FEBS Letters 540 (2003) 188–194 FEBS 27122

Analysis of nucleotide distribution in the genome of *Streptomyces coelicolor* A3(2) using the Z curve method

Hong-Yu Ou, Feng-Biao Guo, Chun-Ting Zhang*

Department of Physics, Tianjin University, Tianjin 300072, PR China

Received 27 November 2002; revised 3 March 2003; accepted 4 March 2003

First published online 18 March 2003

Edited by Takashi Gojobori

Abstract **The nucleotide distribution of all 33 527 open reading frames (ORFs) ($\geq$ 300 bp) in the genome of *Streptomyces coelicolor* A3(2) has been analyzed using the Z curve method. Each ORF is mapped onto a point in a 9-dimensional space. To visualize the distribution of mapping points, the points are projected onto the principal plane based on principal component analysis. Consequently, the distribution pattern of the 33 527 points in the principal plane shows a flower-like shape, in which there are seven distinct regions. In addition to the central region, there are six petal-like regions around the center, one of which corresponds to 7172 coding sequences. The central region and the remaining five petal-like regions correspond to the intergenic sequences and out-of-frame non-coding ORFs, respectively. It is shown that selective pressure produces a remarkable bias of the G+C content among three codon positions, resulting in the interesting phenomenon observed. A similar phenomenon is also observed for other bacterial genomes with high genomic G+C content, such as *Pseudomonas aeruginosa* PA01 (G+C = 66.6%). However, for the genomes of *Bacillus subtilis* (G+C = 43.5%) and *Clostridium perfringens* (G+C = 28.6%), no similar phenomenon was observed. The finding presented here may be useful to improve the gene-finding algorithms for genomes with high G+C content. A set of supplementary materials including the plots displaying the base distribution patterns of ORFs in 12 prokaryotes is provided on the website http://tubic.tju.edu.cn/highGC/.**

Key words: Genome; G+C content; Seven clusters; Z curve; *Streptomyces coelicolor*

1. Introduction

By October 2002, more than 90 complete microbial genomes were available in the GenBank/EMBL/DDJB databases and many more sequencing projects are under way. The availability of these genomic sequences offers an unprecedented chance to systematically study biological functions, organizations and evolution of the genomes. Coding regions of DNA sequences are not random chains, and they are characterized by patterns of specific codons. Shepherd [1] proposed that the preference codon is of the RNY type, where R, Y and N represent purine, pyrimidine and any base, respectively. This non-random usage of codons can be used to find coding sequences (CDSs) [2]. The codon usage of more than 10 000 genes was analyzed in 1990 by Ikemura and coworkers [3], and the CUTG database which contains the updated codon usage table for each organism has been available since 1996 [4], using the DNA sequences obtained from the latest major releases of GenBank. It has also been found that the codon choice pattern contains more information than is necessary for encoding proteins. The difference in codon usage may be related to gene expression level [5] and some horizontally transferred genes [6,7]. Therefore, the three positions of codons are associated with different biological functions and the base choices at different positions are usually specific [8–10]. The whole genomic G+C content in bacterial genomes ranges from around 25% to around 75% [11,12]. The pioneering work of Bibb and co-workers revealed that the G+C content at different codon positions is conspicuously different in dozens of *Streptomyces* genes [13–15]. This finding is useful for predicting CDSs in genomes with a high G+C content [16].

Recently, the whole genome of *Streptomyces coelicolor* A3(2) has been sequenced (www.sanger.uk/Projects/S_coelicolor). As the model representative of a group of soil-dwelling organisms with a complex lifecycle involving mycelial growth and spore formation, *S. coelicolor* A3(2) has a large (8.7 Mb) genome with high (72.1%) G+C content [17]. This paper is devoted to studying the nucleotide distribution of all possible open reading frames (ORFs) ($\geq$ 300 bp) in the genome of *S. coelicolor*. Based on the Z curve method [18], the occurrence frequencies of nucleotides at three codon positions of an ORF are mapped onto a point in a 9-dimensional (9-D) space. In order to visualize the distribution pattern of the mapping points for all possible ORFs ($\geq$ 300 bp), all mapping points are projected onto a 2-dimensional (2-D) plane spanned by the first two principal axes of the principal component analysis (PCA). Interestingly, the projected points are clustered into seven distinct regions. Since the first and second principal components account for 87% of the total inertia of the nine components, it is concluded that the phenomenon of seven clusters occurs in the 9-D space, too. For comparison, similar analyses were performed for the genomes of *Bacillus subtilis* and *Clostridium perfringens* strain 13, with G+C contents of 43.5% and 28.6%, respectively. No similar phenomenon of seven clusters was observed. Further analysis is presented in this paper to explain the origin of this interesting phenomenon. It is shown that the phenomenon of seven clusters is related to the specific organization of genomes with high G+C content.

*Corresponding author. Fax: (86)-22-27402697.
E-mail address: ctzhang@tju.edu.cn (C.-T. Zhang).

0014-5793/03/$22.00 © 2003 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00263-1

2. Materials and methods

2.1. Data sets

The genome DNA sequence and the annotation information of *S. coelicolor* A3(2) [17] were downloaded from GenBank release 131.0 (accession number AL645882). The complete genome contains 7512 predicted CDSs, of which 7172 are longer than 300 bp. In the present study, all the possible ORFs ($\geq$300 bp) in each of the six frames of the double-stranded DNA are extracted. Here, an ORF is defined as a fragment of DNA sequence beginning with one of the initiation codons ATG, CTG, GTG and TTG and ending with one of the in-frame termination codons. Consequently, 33 527 ORFs ($\geq$300 bp) are found, of which 7172 ORFs are CDSs, whereas the remaining 26 355 ORFs are non-coding. The set of 33 527 ORFs ($\geq$300 bp) is called Set 1 hereafter.

In addition to Set 1, in order to study the nucleotide distribution of these ORFs, a theoretical data set needs to be constructed, which contains (i) the DNA sequences in the intergenic regions and (ii) a part of ORFs in the six reading frames of CDSs. For (i), 920 intergenic sequences ($\geq$300 bp) are extracted from the potentially untranslated regions of the *S. coelicolor* linear chromosome, based on the locations of the 7512 annotated CDSs. Note that the 'codon' in an intergenic sequence is meaningless. To calculate the base composition of an intergenic sequence, for example, GAGTGCACT..., then G, T, A... are defined as bases at the first 'codon' position and so forth. For (ii), we seek the ORFs ($\geq$300 bp) in all the six possible reading frames associated with each of the 7172 CDSs ($\geq$300 bp). It is well known that for a DNA sequence, there are three forward frames, i.e. Forward 0, 1 and 2, and three reverse frames, i.e. Reverse 0, 1 and 2, respectively. The six reading frames lead to six possible protein coding sequences, of which usually only one is likely to encode a protein. For simplicity, all 7172 CDSs are assumed to be translated in the frame of Forward 0. For each of the 7172 CDSs, look for at most one ORF $\geq$300 bp within the coding region of the CDS being studied, if there is any such ORF, in each of the six frames except the frame of Forward 0. For example, suppose that one of the 7172 CDSs begins with the codon ATG, and then look for the possible ORF $\geq$300 bp starting from the second base T. If such an ORF is found, it is assigned to the group of the frame of Forward 1. In this way, we obtained a total of 13 281 ORFs, of which 7172, 1243, 1215, 638, 2455 and 558 ORFs are in the frames of Forward 0, 1, 2 and Reverse 0, 1, 2, respectively. Finally, we construct a second data set, called Set 2, consisting of 920 intergenic sequences and 13 281 ORFs obtained in the above way.

2.2. Data analysis methods

In this study, the method of the Z curve [18] is used to analyze the base composition at three codon positions. Two other data analysis techniques, PCA and fuzzy c-means (FCM) clustering analysis, are used, too. The graphic method based on the Z curve has been used to analyze the codon usage of CDSs in several genomes, including human [19] and *Escherichia coli* [8]. Suppose that the occurrence frequencies of the bases A, C, G and T at the first, second and third codon positions in an ORF are denoted by a_i, c_i, g_i and t_i, respectively, where $i=1, 2, 3$. The four numbers, a_i, c_i, g_i and t_i, are mapped onto a point in a 3-dimensional (3-D) space V_i with the coordinates

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

Then, each ORF may be represented by a point or a vector in a 9-D space V, where $V = V_1 \oplus V_2 \oplus V_3$, here the symbol $\oplus$ denotes the direct sum of two subspaces. The nine components u_1–u_9 of the space V are defined as follows

$$\begin{cases} u_1 = x_1 - \langle x\rangle,\ u_2 = y_1 - \langle y\rangle,\ u_3 = z_1 - \langle z\rangle, \\ u_4 = x_2 - \langle x\rangle,\ u_5 = y_2 - \langle y\rangle,\ u_6 = z_2 - \langle z\rangle, \\ u_7 = x_3 - \langle x\rangle,\ u_8 = y_3 - \langle y\rangle,\ u_9 = z_3 - \langle z\rangle, \end{cases} \quad (2)$$

where $\langle x\rangle = (x_1+x_2+x_3)/3$, $\langle y\rangle = (y_1+y_2+y_3)/3$ and $\langle z\rangle = (z_1+z_2+z_3)/3$. The components u_1, u_4, u_7; u_2, u_5, u_8; and u_3, u_6, u_9 display the distribution of bases of the purine/pyrimidine (A or G/C or T) type; the amino/keto (A or C/G or T) type and the weak H-bond/strong H-bond (A or T/G or C) type at the first, second and third codon positions, respectively.

PCA defines a rotation of the variables of a given data set. A new set of variables is derived from the linear combination of the original variables. The first principal axis is chosen to maximize the standard deviation of the derived variable and the second principal axis is to maximize the standard deviation among directions not correlated with the first, and so forth. For details about this method, refer to [20].

To present the FCM clustering algorithm [21], we suppose that the mapping points are clustered into g groups in the 9-D space V. The extent of a point belonging to a group is determined by the membership degree. Denote the membership degree of a point belonging to the kth group by m_k ($k=1,...,g$). Here, $m_k \in [0,1]$ and $\Sigma m_k = 1$, $k=1,...g$. To find the fuzzy clustering centroid for each of the g groups, the objective function J_q defined below is minimized

$$J_q = \sum_{k=1}^{g}\sum_{i=1}^{N} [m_k(i)]^q d^2\ (\mathbf{U}_i, \mathbf{W}_k), \quad (3)$$

where N is the number of points, $m_k(i)$ is the membership degree of the ith point belonging to the kth group, and $d^2(\mathbf{U}_i,\mathbf{W}_k)$ is the square of the Euclidean distance between the ith point (or ith vector $\mathbf{U}_i$) and the kth clustering centroid $\mathbf{W}_k$. The exponent q controls the degree of fuzziness, and is usually taken to be slightly greater than 1 ($q=1.3$ is adopted here). The fuzzy clustering is carried out iteratively by minimizing the objective function J_q according to the following algorithm (see, e.g. [22]).

1. Choose initial centroids $\mathbf{W}_k$, $k=1,...,g$, and compute the membership degrees for each group and for all N points by the following formula

$$m_k(i) = \frac{[1/d^2(\mathbf{U}_i, \mathbf{W}_k)]^{1/(q-1)}}{\sum_{k=1}^{g}[1/d^2(\mathbf{U}_i, \boldsymbol{W}_k)]^{1/(q-1)}},\ i = 1,\ 2, ..., N,\ k = 1, ..., g, \quad (4)$$

   and then evaluate the objective function $J_q^{(0)}$ by Eq. 3;
2. Compute the new centroids by

$$\mathbf{W}_k = \frac{\sum_{i=1}^{N}[m_k(i)]^q\ \mathbf{U}_i}{\sum_{i=1}^{N}[m_k(i)]^q},\ k = 1, ..., g. \quad (5)$$

3. Update the membership degrees by substituting Eq. 5 into Eq. 4 and update the value of the objective function by $J_q^{(1)}$. The iteration goes on until $|J_q^{(1)} - J_q^{(0)}| < \varepsilon$, where ε is a given small number.

3. Results and discussion

3.1. The phenomenon of seven clusters

For each ORF in Set 1 and Set 2, the nine variables u_1–u_9 were calculated, which correspond to a point in the 9-D space V. To allow a direct comparison between Set 1 and Set 2, PCA was performed for both data sets simultaneously. For Set 1, there are 33 527 ORFs, corresponding to 33 527 mapping points in the 9-D space V. To visualize the distribution of mapping points in the 9-D space, the mapping points were projected onto a 2-D plane spanned by the first and second principal axes, using the PCA method. The first and second principal components account for 52.4% and 34.6% of the total inertia of the 9-D space, respectively, and no other component accounts for more than 6%. Therefore, the distribution pattern of the points in the 2-D plane basically reflects that in the 9-D space. Fig. 1 shows the distribution pattern of the 33 527 points in the 2-D plane. Two remarkable features in Fig. 1 need to be emphasized. First, a phenomenon of seven clusters is observed. The points are clustered into seven distinct regions. The smallest group is located at the center and the six petal-like regions are distributed symmetrically around

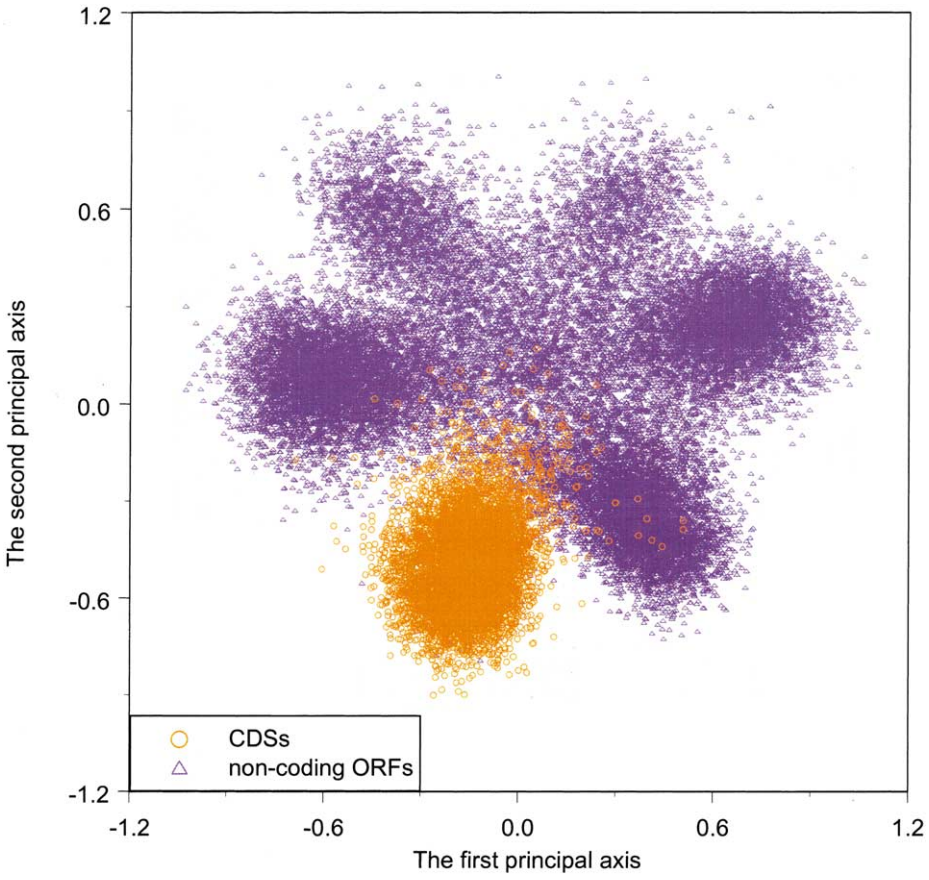

Fig. 1. Each of the 33 527 ORFs ($\geq$ 300 bp) in the genome of *S. coelicolor* is mapped onto a point in a 9-D space derived from the nine variables u_1–u_9 defined in Eq. 2. To visualize the distribution of the 33 527 mapping points in the 9-D space, the mapping points are projected onto the principal plane spanned by the first two principal axes based on the principal components analysis. The first and second principal components account for 52.4% and 34.6% of the total inertia of the 9-D space, respectively. Note that the pattern of distribution shows a flower-like shape, in which there are seven distinct regions. The petal-like region consisting of orange open circles corresponds to 7172 CDSs ($\geq$ 300 bp) in the organism, while the other five petal-like regions and the central region consisting of purple open triangles correspond to 26 355 non-coding ORFs.

the center. Second, the points corresponding to the 7172 CDSs are mainly situated in one of the petal-like regions, denoted by orange open circles. The six petal-like regions are located relatively far from the center. To confirm this observation, the FCM clustering algorithm was performed in the 9-D space. Here the number of groups $g = 6$ is studied, without considering the smallest central region. Consequently, in one of the six clusters found by the algorithm there are 7684 points, of which 7007 correspond to CDSs and 677 correspond to non-coding ORFs. Therefore, 7007/7172 = 97.7% of CDSs ($\geq$ 300 bp) are found by this method. It is also found that the six petal-like regions are located relatively far from the center of the 9-D space. The points in the central region and the remaining five petal-like regions correspond to non-coding ORFs. In the following we will give a convincing explanation about the two features of the distribution pattern.

3.2. The origin of the phenomenon of seven clusters

To understand the biological implication of the phenomenon of seven clusters, PCA was performed on Set 2. Note that Set 2 consists of seven kinds of DNA sequences, corresponding to the intergenic sequences and the ORFs in the frames of Forward 0, 1, 2 and Reverse 0, 1, 2, respectively. Fig. 2 shows the distribution pattern of a total of 13 281 points in the 2-D plane. Interestingly, the points corresponding to the intergenic sequences are situated mainly at the central region, whereas the points associated with one of the six petal-like regions, consisting of orange open circles, correspond mainly to the 7172 CDSs, and the remaining five petal-like regions correspond to the ORFs in the frames of Forward 1, 2 and Reverse 0, 1, 2, respectively. When Fig. 1 and Fig. 2 are compared, the origin of appearance of seven clusters in Fig. 1 is self-explanatory. However, the fact that the six petal-like regions are situated relatively far from the center needs to be explained. This is discussed in the next two sections.

3.3. The relationship between the G+C content at three codon positions and the genomic G+C content

The fact that the petal-like region containing most genes is located far from the center indicates that the base distribution in coding regions is significantly biased. To analyze this problem, consider the contributions of each of the variables u_1–u_9 in the PCA. The cumulative proportion of the first two principal components is 87.0% of the total inertia of the 9-D space V. For the first principal component, the two most important original variables are u_3 and u_9, which contribute 44.8% and 32.9% of the total inertia, respectively. Similarly, for the second principal component, the two most important original

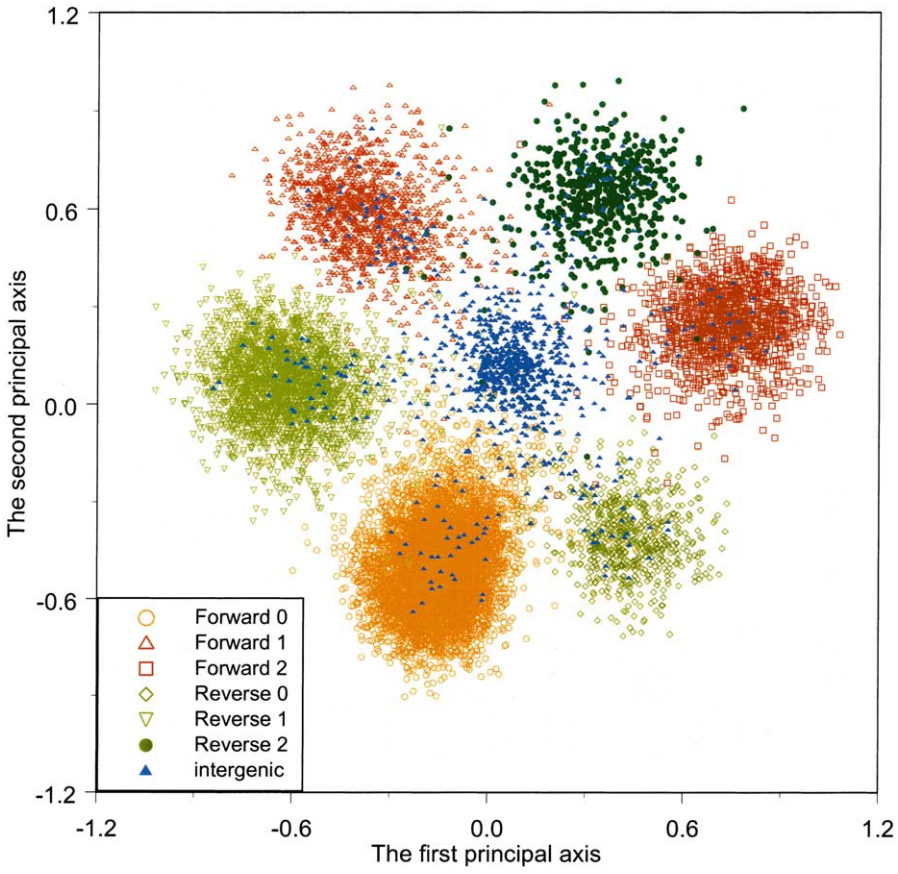

Fig. 2. PCA was performed on the 13 281 mapping points in Set 2, consisting of 920 intergenic sequences, and 7172, 1243, 1215, 638, 2455 and 558 ORFs ($\geq$ 300 bp), corresponding to the reading frames of Forward 0, 1, 2 and Reverse 0, 1, 2, respectively. The plot is similar to Fig. 1. Note that the mapping points for the intergenic sequences are situated around the center, whereas those corresponding to the 7172, 1243, 1215, 638, 2455 and 558 ORFs are situated at the six petal-like regions, respectively. The mapping points corresponding to 7172 CDSs are situated at one petal-like region consisting of orange open circles, whereas those corresponding to all out-of-frame non-coding ORFs are situated basically at the remaining five petal-like regions. When this figure is compared with Fig. 1, the origin of appearance of the phenomenon of seven clusters is self-explanatory.

variables are u_6 and u_9, which contribute 46.4% and 19.3% of the total inertia, respectively. Thus, the three variables u_3, u_6 and u_9 contribute more significantly than other original variables for the formation of seven clusters in the 9-D space. According to Eq. 1, the variables u_3, u_6 and u_9 are related to the G+C content at the first, second and third codon positions, respectively. For all 7512 annotated CDSs of *S. coelicolor*, the mean values of u_3, u_6 and u_9 averaged over all annotated genes are -0.01, 0.411 and -0.40, respectively. Using Eqs. 1 and 2, the mean G+C content at the first, second and third codon positions is 72.6%, 51.5% and 92.2%, respectively. It is notable that these figures are essentially unchanged compared with the previous analysis [15], in which only 64 genes from the genus *Streptomyces* were involved.

The above analysis suggests that the strongly biased G+C content at three codon positions of *S. coelicolor* accounts for the observed bias of the petal-like region from the center. To clarify the phenomenon, it is necessary to study the relationship between the mean G+C content at three codon positions and the genomic G+C content. The genome sequences and the related annotation files of 33 sequenced bacterial and archaeal genomes were downloaded from GenBank release 131.0. The names of these organisms are listed in the figure legend of Fig. 4. For each genome, the mean values of u_3, u_6 and u_9 averaged over all the annotated CDSs, denoted by $\overline{u}_3$, $\overline{u}_6$ and $\overline{u}_9$, respectively, were calculated. It has been observed that for a wide range of species, the genomic G+C content has a positive linear correlation with the G+C content at the three codon positions, although the slopes differ – the steepness rank order being the third, first and second codon positions, respectively [12]. Here, we further analyze the correlation of $\overline{u}_3$, $\overline{u}_6$ and $\overline{u}_9$ with the genomic G+C content using the data from 33 organisms. In Fig. 3, the y-axis denotes the values of $\overline{u}_3$, $\overline{u}_6$ and $\overline{u}_9$ for each genome, whereas the x-axis denotes the overall G+C content of the genome concerned. The data points for three codon positions can be fitted with three straight lines using the least square method, with the correlation coefficients $R = 0.899$, 0.979 and -0.968, respectively.

Using Eqs. 1 and 2, we have

$$\overline{u}_3 = \overline{z}_1 - \langle z \rangle = 2\langle G+C \rangle - 2\overline{(G+C)_1}, \quad (6)$$

and so forth for $\overline{u}_6$ and $\overline{u}_9$, where $\overline{(G+C)_1}$ is the mean G+C content at the first codon position averaged over all the annotated CDSs, and $\langle G+C \rangle$ is the mean value of the mean G+C content averaged over three codon positions, i.e. $\langle G+C \rangle = (\overline{(G+C)_1} + \overline{(G+C)_2} + \overline{(G+C)_3})/3$. $\overline{u}_6$ and $\overline{u}_9$ are defined similarly. It is seen from Fig. 3 that when the genomic G+C content increases from 25.5% (*Ureaplasma urealyticum*)

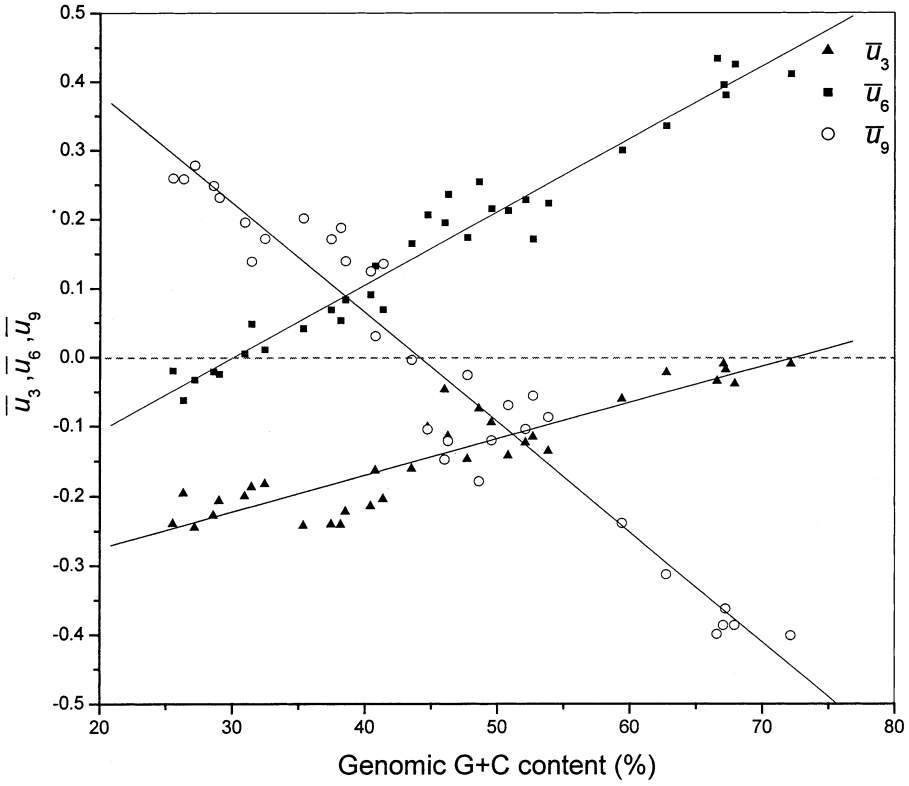

Fig. 3. Correlation of the genomic G+C content and the mean values of u_3, u_6 and u_9, denoted by $\overline{u}_3$, $\overline{u}_6$ and $\overline{u}_9$, respectively, which are averaged over all the annotated CDSs for each of the 33 prokaryotic genomes taken from GenBank release 131.0. The names of these organisms are listed in the legend of Fig. 4. The correlation of each of $\overline{u}_3$, $\overline{u}_6$ and $\overline{u}_9$ with the genomic G+C content is well fitted by a straight line, with correlation coefficients $R = 0.90$, 0.98 and -0.97, respectively.

to 72.1% (*S. coelicolor*), the mean G+C content at the first codon position, i.e. $\overline{(G+C)_1}$, increases gradually. However, for almost all genomes, $\overline{(G+C)_1} > \langle G+C \rangle$, as reflected by the fact that almost all the triangle points are situated below the horizontal dashed line in Fig. 3. The quantity $\langle G+C \rangle$ represents the average G+C content of the whole coding sequences. Since most parts of bacterial and archaeal genomes are used to encode proteins, the G+C content of the whole coding sequences is roughly equal to the genomic G+C content. Therefore, this fact suggests that for almost all 33 genomes studied here the mean G+C content at the first codon position is greater than the genomic G+C content. Similarly, the mean G+C content at the second codon position, i.e. $\overline{(G+C)_2}$, also increases gradually, but for most genomes $\overline{(G+C)_2}$ is smaller than the genomic G+C content. This is reflected by the fact that almost all square points except five are situated above the horizontal dashed line. The five exceptions belong to those genomes in which the genomic G+C content is less than 30%, which may be considered a critical point for the G+C content at the second codon position. Finally, the mean G+C content at the third codon position, i.e. $\overline{(G+C)_3}$, increases quickly with the increase of the genomic G+C content. It is seen from Fig. 3 that there is another critical point at the x-axis where the genomic G+C content is equal to about 44%. For the genome with its genomic G+C content smaller (greater) than 44%, the mean G+C content at the third codon position, i.e. $\overline{(G+C)_3}$, is smaller (higher) than the genomic G+C content.

To investigate the relationship between the mean G+C content at the three codon positions and the genomic G+C content further, we define the standard deviation s

$$s^2 = \frac{1}{2} \sum_{n=3,6,9} (\overline{u}_n - (\overline{u}_3 + \overline{u}_6 + \overline{u}_9)/3)^2. \quad (7)$$

Fig. 4 shows the distribution of s as a function of the genomic G+C content for the 33 bacterial and archaeal genomes. For the genomes with lower genomic G+C content close to 25%, s is equal to about 0.25. When the genomic G+C content ranges from 40% to 50%, the value of s ranges from about 0.2 to 0.15. As the genomic G+C content increases from about 55% to 72%, the value of s increases quickly. In the case of *S. coelicolor* with a genomic G+C content of 72.1%, s is equal to 0.41, indicating the highly biased G+C content at three codon positions. Now it is clear that the highly biased G+C content at three codon positions is the key reason resulting in the phenomenon of seven clusters. The fact that the petal-like region associated with genes in Figs. 1 and 2 is located far from the center is due to the same reason. Fitting the points in Fig. 4 by a parabola, it is found that the parabola has a global minimum point at about G+C = 42%. The genomic G+C content of *B. subtilis* is 43.5% ($s = 0.16$), which is close to the minimum point. In an analysis similar to that for the genome of *S. coelicolor*, we examined the base distribution pattern of ORFs in the genome of *B. subtilis*. All 11 228 ORFs (≥ 210 bp) were extracted from the complete genome. The corresponding mapping points are projected onto the 3-D PCA space spanned by the first three principal axes, of which the cumulative percentage of the total

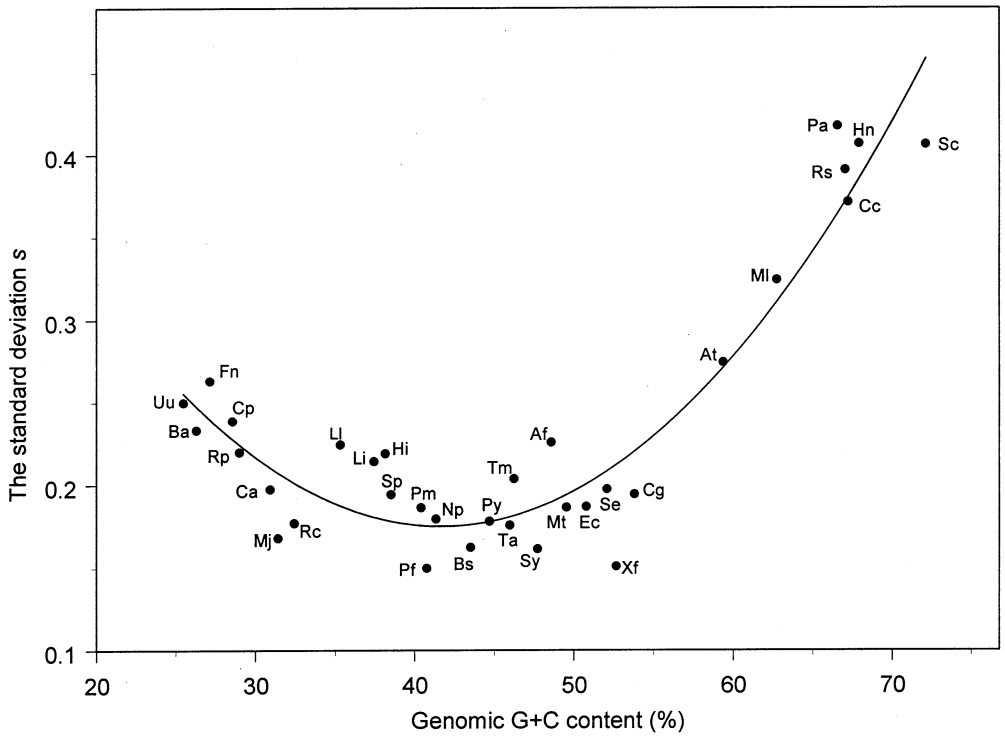

Fig. 4. Relation between the genomic G+C content and the standard deviation of $\overline{u}_3$, $\overline{u}_6$ and $\overline{u}_9$, denoted by s (see Eq. 7). There are 33 data points corresponding to the 33 bacterial and archaeal genomes studied here. Fitting the points by a parabola with the polynomial regression, we find that the regression coefficient $R^2 = 0.89$. Note that there is a global minimum point situated at about 42% (genomic G+C content). See the text for more detailed explanation about the meaning of the plot. Non-standard abbreviations of the names of the 33 sequenced prokaryotes under analysis are: Uu, *Ureaplasma urealyticum*; Ba, *Buchnera* sp. APS; Fn, *Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586; Cp, *Clostridium perfringens* strain 13; Rp, *Rickettsia prowazekii* strain Madrid E; Ca, *Clostridium acetobutylicum* ATCC824; Mj, **Methanococcus jannaschii*; Rc, *Rickettsia conorii*; Ll, *Lactococcus lactis* subsp. *lactis* IL1403; Li, *Listeria innocua* Clip11262; Hi, *Haemophilus influenzae* Rd; Sp, *Streptococcus pyogenes* strain SF370 serotype M1; Pm, *Pasteurella multocida* PM70; Pf, **Pyrococcus furiosus* DSM 3638; Np, *Nostoc* sp. PCC 7120; Bs, *Bacillus subtilis*; Py, **Pyrococcus abyssi*; Ta, **Thermoplasma acidophilum*; Tm, *Thermotoga maritima*; Sy, *Synechocystis* PCC6803; Af, **Archaeoglobus fulgidus*; Mt, **Methanobacterium thermoautotrophicum* delta H; Ec, *Escherichia coli* K12; Se, *Salmonella enterica* serovar Typhi strain CT18; Xf, *Xylella fastidiosa*; Cg, *Corynebacterium glutamicum*; At, *Agrobacterium tumefaciens* strain C58; Ml, *Mesorhizobium loti*; Pa, *Pseudomonas aeruginosa* PA01; Rs, *Ralstonia solanacearum* GMI1000; Cc, *Caulobacter crescentus*; Hn, **Halobacterium* sp. NRC_1; Sc, *Streptomyces coelicolor* A3(2). The asterisk denotes an archaeon.

inertia is 74.8%. The mapping points cannot be grouped into several recognizable clusters (see the supplementary material). Furthermore, the points corresponding to the 3946 CDSs are located near the distribution center, therefore, most CDSs cannot be separated from the non-coding ORFs based on the point distribution. In addition, the base composition bias of ORFs in *C. perfringens* strain 13 (genomic G+C content = 28.6%; $s = 0.24$) is also analyzed. All the mapping points corresponding to the 6078 ORFs (≥ 150 bp) are gathered into three clusters in the 3-D PCA space spanned by the first three principal axes (see the supplementary material). The cumulative percentage of the total inertia for the first three principal components is 77.3%. However, in one of the three clusters, the points corresponding to more than half of the 2657 CDSs overlap with those corresponding to non-coding ORFs, which are in the reading frame of Reverse 1. For genomes with higher values of s, a phenomenon similar to that of *S. coelicolor* has been observed. For example, we have observed a similar base distribution pattern in the 9-D space V, i.e. the occurrence of seven clusters for the ORFs (≥ 300 bp) in the genome of *Pseudomonas aeruginosa* PA01 (genomic G+C content = 66.6%; $s = 0.41$). For more details, refer to the supplementary material. In summary, the reason for the phenomenon of seven clusters and the phenomenon that the petal-like regions are located far from the center is the higher genomic G+C content with higher values of s. The parameter s introduced here appears to be useful to analyze the bias of the G+C content among three codon positions.

3.4. The relationship between selective pressure and genomic G+C content

A great amount of studies have shown that coding sequences are not random, instead, specific patterns are hidden within them. It has been well known for a long time that purine bases are predominant at the first codon position [18], bases at the second codon position are relatively short of G [23], while bases at the third codon position are species-dependent [8]. This pattern is valid not only for prokaryotic but also for eukaryotic genomes. This pattern is also valid regardless of the genomic G+C content. The bacterium *S. coelicolor* is an organism with very high genomic G+C content (72.1%). However, the codon usage of CDSs in this genome should also obey the pattern mentioned above. The demand that there are not many G bases at the second codon position leads to the constraint that the mean G+C content at the second position should be lower. To keep an appropriate ratio between the numbers of A and G bases at the first codon position, there cannot be too many G bases there. Consequently, selective

pressure makes the G+C content at the third codon position much higher. As mentioned previously, the mean G+C content at the first, second and third codon positions for the genome of *S. coelicolor* is 72.6%, 51.5% and 92.2%, respectively. These figures confirm the above analysis. Since the genomic G+C content of *S. coelicolor* is as high as 72.1%, the strong selective pressure exerts a great influence on the codon usage of genes in this genome. In summary, the phenomenon observed in this paper is caused by the higher genomic G+C content of *S. coelicolor*. The genomic G+C content is so high that selective pressure forces a strong bias of the G+C content among three codon positions. This bias leads to the occurrence of seven clusters in the 9-D space spanned by u_1–u_9 derived from the Z curve method. The bias also makes the petal-like regions be situated far from the center. The existence of five petal-like regions in addition to the one corresponding to CDSs clearly shows that there are many non-coding ORFs entirely or partly overlapping with CDSs. A detailed calculation shows that only 26 ORFs ($\geq$300 bp) in the genome of *S. coelicolor* do not overlap with other ORFs out of all the 33 527 ORFs ($\geq$300 bp). This fact may cause difficulties for computer-aided gene-finding algorithms, due to the potential high rate of false positive prediction.

4. Conclusion

In this paper, it is shown that selective pressure causes a strong bias of the G+C content among three codon positions, whereas the bias accounts for the formation of seven clusters for the ORFs ($\geq$300 bp) in the genome of *S. coelicolor* in a 9-D space. Most CDSs are gathered into one of the six petal-like regions around the distribution center. Using the FCM clustering algorithm, 97.7% of the 7172 CDSs can be separated out of all the 33 527 ORFs ($\geq$300 bp). Based on a similar analysis, it is found that for the genomes of *B. subtilis* and *C. perfringens*, the distinct seven clusters do not occur, due to the fact that variations of base distributions at three positions of codon are relatively weak. It is also found that the phenomenon of seven clusters appears for all other bacterial and archaeal genomes with a high genomic G+C content. For example, mapping points for ORFs in the genome of *P. aeruginosa* show a similar pattern in the 9-D space. It is hoped that this work could be useful to improve the computer-aided gene-finding algorithms for genomes with a high genomic G+C content.

5. Supplementary materials

A list of the 2-D and 3-D rotational plots for visualizing the base distribution patterns of ORFs in the 2-D PCA plane and 3-D PCA space is provided on the website http://tubic.tju.edu.cn/highGC/. Relevant plots for 12 bacterial or archaeal genomes with different genomic G+C content are also provided and analyzed therein.

Acknowledgements: We thank Ren Zhang, Ling-Ling Chen and Sheng-Yun Wen for invaluable assistance. The present study was supported in part by the 973 Project of China (Grant G1999075606).

References

[1] Shepherd, J.C. (1981) Proc. Natl. Acad. Sci. USA 78, 1596–1600.
[2] Fickett, J.W. and Tung, C.S. (1992) Nucleic Acids Res. 20, 6441–6450.
[3] Wada, K., Aota, S., Tsuchiya, R., Ishibashi, F., Gojobori, T. and Ikemura, T. (1990) Nucleic Acids Res. 18 (Suppl.), 2367–2411.
[4] Nakamura, Y., Wada, K., Wada, Y., Doi, H., Kanaya, S., Gojobori, T. and Ikemura, T. (1996) Nucleic Acids Res. 24, 214–215.
[5] Ikemura, T. (1981) J. Mol. Biol. 151, 389–409.
[6] Medigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991) J. Mol. Biol. 222, 851–856.
[7] Karlin, S. (2001) Trends Microbiol. 9, 335–343.
[8] Zhang, C.T. and Chou, K.C. (1994) J. Mol. Biol. 238, 1–8.
[9] Mrazek, J. and Karlin, S. (1998) Proc. Natl. Acad. Sci. USA 95, 3720–3725.
[10] Li, W. (1999) Comput. Chem. 23, 283–301.
[11] Sueoka, N. (1962) Proc. Natl. Acad. Sci. USA 48, 582–591.
[12] Muto, A. and Osawa, S. (1987) Proc. Natl. Acad. Sci. USA 84, 166–169.
[13] Bibb, M.J., Findlay, P.R. and Johnson, M.W. (1984) Gene 30, 157–166.
[14] Bibb, M.J., Ward, J.M. and Cohen, S.N. (1985) Mol. Gen. Genet. 199, 26–36.
[15] Wright, F. and Bibb, M.J. (1992) Gene 113, 55–65.
[16] Ishikawa, J. and Hotta, K. (1999) FEMS Microbiol. Lett. 174, 251–253.
[17] Bentley, S.D. et al. (2002) Nature 417, 141–147.
[18] Zhang, C.T. and Zhang, R. (1991) Nucleic Acids Res. 19, 6313–6317.
[19] Zhang, C.T. and Chou, K.C. (1993) J. Protein Chem. 12, 329–335.
[20] Dillon, W.R. and Goldstein, M. (1984) Multivariate Analysis, Methods and Applications, Wiley, New York.
[21] Bezdek, J.C. (1981) Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York.
[22] Zhang, C.T., Chou, K.C. and Maggiora, G.M. (1995) Protein Eng. 8, 425–435.
[23] Trifonov, E.N. (1987) J. Mol. Biol. 194, 643–652.