第37卷 第14期 计 算 机 工 程 2011年7月
Vol.37 No.14 Computer Engineering July 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)14—0180—03 文献标识码：A 中图分类号：TP393.06

基于 Gibbs 采样与遗传算法的模体识别

刘文远，田陆芳，王常武，王宝文

(燕山大学信息科学与工程学院，河北 秦皇岛 066004)

摘 要：借鉴 Gibbs 采样思想，将序列峰值所对应的候选模体作为遗传算法的初始种群，提出一种改进的模体识别算法。将模体在序列中的出现次数作为变量加入到适应度函数中，使其更符合生物数据的特性。在算法变异操作中加入 IUPAC 简并码保持种群的多样性。对 DBTSS 数据库中的真实数据进行测试，结果表明该算法具有较高的识别精度和较快的搜索速度。

关键词：模体识别；遗传算法；Gibbs 采样；IUPAC 简并码

Motif Identification Based on Gibbs Sampling and Genetic Algorithm

LIU Wen-yuan, TIAN Lu-fang, WANG Chang-wu, WANG Bao-wen

(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

【Abstract】 Based on the idea of Gibbs sampling, this paper initializes the first population by selecting the candidate motifs corresponding to the peaks, and proposes an improved motif identification algorithm. The definition of the fitness function adds a parameter, the number of occurrences of a motif, more in line with the characteristics of biological data. In order to maintain the diversity of population, the algorithm uses the IUPAC degenerate code for mutation. Test result of real data in the DBTSS database shows that this algorithm has higher identification precision and quick search speed.

【Key words】 motif identification; Genetic Algorithm(GA); Gibbs sampling; IUPAC degenerate code

DOI: 10.3969/j.issn.1000-3428.2011.14.060

1 概述

揭示基因组水平上的基因表达调控规律是现代生物信息学面临的重大挑战之一。转录调控是基因表达的关键步骤，转录调控因子有序地结合在目标基因启动子序列中的特殊位点，启动基因的转录和控制基因的转录效率。这些位点被称为转录因子结合位点，长度从几个到十几个碱基对不等。每个转录因子的结合位点通常都有特定的模式，称为模体[1]。识别转录因子结合位点对研究基因的转录调控有重要的意义。

模体识别已经有很多有效的算法。其中，具有代表性的有基于期望最大化算法的 MEME[2]、Gibbs Motif Sampler[3]算法等。近年来，遗传算法(Genetic Algorithm, GA)被应用于模体识别中。例如，文献[4]提出 st-GA 算法，该算法可以识别不同长度的模体。文献[5]设计 FMGA(Finding Motifs by Genetic Algorithm)算法，它使用一般遗传算法的框架，并且设计了加快收敛速度的遗传算子。文献[6]提出 MDGA(Motif Discovery using a Genetic Algorithm)算法，该算法能够有效地为同源基因预测结合位点，其适应度是通过将结合位点上每一列的信息相加得到。文献[7]针对不同的种群设计 GAMI(Genetic Algorithm approach to Motif Inference)算法，该算法对长序列的模体识别很有效。文献[8]是针对(l, *d*)问题提出的，该算法可以解决较长模体的识别问题，并且对于保守性较弱的序列也能很好地进行识别。在这些算法中，初始种群是随机生成的，并且试图从所有的序列中寻找模体，但在实际中，有些序列中不含模体，而一些序列中含有多条模体[9]。本文提出一个基于遗传算法的模体识别方法。该算法中的初始种群不是随机生成，借鉴 Gibbs 采样思想对原始序列打分，挑选出优秀的个体作为遗传算法的初始种群。不假设每条序列中都含有模体，允许序列中不含有模体。

2 模体识别算法设计

2.1 个体编码

本文采用模体的真实表达序列作为个体的编码方式。如模体 GCGGCC 的编码方式是 GCGGCC。基因序列在表达过程中会发生变异，在所有序列中找到完全匹配的子序列是很困难的，当所求模体的长度 w 超过 20 时更是如此。在算法中用误配数判断子序列是否是模体。例如，在模体 GCGGCC 的识别过程中，允许有一个碱基不匹配，则 GAGG CC 被认为是模体。

2.2 初始种群的产生

本文遗传算法利用 Gibbs 采样思想生成初始种群。具体做法：输入一组序列 $S=\{S_1,S_2,\cdots,S_m\}$，在这组序列中随机地选择起始位置 $s=(s_1,s_2,\cdots,s_m)$，在这些位置构建 w-子串集合，其中，w 是所求模体长度。这 m 条长度为 w 的子串构造位置频率矩阵(Position Frequency Matrix, PFM)。对于 DNA 序列，PFM 是一个 $4\times w$ 的矩阵。用 PFM 对每条序列中的长度为 w 的候选模体打分，打分函数为：

$$p_i=\prod_{i=1}^{w}q_{a_i},\ i,\ i=1,2,\mathrm{L}\ ,n$$

其中，q_{a_i} 表示 a_i 在第 i 位置出现的频率值。图 1 是对其中一段序列的打分结果。从图 1 可以看出，不同的候选模体对应的打分值不同。为了扩大搜索范围从而找到最优模体，对所

基金项目：河北省教育厅自然科学研究计划基金资助项目(2009339)

作者简介：刘文远(1968－)，男，教授、博士生导师，主研方向：软计算，数据库技术，生物信息学；田陆芳，硕士研究生；王常武，教授；王宝文，副教授

收稿日期：2011-02-18 **E-mail：**tianlf2009@yahoo.cn

有序列打分。将每条序列中峰值所对应的候选模体加入到种群中，作为遗传算法的初始种群。

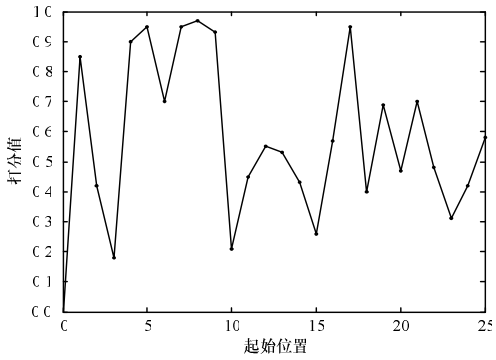

图 1　序列中候选模体对应的打分值

2.3　适应度函数的定义

定义 1 模体 P_n 与具有相同长度的子序列 S_m^i 的适应值为：

$$FS_m^i(S_m^i,P_n)=\frac{\sum_{j=1}^{w} match(S_m^{i,j},P_n^j)}{w}$$

本文将 IUPAC 简并码加入到匹配函数中，函数具体为：

$$match(s_m^{i,j},P_n^j)=\begin{cases}1.0 & s_m^{i,j}=P_n^j \text{ for } P_n^j\in\{A,T,G,C\}\\ 0.5 & s_m^{i,j}=P_n^j \text{ for } P_n^j\in\{W,R,K,S,Y,M\}\\ 0.0 & s_m^{i,j}=P_n^j \text{ for else}\end{cases}$$

其中，w 是模体 P_n 的长度；$s_m^{i,j}$ 是子序列 S_m^i 第 j 个位置上的元素；P_n^j 是模体 P_n 第 j 个位置上的元素。IUPAC 简并码如表 1 所示。

表 1　IUPAC 简并码

IUPAC 码	碱基
W	A or T
R	A or G
K	G or T
S	C or G
Y	C or T
M	A or C
B	C, G or T
D	A, G or T
H	A, C or T
V	A, C or G
N	A, C, G or T

定义 2 假设在序列 S_m 中与模体 P_n 等长的子序列有 k 条，即 $S_m^1,S_m^2,\mathrm{L}\ ,S_m^k$，则模体 P_n 与序列 S_m 的适应值为：

$$FS_m(S_m,P_n)=\max_{l=1}^{k}\{FS_m^l(S_m^l,P_n)\}\times\left|\{S_m^l:d(P_n,S_m^l)\leqslant t\}\right|$$

其中，t 表示允许的误配数，由用户输入；$d(P_n,S_m^l)$ 表示模体 P_n 与子序列 S_m^l 的海明距离；$\left|\{S_m^l:d(P_n,S_m^l)\leqslant t\}\right|$ 表示与模体 P_n 的海明距离小于 t 的子序列的个数。

实例 有 4 条序列 S_1、S_2、S_3、S_4，假设模体 P_1 的误配数是 1，则：

P_1：AGGAGGR

S_1：GTAAGGAAGGTATAGAGGAGGATTTAAG

S_2：TGGAGGGTGCACGAGGATTATATATGGT

S_3：CTGGATAGAGGAGAGCATATGGAGATAT

S_4：GCGAAGGAGGAGCTGCATATGCTAGAAT

$FS_1(S_1,P_1)$=6.5/7×2=13/7

$FS_2(S_2,P_1)$=5.5/7×1=5.5/7

$FS_3(S_3,P_1)$=0

$FS_4(S_4,P_1)$=6.5/7×1=6.5/7

定义 3 模体 P_n 与序列组 $S=\{S_1,S_2,\mathrm{L}\ ,S_m\}$ 的适应值为：

$$TFS(S,P_n)=\frac{\sum_{i=1}^{m} FS_l(S_l,P_n)}{m}$$

由定义可知，适应度越大，说明模体 P_n 与真正的模体越接近。

针对实例中的 4 条序列，模体 P_1 的适应值为：

$$TFS(S,P_1)=\frac{\frac{13}{7}+\frac{5.5}{7}+0+\frac{6.5}{7}}{4}=\frac{15}{28}$$

2.4　模体识别算法描述

算法思想是利用 Gibbs 采样思想生成遗传算法的初始种群，得到一个基于遗传算法的模体识别。算法描述如下：

```
Begin
  (1) InitPopulation; /*初始化种群*/
  (2)    SelectStartposition s=(s1,s2,…,sm); /*起始位置*/
  (3)    ConstructPFM(s,w);
  (4)    for(int i=1;i≤n;i++)
  (5)      for(int j=1;j≤L–w+1;j++)
  (6)         score[i,j]=p;
  (7)      end for;
  (8)    end for;
  (9) 选择峰值对应的子串作为初始种群(n);
  (10) While(iteration number≤M)
  (11)    for(int k=1;k≤n;k++)
  (12)       TFS[k];
  (13)    end for;
  (14)    Sort TFS;
  (15)    Selection;
  (16)    Mutation;
  (17)    Crossover; /*单点交叉*/
  (18) end while;
  (19) return the transcription factor binding sites.
End
```

算法描述中的步骤(1)~步骤(9)是初始化种群的步骤，步骤(4)~步骤(8)是用打分函数对每条启动子序列中的长度为 w 的子串进行打分，步骤(10)~步骤(19)是遗传算法的迭代过程，步骤(11)~步骤(13)是用前面定义的适应度函数计算种群中所有个体的适应度。

3　实验结果

实验数据来自于数据库 DBTSS(Database of Transcriptional Start Sites)。测试序列从转录起始位点的–2 000 bp~1 000 bp。分别用 DLD1 和 MCF7 2 组数据集进行测试。设置模体的长度为 8 和 13。为了描述方便，定义误配度，即误配数在模体中所占的比例。误配度分别设置为 25%和 18%。用本文算法分别对这 2 组数据进行测试，测试结果如表 2 和表 3 所示。

表 2　本文算法对 DLD1 的模体测试结果

模体长度	预测出的模体	模体出现次数		含有模体的序列/总的序列数 (25%误配)
		误配度=25%	误配度=18%	
8	CATGCTTG	24	20	24/50
8	ACACACAC	15	15	13/50
13	CACACACACACAC	13	12	12/50

表 3　GBGA 算法对 MCF7 的模体测试结果

模体长度	预测出的模体	模体出现次数		含有模体的序列/总的序列数 (25%误配)
		误配度=25%	误配度=18%	
8	CCCGGCCG	21	20	20/50
8	TCCTCCCA	19	15	16/50
13	CCCGGCCGCGCCC	10	7	10/50

将本文算法与 MEME 和 Gibbs Sampler 的实验结果进行比较，结果如表 4 和表 5 所示。

表 4 3 种算法对 DLD1 的识别结果比较

模体长度	算法	预测出的模体	含有模体的序列/总的序列数
8	本文算法	CATGCTTG	24/50
		ACACACAC	13/50
	MEME	CATGCTTG	24/50
	Gibbs Sampler	CAAGCTTG	11/50
13	本文算法	CACACACACACAC	12/50
	MEME	ACACACACACACA	12/50
	Gibbs Sampler	TGTGTGTGTGTGT	10/50

表 5 3 种算法对 MCF7 的识别结果比较

模体长度	算法	预测出的模体	含有模体的序列/总的序列数
8	GBGA	CCCGGCCG	20/50
		TCCTCCCA	16/50
	MEME	TCCTCCCA	15/50
	Gibbs 采样	GCTGGGCC	13/50
13	GBGA	CCCGGCCGCGCCC	10/50
	MEME	AAAGAAAATGGGA	7/50
	Gibbs 采样	GGCCCAGACGCCG	9/50

可以看出，本文算法的识别精度优于 MEME 和 Gibbs Sampler。当模体长度为 8 时，本文算法识别的结果是模体 CATGCTTG 和模体 ACACACAC，模体出现次数分别为 24 和 15。MEME 能识别出模体 CATGCTTG，模体出现次数为 24，与本文算法的识别结果相同，但在识别种类上本文算法较好；Gibbs Sampler 的识别结果是 CAAGCTTG，模体出现次数为 11，无论是模体种类还是出现次数都不如本文算法。从表 5 中可以看出，本文算法的识别结果优于 MEME 和 Gibbs Sampler。每一代中最优候选模体与真实模体的匹配见图 2。

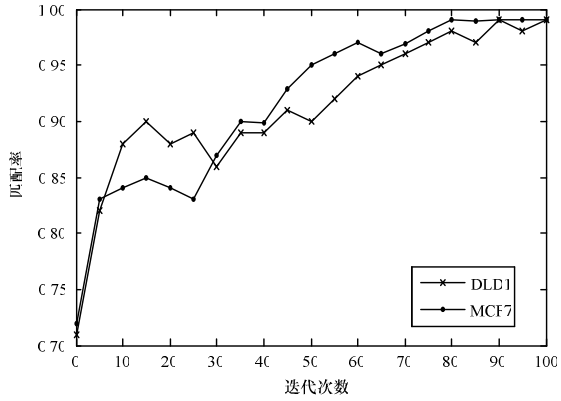

图 2 每一代中最优候选模体与真实模体的匹配

图 2 表示每一代中适应度最高的候选模体与真实模体的匹配度。可以看出，在迭代过程中，个体的适应度和匹配度逐代增加，最终达到一个最大值并基本保持不变。本文算法大约迭代 100 次后能达到最优，与 FMGA 算法和 MOGAMOD (Multi-objective Genetic Algorithm for Motif Discovery)算法相比，该算法的收敛速度有较大的提高。

4 结束语

在对 Gibbs 采样算法和遗传算法研究的基础上提出本文算法。该算法通过对所有序列打分挑选出峰值对应的候选模体作为遗传算法的初始种群；将模体出现次数加入到适应度函数中，使用 IUPAC 简并码进行变异。从实验结果可以看出，本文算法在遗传迭代过程中收敛速度较快，识别精度比 Gibbs Sampler 高，搜索速度比 MEME 快。

参考文献

[1] D'heaseleer P. What Are DNA Sequence Motifs?[J]. National Biotechnology, 2006, 24(4): 423-425.

[2] Timothy L, Bailey C E. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization[J]. Machine Learning, 1995, 21(1/2): 51-80.

[3] Thompson W, Rouchka E C, Lawrence C E. Gibbs Recursive Sampler: Finding Transcription Factor Binding Sites[J]. Nucleic Acids Research, 2003, 31(13): 3580-3585.

[4] Stine M. Motif Discovery in Upstream Sequences of Coordinately Expressed Genes[C]//Proc. of CEC'03. Memphis, USA: [s. n.], 2003: 1596-1603.

[5] Liu F F M. FMGA: Finding Motifs by Genetic Algorithm[C]//Proc. of BIBE'04. Taichung, Taiwan, China: IEEE Press, 2004: 459-466.

[6] Che Dongsheng. MDGA: Motif Discovery Using a Genetic Algorithm[C]//Proc. of Conference on Genetic and Evolutionary Computation. Washington D. C., USA: [s. n.], 2005: 447-452.

[7] Congdon C B. Preliminary Results for GAMI: A Genetic Algorithms Approach to Motif Inference[C]//Proc. of Symposium on Computational Intelligence in Bioinformatics and Computational Biology. [S. l.]: IEEE Press, 2005: 1-8.

[8] Paul T K, Iba H. Identification of Weak Motifs in Multiple Biological Sequences Using Genetic Algorithm[C]//Proc. of GECCO'06. Seattle, USA: [s. n.], 2006: 271-278.

[9] 张 菲, 谭 军, 谢竞博. 基于不同算法的 Motif 预测比较分析与优化[J]. 计算机工程, 2009, 35(22): 94-96.

编辑 陆燕菲

(上接第 179 页)

在不同领域的不平衡样本集上的仿真实验表明，该方法能够提高少数类的分类准确率，有效处理不平衡数据。今后的工作将集中在 3 种基分类器的参数优化和分类器集成与投票机制中基分类器的权重设置方面。

参考文献

[1] 马 捷, 樊 玮, 袁红玉. 基于数据场的 SVM 技术在雷暴预报中的应用[J]. 计算机工程, 2009, 35(19): 263-265.

[2] 叶志飞, 文益民, 吕宝粮. 不平衡分类问题研究综述[J]. 智能系统学报, 2009, 4(2): 148-156.

[3] Cen Li. Classifying Imbalanced Data Using a Bagging Ensemble Variation(BEV)[C]//Proc. of the 45th ACM Annual Southeast Regional Conference. Winston-Salem, USA: ACM Press, 2007.

[4] Zhu Xingquan. Lazy Bagging for Classifying Imbalanced Data[C]//Proc. of ICDM'07. Omaha, Nebraska, USA: IEEE Computer Society, 2007: 763-768.

[5] Chawla N, Bowyer K, Hall L, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(2): 321-357.

[6] Breiman L. Bagging Predictors[J]. Machine Learning, 1996, 24(2): 123-140.

编辑 张正兴