学校编码：10384　　分类号_____密级 _____

学　　号：200431032　　UDC_____

厦　门　大　学

硕　士　学　位　论　文

基于判别分析的植物 poly(A)位点识别研究

Research of Plant poly(A) Site Identification Based on Discriminant Analysis

陈 舒 婷

指导教师姓名：吉　国　力　教授

专　业　名　称：系　统　工　程

论文提交日期：　2007　年　6　月

论文答辩日期：　2007　年　6　月

学位授予日期：　2007　年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2007 年 6 月

厦门大学学位论文原创性声明

兹呈交的学位论文，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的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1.保密（），在年解密后适用本授权书。

2.不保密（）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

摘　要

植物 mRNA 序列中多聚腺苷化位点（简称 poly(A)位点）识别是基因识别的重要组成部分，在基因组分析中，对 poly(A)位点的正确识别有助于确定基因编码的终止位置，对分析基因的转录过程及探索基因表达的调控机制都起着十分重要的作用。大量的研究人员已经对不同生物的 poly(A)位点识别问题进行了研究，但由于植物的 poly(A)位点表现出分散性、多样性以及复杂性的特点，所以在植物 mRNA 序列中关于 poly(A)位点选择的理解仍十分有限。

判别分析是根据判别对象若干个指标的观测结果判定其应属于哪一类的统计学方法。逐步判别分析是对进入判别模型的特征根据对判别贡献的大小进行逐步选择，最后根据筛选出的特征建立判别模型。

本文根据拟南芥 poly(A)位点上下游周围序列顺式作用元件的特征，运用逐步判别分析的方法来建立 poly(A)位点的识别模型。对建立模型采用的训练集数据，使用 k-gram 核苷酸模式、Z 曲线、位置特异性分数矩阵、一阶异构马尔可夫模型、阶乘矩等方式表示提取的生物特征；首先使用基于信息增益、熵等多种属性选择算法对特征空间进行初步的筛选，获得若干重要特征。而后对得到的序列特征的数值编码作为逐步判别分类的输入，针对训练数据建立判别模型。本文使用建立的判别模型对测试数据进行预测，并对各测试组的预测结果进行分析，发现逐步判别分类在识别精度上基本取得了令人满意的结果。逐步判别在位点识别模型的建立过程中可以进一步筛选出对位点预测有显著作用的特征，选择出的变量更能够反应类间差异，大大减少了新序列测定位点所需抽取的特征量。模型的训练和测试结果表明，拟南芥 poly(A)位点的逐步判别模型是一种有效且高性能的位点预测模型。

关键词： poly(A)位点识别；特征提取；逐步判别模型

ABSTRACT

Messenger RNA (mRNA) polyadenylation is a crucial step during the maturation of most eukaryotic mRNA, in which a polyadenine [poly(A)] tract is added to the cleaved 3'end of a precursor-mRNA post-transcriptionally. And predicting the poly(A) site of mRNA encoded by a gene would help to predict gene boundaries. Many researchers have done research on this problem in different species. However, because of diversity and complexity, plant mRNA poly(A) site selection only gain very limited understanding, and there is no formal report on the prediction of the poly(A) sites using a computer algorithm.

Discriminant Analysis is a statistic method to predict the type of the Object base on Indicators of the Object. Stepwise Discriminant Analysis is to build the model base on Screening character, which is selected from characters' contribution to Discriminant.

In this thesis, I build a Discriminant model base on Nucleotide Distributing Character Around the Arabidopsis poly(A) Site. I get the training data from k-gram Nucleotide mode, Z-curve, score matrix of Location Specific, A band Heterogeneous Markov Model, Factorial Moment, etc. Firstly, I select the character space base on information gain, Entropy and get the important character; then I translate the characters into Digital and build the model. Finally, I test my model through test data and analyze the result. It is satisfy about the Recognition Accuracy of Stepwise Discriminant Analysis. Stepwise Discriminant Analysis can select characters which are useful to predict poly(A) site, find Difference of Variables, Gradually Reduce the character to predict poly(A) site. The result of training and test show that Stepwise Discriminant Analysis of Arabidopsis poly(A) site is feasible and effective.

Key Words: poly(A) site identification; Feature extraction; Stepwise discriminant model

目　录

Contents

第一章　绪论

1.1 前言

近年来人类基因组计划和水稻基因组计划等大型国际合作研究项目的实施，使人类在生命科学领域尤其是核酸和蛋白质等生物大分子的序列、结构与功能等方面迅速积累了大量的数据和信息。迄今为止，已有一万多种蛋白质的空间结构以不同的分辨率被测定。基于互补DNA序列测序所建立起来EST数据库其记录已达数百万条。在这些数据基础上派生、整理出来的数据库已达500余个。这一切构成了一个生物学数据的海洋。这种科学数据的急速和海量积累，在人类的科学研究历史中是空前的。数据并不等于信息和知识，但却是信息和知识的源泉，如何处理、分析、解释和利用这些数据是一个迫切需要解决的问题。同时与正在以指数方式增长的生物学数据相比，人类相关知识的增长却十分缓慢。这构成了一个极大的矛盾，由此催生了一门新兴的交叉学科——生物信息学[1,2,3]。生物信息学是生物学与计算机科学以及应用数学等学科相互交叉而形成的一门边缘学科。它通过对生物学实验数据的获取、加工、存储、检索与分析，进而达到揭示数据所蕴含的生物学意义的目的。

生物信息学是内涵非常丰富的学科，其核心是基因组信息学，它希望通过对DNA、RNA和蛋白质的研究，分析生物序列中的结构、功能、进化，以及生物序列间的关系。生物信息学研究所要达到的目标主要包括：

1 识别出基因的精确外显子－内含子结构，以及对各部分的认识，识别和搜索，其中包括一些控制信号，例如 promoter，enhancer 等。

2 从氨基酸的序列预测蛋白质的高级结构（二级和三级）。

3 了解基因表达的调控机理及对其功能的研究和分析。

其中了解基因表达的调控机理是生物信息学的重要内容，根据生物分子在基因调控中的作用，描述人类疾病的诊断、治疗内在规律。它的研究目标是揭示基因组信息结构的复杂性及遗传语言的根本规律。近来的研究表明，基因组不仅是基因的简单排列，它有其特有的组织结构和信息结构，这种结构是在长期的演化过程中产生的，也是基因发挥其功能所必须的。弄清楚生物体基因组特有的组织

结构和信息结构，是解释生命的遗传语言的关键。

基因表达的第一步是从 DNA 上的遗传密码转录成信使 RNA（mRNA），转录的启动是基因表达的一个主要调控点，而转录后水平的调控在整个基因的表达调控网络中也处于非常重要的地位。真核生物中，成熟的有功能的 mRNA 要经过原初转录本(pre-mRNA)5'帽子的形成、内含子的剪切及 3'末端的加工才能形成。而关于 3'末端的加工包括两个过程，加工首先在 3'非编码区内某一特定的多聚腺苷化位点（简称 poly(A)位点）处切割，产生断裂，随后在断裂末端进行多聚腺苷化。多聚腺苷化后的成熟的 mRNA 才能保证被运送到细胞质中进入核糖体被翻译，同时，多聚腺苷化对 mRNA 的稳定性有很大影响。多聚腺苷化有两个主要的问题，一是由在 pre-mRNA 的 3'-UTR 区的一组特定的信号来决定哪里是 poly(A)位点，这一信号是由基因组信息所决定。另一个是由一组蛋白质与酶来识别这些信号，然后在 poly(A)位点上切割，加上一大串腺嘌呤（adenine）。这篇论文主要是研究如何在植物 mRNA 序列中识别 poly(A)位点的问题。

1.2 一些相关的生物学方面的基础知识

1.2.1 遗传物质

遗传是物种延续和进化的前提，携带了遗传信息，生命体才能按照指令正确地生长、发育并维持其自身结构和功能，并且把这种遗传信息从亲代传递给子代。基因是遗传的基本单位，现代分子生物学研究已经证实 DNA 是遗传物质的主要载体，每个基因都是由代表一种特殊蛋白质信息的 DNA 序列组成的。几乎所有生物的遗传物质都是 DNA，只有少数噬菌体、植物病毒和动物病毒的遗传物质是 RNA (ribonucleic acid，核糖核酸)。生物体的形态是生物体所具备的全部基因及其发育环境相互作用结果，而当 DNA 序列发生变化而引起表型改变时，就揭示了基因对生物体的影响。

生物可以分成两大类－－原核生物和真核生物。原核生物 (prokaryote) 是单个细胞，其遗传物质分布在整个细胞中。真核生物 (eukaryote) 中，遗传物质被组织在细胞核这个轮廓分明的结构部分中，在细胞分裂前及准备细胞分裂时，DNA 被暂时组织成一种紧密结构，称为染色体 (chromosome)。DNA 是染色体中最重要的组成部分，它是一种很长的多聚体，叫多核苷酸 (polynucleotide)。核苷酸由核糖、磷酸基团及碱基 (base) 三部分组成。如图 1.1 所示，碱基是腺嘌呤 (adenine, A) 、鸟嘌呤 (guanine, G) 、胞嘧啶 (cytosine, C) 、胸腺嘧啶 (thymine, T) 中的一种。ATCG 在结构上是以成对的方式存在的，A 只与 T 配对，C 只与 G 配对，反之亦然。因此，通常称 DNA 序列中的一个字符为一个碱基对 (base pair, bp)，以此作为 DNA 序列的长度单位，例如人类基因组大约共有 30 亿个碱基对。核苷酸之间相互由磷酸二酯键连接起来成为长链的 DNA 分子。位于核酸链一端的末端核苷酸有一个游离的磷酸（5’端），另一端的末端核苷酸有一个游离的羟基（3'端）。生物学上对多核苷酸序列的记录通常按照从 5’端到 3’端的顺序进行，例如 5’- ATGGTCAACTG -3’ 。

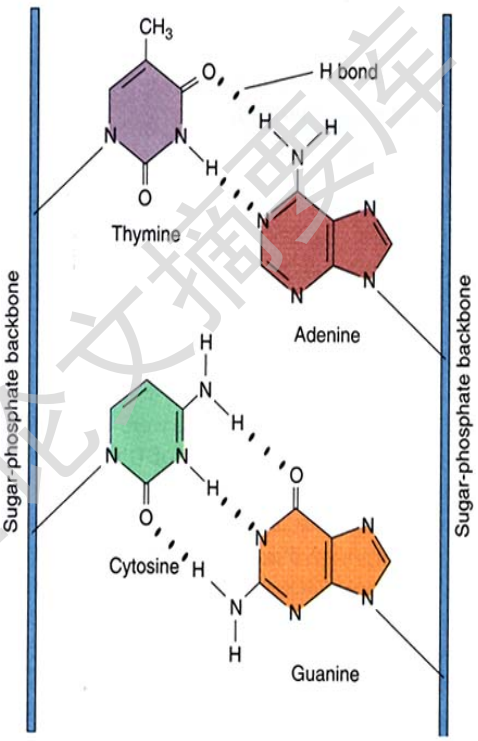

图 1.1 核苷酸结构及四种碱基的配对

Watson 和 Crick 提出的 DNA 双螺旋 (double-helix) 结构模型[4]为遗传信息传递奠定了物质结构基础。对于 DNA 的结构来说，Watson 和 Crick 模型的重要特点是：DNA 分子（通常）并不是一条多核苷酸链，而是两条。这两条链以双螺旋的方式彼此缠绕在一起，就像拧在一起的两股绳子一样，如图 1.2 所示：

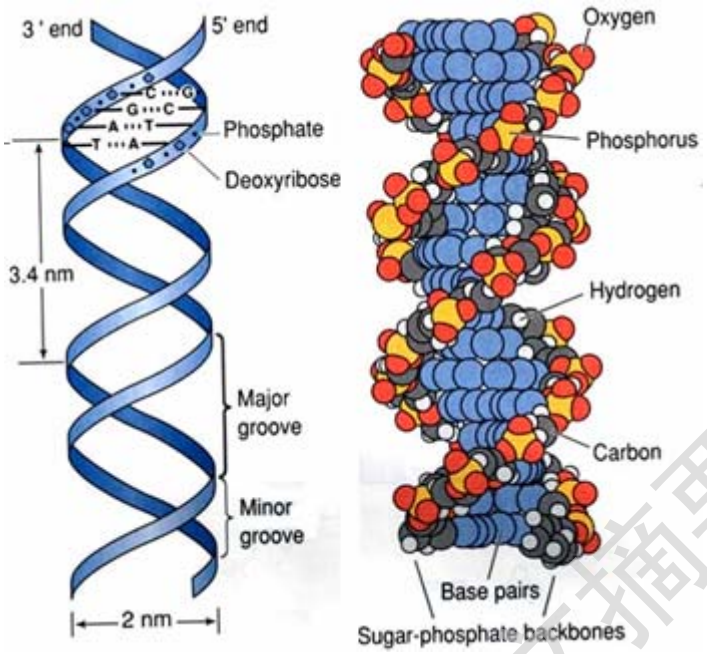

图 1.2 DNA 的双螺旋结构

图片来源：Moises Burset, Roderic Guigo. Evaluation of Gene Structure Prediction Programs.Genomics. 1996.34(3):353~367

上图中，两条 DNA 单链上的碱基互补形成双螺旋结构，一条链的 3’端到 5’端对应于另一条链的 5’端到 3’端。右图为双螺旋的分子结构图。

每条链的基本骨架是交替的糖－磷酸基团，两条链的极性是相反的，也就是说，一条链上的原子序列与另一条相反。因此，一条链对另一条链来讲是倒置的，也叫反向平行。碱基排列跟基本骨架成直角，并伸入分子中央。一条链上的碱基总是跟另一条链上同一水平的碱基配对。因此，两条链沿其全长通过碱基对之间的氢键结合在一起。其全部结构就像是沿着轴心旋转的绳梯一样，边上的绳子相当于糖－磷酸的基本骨架；梯级相当于配对的碱基。

双螺旋分子两条链的严格互补性，是指一条链的核苷酸顺序，无例外地取决于另一条链。每条 DNA 链都能作为模板，以合成一条准确地限定核苷酸顺序的新链。图 1.3 显示的 DNA 复制 (DNA replication) 机理，即根据互补规则，解释 DNA 的两条链如何指导互补链的合成，从而产生两个与亲本 DNA 相同的分子的。图的下部表示亲代双链体，上部表示正在互补的碱基配对产生的两个子代双链体。亲代的两条链已经分开，因此每条链都能作为互补合成的模板，每一个子代双链体在序列上与原先的亲代完全相同，而且含有一条亲代链和一条新合成的链。细胞的每一次分裂都会产生一个完整的基因组拷贝，这是遗传信息从一个细