员会审查核定的保密学位论文，于年　　月　　日解密，解密后适用上述授权。

（　√　）2. 不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

随着测序技术的发展，生物大分子序列数量快速积累，迫切需要了解序列所蕴含的重要生命信息。近年来，生物大分子序列的结构与功能研究已经成为生物信息学领域研究的热点问题。目前，基于生物大分子序列和机器学习模型的方法是生物信息学领域中预测序列结构和功能的重要研究手段。本文从如何构建有效的序列向量化方法、分类算法、以及高质量数据集角度出发，对生物大分子序列预测的几个具体问题进行了深入研究，包括蛋白质结构类预测、蛋白质折叠模式类预测、细胞因子与受体相互作用预测、细胞穿透肽预测、以及 microRNA 前体预测。本文的研究内容包括以下几个方面：

第一，针对蛋白质结构类预测问题，目前现有预测方法普遍存在的问题是特征中包含信息单一导致特征的表达能力较低。为了克服这一问题，本文提出了基于序列与结构特征的蛋白质结构类预测方法 RF_PSCP。在该方法中，首先利用了基于多信息融合的特征提取方法，将蛋白质的初级序列信息、二级结构信息和序列结构信息融合到特征向量中，从不同角度更加全面刻画不同结构类间蛋白质序列的差异性；然后，将特征向量输入随机森林进行结构类预测。在 10 折交叉验证中，本文提出的方法 RF_PSCP 在多个基准数据集上的预测准确率上均显著优于现有的方法，表明了方法的有效性。此外，在多个更新数据集上稳定的预测效果表明了方法良好的鲁棒性。

第二，在蛋白质折叠模式类预测领域中，目前基于机器学习的预测方法实际的预测效果并不理想。为了进一步提高方法的预测性能，本文提出了基于集成学习的蛋白质折叠模式类预测方法 PFPA，从序列向量化方法与分类算法两个方面做了相应改进，从而提升了预测效果。在序列向量化方面，利用了两种新的向量化方法：基于 PSI-BLAST 和基于 PSI-PRED 的特征算法，使得特征向量充分包含初级序列信息、进化信息、以及局部和全局二级结构信息。在分类算法方面，本文采用了平均概率的集成策略将五种不同的基分类器结合，从而形成集成分类器对蛋白质序列进行折叠模式类预测。与现有方法在基准数据集上的比较，表明了本文提出的方法的优越性。

第三，针对细胞因子与受体相互作用预测问题，本文从蛋白质相互作用具有局

部性特点出发，提出了基于局部进化特征的细胞因子与受体相互作用预测方法 CRI-Pred。在该方法中，首次引入了蛋白质序列局部信息的概念。为了提取局部信息，利用平均分割的方法将位置特异性得分矩阵分成多个子矩阵，将两个进化特征模型（*Pse-PSSM* 和 *AAC-PSSM-AC*）应用于子矩阵中将蛋白质序列向量化，从而使得特征向量融合了蛋白质序列的局部保守信息、进化信息、以及序列的顺序信息。在分类器方面，本文采用随机森林作为分类器进行预测。实验结果表明，本文提出的方法在整体预测准确率指标上比现有预测方法高 5.1%。

第四，在细胞穿透肽预测领域中，本文针对现有方法的一些不足做了相应改进，从而提出了基于随机森林的细胞穿透肽预测方法 SkipCPP-Pred。在该方法中，本文提出了自适应 *k-skip-n-gram* 特征向量化方法，在 *n-gram* 模型基础上增加更多的距离和序列氨基酸间相关性，从而一定程度上解决了传统 *n-gram* 方法造成的特征空间稀疏问题。其次，在数据集构建方面，本文重新构建了一个新的数据集：降低样本的冗余，增加数据集样本量，提升正反例样本相似性分布，从而克服基于现有数据集构建的预测方法出现的“过预测”问题。为了验证方法的有效性，本文比较了 SkipCPP-Pred 与现有方法的预测效果。实验结果表明，SkipCPP-Pred 比现有方法能够更加准确预测序列是否具有细胞穿透功能。

第五，在 microRNA 前体预测领域中，目前现有的预测方法普遍存在训练集中反例样本不具有代表性，导致预测方法泛化能力差的问题。本文提出了基于高质量反例的人类 microRNA 前体预测方法 miRNAPre。该方法的研究重点是从反例选择的角度出发，提出了高质量反例挖掘方法，通过反复迭代的深度挖掘，从而克服现有反例样本过度依赖参数选择导致与正例样本差异性较大的问题。在预测方法的构建方面，基于多信息融合的方法将序列向量化为包含了多种不同信息的特征，以支持向量机分类器作为特征向量输入进行预测。与现有方法在多个的独立测试集上的比较结果显示 miRNAPre 均取得了更高的敏感性和特异性，实验表明了 miRNAPre 能够为生物实验提供可靠的 microRNA 前体候选预测服务。

关键词：生物信息学；机器学习；生物大分子序列预测；多信息融合

Abstract

With the development of next-generation sequencing techniques, the number of biological sequences is in the explosive growth. A majority of these sequences are not characterized. Facing with such numerous biological sequences, traditional experimental methods are time-consuming and cost-consuming. Thus, it is an urgent demand to develop computational methods to mine the important information embedded in biological sequences, such as structural and functional information. In this dissertation, we mainly focus on the following five aspects：

Firstly, considering the problems of existing features in characterizing only single-view information for depicting protein sequences, a novel feature representation method is proposed by integrating multiple types of information from the following three views: primary protein sequence, secondary structure, and sequence-structure. Experimental results show that the features based on multi-information fusion are more effective than the features using single-view information for protein structural class prediction. Based on the proposed features, we develop a novel prediction method using Random Forest classifier, namely RF_PSCP. 10-fold cross validation results on benchmark datasets show that the proposed predictor is more accurate than the state-of-the-art methods for protein structural class prediction. Moreover, the proposed predictor also provides promising prediction results for predicting those newly discovered proteins in updated datasets. This demonstrates the proposed predictor has the potential to be a useful tool for researchers working in this area.

Secondly, within the field of protein fold prediction, numerous taxonomic methods have been developed, and much progress has been made in recent years. Unfortunately, the overall prediction accuracies of existing methods are not satisfactory. To improve the prediction performance, we propose a novel ensemble learning protein fold prediction method, namely PFPA. In PFPA, we make some improvements in the following two aspects: the feature representation method and classification algorithm. For feature representation, two protein vectorization algorithms are presented. They are

based on PSI-BLAST and PSI-PRED, respectively. The PSI-BLAST-based algorithm uses the evolutionary information embedded in PSI-BLAST profiles to transform proteins sequences into consensus sequences. Then it employs the traditional n-gram model to extract sequence-based features from the consensus sequences containing rich evolutionary information. The PSI-PRED-based algorithm uses the secondary structure information from secondary structure sequences and local and global structural evolutionary information from PSI-BLAST profiles to vectorize proteins. For classification algorithm, a novel ensemble classifier is presented by using an averaging probability strategy to combine five basic classifiers. Experimental results on multiple datasets demonstrate the superiority of the proposed predictor as compared with the state-of-the-art predictors.

Thirdly, in the field of cytokine-receptor interaction prediction, a novel protein vectorization algorithm is presented based on the local evolutionary conservation of cytokine and receptor interactions. To capture the local information, the PSI-BLAST profile, also known as Position-Specific Scoring Matrix (PSSM), is fragmented into several sub-PSSMs by rows. Then, two protein vectorization models, *Pse-PSSM* and *AAC-PSSM-AC*, are employed to extract local features from each sub-PSSM. By combining the features from all the sub-PSSMs, a feature vector is yielded sufficiently containing local information, evolutionary information and sequence-order information. Furthermore, a novel cytokine-receptor interaction predictor, namely CRI-Pred, is presented by integrating the resulting feature vector and the random forest classifier. Experimental results show that the proposed predictor is 5.1% more accurate than existing predictors.

Fourthly, considering the feature sparse problems of the n-gram features in cell-penetrating peptide prediction, an improved feature representation algorithm is proposed by using an adaptive k-skip-n-gram model. The adaptive k-skip-n-gram model considers not only contiguous amino acids as the traditional n-gram model did, but also incorporates the extra distance information of the amino acids. On the other hand, we construct a new high-quality dataset by reducing the data redundancy and enhancing the similarity between the positive and negative classes. Using this dataset and the adaptive

k-skip-n-gram features, we train a novel computational predictor based on random forest classifier, namely SkipCPP-Pred. Jackknife results show that the proposed predictor SkipCPP-Pred is more accurate to predict whether sequences are cell penetrating or not, as compared with existing predictors.

Finally, within the field of microRNA precursor prediction, the problem lying in existing prediction methods is that negative samples for model training are not sufficiently representative. To address this problem, we present a novel negative sample selection technique by using a multi-level mining strategy, and successfully collect high-quality negative samples. Two recent classifiers rebuilt with the collected negative set achieve improved performance, which demonstrate that the negative samples are important for prediction models. Based on the high-quality negative samples, we propose a Support Vector Machine (SVM)-based predictor, namely miRNAPre. In independent test, the proposed predictor achieves better performance in terms of sensitivity and specificity as compared with existing methods. This indicates that our method is capable to provide promising prediction in this field.

Keywords: Bioinformatics; Machine Learning; Biological Macromolecule Sequence Prediction; Multi-Information Fusion

目 录

Contents