International Journal of Molecular Medicine and Advance Sciences 2(3): 251-263, 2006

Protein Folding Information in Nucleic Acids Which Is Not Present in the Genetic Code

Jan C. Biro
Homulus Foundation, San Francisco, CA, USA

Abstract: All the information necessary for protein folding is supposed to be present in the amino acid sequence. It is still not possible to provide specific ab initio structure predictions by bioinformatical methods. It is suspected that additional folding information is present in protein coding nucleic acid sequences, which is not represented by the known genetic code. Nucleic acid subsequences comprising the 1st and/or 3rd codon residues in mRNAs express significantly higher Free Folding Energy (FFE) than the subsequence containing only the 2nd residues ($p<0.0001$, $n=81$). This periodic FFE difference is not present in introns and therefore it is a specific physico-chemical characteristic of coding sequences and it might contribute to unambiguous definition of codon boundaries during translation. The FFE in the 1st and 3rd residues is additive, which suggests that these residues contain a significant number of complementary bases and contribute to selection for local RNA secondary structures in coding regions. This periodic, codon-related structure-forming of mRNAs indicates a connection between the structure of exons and the corresponding (translated) proteins. The folding energy dot plots of RNAs and the residue contact maps of the coded proteins are indeed similar. Residue contact statistics using 81 different protein structures confirmed that amino acids that are coded by partially reverse and complementary codons (Watson-Crick (WC) base pairs at the 1st and 3rd codon positions and translated in reverse orientation) are preferentially co-located in protein structures. Exons are distinguished from introns and codon boundaries are physico-chemically defined by periodically distributed FFE differences between codon positions. There is a selection for local RNA secondary structures in coding regions and this nucleic acid structure resembles the folding profiles of the coded proteins. The preferentially (specifically) interacting amino acids are coded by partially complementary codons, which strongly supports the connection between mRNA and the corresponding protein structures and indicates that there is protein folding information in nucleic acids that is not present in the genetic code. This might give some additional explanation of codon redundancy.

Key words: Codon, translation, protein folding, RNA folding, protein interaction, complementarity, protein design, anfinsen, protein interaction

INTRODUCTION

The protein folding problem has been one of the grand challenges in computational molecular biology. The problem is to predict the native three-dimensional structure of a protein from its amino acid sequence. It is widely believed that the amino acid sequence contains all the necessary information to make up the correct three-dimensional structure, since protein folding is apparently thermodynamically determined; i.e., given a proper environment, a protein will fold up spontaneously. This is called Anfinsen's thermodynamic principle[1].

The thermodynamic principle has been confirmed many times on many different kinds of proteins in vitro. Critics says that the *in vivo* chemical conditions are different from those *in vitro*, the correct folding is determined by interactions with other molecules (chaperons, hormones, substrate, etc.) and protein folding is much more complex than re-naturation of de-natured poly amino acids. The fact that many naturally occurring proteins fold reliably and quickly to their native state, despite the astronomical number of possible configurations, has come to be known as Levinthal's Paradox[2].

Anfinsen's principle was formulated in the 1960s using purely chemical experiments and a lot of intuition. Today, we have a lot of sequences and structures available to establish a logical and understandable link between sequence, structure and function. But it is still not possible to correctly predict the structure (or a range of possible structures) purely from the sequence, ab initio and in silico[3].

Corresponding Author: Jan C. Biro, 88 Howard, #1205, San Francisco, 94195 CA, USA

There are two potential, external sources of additional and specific protein folding information: (a) the chaperons (other proteins that assist in the folding of proteins and nucleic acids[4] and (b) the protein coding nucleic acid sequences themselves (which are templates of the protein syntheses, but are not defined as chaperons).

The idea that the nucleotide sequence itself could modulate translation and hence affect co-translational folding and assembly of proteins has been investigated in a number of studies[5-7]. Studies on the relationships between synonymous codon usage and protein secondary structural units are especially popular[8-10]. The genetic code is redundant (61 codons code 20 amino acids) and as many as 6 synonymous codons can code the same amino acid (Arg, Leu, Ser). The "wobble" base has no effect on the meaning of most codons but still the codon usage (wobble usage) is not randomly defined[11,12] and there are well known, stable species-specific differences in the codon usage. It seems to be logical to search for some meaning (biological purpose) of the wobble bases and try to associate them with protein folding.

Another observation concerning the code redundancy dilemma is that there is a widespread selection (preference) for local RNA secondary structure in protein coding regions[13]. A given protein can be encoded by a large number of distinct mRNA species, potentially allowing mRNAs to simultaneously optimize desirable RNA structural features in addition to their protein coding function. The immediate question is whether there is some logical connection between the possible, optimal RNA structures and the possible, optimal biologically active protein structures.

MATERIALS AND METHODS

Single-stranded RNA molecules can form local secondary structures through the interactions of complementary segments. WC base pair formation lowers the average free energy, dG, of the RNA and the magnitude of change is proportional to the number of base pair formations. Therefore the Free Folding Energy (FFE) is used to characterize the local complementarity of nucleic acids[13]. The free folding energy is defined as $FFE=(dG_{shuffled}- dG_{native})/L\times100$, where L is the length of the nucleic acid, i.e., free energy difference between native and shuffle (randomized) nucleic acids per 100 nucleotides. Higher positive values indicate stronger bias toward secondary structure in the native mRNA and negative values indicate bias against secondary structure in the native mRNA.

We used a nucleic acid secondary structure predicting tool, the mfold[14] to obtain dG values and the lowest dG was used to calculate the FFE. The mfold also provided the folding energy dot plots, which are very useful to visualize the energetically most favored structures in a 2D matrix.

A series of JAVA tools were used: SeqX to visualize the protein structures in 2D as amino acid residue contact maps[15]; SeqForm for selection of sequence residues in predefined phases (every third in our study)[16]; SeqPlot for further visualization and statistical analyses of the dot-plot views[17]; Dotlet as a standard dot-plot viewer[18]. Structural data were downloaded from PDB[19], NDB[20] and from a wobble base oriented database called Integrated Sequence-Structure Database (ISSD)[21].

Structures were generally randomly selected regarding species and biological function (a few exceptions are mentioned in the Results). Care was taken to avoid very similar structures in the selections. A propensity for alpha helices was monitored during selection and structures with very high and very low alpha helix content were also selected to make sure of a wide range of structural representation.

Linear regression analyses and Student's t-tests were used for statistical analyses of the results.

RESULTS

Observations were made on human peptide hormone structures. This group of proteins is very well defined and annotated, the intron-exon boundaries are known and even intron data are easily accessible. The coding sequences were phase separated by SeqForm into three

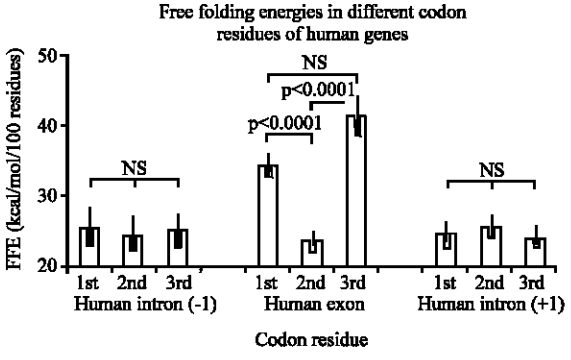

Fig. 1: Free Folding Energies (FFE) in different codon residues of human genes. The coding sequences (exons) of 18 human hormone genes and the preceding (–1) and following (+1) sequences (introns) were phase separated into three subsequences each corresponding to the 1st, 2nd and 3rd codon positions in the coding sequence. The dG values were determined by mfold and the FFE was calculated. Each bar represents the mean±SEM, n=18

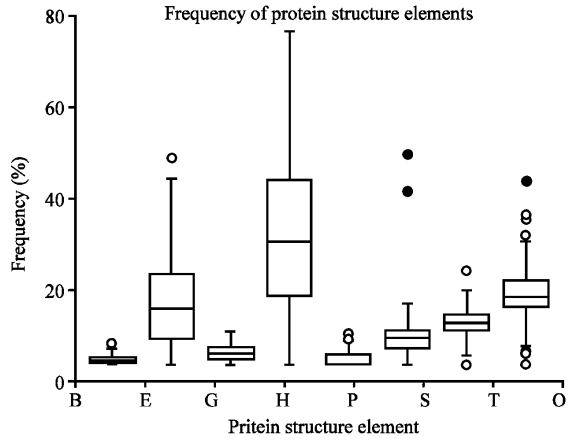

Fig. 2: Frequency of protein structure elements. Box plot representation of protein secondary structure elements in 81 structures. L=317±20 (mean±SEM, n=81). Secondary structure codes: H, alpha helix; B, residue in isolated beta bridge; E, extended strand, participates in beta ladder; G, 3-helix (3/10 helix); I, 5 helix (pi helix); P, polyproline type II helix (left-handed); T, hydrogen bonded turn; S, bend

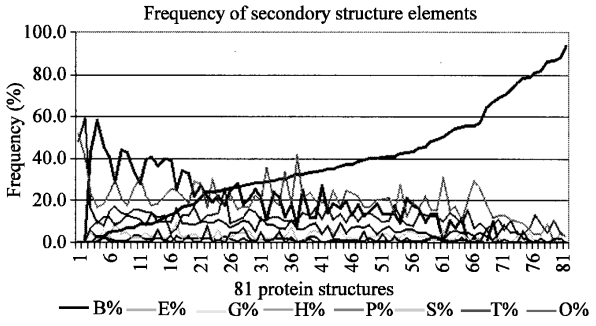

Fig. 3: Frequency of secondary structure elements. The propensity of different structural elements in 81 different proteins is shown. L=317±20 (mean±SEM, n=81). Secondary structure codes: H, alpha helix; B, residue in isolated beta bridge; E, extended strand, participates in beta ladder; G, 3-helix (3/10 helix); I, 5 helix (pi helix); P, polyproline type II helix (left-handed); T, hydrogen bonded turn; S, bend

subsequences, each containing only the 1st, 2nd or 3rd letters of the codons. Similar phase separation was made for intronic sequences immediately before and after the exon. There are, of course, no known codons in the intronic sequences, therefore we continued the same phase that we applied for the exon, assuming that this kind of selection is correct and maintained the name of the phase denotation even for non-coding regions. Subsequences corresponding to the 1st and 3rd codon

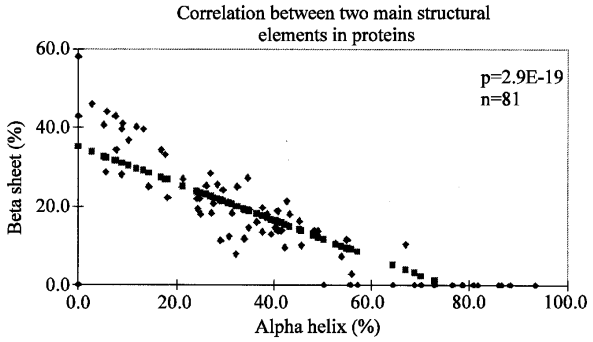

Fig. 4: Correlation between two main structural elements in proteins. Data was taken from Fig. 3 (H, alpha helix; E, beta sheet)

letters in the coding regions had significantly higher FFEs than subsequences corresponding to the 2nd codon letters. No such difference was seen in non-coding regions (Fig. 1).

In a larger selection of 81 different protein structures, the corresponding protein and coding sequences were used to extend the observations. These 81 proteins were represented different (randomly selected) species and different (also randomly selected) protein functions and therefore the results might be regarded as more generally valid. The propensity of different secondary structure elements was recorded (as annotated in different databases) (Fig. 2).

The proportion of alpha helices varied from 0 to 90% in the 81 proteins and showed a significant negative correlation to the proportion of beta sheets (Fig. 3 and 4).

The original observation made on human hormone proteins, that significantly more free folding energy is associated with the 1st and 3rd codon residues than with the 2nd was confirmed on a larger and more heterogeneous protein selection. A significant difference showed up even between the 1st and 3rd residues in this larger selection (Fig. 5).

There is a correlation between the protein structure and the FFE associated with codon residues. The correlation is negative between FFE associated with the 2nd (middle) codon residues and the alpha helix content of the protein structure. The correlation is especially significant when the FFE ratios are compared to the helix/sheet ratios (Fig. 6 and 7).

The alpha helix is the most abundant structure element in proteins. It shows negative correlation to the frequency of the second most prominent protein structure, the beta sheet. The propensity of some amino acids and the major physico-chemical characteristics (charge and polarity) shows significant correlation (positive or negative) to this structural feature. We

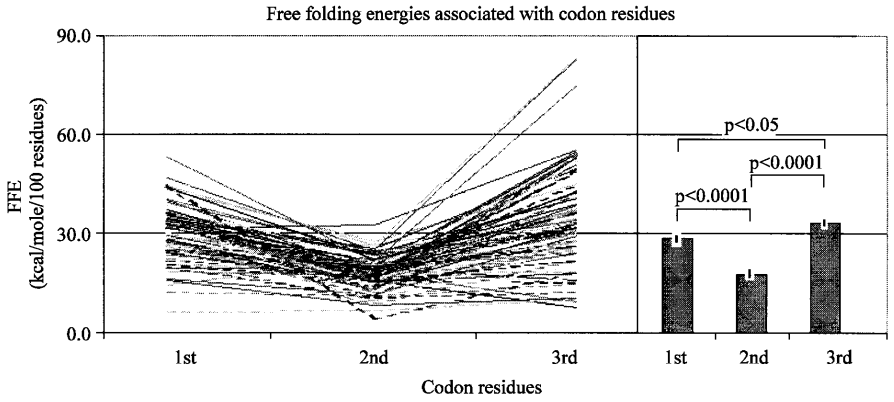

Fig. 5: Free folding energies associated with codon residues. Free Folding Energies (FFE) ware determined in phase-selected subsequences of 81 different protein coding nucleic acids. The lines indicate individual values (left part of the Figure), while the bars (right part of the Figure) indicate the mean±SEM (n=81)

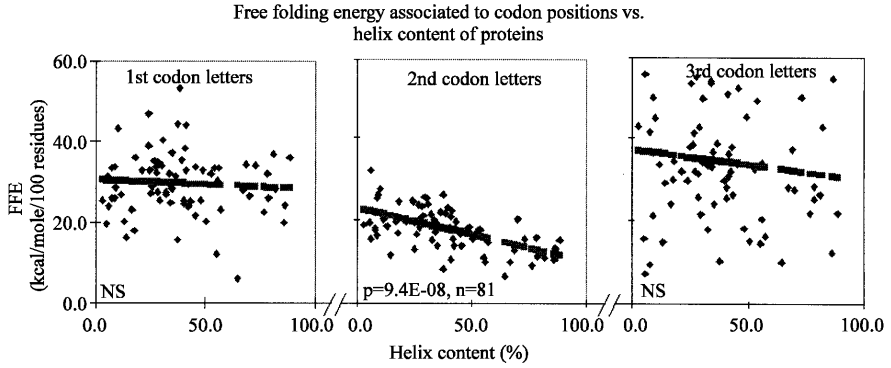

Fig. 6: Free folding energy associated with codon positions vs. helix content of proteins. Linear regression analyses where pink symbols represent the linear regression line

include statistical analyses of alpha helix content and other protein characteristics in this article to show the complexity behind the term alpha helix and to show the insecurity in interpreting any correlation to this structural feature (Fig. 8 and 9). Detailed analyses of these data are outwith the scope of this study.

Higher FFE in subsequences of 1st and 3rd codon residues than in the 2nd indicates the presence of a larger number of complementary bases at the right positions of these subsequences. However, this might be the study only because the first and last codons form simpler subsequences and contain longer repeats of the same nucleotide than the 2nd codons. This would not be surprising for the 3rd (wobble) base but would not be expected for the 1st residue, even though it is known that the central codon letters are the most important to distinguish between amino acids (as shown in the in the Common Periodic Table of Codons and Amino Acids[22]. It is more significant to see that the FFEs in 1st and 3rd residues are additive and together they represent the entire FFE of the intact mRNA (Fig. 10).

Higher FFE at the 1st and 3rd codon positions than at 2nd indicates that the number of complementary bases (a-t and g-t) is higher in the 1st and 3rd subsequences than in the second. This is possible only if more complementers are in 1-1, 1-3, 3-1, 3-3 position pairs than in 1-2, 2-1, 2-3, 3-2 position pairs. We wanted to know whether the 1-1, 3-3 (complement) or the 1-3, 3-1 (reverse-complement) pairing is more predominant.

The length of phase-separated nucleic acid subsequences (l) is a third of the original coding sequence (L). The number of different residues (a, t, g and c) varies at different codon positions (1, 2, 3).

$$a1+u1+g1+c1=a2+t2+g2+c2=a3+t3+g3+c3=l=L/3$$

The highest number of complementary pairs might occur in the 1st subsequence if

$$a1=t1,\ g1=c1 \text{ and } a1/t1=g1/c1=1$$

If, for example, a1>t1, g1=c1 an excess of unpaired a1 occurs and a1/t1>g1/c1=1 and the possible FFE in

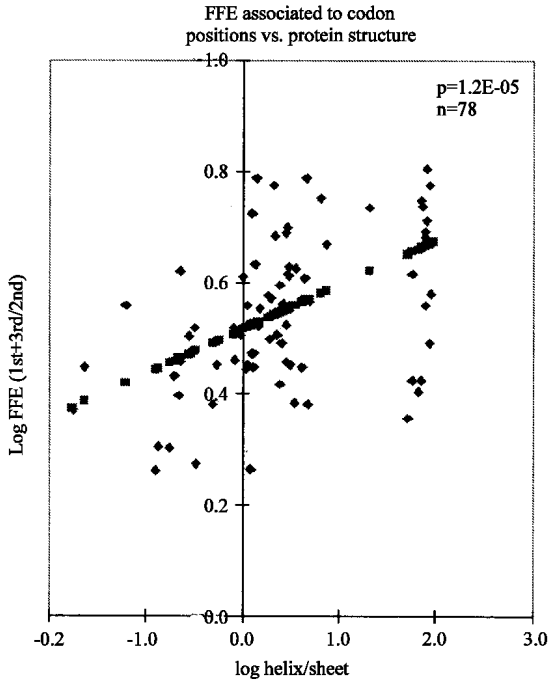

Fig. 7: FFE associated with codon positions vs. protein structure. Same data as in Fig. 6 after calculating ratios and log transformation. Linear regression analyses where pink symbols represent the linear regression line

subsequence 1 will be less. Following the same logic for other pairs in other subsequences we can conclude that any deviation from a/t=g/c=1 is suboptimal regarding the FFE. Counting the different residue ratios and combinations indicates that the optima are obtained if the residues in the first position form WC pairs with residues at the third positions (1-3) and vice versa (3-1). This is consistent with the expectation that mRNA will form local loops, in which the direction of more or less double stranded sequences is reversed and (partially) complemented. (Fig. 11).

The partial (suboptimal) reverse complementarity of codon-related positions in nucleic acids suggested some similarity between protein structures and the possible structures of the coding sequences. This possibility was examined by visual comparison of 16 randomly selected protein residue contact maps and the energy dot plots of the corresponding RNAs. We could see similarities between the two different kinds of maps (Fig. 12). However, this type of comparison is not quantitative and statistical evaluation is not directly possible.

Another similar, but still not quantitative, comparison of protein and coding structures was performed on four proteins that are known to have very similar 3D structures but their primary structure (the sequence) is less than 30% similar, as well as the

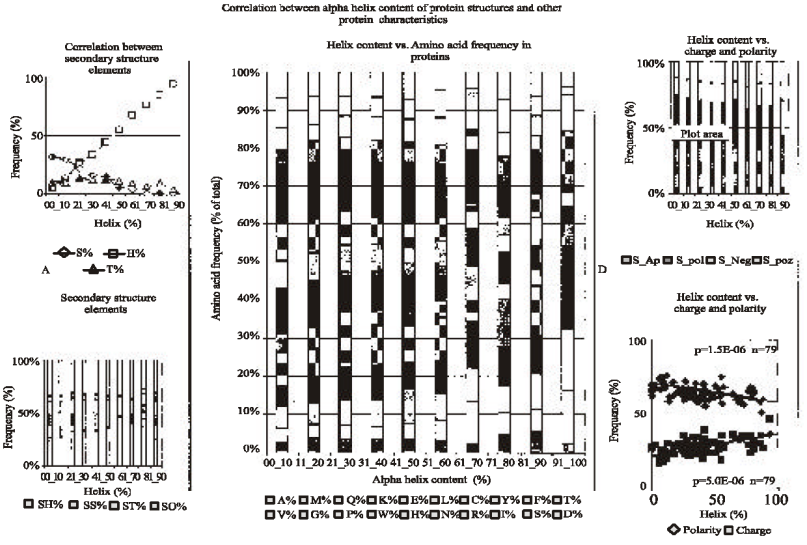

Fig. 8: Correlation between alpha helix content of protein structures and other protein characteristics. The alpha helix content of 80 protein structures was compared to the frequency of other major structural elements (A,B), the frequency of individual amino acids (C) and the frequency of charged and hydrophobe residues (D,E). (A) The correlation between helix (H), beta sheet (S) and turn (T); (B) the proportions between the sum of helices (SH), beta strands (SS), turns (ST) and all other structural elements (TO). (D) The proportion between the sums of apolar (S_Ap), polar (S_Pol), negatively charged (S_Neg) and positively charged (S_Poz) amino acids. (E) The linear regression analyses correlation between helix content and the percentage of polar + apolar (Polarity) and positively + negatively charged (Charge) residues

Correlation between frequency of individual amino acids and the main secondary structure elements in proteins

	A%	C%	D%	E%	F%	G%	H%	I%	K%	L%	M%	N%	P%	Q%	R%	S%	T%	V%	W%	Y%
Helix (%)	3.E-04	1.E-04	2.E-01	1.E-05	2.E-06	2.E-06	4.E-01	2.E-01	1.E-04	9.E-04	2.E-02	1.E-01	2.E-06	1.E-04	8.E-02	2.E-01	3.E-03	4.E-03	1.E-02	6.E-02
Helix correlation	POS	NEG	NS	POS	NEG	NEG	NS	NS	POS	POS	POS	NS	NEG	POS	NS	NS	NEG	NEG	NEG	NS
Sheet (%)	2.E-03	6.E-05	6.E-02	3.E-05	5.E-06	6.E-06	5.E-01	3.E-01	1.E-03	5.E-05	4.E-02	6.E-02	9.E-02	5.E-03	4.E-02	5.E-02	8.E-04	8.E-03	2.E-01	4.E-02
Sheet correlation	NEG	POS	NS	NEG	POS	POS	NS	NS	NEG	NEG	NEG	NS	NS	NEG	NEG	NS	POS	POS	NS	POS
Turn (%)	9.E-01	2.E-01	1.E-01	2.E-01	1.E-01	2.E-03	5.E-01	3.E-01	7.E-02	2.E-02	6.E-01	7.E-01	1.E-03	1.E-01	1.E-01	3.E-01	9.E-02	9.E-02	1.E-01	5.E-01
Turn correlation	NS	NS	NS	NS	NS	POS	NS	NS	NS	NEG	NS	NS	POS	NS	NS	NS	NS	NS	NS	NS

A-Y: One letter code of amino acid, POS: positive correlation, NEG: negative correlation, NS: not significant, linear regression analyses, n =79

Fig. 9: Correlation between frequency of individual amino acids and the main secondary structure elements in proteins. See results for explanation

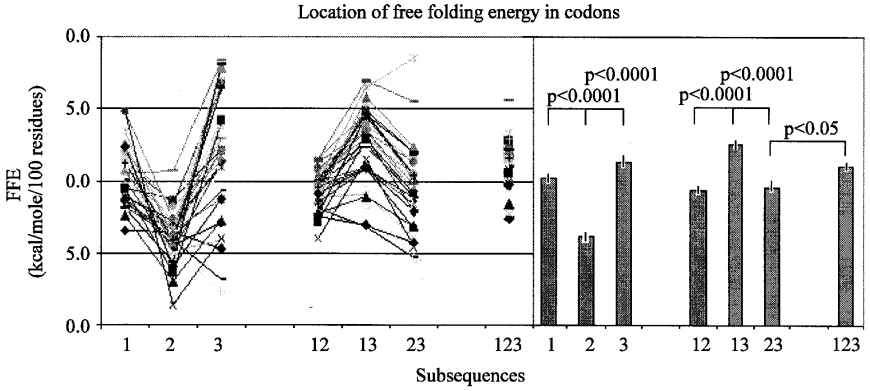

Fig. 10: Location of free folding energy in codons. Free Folding Energies (FFE) were determined in phase-selected subsequences of 31 different protein coding nucleic acids. The original intact RNA contained the intact three-letter codons (123). Subsequences were constructed by periodical removal of one letter from the codon and maintaining the other two (12, 13, 23) or removing two letters and maintaining only one (1, 2, 3). The lines indicate individual values (left), while the bars (right) indicate the mean±SEM (n=31)

sequence of their mRNA. These four proteins are examples of the fact that the tertiary structure of proteins is much more conserved than the amino acid sequence. We asked the question whether this is true for the RNA structures and sequence? We found that there are signs of conservation even of the RNA secondary structure (as indicated by the energy dot plots) and there are similarities between the protein and nucleic acid structures (Fig. 13).

Comparisons of the protein residue contact map with the nucleic acid folding maps suggest similarities between the 3D structures of these different kinds of molecules. However, this is a semi-quantitative method.

A more direct statistical support might be obtained by analyzing and comparing residue co-locations in these structures. Assume that the structural unit of mRNA is a tri-nucleotide (codon) and the structural unit of the protein is the amino acid. The codon may form a secondary structure by interacting with other codons accordingly to the WC base complementary rules and contribute to the formation of a local double helix. The 5′-A1U2G3-3′ sequence (Met, M codon) forms a perfect double string with the 3′-U3A2C1-5′ sequence (His, H codon, reverse and complementary reading). Suboptimal complexes are 5′-A1X2G3-3′ partially complemented by 3′-U3X2C1-5′ (AAG, Lys; AUG, Met; AGG, Arg; ACG, Pro; and CAU, His; CUU, Leu; CGU, Arg; CCU, Pro, respectively).

Our experiments with FFE indicate that local nucleic acid structures are formed under this suboptimal condition, i.e., when the 1st and 3rd codon residues are complementary but the 2nd is not. If this is the study and there is a connection between nucleic acid and protein 3D structure, one might expect that the 4 amino acids coded by 5′-A1X2G3-3′ codons will preferentially co-locate with other 4 amino acids coded by 3′-U3X2C1-5′ codons. We have constructed 8 different complementary codon combinations and found that the codons of co-locating amino acids are often complementary at the 1st and 3rd positions and follow the D-1X3/RC-3X1 formula but not the 7 other formulas (Fig. 14 and 15).

These special amino acid pairs and their frequency are indicated and summarized in a matrix (Fig. 16).

DISCUSSION

It is well known that coding and non-coding DNA sequences (exon/intron) are different and this difference is somehow related to the asymmetry of the codons, i.e.. that the third codon letter (wobble) is poorly defined. Many Markov models have been formulated to find this

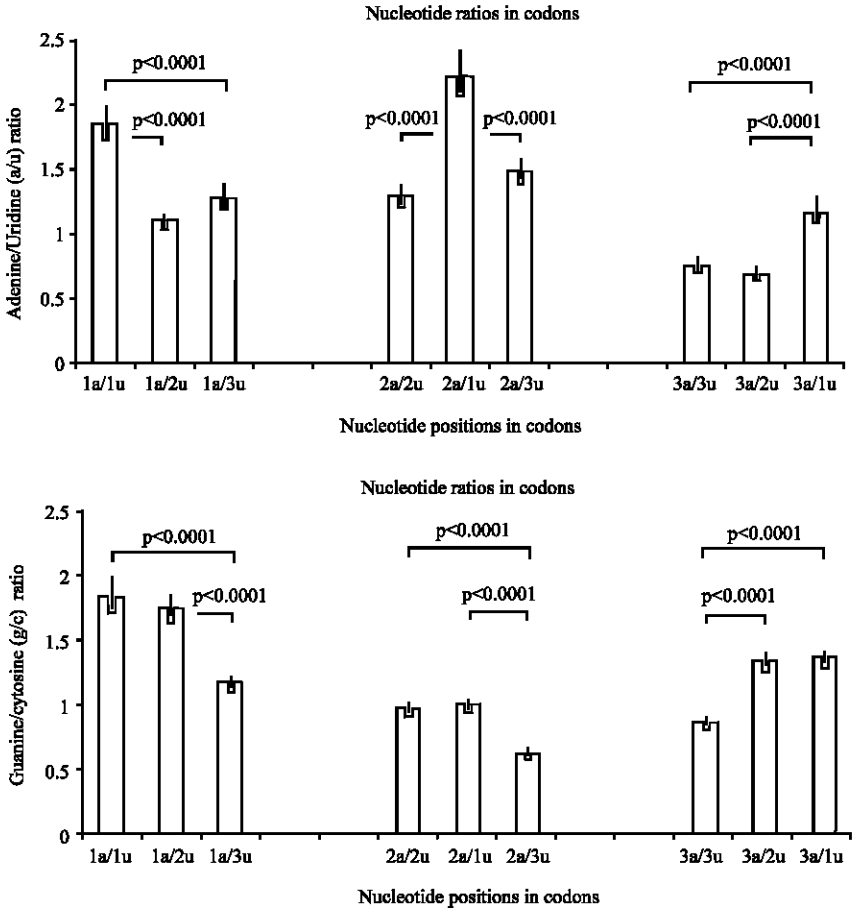

Fig. 11: Nucleotide ratios in codons. The number of the 4 different nucleotide bases was counted at the 1st, 2nd and 3rd codon positions in 30 different protein coding RNA sequences. The ratio of the Watson–Crick pairs at different codon positions are indicated by bars (±SEM, n=30). Ideally, the ratio of complementary base pairs is ~1.0. This ideal situation was mostly satisfied when one of the complementary bases was located at codon position 1 with the other at codon position 3 (pink) or both complements at codon position 2 (violet)

asymmetry and de novo predict coding sequences (genes).

These in silico methods work rather well but not perfectly and some scientists remain unconvinced that the codon asymmetry explains the exon-intron differences satisfactorily.

Another codon-related problem is that the well known, non-overlapping, triplet codon translation is extremely phase-dependent and there is theoretically no tolerance for any phase shift. There are famous examples of how single nucleotide deletion might destroy the meaningful translation of a sequence and which are incompatible with life. However, considering the magnitude and complexity of the eukaryotic proteome, the precision of translation is astonishingly good. Such physical precision is not possible withou massive and consistent physico-chemical fundaments. Therefore, discovery of the existence of secondary structure bias (folding energy differences) in coding regions of many organisms[13] was a very welcome observation because it differentiates exons from introns on a physico-chemical basis.

Our experiments with Free Folding Energy (FFE) confirmed that this bias exists. In addition, there is a very consistent and very significant pattern of FFE distribution along the nucleotide sequence. Comparing the FFE of phase-selected subsequences, subsequences comprised of only the 1st or only the 3rd codon letters showed significantly higher FFE than those consisting only of the 2nd letters. This FFE difference was not present in intronic sequences preceding and following the exons, but it was present in exons from different species including viruses. This is an interesting observation because this phenomena might not only distinguish between exons and introns on a physico-chemical basis,

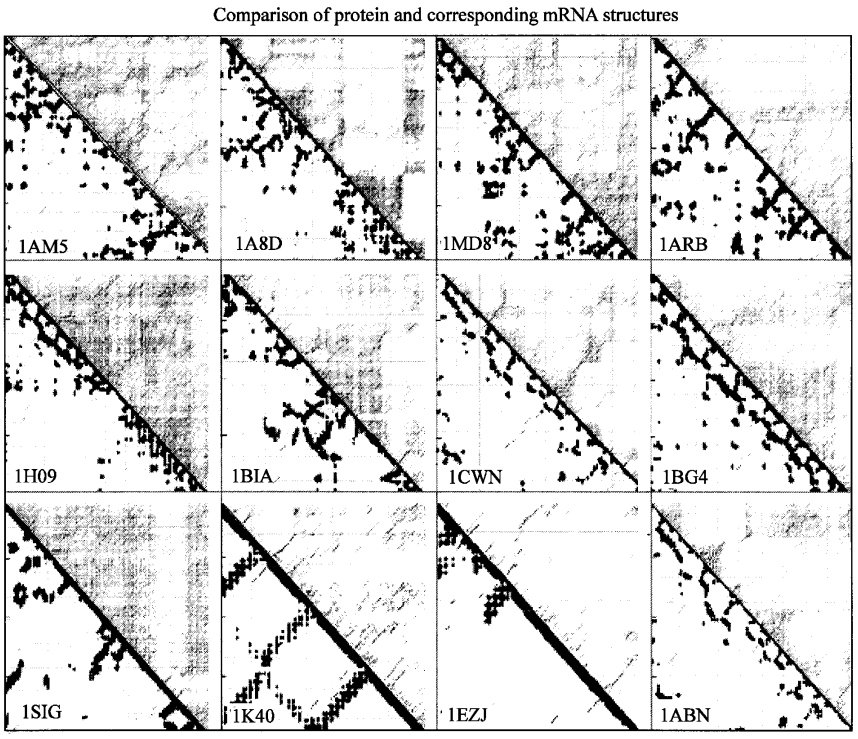

Fig. 12: Comparison of protein and corresponding mRNA structures. Residue Contact Maps (RCM) were obtained from the PBD files of protein structures using the SeqX tool (left triangles). Energy dot plots (EDP) for the coding sequences ware obtained using the mfold tool (right triangles). The two kinds of maps were aligned along a common left diagonal axis to make an easy visual comparison of the different kind of representations possible. The black dots in the RCMs indicate amino acids that are within 6 A 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)

but it might even clearly define the tri-nucleotide codons and thus the phase of the translation. This codon-related phase-specific variation in FFE may explain why mRNAs have greater negative free folding energies than shuffled or codon choice randomized sequences[23].

Free folding energy in nucleic acids is always associated with WC base pair formation. Higher FFE indicates more WC pairs (presence of complementarity) and lower FFE indicates fewer WC pairs (less complementarity). The FFE in the 1st and 3rd codon positions was additive, while the 2nd letter did not contribute to the total FFE; the total FFE of the entire (intact) nucleic acid was the same as subsequences containing only the 1st and 3rd codon letters (2nd deleted). This is an indication for that the local RNA secondary structure bias is caused by complementarity of the 1st and 3rd codon residues in local sequences. This partial, local complementarity is more optimal in reverse orientation of the local sequences as expected with loop formations.

It is known that single stranded RNA molecules can form local secondary structures through the interactions of complementary segments. The novel observation here is that these interactions preferentially involve the 1st and 3rd codon residues. This connection between the RNA secondary structure and codons immediately directed attention toward the question of protein folding and its long suspected connection to RNA folding[24,25].

Only about one-third (20/64) of the genetic code is used for protein coding, i.e., there is a great excess of information in the mRNA. At the same time, the information carried by amino acids seems to be insufficient (as stated by some scientists) to complete unambiguous protein folding. Therefore, it is believed that the third codon residue (wobble base) carries some additional information to that already present in the genetic code. A specialized wobble base oriented database, the ISSD[21], was established in an effort to connect different features of protein structure to wobble bases[26] with more or less success.

We found a significant negative correlation between FFE of the 2nd codon residue and the helix content of protein structures, which was not expectedeven though this possibility is mentioned in the literature[9]. Our

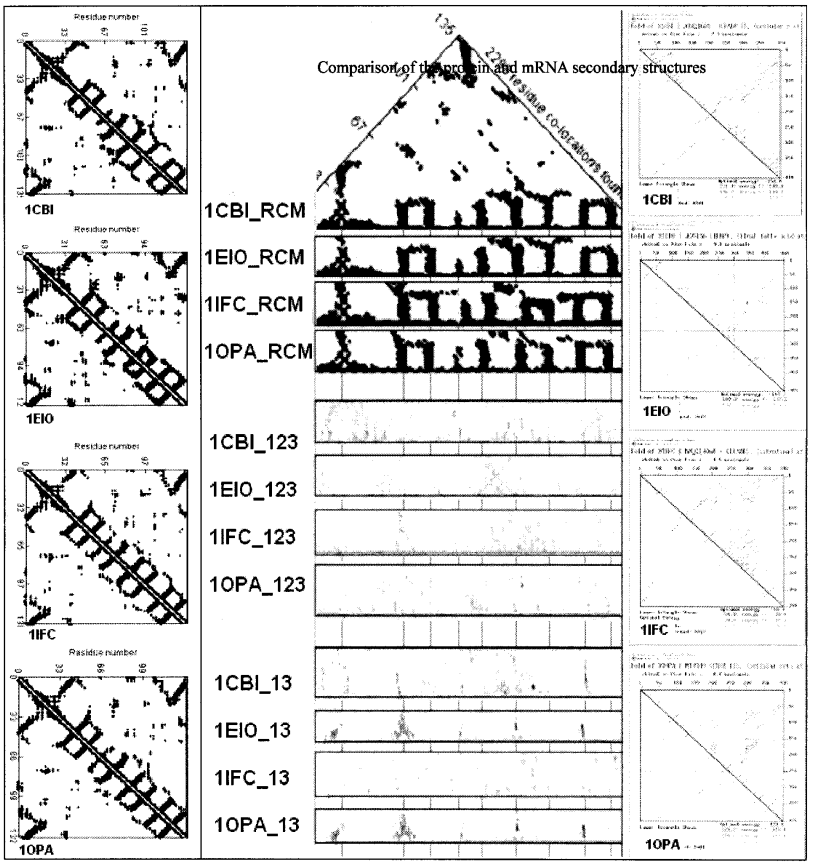

Fig. 13: Comparison of the protein and mRNA secondary structures. Residue contact maps (RCM) ware obtained from the PBD files of 4 protein structures (1CBI, 1EIO, 1IFC, 1OPA) using the SeqX tool (left column). Energy Dot Plots (EDP) for the coding sequences were obtained using the mfold tool (right column). The left diagonal portion of these two kinds of maps was compared in the central part of the figure. Blue horizontal lines in the background correspond to the main amino acid co-location sites in the RCM. Intact RNA (123) as well as subsequences containing only the 1st and 3rd codon letter (13) are compared. The black dots in the RCMs indicate amino acids that are within 6 A 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)

previous study on a Common Periodic Table of Codons and Nucleic Acids[22] indicated that the second codon residue is intimately coupled with the known physico-chemical properties of the amino acids. Almost all amino acids show significant positive or negative correlation to the helix content of proteins. Therefore, the real biological meaning and significance of any connection between FFE of the 2nd codon residue and the propensity of a protein structural element is highly questionable.

It was possible to make direct visual comparison of mRNA structure (as statistically predicted by mfold energy dot-plot) and protein structures (as 2D residue contact maps). This method suggests similarity between nucleic acid and protein structures. It is known that some complex protein structures are very similar even if there is less than 30% sequence similarity. It was interesting to see that the same principle might apply for nucleic acids and structural similarity might exist even when the sequence similarity is low. Furthermore, significant similarity between nucleic acid and protein structures might exist even without translational connection.

Structure seems to be more preserved, even in nucleic acids, than sequence.

However, even if the matrix comparisons are suggestive, they remain semi-quantitative methods. Better support was necessary.

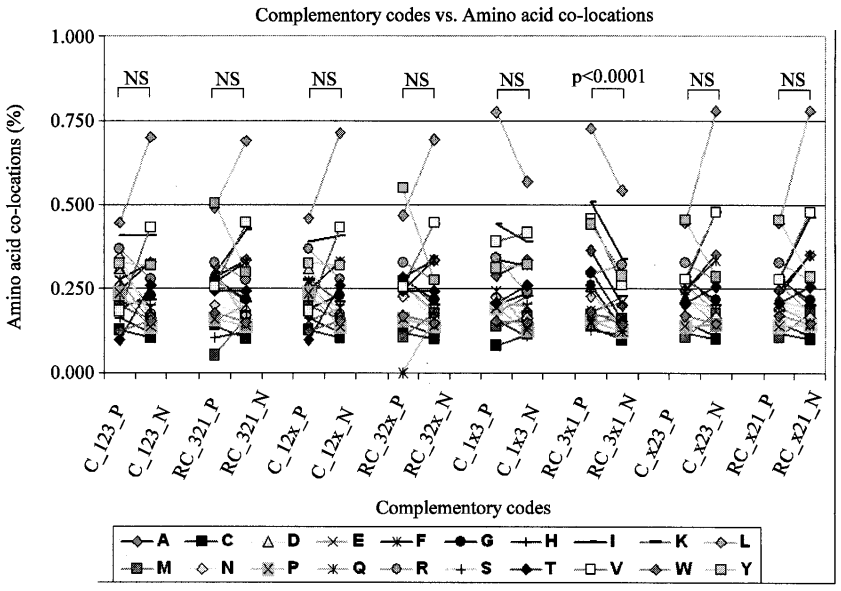

Fig. s14: Complementary codes vs. amino acid co-locations. The propensity of the 400 possible amino acid pairs was monitored in 81 different protein structures with the SeqX tool. The tool detected co-locations when two amino acids were within 6 A of each other (neighbors on the same strand were excluded). 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 the not the second and the possible codons are read in reverse orientation. The 400 co-locations were divided into 20 subgroups corresponding to 20 amino acids (one of the co-locating pairs), each group containing the 20 amino acids (corresponding to the other amino acid in the co-locating pair). If the codons of the amino acid pairs followed the predefined complementary code the co-location was regarded as positive (P); if not, the co-location was regarded as negative (N). Each symbol represents the mean frequency of P or N co-locations corresponding to the indicated amino acid. Paired Student's t-test, n=20

A working hypotheses grew out of these observations, namely that (a) partial, local reverse-complementarity exists in nucleic acids that form the nucleic acid structure; (b) there is some degree of similarity between the folding of nucleic acids and proteins; (c) protein structure determines the amino acid co-locations; (4) as a consequence, amino acids coded by the interacting (partially reverse complementary) codons might show preferential co-locations in the protein structures.

And it seems to be the study: codons which contain complementary bases at the 1st and 3rd positions and are translated in reverse orientation result in amino acids which are preferentially co-located (interacting) in the 3D protein structure. Other complementary residue combinations or translation in the same (not reverse) direction (as much as seven combinations in total) did not result in any preferentially co-locating subset of amino acid pairs.

Construction of residue contact maps for protein structures and statistical evaluation of residue co-locations is a frequently used method for visualization and analyses of spatial connections between amino acids[27-29]. The amino acid co-locations in real protein structures is clearly not random[30,31] and therefore residue co-location matrices are often used to assist in the prediction of novel protein structures[32,33]. We have carefully examined the physico-chemical properties of specifically interacting amino acids in and between protein structures and we concluded that these interactions follows the well known physico-chemical rules of size, charge and hydrophobe compatibility (unpublished data) well in line with Anfinsen's prediction. The recent study supports the fact that there is a previously unknown connection between the codons of specifically interacting amino acids; those codons are

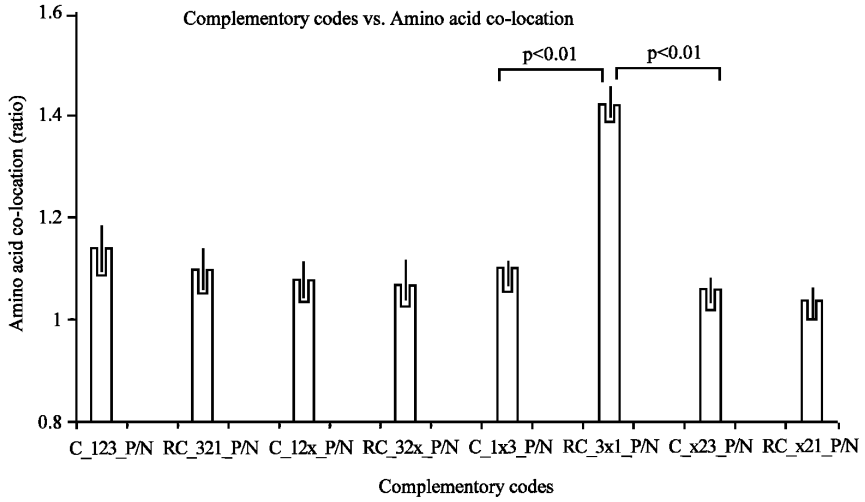

Fig. s15: Complementary codes vs. amino acid co-locations (ratios). The ratio of positive (P) and negative (N) co-locations was calculated on data from Fig. 13. Each bar represents the mean±SEM (n=20)

Complementory codes vs. Amino acid co-locations

SeqX 80

1st	2nd	3rd	AA	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
			1st	G	T	G	G	T	G	C	A	A	CT	A	A	C	C	AC	AT	A	G	T	T
			2nd	C	G	A	A	T	G	A	T	A	T	T	A	C	A	G	CG	C	T	G	A
			3rd	X	CT	CT	AG	CT	X	CT	ACT	AG	XAG	G	CT	X	AG	AGX	CTX	X	X	G	CT
G	C	X	A	7.5	1.4	2.6	2.9	4.1	3.9	1.5	6.3	3.3	9.9	2.6	2.6	2.0	2.5	3.3	3.7	4.5	7.5	1.6	4.7
T	G	CT	C	1.4	3.1	0.7	0.6	1.3	1.4	0.4	1.2	0.7	1.7	0.5	0.9	0.6	0.4	0.7	1.1	1.3	1.9	0.5	1.3
G	A	CT	D	2.6	0.7	1.6	1.3	1.8	2.7	1.4	2.3	3.1	3.6	0.8	1.5	1.3	1.3	3.5	2.3	2.7	3.0	1.0	2.2
G	A	AG	E	2.9	0.6	1.3	2.2	2.1	2.1	1.2	2.9	3.5	4.1	1.0	1.4	1.3	1.6	3.6	2.3	2.2	3.4	1.0	2.2
T	T	CT	F	4.1	1.3	1.6	2.1	3.3	2.8	1.3	4.8	2.4	6.5	1.9	1.8	2.2	1.7	3.3	2.4	2.4	4.9	1.6	3.0
G	G	X	G	3.9	1.4	2.7	2.1	2.8	4.2	1.3	4.2	2.3	5.2	1.4	2.3	2.1	1.7	3.0	3.0	3.3	4.4	1.3	2.7
C	A	CT	H	1.5	0.4	1.4	1.2	1.3	1.3	0.6	1.5	0.8	2.4	0.5	0.7	0.9	0.7	1.2	1.3	1.2	1.5	0.6	1.3
A	T	ACT	I	6.3	1.2	2.3	2.9	4.8	4.2	1.5	7.0	2.7	12.2	2.5	2.1	1.8	2.2	3.1	3.2	3.7	7.7	1.8	4.4
A	A	AG	K	3.3	0.7	3.1	3.5	2.4	2.9	0.8	2.7	1.5	4.2	1.3	1.5	1.3	1.7	1.7	1.9	2.9	3.4	1.0	2.5
CT	T	XAG	L	9.9	1.7	3.6	4.1	6.5	5.2	2.4	12.2	4.2	19.7	3.9	3.7	3.2	3.5	5.8	4.4	5.0	11.2	3.3	6.4
A	T	G	M	2.8	0.5	0.8	1.0	1.9	1.4	0.5	2.5	1.3	3.9	1.5	0.6	0.9	0.9	1.5	1.2	1.2	2.6	0.6	1.8
A	A	CT	N	2.6	0.9	1.5	1.4	1.9	2.3	0.7	2.1	1.6	3.7	0.8	1.6	1.5	1.4	1.8	2.2	2.1	2.1	1.0	1.9
C	C	X	P	2.0	0.8	1.3	1.3	2.2	2.1	0.9	1.8	1.3	3.2	0.9	1.5	1.5	1.0	2.2	1.4	1.6	2.8	1.2	2.1
C	A	AG	Q	2.5	0.4	1.3	1.6	1.7	1.7	0.7	2.2	1.7	3.5	0.9	1.4	1.0	0.9	1.6	1.3	1.7	2.5	0.7	1.6
AC	G	AGX	R	3.3	0.7	3.6	3.6	3.3	3.0	1.2	3.1	1.7	5.8	1.5	1.8	2.2	1.6	2.4	2.5	2.6	3.7	0.8	2.5
AT	CG	CTX	S	3.7	1.1	2.3	2.3	2.4	3.0	1.3	3.2	1.9	4.4	1.2	2.2	1.4	1.3	2.5	2.5	2.6	3.3	1.2	2.3
A	C	X	T	4.5	1.3	2.7	2.2	2.4	3.3	1.2	3.7	2.3	5.0	1.2	2.1	1.6	1.7	2.6	2.6	3.8	4.4	0.9	2.5
G	T	X	V	7.5	1.9	3.0	3.4	4.9	4.4	1.5	7.7	3.4	11.2	2.8	2.1	2.8	2.5	3.7	3.3	4.4	8.7	2.0	4.6
T	G	G	W	1.6	0.5	1.0	1.0	1.6	1.3	0.5	1.8	1.0	3.3	0.6	1.0	1.2	0.7	0.8	1.2	0.9	2.0	0.9	1.5
T	A	CT	Y	4.7	1.3	2.2	2.2	3.0	2.7	1.3	4.4	2.5	6.4	1.8	1.9	2.1	1.6	2.5	2.3	2.5	4.6	1.5	3.6

RC_3X1 code: 1st and 3rd codon letters are complementary in reverse order, indicated by complementary colors (red, blue); unfavored colocations are gray-colored.
X: any residue; AA: amino acids, one-letter code; Co-location frequency is multiplied by 10, Sum=1000

Fig. 16: Complementary codes vs. amino acid co-locations. See results for explanation

complementary at the 1st and 3rd (but not the 2nd) codon positions.

The idea that sequence complementarity might explain the nature of specific protein-protein interactions is not new and was suggested already in 1981[34].

I was never able to experimentally confirm my own original theory, which suggested a perfect complementarity between codons of interacting amino acids[34,35], in contrast to others[36]. The explanation is that this codon complementarity is suboptimal and does not involve the 2nd codon residue. Experimental in vitro confirmation is required to validate this recent theoretical and in silico prediction.

Availability: http://www.janbiro.com/downloads: SeqX, SeqForm, SeqPlot, Dotlet.

REFERENCES

1. Anfinsen, C.B., R.R. Redfield, W.I. Choate, J. Page and W.R. Carroll, 1954. Studies on the gross structure, cross-linkages and terminal sequences in ribonuclease. J. Bio. Chem., 207: 201-210.

2. Levinthal, C., 1969. How to fold graciously in Mossbauer spectroscopy in biological systems. In proceedings of a meeting held at allerton house, monticello, IL. Edited by Debrunner P, Tsibris JCM, Munck E. Urbana, IL: University of Illinois Press, pp: 22-24.
3. Klepeis, J.L. and A.C. Floudas, 2003. ASTRA-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. Biochem. J., 85: 2119-2146.
4. Walter, S. and J. Buchner, 2002. Molecular chaperones-cellular machines for protein folding. Angew Chem. Intl. Ed. Engl., 41: 1098-1113.
5. Komar, A.A., A. Kommer, I.A. Krasheninnikov and A.S. Spirin, 1997. Cotranslational folding of globin. J. Biol. Chem., 272: 10646-10651.
6. Thanaraj, T.A. and P. Argos, 1996. Protein secondary structural types are differentially coded on messenger RNA. Protein Sci., 5: 1973-1983.
7. Brunak, S. and J. Engelbrecht, 1996. Protein structure and the sequential structure of mRNA: Alpha-helix and beta-sheet signals at the nucleotide level. Proteins, 25: 237-252.
8. Gupta, S.K., S. Majumdar, T.K. Bhattacharya and T.C. Ghosh, 2000. Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem. Biophys. Res. Commun., 269: 692-696.
9. Chiusano, M.L., F. Alvarez-Valin, M. Di Giulio, G. D'Onofrio, G. Ammirato, G. Colonna and G: Bernardi, 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene., 261: 63-69.
10. Gu, W., T. Zhou, J. Ma, X. Sun and Z. Lu, 2004. The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. Biosystems, 73: 89-97.
11. Ermolaeva, O., 2001. Synonymous codon usage in bacteria. Curr. Issues Mole. Biol., 3: 91-97.
12. Biro, J.C., J.M. Biro and A.M. Biro, 2005. Hidden massages in hidden sub-sequences: a study on collagens. In 30th FEBS Congress-9th IUBMB Conference, Budapest, Hungary.
13. Katz, L. and C.B. Burge, 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome. Res., 13: 2042-2051.
14. Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31: 3406-3415.
15. Biro, J.C. and G. Fordos, 2005. SeqX: A tool to detect, analyze and visualize residue co-locations in protein and nucleic acid structures. BMC Bioinformatics 2005 [http://www.janbiro.com/downloads].
16. Biro, J.C., 2005. SeqForm. 2005 [http://www.janbiro.com/downloads].
17. Biro, J.C., 2005. SeqPlot. 2005 [http://www.janbiro.com/downloads].
18. Junier, T. and M. Pagni, 2000. Dotlet: Diagonal plots in a web browser. Bioinformatics, 16: 178-179 [http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html].
19. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, 2000. The Protein Data Bank. Nucleic Acids Res., 28: 235-242 [http://www.pdb.org/].
20. Berman, H.M., W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.H. Hsieh, A.R. Srinivasan and B. Schneider, 1992. The Nucleic Acid Database: A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J., 63: 751-759 [http://ndbserver.rutgers.edu/index.html].
21. Adzhubei, I.A. and A.A. Adzhubei, 1999. ISSD Version 2.0: Taxonomic range extended. Nucleic Acids Res., 27: 268-271 [http://www.protein.bio.msu.su/issd/].
22. Biro, J.C., B. Benyo, C. Sansom, A. Szlavecz, G. Fordos, T. Micsik and Z. Benyo, 2003. A common periodic table of codons and amino acids. Biochem. Biophys. Res. Commun., 306: 408-415.
23. Seffens, W. and D. Digby, 1999. mRNA have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res., 27: 1578-1584.
24. Oresic, M., M. Dehn, D. Korenblum and D. Shalloway, 2003. Tracing specific synonymous codon-secondary structure correlations through evolution. J. Mole. Evol., 56: 473-4840.
25. D'Onofrio, G., T.C. Ghosh and G. Bernardi, 2002. The base composition of the genes is correlated with the secondary structures of the encoded proteins. Gene., 300: 179-187.
26. Xie, T. and D. Ding, 1998. The relationship between synonymous codon usage and protein structure. FEBS Lett., 434: 93-96.
27. Kumarevel, T.S., M.M. Gromiha and M.N. Ponnuswamy, 2001. Distribution of amino acid residues and residue-residue contacts in molecular chaperons. Prep. Biochem. Biotech., 31: 163-183.
28. Eilers, M., A.B. Patel, W. Liu and S.O. Smith, 2002. Comparison of helix interactions in membrane and soluble alpha-bundle proteins. Biochem. J., 82: 2720-2736.

29. Glaser, F., D.M. Steinberg, I.A. Vakser and N. Ben-Tal, 2001. Residue frequencies at protein-protein Interfaces. Proteins Struct. Funct. Genet., 43: 89-102.
30. Naor, D., D. Fisher, R.L. Jernigan, H. Wolfson and R. Nussinov, 1996. Amino acid pair interchanges at spatially conserved locations. J. Mole. Biol., 256: 924-938.
31. Azarya-Sprinzak, E., D. Naor, H.J. Wolfson and R. Nussinov, 1997. Interchanges of spatially neighboring residues in structurally conserved environment. Protein Eng., 10: 1109-1122.
32. Singer, M.S., G. Vriend and R.P. Bywater, 2002. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Eng., 15: 721-725.
33. Shao, Y. and C. Bystroff, 2003. Predicting inter-residue contacts using templates and pathways. Proteins Struct. Funct. Genet., 53: 497-502.
34. Biro, 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.
35. Segersteen, U., H. Nordgren and J.C. Biro, 1986. Frequent occurrence of short complementary sequences in nucleic acids. Biochem. Biophys. Res. Commun., 139: 94-101.
36. Hela, J.R., G.W. Roberts, J.G. Raynes, A. Bhakoo and A.D. Miller, 2002. Specific interactions between sense and complementary peptides: the basics for the proteomic code. Chembiochem., 3: 136-151.