

Graph Theory Patterns in the Genetic Codes

William SEFFENS

*Department of Biological Sciences and Center for Theoretical Study of Physical Systems,
Clark Atlanta University, 223 James Brawley Dr., S.W., Atlanta, GA 30134, U.S.A.
E-mail address: WSeffens@cau.edu*

(Received September 7, 1998; Accepted January 17, 2003)

Keywords: Genetic Code, Graph Theory, mRNA, Codons

Abstract. The genetic code in biology describes how genes that are composed of DNA are translated into proteins composed of amino acids. There are twelve known genetic codes, the standard code used by most organisms, and alternate codes used by mitochondria and some lower organisms. These genetic codes can be represented as graphs, with vertices labeled by the amino acids and lines representing complementary DNA bases arranged as three-letter codons paired with reverse-complement codons. The resultant genetic graphs present forms that suggest an underlying mathematical symmetry that has been shaped by evolutionary forces. The forms of the graphs are discussed relating mathematics and biology.

1. Introduction

The flow of genetic information in biology is from DNA to a large number of mRNAs via a process called transcription, then from the mRNA molecules to proteins in a process called translation. The DNA and RNA are composed of long strings of nucleotide bases, represented as labels with the letters A, C, G, and T (except U for RNA sequences). The genetic code specifies how three DNA bases, as a group called a codon, are translated into an amino acid in a polypeptide or protein. In the standard genetic code, from one to six codons can specify any particular amino acid. The average codon degeneracy for all of the amino acids is three codons. Thus for a small mRNA coding for 100 amino acids, there are about 3^{100} or 5×10^{47} different combinations of bases using synonymous codons that code for the same polypeptide. The actual number of combinations would depend upon the frequency of occurrence of the amino acids in the protein, which varies with the organism considered (NAKAMURA *et al.*, 1997).

There are a total of twelve different genetic codes, the standard and eleven alternate codes (JUKES and OSAWA, 1993). The standard code is utilized in bacteria and in most higher organisms where the cellular nucleus contains the chromosomes of DNA. Sub-cellular organelles called mitochondria contain a much smaller amount of DNA than the nucleus, and implements one of the several alternate genetic codes. All of the genetic codes translate into exactly the same set of twenty amino acids. Differences in the genetic codes

occur by different assignments for the codons into one of the amino acids, or a termination codon, or as being unused.

The underlying structure of the genetic code has been suggested to influence mRNA sequences through evolutionary selection of RNA secondary structures (SEFFENS and DIGBY, 1999). That study found that mRNA sequences are thermodynamically more stable than expected for secondary structure folding. Calculated folding free energies are 80% more likely to be more negative than expected compared to mononucleotide shuffled sequences. Randomization while preserving dinucleotide compositions instead yields control sets of nearly the same free energy (WORKMAN and KROGH, 1999). That work demonstrates that preserving dinucleotide composition biases will increase the magnitude of folding free energies in shuffled sequences. Single mutational events that preserve dinucleotide frequency though are extremely rare.

For 79 mRNA sequences selected from a yeast SAGE library, the free energy minimization calculations of native mRNA sequences are also usually more negative than randomized mRNA sequences (SEFFENS *et al.*, 2002). If this yeast SAGE data is grouped according to expression levels, the mean folding free energy bias is different between the high, average, and low expression-level genes. A t-Test for paired two-samples of means showed a significant difference in folding free energies between high and low expression yeast genes (SEFFENS *et al.*, 2002). Thus the sequence of these yeast genes typically give rise to more stable secondary mRNA structures in high expression genes than in single-copy genes.

Genes could then be classified according to whether they are more or less stable in calculated folding free energy compared to mononucleotide-randomized sequences (SEFFENS, 1999). The excess RNA secondary structures may be involved in gene regulation mechanisms, intron splicing, or steady state mRNA levels. Structural elements of mRNA are known to play integral roles in mechanisms regulating translation and mRNA stability, which in turn directly affect translation efficiency and turnover rate of message, and therefore the amount of a specific protein that is synthesized. The regulation of mRNA turnover is an essential step in controlling message abundance and therefore gene expression in cells. Message degradation or stability plays a critical role in cell proliferation or cellular differentiation, and is crucial in mechanisms that maintain normal biological functions of individual cells and tissues. Aberrant mRNA turnover usually leads to altered levels of proteins, which can dramatically modify cellular properties. For example, oncogene or growth factor over-expression is often associated with abnormal cell proliferation and malignant transformation. Since message turnover is an important component of gene regulation, it is not surprising to find that message stability characteristics of key growth regulatory genes are tightly controlled.

Interestingly, an *in silico* study of mRNA secondary structure has found a bias within the coding sequences of genes that favors “in-frame” pairing of nucleotides (SEFFENS and DIGBY, 2000). This pairing of codons in mRNA, each hydrogen bonded with its reverse-complement, can be performed as a theoretical exercise on the 64 codons in the genetic code. This graph-drawing exercise partitions the 20 amino acids into three subsets as represented by a three-component graph. The composition of proteins in terms of amino acid membership in the three subgroups has been measured, and runs of members within the same subgroup have been analyzed. One can identify proteins having very long runs,

or count the number of runs in each protein. The latter category includes a runs-test statistic applied to the number of runs, relative to the number of runs to be expected from a random arrangement of the same elements (WALLIS and ROBERTS, 1956). As compared to randomized versions of the same protein sequences, the distribution of the runs-test statistic over the native protein sequences is negatively skewed (DIGBY *et al.*, 2002). To assess whether this statistical bias was due to a chance grouping of the amino acids in the real genetic code, several alternate groupings were examined by permuting the assignment of amino acids to groups. A metric was constructed to define the difference, or “distance”, between any two such groupings, and an exhaustive search was conducted among alternate groupings maximally distant from the real genetic code, to select sets that were also maximally distant from one another. To determine if this difference between native and randomized protein sequences is unique to the partition based on the genetic graph, other sets of amino acids were examined. Interestingly the calculated skewness for the alternate partition (skewness = -0.210) is less than for the real genetic code partition (skewness = -0.376) from DIGBY *et al.* (2002).

This genetic graph then is constructed by pairing codons with their reverse complement for all 20 amino acids. The codons, in turn, are grouped together according to the respective amino acids for which they code. Lines with the amino acids as vertices on a graph then represent reverse complement relationships. This genetic graph suggest a mathematical symmetry is present which may aid in understanding the evolution of the genetic code and codon usage. This work examines the graph theoretical form of the standard and alternate genetic codes to establish a framework for further work. The graphical nature of the standard genetic code was first noted by an argument considering anti-proteins to explain receptor-ligand modeling in biochemistry (ZULL and SMITH, 1990).

2. Materials and Methods

Analysis of genetic codes utilized standard graph theory methodology (HARARY, 1972). Calculations were performed using Mathematica (Wolfram Research Inc., Champaign, IL) and the discrete mathematics package “Combinatorica” by SKIENA (1990). The notebooks used to generate the genetic graphs and their graph theory analysis are available from the author. The genetic codes were obtained from a list compiled by A. Elzanowski and J. Ostell at the National Center for Biotechnology Information (<http://www3.ncbi.nlm.gov/taxonomy>). This list is based primarily on the reviews by OSAWA *et al.* (1992) and JUKES and OSAWA (1993). The genetic codes are identified by an ID number for use in GenBank and the DNA Data Bank of Japan database annotation (TATENO and GOJOBORI, 1997). The standard genetic code (ID = 1) is utilized in most organisms for translating the information contained in the chromosome(s). The other codes are: vertebrate mitochondrial (ID = 2), yeast mitochondrial (ID = 3), protozoan mitochondrial (ID = 4), invertebrate mitochondrial (ID = 5), ciliate nuclear (ID = 6), echinoderm mitochondrial (ID = 9), euplotid nuclear (ID = 10), bacterial (ID = 11), alternate yeast nuclear (ID = 12), ascidian mitochondrial (ID = 13), flatworm mitochondrial (ID = 14), and blepharisma nuclear (ID = 15).

The graph representation of the genetic code uses vertices for the twenty amino acids, and lines to represent codon-reverse complement codon pairs. Stop codons are not

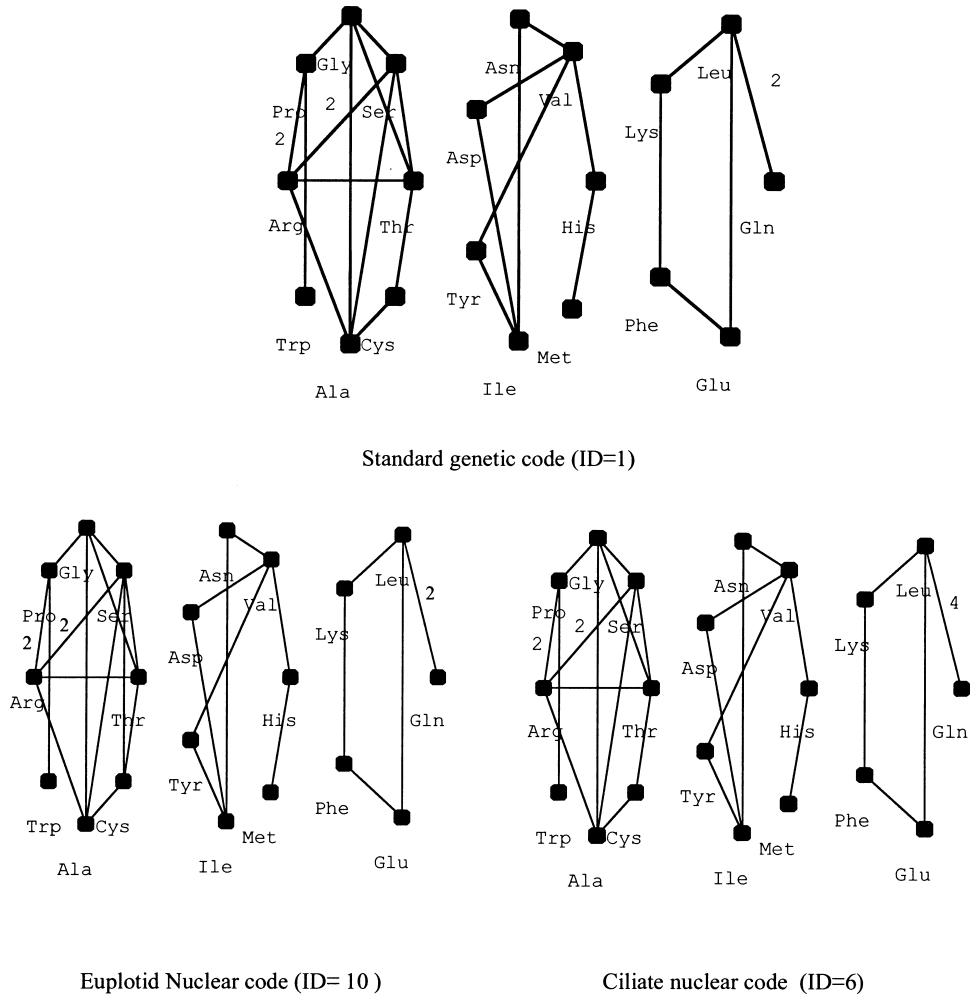


Fig. 1. Nuclear genetic codes with three components. Multilines are shown with numbers.

represented since they do not form a group in the same sense as the amino acids. Hence the reverse complement of the stop codons are not represented in the graphs. This will cause some vertices to have a lower degree or valence than expected by the known number of codons (degeneracy) for each amino acid. This allows a fixed set comparison with the alternate genetic codes where the number of stop codons varies and even some codons are unused (OSAWA *et al.*, 1992; JUKES and OSAWA, 1993). Since ignoring stop codons is equivalent to contracting the genetic graph, non-trivial graph theory properties of the resulting graphs will not change.

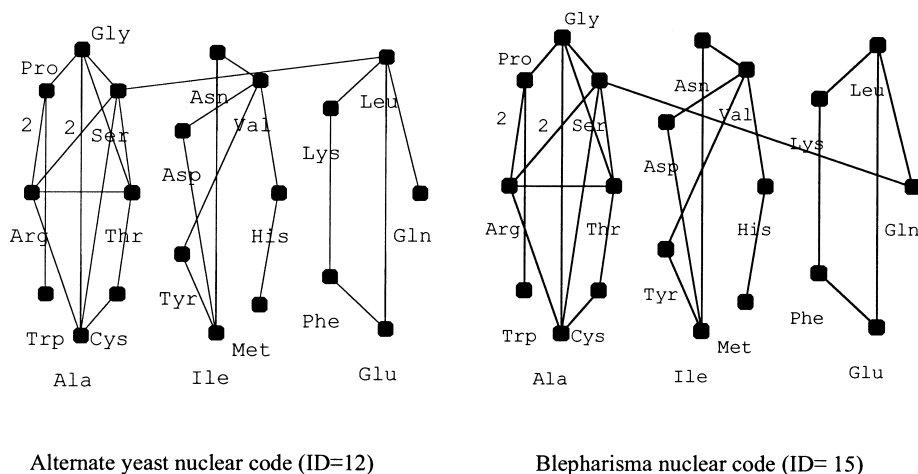
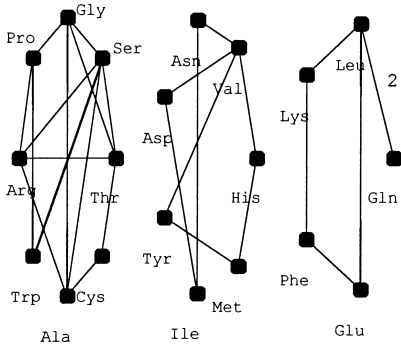


Fig. 2. Nuclear genetic codes that have only 2 connected components. Multilines are shown with numbers.

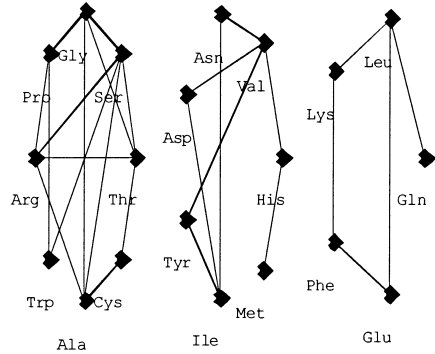
3. Results and Discussion

A graph-drawing exercise can be performed by listing the codons of each amino acid in the genetic code, then drawing a line to link codon: reverse-complement codons together. This exercise results in three independent graph components or families, each comprising a subset of the twenty amino acids (Fig. 1). The three amino acid families can be identified by the member with the greatest degeneracy or graph vertex degree. The largest amino acid family, Serine, contains a C or G at the second codon position, while the Valine and Leucine families contain A or T at this second codon position. This decomposition of the twenty amino acids into three families is a structural feature of the genetic code. If Arginine, Serine and Leucine are excluded for the moment, the degeneracy in the genetic code can be expressed as N_1N_2X , where $X = \{N = (A,C,G,T), \text{ or } Y = (C,T), \text{ or } R = (A,G), \text{ or } H = (A,C,T)\}$. For Arginine and Leucine, the N_2 position is also fixed, and N_1 is one of two different bases. For Serine, the choice for N_2 is the complement of the other (C or G). Since the second codon position is constant (except for Serine) among all synonymous codons, the twenty amino acids (including Serine due to $N_2 = \{C \text{ or } G\}$) must group into at least two families based on the pairing of reverse complement codons. One family contains $N_2 = C$ and G, and the other family contains $N_2 = A$ and T. It is interesting that the $N_2 = A$ and T family has decomposed further into two smaller graphs as shown in Fig. 1 for three nuclear genetic codes.

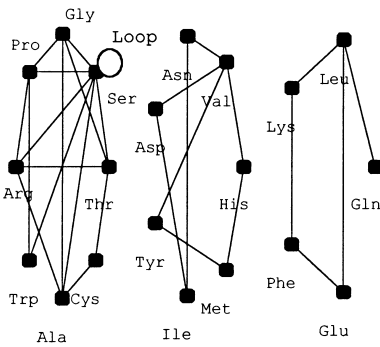
The stop codons, if treated as a coherent group like those coding for amino acids, would still not generate a completely connected graph. Only two components, the Serine and Leucine Family would be connected in this case. The functional significance of stop codons may not require that they be treated as an equivalence group in the same sense as the rest of the genetic code. In fact, they are the least preserved features among the several code variations (OSAWA *et al.*, 1992).



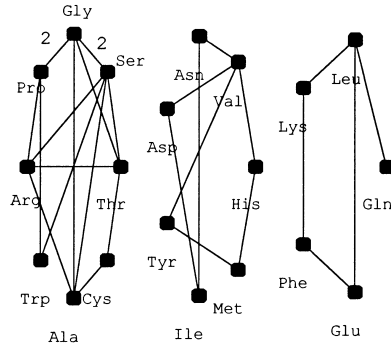
Vertebrate Mitochondrial code (ID= 2)



Protozoan mitochondrial code (ID=4)



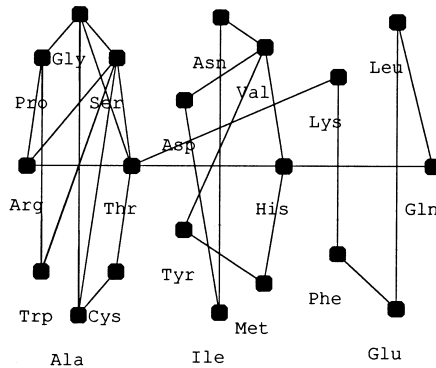
Invertebrate Mitochondrial code (ID= 5)



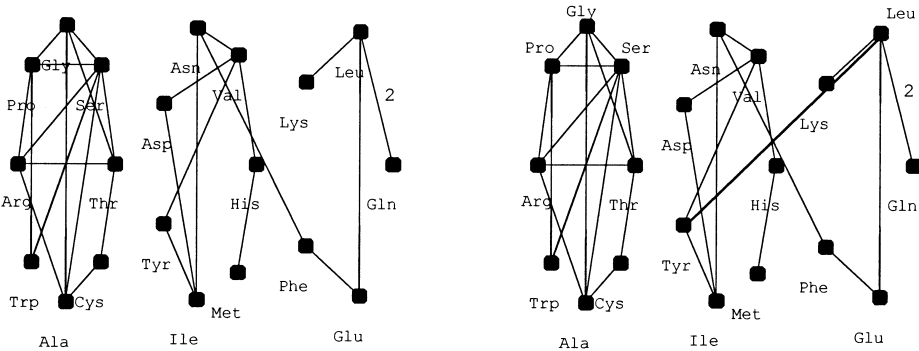
Ascidian mitochondrial code (ID=13)

Fig. 3. Mitochondrial genetic codes with three components. Multilines are shown with numbers, and pseudographs with a small loop.

There are a total of twelve known genetic codes, each identified by an ID number for use in database annotation. The standard genetic code (ID = 1) is utilized in most organisms for translating the information contained in the chromosome(s). The alternate codes are identified with their ID numbers in Sec. 2 (Materials and Methods). The modified genetic codes yield similar multi-component graphs (Figs. 1–4). Note that the Yeast mitochondrial code in Fig. 4 contains a line that appears to include four vertices, but only Arg and Gln are actually incident due to the graphing routine in Mathematica. All of the genetic codes have either two or three components (Table 1). These are identified as “gc(x)Fam(y)”, where (x) is the genetic code ID number and (y) is a component label composed of S, V, or L. All of the genetic code components appear to be built from the three amino acid families in the



Yeast mitochondrial code (ID=3)



Echinoderm mitochondrial code (ID=9)

Flatworm mitochondrial code (ID=14)

Fig. 4. Mitochondrial genetic codes that have only 2 connected components. Multilines are shown with a number.

standard code. These are the Serine Family (gc1FamS), the Valine Family (gc1FamV), and the Leucine Family (gc1FamL). There are five instances where the genetic code forms only two components, the Leucine Family is either joined to the Serine Family (three instances identified as gc(x)FamSL) or to the Valine Family (two instances identified as gc(x)FamVL). These instances are shown in Fig. 2 for the nuclear codes, and Fig. 4 for the mitochondrial codes. If there were a genetic code with Serine joined to Valine family, the distribution of vertices between the two resulting components would be the most uneven of the three possible permutations. This may indicate a biological selective pressure for the components to be of near equal size.

Table 1. Component composition of genetic code graphs.

Genetic code	ID	Components	Multi	Pseudo	Planar	Notes
Standard	1	gc1FamS gc1FamV gc1FamL	T F T	F F F	F T T	
Vertebr. mit.	2	gc2FamS gc2FamV gc1FamL	F F	F F	F T	unmodified
Yeast mit.	3	gc3FamSL gc2FamV	T	F	T	Ser and Leu joined same as code 2
Prot. mit.	4	gc4FamS gc1FamV gc1FamL	T	F	F	unmodified unmodified
Invertebr. mit.	5	gc5FamS gc2FamV gc1FamL	T	T	F	pseudograph same as code 2 unmodified
Ciliate nuc.	6	gc1FamS gc1FamV gc6FamL	T	F	T	unmodified unmodified
Echino. mit.	9	gc5FamS gc9FamVL	T	F	T	same as code 5 Val and Leu joined
Euplotid nuc.	10	gc10FamS gc1FamV gc1FamL	T	F	F	unmodified unmodified
Alt. Yeast nuc.	12	gc12FamSL gc1FamV	T	F	F	Ser and Leu joined unmodified
Ascidian mit.	13	gc13FamS gc2FamV gc1FamL	T	F	F	same as code 2 unmodified
Flatworm mit.	14	gc5FamS gc14FamVL	T	F	T	same as code 5 Val and Leu joined
Bleph. nuc.	15	gc15FamSL gc1FamV	T	F	F	Ser and Leu joined unmodified

Notes: Unmodified refers to Universal genetic code. Abbreviations: Vertebr., Vertebrate; Mit., mitochondrial; Prot., protozoan; Invertebr., invertebrate; nuc., nuclear; Echino., echinoderm; Alt., alternate; Bleph., blepharisma. Multi (multi-graph); and Pseudo (pseudo-graph).

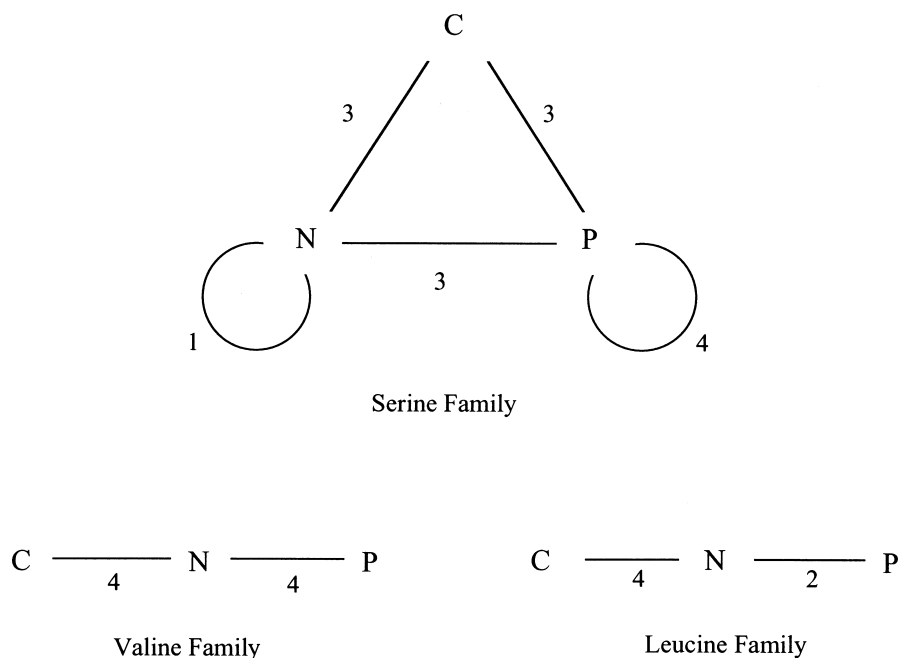


Fig. 5. Relabeled vertices for standard genetic code to show symmetry. Multilines are shown with numbers and pseudographs with a loop.

Except for the vertebrate mitochondrial code ($ID = 2$), all of the examined genetic codes have the Serine and Leucine families configured as multigraphs in the terminology of graph theory (Table 1). A multigraph possesses vertices for which multiple lines connect the same two vertices (HARARY, 1972). The vertebrate mitochondrial graph has only the Leucine family configured as a multigraph. Since the set of amino acids does not change for any of the genetic codes, the number of vertices for the components remains constant, except for the joined components. The number of lines for each component will change depending on the addition or loss of stop codons. The six different Serine components are all non-planar graphs, while the four different Valine and Leucine components are planar. A planar graph can be drawn without any lines crossing. The three Serine joined components retain the non-planarity of the basic Serine Family, and the Valine joined components are still planar.

Few graph theoretical properties are invariant for the genetic codes and their components that could be used to differentiate possible genetic codes from random graphs of the same number of vertices and lines. The genetic graphs are not simple since several components are multigraphs, as mentioned earlier. Even one component, *gc5FamS*, is a pseudograph with a line incident to the same vertex. This would be represented as a loop on that vertex. The genetic graphs all possess cycles, several components are Hamiltonian and some are also Eulerian. A Hamiltonian graph has a cycle or path that visits each vertex exactly once,

while an Eulerian graph has a path that includes each line exactly once. Some of the Serine components are biconnected, meaning that the graph has no articulation vertices, that is, vertices whose removal will disconnect the graph.

An interesting graph theory property that has a biological interpretation is that all of the graphs are either bipartite or nearly bipartite. A graph is bipartite if the vertices can be partitioned into two disjoint sets such that there exists no edge between vertices of different sets. An alternate definition is that a graph is bipartite if and only if all cycles are of even length. This graph property is related to the genetic code having complement pairs, A with T, and C with G, in the fundamental composition of DNA and, with minor modification, in RNA. The smaller Valine and Leucine components are bipartite, while the Serine components have added lines that make them non-bipartite.

The decomposition of the twenty amino acids into the three independent families has unique properties that are related to symmetry breaking in graphs. Each of the three families includes charged, polar, and non-polar members (ZULL and SMITH, 1990). This property suggests that it is possible to find proteins comprised of amino acids from only a single family. A preliminary survey has identified proteins that are over 90% from the Serine family (data not shown). Graph symmetry of the genetic codes becomes evident upon a relabeling of the vertices. The construction of the genetic graphs in Figs. 1–4 involves the grouping of codons into sets that are represented by vertices. Each vertex contains codons that code for a unique amino acid. The amino acids can then be grouped into biochemical functional sets that again can be represented by vertices. This process then involves a contraction of the vertices of the genetic graphs. If three biochemical groups are used, a three-vertex graph will result. As an example, the three biochemical groups could be polar (Ser, Gly, Thr, Cys, Asn, Tyr, Gln), nonpolar (Ala, Pro, Trp, Val, Ile, Met, Leu, Phe), and charged (Arg, Asp, His, Glu, Lys). The resulting graph contraction of the standard genetic code is shown in Fig. 5. The three graph components are multi- and pseudo-graphs that possess considerable symmetry. Therefore as the identity of the amino acids that are coded by the codons becomes revealed, the symmetry in the genetic codes is broken.

The graph theory properties of the genetic graphs may be useful to correlate with biological parameters to uncover new relationships. Graph theory properties that differentiate the vertices (amino acids) into subsets, such as those vertices that form bridges, may be useful to classify protein sequences. There may be protein features or properties (KARLIN and BUCHER, 1992) for which correlation to graph theory properties may have a useful biological interpretation. Biological properties such as the propensity of amino acids to appear between protein domains (SEFFENS, 1994), amino acid frequency of use, sets of amino acids at certain positions in a gene (such as signal sequences or intron/exon boundaries), could exhibit correlation to graph theoretic sets.

This graph theory treatment may help to uncover the origins of the genetic code and some of its properties. Since the unraveling of the genetic code a large number of proposals have attempted to explain its origin and evolution (CRICK, 1968). The problem of codon-to-amino-acid assignment has been viewed in terms of chemical interactions (WOESE, 1973), whereas CRICK (1968) postulated that the code was the result of a frozen accident at an evolutionary stage from which any change in the code was prohibited for an ecosystem of protoorganisms. Even with the discovery of the alternate genetic codes in mitochondria

and some organisms (OSAWA and JUKES, 1988), the frozen accident hypothesis appears reasonable. The discovery of catalytic RNA (CECH, 1990) leads to the inference of a protobiotic RNA world and adds to the possible codon assignment mechanism. Such an RNA-determined scenario suggests that, in the transition to a protein-assisted replication, many codes could have coexisted, resulting in selective competition (DIGBY and SEFFENS, 1999).

This graph theory treatment of the genetic codes is meant to define the code in mathematical terms, and to outline a means of uncovering biological relationships in DNA or protein sequences. If the genetic codes can be sufficiently defined or characterized, one could study all possible genetic codes to determine if the current ones are a frozen accident or if they are optimal in some thermodynamic parameter. It is hoped that this communication will stimulate a discussion between mathematicians and molecular biologists for this end.

4. Conclusion

When codons are graphically paired to their reverse complement codons, the twenty amino acids group into three independent families. All of the alternative genetic codes preserve sets of disjoint subgraphs. Data on mRNA folding has supported the biological relevance of the graph representation of the genetic code. A most important support being an *in silico* study of mRNA secondary structure that found a bias within the coding sequences of genes that favor in-frame pairing of the nucleotides (SEFFENS and DIGBY, 2000). The pairing of codons with their reverse complement in mRNA stem-loop structures can be performed as a graph drawing exercise on all codons in the genetic code. This graph drawing exercise generates a three-component graph, thus partitioning the amino acids into three groups. Eight amino acids are in one group, while seven and five are in two smaller groups. A search of 430,000 proteins found sequences with statistically long and short runs of amino acids in one of the graph theory groups (DIGBY *et al.*, 2002). The mRNA sequences corresponding to these unusual proteins were *in silico* folded and compared to randomized sequences to calculate a folding bias score. Proteins with long runs of amino acids from one group were coded by mRNAs that had large negative scores. These mRNAs thus possessed greater potential for forming secondary structures than expected by chance. Proteins with short runs of amino acids from one group were coded by mRNAs that had small positive folding scores. This supports the biological relevance for the graph theory partition of the genetic code and the three groups of amino acids. This suggests a graph symmetry that may aid in understanding the evolution of the genetic code and codon usage. This work presents a tool, graph theory, to ask new questions concerning the genetic code.

This work was supported (or partially supported) by NIH grant GM08247, by a Research Centers in Minority Institutions award, G12RR03062, from the Division of Research Resources, National Institutes of Health and NSF CREST Center for Theoretical Studies of Physical Systems (CTSPS) Cooperative Agreement #HRD-9632844

REFERENCES

- CECH, T. (1990) Self-splicing and enzymatic activity of an intervening sequence RNA from Tetrahymena, *Angew. Chem. Int. Rd. Engl.*, **29**, 759–768.

- CRICK, F. (1968) The origin of the genetic code, *J. Mol. Biol.*, **38**, 367–379.
- DIGBY, D. and SEFFENS, W. (1999) Evolutionary algorithm analysis of the biological genetic codes, in *Proceedings of 1999 Genetic and Evolutionary Computation Conference*, Orlando, FL, 1440 pp.
- DIGBY, D., ABEBE, F. and SEFFENS, W. (2002) Runs of amino acids are longer than expected in proteins based on a graph theory representation of the genetic code, *J. Biological Systems*, **10**(4), 10–23.
- HARARY, F. (1972) *Graph Theory*, Addison-Wesley, Canada.
- JUKES, T. H. and OSAWA, S. (1993) Evolutionary changes in the genetic code, *Comp. Biochem. Physiol.*, **106B**, 489–494.
- KARLIN, S. and BUCHER, P. (1992) Correlation analysis of amino acid usage in protein classes, *Proc. Natl. Acad. Sci. USA*, **89**, 12165–12169.
- NAKAMURA, Y., GOJOBORI, T. and IKEMURA, T. (1997) Codon usage tabulated from the international DNA sequence databases, *Nucleic Acids Research*, **25**, 244–245.
- OSAWA, S. and JUKES, T. (1988) Evolution of the genetic code as affected by anticodon content, *Trends Genet.*, **4**, 191–198.
- OSAWA, S., JUKES, T. H., WATANABE, K. and MUTO, A. (1992) Recent evidence for evolution of the genetic code, *Microbiol. Rev.*, **56**, 229–264.
- SEFFENS, W. (1994) Calculated flexibility correlates with linker preference between protein domains, *J. Protein Chemistry*, **13**, 487–488.
- SEFFENS, W. (1999) mRNA classification based on calculated folding free energies, *WWW Journal of Biology* (<http://epress.com/w3jbio/>), Vol. 4-3.
- SEFFENS, W. and DIGBY, D. (1999) mRNAs have greater calculated folding free energies than shuffled or codon choice randomized sequences, *Nucleic Acids Research*, **27**, 1578–1584.
- SEFFENS, W. and DIGBY, D. (2000) Gene sequences are locally optimized for global mRNA folding, in *Optimization in Computational Chemistry and Molecular Biology* (eds. C. A. Floudas and P. M. Pardalos), pp. 131–140, Kluwer Academic Publishers, NL.
- SEFFENS, W., HUD, Z. and DIGBY, D. (2002) Yeast SAGE expression levels are related to calculated mRNA folding free energies, in *Biocomputing* (eds. P. Pardalos, J. Principe and S. Rajasekaran), pp. 175–184, Kluwer Academic Publishers.
- SKIENA, S. (1990) *Implementing Discrete Mathematics*, Addison-Wesley Co., Redwood City, CA.
- TATENO, Y. and GOJOBORI, T. (1997) DNA data bank of Japan in the age of information biology, *Nucleic Acids Research*, **25**, 14–17.
- WALLIS, W. and ROBERTS, H. (1956) *Statistics: A New Approach*, Addison-Wesley, Canada.
- WOESE, C. (1973) Evolution of the genetic code, *Naturwissenschaften*, **60**, 447–459.
- WORKMAN, C. and KROGH, A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution, *Nucleic Acids Res.*, **27**(24), 4816–4822.
- ZULL, J. E. and SMITH, S. K. (1990) Is genetic code redundancy related to retention of structural information in both DNA strands?, *Trends Biochem. Sci.*, **15**, 257–261.