Journal of Southeast University (English Edition) Vol.19 No.4 Dec. 2003 ISSN 1003—7985

Auto-selection order of Markov chain for background sequences with chi-square test

Xie Xueying Sun Xiao Lu Zuhong

(Chien-Shiung Wu Laboratory, Southeast University, Nanjing 210096, China)

Abstract: Modeling non-coding background sequences appropriately is important for the detection of regulatory elements from DNA sequences. Based on the chi-square statistic test, some explanations about why to choose higher-order Markov chain model and how to automatically select the proper order are given in this paper. The chi-square test is first run on synthetic data sets to show that it can efficiently find the proper order of Markov chain. Using chi-square test, distinct higher order context dependences inherent in ten sets of sequences of yeast *S. cerevisiae* from other literature have been found. So the Markov chain with higher-order would be more suitable for modeling the non-coding background sequences than an independent model.

Key words: non-coding sequences; regulatory elements; chi-square test; Markov chain

It is widely noticed that the control or regulation of gene expression is primarily determined by relatively short sequences (termed with regulatory elements, transcription factors binding sites or motifs) in upstream or non-coding regions surrounding a gene. But information about these motifs (for example where they locate and how they work) is not well-known. Extracting functional motifs hidden in voluminous genome sequences is one of the grand challenges to computational biology, but first of all it is very important to model background distribution suitably.

Most popular probabilistic models published so far have applied a simple independent model[1-4], which used the frequencies of nucleotides A, C, G, and T in intergenic or input sequences to represent background sequences. However, such an independent model based on a single nucleotide can't reflect the complex structure of genome sequences and the context of relationships between nucleotides.

In DNA, the presence of a particular nucleotide usually has influence on its neighboring positions. So a better way to evaluate the probability of generating the site from the background model is based on Markov chain.

Recently, some researchers have successfully applied higher-order Markov chain to model their background sequences for gene recognition[5,6] and motif detection[7,8]. Here, only the specific issues associated with background model in motif detection will be addressed. Liu, et al.[7] developed an extended version of the Gibbs sampler called BioProspector. They proposed the use of a context dependent Markov background model with order from zero to three in a Gibbs sampling algorithm. The probability of a site being generated with their 3-order background model is computed as

$$P(b_i, b_{i+1}, \cdots, b_{i+L-1}) = P(b_i)P(b_{i+1} \mid b_i) \times P(b_{i+2} \mid b_i b_{i+1}) P(b_{i+3} \mid b_i b_{i+1} b_{i+2}) \cdots \times P(b_{i+L-1} \mid b_{i+L-4} b_{i+L-3} b_{i+L-2})$$

They illuminated the better performance of 3-order Markov background model in detecting the RAP1 sites of *S. cerevisiae*. Thijs, et al.[8] also investigated the improvement of Gibbs sampling performance to discover the promoter regulatory elements of *Arabidopsis thaliana* with a higher-order Markov background model built on a set of carefully selected intergenic sequences. Different from Liu's calculation of the probability of site being generated from the background model, they took m preceding bases of the site into account

$$P(b_i, b_{i+1}, \cdots, b_{i+L-1}) = P(b_i b_{i+1} \cdots b_{i+m-1}) \times \prod_{j=m}^{L-1} P(b_{l+j} \mid b_{i+j-m} b_{i+j-m+1} \cdots b_{i+j-1})$$

Via the test on several synthetic datasets, they showed good abilities of their polished program to deal with noisy data and to detect the less conserved motifs. But they didn't explain why this higher order background model could do better, and how the appropriate order could be auto-determined.

Received 2003-05-15.

Foundation items: The National High Technology Research and Development Program (863 Program) (2002AA231071) and the Natural Science Foundation of Jiangsu Province (BK2002057).

Biographies: Xie Xueying (1977—), female, graduate; Sun Xiao (corresponding author), male, doctor, professor, xsun@seu.edu.cn.

Furthermore, many papers have studied the over- and under-represented words in DNA sequences using various statistics based on the Markov model for DNA sequences[9,10].

We will attempt to use χ^2 statistic to determine the proper order of Markov chain for modeling background sequences. Applications of χ^2 statistic to test context dependence may be reviewed in Ref.[11].

Intergenic sequences or a set of selected upstream sequences from yeast *S.cerevisiae* were often used to serve as reliable background models in many motif discovery algorithms[3,4,7,12]. The context dependence implicit in those sequences is important and will be investigated in this paper. In order to validate the feasibility of using χ^2 statistic to test higher order dependence, synthetic random sequences defined with 4-letter DNA alphabet $\Sigma = \{A, G, C, T\}$ are constructed according to m-order Markov chain models (MC_m). Then by virtue of χ^2 test it will be discussed what is the appropriate order for modeling real non-coding sequences of *S.cerevisiae* with Markov chain.

1 χ^2 Test for Choosing Order of Markov Chain

χ^2 statistic determines that the difference between observed O and expected E scores should be attributed to some actual difference in nature or to chance; it can be computed by

$$\chi^2 = \sum_{i=0}^{I} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(F_{\mathrm{dof}})$$

where F_{dof} is the degree of freedom, and it is the total number of free parameters.

For consistency, we denote the independent background model as a 0-order Markov chain. Consider an m-order Markov model for DNA sequences, the probability of nucleotide b is determined by its m preceding nucleotides.

In this paper, the idea behind χ^2 test is to compare observed frequencies of oligo-nucleotide with those that would be expected if null hypothesis H_0 of m-order Markov chain were true. If $\chi^2 > \chi^2_{1-\alpha}$, H_0 can be rejected at a given confidence level $\alpha = 0.05$, which is the most typical value. $\chi^2_{1-\alpha}$ is the critical value that cuts off the upper 5% of the distribution with a particular degree of freedom. If H_0 is rejected, it can be concluded that the difference between observed frequencies and expected frequencies is more than what might occur by chance, and it's not suitable for modeling the sequences with m-order Markov chain. In such a situation, a higher order model should be tried. As a rule of thumb, the use of χ^2 test should be avoided if any expected frequency is less than 5. We also apply p-value to evaluate the statistical significance of the result. A low p-value for the statistical test points to rejection of the null hypothesis because it indicates how unlikely it is that a test statistic which is as extreme as or more extreme than the one given by this data, will be observed from this population if the null hypothesis is true. When α is set to 0.05, any test resulting in a p-value under 0.05 (corresponding to $\chi^2 > \chi^2_{1-\alpha}$) would be significant; therefore, the null hypothesis should be rejected to be in favor of the alternative hypothesis.

First of all, we can consider the DNA as a random sequence $\{b_i\}$ with state space Σ. Let $N(b_0 b_1 \cdots b_m b_{m+1})$ be the occurrence count of oligo-nucleotides $b_0 b_1 \cdots b_m b_{m+1}$ in sequences, and $f(b_0 b_1 \cdots b_m b_{m+1})$ be its frequency. Obviously, $f(b_0 b_1 \cdots b_m b_{m+1}) = \dfrac{N(b_0 b_1 \cdots b_m b_{m+1})}{n - m - 1}$.

A Markov model is specified in whole by its parameters: transition probability matrix and initial distribution. In this paper, the initial distribution is set to $\dfrac{1}{|\Sigma|}$ equally, and its transition probability $\pi(b_1 \cdots b_m, b_{m+1})$, which is the occurrence frequency of b_{m+1} given that its m preceding letters $b_1 \cdots b_m$ occurred, is estimated by the maximum likelihood as $\dfrac{N(b_1 \cdots b_m b_{m+1})}{N(b_1 \cdots b_m \ \cdot)}$[11]. It is worth noticing that $N(b_1 \cdots b_m \ \cdot) = \sum_{b \in \Sigma} N(b_1 \cdots b_m b)$ is the boundary count, it equals $N(b_1 \cdots b_m)$ except for the last m letters of the sequence, for which the count differs by 1 at most. So when n is large, $\pi(b_1 \cdots b_m, b_{m+1})$ can be approximated by $\dfrac{f(b_1 \cdots b_m b_{m+1})}{f(b_1 \cdots b_m)}$.

Suppose the null hypothesis H_0 is "DNA sequences can be modeled with m-order Markov chain ($m > 0$)", and the alternative H_1 is to be "DNA couldn't be modeled with m-order Markov chain". Under H_0, the frequency of $b_0 b_1 \cdots b_m b_{m+1}$ can be described as

$$\begin{aligned}
&\hat{f}(b_0 b_1 \cdots b_m b_{m+1}) = f(b_0 b_1 \cdots b_m) \pi(b_0 b_1 \cdots b_m, b_{m+1}) = \\
&f(b_0 b_1 \cdots b_{m-1}) \times \pi(b_0 b_1 \cdots b_{m-1}, b_m) \pi(b_1 \cdots b_m, b_{m+1}) \approx \\
&f(b_0 b_1 \cdots b_{m-1}) \times \frac{f(b_0 b_1 \cdots b_{m-1} b_m)}{f(b_0 b_1 \cdots b_{m-1})} \frac{f(b_1 \cdots b_m b_{m+1})}{f(b_1 \cdots b_m)} = \\
&\frac{f(b_0 b_1 \cdots b_m) \times f(b_1 \cdots b_m b_{m+1})}{f(b_1 \cdots b_m)} \qquad (1)
\end{aligned}$$

so expected number equals $n \times \hat{f}(b_0 \cdots b_{m+1})$.

With H_1, $N(b_0 b_1 b_2 \cdots b_m b_{m+1})$ is the observed number in sequences. Therefore,

$$\chi^2 = \sum_{b_0 \in \Sigma} \cdots \sum_{b_{m+1} \in \Sigma} \frac{(N(b_0 \cdots b_{m+1}) - n\,\hat{f}(b_0 \cdots b_{m+1}))^2}{n\,\hat{f}(b_0 \cdots b_{m+1})} \tag{2}$$

In order to test m-order context dependence, the $(m+2)$-letter should be taken into consideration. Its frequency is determined by anterior $(m+1)$-letter's frequency and its transition frequency to $(m+2)$-th letter. So the degree of freedom is

$$F_{\text{dof}} = (4^{m+1} - 1) \times (4 - 1) = 3 \times (4^{m+1} - 1) \tag{3}$$

F_{dof} and $\chi^2_{1-\alpha}$ with $\alpha = 0.05$ corresponding to different orders m are listed in Tab.1.

Tab.1 F_{dof} and $\chi^2_{1-\alpha}$ with regard to order m

m	F_{dof}	α	$\chi^2_{1-\alpha}$
0	9	0.05	16.92
1	45	0.05	61.66
2	189	0.05	222.08
3	765	0.05	830.46
4	3 069	0.05	3 199.00
5	49 149	0.05	12 544.00

To compensate for zero occurrences of certain oligo-nucleotides, a pseudo-count should be added to its occurrence count. Just as Thijs, et al.[8], we choose the pseudo-counts proportional to single nucleotide frequency and in inverse proportion to the square root of the size of the dataset.

We will test Markov model beginning with lower order. So before testing m-order Markov chain, we will firstly test whether we can assume the sequence consists of independent letters (0-order) just as many algorithms have done. To be a little different from m-order Markov chain, the expected frequency of dinucleotide is $\hat{f}(b_0 b_1) = f(b_0) f(b_1) = \dfrac{N(b_0\ \cdot)}{n-1} \times \dfrac{N(\cdot\ b_1)}{n-1}$.

2 Data Sets

Context dependence in non-coding sequences of yeast *S. cerevisiae* will be investigated. *S. cerevisiae* is the simplest eukaryote organism; its whole genome has been sequenced and all of its open reading frames (ORFs) have been determined[13]. First of all, we will apply χ^2 test on synthetic data.

2.1 Synthetic data

Sequences are randomly generated on Σ according to model MC_m $(m = 0,1,2,3)$. To simulate the real size of sequences used in many motif discovery algorithms, we generate 10 sequences each with length 800 for every model per trial. From the statistical point of view, every trial is executed 1 000 times to check if the behavior follows the theoretical chi-square distribution. Although synthetic data do not fully resemble biological sequences, they can be used to validate the feasibility of applying χ^2 test to find the context dependence in sequences.

2.2 Real data from *S. cerevisiae*

Because the objective of this research is to discuss which order should be selected when using higher-order Markov chain to model the background distribution applied in many motif detection algorithm, it's necessary to take real DNA sequences into consideration. We have built a dataset with intergenic sequences from *S. cerevisiae*. It consists of upstream sequences of ten families of genes which were firstly defined by van Helden[12]. Upstream sequences with length 800 of those genes, where most regulatory sites or motifs would locate[3], have been retrieved. The intergenic sequence was retrieved if its length was below 800 base pairs.

For more details about genes of the dataset, readers can refer to the literature of van Helden[12] and Saccharomyces Genome Database at Stanford[14].

3 Results and Discussion

3.1 Synthetic sequences

We have tested the null hypothesis of context dependence with distinct order aiming to validate the m-letter context dependence in sequences generated by MC_m. P-values and chi-square values are listed in Tab.2. Just as assumed, for sequences generated by

Tab.2 p-value and mean of χ^2 of 1 000 runs in every trial

MC_m	Order	F_{dof}	$\overline{\chi^2}$	p	R
MC_0	0	9	8.92	0.445	0.03
MC_1	0	9	72.91	4.09×10^{-12}	1
	1	45	35.91	0.831	0.007
MC_2	0	9	74.24	2.23×10^{-12}	1
	1	45	120.58	7.95×10^{-9}	1
	2	189	144.87	0.993	0
MC_3	0	9	72.04	6.05×10^{-12}	1
	1	45	121.40	6.08×10^{-9}	1
	2	189	304.08	2.17×10^{-7}	0.102
	3	765	583.02	1	0

Note: Order is the context dependence for which chi square test with sequences generated by MC_m; $R = N(\chi^2 > \chi^2_{1-\alpha})/1\,000$.

MC_m, Tab.2 shows that $\overline{\chi^2} > \chi^2_{0.95}$ are tenable and all p-values are smaller than 0.05 when testing for context dependence less than m; but for m-order test its p-value is much bigger than 0.05 (0.831, 0.993 and 1 respectively for MC_0, MC_1 and MC_2). Since every trial was run 1 000 times, we substituted χ^2 with its mean value. For example, concerning sequences generated with MC_2, $\overline{\chi^2}$ corresponding to order 0 and 1 test are bigger than the critical value 16.92 and 61.66 respectively and their p-values (2.23×10^{-12}, 7.95×10^{-9}) are both much smaller than 0.05; $\overline{\chi^2}$ of 2-order test is smaller than critical value 222.08 and its p-value (0.993) is much bigger than 0.05. Therefore it can be concluded that those sequences have implicit 2-order context dependence, which is accordant with their generator-MC_2.

Fig.1, Fig.2, and Fig.3 depict the curves of chi-square value frequency distribution in trials for model MC_2 and the theoretical chi-square distribution with identical degrees of freedom. These curves show that the computed chi-square values (denoted with experiment) almost followed the standard distribution. So our trials have their statistical significance.

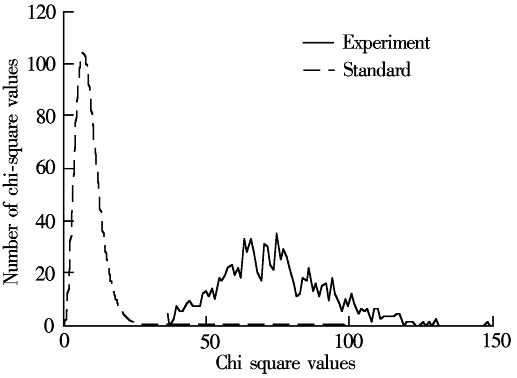

Fig.1 Curves of computed chi-square values for testing 0-order context dependence and standard distribution

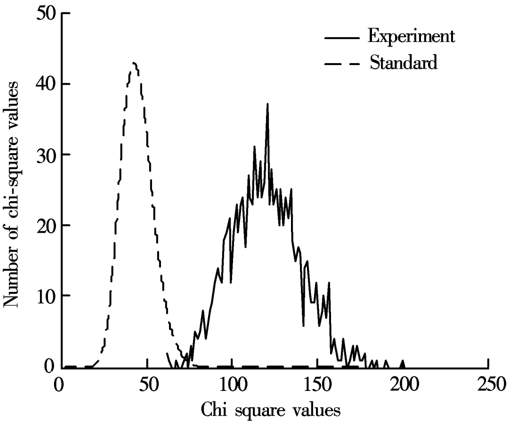

Fig.2 Curves of computed chi-square values for testing 1-order context dependence and standard distribution

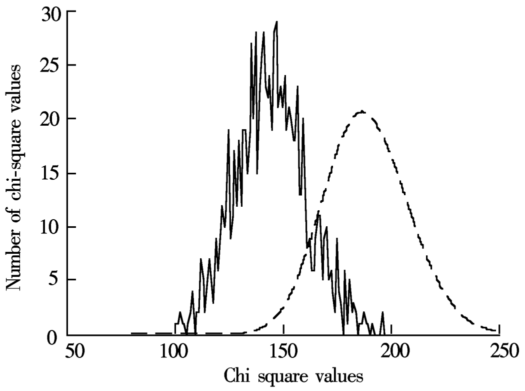

Fig.3 Curves of computed chi square values for testing 2-order context dependence and standard distribution

3.2 Results on real sequences of *S.cerevisiae*

We have analyzed the upstream sequences of ten families of genes defined first by van Helden[12]. As shown in Tab.3, all χ^2 are bigger than the corresponding critical values at 5% significance level (16.92 and 61.66) when testing the independence and one letter dependence, and all p-values are smaller than 0.05 (columns 1 to 4). These results indicate that hypothesis of independence and one previous letter's dependence should be rejected and there is a two or higher order dependence in upstream sequences. The χ^2 and p values for testing 2-order Markov chain of 5 families are respectively bigger than the critical value 222.08 and 0.05 (in bold type of the columns 5 to 6). We have studied the datasets and found that those families almost have larger size than those with $\chi^2 < 222.08$. Besides the above reasons, this phenomenon might be explained with the following possible reasons: ① Different gene families may have different context dependences; ② There are different context relationships in the same family, it may not be suitable for modeling the sequences with a fixed order. Only one element in bold type occurs when testing for 3-order dependence (in columns 7 to 8), so it's feasible to model the background distribution of those data sets using a 3-order Markov model. Since using higher-order Markov chain to model the lower distribution will overestimate context correlation and lower order model cannot reflect the complexity in the sequence, it's recommended to apply an appropriate background model while analyzing different datasets.

Tab.3 Chi-square values for upstream sequences with a length of 800 of ten families of genes defined by van Helden[12]

CP	0		1		2		3	
	χ^2	p	χ^2	p	χ^2	p	χ^2	p
GAL	**77.91**	**0**	**64.93**	**0.027 4**	163.36	0.911 3	550.01	1
GCN	**370.67**	**0**	**298.19**	**0**	441.29	0	760.79	0.536 2
HAP	**82.34**	**0**	**110.13**	**0**	175.00	0.759 3	602.36	1
INO	**59.89**	**0**	**136.24**	**0**	**232.48**	**0.017 1**	669.93	0.994 2
MET	**91.05**	**0**	**175.85**	**0**	**261.01**	**0.000 4**	729.47	0.817 4
NIT	**48.60**	**0**	**129.39**	**0**	208.85	0.153 6	638.70	0.999 7
PDR	**74.01**	**0**	**190.51**	**0**	183.66	0.596 0	599.89	1
PHO	**68.30**	**0**	**74.54**	**0.003 7**	202.90	0.232 0	644.31	0.999 4
TUP	**541.55**	**0**	**464.17**	**0**	**319.59**	**0**	**848.90**	**0.018 4**
YAP	**166.50**	**0**	**135.11**	**0**	226.58	0.032 0	630.90	0.999 9

4 Conclusion

Using the χ^2 test we attempt to explain the better performance of higher-order Markov chain in modeling background sequences than that of single nucleotide frequencies. It is the complicated context dependences inherent in intergenic sequences that can be modeled with a higher-order Markov chain. As shown in the above results different qualities and sizes of intergenic sequences would possess distinct dependence, χ^2 test can help to determine the appropriate order of the Markov chain automatically. Results also suggest that intergenic sequences may own different context dependences and future work can concentrate on using interpolated Markov chains to model background sequences; it can catch various context dependences in sequences at one time.

Although only *S. cerevisiae* is investigated in this paper, this method can easily be extended to study other organisms.

References

[1] Lawrence C E, Altschul S F, Boguski M S, et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment [J]. *Science*, 1993, **262**(5131): 208-214.

[2] Bailey T L, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization [J]. *Mach Learn*, 1995, **21**(1,2): 51-83.

[3] Roth F P, Hughes J D, Estep P W, et al. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation [J]. *Nat Biotechnol*, 1998, **16**(10): 939-945.

[4] Hughes J D, Estep P W, Tavazoie S, et al. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Sacchaomyces cerevisiae [J]. *J Mol Biol*, 2000, **296**(5): 1205-1214.

[5] Delcher A L, Harmon D, Kasif S, et al. Improved microbial gene identification with Glimmer [J]. *Nucleic Acids Res*, 1999, **27**(23): 4636-4641.

[6] Lukashin A V, Borodovsky M. GeneMark.hmm: new solutions for gene finding [J]. *Nucleic Acids Res*, 1998, **26**(4): 1107-1115.

[7] Liu X, Brutlag D L, Liu J S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes [A]. In: Altman R B, ed. *Proceedings of the 6th Pacific Symposium on Biocomputing* [C]. USA: World Scientific Pub Co, 2001. 127-138.

[8] Thijs G, Lescot M, Marchal K, et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling [J]. *Bioinformatics*, 2001, **17**(12): 1113-1122.

[9] Schbath S, Prum B, Turckheim é de. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences [J]. *J Comput Biol*, 1995, **2**(3): 417-437.

[10] Schbath S. An efficient statistic to detect over- and under-represented words in DNA sequences [J]. *J Comput Biol*, 1997, **4**(2): 189-192.

[11] Reinert G, Schbath S, Waterman M S. Probabilistic and statistical properties of words: an overview [J]. *J Comput Biol*, 2000, **7**(1,2): 1-46.

[12] van Helden J, Andre B, Collado-Vides J. Extracting regulatory sites from upstream region of yeast genes by computational analysis of oligo-nucleotide frequencies[J]. *J Mol Biol*, 1998, **281**(5): 827-842.

[13] Goffeau A, Barrell B G, Bussey H, et al. Life with 6000 genes [J]. *Science*, 1996, **274**(5287): 546-567.

[14] Dolinski K, Balajrushnan R, Christil K R, et al. Saccharomyces genome database [EB/OL]. http:// genome-www. stanford.edu/saccharomyces/, 2002.

卡方检验确定背景序列模型 Markov chain 的阶数

谢雪英　孙　啸　陆祖宏

（东南大学吴健雄实验室，南京 210096）

摘　要　合理建模非编码序列对正确识别 DNA 序列中的调控元件非常重要.基于卡方统计检验,给出了选用 Markov chain 模型来模拟序列背景分布的原因及如何确定 Markov chain 阶数的方法.卡方测试分析模拟数据发现它能有效地确定模型阶数.选择分析啤酒酵母中 10 类基因的上游序列集发现:所有序列集至少具有一阶以上的上下文相关性,除 1 组基因外,其余 9 组数据集具有二阶或三阶的上下文相关性.这说明用高阶 Markov chain 来建模背景序列比单碱基模型(零阶)更合理.

关键词　非编码序列；调控元件；卡方测试；马尔可夫链

中图分类号　Q52