SPECIAL TOPIC — 8th IUPAP International Conference on Biological Physics

Computational prediction of over-annotated protein-coding genes in the genome of *Agrobacterium tumefaciens* strain C58*

Yu Jia-Feng(于家峰)[a)b)†], Sui Tian-Xiang(隋天翔)[a)d)], Wang Hong-Mei(王红梅)[c)], Wang Chun-Ling(王春玲)[c)], Jing Li(荆 莉)[c)], and Wang Ji-Hua(王吉华)[a)c)]

[a)] *Shandong Provincial Key Laboratory of Functional Macromolecular Biophysics, Institute of Biophysics, Dezhou University, Dezhou* 253023, *China*
[b)] *State Key Laboratory of Bioelectronics, Southeast University, Nanjing* 210096, *China*
[c)] *College of Physics and Electronic Information, Dezhou University, Dezhou* 253023, *China*
[d)] *College of Life Science, Shandong Normal University, Jinan* 250014, *China*

(Received 22 January 2015; revised manuscript received 2 April 2015; published online 20 October 2015)

Agrobacterium tumefaciens strain C58 is a type of pathogen that can cause tumors in some dicotyledonous plants. Ever since the genome of *A. tumefaciens* strain C58 was sequenced, the quality of annotation of its protein-coding genes has been queried continually, because the annotation varies greatly among different databases. In this paper, the questionable hypothetical genes were re-predicted by integrating the TN curve and *Z* curve methods. As a result, 30 genes originally annotated as "hypothetical" were discriminated as being non-coding sequences. By testing the re-prediction program 10 times on data sets composed of the function-known genes, the mean accuracy of 99.99% and mean Matthews correlation coefficient value of 0.9999 were obtained. Further sequence analysis and COG analysis showed that the re-annotation results were very reliable. This work can provide an efficient tool and data resources for future studies of *A. tumefaciens* strain C58.

Keywords: *Agrobacterium tumefaciens* strain C58, protein-coding gene, genome re-annotation, graphical representation

PACS: 82.39.Pj, 87.14.gk **DOI:** 10.1088/1674-1056/24/12/128202

1. Introduction

At present, the rapid development of sequencing technology is accelerating the explosive accumulation of genomic sequences in bioinformatics databases. When faced with these abstract character strings composed of A, G, C, and T, it is difficult to obtain useful information directly. Thus, one of the most important tasks at hand is to develop efficient mathematical algorithms for annotating the genomic components precisely. Among programs that mine genomic components, protein-coding gene prediction may be the most basic one.

Because of the lack of introns in a prokaryotic genome, it would seem that gene prediction would be much easier in prokaryotes than in eukaryotes. However, more and more studies have indicated that gene prediction in prokaryotic genomes is far from being completely accomplished, and the annotation errors seem to be very universal in microbial genomes, a reality that has seriously depressed the reliability of databases and even caused false scientific conclusions.[1,2] Some genome re-annotation algorithms have been put forward to filter out the falsely predicted protein-coding genes (i.e., the over-annotated protein-coding genes). However, because the numerical parameters employed in the prediction algorithms only partially exhibit the specific features of protein-coding genes, the predicted results produced by different programs differ greatly in most cases.[3] Thus, it is essential to seek better numerical parameters in order to depict features of protein-coding genes in prokaryotic genomes comprehensively. In this work, the over-annotated protein-coding genes in the genome of *Agrobacterium tumefaciens* strain C58 were re-annotated on the basis of two kinds of graphical representations;[4] namely, the *Z* curve[5] and the TN curve[6] methods. In order to assess the reliability of the predicted results, the concepts of relative G + C content and purine/pyrimidine disparity at the three codon positions were proposed, which can exhibit the uneven usage properties of protein-coding sequences. Finally, circumspect bioinformatics analysis based on the above-mentioned evaluation indexes and COG annotation, as well as other evidence, showed that the predicted results are very reliable.

This paper is organized as follows. In Section 2, the dataset and the re-annotation algorithm were introduced. In section 3, the re-annotation results and discussion were presented. Finally, we presented a summary of our results in section 4.

*Project supported by the National Natural Science Foundation of China (Grant Nos. 61302186 and 61271378) and the Funding from the State Key Laboratory of Bioelectronics of Southeast University.

†Corresponding author. E-mail: jfyu1979@126.com

© 2015 Chinese Physical Society and IOP Publishing Ltd

2. Materials and methods

2.1. The *A. tumefaciens* strain C58 genome

A. tumefaciens strain C58, originally isolated from a cherry tree tumor, is a biovar 1 nopaline-producing strain, with a genome consisting of one circular and one linear chromosome and two plasmids.[7] The genomic sequence was downloaded from the relatively well-annotated RefSeq database,[8] where the accession numbers of the four replicons of strain C58's genome are NC_003062, NC_003063, NC_003064, and NC_003065, respectively. According to the current annotation of strain C58's genome, 5355 open reading frames (ORFs) are protein-coding genes that can be classified into three groups. The first group includes the function-known genes, which are genuine protein-coding genes that can be used as the training set and testing set in the re-annotation program. The third group is composed of the ORFs that are marked with "hypothetical protein," which need to be further verified because some may be random sequences that are falsely predicted to be protein-coding genes. Other annotated genes in the second group are marked with prefixes, such as "putative" and "conserved." Although the annotated genes in the second group are much more reliable than those in the third group, some are still questionable.

2.2. The Re-annotation algorithm

2.2.1. Numerical descriptors

The re-annotation program was developed by integrating the TN curve and the Z curve methods[4] (for details, please refer to www.cbi.seu.edu.cn/RPGM). To numerically display the intrinsic properties that differentiate between protein-coding and non-coding sequences, 75 numerical parameters were derived from the two kinds of graphical representations.

The TN curve can project 64 kinds of trinucleotides into a two-dimensional space,[5] and then each trinucleotide is numerically decided by coordinates (x,y) on the basis of which $z = x * y$ is defined. Based on the encoding strategy of the TN curve, 54 character parameters corresponding to the geometric center of each variable are derived to exhibit the specific property of composition and order of trinucleotides along the DNA sequences, which is written as Eqs. (1), (2), and (3).

$$
\begin{aligned}
u_1^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+0\}}} x_i^{\{+0\}}\right)/N^{\{+0\}},\\
u_2^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+0\}}} y_i^{\{+0\}}\right)/N^{\{+0\}},\\
u_3^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+0\}}} z_i^{\{+0\}}\right)/N^{\{+0\}},\\
u_4^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+0\}}} x_i^{\{+0\}\prime}\right)/N^{\{+0\}},
\end{aligned}
\tag{1a}
$$

$$
\begin{aligned}
u_5^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+0\}}} y_i^{\{+0\}\prime}\right)/N^{\{+0\}},\\
u_6^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+0\}}} z_i^{\{+0\}\prime}\right)/N^{\{+0\}},\\
u_7^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+1\}}} x_i^{\{+1\}}\right)/N^{\{+1\}},\\
u_8^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+1\}}} y_i^{\{+1\}}\right)/N^{\{+1\}},\\
u_9^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+1\}}} z_i^{\{+1\}}\right)/N^{\{+1\}},
\end{aligned}
\tag{1b}
$$

$$
\begin{aligned}
u_{10}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+1\}}} x_i^{\{+1\}\prime}\right)/N^{\{+1\}},\\
u_{11}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+1\}}} y_i^{\{+1\}\prime}\right)/N^{\{+1\}},\\
u_{12}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+1\}}} z_i^{\{+1\}\prime}\right)/N^{\{+1\}},
\end{aligned}
\tag{1c}
$$

$$
\begin{aligned}
u_{13}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+2\}}} x_i^{\{+2\}}\right)/N^{\{+2\}},\\
u_{14}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+2\}}} y_i^{\{+2\}}\right)/N^{\{+2\}},\\
u_{15}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+2\}}} z_i^{\{+2\}}\right)/N^{\{+2\}},
\end{aligned}
\tag{1d}
$$

$$
\begin{aligned}
u_{16}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+2\}}} x_i^{\{+2\}\prime}\right)/N^{\{+2\}},\\
u_{17}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+2\}}} y_i^{\{+2\}\prime}\right)/N^{\{+2\}},\\
u_{18}^{\mathrm{I}} &= \left(\sum_{i=1}^{N^{\{+2\}}} z_i^{\{+2\}\prime}\right)/N^{\{+2\}};
\end{aligned}
\tag{1e}
$$

$$
\begin{aligned}
u_{19}^{\mathrm{II}} &= \left(\sum_{i=1}^{N^{\{+0\}}} x_i^{\{+0\}}\right)/N^{\{+0\}},\\
u_{20}^{\mathrm{II}} &= \left(\sum_{i=1}^{N^{\{+0\}}} y_i^{\{+0\}}\right)/N^{\{+0\}},\\
u_{21}^{\mathrm{II}} &= \left(\sum_{i=1}^{N^{\{+0\}}} z_i^{\{+0\}}\right)/N^{\{+0\}},
\end{aligned}
\tag{1f}
$$

$$
\begin{aligned}
u_{22}^{\mathrm{II}} &= \left(\sum_{i=1}^{N^{\{+0\}}} x^{\{+0\}\prime}\right)/N^{\{+0\}},\\
u_{23}^{\mathrm{II}} &= \left(\sum_{i=1}^{N^{\{+0\}}} y^{\{+0\}\prime}\right)/N^{\{+0\}},\\
u_{24}^{\mathrm{II}} &= \left(\sum_{i=1}^{N^{\{+0\}}} z^{\{+0\}\prime}\right)/N^{\{+0\}},
\end{aligned}
\tag{2a}
$$

$$
u_{25}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+1\}}} x_i^{\{+1\}}\right)/N^{\{+1\}},
\tag{2b}
$$

$$u_{26}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+1\}}} y_i^{\{+1\}}\right)/N^{\{+1\}},$$
$$u_{27}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+1\}}} z_i^{\{+1\}}\right)/N^{\{+1\}}, \quad (2c)$$
$$u_{28}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+1\}}} x^{\{+1\}\prime}\right)/N^{\{+1\}},$$
$$u_{29}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+1\}}} y^{\{+1\}\prime}\right)/N^{\{+1\}},$$
$$u_{30}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+1\}}} z^{\{+1\}\prime}\right)/N^{\{+1\}}, \quad (2d)$$
$$u_{31}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+2\}}} x_i^{\{+2\}}\right)/N^{\{+2\}},$$
$$u_{32}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+2\}}} y_i^{\{+2\}}\right)/N^{\{+2\}},$$
$$u_{33}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+2\}}} z_i^{\{+2\}}\right)/N^{\{+2\}}, \quad (2e)$$
$$u_{34}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+2\}}} x^{\{+2\}\prime}\right)/N^{\{+2\}},$$
$$u_{35}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+2\}}} y^{\{+2\}\prime}\right)/N^{\{+2\}},$$
$$u_{36}^{\mathrm{II}} = \left(\sum_{i=1}^{N^{\{+2\}}} z^{\{+2\}\prime}\right)/N^{\{+2\}}; \quad (2f)$$
$$u_{37}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+0\}}} x_i^{\{+0\}}\right)/N^{\{+0\}},$$
$$u_{38}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+0\}}} y_i^{\{+0\}}\right)/N^{\{+0\}},$$
$$u_{39}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+0\}}} z_i^{\{+0\}}\right)/N^{\{+0\}}, \quad (3a)$$
$$u_{40}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+0\}}} x^{\{+0\}\prime}\right)/N^{\{+0\}},$$
$$u_{41}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+0\}}} y^{\{+0\}\prime}\right)/N^{\{+0\}},$$
$$u_{42}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+0\}}} z^{\{+0\}\prime}\right)/N^{\{+0\}}, \quad (3b)$$
$$u_{43}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+1\}}} x_i^{\{+1\}}\right)/N^{\{+1\}},$$
$$u_{44}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+1\}}} y_i^{\{+1\}}\right)/N^{\{+1\}},$$
$$u_{45}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+1\}}} z_i^{\{+1\}}\right)/N^{\{+1\}}, \quad (3c)$$
$$u_{46}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+1\}}} x^{\{+1\}\prime}\right)/N^{\{+1\}},$$
$$u_{47}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+1\}}} y^{\{+1\}\prime}\right)/N^{\{+1\}},$$
$$u_{48}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+1\}}} z^{\{+1\}\prime}\right)/N^{\{+1\}}, \quad (3d)$$
$$u_{49}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+2\}}} x_i^{\{+2\}}\right)/N^{\{+2\}},$$
$$u_{50}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+2\}}} y_i^{\{+2\}}\right)/N^{\{+2\}},$$
$$u_{51}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+2\}}} z_i^{\{+2\}}\right)/N^{\{+2\}}, \quad (3e)$$
$$u_{52}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+2\}}} x^{\{+2\}\prime}\right)/N^{\{+2\}},$$
$$u_{53}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+2\}}} y^{\{+2\}\prime}\right)/N^{\{+2\}},$$
$$u_{54}^{\mathrm{III}} = \left(\sum_{i=1}^{N^{\{+2\}}} z^{\{+2\}\prime}\right)/N^{\{+2\}}. \quad (3f)$$

It is noted that in Eq. (3), $+0$, $+1$, and $+2$ represent the three forward-reading frames from $5'$ to $3'$ of each sequence, respectively, and $N^{\{\}}$ is the total number of trinucleotides in each reading frame; I, II, and III denote three encoding strategies to decide (x,y), which is dependent on the first, second, and third base category, respectively; x', y', and z' display the cumulative effect of x, y, and z, respectively.

In contrast to the TN curve, the 21 parameters derived from the Z curve exhibit the statistical properties of each codon position,[9] including nine position-specific parameters (Eq. (4)):

$$\begin{cases} x_i = (a_i + g_i) - (c_i + t_i), \\ y_i = (a_i + c_i) - (g_i + t_i), \\ z_i = (a_i + t_i) - (g_i + c_i), \end{cases} \quad (4)$$

where $i = 1, 2, 3$ denotes the three codon positions. Besides this, 12 phase-specific dinucleotides are also illustrated by the following Z-transform (Eq. (5)):

$$\begin{cases} x^X = [p(XA) + p(XG)] - [p(XC) + p(XT)], \\ y^X = [p(XA) + p(XC)] - [p(XG) + p(XT)], \\ z^X = [p(XA) + p(XT)] - [p(XG) + p(XC)]. \end{cases} \quad (5)$$

In Eq. (5), $X = A, C, G, T$; $p(AA), p(AC), \ldots, p(TT)$ represent the occurring frequencies of the 16 dinucleotides AA, $AC, \ldots, TT$, respectively.

As is well known, the intrinsic difference between protein-coding genes and non-coding sequences lies in the former having regularly specific features, such as asymmetric nucleotide distributions at the three codon positions and codon usage bias, whereas the latter does not have such features. Thus, the ability to propose efficient numerical descriptors to exhibit the specific features of protein-coding genes is

the core for gene prediction programs. In this work, 54 numerical descriptors derived from the TN curve and 21 statistic numerical descriptors derived from the Z curve method were used to demonstrate the specific features of protein-coding genes. As has been introduced before,[4,10,11] the numerical descriptors derived from the TN curve can exhibit information about the composition and distribution of trinucleotides along DNA sequences. In contrast, the numerical descriptors derived from the Z curve provide statistical information about the base compositions at the three codon positions as well as the dinucleotide compositions.[9] Therefore, the two groups of parameters can complement each other, displaying comprehensive sequence properties of protein-coding genes from different angles. On the other hand, the differences between protein-coding genes and non-coding sequences are universal in all species, which has been the foundation of gene prediction algorithms. In this work, the 75 derived parameters were used to demonstrate these universal properties that differentiate protein-coding sequences from non-coding sequences; therefore, it is conceivable that the present method is applicable to other prokaryotic genomes.

2.2.2. The fisher discriminant algorithm

The Fisher discriminant algorithm has been used extensively in gene prediction-related problems (for details, please consult the work by Zhang and Wang).[12] The function-known protein-coding genes in the first group mentioned above were used as the positive training set, and corresponding shuffling sequences were used as the negative training set. Thus, a 75-D coefficient (C) for differentiating the positive and negative samples was obtained by the Fisher linear equation. Furthermore, the threshold C_0 was determined by strictly allowing the rates of false-negatives and false positives to be identical. Then, a $T_\text{score} = C \cdot V - C_0$ was assigned to each sequence, which can decide its coding/non-coding status by $T_\text{score} > 0$ or $T_\text{score} < 0$.

2.3. Evaluation index

To evaluate the efficiency of the predicted results, the accuracy (A_c), sensitivity (s_n), and specificity (s_p) parameters proposed by Burset and Guigo[13] were adopted, which are described as

$$\begin{cases} s_\text{n} = \dfrac{TP}{TP+FN}, \\ s_\text{p} = \dfrac{TN}{TN+FP}, \\ A_\text{c} = \dfrac{s_\text{n}+s_\text{p}}{2}, \end{cases} \tag{6}$$

where TP and TN denote the number of coding and non-coding ORFs that have been correctly predicted, respectively. By contrast, FP and FN denote the coding and non-coding ORFs that have been falsely predicted, respectively.

In addition, the MCC value was also used to describe the agreement of predictions and annotation with a single value in the range of $[-1, 1]$. The MCC can be represented as

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}}. \tag{7}$$

3. Results and discussion

3.1. Predicting the over-annotated protein-coding genes

On the basis of the re-annotation algorithm introduced above, the questionable protein-coding genes in the second and third groups were re-predicted. It is noted here that because of the lack of non-coding sequences in prokaryotic genomes, the negative samples in the training set were produced by shuffling the genuine protein-coding genes. Then, the re-annotation algorithm was run 10 times to provide objective predictions, and those items that were predicted to be non-coding at least eight times were filtered as over-annotated genes. To evaluate the performance of the present method, the mean accuracy and MCC values were calculated. After performing the re-annotation algorithm, almost all the protein-coding genes in the first group could be correctly predicted each time, and only one protein-coding gene (Atu3016) in NC_003062 was falsely predicted twice. The average accuracy and MCC values of 99.99% and 0.9999 were obtained, respectively. Using the discriminant coefficient C and threshold C_0 trained by the training set, all the annotated protein-coding genes in the second and third groups were re-predicted. Finally, all of the items in the second group were predicted to be protein-coding genes, and 30 hypothetical genes in the third group were predicted to be over-annotated genes, including 11 items in the positive strand and 19 items in the negative strand. In Table 1, the 30 over-annotated genes are presented.

Table 1. The 30 hypothetical genes that were predicted to be non-coding sequences.

Atu0152	Atu0532	Atu0805	Atu1092	Atu1223	Atu1466	Atu1566	Atu1638	Atu1660	Atu1666
Atu1724	Atu1970	Atu2146	Atu2192	Atu2367	Atu2368	Atu2491	Atu2704	Atu3120	Atu5025
Atu5146	Atu5161	Atu5179	Atu5286	Atu5486	Atu5549	Atu6137	Atu8020	Atu8062	Atu8205

3.2. Reliability analysis of the annotation results

After performing the re-annotation algorithm on the *A. tumefaciens* strain C58 genome, the questionable protein-coding genes were re-annotated, with high accuracy and *MCC* values. Our recent work has demonstrated that most gene-finding programs can produce very different prediction results even with similar accuracy.[3] Thus, it is still not certain whether the re-annotation results are reliable enough. It would be necessary to provide external evidence for why the filtered hypothetical genes do not code proteins.

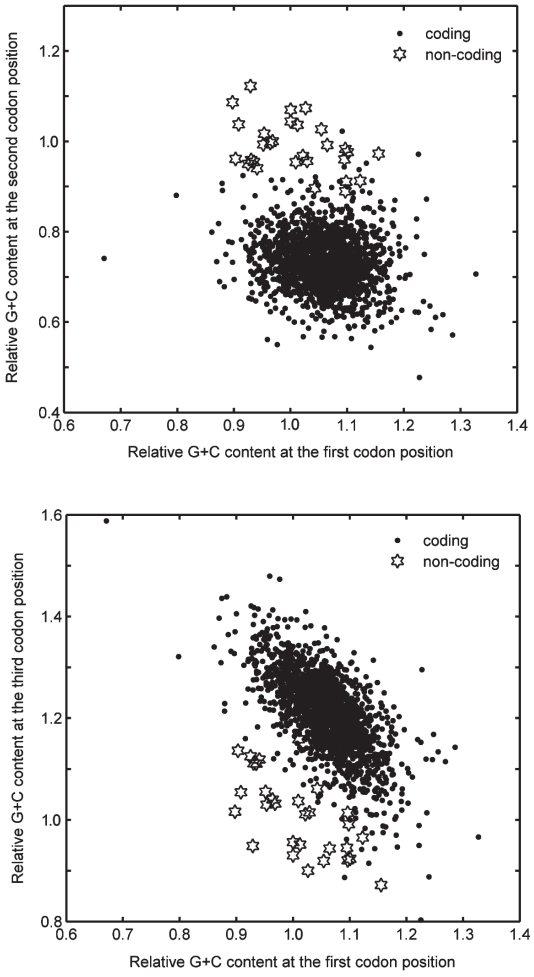


Fig. 1. The relative G + C content at the three codon positions. As can be seen, the locating regions of the predicted non-coding genes are obviously different from those of the other genes. Careful observation indicates that the G + C content at the second and third codon positions of the protein-coding genes exhibit high usage bias. On the contrary, the values of relative G + C contents at the three codon positions of the predicted non-coding genes are about 1, respectively, which indicate that they are likely to be random sequences.

As is well known, the assumption that the specific sequence structures differentiate protein-coding genes from non-coding sequences forms the theoretical basis of gene-finding programs. Many kinds of statistical methods have been proposed to describe the specific features of protein-coding genes from different angles, which may be the key factor that causes the various predictions by different gene-finding programs. In this work, the 75 derived parameters reflect information about the composition and distribution of both nucleotides and trinucleotides. In the scatter diagram of Fig. 1, a simple index was used to reflect the relative G + C content at the three codon positions of the predicted over-annotated genes and the genuine genes composed of the remaining ORFs. The relative G + C content is defined as the ratio of the G + C content at each codon position to the G + C content of the whole sequence. Obviously, the relative G + C content equals 1 if the G + C content at the corresponding position is identical to that of the whole sequence, which means less usage bias. As can be seen from Fig. 1, the locating regions of the predicted over-annotated genes differed greatly from those of the protein-coding genes. Close observation shows that the over-annotated genes are located around coordinate (1, 1), whereas the genuine protein-coding genes are far from this coordinate. Therefore, figure 1 indicates that the 30 over-annotated genes are likely to be random sequences.

Previous studies on protein-coding sequences have shown that strict constraints are exerted on each codon position to encode functional proteins, a fact that has been used as the foundation of gene-finding algorithms. Generally, it was found that the constraints are universal, in that the first codon position prefers purine bases.[14] On the other hand, the flexible third position determines the codon usage bias, which is coincident with Fig. 1. In Fig. 2, the purine/pyrimidine disparity obtained by the difference of A+G content and T+C content at the three codon positions was calculated. A close observation of this figure shows that the distributions of the predicted over-annotated genes are obviously different from those of the genuine protein-coding genes. It is obvious that the purine bases are absolutely predominant at the first codon position in genuine protein-coding genes, whereas the purine/pyrimidine disparity at the first position of the over-annotated genes is below zero. Further analysis of the genuine protein-coding genes and the over-annotated genes showed the mean values of purine/pyrimidine disparity at the first codon position to be 0.24 and -0.03, respectively. In Fig. 1, we have demonstrated that the G + C content at the third codon position of the over-annotated protein-coding genes exhibit less bias. A similar tendency can be found from Fig. 2, as the values of purine/pyrimidine disparity of the over-annotated genes are close to zero. Further analysis showed the mean values of purine/pyrimidine disparity at the third position for the genuine and over-annotated genes to be -0.164 and 0.02, respectively. Therefore, figure 2 provides additional evidence that the 30 over-annotated genes do not encode proteins.

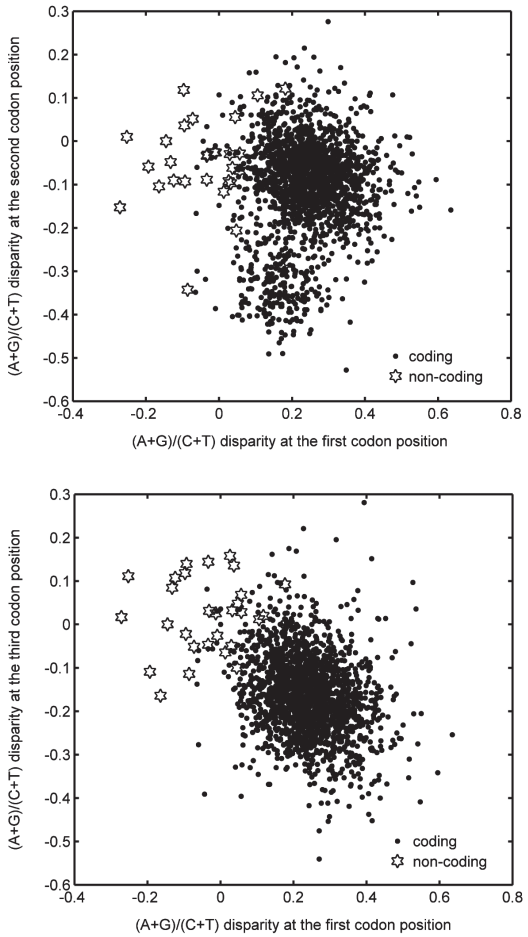


Fig. 2. Purine/pyrimidine disparity at the three codon positions. This figure shows that the locating regions of the over-annotated genes are obviously different from those of the genuine protein-coding genes. It is found that the purine bases are absolutely predominant at the first codon position in protein-coding genes, whereas the purine/pyrimidine disparity values at the first position of the over-annotated genes are below zero.

In the past several years, the Clusters of Orthologous Groups of proteins (COG) database[15] has been used widely in gene recognition-related problems. Each COG represents a group of three or more proteins that have evolved from a common ancestor. The ORF marked with a COG code is believed to be highly likely to encode proteins. According to the current annotation of the *A. tumefaciens* strain C58 genome, almost 97% of the function-known genes are assigned with COG codes, whereas no more than 60% of the hypothetical genes are marked with COG codes. Among the 30 over-annotated protein-coding genes, only Atu5025 and Atu8205 have COG codes, which are COG0856F and COG4991S, respectively. The COG code letters "F" and "S" designate the metabolism group and the poorly characterized group, respectively. COG0856F stands for "orotate phosphoribosyltransferase homologs" and COG4991S stands for "uncharacterized protein with a bacterial SH3 domain homolog"; hence, both ORFs have no explicit matches, even though they have been assigned COG codes. It was noted that the genome annotation of *A. tumefaciens* strain C58 had been updated during the preparation of this revised work. Unsurprisingly, the COG codes of Atu5025 and Atu8205 have been cancelled in the updated version. Therefore, the COG analysis indicates that our prediction is much more reliable in another sense. Previous studies have shown that some annotated short ORFs may be not true protein-coding genes, and these short ORFs also account for most of the over-annotation of the protein-coding genes in prokaryotic genomes. In strain C58's genome, the average length of the 30 over-annotated genes is 340 bp, whereas that of the genuine protein-coding genes is about 950 bp. In a recent work by Wang *et al.*,[16] 29 ORFs were predicted to be over-annotated genes in strain C58's genome. Importantly, although there are only 13 common items with our work, it was found that 10 of these could be verified by molecular biological experiments. Therefore, the present work is an efficient complement for future genome annotation research.

4. Conclusions

Up to now, almost 6887 genomic projects have been completed, among which about 96% are from prokaryotic genomes.[17] As genome annotation is a multistep process,[18,19] the quick advance of these increasing sequence data provides severe challenges for bioinformatics, particularly with regards how to accurately annotate the genomic components. In the past 30 years, gene finding has been one of the most important topics of the life sciences. Even so, the problem of gene annotation errors has been deemed a universal phenomenon in microbial genomes,[20,21] and different re-annotation strategies have been outlined.[3] Despite that most programs have claimed to achieve high accuracies on their training sets and testing sets, more and more recent studies using different gene-finding algorithms have shown that the predicted results differ greatly.[3,4,22,23] Therefore, there is an urgent need to develop more comprehensive algorithms for reliable predictions. In this work, with high accuracy, we re-annotated the *A. tumefaciens* strain C58 genome on the basis of 75 efficient parameters derived from the TN curve and *Z* curve methods. The reliability analysis based on the proposed position-specific indexes, COG annotation, and other analyses support our predictions as being credible. We hope that this computational method will be a helpful tool for future genome annotation studies.

References

[1] Kyrpides N C 2009 *Nat. Biotechnol.* **27** 627
[2] Petty N K 2010 *Nat. Rev. Microbiol.* **8** 762

[3] Yu J F, Guo Z Z, Sun X and Wang J H 2014 *Curr. Bioinformatics* **9** 147
[4] Yu J F, Xiao K, Jiang D K, Guo J, Wang J H and Sun X 2011 *DNA Res.* **18** 435
[5] Zhang C T and Zhang R 1991 *Nucleic Acids Res.* **19** 6313
[6] Yu J F, Sun X and Wang J H 2009 *J. Theor. Biol.* **261** 459
[7] Wood D, Setubal J, Kaul R, *et al.* 2001 *Science* **294** 2317
[8] Pruitt K D, Tatusova T and Maglott D R 2007 *Nucleic Acids Res.* **35** 61
[9] Gao F and Zhang C T 2004 *Bioinformatics* **20** 673
[10] Yu J F and Sun X 2010 *J. Comput. Chem.* **31** 2126
[11] Yu J F, Guo J, Liu Q B, Hou Y, Xiao K, Chen Q L, Wang J H and Sun X 2015 *Genes Genom.* **37** 347
[12] Zhang C T and Wang J 2000 *Nucleic Acids Res.* **28** 2804
[13] Burset M and Guigo R 1996 *Genomics* **34** 353
[14] Trifonov E N 1987 *J. Mol. Biol.* **194** 643
[15] Tatusov R L, Galperin M Y, Natale D A and Koonin E V 2000 *Nucleic Acids Res.* **28** 33
[16] Wang Q, Lei Y, Xu X, Wang G J and Chen L L 2013 *PLoS One* **7** e43176
[17] Liolios K, Chen I A, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz V M and Kyrpides N C 2010 *Nucleic Acids Res.* **38** D346
[18] Reed J L, Famili I, Thiele I and Palsson B O 2006 *Nat. Rev. Genet.* **7** 130
[19] Reeves G A, Talavera D and Thornton J M 2009 *J. R. Soc. Interface* **6** 129
[20] Skovgaard M, Jensen L J, Brunak S, Ussery D and Krogh A 2001 *Trends Genet.* **17** 425
[21] Salzberg S L 2007 *Genome Biol.* **8** 102
[22] Bakke P, Carney N, DeLoache W, Gearing M, Ingvorsen K, Lotz M, McNair J, Penumetcha P, Simpson S, Voss L, Win M, Heyer L and Campbell A 2009 *PLoS One* **4** e6291
[23] Yu J F, Jiang D K, Jin Y, Wang J H and Sun X 2012 *MATCH. Commun. Math. Comput. Chem.* **67** 845