Distinguishing coding from non-coding sequences in a prokaryote complete genome based on the global descriptor

Guo-Sheng Han[1], Zu-Guo Yu[1,2*]
[1]School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China.

Vo Anh[2]
[2]School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia.

Raymond H. Chan[3]
[3]Department of Mathematics, Chinese University of Hong Kong, Shatin, N.T.,,Hong Kong, China.

Abstract

Recognition of coding sequences in a complete genome is an important problem in DNA sequence analysis. Their rapid and accurate recognition contributes to various relevant research and application. In this paper, we aim to distinguish the coding sequences from the non-coding sequences in a prokaryote complete genome. We select a data set of 51 available bacterial genomes. Then, we use the global descriptor method on the coding/non-coding primary sequences and obtain 36 parameters for each coding/non-coding primary sequence. These parameters are used to generate some spaces, whose points represent coding/non-coding sequences in our selected data set. In order to evaluate this method, we perform Fisher's linear discriminant algorithm on it and get relative satisfactory discriminant accuracies. The average accuracies of the global descriptor method (36 parameters) for the training and test sets are 97.81% and 97.49%, respectively. Finally, a comparison with Z curve methods using the same data set is undertaken. When we combine our method with the Z curve method, higher accuracies are obtained. This good performance indicates that the global descriptor method of this paper may complement the existing methods for the gene finding problem.

1. Introduction

As more and more genomic sequences, including those of bacteria and archaea, are available in public databases such as Genbank at ftp://ncbi.nlm.nih.gov/genbank/genomes/, many research works have been carried out on genomic sequences. Recognition of coding sequences in a genome contributes to the automatic genome annotation process. So there is a demand for prediction techniques that can rapidly and accurately distinguish the coding sequences from the non-coding sequences in a complete genome [1].

Existing well-known protein-coding prediction methods are generally divided into two categories [2]: one is based on statistic analysis, such as ZCURVE [1], GeneMarks [3] and Glimmer [4]; the other is based on similarity alignment, such as CRITICA [5] and ORPHEUS [6]. Recently, Tech and Merkl [7] combined ZCURVE, Glimmer and CRITICA into a technique named YACOP. Chen *et al.* [8] proposed a new system to recognize protein coding genes in the coronavirus genomes. If they are used separately, the Z curve method has a better performance as noted in [1,9]. Hence we will compare our method with the Z curve method only.

In this paper, we attempt to recognize the coding sequences from the non-coding sequences in a prokaryote complete genome. We use the 51 bacteria genomes which were used previously by our group [10] to test our methods. The Z curve method has achieved much success in recent years [1,11]. Dubchak *et al.* [12] first proposed the protein-chain descriptor method to predict protein folding classes. Recently our group proposed a global descriptor (GD) for DNA sequences and used it to identify promoters [13]. Here we propose to use this technique to distinguish coding and non-coding sequences in a genome. There are 36 parameters from the global descriptor method. Fisher's linear discriminant algorithm shows that the average accuracies of the global descriptor method (36 parameters) for the training and test sets are 97.81% and 97.49%, respectively. From a comparison of sensitivity, specificity, accuracy and correlation coefficient with the Z curve method [1],

*Corresponding author Zu-Guo Yu, e-mail: yuzg1970@yahoo.com or z.yu@qut.edu.au

our method proposed here may play a complementary role to the existing methods for the gene finding problem.

2 Materials

We selected 51 complete genomes of archaea and eubacteria available from the public databases Genbank at the web site ftp://ncbi.nlm.nih.gov/genbank/genomes/. We use all coding and non-coding sequences with length greater than 300 bp in the complete genomes of these 51 prokaryotes. This data set is the same as that used in Ref. [10]. For information about the categories, species names and abbreviation of names, number of coding and non-coding regions of the selected 51 prokaryote complete genomes, one can refer the Table 1 in Ref. [10]. Then we divide the data into training and test sets randomly. A set of 80% of coding/noncoding sequences is regarded as a training set, and the set of the remaining 20% of coding/noncoding sequences as a test set.

3 Method

Global descriptor of coding/non-coding sequence (GD)

The global descriptor method was proposed first by Dubchak *et al.* [12] for predicting protein folding classes based on a global protein chain description. The protein-chain descriptor consists of overall composition, transition, and distribution of amino acid attributes. Relevant further research has also been performed in Refs. [15-18]. Recently our group proposed a global descriptor (GD) for DNA sequences and used it to identify promoters [13]. Here we outline the global descriptor of DNA sequences in Ref. [13] in the following.

The global description contains three parts: composition ($Comp$), transition ($Tran$) and distribution ($Dist$). In order to explain the method, we suppose that a sequence consists of only two kinds of letters (A and B). The composition is used to measure the frequency of occurrence of each kind of letters in the sequences. For example, for the sequence: BABBABABBABBAABABABBAAABBABABA, there are 14 As and 16 Bs, hence the frequencies for A and B are $100.00\times14/(14+16) = 46.67$, $100.00 \times 16/(14 + 16) = 53.33$, respectively. These two numbers represent the first part of the global description, $Comp$. The second part, $Tran$, characterizes the percent frequency with which A is followed by B or B is followed by A. For example, for the above sequence, there are 21 transitions of this type, that is, $(21/29) \times 100.00 = 72.14$. The third part of the global description, $Dist$, measures the chain length within which the first, 25%, 50%, 75% and 100% of certain type of letters are located, respectively. For example, for the above sequence, the first, 25%, 50%, 75% and 100% of Bs are located within the first, 6th, 12th, 20th and 29th nucleotides, respectively. The $Dist$ descriptor for Bs is thus: $1/30 \times 100.00 = 3.33$, $6/30 \times 100.00 = 20.00$, $12/30 \times 100.00 = 40.00$, $20/30 \times 100.00 = 66.67$ and $29/30 \times 100.00 = 96.67$. Likewise, the $Dist$ descriptor for As is 6.67, 23.33, 53.33, 73.33 and 100.00. As a result, the global description for the above sequence is $(Comp; Tran; Dist) = (46.67, 53.33; 72.14; 6.67, 23.33, 53.33, 73.33, 100.00, 3.33, 20.00, 40.00, 66.67, 96.67)$. A more detailed description of global description of sequences can be found in Refs. [12,15-18].

The global description for the coding/non-coding sequences can be computed by similar procedures. As the coding/non-coding sequences consist of four types of nucleotides (A, C, G and T), there are 4 parameters for $Comp$, 6 parameters for $Tran$ (6 cases: A is followed by C, by G, by T respectively; C is followed by G, by T respectively; and G is followed by T) and 20 parameters for $Dist$. Overall, a total of 30 parameters are used to construct a global description of a coding/noncoding sequence.

The Entropy Density Profile (EDP) model is a global statistical description for a DNA sequence, which employs Shannon's artificial linguistic description for a DNA sequence of finite length like an open reading frame (ORF) [19]. Zhu *et al.* [19] developed a new non-supervised gene prediction algorithm for bacterial and archaeal genomes based on EDP. Here we describe such method briefly. If p_i $(i = 1, 2, 3, 4)$ are the frequencies for the four types of nucleotides of a coding/non-coding sequence, then an EDP vector $S = \{s_i\}$ inferred from $\{p_i\}$ is used to represent the sequence with an emphasis on the information content, where i is the index of the four kinds of nucleotides. The EDP s_i are defined as [19]

$$s_i = 100.00 \times (-\frac{1}{E} p_i \log p_i),\ i = 1, 2, 3, 4, \qquad (1)$$

where $E = -\sum_{i=1}^{4} p_i \log p_i$ is the Shannon entropy.

It was shown that $P = p_1^2 + p_2^2 + p_3^2 + p_4^2$ is a useful statistical quantity for analysis of DNA sequences [20,21], which was called a nucleotide composition constraint of genomes [8]. As a result, we obtain 6 parameters s_1, s_2, s_3, s_4, E and P from EDP.

Overall, combining the above two description systems, we get 36 parameters from the global descriptor for a coding/non-coding sequence.

4 Results and discussion

From the method described in the Methods section, we get a total of 36 parameters. We use Fisher's linear discriminant algorithm [22-24] to calculate the discriminant accuracies in order to evaluate our methods. We randomly divide all coding and non-coding sequences into two sets respectively. A set of 80% of coding/non-coding sequences is regarded as a training set, and the set of the remaining 20% of coding/noncoding sequences as a test set.

Fisher's discriminant algorithm is used to find a classifier in the parameter space for a training set. The given training set $H = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ is partitioned into $n_1 \leq n$ training vectors in a subset H_1 and $n_2 \leq n$ training vectors in a subset H_2, where $n_1 + n_2 = n$ and each $\mathbf{x}_i$ is a κ-dimensional vector, represented by one point in the κ-dimensional parameter space. Then $H = H_1 \cup H_2$. We need to find a parameter vector $\mathbf{w} = (w_1, w_2, \cdots, w_\kappa)^T$ for the κ-dimensional space such that $\{y_i = \mathbf{w}^T\mathbf{x}_i\}_{i=1}^n$ can be classified into two classes in the space of real numbers. If we denote

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x}_i \in H_j} \mathbf{x}_i, \quad j = 1, 2, \tag{2}$$

$$\mathbf{S}_j = \sum_{\mathbf{x}_i \in H_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T, \qquad j = 1, 2, \tag{3}$$

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2, \tag{4}$$

then the parameter vector $\mathbf{w}$ is estimated as $\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ [14]. As a result, Fisher's discriminant rule becomes: "assign $\mathbf{x}$ to H_1 if $\mathbf{Z}(\mathbf{x}) = (\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{S}_w^{-1}[\mathbf{x} - \frac{1}{2}(\mathbf{m}_1 + \mathbf{m}_2)] > 0$ and to H_2 otherwise" [14].

The discriminant accuracies for resubstitution analysis are defined as

$$p_c = \frac{\text{No. of all correct coding discriminations}}{\text{No. of coding sequences in the training set}}, \tag{5}$$

$$p_{nc} = \frac{\text{No. of all correct non-coding discriminations}}{\text{No. of non-coding sequences in the training set}}. \tag{6}$$

For the tests, the discriminant accuracies q_c and q_{nc} are defined similarly by changing "training set" to "test set" in Eqs. (5) and (6), respectively.

In order to evaluate the correct prediction rate and reliability of a predictive method, the sensitivity (S_n), specificity (S_p), accuracy (A_c) and correlation coefficient (CC) are also used [1]:

$$S_n = TP/(TP + FN), \tag{7}$$

$$S_p = TP/(TP + FP), \tag{8}$$

$$A_c = (S_n + S_p)/2, \tag{9}$$

$$CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}}, \tag{10}$$

where TP denotes the number of correctly recognized coding sequences, FN the number of coding sequences recognized as non-coding sequences, FP the number of non-coding sequences recognized as coding sequences, and TN the number of correctly recognized non-coding sequences.

Then we use the algorithm (eqs. (5) and (6)) to calculate the discriminant accuracies for the different methods. The discriminant accuracies for all 51 prokaryote genomes using the GD method are listed in Table 1. From Table 1, we can see that the GD method works well in the coding sequence recognition problem. However, from Table 1 in Ref. [10], we also can see that there are a few genomes with less than 40 non-coding regions. So on these genomes in Table 1, the results are over-fitting which is visible in lower testing than training performance. This is the limitation of our method.

From Fisher's discriminant algorithm, we calculate the four quantities defined in Eqs. (7)-(10). The average results are listed in Table 2.

In the past few years, multifractal analysis has been successfully applied in many fields [25-27]. Zhou *et al.* [10] applied multifractal analysis in the distinction of coding and non-coding sequences in complete genomes. So we compare our method with the multifractal analysis method proposed in Ref. [10] using the same data set. In Ref. [10], for all 51 prokaryotes, their average discriminant accuracies p_c, p_{nc}, q_c and q_{nc} can reach as high as 72.28%, 88.65%, 72.53% and 84.18% respectively. From Table 1, it is clear that the GD method has a better performance. The average discriminant accuracies for the 51 prokaryote genomes using the GD method can reach 97.81%, 81.99%, 97.49% and 78.47%, respectively. We also perform the method proposed in Ref. [14] on our data set. The results are not satisfactory. Their average discriminant accuracies p_c, p_{nc}, q_c and q_{nc} are 67.93%, 60.43%, 67.79% and 60.06%. So this method is not useful for our problem.

The Z curve method [1] has achieved much success in the gene finding field. Consequently, we perform a comparison with the Z curve method. It can be seen from Table 2 that our methods have similar prediction accuracies to those of the Z curve method (33 parameters) which Guo *et al.* used to recognize protein-coding genes in bacterial and archaeal genomes [1]. It means that our method proposed here is comparable with the best existing methods for the gene finding problem.

The main aim of this work is to develop efficient algorithms that can distinguish coding sequences from non-coding sequences in a genome. Once all open reading

Table 1. The discriminant accuracies for the 51 organisms using the GD method.

Species	p_c	p_{nc}	q_c	q_{nc}
Paer	99.86%	89.24%	99.33%	89.09%
NmenA	87.50%	90.08%	87.29%	87.50%
Smel	99.43%	90.20%	99.19%	93.33%
Mpul	97.70%	93.06%	99.30%	94.74%
Rpro	92.16%	67.74%	90.15%	60.32%
Buch	99.76%	81.48%	98.10%	92.86%
CpneuJ	98.09%	88.42%	98.98%	79.17%
Ctra	98.05%	80.56%	97.60%	83.33%
Synecho	98.15%	86.17%	98.80%	85.90%
Tpal	99.73%	73.33%	99.46%	50.00%
Ccre	99.64%	84.00%	99.13%	94.00%
EcolKM	97.41%	87.00%	97.95%	84.00%
EcolOH	97.43%	85.39%	96.98%	78.87%
Cjej	99.50%	85.19%	93.34%	57.14%
Mpneu	97.54%	86.46%	93.94%	80.00%
pNGR234	98.35%	78.57%	100%	60.71%
Mthe	99.54%	85.71%	99.09%	68.75%
HaloNRC	99.64%	86.39%	99.15%	83.72%
Atum	99.15%	83.92%	99.01%	84.00%
Bhal	98.59%	83.03%	98.87%	82.73%
Mjan	98.59%	81.43%	98.35%	80.00%
Nmen	97.65%	79.80%	96.77%	77.63%
Pmul	99.08%	80.34%	98.16%	80.00%
CaceA	98.66%	82.51%	99.08%	76.74%
Xfas	93.60%	81.85%	95.57%	78.82%
MtubH	99.18%	78.07%	99.32%	81.58%
MtubC	99.55%	68.33%	99.04%	64.44%
Uure	98.87%	77.14%	96.43%	88.89%
CpneuA	97.27%	84.15%	96.88%	80.95%
Cpneu	97.67%	86.73%	97.42%	72.00%
Spyo	97.10%	75.00%	98.01%	77.55%
Aful	98.14%	81.44%	96.64%	64.00%
Tmar	98.89%	75.00%	98.22%	53.33%
Llac	97.77%	79.80%	97.77%	82.00%
Ssol	95.26%	84.05%	94.35%	84.85%
Pabyssi	98.37%	76.58%	98.52%	78.57%
Hinf	96.34%	78.62%	95.13%	83.78%
Hpyl	97.57%	77.45%	97.85%	73.08%
Nost	96.31%	84.98%	97.09%	84.68%
Mgen	96.67%	79.49%	94.44%	60.00%
Bbur	98.06%	85.71%	97.42%	100%
Aquae	98.40%	79.27%	96.64%	71.43%
SaurN	97.76%	86.52%	97.82%	88.24%
Tvol	95.66%	82.91%	96.75%	72.50%
Aero	98.71%	74.26%	99.24%	78.82%
Taci	96.30%	86.81%	95.68%	75.00%
SaurM	98.04%	87.07%	96.62%	86.41%
Bsub	98.76%	83.24%	99.03%	81.40%
Spne	95.52%	73.11%	94.72%	68.33%
Phor	92.16%	67.74%	90.15%	60.32%
Mlep	80.82%	76.89%	80.62%	75.97%

Table 2. The average prediction accuracies for the testing sets made up of all coding sequences and non-coding sequences for 51 bacteria genomes using three methods.

Tool	$S_n(\%)$	$S_p(\%)$	$A_c(\%)$	CC
Z curve method(33 parameters)	97.08	99.69	98.39	0.73
GD method(36 parameters)	97.86	98.94	98.40	0.70
GD+Z curve method(69 parameters)	98.28	99.41	98.85	0.83

frames (ORFs) in a genome are extracted by certain methods, our methods can determine with high accuracy which ORF is a coding sequence. So our methods may play a complementary role to the existing methods for the gene finding problem.

Acknowledgement

Financial support was provided by the Chinese National Natural Science Foundation (grant no. 30570426) and the Fok Ying Tung Education Foundation (grant no. 101004) (Z.-G. Yu), the Australian Research Council (grant no. DP0559807) (V.V. Anh), and the Hong Kong Research Grants Council (grant no. CUHK 400508) and KDD(HK) Limited Research Fellowship (R.H. Chan).

References

[1] Guo FB , Ou HY and Zhang CT, ZURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.*, 31: 1780-1789, 2003.

[2] Guo FB and Zhang CT: ZCURVE_V, a new self-training system for recognizing protein-coding genes in viral and phage genomes. *BMC Bioinformatics*, 7: 9(1-11), 2006.

[3] Besemer J, Lomsadze A, Borodovsky M, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, 29:2607-2618, 2001.

[4] Delcher AL, Harmon D, Kasif S, White O, Salzberg SL, Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, 27:4636-4641, 1999.

[5] Badger JH, Olsen GJ, CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.*, 16:512-24, 1999.

[6] Frishman D, Mironov A, Mewes HW, Gelfand M, Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.*, 26:2941-2947, 1998.

[7] Tech M, Merkl R, YACOP: Enhanced gene prediction obtained by a combination of existing methods.*In Silico Biol*, 3:441-51, 2003.

[8] Chen LL, Ou HY, Zhang R and Zhang CT, ZCURVE_CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. *Biochemical and Biophysical Research Communications*, 307: 382-388, 2003.

[9] Zhang CT and Zhang R, A nucleotide composition constraint of genome sequences. *Comput. Biol. Chem.*, 28: 149-153, 2004.

[10] Zhou LQ, Yu ZG, Deng JQ, Anh V and Long SC, A fractal method to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation. *J. Theor. Biol.*, 232: 559-567, 2005.

[11] Yang JY, Yu ZG and Anh V, Clustering structures of large proteins using multifractal analyses based on a 6-letter model and hydrophobicity scale of amino acids. *Chaos, Solitons and Fractals*, doi:10.1016/j.chaos.2007.08.014, 2009.

[12] Dubchak I, Muchanikt I, Holbrook SR and Kim SH, Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci.*, 92: 8700-8704, 1995.

[13] Yang JY, Zhou Y, Yu ZG, Anh V and Zhou LQ, Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. *BMC Bioinformatics*, 9(1):113, 2008.

[14] Yan M, Lin ZS, and Zhang CT, A new fourier transform approach for protein. coding measure based on the format of the Z curve. *Bioinformatics*, 14: 685-690, 1998.

[15] Carter RJ, Dubchak I and Holbrook SR, A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res.*, 29: 3928-3938, 2001.

[16] Cai CZ, Han LY, Ji ZL, Chen X and Chen YZ, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, 31: 3692-3697, 2003.

[17] Zhang Z, Kochhar S and Grigorov MG, Descriptor-based protein remote homology identification. *Protein Sci.*, 14: 431-444, 2005.

[18] Li ZR, Lin HH, Han LY, Jiang L, Chen X and Chen YZ, PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, 34: W32-W37, 2006.

[19] Zhu HQ, Hu GQ, Yang YF, Wang J and She ZS, MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *BMC Bioinformatics.*, 8: 97(1-14), 2007.

[20] Zhang CT, and Zhang R, Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.*, 19: 6313-6317, 1991.

[21] Zhang CT and Wang J, Recognition of Protein Coding Genes in the Yeast Genome at Better Than 95% Accuracy Based on the Z curve. *Nucleic Acids Res.*, 28: 2804-2814, 2000.

[22] Mardia KV, Kent JT, and Bibby JM, *Multivariate Analysis.* Academic Press, London,1979.

[23] Duda RO, Hart PE, and Stork DG, *Pattern Classification*, 2nd ed., John Wiley & Sons, New York, 2001.

[24] Sneath PH and Sokal RR, *Numerical Taxonomy*, Freeman, San Francisco, 1973.

[25] Yu ZG, Anh V, Lau KS and Zhou LQ, Clustering of protein structures using hydrophobic free energy and solvent accessibility of proteins. *Phys. Rev. E*, 73: 031920, 2006.

[26] Yu ZG, Anh V and Lau KS, Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome. *Physica A*, 301: 351-361, 2001.

[27] Yu ZG, Anh V and Lau KS, Measure representation and multifractal analysis of complete genomes. *Phys. Rev. E*, 64: 031903, 2001.