Proceedings of the 2006 WSEAS International Conference on Mathematical Biology and Ecology, Miami, Florida, USA, January 18-20, 2006 (pp161-167)

A method to find protein coding genes in the yeast genome based on a 3D graphical representation of DNA sequence

Chun-xin Yuan[1*], Chun Li[1], Da-chao Li[2]
[1]Department of Applied Mathematics,
Dalian University of Technology,
Dalian 116024, China
[2] Department of Mathematics,
Hainan Normal University,
Haikou, 571158, China
It is supported by National Natural Science Foundation of China(10571019)

Abstract: - We develop a method to find protein coding genes based on a 3D graphical representation of DNA sequence. The method is simple and robust. We illustrate it on the yeast genome and it may be extended to find genes in prokaryotic genomes or eukaryotic genomes with less introns. Three-fold cross-validation tests have demonstrated that the accuracy of the algorithm is better than 96%. Based on this, it is found that the total number of protein coding genes in the yeast genome is 5891～5920. Among the ORFs annotated in the MIPS database, those recognized as non-coding by the present algorithm are listed in this paper in detail.

Key-Words: - Gene-finding; Yeast genome; 3D graphical representation of DNA sequence

1 Introduction

One of the most critical steps of genome annotation is the process of predicting genes that code for proteins. Generally, there are two algorithmic concepts appropriate to recognize genes [1]: 1) A sequence can be classified as a gene, if it shows significant similarity to a sequence, which was annotated as coding and deposited in a database. 2) A statistical analysis of a sequence may indicate its coding potential. This concept is based on the fact that the distributions of nucleotides in coding and non-coding sequences differ statistically significantly [2, 3].

The budding yeast *Saccharomyces cerevisiae* is an important model organism for the Human Genome Project. As the first sequenced genome of a eukaryotic organism, *S. cerevisiae*, much work has been done on this aspect. The number of protein coding genes in the yeast genome was estimated to be 5800-6000 [4-6], however, some researchers believe that the number should be less than 4800[7] or 5579[8]. But the prediction of protein coding genes is still far from being a trivial problem.

In this paper we present a simple gene-finding algorithm based on the 3D graphical representation of DNA sequence proposed in [9]. The algorithm utilizes an angle discriminant method to separate the object (ORFs) into two classes of positives (genes) and negatives (non coding ORFs). This simple gene-finding algorithm can perform quickly and it may be complementary with other existing methods.

2 Databases and methods

2.1 The database

In this paper, all the *S .cerevisiae* genome DNA sequences are taken from http://pedant.gsf.de/ of the Munich Information Center for Protein Sequences (MIPS) released on October 10, 2001. 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 3410 (18), 229, 820(2), 1003, 516, and 471(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 mitochondrial genetic code differs from the universal genetic code. So in each of the six classes, 3392, 229, 818,1003, 516, and 463 ORFs are contained, respectively.

2.2 The 3D graphical representation of DNA sequence

The three-dimensional graphical representation of DNA sequences provides a visual inspection of DNA data. Several researchers have proposed different graphical representations of DNA sequence [9,14-16].

Our gene-finding algorithm base on the3D

graphical representation of DNA sequence outlined recently in [9]. We present it briefly as follows. We assign A (adenine), G (guanine), T (thymine), and C (cytosine) to –x, +x, -y, and +y, respectively, while the corresponding curve extend along with z-axes. In detail, let $B = b_1 b_2 b_3 \cdots b_n$ be an arbitrary DNA sequence. Then we have a map Φ_1, which maps B into a plot set. Explicitly,

$$\Phi_1(B) = \Phi_1(b_1)\Phi_1(b_2)\Phi_1(b_3)\cdots\Phi_1(b_n)$$

$$\text{where } \Phi_1(b_i) = \begin{cases} (-1,0,i) & if\ b_i = A \\ (1,0,i) & if\ b_i = G \\ (0,-1,i) & if\ b_i = T \\ (0,1,i) & if\ b_i = C \end{cases}$$

Connecting adjacent points, we obtain a 3-D curve. In addition, we have another two maps Φ_2, Φ_3, where

$$\Phi_2(b_i) = \begin{cases} (-1,0,i) & if\ b_i = A \\ (1,0,i) & if\ b_i = T \\ (0,-1,i) & if\ b_i = G \\ (0,1,i) & if\ b_i = C \end{cases},$$

$$\Phi_3(b_i) = \begin{cases} (-1,0,i) & if\ b_i = A \\ (1,0,i) & if\ b_i = C \\ (0,-1,i) & if\ b_i = T \\ (0,1,i) & if\ b_i = G \end{cases}.$$

So, for one DNA sequence there are three curves that represent it.

2.3. The gene-finding algorithm

Based on two facts (1) amino acid are encoded by triplets of nucleotides of DNA and (2) each nucleotide base does not appear with equal probability at each codon position, comes a conclusion that both the four base (A,C,G, and T) and the three positions are likely to be related with the genetic code [10,11] .The curve for the subsequence in an ORF with bases at positions $1,4,7\cdots$, forms a phase-specific curve. We call this the phase-1 curve. Similarly, the curves with bases at positions $2,5,8\cdots$, and $3,6,9\cdots$, are called the phase-2 and phase-3 curve, respectively. For an ORF sequence, the phase-1, phase-2, and phase-3 curves describe the distributions of bases at first, second, and third codon positions, respectively. For each phase-specific subsequence, there are three maps Φ_1, Φ_2, Φ_3, as for the ordinary DNA sequence. The coordinates of the *ith* point of phase-j (j=1,2,3) curve under the map of Φ_k (k=1,2,3) are denoted by $(x_{i,j}^k, y_{i,j}^k, z_{i,j}^k)$.

We define

$$v_{1,j} = \frac{\sum_{i=1}^{N} x_{i,j}^1}{z_{N,j}^1}, v_{2,j} = \frac{\sum_{i=1}^{N} y_{i,j}^1}{z_{N,j}^1},$$

$$v_{3,j} = \frac{\sum_{i=1}^{N} x_{i,j}^2}{z_{N,j}^2}, v_{4,j} = \frac{\sum_{i=1}^{N} y_{i,j}^2}{z_{N,j}^2},$$

$$v_{5,j} = \frac{\sum_{i=1}^{N} x_{i,j}^3}{z_{N,j}^3}, v_{6,j} = \frac{\sum_{i=1}^{N} y_{i,j}^3}{z_{N,j}^3},$$

$$v_{7,j} = \bar{a}^2 + \bar{c}^2 + \bar{g}^2 + \bar{t}^2, \text{ where } \bar{a}, \bar{c}, \bar{g},$$

and $\bar{t}$ are the average occurrence frequencies of bases A, C, G, and T in the DNA subsequence studied. That is, $\bar{a} = A_N/N$, $\bar{c} = C_N/N$, $\bar{g} = G_N/N$, $\bar{t} = T_N/N$, where A_N, C_N, G_N, and T_N are the occurrence numbers of bases A, C, G, and T, respectively, in the subsequences, and N is the total length of the subsequence studied. The variable $v_{7,j}$ was found to be a useful statistical quantity for the analysis of DNA sequence [12].

So, for each phase-specific subsequence, there is a seven-dimensional vector $V_j = (v_{1,j}, v_{2,j}, \cdots, v_{7,j})$, which corresponds to it. We define a 21-dimensional vector $U=(u_1,u_2,u_3,\cdots,u_{21})$, where

$$\begin{cases} u_1 = v_{1,1} & u_2 = v_{2,1} & u_3 = v_{3,1} \\ u_4 = v_{4,1} & u_5 = v_{5,1} & u_6 = v_{6,1} \\ u_7 = v_{7,1} & u_8 = v_{1,2} & u_9 = v_{2,2} \\ u_{10} = v_{3,2} & u_{11} = v_{4,2} & u_{12} = v_{5,2} \\ u_{13} = v_{6,2} & u_{14} = v_{7,2} & u_{15} = v_{1,3} \\ u_{16} = v_{2,3} & u_{17} = v_{3,3} & u_{18} = v_{4,3} \\ u_{19} = v_{5,3} & u_{20} = v_{6,3} & u_{21} = v_{7,3} \end{cases} \quad (1)$$

Therefore, each of the coding ORFs or non-coding DNA sequences is represented by a 21-dimensional vector.

According to the ergodicity principle, we randomly divide the 3392 genes into two unequal parts, in which the larger part consists of 2000 genes, and the smaller consists of 1392 genes. The former serves as a training set; whereas the latter serves as a test set. Both the training and test sets should be accompanied by the counterparts of negative samples. Considering that the intergenic sequence with length longer than 300bp, which starts with ATG and ends with one of the stop codons, is

unlikely to be ORF[8,12], we randomly select about 7600 such intergenic sequences from the 16 yeast chromosomes to produce the negative samples. We randomly selected 2000 and 1392 intergenic sequences from the above 7600 sequences, which form the training and test sets of the negative samples, respectively.

The training set of samples (ORFs) is divided into two parts: one includes the positive samples composed of true protein coding genes, the other includes negative samples composed of non-coding DNA sequences. In the positive set the i-th true coding ORF is described by a vector $(u^1_{i,1}, u^1_{i,2}, \cdots, u^1_{i,21})$, where $u^1_{i,s}$ are the s-component of the vector (s=1,2, ⋯21). Similarly, in the negative set the i-th non-coding DNA sequences is described by a vector $(u^2_{i,1}, u^2_{i,2}, u^2_{i,3}, \cdots, u^2_{i,21})$, where $u^2_{i,s}$ are the s-component of the vector (s=1,2,⋯21). Suppose the positive and negative sets both include M samples, then we denote the geometric centers of theirs by $\overline{U^1}$ and $\overline{U^2}$, respectively, where $\overline{U^1} = (\overline{u^1_1}, \overline{u^1_2}, \cdots \overline{u^1_{21}}), \overline{U^2} = (\overline{u^2_1}, \overline{u^2_2}, \cdots, \overline{u^2_{21}})$ and $\overline{u^1_s} = \frac{1}{M}\sum_{i=1}^{M} u^1_{i,s}, \overline{u^2_s} = \frac{1}{M}\sum_{i=1}^{M} u^2_{i,s}$ (s=1,2,⋯21).

A query ORF is indicated by a 21-dimensional vector $U = (u_1, u_2, \cdots, u_{21})$. To judge whether this ORF is a true protein coding gene or not, calculate the angle $\langle U, \overline{U^1} \rangle$ between U and $\overline{U^1}$, and the angle $\langle U, \overline{U^2} \rangle$ between U and $\overline{U^2}$, where $\langle U, \overline{U^1} \rangle = \cos^{-1} \frac{(U, \overline{U^1})}{|U \| \overline{U^1}|}$, $\langle U, \overline{U^2} \rangle = \cos^{-1} \frac{(U, \overline{U^2})}{|U \| \overline{U^2}|}$. A codingness index Δ is defined as

$$\Delta = \langle U, \overline{U^2} \rangle - \langle U, \overline{U^1} \rangle + c \qquad (2)$$

where c is a constant determined by making 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 Criteria for the evaluation of the algorithm

For the evaluation of the performance of the algorithm, we have to discuss the definitions of sensitivity, specificity and selectivity. Denoted by TP the number of coding ORFs that have been correctly predicted as coding, and FN the number of coding ORFs that have been predicted as non-coding. Let TN denote the number of non-coding sequences that have been predicted as non-coding and FP denote the number of non-coding sequences that have been predicted as coding. Then we can define the following term:

$$S_p(specificity) = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{S_n}(sensitivity) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{S_l}(selectivity) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

That is, S_n is the proportion of coding ORFs that have been correctly predicted as coding, S_p is the proportion of non-coding sequences that have been correctly predicted as non-coding, and S_l is the fraction of correctly predicted positive cases among all cases predicted as positive.

The accuracy is defined as the average of S_n and S_p. The definition of accuracy is the same as in [8,12,13]: $AC = (S_n + S_p)/2$

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

Test set	1	2	3
Sensitivity(%)	0.974119	0.974108	0.976276
Specificity(%)	0.957585	0.950395	0.961179
Accuracy(%)	0.965852	0.9622515	0.968728

3.2 Self-consistency and cross-validation tests

To test the new algorithm, the resubstitution and cross-validation tests are performed. In the version of MIPS database, released on October 10, 2001, the ORFs were classified into six classes, in which the first class consists of 3410 entries corresponding to the known proteins. Excluding the protein coding genes from the mitochondria, 3392 protein genes of the first class residing at the 16 yeast chromosomes remain. The mitochondrial genes are excluded from the present study because the mitochondrial genetic code differs from the universal genetic code.

Using the sequences in the training set, the average vectors $\overline{U^1}, \overline{U^2}$ and the parameter c 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 3392 ORFs into two parts (2000 and 1392) is randomly. 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 test is always greater than 96%, which is higher than that reported in [8,12] and is comparable to that obtained in [13], however, this method is much faster than the method utilized in [13]. In table 1, the sensitivity, specificity and accuracy of each test are listed.

3.3 Apply the algorithm to recognize yeast genes

After performing the resubstitution and cross-validation tests, the 2000 and 1392 positive samples (true genes) are then merged. The 3392 negative samples are selected randomly from the 7600 intergenic sequences mentioned above. These 3392 positive and 3392 negative samples form a new training set. The vectors $\overline{U^1}, \overline{U^2}$, and the parameter c are obtained.

$\overline{U^1}$ = (0.174627, 0.081876, 0.037818, -0.054932, 0.276077, 0.119695, 0.136808, 0.130326, -0.133519, 0.209294, -0.054551, 0.282182, 0.075775, -0.078968, 0.087635, -0.131975, 0.097351, -0.122259, 0.271437, -0.034624, -0.009716),

$\overline{U^2}$ = (0.141616, -0.144077, 0.141884, -0.143808, 0.275982, -0.002193, -0.000269, 0.146585, -0.138505, 0.144128, -0.140963, 0.275980, 0.005623, 0.002458, 0.145024, -0.127824, 0.136119, -0.136730, 0.274484, 0.008294, 0.008905), c=0.068875

We then apply the vectors $\overline{U^1}, \overline{U^2}$, and c listed above to recognizing genes in the ORFs of the 2nd-6th classes in the MIPS database. For each ORF calculate the vector $U = (u_1, u_2, \cdots, u_{21})$, where $u_1, \cdots u_{21}$ are defined in Eq. (1). Based on the vectors $U, \overline{U^1}, \overline{U^2}$, and the parameter c, calculate the codingness index Δ using Eq. (2). If $\Delta > 0$, the query ORF is recognized as a coding gene, if $\Delta < 0$, the ORF or DNA sequence is recognized as a non-coding one. According to the MIPS database, there are 229, 818,1003, 516, and 463 entries of the 2nd-6th classes in the yeast genome. Consequently, there are 7, 49, 118, 113, 300 entries in the five classes that are recognized as non-coding ORFs. The detailed results are listed in Table 2-6.

Table 2. The 7 ORFs of the 2nd class (strong similarity to known protein) in the MIPS database, which are recognized as non-coding

ybr210w	ymr040w	yel004w	ylr046c	yar061w	yll051c	ypl141c

Table 3. The 49 ORFs of the 3rd class (similarity or weak similarity to known protein) in the MIPS database, which are recognized as non-coding

ydl199c	yfl040w	yhr130c	yil040w	yjr136c	ylr064w	ylr311c
ymr088c	yor053w	yor286w	ybl089w	ybr293w	ydr249c	yer097w
yfr057w	ygl160w	yhr181w	yjl091c	yjl193w	ylr365w	ymr221c
ymr306w	ynl109w	yol163w	yol079w	ycr001w	ydl206w	ydr119w
ydr307w	ydr413c	yel045c	yer113c	yll005c	ylr050c	ylr184w
ymr245w	yol107w	yor350c	ykl037w	yal066w	ydr319c	ygr101w
ykr030w	ylr283w	ydr115w	ydr366c	ygl104c	ygr284c	yil025c

Table 4. The 118 ORFs of the 4th class (similarity to unknown protein) in the MIPS database, which are recognized as non-coding

yar060c	ybr099c	ybr147w	ycr038w-a	ydl240c-a	ydr210w	yer079c-a
yer140w	ygl263w	yhl034w-a	yhl045w	yil090w	yir040c	yjr162c
ykl225w	ylr149c-a	ylr161w	ylr414c	yml007c-a	ymr010w	ynl156c
ynr077c	yol048c	ybl049w	ycl002c	ycl065w	ydr504c	yfr012w
ygl041c	ygr016w	yhr017w	ykl106c-a	ymr013w-a	ymr119w	yol002c
yol159c-a	yor044w	yor365c	ypl264c	ypr016w-a	yar068w	ybr103c-a
ydl027c	ydr084c	ydr438w	ydr492w	yfl015c	yfl062w	yfl068w
ygl010w	ygl084c	ygr293c	yhl041w	yhr069c-a	ykl219w	ykr051w

yll065w	ylr023c	yml047c	yml132w	ymr326c	ynl326c	yol003c
yol162w	ypr071w	yal018c	yal047w-a	ybl029c-a	ybl108w	ybr004c
ybr168w	ybr300c	ycr097w-a	ydl185c-a	ydl248w	ydr525w-a	yel053w-a
yhl042w	yhr212c	yil174w	yjl097w	ykl223w	ylr156w	ynl067w-a
ypr074w-a	ybl109w	ybr191w-a	ybr302c	ycr102w-a	ydl054c	ydl114w-a
ydl159w-a	ydr126w	ydr367w	yel033w	yel067c	ygl260w	ygr149w
ygr295c	yhl044w	yhr214w-a	yil029c	yil089w	yil175w	yir030w-a
yir044c	yjl052c-a	yjr013w	yjr044c	yjr161c	ykl165c-a	ylr036c
ylr159w	ynl336w	yol047c	yol101c	yor314w-a	ypl165c	

Table 5. The 113 ORFs of the 5th class (no similarity) in the MIPS database, which are recognized as non-coding.

yar047c	ycl056c	ydr042c	ydr524w-a	yel010w	yfl021c-a	yfr042w
ygr168c	ylr111w	ymr151w	ynl324w	yor248w	ypr170w-a	yar053w
ybl048w	ybr056w-a	ycl058c	ycr085w	ydl196w	ydr015c	ydr102c
ydr274c	ydr396w	yel014c	yel059w	yer135c	ygl188c	yhr139c-a
yjl077c	yjl215c	ykr032w	yll030c	ylr112w	yml084w	yml122c
ymr057c	ymr320w	yor029w	yor072w	yor314w	yor364w	ypr012w
ybl071c	ybr144c	ydr278c	ydr344c	ydr535c	yer066c-a	yer172c-a
ygr290w	yhl037c	yhr095w	yir020c-b	yjl028w	yjr157w	ykr073c
ylr122c	ylr366w	ylr400w	yml090w	ymr003w	ymr141c	ynl143c
ynl211c	yol160w	ypl056c	ypr014c	yal064w	ybr027c	ybr292c
ycr022c	ydr024w	ydr179w-a	ydr350c	yer091c-a	ygr026w	ygr291c
yil012w	yir020c	yjl136w-a	ykl158w	ylr124w	ylr264c-a	ymr254c
ynl146w	ynl174w	ynl303w	yor152c	ypl200w	ypr153w	yar030c
yar070c	ycl021w-a	ycr025c	ydr029w	yfl019c	yfr035c	ygl006w-a
yhl005c	yjr023c	ykl044w	yll059c	ylr381w	ylr404w	ymr187c
ynl150w	ynl179c	yor015w	yor268c	yor392w	ypl041c	ypr064w
ypr170c						

Table 6. The 296 ORFs of the 6th class (questionable ORFs) in the MIPS database, which are recognized as non-coding.

ybl012c	ybl073w	ybr090c	ybr124w	ybr266c	ycr018c-a	ycr087w
ydl026w	ydl062w	ydr034c-a	ydr112w	ydr154c	ydr203w	ydr269c
ydr355c	ydr431w	ydr467c	ydr526c	yer138w-a	yer181c	yfl032w
ygl024w	ygl118c	ygl168w	ygl204c	ygr039w	ygr069w	ygr122c-a
ygr176w	yhl006w-a	yhr125w	yil020c-a	yil066w-a	yjl086c	yjl150w
yjr018w	ykl030w	ykl136w	ylr101c	ylr198c	ylr322w	ylr428c
yml009c-a	yml047w-a	ymr075c-a	ymr316c-a	ynl205c	ynl276c	yor041c
yor121c	yor170w	yor225w	yor282w	ypl034w	ypl114w	ypr053c
ypr177c	q0143	yal056c-a	ybl053w	ybl077w	ybl107w-a	ybr224w
ycl023c	ycr041w	ydl009c	ydl032w	ydl068w	ydl172c	ydr034w-b
ydr114c	ydr157w	ydr360w	yer076w-a	yer107w-a	yer145c-a	yfr036w-a
ygl074c	ygl132w	ygl177w	ygr011w	ygr045c	ygr073c	ygr137w
ygr182c	ygr259c	yhl019w-a	yil029w-a	yil068w-a	yir023c-a	yjl022w
yjr087w	ykl036c	ykl115c	ykl147c	ykl202w	yll020c	ylr123c
ylr202c	ylr252w	ylr282c	ylr358c	ylr434c	yml116w-a	ymr086c-a
ymr158w-b	ymr290w-a	ynl089c	ynl170w	ynl226w	yol013w-b	yol099c
yor135c	yor199w	yor235w	yor300w	yor345c	ypl035c	ypl205c
ypr038w	ypr092w	ypr136c	yjr038c	ypr150w	yal026c-a	yal059c-a
ybl062w	ybr051w	ybr109w-a	ybr226c	ycl041c	ydl151c	ydl187c
ydr048c	ydr133c	ydr220c	ydr290w	ydr401w	ydr442w	ydr509w
yel009c-a	yer084w-a	yer148w-a	yfl012w-a	yfr052c-a	ygl042c	ygl088w

ygl149w	ygl182c	ygl217c	ygr018c	ygr107w	ygr139w	ygr265w
yhl030w-a	yhr063w-a	yhr145c	yil030w-a	yil071w-a	yil163c	yjl032w
yjl120w	yjl202c	ykl053w	ykl118w	ylr261c	ylr294c	ylr334c
ylr444c	yml012c-a	ymr119w-a	ymr172c-a	ynl013c	ynl105w	ynl171c
ynl228w	yol035c	yol106w	yor082c	yor200w	yor309c	ypl044c
ypl238c	ypr039w	ypr099c	ypr142c	ypr087w	ypr050c	yal031w-a
ybl065w	ybl094c	ybr064w	ybr178w	ydl016c	ydl152w	ydl221w
ydr053w	ydr136c	ydr230w	ydr445c	yel018c-a	yer046w-a	yer133w-a
yfl013w-a	yfr056c	ygl152c	ygl193c	ygl218w	ygr114c	ygr151c
yhl046w-a	yil047c-a	yil100c-a	yjl009w	yjl135w	yjl175w	yjr128w
ykl076c	ykr033c	ylr169w	ylr230w	ylr269c	ylr302c	ylr458w
yml094c-a	ymr046w-a	ymr304c-a	ynl028w	ynl114c	ynl235c	ynr005c
yol037c	yol134c	yor102w	yor146w	yor263c	ypl073c	ypr077c
ypr146c	q0092	yal034c-b	ybl070c	ybr089w	ybr116c	ycr064c
ydl050c	ydl094c	ydl158c	ydr008c	ydr149c	ydr199w	ydr241w
ydr426c	ydr455c	ydr521w	yel075w-a	yer067c-a	yer087c-a	yer137w-a
yer165c-a	ygl109w	ygl165c	ygr025w	ygr064w	ygr115c	ygr228w
yhl002c-a	yhr028w-a	yhr071c-a	yil060w	yil115w-a	yir017w-a	yjl015c
yjl142c	yjr071w	ykl083w	ykl131w	ylr171w	ylr232w	ylr317w
ylr339c	ymr052c-a	ymr153c-a	ymr193c-a	ymr306c-a	ynl120c	ynl198c
ynl266w	ynr025c	yol150c	yor169c	yor277c	yor331c	ypl102c
ypl185w	ypl261c					

Of the entries in above lists, statistically, FN (in list 7) are actually coding. Unfortunately, we cannot identify them at present due to the limited recognition accuracy achieved.

Based on the above result and the sensitivity and specificity, the four quantities TP, TN, FP, and FN can be calculated. Take the 5th class ORFs as an example. The total number of the 5th class ORFs is 516, in which 113 ones are recognized as non-coding. Assume that both the sensitivity and specificity are equal to 96%. We have a system of linear equations as follows:

$$\begin{cases} TP/(TP + FN) = 0.96 \\ TN/(TN + FP) = 0.96 \\ TN + FN = 113 \\ TP + FN + TN + FP = 516 \end{cases}$$

solving the above set of equations, we find $TP \approx 399$, $TN \approx 96$, $FP \approx 4$, and $FN \approx 17$.Therefore, the number of real coding ORFs of the 5th class equals to TP+FN=399+17=416. Similar calculations for the others are performed. Note that for the 2nd class, the above system has negative solutions: $TP \approx 222$, $TN \approx -2$, $FP \approx 0$, $FN \approx 9$. In this case, we prefer FN=7, TN=0. The results are listed in table 7.

Table 7 The numbers of predicted coding and non-coding ORFs of the 2nd-6th classes

	2	**3**	**4**	**5**	**6**
Total number of ORFs	229	818	1003	516	463
TP	222	769	882	399	155
TN	0	16	81	96	289
FP	0	0	3	4	12
FN	7	33	37	17	7
TP+FN	229	802	919	416	162
TN+FP	0	18	84	100	302

We estimate the number of protein coding genes in the 16 yeast chromosomes. The total number should be equal to 5920, the sum of the number of the 1st class and the number of those in the 2nd-6th classes recognized by the present method. Note that the accuracy is actually greater than 96%, so, this figure should be considered as an upper bound of the number of genes in the yeast genome. Assume that both the sensitivity and specificity are equal to 97%. We also have a system of linear equations. According to the solutions to these system of equations, we can estimate a lower bound of the number of genes in the yeast genome, which is 5891. The above estimate is based on error analysis, i.e. we have considered the false negative and false positive events in the prediction for each class. So it should be statistically reliable.

4 Conclusion

In this paper, a novel gene recognizing method based on a 3D graphical representation of DNA sequence is proposed. As a satisfied result, the successful rates by both self-consistency and cross-validation tests very high and the total number of genes estimated here is 5891～5920, coincident with 5800-6000, which is widely accepted. As should be pointed out, to extend the method to more complicated structures, we have not excluded intron-containing genes. The present work is based on an assumption that the unknown genes have the same statistical properties as the known genes. This might not be so in some special cases, for example, for some low-expressed genes. In this case, the results should be referred to with caution.

References:

[1] M. Tech and R. Merkl, YACOP: Enhanced gene prediction Obtained by a combination of existing methods, In Silico Biology 2003,3, 0037

[2] J. W. Fickett, Recognition of protein coding regions in DNA sequences, Nucleic Acids Res. 1982,10, 5303-5318

[3] Staden, R. Computer methods to aid the determination and analysis of DNA sequences. Biochem. Soc. Trans. 1984,12,1005-1008

[4] A. Goffeau, B. G. Barrel, H. Bussey, R. W. Davis, B. Dujon, H Feldmann, F Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettlin, and S. G. Oliver, Life with 6000 genes. Science 1996, 274, 546.

[5] E. A. Winzeler, and R.W. Davis, Functional analysis of the yeast genome, Curr. Opin. Genet. Dev. 1997, 7, 771-776

[6] H. W. Mewes, K. Albermann, M. Bahr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S. G. Oliver, F. Pfeiffer and A. Zollner, Overview of yeast Genome, Nature. 1997, 387, 7-8.

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

[8] C. T. Zhang, J. Wang and R. Zhang, Using a Euclid distance discriminant method to find protein coding genes in the yeast genome, Comput. Chem. 2002, 26,195-206

[9] C. Yuan, B. Liao, T. M. Wang, New 3-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 2003,379 ,412-417

[10] W. Li, P. Bernaola-Galvan, F. Haghighi and I. Grosse, Applications of recursive segmentation to the analysis of DNA sequence, Comput. Chem. 2002, 26, 491-510

[11] R. Staden, Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes, Nucleic Acids Res. 1984, 12, 551-567

[12] C. T. Zhang and J. Wang, Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve, Nucleic Acids Res. 2000, 28, 2804-2814

[13] C. Li, P. A. He and J. Wang, Artificial neural network method for predicting protein coding genes in the yeast genome, Internet Electron. J. Mol. Des. 2003, 2. 527-538

[14] M. Randic, M. Vracko, A. Nandy, S. C. Basak, On 3-D graphical representation of DNA primary sequence and their numerical characterization, J. Chem. Inf. Comput. Sci, 2000,40,1235-1244

[15] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem, 1983, 258, 1318-1327

[16] R. Zhang and C.T. Zhang, Z-curve, an intuitive tool for visualizing and analyzing the DNA sequences, J. Biomol. Str. Dyn, 1994, 11(4), 767-782