Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, 2014, *15*, 95-103

Recognition of Protein-coding Genes Based on *Z*-curve Algorithms

Feng-Biao Guo[a], Yan Lin[b] and Ling-Ling Chen[c,*]

[a]*Center of Bioinformatics and Key Laboratory for NeuroInformation of the Ministry of Education, University of Electronic Science and Technology of China, Chengdu, 610054, China;* [b]*Department of Physics, Tianjin University, Tianjin 300072, China;* [c]*College of Life Science and Technology, Huazhong Agricultural University, Wuhan, 430070, China*

Abstract: Recognition of protein-coding genes, a classical bioinformatics issue, is an absolutely needed step for annotating newly sequenced genomes. The *Z*-curve algorithm, as one of the most effective methods on this issue, has been successfully applied in annotating or re-annotating many genomes, including those of bacteria, archaea and viruses. Two *Z*-curve based *ab initio* gene-finding programs have been developed: ZCURVE (for bacteria and archaea) and ZCURVE_V (for viruses and phages). ZCURVE_C (for 57 bacteria) and Zfisher (for any bacterium) are web servers for re-annotation of bacterial and archaeal genomes. The above four tools can be used for genome annotation or re-annotation, either independently or combined with the other gene-finding programs. In addition to recognizing protein-coding genes and exons, *Z*-curve algorithms are also effective in recognizing promoters and translation start sites. Here, we summarize the applications of *Z*-curve algorithms in gene finding and genome annotation.

Received on: September 15, 2013- Revised on: November 19, 2013- Accepted on: November 20, 2013

Keywords: Genome annotation, Genome re-annotation, *Z*-curve algorithm, ZCURVE, ZCURVE_V.

1. INTRODUCTION

Recognition of protein-coding genes is one of the most classical bioinformatics issues, and is an absolutely needed step for annotating newly sequenced genomes. Since the late 1970s, thousands of papers have been published on this issue. As for eukaryotic gene recognition, although great advances have been made, there is still plenty of room for improvement [1]. Compared with eukaryotic genomes, prokaryotes have a simpler gene structure and hence gene recognition is relatively straightforward [2]. Finding open reading frames (ORFs) is the first step in gene recognition in prokaryotes; however, arbitrarily assigning an ORF as a coding gene leads to a high rate of false positive prediction [3]. It is necessary to use computational methods to choose *bona fide* genes from these candidate ORF sets [4]. Pioneer researchers generally adopted a Markov chain to describe protein-coding genes [5, 6].

In 1991, Zhang and Zhang proposed algorithms to describe protein-coding genes and non-coding sequences using a graphical method, the *Z*-curve method [7, 8]. In 2000, *Z*-curve based algorithms were developed to recognize genes particularly in prokaryotes or intron-rare eukaryotes, such as yeast. Through the *Z*-curve algorithm, the complete gene set in the *Saccharomyces cerevisiae* was refined and was estimated to contain a total of about 5600 genes [9]. Also, this method was used in gene re-annotation of *Vibrio cholerae* and satisfactory results were obtained [10]. In 2003, a more advanced form of the method was developed, designated as ZCURVE 1.0, which could be used to perform *ab initio* gene finding in any newly sequenced bacterial or archaeal genomes [11]. In 2006, we developed ZCURVE_V, a ZCURVE-based program that specially performs *ab initio* gene finding in viral genomes [12]. Besides prokaryotes, the *Z*-curve algorithm could also be used to predict exons in eukaryotic genomes with high accuracy [13]. After extending the *Z*-curve method to include thousands of *Z*-curve parameters, it could also be used to predict human [14] or prokaryotic promoters [15] with very high accuracy.

*Address correspondence to this author at the College of Life Science and Technology, Huazhong Agricultural University, Wuhan, 430070, China; Tel/Fax: +86 27 87280877; E-mail: llchen@mail.hzau.edu.cn

2. THE *Z*-CURVE ALGORITHM

In every gene recognition method, there are two main parts, and the *Z*-curve method is not an exception. One is the recognizing features and the other is the discriminating (or classifying) method. In the *Z*-curve method, a series of features are derived based on the *Z*-curve theory of DNA sequences. Here we summarize the method as follows. The frequencies of bases A, C, G and T occurring in an ORF or a fragment of DNA sequence with bases at positions 1, 4, 7, …; 2, 5, 8, …, and 3, 6, 9, ..., are denoted by a_1, c_1, g_1, t_1; a_2, c_2, g_2, t_2; a_3, c_3, g_3, t_3, respectively. They are in fact the frequencies of bases at the 1st, 2nd and 3rd codon positions. Based on the *Z*-curve [7, 8], a_i, c_i, g_i, t_i are mapped onto a point P_i in a 3-dimensinal space V_i, i = 1, 2, 3. The coordinates of P_i, denoted by x_i, y_i, z_i, are determined by the Z-transform of the DNA sequence [9].

$$\begin{cases} x_i = (a_i + g_i) - (c_i + t_i), \\ y_i = (a_i + c_i) - (g_i + t_i), \quad x_i,\ y_i,\ z_i \in [-1,\ 1],\ i = 1,\ 2,\ 3. \\ z_i = (a_i + t_i) - (g_i + c_i). \end{cases} \quad (1)$$

1875-5488/14 $58.00+.00

The above 9 coordinates denote 9 classifying features. If we consider di-nucleotides at different codon positions, there will be 4×4×3×(3/4) = 36 features, which can be denoted by equation (2).

$$\begin{cases} x_k^X = (p_k(\mathrm{XA}) + p_k(\mathrm{XG})) - (p_k(\mathrm{XC}) + p_k(\mathrm{XT})), \\ y_k^X = (p_k(\mathrm{XA}) + p_k(\mathrm{XC})) - (p_k(\mathrm{XG}) + p_k(\mathrm{XT})), \quad \mathrm{X} = \mathrm{A,C,G,T}, \quad k = 12,23,31 \\ z_k^X = (p_k(\mathrm{XA}) + p_k(\mathrm{XT})) - (p_k(\mathrm{XG}) + p_k(\mathrm{XC})). \end{cases} \tag{2}$$

They are called phase dependent di-nucleotide parameters [11]. If all three codon positions are considered as a whole, there will be only 12 phase independent parameters [13], which could be described by equation (**3**).

$$\begin{cases} x_X = (p(\mathrm{XA}) + p(\mathrm{XG})) - (p(\mathrm{XC}) + p(\mathrm{XT})), \\ y_X = (p(\mathrm{XA}) + p(\mathrm{XC})) - (p(\mathrm{XG}) + p(\mathrm{XT})), \quad \mathrm{X} = \mathrm{A,C,G,T}, \\ z_X = (p(\mathrm{XA}) + p(\mathrm{XT})) - (p(\mathrm{XG}) + p(\mathrm{XC})). \end{cases} \tag{3}$$

The above *Z*-curve parameters could serve as classifying features when performing gene prediction in genomes. For convenience, we express these parameters by the united symbol u_n as follows.

$$\begin{cases} u_1 = x_1, \ u_2 = y_1, \ u_3 = z_1, \\ u_4 = x_2, \ u_5 = y_2, \ u_6 = z_2, \\ u_7 = x_3, \ u_8 = y_3, \ u_9 = z_3 \\ u_{10} = x_{12}^A, \ u_{11} = y_{12}^A, \ u_{12} = z_{12}^A, \\ u_{13} = x_{12}^C, \ u_{14} = y_{12}^C, \ u_{15} = z_{12}^C, \\ u_{16} = x_{12}^G, \ u_{17} = y_{12}^G, \ u_{18} = z_{12}^G, \\ u_{19} = x_{12}^T, \ u_{20} = y_{12}^T, \ u_{21} = z_{12}^T, \\ u_{22} = x_{23}^A, \ u_{23} = y_{23}^A, \ u_{24} = z_{23}^A, \\ u_{25} = x_{23}^C, \ u_{26} = y_{23}^C, \ u_{27} = z_{23}^C, \\ u_{28} = x_{23}^G, \ u_{29} = y_{23}^G, \ u_{30} = z_{23}^G, \\ u_{31} = x_{23}^T, \ u_{32} = y_{23}^T, \ u_{33} = z_{23}^T. \\ u_{34} = x_{31}^A, \ u_{35} = y_{31}^A, \ u_{36} = z_{31}^A, \\ u_{37} = x_{31}^C, \ u_{38} = y_{31}^C, \ u_{39} = z_{31}^C, \\ u_{40} = x_{31}^G, \ u_{41} = y_{31}^G, \ u_{42} = z_{31}^G, \\ u_{43} = x_{31}^T, \ u_{44} = y_{31}^T, \ u_{45} = z_{31}^T, \\ u_{46} = x_A, u_{47} = y_A, u_{48} = z_A, \\ u_{49} = x_C, u_{50} = y_C, u_{51} = z_C, \\ u_{52} = x_G, u_{53} = y_G, u_{54} = z_G, \\ u_{55} = x_T, u_{56} = y_T, u_{57} = z_T, \end{cases} \tag{4}$$

Usually, one classification method, such as the Fisher linear discriminant or Support Vector Machine, is also required to form a complete gene finding model. When a sufficient number of positive and negative samples have been prepared, we need to calculate values of all the Z parameters. With these values, one sample can correspond to a unique point in the high dimension space. The Fisher linear discriminant method can then be applied to locate a super-plane that differentiates the two kinds of samples as significantly as possible, in the high dimension space spanned by the *Z*-curve parameters. See details in [9] for how to determine the equation of the super-plane. After obtaining the super-plane equation, the distance from each new point to the super-plane can be computed and the new sample is determined to be positive or negative based on the distance value.

An example of *Ralstonia solanacearum*, a bacterium with high G+C content, is shown in (Fig. **1**). As can be seen, coding ORFs and non-coding ORFs are distributed in separate regions with minor overlapping. The space is spanned by the three most important axes using the principal component analysis of the variables u_1-u_{33} defined in equation (4). To obtain optimal prediction, in fact, we performed Fisher linear discriminant analysis for multiple times to choose coding genes from the ORF congregation for such high G+C content bacteria. Note that the figure shows the classifying schema in a 3-dimensional space, but the actual dimension is much larger.

3. *AB INITIO* GENE FINDING IN BACTERIAL AND ARCHAEAL GENOMES

The *ab initio* gene finding program that we developed was originally called ZCURVE 1.0 [11]. When implementing the method, 33 classifying variables, which correspond to u_1-u_{33} in equation (4), and the Fisher discriminant were used. The whole program contains five modules: (i) Choosing seed ORFs; (ii) Training the model; (iii) Finding all ORFs; (iv) Determining the coding potentials of all ORFs; (v) Eliminating the error prediction due to overlapping. Seed ORFs are those non-overlapping ORFs larger than 500 bp. These ORFs have a very high possibility (generally, >98%) to encode proteins. An ORF is defined here as the DNA sequence between a pair of in-frame start codon (ATG, GTG, or TTG) and stop codon (TAA, TAG, or TGA). At first, we seek all ORFs longer than 500 bp on the two strands. Subsequently, those ORFs with one or more bases overlapping with each other will be discarded. The retained non-overlapping long ORFs will constitute the reliable positive samples used in the training set. However, this rule does not remain true for the bacterial genomes with G+C content greater than 56%. We therefore have to seek another method called 9-D super sphere to obtain seed ORFs. Negative samples in the training set were obtained by randomly shuffling positive samples and destroying their natural structures. The Fisher discriminant algorithm is used to differentiate the positive and negative samples. The parameters of u_1-u_{33} are taken as classifying features. The decision of coding/non-coding for each ORF is simply made by the criterion of $\mathbf{c}\cdot\mathbf{u} > c_0 / \mathbf{c}\cdot\mathbf{u} < c_0$, where $\mathbf{c} = (c_1, \ c_2, \ ..., \ c_{33})^{\mathrm{T}}$, $\mathbf{u} = (u_1, \ u_2, \ ..., \ u_{33})^{\mathrm{T}}$, and "T" indicates the transpose of a matrix. The vector c constituting 33 Fisher coefficients could be determined by the training process. In ZCURVE 1.0, 90 bp is set as the minimum length of ORFs. However, users can change it to any integral value. All ORFs longer than the set length will be determined to be coding or non- coding ORFs. In fact, even these ORFs meeting $\mathbf{c}\cdot\mathbf{u} > c_0$ are not all genes, but are still needed to check for error prediction possibly due to overlapping with longer genes. If two ORFs are predicted as potential genes and have short overlapping regions, both will be retained as genes. Otherwise, one ORF must be disregarded due to either a lower Z score or smaller size.

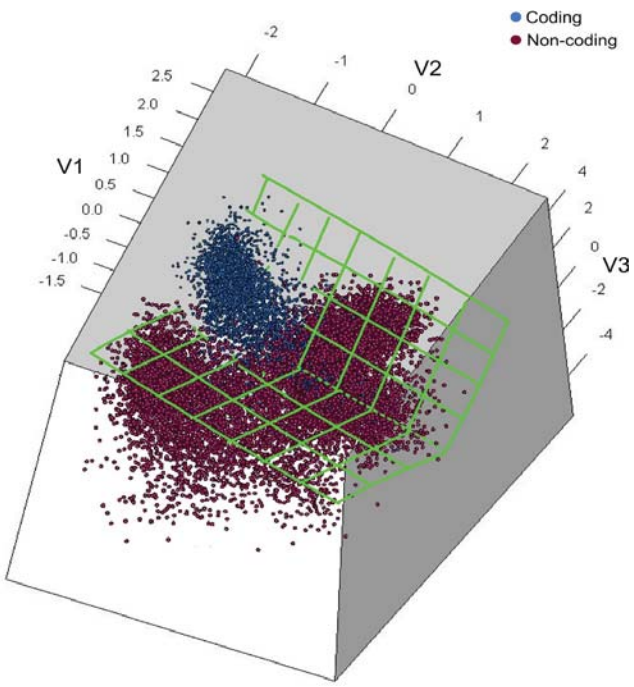

Fig. (1). An example illustrating the principle of the *Z*-curve algorithm for gene recognition. Each of the ORFs in *Ralstonia solanacearum* is mapped onto a point in the 33-D space derived from the variables u_1-u_{33}. To visualize the distribution pattern of the 14276 mapping points in the high dimensional space, the mapping points are projected onto the 3-D space spanned by the first three principal axes after the principal component analysis. In the figure, the blue circles denote 3244 coding ORFs and the purple circles indicate 11032 non-coding ORFs. The light green planes correspond to the two Fisher classifying planes, which can optimally differentiate the two types of points. Note that the all coding ORFs are clustered into a narrow region and very few of them overlap with non-coding ORFs.

ZCURVE 2.0 beta is the latest version of the software, which can be downloaded from http://tubic.tju.edu.cn/Zcurve_B. In this new version, the support vector machine is used instead of the Fisher linear discriminant because the former is more sensitive in gene classification. When running the ZCURVE program, users only need to provide the genomic sequence of the investigated bacterial strains. Finally, the program will output a file containing the chromosomal coordinates of all predicted genes. As of now, the ZCURVE program has over one hundred registered users such as University of Pennsylvania, Broad Institute of MIT, NITE institute in Japan, University of Warwick in England, and Max Planck Institute for Marine Microbiology in Germany. ZCURVE has been widely used in annotation of newly sequenced genomes and prediction of coding potentials of some genes of interest. For example, Egan and Waldor used ZCURVE prediction to confirm their result in the genome of *Vibrio cholerae* [16]. ZCURVE has been combined with other gene-finding programs into some metatools of bacterial gene finders, such as YACOP [17] and MORF [J. Waldmann and H. Teeling, unpublished], and is reviewed in [18]. ZCURVE has been used in at least 31 genomic projects independently or by combining with other well-known gene finders, or integrated into metatools [19-49]. (Table **1**) lists information about the 31 projects, including six metagenomic projects, two large phages and one large plasmid, as well as 28 bacterial genomes in 23 projects. Note that the actual number of genome projects using ZCURVE can be much higher than 31 because many using YACOP and MORFind without quoting the ZCURVE paper or website are not included in the list. Generally, over 97% of genuine genes could be found and the false positive rate would be less than 10%. To achieve more reliable results, the authors strongly suggest combining our method with one or two other *ab initio* gene finders when automatically annotating newly sequenced bacterial or archaeal genomes.

4. *AB INITIO* GENE FINDING IN VIRAL AND PHAGE GENOMES

Although virus genomes are much smaller than bacterial ones, annotation of viral genomes is still a difficult task. One of the problems in recognizing protein-coding genes in viruses is that the training set is usually unavailable. In the pioneer program GeneMarks for viruses [50], one heuristic method is used to collect seed ORFs. In each genome, functional proteins and other nucleotide sequences have quite different sequence composition. Furthermore, almost all the functional proteins, particularly those conserved proteins, have similar amino acid composition in one specific genome. Therefore, seed ORFs are selected based on the composition of amino acids.

We developed a method that uses only one seed ORF as the training set [12]. This ORF is the one that has the most bases in one specific virus and is very likely to be a protein-coding gene. We call this ORF the 'maximum ORF'. Investigation of more than 100 viral genomes proved that all maximum ORFs encode proteins. Therefore, the maximum ORF can be regarded as a reliable training set. Considering that there are so few seed ORFs, we used the Euclidean distance discriminant to choose genes from all found ORFs. For one candidate ORF and the maximum ORF, we calculate 33 features for all of them. Subsequently, the Euclidean distance between them is computed based on the 33 features. If the distance is shorter than $\sqrt{6.90}$, the candidate will preliminarily be predicted as a gene and otherwise not. In fact, we will furthermore determine whether the ORF is falsely predicted because it has significant overlapping with longer genes. Finally, we will obtain all predicted genes after discarding overlapping ORFs. We implement the above method in the program ZCURVE_V. Similar to ZCURVE, the virus version uses 33 parameters derived from the Z-transform of the DNA sequence. However, the methods for differentiating positive and negative samples and for generating seed ORFs are very different.

As a gene finder specially designed for viruses, phages and plasmids, ZCURVE_V can be used to annotate any anonymous genomes belonging to them. Scientists from the NCBI viral genome section listed ZCURVE_V as one of the three standard programs for viral gene finding [51]. The program has been integrated into the bacteriocin mining tool as an easily used module for ORF finding. It has been used to annotate genomes of at least two bacteria, one virus and sixteen phages [52-59] (Table **2**). A prominent advantage of

Table 1. Genomic projects involving the ZCURVE system.

Genome	Tool	Year	Reference
Lactobacillus salivarius	YACOP	2006	[19]
Escherichia coli phage, named Rtp	GeneMarks, ZCURVE	2006	[20]
Symbiont metagenome in *Olavius algarvensis*	MORFind	2006	[21]
Magnetospirillum magnetotacticum MS-1 and *M. gryphiswaldense* MSR-1	MORFind	2007	[22]
Human gut mobile metagenome	ZCURVE, Glimmer	2007	[23]
The filamentous *Beggiatoa*	MORFind	2007	[24]
Fosmid of marine *Planctomycetes*	MORFind	2007	[25]
Mycobacterium tuberculosis H37Ra	ZCURVE	2008	[26]
Fosmid of methanotrophic Archaea (ANME)	MORFind	2009	[27]
Desulfobacterium autotrophicum HRM2	YACOP	2009	[28]
Phaeobacter gallaeciensis DSM 17395	YACOP	2009	[29]
Amycolatopsis mediterranei U32	ZCURVE, Glimmer, GeneMark	2010	[30]
Bacillus thuringiensis BMB171	Glimmer, ZCURVE	2010	[31]
Variovorax paradoxus S110	YACOP	2011	[32]
Bacillus megaterium WSH-002	ZCURVE, Glimmer	2011	[33]
Ketogulonicigenium vulgare WSH-001	ZCURVE, Glimmer	2011	[34]
Haloarcula hispanica	ZCRUVE, Glimmer	2011	[35]
Mycoplasma bovis Hubei-1	ZCURVE, Glimmer	2011	[36]
Brucella melitensis M28 and M5-90	ZCURVE, Glimmer	2011	[37]
Acinetobacter baumannii MDR-TJ	ZCURVE, Glimmer, GeneMark	2012	[38]
Staphylococcus aureus D139, H19, E1410, M809, and WW2703/97	ZCURVE, Glimmer, GeneMark MetaGene	2012	[39]
Cluster of myxobacteria	YACOP	2012	[40]
Haloferax mediterranei	ZCURVE, Glimmer	2012	[41]
Bifidobacterium longum JDM301	ZCURVE, Glimmer	2012	[42]
Siphophage VHS1 from *Vibrio harveyi*	Zcurve, GeneMarkS, EasyGene, MetaGene, Genewise, Glimmer	2012	[43]
Oceaniovalibus guishaninsula JLT2003	ZCURVE, Glimmer	2012	[44]
Streptomyces hygroscopicus 5008	ZCURVE, Glimmer	2012	[45]
Tistrella mobilis KA081020-065	ZCURVE, Glimmer	2012	[46]
Glaciecola psychrophila 170T	ZCURVE, Glimmer	2013	[47]
Moraxella catarrhalis	ZCURVE, Prodigal, GeneMarkHMM, Glimmer	2013	[48]
Klebsiella pneumoniae plasmid pKF3-140	ZCURVE, Glimmer	2013	[49]

ZCURVE_V is that it can accurately predict genes in viral genomes even as short as about 1000 nucleotides, in addition to the advantage of being able to run online (http://tubic.tju.edu.cn/Zcurve_V), without the need to install locally. Therefore, ZCURVE_V can be preferably used when annotating short viral genomes. Alternatively, users can combine ZCURVE_V and GeneMarks or homology search to gain more reliable results.

Table 2. Genomic projects involving the ZCURVE_V system.

Genome	Tool	Year	Reference
Me Tri virus	ZCURVE_V	2008	[52]
VP882 phage of *Vibrio parahaemolyticus O3:K6*	ZCURVE_V GeneMark.hmm, Glimmer	2009	[53]
Clostridium acetobutylicum EA 2018	ZCURVE_V, Glimmer	2011	[54]
Escherichia coli O157:H7 Lytic Phage AR1	ZCURVE_V, GeneMark.hmm	2001	[55]
Pseudomonas aeruginosa Strain AH16	ZCURVE_V, Glimmer	2012	[56]
Lactococcal phages Q33 and BM13	ZCURVE_V, ORFinder, GenMark	2013	[57]
VP3 phage of *Vibrio cholerae*	ZCURVE_V, GeneMark, Glimmer	2013	[58]
Eleven lactococcal 936-type phages	ZCRUVE_V, GeneMark.hmm	2013	[59]

5. GENOME RE-ANNOTATION IN BACTERIAL AND VIRAL GENOMES

Considering that the protein-coding genes in sequenced genomes are annotated with gene-finding programs, only a few are verified with experiments. The sequenced genomes often contain false-positive and false-negative annotations, especially in GC-rich genomes [60-66]. False-positive annotation means that some non-coding ORFs were incorrectly predicted as protein-coding genes (most of them are short ORFs without functional information), and false-negative annotation indicates that protein-coding genes are missed in the sequenced genomes. Most of the gene-identification programs achieve good results in low GC content genomes, however, the recognition accuracy drops rapidly in high GC content genomes since these genomes contain fewer stop codons and more spurious ORFs.

Generally, ORFs in annotation files of microbial genomes are divided into two groups. The first group contains genes with known functions, and the second group contains "hypothetical", "unknown" or "predicted" ORFs, which involve false-positive prediction. Based on the assumption that the statistical features of DNA sequences of the two groups are similar, Wang and Zhang identified 172 annotated genes as non-coding ORFs in *V. cholerae* based on the *Z*-curve method [10]. Chen and Zhang combined 18 Z parameters (u_1-u_6, u_{46}-u_{57}) and the Fisher discriminant into a program, called ZCURVE_C, to recognize the "hypothetical ORFs" in 57 microbial genomes [67], and the program is available at the website http://tubic.tju.edu.cn/ZCURVE_C/Default.cgi. Guo *et al.* re-annotated the genome of a hyper-thermophilic crenarchaeon *Aeropyrum pernix* K1 by combining the 9 Z parameters (u_1-u_9) and K-means clustering and identified many false-positive ORFs [68, 69]. *Amsacta moorei* entomopoxvirus is a typical over-annotated virus, Guo and Yu suggested that 38 of 294 originally annotated genes did not encode proteins based on the 9 *Z*-curve parameters (u_1-u_9) and the Fisher discriminant method [70]. In 2008, Chen *et al.* re-annotated the plant pathogen genome *Erwinia carotovora* subsp. *atroseptica* SCRI1043 and identified that 49 originally annotated 'hypothetical genes' should be non-coding ORFs based on the *Z*-curve method with 21 parameters (u_1-u_9, u_{46}-u_{57}). Theoretical evidence of principal component analysis (PCA), clusters of orthologous groups of proteins (COG) occupation, and average length distribution showed that the identified non-coding ORFs were highly unlikely to encode proteins [71]. Using sequence alignment tools and some functional resources, they also predicted the functions of hundreds of 'hypothetical genes' [71]. In 2011, Du *et al.* performed a re-annotation in the genome of *Pyrobaculum Aerophilum*. Consequently, 25 hypothetical ORFs were eliminated by using the method of the 33 Z parameters (u_1-u_{33}) with the Fisher discriminant. Recently, Wang *et al.* re-annotated *Agrobacterium tumefaciens* strain C58 genome, and 29 originally annotated 'hypothetical genes' were recognized as non-coding ORFs by using the *Z*-curve method with 21 parameters ((u_1-u_9, u_{46}-u_{57}) [72]. Wang *et al.* also used reverse transcription-PCR (RT-PCR) experiments to verify their prediction. Nearly 80% of the non-coding ORFs or newly predicted protein-coding genes were verified with RT-PCR experiments [73]. Very recently, Guo *et al.* performed a comprehensive analysis to re-annotate 10 complete genomes of the *Neisseriaceae* family [74]. Transcriptions of over 80% of genes newly found by the ZCURVE program could be experimentally validated by RT-PCR. In the work, the authors constructed a new web server Zfisher, which can be used to examine coding potentials of hypothetical proteins in any annotated genomes and is freely available at http://147.8.74.24/Zfisher/. All the above cases showed that the *Z*-curve method is highly accurate in re-annotating microbial genomes.

6. RECOGNITION OF HUMAN SHORT EXONS

Accurately predicting short exons in genomes of human and other eukaryotes is a rather difficult issue. To improve the accuracy in predicting human exons, our group compared 19 algorithms including the *Z*-curve algorithm [13]. Based on a standard human dataset, the *Z*-curve algorithm with 69 parameters could differentiate coding and non-coding sequences as short as 192 bp with an accuracy of 96.2%, and the accuracy was above 82% for 72 bp sequences. Such accuracy was even higher than the 5th order Markov model, and consistent results were obtained by other groups [75]. Kellis and coworkers compared four types of single species methods, including Fourier transform, codon bias, interpolated context models and the *Z*-curve algorithm, and found that the *Z*-curve method showed the best performance in the

Drosophila [76]. Recently, Song *et al.* improved the accuracy to over 98% for 192 bp human sequences by combining Z-curve parameters and kernel partial least squares [77].

7. RECOGNITION OF PROMOTERS AND TRANSLATION START SITES

The *Z*-curve algorithm may be used to recognize promoters, which play an essential role in determining where the transcription of a particular gene should be initiated. By combining the Z parameters of phase-dependent single nucleotide and di-nucleotides, which are listed in equations (1) and (2), and the Fisher discriminant analysis, Yang *et al.* obtained a satisfactory accuracy (over 85%) for classifying human Pol II promoters [14]. Song further extended the number of Z parameters by considering phase independent single nucleotide, di-nucleotides, tri-nucleotides, …… and w-nucleotides [15]. Song referred to the unlimited Z parameters as variable-window *Z*-curve features. When considering six nucleotides, a total of 4095 variables were obtained. Using partial least square to filter features, a subset of 220 parameters could be used to obtain an accuracy of about 95% in prokaryotes [15].

In addition to identifying promoters, a novel algorithm that is based on characteristic *Z*-curve patterns around translation start sites (TSS) was developed to identify TSSs [78]. For instance, three *Z*-curve components for nucleotides around *E. coli* and *B. subtilis* TSSs validated experimentally show distinct patterns from those of false TSSs (Fig. **2**). Taking the *x* component as an example, true TSSs, but not false ones, have a jump in the region of -14 to -7, and have apparent three-base periodic patterns in sequences only downstream of TSS (Fig. **2**). These mononucleotide distribution patterns around TSS were used to recognize bacterial TSSs, and an online program, GS-finder, was developed to implement this algorithm [78]. The *Z*-curve method has also been used to study nucleosome positioning in the yeast genome [79]. Therefore, the *Z*-curve algorithm can find applications in a wide array of areas including promoter and TSS identification.

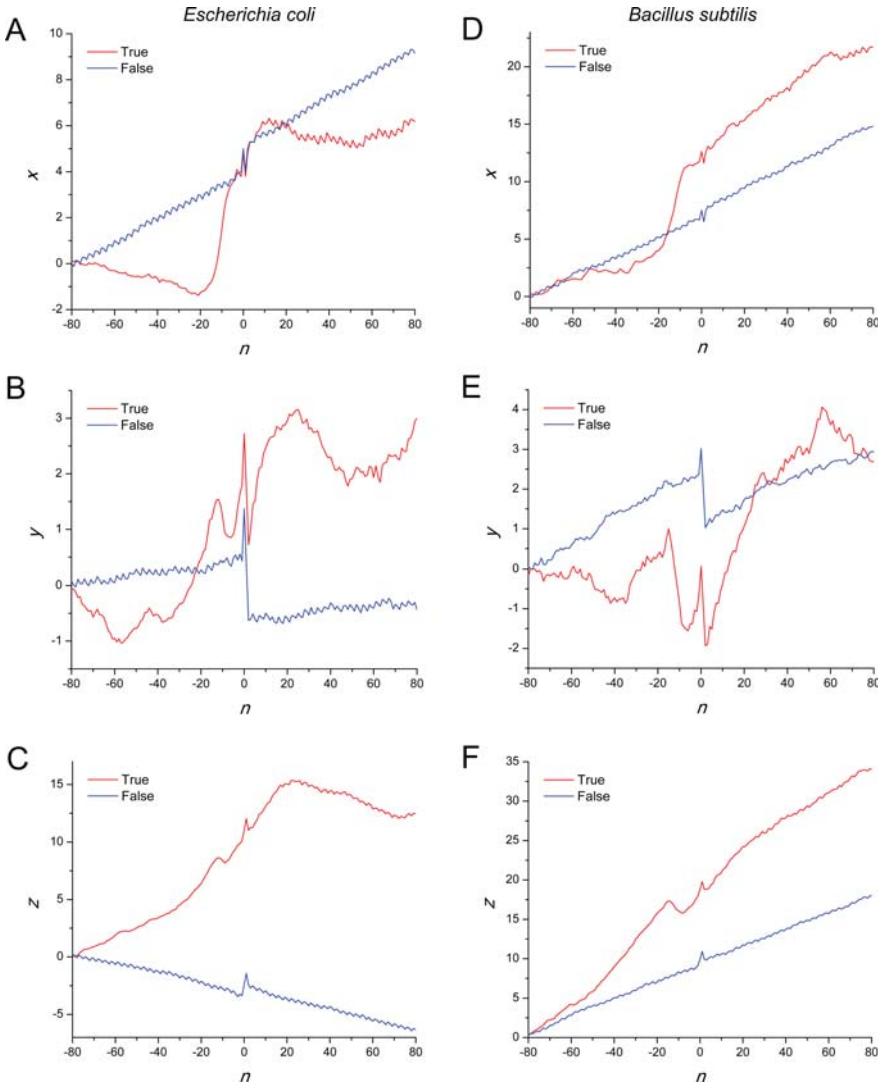

Fig. (2). Three *Z*-curve components show distinct patterns around translation start sites (TSSs). Averaged *x*, *y* and *z* components (**A**, **B** and **C**, respectively), for nucleotides around 195 experimentally verified TSSs in *E. coli*. Averaged *x*, *y* and *z* components (**D**, **E** and **F**, respectively), for 58 experimentally verified TSSs in *B. subtilis*.

8. CONCLUSION

In this review paper we summarize the principle of the *Z*-curve method and its wide applications in eukaryotic and prokaryotic gene recognition. Two versatile programs, ZCURVE, for automatic annotation of bacterial and archaeal genomes and ZCURVE_V, for automatic annotation of viral and phage genomes, are extensively described. Considering the excellent performance of the method in gene recognition, we hope that the *Z*-curve algorithm will find more and more applications in genome analysis.

NOTE ADDED IN PROOF

Recently Weissman and coworkers discovered that ribosomes can read through stop codons in a regulated manner, and nucleotide compositions characterized by Z-curve parameters are distinct among coding regions, UTRs and novel extensions [80].

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

We are very grateful to Prof. Chun-Ting Zhang for inspiring discussions. We thank Mr. Zhong-Shan Cheng, Zhi-Gang Hua and Yuan-Nong Ye for help in making the figures. This study was supported by the program for New Century Excellent Talents in University (grant NCET-11-0059), the National Natural Science Foundation of China (grants 31071109, 31071659, 31271406).

REFERENCES

[1] Maji, S.; Garg, D. Progress in Gene Prediction: Principles and Challenges. *Curr. Bioinform.*, **2013**, *8*, 226 – 243

[2] Richardson, E.J.; Watson, M. The automatic annotation of bacterial genomes. *Brief. Bioinform.*, **2013**, *14*, 1-12.

[3] Besemer, J.; Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*., **2005**, *33*, W451-4.

[4] Delcher, A.L.; Bratke, K.A.; Powers, E.C.; Salzberg, S.L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinform.*, **2007**, 23, 673-9.

[5] Borodovsky, M.; Mcininch, J. GeneMark: parallel gene recognition for both DNA strands. *Comput. & Chem.*, **1993**, *17*, 123-133.

[6] Salzberg, S.L.; Delcher, A.L.; Kasif, S.; White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **1998**, *26*, 544-8.

[7] Zhang, C.T.; Zhang, R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res*., **1991**, *19*, 6313-7.

[8] Zhang, C.T.; Zhang, R. Diagrammatic representation of the distribution of DNA bases and its applications. *Int. J. Biol. Macromol*., **1991**, *13*, 45-9.

[9] Zhang, C.T.; Wang, J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the *Z*-curve. *Nucleic Acids Res.*, **2000**, *28*, 2804-14.

[10] Wang, J.; Zhang, C.T. Identification of protein-coding genes in the genome of *Vibrio cholerae* with more than 98% accuracy using occurrence frequencies of single nucleotides. *Eur. J. Biochem.*, **2001,** *268*, 4261-8.

[11] Guo, F.B.; Ou, H.Y.; Zhang, C.T. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res*., **2003,** *31*, 1780-9

[12] Guo, F.B.; Zhang, C.T. ZCURVE_V: a new self-training system for recognizing protein-coding genes in viral and phage genomes. *BMC Bioinform.*, **2006**, *10*, 7-9.

[13] Gao, F; Zhang, C.T. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinform.*, **2004**, *20*, 673-81.

[14] Yang, J.Y.; Zhou, Y.; Yu, Z.G.; Anh, V.; Zhou, L.Q. Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. *BMC Bioinform.*, **2008**, *9*, 113.

[15] Song, K. Recognition of prokaryotic promoters based on a novel variable-window *Z*-curve method. *Nucleic Acids Res*., **2012**, *40*, 963-71.

[16] Egan, E.S.; Waldor, M.K. Distinct replication requirements for the two *Vibrio cholerae* chromosomes. *Cell*, **2003**, *114*, 521-30.

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

[18] Overbeek, R.; Bartels, D.; Vonstein, V.; Meyer, F. Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem. Rev*., **2007**, *107*, 3431-47.

[19] Claesson, M.J.; Li, Y.; Leahy, S.; Canchaya, C.; van Pijkeren. J.P.; Cerdeno-Tárraga, A.M.; Parkhill, J.; Flynn, S.; O'Sullivan, GC.; Collins, J.K.; Higgins, D.; Shanahan, F.; Fitzgerald, G.F.; van Sinderen, D.; O'Toole, P.W. Multireplicon genome architecture of *Lactobacillus salivarius*. *Proc. Natl. Acad. Sci. U.S.A*., **2006**, *103*, 6718-23.

[20] Wietzorrek, A.; Schwarz, H.; Herrmann, C.; Braun, V. The genome of the novel phage Rtp, with a rosette-like tail tip, is homologous to the genome of phage T1. *J. Bacteriol.*, **2006**, *188*, 1419-36.

[21] Woyke, T.; Teeling, H.; Ivanova, N.N.; Huntemann, M.; Richter, M.; Gloeckner, F.O.; Boffelli, D.; Anderson, I.J.; Barry, K.W.; Shapiro, H.J.; Szeto, E.; Kyrpides, N.C.; Mussmann, M.; Amann, R; Bergin, C.; Ruehland, C.; Rubin, E.M.; Dubilier, N. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, **2006**, *443*, 950-5.

[22] Richter, M.; Kube, M.; Bazylinski, D.A.; Lombardot, T.; Glockner, F.O.; Reinhardt, R.; Schüler, D. Comparative genome analysis of four magnetotactic bacteria reveals a complex set of group-specific genes implicated in magnetosome biomineralization and function. *J. Bacteriol*., **2007**, *189*, 4899-910.

[23] Jones, B.V.; Marchesi, J.R.; Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods*, **2007**, *4*, 55-61.

[24] Mussmann, M.; Hu, F.Z.; Richter, M.; de Beer, D.; Preisler, A.; Jorgensen, B.B.; Huntemann, M.; Glockner, F.O.; Amann, R.; Koopman, W.J.; Lasken, R.S.; Janto, B.; Hogg, J.; Stoodley, P.; Boissy, R.; Ehrlich, G.D. Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biol.,* **2007**, *5*, e230.

[25] Woebken, D.; Teeling, H.; Wecker, P.; Dumitriu, A.; Kostadinov, I.; Delong, E.F.; Amann, R.; Glockner, F.O. Fosmids of novel marine *Planctomycetes* from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes. *ISME J*., **2007**, *1*, 419-35.

[26] Zheng, H.; Lu, L.; Wang, B.; Pu, S.; Zhang, X.; Zhu, G.; Shi, W.; Zhang, L.; Wang, H.; Wang, S.; Zhao, G.; Zhang, Y. Genetic basis of virulence attenuation revealed by comparative genomic analysis of *Mycobacterium tuberculosis* strain H37Ra versus H37Rv. *PLoS One*, **2008**, *3*, e2375.

[27] Meyerdierks, A.; Kube, M.; Kostadinov, I.; Teeling, H.; Glockner, F.O, Reinhardt, R.; Amann, R. Metagenome and mRNA expression analyses of anaerobic methanotrophic archaea of the ANME-1 group. *Environ. Microbiol*., **2010**, *12*, 422-39.

[28] Strittmatter, A.W.; Liesegang, H.; Rabus, R.; Decker, I.; Amann, J.; Andres, S.; Henne, A.; Fricke, W.F.; Martinez-Arias, R.; Bartels, D.; Goesmann, A.; Krause, L.; Pühler, A.; Klenk, H.P.; Richter, M.; Schüler, M.; Glockner, F.O.; Meyerdierks, A.; Gottschalk, G.; Amann, R. Genome sequence of *Desulfobacterium autotrophicum* HRM2, a marine sulfate reducer oxidizing organic carbon completely to carbon dioxide. *Environ. Microbiol.*, **2009**, *11*, 1038-55.

[29] Zech, H.; Thole, S.; Schreiber, K.; Kalhofer, D.; Voget, S.; Brinkhoff, T.; Simon, M.; Schomburg, D.;, Rabus, R.; Growth phase-dependent global protein and metabolite profiles of *Phaeobacter gallaeciensis* strain DSM 17395, a member of the marine Roseobacter-clade. *Proteomics*, **2009**, *9*, 3677-97.

[30] Zhao, W.; Zhong, Y.; Yuan, H.; Wang, J.; Zheng, H.; Wang, Y.; Cen, X.; Xu, F.; Bai, J.; Han, X.; Lu, G.; Zhu, Y.; Shao, Z.; Yan, H.; Li, C.; Peng, N.; Zhang, Z.; Zhang, Y.; Lin, W.; Fan, Y.; Qin, Z.; Hu, Y.; Zhu, B.; Wang, S.; Ding, X.; Zhao, G.P. Complete ge-

nome sequence of the rifamycin SV-producing *Amycolatopsis mediterranei* U32 revealed its genetic characteristics in phylogeny and metabolism. *Cell Res.*, **2010**, *20*, 1096-108.

[31] He, J.; Shao, X.; Zheng, H.; Li, M.; Wang, J,; Zhang, Q.; Li, L.; Liu, Z.; Sun, M.; Wang, S.; Yu, Z. Complete genome sequence of Bacillus thuringiensis mutant strain BMB171. *J. Bacteriol*., **2010**, 192, 4074-5.

[32] Han, J.I.; Choi, H.K.; Lee, S.W.; Orwin, P.M.; Kim, J.; Laroe, S.L.; Kim, T.G.; O'Neil, J.; Leadbetter, J.R.; Lee, S.Y.; Hur, C.G.; Spain, J.C.; Ovchinnikova, G.; Goodwin, L.; Han, C. Complete genome sequence of the metabolically versatile plant growth-promoting endophyte *Variovorax paradoxus* S110. *J. Bacteriol.,* **2011**, *193*, 1183-90.

[33] Liu, L.; Li, Y.; Zhang, J.; Zou, W.; Zhou, Z.; Liu, J.; Li, X.; Wang, L.; Chen, J. Complete genome sequence of the industrial strain *Bacillus megaterium* WSH-002. *J. Bacteriol.*, **2011**, *193*, 6389-90.

[34] Liu, L.; Li, Y.; Zhang, J.; Zhou, Z.; Liu, J.; Li, X.; Zhou, J.; Du, G.; Wang, L.; Chen, J. Complete genome sequence of the industrial strain *Ketogulonicigenium vulgare* WSH-001. *J. Bacteriol*., **2011**, *193*, 6108-9.

[35] Liu, H.; Wu, Z.; Li, M.; Zhang, F.; Zheng, H.; Han, J.; Liu, J.; Zhou, J.; Wang, S.; Xiang, H. Complete genome sequence of *Haloarcula hispanica*, a Model *Haloarchaeon* for studying genetics, metabolism, and virus-host interaction. *J. Bacteriol.*, **2011**, *193*, 6086-7.

[36] Li, Y.; Zheng, H.; Liu, Y.; Jiang, Y.; Xin, J.; Chen, W.; Song, Z. The complete genome sequence of *Mycoplasma bovis* strain Hubei-1. *PLoS One*., **2011**, *6*, e20999.

[37] Wang, F.; Hu, S.; Gao, Y.; Qiao, Z.; Liu, W.; Bu, Z. Complete genome sequences of *Brucella melitensis* strains M28 and M5-90, with different virulence backgrounds. *J. Bacteriol.*, **2011**, *193*, 2904-5.

[38] Huang, H.; Yang, Z.L.; Wu, X.M.; Wang, Y.; Liu, Y.J.; Luo, H.; Lv, X.; Gan, Y.R.; Song, S.D.; Gao, F. Complete genome sequence of *Acinetobacter baumannii* MDR-TJ and insights into its mechanism of antibiotic resistance. *J. Antimicrob. Chemother*., **2012**, *67*, 2825-32.

[39] Thomas, J.C.; Godfrey, P.A.; Feldgarden, M.; Robinson, D.A. Candidate targets of balancing selection in the genome of *Staphylococcus aureus*. *Mol. Biol. Evol*., **2012**, *29*, 1175-86.

[40] Brinkhoff, T.; Fischer, D.; Vollmers, J.; Voget, S.; Beardsley, C.; Thole, S.; Mussmann, M.; Kunze, B.; Wagner-Dobler, I.; Daniel, R.; Simon, M. Biogeography and phylogenetic diversity of a cluster of exclusively marine myxobacteria. *ISME J*., **2012**, *6*, 1260-72.

[41] Han, J.; Zhang, F.; Hou, J.; Liu, X.; Li, M.; Liu, H.; Cai, L.; Zhang, B.; Chen, Y.; Zhou, J.; Hu, S.; Xiang, H. Complete genome sequence of the metabolically versatile halophilic archaeon *Haloferax mediterranei,* a poly(3-hydroxybutyrate-co-3-hydroxyvalerate) producer. *J. Bacteriol.*, **2012**, *194*, 4463-4.

[42] Wei, Y.X.; Zhang, Z.Y.; Liu, C.; Malakar, P.K.; Guo, X.K. Safety assessment of *Bifidobacterium longum* JDM301 based on complete genome sequences. *World J. Gastroenterol*., **2012**, *18*, 479-88.

[43] Khemayan, K.; Prachumwat, A.; Sonthayanon, B.; Intaraprasong, A.; Sriurairatana. S.; Flegel, T.W. Complete genome sequence of virulence-enhancing Siphophage VHS1 from *Vibrio harveyi. Appl. Environ. Microbiol*., **2012**,*78*, 2790-6.

[44] Tang, K.; Liu, K.; Jiao, N. Draft genome sequence of *Oceaniovalibus guishaninsula* JLT2003T. *J. Bacteriol*., **2012**, *194*(23), 6683.

[45] Wu, H.; Qu, S.; Lu, C.; Zheng, H.; Zhou, X.; Bai, L.; Deng, Z. Genomic and transcriptomic insights into the thermo-regulated biosynthesis of validamycin in *Streptomyces hygroscopicus* 5008. *BMC Genomics*, **2012**, *13*, 337.

[46] Xu, Y.; Kersten, R.D.; Nam, S.J.; Lu, L.; Al-Suwailem, A.M.; Zheng, H.; Fenical, W.; Dorrestein, P.C.; Moore, B.S.; Qian, P.Y. Bacterial biosynthesis and maturation of the didemnin anti-cancer agents. *J. Am. Chem. Soc*., **2012**, *134*, 8625-32.

[47] Yin, J.; Chen, J.; Liu, G.; Yu, Y.; Song, L.; Wang, X.; Qu, X. Complete Genome Sequence of Glaciecola psychrophila Strain 170T. *Genome Announc*., **2013**, *1*, e00199-13.

[48] de Vries, S.P.; Burghout, P.; Langereis, J.D.; Zomer, A.; Hermans, P.W.; Bootsma, H.J. Genetic requirements for *Moraxella catarrhalis* growth under iron-limiting conditions. *Mol. Microbiol*., **2013**, *87*, 14-29.

[49] Bai, J.; Liu, Q.; Yang, Y.; Wang, J.; Yang, Y.; Li, J.; Li, P.; Li, X.; Xi, Y.; Ying, J.; Ren, P.; Yang, L.; Ni, L.; Wu, J.; Bao, Q.; Zhou, T.G. Insights into the evolution of gene organization and multidrug resistance from *Klebsiella pneumoniae* plasmid pKF3-140. *Gene*, **2013**, *519*, 60-6.

[50] 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*., **2001**, *29*, 2607-18.

[51] Brister, J.R.; Bao, Y.; Kuiken, C.; Lefkowitz, E.J.; Le Mercier, P.; Leplae, R.; Madupu, R.; Scheuermann, R.H.; Schobel, S.; Seto, D.; Shrivastava, S.; Sterk, P.; Zeng, Q.; Klimke, W.; Tatusova, T; Towards Viral Genome Annotation Standards, Report from the 2010 NCBI Annotation Workshop. *Viruses*, **2010**, *2*, 2258-68.

[52] Tan le, V.; Ha do, Q.; Hien, V.M.; van der Hoek, L.; Farrar, J.; de Jong, M.D, Me Tri virus: a Semliki Forest virus strain from Vietnam? *J. Gen. Virol*., **2008,** *89*, 2132-5.

[53] Lan, S.F.; Huang, C.H.; Chang, C.H.; Liao. W.C.; Lin, I.H.; Jian, W.N.; Wu, Y.G.; Chen, S.Y.; Wong, H.C. Characterization of a new plasmid-like prophage in a pandemic *Vibrio parahaemolyticus* O3:K6 strain. *Appl. Environ. Microbiol*., **2009**, *75*, 2659-67.

[54] Hu, S.; Zheng, H.; Gu, Y.; Zhao, J.; Zhang, W.; Yang, Y.; Wang, S.; Zhao, G.; Yang, S.; Jiang, W. Comparative genomic and transcriptomic analysis revealed genetic characteristics related to solvent formation and xylose utilization in *Clostridium acetobutylicum* EA 2018. *BMC Genomics*, **2011**, *12*, 93.

[55] Liao, W.C.; Ng, W.V.; Lin, I.H.; Syu, W.J.; Liu, T.T., Chang, C.H. T4-Like genome organization of the *Escherichia coli* O157:H7 lytic phage AR1. *J. Virol.*, **2011**, *85*, 6567-78.

[56] Wu, D.Q.; Cheng, H.; Wang, C.; Zhang, C.; Wang, Y.; Shao, J.; Duan, Q. Genome sequence of *Pseudomonas aeruginosa* strain AH16, isolated from a patient with chronic pneumonia in China. *J. Bacteriol*., **2012**, *194*, 5976-7.

[57] Mahony, J.; Martel, B.; Tremblay, D.M.; Neve, H.; Heller, K.J.; Moineau, S.; van Sinderen, D. Molecular analysis of lactococcal phages Q33 and BM13: Identification of a new P335 subgroup. *Appl. Environ. Microbiol*., **2013**, *79*, 4401-9.

[58] Li, W.; Zhang, J.; Chen, Z.; Zhang, Q.; Zhang, L.; Du, P.; Chen, C.; Kan, B. The genome of VP3, a T7-like phage used for the typing of *Vibrio cholerae*. *Arch. Virol.*, **2013**, *158*, 1865-76.

[59] Mahony, J.; Kot, W.; Murphy, J.; Ainsworth, S.; Neve, H.; Hansen, L.H.; Heller, K.J.; S?rensen, S.J.; Hammer, K.; Cambillau, C.; Vogensen, F.K.; van Sinderen, D. Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. *Appl. Environ. Microbiol.*, **2013**, *79*, 4385-92.

[60] Nielsen, P.; Krogh, A. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinform.*, **2005**, *21*, 4322-4329.

[61] Jones, C.E.; Brown, A.L.; Baumann, U. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinform.*, **2007**, *8*, 170.

[62] Salzberg, S.L. Genome re-annotation: a wiki solution? *Genome Biol*., **2007**, *8*, 102.

[63] Nagy, A.; Hegyi, H.; Farkas, K.; Tordai, H.; Kozma, E.; Bányai, L.; Patthy, L. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. *BMC Bioinformatics*, **2008**, *9*, 353.

[64] Poptsova, M.S.; Gogarten, J.P. Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiol. SGM*, **2010**, *156*, 1909–1917.

[65] Yu, J.F.; Xiao, K.; Jiang, D.K.; Guo, J.; Wang, J.H.; Sun, X. An integrative method for identifying the over-annotated protein-coding genes in microbial genomes. *DNA Res*., **2011**, *18*, 435–449.

[66] Yu, J.F.; Guo, Z.Z.; Sun, X.; Wang, J.H. A review of the computational methods for identifying the over-annotated genes and missing genes in microbial genomes. *Curr. Bioinform*., **2013**, 9(2), 147-154.

[67] Chen, L.L.; Zhang, C.T. (2003) Gene recognition from questionable ORFs in bacterial and archaeal genomes. *J. Biomol. Struct. Dyn.*, **2003**, *21*, 99-110.

[68] Guo, F.B.; Wang, J.; Zhang, C.T. Gene recognition based on nucleotide distribution of ORFs in a hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res*., **2004**, *11*, 361-70.

[69] Guo, F.B.; Lin, Y. Identify protein-coding genes in the genomes of *Aeropyrum pernix* K1 and *Chlorobium tepidum* TLS. *J. Biomol. Struct. Dyn.*, **2009**, *26*, 413-20.

[70] Guo, F.B.; Yu, X.J. Re-prediction of protein-coding genes in the genome of *Amsacta moorei entomopoxvirus*. *J. Virol. Methods.*,

2007, *146*, 389-92.

[71] Chen, L.L.; Ma, B.G.; Gao, N. (2008) Reannotation of hypothetical ORFs in plant pathogen *Erwinia carotovora* subsp. atroseptica SCRI1043. *FEBS J.*, **2008**, *275*, 198–206.

[72] Du, M.Z.; Guo, F.B.; Chen, Y.Y. Gene re-annotation in genome of the extremophile *Pyrobaculum aerophilum* by using bioinformatics methods. *J. Biomol. Struct. Dyn*., **2011**, *29*, 391-401.

[73] Wang, Q.; Lei, Y.; Xu, X.W.; Wang G.; Chen L.L. (2012) Theoretical prediction and experimental verification of protein-coding genes in plant pathogen genome *Agrobacterium tumefaciens* strain C58. *PLoS One*, **2012**, *7*, e43176.

[74] Guo, F.B.; Xiong, L.; Teng, J.L.; Yuen, K.Y.; Lau, S.K.; Woo PC. Re-annotation of protein-coding genes in 10 complete genomes of *Neisseriaceae* family by combining similarity-based and composition-based methods. *DNA Res.*, **2013**, *20*, 273-86.

[75] Saeys, Y.; Rouzé, P.; Van de Peer, Y. In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. *Bioinformatics*., **2007**, *23*, 414-20.

[76] Lin, M.F.; Deoras, A.N.; Rasmussen, M.D.; Kellis, M. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput Biol*., **2008**, *4*, e1000067.

[77] Song, K.; Zhang, Z.; Tong, T.P.; Wu, F. Classifier assessment and feature selection for recognizing short coding sequences of human genes. *J. Comput. Biol*., **2012**, *19*, 251-60.

[78] Ou HY, Guo FB, Zhang CT. GS-Finder: a program to find bacterial gene start sites with a self-training method. Int J Biochem Cell Biol. **2004**, 36: 535-44.

[79] Wu X, Liu H, Liu H, Su J, Lv J, Cui Y, Wang F, Zhang Y. *Z*-curve theory-based analysis of the dynamic nature of nucleosome positioning in Saccharomyces cerevisiae. *Gene*., **2013**, *530*, 8-18.

[80] Dunn, J.G.; Foo, C.K.; Belletier, N.G.; Gavis, E.R.; Weissman, J.S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *Elife*., **2013**, *2*, e01179.