**Inter*net* Electronic Journal of Molecular Design**

August 2003, Volume 2, Number 8, Pages 527–538

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday
Part 2

Guest Editors: Douglas J. Klein and Sonja Nikolić

# Artificial Neural Network Method for Predicting Protein Coding Genes in the Yeast Genome

Chun Li,[1] Ping–an He,[1] and Jun Wang[1,2]

[1] Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, P. R. China
[2] College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, P. R. China

Received: March 14, 2003; Revised: May 12, 2003; Accepted: May 19, 2003; Published: August 31, 2003

**Citation of the article:**
C. Li, P. 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, http://www.biochempress.com.

**Inter*net* Electronic Journal of Molecular Design**

**BioChem Press**
http://www.biochempress.com

# Artificial Neural Network Method for Predicting Protein Coding Genes in the Yeast Genome[#]

Chun Li,[1,*] Ping–an He,[1] and Jun Wang[1,2]

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

[2] College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, P. R. China

Received: March 14, 2003; Revised: May 12, 2003; Accepted: May 19, 2003; Published: August 31, 2003

*Internet Electron. J. Mol. Des*. **2003**, *2* (8), 527–538

**Abstract**

**Motivation.** The rapid growth of DNA sequences data in various DNA databanks has made analyzing these sequences, especially, finding genes in them very important, and it is even a more critical task at present to clarify the number of genes. The motivation of this paper is to suggest an artificial neural network method specific for predicting protein–coding genes in the yeast genome.

**Method.** We first obtain a 12–dimensional vector from a DNA primary sequence, and then construct a 12×21×1 three–layer feedforward neural network. After being trained in a supervised manner with the error back–propagation algorithm by sufficient samples, the network is examined by the cross–validation test.

**Results.** As a result, the average absolute error $\delta$ and the average variance $\sigma^2$ are 0.0084 and 0.0077, respectively, and the accuracy of the prediction is better than 96%. Based on this, it was found that the numbers of coding ORFs in the 2nd–6th classes are 393, 189, 803, 924 and 229, respectively. Thus, the total number of protein coding genes in the yeast genome is equal to 5930 coincident with a widely accepted range 5800–6000. The names of putative non–coding ORFs are listed in detail.

**Conclusions.** The results imply that the current artificial neural network method is a useful computer technique for predicting protein–coding genes, and can be extended to find genes with more complicated structures.

**Keywords.** DNA sequence; neural network; gene prediction; gene recognition; Yeast genome.

## 1 INTRODUCTION

With the development of biotechnologies, the analysis of sequences, especially, gene finding become more and more important in bioinformatics. Essentially, there are two different gene prediction methods [1]. One is the signal sensor, by which we detect the presence of functional sites specific to a gene [2–7], such as splicing sites, poly (A) sites (in 3'–UTRs), promoters and start/stop codon, etc. The other is the content sensor, a measure to classify DNA regions by a sufficient

[#] Dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday.

* Correspondence author; E–mail: lchlmb@yahoo.com.cn.

similarity [8–13], *e.g*. coding versus non–coding. Although much work has been done on this aspect, the prediction of protein coding genes is still far from being a trivial problem, and it is even a more critical task at present to clarify the number of genes [10].

The budding yeast (*Saccharomyces cerevisiae*) is an important model organism for the Human Genome Project. The number of protein coding genes in the yeast genome is estimated to be 5800–6000 [14–16], which is widely accepted at present. However, some researchers believe that the number should be less than 4800 [17]. The two results are obviously controversial.

In this paper, we first obtain a 12–dimensional vector from a DNA primary sequence. Then, based on the idea that the unknown genes have similar statistical properties to the known ones [9,10], we apply a multilayer feedforward artificial neural network (MLF ANN) method to predict protein–coding genes in the yeast genome, in which the network is trained in a supervised manner with the error back–propagation (BP) algorithm. As a satisfied result, the successful rates by both self–consistency and cross–validation tests are very high and the total number of genes estimated here is 5930, exactly coincident with 5800–6000.

## 2 MATERIALS AND METHODS

### 2.1 The Database

In this paper, all the *S. cerevisiae* genome DNA primary sequences are taken from http://mips.gsf.de of MIPS (the Munich Information Center for Protein Sequences) released on 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.

### 2.2 The 12–Dimensional Vectors

Based on two facts: (1) amino acids are encoded by triplets of nucleotides of DNA (codons) and (2) each nucleotide base does not appear with equal probability at each codon position, comes a conclusion that both the four bases (A, C, G and T) and the three codon positions are likely to be related with the genetic code [6,18,19]. By denoting the bases A, C, G and T at the 1st, 2nd and 3rd codon positions in an ORF with $A_i$, $C_i$, $G_i$ and $T_i$, respectively, where $i$ = 1, 2, 3, we write a DNA primary sequence as a sequence over $\Omega_{12} = \{A_1, A_2, A_3, C_1, C_2, C_3, G_1, G_2, G_3, T_1, T_2, T_3\}$

naturally. For example, we write a fragment ATGTCACTC… as $A_1T_2G_3T_1C_2A_3C_1T_2C_3$…. Then, based on the 12–symbol sequence, we obtain a 12–dimensional vector $v = \{f_1^A, f_2^A, f_3^A, f_1^C, \cdots, f_3^T\}$, where $f_i^X$ ($i$ = 1, 2, 3, $X \in \Omega = \{A, C, G, T\}$) represents the occurrence frequency of the symbol $X_i$ in the 12–symbol sequence. For convenience, the vector is rewritten by $v = \{f_1, f_2, \cdots, f_{12}\}$. Since the sum $\sum_{j=1}^{12} f_j$ is 1, it is enough to compute only 11 such $f_j$'s. In addition, the redundancy γ (defined later) reflects the difference of hereditary capacity between the coding ORF and non–coding DNA sequence, and was found to be a useful statistical quantity for the analysis of DNA sequences. Taking into account of these two aspects above, we replace the first component of ν by the redundancy γ, thus get a new 12–dimensional vector written as $v' = \{\gamma, f_2, \cdots, f_{12}\}$, where γ is given as follows [20,21]:

$$H_0 = \ln 12,\ H = -\sum_{j=1}^{12} f_j \ln f_j,\ \gamma = 1 - H/H_0, \quad (1)$$

where $H_0$ is the maximum entropy of all possible 12–symbol sequences and $H$ is the entropy of the 12–symbol sequence considered.

As an example, we take the sequence "YHR*099*W" (from yeast chromosome 8, positions 302763–313997) of the 1st class in the MIPS database, its 12–dimensional vector is:

$v'$ = (0.020276, 0.116364, 0.108529, 0.061610, 0.067931, 0.058583, 0.083244, 0.035791, 0.060719, 0.078170, 0.113248, 0.105502).

## 2.3 The MLF ANN Method

With its many features and capabilities for recognition, generalization and classification, artificial neural network technology has been applied successfully in bioinformatics [7,22–27]. Generally speaking, a neural network is characterized by (*a*) its pattern of connections between the neurons, *i.e.* its architecture, (*b*) its method of determining the weights on the connections, *i.e.* its training, or learning, algorithm, and (*c*) its activation function [22,28,29].

In this paper, we use the MLF ANN to predict protein–coding genes in the yeast genome. Typically, this network consists of a set of sensory units (source nodes) that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computation nodes. The three–layer feedforward neural network is shown in Figure 1. The depicted network has an *I*×*J*×*K* architecture (with *I* input units, *J* hidden units and *K* output units), which has two bias units, both having an input signal of 1 (*i.e.* $x_0$ and $z_0$), for the input and hidden layers, respectively. There are two layers of adaptive weights, and the term $w_{ji}$ ($w_{kj}$) is the weight of the *j*–th hidden unit (*k*–th output unit) associated with the input (hidden) signal $x_i$ ($z_j$). In this paper, since the recognition of protein coding genes is a 2–class classification problem, a single output neuron is enough, that is, *K*

= 1 [28], and the weight $w_{kj}$ can be denoted by $w_j$ simply. In addition, considering that the vector obtained from a DNA primary sequence is 12–dimensional, we naturally design a 12×21×1 three–layer feedforward ANN.

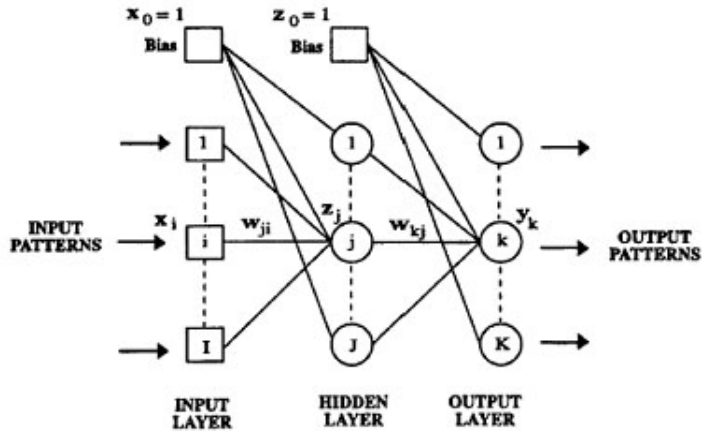


**Figure 1.** A three-layer feedforward neural network.

The activation function we used here is a sigmoidal nonlinearity defined by the logistic function [28,29]:

$$f(u) = \frac{1}{1+\exp(-u)} \tag{2}$$

According to Eq. (2), the output of each layer unit can be calculated by:

$$\begin{aligned} z_j &= f(\sum_{i=0}^{12} w_{ji} x_i)\ , \quad \mathrm{j} = 1, \cdots, 21 \\ y &= f(\sum_{j=0}^{21} w_j z_j)\ . \end{aligned} \tag{3}$$

Obviously, *y*, the output of the output unit, is exactly the actual output of a sample.

To optimize the weights of the network, the back–propagation algorithm is used in the three–layer feedforward ANN. For a given training set, corresponding to the supervised learning, we denote its *n*–th sequence by the sample pair $(v'_n, O_n)$, where $v'_n \in R^{12}$ is as pointed in Section 2.2, $O_n \in \{0,1\}$ $(n = 1,\ 2,\ \cdots,\ N)$. Here 0 and 1 are used to stand respectively for the two classes: the non–coding and coding sequences. The algorithm is formulated as follows:

Step 1: Initialize the weights and learning rate of the network and $n = 1$.

Step 2: Present the sample *n*.

Step 3: Compute outputs by Eq. (3).

Step 4: Update weights.

$$w_j(t+1) = w_j(t) - \eta(t) \times d \times z_j$$

$$w_{ji}(t+1) = w_{ji}(t) - \eta(t) \times d_j \times x_i$$

where $\eta(t)$ is the learning–rate that decreases automatically as the number of training iterations increases; $d = (y - O_n) \times f'(\sum_{j=0}^{21} w_j(t) z_j)$, and $d_j = w_j \times d \times f'(\sum_{i=0}^{12} w_{ji}(t) x_i)$.

Step 5: If $n < N$, $n = n + 1$, go to step 2.

Step 6: Go to step 7 if the absolute value of each weight adjustment for every sample n is smaller than a specified value $\varepsilon$ (*e.g.* 0.0001); Otherwise, $n = 1$, go to step 2.

Step 7: End.

# 3 RESULTS AND DISCUSSION

## 3.1 Self–consistency and Cross–validation Tests

We examine the results by a self–consistency test and a cross–validation test. In the self–consistency test, two criterion parameters, the average absolute error δ and the variance $\sigma^2$, are introduced to evaluate the training quality:

$$\delta = \frac{1}{N} \sum_{n=1}^{N} | O_n - y_n | \ , \ \sigma^2 = \frac{1}{N} \sum_{n=1}^{N} (\delta - | O_n - y_n |)^2 \ , \tag{4}$$

where $O_n$ and $y_n$ are the expected output and the actual output for the $n$–th sample, respectively.

In the cross–validation test, we discuss the accuracy, sensitivity and specificity, which are widely used to evaluate the performance of an algorithm. The notations used here are the same as that in [3,9,10,30]. Let *TP* (*FN*) denote the number of coding ORFs that have been predicted as coding (non–coding), and *TN* (*FP*) the number of non–coding sequences that have been predicted as non–coding (coding). The three parameters, $S_n$ (sensitivity), $S_p$ (specificity) and $\widetilde{S}$ denoting the accuracy of the prediction, are given by

$$S_n = \frac{TP}{TP + FN} \ , \ S_p = \frac{TN}{TN + FP} , \ \widetilde{S} = \frac{1}{2}(S_n + S_p) . \tag{5}$$

Namely, $S_n$ 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 $\widetilde{S}$ is the average of the sensitivity and specificity. The definition of $S_p$ in Eq. (5) may cause problems in recognizing genes along the genomic DNA sequence. Because the frequency of non–coding nucleotides is generally much greater than that of coding ones (both in reality and in the predictions), *TN* >> *FP*, and therefore $S_p$ tends towards 1. To solve this problem, instead of using

the definition of $S_p$ in Eq. (5), one used the refined definition [30]:

$$S'_p = \frac{TP}{TP + FP} \tag{6}$$

However, in this study, the test set consists of 2400 coding ORFs and 2400 non–coding sequences, respectively, and it is therefore appropriate to use $S_p$ as defined in Eq. (5) rather than in Eq. (6). To perform the self–consistency and cross–validation tests, two independent sets are needed, namely a training set and a test set. They both consist of two parts: one part includes the positive samples composed of true protein coding genes, and the other part includes the negative samples composed of non–coding DNA sequences.

In this paper, we use the 3392 known genes of the 1st class in the MIPS database as the positive samples. As should be pointed out, to extend the method to more complicated structures, we haven't excluded intron–containing genes of the 1st class. Considering that the intergenic sequence with length longer than 300 bp, which starts with ATG and ends with one of the stop codons, is unlikely to be ORF [9,10], we randomly select about 7600 such intergenic sequences from the 16 yeast chromosomes to produce the negative samples.

According to the ergodicity principle, we randomly select 992 representative positive samples and 992 representative negative samples to form the training set. To form the test set, we use the remaining 2400 positive samples and 2400 negative samples randomly selected from the remaining ones. Then, using the sequences in the training set, the average absolute error $\delta$ and the variance $\sigma^2$ (see Eq. (4)) are calculated, while by using the sequences in the test set, we obtain the $S_n$, $S_p$ and $\widetilde{S}$ (see Eq. (5)). We perform both the self–consistency test and the cross–validation test five times in this way, and list the results in Table 1 and Table 2. Observing Table 1, we find that the average absolute error $\delta$ and the variance $\sigma^2$ in each self–consistency test are fairly small, whose averages are only 0.0084 and 0.0077, respectively. This result indicates that the neural network has strong self–organizing and self–adaptability ability. In addition, as can be seen from Table 2, the accuracy in each cross–validation test is always greater than 96.0%, which is higher than that reported by Zhang *et al*. [9,10].

**Table 1.** The average absolute error $\delta$ and the variance $\sigma^2$ for 5 different training sets

| | 1 | 2 | 3 | 4 | 5 | average |
|---|---|---|---|---|---|---|
| $\delta$ | 0.0073 | 0.0089 | 0.0092 | 0.0083 | 0.0082 | 0.0084 |
| $\sigma^2$ | 0.0071 | 0.0080 | 0.0085 | 0.0075 | 0.0075 | 0.0077 |

**Table 2.** The accuracy of the prediction for 5 different test sets

| | 1 | 2 | 3 | 4 | 5 | average |
|---|---|---|---|---|---|---|
| $S_n$ | 95.75 | 97.70 | 97.12 | 96.67 | 96.00 | 96.65 |
| $S_p$ | 96.50 | 95.67 | 96.25 | 96.29 | 96.17 | 96.18 |
| $\widetilde{S}$ | 96.13 | 96.69 | 96.69 | 96.48 | 96.09 | 96.42 |

**Table 3.** The 134 ORFs of the 2nd class (no similarity) in the MIPS database, which are recognized as non–coding

| | | | | |
|---|---|---|---|---|
| yal064w | ydr042c | yhr095w | ylr124w | ynl211c |
| yar030c | ydr095c | yhr139c–a | ylr162w | ynl269w |
| yar047c | ydr102c | yhr173c | ylr296w | ynl303w |
| yar053w | ydr215c | yil012w | ylr366w | ynl324w |
| yar064w | ydr274c | yil058w | ylr400w | yol038c–a |
| yar070c | ydr278c | yil086c | ylr402w | yol118c |
| ybl044w | ydr344c | yil152w | ylr404w | yol160w |
| ybl048w | ydr350c | yir020c | ylr416c | yor015w |
| ybl071c | ydr524w–a | yjl027c | yml084w | yor024w |
| ybr027c | ydr525w | yjl028w | yml090w | yor029w |
| ybr056w–a | ydr535c | yjl064w | yml107c | yor183w |
| ybr126w–a | yel010w | yjl077c | yml122c | yor248w |
| ybr144c | yel014c | yjl118w | ymr003w | yor268c |
| ybr157c | yel059w | yjl136w–a | ymr007w | yor314w |
| ybr292c | yer066c–a | yjl215c | ymr057c | yor343c |
| ycl056c | yer091c–a | yjr023c | ymr082c | yor364w |
| ycl075w | yer135c | yjr120w | ymr103c | yor376w |
| ycr006c | yer172c–a | yjr157w | ymr122c | yor392w |
| ycr022c | yfr035c | ykl044w | ymr141c | ypl041c |
| ycr025c | yfr042w | ykl097c | ymr151w | ypl055c |
| ycr043c | yfr054c | ykl158w | ymr187c | ypl056c |
| ydl162c | ygl006w–a | ykr032w | ymr320w | ypl080c |
| ydl196w | ygl015c | ykr073c | ynl017c | ypr064w |
| ydr010c | ygl188c | yll030c | ynl122c | ypr096c |
| ydr015c | ygr290w | ylr111w | ynl150w | ypr170c |
| ydr024w | ygr291c | ylr112w | ynl174w | ypr170w–a |
| ydr029w | yhl037c | ylr122c | ynl179c | |

**Table 4.** The 271 ORFs of the 3rd class, questionable ORFs, in the MIPS database, which are recognized as non–coding

| | | | | |
|---|---|---|---|---|
| yal019w–a | ydr355c | ygr228w | ykl136w | ynl171c |
| yal031w–a | ydr360w | ygr259c | ykl147c | ynl184c |
| yal056c–a | ydr401w | ygr265w | ykl153w | ynl198c |
| yal059c–a | ydr431w | yhl002c–a | ykl169c | ynl205c |
| ybl012c | ydr467c | yhl006w–a | ykl202w | ynl226w |
| ybl053w | ydr521w | yhl019w–a | ykr033c | ynl228w |
| ybl070c | ydr526c | yhl030w–a | ykr047w | ynl235c |
| ybl073w | yel009c–a | yhl046w–a | yll020c | ynl266w |
| ybl077w | yel018c–a | yhr028w–a | ylr101c | ynl296w |
| ybl083c | yel075w–a | yhr049c–a | ylr123c | ynr025c |
| ybl094c | yer014c–a | yhr056w–a | ylr140w | yol013w–a |
| ybl107w–a | yer046w–a | yhr063w–a | ylr169w | yol150c |
| ybr090c | yer076w–a | yhr071c–a | ylr171w | yor041c |
| ybr113w | yer084w | yhr125w | ylr198c | yor082c |
| ybr116c | yer084w–a | yhr145c | ylr202c | yor102w |
| ybr178w | yer087c–a | yil020c–a | ylr230w | yor121c |
| ybr206w | yer119c–a | yil030w–a | ylr232w | yor146w |
| ybr224w | yer133w–a | yil047c–a | ylr252w | yor169c |
| ybr226c | yer137w–a | yil060w | ylr261c | yor199w |
| ybr266c | yer138w–a | yil066w–a | ylr269c | yor218c |
| ycl023c | yer145c–a | yil068w–a | ylr279w | yor225w |
| ycl041c | yer148w–a | yil071w–a | ylr280c | yor235w |
| ycr041w | yer152w–a | yil100c–a | ylr282c | yor263c |
| ycr049c | yer181c | yil163c | ylr294c | yor282w |

**Table 4.** (Continued)

| | | | | |
|---|---|---|---|---|
| ycr064c | yfl012w–a | yil171w–a | ylr302c | yor300w |
| ycr087w | yfl032w | yir017w–a | ylr317w | yor309c |
| ydl009c | yfr036w–a | yjl009w | ylr358c | yor325w |
| ydl016c | yfr052c–a | yjl015c | ylr428c | yor333c |
| ydl032w | yfr056c | yjl022w | ylr434c | yor345c |
| ydl050c | ygl052w | yjl032w | ylr444c | yor379c |
| ydl062w | ygl072c | yjl067w | ylr458w | ypl025c |
| ydl151c | ygl088w | yjl086c | ylr465c | ypl034w |
| ydl152w | ygl109w | yjl120w | yml009c–a | ypl035c |
| ydl158c | ygl149w | yjl135w | yml034c–a | ypl102c |
| ydl172c | ygl152c | yjl150w | yml047w–a | ypl114w |
| ydl187c | ygl165c | yjl152w | yml089c | ypl185w |
| ydl221w | ygl177w | yjl175w | yml094c–a | ypl205c |
| ydr008c | ygl182c | yjl182c | yml116w–a | ypl261c |
| ydr034w–b | ygl193c | yjl188c | ymr031w–a | ypr039w |
| ydr048c | ygl204c | yjl202c | ymr075c–a | ypr044c |
| ydr053w | ygr011w | yjl211c | ymr086c–a | ypr050c |
| ydr094w | ygr018c | yjl220w | ymr135w–a | ypr053c |
| ydr112w | ygr025w | yjr018w | ymr158w–b | ypr077c |
| ydr114c | ygr039w | yjr020w | ymr172c–a | ypr087w |
| ydr133c | ygr064w | yjr038c | ymr193c–a | ypr092w |
| ydr136c | ygr069w | yjr087w | ymr290w–a | ypr099c |
| ydr149c | ygr073c | yjr128w | ymr304c–a | ypr136c |
| ydr157w | ygr107w | ykl030w | ymr306c–a | ypr142c |
| ydr199w | ygr115c | ykl036c | ymr316c–a | ypr146c |
| ydr203w | ygr137w | ykl053w | ynl013c | ypr150w |
| ydr220c | ygr139w | ykl083w | ynl028w | ypr177c |
| ydr230w | ygr151c | ykl111c | ynl105w | |
| ydr241w | ygr164w | ykl115c | ynl114c | |
| ydr290w | ygr176w | ykl118w | ynl120c | |
| ydr327w | ygr182c | ykl131w | ynl170w | |

**Table 5.** The 50 ORFs of the 4th class (similarity or weak similarity to known proteins) in the MIPS database, which are recognized as non–coding

| | | | | |
|---|---|---|---|---|
| ybr239c | ydr411c | ygl160w | ykl037w | ynl176c |
| ybr293w | ydr413c | ygl186c | ykr030w | ynr059w |
| ycl001w–a | yel045c | ygr101w | yll005c | yol163w |
| ycr001w | yer048w–a | ygr284c | ylr064w | yor053w |
| ydl119c | yer097w | yhr035w | ylr184w | yor247w |
| ydl228c | yer113c | yhr130c | ylr283w | ypl072w |
| ydr115w | yfl040w | yhr143w | ylr311c | ypr013c |
| ydr307w | yfl067w | yil025c | ylr365w | ypr015c |
| ydr319c | yfr057w | yjl091c | ymr158w | ypr079w |
| ydr393w | ygl046w | yjl193w | ymr245w | ypr094w |

## 3.2 Predict Genes in the ORFs of the 2nd–6th Classes

After performing the self–consistency and cross–validation tests, we randomly selected a group of weights from the above five different training sets to calculate the output $y$ of each query DNA sequence of the $2^{nd}$–$6^{th}$ classes in the MIPS database. If $y \geq 0.5$, the sequence is recognized as a true protein–coding gene; otherwise, it is recognized as a non–coding sequence. As a result, there are 134, 271, 50, 113 and 6 sequences in the $2^{nd}$–$6^{th}$ classes that are recognized as non–coding ORFs, respectively. We listed them in Tables 3–7.

**Table 6.** The 113 ORFs of the 5th class (similarity to unknown proteins) in the MIPS database, which are recognized as non–coding

| | | | | |
|---|---|---|---|---|
| yal034c | ydl183c | ygl260w | ykl018c–a | ynl034w |
| yal047w–a | ydl185c–a | ygl263w | ykl106c–a | ynl067w–a |
| yar060c | ydr084c | ygr149w | ykl165c–a | ynl074c |
| ybl029c–a | ydr210w | yhl034w–a | ykl223w | ynl156c |
| ybl049w | ydr306c | yhl041w | ykl225w | ynl162w–a |
| ybl059w | ydr367w | yhl044w | ykr051w | ynl194c |
| ybl108w | ydr425w | yhl045w | ykr065c | ynl326c |
| ybl109w | ydr459c | yhr017w | yll065w | ynl337w |
| ybr004c | ydr504c | yhr067w | ylr023c | ynl338w |
| ybr096w | ydr543c | yhr069c–a | ylr036c | ynr014w |
| ybr099c | ydr544c | yhr162w | ylr149c–a | ynr075w |
| ybr103c–a | yel033w | yhr212c | ylr156w | ynr077c |
| ybr168w | yel067c | yhr217c | ylr159w | yol002c |
| ybr250w | yel074w | yil029c | ylr161w | yol003c |
| ybr300c | yer140w | yil080w | ylr463c | yor044w |
| ycl002c | yer188c–a | yil090w | yml007c–a | ypl165c |
| ycl065w | yfl015c | yil174w | yml047c | ypl229w |
| ycr038w–a | yfl062w | yil175w | ymr013w–a | ypr071w |
| ycr097w–a | yfr012w | yir030w–a | ymr119w | ypr074w–a |
| ycr102w–a | ygl041c | yir040c | ymr155w | ypr100w |
| ydl027c | ygl124c | yjl052c–a | ymr181c | ypr151c |
| ydl114w–a | ygl219c | yjr013w | ymr326c | |
| ydl159w–a | ygl231c | yjr162c | ynl018c | |

**Table 7.** The 6 ORFs of the 6th class (strong similarity to known proteins) in the MIPS database, which are recognized as non–coding

| | | | | | |
|---|---|---|---|---|---|
| yar061w | ybl009w | ycr063w | yel004w | ypl032c | ypl183w–a |

Furthermore, we re–estimate the number of protein coding genes in the 16 yeast chromosomes based on the above results. Take the 2nd class ORFs as an example, we calculate *FP*, *FN*, *TN* and *TP*. The total number of the 2nd class ORFs is 516, in which 134 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:

$$\begin{cases} TP/(TP+FN)=0.96 \\ \mathrm{TN}/(TN+FP)=0.96 \\ TN+FN=134 \\ TP+FN+TN+FP=516 \end{cases}$$

Solving the above system of equations, we can obtain $FP \approx 5$, $FN \approx 16$, $TN \approx 118$, and $TP \approx 377$. Therefore, the number of real coding ORFs of the 2nd class equals to $TP + FN = 377 + 16 = 393$. Similar calculations for the others are performed. Note that for the 6th–class, the above system has negative solutions: $FP \approx 0$, $FN \approx 9$, $TN \approx -3$, $TP \approx 223$. The reason is that the number predicted as non–coding sequences is only 6, which is too small. In this case, we prefer $FN = 6$, $TN = 0$.Then we listed the "revised" results in Table 8.

**Table 8.** The numbers of predicted coding and non–coding ORFs of the 2nd–6th classes

| Class | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Total number of ORFs | 516 | 463 | 818 | 1003 | 229 |
| *TP* | 377 | 181 | 767 | 887 | 223 |
| *FN* | 16 | 8 | 36 | 37 | 6 |
| *TN* | 118 | 263 | 14 | 76 | 0 |
| *FP* | 5 | 11 | 1 | 3 | 0 |
| *TP* + *FN* | 393 | 189 | 803 | 924 | 229 |
| *TN* + *FP* | 123 | 274 | 15 | 79 | 0 |

Thus, the total number of protein coding genes should be equal to 5930, the sum of the number of the 1st class (3392) and the number of those in the 2nd–6th classes recognized by the present method (393 + 189+ 803+ 924 + 229 = 2538, see Table 8). Note that the accuracy is actually greater than 96%, so, this figure should be considered as an upper limit of the number of 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 [14–16].

## 4 CONCLUSIONS

In this paper, based on the single nucleotide frequencies at three codon positions in the ORFs and the redundancy $\gamma$ of the entropy, we obtain a 12–dimensional vector from a DNA primary sequence. Then, we apply a 12×21×1 three–layer feedforward ANN method to predict protein–coding genes in the yeast genome by training the network in a supervised manner with a highly popular algorithm known as the error back–propagation algorithm. By this method, we find that the numbers of coding sequences (*i.e*. *TP* + *FN* in Table 8) in the 2nd–6th classes are at most 393, 189, 803, 924 and 229, respectively. Thus, the total number of protein coding genes in the 16 yeast chromosomes is estimated to be less than or equal to 5930. This method is based on the assumption that the DNA sequences coding for proteins in the 1st class ORFs have similar statistical properties to those coding for proteins in the 2nd–6th class ORFs. The prediction is examined by self–consistency test and cross–validation test. As a result, the averages of the average absolute error $\delta$ and the variance $\sigma^2$ through the self–consistency test are 0.0084 and 0.0077, respectively, which indicates that the neural network has strong ability of self–organizing and self–adaptability. As can be obtained through cross–validation test, the accuracy of the prediction, 96.0%, is higher than that reported by Zhang *et al*. [9,10]. In a word, the successful rates of both self–consistency and cross–validation tests are quite high. Worthy of mentioning is that both the training set and the test set in this paper do not exclude the intron–containing genes. This fact may imply that the current artificial neural network method is a useful computer technique for predicting protein–coding genes, and can be extended to recognize genes with more complicated structures.

## Acknowledgment

We thank Dr. Ovidiu Ivanciuc for sending us some references. This work is supported in part by the National Natural Science Foundation of China.

# 5 REFERENCES

[1] C. Mathe, M. F. Sagot, T. Schiex and P. Rouze, Survey and Summary: Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Res*. **2002**, *30*, 4103–4117.
[2] J. Besemer, A. Lomsadze, and M. Borodovsky, GeneMarkS: a self–training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions, *Nucleic Acids Res*. **2001**, *29*, 2607–2618.
[3] R. J. Carter, I. Dubchak and S. R. Holbrook, A computational approach to identify genes for functional RNAs in genomic sequences, *Nucleic Acids Res*. **2001**, *29*, 3928–3938.
[4] R. Guigo, Computational gene identification: an open problem, *Comput*. *Chem*. **1997**, *21*, 215–222.
[5] T. A. Thanaraj, Positional characterization of false positives from computational prediction of human splice sites, *Nucleic Acids Res*. **2000**, *28*, 744–754.
[6] W. Li, P. Bernaola–Galvan, F. Haghighi, and I. Grosse, Applications of recursive segmentation to the analysis of DNA sequences, *Comput*. *Chem*. **2002**, *26*, 491–510.
[7] Y. D. Cai and P. Bork, Homology–Based Gene Prediction Using Neural Nets, *Analytical Biochem*. **1998**, *265*, 269–274.
[8] C. Burge and S. Karlin, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol*. **1997**, *268*, 78–94.
[9] 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.
[10] 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.
[11] F. B. Guo, H. Y. Ou, and C. T. Zhang, ZCURVE: a new system for recognizing protein–coding genes in bacterial and archaeal genomes, *Nucleic Acids Res*. **2003**, *31*, 1780–1789.
[12] Q. Liu, Y.S. Zhu, B. H. Wang and Y. X. Li, A HMM–based method to predict the transmembrane regions of $\beta$–barrel membrane proteins, *Computational Biol*. *Chem*. **2003**, *27*, 69–76.
[13] S. L. Salzberg, M. Pertea, A. L. Delcher, M. J. Gardner, and H. Tettelin, Interpolated Markov models for eukaryotic gene finding, *Genomics* **1999**, *59*, 24–31.
[14] 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, *Science* **1996**, *274*, 546.
[15] E. A. Winzeler, and R. W. Davis, Functional analysis of the yeast genome, *Curr*. *Opin*. *Genet*. *Dev*. **1997**, *7*, 771–776.
[16] 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 the yeast genome, *Nature*. **1997**, *387*, 7–8.
[17] 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.
[18] J. W. Fickett, Recognition of protein coding regions in DNA sequences, *Nucleic Acids Res*. **1982**, *10*, 5303–5318.
[19] 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.
[20] S. Guiasu, Information Theory with Application, *McGraw–Hill*. **1997**.
[21] Z. Y. Fu, Information Theory: Elementary Theory and Application, *Publishing House of Electronics Industry*. **2001**. (in Chinese).
[22] C. H. Wu, Artificial neural networks for molecular sequence analysis, *Comput*. *Chem*. **1997**, *21*, 237–256.
[23] Y. D. Cai, J. Hu, Y. X. Li, and K. C. Chou, Prediction of Protein Structural Classes by a Neural Network Method, *Internet Electron*. *J. Mol*. *Des*. **2002**, *1*, 332–338, http://www.biochempress.com.
[24] Y. D. Cai, X. J. Liu, X. Xu, and K. C. Chou, Support Vector Machines for Predicting Membrane Protein Types by Incorporating Quasi–Sequence–Order Effect, *Internet Electron*. *J. Mol*. *Des*. **2002**, *1*, 219–226, http://www.biochempress.com.
[25] Y. D. Cai, X. J. Liu, X. Xu, and K. C. Chou, Short Communication: Prediction of protein structural classes by support vector machines, *Comput*. *Chem*., **2002**, *26*, 293–296.
[26] Y. D. Cai, X. J. Liu, X. Xu, and K. C. Chou, Artificial neural network method for predicting protein secondary

structure content, *Comput*. *Chem*. **2002**, *26*, 347–350.
[27] Z. H. Chen and Z. Z. Yan, Artificial neural networks applications for genome informatics, *Foreign medical: Biomedical engineering fascicle* **2002**, *25*, 145–149, (in Chinese).
[28] Z. R. Yuan, Artificial Neural Networks with Application, *Tsinghua University Press*, **1999**, (in Chinese).
[29] S. Haykin, Neural Networks: A Comprehensive Foundation, Second Edition. *Tsinghua University Press*, 2001.
[30] M. Burset and R. Guigo, Evaluation of Gene Structure Prediction Programs, *Genomics* **1996**, 34, 353–367.

## Biographies

**Chun Li** is a M.S student of Applied Mathematics at the Dalian University of Technology. His main research interests include combinatorics, information theory and bioinformatics.

**Ping–an He** is a PhD student of Applied Mathematics at the Dalian University of Technology. His main research interests include combinatorics, graph theory and bioinformatics.

**Jun Wang** is a Professor of Applied Mathematics at the Dalian University of Technology, the advisor of the first author.