Several remarks on the metric space of genetic codes

David Weisman*

Department of Biology,
University of Massachusetts Boston,
100 Morrissey Blvd., Boston,
Massachusetts 02125, USA
E-mail: David.Weisman@acm.org
*Corresponding author

Dan A. Simovici

Department of Computer Science,
University of Massachusetts Boston,
100 Morrissey Blvd., Boston,
Massachusetts 02125, USA
E-mail: dsim@cs.umb.edu

Abstract: A genetic code, the mapping from trinucleotide codons to amino acids, can be viewed as a partition on the set of 64 codons. A small set of non-standard genetic codes is known, and these codes can be mathematically compared by their partitions of the codon set. To measure distances between set partitions, this study defines a parameterised family of metric functions that includes Shannon entropy as a special case. Distances were computed for 17 curated genetic codes using four members of the metric function family. With these metric functions, nuclear genetic codes had relatively small inter-code distances, while mitochondrial codes exhibited greater variance. Hierarchical clustering using Ward's algorithm produced a tight grouping of nuclear codes and several distinct clades of mitochondrial codes. This family of functions may be employed in other biological applications involving set partitions, such as analysis of microarray data, gene set enrichment and protein–protein interaction mapping.

Keywords: non-standard genetic codes; metric space; set partition; Shannon entropy; data mining; clustering; classification; discretisation; Gini index.

Reference to this paper should be made as follows: Weisman, D. and Simovici, D.A. (2012) 'Several remarks on the metric space of genetic codes', *Int. J. Data Mining and Bioinformatics*, Vol. 6, No. 1, pp.17–26.

Biographical notes: David Weisman is a PhD candidate in the Department of Biology at University of Massachusetts Boston. His current research is in the areas of computational biology, synthetic and systems biology, and the laboratory-directed evolution of a receptor protein. Prior to his biological research, he consulted in compiler and language design, operating system development, distributed systems architecture, and secure network protocols.

Dan A. Simovici is a Professor of Computer Science at the University of Massachusetts Boston and an Associate of Dana-Farber Cancer Institute in Boston. He obtained his PhD in Mathematics from the University of Bucharest, Romania in July 1974. His current research is in the area of data mining.

He has published more than 120 research papers, as well as several books including Mathematical Tools for Data Mining (Springer 2008). In addition, he has served on program committees of the major data mining conferences.

1 Introduction

The genetic code, the mapping from trinucleotide codons to the amino acids, is a central feature of present-day biological systems. As the code was elucidated, its highly regular organisation became plainly manifest, raising fundamental questions surrounding the code's origin and evolution. Most prominently, the mapping appears to reduce the harmful effects of point mutations and mistranslations (Woese, 1965). For example, all codons with uracil in the second position translate to hydrophobic amino acids, thereby providing some resilience to mutations in the first and third positions. Similarly, much of the code redundancy is organised to allow nucleotide changes in the third codon 'wobble' position without changing the resulting amino acid.

A variety of computational experiments have quantified genetic code robustness to DNA point mutations. These experiments typically generate a non-standard code and then measure the burden of random DNA point mutations on a translated protein, modelling the mutational cost as a function of changes in a biophysical amino acid property such as hydrophobicity or molecular volume. In an early work, Alff-Steinberger (1969) tested 200 codes and found that the standard code was significantly more robust than the random codes when measured by tolerance to mutations in the first and third codon positions. Freeland and Hurst (1998) further explored robustness by generating 10^6 variant codes that preserved the standard degeneracy pattern, and tested these codes against a mutational model with parameterised probabilities of transition and transversion errors. Their work found that the vast majority of random amino acid assignments were inferior to the standard code. More recently, Itzkovitz and Alon (2007) examined the ability of the standard code to represent additional non-coding information such as transcription factor binding sites and mRNA secondary structure, and found that the standard code was far more capable than the majority of perturbed codes. This work also found that the stop codon assignments of the standard code were nearly optimal for truncating proteins following frame shift errors. It is certainly plausible that these properties are biologically advantageous and contributed to code evolution.

A variety of hypotheses attempt to describe the origin of the genetic code. In one of the earliest models, Crick (1968) proposed that primitive genetic codes translated a subset of amino acids and eventually grew to specify the full set. Crick's 'frozen accident' model assumes that any subsequent code evolution would be gravely deleterious to the majority of proteins, and, therefore, highly unlikely to occur. In another model, Wong (1975) argued that the code organisation reflected the emergence of new amino acid biosynthetic pathways, by observing that neighbouring codons frequently represent amino acids related by pathway. More recently, Vetsigian et al. (2006) proposed that in the early stages of protein-based biology, the code became standardised as a consequence of horizontal gene transfer, in which the ability to share genetic information between species was selectively advantageous to the group. Such information sharing is possible only when group members have compatible genetic codes, and this information sharing, along with code fitness, drove evolution towards a standard code.

The discovery of alternate codes in present-day organisms invalidated Crick's frozen limitation (Knight et al., 2001), particularly because identical non-standard codes have emerged more than once. For example, AUA normally codes for isoleucine but specifies methionine in some metazoan and fungal mitochondria, which do not share a common ancestor having this variant code. Sixteen alternate codes have been curated and others have been hypothesised through computational analysis of mitochondrial genomes (Abascal et al., 2006). All reassigned triplets have been observed in various species' mitochondria, but only a subset of these triplets has been reassigned in nuclear genetic codes. For a given triplet codon, a reassignment is not necessarily universal; for example, UGA is normally a stop codon but has been reassigned to both cysteine and tryptophan. These data suggest that general principles permit code evolution along initially similar trajectories. At the same time, it appears that mitochondria are relatively freer to explore a slightly broader range of evolutionary paths.

To accommodate the existence of multiple present-day genetic codes, several hypotheses propose evolutionary mechanisms that obviate the harsh limitation of the frozen accident model. The codon capture hypothesis (Osawa et al., 1992) posits that a shift to high G-C or A-T genomic content reduces the frequency of certain codons, eventually allowing these codons to become unassigned. In that state, relatively neutral evolution is free to explore reassignment of these codons to other amino acids, eventually settling on a beneficial assignment. Another hypothesis, the ambiguous intermediate model, proposes that a particular codon can represent more than one amino acid without causing fatal effects to the proteome. Then, selective pressures on the translational machinery eliminate the ambiguity, resulting in the codon specifying a single amino acid. Evidence supporting this model comes from *Candida*, in which the CUG codon predominantly translates to serine but also produces small amounts of leucine (Suzuki et al., 1997). Broadly supporting the notion of code evolvability, Wong (1983) isolated a strain of *Bacillus subtilis* that preferentially translates 4-fluorotryptophan into its proteome rather than standard tryptophan.

The set of genetic codes can also be viewed from a mathematical perspective. Given the set C of 64 trinucleotide codons, and given the union A of 20 amino acids along with the stop signal, each genetic code defines a mapping $r : C \to A$. A genetic code uniquely partitions C into a set of codon blocks, with each block representing a particular $a \in A$. Given a distance metric on two partitions of C, genetic codes can be compared and clustered to elucidate their differences. The remainder of this paper is as follows. First, we develop a family of parameterised distance functions based on a generalisation of entropy. Next, we apply several members of this family to a set of genetic codes found in extant organisms. Finally, we demonstrate that these distance functions reveal biologically interesting clusters of codes.

2 The metric space of finite functions

The notion of equivalence relation and partition is used repeatedly in this section. We refer the reader to Simovici and Djeraba (2008) for a comprehensive treatment of these concepts.

The set of equivalence relations on a set S will be denoted by $\mathsf{EQ}\ (S)$. Let S, T be two finite sets and let $f : S \to T$ be a function. The *kernel equivalence* of f is defined as

$$\mathbf{ker}(f) = \{(x, y) \in S^2 \mid f(x) = f(y)\}.$$

We denote by π_f the partition of the set S that corresponds to the equivalence relation $\mathbf{ker}(f)$. The blocks of the partition π_f are the equivalence classes of $\mathbf{ker}(f)$ and have the form

$$[x] = \{y \in S \mid f(y) = f(x)\}.$$

Each such block corresponds to a value $f(x)$ of the function and any two distinct blocks are disjoint.

Two functions $f : S \to T$ and $g : S \to T$ have the same kernel equivalence, i.e., $\mathbf{ker}(f) = \mathbf{ker}(g)$ if and only if there exists a permutation p of T such that $f = pg$. Indeed, if such a permutation exists, then $f(x) = f(z)$ if and only if $g(x) = g(z)$ for all $x, z \in S$. Conversely, if $\mathbf{ker}(f) = \mathbf{ker}(g)$ we define the mapping $p : T \to T$ as

$$p(t) = \begin{cases} g(x) & \text{if there is } x \in S \text{ such that } f(x) = t, \\ t & \text{if } x \notin f(S) \end{cases}$$

for $t \in T$. It is easy to verify that p is a well-defined permutation of the set T and that $pf(x) = g(x)$ for every $x \in S$. Thus, an equivalence relation on S characterises a class of functions from S to T such that any of these functions can be obtained from any other function by a permutation of its values.

Define the function $d : (\mathsf{EQ}\,(S))^2 \to \mathbb{N}$ as

$$d(\rho, \sigma) = |\rho - \sigma| + |\sigma - \rho|,$$

for $\rho, \sigma \in \mathsf{EQ}\,(S)$. In other words, $d(\rho, \sigma)$ equals the cardinality of the symmetric difference of the relations ρ and σ, i.e., the number of pairs in ρ but not in σ, plus the number of pairs in σ but not in ρ. It is easy to verify the following
properties:

i $d(\rho, \sigma) = 0$ if and only if $\rho = \sigma$

ii $d(\rho, \sigma) = d(\sigma, \rho)$

iii $d(\rho, \sigma) \leq d(\rho, \tau) + d(\tau, \sigma)$,

for every $\rho, \tau, \sigma \in \mathsf{EQ}\,(S)$. Thus, d is a metric on $\mathsf{EQ}\,(S)$. Starting from d, we can define a semimetric on the set of functions T^S by

$$D(f, g) = d(\mathbf{ker}(f), \mathbf{ker}(g)),$$

for $f, g : S \to T$. Note that D is only a semimetric because $D(f, g) = 0$ implies only that $\mathbf{ker}(f) = \mathbf{ker}(g)$, i.e., f and g can be obtained from each other composition with a permutation, as we have shown earlier.

The value of the semimetric $D(f, g)$ can be expressed using the partitions $\pi_f = \{B_1, \ldots, B_m\}$ and $\pi_g = \{C_1, \ldots, C_n\}$ of the functions f and g. Indeed, since

$$|(\mathbf{ker}(f) - \mathbf{ker}(g)) \cup (\mathbf{ker}(g) - \mathbf{ker}(f))|$$
$$= |\mathbf{ker}(f)| + |\mathbf{ker}(g)| - 2 \cdot |\mathbf{ker}(f) \cap \mathbf{ker}(g)|,$$

we have

$$D(f,g)=\sum_{i=1}^{m}|B_i|^2+\sum_{j=1}^{n}|C_j|^2-2\cdot\sum_{i=1}^{m}\sum_{j=1}^{n}|B_i\cap C_j|. \tag{1}$$

Let S be a finite set and let $\pi=\{B_1,\ldots,B_m\}$ be a partition of S. The *Shannon entropy* of π is the number

$$\mathcal{H}(\pi)=-\sum_{i=1}^{m}\frac{|B_i|}{|S|}\log_2\frac{|B_i|}{|S|}.$$

The *Gini index* of π is the number

$$\text{Gini}(\pi)=1-\sum_{i=1}^{m}\left(\frac{|B_i|}{|S|}\right)^2.$$

Both numbers can be used to evaluate the uniformity of the distribution of the elements of S in the blocks of π because both values increase with the uniformity of the distribution of the elements of S.

It is interesting to observe that an equivalent metric can be obtained starting from generalised entropies (Daróczy, 1970; Havrda and Charvat, 1967). The algebraic axiomatisation of partition entropy was done by Jaroszewicz and Simovici (1999), and various applications of Shannon and generalised entropies in data mining were considered by Simovici and Jaroszewicz (2000, 2003).

Let π be a partition of S and let $C\subseteq S$. Denoted by π_C, the 'trace' of π on C is given by

$$\pi_C=\{B\cap C\,|\,B\in\pi \text{ such that } B\cap C\neq\phi\}.$$

Let π, σ be two partitions of S and let $\sigma=\{C_1,\ldots,C_n\}$. The *$\beta$-conditional entropy* of the partitions π and σ is the function $\mathcal{H}_\beta$ defined by

$$\mathcal{H}_\beta(\pi/\sigma)=\sum_{j=1}^{n}\left(\frac{|C_j|}{|S|}\right)^{\beta}\mathcal{H}_\beta(\pi_{C_j}).$$

For $\pi=\{B_1,\ldots,B_m\}$ and $\sigma=\{C_1,\ldots,C_n\}$, the conditional entropy can be written explicitly as

$$\begin{aligned}\mathcal{H}_\beta(\pi/\sigma)&=\sum_{j=1}^{n}\left(\frac{|C_j|}{|S|}\right)^{\beta}\sum_{i=1}^{m}\frac{1}{1-2^{1-\beta}}\left[1-\left(\frac{|B_i\cap C_j|}{|C_j|}\right)^{\beta}\right]\\&=\frac{1}{1-2^{1-\beta}}\sum_{j=1}^{n}\left(\left(\frac{|C_j|}{|S|}\right)^{\beta}-\sum_{i=1}^{m}\left(\frac{|B_i\cap C_j|}{|S|}\right)^{\beta}\right).\end{aligned}$$

For every $\beta>1$, the mapping d_β defined by

$$d_\beta(\pi,\sigma)=\mathcal{H}_\beta(\pi/\sigma)+\mathcal{H}_\beta(\sigma/\pi)$$

is a metric on the set of partitions of S. An explicit expression of the metric between two partitions can now be obtained using the values of conditional entropies given by the previous equality:

$$d_\beta(\pi,\sigma)=\frac{1}{(1-2^{1-\beta})\,|S|^\beta}\left(\sum_{i=1}^{m}|B_i|^\beta+\sum_{j=1}^{n}|C_j|^\beta\right.$$
$$\left.-2\cdot\sum_{i=1}^{m}\sum_{j=1}^{n}|B_i\cap C_j|^\beta\right), \qquad (2)$$

where $\pi=\{B_1,\ldots,B_m\}$ and $\sigma=\{C_1,\ldots,C_n\}$ are two partitions of S.

In the special case $\beta=2$, we have

$$d_2(\pi,\sigma)=\frac{2}{|S|^2}\left(\sum_{i=1}^{m}|B_i|^2+\sum_{j=1}^{n}|C_j|^2\right.$$
$$\left.-\sum_{i=1}^{m}\sum_{j=1}^{n}2\,|B_i\cap C_j|^2\right),$$

which implies

$$d_2(\pi,\sigma)=\frac{2}{|S|^2}d(\pi,\sigma),$$

where d is the distance introduced by using the symmetric difference in equality (1).

It is interesting to note that when β approaches 1, we have

$$\lim_{\beta\to1}\mathcal{H}_\beta(\pi)=-\sum_{i=1}^{m}\frac{|B_i|}{|S|}\log_2\frac{|B_i|}{|S|},$$

which is precisely the Shannon entropy. Furthermore,

$$\lim_{\beta\to1}d_\beta(\pi,\sigma)=-\sum_{i=1}^{m}\frac{|B_i|}{|S|}\log_2\frac{|B_i|}{|S|}-\sum_{j=1}^{n}\frac{|C_j|}{|S|}\log_2\frac{|C_j|}{|S|}$$
$$+2\cdot\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|B_i\cap C_j|}{|S|}\log_2\frac{|B_i\cap C_j|}{|S|}$$

a metric that was used in de Màntaras (1991) for comparing clusterings.

Using a naïve algorithm to compute equality (2), the running time is bounded by the set intersection term, and is $\mathcal{O}(mn|S|)$.

3 Results and discussion

Using the metric on set partitions described in equality (2), we computed pairwise distance matrices between the 17 genetic codes curated at NCBI. Four distance matrices were derived corresponding to the functions $d_\beta(\pi,\sigma)$ with $\beta\in\{1.01, 2, 3, 5\}$.

To visualise these four distance matrices, we performed metric multidimensional scaling on the data set. These results, shown in Figure 1, indicate that all nuclear-only codes are relatively similar, and these also cluster closely with codes found in both the nucleus and the mitochondria. Conversely, the mitochondrial-only genetic codes exhibit a higher level of diversity. These findings are consistent across different levels of the non-linear parameter β. For all values of β examined, the mitochondrial outliers such as

M3, M14 and M22 remained near the boundaries of the projected space, while the diffusion and relative distances between points varied considerably.

To evaluate equality (2) as a metric for clustering applications, we computed pairwise distances of the genetic codes using $\beta = 3$, and performed hierarchical clustering of the distance matrix using Ward's minimum variance method. Figure 2 shows the resulting heatmap and dendrogram. To validate that the tree topology reflects the pairwise genetic code distances, we computed the Spearman cophenetic correlation between the code distance matrix and the tree, resulting in $\rho = 0.85$. Genetic codes found in the nucleus or mitochondria formed a nearly homogeneous single clade, whereas the mitochondrial-only codes exhibited local clustering as well as greater overall diversity. These results are consistent with the findings from multidimensional scaling.

Figure 1 Distances between genetic codes for several values of the β parameter. Distances are projected onto $\mathbb{R}^2$ via metric multidimensional scaling. Black circles and labels beginning with *M* represent mitochondrial-only codes; blue triangles labelled *N* are nuclear-only codes; and, red squares labelled *B* are codes found in both organelles. The integers are genetic code identifiers defined by NCBI, and are summarised in Figure 2 (see online version for colours)

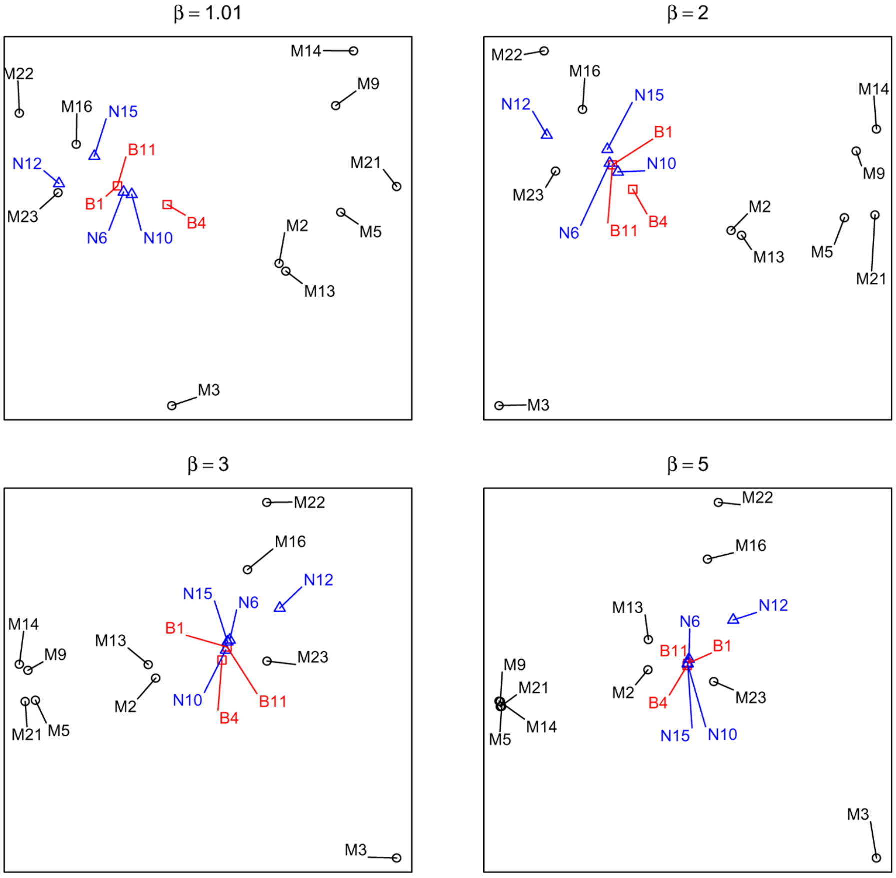
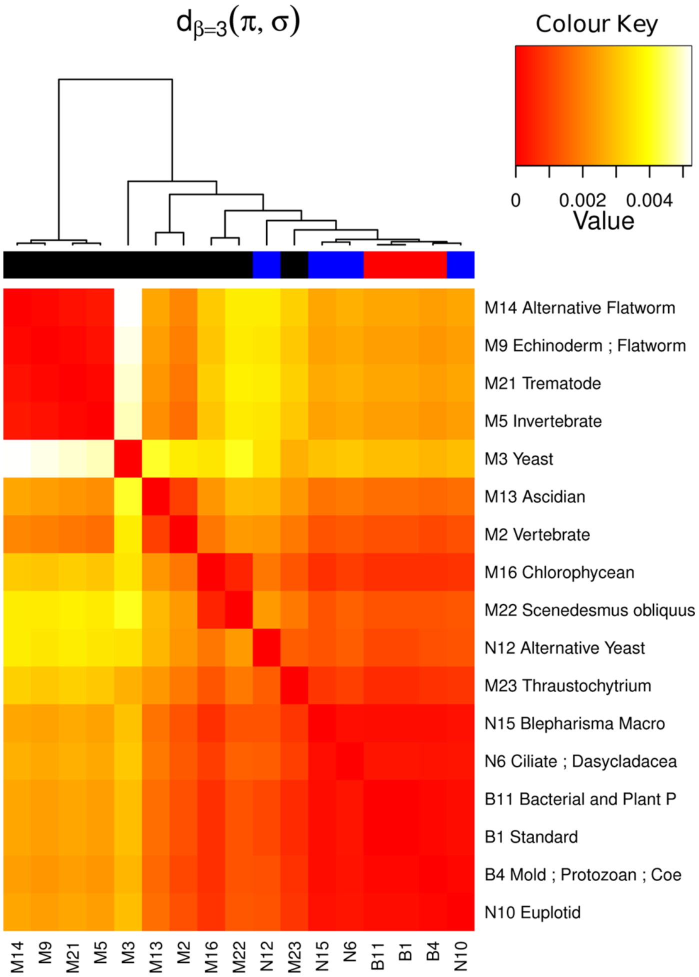


Figure 2 Hierarchical clustering of genetic codes. Genetic code distances are computed with $d_{\beta=3}(\pi, \sigma)$, and clustered using Ward's method. Heatmap colours represent values of *d*. In the top row, blue represents nuclear-only codes, black codes are mitochondrial-only, and red codes are found in both organelles (see online version for colours)



Overall, the results from these experiments support a hypothesis that mitochondria and nuclear genetic codes are under different selection pressures (Knight et al., 2001). The tight clustering of nuclear codes suggests that evolution has relatively little flexibility to modify this mapping. At the same time, if measured by the diversity of pairwise distances, mitochondrial genetic codes appear to have somewhat greater freedom to evolve.

The analysis was performed with 17 known genetic codes. Deeper biological interpretation, however, was limited by this small set. To move beyond this constraint, it would be valuable to mine the immense collection of genomic sequence data and search for evidence of other genetic codes (Abascal et al., 2006). If new codes are found, computing their pairwise distances would cast further light on the biological forces that drive the emergence and evolution of genetic codes.

The family of distance metrics described here can be widely applied in other data mining and bioinformatics research. As discretisation of a variable inherently partitions a set of observations, a number of biological assays fall naturally within this framework. Microarray analysis, mass spectroscopy, protein–protein interaction, as well as metagenomic marker data sets can be interpreted as discretised set partitions, thereby facilitating pairwise comparison, clustering and classification of these biological data.

4 Methods

All computations were performed in *R* version 2.10.1 on generic x86 hardware, running Debian Linux version 5.0. The archive ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz, dated 11/01/2008, was downloaded and the genetic codes file `gencode.dmp` was extracted. A custom Python script reformatted this file into `.csv` for reading from *R*, and added annotations specifying the code location $\{B, M, N\}$. These annotations were manually determined from the NCBI descriptions at http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi and from Knight et al. (2001).

Four sets of pairwise genetic code distances were computed per equality (2) using $\beta \in \{1.01, 2, 3, 5\}$. Hierarchical clustering was performed by the *R* function `hclust` using Ward's method, and multidimensional scaling was performed with the *R* function `cmdscale`. The heatmap was produced with *R* function `heatmap.2`.

An *R* package, `partitionMetric`, implementing this metric is available from the The Comprehensive *R* Archive Network at http://www.r-project.org/.

References

Abascal, F., Posada, D., Knight, R.D. and Zardoya, R. (2006) 'Parallel evolution of the genetic code in arthropod mitochondrial genomes', *PLoS. Biol.*, Vol. 4, No. 5, pp.711–718.

Alff-Steinberger, C. (1969) 'The genetic code and error transmission', *Proc. Natl. Acad. Sci. USA*, Vol. 64, No. 2, pp.584–591.

Crick, F.H. (1968) 'The origin of the genetic code', *J. Mol. Biol.*, Vol. 38, No. 3, pp.367–379.

Daróczy, Z. (1970) 'Generalized information functions', *Inform. and Control*, Vol. 16, pp.36–51.

de Màntaras, R.L. (1991) 'A distance-based attribute selection measure for decision tree induction', *Mach. Learn.*, Vol. 6, pp.81–92.

Freeland, S.J. and Hurst, L.D. (1998) 'The genetic code is one in a million', *J. Mol. Evol.*, Vol. 47, No. 3, pp.238–248.

Havrda, J.H. and Charvat, F. (1967) 'Quantification methods of classification processes: concepts of structural α-entropy', *Kybernetica*, Vol. 3, pp.30–35.

Itzkovitz, S. and Alon, U. (2007) 'The genetic code is nearly optimal for allowing additional information within protein-coding sequences', *Genome. Res.*, Vol. 17, No. 4, pp.405–412.

Jaroszewicz, S. and Simovici, D.A. (1999) 'On axiomatization of conditional entropy', *Proceedings of the 29th International Symposium for Multiple-Valued Logic*, *Freiburg, Germany*, IEEE Computer Society, Los Alamitos, CA, pp.24–31.

Knight, R.D., Freeland, S.J. and Landweber, L.F. (2001) 'Rewiring the keyboard: evolvability of the genetic code', *Nat. Rev. Genet.*, Vol. 2, No. 1, pp.49–58.

Osawa, S., Jukes, T.H., Watanabe, K. and Muto, A. (1992) 'Recent evidence for evolution of the genetic code', *Microbiol. Rev.*, Vol. 56, No. 1, pp.229–264.

Simovici, D.A. and Jaroszewicz, S. (2000) 'On information-theoretical aspects of relational databases', in Calude, C. and Paun, G. (Eds.): *Finite Versus Infinite*, Springer-Verlag, London, pp.301–321.

Simovici, D. and Jaroszewicz, S. (2003) 'Generalized conditional entropy and decision trees', *Extraction et Gestion des connaissances – EGC*, Lavoisier, Paris, pp.363–380.

Simovici, D.A. and Djeraba, C. (2008) *Mathematical Tools for Data Mining*, Springer-Verlag, London.

Suzuki, T., Ueda, T. and Watanabe, K. (1997) 'The 'polysemous' codon: a codon with multiple amino acid assignment caused by dual specificity of tRNA identity', *EMBO J.*, Vol. 16, No. 5, pp.1122–1134.

Vetsigian, K., Woese, C. and Goldenfeld, N. (2006) 'Collective evolution and the genetic code', *Proc. Natl. Acad. Sci. USA*, Vol. 103, No. 28, pp.10696–10701.

Woese, C.R. (1965) 'On the evolution of the genetic code', *Proc. Natl. Acad. Sci. USA*, Vol. 54, No. 6, pp.1546–1552.

Wong, J.T. (1975) 'A co-evolution theory of the genetic code', *Proc. Natl. Acad. Sci. USA*, Vol. 72, No. 5, pp.1909–1912.

Wong, J.T. (1983) 'Membership mutation of the genetic code: loss of fitness by tryptophan', *Proc. Natl. Acad. Sci. USA*, Vol. 80, No. 20, pp.6303–6306.