Pattern Extraction from a Sequence Surrounding a Transcriptional Starting Site Using Texture Analysis : The Utility of Texture Analysis Based on the Gray Level Run Length Matrix

Haruhiko MURASE and Kazuhiro FUKUI

*Laboratory of Phytotronics and Sensibility Engineering, Graduate School of Agriculture and Biological
Sciences, Osaka Prefecture University, 1-1 Gakuen-cho, Sakai Osaka 599-8531, Japan*

(Received October 16, 2003)

The computational methods to detect the coding regions of genes and specific sequences of the transcriptional control region using stochastic models such as the hidden Markov Model have been developed. But there are still aspects of the control of the gene expression based on a sequence of the transcriptional control region that have not been described. We report here our attempt to apply texture analysis to a sequence surrounding a transcriptional starting site to extract the features. We developed the features extraction program by employing texture analysis based on the gray level run length matrix and applied it to sequences surrounding transcriptional starting sites of genes coding the heat shock proteins. Thus we could successfully represent features extracted from a sequence surrounding transcriptional starting sites as specific patterns of them.

Keywords : heat shock protein, pattern extraction, promoter, texture analysis

INTRODUCTION

The prediction of gene function is an important aspect of bioinformatics. It is well known that the gene expression is regulated by multiple sigma factors, each of which has its promoter specificity in a bacterial genome. There are two relatively conserved sequence elements around the position -35 and -10 in the promoter. However the rule in the promoter sequences has not been solved yet. There are great demands for new computational methods to extract features from DNA sequences and stochastic models such as hidden Markov Model which has recently been utilized for the analysis of the sequence patterns and the gene finding in DNA sequences (Krogh et al., 1994; Yada et al., 1997a, 1998). As for the promoter sequences, the prediction of the transcription units has been analyzed using hidden Markov Model (Yada et al., 1997b, 1999). In this paper, we propose a new approach for the extraction of features from a sequence surrounding a transcriptional starting site in bacterial genome and detection of the specific pattern from them using texture analysis based on the gray level run length matrix.

Corresponding author : Haruhiko Murase, fax : +81-72-254-9918,
e-mail : hmurase@bics.envi.osakafu-u.ac.jp

MATERIALS AND METHODS

Heat shock protein. Sequences surrounding a transcriptional starting site, including the promoter, which controls the expression of genes coding the heat shock protein (abbreviated as Hsp), were prepared from ReguronDB ([http://www.cifn.unam.mx/Computational Genomics/regurondb/](http://www.cifn.unam.mx/Computational_Genomics/regurondb/)), Biocyc (<http://biocyc.org/ecoli/>) and DBTBS (<http://dbtbs.hgc.jp/>). The Hsp is one of the proteins induced in stress response and it is essential for every living

dnaK 12040
GCGCAAAAGCACAAAAAAT TTTGCACTCCOCTTGATG ACGTGGTTACGACCCCAIT tAGTAGTCAACCGCAGTGAGT

dnaKp2
GAGTCTGCAAAAAATGAAA TTGGCAGTTGAAACCAGAC GTTGGCCCTATTACAGAC ICACAACCACATGATGACCGA

dnaKp3
GGCAGTTGAAACCAGACGTT TGGCCCTATTACAGACTCA CAACCACATGATGACCGAAT aTATAGTGGAGACGTTTAGAT

dnaKp1
AGCAGAAAAATTTTTCAT CTCCOCTTGATGACGTTGT TTAGACCCCAITTAGTAGT cAACCGCAGTGAGTGAGTCTG

grpE
GATTGATGACATGTGAGTG CTTCOCTTGAAACCTGAAA CTGATCCCATATAAGCGA aGTTAGCGAGATGAATGCGAA

hscA
AGACGTGGGACGGCGCGG GATACGTTGCGTAAGCTCG TTTTCTGATAAACTGCGAA gCAGTGCCGAACAACCTGAAG

hscC
GGGCTTTGTGATTACTGG CTGATCGTCTGAAATTTAC GCTTAACACACGTCATTGAT tAAGATCACCAACGCTTAOCT

clpB
TTTTCACTAATCTGGTCA ATAACTTGAAATAITGAGG GATGACCTCAITTAATCTC aGTAGCAACTTTGATCCGTGA

hslV
ATACAACTGCATTGGCTC GCGCGCGGGGTTGAAACCC TCAAAATCCCCCACTAT aATTGCATTATGCCCCGTACT

hspGp1
ACGAGACGAGGATCACTGC TCTGCTTGAAATTAITCTC OCTGTGCCCAITCTCCCA cATCTGTTTTTAACCTTAA

hspGp2
GGATCACTGCTCTGCTTG AAATTATCTOCTTGTOCC CATCTCTCCACATCTGTT tTTAACTTAAATGGCATT

mopB 4368212
TCCCCCTGAAGGGGGGAAG CCTCATCCCCATTCTCTGG TCAACAGCCGGGAAACCAAG tAAGCTCCGGGTCAACCCATA

htrC
AGATGACTAATCTCTGAAA CTGACTGAACATAITGAGT AAACTCGCAAGGATTGAT aCTATTCTGTGTAACTTTCT

ibpA
TGCAAAAAAAGTCCGCTGA TAAGCTTGAAAAGTTCAIT TCCAGACCAATTTTACATC gTAGCCGATGAGGACGCGCT

hslJ 1439813
GCGCTCCCTGCATTCCAG GGGAGCTGATCAGATAAT CCAATGACCTTTCACTCT tATTCTTAAATAGTCTGTAG

hspX
TCTGGTGACTTAOCCTAT CCAGACTTGAAAATAGTGC GTAAACCATACGATGTGGT aTGCATATTGGGTTTGTGA

clpPp1
ACCGTGGGAAAAGCTCTTT CGGTGTAGGTAAACAACA AAGATTGTATGCTTGAAAT aTGGTGATGCGTACCCATAA

clpPp2
CCGTACCATACACAGGGA CTAGCTGATAATCGTCCAT AAGGTACAACTGGTACAGC aGGTTTTTCAATTTTATCCA

yrfF 3523945
AAATAATGCTCCATAATCA AATTAAAAGTCAATAGGTG AAATAACTCCAGGAATTGC tGATATTCCOCTTCAGAGGG

the promoter of dnaK gene in *Bacillus subtilis*
GGGCTTTTTTTGGTGAGT TATAATTGACATTTTCTTG TGGTTTGATCTTTTGTAT aGAATAGCACTGCTTA

Fig. 1 The promoter sequences of genes coding the Hsp. The 20 sequences surrounding transcriptional starting sites are listed with the promoter name. Nineteen sequences belong to *Escherichia coli* and a sequence written the promoter of dnaK gene belongs to *Bacillus subtilis*. A space is inserted between four domains. Transcriptional starting points are written in lower-case letters. The length of nineteen sequences in *Escherichia coli* is 81 bp and the length of a sequence in *Bacillus subtilis* is 78 bp.

thing. The Hsp helps the correct folding of a protein, prevents a stressed protein from misfolding and resolves misfolding proteins. The selection of Hsp is derived from the conjecture that proteins induced in a common stress have elements within sequences in common and the fact that information has been accumulating experimentally is larger. Sequences used in this work were twenty in total; they were composed in nineteen sequences in *Escherichia coli* K12 W3110 and a sequence in *Bacillus subtilis* (Fig. 1). Twenty nucleotide sequences, whose lengths are 81 bp, surrounding transcriptional starting sites (from position -60 to $+21$) were used. The position of a transcriptional starting point is represented at $+1$.

Each of the twenty nucleotide sequences was divided into four domains, namely first domain (positions -60 to -41), second domain (positions -40 to -21), third domain (positions -20 to -1) and fourth domain (positions $+1$ to $+21$) (Fig. 2). Generally speaking, the specific sequences appear around the position -35 and -10 in the promoter. The transcription bubble is usually formed on 18 bp. Therefore, the length of the division treated in this experiment was fixed at 20 bp to divide the promoter sequences into four domains.

Features were extracted from the four domains using texture analysis based on the gray level run length matrix respectively. Namely four values were extracted from a sequence. The transition of four values from the four domains was represented as a pattern of a sequence.

Texture analysis based on the gray level run length matrix. Texture analysis is the method to analyze the density distribution in the image statistically. Texture analysis based on the gray level run length matrix is one such analysis and is a method based on the run which is a series of pixels having the same gray level on a definite direction in the image (Galloway, 1975). Since density, length and direction in the run are included, the aggregation of the run from the image represents features of the texture.

In this work, texture analysis based on the gray level run length matrix was applied to

Fig. 2 The division of a sequence into four domains.

CGGCAAAATG

Length of the run

	1	2	3	4
A	0	0	0	1
T	1	0	0	0
C	2	0	0	0
G	1	1	0	0

Density

Fig. 3 The instance of the gray level run length matrix. The runs from a sequence "CGGCAAAATG" are stored in the gray level run length matrix.

one-dimensional sequences. To use texture analysis four nucleotides, which are A, T, C and G, were converted into four figures, which are 0, 1, 2 and 3. The runs from the sequences were stored in the gray level run length matrix (Fig. 3). There were Short Runs Emphasis, Long Runs Emphasis, Gray Level Non-uniformity, Run Length Non-uniformity and Run Percentage in texture analysis based on the gray level run length matrix. In this study, the Long Runs Emphasis as one of the prime features was used to fit those fairly long DNA sequences (Fig. 4).

The transformation of four values of the Long Runs Emphasis into a single scalar value. An example of four values of the Long Runs Emphasis is shown in Fig. 5. The pattern shown in Fig. 5 implies characteristics of the analyzed domain.

We calculated a single scalar value (L) from four values using Eq. (1).

$$L = \sum_{i=1}^3 D_i \cdot \left(\frac{\tan |\theta_i|}{\tan \theta_i} \right) \quad (1)$$

where D_i is the distance between node i and node $i+1$ and θ_i is the angle as shown in Fig. 6.

Short Runs Emphasis	$l_1 = \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} \frac{p(i, j)}{j^2} / \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} p(i, j)$
Long Runs Emphasis	$l_2 = \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} j^2 p(i, j) / \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} p(i, j)$
Gray Level Non-uniformity	$l_3 = \sum_{i=1}^{Ng} \left(\sum_{j=1}^{Nr} p(i, j) \right)^2 / \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} p(i, j)$
Run Length Non-uniformity	$l_4 = \sum_{j=1}^{Nr} \left(\sum_{i=1}^{Ng} p(i, j) \right)^2 / \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} p(i, j)$
Run Percentage	$l_5 = \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} p(i, j) / P$

$p(i, j)$ is the (i, j) th entry in the given run length matrix.

i and j is the density and the length of runs respectively.

Ng is the number of gray level in the sequence.

Nr is the number of different run lengths.

P is the length of sequences.

Fig. 4 Features extracted from sequences using texture analysis based on the gray level run length matrix.

Fig. 5 A possible pattern represented by four values of the Long Runs Emphasis.

Fig. 6 Description of parameters used in Eq. (1).

RESULTS AND DISCUSSION

The run length features were extracted from the pattern of DNA sequence from left to right. The results are indicated in Fig. 7. The longer the runs there are in the sequences, the larger the value they indicate in Fig. 7. For example, the first node of dnaK_12040 has the large value because of the existence of a series of six identical bases in the analyzed domain. The Long Runs Emphasis shown in Fig. 7 was described as a single scalar value defined by Eq. (1). The calculated scalar value for each protein was plotted in Fig. 8.

DnaK_12040, dnaKp2, dnaKp3 and dnaKp1 control a dnaKJ transcription unit. It is known that transcripts from dnaKp1 and dnaKp2 are the major products of in vitro transcription of dnaKJ operon by Sigma32, and this sigma holoenzyme transcribes dnaKp3 poorly in the ReguronDB ([http://www.cifn.unam.mx/Computational Genomics/regurondb/](http://www.cifn.unam.mx/Computational_Genomics/regurondb/)). It seems that the differences of the value of L among dnaKp2, dnaKp1, dnaKp3 assure the experimental results. In general it is thought that Hsp104, Hsp90, Hsp70, Hsp60 and small Hsp

The promoter sequences of genes coding the Hsp
Fig. 7 Patterns of promoters of genes coding the Hsp in the Long Runs Emphasis.

The promoter sequences of genes coding the Hsp
Fig. 8 The single scalar value calculated from four features of a pattern.

Table 1 The list of promoters, genes and the function of proteins.

Promoter name	Gene name	1. Definition of protein	2. Function of protein
dnaK_12040 dnaKp2 dnaKp3 dnaKp1	dnaK dnaJ	1 chaperone Hsp70, DNA biosynthesis, autoregulated heat shock proteins 2 factor, chaperones 1 chaperone with DnaK, heat shock protein 2 factor, chaperones	
grpE	grpE	1 phage lambda replication, host DNA synthesis, heat shock protein, protein repair 2 IS, phage, Tn, phage-related functions and prophages	
hscA	hscA	1 heat shock protein, chaperone, member of Hsp70 protein family 2 factor, chaperones	
hscC	hscC	1 putative dnaK protein 2 putative enzyme, not classified	
clpB	clpB	1 ClpB protease, ATP dependent, heat shock protein 2 macromolecule degradation, degradation of proteins, peptides, sugars	
hslV	hslU hslV	1 heat shock protein hslVU, ATPase subunit, homologous to chaperones 2 adaptation, adaptations, atypical conditions 1 heat shock protein VU, proteasome-related peptidase subunit 2 macromolecule degradation, degradation of proteins, peptides, sugars	
htpGp1 htpGp2	htpG	1 chaperone Hsp90, heat shock protein C 62.5 2 folding and ushering proteins, chaperones	
mop	mopA mopB	1 GroEL, chaperone Hsp60, peptide-dependent ATPase, heat shock protein, affects cell division 2 cell division, cell division 1 GroES, 10 Kd chaperone binds to Hsp60, suppressing its ATPase activity, affects cell division 2 cell division, cell division	
htrC	htrC	1 heat shock protein htrC 2 adaptation, adaptations, atypical conditions	
ibpA	ibpA ibpB	1 heat shock protein 2 adaptation, adaptations, atypical conditions 1 heat shock protein 2 adaptation, adaptations, atypical conditions	
hslJ_1439813	hslJ	1 heat shock protein hslJ 2 adaptation, adaptations, atypical conditions	
htpX	htpX	1 heat shock protein, integral membrane protein 2 adaptation, adaptations, atypical conditions	
clpPp1 clpPp2	clpP	1 proteolytic subunit of clpA-clpP ATP-dependent serine protease, heat shock protein F21.5 2 macromolecule degradation, degradation of proteins, peptides	
yrfF_3523945	yrfH yrfI	1 heat shock protein 15, DNA/RNA binding 2 adaptation, adaptations, atypical conditions 1 heat shock protein 33, redox regulated chaperone 2 folding and ushering proteins, chaperones	
unknown (promoter of dnaK in <i>Bacillus subtilis</i>)	dnaK	1 chaperone protein dnaK (heat shock protein 70) (heat shock 70 kDa protein) (HSP70)	

reside in Bacteria and Eucarya (Yoshida et al., 2003). Regardless of the absence of the homology in the sequences, similar values to dnaK_12040 are found in Fig. 8 such as dnaKp2, dnaKp1, mopB_4368212 and ibpA in *Escherichia coli* and the promoter of dnaK gene in *Bacillus subtilis*. In addition, genes controlled by dnaK_12040, dnaKp2, dnaKp1 and the promoter of *Bacillus subtilis* compound Hsp70 have a similar function (Table 1). Therefore it is inferred that there is common control information about the expression condition or timing of genes in their promoter sequences and that this information is conserved as the appearance pattern of long runs at twenty nucleotides intervals between two living things. A series of the same nucleotide within first domain appears as the especially remarkable feature. It supports a report that it seems very common that a transcript can be started from, or terminated at, multiple positions in *Bacillus subtilis* (Yada et al., 1999). It is inferred that not only specific sequences such as the -35 and -10 regions but also a series of the same nucleotide are important for the recognition by RNA polymerase in the control of the transcription.

CONCLUSION

Features could be extracted from sequences surrounding transcriptional starting sites by using texture analysis based on the gray level run length matrix. It is possible that the recognition of the transcriptional control region by RNA polymerase is involved in the appearance of a series of the same nucleotide. Texture analysis based on the gray level run length matrix is an effective features extraction from the transcriptional control region.

REFERENCES

- Galloway, M. M. 1975. Texture analysis using gray level run lengths. *Comput. Graphics Image Processing* **4**: 172-179.
- Krogh, A., Mian, I. S., Haussler, D. 1994. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res.* **22**: 4768-4778.
- Yada, T., Sazuka, T., Hirose, M. 1997a. Analysis of sequence patterns surrounding the translation initiation sites on Cyanobacterium genome using the hidden Markov model. *DNA Res.* **4**: 1-7.
- Yada, T., Totoki, Y., Ishii, T., Nakai, K. 1997b. Functional prediction of *B. subtilis* genes from their regulatory sequences. *Intell. Syst. Mol. Biol.* **5**: 354-357.
- Yada, T., Totoki, Y., Ishikawa, M., Asai, K., Nakai, K. 1998. Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. *Bioinformatics* **14**: 317-325.
- Yada, T., Nakao, M., Totoki, Y., Nakai, K. 1999. Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics* **15**: 987-993.
- Yoshida, T., Yohda, M., Tadashi, M. 2003. Archaeal molecular chaperones: protein folding mechanism of the archaeal chaperonin. (Japanese text with English abstract) *Protein, Nucleic Acid and Enzyme* **48**: 33-39.

〈和文抄録〉

テクスチャー解析法を用いた転写開始部位周辺配列からのパターン抽出

村瀬治比古・福井一広

大阪府立大学大学院農学生命科学研究科

遺伝子のコード領域や転写制御配列の特異的配列を特定するために隠れマルコフモデルなどの計算手法が開発され適用されている。しかし、転写制御領域の塩基配列と遺伝子発現の関係については多くの異なった見解がある。そこで、転写開始部位の周辺配列について何らかの物理的な特徴の存在を期待してテクスチャー解析を適用し、テクスチャ特徴量を抽出した。ランレングス行列によるテクスチャー解析を用い、熱ショック蛋白質の転写制御領域に適用した。そして特徴量の変化を一つのパターンと捉えて、転写開始部位の周辺配列を表すことにした。