

基于吉布斯采样的 TFBS 识别算法研究^{*}

何彩升^{1,2} 戴宪华¹ 向倩¹ 王江¹ 邓泱泱¹ 戴智明¹

(中山大学电子与通信工程系 广州 510275)¹

(华南肿瘤学国家重点实验室 中山大学肿瘤防治中心 广州 510060)²

摘要 计算机方法识别转录因子结合位点(TFBS,也称“模式”)是目前生物信息学的一个很有吸引性和挑战性的课题。吉布斯采样识别模式的算法本质上是一个启发式搜索方法,容易陷入非全局最优的局部最大值。为此,提出了一种改进的吉布斯采样策略 YGMS(Yeast Gibbs Motif Sampler)来识别酿酒酵母共表达基因调控区域转录因子结合位点。在酵母的共调控基因序列的数据集测试中,YGMS 比其他几个基于吉布斯采样算法更有效地识别出真实模式序列,在一定程度上提高了算法的性能。

关键词 生物信息学,吉布斯采样,转录因子结合位点

Research of Algorithm for Finding TFBS Based on Gibbs Sampling

HE Cai-Sheng^{1,2} DAI Xian-Hua¹ XIANG Qian¹ WANG Jiang¹ DENG Yang-Yang¹ DAI Zhi-Ming¹

(Department of Electronic & Communication Eng., School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275)¹

(State Key Laboratory of Oncology in Southern China, Cancer Center, Sun Yat-sen University, Guangzhou 510060)²

Abstract Computational methods detecting the transcription factor binding sites (TFBS) remain one of the most intriguing and challenging subjects in bioinformatics. Gibbs sampling is essentially a heuristic method, and it is easy to trap into a non-optimal “local maximum”. To overcome it and to improve the performance of the algorithm, an improved Gibbs sampling strategy YGMS (Yeast Gibbs Motif Sampler) for finding motifs in gene sequences of yeast is present. YGMS and other existing Gibbs sampling algorithms were tested on real biological data sets with yeast regulatory elements. The results show that YGMS has better performance than other Gibbs sampling methods to a great extent in accuracy and sensitivity of finding true motifs.

Keywords Bioinformatics, Gibbs sampling, Transcription factor binding sites

1 前言

随着生物学和相关技术的发展,全基因组学研究取得了蓬勃发展,同时快速产生了海量的实验数据。目前,生物信息学研究重点已从积累数据转移到如何解释这些数据,并从数据中挖掘潜在的知识。采用有效的算法识别真核生物转录因子结合位点是其中一个热点研究课题。

基因表达调控是进化分子生物学的一个重要研究方面。大部分的基因表达调控是发生在转录水平上。在真核生物中,三种主要的 RNA 聚合酶均有自己的特异性启动子类型,通过识别和结合其相应的启动子区域,引导完整的 RNA 聚合酶及其他调节因子(主要包括转录因子)结合形成转录起始复合物,从而开始转录活动。转录因子具有结构方面的特征,使其能够识别和结合特定的 DNA 序列^[1]。转录因子结合位点(Transcription Factor Binding Sites, TFBS),经常是启动子区域中的一小段长度约为 5~20bp 的核苷酸序列,这种结构模式常称模式或 Motif。生物学功能分析揭示,主要 TFBS 的变异相对缓慢,因此,模式序列在结构上相对保守。同时,相似表达图谱的共调控基因组在转录调控机制上也高度相似^[2]。

计算机方法识别转录因子结合位点是目前生物信息学的一个很有吸引性和挑战性的课题。最近几年来,很多学者一直都在研究识别 TFBS 的计算机方法。这些方法,总的来可分为两大类:一类是基于字串的方法。这种方法主要是通过多联核苷酸短序列的计数和频率的统计来识别,其中最常用的方法是一致性序列模型(Consensus Model,或称共有序列)。第二类是基于概率序列模型的方法,如期望最大化(Expectation Maximization, EM)和吉布斯采样(Gibbs Sampling)方法。由于吉布斯采样方法的敏感性和快速收敛特性,该方法是目前启发式识别序列模式的较有效方法之一。

2 吉布斯采样识别模式算法

先简要地介绍一下模式的建模方法。假设有一个包含 N 个 DNA 序列的集合 S 。每个序列包含有且仅有一个长度为 W 的模式。为了识别模式,可将序列分成两个概率模型:一个为背景模型。假设除了模式序列外的序列是由核苷酸依据离散分布 $\theta_0 = (q_0^A, q_0^C, q_0^G, q_0^T)^T$ 随机产生的;另一个为模式模型。假设模式是由不同的离散分布 q_i^j 在相应位置上独立产生,如式(1)所示。这个模式模型常称为位置权重矩阵(Position Weight Matrix, PWM)。计算 PWM 时,首先依据模式

^{*} 基金项目:国家自然科学基金项目(60474075)。何彩升 工程师,博士研究生,主要研究方向为生物信息处理;戴宪华 教授,博士,博士生导师,主要研究方向为无线通信、生物信息处理,非线性信号处理;向倩 讲师,博士,主要研究方向为生物信息处理,信号处理;王江、邓泱泱、戴智明 博士研究生,主要方向为生物信息处理,信号处理。

在各序列的相应位置的核苷酸频数产生一个位置频数矩阵 (Position Frequency Matrix, PFM), 然后利用式(2)将 PFM 转为 PWM。其中, f_i^b 是模式模型中第 i 个位置上碱基 b 的出现频数, N 为模式的序列个数, $s(b)$ 为计数函数。

$$\theta_w = \begin{pmatrix} q_1^A & q_1^C & \cdots & q_1^W \\ q_2^A & q_2^C & \cdots & q_2^W \\ \vdots & \vdots & \ddots & \vdots \\ q_n^A & q_n^C & \cdots & q_n^W \end{pmatrix} \quad (1)$$

$$q_i^b = \frac{f_i^b + s(b)}{N + \sum_{b' \in \{A, C, G, T\}} s(b')} \quad (2)$$

吉布斯采样,也叫热浴(heatbath)方法,是一种马尔可夫链蒙特卡罗方法(Markov Chain Monte Carlo, MCMC)。在吉布斯采样过程中,依据所有其他变量的当前值,迭代地对其中每一个变量进行采样。从 $(x_1^1, \dots, x_n^1)$ 开始,

1. 依据 $P(X_1 | x_2^1, x_3^1, \dots, x_n^1)$ 选取 x_1^{1+1} 的值;
2. 依据 $P(X_2 | x_1^{1+1}, x_3^1, \dots, x_n^1)$ 选取 x_2^{1+1} 的值;
3.
- n . 依据 $P(X_n | x_1^{1+1}, x_2^{1+1}, \dots, x_{n-1}^{1+1})$ 选取 x_n^{1+1} 的值。

据此,依次对变量进行循环操作。可以证明,吉布斯采样算法最终可以得到一个平稳分布^[3]。

吉布斯采样最早是由 Lawrence 等引入蛋白质序列中模式的识别^[4]。其基本算法详见文[4]描述。由于吉布斯采样的敏感性和快速收敛特性,该算法已成为生物序列模式的识别和提取的主要算法之一。

最近几年,陆续提出了一些改进的基于吉布斯采样的算法,如 AlignACE^[5], BioProspector^[6], GLAM^[7], Motif sampler^[8], SeSiMCMC^[9], Gibbs Recursive Sampler^[10] 等。其中, AlignACE 算法返回输入序列集中的多个候选模式集合。该算法采用 MAP(maximum a priori log likelihood)得分函数; BioProspector 应用 0 至 3 阶的马尔可夫背景模型进行采样,并定义得分阈值范围来加速算法的收敛;在 GLAM 中,自动优化模式序列长度;Motifs Sampler 是采用一个高阶马尔可夫背景模型来实现在有噪声的数据集中采样;SeSiMCMC 在算法的更新步骤中采用贝叶斯方法来识别中间可能含有空隙型的两个重复或互文结构的模式;Gibbs Recursive Sampler 是在文[4]和[12]的算法基础上应用(rigorous)贝叶斯方法加以改进的。这些改进都在不同程度上提高了算法的模式识别性能。

吉布斯采样算法,本质上是一个启发式的方法,不是穷尽搜索的方法。因此,这些算法是启发式地去搜索一个数学上的最大值,但却容易陷入非全局最优的局部最大值。为了解决这个问题,经常需要多次运行算法,通过从多次独立的随机初始位置运行算法来寻找全局最佳值。

为了克服以上问题,提高算法的敏感度和准确度,我们提出了一种改进的吉布斯采样算法 YGMS(Yeast Gibbs Motif Sampler)来识别酿酒酵母的共表达基因调控区域的转录因子结合位点。

3 YGMS 算法的得分函数

信息含量(Information Content)是采用信息论中不确定性的降低来描述模式的保守性。在较多识别算法中,常用信息含量作为主要的得分函数,如式(3)所示。其中, q_0^b 是碱基 b 在背景模型 θ_0 中的出现频率, q_i^b 则是模式模型 θ_w 中第 i 个位置上碱基 b 的出现频率, W 为模式序列的长度。

$$IC = \sum_{i=1}^W \sum_{b \in \{A, C, G, T\}} q_i^b \log_2 \frac{q_i^b}{q_0^b} \quad (3)$$

虽然,信息含量能够很有效地衡量模式保守程度,但在真核生物中,模式的长度常为 5~10 个碱基,利用启发式算法也可能识别到基因序列中高度保守的“假模式”序列,而非高度保守的真实模式序列却未能有效识别。为此,我们引入一个新的得分函数,以提高算法识别真实模式的概率。

首先假定,生物序列可以描述为一个符合马尔可夫链随机过程的输出,即对于一个特定的核酸,在位置 p 出现的概率依赖于前面 k (在这只考虑 $k=1$ 的情况)个位置的核酸。因而模式的识别问题可表达为:在一个长度为 L 的 DNA 序列 $s = s_1, s_2, \dots, s_L$ 中,判别 s 是否为模式序列。对于 $A = \{A, C, G, T\}$ 中的任何两个字符 a, b , 定义其转换概率为 $t_{ab} = t_{i-1, i} = p(s_i = b | s_{i-1} = a)$ 。定义序列的转移矩阵为 $T = \{t_{ab}\}$ 。假设 $\{x_i\}$ 是一个随机过程,随机变量 x_i 的取值仅依赖于 x_{i-1} , 对于所有 $s_1, s_2, \dots, s_i \in A$, 有

$$\begin{aligned} p(x_i = s_i | x_1 = s_1, x_2 = s_2, \dots, x_{i-1} = s_{i-1}) \\ = p(x_i = s_i | x_{i-1} = s_{i-1}) \\ = t_{s_{i-1}, s_i} \end{aligned} \quad (4)$$

则整个序列 s 的发生概率为 $P(s) = \prod_{i=1}^L t_{s_{i-1}, s_i}$ (设 $s_0 = \text{"Begin"}$, $s_{L+1} = \text{"End"}$ 两个特殊字符)。

设 t_{ab}^+ 为模式模型的字符转换概率, t_{ab}^- 为背景模型的字符转换概率,则序列 s 的对数似然得分函数如式(5)。

$$Score(s) = \log_2 \frac{p(s | \text{motifs})}{p(s | \text{background})} = \sum_{i=1}^L \log_2 \frac{t_{s_{i-1}, s_i}^+}{t_{s_{i-1}, s_i}^-} \quad (5)$$

结合式(3)和式(5),形成新的得分函数 $Score(M)$ 如式(6)所示。其中, N 为序列集中具有模式的序列个数。

$$Score(M) = \sum_{i=1}^W \sum_{b \in A} q_i^b \log_2 \frac{q_i^b}{q_0^b} + \left(\sum_{j=1}^N \sum_{i=1}^L \log_2 \frac{t_{s_{i-1}, s_i}^+}{t_{s_{i-1}, s_i}^-} \right) / N \quad (6)$$

我们从 SCPD 和 TRANSFAC 中收集已经生物实验证明的酵母模式 PWM 共 26 个,形成如表 1 的酵母模式序列一阶转移概率矩阵。同时,根据文[11]的酵母基因组两联核苷酸频率数据计算形成如表 2 的酵母背景序列的一阶转移概率矩阵。

表 1 酵母模式序列转移概率矩阵

	A	C	G	T
A	0.305405	0.261070	0.204366	0.229159
C	0.228878	0.292705	0.265077	0.213340
G	0.350233	0.166240	0.264768	0.170139
T	0.220836	0.276772	0.238538	0.314958

表 2 酵母背景序列转移概率矩阵

	A	C	G	T
A	0.367444	0.160294	0.160294	0.300295
C	0.332960	0.185965	0.162129	0.318946
G	0.318289	0.198667	0.185881	0.297164
T	0.281623	0.171623	0.179453	0.367302

4 YGMS 算法描述

YGMS 算法在文[4]的吉布斯采样基本算法的基础上进行了几方面的改进:

(1)传统的采样方法均是完全随机选择序列采样的起始位点。在 YGMS 中,根据不同位点的 $Exp(Score(s))$ 权重,按

照轮盘赌原理选取起始位点。同时,设定一个阈值如 $Score(s) > 1$ 来提高算法的计算效率。

(2)算法采用同时多列并行采样的方法。在每次吉布斯采样迭代后进行位置权重矩阵更新后进入下一循环迭代。更新的方法为:设共有 C 列进行并行采样, $\theta_{i,w}$ 和 $Score(M_i)$ 分别为每列采样模式的位置权重矩阵和得分。更新后的位置权重矩阵 $\theta_{i,w}$ 如式(7)所示,其中, α 为一系数, $0 < \alpha < 1$ 。在 YGMS 中,设 $\alpha = 0.5$; p_i 的计算如公式(8)所示。

$$\theta'_{i,w} = (1-\alpha)\theta_{i,w} + \alpha \sum_{i=1}^C p_i \theta_{i,w} \quad (7)$$

$$p_i = \exp(score(M_i)) / \sum_{i=1}^C \exp(score(M_i)) \quad (8)$$

(3)算法还加强了相移(phase shift)。在每隔 M 次吉布斯采样后根据模式序列的得分值相应向左或向右进行相移。选择相应位点的相移权重为 $Exp(Score(M))$ 。

通过以上几个改进,结合生物学的先验统计知识,有效地提高了算法识别敏感度和准确率,大大提高了算法的性能。

5 测试数据和结果

为了测试算法的性能,我们从 SCPD 获取酵母的 9 个不同共调控基因序列数据集。数据集的概况包括调控因子、基因序列数目和模式的共有序列等,如表 3 所示。

表 3 测试数据集的概况。所有基因上游序列长度均为 800bp

Dataset Name	Regulator	Number of sequences	Consensus
Dataset1	BAS1	4	TGACTC
Dataset2	GCN4	9	TGANTN
Dataset3	GCR1	5	CWTCC
Dataset4	MATalpha2	5	CRTGTWWWW
Dataset5	MCB	6	WCGCGW
Dataset6	PDR3	7	TCCGYGGA
Dataset7	PHO4	4	CACGTK
Dataset8	REB1	10	YYACCCG
Dataset9	STE12	4	TGAAACM

我们采用多个吉布斯采样算法(Gibbs95^[12]: <http://rsat.scmbb.ulb.ac.be/rsat/>, Gibbs Recursive Sampler^[10]: <http://bayesweb.wadsworth.org/gibbs/gibbs.html>, Motif Sampler^[8]: <http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html>)和 YGMS 对以上数据集进行了测试比较。所有测试是用其默认参数重复运行 20 次。其结果如表 4 所示:Gibbs Recursive Sampler 在原 Gibbs95 的基础上有了较大的性能改善。Motif sampler 在采用了一阶背景模型的情况下,其性能也不太理想。我们的算法 YGMS 虽然在 GCR1 和 STE12 数据集上未能体现出性能的提高,但在其他 7 个数据集中均反映了较强的性能改善。我们也运行了 AlignACE^[5],由于该算法是返回一个候选模式集合,我们对其 MAP 得分最高的前三个模式序列进行了比较。相对而言,其性能和 Gibbs Recursive Sampler 比较接近。由于方法返回结果的不可比较性,未列入表 4 进行比较。

总的来说,在这些数据集的测试中,算法 YGMS 的性能相对比其他 3 个算法有较大幅度的提高。

结论和研究展望 为了提高转录因子结合位点的识别性能,我们在吉布斯采样识别模式的基本算法引入了几个改进,并结合生物学的先验知识引入了新的得分函数。改进后的算法 YGMS 在酵母共调控基因组数据集的测试中表现出比其

他几个基于吉布斯采样算法更高的准确率和敏感度,提高了算法的模式识别性能。

表 4 各种算法的测试结果表。表格的数值分别是各算法重复运行 20 次找到相应模式的次数和成功率。

Algorithm	Gibbs95	Gibbs Recursive Sampler	Motif Sampler	YGMS
BAS1	6(30%)	10(50%)	2(10%)	15(75%)
GCN4	0(0%)	2(10%)	1(5%)	4(20%)
GCR1	2(10%)	6(30%)	1(5%)	6(30%)
MATa2	6(30%)	12(60%)	3(15%)	17(85%)
MCB	5(30%)	12(60%)	4(20%)	20(100%)
PDR3	20(100%)	20(100%)	8(40%)	20(100%)
PHO4	10(50%)	8(40%)	5(25%)	13(65%)
REB1	3(15%)	16(80%)	1(5%)	20(100%)
STE12	2(10%)	6(30%)	2(10%)	6(30%)

我们下一步研究将在序列水平分析的基础上结合生物学基因表达分析,来提高对基因表达调控的研究。

参考文献

- Klug W S, Cummings M R. Essentials of Genetics. (2nd edn). New Jersey: Prentice Hall, 1996
- Wray G A, Hahn M W, Abouheif E, et al. The Evolution of Transcription Regulation in Eukaryotes. *Mol Biol Evol*, 2003, 20(9): 1377~1419
- Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach (2nd edn). Cambridge MIT Press, 2001
- Lawrence C E, Altschul S F, Boguski M S, et al. Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science*, 1993, 262: 208~214
- Hughes J D, Estep P W, Tavazoie S, et al. Computational Identification of Cis-regulatory Elements Associated with Groups of Functionally Related Genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, 2000, 296: 1205~1214
- Liu X, Brutlag D L, Liu J S. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In: *Proc. Pacific Symp. Biocomputing*, 2001, 6: 127~138
- Frith M C, Hansen U, Spouge J L, et al. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research*, 2004, 32(1): 189~200
- Moreau Y, De Smet F, Thijs G, et al. Functional Bioinformatics of Microarray Data: From Expression to Regulation. *Proceedings of The IEEE*, 2002, 90(11): 1722~1743
- Favorov A V, Gelfand M S, Gerasimova A V, et al. Gibbs Sampler For Identification Of Symmetrically Structured, Spaced Dna Motifs With Improved Estimation Of The Signal Length And Its Validation On The Arca Binding Sites. In: *Proceedings of BGRS 2004*, 2004, 269~272
- Thompson W, Rouchka E C, Lawrence C E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Research*, 2003, 31(13): 3580~3585
- 孙啸, 陆祖宏, 谢建明. 生物信息学基础. 北京: 清华大学出版社, 2005
- Neuwald A F, Liu J S, Lawrence C E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science*, 1995, 4(8): 1618~32