学校编码：10384　　分类号密级

学号：23220141153362　　UDC

厦门大学

硕士学位论文

基于特征构造与特征选择的 DNA 结合蛋白识别

Identifying DNA-binding Proteins Based on Feature Construction andFeature Selection

林杨

指导教师姓名：吉国力教授

专业名称：控制工程

论文提交日期：2017 年 4 月

论文答辩时间：2017 年 5 月

学位授予日期：2017 年 5 月

答辩委员会主席：

评阅人：

2017 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）：

2017 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）：

2017 年　月　日

摘要

DNA 结合蛋白在分子生物学中扮演重要角色，其影响基因表达调控、DNA 复制等活动。然而以生物实验的方式识别 DNA 结合蛋白，耗时且昂贵。数据分析的方式已成为识别 DNA 结合蛋白的重要途径。为提高 DNA 结合蛋白的识别性能，针对蛋白质序列数据特点，研究 DNA 结合蛋白识别问题中的特征工程，包括特征构造与特征选择。该特征工程能有效提高 DNA 结合蛋白的识别性能，为识别 DNA 结合蛋白提供了一种简便、高效的方法。论文主要包括以下几个方面。

1）特征构造：针对 DNA 结合蛋白序列之间低相似性、独立性、序列长度不同等特点，本文从物理化学特性、混沌游戏表示、分形维数、位置特异性得分矩阵和频谱分析五种不同角度的特征构造方法来描述 DNA 结合蛋白，并对其进行组合，构造生成高维特征。

2）特征选择：针对所生成的高维特征，为减少冗余特征，降低特征维度，使用多个经典的特征选择方法对其进行降维，包括基于 SVM 和 PLS 的递归特征消除法和基于分类间隔的 ReliefF 算法。实验并分析了 SVM-RFE 优于其他特征选择方法的原因。

3）识别分析：采用 SVM 分类器实现 DNA 结合蛋白的识别，使用三个独立的公共数据集对所提方法进行验证，通过 30 次十折交叉验证评估算法的性能。从准确度、特异性、敏感度、马修斯相关系数四个性能指标对分类器识别性能进行评估，并从离散程度、特征分布等角度分析实验结果。

通过对不同的特征构造和特征选择的分析比较，实验结果表明，本文组合的多种特征构造方法，结合 SVM-RFE 特征选择方法能更加有效地提升 DNA 结合蛋白的识别性能。通过对所选特征子集的分布来看，蛋白质多种物化特征和 PSSM 矩阵对 DNA 结合蛋白的识别起着重要作用。与 DNA-Prot 和混合分形方法进行对比，也表明所提方法识别效果更优。同时，所提方法不仅适合于 DNA 结合蛋白分析的特征工程（特征构造和特征选择）中，同样也适用于其他蛋白

质序列的分析，包括蛋白质结构类预测及其他蛋白的识别。

关键词：DNA 结合蛋白；特征构造；特征选择；识别性能

Abstract

DNA-binding proteins play a vital important role in molecular biology, it has an impact on genes regulation, DNA replication and so on. However, DNA-binding proteins identified by experimental techniques, which are time-consuming and expensive.The method of data analysis has been becoming animportant way to identify DNA-binding proteins.In order to improving the identificationperformance, the feature engineering of DNA-binding proteins identification problem, including feature construction and feature selection, was studied according to its data characteristics, the feature engineering can effectively improve the DNA-binding proteins identification performanceand provide a simple and efficient method for identifying DNA-binding proteins. The main work of this paper is as followed.

1) Feature construction: aiming at the characteristics of low similarity, independence and different sequence length between DNA-binding proteins,we describe DNA-binding proteins from five aspects: physicochemical properties, chaos game representation, fractal dimension, position specificity score matrix and spectrum analysis to construct high dimensional features.

2) Feature selection: In order to reduce the redundant features and feature dimension, we use several classical feature selection method includingrecursive feature elimination method based on SVM and PLS, ReliefF algorithm based on classification interval to reduce its dimension. At the same time, the reason why SVM-RFE is superior to other feature selection methods is also analyzed.

3)Identification analysis: We adopt SVM to realize the identification of DNA-binding proteins.The proposed method is validated through three independent public datasets, and the algorithm’s veracity is tested by 30times 10-fold cross validation. Furthermore, the performance of the classifier is evaluated from the four performance indexes of prediction, which are accuracy, specificity, sensitivity and

Matthews correlation coefficient.And the experimental results are analyzed from three different perspectives, which are performance index, discrete degree and feature distribution.

Through the analysis and comparison of different feature construction and feature selection, the results show that the combination of multi-class feature construction and SVM-RFE feature selection method can effectively improve the identification accuracy of DNA-binding proteins.By analysis the distribution of selected feature subsets, various physicochemical characteristics and PSSMmatrix play an important role in the identification of DNA-binding proteins. Compared with DNA-Prot and hybrid fractal features, it also show that the proposed method is superior to them. In addition, the proposed method is not only suitable for the feature engineering of DNA-binding proteins, but also for the analysis of protein structure prediction and other proteins identification.

Keywords: DNA-binding proteins; feature construction; feature selection; identification performance

目录

CONTENTS