# Global Topology of Codon Usage Equality Networks of Escherichia Coli Essential Genes

Mohammad-Hadi Foroughmand-Araabi[1], Sama Goliaei[2], and Bahram Goliaei[1]

[1] Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.
[2] Faculty of New Sciences & Technologies, University of Tehran, Tehran, Iran.
foroughmand@ut.ac.ir, sgoliaei@ut.ac.ir, goliaei@ibb.ut.ac.ir

**Abstract.** Relation between codon usage and other biological phenomena in living organisms, has been proved in previous researches. In advance, recently, it is shown that the codon usage of co-expressed and co-function genes are similar. In this paper, we introduced a set of networks called codon usage equality networks, each network for an amino acid. These networks represent codon usage similarities between genes. We showed that, at least some of these networks have scale-free and small world properties. Also we showed that betweenness centrality and degree of network nodes are related proportionally. We also compared the networks for different amino acids. We showed that among all codon usage equality networks of different amino acids, networks corresponding to amino acids "Proline", "Valine", and "Arginine" are more similar to each other.

**Keywords:** Codon usage, Genome sequence pattern, Biological network, Biological network topology

## 1 Introduction

The codon degeneracy phenomenon, which is the phenomenon that more than one codon may code for an amino acid, is discovered in 1965 [18]. It is shown that there are tendencies in organisms to prefer some codons over the others [17]. The frequency of appearance of a codon in comparison to other codons, which are coding the same amino acid (synonymous codons), is called the codon usage.

Although many researchers assign one codon usage to each organism, however, it is known that the codon usage for different genes within an organism are different. This phenomenon was a motivation for a lot of further studies in this area.

Many researches have reported the synonymous codon selection in human [5]. Three main mechanism are provided for this phenomenon, namely, translation efficiency [12, 19], mRNA stability [6], and splicing control [19, 4]. It is known that the speed of translation is proportionally related to the amount of available tRNAs for the codons which are presented in the gene. That is because the

translation process consists of some consecutive operations of finding appropriate tRNAs for codons which are placed on mRNAs. Thus, codon usage and tRNA abundance together control the translation rate [12].

Recently, Najafabadi, et al. showed that co-expressed and co-function genes use similar ratios of codons for amino acids. They claimed that the organism changes the abundance of tRNAs, and by this change controls the translation rate of proteins [16].

The concept of networks is widely used in various aspects of biological studies, such as topological properties of protein-protein interaction (PPI) network, metabolic network, transcription regulation network, signal transduction network, and functional association networks [1]. The concept of networks is also useful in some less popular biological areas such as, protein domain networks which represent appearance of protein domains in different proteins [22], and amino acid bounding networks which represent chemical bounds between amino acids in a three-dimensional structure of a protein [10].

In this paper, we introduced the concept of "codon usage equality network" for the first time. In contrast to the network representation of codon usage equalities, previous studies focus on the similarities between genes, independently. By considering similarities between gene codon usages independently, the distribution of equal codon usages and their relations are omitted. For example, connected components in networks represent connectedness of genes with similar codon usages, which is defined only when considering similarities as a network. Also, topological analysis like analysis of clustering coefficients and betweenness centralities is only possible for networks.

In this paper, we analyzed topological properties of codon usage equality networks for different amino acids. Based on codon usage equality networks for different amino acids, we provided a hypothesis that explains benefits of using these codon usage equality networks in protein expression regulation.

## 2 Materials and Methods

### 2.1 Codon Usage Equality Network

The "codon usage equality network" is constructed based on the similarity between the codon usages of the genes. For example, consider the network for amino acid $a$. In this network, nodes represent genes, and two genes are neighbors if and only if they use similar ratios of codons for amino acid $a$.

In order to exactly specify the edges of the network, we need to clearly define the concept of "using similar ratios of codons". Let $u$ and $v$ be two nodes of the network for amino acid $a$, which are representing genes $g_u$ and $g_v$, respectively. Also, suppose that, amino acid $a$ has codons $c_1, \ldots, c_k$. Number of occurrences of codon $c_i$ in gene $g_v$ is represented as $n_{v,c_i}$. The codon usage corresponding to gene $g_v$ is represented as $cu_v$ which is a vector of length $k$ (the number of codons that code for amino acid $a$) which is $cu_v[i] = n_{v,c_i} / \sum_j n_{v,c_j}$ . In this equation, $\sum_j n_{v,c_j}$ is the number of occurrences of all codons of amino acid $a$,

thus, $cu_v[i]$ represents the percentage of occurrence of codon $c_i$. Thus, we have $\sum_i cu_v[i] = 1$.

The likelihood ratio test statistic for two codon usages $cu_v$ and $cu_u$ is $D(v,u) = -2\ln(L_n/L_a)$, where $L_n$ and $L_a$ are maximum likelihood of the null and the alternative models, respectively. In the alternative model, there is one parameter for each variable $cu_v[i]$ and $cu_u[i]$. According to the definition, maximum likelihood for variables $cu_v[i]$ and $cu_u[i]$ are $cu_v[i]^{n_{v,c_i}}$ and $cu_u[i]^{n_{u,c_i}}$, respectively. In the null model for each codon $c_i$, there is one parameter which represents average behavior of $cu_v[i]$ and $cu_u[i]$. The maximum likelihood for this parameter in the alternative model is

$$\left(\frac{n_{v,c_i} + n_{u,c_i}}{\sum_j n_{v,c_j} + n_{u,c_j}}\right)^{n_{v,c_i}+n_{u,c_i}} \tag{1}$$

The maximum likelihood for each model is the product of maximum likelihood for its parameters.

We computed the probability of a chi-squared distribution with $k-1$ degrees of freedom to get a value less than $D(v,u)$, and name this probability as $p(v,u)$. This is the p-value of not considering $v$ and $u$ as non-equal probability vectors. We choose $k-1$ as the degrees of freedom, because, we are testing equality of two vectors of probabilities of size $k$, and in each vector of probability, the sum of values is 1. Thus, the vector have $k-1$ degrees of freedom. We say that these two codon usages are similar if and only if $p(v,u) < t$ for some threshold $t$. In this paper, we used the threshold $t = \%5$.

Note that, in analyzing codon usage equalities we ignored amino acids with only one code. The codon usage for these amino acids, which are "Methionine" and "Tryptophan", is meaningless. Finally, we obtained 18 networks for 18 amino acids.

### 2.2 Network Analysis

A network $N$ is represented as $N = (V, E)$ where $V$ is the set of nodes (vertices) and $E$ is the set of edges. In this paper, we modeled the network as an undirected network. Degree of a node $v$ is the number of its neighbors and is represented as $d_v$. Network density is defined as the proportion of all possible edges which are presented in the network, i.e. the density is $e/(n(n-1)/2)$, where $e$ is the number of edges and $n$ is the number of nodes of the network. A connected component of a network is a subset of its nodes which are connected with one or more edges.

Distance between two nodes $v$ and $u$ is the number of edges in the shortest path between these two nodes and is represented as $d(v,u)$. Diameter of a network is the maximum of all $d(v,u)$ for $v \neq u$. Let eccentricity of a node $v$ to be the maximum of shortest paths between node $v$ and all other nodes. Radius of a network is the minimum of eccentricity of the nodes in the network. Characteristic path length of a network is the average of shortest paths between all possible pairs of nodes. Network centralization represents the extent to which

nodes are more central than others and is computed as (see [14])

$$\frac{n}{n-2}\left(\max_v\{d_v\}/(n-1) - e/(n(n-1)/2)\right) \tag{2}$$

Closeness centrality of a node $v$ is

$$C_v = 1/\sum_{u\neq v} d(v,u)$$

The closeness centrality represents that how far the node is from other nodes. A node with high closeness centrality tends to be a node which has short distances to all other nodes, while a node with low closeness centrality is a node which is far from other nodes. Roughly speaking, the closeness centrality is high when the node is almost the center of the network. The closeness centrality of a network is the average of the closeness centralities of its nodes. Betweenness of a node $v$ represents the importance of this node in the formation of shortest paths in the network. The betweenness $B(v)$ is

$$B(v) = \sum_{s,t} \sigma_{s,t|v}/\sigma_{s,t} \tag{3}$$

where $\sigma_{s,t}$ is the number of shortest paths between $s$ and $t$, and $\sigma_{s,t|v}$ is the number of shortest paths between $s$ and $t$ which are passing through $v$. The betweenness of a network is the average of the betweenness of its nodes.

The clustering coefficient of a node $v$ is the ratio of the pairs of its neighbors which are also neighbors. Thus, the clustering coefficient of $v$ is

$$CC(v) = \frac{|\{s,t|s\to v, t\to v, s\to t\}|}{d_v(d_v-1)/2} \tag{4}$$

where $v \to u$ indicates that $v$ is neighbor of $u$. Two neighbors of a node with high clustering coefficient are highly likely to be neighbors. The clustering coefficient of a network is defined as the average of clustering coefficients of its nodes, i.e. $CC = \sum_v CC(v)/n$, where $n$ is the number of nodes of the network.

Heterogeneity of a network represents how much the network tends to have hubs, which are the nodes with high degrees in a network with low average degree. The heterogeneity of a network is defined as (see [9])

$$\sqrt{\frac{\sum_v d_v^2}{(\sum_v d_v)^2/n} - 1} \tag{5}$$

A network is called scale-free if its degree distribution follows the power law, i.e. $P(k) \propto k^\gamma$ for $k$ greater than some fixed value $k_0$, where $P(k)$ is the proportion of the nodes with degree $k$.

We used the Gephi [3] software and developed python scripts and C++ applications to compute above mentioned statistics for the networks.

### 2.3 Comparing Codon Usage Equality Networks of Amino Acids

We compared structure of codon usage equality networks of different amino acids. First, for each pair of networks, we evaluated the correlation between presences of edges in two networks, $n_1$ and $n_2$. We considered the presence of edges in the first (second) network as a random variable $X$ ($Y$). Thus, $x_e = 1$ ($y_e = 1$) if and only if the edge $e$ is presented in the first (second) network.

Value of the correlation coefficient is a number in the range of -1 to 1. The closer the correlation coefficient is to 1 or -1, the stronger the relation between two random variables is. Positive values indicate positive linear relation, and negative values indicate negative linear relation.

### 2.4 Dataset

In this paper, we studied the topology of the "codon usage equality network" which is constructed based on codon usages of essential genes of Escherichia Coli. First, we started by candidate essential genes of *Escherichia coli* which are the genes unable to be deleted from the organism's chromosomes [2]. 300 *E. Coli* known genes are reported to be essential for it [2]. For these genes, we computed the codon usage of each amino acid. Thus, for each gene, we had 20 codon usages, for 20 amino acids. The sequence of the genes are gathered from the CDS database of the EBI. In order to remove the bias of considering only one strain of the *E. Coli*, we computed average codon usages of the sequenced genes of all available strains.

## 3 Results

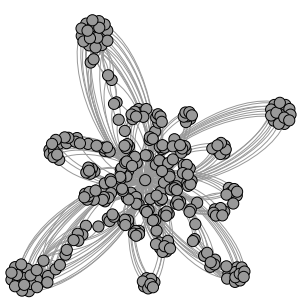

**Fig. 1.** Codon usage equality network for amino acid "Glutamic acid".

The codon usage equality network for amino acid "Glutamic acid" is presented in Fig. 1. Supplementary materials of this paper is available online at `http://cbp.ut.ac.ir/cuen/`.

### 3.1 Codon Usage Equality Networks Are Scale-Free and Small-World for Some Amino Acids

For each of 18 amino acids with more than one codon, we built a network. For each, number of nodes, number of edges, network density, network heterogeneity, number of connected components, network diameter, network radius, network centralization, characteristic path length, average number of neighbors, and network clustering coefficient are presented in Table 1. These network statistics represent global topology of the network. As it could be seen in the table, clustering coefficients of the networks are more than 0.6 for all amino acids, except the amino acid "Alanine". It means, the networks consist of some clusters with high density of edges.

**Table 1.** General topological measures of the codon usage equality networks for different amino acids.

| AA | Number of occurrences | Number of codons | Number of nodes | Number of edges | Density | Heterogeneity | Connected components | Diameter | Radius | Centralization | Characteristic path length | Average degree | Clustering Coefficient | Closeness | Betweenness |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 3682 | 2 | 300 | 19405 | 0.433 | 0.671 | 1 | 2 | 1 | 0.571 | 1.567 | 129.367 | 0.886 | 1.567 | 84.817 |
| D | 23005 | 2 | 300 | 1218 | 0.027 | 2.164 | 1 | 2 | 1 | 0.979 | 1.973 | 8.120 | 0.898 | 1.973 | 145.440 |
| E | 29127 | 2 | 300 | 1396 | 0.031 | 1.912 | 1 | 2 | 1 | 0.975 | 1.969 | 9.307 | 0.909 | 1.969 | 144.847 |
| F | 13933 | 2 | 300 | 2346 | 0.052 | 2.167 | 1 | 2 | 1 | 0.954 | 1.948 | 15.640 | 0.974 | 1.948 | 141.680 |
| H | 8642 | 2 | 300 | 4335 | 0.097 | 1.596 | 1 | 2 | 1 | 0.909 | 1.903 | 28.900 | 0.971 | 1.903 | 135.050 |
| K | 21057 | 2 | 300 | 1056 | 0.024 | 0.729 | 44 | 13 | 0 | 0.044 | 3.710 | 7.040 | 0.909 | 2.074 | 23.803 |
| N | 15162 | 2 | 300 | 1550 | 0.035 | 1.088 | 62 | 8 | 0 | 0.080 | 1.664 | 10.333 | 0.917 | 1.323 | 4.667 |
| Q | 17686 | 2 | 300 | 2018 | 0.045 | 1.867 | 1 | 2 | 1 | 0.961 | 1.955 | 13.453 | 0.968 | 1.955 | 142.773 |
| Y | 10488 | 2 | 300 | 4526 | 0.101 | 1.620 | 1 | 2 | 1 | 0.905 | 1.899 | 30.173 | 0.970 | 1.899 | 134.413 |
| I | 24257 | 3 | 300 | 1849 | 0.041 | 1.995 | 1 | 2 | 1 | 0.965 | 1.959 | 12.327 | 0.903 | 1.959 | 143.337 |
| A | 40295 | 4 | 300 | 75 | 0.002 | 1.510 | 232 | 5 | 0 | 0.012 | 1.518 | 0.500 | 0.468 | 0.465 | 0.193 |
| G | 31098 | 4 | 300 | 334 | 0.007 | 1.786 | 175 | 12 | 0 | 0.053 | 3.403 | 2.227 | 0.675 | 1.104 | 10.783 |
| P | 17965 | 4 | 300 | 1599 | 0.036 | 3.173 | 1 | 2 | 1 | 0.971 | 1.964 | 10.660 | 0.965 | 1.964 | 144.170 |
| T | 20958 | 4 | 300 | 193 | 0.004 | 1.394 | 178 | 8 | 0 | 0.029 | 3.207 | 1.287 | 0.631 | 1.109 | 6.173 |
| V | 32623 | 4 | 300 | 431 | 0.010 | 5.987 | 1 | 2 | 1 | 0.997 | 1.990 | 2.873 | 0.916 | 1.990 | 148.063 |
| L | 40368 | 6 | 300 | 90 | 0.002 | 3.003 | 254 | 3 | 0 | 0.038 | 1.435 | 0.600 | 0.720 | 0.287 | 0.200 |
| R | 26438 | 6 | 300 | 1905 | 0.042 | 2.148 | 1 | 2 | 1 | 0.964 | 1.958 | 12.700 | 0.915 | 1.958 | 143.150 |
| S | 20913 | 6 | 300 | 378 | 0.008 | 6.817 | 1 | 2 | 1 | 0.998 | 1.992 | 2.520 | 0.912 | 1.992 | 148.240 |

Rows represent amino acids, sorted by the number of codons that code for the amino acid. AA indicates amino acid. Two first columns, i.e. "Number of occurrences" and "Number of codons" represent amino acid related properties, while the others represent properties of the codon usage equality networks.

As it is shown in Table 1, number of edges of the networks for amino acids "Alanine", "Leucine", and "Threonine" are few, i.e. less than number of vertices minus one. As a consequence, these networks have more than one connected components. The network for amino acid "Glycine" has also only a few number

of edges more than 300. Interestingly, while two networks corresponding to amino acids "Lysine" and "Asparagine" have a lot of edges, but they have more than one connected components. Also, in comparison to other networks, these two networks have low betweenness centralities. Note that, radius 0 for disconnected networks represent undefined values.

The diameter of the networks for all the amino acids except "Glycine" and "Lysine" are low, i.e. less than or equal to 8 which is the $\log_2(300)$. This observation represents the small-world property of the codon usage equality networks.

We tested whether the networks are scale-free or not. We fitted a power law function $P(k) \propto k^{\gamma}$ to the degree distribution and evaluated the goodness of fit. We used the method which is provided by Clauset, et al. [7] to check whether degrees of a network is power law distributed or not. We tested this property for each codon usage equality network and reported $\gamma$ and p-values in Table 2. Low p-value, for example less than %5, indicates that the test rejects the hypothesis that the original data could have drawn from a power-law distribution. Thus, for 10 amino acids "Alanine", "Aspartic acid", "Glutamic acid", "Glycine", "Lysine", "Leucine", "Proline", "Serine", "Threonine", and "Valine", we cannot reject the hypothesis that the degree distribution is power-law. Among these 10 networks, 4 networks are the networks with a few number of edges (see Table 1). From two disconnected networks with high number of edges, the network for amino acid "Lysine" follows, and the network for amino acid "Asparagine" does not follow the power-law. The high value of $\gamma$ for the network of amino acid "Lysine" may be due to its high number of components.

**Table 2.** Fitting the power law function $P(k) \propto k^{\gamma}$ to the degree distribution of codon usage equality networks for different amino acids.

| AA | $\gamma$ | p-value | AA | $\gamma$ | p-value | AA | $\gamma$ | p-value |
|---|---|---|---|---|---|---|---|---|
| A | 3.00 | 1.00 | H | 2.64 | 0.00 | Q | 3.90 | 0.00 |
| C | 2.00 | 0.00 | I | 5.28 | 0.01 | R | 2.61 | 0.00 |
| D | 6.68 | 0.22 | K | 141.79 | 1.00 | S | 1.63 | 0.51 |
| E | 4.93 | 0.14 | L | 3.00 | 1.00 | T | 3.00 | 1.00 |
| F | 2.30 | 0.00 | N | 3.27 | 0.00 | V | 1.71 | 0.96 |
| P | 2.74 | 0.56 | G | 3.00 | 1.00 | Y | 2.43 | 0.00 |

AA indicates amino acid. P-value shows the level of confidence for rejecting the hypothesis that the degrees are drawn from a power-law distribution.

The distribution of degrees for 18 amino acids are presented in Fig. 2.

### 3.2 Proportional Relation of Betweenness Centrality and Degree

Betweenness centrality of the networks are computed for each node. We fitted the function $y = ax^b$ to the graphs representing betweenness centrality versus node degrees for different amino acids, where $y$ is the node's betweenness centrality and $x$ is the node's degree. P-values and R-squared value of the fitness is presented in Table 3. Resulting p-values show that, for all amino acid except

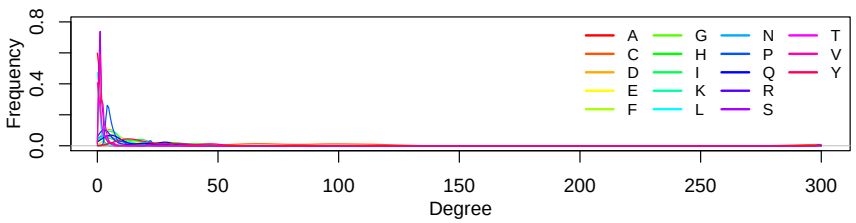


**Fig. 2.** Degree distribution of 18 amino acids in codon usage equality network of *E. Coli* essential genes.

"Alanine" and "Asparagine" the betweenness of the network nodes are proportionally related to the degrees, with p-value less than %1. As it is shown in this table, for networks corresponding to amino acids "Cysteine", "Phenylalanine", "Histidine", "Leucine", "Proline", "Serine", "Valine", and "Tyrosine" the value $R^2$ is higher than 0.8, i.e. %80 of the relation between betweenness centralities and degrees of nodes could be evaluated for these networks. Among the disconnected networks, only the network for amino acid "Leucine" have $R^2$ more than 0.3.

**Table 3.** Fitting a power law function, in the form $y = ax^b$, to the relation between betweenness centralities and degrees of nodes.

| AA | a | b | p-value | $R^2$ | a | b | p-value | $R^2$ | a | b | p-value | $R^2$ | a | b |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2.719 | -0.123 | 0.903 | 0.001 | 0.000 | 3.263 | 0.000 | 0.891 | 0.000 | 2.236 | 0.000 | 0.772 | 0.007 | 1.887 |
| C | 0.000 | 6.416 | 0.000 | 0.994 | 0.009 | 1.989 | 0.000 | 0.546 | 0.040 | 1.833 | 0.000 | 0.736 | 0.000 | 1.627 |
| D | 0.067 | 1.361 | 0.000 | 0.328 | 0.267 | 1.842 | 0.002 | 0.095 | 6.604 | 2.483 | 0.000 | 0.943 | 0.000 | 2.057 |
| E | 0.064 | 1.227 | 0.000 | 0.200 | 0.633 | 1.388 | 0.000 | 0.868 | 0.000 | 2.083 | 0.000 | 0.322 | 0.958 | 0.814 |
| F | 0.001 | 2.727 | 0.000 | 0.831 | 5.899 | -0.152 | 0.771 | 0.002 | 1.938 | 2.600 | 0.000 | 0.952 | 0.016 | 1.412 |
| G | 0.991 | 1.298 | 0.000 | 0.231 | 0.001 | 2.778 | 0.000 | 0.930 | 0.005 | 3.297 | 0.000 | 0.899 | 0.000 | 2.401 |

AA indicates amino acid. P-value shows the level of confidence for the data to be fitted to the power-law distribution.

### 3.3 Similarity of Codon Usage Equality Networks

We computed the correlation coefficient between codon usage equality networks, and presented the results in Table 4. No entry in this table is less than -0.01, it means that there is almost no negative relation between presences of edges between any two networks.

In Table 4, networks of amino acids "Alanine", "Glycine", "Leucine", and "Threonine" which have a few number of edges (see Table 1) have low value (less than 0.1) of correlation coefficients with other networks. Also, networks for amino acids "Asparagine" and "Lysine" which are disconnected networks are not similar to any other network with correlation coefficient more than 0.1. Interestingly, the correlation coefficients between these 6 networks are also less

than 0.1. The only connected network which is not similar to any other network with correlation coefficient more than 0.1 is the network corresponding to amino acid "Serine". On the other hand, networks for amino acids "Aspartic acid", "Phenylalanine", "Isoleucine", "Histidine", "Proline", "Arginine", "Valine", and "Tyrosine" are similar to 1, 1, 1, 3, 1, 2, 3 and 1 other networks with correlation coefficient more than 0.2, respectively. Highest values of correlation coefficients are between networks of amino acids "Valine" and "Proline", and amino acids "Valine" and "Arginine", which are 0.35 and 0.32.

## 4 Discussion

In this paper, for the first time, we introduced the concept of codon usage equality network. We built the codon usage equality network of *E. Coli* essential genes. These genes could represent elements of biological pathways which do not have alternatives. Consequently, their biological properties could be a representative of all the genes, and all the essential pathways may contain at least one of these genes.

Although there are some objections about the protein-protein interaction networks of *C. elegans*, *D. melanogaster*, and *E. Coli* to be best fitted to a power law function, but, this property holds for these networks at least approximately [20]. Also, properties of the protein-protein interaction network between essential genes of *E. Coli* is studied and it is shown that this network is also approximately scale-free [13]. The facts that clustering coefficients of the networks are high, diameters of the networks are low, and average distances between nodes are low (see Table 1) show that codon usage equality networks of *E. Coli* essential genes have small-world properties, at least for all amino acids except "Glycine", "Lysine", and "Threonine".

As it is shown in Table 4, codon usage equality networks are not very similar to each other. On the other hand, since the correlation coefficient between networks are not less than -0.01, networks are not negative of each other. As a conclusion, we can state that the networks for different amino acids contain almost independent set of edges.

### 4.1 Comparison with previous works

Some previous works on codon usage analysis have shown relations between codon usages and properties of genes and proteins. For example, Najafabadi, et al. considered co-expressed proteins. They showed that the similarity between codon usages of co-expressed genes is more than the similarity between two random genes. This result could be restated as a fact about the networks. Consider two networks, first, the codon usage similarity network between genes, and second, the network between genes representing co-expressed genes. Najafabadi, et al. [16] showed that if we partition the co-expressed genes network to some sub-networks with maximum edge densities, the density of edges of the codon usage similarity network in these sub-networks are higher than other random

sub-network partitioning. Roughly speaking, they showed that the codon usage equality network is more similar to co-expressed genes network than a random network. Also, they found the same result for co-function genes. In this study, in contrast, we directly considered the codon usage equality network, and evaluated its topological properties.

Some previous research works consider similarities/dissimilarities between codon usages of genes independently [15]. However, relations between codon usage similarities are not studied yet. We showed in this paper that codon usage equality networks have scale-free and small-world properties. These are the properties which are only present in networks.

In contrast to the non-network based approach, considering similarities as a network may reveal important properties of codon usages. For example, distribution of codon usage may affect structure of the networks. Betweenness centralities are different between a network which is constructed from completely random codon usages and a network which is constructed from codon usages with some specific preferences. Thus, by this method we can understand codon usage preferences of organisms.

**Table 4.** Correlation coefficient between codon usage equality networks for different amino acids.

| | A | C | E | D | G | F | I | H | K | L | N | Q | P | S | R | T | V | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| C | 0.00 | 1.00 | 0.05 | 0.06 | 0.04 | 0.09 | 0.10 | 0.10 | 0.00 | 0.04 | 0.07 | 0.11 | 0.19 | 0.10 | 0.11 | 0.02 | 0.07 | 0.16 |
| E | 0.00 | 0.05 | 1.00 | 0.23 | 0.00 | 0.03 | 0.17 | 0.02 | 0.00 | 0.03 | 0.01 | 0.02 | 0.02 | -0.01 | 0.01 | 0.00 | -0.01 | 0.11 |
| D | 0.00 | 0.06 | 0.23 | 1.00 | 0.00 | 0.03 | 0.20 | 0.03 | 0.01 | 0.02 | 0.02 | 0.01 | 0.02 | -0.01 | 0.02 | 0.00 | 0.02 | 0.14 |
| G | 0.00 | 0.04 | 0.00 | 0.00 | 1.00 | 0.01 | 0.03 | 0.05 | 0.00 | 0.03 | 0.04 | 0.00 | 0.05 | 0.05 | 0.03 | 0.02 | 0.00 | 0.04 |
| F | 0.00 | 0.09 | 0.03 | 0.03 | 0.01 | 1.00 | 0.02 | 0.21 | 0.02 | 0.00 | 0.04 | 0.02 | 0.03 | 0.00 | 0.13 | 0.00 | 0.00 | 0.03 |
| I | 0.00 | 0.10 | 0.17 | 0.20 | 0.03 | 0.02 | 1.00 | 0.01 | 0.01 | 0.04 | 0.04 | 0.01 | 0.18 | 0.00 | 0.04 | 0.02 | -0.01 | 0.21 |
| H | 0.01 | 0.10 | 0.02 | 0.03 | 0.05 | 0.21 | 0.01 | 1.00 | 0.03 | 0.02 | 0.07 | 0.03 | 0.14 | 0.02 | 0.24 | 0.01 | 0.21 | 0.11 |
| K | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.02 | 0.01 | 0.03 | 1.00 | 0.00 | 0.01 | -0.01 | 0.00 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 |
| L | 0.00 | 0.04 | 0.03 | 0.02 | 0.03 | 0.00 | 0.04 | 0.02 | 0.00 | 1.00 | 0.08 | 0.01 | 0.04 | 0.01 | 0.02 | 0.02 | 0.02 | 0.05 |
| N | 0.00 | 0.07 | 0.01 | 0.02 | 0.04 | 0.04 | 0.04 | 0.07 | 0.01 | 0.08 | 1.00 | 0.01 | 0.05 | 0.04 | 0.05 | 0.00 | 0.03 | 0.08 |
| Q | -0.01 | 0.11 | 0.02 | 0.01 | 0.00 | 0.02 | 0.01 | 0.03 | -0.01 | 0.01 | 0.01 | 1.00 | 0.19 | 0.00 | 0.03 | 0.02 | 0.02 | 0.04 |
| P | 0.00 | 0.19 | 0.02 | 0.02 | 0.05 | 0.03 | 0.18 | 0.14 | 0.00 | 0.04 | 0.05 | 0.19 | 1.00 | 0.00 | 0.19 | 0.02 | 0.35 | 0.13 |
| S | 0.00 | 0.10 | -0.01 | -0.01 | 0.05 | 0.00 | 0.00 | 0.02 | 0.00 | 0.01 | 0.04 | 0.00 | 0.00 | 1.00 | 0.04 | 0.01 | 0.00 | 0.00 |
| R | 0.00 | 0.11 | 0.01 | 0.02 | 0.03 | 0.13 | 0.04 | 0.24 | 0.02 | 0.02 | 0.05 | 0.03 | 0.19 | 0.04 | 1.00 | 0.04 | 0.32 | 0.02 |
| T | 0.00 | 0.02 | 0.00 | 0.00 | 0.02 | 0.00 | 0.02 | 0.01 | 0.00 | 0.02 | 0.00 | 0.02 | 0.02 | 0.01 | 0.04 | 1.00 | -0.01 | 0.02 |
| V | 0.00 | 0.07 | -0.01 | 0.02 | 0.00 | 0.00 | -0.01 | 0.21 | 0.01 | 0.02 | 0.03 | 0.02 | 0.35 | 0.00 | 0.32 | -0.01 | 1.00 | 0.00 |
| Y | 0.00 | 0.16 | 0.11 | 0.14 | 0.04 | 0.03 | 0.21 | 0.11 | 0.00 | 0.05 | 0.08 | 0.04 | 0.13 | 0.00 | 0.02 | 0.02 | 0.00 | 1.00 |

### 4.2 Benefits of using non-similar codon usage equality networks in protein expression regulation

The codon usage is known to be a factor for controlling the level of proteins [12, 8, 11, 19]. The speed of translation is proportional to the number of codons and amount of available tRNAs for those codons. For example, the translation of a gene containing a lot of codons with low amount of available tRNAs would be very slow. Thus, an organism may control the speed of translation by changing

the amount of available tRNAs for different codons [16]. However, two genes with similar ratios of codon for an amino acid, could not be distinguished by this mechanism. These genes, in our study, are modeled as neighboring nodes in a codon usage equality network. Besides, if two genes have different ratios of codons for an amino acid, by controlling the amount of available tRNAs for that amino acid, the organism may control the amount of available corresponding proteins. Considering the results we presented in this paper, which show that the networks for different amino acids have independent set of edges, the organism have the freedom to change the ratio of codons for different amino acids independently, and differentiate between more genes.

We presented an example to illustrate the benefits of having non-equal codon usage equality networks, for different amino acids. Consider an amino acid $X$, with four codons. The organism may change the amount of available tRNAs for these four codons, and consequently, may control the expression level of corresponding proteins. Note that, the organism is not able to change the ratio of protein expression levels for proteins with similar codon usage, in amino acid $X$, by changing the amount of tRNAs. According to the definition of the codon usage equality network, the genes with similar codon usages form a cluster in the codon usage equality network of the amino acid $X$. In this case we say the organism is able to differentiate between clusters of genes, in the ratios of expression levels, by controlling the amount of available tRNAs for different codons.

Suppose that the organism can differentiate between three clusters of genes by amino acid $X$, or equally, suppose that there are three clusters of genes in the codon usage equality network of amino acid $X$. Also, suppose that the organism can differentiate between 10 clusters of genes by controlling the amount of tRNAs for amino acid $Y$, and networks of amino acid $X$ and $Y$ are independent. Since the codon usage equality network for amino acid $Y$ is independent of the network of the codon usage equality for amino acid $X$, by controlling the ratio of tRNAs for amino acids $X$ and $Y$ simultaneously, the organism is able to distinguish between 30 clusters of genes. This is a very good advantage which is a result of using independent codon usage equality networks. Although this statement is true in theory, since the networks are not perfectly independent, the organism may not practically be able to distinguish between exactly 30 clusters of genes, but, having more independent gene equality networks gives the organism more freedom to distinguish between more clusters of genes.

As future works, we will study mechanisms behind formation of codon usage equality networks. Mechanisms that lead to formation of networks with similar properties as codon usage equality networks is studied previously [21]. We will check these formation mechanism and their validities for codon usage equality networks.

## References

1. Albert, R.: Scale-free networks in cell biology. J Cell Sci 118(21), 4947–4957 (2005)

2. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H.: Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. Mol Syst Biol 2(1) (2006)
3. Bastian, M., Heymann, S., Jacomy, M.: Gephi: An open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media (2009)
4. Cartegni, L., Chew, S.L., Krainer, A.R.: Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3(4), 285–298 (2002)
5. Chamary, J.V., Parmley, J.L., Hurst, L.D.: Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7(2), 98–108 (2006)
6. Chamary, J., Hurst, L.: Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6(9), R75 (2005)
7. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Rev 51(4), 661–703 (2009)
8. Comeron, J.M.: Selective and mutational patterns associated with gene expression in humans. Genetics 167(3), 1293–1304 (2004)
9. Dong, J., Horvath, S.: Understanding network concepts in modules. BMC Syst Biol 1(1), 24 (2007)
10. Greene, L.H., Higman, V.A.: Uncovering network systems within protein structures. J Mol Biol 334(4), 781 – 791 (2003)
11. Kotlar, D., Lavner, Y.: The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. BMC Genomics 7(1), 67 (2006)
12. Lavner, Y., Kotlar, D.: Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 345(1), 127 – 138 (2005)
13. Lin, C.C., Juan, H.F., Hsiang, J.T., Hwang, Y.C., Mori, H., Huang, H.C.: Essential core of protein-protein interaction network in escherichia coli. J Proteome Res 8(4), 1925–1931 (2009)
14. Ma, H.W., Zeng, A.P.: The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics 19(11), 1423–1430 (2003)
15. Najafabadi, H., Salavati, R.: Sequence-based prediction of protein-protein interactions by means of codon usage. Genome Biol 9(5), R87 (2008)
16. Najafabadi, H.S., Goodarzi, H., Salavati, R.: Universal function-specificity of codon usage. Nucleic Acids Res 37(21), 7014–7023 (2009)
17. Nakamura, Y., Gojobori, T., Ikemura, T.: Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res 28(1), 292 (2000)
18. Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., O'Neal, C.: RNA codewords and protein synthesis, VII. on the general nature of the RNA code. Proc Natl Acad Sci 53(5), 1161–1168 (1965)
19. Stoletzki, N., Eyre-Walker, A.: Synonymous codon usage in Escherichia coli: Selection for translational accuracy. Mol Biol Evol 24(2), 374–381 (2007)
20. Stumpf, M.P., Ingram, P.J., Nouvel, I., Wiuf, C.: Statistical model selection methods applied to biological networks. In: Transactions on Computational Systems Biology III, pp. 65–77. Springer (2005)
21. Wang, X.F., Chen, G.: Complex networks: small-world, scale-free and beyond. Circ Syst Mag, IEEE 3(1), 6–20 (2003)
22. Wuchty, S.: Scale-free behavior in protein domain networks. Mol Biol Evol 18(9), 1694–1702 (2001)