Int. J. Mol. Sci. **2008**, *9*, 2424-2446; DOI: 10.3390/ijms9122424

OPEN ACCESS

International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms/

Review

Discovery of Proteomic Code with mRNA Assisted Protein Folding

Jan C. Biro

Homulus Foundation, 612 S Flower St, Los Angeles, 90 017 CA, USA. E-Mail: jan.biro@att.net; Tel. +1-213-627-6134

Received: 18 August 2008; in revised form: 24 November 2008 / Accepted: 2 December 2008 / Published: 3 December 2008

Abstract: The 3x redundancy of the Genetic Code is usually explained as a necessity to increase the mutation-resistance of the genetic information. However recent bioinformatical observations indicate that the redundant Genetic Code contains more biological information than previously known and which is additional to the 64/20 definition of amino acids. It might define the physico-chemical and structural properties of amino acids, the codon boundaries, the amino acid co-locations (interactions) in the coded proteins and the free folding energy of mRNAs. This additional information, which seems to be necessary to determine the 3D structure of coding nucleic acids as well as the coded proteins, is known as the *Proteomic Code* and *mRNA Assisted Protein Folding*.

Keywords: Gene, code, codon, translation, wobble-base.

1. Introduction

Mapping between messages in nucleic acid and protein alphabet is a fascinating story, a story that still unfolding. It is about to understand the rules of information transfer between DNA and proteins. First of all it is not only a biochemical puzzle and much of the early methods for devising codes came from combinatorics, information theory. Four and 20 (number of bases and amino acids) seems to be magical numbers with amazingly many possible mathematical connections between them [1].

The existence of a *Genetic Code* became obvious immediately after the discovery of DNA structure [2, 3]. The first suggestion for a code came from George Gamow, not even a biologist, but a physicist who became most famous as the chief proponent of the Big Bang theory in cosmology. He proposed a *Diamond Code* [4-6] where DNA acted directly as a template for assembling amino acids into proteins.

Various combinations of bases along one of the grooves in the double helix could form distinctively shaped cavities into which the side chains of amino acids might fit. Each cavity would attract a specific amino acid; when all the amino acids were lined up in the correct order along the groove, an enzyme would come along to polymerize them. Gamow's code turned out to be an overlapping triplet code which provided exactly the desired 20 combinations. There are many beautiful aspects of the overlapping codon: a) it maximizes the density of information storage and b) even though three bases are needed to specify any single amino acids, the overall ratio of bases to amino acids approaches 1:1; c) it supposes that the distance between base pairs in DNA and the distance between amino acids in the proteins is raptly similar, which is exactly the case; 4) and avoids the possibility and consequences of frame shift. Unfortunately this code has serious constrains: only $4^4 = 256$ overlapping codon combinations are possible, while with 20 kinds of amino acids, there are $20^2 = 400$ possible dipeptides. The 144 "impossible" dipeptides were found in real proteins by Sidney Brenner [7] and it ruled out Gamow's codon "diamonds".

Another brilliant code was created by Francis Crick [8] the *Comma-Free Code*. Crick was speculating about an adaptor hypothesis (that he never published) the idea that amino acids do not interact directly with mRNA, but there is a mediator which recognizes the codons. Codons line up continuously along the DNA, like pearls on a neckless, and are recognized by a specific mediator. (This mediator is known today as tRNA). However, unlike a neckless, there is no space ("comma") between the codons which would indicate the codon boundaries. How is it still possible to distinguish between meaningful (those necessary 20 which lined up prettily on the DNA) and meaningless (the remaining 44 overlapping) codons? The answer was original: there is no mediator for meaningless codons. This solution was so simple and elegant, that it got the name the "prettiest wrong idea in all of 20^{th}-century science".

There were many theoreticians involved in the invention of the Genetic Code. Finally, *M. W. Nirenberg* and *J H. Matthaei*, two laboratory bench scientists published the real *Genetic Code* [9, 10] and provided a huge surprise: *it was redundant*. It was not compact, it was seemingly chaotic, there were no signs of any protection against frame shifts, and there were no signs of any connection between the codons and any characteristics of the coded amino acids.

Why do we have 64 triplets for coding 20 amino acids - more than three times the number needed? Explaining away this excess became a major preoccupation of coding theorists. One intelligent explanation is that the redundancy confers a kind of error tolerance, in that many mutations convert between synonymous codons. When a mutation does alter an amino acid, the substitute is likely to have properties to those of the original. Alternatively the mutation is likely to be a stop codon which completely aborts the wrong translation. Another possible explanation is that the *Genetic Code* is developing, say started with 4, one letter codon coding 4 amino acids; continued to, say 16 two letter codons coding 16 amino acids. Recently we happen to have 64, three-letters-codons and they are coding 20 amino acids. However there is a potential to end up in the future with a, say 64, three-letters-codons coding 64 amino acids system.

Crick had, of course, his own explanation: forget any logical connection between codons and coded amino acids, it is just a "frozen accident" or by other words "that is what we got, lets we like it…"

2. Second Look at the Nirenberg Code

2.1. The 3D structure of mRNA

Evolution often occurs in stages, one step is followed by a plateau before the next step is possible to take. It is true even for the development of scientific thinking and understanding. The 50-es and 60-es were very fertile for biology and the foundation of the recent molecular biology was laid. It took 30+ years to recognize, that the discovery of DNA was not the discovery of *The secret of Life*, the Life has much more secrets. Some early ideas were and some still are wrong, very wrong. The most embarrassing mistake of the modern genomics was probably the concept of sense and non-sense DNA strands. This was a concept deeply rooted in the mind of even the most brilliant scientists. The possibility of whole genome sequencing finally opened the gates (in late 90-es) to get rid of this nonsense idea and begin to see that both DNA strands are equally important for protein expression. It is no longer feasible, that one strand is expressed, while the other (complementary) strand serves only as a reproductive template (backup). Complementarity of bases is fundamental for helix formation of dsDNA, but more than that it makes possible formation of much more complex and sophisticated 3D structures than the known monotone helix. The unique, signature-structure of tRNA is well recognized and accepted, meanwhile the possibility and significance of mRNA 3D structure formation is widely denied [11]. The mRNA is passing the translation machinery on the surface of ribosomes, codon after codon, like a tape passes the magnets of a tape recorder. More linear mRNA is expected to be more suited for translation because a structured mRNA only slows down the machinery.

The number of mRNA images in Nucleic Acid Structure Database (NDB) is very limited [12]. Fortunately the known rules of codon base pairing makes it possible to predict the thermodynamically most likely structures of any nucleic acid [13]. (Such tools and predictions are not yet available for proteins). It is widely accepted today, that mRNAs have secondary structure even if there are numerous methodological considerations how to measure the folding energy content of nucleic acids [14-16].

The functional significance of the mRNA secondary structure is not known, therefore wobble base replacement is a widely used and accepted method to eliminate the folding energy of mRNA and achieve higher translational efficiency [17-18].

2.2. Physico-chemical definition of codon boundaries

Frame-shift is a major concern regarding the translation. Nirenberg's Genetic Code seems not to give any protection against the possible occurrence of frame-shifts even if it gives some promises to reduce the catastrophic consequences of wrong codon readings. However a second look at the base composition of codons (64 as it is), or the usage-weighed variants from different Species Specific Codon Usage Tables [19], reveals that it is not completely random. The first and third codon positions contain significantly more G and C bases, than the second (middle) codon positions. This bias has remarkable consequences. There are 3 H bonds between C and G (dG=-1524 kcal/1k bases) while only two between A and T bases (dG=-365 kcal/1k bases). Therefore the GC content is the major determinant of folding energy, and by that way the mRNA's 3D structure. The higher GC content indicates that the 1st and 3rd codon residues have significantly larger effect on the mRNA structure than

the 2nd one. Indeed, wobble bases were found to be the most important codon residues to determine mRNA structure [14].

This codon related periodic variation of GC content means that there is a periodic pattern of folding energy (dG) along the mRNA, which distinguishes the central codon base from the 1st and 3rd and forms a physico-chemical barrier or boundary between the codons. This is a statistical rule which doesn't apply for every single codon, but still shows a general tendency that there is some potential protection against frame shifts in the Nirenberg's *Genetic Code* [14, 20-21]. Manipulation of wobble bases to eliminate mRNA secondary structures (and speed up the translation) will destroy this energy pattern and by that way increase the translation errors. Native synonymous codon usage (and codon bias) seems to prefer selection for translational accuracy versus velocity [22-25].

2.3. The Common Periodic Table of Codons and Amino Acids

There is even another completely separate line of evidence suggesting that codon positions are different, and the central codon position has a very special role. There always has been an effort to connect codons to their coded amino acids. The wobble base lost its importance because of its interchangeability. Most scientific efforts focused on to find stereo-chemical compatibility (spatial fitting) between the atomic geometry defined by 2 or 3 nucleic acid bases and the corresponding geometry defined by the residue of the coded amino acid [26, 27]. Crick furiously attached these efforts stating that any connection between codons and amino acids is only accidental and there is no underlying chemical rationale [28]. On the other hand Carl Woese argued that the *Genetic Code* developed in a way that was very closely connected to the development of the amino acid repertoire, and that this close biochemical connection is a fundamental of specific protein–nucleic acid interactions [29].

A common regularity in an arrangement of codons and amino acids provides a strong support for the evolutionary connection between mRNA and coded proteins.

A *Periodic Table of Codons* (Table 1a) has been designed in which the codons are in regular locations. The *Table* has four fields (16 places in each), one with each of the four nucleotides (A, U, G, C) in the central codon position. Thus, AAA (lysine), UUU (phenylalanine), GGG (glycine) and CCC (proline) are positioned in the corners of the fields as the main codons (and amino acids). They are connected to each other by six axes (2 horizontal, 2 vertical, 2 axial). The resulting nucleic acid periodic table shows perfect axial symmetry for codons. The corresponding *Amino Acid Table* (Table 1b) also displays periodicity regarding the biochemical properties (charge and hydropathy) of the 20 amino acids, and the positions of the stop signals. Table 1 emphasizes the importance of the central nucleotide in the codons, and predicts that purines control the charge while pyrimidines determine the polarity of the amino acids [30]. In addition to this correlation between the codon sequence and the physico-chemical properties of the amino acids, there is a correlation between the central residue and the chemical structure of the amino acids. A central *uridine* correlates with the functional group –$C(C)_2$–; a central *cytosine* correlates with a single carbon atom, in the C_1 position; a central *adenine* coincides with the functional groups –CC=N and –CC=O; and finally a central *guanine* coincides with the functional groups –CS, –C=O, and C=N, and with the absence of a side chain (glycine) (Table 2).

Table 1. The Common Periodic Table of Codons and Amino Acids.

a. Periodic Table of Codons.

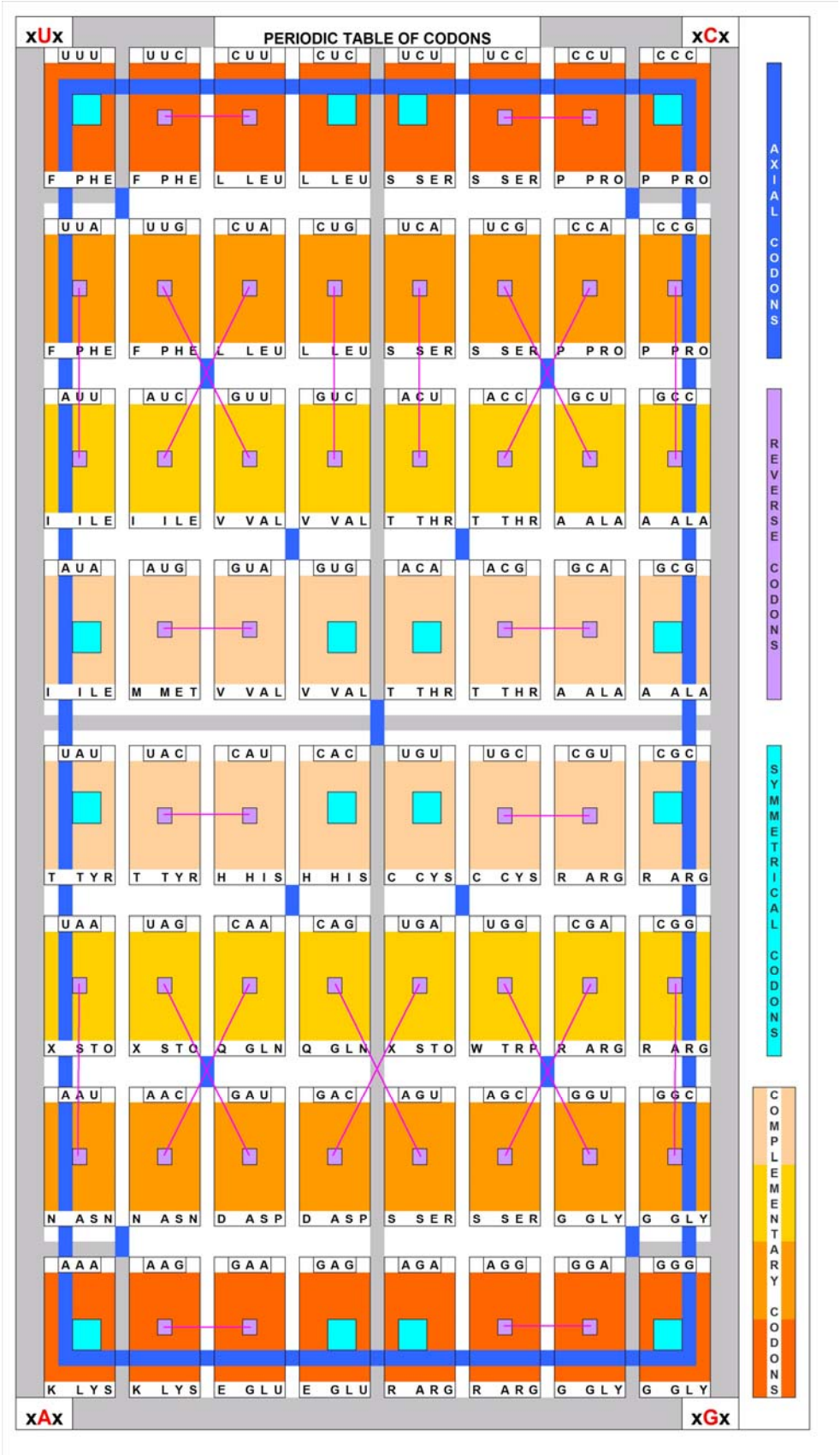


Table 1. Cont.

b. Periodic Table of Amino Acids.

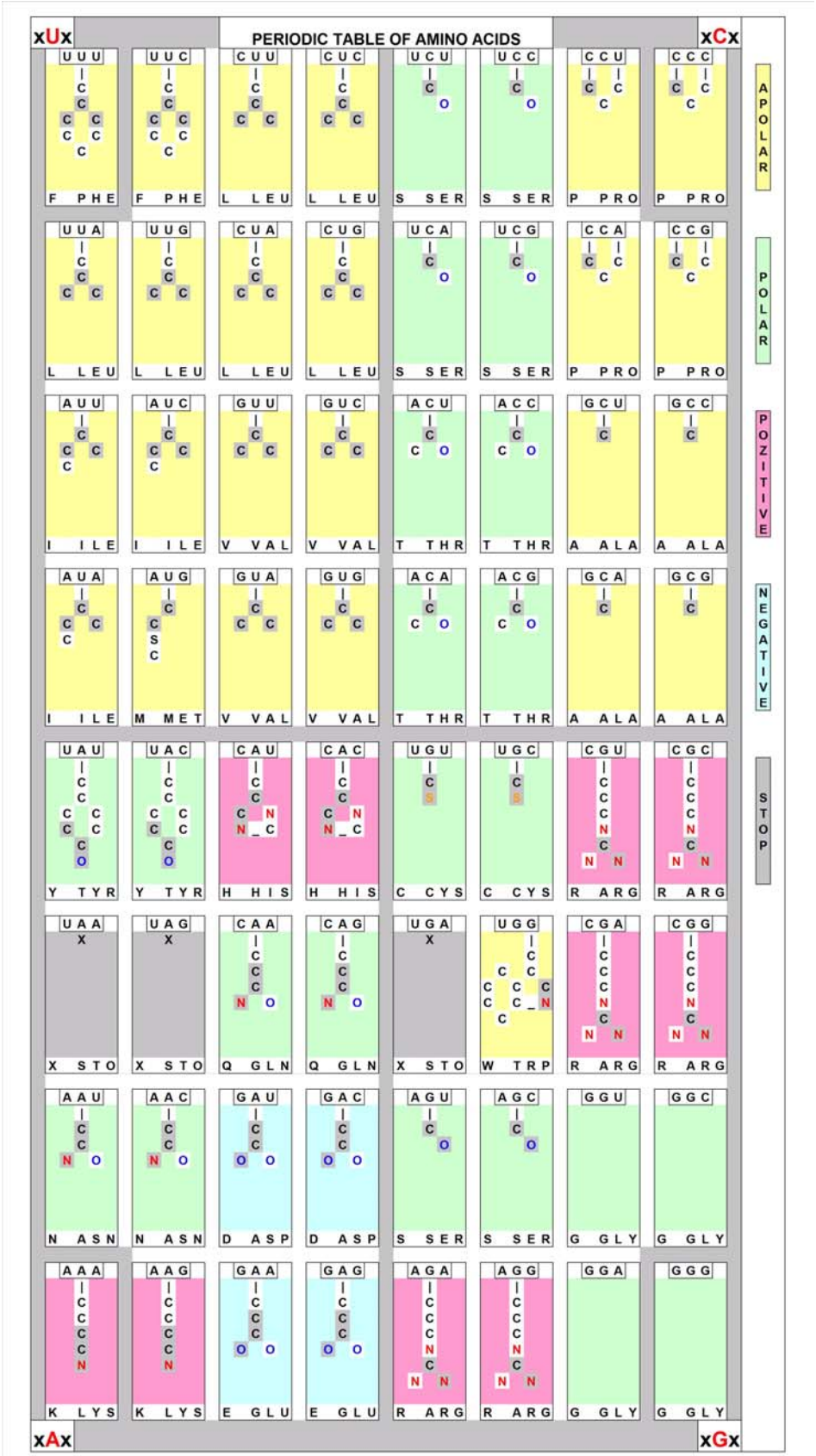
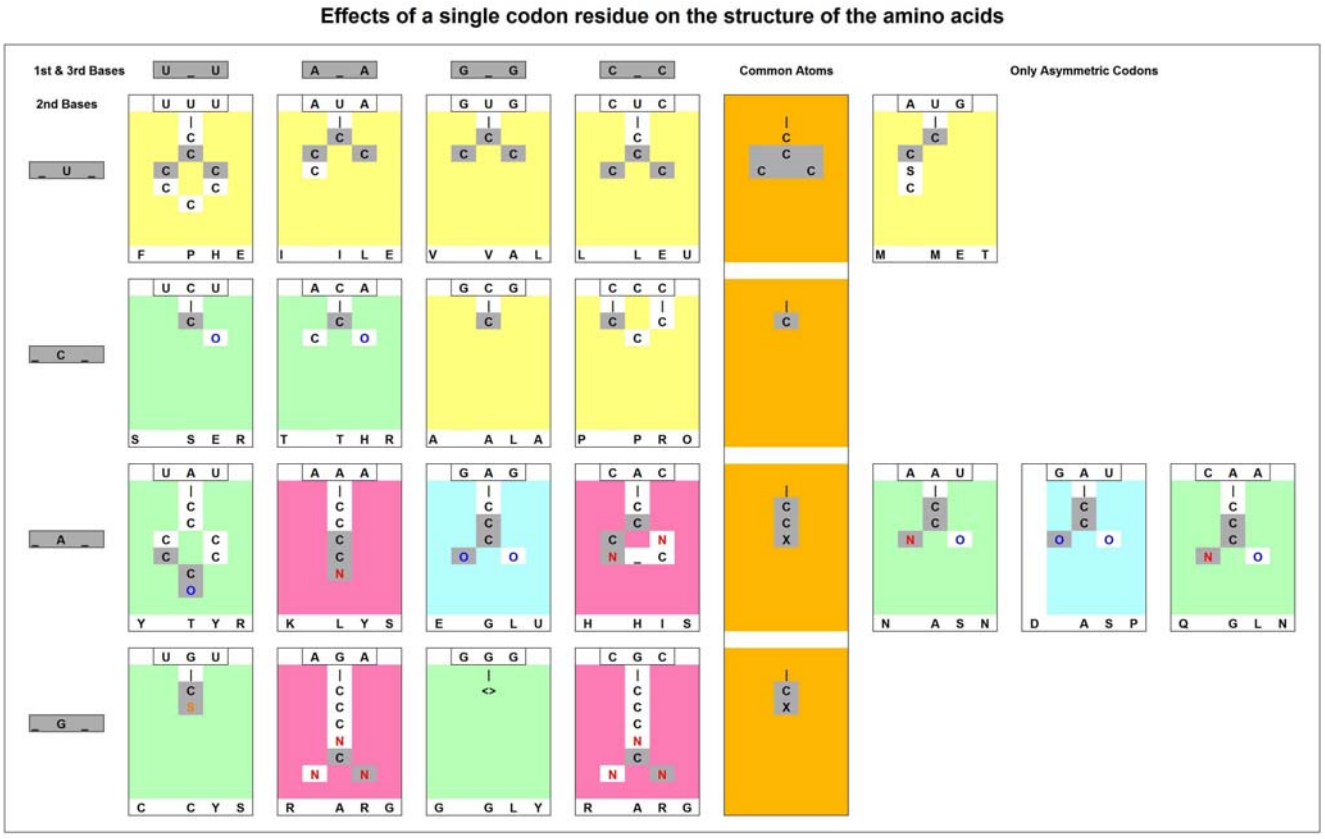


Table 2. Effects of a single codon residue on the structure of the amino acids.



The 16 amino acids coded by symmetric codons were sorted into *rows* accordingly to their central codon residues (_A_: adenine, _U_: uridine, _C_: cytosine and _G_: guanine) and into *columns* accordingly to their 1st and 3rd codon residues (U_U, A_A, G_G, C_C). The 4 amino acids coded by only asymmetric codons are also included into the table (last columns). Some structural features in the amino acids could be related to the central codon residues (atoms in grey boxes and the yellow column in the middle of the table).

I interpret these results as a clear-cut answer for the Woese vs. Crick dilemma: there *is* a connection between the codon structure and the properties of the coded amino acids. The second (central) codon base is the most important determinant of the amino acid property. It explains why the reading orientation of translation has so little effect on the hydropathy profile of the translated peptides. Note that 24 of 32 codons (U or C in the central position) code apolar (hydrophobic) amino acids, while only 1 of 32 codons (A or G in the central position) codes non-apolar (non-hydrophobic, charged or hydrophilic) amino acids. It explains why complementary amino acid sequences have opposite hydropathy, even if the binary hydropathy profile is the same.

Although there can be seen a quite large correlation between amino acid codons and amino acid properties (an interesting finding as such), one has to be cautious in saying that this correlation has a functional significance. Such significance could eventually exist there (e.g. in determining the structure of mRNA), but more evidence is needed. Before that this correlation still can be a fruitful hypothesis for further research.

2.4. Visualization of specific nucleic acid – protein interactions

The strong connection between codon structure and physicochemical properties of coded amino acids (the existence of *The Common Periodic Table of Codons and Amino Acids*) suggests that specific interaction between codons and coded amino acids might occur. There is no doubt, that specific interaction between nucleic acids and proteins is an absolute necessity for many vital functions, for example the regulation of gene expression. While should the codon / coded amino acid interaction be the only forbidden possibility to accomplish this function?

The interaction between restriction enzymes (RE) and their recognition sequences (RS) are known to be very specific and fortunately numerous such interactions are visualized and available from PDB. A review of the seven available crystallographic studies [31] showed that the amino acids coded by codons that are subsets of recognition sequences were often closely located to the RS itself and they were in many cases directly adjacent to the codon-like triplets in the RS. Fifty-five examples of this codon-amino acid co-localization were found and analyzed, which represents 41.5% of total 132 amino acids which are localized within 8 Å distance to the C1' atoms in the DNA. The average distance between the closest atoms in the codons and amino acids is 5.5 +/- 0.2 Å (mean +/- S.E.M, n = 55), while the distance between the nitrogen and oxygen atoms of the co-localized molecules is significantly shorter, (3.4 +/- 0.2 Å, $p < 0.001$, n = 15), when positively charged amino acids are involved. This is indicating that a direct interaction between nucleic- and amino acids might occur. We interpret these results in favor of Woese and suggest that the *Genetic Code* is "rational" and there is a stereo-specific relationship between the codons and the coded amino acids (Figure 1). But how well

Figure 1. Co-location of Codon-like Triplets and Amino Acids in RE-RS Complexes. Examples for co-locations of amino acids in restrictions endonucleases (RE) and codon-like triplets in restrictions enzyme recognition sites (RS). The name of enzyme, the name and position of nucleic acid bases and amino acids are indicated. Four amino acids located at overlapping-codon like base sequence in *Eco*RI is indicated in the yellow box.

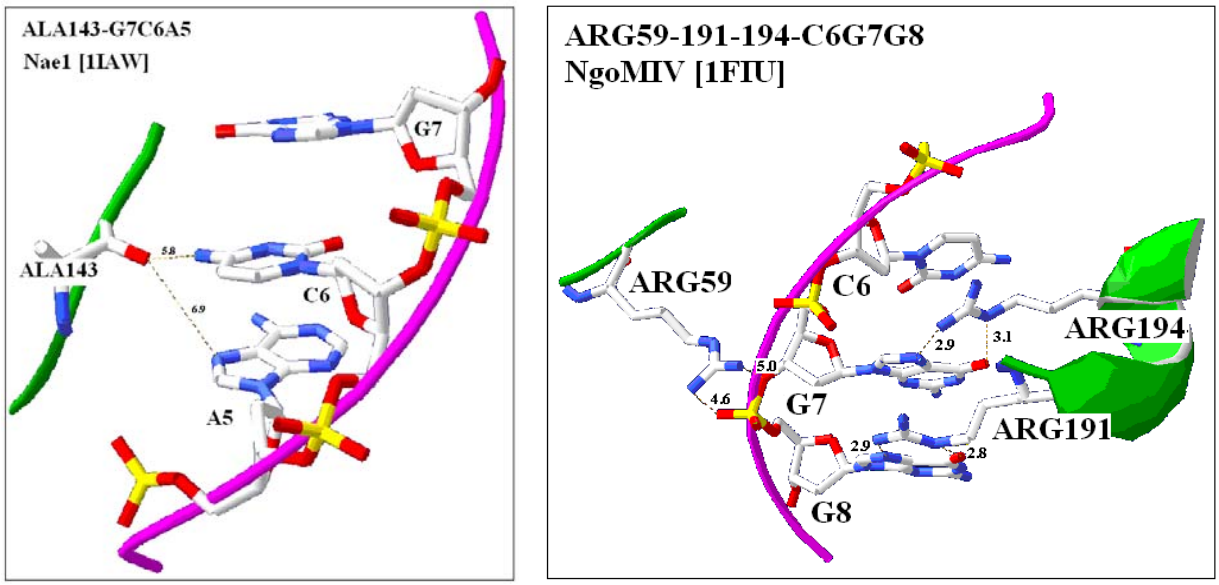


Figure 1. Cont.

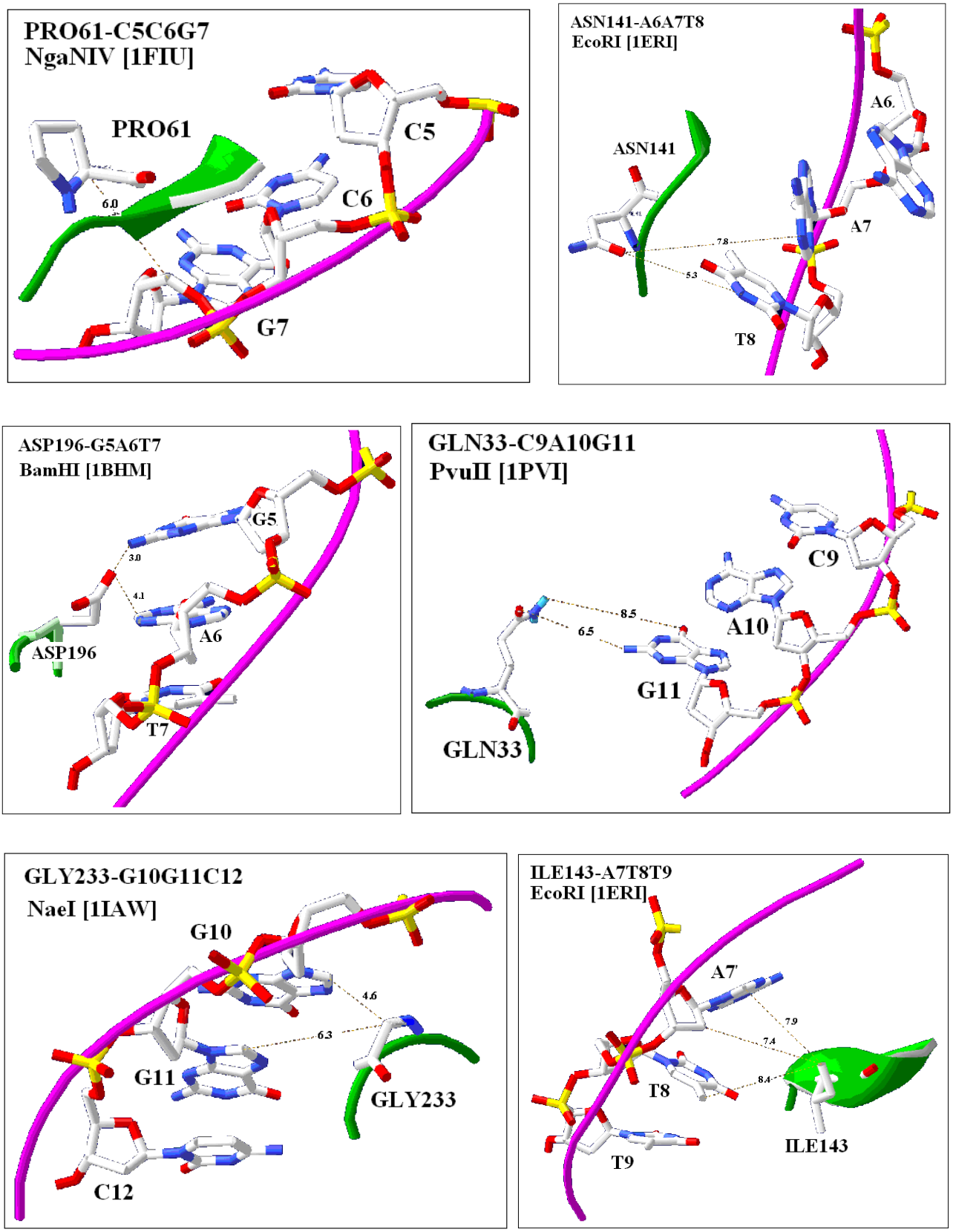
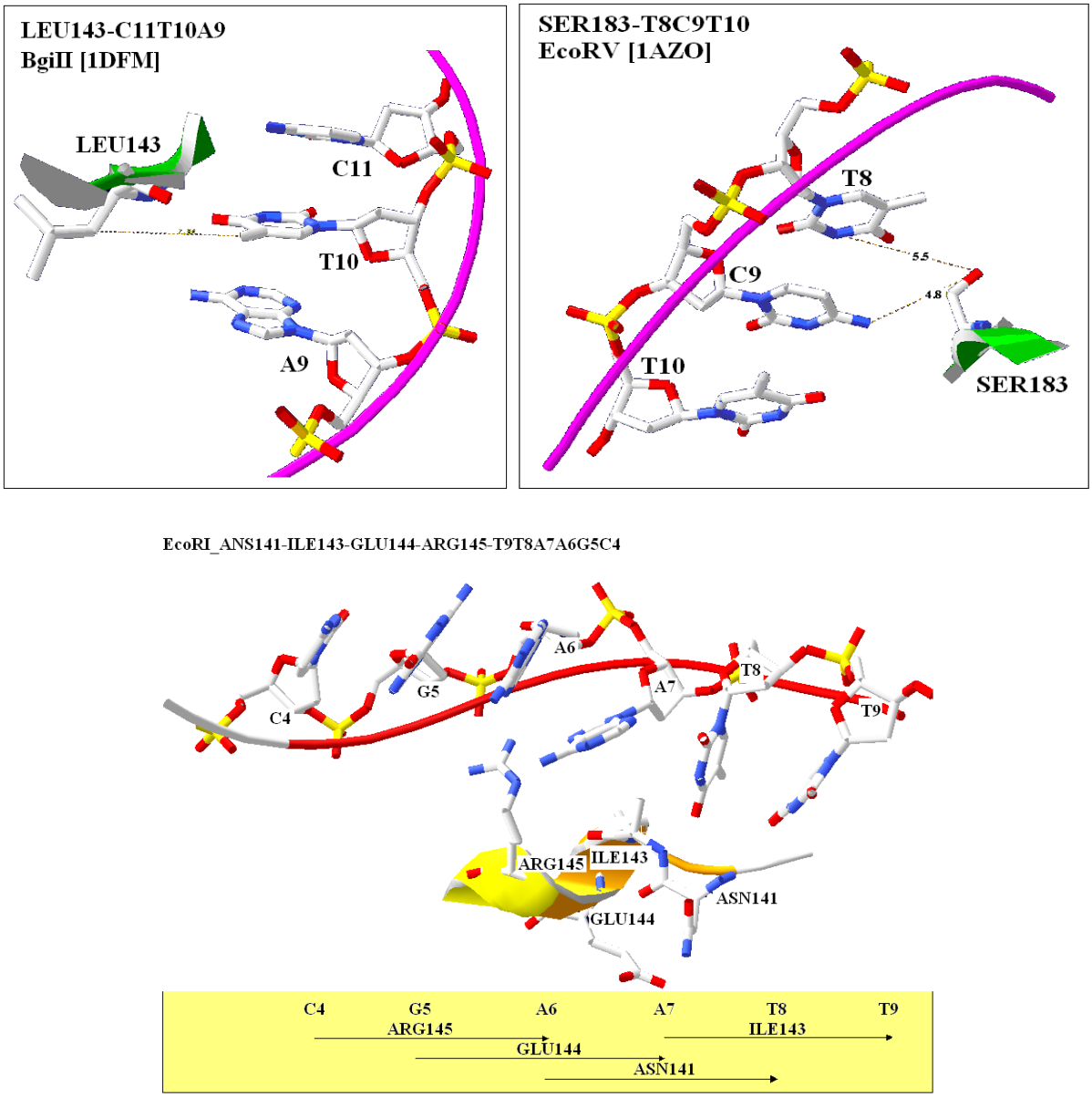


Figure 1. Cont.



this finding supports Woese is still open. Although the finding that amino acids and their codons in restriction enzymes are closely located can give ideas for studying the genetic code from the same perspective, it is not sifficient evidence that the same occurs in the general genetic code. However the close location of codons and coded amino acids in restriction enzymes is an indirect indication for the possible existence of specific affinity between coding- (mRNA) and coded- (protein) sequences, which has a special functional meaning for the concept of the RNA assisted protein folding.

2.5. Partial Complementary Coding of Co-locating Amino Acids

There was an idea published in early 80-s [32-37] suggesting that specifically interacting proteins are coded by complementary nucleic acids. The idea was, of course, rejected because it proposed that even the "non-sense" DNA strand might be expressed and makes some sense, which was an absurdity at that time. An early effort to confirm this theory, using (rather undeveloped) bioinformatical

methods, failed [38]. There was a short come-back of this idea (called today as the *Proteomic Code*) and several research groups confirmed that proteins derived from complementary nucleic acid strands have specific, high affinity attraction to each other [39-43]. However it turned out that there is some problem with the consistency of the results: the method sometimes worked sometimes not.

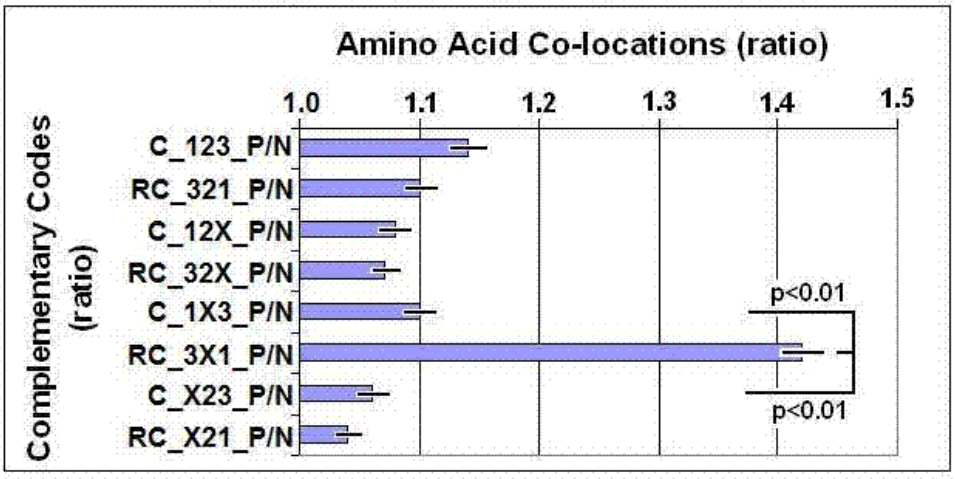
We constructed a bioinformatics tool to collect data of co-locating amino acids from known protein structures, listed in the PDB, for statistical analyses [44]. (The immediate neighborhood on the same peptide chain was not counted as co-location). These analyses provided some rather novel observations.

1) Co-locating amino acids are physico-chemically compatible with each other, i.e. large and small, positive and negative, hydrophobe and hydrophobe amino acids are preferentially co-located with each other. The novelty of this observation is that physico-chemical rules apply already on residue level and do not necessarily need large, complex interfaces of interacting proteins [45].

2) Co-locating amino acids are preferentially coded by partially complementary codons, where the 1st and 3rd bases are complementary but the 2nd may but not necessarily are complementary to each other (Figure 2).

The propensities for the 20x20=400 possible amino acid pairs were monitored in 81 different protein structures with the SeqX tool. The tool detected co-locations when two amino acids were within 6 Å distance of each other (neighbors on the same strand were excluded). The total number of co-locations was 34,630. Eight different complementary codes were constructed for the codons (2 optimal and 6 suboptimal). In the two optimal codes, all three codon residues (123) were complementary (C) or reverse complementary (RC) to each other. In the suboptimal codes, only two of three codon residues were C or RC to each other (12, 13, 23), while the third was not necessarily complementary (X). (For example, Complementary Code RC_1X3 means that the first and third codon letters are always complementary, but not the second and the possible codons are read in reverse orientation. P/N (positive/negative) ratio indicates the proportion of co-locating amino acids coded by the defined codon complementary rules.

Figure 2. Complementary codes vs. amino acid co-locations (modified from [47]).



These observations lead us to conclude that there are significant additional functional and structural connections between codons and coded amino acids to that what was described by Nirenberg and is known as the *Genetic Code*. We formulated the recent concept of *Proteomic Code* to describe this additional connection (for review see [46, 47]).

2.6. The Role and Predictability of Wobble bases

The 64/20 Genetic Code is redundant, mainly because the 3rd codon bases, in most codons, are interchangeable without any consequence on the sequence of the coded protein. The only information expected from the DNA to the protein syntheses is the coding of amino acids, because it is believed, that the only information necessary to correct protein folding is only the correct amino acid sequence itself.

These believe is based on *Anfinsen's thermodynamic principle* [48] which states that the amino acid sequence contains all the necessary information to correct and unambiguous protein folding or, by other words, folding/refolding is in principle a reversible, thermodynamically driven process. In practice, however, the reversibility of denaturation is not observed in all cases. Most proteins in an environment close to their neutral, physiological conditions (neutral pH, ~ 150 mM ionic strength) cannot fold back after complete or partial unfolding. This is particularly true for large proteins [49]. Folding and refolding might be visualized as navigation of an ensemble of unfolded molecules through the bumpy energy landscape in search of the native state [50]. A fraction of molecules reaches the native state directly, whereas the remaining fraction gets kinetically trapped in metastable conformations. Therefore the general validity of *Anfinsen's thermodynamic principle* is the subject of returning criticism (like the Levinthal paradox [52] which refers to kinetic aspects of protein folding and does not disprove the existence nor the stability of the native state). However, even if no protein has yet been identified that cannot fold spontaneously under permissive conditions, most protein chemists agree, that, in many cases, additional molecular information, provided by *chaperons*, is necessary to guide or facilitate the protein folding under physiological (neutral) conditions.

The annealing action of these (mainly protein) chaperons is described as providing supplemental information for systems that do not otherwise have a definite ground state. Extensive research into chaperonin assisted protein folding [50, 51, 53-55] by indicates that chaperonins rescue misfolded proteins from kinetic traps and that the native state of the substrate protein is not altered by the chaperonin. A chaperon, for example, interact with it's denatured substrate protein (SP), removes it from it's kinetic trap by stretching it [51], stabilize elongated chains. Correct folding in the cytosol is achieved either on controlled chain release from these primary chaperons or after transfer of newly synthesized proteins to downstream chaperones, such as the chaperonins [54]. Protein chaperons are highly conserved, and the coexistence of these chaperones in the same cytosol suggests that certain chaperone-cochaperone interactions are also permitted.

By other words there is some 3x excess of information before translation, and there seems to be a shortage of information after translation. It is logical to assume, that folding information is stored in the redundant codons, more concretely in the wobble bases. The literature is actually rather rich with observations connecting the wobble bases to some structural feature of the coded proteins [56-61]. Even a wobble base centric sequence – structure database was constructed [62].

The preferential coding of co-locating amino acids by partially complementary nucleic acids, (for example by 5'>ANG>3'/3'<TNC<5' pattern) immediately suggests a role for the wobble bases. They are integrated parts of codons, defining amino acid co-locations. They are not randomly chosen, but logically selected: the wobble base of Xc codon (defining Xa amino acid) is defined by the first residue of codon Yc (which is coding Ya amino acid) if that two amino acids (Xa and Ya) are co-locating and *vice versa* the 3rd residue of Yc is defined by the 1st base in codon Xc (A defines T & U, G defines C).

Protein structures contain many amino acid co-locations (immediate neighbors on the same chain are excluded). Suppose that preferential partial complementarity coding of amino acids is not a rarity, but it is a rule. In that case the signs of non-randomness of wobble base selection should be seen not only in a small subset of proteins but even in very large data sets, like the species specific codon usage frequency tables.

Statistical analyses of A, T, G, C frequencies at 1st, 2nd, and 3rd codon positions in 113 species specific Codon Usage Frequency Tables and 87 protein structures showed strongly significant internal correlation between the frequency of nucleic acid bases at different codon positions. This strong relationship made it possible to predict the frequency of all possible wobble bases in all the 64 codons in all the 113 species (P<1.3E-64, N=113) and all the 87 proteins (p<1.1E-28, n=87) [63].

These strong correlations wouldn't be possible with random selection of wobble bases. Therefore we concluded, that synonymous codons are not interchangeable with each other without disturbing the internal order of bases in integrated codon systems like a native mRNA or a species specific Codon Usage Frequency Order.

2.7. Integrated Codon Systems

There are more than observations provided by theoretical and computational biology which are indicating, that native, natural proteins, as well as their coding sequences, are much more than the sequential collection of their building blocks. They are an integrated, interconnected system.

1) Wobble base mutations are expected to be "silent" without any consequences for the biological functions or phenotypes. They are often not. "Silent" polymorphism or mutation affects a) substrate specificity [64], b) drug pharmacokinetics and multidrug resistance in human cancer cells [65], c) mRNA stability and synthesis of the receptor [66], d) splicing [67, 68], e) different functions [69-71].

2) It has recently become clear that the classical notion of the random nature of mutations does not hold for the distribution of mutations among genes: most collections of mutants contain more isolates with two or more mutations than predicted by the mutant frequency on the assumption of a random distribution of mutations. Excesses of multiples are seen in a wide range of organisms, including riboviruses, DNA viruses, prokaryotes, yeasts, and higher eukaryotic cell lines and tissues. In addition, such excesses are produced by DNA polymerases *in vitro*. These "multiples" appear to be generated by transient, localized hypermutations rather than by heritable mutator mutations. The components of multiples are sometimes scattered at random and sometimes display an excess of smaller distances between mutations [72, 73].

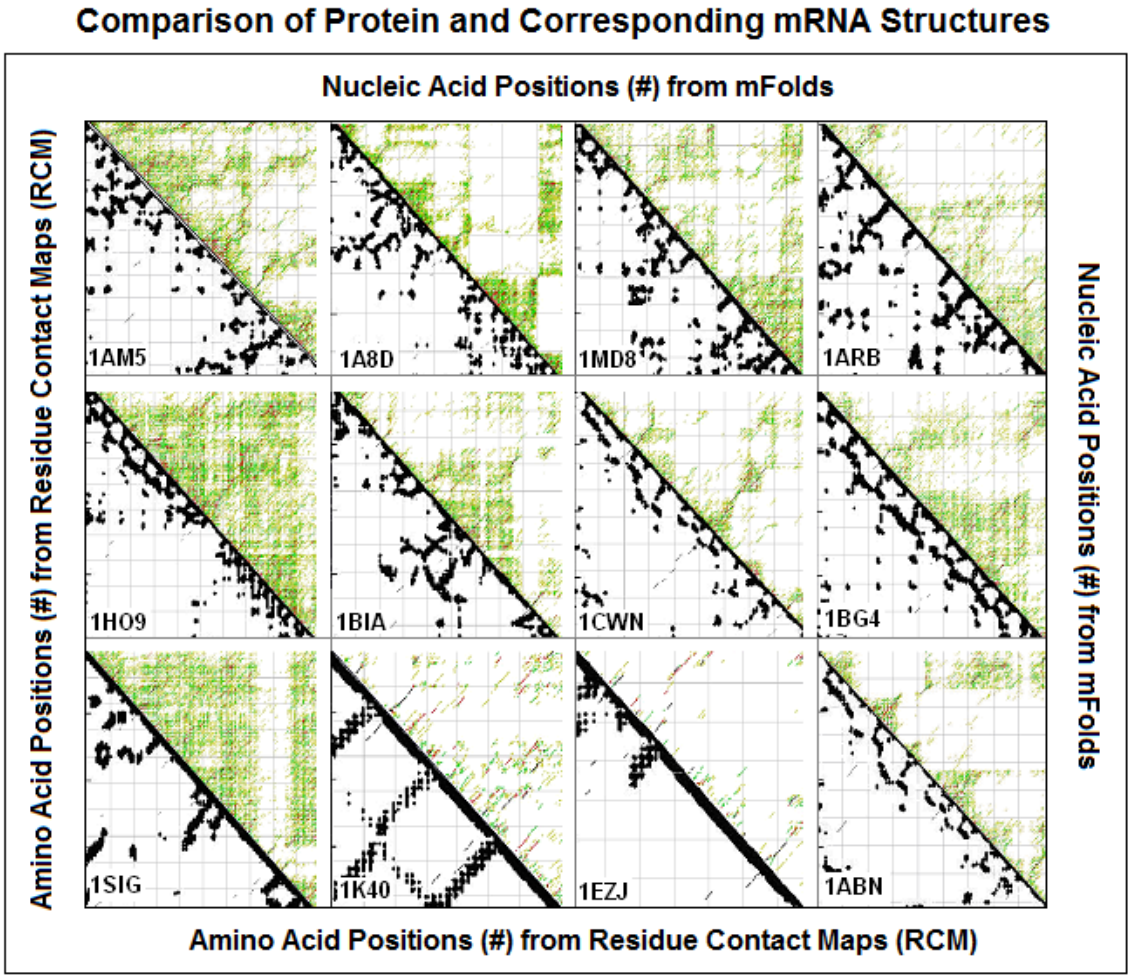
3) A compensatory mutation occurs when the fitness loss caused by one mutation is remedied with a second mutation at a different site in the genome.

Often it occurs in the same gene, alters the protein sequence [74, 75] but saves the protein's secondary and tertiary structure. Compensatory mutations (in the same genes) seem to work through restoring the protein's structure (allosterism) which saves the protein's function [76].

Uneven concentration of mutations of smaller distances and their compensatory character are further indications of the integration and interconnectivity of codons in the same gene and consequently, conservation of structurally critical amino acid connections (but the amino acids) in the coded proteins.

However it should be noticed, that the distribution of mutations in directed evolution experiments is so complex phenomenon that it cannot directly be used to support the proteomic code idea.

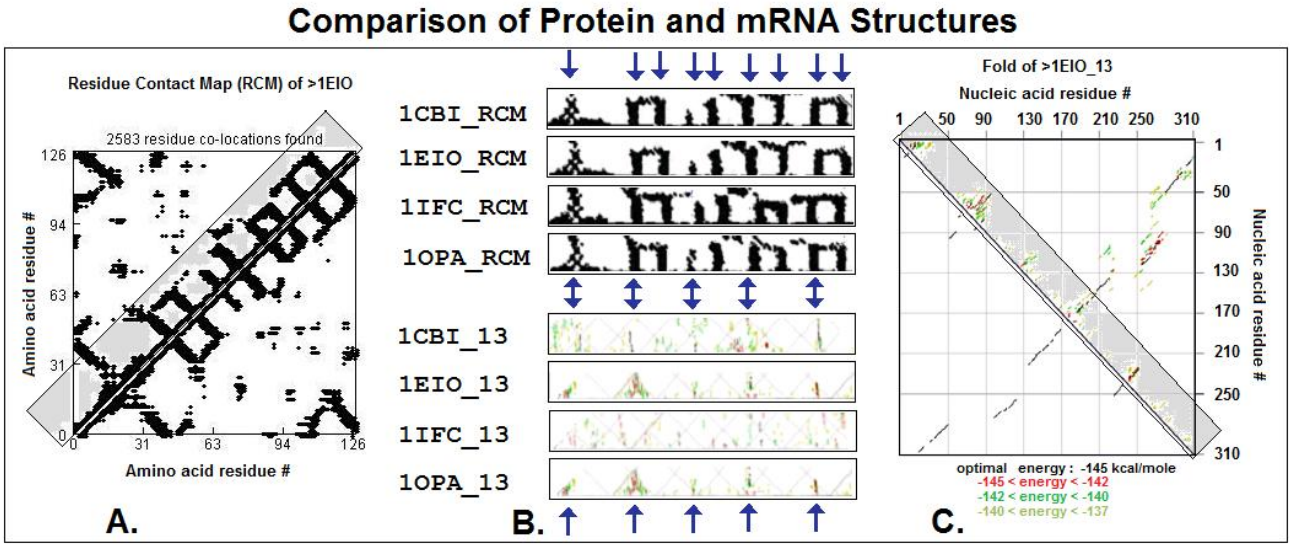
Figure 3. Comparison of 12 randomly selected protein and corresponding mRNA structures (modified from [47]). Residue contact maps (RCM) were obtained from the PBD files of the protein structures using the SeqX tool (left triangles). Energy dot plots (EDP) for the coding sequences were obtained using the mfold tool (right triangles). The two maps were aligned along a common left diagonal axis to facilitate visual comparison between the different possible representations. The black dots in the RCMs indicate amino acids that are within 6 Å of each other in the protein structure. The colored (grass-like) areas in the EDPs indicate the energetically mostly likely RNA interactions (color code in increasing order: yellow, green red, black). The coordinates indicate the number of amino acid and the corresponding nucleic acid residues.



2.8. The RNA-assisted Protein Folding

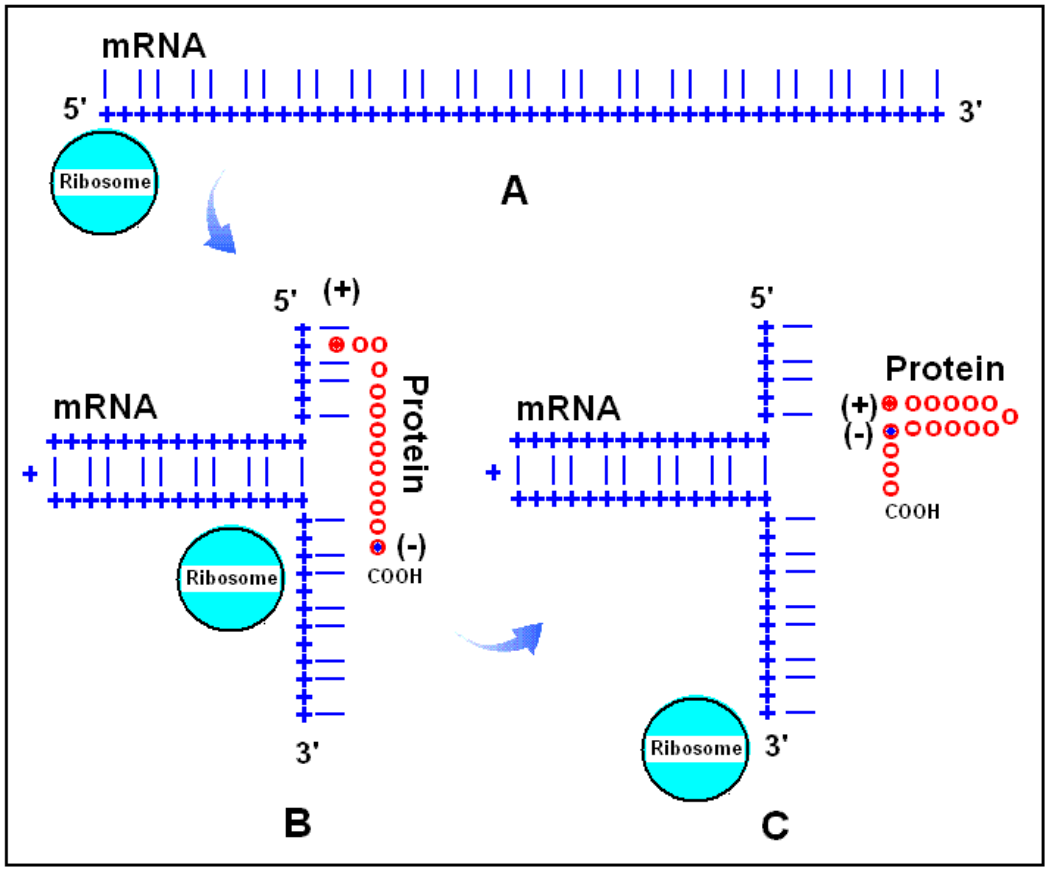
The preferential partial complementarity coding of co-locating amino acids (*The Proteomic Code*) suggests the possibility that mRNA and coded proteins may share some common structural features. Side by side comparison of 2D projections of mRNA and coded proteins seems to confirm this possibility (Figure 3, 4).

Figure 4. Comparison of the protein and mRNA structures (modified from [47]). Residue contact maps (RCM) were obtained from the PBD files of four protein structures (1CBI, 1EIO, 1IFC, 1OPA) using the SeqX tool (example on the left side of the figure). Energy dot plots (EDP) for the coding sequences were obtained using the mfold tool (example on the right side of the figure). The left diagonal portions of these two maps (in gray boxes in the examples) are compared in the central part of the figure. RNA subsequences containing only the 1st and 3rd codon letters (13) are compared. The black dots in the RCMs indicate amino acids that are within 6 Å of each other in the protein structure. The colored (grass-like) areas in the EDPs indicate the energetically most likely RNA interactions (color code in increasing order: yellow, green red, black). The blue arrows indicate the main similarities between the protein and nucleic acid structures.



Biological rules are, of course, always statistical rules, probabilities and tendencies. Nucleic acids as well as proteins have many possible configurations where one or a few are expected to dominate and define the main and characteristic configuration. Coding- and coded sequences might have their own range of more or less different folding potentials. However when a protein is generated on the surface of ribosome the coding- and coded sequences are very close to each other. This temporary intimate closeness is a possibility for coding sequences for transferring folding information to coded proteins, information that is additional to that these proteins already have in their amino acid sequences. Some ideas how it is possible are sketched in Figure 5 and 6).

Figure 5. RNA assisted protein loop formation**.** Translation begins with the attachment of the 5' end of a mRNA to the ribosome (A). Ribonucleotides are indicated by blue + and the 1st and 3rd bases in the codons by blue lines, while the 2nd base positions are left empty. A positively charged amino acid [(+) and red dots], for example arginine, remains attached to its codon. The mRNA forms a loop because the 1st and 3rd bases are locally complementary to each other in reverse orientation (B). The growing protein is indicated by red circles (o). When translation proceeds to an amino acid with especially high affinity to the mRNA-attached arginine, for example a negatively charged Glu or Asp [(-) and blue dot], the charge attraction removes the Arg from its mRNA binding site and the entire protein is released from the mRNA and completes a protein loop (C). The protein continues to grow toward the direction of its carboxy terminal (COOH). (Figure is reproduced from [77]).



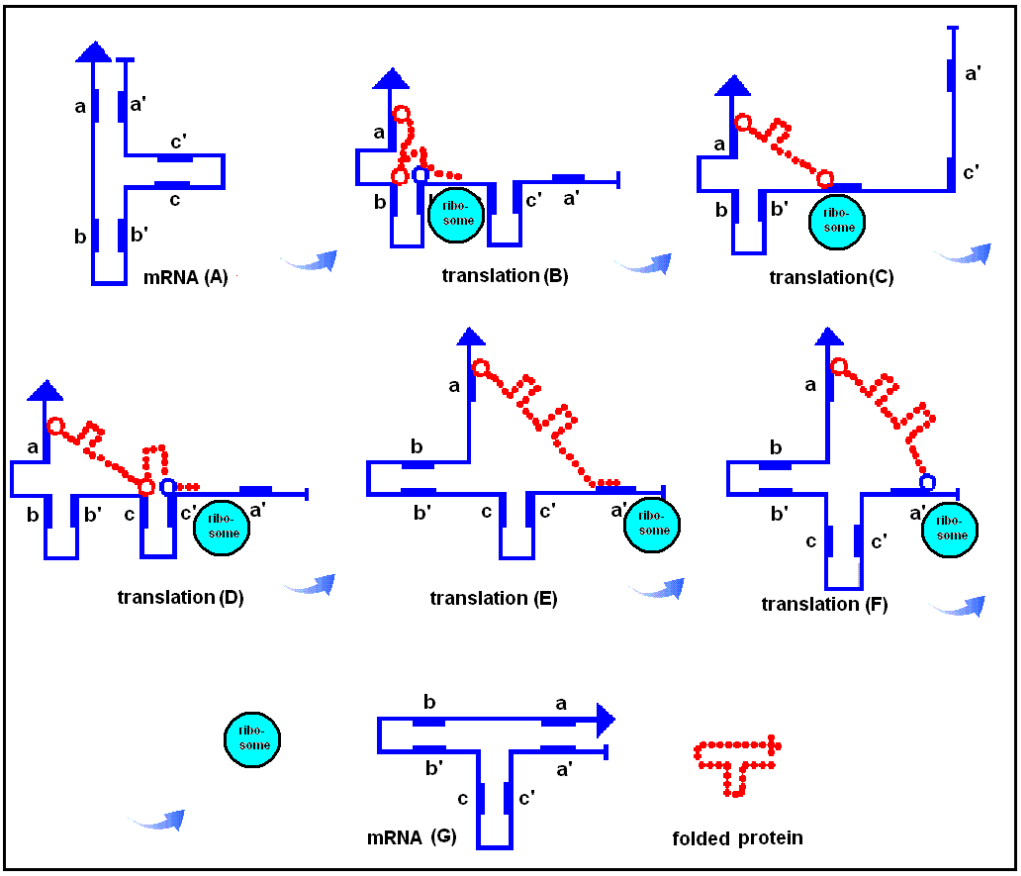
The RNA-assisted protein folding is an interesting theory which derived recently [77] from the much older theory of Proteomic Code [46] and is mainly based on bioinformatical observations. Laboratory evidence is still missing. Mutagenesis (for example) should in principle be able to give some light to this topic. It remains to see how the structural features of translational machinery allow this kind of mechanism. The involvement and role of transfer RNA (tRNA) is of course not to forget.

2.9. Evolutionary Aspects of the Redundant Genetic and Proteomic Codes.

The *Common Periodic Table of Codons and Amino Acids* is convincing evidence for the non-randomness of codon / coded amino acid connection. This connection is further emphasized by the

preferential complementary coding of co-locating amino acids. The predictability of wobble base frequency is another indicator of non-random wobble base selection and a functional network between codons. Order is traditionally seen as the result of development from the chaotic to organized, from simple to complex. This process is called evolution.

Figure 6. RNA-assisted (translational) protein folding. There are three reverse and complementary regions in a mRNA (blue line, A): a-a', b-b', c-c', which fold the mRNA into a T-like shape. During the translation process the mRNA unfolds on the surface of the ribosome, but subsequently refolds, accompanied by its translated and lengthening peptide (red dotted line, B-F). The result of translation is a temporary ribonucleotide complex, which dissociates into two T-shape-like structures: the original mRNA and the properly folded protein product (G). The red circles indicate the specific, temporary attachment points between the RNA and protein (for example a basic amino acid) while the blue circles indicate amino acids with exceptionally high affinity for the attachment points (for example acidic amino acids); these capture the amino acids at the attachment point and dissociate the ribonucleoprotein complex. Transfer-RNAs are of course important participants in translation, but they are not included in this scenario. (Figure is reproduced from [77]).



The theory of Ikehara [78, 79] about the origins of gene, genetic code, protein and life is especially interesting regarding the *Proteomic Code*. Ikehara suggests (and support with experimental evidence) that "geneprotein" system, comprised of 64 codons and 20 amino acids developed successfully during

the evolution. The development started with a GNC-type primeval genetic code (G: guanine, C: Cytosine, N: any of the four nucleotides), coding only four amino acids (Gly: [G], Ala: [A], Asp: [D], Val: [V]) forming the so called [GADV]-proteins. This minimal set of only four amino acids and the [GADV]-proteins are able to represent the 6 major (and characteristic) protein moieties/indices (hydropathy, α-helix, β-sheet and β-turn forms, acidic amino acid content and basic amino acid content) which are necessary for appropriate three-dimensional structure formation of globular, water-soluble proteins on the primitive earth. The [GADV]-proteins (even randomized) have catalytic properties and able to facilitate the syntheses of other [GADV]-proteins (also random).

The primeval genetic code continued to develop toward a more complex SNS-type primitive genetic code (S: G or C) containing 16 codons and encoding 10 amino acids (L, P, H, Q, R, V, A, D, E, G) before the recent 64 codon/20 amino acid-type genetic code became established.

Furthermore, Ikehara concluded from the analysis of microbial genes that newly-born genes are products of nonstop frames (NSF) on antisense strands of microbial GC-rich genes [GC-NSF(antisense)] and from SNS repeating sequences [(SNS)n] similar to the GC-NSF(antisense).

The similarity between GNC/SNS-type primitive codons (which are expressed even from the reverse-complement strands as GC-rich non-stop genes) and the *Proteomic Code* is obvious. Both concepts suggest and agree with each other regarding a) the connection between 2nd codon residue and the fundamental physicochemical properties of the coded amino acids, b) the importance of 1st and 3rd codon letters in determining the nucleic acid (as well as protein) structure, c) the importance of compositional difference between 1st, 3rd and central codon residues (to emphasize the codon boundaries), d) the importance of complementarity (even in the mRNA) in development of protein structure and function, e) the importance of GC at the 1st and 3rd codon positions (as the source of lower Gibbs energy (dG), than central codon positions have, where even AT are permitted). I think that the concept of GNC/SNS-type primitive codons and the Proteomic Code are convergent ideas, both reflecting the same fundamental aspects of the connection between nucleic acid and protein structure and function.

3. Conclusions

The Nirenberg's *Code* seems to have another face that was in the shadow some 40+ years. We start to see, that the *Genetic Code* actually might contain all characteristics that was expected and promised by the early theoretical code models.

1) Codon boundaries are physico-chemically defined to a certain degree, which theoretically should give some protection against frame-shifts.

2) Codon residues are not randomly assigned, but there is a connection between codon architecture and the physicochemical properties of the coded proteins.

3) Amino acids preferentially interact with their codons (studied in restrictions endonucleases).

4) Co-locating amino acids are preferentially coded by partially complementary codons which create inter-connectivity between structurally important amino acids.

5) Wobble bases are not randomly assigned at all, their frequency is statistically well predictable from the frequency of bases at other codon positions.

6) Wobble base redundancy makes it possible the development of codon integration in coding sequences which might be used for compensatory mutations. This is the second line of defense against mutations (after the known tolerance provided by the coding redundancy).

7) The internally inter-connected and integrated system of codons makes it possible that coding sequences provide a mold for structure forming of coded proteins and function as *nucleic acid chaperons* providing the missing molecular information to correct protein folding.

It is concluded that the redundant *Genetic Code* contains biological information which is additional to the 64/20 definition of amino acids. This additional information is used to define the 3D structure of coding nucleic acids and coded proteins and is called the *Proteomic Code*.

Acknowledgements

The author wishes to thank for the friendly attention and advices of Dr. G.L.G. Miklos (Secure Genetics Pty Ltd, Australia). The continuous support and help of Dr. P. S. Agutter (BMC, *Theor. Biol. and Med. Modelling*) is very much acknowledged.

References and Notes

1. Hayes, B. The invention of the Genetic Code. *Amer. Sci.* **1998**, *1*, 8-14.
2. Franklin, R.; Gosling, R. Molecular configuration in sodium thymonucleate. *Nature* **1953**, *171*, 740-741.
3. Watson, J.D.; Crick, F.F.C. A structure for deoxyribose nucleic acids. *Nature* **1953**, *171*, 737-739.
4. Gamow, G. Possible relation between deoxyribonucleic acid and protein structures. *Nature* **1954**, *173*, 318-318.
5. Gamow, G. Possible mathematical relation between deoxyribonucleic acid and proteins. *Det Kongelige Danske Videnskabernes Selskab, Biologiske Meddelelser* **1954**, *22*, 1-13.
6. Gamow, G.; Rich, A.; Ycas, M. The problem of information transfer from nucleic acids to proteins. *Adv. Bio. Med. Phys.* **1956**, *4*, 23-68.
7. Brenner, S. On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. *Proc. Nat. Acad. Sci. USA* **1957**, *43*, 687-694.
8. Crick, F.H.C.; Griffith, J.S.; Orgel, L.E. Codes without commas. *Proc. Nat. Acad. Sci. USA* **1957**, *43*, 416-421.
9. Nirenberg, M.W.; Matthaei, J.H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Nat. Acad. Sci. USA* **1961**, *47*, 1588-1602.
10. Leder, P.; Nirenberg, M. RNA code words and protein synthesis, II. Nucleotide sequence of a valine RNA codeword. *Proc. Nat. Acad. Sci. USA* **1964**, *52*, 420-427.
11. Workman, C.; Krogh, A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* **2002**, *24*, 4816-4822.
12. *The Nucleic Acid Database Project*. Rutgers, The State University of New Jersey 2008. http://ndbserver.rutgers.edu/index.html. Accessed November 26, 2008.

13. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **2003**, *31*, 3406-3415.
14. Biro, J.C. Correlation between nucleotide composition and folding energy of coding sequences with special attention to wobble bases. *Theor. Biol. Med. Model.* **2008**, *5*, 14.
15. Meyer, I.M.; Miklós, I. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.* **2005**, *33*, 6338-6348.
16. Katz, L.; Burge, C.B. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* **2003**, *13*, 2042-2051.
17. Hatfield, G.W.; Roth, D.A. Optimizing scaleup yield for protein production: Computationally Optimized DNA Assembly (CODA) and translation engineeringtrade mark. *Biotechnol. Annu. Rev*. **2007**, *13*, 27-42.
18. Zhang, W.; Xiao, W.; Wei, H.; Zhang, J.; Tian, Z. mRNA secondary structure at start AUG codon is a key limiting factor for human protein expression in Escherichia coli. *Biochem. Biophys. Res. Commun*. **2006**, *349*, 69-78.
19. *Codon Usage Database, NCBI-GenBank Flat File Release 160.0* [June 15, 2007]. http://www.kazusa.or.jp/codon/. Accessed November 26, 2008.
20. Biro, J.C. Indications that "codon boundaries" are physico-chemically defined and that protein-folding information is contained in the redundant exon bases. *Theor. Biol. Med. Model.* **2006**, *3*, 28.
21. Shabalina, S.A.; Ogurtsov, A.Y.; Spiridonov, N.A. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* **2006**, *34*, 82428-2437.
22. Stoletzki, N.; Eyre-Walker, A. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol. Biol. Evol*. **2007**, *24*, 374-381.
23. Smith, N.G.; Eyre-Walker, A. Why are translationally sub-optimal synonymous codons used in Escherichia coli? *J. Mol. Evol.* **2001**, *53*, 225-236.
24. Akashi, H. Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy. *Genetics* **1994**, *136*, 927-935.
25. Zhang, J.; Long, M.; Li, L. Translational effects of differential codon usage among intragenic domains of new genes in *Drosophila. Biochim. Biophys. Acta (BBA) - Gene Struct. Express.* **2005**, *1728*, 135-142.
26. Pelc, S.R.; Welton, M.G.E. Stereochemical relationship between coding triplets and amino-acids. *Nature* **1966**, *209*, 868-870.
27. Welton, M.G.E.; Pelc, S.R. Specificity of the Stereochemical relationship between ribonucleic acid-triplets and amino-acids. *Nature* **1966**, *209*, 870-872.
28. Crick, F.H.C. An Error in Model Building. *Nature* **1967**, *213*, 798-798.
29. Woese, C.R. The molecular basis for gene expression. In *The Genetic Code*; Harper & Row: New York, 1967; Chapters 6-7, pp. 156-160.
30. Biro, J.C.; Benyo, B.; Sansom, C.; Szlavecz, A.; Fordos, G.; Micsik, T.; Benyo, Z. A common periodic table of codons and amino acids. *Biochem. Biophys. Res. Commun.* **2003**, *306*, 408-415.
31. Biro, J.C.; Biro, J.M.K. Frequent occurrence of recognition Site-like sequences in the restriction endonucleases. *BMC Bioinformat.* **2004**, *5,* 30.

32. Biro, J. Comparative analysis of specificity in protein-protein interactions. Part I: A theoretical and mathematical approach to specificity in protein-protein interactions. *Med. Hypotheses* **1981**, *7*, 969-979.
33. Biro, J. 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* **1981**, *7*, 981-993.
34. Biro, J. Comparative analysis of specificity in protein-protein interactions. Part III: Models of the gene expression based on the sequential complementary coding of some pituitary proteins. *Med. Hypotheses* **1981**, *7*, 995-1007.
35. Mekler, L.B. Specific selective interaction between amino acid groups of polypeptide chains. *Biofizika* **1969**, *14*, 581-584.
36. Mekler, L.B.; Idlis, R.G. *VINITI Deposited Doc* **1981**, 1476-1481.
37. Blalock, J.E.; Bost, K.L. Binding of peptides that are specified by complementary RNAs. *Biochem. J.* **1986**, *234*, 679-683.
38. Segerstéen, U.; Nordgren H.; Biro, J.C. Frequent occurrence of short complementary sequences in nucleic acids. *Biochem. Biophys. Res. Commun.* **1986**, *139*, 94-101.
39. Blalock, J.E.; Smith, E.M. Hydropathic anti-complementarity of amino acids based on the genetic code. *Biochem. Biophys. Res. Commun.* **1984**, *121*, 203-207.
40. Root-Bernstein, R.S. Amino acid pairing. *J. Theor. Biol.* **1982**, *94*, 885-894.
41. Siemion, I.Z.; Stefanowicz, P. Periodical changes of amino acid reactivity within the genetic code. *Biosystems* **1992**, *27*, 77-84.
42. Heal, J.R.; Roberts, G.W.; Raynes, J.G.; Bhakoo, A.; Miller, A.D. Specific interactions between sense and complementary peptides: the basis for the proteomic code. *Chembiochemistry* **2002**, *3*, 136-151. (Review. Erratum in: *Chembiochemistry* **2002**, *3,* 271).
43. Baranyi, L.; Campbell, W.; Ohshima, K.; Fujimoto, S.; Boros, M.; Okada, H. The antisense homology box: A new motif within proteins that encodes biologically active peptides. *Nat. Med.* **1995**, *1*, 894-901.
44. Biro, J.C.; Fördös, G. SeqX a tool to detect, analyze and visualize residue co-locations in protein and nucleic acid structures. *BMC Bioinformat.* **2005**, *6*, 170.
45. Biro, J.C. Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theor. Biol. Med. Model.* **2006**, *3*, 15.
46. Biro, J.C.; The Proteomic Code: A molecular recognition code for proteins. Review. *Theor. Biol. Med. Model.* **2007**, *4*, 45.
47. Biro, J.C. Protein folding information in nucleic acids which is not present in the genetic code. *Ann. NY Acad. Sci.* **2006**, *1091*, 399-411.
48. Anfinsen, C.B.; Redfield, R.R.; Choate, W.I.; Page, J.; Carroll, W.R. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *J. Biol. Chem.* **1954**, *207*, 201-210.
49. Makhatadze, G.I.; Privalov, P.L. Energetics of Protein Structure. In *Advances in Protein Chemistry*. Anfinsen, C.B.; Edsall, J.T., Richards, F.M., Eds.; Academic Press: New York, 1995; Volume 47, pp. 308-405.
50. Thirumalai. D.; Hyeon C. RNA and protein folding: common themes and variations. *Biochemistry* **2005**, *44*, 4957-4970.

51. Thirumalai, D.; Lorimer G.H. Chaperonin-mediated protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 245-69.
52. Levinthal, C. How to fold graciously in Mossbauer spectroscopy in biological systems. *In Proceedings of a Meeting held at Allerton House*; Debrunner, P., Tsibris, J.C.M., Munck, E., Urbana, I.L., Eds.; University of Illinois Press: Monticello; 1969; pp. 22-24.
53. Grantcharova, V.; Alm, E.J.; Baker, D.; Horwich, A.L. Mechanisms of protein folding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 70-82.
54. Hartl, F.U.; Hayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **2002**, *295*, 1852-1858.
55. Genevaux, P.; Georgopoulos, C.; Kelley, W.L. The Hsp70 chaperone machines of Escherichia coli: a paradigm for the repartition of chaperone functions. *Mol. Microbiol.* **2007**, *66*, 840-857.
56. Thanaraj, T.A.; Argos, P. Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci.* **1996**, *5*, 1973-1983.
57. Brunak, S.; Engelbrecht, J. Protein structure and the sequential structure of mRNA: alpha-Helix and beta-sheet signals at the nucleotide level. *Proteins* **1996**, *25*, 237-252.
58. Gupta, S.K.; Majumdar, S.; Bhattacharya, T.K.; Ghosh. T.C. Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem. Biophys. Res. Commun.* **2000**, *269*, 692-696.
59. Chiusano, M.L.; Alvarez-Valin, F.; DI Giulio, M. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene* **2000**, *261*, 63-69.
60. Gu, W.; Zhoa, T.; Ma, J.; Sun, X.; Lu, Z. The relationship between synonymous codon usage and protein structure in *Escherichia coli* and Homo sapiens. *Biosystems* **2004**, *73*, 89-97.
61. Ermolaeva, O. Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* **2001,** *3,* 91-97.
62. Adzhubei, I.A.; Adzhubei, A.A. ISSD Version 2.0: Taxonomic range extended. *Nucleic Acids Res.* **1999**, *27*, 268-271.
63. Biro, J.C. Does codon bias have an evolutionary origin? *Theor. Biol. Med. Model.* **2008**, *5*, 16.
64. Kimchi-Sarfaty, C.; Oh, J.M.; Kim, I.W.; Sauna, Z.E.; Calcagno, A.M.; Ambudkar, S.V.; Gottesman, M.M. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **2007**, *315*, 525-528.
65. Sauna, Z.E.; Kimchi-Sarfaty, C.; Ambudkar, S.V.; Gottesman, M.M. Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer. *Cancer Res.* **2007**, *67*, 9609-9612.
66. Duan, J.; Wainwright, M.S.; Comeron, J.M.; Saitou, N.; Sanders, A.R.; Gelernter, J.; Gejman, P.V. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **2003**, *12*, 205-216.
67. Pagani, F.; Raponi, M.; Baralle, F.E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Nat. Acad. Sci. USA* **2005**, *102*, 6368-6372.
68. Nielsen, K.B.; Sorensen, S.; Cartegni, L.; Corydon, T.J.; Doktor, T.K.; Schroeder, L.D.; Reinert, L.S.; Elpeleg, O.; Krainer, A.R.; Gregersen, N.; Kjems, J.; Andresen, B.S. Seemingly neutral polymorphic variants may confer immunity to splicing inactivating mutations: A synonymous SNP in exon 5 of MCAD protects from deleterious mutations in a flanking exonic splicing enhancer. *Am. J. Hum. Genet.* **2007**, *80*, 416-432. Erratum in: *Am. J. Hum. Genet.* **2007**, *80*, 816.

69. Sauna, Z.E.; Kimchi-Sarfaty, C.; Ambudkar, S.V.; Gottesman, M.M. The sounds of silence: Synonymous mutations affect function. *Pharmacogenomics* **2007**, *8*, 527-532.
70. Komar, A.A.; Genetics. SNPs, silent but not invisible. *Science* **2007**, *315*, 466-467.
71. Soares, C. Codon spell check. Silent mutations are not so silent after. *Sci. Am.* **2007**, *296*, 23-24.
72. Drake, J.W. Too many mutants with multiple mutations. *Crit. Rev. Biochem. Mol. Biol.* **2007**, *42*, 247-258.
73. Drake, J.W.; Bebenek, A.; Kissling, G.E.; Peddada, S. Clusters of mutations from transient hypermutability. *Proc. Nat. Acad. Sci. USA* **2005**, *102*, 12849-12854.
74. Poon, A.; Chao. L. The rate of compensatory mutation in the DNA bacteriophage φX174. *Genetics* **2005**, *170*, 989–999.
75. Plotnikova, O.V.; Kondrashov, F.A.; Vlasov, P.K.; Grigorenko, A.P.; Ginter, E.K.; Rogaev, E.I. Conversion and compensatory evolution of the gamma-crystallin genes and identification of a cataractogenic mutation that reverses the sequence of the human CRYGD gene to an ancestral state. *Am. J. Hum. Genet*. **2007**, *81*, 32-43.
76. Kim, H.; Shen, T.; Sun, D.P.; Ho, N.T.; Madrid, M.; Tam, M.F.; Zou, M.; Cottam, P.F.; Ho, C. Restoring allosterism with compensatory mutations in hemoglobin. *Proc. Nat. Acad. Sci. USA* **1994**, *91*, 11547-11551.
77. Biro, J.C. Nucleic acid chaperons: a theory of an RNA-assisted protein folding. *Theor. Biol. Med. Model.* **2005**, *2*, 35.
78. Ikehara, K. Origins of gene, genetic code, protein and life: Comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis. *J. Biosci.* **2002**, *27*, 165-186.
79. Ikehara, K.; Omori, Y.; Arai, R.; Hirose, A. A novel theory on the origin of the genetic code: A GNC-SNS hypothesis. *J. Mol. Evol.* **2002**, *54*, 530-538.