**代号**

**分类号**

**学号**

**密级**

**10701**

**TP301.6**

**公开**

**1103121608**

**题（中、英文）目**

**基于质心估计的模体发现算法**

**及其在ChIP-seq数据上的应用**

**A Centroid Estimation Based Motif Discovery Algorithm**

**and Its Aplication on ChIP-seq Data**

**作者姓名**

刘源

**指导教师姓名、职务**

霍红卫 教授

**学科门类**

工学

**提交论文日期**

**学科、专业**

计算机软件与理论

**西安电子科技大学**

**学位论文独创性(或创新性)声明**

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切的法律责任。

本人签名： 日期

**西安电子科技大学**

**关于论文使用授权的说明**

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。同时本人保证，毕业后结合学位论文研究课题再撰写的文章一律署名单位为西安电子科技大学。

（保密的论文在解密后遵守此规定）

本学位论文属于保密，在 年解密后适用本授权书。

本人签名： 日期

导师签名： 日期

# 摘要

模体发现是生物信息学领域中的重要问题，模体中蕴含着重要的遗传信息，在研究基因转录和调控机制方面有着重要的意义。通过计算类方法来寻找联合调控基因片段中包含的模体已经成为了一种普遍的模体发现方式，计算类的模体发现算法和工具目前已经超过了100种。

为了描述模体发现中模体数量不确定，实例间存在不确定性变异的特征，本文采用统计学模型来更好的表示模体，并分为两步来求解模体发现问题。第一步，通过给定的模体碱基构成矩阵得到模体起始位点集的后验概率，并通过质心估计选取最具代表性的起始位点；第二步通过给定模体起始位点集来更新模体碱基构成矩阵。随后，算法将以上两步整合到吉布斯采样的框架中，通过多次迭代，使最终的结果趋于收敛。ChIP-seq上的序列数据数据量庞大，序列非常短，这样的数据很难直接用于传统的模体发现算法。为了解决这个问题，本文设计了基于*k*-均值聚类的聚类方法对ChIP-seq数据进行预处理，在精简序列数量的同时最大程度保留了模体相关信息，随后将处理后的数据作为基于质心估计的模体发现算法的输入，得到了预期的效果。

本文同时在Tompa标准测试集和ChIP-seq数据上对算法的有效性进行了验证。在Tompa测试集上采用性能系数对算法的结果进行了评价，并对比了算法在不同物种基因序列上的效果，发现在酵母上可以获得比其他物种更好的效果；使用转录因子Oct4在老鼠胚胎干细胞上的ChIP-seq数据作为输入，成功的找到了Oct4的模体。

关键字：模体发现 质心估计 k-均值聚类 贝叶斯统计 ChIP-seq

# Abstract

Motif discovery is an important subject in Bioinformatics, motif is a medium of a vital piece of genetic information, and plays a key role in the research on gene transcription and regulation. Computational methods of targeting motifs among co-regulate DNA sequences have been widely used, and over 100 tools of this kind have been presented in recent years.

To better express the uncertainty of motif number and the variation between motif instances, this thesis fits motif in a statistical model, the solution for motif discovery is devided in 2 steps. First, find the motif starting positions in sequences on condition of a given motif composition matrix. Second, update motif composition matrix by the starting positions found in step 1. The 2 steps above are then fit into a Gibbs sampling framework and iterating until the final results are convergence. ChIP-seq data always contains a large number of short sequences, and usually can’t be used in traditional motif discovery methods. In this thesis, a k-means clustering algorithm is designed to preprocess the ChIP-seq peak region sequences, the amount of sequence is successfully reduced with most useful information about motif preserved, the preprocessed sequence is then used as the input of the centroid estimation based algorithm and the result is as expected.

The dataset presented by Tompa el al. as well as a group of ChIP-seq data is used to test our centroid estimation based algorithm. On Tompa dataset, performance coefficient is used to evaluate the result and to compare performance on varied species, the result shows that our algorithm works better on Yeast than the others. The ChIP-seq data of the Oct4 transcript factor conducted on mouse ES cell is then used as another input, and the Oct4 motif is successfully found.

**Keywords**: **Motif Discovery Centroid Estimation k-means clustering**

**Bayesian Statistics ChIP-seq**

# 目录

[**第一章 绪论 3**](#_Toc375845061)

[1.1研究意义及其背景 3](#_Toc375845062)

[1.2国内外研究历史及现状 4](#_Toc375845063)

[1.3本文的主要内容及组织结构 5](#_Toc375845064)

[**第二章 模体发现算法的统计学模型 6**](#_Toc375845065)

[2.1 贝叶斯统计学 6](#_Toc375845066)

[2.2 共轭先验 7](#_Toc375845067)

[2.3 数据缺失模型 8](#_Toc375845068)

[2.4 模体发现模型实例 8](#_Toc375845069)

[**第三章 基于统计分析的模体发现算法基本结构 10**](#_Toc375845070)

[3.1 模体表示 10](#_Toc375845071)

[3.2 基于统计分析的模体发现算法基本结构 13](#_Toc375845072)

[3.3 模体模型 14](#_Toc375845073)

[3.3.1 匹配得分函数 14](#_Toc375845074)

[3.3.2 模体起始位点先验分布 15](#_Toc375845075)

[3.4 背景模型 16](#_Toc375845076)

[3.5 模体评价策略（模体统计学意义评估） 17](#_Toc375845077)

[3.6 求解策略 19](#_Toc375845078)

[3.6.1 吉布斯采样策略 19](#_Toc375845079)

[3.6.2 EM策略 22](#_Toc375845080)

[3.6.3 其他策略 23](#_Toc375845081)

[**第四章 基于质心估计的模体发现算法 24**](#_Toc375845082)

[4.1 质心估计 24](#_Toc375845083)

[4.2 基于质心估计的模体发现算法介绍 27](#_Toc375845084)

[4.2.1 模型建立 27](#_Toc375845085)

[4.2.2 单序列单模体 28](#_Toc375845086)

[4.2.3 单序列多模体 29](#_Toc375845087)

[4.2.4 多序列多模体 33](#_Toc375845088)

[**第五章 面向ChIP-seq数据的应用 35**](#_Toc375845089)

[5.1. ChIP-seq实验及ChIP-seq数据特征 35](#_Toc375845090)

[5.1.1染色质免疫共沉淀实验 35](#_Toc375845091)

[5.1.2高通量测序技术及ChIP-seq数据特征 36](#_Toc375845092)

[5.2 面向ChIP-seq数据的应用 38](#_Toc375845093)

[**第六章 算法测试与实验结果分析 41**](#_Toc375845094)

[6.1 模体准确性评价标准 42](#_Toc375845095)

[6.2 在同源基因序列上的测试 43](#_Toc375845096)

[6.3 在ChIP-seq数据上的测试 47](#_Toc375845097)

[**第七章 总结 50**](#_Toc375845098)

[**参考文献 51**](#_Toc375845099)

# 第一章 绪论

## 1.1研究意义及其背景

蛋白质是生命的物质基础，这种线性大分子由氨基酸构成，每一种蛋白质都有自己独特的氨基酸序列。现代生物学已经通过实验手段揭示了分子中蛋白质的合成过程：首先，DNA上一个特定的片段（即基因）被复制到信使RNA上，然后信使RNA携带该基因的信息到核糖体上合成特定的蛋白质。真实的过程要更为复杂。DNA（脱氧核糖核酸）是一种由四种碱基——腺嘌呤（A），胸腺嘧啶（T），鸟嘌呤（C）和胞嘧啶（G）组成的简单线性分子，DNA多数时候是惰性的，而且总是呈双链螺旋结构，这样有助于DNA作为遗传信息的稳定存储介质。RNA与DNA组成类似，不同的是没有胸腺嘧啶（T）——由与其类似的碱基尿嘧啶（U）来代替，糖化物上增加了一个氧原子。另外RNA通常是单链形式存在的，所以化学性质比较活跃。在真实的蛋白质合成过程中， 一些称作转录因子的蛋白质结合到DNA上的某些特殊片段上，促使RNA聚合酶合成RNA复制下游基因，这个过程称作转录。转录完成后带有基因信息的RNA会通过复杂的翻译过程指导蛋白质的合成。

生物学实验确定了以上的蛋白质合成过程，但这个过程中有众多的细节问题仍然是不明确的，其中一个重要的问题就是模体发现问题。在转录过程中，DNA上与转录因子结合的区域被称作模体（Motif），模体与其下游被RNA复制的片段合称启动子区。尽管承受着进化压力，启动子区较之DNA的其它片段还是更为保守的，模体上包含着大量与转录相关的信息。目前对于模体的认识和研究是相当不充分的，想要进一步的研究这些相对保守的DNA片段，首要的任务是在现有的DNA序列中找出这些模体，这就是模体发现问题。

我们可以将DNA抽象成由字母A、C、G、T构成的字符串，自然而然的，模体就是DNA的子串。所以模体发现问题可以简单的归纳为：给定一条或多条字符串，从中找出一些重复出现的子串，并且这些子串的模式以及存在性都是未知的。

由于模体发现对于研究基因调控具有重要的意义，这个问题多年来吸引了大量的生物学家和计算机学家。生物学家通过实验室手段发现了大量的模体，同时计算机学家开发的大量计算类方法也为这个领域做出了重要的贡献。

## 1.2国内外研究历史及现状

近30年内，计算类的模体发现方法一直被广泛的研究。最早期的模体发现算法之一可见于文献[1]，该算法实现了三种主要功能：1）搜索和计数；2）检测重复出现的子串和回文串；3）对多条序列进行两两对比，发现其共有序列。另外该算法有一个需要注意的特征就是它能够一定程度的识别存在变异或碱基缺失的模体。随后在1982年，文献[2]进一步推广了该算法，使其可以同时对比多条序列。这篇论文同样对于模体的特征进行了精确定义，同时限定了模体允许的最大变异，以及不同结合位点间的最大距离。同年，文献[3]使用了一个神经网络方法来解决模体发现问题，并首次引入了位置权重矩阵（不过这个矩阵在这篇文章中是用于RNA序列的）。与此类似，文献[4]引入了以对数频率为权值的位置权值矩阵。随后，计算机科学家在模体的表示方法、模体的评价方法、模体的寻找策略等多方面进行了广泛而深入的研究。目前已有超过一百种的模体发现算法和相应的工具可供使用，为生物学家对于基因调控的研究提供了非常大的帮助。

在最近十年，一方面众多的计算机科学家仍然在探索解决模体发现问题的新的计算途径，例如Gibbs采样算法，EM算法，以及基于这两种算法的众多改进版本。另一方面，也有相当大的研究精力转向了对已有模体发现方法的分类、对比、测评和综合研究上。这类研究一方面是面向生物学家的，旨在通过对现有模体发现工具的全面测评为生物学家的选择提供多方面的帮助，例如文献[5]综合测评了13种最具代表性的模体发现工具，提出了众多的测评标准，并构建了一份用于测试模体发现算法的通用数据集；另一方面，对于现有模体发现算法的综合分析旨在为算法的设计者提供参考和帮助，例如文献[6]分析了大量的模体发现算法，将这些算法分为两类：1）基于字符串的方法，这类算法主要采用穷举和搜索的策略，而且后期大多需要使用聚类或其他方法对结果进行进一步的处理。2）基于统计分析的方法，这类算法通过对模体模型和背景模型的方法，使用最大化后验概率或者贝叶斯推断来估计模型参数。另外，文献[7]建立了一个综合的算法框架，并依照这个框架分析了大量的模体发现算法。而文献[8]通过分析五种主流模体发现算法，分析了模体发现算法的局限性和潜在优势，并使用聚类方法对这五种算法找到的模体进行了进一步的筛选求精，得到了比较满意的结果。

此外，在计算类方法不断创新的同时，生物学实验手段的改进也为模体发现方法提供了新的思路和方法。从2007年以来，ChIP-seq实验通过将染色体免疫共沉淀方法和下一代高通量测序结合起来，为全基因组的序列分析提供了新的途径，对于ChIP-seq实验所产生的数据的分析也称为研究的热点之一，目前已经出现了BayesPeak[9]、MACS[10]、QuEST[11]以及DREME[12]等被广泛使用的工具。

## 1.3本文的主要内容及组织结构

现有的模体发现算法种类繁多，国内外学者通过各种方法来尝试解决模体发现问题，本文将研究范围限定在基于统计分析的模体发现算法上，首先总结了此类算法的常见结构，然后依照该结构，对多种算法进行了详细的分析和对比。随后，本文提出了基于质心估计的模体发现算法，并同时在4个物种的同源序列和一组ChIP-seq数据上测试了该算法，取得了较好的效果。

本文的内容按照以下方式组织：

第一章介绍了模体发现问题的背景、研究历史和现状；

第二章介绍了贝叶斯统计学的特征，以及与模体发现问题相关的统计模型及常用的计算技术。另外，本章通过一个简单的例子，从统计学的角度为模体发现建立了模型，然后根据统计学方法，为解决模体发现方法给出了简要的求解策略。值得注意的是，本章归纳出的模体发现问题的模型和求解方法是适用于现有的所有基于统计分析的模体发现方法的。

第三章首先介绍了常用的模体表示方式，在后文的论述中，这些模体表示方式将多次用到。然后，在第二章的基础上，本章的第二节提出了总结了统计学模体发现算法的基本结构。本章的后续章节按照该结构的构成，逐个模块对多种算法的特定部分进行横向对比。这样的对比方式可以更细致的凸显出同类算法的不同之处，同时也会揭示出算法设计者在算法不同部分所作出的刻意选择的原因。

第四章详细的介绍了基于质心估计的模体发现算法。

第五章给出了将基于质心估计的模体发现应用于ChIP-seq数据的方法。

第六章分别在4个物种的同源序列和ChIP-seq数据集上对算法进行了详细的测试，并对测试结果进行了分析评价。

第六章总结了本文的主要贡献和不足，同时提出了未来的工作方向。

# 第二章 模体发现算法的统计学模型

基于统计分析的模体发现算法近年来在模体发现领域一直居于主导地位，例如EM算法[13]，Gibbs采样算法[14]，以及这两种主要算法的众多扩展版本，MEME[15]，AlignACE[16]，Gibbs Recursive Sampler[17]，等等。另外还有大量的基于字符串的方法为了增强适应性也加入了统计学模型，例如FMGA算法[18]在遗传算法的基础上加入了位置权重矩阵来增强对模体变异的适应性。

本章试图通过一些简单的例子，从贝叶斯统计学的角度来为模体发现问题建立数学模型。

## 2.1 贝叶斯统计学

贝叶斯统计学的根本目的在于做出更准确的推断，即根据事实或假设来得出结论或者决策。这里的事实在统计学中即为观察数据，假设是根据相关系统建立的概率模型，结论或者决策通常涉及到不可观测的量的取值。

经典统计学在推断中使用点估计方法（最常见的是最大似然估计）得到未知量的值。为了保留估计的不确定性，通常为未知参数设立一个置信区间。贝叶斯统计学采用更加一般化的方法，为所有不确定的量建立分布模型，然后再通过观察数据来对所有这些量求其后验概率。贝叶斯推断只使用经过数据修正的概率分布来得出结论或决策。

贝叶斯统计将所有的量都纳入建模范围，不论是观测到的数据，未知的参数，或者缺失数据，在贝叶斯统计学中都被视为随机变量，贝叶斯统计学的分析行为被文献[19]划分为了3步：（1）使用全概率模型。求出所有量的联合概率。（2）在观察数据后，对特定的量进行后验概率计算。（3）评价模型的适用性并作出改进。

令，，分别表示观测到的数据，不可观测的参数和参数的先验分布。进行第（1）步的标准步骤就是计算数据的似然值，即将所有的未知变量和参数的概率求乘积。然后使用 来求出联合概率。第（2）步的直观做法是通过观测数据，应用贝叶斯定理来计算要求的随机变量的后验概率。



当为离散变量的时候，将积分换成求和，当为多个变量组成的向量的时候，采用多重积分（或求和）对每个的分量进行积分。需要注意的是完成贝叶斯定理中的计算需要对联合概率中每个未知变量进行积分，除了要求的量。尽管贝叶斯公式表面看起来是很简单的，使用贝叶斯统计仍然存在两个难点：（1）根据实际情形建立适当的统计模型非常困难，必须使得能够精确的表达实际场景的需求。（2）贝叶斯公式中的计算在大多数情况下是非常困难的。

## 2.2 共轭先验

对于应用贝叶斯统计的两个难点，建立模型需要对实际情景有细致的了解和分析，以及适当直觉和经验；而为了简化计算，更具体的说，为了使得贝叶斯公式中的积分可以计算，文献[19]提出了共轭先验分布。简单的说，共轭先验就是一组与似然函数具有相同形式的分布函数这样，使用共轭先验的贝叶斯公式，化简后就会与先验函数具有相同的形式。

在计算生物学中，因为被分析的数据通常形式比较固定——由四种字母组成的DNA序列或者由20种字母组成的蛋白质序列。所以最常用的分布模型是二项分布（伯努利分布）或者多项分布（二项分布的更一般化的形式），未知参数通常与组成序列的字母的出现频率有关，与多项分布族对应的共轭先验分布是狄利克雷分布，其中Beta分布是一种特殊形式，与二项分布对应。

## 2.3 数据缺失模型

在大多数模型中，分清未知参数和缺失数据是非常重要的。本质上说未知参数和缺失数据是没有显著区别的，在联合概率中它们均以未知变量的形式出现，在按照贝叶斯公式求解后验概率时，它们也是处于相同地位，可以根据需要通过积分消去。但另一方面，未知参数通常与全局数据的特征有关，而且未知参数的数量是固定的；而缺失数据的数量会随着观测数据规模的变化而变化。例如在模体发现问题中，输入数据为DNA序列，未知参数为A、C、G、T四种碱基的出现频率，而缺失数据就是每条序列中的模体起始位点（或称序列对准数据）。随着输入序列数量的增加，中元素的数量保持不变，而序列对准数据的数量在随之增加。

当模型中存在缺失数据时，要进行推断，可以通过求“观测数据似然值”来进行：



因为这个公式中的积分通常是难以计算的，所以通常需要一些复杂的计算方法辅助计算，例如EM算法[13]和数据增广[20]。

## 2.4 模体发现模型实例

首先以一个简单的硬币投掷举例。假设有*n*枚硬币，这些硬币质地是不均匀的，即投掷之后正反面朝上的概率不同。依次投掷这些硬币并将其落地后的结果摆成一排。现在需要估计这一排中国徽朝上的硬币数量。

假定这*n*枚硬币国徽朝上的概率都相同，且为，令表示国徽朝上的硬币数量，表示数字朝上的硬币数量。观察到的硬币序列的的似然函数就可以写作：

 (2-1)

为参数指定二项分布对应的共轭先验分布Beta分布，则的先验分布为：

** (2-2)

联合概率为：



把式（2-2）代入上式：



 (2-3)

则的后验概率为

 (2-4)

这个后验概率刚好与共轭先验分布同分布，仍然是一个Beta分布，只不过分布参数被更新了，即变成了，注意到在这个后验分布中可以和交换位置，它们是等价的，所以通常也被称作伪计数。

现在假设以上例子中的硬币有*D*个面，对上式中的记法稍作扩展，，二项分布扩展为多项分布，共轭先验从Beta分布扩展为狄利克雷分布，按照同样的方法，根据(2-3)，有：



其中是狄利克雷分布的参数，可以认为是伪计数，是*d*面朝上的计数，。与(2-4)类似，可得后验概率：



这样，令*D*=4，上例就解决了计算无模体序列中各个碱基在某个位点的出现频率。

现在更进一步，假设上例中*n*枚*D*面的硬币实际上分为两种，这两种的构造不同，所以每种出现面*d*的概率不同，假设其中一类是A，有*w*个，那么A的其实位置就是缺失数据，其他条件与上例相同，我们根据数据缺失模型一节的论证，以及上例中求后验概率的方法，可以：(1)将上例中的参数换为缺失数据，得出缺失数据的后验概率，即找出模体位置；(2)在上例中添加缺失数据，与可观测数据一起求出参数的后验估计，即求出模体构成。实际上EM算法就是交替的执行（1）和（2）最终得到缺失数据和参数的值。而Gibbs采样方法是通过交替对这两个值进行采样来找出模体。总的来说模体发现问题可以归约为一个包含缺失数据的贝叶斯推断问题，采用上例中的策略可以解决，但具体的模型非常复杂，且操作难度很大。

# 第三章 基于统计分析的模体发现算法基本结构

## 3.1 模体表示

DNA序列可以使用字符串来表示，而模体作为DNA序列的子串，自然也可以表示为一个字符串，然而因为变异的存在，同一个模体可能以不同的形态出现，为同一模体的不同形态找到一个统一的表现形式是十分必要的。

常见的模体表示方法主要有错配模型，IUPAC(International Union of Pure and Applied Chemistry)模型，位置权重矩阵和Sequence Logo。错配模型是由一个元组表示的，，其中表示长度为*n*的共有序列，表示*cs*中最大的汉明距离，可理解为允许的最大变异个数，这个模型中通常*n*取6到20，*d*取1到4。通常这种模型可以使用元组(*n,d*)来简单表示。

IUPAC模型是一个加入了退化字符的一致序列，表达能力更强，的每一位都是的非空子集。它用单个字母来表示的任意子集。

表3.1 IUPAC模型中各字母的含义

|  |  |
| --- | --- |
| 字母 | 代表的子集 |
| G | G |
| A | A |
| T | T/U |
| C | C |
| U | U, T |
| R | G, A |
| Y | T/U, C |
| M | A, C |
| K | G, T/ U |
| S | G, C |
| W | A, T/ U |
| H | A, C, T/ U |
| B | G, T/U, C |
| V | G, C, A |
| D | G, A, T/U |
| N | G, A, T/U, C |

位置权重矩阵（Position Weighted Matrix, PWM）模型的表达能力更强，它包含一个阶的矩阵，*n*表示模体长度，是矩阵的列数，矩阵每一列代表模体中的一位。矩阵的元素是指碱基*b*在模体第*i*位出现的概率。



图3.1. PWM模型

另外一种模体表示方法是Sequence Logo，这是一种新颖而直观的表示方法，由Schneider和Stephens提出[21]。

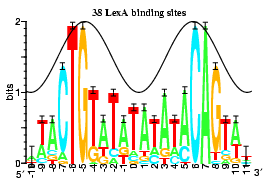


图3.2. Sequence Logo示例

图3.2展示了一个Sequence Logo，图中横坐标表示模体位置，从5’端到3’端，纵坐标表示每个位点的信息量（IC, Information Content）:

 (3-1)

其中 即为PWM中的元素，根据上式，完全保守的位点的信息量为2，某个位点一共有两种碱基出现，各占一半的概率，则该位点信息量为1，如果某个位点4个位点以等概率出现，则该位点的信息量为0。

公式(3.1)有两个缺陷：1）用于计算的模体样本较小，这样有可能从所选样本看来模体的某个位点是完全保守的，而加大样本容量以后则不能保证。2）该公式假定在背景中四种碱基是均匀分布的，实则不然。例如在秀丽隐杆线虫中GC的比例只有36%，而在恶性疟原虫中GC的比例只有19%。 所以更具有现实意义的做法是把公式（3.1）推广成相对熵的形式（也叫Kullback-Leibler散度），这其中就可以加入对背景概率的考虑：

 (3-2)

其中是背景中碱基*b*的概率。

本节中所讲的模体表示方法，与下节中的模体模型在形式上有部分重复。然而模体表示方法并不是本文提出的统一框架的一部分，它只是单纯的模体表现方式，有可能被用于算法最终结果或中间结果的表示，而模体模型是模体发现算法的重要元素和基础数据结构，在算法的运行过程中具有非常重要的作用。

## 3.2 基于统计分析的模体发现算法基本结构

现有的基于统计分析的模体发现算法都是非常复杂的。算法的设计者通过使用各种各样的求解策略来寻找给定的基因序列中重复出现的模式，并通过定义多种评价策略来计算这些重复模式的统计学意义，以期对这些预测模式做进一步的选择。由于模体的众多不确定性的特性，模体发现算法的结果通常也具有较大的不确定性。

通过计算方法得到的模体模式通常具有突出的统计学意义——出现次数上明显的重复，或者与背景序列的较大差异，抑或依据某种测度时较为明显的保守性。但这些统计学意义并非总是能够代表其生物学特性。为了解决这一问题，算法的设计者通过为猜测基因序列和模体在自然形成的过程中所采用随机过程和分布模型，试图为这些序列建立更具生物学意义的统计模型。这些模型已被证明在部分物种上是有效的[22]。一旦贴近生物学意义的合理统计学模型被建立，模体发现问题就可以使用有效的计算方法来得到更准确的模体。

通过对现有的大量基于统计分析的模体发现算法的分析，本文为总结了此类算法的基本结构，为了尽可能的涵盖广泛并具有一般性，这个框架由四部分组成，模体模型，背景模型，模体评价策略和求解策略。



图3.1 基于统计分析的模体发现算法的基本结构

图3.1展示了基于统计分析的模体发现算法的基本构成和各个组件间的合作机制。模体模型（Motif Model）和背景模型（Background Model）通过一组公用的参数（Model Parameters）紧密的联系，求解策略（Solution Strategy）是整个框架的驱动，其输入为现有模型参数，通过计算新的模体得分（Motif Score）来更新模型参数，经过一定的迭代次数取得最优结果。

本章的剩余部分将按照以上基本结构对模体发现算法的主要元素进行逐一解释，并会对每个元素上的不同选择进行详细对比。

## 3.3 模体模型

模体模型是模体发现算法中最基本的元素。常见的模体模型都可以归纳为以下形式：

**

*mg*表示从自然数到实数的映射，将模体起始位置自然数*p*映射到一个表示模体得分*mg(p)*的实数。这个映射函数包括两部分：匹配得分*m\*(p)*和模体起始位置为*p*的先验概率*og(p)*。

### 3.3.1 匹配得分函数

匹配得分函数*m\*(p)*就是为不同的模体给出不同的得分。现在匹配函数模型已经有非常多，主要可以分为两类：确定性模型和概率模型。

概率模型的主要代表是位置权重矩阵或称位置得分矩阵（Postion Specific Scoring Matrix, PSSM）。这种矩阵表示方式通常是假定模体的各个位置间的概率分布是相互独立的[3]。在这种模型下，一个模体的得分就是字符串的似然值的对数。位置权重矩阵得分可以自然的被当做模体与转录因子的结合能量（即结合时释放的能量）： 假定转录因子与模体*S*的结合能量来自模体每一位碱基贡献的结合能量之和，则有

**

其中*W(b,i)*表示矩阵第*b*行第*i*列元素的取值。

位置权重矩阵目前已经有很多扩展版本。这些扩展的主要思想是去除模体中各个位点之间的独立性。关于此更详细的讨论可参见文献[23][24][25]。为模体间各位点注入依赖的最直接的方式就是在矩阵中加入互相影响的位点对，例如文献[24][25]所做的。另一种直接的途径就是采用混合模型，这样模体的形成过程就可以用几个主要的随机过程来描述[26]，同理模体的分布也可以根据随机过程的终态分布得到。第三种扩展的途径是把模体当做一个*n*阶马尔科夫链来处理[27]。但这种方法面临的一个困境是*n*的取值如果太高，那就有太多的参数需要估计，而太低的话会又会降低位点之间的依赖（从自然进化，以及化学原理上讲，位点之间的依赖是完全可能存在的）。另外，假设位点之间的依赖关系是变化的，即某些位点之间的依赖被视为重要的，某些位点之间的依赖则不然，那么在这种情况下，采用变长马尔科夫模型是一个合理的选择[28][329]。

另外一个为模体位点间的依赖建模的方式是使用贝叶斯网络，文献[26]通过对不同的贝叶斯网络模型进行讨论，归纳出使用贝叶斯树（贝叶斯网络的一种形式），或者多种树的混合模型，能够很好的在参数个数，和对依赖程度的表达之间取得平衡，另外在计算上也有一定的便利。于此类似，文献[30]讨论了一个变阶贝叶斯网络。

除了在位点间的依赖中寻求改进，文献[30]研究了为模体中保守位点的分布建立模型。在他们的模型中，首先在底层对位点原型建立了一个马尔科夫链，位点原型上每个位点上可能的不同碱基呈多项分布，参数的超分布为狄利克雷分布。

确定型匹配得分函数通常将一个假设的模体起始位置*p*映射到一个布尔常量*b*上，如果认为这个位点确定可以作为模体起始位点，则返回1，否则返回0。这种确定性的方式普遍用于早起的基于字符串的模体发现方法，故此处不再赘述。

### 3.3.2 模体起始位点先验分布

模体的上下文环境对于其活动是具有重要影响的。距离转录起始位点（Transcription Starting Site, TSS。模体下游片段，从该位点起始会有一段序列被RNA复制作为合成蛋白质的重要信息）的距离，同源基因中序列的保守性，DNA结构和CpG-岛等都可能对模体行为产生影响。这些信息被综合的表现在*og*(*p*)中，*og*(*p*)表示某个未指定模体起始于位置*p*的先验信念。

最简单的模体起始位置先验就是模体富集率。这个比率只影响被算作模体的子序列的序列个数。另一个简单先验是序列偏差，就是表示某个模体在某条序列中出现的概率是高于其他序列的。包括Bioprospector[32]和TFBScluster[33]在内的少数工具提供了将搜索限定在某条序列中的选项，这对应着一个二项序列偏差先验。

在高等生物中，调控基序（即模体）出现的位置规律性并不强，可能在序列的上游远端，下游，内含子中，甚至在外显子中也有可能出现。然而，大多已知模体都是在TSS的上游位置出现，并且紧邻TSS。通常情况下模体起始位点的空间先验分布可以根据它与TSS的相对位置来建立。对于一些被广泛研究的基因，人们对于其模体起始位点的分布已经有了一定的经验信息，文献[34][35]就利用大肠杆菌上积累的经验信息建立了模体起始位点的分布模型。然而迄今为止，绝大多数的模体发现算法都只是直接在TSS的紧邻上游区域搜索模体。

DNA结构也是在设计先验模型中应该考虑的因素之一，例如序列的曲度及其在双螺旋结构中的位置等也会影响其包含模体的概率。但只有很少的模体发现算法将这些因素纳入考虑范围。

另外高GC区域和出现CpG-岛的区域通常会暗示模体的存在，有少量的算法考虑到了这一点，例如文献[35]。

## 3.4 背景模型

如果忽略生物学上的特定意义，那么背景和模体可以使用相同的模型。然而背景与模体又有以下区别：1）背景不需要起始位置的先验分布；2）背景不需要匹配得分函数。所以背景模型的建立采用与模体模型不同的方式。背景模型指背景中碱基的分布和背景的形成过程。

通常情况下背景中碱基的分布被视为多项分布[19][32]，由于贝叶斯理论在这一领域的广泛使用，采用共轭分布能够大幅的简化计算[33]，所以大多数算法中对于背景中碱基的分布采用多项分布对应的共轭先验分布，即狄利克雷分布。令，用来表示四种碱基出现在背景中的未知概率，则背景序列出现的概率（似然值）可以表示为，其中是序列中四种碱基的数量。这样，令的先验为，就是一个狄利克雷分布，参数为，常被称作伪计数。文献[33]提出了一个贝叶斯推断算法来计算背景中的碱基分布，并且考虑了背景序列的异质性，以及异质片段分布的不确定性。

背景的形成过程有两种常用模型，一种是完全随机的生成，另一种是根据马尔科夫过程生成。早期的模体发现方法多认为背景是随机生成的[19][33]，文献[14]提出了一个高阶马尔科夫模型。假定输入序列为，即长度为*n*的随机序列，其中*A1*到*An*为随机变量。令*a={a1,…,an}*为*A*的一个实例，其中*a1*到*an*为某个随机过程中*A1*到*An*的取值。如果假定序列中存在一个模体，令其起始点为*J*，*J*也是一个随机变量，它的取值是不确定的，假定生成这个序列的随机过程为*m*阶马尔科夫过程，在这样的前提下，观察到序列*a*的概率为：

**



其中*p(u,v)*表示给定状态*u*，转移到状态*v*的概率。自然情况下，由于使用*m*阶马尔科夫，所以*u*的最后*m-1*个字符和*v*的前*m-1*个字符是相同的，这样能保证*p(u,v)*总是大于0的。*qj(a)*是在模体的第*j*个位置找到*a*的概率。是马尔科夫链的终态分布，也叫平衡分布，同时也是初始状态，即序列*a*的前*m*个字符的分布，是人为指定的，通常为上文中的狄利克雷分布。经过一系列的实验，文献[22]的结果表明在真核生物（人，果蝇，老鼠，酵母）基因中，采用3阶马尔科夫背景会取得比较好的效果。

## 3.5 模体评价策略（模体统计学意义评估）

模体的统计学意义评估在模体发现算法中处于非常重要的地位。通常情况下，基于统计学的模体发现算法在初始状态下会先随机的选择一些位点来建立一个初始模型，然后经过一系列迭代，每步迭代中算法会判断是否要为原有的模型加入或删除某些模体，决策的依据就是加入或删除某个新模体后现有模体模型的得分会不会增加，如果增加，则更新模体模型及其得分。经过足够多的迭代之后，算法的结果会收敛在一个或多个（局部或者全局）最优得分模型上。

所以算法迭代过程中对模体模型的评价策略直接影响着算法收敛的方向，是模体预测过程的重要导向。

在统计学模型中，最常用的模体评价策略主要有三种：信息量、相对熵、对数似然值和最大化后验概率。

在2.1节中讲述Sequence Logo的时候已经介绍了信息量，简而言之，信息量就是一种衡量序列保守性的策略，根据公式（2.1），如果某个位点上四种碱基等概率出现，则该位点的信息量为零，即它是完全不保守的。相对熵是对信息量评价方法的一个扩展和补充，由于公式(2.1)有两个缺陷：1）用于计算的模体样本较小，这样有可能从所选样本看来模体的某个位点是完全保守的，而加大样本容量以后则不能保证。2）该公式假定在背景中四种碱基是均匀分布的，实则不然。例如在秀丽隐杆线虫中GC的比例只有36%，而在恶性疟原虫中GC的比例只有19%。 所以更具有现实意义的做法是把公式（2.1）推广成相对熵的形式（也叫Kullback-Leibler距离），这其中就可以加入对背景概率的考虑：

**

其中是背景中碱基*b*的概率。

似然函数模型是比较常见的，它只在给定模体模型的情况下，生成输入序列的概率的大小。通常情况下都是使用似然值的对数作为模体评价的测度，而不是似然值本身。最大对数似然值（MAL，Maximize A Log-likelihood）估计就是通过调整模型参数，使得根据当前的模体模型生成当前的输入序列的概率为最大：

**

对数似然值模型允许计算更为复杂的背景模型（特别是3阶或更高阶的马尔科夫模型），使用对数似然值模型使得排除一些低复杂度的重复模式变得较为简单，例如连续A或者连续T的串，在信息量模型中，这些重复模式有可能被识别为高信息量模式而保留，而在对数似然值模型中就会将这些模式排除。实际上，对数似然值约等于模体模型信息量乘以模体个数。所以这种评价模型除了表达了模体的保守性之外，还可以将模体的个数考虑在内。

最大化后验概率（MAP, Maximum A Posterior probability）模型是贝叶斯理论中常用的估计方法，是指在给定数据的条件下，选取后验概率最大的模型。如果假定每个模型的先验概率相同，那最大化后验概率本质上和最大似然估计相同。然而如果对于模型的分布有一定的先验知识，那么根据贝叶斯公式

**

可以得出

**

即MAP得分包括了对数似然值得分，另外增加了模体模型的先验知识来评价模体质量。

虽然这些模体评价模型已经在发现新模体的工作中起到了重要作用，但这些评价方法仍然不能做到很好的区别真实的模体和伪模体，所以优化和调整现有的模体评价方法，以及开发新的模体评价方法，仍然是十分必要的。

## 3.6 求解策略

模体发现问题在统计学上可以认为是一个贝叶斯数据缺失问题[36]。解决这个问题的基本方案是数据增广（Data Augmentation）[20]， 通常的数据增广策略有吉布斯采样策略（Gibbs Sampling）和期望最大化（EM，Expectation Maximization）策略。另外，在基于字符串的模体发现方法中，搜索策略被广泛应用，而少数的算法还用到了贪心算法和动态规划算法等常用的最优化策略。

### 3.6.1 吉布斯采样策略

吉布斯采样是马尔科夫链蒙特卡洛（Markov Chain Monte Carlo, MCMC）策略的一种常用实现。MCMC策略是一个基于采样的模拟技术，目的是产生一个符合目标分布的随机独立样本。MCMC策略的第一步是指定一个非周期性、不可归约的马尔科夫链，而且该马尔科夫链的终态分布就是样本的目标分布。下一步就是模拟第一步产生的马尔科夫链的一个或多个实现过程（即多次模仿随机过程）。每次模拟都会产生一组满足目的分布的随机变量（这个过程可以理解为抽样）。

吉布斯采样的过程可以大致划分为下面4步：

1. 选择初始值。
2. 设置计数器*i*=0。
3. 模拟随机抽取序列（模拟随机过程）：

，

，

**，

… …

**

得到



1. ，回到3。

以上是通用的吉布斯采样步骤，在不同领域的具体实现上可能有不同的形式，但基本的原理是一致的，都是通过采样来得到指定分布下的样本，补全缺失数据，然后跟可观测的数据一起对模型参数进行估计。

吉布斯采样在生物序列分析中的使用是非常广泛的，最早在这个领域使用此种策略的是文献[14]。但他们当时是将吉布斯采样使用在蛋白质序列而不是DNA序列中。吉布斯采样之所以被当做MCMC方法，主要因为：1）每一步的结果都是基于前一步的结果得出的；2）选择下一步的方式不是确定的，而是根据随机抽样的。这满足了MCMC的两个特点。

假设给定了*N*个序列*S1,S2,…,SN*，我们在这些序列中寻找相似的宽度为*W*的片段。Lawrence等人提出的吉布斯采样算法是这样的：维护两个数据结构——模式描述（即一个位置权值矩阵，包括了各个碱基的背景概率）和对准信息（即位置矢量，描述每条序列中的模体起始位置），执行以下两步的求解策略：1）预测更新。随机的，或者按照某种顺序，从*N*条序列中选出一条*z*。根据现有的*ak*，计算模式描述。2）采样。将序列*z*中每个长度为*W*的字串都看做可选模体。根据现有的模式描述来产生模体*x*的概率为*Qx*，根据现有模式描述中各碱基的背景概率来生成模体*x*的概率即为*Px*，则为*x*打分*Ax=Qx/Px*，等为*z*中的所有可选模体都打分之后，从中任选一个模体*y*，假定其起始位置为*i*，更新*az=i*。

该算法的中心思想是第一步建立的模式描述越精确，第二步中的对准信息就越准确，反之亦然。给定随机的*ak*，在第二步中，模式描述*qi,j*将确保无偏好性。一旦某个正确的*ak*被选定，*qi,j*就会开始反应一些可能的模式（虽然不是完美的）。这个过程通过不断的迭代改进*ak*，然后*ak*又反过来修正模式描述，使得所描述的模式跟背景的差异越来越大。在算法过程的第一步中提到了根据*ak*来计算*qi,j*。对于选定模式的第*i*个位点，我们有*N-1*个观测到的碱基，因为序列*z*在这个计算中是被排除的；令*ci,j*表示位置*i*上碱基*j*的个数。通过贝叶斯理论可知，要估计模式描述，可以为*ci,j*加上一个伪计数*bj*，模式概率，其中*B*指*bj*的和。碱基*j*的背景概率*pj*是按照无模型背景来计算，即所有非模体串中*j*的个数与所有非模体位点数的比值。经过归一化之后，*Ax*表示序列*z*中的模式起始于位点*x*的概率，算法通过一系列迭代找到集合*ak*，使得*Ax*，的乘积最大。同理，也可以使*Ax*，的对数和*F*取最大。保持以上各个量的记法不变，*F*定义为



其中*ci,j*和*qi,j*都是根据最终的对准信息*ak*计算出来的。

使用吉布斯采样策略的算法数量众多，应用较为广泛的还有AlignACE [16]，MotifSampler[37]，BioProspector[32]，GibbsST[38]，Gibbs Recursive Sampler[17]。下表展示了吉布斯采样方法的主要改进版本，这些算法的改进多集中在分布模型的改进和模体评价方法的改进上，对吉布斯采样策略本身的重大改进并不多见。

表3.2 使用吉布斯采样策略的算法对比

|  |  |
| --- | --- |
| 算法 | 描述 |
| AlignACE | 1. 模型改进。改进了背景模型，采用了关于基因组中碱基分布的先验信息。 2. 同时考虑了每条输入序列的互补序列。 3. 最终同时发现多个模体的策略被替换为逐个发现每个模体。 4. 使用MAP(Maximum A Posterior log-likelihood)估计。 |
| MotifSampler | 1. 考虑了每条序列中相同模体重复出现的分布模型。 2. 背景采用了高阶马尔科夫模型。 |
| BioProspector | 1. 使用了3阶马尔科夫背景模型，模型的参数可由用户指定，也可以根据某条指定的序列进行估计。 2. 每个模体的评价基于一个采用Monte Carlo方法得出的得分分布。 3. 提供了对回文串的建模。 |
| GibbsST | 1. 在吉布斯采样的基础上还使用了模拟退火，以增加全局最优的概率。 |
| Gibbs Recursive Sampler | 1. 采用严格的贝叶斯推断来推断模体位置和模体数量。 2. 考虑到了背景模型中的异质性。 3. 考虑了模体分布的先验信息。 |

### 3.6.2 EM策略

EM策略是一种在含有缺失数据的随机过程中学习参数化模型的通用方法。这个过程中会出现两种数据，可观测数据*o*和不可观测数据*h*，*o*和*h*都是随机变量，它们的联合分布取决于一些模型参数a，即这三个数据符合分布*P(o,h|a)*。更具一般性的，*o*，*h*和*a*都可以是向量。

EM的目标是通过观察数据来找到最有可能的参数*a*，即希望通过调整参数，使模型最好的表达数据。因为数据*h*是不可观测的，所以估计*a*只能依靠可观测数据*o*。所以EM的目标可以形式化的表述为：给定观测数据*o*，找到可以使联合概率

**

最大的参数*a*。

EM策略通过迭代的计算{*ap*}来最大化*P*(*o|a*)，最后*a*会收敛到一个局部最优解*a\**上。在EM策略的第*p*步迭代中，通过*ap*计算*ap+1*的具体过程分作两步：E步，估计缺失数据*h*，估计的方法是根据可观测数据*o*和参数*ap*来估计*h*的期望，将期望当做数据本身：



M步，使用新近估计的*h*来改进对*a*的估计。选择*ap+1*的标准是使*o*和的联合概率取得最大，即：

**

EM策略通过迭代地使用缺失信息的期望来代替缺失信息本身，简化了数据缺失问题的分析。在模体发现中，缺失的信息是模体的真实位置信息。所以，按照以上的EM框架，缺失数据*h*为模体位置信息，

可观测数据*o*为DNA序列*Si*，参数集*a*为位置权重矩阵，其中表示各个碱基在背景中的分布。

EM算法的改进版本中使用最广泛的是Bailey和Elkan提出的MEME算法[32]，MEME有三个亮点：1）使用序列的子串作为EM提供迭代的起始信息来增加找到全局最优解的可能性；2）去掉了最初的EM算法中每条序列中有且仅有一个模体的限制；3）用到了一种在迭代中动态的删除已发现的共享模体的统计学方法，这使得发现多个模体成为可能。

另外，使用EM策略求解模体发现问题的算法还有LOGOS[39]，Improbizer[40]，PhyME[41]，OrthoMEME[42]等。

### 3.6.3 其他策略

以上介绍了两类在基于统计的模体发现方法中较为常见的求解策略，但模体发现方法种类繁多，除以上两种常见的求解策略之外，搜索也是一种非常常见的策略，不过搜索通常用在基于字符串的模体发现算法中。这种策略就是在给出的基因序列中找出所有符合模体长度的字符串，按照启发性的方法搜索，或者暴力搜索得到在一定得分阈值之上的子串视为模体，这种策略的缺点是效率低下，准确度差，而且一般只适用于短模体（小于或等于12bp），较小数据集，以及植入模体数据，对于生物数据中的变异适应性较差。

遗传算法是一种启发式的搜索方法，在模体发现算法中，也是被尝试过的求解策略。文献[18]中提出了一种基于遗传算法的FMGA算法。该算法为了解决搜索算法的固有缺陷，采用了PWM模型以允许模体中的变异，另外又提出了一种特殊的罚分策略（属于模体评价模型）来允许Gap。并且在该文献的测试中，FMGA的预测准确度要优于MEME和Gibbs采样算法。

文献[43]中使用神经网络来寻找模体。该网络包含若干层次，每层都进行不同程度的分类操作。在该算法的层次结构中作者设法让每层的计算复杂度都维持在比较低的水平，这样每层进行的分类操作只需要在整个输入数据的一个很小的子空间中进行。

表3.3 基本结构下多种基于统计的模体发现方法的构成

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 算法 | 模体模型 | 背景模型 | 模体评价策略 | 求解策略 |
| MEME | PWM | High-Order Markov | MAL(maximize a log likelihood) | EM |
| AlignACE | PWM | High-Order Markov | Information Content | Gibbs Sampling |
| SemiMCMC | PWM | High-Order Markov | Improved MAL | Gibbs Sampling |
| PhyME | PWM | Phylogenetic Tree | MAL | EM |
| Consensus | PWM | Multinomial Distribution | Information Content; P-Value | Greedy Algorithm |
| Gibbs Recursive Sampler | PWM | High-Order Markov | MAL | Improved Gibbs Sampling |

# 第四章 基于质心估计的模体发现算法

## 4.1 质心估计

在统计学中，最大似然估计等以最优化方案为导向的估计策略有着非常普遍的应用，上文中分析的众多算法基本上全是建立在这些传统的估计方法的基础上。统计推断理论表明了最大似然估计以及与其相似的估计方法都只是使用观测数据来获得单个可能性最大的解，而实际上，除非其余的候选解都与这个解距离很近，或者这个解的概率大到接近1，并没有任何理论原则可以证明这个解可以很好的代表整个以后验概率为权重的解空间。文献[44]提出了一个新的估计方法，称作质心估计，通过使用贝叶斯推断，定义一系列损失函数，来找出解空间中最具代表性的解。

下面在离散多维空间中对质心估计和最大后验概率估计加以对比。

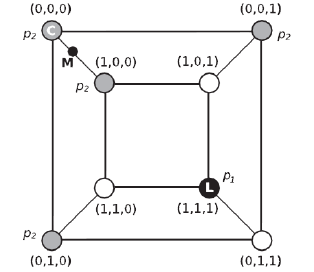


图3.1 三个随机变量组成的概率空间[44]

假设某个问题中有三个布尔值的随机变量（即服从二项分布）*x1*，*x2*，*x3*，他们所有取值组合表现为图3.1中的8个顶点，假设它们的先验信息与问题无关，经过观察数据的修正，它们的后验概率分布如图3.1所示。其中*p2*<*p1*<2*p2*。最大后验概率估计(Maximum A Posterior, MAP)是指求出，那么依照图3.1，*L*(1,1,1)为最大化后验概率估计。但是从这个后验概率空间看，*L*并不能很好的代表数据。首先，图中除*L*之外，所有概率不为0的点的概率之和为*L*的概率的2到4倍。其次，在这个分布空间中，所有与*L*距离最小的点（图3.1中所有与*L*相邻的点与其海明距离都为1），他们的概率全部为0。所以在这个点估计问题中，*L*其实是一个被孤立的点，然后由于它碰巧有最大的后验概率，按照最大后验概率的估计方法，就会将它选出来代表数据。其实，如果做一个平均值*M*(*p1*+*p2*, *p1*+*p2*,*p1*+*p2*)，下面可以证明*p1*+*p2*<1/2，所以，*M*更靠近*C*（0,0,0）。

更一般化的，对于*n*个服从0-1分布的随机变量，假设*p2*<*p1*<(*n*-1)*p2*，与上例类似，我们假设与某个点*C*相邻的*n* 个点与*C*的海明距离最大为1，而且连同*C*一起，这*n*+1 个点的后验概率均为*p2*，在这个超立方体中，位于*C* 的对角的*L* 的后验概率为*p1*，而*L*的所有相邻点的概率均为0，而且*L*与所有剩余的后验概率不为零的点的海明距离至少为*n*-1。 由于*p*1+(*n*+1)*p*2=1，所以，1/(2*n*)<*p*2<1/(*n*+2)，如果令*p*2=1/(*n*+3)，那么*p*1=2/(*n*+3)，那么整个(*x*1,*x*2,..*x*n) 的取值空间中，*C*周围（包括*C*）的概率质量与*L*周围的概率质量（包括*L*）的比则为(*n*+1)*p*2/*p*1=(*n*+1)/2，这样，如果问题的维度*n* 变的很大，那这个比值也就很大，同时*C*与*L* 的距离也会更大，其中*C*为质心估计，而根据最大化后验概率估计，因为*p*2<*p*1，*L* 总是会被选为最优估计。

另外，因为最大化后验概率或者最大似然估计只将注意力集中在概率最大的那个点上，而忽略了其它的点，而这些其它的点也很有可能恰好跟概率最大的点差异很大（距离很远），加之这些概率最大的点，它们的概率通常也很小，所以在高位离散空间中，最大化后验概率估计只能很好的代表估计本身，而不能真实的反映数据。

质心估计是建立在统计推断理论上的。统计推断使用损失函数来评价估计的损失（或风险），选取损失最小的估计。其中表示对参数的估计，表示该参数的真实值。通常在某个估计的损失时，采用期望损失，定义如下：

质心估计着力于真实的反映概率分布空间，所以找到的估计点应该不是被孤立的，而是周围聚集了大量概率和距离都很相近的点。为了达到这样的目的，质心估计选取那些会随着维度（随机变量个数）增加而快速增长的损失函数。首先，考虑海明损失，



其中*I*为判别函数，其参数为真时*I*取值为1，否则为0。*z*和*y*的长度都为*n*，海明损失单纯的衡量了离散空间中两个随机变量差异的位数。对于某个估计，其后验风险为









由此易得，当



时，后验风险最小，称为质心估计。

## 4.2 基于质心估计的模体发现算法介绍

### 4.2.1 模型建立

此处，基于第二章和第三章的背景，建立一个简单的模型，在此基础上对本文提出的模体发现算法进行描述。

定义序列集，，在任意序列上定义模体位点集，另外，定义，为序列中的碱基构成比例，其中为背景中的碱基构成，在下文的推导过程中，使用BG来标记背景序列，为模体中的碱基构成，*L*为模体长度，定义字母集，***R***中的所有序列都是基于字母集*S*。另外，此处沿用4.1节中的损失函数*H*和指示函数*I*的定义。根据2.3节的数据缺失模型可知，在模体发现问题中可以将看做缺失数据，将看成隐含参数，而为观察数据。下面我们将用适当的数据增广方法来求解这个数据缺失问题。

### 4.2.2 单序列单模体

为了使情境更简单，更清楚的阐明算法的操作步骤，本节先假设输入数据中只包含一条序列，而且该序列中只有一个模体。此时，。为简单起见，此处用*R*和*Y*来代表和，并且，我们为*Y*设置一个最简单的平均分布先验，根据2.2节的介绍，我们的目的是求得*Y*的后验概率：

 (4-1)

假设*Y*是给定的，则序列的似然函数如下：

 (4-2)

如前文所述，如果给定一个损失函数*H*，则*Y*的质心估计可以表示为：

 (4-3)

而*Y*的最大化后验概率估计可以表示为：



根据式(4-1)和(4-2)，可以比较容易的计算出*Y*的后验概率分布，从而得到，而的计算比较困难，下面通过定义合适的损失函数来简化质心估计的计算过程。

此处采用汉明损失来作为损失函数*H*，定义



其中是位置*i*的状态，当位置*i*为背景时，其值为0，当位置*i*为模体时其值为*Y-i+1*，即该位在模体中的相对位置。对于*h*的选择，这里采用对称Kullback-Leibler距离，即从位置*i*到位置*j*的信息熵加上从*j*到*i*的信息熵，





化简后代入*H*，可得：



可以看出，*H*是的函数。

为了进一步简化计算，现在把损失函数转化为收益函数，其中，表示时的值。所以，当时没有收益，。所以，根据收益函数的定义，以及等式(4-3)，有。*G*与*H*一样，也是的函数。

在以上的基础上，现在来证明：

 (4-4)

即质心估计是G和Y的后验分布的卷积。

可以直接根据等式(4.3)来推导：

**

**

**

**

得证。这样就可以方便的计算出。

### 4.2.3 单序列多模体

本节基于上一小节进行拓展，现在假设只有一条序列，但包含多个模体。此时，序列的似然函数与单模体时相似：

**

在上一节中，由于*Y*是标量，所以我们为*Y*设定了一个平均先验分布，此处*Y*变为向量，需要设定一个熵值较小的先验：令，有，然后，令，得到



其中，表示序列*R*中结合位点的最大值。

以上先验概率中，对于*c*(*Y*)的分布仍然无法确定，但是，我们可以采用一个Markov链模型，只有两种状态，背景和模体，其中背景和模体之间的状态转移概率（从任意一个状态到另外一个），以及起始状态为背景的概率均假设为*p*，这样可得

**

因为一条序列中有*c(Y)*条模体，所以就有*c(Y)*次从背景到模体的状态转移。通过*p* 可以对这个先验结构进行更灵活的控制：可以对*p*的取值设定一个超先验分布（参数的分布通常称为超分布），也可以直接根据模体数的期望*b*来得到*p*的合理取值；如果*n*远远大于*b*，那么，因为*c*(*Y*) 可以近似取均值为*n*(1-*p*)的泊松分布，所以是一个合理的选择。

然后，有后验概率

 (4-5)

以及，

 (4-6)

以及，

**

观察等式(4-5)和(4-6)，只要算出常量的值，就可以得到*c*(*Y*)和*Y*的条件后验概率，以及后验概率。

另外，在多模体序列中，我们仍然采用海明损失作为推断的损失函数（如第4.1节所述，质心估计是基于统计推断的）。但是此处因为每个样本点的长度都不相同，这为采用统一的海明损失造成了一定的困难。为此，可以对每个相同长度的样本组采用一个局部估计，然后使用下面的三角不等式：



其中*Y*是为任意一个长度为*c*的模体起始位置序列，是所有的长度为*c*的模体起始位置序列的局部质心估计，为全局质心估计，即最终的模体起始位置序列的估计。

有了以上的准备工作，下面我们提出多模体质心估计的求解策略。因为序列中的模体数是不确定的，所以向量的长度是无法确定的，质心估计要从整个解空间中的点中选出符合条件的解，而这些解是长度不同的向量，其长度*c*(*Y*)=1,...,*C*。称模体个数固定为*c*时的质心估计为局部质心估计，相应的，模体个数不定时的质心估计为全局质心估计，下面的策略可以用来求出全局质心估计：

1. 对于每个模体数*c*=1,2,…,*C*，找到其局部质心估计：

 (4-7)

给出，计算全局质心估计：

 (4-8)

下面先推导局部质心估计的计算方法。

定义配对海明损失：，其中，从其定义中可以看出，即是的上界。

下面根据等式(4-3)来推导配对局部质心估计（即以配对海明损失作为损失函数的质心估计）。

**定理1** 如果的边缘后验概率为，那么配对局部质心估计为：



证明：跟证明公式(4-4)的思路一样：





**





得证。

从定理1中同样可以看到卷积结构。在单序列单模体问题中我们可以根据这个卷积公式直接计算出估计值，但是现在有了后验联合概率，并且卷积的定义域要求更加严格。为了计算方便，定义为联合后验概率的概率，即有

 (4-9)

这个结构非常关键，使得我们可以通过动态规划来得到。

通过上面的方法得到局部质心估计后，下面推导全局质心估计。

局部质心估计已经可以大体的反映所有的结合位点配置形成的空间的后验概率分布情况，最终我们要从这些局部估计中选取一个能较好的代表整体空间的点。有了以上的原则，我们首先指定一个估计，像等式(4-8)一样使得条件海明损失达到最小。但这种方法仍然有一个弊端——像直接从全部空间中求质心估计（即不用先局部质心，后全局质心。）一样，空间内的点存在长度不同的问题。

求解全局质心估计时，可以从所有的模体起始位点配置所构成的解空间中搜索后验损失的期望最小的点。但这种暴力搜索的方法显然效率很低，启发式搜索是个不错的选择，可以采用一个简单的启发函数吧最终的全局估计限定在已经取得的局部估计之中，



而此处可以使用另一种更简单的策略，就是在已知局部估计的前提下，求序列中模体数量的估计，，然后令全局估计。这里对*c*的估计用的是最大化后验概率估计，因为它是一维的，所以两种估计的结果是相同的。

### 4.2.4 多序列多模体

在前两小节中，所有的讨论都是基于预先给定的，但是的初始值与算法最终发现的模体之间有着紧密的联系。所以如果一开始为给出了一个不合理的值，那么求出的模体也会与真实模体相差较远。的取值是至关重要的，而且也没有统一的经验值可供使用，所以需要通过采样的方法来获得的值。本文中使用吉布斯采样的方法，根据第三章中提出的此类算法的基本结构，吉布斯采样即为本算法的求解策略。

现在我们的数据集包含*m*条序列，这样模体起始位置就多了一个序列索引，。与前文相同，先假设*Y*和不是相互独立的，即有

 (4-10)

我们可以继续采用前文中的方法为每个序列分别计算其局部质心和全局质心估计。但是此处，先假设是随机的，假定它服从Dirichlet分布：

 (4-11)

本节应用吉布斯采样，先给出作为条件来采样*Y*，然后再用*Y*作为条件来采样，如此迭代多次直到采样的值收敛。我们通过吉布斯采样可以将前两小节的算法串联起来，形成完整的可以处理多序列多模体的模体发现算法。

算法：基于质心估计的模体发现算法

输入：

输出：，

1. （采样）。对于每条序列：令，然后采样。

（1）（初始化）设置前向和，然后，对于，如果，令。

（2）（计算前向和）对于以及：令

，

其中。

（3）（采样）对于：计算的边缘后验概率，



然后后根据来采样，。

（4）（采样）对于：采样，



**

1. （采样）对于计算，然后采样



# 第五章 面向ChIP-seq数据的应用

传统模体发现方法通常利用多条可能包含某个模体的同源序列作为输入，通过在这些序列中进行模式查找，或者利用相关模体的先验信息，通过统计学方法为该模体建模，随后利用输入数据不断修正优化模型，直到模型的统计学意义满足预先设定的阈值。

染色质免疫共沉淀（ChIP，Chromatin Immunoprecipitation）实验为研究蛋白质与DNA分子的相互总用提供了新的途径。将ChIP实验与高通量测序技术结 合起来（ChIP-seq，Chromatin Immunoprecipitation followed by Sequencing）可以高效的在全基因组范围内检测组蛋白或者转录因子与DNA分子相互作用的区域信息。通过ChIP-seq数据进行模体发现问题的研究已经成为近年来该领域的主流趋势。

## 5.1. ChIP-seq实验及ChIP-seq数据特征

### 5.1.1染色质免疫共沉淀实验

特定的DNA-蛋白质结合位点，可以用染色质免疫共沉淀（Chromatin Immunoprecipitation, ChIP）方法分离出来。ChIP实验的基本流程如下[45]：

1. 甲醛交联整个细胞系（组织），将目标蛋白质（例如转录因子）与染色质联结起来；
2. 分离基因组DNA，并用超声波将其打断成一定长度的小片段；
3. 添加与目标蛋白质特异的抗体，该抗体与目标蛋白质形成免疫沉淀结合复合体；
4. 去交联，纯化DNA即得到染色质免疫沉淀的DNA样本，准备测序。

通过以上步骤，可以得到大量的DNA片段，通常的长度为（0.2~1kbp），虽然这些片段中仍然存在着变异和重复，但这些片段中包含特定蛋白质（实验中采用的蛋白质，例如转录因子）结合位点的几率非常大。在ChIP实验完成后，这些大量的短片段被直接使用高通量测序技术进行测序。

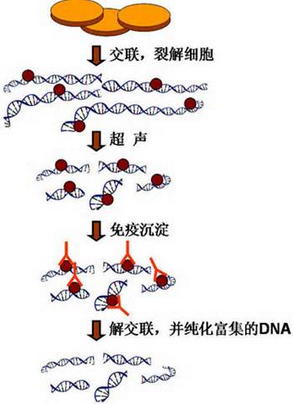


图5.1 ChIP实验示意图

### 5.1.2高通量测序技术及ChIP-seq数据特征

高通量测序技术（High-throughput sequencing）又称“下一代”（Next Generation Sequencing,NGS）测序技术[45]。能同时并行地对几十万到几百万条DNA分子进行序列测定。现有的高通量测序技术主要有：454焦磷酸测序(454 Pyrosequencing)、Illumina（Solexa）测序和ABI SOLiD测序等。虽然这些测序技术各有不同，不过其测序原理都遵循以下的流程：（1）样本准备；（2）文库构建；（3）测序反应；（5）数据分析。

从本质上说，ChIP-seq数据就是选用特定蛋白质进行ChIP实验后得到的染色质片段，在经过高通量测序后得到的序列片段。与传统模体发现算法中使用的启动子区序列相比，ChIP-seq数据具有如下特征：

1. 从序列数量上来说，ChIP-seq数据中包含的序列条数远远大于启动子区序列数[46]。具体的数量从几万条到几百万条不等，主要取决于ChIP实验中使用的染色质物种，以及与该染色质产生交联的具体蛋白质。例如文献[10]中用与ChIP-seq实验分析的人类基因转录因子GABP、SRF和NRSF，它们分别产生的序列片段为786221条、8721730条和8813398条。
2. 序列长度很短。由于高通量测序的局限，通常对染色质从5’端开始定序，而且只能对起始的大约25到50bp定序，这导致基于ChIP-seq数据的基因分析方法都只有这些短序列可供利用。
3. 包含控制集（对照集）。ChIP-seq实验中通常都包含一组控制集。控制集在后续的分析中主要有三种用途：1）通过与实验集的对照，去除实验集中的假阳性峰值（Peak），如图5.2[11]，图5.2中左侧被全出来的峰被认为是有实际意义的，因为控制集的对应部分并没有呈现出峰，而右侧的峰被认为是无效的，因为控制集的对应区域也呈现出峰； 2）对照集的一部分被用来充当ChIP实验的结果，在确定峰值阈值的时候，可以参考该部分对照集中标签（Tag，或者Read，指ChIP实验中产出的染色质被NGS定序的那一部分，通常位于5’断，可以认为是ChIP-seq数据中的一条序列）密度与实验集中标签密度的比值来定义峰值的阈值；3）对照集的另一部分被当做背景。与传统模体发现方法类似，在峰值之外的序列被认为是背景，在Peak-Calling步骤中，部分算法考虑到了背景的分布，这时被当做背景的对照集中的Tag密度分布可以被作为背景的先验信息用于建立统计模型。

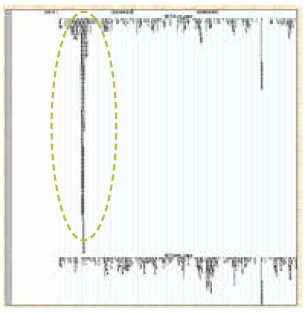


图5.2 实验集与控制集的tag密度分布情况。

1. 需要参照基因组。实际上在传统的模体发现算法和实验中，也是需要参照基因组的。参照基因组的主要作用是将用于模体发现的基因片段映射到基因组的特定位置中，从而也就把实验或者计算方法得到的模体从序列片段中的位置映射到了基因组中的绝对位置，使得验证模体发现的结果成为可能。而在基于ChIP-seq数据的模体发现中，将ChIP-seq数据映射到参考基因组中是首要的步骤，在此基础上，才能进行后续的分析和计算。将ChIP-seq映射回参考基因组可以使用Eland，SOAP，MAQ，BOWTIE等工具。

5. 结果集增大。对于模体发现问题而言，结果集即为算法找出的模体数量（或者峰的数量）。这一点并非ChIP-seq数据本身的特征，但由于ChIP-seq数据集的庞大序列基数，导致大部分Peak-Calling算法发现的峰有上千个之多，远高于传统模体发现算法所能发现的模体数量。另外，每个峰值区间的序列数量都有上千条，也远高于普通的传统模体发现算法所能处理的规模。

## 5.2 面向ChIP-seq数据的应用

基于ChIP-seq数据的模体发现通常可以分为3个阶段：寻长定位，峰值提取和模体发现。寻长定位是指将ChIP实验产生的短序列映射回参考基因组，这是所以基于ChIP-seq数据模体发现方法的基础工作，通常这一步工作可以使用Bowtie等工具完成；峰值提取是基于ChIP-seq数据模体发现方法最重要的阶段，目前这一阶段的任务已经得到了广泛的研究，BayesPeak、MACS、QuEST等工具都能够很好的完成很多物种上峰值提取的任务。而第三阶段的模体发现，大部分工具都直接采用MEME来完成，但MEME最初是设计用于解决同源序列上的模体发现，一旦序列数量增加到一定量级以上速度就会变的非常慢，另外，ChIP-seq数据序列长度很短，相比于常用于传统模体发现的同源序列，其模体与背景的比例有较大的差异。Bailey等人提出了DREME来解决专门面向ChIP-seq数据的模体发现问题，该方法在性能和准确性方面均比MEME有更好的表现，但该算法是基于模式匹配和搜索，由于模体长度增加，算法的搜索空间就会快速放大，所以DREME只能用于寻找4~8bps的短模体。

本节将上一章提出的基于质心估计的模体发现算法应用于ChIP-seq数据。由于ChIP-seq峰值区间落入的短序列数量巨大，直接作为算法的输入会严重影响算法的性能，而且这些短序列中会有很多冗余信息：首先，峰值区间可能同时包含正负向序列；其次，可能会有多条序列起始于同一个位点，但长度不同；最后，可能有多条序列终止于相同位点，但起始点不同。

为了使峰值区间的序列满足作为算法输入的要求，需要对序列数量进行精简。但减少序列数量，不可避免的会使序列集中包含的信息量减少，并且如果在选择过程中过滤掉了包含模体的序列而留下了大量只含有背景的序列，势必会影响最终模体发现的效果。由于ChIP-seq数据是来源于ChIP实验的结果，所以得到的序列本身大部分都是可以和模体成功结合的基因片段，除去在读长定位阶段引入的噪音信息，以及在测序时产生的错误和序列截断之外，序列集中含有的背景信息应该远小于模体信息。所以应该从最初的序列集中选择最具代表性的序列。

k-均值聚类算法是一种广泛使用的基于划分的聚类算法，把n个对象分为k个簇，以使簇内的对象具有最高的相似度。相似度通过定义对象间的距离来度量。k-均值聚类算法首先按照一定准则（默认是随机）选择k个对象代表k个簇的中心。对剩余的每个对象，根据其与各个簇中心的距离，将它付给最近的簇，然后重新计算每个簇的平均距离。这样一直迭代，直到准则函数收敛。目标函数的形态如下：

 (5-1)

它假设对象属性来自于空间向量，并且目标是使各个簇内部的方差总和最小。假设簇为。是簇中所有元素的重心，或者叫中心点。

算法的基本操作步骤如下：

**算法2**：k-均值聚类算法

**输入**：簇的数目*k*；包含*n*个对象的数据集*D*

**输出**：*k*个簇的集合

1. 从*D*中选择*k*个对象，作为初始簇中心，同时设为空。
2. **for** *x* **in** *D*
3. **do**
4. *minDist*
5. *dest* 0
6. **for** *i* **from** 1 **to** *k*
7. **do**
8. **if** *dist*(*x*, ) < *minDist*
9. **do** *minDist* *dist*(*x*, )
10. *dest* *i*
11. *add*(*Sdest*, *x*)
12. *update*(*Sdest*, )

其中第11行的*add*操作将元素*x*加入到簇*Sdest*中，第12行的*update*操作根据新加入的元素重新计算簇的中心的取值，而第8行的*dist*操作为计算元素*x*到第*i*个簇的中心的距离，这个距离根据应用场景的不同有不同的实现。

下面讨论如何对ChIP-seq数据中峰值区间的短序列进行k-均值聚类，以减少序列条数并尽可能的保存有用的模体信息。

首先，需要选择*k*个簇的初始中心点。初始中心的选择对最终聚类的效果有很大的影响，选择的原则是尽可能的让区间内的序列平均分布，如果最终只有一个有效簇（基数大于一定阈值的簇）产生，那聚类就是失败的。

假设峰值区间的长度为*L*，首先将区间分为*k*个子区间，然后将所有序列的起始位置映射回峰值区间，为每个子区间求出中心点，这里求中心点的方法与k-均值聚类算法中求簇的中心点的方法类似。设目标函数为



其中*xj*和均为中任意序列的起始点坐标，能使取得最小值的为子区间的中心，如果有多条序列以为起始点，则选中长度最长的一条。

选定*k*个簇的初始中心点后，就可以开始聚类，这里选择编辑距离作为序列之间的距离度量，之所以使用编辑距离，而不是海明距离，因为落入同一个峰值区间的短序列长度是不一定相同的。另外，聚类的目标函数与式（5-1）稍有不同，新的目标函数为



因为此处的距离都是正向的，所以不用再求平方。

下面给出算法2中的*update*(,)函数。

**算法3**：更新簇内中心点

**输入**：簇内元素集合，原中心点

**输出**：更新后的中心点

1. *dist*
2. *pos* 0
3. **for** *j* **from** 1**to** 
4. **do**
5.  
6. *sum*
7. **for** *k* **from** 1 **to** 
8. **do**
9. *sum* *sum* + *E*(,)
10. **if** *sum* < *dist*
11. **do**
12. *dist* *sum*
13. *pos* *j*
14.  

这个算法中编辑距离*E*(,)使用动态规划方法来计算，此处不再赘述。

经过以上的聚类过程之后，得到的*k*条序列就可以作为第四章提出的基于质心估计的模体发现算法的输入，从而找出模体。

# 第六章 算法测试与实验结果分析

本章将分别在传统生物数据和ChIP-seq数据集上对本文提出的算法进行测试，并且将于其他同类算法的结果进行对比分析。

## 6.1 模体准确性评价标准

本节首先定义一组性能系数，用于在已知模体时评价模体发现算法的结果。假设有工具*T*和数据集*D*，现在有*D*上的已知模体集和*T*在*D*上的预测模体集。对于*T*预测的准确度，我们可以按照以下的方法进行碱基水平和模体水平的评价，首先，定义：

*nTP*（真阳性碱基数）: 预测模体与已知模体重合部分的碱基数目。

*nFN*（假阴性碱基数）: 在已知模体中但未出现在预测模体中的碱基数目。

*nFP*（假阳性碱基数）: 在预测模体中，但不在已知模体中的碱基数目。

*nTN*（真阴性碱基数）: 既不在已知模体中，又不在预测模体中的碱基数目。

模体重合：如果模体*A*和模体*B*的重合碱基数大于等于模体*B*长度的四分之一，则称模体*A*是与模体*B*重合的。

*sTP*（模体真阳性）：与已知模体重合的预测模体个数。

*sFN*（模体假阴性）：与预测模体不重合的已知模体的个数。

*sFP*（模体假阳性）：与已知模体不重合的预测模体个数。

以下定义对于碱基级别(x=n)和模体级别(x=s)都是适用的：

*xSn*（敏感性）：*xSn*=*xTP*/(*xTP*+*xFN*)。敏感性给出了已知模体（或属于已知模体的碱基）占预测模体（或属于预测模体的碱基）的比例。

*xPPV*（阳性预测度）：*xPPV*=*xTP*/(*xTP*+*xFP*)。阳性预测度给出了给出了预测模体中已被证实的模体的比例，或者预测模体中被证实属于真实模体的碱基比例。在碱基水平上还可以定义：

*nSP*（确定性）：*nSP*=*nTN*/(*nTN*+*nFP*)。

最后，基于以上定义的众多测度，定义碱基层面综合的准确性测度：

*nPC* （准确度系数）：*nPC*=*nTP*/(*nTP*+*nFN*+*nFP*)，以及

*nCC*（相关系数）：以及，

*sASP*（模体水平平均准确度）：*sASP*=(*sSn*+*sPPV*)/2

需要注意的是，对于不同的测试集，以上评价标准并不都是可用的，例如在没有模体的序列中，如果工具*T*预测到了模体，那么*TP*+*FN*=0，这样的话*xSn*就取值无穷大，即是不可用的。遇到这种情况，或者选择其他有实际值的测度，或者测试多组数据并在这些数据上求平均值。

## 6.2 在同源基因序列上的测试

本节采用文献[5]建立的标准测试集。这组测试数据包含了真实的生物数据和依据不同模型建立的人造数据，目前已经被广泛应用于众多模体发现算法的测试中。表6.2给出了所选测试数据的详细参数。数据名称以’r’结尾表示改组数据来自于真实的基因调控区域；名称以’m’和’g’结尾的表示改组数据位人工生成的数据，并随即的植入模体。其中’m’表示数据的背景是根据3阶马尔科夫随机过程生成的，而’g’则表示背景时来自于同物种的若干随机选取的调控区域。

表6.1 测试数据详情

|  |  |  |
| --- | --- | --- |
| 物种/数据名称 | 序列条数/长度 | 模体实例个数/模体实例最大长度 |
| Fly/dm03m | 3/2000 | 9/17 |
| Fly/dm02r | 1/2000 | 5/13 |
| Fly/dm08m | 3/2000 | 0/- |
| Fly/dm05g | 3/2500 | 14/21 |
| Human/hm05r | 3/1000 | 11/32 |
| Human/hm03r | 10/1500 | 15/46 |
| Human/hm02r | 9/1000 | 11/36 |
| Human/hm01g | 18/2000 | 16/54 |
| Human/hm17g | 11/500 | 10/18 |
| Mouse/mus02r | 9/1000 | 12/33 |
| Mouse/mus06g | 3/500 | 5/18 |
| Mouse/mus03g | 5/500 | 9/29 |
| Mouse/mus11m | 12/500 | 15/27 |
| Yeast/yst02g | 4/500 | 5/23 |
| Yeast/yst08r | 11/1000 | 14/49 |
| Yeast/yst09g | 16/1000 | 13/19 |
| Yeast/yst10m | 5/1000 | 0/- |
| Yeast/yst06g | 7/500 | 7/26 |

以上测试集包含四个物种的数据，在本文的测试中，针对每个物种设置统一的参数，设置的原则是模体数量和模体长度都取本物种组内数据的最大值：果蝇模体个数*M*=14，模体长度*L*=21；人类*M*=16，*L*=54；老鼠*M*=15，*L*=33；酵母*M*=14，*L*=49。对于迭代次数，统一设为*k*=10000，而在实际测试中，果蝇的4组序列需要的迭代次数最多，平均需要大约8000次才能收敛，而其他数据则均在不到6000次迭代后就已收敛。对于矩阵，本实验中统一采用四种碱基的平均先验作为初始值。对于模体模型，本实验中设定每条序列中只包含一个模体实例。另外，本实验中对于每个组输入数据，限定只输出5个候选模体，按照其相对熵排序，而在以下结果分析中只选取5个候选模体中*nPC*值最大的模体作为结果进行测评。

算法在各组数据上得出的模体的Sequence Logo和相对熵如表6.2所示。图6.1给出了算法在各组数据上的*nPC*曲线，图6.2给出了算法在四个物种上的性能系数指标。

表6.2 算法在各组数据上得到的模体

|  |  |  |
| --- | --- | --- |
| 物种/数据名称 | Sequence Logo | 相对熵 |
| Fly/dm03m |  | 9.26 |
| Fly/dm02r | — | — |
| Fly/dm08m | — | — |
| Fly/dm05g | — | — |
| Human/hm05r | — | — |
| Human/hm03r |  | 15.27 |
| Human/hm02r |  | 9.28 |
| Human/hm01g | — | — |
| Human/hm17g |  | 16.38 |
| Mouse/mus02r | — | — |
| Mouse/mus06g |  | 9.3 |
| Mouse/mus03g |  | 12.71 |
| Mouse/mus11m |  | 11.24 |
| Yeast/yst02g | — | — |
| Yeast/yst08r |  | 9.44 |
| Yeast/yst09g |  | 11.72 |
| Yeast/yst10m |  | 11.38 |
| Yeast/yst06g | — | — |

表6.2中展示了本实验在Tompa数据集的部分数据上的结果模体，并且给出了求出的每个模体的相对熵，然而相对熵的值并不能说明所求出的模体的准确性，只表明了算法找出的所有模体实例上序列的保守性。从Sequence Logo上也可以看出，找出的模体中有一部分并不具有很好的保守性，例如在Yeast/yst09g上找出的模体中，第2个位点几乎是完全不保守的，而在Yeast/yst10m上找到的模体中，第12、13、15个位点也完全不保守。在表6.2中标识为‘—’未找到模体。在本实验中，如果得到的结果模体集中，超过模体长度一半的位点的相对熵小于0.1，则认为这个结果模体是无效的，因为这样的模体是不具有保守性的。

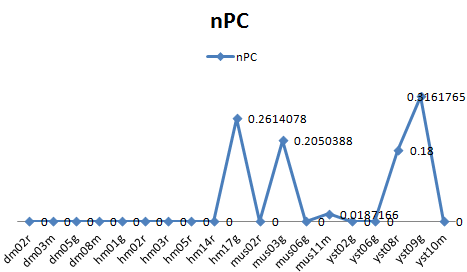


图6.1 算法在测试集上的nPC曲线

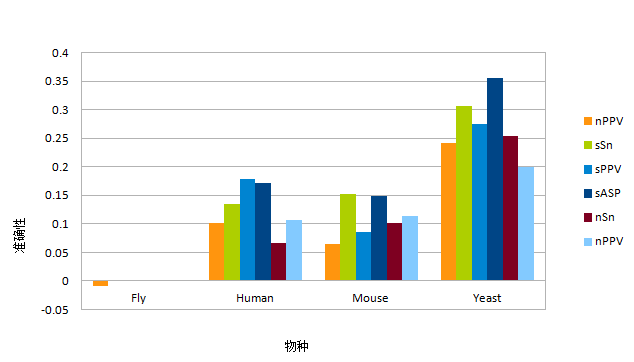


图6.2 算法在4个物种上的预测准确性

从以上的实验数据中可以看出算法在植入模体的模拟数据上准确性要高于真实生物数据，这是因为在本算法的实现中，所有的序列背景都是按照1阶马尔科夫模型实现的，而实际的生物数据中背景模型是不可知的。另外，由图6.2可知，算法在酵母上的准确性要高于其他3个物种，因为酵母为单细胞真菌，生物结构和基因序列中的背景模型相比其他3个物种都更加简单；而果蝇数据集上的模体预测准确性远低于人类和老鼠，这是不能用以上原理解释的。观察测试数据集，可以看出果蝇的4组数据中序列长度与模体长度的比值是最大的，因为本实验中假定所有的序列都只包含一个模体实例，所以根据4.2.4节建立的模型，当*n/L*很大时，模体起始位点数量*c*(*Y*) 的取值范围大大增加，导致在这个区间中采样到正确的Y的概率相应减小，从而更难找到正确的模体。

## 6.3 在ChIP-seq数据上的测试

由于本文的研究重点在ChIP-seq数据分析的第三个阶段，所以数据集的准备需要在准备原始的ChIP-seq数据的基础上先完成前两个阶段。本实验中选用的实验数据来自于Chen等人在2008年发布的公共测试数据[47]。数据详细描述见表6.3。

表 6.3 实验数据详细参数

|  |  |
| --- | --- |
| NCBI编号 | GSE11431 |
| 发布时间 | 2008-06-12 |
| 实验平台 | GPL9185 Illumina Genome Analyzer |
| 生物组织 | 老鼠 胚胎干细胞 （Mus musculus） |
| 参照基因组 | Mouse July 2007 (NCBI37/mm9) |
| 实验采用的转录因子 | Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1, CTCF |
| 实验类型 | ChIP-seq |

本实验中只使用了数据集中针对转录因子Oct4生成ChIP-seq数据。因为已知Illumina工具在序列尾部会产生更多的错误，所以在实验之前要先对原始数据的数据质量进行验证。

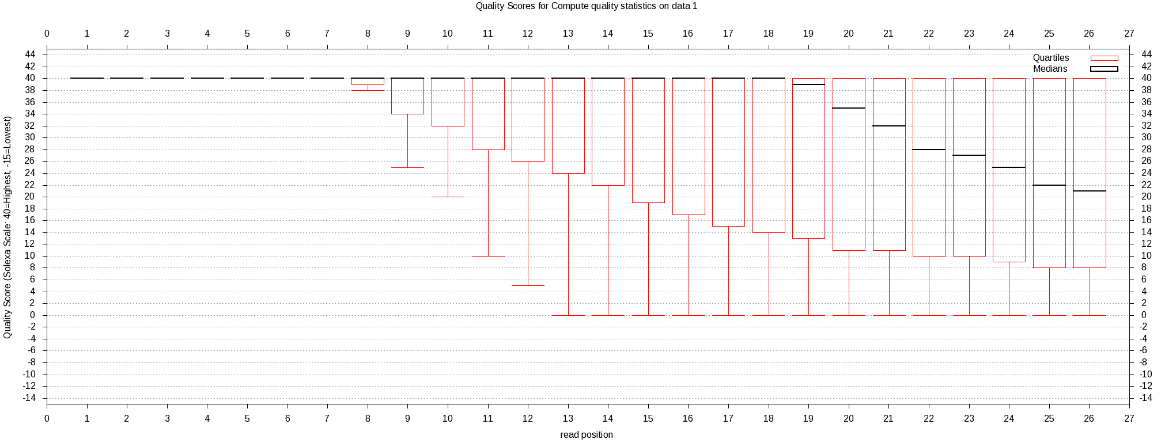
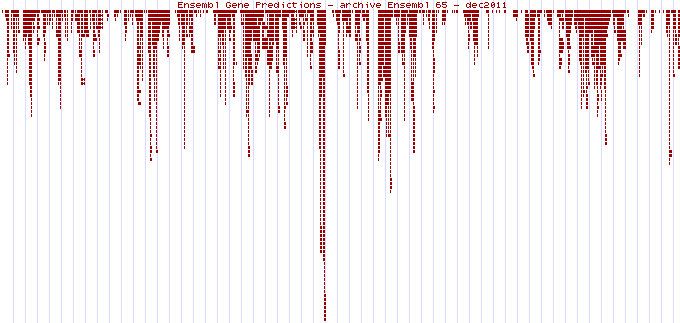


图6.3 Oct4标签质量箱线图

图6.3给出了Oct4标签的序列质量箱线图。Oct4标签的平均长度为26bp。该图由NGS标准工具集产出。序列的得分为0~40分，黑色横线表示位点上的平均得分，红色横线表示得分的浮动区间，由图6.3可以看出，在标签尾部得分确实有一定的下降。

在对标签质量进行检测之后，本实验使用Bowtie将标签映射回参照基因组，并得到标签在基因组的坐标。随后根据映射好的数据进行峰值提取。本实验在Oct4上共产生了1886个峰，图6.4展示了这其中的一部分。图6.4 b)是对a)中的一个峰值放大后的结果。该图使用UCSC Genome Browser生成。



a)



b)

图6.4 Oct4上的部分峰值示例

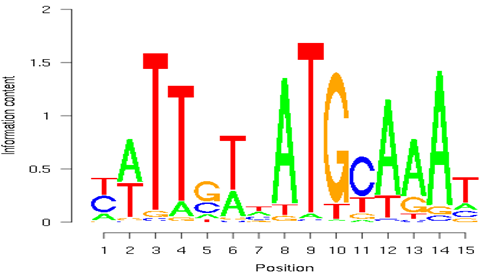
在得到峰值区间的坐标之后，本实验只选择了每种模体上得分最高的一个峰进行处理。对峰值的处理方法为确定峰值中心坐标之后，向峰值中心左右各扩展200bp，整个峰值区间为400bp，然后将该区域内所有的标签作为k-均值聚类算法的输入。

聚类的目的是将距离相近的元素尽可能的归为一类，其评价的标准是簇中元素相似度越高越好，而簇间差异越大越好。此处使用一下聚类评价函数来评价聚类效果：

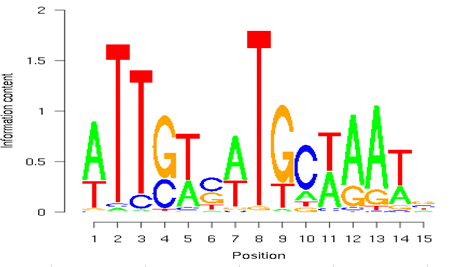


其中表示每个簇的中心元素到其他元素的编辑距离和，这个值越小，说明簇的内聚性越好；表示*k*个簇的中心元素间的编辑距离之和，这个值越大，说明簇间的差异越大。综上所述，*Q*的值越大，则聚类的效果越好。

本实验中*k*=100，序列平均长度为26，在选定的峰值区间，共有10087个标签，通过聚类最终得到100条序列作为4.2.4节算法的输入。最终得到的模体如图6.5 a)所示，图6.5 b)给出了Oct4的已知模体的Sequence Logo。



a)



b)

图6.5 实验找到的模体与已知模体Sequence Logo对比

由图6.5可见，本实验成功的找出了转录因子Oct4的模体，而且效果比在同源序列上要好很多。主要原因有两点：1）因为基于ChIP-seq数据的序列分析在模体发现进行之前已经进行了大量的工作，尤其是在峰值提取阶段已经大大缩小了模体出现的区间，并且通过峰的选择，可以将大部分与模体无关的标签过滤掉，这些都是ChIP-seq数据的优势所在，具有很强的模体针对性，通过生物实验，可以将与模体相关度更高的信息保留下来，从而让后续的模体发现工作更容易进行；2）如6.2节对果蝇数据测试结果的分析，如果序列长度与模体长度的比值越小，或者说模体相对序列来说越长，那么找到模体的可能性就越大。在本节的实验中，序列长度平均为26，而模体长度只有15，这个很小的比例也是最终能够找到准确模体的重要原因。

# 第七章 总结

尽管模体发现算法已经被广泛的注意和研究，但通过对众多模体发现算法的研究，我们认为现有的模体发现方法仍然面临着以下的困境：

1. 生物学上对于模体构成的认识仍然比较缺乏，导致可以用来为模体建立数学模型的先验知识比较匮乏，现有的模体模型尽管不断的在数学上完善，但难以保证更具数学上优秀特征的模体模型会具有更好的生物学意义。

2. 模体发现算法的预测准确性难以验证。首先由于生物物种的多样性，不同物种的DNA序列构成和复杂程度差异较大，同一个模体发现工具面临不同物种的基因调控序列，预测的准确性差异非常大。其次，建立用于测试的公共数据非常困难，这主要是由于现在发现证实的模体数量有限，很难确保公共数据中没有未知模体，所以无法未被测工具提供完全准确的“答案”，这样如果某个工具在这组数据上找到了未知的模体，将因此被罚分。

3. 模体发现工具的使用是非常困难的。基于统计的模体发现算法都是具有不确定性的，每个工具运行多次会产生多个不同的结果；而每次运行一搬都会产生多个结果。如何从这些众多的结果中选取可以采纳的预测并没有可供参考的可靠原则。这使得使用这些工具去预测未被证实的模体变的比较困难。

本文的主要贡献在于：1）从结构和模型选择上分析对比了多种基于统计推断的模体发现方法；2）提出了基于质心估计的模体发现方法并进行了全面的测试。3）将提出的算法成功的应用到ChIP-seq数据集上，并获得了较好的效果。

另外，本文还存在一些不足，例如算法对于短模体的效果并不好（指序列长度与模体长度的比值过大）；算法需要指定模体长度，而这在有些情况下是无法做到的；另外，在ChIP-seq数据上测试时程序的运行速度较慢。

针对本文的不足，还需要进行多方面的改进，例如改进算法模型，取出算法结果对于模体长度的依赖；增加模体长度自适应策略，根据数据特征来自动学习适当的模体长度；另外在算法性能方面，可以改进聚类的方法，使得聚类阶段得到的序列更接近模体，从而减小后续算法的迭代次数。

# 致谢

首先非常感谢我的导师霍红卫教授，本篇论文从选题到最终完稿，期间的每个环节都离不开霍老师的悉心指导和启发。霍老师严谨的治学态度，持之以恒，精益求精的科研精神一直深深地影响着我。在此谨向霍老师致以诚挚的谢意。

十分感谢徐学洲老师、于强博士、张懿璞博士和郭鸿志老师，他们在我的论文撰写过程中提出了很多宝贵的意见和建议。感谢我的同学乔科、柴明辉、石峰、郭敏、柳娜娜和李龙，他们在我平时的生活和学习中提供了诸多帮助。感谢同实验室的陈龙刚、李双江、赵恒、聂绎静、赵玉豪、王哲、陈晓阳、赵睿醒，他们为实验室营造了良好的科研氛围，并在我的论文撰写过程中提供了诸多帮助。

感谢我的父母和姐姐，他们在我的生活和学习中给予了无微不至的关怀和照顾，在我遇到困难和挫折时给予无尽的鼓励和帮助，衷心地感谢他们的无私付出。

感谢所有关心、支持和帮助过我的老师、同学和朋友们。

最后，对所有参加论文评审的各位专家、教授表示衷心的感谢。

# 参考文献

[1] Korn, Lawrence J., Cary L. Queen, and Mark N. Wegman. "Computer analysis of nucleic acid regulatory sequences." Proceedings of the National Academy of Sciences 74.10 (1977): 4401-4405.

[2] Queen, Cary, Mark N. Wegman, and Laurence Jay Korn. "Improvements to a program for DNA analysin: a procedure to find homologies among many sequences." Nucleic Acids Research 10.1 (1982): 449-456.

[3] Stormo, Gary D., et al. "Use of the ‘Perceptron’algorithm to distinguish translational initiation sites in E. coli." Nucleic Acids Research 10.9 (1982): 2997-3011.

[4] Staden, Rodger. "Computer methods to locate signals in nucleic acid sequences." Nucleic acids research 12.1Part2 (1984): 505-519.

[5] Tompa, Martin, et al. "Assessing computational tools for the discovery of transcription factor binding sites." Nature biotechnology 23.1 (2005): 137-144.

[6] Das, Modan, and Ho-Kwok Dai. "A survey of DNA motif finding algorithms." BMC bioinformatics 8.Suppl 7 (2007): S21.

[7] Sandve, Geir Kjetil, and Finn Drablos. "A survey of motif discovery methods in an integrated framework." Biol Direct 1.11 (2006).

[8] Hu, Jianjun, Bin Li, and Daisuke Kihara. "Limitations and potentials of current motif discovery algorithms." Nucleic acids research 33.15 (2005): 4899-4913.

[9] Spyrou, Christiana, et al. "BayesPeak: Bayesian analysis of ChIP-seq data." BMC bioinformatics 10.1 (2009): 299.

[10] Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, Nussbaum C, Myers R, Brown M, Li W, Liu X: Model-based analysis of ChIP-Seq (MACS). Genome Biology 2008, 9(9):.

[11] Valouev, Anton, et al. "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." Nature methods 5.9 (2008): 829-834.

[12] Bailey, Timothy L. "DREME: motif discovery in transcription factor ChIP-seq data." Bioinformatics 27.12 (2011): 1653-1659.

[13] Bailey, Timothy L., and Charles Elkan. "Fitting a mixture model by expectation maximization to discover motifs in bipolymers." (1994): 28-36.

[14] Lawrence, Charles E., et al. "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." SCIENCE-NEW YORK THEN WASHINGTON- 262 (1993): 208-208.

[15] Bailey, Timothy L., et al. "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic acids research 34.suppl 2 (2006): W369-W373.

[16] Roth, Frederick P., et al. "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." Nature biotechnology 16.10 (1998): 939-945.

[17] Thompson, William, Eric C. Rouchka, and Charles E. Lawrence. "Gibbs Recursive Sampler: finding transcription factor binding sites." Nucleic Acids Research 31.13 (2003): 3580-3585.

[18] Liu, Falcon FM, et al. "FMGA: finding motifs by genetic algorithm." Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on. IEEE, 2004.

[19] Gelman, Andrew, and Donald B. Rubin. "Avoiding model selection in Bayesian social research." Sociological Methodology 25 (1995): 165-174.

[20] Tanner, Martin A., and Wing Hung Wong. "The calculation of posterior distributions by data augmentation." Journal of the American statistical Association 82.398 (1987): 528-540.

[21] Schneider, Thomas D., and R. Michael Stephens. "Sequence logos: a new way to display consensus sequences." Nucleic acids research 18.20 (1990): 6097-6100.

[22] Kim, Nak-Kyeong, Kannan Tharakaraman, and John L. Spouge. "Adding sequence context to a Markov background model improves the identification of regulatory elements." Bioinformatics 22.23 (2006): 2870-2875.

[23] Benos, Panayiotis V., Martha L. Bulyk, and Gary D. Stormo. "Additivity in protein–DNA interactions: how good an approximation is it?." Nucleic acids research 30.20 (2002): 4442-4451.

[24] Zhou, Qing, and Jun S. Liu. "Modeling within-motif dependence for transcription factor binding site predictions." Bioinformatics 20.6 (2004): 909-916.

[25] Stormo, Gary D., Thomas D. Schneider, and Larry Gold. "Quantitative analysis of the relationship between nucleotide sequence and functional activity." Nucleic acids research 14.16 (1986): 6661-6679.

[26] Barash, Yoseph, et al. "Modeling dependencies in protein-DNA binding sites." Proceedings of the seventh annual international conference on Research in computational molecular biology. ACM, 2003.

[27] Lim, Lee P., and Christopher B. Burge. "A computational analysis of sequence features involved in recognition of short introns." Proceedings of the National Academy of Sciences 98.20 (2001): 11193-11198.

[28] Cawley, S. "Statistical models for DNA sequencing and analysisspliceosome: motors, clocks, springs, and things. Cel 1, Statistical models for DNA sequencing and analysis." PhD thesis (2000).

[29] Zhao, Xiaoyue, Haiyan Huang, and Terence P. Speed. "Finding short DNA motifs using permuted Markov models." Journal of Computational Biology 12.6 (2005): 894-906.

[30] Ben-Gal, I., et al. "Identification of transcription factor binding sites with v ariable-order Bayesian networks." Bioinformatics 21.11 (2005): 2657-2666.

[31] Xing, Eric P., et al. "A hierarchical Bayesian Markovian model for motifs in biopolymer sequences." Advances in Neural Information Processing Systems 15 (2002): 1489-1496.

[32] Liu, X., Douglas L. Brutlag, and Jun S. Liu. "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput. Vol. 6. 2001.

[33] Donaldson, Ian John, Michael Chapman, and Berthold Göttgens. "TFBScluster: a resource for the characterization of transcriptional regulatory networks." Bioinformatics 21.13 (2005): 3058-3059.

[34] McCue, Lee Ann, et al. "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes." Nucleic acids research 29.3 (2001): 774-782.

[35] Pudimat, Rainer, Ernst-Günther Schukat-Talamazzini, and Rolf Backofen. "Feature based representation and detection of transcription factor binding sites." Proc. German Conference on Bioinformatics. 2004.

[36] Baldi, Pierre, et al. "Hidden Markov models of biological primary sequence information." Proceedings of the National Academy of Sciences 91.3 (1994): 1059-1063.

[37] Thijs, Gert, et al. "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes." Journal of Computational Biology 9.2 (2002): 447-464.

[38] Shida, Kazuhito. "GibbsST: a Gibbs sampling method for motif discovery with enhanced resistance to local optima." BMC bioinformatics 7.1 (2006): 486.

[39] Xing, Eric P., et al. "Logos: a modular bayesian model for de novo motif detection." Journal of Bioinformatics and Computational Biology 2.01 (2004): 127-154.

[40] Ao, Wanyuan, et al. "Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR." Science Signaling 305.5691 (2004): 1743.

[41] Sinha, Saurabh, Mathieu Blanchette, and Martin Tompa. "PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences." BMC bioinformatics 5.1 (2004): 170.

[42] Prakash, Amol, et al. "Motif discovery in heterogeneous sequence data." Pac Symp Biocomput. Vol. 9. 2004.

[43] Liu, Derong, et al. "Motif discoveries in unaligned molecular sequences using self-organizing neural networks." Neural Networks, IEEE Transactions on 17.4 (2006): 919-928.

[44] Carvalho, Luis E., and Charles E. Lawrence. "Centroid estimation in discrete high-dimensional spaces with applications in biology." Proceedings of the National Academy of Sciences 105.9 (2008): 3209-3214.

[45] Mardis ER: ChIP-seq: welcome to the new frontier. Nature Methods 2008, 4(8):613-614.

[46] Kim TH, Ren B: Genome-wide analysis of protein-DNA interactions. Annual Review of Genomics and Human Genetics 2006, 7:81-102.

[47] Chen X, Xu H, Yuan P, Fang F et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 2008 Jun 13;133(6):1106-17.