Classification of Genes Based on Synonymous Codon Usage

Christina Fan

CS229 Project Report
Fall 2009

Introduction

The prediction of the function of a novel gene remains to be a challenging problem. Given a piece of coding sequence, one can deduce its function by finding homologous genes using sequence or protein structural alignment. One can also perform gene expression measurements or gene knockout experiments to determine the function of a gene. However, a gene may not have homology with any known gene and experiments can be expensive. It would therefore be beneficial if the function of a gene can be predicted from the characteristics of its coding sequence. One such characteristics is synonymous codon usage.

Proteins are strings of amino acids and their sequences are translated from nucleotide sequences using the genetic code, which is a set of trinucleotide sequences. Each trinucleotide is termed a codon. Proteins are constructed from 20 different amino acids, yet there are $4^3 = 64$ different codons. The genetic code is therefore redundant - most amino acids are encoded by more than one codon (Table 1). Trinucleotide sequences coding for the same amino acids are referred to as synonymous codons. The usage of synonymous codons is not random – some codons are preferred over the others for a given amino acid and such perference varies among species as well as among genes within a species.

It has been shown that for a number of organisms such as yeast, bacteria, plants, insect, and mammals, genes can be clustered into groups that have different expression patterns and functions based on codon usage (1). In this project, I focused on baker's yeast (*Saccharomyces cerevisiae*), which is a unicellular eukaryote that has been studied extensively as a model organism. Sharp et. Al demonstrated that yeast genes can be seggregated into two main clusters with different expression levels based on synonymous codon usage (2). Najafabadi et. Al showed that co-expressed genes in yeast have similar synonymous codon usage and that codon usage can improve the prediction of protein-protein interaction in organisms including yeast (3,4). Here, I would like to build a classifier to predict the function of a gene, using yeast as a model.

Table 1. The Genetic Code.

Amino Acid	**Codons**
Alanine	GCT,GCC,GCA,GCG
Arginine	CGT,CGC,CGA,CGG,AGA,AGG
Asparagine	AAT,AAC
Aspartic acid	GAT,GAC
Cysteine	TGT,TGC
Glutamic acid	GAA,GAG
Glutamine	CAA,CAG
Glycine	GGT,GGC,GGA,GGG
Histidine	CAT,CAC
Isoleucine	ATT,ATC,ATA
Leucine	CTT,CTC,CTA,CTG,TTA,TTG
Lysine	AAA,AAG
Methionine	ATG
Phenylalanine	TTT,TTC
Proline	CCT,CCC,CCA,CCG
Serine	AGT,AGC,TCT,TCC,TCA,TCG
Threonine	ACT,ACC,ACA,ACG
Tryptophan	TGG
Tyrosine	TAT,TAC
Valine	GTT,GTC,GTA,GTG
Stop Codon	TAA,TAG,TGA

Methods

Data

Yeast genes that participated in various molecular pathways were retrieved from the KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database. The categories chosen for this study included those involved in metabolism, DNA replication and transcription, translation, and cellular processes. Details of the chosen pathways are presented in Table 2. Coding sequences of the genes were retrieved from the *Saccharomyces* Genome Database (SGB). Duplicated copies of genes within the same category were removed. Genes that were found in more than one categories were ignored. Only genes with more than 100 codons were considered. The final data set had 891 genes.

Feature Space

There are 64 different codons. However, since methionine and tryptophan have only 1 corresponding codon, these two codons do not contribute to the measurement of codon usage and are removed from the analysis. Also, stop codons that signal the termination of transcription are also removed. Each gene is therefore represented by a vector containing 59 elements, each element corresponds to the relative synonymous codon usage frequency (RSCU) (f_{ij}) (2),

$$f_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}}$$

where x_{ij} is the number of ocurrences of synonymous codon j of amino acid i and n_i is the number of synonymous codons encoding for amino acid i. This measure can be interpreted as the observed number of occurrence of codon i versus the expected number of occurrence of codon i given an amino acid j assuming uniform distribution. It has the advantage of removing the effect of amino acid composition on the codon usage profile of a gene. Since a gene does not necessarily carry all 20 amino acids, a pseudo-count (+1) was added for every codon.

Data Analysis

Calculation of codon usage and building of learning algorithms were done in Matlab.

Supervised Learning Algorithms

1. Softmax Regression

Softmax regression is an example of Generalized Linear Model applied to multinomial data. The model takes the feature vector of each gene with its label in the training set and output the probabilities of a test sample being in the different classes. The parameters were found by maximizing the log-likelihood with batch gradient ascent. The class with the higest probability was chosen to be the predicted class.

2. Support vector machines (SVM)

Ma et. Al (5) demonstrated the use of SVM for the classification of human leukocyte antigens using codon usage. SVMs are inherently a two-class model but they can be extended to multiple classes. Two common ways are the one-versus-all method, where k SVM models are built, one for each of the k classes. A test sample is assigned to the class that classifies it with the largest margin. The second approach is the one-versus-one model, where $k(k-1)/2$ models are built. Each model is built for a pair of the classes. The label of a test sample will be the class that are chosen by most models. Here, I used the later approach.

Evaluation of Learning Algorithms

Ten-fold cross validation was used to evaluate the performance of the classifiers. Briefly, the data set was partitioned into 10 subsets of approximately equal size, each subset containing equal proportion of genes from each category. In each validation, nine subsets were used for training and the remaining subset was used for testing. The testing error was calculated as the average of the testing errors from all ten validations.

Table 2. Yeast genes participating in the listed pathways were included in the analyses.

Category	**Metabolism**	**DNA Replication, Transcription**	**Translation**	**Cellular Processes**
Included Pathways	Glycolysis Starch metabolism Galactose matabolism Citrate cycle Oxidative phosphorylation Pyruvate metabolism Steroid biosynthesis Fatty acid metabolism Fatty acid biosynthesis Valine biosynthesis Lysine biosynthesis Alanine metabolism Tyrosine metabolism Glycine metabolism Lysine degradation Phenylalanine metabolism Tryptophan metabolism Histidine metabolism Cysteine metabolism Valine degradation P450 cytochromes Pyrimidne metabolism Purine metabolism Sulphur metabolism Nitrogen metabolism	DNA replication Base excision repair Nucleotide excision repiar Mismatch repair Ubiquitin mediated proteolysis RNA degradation RNA polymerase Basal Transcription factors Spliceosome	Ribosomes	Cell cycle Meiosis Endocytosis Autophagy
Total number of genes	386	177	123	204

Results

Principal Component Analysis

The data set of 891 yeast genes, each represented by a feature vector of 59 elements, was subject to principal component analysis. The first two principal components explained 37% and 11 % of the total variance, repsectively. Figure 1 shows the representation of each gene in the space spanned by the first two principapl components. It was obvious from Figure 1 that different functional categories of genes occupied different spaces defined by the first two components, except for the categories of DNA replication/transcription and cellular processes (Figure 1, right panels). These two groups were thus grouped together for the building of gene function classifiers. Interestingly, ten of the genes partitcipating in the oxidative phosphorylation pathway were located in unique location in the principal component space (Figure 1, bottom panel). Further exmination of the data set revealed that they were genes in the mitochondria, a DNA continaing organelle that participates in energy metabolism in eukaryotes and is believed to be derived from endosymbiotic prokaryotes. It was therefore not surprising that mitochondrial genes have synonymous codon usage very different from nuclear genes. None of the other pathways contained mitochondrial genes. These ten genes were removed from the data set for training. The final data set thus consisted of 881 nuclear genes.

Figure 2 is a plot of the contribution of each codon to the first two principal components. One interesting observation was that codons ending with A or T contributed positively to the second principal component while those ending with G or C contributed negatively.

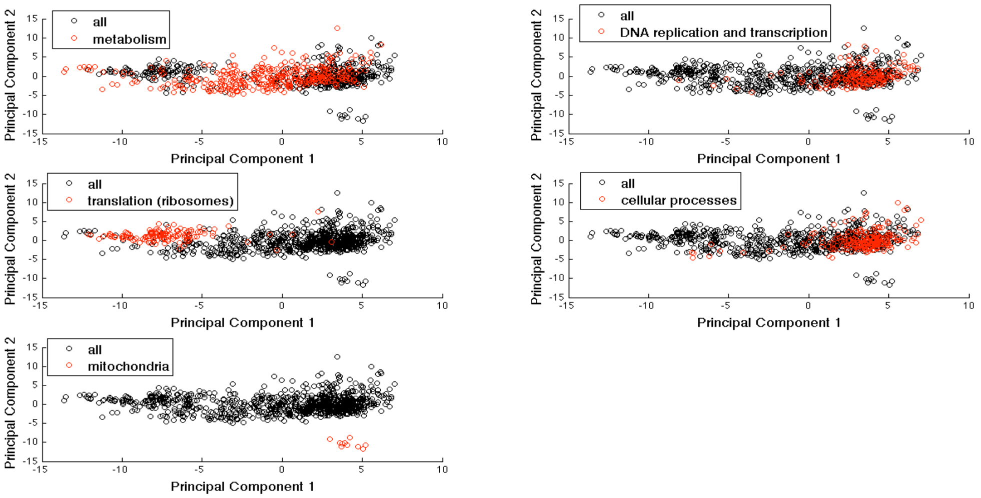

Figure 1. Principal component analysis of synonymous codon usage of yeast genes.

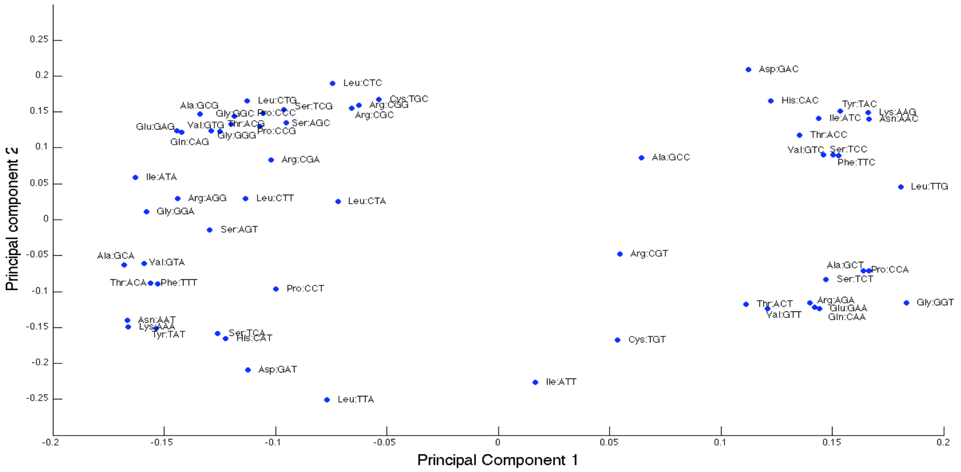

Figure 2. Contribution of each codon to the first two principal components.

Supervised Learning

Classifiers were built for three classes/categories using 881 nuclear genes. Category 1 corresponded to genes involved in metabolism; category 2 corresponded to genes involved in DNA replication, transcription, and cellular processes; category 3 corresponded to ribosomes.

Softmax regression

Using batch gradient descent with a learning rate of 0.0001, yeast genes were classified into 3 categories based on their codon usage. Table 3 shows the confusion matrix summed over all ten validations (i.e. each cross validation generated a confusion matrix and Table 3 is the sum of all 10 confusion matrices, such that each gene was classified once). The average test error (percentage of genes misclassified) was 21.6% , which was close to the training error. Most of the genes in category 1 (metabolism) that were misclassified fell into category 2 (DNA replication, translation, cellular processes). The specificity of category 2 (number of true category 2 genes / number of genes classified into category 2) was quite low,

~62%. This was not surprising, since some of the category 1 genes fell into similar space defined by the first two principal components as category 2 (Figure 1).

Table 3. Confusion matrix of softmax regression.

predicted

		1	2	3	True Positive proportion
	1	270	88	18	0.72
target	2	32	145	0	0.82
	3	24	1	98	0.80
	True Positive proportion	0.83	0.62	0.84	

Average Train Error (%)	19.46
Average Test Error (%)	21.59

Support Vector Machines

Three regularized SVMs were built, one for each category pair (categories 1 vs. 2, 2 vs. 3, 1 vs. 3). A gene was classified into category *i* if both SVMs involving category *i* gave the same prediction *i*. If the prediction of any two SVMs disagreed, no prediction was made for the gene and such event was termed 'uninformative'. Different kernels and values for *C*, the parameter that controls the weighting of the slack variables, were experimented. Tables 4 and 5 list the training error, test error, and uninformative rate of the multi-class SVM using polynomial (linear, quadratic, and cubic) and Gaussian kernels, respectively, after ten-fold cross-validation. Decreasing *C* reduced overfitting, as seen from the increase in training error. Quadratic and cubic kernels tended to overfit the data, as observed from the zero training error and relatively high test error rates. Linear kernels performed better than cubic and quadratic kernels. The best error rate of the multi-class SVM was 18.07%, which was obtained using a Gaussian kernel. Such eror rate was better than that obtained from softmax regression. Table 6 is the confusion matrix with the best error rate.

Table 6. Confusion matrix of the best multi-class SVM model.

Gaussian Kernel
(Sigma=5, C=1.5)

predicted

		1	2	3	True Positive proportion
	1	281	84	11	0.75
target	2	56	325	0	0.85
	3	6	2	115	0.93
	True Positive proportion	0.82	0.79	0.91	

Uninformative rate=0%
Average test error=18.07%

Table 4. Errors of the multi-class SVM using polynomial kernels under different parameters.

Polynomial Kernel

Poly-nomial Degree	**C**	**Average Training Error (%)**			**Average Test Error (%)**	**Average Uninformative Rate (%)**
		1 vs. 2	**2 vs. 3**	**1 vs. 3**		
1	0.002	18.69	3.40	8.46	22.27	0.00
	0.005	17.69	3.11	6.90	21.03	0.57
	0.01	17.23	2.98	5.95	**20.68**	0.57
	0.05	16.70	1.87	6.12	20.92	0.57
	1	16.63	1.63	5.88	21.20	0.80
2	0.002	2.95	0.00	0.96	26.01	1.70
	0.005	1.14	0.00	0.22	28.34	1.82
	0.01	0.32	0.00	0.00	30.91	1.48
	0.05	0.00	0.00	0.00	31.34	1.36
	1	0.00	0.00	0.00	31.34	1.36
3	0.002	0.00	0.00	0.00	23.50	0.91
	0.005	0.00	0.00	0.00	23.50	0.91
	0.01	0.00	0.00	0.00	23.50	0.91
	0.05	0.00	0.00	0.00	23.50	0.91
	1	0.00	0.00	0.00	23.50	0.91

Table 5. Errors of the multi-class SVM using Gaussian kernels under different parameters. The best error rate is highlighted in bold.

Gaussian Kernel

σ	**C**	**Average Training Error (%)**			**Average Test Error (%)**	**Average Uninformative Rate (%)**
		1 vs. 2	**2 vs. 3**	**1 vs. 3**		
5	0.1	17.51	3.02	8.44	22.95	0.00
	0.5	12.95	1.15	3.74	19.32	0.00
	1	8.81	0.53	2.07	18.52	0.00
	1.5	6.16	0.33	1.40	**18.07**	0.00
	2	4.11	0.07	0.96	18.64	0.00
8	0.1	19.67	3.20	9.78	23.18	0.00
	0.5	16.06	2.14	6.01	20.45	0.00
	1	14.59	1.21	3.70	18.98	0.00
	1.5	13.44	0.86	3.16	18.86	0.00
	2	12.39	0.79	3.01	18.41	0.00
10	0.1	20.37	3.44	10.47	23.64	0.00
	0.5	17.00	2.93	6.66	20.45	0.00
	1	15.97	2.07	4.88	19.45	0.11
	1.5	15.07	1.32	3.74	18.98	0.00
	2	14.47	1.01	3.27	19.20	0.00

Conclusion

The function of yeast genes could be predicted fairly well based solely on the coding sequences. In this study, one-versus-one support vector machine with a Gaussian kernel performed better than softmax regression. It would be interesting to investigate whether the use of codon usage for gene prediction could be extended to other organisms including higher level eukaryotes.

References

(1) Gustafsson C et. Al. Codon bias and heterologous protein expression. Trends in Biotechnology. Vol. 22 No. 7. 2004.

(2) Sharp PM et. Al. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Research. 14 (13): 5125-5143. 1986.

(3) Najafabadi HS et. Al. Sequence-based prediction of protein-protein interactions by means of codon usage. Genome Biology. 9:R87. 2008.

(4) Najafabdai HS et. Al. Universal function-specificity of codon usage. Nucleic Acids Research. 1-10. 2009.

(5) Ma J et Al. Gene Classification Using Codon Usage and Support Vector Machine. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 6(1): 134-143. 2009.