Finding Protein Coding Genes in the Yeast Genome Based on the Characteristic Sequences

Ping-an He[a,*] Chun Li[b] and Jun Wang[b]

T-Life Research Center, Fudan University, Shanghai 200433, P.R.China

Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, P.R. China

** Correspondence author: E-mail pinganhe@yahoo.com.cn*

Abstract: The characteristic sequences of a DNA sequence are a group of (0,1) sequences. Each of them is a reduced representation of the given DNA sequence, and two of them can uniquely reconstruct the sequence. Based on the numerical description of the characteristic sequences, a protein coding gene finding algorithm specific for the yeast genome at better 95% accuracy was suggested. Based on this, it is found that the total number of protein coding genes in the yeast genome is 5897, coincident with 5800-6000, which is widely accepted. The names of putative non-coding ORFs are listed here in detail.

INTRODUCTION

Most gene-finding algorithms are based on the differences of statistical properties between DNA sequences in coding and non-coding regions [1-7,13-21]. The phases in one strand of a DNA double helix are heterogeneous in the coding regions, whereas homogeneous in the non-coding regions. This fact constitutes the basis of almost all gene-finding algorithms [1,2]. The prediction of coding sequences has garnered a lot of attention during the last decade [1-7,13-21]. We can distinguish two kinds of methods, one relies on training with sets of example and counter-example sequences, and the other exploits the intrinsic properties of the DNA sequences to be analyzed.

Currently, the most popular approach is to consider a set of candidate exons weighted by some statistical parameters and then construct the optimal gene, defined as a consistent chain of exons using dynamic programming [3,4,5]. The recognition of coding sequences is usually approached by measuring the positional and compositional biases imposed by the genetic code on the DNA sequences in protein-coding regions [6]. Recent developments in the prediction of coding sequences require computation of discriminant functions with parameters that are estimated with a training set composed of examples and counter-examples (coding and non-coding sequences) [6, 7]. For example, Zhang1, et al. [1,2] suggested a gene finding algorithm based on the YZ score index. In their algorithm, a graphical approach was used to explore the difference between coding and non-coding sequences.

An ORF is a DNA stretches that potentially encode protein. They always have a start codon (ATG) at one end and a translation-terminating stop codon at the other end, and with at least 300bases in between. In Human DNA sequences, almost no ORFs representation an actual gene—the may contain pieces of a gene. As a result, locating ORFs in human genome will not accomplish much in terms of gene recognition. However, in bacterial DNA sequences, practically all ORFs are coding sequences, which make the gene recognition easy.

In a previous paper [8], the characteristic sequences were introduced to represent a DNA

sequence and make comparisons of the similarity and dissimilarity of DNA sequences [also see 17]. Based on the ideas of the characteristic sequences and the Euclid distance discriminant method, we propose, in this paper, an algorithm for the recognition of coding ORFs and non-coding ORFs sequences in the yeast *Saccharomyces cerevisiae* genome.

MATERIALS AND METHODS

The Database

The budding yeast *Saccharomyces cerevisiae* is an important model organism for the Human Genome Project. In this paper, we adopt the *S. cerevisiae* genome DNA sequences. The *S. cerevisiae* genome DNA sequences can be obtained from the Munich Information Center for Protein Sequences (MIPS), released in 1997[9, 11]. The data for classification of ORFs in the yeast genome were downloaded from http://mips.gsf.de, release, October 10, 2001. In the MIPS database, all the ORFs are classified into six classes, which correspond to known proteins, no similarity, questionable ORFs, similarity or weak similarity to known proteins, similarity to unknown proteins and strong similarity to known proteins, respectively. The 1st, 2nd, 3rd, 4th, 5th and 6th classes include ***3410(18), 516, 471(8), 820(2), 1003*** and ***229***, 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, **3392, 516, 463, 818, 1003** and **229** ORFs are contained, respectively.

The characteristic sequences and their numerical characterization

Mathematically, a homomorphism in algebra represents and emphasizes a partial mirror of an algebraic system. With this idea in the mind, we introduce the concept of characteristic sequences of a DNA sequence as follows.

According to their chemical structures, there are two ways to divide the four bases A, C, G, T into two classes: purine R={A,G} and pyrimidine Y={C, T}; amino group M={A, C } and keto group K={G, T}. Besides these, the division can also be made according to the strength of the hydrogen bond, i.e., weak H-bonds W={A, T} and strong H-bonds S={G, C}.

By the three divisions we reduce a DNA sequence into three (0,1) sequences, which is stated in mathematical terms as follows. Given a DNA sequence $G = a_1 a_2 a_3 \cdots$, we define three homomorphic maps ϕ_i, $i = 1, 2, 3$ by $\phi_i(G) = \phi_i(a_1)\phi_i(a_2)\cdots\cdots$, where

$$\phi_1 = \begin{cases} 1 & if \quad a_i \in R \\ 0 & if \quad a_i \in Y \end{cases} \quad \phi_2 = \begin{cases} 1 & if \quad a_i \in M \\ 0 & if \quad a_i \in K \end{cases} \text{ and } \phi_3 = \begin{cases} 1 & if \quad a_i \in W \\ 0 & if \quad a_i \in S \end{cases}$$

The $\phi_i(G)$, $i = 1, 2$ and 3, are called (R, Y)-, (M, K)-, and (W, S)-characteristic sequences, respectively.

Given a (0,1)-sequence $S = a_1 a_2 a_3 \cdots$, we define its normalized height function $h_s(p)$ (or $h(p)$ for short) to be p/q, which denotes the frequency of 1's occurring in the prefix of length p of *S*, that is, q is the number of 1's in $a_1 a_2 \ldots a_p$. Let *k* be a fixed positive integer. If *S* has length *n*, then we can divide it into *k* segments and consider their normalized height functions h([n/k]), h([2n/k]), ……, h([n]), where [n/k] denotes the biggest integer less than or equal to n/k.

From a DNA sequence and the above operation we construct its characteristic sequences. We obtain $h_R([in/k])$, $h_M([in/k])$ and $h_W([in/k])$, i=1,2,…, k, where R, M and W denote (R,Y)-, (M,K)- and (W,S)-characteristic sequences, respectively. By comparing these values, we can obtain some information of the DNA sequence.

The gene-finding algorithm

In this section, we suggest a gene-finding algorithm based on the different statistical properties at the three codon positions between protein coding ORFs and non-coding ones. The subsequence in an ORF with bases at positions 3i+1 (i=0,1,2 …) forms a phase-specific sequence, we call it the 1-subsequence. Similarly, we can also define 2-, 3-subsequence with bases at positions 3i+j, i=0,1,2 … and j=2 or 3 in the ORF.

For each phase-specific subsequence, regarded as an ordinary DNA sequence, there are three characteristic sequences. For each of them, taking k=2 and considering its normalized height function, we obtain a 6-dimensional real vector for the phase-specific subsequence. We denote the six components of the i-subsequence by R^1_{ni}, R^2_{ni}, M^1_{ni}, M^2_{ni}, W^1_{ni}, W^2_{ni}, i=1,2,3. Making a union of the three 6-dimensional vectors, we can describe each ORF (or an intergenic DNA sequence) by a point in a 18-dimensional real space.

To complete the algorithm in a computer, we need two groups of samples. Let *P* denote the group of the positive samples consisting of true protein coding genes, and *N* the group of negative samples composed of non-coding DNA sequences. The two groups of samples form the training set used in the protein coding gene-finding algorithm. Let *n* approximate the number of samples in each group. In the positive samples the k-th true coding ORF is described by a vector $(u^P_{k1}, u^P_{k2}, \ldots, u^P_{k18})^T$, where u^P_{ki}'s are the i-component of the vector (i=1, 2, …, 18), and ``T'' denotes the ordinary transpose operator of matrix. Similarly, in the negative samples the k-th non-coding DNA sequence is described by a vector $(u^N_{k1}, u^N_{k2}, \ldots, u^N_{k18})^T$.

We adopt the convention used by Zhang, et al.[1]. By $\bar{U}^P$ and $\bar{U}^N$ we denote the geometric centers of the positive and negative samples in the 18-dimensional space, where

$$\bar{U}^P = (\bar{u}^P_1, \bar{u}^P_2, \ldots, \bar{u}^P_{18})^T, \quad \bar{U}^N = (\bar{u}^N_1, \bar{u}^N_2, \ldots, \bar{u}^N_{18})^T \qquad (1)$$

and

$$\bar{u}^P_k = \frac{1}{n}\sum_{i=1}^{i=n} u^P_{ik}, \quad \bar{u}^N_k = \frac{1}{n}\sum_{i=1}^{i=n} u^N_{ik} \quad k=1,2, \ldots, 18. \qquad (2)$$

By an 18-dimensional vector $(u_1, u_2, \ldots, u_{18})^T$ we denote a query ORF. We calculate the Euclid distances $d(U, \bar{U}^P)$ between U and $\bar{U}^P$, and $d(U, \bar{U}^N)$ between U and $\bar{U}^N$ to judge whether or not this ORF is a true protein coding gene. Here

$$d(U, \bar{U}^P) = \left[\sum_{k=1}^{k=18} (u_k - \bar{u}^P_k)^2\right]^{1/2} \quad \text{and} \quad d(U, \bar{U}^N) = \left[\sum_{k=1}^{k=18} (u_k - \bar{u}^N_k)^2\right]^{1/2} \qquad (3)$$

A coding index Δ is defined as $\Delta=d(U, \bar{U}^P)-d(U, \bar{U}^N)+c$ (4), 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 a true protein coding gene, otherwise, the ORF or DNA sequence is recognized as a non-coding sequence.

EVALUATION AND APPLICATION

Definitions of sensitivity, specificity and accuracy

Sensitivity and specificity measures are widely used to characterize the accuracy of an algorithm or a recognition function. Here, we adopt the definitions and notations in Burset and Guigo [10].

Let TP denote 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. Then we define the sensitivity S_n as,

$S_n = TP/(TP+FN)$ (5).

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,

$S_p= TN/(TN+FP)$ (6).

That is, S_p is the proportion of intergenic sequences that have been correctly predicted as non-coding. In addition to, we define the accuracy T as the average of the sensitivity and specificity, that is

$T=1/2\ (S_n+S_p)$ (7).

Self-consistency and cross-validation tests

Usually, the re-substitution and cross-validation tests are efficient methods to evaluate the algorithm. The former reflects the self-consistency, and the latter reflects the extrapolating effectiveness of the algorithm. In the references [1, 2], the authors used the first class in the MIPS database, and regarded them as the positive samples. From the 16 yeast chromosomes, they randomly selected about 6000 intergenic sequences with length longer than 300 bp, starting with ATG and ending with one of the stop codons, and then, from the 6000 intergenic sequences, they randomly selected 2958 sequences as the negative samples and randomly divided each sample into two samples: training set and test set. Using them, their algorithms were evaluated.

Following Zhang's methodology, in this paper, we still use the MIPS database to evaluate our algorithm. The first class includes 3392 known genes in the 16 yeast chromosomes in the MIPS database. There are some differences between our data and that in Zhang's [1] paper. Data used in treatment was of more recent origin than that used in the Zhang's work.

In the MIPS database released in 2001, the first class included 3392 known genes. We randomly divide the 3392 genes into two parts, one of which includes 2000 genes and the other 1392 genes. The former is regarded as a training set and the latter is regarded as a test set. Using Zhang's [1] method, we randomly select 7691 intergenic sequences (non-coding sequence) from *S. cerevisiae* genome, and randomly select 2000 and 1392 sequences from the above 7691 sequences, which

form the training and test sets of the negative samples, respectively. In summary, the training set includes 2000 positive samples (true genes) and 2000 negative samples (intergenic sequences), and the test set includes 1392 positive samples (true genes) and 1392 negative samples (intergenic sequences).

Using the sequences in the training set, the average vectors $\bar{U}^P$, $\bar{U}^N$ and the parameter c (see Eq. (2) and (4)) are determined. Using these quantities, the accuracy of the gene-finding algorithm in the training and test sets is calculated. Repeating the above random division procedure six times, we perform six re-substitution and cross-validation tests. The results of the cross-validation tests are listed in **Table 1**. As we will see from **Table 1**, the accuracy in each cross-validation test is always greater than 95%.

Table 1 The accuracy of the algorithm for three different tests

	Test1	Test2	Test3	Test4	Test5	Test6
Sensitivity(%)	95.9	94.6	96.6	95.9	95.7	94.4
Specificity(%)	94.8	95.8	94.3	95.0	95.5	96.4
Accuracy(%)	95.35	95.2	95.45	95.45	95.6	95.4

Application of the algorithm to find genes in the ORFs of the 2nd-6th classes

In this section, we recognize genes in the ORFs of the 2nd-6th classes in the MIPS database using the algorithm.

Firstly, we merge the training set and test set of the positive samples into a new training positive set, and randomly select 3392 sequences from the 7691 intergenic sequences as mentioned above to form a new training negative set. In order to counter the particularity of the selected samples, we repeat this process ten times, and every time we calculate the average vectors $\bar{U}_i^P$, $\bar{U}_i^N$ and the parameter c_i, so we obtain ten triples ($\bar{U}_i^P$, $\bar{U}_i^N$, c_i) i=1,2,…,10.

Secondly, by taking the average of the ten triples we obtain a new triple as follows:

U^P=(0.62111, 0.62825, 0.54748 0.54638, 0.49741,0.49147, 0.48988, 0.49839, 0.62634, 0.63190, 0.57953, 0.57735, 0.47751, 0.47784, 0.60762, 0.60980, 0.48249, 0.48755), (11)

U^N=(0.50238,0.49925, 0.64094, 0.64316, 0.50307, 0.49982, 0.50059, 0.50398, 0.64064, 0.64235, 0.49962, 0.50252, 0.50898,0.50913, 0.63127, 0.63606, 0.49709, 0.50002), (12)

and c=0.015360 (13)

Thirdly, we judge each sequence in the ORFs of the 2nd-6th classes in the MIPS database based on the vectors U^P, U^N and the parameter c listed in (11), (12) and (13), respectively. For each ORF, we calculate the vector $U=(u_1, u_2, \ldots, u_{18})^T$, where u_i are defined in (5). Based on the vectors U, U^P, U^N and the parameter c, we calculate each coding-ness index Δ using (7). If $\Delta>0$, the query ORF is recognized as a coding gene, otherwise, non-coding. In each class, the ORFs recognized as non-coding ORFs are listed in **Tables 2-6** corresponding to the 2nd-6th classes in the yeast genome, respectively.

Furthermore, we re-estimate the number of protein coding genes in the 16 yeast chromosomes based on the above results. For example, the total number of the 2nd class ORFs is 516, in which 126 are recognized as non-coding. Suppose both the sensitivity and specificity of our algorithm are 95%, we can obtain a system of four linear equations as follows:

$$\begin{cases} TP/(TP+FN) = 0.95 \\ TN/(TN+FP) = 0.95 \\ TN+FN = 126 \\ TP+FN+TN+FP = 516 \end{cases},$$

from which we obtain that FP ≈ 6, FN ≈20, TP ≈ 384, TN ≈106. The number of the real coding sequences of the 2nd class should be equal to TP+FN=384+20=404. For the 3rd-6th classes, we can treat them in the same way. For the 6th-class, however, the above system has negative solutions. The reason is that the number recognized as non-coding sequences is too small, which is only 5. In this case, taking FP=FN=0, we have TP=224 and TN=5. Then, we list the values of TP, FP, TN, and FN in the 2nd-6th class ORFs in **Table 7**.

Thus, the total number of protein coding genes should be equal to 5897, the sum of the number of the 1st class (3410) and the number of those in the 2nd-6th classes recognized by the present algorithm (3410+404+159+797+903+224, see **Table 7**). Note that the accuracy is actually greater than 95%, so, this sum should be an upper bound of the number of the genes in the yeast genome. The above estimate of protein coding genes in the yeast genome is coincident with 5800-6000, which is widely accepted [9,11,12]. 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.

CONCLUDING REMARKS

In this paper, we propose an algorithm for distinguish coding ORFs and non-coding ORFs in the yeast genome. For complete the algorithm, we take the first class ORFs (known protein) as coding gene sequences and intergenic DNA sequence as non-coding sequences. Using them, we distinguish coding ORFs and non-coding ORFs for 2nd-6th classes ORFs in the yeast genome and obtain the number of coding ORFs in the 2^{nd}-6^{th} classes are at most 404,159, 797, 903 and 224, respectively. As a result, the total number of coding ORFs is estimated to be less than to 5897 in the 16 yeast chromosomes. Besides, we can also observe that the percentage of non-coding ORFs is 17.9% in 2^{nd}-6^{th} classes from **Table 7**, that is most ORFs are indeed genes. However, the percentages in the 2nd and 3rd classes are higher than others, 21.7% and 65.7%, respectively. According to classification of ORFs in the MIPS database, some of these ORFs neither their function nor homology is known. So, their high percentage is no wonder. With the increase in known genes, the number and percentage should be decrease.

As we mentioned, the idea of characteristic sequences comes from algebra, which is a kind of reduced representation for a complicated objects. This idea is applied not only to DNA sequences, but also to protein sequences and others. In practice, we can also concentrate on a single characteristic sequence. For example, in gene-finding algorithm of this paper, we can replace the 18-dimensional real space by a 6-dimensional real space: R^1_{ni}, R^1_{ni}, i=1,2,3, according to the purine-pyrimidine classification. Using the 6-dimensional space, we can perform the same algorithm on the yeast genome to research the biological function of purine-pyrimidine. Similarly, we can also take M^1_{ni}, M^2_{ni} or W^1_{ni}, W^2_{ni}, i=1,2,3, to research the biological functions of amino-keto groups and weak-strong H-bonds. This might provide a possibility to reveal the biological functions of purine-pyrimidine, amino-keto groups and weak-strong H-bonds, respectively.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China and Shanghai Postdoctoral Science Foundation.

REFERENCES AND NOTES

[1] Zhang,C. T.; Wang, J. *Nucleic Acids Res.* **2000,** *28*, 2804-2814.
[2] Zhang, C. T.; Wang, J.; Zhang, R. *Computers & Chem.* **2002,** *26,* 195-206.
[3] Guigo, R. *J. Comput. biol.* **1998,** *5,* 681-702.
[4] Guigo, R. *Bishop M. J. (ed.),* **1999,** 54-80. Lodon: Academic press.
[5] Roytberg, M. A.; Astakhova, T. V.; Gelfand, M. S. *Computers & Chem.* **1997,** *21,* 229-235.
[6] Quentin, Y.; Voiblet, C.; Martin, F.; Fichant, G. *Computers & Chem.* **1999,** *23,* 209-217.
[7] Guigo, R. Computers Chem. **1997,** *21,* 215-222.
[8] He, P. A.; Wang, J. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 1080-1085.
[9] 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. *Science.* **1996,** *274*, 546.
[10] Burset, M.; Guigo, R. *Genomics.* **1996,** *34,* 353-367.
[11] 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. *Nature (Suppl.).* **1997,** *387*, 7-8.
[12] Winzeler, E. A.; Davis, R. W. *Curr. Opin. Genet. Dev.* **1997,** *7,* 771-776.
[13] Chiusano, M. L.; Alvarez-Valin, F.; Giulio, M. D.; D'Onofrio, G.; Ammirato, G.; Colonna, G.; Bernardi, G. *Gene.* **2000,** *261*, 63-69.
[14] Fickett, J. W. *Trends Genet.* **1996,** *12,* 316-320.
[15] Gelfand, M. S. J. Computational Biol. **1995,** *2,* 87-115.
[16] Mackiewicz, P.; Kowalczuk, M.; Gierlik, A.; Dudek, M. R.; Cebrat, S. *Nucleic Acids Res.* **1999,** *27,* 3503-3509.
[17] Buldyrev, S. V.; Goldberger, A. L.; Havlin, S.; Mantegna, R. N.; Matsa, M. E.; Peng, C. K.; Simons, M.; Stanley, H. E. *Phys. Rev. E.* **1995,** *51*(5), 5084-5091.
[18] Salamov, A.; Solovyev, V. *Genome Research.* **2000,** *10*, 516-522.
[19] Shepherd,J. C. W. *Proc. Natl. Acad. Sci. USA.* **1981,** *78*, 1596-1600.
[20] Siemion, I. Z.; Siemion, P. J. *Biosystems.* **1994,** *33,* 39-48.
[21] Solovyev, V. V. *BioSystems.* **1993,** *30*, 137-160.
[22] Thomas, A.; Skolnick, M. *IMAJ. Math. Appl. Med. Biol.* **1994,** *11*, 149-160.
[23] Zhang, M. Q. Proc. *Natl. Acad. Sci. USA.* **1997,** 94, 565-568.

Table 2 The 126 ORFs of the 2nd class (no similarity)in the MIPS database, which are recognized as non-coding

yal037c-a yal064w yar030c yar047c yar053w yar070c ybl048w ybl071c
ybr027c ybr056w-a ybr209w ybr292c ycl056c ycl058c ycr022c ycr025c
ycr085w ydl176w ydl196w ydr015c ydr024w ydr029w ydr042c ydr065w
ydr102c ydr179w-a ydr274c ydr278c ydr344c ydr350c ydr396w ydr524w-a
ydr535c yel010w yel014c yel059w yer066c-a yer091c-a yer135c yer172c-a
yfl019c yfl021c-a yfr042w ygl006w-a ygl138c ygl188c ygr026w ygr168c

ygr226c ygr290w ygr291c yhl005c yhl037c yhr078w yhr095w yhr139c-a
yhr173c yil012w yil027c yil071c yir020c yir020c-b yjl027c yjl028w
yjl064w yjl077c yjl136w-a yjl215c yjr023c yjr157w ykl044w ykl158w
ykl162c ykr032w ykr073c yll007c yll030c yll059c ylr111w ylr112w
ylr122c ylr124w ylr145w ylr264c-a ylr265c ylr366w ylr381w ylr400w
ylr404w yml084w yml090w ymr003w ymr057c ymr082c ymr141c ymr148w
ymr151w ymr163c ymr187c ymr252c ymr254c ymr320w ynl122c ynl143c
ynl146w ynl150w ynl174w ynl179c ynl211c ynl303w ynl324w yol159c
yol160w yor024w yor029w yor097c yor152c yor248w yor255w yor364w
yor392w ypl041c ypl200w ypr012w ypr153w ypr170w-a

Table 3 The 297 ORFs of the 3rd class (questionable ORFs) in the MIPS database, which are recognized as non-coding

yal026c-a yal031w-a yal059c-a ybl053w ybl062w ybl065w ybl070c
ybl073w ybl077w ybl094c ybl107w-a ybr051w ybr064w ybr089w
ybr090c ybr109w-a ybr116c ybr178w ybr206w ybr224w ybr226c
ybr266c ybr277c ycl041c ycr018c-a ycr041w ycr064c ycr087w
ydl009c ydl016c ydl026w ydl032w ydl050c ydl062w ydl068w
ydl094c ydl151c ydl152w ydl158c ydl172c ydl187c ydl221w
ydr008c ydr034c-a ydr048c ydr053w ydr112w ydr114c ydr133c
ydr136c ydr149c ydr154c ydr157w ydr199w ydr203w ydr220c
ydr230w ydr241w ydr269c ydr271c ydr290w ydr355c ydr360w
ydr401w ydr417c ydr426c ydr431w ydr445c ydr467c ydr509w
ydr521w ydr526c yel009c-a yel018c-a yel075w-a yer046w-a yer067c-a
yer076w-a yer084w yer084w-a yer087c-a yer133w-a yer137w-a yer138w-a
yer145c-a yer148w-a yer165c-a yer181c yfl012w-a yfl013w-a yfl015w-a
yfl032w yfr036w-a yfr052c-a yfr056c ygl024w ygl042c ygl052w
ygl072c ygl074c ygl088w ygl109w ygl118c ygl132w ygl149w
ygl152c ygl165c ygl168w ygl177w ygl182c ygl193c ygl204c
ygl214w ygl217c ygl218w ygr011w ygr018c ygr039w ygr050c
ygr051c ygr069w ygr073c ygr107w ygr114c ygr115c ygr122c-a
ygr139w ygr151c ygr176w ygr182c ygr219w ygr228w ygr259c
ygr265w yhl002c-a yhl006w-a yhl019w-a yhl030w-a yhl046w-a yhr028w-a
yhr049c-a yhr063w-a yhr071c-a yhr125w yhr145c yhr193c-a yil020c-a
yil029w-a yil030w-a yil047c-a yil060w yil066w-a yil068w-a yil071w-a
yil100c-a yil163c yir017w-a yir023c-a yjl009w yjl015c yjl022w
yjl032w yjl075c yjl086c yjl120w yjl135w yjl142c yjl150w
yjl175w yjl182c yjl202c yjr018w yjr038c yjr071w yjr087w
ykl030w ykl036c ykl053w ykl076c ykl083w ykl115c ykl118w
ykl131w ykl136w ykl147c ykl202w ykr033c ykr047w yll020c
ylr101c ylr123c ylr140w ylr169w ylr171w ylr198c ylr202c
ylr230w ylr252w ylr261c ylr269c ylr279w ylr282c ylr294c
ylr302c ylr317w ylr322w ylr334c ylr358c ylr428c ylr434c
ylr444c ylr458w ylr465c yml009c-a yml012c-a yml047w-a yml094c-a

yml116w-a ymr046w-a ymr052c-a ymr075c-a ymr086c-a ymr135w-a ymr153c-a
ymr158c-a ymr158w-b 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 ynl319w ynr005c ynr025c yol013w-b
yol035c yol099c yol134c yol150c yor041c yor082c yor102w
yor121c yor146w yor169c yor170w yor199w yor200w yor225w
yor235w yor263c yor277c yor282w yor300w yor309c yor325w
yor331c yor345c yor379c ypl034w ypl035c ypl044c ypl073c
ypl102c ypl114w ypl185w ypl205c ypl238c ypl261c ypr039w
ypr050c ypr053c ypr077c ypr087w ypr099c ypr136c ypr142c
ypr146c ypr150w ypr177c

Table 4 The 60 ORFs of the 4th class (similarity or weak similarity to known proteins)in the MIPS database, which are recognized as non-coding

yal066w ybl089w ybr293w ycr001w ydl073w ydl119c ydl199c ydl206w
ydr100w ydr115w ydr205w ydr249c ydr307w ydr319c ydr366c ydr413c
ydr524c yel045c yer097w yfl040w yfr057w ygl104c ygl160w ygr101w
ygr284c yhl035c yhr035w yhr130c yhr181w yil025c yil040w yil088c
yjl091c yjl170c yjl193w ykr030w ykr103w yll005c yll037w ylr050c
ylr064w ylr184w ylr283w ylr311c ylr365w yml023c ymr088c ymr245w
ymr306w ynl109w ynl176c ynr059w yol079w yol107w yol152w yol163w
yor053w yor080w yor286w yor350c

Table 5 The 140 ORFs of the 5th class (similarity to unknown proteins)in the MIPS database, which are recognized as non-coding

yal018c yar029w yar060c yar068w ybl029c-a ybl049w ybl108w ybl109w
ybr004c ybr096w ybr099c ybr103c-a ybr147w ybr168w ybr191w-a ybr300c
ybr302c ycl002c ycl005w ycl065w ycr038w-a ycr097w-a ycr102w-a ydl027c
ydl054c ydl089w ydl114w-a ydl123w ydl159w-a ydl185c-a ydl240c-a ydl247w-a
ydl248w ydr018c ydr066c ydr084c ydr105c ydr126w ydr131c ydr210w
ydr275w ydr367w ydr437w ydr438w ydr459c ydr492w ydr504c ydr525w-a
yel033w yel053w-ayel067c yer074w-a yer079c-a yer140w yfl015c yfl062w
yfl068w yfr012w ygl010w ygl041c ygl084c ygl260w ygl263w ygr004w
ygr016w ygr149w ygr295c yhl034w-a yhl041w yhl042w yhl044w yhl045w
yhr067w yhr069c-a yhr212c yhr214w-a yil029c yil089w yil090w yil174w
yil175w yir030w-a yir040c yjl003w yjl052c-a yjl097w yjr013w yjr044c
yjr054w yjr161c yjr162c ykl018c-a ykl106c-a ykl165c-a ykl219w ykl223w
ykl225w ykr051w ykr106w yll065w ylr036c ylr047c ylr149c-a ylr368w
ylr408c ylr463c yml007c-a yml047c yml132w ymr010w ymr013w-a ymr071c
ymr119w ymr326c ynl008c ynl067w-a ynl162w-a ynl326c ynl336w ynr061c
ynr062c yol002c yol003c yol047c yol048c yol101c yol159c-a yol162w

yor044w yor147w yor175c yor365c ypl162c ypl165c ypl246c ypl264c ypr016w-a ypr071w ypr074w-a ypr114w

Table 6 The 5 ORFs of the 6th class (strong similarity to known proteins)in the MIPS database, which are recognized as non-coding

ybr210w yel004w yll051c ylr046c ymr040w

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

Class	2	3	4	5	6	Total
Total number of ORFs	516	463	818	1003	229	3029
TP	384	151	757	858	224	2374
FN	20	8	40	45	0	113
TN	106	289	20	95	5	515
FP	6	15	1	5	0	27
Total number of coding ORFs	404	159	797	903	224	2487
Total number of noncoding ORFs	112	304	21	100	5	542
Percentage of noncoding ORFs	21.7%	65.7%	2.6%	10%	2.2%	17.9%