同济大学
TONGJI UNIVERSITY

博士学位论文

面向基因组分析的数据挖掘算法研究

国家自然科学基金项目(No. 61173118)资助

姓　　名：甘杨兰

学　　号：0710080071

所在院系：电子与信息工程学院

学科门类：工学

学科专业：计算机应用技术

指导教师：关佶红 教授

副指导教师：章伟雄 教授 (美国华盛顿大学)

二〇一一年十二月

同濟大學
TONGJI UNIVERSITY

A dissertation submitted to

Tongji University in conformity with the requirements for

the degree of Doctor of Philosophy

Data Mining Algorithms for Genome Analysis

（Supported by the National Natural Science Foundation of China）

Candidate: Yanglan Gan

Student Number: 0710080071

School: College of Electronics and Information Engineering

Discipline: Engineering

Major: Computer Application Technology

Supervisor: Prof. Jihong Guan

Co-Advisor: Prof. Weixiong Zhang (Wash. Univ.)

December, 2011

面向基因组分析的数据挖掘算法研究

甘杨兰

同济大学

学位论文版权使用授权书

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年　　月　　日

同济大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

摘要

随着人类基因组计划的完成和基因组测序技术的日趋成熟，基因组的研究对象不再局限于少数几种模式生物，人们获得了海量的基因组数据，包括序列数据和基因表达数据，巨大的数据积累往往蕴含着突破性发现的可能。我们正处在一个从积累数据向解释数据转变的时代，这为数据挖掘提供了新的研究课题，而生物系统本质上的模型复杂性及在分子层面上完备的生命组织理论的缺乏，给面向基因组分析的的数据挖掘研究带来新的挑战。本文研究面向基因组分析的数据挖掘新方法，主要工作与创新点包括以下四个方面：

（1）研究了基因表达数据的聚类问题，提出了一种面向基因表达数据分析的基于模式树的子空间聚类方法。传统基于距离相似度的聚类算法不能有效地发现具有相似表达的基因，而基于模式相似度的聚类算法能发现距离很远但具有相似模式的簇。但现有基于模式的方法在高维大规模数据库中效率不高，并且对所发现簇的质量缺乏评价标准。本文提出了一种能解决这些问题的基于模式相似度的快速子空间聚类算法，通过引入一种用于频繁模式挖掘的结构---模式树来确定簇的子空间，只需扫描一遍数据库，大大提高了聚类的效率，并给出一种通用的簇质量评价方法，使聚类的结果更有意义，用户能根据不同的需要灵活设置合适的质量控制参数。在基因表达数据上的聚类分析结果表明，本文提出的新算法能有效发现具有相似表达模式的基因，同一个簇中基因具有相似的功能，比已有方法有更好的性能和效率。

（2）进行了基因组非编码区的功能分析，提出了一种基于模式的最近邻搜索算法识别启动子。核心启动子的识别是理解基因非编码区功能和基因转录调控的关键。由于启动子序列的结构复杂，现有启动子识别方法对 CpG 岛不相关的启动子预测准确率比对 CpG 岛相关启动子的预测准确率要低，导致在整个基因组上启动子预测准确率不高。本文首先从新的结构角度，系统分析了这两类启动子的异同点，发现含 CpG 岛相关和 CpG 岛不相关启动子的结构特征值差异很大，但是具有非常相似的结构模式，并且这种模式和非启动子序列的模式不同。基于这一发现，我们提出一个能预测 CpG 岛相关和 CpG 岛不相关启动子的统一的方法，即基于模式最近邻搜索的启动子预测算法(简称为 PNNP)。实验表明本文提出的启动子预测方法能有效的发现启动子的结构特征模式，提高了整个基因组上的启动子预测准确率，特别是对 CpG 岛不相关启动子的预测。

（3）研究了基因组序列的特征，提出了一个特征选择框架用于选择和启动子序列相关的结构特征。首先通过比较启动子和非启动子序列的结构特征，发现启动子区域和非启动子序列具有不同的结构特征和结构特征谱模式，并且各种特征的之间的对比不一样，表明这些结构特征具有不同的启动子和非启动子区分能力。本文采用 4 种基于不同特征重要性衡量标准的 Filter 方法(信息增益、CHI、CFS 和 ReliefF)和两种基于不同分类器的 Wrapper(SVM 和 KNN)方法，选择和启动子相关的特征，解决了启动子识别过程中的高维和过拟合问题。和现有启动子的预测方法比较结果表明，对结构特征的属性选择能大幅提高启动子预测水平。

（4）研究了细胞中基因组的组织结构，提出了基于结构特征的核小体定位模型。首先在酵母全基因组上对与 DNA 序列的弯曲度，柔韧性以及能量相关的 12 个结构特征进行全面分析，系统研究这些特征对 DNA 序列的核小体形成能力、染色体组织和基因表达影响。通过计算基因组上核小体分布和结构特征谱的皮尔逊相关性，我们识别出两类典型的结构特征：一类是促进 DNA 序列上核小体形成的特征，这些特征是核小体密集的着丝粒区域的明显特征；另一类特征抑制核小体形成，阻止核小体在启动子区域的形成，并且这些特征的明显程度和基因表达丰度相关。根据结构谱和核小体分布的相关性，我们训练出集成了这 12 个结构特征的线性模型来预测核小分布。实验结果表明，本文提出的基于结构特征的模型能准确预测基因组上的核小体分布。进一步通过对序列结构特征谱的峰值定位，我们提出了一个新的算法 DLaNe 用于定位核小体的具体位置。与现有算法的比较结果表明 DLaNe 具有更好的准确性，所采用的结构特征具有很好的核小体定位能力。

关键词：基因组分析，基因表达聚类，启动子预测，核小体定位

ABSTRACT

With the completion of the Human Genome Project and the rapid development of genome sequencing technology, the objects of genome analysis are no longer limited to a few model organisms. Actually, we have obtained vast amounts of biological data, including genomic sequence data and gene expression data. Therefore, we are at the turning point from the accumulation of data to the detailed explanation to data, which provides many opportunities in data mining research field. However, due to inherent model complexity of biological systems and lack of complete organization theory of life at the molecular level, it incurs lots of data mining research challenges oriented to genome analysis. In this thesis, we conduct research on genome analysis and propose several novel data mining models and algorithms for use. The main work and contributions are summarized in the following four aspects:

First, we examined clustering issue on gene expression data, and proposed a novel pattern-based approach to subspace clustering oriented to the analysis of gene expression data. Traditional clustering models based on distance similarity are not always effective in capturing correlation among data objects, while pattern-based clustering algorithms can do well in identifying correlation hidden among data objects with long distance. However, the state-of-the-art pattern-based clustering methods are inefficient and provide no standard metric to measure the quality of clusters found. This paper proposed a new pattern-based subspace clustering method, which can tackle the problems mentioned above. Observing the analogy between mining frequent itemsets and discovering subspace clusters, we apply pattern tree, a structure used in frequent itemsets mining, to determine the target subspaces by scanning the database only once, which can be done efficiently in large datasets. Furthermore, we also introduce a general clustering quality evaluation model to guide the identification of meaningful clusters. The proposed new method enables the users to set flexibly proper quality-control parameters to meet different needs. Experimental results on synthetic and real datasets showed that our proposed method can outperform the existing methods both in its efficiency and effectiveness.

Second, we analyzed the functionality of the genome noncoding region, and proposed a novel Pattern-based Nearest Neighbor search algorithm for Promoter

(PNNP). Core promoter identification is a key clue in understanding gene noncoding region functionality and gene transcription regulation. However, due to the diverse nature and complex structures of promoter sequences, the accuracy of existing prediction approaches for non-CpG island (simply as CGI)-related promoters is not as high as that for CGI-related promoters. It can easily lead to low prediction accuracy genome-wide promoter. We first systematically analyze similarities and differences between these two types of promoters (CGI- and non-CGI-related) from a new structural perspective, and find that the structural feature values significantly differ with each other between CGI-related and non-CGI-related promoters, while they share structural patterns with high similarities. Based on above investigations, we devised and proposed a unified framework, called PNNP (Pattern-based Nearest Neighbor search for Promoter), to predict both CGI- and non-CGI-related promoters based on their structural features. The experimental results demonstrated that the proposed PNNP approach is effective in capturing the structural patterns of promoters, and can significantly improve genome-wide prediction performance for promoters, especially non-CGI-related promoter prediction.

Third, we examined the characteristics of genome sequences, and proposed a feature selection framework that is employed to select structural features correlative with promoter sequences. Based on the comparison on the structural features of CGI- and non-CGI-related promoters, we observed that there are different structural features and structural profiles between promoter regions and non promoter sequences. Therefore, these structural features have different capability of promoters and non promoter identification. We applied the Filter method with four different kinds of feature metric standards, including information gain, CHI, CFS and ReliefF, and Wrapper method based on two different classifiers (Support Vector Machine, SVM and K-Nearest Neighbors, KNN) to select those structural features correlative with promoter identification. Comparing to existing methods for promoter prediction, our proposed method can significantly improve the accuracy of promoter prediction by the property selection of structural features.

Fourth, we investigated the genome sequences, and proposed a feature selection framework that is employed to select structural features correlative with promoter sequences. Under the complete analysis of twelve structural features on Nucleosome distribution along chromatin and genomic DNA accessibility, we systematically explored nucleosome formation potential of genomic sequences and

the effect on chromatin organization and gene expression in S. cerevisiae. By the computation of correlation of nucleosome distributions in genome and pearson correlation of structural profiles, we classified these features into two classes, one positively and the other negatively correlated with nucleosome occupancy. These two kinds of structural features facilitated nucleosome binding in centromere regions and repressed nucleosome formation in the promoter regions of protein-coding genes to mediate transcriptional regulation. Based on correlations of structural profiles and nucleosome distributions, we integrated all twelve structural features in a model to predict accurate nucleosome occupancy. The experimental results showed that the proposed structural features based model can greatly improve the accuracy of nucleosome prediction. Through the peak detection of structural profiles, we further developed a novel approach, called DLaNe, which is used to locate specific position of nucleosome. As a comparison, DLaNe can demonstrate its power in predicting nucleosome positions using structural features.

Key Words: Genome analysis, Gene expression clustering, Promoter prediction, Nucleosome location

目录

第 1 章 绪论

随着上世纪八十年代末人类基因组计划 (Human genome project) 的启动，生物信息学 (Bioinformatics) —由生物学、计算机科学以及应用数学等学科相互交叉而形成的一门新兴学科应运而生[1]，并已成为当今自然科学尤其是生命科学的前沿研究领域之一。生物信息学的定义随着研究发展的需要被几经改动，一般认为，生物信息学是研究生物信息的采集、处理、存储、传播、分析和解释等各方面的一门学科，它通过综合利用生物学，计算机科学和信息技术而揭示大量而复杂的生物数据所蕴含的生物学奥秘。这个定义具有双重的含义：一是海量数据的管理，即数据的收集、存储与服务；二是对这些海量数据的分析与探索，从中发现新的规律。虽然在研究过程中生物信息学和计算生物学 (Computational biology) 经常将等同使用，但实际上二者存在着很大的差别。计算生物学通常是指将计算机系统和计算机方法应用于生物现象的模型研究；而生物信息学则是以计算机技术为研究手段和工具，同时采用数学、统计学的模型、研究方法来解决生物科学的问题，因而成为生物学、计算机科学、统计学、数学等多学科之间的交叉领域。生物信息学的发展既依赖于这些学科，同时又反过来为这些学科提供新的研究领域而促进其发展。

在人类基因组计划顺利完成后的十年，DNA 测序技术从以“荧光标记的 Sanger 法”为代表的第一代测序技术发展到高通量的第二代测序技术，不仅实现了测序的自动化、平行性和微型化，而且带动各项生物科学技术的迅猛发展，使得生物序列信息 (包括 DNA、RNA 和蛋白质) 发生了爆炸性的增长。至2011年4月，科学家们相继完成了～250 种真核生物和～4000 种细菌和病毒的测序工作，GenBank 数据库中的序列记录达到了 135,440,924 条，总共包含了 126,551,501,141 个碱基对[2]，并且碱基对的数量还在以每十八个月就增加大约一倍的速度在增长，在此基础上派生和整理出来的各种分子生物学数据库已达 1,000 多个。数据资源的急剧膨胀带来一个严峻的问题，传统的数据收集，管理和分析策略不再适用，如何有效的储存和进一步加工利用海量生物数据成为重要的研究课题。同时数据量的巨大积累往往蕴含着突破性发现的可能，这为数据挖掘方法提供了新的应用领域，我们正在从一个积累数据的时代向解释数据的时代转变。数据挖掘 (Data mining)，就是从存放在数据库，数据仓库或其他

信息库中的大量的数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程[3]。数据挖掘方法目前已广泛用于对海量生物数据的分析，并取得了大量的发现。但是，由于生物系统本质上的模型复杂性及在分子层上完备的生命组织理论的缺乏，给面向生物信息的数据挖掘研究提出了新的课题和挑战。本文主要研究数据挖掘新方法及其在生物序列分析中的应用。

本章首先回顾生物信息学的发展历史，接着概述生物信息学的主要研究内容，以及当前数据挖掘方法在生物信息学研究中的一些重要应用，最后给出本文的研究内容和章节安排。

1.1 生物信息学概述

关于“人类基因组草图”的论文于 2001 年 2 月在 Nature 上的正式发表，标志着生命科学已经由“分子生物学时代”发展到“后基因组时代”[4]。随着基因组测序技术日趋的成熟，基因组的研究对象不再局限于人类和少数几个模式生物，已经完成了从个体基因组到群体基因组的飞跃[5]。同时，DNA 芯片，RNA 测序，质谱等新技术的不断涌现，使得生物信息学的研究内容以基因组图谱为主线不断扩展，在短短十几年间，已经形成了多个研究范畴。

当前生物信息学领域的研究主要集中于核苷酸、氨基酸序列和基因表达数据的存储、分类、检索和分析等方面。一是以基因组 DNA (脱氧核糖核酸) 序列和氨基酸序列为出发点，找出基因组中的蛋白质编码区域，并且了解基因组中大量的非编码区里面蕴含的信息，破译 DNA 序列中的生命遗传规律，同时了解基因序列之间的转录、表达和调控关系，了解蛋白质序列之间的调控和相互作用关系，进而了解生命的代谢、发育、分化和进化的规律。二是以基因表达数据为出发点，找出与某种生物现象相关的关键基因，进而研究这些基因的功能和生命含义，以及这些基因之间的相互调控网络。以下简要介绍一些主要的研究方向。

1.1.1 基因组序列分析

随着大量物种的全基因组测序完成，后基因组时代以功能基因组为研究内容，破译隐藏在生物序列中的生命规律。一个生物体的基因组 (Genome) 是指包含在该生物的 DNA (部分病毒是 RNA) 中的全部遗传信息。基因组不是基因的简单排列，它包括基因和非编码 DNA，并且具有特定的组织结构和信息结构，

这种结构是在长期的进化过程中产生的，已经成为基因发挥功能所必须的要素。基因是具有遗传效应的特定 DNA 序列，即基因就是能编码某种蛋白质的 DNA 片段。根据其功能可分为调节基因 (Regulator gene) 和结构基因 (Structure gene)。调节基因从广义上讲是指任何一种能够调节或限制其它基因活性的一类基因；一般情况下则是指基因产物参与调节别的基因活性的基因，如阻遏基因等。结构基因是指除了调节基因以外的编码任何 RNA 或蛋白质产物的基因。非编码 DNA 是指不能被转录翻译成蛋白质的 DNA 片段。此类 DNA 序列在真核生物的基因组中占有大多数，通常认为具有调节基因表达的作用。由于生物的遗传信息主要存储在基因组中，研究生命规律的出发点就是核酸序列，对基因组 DNA 序列进行分析的目标就是从序列之中寻找基因及其表达调控信息，也就是了解基因组的结构和功能。对 DNA 序列的分析需要根据不同的需求进行不同的处理，并不存在一个通用的序列分析方法。

(1) 基因发现

基因是指导合成具有生物功能的多肽或蛋白质分子所必须的核苷酸序列，除了包含编码蛋白质的序列，还包含调控基因转录的序列。DNA 序列中蛋白质编码区域往往从特定的起始密码子处开始，在终止密码子处结束，从起始密码子到终止密码子之间的一段 DNA 序列称为开放读码框 (Open Reading Frame，简称为ORF)。发现基因的基本任务就是在基因组中找出所有可能的开放读码框，并进一步对这些开放读码框中的序列进行识别，确认其是否能够编码蛋白质。真核生物的 DNA 序列中蛋白质编码区域的结构比原核生物编码区域结构复杂得多，其蛋白质编码区域被一些称为内含子的核酸序列片段分割开来，被分开的编码序列片段称为外显子。因此，基因发现的另一个任务就是找出基因组中的内含子和外显子，而识别外显子和内含子的关键问题是识别它们之间的分界点，即剪切位点。

(2) 基因非编码区域的分析

在基因组中，除了编码蛋白质的区域，还存在大量的非编码区域，这些区域通常决定了 DNA 与蛋白质或者 DNA 与 RNA 的相互作用。如 DNA 转录起始位置上游的特定区域通常会有一些特定的核苷酸序列存在，特定的 RNA 聚合酶在转录前结合在这些核苷酸序列上，引起 DNA 序列的解链并转录，这些特定的区域被称为基因的启动子 (Promoter)。包含调控信息的区域还有增强子 (Enhancer)、抑制子 (Inhibitor)、基因终止序列 (Terminator sequence)、转录因子