Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 306 (2003) 310–317
BBRC
www.elsevier.com/locate/ybbrc

Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages

Ling-Ling Chen[a,b] and Chun-Ting Zhang[a,*]

[a] *Department of Physics, Tianjin University, Tianjin 300072, China*
[b] *Department of Biology, Shandong University of Technology, Zibo 255049, China*

Received 8 May 2003

Abstract

Seven GC-rich (group I) and three AT-rich (group II) microbial genomes are analyzed in this paper. The seven microbes in group I belong to different phylogenetic lineages, even different domains of life. The common feature is that they are highly GC-rich organisms, with more than 60% genomic GC content. Group II includes three bacteria, which belong to the same subdivision as *Pseudomonas aeruginosa* in group I. The genomic GC content of the three bacteria is in the range of 26–50%. It is shown that although the phylogenetic lineages of the organisms in group I are remote, the common feature of highly genomic GC content forces them to adopt similar codon usage patterns, which constitutes the basis of an algorithm using a set of universal parameters to recognize known genes in the seven genomes. The common codon usage pattern of function known genes in the seven genomes is G$\bar{G}$S type, where G, $\bar{G}$, and S are the bases of G, non-G, and G/C, respectively. On the contrary, although the phylogenetic lineages of the three bacteria in group II are quite close, the codon usage patterns of function known genes in these genomes are obviously distinct. There are no universal parameters to identify known genes in the three genomes in group II. It can be deduced that the genomic GC content is more important than phylogenetic lineage in gene recognition programs. We hope that the work might be useful for understanding the common characteristics in the organization of microbial genomes.

Keywords: Codon usage; GC content; Phylogenetic lineage; Microbial genomes

Since the first bacterial genome *Haemophilus influenzae* was sequenced in 1995 [1], many microbial sequencing projects have been completed. The availability of the sequences and annotations of many prokaryotic species offers an extraordinary opportunity for analyzing the relationships among these genomes in a comprehensive and precise fashion. Protein coding regions are not random sequences and they are characterized by patterns of specific codons. In general, the composition of sequences is analyzed based on the GC content. The genomic GC content in prokaryotes ranges from about 25% to 75% [2], much wider than that in eukaryotes. Since 1980s, many scientists have studied the codon usage patterns of several organisms [3–12] and the generally accepted pattern is R$\bar{G}$N type, where R, $\bar{G}$, and N indicate the bases of purine, non-G, and any base at the first, second, and third codon positions, respectively. The severe restrictions on the base frequencies at the first two codon positions are universal in protein coding sequences and independent of species [8]. In contrast, the distribution of bases at the third codon position is species-dependent, and thus it might be possible to distinguish one species from another by using such distributions [8]. It has been suggested that the first, second, and third codon positions are associated with the biosynthetic pathway, hydrophobicity pattern, and the α-helix or β-strand forming potentiality of the coded amino acid, respectively [13,14].

In this paper, the codon usage patterns of function known genes in 10 microbial genomes are analyzed using the GC content at three codon positions. The 10 genomes can be divided into two groups. Group I contains seven microbes, which belong to different domains of life. Among them, six are bacteria and one is an archaeon, and the six bacteria belong to different lineages.

* Corresponding author. Fax: +86-22-2740-2697.
E-mail address: ctzhang@tju.edu.cn (C.-T. Zhang).

0006-291X/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0006-291X(03)00973-2

The common feature is that they are highly GC-rich organisms, with more than 60% genomic GC content. Group II includes three bacteria with close lineages, which belong to the same subdivision as *P. aeruginosa* in group I. Although the phylogenetic lineages of the seven genomes in group I are remote, the common feature of highly genomic GC content forces them to adopt similar codon usage patterns. Using the three parameters and Fisher discriminant method, function known genes in the seven genomes can be identified with high accuracy. Most importantly, the Fisher coefficients are universal in the seven genomes. Each set of Fisher coefficients derived from any individual genome can be used to identify most of known genes in the seven genomes with high accuracy. The similar distribution patterns of bases and the significant difference in the GC content at three codon positions form the basis of the high recognition accuracy of the current method. Though the lineages of the bacteria in group II are quite close, function known genes in these genomes adopt different codon usage patterns because of their diverse genomic GC content. In group II, the Fisher coefficients obtained from one genome cannot identify known genes in other genomes with high accuracy. This work might be helpful for the current gene recognition programs, which are based on the statistical properties of training samples. The predominant models for microbial sequence analysis are Markov models [15,16] and Z curve methods [17–19]. Markov models use thousands of probabilities and these parameters are species-specific [15]. Currently, there are few algorithms which are species-independent, i.e., the parameters used are universal for different species. In this paper, three universal parameters for identifying function known genes in the seven GC-rich genomes in group I are presented and the recognition accuracy is satisfactory.

Materials and methods

The 10 microbial genomes analyzed are *Caulobacter crescentus* CB15 (GenBank Accession No. AE005673), *Deinococcus radiodurans* (chromosomes 1) (AE000513), *Halobacterium* sp. NRC-1 (AE004437), *Mesorhizobium loti* (BA000012), *Mycobacterium tuberculosis* H37Rv (AL123456), *Sinorhizobium meliloti* 1021 (AL591688), *Pseudomonas aeruginosa* PA01 (AE004091), *Buchnera* sp. APS (BA000003), *Haemophilus influenzae* Rd (L42023), and *Vibrio cholerae* (chromosome 1) (AE003852). The data were downloaded from GenBank, release 132.0. Group I comprises the former seven microbes, which belong to different phylogenetic lineages. Among them, *Halobacterium* is an archaeon and the other six are bacteria. *M. loti* and *S. meliloti* belong to the same family and the other four belong to different divisions. The common feature is that their genomic GC content is greater than 60%. Group II contains the latter three microbes, which belong to proteobacteria, gamma subdivision. *P. aeruginosa* in group I also belongs to this subdivision. The genomic GC content of the three genomes in group II ranges from 26% to 50%.

The method used here is based on the GC content at three codon positions. Let GC_1, GC_2, and GC_3 denote the GC content at the first, second, and third codon positions, respectively. And the three-dimensional space V is spanned by u_1, u_2, and u_3, which are defined by

$$\begin{aligned} u_1 &= GC_1, \\ u_2 &= GC_2, \\ u_3 &= GC_3. \end{aligned} \tag{1}$$

Therefore, an ORF or a fragment of DNA sequence can be represented by a point or a vector in the three-dimensional space V. To complete the algorithm, two sets of samples are needed. One is a set of positive samples corresponding to the protein coding genes; the other is a set of negative samples corresponding to the non-coding sequences. The two sets of samples constitute the training set used in the Fisher discrimination algorithm. The Fisher linear discrimination equation in this case represents a plane in the three-dimensional space V, described by a vector $\mathbf{c}$, which has three components c_1, c_2, and c_3. For details to determine the vector $\mathbf{c}$, refer to [20]. The vector $\mathbf{c}$ is not unique in the sense that $\mathbf{c}$ multiplied by a constant is still acceptable. Without losing generality, the constant is chosen such that $|\mathbf{c}|^2 = 1$. Based on the data in the training set, an appropriate threshold c_0 is determined to make the coding/non-coding decision. The threshold c_0 is uniquely determined by letting the false negative rate equal to the false positive rate. Once the vector $\mathbf{c}$ and the threshold c_0 are obtained, the decision of coding/non-coding for each ORF in the test set is simply made by the criterion of $\mathbf{c}\cdot\mathbf{u} > c_0 / \mathbf{c}\cdot\mathbf{u} < c_0$, where $\mathbf{c} = (c_1, c_2, c_3)^{\mathrm{T}}$ and $\mathbf{u} = (u_1, u_2, u_3)^{\mathrm{T}}$, and 'T' indicates the transposition of a matrix. If the training set and test set are the same, the procedure is called self-evaluation test, otherwise it is called cross-validation test. In order to evaluate the gene identification algorithm, the sensitivity s_n and specificity s_p are used. They are defined as: $s_n = TP/(TP + FN)$, $s_p = TN/(TN + FP)$, where TP, TN, FP, and FN are fractions of positive correct, negative correct, false positive, and false negative predictions, respectively. In other words, s_n is the proportion of coding ORFs that have been correctly predicted as coding, and s_p is the proportion of non-coding sequences that have been correctly predicted as non-coding. The accuracy is defined as the average of s_n and s_p. The criterion of $\mathbf{c}\cdot\mathbf{u} > c_0 / \mathbf{c}\cdot\mathbf{u} < c_0$ for making the decision of coding/non-coding can be rewritten as $Z(\mathbf{u}) > 0 / Z(\mathbf{u}) < 0$, where $Z(\mathbf{u}) = \mathbf{c}\cdot\mathbf{u} - c_0$. $Z(\mathbf{u})$ is called the Z score for an ORF or a fragment of DNA sequence.

Results and discussion

Recognition accuracy of the algorithm and the generality of Fisher coefficients in group I

In the GenBank (Release 132.0), a total of 3737, 2579, 2058, 6752, 3918, 5565, 3341, 564, 1709, and 2736 ORFs for *C. crescentus*, *D. radiodurans* (chromosome 1), *Halobacterium* sp. NRC, *M. loti*, *M. tuberculosis*, *P. aeruginosa*, *S. meliloti*, *Buchnera*, *H. influenzae*, and *V. cholerae* (chromosome 1) are annotated, respectively. According to the annotation, they can be divided into two classes: the first class contains genes with known functions, whereas the second class includes 'putative,' 'possible,' 'similar,' 'probable,' 'conserved hypothetical,' 'hypothetical,' and 'unknown' ORFs. Whether these ORFs are coding sequences or not are not known yet. We only consider the genes in the first class, which correspond to proteins with known functions. Thus, 1953, 1017, 877, 2983, 1470, 1433, 187, 477, 1004, and 1563 known genes are selected from the above genomes, respectively. They constitute the positive samples of

each genome. Although the amount of intergenic sequences in prokaryotes is much fewer than those in eukaryotes, it is sufficient for this study because of the small number of parameters used. Except protein and RNA coding regions, the intergenic sequences with length greater than 150 bp are chosen from the above genomes, respectively. Thus, 865, 538, 447, 2052, 900, 1504, 1277, 157, 430, and 388 intergenic sequences are selected, respectively. They serve as the negative samples of each genome.

Using the samples in the positive and negative sets in each genome, the Fisher coefficients c_1, c_2, c_3, and threshold c_0 are determined. After this, the parameters u_1, u_2, and u_3 for each positive or negative sample are calculated and then the Z score for each sample is obtained. The accuracy of the algorithm is obtained using the criterion of $Z(\mathbf{u}) > 0/Z(\mathbf{u}) < 0$. This is a self-evaluation test. The sensitivity, specificity, and accuracy for each genome are calculated and listed in Table 1. It can be seen that the recognition accuracy is satisfactory using only the three parameters. Table 2 lists the Fisher coefficients c_1, c_2, c_3, and threshold c_0 for the two groups of genomes. For each coefficient, the arithmetic mean and standard deviation in group I are calculated and listed in Table 2. As can be seen, the corresponding coefficient for different species is similar and the standard deviation is less than 0.1 for the seven genomes in group I.

To test the universality of the Fisher coefficients in group I, the following experiment is conducted. Using each set of Fisher coefficients in Table 2 to recognize known genes in each of the other six genomes, the recognition accuracy is shown in Table 3. In this table, the genomes used to train the Fisher coefficients are shown in the first column, and the genomes used to evaluate the accuracy of the algorithm are listed in the first row. The elements in the leading diagonal with bold style are the self-evaluation accuracy of each genome. The other elements are the accuracy of using the Fisher coefficients obtained from one genome to recognize known genes in others. For example, the second element in the first row is the accuracy of using the Fisher coefficients derived from *C. crescentus* to identify known genes in *D. radiodurans* and the recognition accuracy is 96.59%. It can be seen that for each genome, the recognition accuracy using the coefficients from other genomes is almost as high as that using its own coefficients. Most of known genes in the seven genomes can be correctly identified using the Fisher coefficients derived from any of the seven genomes. All known genes and intergenic sequences in the seven genomes are combined, forming two larger sets of positive and negative samples. The Fisher coefficients and threshold for the larger sets are calculated and listed in the tenth row in Table 2. The coefficients are called 'total self-evaluation coefficients' and used to recognize the 9920 known genes ($9920 = 1953 + 1017 + 877 + 2983 + 1470 + 1433 + 187$) and 7583 intergenic sequences. Consequently, the accuracy is 96.64%. The 'total self-evaluation coefficients' are similar to the arithmetic mean ones averaged over the coefficients derived from the seven genomes, respectively. Using the arithmetic mean coefficients to identify all the 9920 known genes, the recognition accuracy is 96.60%, which is very close to the self-evaluation accuracy. Therefore, a set of universal Fisher coefficients is obtained, which can be used to recognize genes in the seven genomes with satisfactory accuracy.

The same procedure is performed for the three bacteria in group II. However, the results are quite diverse. The self-evaluation accuracy for each genome is satis-

Table 1
The sensitivity, specificity, and accuracy for recognizing function known genes in two groups of microbial genomes using the present algorithm

	Species[a]	GC[b] (%)	Number of function known genes[c]	Number of intergenic sequences[d]	Sensitivity (%)	Specificity (%)	Accuracy[e] (%)
I	*C. cre*	67.21	1953	865	98.26	98.27	98.26
	*D. rad*1	67.01	1017	538	97.84	97.77	97.80
	H. sp.NRC[f]	67.91	877	447	94.64	94.63	94.64
	M. loti	62.75	2983	2052	97.25	97.27	97.26
	M. tub	65.61	1470	900	94.76	94.67	94.72
	P. aer	66.56	1433	1504	97.56	97.61	97.58
	S. meli	62.73	187	1277	98.93	98.90	98.92
II	*Buchnera*	26.31	477	157	97.27	97.45	97.36
	H. inf	38.15	1004	430	94.92	94.65	94.78
	*V. cho*1	47.70	1563	388	94.37	94.07	94.22

[a] The arrangement of the microbes studied here is according to the alphabetical order of their names. To save printing space, the abbreviation names of the microbes are used. For example, *Caulobacter crescentus* is abbreviated as *C. cre*, and so forth.
[b] The genomic GC content.
[c] The number of function known genes in the first class.
[d] The number of intergenic sequences with length greater than 150 bp.
[e] Accuracy is defined as the average of the sensitivity and specificity.
[f] Archaeal genome.

Table 2
Fisher coefficients and threshold for each genome and the arithmetic means and standard deviations averaged over the seven GC-rich microbial genomes in group I

	Species	c_0	c_1	c_2	c_3
I	*C. cre*	−0.4262	0.2872	−0.5688	0.7707
	*D. rad*1	−0.3951	0.2282	−0.5675	0.7911
	H. sp.NRC[a]	−0.4405	0.1792	−0.5219	0.8340
	M. loti	−0.3744	0.3738	−0.6405	0.6708
	M. tub	−0.5003	0.3040	−0.5059	0.8073
	P. aer	−0.2161	0.1971	−0.7236	0.6614
	S. meli	−0.3206	0.1760	−0.6290	0.7572
	Mean[b]	**−0.3819**	**0.2494**	**−0.5939**	**0.7561**
	Deviation[c]	0.0921	0.0746	0.0758	0.0663
	Total-self[d]	**−0.3843**	**0.2642**	**−0.6433**	**0.7186**
II	*Buchnera*	−0.1466	0.7393	0.2189	−0.6368
	H. inf	−0.3318	0.9573	0.0036	−0.2890
	*V. cho*1	−0.2979	0.8711	−0.4771	0.1165

[a] Archaeal genome.
[b] The arithmetic mean of each Fisher coefficient in the same column, which is emphasized in bold style.
[c] The standard deviation of each Fisher coefficient in the same column.
[d] The 'total self-evaluation coefficients' obtained using all function known genes and intergenic sequences in the seven genomes listed in Table 1, which are emphasized in bold style.

Table 3
The recognition accuracy using the Fisher coefficients trained in one genome to recognize known genes in all the seven genomes in group I

	C. cre (%)	*D. rad*1 (%)	*H.* sp.NRC[a] (%)	*M. loti* (%)	*M. tub* (%)	*P. aer* (%)	*S. meli* (%)
C. cre	**98.26**[b]	96.59	93.45	96.70	94.31	96.93	99.15
*D. rad*1	97.72	**97.80**	93.50	96.96	94.56	96.84	98.92
H. sp.NRC[a]	98.08	96.63	**94.64**	95.91	93.84	96.92	99.19
M. loti	97.95	96.40	93.40	**97.26**	93.86	96.94	98.72
M. tub	98.04	96.59	93.56	96.00	**94.72**	96.82	99.26
P. aer	98.24	96.91	94.12	97.08	93.20	**97.58**	98.46
S. meli	98.22	96.72	94.01	96.58	93.67	97.26	**98.92**

[a] Archaeal genome.
[b] The percent in the leading diagonal with bold style is the self-evaluation accuracy for each genome.

factory. But using the coefficients to recognize known genes in other genomes, the average recognition accuracy is less than 60%. No universal parameters exist in the three genomes, although they are closely related in phylogenetic lineages. The mechanism of these phenomena will be analyzed in the following section.

The base distribution patterns at three codon positions of the 10 genomes

As viewed from genome evolution, the changes of nucleotide compositions are affected by selective and mutational mechanism [12]. Owing to selective constraints, nucleotide compositions in protein coding regions are relatively conservative under the evolutionary pressure, and the base contents at three codon positions are correlated. The generally accepted codon usage pattern is the $R\bar{G}N$ type. The severe restrictions on the base frequencies at the first two codon positions are universal in protein coding sequences and independent of species. This pattern reflects the formation of a stable and native folding structure of proteins [8]. Trifonov [4] proposed that the pattern is responsible for monitoring the correct reading frame during translation. In contrast, the distributions of bases at the third codon position are species-dependent [8]. Ikemura [3] pointed out that the base frequencies at the third codon position might reflect mutational biases or selection for favored codons, and it is actually relevant to the tRNA content in cells. The phenomenon of species-independent at the first two codon positions and species-specific at the third codon position implies that though the base usage at three codon positions reveals a selection-mutation balance, the influence of mutational pressure is much stronger at the third than at the first two codon positions, where some universal selection rules may dominate over the mutational bias [11].

The distributions of bases at three codon positions of function known genes in the 10 genomes are shown in Table 4. The base distribution patterns are similar in the seven genomes in group I. The 'GC disparity' column lists the values of differences between the genomic GC content and the GC content at each codon position, respectively. Positive value of 'GC disparity' means the

Table 4
The distributions of bases A, G, C, and T and the GC content, GC disparity at three codon positions of two groups of genomes

	Species		A	G	C	T	GC content	GC disparity[b]
I	*C. cre*	1st[a]	0.2018	0.4182	0.2622	0.1178	0.6804	0.0083
		2nd[a]	0.2500	0.1970	0.2712	0.2818	0.4682	−0.2039
		3rd[a]	0.0430	0.3833	0.4922	0.0815	0.8755	0.2034
	*D. rad*1	1st	0.2011	0.4038	0.2916	0.1035	0.6954	0.0253
		2nd	0.2569	0.2133	0.2502	0.2796	0.4635	−0.2066
		3rd	0.0659	0.3785	0.4799	0.0757	0.8584	0.1883
	H. sp.NRC[c]	1st	0.1863	0.4704	0.2288	0.1144	0.6992	0.0201
		2nd	0.2828	0.1836	0.2701	0.2635	0.4537	−0.2254
		3rd	0.0551	0.3677	0.5295	0.0477	0.8972	0.2181
	M. loti	1st	0.2236	0.3934	0.2500	0.1330	0.6434	0.0159
		2nd	0.2499	0.1928	0.2606	0.2968	0.4534	−0.1741
		3rd	0.0772	0.3529	0.4544	0.1155	0.8073	0.1798
	M. tub	1st	0.1949	0.4303	0.2445	0.1304	0.6748	0.0187
		2nd	0.2298	0.2113	0.2867	0.2723	0.4980	−0.1581
		3rd	0.0870	0.3731	0.4281	0.1118	0.8012	0.1451
	P. aer	1st	0.2080	0.3863	0.2852	0.1206	0.6715	0.0059
		2nd	0.2733	0.2025	0.2315	0.2927	0.4340	−0.2316
		3rd	0.0601	0.3651	0.5038	0.0710	0.8689	0.2033
	S. meli	1st	0.2300	0.3927	0.2481	0.1293	0.6408	0.0135
		2nd	0.2682	0.1909	0.2518	0.2891	0.4427	−0.1846
		3rd	0.0832	0.3467	0.4543	0.1158	0.8010	0.1743
II	*Buchnera* sp.NRC	1st	0.3887	0.2480	0.1313	0.2320	0.3793	0.1162
		2nd	0.3598	0.1349	0.1746	0.3307	0.3095	0.0464
		3rd	0.4222	0.0801	0.0635	0.4342	0.1436	−0.1195
	H. inf	1st	0.2798	0.3397	0.1811	0.1994	0.5208	0.1393
		2nd	0.3253	0.1563	0.2095	0.3088	0.3658	−0.0157
		3rd	0.3359	0.1436	0.1434	0.3771	0.2870	−0.0945
	*V. cho*1	1st	0.2563	0.3516	0.2255	0.1667	0.5771	0.1001
		2nd	0.3092	0.1674	0.2189	0.3046	0.3863	−0.0907
		3rd	0.2206	0.2482	0.2395	0.2916	0.4877	−0.0107

[a] 1st, 2nd, and 3rd indicate the first, second, and third codon positions, respectively.
[b] The value is the difference between the GC content at each codon position and the genomic GC content. The positive value means the GC content at this codon position is greater than the genomic GC content and vice versa.
[c] Archaeal genome.

GC content at this codon position is greater than the genomic GC content and vice versa. It can be seen that the base distribution patterns in the two groups are obviously distinct.

In group I, G is the most dominant base at the first codon position, its average content in the seven genomes is more than 40%, and T is the least dominant base. The GC content at the first codon position is similar to the genomic GC content. At the second codon position, the distribution of bases is in a state of equilibrium and G is the least dominant base. The GC content at this position is much lower than the genomic GC content, so the 'GC disparity' is negative. At the third codon position, the GC content is very high, more than 80% on average in the seven genomes studied, much higher than the genomic GC content and that of the first two codon positions. In summary, the preferred codon usage pattern in the seven genomes is the $G\bar{G}S$ type, where G, $\bar{G}$, and S are the bases of G, non-G, and G/C, respectively. The negative samples are intergenic sequences that do not code for proteins, and the frequencies of bases at three 'codon' positions are almost identical. Note that the 'codon' in a negative sample is meaningless. The significant difference of GC content at three codon positions forms the basis of the high recognition accuracy of the current algorithm. Besides the codon usage patterns, the average gene length and amino acid usage for the seven genomes are calculated, and no significant deviations are found. The similar distribution patterns of bases in the seven genomes might explain why the parameters are universal.

In contrast, for the three bacteria in group II, the codon usage patterns show different features. As mentioned above, the base frequencies at the first two codon

positions are severely restricted. The preferred bases are R and non-G at the first and second codon positions, respectively. At the third codon position, the base distribution pattern is species-specific. *Buchnera* sp. NRC is a highly AT-rich genome, the genomic GC content is 26.31%, and AT content accounts for more than 85% at the third codon position. With the genomic GC content increasing from 38.15% to 47.70% in *H. influenzae* and *V. cholerae*, the AT content at the third codon position decreases to 71% and 51%, respectively. Owing to the diverse genomic GC content in the three bacteria in group II, the GC content and 'GC disparity' at each codon position do not obey the similar rule as that in group I.

To intuitively show the difference between the codon usage patterns of the two groups, three genomes *V. cholerae*, *P. aeruginosa*, and *Halobacterium* sp. NRC are chosen to draw Figs. 1A and B. *P. aeruginosa* belongs to the same subdivision as *V. cholerae* and *Halobacterium* sp. NRC is an archaeon. Though the phylogenetic lineages of *P. aeruginosa* and *Halobacterium* sp. NRC are remote, their genomic GC content is similar and the genomic GC content of *V. cholerae* is much lower. In Figs. 1A and B, the x axis represents the value of GC_2 and y axis denotes the value of GC_3. The open cycles represent known genes of *P. aeruginosa* in each plot, while the filled cycles denote known genes of *Halobacterium* sp. NRC and *V. cholerae* in A and B, respectively. In Fig. 1A, two clusters of points largely overlap with each other, indicating that the codon usage patterns of *P. aeruginosa* and *Halobacterium* sp. NRC are similar. While in Fig. 1B, two clusters of points are well separated with little overlap, indicating that the codon usage patterns are various in the two genomes. It illustrates that the codon usage patterns are affected more by the genomic GC content, rather than by phylogenetic lineages.

The rank of importance of the three parameters in group I and the GC_2–GC_3 graph for C. crescentus and Halobacterium sp. NRC

As mentioned above, the Fisher linear discrimination equation represents a plane in the three-dimensional space. The equation can be denoted by the vector $\mathbf{c}$, which has three components c_1, c_2, and c_3, where $|\mathbf{c}|^2 = 1$. The square of each Fisher coefficient c_i^2 indicates the contribution of variable u_i to the discrimination. So the absolute value c_i can represent the importance of the variable u_i in the Fisher discrimination. In Table 2, the absolute values c_2 and c_3 are roughly equal and much greater than that of c_1 in group I, and the average value of $c_2^2 + c_3^2$ in the seven genomes is greater than 85%. So the variables u_2 and u_3 are much more important than u_1 in the seven genomes in group I. Figs. 2A and B show the distribution of points based on the variables u_2 (GC_2) and u_3 (GC_3) for *C. crescentus* and *Halobacterium* sp. NRC. The plots of other five genomes in group I, not shown here, are similar to that of *C. crescentus*. The x axis represents the value of GC_2 and y axis denotes the value of GC_3. In each plot, the open cycles represent known genes and the filled triangles indicate intergenic sequences. The two clusters can be well separated with little overlap. It can be clearly

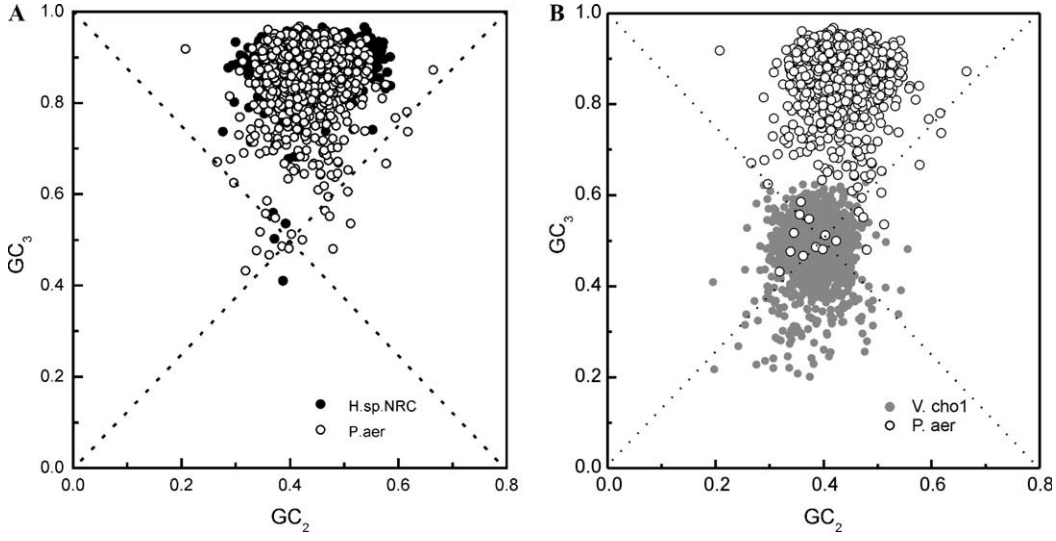

Fig. 1. (A,B) The distributions of points based on GC_2 verse GC_3 for function known genes in *P. aeruginosa*, *Halobacterium* sp. NRC, and *V. cholerae*, respectively. The x axis indicates the value of GC_2 and the y axis denotes the value of GC_3. In each plot, the open cycles represent the function known genes in *P. aeruginosa*. The filled cycles denote the function known genes in *Halobacterium* sp. NRC and *V. cholerae* in plots (A) and (B), respectively. In plot (A), two clusters of points largely overlap with each other, indicating that the codon usage patterns of the function known genes in *P. aeruginosa* and *Halobacterium* sp. NRC are very similar. While in plot (B), two clusters of points are well separated with little overlap, indicating that the codon usage patterns of the function known genes are various in *P. aeruginosa* and *V. cholerae*.

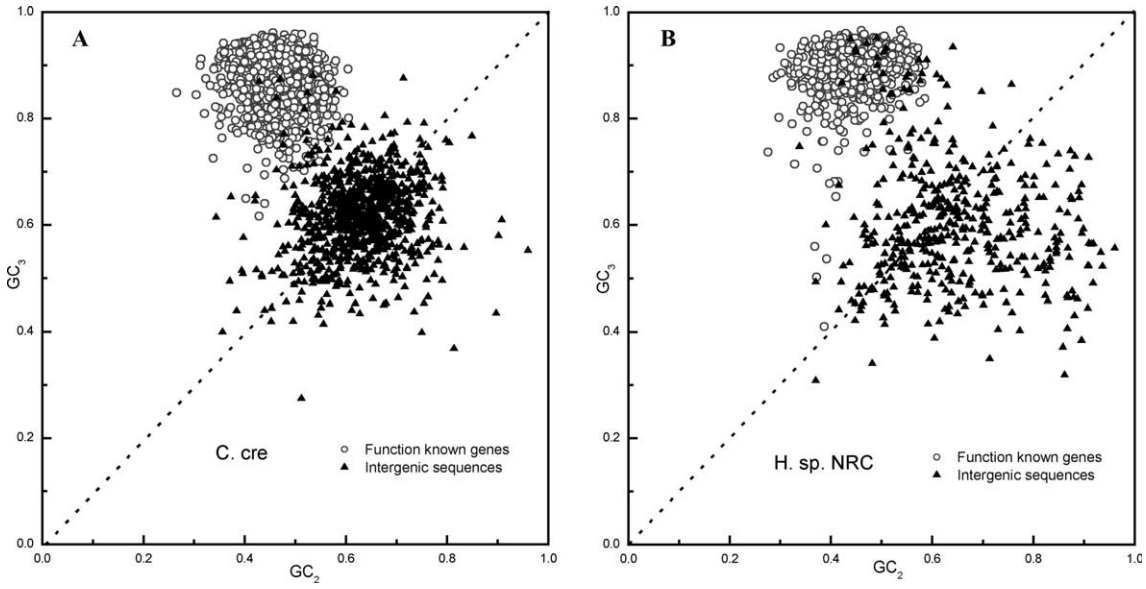

Fig. 2. (A,B) The distributions of points based on GC_2 verse GC_3 for the function known genes and intergenic sequences with length greater than 150 bp in *C. crescentus* and *Halobacterium* sp. NRC. The *x* axis indicates the value of GC_2 and the *y* axis denotes the value of GC_3. In each plot, the open cycles represent the positive samples (function known genes) and the filled triangles denote the negative samples (intergenic sequences), and the two clusters of points can be well separated with little overlap.

seen that the points corresponding to function known genes are clustered at the top left corner while those corresponding to intergenic sequences are separated and symmetrically flank the diagonal, indicating that genes have specific base usage patterns and intergenic regions are similar to random sequences. In Fig. 2B, the point distribution patterns of intergenic sequences in *Halobacterium* sp. NRC are more scattered than those in other genomes, which indicate that the base distributions of intergenic regions in archaea are different from those of bacteria.

In summary, two groups of microbial genomes are analyzed in this paper using the GC content at three codon positions. Group I contains seven microbes. Although they belong to different phylogenetic lineages, even different domains of life, the common feature of high genomic GC content forces them to adopt similar base and codon usage patterns. Therefore, it is possible to use a set of universal parameters to recognize genes in the seven genomes. Using only three parameters and the Fisher discriminant method, the accuracy for finding function known genes in each genome is satisfactory. Similar distribution patterns of bases and the significant difference of the GC content at three codon positions constitute the basis of the present algorithm. Though the phylogenetic lineages of the three genomes in group II are very close, the diverse genomic GC content compels them to adopt different codon usage patterns. So the genomic GC content is more important than phylogenetic lineage in gene recognition programs. We hope that the work might be useful for understanding the common characteristics of the organization in microbial genomes.

Acknowledgments

We thank Feng-Biao Guo and Hong-Yu Ou for invaluable assistance. The present study was supported in part by the 973 Project of China (Grant 1999075606).

References

[1] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.-F. Tomb, B.A. Dougherty, J.M. Merrick, et al., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science 269 (1995) 496–512.

[2] A. Muto, S. Osawa, The guanine and cytosine content of genomic DNA and bacterial evolution, Proc. Natl. Acad. Sci. USA 84 (1987) 166–169.

[3] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.

[4] E.N. Trifonov, Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences, J. Mol. Biol. 194 (1987) 643–652.

[5] P.M. Sharp, K.M. Devine, Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons, Nucleic Acids Res. 17 (1989) 5029–5039.

[6] S.G.E. Anderson, C.G. Kurland, Codon preferences in free-living micro organisms, Microbiol. Rev. 54 (1990) 198–210.

[7] F. Wright, M.J. Bibb, Codon usage in the G + C-rich *Streptomyces* genome, Gene 113 (1992) 55–65.

[8] C.-T. Zhang, K.C. Chou, A graphic approach to analyzing codon usage in 1562 *Escherichia coli* protein coding sequences, J. Mol. Biol. 238 (1994) 1–8.

[9] G. Gutierrez, L. Marquez, A. Marin, Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency, Nucleic Acids Res. 24 (1996) 2525–2527.

[10] J. Wang, The base contents of A, C, G or U for the three codon positions and the total coding sequences show positive correlation, J. Biomol. Struct. Dyn. 16 (1998) 51–57.
[11] A. Pan, C. Dutta, J. Das, Codon usage in highly expressed genes of *Haemophilus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias, Gene 215 (1998) 405–413.
[12] A.C. Frank, J.R. Lobry, Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, Gene 238 (1999) 65–77.
[13] I.Z. Siemion, P.J. Siemion, The informational context of the third base in amino acid codons, Biosystems 33 (1994) 39–48.
[14] F.J. Taylor, D. Coates, The code within the codons, Biosystems 22 (1989) 177–187.
[15] S.L. Salzberg, A.L. Delcher, S. Kasif, O. White, Microbial gene identification using interpolated Markov models, Nucleic Acids Res. 26 (1998) 544–548.
[16] M. Borodovsky, J. Mclninch, GenMark: parallel gene recognition for both DNA strands, Comput. Chem. 17 (1993) 123–133.
[17] C.-T. Zhang, R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic technique, Nucleic Acids Res. 19 (1991) 6313–6317.
[18] R. Zhang, C.-T. Zhang, *Z* curves, an intuitive tool for visualizing the DNA sequences, J. Biomol. Struct. Dyn. 11 (1994) 767–782.
[19] F.B. Guo, H.Y. Ou, C.-T. Zhang, ZCURVE: a new system for recognizing protein coding genes in bacterial and archaeal genomes, Nucleic Acids Res. 31 (2003) 1780–1789.
[20] W.R. Dillon, M. Goldstein, Multivariate Analysis, Methods and Applications, Wiley, New York, USA, 1984.