Biro, J Theor Comput Sci 2013, 1:1
DOI: 10.4172/2376-130X.1000103

Journal of
Theoretical & Computational Science

Review Article Open Access

Suggestion to Complete the Canonical Genetic Code: The Proteomic Code and Nucleic Acid Assisted Protein Folding

Jan Charles Biro*

Homulus Foundation, Los Angeles, CA, USA

Abstract

Background: This research was carried out to provide a summary of a series of bioinformatical observations between 2002 and 2012 concerning the structure of nucleic acids and codons, the interaction between codons and nucleic acids and the general concept of *translation*.

Methods: Public databases and resources, together with assays to determine the free folding energies in various codon residues, were utilized during this study.

Results: This study demonstrates that widely-held paradigms within the field of molecular biology should be modified. In particular, it suggests that codons developed in association with the encoded amino acids, and that wobble bases are not randomly chosen in synonymous codons, since these have well-defined roles in determining the structure of nucleic acids and their folding energies. Furthermore, the proteomic code determines that co-locating amino acids are preferentially encoded by complementary codons (at least at the 1st and 3rd codon positions), and structural information transfer between nucleic acids and proteins during translation requires direct contact between "dedicated" amino acids and their codons. In addition, this study highlights the fact that there is a tRNA cycle that allows the possibility of direct codon amino acid contact.

Conclusions: These observations provide a more complete understanding of the redundant *Genetic Code* and the mechanism of *protein folding*.

Furthermore, the proteomic code provides the first real possibility for scientists to design interacting peptides that have high affinity and specificity for target peptides, with the potential to accelerate the growth of affinity assays and the integration of large numbers of affinity tests into chips.

Keywords: Genetic code; Molecular biology

Abbreviations: ds: double stranded; ss: single stranded; tRNA: transfer RNA; CDS: nucleic acid; FE: Folding Energy

Background

A biological revolution began approximately at the beginning of this millennium, triggered by the simultaneous development of genome sequencing, computational technology and bioinformatics. Fundamental, primary, un-interpreted (un-corrupted?) data began to accumulate in public databases and were available to every scientist. Consequently, a unique possibility opened for biologists to practice their discipline using only "paper and pen (computer)" in a comparable way to mathematicians and theoretical physicists. Previously, biology had been a labor intensive branch of science that required expensive personal and laboratory resources, and the "human factor" (competition for resources) has always been a key factor for success. Unfortunately, the consequence was that fundamental biological concepts were developed around paradigms (and dominant personalities) that have "bent" the objective and scientific views. We recognized that the development of bioinformatics could present researchers, with an exceptional and unprecedented tool that provided the opportunity to re-address (re-evaluate) several fundamental concepts within molecular biology, allowing a fuller understanding to be achieved than has been possible since the structure of DNA was discovered in 1953 and the universal Genetic Code in 1962.

This article is a compact review of approximately 10 years' work and numerous associated peer-reviewed publications. The intention is to provide a comprehensive view of a very large field. Therefore, the reader is respectfully advised to consult the cited, original publications for details.

Molecular Biology

Molecular biology is a young scientific field with a history of approximately 60 years, but it is rich in paradigms, i.e. "*universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of researchers*", and informs us (rather imperatively) *what* is to be observed and scrutinized, the kind of *questions* that should be asked, and *how* the results of scientific investigations should be interpreted. Several of these powerful paradigms were established by Francis Compton Crick, the founding father of molecular biology.

Major paradigms in molecular biology

The historical and recent paradigms are as follows. (1) Biological information is "bites in molecules". (2) A functional distinction between the strands of double stranded (ds) DNA is possible and necessary. (3) Connections between codons and amino acids, if any, are accidental. (4) Codons have no structure. (5) The Genetic Code is redundant, since synonymous codons have the same meaning (i.e. the choice of

***Corresponding author:** Jan Charles Biro, Homulus Foundation, Los Angeles, CA, USA, E-mail: jan.biro@att.net

Received August 24, 2013; **Accepted** October 11, 2013; **Published** October 22, 2013

Citation: Biro JC (2013) Suggestion to Complete the Canonical Genetic Code: The Proteomic Code and Nucleic Acid Assisted Protein Folding. J Theor Comput Sci 1: 103. doi: 10.4172/2376-130X.1000103

Copyright: © 2013 Biro JC. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

wobble bases is accidental) (6) Specific protein-protein interactions are completely different from specific nucleic acid interactions. (7) Codons and encoded amino acids do not interact with each other, and specific interactions between nucleic acids and proteins have nothing to do with codons and encoded amino acids. (8) The structure of DNA is well known, and no new insights into it are to be expected. (9) Translation is like "tape reading": mRNA has no preferred 3D structure. (10) All information for correct protein folding is present in the sequence of amino acids. (11) Codon redundancy is a protection against the unwanted consequences of mutations. (12) Transfer-RNAs are only "adaptors" between codons and amino acids.

In this article, these paradigms will be discussed and modifications suggested, enabling a consistent and meaningful picture of the process of *translation* to be constructed.

Paradigm 1: Definition of biological information

The most important characteristic of information is that it has meaning; "Understanding" discriminates information from data. Biological information concerns signals, usually carried by molecules in 3D forms, which have some biological meaning [1]. The biological information manifests itself through the signal interacting with a receiver (receptor), and the signal-receiver interaction is usually specific and selective. Typical examples of signal-receiver interactions include high affinity and high specificity interactions between receptors and ligands, antigens and antibodies, and the Watson-Crick type of complementary interactions between nucleic acid strands. Complementarity is a very important concept underlying understanding of the nature and processing of biological information. Furthermore, it ensures that biological information has a duality: it exists in duplicates, which are physico-chemical mirror images or molds of each other.

A simple method to measure "information entropy" was proposed by Shannon [2]. He suggested that if the sending device is equally likely to send any one of a set of N messages, then the preferred measure of 'the information produced when one message is chosen from the set' is the base two logarithm of N. This formula is useful for estimating the possible "information content" of biological macromolecules; $N=4^n$ for nucleic acids and $N=20^n$ for proteins, where n is the number of nucleotides or amino acids, respectively. For example, the calculated information content of a 15 nucleotide long nucleic acid is 90 bits, and the corresponding five amino acid long oligopeptide is 64.8 bits (the Shannon formula does not calculate the information content of a message, as information manifests itself only upon reception. However, the formula provides a maximum estimate of possible information in a message if it is received and understood). Figure 1 presents a schematic representation of information processing.

Biological information is often redundant; it exists in several identical or mirror (complementary) copies. Redundancy does not increase the information content of a message. The calculated information content of dsDNA is not twice as much as the information content of its single stranded (ss) DNA variants.

The redundancy of Nirenberg's universal Genetic Code creates a unique dilemma. Twenty amino acids (and stop/start signals) are encoded by 64 codons, which suggest an approximate 3-fold redundancy. However, the synonymous codons are not identical or mirror images of one another; they differ in the third wobble bases. This indicates that codon redundancy is not a simple (or true) repetition of the same messages, but it is the source of additional information. This additional information (that is not necessary for the unambiguous encoding of all amino acids) is as much as 28% of the total information stored and carried in the coding nucleic acid sequences.

An additional characteristic of biological information is that it is "feedback" regulated. The feedback mechanisms ensure that biological signals continue as long as is necessary for reception and interpretation of the signal (i.e. generation of information from the signal).

Paradigm 2: Expression of dsDNA

As recently as 10 years ago, molecular biologists believed and stated that the leading strand of dsDNA is never expressed (i.e. translated into RNA), and that all RNAs are synthesized using the lagging strand. Much time was wasted trying to identify and characterize the distinguishing signature between these complementary but otherwise perfectly identical strands. Only genome-wide sequencing projects convinced the educated public that transcription occurs on each DNA strand (50-50%), and consequently, complementary DNA strands are functionally identical. The stated functional difference concerning DNA strands was analogous to the "The Emperor's New Clothes" and the most embarrassing fallacy of molecular biology.

Paradigm 3: Connection between codons and amino acids

Two alternative hypotheses have been posed to explain the origin of the genetic code. One hypothesis was championed by Woese [3], who argued that there was stereochemical matching, i.e. affinity between amino acids and certain triplet sequences. He proposed that the genetic code developed in a way that was closely connected to the development of the amino acid repertoire, and that this close biochemical connection is fundamental to specific protein–nucleic acid interactions.

The alternative hypothesis was championed by Crick [4], who considered that the basis of the code could be a "frozen accident", with no underlying chemical rationale. Crick [4], the first to suggest and promote the idea of an "adaptor" (transfer RNA; tRNA) between nucleic acids and proteins, furiously attacked any attempt to propose or model any direct codon-amino acid connection.

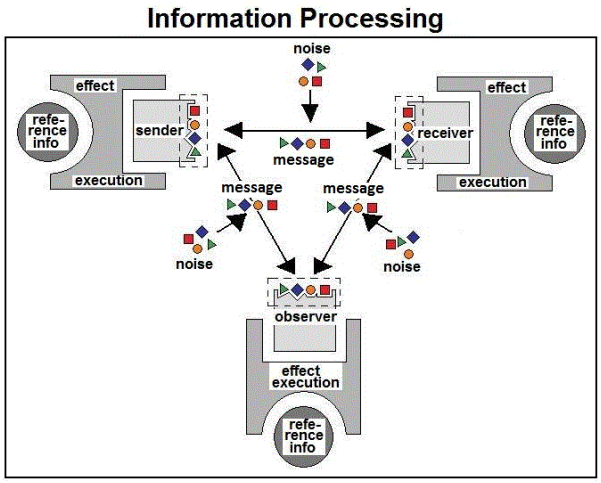

Figure 1: Information processing.

Information (a given order of elements) exists at three places: the sender, receiver and observer. They are spatially separated, but their construction is similar. They uniformly contain reference information to distinguish signals from noise. There is an executive function in each that creates or stores the order in the signal, and is responsible for a response to the message. Noise is the un-ordered occurrence of elements [1].

Citation: Biro JC (2013) Suggestion to Complete the Canonical Genetic Code: The Proteomic Code and Nucleic Acid Assisted Protein Folding. J Theor Comput Sci 1: 103. doi: 10.4172/2376-130X.1000103

However, there is now very strong evidence for a "logical" connection between codons and amino acids, as demonstrated through the construction of a *Common Periodic Table of Codons and Nucleic Acids* [5]. This Table demonstrates the connection between the "RNA World" and "Protein World", and clearly indicates that codons have structure (Figure 2).

Paradigm 4: The structure of codons

The three nucleotides in codons have different functions that clearly distinguish them from one another. First, they have a preferred order of reading that defines the 1st, 2nd and 3rd codon positions. The third "wobble" nucleotides have little importance in defining amino acids and several of them are interchangeable. *The Common Periodic Table of Codons and Nucleic Acids* reveals that 2nd codon nucleotides have a preeminent role in defining the molecular structure and physico-chemical characteristics of the encoded amino acids [5]. The selection of wobble bases is not random. Codon usage frequency tables clearly indicate that synonymous codons are not equally utilized (as would be expected if selection were random). The 1st and 3rd codon positions in exons (but not introns) contain more G or C bases than the 2nd positions. There are three hydrogen bonds between C and G (dG=-1524 kcal/1000 bases), but only two between A and T (dG=-365 kcal/1000 bases). Consequently, the GC-rich 1st and 3rd codon residues contribute more to the thermodynamic force between complementary nucleic acid sequences (including the codon–anticodon interactions) than the 2nd codon residues (this is a statistically derived conclusion that is valid for large numbers of interactions, and not for every interacting codon). This difference between thermodynamic potential of codon residues could be interpreted as a virtual, physicochemical definition of codon boundaries. Such a definition is useful in terms of gaining a better understanding of how the correct codon reading is achieved and translation is protected against frame-shifts [6]. Figure 3 presents the free folding energies for various codon residues.

Paradigm 5: Synonymous codons

Codon Usage Frequency Tables provided further insights into the structure and organization of the universal Genetic Code and coding sequences. These tables are available for more than 100 species, and each suggests that synonymous codons are not equally preferred, i.e. not randomly chosen. In addition, the literature suggests a correlation between the choice of wobble bases in synonymous codons and some structural aspects of encoded peptides. Statistical studies carried out by our research group, where correlations between the frequencies of the four possible bases in the three possible codon positions were searched for using the codon usage tables of the 113 available species, strongly confirm the non-random selection of wobble bases in synonymous codons [7]. The codon usage frequencies indicate the existence of a pan-genomic network of codons, nucleic acids and proteins (Figure 4).

Figure 2: Periodic table of codons and amino acids.

Codons were sorted according to the order, symmetry and complementarity of their bases (left). The corresponding order of amino acids reveals periodicity of the physicochemical properties (polarity, charge, and molecular structure) of the encoded amino acids (right). Note that the periodic tables distinguish four separate fields, each corresponding to the four bases at the central codon positions. The frames of amino acid residues are rooted to the codons (boxes). The names of amino acids are indicated by one and three letters [5].

Citation: Biro JC (2013) Suggestion to Complete the Canonical Genetic Code: The Proteomic Code and Nucleic Acid Assisted Protein Folding. J Theor Comput Sci 1: 103. doi: 10.4172/2376-130X.1000103

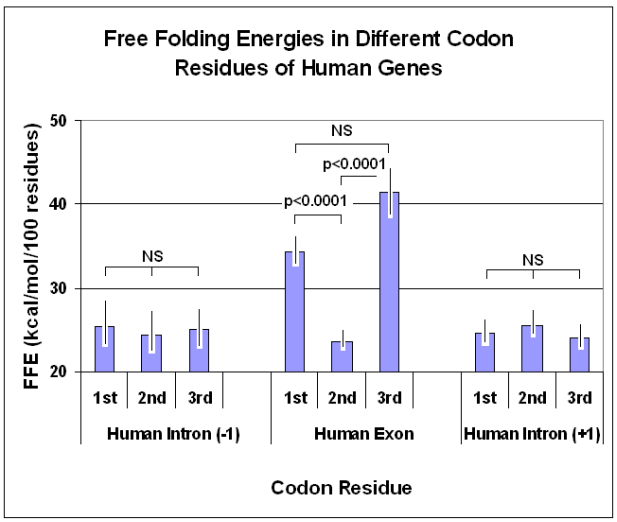

Figure 3: Free folding energies in different codon residues.

Free folding energies (FFE) were determined in phase-selected sub-sequences of 81 genes. The original nucleic acids contained intact three-letter codons (1st+2nd+3rd). Sub-sequences were constructed by periodic removal of one letter from the codon and maintaining the other two (1st+2nd, 1st+3rd, 2nd+3rd), or removing two letters and maintaining only one (1st, 2nd, 3rd). Distinctions were made between exons (B and D) and the preceding (-1, A) and following (+1, C) sequences (introns). The d*G* values were determined using *mfold* and the FFE was calculated. Each bar represents the mean ± SEM, *n*=81 [6].

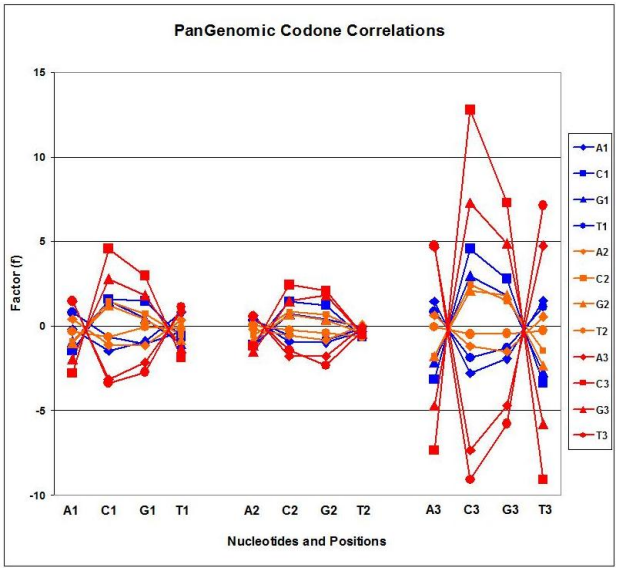

Figure 4: CUF–Pan-genomic codon correlations.

Codon frequencies were collated from 113 Codon Usage Frequency (CUF) Tables and the correlation coefficients (C, 64×64) were calculated. f=-log C. A–the sign was added to indicate negative correlations. The figure presents the f values between 4×4 codon letter combinations in 3×3 codon positions. Each symbol represents the mean f value (n=113). f<-2 and f>2 correspond to statistically significant correlations [7].

The possible existence of such a network is consistent with the view that various functions of an organism are connected to one another in metabolic networks established by specific protein-protein interactions, with a significant conceptual addition: The specific protein-protein interactions (networks) are already present at the CDS (nucleic acid) level.

The usage frequencies of individual codons (including every synonymous codon) are predictable from the usage frequencies of other codons (Figure 5).

Paradigm 6: Residue to residue type interactions

Specific nucleic acid interactions are established by individual complementary bases, i.e. the size and number of possible hydrogen bonds "fit" with each other in AT and GC base pairs, but not in other nucleotide pairs. Residue-to-residue type interactions certainly exist within and between proteins (as parallel and anti-parallel beta sheets). However, their importance in terms of providing specifically interacting structures is not established. Important peptide interactions are suggested to be formed by large, 3D surfaces that are akin to molds of one another. Consequently, the primary functional form is two dimensional (a string) for nucleic acids and 3D (spatial) for proteins. This fundamental, functional distinction between nucleic acids and proteins (well-founded or not) creates a serious methodological problem: 3D interactions are very difficult to study using recent computational methods. Therefore, it was theoretically and methodologically important to bypass this problem and "reduce" 3D interactions to a series of 2D (sequential) interactions, as even large and complex special interactions are formed by small and single amino acid to amino acid interactions (Figures 6 and 7).

To study the rules of residue to residue type interactions in proteins, a tool called SeqX was developed [8]. The tool provides statistical data concerning amino acid co-locations in known protein structures.

Amino acid co-location studies [9] confirmed that charge and hydropathy compatibility rules apply at the individual amino acid level (i.e. hydrophobe attracts hydrophobe, hydrophile attracts hydrophile, and opposite charges attract (Figure 8). However, the existence of size compatibility (i.e. large amino acids attracting small amino acids) between individual co-locating amino acids is a novel observation that further focuses attention on amino acid complementarity (broadly analogous to the base complementarity in nucleic acids).

Paradigm 7: The codon amino acid interaction

The existence of connection between codons and the physico-chemical properties of encoded amino acids is well supported by *The Common Periodic Table of Codons and Nucleic Acids* [5]. This connection suggests co-evolution and possible specific spatial compatibility between codons and amino acids. This question was studied in relation to well-known examples of highly specific protein-nucleic acid interactions provided by restriction endonucleases (RE) and their nucleic acid cut sites (RS) [10]. Such studies confirmed that codons and encoded amino acids preferentially co-locate with one another in these structures, suggesting a stereochemical connection between nucleic acids and proteins (Figure 9).

Paradigm 8: Alternative structures of nucleic acids

An elegant molecular model of DNA was provided by Watson, Crick, Wilkins and Franklin in 1953. The model was experimentally confirmed and honored with the prestigious Nobel Prize in 1962. This model has confirmed the structure of nucleic acids for the last 60 years, with the exception of interesting and rare structural variations including Z-DNA. However, there are major problems concerning ds-DNA. The structure is "closed" with the bases facing towards the central axis of the spiral: the sequence information is "protected". Another major problem concerns how the spiral structure is read during translation without breaking the chain of nucleotides. Consequently, ds-DNA is a logical

Citation: Biro JC (2013) Suggestion to Complete the Canonical Genetic Code: The Proteomic Code and Nucleic Acid Assisted Protein Folding. J Theor Comput Sci 1: 103. doi: 10.4172/2376-130X.1000103

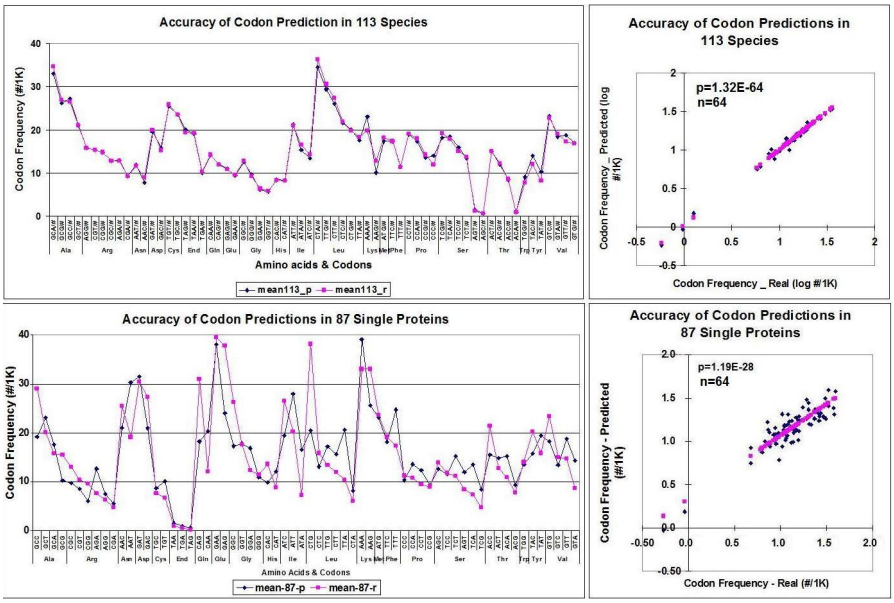

Figure 5: Accuracy of codon predictions in species and proteins.

Codon frequencies were predicted in 113 species (A, B) and in 87 individual proteins (C, D). The average real (r) and predicted (p) codon frequencies were plotted (A, C) and correlations were analyzed (B, D) [7].

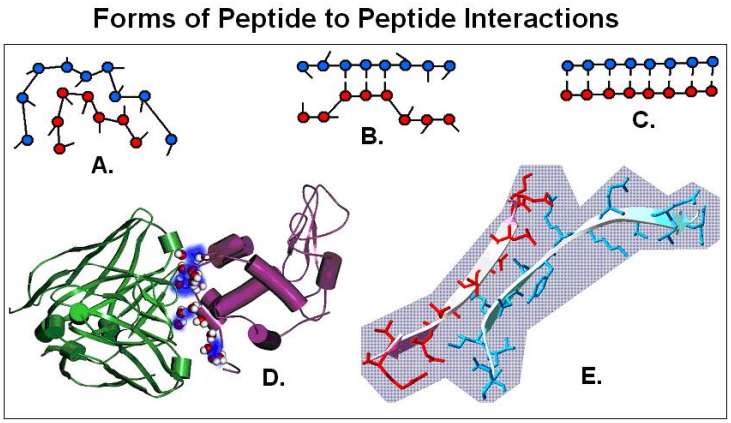

Figure 6: Forms of peptide to peptide interactions.

There are two main forms of amino acid co-locations (interactions), the "docking form" (A, D) and "residue-to-residue form" (C, E). Intermediate form is indicated by B.

and well suited structure for storing valuable genetic information. However, the structure makes no sense in the context of information expression. Therefore, a series of molecular modeling studies were carried out to research and reconsider the earlier idea from F. Crick and L. Pauling that DNA might be an opened, non-helical structure with outward-facing bases, suited to expressing its sequential information. A classical, simple, manual modeling method was utilized, which is more suitable for human perception than computerized 3D model building. This allowed us to observe several stereochemical alternatives to the double spiral (Figure 10), which are more suitable for expression (and storage) of genetic information [11].

There is an immediate concern when presented with these science-fiction-like models: how is it possible that these forms have never been observed previously, when nucleic acids have been at the center of research for half a century? One possible answer is that these non-helical, alternative structures appear very fragile compared with the canonical model and they do not exist in pure DNA extracts, possibly because they require structural support from proteins to maintain structural integrity. Closer examination of gently extracted nucleoproteins might reveal these "functional" nucleic acids variants.

Paradigm 9: The 3D structure of mRNA

According to research, mRNA does not have any structure, as the recent concept of translation is stated as being analogous to a tape recording: mRNA freely passes through ribosomes, where it is read and translated into protein; or on the codon to amino acid basis, with the help of tRNAs as adaptors. Some mRNA structure is not necessary for this process, and some view mRNA as an unnecessary complication that slows down translation and reduces protein yield. However, thermodynamic studies oppose this view and determine that mRNA does have 3D structure [12]. During this study, it was demonstrated that the FE (folding energy, dG) associated with coding sequences is

Citation: Biro JC (2013) Suggestion to Complete the Canonical Genetic Code: The Proteomic Code and Nucleic Acid Assisted Protein Folding. J Theor Comput Sci 1: 103. doi: 10.4172/2376-130X.1000103

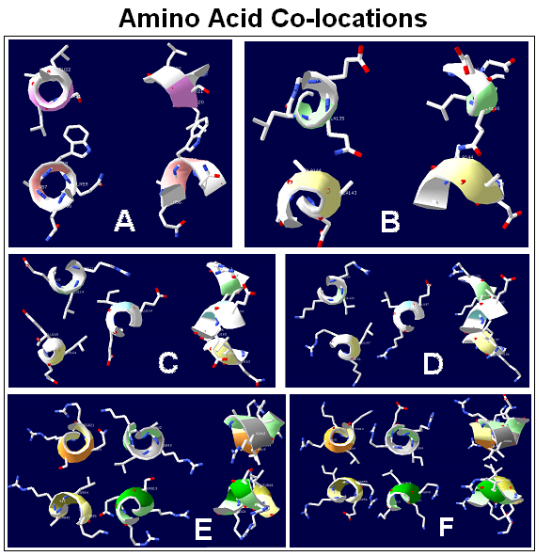

Figure 7: Amino acid co-locations.

Examples of residue-to-residue type amino acid co-locations from real proteins listed in the Protein Data Base (PDB). The pictures illustrate the co-location (interaction) of two (A, B), three (C, D) and four (E, F) amino acids. Pictures present the axial (left part) and side views (right part) of the co-locations.

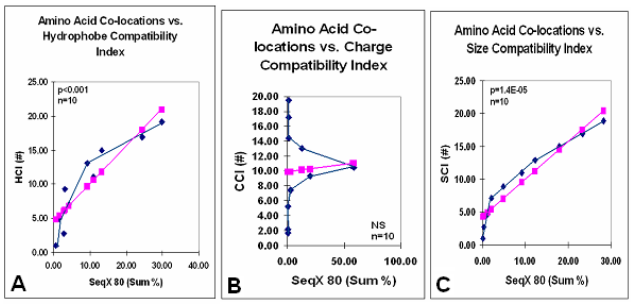

Figure 8: Amino acid co-locations vs. size, charge and hydrophobe compatibility indexes (SCI, C; CCI, B; HCI, A) in major subgroups. Compatibility indexes were calculated for the 20×20 possible amino acid co-locations. Higher index indicates greater compatibility. A sample containing a total of 34,630 co-locations in 80 different protein structures (SeqX80) were divided into 10 major subgroups and the average frequency (Sum %) of co-locations in each subgroup were plotted against the compatibility indexes. The group averages are connected by the blue lines, while the pink symbols and lines indicate the calculated linear regression [9].

significant and negative (-407 kcal/1000 bases, mean value), indicating that these sequences can form structures. However, the FE only has a small free component, less than 10% of the total. The contributions of the 1st and 3rd codon bases to the FE are larger than the contribution of the 2nd (central) bases. It is possible to achieve an approximately 4-fold change in FE by altering the wobble bases in synonymous codons (Figure 11).

These observations suggest the importance (non-randomness) of wobble bases.

Paradigm 10: Location of protein folding information

Proteins are assumed to contain all necessary information for unambiguous folding (Anfinsen's principle). However, *ab initio* structure prediction is often unsuccessful, as the amino acid sequence itself is not sufficient to guide among endless folding possibilities. It seems logical to attempt to find the "missing" information in nucleic acids, specifically in redundant codons.

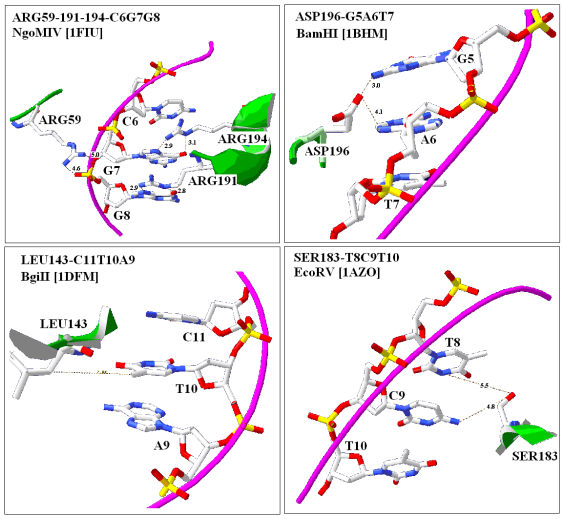

Figure 9: Examples of codon–amino acid co-locations from restrictions enzyme (RE) and specific cut site (RS) complexes [10].

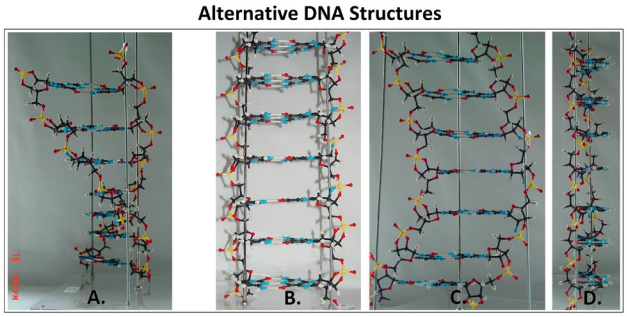

Figure 10: Alterative DNA structures.

Molecular models of dsDNA: spiral (A), linear-ladder (B), twisted-ladder (C) and side view of linear models (D). Atoms are color coded: P (yellow), O (red), C (black), N (blue), h-bond (white) [11].

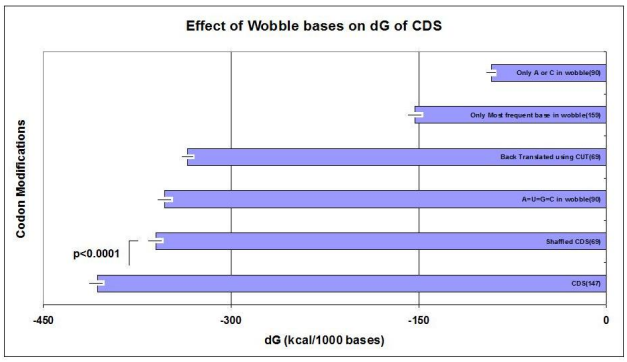

Figure 11: Effect of wobble bases on the dG of CDS.

The TFE of mRNA is indicated in native sequences (CDS), after residue randomization (shuffle) and the indicated manipulation of the wobble bases (see the text for details). Each column represents the mean ± S.E.M.; n is indicated in the columns [12].

Citation: Biro JC (2013) Suggestion to Complete the Canonical Genetic Code: The Proteomic Code and Nucleic Acid Assisted Protein Folding. J Theor Comput Sci 1: 103. doi: 10.4172/2376-130X.1000103

Messenger-RNA energy dot plots and protein residue contact maps were comparable (Figure 12). The structure of mRNA is conserved if the protein structure is conserved, even if sequence similarity is low. These observations led us to propose that similarity may exist between nucleic acid and protein folding [13].

Paradigm 11: Making the "equation" right between nucleic acids and proteins

Nucleic acids contain much excess information owing to codon redundancy. The paradigm insists that this excess information is used to provide protection (security backup) against mutations, i.e. alterations in wobble bases should not affect the correct sequence of amino acids in encoded proteins. However, we should expect redundancy of the 1st or 2nd codon residue, and this is certainly not the case. It is especially strange if we consider that some proteins are not able to fold correctly, as the amino acid sequence is often insufficient to provide correct protein folding. Therefore, research concerning connections between nucleic acid and protein structures and interactions was initiated.

We demonstrated that *co-locating amino acids are preferentially encoded by partially complementary codons, where the 1st and 3rd codon residues are complementary to each other in reverse orientation, but the 2nd codon residues may but not necessarily do complement one another.* This connection between codon co-locations (partially complementary) and amino acid co-locations (interactions) allows the possibility of transfer of spatial (folding) information from nucleic acids to proteins. This is called the 'Proteomic Code', and is missing from the redundant, universal Genetic Code [14].

In 1981, we proposed the idea that specifically interacting peptides are encoded by complementary codons [15]. This was the *first generation of the Proteomic Code*. Several scientists found the idea useful during the design and production of interacting peptides with specific high affinity [16]. However, it became apparent that not all peptides encoded by complementary codons do interact with each other. Fortunately, a modification of the original concept, where complementarity of the 2nd codon residues is permitted, but not obligatory (the *second generation of the Proteomic Code*, [14]), solves this problem (Figure 13).

Paradigm 12: The extended role of tRNA in translation

tRNA was proposed by Crick as a necessity to "adapt" the codon to

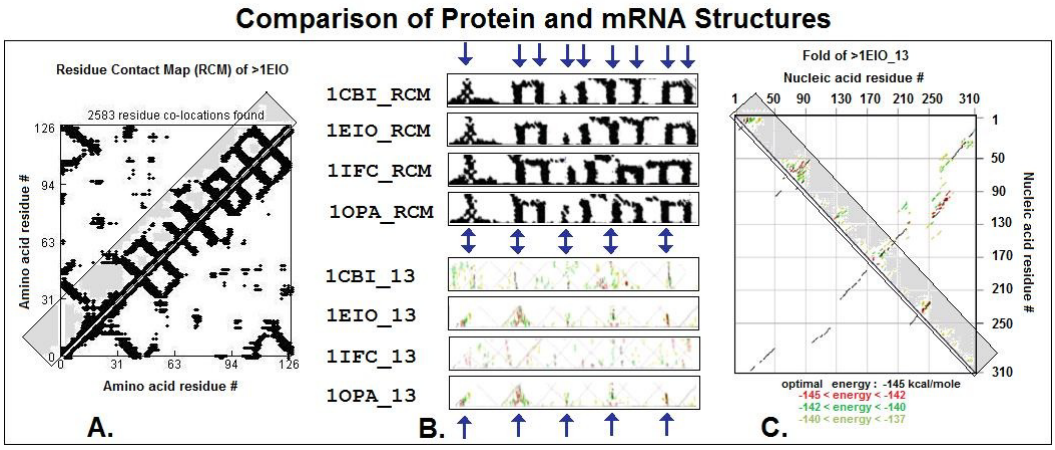

Figure 12: Comparison of protein and mRNA structures.

2D projections of proteins and corresponding mRNAs of four sequences were obtained using SeqX (RCM, A. [8]) and mfold tool (energy dot plots, C). The central, axial segments of these projections (grey areas) were compared (B). The sites of structural similarity are indicated (blue arrows) [13].

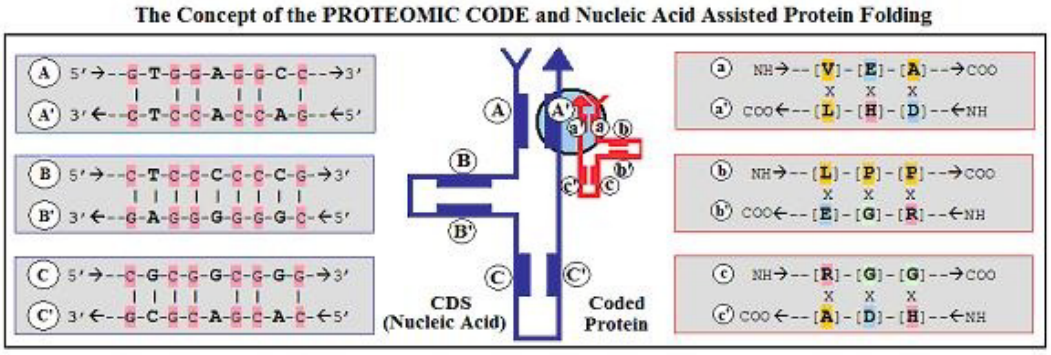

Figure 13: The Concepts of proteomic code and nucleic acid assisted protein folding.

The 3D structure of an encoded protein (red) is established and maintained by segments with specifically interacting domains that contain numerous amino acid co-locations (a-a', b-b', c-c'). Co-locating amino acids (X between their one letter names) are preferentially encoded by partially complementary codons, where the 1st and 3rd codon residues (pink letters connected by |) are complementary to one another (A-T or G-C), but the 2nd codon residues may be, but is not necessarily, complementary to each other. This rule is called the PROTEOMIC CODE. The complementary sites in nucleic acids define segments in the CDS (Nucleic Acid, blue, A-A', B-B', C-C'), which provide a 3D nucleic acid structure similar to the structure of the encoded protein. Codon amino acid interactions transfer the spatial information in CDS to proteins during translation. This process is called NUCLEIC ACID ASSISTED PROTEIN FOLDING [16,19,20].

Citation: Biro JC (2013) Suggestion to Complete the Canonical Genetic Code: The Proteomic Code and Nucleic Acid Assisted Protein Folding. J Theor Comput Sci 1: 103. doi: 10.4172/2376-130X.1000103

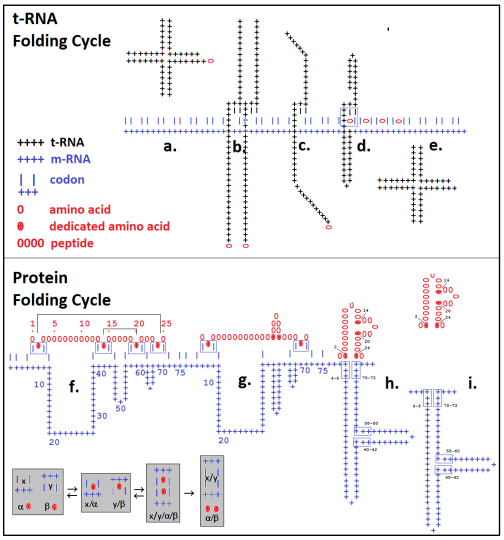

Figure 14: Concept of RNA assisted protein folding.

The model comprises tRNA (upper part) and protein (lower part) folding cycles. During the tRNA cycle, the aminoacyl-tRNA (clover-leaf form, (a)) unfolds, interacts with its codon, and the previously attached tRNA (b) refolds to a configuration that brings the amino acid tail into close proximity with the codon-anticodon site (c, d), loses the amino acid, refolds to its original cloverleaf configuration (e) and is recycled.

The protein folding cycle begins when the peptide synthetase forms peptide bonds between individual amino acids. Some "dedicated" amino acids remain attached to their codons, but most are displaced. The difference in length between the peptide and mRNA creates mRNA folds (f) and the interaction between complementary codons creates peptide folds (g), one after the other (h). The growing peptide-mRNA complex dissociates after "pairing" the last "dedicated" amino acid pair with its corresponding codon pair (i) and the mRNA is recycled. The numbers indicate the positions of the dedicated amino acids and their codons in a 25 amino acid-long peptide and its 75 nucleotide-long mRNA.

The inserted gray boxes depict the rules of the Proteomic Code [2]: co-locating amino acids (α and β) and are encoded by codons (x and y) which are complementary to each other at the 1st and 3rd nucleotide positions; they form different complexes with each other (x/α, y/β, x/y/α/β, x/y, α/β).

the encoded amino acid (a codon is three times longer than an amino acid). However, it became apparent that tRNA is approximately 20-times larger than necessary for this role. The "adaptor" became a clumsy "barrier" between the nucleic acid and protein worlds [3]. Therefore, radical revision (research) of the function of tRNA was necessary to understand the transfer of spatial information from nucleic acids to peptides, and to make sense of the size and frequency of tRNAs [17]. Thermodynamic studies and the literature indicate that tRNA has several possible configurations (in addition to the canonical cloverleaf form). Furthermore, side-by-side interactions between tRNAs are thermodynamically favored. Consequently, we concluded and suggested that there is a tRNA cycle involving unfolding, interaction and refolding of tRNAs, and that this cycle brings codon-anticodon sites into the proximity of the corresponding amino acids [18,19]. Some "dedicated" amino acids remain in contact with their codons after polymerization of amino acids and release of the newly synthesized peptide. This temporary contact is necessary for nucleic acid-assisted protein folding, and this direct codon-amino acid contact is established by tRNAs (Figure 14).

Conclusions

On the basis of the critical review and re-evaluation of recent major paradigms in molecular biology outlined above, we conclude the following, with the suggestion that this novel overview of translation and protein folding provides a complete picture of the Genetic Code and the process of protein synthesis:

1. Codons developed in association with encoded amino acids. There is a stereochemical connection between several (if not all) amino acids and their codons.

2. Wobble bases are not randomly chosen in synonymous codons, but each has well defined roles in determining the structure of nucleic acids and their folding energies.

3. Co-locating amino acids are preferentially encoded by complementary codons (at least at the 1st and 3rd codon positions). This rule is called the Proteomic Code.

4. Structural information contained within nucleic acids is transferred to proteins during translation. This transfer requires direct contact between "dedicated" amino acids and their codons. This process is called nucleic acid assisted protein folding or the concept of mRNA chaperons.

5. There is a tRNA cycle that allows direct codon amino acid contact to be possible.

Biotechnological consequences and applications

There are two main consequences of this research for the biotechnological industry. First, cloning and mass-production of proteins for biotechnological applications is common practice. Modification of the aboriginal CDS of naturally-occurring peptides is normal, as replacing one codon with a synonymous codon often has beneficial effects on the peptide synthesis yield. These modifications have no effect on the amino acid sequence of the product, but may compromise the folding (and functionality) of the protein product.

Second, the proteomic code provides the first real possibility for scientists to design interacting peptides that have high affinity and specificity for their target peptides. The procedure is simple, quicker and more economical than classical antibody production methods. Therefore, it has the potential to accelerate the growth of affinity assays and integrate large numbers of affinity tests into chips. An example of this new, Proteomic Code–based technology is the Affiseq®, which is based on a recent US Patent [20].

Apology

We are aware that our novel picture of translation, the suggestion of Proteomic Code, the nucleic acid chaperons, the reinterpretation of the redundancy of the original Genetic Code and the role of tRNA is controversial for several molecular biologists. It contains elements that have never been tested before. However, the methods employed in bioinformatics are fundamentally different from laboratory methods of classical molecular biology. Easy access to databases and the relatively low cost of tools required to practice this discipline provide a unique opportunity to work in exceptionally wide fields of scientific interest and obtain insights before laboratory scientists can carry out such work. Therefore, we believe that our hypotheses will fulfill a useful function and provide direction in the world of laboratory research and the generous flow of data and publications. Our main reason for this journey "outside the box" is simple. As there is no religion without belief, there is no science without disbelief! To be a good scientist, one must learn the principal dogmas of his discipline, but exceptionally scientists can learn how to discount some of them.

Competing Interests

The author states that he has no competing interests.

Acknowledgement

This work was supported by grants from the Homulus Foundation, Los Angeles, CA, USA. We used public databases and software during this study. We would like to thank the thousands of scientists who have created and maintained these resources, supported by generous grants from nations and individuals. JCB has been a "scientist in the US National Interest" since 2006. JCB wishes to thank the trust and the support of this great Nation. The author also wishes to thank Dr. Paul Agutter (the editor in chief of the *Theoretical Biology and Medical Modelling*) for encouragement and support; and he very much appreciates the continuous attention and advice of George L Gabor MIKLOS PhD (Director of Secure Genetics, Sydney, Australia). The pioneering works and views of Prof. Carl Woese were very useful and indeed essential for parts of our work. We respectfully recognize and acknowledge this.

References

1. Biró JC (2011) Biological information—Definitions from a biological perspective. Information 2: 117-139.
2. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27: 379-423.
3. Woese CR (1967) The Genetic Code: The Molecular Basis for Gene Expression. Harper & Row, New York, USA.
4. Crick FH (1968) The origin of the genetic code. J Mol Biol 38: 367-379.
5. Biro JC, Benyó B, Sansom C, Szlávecz A, Fördös G, et al. (2003) A common periodic table of codons and amino acids. Biochem Biophys Res Commun 306: 408-415.
6. Biro JC (2006) Indications that "codon boundaries" are physico-chemically defined and that protein-folding information is contained in the redundant exon bases. Theor Biol Med Model 3: 28.
7. Biro JC (2008) Does codon bias have an evolutionary origin? Theor Biol Med Model 5: 16.
8. Biro JC, Fördös G (2005) SeqX: A tool to detect, analyze and visualize residue co-locations in protein and nucleic acid structures. BMC Bioinformatics 6: 170.
9. Biro JC (2006) Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. Theor Biol Med Model 3: 15.
10. Biro JC, Biro JM (2004) Frequent occurrence of recognition site-like sequences in the restriction endonucleases. BMC Bioinformatics 5: 30.
11. Biro JC (2003) Speculations about alternative DNA structures. Med Hypotheses 61: 86-97.
12. Biro JC (2008) Correlation between nucleotide composition and folding energy of coding sequences with special attention to wobble bases. Theor Biol Med Model 5: 14.
13. Biro JC (2005) Nucleic acid chaperons: A theory of an RNA-assisted protein folding. Theor Biol Med Model 2: 35.
14. Biro JC (2006) A novel intra-molecular protein-protein interaction code based on partial complementary coding of co-locating amino acids. Med Hypotheses 66: 137-142.
15. Bíró J (1981) Comparative analysis of specificity in protein-protein interactions. Part II.: The complementary coding of some proteins as the possible source of specificity in protein-protein interactions. Med Hypotheses 7: 981-993.
16. Biro JC (2007) The Proteomic Code: A molecular recognition code for proteins. Theor Biol Med Model 4: 45.
17. Woese CR (2001) Translation: In retrospect and prospect. RNA 7: 1055-1067.
18. Biro JC (2012) The concept of RNA-assisted protein folding: the role of tRNA. Theor Biol Med Model 9: 10.
19. Biro JC, Biro JM (2013) The concept of RNA-assisted protein folding: representation of amino acid kinetics at the tRNA level. J Theor Biol 317: 168-174.
20. Biro JC (2012) System and method to obtain oligo-peptides with specific high affinity to query proteins. Patent No: US 8,145,437 B2.