ELSEVIER

Computers and Chemistry 26 (2002) 195–206

Computers & Chemistry

www.elsevier.com/locate/compchem

# Using a Euclid distance discriminant method to find protein coding genes in the yeast genome

Chun-Ting Zhang [a,*], Ju Wang [a], Ren Zhang [b]

[a] *Department of Physics, Tianjin University, Tianjin 300072, China*

[b] *Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, China*

Received 11 June 2000; received in revised form 24 May 2001; accepted 18 June 2001

**Abstract**

The Euclid distance discriminant method is used to find protein coding genes in the yeast genome, based on the single nucleotide frequencies at three codon positions in the ORFs. The method is extremely simple and may be extended to find genes in prokaryotic genomes or eukaryotic genomes with less introns. Six-fold cross-validation tests have demonstrated that the accuracy of the algorithm is better than 93%. Based on this, it is found that the total number of protein coding genes in the yeast genome is less than or equal to 5579 only, about 3.8–7.0% less than 5800–6000, which is currently widely accepted. The base compositions at three codon positions are analyzed in details using a graphic method. The result shows that the preference codons adopted by yeast genes are of the R$\bar{G}$W type, where R, $\bar{G}$ and W indicate the bases of purine, non-G and A/T, whereas the 'codons' in the intergenic sequences are of the form NNN, where N denotes any base. This fact constitutes the basis of the algorithm to distinguish between coding and non-coding ORFs in the yeast genome. The names of putative non-coding ORFs are listed here in detail.

*Keywords:* Gene-finding; Yeast genome; Z curve; Euclid distance discriminant method

## 1. Introduction

The budding yeast *Saccharomyces cerevisiae* is an important model organism for the Human Genome Project. Due to the efforts of more than 600 scientists worldwide, the first sequenced genome of an eukaryotic organism, *S. cerevisiae*, has been completed (Oliver et al., 1992; Dujon et al., 1994, 1997; Bowman et al., 1997; Feldmann et al., 1994; Galibert et al., 1996; Johnston et al., 1994, 1997; Bussey et al., 1995, 1997; Murakami et al., 1995; Churcher et al., 1997; Dietrich et al., 1997; Jacq et al., 1997; Philippsen et al., 1997; Tettelin et al., 1997). Although this is a great scientific achievement, much work remains to be done. The completion of the Yeast Genome Project may be deemed as the first step in a 'Long March' towards understanding the genetic secret of this relatively simple organism. It is necessary to clarify functions of genes and relationships of them. However, to clarify the number of genes is even a more critical task at present. The number of protein coding genes in the yeast genome was estimated to be 5800–6000 (Goffeau et al., 1996; Winzeler and Davis, 1997; Mewes et al., 1997), which is currently widely accepted. On the contrary, another group estimated recently that the number should be less than 4700 (Mackiewicz et al., 1999). The results are obviously controversial.

Historically, the codingness of an ORF or a fragment of DNA sequence in the yeast genome was described by using the Codon Bias Index (CBI) (Benetzen and Benjamin, 1982) or the Codon Adaptation Index (CAI)

* Corresponding author. Tel.: +86-22-2740-2987; fax: +86-22-2335-8329.

*E-mail address:* ctzhang@tju.edu.cn (C.-T. Zhang).

0097-8485/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.

PII: S0097-8485(01)00107-3

(Sharp and Li, 1987). Although these indices were used widely (Dujon et al., 1994), the coding properties of a coding sequence are not sufficiently reflected by them. For example, some ORFs shorter than 150 codons with CAI < 0.11 have identified phenotypes (Mackiewicz et al., 1999). During the past decade, numerous advanced gene-finding algorithms have been developed. See, e.g., a recent review paper written by Fickett (1996). A set of 745 sequences in the yeast genome was selected to evaluate the gene-finding algorithm based on a correspondence analysis (Quentin et al., 1999). Genes in the same set were also predicted using the GeneMark program (Borodovsky and McIninch, 1993) (see the discussion of Quentin et al. (1999)). In this paper an extremely simple gene-finding algorithm based on the Euclid distance discriminant method is proposed. The algorithm utilizes a graphic approach to explore the difference between coding and non-coding sequences. In addition, this simple gene-finding algorithm is useful because it may be complementary with other existing methods. Therefore, by joining them, more reliable gene recognition results could be expected. Based on the algorithm, a new index is proposed to describe the codingness of yeast ORFs, which may be an appropriate complement to CBI or CAI, which are already widely used.

## 2. Materials and methods

### 2.1. The database

The *S. cerevisiae* genome DNA sequences were obtained from a CD-ROM distributed from the Munich Information Centre for Protein Sequences (MIPS), released in 1997. The data for classification of ORFs in the yeast genome were downloaded from http://speedy.mips.biochem.mpg.de, release, September 27, 1999 (Mewes et al., 1999) (the database is referred to as MIPS database, hereafter). In the MIPS database, all the ORFs are classified into six classes, which correspond to known proteins, strong similarity to known proteins, similarity or weak similarity to known proteins, similarity to unknown proteins, no similarity and questionable ORFs, respectively. The 1st, 2nd, 3rd, 4th, 5th and 6th classes include 3199 (18), 248, 869, 789 (1), 805 and 447 (8) entries, respectively, where the figures in the parentheses indicate the numbers of ORFs in the mitochondrial genome. The mitochondrial ORFs are excluded here since the samples are too few to have statistical significance. So in each of the six classes, 3181, 248, 869, 788, 805 and 439 ORFs are contained, respectively.

### 2.2. The gene-finding algorithm

The gene-finding algorithm presented in this paper is based on the differences of single nucleotide frequencies at the three codon positions between protein coding ORFs and non-coding ones. Suppose that the occurrence frequencies of the bases A, C, G and T at the 1st, 2nd and 3rd codon positions in an ORF are denoted by $a_i$, $c_i$, $g_i$ and $t_i$, respectively, where $i = 1, 2, 3$. Since $a_i + c_i + g_i + t_i = 1$, the four real numbers $a_i$, $c_i$, $g_i$and $t_i$may be mapped onto a point $P_i$ in a three-dimensional space $V_i$. The coordinates $x_i$, $y_i$ and $z_i$ of $P_i$ are determined by the so-called *Z-transform* for DNA sequence, which transforms the nucleotide frequencies into a three-dimensional curve, the Z curve (Zhang and Zhang, 1991)

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

Define $V = V_1 \oplus V_2 \oplus V_3$, i.e. the nine-dimensional space V is the direct sum of the subspaces $V_1$, $V_2$ and $V_3$. Denoting the nine bases of V by $u_1$–$u_9$, we define

$$\begin{cases} u_1 = x_1, \quad u_2 = y_1, \quad u_3 = z_1, \\ u_4 = x_2, \quad u_5 = y_2, \quad u_6 = z_2, \\ u_7 = x_3, \quad u_8 = y_3, \quad u_9 = z_3. \end{cases} \tag{2}$$

Therefore, each coding ORF or non-coding DNA sequence is represented by a point or a vector, respectively, in the nine-dimensional space V.

To complete the algorithm in a computer, usually a training set of samples (ORFs) is needed. The training set consists of two parts: one includes the positive samples composed of true protein coding genes, whereas the other includes negative samples composed of non-coding DNA sequences. Suppose that there are $N$ samples in each part. In the positive samples the $i$-th true coding ORF is described by a vector $(u^1_{i,1}, u^1_{i,2}, \ldots, u^1_{i,9})^T$, where $u^1_{i,s}$ are the $s$-component of the vector ($s = 1, 2, \ldots, 9$), and 'T' indicates a transpose operator for a matrix. Similarly, in the negative samples the $i$-th non-coding DNA sequence is described by a vector $(u^2_{i,1}, u^2_{i,2}, \ldots, u^2_{i,9})^T$, where $u^2_{i,s}$ are the $s$-component of the vector ($s = 1, 2, \ldots, 9$). The geometric centers for the positive and negative samples in the 9-dimensional space V are denoted by $\bar{U}^1$ and $\bar{U}^2$, respectively, where

$$\bar{U}^1 = (\bar{u}^1_1, \bar{u}^1_2, \ldots, \bar{u}^1_9)^T, \quad \bar{U}^2 = (\bar{u}^2_1, \bar{u}^2_2, \ldots, \bar{u}^2_9)^T, \tag{3}$$

and

$$\bar{u}^1_s = \frac{1}{N}\sum_{i=1}^{N} u^1_{i,s}, \quad \bar{u}^2_s = \frac{1}{N}\sum_{i=1}^{N} u^2_{i,s}, \quad s = 1, 2, \ldots, 9. \tag{4}$$

Suppose that a query ORF is indicated by a nine-dimensional vector $\mathbf{U} = (u_1, u_2, \ldots, u_9)^{\mathrm{T}}$. To judge whether this ORF is a true protein coding gene or not, calculate the Euclid distance $d(\mathbf{U}, \bar{\mathbf{U}}^1)$ between $\mathbf{U}$ and $\bar{\mathbf{U}}^1$, and the Euclid distance $d(\mathbf{U}, \bar{\mathbf{U}}^2)$ between $\mathbf{U}$ and $\bar{\mathbf{U}}^2$, where

$$d(\mathbf{U}, \bar{\mathbf{U}}^1) = \left[ \sum_{s=1}^{9} (u_s - \bar{u}_s^1)^2 \right]^{1/2},$$
$$d(\mathbf{U}, \bar{\mathbf{U}}^2) = \left[ \sum_{s=1}^{9} (u_s - \bar{u}_s^2)^2 \right]^{1/2}. \tag{5}$$

A codingness index $\Delta$ is defined as

$$\Delta = d(\mathbf{U}, \bar{\mathbf{U}}^2) - d(\mathbf{U}, \bar{\mathbf{U}}^1) + c, \tag{6}$$

where $c$ is a constant determined by making the false positive rate and false negative rate identical in the training set. If $\Delta > 0$, the query ORF is recognized as coding gene, otherwise, if $\Delta < 0$, the ORF or DNA sequence is recognized as a non-coding one.

## 3. Results and discussions

### 3.1. Definitions of sensitivity, specificity and accuracy

To evaluate the performance of the algorithm, we have to discuss the definitions of the accuracy, sensitivity and specificity. The notations used here are the same as in Burset and Guigo (1996). Denoted by TP the number of coding ORFs that have been correctly predicted as coding, and denoted by FN the number of coding ORFs that have been predicted as non-coding, we define the sensitivity $s_n$ as

$$s_n = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{7}$$

That is, $s_n$ is the proportion of coding ORFs that have been correctly predicted as coding. Similarly, denoted by TN the number of intergenic sequences that have been correctly predicted as non-coding, and denoted by FP the number of intergenic sequences that have been predicted as coding, we define the specificity $s_p$ as

$$s_p = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{8}$$

That is, $s_p$ is the proportion of intergenic sequences that have been correctly predicted as non-coding. The accuracy is defined as the average of $s_n$ and $s_p$.

Table 1
The accuracy of the algorithm for three different test sets

| Test set | 1 | 2 | 3 |
|---|---|---|---|
| Sensitivity (%) | 93.00 | 93.60 | 94.60 |
| Specificity (%) | 93.10 | 93.40 | 93.00 |
| Accuracy[a] (%) | 93.05 | 93.50 | 93.80 |

[a] Accuracy is defined as the average of the sensitivity and specificity.

### 3.2. Self-consistency and cross-validation tests

How to evaluate a gene-finding algorithm is an important issue. Usually, the accuracy of a gene-finding algorithm is evaluated by the resubstitution and cross-validation tests, respectively. The former reflects the self-consistency, and the latter reflects the extrapolating effectiveness of the algorithm. To evaluate the algorithm, a training set and a test set are needed, which should be independent of one another. In the MIPS database, the first class includes 3181 known genes residing in the 16 yeast chromosomes. Among them, 223 are intron-containing genes and the remaining 2958 are intronless. Randomly divide the 2958 intronless genes into two unequal parts, in which the larger part includes 1958 genes, whereas the smaller includes 1000 genes. The former is served as a training set, whereas the latter is served as a test set. Both the training and test sets should be accompanied by the counterparts of negative samples. We have randomly selected about 6000 intergenic sequences with lengths longer than 300 bp from the 16 yeast chromosomes, and each of them starts with ATG and ends with one of the stop codons. Note that such sequences are unlikely to be ORFs, because there are usually several stop codons within the sequences. We randomly selected 1958 and 1000 intergenic sequences from the above 6000 sequences, which form the training and test sets of the negative samples, respectively. In summary, the training set includes 1958 positive samples (true genes) and 1958 negative samples (intergenic sequences). The test set includes 1000 positive samples (true genes) and 1000 negative samples (intergenic sequences). Using the sequences in the training set, the average vectors $\bar{\mathbf{U}}^1$, $\bar{\mathbf{U}}^2$ and the parameter $c$ (see Eq. (6)) are determined. Using these quantities, the accuracy of gene-finding algorithm in the training and test sets is calculated, which reflects the self-consistency and extrapolating effectiveness of the algorithm. The division of 2958 ORFs into two parts (1958 and 1000) is random. Repeating the above random division procedure three times, we have performed three resubstitution and cross-validation tests. In each case, the constant $c$ is determined by making the false positive rate and false negative rate identical in the resubstitution test. The results of the cross-validation tests are listed in Table 1, where the accuracy is defined as the average of the sensitivity and the specificity. As can be seen from Table 1, the accuracy in each cross-validation test is always greater than 93%. This accuracy is comparable to that obtained by Quentin et al. (1999), based on 745 sequences. Their algorithm performed slightly better than our method, however, so far as we know, no tests on recognizing all known yeast genes are reported by them.

It should be pointed out that among the 6000 randomly selected intergenic sequences, some sequences may be coding regions containing introns, or code for small proteins or peptides. To examine their influence on the gene-finding result, we shuffle each of the 6000 intergenic sequence 20 000 times to destroy its possible coding structure (yet the overall base composition is the same as the original sequence). Then the shuffled sequences are used as the new negative samples in the algorithm. It is found that the average vector for the negative samples, the constant $c$, as well as the predictive results are very similar to those when the original negative samples are used. This means that the few possible short coding sequences among the negative samples do not affect the predictive result.

There are 223 intron-containing genes of the 1st class in the MIPS database, whose introns have been removed in advance. These ORFs are used as an independent test set to perform another three-fold cross-validation tests. Using the average vectors $\bar{\mathbf{U}}^1$, $\bar{\mathbf{U}}^2$ and the parameter $c$ obtained for each of the three training sets discussed above, the recognition for the 223 sequences is performed. Consequently, the accuracy (defined as the sensitivity in this case) in each test is always greater than 93%, based on the parameters derived from the above three training tests. In other words, a total of six cross-validation tests confirm that the accuracy of the algorithm presented is better than 93%.

### 3.3. Apply the algorithm to find genes in the ORFs of the 2nd–6th classes

After performing the resubstitution and cross-validation tests, the 1958 and 1000 positive samples (true genes) are then merged. The 2958 negative samples are selected randomly from the 6000 intergenic sequences mentioned above. These 2958 positive and 2958 negative samples form a new training set. The random selection is repeated three times. Consequently, we have three combinations. For each combination the positive samples are identical, whereas the negative samples are different each time. The average vectors $\bar{\mathbf{U}}^1$, $\bar{\mathbf{U}}^2$, and the parameter $c$ obtained in each combination are averaged over the three combinations, we find

$$\bar{\mathbf{U}}^1 = (0.2565 \quad -0.0182 \quad 0.0910 \quad -0.0038 \quad 0.1553 \quad 0.2644 \quad -0.0438 \quad -0.0259 \quad 0.2184), \tag{9}$$

$$\bar{\mathbf{U}}^2 = (0.0144 \quad 0.0142 \quad 0.2768 \quad -0.0139 \quad -0.0120 \quad 0.2824 \quad 0.0078 \quad -0.0150 \quad 0.2605), \tag{10}$$

$$c = -0.017. \tag{11}$$

We then apply the average vectors $\bar{\mathbf{U}}^1$, $\bar{\mathbf{U}}^2$, and the parameter $c$ listed in Eqs. (9)–(11) to recognizing genes in the ORFs of the 2nd–6th classes in the MIPS database. For each ORF calculate the vector $\mathbf{U} = (u_1, u_2, \ldots, u_9)^{\mathrm{T}}$, where $u_1$–$u_9$ are defined in Eqs. (1) and (2). Based on the vectors $\mathbf{U}$, $\bar{\mathbf{U}}^1$, $\bar{\mathbf{U}}^2$, and the parameter $c$, calculate the codingness index $\Delta$using Eq. (6). If $\Delta > 0$, the query ORF is recognized as a coding gene, otherwise, if $\Delta < 0$, the ORF or DNA sequence is recognized as a non-coding one.

It should be pointed out that using the algorithm and the parameters derived from the 1st class ORFs to find genes in the 2nd–6th classes is based on an assumption that both DNA sequences have similar statistical behaviors. This might not be so in some special cases, for example, for some low-expressed genes. In this case, the results of gene-finding in the 2nd–6th class ORFs should be referred to with caution. We hope to see to what extent the assumption is valid, based on a comparison between the results presented here and some future related experiments.

According to the MIPS database, there are 248, 869, 788, 805 and 439 nuclear ORFs of the 2nd–6th classes in the yeast genome. Consequently, 28, 112, 157, 215 and 355 are recognized as non-coding ORFs. The four quantities TP, TN, FP and FN mentioned above can be calculated, based on the above results and the sensitivity and specificity obtained. Compute TP, TN, FP and FN for the 2nd class ORFs in the MIPS database first. The total number of the 2nd class ORFs is 248, in which 28 are recognized as non-coding. Assume that both the sensitivity and specificity are all equal to 93%. We have a set of four linear equations as follows: $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) = 0.93$; $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP}) = 0.93$; $\mathrm{TN} + \mathrm{FN} = 28$ and $\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN} = 248$. Solving the above set of equations, we find $\mathrm{TP} \approx 219$; $\mathrm{TN} \approx 12$; $\mathrm{FP} \approx 1$ and $\mathrm{FN} \approx 16$. The number of real coding ORFs of the 2nd class should be equal to $\mathrm{TP} + \mathrm{FN} \approx 235$. Of the 28 ORFs recognized as non-coding, statistically, 16 (FN) are actually coding. Similar calculations for the 3rd–6th class ORFs are performed. The results are listed in Table 2.

Based on the above results, we re-estimate the number of protein coding genes in the 16 yeast chromosomes. The total number should be equal to: the number of intronless genes in the 1st class (2958) + the number of intron-containing genes in the 1st class (223) + the number of those in the 2nd–6th classes, including intronless and intron-containing genes, recognized by the present algorithm (235 + 810 + 670 + 620 + 63 = 2398, see Table 2). The sum is 5579. Note that the accuracy is actually greater than 93%, so, this

Table 2
The numbers of predicted coding and non-coding ORFs of the 2nd–6th classes

| Class | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Total number of ORFs ($n$)[a] | 248 | 869 | 788 | 805 | 439 |
| TP | 219 | 753 | 623 | 577 | 59 |
| TN | 12 | 55 | 110 | 172 | 350 |
| FP | 1 | 4 | 8 | 13 | 26 |
| FN | 16 | 57 | 47 | 43 | 4 |
| TP+FN[b] | 235 | 810 | 670 | 620 | 63 |
| TN+FN[b] | 28 | 112 | 157 | 215 | 354 |
| TN+FP[b] | 13 | 59 | 118 | 185 | 376 |
| (TN+FP)/$n$[c] | 13/248 = 5.2% | 59/869 = 6.8% | 118/788 = 15.0% | 185/805 = 23.0% | 376/439 = 85.6% |

[a] The mitochondrial ORFs are not included.
[b] TP+FN, TN+FN and TN+FP indicate the numbers of real coding, predicted non-coding and real non-coding ORFs, respectively, based on the assumption that both the sensitivity and specificity of the gene-finding algorithm are equal to 93.0%.
[c] The percentage of the real non-coding ORFs over the total ORFs in this class.

Table 3
The 28 ORFs of the 2nd class in the MIPS database, which are recognized as non-coding[a]

| | | | | | |
|---|---|---|---|---|---|
| YAL004w | YCL069w | YEL004w | YKL008c | YLR034c | YMR118c |
| YAR061w | YDR033w | YER039c | YKL033w-a | YLR046c | YMR279c |
| YBL009w | YDR107c | YER185w | YKR027w | YLR164w | YNL320w |
| YBR161w | YDR276c | YGL054c | YKR105c | YLR176c | |
| YBR210w | YDR384c | YGR131w | YLL051c | YMR040w | |

[a] Of the 28 ORFs listed, statistically, 16 are actually coding. Unfortunately, we cannot identify them at present due to the limited recognition accuracy achieved.

figure should be considered as an up-limit of gene number in the yeast genome. The above estimate of protein coding genes in the yeast genome is about 3.8–7.0% less than that of 5800–6000, which is widely accepted (Goffeau et al., 1996; Winzeler and Davis, 1997; Mewes et al., 1997). The above estimate is based on error analysis, i.e. we have considered the false positive and false negative events in the prediction for each class. So, it should be statistically reliable. As we can see in Table 2, the ratio of non-coding ORFs in each of the 2nd–6th classes is different, which is about 5.2% in 2nd class and 85.6% in 6th class. Meanwhile our estimate (5579) is about 18.7% larger than 4700, estimated recently by another group (Mackiewicz et al., 1999).

The names of the non-coding ORFs of the 2nd–6th classes recognized by the present algorithm are listed in Tables 3–7, respectively. Some of them are actually coding ORFs, but we cannot identify them at present due to the limited accuracy achieved here. We list only the number of such coding ORFs (i.e. FN) in the footnotes of Tables 3–7, respectively, for the 2nd–6th class ORFs in the MIPS database.

### 3.4. Graphic analysis of base composition at different codon positions

As described in Section 2.2, the base composition of an ORF at each codon position can be represented by a point with coordinate ($x_i$, $y_i$, $z_i$) in a three-dimensional space $V_i$, where $i = 1$, 2, 3. Therefore, a set of ORFs is associated with three sets of three-dimensional mapping points. The distribution pattern of these points can be studied in a graphic approach (Zhang and Zhang, 1991; Zhang and Chou, 1994). Using the graphic method, we would like to get some insight into the working mechanism of the algorithm and find the reason why a considerable part of the ORFs in the 2nd–6th classes are actually not coding genes.

As mentioned above, the number of known genes is 3181, and the total number of the ORFs in the 2nd–6th classes is equal to $248 + 869 + 789 + 805 + 447$, i.e. 3158. For comparison, the 6000 intergenic sequences are used as the negative samples. The graphs (Zhang and Zhang, 1991; Zhang and Chou, 1994) corresponding to these data sets are drawn and compared. To save printing space, only the projections onto the $x$–$y$ or $x$–$z$ planes are shown here. Consider the known genes

Table 4
The 112 ORFs of the 3rd class in the MIPS database, which are recognized as non-coding[a]

YBL089w
YBL091c-a
YBR074w
YBR180w
YBR220c
YBR293w
YCL001w-a
YCR023c
YCR062w
YCR087c-a
YDL015c
YDL119c
YDL199c
YDL206w
YDL228c
YDR100w
YDR115w
YDR119w
YDR205w

YDR249c
YDR302w
YDR303c
YDR307w
YDR366c
YDR387c
YDR411c
YDR413c
YEL045c
YEL064c
YER048w-a
YER097w
YER113c
YER119c
YER184c
YFL027c
YFL040w
YFR057w
YGL104c

YGL160w
YGL186c
YGR023w
YGR065c
YGR067c
YGR077c
YGR101w
YGR284c
YHL035c
YHR002w
YHR035w
YHR048w
YHR130c
YHR142w
YHR181w
YIL025c
YIL040w
YIL054w
YIL088c

YIL166c
YJL091c
YJL170c
YJL193w
YJR036c
YJR124c
YJR136c
YKL037w
YKL174c
YKL221w
YKL222c
YKR030w
YKR088c
YKR103w
YLL005c
YLL037w
YLL054c
YLR010c
YLR050c

YLR064w
YLR184w
YLR251w
YLR266c
YLR283w
YLR311c
YLR365w
YLR394w
YML023c
YMR088c
YMR221c
YMR245w
YMR306w
YNL065w
YNL109w
YNL176c
YNL203c
YNL275w
YNL305c

YNR056c
YNR059w
YNR063w
YOL079w
YOL107w
YOL119c
YOL137w
YOL152w
YOL163w
YOR049c
YOR053w
YOR292c
YOR297c
YOR350c
YPL125w
YPL244c
YPR094w

[a] Of the 112 ORFs listed, statistically, 57 are actually coding. Unfortunately, we cannot identify them at present due to the limited recognition accuracy achieved.

Table 5
The 157 ORFs of the 4th class in the MIPS database, which are recognized as non-coding[a]

YAL018c
YAL034c
YAR060c
YAR068w
YBL108w
YBL109w
YBR004c
YBR099c
YBR147w
YBR168w
YBR183w
YBR300c
YBR302c
YCL002c
YCL038c
YCL073c
YCR102w-a
YCR103c
YDL123w
YDL183c
YDL248w
YDR018c
YDR066c
YDR084c
YDR105c
YDR126w
YDR131c

YDR306c
YDR438w
YDR459c
YDR492w
YDR504c
YDR524c
YDR525w-a
YDR543c
YDR544c
YEL033w
YEL067c
YER072w
YER188c-a
YFL015c
YFL062w
YFL063w
YFL065c
YFL068w
YFR012w
YGL010w
YGL041c
YGL084c
YGL124c
YGL260w
YGL263w
YGR015c
YGR016w

YGR071c
YGR110w
YGR125w
YGR212w
YGR293c
YGR295c
YHL041w
YHL042w
YHL044w
YHL045w
YHL048w
YHR054c
YHR133c
YHR162w
YHR212c
YHR214w-a
YHR217c
YIL029c
YIL089w
YIL090w
YIL174w
YIL175w
YIR040c
YIR043c
YIR044c
YJL062w
YJL097w

YJL108c
YJL147c
YJR013w
YJR044c
YJR116w
YJR161c
YJR162c
YKL034w
YKL219w
YKL223w
YKL225w
YKR051w
YKR106w
YLL023c
YLL031c
YLR023c
YLR036c
YLR047c
YLR156w
YLR159w
YLR161w
YLR241w
YLR246w
YLR414c
YLR463c
YML033w
YML036w

YML047c
YML132w
YMR010w
YMR034c
YMR101c
YMR119w
YMR155w
YMR253c
YMR324c
YMR326c
YNL008c
YNL026w
YNL101w
YNL156c
YNL297c
YNL326c
YNL336w
YNL337w
YNR062c
YNR075w
YNR077c
YOL002c
YOL003c
YOL048c
YOL092w
YOL101c
YOL129w

YOL162w
YOR044w
YOR147w
YOR175c
YOR193w
YOR228c
YOR245c
YOR365c
YOR390w
YPL087w
YPL165c
YPL189w
YPL229w
YPL246c
YPL257w
YPL264c
YPL279c
YPR071w
YPR114w
YDR367w*
YMR292w*
YOL047c*

[a] Of the 157 ORFs listed above, 154 are intronless and three are intron-containing (marked with *). Note that of the 157 ORFs listed, statistically, 47 are actually coding. Unfortunately, we cannot identify them at present due to the limited recognition accuracy achieved.

Table 6
The 215 ORFs of the 5th class in the MIPS database, which are recognized as non-coding[a]

YAL008w
YAL064w
YAL066w
YAR030c
YAR040c
YAR047c
YAR053w
YAR064w
YAR069c
YAR070c
YBL048w
YBL049w
YBL071c
YBR013c
YBR027c
YBR058c-a
YBR085c-a
YBR096w
YBR126w-a
YBR144c
YBR292c
YCL056c
YCL057c-a
YCL058c
YCR001w
YCR006c
YCR022c
YCR025c
YCR043c
YCR085w
YDL027c
YDL054c
YDL089w
YDL162c
YDL180w
YDL196w
YDL231c
YDR015c
YDR024w
YDR029w
YDR042c
YDR065w
YDR102c
YDR141c
YDR179w-a
YDR215c
YDR274c
YDR278c
YDR315c
YDR319c
YDR344c
YDR350c
YDR396w
YDR437w
YDR524w-a
YDR525w
YEL010w
YEL014c
YEL059w
YER044c
YER050c
YER066c-a
YER091c-a
YER135c
YER140w
YER172c-a
YFL019c
YFL021c-a
YFR035c
YFR042w
YFR054c
YGL006w-a
YGL057c
YGL138c
YGL188c
YGL230c
YGR026w
YGR149w
YGR168c
YGR226c
YGR290w
YGR291c
YHL005c
YHL037c
YHR067w
YHR078w
YHR095w
YHR139c-a
YHR140w
YHR173c
YIL012w
YIL028w
YIL037c
YIL071c
YIL086c
YIL152w
YIR020c
YIR020c-a
YJL027c
YJL028w
YJL064w
YJL077c
YJL118w
YJL136w-a
YJL163c
YJL215c
YJR011c
YJR023c
YJR041c
YJR120w
YJR157w
YKL044w
YKL051w
YKL097c
YKL102c
YKL158w
YKL162c
YKR032w
YKR065c
YKR073c
YLL007c
YLL014w
YLL030c
YLL033w
YLL042c
YLL059c
YLR042c
YLR049c
YLR111w
YLR112w
YLR122c
YLR124w
YLR199c
YLR255c
YLR264c-a
YLR265c
YLR267w
YLR296w
YLR312c
YLR366w
YLR376c
YLR381w
YLR400w
YLR402w
YLR404w
YML003w
YML038c
YML084w
YML090w
YML107c
YML122c
YMR003w
YMR007w
YMR057c
YMR071c
YMR082c
YMR103c
YMR122c
YMR141c
YMR148w
YMR151w
YMR163c
YMR187c
YMR191w
YMR252c
YMR254c
YMR258c
YMR259c
YMR320w
YNL017c
YNL038w
YNL122c
YNL143c
YNL146w
YNL150w
YNL174w
YNL179c
YNL211c
YNL269w
YNL303w
YNL311c
YNL324w
YOL024w
YOL038c-a
YOL072w
YOL118c
YOL160w
YOL166c
YOR015w
YOR024w
YOR029w
YOR068c
YOR072w
YOR080w
YOR152c
YOR183w
YOR268c
YOR314w
YOR364w
YOR376w
YOR392w
YPL041c
YPL052w
YPL056c
YPL066w
YPL162c
YPL200w
YPR012w
YPR014c
YPR064w
YPR098c
YPR153w
YPR170c
YPR170w-a
YDR535c*

[a] Of the 215 ORFs listed above, 214 are intronless and one is intron-containing (marked with *). Note that of the 215 ORFs listed, statistically, 43 are actually coding. Unfortunately, we cannot identify them at present due to the limited recognition accuracy achieved.

first. The projections of the 3181 mapping points onto the $x$–$y$ planes for the base composition at the 1st and 2nd codon positions are shown in Fig. 1(a) and (b), respectively. The projection of these mapping points onto the $x$–$z$ plane for the base composition at the 3rd codon position is shown in Fig. 1(c). As can be seen, the distribution patterns of the projection points of the three plots are considerably different. In Fig. 1(a), most of the points are situated at the region where $a_1 > t_1$ and $g_1 > c_1$, i.e. purine bases are predominant at the first codon position. In Fig. 1(b), most of the points are situated at the region lacking G. In Fig. 1(c), most of the points are at the region where A and T are predominant. In summary, the preference codons of the yeast protein coding genes are of the R$\bar{\text{G}}$W type, where R, $\bar{\text{G}}$ and W are the bases of purine, non-G and A/T, respectively. For comparison, the projection of the mapping points onto the $x$–$y$ plane for the base composition at the 1st 'codon' position of the 6000 intergenic sequences is shown in Fig. 2. Note that the 'codon' in an intergenic sequence is meaningless. Because all the intergenic sequences selected here begin with ATG, for example, ATGGCGCAT…, the bases A, G, C… are defined as at the first 'codon' position and so forth. Since the distribution patterns of the points at the 2nd and 3rd 'codon' positions are almost identical with that at the 1st 'codon' position, they are not shown here. Therefore, the 'codons' of the intergenic sequences are of the type NNN, where N indicates any base.

It has been suggested that the first, second and third position of the codons are associated, respectively, with the biosynthetic pathway, hydrophobicity pattern, and

Table 7
The 355 ORFs of the 6th class in the MIPS database, which are recognized as non-coding[a]

YAL034c-b
YAL042c-a
YAL056c-a
YBL012c
YBL053w
YBL062w
YBL065w
YBL070c
YBL073w
YBL077w
YBL083c
YBL094c
YBL107w-a
YBR051w
YBR064w
YBR089w
YBR109w-a
YBR113w
YBR116c
YBR124w
YBR134w
YBR178w
YBR206w
YBR224w
YBR226c
YBR232c
YBR266c
YBR277c
YCL006c
YCL023c
YCL041c
YCL065w
YCR018c-a
YCR041w
YCR064c
YCR087w
YDL009c
YDL011c
YDL016c
YDL023c
YDL026w
YDL032w
YDL034w
YDL041w
YDL050c
YDL062w
YDL068w
YDL071c
YDL094c
YDL151c
YDL152w
YDL158c
YDL172c
YDL187c
YDL221w
YDR008c
YDR034c-a
YDR048c
YDR053w
YDR112w
YDR114c
YDR133c
YDR136c
YDR149c
YDR154c
YDR157w
YDR187c
YDR199w
YDR203w
YDR220c
YDR230w
YDR241w
YDR269c
YDR271c
YDR290w
YDR340w
YDR355c
YDR360w
YDR401w
YDR417c
YDR426c
YDR431w
YDR442w
YDR445c
YDR455c
YDR467c
YDR491c
YDR509w
YDR521w
YDR526c
YEL075w-a
YER006c-a
YER046w-a
YER067c-a
YER084w
YER138w-a
YER145c-a
YER148w-a
YER165c-a
YER181c
YFL012w-a
YFL013w-a
YFL032w
YFR036w-a
YFR056c
YGL024w
YGL042c
YGL052w
YGL072c
YGL074c
YGL088w
YGL102c
YGL109w
YGL118c
YGL132w
YGL149w
YGL152c
YGL165c
YGL168w
YGL177w
YGL182c
YGL193c
YGL204c
YGL214w
YGL217c
YGL218w
YGL239c
YGR011w
YGR018c
YGR025w
YGR039w
YGR045c
YGR050c
YGR051c
YGR064w
YGR069w
YGR073c
YGR107w
YGR114c
YGR115c
YGR122c-a
YGR137w
YGR139w
YGR151c
YGR164w
YGR176w
YGR182c
YGR190c
YGR219w
YGR228w
YGR242w
YGR259c
YGR265w
YHL002c-a
YHL006w-a
YHL030w-a
YHL046w-a
YHR049c-a
YHR056w-a
YHR063w-a
YHR125w
YHR145c
YIL060w
YIL066w-a
YIL068w-a
YIL071w-a
YIL100c-a
YIL156w-a
YIL163c
YIL171w-a
YIR017w-a
YIR023c-a
YJL009w
YJL015c
YJL022w
YJL032w
YJL067w
YJL075c
YJL086c
YJL120w
YJL135w
YJL142c
YJL150w
YJL152w
YJL169w
YJL175w
YJL182c
YJL188c
YJL202c
YJL220w
YJR018w
YJR020w
YJR037w
YJR038c
YJR071w
YJR087w
YJR128w
YKL030w
YKL036c
YKL053w
YKL076c
YKL083w
YKL111c
YKL115c
YKL118w
YKL123w
YKL131w
YKL136w
YKL147c
YKL153w
YKL162c-a
YKL169c
YKL177w
YKL202w
YKR033c
YKR040c
YKR047w
YLL020c
YLR101c
YLR123c
YLR140w
YLR169w
YLR171w
YLR198c
YLR217w
YLR230w
YLR232w
YLR252w
YLR261c
YLR269c
YLR279w
YLR282c
YLR294c
YLR302c
YLR317w
YLR322w
YLR334c
YLR339c
YLR358c
YLR379w
YLR428c
YLR434c
YLR444c
YLR458w
YLR465c
YML009w-a
YML012c-a
YML031c-a
YML034c-a
YML047w-a
YML057c-a
YML089c
YML094c-a
YML099w-a
YML116w-a
YMR046w-a
YMR052c-a
YMR075c-a
YMR086c-a
YMR119w-a
YMR135w-a
YMR153c-a
YMR158c-b
YMR158w-a
YMR172c-a
YMR193c-a
YMR290w-a
YMR304c-a
YMR306c-a
YMR316c-a
YNL013c
YNL028w
YNL089c
YNL105w
YNL114c
YNL120c
YNL170w
YNL171c
YNL184c
YNL198c
YNL205c
YNL226w
YNL228w
YNL235c
YNL266w
YNL276c
YNL296w
YNL319w
YNR005c
YNR025c
YOL013w-a
YOL013w-a
YOL035c
YOL037c
YOL099c
YOL106w
YOL134c
YOL150c
YOR041c
YOR082c
YOR102w
YOR105w
YOR121c
YOR135c
YOR146w
YOR169c
YOR170w
YOR199w
YOR200w
YOR203w
YOR218c
YOR225w
YOR235w
YOR248w
YOR263c
YOR277c
YOR282w
YOR300w
YOR309c
YOR325w
YOR331c
YOR333c
YOR345c
YOR379c
YPL034w
YPL035c
YPL044c
YPL073c
YPL102c
YPL114w
YPL136w
YPL182c
YPL185w
YPL205c
YPL238c
YPL261c
YPR002c-a
YPR038w
YPR039w
YPR050c
YPR053c
YPR077c
YPR087w
YPR092w
YPR099c
YPR126c
YPR130c
YPR136c
YPR142c
YPR146c
YPR150w
YPR177c
YBR090c*
YER014c-a*
YLR202c*

[a] Of the 355 ORFs listed above, 352 are intronless and three are intron-containing (marked with *). Note that of the 355 ORFs listed, statistically, four are actually coding. Unfortunately, we cannot identify them at present due to the limited recognition accuracy achieved.

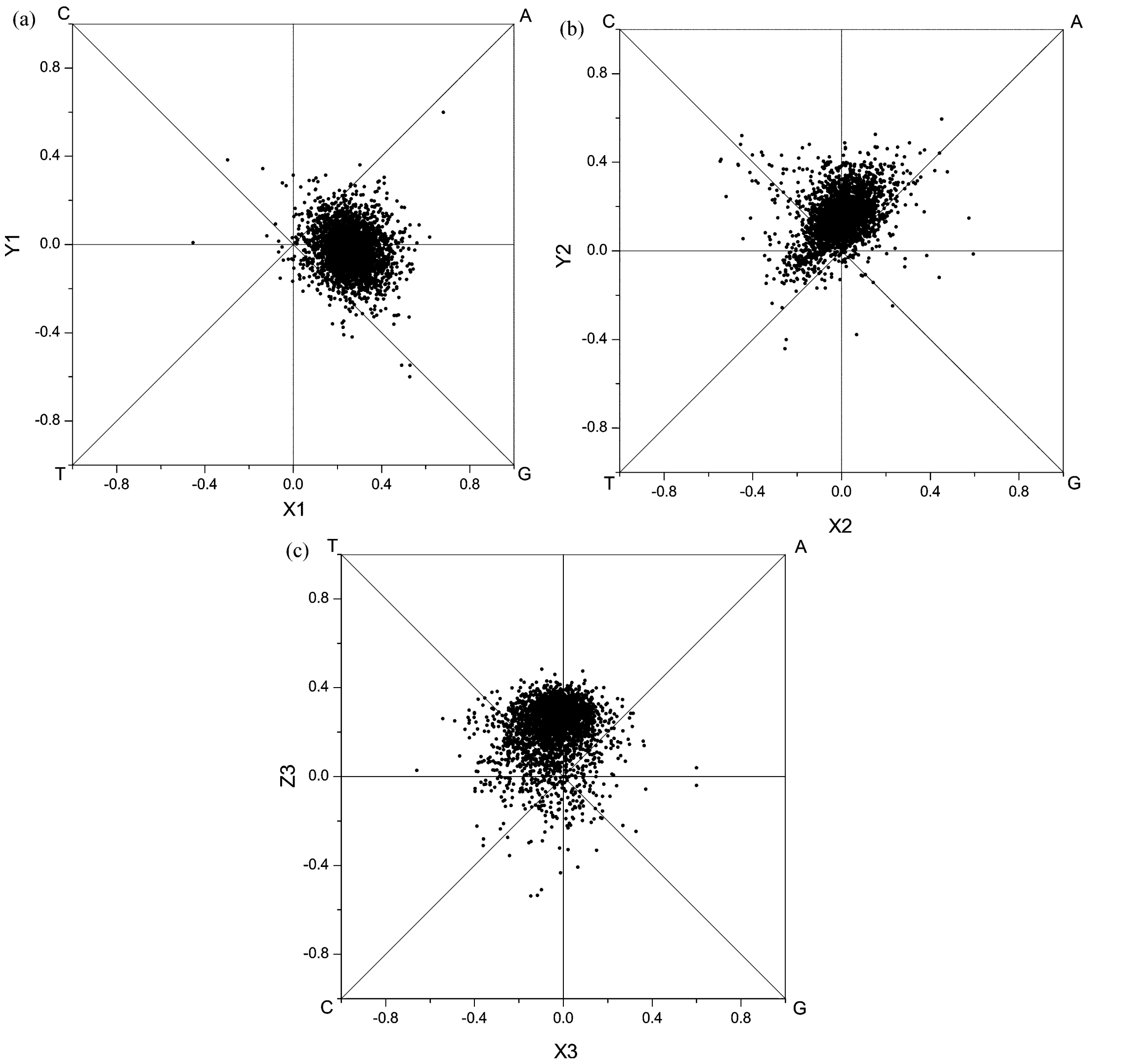

Fig. 1. Distributions of the mapping points for base composition at three codon positions of the 3181 protein coding genes in the yeast genome in a three-dimensional space. (a) Projection onto the $x$–$y$ plane for the 1st codon position; (b) projection onto the $x$–$y$ plane for the 2nd codon position; and (c) projection onto the $x$–$z$ plane for the 3rd codon position. For more detailed explanation regarding the graph, refer to Zhang and Zhang (1991) or Zhang and Chou (1994).

the alpha helix or beta strand forming potentiality of the coded amino acids (Siemion and Siemion, 1994; Taylor and Coates, 1989). Recently, it is reported that there is strong correlation between the base frequencies in the second codon position of genes and the corresponding secondary structures in the encoded proteins (Gupta et al., 2000; Chiusano et al., 2000). Chiusano et al. (2000) attributed this relation to the hydrophobic and hydrophilic amino acids encoded by codons having U or A, respectively, in their second codon site. It is also supposed that the specific codon choice is functionally needed in mRNA–rRNA interaction in ribosome, which is responsible for monitoring the correct reading frame during translation (Trifonov, 1987; Lagunez-Otero and Trifonov, 1992; Lobry and Gautier, 1994). These mean that the three codon positions are associated with different biological functions and the base choices at these positions are usually specific. In the case of yeast genome, the preferred codon usage pattern is R$\bar{G}$W, and the difference of the two codon types, i.e. R$\bar{G}$W and NNN, forms the basis to distinguish between coding and non-coding sequences. The present algorithm is based on the difference of mapping point distribution patterns between the two kinds of sequences. As we can see, the present algorithm works well.

For comparison, the projection of the mapping points for the base composition at the 1st codon position of the 3158 ORFs of the 2nd–6th classes in the MIPS database, is shown in Fig. 3. Obviously, the distribution pattern of Fig. 3 is in between those of Fig. 1(a) and Fig. 2, indicating that a part of ORFs of the 2nd–6th classes in the MIPS database are actually of non-coding ones. This fact implies that the codons in some ORFs of the 2nd–6th classes in the MIPS database are of the R$\bar{G}$W type, whereas the 'codons' in some other ORFs are of the NNN type. The latter 'codons ' do not code for any realistic proteins. This is the reason why 23% (refer to the last row in Table 2) of the ORFs of the 2nd–6th classes in the MIPS database are non-coding.

In fact, the division between coding and non-coding regions can be seen in a more intuitive manner by a principal component analysis (PCA). PCA defines a rotation of the variables of given data. The first derived direction (a linear combination of the variables) is chosen to maximize the standard deviation of the derived variable, the second to maximize the standard deviation among directions un-correlated with the first, and so forth. For details about this method refer to Dillon and Goldstein (1984). For the data set comprising the values of $u_1$–$u_9$ of 3181 known genes and 6000 intergenic sequences, a plot based on the two most important axes using the PCA is shown in Fig. 4. The first and the second axis account for 34.3 and 15.9% of the total inertia of the nine-dimensional space, respectively, and no other axis accounts for more than 10%. The variation in the second axis is mainly because of a small number of outliers, which are mostly short genes or ORFs. The two principal axes are responsible for separating the coding and non-coding sequences into two clusters. The coding sequences are represented by close cycles and the non-coding sequences are represented by open cycles. As we can see, the two clusters appear to be distinct, with quite little overlap. The closeness of any two regions in Fig. 4 reflects the similarities of their base frequencies at the three codon positions, implying that the base choices at the three codon positions for most genes are quite different from those of non-coding sequences.

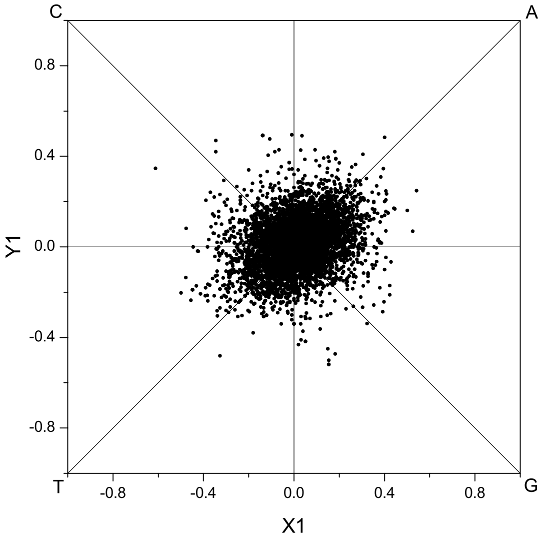

Fig. 2. Distributions of the mapping points for base composition of about 6000 intergenic sequences in the yeast genome in a three-dimensional space. Projection onto the $x$–$y$ plane for the 1st 'codon' position. It should be pointed out that the distribution patterns for the 2nd and 3rd 'codon' positions (not shown here) are almost identical to that for the 1st 'codon' position. Note that the 'codon' in an intergenic sequence is meaningless. The definition of 'codon' positions in an intergenic sequence is arbitrary. See the legend of Fig. 1 for explanation of the graph.

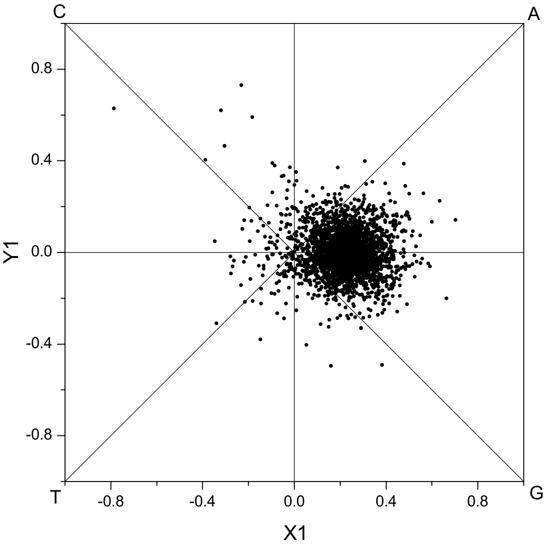

Fig. 3. Distributions of the mapping points for base composition of the 3158 ORFs of the 2nd–6th classes in the MIPS database in a three-dimensional space. Projection onto the $x$–$y$ plane for the 1st codon position. Specially note that the distribution pattern here is in between those in Fig. 1(a) and Fig. 2, indicating that a part of the ORFs of the 2nd–6th classes in the MIPS database is actually non-coding. See the legend of Fig. 1 for explanation of the graph.

## 4. Conclusions

A simple gene-finding algorithm with high accuracy (93%) for the yeast genome is presented in this paper. Six-fold cross-validation tests confirm the above accuracy. Using the algorithm, it is found that 751 ORFs (about 23.8%) of the 2nd–6th classes classified in the MIPS database are likely non-coding. The total number of protein coding genes in the 16 yeast chromosomes is estimated to be less than or equal to 5579. This estimate is based on the assumption that the DNA sequences coding for proteins in the 1st class ORFs have similar

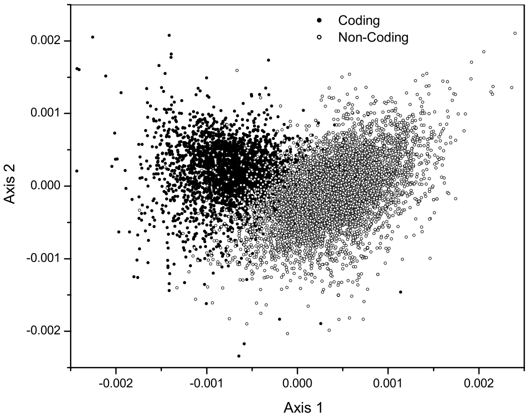

Fig. 4. The distribution of points based on the two most important axes using the principal component analysis of the nine variables $u_1$–$u_9$ for the 3181 known nuclear genes and 6000 intergenic sequences. Axis 1 is a derived direction (a linear combination of the variables $u_1$–$u_9$) chosen to maximize the standard deviation of the derived variable, and Axis 2 is the direction to maximize the standard deviation among directions un-correlated with the first. For the values of $u_1$–$u_9$ of 3181 known genes and 6000 intergenic sequences, Axes 1 and 2 account for 34.3 and 15.9% of the total inertia of the nine-dimensional space, respectively. The close cycles indicate the known genes while the open cycles represent the non-coding sequences. Note that the two clusters appear to be distinct, with quite little overlap. This fact constitutes the basis of the present algorithm.

statistical properties to those coding for proteins in the 2nd–6th class ORFs. We hope to see to what extent the above assumption is valid, relying on a comparison between the results presented here and some future related experiments. The working mechanism of the present algorithm is studied in detail by a graphic approach. It is found that the preference codons of the yeast protein coding genes are of the R$\bar{\text{G}}$W type, where R, $\bar{\text{G}}$ and W are the bases of purine, non-G and A/T, respectively, whereas the 'codons' of the intergenic sequences are of the type NNN, where N indicates any base. The difference of the two codon types, i.e. R$\bar{\text{G}}$W and NNN, forms the basis of the algorithm to distinguish between coding and non-coding sequences. Mathematically, the present algorithm is based on the difference of mapping point distributions between coding and non-coding sequences. The algorithm can be extended to find genes in any prokaryotic genome, but it cannot be directly applied to higher eukaryotic genomes with more introns.

## Acknowledgements

We are grateful to both anonymous referees for carefully evaluating the manuscript and helpful suggestions, which are very important for improving the quality of the manuscript. We thank Mr. F.-B. Guo for doing the principal component analysis. The present study was supported in part by the 973 Project of China (grant G1999075606).

## References

Benetzen, J., Benjamin, D.H., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.

Borodovsky, M.Y., McIninch, J.D., 1993. GeneMark: parallel gene recognition for both DNA strands. Comput. Chem. 17, 123–153.

Bowman, S., Churcher, C., Badcock, K., Brown, D., Chillingworth, T., Connor, R., Dedman, K., Gentels, S., Hamlin, N., Hunt, S., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIII. Nature. (Suppl.) 387, 90–93.

Burset, M., Guigo, R., 1996. Evaluation of gene structure prediction programs. Genomics 34, 353–367.

Bussey, H., Kabak, D.B., Zhong, W., Vo, D.T., Cloak, M.W., Fortin, N., Hall, J., Ouellette, B.F., Keng, T., Barton, A.B., et al., 1995. The nucleotide sequence of chromosome I from *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. 92, 3809–3813.

Bussey, H., Storms, R.K., Ahmed, A., Albermann, K., Allen, E., Ansorge, W., Araujo, R., Aparicio, A., Barrell, B., Badcock, K., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XVI. Nature. (Suppl.) 387, 103–105.

Chiusano, M.L., Alvarez-Valin, F., Giulio, M.D., D'Onofrio, G., Ammirato, G., Colonna, G., Bernardi, G., 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene 261, 63–69.

Churcher, C., Bowman, S., Badcock, K., Bankier, A., Brown, D., Chillingworth, T., Connor, R., Delvin, K., Gentles, S., Hamlyn, N., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IX. Nature. (Suppl.) 387, 84–87.

Dietrich, F.S., Mulligan, J., Hennessy, K., Yelton, M.A., Allen, E., Araujo, R., Aviles, E., Berno, A., Brennan, T., Carpenter, J., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome V. Nature. (Suppl.) 387, 78–81.

Dillon, W.R., Goldstein, M., 1984. Multivariate Analysis, Methods and Applications. Wiley, New York.

Dujon, B., Alexandraki, D., Andre, B., Ansorge, W., Baladron, V., Ballesta, J.P.G., Banrevi, A., Bolle, P.A., Bolotin-Fukuhara, M., Bossier, P., et al., 1994. Complete DNA sequence of yeast chromosome XI. Nature. (Suppl.) 369, 371–378.

Dujon, B., Albermann, K., Aldea, M., Alexandraki, D., Ansorge, W., Arino, J., Benes, V., Bohn, C., Bolotin-Fukuhara, M., Bordonne, R., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XV. Nature. (Suppl.) 387, 98–102.

Feldmann, H., et al., 1994. Complete DNA sequence of yeast chromosome II. EMBO J. 13, 5793–5809.

Fickett, J.W., 1996. Finding genes by computer: the state of the art. Trends Genet. 12, 316–320.

Galibert, F., Alexandraki, D., Baur, A., Boles, E., Chalwatzis, N., Chuat, J.-C., Coster, F., Cziepluch, C., De Haan, M., Domde, H., et al., 1996. Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X. EMBO J. 15, 2031–2049.

Goffeau, A., Barrel, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettlin, H., Oliver, S.G., 1996. Science 274, 546.

Gupta, S.K., Majumdar, S., Bhattacharya, T.K., Ghosh, T.C., 2000. Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem. Biophys. Res. Commun. 269, 692–696.

Jacq, C., Alt-Morbe, J., Andre, B., Arnold, W., Bahr, A., Ballesta, J.P.G., Bargues, M., Baron, L., Becker, A., Biteau, N., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV. Nature. (Suppl.) 387, 75–78.

Johnston, M., Andrews, S., Brinkman, R., Cooper, J., Ding, H., Dover, J., Du, Z., Favello, A., Fulton, L., Gattung, S., et al., 1994. Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. Science 265, 2077–2082.

Johnston, M., Hillier, L., Riles, L., Albermann, K., Andre, B., Ansorge, W., Benes, V., Bruckner, M., Delius, H., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. Nature. (Suppl.) 387, 87–90.

Lagunez-Otero, J., Trifonov, E.N., 1992. mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. J. Biomolec. Struct. Dyn. 10, 455–464.

Lobry, J.R., Gautier, C., 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. 22, 3174–3180.

Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M.R., Cebrat, S., 1999. Origin and properties of non-coding ORFs in the yeast genome. Nucleic Acids Res. 27, 3503–3509.

Mewes, H.W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., Pfeiffer, F., Zollner, A., 1997. Overview of the yeast genome. Nature. (Suppl.) 387, 7–8.

Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., Frishman, D., 1999. MIPS: a database for protein sequences and complete genomes. Nucleic Acids Res. 27, 44–48.

Murakami, Y., Naitou, M., Hagiwara, H., Shibata, T., Ozawa, M., Sasanuma, S.I., Sasanuma, M., Tsuchiya, Y., Soeda, E., Yokoyama, K., et al., 1995. Analysis of the nucleotide sequence of chromosome VI from *Saccharomyces cerevisiae*. Nature Genet. 10, 262–268.

Oliver, S.G., van der Aart, Q.J.M., Agosoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P.G., Benit, P., et al., 1992. The complete DNA sequence of yeast chromosome III. Nature 357, 38–46.

Philippsen, P., Kleine, K., Pohlmann, R., Dusterhoft, A., Hamberg, K., Hagemann, J.H., Obermaier, B., Urrestarazu, L.A., Aert, R., Albermann, K., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV and its evolutionary implications. Nature. (Suppl.) 387, 93–98.

Quentin, Y., Voiblet, C., Martin, F., Fichant, G., 1999. Protein-coding region discovery in organisms underrepresented in databases. Comput. Chem. 23, 209–217.

Sharp, P.M., Li, W.-H., 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential application. Nucleic Acids Res. 15, 1281–1295.

Siemion, I.Z., Siemion, P.J., 1994. The informational context of the third base in amino acid codons. Biosystems 33, 39–48.

Taylor, F.J., Coates, D., 1989. The code within the codons. Biosystems 22, 177–187.

Tettelin, H., Agostoni-Carbone, M.L., Albermann, K., Albers, M., Arroyo, J., Backes, U., Barreiros, T., Bertani, I., Bjourson, A.J., Brucker, M., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome VII. Nature. (Suppl.) 387, 81–84.

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

Winzeler, E.A., Davis, R.W., 1997. Functional analysis of the yeast genome. Curr. Opin. Genet. Dev. 7, 771–776.

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

Zhang, C.T., Zhang, R., 1991. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. Nucleic Acids Res. 19, 6313–6317.